Высшее образование

В. Н. Калинина

Теория вероятностей и математическая статистика

Компьютерно-ориентированный курс

ДРОФА

Высшее образование

В. Н. Калинина

Теория вероятностей и математическая статистика

Компьютерно-ориентированный курс

Учебное пособие для высших учебных заведений

Допущено Советом Учебно-методического объединения ВУЗов России по образованию в области менеджмента в качестве учебного пособия по специальности «Менеджмент организации»

ДРОФА

УДК 519.2(075.32)
ББК 22.17я723
К17

Рецензенты:

д-р эконом. наук, проф. *В. С. Мхитарян*
(зав. кафедрой математической статистики и эконометрики
Московского государственного университета экономики,
статистики и информатики);

д-р эконом. наук, проф. *К. А. Раицкий*
(декан факультета экономики и менеджмента
Института гуманитарного образования)

Калинина, В. Н.

К17 Теория вероятностей и математическая статистика. Компьютерно-ориентированный курс : учеб. пособие для вузов / В. Н. Калинина. — М. : Дрофа, 2008. — 471, [9] с. : ил.

ISBN 978-5-358-04757-0

Изложенный в пособии материал соответствует требованиям к содержанию и уровню математического образования, приведенным в Государственном образовательном стандарте высшего профессионального образования по специальности «Менеджмент организации».

Учтены особенности подготовки бакалавров и специалистов по направлению «Менеджмент» и другим направлениям. Многочисленные примеры и задачи раскрывают возможности использования вероятностных и математико-статистических методов в управлении и экономике; содержатся рекомендации по применению Microsoft Excel.

Для студентов вузов, обучающихся по программам подготовки бакалавров и специалистов. Может быть использовано студентами образовательных учреждений среднего профессионального образования, учителями и учащимися лицеев и колледжей.

УДК 519.2(075.32)
ББК 22.17я723

ISBN 978-5-358-04757-0

ПРЕДИСЛОВИЕ

При написании учебного пособия автор руководствовался следующими целями:

— изложить в доступной для студентов форме вероятностные и математико-статистические модели и методы, широко используемые на практике;

— проиллюстрировать рассмотренные методы и их возможности на задачах, возникающих в различных областях деятельности, причем особое внимание уделить интерпретации результатов применения методов;

— ориентировать студентов при решении задач на использование процессора Microsoft Excel. В книге множество иллюстраций использования вероятностных и математико-статистических методов и команд Microsoft Excel при решении управленческих и экономических задач: управления финансовыми операциями, имитации работы системы управления запасами, сравнении результатов управленческих решений, оценивании влияния тех или иных факторов на результат деятельности.

Пособие состоит из четырех частей:

Часть 1. Случайные события и их вероятности.

Часть 2. Случайные величины и модели законов распределения вероятностей.

Часть 3. Изучение случайной величины по результатам наблюдений.

Часть 4. Изучение зависимостей.

В первой и второй частях рассмотрены вероятностные модели случайных событий и случайных величин (одномерных), с помощью предельных теорем обоснована возможность изучения случайных событий и случайных величин по результатам наблюдений. Математико-статистические методы такого изучения изложены в третьей части. В четвертой части при рассмотрении зависимостей между величинами параллельно излагаются вероятностные и математико-статистические модели и методы. Такая последовательность изложения, как показал многолетний опыт

преподавания автором «Теории вероятностей и математической статистики» в Государственном университете управления, позволяет проследить аналогии, имеющие место при вероятностном и статистическом изучении случайных величин, показать неразрывность связи теории вероятностей и математической статистики. В пособии сделан акцент на следующих важных теоретических положениях: свойстве статистической устойчивости и условиях закона больших чисел; механизме формирования значений случайной величины с тем или иным законом распределения; условиях использования той или иной модели; понятиях стохастической и корреляционной зависимости, функции регрессии; свойствах коэффициента корреляции и корреляционного отношения, позволяющих четко сформулировать последовательность ответов на вопросы о существовании корреляционной зависимости и ее виде.

Все выводы и доказательства, приводящие к определенному результату, а также замечания набраны петитом. Начало и конец выводов и доказательств отмечены соответственно знаками ≫ и ≪; начало и конец примеров — знаками ▷ и ◁; начало и конец задач — знаками ⏩ и ⏪.

Автор выражает искреннюю признательность рецензентам: доктору экономических наук, профессору В. С. Мхитаряну и доктору экономических наук, профессору К. А. Раицкому, а также И. Ю. Айрапетян, В. И. Зариповой, Т. С. Лобановой, Е. А. Лобовой, К. П. Марковой, М. А. Скворцовой, принявшим участие в подготовке рукописи к изданию.

Автор

ВВЕДЕНИЕ

События в материальном мире можно разбить на три категории: достоверные, невозможные и случайные. Например, если бросать игральную кость[1], то достоверно, что число выпавших очков — натуральное число; невозможно, чтобы это число было равно 7, и возможно, что оно равно 4. Однако в одних случаях оно равно 4, а в других выпадают числа 1, 2, 3, 5, 6. Событие, состоящее в том, что при бросании игральной кости выпадет 4, — случайное событие. Потребности практики привели математиков к изучению случайных событий, случайных явлений (в подтверждение далее рассмотрены соответствующие примеры).

Но как можно изучать случайные явления, если нельзя заранее знать их исход? Действительно, предвидеть результат единичного бросания игральной кости не представляется возможным, но оказывается, что при многократном бросании примерно в одинаковых условиях проявляются некоторые закономерности: маловероятно, что во всех испытаниях выпадет 1, гораздо вероятнее, что значения 1, 2, 3, 4, 5, 6 будут появляться примерно с одинаковой частотой и т. д. Если результат отдельного наблюдения случайного явления предсказать нельзя, а результаты многих наблюдений этого явления в типичных условиях как бы теряют свойства случайности, т. е. становятся предсказуемыми с большой степенью надежности, то говорят, что такое случайное явление обладает свойством ***статистической устойчивости***. Проиллюстрируем это свойство двумя примерами из финансового рынка.

• Отклонения курсовой стоимости акции от ее номинальной стоимости даже при устойчивой экономике подвержены случайным колебаниям, и предвидеть конкретное значение отклонения нельзя. Однако если определенным способом систематизировать достаточно большое

[1] Игральная кость — однородный кубик, грани которого размечены цифрами от 1 до 6; обычно сумма противоположных граней равна 7.

число этих отклонений, то в их изменении можно найти некоторую закономерность, зная которую можно рассчитать вероятность того, что это отклонение будет лежать в том или ином интервале.

• Инвестор формирует портфель ценных бумаг. Допустим, что он вложит деньги в акции только одной компании. Даже при устойчивой экономике предсказать эффективность вложений на момент заключения сделки нельзя — она случайна и зависит от колебаний курсовой стоимости акций. Однако если инвестор вложит деньги в акции нескольких компаний, то эффективность вложений также зависит от курсовых колебаний, но не столько от колебания каждого курса, сколько от усредненного колебания. Средний же курс, как правило, колеблется меньше, поскольку при повышении курса акций одних компаний курс акций других компаний может понизиться и колебания могут взаимно погаситься. При увеличении числа компаний-акционеров средний курс как бы теряет свойства случайности, становится статистически устойчивым. Именно поэтому опытный инвестор является держателем не одного вида ценных бумаг, а нескольких (акции разных компаний, векселя, контракты и т. д.).

Условия, при которых случайное явление обладает статистической устойчивостью, впервые сформулированы Я. Бернулли (XVII, XVIII вв.), впоследствии подтверждены и развиты в работах А. Муавра, П. Лапласа, К. Гаусса, С. Пуассона (XVIII, XIX вв.), а затем в работах русских математиков А. М. Ляпунова, А. А. Маркова, П. Л. Чебышёва, С. Н. Бернштейна, В. И. Романовского, А. Н. Колмогорова, А. Я. Хинчина, Б. В. Гнеденко (XIX, XX вв.).

Математические модели случайных явлений, обладающих статистической устойчивостью, изучает теория вероятностей; предметом математической статистики является изучение таких явлений по результатам наблюдений. Математическая статистика неразрывно связана с теорией вероятностей: если теория вероятностей предоставляет исследователю набор математических моделей случайных явлений, то методы математической статистики позволяют среди множества возможных теоретико-вероятностных моделей выбрать такую, которая наилучшим образом соответствует имеющимся в распоряжении исследователя результатам наблюдений. Правомерность принятия модели в качестве закономерности явления может подтвердить только практика ее дальнейшего использования.

Знание теории вероятностей и математической статистики не только свидетельствует об эрудиции современного человека, но и совершенно необходимо для его успешной

деятельности в любой области науки и практики. Приведем несколько задач, решение которых базируется на теории вероятностей и математической статистике.

• Два инвестора решили приобрести одинаковое количество акций определенной фирмы. Один из них вкладывает в покупку акций часть капитала, оставив остальную часть в банке. Другой, не имея капитала, берет деньги взаймы под некоторый процент под залог. Интуитивно ясно, что меньше рискует второй. Как количественно оценить степень риска каждого?

• Фирма, занимающаяся рыночными исследованиями, должна установить степень известности некоторого моющего средства среди жителей крупного города. Излишне опрашивать каждого жителя, если было бы можно надежно определить степень известности этого средства во всем городе, выбрав для опроса лишь небольшое количество жителей, например 100 человек. Каким должен быть объем выборки? Как предсказать степень известности моющего средства среди жителей всего города, зная ее только для выборки? Каковы ошибка и надежность предсказания?

• Главным «барометром», позволяющим предсказывать судьбу ценных бумаг, является ситуация на рынке в целом, которая отражается суммой курсов важнейших видов ценных бумаг, взвешенных с учетом акционерного капитала каждой корпорации. Поэтому стоимость важнейших ценных бумаг — объект постоянных наблюдений. Однако финансовый аналитик обязан давать рекомендации не только по немногим ведущим, но и по возможно большему числу компаний, выдвигающих свои ценные бумаги на рынок. Какими показателями измерить влияние ситуации на рынке в целом на курсовую стоимость ценных бумаг? Как предсказать курсовую стоимость ценных бумаг, предвидя ситуацию на рынке в целом?

• Из криминологии известно, что динамика преступности связана с динамикой социально-демографических факторов. Как построить математическую модель, отражающую эту связь, и, используя ее, оценить степень влияния различных факторов на преступность?

На эти и ряд других вопросов дают ответы теория вероятностей и математическая статистика.

Случайные события и их вероятности

ГЛАВА 1

Понятие вероятности

В главе кратко рассмотрена эволюция теории вероятностей, начиная с классической схемы равновозможных исходов. На первых порах разработка теории вероятностей была обусловлена интересом математиков XVII в. к исследованию азартных игр. Примеры из азартных игр просты и позволяют продемонстрировать основные положения теории вероятностей. Именно этим объясняется то, что в этой главе большинство примеров относится к азартным играм.

Вводятся важнейшие понятия теории вероятностей. Большое внимание уделено использованию комбинаторных формул при нахождении вероятностей.

§ 1.1. Виды случайных событий. Дискретное множество элементарных событий. Множество исходов опыта

Пусть многократно проводится некоторый опыт (испытание, эксперимент, наблюдение) и можно считать типичные условия проведения опыта не изменяющимися при его повторении. В результате опыта происходят или не происходят те или иные события.

Определения. *Событие, которое каждый раз происходит в опыте при его повторении, называется достоверным в этом опыте.*

Событие, которое не наступает никогда в опыте, сколько бы раз этот опыт ни повторили, называется невозможным.

Событие, которое то происходит, то не происходит при повторении опыта, называется случайным.

▶ **ПРИМЕР 1.1.** Опыт: одиночное подбрасывание монеты. Случайными событиями являются: выпадение «герба» (Г), выпадение цифры (Ц). Событие «невыпадение ни «герба»,

ни цифры» можно считать невозможным, если, конечно, условия опыта исключают потерю монеты или падение ее на ребро. Событие «выпадение «герба» или цифры» достоверное.

ПРИМЕР 1.2. Опыт: одиночное подбрасывание игральной кости. Случайными событиями являются, например, выпадение трех очков, выпадение четного числа очков, выпадение числа, не превышающего пяти, и т. д. Событие: выпадение семи очков — невозможное; событие: выпадение четного или нечетного числа — достоверное. ◀

Один и тот же результат опыта может означать одновременно несколько случайных событий. Так, если при бросании игральной кости выпадет три очка, то одновременно произойдут такие события: выпадет нечетное число очков; выпадет число очков, не превышающее трех; выпадет число очков между двумя и пятью; выпадет число очков, не меньшее трех, и т. д. Эти события не исключают друг друга и поэтому могут произойти одновременно. Напротив, при бросании монеты события: выпадение «герба» и выпадение цифры, не могут произойти одновременно. Также при бросании игральной кости не могут произойти одновременно события: выпадение числа очков, не превышающего двух; и выпадение числа очков, не меньшего четырех.

Определения. *Два события, которые одновременно не могут произойти в одном и том же опыте, называются несовместными.*

Три или более событий называются попарно несовместными, если никакие два из них не могут произойти одновременно в одном и том же опыте.

Обратим внимание, что, если в группе событий никакие два события не могут произойти одновременно, то и никакие три, и никакие четыре и т. д. не могут произойти одновременно. Таким образом из попарной несовместности вытекает несовместность событий «по три», «по четыре» и т. д. Однако из несовместности событий «по три», «по четыре» и т. д. вовсе не следует их попарная несовместность.

Определение. *Группа событий, связанных с одним и тем же опытом, хотя бы одно (по крайней мере одно, не менее одного) из которых обязательно происходит в опыте, называется полной группой событий.*

При бросании монеты одно из событий: выпадение «герба» или выпадение цифры, обязательно происходит; эти события образуют полную группу. Выпадение 1, 2, 3, 4, 5, 6

очков при бросании кости — также полная группа событий.

Особо следует выделить *полную группу попарно несовместных событий*. Такой группой является: выпадение «герба» и выпадение цифры при подбрасывании монеты; выпадение 1, 2, 3, 4, 5, 6 очков при бросании кости. События: выпадение числа очков, большего двух; выпадение числа очков, меньшего пяти, не образуют полной группы несовместных событий, хотя и образуют просто полную группу.

Среди приведенных примеров случайных событий есть события, которые можно разложить на более простые. Так, если при бросании игральной кости выпадет четное число очков, то это означает, что произойдет одно из трех событий: выпадет двойка, выпадет четверка, выпадет шестерка. Напротив, каждое из шести событий: выпадет единица, выпадет двойка и т. д., является неразложимым событием. Неразложимое событие естественно назвать ***элементарным***. Очевидно, что элементарные события попарно несовместны в одном и том же опыте, и группа всех элементарных событий, которые могут произойти в опыте, является полной.

Пусть опыт имеет конечное число элементарных событий ω_i или бесконечное, но счетное число элементарных событий (последнее означает, что хотя элементарных событий и бесконечно много, однако они поддаются пересчету). В каждом из этих случаев множество всех элементарных событий называют ***дискретным*** и обозначают так:

$$\Omega = \{\omega_i\}, \text{ где } i = \begin{cases} 1, 2, \ldots, N, \text{ если число событий конечно;} \\ 1, 2, \ldots, \text{ если число событий бесконечно, но счетно}^1. \end{cases}$$

Приведем примеры дискретных множеств элементарных событий.

▶ **ПРИМЕР 1.3.** Опыт: одиночное подбрасывание монеты (исключающее потерю монеты или падение ее на ребро). Элементарными событиями являются: ω_1 — выпадение «герба» (Г); ω_2 — выпадение цифры (Ц). Множество элементарных событий: $\Omega = \{\omega_1, \omega_2\}$, или $\Omega = \{\text{Г}, \text{Ц}\}$, конечно.

ПРИМЕР 1.4. Опыт: бросание монеты до первого выпадения «герба». Элементарные события здесь таковы:

ω_1 — выпадение «герба» при первом бросании монеты (Г);

[1] Ω (ω) — прописная (строчная) буква «омега» греческого алфавита.

ω_2 — выпадение цифры при первом бросании монеты и «герба» при втором бросании монеты (ЦГ);

ω_3 — выпадение цифры при первом и втором бросании монеты и «герба» при третьем бросании (ЦЦГ);

ω_4 — выпадение цифры при первом, втором и третьем бросании монеты и «герба» при четвертом бросании монеты (ЦЦЦГ) и т. д.

Множество элементарных событий $\Omega = \{\omega_i\}$, где $i = 1, 2, \ldots$ бесконечно, но счетно.

ПРИМЕР 1.5. Опыт: бросание наудачу (в § 1.2 смысл термина «наудачу» будет уточнен) точки на некоторый отрезок, при этом предполагается, что точка попадет на этот отрезок. В этом опыте множество элементарных событий — это множество всех точек отрезка: оно бесконечно и более чем счетно (нельзя пересчитать точки отрезка). Такое множество элементарных событий не является дискретным. ◀

Вернемся к рассмотрению дискретного множества элементарных событий. Любое событие, появляющееся в результате опыта S, можно рассматривать как некоторое множество, являющееся подмножеством множества Ω (напомним, что множество X является подмножеством множества Y, это записывается так: $X \subset Y$, если каждый элемент множества X является в то же время элементом множества Y).

▶ **ПРИМЕР 1.6.** При одиночном бросании монеты множество элементарных событий $\Omega = \{\omega_1, \omega_2\}$, где ω_1 — выпадение «герба», ω_2 — выпадение цифры.

События:

A — выпадение «герба» можно рассматривать как множество, состоящее из одного элемента ω_1: $A = \{\omega_1\}$, которое является подмножеством множества Ω: $A \subset \Omega$;

B — выпадение цифры можно рассматривать как множество, состоящее из одного элемента ω_2: $B = \{\omega_2\}$ и $B \subset \Omega$;

C — выпадение «герба» или цифры можно рассматривать как множество, состоящее из двух элементов ω_1 и ω_2: $\Omega = \{\omega_1, \omega_2\}$ — оно совпадает с Ω. ◀

Событие, соответствующее всему множеству Ω элементарных событий, является достоверным, поскольку оно наступает при наступлении любого из элементарных событий, входящих в Ω, а они образуют полную группу. Именно поэтому достоверное событие обозначают так же, как и множество элементарных событий, буквой Ω. Невозможное событие обозначают символом пустого множества: $\varnothing$. Условимся не только достоверное, но и невозможное собы-

тие считать подмножеством множества Ω элементарных событий.

Сколько всего событий, включая и достоверное, и невозможное события, можно связать с опытом S, множество элементарных событий которого $\Omega = \{\omega_i\}$, $i = 1, 2, \ldots, N$, конечно? Поскольку для каждого ω_i возможны два варианта: либо ω_i входит в число элементарных событий, влекущих появление задуманного события, либо не входит, и так как число элементарных событий равно N, общее число событий, связанных с опытом S, равно 2^N. Действительно, в примере 1.3 опыт состоял в подбрасывании монеты и $\Omega = \{\omega_1, \omega_2\}$, или $\Omega = \{\text{Г}, \text{Ц}\}$. С этим опытом связаны события: выпадение «герба», выпадение цифры, выпадение «герба» или цифры (это достоверное событие); добавив к ним невозможное событие, получим, что всего событий четыре, или 2^N, где $N = 2$ — число элементарных событий.

ПРИМЕР 1.7. При двукратном бросании монеты (или одновременном бросании двух монет) множество элементарных событий включает в себя четыре комбинации результатов при первом и втором бросании соответственно: $\Omega = \{\text{ЦЦ}, \text{ЦГ}, \text{ГЦ}, \text{ГГ}\}$. С этим опытом связано $2^4 = 16$ событий, включая и невозможное: {ЦЦ}, {ЦГ}, {ГЦ}, {ГГ}, {ЦЦ, ЦГ}, {ЦЦ, ГЦ}, {ЦЦ, ГГ}, {ЦГ, ГЦ}, {ЦГ, ГГ}, {ГЦ, ГГ}, {ЦЦ, ЦГ, ГЦ}, {ЦЦ, ЦГ, ГГ}, {ЦЦ, ГЦ, ГГ}, {ЦГ, ГЦ, ГГ}, Ω, $\varnothing$.

При трехкратном бросании монеты (или при одновременном бросании трех монет) множество элементарных событий включает $2^3 = 8$ событий: $\Omega = \{\text{ЦЦЦ}, \text{ГЦЦ}, \text{ЦГЦ}, \text{ЦЦГ}, \text{ГЦГ}, \text{ЦГГ}, \text{ГГЦ}, \text{ГГГ}\}$, а общее число событий, связанных с этим опытом, равно 2^8.

ПРИМЕР 1.8. При одиночном подбрасывании игральной кости множество элементарных событий $\Omega = \{\omega_1, \omega_2, \ldots, \omega_6\}$, где ω_i — выпадение i очков, $i = 1, 2, \ldots, 6$. С этим опытом связано $2^6 = 64$ события, включая и невозможное.

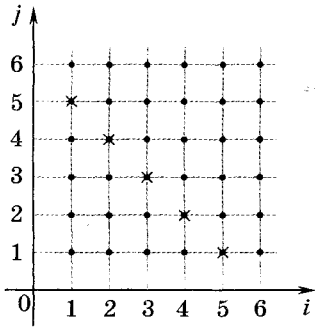


Рис. 1.1

ПРИМЕР 1.9. При одновременном однократном бросании двух игральных костей элементарными событиями являются упорядоченные пары (i, j) чисел, выпавших на костях (i — на первой, j — на второй кости). Всего таких пар будет $N = 6 \cdot 6 = 36$. На рисунке 1.1 множество пар (i, j) представлено на плоскости в виде совокупности точек с координатами i, j; множество элементарных событий $\Omega = \{(i, j)\}$,

$i, j = 1, 2, \ldots, 6$. Количество всех событий, включая и невозможное, равно 2^{36}. В их числе, например, имеется событие A, состоящее в том, что сумма выпавших на обеих костях очков равна 6. Этому событию соответствуют пять точек на рисунке 1.1, отмеченных «крестиком». Используя язык множеств, событие A можно записать так:

$$A = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}. \blacktriangleleft$$

Итак, если множество элементарных событий конечно, т. е. $\Omega = \{\omega_i\}$, $i = 1, 2, \ldots, N$, то число подмножеств множества Ω, включая само Ω и пустое множество $\varnothing$, так же как и число событий, связанных с этим опытом, включая достоверное и невозможное события, равно 2^N.

В случае, когда множество Ω элементарных событий опыта бесконечно (см. пример 1.4), бесконечными будут и множество подмножеств множества Ω, и множество событий, связанных с этим опытом.

В заключение подчеркнем следующее.

— Множество элементарных событий является исходным понятием теории вероятностей и представляет собой вероятностную модель реального опыта, при построении которой вводится ряд допущений. Так, уже в самом простом случае при бросании монеты предполагалось, что монета не может ни затеряться, ни встать на ребро, поэтому множество элементарных событий включает лишь два события: выпадение цифры и выпадение «герба».

— Множество $\Omega = \{\omega_i\}$ элементарных событий — это множество событий, неразложимых на более простые, попарно несовместных и образующих полную группу, используя которые можно любое событие, происходящее в опыте, представить как подмножество множества Ω. Однако если понятия попарной несовместности и полной группы вполне определены, то понятие неразложимости «расплывчато». Более простое решение ряда вероятностных задач можно получить, ориентируясь не на множество Ω неразложимых событий, а на множество попарно несовместных, образующих полную группу событий, которое обозначим через Ω'. Требование неразложимости событий при этом второстепенно, важно лишь, чтобы интересующее нас в опыте событие можно было представить как подмножество множества Ω'. Поясним сказанное примером.

▶ **ПРИМЕР 1.10.** Опыт: одиночное бросание игральной кости. В примере 1.8 было выяснено, что множество Ω элементарных (неразложимых) событий этого опыта состоит из шести чисел: $\Omega = \{1, 2, 3, 4, 5, 6\}$, и любому событию в этом опыте соответствует некоторое подмножество множества Ω

(всего таких подмножеств 2^6). В частности, событию A, состоящему в появлении четного числа очков, соответствует подмножество $A = \{2, 3, 6\}$ множества Ω.

Но этому же событию A соответствует подмножество $A = \{Ч\}$ множества $\Omega' = \{Ч, Н\}$, где Ч — выпадение при подбрасывании кости четного числа очков, Н — выпадение нечетного числа. Обратим внимание на то, что события Ч и Н образуют полную группу и несовместны в одном опыте, однако свойством неразложимости, в отличие от событий множества Ω, не обладают. Конечно, на множестве $\Omega' = \{Ч, Н\}$ нельзя изучать все события, которые могут произойти при бросании игральной кости и которые можно изучить на множестве $\Omega = \{1, 2, 3, 4, 5, 6\}$ неразложимых событий. В этом смысле множество Ω' «беднее» множества Ω. Однако событие A можно изучить как на множестве Ω, так и на более «бедном» множестве Ω'. ◄

Таким образом, если к элементарным событиям не предъявлять требования неразложимости, выбор множества событий при изучении некоторого опыта не однозначен: он зависит не только от самого опыта, но и от тех случайных событий, которые предполагается изучить в опыте.

Поэтому наряду с понятием «множество элементарных событий» введем понятие «множество исходов опыта» и, по определению, будем считать, что ***множество исходов опыта*** — это полная группа попарно несовместных событий, не обязательно неразложимых, таких, что любое интересующее нас событие в опыте является подмножеством множества исходов. Условимся в дальнейшем множество исходов опыта обозначать той же буквой Ω, что и множество элементарных событий.

§ 1.2. Вероятность исхода опыта и произвольного события. Классический, эмпирический и геометрический подходы к нахождению вероятности

Пусть множество исходов некоторого опыта S: $\Omega = \{\omega_i\}$ (напомним, что события ω_1, ω_2, ... попарно несовместны и образуют полную группу). Пусть A — произвольное событие в опыте S; событию A соответствует некоторое подмножество A множества Ω: $A = \{\omega_j\} \subset \Omega$.

Определения. *Вероятностями $P(\omega_i)$ исходов ω_i, принадлежащих множеству Ω, называют неотрицательные числа*

$$P(\omega_i) \geqslant 0 \qquad (1.1)$$

такие, сумма которых по всем исходам, составляющим множество Ω, *равна единице*:

$$\sum_{\omega_i \in \Omega} P(\omega_i) = 1^1. \tag{1.2}$$

Вероятностью события A, $A \subset \Omega$, *называют сумму вероятностей всех исходов, входящих в* A:

$$P(A) = \sum_{\omega_j \in A} P(\omega_j). \tag{1.3}$$

Нетрудно убедиться, что из соотношений (1.1) — (1.3) вытекает неравенство

$$0 \leqslant P(A) \leqslant 1. \tag{1.4}$$

В частном случае, когда A — достоверное событие (достоверное событие мы обозначили буквой Ω), получим

$$P(\Omega) \underset{(1.3)}{=} \sum_{\omega_i \in \Omega} P(\omega_i) \underset{(1.2)}{=} 1 \Rightarrow P(\Omega) = 1, \tag{1.5}$$

т. е. вероятность достоверного события равна единице; если A — невозможное событие (невозможное событие мы обозначили символом $\varnothing$), то

$$P(\varnothing) = 0, \tag{1.6}$$

т. е. вероятность невозможного события равна нулю.

1. Классическая формула вероятности события. Пусть множество исходов опыта *конечно*, т. е. $\Omega = \{\omega_i\}$, $i = 1, 2, \ldots, N$, и ***исходы равновозможны***, т. е. их вероятности равны друг другу: $P(\omega_1) = P(\omega_2) = \ldots = P(\omega_N)$. Тогда, учитывая соотношение (1.2), которое принимает вид $P(\omega_1) + P(\omega_2) + \ldots + P(\omega_N) = 1$, получим

$$P(\omega_i) = 1/N, \quad i = 1, 2, \ldots, N. \tag{1.7}$$

Далее, если событие A наступает, когда происходит один из M равновозможных исходов ω_i, то, согласно (1.3), имеем, что вероятность этого события

$$P(A) = M/N. \tag{1.8}$$

Формулу (1.8) читают так: «если множество Ω исходов опыта конечно и исходы равновозможны, то вероятность события A, определенного на множестве Ω, равна отношению числа M исходов, благоприятствующих наступлению события A, к общему числу N исходов». Формулу (1.8) называют ***классической формулой вероятности собы-***

[1] Прописная (строчная) буква P (p) латинского алфавита используется для обозначения вероятности (от англ. *probability* — вероятность).

тия. (Именно по этой формуле, предложенной Б. Паскалем в 1654 г., рассчитывали вероятности классики теории вероятностей на начальном этапе формирования теории.)

▶ **ПРИМЕР 1.6** (продолжение). При однократном бросании монеты исходов два ($N = 2$): ω_1 — выпадение «герба», ω_2 — выпадение цифры. В силу симметрии монеты, эти исходы равновозможны, вероятность каждого из них $P(\omega_1) = P(\omega_2) = 1/2$.

ПРИМЕР 1.7 (продолжение). При двукратном бросании монеты равновозможных исходов четыре ($N = 4$). Событию A, состоящему в появлении цифры хотя бы при одном бросании, благоприятствуют три ($M = 3$) исхода: ЦЦ, ЦГ, ГЦ, поэтому вероятность $P(A) = 3/4$.

ПРИМЕР 1.8 (продолжение). При однократном бросании игральной кости шесть ($N = 6$) равновозможных исходов. Событию A, состоящему в появлении не менее двух и не более пяти очков, благоприятствуют четыре ($M = 4$) исхода: ω_2, ω_3, ω_4, ω_5, поэтому $P(A) = 4/6 = 2/3$.

ПРИМЕР 1.9 (продолжение). При одновременном бросании двух игральных костей равновозможных исходов $N = 36$. Событию A, состоящему в том, что сумма выпавших очков на обеих костях равна 6, благоприятствуют $M = 5$ исходов, поэтому $P(A) = 5/36$. Событию B, которое состоит в том, что на обеих костях выпадет одинаковое число очков, благоприятствуют $M = 6$ исходов: (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), поэтому $P(B) = 6/36 = 1/6$.

ПРИМЕР 1.10 (продолжение). Событие A, состоящее в появлении четного числа очков при одиночном бросании игральной кости, может быть определено как на множестве равновозможных исходов $\Omega = \{1, 2, 3, 4, 5, 6\}$ и тогда $P(A) = 3/6$, так и на множестве равновозможных исходов $\Omega' = \{\text{Ч}, \text{Н}\}$ и тогда $P(A) = 1/2$.

Как и следовало ожидать, оба результата совпали. ◀

Обратим внимание на то, что классический подход к определению вероятности не требует проведения опыта. Однако обязательными условиями применения этого подхода являются конечность множества исходов опыта и их равновозможность, что существенно ограничивает область его применения.

2. Эмпирический подход к нахождению вероятности. Согласно эмпирическому (опытному или статистическому) подходу, вероятность находится после многократного повторения опыта S в типичных условиях. В основе подхода

лежит понятие ***относительной частоты***: если событие A появилось в m из n опытов, то относительная частота $\hat{p}(A)$ этого события равна

$$\hat{p}(A) = m/n.$$

Конечно, если A — случайное событие, то при переходе от одной серии n опытов к другой, так же как и при изменении самого числа n, относительная частота будет изменяться. Это подтверждают следующие результаты, полученные при подбрасывании монеты:

Экспериментатор	Число подбрасываний n	Число выпадений «герба» m	Относительная частота $\hat{p} = m/n$
Ж. Бюффон (1707—1788)	4 040	2 048	0,5069
К. Пирсон (1857—1936)	12 000	6 019	0,5016
К. Пирсон	24 000	12 012	0,5005

Относительная частота выпадения «герба» изменяется, но она близка к числу $p = 0{,}5$. В приведенных результатах оказалось, что чем больше n, тем ближе m/n к числу 0,5. Однако считать, что это всегда так, нельзя. Например, если монету вновь подбросить 24 000 раз, то относительная частота m/n может оказаться равной не 0,5005, как получил К. Пирсон, а 0,5020 и отстоять от 0,5 дальше, чем относительная частота 0,5016, полученная при 12 000 подбрасываний.

Я. Бернулли доказал теорему, согласно которой, *чем больше число опытов n, тем больше (при выполнении достаточно общих условий) уверенность в незначительном отклонении относительной частоты события от некоторого постоянного числа p, называемого вероятностью этого события.* (В § 6.2 будет доказана теорема Я. Бернулли.) Устойчивость (или малая «колеблимость» значений относительной частоты при большом числе испытаний) была подмечена во многих явлениях задолго до XVIII в. Так, еще в Древнем Китае было обнаружено, что для государств и больших городов отношение числа родившихся мальчиков к числу всех родившихся из года в год почти неизменно (чуть больше 0,5).

Из теоремы Я. Бернулли вытекает, что значение относительной частоты m/n при большом числе испытаний

предсказуемо, т. е. m/n обладает статистической устойчивостью, и именно поэтому относительную частоту $\hat{p} = m/n$ появления события в большом числе n испытаний можно принять за вероятность этого события; ее называют ***эмпирической,*** или ***опытной,*** или ***статистической вероятностью.*** Эмпирическая вероятность, так же как и вероятность, рассчитанная по классической формуле, удовлетворяет соотношению (1.4).

Эмпирический подход к нахождению вероятности требует проведения большого числа опытов. Но как можно, например, провести многократный запуск ракеты в определенную область G для того, чтобы определить вероятность попадания ракеты в некоторую подобласть g этой области? Для нахождения такой вероятности нельзя использовать и классический подход, предполагающий конечность множества исходов опыта, так как множество исходов опыта: запуск ракеты в область G — это множество всех точек области G (полагаем, что ракета обязательно попадет в область G), а оно не является конечным.

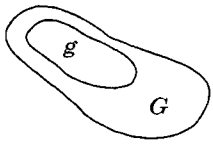


Рис. 1.2

3. Геометрический подход к нахождению вероятности. Проиллюстрируем этот подход, вычисляя вероятность события, состоящего в том, что точка O, брошенная наудачу на область G (рис. 1.2), попадет в подобласть g.

Будем считать, что:

— если точка O брошена на область G, то попадание точки на эту область — достоверное событие, т. е.

$$P(\text{т. } O \in G) = 1; \tag{1.9}$$

— если точка O брошена наудачу на G, то вероятность попадания точки в любую подобласть w области G не зависит от расположения w внутри G и пропорциональна площади S_w подобласти w, т. е.

$$P(\text{т. } O \in w) = cS_w, \tag{1.10}$$

где $c > 0$ — коэффициент пропорциональности.

Принимая в качестве w всю область G, получим

$$1 \underset{(1.9)}{=} P(\text{т. } O \in G) \underset{(1.10)}{=} cS_G.$$

Отсюда имеем $c = 1/S_G$ и $P(\text{т. } O \in w) = S_w/S_G$, поэтому

$$P(\text{т. } O \in g) = S_g/S_G. \tag{1.11}$$

В примере область G располагалась на плоскости; если G — отрезок, g — подотрезок отрезка G, то вероятность

$$P(\text{т. } O \in g) = l_g/l_G,$$

где l_g и l_G — длины соответствующих отрезков. Если G — область в трехмерном пространстве, то вероятность приравнивается отношению соответствующих объемов

$$P(\text{т. } O \in g) = V_g/V_G.$$

Используя геометрический подход, решим следующую задачу.

ЗАДАЧА 1.1 (задача о встрече). А и В решили встретиться в определенном месте в течение определенного часа. Пришедший первым ждет другого 15 мин, после чего уходит. Какова вероятность встречи А и В, если приход каждого в течение оговоренного часа происходит наудачу и момент прихода одного лица не влияет на момент прихода другого?

Решение. Пусть x (мин) и y (мин) — моменты прихода на встречу соответственно А и В. Все возможные варианты значений пары (x, y) — это точки квадрата $OPQC$, изображенного на рисунке 1.3.

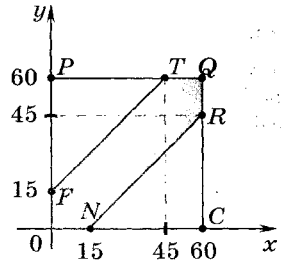


Рис. 1.3

Необходимое и достаточное условие встречи А и В таково:

$$|x - y| \leqslant 15,\ 0 \leqslant x \leqslant 60,\ 0 \leqslant y \leqslant 60.$$

Точки $(x; y)$, удовлетворяющие этой системе неравенств, принадлежат шестиугольнику $OFTQRN$. Вероятность встречи А и В равна отношению площадей

$$S_{OFTQRN}/S_{OPQC} = (S_{OPQC} - 2S_{NRC})/S_{OPQC} = \\ = (60^2 - 45^2)/60^2 = 0{,}4375. \ \blacktriangleleft$$

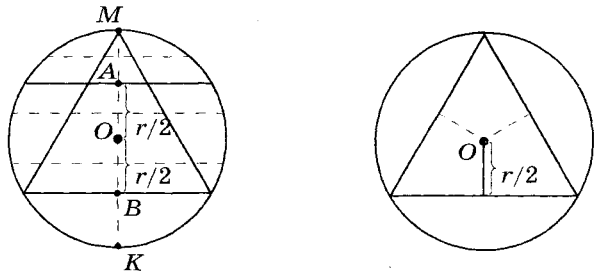
Рассмотрим еще одну задачу, предполагающую использование геометрического подхода, которая иллюстрирует его неоднозначность.

ЗАДАЧА 1.2 (парадокс Ж. Бертрана, XIX в.). Наудачу в круге радиуса r проводится хорда. Какова вероятность того, что ее длина превзойдет длину стороны вписанного равностороннего треугольника?

Решение. Рассмотрим только два варианта толкования выражения «наудачу в круге проводится хорда».

а) Наудачу выбираем точку на диаметре, перпендикулярном фиксированной стороне вписанного равностороннего треугольника, и через эту точку проводим хорду, перпендикулярную этому же диаметру (рис. 1.4, *а*). При таком толковании вероятность того, что хорда превзойдет длину стороны, равна отношению длин отрезков AB и MK:

$$P = l_{AB}/l_{MK} = 2l_{OA}/l_{MK} = (2r/2)/(2r) = 1/2.$$



а) *б)*

Рис. 1.4

б) Наудачу выбираем точку внутри круга и считаем, что эта точка — середина хорды. В этом случае хорда больше стороны треугольника, если ее середина принадлежит вписанному в треугольник кругу. Поэтому искомая вероятность равна отношению площадей вписанного и описанного кругов

$$P = \pi(r/2)^2/(\pi r^2) = 1/4.$$

Получены разные ответы. Это объясняется тем, что за решение одной и той же задачи на основании того, что не определено понятие «наудачу в круге проводится хорда», выдаются решения двух различных задач. ◀

Мы рассмотрели следующие три подхода к вычислению вероятности:

— классический; его используют только при решении задач, в которых множество исходов опыта конечно и исходы равновозможны;

— эмпирический (опытный или статистический), предполагающий возможность многократного повторения опыта в типичных условиях; его применяют более широко, однако, в отличие от «классической вероятности», которая всегда определяется однозначно, «статистическая вероятность» не однозначна: ее значение может меняться и при другом количестве проведенных опытов, и при повторном проведении того же количества опытов;

— геометрический подход; здесь для ситуации, рассматриваемой в конкретной задаче, подбирается некоторая «геометрическая модель», и в ее рамках определяется значение «геометрической вероятности».

Наряду с перечисленными подходами (последние два имеют скорее описательный характер) существует формально-логический, аксиоматический подход к построению теории вероятностей, предложенный русскими математиками С. Н. Бернштейном (1917) и А. Н. Колмогоро-

вым (1933). Не приводя общую аксиоматику теории вероятностей, заметим, что определение вероятности соотношениями (1.1)—(1.3) равносильно определению вероятности, принятому в аксиоматической теории.

Возможны два варианта толкования числового значения вероятности. Это:

— показатель степени уверенности в справедливости суждения, касающегося появления события (субъективный вариант толкования);

— ожидаемая относительная частота появления события в большом числе испытаний (объективный вариант толкования).

Выбор варианта зависит от конкретной ситуации. Например, если имеется информация о том, что вероятность коммерческих затруднений при производстве какого-то изделия равна 0,2, причем производство включает достаточно большое число фирм, то число 0,2 можно истолковать как ожидаемую относительную частоту коммерческих затруднений:

$$\left(\frac{\text{число фирм, испытывающих затруднения}}{\text{общее число фирм}}\right).$$

Однако если значение вероятности, равное 0,2, относится только к одной фирме, то истолкование числа 0,2 как показателя степени уверенности может оказаться более уместным.

Таким образом, если речь идет об одиночном испытании, то толкование вероятности как ожидаемой частоты представляется неуместным. Поскольку управленческие решения часто относятся к особым неопределенным ситуациям, а не к ряду идентичных, в теории управленческих решений большее применение находит субъективный вариант толкования вероятности, т. е. толкование вероятности события как степени уверенности в появлении этого события.

§ 1.3. Комбинаторика при классическом подходе к нахождению вероятности

При использовании классической формулы вероятности (1.8) в решении конкретных задач числовые значения входящих в эту формулу величин N и M не всегда очевидны. Часто для их определения требуется применить правила и формулы комбинаторики, позволяющие подсчитывать число определенных комбинаций из заданных элементов.

Приведем два основных правила такого подсчета.

1) **Правило суммы:** *если элемент a можно выбрать n способами и если после его выбора элемент b можно выбрать m способами, то выбор «либо a, либо b» можно осуществить $n + m$ способами.*

≫ Действительно, представим, что $n + m$ различающихся только цветом шариков распределены по двум ящикам: в первом n шариков, а втором m. Шарик, случайным образом выбранный из первого ящика (так как шарики в ящике различаются только цветом, то это можно сделать следующим образом: перемешать шарики и, закрыв глаза, вытащить один шарик), может иметь n вариантов цвета, а из второго — m вариантов. Тогда число вариантов цвета для шарика, вынутого либо из первого, либо из второго ящика, равно $n + m$. ≪

2) **Правило произведения:** *если элемент a можно выбрать n способами и если после его выбора элемент b можно выбрать m способами, то выбор упорядоченной пары элементов (a, b) можно осуществить nm способами.*

≫ Действительно, представим, что $n + m$ различающихся только цветом шариков распределены по двум ящикам: в первом n шариков, а во втором m. Случайным образом выберем из первого ящика шарик, а затем также случайным образом выберем шарик из второго ящика. Сколько существует различных упорядоченных пар цветов для двух вынутых шариков (упорядоченность пар цветов означает, что, например, пары (синий, белый) и (белый, синий) различны)? Так как шарик, вынутый из первого ящика, может иметь n вариантов цвета и при каждом из этих вариантов шарик, вынутый из второго ящика, может иметь m вариантов цвета, то для двух шариков число различных упорядоченных пар цветов равно nm. ≪

Это правило мы использовали и ранее, не делая на этом акцент. Так, в примере 1.9 на рисунке 1.1, по существу, дана графическая иллюстрация правила произведения: для количества i выпавших очков на первой игральной кости возможно 6 вариантов, для числа j выпавших очков на второй игральной кости — тоже 6 вариантов, поэтому для упорядоченной пары (i, j) существует $6 \cdot 6 = 36$ вариантов.

Правила суммы и произведения можно обобщать на случаи выбора более двух элементов.

К комбинаторным формулам относятся *формулы подсчета размещений, перестановок и сочетаний*. Приведем эти формулы, доказав некоторые из них.

Определение. *Размещениями без повторений из n различных элементов по m элементов $(m < n)$ называются все такие последовательности m различных элементов, выбранных из исходных n, которые отличаются друг от друга или порядком следования элементов, или составом элементов.*

Число размещений без повторений из n элементов по m обозначают символом A_n^m и вычисляют по следующей формуле:

$$A_n^m = n(n-1)\ldots(n-m+1) \equiv n!/(n-m)!,\ m < n, \tag{1.12}$$

где $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$, а $(n-m)! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n-m)$[1].

▶ **ПРИМЕР 1.11.** Размещения без повторений из трех ($n = 3$) различных элементов a, b, c по два ($m = 2$) элемента таковы: a, b; b, a; a, c; c, a; b, c; c, b. Размещений — шесть; и, согласно (1.12), получим такой же результат:

$$A_3^2 = 3 \cdot \ldots \cdot (3-2+1) = 3 \cdot 2 = 6,$$

или

$$A_3^2 = 3!/(3-2)! = 3!/1! = 1 \cdot 2 \cdot 3/1 = 6. \quad ◀$$

≫ Докажем соотношения (1.12). При формировании из n различных элементов произвольной последовательности, включающей m различных элементов, отобрать первый элемент в последовательность можно n способами (им может быть любой из имеющихся n различных элементов); после выбора первого элемента второй, который не может быть таким же, как первый, можно выбрать $n-1$ способом; после выбора первого и второго элементов третий можно выбрать $n-2$ способами и т. д.; и наконец, m-й элемент последовательности можно выбрать $n-(m-1) = n-m+1$ способами. Поэтому в соответствии с правилом произведения, формирование упорядоченной последовательности m элементов можно осуществить $n(n-1)(n-2)\ldots(n-m+1)$ способами, т. е.

$$A_n^m = n(n-1)(n-2)\ldots(n-m+1).$$

Произведем тождественные преобразования полученного для A_n^m выражения. Имеем

$$A_n^m = \frac{A_n^m (n-m)!}{(n-m)!} =$$

$$= \frac{[n(n-1)(n-2)\ldots(n-m+1)] \cdot 1 \cdot 2 \cdot \ldots \cdot (n-m)}{(n-m)!} =$$

$$= \frac{1 \cdot 2 \ldots \cdot (n-m)(n-m+1)\ldots(n-2)(n-1)n}{(n-m)!} = \frac{n!}{(n-m)!}.$$

Соотношения (1.12) доказаны. ≪

Определение. *Размещениями с повторениями из элементов k типов по m элементов (k и m могут быть в любых соотношениях: $m < k$, $m \geqslant k$) называются все та-*

[1] «≡» — знак тождества; «$n!$» читается «эн факториал» (от англ. *factor* — множитель) и означает произведение целых чисел от 1 до n.

кие последовательности m элементов, принадлежащих исходным типам, которые отличаются одна от другой или порядком следования элементов или составом элементов.

Число размещений с повторениями из k типов элементов по m элементов обозначают $\overline{A}_k^m$ и вычисляют по формуле

$$\overline{A}_k^m = k^m. \qquad (1.13)$$

▶ **ПРИМЕР 1.12.** Размещения с повторениями из элементов двух ($k = 2$) типов: тип a и тип b, по три ($m = 3$) элемента, таковы: a, a, a; b, a, a; a, b, a; a, a, b; b, b, a; b, a, b; a, b, b; b, b, b. Размещений с повторениями восемь; и согласно (1.13) получим такой же результат: $\overline{A}_2^3 = 2^3 = 8$. ◀

≫ Докажем формулу (1.13). При формировании из k типов элементов последовательности m элементов, к которым не предъявлено требование их обязательного различия, каждый элемент этой последовательности может быть любого из k типов, т. е. его можно выбрать k способами. Поэтому в соответствии с правилом произведения формирование последовательности m элементов можно осуществить $\underbrace{k \cdot k \cdot \ldots \cdot k}_{m} = k^m$ способами, т. е. $\overline{A}_k^m = k^m$. ≪

▶▶ **ЗАДАЧА 1.3.** В фирме работают восемь человек одинаковой квалификации. Случайно выбранным трем из них поручают три различных вида работ (первому выбранному — работу 1-го вида, второму — 2-го вида, третьему — 3-го вида). Какова вероятность того, что работа каждого вида будет поручена определенным лицам?

Р е ш е н и е. При отборе трех человек из восьми важен не только состав отобранных, но и в каком порядке они отобраны: от порядка отбора зависит распределение работ. Поэтому варианты отбора трех человек из восьми — это размещения и размещения без повторений, так как и среди восьми человек нет одинаковых людей и среди трех их тоже нет. Общее число таких вариантов $A_8^3 = 8!/(8 - 3)! = 336$.

Очевидно, что эти варианты:

— попарно несовместны в одном опыте, состоящем в однократном выборе трех человек из восьми;

— образуют полную группу, так как по крайней мере один из вариантов в опыте произойдет (если учесть попарную несовместность вариантов, в опыте произойдет ровно один из вариантов, а «не по крайней мере один»);

— равновозможны, так как никакая последовательность «троек» не имеет никаких преимуществ в своем появлении.

Поэтому множество Ω, состоящее из 336 вариантов, — это множество равновозможных исходов. Событие же A, состоящее в том, что работа каждого вида будет поручена определенным лицам, или, иначе, состоящее в том, что эти определенные лица будут отобраны в фиксированном порядке, является подмножеством множества Ω, а именно A происходит, когда происходит ровно один исход множества Ω. Таким образом, есть все основания использовать для подсчета вероятности $P(A)$ классическую формулу вероятности (1.8), в которой $N = 336$, $M = 1$; $P(A) = 1/336$.

ЗАДАЧА 1.4. Замок камеры хранения имеет четыре диска, каждый из которых разделен на 10 секторов; на секторах каждого из дисков написаны цифры 0, 1, 2, ... , 9. Какова вероятность открыть закрытую камеру для человека: а) забывшего все, что он набрал на дисках, закрывая камеру; б) запомнившего только цифру, набранную на первом диске; в) запомнившего только, что ни на втором, ни на третьем, ни на четвертом диске он не набирал цифры 6?

Р е ш е н и е. а) Закрывая камеру, замок которой имеет четыре диска, человек, по сути, выбирает последовательность четырех ($m = 4$) цифр из цифр десяти ($k = 10$) типов: тип 0, тип 1, ... , тип 9. При этом важен и состав четырех цифр, и порядок их следования. Поэтому число возможных наборов равно числу размещений с повторениями из элементов $k = 10$ типов по $m = 4$ элемента: $N = \overline{A}_{10}^{4} = 10^4$. И только при одном наборе из 10^4 наборов закрытая камера будет открыта. Искомая вероятность равна $1/10^4$.

б) При известной цифре на первом диске число возможных наборов на трех остальных равно $\overline{A}_{10}^{3} = 10^3$ и искомая вероятность равна $1/10^3$.

в) Число возможных наборов на первом диске равно 10. Число наборов на трех остальных равно числу размещений с повторениями по $m = 3$ элемента из элементов $k = 9$ типов (исходные 10 цифр минус одна цифра 6): $\overline{A}_{9}^{3} = 9^3$.

Общее число возможных наборов на четырех дисках, в соответствии с правилом произведения, равно $10 \cdot 9^3$ и искомая вероятность равна $1/(10 \cdot 9^3)$. ◄

О п р е д е л е н и е. *Перестановками без повторений из n различных элементов называются все возможные последовательности этих n элементов.*

Число перестановок без повторений из n элементов обозначают символом P_n и вычисляют по формуле

$$P_n = n! = 1 \cdot 2 \cdot \ldots \cdot n. \qquad (1.14)$$

ПРИМЕР 1.13. Перестановки без повторений из трех ($n = 3$) различных элементов a, b, c таковы: a, b, c; b, a, c; b, c, a; a, c, b; c, b, a; c, a, b. Число перестановок шесть. Согласно формуле (1.14), получим такой же результат: $P_3 = 3! = 1 \cdot 2 \cdot 3 = 6$. ◀

≫ Докажем формулу (1.14). Перестановки без повторений из n различных элементов можно рассматривать как частный случай размещений без повторений из n различных элементов по $m = n$ элементов. Тогда, учитывая соотношения (1.12), при $m = n$ получим

$$P_n = A_n^n = n(n-1)\ldots(n-n+1) = n(n-1)\ldots 1 = n!,$$

или (по определению $0! = 1$)

$$P_n = A_n^n = n!/(n-n)! = n!/0! = n! \quad \ll$$

Определение. *Перестановками с повторениями из n элементов k типов (число элементов первого типа — n_1, число элементов второго типа — n_2, ... , число элементов k-го типа n_k; $k < n$; $\sum_{i=1}^{k} n_i = n$) называются все возможные последовательности исходных n элементов.*

Число перестановок с повторениями

$$\overline{P}_{n=n_1+n_2+\ldots+n_k} = \frac{n!}{n_1!n_2!\ldots n_k!} \qquad (k < n). \qquad (1.15)$$

ПРИМЕР 1.14. Перестановки с повторениями из трех ($n = 3$) элементов a, a, b двух ($k = 2$) типов: тип a повторяется $n_1 = 2$ раза, тип b повторяется $n_2 = 1$ раз, таковы: a, a, b; a, b, a; b, a, a. Число перестановок равно трем и, согласно формуле (1.15), получим такой же результат

$$\overline{P}_{3=2+1} = \frac{3!}{2!1!} = 3. \quad ◀$$

Замечания. **1.** Если все n элементов разных типов, т. е. $k = 1 + 1 + \ldots + 1 = n$, то число перестановок с повторениями равно числу перестановок без повторений. Действительно,

$$\overline{P}_{n=1+1+\ldots+1} = \frac{n!}{1!1!\ldots 1!} = n! = P_n.$$

2. При любом виде перестановок (и без повторений, и с повторениями) каждая перестановка включает все n исходных элементов и одна перестановка отличается от другой только порядком следования элементов.

ЗАДАЧА 1.5. Какова вероятность получить слово «содержать», переставляя в случайном порядке буквы этого слова? Какова вероятность получить слово «математика», переставляя в случайном порядке буквы этого слова?

Р е ш е н и е. В слове «содержать» все девять букв разные; число перестановок этих букв равно $N = P_9 = 9!$ и лишь $M = 1$ вариант из 9! вариантов дает слово «содержать». Поэтому вероятность получить это слово равна $1/9!$

В слове «математика» $n = 10$ букв, однако различных букв 6:

«м» повторяется $n_1 = 2$ раза, «а» повторяется $n_2 = 3$ раза, «т» повторяется $n_3 = 2$ раза, «е» повторяется $n_4 = 1$ раз, «и» повторяется $n_5 = 1$ раз, «к» повторяется $n_6 = 1$ раз.

Поэтому перестановки букв слова «математика» — это перестановки с повторениями из $n = 10$ элементов $k = 6$ типов, и, в соответствии с (1.15), общее число таких перестановок

$$\overline{P}_{10=2+3+2+1+1+1} = \frac{10!}{2!3!2!1!1!1!} = 151\,200.$$

Из них только одна перестановка дает слово «математика»; вероятность получить это слово, случайно переставляя его буквы, равна $1/151200$. ◀

О п р е д е л е н и е. *Сочетаниями без повторений из n различных элементов по m элементов* ($m < n$) *называются все такие последовательности из m различных элементов, выбранных из исходных n, которые отличаются одна от другой составом элементов.*

Число сочетаний без повторений из n элементов по m обозначают символом C_n^m и вычисляют по формуле

$$C_n^m = \frac{n!}{m!(n-m)!} \qquad (m < n). \tag{1.16}$$

▶ **ПРИМЕР 1.15.** Сочетания без повторений из трех ($n = 3$) различных элементов a, b, c по два ($n = 2$) элемента таковы: a, b; a, c; c, b (сочетания отличаются друг от друга только составом элементов, поэтому, например, последовательности a, b и b, a — это одно и то же сочетание). Число сочетаний без повторений равно трем. Согласно формуле (1.16), получим такой же результат:

$$C_3^2 = \frac{3!}{2!(3-2)!} = 3. \quad ◀$$

≫ Докажем формулу (1.16). Напомним, что размещения без повторений из n элементов по m отличаются друг от друга или составом m элементов, или порядком их следования. Порядков же следования фиксированных m различных элементов столько, сколько можно составить перестановок без повторений из этих m элементов, т. е. $P_m = m!$. Сочетания без повторений отличаются друг от друга только составом элементов, поэтому число сочетаний без повторений из n элементов по m меньше числа размещений без повторений из n элементов по m в $m!$ раз

$$C_n^m = \frac{A_n^m}{m!} = \frac{n!}{m!(n-m)!}.$$

Формула (1.16) доказана. ≪

Определение. *Сочетаниями с повторениями из элементов k типов по m элементов (k и m могут быть в любых соотношениях: $m < k$, $m \geqslant k$) называются все такие последовательности m элементов, принадлежащие исходным типам, которые отличаются друг от друга составом элементов.*

Число сочетаний с повторениями из элементов k типов по m элементов равно

$$\overline{C}_k^m = C_{k+m-1}^m \equiv \frac{(k+m-1)!}{m!(k-1)!}. \qquad (1.17)$$

▶ **ПРИМЕР 1.16.** Сочетания с повторениями из элементов двух ($k = 2$) типов (тип a и тип b) по три ($m = 3$) элемента таковы: a, a, a; b, a, a; b, b, a; b, b, b (сочетания отличаются друг от друга составом элементов, поэтому, например, последовательности: b, a, a; a, b, a и a, a, b — это одно и то же сочетание). Число сочетаний с повторениями — четыре. Согласно формуле (1.17), получим такой же результат:

$$\overline{C}_2^3 = C_{2+3-1}^3 = C_4^3 = \frac{4!}{3!(4-3)!} = 4. \quad ◀$$

≫ **ЗАДАЧА 1.6.** В совокупности имеется K элементов, среди которых L элементов первого вида и $K - L$ элементов второго вида (например, $K = 7$, $L = 4$, $K - L = 3$). Из этой совокупности наугад без возвращения берут k элементов (например, $k = 5$). Какова вероятность того, что в выборке окажется ровно l (например, $l = 3$) элементов первого вида (рис. 1.5)?

Решение. Очевидно, что числа всевозможных выборок из K элементов по k элементов (по условию задачи порядок элементов в выборке не важен) равно $N = C_K^k = C_7^5$.

Подсчитаем число M выборок, в которых содержится l элементов первого вида. Число выборок l элементов из об-

щего числа L элементов первого вида равно $C_L^l = C_4^3$. Каждый такой выбор комбинируется с выбором $k - l$ элементов второго вида из их общего числа $K - L$, и таких вариантов $C_{K-L}^{k-l} = C_3^2$. Поэтому, в соответствии с правилом умножения,

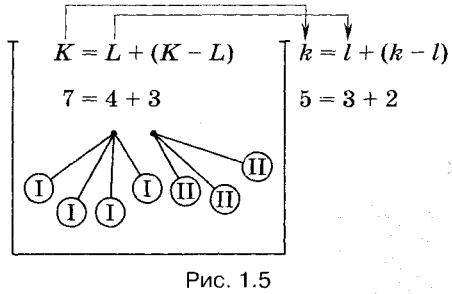


Рис. 1.5

$$M = C_L^l \, C_{K-L}^{k-l} = C_4^3 \, C_3^2 .$$

Искомая вероятность

$$P = M/N = C_L^l \cdot C_{K-L}^{k-l} / C_K^k = C_4^3 \, C_3^2 / C_7^5 = 4/7. \quad \blacktriangleleft$$

Задача подобного типа имеет решение только в том случае, если:

а) объем выборки k не превышает объема исходной совокупности K, $1 \leqslant k \leqslant K$ (в условиях задачи: $1 < (k = 5) < 7$);

б) число l элементов первого вида в выборке:

— с одной стороны, не может быть больше объема выборки k и не может быть больше числа L элементов первого вида в исходной совокупности, т. е. $l \leqslant \min(k, L)$ (в условиях задачи: $(l = 3) < \min(k = 5, L = 4) = 4$);

— с другой стороны, не может быть меньше нуля и не может быть меньше минимального числа элементов первого вида в выборке, которое равно разности между объемом выборки k и общим числом $K - L$ элементов второго вида, т. е. $l \geqslant \max\{0, k - (K - L)\}$ (в условиях задачи: $(l = 3) > > \max\{0; 5 - (7 - 4)\} = 2$). $\blacktriangleleft$

Итак, в задачах типа (1.6) вероятности находятся по формуле

$$P = C_L^l \cdot C_{K-L}^{k-l} / C_K^k, \quad 1 \leqslant k \leqslant K,$$

$$\max\{0, k - (K - L)\} \leqslant l \leqslant \min(k, L) \qquad (1.18)$$

и называются *гипергеометрическими*.

ЗАДАЧА 1.7. Известно, что пять из 40 пассажиров автобуса замешаны в похищении крупной суммы денег. На остановке к автобусу подошел инспектор уголовного розыска и заявил, что ему для обнаружения по крайней мере одного преступника достаточно произвести обыск у шести наугад выбранных пассажиров. Что руководит инспектором: риск или трезвый расчет?

Р е ш е н и е. В задаче объем исходной совокупности $K = 40$ пассажиров, среди них $L = 5$ похитителей и $K - L = 35$ пассажиров, не имеющих отношения к похищению.

Объем выборки $k = 6$; значит, число всевозможных выборок равно $N = C_K^k = C_{40}^6$.

По условию в выборке должен быть по крайней мере один преступник, т. е.

— либо ровно один преступник ($l = 1$), число таких выборок $C_L^l C_{K-L}^{k-l} = C_5^1 C_{35}^5$,

— либо ровно два преступника ($l = 2$), число таких выборок $C_5^2 C_{35}^4$,

— либо ровно три преступника ($l = 3$), число таких выборок $C_5^3 C_{35}^3$,

— либо ровно четыре преступника ($l = 4$), число таких выборок $C_5^4 C_{35}^2$,

— либо ровно пять преступников ($l = 5$), число таких выборок $C_5^5 C_{35}^1$.

Тогда, в соответствии с правилом суммы, число выборок, в которых имеется по крайней мере один преступник, равно

$$M = C_5^1 C_{35}^5 + C_5^2 C_{35}^4 + C_5^3 C_{35}^3 +$$

$$+ C_5^4 C_{35}^2 + C_5^5 C_{35}^1 = 2215220.$$

Вероятность обнаружения по крайней мере одного преступника в выборке из шести пассажиров равна

$$P = M/N = 2215220/C_{40}^6 = 2215220/3838380 = 0{,}5771;$$

вероятность превысила 0,55 (если бы инспектор произвел обыск не шести, а пяти пассажиров, то вероятность обнаружения среди них по крайней мере одного преступника была бы равной 0,5066). По-видимому, это и дало основание инспектору назвать цифру 6.

ЗАДАЧА 1.8. Инвестор формирует портфель ценных бумаг. Он может вложить свои деньги в акции пяти различных фирм. Сколькими способами инвестор может образовать набор из семи акций и какова вероятность того, что в набор попадут четыре акции, принадлежащие различным фирмам?

Р е ш е н и е. По условию из акций пяти ($k = 5$) типов инвестор составляет набор из семи ($m = 7$) акций (в число

таких наборов может в том числе входить и набор, все семь акций которого принадлежат какой-то одной фирме). Очевидно, что для инвестора важен только состав набора: акции каких фирм и в каком количестве входят в набор, и совсем не важен порядок следования отобранных акций. Поэтому количество таких наборов равно числу сочетаний с повторениями из элементов $k = 5$ типов по $m = 7$ элементов: $N = \overline{C}_5^7$. Учитывая формулу (1.17), получим

$$N = \overline{C}_5^7 = C_{5+7-1}^7 = C_{11}^7 = \frac{11!}{7!4!} = 330.$$

Среди этих наборов количество наборов, в каждом из которых четыре акции принадлежат различным фирмам, равно числу сочетаний без повторений из пяти элементов (5 различных фирм) по 4:

$$M = C_5^4 = \frac{5!}{4!1!} = 5.$$

Искомая вероятность $P = M/N = C_5^4/\overline{C}_5^7 = 1/66$. ◀

УПРАЖНЕНИЯ

1. Для испытания, состоящего в однократном подбрасывании трех монет (или в трехкратном подбрасывании одной монеты), запишите множество Ω элементарных событий. Каковы формулировки достоверного и невозможного событий? Сколько всего событий, включая достоверное и невозможное, можно образовать на множестве Ω? Назовите некоторые из них и запишите их как подмножества множества Ω.

2. Образуют ли события «на обеих монетах выпала цифра», «на обеих монетах выпал «герб» множество исходов опыта, состоящего в однократном подбрасывании двух монет? Какие события надо добавить к этим двум, чтобы получилось множество элементарных событий?

3. Образуют ли перечисленные события множество исходов: а) выигрыш, проигрыш в шахматной партии; б) выигрыш, проигрыш финальной встречи по баскетболу; в) выпадение не более трех и выпадение не менее четырех очков при однократном бросании игральной кости?

4. На складе хранятся 1000 аккумуляторов. Известно, что после года хранения 100 штук выходят из строя. Какова вероятность того, что взятый наудачу после года хранения аккумулятор: а) окажется исправным; б) окажется исправным, если известно, что после пяти месяцев хранения были удалены 20 аккумуляторов, ставшими неисправными? Как бы вы провели отбор аккумулятора наудачу?

5. Фирма, занимающаяся рыночными исследованиями, установила, что из 1000 наудачу выбранных жителей города 200 человек знакомо с ее продукцией. Сколько примерно человек окажется знакомыми с продукцией фирмы среди всех 20 000 жителей города?

6 (задача Б. Паскаля). Какова вероятность того, что, проснувшись внезапно ночью, вы обнаружите минутную стрелку между 20 и 40 минутами (предполагается, что ночь длится с нуля до восьми часов)?

7. Авиакомпания выполняет шесть рейсов между Ростовом-на-Дону и Москвой и два рейса между Москвой и Нью-Йорком. Сколькими способами можно заказать билет из Ростова-на-Дону до Нью-Йорка, если все рейсы осуществляются в разные дни?

8. Сколько чисел, меньших 10^5, можно составить из следующих цифр: 1, 3, 5, 7, 9?

9. Сколько четырехзначных чисел можно составить из цифр 3, 4? Какова вероятность составления из этих цифр четырехзначного числа, число тысяч которого равно 3, а число единиц 4?

10 (задача-шутка). В некотором сказочном королевстве не было двух человек с одинаковым набором зубов. Какое могло бы быть наибольшее число жителей этого королевства, если у человека 32 зуба?

11. Собрание, на котором присутствует 20 человек, избирает двух делегатов на две конференции (на каждую по одному делегату). Сколькими способами это можно сделать? Сколькими способами можно отобрать двух кандидатов на одну конференцию?

12. Слово «юрист» разрезали на буквы и приставили в случайном порядке друг к другу все пять букв. Какова вероятность получить слово «юрист»?

Слово «социолог» разрезали на буквы. а) В случайном порядке приставили друг к другу все 8 букв. Какова вероятность получить слово «социолог»? б) В случайном порядке приставили друг к другу пять из 8 букв. Какова вероятность получить слово «голос»? в) В случайном порядке приставили друг к другу четыре из 8 букв. Какова вероятность получить слово «слог»?

13. Для полета в космическом корабле укомплектовывается экипаж: командир корабля, первый и второй помощник (эту тройку отбирают из 10 космонавтов), два бортинженера примерно с одинаковыми обязанностями (их отбирают из восьми специалистов) и врач (его выбирают из трех медиков). Сколькими способами можно укомплектовать экипаж?

14. Среди кандидатов в студенческий совет три первокурсника, четыре второкурсника и семь третьекурсников. Из этого состава наудачу выбирают пять человек. Найдите вероятности следующих событий: а) выбраны одни третьекурсники; б) все первокурсники будут выбраны; в) не выбрано ни одного второкурсника.

15. Покупая карточку лотерии «Спортлото», игрок должен зачеркнуть шесть из 49 возможных чисел от 1 до 49. а) Сколько вариантов вычеркивания шести чисел существует? б) Чему равна вероятность того, что игрок угадает все шесть чисел (при угадывании шести чисел игрок выигрывает значительную сумму денег)? в) Каковы вероятности того, что игрок угадает только четыре числа; только два числа; ни одного числа?

16. В продажу поступили калькуляторы шести различных видов. Сколькими способами можно образовать набор из 10 калькуляторов? Какова вероятность того, что в набор войдут ровно два калькулятора первого вида и три калькулятора второго вида?

ГЛАВА 2

Простейшие теоремы теории вероятностей

В главе 1 мы познакомились со способами непосредственного нахождения вероятностей событий. Иногда эти способы довольно сложные. Чаще для подсчета вероятностей применяют не непосредственные, прямые методы, а косвенные, позволяющие по известным вероятностям одних событий находить вероятности других событий. Для использования косвенных методов надо уметь выражать одни события через другие; этот вопрос и рассматривается в параграфе 2.1. В следующих параграфах излагаются простейшие теоремы теории вероятностей, используемые при косвенном подсчете вероятностей.

§ 2.1. Отношения между событиями. Операции над событиями. Диаграмма Вьенна

Пусть множество исходов некоторого опыта S: $\Omega = \{\omega_i\}$; A и B — два произвольных события, связанных с опытом S, которым соответствуют подмножества A и B множества Ω: $A = \{\omega_j\} \subset \Omega$, $B = \{\omega_k\} \subset \Omega$.

В частности, любое из событий A и B может быть не только случайным, но и достоверным, его обозначают буквой Ω (ему соответствует все множество Ω), и невозможным, его обозначают символом $\varnothing$ (символ пустого множества).

Определяя то или иное отношение между событиями A и B или операцию над ними, будем одновременно давать наглядное представление этого отношения или операции, используя диаграммы Вьенна[1]. На диаграмме Вьенна множество Ω всех исходов опыта условно изображается в виде некоторой области Ω на плоскости, сами исходы — это точки области Ω, событие A — некоторая подобласть A области Ω, а исходы, благоприятствующие событию A, — это точки области A (рис. 2.1).

Событие A включено в событие B, если множество исходов, составляющих A, является подмножеством множества исходов, составляющих B. Отношение включения A в B обозначается так: $A \subset B$ (рис. 2.1, *а*). Очевидно, что если $A \subset B$, то наступление события A в опыте S обязательно ведет к наступлению и события B; обратное утверждение неверно: при наступлении события B событие A может и не произойти.

[1] Дж. Вьенн — английский логик (1834—1923).

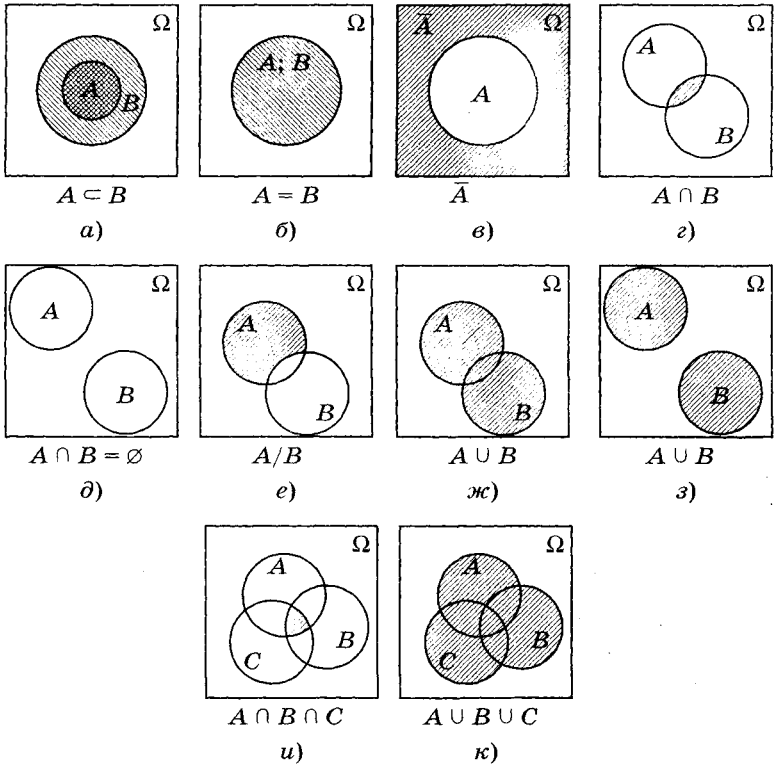


Рис. 2.1

События A и B равносильны, или ***эквивалентны,*** если A включено в B, $A \subset B$, а B включено в A, т. е. $B \subset A$. Равносильность событий A и B обозначается так: $A = B$ (рис. 2.1, *б*). Очевидно, что если $A = B$, то наступлению события A и события B благоприятствуют одни и те же исходы.

Событие называют противоположным событию A, или ***дополнением события*** A, или ***отрицанием события*** A, если оно состоит из всех тех исходов множества Ω, которые не принадлежат A. Событие, противоположное событию A, обозначают символом $\bar{A}$ (читают: «не A» (рис. 2.1, *в*)). Очевидно, что A и $\bar{A}$ — несовместные в одном опыте события, т. е. если в опыте происходит A, то $\bar{A}$ не происходит, если же происходит $\bar{A}$, то A не происходит. Используя диаграммы Вьенна, нетрудно убедиться в справедливости следующих утверждений.

1. $\overline{\Omega} = \varnothing$.

2. $\overline{\overline{A}} = A$.

3. Если $A \subset B$, то $\overline{B} \subset \overline{A}$.

Пересечением или ***произведением событий*** A и B называют событие (обозначение $A \cap B$, где $\cap$ — знак логического умножения, или AB), состоящее из всех тех исходов множества Ω, которые одновременно принадлежат A и B (рис. 2.1, *г*). Очевидно, что событие AB наступает лишь при одновременном, совместном наступлении события A и события B. Используя диаграммы Дж. Вьенна, нетрудно убедиться в справедливости следующих утверждений.

1. Если A и B несовместны, то $A \cap B = \varnothing$ (рис. 2.1, *д*) (в частности, $A \cap \overline{A} = \varnothing$).

2. $A \cap \Omega = A$.

3. $A \cap \varnothing = \varnothing$.

4. Если $A \subset B$, то $A \cap B = A$.

Разностью событий A и B называют событие (его обозначение $A \backslash B$, читают: «A без B»), состоящее из всех тех исходов множества Ω, которые входят в A, но не входят в B (рис. 2.1, *е*). Очевидно, что событие $A \backslash B$ наступает лишь тогда, когда событие A наступает в опыте, а событие B не наступает в этом же опыте, т. е. событие $A \backslash B$ равносильно событию $A \cap \overline{B}$, $A \backslash B = A \cap \overline{B}$. Убедимся в справедливости последнего соотношения, используя диаграммы Дж. Вьенна (рис. 2.2).

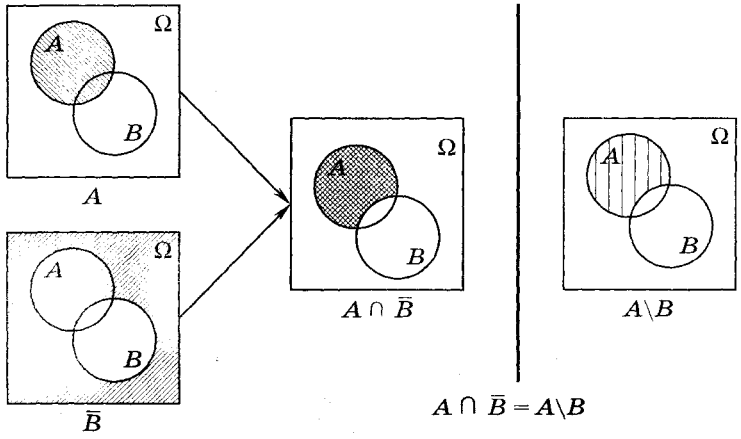


Рис. 2.2

Используя диаграммы Вьенна, нетрудно убедиться и в справедливости следующих утверждений.

1. $A \setminus A = \varnothing$.

2. $\Omega \setminus A = \overline{A}$.

3. Если $A \subset B$, то $A \setminus B = \varnothing$.

Объединением или ***суммой событий*** A и B называют событие (обозначение $A \cup B$, где $\cup$ — знак логического сложения, или $A + B$), состоящее из всех тех исходов множества Ω, которые принадлежат по крайней мере одному из событий A, B (или, иначе, хотя бы одному из событий A, B; или, иначе, не менее, чем одному из событий A, B); из этого следует, что событие $A \cup B$ наступает при наступлении A, при наступлении B и при совместном, одновременном наступлении и A, и B (конечно, если одновременное наступление этих событий возможно). Событие $A \cup B$ изображено на рисунке 2.1, *ж*, — где A и B — совместные события, и на рисунке 2.1, *з*, где A и B — несовместные события.

Используя диаграммы Дж. Вьенна, нетрудно убедиться в справедливости следующих утверждений.

1. $A \cup A = A$.

2. $A \cup \Omega = \Omega$.

3. $A \cup \varnothing = A$.

4. Если $A \subset B$, то $A \cup B = B$.

5. $A \cup \overline{A} = \Omega$.

Замечание. Операции пересечения и объединения могут быть применены и к большему, чем два, числу событий, связанных с одним и тем же опытом S. В этих случаях используют следующие обозначения:

$A_1 \cap A_2 \cap \ldots = \bigcap_{i=1}^{\infty} A_i$ — пересечение событий A_1, A_2, ... , или, иначе, событие, наступающее при одновременном, совместном их наступлении;

$A_1 \cup A_2 \cup \ldots = \bigcap_{i=1}^{\infty} A_i$ — объединение событий A_1, A_2, ... , или, иначе, событие, наступающее тогда, когда наступает хотя бы одно, или по крайней мере одно, или не менее одного из событий A_1, A_2, ...

Пересечение и объединение трех событий: A, B, C изображены соответственно на рисунке 2.1, *и* и *к*.

▶ **ПРИМЕР 2.1.** При однократном подбрасывании двух игральных костей:

1) Событие A — «сумма очков на выпавших гранях равна трем», влечет за собой событие B — «нечетная сумма очков на выпавших гранях», $A \subset B$.

2) События: C — «четная сумма очков» и D — «выпадение числа очков одной и той же четности на обеих костях», равносильны, $C = D$.

3) Событие E — «сумма очков не превышает трех» и событие F — «сумма очков больше трех», противоположны, $F = \overline{E}$.

4) Пересечением событий K — «сумма очков не меньше четырех и не больше восьми» и L — «сумма очков не меньше шести и не больше девяти», будет событие $K \cap L$ — «сумма очков не меньше шести и не больше восьми».

5) Разностью событий K и L является событие $K \setminus L$ — «сумма очков равна четырем или пяти».

6) Объединением событий K и L будет событие $K \cup L$ — «сумма очков не меньше четырех и не больше девяти». ◀

Операции пересечения и объединения событий обладают свойствами, в основном аналогичными свойствам алгебраических операций умножения и сложения:

Свойства операций над событиями	Свойства алгебраических операций
Свойство коммутативности	
$A \cap B = B \cap A$	$ab = ba$
$A \cup B = B \cup A$	$a + b = b + a$
Свойство ассоциативности	
$A \cap B \cap C = A \cap (B \cap C) = (A \cap B) \cap C = B \cap (A \cap C)$	$abc = a(bc) = (ab)c = b(ac)$
$A \cup B \cup C = A \cup (B \cup C) = (A \cup B) \cup C = B \cup (A \cup C)$	$a + b + c = a + (b + c) = (a + b) + c = b + (a + c)$
Распределительное или дистрибутивное свойство операций по отношению друг к другу	
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	$a(b + c) = ab + ac$
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (*)	в алгебре соотношение, аналогичное (*), неверно: $a + (bc) \neq (a + b)(a + c)$

В справедливости приведенных свойств операций над событиями нетрудно убедиться, используя диаграммы Дж. Вьенна. Убедимся, например, в справедливости свойства (*), не имеющего алгебраического аналога (рис. 2.3).

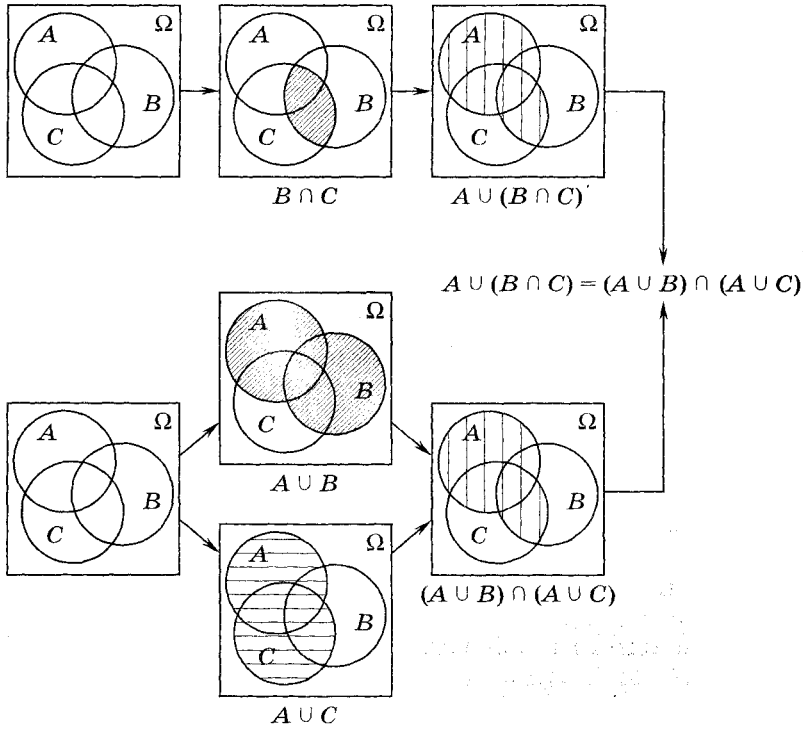


Рис. 2.3

В заключение еще раз остановимся на операции объединения событий. Напомним, $A \cup B$ — это событие, состоящее в появлении хотя бы одного из событий A, B. Содержание этой фразы «высвечивают» следующие соотношения:

$$A \cup B = (A \cap \overline{B}) \cup (\overline{A} \cap B) \cup (A \cap B); \qquad (2.1)$$

$$A \cup B = \overline{\overline{A} \cap \overline{B}}. \qquad (2.2)$$

≫ Убедимся в справедливости соотношения (2.1). Проанализируем его правую часть:

$A \cap \overline{B}$ — одновременное появление событий A и $\overline{B}$, или, иначе, появление A и не появление B;

$\overline{A} \cap B$ — не появление A и одновременно с этим появление B;

$A \cap B$ — одновременное появление и A, и B.

$A \cap \overline{B}$, $\overline{A} \cap B$, $A \cap B$ — попарно несовместные в одном опыте события, поэтому объединение этих событий $(A \cap \overline{B}) \cup (\overline{A} \cap B) \cup (A \cap B)$ — событие, состоящее в появлении или только A, или только B, или одновременном появлении и A, и B; иначе, $(A \cap \overline{B}) \cup (\overline{A} \cap B) \cup (A \cap B)$ — это

событие, состоящее в появлении хотя бы одного из событий A, B; следовательно, оно равносильно событию $A \cup B$.

Теперь убедимся в справедливости соотношения (2.2).

≫ Событие $A \cup B$ — появление хотя бы одного из событий A, B.

Событие $\overline{A} \cap \overline{B}$ — не появление ни A, ни B; тогда событие, противоположное событию $\overline{A} \cap \overline{B}$, есть $\overline{\overline{A} \cap \overline{B}}$ — появление хотя бы одного из событий A, B.

Таким образом, события $A \cup B$ и $\overline{\overline{A} \cap \overline{B}}$ эквиваленты, т. е. $A \cup B = \overline{\overline{A} \cap \overline{B}}$. ≪

§ 2.2. Теоремы о вероятности объединения событий

Напомним, что на диаграмме Дж. Вьенна множество Ω всех исходов опыта изображается в виде некоторой области Ω на плоскости, а исходы, которые благоприятствуют наступлению случайного события A, связанному с этим опытом, образуют подобласть A области Ω (рис. 2.4).

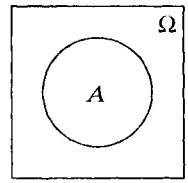


Рис. 2.4

Используем геометрический подход к определению вероятностей. Будем считать, что:

— опыт — это бросание наудачу точки O на область Ω, при этом примем площадь этой области

$$S_\Omega = 1; \qquad (2.3)$$

— попадание точки O на область Ω — это достоверное событие;

— случайное событие A — это попадание точки O в область A.

Тогда, используя формулу (1.11), получим, что вероятность события A равна площади S_A области A

$$P(A) = P(\text{т. } O \in A) = S_A / S_\Omega \underset{(2.3)}{=} S_A / 1 = S_A.$$

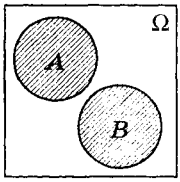
Интерпретация вероятности события как площади области, соответствующей этому событию на диаграмме Дж. Вьенна, существенно упрощает доказательство простейших теорем теории вероятностей.

Теорема 1. *Вероятность объединения (суммы) двух несовместных событий A и B равна сумме вероятностей этих событий, т. е.*

$$P(A \cup B) \underset{A \cap B \neq \varnothing}{=} P(A) + P(B). \qquad (2.4)$$

Замечания. 1. Напомним, что запись $A \cap B = \varnothing$ означает, что события A и B несовместны; расположение этой записи под знаком равенства (2.4) говорит о том, что оно верно только для несовместных событий.

2. Если события A и B несовместны, то событие $A \cup B$ — это появление одного из двух событий: либо A, либо B, неважно какого именно. Поэтому теорема 1 имеет и такую формулировку: *вероятность появления одного из двух несовместных событий, неважно какого, равна сумме вероятностей этих событий.*



$A \cup B$

Рис. 2.5

≫ Несовместные события A и B на диаграмме Дж. Вьенна изображаются как непересекающиеся подобласти A и B области Ω. Поэтому объединение этих событий $A \cup B$ — это и область A, и область B. На рисунке 2.5 область, соответствующая событию $A \cup B$, выделена. Площадь выделенной области обозначим $S_{A \cup B}$. Истолковывая вероятность как площадь, получаем, что и требовалось доказать

$$P(A \cup B) = S_{A \cup B} = S_A + S_B = P(A) + P(B). \quad \ll$$

Очевидным обобщением теоремы 1 является следующая теорема.

Теорема 2. *Вероятность объединения (суммы) попарно несовместных событий $A_1, A_2, \ldots, A_n$, или, иначе, вероятность появления одного из попарно несовместных событий, неважно какого, равна сумме вероятностей этих событий, т. е.*

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n),$$
$$A_i \cap A_j = \varnothing,\ i, j = 1, 2, \ldots, n;\ i \neq j. \qquad (2.5)$$

Замечание. Запись «$A_i \cap A_j = \varnothing$; $i, j = 1, 2, \ldots, n$; $i \neq j$» означает, что никакие два разных события A_i и A_j из исходных n событий не совместны.

Следствие 1. *Сумма вероятностей попарно несовместных событий $A_1, A_2, \ldots, A_n$, образующих полную группу, равна единице.*

≫ Докажем это следствие. С одной стороны, из попарной несовместности событий $A_1, A_2, \ldots, A_n$ вытекает, что:

— $A_1 \cup A_2 \cup \ldots \cup A_n$ — появление одного из событий $A_1, A_2, \ldots, A_n$, неважно какого;

— согласно теореме 2,

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

С другой стороны, события $A_1, A_2, \ldots, A_n$ образуют полную группу, а это означает, что событие, состоящее в появлении одного из них,

т. е. событие $A_1 \cup A_2 \cup \ldots \cup A_n$, достоверное, $P(A_1 \cup A_2 \cup \ldots \cup A_n) = 1$. Поэтому имеет место следующая цепочка равенств:

$$1 = P(A_1 \cup A_2 \cup \ldots \cup A_n) =$$

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

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

Замечание. Доказать следствие можно и используя диаграмму Вьенна. Подобласти, соответствующие попарно несовместным событиям A_1, A_2, ..., A_n, образующим полную группу, не пересекаются и заполняют всю область Ω. Для трех попарно несовместных событий A_1, A_2, A_3, образующих полную группу, диаграмма Дж. Вьенна изображена на рисунке 2.6. Истолковывая вероятность события как площадь соответствующей области, получим

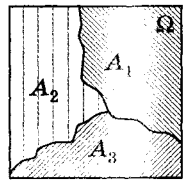


Рис. 2.6

$$P(A_1) + P(A_2) + P(A_3) = S_{A_1} + S_{A_2} + S_{A_3} = S_\Omega = 1. \quad \ll$$

Следствие 2. *Вероятность события* $\overline{A}$, *противоположного событию* A,

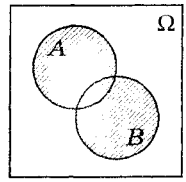
$$P(\overline{A}) = 1 - P(A). \qquad (2.6)$$

$\gg$ События A и $\overline{A}$, будучи противоположными, несовместны и образуют полную группу. Поэтому, согласно следствию 1: $P(A) + P(\overline{A}) = 1$, или $P(\overline{A}) = 1 - P(A)$. $\ll$

Теорема 3. *Вероятность объединения (суммы) совместных событий A и B*

$$P(A \cup B) \underset{A \cap B \neq \varnothing}{=} P(A) + P(B) - P(A \cap B). \qquad (2.7)$$

$\gg$ Изобразив совместные события A и B на диаграмме Дж. Вьенна (рис. 2.7) и истолковывая вероятности событий как площади подобластей, соответствующих этим событиям, получим



$A \cup B$
Рис. 2.7

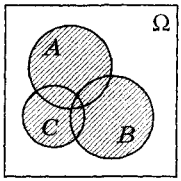
$$P(A \cup B) = S_{A \cup B} = S_A + S_B - S_{A \cap B} =$$

$$= P(A) + P(B) - P(A \cap B),$$

где $S_{A \cup B}$ — площадь всей выделенной области; $S_{A \cap B}$ — площадь той части выделенной области, которая одновременно входит и в область A, и в область B.

Итак, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. $\ll$

Замечание. Если события A и B несовместны, то пересечение этих событий невозможно, т. е. $A \cap B = \varnothing$. Тогда $P(A \cap B) = 0$ и равенство (2.7) принимает вид (2.4). Таким образом, теорема 1 является как бы частным случаем теоремы 3.



$A \cup B \cup C$

Рис. 2.8

Обобщением равенства (2.7) на случай трех событий A, B и C, изображенных на рисунке 2.8, является следующее равенство:

$$P(A \cup B \cup C) = \\ = P(A) + P(B) + P(C) - P(A \cap B) - \\ - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C). \quad (2.8)$$

≫ Действительно, $P(A \cup B \cup C) = S_{A \cup B \cup C} = S_A + S_B + S_C - S_{A \cap B} - S_{A \cap C} - S_{B \cap C} + S_{A \cap B \cap C} = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$,

где $S_{A \cup B \cup C}$ — площадь всей выделенной области; $S_{A \cap B}$ — площадь той части выделенной области, которая одновременно входит и в область A, и в область B; $S_{A \cap C}$ — площадь той части выделенной области, которая одновременно входит и в область A, и в область C; $S_{B \cap C}$ — площадь той части выделенной области, которая одновременно входит и в область B, и в область C; $S_{A \cap B \cap C}$ — площадь той части выделенной области, которая одновременно входит во все три области A, B и C. ≪

≫ **ЗАДАЧА 2.1** (парадокс Р. Мизеса). Некий теннисист может поехать на турнир либо в Москву, либо в Лондон, причем турниры проходят одновременно. Вероятность того, что он займет первое место в Москве, равна 0,9 (если, конечно, он туда поедет), а в Лондоне 0,6. Какова вероятность того, что он займет где-либо первое место?

Р е ш е н и е. Событие A — выигрыш теннисистом турнира в Москве и событие B — выигрыш тем же теннисистом турнира в Лондоне несовместны. Поэтому вероятность выигрыша либо в Москве, либо в Лондоне

$$P(A \cup B) = P(A) + P(B) = 0{,}9 + 0{,}6 = 1{,}5.$$

Получили: вероятность события больше единицы, $P(A \cup B) > 1$, чего быть не может. В чем причина этого парадокса?

Напомним, несовместные события — это события, которые одновременно не могут произойти в одном и том же опыте. В задаче речь идет не об одном, а о двух «опытах»: первый опыт — поездка теннисиста в Москву и, если этот опыт реализовать, то вероятность выигрыша для теннисиста равна 0,9; второй опыт — поездка в Лондон, при реализации этого опыта вероятность выигрыша 0,6. Игнорирование обстоятельства, что несовместность событий рассматривается только в их привязке к одному и тому же опыту, и привело к парадоксальному результату.

ЗАДАЧА 2.2. Результаты опроса 1000 случайно выбранных молодых людей таковы: 811 из них работают; 752 проживают в Москве; 418 учатся; 570 работающих москвичей; 356 молодых людей работают и учатся одновременно,

348 учащихся москвичей, 297 работающих и учащихся москвичей. Содержится ли в этой информации ошибка?

Р е ш е н и е. Пусть A — событие, состоящее в том, что случайно выбранный молодой человек работает, B — проживает в Москве, C — учится. Тогда событие, состоящее в том, что случайно выбранный молодой человек работает и проживает в Москве, — это произведение событий A и B, т. е. $A \cap B$. Аналогично $A \cap C$ — это событие, состоящее в том, что случайно выбранный молодой человек работает и учится; $B \cap C$ — молодой человек — учащийся москвич; $A \cap B \cap C$ — молодой человек работает, учится и живет в Москве.

Подсчитаем вероятности перечисленных событий. Было опрошено достаточно большое число молодых людей — 1000, поэтому, используя эмпирический подход к нахождению вероятности, получим: $P(A) = 811/1000 = 0{,}811$, $P(B) = 0{,}752$, $P(C) = 0{,}418$, $P(A \cap B) = 0{,}570$, $P(A \cap C) = 0{,}356$, $P(B \cap C) = 0{,}348$, $P(A \cap B \cap C) = 0{,}297$.

Теперь, используя формулу (2.8), найдем вероятность события $P(A \cup B \cup C)$:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) = 0{,}811 + 0{,}752 + 0{,}418 - 0{,}570 - 0{,}356 - 0{,}348 + 0{,}297 = 1{,}004.$$

Получили: $P(A \cup B \cup C) = 1{,}004 > 1$, чего быть не может. Следовательно, в информации задачи содержится ошибка.

ЗАДАЧА 2.3 (задача-шутка). В ожесточенном бою не менее 70% бойцов потеряли один глаз, не менее 75% — одно ухо, не менее 80% — одну руку и не менее 85% — одну ногу. Каково минимальное число потерявших одновременно глаз, ухо, руку и ногу?

Р е ш е н и е. Введем следующие события. Случайно выбранный боец потеряет: A — глаз, B — ухо, C — руку, D — ногу, F — одновременно все четыре «предмета». Тогда событие $\overline{F}$ заключается в том, что случайно выбранный боец не потеряет хотя бы один из четырех «предметов» и, в соответствии с определением объединения событий,

$$\overline{F} = \overline{A} \cup \overline{B} \cup \overline{C} \cup \overline{D}.$$

Учитывая это равенство, получим следующую цепочку соотношений:

$$P(F) = 1 - P(\overline{F}) = 1 - P(\overline{A} \cup \overline{B} \cup \overline{C} \cup \overline{D}) \geqslant 1 - [P(\overline{A}) + P(\overline{B}) + P(\overline{C}) + P(\overline{D})] =$$

$$= 1 - [(1 - P(A)) + (1 - P(B)) + (1 - P(C)) + (1 - P(D))] \geqslant$$
$$\geqslant 1 - [(1 - 0{,}7) + (1 - 0{,}75) + (1 - 0{,}80) + (1 - 0{,}85)] = 0{,}1.$$

Итак, не менее 10% бойцов потеряли одновременно глаз, ухо, руку и ногу. ◀

§ 2.3. Теоремы о вероятности пересечения событий. Условная вероятность. Независимые события

Прежде чем рассматривать теоремы, введем несколько новых понятий.

Возможность наступления некоторого события A может зависеть от наступления или ненаступления некоторого другого события B.

Определение. *Вероятность наступления события A при условии наступления события B называется условной вероятностью события A и обозначается $P(A|B)$.*

Замечание. Не следует путать обозначение $A|B$ (читается: наступление события A при условии наступления события B) с обозначением $A \setminus B$ (читается: A без B).

Обычную вероятность события A, в отличие от условной вероятности $P(A|B)$, иногда называют безусловной.

▶ **ПРИМЕР 2.2.** Опыт: одиночное подбрасывание игральной кости. Случайные события: A — выпадение одного очка, B — выпадение трех очков. Безусловная вероятность события A равна $P(A) = 1/6$, а условная вероятность $P(A|B) = 0$.

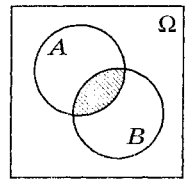
Действительно, если при подбрасывании кости выпадут три очка (произойдет событие B), то выпасть одно очко (появится событие A) в том же самом подбрасывании не может, поэтому вероятность появления события A при условии появления события B равна нулю.

Также и вероятность наступления события B при условии наступления события A равна нулю: $P(B|A) = 0$.

Если же при подбрасывании игральной кости событие B не произойдет (три очка не выпадут), то произойдет событие $\overline{B}$ — выпадут либо одно, либо два, либо четыре, либо пять, либо шесть очков; из перечисленных пяти исходов наступлению события A благоприятствует только один — выпадение одного очка. Поэтому, согласно классической формуле вычисления вероятности, $P(A|\overline{B}) = 1/5$. Нетрудно убедиться, что и $P(B|\overline{A}) = 1/5$. ◀

Итак, если события A и B несовместны, то наступление в опыте события A исключает появление в том же опыте со-

бытия B, т. е. $P(B|A) = 0$; так же как наступление в опыте события B исключает появление в том же опыте события A, т. е. $P(A|B) = 0$.



$A \cap B$
Рис. 2.9

Как вычислить условные вероятности $P(A|B)$ и $P(B|A)$, если события A и B совместны? Обратимся к диаграмме Дж. Вьенна (рис. 2.9) и используем геометрический подход к нахождению вероятности; напомним, что в этом подходе опыт — это бросание точки наудачу на область Ω.

Вероятность $P(A|B)$ вычисляется при условии, что в опыте произойдет событие B, т. е. брошенная точка обязательно попадет в область B. Появление события A при выполнении этого условия означает, что брошенная точка попадет в выделенную область $A \cap B$. Поэтому $P(A|B) = S_{A \cap B}/S_B$, где $S_{A \cap B}$ — площадь выделенной области. Но так как $S_{A \cap B} = P(A \cap B)$, а $S_B = P(B)$, то

$$P(A|B) = P(A \cap B)/P(B), \qquad (2.9)$$

естественно предполагается, что $P(B) \neq 0$.

Заменив в формуле (2.9) A на B и B на A и воспользовавшись свойством коммутативности операции пересечения событий, откуда следует, что $P(B \cap A) = P(A \cap B)$, получим

$$P(B|A) = P(A \cap B)/P(A). \qquad (2.10)$$

В частности, если A и B несовместны, то $P(A \cap B) = 0$ и $P(A|B) = 0$, также $P(B|A) = 0$.

На условную вероятность переносятся все полученные ранее результаты. В частности:

если события A и B — несовместны, то, аналогично формуле (2.4),

$$P((A \cup B)|C) = P(A|C) + P(B|C);$$

по аналогии с (2.6),

$$P(\overline{A}\,|B) = 1 - P(A|B). \qquad (2.11)$$

Теорема 4. *Вероятность пересечения (произведения) событий A и B равна произведению вероятности одного из них на условную вероятность другого, т. е.*

$$P(A \cap B) = P(A)P(B|A) \qquad (2.12)$$

или

$$P(A \cap B) = P(B)P(A|B). \qquad (2.13)$$

≫ Справедливость равенства (2.12) вытекает из соотношения (2.10), а равенства (2.13) — из соотношения (2.9). ≪

Обобщением теоремы 4 является следующая теорема.

Теорема 5. *Вероятность пересечения (произведения) событий* $A_1, A_2, A_3, \ldots, A_n$ *равна*

$$P(A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_n) =$$

$$= P(A_1)P(A_2|A_1)P(A_3|(A_1 \cap A_2))\ldots P(A_n|(A_1 \cap A_2 \cap \ldots \cap A_{n-1})). \quad (2.14)$$

Здесь $P(A_3|(A_1 \cap A_2))$ — вероятность появления события A_3 при условии совместного появления событий A_1, A_2; $P(A_n|(A_1 \cap A_2 \cap \ldots \cap A_{n-1}))$ — вероятность появления события A_n при условии совместного появления событий $A_1, A_2, \ldots, A_{n-1}$.

В случае трех событий

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|(A_1 \cap A_2)). \quad (2.15)$$

Замечание. В силу свойства коммутативности операции пересечения событий, вероятность $P(A_1 \cap A_2 \cap \ldots \cap A_n)$ можно рассчитать по нескольким тождественным формулам (число этих формул равно $n!$). Например, при $n = 3$, так:

$$P(A_1 \cap A_2 \cap A_3) = P(A_2 \cap A_1 \cap A_3) = P(A_2)P(A_1|A_2)P(A_3|(A_1 \cap A_2))$$

или так:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1 \cap A_3 \cap A_2) =$$

$$= P(A_1)P(A_3|A_1)P(A_2|(A_1 \cap A_3)) \text{ и т. д.}$$

Определение. *Событие A не зависит от события B, если условная вероятность события A равна его безусловной вероятности, т. е. если*

$$P(A|B) = P(A). \quad (2.16)$$

Событие A зависит от события B, если

$$P(A|B) \neq P(A). \quad (2.17)$$

Убедимся в том, что *независимость событий взаимна*: если A не зависит от B, то и B не зависит от A.

≫ Так как A не зависит от B, то $P(A|B) = P(A)$ и тогда, согласно формуле (2.13), $P(A \cap B) = P(B)P(A)$. Вместе с тем согласно формуле (2.12) $P(A \cap B) = P(A)P(B|A)$. В итоге получаем $P(A \cap B) = P(B)P(A) = P(A)P(B|A)$. Отсюда следует, что $P(B|A) = P(B)$, т. е. B не зависит от A. ≪

Итак, если событие A не зависит от события B, то и событие B не зависит от события A. Поэтому в дальнейшем, сталкиваясь с «независимостью», будем говорить: «A и B независимые события», не указывая при этом «направление» независимости.

Также и *зависимость событий взаимна*: если событие A зависит от события B, то и событие B зависит от события A; в этом случае будем говорить: «A и B зависимые события».

Из независимости (зависимости) событий A и B следует независимость (зависимость) событий: $\overline{A}$ и B, A и $\overline{B}$, $\overline{A}$ и $\overline{B}$.

Докажем только одно из этих утверждений: если A и B независимы, то независимы и события $\overline{A}$ и B.

≫ События A и B, независимы, поэтому $P(A|B) = P(A)$. Учитывая это равенство, получим

$$P(\overline{A}\,|B) \underset{(2.11)}{=} 1 - P(A|B) = 1 - P(A) = P(\overline{A}\,).$$

Итак, $P(\overline{A}\,|B) = P(\overline{A}\,)$, а это означает независимость событий $\overline{A}$ и B. ≪

Остальные утверждения доказываются аналогичным образом.

Замечание. Смешение понятий независимости и несовместности — грубейшая ошибка. Несовместность не только не то же самое, что и независимость, а наоборот, пример очень сильной зависимости событий. Если A и B — случайные несовместные события, то наступление одного из них, например A, исключает наступление другого, поэтому $P(B|A) = 0$. С другой стороны, событие B, будучи случайным, имеет вероятность $P(B) \neq 0$. Окончательно имеем $P(B|A) \neq P(B)$, т. е. события A и B зависимы.

Теорема 6. *Вероятность пересечения (произведения) независимых событий A и B равна произведению их вероятностей, т. е.*

$$P(A \cap B) = P(A)P(B). \qquad (2.18)$$

≫ Так как A и B независимы, то $P(B|A) = P(B)$, и заменив в (2.12) $P(B|A)$ на $P(B)$, получим: $P(A \cap B) = P(A)P(B)$, что и требовалось доказать. ≪

Верно и утверждение, обратное содержащемуся в теореме 6: *если $P(A \cap B) = P(A)P(B)$, то события A и B независимы.*

≫ Действительно, сопоставив равенство $P(A \cap B) = P(A)P(B)$ с равенством (2.12): $P(A \cap B) = P(A)P(B|A)$, получим, что $P(B|A) = P(B)$, а это означает, что события A и B независимы. ≪

Так как из независимости событий A и B следует равенство (2.18) и, наоборот, из равенства (2.18) вытекает независимость событий A и B, то часто A и B называют независимыми, если $P(A \cap B) = P(A)P(B)$, и зависимыми, если $P(A \cap B) \neq P(A)P(B)$.

Определение. *События $A_1, A_2, \ldots, A_n$ независимы в совокупности или просто независимы, если, наряду с их*

попарной независимостью, независимы любое из них и пересечение (произведение) любого числа из остальных; в противном случае события $A_1, A_2, \ldots, A_n$ *зависимы.*

Например, события A_1, A_2, A_3 независимы в совокупности, если независимы события A_1 и A_2; A_1 и A_3; A_2 и A_3; A_1 и $A_2 \cap A_3$; A_2 и $A_1 \cap A_3$; A_3 и $A_1 \cap A_2$.

Можно доказать, что из независимости в совокупности событий $A_1, A_2, \ldots, A_n$ следует независимость в совокупности событий $\overline{A}_1, A_2, \ldots, A_n$; событий $A_1, \overline{A}_2, \ldots, A_n, \ldots$; событий $\overline{A}_1, \overline{A}_2, \ldots, \overline{A}_n$.

Очевидным обобщением теоремы 6 является следующая теорема.

Теорема 7. *Вероятность пересечения (произведения) независимых в совокупности событий* $A_1, A_2, \ldots, A_n$ *равна произведению их вероятностей, т. е.*

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2)\ldots P(A_n). \qquad (2.19)$$

Можно доказать, что верно утверждение, обратное содержащемуся в теореме 7: *если выполняется равенство* (2.19), *то события* $A_1, A_2, \ldots, A_n$ *независимы в совокупности.* Поэтому часто события $A_1, A_2, \ldots, A_n$ называют независимыми в совокупности, если имеет место равенство (2.19), и зависимыми, если равенство (2.19) неверно.

Убедимся в том, что из попарной независимости событий вовсе не следует их независимость в совокупности.

▶ **ПРИМЕР 2.3** (пример С. Н. Бернштейна). Опыт: одиночное подбрасывание правильного тетраэдра, одна грань которого раскрашена в синий цвет, другая — в красный, третья — в зеленый, а четвертая — частично во все три цвета. Случайные события С, К, З — тетраэдр упадет на плоскость гранью, содержащей соответственно синий, красный, зеленый цвета. Очевидно,

$$P(\text{С}) = P(\text{К}) = P(\text{З}) = 2/4 = 1/2.$$

Вероятность пересечения событий С и К

$$P(\text{С} \cap \text{К}) = 1/4.$$

Так как

$$P(\text{С} \cap \text{К}) = P(\text{С})P(\text{К}),$$

то события С и К независимые. Нетрудно убедиться и в том, что независимы события С и З; К и З. Далее имеем $P(\text{С} \cap \text{К} \cap \text{З}) = 1/4$ и

$$P(\text{С} \cap \text{К} \cap \text{З}) \neq P(\text{С})P(\text{К})P(\text{З}),$$

поэтому события С, К, З зависимы в совокупности, хотя независимы попарно. ◄

Итак, из независимости в совокупности вытекает попарная независимость. Однако из попарной независимости вовсе не следует независимость в совокупности. Понятие несовместности обладает прямо противоположным свойством: из несовместности группы событий $A_1, A_2, \ldots, A_n$ не вытекает их попарная несовместность, но из попарной несовместности следует несовместность по «три», по «четыре» и т. д.

►► **ЗАДАЧА 2.4.** Для двух случайных событий A, B известны вероятности: $P(A) = 0{,}8$, $P(A \cup B) = 0{,}9$, $P(B|A) = 0{,}6$. Найти $P(B)$, $P(A|B)$, $P(A \cap B)$ и выяснить, зависимы ли события A и B.

Р е ш е н и е. Согласно (2.12),

$$P(A \cap B) = P(A)P(B|A) = 0{,}8 \cdot 0{,}6 = 0{,}48.$$

На основании (2.7),

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Отсюда

$$P(B) = P(A \cup B) - P(A) + P(A \cap B) = 0{,}9 - 0{,}8 + 0{,}48 = 0{,}58;$$

$$P(A|B) = P(A \cap B)/P(B) = 0{,}48/0{,}58 = 24/29.$$

Так как $P(A|B) \neq P(A)$, или поскольку $P(A \cap B) \neq P(A)P(B)$, события A и B зависимы. ◄◄

§ 2.4. Решение задач на применение теорем о вероятностях объединения и пересечения событий. Теорема о вероятности хотя бы одного из независимых в совокупности событий

►► **ЗАДАЧА 2.5.** Жюри состоит из трех человек X, Y, Z; X и Y каждый с вероятностью 0,8 принимают правильное решение, а Z для вынесения решения подбрасывает монету. Члены жюри действуют *независимо*. Решение принимается большинством голосов. Какова вероятность правильного решения?

Р е ш е н и е. Введем следующие события:

A — «X примет правильное решение», $P(A) = 0{,}8$, $P(\overline{A}) = 0{,}2$;

B — «Y примет правильное решение», $P(B) = 0{,}8$, $P(\overline{B}) = 0{,}2$;

C — «Z примет правильное решение», $P(C) = 0{,}5$, $P(\overline{C}) = 0{,}5$.

Правильное решение будет принято, если правильное решение примут X, Y, но не Z; или X, Z, но не Y; или Y, Z,

но не X; или X, Y и Z. Поэтому событие D — «жюри примет правильное решение» — выражается через исходные события так:

$$D = (A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C) \cup (A \cap B \cap C).$$

Здесь четыре события, взятые в скобки, попарно несовместны, и это дает основание воспользоваться теоремой 2 (формулой (2.5)); события, стоящие внутри любой из скобок, независимы (члены жюри действуют независимо), и это дает основание воспользоваться теоремой 7 (формулой (2.19)). Получим

$$P(D) = P((A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C) \cup (A \cap B \cap C)) \underset{(2.5)}{=} P(A \cap B \cap \bar{C}) + P(A \cap \bar{B} \cap C) + P(\bar{A} \cap B \cap C) + P(A \cap B \cap C) \underset{(2.19)}{=} P(A)P(B)P(\bar{C}) + P(A)P(\bar{B})P(C) + P(\bar{A})P(B)P(C) + P(A)P(B)P(C) = 0{,}8 \cdot 0{,}8 \cdot 0{,}5 + 0{,}8 \cdot 0{,}2 \cdot 0{,}5 + 0{,}2 \cdot 0{,}8 \cdot 0{,}5 + 0{,}8 \cdot 0{,}8 \cdot 0{,}5 = 0{,}8.$$

ЗАДАЧА 2.6. События A, B, C независимы в совокупности, их вероятности известны. Найти вероятность появления хотя бы одного из трех событий.

Р е ш е н и е. Пусть событие D — «появление хотя бы одного из событий A, B, C». Возможны три способа решения задачи.

1-й способ. Синоним выражения «появление хотя бы одного из трех событий» есть «появление одного из трех или двух из трех или всех трех». Поэтому

$$P(D) = P[\underbrace{(A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap \bar{C}) \cup (\bar{A} \cap \bar{B} \cap C)}_{\text{одно из трех появится}} \cup \underbrace{(A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C)}_{\text{два из трех появятся}} \cup \underbrace{(A \cap B \cap C)}_{\text{три появятся}}] \underset{(*)}{=} P(A \cap \bar{B} \cap \bar{C}) + P(\bar{A} \cap B \cap \bar{C}) + \ldots + P(A \cap B \cap C) \underset{(**)}{=} P(A)[1 - P(B)][1 - P(C)] + [1 - P(A)]P(B)[1 - P(C)] + \ldots + P(A)P(B)P(C).$$

Возможность перехода (*) объясняется попарной несовместностью событий — «круглых скобок»; возможность перехода (**) объясняется независимостью в совокупности событий, стоящих внутри каждой круглой скобки.

2-й способ. Из определения понятия объединения событий следует: $D = A \cup B \cup C$. Требование попарной несов-

местности к событиям A, B, C не предъявлено, поэтому в соответствии с формулами (2.8) и учитывая далее, что события A, B, C независимы в совокупности, получим

$$P(D) = P(A \cup B \cup C) = P(A) + P(B) + P(C) -$$
$$- P(A)P(B) - P(A)P(C) - P(B)P(C) + P(A)P(B)P(C).$$

3-й способ. Если событие D — появление хотя бы одного из трех событий A, B, C, то $\overline{D}$ — не появление ни A, ни B, ни C, т. е. $\overline{D} = \overline{A}\,\overline{B}\,\overline{C}$. Учитывая независимость в совокупности событий A, B, C, потому и событий $\overline{A}$, $\overline{B}$, $\overline{C}$, получим

$$P(D) = 1 - P(\overline{D}) = 1 - P(\overline{A}\,\overline{B}\,\overline{C}) = 1 - P(\overline{A})P(\overline{B})P(\overline{C}) =$$
$$= 1 - [1 - P(A)][1 - P(B)][1 - P(C)]. \quad \blacktriangleleft$$

Нетрудно видеть, что третий способ наименее трудоемкий. Обобщим его на случай n событий. Сформулируем следующую теорему.

Теорема 8. *Вероятность появления хотя бы одного из n независимых в совокупности событий A_1, A_2, ..., A_n равна единице минус произведение вероятностей противоположных событий $\overline{A}_1$, $\overline{A}_2$, ..., $\overline{A}_n$, т. е.*

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = 1 - P(\overline{A}_1)P(\overline{A}_2)\ldots P(\overline{A}_n). \quad (2.20)$$

Следствие. *Вероятность появления хотя бы одного из n независимых в совокупности событий A_1, A_2, ..., A_n, вероятности появления которых одинаковы и равны числу p, подсчитывается так:*

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = 1 - (1 - p)^n. \quad (2.21)$$

▶▶ **ЗАДАЧА 2.7.** Первый прибор состоит из n блоков, соединенных последовательно (рис. 2.10, *а*), второй из n параллельно соединенных блоков (рис. 2.10, *б*). Блоки выходят из строя независимо друг от друга. Надежность (вероятность безотказной работы в течение времени T) каждого блока равна p. Найти надежность P каждого прибора в целом. Какой должна быть надежность p каждого блока, чтобы обеспечить на-

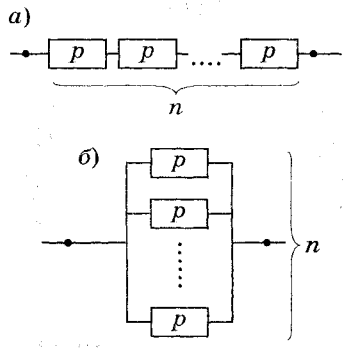


Рис. 2.10

дежность прибора в целом не меньшую, чем $P_{\min} = 0{,}9$, если число блоков $n = 5$?

Р е ш е н и е. Первый прибор будет безотказно работать в том и только том случае, если безотказно будут работать все n блоков. Так как блоки выходят из строя независимо друг от друга, то, в соответствии с теоремой 7 (о вероятности пересечения независимых в совокупности событий), надежность прибора $P_1 = \underbrace{p \cdot p \cdot \ldots \cdot p}_{n} = p^n$. Второй прибор будет безотказно работать в том и только том случае, если безотказно будет работать хотя бы один из n блоков. По формуле (2.21) надежность прибора $P_2 = 1 - (1-p)^n$.

Конечно, при числе блоков $n \geqslant 2$ надежность P_2 второго прибора больше надежности P_1 первого. В этом нетрудно убедиться, если учесть, что $0 < p < 1$.

Найдем надежность p каждого блока при фиксированном числе n, обеспечивающую надежность прибора в целом, не меньшую $P_{\min}$. В приведенных ниже выкладках используется операция логарифмирования. Напомним, что рассматриваемые в задаче вероятности — положительные числа, поэтому их логарифмирование допустимо. Далее, поскольку основанием логарифма является число $10 > 1$, неравенства $x_1 \geqslant x_2 \geqslant 0$ и $\lg x_1 \geqslant \lg x_2$ равносильны, также равносильны неравенства $z_1 \geqslant z_2$ и $10^{z_1} \geqslant 10^{z_2}$. Все необходимые вычисления приведены в следующей таблице:

Первый прибор	Второй прибор
Имеем $P_1 \geqslant P_{\min}$, $p^n \geqslant P_{\min}$, $\lg p^n \geqslant \lg P_{\min}$, $n \lg p \geqslant$ $\geqslant \lg P_{\min}$, $\lg p \geqslant (\lg P_{\min})/n$. Подставив в последнее неравенство числовые значения $P_{\min}$ и n, получим $\lg p \geqslant (\lg 0{,}9)/5$, $\lg p \geqslant -0{,}00915$, $p \geqslant 10^{-0{,}00915}$, $p \geqslant 0{,}979$, $p_{\min} = 0{,}979$	Имеем $P_2 \geqslant P_{\min}$, $1 - (1-p)^n \geqslant$ $\geqslant P_{\min}$, $1 - P_{\min} \geqslant (1-p)^n$, $\lg(1 - P_{\min}) \geqslant n \lg(1-p)$, $\frac{\lg(1 - P_{\min})}{n} \geqslant \lg(1-p)$. Подставив числовые значения, получим $\frac{\lg(1 - 0{,}9)}{5} \geqslant \lg(1-p)$, $-0{,}2 \geqslant \lg(1-p)$, $10^{-0{,}2} \geqslant 1-p$, $0{,}631 \geqslant 1-p$, $p \geqslant 0{,}369$, $p_{\min} = 0{,}369$

Как и следовало ожидать, при параллельном соединении блоков заданная надежность прибора в целом обеспечивается меньшей надежностью каждого блока, чем при последовательном соединении.

ЗАДАЧА 2.8. Вероятность p безотказной работы лифта в течение суток равна 0,9. Каким должно быть минимальное число таких лифтов в доме, чтобы вероятность P безотказной работы в течение суток хотя бы одного из них была не меньше 0,95. Предполагается, что лифты работают независимо друг от друга.

Р е ш е н и е. В соответствии с формулой (2.21) и условием задачи вероятность безотказной работы хотя бы одного из n лифтов $1-(1-p)^n \geqslant P$. Отсюда

$$1-P \geqslant (1-p)^n;\ \lg(1-P) \geqslant n\lg(1-p).$$

Далее, так как $0<1-p<1$ и, следовательно, $\lg(1-p)<0$, то

$$\frac{\lg(1-P)}{\lg(1-p)} \leqslant n.$$

Подставив в последнее неравенство числовые значения вероятностей, получим

$$n \geqslant \frac{\lg(1-0{,}95)}{\lg(1-0{,}9)} = \frac{\lg 0{,}05}{\lg 0{,}1} = 1{,}3; \quad n_{\min} = 2. \ \blacktriangleleft$$

§ 2.5. Формулы полной вероятности и Байеса

Теорема 9. *Если выполняются следующие условия:*

— *событие A наступает только вместе с каким-либо из событий $H_1, H_2, \ldots, H_n$ (называемых гипотезами);*
— *гипотезы $H_1, H_2, \ldots, H_n$ попарно несовместны;*
— *гипотезы $H_1, H_2, \ldots, H_n$ образуют полную группу,* (2.22)

то имеет место формула полной вероятности:

$$P(A) = P(H_1)P(A|H_1) + \ldots + P(H_n)P(A|H_n). \qquad (2.23)$$

$\gg$ Докажем теорему для случая $n=3$, используя геометрический подход к определению вероятностей. На диаграмме Вьенна (рис. 2.11) гипотезы H_1, H_2, H_3, образуя в соответствии с условиями (2.22) полную группу попарно несовместных событий, заполнят, не пересекаясь, всю область Ω, площадь которой, согласно (2.3), равна 1.

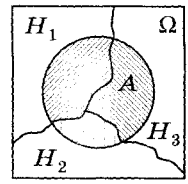


Рис. 2.11

На этом же рисунке изображено (кругом) событие A, которое, согласно (2.22), наступает лишь совместно с какой-либо из гипотез H_1, H_2, H_3. Площадь области A

$$S_A = S_{A\cap H_1} + S_{A\cap H_2} + S_{A\cap H_3},$$

где, например, $S_{A \cap H_2}$ — площадь области, заштрихованной вертикальной штриховкой. Поэтому

$$P(A) = P(A \cap H_1) + P(A \cap H_2) + P(A \cap H_3).$$

Отсюда, учитывая теорему 4 о вероятности пересечения двух событий, получим формулу полной вероятности при $n = 3$

$$P(A) = P(H_1)P(A|H_1) + P(H_2)P(A|H_2) + P(H_3)P(A|H_3).$$

Замечание. Если событие A наступает при наступлении либо события H_1, либо события H_2, но не H_3, то $P(A|H_3) = 0$. ≪

Теорема 10. *Если в дополнение к условиям* (2.22) *имеется сообщение о том, что в опыте событие A произойдет, то для каждой гипотезы имеет место формула Т. Байеса*

$$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{P(A)}, \qquad i = 1, 2, \ldots, n. \qquad (2.24)$$

≫ Доказательство теоремы вытекает из того, что, с одной стороны,

$$P(A \cap H_i) = P(A)P(H_i|A),$$

а с другой,

$$P(A \cap H_i) = P(H_i)P(A|H_i).$$

Приравняв правые части обоих равенств, получим

$$P(A)P(H_i|A) = P(H_i)P(A|H_i),$$

откуда и вытекает формула (2.24). ≪

Замечания. 1. Условные вероятности $P(H_i|A)$ называют ***апостериорными, послеопытными*** (опыт будет произведен, и в нем появится событие A) вероятностями гипотез; безусловные вероятности $P(H_i)$ называют ***априорными, доопытными***. Формула Байеса позволяет «пересмотреть» вероятности $P(H_i)$ гипотез, выдвинутых для объяснения некоторого явления перед началом опыта. «Пересмотр» состоит в подсчете условных вероятностей $P(H_i|A)$ после проведения опыта, результатом которого будет появление события A.

2. Поскольку $H_1, H_2, \ldots, H_n$ — попарно несовместные события, образующие полную группу, сумма априорных вероятностей равна единице:

$$P(H_1) + P(H_2) + \ldots + P(H_n) = 1.$$

Аналогичное равенство имеет место и для апостериорных вероятностей:

$$P(H_1|A) + P(H_2|A) + \ldots + P(H_n|A) = 1.$$

▶ **ЗАДАЧА 2.9.** В пирамиде 10 винтовок, четыре из которых с оптическим прицелом. Вероятность того, что стрелок поразит мишень при выстреле из винтовки с оптическим прицелом, равна 0,95; для винтовки без оптического прицела эта вероятность равна 0,8. Стрелок наудачу берет винтов-

ку. Каковы вероятности того, что: а) стрелок возьмет винтовку с оптическим прицелом и поразит мишень; б) стрелок возьмет винтовку без оптического прицела и поразит мишень; в) стрелок поразит мишень; г) допустим, что стрелок поразит мишень из наудачу взятой винтовки. Что вероятнее: он будет стрелять из винтовки с оптическим прицелом или без него?

Р е ш е н и е. Введем события: A — стрелок поразит мишень из наудачу взятой винтовки, H_1 — наудачу взята винтовка с оптическим прицелом, H_2 — наудачу взята винтовка без оптического прицела. По условию $P(H_1) = 0{,}4$, $P(H_2) = 0{,}6$, $P(A|H_1) = 0{,}95$, $P(A|H_2) = 0{,}8$.

а) Требуется найти вероятность совместного появления событий H_1 и A:

$$P(H_1 \cap A) = P(H_1)P(A|H_1) = 0{,}4 \cdot 0{,}95 = 0{,}38.$$

б) $P(H_2 \cap A) = P(H_2)P(A|H_2) = 0{,}6 \cdot 0{,}8 = 0{,}48.$

в) Событие A наступает лишь при наступлении события H_1 или события H_2, которые несовместны и образуют полную группу. Поэтому, в соответствии с (2.23),

$$P(A) = P(H_1)P(A|H_1) + P(H_2)P(A|H_2) =$$

$$= 0{,}4\ 0{,}95 + 0{,}6 \cdot 0{,}8 = 0{,}86.$$

г) В результате выстрела мишень будет поражена, т. е. произойдет событие A. Пересчитаем вероятности гипотез, используя формулу (2.24). Имеем

$$P(H_1|A) = P(H_1)P(A|H_1)/P(A) = 0{,}4 \cdot 0{,}95/0{,}86 = 19/43;$$

$$P(H_2|A) = P(H_2)P(A|H_2)/P(A) = 0{,}6 \cdot 0{,}8/0{,}86 = 24/43.$$

Вероятнее, что выстрел будет произведен из винтовки без оптического прицела. В задаче две гипотезы, поэтому $P(H_1|A) + P(H_2|A) = 1$. ◀

УПРАЖНЕНИЯ

1. Пусть A, B, C — три совместных события, связанные с одним опытом. Запишите с помощью операций над событиями события, состоящие в том, что из этих трех событий: а) произойдет только A; б) произойдет одно и только одно событие; в) произойдет по крайней мере одно событие; г) A и B произойдут, а C не произойдет; д) произойдут ровно два события; е) произойдут по крайней мере два события; ж) произойдут все три события; з) ни одно событие не произойдет; и) произойдут не более двух событий. События A, B, C и искомые события изобразите на диаграмме Вьенна.

2. Используя диаграммы Дж. Вьенна, проверьте справедливость следующих соотношений: а) $\overline{A \cap B} = \overline{A} \cup \overline{B}$; б) $\overline{A \cup B} = \overline{A} \cap \overline{B}$ [соот-

ношения а) и б) называются *законами Де-Моргана*]; в) $\bar{A} \cap \bar{B} = \overline{A \cap B}$; г) $\overline{A \cup B} = \bar{A} \cup \bar{B}$.

3. Установка состоит из двух котлов и одной машины. Пусть событие A — машина исправна; события B_1 и B_2 — исправны соответственно первый и второй котлы; C — установка работоспособна, что возможно лишь в том случае, когда исправна машина и хотя бы один из котлов. Выразите события C и $\bar{C}$ через события A, B_1 и B_2.

4. Староста группы дал следующие сведения о студентах группы: в группе 45 студентов, в том числе 25 юношей; 30 студентов учатся на «хорошо» и «отлично», в том числе 16 юношей; спортом занимаются 28 человек, в их числе 18 юношей и 17 студентов, учащихся на «хорошо» и «отлично»; 15 юношей учатся на «хорошо» и «отлично» и занимаются спортом. Содержится ли в этих сведениях ошибка?

5. В фирме 21% работников получают высокую заработную плату. Известно также, что 40% работников фирмы — женщины, а 6,4% работников — женщины, получающие высокую заработную плату. Можно ли утверждать, что на фирме существует дискриминация женщин в оплате труда?

6. Студент пришел на экзамен, зная лишь 20 из 25 вопросов. Экзаменатор задает ему три вопроса. Используя понятие условной вероятности, найдите вероятность того, что студент знает все эти вопросы. Найдите ту же вероятность, используя формулу гипергеометрической вероятности.

7. Вероятность для компании получить контракт в стране А равна 0,4; вероятность получить его в стране В равна 0,3. Вероятность того, что контракты будут заключены и в стране А, и в стране В, равна 0,1. Чему равна вероятность получения контракта хотя бы в одной стране?

8. Вероятность того, что наудачу опрошенный человек поддержит правительственную программу, равна 0,65. Какова минимальная численность опрошенных, в которой с вероятностью, не меньшей 0,9, по крайней мере один человек не поддержит эту программу?

9. Экономист компании полагает, что вероятность роста стоимости акций его компании в следующем году равна 0,75, если экономика страны будет на подъеме; та же вероятность равна 0,3, если экономика страны не будет успешно развиваться. По его мнению, вероятность экономического подъема в будущем году равна 0,4. Каковы вероятности того, что: а) экономика страны будет на подъеме и стоимость акций возрастет; б) акции компании поднимутся в цене; в) если акции поднимутся в цене, то что вероятнее: экономика будет на подъеме или нет?

ГЛАВА 3

Независимые повторные испытания

В главе изучаются основные закономерности, относящиеся к одной из важнейших схем теории вероятностей — схеме последовательности независимых испытаний. Эти закономерности позволяют определять вероятности любого заданного числа появлений события.

§ 3.1. Испытания и формула Бернулли. Биномиальное распределение

Довольно часто в вероятностных задачах постулируются следующие условия:

— испытание, результат которого зависит от случая, повторяется n раз;

— в каждом испытании имеются лишь две возможности: A и $\overline{A}$ (появление A назовем ***успехом***, а появление $\overline{A}$ — ***неудачей***);

— испытания независимы (результат любого из них никак не связан с результатами любого числа других)[1];

— вероятность успеха в каждом отдельно взятом или единичном испытании имеет одно и то же значение p (испытания проводятся в одинаковых вероятностных условиях). (3.1)

Испытания, удовлетворяющие условиям (3.1), называют ***испытаниями Я. Бернулли*** (по имени итальянского математика, жившего в 1654—1705 гг.). Приведем пример таких испытаний.

▶ **ПРИМЕР 3.1.** В урне a белых и b черных шаров. Шары отличаются только цветом. Испытание состоит в извлечении наудачу шара из урны. Появление белого шара — успешное испытание, черного — неудача. Пусть проводятся $n = 2$ испытаний. Можно ли назвать их испытаниями Бернулли, т. е. являются ли они независимыми и проведенными в одинаковых вероятностных условиях? Ответим на этот вопрос для случаев выборки с возвратом и без возврата.

Пусть события A_1 и $\overline{A}_1$ — соответственно успех (извлечение белого шара) и неудача (извлечение черного шара) в первом испытании. Для обоих типов выборок $P(A_1) = a/(a + b)$ и $P(\overline{A}_1) = b/(a + b)$. Пусть A_2 и $\overline{A}_2$ — соответственно успех и неудача во втором испытании. Найдем вероятности этих событий.

Выборка с возвратом. В этом типе выборки перед каждым следующим выбором шар, отобранный на предыдущем шаге, возвращают в урну. Поэтому перед вторым испытанием в урне будет a белых и b черных шаров, как и перед первым испытанием, и $P(A_2) = a/(a + b)$, а $P(\overline{A}_2) = b/(a + b)$.

[1] Понятие независимости испытаний уточнено в примере 3.1.

Для этой выборки условная вероятность $P(A_2|A_1)$ равна безусловной вероятности $P(A_2)$: $P(A_2|A_1) = P(A_2) = a/(a + b)$, поэтому события A_1 и A_2 независимы, следовательно, независимы $\overline{A}_1$ и A_2; A_1 и $\overline{A}_2$; $\overline{A}_1$ и $\overline{A}_2$. Иначе, каким бы ни был результат одного испытания, он никак не связан с результатом другого испытания, т. е. испытания независимы. Далее, так как вероятность успеха в первом испытании равна вероятности успеха во втором: $P(A_1) = P(A_2) = a/(a + b)$, то испытания проводятся в одинаковых вероятностных условиях.

Таким образом, извлечение с возвратом последовательно двух (аналогично трех и более) шаров — это испытания Бернулли.

Выборка без возврата. В этом типе выборки перед каждым следующим выбором шар, отобранный на предыдущем шаге, в урну не возвращают. Поэтому перед вторым испытанием в урне $(a + b - 1)$ шаров, из которых число белых шаров равно:

$a - 1$, если результат первого отбора — белый шар (появление события A_1);

a, если результат первого отбора — черный шар (появление события $\overline{A}_1$).

Используя формулу полной вероятности (2.23) (для данной ситуации событие A — это A_2, гипотезы H_1 и H_2 — это события A_1 и $\overline{A}_1$), получим

$$P(A_2) = P(A_1)P(A_2|A_1) + P(\overline{A}_1)P(A_2|\overline{A}_1) = \\ = \frac{a}{a+b}\,\frac{a-1}{a+b-1} + \frac{b}{a+b}\,\frac{a}{a+b-1} = \frac{a}{a+b}.$$

Для этой выборки условная вероятность $P(A_2|A_1) = (a - 1)/(a + b - 1)$, что не равно безусловной вероятности $P(A_2) = a/(a + b)$, $P(A_2|A_1) \neq P(A_2)$, поэтому события A_1 и A_2 зависимы, следовательно, зависимы и два рассматриваемых испытания. Таким образом, они не являются испытаниями Бернулли, хотя вероятность успеха в первом испытании и вероятность успеха во втором имеют, как и при выборке с возвратом, одно и то же значение: $P(A_1) = P(A_2) = a/(a + b)$. Вероятность успеха не изменяется при переходе от одного испытания к другому. Она останется равной $a/(a + b)$ и для третьего, и для всех последующих испытаний, включая и последнее, с номером $(a + b)$.

Итак, выборка с возвратом гарантирует соблюдение требований (3.1) к испытаниям Бернулли; выборка без воз-

врата не обеспечивает независимости испытаний, хотя и гарантирует неизменность вероятностных условий проведения испытаний. ◂

Найдем вероятность $P_n(m)$ того, что среди n испытаний Бернулли успешных будет ровно m (и, как следствие, неуспешных $n - m$). Выведем формулу для случая $n = 4$ и $m = 3$.

≫ Введем следующие события:

B — появление трех успехов в четырех испытаниях;

A_i — успех в i-м испытании, $i = 1, 2, 3, 4$;

$\overline{A}_i$ — неудача в i-м испытании, $i = 1, 2, 3, 4$.

Напомним, что в соответствии с требованиями (3.1):

— испытания независимы — это означает независимость в совокупности событий A_1, A_2, A_3, A_4, следовательно, и независимость в совокупности любой комбинации из «успехов» и «неудач» в четырех испытаниях;

— вероятности успеха в испытаниях имеют одно и то же значение p, т. е.

$$P(A_1) = P(A_2) = P(A_3) = P(A_4) = p.$$

Отсюда следует, что

$$P(\overline{A}_1) = P(\overline{A}_2) = P(\overline{A}_3) = P(\overline{A}_4) = 1 - p.$$

Далее будем называть число p *вероятностью успеха в единичном испытании*, а $q = 1 - p$ — *вероятностью неудачи в единичном испытании*.

Выразим событие B — появление трех успехов в четырех испытаниях — через «успехи» и «неудачи»:

$$B = \underbrace{(A_1 \cap A_2 \cap A_3 \cap \overline{A}_4)}_{\substack{\text{успех в 1-м, 2-м и 3-м} \\ \text{испытаниях и неудача в 4-м}}} \cup \underbrace{(A_1 \cap A_2 \cap \overline{A}_3 \cap A_4)}_{\substack{\text{успех в 1-м, 2-м и 4-м} \\ \text{испытаниях и неудача в 3-м}}} \cup$$

$$\cup \underbrace{(A_1 \cap \overline{A}_2 \cap A_3 \cap A_4)}_{\substack{\text{успех в 1-м, 3-м и 4-м} \\ \text{испытаниях и неудача в 2-м}}} \cup \underbrace{(\overline{A}_1 \cap A_2 \cap A_3 \cap A_4)}_{\substack{\text{успех в 2-м, 3-м и 4-м} \\ \text{испытаниях и неудача в 1-м}}}.$$

Учитывая попарную несовместность «скобок» в одном опыте, состоящем из четырех испытаний Я. Бернулли, и независимость в совокупности событий внутри каждой скобки, используя последовательно теоремы о вероятности объединения попарно несовместных событий и о вероятности произведения независимых в совокупности событий, получим

$$P(B) = P(A_1)P(A_2)P(A_3)P(\overline{A}_4) + P(A_1)P(A_2)P(\overline{A_3})P(A_4) +$$

$$+ P(A_1)P(\overline{A_2})P(A_3)P(A_4) + P(\overline{A}_1)P(A_2)P(A_3)P(A_4) =$$

$$= pppq + ppqp + pqpp + qppp = 4p^3q. \quad ≪$$

Итак, вероятность того, что в $n = 4$ испытаниях Бернулли успех произойдет $m = 3$ раза, равна $P_4(3) = 4p^3q$. Из вывода этой формулы нетрудно видеть, что:

— множитель 4 — это количество способов, которыми на «четырех местах можно разместить три успеха»; оно равно числу сочетаний из четырех ($n = 4$) по три ($m = 3$):

$$C_4^3 = \frac{4!}{3!(4-3)!} = 4;$$

— показатель степени числа p равен числу успехов $m = 3$;

— показатель степени числа q равен числу неудач $n - m = 4 - 3 = 1$.

В общем случае вероятность того, что в n испытаниях Бернулли m успешных находят по ***формуле Бернулли***:

$$P_n(m) = C_n^m p^m q^{n-m}, \qquad m = 0, 1, 2, \ldots, n, \qquad (3.2)$$

где $C_n^m = \frac{n!}{m!(n-m)!}$ — число сочетаний из n элементов по m; p — вероятность успеха в единичном испытании; $q = 1 - p$ — вероятность неудачи в единичном испытании.

Используя формулу Бернулли, найдем вероятности того, что число успехов $m = 0, 1, 2, \ldots, n$, и поместим их в таблицу 3.1.

Таблица 3.1

Число успехов m	0	1	2	...	n	
Вероятность $P_n(m) = C_n^m p^m q^{n-m}$	$C_n^0 p^0 q^{n-0} = q^n$	$C_n^1 pq^{n-1} = npq^{n-1}$	$C_n^2 p^2 q^{n-2}$	...	$C_n^n p^n q^{n-n} = p^n$	$\Sigma = 1$

«Вероятности Бернулли» в таблице 3.1 являются коэффициентами при $x^0, x^1, \ldots, x^n$ в разложении бинома $(q + px)^n$ по степеням x:

$$(q+px)^n = \underbrace{C_n^0 q^n p^0 x^0}_{P_n(0)} + \underbrace{C_n^1 q^{n-1} p^1 x^1}_{P_n(1)} + \underbrace{C_n^2 q^{n-2} p^2 x^2}_{P_n(2)} + \ldots + \underbrace{C_n^n q^{n-n} p^n x^n}_{P_n(n)}. \qquad (3.3)$$

В силу этого заданное таблицей 3.1 распределение вероятностей Бернулли по числу успехов называют ***биномиаль-***

ным распределением или ***биномиальным рядом распределения.***

Сумма вероятностей в таблице 3.1 равна единице. В этом можно убедиться двумя способами:

— сумма вероятностей — это вероятность достоверного события: «при проведении n испытаний число успехов равно 0, или 1, или 2 ... или n», которая всегда равна единице;

— полагая в (3.3) $x = 1$, получим

$$C_n^0 q^n p^0 + C_n^1 q^{n-1} p^1 + \ldots + C_n^n q^0 p^n = (p+q)^n = 1^n = 1.$$

Замечание. Если понятие испытаний Бернулли (3.1) обобщить на случай, когда в каждом испытании имеется $k > 2$ возможностей: A_1, A_2, ... , A_k, причем p_i — вероятность появления A_i в единичном испытании, $i = 1, 2, \ldots, k$, то вероятность появления в n испытаниях m_1 раз события A_1, m_2 раз события A_2, ... , m_k раз события A_k ($m_1 + m_2 + \ldots + m_k = n$) равна

$$P_n(m_1, m_2, \ldots, m_k) = \frac{n!}{m_1! m_2! \ldots m_k!} p_1^{m_1} p_2^{m_2} \ldots p_k^{m_k}. \qquad (3.4)$$

Эта вероятность является коэффициентом при $x_1^{m_1} x_2^{m_2} \ldots x_k^{m_k}$ в разложении полинома $(p_1 x_1 + p_2 x_2 + \ldots + p_k x_k)^n$ по степеням «икс»; поэтому вероятность (3.4) называют ***полиномиальной***. В формуле (3.4) коэффициент при произведении вероятностей есть не что иное, как число перестановок $\overline{P}_{n=m_1+m_2+\ldots+m_k}$ с повторениями из n элементов k типов (см. (1.15)).

Число успехов m^*, которому соответствует наибольшая вероятность в таблице 3.1, называют ***наивероятнейшим числом успехов***. Его можно найти, не вычисляя всех вероятностей, следующим образом:

если $np + p$ — дробное число, то m^* — целое число из интервала $(np - q, np + p)$,
если $np + p$ — целое число, то $np - q = np + p - 1$ — тоже целое, и наивероятнейших чисел будет два: $m_1^* = np - q$ и $m_2^* = m_1^* + 1 = np + p$. (3.5)

≫ Убедимся в правомерности этого алгоритма. Рассматривая вероятность $P_n(m)$ в формуле (3.2) как функцию аргумента m (m — целое число), найдем области ее возрастания и убывания.

Область возрастания. Из неравенства $P_n(m) > P_n(m-1)$ следует:

$$C_n^m p^m q^{n-m} > C_n^{m-1} p^{m-1} q^{n-m+1},$$

$$\frac{n!p}{m!(n-m)!} > \frac{n!q}{(m-1)!(n-m+1)!}, \qquad \frac{p}{m} > \frac{q}{n-m+1},$$

$$pn - pm + p > qm, \; pn + p > m(p+q), \; m < np + p.$$

Итак, $[0, np + p)$ — область возрастания вероятности $P_n(m)$.

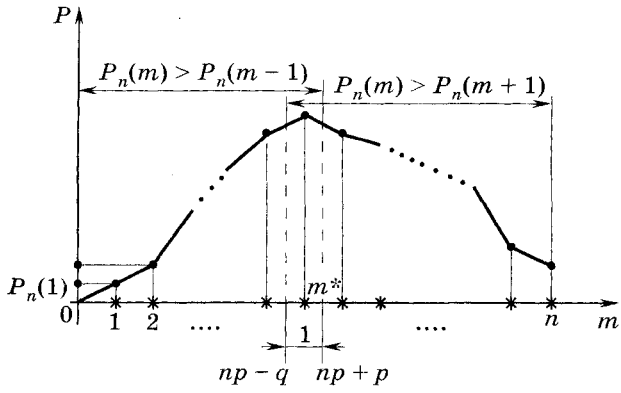


Рис. 3.1

Область убывания. Аналогично можно убедиться в том, что из неравенства $P_n(m) > P_n(m+1)$ следует, $m > np - q$, т. е. $(np - q, n]$ — область убывания вероятности $P_n(m)$.

Пусть $(np - q)$ — дробное число (в этом случае число $np - q + 1 = np + p$ — тоже дробное). Тогда, учитывая области возрастания и убывания функции $P_n(m)$, получим: если целое число $m \in [0, np + p)$, то вероятность $P_n(m)$ больше «предыдущей» вероятности $P_n(m-1)$; если же $m \in (np - q, n]$, то вероятность $P_n(m)$ больше «последующей» вероятности $P_n(m+1)$. Графическая иллюстрация этих результатов приведена на рисунке 3.1 (целые числа помечены крестиком). Из рисунка 3.1 видно, что наибольшую вероятность имеет целое число m^* из интервала $(np - q, np + p)$.

Если же $np - q$ (стало быть, и $np + p$) — целое число, то наибольшую вероятность имеют два числа: $m_1^* = np - q$ и $m_2^* = m_1^* + 1 = np + p$ и, конечно, вероятности этих чисел одинаковы, или, иначе, $P_n(m_1^*)/P_n(m_2^*) = 1$. Убедимся в этом. Имеем

$$\frac{P_n(m_1^*)}{P_n(m_2^*)} = \frac{P_n(m_1^*)}{P_n(m_1^*+1)} = \frac{C_n^{m_1^*} p^{m_1^*} q^{n-m_1^*}}{C_n^{m_1^*+1} p^{m_1^*+1} q^{n-m_1^*-1}} = \frac{n!(m_1^*+1)!(n-m_1^*-1)!}{m_1^*!(n-m_1^*)!n!} \frac{q}{p} =$$

$$= \frac{m_1^*+1}{n-m_1^*} \frac{q}{p} = \frac{np-q+1}{n-np+q} \frac{q}{p} = \frac{np+p}{nq+q} \frac{q}{p} = 1. \quad \ll$$

Часто требуется найти вероятность $P_n(m_1 \leqslant m \leqslant m_2)$ того, что в n испытаниях Бернулли число m успешных — целое число из отрезка $[m_1, m_2]$, где m_1, m_2 — целые числа, $0 \leqslant m_1 < m_2 \leqslant n$. Появление в n испытаниях успехов m_1 или $m_1 + 1$, ... или m_2 — попарно несовместные события, поэтому

$$P_n(m_1 \leqslant m \leqslant m_2) =$$
$$= P_n(m_1) + P_n(m_1+1) + \ldots + P_n(m_2). \qquad (3.6)$$

В ряде случаев вычисления по формуле (3.6) можно заменить более простыми расчетами. Например, вероятность того, что в n испытаниях будет хотя бы один успех, или, иначе, число успехов m — целое число из отрезка $[1, n]$

$$P_n(1 \leqslant m \leqslant n) = P_n(1) + P_n(2) + \ldots + P_n(n) = 1 - P_n(0); \tag{3.7}$$

вероятность того, что в n испытаниях число m успехов не меньше двух

$$P(2 \leqslant m \leqslant n) = P_n(2) + P_n(3) + \ldots + P_n(n) = 1 - P_n(0) - P_n(1). \tag{3.8}$$

ЗАДАЧА 3.1. Примерно 20% судебных дел — это дела по обвинению в краже. В порядке прокурорского надзора проверено четыре наудачу отобранных дела. а) Какова вероятность появления среди отобранных дел хотя бы одного дела о краже? б) Каково наивероятнейшее число дел о краже среди отобранных и какова вероятность этого числа?

Р е ш е н и е. По условию число испытаний $n = 4$; «успех» — наугад взятое дело — это дело о краже, вероятность успеха $p = 0{,}2$, вероятность «неудачи» $q = 0{,}8$.

а) Вероятность появления среди $n = 4$ дел хотя бы одного дела о краже в соответствии с формулой (3.7) равна

$$P_4(1 \leqslant m \leqslant 4) = 1 - P_4(0) = 1 - C_4^0 p^0 q^4 =$$

$$= 1 - 0{,}8^4 = 1 - 0{,}4096 = 0{,}5904.$$

б) Воспользуемся алгоритмом (3.5). Имеем: $np + p = 4 \cdot 0{,}2 + 0{,}2 = 1$ — целое число, поэтому наивероятнейших чисел два: $m_1^* = np - q = 4 \cdot 0{,}2 - 0{,}8 = 0$ и $m_2^* = np + p = 1$. Вероятности этих чисел: $P_4(0) = 0{,}4096$, $P_4(1) = C_4^1 \cdot 0{,}2^1 \times \times 0{,}8^3 = 0{,}4096$.

Как и следовало ожидать, вероятности одинаковы, и они наибольшие. В этом нетрудно убедиться, составив биномиальный ряд распределения:

m	0	1	2	3	4	
$P_4(m) = C_4^m 0{,}2^m 0{,}8^{4-m}$	0,4096	0,4096	0,1536	0,0256	0,0016	$\Sigma = 1$

◀◀

§ 3.2. Формула и распределение Пуассона

Если число n испытаний Бернулли *велико* (несколько сотен), а вероятность p успеха в единичном испытании *мала* (близка к нулю), то хорошее приближение к вероятностям, найденным по формуле Бернулли (3.2), дают вероятности, рассчитанные по ***формуле Пуассона***. Согласно этой формуле, вероятность $P(m)$ появления m успехов в n испытаниях

$$P(m) = \frac{\lambda^m}{m!} e^{-\lambda}, \qquad m = 0, 1, 2, \ldots, n, \tag{3.9}$$

где $\lambda = np$, $e = 2{,}71828\ldots$ — основание натурального логарифма, $m! = 1 \cdot 2 \cdot \ldots \cdot m$.

В силу «малости» вероятности p, формулу Пуассона называют также ***формулой «редких явлений»***.

Замечание. Строго говоря, формула Пуассона — предельный случай формулы Бернулли при $n \to \infty$, если при этом произведение np остается постоянной величиной (обозначим ее буквой λ), т. е.

$$\lim_{\substack{n \to \infty \\ np = \lambda}} C_n^m p^m (1-p)^{n-m} = \frac{\lambda^m}{m!} e^{-\lambda}.$$

Практика показывает допустимость замены формулы Бернулли формулой Пуассона при n, равном нескольким сотням, и $np < 10$. Различия между вероятностями Бернулли и Пуассона тем меньше, чем больше n и меньше p.

Используя формулу Пуассона, найдем вероятности различного числа успехов при бесконечно большом числе испытаний и поместим их в таблицу 3.2.

Таблица 3.2

Число успехов m	0	1	2	3	...	
Вероятность $P(m) = \frac{\lambda^m}{m!} e^{-\lambda}$	$\frac{\lambda^0}{0!} e^{-\lambda} = e^{-\lambda}$	$\frac{\lambda^1}{1!} e^{-\lambda} = \lambda e^{-\lambda}$	$\frac{\lambda^2}{2!} e^{-\lambda}$	$\frac{\lambda^3}{3!} e^{-\lambda}$	...	$\Sigma = 1$

Заданное таблицей 3.2 распределение вероятностей Пуассона по числу успехов называют ***пуассоновским распределением*** или ***пуассоновским рядом распределения***.

В таблице 3.2 в отличие от таблицы 3.1 вероятностей бесконечно много, но сумма их, как и в таблице 3.1, рав-

на 1. Докажем это, воспользовавшись известным из курса математики функциональным рядом

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots, \qquad |x| < \infty, \tag{3.10}$$

и тем, что хотя число n испытаний и бесконечно большое, но вероятность p настолько мала, что $\lambda = np$ — конечное число, т. е. $\lambda < \infty$.

≫ Сумма $P(0) + P(1) + P(2) + P(3) + \dots =$

$$= e^{-\lambda} + \frac{\lambda}{1!}e^{-\lambda} + \frac{\lambda^2}{2!}e^{-\lambda} + \frac{\lambda^3}{3!}e^{-\lambda} + \dots = e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots\right) \underset{(3.10)}{=}$$

$$\underset{(3.10)}{=} e^{-\lambda}e^{\lambda} = e^0 = 1. \quad \ll$$

В случае, когда вероятности находят по формуле Пуассона, число успехов m^*, которому соответствует наибольшая вероятность, или, иначе, наивероятнейшее число успехов, находят следующим образом:

если в формуле Пуассона число λ — дробь, то m^* — целое число из интервала $(\lambda - 1, \lambda)$;

если λ — целое число, то наивероятнейших чисел два: $m_1^* = \lambda - 1$ и $m_2^* = \lambda$. (3.11)

Обоснование этого алгоритма аналогично обоснованию алгоритма (3.5). Убедимся лишь в том, что если наивероятнейших чисел два: $m_1^* = \lambda - 1$ и $m_2^* = \lambda$, то их вероятности одинаковы, или отношение этих вероятностей равно единице:

$$\frac{P(m_1^*)}{P(m_2^*)} = \frac{P(\lambda - 1)}{P(\lambda)} \underset{(3.9)}{=} \frac{\lambda^{\lambda-1}e^{-\lambda}\lambda!}{(\lambda - 1)!\lambda^{\lambda}e^{-\lambda}} = \frac{\lambda}{\lambda} = 1.$$

≫ **ЗАДАЧА 3.2.** Примерно 0,1% судебных дел — это дела по обвинению в убийстве. Проверено 200 наудачу взятых судебных дел. Какова вероятность того, что среди них дел об убийстве: а) 0; 1; 2; 3; б) хотя бы одно; в) более трех. Каково наивероятнейшее число дел об убийстве среди 200 дел и вероятность этого числа?

Р е ш е н и е. По условию $n = 200$, а вероятность успеха в единичном испытании («успех» — случайно взятое дело — это дело по обвинению в убийстве) $p = 0{,}001$. Поскольку n достаточно велико, а $np = 0{,}2 < 10$, есть основания использовать формулу Пуассона (3.9), в которой $\lambda = np = 200 \cdot 0{,}001 = 0{,}2$.

а) Требуемые вероятности вычислим (с точностью до четырех десятичных знаков) по формуле Пуассона и для сравнения по формуле Бернулли (3.2). Имеем

m	0	1	2	3	
$P(m)=\frac{0{,}2^m}{m!}\mathrm{e}^{-0{,}2}$	0,81873	0,16375	0,01637	0,00109	$\Sigma=0{,}9999$
$P_{200}(m)=$ $=C_{200}^m(0{,}001)^m\times$ $\times(0{,}999)^{200-m}$	0,81865	0,16389	0,01632	0,00108	$\Sigma=0{,}9999$

Вероятности Пуассона и Бернулли практически не отличаются. В данной задаче сумма вероятностей, найденных по формуле Бернулли, не может быть равна единице, поскольку число дел об убийстве среди $n = 200$ может быть равным не только 0, 1, 2, 3, но и 4, 5, ... , 200. Сумма вероятностей, найденных по формуле Пуассона, также не равна единице. Строго говоря, сумма «пуассоновских» вероятностей равна единице только при бесконечно большом числе испытаний. В рассматриваемом случае $n = 200$ — это конечное число, но достаточно большое. Поэтому сумма пуассоновских вероятностей при $m = 0, 1, 2, \ldots, 200$ практически равна единице.

б) Учитывая сказанное о сумме пуассоновских вероятностей, находим, что вероятность появления хотя бы одного дела об убийстве среди $n = 200$ дел равна

$$P(1 \leqslant m \leqslant 200) \underset{(3.7)}{=} 1 - P(0) \underset{(3.9)}{=} 1 - \frac{0{,}2^0}{0!}\mathrm{e}^{-0{,}2} =$$

$$= 1 - 0{,}8187 = 0{,}1812.$$

в) Имеем

$$P(3 < m \leqslant 200) = P(4 \leqslant m \leqslant 200) = 1 - P(0 \leqslant m \leqslant 3) =$$

$$= 1 - [P(0) + P(1) + P(2) + P(3)] \underset{(3.9)}{=}$$

$$\underset{(3.9)}{=} 1 - \left(\frac{0{,}2^0}{0!}\mathrm{e}^{-0{,}2} + \ldots + \frac{0{,}2^3}{3!}\mathrm{e}^{-0{,}2}\right) = 1 - 0{,}9999 = 0{,}0001.$$

В задаче $\lambda = np = 0{,}2$, поэтому, в соответствии с алгоритмом (3.11), наивероятнейшее число m^* дел о краже — это целое число из интервала $(0{,}2 - 1;\ 0{,}2)$, т. е. $m^* = 0$ и вероятность этого числа

$$P(0) = \frac{0{,}2^0}{0!}\mathrm{e}^{-0{,}2} = 0{,}8187. \quad \blacktriangleleft$$

Формулу Пуассона в виде

$$P_t(m) = \frac{(\lambda t)^m}{m!}\mathrm{e}^{-\lambda t}, \qquad m = 0, 1, 2, \ldots, \qquad (3.12)$$

используют для нахождения $P_t(m)$ — вероятности того, что за промежуток времени длиной t наступит m событий ***простейшего потока*** — потока однородных событий, происходящих в случайные моменты времени. Этот поток имеет типичные для многих ситуаций свойства:

— поток ординарный, т. е. одновременное наступление двух или более событий практически невозможно;

— поток, установившийся, стационарный с интенсивностью, равной λ (***интенсивность*** — это среднее число событий потока, происходящих в единицу времени; в стационарном потоке интенсивность — постоянная величина, не зависящая от расположения единицы времени на оси времени);

— поток без последействия, т. е. на вероятность появления любого числа событий в любой промежуток времени не влияет ни число событий, ни моменты их появления вне этого промежутка.

Формулы (3.12) и (3.9) идентичны: заменив в (3.12) произведение λt буквой λ, получим формулу (3.9).

ЗАДАЧА 3.3. При установившейся на протяжении суток криминогенной обстановке в городе в среднем за сутки происходят 15 правонарушений. Каково наивероятнейшее число правонарушений за сутки, за час и каковы вероятности этих чисел? Предполагается, что поток правонарушений простейший.

Р е ш е н и е. По условию $\lambda = 15$ (правонарушений в сутки). При $t = 1$ (сутки) $\lambda t = 15$ и наивероятнейших чисел правонарушений за сутки в соответствии с алгоритмом (3.11) два: $m_1^* = \lambda t - 1 = 14$ и $m_2^* = \lambda t = 15$. Используя (3.12), найдем вероятности этих чисел:

$$P_1(14) = \frac{15^{14}}{14!}\,e^{-15} = 0{,}1024;$$

$$P_1(15) = \frac{15^{15}}{15!}\,e^{-15} = 0{,}1024.$$

Как и следовало ожидать, вероятности одинаковы.

Поскольку один час составляет 1/24 часть суток, при $t = 1/24$ (суток) $\lambda t = 15/24$, и наивероятнейшее число правонарушений за час, в соответствии с алгоритмом (3.11), равно целому числу из интервала $(\lambda t - 1;\ \lambda t) = (-9/24;\ 15/24)$, $m^* = 0$. Вероятность этого числа

$$P_{1/24}(0) = \frac{(15/24)^0}{0!}\,e^{-15/24} = 0{,}5353. \quad \blacktriangleleft$$

§ 3.3. Локальная и интегральная формулы Муавра — Лапласа. Функция Лапласа

Французский математик Муавр в 1730 г. предложил формулу вычисления вероятности $P_n(m)$ появления m успехов в n испытаниях Бернулли для случая большого числа n и вероятности успеха в единичном испытании $p = 1/2$, отличающуюся от формулы Бернулли (3.2). Впоследствии французский математик Лаплас обобщил ее на случай произвольного p, отличного от 0 и 1. Эту формулу называют ***локальной формулой Муавра — Лапласа***: при большом числе n испытаний Бернулли с вероятностью успеха в единичном испытании, равной p, $0 < p < 1$, вероятность появления m успехов

$$P_n(m) = \frac{1}{\sqrt{2\pi npq}} e^{-(m-np)^2/(2npq)}. \qquad (3.13)$$

Вероятности, найденные по локальной формуле Муавра — Лапласа (3.13), тем меньше отличаются от вероятностей Бернулли (3.2), чем больше n и чем ближе p к 1/2. Практика показывает, что при n порядка нескольких сотен или еще большем, а также при p, не слишком близкой к 0 или 1, использование формулы (3.13) приводит к удовлетворительным результатам.

При большом числе n и малой вероятности p (близкой к нулю) хорошее приближение к вероятностям Бернулли (3.2) дают вероятности, рассчитанные по формуле Пуассона (3.9).

Рассмотрим нахождение вероятностей по формуле (3.13). В приложении П. 1 приведена таблица значений функции

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad (3.14)$$

при $z \geqslant 0$ и ее график. Эта функция четная, т. е. $\varphi(-z) = \varphi(z)$, что нетрудно видеть из (3.14). Поэтому в таблице содержатся значения $\varphi(z)$ при $z \geqslant 0$; с ростом $z \geqslant 0$ функция $\varphi(z)$ уменьшается, и при $z > 3{,}5$ значения функции $\varphi(z)$ практически равны нулю (оставаясь «чуть» больше нуля). Сопоставив формулы (3.13) и (3.14), получим алгоритм вычисления $P_n(m)$:

$$\left.\begin{array}{l} n \\ p \\ q = 1 - p \\ m \end{array}\right\} \to z = \frac{m - np}{\sqrt{npq}} \xrightarrow[\text{П.1}]{} \varphi(z) \longrightarrow P_n(m) = \frac{1}{\sqrt{npq}} \varphi(z). \qquad (3.15)$$

» **ЗАДАЧА 3.4.** Какова вероятность того, что при $n = 1200$-кратном подбрасывании монеты «герб» выпадет: а) 550 раз; б) наивероятнейшее число раз?

Р е ш е н и е. В задаче $n = 1200$ (велико), «успех» в единичном испытании — это выпадение «герба», при единичном подбрасывании монеты вероятность успеха $p = 1/2$ (не мала).

а) В соответствии с алгоритмом (3.15) получим

$$\left.\begin{array}{l} n = 1200 \\ p = 0{,}5 \\ q = 0{,}5 \\ m = 550 \end{array}\right\} \to z = \frac{550 - 1200 \cdot 0{,}5}{\sqrt{1200 \cdot 0{,}5 \cdot 0{,}5}} = -2{,}89 \xrightarrow[\text{п.1}]{} \varphi(-2{,}89) =$$

$$= 0{,}0060 \to P_{1200}(550) = \frac{0{,}0060}{\sqrt{1200 \cdot 0{,}5 \cdot 0{,}5}} = 0{,}00035.$$

б) Найдем число m^*, при котором вероятность (3.13) принимает максимальное значение. Будем рассматривать $P_n(m)$ как непрерывную функцию аргумента m. Так как показатель степени при числе e не больше нуля, то $P_n(m)$ принимает максимальное значение, если показатель степени равен нулю, т. е. если $m = np$. Итак, $m^* = np$.

В задаче $m^* = 1200 \cdot 0{,}5 = 600$ и в соответствии с алгоритмом (3.15)

$$P_{1200}(600) = \varphi(0)/\sqrt{1200 \cdot 0{,}5 \cdot 0{,}5} = 0{,}3989/7{,}0711 = 0{,}023$$

— и это наибольшая вероятность; вероятности других чисел ($m = 0, 1, 2, \ldots, 599, 601, \ldots, 1200$) меньше. Так, $P_{1200}(550) = 0{,}00035 < 0{,}023$. «

Значения вероятностей, рассчитываемых по локальной формуле Муавра — Лапласа, малы, поэтому более ценную информацию дает значение вероятности $P_n(m_1 < m < m_2)$ того, что в n испытаниях Бернулли число успехов m — целое число из интервала $(m_1; m_2)$. Эту вероятность при большом числе n испытаний Бернулли (порядка нескольких сотен и более) и вероятности p успеха в единичном испытании, не слишком близкой к 0 или 1, находят по ***интегральной формуле Муавра — Лапласа***:

$$P_n(m_1 < m < m_2) = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-t^2/2}\, dt, \qquad (3.16)$$

где $z_1 = (m_1 - np)/\sqrt{npq}$, $z_2 = (m_2 - np)/\sqrt{npq}$.

Рассмотрим нахождение вероятностей по формуле (3.16). Можно убедиться в том, что какие бы знаки ни име-

ли пределы интегрирования z_1 и z_2, всегда имеет место следующая цепочка равенств:

$$\frac{1}{\sqrt{2\pi}}\int_{z_1}^{z_2} e^{-t^2/2}\,dt =$$

$$= \frac{1}{\sqrt{2\pi}}\int_{z_1}^{0} e^{-t^2/2}\,dt + \frac{1}{\sqrt{2\pi}}\int_{0}^{z_2} e^{-t^2/2}\,dt =$$

$$= \frac{1}{\sqrt{2\pi}}\int_{0}^{z_2} e^{-t^2/2}\,dt - \frac{1}{\sqrt{2\pi}}\int_{0}^{z_1} e^{-t^2/2}\,dt = \Phi(z_2) - \Phi(z_1), \quad (3.17)$$

где функция

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{0}^{z_2} e^{-t^2/2}\,dt \quad (3.18)$$

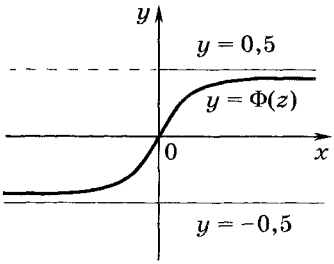


Рис. 3.2

называется ***функцией (интегралом) Лапласа***. Эта функция нечетная: $\Phi(-z) = -\Phi(z)$, поэтому ее график (рис. 3.2) симметричен относительно начала координат; $\Phi(z) \to 0{,}5$, если $z \to \infty$.

Таблица значений функции $\Phi(z)$ при $z \geqslant 0$ приведена в приложении П. 1. Обратим внимание, что при $z > 3{,}5$ значения функции $\Phi(z)$ практически равны 0,5 (оставаясь «чуть» меньше 0,5).

Учитывая преобразования (3.17), запишем интегральную формулу Муавра — Лапласа (3.16) в виде

$$P_n(m_1 < m < m_2) = \Phi\left(\frac{m_2 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{m_1 - np}{\sqrt{npq}}\right). \quad (3.19)$$

Соответствующий алгоритм таков:

$$\left.\begin{array}{l} n \\ p \\ q = 1 - p \\ m_1 \\ m_2 \end{array}\right\} \begin{array}{l} \nearrow z_1 = \dfrac{m_1 - np}{\sqrt{npq}} \xrightarrow[\text{П.1}]{} \Phi(z_1) \\ \searrow z_2 = \dfrac{m_2 - np}{\sqrt{npq}} \xrightarrow[\text{П.1}]{} \Phi(z_2) \end{array} \begin{array}{l} \searrow \\ \nearrow \end{array} \begin{array}{l} P_n(m_1 < m < m_2) = \\ = \Phi(z_2) - \Phi(z_1). \end{array} \quad (3.20)$$

ЗАДАЧА 3.5. Для лица, дожившего до 70-летнего возраста, вероятность заболевания на 71-м году равна 0,1. Застрахована группа в 1000 человек 70-летнего возраста; годовой страховой взнос 12 ден. ед. В случае заболевания застрахованный получит 100 ден. ед. Какова вероятность того, что:

а) страховое учреждение к концу года окажется в убытке;
б) его доход превысит 2000 ден. ед.?

Р е ш е н и е. По условию $n = 1000$ (чел.) (велико), а вероятность травматизма $p = 0{,}1$ (не мала), $q = 0{,}9$. Если m — количество застрахованных, заболевших на 71-м году жизни, то:

а) страховое учреждение окажется в убытке, если $100\,m > 12 \cdot 1000$, или если $m > 120$. Поэтому, учитывая, что в группе 1000 человек, имеем

$$P_{1000}(m > 120) =$$

$$= P_{1000}(120 < m < 1001) \underset{(3.19)}{=} \Phi\left(\frac{1001 - 100}{\sqrt{1000 \cdot 0{,}1 \cdot 0{,}9}}\right) -$$

$$- \Phi\left(\frac{120 - 100}{\sqrt{1000 \cdot 0{,}1 \cdot 0{,}9}}\right) = \Phi(94{,}97) - \Phi(2{,}11) \underset{\text{П. 1}}{=}$$

$$\underset{\text{П. 1}}{=} 0{,}49(9) - 0{,}4821 = 0{,}0187;$$

б) доход страхового учреждения превысит 2000 ден. ед., если $12 \cdot 1000 - 100m > 2000$, или если $100 > m$. Имеем

$$P_{1000}(m < 100) = P_{1000}(-1 < m < 100) \underset{(3.19)}{=} \Phi\left(\frac{100 - 100}{9{,}49}\right) -$$

$$- \Phi\left(\frac{-1 - 100}{9{,}49}\right) = \Phi(0) - \Phi(-10{,}6) = \Phi(0) + \Phi(10{,}6) \underset{\text{П. 1}}{=}$$

$$\underset{\text{П. 1}}{=} 0 + 0{,}49(9) = 0{,}49(9). \quad \blacktriangleleft$$

§ 3.4. Формула геометрической вероятности

Ранее нас интересовала вероятность $P_n(m)$ того, что при проведении определенного числа n испытаний Бернулли будет m успешных ($m = 0, 1, \ldots, n$). Возможна и другая постановка вопроса: проводятся испытания Бернулли (с вероятностью успеха p в каждом из испытаний) до тех пор, пока очередное испытание не будет успешным; какова вероятность того, что впервые успех произойдет в k-м испытании ($k = 1, 2, \ldots$)?

Выведем формулу для нахождения этой вероятности.

$\gg$ Введем события A_i и $\overline{A}_i$ — соответственно успех и неудача в i-м испытании, $i = 1, 2, \ldots$. Так как речь идет об испытаниях Бернулли, то

$$P(A_i) = p \text{ и } P(\overline{A}_i) = 1 - p = q, \qquad i = 1, 2, \ldots .$$

Учитывая независимость испытаний, рассчитаем вероятности следующих событий:

Формулировка события	Символьная запись события	Вероятность события
Впервые успех произойдет: в первом испытании	A_1	$P(A_1) = p$
во втором испытании	$\bar{A}_1 \cap A_2$	$P(\bar{A}_1 \cap A_2) = P(\bar{A}_1) \cdot P(A_2) = qp$
в третьем испытании	$\bar{A}_1 \cap \bar{A}_2 \cap A_3$	$P(\bar{A}_1 \cap \bar{A}_2 \cap A_3) = P(\bar{A}_1)P(\bar{A}_2)P(A_3) = q^2p$
в четвертом испытании	$\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3 \cap A_4$	$P(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3 \cap A_4) = q^3p$
.....	.....	.....
в k-м испытании	$\bar{A}_1 \cap \bar{A}_2 \cap \ldots \ldots \cap \bar{A}_{k-1} \cap A_k$	$P(\bar{A}_1 \cap \bar{A}_2 \cap \ldots \ldots \cap \bar{A}_{k-1} \cap A_k) = q^{k-1}p$
.....	.....	.....

≪

Итак, вероятность $P_k^{(1)}$ того, что впервые успех произойдет в k-м испытании Бернулли

$$P_k^{(1)} = q^{k-1}p, \qquad k = 1, 2, \ldots, \tag{3.21}$$

где p — вероятность успеха в единичном испытании.

Формула (3.21) называется ***формулой геометрической вероятности*** («геометрической» в силу того, что последовательность вероятностей p, qp, q^2p, q^3p,... является геометрической прогрессией). ***Геометрический ряд распределения*** вероятностей имеет вид:

Таблица 3.3

Число испытаний k до первого успешного, включая его	1	2	3	4	...
Вероятность $P_k^{(1)} = q^{k-1}p$	p	qp	q^2p	q^3p	...

$\Sigma = 1$

Ряд бесконечный, но сумма вероятностей равна 1.

≫ Действительно,

$$p + qp + q^2p + q^3p + \ldots = p(1 + q + q^2 + q^3 + \ldots) = p\frac{1}{1-q} = p\frac{1}{p} = 1.$$

Здесь учтено, что $0 < q < 1$; следовательно, бесконечная геометрическая прогрессия $1, q, q^2, q^3 \ldots$ со знаменателем q — убывающая, а сумма членов такой прогрессии равна $1/(1-q)$. ≪

Очевидно, что самую большую вероятность имеет появление успеха в первом испытании ($p > qp > q^2p > \ldots$).

Вероятность $P_k^{(m)}$ того, что потребуется провести k испытаний ($k = 1, 2, \ldots$) в ожидании появления успеха m раз (m-й раз успех должен появиться в k-м испытании), находится по формуле

$$P_k^{(m)} = C_{k-1}^{m-1} p^m q^{k-m}, \qquad k = m, m+1, m+2, \ldots . \tag{3.22}$$

≫ Действительно, вероятность любого фиксированного размещения m успехов на k местах равна $p^m q^{k-m}$, но так как m-й успех должен произойти в k-м испытании, то для размещения $m-1$ успехов остается $k-1$ мест и число вариантов таких размещений равно C_{k-1}^{m-1} — числу сочетаний из $k-1$ по $m-1$. ≪

При $m = 1$

$$P_k^{(m)} = C_{k-1}^{m-1} p^m q^{k-m} \underset{m=1}{=} q^{k-1}p, \qquad k = 1, 2, 3, \ldots .$$

Мы получили формулу геометрической вероятности.

ЗАДАЧА 3.6. Вероятность того, что взрослый человек поддерживает правительственную программу, равна 0,6. Какова вероятность того, что: а) третий опрошенный будет первым, поддерживающим программу; б) шестой опрошенный будет третьим человеком, который поддерживает программу.

Р е ш е н и е. По условию $p = 0{,}6$, $q = 0{,}4$.

а) Здесь $m = 1$, $k = 3$ и, в соответствии с (3.21),

$$P_3^{(1)} = 0{,}4^{3-1} \cdot 0{,}6 = 0{,}096.$$

б) Здесь $m = 3$, $k = 6$ и, в соответствии с (3.22),

$$P_6^{(3)} = C_5^2 \cdot 0{,}6^3 \cdot 0{,}4^3 = \frac{5!}{2!3!} \cdot 0{,}0138 = 0{,}138. \quad \ll$$

УПРАЖНЕНИЯ

1. Считая вероятность рождения мальчика равной 1/2, найдите вероятность того, что в семье, планирующей иметь семь детей, будет: а) четыре мальчика и три девочки; б) число мальчиков от трех до пяти; в) число мальчиков равно наивероятнейшему числу; г) мальчиков больше, чем девочек; д) хотя бы один мальчик.

2. Случайно встреченное лицо с вероятностью, близкой к 0,2, может оказаться брюнетом, с вероятностью 0,3 — шатеном, с вероятно-

стью 0,4 — блондином и с вероятностью 0,1 — рыжим. Какова вероятность того, что среди пяти случайно встреченных лиц: а) три блондина и два шатена; б) хотя бы один рыжий; в) не меньше четырех блондинов?

3. Сколько раз придется бросить игральную кость, чтобы наивероятнейшее число появлений единицы было бы равно 32?

4. Учебник издан тиражом 10 000 экземпляров. Вероятность того, что наугад взятый экземпляр сброшюрован неправильно, равна 0,0001. Найдите вероятность того, что тираж содержит хотя бы один неправильно сброшюрованный экземпляр. Каково наивероятнейшее число неправильно сброшюрованных экземпляров и вероятность этого числа?

5. Французский естествоиспытатель Ж. Бюффон подбросил монету 4040 раз, при этом «герб» выпал 2048 раз. Какова вероятность того, что при повторном проведении 4040 подбрасываний: а) «герб» появится вновь 2048 раз; б) «герб» появится менее 2048 раз; в) число выпадений «герба» будет отличаться от наивероятнейшего числа по абсолютной величине менее, чем на 100?

6. Среди выпускаемых заводом автомобилей 5% некомплектны. Какова вероятность того, что десятый проверенный автомобиль будет: а) первым некомплектным; б) вторым некомплектным?

7. Двое поочередно бросают игральную кость. Выиграет тот, у кого впервые появится шестерка. Найдите вероятность выигрыша для каждого игрока.

ЧАСТЬ 2

Случайные величины и модели законов распределения вероятностей

ГЛАВА 4

Способы задания и числовые характеристики случайной величины

Наряду со случайным событием и вероятностью понятие случайной величины является важнейшим в теории вероятностей. В этой главе изучаются способы, с помощью которых случайные величины могут быть описаны и охарактеризованы; приводятся примеры использования случайных величин в решении задач рыночной экономики.

§ 4.1. Понятие случайной величины. Функция распределения вероятностей и ее свойства

Ограничимся качественным определением понятия случайной величины. ***Случайной величиной***, связанной с данным опытом, называется переменная величина, которая в результате опыта может принять то или иное числовое значение, причем заранее неизвестно какое именно[1]. Условимся в дальнейшем случайные величины обозначать прописными конечными буквами латинского алфавита: ... X, Y, Z. Напомним, для обозначения случайных событий ранее были использованы начальные буквы этого алфавита: A, B, C, ...

Различают случайные величины следующих двух основных типов:

— ***дискретная*** — возможные значения которой поддаются перечислению, счету, или, иначе, множество значений которой счетно;

— ***непрерывная*** — возможные значения которой не поддаются перечислению, они «непрерывно» заполняют некоторый отрезок (или отрезки); множество значений этой величины несчетно.

[1] Строгое математическое определение понятия случайной величины можно найти в [11, 13].

Примеры дискретных случайных величин: число успехов в n испытаниях Бернулли (возможные значения это числа — 0, 1, 2, ... , n); относительная частота числа успехов в n испытаниях Бернулли $\left(\text{ее возможные значения } 0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\right)$; число испытаний Бернулли до появления успеха первый раз (его возможные значения 1, 2, ...).

Если в первых двух примерах число значений случайной величины конечно, то в последнем число значений бесконечно, но и в том, и в другом случае эти значения можно перечислить.

Частным случаем дискретной случайной величины является ***альтернативная,*** или ***булевская*** (названная по имени математика Дж. Буля, XIX в.), принимающая только два значения: 1 — при успешном испытании, 0 — при неудачном, с одинаковыми или различными вероятностями.

Примеры непрерывных случайных величин: время безотказной работы лампы после ее включения; абсцисса или ордината точки, брошенной наудачу на плоскость (прямоугольную систему координат); размер уклонения точки попадания снаряда от центра цели.

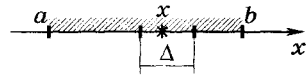
Для того чтобы задать случайную величину, необходимо указать все ее значения и уметь находить все возможные вероятности.

Для дискретной случайной величины X говорить о вероятности $P(X = x)$ того, что X примет значение x из множества всех возможных, имеет смысл, и всегда $0 < P(X = x) < 1$. Так, если X — число успехов в n испытаниях Бернулли, то, в соответствии с формулой Бернулли (3.2), вероятность того, что X примет значение x ($x = 0, 1, \ldots, n$), равна $P(X = x) = P_n(x) = C_n^x p^x (1-p)^{n-x}$, где p — вероятность успеха в единичном испытании, и всегда $0 < P(X = x) < 1$. Если X — число испытаний Бернулли до появления успеха первый раз, включая и успешное испытание, то, в соответствии с формулой геометрической вероятности (3.21),

$$P(X = x) = P_x^{(1)} = (1-p)^{x-1} p,$$

где $x = 1, 2, 3, \ldots$, и всегда $0 < P(X = x) < 1$.

Для большого класса непрерывных случайных величин говорить о вероятности $P(X = x)$ того, что непрерывная величина X, значениями которой являются все точки отрезка $[a, b]$, примет фиксированное значение x из этого отрезка, бессмысленно: для фиксированного $x \in [a, b]$ вероятность $P(X = x) = 0$.



$P(X \in \Delta) = cl_\Delta$

Рис. 4.1

Убедимся в этом на примере *равномерно распределенной на отрезке* $[a, b]$ *случайной величины* — непрерывной величины X, для которой вероятность $P(X \in \Delta)$ того, что X примет любое значение из подотрезка Δ отрезка $[a, b]$, не зависит от расположения Δ на этом отрезке и пропорциональна длине отрезка Δ. Аналогом таких требований к вероятности $P(X \in \Delta)$ является выражение «точка X наудачу брошена на отрезок $[a, b]$» (рис. 4.1). В соответствии с геометрическим подходом к нахождению вероятности (см. § 1.2) $P(X \in \Delta) = l_\Delta/(b - a) \neq 0$, где $(b - a)$ — длина отрезка $[a, b]$, а l_Δ — длина отрезка Δ. Но тогда $P(X = x) = (\text{«длина точки } x\text{»})/(b - a) = 0$. Имеет место парадокс: случайное событие, состоящее в том, что X примет значение, равное x, возможно (ведь брошенная на $[a, b]$ точка X может попасть в точку x), а вероятность этого события равна нулю. На самом деле это не более парадоксально, чем представление об отрезке, имеющем определенную длину, тогда как ни одна точка отрезка длиной не обладает. Сколь угодно малый отрезок Δ, содержащий точку x, длину имеет, а точка x длины не имеет. Аналогично, $P(X \in \Delta) \neq 0$, а $P(X = x) = 0$. Напомним, что в силу теоремы Бернулли, с ростом числа n «бросаний» точки X на $[a, b]$ уверенность в незначительном отклонении относительной частоты m/n попадания т. X в т. x от вероятности $P(X = x)$ увеличивается. Отсюда, если учесть, что $P(X = x) = 0$, следует, что при неограниченном повторении бросаний событие $X = x$ будет появляться сколь угодно редко, и совсем не следует, что это событие невозможно.

Итак, говорить о вероятности $P(X = x)$ того, что X примет фиксированное значение x из своего множества значений, имеет смысл только для дискретной, но не для непрерывной, случайной величины. Для обоих типов случайных величин имеет смысл говорить о вероятности попадания величины X на отрезок, на интервал, на полуинтервал. В теории вероятностей используют вероятность попадания X на интервал $(-\infty, x)$, где x — произвольное действительное число, т. е. вероятность $P(X < x)$, которая зависит от x и является его некоторой функцией. Эту функцию называют ***функцией распределения вероятностей*** (или просто функцией распределения) случайной величины X и обозначают $F_X(x)$:

$$F_X(x) = P(X < x), \qquad (4.1)$$

где x — произвольное действительное число.

Функция распределения — универсальная характеристика случайной величины: она существует и для дискретных, и для непрерывных случайных величин и полностью определяет случайную величину.

Далее приведем примеры, подтверждающие, что при известной функции распределения случайная величина полностью определена, т. е. можно указать множество ее значений и находить все возможные вероятности. Но сначала рассмотрим свойства функции распределения:

1^0. $0 \leqslant F_X(x) \leqslant 1$ *при любом действительном* x (свойство вытекает из того, что $F_X(x)$ — это вероятность).

2^0. $F_X(-\infty) = 0$ и $F_X(+\infty) = 1$.

Не приводя строгого доказательства, проиллюстрируем правомерность равенств 2^0 на «прагматичном» уровне. Так как неравенство $X < -\infty$ невозможно, то $P(X < -\infty) = 0$; и $F_X(-\infty) = P(X < -\infty) = 0$. Аналогично, так как неравенство $X < +\infty$ достоверно, то $P(X < +\infty) = 1$, и $F_X(+\infty) = P(X < +\infty) = 1$.

3^0. $F_X(x)$ — *неубывающая функция, т. е. при любых* x_1 *и* x_2 $(x_2 > x_1)$ *имеет место неравенство* $F_X(x_2) \geqslant F_X(x_1)$.

≫ Рассмотрим (рис. 4.2) три события: событие A, состоящее в том, что $X < x_2$; событие B, состоящее в том, что $X < x_1$; событие C, состоящее в том, что $x_1 \leqslant X < x_2$.

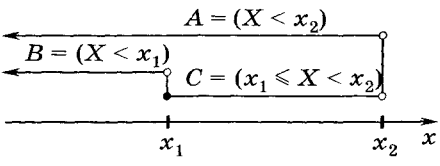


Рис. 4.2

На рисунке • — включенная точка, а ∘ — исключенная. Учитывая, что событие $A = B \cup C$, а B и C несовместны, по теореме о вероятности объединения несовместных событий получим

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

или

$$P(X < x_2) = P(X < x_1) + P(x_1 \leqslant X < x_2).$$

Воспользовавшись формулой (4.1), имеем

$$F_X(x_2) = F_X(x_1) + P(x_1 \leqslant X < x_2),$$

или

$$P(x_1 \leqslant X < x_2) = F_X(x_2) - F_X(x_1). \qquad (4.2)$$

Так как $P(x_1 \leqslant X < x_2) \geqslant 0$, то и $F_X(x_2) - F_X(x_1) \geqslant 0$, или $F_X(x_2) \geqslant F_X(x_1)$, что и требовалось доказать. ≪

Рассмотрим равенство (4.2), которое определяет вероятность того, что величина X примет значение из полуинтер-

вала $[x_1, x_2)$. Будем неограниченно уменьшать этот интервал, полагая, что $x_2 \to x_1$. В пределе вместо вероятности попадания X на полуинтервал получим вероятность того, что X примет значение x_1:

$$P(X = x_1) = \lim_{x_2 \to x_1} P(x_1 \leqslant X < x_2) \underset{(4.2)}{=}$$

$$\underset{(4.2)}{=} \lim_{x_2 \to x_1} [F_X(x_2) - F_X(x_1)] = \lim_{x_2 \to x_1} F_X(x_2) - F_X(x_1). \quad (4.3)$$

Значение этой разности зависит от того, терпит ли функция $F_X(x)$ в точке x_1 разрыв (рис. 4.3, *а*) или же она в этой точке непрерывна (рис. 4.3, *б*). Если в точке x_1 функция $F_X(x)$ имеет разрыв, скачок, то разность равна размеру скачка ΔF функции в точке x_1, и, согласно (4.3), $P(X = x_1) = \Delta F$. Если в точке x_1 функция $F_X(x)$ непрерывна, то разность равна нулю, и $P(X = x_1) = 0$.

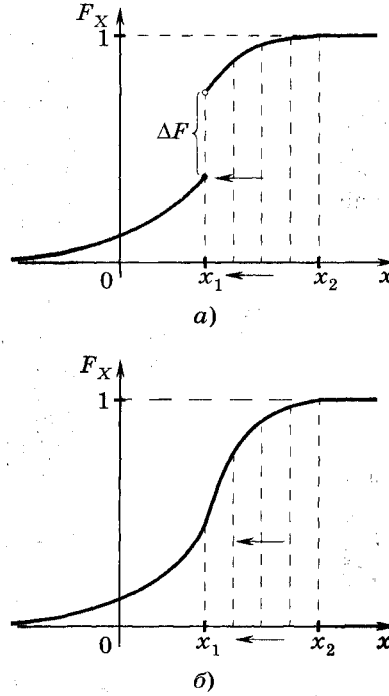


Рис. 4.3

Исходя из сказанного делаем вывод, что вероятность «значения x», в котором происходит разрыв, скачок функции распределения случайной величины, равна размеру скачка, а вероятность «значения x», в котором функция непрерывна, равна нулю.

В дальнейшем убедимся, что функция распределения дискретной случайной величины разрывная, скачкообразная, причем скачки происходят в точках, равных значениям величины, а размеры скачков равны вероятностям соответствующих значений. Будем в дальнейшем предполагать, что непрерывная случайная величина, значениями которой являются все точки некоторого $[a, b]$, имеет функцию распределения, непрерывную на этом отрезке. Следовательно, вероятность того, что непрерывная случайная величина примет фиксированное значение x, принадлежащее $[a, b]$, равна нулю. Правомерность этого положения обоснована в начале параграфа для равномерно распределенной на $[a, b]$ случайной вели-

чины (в § 5.2 будет показано, что функция распределения этой величины непрерывна).

Так как для непрерывной случайной величины X (с непрерывной функцией распределения) $P(X = x) = 0$ для любого фиксированного x, то $P(X \leqslant x) = P(X < x)$. Учитывая равенство (4.2), имеющее место для случайной величины любого типа, для непрерывной величины X получим

$$P(x_1 < X < x_2) \underset{P(X=x_2)=0}{=} P(x_1 < X \leqslant x_2) \underset{P(X=x_1)=0}{=} P(x_1 \leqslant X \leqslant x_2) \underset{P(X=x_2)=0}{=} P(x_1 \leqslant X < x_2) \underset{(4.2)}{=} F_X(x_2) - F_X(x_1). \qquad (4.4)$$

Принимая во внимание эту цепочку равенств, в дальнейшем при нахождении всевозможных вероятностей для непрерывной случайной величины не будем делать различий между отрезком $[x_1, x_2]$, интервалом (x_1, x_2) и полуинтервалами $(x_1, x_2]$ и $[x_1, x_2)$. Напомним, если X дискретная, то только $P(x_1 \leqslant X < x_2) = F_X(x_2) - F_X(x_1)$.

▶ **ПРИМЕР 4.1.** Студенту предстоит сдать два ($n = 2$) экзамена. Вероятность сдачи экзамена $p = 0{,}6$. Пусть случайная величина X — число экзаменов, сданных (в будущем) студентом. Здесь X — дискретная случайная величина, ее значения 0, 1, 2. Найдем их вероятности, используя формулу Бернулли

$$P(X = x) = C_n^x p^x (1-p)^{n-x}, \qquad x = 0, 1, 2.$$

Получим $P(X = 0) = C_2^0 \cdot 0{,}6^0 \cdot 0{,}4^2 = 0{,}16$; $P(X = 1) = 0{,}48$; $P(X = 2) = 0{,}36$.

Значения x величины X и вероятности $P(X = x)$ приведены в таблице 4.1. В таблице 4.2 приведены значения функции распределения $F_X(x)$ величины X, найденные при любом действительном x.

Таблица 4.1

x	0	1	2
$P(X = x)$	0,16	0,48	0,36

Таблица 4.2

x	$(-\infty; 0]$	$(0; 1]$	$(1; 2]$	$(2; +\infty)$
$F_X(x) = P(X < x)$	0	0,16	0,16 + 0,48 = 0,64	0,64 + 0,36 = 1

Поясним расчет значений функций распределения $F_X(x)$. Пусть $x \in (-\infty, 0]$, например $x = 0$. Тогда согласно (4.1), $F_X(0) = P(X < 0)$, но так как величина X не может быть меньше нуля (см. табл. 4.1), то $P(X < 0) = 0$, следовательно, $F_X(0) = 0$. Очевидно, что и при любом другом $x \in (-\infty, 0]$ значение $F_X(x) = 0$.

Пусть $x \in (0, 1]$, например $x = 0{,}01$. Тогда

$$F_X(0{,}01) = P(X < 0{,}01) = P(X = 0) = 0{,}16.$$

Пусть $x \in (1, 2]$, например $x = 2$. Тогда, учитывая, что величина X имеет два значения, меньших числа 2, — это 0 и 1, получим

$$F_X(2) = P(X < 2) = P((X = 0) \cup (X = 1)).$$

Последняя вероятность — это вероятность объединения двух несовместных событий: события A, состоящего в том, что «X примет значение 0», и события B — «величина X примет значение 1», поэтому

$$P((X = 0) \cup (X = 1)) = P(A \cup B) = P(A) + P(B) =$$
$$= P(X = 0) + P(X = 1) = 0{,}16 + 0{,}48 = 0{,}64.$$

И, наконец, при любом $x \in (2, +\infty)$ значение $F_X(x) = 1$, например, при $x = 2 + \varepsilon$, где $\varepsilon > 0$,

$$F_x(2 + \varepsilon) = P(X < 2 + \varepsilon) =$$
$$= P((X = 0) \cup (X = 1) \cup (X = 2)) =$$
$$= P(X = 0) + P(X = 1) +$$
$$+ P(X = 2) = 0{,}16 + 0{,}48 +$$
$$+ 0{,}36 = 1.$$

Аналитическая запись функции распределения, заданной таблицей 4.2, такова:

$$F_X(x) = \begin{cases} 0 & \text{при } x \leqslant 0, \\ 0{,}16 & \text{при } 0 < x \leqslant 1, \\ 0{,}64 & \text{при } 1 < x \leqslant 2, \\ 1 & \text{при } x > 2. \end{cases}$$

Графическое изображение ряда распределения вероятностей, заданного таблицей 4.1, представлено на рисунке 4.4, *а*; график функции распределения приведен на рисунке 4.4, *б*. Из рисунка 4.4, *б* видно, что: функция распределения $F_X(x)$

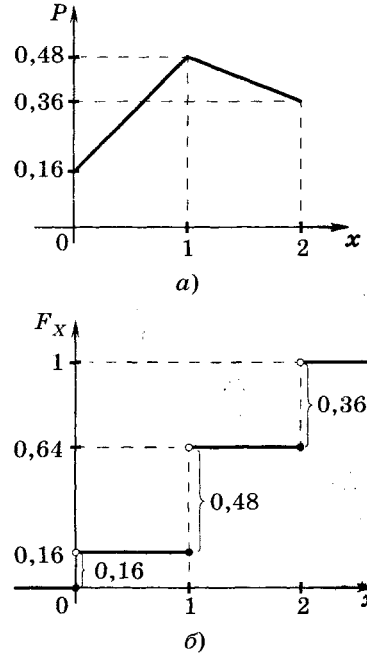


Рис. 4.4

скачкообразная; скачки функции имеют место в точках 0, 1, 2, соответствующих значениям величины X (табл. 4.1), а величины скачков, равные 0,16, 0,48, 0,36, совпадают с вероятностями этих значений, что соответствует сделанному выше утверждению относительно поведения функции распределения дискретной случайной величины. ◀

▶▶ **ЗАДАЧА 4.1.** Функция

$$F_X(x) = \begin{cases} 0 & \text{при } x \leqslant -1, \\ 0{,}3 & \text{при } -1 < x \leqslant 2, \\ 0{,}8 & \text{при } 2 < x \leqslant 4, \\ 1 & \text{при } x > 4. \end{cases}$$

Постройте график функции. Можно ли эту функцию назвать функцией распределения? При положительном ответе на вопрос найдите:

а) все возможные значения случайной величины и их вероятности;

б) $P(X < 2)$, $P(X > 1)$, $P(0{,}5 \leqslant X < 5)$, $P(1 < X \leqslant 4)$.

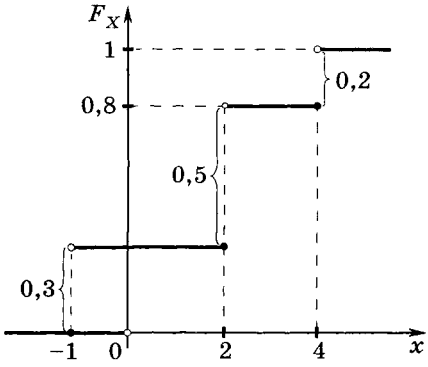


Рис. 4.5

Р е ш е н и е. График функции $F_X(x)$ изображен на рисунке 4.5. Функция $F_X(x)$ обладает всеми свойствами функции распределения, а именно: $0 \leqslant F_X(x) \leqslant 1$; $F_X(-\infty) = 0$, а $F_X(+\infty) = 1$; $F_X(x)$ — неубывающая функция. Поэтому $F_X(x)$ — это функция распределения.

а) $F_X(x)$ — скачкообразная функция со скачками в точках −1, 2, 4 — это и есть значения случайной величины X; вероятности этих значений равны величинам соответствующих скачков: 0,3; 0,5; 0,2. Таким образом, X — дискретная случайная величина, и ряд распределения вероятностей имеет вид

x	−1	2	4	
$P(X = x)$	0,3	0,5	0,2	$\Sigma = 1$

б) Требуемые вероятности найдем двумя способами: в первом случае используем функцию распределения; во втором — ряд распределения. Итак,

$$P(X < 2) \underset{(4.1)}{=} F_X(2) = 0{,}3;$$

с другой стороны,

$$P(X < 2) = P(X = -1) = 0{,}3$$

(судя по ряду распределения имеется только одно значение величины X, меньшее 2, это -1, поэтому $P(X < 2) = P(X = -1)$.

Далее имеем

$$P(X > 1) = 1 - P(X \leqslant 1) = 1 - [P(X < 1) + P(X = 1)] = 1 - [F_X(1) + P(X = 1)] = 1 - [0{,}3 + 0] = 0{,}7$$

(при $x = 1$ функция $F_X(x)$ не имеет скачка, поэтому $x = 1$ не является значением величины X и $P(X = 1) = 0$); с другой стороны,

$$P(X > 1) = P((X = 2) \cup (X = 4)) = P(X = 2) + P(X = 4) = 0{,}5 + 0{,}2 = 0{,}7$$

(среди значений величины X есть два значения, больших 1, это 2 и 4, поэтому $P(X > 1) = P((X = 2) \cup (X = 4))$.

Аналогично

$$P(0{,}5 \leqslant X < 5) \underset{(4.2)}{=} F_X(5) - F_X(0{,}5) = 1 - 0{,}3 = 0{,}7;$$

с другой стороны,

$$P(0{,}5 \leqslant X < 5) = P((X = 2) \cup (X = 4)) = P(X = 2) + P(X = 4) = 0{,}5 + 0{,}2 = 0{,}7.$$

И, наконец,

$$P(1 < X \leqslant 4) = P(1 \leqslant X < 4) - P(X = 1) + P(X = 4) = [F_X(4) - F_X(1)] - 0 + 0{,}2 = (0{,}8 - 0{,}3) + 0{,}2 = 0{,}7$$

(при $x = 1$ функция $F_X(x)$ не имеет скачка, поэтому $P(X = 1) = 0$; при $x = 4$ функция $F_X(x)$ делает скачок, величина которого равна 0,2, поэтому $P(X = 4) = 0{,}2$); с другой стороны,

$$P(1 < X \leqslant 4) = P((X = 2) \cup (X = 4)) = P(X = 2) + P(X = 4) = 0{,}5 + 0{,}2 = 0{,}7.$$

ЗАДАЧА 4.2. Функция

$$F_X(x) = \begin{cases} 0 & \text{при } x \leqslant 0, \\ 4x^2 & \text{при } 0 < x \leqslant 0{,}5, \\ 1 & \text{при } x > 0{,}5. \end{cases} \qquad (4.5)$$

Постройте график этой функции. Можно ли ее назвать функцией распределения? Найдите:

а) множество значений случайной величины X;

б) $P(X = 2)$, $P(0{,}2 \leqslant X < 0{,}3)$, $P(0{,}4 \leqslant X \leqslant 0{,}6)$ и дайте их геометрическую интерпретацию, используя график функции $F_X(x)$.

Замечание. Обратим внимание на то, что функция (4.5) непрерывна на всей числовой оси $(-\infty, +\infty)$, в том числе и в «граничных» точках $x = 0$ и $x = 0{,}5$. Поэтому записи (4.5) идентична, например, следующая запись:

$$F_X(x) = \begin{cases} 0 & \text{при } x < 0, \\ 4x^2 & \text{при } 0 \leqslant x \leqslant 0{,}5, \\ 1 & \text{при } x > 0{,}5. \end{cases} \tag{4.6}$$

Решение. График функции $F_X(x)$ изображен на рисунке 4.6. Так как $F_X(x)$ обладает всеми свойствами функции распределения: $0 \leqslant F_X(x) \leqslant 1$; $F_X(-\infty) = 0$, а $F_X(+\infty) = 1$; $F_X(x)$ — неубывающая функция, это функция распределения.

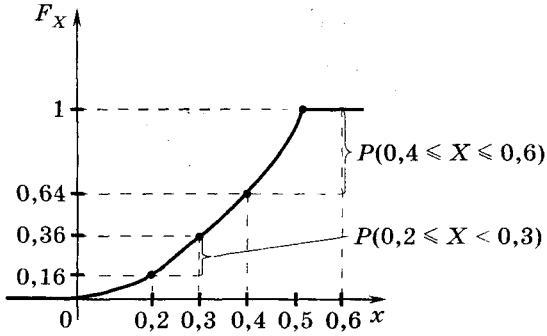


Рис. 4.6

а) Множество значений величины X — это все точки отрезка $[0; 0{,}5]$: каким бы малым ни был подотрезок $[a, b]$ этого отрезка

$$P(a \leqslant X < b) \underset{(4.2)}{=} F_X(b) - F_X(a) \underset{(4.6)}{=}$$

$$\underset{(4.6)}{=} 4b^2 - 4a^2 \neq 0, \ (a, b \in [0; 0{,}5],$$

поэтому $F_X(a) = 4a^2$, $F_X(b) = 4b^2$).

Для любого отрезка из интервалов $(-\infty, 0)$ и $(0{,}5, +\infty)$ аналогичная вероятность равна нулю. Итак, X — непрерывная случайная величина, заданная на отрезке $[0; 0{,}5]$.

б) Так как $F_X(x)$ — непрерывная функция распределения случайной величины, заданной на $[0; 0{,}5]$, то $P(X = x) = 0$ для любого фиксированного $x \in [0; 0{,}5]$ и имеет место равенство (4.4). Учитывая это, получим $P(X = 2) = 0$; $P(0{,}2 \leqslant X < 0{,}3) = F_X(0{,}3) - F_X(0{,}2) = 4 \cdot 0{,}3^2 - 4 \cdot 0{,}2^2 = 0{,}2$. Эта вероятность на рисунке 4.6 равна разности ординат в точках $x = 0{,}3$ и $x = 0{,}2$.

Имеем $P(0,4 \leqslant X \leqslant 0,6) = F_X(0,6) - F_X(0,4) = 1 - 4 \cdot 0,4^2 = 0,36$ (в силу (4.4) $P(0,4 < X < 0,6) = P(0,4 < X \leqslant 0,6) = P(0,4 \leqslant X \leqslant 0,6) = P(0,4 \leqslant X < 0,6) = F_X(0,6) - F_X(0,4) = 0,36$). Эта вероятность на рисунке 4.6 определяется разностью ординат в точках $x = 0,6$ и $x = 0,4$. ◀

§ 4.2. Ряд распределения вероятностей и функция плотности вероятности

В § 4.1 было показано, что случайная величина, как дискретная, так и непрерывная, полностью определена (т. е. известны ее значения и можно найти все возможные вероятности), если задана ее функция распределения. Однако существуют и другие способы задания случайной величины: дискретная случайная величина задается рядом распределения вероятностей, а непрерывная — функцией плотности вероятностей. Зная ряд распределения, всегда можно получить функцию распределения (подтверждение этому — пример 4.1); и из функции плотности можно получить, как будет показано ниже, функцию распределения.

Функцию распределения вероятностей; или ряд распределения вероятностей для дискретной случайной величины и функцию плотности вероятности для непрерывной случайной величины называют ***законом распределения вероятностей.***

Дискретная случайная величина. Ранее неоднократно приводились примеры ряда распределения для конкретных случайных величин. В общем случае ***ряд распределения вероятностей*** случайной величины X показывает «распределение» вероятностей по значениям случайной величины и задается в форме таблицы 4.3, в которой x_1, x_2, ... , x_n — *расположенные в возрастающем порядке все возможные значения случайной величины*; p_1, p_2, ... , p_n — вероятности этих значений: $p_1 = P(X = x_1)$, $p_2 = P(X = x_2)$, ... , $p_n = P(X = x_n)$.

Таблица 4.3

x	x_1	x_2	...	x_n
$P(X = x)$	p_1	p_2	...	p_n

В таблице 4.3 число значений величины X конечно (равно n); при бесконечном числе значений ряд распределения выглядит так:

Таблица 4.4

x	x_1	x_2	...	x_n	...
$P(X = x)$	p_1	p_2	...	p_n	...

Из определения ряда распределения вытекают следующие его свойства:

1^0. $0 < p_i < 1$, $i = 1, 2, \ldots, n$,

так как $p_i = P(X = x_i)$, а событие $X = x_i$, состоящее в том, что случайная величина X примет значение x_i, — случайное, но не достоверное и не невозможное.

2^0. $\sum\limits_{i=1}^{n} p_i = 1$, *или* $\sum\limits_{i=1}^{n} P(X = x_i) = 1$,

так как события $x = x_1$, $x = x_2$, ... , $x = x_n$ образуют полную группу попарно несовместных событий.

Графическое изображение ряда распределения вероятностей называют ***многоугольником распределения вероятностей***. Это ломаная линия, звенья которой соединяют соседние точки с координатами $(x_i; p_i)$, $i = 1, 2, \ldots, n$.

Функция распределения $F_X(x)$ величины X, ряд распределения которой задан таблицей 4.5, такова:

Таблица 4.5

x	$(-\infty, x_1]$	$(x_1, x_2]$	$(x_2, x_3]$	$(x_3, x_4]$	...	$(x_{n-1}, x_n]$	$(x_n, +\infty)$
$F_X(x) = P(X < x)$	0	p_1	$p_1 + p_2$	$p_1 + p_2 + p_3$	...	$p_1 + p_2 + \ldots + p_{n-1}$	$p_1 + p_2 + \ldots + p_n = 1$

Напомним, что $F_X(x)$ — скачкообразная функция; ее часто называют ***функцией накопленных вероятностей***. Зная $F_X(x)$, можно «восстановить» ряд распределения: скачки функции $F_X(x)$ происходят в точках x_1, x_2, ... , x_n — значениях величины X, а величины скачков равны вероятностям p_1, p_2, ... , p_n этих значений.

Непрерывная случайная величина. Будем рассматривать непрерывные случайные величины, функции распределения которых не только всюду непрерывны, но и всюду дифференцируемы (такие величины представляют наибольший практический интерес). Удобной формой задания подобной случайной величины X является ***функция плотности вероятности*** $f_X(x)$, определяемая следующим образом.

Рассмотрим отношение вероятности попадания величины X на участок $[x, x+\Delta x)$, $P(x \leqslant X < x+\Delta x) = F_X(x+\Delta x) - F_X(x)$, к длине Δx участка, т. е. «среднюю вероятность», приходящуюся на единицу длины участка, и Δx устремим к нулю. В пределе получим производную функции $F_X(x)$, которую обозначим $f_X(x)$:

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

$$f_X(x) = F'_X(x); \tag{4.7}$$

$$f_X(x) = \lim_{\Delta x \to 0} \frac{P(x \leqslant X < x+\Delta x)}{\Delta x}, \tag{4.8}$$

т. е. $f_X(x)$ — это как бы плотность вероятности в точке x. Поэтому функцию $f_X(x)$ называют плотностью вероятности величины X, реже — дифференциальной функцией распределения (при использовании такого названия для $f_X(x)$ функцию $F_X(x)$ называют интегральной функцией распределения). График функции плотности $f_X(x)$ называют ***кривой распределения.***

Будем предполагать, что функция $f_X(x)$ интегрируема (по Риману) не только в конечных, но и в бесконечных пределах (что верно для непрерывных величин, представляющих практический интерес). Тогда, учитывая, что первообразной для $f_X(x)$ является функция $F_X(x)$ (см. (4.7)), в соответствии с формулой Ньютона—Лейбница получим

$$\int_{x_1}^{x_2} f_X(x)\,dx = F_X(x_2) - F_X(x_1).$$

Но так как $F_X(x_2) - F_X(x_1) = P(x_1 \leqslant X < x_2)$, то

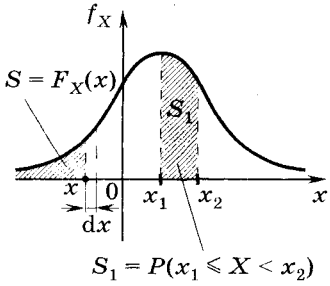
$$\int_{x_1}^{x_2} f_X(x)\,dx = P(x_1 \leqslant X < x_2). \tag{4.9}$$

Итак, найти вероятность попадания непрерывной случайной величины X в полуинтервал $[x_1, x_2)$ можно, используя либо функцию распределения $F_X(x)$, либо функцию плотности $f_X(x)$:

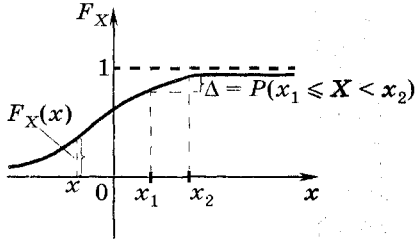
$$P(x_1 \leqslant X < x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)\,dx. \tag{4.10}$$

Напомним, что для непрерывной величины

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



а)



б)

Рис. 4.7

Из соотношения (4.10) вытекает, что $P(x_1 \leqslant X < x_2)$ имеет двоякую геометрическую интерпретацию. Эта вероятность равна:

— площади S_1 криволинейной трапеции, ограниченной кривой распределения, осью Ox и прямыми $x = x_1$ и $x = x_2$ (рис. 4.7, *а*);

— разности ординат $F_X(x_2)$ и $F_X(x_1)$ функции $F_X(x)$ в точках $x = x_1$ и $x = x_2$ (рис. 4.7, *б*).

В равенстве (4.9) приравняем x_2 числу x, заменив при этом переменную интегрирования x на t, а x_1 устремим к $-\infty$. В пределе, учитывая, что $P(X < x) = F_X(x)$, получим

$$\int_{-\infty}^{x} f_X(t)\,dt = P(-\infty < X < x) = P(X < x) = F_X(x),$$

или

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt. \qquad (4.11)$$

Формулы (4.7) и (4.11) позволяют заключить: *если известна одна из функций — функция распределения $F_X(x)$ или плотность вероятности $f_X(x)$, то другая определяется однозначно.*

Геометрическая интерпретация равенства (4.11) дана на рисунке 4.7: площадь S криволинейной трапеции, опирающейся на интервал $(-\infty, x)$, численно равна ординате $F_X(x)$ в точке x.

Рассмотрим свойства функции плотности $f_X(x)$.

1^0. $f_X(x) \geqslant 0$ *при любом действительном* x. (4.12)

≫ Функция $f_X(x)$ является производной функции распределения $F_X(x)$, которая, как было доказано в § 4.1, не убывает, а производная неубывающей функции неотрицательна. ≪

Из свойства 1^0 следует, что кривая распределения — график функции $f_X(x)$ — лежит не ниже оси абсцисс.

$$2^0.\ \int_{-\infty}^{+\infty} f_X(x)\,dx = 1. \qquad (4.13)$$

≫ В равенстве (4.9) устремим x_1 к $-\infty$, а x_2 — к $+\infty$. В пределе, учитывая, что событие $-\infty < X < +\infty$ достоверно, получим

$$\int_{-\infty}^{+\infty} f_X(x)\,dx = P(-\infty < X < +\infty) = 1,$$

что и требовалось доказать. ≪

Геометрическая интерпретация этого свойства такова: площадь, ограниченная кривой распределения и осью абсцисс, равна единице.

Замечание. Вероятность попадания случайной величины X на элементарный, малый участок dx равна (с точностью до бесконечно малых более высокого порядка) $f_X(x)\,dx$ — площади прямоугольника, опирающегося на dx (см. рис. 4.7, *а*); величину $f_X(x)\,dx$ называют ***элементом вероятности***. Если вспомнить, что определенный интеграл (интеграл Римана) — это предел интегральной суммы при $dx \to 0$, то допустима такая трактовка равенства (4.13): сумма элементов вероятности равна единице, т. е. свойство 2^0 является как бы «непрерывным» аналогом свойства 2^0 ряда распределения: сумма вероятностей значений дискретной случайной величины равна единице.

ЗАДАЧА 4.3. В условиях задачи 4.2 найдите функцию плотности и, используя ее, определите $P(0{,}2 \leqslant X \leqslant 0{,}3)$, $P(0{,}4 \leqslant X \leqslant 0{,}6)$, $P(X < 0{,}6)$. Дайте геометрическую интерпретацию этих вероятностей.

Решение. а) Зная функцию распределения $F_X(x)$, найдем функцию плотности вероятности

$$F_X(x) = \begin{cases} 0 & \text{при } x < 0, \\ 4x^2 & \text{при } 0 \leqslant x \leqslant 0{,}5, \\ 1 & \text{при } x > 0{,}5, \end{cases}$$

$$f_X(x) = F'_X(x) = \begin{cases} 0 & \text{при } x < 0, \\ 8x & \text{при } 0 \leqslant x \leqslant 0{,}5, \\ 0 & \text{при } x > 0{,}5. \end{cases}$$

График функции $f_X(x)$ изображен на рисунке 4.8.

Обратим внимание на то, что при x, не принадлежащем множеству значений величины X — в данном случае при $x \notin [0; 0{,}5]$, — функция $f_X(x) = 0$. Тогда, учитывая (4.9), получим, что для любого отрезка $[x_1, x_2] \notin [0; 0{,}5]$

$$P(x_1 \leqslant X < x_2) = \int_{x_1}^{x_2} 0\,dx = 0.$$

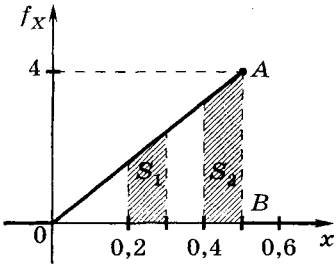


Рис. 4.8

Теперь найдем искомые вероятности. Отрезок $[0{,}2;\ 0{,}3] \in [0;\ 0{,}5]$, на котором $f_X(x) = 8x$. Поэтому, согласно (4.9),

$$P(0{,}2 \leqslant X \leqslant 0{,}3) = \int_{0{,}2}^{0{,}3} f_X(x)\,\mathrm{d}(x) = \int_{0{,}2}^{0{,}3} 8x\,dx = 4x^2\Big|_{0{,}2}^{0{,}3} =$$

$$= 4 \cdot 0{,}3^2 - 4 \cdot 0{,}2^2 = 0{,}2$$

— найденная вероятность равна площади S_1 (см. рис. 4.8).

Сравнив отрезок $[0{,}4;\ 0{,}6]$ с множеством значений величины X — отрезком $[0;\ 0{,}5]$, видим, что последнему принадлежит лишь часть отрезка $[0{,}4;\ 0{,}6]$, а именно $[0{,}4;\ 0{,}5]$, поэтому

$$P(0{,}4 \leqslant X \leqslant 0{,}6) = P(0{,}4 \leqslant X \leqslant 0{,}5) = \int_{0{,}4}^{0{,}5} 8x\,dx = 0{,}36$$

— площадь S_2.

Интервал $(-\infty;\ 0{,}6)$ полностью включает отрезок $[0;\ 0{,}5]$ — множество значений величины X, поэтому

$$P(X < 0{,}6) = P(-\infty < X < 0{,}6) = P(0 \leqslant X \leqslant 0{,}5) = \int_{0}^{0{,}5} 8x\,dx = 1$$

— площадь треугольника OAB. ◄

В заключение отметим еще раз: если множество значений случайной величины X — это все точки отрезка $[\alpha, \beta]$, т. е. величина X задана на отрезке $[\alpha, \beta]$, то при $x \notin [\alpha, \beta]$ функция $f_X(x) = 0$, поэтому

$$\int_{-\infty}^{\alpha} f_X(x)\,dx = \int_{\beta}^{+\infty} f_X(x)\,dx = 0;$$

равенство (4.13) в этом случае принимает вид

$$\int_{-\infty}^{+\infty} f_X(x)\,dx = \int_{\alpha}^{\beta} f_X(x)\,dx = 1. \qquad (4.14)$$

§ 4.3. Числовые характеристики случайной величины

В § 4.2 было показано, что полной характеристикой случайной величины является ее закон распределения (функция распределения; или ряд распределения — для дискретной величины и плотность вероятности — для непрерывной).

Действительно, если известен закон распределения, то известны все значения случайной величины и можно найти различные вероятности. Однако в ряде случаев о случайной величине требуется знать гораздо меньше: достаточно указать некоторые постоянные числа — числовые характеристики случайной величины, получаемые по определенным правилам из закона распределения (предполагаем, что дискретная величина задана рядом распределения, а непрерывная — плотностью вероятности). Числовые характеристики дают общее представление о случайной величине.

Среди числовых характеристик различают характеристики положения (математическое ожидание, мода, медиана, квантили различных порядков) и характеристики рассеивания (дисперсия, среднее квадратическое отклонение, коэффициент вариации), а также моменты различных порядков и коэффициенты асимметрии и эксцесса.

4.3.1. Характеристики положения. ***Математическое ожидание***, или среднее значение случайной величины, обозначают прописной латинской буквой M, поставленной перед обозначением случайной величины: MX — математическое ожидание случайной величины X; MX — это постоянная величина, способ нахождения которой определяется видом величины X (см. табл. 4.6).

Таблица 4.6

<table>
<tr><td colspan="5">X — дискретная случайная величина с рядом распределения</td><td rowspan="2">X — непрерывная случайная величина с плотностью $f_X(x)$, заданная на отрезке $[\alpha, \beta]$</td></tr>
<tr><td>x</td><td>x_1</td><td>x_2</td><td>...</td><td>x_n</td></tr>
<tr><td>$P(X = x)$</td><td>p_1</td><td>p_2</td><td>...</td><td>p_n</td><td></td></tr>
<tr><td colspan="5">$$MX = \sum_{i=1}^{n} x_i p_i. \quad (4.15)$$
Замечание. Здесь и далее предполагается, что число значений величины X конечно.
При бесконечном числе значений величины X
$$MX = \sum_{i=1}^{\infty} x_i p_i$$</td><td>$$MX = \int_{\alpha}^{\beta} x f_X(x)\,dx. \quad (4.16)$$
Замечание. Вспомнив, что при малом dx элемент вероятности
$f_X(x)\,dx \approx P(x \leqslant X < x + dx)$,
нетрудно понять, что (4.16) является непрерывным аналогом (4.15)</td></tr>
</table>

Математическое ожидание MX измеряется в тех же единицах, что и величина X (вероятности p_i и элемент веро-

ятности $f_X(x)\,dx$ — безразмерные величины). Если X измеряется, например, в ден. ед., то и MX измеряется в ден. ед.

Выше математическое ожидание характеризовалось как среднее значение случайной величины. Что это означает, поясним на примере дискретной величины.

▶ **ПРИМЕР 4.2.** Случайная величина X задана рядом

x	-2	2	5
$P(X=x)$	0,6	0,1	0,3

В соответствии с (4.15),

$$MX = -2 \cdot 0{,}6 + 2 \cdot 0{,}1 + 5 \cdot 0{,}3 = 0{,}5.$$

На рисунке 4.9 дано графическое изображение ряда и MX: «вес (размер)» точки — значения величины X равен вероятности этого значения, крестиком отмечено $MX = 0{,}5$.

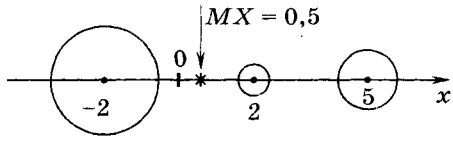


Рис. 4.9

Найдем координату x_0 центра тяжести точек $x_1 = -2$, $x_2 = 2$, $x_3 = 5$ с «весами» $p_1 = 0{,}6$; $p_2 = 0{,}1$; $p_3 = 0{,}3$. Имеем

$$x_0 = (x_1p_1 + x_2p_2 + x_3p_3)/(p_1 + p_2 + p_3) \underset{\sum p_i = 1}{=}$$

$$\underset{\sum p_i = 1}{=} x_1p_1 + x_2p_2 + x_3p_3 = MX = 0{,}5.$$

Из этой цепочки равенств видно, что MX — это «центр тяжести» значений величины X с весами, равными их вероятностям, или, иначе, MX — это среднее арифметическое значений величины X, взвешенных их вероятностями (если значения x_i не взвешивать вероятностями p_i, то среднее значений равно $(-2 + 2 + 5)/3 = 5/3 \neq MX$); короче, MX — среднее значение случайной величины. ◀

Рассмотрим основные свойства математического ожидания, имеющие место как для дискретных, так и для непрерывных величин (докажем свойства для случая дискретных величин).

1^0. *Математическое ожидание постоянной c равно этой постоянной*

$$Mc = c. \qquad (4.17)$$

≫ Рассматривая постоянную c как частный вид случайной величины, принимающей единственное значение c с вероятностью единица, по формуле (4.15) для математического ожидания получим $Mc = c \cdot P(c = c) = c \cdot 1 = c$. ≪

2^0. *Константу c выносят за знак математического ожидания*

$$M(cX) = cMX. \tag{4.18}$$

≫ Если случайная величина X имеет ряд распределения

x	x_1	x_2	...	x_n
$P(X = x)$	p_1	p_2	...	p_n

то ряд распределения случайной величины, равной произведению cX, таков:

cx	cx_1	cx_2	...	cx_n
$P(cX = cx)$	p_1	p_2	...	p_n

Действительно, событие $cx = cx_i$, $i = 1, 2, \ldots, n$, состоящее в том, что величина cX примет значение cx_i, равносильно событию $x = x_i$, поэтому $P(cX = cx_i) = P(X = x_i) = p_i$. Отсюда получим

$$M(cX) = \sum_{i=1}^{n} (cx_i)p_i = c \sum_{i=1}^{n} x_i p_i = cMX. \quad \ll$$

3^0. *Математическое ожидание суммы случайных величин X и Y равно сумме их математических ожиданий*

$$M(X + Y) = MX + MY. \tag{4.19}$$

≫ Докажем свойство, ограничившись случаем, когда величина X задана рядом распределения

x	x_1	x_2	x_3
$P(X = x)$	p_1	p_2	p_3

а величина Y — рядом

y	y_1	y_2
$P(Y = y)$	q_1	q_2

При таких X и Y множество значений случайной величины $(X + Y)$: $\{x_1 + y_1, x_1 + y_2, x_2 + y_1, x_2 + y_2, x_3 + y_1, x_3 + y_2\}$. Значения величины $(X + Y)$ и их вероятности приведены в таблице 4.7.

Таблица 4.7

$x+y$	x_1+y_1	x_1+y_2	x_2+y_1	x_2+y_2	x_3+y_1	x_3+y_2
$P(X+Y=x+y)$	p_{11}	p_{12}	p_{21}	p_{22}	p_{31}	p_{32}

Убедимся в том, что сумма вероятностей в таблице 4.7 равна единице. Введем следующие события:

A_i — величина X примет значение x_i, $i = 1, 2, 3$; B_j — величина Y примет значение y_j, $j = 1, 2$; $C_{ij} = A_i \quad B_j$ — величина X примет значение x_i, а Y — значение y_j; как следствие, величина $X + Y$ примет значение $x_i + y_j$.

В этих обозначениях:

$$p_{11} + p_{12} = P(C_{11}) + P(C_{12}) = P(A_1 \quad B_1) + P(A_1 \quad B_2) =$$
$$= P(B_1)P(A_1|B_1) + P(B_2)P(A_1|B_2) \underset{(*)}{=} P(A_1) = p_1.$$

При переходе (*) была использована формула полной вероятности (2.23), поскольку здесь условия (2.22) выполняются, а именно:

— событие A_1 (появление x_1) наступает при наступлении либо события B_1 (появление y_1), либо события B_2 (появление y_2);

— события B_1 и B_2 несовместны;

— события B_1 и B_2 образуют полную группу (так как $P(B_1) + P(B_2) = P(Y = y_1) + P(Y = y_2) = q_1 + q_2 = 1$).

Итак, $p_{11} + p_{12} = p_1$. Аналогично можно убедиться в том, что $p_{21} + p_{22} = p_2$, а $p_{31} + p_{32} = p_3$; а также в том, что $p_{11} + p_{21} + p_{31} = q_1$ и $p_{12} + p_{22} + p_{32} = q_2$. Отсюда в таблице 4.7 сумма вероятностей $p_{11} + p_{12} + \ldots + p_{32} = p_1 + p_2 + p_3 = q_1 + q_2 = 1$.

Найдем математическое ожидание случайной величины $(X + Y)$. Имеем

$$M(X+Y) = (x_1+y_1)p_{11} + (x_1+y_2)p_{12} + \ldots + (x_3+y_2)p_{32} =$$
$$= x_1(p_{11}+p_{12}) + x_2(p_{21}+p_{22}) + x_3(p_{31}+p_{32}) + y_1(p_{11}+p_{21}+p_{31}) +$$
$$+ y_2(p_{12}+p_{22}+p_{32}) = (x_1p_1 + x_2p_2 + x_3p_3) + (y_1q_1 + y_2q_2) = MX + MY.$$

Итак, $M(X + Y) = MX + MY$, что и требовалось доказать. ≪

4^0. *Математическое ожидание произведения независимых случайных величин X и Y равно произведению их математических ожиданий*:

$$M(XY) = MX \cdot MY, \qquad (4.20)$$

если x и y независимы.

Прежде чем доказать это свойство, введем понятие независимых случайных величин.

Определение. *Случайные величины X и Y независимы, если события $X < x$ и $Y < y$ — независимы, т. е.*

$$P((X < x) \quad (Y < y)) = P(X < x)P(Y < y), \qquad (4.21)$$

где x и y — любые действительные числа; в противном случае — величины X и Y зависимы.

Обобщение определения на случай более двух величин очевидно.

Для дискретных величин X и Y данное определение тождественно следующему: *дискретные случайные величины X и Y независимы, если событие X примет значение x_i, а Y — значение y_j, независимы, т. е.*

$$P((X = x_i) \cap (Y = y_j)) = P(X = x_i)P(Y = y_j), \qquad (4.22)$$

где x_i и y_j — любые из возможных значений соответственно величин X и Y; иначе — X и Y зависимы.

≫ Докажем свойство 4^0, ограничившись случаем, когда величина X задана рядом

x	x_1	x_2	x_3
$P(X = x)$	p_1	p_2	p_3

а величина Y — рядом

y	y_1	y_2
$P(Y = y)$	q_1	q_2

При таких X и Y множество значений величины XY: $\{x_1y_1, x_1y_2, x_2y_1, x_2y_2, x_3y_1, x_3y_2\}$; в силу независимости X и Y

$$P((X = x_i) \cap (Y = y_j)) = P(X = x_i)P(Y = y_j) = p_iq_j,$$

$$i = 1, 2, 3; \qquad j = 1, 2.$$

Поэтому ряд распределения величины XY таков:

xy	x_1y_1	x_1y_2	x_2y_1	x_2y_2	x_3y_1	x_3y_2
$P(XY = xy)$	p_1q_1	p_1q_2	p_2q_1	p_2q_2	p_3q_1	p_3q_2

Сумма вероятностей ряда действительно равна единице

$$p_1q_1 + p_1q_2 + \ldots + p_3q_2 = p_1(q_1 + q_2) + p_2(q_1 + q_2) +$$

$$+ p_3(q_1 + q_2) = p_1 \cdot 1 + p_2 \cdot 1 + p_3 \cdot 1 = 1.$$

Тогда

$$M(XY) = x_1y_1p_1q_1 + x_1y_2p_1q_2 + \ldots + x_3y_2p_3q_2 =$$

$$= x_1p_1(y_1q_1 + y_2q_2) + x_2p_2(y_1q_1 + y_2q_2) + x_3p_3(y_1q_1 + y_2q_2) =$$

$$= (y_1q_1 + y_2q_2)(x_1p_1 + x_2p_2 + x_3p_3) = MY \cdot MX.$$

Итак, $M(XY) = MY \cdot MX$, что и требовалось доказать. ≪

Приведем некоторые *следствия свойств математического ожидания.*

1) $M(X - Y) = MX - MY.$ (4.23)

≫ Имеем

$$M(X - Y) = M(X + (-1 \cdot Y)) \underset{(4.19)}{=} MX + M(-1 \cdot Y) \underset{(4.18)}{=}$$

$$\underset{(4.18)}{=} MX - 1MY = MX - MY. \ll$$

2) $M(a + bX) = a + bMX,$ (4.24)

где a *и* b — *неслучайные величины* (действительные числа).

≫ Имеем

$$M(a + bX) \underset{(4.19)}{=} Ma + M(bX) \underset{(4.17,4.18)}{=} a + bMX. \ll$$

3) *Математическое ожидание суммы случайных величин* $X_1, X_2, \ldots, X_k$, $k > 2$, *равно сумме их математических ожиданий*:

$$M \sum_{i=1}^{k} X_i = \sum_{i=1}^{k} MX_i. \quad (4.25)$$

≫ Доказательство основано на использовании свойства (4.19), утверждающего, что математическое ожидание суммы двух случайных величин равно сумме их математических ожиданий:

$$M \sum_{i=1}^{k} X_i = M(X_1 + \sum_{i=2}^{k} X_i) \underset{(4.19)}{=} MX_1 + M \sum_{i=2}^{k} X_i = MX_1 + M(X_2 +$$

$$+ \sum_{i=3}^{k} X_i) \underset{(4.19)}{=} MX_1 + MX_2 + M \sum_{i=3}^{k} X_i = \ldots = \sum_{i=1}^{k} MX_i. \ll$$

4) Математическое ожидание

$$M(a + \sum_{i=1}^{k} b_i X_i) = a + \sum_{i=1}^{k} b_i MX_i, \quad (4.26)$$

где $X_1, X_2, \ldots, X_k$ — *произвольные случайные величины*; $a, b_1, b_2, \ldots, b_k$ — *неслучайные величины* (действительные числа).

≫ Имеем

$$M(a + \sum_{i=1}^{k} b_i X_i) \underset{(4.19)}{=} Ma + M \sum_{i=1}^{k} b_i X_i \underset{(4.17)}{=}$$

$$\underset{(4.17)}{=} a + M \sum_{i=1}^{k} b_i X_i \underset{(4.25)}{=} a + \sum_{i=1}^{k} M(b_i X_i) \underset{(4.18)}{=} a + \sum_{i=1}^{k} b_i MX_i. \ll$$

5) *Математическое ожидание произведения независимых в совокупности случайных величин* $X_1, X_2, \ldots, X_k$,

$k > 2$, равно произведению их математических ожиданий:

$$M(X_1X_2...X_k) = MX_1 \cdot MX_2 \cdot ... \cdot MX_k,$$

или

$$M\prod_{i=1}^{k} X_i = \prod_{i=1}^{k} MX_i. \qquad (4.27)$$

Предварительно введем понятие случайных величин, независимых в совокупности.

Определение. *Случайные величины X_1, X_2, ..., X_k независимы в совокупности, или просто независимы, если наряду с их попарной независимостью независимы любая из этих величин и произведение любого числа из оставшихся величин.*

≫ Доказательство следствия 5 основано на использовании свойства (4.20), утверждающего, что математическое ожидание произведения двух независимых случайных величин равно произведению их математических ожиданий:

$$M(X_1X_2...X_k) = M(X_1\underbrace{(X_2...X_k)}_{Y_1}) \underset{(*)}{=} MX_1 \cdot MY_1 = MX_1 \cdot M(X_2...X_k) =$$

$$= MX_1 \cdot M(X_2\underbrace{(X_3...X_k)}_{Y_2}) \underset{(**)}{=} MX_1 \cdot MX_2 \cdot MY_2 =$$

$$= MX_1 \cdot MX_2 \cdot M(X_3...X_k) = ... = MX_1 \cdot MX_2 ... MX_k.$$

При переходах (*) и (**) учтена независимость величин X_1 и Y_1 и величин X_2 и Y_2, вытекающая из определения независимых в совокупности величин X_1, X_2, ..., X_k; и использовано свойство (4.20). ≪

▶ **ПРИМЕР 4.3.** Случайная величина x задана рядом распределений

x	-2	2	5
$P(X = x)$	0,6	0,1	0,3

$$MX = -2 \cdot 0{,}6 + 2 \cdot 0{,}1 + 5 \cdot 0{,}3 = 0{,}5.$$

Составим ряд распределения величины $Y = -3X + 6$:

y	$-3 \cdot (-2) + 6$	$-3 \cdot 2 + 6$	$-3 \cdot 5 + 6$
$P(Y = y)$	0,6	0,1	0,3

y	12	0	-9
$P(Y = y)$	0,6	0,1	0,3

y	-9	0	12
$P(Y=y)$	0,3	0,1	0,6

(Напомним, что в ряду распределения значения случайной величины располагаются в возрастающем порядке.) Найдем MY. Используя ряд распределения величины Y, имеем:

$$MY = -9 \cdot 0{,}3 + 0 \cdot 0{,}1 + 12 \cdot 0{,}6 = 4{,}5;$$

используя равенство (4.24), получаем

$$MY = M(-3X + 6) = -3MX + 6 = -3 \cdot 0{,}5 + 6 = 4{,}5.$$

Получены одинаковые результаты. ◀

▶▶ **ЗАДАЧА 4.4.** Независимые случайные величины X и Y имеют соответственно ряды распределения

x	-1	0	1
$P(X=x)$	0,1	0,1	0,8

и

y	1	2
$P(Y=y)$	0,7	0,3

Составьте: а) ряд распределения случайной величины $Z = X - Y$, и найдите ее математическое ожидание; б) ряд распределения случайной величины $V = \min(X, Y)$, и найдите $P(V < MV)$.

Р е ш е н и е. а) Если X принимает значение -1, а Y — значение 1, то $Z = X - Y$ принимает значение, равное $(-1 - 1)$, и, в силу независимости величин X и Y,

$$P(X - Y = -1 - 1) = P(X = -1) \cdot P(Y = 1) = 0{,}1 \cdot 0{,}7 = 0{,}07.$$

Рассуждая аналогично для других комбинаций значений случайных величин X и Y, получим ряд распределения величин Z

$z = x - y$	$-1-1$	$-1-2$	$0-1$	$0-2$	$1-1$	$1-2$
$P(Z = x-y)$	$0{,}1 \cdot 0{,}7$	$0{,}1 \cdot 0{,}3$	$0{,}1 \cdot 0{,}7$	$0{,}1 \cdot 0{,}3$	$0{,}8 \cdot 0{,}7$	$0{,}8 \cdot 0{,}3$

z	-2	-3	-1	-2	0	-1
$P(Z=z)$	0,07	0,03	0,07	0,03	0,56	0,24

z	-3	-2	-1	0	
$P(Z=z)$	0,03	0,1	0,31	0,56	$\Sigma=1$

В ряду $P(Z=-2)=0{,}1$, так как событие, состоящее в том, что Z примет значение -2, может произойти лишь, когда произойдет одно из двух несовместных событий: (X примет значение -1, а Y — значение 1) или (X примет значение 0, а Y — значение 2). Поэтому

$$P(Z=-2)=P((X-Y)=-2)=P((X=-1)\cap(Y=1))+P((X=$$
$$=0)\cap(Y=2))=0{,}1\cdot0{,}7+0{,}1\cdot0{,}3=0{,}07+0{,}03=0{,}1.$$

Аналогично находим

$$P(Z=-1)=P(X-Y=-1)=$$
$$=P((X=0)\cap(Y=1))+P((X=1)\cap$$
$$\cap(Y=2))=0{,}1\cdot0{,}7+0{,}8\cdot0{,}3=0{,}07+0{,}24=0{,}31.$$

Используя ряд распределения величины Z, имеем

$$MZ=-3\cdot0{,}03-2\cdot0{,}1-1\cdot0{,}31+0\cdot0{,}56=-0{,}6.$$

Такой же результат получим, используя равенство (4.23), имеющее место как для независимых, так и для зависимых случайных величин:

$$MZ=M(X-Y)=MX-MY=(-1\cdot0{,}1+1\cdot0{,}8)-$$
$$-(1\cdot0{,}7+2\cdot0{,}3)=-0{,}6.$$

б) Составим ряд распределения величины $V=\min(X, Y)$:

$v=\min(x, y)$	$\min(-1;1)$	$\min(-1;2)$	$\min(0;1)$	$\min(0;2)$	$\min(1;1)$	$\min(1;2)$
$P(V=v)$	$0{,}1\cdot0{,}7$	$0{,}1\cdot0{,}3$	$0{,}1\cdot0{,}7$	$0{,}1\cdot0{,}3$	$0{,}8\cdot0{,}7$	$0{,}8\cdot0{,}3$

v	-1	-1	0	0	1	1
$P(V=v)$	0,07	0,03	0,07	0,03	0,56	0,24

v	-1	0	1	
$P(V=v)$	0,1	0,1	0,8	$\Sigma=1$

$$MV=0{,}7 \text{ и } P(V<MV)=P(V<0{,}7)=$$
$$=P(V=-1)+P(V=0)=0{,}1+0{,}1=0{,}2.$$

ЗАДАЧА 4.5. Плотность вероятности случайной величины X

$$f_X(x) = \begin{cases} ax \text{ при } x \in [2, 4], \\ 0 \quad \text{при } x \notin [2, 4], \end{cases} \tag{4.28}$$

где a — постоянная величина. Найдите MX и число x_1, при котором $P(X < x_1) = 0{,}5$.

Р е ш е н и е. Определим числовое значение постоянной a. Так как величина X задана на отрезке [2; 4], то, учитывая (4.14), имеем

$$\int_2^4 f_X(x)\,dx = \int_2^4 ax\,dx = \frac{ax^2}{2}\Big|_2^4 = \frac{a \cdot 4^2}{2} - \frac{a \cdot 2^2}{2} = \frac{12}{2}a = 1.$$

Отсюда $a = 1/6$ и для $x \in [2, 4]$ $f_X(x) = x/6$. Согласно (4.16),

$$MX = \int_2^4 x \cdot \frac{x}{6}\,dx = \frac{x^3}{18}\Big|_2^4 = \frac{30}{9} = 3\frac{1}{9} \approx 3{,}1(1).$$

Далее, учитывая, что X — непрерывна, и принимая во внимание (4.9), получим

$$P(X < x_1) = P(-\infty < X < x_1) = \int_{-\infty}^{x_1} f_X(x)\,dx =$$

$$= \int_2^{x_1} \frac{x}{6}\,dx = \frac{x^2}{12}\Big|_2^{x1} = \frac{x_1^2}{12} - \frac{1}{3} = 0{,}5.$$

Отсюда x_1 — корень уравнения $\frac{x^2}{12} - \frac{5}{6} = 0$, но поскольку величина X задана на [2; 4], то $x_1 = +\sqrt{10} \approx 3{,}2$.

Графическое изображение функции $f_X(x)$ и числа x_1 дано на рисунке 4.10. ◀

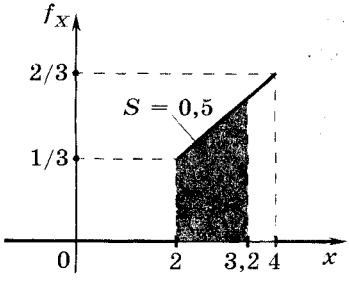


Рис. 4.10

В качестве характеристик «положения» случайной величины наряду с математическим ожиданием (наиболее употребительной характеристикой) используются: мода, медиана, квантили и процентные точки различных порядков.

Мода случайной величины X («наиболее вероятное значение» величины) обозначается символом x_{mod}.

Медиана («серединное значение» случайной величины X) обозначается символом x_{med}. Покажем на примерах, как найти моду и медиану дискретной и непрерывной случайной величины.

▶ **ПРИМЕР 4.4.** Найдем моду и медиану случайной величины x со следующим рядом распределения:

x	$x_1 = -1$	$x_2 = 2$	$x_3 = 4$	$x_4 = 7$
$P(X = x)$	$p_1 = 0{,}1$	$p_2 = 0{,}35$	$p_3 = 0{,}35$	$p_4 = 0{,}2$

Мода x_{mod} — это такое значение x_k величины X, которому соответствует наибольшая вероятность, т. е.

$$x_{\text{mod}} = x_k, \text{ если } p_k = \max_{1 \leqslant i \leqslant n} \{p_i\}. \tag{4.29}$$

В соответствии с (4.29), X имеет две моды: $x_{\text{mod}}^{(1)} = x_2 = 2$ и $x_{\text{mod}}^{(2)} = x_3 = 4$, так как вероятность каждого из этих значений равна $0{,}35 = \max\{0{,}1; 0{,}35; 0{,}35; 0{,}2\}$.

Медиана x_{med} — это такое действительное число, при котором

$$P(X \leqslant x_{\text{med}}) = 0{,}5. \tag{4.30}$$

Чтобы найти медиану, для каждого значения величины X подсчитаем накопленную вероятность:

x	$x_1 = -1$	$x_2 = 2$	$x_3 = 4$	$x_4 = 7$
$p^{\text{нак}}$	$p_1^{\text{нак}} = 0{,}1$	$p_2^{\text{нак}} = 0{,}1 + 0{,}35 = 0{,}45$	$p_3^{\text{нак}} = 0{,}45 + 0{,}35 = 0{,}8$	$p_4^{\text{нак}} = 0{,}8 + 0{,}2 = 1$

и найдем медианный интервал $[x_l, x_{l+1})$ — интервал, в котором лежит x_{med};

медиана

$$x_{\text{med}} = x_l + \frac{x_{l+1} - x_l}{p_{l+1}}(0{,}5 - p_l^{\text{нак}}). \tag{4.31}$$

Так как $p_2^{\text{нак}} < 0{,}5$, а $p_3^{\text{нак}} > 0{,}5$, то интервал $[x_2, x_3)$, т. е. интервал $[2; 4)$ является медианным. Поэтому в формуле (4.31) $l = 2$, $x_l = 2$, $x_{l+1} = 4$, $p_{l+1} = P(X = x_{l+1}) = P(X = 4) = 0{,}35$, а $p_l^{\text{нак}} = 0{,}45$. Окончательно получаем

$$x_{\text{med}} = 2 + \frac{4 - 2}{0{,}35} \cdot (0{,}5 - 0{,}45) = 2{,}28.$$

Как и следовало ожидать, медиана принадлежит медианному интервалу: $x_{\text{med}} = 2{,}28 \in [2; 4)$ и располагается бли-

же к его началу: для концов интервала $p^{\text{нак}}$ равна соответственно 0,45 и 0,8, а в точке x_{med} по определению (4.30) накопленная вероятность, или $P(X \leqslant x_{\text{med}})$, равна 0,5.

ПРИМЕР 4.5. Найдем моду и медиану случайной величины X с плотностью вероятности

$$f_X(x) = \begin{cases} \cos x \text{ при } x \in [0, \pi/2], \\ 0 \quad \text{при } x \notin [0, \pi/2]. \end{cases} \tag{4.32}$$

Предварительно обратим внимание на то, что случайная величина X задана на отрезке $[0, \pi/2]$ и что $f_X(x)$ — действительно плотность вероятности, поскольку выполняются свойства (4.12) и (4.13): $f_X(x) \geqslant 0$ при любом действительном x и

$$\int_{-\infty}^{+\infty} f_X(x)\,dx = \int_0^{\pi/2} \cos x\,dx = 1.$$

Мода x_{mod} — это действительное число x, при котором функция плотности, заданная на отрезке $[\alpha, \beta]$, достигает максимального значения.

Для случайной величины X $x_{\text{mod}} = 0$, так как $x = 0$ — точка максимума функции (4.32) на отрезке $[0, \pi/2]$.

Медиана x_{med} — это действительное число, при котором

$$P(X \leqslant x_{\text{med}}) = 0{,}5, \tag{4.33}$$

или (учитывая, что для непрерывной величины X функция распределения $F_X(x) = P(X < x) = P(X \leqslant x)$), при котором

$$F_X(x_{\text{med}}) = 0{,}5, \tag{4.34}$$

или при котором

$$\int_{\alpha}^{x_{\text{med}}} f_X(x)\,dx = 0{,}5. \tag{4.35}$$

Используя (4.35) и учитывая при этом, что в рассматриваемом случае X задана на отрезке $[0, \pi/2]$, найдем медиану:

$$\int_{\alpha}^{x_{\text{med}}} f_X(x)\,dx = \int_0^{x_{\text{med}}} \cos x\,dx = \sin x\Big|_0^{x_{\text{med}}} =$$

$$= \sin x_{\text{med}} - \sin 0 = \sin x_{\text{med}} = 0{,}5,$$

отсюда $x_{\text{med}} = \arcsin 0{,}5 = \pi/6$. Согласно (4.33), $P(X \leqslant \pi/6) = 0{,}5$, следовательно, $P(X > \pi/6) = 1 - P(X \leqslant \pi/6) = 0{,}5$, т. е. прямая $x = \pi/6$ делит площадь под кривой распределения, равную единице, на две равные части S_1 и S_2: $S_1 = S_2 =$

$= 0{,}5$. Кривая распределения, x_{mod} и x_{med} изображены на рисунке 4.11. ◂

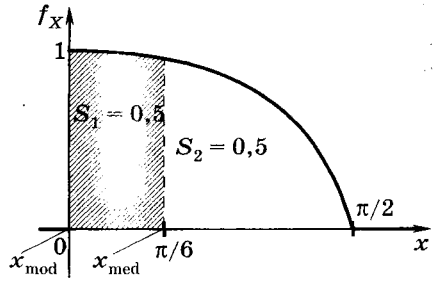


Рис. 4.11

Наконец, определим для непрерывной величины X понятия квантиля и процентной точки.

Квантилем порядка q, или $q \cdot 100\%$ квантилем, $0 < q < 1$, непрерывной величины X называется такое действительное число x_q^*, для которого

$$P(X < x_q^*) = q. \tag{4.36}$$

Процентной точкой порядка p, или $p \cdot 100\%$ точкой, $0 < p < 1$, непрерывной величины X называется такое действительное число x_p, для которого

$$P(X > x_p) = p. \tag{4.37}$$

Убедимся в том, что для непрерывной величины X квантиль порядка $1 - p$ совпадает с процентной точкой порядка p, т. е.

$$x_{1-p}^* = x_p. \tag{4.38}$$

≫ Положив в (4.36) $q = 1 - p$, получим $P(X < x_{1-p}^*) = 1 - p$, или $P(X \geqslant x_{1-p}^*) = p$; поскольку X — непрерывна, $P(X > x_{1-p}^*) = p$. Сравнив последнее равенство с (4.37), заключаем, что $x_{1-p}^* = x_p$. ≪

Убедимся в том, что медиана непрерывной величины X является квантилем порядка 0,5 (или 50% квантилем). Действительно, согласно определению (4.33) медианы x_{med} непрерывной величины X, $P(X < x_{\text{med}}) = 0{,}5$. Сопоставив это равенство с (4.36), получим: $x_{\text{med}} = x_{0,5}^*$, т. е. медиана — это квантиль порядка 0,5. Будучи квантилем порядка $0{,}5 = 1 - 0{,}5$, медиана, если учесть (4.38), также является процентной точкой порядка 0,5: $x_{1-0,5}^* = x_{0,5}$.

4.3.2. Характеристики рассеивания. ***Дисперсия*** — это неотрицательное число, характеризующее рассеивание значений случайной величины X относительно ее математического ожидания MX; дисперсию обозначают прописной латинской буквой D, поставленной перед обозначением случайной величины: DX — дисперсия случайной величины X; DX — это постоянная величина.

Определение. *Дисперсией случайной величины X называется математическое ожидание квадрата отклонения величины от ее математического ожидания,*

$$DX = M(X - MX)^2. \tag{4.39}$$

Дисперсия величины X, будучи равной $M(X - MX)^2$, является центром группирования значений случайной величины $(X - MX)^2$ — это и имеют в виду, когда говорят, что дисперсия характеризует рассеивание значений величины X относительно MX.

Замечания. 1. Дисперсия не определяется как математическое ожидание отклонения величины X от MX. Учитывая, что MX — постоянная величина, получим

$$M(X - MX) \underset{(4.23)}{=} MX - M(MX) \underset{(4.17)}{=} MX - MX = 0.$$

2. Можно доказать, что если в формуле (4.39) MX заменить на любую постоянную величину $c \neq MX$, то

$$M(X - c)^2 > M(X - MX)^2, \qquad c \neq MX; \tag{4.40}$$

говорят, что DX, определяемая формулой (4.39), обладает свойством минимальности.

Из (4.39) следует, что DX измеряется в квадратных единицах. Например, если X измеряется в ден. ед., то и MX измеряется в ден. ед., но тогда единицей измерения величины $(X - MX)^2$, следовательно, и $M(X - MX)^2$ будет (ден. ед.)^2. Поэтому, наряду с DX, в качестве характеристики рассеивания значений случайной величины X относительно MX используют постоянную величину $\sigma_X = + \sqrt{DX}$ (σ — греческая буква «сигма»), которая измеряется в тех же единицах, что и величина X.

Определение. *Средним квадратическим отклонением (стандартным отклонением) случайной величины X называют арифметическое значение корня квадратного из дисперсии:*

$$\sigma_X = + \sqrt{DX}\,.$$

Теорема. *Дисперсия случайной величины X равна разности между математическим ожиданием квадрата случайной величины и квадратом ее математического ожидания, т. е.*

$$DX = M(X^2) - (MX)^2. \tag{4.41}$$

≫ Используя свойства математического ожидания, проведем тождественные преобразования:

$$DX = M(X - MX)^2 = M(X^2 - 2XMX + (MX)^2) \underset{(4.25)}{=}$$

$$\underset{(4.25)}{=} M(X^2) - M(2XMX) + M(MX)^2 \underset{\substack{(4.18)\\(4.17)}}{=}$$

$$\underset{\substack{(4.18)\\(4.17)}}{=} M(X^2) - 2MX \cdot MX + (MX^2) = M(X)^2 - (MX)^2,$$

что и требовалось доказать. ≪

Конкретный вид формул (4.39) и (4.41) для дискретной и непрерывной случайных величин приведен в таблице 4.8.

Таблица 4.8

<table>
<tr><td colspan="5">X — дискретная величина с рядом распределения</td><td rowspan="3">X — непрерывная величина с плотностью $f_X(x)$, заданная на отрезке $[\alpha, \beta]$</td></tr>
<tr><td>x</td><td>x_1</td><td>x_2</td><td>...</td><td>x_n</td></tr>
<tr><td>$P(X = x)$</td><td>p_1</td><td>p_2</td><td>...</td><td>p_n</td></tr>
<tr><td colspan="6">Вычисление дисперсии, основанное на использовании формулы (4.39): $DX = M(X - MX)^2$</td></tr>
<tr><td colspan="5">Ряд распределения случайной величины $(X - MX)^2$ имеет вид
<table><tr><td>$(x - MX)^2$</td><td>$(x_1 - MX)^2$</td><td>...</td><td>$(x_n - MX)^2$</td></tr><tr><td>P</td><td>p_1</td><td>...</td><td>p_n</td></tr></table>поэтому
$M(X - MX)^2 = \sum_{i=1}^{n} (x_i - MX)^2 p_i,$
следовательно,
$DX = \sum_{i=1}^{n} (x_i - MX)^2 p_i.$ (4.42)</td><td>$DX = \int_{\alpha}^{\beta} (x - MX)^2 f_X(x)\,dx$ (4.43)
(эта формула является «непрерывным аналогом» формулы (4.42)).</td></tr>
<tr><td colspan="6">Вычисление дисперсии, основанное на использовании формулы (4.41): $DX = M(X^2) - (MX)^2$</td></tr>
<tr><td colspan="5">Ряд распределения случайной величины X^2 имеет вид
<table><tr><td>x^2</td><td>x_1^2</td><td>x_2^2</td><td>...</td><td>x_n^2</td></tr><tr><td>$P(X^2 = x^2)$</td><td>p_1^2</td><td>p_2^2</td><td>...</td><td>p_n^2</td></tr></table></td><td></td></tr>
</table>

поэтому $M(X^2) = \sum_{i=1}^{n} x_i^2 p_i$, тогда $DX = \sum_{i=1}^{n} x_i^2 p_i - (MX)^2$. (4.44) В (4.42) и (4.44) $MX = \sum_{i=1}^{n} x_i p_i$	$DX = \int_{\alpha}^{\beta} x^2 f_X(x)\,dx - (MX)^2$ (4.45) Формула (4.45) является «непрерывным аналогом» формулы (4.44). В формулах (4.43) и (4.45) $MX = \int_{\alpha}^{\beta} x f_X(x)\,dx$

▶ **ПРИМЕР 4.6.** Вычислим DX и σ_X для случайной величины X с рядом распределения

x	-2	2	5
$P(X = x)$	0,6	0,1	0,3

В примере 4.3 было найдено $MX = 0{,}5$. Согласно (4.42), дисперсия $DX = \sum_{i=1}^{3} (x_i - MX)^2 p_i = (-2 - 0{,}5)^2 \cdot 0{,}6 + (2 - -0{,}5)^2 \cdot 0{,}1 + (5 - 0{,}5)^2 \cdot 0{,}3 = 10{,}05$.
Согласно (4.44),

$$DX = \sum_{i=1}^{3} x_i^2 p_i - (MX)^2 =$$

$$= (-2)^2 \cdot 0{,}6 + 2^2 \cdot 0{,}1 + 5^2 \cdot 0{,}3 - 0{,}5^2 = 10{,}05,$$

результаты формул (4.42) и (4.44), как и следовало ожидать, совпадают. Среднее квадратическое отклонение $\sigma_X = \sqrt{DX} = \sqrt{10{,}05} = 3{,}17$. ◀

Вычисление дисперсии по формуле (4.44) менее трудоемко, чем по формуле (4.42). Поэтому при нахождении дисперсии дискретной величины чаще используют формулу (4.44). Аналогично при нахождении дисперсии непрерывной величины обычно применяют формулу (4.45), а не формулу (4.43).

▶ **ПРИМЕР 4.7.** Функция плотности

$$f_X(x) = \begin{cases} 0 & \text{при } x < 0, \\ \lambda e^{-\lambda x} & \text{при } x \geqslant 0, \text{ где } \lambda > 0. \end{cases}$$

Найдем дисперсию, используя формулу (4.45). Предварительно заметим: величина X задана на отрезке $[0, +\infty)$; и функция $f_X(x)$ — действительно плотность вероятности, поскольку при $\lambda > 0$ выполняются свойства функции плотности (4.12) и (4.13):

$$f_X(x) \geqslant 0$$

и

$$\int\limits_{-\infty}^{+\infty} f_X(x)\,dx = \int\limits_{0}^{+\infty} \lambda e^{-\lambda x}\,dx = \lambda\left(-\frac{1}{\lambda}\right)e^{-\lambda x}\Big|_0^{+\infty} =$$

$$= -e^{-\lambda x}\Big|_0^{+\infty} = \lim_{x\to\infty}(-e^{-\lambda x}) - (-1) = 0 + 1 = 1.$$

Для нахождения дисперсии DX надо знать математическое ожидание MX.

Согласно (4.16),

$$MX = \int\limits_{0}^{+\infty} x f_X(x)\,dx = \int\limits_{0}^{+\infty} x\lambda e^{-\lambda x}\,dx.$$

Проинтегрируем формулу по частям. Положим $\boldsymbol{u = x}$, $dv = \lambda e^{-\lambda x}\,dx$; тогда $du = dx$, $v = \int \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}$, и

$$MX = \int\limits_{0}^{+\infty} x\lambda e^{-\lambda x}\,dx = uv\Big|_0^{+\infty} - \int\limits_{0}^{+\infty} v\,du =$$

$$= -xe^{-\lambda x}\Big|_0^{+\infty} - \int\limits_{0}^{+\infty} -e^{-\lambda x}\,dx = [\lim_{x\to+\infty}(-xe^{-\lambda x}) - 0] - \frac{1}{\lambda}e^{-\lambda x}\Big|_0^{+\infty} =$$

$$= 0 - \frac{1}{\lambda}(\lim_{x\to+\infty} e^{-\lambda x} - 1) = \frac{1}{\lambda}.$$

Итак, $MX = 1/\lambda$.

Теперь найдем дисперсию. Согласно (4.45),

$$DX = \int\limits_{0}^{+\infty} x^2 f_X(x)\,dx - (MX)^2 = \int\limits_{0}^{+\infty} x^2\lambda e^{-\lambda x}\,dx - \frac{1}{\lambda^2}.$$

Проинтегрируем формулу по частям. Положим $u = x^2$, $dv = \lambda e^{-\lambda x}\,dx$; тогда $du = 2x\,dx$, а $v = -e^{-\lambda x}$ и

$$DX = uv\Big|_0^{+\infty} - \int\limits_{0}^{+\infty} v\,du - \frac{1}{\lambda^2} = -x^2e^{-\lambda x}\Big|_0^{+\infty} - \int\limits_{0}^{+\infty} -2xe^{-\lambda x}\,dx - \frac{1}{\lambda^2} =$$

$$= 0 + 2\int\limits_{0}^{+\infty} xe^{-\lambda x}\,dx - \frac{1}{\lambda^2} = 2\frac{1}{\lambda}\underbrace{\int\limits_{0}^{+\infty} x\lambda e^{-\lambda x}\,dx}_{MX} - \frac{1}{\lambda^2} = \frac{2}{\lambda}MX - \frac{1}{\lambda^2} =$$

$$= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Итак, $DX = \frac{1}{\lambda^2}$. ◀

Рассмотрим *свойства дисперсии* (среднего квадратического отклонения), имеющие место как для дискретных, так и для непрерывных величин. При их доказательстве используют свойства математического ожидания.

Основные свойства дисперсии

1^0. *Дисперсия постоянной с равна нулю*

$$D(c) = 0. \tag{4.46}$$

≫ Рассматривая постоянную c как частный вид случайной величины и учитывая, что $Mc = c$, получим

$$Dc \underset{(4.39)}{=} M(c - Mc)^2 = M(c - c)^2 = M0 = 0. \quad \ll$$

Учитывая (4.46), получим, что среднее квадратическое отклонение постоянной c равно нулю:

$$\sigma_c = \sqrt{Dc} = 0. \tag{4.47}$$

2^0. *Константу с выносят за знак* ***дисперсии*** *в квадрате*

$$D(cX) = c^2 DX. \tag{4.48}$$

≫ Имеем

$$D(cX) \underset{(4.39)}{=} M(cX - M(cX))^2 \underset{(4.18)}{=} M(cX - cMX)^2 =$$

$$= M[c^2(X - MX)^2] \underset{(4.18)}{=} c^2 M(X - MX)^2 = c^2 DX. \quad \ll$$

Учитывая (4.48), получим

$$\sigma_{cX} = \sqrt{D(cX)} = \sqrt{c^2 DX} = |c|\sigma_X. \tag{4.49}$$

3^0. *Дисперсия суммы независимых случайных величин X и Y равна сумме их дисперсий*:

$$D(X + Y) = DX + DY. \tag{4.50}$$

≫ Имеем

$$D(X + Y) = M[(X + Y) - M(X + Y)]^2 \underset{(4.19)}{=} M[(X + Y) - (MX + MY)]^2 =$$

$$= M[(X - MX) + (Y - MY)]^2 = M[(X - MX)^2 + (Y - MY)^2 +$$

$$+ 2(X - MX)(Y - MY)] \underset{(4.25)}{=} M(X - MX)^2 + M(Y - MY)^2 +$$

$$+ M[2(X - MX)(Y - MY)] \underset{\substack{(4.39)\\(4.18)}}{=} \underbrace{DX + DY + 2M[(X - MX)(Y - MY)]} =$$

$$= DX + DY + 2M(XY - XMY - YMX + MX \cdot MY) \underset{(4.25)}{=}$$

$$\underset{(4.25)}{=} DX + DY + 2[M(XY) - M(X \cdot MY) - M(Y \cdot MX) + M(MX \cdot MY)] \underset{\substack{(4.18)\\(4.17)}}{=}$$

$$\underset{\substack{(4.18)\\(4.17)}}{=} DX + DY + 2[M(XY) - MY \cdot MX - MX \cdot MY + MX \cdot MY] =$$

$$= DX + DY + 2(M(XY) - MY \cdot MX) \underset{(*)}{=} DX + DY$$

(так как X и Y — независимы, то согласно (4.20), $M(XY) = MX \cdot MY$). ≪

Учитывая (4.50), получим, что среднее квадратическое отклонение суммы *независимых* случайных величин

$$\sigma_{X+Y} = \sqrt{D(X+Y)} = \sqrt{DX + DY} = \sqrt{\sigma_X^2 + \sigma_Y^2}. \quad (4.51)$$

4^0. *Дисперсия суммы любых случайных величин* X *и* Y

$$D(X+Y) = DX + DY + 2r_{X,Y}\sigma_X\sigma_Y, \quad (4.52)$$

где

$$r_{X,Y} = \frac{M[(X - MX)(Y - MY)]}{\sigma_X\sigma_Y} \quad (4.53)$$

— *числовая характеристика взаимосвязи двух величин* X и Y, *которую называют* ***коэффициентом корреляции*** (его смысл и свойства изучаются в § 11.2).

≫ При доказательстве свойства 3^0 независимость величин X и Y была использована лишь при переходе (*) (для независимых величин, согласно (4.20), $M(XY) = MX \cdot MY$). Если к X и Y не предъявлено требование независимости, то, «остановив» доказательство свойства 3^0 подчеркнутым соотношением, получим

$$D(X+Y) = DX + DY + 2\underline{M[(X - MX)(Y - MY)]} \underset{(4.53)}{=}$$

$$\underset{(4.53)}{=} DX + DY + 2r_{X,Y}\sigma_X\sigma_Y. \quad ≪$$

Замечание. При доказательстве свойства 3^0 было получено

$$M[(X - MX)(Y - MY)] = M(XY) - MX \cdot MY,$$

следовательно, формуле (4.53) тождественна формула

$$r_{X,Y} = \frac{M(XY) - MXMY}{\sigma_X\sigma_Y}, \quad (4.54)$$

из которой видно, что для независимых величин X и Y (в этом случае $M(XY) = MX \cdot MY$) выполнено $r_{X,Y} = 0$. Поэтому равенство (4.50), верное только для независимых величин, является частным случаем равенства (4.52).

Учитывая (4.52), получаем, что для любых случайных величин X и Y

$$\sigma_{X+Y} = \sqrt{DX + DY + 2r_{X,Y}\sigma_X\sigma_Y}. \quad (4.55)$$

Приведем *следствия свойств дисперсии* (в их справедливости предлагаем убедиться самостоятельно).

Некоторые следствия свойств дисперсии

1) $D(-X) = DX$; $\sigma_{-X} = \sigma_X$; (4.56)

2) $D(a + bX) = b^2 DX$; $\sigma_{a+bX} = |b|\sigma_X$, (4.57)

где a и b — неслучайные величины.

3) $D(X - Y) = DX + DY$; $\sigma_{X-Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$, (4.58)

где X и Y — независимые случайные величины.

4) $D\sum\limits_{i=1}^{k} X_i = \sum\limits_{i=1}^{k} DX_i$, (4.59)

где $X_1, X_2, \ldots, X_k$ — независимые в совокупности случайные величины.

5) $D(a + \sum\limits_{i=1}^{k} b_i X_i) = \sum\limits_{i=1}^{k} b_i^2 DX_i$, (4.60)

где $X_1, X_2, \ldots, X_k$ — независимые в совокупности случайные величины, $a, b_1, b_2, \ldots, b_k$ — неслучайные величины.

6) $D(X - Y) = DX + DY - 2r_{X,Y}\sigma_X\sigma_Y$, (4.61)

где X и Y — произвольные случайные величины.

7) $D\sum\limits_{i=1}^{3} X_i = \sum\limits_{i=1}^{3} DX_i + 2\sum\limits_{\substack{i,j=1 \\ j>i}}^{3} r_{X_i,X_j}\sigma_{X_i}\sigma_{X_j}$, (4.62)

где X_1, X_2, X_3 — произвольные случайные величины.

Равенство (4.62) можно легко обобщить на случай $k > 3$ случайных величин. (Напомним, если величины независимы, то коэффициент корреляции любых двух из них равен нулю.)

8) $D(a + \sum\limits_{i=1}^{3} b_i X_i) = \sum\limits_{i=1}^{3} b_i^2 DX_i + 2\sum\limits_{\substack{i,j=1 \\ j>i}}^{3} b_i b_j r_{X_i,X_j}\sigma_{X_i}\sigma_{X_j}$, (4.63)

где X_1, X_2, X_3 — произвольные случайные величины.

Равенство (4.63) можно легко обобщить на случай $k > 3$ случайных величин.

▶ **ПРИМЕР 4.7.** Найти двумя способами дисперсию случайной величины $Y = -3X + 6$; ряд распределения величины X приведен в примере 4.3.

1-й способ. Используем найденный в примере 4.3 ряд распределения величины $Y = -3X + 6$:

y	-9	0	12
$P(Y = y)$	0,3	0,1	0,6

и уже найденные значения $MY = 4{,}5$, $DX = 10{,}05$ (см. пример 4.6). Воспользуемся формулой (4.42). Имеем

$$DY = \sum_{i=1}^{3} y_i^2 p_i - (MY)^2 =$$

$$= (-9)^2 \cdot 0{,}3 + 0^2 \cdot 0{,}1 + 12^2 \cdot 0{,}6 - 4{,}5^2 = 90{,}45.$$

2-й способ. Воспользуемся формулой (4.57),

$$DY = D(-3X + 6) = (-3)^2 DX = 9 \cdot 10{,}05 = 90{,}45.$$

Получены одинаковые результаты. ◀

▶▶ **ЗАДАЧА 4.6.** Найти математическое ожидание, дисперсию и среднее квадратическое отклонение случайной величины $\dot{X} = \dfrac{X - MX}{\sigma_X}$.

Р е ш е н и е. Учитывая, что MX и σ_X — это неслучайные (постоянные) величины и используя свойства математического ожидания, имеем

$$M\dot{X} = M\left(\frac{X - MX}{\sigma_X}\right) = \frac{1}{\sigma_X} M(X - MX) =$$

$$= \frac{1}{\sigma_X}(MX - M(MX)) = \frac{1}{\sigma_X}(MX - MX) = 0.$$

Используя свойства дисперсии, получим

$$D\dot{X} = D\left(\frac{X - MX}{\sigma_X}\right) = \frac{1}{\sigma_X^2} D(X - MX) = \frac{1}{\sigma_X^2}(DX - D(MX)) =$$

$$= \frac{1}{\sigma_X^2}(DX - 0) = \frac{DX}{DX} = 1.$$

Итак, $M\dot{X} = 0$, $D\dot{X} = 1$ и $\sigma_{\dot{X}} = 1$. ◀◀

Случайную величину $\dot{X}$, у которой $M\dot{X} = 0$, а $D\dot{X} = 1$, называют ***стандартной***, а переход от величины X к величине $\dot{X} = (X - MX)/\sigma_X$ — ***стандартизацией*** случайной величины X.

Величину $(X - MX)$ называют ***центрированной***; ее математическое ожидание равно нулю:

$$M(X - MX) = MX - M(MX) = MX - MX = 0.$$

Величину X/σ_X называют ***нормированной***; ее дисперсия равна единице:

$$D(X/\sigma_X) = \frac{1}{\sigma_X^2} DX = 1.$$

Коэффициент вариации случайной величины X

$$V_X = \sigma_X/|MX| \qquad (4.64)$$

является характеристикой рассеивания значений величины X (около MX), сопоставленного с математическим ожиданием MX. Чем меньше коэффициент вариации, тем более «представительно» математическое ожидание MX (в смысле замены значений величины X ее математическим ожиданием).

4.3.3. Начальные и центральные моменты. Коэффициенты асимметрии и эксцесса. ***Моментом k-го порядка*** случайной величины X относительно числа a называют $M(X - a)^k$ — математическое ожидание случайной величины $(X - a)^k$.

Если $a = 0$, то момент называют ***начальным***; обозначение: $\nu_k(X)$;

$$\nu_k(X) = M(X^k). \qquad (4.65)$$

Если $a = MX$, то момент называют ***центральным***; обозначение: $\mu_k(X)$;

$$\mu_k(X) = M(X - MX)^k. \qquad (4.66)$$

Начальные и центральные моменты находят по следующим формулам:

если X — дискретная величина с рядом распределения

x	x_1	x_2	...	x_n
$P(X = x)$	p_1	p_2	...	p_n

то

$$\nu_k(X) = \sum_{i=1}^{n} x_i^k p_i, \qquad \mu_k(X) = \sum_{i=1}^{n} (x_i - MX)^k p_i;$$

если X — непрерывная величина с плотностью $f_X(x)$, заданная на $[\alpha, \beta]$, то

$$\nu_k(X) = \int_{\alpha}^{\beta} x^k f_X(x)\,dx, \qquad \mu_k(X) = \int_{\alpha}^{\beta} (x - MX)^k f_X(x)\,dx.$$

В таблице 4.9 приведены начальные и центральные моменты порядков $k = 0, 1, 2, 3, 4$ и выражения центральных моментов через начальные.

Таблица 4.9

k	$\nu_k(X)$	$\mu_k(X)$
0	$\nu_0 = M(X^0) = 1$	$\mu_0 = M(X - MX)^0 = 1$
1	$\nu_1 = MX$	$\mu_1 = M(X - MX) = 0$
2	$\nu_2 = M(X^2)$	$\mu_2 = \underbrace{M(X - MX)^2}_{DX} = M(X^2) - (MX)^2 = \nu_2 - \nu_1^2$ (4.67)
3	$\nu_3 = M(X^3)$	$\mu_3 = M(X - MX)^3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3$ (4.68)
4	$\nu_4 = M(X^4)$	$\mu_4 = M(X - MX)^4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4$ (4.69)

Доказательство соотношений (4.68) и (4.69) основано на свойствах математического ожидания. Докажем, например, (4.68).

≫ Имеем

$$M(X - MX)^3 = M(X^3 - 3X^2MX + 3X \cdot (MX)^2 - (MX)^3) =$$
$$= M(X^3) - M(3X^2MX) + M(3X(MX)^2) - M(MX)^3 =$$
$$= M(X^3) - 3(MX)M(X^2) + 3(MX)^2MX - (MX)^3 =$$
$$= M(X^3) - 3(MX)M(X^2) + 2(MX)^3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3. \quad \ll$$

Аналогично доказывается соотношение (4.69).

Центральный момент k-порядка имеет следующее свойство:

$$\mu_k(a + bX) = b^k\mu_k(X), \qquad (4.70)$$

где a и b — неслучайные величины.

≫ Введем обозначение $Y = a + bX$ и вспомним, что $M(a + bX) = a + bMX$. Тогда

$$\mu_k(a + bX) = \mu_k(Y) \underset{(4.66)}{=} M(Y - MY)^k = M(a + bX - a - bMX)^k =$$
$$= M[b^k(X - MX)^k] = b^kM(X - MX)^k = b^k\mu_k(X),$$

что и требовалось доказать. ≪

С центральными моментами третьего и четвертого порядка связаны коэффициенты асимметрии и эксцесса.

Коэффициент асимметрии:

$$A_X = \mu_3(X)/\sigma_X^3 . \qquad (4.71)$$

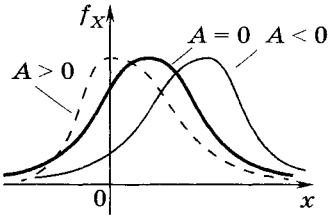


Рис. 4.12

Для симметричных распределений $A = 0$; для распределений с «длинной частью» кривой, расположенной справа от ее вершины, $A > 0$; а распределения с «длинной частью» кривой, расположенной слева от ее вершины, $A < 0$ (рис. 4.12). Обычно $|A| < 2$; A — безразмерная характеристика.

Коэффициент эксцесса, или островершинности распределения:

$$E_X = \mu_4(X)/\sigma_X^4 - 3. \qquad (4.72)$$

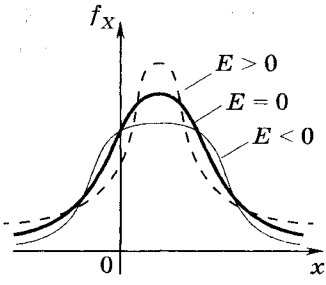


Рис. 4.13

Началом отсчета в измерении «островершинности» служит нормальное распределение (см. 5.2.3), для которого $E_X = 0$. Как правило, для распределений с более высокой и более острой вершиной кривой распределения (многоугольника вероятностей) $E > 0$; а с менее острой $E < 0$ (рис. 4.13). Обычно $-1 \leqslant E \leqslant 6$; E — безразмерная величина.

Коэффициент асимметрии имеет следующее свойство:

$$A_{a+bX} = \begin{cases} A_X, & \text{если } b > 0, \\ -A_X, & \text{если } b < 0. \end{cases} \qquad (4.73)$$

≫ Согласно (4.71),

$$A_{a+bX} = \mu_3(a + bX)/\sigma_{a+bX}^3 ;$$

на основании (4.57) и (4.70),

$$\sigma_{a+bX}^3 = |b|^3\sigma_X^3 , \ \mu_3(a + bX) = b^3\mu_3(X).$$

Поэтому

$$A_{a+bX} = \frac{b^3\mu_3(X)}{|b|^3\sigma_X^3} = \frac{b^3}{|b|^3} A_X = \begin{cases} A_X, & \text{если } b > 0, \\ -A_X, & \text{если } b < 0. \end{cases} \quad \ll$$

Коэффициент эксцесса имеет следующее свойство:

$$E_{a+bX} = E_X. \qquad (4.74)$$

≫ Учитывая (4.57) и (4.70), получим

$$E_{a+bX} = \frac{\mu_4(a+bX)}{\sigma^4_{a+bX}} - 3 = \frac{b^4\mu_4(X)}{|b|^4\sigma^4_X} - 3 = E_X. \quad \ll$$

§ 4.4. Математическое ожидание и среднее квадратическое отклонение как характеристики финансовых операций

Финансовая операция может быть рискованной, если получаемый в результате ее проведения доход случаен, не известен на момент начала операции, и безрисковой, если доход на момент начала операции точно известен. В условиях рыночной экономики, непрогнозируемости инфляции и обменного курса валют практически все финансовые операции рискованные.

Пусть X (ден. ед.) — случайный доход финансовой операции O_X; MX (ден. ед.) — ожидаемый в среднем ее доход (напомним, что MX — постоянная величина). За риск операции принимают среднее квадратическое отклонение $\sigma_X = \sqrt{DX}$ (ден. ед.) — характеристику разброса дохода X вокруг MX — ожидаемого в среднем дохода. Чем больше σ_X, тем больше «нервное напряжение» финансиста: ведь при большом σ_X наряду с возможностью большого дохода существует возможность и большого убытка; при малом σ_X нет больших доходов, но нет и больших убытков.

≫ **ЗАДАЧА 4.7.** Проводятся две независимые рискованные финансовые операции O_1 и O_2 (независимость операций означает, что их доходы X_1 и X_2 — независимые случайные величины). Как часто, или, иначе, с какими вероятностями следует пользоваться этими операциями для сведения риска к минимуму?

Замечание. Задача имеет смысл, если среди операций O_1 и O_2 нет такой, у которой ожидаемый в среднем доход был бы больше, а риск — меньше, чем у другой. При наличии таковой именно ее и следует использовать. В дальнейшем будем предполагать, что средний доход $MX_2 = m_2$ второй операции больше среднего дохода $MX_1 = m_1$ первой ($m_1 < m_2$), но и риск второй операции больше риска первой ($\sigma_1 < \sigma_2$).

Решение. Пусть p — вероятность использования операции O_1, а $(1 - p)$ — вероятность использования операции O_2 (например, при $p = 1/5$ операция O_1 будет проводиться в одном случае, а операция O_2 — в четырех случа-

ях из пяти). При такой комбинации операций случайный доход

$$Y = pX_1 + (1-p)X_2,$$

ожидаемый в среднем доход

$$MY = M(pX_1 + (1-p)X_2) = pMX_1 + (1-p)MX_2 = \\ = pm_1 + (1-p)m_2, \qquad (4.75)$$

а риск с учетом независимости величин X_1 и X_2

$$\sigma_Y = \sqrt{D(pX_1 + (1-p)X_2)} \underset{(4.60)}{=} \\ \underset{(4.60)}{=} \sqrt{p^2 DX_1 + (1-p)^2 DX_2} = \sqrt{p^2\sigma_1^2 + (1-p)^2\sigma_2^2}. \qquad (4.76)$$

Обратим внимание на то, что при любом $p \in [0, 1]$ средний доход (4.75) комбинации операций O_1 и O_2 заключен между m_1 и m_2 (напомним, что в задаче $m_1 < m_2$), $m_1 \leqslant MY \leqslant m_2$.

≫ Действительно, проведем тождественные преобразования системы

$$\begin{cases} MY \geqslant m_1, \\ MY \leqslant m_2, \end{cases} \text{ или системы } \begin{cases} pm_1 + (1-p)m_2 \geqslant m_1, \\ pm_1 + (1-p)m_2 \leqslant m_2. \end{cases}$$

Получим

$$\begin{cases} (1-p)m_2 \geqslant (1-p)m_1, \\ pm_1 \leqslant pm_2; \end{cases}$$

что очевидно, если учесть, что $p \in [0, 1]$ и $m_1 < m_2$. ≪

Найдем p, $0 \leqslant p \leqslant 1$, при котором риск σ_Y (4.76) минимален. Для этого рассмотрим σ_Y^2 как функцию от p на отрезке $[0, 1]$

$$\sigma_Y^2 = g(p) \underset{(4.76)}{=} p^2\sigma_1^2 + (1-p)^2\sigma_2^2 = (\sigma_1^2 + \sigma_2^2)p^2 - 2\sigma_2^2 p + \sigma_2^2.$$

На рисунке 4.14 изображен график функции $g(p)$. Это парабола, ветви которой направлены вверх, так как $\sigma_1^2 + \sigma_2^2 > 0$; она не пересекает ось абсцисс, поскольку дискриминант

$$D = (-2\sigma_2^2)^2 - 4(\sigma_1^2 + \sigma_2^2)\sigma_2^2 = -4\sigma_1^2\sigma_2^2 < 0;$$
$$g(0) = \sigma_2^2,\ g(1) = (\sigma_1^2 + \sigma_2^2) - 2\sigma_2^2 + \sigma_2^2 = \sigma_1^2.$$

Так как

$$\frac{dg}{dp} = 2p(\sigma_1^2 + \sigma_2^2) - 2\sigma_2^2 = 0 \text{ при } p^* = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \text{ и } \left.\frac{d^2g}{dp^2}\right|_{p^*} > 0,$$

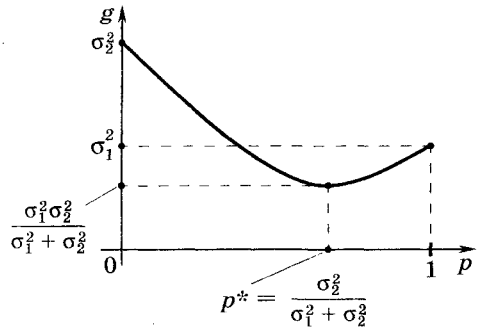


Рис. 4.14

то в точке $p = p^*$ значение функции $g(p)$ минимально и равно

$$g(p^*) = (\sigma_1^2 + \sigma_2^2)(p^*)^2 - 2\sigma_2^2 p^* + \sigma_2^2 = \sigma_1^2 \sigma_2^2 / (\sigma_1^2 + \sigma_2^2)$$

(нетрудно убедиться в том, что это значение меньше, чем σ_1^2, и тем более меньше, чем σ_2^2).

Итак, если операцию O_1, средний доход которой равен m_1, а риск σ_1, использовать с вероятностью p, а операцию O_2, средний доход которой равен m_2 ($m_2 > m_1$), а риск σ_2 ($\sigma_2 > \sigma_1$), использовать с вероятностью $(1 - p)$, то в этом случае средний доход будет заключен между m_1 и m_2, а дисперсия дохода, следовательно, и риск (как корень квадратный из дисперсии) будут наименьшими при $p = \sigma_2^2 / (\sigma_1^2 + \sigma_2^2)$. Риск будет равен $\sigma_1 \sigma_2 / \sqrt{\sigma_1^2 + \sigma_2^2}$, что меньше рисков σ_1 и σ_2 исходных операций.

ЗАДАЧА 4.8 (эффект диверсификации). Средний доход каждой из n независимых рискованных операций заключен между числами a и b, а риск — между числами c и d (c, $d > 0$). В каких пределах заключен средний доход и риск операции, состоящей в использовании исходных операций одинаково часто? Снижается ли «нервное напряжение» финансиста при увеличении n?

Р е ш е н и е. Пусть X_i — случайный доход операции O_i, $MX_i = m_i$ и σ_i — ее средний доход и риск; по условию $a \leqslant m_i \leqslant b$, а $c \leqslant \sigma_i \leqslant d$, $i = 1, 2, \ldots, n$. Если операции $O_1, O_2, \ldots, O_n$ используются одинаково часто, т. е. каждая с вероятностью $1/n$, то случайный доход

$$Y = \sum_{i=1}^{n} \frac{1}{n} X_i = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Ожидаемый в среднем доход равен

$$MY = M\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} MX_i = \frac{1}{n}\sum_{i=1}^{n} m_i,$$

а риск, с учетом независимости величин X_1, X_2, ... , X_n, равен

$$\sigma_Y = \sqrt{DY} = \sqrt{D\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} DX_i} =$$

$$= \frac{1}{n}\sqrt{\sum_{i=1}^{n} DX_i} = \frac{1}{n}\sqrt{\sum_{i=1}^{n} \sigma_i^2}.$$

Итак,

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i,\ MY = \frac{1}{n}\sum_{i=1}^{n} m_i,\ DY = \sigma_Y^2 = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2. \quad \textbf{(4.77)}$$

Так как $a \leqslant m_i \leqslant b$, $i = 1, 2, \ldots, n$, то $na \leqslant \sum_{i=1}^{n} m_i \leqslant nb$ и $a \leqslant$ $\leqslant \frac{1}{n}\sum_{i=1}^{n} m_i \leqslant b$, или $a \leqslant MY \leqslant b$, т. е. средний доход, получаемый при одинаково частом использовании исходных операций, лежит в тех же пределах, что и средний доход каждой из них.

Так как $c^2 \leqslant \sigma_i^2 \leqslant d^2$, $i = 1, 2, \ldots, n$, то $nc^2 \leqslant \sum_{i=1}^{n} \sigma_i^2 \leqslant nd^2$ и $\frac{nc^2}{n^2} \leqslant \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 \leqslant \frac{nd^2}{n^2}$. Учитывая формулу (4.77), имеем $\frac{c^2}{n} \leqslant \sigma_Y^2 \leqslant \frac{d^2}{n}$ и $\frac{c}{\sqrt{n}} \leqslant \sigma_Y \leqslant \frac{d}{\sqrt{n}}$, т. е. границы риска при одинаково частом использовании исходных операций в $\sqrt{n}$ раз меньше границ риска каждой из них, с увеличением n границы риска сужаются, и «нервное напряжение» финансиста уменьшается. В этом и состоит *эффект диверсификации* (разнообразия) — чем больше проводится независимых рискованных операций, тем уже границы риска такого действия.

ЗАДАЧА 4.9. (задача Д. Тобина[1] формирования портфеля ценных бумаг). Инвестор решил вложить деньги в три вида ценных бумаг (набор ценных бумаг называют портфелем ценных бумаг): безрисковые бумаги с гарантированной доходностью m_0 (%) и два вида бумаг со случайными зависи-

[1] *Д. Тобин* — американский экономист, лауреат Нобелевской премии, изучающий проблемы оптимальных портфелей ценных бумаг.

мыми доходностями X_1 (%) и X_2 (%), коэффициент корреляции между которыми $r_{X_1, X_2} = r_{12}$. Ожидаемые в среднем доходности от рисковых бумаг $MX_1 = m_1$ и $MX_2 = m_2$, а риски бумаг $\sigma_{X_1} = \sigma_1$ и $\sigma_{X_2} = \sigma_2$. Какую долю денег инвестор должен вложить в каждый вид бумаг, чтобы в среднем доходность портфеля составила $m_{\text{п}}$ (%), но при этом риск портфеля был бы минимальным?

Р е ш е н и е. Пусть λ_0, λ_1 и λ_2 — доли вложений соответственно в безрисковые и два вида рисковых бумаг. Тогда $Y = \lambda_0 m_0 + \lambda_1 X_1 + \lambda_2 X_2$ — случайная доходность портфеля в целом ($\lambda_0 m_0$ — неслучайная составляющая доходности); средняя доходность портфеля

$M(Y) = M(\lambda_0 m_0 + \lambda_1 X_1 + \lambda_2 X_2) = \lambda_0 m_0 + \lambda_1 MX_1 + \lambda_2 MX_2 = \lambda_0 m_0 + \lambda_1 m_1 + \lambda_2 m_2$ — она должна быть равна заданному числу $m_{\text{п}}$; риск портфеля с учетом зависимости доходов X_1 и X_2

$$\sigma_Y = \sqrt{DY} = \sqrt{D(\lambda_0 m_0 + \lambda_1 X_1 + \lambda_2 X_2)} \underset{(4.63)}{=}$$

$$\underset{(4.63)}{=} \sqrt{\lambda_1^2 DX_1 + \lambda_2^2 DX_2 + 2\lambda_1\lambda_2 r_{X_1, X_2}\sigma_{X_1}\sigma_{X_2}} =$$

$$= \sqrt{\lambda_1^2\sigma_1^2 + \lambda_2^2\sigma_2^2 + 2\lambda_1\lambda_2 r_{12}\sigma_1\sigma_2}.$$

В задаче требуется найти такие λ_0, λ_1 и λ_2, которые удовлетворяют следующим требованиям:

$$\begin{cases} \lambda_0 + \lambda_1 + \lambda_2 = 1, \\ \lambda_0 m_0 + \lambda_1 m_1 + \lambda_2 m_2 = m_{\text{п}}, \\ f(\lambda_1, \lambda_2) = \lambda_1^2\sigma_1^2 + \lambda_2^2\sigma_2^2 + 2\lambda_1\lambda_2 r_{12}\sigma_1\sigma_2 \to \min. \end{cases} \tag{4.78}$$

Схема решения задачи (4.78), например, может быть такой: из первых двух уравнений выразить λ_1 и λ_2 через λ_0; полученные выражения подставить в минимизируемую функцию и, рассматривая ее как функцию от одной неизвестной λ_0, найти минимум. Предоставляем самостоятельно получить значения λ_0^*, λ_1^*, λ_2^* долей λ_0, λ_1, λ_2. Отметим, что в случае *независимых* рисковых бумаг $r_{12} = 0$,

$$\lambda_1^* = \frac{(m_1 - m_0)(m_{\text{п}} - m_0)}{\sigma_1^2 d^2}, \qquad \lambda_2^* = \frac{(m_2 - m_0)(m_{\text{п}} - m_0)}{\sigma_2^2 d^2},$$

$$\lambda_0^* = 1 - \lambda_1^* - \lambda_2^*,$$

где $d^2 = (m_1 - m_0)^2/\sigma_1^2 + (m_2 - m_0)^2/\sigma_2^2$. ◀

УПРАЖНЕНИЯ

1. Абитуриент при поступлении в институт сдает четыре экзамена. Вероятность сдачи единичного экзамена равна 0,8. Составьте ряд распределения случайной величины X — числа сданных абитуриентом экзаменов, предположив, что сдача экзаменов — независимые испытания. Каким будет ряд распределения, если решить аналогичную задачу не для абитуриента, а для студента, сдающего четыре семестровых экзамена?

2. Ряд распределения случайной величины X имеет вид

x	1	2	7	13
$P(X = x)$	0,1	0,3	0,2	?

Требуется:

а) найти MX, x_{mod}, x_{med}, DX, σ_X, A_X, E_X;

б) построить многоугольник распределения вероятностей, найти функцию распределения $F_X(x)$ и построить ее график;

в) вычислить следующие вероятности: $P(X = 6)$, $P(X < 6)$, $P(X < 14)$, $P(1 \leqslant X < 13)$, $P((X = 2) | (X < 13))$.

3. Ряды распределения независимых случайных величин X и Y такие:

x	-1	2
$P(X = x)$	0,3	0,7

y	-1	2
$P(Y = y)$	0,3	0,7

Составьте ряды распределения случайных величин $Z = 2X$, $U = X + Y$, $V = XY$, $W = \max(X, Y)$. Найдите математические ожидания и дисперсии случайных величин Z, U двумя способами: используя их ряды распределения и на основании свойств математического ожидания и дисперсии.

4. Известны следующие характеристики случайной величины X: $MX = -4$; $\sigma_X = 1$; $x_{\text{med}} = -1{,}5$; $x_{\text{mod}} = 1$; $\mu_3(X) = 1{,}5$; $E_X = -1$. Найдите A_X и $\mu_4(X)$. Найдите также восемь алогичных характеристик для случайной величины $Y = -3X - 5$.

5. Функция плотности

$$f_X(x) = \begin{cases} ax \text{ при } x \in [0, 2], \\ 0 \quad \text{при } x \notin [0, 2]. \end{cases}$$

Найдите a, $F_X(x)$; постройте графики функций $f_X(x)$ и $F_X(x)$. Определите MX, x_{mod}, x_{med}, 95%-й квантиль, 5%-ю точку. Вычислите $P(X \leqslant 0{,}3)$, $P(|X - MX| < 0{,}5)$ двумя способами: используя $f_X(x)$ и $F_X(x)$; укажите эти вероятности на обоих графиках.

6. Случайные доходы X_1 и X_2 независимых финансовых операций O_1 и O_2 заданы соответственно рядами распределения

x_1	−5	25
$P(X_1 = x_1)$	0,01	0,99

x_2	15	25
$P(X_2 = x_2)$	0,5	0,5

Найдите средний доход и риск каждой операции. Какая из операций предпочтительнее и почему?

7. Средние доходы и риски двух независимых операций O_1 и O_2 такие: $m_1 = 15$, $m_2 = 20$, $\sigma_1 = 2$, $\sigma_2 = 4$. С какими вероятностями следует использовать эти операции для сведения риска к минимуму? Каков средний доход и риск найденной комбинации операций?

8. Сформируйте портфель Д. Тобина минимального риска из двух видов ценных бумаг: безрисковой с доходностью $m_0 = 2$ (%) и рисковой со средней доходностью $m_1 = 10$ (%) и риском $\sigma_1 = 5$ (%), если средняя доходность портфеля должна быть равной $m_п = 8$ (%).

9. Сформируйте портфель Д. Тобина минимального риска из трех видов ценных бумаг: безрисковой с доходностью $m = 2$ (%) и двух независимых рисковых со средними доходностями $m_1 = 4$ (%) и $m_2 = 5$ (%) и рисками $\sigma_1 = 1$ (%) и $\sigma_2 = 2$ (%), если средняя доходность портфеля должна быть равной $m_п = 3$ (%).

ГЛАВА 5

Модели законов распределения вероятностей и их реализация в Microsoft Excel. Метод статистических испытаний

В главе 4 установлено, что полной характеристикой случайной величины является ее закон распределения (функция распределения; или ряд распределения — для дискретной величины и плотность вероятности — для непрерывной). Приведем основные, наиболее распространенные на практике модели (формулы) законов распределения вероятностей случайных величин и реализацию этих моделей в Microsoft Excel — наиболее популярном средстве работы с электронными таблицами. Затем рассмотрим вопросы компьютерной имитации реальных процессов, в которых присутствуют элементы случайности.

§ 5.1. Основные модели дискретных распределений

Основные модели дискретных распределений были рассмотрены в гл. 3 и задаче 1.6. Вспомним эти модели, используя введенную в гл. 4 терминологию случайных величин.

5.1.1. Биномиальный закон. ***Биномиальная случайная величина*** X, или X_{Bi}[1], — это число успехов в n испытаниях Бернулли, когда вероятность успеха в единичном испытании равна p.

Биномиальный закон распределения задается формулой Бернулли (3.2), которая в терминах случайной величины записывается следующим образом:

$$P(X_{Bi} = x) = C_n^x p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n, \quad (5.1)$$

где $C_n^x = \dfrac{n!}{x!(n-x)!}$ — число сочетаний из n элементов по x $\left(\text{вместо } C_n^x \text{ иногда используют обозначение } \binom{n}{x}\right)$, а число p называют ***параметром*** биномиального закона.

Ряд распределения вероятностей биномиальной случайной величины:

x	0	1	2	...	n
$P(X_{Bi} = x)$	$(1-p)^n$	$C_n^1 p(1-p)^{n-1}$	$C_n^2 p^2 (1-p)^{n-2}$	...	p^n

(5.2)

Найдем математическое ожидание и дисперсию биномиальной случайной величины, не прибегая к формуле (4.15) математического ожидания и формуле (4.42) или (4.44) дисперсии, верным для любой дискретной случайной величины.

≫ Воспользуемся тем, что биномиальную величину можно представить как сумму n альтернативных, или булевских случайных величин.

Альтернативная, или булевская случайная величина X_{Al}[2] — это число успехов в одном испытании Бернулли, когда вероятность успешности этого испытания равна p.

---

[1] Индекс «*Bi*» от англ. *binomial* — биномиальный, двучленный.

[2] Индекс «*Al*» от англ. *alternative* — альтернативный, взаимноисключающий.

Ряд распределения вероятностей альтернативной случайной величины

x	0	1
$P(X_{Al} = x)$	$1 - p$	p

называют ***распределением Бернулли***. Найдем характеристики альтернативной случайной величины:

$$MX_{Al} = 0 \cdot (1 - p) + 1 \cdot p = p,$$

$$DX_{Al} = M(X_{Al})^2 - (MX_{Al})^2 =$$

$$= 0^2(1 - p) + 1^2 p - p^2 = p - p^2 = p(1 - p).$$

Итак, для альтернативной случайной величины имеем

$$MX_{Al} = p,\ DX_{Al} = p(1 - p). \qquad (5.3)$$

Число успехов в любом испытании Бернулли — альтернативная случайная величина, поскольку это число равно нулю с вероятностью $1 - p$ или равно единице с вероятностью p. Поэтому число успехов в i-м испытании Бернулли обозначим $X_{Al}^{(i)}$, $i = 1, 2, \ldots, n$. Тогда число успехов в n испытаниях Бернулли, или биномиальная случайная величина

$$X_{Bi} = X_{Al}^{(1)} + X_{Al}^{(2)} + \ldots + X_{Al}^{(n)} = \sum_{i=1}^{n} X_{Al}^{(i)}.$$

Отсюда, учитывая формулу (5.3), получим

$$MX_{Bi} = M\left(\sum_{i=1}^{n} X_{Al}^{(i)} \right) = \sum_{i=1}^{n} MX_{Al}^{(i)} = \sum_{i=1}^{n} p = np,$$

$$DX_{Bi} = D\left(\sum_{i=1}^{n} X_{Al}^{(i)} \right) \underset{(*)}{=} \sum_{i=1}^{n} DX_{Al}^{(i)} = \sum_{i=1}^{n} p(1 - p) = np(1 - p)$$

(переход (*) верен только для независимых случайных величин $X_{Al}^{(1)}$, $X_{Al}^{(2)}, \ldots, X_{Al}^{(n)}$, а они независимы в силу того, что испытания Бернулли, по определению, независимы). ≪

Таким образом, для биномиальной случайной величины X_{Bi} среднее значение, или среднее число успехов в n испытаниях,

$$MX_{Bi} = np, \qquad (5.4)$$

где p — вероятность успешности испытания; дисперсия и среднее квадратическое отклонение (характеристики среднего разброса значений величины вокруг математического ожидания) соответственно равны

$$DX_{Bi} = np(1 - p), \qquad \sigma_{X_{Bi}} = \sqrt{np(1 - p)}. \qquad (5.5)$$

Примеры биномиальной случайной величины:

1) число дефектных изделий в партии объема n, отобранной из массовой продукции, производимой в установившемся, стационарном режиме;

2) число объектов, обладающих заданным свойством, оказавшихся среди n случайно отобранных из бесконечно большой совокупности объектов.

Рассмотрим реализацию модели в Microsoft Excel. Для запуска Microsoft Excel следует:

- нажать кнопку **Пуск**, передвинуть указатель мышки на команду **Программы**, затем на Microsoft Excel и сделать щелчок.

Для работы с биномиальной моделью надо:

- на стандартной панели инструментов щелкнуть по кнопке **Вставка функций** (f_x);
- в открывшемся окне **Категория** выбрать **Статистические**, в окне **Функция** выбрать БИНОМРАСП и щелкнуть по кнопке ОК;
- заполнить аргументы функции БИНОМРАСП (число s; число испытаний; вероятность s; интегральный), где:

— число s (первая буква англ. *success* — успех) — это заданное число x успешных испытаний или заданная верхняя граница числа успешных испытаний;

— число испытаний — это число n испытаний Бернулли;

— вероятность s — это вероятность p успеха;

— интегральный — это логический аргумент, который имеет одно из двух значений: ИСТИНА или ЛОЖЬ,

и щелкнуть по кнопке ОК.

Функция БИНОМРАСП при заданных значениях аргумента, БИНОМРАСП (x; n; p; интегральный), возвращает:

— вероятность

$$P(X_{Bi} = x) = \binom{n}{x} p^x (1-p)^{n-x} = C_n^x p^x (1-p)^{n-x}$$

появления x успехов в n испытаниях, если значение аргумента «интегральный» — ЛОЖЬ;

— интегральную (накопленную) вероятность

$$P(X_{Bi} \leqslant x) = P(X_{Bi} = 0) + P(X_{Bi} = 1) + \ldots + P(X_{Bi} = x) =$$

$$= C_n^0 p^0 (1-p)^{n-0} + C_n^1 p^1 (1-p)^{n-1} + \ldots + C_n^x p^x (1-p)^{n-x} =$$

$$= \sum_{m=0}^{x} C_n^m p^m (1-p)^{n-m},$$

если значение аргумента «интегральный» — ИСТИНА.

▶ **ПРИМЕР 5.1.** Пусть X — число дефектных изделий среди трех ($n = 3$) отобранных. Вероятность того, что изделие имеет дефект, $p = 0{,}1$. Составим ряд распределения величины X, используя функцию БИНОМРАСП (x; n; p; ЛОЖЬ). Получим

x	0	1	2	3
$P(X = x)$	0,729	0,243	0,027	0,001

где, например, 0,729 = БИНОМРАСП (0; 3; 0,1; ЛОЖЬ) = $= C_3^0 0{,}1^0(1 - 0{,}1)^{3-0}$, а 0,027 = БИНОМРАСП (2; 3; 0,1; ЛОЖЬ) = $C_3^2 0{,}1^2(1 - 0{,}1)^{3-2}$.

Составим ряд накопленных (интегральных) вероятностей, используя функцию БИНОМРАСП (x; n; p; ИСТИНА). Получим

x	0	1	2	3
$P(X \leqslant x)$	0,729	0,972	0,999	1

где, например:

$$0{,}972 = \text{БИНОМРАСП (1; 3; 0,1; ИСТИНА)} =$$
$$= P(X = 0) + P(X = 1) = C_3^0 0{,}1^0(1 - 0{,}1)^{3-0} + C_3^1 0{,}1^1(1 - 0{,}1)^{3-1},$$
а
$$1 = \text{БИНОМРАСП (3; 3; 0,1; ИСТИНА)} =$$
$$= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) =$$
$$= \sum_{m=0}^{3} C_3^m 0{,}1^m(1 - 0{,}1)^{3-m}. \quad \blacktriangleleft$$

5.1.2. Закон Пуассона. Смысл ***пуассоновской случайной величины*** определяется содержанием конкретной задачи. Это может быть X, или X_{Po}[1], — число успехов в бесконечно большом числе испытаний Бернулли, когда вероятность успеха в единичном испытании мала. В этом случае ***закон Пуассона*** задается формулой Пуассона

$$P(X_{Po} = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, 2, \ldots, \tag{5.6}$$

а число λ называют ***параметром*** закона Пуассона. Или X, или $X_{Po}(t)$, — число событий простейшего потока, имеющего интенсивность λ, которое наступит за время t. В этом случае закон Пуассона задается формулой

$$P(X_{Po}(t) = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}, \qquad x = 0, 1, 2, \ldots. \tag{5.7}$$

[1] Индекс «*Po*» от англ. *Poisson* — Пуассон.

(Формула (5.7) не отличается от (5.6), если произведение λt заменить буквой λ.)

Ряд распределения вероятностей пуассоновской случайной величины X_{Po} имеет вид

x	0	1	2	3	...
$P(X_{Po}=x)$	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2}{2!}e^{-\lambda}$	$\frac{\lambda^3}{3!}e^{-\lambda}$	...

(5.8)

Сумма бесконечного числа вероятностей этого ряда равна единице (см. § 3.2). Математическое ожидание пуассоновской величины X_{Po}

$$MX_{Po} = 0\cdot e^{-\lambda} + 1\cdot\lambda e^{-\lambda} + 2\cdot\frac{\lambda^2}{2!}e^{-\lambda} + 3\cdot\frac{\lambda^3}{3!}e^{-\lambda} + \ldots =$$

$$= \lambda e^{-\lambda}\left(0 + 1 + 2\frac{\lambda}{2!} + 3\frac{\lambda^2}{3!} + \ldots\right) \underset{(*)}{=}$$

$$\underset{(*)}{=} \lambda e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right)'_{\lambda} \underset{(3.10)}{=} \lambda e^{-\lambda}e^{\lambda} = \lambda e^{0} = \lambda$$

(в правой части равенства (*) стоит производная суммы по переменной λ). Дисперсия пуассоновской случайной величины X_{Po}

$$DX_{Po} \underset{(4.44)}{=} M(X_{Po})^2 - (MX_{Po})^2 = 0^2e^{-\lambda} + 1^2\lambda e^{-\lambda} + 2^2\frac{\lambda^2}{2!}e^{-\lambda} +$$

$$+ 3^2\frac{\lambda^3}{3!}e^{-\lambda} + \ldots - \lambda^2 = \lambda e^{-\lambda}\left(0 + 1 + 2^2\frac{\lambda}{2!} + 3^2\frac{\lambda^2}{3!} + \ldots\right) - \lambda^2 =$$

$$= \lambda e^{-\lambda}\left(1 + \lambda + 2\frac{\lambda^2}{2!} + 3\frac{\lambda^3}{3!} + \ldots\right)'_{\lambda} - \lambda^2 = \lambda e^{-\lambda}(MX_{Po} : e^{-\lambda})'_{\lambda} - \lambda^2 =$$

$$= \lambda e^{-\lambda}(\lambda e^{\lambda})'_{\lambda} - \lambda^2 = \lambda e^{-\lambda}(e^{\lambda} + \lambda e^{\lambda}) - \lambda^2 = \lambda e^{0} + \lambda^2 e^{0} - \lambda^2 = \lambda.$$

Таким образом, для пуассоновской случайной величины X_{Po} среднее ее значение, или среднее числа успехов в бесконечно большом числе испытаний (с малой вероятностью p в единичном испытании), равно дисперсии этой величины

$$MX_{Po} = DX_{Po} = \lambda, \text{ а } \sigma_{X_{Po}} = \sqrt{\lambda}. \qquad (5.9)$$

Аналогично для пуассоновской величины $X_{Po}(t)$ среднее ее значение, или среднее число появлений событий простейшего потока за время t, равно дисперсии этой величины

$$MX_{Po}(t) = DX_{Po}(t) = \lambda t. \qquad (5.10)$$

Обратим внимание, что при $t = 1$ из этой формулы имеем

$$MX_{Po}(1) = \lambda, \qquad (5.11)$$

т. е. интенсивность λ — это среднее число появлений событий простейшего потока в единицу времени.

Примеры пуассоновской случайной величины:

1) число несчастных случаев и редких заболеваний в единицу времени;

2) число сбоев отлаженного производственного процесса в единицу времени;

3) число требований, поступивших в единицу времени в некоторую систему обслуживания (магазин, парикмахерскую, цех и т. д.).

В Microsoft Excel пуассоновская модель реализована как **Статистическая функция** ПУАССОН (x; среднее; интегральный), где

— x — значение пуассоновской величины (число успехов или событий простейшего потока за время t) или заданная верхняя граница значений пуассоновской величины;

— среднее — математическое ожидание пуассоновской величины (число λ или число λt);

— интегральный — логический аргумент, который имеет одно из двух значений: ИСТИНА или ЛОЖЬ.

ПУАССОН (x; λ; интегральный) возвращает

— вероятность

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

если значение аргумента «интегральный» — ЛОЖЬ;

— интегральную (накопленную) вероятность

$$P(X_{Po} \leqslant x) = P(X_{Po} = 0) + P(X_{Po} = 1) + \ldots + P(X_{Po} = x) =$$
$$= \frac{\lambda^0}{0!} e^{-\lambda} + \frac{\lambda^1}{1!} e^{-\lambda} + \ldots + \frac{\lambda^x}{x!} e^{-\lambda} = \sum_{m=0}^{x} \frac{\lambda^m}{m!} e^{-\lambda},$$

если значение аргумента «интегральный» — ИСТИНА.

▶ **ПРИМЕР 5.2.** Пусть X — число заказов такси, поступающих в диспетчерскую таксопарка в дневное время за 2 мин. Поток заказов в дневное время — простейший поток событий с интенсивностью $\lambda = 1{,}2$ заказ/мин. Найдем $P(X = 3)$ и $P(X \leqslant 3)$, используя функцию ПУАССОН.

Поскольку $\lambda = 1{,}2$ — среднее число заказов в одну минуту, среднее число заказов за $t = 2$ (мин) равно $\lambda t = 2{,}4$; это число и является значением аргумента «среднее» функции ПУАССОН. Поэтому $P(X = 3)$ = ПУАССОН (3; 2,4; ЛОЖЬ) = = 0,20901, а $P(X \leqslant 3)$ = ПУАССОН (3; 2,4; ИСТИНА) =

$= 0{,}77872$. Такие же результаты будут получены, если воспользоваться формулой (5.7):

$$P(X_{P_0}(2) = 3) = \frac{(1{,}2 \cdot 2)^3}{3!} e^{-1{,}2 \cdot 2} = 0{,}20901,$$

$$P(X_{P_0}(2) \leqslant 3) = P(X_{P_0}(2) = 0) + P(X_{P_0}(2) = 1) + P(X_{P_0}(2) = 2) +$$

$$+ P(X_{P_0}(2) = 3) = \frac{2{,}4^0}{0!} e^{-2{,}4} + \frac{2{,}4^1}{1!} e^{-2{,}4} + \frac{2{,}4^2}{2!} e^{-2{,}4} +$$

$$+ \frac{2{,}4^3}{3!} e^{-2{,}4} = 0{,}77872. \quad \blacktriangleleft$$

5.1.3. Геометрический и отрицательный биномиальный закон. ***Геометрическая случайная величина*** X, или X_G[1], — это число испытаний Бернулли (с вероятностью p успеха в единичном испытании), которые придется произвести до первого успеха, включая и успешное испытание.

Геометрический закон распределения задается формулой (3.21) геометрической вероятности

$$P(X_G = x) = (1 - p)^{x-1} p, \qquad x = 1, 2, \ldots, \qquad (5.12)$$

в которой число p называют ***параметром*** геометрического закона.

Ряд распределения вероятностей геометрической случайной величины X_G:

x	1	2	3	...
$P(X_G = x)$	p	$(1-p)p$	$(1-p)^2 p$	...

(5.13)

Сумма бесконечного числа вероятностей этого ряда равна единице (см. § 3.4). Найдем математическое ожидание и дисперсию геометрической величины, воспользовавшись обозначением $q = 1 - p$. Имеем

$$MX_G = 1p + 2qp + 3q^2 p + \ldots = p(1 + 2q + 3q^2 + \ldots) =$$

$$= p(q + q^2 + q^3 + \ldots)'_q \underset{(*)}{=} p\left(\frac{q}{1-q}\right)'_q = p\frac{(1-q)+q}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}$$

(при переходе (*) учитывалось, что $0 < q < 1$ и, следовательно, прогрессия q, q^2, q^3, ... является убывающей, а сумма членов такой прогрессии равна $\frac{q}{1-q}$);

$$DX_G = M(X_G)^2 - (MX_G)^2 =$$

$$= 1^2 p + 2^2 qp + 3^2 q^2 p + \ldots - (1/p)^2 =$$

$$= p(1 + 2^2 q + 3^2 q^2 + \ldots) - (1/p)^2 =$$

[1] Индекс «G» от англ. *geometric* — геометрический.

$$= p(q + 2 \cdot q^2 + 3 \cdot q^3 + \ldots)'_q - (1/p)^2 = p(MX_G q/p)'_q - (1/p)^2 =$$

$$= p\left(\frac{1}{p}\frac{q}{p}\right)'_q - (1/p)^2 = p\left(\frac{q}{(1-q)^2}\right)'_q - (1/p)^2 =$$

$$= p\frac{(1-q)^2 + 2(1-q)q}{(1-q)^4} - (1/p)^2 = q/p^2 = (1-p)/p^2.$$

Итак, для геометрической случайной величины X_G среднее ее значение или среднее число испытаний Бернулли до появления первого успешного, включая и само успешное испытание, равно

$$MX_G = 1/p, \tag{5.14}$$

где p — вероятность успеха в единичном испытании; дисперсия

$$DX_G = (1-p)/p^2. \tag{5.15}$$

Обобщением геометрической случайной величины X_G является случайная величина $X_{N.\,Bi}$[1] с отрицательным биномиальным законом распределения: $X_{N.\,Bi}$ — это число испытаний Бернулли (с вероятностью p успеха в единичном испытании), которые придется провести в ожидании появления успеха заданное число m раз.

Отрицательный биномиальный закон задается формулой (3.22), которая в терминах случайной величины имеет вид

$$P(X_{N.\,Bi} = x) = C_{x-1}^{m-1} p^m (1-p)^{x-m},$$

$$x = m, m+1, m+2, \ldots, \tag{5.16}$$

где $C_{x-1}^{m-1} = \frac{(x-1)!}{(m-1)!(x-m)!}$ — число сочетаний из $(x-1)$ по $(m-1)$, а число p — ***параметр*** отрицательного биномиального закона.

Так как $C_{x-1}^{1-1} = 1$, то при $m = 1$ из формулы (5.16) следует формула (5.12) геометрического закона.

Для нахождения математического ожидания и дисперсии отрицательной биномиальной величины $X_{N.\,Bi}$ воспользуемся тем, что ее можно представить как сумму m геометрических случайных величин $X_G^{(1)}, X_G^{(2)}, \ldots, X_G^{(m)}$:

$$X_{N.\,Bi} = X_G^{(1)} + X_G^{(2)} + \ldots + X_G^{(m)} = \sum_{i=1}^{m} X_G^{(i)},$$

[1] Индекс «*N. Bi*» от англ. *negative binomial* — отрицательный биномиальный.

Рис. 5.1

где $X_G^{(1)}$ — число испытаний в ожидании появления успеха в первый раз; $X_G^{(2)}$ — число испытаний, которое потребуется провести после появления успеха первый раз в ожидании появления успеха второй раз; ... ; $X_G^{(m)}$ — число испытаний, которое потребуется провести после появления успеха $(m-1)$-й раз в ожидании появления успеха m-й раз (рис. 5.1, успешные испытания отмечены «крестиком», не успешные — точками). Тогда математическое ожидание отрицательной биномиальной величины

$$MX_{N.Bi} = M\sum_{i=1}^{m} X_G^{(i)} = \sum_{i=1}^{m} MX_G^{(i)} \underset{(5.14)}{=} \sum_{i=1}^{m} 1/p = m/p;$$

дисперсия

$$DX_{N.Bi} = D\sum_{i=1}^{m} X_G^{(i)} \underset{(*)}{=} \sum_{i=1}^{m} DX_G^{(i)} \underset{(5.15)}{=} \sum_{i=1}^{m} (1-p)/p^2 = m(1-p)/p^2$$

(переход (*) верен только для следующих независимых величин $X_G^{(1)}$: $X_G^{(2)}$, ... , $X_G^{(m)}$, которые независимы в силу того, что испытания Бернулли, по определению, независимы).

Итак, для отрицательной биномиальной величины $X_{N.Bi}$ ее среднее значение, или среднее число испытаний Бернулли, которое потребуется провести в ожидании появления успеха заданное число m раз, равно

$$MX_{N.Bi} = m/p, \qquad (5.17)$$

где p — вероятность успеха в единичном испытании; дисперсия

$$DX_{N.Bi} = m(1-p)/p^2. \qquad (5.18)$$

Примеры отрицательной биномиальной (геометрической при $m = 1$) случайной величины:

1) число циклов функционирования системы, имеющей $(m-1)$ резервных (автоматически подключающихся) элементов;

2) объем выборки, необходимый для получения m объектов с заданным свойством при их случайном извлечении из бесконечно большой совокупности объектов.

В Microsoft Excel отрицательная биномиальная модель реализована как **Статистическая функция** ОТРБИНОМРАСП (число f; число s; вероятность s), где:

— число f (первая буква англ. *failure* — неудача) — число неудач, которое равно разности между числом x испытаний в ожидании появления успеха m раз и числом m успехов, т. е. $f = x - m$ ($x = m, m + 1, \ldots$);

— число s — число m успехов;

— вероятность s — вероятность p успеха.

Функция ОТРБИНОМРАСП ($x - m$; m; p) **возвращает** вероятность

$$P(X_{N.\,Bi} = x) = C_{x-1}^{m-1} p^m (1-p)^{x-m}.$$

▶ **ПРИМЕР 5.3.** Устройство содержит работающий элемент и два резервных автоматически подключающихся элемента. Вероятность выхода из строя любого работающего элемента $p = 0{,}7$. Найдем вероятность того, что устройство проработает в течение шести производственных циклов, а затем выйдет из строя.

Здесь успех в испытании (производственном цикле) — это выход из строя работающего элемента. Для того чтобы устройство вышло из строя, должны отказать все три элемента и, судя по условию, третий элемент должен отказать в течение седьмого производственного цикла. Исходя из этого, искомая вероятность — это вероятность того, что потребуется провести $x = 7$ испытаний (циклов) в ожидании трех ($m = 3$) успехов (отказов трех элементов).

Аргументы функции ОТРБИНОМРАСП в данном случае таковы: $x - m = 7 - 3 = 4$, $m = 3$, $p = 0{,}7$; получим ОТРБИНОМРАСП (4; 3; 0,7) = 0,04167 — это и есть вероятность выхода устройства из строя после шести производственных циклов (такую же вероятность получим и по формуле (5.16):

$$P(X_{N,\,Bi} = 7) =$$

$$= C_{7-1}^{3-1} \cdot 0{,}7^3 \cdot 0{,}3^4 = 0{,}04167). \quad ◀$$

5.1.4. Гипергеометрический закон. ***Гипергеометрическая случайная величина*** X, или X_{Hg}[1], — это число объектов, обладающих заданным свойством, оказавшихся среди k объектов, случайно извлеченных (без возвращения) из общей совокупности K объектов, L из которых обладают этим свойством.

[1] Индекс «*Hg*» от англ. *hypergeometric* — гипергеометрический.

Гипергеометрический закон распределения задается формулой (1.18), которая в терминах случайной величины записывается следующим образом:

$$P(X_{Hg} = x) = \frac{C_L^x C_{K-L}^{k-x}}{C_K^k}, \qquad (5.19)$$

$$1 \leqslant k \leqslant K,$$

$$x = \underbrace{\max\{0; k - (K - L)\}}_{b}, b + 1; b + 2, \ldots, \min(k, L).$$

Примем без доказательства следующие формулы:

$$MX_{Hg} = kL/K; \qquad (5.20)$$

$$DX_{Hg} = k\frac{L}{K-1}\left(1 - \frac{L}{K}\right)\left(1 - \frac{k}{K}\right). \qquad (5.21)$$

В Microsoft Excel гипергеометрическая модель реализована как **Статистическая функция** ГИПЕРГЕОМЕТ (пример s; размер выборки; генеральная совокупность s; размер генеральной совокупности), где:

— пример s — это число x «успехов» в выборке;

— размер выборки — это k;

— генеральная совокупность s — это число L «успехов» в общей совокупности;

— размер генеральной совокупности — это число K.

Функция возвращает ГИПЕРГЕОМЕТ (x; k; L; K), равную вероятности (5.19).

▶ **ПРИМЕР 5.4.** Составим ряд распределения случайной величины X — числа дефектных изделий в выборке объема $k = 4$, случайно извлеченной (без возвращения) из партии изделий объема $K = 10$, в которой содержалось $L = 8$ дефектных изделий.

Установим, какие значения может принять случайная величина X. В соответствии с формулой (5.19) получим: $\max\{(0; 4 - (10 - 8)\} = \max\{0; 2\} = 2$, а $\min(4; 8) = 4$. Поэтому $x = 2, 3, 4$. Для нахождения вероятностей этих значений воспользуемся функцией ГИПЕРГЕОМЕТ (x; 4; 8; 10), где $x = 2, 3, 4$. Получим ряд

x	2	3	4
$P(X_{Hg} = x)$	0,13(3) = 2/15	0,53(3) = 8/15	0,3(3) = 5/15

(Самостоятельно найдите вероятности по формуле (5.19) и убедитесь в том, что они совпадут с найденными.)

Среднее число дефектных изделий в выборке:

$$MX = 2 \cdot 2/15 + 3 \cdot 8/15 + 4 \cdot 5/15 = 48/15 = 3{,}2,$$

что совпадает со средним числом, найденным по формуле (5.20):

$$MX = 4 \cdot 8/10 = 3{,}2.$$

Дисперсия числа дефектных изделий в выборке

$$DX = M(X)^2 - (MX)^2 =$$

$$= 2^2 \cdot 2/15 + 3^2 \cdot 8/15 + 4^2 \cdot 5/15 - 3{,}2^2 = 32/75,$$

что совпадает с дисперсией, найденной по формуле (5.21):

$$DX = 4 \cdot \frac{8}{10-1}\left(1 - \frac{8}{10}\right)\left(1 - \frac{4}{10}\right) = \frac{32}{75}. \quad \blacktriangleleft$$

§ 5.2. Основные модели непрерывных распределений

5.2.1. Равномерный (прямоугольный) закон. ***Равномерно распределенная на отрезке*** $[a, b]$ ***случайная величина*** X, или $R(a, b)$[1] — это случайная величина, для которой вероятность того, что она примет значение из любого подотрезка заданного отрезка $[a, b]$ ее возможных значений, пропорциональна длине подотрезка и не зависит от места его расположения на отрезке $[a, b]$. Из этого определения следует

$$P(R(a, b) \in \Delta) = \begin{cases} c \cdot |\Delta|, & \text{если } \Delta \subset [a, b], \\ 0, & \text{если } \Delta \cap [a, b] = \varnothing, \end{cases} \qquad (5.22)$$

где c — коэффициент пропорциональности, $|\Delta|$ — длина отрезка Δ.

Вспомнив смысл функции плотности (см. формулу(4.8)) и учитывая, что для равномерно распределенной величины X вероятность $P(x \leqslant X < x + \Delta x) = c\,\Delta x$ при $x, x + \Delta x \in [a, b]$, получим

$$f_X(x) = \lim_{\Delta x \to 0} \frac{P(x \leqslant X < x + \Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{c\Delta x}{\Delta x} = c,$$

или

$$f_{R(a, b)}(x) = \begin{cases} c, & \text{если } x \in [a, b], \\ 0, & \text{если } x \notin [a, b]. \end{cases}$$

Найдем постоянную c, используя свойство функции плотности:

$$1 = \int_a^b f_{R(a, b)}(x)\,dx = \int_a^b c\,dx = cx\Big|_a^b = c(b - a);$$

$$c = 1/(b - a).$$

[1] Символ «R» от англ. *rectangular* — равномерный, прямоугольный.

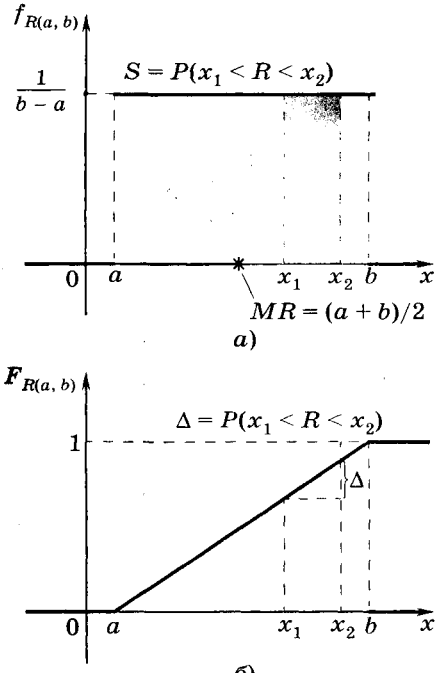


Рис. 5.2

Итак, равномерный на отрезке $[a, b]$ закон задается функцией плотности

$$f_{R(a,b)}(x) = \begin{cases} 1/(b-a), & \text{если } x \in [a, b], \\ 0, & \text{если } x \notin [a, b]. \end{cases} \quad (5.23)$$

График функции изображен на рисунке 5.2, *а*.

Числа a и b называют ***параметрами*** равномерного на отрезке $[a, b]$ закона распределения.

Найдем, используя формулу (5.23), функцию распределения равномерной на отрезке $[a, b]$ случайной величины. Имеем:

при $x < a$

$$F_{R(a,b)}(x) = P(R(a, b) < x) = P(R(a, b) \in (-\infty, x)) = 0;$$

при $a \leqslant x \leqslant b$

$$F_{R(a,b)}(x) = P(R(a, b) < x) = P(R(a, b) < a) + P(R(a, b) \in [a, x)) = 0 + c(x - a) = \frac{1}{b-a}(x - a);$$

при $x > b$

$$F_{R(a,b)}(x) = P(R(a, b) < x) = P(R(a, b) < a) + P(R(a, b) \in [a, b]) + P(b < R(a, b) < x) = 0 + c(b - a) + 0 = 1.$$

Таким образом, функция распределения равномерной на отрезке $[a, b]$ случайной величины

$$F_{R(a,b)}(x) = \begin{cases} 0 & \text{при } x < a, \\ \frac{x-a}{b-a} & \text{при } x \in [a, b], \\ 1 & \text{при } x > b. \end{cases} \quad (5.24)$$

Аналогичный результат можно было получить, если воспользоваться выражением (4.11) функции распределения через функцию плотности (5.23). График функции (5.24) изображен на рисунке 5.2, *б*.

Вероятность $P(x_1 < R(a, b) < x_2)$ численно равна площади S (см. рис. 5.2, *а*) и разности Δ ординат функции $F_{R(a, b)}(x)$ в точках $x = x_2$ и $x = x_1$ (см. рис. 5.2, *б*); $P(x_1 < R(a, b) < x_2) = (x_2 - x_1)/(b - a)$.

Найдем математическое ожидание и дисперсию величины $R(a, b)$. Имеем

$$MR(a, b) \underset{(4.16)}{=} \int_a^b x f_{R(a, b)}(x)\,dx \underset{(5.23)}{=} \int_a^b x \frac{1}{b-a}\,dx =$$

$$= \frac{1}{b-a}\frac{x^2}{2}\Big|_a^b = \frac{a+b}{2}$$

— **это середина** отрезка $[a, b]$ (см. рис. 5.2, *а*);

$$DR(a, b) \underset{(4.45)}{=} \int_a^b x^2 f_{R(a, b)}(x)\,dx - [MR(a, b)]^2 =$$

$$= \int_a^b x^2 \cdot \frac{1}{b-a}\,dx - \left(\frac{a+b}{2}\right)^2 = \frac{x^3}{3(b-a)}\Big|_a^b - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

Таким образом, для равномерно распределенной на отрезке $[a, b]$ случайной величины $R(a, b)$

$$MR(a, b) = \frac{a+b}{2}, \quad DR(a, b) = \frac{(b-a)^2}{12}, \quad \sigma_{R(a, b)} = \frac{b-a}{2\sqrt{3}}. \quad (5.25)$$

Нетрудно видеть, что для случайной величины $R(0; 1)$, равномерно распределенной на отрезке $[0; 1]$, $(a = 0, b = 1)$

$$f_{R(0; 1)}(x) = \begin{cases} 1 & \text{при } x \in [0; 1], \\ 0 & \text{при } x \notin [0; 1]; \end{cases}$$

$$F_{R(0; 1)}(x) = \begin{cases} 0 & \text{при } x < 0, \\ x & \text{при } x \in [0; 1], \\ 1 & \text{при } x > 1 \end{cases} \quad (5.26)$$

(графики этих функций изображены на рисунке 5.3);

$$MR(0; 1) = 1/2,\ DR(0; 1) = 1/12,\ \sigma_{R(0; 1)} = 1/(2\sqrt{3}).$$

▶▶ **ЗАДАЧА 5.1.** Пусть $F_X(x)$ — непрерывно возрастающая функция распределения случайной величины X. Докажите, что случайная величина $Y = F_X(X)$ равномерно распределена на отрезке $[0; 1]$.

Р е ш е н и е. Обратим внимание на следующее: так как $Y = F_X(X)$, а $0 \leqslant F_X(x) \leqslant 1$, то $Y \in [0; 1]$. Найдем $F_Y(x)$ — функцию распределения величины Y. Имеем:

при $x < 0$ событие $Y < x$ — невозможное, поэтому $F_Y(x) = P(Y < x) = 0$;

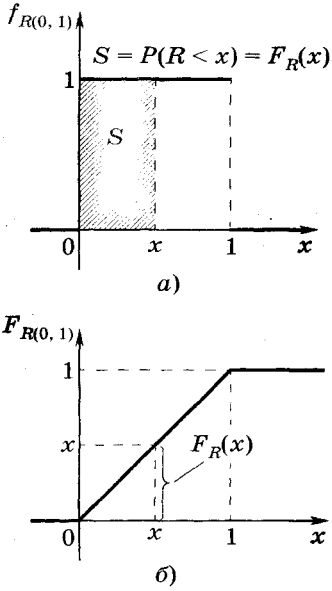


Рис. 5.3

при $x > 1$ событие $Y < x$ — достоверное, поэтому $F_Y(x) = P(Y < x) = 1$;

при $x \in [0; 1]$ получаем

$$F_Y(x) = P(Y < x) = P(F_X(X) < x) \underset{(*)}{=}$$

$$\underset{(*)}{=} P(F_X^{-1}(F_X(X)) < F_X^{-1}(x)) =$$

$$= P\,(X < \underbrace{F_X^{-1}(x)}_{d}) = F_X(d) =$$

$$= F_X(F_X^{-1}(x)) = x$$

(переход (*) основан на том, что функция F_X^{-1}, будучи обратной к непрерывно возрастающей функции $F_X(x)$, является возрастающей).

Итак,

$$F_Y(x) = \begin{cases} 0 & \text{при } x < 0, \\ x & \text{при } x \in [0; 1], \\ 1 & \text{при } x > 1. \end{cases}$$

Сравнив эту функцию с функцией (5.26), заключаем, что случайная величина $Y = F_X(X)$ равномерно распределена на отрезке $[0; 1]$. ◀

Пример равномерно распределенной случайной величины: ошибка округления при проведении числовых расчетов с фиксированным числом десятичных знаков.

▶ **ЗАДАЧА 5.2.** Взвешивание драгоценных металлов производится на весах, цена деления которых равна 0,1, а показания весов округляются при взвешивании до ближайшего деления их шкалы. Какова вероятность возникновения ошибки округления, большей, чем 0,03 г? Каковы средняя ошибка и среднее квадратическое отклонение ошибки?

Р е ш е н и е. Пусть X — ошибка округления при взвешивании (разность между показаниями весов и результатом округления); областью ее значений является отрезок $[-0{,}05; 0{,}05]$. Предположив равномерность распределения величины X на этом отрезке, запишем, согласно формулам (5.23) и (5.24), ее плотность $f_X(x)$ и функцию распределения $F_X(x)$:

$$f_X(x) = \begin{cases} 10 & \text{при } x \in [-0{,}05; 0{,}05], \\ 0 & \text{при } x \notin [-0{,}05; 0{,}05]; \end{cases}$$

$$F_X(x) = \begin{cases} 0 & \text{при } x < -0{,}05, \\ 10(x + 0{,}05) & \text{при } x \in [-0{,}05; 0{,}05], \\ 1 & \text{при } x > 0{,}05. \end{cases}$$

Найдем P(x > 0,03) двумя способами:

$$P(X > 0{,}03) = \int_{0{,}03}^{0{,}05} 10\,dx = 10x\Big|_{0{,}03}^{0{,}05} = 0{,}2;$$

$$P(X > 0{,}03) = 1 - P(X < 0{,}03) = 1 - F_X(0{,}03) =$$

$$= 1 - 10(0{,}03 + 0{,}05) = 0{,}2.$$

Согласно формулам (5.25), средняя ошибка округления $MX = 0$, среднее квадратическое отклонение $\sigma_X = 1/(20\sqrt{3})$. ≪

5.2.2. Показательный (экспоненциальный) закон. ***Показательно (экспоненциально) распределенная случайная величина*** T, или $T(\lambda)$ — это промежуток времени между двумя последовательными событиями простейшего потока, имеющего интенсивность λ (напомним, интенсивностью называют математическое ожидание числа событий потока в единицу времени); $T(\lambda) > 0$, так как одновременное наступление двух и более событий в простейшем потоке практически невозможно.

Найдем функцию распределения $F_{T(\lambda)}(t)$ показательно распределенной величины.

≫ Напомним, для простейшего потока с интенсивностью λ вероятность $P_t(m)$ того, что за время t наступит m его событий, находится по формуле (3.12)

$$P_t(m) = \frac{(\lambda t)^m}{m!} e^{-\lambda t},\ m = 0, 1, 2\ \ldots .$$

Отсюда вероятность того, что за время t не произойдет ни одного события $P_t(0) = e^{-\lambda t}$. С другой стороны, $P_t(0)$ — это вероятность $P(T(\lambda) > t)$ того, что промежуток времени $T(\lambda)$ между двумя последовательными событиями потока превысит t. Поэтому

$$P(T(\lambda) > t) = P_t(0) = e^{-\lambda t}.$$

Учитывая это равенство, запишем функцию распределения

$$F_{T(\lambda)}(t) = P(T(\lambda) < t) = 1 - P(T(\lambda) > t) = 1 - e^{-\lambda t} \text{ при } t \geqslant 0. \quad ≪$$

Итак, показательный (экспоненциальный) закон распределения задается функцией распределения

$$F_{T(\lambda)}(t) = \begin{cases} 0 & \text{при } t < 0, \\ 1 - e^{-\lambda t} & \text{при } t \geqslant 0, \end{cases} \qquad (5.27)$$

или, принимая во внимание, что плотность $f_X(x) = F'_X(x)$, функцией плотности

$$f_{T(\lambda)}(t) = \begin{cases} 0 & \text{при } t < 0, \\ \lambda e^{-\lambda t} & \text{при } t \geqslant 0. \end{cases} \qquad (5.28)$$

Число λ называют ***параметром*** показательного закона.

Графики обеих функций изображены на рисунке 5.4.

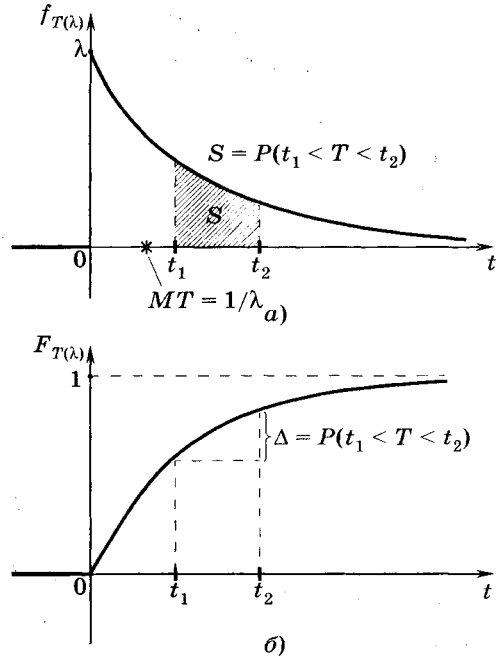


Рис. 5.4

При построении графиков учтены следующие соотношения:

t	0	$t \to \infty$
$f_{T(\lambda)}(t)$	λ	$f \to 0$

t	0	$t \to \infty$
$F_{T(\lambda)}(t)$	0	$F \to 1$

Найдем значение вероятности, указанной на рисунке 5.4. Имеем

$$P(t_1 < T(\lambda) < t_2) = \int_{t_1}^{t_2} \lambda e^{-\lambda t}\, dt = \lambda\left(-\frac{1}{\lambda}\right)e^{-\lambda t}\Big|_{t_1}^{t_2} = e^{-\lambda t_1} - e^{-\lambda t_2},$$

или

$$P(t_1 < T(\lambda) < t_2) = F_{T(\lambda)}(t_2) - F_{T(\lambda)}(t_1) =$$

$$= (1 - e^{-\lambda t_2}) - (1 - e^{-\lambda t_1}) = e^{-\lambda t_1} - e^{-\lambda t_2}.$$

Итак,

$$P(t_1 < T(\lambda) < t_2) = e^{-\lambda t_1} - e^{-\lambda t_2};\ t_1, t_2 \geqslant 0. \qquad (5.29)$$

Выше из предположения о том, что поток событий — простейший с интенсивностью λ (или, иначе, что число событий потока за время t — пуассоновская случайная величина с математическим ожиданием, равным λ), был получен показательный закон распределения (с параметром λ) промежутка времени между последовательными событиями. Можно доказать и обратное: если промежуток времени между последовательными событиями потока имеет показательный закон распределения с параметром λ, то этот поток — простейший с интенсивностью λ. Именно в силу справедливости «прямого» и обратного» утверждений показательно распределенная случайная величина была определена как промежуток времени между последовательными событиями простейшего потока.

Найдем математическое ожидание и дисперсию случайной величины $T(\lambda)$. Воспользуемся результатами примера 4.7, в котором функция плотности совпадает с функцией плотности (5.28). Для показательной (экспоненциальной) случайной величины $T(\lambda)$ имеем:

— среднее значение, или средний промежуток времени между двумя последовательными событиями простейшего потока с интенсивностью λ равен

$$MT(\lambda) = 1/\lambda \qquad (5.30)$$

(это равенство очевидно: если в единицу времени происходит в среднем λ событий потока, то промежуток времени между последовательными событиями в среднем равен $1/\lambda$; заметим также: из того, что $T(\lambda) > 0$, вытекает, что и $MT(\lambda) > 0$, поэтому и параметр показательного закона $\lambda > 0$);

— дисперсия и среднее квадратическое отклонение соответственно равны

$$DT(\lambda) = 1/\lambda^2, \qquad \sigma_{T(\lambda)} = 1/\lambda. \qquad (5.31)$$

Показательная случайная величина T обладает ***свойством отсутствия последействия***, или ***свойством марковости*** (по имени русского математика А. А. Маркова): если промежуток времени T между последовательными событиями простейшего потока уже длился некоторое время τ, то это не влияет на закон распределения оставшейся части $T - \tau$ промежутка: он является показательным с параметром λ.

≫ Так как промежуток T между последовательными событиями уже длился некоторое время $\tau > 0$, то $T \geqslant \tau$. Найдем при $t \geqslant 0$ функцию рас-

пределения оставшейся части $T - \tau$ промежутка при условии, что $T \geqslant \tau$. При этом для простоты выкладок введем следующие события: событие A, состоящее в том, что $T - \tau < t$; событие B, состоящее в том, что $T \geqslant \tau$.

Функция распределения

$$F_{(T-\tau)|(T\geqslant\tau)}(t) = P((T-\tau<t)|(T\geqslant\tau)) = P(A|B) = \frac{P(A\quad B)}{P(B)} =$$

$$= \frac{P(\tau \leqslant T < t+\tau)}{P(\tau \leqslant T < \infty)} \underset{(5.29)}{=} \frac{e^{-\lambda\tau} - e^{-\lambda(t+\tau)}}{e^{-\lambda\tau}} = 1 - e^{-\lambda t}.$$

Сравнив полученное выражение с формулой (5.27), заключаем, что оставшаяся часть $T - \tau$ промежутка времени T между последовательными событиями простейшего потока, длившегося некоторое время τ, имеет, как и T, показательный закон с параметром λ. ≪

ЗАДАЧА 5.3. В дачном поселке иногда отключают электричество на случайное время, распределенное по показательному закону, в среднем на 3 ч. На этот раз электричества нет уже 2 ч. Какова вероятность того, что: а) его дадут в ближайшие полчаса; б) его не дадут еще 1 ч?

Р е ш е н и е. Пусть T — случайный промежуток времени между последовательными отключениями электричества. По условию $MT = 3$ (ч), а согласно формуле (5.30), $MT(\lambda) = 1/\lambda$. Отсюда $\lambda = 1/3$. Используя свойство отсутствия последействия показательной величины T, получаем, что случайная величина $X = T - 2$ так же, как и T, подчиняется показательному закону с параметром $\lambda = 1/3$, т. е. при $t > 0$

$$F_{X|(T\geqslant 2)}(t) = P((X<t)|(T\geqslant 2)) = 1 - e^{-t/3}.$$

Поэтому:

а) $P((X < 0{,}5)|(T \geqslant 2)) = 1 - e^{-0{,}5/3} = 1 - e^{-1/6} = 0{,}15$;

б) $P((X > 1)|(T \geqslant 2)) = 1 - P((X < 1)|(T \geqslant 2)) = 1 - (1 - e^{-1/3}) =$
$= e^{-1/3} = 0{,}72$. ≪

В Microsoft Excel показательная (экспоненциальная) модель реализована как **Статистическая функция** ЭКСПРАСП (x; лямбда; интегральный), где:

— $x > 0$ — заданное значение показательной случайной величины или заданная верхняя граница значений этой величины;

— лямбда — параметр λ;

— интегральный — логический аргумент, который принимает одно из двух значений: ИСТИНА или ЛОЖЬ.

ЭКСПРАСП (x; λ; ЛОЖЬ) возвращает $f_{T(\lambda)}(x) = \lambda e^{-\lambda x}$, а ЭКСПРАСП ($x$; λ; ИСТИНА) возвращает $F_{T(\lambda)}(x) = P(0 < T(\lambda) < x) = 1 - e^{-\lambda x}$.

Например,

ЭКСПРАСП (0,2; 10; ЛОЖЬ) $= f_{T(\lambda=10)}(0,2) = 10e^{-10\cdot 0,2} =$ $= 1,3534$; а ЭКСПРАСП (0,2; 10; ИСТИНА) $= F_{T(\lambda=10)}(0,2) =$ $= P(0 < T(\lambda) < 0,2) = 1 - e^{-10\cdot 0,2} = 0,8647$.

5.2.3. Нормальный закон. Это основной закон теории и практики вероятностно-статистических исследований. Механизм формирования нормально распределенных случайных величин был изучен французскими учеными К. Гауссом и П. Лапласом (начало XVIII в.) Их идея состояла в том, что если:

— каждое значение непрерывной случайной величины X формируется под воздействием большого числа независимых случайных факторов $W_1, W_2, \ldots, W_N$;

— характер воздействия факторов — аддитивный (от англ. *add* — прибавлять), т. е.

$$X = a + \Delta(W_1) + \Delta(W_2) + \ldots + \Delta(W_N),$$

где a — неслучайная компонента величины X (т. е. как бы «истинное» значение X в идеализированной схеме, когда устранено влияние всех случайных факторов); $\Delta(W_i)$ — случайная добавка к a за счет воздействия фактора W_i, $i = 1, 2, \ldots, N$;

— каждая случайная добавка мала, равновероятна по знаку и не может превалировать (от англ. *prevail* — преобладать) среди остальных,

то величину X имеет смысл назвать «нормальной» (перечисленные условия формирования нормальны, типичны для многих случайных величин, но далеко не для всех).

Примеры нормально распределенных случайных величин:

1) ошибка измерения;

2) отклонение от номинального значения какого-то параметра изделий, производимых в условиях отлаженного, стационарного режима;

3) ошибка при стрельбе по цели.

Функция плотности нормальной величины, которую будем обозначать через $N(a, \sigma)$, имеет вид

$$f_{N(a,\sigma)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/(2\sigma^2)}, \quad -\infty < x < \infty. \qquad (5.32)$$

Можно доказать, что:

1. Математическое ожидание и среднее квадратическое отклонение нормальной величины $N(a, \sigma)$ соответственно равны числам a и σ в формуле (5.32):

$$MN(a, \sigma) = a, \qquad \sigma_{N(a,\sigma)} = \sigma. \qquad (5.33)$$

Числа a и σ называют ***параметрами*** нормального закона.

2. Мода и медиана нормального распределения равны математическому ожиданию

$$x_{\text{mod}} = x_{\text{med}} = a.$$

3. Асимметрия и эксцесс равны нулю

$$A_{N(a,\,\sigma)} = E_{N(a,\,\sigma)} = 0.$$

Воспользовавшись выражением (4.11) функции распределения через функцию плотности, найдем функцию распределения нормальной случайной величины

$$F_{N(a,\,\sigma)}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-(t-a)^2/(2\sigma^2)} dt. \qquad (5.34)$$

Построим график функции плотности (5.32) — кривую нормального распределения, называемую ***кривой Гаусса*** (рис. 5.5, *а*). Нетрудно убедиться, что:

1) кривая распределения симметрична относительно прямой $x = a$;

2) точка $\left(a;\ \frac{1}{\sigma\sqrt{2\pi}}\right)$ является точкой максимума;

3) при $x < a$ плотность — возрастающая функция, а при $x > a$ плотность — убывающая функция;

4) $\lim\limits_{x \to \pm\infty} f_{N(a,\,\sigma)}(x) = 0$;

5) точки $\left(a - \sigma,\ \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2}\right)$ и $\left(a + \sigma,\ \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2}\right)$ являются точками перегиба; при $x < a - \sigma$ и $x > a + \sigma$ кривая распределения выпукла вниз; при $x \in (a - \sigma,\ a + \sigma)$ кривая выпукла вверх.

При построении графика функции распределения нормальной случайной величины (рис. 5.5, *б*) учитывалось следующее:

$$F_{N(a,\,\sigma)}(x) = P(N(a, \sigma) < x) \to 0 \text{ при } x \to -\infty;$$

$$F_{N(a,\,\sigma)}(x) = P(N(a, \sigma) < x) \to 1 \text{ при } x \to \infty;$$

$$F_{N(a,\,\sigma)}(a) = P(N(a, \sigma) < a) = \int\limits_{-\infty}^{a} f_{N(a,\,\sigma)}(x)\, dx =$$

$$= \frac{1}{2} \int\limits_{-\infty}^{\infty} f_{N(a,\,\sigma)}(x)\, dx = \frac{1}{2} \cdot 1 = \frac{1}{2}.$$

Установим, как влияет изменение математического ожидания a и среднего квадратического отклонения σ на вид кривой Гаусса. Очевидно, что изменение a приводит к сдвигу кривой: если a увеличивается, то кривая сдвигается вправо, если a уменьшается, то кривая сдвигается влево;

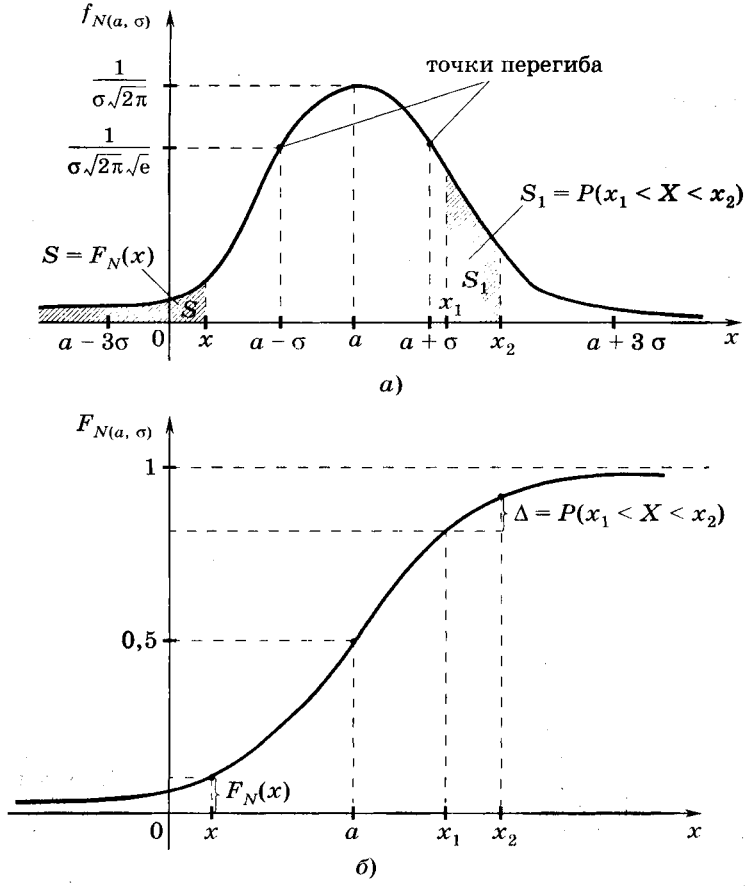


Рис. 5.5

форма кривой при этом не меняется (рис. 5.6, *а*). При изменении σ изменяется форма кривой: при уменьшении σ вершина кривой «заостряется», поднимается вверх $\left(\frac{1}{\sigma\sqrt{2\pi}}\right.$ увеличивается ! $\left.\right)$, а ее ветви «стягиваются» к прямой $x = a$ (площадь под кривой распределения должна остаться равной 1); при увеличении σ вершина кривой становится все более «пологой», опускается, а ее ветви удаляются от прямой $x = a$ (рис. 5.6, *б*).

В Microsoft Excel нормальная модель реализована двумя **Статистическими функциями**:

— НОРМРАСП (x; a; σ; интегральный), которая возвращает либо значение функции распределения $F_{N(a,\sigma)}(x)$ =

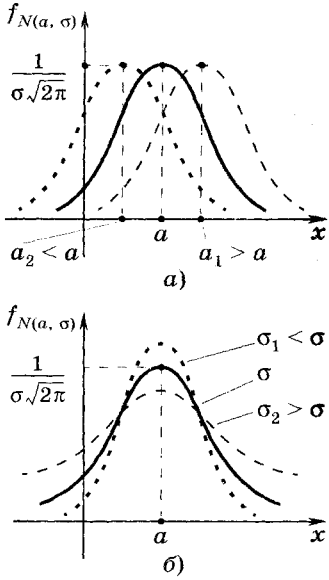


Рис. 5.6

$= P(N(a, \sigma) < x)$, если логический аргумент «интегральный» есть ИСТИНА, либо значение плотности $f_{N(a,\sigma)}(x)$, если аргумент «интегральный» — ЛОЖЬ;

— НОРМОБР (p; a; σ), которая возвращает значение функции $x = F^{-1}_{N(a,\sigma)}(p)$, обратной к функции распределения $F_{N(a,\sigma)}(x)$; или, иначе, возвращает квантиль порядка p, т. е. такое число x, при котором $P(N(a, \sigma) < x) = p$ (см. формулу (4.36)). Например,

НОРМРАСП (42; 40; 1,5; ИСТИНА) $= F_{N(40;\,1,5)}(42) = P(N(40; 1,5) < 42) = \frac{1}{1,5\sqrt{2\pi}} \int\limits_{-\infty}^{42} e^{-(x-40)^2/(2\cdot 1,5^2)}\,dx = 0,908789$;

НОРМОБР (0,908789; 40; 1,5) $= 42$, так как $F^{-1}_{N(40;\,1,5)}(0,908789) = 42$, или $P(N(40; 1,5) < 42) = 0,908789$.

Введем в рассмотрение ***стандартную нормальную величину***. Стандартизируем нормальную случайную величину $X = N(a, \sigma)$ (см. задачу 4.6), т. е. перейдем к величине $Z = (X - MX)/\sigma_X = (X - a)/\sigma$. Напомним, что для стандартной величины Z: $MZ = 0$, а $\sigma_Z = 1$. Убедимся в том, что Z, так же как и $X = N(a, \sigma)$, подчиняется нормальному закону.

≫ Для функции распределения величины Z справедлива следующая цепочка равенств:

$$F_Z(z) = P(Z < z) = P\left(\frac{X - a}{\sigma} < z\right) = P(X < a + \sigma z) = F_X(a + \sigma z).$$

Тогда функция плотности величины Z

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{dF_X(a + \sigma z)}{dz} \underset{a+\sigma z = x}{=} \frac{dF_X(x)}{dz} = \frac{dF_X(x)}{dx}\frac{dx}{dz} =$$

$$= f_X(x)\sigma = f_X(a + \sigma z)\sigma \underset{(5.32)}{=}$$

$$\underset{(5.32)}{=} \frac{1}{\sigma\sqrt{2\pi}} e^{-(a+\sigma z-a)^2/(2\sigma^2)}\sigma = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

Итак, функция плотности величины $Z = (X - a)/\sigma$, где $X = N(a, \sigma)$ — нормальная случайная величина, имеет вид

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \qquad (5.35)$$

Сравнив формулы (5.35) и (5.32), видим, что если в (5.32) положить $a = 0$ и $\sigma = 1$, то получим (5.35); следовательно, Z имеет нормальный закон распределения с параметрами $a = 0$ и $\sigma = 1$, т. е. $Z = N(0; 1)$, что и требовалось доказать. ≪

Учитывая, что функция распределения

$$F_Z(z) = \int_{-\infty}^{z} f_Z(t)\,dt,$$

получаем, что стандартная нормальная величина $Z = N(0; 1)$ имеет функцию распределения

$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\,dt. \qquad (5.36)$$

Графики функций (5.35) и (5.36) изображены на рисунке 5.7.

Итак, если исходная величина имеет нормальное распределение, то и соответствующая ей стандартная величина нормально распределена

$$Z = (N(a, \sigma) - a)/\sigma = N(0; 1). \qquad (5.37)$$

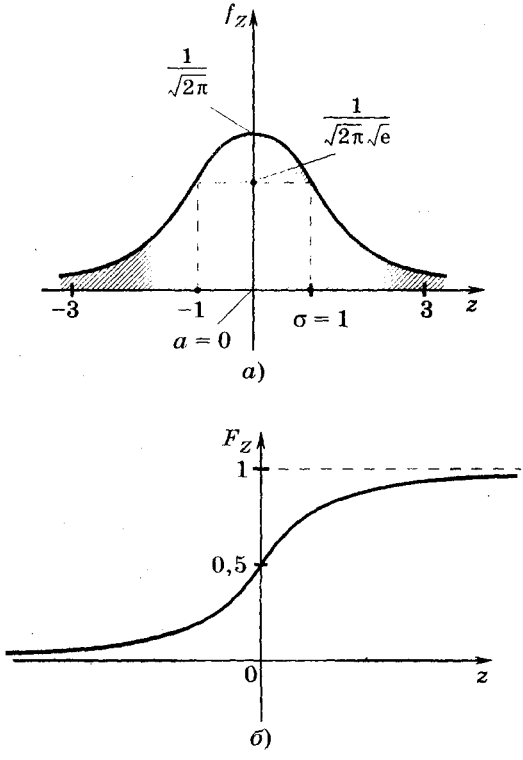


Рис. 5.7

Верно и обратное утверждение, а именно из равенства $(Y - MY)/\sigma_Y = N(0; 1)$ следует, что $Y = N(MY; \sigma_Y)$.

Значения функций (5.35) и (5.36) можно получить, воспользовавшись **Статистической функцией** НОРМРАСП (x; a; σ; интегральный), если принять $a = 0$, а $\sigma = 1$, где «интегральный» принимает значение ЛОЖЬ для функции (5.35) и ИСТИНА для (5.36). Для получения значений функции (5.36) в Microsoft Excel существует также специальная функция НОРМРАСП (z), которая при заданном числе z находит значение $F_Z(z) = P(N(0; 1) < z)$.

Значения функции плотности $f_Z(z)$ (5.35) стандартной нормальной величины (в дальнейшем функцию $f_Z(z)$ будем обозначать $\varphi(z)$) можно определить по таблице приложения П. 1[1]. Причем так как $\varphi(z) = f_Z(z)$ — четная функция, т. е. $\varphi(-z) = \varphi(z)$, что следует из (5.35), то в таблице приведены значения $\varphi(z)$ при $z \geqslant 0$.

В приложении П. 1 приведена также таблица значений функции Лапласа

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-t^2/2}\, dt \qquad (5.38)$$

при $z \geqslant 0$. Нет необходимости приводить значения $\Phi(z)$ при $z < 0$, так как из (5.38) видно, что $\Phi(z)$ — нечетная функция, т. е. $\Phi(-z) = -\Phi(z)$.

Проследим за изменениями табличных значений $\Phi(z)$. Поскольку $z \geqslant 0$, геометрически интеграл (5.38) — это площадь заштрихованной криволинейной трапеции (см. П. 1), основанием которой является интервал $(0; z)$. С ростом z эта площадь увеличивается, оставаясь меньше 0,5 (вся площадь под кривой равна 1). Поэтому и табличные значения функции $\Phi(z)$ растут при увеличении z, оставаясь меньше 0,5.

Докажем, что функция распределения $F_Z(z)$ (5.36) стандартной нормальной величины и функция Лапласа (5.38) связаны следующим соотношением:

$$F_Z(z) = 0{,}5 + \Phi(z). \qquad (5.39)$$

≫ Каким бы ни было число z, положительным или отрицательным, всегда

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-t^2/2}\, dt + \frac{1}{\sqrt{2\pi}} \int_0^z e^{-t^2/2}\, dt$$

[1] Напомним, функции $f_z(z)$ и $\Phi(z)$ уже встречались при использовании локальной и интегральной формул Муавра — Лапласа (см. § 3.3).

(убедитесь в этом самостоятельно), а поскольку в правой части равенства первое слагаемое равно 0,5 — это половина площади под кривой распределения, изображенной на рисунке 5.7, *а*, а второе слагаемое — это $\Phi(z)$, то, учитывая, что левая часть равенства — это функция $F_Z(z)$, получаем соотношение (5.39). ≪

Приведем формулы вычисления вероятностей для нормальной случайной величины с использованием таблиц значений функции плотности стандартной нормальной величины и функции Лапласа.

1. $$P(x < N(a, \sigma) < x + \mathrm{d}x) \approx \frac{1}{\sigma} f_Z\left(\frac{x-a}{\sigma}\right)\mathrm{d}x = \frac{1}{\sigma}\varphi\left(\frac{x-a}{\sigma}\right)\mathrm{d}x, \tag{5.40}$$

где $\mathrm{d}x$ — элементарный (малый) участок.

≫ Напомним, что при малом $\mathrm{d}x$ искомая вероятность равна (с точностью до бесконечно малых высшего порядка) элементу вероятности $f_{N(a,\sigma)}(x)\,\mathrm{d}x$:

$$P(x < N(a, \sigma) < x + \mathrm{d}x) \approx f_{N(a\,\sigma)}(x)\,\mathrm{d}x. \tag{5.41}$$

Сопоставляя выражения (5.32) и (5.35), увидим, что

$$f_{N(a,\sigma)}(x) = \frac{1}{\sigma} f_Z\left(\frac{x-a}{\sigma}\right) = \frac{1}{\sigma}\varphi\left(\frac{x-a}{\sigma}\right). \tag{5.42}$$

Заменив в приближенном равенстве (5.41) функцию $f_{N(a,\sigma)}(x)$ ее выражением (5.42), получим формулу (5.40). ≪

Приведем алгоритм вычисления значения функции $f_{N(a,\sigma)}(x)$:

$$x \to z = \frac{x-a}{\sigma} \xrightarrow[\text{П.1}]{} \varphi(z) \to \frac{1}{\sigma}\varphi(z) = f_{N(a\,\sigma)}(x). \tag{5.43}$$

2. $$P(N(a, \sigma) < x) = 0{,}5 + \Phi\left(\frac{x-a}{\sigma}\right). \tag{5.44}$$

≫ Действительно,

$$P(N(a, \sigma) < x) = P\left(\frac{N(a,\sigma)-a}{\sigma} < \frac{x-a}{\sigma}\right) \underset{(5.37)}{=} P\left(Z < \frac{x-a}{\sigma}\right) =$$

$$= F_Z\left(\frac{x-a}{\sigma}\right) \underset{(5.39)}{=} 0{,}5 + \Phi\left(\frac{x-a}{\sigma}\right). \quad ≪$$

Алгоритм вычислений по формуле (5.44) таков:

$$x \to z = \frac{x-a}{\sigma} \xrightarrow[\text{П.1}]{} \Phi(z) \to 0{,}5 + \Phi(z) =$$

$$= P(N(a, \sigma) < x). \tag{5.45}$$

3. $$P(x_1 < N(a, \sigma) < x_2) = \Phi\left(\frac{x_2-a}{\sigma}\right) - \Phi\left(\frac{x_1-a}{\sigma}\right). \tag{5.46}$$

≫ В самом деле,

$$P(x_1 < N(a, \sigma) < x_2) \underset{(4.10)}{=} F_{N(a,\sigma)}(x_2) - F_{N(a,\sigma)}(x_1) = P(N(a, \sigma) <$$

$$< x_2) - P(N(a, \sigma) < x_1) \underset{(5.44)}{=} 0{,}5 + \Phi\left(\frac{x_2 - a}{\sigma}\right) - \left[0{,}5 + \Phi\left(\frac{x_1 - a}{\sigma}\right)\right] =$$

$$= \Phi\left(\frac{x_2 - a}{\sigma}\right) - \Phi\left(\frac{x_1 - a}{\sigma}\right). \quad \ll$$

4. $P(|N(a, \sigma) - a| < \varepsilon) = 2\Phi(\varepsilon/\sigma)$, (5.47)

где ε — любое положительное число.

≫ Действительно,

$$P(|N(a, \sigma) - a| < \varepsilon) = P(-\varepsilon < N(a, \sigma) - a < \varepsilon) =$$

$$= P(\underbrace{-\varepsilon + a}_{x_1} < N(a, \sigma) < \underbrace{\varepsilon + a}_{x_2}) \underset{(5.47)}{=} \Phi\left(\frac{\varepsilon + a - a}{\sigma}\right) - \Phi\left(\frac{-\varepsilon + a - a}{\sigma}\right) =$$

$$= \Phi(\varepsilon/\sigma) - \Phi(-\varepsilon/\sigma) \underset{\substack{\Phi\text{ — нечетная} \\ \text{функция}}}{=} \Phi(\varepsilon/\sigma) - [-\Phi(\varepsilon/\sigma)] = 2\Phi(\varepsilon/\sigma). \quad \ll$$

5. $P(|N(a, \sigma) - a| < 3\sigma) = 0{,}9972.$ (5.48)

≫ Равенство (5.48) вытекает из (5.47), если в последнем взять $\varepsilon = 3\sigma$ и воспользоваться таблицей функции Лапласа (П. 1):

$$P(|N(a, \sigma) - a| < \underbrace{3\sigma}_{\varepsilon}) \underset{(5.47)}{=} 2\Phi(3\sigma/\sigma) = 2\Phi(3) \underset{\text{П. 1}}{=} 0{,}9972. \quad \ll$$

Равенство (5.48) называют ***«правилом трех сигм»***: *если случайная величина подчиняется нормальному закону, то с вероятностью, близкой к единице, можно утверждать, что все ее значения находятся в трехсигмовом интервале относительно математического ожидания величины.* Геометрически это означает, что площадь под криволинейной трапецией (см. рис. 5.5, *а*), опирающейся на интервал $(a - 3\sigma; a + 3\sigma)$, равна 0,9972; площадь криволинейных трапеций, опирающихся на интервалы $(-\infty, a - 3\sigma)$ и $(a + 3\sigma, +\infty)$, равна 0,0028, т. е. только 0,28% значений нормальной величины $N(a, \sigma)$ лежат вне интервала $(a - 3\sigma; a + 3\sigma)$.

Применив правило 3σ к стандартной нормальной величине $Z = N(0; 1)$, у которой $a = 0$ и $\sigma = 1$, получим

$$P(|Z| < 3) \underset{(5.48)}{=} 0{,}9972, \text{ или } P(-3 < Z < 3) = 0{,}9972,$$

т. е. площадь криволинейной трапеции под кривой стандартного нормального распределения (см. рис. 5.7, *а*), опирающейся на интервал $(-3; 3)$, равна 0,9972, а половина этой площади

$$\frac{1}{\sqrt{2\pi}} \int_0^3 e^{-t^2/2}\, dt \underset{(5.38)}{=} \Phi(3) = 0{,}4986.$$

При $z > 3$ значения $\Phi(z)$ увеличиваются, оставаясь при этом незначительно больше, чем $\Phi(3)$, и меньше, чем 0,5 (см. П. 1).

Замечание. Из правила «трех сигм» вытекает, что для нормально распределенной величины $N(a, \sigma)$, практически все значения которой положительны, коэффициент вариации $V_{N(a,\sigma)} < 0{,}3(3)$.

≫ Действительно, в соответствии с правилом «трех сигм» $P(|N(a, \sigma) - a| < 3\sigma) = 0{,}9972$, т. е. практически достоверным является событие $a - 3\sigma < N(a, \sigma) < a + 3\sigma$. Далее, поскольку практически все значения величины $N(a, \sigma)$ положительны, можно считать, что $0 < a - 3\sigma$ или $\sigma < a/3$. Тогда коэффициент вариации

$$V_{N(a,\sigma)} \underset{(4.64)}{=} \sigma_{N(a,\sigma)}/|MN(a, \sigma)| = \sigma/|a| < (a/3)/a = 1/3. \quad \ll$$

ЗАДАЧА 5.4. Текущая цена акции распределена нормально с математическим ожиданием $a = 15{,}28$ ден. ед. и средним квадратическим отклонением $\sigma = 0{,}12$ ден. ед. Найдите вероятности того, что цена акции окажется: а) не выше 15,00 ден. ед.; б) не ниже 15,50 ден. ед.; в) между 15,10 ден. ед. и 15,20 ден. ед.; г) будет отличаться от средней цены менее чем на 0,25 ден. ед.; д) найдите цену акции, превышение которой возможно лишь в 10% случаев.

Замечание. Нормальная случайная величина $N(a, \sigma)$ может принять любое значение, в том числе и отрицательное; с другой стороны, в соответствии с правилом 3σ в 99,72% значений принадлежат интервалу $(a - 3\sigma; a + 3\sigma)$ и лишь в $(100 - 99{,}72)/2 = 0{,}14\%$ значений меньше $a - 3\sigma$. В условиях задачи $a - 3\sigma = 14{,}92 > 0$, поэтому вероятность получения отрицательных значений значительно меньше 0,0014. Таким образом, предположение о нормальном законе распределения цены акций — положительной величины, не вступает в противоречие с тем, что нормальная величина определена на всей числовой оси.

Решение. Имеем

$$\text{а) } P(N(a, \sigma) \leqslant 15) \underset{(5.44)}{=} 0{,}5 + \Phi\left(\frac{15 - a}{\sigma}\right) =$$

$$= 0{,}5 + \Phi\left(\frac{15 - 15{,}28}{0{,}12}\right) = 0{,}5 + \Phi(-2{,}33) =$$

$$= 0{,}5 - \Phi(2{,}33) \underset{\text{П. 1}}{=} 0{,}5 - 0{,}4906 = 0{,}0094.$$

Рассчитаем эту же вероятность, используя функцию НОРМРАСП: НОРМРАСП (15; 15,28; 0,12; ИСТИНА) вернет $P(N(15{,}28;\ 0{,}12) < 15) = 0{,}0098$. Различие результатов объясняется бóльшей точностью компьютерных вычислений по сравнению с ручными (это следует иметь в виду и в дальнейшем, при сравнении ручного и компьютерного способов решения задачи).

б) $P(N(a, \sigma) \geqslant 15{,}50) = 1 - P(N(a, \sigma) < 15{,}50) \underset{(5.44)}{=}$

$$= 1 - \left[0{,}5 + \Phi\left(\frac{15{,}50 - 15{,}28}{0{,}12}\right)\right] = 0{,}5 - \Phi(1{,}83) \underset{\text{П. 1}}{=}$$

$$\underset{\text{П. 1}}{=} 0{,}5 - 0{,}4678 = 0{,}0322.$$

в) $P(15{,}10 < N(a, \sigma) < 15{,}20) \underset{(5.46)}{=} \Phi\left(\frac{15{,}2 - a}{\sigma}\right) -$

$$- \Phi\left(\frac{15{,}1 - a}{\sigma}\right) = \Phi\left(\frac{15{,}2 - 15{,}28}{0{,}12}\right) - \Phi\left(\frac{15{,}1 - 15{,}28}{0{,}12}\right) =$$

$$= \Phi(-0{,}67) - \Phi(-1{,}5) = -\Phi(0{,}67) + \Phi(1{,}5) \underset{\text{П. 1}}{=}$$

$$\underset{\text{П. 1}}{=} -0{,}2422 + 0{,}4332 = 0{,}191.$$

Теперь для расчета той же вероятности воспользуемся приближенной формулой (5.40):

$$P(15{,}10 < N(a, \sigma) < 15{,}20) =$$

$$= P(\underbrace{15{,}10}_{x} < N(a, \sigma) < 15{,}10 + \underbrace{0{,}10}_{dx}) \underset{(5.40)}{\approx}$$

$$\underset{(5.40)}{\approx} \frac{1}{\sigma} f_Z\left(\frac{15{,}10 - a}{\sigma}\right) 0{,}1 = \frac{1}{0{,}12} f_Z\left(\frac{15{,}10 - 15{,}28}{0{,}12}\right) 0{,}1 =$$

$$= \frac{0{,}1}{0{,}12} f_Z(-1{,}5) = \frac{0{,}1}{0{,}12} \varphi(1{,}5) \underset{\text{П. 1}}{=} \frac{0{,}1}{0{,}12} \cdot 0{,}1295 = 0{,}1079$$

— результат отличается от вероятности 0,191, рассчитанной по формуле (5.46), более чем на 0,08. Такое не малое отличие объясняется тем, что приближенная формула (5.40) дает небольшую погрешность лишь при малом dx, а в данном случае $dx = 0{,}1 \approx \sigma$. Напомним, σ — это характеристика среднего разброса значений случайной величины вокруг математического ожидания, и в силу этого участок длиной $dx \approx \sigma$ не мал. Формула же (5.46) абсолютно точная.

г) $P(|N(a, \sigma) - a| < \underbrace{0{,}25}_{\varepsilon}) \underset{(5.47)}{=} 2\Phi(0{,}25/\sigma) =$

$$= 2\Phi(0{,}25/0{,}12) = 2\Phi(2{,}08) \underset{\text{П. 1}}{=} 2 \cdot 0{,}4821 = 0{,}9642.$$

д) Здесь требуется найти такое значение x величины $N(a, \sigma)$, при котором $P(N(a, \sigma) > x) = 0{,}1$ или $P(N(a, \sigma) < x) = 0{,}9$ (если вспомнить понятия процентной точки и квантиля, то речь идет о нахождении 10%-й точки, или, что то же самое, 90%-го квантиля).

Воспользовавшись формулой (5.44), получим

$$P(N(a, \sigma) < x) = 0{,}5 + \Phi\left(\frac{x - 15{,}28}{0{,}12}\right) = 0{,}9 \text{ и } \Phi\left(\frac{x - 15{,}28}{0{,}12}\right) = 0{,}4.$$

Найдем значение аргумента $z = (x - 15{,}28)/0{,}12$ функции Ф, при котором $\Phi(z) = 0{,}4$. Обратимся к П. 1; в «столбцах $\Phi(z)$» ближайшим к 0,4 является число 0,4032, ему соответствует $z = 1{,}30$. Из уравнения $(x - 15{,}28)/0{,}12 = 1{,}30$ получаем $x = 15{,}44$, т. е. в 10% случаев цена акции превысит 15,44 ден. ед.

Найдем значение x, используя функцию НОРМОБР: НОРМОБР (0,9; 15,28; 0,12) вернет $x = 15{,}43$. ≪

5.2.4. Логарифмически нормальный закон. Случайная величина называется ***логарифмически нормальной*** ($X_{л.н}$), если ее логарифм ($\ln X_{л.н}$) — нормально распределенная величина. Из этого определения вытекает, что $X_{л.н} > 0$.

Нормальность распределения величины $\ln X_{л.н}$ гарантируется при выполнении следующих условий формирования значений величины $X_{л.н}$:

— каждое значение формируется под воздействием большого числа независимых случайных факторов W_1, $W_2, \ldots, W_N$;

— характер воздействия факторов — мультипликативный (от англ. *multiple* — умножать). Это означает, что случайный прирост, вызываемый действием каждого следующего фактора, пропорционален уже достигнутому к этому моменту значению величины; т. е. если c — неслучайная компонента величины $X_{л.н}$, а $\Delta(W_i)$, $i = 1, 2, \ldots, N$, — случайные эффекты воздействия факторов, то

$$X_{л.н} = c + c\,\Delta(W_1) + Y_1\,\Delta(W_2) + Y_2\,\Delta(W_3) + \ldots + Y_{n-1}\,\Delta(W_N);$$

(Y_1, Y_2, Y_3, Y_{n-1})

— воздействие $\Delta(W_i)$ каждого фактора мало, равновероятно по знаку и не может превалировать среди воздействий остальных факторов.

Такой механизм формирования значений величины характерен для многих конкретных социально-экономических и физических ситуаций. Примерами логарифмически нормальной величины являются:

1) заработная плата в совокупности работников;

2) среднедушевой доход в совокупности семей некоторой социальной группы;

3) долговечность изделия, эксплуатируемого в режиме износа и старения;

4) размеры частиц при дроблении.

Можно доказать, что функция плотности логарифмически нормальной величины имеет вид

$$f_{X_{\text{л.н.}}}(x) = \frac{1}{\sigma_{\ln X_{\text{л.н.}}}\sqrt{2\pi}x}\, e^{-(\ln x - \ln c)^2/(2\sigma^2_{\ln X_{\text{л.н.}}})},\ x > 0, \quad (5.49)$$

где $c = x_{\text{med}}$ — медиана логарифмически нормального распределения (чуть позже убедимся в том, что для этого распределения $\ln x_{\text{med}}$ равен математическому ожиданию случайной величины $\ln X_{\text{л.н}}$, $\ln x_{\text{med}} = M \ln X_{\text{л.н}}$), $\sigma_{\ln X_{\text{л.н.}}}$ — среднее квадратическое отклонение величины $\ln X_{\text{л.н.}}$

График функции (5.49) имеет правостороннюю асимметрию (коэффициент асимметрии $A > 0$): правая ветвь кривой, начиная от вершины, длиннее левой. При замене x на $\ln x$ график трансформируется в кривую Гаусса (см. рис. 5.5, *а*), у которой $a = M \ln X_{\text{л.н.}}$, а $\sigma = \sigma_{\ln X_{\text{л.н.}}}$.

Рассчитывая значения функции распределения $F_{X_{\text{л.н.}}}(x)$ логарифмически нормальной величины $X_{\text{л. н}}$ (напомним, что всегда $X_{\text{л. н}} > 0$), или, иначе, $P(X_{\text{л. н}} < x)$), пользуются тем, что величина $Y = \ln X_{\text{л. н}}$ имеет нормальный закон распределения:

$$F_{X_{\text{л.н.}}}(x) = P(X_{\text{л. н}} < x) \underset{(*)}{=} P(\underbrace{\ln X_{\text{л.н.}}}_{Y} < \underbrace{\ln x}_{y}) \underset{(5.44)}{=}$$

$$\underset{(5.44)}{=} 0{,}5 + \Phi\left(\frac{y - MY}{\sigma_Y}\right) = 0{,}5 + \Phi\left(\frac{\ln x - M\ln X_{\text{л.н}}}{\sigma_{\ln X_{\text{л.н}}}}\right)$$

(при переходе (*) учитывался возрастающий характер функции $\ln X$). Итак,

$$F_{X_{\text{л.н.}}}(x) = P(X_{\text{л. н}} < x) = 0{,}5 + \Phi\left(\frac{\ln x - M\ln X_{\text{л.н}}}{\sigma_{\ln X_{\text{л.н}}}}\right). \quad (5.50)$$

Тогда

$$P(x_1 < X_{\text{л. н}} < x_2) = F_{X_{\text{л.н.}}}(x_2) - F_{X_{\text{л.н.}}}(x_1) \underset{(5.50)}{=}$$

$$\underset{(5.50)}{=} \Phi\left(\frac{\ln x_2 - M\ln X_{\text{л.н}}}{\sigma_{\ln X_{\text{л.н}}}}\right) - \Phi\left(\frac{\ln x_1 - M\ln X_{\text{л.н}}}{\sigma_{\ln X_{\text{л.н}}}}\right). \quad (5.51)$$

Воспользовавшись формулой (5.50), убедимся в том, что $\ln x_{\text{med}} = M \ln X_{\text{л. н}}$, где x_{med} — медиана логарифмически нормальной величины $X_{\text{л. н}}$.

≫ Для непрерывной величины медиана определяется как корень уравнения (4.34), т. е. $F_{X_{\text{л.н}}}(x_{\text{med}}) = 0{,}5$.

Учитывая формулу (5.50), имеем

$$0{,}5 + \Phi\left(\frac{\ln x_{med} - M \ln X_{л.н.}}{\sigma_{\ln X_{л.н}}}\right) = 0{,}5.$$

Далее, поскольку функция $\Phi(z) = 0$ при $z = 0$ получаем, что $\ln x_{med} - M \ln X_{л.н} = 0$, или $\ln x_{med} = M \ln X_{л.н}$. ≪

В Microsoft Excel логарифмически нормальная модель реализована **Статистическими функциями:**

— ЛОГНОРМРАСП (x; $M \ln X_{л.н}$; $\sigma_{\ln X_{л.н.}}$), которая возвращает $P(X_{л.н} < x)$, рассчитываемую по формуле (5.50)

— ЛОГНОРМОБР (p; $M \ln X_{л.н}$; $\sigma_{\ln X_{л.н.}}$), которая возвращает такое значение x логарифмически нормальной величины $X_{л.н}$, при котором $P(X_{л.н} < x) = p$.

Замечание. Обычно при изучении случайной величины в первую очередь находят ее математическое ожидание и дисперсию (среднее квадратическое отклонение). В формулах (5.50) и (5.51) и в функциях ЛОГНОРМРАСП и ЛОГНОРМОБР фигурируют не математическое ожидание $MX_{л.н}$ и среднее квадратическое отклонение $\sigma_{X_{л.н}}$ исходной величины $X_{л.н}$, а аналогичные характеристики логарифма исходной величины: $M \ln X_{л.н}$ и $\sigma_{\ln X_{л.н}}$. Однако можно доказать, что последние выражаются через $MX_{л.н}$ и $\sigma_{X_{л.н}}$ по следующим формулам:

$$M \ln X_{л.н} = \ln \frac{(MX_{л.н})^2}{\sqrt{\sigma^2_{X_{л.н}} + (MX_{л.н.})^2}}; \qquad (5.52)$$

$$\sigma^2_{\ln X_{л.н}} = \ln \frac{\sigma^2_{X_{л.н}} + (MX_{л.н.})^2}{(MX_{л.н.})^2}. \qquad (5.53)$$

ЗАДАЧА 5.5. Месячный среднедушевой доход случайно выбранной семьи из некоторой социальной группы является логарифмически нормальной величиной с математическим ожиданием 500 ден. ед. и средним квадратическим отклонением 300 ден. ед. Найти:

а) долю семей, имеющих среднедушевой доход менее 700 ден. ед.;

б) размер среднедушевого дохода, превышение которого возможно лишь у 10% семей.

Решение. Пусть $X_{л.н}$ — месячный среднедушевой доход случайно выбранной семьи. По условию $MX_{л.н} = 500$; $\sigma_{\ln X_{л.н.}} = 300$. Используя формулы (5.52) и (5.53), получим

$$M \ln X_{л.н} = \ln (500^2 / \sqrt{300^2 + 500^2}) = 6{,}06,$$

$$\sigma_{\ln X_{л.н}} = \sqrt{\ln \frac{300^2 + 500^2}{500^2}} = 0{,}55.$$

а) $P(X_{\text{л. н}} < 700) \underset{(5.50)}{=} 0{,}5 + \Phi\left(\frac{\ln 700 - 6{,}06}{0{,}55}\right) =$

$= 0{,}5 + \Phi(0{,}89) \underset{\text{П. 1}}{=} 0{,}5 + 0{,}3159 = 0{,}8159.$

Для расчета вероятности воспользуемся функцией ЛОГНОРМРРАСП: ЛОГНОРМРАСП (700; 6,06; 0,55) вернет $P(X_{\text{л. н}} < 700) = 0{,}81404$.

б) Согласно условию, надо найти среднедушевой доход x, при котором $P(X_{\text{л. н}} > x) = 0{,}1$, или $P(X_{\text{л. н}} < x) = 0{,}9$ (т. е. речь идет о нахождении 10%-й точки, или 90%-го квантиля логарифмически нормального распределения).

По формуле (5.50) получим

$$P(X_{\text{л. н}} < x) = 0{,}5 + \Phi\left(\frac{\ln x - 6{,}06}{0{,}55}\right) = 0{,}9.$$

Отсюда $\Phi\left(\frac{\ln x - 6{,}06}{0{,}55}\right) = 0{,}4.$

Воспользуемся П. 1; найдем в «столбцах $\Phi(z)$» таблицы число, ближайшее к 0,4, — это 0,4032; ему соответствует $z = 1{,}30$. Из уравнения $(\ln x - 6{,}06)/0{,}55 = 1{,}30$ получим $\ln x = 6{,}77$, $x = 871{,}3$. Итак, у 10% семей среднедушевой доход превышает 871,3 ден. ед.

Для определения x воспользуемся функцией ЛОГНОРМОБР: ЛОГНОРМОБР (0,9; 6,06; 0,55) вернет $x = 866{,}8$. ◀

§ 5.3. Законы распределения, используемые как техническое средство при получении статистических выводов

Рассмотрим три закона, основное назначение которых — представить исследователю необходимый технический аппарат при обобщении выводов, полученных по результатам обработки ограниченного числа наблюдений, на всю мыслимую их совокупность. Эти законы связаны с распределением стандартной нормальной величины $N(0; 1)$ и называются: χ^2-распределение («хи-квадрат»-распределение), распределение Стьюдента, F-распределение.

1. χ^2-распределение. ***χ^2-распределением с k степенями свободы*** называется распределение суммы квадратов k независимых стандартных нормальных величин

$$\chi^2(k) = \sum_{i=1}^{k} N_i^2(0; 1). \qquad (5.54)$$

Сама величина (5.54) называется ***χ^2-величиной с k степенями свободы***. Обратим внимание на то, что, в силу равенства (5.54), величина $\chi^2(k) \geqslant 0$ и $k \geqslant 1$.

Число степеней свободы определяют как разность между числом суммируемых величин и числом линейных связей, ограничивающих свободу изменения этих величин. В выражении (5.54) суммируются k величин, и они независимы. Поэтому говорят, что величина χ^2, определяемая этим выражением, имеет k степеней свободы.

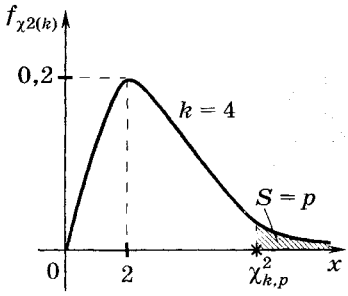


Рис. 5.8

Формула плотности $\chi^2(k)$-распределения довольно сложная и здесь не рассматривается. Приведем лишь некоторые числовые характеристики величины $\chi^2(k)$:

$$M\chi^2(k) = k,\ D\chi^2(k) = 2k,\ x_{\text{mod}} = k - 2 \text{ при } k \geqslant 2.$$

Вид графика функции плотности $\chi^2(k)$-распределения зависит от числа k. На рисунке 5.8 изображен график при $k = 4$ (в этом случае $x_{\text{mod}} = 2$).

Таблица процентных точек порядка $p = 0{,}995;\ 0{,}99;\ \ldots$ $\ldots;\ 0{,}005$ распределения $\chi^2(k)$ приведена в приложении П. 2. Она позволяет при заданных числе k степеней свободы и вероятности p найти процентную точку $\chi^2_{k,p}$, т. е. точку, для которой $P(\chi^2(k) > \chi^2_{k,p}) = p$ (рис. 5.8). Используя эту таблицу, можно, задавшись числом k и числом x в k-й строке таблицы, найти $p = P(\chi^2(k) > x)$.

С $\chi^2(k)$-распределением связана также таблица в П. 3. Ее назначение выясняется в § 8.3.

В Microsoft Excel $\chi^2(k)$-распределение реализовано **Статистическими функциями:**

— ХИ2РАСП (x; k), возвращающей при $x \geqslant 0$ вероятность $p = P(\chi^2(k) > x)$;

— ХИ2ОБР (p; k), возвращающей $p \cdot 100\%$ точку x, т. е. x, при котором $P(\chi^2(k) > x) = p$.

2. Распределение Стьюдента[1] **(T-распределение).** ***Распределением Стьюдента (или T-распределением) с k степенями свободы*** называется распределение случайной величины

$$T(k) = N(0,1)/\sqrt{\chi^2(k)/k}, \qquad (5.55)$$

где $N(0,1)$ и $\chi^2(k)$ — независимые случайные величины; сама величина (5.55) называется **T-величиной (величиной**

[1] *Стьюдент* — псевдоним английского статистика В. Госсета, исследовавшего это распределение.

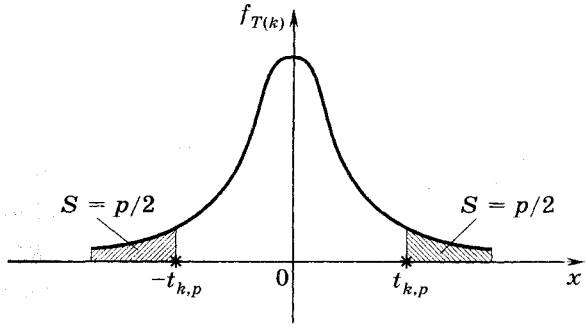


Рис. 5.9

Стьюдента) с k степенями свободы. Величина $T(k)$ задана на всей числовой оси, $k \geqslant 1$.

График функции плотности $f_{T(k)}(x)$ (кривая Стьюдента) симметричен относительно оси ординат (рис. 5.9) и напоминает кривую распределения стандартной нормальной величины $z = N(0; 1)$ (см. рис. 5.7, *а*). Однако это сходство сводится лишь к тому, что у обоих распределений математическое ожидание, так же как мода, медиана и коэффициент асимметрии, равны нулю. Другие характеристики (и, конечно, функции плотности) различны. Например, эксцесс $E_{N(0;\,1)} = 0$, а эксцесс $E_{T(k)} = 6/(k - 4) > 0$, $k > 4$. Это говорит о том, что кривая Стьюдента при $k > 4$ имеет более высокую и острую вершину по сравнению с вершиной кривой, изображенной на рисунке 5.7, *а*.

В приложении П. 4 приведена таблица, позволяющая при заданных k и «двусторонней» вероятности p (площадь каждого «хвоста» распределения равна $p/2$ (см. рис. 5.9)) найти число $t_{k,\,p}$, при котором $P(|T(k)| > t_{k,\,p}) = p$. Но тогда, в силу симметрии, $P(T(k) > t_{k,\,p}) = p/2$, т. е. $t_{k,\,p}$ — это процентная точка порядка $p/2$ распределения $T(k)$. «Односторонняя» вероятность, равная $p/2$, указана внизу таблицы П. 4. Используя эту таблицу, можно также, задавшись числом k и в k-й строке таблицы числом x, найти в ее верхней строке вероятность $p = P(|T(k)| > x)$, а в нижней — вероятность $p/2 = P(T(k) > x)$.

В Microsoft Excel $T(k)$-распределение реализовано следующими **Статистическими функциями**:

— СТЬЮДРАСП (x; k; хвосты), где $x \geqslant 0$. Если значение аргумента «хвосты» приравнять 1 (единице), то функция возвращает одностороннюю вероятность $P(T(k) > x)$; если «хвосты» приравнять 2 (двойке), то функция возвращает двухстороннюю вероятность $P(|T(k)| > x)$;

— СТЬЮДРАСПОБР (p; k) возвращает число $x \geqslant 0$, при котором $P(|T(k)| > x) = p$.

3. *F*-распределение. *F*-распределением с числами степеней свободы k_1 и k_2 называется распределение случайной величины

$$F(k_1k_2) = \frac{\chi^2(k_1)/k_1}{\chi^2(k_2)/k_2}, \qquad (5.56)$$

где $\chi^2(k_1)$ и $\chi^2(k_2)$ — независимые случайные величины; сама величина (5.58) называется ***F*-величиной с k_1 и k_2 степенями свободы.** Обратим внимание на то, что величина $F(k_1k_2) \geqslant 0$, а $k_1, k_2 \geqslant 1$.

Замечание. При $k = 1$ под знаком суммы в (5.54) одно слагаемое $N^2(0; 1)$, поэтому $\chi^2(1) = N^2(0; 1)$. Тогда

$$F(1; k) \underset{(5.56)}{=} \frac{\chi^2(1)/1}{\chi^2(k)/k} = \frac{N^2(0;1)}{\chi^2(k)/k} \underset{(5.55)}{=} T^2(k);$$

или, окончательно,

$$F(1; k) = T^2(k), \qquad (5.57)$$

т. е. случайная величина $F(1; k)$ равна квадрату величины Стьюдента с k степенями свободы.

График функции плотности F-распределения зависит от чисел k_1 и k_2. На рисунке 5.10 изображен график при $k_1 = 4$ и $k_2 = 40$. Мода $F(k_1, k_2)$-распределения при $k_1 > 1$ равна $x_{\text{mod}} = k_2(k_1 - 2)/(k_1(k_2 + 2))$; если $k_1 = 4$ и $k_2 = 40$, то $x_{\text{mod}} = 0{,}48$.

В таблице, приведенной в П. 5, даны 5% ($p = 0{,}05$) точки распределения $F(k_1, k_2)$, т. е. точки $f_{k_1; k_2; 0{,}05}$, для которых $P(F(k_1, k_2) > f_{k_1; k_2; 0{,}05}) = 0{,}05$ (рис. 5.10).

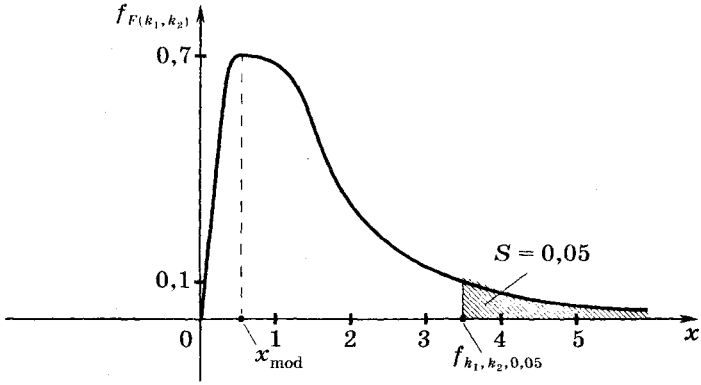


Рис. 5.10

В Microsoft Excel $F(k_1, k_2)$-распределение реализовано следующими **Статистическими функциями:**

— FРАСП (x; k_1; k_2), возвращающей при $x \geqslant 0$ вероятность $P(F(k_1, k_2) > x)$;

— FРАСПОБР (p; k_1; k_2), возвращающей при любой вероятности p $p \cdot 100\%$-ю точку распределения $F(k_1, k_2)$, т. е. число x, при котором $P(F(k_1, k_2) > x) = p$.

§ 5.4. Понятие о методе статистических испытаний. Генерация случайных чисел в Microsoft Excel

В § 5.1 и 5.2 были рассмотрены наиболее распространенные на практике модели законов распределения случайных величин.

В терминах случайных величин формулируется большинство научных и практических задач. К ним относятся задачи исследования различных систем управления со случайными воздействиями, многие производственно-технические и экономические задачи и т. д. Решить их аналитическими методами и установить формульные зависимости удается далеко не всегда. Если аналитическое решение задачи трудноосуществимо, то используют ***метод статистических испытаний.***

Основная идея метода состоит в следующем. Реальный процесс с присутствующими в нем случайностями имитируют, моделируют другим процессом, имеющим с исходным одинаковую вероятностную структуру. По результатам достаточно большого числа испытаний — многоразовых имитаций оценивают характеристики реального процесса.

Метод статистических испытаний часто называют ***методом Монте-Карло*** по названию города Монте-Карло, известного играми в рулетку, которая является одним из способов имитации случайности.

Применение метода статистических испытаний часто дает большой экономический эффект: вместо проведения дорогостоящих или трудноосуществимых натурных наблюдений реального процесса, его «проигрывают», имитируют, используя компьютерные системы программ. Процесс получения значений случайной величины с тем или иным законом распределения называют ***моделированием случайной величины.***

В Microsoft Excel для моделирования случайных величин используется программа **«Генерация случайных чисел»** пакета **«Анализ данных»**. Используя ее, можно по-

лучить последовательность значений случайной величины, имеющей следующее распределение:

— дискретное, задаваемое рядом распределения (в виде двух столбцов: левого, содержащего значения величины, и правого, в котором приведены вероятности значений);

— Бернулли с параметром p (это распределение альтернативной величины, принимающей с вероятностью p значение 1 при успешном испытании и с вероятностью $1 - p$ значение 0 при неудачном испытании);

— биномиальное с заданным числом испытаний n и параметром p (вероятность успеха в единичном испытании);

— Пуассона с параметром λ (среднее число событий простейшего потока в единицу времени), если в задаче речь идет о числе событий, наступающих в единицу времени; или с параметром λt, если речь идет о числе событий, наступающих за время t;

— равномерное на отрезке $[a, b]$, его параметры — числа a и b;

— нормальное с параметрами a (среднее значение величины) и σ (среднее квадратическое, или стандартное отклонение).

Выбрав в меню Microsoft Excel последовательно пункты **Сервис/Анализ данных/Генерация случайных чисел** и щелкнув по кнопке OK, получим диалоговое окно **«Генерация случайных чисел»**, в котором:

число переменных — число моделируемых случайных величин (с одинаковым распределением);

число случайных чисел — число значений случайной величины, которое надо получить на выходе;

распределение — осуществляет выбор одного из перечисленных выше распределений;

параметры — это параметры выбранного распределения (о них речь шла выше);

случайное рассеивание — произвольное целое число (если впоследствии вновь задать это число, то будет получена та же последовательность случайных чисел, которая была получена при первоначальном его задании);

выходной интервал — предполагает введение ссылки на левую верхнюю ячейку, начиная с которой будут размещаться случайные числа (если число переменных равно 1, то будет один столбик; при числе переменных, равном 2, — два столбика и т. д.).

Заполнив диалоговое окно и щелкнув по кнопке OK, получим последовательность (одну или несколько, что определяется числом переменных) значений случайной величины с выбранным распределением.

Покажем на примере, как используются эти последовательности при имитации реальных процессов. Рассмотрим систему управления запасами сыпучего материала, характеризующуюся следующими данными.

- Система состоит из одного склада, на котором хранится сыпучий материал одного вида.

- Ежедневный спрос q на материал является случайной величиной, имеющей нормальное распределение с математическим ожиданием (ежедневным средним спросом), равным $a = 2{,}5$ т, и средним квадратическим отклонением $\sigma = 0{,}7$ т: $q = N(2{,}5;\ 0{,}7)$.

- Пополнение запасов на складе происходит по следующему правилу: если сумма запасов i после удовлетворения спроса и объема $Q^{зак}$ заказа на пополнение складских запасов меньше некоторого критического уровня S или равна ему, то оформляется заказ на Q т материала; в противном случае заказ на пополнение складских материалов не оформляется.

- Продолжительность L исполнения заказа на пополнение запасов измеряется в днях и является случайной величиной, принимающей два равновероятных значения: 2 или 3 дня (если заказ на пополнение запасов оформляется в t-й день, то заказанный материал поступает на склад в день $t^{зак} = t + L + 1$).

- Один день функционирования системы характеризует следующая последовательность событий: вначале учитывается материал в объеме $Q^{зак}$, который в этот день должен поступить на склад (после этого $Q^{зак}$ приравнивают нулю). Затем удовлетворяется спрос на продукцию (если это возможно; неудовлетворенная часть спроса теряется); определяется количество i оставшегося на складе материала; и наконец, если $i + Q^{зак} \leqslant S$, то оформляется заказ на Q т материала ($Q^{зак}$ приравнивается Q), если $i + Q^{зак} > S$, то заказ на пополнение запасов не оформляется.

Проимитируем работу системы в течение 25 дней, начиная с первого дня ($t = 1$), приняв:

— уровень запасов на начало первого дня, равным 8 т;

— на начало первого дня заказ на пополнение запасов $Q^{зак}$ отсутствует ($Q^{зак} = 0$);

— объем заказа на пополнение запасов, если в этом возникает необходимость, $Q = 8$ т;

— критический уровень запасов на складе $S = 5$ т.

«Процесс имитации» отражен в таблице 5.1. Дадим некоторые пояснения:

— в графе (2) приведены значения случайного спроса $q = N(a = 2{,}5;\ \sigma = 0{,}7)$, взятые (с двумя знаками после запятой) с «выхода» работы программы генерации 25 случайных чисел нормального распределения;

— в графе (7) содержатся значения случайной продолжительности L исполнения заказа, взятые с «выхода» работы программы генерации случайных чисел дискретного распределения, заданного таблицей

l	2	3
$P(L = l)$	0,5	0,5

— значение продолжительности L исполнения заказа определяется только в том случае, когда возникает потребность в пополнении запасов на складе, т. е. когда $i + Q^{\text{зак}} \leqslant 5$;

— проследим работу системы, например, с 3-го по 6-й день. К концу 3-го дня $i + Q^{\text{зак}} = 3{,}12 + 0 < 5$, поэтому делаем заказ $Q^{\text{зак}} = Q = 8$ т на пополнение складских запасов; продолжительность исполнения заказа 2 дня, срок его исполнения — 6-й день (в начале 6-го дня к складским запасам добавится 8 т материала). Спрос на материал в 4-й день полностью удовлетворен и к концу дня $i + Q^{\text{зак}} = 0{,}01 + 8 > 5$ — пополнять запасы не надо. Спрос в 5-й день нельзя полностью удовлетворить; спрос равен 2,65 т, а на складе 0,01 т материала; неудовлетворенный спрос равен 2,64 т и он «забывается»; к концу этого дня $i + Q^{\text{зак}} = 0 + 8 > 5$ — пополнять запасы не надо. В начале 6-го дня на склад прибудет заказанный материал в объеме $Q^{\text{зак}} = 8$, наличные запасы составят (0 + 8) т и $Q^{\text{зак}}$ станет равным 0 и т. д.

По результатам имитации работы реальной системы можно оценить ее эффективность. Так, в рассмотренном примере объем неудовлетворенного спроса по отношению к общему спросу за 25 дней составил $14{,}7 \cdot 100/55{,}54 = 26\%$; средний объем неудовлетворенного спроса за день составил $14{,}7/25 = 0{,}6$ т; относительная частота появления среди 25 дней дня, когда спрос не удовлетворяется, $\hat{p} = 10/25 = 0{,}4$. Чтобы повысить доверие к числовым значениям показателей работы системы, следует увеличить продолжительность имитации.

Если показатель эффективности работы системы — вероятность p появления некоторого события (например, появление дня с неудовлетворенным спросом), то, используя формулу (6.22) (ее вывод дан в § 6.1), можно рассчитать продолжительность имитации — число дней n, гарантирующее с вероятностью, не меньшей $\underline{P}$, отличие полученной в результате имитации относительной частоты $\hat{p}$ появ-

ления этого события от (неизвестной) вероятности p, меньшее (по абсолютной величине) заданного числа $\varepsilon > 0$. Приняв в формуле (6.22) $\underline{P} = 0{,}85$, $\varepsilon = 0{,}1$, $c = 0{,}25$, получим

$$n \geqslant \frac{0{,}25}{0{,}1^2(1 - 0{,}85)} = 166{,}(6).$$

Промоделировав работу системы в течение 167 дней, можно с вероятностью, не меньшей 0,85, ожидать, что ошибка «использования» полученной по результатам моделирования относительной частоты $\hat{p}$ появления дня с неудовлетворенным спросом вместо (неизвестной) вероятности p появления такого дня меньше 0,1.

Если показатель эффективности работы системы — среднее значение некоторой случайной величины X, или, иначе, ее математическое ожидание MX, то необходимую продолжительность имитации можно рассчитать по формуле (6.19). Пусть, например, X — объем неудовлетворенного спроса в течение дня; по 25-дневной имитации системы средний объем неудовлетворенного спроса за день $\bar{x} = 14{,}7/25 = 0{,}6$ т. Найдем число дней n по формуле (6.19), положив $\underline{P} = 0{,}85$, $\varepsilon = 0{,}1$, $c = 1{,}1$ (c — предполагаемая верхняя граница неизвестной DX). Получим

$$n \geqslant \frac{0{,}11}{0{,}1^2(1 - 0{,}85)} = 73(3).$$

Промоделировав работу системы в течение 74 дней, можно ожидать с вероятностью, не меньшей 0,85, что ошибка «использования» полученного по результатам моделирования среднеежедневного объема неудовлетворенного спроса вместо неизвестного значения MX будет меньше 0,1.

Таблица 5.1

День (t)	Объем спроса в t-й день (q)	Наличный объем запасов на начало t-го дня	Объем неудовлетворенного спроса	Объем запасов, оставшихся после удовлетворения спроса (i)	Объем запасов после удовлетворения спроса плюс объем заказа ($i + Q^{зак}$)	Продолжительность исполнения заказа (L)	День поступления заказа на склад ($t^{зак} = t + L + 1$)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	0,38	8	0	7,62	$7{,}62 + 0 > 5$		
2	2,61	7,62	0	5,01	$5{,}01 + 0 > 5$		

Окончание табл.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
3	1,89	5,01	0	3,12	3,12 + 0 < 5	2	3 + 2 + 1 = 6
4	3,11	3,12	0	0,01	0,01 + 8 > 5		
5	2,65	0,01	2,64	0	0 + 8 > 5		
6	2,46	0 + 8	0	5,54	5,54 + 0 > 5		
7	2,23	5,54	0	3,31	3,31 + 0 < 5	2	7 + 2 + 1 = 10
8	3,38	3,31	0,07	0	0 + 8 > 5		
9	3,15	0	3,15	0	0 + 8 > 5		
10	2,96	0 + 8	0	5,04	5,04 + 0 > 5		
11	1,84	5,04	0	3,2	3,2 + 0 < 5	3	11 + 3 + 1 = 15
12	3,25	3,2	0,05	0	0 + 8 > 5		
13	2,89	0	2,89	0	0 + 8 > 5		
14	2,52	0	2,52	0	0 + 8 > 5		
15	2,14	0 + 8	0	5,86	5,86 + 0 > 5		
16	0,98	5,86	0	4,88	4,88 + 0 < 5	3	16 + 3 + 1 = 20
17	1,57	4,88	0	3,31	3,31 + 8 > 5		
18	2,26	3,31	0	1,05	1,05 + 8 > 5		
19	1,77	1,05	0,72	0	0 + 8 > 5		
20	1,82	0 + 8	0	6,18	6,18 + 0 > 5		
21	4,09	6,18	0	2,09	2,09 + 0 < 5	3	21 + 3 + 1 = 25
22	2,4	2,09	0,31	0	0 + 8 > 5		
23	1,67	0	1,67	0	0 + 8 > 5		
24	0,68	0	0,68	0	0 + 8 > 5		
25	0,84	0 + 8	0	7,16	7,16 > 5		
Σ	55,54		14,7				

УПРАЖНЕНИЯ

1. Студент сдает четыре семестровых экзамена. Вероятность успешной сдачи каждого экзамена равна 0,8. Предположив, что сдача экзаменов — независимые испытания, составьте ряд распределения случайной величины X — числа сданных студентом экзаменов; найдите ожидаемое в среднем число экзаменов, которое сдаст студент, и средний разброс числа экзаменов, которое сдаст студент, относительно среднего числа сданных экзаменов.

2. В среднем левши составляют 1% всего населения. Сколько в среднем надо опросить людей, чтобы: а) встретить левшу; б) набрать десятерых левшей?

3. Составьте ряд распределения числа незнакомцев, которых придется опросить для нахождения человека, день рождения которого совпадает с вашим, и найдите среднее число опрошенных незнакомцев.

4. В банк поступило 20 авизо, среди которых три фальшивых. Тщательной проверке (которая гарантировано выявляет фальшивые документы) подвергают пять случайно выбранных авизо. Составьте ряд распределения случайной величины X — числа фальшивых авизо среди пяти выбранных. Найдите среднее число выявленных фальшивых авизо и среднее квадратическое отклонение числа выявленных фальшивок.

5. Случайная величина X равномерно распределена на отрезке $[2; 4]$, а величина Y равномерно распределена на отрезке $[0; 5]$; X и Y независимы. Найдите: $M(XY)$, $D(2X - 3Y)$, $P((X < 3) \quad (1 < Y < 4))$.

6. Длительность междугородных телефонных разговоров распределена по показательному закону, разговор продолжается в среднем 3 мин. Найдите вероятность того, что:

а) разговор продлится более 3 мин;

б) разговор продлится от 1 мин до 2 мин; от 4 мин до 5 мин (найденным вероятностям дайте геометрическую интерпретацию, используя график функции плотности показательного закона);

в) разговор, который уже длится 2 мин, закончится в течение ближайшей минуты.

7. Докажите, что величина $T = \min(T_1, T_2)$, где T_1 и T_2 — независимые показательно распределенные случайные величины соответственно с параметрами λ_1 и λ_2, имеет показательный закон с параметром $\lambda = \lambda_1 + \lambda_2$.

8. В магазине два входа; потоки покупателей на этих входах независимы и являются простейшими, соответственно с интенсивностью $\lambda_1 = 1{,}5$ чел./мин и $\lambda_2 = 0{,}5$ чел./мин. Определите вероятность того, что:

а) в наугад выбранную минуту ни один человек не посетит магазин;

б) промежуток времени между двумя следующими друг за другом покупателями превысит 3 мин.

9. Проверка с помощью некоторого теста способностей специалистов в определенной области знаний показала, что «тестовая» оценка распределена по нормальному закону с параметрами 500 и 100.

а) Какая доля специалистов будет иметь оценку от 300 до 700?

б) Сколько специалистов из 100 будут иметь оценку, меньшую 225, большую 675?

в) Найдите 90%-й квантиль распределения. Каков его смысл в терминах задачи?

10. Статистика по вкладам населения в банк свидетельствует, что размер вклада X — логарифмически нормальная величина с медианой $x_{med} = 1200$ ден. ед. и $\sigma_{\ln X} = 200$ ден. ед. Определите долю клиентов, размер вклада которых составит не менее 1000 ден. ед.

11. Найдите 10% точки распределений $\chi^2(10)$, $T(10)$, $F(2;10)$, используя соответствующие таблицы и статистические функции Microsoft Excel.

12. Проимитируйте работу системы управления запасов, рассмотренную в § 5.4, в течение 10 дней. Какова доля неудовлетворенного спроса в общем спросе за 10 дней?

ГЛАВА 6

Закон больших чисел и центральная предельная теорема

Закон больших чисел — это общее название нескольких теорем, в каждой из которых для тех или иных условий устанавливается факт приближения среднего значения большого числа случайных величин к некоторой постоянной величине. Иначе говоря, устанавливается факт статистической устойчивости, заключающийся в предсказуемости среднего значения большого числа случайных величин почти с полной определенностью. В число этих теорем входит и теорема Я. Бернулли, формулирующая условия статистической устойчивости относительной частоты появления события. Эта теорема и теорема П. Л. Чебышёва, также входящая в закон больших чисел, рассматриваются в этой главе.

Центральная предельная теорема включает теоремы, формулирующие условия предсказуемости «в пределе» закона распределения случайной величины, или, иначе, условия статистической устойчивости закона распределения. Оказывается, что в ряде случаев таким предельным законом является нормальный. В этой главе выясняется смысл центральной предельной теоремы и, как ее следствия, приводятся теоремы Муавра—Лапласа (вытекающие из этих теорем формулы Муавра—Лапласа приведены в § 3.3); рассматриваются практические приложения центральной предельной теоремы.

§ 6.1. Неравенство Чебышёва и его приложения

Приведем неравенство Чебышёва (имеющее само по себе практический интерес), на котором базируется доказательство теоремы Чебышёва.

Для любой случайной величины X, имеющей конечную дисперсию, при любом $\varepsilon > 0$ имеет место неравенство

$$P(|X - MX| < \varepsilon) \geqslant 1 - DX/\varepsilon^2. \qquad (6.1)$$

≫ Для простоты ограничимся доказательством неравенства для дискретной случайной величины X. В выражении $DX = \sum_{i=1}^{n} (x_i - MX)^2 p_i$ со-

храним под знаком суммы те слагаемые, у которых $(x_i - MX)^2 \geqslant \varepsilon^2$ или, иначе, у которых $|x_i - MX| \geqslant \varepsilon$. Это не изменит суммы или приведет к ее уменьшению:

$$\sum_{i=1}^{n} (x_i - MX)^2 p_i \geqslant \sum_{|x_i - MX| \geqslant \varepsilon} (x_i - MX)^2 p_i.$$

Заменив затем стоящие в правой части формулы квадраты на ε^2, получим цепочку неравенств

$$DX = \sum_{i=1}^{n} (x_i - MX)^2 p_i \geqslant \sum_{|x_i - MX| \geqslant \varepsilon} (x_i - MX)^2 p_i \geqslant$$

$$\geqslant \sum_{|x_i - MX| \geqslant \varepsilon} \varepsilon^2 p_i = \varepsilon^2 \sum_{|x_i - MX| \geqslant \varepsilon} p_i = \varepsilon^2 P(|X - MX| \geqslant \varepsilon),$$

или, окончательно,

$$DX \geqslant \varepsilon^2 \cdot P(|X - MX| \geqslant \varepsilon).$$

Отсюда

$$P(|X - MX| \geqslant \varepsilon) \leqslant DX/\varepsilon^2,$$

и

$$P(|X - MX| < \varepsilon) = 1 - P(|X - MX| \geqslant \varepsilon) \geqslant 1 - DX/\varepsilon^2.$$

Таким образом,

$$P(|X - MX| < \varepsilon) \geqslant 1 - DX/\varepsilon^2. \quad \ll$$

Неравенство Чебышёва (6.1), не предполагая знания закона распределения случайной величины, указывает лишь нижнюю границу вероятности $P(|X - MX| < \varepsilon)$. При известном законе распределения (функции распределения $F_X(x)$) можно найти точное значение этой вероятности. Так, если X — непрерывна, то

$$P(|X - MX| < \varepsilon) = P(MX - \varepsilon < X < MX + \varepsilon) =$$
$$= F_X(MX + \varepsilon) - F_X(MX - \varepsilon),$$

и оно не меньше $1 - DX/\varepsilon^2$. Например, если $X = N(a, \sigma)$, то $MX = a$, и, согласно (5.47),

$$P(|N(a, \sigma) - a| < \varepsilon) = 2\Phi(\varepsilon/\sigma).$$

В частности, при $\varepsilon = 3\sigma$

$$P(|N(a, \sigma) - a| < 3\sigma) \underset{(5.48)}{=} 0{,}9972 \geqslant 1 - DX/(3\sigma)^2 =$$
$$= 1 - \sigma^2/(9\sigma^2) = 0{,}8(8).$$

Рассмотрим частный случай неравенства Чебышёва, когда в качестве случайной величины X берется среднее арифметическое

$$\frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

где

$X_1, X_2, \ldots, X_n$ — независимые случайные величины, (6.2)
дисперсии которых конечны и ограничены числом c:

$$DX_1 \leqslant c, DX_2 \leqslant c, \ldots, DX_n \leqslant c. \tag{6.3}$$

Найдем характеристики среднего арифметического.

≫ Имеем

$$M\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} MX_i; \tag{6.4}$$

$$D\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2} D\sum_{i=1}^{n} X_i \underset{(6.2)}{=} \frac{1}{n^2}\sum_{i=1}^{n} DX_i. \tag{6.5}$$

Далее, учитывая (6.3), получим

$$D\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n} DX_i \underset{(6.3)}{\leqslant} \frac{1}{n^2}\sum_{i=1}^{n} c = \frac{nc}{n^2} = \frac{c}{n},$$

т. е. *дисперсия среднего конечна* (ограничена сверху числом c/n). Поэтому для среднего имеет место неравенство Чебышёва. Заменив в (6.1) X средним, имеем

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - M\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)\Big| < \varepsilon\Big) \geqslant 1 - \frac{D\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)}{\varepsilon^2}. \tag{6.6}$$

Заменив в этом неравенстве математическое ожидание и **дисперсию** среднего их выражениями (6.4) и (6.5) и учитывая (6.3), **придем** к следующему соотношению:

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} MX_i\Big| < \varepsilon\Big) \geqslant$$

$$\geqslant 1 - \frac{\sum_{i=1}^{n} DX_i}{n^2\varepsilon^2} \geqslant 1 - \frac{c}{n\varepsilon^2}, (DX_i \leqslant c). \quad \ll \tag{6.7}$$

Рассмотрим две часто используемые на практике математические модели, приводящие к неравенству Чебышёва.

Модель I. Пусть $X_1, X_2, \ldots, X_n$ — результаты наблюдений случайной величины X с ограниченной дисперсией $(DX \leqslant c)$.

Замечание. В зависимости от контекста различают две интерпретации результатов наблюдений случайной величины X. Это:

1) конкретные числа, их обозначают следующими строчными буквами: $x_1, x_2, \ldots, x_n$;

2) обозначения тех значений, зависящих от случая, которые могли бы быть получены при наблюдениях. В этом случае для обозначения результатов наблюдений используют прописные буквы $X_1, X_2, \ldots, X_n$, подчеркивая тем самым случайность результатов.

В этой главе используется вторая интерпретация.

Предположим, что:

наблюдения независимы — это означает, что результаты наблюдений $X_1, X_2, \ldots, X_n$ — независимые случайные величины; (6.8)

наблюдения проводятся в одинаковых вероятностных условиях, или, короче, *в типичных условиях* — это означает, что каждый из результатов наблюдений имеет такой же закон распределения, что и наблюдаемая величина X, т. е. (6.9)

$$F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x) = F_X(x),$$

и, как следствие этого (и того, что $DX \leqslant c$),

$$MX_1 = MX_2 = \ldots = MX_n = MX; \qquad (6.10)$$

$$DX_1 = DX_2 = \ldots = DX_n = DX \leqslant c. \qquad (6.11)$$

Замечание. Результаты наблюдений $X_1, X_2, \ldots, X_n$ можно трактовать как значения случайной величины, или признака, X у n объектов (*случайно*) отобранных из совокупности N объектов (n — объем выборки, N — объем генеральной совокупности). Выполнение условий (6.8) и (6.9) гарантируется при случайной выборке с возвращением; при выборке без возвращения нарушается условие (6.8) (см. пример 3.1). Однако этим нарушением можно пренебречь, если объем N генеральной совокупности достаточно большой ($N \to \infty$), а объем выборки n мал (по отношению к N). Поэтому в дальнейшем, используя понятие выборки, будем иметь в виду выборку с возвращением или выборку без возвращения, но из бесконечно большой генеральной совокупности. Формулируя задачу в терминах выборки, будем называть: $\frac{1}{n}\sum_{i=1}^{n} X_i$ — ***выборочным средним***, а MX — ***генеральным средним*** случайной величины, или признака, X.

Найдем в условиях рассматриваемой модели характеристики среднего:

$$M\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} MX_i \underset{(6.10)}{=} \frac{1}{n}\sum_{i=1}^{n} MX = \frac{nMX}{n} = MX;$$

$$D\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2} D\sum_{i=1}^{n} X_i \underset{(6.8)}{=} \frac{1}{n^2}\sum_{i=1}^{n} DX_i \underset{(6.11)}{=}$$

$$\underset{(6.11)}{=} \frac{1}{n^2}\sum_{i=1}^{n} DX = \frac{nDX}{n^2} = \frac{DX}{n},$$

или, окончательно,

$$M\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = MX; \qquad (6.12)$$

$$D\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{DX}{n},\ \sigma_{\left(\sum_{i=1}^{n} X_i\right)}/n = \frac{\sigma_X}{\sqrt{n}}. \qquad (6.13)$$

В рассматриваемой модели можно использовать неравенство Чебышёва (6.6) для среднего, имеющее место при выполнении условий (6.2) и (6.3), поскольку здесь эти условия также выполняются (см. (6.8) и (6.11)).

Заменив в (6.6) математическое ожидание и дисперсию среднего их выражениями (6.12) и (6.13) и учитывая, что $DX \leqslant C$, получим

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon\right) \geqslant 1 - \frac{DX}{n\varepsilon^2} \geqslant$$

$$\geqslant 1 - \frac{c}{n\varepsilon^2}, \quad (DX \leqslant c). \qquad (6.14)$$

Поясним смысл равенств (6.12) и (6.13) и неравенства Чебышёва (6.14) для среднего арифметического. Напомним варианты интерпретации математического ожидания случайной величины: это среднее значение случайной величины, или генеральное среднее, или постоянная неслучайная компонента случайной величины, т. е. как бы ее истинное значение в идеализированной схеме, когда устранено влияние всех случайных факторов. Используя любой из этих вариантов, будем иметь в виду, что его можно заменить любым другим.

Из равенств (6.12) и (6.13) следует:

— «истинное» значение среднего арифметического результатов наблюдений величины X равно MX — истинному значению этой величины;

— среднее квадратическое отклонение среднего арифметического результатов n наблюдений величины X в $\sqrt{n}$ раз меньше σ_X, или, иначе, «средний разброс» значений среднего арифметического результатов n наблюдений величины X вокруг MX (именно вокруг MX, см. (6.12)) в $\sqrt{n}$ раз меньше «среднего разброса» значений величины X вокруг MX. Но если учесть, что $MX_i = MX$, а $\sigma_{X_i} = \sigma_X$, $i = 1, 2, \ldots, n$ (см. (6.10) и (6.11)), то можно сделать следующее заключение: «средний разброс» значений среднего арифметического результатов n наблюдений величины X вокруг MX в $\sqrt{n}$ раз меньше σ_{X_i} — «среднего разброса» значений каждого отдельного наблюдения вокруг MX, т. е. среднее результатов n наблюдений величины X, или выборочное среднее, как бы в $\sqrt{n}$ раз ближе к MX — генеральному среднему, чем результат каждого отдельного наблюдения.

Неравенство (6.14) устанавливает нижнюю границу вероятности, с которой можно ожидать, что при замене (не-

известного) генерального среднего (MX) выборочным средним $\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$ погрешность меньше заданного числа $\varepsilon > 0$.

Модель II (частный случай модели I). Пусть $X_{Al,1}$, $X_{Al,2}$, ..., $X_{Al,n}$ — результаты наблюдений альтернативной случайной величины X_{Al}:

x	0	1
$P(X_{Al} = x)$	$1 - p$	p

$$MX_{Al} = p, \qquad DX_{Al} = p(1 - p).$$

Убедимся в том, что DX_{Al} ограничена сверху.

≫ Рассмотрим функцию $f(p) = p(1 - p)$ на отрезке $[0; 1]$ и найдем ее наибольшее значение: $f'(p) = 1 - 2p$; $p_{крит} = 0{,}5$; $f(p_{крит}) = 0{,}25$, что больше значений $f(0) = 0$ и $f(1) = 0$ функций на концах отрезка $[0; 1]$. Следовательно, $f(p) \leqslant 0{,}25$, или $DX_{Al} \leqslant 0{,}25$. ≪

Как и в модели I, будем предполагать, что наблюдения величины X_{Al} независимы и проводятся в одинаковых вероятностных или в типичных условиях, т. е. выполняются условия (6.8) и (6.9) модели I. Поэтому имеет место неравенство (6.14), которое принимает вид

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{Al,i} - MX_{Al}\right| < \varepsilon\right) \geqslant 1 - \frac{DX_{Al}}{n\varepsilon^2} \geqslant$$

$$\geqslant 1 - \frac{c}{n\varepsilon^2}, \qquad DX_{Al} \leqslant c \leqslant 0{,}25. \qquad (6.15)$$

Если событие, состоящее в том, что величина $X_{Al,i}$ примет значение 1, трактовать как «успешное наблюдение», а событие, состоящее в том, что $X_{Al,i}$ примет значение 0, как «неудачное», то нетрудно понять, что $\sum_{i=1}^{n} X_{Al,i}$ — это число «успехов» в n наблюдениях. Но поскольку наблюдения независимы и вероятность успешности любого из n наблюдений равна p ($P(X_{Al,1} = 1) = P(X_{Al,2} = 1) = \ldots = P(X_{Al,n} = 1) = P(X_{Al} = 1) = p$ — это следует из того, что наблюдения проводятся в типичных условиях), эти n наблюдений являются испытаниями Бернулли. Число успехов в n испытаниях Бернулли мы обозначали буквой m; используем и здесь это обозначение: $\sum_{i=1}^{n} X_{Al,i} = m$ (здесь m — не конкретное число успехов в n испытаниях Бернулли, а случайное число успехов, т. е., по сути, здесь m — это бино-

миальная случайная величина X_{Bi}). Заменив в (6.15) $\sum_{i=1}^{n} X_{Al,i}$ на m, а MX_{Al} и DX_{Al} — их значениями $MX_{Al} = p$ и $DX_{Al} = p(1-p)$, получим

$$P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) \geqslant 1 - \frac{p(1-p)}{n\varepsilon^2} \geqslant 1 - \frac{c}{n\varepsilon^2}, \qquad (6.16)$$

$$p(1-p) \leqslant c \leqslant 0{,}25.$$

Из проведенных рассуждений следует, что условия модели II и вытекающее из них неравенство (6.16) можно переформулировать так: если m — *случайное число успехов в n испытаниях Бернулли, то при любом* $\varepsilon > 0$ *имеет место неравенство* (6.16).

Неравенство (6.16) устанавливает нижнюю границу вероятности, с которой можно ожидать, что при замене (неизвестной) вероятности p успеха в единичном испытании Бернулли относительной частотой m/n появления успеха в n испытаниях Бернулли, погрешность будет меньше заданного числа $\varepsilon > 0$.

Итак, модель II (неравенство (6.16)) является частным случаем *модели* I (неравенство (6.14)).

Рассмотрим три типа задач, которые можно решать в условиях модели I, используя неравенство (6.14), и приведем их решения.

1. Найти нижнюю границу $\underline{P} > 0$ вероятности того, что $\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon$, если известны $\varepsilon > 0$, n и число $c \geqslant DX$.

Р е ш е н и е. Из формулы (6.14) получаем

$$\underline{P} = 1 - c/(n\varepsilon^2), \qquad c \geqslant DX. \qquad (6.17)$$

2. Найти число $\varepsilon > 0$ такое, при котором $P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon\right) \geqslant \underline{P}$, если известны $\underline{P} > 0$, n и число $c \geqslant DX$.

Р е ш е н и е. Полагая в (6.14) $1 - c/(n\varepsilon^2) \geqslant \underline{P}$, получаем

$$\varepsilon \geqslant \sqrt{\frac{c}{n(1-\underline{P})}}, \qquad c \geqslant DX. \qquad (6.18)$$

3. Найти n, если известны $\varepsilon > 0$, нижняя граница $\underline{P} > 0$ вероятности того, что $\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon$, и число $c \geqslant DX$.

Р е ш е н и е. Из формулы (6.18) получаем

$$n \geqslant \frac{c}{\varepsilon^2(1-\underline{P})}, \qquad c \geqslant DX. \qquad (6.19)$$

В условиях *модели* II, используя неравенство (6.16), можно решать следующие аналогичные типы задач.

1. Найти нижнюю границу $\underline{P} > 0$ вероятности того, что $|m/n - p| < \varepsilon$, если известны $\varepsilon > 0$, n и число c такое, что $p(1-p) \leqslant c \leqslant 0{,}25$.

Р е ш е н и е. Из формулы (6.16) получаем

$$\underline{P} = 1 - c/(n\varepsilon^2), \qquad p(1-p) \leqslant c \leqslant 0{,}25. \tag{6.20}$$

2. Найти число $\varepsilon > 0$ такое, при котором $P(|m/n - p| < \varepsilon) \geqslant \underline{P}$, если известны $\underline{P} > 0$, n и число c такое, что $p(1-p) \leqslant c \leqslant 0{,}25$.

Р е ш е н и е. Полагая в (6.16) $1 - c/(n\varepsilon^2) \geqslant \underline{P}$, получаем

$$\varepsilon \geqslant \sqrt{\frac{c}{n(1-\underline{P})}}, \qquad p(1-p) \leqslant c \leqslant 0{,}25. \tag{6.21}$$

3. Найти n, если известны $\varepsilon > 0$, нижняя граница $\underline{P} > 0$ вероятности того, что $|m/n - p| < \varepsilon$, и число c такое, что $p(1-p) \leqslant c \leqslant 0{,}25$.

Р е ш е н и е. Из формулы (6.21) получаем

$$n \geqslant \frac{c}{\varepsilon^2(1-\underline{P})}, \qquad p(1-p) \leqslant c \leqslant 0{,}25. \tag{6.22}$$

ЗАДАЧА 6.1. Определите отклонение доли мальчиков среди 300 новорожденных от вероятности рождения мальчика, которое можно гарантировать с вероятностью, не меньшей 0,95. Статистическая вероятность рождения мальчика равна 0,515.

Р е ш е н и е. Введем следующие обозначения: m/n — доля мальчиков среди $n = 300$ новорожденных; p — вероятность рождения мальчика (ее истинное значение неизвестно, известна лишь опытная, статистическая вероятность $\hat{p} = 0{,}515$). Требуется найти $\varepsilon > 0$, при котором $P(|m/n - p| < \varepsilon) \geqslant 0{,}95$. Заключаем, что нижняя граница вероятности равна $\underline{P} = 0{,}95$. Воспользуемся формулой (6.21) в следующих двух вариантах:

1) полагая $c = \hat{p}(1-\hat{p})$, имеем

$$\varepsilon \geqslant \sqrt{\frac{0{,}515(1-0{,}515)}{300(1-0{,}95)}} = 0{,}12904;$$

2) полагая $c = 0{,}25$ (так поступают при отсутствии информации о приблизительном значении вероятности p), находим $\varepsilon \geqslant 0{,}1291$. Как и следовало ожидать, отсутствие информации приводит к увеличению нижней границы для ε.

ЗАДАЧА 6.2. Выборочным путем требуется оценить средний вес зерна пшеницы во всей партии. Сколько нужно обследовать зерен, чтобы с вероятностью, не меньшей 0,9, можно было ожидать, что средний вес отобранных зерен будет отличаться от среднего веса зерна во всей партии (по модулю) менее чем на 0,001 г? Установлено, что среднее квадратическое отклонение веса зерна не превышает 0,04 г.

Р е ш е н и е. Введем следующие обозначения: X — вес случайно отобранного зерна; MX — средний вес зерна во всей партии; n — число отобранных зерен; $X_1, X_2, \ldots, X_n$ — вес отобранных зерен. Требуется найти n, при котором

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < 0{,}001\right) \geqslant 0{,}9.$$

Заключаем, что $\varepsilon = 0{,}001$, $\underline{P} = 0{,}9$; по условию также известно, что $\sigma_X \leqslant 0{,}04$, или $DX \leqslant 0{,}04^2 = 0{,}0016$. Воспользуемся формулой (6.19), полагая $c = 0{,}0016$. Получаем

$$n \geqslant \frac{0{,}0016}{0{,}001^2 \cdot (1 - 0{,}9)} = 16\,000.$$

ЗАДАЧА 6.3. Сколько раз следует подбросить монету, чтобы относительная частота появления «герба» отличалась от вероятности его появления (по модулю) менее чем на 0,01, и этот результат гарантировался бы с надежностью, не меньшей 95%? не меньшей 80%?

Р е ш е н и е. Пусть m/n — относительная частота появления «герба» в n подбрасываниях монеты; $p = 0{,}5$ — вероятность появления «герба» в единичном подбрасывании. По условию $\varepsilon = 0{,}01$, $\underline{P} = 0{,}95$. Воспользуемся формулой (6.22), считая $c = 0{,}25$. Получаем

$$n \geqslant \frac{0{,}25}{0{,}01^2(1 - 0{,}95)} = 50\,000.$$

Если надежность уменьшить ($\underline{P} = 0{,}8$), то и n уменьшится ($n \geqslant 12\,500$). ◀

§ 6.2. Теорема Чебышёва. Теорема Бернулли

Теорема Чебышёва. *Если $X_1, X_2, \ldots, X_n, \ldots$ — независимые случайные величины, имеющие конечные дисперсии, ограниченные одной и той же постоянной, то среднее арифметическое этих величин сходится по вероятности к среднему арифметическому их математических ожиданий, т. е. при любом $\varepsilon > 0$*

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} MX_i\right| < \varepsilon\right) = 1. \qquad (6.23)$$

≫ Для независимых случайных величин с конечными дисперсиями, ограниченными одной и той же постоянной c, имеет место неравенство (6.7). Переходя в неравенстве (6.7) к пределу при $n \to \infty$, получаем

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} MX_i\right| < \varepsilon\right) \geqslant \lim_{n\to\infty}\left(1 - \frac{c}{n\varepsilon^2}\right) = 1,$$

а так как вероятность P не может быть больше единицы, то отсюда и следует утверждение теоремы. ≪

Рассмотрим частные случаи теоремы Чебышёва в условиях моделей, приведенных в § 6.2.

Модель I. Пусть $X_1, X_2, \ldots, X_n, \ldots$ — последовательность независимых наблюдений случайной величины X с ограниченной дисперсией ($DX \leqslant c$), проводимых в типичных условиях, т. е. выполняются требования (6.8) и (6.9). В этом случае имеет место неравенство (6.14). Переходя в неравенстве (6.14) к пределу при $n \to \infty$, получаем

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon\right) \geqslant \lim_{n\to\infty}\left(1 - \frac{c}{n\varepsilon^2}\right) = 1,$$

а так как вероятность P не может быть больше единицы, то

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon\right) = 1, \qquad (6.24)$$

т. е. при выполнении условий (6.8) и (6.9) среднее арифметическое результатов n наблюдений величины X (выборочное среднее) *сходится по вероятности к* MX (генеральному среднему).

Замечание. Нередко из формулы (6.24) делают необоснованный вывод, что выборочное среднее при увеличении n сходится к MX, т. е. что

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = MX.$$

На самом деле, в (6.24) речь идет о *«сходимости по вероятности»*, т. е. о том, что при любом сколь угодно малом $\varepsilon > 0$ наступает такой «момент» n_0, начиная с которого (т. е. при всех $n \geqslant n_0$) справедливо неравенство

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon\right) > 1 - \delta,$$

где $\delta > 0$ — сколь угодно малое число.

Суть закона больших чисел в рамках предельного равенства (6.24) такова: при увеличении числа n независимых наблюдений величины X (при увеличении объема выборки), проводимых в типичных условиях, возрастает уверенность (вероятность $P \to 1$) незначительного отклонения

(число $\varepsilon > 0$ может быть сколь угодно малым!) среднего арифметического результатов наблюдений (выборочного среднего) от MX (генерального среднего, или «истинного» значения величины X). А так как MX — постоянная величина, то получаем, что случайное значение среднего $\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$ при большом n предсказуемо: оно сколь угодно близко к постоянному числу, и это утверждение имеет вероятность, близкую к единице. Таким образом, *среднее обладает свойством статистической устойчивости.*

Модель II. Напомним, что в рамках этой модели получен следующий результат: если m — случайное число успехов в n испытаниях Бернулли, то при любом $\varepsilon > 0$ имеет место неравенство (6.16). Перейдя в неравенстве (6.16) к пределу при $n \to \infty$, получаем

$$\lim_{n\to\infty} P(|m/n - p| < \varepsilon) \geqslant \lim_{n\to\infty}\left(1 - \frac{c}{n\varepsilon^2}\right) = 1.$$

Так как вероятность P не может быть больше единицы, то отсюда имеем

$$\lim_{n\to\infty} P(|m/n - p| < \varepsilon) = 1. \qquad (6.25)$$

Предельное равенство (6.25), доказанное с использованием неравенства Чебышёва (6.16) для относительной частоты, является выводом теоремы Я. Бернулли, сформулированной и доказанной в начале XVIII в.

Теорема Бернулли. *Если m — число успешных испытаний (число появлений некоторого события A) в n независимых испытаниях и p — вероятность успешности любого из этих испытаний, то относительная частота $\hat{p} = m/n$ появления успешного испытания сходится по вероятности к вероятности p, т. е. при любом $\varepsilon > 0$ имеет место предельное равенство* (6.25).

Отметим, что из (6.25) не следует, что $\lim\limits_{n\to\infty} \frac{m}{n} = p$. Суть равенства (6.25) состоит в следующем: при увеличении числа n независимых испытаний, проводимых в типичных условиях, возрастает уверенность (вероятность $P \to 1$) в незначительном отклонении (ведь число $\varepsilon > 0$ может быть сколь угодно малым!) относительной частоты m/n появления успеха в этих испытаниях от вероятности p успешности любого из испытаний. А так как p — постоянное число, получаем, что случайное значение относительной частоты $\frac{m}{n}$ при большом n предсказуемо: оно сколь угодно

близко к постоянному числу, и это утверждение имеет вероятность, близкую к единице. Таким образом, *относительная частота обладает свойством статистической устойчивости.*

§ 6.3. Центральная предельная теорема

Смысл результатов, полученных в § 6.2, состоит в том, что при увеличении числа n случайных слагаемых их среднее арифметическое как бы утрачивает случайный характер и при $n \to \infty$ «случайность» практически исчезает. Однако при любом конечном числе n слагаемых случайный разброс среднего арифметического этих слагаемых остается. Ответ на вопрос о характере этого разброса (при $n \to \infty$) дает *центральная предельная теорема.*

Существует несколько ее вариантов, различающихся условиями, выполнение которых гарантирует нормальность распределения суммы достаточно большого числа случайных слагаемых. Поясним смысл центральной предельной теоремы.

≫ Пусть $X_1, X_2, \ldots, X_n, \ldots$ — *последовательность независимых случайных величин, имеющих конечные математические ожидания и дисперсии.* Так как случайные величины независимы, то $D\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} DX_i$. Стандартизируем случайную величину, равную $\sum_{i=1}^{n} X_i$:

$$Z_{(n)} = \frac{\sum_{i=1}^{n} X_i - M\sum_{i=1}^{n} X_i}{\sqrt{D\sum_{i=1}^{n} X_i}} = \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} MX_i}{\sqrt{\sum_{i=1}^{n} DX_i}} =$$

$$= \sum_{i=1}^{n} (X_i - MX_i)/\sqrt{\sum_{i=1}^{n} DX_i}. \qquad (6.26)$$

Напомним, что математическое ожидание и дисперсия стандартизованной величины равны соответственно 0 и 1: $MZ_{(n)} = 0$, $DZ_{(n)} = 1$, $\sigma_{Z_{(n)}} = 1$.

Различные варианты центральной предельной теоремы формулируют условия (в дополнение к названным выше), соблюдение которых гарантирует при $n \to \infty$ приближение распределения величины

$$Z_{(n)} = \sum_{i=1}^{n} \frac{X_i - MX_i}{\sqrt{\sum_{i=1}^{n} DX_i}}$$

к стандартному нормальному, или, иначе, формулируют условия, соблюдение которых позволяет при *большом n* считать, что

$$\sum_{i=1}^{n} \frac{X_i - MX_i}{\sqrt{\sum_{i=1}^{n} DX_i}} \approx N(0;\ 1). \qquad (6.27)$$

Качественная суть этих условий такова: каждое из слагаемых $(X_i - MX_i)/\sqrt{\sum_{i=1}^{n} DX_i}$ должно быть *малым, равновероятным по знаку и не должно превалировать среди остальных* (напомним, впервые эти условия введены в разделе 5.2.3 при описании механизма формирования значений нормально распределенной величины). Строгая математическая формализация условий дана в исследованиях А. А. Маркова и А. Н. Ляпунова. ≪

Сумма, стоящая в левой части равенства (6.27), — это результат стандартизации величины, равной $\sum_{i=1}^{n} X_i$. А если стандартизованная величина подчиняется нормальному закону, то и исходная величина $\sum_{i=1}^{n} X_i$ имеет нормальное распределение (см. раздел 5.2.3). Поэтому из равенства (6.27), верного при *большом n*, следует, что при *большом n*

$$\sum_{i=1}^{n} X_i \approx N\Big(M \sum_{i=1}^{n} X_i,\ \sqrt{D \sum_{i=1}^{n} X_i}\Big),$$

или

$$\sum_{i=1}^{n} X_i \approx N\Big(\sum_{i=1}^{n} MX_i,\ \sqrt{\sum_{i=1}^{n} DX_i}\Big). \qquad (6.28)$$

Замечания. 1. В рассмотренном варианте теоремы тип величин $X_1, X_2, \ldots, X_n, \ldots$ не оговаривается: они могут быть как непрерывными, так и дискретными. Если они дискретные, то результат (6.28) любопытен: сумма большого числа дискретных величин ведет себя как нормально распределенная величина, которая, как известно, является непрерывной.

2. Можно доказать, что если каждая из следующих независимых величин: $X_1, X_2, \ldots, X_n, \ldots$ имеет нормальный закон распределения, т. е. $X_i = N(MX_i, \sqrt{DX_i})$, $i = 1, 2, \ldots, n, \ldots$, то сумма этих величин имеет (при *любом n*) нормальный закон распределения с математическим ожиданием и дисперсией, равными соответственно сумме математических ожиданий и сумме дисперсий величин:

$$\sum_{i=1}^{n} X_i \underset{X_i = N(MX_i,\ \sqrt{DX_i})}{=} N\Big(\sum_{i=1}^{n} MX_i,\ \sqrt{\sum_{i=1}^{n} DX_i}\Big). \qquad (6.29)$$

Сформулированное утверждение отношения к центральной предельной теореме не имеет: равенство (6.29) — строгое, а (6.28) — приблизительное, имеющее место лишь при больших n. Однако практика показывает, что если число складываемых независимых случайных величин порядка десяти (иногда и меньше), то погрешность приблизительного равенства (6.28) невелика.

» ЗАДАЧА 6.4. Игральную кость подбрасывают 1000 раз. Найти пределы, в которых с вероятностью 0,95 лежит сумма выпавших очков.

Р е ш е н и е. Пусть X_i — число выпавших очков при i-м подбрасывании кости. Нетрудно убедиться в том, что $MX_i = 21/6$, а $DX_i = 105/36$. Так как число испытаний велико, то, согласно (6.28), сумма очков, выпавших при 1000 подбрасываниях,

$$\sum_{i=1}^{1000} X_i \approx N\Big(\sum_{i=1}^{1000} MX_i = 3500;\ \sqrt{\sum_{i=1}^{1000} DX_i} = 54{,}0\Big) = N(3500;\ 54{,}0).$$

Предположим, что 95% пределы суммы выпавших очков симметричны относительно ее математического ожидания, равного 3500. В этом случае задача сведется к нахождению числа $\varepsilon > 0$ такого, при котором

$$P(3500 - \varepsilon < \sum_{i=1}^{1000} X_i < 3500 + \varepsilon) = P\Big(\Big|\sum_{i=1}^{1000} X_i - 3500\Big| < \varepsilon\Big) \approx$$

$$\approx P\Big(\big|N(3500;\ 54{,}0) - 3500\big| < \varepsilon\Big) \underset{(5.47)}{=} 2\Phi(\varepsilon/54{,}0) = 0{,}95,$$

или $\Phi(\varepsilon/54) = 0{,}475$. В приложении П. 1 найдем число z, при котором $\Phi(z) = 0{,}475$; $z = 1{,}95$. Тогда $\varepsilon = 1{,}95 \cdot 54{,}0 = 105{,}3$ и пределы таковы: $3500 - \varepsilon = 3394{,}7$ и $3500 + \varepsilon = 3605{,}3$, а $P(3394{,}7 < \sum_{i=1}^{1000} X_i < 3605{,}3) = 0{,}95$. Если учесть, что сумма выпавших очков — целое число, то $P(3394 < \sum_{i=1}^{1000} X_i < 3606) \geqslant 0{,}95$. «

§ 6.4. Интегральная и локальная теоремы Муавра—Лапласа

Рассмотрим частный случай приблизительного равенства (6.28), когда в роли независимых случайных величин X_i выступают независимые альтернативные величины $X_{Al,i}$:

x	0	1
$P(X_{Al,i} = x)$	$1 - p$	p

$$MX_{Al,i} = p,\ DX_{Al,i} = p(1-p),$$

$\sum_{i=1}^{n} X_{Al,i} = m$ — число успехов в n испытаниях Бернулли (последнее равенство обосновано ранее). Произведя соответствующие замены в (6.28), получим при *большом n*

$$m \approx N\Big(\sum_{i=1}^{n} p = np,\ \sqrt{\sum_{i=1}^{n} p(1-p)} = \sqrt{np(1-p)}\Big),$$

или

$$m \underset{n\text{ велико}}{\approx} N(np,\ \sqrt{np(1-p)}), \qquad (6.30)$$

т. е. случайное число успехов m в большом числе n испытаний Бернулли ведет себя приблизительно как нормально распределенная величина с математическим ожиданием, равным np, и средним квадратическим отклонением, равным $\sqrt{np(1-p)}$. Отсюда при *большом n*

$$P_n(m_1 < m < m_2) \approx P(m_1 < N(np,\ \sqrt{np(1-p)} < m_2) \underset{(5.46)}{=} \Phi\Big(z_2 = \frac{m_2 - np}{\sqrt{np(1-p)}}\Big) - \Phi\Big(z_1 = \frac{m_1 - np}{\sqrt{np(1-p)}}\Big) \underset{(5.38)}{=}$$

$$\underset{(5.38)}{=} \frac{1}{\sqrt{2\pi}} \int_0^{z_2} e^{-t^2/2}\,dt - \frac{1}{\sqrt{2\pi}} \int_0^{z_1} e^{-t^2/2}\,dt = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-t^2/2}\,dt.$$

Полученный результат совпадает с интегральной формулой Муавра—Лапласа (см. (3.16) и (3.19)) и составляет содержание следующей теоремы.

Интегральная теорема Муавра—Лапласа. *При больших числах n независимых испытаний с одинаковой вероятностью p (0 < p < 1) появления успеха в любом из них для вероятности того, что число успехов m заключено в интервале (m_1, m_2), имеет место приблизительное равенство (оно тем точнее, чем больше n):*

$$P_n(m_1 < m < m_2) \approx \frac{1}{2\pi} \int_{z_1}^{z_2} e^{-t^2/2}\,dt, \qquad (6.31)$$

где $z_1 = (m_1 - np)/\sqrt{npq}$, $z_2 = (m_2 - np)/\sqrt{npq}$, $q = 1 - p$.

Вычислим вероятность того, что в большом числе n испытаний число успешных испытаний равно k. Взяв $z_1 = (k - np)/\sqrt{npq}$ и $z_2 = (k + 1 - np)/\sqrt{npq}$, из (6.31) имеем

$$P_n(k) = P(k \leqslant m < k+1) \approx \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-t^2/2}\,dt \underset{(*)}{\approx}$$

$$\underset{(*)}{\approx} \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2}(z_2 - z_1) = \frac{1}{\sqrt{2\pi}} e^{-(k-np)^2/(2npq)} \frac{1}{\sqrt{npq}}.$$

При переходе (*) определенный интеграл приравнен площади прямоугольника, опирающегося на отрезок $[z_1, z_2]$ и имеющего высоту, равную $e^{-z_1^2/2}$. Длина отрезка $[z_1, z_2]$ равна $1/\sqrt{npq}$ и мала при больших n, поэтому погрешность такой замены при больших n мала.

Итак, *при больших числах n независимых испытаний с одинаковой вероятностью p* $(0 < p < 1)$ *появления успеха в любом из них для вероятности того, что число успехов будет равно m, имеет место приблизительное равенство* (*оно тем точнее, чем больше n*):

$$P_n(m) \approx \frac{1}{\sqrt{2\pi}\sqrt{npq}} e^{-(m-np)^2/(2npq)}. \qquad (6.32)$$

Это утверждение составляет содержание **локальной теоремы Муавра—Лапласа**, а формула (6.32), совпадающая с формулой (3.13), называется локальной формулой Муавра—Лапласа.

Примеры задач, при решении которых используют интегральную и локальную формулы Муавра—Лапласа, были приведены в § 3.3.

§ 6.5. О распределении среднего арифметического и относительной частоты

О распределении среднего арифметического *n* случайных величин. Вернемся к рассмотрению приблизительного равенства (6.28), имеющего место при выполнении достаточно общих условий центральной предельной теоремы, предъявляемых к независимым случайным величинам X_1, X_2, ..., X_n, ... (с произвольным законом распределения), и при *больших n*. Из (6.28) следует, что

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx N\Big(\frac{1}{n}\sum_{i=1}^{n} MX_i, \sqrt{\sum_{i=1}^{n} DX_i/n^2}\Big). \qquad (6.33)$$

Замечание. Для независимых и нормально распределенных величин X_1, X_2, ..., X_n, ... при любом n имеет место вытекающее из (6.29) строгое равенство

$$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{X_i=N(MX_i,\sqrt{DX_i})}{=} N\Big(\frac{1}{n}\sum_{i=1}^{n} MX_i, \sqrt{\sum_{i=1}^{n} DX_i/n^2}\Big). \qquad (6.34)$$

В условиях **модели I**, рассмотренной в разделе 6.2 (где $X_1, X_2, \ldots, X_n, \ldots$ — результаты независимых наблюдений случайной величины X с ограниченной дисперсией и про-

извольным законом распределения, проведенные в типичных условиях), приближенное равенство (6.33), учитывая (6.10) и (6.11), можно записать в виде

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx N(MX,\ \sqrt{DX/n}), \tag{6.35}$$

т. е. для выборочного среднего $\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$ при *больших* n имеют место (приближенно) все формулы нормального закона и, в частности, формула (5.47), которая с учетом (6.35), принимает вид

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon\right) \approx 2\Phi\left(\frac{\varepsilon}{\sqrt{DX/n}}\right) \geqslant$$

$$\geqslant 2\Phi\left(\frac{\varepsilon}{\sqrt{c/n}}\right), \quad DX \leqslant c. \tag{6.36}$$

Замечание. Если же X — нормально распределенная величина, то из (6.34) получаем, что при любом n имеет место строгое равенство

$$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{X=N(MX,\ \sqrt{DX_i})}{=} N(MX,\ \sqrt{DX/n}),$$

тогда

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon\right) = 2\Phi\left(\frac{\varepsilon}{\sqrt{DX/n}}\right) \geqslant$$

$$\geqslant 2\Phi\left(\frac{\varepsilon}{\sqrt{c/n}}\right), \quad DX \leqslant c. \tag{6.37}$$

Приведем три типа задач, которые можно решать, используя формулу (6.36) [формулу (6.37)]. Они аналогичны типам задач, рассмотренным в § 6.1, которые были решены в условиях модели I с использованием неравенства (6.14). Следует иметь в виду, что:

— в формуле (6.36) n *должно быть достаточно большим числом*, а случайная величина X может иметь любой закон распределения;

— в формуле (6.37) n — любое число, но X — нормально распределенная величина, тогда как в формуле (6.14) и n — любое число и X может иметь любой закон распределения.

1. Найти нижнюю границу $\underline{P} > 0$ вероятности того, что $\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon$, если известны $\varepsilon > 0$, n и число $c \geqslant DX$.

Решение. Из (6.36), так же как и из (6.37), получаем

$$\underline{P} = 2\Phi\left(\frac{\varepsilon}{\sqrt{c/n}}\right), \quad c \geqslant DX. \tag{6.38}$$

2. Найти число $\varepsilon > 0$ такое, при котором $P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon\right) \geqslant \underline{P}$, если известны $\underline{P} > 0$, n и $c \geqslant DX$.

Решение. Полагая в (6.36) так же, как и в (6.37), $2\Phi\left(\frac{\varepsilon}{\sqrt{c/n}}\right) \geqslant \underline{P}$, найдем в П. 1 такое $z_{\underline{P}/2}$, при котором $\Phi(z_{\underline{P}/2}) = \underline{P}/2$. Тогда из неравенства $\frac{\varepsilon}{\sqrt{c/n}} \geqslant z_{\underline{P}/2}$ получим

$$\varepsilon \geqslant z_{\underline{P}/2}\sqrt{c/n}, \qquad c \geqslant DX. \tag{6.39}$$

3. Найти n, если известны $\varepsilon > 0$, нижняя граница $\underline{P}$ вероятности того, что $\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon$ и $c \geqslant DX$.

Решение. Из формулы (6.39) получаем

$$n \geqslant \frac{c z^2_{\underline{P}/2}}{\varepsilon^2}, \qquad c \geqslant DX. \tag{6.40}$$

О распределении относительной частоты m/n при большом числе n испытаний Бернулли. Из приблизительного равенства (6.30), имеющего место при *большом* n, следует, что

$$m/n \approx N(p, \sqrt{p(1-p)/n}), \tag{6.41}$$

т. е. относительная частота появления успеха в большом числе n испытаний Бернулли (рассматриваемая при фиксированном n как случайная величина) ведет себя приблизительно как нормально распределенная величина с математическим ожиданием, равным p, и средним квадратическим отклонением, равным $\sqrt{p(1-p)/n}$. Поэтому для относительной частоты m/n при *больших* n имеют место (приближенно) все формулы нормального закона и, в частности, формула (5.47), которую, учитывая (6.41), можно записать в виде

$$P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) \approx 2\Phi\left(\frac{\varepsilon}{\sqrt{p(1-p)/n}}\right) \geqslant$$

$$\geqslant 2\Phi\left(\frac{\varepsilon}{\sqrt{c/n}}\right), \qquad p(1-p) \leqslant c \leqslant 0{,}25. \tag{6.42}$$

Используя формулу (6.42), можно решать задачи трех типов, аналогичных рассмотренным в § 6.1, решаемых в условиях **модели II** с использованием неравенства (6.16). Но при этом надо иметь в виду, что *формула* (6.42) *предпо-*

лагает, что n достаточно велико, тогда как в формуле (6.16) *n — любое число.*

1. Найти нижнюю границу $\underline{P} > 0$ вероятности того, что $|m/n - p| < \varepsilon$, если известны $\varepsilon > 0$, n и число c такое, что $p(1-p) \leqslant c \leqslant 0{,}25$.

Р е ш е н и е. Из (6.42) получаем

$$\underline{P} = 2\Phi\left(\frac{\varepsilon}{\sqrt{c/n}}\right), \qquad p(1-p) \leqslant c \leqslant 0{,}25. \tag{6.43}$$

2. Найти число $\varepsilon > 0$ такое, при котором $P(|m/n - p| < \varepsilon) \geqslant \underline{P}$, если известны $\underline{P} > 0$, n и число c такое, что $p(1-p) \leqslant c \leqslant 0{,}25$.

Р е ш е н и е. Полагая в (6.42) $2\Phi\left(\frac{\varepsilon}{\sqrt{c/n}}\right) \geqslant \underline{P}$, найдем в приложении П. 1 такое число $z_{\underline{P}/2}$, при котором $\Phi(z_{\underline{P}/2}) = \underline{P}/2$. Тогда из неравенства $\varepsilon/\sqrt{c/n} \geqslant z_{\underline{P}/2}$ получим

$$\varepsilon \geqslant z_{\underline{P}/2}\, c/n, \qquad p(1-p) \leqslant c \leqslant 0{,}25. \tag{6.44}$$

3. Найти n, если известны $\varepsilon > 0$, нижняя граница $\underline{P} > 0$ вероятности того, что $|m/n - p| < \varepsilon$, и число c такое, что $p(1-p) \leqslant C \leqslant 0{,}25$.

Р е ш е н и е. Из формулы (6.44) получаем

$$n \geqslant \frac{c z^2_{\underline{P}/2}}{\varepsilon^2}, \qquad p(1-p) \leqslant c \leqslant 0{,}25. \tag{6.45}$$

Решим задачи 6.1—6.3, ориентируясь на нормальный закон распределения, и полученные результаты сравним с результатами, найденными при использовании неравенства Чебышёва.

ЗАДАЧА 6.1 (продолжение). По условию $n = 300$; $P(|m/n - p| < \varepsilon) \geqslant 0{,}95$, т. е. $\underline{P} = 0{,}95$; $\hat{p} = 0{,}515$. Найти ε.

Р е ш е н и е. Полагая $c = \hat{p}(1-\hat{p})$, найдем ε по формуле (6.44). При $\underline{P}/2 = 0{,}475$ в П. 1 найдем $z_{0{,}475} = 1{,}95$; $\varepsilon \geqslant 1{,}95\sqrt{\frac{0{,}515 \cdot 0{,}485}{300}} \geqslant 0{,}056$.

Нижняя граница для ε стала значительно меньше (при использовании неравенства Чебышёва $\varepsilon \geqslant 0{,}12904$), что следовало ожидать. Здесь учтена информация о законе распределения относительной частоты: при большом n он приблизительно нормальный, а при использовании неравенства Чебышёва эта информация не учитывалась.

ЗАДАЧА 6.2 (продолжение). По условию $DX \leqslant 0{,}0016$, т. е. $c = 0{,}0016$; $P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < 0{,}001\right) \geqslant 0{,}9$, т. е. $\varepsilon = 0{,}001$, а $\underline{P} = 0{,}9$. Найти n.

Р е ш е н и е. Воспользуемся формулой (6.40). При $\underline{P}/2 = 0{,}45$ в П. 1 найдем $z_{0{,}45} = 1{,}65$; $n \geqslant 0{,}0016 \cdot 1{,}65^2/0{,}001^2 = 4356$ (при использовании неравенства Чебышёва $n \geqslant 16\,000$). Значительное уменьшение нижней границы для n объясняется, как и в задаче 6.1, тем, что здесь учтена информация о законе распределения выборочного среднего: закон приблизительно нормальный. При $n \geqslant 4356$ (это не мало!) допущение о нормальности распределения выборочного среднего имеет основание.

ЗАДАЧА 6.3 (продолжение). По условию $P(|m/n - p| < 0{,}01) \geqslant 0{,}95$, т. е. $\varepsilon = 0{,}01$, а $\underline{P} = 0{,}95$. Найти n.

Р е ш е н и е. Полагая $c = 0{,}25$, воспользуемся формулой (6.45), в которой $z_{P/2} = z_{0{,}475} = 1{,}95$. Получаем $n \geqslant 0{,}25 \cdot 1{,}95^2/0{,}01^2 = 9606{,}2$ что также существенно меньше результата, полученного при использовании неравенства Чебышёва ($n \geqslant 50\,000$), но вполне достаточно для допущения нормальности распределения относительной частоты m/n. ◀

УПРАЖНЕНИЯ

1. Какова вероятность того, что модуль отклонения случайной величины от ее математического ожидания меньше двух средних квадратических отклонений? Ответьте на этот же вопрос, если случайная величина имеет нормальный закон распределения.

2. Средние ежедневные расходы на покупку канцелярских принадлежностей для офиса банка составляют 1000 ден. ед. Оцените вероятность того, что расходы на канцелярские принадлежности в любой наугад выбранный день меньше 2000 ден. ед., если среднее квадратическое отклонение ежедневных расходов не превышает 200 ден. ед.

3. Для определения среднегодового дохода налогоплательщиков города налоговой инспекцией проверены 250 случайно отобранных его жителей. Оцените вероятность того, что среднегодовой доход отобранных жителей отклонится от среднегодового дохода жителей всего города менее (по модулю) чем на 1000 ден. ед., если известно, что среднее квадратическое отклонение годового дохода не превышает 2500 ден. ед.? Сколько надо отобрать жителей, чтобы отклонение в 1000 ден. ед. гарантировать с надежностью, не меньшей 99%?

4. Для ориентировочного определения средней урожайности сахарной свеклы на поле площадью 1500 га выбраны случайным образом участки площадью 1 м^2 на каждом гектаре поля, и на этих участках определена урожайность (ц). Оцените предельное отклонение выборочной средней от средней урожайности на всем поле, которое можно га-

рантировать с вероятностью, не меньшей 0,8. Установлено, что дисперсия урожайности на каждом гектаре не превышает 2500 ц^2.

5. Вероятность изготовления нестандартной радиолампы не превышает 0,03. Какое наименьшее количество радиоламп следует отобрать, чтобы с вероятностью, не меньшей 0,8, можно было ожидать, что доля нестандартных радиоламп среди отобранных будет отличаться от вероятности изготовления нестандартной лампы по абсолютной величине менее чем на 0,005?

6. Для экспериментальной проверки статистической устойчивости относительной частоты выпадения «герба» в различное время проведены следующие опыты:

1) монета была брошена 4040 раз, «герб» выпал 2948 раз (опыт Бюффона);

2) монета была брошена 12 000 раз, «герб» выпал 6019 раз (опыт К. Пирсона).

Для каждой из этих ситуаций найдите:

а) вероятность того, что при повторении опыта будет получен тот же результат;

б) вероятность того, что при повторении опыта модуль отклонения относительной частоты выпадения «герба» от вероятности, равной 0,5, не превзойдет зафиксированного в опыте отклонения;

в) границу модуля отклонения относительной частоты от вероятности, которую можно гарантировать с вероятностью 0,99.

Сколько раз нужно подбросить монету, чтобы с вероятностью 0,99 гарантировать, что модуль отклонения относительной частоты от вероятности, равной 0,5, не превзойдет 0,001?

Изучение случайной величины по результатам наблюдений

ГЛАВА 7

Первичная обработка выборочных данных

В первой и второй частях были рассмотрены математические модели случайных событий и случайных величин, а в приведенных там примерах и задачах предполагалась выполнимость некоторых «идеальных» условий, достаточных для использования той или иной модели (формулы). Предполагались также известными вероятности некоторых исходных событий или законы распределения случайных величин, или числовые характеристики величин и т. д. Однако правомерность этих предположений может быть подтверждена только опытом, наблюдениями. *Изучение случайных величин по результатам наблюдений — предмет математической статистики*[1]. Определим предмет математической статистики в терминах «генеральной совокупности» и «выборки», предварительно уточнив их.

Далее под ***генеральной совокупностью*** будем понимать совокупность всех мыслимых результатов наблюдений случайной величины X (или всех мысленно возможных объектов интересующего нас типа, у которых фиксируются значения признака — величины X). Поскольку в данном определении речь идет о всех мысленно возможных результатах наблюдений (или объектах), понятие генеральной совокупности является абстрактным, условно-математическим.

Определенную таким образом генеральную совокупность не следует смешивать с реальными совокупностями. Так, обследовав даже все семьи данного района с точки зрения их среднедушевого годового дохода, можно обследованную совокупность рассматривать лишь как представителя гипотетически возможной более широкой совокупности семей.

Введенное понятие генеральной совокупности, как совокупности всех мыслимых результатов наблюдений случайной величины, будем считать синонимом понятия случайной величины и в дальнейшем не различать эти понятия. Однако следует иметь в виду, что множество всех мыслимых результатов наблюдений исследуемой величины X, вообще говоря, «больше»

[1] Термин «статистика» может употребляться в трех значениях. Это: название научной дисциплины; собранная числовая информация; название случайной величины, являющейся функцией $\varphi(X_1, X_2, \ldots, X_n)$ случайных результатов наблюдений $X_1, X_2, \ldots, X_n$ исследуемой величины X.

множества всех значений величины: каждое фиксированное ее значение может наблюдаться неоднократно.

Выборка из данной генеральной совокупности — это результаты ограниченного числа n наблюдений случайной величины X (значения признака — величины X, зафиксированные у n объектов интересующего нас типа), или, короче, выборка — это обследованная часть генеральной совокупности; n — объем выборки.

Изучение генеральной совокупности (в целом) по выборке из нее — предмет математической статистики.

Выборочные данные (результаты наблюдения изучаемой случайной величины X) обычно труднообозримы. Для того чтобы, используя эти данные, сделать выводы относительно генеральной совокупности (относительно величины X), надо провести первичную их обработку, включающую:

— придание выборочным данным наглядного вида для получения представления о законе распределения генеральной совокупности (о законе распределения величины X);

— вычисление выборочных числовых характеристик для получения представления о числовых характеристиках генеральной совокупности (о числовых характеристиках величины X).

Методы первичной обработки выборочных данных и их реализации в Microsoft Excel рассматриваются в этой главе.

§ 7.1. Выборочные аналоги функции распределения, ряда распределения и функции плотности

Числовые результаты n наблюдений случайной величины X или выборочные данные обозначим через $x_1, x_2, \ldots, x_n$.

Вариационным рядом называют ряд выборочных данных, расположенных в порядке неубывания; вариационный ряд обозначают так:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}, \qquad (7.1)$$

где $x_{(1)} \geqslant x_{(2)} \geqslant \ldots \geqslant x_{(n)}$.

Выборочной «случайной» величиной называют дискретную случайную величину $\hat{X}$, ряд распределения которой имеет вид таблицы 7.1.

Таблица 7.1

$x_{(i)}$	$x_{(1)}$	$x_{(2)}$	...	$x_{(n)}$	
$P(\hat{X} = x_{(i)})$	$1/n$	$1/n$	...	$1/n$	$\Sigma = 1$

При наличии в вариационном ряду (7.1) повторяющихся чисел имеет смысл объединить их в группы. Введем следующие обозначения:

$x'_1, x'_2, \ldots, x'_v$ — расположенные *в порядке возрастания* различные наблюдавшиеся значения величины X — ***варианты***;

ν — количество вариантов;

m_i — число, показывающее, сколько раз встречается вариант x_i' в ряду наблюдений, — ***частота варианта*** x_i, $i = 1, 2, \ldots, \nu$;

$\hat{p}_i = m_i/n$ — ***относительная частота*** (частость, опытная, или статистическая вероятность) ***варианта*** x_i', $i = 1, 2, \ldots, \nu$.

Составим таблицу 7.2, в первой строке которой поместим в возрастающем порядке различающиеся значения выборочной величины $\hat{X}$, или, иначе, варианты x_i', $i = 1, 2, \ldots, \nu$, а во второй строке $P(\hat{X} = x_i') = \hat{p}_i$:

Таблица 7.2

x_i'	x_1'	x_2'	...	x_ν'
$\hat{p}_i$	$\hat{p}_1 = m_1/n$	$\hat{p}_2 = m_2/n$	...	$\hat{p}_\nu = m_\nu/n$

Эта таблица показывает распределение опытных, или статистических, вероятностей между вариантами и называется ***статистическим рядом распределения.***

Очевидны следующие соотношения:

$$\sum_{i=1}^{\nu} m_i = n; \quad 0 < \hat{p}_i < 1, \quad i = 1, 2, \ldots, \nu; \quad \sum_{i=1}^{\nu} \hat{p}_i = 1. \quad (7.2)$$

Статистический ряд распределения — выборочный аналог ряда распределения вероятностей (см. табл. 4.3). Различие между ними состоит в следующем: в вероятностном ряду указывают *все возможные* значения величины X и их *истинные вероятности*, а в статистическом — *различные наблюдавшиеся* значения величины X и *опытные вероятности* этих значений. Ломаная линия, звенья которой соединяют соседние точки с координатами $(x_i', \hat{p}_i)$, $i = 1, 2, \ldots, \nu$, называется ***многоугольником распределения относительных частот*** (частостей, опытных, или статистических вероятностей) — это выборочный аналог многоугольника распределения вероятностей.

Введем понятие выборочной функции распределения. Напомним, что функцией распределения случайной величины X мы называли функцию $F_X(x) = P(X < x)$, где x — любое действительное число (см. (4.1)). По аналогии ***выборочной функцией распределения случайной величины*** X назовем функцию

$$\hat{F}_X(x) = \hat{P}(X < x), \quad (7.3)$$

где x — любое действительное число; $\hat{P}(X < x)$ — опытная вероятность того, что величина X примет значение, меньшее числа x.

Выборочную функцию распределения $\hat{F}_X(x)$ случайной величины X иногда называют функцией распределения выборочной случайной величины $\hat{X}$. В этом случае ее обозначают $F_{\hat{X}}(x)$ и пишут: $F_{\hat{X}}(x) = P(\hat{X} < x)$. Определенная таким образом функция совпадает с функцией (7.3), поскольку $P(\hat{X} < x) = \hat{P}(X < x)$.

Пусть наблюдения сгруппированы в статистический ряд (см. табл. 7.2). Тогда табличная форма задания выборочной функции распределения (7.3) такова:

Таблица 7.3

x	$(-\infty, x_1']$	$(x_1', x_2']$	$(x_2', x_3']$	...	$(x_{v-1}', x_v']$	$(x_v', +\infty)$
$\hat{F}_X(x)$	0	$\hat{p}_1$	$\hat{p}_1 + \hat{p}_2$	...	$\hat{p}_1 + \hat{p}_2 + ... + \hat{p}_{v-1}$	$\hat{p}_1 + \hat{p}_2 + ... + \hat{p}_v = 1$

Поясним, как заполняется таблица 7.3. Пусть $x = x_1'$. Так как не было наблюдений, меньших числа x_1' (см. табл. 7.2), то опытная вероятность $\hat{P}(X < x_1') = 0$, поэтому и $\hat{F}_X(x_1') = \hat{P}(X < x_1') = 0$. Таким же образом объясняется и то, что $\hat{F}_X(x) = 0$ для всех $x < x_1'$.

Пусть $x \in (x_1', x_2']$. Поскольку в ряду n наблюдений чисел, меньших x, было m_1 и все они равны варианту x_1' (см. табл. 7.2), $\hat{P}(X < x) = \hat{P}(X = x_1') = m_1/n = \hat{p}_1$, следовательно, $\hat{F}_X(x) = \hat{p}_1$ и т. д.

Функция $\hat{F}_X(x)$, заданная таблицей 7.3, — выборочный аналог функции $F_X(x)$, заданной таблицей 4.5; как и последняя, она является скачкообразной: скачки имеют место в точках — вариантах $x_1', x_2', \ldots, x_v'$, а величины скачков равны относительным частотам, — опытным вероятностям $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_v$. График функции $\hat{F}_X(x)$, заданной таблицей 7.3, называется ***кумулятивной кривой*** (кривой накопленных частот).

▶ **ПРИМЕР 7.1.** У $n = 100$ случайно выбранных студентов, каждый из которых сдавал четыре экзамена, собраны сведения о величине X — количестве сданных экзаменов:

4, 0, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 4, 4, 4, 4, 1, 2, 3, 4, 4, 4, 4, 3, 4, 4, 3, 4, 3, 3, 4, 3, 3, 4, 4, 3, 3, 4, 4, 4, 4, 3, 4, 4, 4, 3, 4, 4, 4, 4, 4, 2, 3, 4, 4, 4, 4, 4, 3, 3, 3, 4, 4, 4, 3, 4, 4, 4, 4, 3, 3, 4, 4, 4, 3, 3, 3, 4, 4, 4, 2, 3, 3, 4, 4, 3, 3, 3, 4, 4, 4, 4, 4, 3, 3, 3.

Среди этих чисел есть повторяющиеся. Сгруппируем данные в статистический ряд (см. табл. 7.2). Вариантами являются числа $x_1' = 0$, $x_2' = 1$, ... , $x_5' = 4$ — различное число сданных экзаменов; частоты (m_i) и относительные частоты $(\hat{p}_i)$ этих вариантов приведены в таблице 7.4.

Таблица 7.4[1]

i	1	2	3	4	5	
Число сданных экзаменов (x_i')	0	1	2	3	4	
Число студентов (m_i)	1	1	3	35	60	$\Sigma = 100$
$\hat{p}_i = m/n$	0,01	0,01	0,03	0,35	0,60	$\Sigma = 1$
$p_i = C_4^{x_i'} 0{,}88^{x_i'} \times 0{,}12^{4-x_i'}$	0,00021	0,00608	0,06691	0,32711	0,59969	$\Sigma = 1$

Выборочная функция распределения $\hat{F}_X(x) = \hat{P}(X < x)$ приведена в таблице 7.5.

Таблица 7.5

x	$(-\infty; 0]$	$(0; 1]$	$(1; 2]$	$(2; 3]$	$(3; 4]$	$(4; +\infty)$
$\hat{F}_X(x)$	0	$0 + 0{,}01 = 0{,}01$	$0{,}01 + 0{,}01 = 0{,}02$	$0{,}02 + 0{,}03 = 0{,}05$	$0{,}05 + 0{,}35 = 0{,}4$	$0{,}4 + 0{,}6 = 1$

Многоугольник распределения относительных частот, заданный таблицей 7.4, изображен на рисунке 7.1, *а* сплошной линией. График выборочной функции распределения, заданной таблицей 7.5 (кумулятивная кривая), изображен на рисунке 7.1, *б*.

Если просмотр первичных данных не позволял составить представление о варьировании количества сданных экзаменов, то, построив статистический ряд распределе-

[1] Содержание последней строки таблицы выясняется в примере 7.11.

ния и выборочную функцию распределения $\hat{F}_X(x)$, можно сделать вывод: студентов, сдавших четыре экзамена, примерно в два раза больше студентов, которые сдали три экзамена; с ростом количества сданных экзаменов растет и число сдавших их студентов; 40% студентов сдали менее четырех экзаменов и т. д. ◀|

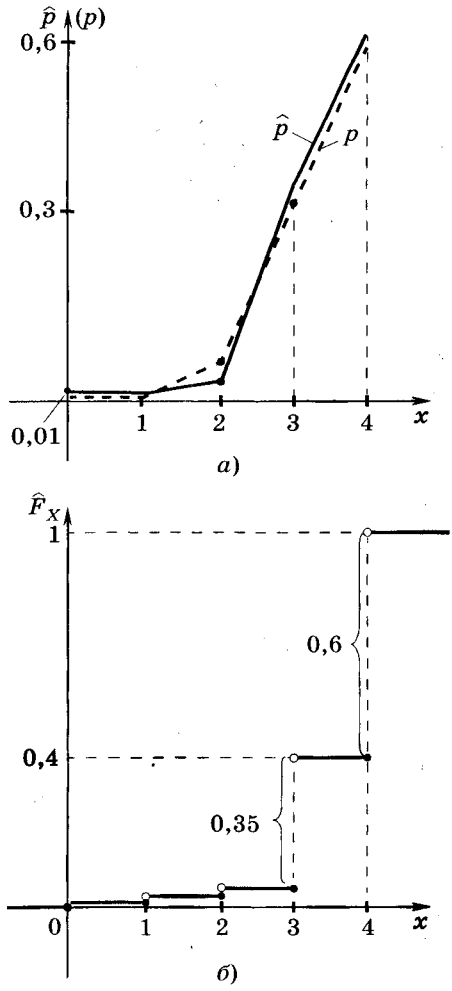


Рис. 7.1

В примере 7.1 изучалась дискретная случайная величина X — количество сданных экзаменов из четырех; число ее возможных значений (0, 1, 2, 3, 4) мало. При изучении по наблюдениям дискретной величины с большим числом возможных значений или непрерывной величины (ее значения — это все точки некоторого отрезка) группировка наблюдений в статистический ряд зачастую не позволяет уловить характерные черты их варьирования: слишком много может оказаться вариантов — различающихся чисел в ряду наблюдений — и малой относительная частота вариантов. Поэтому обычно результаты наблюдений непрерывной величины (и дискретной с большим числом возможных значений) группируют в ***интервальный ряд*** (см. табл. 7.6).

Таблица 7.6

$[a_i, a_{i+1})$	$[a_1, a_2)$	$[a_2, a_3)$	...	$[a_v, a_{v+1})$	
$\hat{p}_i$	$\hat{p}_1 = m_1/n$	$\hat{p}_2 = m_2/n$	...	$\hat{p}_v = m_v/n$	$\Sigma = 1$

В таблице 7.6: $[a_i, a_{i+1})$, $i = 1, 2, \ldots, \nu$, — интервалы группирования одинаковой длины; m_i — число наблюдений, попавших в интервал $[a_i, a_{i+1})$, иначе, число наблюдений, бóльших или равных левой границе a_i интервала и меньших его правой границы a_{i+1}; числа $m_1, m_2, \ldots, m_\nu$ называют ***интервальными частотами***; $\hat{p}_i = m_i/n$, $i = 1, 2, \ldots, \nu$, — ***интервальные относительные частоты*** (частости, опытные, или статистические вероятности); $\hat{p}_i = \hat{P}(a_i \leqslant X < a_{i+1})$.

Установим, каким должно быть число интервалов и как следует формировать их границы.

При слишком большом, как и при слишком малом, числе ν интервалов, уловить характерные черты варьирования наблюдений нельзя. Ориентиром в выборе оптимального числа интервалов служит формула Стерджеса: число интервалов $\nu = 1 + 3{,}322 \lg n$, где n — число наблюдений. Тогда длина каждого интервала

$$h = (x_{(n)} - x_{(1)})/(1 + 3{,}322 \lg n), \qquad (7.4)$$

где $x_{(1)}$ и $x_{(n)}$ — соответственно минимальный и максимальный результат наблюдений (см. (7.1)).

Обычно при расчете h после запятой оставляют столько десятичных знаков, столько их (или на один больше) в исходных данных.

Границы интервалов формируют так: начало первого интервала $a_1 = x_{(1)} - h/2$; конец первого или начало второго интервала $a_2 = a_1 + h$; конец второго или начало третьего интервала $a_3 = a_2 + h$ и т. д.; формирование интервалов заканчивают, как только конец очередного интервала больше максимального наблюдения $x_{(n)}$.

Если наблюдения величины X сгруппированы в интервальный ряд (см. табл. 7.6), то точные значения выборочной функции распределения $\hat{F}_X(x) = \hat{P}(X < x)$ можно найти для тех x, которые указаны в таблице 7.7.

Таблица 7.7

x	$(-\infty, a_1]$	a_2	a_3	...	a_ν	$[a_{\nu+1}, +\infty)$
$\hat{F}_X(x)$	0	$\hat{p}_1$	$\hat{p}_1 + \hat{p}_2$	...	$\hat{p}_1 + \hat{p}_2 + \ldots$ $\ldots + \hat{p}_{\nu-1}$	$\hat{p}_1 + \hat{p}_2 + \ldots$ $\ldots + \hat{p}_\nu = 1$

Поясним эту таблицу. Например, при $x = a_3$

$$\hat{F}_X(a_3) \underset{(7.3)}{=} \hat{P}(X < a_3) = \hat{P}(-\infty < X < a_3) =$$
$$= \hat{P}(X \in [a_1, a_2)) + \hat{P}(X \in [a_2, a_3)) = \hat{p}_1 + \hat{p}_2;$$

при $x = a_{\nu+1}$

$$\widehat{F}_X(a_{\nu+1}) = \widehat{P}(X < a_{\nu+1}) = \widehat{P}(-\infty < X < a_{\nu+1}) =$$

$$= \widehat{p}_1 + \widehat{p}_2 + \ldots + \widehat{p}_\nu = 1.$$

Внутри интервалов $[a_i, a_{i+1})$, $i = 1, 2, \ldots, \nu$, предполагается равномерное «накапливание» вероятностей (см. кумулятивную кривую, изображенную на рисунке 7.2, *а*).

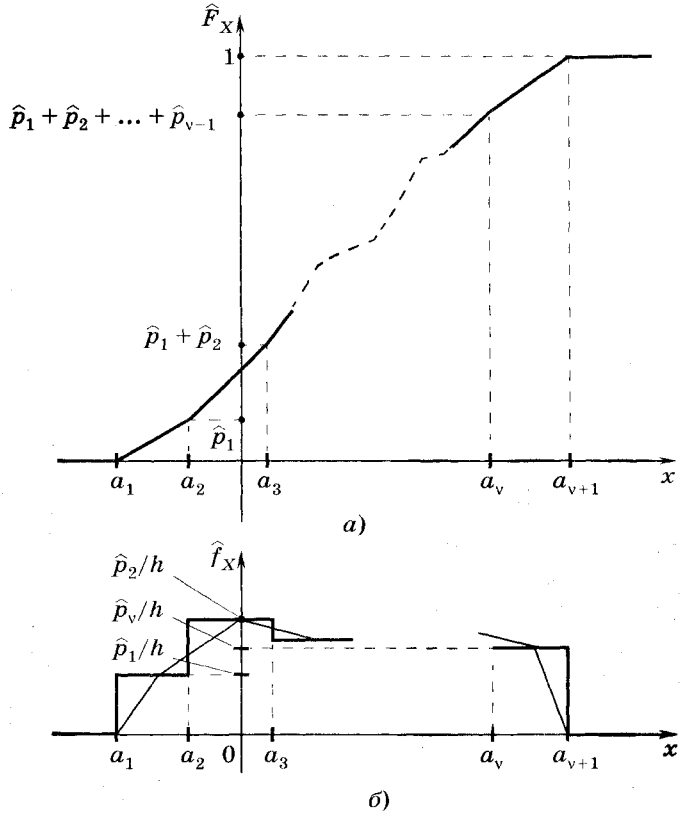


Рис. 7.2

Пусть наблюдения непрерывной случайной величины X сгруппированы в интервальный ряд (см. табл. 7.6). Введем понятие выборочной функции плотности. Из определения (4.8) функции плотности $f_X(x)$ вероятности случайной величины X следует, что при малом Δx

$$f_X(x)\,\Delta x \approx P(x \leqslant X < x + \Delta x).$$

Аналогично ***выборочной функцией плотности*** (или функцией плотности опытной вероятности) назовем

функцию $\hat{f}_X(x)$, которая при $x \in [a_i, a_{i+1})$ удовлетворяет равенствам

$$\hat{f}_X(x)h = \hat{P}(a_i \leqslant X < a_{i+1}) = \hat{p}_i,\ i = 1, 2, \ldots, \nu,$$

где h — длина интервала. Отсюда получаем, что

$$\hat{f}_X(x) = \hat{p}_i/h,\ x \in [a_i, a_{i+1}),\ i = 1, 2, \ldots, \nu.$$

Вне интервалов интервального ряда (см. табл. 7.6) $\hat{f}_X(x) = 0$.

Табличная форма задания выборочной функции плотности такова:

Таблица 7.8

x	$(-\infty, a_1)$	$[a_1, a_2)$	$[a_2, a_3)$	...	$[a_\nu, a_{\nu+1})$	$[a_{\nu+1}, +\infty)$
$\hat{f}_X(x)$	0	$\hat{p}_1/h$	$\hat{p}_2/h$	...	$\hat{p}_\nu/h$	0

Графически выборочную функцию плотности $\hat{f}_X(x)$ изображают в виде столбчатой диаграммы, называемой ***гистограммой***, которую часто «сглаживают» ***полигоном*** (рис. 7.2, *б*).

Площадь под выборочной функцией плотности (площадь столбчатой диаграммы) равна $h\hat{p}_1/h + h\hat{p}_2/h + \ldots$ $\ldots + h\hat{p}_\nu/h = 1$ (площадь под функцией плотности вероятности также равна 1). Полигон состоит из отрезков, соединяющих середины соседних «площадок».

Замечание. Часто по оси ординат откладывают интервальные частоты m_i или интервальные частости $\hat{p}_i = m_i/n$, и полученные столбчатые диаграммы также называют гистограммами. По форме эти гистограммы не отличаются от ранее определенной, однако площади под ними не равны единице.

Рассмотрим пример построения по наблюдениям интервального ряда, при этом воспользуемся пакетом прикладных программ Microsoft Excel.

▶ **ПРИМЕР 7.2.** Служба маркетинга собрала сведения об объеме ежедневных продаж товара (в ден. ед.) дилером за последние 100 дней:

47,0; 37,2; 52,4; 62,8; 62,0; 67,3; 28,2; 47,7; 61,0; 39,1;
46,7; 46,3; 63,4; 49,1; 48,1; 44,9; 69,7; 58,7; 73,8; 43,5;
35,6; 41,5; 34,8; 46,4; 49,7; 50,3; 46,8; 71,9; 32,6; 42,6;
56,9; 53,2; 40,6; 47,6; 51,3; 55,6; 51,4; 40,9; 68,8; 54,9;
72,1; 64,4; 63,0; 51,1; 50,0; 54,5; 49,7; 39,5; 32,3; 58,3;
43,1; 33,1; 31,5; 40,2; 42,3; 28,8; 44,3; 46,0; 51,3; 46,3;

66,6; 33,9; 55,4; 59,0; 69,2; 49,2; 44,8; 56,8; 46,2; 57,6; 24,2; 64,5; 37,2; 43,5; 57,6; 54,7; 58,7; 56,0; 36,3; 38,8; 50,7; 58,3; 58,6; 43,6; 40,8; 61,1; 38,0; 34,4; 57,1; 56,4.

В условиях примера наблюдаемая случайная величина X — ежедневный объем продаж; число наблюдений $n = 100$. Данные об объеме продаж введем в рабочий лист (таблицу) Microsoft Excel. Воспользовавшись **Статистическими функциями** МИН и МАКС, найдем минимальный и максимальный объем продаж: $x_{(1)} = 24{,}2$; $x_{(100)} = 73{,}8$. Вычислим по формуле (7.4) длину интервала h:

$$h = (73{,}8 - 24{,}2)/(1 + 3{,}322 \lg 100) \approx 6{,}5.$$

Сформируем интервалы: начало первого интервала

$$a_1 = x_{(1)} - h/2 = 24{,}2 - 3{,}25 = 20{,}95;$$

его конец

$$a_2 = a_1 + h = 20{,}95 + 6{,}5 = 27{,}45;$$

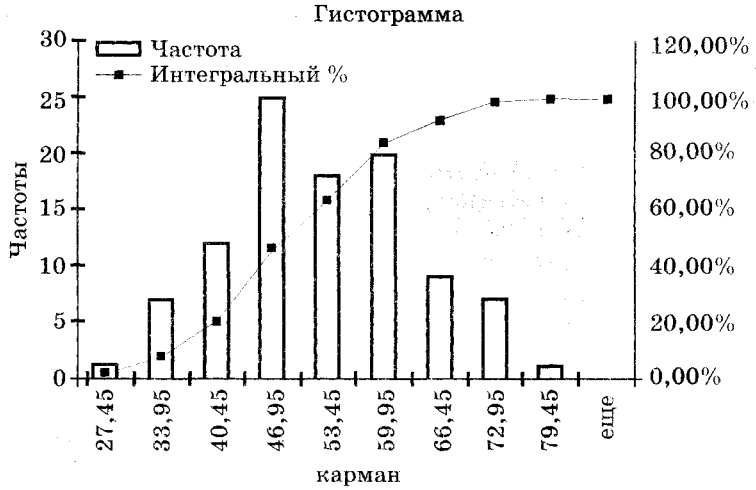
конец второго интервала

$$a_3 = a_2 + h = 27{,}45 + 6{,}5 = 33{,}95 \text{ и т. д.}$$

Формирование интервалов закончим, как только конец очередного интервала будет больше максимального наблюдения $x_{(100)} = 73{,}8$; этим «концом» является конец девятого интервала, равный 79,45 (см. табл. 7.9, столбец 2).

Теперь, просматривая наблюдения, надо определить, сколько из них попадет в каждый интервал, т. е. найти интервальные частоты. В рассматриваемом примере концы интервалов содержат два десятичных знака, а наблюдения — один десятичный знак. Поэтому принятая ранее договоренность о включении в интервал наблюдений, бóльших или равных его левой границе и меньших правой, здесь опускается. Эту работу можно существенно облегчить, если воспользоваться программой «**Гистограмма**» из пакета «**Анализ данных**» Microsoft Excel. Последовательность действий такая:

- в рабочий лист в дополнение к 100 имеющимся там числам введем правые границы интервалов (эти границы называют *карманами*);
- выберем в меню «**Сервис**» пункт «**Анализ данных**», а в нем пункт «**Гистограмма**»;
- в окне ввода исходных данных программы «**Гистограмма**» укажем *входной интервал* (ссылка на ячейки, содержащие 100 наблюдений), *интервал карманов* (ссылка на ячейки, содержащие правые границы интервалов), *выходной интервал* (ссылка на ячейку, начиная с которой будут располагаться результаты работы программы) и ак-



Карман	Частота	Интегральный %
27,45	1	1,00%
33,95	7	8,00%
40,45	12	20,00%
46,95	25	45,00%
53,45	18	63,00%
59,95	20	83,00%
66,45	9	92,00%
72,95	7	99,00%
79,45	1	100,00%
Еще	0	100,00%

Рис. 7.3

тивизируем (установим флажок) «*Интегральный процент*» (для подсчета процентных значений выборочной функции распределения в правых границах интервалов, $\hat{F}_X(a_{i+1})100\% = \hat{P}(X < a_{i+1})100\%$) и «*Вывод графика*» (для построения гистограммы, в которой по оси ординат откладываются интервальные частоты m_i, и кумулятивной кривой).

Результаты работы программы «**Гистограмма**» представлены на рисунке 7.3[1].

[1] Программа «Гистограмма» к «текущему карману» относит наблюдения, меньшие и равные этого «кармана» и большие «предыдущего кармана», что, вообще говоря, не согласуется с принятой нами договоренностью о наблюдениях, включенных в интервал. В условиях примера эта несогласованность не сказывается на результатах группировки, поскольку правые концы интервалов — карманы имеют на один десятичный знак больше, чем исходные.

Интервальные частости $\hat{p}_i$, значения выборочной функции плотности $\hat{f}_X(x_i')$ в серединах интервалов и значения выборочной функции распределения $\hat{F}_X(a_{i+1})$ в правых границах интервалов приведены в таблице 7.9 (столбцы 5, 6 и 7). Гистограмма (по оси ординат отложены значения $\hat{f}_X(x_i')$) и полигон изображены на рисунке 7.4, *а*, кумулята на рисунке 7.4, *б*. О графиках функций $f_N(x)$ и $F_N(x)$, изображенных на этом рисунке, речь пойдет в примере 7.13.

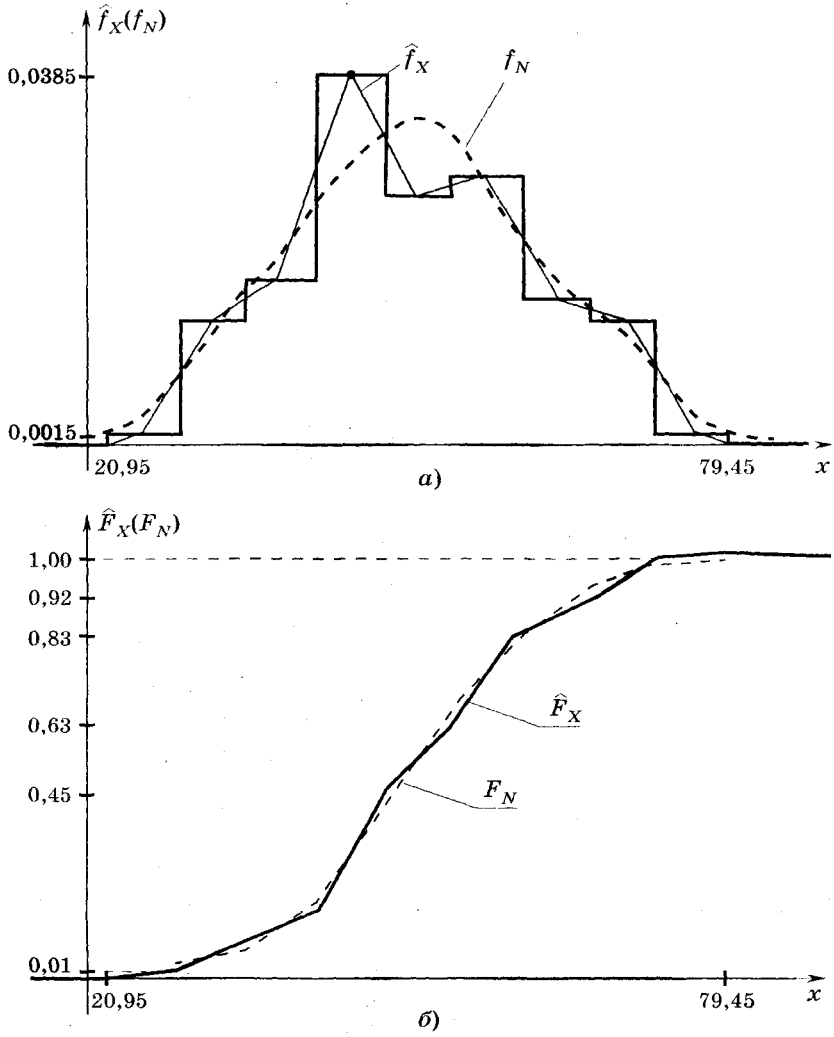


Рис. 7.4

i	$[a_i, a_{i+1})$	Середина интервала x_i'	Интервальная частота m_i	Интервальная частость $\hat{p}_i = m_i/n$	Выборочная функция плотности $\hat{f}_X(x_i') = \hat{p}_i/h = \hat{p}_i/6{,}5$	Выборочная функция распределения $\hat{F}_X(a_{i+1})$
1	2	3	4	5	6	7
1	[20,95; 27,45)	24,2	1	0,01	0,0015	0,01
2	[27,45; 33,95)	30,7	7	0,07	0,0108	0,08
3	[33,95; 40,45)	37,2	12	0,12	0,0185	0,20
4	[40,45; 46,95)	43,7	25	0,25	0,0385	0,45
5	[46,95; 53,45)	50,2	18	0,18	0,0277	0,63
6	[53,45; 59,95)	56,7	20	0,20	0,0308	0,83
7	[59,95; 66,45)	63,2	9	0,09	0,0138	0,92
8	[66,45; 72,95)	69,7	7	0,07	0,0108	0,99
9	[72,95; 79,45)	76,2	1	0,01	0,0015	1,00
Σ			100	1,00		

Таблица 7.9[1]

$z_i = \frac{x_i' - \overline{x'}}{\hat{\sigma}_{(ш)}} = \frac{x_i' - 49{,}5}{10{,}83}$	$\varphi(z_i)$ (см. П. 1)	Плотность нормального распределения $f_N(x_i') = \varphi(z_i)/\hat{\sigma}_{(ш)} = \varphi(z_i)/10{,}83$	$z_{i+1} = \frac{a_{i+1} - \overline{x'}}{\hat{\sigma}_{(ш)}} = \frac{a_{i+1} - 49{,}5}{10{,}83}$	$\Phi(z_{i+1})$ (см. П. 1)	Функция нормального распределения $F_N(a_{i+1}) = 1/2 + \Phi(z_{i+1})$
8	9	10	11	12	13
−2,34	0,0252	0,0023	−2,04	−0,4798	0,0202
−1,74	0,0863	0,0080	−1,44	−0,4265	0,0735
−1,14	0,2059	0,0190	−0,84	−0,3023	0,1977
−0,54	0,3429	0,0317	−0,24	−0,0987	0,4013
0,06	0,3984	0,0368	0,36	0,1368	0,6368
0,66	0,3230	0,0298	0,96	0,3289	0,8289
1,26	0,1826	0,0169	1,56	0,4394	0,9394
1,86	0,0721	0,0066	2,16	0,4842	0,9842
2,46	0,0198	0,0018	2,76	0,4970	0,9970

[1] Содержание столбцов 8—13 выясняется в примере 7.13. При использовании таблицы П.1 (столбцы 9 и 12) значения аргументов z_i и z_{i+1} (столбцы 8 и 11) приравнивались к ближайшему z, заданному в этой таблице.

§ 7.2. Выборочные аналоги числовых характеристик случайных величин

Напомним, что выборочной «случайной» величиной называют величину $\hat{X}$, ряд распределения которой имеет вид таблицы 7.1.

Числовые характеристики выборочной «случайной» величины $\hat{X}$ называют выборочными аналогами соответствующих числовых характеристик случайной величины X (выборочными аналогами генеральных характеристик).

Они дают общее представление о результатах наблюдений величины X. Среди выборочных характеристик, так же как и среди генеральных, различают характеристики положения и характеристики рассеивания.

Выборочные характеристики положения

Математическое ожидание выборочной «случайной» величины $\hat{X}$, ряд распределения которой задан таблицей 7.1,

$$M\hat{X} \underset{(4.15)}{=} \sum_{i=1}^{n} x_{(i)} P(\hat{X} = x_{(i)}) = \sum_{i=1}^{n} x_{(i)} \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i{}^{1}.$$

Эта характеристика является выборочным аналогом математического ожидания MX случайной величины X (или выборочным аналогом генеральной средней). Математическое ожидание $M\hat{X}$ обозначают обычно символом $\bar{x}$ и называют ***выборочным средним***;

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Так вычисляется выборочное среднее в случае, когда исходные данные представлены в форме ряда наблюдений величины X: $x_1, x_2, \ldots, x_n$. Однако часто исходные данные представляются сгруппированными или в статистический ряд (см. табл. 7.2), или в интервальный ряд (см. табл. 7.6). Формулы вычисления среднего при различных формах задания исходных данных приведены в таблице 7.10.

---

[1] Так как $\sum_{i=1}^{n} x_{(i)} = \sum_{i=1}^{n} x_i$ (сумма вариационного ряда равна сумме результатов наблюдений), то символ $x_{(i)}$ в последнем равенстве заменен на x_i.

Таблица 7.10

Форма задания исходных данных	Выборочное среднее
Ряд наблюдений $x_1, x_2, \ldots, x_n$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (7.5)
Статистический ряд (см. таблицу 7.2)	$\bar{x}' = \frac{1}{n}\sum_{i=1}^{v} x_i' m_i = \sum_{i=1}^{v} x_i' \hat{p}_i$, (7.6) где x_i', $i = 1 \div v$, — варианты, m_i — частоты, $\hat{p}_i = m_i/n$ — частости вариантов, $n = \sum_{i=1}^{v} m_i$ — общее число наблюдений
Интервальный ряд (см. таблицу 7.6)	$\bar{x}' = \frac{1}{n}\sum_{i=1}^{v} x_i' m_i = \sum_{i=1}^{v} x_i' \hat{p}_i$, (7.7) где x_i', $i = 1 \div v$, — середины интервалов; m_i — интервальные частоты, $\hat{p}_i = m_i/n$ — интервальные частости, $n = \sum_{i=1}^{v} m_i$ — общее число наблюдений

Формулы (7.5) и (7.6) дают одно и то же значение выборочного среднего. Если же наблюдения сгруппированы в интервальный ряд, то значения выборочного среднего, полученные по формулам (7.7) и (7.5), будут отличаться: при использовании (7.7) наблюдения, попавшие в тот или иной интервал, «заменяются» его серединой.

Работу, связанную с вычислением выборочного среднего (и других характеристик), можно существенно облегчить, если воспользоваться программой **«Описательная статистика»** из пакета **«Анализ данных»** Microsoft Excel. В окне ввода исходных данных программы **«Описательная статистика»** указывается: *входной интервал* (ссылка на ячейки, содержащие исходные данные — результаты n наблюдений величины X); *выходной интервал* (ссылка на ячейку, начиная с которой будут располагаться результаты работы программы); активизируется (устанавливается флажок) *«Итоговая статистика»*, *«уровень надежности»* (по умолчанию 0,95, его смысл выясняется в § 8.3), *«1-й наименьший»*, *«1-й наибольший»*.

▶ **ПРИМЕР 7.3.** Вычислим выборочные средние по данным, приведенным в примерах 7.1 и 7.2.

В рабочий лист Microsoft Excel введем два массива данных:

— 100 значений числа сданных экзаменов (см. пример 7.1) и

— 100 значений объема продаж (см. пример 7.2).

Результаты работы программы **«Описательная статистика»** представлены соответственно на рисунках 7.5, *а* и 7.5, *б* (смысл этих результатов выясняется в этом и следующем параграфе).

Характеристика	Значение характеристики для числа сданных экзаменов	Значение характеристики для объема продаж
Среднее	3,52	49,596
Стандартная ошибка	0,07032	1,08542
Медиана	4	49,15
Мода	4	37,2
Стандартное отклонение	0,70324	10,8542
Дисперсия выборки	0,49454	117,813
Эксцесс	6,36796	–0,4744
Асимметричность	–2,02829	0,09077
Интервал	4	49,6
Минимум	0	24,2
Максимум	4	73,8
Сумма	352	4959,6
Счет	100	100
Наибольший	4	73,8
Наименьший	0	24,2
Уровень надежности	0,13954	2,15371

а) *б*)

Рис. 7.5

Воспользуемся формулами выборочного среднего, содержащимися в таблице 7.10.

Данные примера 7.1. Имеем

$$\bar{x} \underset{(7.5)}{=} (4 + 0 + 4 + \ldots + 3)/100 = 3{,}52$$

— среднее число сданных экзаменов одним студентом; это среднее и приведено в результатах **«Описательной статистики»** (рис. 7.5, *а*).

Обратившись к статистическому ряду (табл. 7.4), вычислим

$$\overline{x'} \underset{(7.6)}{=} 0 \cdot 0{,}01 + 1 \cdot 0{,}01 + 2 \cdot 0{,}03 + 3 \cdot 0{,}35 + 4 \cdot 0{,}6 = 3{,}52.$$

Формулы (7.5) и (7.6), как и следовало ожидать, дали одинаковый результат.

Данные примера 7.2. Имеем

$$\bar{x} \underset{(7.5)}{=} (47{,}0 + 37{,}2 + \ldots + 56{,}4)/100 = 49{,}596,$$

— таков в среднем ежедневный объем продаж товара; это среднее и приведено в результатах **«Описательной статистики»** (рис. 7.5, *б*).

Воспользовавшись интервальным рядом (табл. 7.9), получим

$$\bar{x}' \underset{(7.7)}{=} 24{,}2 \cdot 0{,}01 + 30{,}7 \cdot 0{,}07 + \ldots + 76{,}2 \cdot 0{,}01 = 49{,}485 \approx 49{,}5$$

(результат отличается от предыдущего примерно на 0,1). ◀

Сформулируем основные свойства выборочного среднего $\bar{x}$, рассматривая его как математическое ожидание выборочной величины $\hat{X}$, $\bar{x} = M\hat{X}$, т. е. предполагая, что оно вычисляется непосредственно по ряду наблюдений (см. (7.5)). Доказательства свойств не приводим, поскольку свойства математического ожидания подробно изучались в п. 4.3. [1].

1^0. *Среднее постоянной* (c) *равно этой постоянной*:

$$\bar{c} = c; \tag{7.8}$$

свойство является выборочным аналогом свойства (4.17).

2^0. *Константа* c *выносится за знак среднего*:

$$\overline{cx} = c\bar{x}, \tag{7.9}$$

где $\overline{cx}$ *— среднее из чисел* $cx_1, cx_2, \ldots, cx_n$; свойство является выборочным аналогом свойства (4.18).

3^0. *Если имеется* n_X *наблюдений величины* X: $x_1, x_2, \ldots, x_{n_X}$ *и* n_Y *наблюдений величины* Y: $y_1, y_2, \ldots, y_{n_Y}$, *то*

$$\overline{x + y} = \bar{x} + \bar{y}, \tag{7.10}$$

где $\overline{x + y}$ *— среднее* $n_X n_Y$ *чисел, каждое из которых равно сумме двух чисел — результата наблюдения величины* X *и*

[1] На выборочное среднее $\bar{x} = M\hat{X}$ не распространяются свойства математического ожидания, в которых предполагается независимость случайных величин (см. (4.20) и (4.27)): ведь в определении выборочной величины $\hat{X}$ (см. табл. 7.1), условно названной случайной, фигурируют опытные вероятности, которые зависят от числа наблюдений n, а понятие независимости случайных величин формулируется в терминах истинных вероятностей (см. (4.21)).

результата наблюдения величины Y, *т. е.* $\overline{x+y}$ — *среднее чисел* $x_1+y_1, x_1+y_2, \ldots, x_1+y_{n_Y}, x_2+y_1, x_2+y_2, \ldots, x_2+y_{n_Y}, \ldots, x_{n_X}+y_{n_Y}$.

▶ **ПРИМЕР 7.4.** Пусть наблюдения величины X таковы: 5, 6, 7 ($\bar{x}=6$), а наблюдения Y: 2, 2, 3, 1 ($\bar{y}=2$). Тогда наблюдениями величины $X+Y$ будут $3\cdot 4=12$ чисел: 7, 7, 8, 6, 8, 8, 9, 7, 9, 9, 10, 8. Среднее этих чисел $\overline{x+y}=8$, что равно $\bar{x}+\bar{y}=6+2$. ◀

Свойство (7.10) является выборочным аналогом свойства (4.19).

Основные *следствия* сформулированных свойств таковы:

$$1)\ \overline{x-y}=\bar{x}-\bar{y}, \tag{7.11}$$

где $\overline{x-y}$ — среднее n_Xn_Y чисел, каждое из которых равно разности двух чисел: уменьшаемым является наблюдение величины X (число ее наблюдений равно n_X), а вычитаемым — наблюдение величины Y (число ее наблюдений равно n_Y). Равенство (7.11) — выборочный аналог равенства (4.23).

$$2)\ \overline{a+bx}=a+b\bar{x}, \tag{7.12}$$

где a и b — постоянные числа, а $\overline{a+bx}$ — среднее чисел $a+bx_1, a+bx_2, \ldots, a+bx_n$.

Равенство (7.12) — выборочный аналог равенства (4.24).

3) Если имеется n_1 наблюдений величины X_1: $x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_1}^{(1)}$, среднее которых $\overline{x^{(1)}}$); n_2 наблюдений величины X_2: $x_1^{(2)}, x_2^{(2)}, \ldots, x_{n_2}^{(2)}$, среднее которых $\overline{x^{(2)}}$; ...; n_k наблюдений величины X_k: $x_1^{(k)}, x_2^{(k)}, \ldots, x_{n_k}^{(k)}$, среднее которых $\overline{x^{(k)}}$, то

$$\overline{x^{(1)}+x^{(2)}+\ldots+x^{(k)}}=\overline{x^{(1)}}+\overline{x^{(2)}}+\ldots+\overline{x^{(k)}},$$

или

$$\overline{\sum_{i=1}^{k}x^{(i)}}=\sum_{i=1}^{k}\overline{x^{(i)}}, \tag{7.13}$$

где $\overline{x^{(1)} + x^{(2)} + \ldots + x^{(k)}}$ — среднее $n_1 n_2 \ldots n_k$ чисел, каждое из которых равно сумме k чисел — результатов наблюдений различных случайных величин.

Равенство (7.13) является выборочным аналогом равенства (4.25) и обобщением свойства (7.10).

▶ **ПРИМЕР 7.5.** Пусть наблюдения величины X_1: 5, 6, 7 ($\overline{x^{(1)}} = 6$), наблюдения величины X_2: 2, 2, 3, 1 ($\overline{x^{(2)}} = 2$), а наблюдения величины X_3: 0, −2 ($\overline{x^{(3)}} = -1$). Тогда наблюдениями величины $X_1 + X_2 + X_3$ будут $3 \cdot 4 \cdot 2 = 24$ числа: 7, 5, 7, 5, 8, 6, 6, 4, 8, 6, 8, 6, 9, 7, 7, 5, 9, 7, 9, 7,10, 8, 8, 6. Среднее этих чисел $\overline{x^{(1)} + x^{(2)} + x^{(3)}} = 168/24 = 7$, что равно $\overline{x^{(1)}} + \overline{x^{(2)}} + \overline{x^{(3)}} = 7$. ◀

Выборочная мода $\hat{x}_{\text{mod}}$; способы ее нахождения зависят от формы задания исходных данных.

— Если дан ряд наблюдений, то $\hat{x}_{\text{mod}}$ — наиболее часто встречающееся наблюдение (так она определяется в программе **«Описательная статистика»**).

— Если дан статистический ряд (см. табл. 7.2), то $\hat{x}_{\text{mod}}$ — это вариант, которому соответствует наибольшая частость (частота):

$$\hat{x}_{\text{mod}} = x'_k, \text{ если } \hat{p}_k = \max_{1 \leqslant i \leqslant v} \{\hat{p}_i\}, \qquad (7.14)$$

(ср. (7.14) с (4.29)).

— Если наблюдения сгруппированы в интервальный ряд (см. табл. 7.6), то выборочная мода вычисляется по следующей формуле:

$$\hat{x}_{\text{mod}} = a_{\text{mod}} + h \frac{\hat{p}_{\text{mod}} - \hat{p}_{\text{mod}-1}}{2\hat{p}_{\text{mod}} - \hat{p}_{\text{mod}-1} - \hat{p}_{\text{mod}+1}}, \qquad (7.15)$$

где a_{mod} — начало ***модального интервала***, т. е. такого, которому соответствует наибольшая частость (частота) или наибольшее значение выборочной функции плотности; h — длина интервала группирования; $\hat{p}_{\text{mod}}$ — частость модального интервала; $\hat{p}_{\text{mod}-1}$ — частость интервала, предшествующего модальному; $\hat{p}_{\text{mod}+1}$ — частость интервала, следующего за модальным.

▶ **ПРИМЕР 7.6.** Вычислим моду по данным, приведенным в примерах 7.1 и 7.2.

Данные примера 7.1. Имеем: среди 100 значений числа сданных экзаменов наиболее часто встречается число 4, поэтому $\hat{x}_{\text{mod}} = 4$ (см. рис. 7.5, *а*).

По статистическому ряду (см. табл. 7.4), согласно (7.14), $\hat{x}_{\text{mod}} = x'_5 = 4$, так как $\hat{p}_5 = \max\{\hat{p}_1 = 0{,}01;\ \hat{p}_2 = 0{,}01;\ \hat{p}_3 = 0{,}03;\ \hat{p}_4 = 0{,}35;\ \hat{p}_5 = 0{,}6\}$. Результаты, как и следовало ожидать, совпали.

Данные примера 7.2. Если использовать несгруппированные наблюдения, то $\hat{x}_{\text{mod}} = 37{,}2$ (см. рис. 7.5, *б*), поскольку такой объем продаж был зафиксирован два раза, а остальные — по одному разу.

Если ориентироваться на интервальный ряд (см. табл. 7.9) и использовать формулу (7.15), в которой $a_{\text{mod}} = a_4 = 40{,}45$ (наибольшая частость, равная 0,25, соответствует четвертому интервалу, поэтому этот интервал — модальный); $h = 6{,}5$; $\hat{p}_{\text{mod}} = \hat{p}_4 = 0{,}25$; $\hat{p}_{\text{mod}-1} = \hat{p}_3 = 0{,}12$; $\hat{p}_{\text{mod}+1} = \hat{p}_5 = 0{,}18$, то

$$\hat{x}_{\text{mod}} = 40{,}45 + 6{,}5(0{,}25 - 0{,}12)/(2 \cdot 0{,}25 - 0{,}12 - 0{,}18) = 44{,}675.$$

Полученный результат отличен от предыдущего (37,2), однако он имеет гораздо больше смысла:

— $\hat{x}_{\text{mod}} = 44{,}675$ принадлежит модальному интервалу (40,45; 46,95), которому соответствует наибольшее значение выборочной функции плотности $\hat{f}_X(x)$; это согласуется с определением моды x_{mod} непрерывной случайной величины X, наблюдения которой обычно и группируются в интервальный ряд;

— рассчитываемая по формуле (7.15) мода $\hat{x}_{\text{mod}}$ реагирует на значение частостей интервалов, соседних с модальными; в рассматриваемом примере частость $\hat{p}_5$ интервала, следующего за модальным, больше частости $\hat{p}_3$ интервала, предшествующего модальному ($\hat{p}_5 = 0{,}18 > \hat{p}_3 = 0{,}12$) и $\hat{x}_{\text{mod}} = 44{,}675$ «сдвинута» к правому концу модального интервала. ◀

Выборочная медиана $\hat{x}_{\text{med}}$; способ ее нахождения зависит от формы задания исходных данных.

— Если дан ряд наблюдений, то $\hat{x}_{\text{med}}$ — число, приходящееся на середину вариационного ряда (7.1), — это ряд наблюдений, расположенных в неубывающем порядке. При нечетном числе наблюдений $n = 2q - 1$ на середину вариационного ряда $x_{(1)}, x_{(2)}, \ldots, x_{(q-1)}, x_{(q)}, x_{(q+1)}, \ldots, x_{(n)}$ приходится наблюдение $x_{(q)}$, следовательно,

$$\hat{x}_{\text{med}} = x_{(q)}, \text{ если } n = 2q - 1. \qquad (7.16)$$

При четном числе наблюдений $n = 2q$ на середину вариационного ряда $x_{(1)}, x_{(2)}, \ldots, x_{(q)}, x_{(q+1)}, \ldots, x_{(n)}$ приходятся наблюдения $x_{(q)}$ и $x_{(q+1)}$ и

$$\hat{x}_{\text{med}} = (x_{(q)} + x_{(q+1)})/2, \text{ если } n = 2q. \qquad (7.17)$$

По такому алгоритму рассчитывается $\hat{x}_{\text{med}}$ в программе **«Описательная статистика»**.

— Если наблюдения сгруппированы в статистический или интервальный ряд (см. табл. 7.2 и 7.6), то за медиану принимают число $\hat{x}_{\text{med}}$, при котором

$$\hat{P}(X \leqslant \hat{x}_{\text{med}}) = 0{,}5; \qquad (7.18)$$

найденная таким образом $\hat{x}_{\text{med}}$ имеет гораздо больший смысл по сравнению с $\hat{x}_{\text{med}}$, найденной по формулам (7.16) или (7.17). Кроме того, формула (7.18) является выборочным аналогом формулы (4.30) (и (4.33)).

▶ **ПРИМЕР 7.7.** Вычислим медиану по данным, приведенным в примерах 7.1 и 7.2.

Данные примера 7.1. Расположим сведения о числе экзаменов, сданных 100 студентами, в вариационный ряд

$$(x_{(1)} = 0) < (x_{(2)} = 1) < (x_{(3)} = 2) = (x_{(4)} = 2) = (x_{(5)} = 2) < \\ < (x_{(6)} = 3) = \ldots = (x_{(40)} = 3) < (x_{(41)} = 4) = \ldots = (x_{(100)} = 4).$$

Так как $n = 100 = 2 \cdot 50$, то, согласно (7.17), $q = 50$ и $\hat{x}_{\text{med}} = (x_{(50)} + x_{(51)})/2 = (4 + 4)/2 = 4$ (см. рис. 7.5, *а*).

Теперь воспользуемся статистическим рядом (см. табл. 7.4) и для каждого варианта найдем накопленную частость:

x_i'	$x_1' = 0$	$x_2' = 1$	$x_3' = 2$	$x_4' = 3$	$x_5' = 4$
$\hat{p}_i^{\text{нак}}$	$\hat{p}_1^{\text{нак}} = 0{,}01$	$\hat{p}_2^{\text{нак}} = 0{,}02$	$\hat{p}_3^{\text{нак}} = 0{,}05$	$\hat{p}_4^{\text{нак}} = 0{,}4$	$\hat{p}_5^{\text{нак}} = 1$

Интервал $[x_l', x_{l+1}')$, в левом конце которого $\hat{p}_l^{\text{нак}} \leqslant 0{,}5$, а в правом $\hat{p}_{l+1}^{\text{нак}} > 0{,}5$, называют ***медианным***; таким интервалом является интервал $[x_4', x_5') = [3, 4)$. Предположив линейный характер накопления частостей в медианном интервале, найдем в нем точку $\hat{x}_{\text{med}}$, которая удовлетворяет равенству (7.18), по форме, аналогичной (4.31):

$$\hat{x}_{\text{med}} = x_l' + \frac{x_{l+1}' - x_l'}{\hat{p}_{l+1}}(0{,}5 - \hat{p}_l^{\text{нак}}). \qquad (7.19)$$

В рассматриваемом примере $l = 4$, $x'_l = x'_4 = 3$, $x'_{l+1} = x'_5 = 4$, $\hat{p}_{l+1} = \hat{p}_5 = 0{,}6$ (см. табл. 7.4), $\hat{p}_l^{\text{нак}} = \hat{p}_4^{\text{нак}} = 0{,}4$ и

$$\hat{x}_{\text{med}} = 3 + \frac{4-3}{0{,}6}(0{,}5 - 0{,}4) = 3{,}17$$

(половина наблюдений меньше или равна числу 3,17, а половина — больше этого числа).

Данные примера 7.2. Расположим сведения об объеме продаж за 100 дней в вариационный ряд:

$$x_{(1)} = 24{,}2 < x_{(2)} = 28{,}2 < \ldots < x_{(100)} = 73{,}8.$$

На середину ряда придутся значения $x_{(50)} = 49{,}1$ и $x_{(51)} = 49{,}2$ и $\hat{x}_{\text{med}} = (49{,}1 + 49{,}2)/2 = 49{,}15$ (см. рис. 7.5, *б*).

Теперь воспользуемся интервальным рядом (см. табл. 7.9). Медианным интервалом является 5-й интервал $[a_5, a_6) = [46{,}95;\ 53{,}45)$: в его левом конце $\hat{F}_X(46{,}95) = \hat{P}(X < 46{,}95) = 0{,}45 < 0{,}5$, а в правом $\hat{F}_X(53{,}45) = \hat{P}(X < 53{,}45) = 0{,}63 > 0{,}5$ (см. столбец 7), следовательно, внутри этого интервала находится точка $\hat{x}_{\text{med}}$, для которой выполняется соотношение (7.18); $\hat{x}_{\text{med}}$ находится по формуле

$$\hat{x}_{\text{med}} = a_{\text{med}} + \frac{h}{\hat{p}_{\text{med}}}(0{,}5 - \hat{F}_X(a_{\text{med}})), \qquad (7.20)$$

где a_{med} — начало медианного интервала, т. е. такого интервала $[a_{\text{med}}, a_{\text{med}+1})$, что $\hat{F}_X(a_{\text{med}}) \leqslant 0{,}5$, а $\hat{F}_X(a_{\text{med}+1}) > 0{,}5$; $\hat{p}_{\text{med}}$ — частость медианного интервала.

В рассматриваемом примере медианным будет интервал $[a_5, a_6) = [46{,}95;\ 53{,}45)$, $a_{\text{med}} = a_5 = 46{,}95$; $h = 6{,}5$, $\hat{p}_{\text{med}} = \hat{p}_5 = 0{,}18$, $\hat{F}_X(a_{\text{med}}) = \hat{F}_X(a_5) = \hat{F}_X(46{,}95) = 0{,}45$ (см. столбец 7 табл. 7.9) и $\hat{x}_{\text{med}} = 46{,}95 + 6{,}5(0{,}5 - 0{,}45)/0{,}18 = 48{,}75(5)$ (в 50 из 100 дней объем продаж меньше или равен 48,75 (ден. ед.), а в 50 — больше 48,75). ◂

Выборочные характеристики рассеивания

Дисперсия выборочной «случайной» величины $\hat{X}$, ряд распределения которой задан таблицей (7.1), вычисляется по одной из двух тождественных формул (4.42) или (4.44). Имеем

$$D\hat{X} \underset{(4.42)}{=} \sum_{i=1}^{n}(x_{(i)} - M\hat{X})^2 p_i = \sum_{i=1}^{n}(x_{(i)} - \bar{x})^2 \frac{1}{n} \underset{(*)}{=} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 =$$

$$= \overline{(x - \bar{x})^2},$$

$$D\hat{X} \underset{(4.44)}{=} \sum_{i=1}^{n} x_{(i)}^2 p_i - (M\hat{X})^2 = \sum_{i=1}^{n} x_{(i)}^2 \frac{1}{n} - (\bar{x})^2 \underset{(*)}{=}$$

$$\underset{(*)}{=} \sum_{i=1}^{n} x_i^2 \frac{1}{n} - (\bar{x})^2 = \overline{x^2} - (\bar{x})^2,$$

где $M\hat{X} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. В равенствах, отмеченных (*), символ $x_{(i)}$ заменен на x_i; это не скажется на результатах вычисления $D\hat{X}$.

Дисперсию $D\hat{X}$ обычно обозначают символом $\hat{D}X$ и называют ***выборочной дисперсией***. Итак,

$$\hat{D}X = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \overline{(x - \bar{x})^2} \qquad (7.21)$$

(читается: «выборочная дисперсия равна среднему из квадратов отклонений наблюдений от их среднего»), или

$$\hat{D}X = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \overline{x^2} - (\bar{x})^2 \qquad (7.22)$$

(читается «выборочная дисперсия равна разности между средним из квадратов наблюдений и квадратом среднего»). Здесь $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2$; $x_1, x_2, \ldots, x_n$ — результаты n наблюдений величины X.

Замечания. 1. $\hat{D}X$ не определяется как среднее из отклонений наблюдений от их среднего, так как, учитывая, что $\bar{x}$ — постоянная величина, получим

$$\overline{x - \bar{x}} \underset{(7.11)}{=} \bar{x} - \bar{\bar{x}} \underset{(7.8)}{=} \bar{x} - \bar{x} = 0.$$

2. Нетрудно убедиться в том, что если в формуле (7.21) $\bar{x}$ заменить на любую постоянную величину $c \neq \bar{x}$, то $\overline{(x - c)^2} > \overline{(x - \bar{x})^2}$; говорят, что $\hat{D}X$, определяемая формулой (7.21), обладает свойством минимальности.

Если наблюдения сгруппированы в статистический (интервальный ряд), то выборочная дисперсия

$$\hat{D}X = \sum_{i=1}^{\nu} (x_i' - \overline{x'})^2 \hat{p}_i = \overline{(x' - \bar{x}')^2}; \qquad (7.23)$$

$$\hat{D}X = \sum_{i=1}^{\nu} (x_i')^2 \hat{p}_i - (\overline{x'})^2 = \overline{(x')^2} - (\overline{x'})^2, \qquad (7.24)$$

где x_i' — варианты статистического ряда (середины интервалов интервального ряда); $\hat{p}_i$ — частости вариантов (интервальные частости); ν — число вариантов (интервалов); $\overline{x'} = \sum_{i=1}^{\nu} x_i' \cdot \hat{p}_i$ (см. формулы (7.6) и (7.7)).

Вычисляя дисперсию по ряду наблюдений и по статистическому ряду, получаем один и тот же результат. Результаты вычисления дисперсии по ряду наблюдений и по интервальному ряду отличаются. Чтобы их «сблизить», обычно из дисперсии, рассчитанной по интервальному ряду, вычитают «поправку В. Шеппарда» (начало XIX в.), равную $h^2/12$, где h — длина интервала.

Программа «**Описательная статистика**» работает с рядом наблюдений величины X, но вычисляет не выборочную дисперсию, определяемую формулой (7.21) (или (7.22)), а «*исправленную выборочную дисперсию*», называя ее «*дисперсией выборки*» (см. рис. 7.5). Будем обозначать эту дисперсию s_X^2; она связана с $\hat{D}X$, рассчитанной по формуле (7.21) (или (7.22)), соотношением

$$s_X^2 = \hat{D}X \cdot n/(n-1). \qquad (7.25)$$

Отсюда получаем

$$\hat{D}X = s_X^2 (n-1)/n. \qquad (7.26)$$

Отметим, что s_X^2 дает «лучшее» представление о дисперсии DX случайной величины X, чем $\hat{D}X$.

▶ **ПРИМЕР 7.8.** Вычислим дисперсию по данным, приведенным в примере 7.2. Для несгруппированных наблюдений $\bar{x} = 49{,}596$ (см. рис. 7.5, *б*) и $\hat{D}X \underset{(7.22)}{=} (47{,}0^2 + 37{,}2^2 + \ldots + 56{,}4^2)/100 - 49{,}596^2 = 116{,}6$. Такой же результат получим, воспользовавшись формулой (7.26), в которой $s_X^2 = 117{,}8$ (см. рис. 7.5, *б*, «*дисперсия выборки*»): $\hat{D}X = 117{,}8 \cdot 99/100 = 116{,}6$.

Для наблюдений, сгруппированных в интервальный ряд, $\overline{x'} = 49{,}485$ (см. пример 7.3) и $\hat{D}X \underset{(7.22)}{=} 24{,}2^2 \cdot 0{,}01 + \ldots + 76{,}2^2 \cdot 0{,}01 - 49{,}485^2 = 120{,}746$. Скорректировав этот результат на поправку Шеппарда, получим $\hat{D}X_{(\text{Ш})} = 120{,}746 - 6{,}5^2/12 = 117{,}2$.

Сравним значения дисперсий:

Несгруппированные наблюдения	$s_X^2 = 117{,}8$	$\hat{D}X = 116{,}6$
Наблюдения, сгруппированные в интервальный ряд	$\hat{D}X_{(\text{Ш})} = 117{,}2$	$\hat{D}X = 120{,}7$

Скорректированная на поправку Шеппарда дисперсия $\hat{D}X_{(\text{Ш})}$ отличается от $\hat{D}X = 116{,}6$, рассчитанной по несгруппированным наблюдениям, меньше, чем нескорректированная дисперсия $\hat{D}X = 120{,}7$; кроме того, $\hat{D}X_{(\text{Ш})}$ меньше всего отличается от «исправленной» выборочной дисперсии s_X^2. ◀|

Дисперсия $\hat{D}X$ — выборочный аналог дисперсии DX случайной величины X (или выборочный аналог генеральной дисперсии); выборочное среднее квадратическое отклонение

$$\hat{\sigma}_X = \sqrt{\hat{D}X} \tag{7.27}$$

является выборочным аналогом среднего квадратического отклонения σ_X величины X.

Выборочные дисперсия $\hat{D}X$ и среднее квадратическое отклонение $\hat{\sigma}_X$ характеризуют средний разброс наблюдений вокруг выборочного среднего. Сформулируем их основные свойства. Они аналогичны рассмотренным в п. 4.3.2 свойствам генеральной дисперсии DX и генерального среднего квадратического отклонения σ_X[1].

1^0. *Выборочная дисперсия (выборочное среднее квадратическое отклонение) постоянной с равно нулю*:

$$\hat{D}c = 0 \ (\hat{\sigma}_c = 0). \tag{7.28}$$

Свойство является выборочным аналогом свойств (4.46) (и (4.47)).

2^0. *Константу с выносят за знак выборочной дисперсии в квадрате*:

$$\hat{D}(cX) = c^2 \hat{D}X, \tag{7.29}$$

где $\hat{D}(cX)$ — *дисперсия ряда чисел* cx_1, cx_2, ..., cx_n, а x_1, x_2, ..., x_n — *наблюдения величины* X. Учитывая (7.27), получим

$$\hat{\sigma}_{cX} = \sqrt{\hat{D}(cX)} = \sqrt{c^2 \hat{D}X} = |c| \hat{\sigma}_X . \tag{7.30}$$

Свойства (7.29) и (7.30) являются выборочными аналогами свойств (4.48) и (4.49).

---

[1] Выборочная дисперсия $\hat{D}X$ не обладает свойствами, аналогичными свойствам дисперсии DX, в которых предполагается независимость случайных величин [см. формулы (4.50), (4.58)—(4.60)], поскольку понятие независимости формулируется в терминах истинных, а не опытных вероятностей.

3^0. *Если имеется* n *парных наблюдений* (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), *т. е. с каждого из* n *объектов «сняты» значения двумерной величины* (X, Y), *то дисперсия* $\hat{D}(X+Y)$ *ряда чисел* x_1+y_1, x_2+y_2, ..., x_n+y_n

$$\hat{D}(X+Y) = \hat{D}X + \hat{D}Y + 2\hat{r}_{X,Y}\hat{\sigma}_X\hat{\sigma}_Y, \qquad (7.31)$$

где

$$\hat{r}_{X,Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\overline{(x-\bar{x})(y-\bar{y})}}{\hat{\sigma}_X\hat{\sigma}_Y} \qquad (7.32)$$

— ***выборочный коэффициент корреляции*** (*выборочный аналог коэффициента корреляции* (4.53)).

▶ **ПРИМЕР 7.9.** Результаты парных наблюдений величины (X, Y) приведены в таблице 7.11; там же приведены и некоторые промежуточные расчеты.

Таблица 7.11

i	x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})\times(y_i-\bar{y})$	x_i+y_i	$((x_i+y_i)-\overline{x+y})^2$
1	5	3	−1	−1	1	1	1	8	4
2	6	3	0	−1	0	1	0	9	1
$n=3$	7	6	1	2	1	4	2	13	9
Σ	18	12	0	0	2	6	3	30	14

Из таблицы 7.11 получаем: $\bar{x} = 18/3 = 6$, $\bar{y} = 12/3 = 4$, $\hat{D}X = \overline{(x-\bar{x})^2} = 2/3$, $\hat{D}Y = \overline{(y-\bar{y})^2} = 6/3 = 2$, $\overline{(x-\bar{x})(y-\bar{y})} = 3/3 = 1$. Тогда $\hat{\sigma}_X = \sqrt{2/3}$, $\hat{\sigma}_Y = \sqrt{2}$ и, согласно (7.32), $\hat{r}_{X,Y} = 1/(\sqrt{2/3}\cdot\sqrt{2}) = \sqrt{3}/2$. Правая часть равенства (7.31) равна

$$\hat{D}X + \hat{D}Y + 2\hat{r}_{X,Y}\hat{\sigma}_X\hat{\sigma}_Y = \frac{2}{3} + 2 + 2\frac{\sqrt{3}}{2}\sqrt{2/3}\sqrt{2} = 4\frac{2}{3}.$$

Вычислим его левую часть. Наблюдениями величины $X+Y$ являются три числа: $8 = 5+3$, $9 = 6+3$, $13 = 7+6$, поскольку пара чисел (x, y) «снята» с одного объекта. Среднее $\overline{x+y} = 10$, дисперсия

$$\hat{D}(X+Y) = \overline{((x+y) - \overline{x+y})^2} = 14/3 = 4\frac{2}{3}.$$

Левая и правая части равенства (7.31) одинаковы. ◂

Свойство (7.31), имеющее место только при парных наблюдениях $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, «снятых» с n объектов, является выборочным аналогом свойства (4.52).

Замечание. Выборочный коэффициент корреляции $\hat{r}_{X,Y}$ чаще вычисляют не по формуле (7.32), а по формуле

$$\hat{r}_{X,Y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\hat{\sigma}_X \hat{\sigma}_Y}. \qquad (7.33)$$

Убедимся в тождественности формул (7.32) и (7.33). Для этого докажем равенство их числителей.

≫ Имеем

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\,y_i - \bar{y}\,x_i + \bar{x}\,\bar{y}) =$$

$$= \frac{1}{n}\Big(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}\,\bar{y}\Big) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\frac{1}{n}\sum_{i=1}^{n} y_i - \bar{y}\,\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\,n\bar{x}\,\bar{y} =$$

$$= \overline{xy} - \bar{x}\,\bar{y} - \bar{y}\,\bar{x} + \bar{x}\,\bar{y} = \overline{xy} - \bar{x}\,\bar{y}. \quad ≪$$

Приведем основные следствия сформулированных свойств.

1) $\hat{D}(-X) = \hat{D}X$, $\hat{\sigma}_{-X} = \hat{\sigma}_X$ [ср. с (4.56)]; (7.34)

2) $\hat{D}(a+bX) = b^2\hat{D}X$, $\hat{\sigma}_{a+bX} = |b|\hat{\sigma}_X$, (7.35)

где a и b — постоянные числа, а $\hat{D}(a+bX)$ — *дисперсия ряда чисел* $a+bx_1, a+bx_2, \ldots, a+bx_n$ [ср. с (4.57)];

3) если $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ — *парные наблюдения двумерной величины* (X, Y), *то дисперсия* $\hat{D}(X-Y)$ *ряда чисел* $(x_1-y_1), (x_2-y_2), \ldots, (x_n-y_n)$

$$\hat{D}(X-Y) = \hat{D}X + \hat{D}Y - 2\hat{r}_{X,Y}\hat{\sigma}_X\hat{\sigma}_Y \text{ [ср. с (4.61)]}. \qquad (7.36)$$

Выборочный коэффициент вариации случайной величины X

$$\hat{V}_X = \frac{\hat{\sigma}_X}{\bar{x}} \qquad (7.37)$$

является характеристикой рассеивания наблюдений величины X (около $\bar{x}$), сопоставленного со средним. Чем меньше $\hat{V}_X$, тем более «представительным» является выборочное среднее (в смысле замены наблюдений средним).

Для нормально распределенной величины $N(a, \sigma)$, практически все значения которой положительные, коэффициент вариации $V_{N(a,\sigma)} = \sigma/a < 0{,}3(3)$. Поэтому, если рассчитанный по наблюдениям положительной величины X коэффициент вариации $\hat{V}_X < 0{,}3(3)$, то это может служить основанием (но не достаточным) для предположения о нормальности распределения величины X.

Выборочные моменты и коэффициенты асимметрии и эксцесса

Выборочный начальный момент k-го порядка $\hat{\nu}_k(X)$ случайной величины X является выборочным аналогом начального момента k-го порядка $\nu_k(X)$ (см. (4.65)). Способ его вычисления зависит от формы задания исходных данных:

— для несгруппированных наблюдений $x_1, x_2, \ldots, x_n$

$$\hat{\nu}_k(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^k = \overline{x^k}; \qquad (7.38)$$

— для наблюдений, сгруппированных в статистический (или интервальный) ряд,

$$\hat{\nu}_k(X) = \sum_{i=1}^{\nu} (x_i')^k \hat{p}_i = \overline{(x')^k}, \qquad (7.39)$$

где ν — число групп (интервалов), x_i' — варианты (центры интервалов), $\hat{p}_i$ — частости вариантов (интервальные частости).

Выборочный центральный момент k-го порядка $\hat{\mu}_k(X)$ случайной величины X является выборочным аналогом центрального момента k-го порядка $\mu_k(X)$ [см. (4.66)]. Способ его вычисления зависит от формы задания исходных данных:

— для несгруппированных наблюдений $x_1, x_2, \ldots, x_n$

$$\hat{\mu}_k(X) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k = \overline{(x - \bar{x})^k}, \qquad (7.40)$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$;

— для наблюдений, сгруппированных в статистический (или интервальный) ряд

$$\hat{\mu}_k(X) = \sum_{i=1}^{v} (x'_i - \overline{x'})^k \hat{p}_i = \overline{(x' - \overline{x'})^k}, \qquad (7.41)$$

где v — число групп (интервалов), x'_i — варианты (центры интервалов), $\hat{p}_i$ — частости вариантов (интервальные частости), $\overline{x'} = \sum_{i=1}^{v} x'_i \hat{p}_i$.

Для выборочных начальных и центральных моментов (рассчитанных по несгруппированным наблюдениям) имеют место соотношения, приведенные в табл. 7.12 (аналогичные соотношения справедливы и для моментов, рассчитанных по сгруппированным данным). Они являются выборочными аналогами соотношений, приведенных в таблице 4.9.

Таблица 7.12

k	$\hat{\nu}_k(X)$	$\hat{\mu}_k(X)$	
0	$\hat{\nu}_0 = \overline{x^0} = 1$	$\hat{\mu}_0 = \overline{(x - \bar{x})^0} = 1$	
1	$\hat{\nu}_1 = \bar{x}$	$\hat{\mu}_1 = \overline{(x - \bar{x})^1} = 0$	(7.42)
2	$\hat{\nu}_2 = \overline{x^2}$	$\hat{\mu}_2 = \underbrace{\overline{(x - \bar{x})^2}}_{\hat{D}X} = \overline{x^2} - (\bar{x})^2 = \hat{\nu}_2 - \hat{\nu}_1^2$	(7.43)
3	$\hat{\nu}_3 = \overline{x^3}$	$\hat{\mu}_3 = \overline{(x - \bar{x})^3} = \hat{\nu}_3 - 3\hat{\nu}_2\hat{\nu}_1 + 2\hat{\nu}_1^3$	(7.44)
4	$\hat{\nu}_4 = \overline{x^4}$	$\hat{\mu}_4 = \overline{(x - \bar{x})^4} = \hat{\nu}_4 - 4\hat{\nu}_3\hat{\nu}_1 + 6\hat{\nu}_2\hat{\nu}_1^2 - 3\hat{\nu}_1^4$	(7.45)

Доказательство соотношений (7.42)—(7.45) основывается на свойствах выборочного среднего.

≫ Например, для (7.43) имеем

$$\hat{\mu}_2 = \overline{(x - \bar{x})^2} = \overline{x^2 - 2\bar{x}x + (\bar{x})^2} \underset{(7.13)}{=}$$

$$= \overline{x^2} + \overline{(-2\bar{x}x)} + \overline{(\bar{x})^2} \underset{\substack{(7.9)\\(7.8)}}{=} \overline{(x)^2} - 2\bar{x}\,\overline{(x)} + (\bar{x})^2 =$$

$$= \overline{(x)^2} - 2(\bar{x})^2 + (\bar{x})^2 = \overline{x^2} - \bar{x}^2 = \hat{\nu}_2 - \hat{\nu}_1^2 . \quad \ll$$

Аналогично доказываются и другие соотношения таблицы 7.12.

Для выборочного центрального момента k-го порядка имеет место следующее свойство:

$$\hat{\mu}_k(a + bX) = b^k \hat{\mu}_k(X), \qquad (7.46)$$

где a и b — постоянные величины [ср. с (4.70)].

Выборочный коэффициент асимметрии

$$\hat{A}_X = \hat{\mu}_3(X)/\hat{\sigma}_X^3; \qquad (7.47)$$

выборочный коэффициент эксцесса

$$\hat{E}_X = \hat{\mu}_4(X)/\hat{\sigma}_X^4 - 3. \qquad (7.48)$$

[Ср. (7.47) и (7.48) соответственно с (4.71) и (4.72).]

Свойства выборочных коэффициентов асимметрии и эксцесса аналогичны свойствам (4.73) и (4.74) генеральных коэффициентов:

$$\hat{A}_{a+bX} = \begin{cases} \hat{A}_X, & \text{если } b > 0, \\ -\hat{A}_X, & \text{если } b < 0; \end{cases} \qquad (7.49)$$

$$\hat{E}_{a+bX} = \hat{E}_X, \qquad (7.50)$$

где a и b — постоянные величины.

Лучшие приближения к значениям генеральных коэффициентов асимметрии и эксцесса (A_X и E_X) случайной величины X (особенно при нормальном распределении) дают не выборочные коэффициенты ($\hat{A}_X$ и $\hat{E}_X$), а связанные с ними «исправленные» выборочные коэффициенты асимметрии и эксцесса:

$$\hat{A}_X^{\text{исп}} = \hat{A}_X \sqrt{n(n-1)}\,/(n-2); \qquad (7.51)$$

$$\hat{E}_X^{\text{исп}} = (n-1)[(n+1)\hat{E}_X + 6]/[(n-2)(n-3)]. \qquad (7.52)$$

Именно эти коэффициенты вычисляет по несгруппированным данным программа **«Описательная статистика»** пакета **«Анализ данных»** Microsoft Excel.

Для нормально распределенной величины $N(a, \sigma)$ асимметрия и эксцесс равны нулю: $A_{N(a,\sigma)} = E_{N(a,\sigma)} = 0$. Поэтому близость к нулю коэффициентов $\hat{A}_X^{\text{исп}}$ и $\hat{E}_X^{\text{исп}}$ может служить основанием (но не достаточным) для предположения о нормальности распределения величины X.

▶ **ПРИМЕР 7.10.** По результатам наблюдений $x_1, x_2, \ldots, x_n$ величины X вычислены следующие характеристики: $\bar{x} = 15$, $\hat{x}_{\text{med}} = 10$, $\hat{\sigma}_X^2 = 4$, $\hat{\mu}_3(X) = 8$. Найдем значения соответствующих характеристик и асимметрию для ряда чисел $y_i = -3x_i + 5$, $i = 1, 2, \ldots, n$. Имеем

$$\bar{y} = \overline{-3x + 5} \underset{(7.12)}{=} 5 - 3\bar{x} = 5 - 3 \cdot 15 = -40;$$

$$\hat{y}_{\text{med}} = -3\hat{x}_{\text{med}} + 5 = -3 \cdot 10 + 5 = -25;$$

$$\hat{\sigma}_Y^2 = \hat{D}_Y = \hat{D}(-3X + 5) \underset{(7.35)}{=} (-3)^2 \hat{D}X = 9\hat{\sigma}_X^2 = 9 \cdot 4 = 36;$$

$$\hat{\mu}_3(Y) = \hat{\mu}_3(-3X + 5) \underset{(7.46)}{=} (-3)^3 \hat{\mu}_3(X) = -27 \cdot 8 = -216.$$

Асимметрию $\hat{A}_Y$ вычислим двумя способами:

1) $\hat{A}_Y \underset{(7.47)}{=} \hat{\mu}_3(Y)/\hat{\sigma}_Y^3 = -216/(\sqrt{36})^3 = -1;$

2) $\hat{A}_Y = \hat{A}_{-3X+5} \underset{(7.49)}{=} -\hat{A}_X = -\hat{\mu}_3(X)/\hat{\sigma}_X^3 = -8/2^3 = -1.$ ◀

§ 7.3. Примеры сглаживания выборочных распределений

Результаты первичной обработки выборочных данных (графические изображения выборочных рядов распределения, значения выборочных характеристик), дополненные сведениями о природе изучаемой случайной величины и механизме формирования ее значений, зачастую достаточны для того, чтобы сформулировать гипотезу о модели закона распределения случайной величины: нормальный ли этот закон, биноминальный или какой-либо другой. Используя наблюдения, можно найти числовые значения выборочных аналогов тех параметров предполагаемой модели, истинные (генеральные) значения которых неизвестны; а затем, заменив неизвестные значения параметров их выборочными значениями, рассчитать вероятности появления тех или иных выборочных данных. Такую процедуру называют ***сглаживанием выборочного распределения***. Приведем три примера этой процедуры.

▶ **ПРИМЕР 7.11.** В примере 7.1 были приведены сведения по 100 студентам о числе сданных каждым студентом экзаменов из четырех сдаваемых и построен статистический ряд распределения — ряд распределения относительных частот, или частостей, $\hat{p}_i$ между числами x_i' сданных экзаменов (см. табл. 7.4, строки 2 и 4).

Сформулируем гипотезу о модели закона распределения случайной величины X — числа сданных экзаменов из четырех сдаваемых. Процесс сдачи четырех экзаменов представим как четыре испытания, относительно которых сделаем следующие допущения:

— испытания независимы, т. е. вероятность сдачи любым студентом любого экзамена не зависит от того, будет сдано или нет любое количество других экзаменов;

— вероятность сдачи студентом любого отдельно взятого экзамена одна и та же и равна p, а вероятность «не сдачи» равна $(1 - p)$.

Конечно, эти допущения могут вызвать сомнения, но, возможно, они не будут противоречить результатам наблюдений. При этих допущениях мы имеем дело с испытаниями Бернулли, и число сданных экзаменов из четырех сдаваемых имеет биноминальный закон распределения, т. е. вероятность того, что студент сдаст x экзаменов, равна

$$P(X = x) = C_4^x p^x (1-p)^{4-x}, \qquad x = 0, 1, 2, 3, 4. \quad (7.53)$$

Найдем оценку параметра p, входящего в модель (7.53). В рассматриваемом примере p — вероятность того, что студент сдаст экзамен; частость $\hat{p}$ этого события, учитывая, что имеются сведения об успеваемости 100 студентов, вычислим следующим образом:

$$\hat{p} = \frac{\text{число экзаменов, сданных 100 студентами}}{\text{число экзаменов, сдаваемых 100 студентами}} = \frac{\sum_{i=1}^{5} x_i' m_i}{4 \cdot 100} =$$

$$= \frac{0 \cdot 1 + 1 \cdot 1 + 2 \cdot 3 + 3 \cdot 35 + 4 \cdot 60}{100 \cdot 4} = 0{,}88.$$

Так как $\sum_{i=1}^{5} x_i' m_i / 100 = \overline{x'} = 3{,}52$ (см. рис. 7.5, *а*) — это среднее число экзаменов, сданных одним студентом, то $\hat{p}$ можно было бы определить и так:

$$\hat{p} = \frac{\text{среднее число экзаменов, сданных одним студентом}}{\text{число экзаменов, сдаваемых одним студентом}} =$$

$$= \frac{\overline{x'}}{4} = 0{,}88.$$

Подставив в модель (7.53) вместо неизвестного значения параметра p значение $\hat{p} = 0{,}88$, рассчитаем «биномиальные» вероятности сдачи x' экзаменов из четырех сдаваемых по формуле

$$P(X = x') = C_4^{x'} \cdot 0{,}88^{x'} \cdot 0{,}12^{4-x'}, \; x' = 0, 1, 2, 3, 4$$

(эти вероятности приведены в таблице 7.4). Обратим внимание на то, что сумма биномиальных вероятностей точно равна 1: множество значений величины X — числа сданных экзаменов из четырех сдаваемых состоит из чисел 0, 1, 2, 3, 4, и они все были зафиксированы в наблюдениях.

Многоугольник распределения биномиальных вероятностей изображен на рисунке 7.1, *а* пунктирной линией. Сравнив опытные вероятности (частости) $\hat{p}_i$ с биномиальными p_i, можно убедиться в том, что различия между ними невелики и на основании этого сделать предварительное за-

ключение о приемлемости биномиальной модели. Более серьезное обоснование приемлемости биномиальной модели для распределения числа сданных экзаменов среди четырех сдаваемых дается в § 9.6 (см. пример 9.1).

ПРИМЕР 7.12 (задача Л. Борткевича). В таблице 7.13 приведены ставшие классическими данные польского исследователя Борткевича о числе лиц, убитых ударом копыта в 10 прусских армейских корпусах за 20 лет (1875—1894).

Таблица 7.13

i	1	2	3	4	5	
Число смертей в одном корпусе за год (x_i')	0	1	2	3	4	
Число случаев (m_i)	109	65	22	3	1	$n = 200$
$\hat{p}_i = m_i/n$	0,545	0,325	0,110	0,015	0,005	$\Sigma = 1$
$p_i = \frac{0{,}61^{x_i'}}{x_i'!} e^{-0{,}61}$	0,543	0,331	0,101	0,020	0,003	$\Sigma = 0{,}998$

В пользу гипотезы о пуассоновском законе распределения случайной величины X — числа убитых ударом копыта в армейском корпусе за год говорит следующее:

— вероятность гибели единичного лица от удара копыта, конечно, мала, тогда как численность армейского корпуса велика — это характерные условия пуассоновского закона;

— практически одинаковы выборочное среднее ($\bar{x} = 0{,}61$) и выборочная дисперсия ($\hat{\sigma}_X^2 = 0{,}6079$); напомним, в пуассоновском законе $MX = DX$.

Итак, предположим, что имеет место закон Пуассона, т. е.

$$P(X = x) = \frac{\lambda^x}{x!} e^{-a}, \qquad x = 0, 1, 2, \ldots .$$

Неизвестное значение параметра λ (напомним, $\lambda = MX$) заменим значением выборочного среднего $\bar{x} = 0{,}61$. Рассчитанные по формуле

$$P(X = x') = \frac{0{,}61^{x'}}{x'!} e^{-0{,}61}$$

вероятности при $x' = 0, 1, 2, 3, 4$ приведены в последней строке таблицы 7.12. Обратим внимание на то, что их сум-

ма не равна единице: множество значений пуассоновской величины бесконечно — это числа 0, 1, 2, ... , а в таблице 7.13 вычислены пуассоновские вероятности только для чисел, не больших четырех. Различия между опытными вероятностями $\hat{p}_i$ и пуассоновскими p_i малы. Более серьезное обоснование приемлемости пуассоновской модели дано в § 9.6 (см. пример 9.2).

ПРИМЕР 7.13. В примере 7.2 приведены ежедневные сведения об объеме продаж товара дилером за 100 дней, построен интервальный ряд распределения (см. табл. 7.9, столбцы 2 и 5) и рассчитаны значения $\hat{f}_X(x_i')$ выборочной функции плотности в серединах x_i' интервалов.

Сформулируем гипотезу о модели закона распределения случайной величины X — объеме продаж товара за день. В пользу гипотезы о нормальном законе распределения говорит следующее:

— график выборочной функции плотности — полигон (см. рис. 7.4, *а*, ломаная линия) имеет форму, достаточно близкую к куполообразной кривой нормального распределения (см. рис. 5.5, *а*);

— числовые значения «исправленных» выборочных коэффициентов асимметрии и эксцесса (см. формулы (7.51) и (7.52)), приведенные в результатах работы программы **«Описательная статистика»**, не сильно отличаются от нуля: $\hat{A}_X^{\text{исп}} = 0{,}091$, $\hat{E}_X^{\text{исп}} = -0{,}47$ (см. рис. 7.5, *б*); напомним, для нормально распределенной величины коэффициенты асимметрии и эксцесса равны нулю;

— выборочное среднее $\bar{x} = 49{,}6$, «исправленная» выборочная дисперсия (дисперсия выборки) $s_X^2 = 117{,}8$ (см. рис. 7.5, *б*), выборочная дисперсия $\hat{\sigma}_X^2 \underset{(7.26)}{=} s^2(n-1)/n = 117{,}8 \cdot 99/100 = 116{,}6$, а $\hat{\sigma}_X = 10{,}8$. Отсюда получим выборочный коэффициент вариации $\hat{V}_X = \hat{\sigma}_X / |\bar{x}| = 10{,}8/49{,}6 = 0{,}2 < 0{,}3(3)$; напомним, для нормально распределенной величины, практически все значения которой положительны (а объем продаж — неотрицательная величина), коэффициент вариации меньше 0,3(3);

— ежедневный объем продаж формируется под воздействием большого числа факторов, что является, в силу центральной предельной теоремы, доводом в пользу нормального закона распределения.

Итак, предположим, что величина X — ежедневный объем продаж имеет нормальный закон распределения, т. е. $X = N(a, \sigma)$. Неизвестные значения параметров a и σ заменим значениями их выборочных аналогов, рассчитанными

по интервальному ряду: a заменим $\overline{x'} = 49{,}485 \approx 49{,}5$, а σ заменим $\hat{\sigma}_{(\text{Ш})} = \sqrt{\hat{D}X_{(\text{Ш})}} = \sqrt{117{,}2} = 10{,}83$ ($\hat{D}X_{(\text{Ш})}$ — скорректированна на поправку Шеппарда дисперсия $\hat{D}X$, рассчитанная по интервальному ряду; значения $\overline{x'}$ и $\hat{D}X_{\text{Ш}}$ приведены в примере 7.8).

Тогда функция плотности нормального распределения (5.32) принимает вид

$$f_N(x) = \frac{1}{10{,}83\sqrt{2\pi}}\, e^{-(x-49{,}5)^2/(2\cdot 117{,}2)}, \qquad (7.54)$$

а функция распределения (5.34)

$$F_N(x) = \frac{1}{10{,}83\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t-49{,}5)^2/(2\cdot 117{,}2)}\, dt. \qquad (7.55)$$

Рассчитаем значения $f_N(x'_i)$ функции плотности нормального распределения (7.54) в серединах x'_i интервалов и сравним их со значениями $\hat{f}_X(x'_i)$ выборочной функции плотности.

Значения $f_N(x'_i)$ можно определить:

— используя алгоритм (5.43), который принимает вид

$$x'_i \to z_i = \frac{x'_i - \overline{x'}}{\hat{\sigma}_{(\text{Ш})}} = \frac{x'_i - 49{,}5}{10{,}83} \xrightarrow[\text{п.1}]{} \varphi(z_i) \to \frac{}{\hat{\sigma}_{(\text{Ш})}}\varphi(z_i) = f_N(x'_i)$$

(расчеты приведены в табл. 7.9, столбцы 8, 9, 10);

— используя **Статистическую функцию** Microsoft Excel НОРМРАСП (x; a; σ; интегральный), где $x = x'_i$, $a = 49{,}5$, $\sigma = 10{,}83$, а аргумент «интегральный» есть ЛОЖЬ.

Куполообразная кривая нормального распределения, проведенная через точки $(x'_i, f_N(x'_i))$, где $i = 1, 2, \ldots, 9$, изображена на рисунке 7.4, *а*. Она достаточно хорошо сглаживает полигон (ломаную линию); различия между значениями $f_N(x'_i)$ и $\hat{f}_X(x'_i)$ почти во всех точках небольшие.

Рассчитаем значения $F_N(a_{i+1})$ функции нормального распределения (7.55) в концах интервалов и сравним их со значениями $\hat{F}_X(a_{i+1})$ выборочной функции распределения.

Значения $F_N(a_{i+1})$ можно определить:

— используя алгоритм (5.45), который принимает вид

$$a_{i+1} \to z_{i+1} = \frac{a_{i+1} - \overline{x'}}{\hat{\sigma}_{(\text{Ш})}} = \frac{a_{i+1} - 49{,}5}{10{,}83} \xrightarrow[\text{п.1}]{} \Phi(z_{i+1}) \to$$

$$\to \frac{1}{2} + \Phi(z_{i+1}) = F_N(a_{i+1})$$

(расчеты приведены в табл. 7.9, столбцы 11, 12, 13);

— используя **Статистическую функцию** НОРМРАСП (x; a; σ; интегральный), где $x = a_{i+1}$, $a = 49{,}5$, $\sigma = 10{,}83$, а аргумент «интегральный» есть ИСТИНА.

Кривая функции нормального распределения, проведенная через точки $(a_{i+1}; F_N(a_{i+1}))$, где $i = 1, 2, \ldots, 9$, изображена на рисунке 7.4, *б*; различий между значениями $F_N(a_{i+1})$ и $\widehat{F}_X(a_{i+1})$ практически нет.

Таким образом, проведенное сглаживание выборочного распределения ежедневного объема продаж (за 100 дней) нормальным распределением позволяет сделать предварительное заключение о приемлемости нормального закона. Более серьезное обоснование этого заключения дается в § 9.6 (см. пример 9.3). ◄

УПРАЖНЕНИЯ

1. Ежедневные суммарные денежные вклады населения (тыс. руб.) в отделение банка в течение 20 рабочих дней таковы: 60, 20, 70, 70, 30, 20, 50, 50, 40, 60, 30, 40, 30, 50, 50, 60, 50, 60, 40, 40. Сгруппируйте данные в статистический ряд распределения; задайте таблично выборочную функцию распределения величины X — суммарного денежного вклада за день и постройте ее график. По статистическому ряду вычислите размер среднего суммарного вклада за день, дисперсию и исправленную дисперсию вклада, моду и медиану, асимметрию и эксцесс и исправленные асимметрию и эксцесс. Найденные значения характеристик сравните с результатами работы программы **«Описательная статистика»** Microsoft Excel.

2. Данные о росте 30 студентов (см): 182, 171, 186, 175, 188, 177, 176, 178, 183, 187, 167, 180, 182, 179, 176, 179, 172, 173, 183, 168, 180, 179, 172, 177, 175, 173, 189, 176, 190,6, 172 сгруппируйте в интервальный ряд, задайте таблично выборочные функции плотности и функцию распределения величины X — роста студента; постройте гистограмму, полигон, кумулятивную кривую. Используя интервальный ряд, вычислите средний рост, дисперсию роста с учетом поправки Шеппарда, моду и медиану. Значения найденных характеристик сравните со средним, дисперсией выборки, модой и медианой, рассчитанными программой **«Описательная статистика»** по исходным 30 данным. При расхождении объясните их причину.

3. По результатам наблюдений $x_1, x_2, \ldots, x_n$ величины X найдены ее следующие характеристики: $\bar{x} = 25$, $\widehat{\sigma}_X^2 = 0{,}6$, $s_X^2 = 0{,}55$, $\widehat{A}_X = -0{,}5$, $\widehat{E}_X = 1{,}2$, $\hat{x}_{\text{mod}} = 20$. Найдите соответствующие характеристики ряда чисел $y_i = -4x_i - 15$, $i = 1, 2, \ldots, n$.

4. По следующим парным наблюдениям двумерной величины (X, Y) постройте графики и рассчитайте коэффициенты корреляции, используя формулы (7.32) и (7.33):

а)

x_i	3	4	6	7
y_i	5	10	9	12

б)

x_i	3	4	6	7
y_i	12	9	10	5

в)

x_i	−2	−1	0	1	2
y_i	4	1	0	1	4

Можно ли считать коэффициент корреляции характеристикой степени линейной зависимости? О чем говорит знак коэффициента корреляции?

5. Распределение 200 погибших от несчастных случаев по возрасту таково:

i	1	2	3	4	5	6
Возраст $[a_i, a_{i+1})$	16—21	21—26	26—31	31—36	36—41	41—46
Число погибших (m_i)	133	45	15	4	2	1

Найдите выборочную функцию плотности и функцию распределения и постройте их графики. Предположив, что случайная величина X (продолжительность жизни погибшего) имеет показательный закон распределения, найдите значения функции плотности в серединах интервалов и функции распределения в концах интервалов; постройте их графики. Удовлетворены ли вы результатами «сглаживания» выборочных функций плотности и распределения соответствующими функциями показательного закона?

6. Промежуток времени X между приходом двух автобусов не более 20 мин. Наблюдения дали следующие результаты:

i	1	2	3	4
Промежуток времени $[a_i, a_{i+1})$	0—5	5—10	10—15	15—20
Число случаев (m_i)	22	27	23	28

Постройте график выборочной функции плотности и, предположив, что случайная величина X равномерно распределена на отрезке [0; 20], «сгладьте» выборочную функцию плотности функцией плотности равномерного закона. Каковы теоретические частоты $m_i^{\text{теор}}$ при равномерном законе?

7. По данным социологического опроса получено распределение числа групп по числу респондентов в группе, отрицательно отзывающихся о новой рекламной политике фирмы (в каждой группе по 10 респондентов).

i	1	2	3	4	5
Число респондентов, не поддерживающих новую рекламную политику (x_i')	0	1	2	3	4
Число групп (m_i)	132	43	20	3	2

Предположив, что число респондентов в группе, не поддерживающих новую рекламную политику, распределено по закону Пуассона, определите долю групп, в которых все респонденты поддерживают новую рекламную политику.

8. Распределение числа детей, склонных к абстрактному мышлению, в группе из 10 детей таково:

i	1	2	3	4	5
Число детей, склонных к абстрактному мышлению (x_i')	0	1	2	3	4
Число групп (m_i)	1	7	10	4	3

Предположив, что число детей, склонных к абстрактному мышлению, подчиняется биномиальному закону, определите ожидаемое в среднем число таких детей в группе: а) из 10 чел.; б) из 20 чел.

ГЛАВА 8

Точечные и интервальные оценки числовых характеристик случайной величины (параметров распределений)

В гл. 5 введено понятие параметра закона распределения — это постоянная величина (не общематематическая, как, например, π, e), входящая в формулу закона.

Параметры одних законов являются числовыми характеристиками случайной величины (так, параметры нормального закона a и σ — соответственно математическое ожидание MX и среднее квадратическое отклонение σ_X),

параметры других законов не являются числовыми характеристиками, но обычно могут быть выражены через них (так, параметры равномерного на отрезке $[a, b]$ закона — это числа a и b, которые, используя соотношения (5.25), а именно, $MX = (a + b)/2$ и $DX = (b - a)^2/12$, можно выразить через MX и DX).

В данной главе рассматриваются два подхода к нахождению по результатам наблюдений оценок неизвестных числовых характеристик случайной величины (параметров распределений): точечный и интервальный. Точечный подход указывает лишь точку, около которой находится оцениваемая характеристика (параметр); при интервальном находят интервал, который с некоторой, как правило, большой вероятностью, задаваемой исследователем, накрывает неизвестное значение числовой характеристики (параметра). Рассматриваются также показатели качества точечных оценок и методы нахождения этих оценок; строятся интервальные оценки параметров нормального распределения, вероятности события и коэффициента корреляции.

§ 8.1. Понятие точечной оценки, ее свойства. Точечные оценки математического ожидания, дисперсии и вероятности события

В гл. 7 были рассмотрены различные выборочные характеристики случайной величины X: среднее арифметическое $\bar{x}$, выборочная дисперсия $\hat{D}X$, $\hat{p}$ — опытная вероятность и др. Эти характеристики используют в качестве приближенных значений неизвестных числовых характеристик изучаемой случайной величины X (неизвестных генеральных характеристик). Так, среднее $\bar{x}$ используется как приближенное значение математического ожидания MX (генеральной средней), а выборочная дисперсия $\hat{D}X$ — как приближенное значение генеральной дисперсии DX; частость $\hat{p}$ — как приближенное значение вероятности p.

Определение. *Выборочная характеристика, используемая в качестве приближенного значения генеральной характеристики, называется ее точечной статистической оценкой.*

Среднее арифметическое $\bar{x}$, выборочная дисперсия $\hat{D}X$, частость $\hat{p}$ — это точечные статистические оценки соответственно математического ожидания (генерального среднего) MX, дисперсии (генеральной дисперсии) DX, истинной (генеральной) вероятности p.

«Точечная» означает, что оценка представляет отдельное значение, принимаемое за приближенное значение генеральной характеристики; «статистическая» означает, что оценка рассчитывается по результатам наблюдений, или,

иначе, по собранной исследователем статистике. Далее слово «статистическая» будем опускать.

Замечание. Результаты n наблюдений величины X могут иметь два варианта интерпретации; либо это фактически полученные результаты и тогда их обозначают $x_1, x_2, \ldots, x_n$; либо мыслимые, возможные, случайные результаты и тогда их обозначают $X_1, X_2, \ldots, X_n$.

В случае первого варианта среднее $\bar{x} = \sum_{i=1}^{n} x_i/n$, выборочная дисперсия $\hat{D}X = \sum_{i=1}^{n} (x_i - \bar{x})^2/n$, исправленная выборочная дисперсия $s_X^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$ — конкретные числа.

В случае второго варианта $\overline{X} = \sum_{i=1}^{n} X_i/n$, $\hat{D}X = \sum_{i=1}^{n} (X_i - \overline{X})^2/n$, $s_X^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$ — случайные величины.

То же относится и к числу m успехов в n испытаниях Бернулли и к частости $\hat{p} = m/n$: если речь идет о фактически наблюдавшемся числе успехов, то m и $\hat{p} = m/n$ — числа; если о возможном, случайном числе успехов, то m и $\hat{p} = m/n$ — случайные величины.

Какой именно вариант интерпретации результатов наблюдений, следовательно, и выборочной характеристики используется, как правило, ясно из контекста; различий в обозначениях характеристик, обусловленных вариантом их интерпретации, обычно не делается (такое различие принято только в обозначении среднего: $\bar{x}$ — это число, $\overline{X}$ — случайная величина).

Обозначим через Θ некоторую генеральную характеристику (это может быть и MX, и любая другая числовая характеристика случайной величины X). Пусть ее числовое значение неизвестно, однако предложен некоторый алгоритм или формула вычисления точечной оценки $\hat{\Theta}_{(n)}$ этой характеристики по результатам $X_1, X_2, \ldots, X_n$ наблюдений величины X. Обозначая буквой f этот алгоритм, запишем

$$\hat{\Theta}_{(n)} = f(X_1, X_2, \ldots, X_n). \qquad (8.1)$$

Подставив в (8.1) вместо $X_1, X_2, \ldots, X_n$ конкретные результаты $x_1, x_2, \ldots, x_n$ наблюдений (конкретные числа), получим число $\hat{\Theta}$, которое и принимают за приближенное значение неизвестной генеральной характеристики Θ. Найти погрешность этого приближения нельзя, поскольку числовое значение характеристики Θ неизвестно. Чтобы ответить на вопрос, пригодно или нет найденное приближение, рассмотрим оценку $\hat{\Theta}_{(n)}$ с других позиций.

Здесь и далее в этой главе будем предполагать, что наблюдения случайной величины X удовлетворяют условиям (6.8)—(6.11), а именно:

наблюдения независимы, или, иначе, зависящие от случая результаты наблюдений X_1, X_2, ... , X_n — независимые случайные величины; (8.2)

наблюдения проводятся в типичных условиях, или, иначе, закон распределения каждого зависящего от случая результата наблюдения совпадает с законом распределения наблюдаемой величины X, т. е. $F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x) = F_X(x)$, (8.3)

и, как следствие,

$$MX_1 = MX_2 = \ldots = MX_n = MX; \tag{8.4}$$

$$DX_1 = DX_2 = \ldots = DX_n = DX. \tag{8.5}$$

Выполнение условий (8.2) и (8.3) гарантируется при случайной выборке с возвратом из генеральной совокупности любого объема и при случайной выборке без возврата из бесконечно большой генеральной совокупности.

Аргументы X_1, X_2, ... , X_n в формуле (8.1) не конкретные числа, а случайные величины. Поэтому и оценка $\hat{\Theta}_{(n)}$ также величина случайная; следовательно, можно говорить о ее математическом ожидании ($M\hat{\Theta}_{(n)}$), дисперсии ($D\hat{\Theta}_{(n)}$) и законе распределения. Интерпретация оценки $\hat{\Theta}_{(n)}$ как случайной величины позволяет сформулировать свойства, которыми должна обладать оценка, чтобы ее можно было считать хорошим приближением к неизвестной генеральной характеристике Θ. Это свойства состоятельности, несмещенности и эффективности.

Определение. *Оценка $\hat{\Theta}_{(n)}$ генеральной характеристики Θ называется состоятельной, если для любого $\varepsilon > 0$ выполняется равенство*

$$\lim_{n\to\infty} P(|\hat{\Theta}_{(n)} - \Theta| < \varepsilon) = 1. \tag{8.6}$$

Поясним смысл равенства (8.6). Пусть ε — очень малое положительное число. Тогда равенство (8.6) означает, что чем больше число наблюдений n, тем больше уверенность (вероятность) в незначительном по абсолютной величине отклонении оценки $\hat{\Theta}_{(n)}$ от неизвестной характеристики Θ, или короче: чем больше объем исходной информации, тем

«ближе мы к истине». Если это так, то $\hat{\Theta}_{(n)}$ — состоятельная оценка.

«Хорошая» оценка обязательно должна обладать свойством состоятельности. В противном случае оценка не имеет практического смысла: увеличение объема исходной информации не будет «приближать нас к истине». Поэтому свойство состоятельности следует проверять в первую очередь.

Определение. Оценка $\hat{\Theta}_{(n)}$ генеральной характеристики Θ называется несмещенной, если для любого фиксированного числа наблюдений n выполняется равенство

$$M\hat{\Theta}_{(n)} = \Theta, \qquad (8.7)$$

т. е. математическое ожидание оценки равно неизвестной характеристике.

Поясним смысл равенства (8.7) в терминах выборки. Для этого зафиксируем объем выборки n; произведем все возможные выборки с возвратом этого объема из генеральной совокупности; для каждой из них найдем значение оценки $\hat{\Theta}_{(n)}$, а затем среднее этих значений — это $M\hat{\Theta}_{(n)}$. Равенство (8.7) означает: если оценка несмещенная, то при любом фиксированном n среднее из значений оценки, вычисленное для всевозможных выборок объема n, т. е. $M\hat{\Theta}_{(n)}$ совпадает с точным значением генеральной характеристики Θ. Проиллюстрируем это свойство графически. Допустим, имеется два алгоритма расчета оценок $\hat{\Theta}^{(1)}_{(n)}$ и $\hat{\Theta}^{(2)}_{(n)}$ характеристики Θ:

$$\hat{\Theta}^{(1)}_{(n)} = f_1(X_1, X_2, \ldots, X_n) \text{ и } \hat{\Theta}^{(2)}_{(n)} = f_2(X_1, X_2, \ldots, X_n).$$

Значения этих оценок, вычисленные по выборкам объема n, изображены точками на рисунке 8.1. Так как среднее значений оценки $\hat{\Theta}^{(1)}_{(n)}$ совпадает с Θ, то $\hat{\Theta}^{(1)}_{(n)}$ — несмещенная оценка характеристики Θ; $\hat{\Theta}^{(2)}_{(n)}$ — смещенная оценка: среднее ее значений не совпадает с Θ.

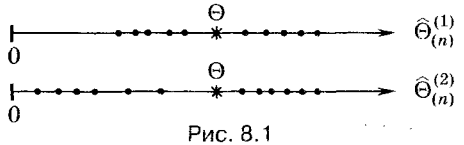


Рис. 8.1

» ЗАДАЧА 8.1. Пусть генеральную совокупность образуют пять чисел: -2, -1, 0, 6, 2. Вычислите генеральное среднее MX и генеральную дисперсию DX Составьте все возможные выборки с возвратом объема $n = 2$; для каждой из них вычислите значения среднего $\overline{X}_{(n=2)}$ и дисперсии $\hat{D}X_{(n=2)}$. Установите, выполняется ли при $n = 2$ равенство

$$M\overline{X}_{(n)} = MX. \qquad (8.8)$$

Является ли выборочная дисперсия несмещенной оценкой генеральной дисперсии?

Р е ш е н и е. Генеральной совокупности -2, -1, 0, 6, 2 соответствует случайная величина X с рядом распределения

x	-2	-1	0	2	6
$P(X=x)$	1/5	1/5	1/5	1/5	1/5

Генеральное среднее

$$MX = -2 \cdot 1/5 - 1 \cdot 1/5 + 0 \cdot 1/5 + 2 \cdot 1/5 + 6 \cdot 1/5 = 1;$$

генеральная дисперсия

$$DX = M(X^2) - (MX)^2 = (4 \cdot 1/5 + 1 \cdot 1/5 + 0 \cdot 1/5 + 4 \cdot 1/5 + 36 \cdot 1/5) - 1^2 = 8.$$

Образуем все возможные выборки с возвратом объема $n = 2$ из данной генеральной совокупности. Они приведены в столбцах 1 и 2 таблицы 8.1; X_1 и X_2 — случайные результаты извлечения соответственно первого числа из пяти данных чисел и второго из тех же пяти чисел.

Например, -2 и -2 означает, что первое число, попавшее в выборку, равно -2 и второе число, попавшее в выборку, также равно -2. Учитывая независимость случайных величин X_1 и X_2 (выборка повторная!), имеем, что вероятность появления такой выборки равна

$$P[(X_1 = -2) \quad (X_2 = -2)] = P(X_1 = -2)P(X_2 = -2) = \frac{1}{5} \cdot \frac{1}{5} = \frac{1}{25}.$$

Нетрудно убедиться в том, что вероятность появления любой другой из перечисленных в таблице 8.1 выборок также равна 1/25.

Если не акцентировать внимание на порядке следования чисел в выборке, то вероятность появления выборки удвоится. Например, вероятность попадания в выборку чи-

сел -1 и -2 равна вероятности появления либо выборки $-2, -1$, либо выборки $-1, -2$, т. е. равна $1/25 + 1/25 = 2/25$, — это число и помещено в таблице 8.1.

Таблица 8.1

Выборка		Вероятность появления выборки	$\overline{X} = \frac{X_1 + X_2}{2}$	$\overline{X^2} = \frac{X_1^2 + X_2^2}{2}$	$\hat{\sigma}^2 = \overline{X^2} - (\overline{X})^2$	$s_X^2 = \frac{n\hat{D}X}{n-1}$ $(n=2)$
X_1	X_2					
1	2	3	4	5	6	7
-2	-2	$1/25$	-2	4	0	0
-2	-1	$1/25 + 1/25 = 2/25$	$-3/2$	$5/2$	$1/4$	$1/2$
-1	-2					
-2	0	$2/25$	-1	2	1	2
0	-2					
-2	6	$2/25$	2	20	16	32
6	-2					
-2	2	$2/25$	0	4	4	8
2	-2					
-1	-1	$1/25$	-1	1	0	0
-1	0	$2/25$	$-1/2$	$1/2$	$1/4$	$1/2$
0	-1					
-1	6	$2/25$	$5/2$	$37/2$	$49/4$	$49/2$
6	-1					
-1	2	$2/25$	$1/2$	$5/2$	$9/4$	$9/2$
2	-1					
0	0	$1/25$	0	0	0	0
0	6	$2/25$	3	18	9	18
6	0					
0	2	$2/25$	1	2	1	2
2	0					
6	6	$1/25$	6	36	0	0
6	2	$2/25$	4	20	4	8
2	6					
2	2	$1/25$	2	4	0	0
		$\Sigma = 1$				

Вычислим для каждой выборки значение среднего $\overline{X}_{(n=2)}$ и дисперсии $\hat{D}X_{(n=2)}$ (см. столбцы 4—6 табл. 8.1). Используя столбцы 4 и 3 таблицы 8.1, построим ряд распределения случайной величины $\overline{X}_{(n=2)}$ (этот ряд приведен в табл. 8.2) и найдем математическое ожидание $M\overline{X}_{(n=2)}$.

Таблица 8.2

$\bar{x}$	-2	$-\frac{3}{2}$	-1	$-\frac{1}{2}$	0	$\frac{1}{2}$	1	2	$\frac{5}{2}$	3	4	6	
$P(\overline{X}_{(n=2)} = \bar{x})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\Sigma = 1$

Математическое ожидание

$$M\overline{X}_{(n=2)} = -2 \cdot 1/25 + (-3/2) \cdot 2/25 + \ldots + 6 \cdot 1/25 = 1 = MX.$$

Таким образом, при $n = 2$ равенство (8.8) выполняется.

Аналогично можно убедиться в том, что равенство (8.8) выполняется и для выборок с возвратом объема $n = 3$, и для выборок с возвратом объема $n = 4$ и т. д. Выполнение равенства (8.8) при любом n означает, что $\overline{X}_{(n)}$ — несмещенная оценка математического ожидания MX. Строгое доказательство этого утверждения приводится дальше.

Чтобы установить, является ли выборочная дисперсия $\hat{D}X_{(n)}$ несмещенной оценкой генеральной дисперсии DX, обратимся к равенству (8.7), в котором положим оценку $\hat{\Theta}_{(n)}$ равной $\hat{D}X_{(n)}$, а генеральную характеристику Θ — равной DX. Оценка $\hat{D}X_{(n)}$ будет несмещенной, если при любом n выполняется равенство $M(\hat{D}X_{(n)}) = DX$. Проверим это, например, при $n = 2$. Используя столбцы 6 и 3 таблицы 8.1, построим ряд распределения случайной величины $\hat{D}X_{(n)}$. Этот ряд приведен в таблице 8.3.

Таблица 8.3

Значение $\hat{\sigma}^2$ дисперсии $\hat{D}X$	0	$1/4$	1	$9/4$	4	$49/4$	9	16	
$P(\hat{D}X_{(n=2)} = \hat{\sigma}^2)$	$5/25$	$4/25$	$4/25$	$2/25$	$4/25$	$2/25$	$2/25$	$2/25$	$\Sigma = 1$

Математическое ожидание

$$M(\hat{D}X_{(n=2)}) = 0 \cdot 5/25 + 1/4 \cdot 4/25 + \ldots + 16 \cdot 2/25 = 4.$$

Так как $M(\hat{D}X_{(n)}) \neq DX$ (напомним, что $DX = 8$) при $n = 2$, то $\hat{D}X$ не обладает свойством несмещенности, т. е. $\hat{D}X$ — смещенная оценка дисперсии DX.

Несмещенной оценкой генеральной дисперсии DX является исправленная выборочная дисперсия s_X^2. Доказательство этого утверждения приводится дальше; здесь же убедимся в том, что при $n = 2$ математическое ожидание исправленной выборочной дисперсии

$$Ms_X^2 = DX. \qquad (8.9)$$

Используя столбцы 7 и 3 таблицы 8.1, построим ряд распределения случайной величины s_X^2, значения которой были найдены для всевозможных выборок с возвратом объема $n = 2$. Он приведен в таблице 8.4.

Таблица 8.4

Значение s^2 дисперсии s_X^2	0	1/2	2	9/2	8	18	49/2	32	
$P(s_X^2 = s^2)$	5/25	4/25	4/25	2/25	4/25	2/25	2/25	2/25	$\Sigma = 1$

Находим

$$Ms_X^2 = 0 \cdot 5/25 + 1/2 \cdot 4/25 + \ldots + 32 \cdot 2/25 = 8 = DX$$

(в задаче 8.1 генеральная дисперсия $DX = 9$).

Аналогично можно убедиться, что и для выборок с возвратом объема $n = 3$ и для выборок с возвратом объема $n = 4$ и т. д. выполняется равенство (8.9). Таким образом, если выборка с возвратом (наблюдения независимы), то s_X^2 — несмещенная оценка дисперсии DX. ◀

Определение. Несмещенная оценка $\hat{\Theta}_{(n)}$ характеристики Θ называется несмещенной эффективной, если она среди всех прочих несмещенных оценок той же самой характеристики обладает наименьшей дисперсией.

Установим смысл этого свойства. Допустим, что имеется два алгоритма нахождения оценок $\hat{\Theta}_{(n)}^{(1)}$ и $\hat{\Theta}_{(n)}^{(2)}$ одной и той же характеристики Θ

$$\hat{\Theta}_{(n)}^{(1)} = \varphi_1(X_1, X_2, \ldots, X_n) \text{ и } \hat{\Theta}_{(n)}^{(2)} = \varphi_2(X_1, X_2, \ldots, X_n),$$

причем обе оценки несмещенные, т. е. при любом n

$$M\hat{\Theta}^{(1)}_{(n)} = \Theta \text{ и } M\hat{\Theta}^{(2)}_{(n)} = \Theta.$$

Выясним, какая из оценок «лучше». Оценка здесь трактуется как случайная величина; показателем разброса значений случайной величины около ее математического ожидания является дисперсия. Так как математические ожидания и той, и другой оценки одинаковы, то естественно считать «лучшей», более эффективной ту оценку, у которой меньше дисперсия. Из рисунка 8.2 видно, что из оценок $\hat{\Theta}^{(1)}_{(n)}$ и $\hat{\Theta}^{(2)}_{(n)}$ более эффективной является $\hat{\Theta}^{(1)}_{(n)}$: разброс ее значений (они изображены точками) около Θ, или, иначе говоря, ее дисперсия меньше.

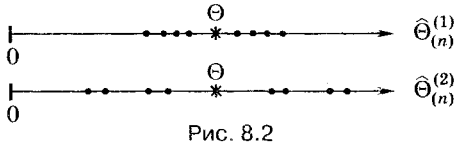


Рис. 8.2

Как же выяснить, является несмещенная оценка эффективной или нет, т. е. имеет ли она по сравнению с другими несмещенными оценками, которых может быть достаточно много, наименьшую дисперсию или нет? В некоторых случаях этот минимум хорошо известен; тогда, сравнив с ним дисперсию рассматриваемой оценки, можно ответить на поставленный вопрос. Так, для случайной величины X, имеющей нормальный закон с дисперсией σ^2_X,

нижняя граница дисперсий различных несмещенных оценок математического ожидания равна σ^2_X/n, (8.10)

нижняя граница дисперсий различных несмещенных оценок дисперсии равна $2\sigma^4_X/n$. (8.11)

В выражениях (8.10) и (8.11) n — число независимых, проводимых в типичных условиях наблюдений случайной величины X.

Нижняя граница дисперсий различных несмещенных оценок вероятности p успешности испытания равна $p(1-p)/n$, (8.12)

где n — число испытаний Бернулли (это независимые испытания, в любом из которых вероятность появления успеха — некоторого события A — равна p).

Выборочное среднее как точечная оценка математического ожидания. Установим, обладает ли выборочное сред-

нее $\overline{X}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (рассматриваемое как случайная величина) перечисленными свойствами состоятельности, несмещенности и эффективности.

1^0. *Состоятельность.* В § 6.2 было доказано, что при проведении в типичных условиях независимых наблюдений случайной величины X с ограниченной дисперсией ($DX \leqslant c$), имеет место равенство (6.24), а именно:

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - MX\right| < \varepsilon\right) = 1,$$

или

$$\lim_{n\to\infty} P(|\overline{X}_{(n)} - MX| < \varepsilon) = 1.$$

Сравнив его с определением (8.6) свойства состоятельности ($\overline{X}_{(n)}$ — это $\hat{\Theta}_{(n)}$, MX — это Θ), заключаем, что *выборочное среднее $\overline{X}_{(n)}$ является состоятельной оценкой математического ожидания MX.*

2^0. *Несмещенность.* В § 6.1 было доказано: если $MX_1 = MX_2 = \ldots = MX_n = MX$ (это соотношение является следствием проведения наблюдений случайной величины X в типичных условиях!), то имеет место равенство (6.12), а именно: $M\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = MX$, или $M\overline{X}_{(n)} = MX$. Сравнивая его с определением (8.7) свойства несмещенности ($\overline{X}_{(n)}$ — это $\hat{\Theta}_{(n)}$, MX — это Θ), заключаем, что $\overline{X}_{(n)}$ — *несмещенная оценка математического ожидания MX.*

3^0. *Эффективность.* Свойство эффективности рассматривается в этой главе только для таких оценок, которые являются несмещенными. Выборочное среднее $\overline{X}_{(n)}$ — несмещенная оценка математического ожидания MX. Чтобы выяснить будет ли эта оценка эффективной, надо найти ее дисперсию $D\overline{X}_{(n)}$ и сравнить $D\overline{X}_{(n)}$ с нижней границей для дисперсий различных несмещенных оценок математического ожидания MX, которая зависит от закона распределения величины X. В § 6.1 было доказано, что при проведении в типичных условиях независимых наблюдений величины X имеет место равенство (6.13), а именно

$$D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = DX/n, \text{ или } D\overline{X}_{(n)} = DX/n.$$

Вместе с тем, если считать величину X нормально распределенной с дисперсией σ_X^2, то, согласно (8.10), нижняя граница для дисперсий различных несмещенных оценок математического ожидания равна σ_X^2/n, что совпадает с $D\overline{X}_{(n)}$.

Таким образом, $\overline{X}_{(n)}$ — *эффективная оценка математического ожидания MX нормально распределенной случайной величины X.* (Свойством эффективности выборочное среднее обладает и для ряда других распределений величины X.)

Точечные оценки генеральной дисперсии. Рассмотрим выборочную дисперсию $\hat{D}X = \sum_{i=1}^{n} (X_i - \overline{X})^2/n$ и исправленную выборочную дисперсию

$$s_X^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1), \tag{8.13}$$

которая связана с $\hat{D}X$ соотношением

$$s_X^2 = n\hat{D}X/(n-1). \tag{8.14}$$

Выясним, какими свойствами обладают $\hat{D}X$ и s_X^2, как точечные оценки дисперсии.

1^0. *Состоятельность.* Можно доказать, что при проведении в типичных условиях независимых наблюдений случайной величины X с ограниченными сверху центральными моментами второго и четвертого порядка имеют место следующие равенства:

$$\lim_{n\to\infty} P(|\hat{D}X - DX| < \varepsilon) = 1,$$

т. е. $\hat{D}X$ — *состоятельная оценка генеральной дисперсии DX, и*

$$\lim_{n\to\infty} P(|s_X^2 - DX| < \varepsilon) = 1,$$

т. е. s_X^2 — также *состоятельная оценка генеральной дисперсии DX.*

2^0. *Несмещенность.* В задаче 8.1 было показано, что $\hat{D}X$ — смещенная оценка дисперсии DX, так как оказалось, что при $n = 2$ математическое ожидание $M(\hat{D}X) \neq DX$.

Найдем $M(\widehat{D}X)$ в общем случае; при этом будем предполагать, что наблюдения независимы и проводятся в типичных условиях, т. е. выполняются соотношения (8.2)—(8.5).

≫ Предварительно обратим внимание на следующее.

— Наблюдения величины X проводятся в типичных условиях, поэтому зависящие от случая результаты этих наблюдений $X_1, X_2, \ldots, X_n$ имеют одинаковые математические ожидания и одинаковые дисперсии, совпадающие соответственно с MX и DX (см. формулы (8.4) и (8.5)), также одинаковы и математические ожидания квадратов результатов наблюдений и квадрата величины X, т. е.

$$M(X_1^2) = M(X_2^2) = \ldots = M(X_n^2) = M(X^2). \qquad (8.15)$$

— Наблюдения независимы, следовательно, согласно (4.20),

$$M(X_iX_j) = MX_iMX_j, \qquad i, j = 1, 2, \ldots, n, \ i \neq j. \qquad (8.16)$$

Найдем $M(\widehat{D}X)$. Имеем

$$M(\widehat{D}X) = M\Big[\sum_{i=1}^{n}(X_i - \overline{X})^2/n\Big] = M\Big[\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)/n\Big] =$$

$$= M\Big(\sum_{i=1}^{n} X_i^2/n\Big) - M\Big(\sum_{i=1}^{n} 2X_i\overline{X}/n\Big) + M\Big(\sum_{i=1}^{n}\overline{X}^2/n\Big) =$$

$$= \sum_{i=1}^{n} M(X_i^2)/n - 2M\Big(\overline{X}\sum_{i=1}^{n} X_i/n\Big) + M(n\overline{X}^2/n) \underset{(8.4)}{=}$$

$$= \sum_{i=1}^{n} M(X^2)/n - 2M(\overline{X}^2) + M(\overline{X}^2) = M(X^2) - M(\overline{X}^2),$$

но

$$M(\overline{X}^2) = M\Big(\frac{X_1 + X_2 + \ldots + X_n}{n}\Big)^2 = \frac{1}{n^2}M\Big(\sum_{i=1}^{n} X_i^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{n} X_iX_j\Big) =$$

$$= \sum_{i=1}^{n} M(X_i^2)/n^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{n} M(X_iX_j)/n^2 \underset{\substack{(8.15)\\(8.16)}}{=}$$

$$= \sum_{i=1}^{n} M(X^2)/n^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{n} (MX_i \cdot MX_j)/n^2 \underset{(8.4)}{=}$$

$$\underset{(8.4)}{=} nM(X^2)/n^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{n} (MX \cdot MX)/n^2 =$$

$$= M(X^2)/n + n(n-1)(MX)^2/n^2 = M(X)^2/n + (n-1)(MX)^2/n.$$

Поэтому

$$M(\widehat{D}X) = M(X^2) - [M(X^2)/n + (n-1)(MX)^2/n] =$$

$$= \frac{n-1}{n}M(X^2) - \frac{n-1}{n}(MX)^2 = \frac{n-1}{n}[M(X^2) - (MX)^2] = \frac{n-1}{n}DX. \quad ≪$$

Таким образом,

$$M(\hat{D}X) = \frac{n-1}{n} DX < DX, \qquad (8.17)$$

т. е. ни при каком объеме выборки n математическое ожидание $M(\hat{D}X)$ не равно DX; $M(\hat{D}X)$ всегда меньше DX. Это означает, что оценка $\hat{D}X$ — *смещенная влево оценка дисперсии* DX, при этом смещение равно

$$|M(\hat{D}X) - DX| = \left|\frac{n-1}{n} DX - DX\right| = \frac{DX}{n},$$

и оно тем меньше, чем больше число наблюдений n.

Несмещенной оценкой дисперсии DX является оценка s_X^2, определяемая по формуле (8.13). Найдем Ms_X^2.

$$\gg \quad Ms_X^2 \underset{(8.14)}{=} M\left(\frac{n\hat{D}X}{n-1}\right) = \frac{n}{n-1} M(\hat{D}X) \underset{(8.17)}{=}$$

$$\underset{(8.17)}{=} \frac{n}{n-1} \frac{n-1}{n} DX = DX. \quad \ll$$

Итак, при любом n

$$Ms_X^2 = DX.$$

Сравнивая это равенство с определением (8.7) свойства несмещенности (s_X^2 — это $\hat{\Theta}_{(n)}$, DX — это Θ), заключаем, что s_X^2 — *несмещенная оценка дисперсии* DX.

Именно несмещенностью оценки s_X^2 объясняется более частое, по сравнению с $\hat{D}X$, ее использование в качестве приближенного значения генеральной дисперсии DX. $\ll$

3^0. *Эффективность.* Напомним, свойство эффективности рассматривалось только для таких оценок, которые являются несмещенными. Чтобы ответить на вопрос, является ли несмещенная оценка s_X^2 дисперсии DX эффективной, надо найти дисперсию Ds_X^2 и сравнить ее с нижней границей для дисперсий различных несмещенных оценок дисперсии DX, которая при нормальном распределении величины X равна, согласно (8.11), $2\sigma_X^4/n$.

Можно доказать, что в случае нормального распределения величины X дисперсия $Ds_X^2 = 2\sigma_X^4/(n-1) \neq 2\sigma_X^4/n$.

Так как Ds_X^2 не совпадает с нижней границей, то s_X^2, *будучи несмещенной оценкой дисперсии* DX, *не является эффективной оценкой.*

Замечание. Несмещенная эффективная и состоятельная оценка дисперсии DX нормально распределенной случайной величины $X = N(a, \sigma)$ имеет вид

$$s_0^2 = \sum_{i=1}^{n} (X_i - a)^2/n.$$

В формулу для s_0^2 входит математическое ожидание a, которое, как правило, заранее не известно, поэтому эта оценка практически не используется.

Относительная частота как точечная оценка вероятности события. Пусть p — неизвестная вероятность появления случайного события A в единичном испытании. Проведем n независимых испытаний, в каждом из которых событие A может произойдет с вероятностью p или не произойти с вероятностью $q = 1 - p$ (напомним, что серия испытаний подобного типа называется последовательностью испытаний Бернулли). Пусть m — количество испытаний, в которых произойдет событие A, или количество успехов. Тогда $\hat{p} = m/n$ — относительная частота, или опытная вероятность, или частость появления события A, и принимается за приближенное значение вероятности p. Выясним, какими свойствами обладает $\hat{p}_{(n)} = m/n$, как точечная оценка вероятности p, при этом будем иметь в виду, что m — не конкретное число успехов, а обозначение зависящего от случая числа успехов, которое могло бы произойти в n испытаниях Бернулли, т. е. m, следовательно, и $\hat{p}_{(n)}$ — случайные величины.

1^0. *Состоятельность.* В § 6.2 была доказана теорема Бернулли, согласно которой при проведении испытаний Бернулли имеет место рвенство (6.25), а именно

$$\lim_{n\to\infty} P(|m/n - p| < \varepsilon) = 1.$$

Сравнив его с определением (8.6) свойства состоятельности (m/n — это $\Theta_{(n)}$, p — это Θ), заключаем, что $\hat{p}_{(n)} = m/n$ — *состоятельная оценка вероятности* p.

2^0. *Несмещенность.* Найдем математическое ожидание частости $\hat{p}_{(n)} = m/n$.

Величина m, будучи случайным числом успехов в n испытаниях Бернулли, является биномиальной случайной величиной X_{Bi}, математическое ожидание которой, согласно (5.4), $MX_{\text{Bi}} = np$; поэтому и математическое ожидание $Mm = np$. Тогда

$$M\hat{p}_{(n))} = M(m/n) \underset{n=\text{const}}{=} M(m)/n = MX_{\text{Bi}}/n = np/n = p.$$

Таким образом, $M\hat{p}_{(n)} = p$, т. е. при любом фиксированном числе n испытаний Бернулли математическое ожидание частости $\hat{p}_{(n)}$ равно вероятности p. Это означает, что $\hat{p}_{(n)} = m/n$ — *несмещенная оценка вероятности* p.

3^0. *Эффективность.* Найдем дисперсию $D\hat{p}_{(n)}$. Учитывая, что в данном случае m — биномиальная случайная величина X_{Bi}, дисперсия которой, согласно (5.5), равна $DX_{\text{Bi}} = np(1-p)$, получим

$$D\hat{p}_{(n)} = D(m/n) \underset{n=\text{const}}{=} D(m)/n^2 = DX_{\text{Bi}}/n^2 =$$

$$= np(1-p)/n^2 = p(1-p)/n.$$

Таким образом, $D\hat{p}_{(n)} = p(1-p)/n$. Так как $D\hat{p}_{(n)}$ совпадает с нижней границей для дисперсий различных несмещенных оценок вероятности p (см. (8.12)), то $\hat{p}_{(n)} = m/n$, *будучи несмещенной оценкой вероятности* p, *является также и ее эффективной оценкой.*

§ 8.2. Методы получения точечных оценок параметров распределений

В § 8.1 изучались точечные оценки основных генеральных характеристик: математического ожидания, дисперсии, вероятности. Однако осталось неясным, каким образом получены эти оценки. Рассмотрим два метода получения точечных оценок: метод моментов и метод максимального правдоподобия. При этом будем предполагать, что оцениваемая генеральная характеристика является параметром того или иного закона распределения.

Метод моментов. Пусть Θ — параметр закона распределения, числовое значение которого неизвестно и надо найти его точечную оценку, располагая результатами $x_1, x_2, \ldots, x_n$ независимых наблюдений величины X, проведенных в типичных условиях.

Напомним, что к числовым характеристикам случайной величины X относятся в том числе и начальные $\nu_k(X)$ и центральные $\mu_k(X)$ моменты порядка $k = 0, 1, 2, \ldots$, изученные в п. 4.3.3. Учитывая, что центральные моменты могут быть выражены через начальные (см. табл. 4.9), сформулируем алгоритм метода моментов в терминах начальных моментов:

— параметр Θ выражают через начальные моменты ν_k, т. е. находят зависимость

$$\Theta = g(\nu_1, \nu_2, \ldots); \qquad (8.18)$$

какие начальные моменты следует включать в функцию (8.18), определяется конкретной моделью закона распределения;

— в функцию (8.18) вместо ν_1, ν_2, ... подставляют рассчитанные по наблюдениям выборочные начальные моменты $\hat{\nu}_1$, $\hat{\nu}_2$, ... (см. § 7.2); получают оценку по методу моментов

$$\hat{\Theta}_{\text{м}} = g(\hat{\nu}_1, \hat{\nu}_2, \ldots) \qquad (8.19)$$

параметра Θ (индекс «м» — первая буква слова «момент»).

Алгоритм нетрудно обобщить на случай, когда число параметров закона распределения больше одного.

▶ **ПРИМЕР 8.1.** Случайная величина X распределена по закону Пуассона. Согласно (5.6),

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, \ldots .$$

По результатам $x_1, x_2, \ldots, x_n$ наблюдений величины X найдем методом моментов оценку $\hat{\lambda}_{\text{м}}$ параметра λ.

Согласно (5.9), параметр λ равен математическому ожиданию пуассоновской величины, $\lambda = MX$, а математическое ожидание любой случайной величины — это начальный момент первого порядка, $MX = \nu_1(X)$; в результате зависимость (8.18) принимает в данном случае вид

$$\lambda = \nu_1(X). \qquad (8.20)$$

Подставив в (8.20) вместо $\nu_1(X)$ выборочный начальный момент первого порядка $\hat{\nu}_1(X) \underset{(7.38)}{=} \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$, получим оценку параметра λ по методу моментов

$$\hat{\lambda}_{\text{м}} = \bar{x}. \qquad (8.21)$$

ПРИМЕР 8.2. Случайная величина X распределена по показательному закону. Согласно (5.28),

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x \geqslant 0.$$

По результатам $x_1, x_2, \ldots, x_n$ наблюдений величины X найдем оценку $\hat{\lambda}_{\text{м}}$ параметра λ.

Согласно (5.30), параметр $\lambda = \frac{1}{MX}$, или, учитывая, что $MX = \nu_1(X)$,

$$\lambda = \frac{1}{\nu_1(X)}. \tag{8.22}$$

Подставив в (8.22) вместо $\nu_1(X)$ выборочный момент $\hat{\nu}_1(X) \underset{(7.38)}{=} \bar{x}$, получим оценку параметра λ по методу моментов:

$$\hat{\lambda}_{\text{м}} = \frac{1}{\bar{x}}. \tag{8.23}$$

ПРИМЕР 8.3. Случайная величина X распределена по нормальному закону, т. е. $X = N(a, \sigma)$. По результатам $x_1, x_2, \ldots, x_n$ наблюдений величины X найдем методом моментов оценки $\hat{a}_{\text{м}}$ и $\hat{\sigma}_{\text{м}}$ параметров a и σ.

Согласно (5.33), $a = MX$ и $\sigma = \sigma_X$, или, учитывая, что $MX = \nu_1(X)$ и $\sigma_X = \sqrt{DX} \underset{(4.67)}{=} \sqrt{\nu_2(X) - [\nu_1(X)]^2}$, получим

$$a = \nu_1(X), \qquad \sigma = \sqrt{\nu_2(X) - [\nu_1(X)]^2}. \tag{8.24}$$

Подставив в (8.24) вместо $\nu_1(X)$ и $\nu_2(X)$ выборочные начальные моменты первого и второго порядков $\hat{\nu}_1(X) = \bar{x}$ и $\hat{\nu}_2(X) = \overline{x^2} = \sum_{i=1}^{n} x_i^2/n$, получим для параметров a и σ оценки по методу моментов:

$$\hat{a}_{\text{м}} = \bar{x}, \; \hat{\sigma}_{\text{м}} = \sqrt{\overline{x^2} - (\bar{x})^2} \underset{(7.43)}{=} \sqrt{\overline{(x - \bar{x})^2}}. \quad \blacktriangleleft \tag{8.25}$$

Оценка, получаемая методом моментов, обладает свойством состоятельности и при большом числе наблюдений n, а в ряде случаев и при любом n, свойством несмещенности. Однако эта оценка даже при большом n, как правило, отличается и порой значительно от эффективной (дисперсия оценки превышает нижнюю границу, допустимую для дисперсии ее несмещенных оценок). Более эффективную оценку (оценку, дисперсия которой меньше дисперсии оценки, полученной методом моментов и, следовательно, ближе к допустимой нижней границе или равна этой границе) или во всяком случае не менее эффективную позволяет получить метод максимального правдоподобия.

Метод максимального правдоподобия. В основе метода максимального правдоподобия лежит понятие функции

правдоподобия результатов наблюдений. Введем это понятие для случая, когда случайная величина X *дискретная*. Предположим, что закон распределения величины X известен с точностью до параметра Θ; это означает, что известна формула $P(X = x)$, по которой можно было бы при известном значении параметра Θ рассчитать вероятность того или иного значения x величины X. Однако значение параметра Θ не известно и наша задача, имея наблюдения величины X, найти оценку параметра Θ.

Пусть $X_1, X_2, \ldots, X_n$ — зависящие от случая результаты n наблюдений величины X, проводимых в типичных условиях. Напомним, что эти предположения означают следующее:

$X_1, X_2, \ldots, X_n$ — независимые случайные величины; (8.26)

закон распределения любой из величин $X_1, X_2, \ldots, X_n$ совпадает с законом распределения величины X, т. е.

$$P(X_1 = x) = P(X_2 = x) = \ldots = P(X_n = x) = P(X = x). \quad (8.27)$$

Пусть $x_1, x_2, \ldots, x_n$ — конкретные результаты n наблюдений величины X. Функция правдоподобия результатов наблюдений дискретной случайной величины X — это функция

$$L(x_1, x_2, \ldots, x_n) = P[(X_1 = x_1) \quad (X_2 = x_2) \quad \ldots \quad \ldots \quad (X_n = x_n)], \quad (8.28)$$

подсчитывающая вероятность появления при n наблюдениях величины X чисел $x_1, x_2, \ldots, x_n$. Чем больше значение функции (8.28), тем правдоподобнее или более вероятно появление в результате наблюдений чисел $x_1, x_2, \ldots, x_n$. Отсюда и название функции (8.28) — функция правдоподобия результатов наблюдений.

Учитывая условия (8.26) и (8.27), получим

$$P[(X_1 = x_1) \quad (X_2 = x_2) \quad \ldots \quad (X_n = x_n)] \underset{(8.26)}{=}$$

$$\underset{(8.26)}{=} P(X_1 = x_1)P(X_2 = x_2)\ldots P(X_n = x_n) \underset{(8.27)}{=}$$

$$\underset{(8.27)}{=} P(X = x_1)P(X = x_2)\ldots P(X = x_n),$$

тогда

$$L(x_1, x_2, \ldots, x_n) =$$

$$= P(X = x_1)P(X = x_2)\ldots P(X = x_n).$$

Поскольку значения вероятностей $P(X = x_1)$, $P(X = x_2)$, ... , $P(X = x_n)$ соответственно зависят не только от чисел x_1, x_2, ... , x_n, но и от неизвестного значения параметра Θ закона распределения величины X, в число аргументов функции правдоподобия, наряду с x_1, x_2, ... , x_n, включим Θ. В результате *функция правдоподобия результатов наблюдений дискретной случайной величины X принимает следующий вид*:

$$L(x_1, x_2, \ldots, x_n; \Theta) = \\ = P(X = x_1; \Theta)P(X = x_2; \Theta)\ldots P(X = x_n; \Theta). \quad (8.29)$$

Если величина X *непрерывная*, функция плотности $f_X(x)$ которой известна с точностью до параметра Θ, то *функция правдоподобия результатов* x_1, x_2, ... , x_n *наблюдений величины X имеет вид*

$$L(x_1, x_2, \ldots, x_n; \Theta) = f_X(x_1; \Theta)f_X(x_2; \Theta)\ldots f_X(x_n; \Theta). \quad (8.30)$$

Согласно методу максимального правдоподобия, за оценку параметра Θ при заданных результатах x_1, x_2, ... , x_n наблюдений величины X принимают такое число $\widehat{\Theta}_{\text{мп}}$, при котором функция правдоподобия $L(x_1, x_2, \ldots, x_n; \Theta)$, рассматриваемая как функция переменной Θ, достигает максимума на области $\{\Theta\}$ допустимых значений параметра Θ, т. е.

$$L(x_1, x_2, \ldots, x_n; \widehat{\Theta}_{\text{мп}}) = \max_{\{\Theta\}} L(x_1, x_2, \ldots, x_n; \Theta) \quad (8.31)$$

(индекс «мп» — начальные буквы слов «максимальное правдоподобие»).

Естественность такого подхода к определению оценки $\widehat{\Theta}_{\text{мп}}$ вытекает из смысла функции L: $L(x_1, x_2, \ldots, x_n; \Theta)$ при фиксированном значении Θ является мерой правдоподобия получения при n наблюдениях величины X чисел x_1, x_2, ... , x_n. Изменяя значения параметра Θ, можно проследить, при каких значениях появление чисел x_1, x_2, ... , x_n более правдоподобно, и выбрать такое его значение $\widehat{\Theta}_{\text{мп}}$, при котором результаты x_1, x_2, ... , x_n будут наиболее правдоподобными.

В ряде случаев оценку $\widehat{\Theta}_{\text{мп}}$ максимального правдоподобия удобнее находить как точку максимума функции $\ln L$, т. е. из условия

$$\ln L(x_1, x_2, \ldots, x_n; \widehat{\Theta}_{\text{мп}}) = \max_{\{\Theta\}} \ln L(x_1, x_2, \ldots, x_n; \Theta), \quad (8.32)$$

которое идентично условию (8.31): так как $\ln L$ — монотонно возрастающая функция, то функции L и $\ln L$, рассматриваемые как функции переменной Θ, достигают максимума при одном и том же ее значении. Функция $\ln L$ называется ***логарифмической функцией правдоподобия.***

Согласно (8.32), для нахождения оценки $\widehat{\Theta}_{\text{мп}}$ следует:

— найти решения ***уравнения максимального правдоподобия***

$$\frac{\mathrm{d} \ln L(x_1, x_2, \ldots, x_n; \Theta)}{\mathrm{d}\Theta} = 0, \tag{8.33}$$

относительно Θ, зависящие от $x_1, x_2, \ldots, x_n$;

— среди найденных решений, лежащих внутри области $\{\Theta\}$ допустимых значений переменной Θ, выделить точки максимума;

— если уравнение (8.33) не определено, не имеет решений, зависящих от $x_1, x_2, \ldots, x_n$, или среди его решений нет точки максимума внутри области $\{\Theta\}$ допустимых значений переменной Θ, то точки максимума следует искать на границе области $\{\Theta\}$.

В заключение заметим, если закон распределения величины X известен с точностью, например, до двух параметров Θ_1 и Θ_2, то для нахождения их оценок методом максимального правдоподобия следует вместо уравнения (8.33) найти решения системы уравнений максимального правдоподобия

$$\begin{cases} \dfrac{\partial \ln L(x_1, x_2, \ldots, x_n; \Theta_1, \Theta_2)}{\mathrm{d}\Theta_1} = 0, \\ \dfrac{\partial \ln L(x_1, x_2, \ldots, x_n; \Theta_1, \Theta_2)}{\mathrm{d}\Theta_2} = 0 \end{cases} \tag{8.34}$$

относительно Θ_1 и Θ_2.

▶ **ПРИМЕР 8.4.** Случайная величина X распределена по закону Пуассона. Согласно (5.6),

$$P(X = x) = \frac{\lambda^x}{x!} \mathrm{e}^{-\lambda}, \qquad x = 0, 1, 2, \ldots .$$

По результатам $x_1, x_2, \ldots, x_n$ наблюдений величины X найдем методом максимального правдоподобия оценку параметра λ.

Функция правдоподобия (8.29) примет следующий вид:

$$L(x_1, x_2, \ldots, x_n; \lambda) = \frac{\lambda^{x_1}}{x_1!}\mathrm{e}^{-\lambda}\frac{\lambda^{x_2}}{x_2!}\mathrm{e}^{-\lambda}\ldots\frac{\lambda^{x_n}}{x_n!}\mathrm{e}^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n} x_i}}{x_1!x_2!\ldots x_n!}\mathrm{e}^{-n\lambda}.$$

Логарифмическая функция правдоподобия

$$\ln L(x_1, x_2, \ldots, x_n; \lambda) =$$

$$= \sum_{i=1}^{n} x_i \ln \lambda - \ln x_1! - \ln x_2! - \ldots - \ln x_n! - n\lambda;$$

ее производная по λ

$$\frac{d \ln L}{d\lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n$$

и уравнение максимального правдоподобия (8.33) принимает вид $\sum_{i=1}^{n} x_i/\lambda - n = 0$; его решение $\lambda^* = \sum_{i=1}^{n} x_i/n$ — это точка максимума функции $\ln L$, поскольку вторая производная этой функции при $\lambda = \lambda^*$ отрицательна. Итак, оценка максимального правдоподобия параметра λ такова:

$$\hat{\lambda}_{\text{мп}} = \sum_{i=1}^{n} x_i/n = \bar{x}\,. \qquad (8.35)$$

Обратим внимание на то, что оценка максимального правдоподобия $\hat{\lambda}_{\text{мп}}$ совпала с оценкой (8.21) $\hat{\lambda}_{\text{м}}$, полученной методом моментов. Однако совпадение оценок, полученных этими методами, не правило.

ПРИМЕР 8.5. Случайная величина X распределена по показательному закону. Согласно (5.28),

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x \geqslant 0.$$

По результатам $x_1, x_2, \ldots, x_n$ наблюдений величины X найдем методом максимального правдоподобия оценку параметра λ.

Функция правдоподобия (8.30) принимает следующий вид:

$$L(x_1, x_2, \ldots, x_n; \lambda) = \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \ldots \lambda e^{-\lambda x_n} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

Логарифмическая функция правдоподобия

$$\ln L(x_1, x_2, \ldots, x_n; \lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i,$$

ее производная по λ

$$\frac{d \ln L}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

и уравнение максимального правдоподобия (8.33) принимает вид $n/\lambda - \sum_{i=1}^{n} x_i = 0$; его решение $\lambda^* = n/\sum_{i=1}^{n} x_i$ — точка

максимума функции $\ln L$, в чем нетрудно убедиться. Поэтому оценка максимального правдоподобия параметра λ такова:

$$\hat{\lambda}_{\text{мп}} = n / \sum_{i=1}^{n} x_i = 1/\bar{x}\,.$$

Она совпала с оценкой (8.23) $\hat{\lambda}_{\text{м}}$, полученной методом моментов.

ПРИМЕР 8.6. Случайная величина X распределена по нормальному закону. Согласно (5.32),

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-a)^2/(2\sigma^2)}, \qquad -\infty < x < +\infty.$$

По результатам $x_1, x_2, \ldots, x_n$ наблюдений величины X методом максимального правдоподобия найдем оценки параметров a и σ.

Функция правдоподобия (8.30) принимает следующий вид:

$$L(x_1, x_2, \ldots, x_n; a, \sigma) =$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x_1-a)^2/(2\sigma^2)} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x_2-a)^2/(2\sigma^2)} \ldots$$

$$\ldots \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x_n-a)^2/(2\sigma^2)} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n e^{-\sum_{i=1}^{n}(x_i-a)^2/(2\sigma^2)}$$

Логарифмическая функция правдоподобия

$$\ln L(x_1, x_2, \ldots, x_n; a, \sigma) =$$

$$= n \ln(1/\sqrt{2\pi}) + n \ln(1/\sigma) - \sum_{i=1}^{n} (x_i - a)^2/(2\sigma^2), \quad (8.36)$$

ее частные производные по a и σ

$$\frac{\partial \ln L}{\partial a} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - a),$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - a)^2,$$

и система уравнений максимального правдоподобия (8.34) имеет вид

$$\begin{cases} \sum_{i=1}^{n} (x_i - a)/\sigma^2 = 0, \\ -n/\sigma + \sum_{i=1}^{n} (x_i - a)^2/\sigma^3 = 0. \end{cases}$$

Ее решение $a^* = \sum_{i=1}^{n} x_i/n = \bar{x}$, $\sigma^* = \sqrt{\sum_{i=1}^{n}(x_i - a^*)^2/n} = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}$. Можно убедиться в том, что функция $\ln L$, рассматриваемая как функция аргументов a и σ, принимает максимальное значение в точке (a^*, σ^*). Поэтому оценки параметров a и σ, полученные методом максимального правдоподобия, таковы:

$$\hat{a}_{\text{мп}} = \bar{x}, \ \hat{\sigma}_{\text{мп}} = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}. \qquad (8.37)$$

Сравнив эти оценки с оценками метода моментов (8.25), заключаем, что в случае нормального распределения метод максимального правдоподобия и метод моментов дают одинаковые результаты. Но это совпадение, как уже отмечалось, не правило. В подтверждение этого рассмотрим следующий пример.

ПРИМЕР 8.7. Случайная величина X распределена равномерно на отрезке $[a, b]$. Согласно (5.23),

$$f_X(x) = 1/(b-a), \ x \in [a, b].$$

По результатам $x_1, x_2, \ldots, x_n$ наблюдений величины X методами моментов и максимального правдоподобия найдем оценки параметров a и b.

Выразим параметры a и b через начальные моменты. Учитывая, что начальный момент первого порядка $\nu_1(X) = MX$, а центральный момент второго порядка $\mu_2(X) = DX$, и принимая во внимание, что, согласно (5.25), $MX = (a+b)/2$, а $DX = (b-a)^2/12$, получаем систему

$$\begin{cases} (a+b)/2 = \nu_1(X), \\ (b-a)^2/12 = \mu_2(X), \end{cases}$$

решив которую относительно a и b и учитывая, что $b > a$, получим

$$a = \nu_1(X) - \sqrt{3\mu_2(X)}, \qquad b = \nu_1(X) + \sqrt{3\mu_2(X)}.$$

Поскольку $\mu_2(X) = \nu_2(X) - [\nu_1(X)]^2$, имеем

$$a = \nu_1(X) - \sqrt{3[\nu_2(X) - \nu_1^2(X)]},$$

$$b = \nu_1(X) + \sqrt{3[\nu_2(X) - \nu_1^2(X)]}. \qquad (8.38)$$

Подставив в (8.38) вместо $\nu_1(X)$ и $\nu_2(X)$ выборочные начальные моменты $\hat{\nu}_1(X) = \bar{x}$ и $\hat{\nu}_2(X) = \overline{x^2} = \sum_{i=1}^{n} x_i^2/n$, получим оценки параметров a и b по методу моментов:

$$\hat{a}_{\text{м}} = \bar{x} - \sqrt{3(\overline{x^2} - (\bar{x})^2)}, \qquad \hat{b}_{\text{м}} = \bar{x} + \sqrt{3(\overline{x^2} - (\bar{x})^2)}. \tag{8.39}$$

Найдем оценки параметров a и b методом максимального правдоподобия. Функция правдоподобия (8.30) принимает следующий вид:

$$L(x_1, x_2, \ldots, x_n; a, b) = 1/(b-a)^n. \tag{8.40}$$

Логарифмическая функция правдоподобия $\ln L = -n \ln (b-a)$. Взяв частные производные этой функции по a и по b, получим систему уравнений максимального правдоподобия

$$\begin{cases} n/(b-a) = 0, \\ -n/(b-a) = 0, \end{cases}$$

которая при $n \neq 0$ не имеет решения относительно a и b. Поэтому точку максимума функции правдоподобия (8.40) по переменным a и b следует искать на границе области их допустимых значений, которая имеет вид

$$(a \leqslant x_{(1)}) \quad (b \geqslant x_{(n)}), \tag{8.41}$$

где $x_{(1)} = \min(x_1, x_2, \ldots, x_n)$, а $x_{(n)} = \max(x_1, x_2, \ldots, x_n)$. Так как функция (8.40) возрастает при возрастании a и убывании b, то ее максимум на области (8.41) достигается в точке $(a^* = x_{(1)}, b^* = x_{(n)})$. Итак, оценки максимального правдоподобия параметров a и b таковы:

$$\hat{a}_{\text{мп}} = \min(x_1, x_2, \ldots, x_n), \ \hat{b}_{\text{мп}} = \max(x_1, x_2, \ldots, x_n). \tag{8.42}$$

Сравнив (8.39) и (8.42), заключаем, что метод моментов и метод максимального правдоподобия дают различные оценки параметров a и b равномерного закона. ◂

Оценка, полученная методом максимального правдоподобия, так же, как и оценка, полученная методом моментов, обладает свойством состоятельности и при большом n, а в ряде случаев и при любом n, свойством несмещенности. Вместе с тем оценка максимального правдоподобия, как отмечалось выше, зачастую эффективнее (во всяком случае не менее эффективна) оценки, полученной методом моментов, и более того, если существует несмещенная эффективная оценка параметра, то она будет получена методом максимального правдоподобия.

§ 8.3. Понятие интервальной оценки. Интервальные оценки параметров нормального закона, вероятности события и коэффициента корреляции

Вычисляя на основании результатов $x_1, x_2, \ldots, x_n$ наблюдений точечную оценку $\hat{\Theta}$ генеральной характеристики Θ, числовое значение которой не известно, мы понимаем, что $\hat{\Theta}$ является лишь приближением к этому значению. Если для большого числа наблюдений точность приближения бывает достаточной для практических выводов, то для выборок небольшого объема вопрос о точности оценок очень важен. В математической статистике он решается следующим образом.

Точечную оценку рассматривают как функцию случайных результатов $X_1, X_2, \ldots, X_n$ наблюдений случайной величины X, $\hat{\Theta}_{(n)} = f(X_1, X_2, \ldots, X_n)$, затем задаются вероятностью γ и по определенным правилам находят такое число $\varepsilon > 0$, при котором выполняется соотношение

$$P(\underbrace{\hat{\Theta}_{(n)} - \varepsilon}_{\Theta_1} < \Theta < \underbrace{\hat{\Theta}_{(n)} + \varepsilon}_{\Theta_2}) = \gamma. \qquad (8.43)$$

Соотношению (8.43) тождественно соотношение

$$P(|\hat{\Theta}_{(n)} - \Theta| < \varepsilon) = \gamma, \qquad (8.44)$$

из которого видно, что абсолютная погрешность оценки $\hat{\Theta}_{(n)}$ не превосходит числа ε. Это высказывание верно с вероятностью, равной γ. Числа Θ_1 и Θ_2 называются ***доверительными границами***, интервал (Θ_1, Θ_2) — ***доверительным интервалом*** или ***интервальной оценкой*** генеральной характеристики Θ, вероятность γ называется ***доверительной вероятностью*** или ***надежностью*** интервальной оценки.

В соотношении (8.43) случайными величинами являются доверительные границы Θ_1 и Θ_2. Ведь $\hat{\Theta}_{(n)}$ — это случайная величина. Генеральная же характеристика Θ — постоянная величина. Поэтому соотношение (8.43) следует читать так:

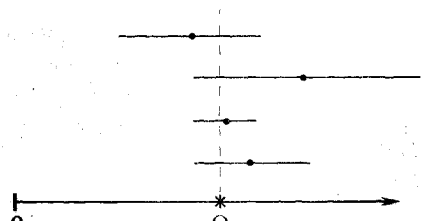


Рис. 8.3

«вероятность того, что интервал (Θ_1, Θ_2) накроет характеристику Θ, равна γ»; именно «интервал накроет характеристику», а не «характеристика попадет в интервал». На рисунке 8.3 друг над другом изображены доверительные интервалы характеристики Θ, построенные для разных выборок; центры интервалов — это выборочные значения оценки $\widehat{\Theta}_{(n)}$.

Надежность γ принято выбирать равной 0,95; 0,99; 0,999. Тогда событие, состоящее в том, что интервал (Θ_1, Θ_2) накроет характеристику Θ, будет практически достоверным; также практически достоверным будет событие, состоящее в том, что погрешность оценки $\widehat{\Theta}_{(n)}$ меньше ε.

В соотношении (8.43) границы Θ_1 и Θ_2 симметричны относительно точечной оценки $\widehat{\Theta}_{(n)}$. Обратим внимание на то, что не всегда удается построить границы с таким свойством.

Поскольку довольно часто встречаются нормально распределенные случайные величины, построим интервальные оценки для параметров нормального распределения — математического ожидания a и среднего квадратического отклонения σ.

Интервальные оценки параметров нормального распределения

Обозначим через X случайную величину, имеющую нормальный закон распределения с параметрами a и σ, т. е. $X = N(a, \sigma)$. Будем предполагать, что наблюдения над этой величиной независимы и проводятся в типичных условиях, т. е. возможные результаты $X_1, X_2, \ldots, X_n$ этих наблюдений обладают следующими свойствами:

$X_1, X_2, \ldots, X_n$ — независимые случайные величины; (8.45)

закон распределения любой из величин $X_1, X_2, \ldots, X_n$ совпадает с законом распределения величины X, т. е.

$$X_1 = N(a, \sigma), \quad X_2 = N(a, \sigma), \ldots, X_n = N(a, \sigma). \quad (8.46)$$

Построим интервальные оценки математического ожидания a и среднего квадратического отклонения σ.

Интервальная оценка математического ожидания нормального распределения при известной дисперсии. Итак, $X = N(a, \sigma)$, причем математическое ожидание a неизвестно, а дисперсия σ^2 известна. По наблюдениям найдем точечную оценку $\overline{X}_{(n)} = \sum_{i=1}^{n} X_i/n$ математического ожидания a.

Зададимся вероятностью γ и определим такое число ε, при котором выполнялось бы соотношение

$$P(\overline{X}_{(n)} - \varepsilon < a < \overline{X}_{(n)} + \varepsilon) = \gamma,$$

или

$$P(|\overline{X}_{(n)} - a| < \varepsilon) = \gamma. \qquad (8.47)$$

Обратим внимание на то, что равенство (8.47) аналогично равенству (8.44): параметр a — это Θ, а средняя $\overline{X}_{(n)}$ — это оценка $\hat{\Theta}_{(n)}$.

≫ Нахождение ε основано на утверждении, сформулированном в § 6.5: для среднего из независимых и нормально распределенных величин $X_1, X_2, \ldots, X_n$ имеет место равенство (6.34), а именно

$$\overline{X}_{(n)} = N\Big(\frac{1}{n}\sum_{i=1}^{n} MX_i, \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} DX_i}\Big).$$

(Заметим, что в силу центральной предельной теоремы, аналогичное равенство, но только при достаточно большом n, имеет место и в том случае, когда распределение величины X отлично от нормального (см. (6.33).)

Учитывая, что, согласно (8.46), $MX_i = a$, а $DX_i = \sigma^2$, $i = 1, 2, \ldots, n$, получим

$$\overline{X}_{(n)} = N\Big(\frac{1}{n}\sum_{i=1}^{n} a, \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2}\Big) = N(a, \sigma/\sqrt{n}). \qquad (8.48)$$

Но для нормально распределенной величины имеет место равенство (5.47), которое для величины $N(a, \sigma/\sqrt{n})$ принимает вид

$$P(|N(a, \sigma/\sqrt{n}) - a| < \varepsilon) = 2\Phi(\varepsilon/(\sigma/\sqrt{n})),$$

или

$$P(|\overline{X}_{(n)} - a| < \varepsilon) = 2\Phi(\varepsilon/(\sigma/\sqrt{n})). \qquad (8.49)$$

Сравнивая (8.49) с (8.47), видим, что $\Phi(\varepsilon/(\sigma/\sqrt{n})) = \gamma/2$. Отсюда получаем следующий алгоритм нахождения числа ε:

— воспользуемся таблицей П.1 функции Лапласа и при вероятности, равной $\gamma/2$; найдем такое число $z_{\gamma/2}$, при котором $\Phi(z_{\gamma/2}) = \gamma/2$;

— выражение $\varepsilon/(\sigma/\sqrt{n})$ приравняем числу $z_{\gamma/2}$, $\varepsilon/(\sigma/\sqrt{n}) = z_{\gamma/2}$, и найдем

$$\varepsilon = z_{\gamma/2}\sigma/\sqrt{n}. \quad \ll \qquad (8.50)$$

Подставив значение ε в (8.47), получим

$$P(|\overline{X}_{(n)} - a| < z_{\gamma/2}\sigma/\sqrt{n}) = \gamma,$$

или, раскрывая модуль,

$$P(\overline{X}_{(n)} - z_{\gamma/2}\sigma/\sqrt{n} < a < \overline{X}_{(n)} + z_{\gamma/2}\sigma/\sqrt{n}) = \gamma. \quad (8.51)$$

Интервал

$$(\overline{X}_{(n)} - z_{\gamma/2}\sigma/\sqrt{n}\,;\ \overline{X}_{(n)} + z_{\gamma/2}\sigma/\sqrt{n}) \quad (8.52)$$

и является интервальной оценкой математического ожидания a, соответствующей надежности γ.

Итак, если случайная величина X имеет нормальный закон распределения (или любой закон распределения, но тогда n должно быть достаточно большим) и значение ее дисперсии σ^2 известно (при большом n значение дисперсии σ^2 может быть и неизвестным, тогда его заменяют выборочной дисперсией $\hat{\sigma}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$), то:

с вероятностью γ можно быть уверенным, что погрешность среднего, т. е. погрешность, возникающая при замене математического ожидания a величины X средним $\overline{X}_{(n)} = \sum_{i=1}^{n} X_i/n$ меньше числа $\varepsilon = z_{\gamma/2}\sigma/\sqrt{n}$;

с вероятностью γ можно быть уверенным, что интервал (8.52) накроет математическое ожидание a величины X.

Приведем схему нахождения числа ε и доверительных границ, отвечающих надежности γ:

$$\gamma \to \Phi(z) = \gamma/2 \xrightarrow[\text{П.1}]{} z_{\gamma/2} \to \varepsilon = z_{\gamma/2}\sigma/\sqrt{n} \begin{cases} \to \overline{X}_{(n)} - \varepsilon \text{ (нижняя граница)}, \\ \to \overline{X}_{(n)} + \varepsilon \text{ (верхняя граница)}. \end{cases} \quad (8.53)$$

Очевидны следующие утверждения (они вытекают и из формул (8.50) и (8.52)):

— при увеличении числа наблюдений n и неизменной γ погрешность ε среднего арифметического уменьшается и границы доверительного интервала (8.52) сближаются;

— при увеличении надежности γ интервальной оценки и неизменном n погрешность ε среднего увеличивается (ведь увеличивается $z_{\gamma/2}$) и границы доверительного интервала (8.52) раздвигаются.

Сформулируем три типа задач, которые можно решать, используя равенство (8.49), и приведем их решения (во всех задачах предполагается, что значение дисперсии σ^2 известно).

1. Найти число $\varepsilon > 0$ такое, при котором $P(|\overline{X}_{(n)} - a| < \varepsilon) = \gamma$, если известны n и γ.

Р е ш е н и е. Имеем $\varepsilon = z_{\gamma/2}\sigma/\sqrt{n}$.

2. Найти вероятность γ того, что $|\overline{X}_{(n)} - a| < \varepsilon$, если известны n и ε.

Р е ш е н и е. Из (8.50) получаем

$$z_{\gamma/2} = \varepsilon\sqrt{n}\,/\sigma. \tag{8.54}$$

По таблице П. 1, найдем значение функции Лапласа при $z = z_{\gamma/2}$, т. е. найдем $\Phi(z_{\gamma/2})$; вероятность

$$\gamma = 2\Phi(z_{\gamma/2}). \tag{8.55}$$

3. Найти число наблюдений n, при котором $P(|\overline{X}_{(n)} - a| < \varepsilon) = \gamma$, если известны ε и γ.

Р е ш е н и е. Из (8.50) получаем

$$n = z^2_{\gamma/2}\,\sigma^2/\varepsilon^2. \tag{8.56}$$

ЗАДАЧА 8.2. Фирма коммунального хозяйства желает на основе выборки оценить среднюю плату за квартиру определенного типа с надежностью не менее 99% и погрешностью, меньшей 10 ден. ед. Предполагая, что квартирная плата имеет нормальное распределение со среднеквадратическим отклонением, не превышающим 35 ден. ед., найти минимальный объем выборки.

Р е ш е н и е. По условию требуется найти такое n, при котором $P(|\overline{X} - a| < 10) \geqslant 0{,}99$, где a и $\overline{X}$ — генеральное и выборочное среднее. Приравняв $\gamma = 0{,}99$, по таблице П. 1 найдем число $z_{\gamma/2}$, при котором $\Phi(z_{\gamma/2}) = \gamma/2 = 0{,}495$; $z_{0{,}495} = 2{,}60$. При $\varepsilon = 10$ и $\sigma = 35$ из (8.56) получаем

$$n = z^2_{0{,}495}\,\sigma^2/\varepsilon^2 \leqslant 2{,}60^2 \cdot 35^2/10^2 = 82{,}81.$$

Но так как с ростом γ и уменьшением ε объем выборки n растет, то при $\gamma \geqslant 0{,}99$ и $\varepsilon < 10$ $n > 82{,}81$ и $n_{\min} = 83$. ◀

З а м е ч а н и е. Если в задаче речь идет о конкретных результатах $x_1, x_2, \ldots, x_n$ наблюдений (т. е. о значениях случайных величин $X_1, X_2, \ldots, X_n$), то в (8.52) и (8.53) вместо случайного среднего $\overline{X}_{(n)}$ используют его фактическое значение $\bar{x}$. В этом случае границы доверительного интервала не случайные величины, а конкретные числа: $\bar{x} - z_{\gamma/2}\sigma/\sqrt{n}$, $\bar{x} + z_{\gamma/2}\sigma/\sqrt{n}$, а поскольку и a — постоянная величина, некорректность записи $P(\bar{x} - z_{\gamma/2}\sigma/\sqrt{n} < a < \bar{x} + z_{\gamma/2}\sigma/\sqrt{n}) = \gamma$ очевидна. Однако, так как *a не известно*, то можно говорить о степени уверенности в том, что интервал $(\bar{x} - z_{\gamma/2}\sigma/\sqrt{n},\ \bar{x} + z_{\gamma/2}\sigma/\sqrt{n})$ накроет a, и эта степень уверенности равна γ. Именно такую интерпретацию доверительного интервала с числовыми границами будем использовать далее.

Интервальная оценка математического ожидания нормального распределения при неизвестной дисперсии. Выше была решена задача построения интервальной оценки для математического ожидания нормального распределения, когда его дисперсия σ^2 известна. Решим эту же задачу, но в условиях, когда дисперсия изучаемого нормального распределения неизвестна.

Итак, случайная величина $X = N(a, \sigma)$, причем неизвестны ни a, ни σ^2. По наблюдениям $X_1, X_2, \ldots, X_n$ определим среднее $\overline{X}_{(n)} = \sum_{i=1}^{n} X_i/n$ и оценку $s^2_{(n)} = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$ дисперсии σ^2. Зададимся доверительной вероятностью γ и найдем такое число ε, при котором выполняется соотношение $P(\overline{X}_{(n)} - \varepsilon < a < \overline{X}_{(n)} + \varepsilon) = \gamma$, или

$$P(|\overline{X}_{(n)} - a| < \varepsilon) = \gamma. \tag{8.57}$$

≫ Нахождение ε основано на следующем утверждении: для среднего из независимых случайных величин, имеющих нормальный закон распределения с математическим ожиданием a и дисперсией σ^2 (здесь эти условия выполняются, см. (8.45) и (8.46)), имеет место следующее равенство:

$$\frac{\overline{X}_{(n)} - a}{s_{(n)}/\sqrt{n}} = T(n-1), \tag{8.58}$$

где $T(n-1)$ — случайная величина Стьюдента с числом степеней свободы, равным $n-1$ (см. § 5.3).

Воспользуемся таблицей П. 4 и найдем при $k = n-1$ и $p = 1-\gamma$ число $t_{k,p}$ такое, при котором

$$P(|T(k)| > t_{k,p}) = p, \text{ или } P(|T(k)| < t_{k,p}) = 1 - p.$$

Подставив в последнее равенство $k = n-1$ и $p = 1-\gamma$, получим $P(|T(n-1)| < t_{n-1,\,1-\gamma}) = \gamma$, или, учитывая (8.58),

$$P\left(\left|\frac{\overline{X}_{(n)} - a}{s_{(n)}/\sqrt{n}}\right| < t_{n-1,\,1-\gamma}\right) = \gamma.$$

Отсюда

$$P(|\overline{X}_{(n)} - a| < t_{n-1,\,1-\gamma} s_{(n)}/\sqrt{n}) = \gamma. \tag{8.59}$$

Сравнив **(8.59)** и (8.57), заключаем, что

$$\varepsilon = t_{n-1,\,1-\gamma} s_{(n)}/\sqrt{n}. \quad \ll \tag{8.60}$$

Следовательно,

$$P(\overline{X}_{(n)} - t_{n-1,\,1-\gamma} s_{(n)}/\sqrt{n} < a < \overline{X}_{(n)} + t_{n-1,\,1-\gamma} s_{(n)}/\sqrt{n}) = \gamma, \tag{8.61}$$

т. е. интервал

$$(\overline{X}_{(n)} - t_{n-1,\,1-\gamma}s_{(n)}/\sqrt{n}\,;\ \overline{X}_{(n)} + t_{n-1,\,1-\gamma}s_{(n)}/\sqrt{n}\,) \tag{8.62}$$

и является интервальной оценкой математического ожидания a, соответствующей вероятности γ.

Итак, если случайная величина X имеет нормальный закон распределения и значение ее дисперсии σ^2 неизвестно, то:

— с вероятностью γ можно быть уверенным в том, что погрешность выборочного среднего меньше числа $\varepsilon = t_{n-1,\,1-\gamma}s_{(n)}/\sqrt{n}$;

— с вероятностью γ можно быть уверенным в том, что интервал (8.62) накроет математическое ожидание a величины X.

Приведем схему нахождения числа ε и доверительных границ, отвечающих надежности γ:

$$\left.\begin{array}{l} n \to k = n-1, \\ \gamma \to p = 1-\gamma \end{array}\right\} \xrightarrow[\text{П.4}]{} t_{k,p} \to \varepsilon = t_{k,p}s_{(n)}/\sqrt{n} \begin{cases} \to \overline{X}_{(n)} - \varepsilon \text{ (нижняя граница)}, \\ \to \overline{X}_{(n)} + \varepsilon \text{ (верхняя граница)}. \end{cases} \tag{8.63}$$

Напомним, что в (8.63)

$$s_{(n)} = \sqrt{\sum_{i=1}^{n}(X_i - \overline{X}_{(n)})^2/(n-1)}\,.$$

Очевидно следующее утверждение (оно вытекает также из формул (8.60) и (8.62)): при увеличении надежности γ интервальной оценки и неизменных n и s погрешность ε выборочного среднего увеличивается (ведь увеличивается $t_{n-1,\,1-\gamma}$), и границы доверительного интервала (8.62) раздвигаются.

Замечание. Как повлияет увеличение n на погрешность среднего, сказать нельзя, поскольку увеличение n может привести как к увеличению, так и к уменьшению значения оценки $s_{(n)}$, вычисляемого по результатам наблюдений ($s_{(n)}$ в отличие от σ не является постоянной величиной).

Используя равенство (8.60), можно решать задачи следующих двух типов (предполагаем, что значение дисперсии σ^2 неизвестно, но известны n и значение оценки $s^2_{(n)}$):

1. Найти число $\varepsilon > 0$ такое, при котором $P(|\overline{X}_{(n)} - a| < \varepsilon) = \gamma$, если известна вероятность γ.

Р е ш е н и е. Имеем $\varepsilon = t_{n-1,\,1-\gamma} s_{(n)} / \sqrt{n}$.

2. Найти вероятность γ того, что $|\overline{X}_{(n)} - a| < \varepsilon$, если известно число ε.

Р е ш е н и е. Из (8.60) получаем

$$t_{n-1,\,1-\gamma} = \varepsilon \sqrt{n} / s_{(n)}. \qquad (8.64)$$

Обратившись к таблице П. 4, в строке $\boldsymbol{k = n - 1}$ найдем число $t_{n-1,\,p}$, тогда $\gamma = 1 - p$.

З а м е ч а н и е. Если в задаче речь идет о конкретных результатах наблюдений, то в (8.62) и (8.63) вместо случайного среднего $\overline{X}_{(n)}$ используют его фактическое значение $\bar{x}$.

Ответим на вопрос, реализовано ли интервальное оценивание математического ожидания в Microsoft Excel. В § 7.2 дано описание работы программы **«Описательная статистика»** пакета **«Анализ данных»**. Напомним, что в окне ввода исходных параметров задается в том числе и *«уровень надежности»* (по умолчанию 0,95) — это и есть надежность интервальной оценки математического ожидания, или доверительная вероятность γ. В последней строке результатов работы программы (см. рис. 7.5) приводится число ε, рассчитанное по формуле (8.60).

▶ **ПРИМЕР 8.8.** По данным примера 7.2 об объемах ежедневных продаж товара дилером за $n = 100$ дней построим 95%-ю интервальную оценку генерального среднего объема продаж за день (a), предположив, что ежедневный объем продаж X — нормально распределенная величина, $X = N(a, \sigma)$.

Воспользуемся результатами работы программы **«Описательная статистика»**, приведенными на рисунке 7.5, *б*. Средний объем продаж — «среднее» $\bar{x} = 49{,}596$, исправленная выборочная дисперсия — «дисперсия выборки» $s^2 = 117{,}813$; «стандартное отклонение» $s = 10{,}8542$. В соответствии с (8.63) найдем ε:

$$\left.\begin{array}{l} n = 100 \to k = 99, \\ \gamma = 0{,}95 \to p = 0{,}05 \end{array}\right\} \xrightarrow[\text{п.4}]{} t_{99;\,0{,}05} \approx 1{,}98 \to \varepsilon =$$

$$= 1{,}98 \cdot 10{,}8542 / \sqrt{100} = 2{,}15.$$

(Такой же результат получен программой **«Описательная статистика»**, см. рисунок 7.5, *б*, строку *«Уровень надежности»* — 2,15371.) 95%-я интервальная оценка такова: (49,596 – 2,154; 49,596 + 2,154), или (47,442; 51,75).

Итак, с вероятностью 0,95 можно ожидать, что интервал (47,442; 51,75) накроет неизвестное значение генерального среднего объема продаж за день. ◀|

▶▶ **ЗАДАЧА 8.3.** Для отрасли, включающей $N = 1200$ фирм, была составлена случайная выборка из $n = 10$ фирм. По выборочным данным оказалось, что в фирме в среднем работают 77,5 человека, а дисперсия равна 360 (чел.^{2}). Используя 90%-й доверительный интервал, оцените: а) среднее число работающих в фирме для отрасли в целом; б) общее число работающих в отрасли. Предполагается, что число X работающих в фирме — нормально распределенная случайная величина, $X = N(a, \sigma)$.

Р е ш е н и е. а) По условию выборочная дисперсия $\hat{\sigma}^2 = 360$; исправленная выборочная дисперсия $s^2 = \hat{\sigma}^2 n/(n-1) = 360 \cdot 10/9 = 400$; $s = 20$ (чел.). В соответствии с (8.63) находим

$$\left.\begin{array}{l} n = 10 \to k = 9, \\ \gamma = 0{,}9 \to p = 0{,}1 \end{array}\right\} \xrightarrow[\text{П.4}]{} t_{9;\,0{,}1} = 1{,}833 \to \varepsilon =$$

$$= 1{,}833 \cdot 20/\sqrt{10} = 11{,}59.$$

Границы доверительного интервала $\bar{x} - \varepsilon = 77{,}5 - 11{,}59 = 65{,}91$ и $\bar{x} + \varepsilon = 77{,}5 + 11{,}59 = 89{,}09$. Итак, с надежностью 0,9 можно ожидать, что интервал (65,91; 89,09) накроет неизвестное среднее число a работающих в фирме для отрасли в целом.

б) Так как число фирм в отрасли $N = 1200$, и для неизвестного числа a работающих в среднем в одной фирме имеет место (с надежностью 0,9) соотношение $65{,}91 < a < 89{,}09$, то для общего числа aN работающих в отрасли имеет место (с такой же надежностью) соотношение $65{,}91N < aN < 89{,}09N$, или $79\,092 < a \cdot 1200 < 106\,908$. Таким образом, с вероятностью 0,9 можно ожидать, что интервал (79 092; 106 908) накроет неизвестное общее число работающих в отрасли.

ЗАДАЧА 8.4. По результатам измерения диаметра 25 корпусов электродвигателей было получено, что $\bar{x} = 100$ мм, $s = 16$ мм. Предполагая нормальное распределение результата измерения, найдите вероятность того, что интервал $(0{,}9\bar{x};\ 1{,}1\bar{x})$ накроет (генеральный) средний диаметр корпуса.

Р е ш е н и е. Приравняв нижнюю границу, равную $0{,}9\bar{x}$, нижней границе в схеме (8.63), получаем $0{,}9\bar{x} = \bar{x} - \varepsilon$. Отсю-

да $\varepsilon = 0{,}1\bar{x} = 0{,}1 \cdot 100 = 10$. Такой же результат будем иметь, если приравнять $1{,}1\bar{x}$ верхней границе, равной $\bar{x} + \varepsilon$.

Используя (8.64), получим $t_{24,\,1-\gamma} = 10 \cdot 5/16 = 3{,}125$. Обратимся к таблице П. 4. В строке $k = 24$ найдем число, ближайшее к 3,125 — это число 3,091; ему соответствует $p = 0{,}005$, т. е. $1 - \gamma = 0{,}005$ и $\gamma \approx 0{,}995$. Таким образом, с вероятностью 0,995 можно быть уверенным в том, что интервал $(0{,}9\bar{x}\,;\, 1{,}1\bar{x})$ накроет неизвестное значение генерального среднего размера диаметра. ◀

Интервальная оценка среднего квадратического отклонения и дисперсии нормального распределения при неизвестном математическом ожидании a. Мы рассматриваем нормально распределенную случайную величину X, дисперсия σ^2 и математическое ожидание a которой неизвестны. Проведено n наблюдений этой величины, возможные результаты которых $X_1, X_2, \ldots, X_n$ обладают свойствами (8.45) и (8.46). По этим результатам найдем среднее $\overline{X}_{(n)} = \sum_{i=1}^{n} X_i/n$, оценку $s^2_{(n)} = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$ неизвестной дисперсии σ^2 и оценку $s_{(n)} = +\sqrt{s^2_{(n)}}$ среднего квадратического отклонения σ.

Для построения интервальной оценки используют два подхода: при первом интервальная оценка среднего квадратического отклонения σ имеет границы, симметричные относительно $s_{(n)}$; при втором — границы таким свойством не обладают. И тот, и другой подход основаны на следующем утверждении: если наблюдения нормально распределенной величины независимы и проводятся в типичных условиях [выполняются условия (8.45) и (8.46)], то для величины $s^2_{(n)}$, рассматриваемой как случайной, имеет место следующее соотношение:

$$(n-1)s^2_{(n)}/\sigma^2 = \chi^2(n-1), \qquad (8.65)$$

где $\chi^2(n-1)$ — случайная величина χ^2 с числом степеней свободы $n-1$ (см. § 5.3). Напомним, что с распределением χ^2 связаны таблицы П. 2 и П. 3.

Первый подход. Зададимся надежностью γ интервальной оценкой и найдем такое число ε, при котором выполняется соотношение $P(|s_{(n)} - \sigma| < \varepsilon) = \gamma$, или

$$P(s_{(n)} - \varepsilon < \sigma < s_{(n)} + \varepsilon) = \gamma. \qquad (8.66)$$

≫ Среднее квадратическое отклонение σ случайной величины всегда положительно ($\sigma > 0$), поэтому ε разумнее находить не из условия (8.66), а из условия

$$P[\max(0; s_{(n)} - \varepsilon) < \sigma < s_{(n)} + \varepsilon] = \gamma. \qquad (8.67)$$

Действительно, если $s_{(n)} - \varepsilon \leqslant 0$, то максимум из чисел 0 и $(s_{(n)} - \varepsilon)$ равен нулю и нижней границей для σ является нуль; если $s_{(n)} - \varepsilon > 0$, то нижней границей для σ является $s_{(n)} - \varepsilon$.

Используя (8.65), можно доказать, что

$$\varepsilon = s_{(n)}\delta_{n-1,\,1-\gamma}, \qquad (8.68)$$

где $\delta_{n-1,\,1-\gamma}$ находится по таблице П. 3 при $k = n - 1$ и $p = 1 - \gamma$. ≪

Следовательно, соотношение (8.67) имеет вид

$$P[\max(0; s_{(n)} - s_{(n)}\delta_{n-1,\,1-\gamma}) < \sigma < s_{(n)} + s_{(n)}\delta_{n-1,\,1-\gamma}] = \gamma,$$

или

$$P[s_{(n)} \max(0; 1 - \delta_{n-1,\,1-\gamma}) < \sigma < s_{(n)}(1 + \delta_{n-1,\,1-\gamma})] = \gamma, \qquad (8.69)$$

т. е. интервал

$$(s_{(n)} \max(0; 1 - \delta_{n-1,\,1-\gamma});\ s_{(n)}(1 + \delta_{n-1,\,1-\gamma})) \qquad (8.70)$$

и является интервальной оценкой среднего квадратического отклонения σ, соответствующей вероятности γ. ≪

Приведем схему нахождения числа ε и границ интервальной оценки среднего квадратического отклонения σ:

$$\left.\begin{array}{l} n \to k = n - 1, \\ \gamma \to p = 1 - \gamma, \end{array}\right\} \xrightarrow[\text{П.3}]{} \delta_{k,\,p} \to \varepsilon = s_{(n)}\delta_{k,\,p} \longrightarrow \begin{cases} s_{(n)} \max(0;\ 1 - \delta_{k,\,p}) \text{ (нижняя граница)}, \\ s_{(n)}(1 + \delta_{k,\,p}) \text{ (верхняя граница)}. \end{cases} \qquad (8.71)$$

Отвечающая надежности γ интервальная оценка дисперсии σ^2 имеет вид

$$(s^2_{(n)} \max^2(0; 1 - \delta_{n-1,\,1-\gamma});\ s^2_{(n)}(1 + \delta_{n-1,\,1-\gamma})^2). \qquad (8.72)$$

С вероятностью, равной γ, можно ожидать, что интервал (8.72) накроет неизвестное значение дисперсии σ^2.

Второй подход. Этот подход ориентирован на использование таблицы П. 2 и предполагает при заданной вероятности γ нахождение таких случайных величин $b_1 > 0$ и $b_2 > 0$, при которых

$$P(b_1 < \sigma < b_2) = \gamma, \qquad (8.73)$$

при этом условие симметричности границ b_1 и b_2 относительно $s_{(n)}$ не считается обязательным.

≫ Используя (8.65), можно доказать, что

$$b_1 = \sqrt{s^2_{(n)}(n-1)/\chi^2_{n-1,(1-\gamma)/2}},$$
$$b_2 = \sqrt{s^2_{(n)}(n-1)/\chi^2_{n-1,(1+\gamma)/2}}, \qquad (8.74)$$

где $\chi^2_{n-1,(1-\gamma)/2}$ и $\chi^2_{n-1,(1+\gamma)/2}$ находят по таблице П. 2 при $k = n-1$ и соответственно при $p = (1-\gamma)/2$ и $p = (1+\gamma)/2$. ≪

Итак, с вероятностью γ можно утверждать, что интервал

$$\left(\sqrt{s^2_{(n)}(n-1)/\chi^2_{n-1,(1-\gamma)/2}};\ \sqrt{s^2_{(n)}(n-1)/\chi^2_{n-1,(1+\gamma)/2}}\right) \qquad (8.75)$$

накроет неизвестное значение среднего квадратического отклонения σ.

Приведем схему нахождения границ интервальной оценки среднего квадратического отклонения σ:

$$\left.\begin{array}{l} n \to k = n-1, \\ \gamma \to p = (1-\gamma)/2 \end{array}\right\} \xrightarrow{\text{П.2}} \chi^2_{k,p} \to b_1 = \sqrt{s^2_{(n)}(n-1)/\chi^2_{k,p}};$$

$$\left.\begin{array}{l} \gamma \to p = (1+\gamma)/2, \\ n \to k = n-1 \end{array}\right\} \xrightarrow{\text{П.2}} \chi^2_{k,p} \to b_2 = \sqrt{s^2_{(n)}(n-1)/\chi^2_{k,p}}. \qquad (8.76)$$

Отвечающая надежности γ интервальная оценка **дисперсии** σ^2 имеет вид

$$\left(s^2_{(n)}(n-1)/\chi^2_{n-1,(1-\gamma)/2};\ s^2_{(n)}(n-1)/\chi^2_{n-1,(1+\gamma)/2}\right). \qquad (8.77)$$

Обратим внимание на следующее: рассмотренные подходы дают различающиеся между собой интервальные оценки; если в задаче речь идет о погрешности оценки $s_{(n)}$, возникающей при замене этой оценкой неизвестного среднего квадратического отклонения σ, то следует использовать первый подход.

▶ **ПРИМЕР 8.9.** По данным примера 7.2 об объемах ежедневных продаж товара дилером за $n = 100$ дней найти верхнюю границу погрешности оценки $s_{(n)}$, гарантируемую с вероятностью 0,95, и оценить с 95%-й надежностью среднее квадратическое отклонение и дисперсию объема ежедневных продаж. Предполагается, что ежедневный объем продаж X — нормально распределенная величина, $X = N(a, \sigma)$.

Воспользуемся результатами работы программы «**Описательная статистика**», приведенными на рисунке 7.5, *б*.

Исправленная выборочная дисперсия — «дисперсия выборки» $s^2 = 117{,}813$; «стандартное отклонение» $s = 10{,}8542$.

Первый подход. В соответствии с (8.71) найдем

$$\left.\begin{array}{l} n = 100 \to k = 99, \\ \gamma = 0{,}95 \to p = 0{,}05 \end{array}\right\} \xrightarrow[\text{П.3}]{} \delta_{99;\,0{,}05} \approx 0{,}146 \to$$

$$\to \varepsilon = 10{,}8542 \cdot 0{,}146 = 1{,}585,$$

т. е. с вероятностью 0,95 можно ожидать, что погрешность оценки s меньше 1,585.

Границы интервальной оценки среднего квадратического отклонения σ таковы:

$$s\max(0;\ 1 - \delta_{99;\,0{,}05}) = 10{,}8542\max(0;\ 1 - 0{,}146) =$$

$$= 10{,}8542 \cdot 0{,}854 = 9{,}269;$$

$$s(1 + \delta_{99;\,0{,}05}) = 10{,}8542 \cdot 1{,}146 = 12{,}439.$$

Итак, с вероятностью 0,95 можно ожидать, что интервал (9,269; 12,439) накроет неизвестное значение среднего квадратического отклонения σ; с такой же вероятностью можно ожидать, что интервал $(9{,}269^2;\ 12{,}439^2)$, или (85,914; 154, 729), накроет неизвестное значение дисперсии σ^2.

Второй подход. В соответствии с (8.76) найдем

$$\left.\begin{array}{l} n = 100 \to k = 99, \\ \gamma = 0{,}95 \to p = (1 - 0{,}95)/2 = 0{,}025 \end{array}\right\} \xrightarrow[\text{П.2}]{} \chi^2_{99;\,0{,}025} =$$

$$= 129{,}56 \to b_1 = \sqrt{117{,}813 \cdot 99/129{,}56} = 9{,}49;$$

$$\left.\begin{array}{l} \gamma = 0{,}95 \to p = (1 + 0{,}95)/2 = 0{,}975 \\ n = 100 \to k = 99 \end{array}\right\} \xrightarrow[\text{П.2}]{} \chi^2_{99;\,0{,}975} =$$

$$= 74{,}22 \to b_2 = \sqrt{117{,}813 \cdot 99/74{,}22} = 12{,}54.$$

Таким образом, с вероятностью 0,95 можно ожидать, что интервалы (9,49; 12,54) и $(9{,}49^2;\ 12{,}54^2)$ накроют соответственно σ и σ^2. ◀|

▶▶ **ЗАДАЧА 8.5.** С какой вероятностью можно ожидать, что при случайной выборке объемом $n = 13$ из нормально распределенной генеральной совокупности погрешность, возникающая при замене среднего квадратического отклонения σ оценкой $s_{(n)}$, меньше $0{,}388s_{(n)}$?

Решение. Речь идет о погрешности оценки $s_{(n)}$, поэтому воспользуемся первым подходом к построению интервальной оценки среднего квадратического отклонения σ. Требуется найти $P(|s_{(n)} - \sigma| < 0{,}388s_{(n)})$. Приравняем $0{,}388s_{(n)}$ числу ε, рассчитываемому по формуле (8.68):

$0{,}388s_{(n)} = s_{(n)}\delta_{n-1,\,1-\gamma}$. Получим $\delta_{12,\,1-\gamma} = 0{,}388$. Обратимся к таблице П. 3: в строке $k = 12$ найдем число, равное 0,388; ему соответствует $p = 0{,}1$, т. е. $0{,}388 = \delta_{12;\,0{,}1} = \delta_{12;\,1-\gamma}$. Отсюда получим $0{,}1 = 1 - \gamma$ и $\gamma = 0{,}9$. Итак, $P(|s_{(n)} - \sigma| < 0{,}388s_{(n)}) = 0{,}9$. ◀

Интервальная оценка вероятности события

Было показано, что хорошей точечной оценкой вероятности p события A является относительная частота появления события A, или частость $\hat{p}_{(n)} = m/n$, где n — общее число независимых испытаний, в каждом из которых событие A может произойти с вероятностью p (напомним, что серия испытаний подобного типа называется последовательностью испытаний Бернулли), а m — число испытаний, в которых произойдет событие A.

Зададимся вероятностью γ и найдем случайные границы p_1 и p_2 такие, чтобы выполнялось соотношение

$$P(p_1 < p < p_2) = \gamma. \tag{8.78}$$

Интервал (p_1, p_2) и является интервальной оценкой вероятности p, отвечающей надежности γ.

Интервальную оценку построим для двух случаев: когда число испытаний Бернулли велико и для любого числа испытаний.

Интервальная оценка вероятности при большом числе испытаний Бернулли. Так как A — случайное событие, то количество m появлений события A в n испытаниях случайно. Будем m — число появлений события A в n испытаниях Бернулли интерпретировать как случайную величину; тогда и частость $\hat{p}_{(n)} = m/n$ — случайная величина. В § 6.3 было показано, что при большом числе n испытаний Бернулли имеет место приближенное равенство (6.42), которое, учитывая, что $\hat{p}_{(n)} = m/n$, запишем в виде

$$\hat{p}_{(n)} \approx N(p,\ \sqrt{p(1-p)/n}),$$

откуда следует [см. (6.42)], что

$$P(|\hat{p}_{(n)} - p| < \varepsilon) \approx 2\Phi(\varepsilon/\sqrt{p(1-p)/n}). \tag{8.79}$$

Найдем границы p_1 и p_2 интервальной оценки вероятности p.

≫ Приравняв правую часть равенства (8.79) заданной вероятности γ, получим

$$\Phi(\varepsilon/\sqrt{p(1-p)/n}) = \gamma/2. \tag{8.80}$$

Используя таблицу П. 1 функции Лапласа, найдем такое число $z_{\gamma/2}$, при котором $\Phi(z_{\gamma/2}) = \gamma/2$. Сравнив последнее равенство с (8.80), заключаем, что $\varepsilon/\sqrt{p(1-p)/n} = z_{\gamma/2}$, или

$$\varepsilon = z_{\gamma/2}\sqrt{p(1-p)/n}\,. \tag{8.81}$$

Учитывая (8.81) и (8.80), запишем равенство (8.79) в виде

$$P(|\hat{p}_{(n)} - p| < z_{\gamma/2}\sqrt{p(1-p)/n}\,) \approx \gamma. \tag{8.82}$$

Неравенство, стоящее в скобках этого выражения, решим относительно p (индекс (n) у частости $\hat{p}_{(n)}$ писать не будем). Для этого возведем его в квадрат:

$$(\hat{p} - p)^2 < z_{\gamma/2}^2\, p(1-p)/n.$$

Возведем $(\hat{p} - p)$ в квадрат и перенесем все члены в левую часть неравенства. Получим

$$(1 + z_{\gamma/2}^2/n)p^2 - (2\hat{p} + z_{\gamma/2}^2/n)p + \hat{p}^2 < 0. \tag{8.83}$$

Найдем корни p_1 и p_2 квадратного трехчлена, стоящего в правой части неравенства (8.83). Имеем

$$\begin{aligned} p_1 &= \frac{\hat{p} + z_{\gamma/2}^2/(2n) - z_{\gamma/2}\sqrt{\hat{p}(1-\hat{p})/n + z_{\gamma/2}^2/(4n^2)}}{1 + z_{\gamma/2}^2/n}, \\ p_2 &= \frac{\hat{p} + z_{\gamma/2}^2/(2n) + z_{\gamma/2}\sqrt{\hat{p}(1-\hat{p})/n + z_{\gamma/2}^2/4(n^2)}}{1 + z_{\gamma/2}^2/n}. \end{aligned} \tag{8.84}$$

Так как коэффициент $(1 + z_{\gamma/2}^2/n)$ квадратного трехчлена в неравенстве (8.83) положителен, то решением этого неравенства является интервал (p_1, p_2). ≪

Таким образом, при определении p_1 и p_2 по формулам (8.84) $P(p_1 < p < p_2) \approx \gamma$.

Приведем схему нахождения границ интервальной оценки вероятности p по формулам (8.84):

$$\left.\begin{array}{l} n, \\ m \end{array}\right\} \longrightarrow \hat{p} = \frac{m}{n}, \qquad \left.\begin{array}{l} \gamma \to \gamma/2 \xrightarrow[\text{П.1}]{} z_{\gamma/2} \end{array}\right\} \longrightarrow \begin{array}{l} \to p_1 \text{ (нижняя граница)}, \\ \to p_2 \text{ (верхняя граница)}. \end{array} \tag{8.85}$$

Формулы (8.84) для нахождения границ p_1 и p_2 интервальной оценки вероятности p обычно используют при n, близком к 100; на практике обычно при n значительно большем 100 ($n \gg 100$) используют получающиеся из этих формул предельные значения границ p_1 и p_2:

$$p_1 = \hat{p} - z_{\gamma/2}\sqrt{\hat{p}(1-\hat{p})/n}\,, \quad p_2 = \hat{p} + z_{\gamma/2}\sqrt{\hat{p}(1-\hat{p})/n}\,. \tag{8.86}$$

В этом случае с вероятностью γ интервал

$$(\hat{p} - z_{\gamma/2}\sqrt{\hat{p}(1-\hat{p})/n}\,,\ \hat{p} + z_{\gamma/2}\sqrt{\hat{p}(1-\hat{p})/n}\,) \qquad (8.87)$$

накроет неизвестную вероятность p, или, иначе, с вероятностью γ можно ожидать, что вычисленная по результатам n испытаний Бернулли частость $\hat{p} = m/n$ определяет значение неизвестной вероятности p с погрешностью, меньшей числа

$$\varepsilon = z_{\gamma/2}\sqrt{\hat{p}(1-\hat{p})/n}\,. \qquad (8.88)$$

Схема нахождения границ интервальной оценки (8.87), отвечающей надежности γ, такова:

$$\gamma \to \gamma/2 \xrightarrow[\text{П.1}]{} z_{\gamma/2} \to \varepsilon = z_{\gamma/2}\sqrt{\hat{p}(1-\hat{p})/n} \begin{cases} \to \hat{p} - \varepsilon \text{ (нижняя граница)}, \\ \to \hat{p} + \varepsilon \text{ (верхняя граница)}, \end{cases} \qquad (8.89)$$

где $\hat{p} = m/n$.

Интервальная оценка вероятности при любом числе n испытаний Бернулли. Построение интервальной оценки основано на следующем известном читателю утверждении, верном при любом n: если проводится n испытаний Бернулли и при этом вероятность появления события A в единичном испытании равна p, то количество испытаний m, в которых появится событие A, является случайной величиной с биномиальным законом распределения, т. е. вероятность

$$P(m = x) = C_n^x p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n. \qquad (8.90)$$

Можно показать, что при использовании биномиального закона (8.90) нижняя граница p_1 интервальной оценки является решением уравнения

$$\sum_{x=0}^{m-1} C_n^x p_1^x (1-p_1)^{n-x} = (1+\gamma)/2, \qquad (8.91)$$

а верхняя граница p_2 — решением уравнения

$$\sum_{x=0}^{m} C_n^x p_2^x (1-p_2)^{n-x} = (1-\gamma)/2. \qquad (8.92)$$

Существуют специальные таблицы для нахождения значений границ p_1 и p_2, удовлетворяющих уравнениям (8.91) и (8.92), по заданным n, $n - m$ и γ. Фрагмент их дан в таблице П. 6.

Этот способ нахождения доверительных границ p_1 и p_2 обычно используют при малом или небольшом числе испы-

таний n, т. е. когда приближенное равенство (8.79) дает значительную погрешность, следовательно, формулы (8.84) и (8.86) не могут быть применены.

▶ **ЗАДАЧА 8.6.** Городская статистика раскрываемости преступлений показала, что из 150 преступлений было раскрыто 100. Постройте 95%-ю интервальную оценку вероятности p раскрываемости преступления.

Р е ш е н и е. Имеем $n = 150$, $m = 100$, $\hat{p} = m/n = 2/3$, $\gamma = 0{,}95$. Построим интервальные оценки изложенными выше способами и сравним полученные результаты:

— воспользуемся формулами (8.91) и (8.92) и найдем точные значения 95%-х границ вероятности p раскрываемости преступления. Из таблицы П. 6 при $m = 100$ и $n - m = 50$ получаем $p_1 = 0{,}585$ и $p_2 = 0{,}741$;

— обратимся к формулам (8.84), которые, как было отмечено, дают хорошие результаты при числе испытаний, близком к 100. Положив в формулах (8.84) $\hat{p} = 2/3$, $n = 150$ и $z_{\gamma/2} = z_{0{,}475} = 1{,}95$, получаем $p_1 = 0{,}588$ и $p_2 = 0{,}737$ (отличия от точных значений границ невелики);

— используем формулы (8.86), которые применяют при n, существенно большем 100. Получаем $p_1 = 0{,}592$ и $p_2 = 0{,}742$ (отличия от точных значений границ также невелики).

ЗАДАЧА 8.7. Из $n = 1000$ случайно отобранных деталей оказалось $m = 50$ нестандартных. Предположив, что при отборе соблюдаются условия испытаний Бернулли, определить вероятность γ того, что интервал (0,04; 0,06) накроет неизвестную вероятность p появления нестандартной детали.

Р е ш е н и е. Здесь $n = 1000$, что значительно больше 100. Поэтому для нахождения вероятности γ воспользуемся схемой (8.89).

В условиях задачи точечная оценка вероятности p равна $\hat{p} = 50/1000 = 0{,}05$, а следовательно, в (8.89) доверительные границы такие: $0{,}05 - \varepsilon$ и $0{,}05 + \varepsilon$. Приравняв их границам заданного интервала (0,04; 0,06), получим $0{,}05 - \varepsilon = 0{,}04$ и $0{,}05 + \varepsilon = 0{,}06$. Оба уравнения дают одинаковое решение $\varepsilon = 0{,}01$ (если решения не одинаковы, то ответить на вопрос задачи используемым здесь способом было бы нельзя). Из формулы (8.88) получаем

$$z_{\gamma/2} = \varepsilon / \sqrt{\hat{p}(1-\hat{p})/n} = 0{,}01/\sqrt{0{,}05 \cdot 0{,}95/1000} = 1{,}45.$$

Зная $z_{\gamma/2} = 1{,}45$, по таблице П. 1 найдем $\gamma/2 = \Phi(1{,}45) = 0.4265$ и $\gamma = 0.85$. Таким образом, с вероятностью 0,85

можно ожидать, что интервал (0,04; 0,06) накроет неизвестное значение вероятности p появлений нестандартной детали. ◀

Интервальная оценка коэффициента корреляции

В п. 4.3.2 введено понятие коэффициента корреляции $r_{X,Y}$ как характеристики взаимосвязи случайных величин X и Y и приведены формулы (4.53) и (4.54) его вычисления. В § 7.2 рассмотрено понятие выборочного коэффициента корреляции $\hat{r}_{X,Y}$ и приведены формулы (7.32) и (7.33) его вычисления по результатам (x_1, y_1), (x_2, y_2), ... , (x_n, y_n) парных наблюдений двумерной случайной величины (X, Y).

Выборочный коэффициент корреляции $\hat{r}_{X,Y}$ — это точечная оценка (генерального) коэффициента корреляции $r_{X,Y}$.

Приведем интервальную оценку коэффициента корреляции $r_{X,Y}$, которую используют при достаточно большом числе n наблюдений ($n > 50$). С вероятностью, равной γ, можно ожидать, что интервал

$$\left(\hat{r}_{X,Y} - \frac{1-\hat{r}^2_{X,Y}}{\sqrt{n}} z_{\gamma/2};\ \hat{r}_{X,Y} + \frac{1-\hat{r}^2_{X,Y}}{\sqrt{n}} z_{\gamma/2}\right) \qquad (8.93)$$

накроет коэффициент корреляции $r_{X,Y}$; $z_{\gamma/2}$ — число, при котором $\Phi(z_{\gamma/2}) = \gamma/2$ (см. таблицу П. 1)[1].

УПРАЖНЕНИЯ

1. Методами моментов и максимального правдоподобия найдите оценки параметров известных вам законов распределения.

2. По выборке объемом $n = 42$ из нормальной генеральной совокупности найдено среднее $\bar{x} = 608$. Предположив, что $\sigma = 15$, определите: а) среднее квадратическое отклонение среднего; каков содержательный смысл этого отклонения; б) 95%-й доверительный интервал для математического ожидания a; в) вероятность того, что интервал $(0{,}992\bar{x};\ 1{,}008\bar{x})$ накроет a; г) вероятность того, что погрешность выборочного среднего меньше 4,5; д) объем выборки, при котором с надежностью 95% погрешность среднего меньше 4,5.

3. Для отрасли, включающей 1200 фирм, составлена случайная выборка из 10 фирм. По выборочным данным оказалось, что в фирме работают в среднем 77,5 человека, а дисперсия равна 360 (чел.^{2}).

---

[1] Строго говоря, это утверждение имеет место, если наблюдения (x_1, y_1), (x_2, y_2), ... , (x_n, y_n) двумерной величины (X, Y) независимы, проведены в типичных условиях, а сама величина (X, Y) подчиняется двумерном нормальному закону [9.13].

а) Используя 95%-й доверительный интервал, оцените среднее число работающих в фирме по всей отрасли и общее число работающих в отрасли; объясните отличие полученных результатов от результатов решения задачи 8.3;

б) используя 90%-й доверительный интервал, оцените среднее квадратическое отклонение работающих в фирме по всей отрасли.

4. Фирма для установления известности ее продукции опросила на каждой из пяти улиц по 40 человек. Количество знакомых с продукцией фирмы оказалось таким: 20, 10, 30, 10, 15.

а) Методами моментов и максимального правдоподобия оцените степень известности продукции фирмы;

б) постройте 90%-й и 95%-й доверительные интервалы для степени известности продукции. Какой из интервалов шире и почему;

в) пользуясь 95%-м интервалом, оцените число жителей среди 2000, знакомых с продукцией фирмы.

5. У 60 юношей одинакового возраста измерены рост и вес и по этим данным рассчитан коэффициент корреляции $\hat{r} = 0{,}75$. Постройте 95%-ю интервальную оценку коэффициента корреляции между ростом и весом для всех юношей того же возраста.

ГЛАВА 9

Проверка статистических гипотез

На практике часто приходится на основе выборочных наблюдений проверять различные предположения относительно генеральной совокупности. Процедура сопоставления выдвинутых гипотез с выборкой и последующее вынесение решения относительно приемлемости этих гипотез называется проверкой гипотез.

§ 9.1. Понятие статистической гипотезы. Основные этапы проверки гипотезы

Под ***статистической гипотезой*** понимают всякое высказывание о генеральной совокупности (случайной величине), проверяемое по выборке (по результатам наблюдений). Примером статистических гипотез являются следующие высказывания: генеральная совокупность, о которой мы располагаем лишь выборочными сведениями, имеет нормальный закон распределения, или генеральное среднее (математическое ожидание случайной величины) равно пяти.

Не располагая сведениями о всей генеральной совокупности, высказанную гипотезу сопоставляют, по определенным правилам, с выборочными сведениями и делают вывод о том, можно принять гипотезу или нет. Процедура сопоставления высказанной гипотезы с выборочными данными называется ***проверкой гипотезы.***

Рассмотрим типичные этапы проверки гипотезы и используемые при этом понятия.

Этап 1. Располагая выборочными данными $x_1, x_2, \ldots, x_n$ и руководствуясь конкретными условиями рассматриваемой задачи, формулируют гипотезу H_0, которую называют ***основной*** или ***нулевой***, и гипотезу H_1, ***конкурирующую*** с гипотезой H_0.

Термин «конкурирующая» означает, что являются противоположными следующие два события:

— по выборке принято решение о справедливости для генеральной совокупности гипотезы H_0;

— по выборке принято решение о справедливости для генеральной совокупности гипотезы H_1.

Гипотезу H_1 называют также ***альтернативной***. Например, если нулевая гипотеза такова: математическое ожидание равно пяти, то альтернативная гипотеза может быть следующей: математическое ожидание меньше пяти, что записывается следующим образом:

$$H_0: MX = 5; \; H_1: MX < 5.$$

Этап 2. Задаются вероятностью α, которую называют ***уровнем значимости***. Поясним ее смысл.

Решение о том, можно ли считать высказывание H_0 справедливым для генеральной совокупности, принимается по выборочным данным, т. е. по ограниченному ряду наблюдений. Следовательно, это решение может быть ошибочным. При этом может иметь место, что:

— отклоняют гипотезу H_0 (принимают альтернативную гипотезу H_1), тогда как на самом деле гипотеза H_0 верна; это ***ошибка первого рода***;

— принимают гипотезу H_0, тогда как на самом деле H_0 неверна (верной является гипотеза H_1); это ***ошибка второго рода***.

Уровень значимости α — это вероятность ошибки первого рода, т. е.

$$\alpha = P(H_1 \mid H_0), \qquad (9.1)$$

где $P(H_1 \mid H_0)$ — вероятность того, что будет принята гипотеза H_1, если на самом деле в генеральной совокупности верна гипотеза H_0. Вероятность α задается заранее малым числом, поскольку это вероятность ошибочного заключения, при этом обычно используют некоторые стандартные значения: 0,05; 0,01; 0,005; 0,001. Например, $\alpha = 0{,}05$ означает следующее: если гипотезу H_0 проверять по каждой

из 100 выборок одинакового объема, то в среднем в пяти случаях из 100 будет совершена ошибка первого рода.

Вероятность ошибки второго рода обозначают β, т. е.

$$\beta = P(H_0 \mid H_1), \qquad (9.2)$$

где $P(H_0 \mid H_1)$ — вероятность того, что будет принята гипотеза H_0, если на самом деле верна гипотеза H_1. В задаче 9.1 будет показано, что, зная α, можно найти вероятность β.

Сказанное иллюстрирует таблица 9.1.

Таблица 9.1

Решение, принимаемое о гипотезе H_0 по выборке / Верна гипотеза H_0 или нет?	Гипотезу H_0 отклоняют (принимают гипотезу H_1)	Гипотезу H_0 принимают
Гипотеза H_0 верна	Ошибка первого рода, ее вероятность $P(H_1 \mid H_0) = \alpha$	Правильное решение, его вероятность $P(H_0 \mid H_0) = 1 - \alpha$
Гипотеза H_0 неверна (верна гипотеза H_1)	Правильное решение, его вероятность $P(H_1 \mid H_1) = 1 - \beta$	Ошибка второго рода, ее вероятность $P(H_0 \mid H_1) = \beta$

В результате проверки гипотезы относительно гипотезы H_0 может быть принято и правильное решение. Существуют два вида правильного решения:

— принимают гипотезу H_0, тогда как и в действительности, в генеральной совокупности, она имеет место; вероятность этого решения $P(H_0 \mid H_0) = 1 - \alpha$;

— не принимают гипотезу H_0 (т. е. принимают гипотезу H_1), тогда как и на самом деле гипотеза H_0 неверна (т. е. верна гипотеза H_1); вероятность этого решения $P(H_1 \mid H_1) = 1 - \beta$.

Этап 3. Находят случайную величину V, являющуюся функцией случайных результатов $X_1, X_2, \ldots, X_n$ наблюдений, $V = V(X_1, X_2, \ldots, X_n)$, такую, что:

— ее значение $v = V(x_1, x_2, \ldots, x_n)$ позволяет судить о «расхождении выборки $x_1, x_2, \ldots, x_n$ с гипотезой H_0»;

— которая, будучи величиной случайной, подчиняется при выполнении гипотезы H_0 некоторому известному закону распределения.

Случайную величину V называют ***критической статистикой.***

Э т а п 4. Так как значения статистики V позволяют судить о «расхождении выборки с гипотезой H_0», то из области допустимых значений критической статистики V следует выделить подобласть ω таких значений, которые свидетельствовали бы о существенном расхождении выборки с гипотезой H_0, следовательно, о невозможности принять гипотезу H_0. Подобласть ω называют ***критической областью***. Допустим, что критическая область выделена. Тогда руководствуются следующим правилом-критерием: *если вычисленное по выборке значение критической статистики V попадает в критическую область, то гипотезу H_0 отклоняют (принимают гипотезу H_1)*. При этом следует понимать, что такое решение может оказаться ошибочным: на самом деле гипотеза H_0 может быть справедливой. Таким образом, ориентируясь на критическую область, можно совершить ошибку первого рода, вероятность которой задана заранее и равна α. Отсюда вытекает следующее требование к критической области ω:

вероятность того, что критическая статистика примет значение из критической области ω, должна быть равна заданному числу α, т. е.

$$P(V \in \omega) = \alpha. \qquad (9.3)$$

Однако критическая область равенством (9.3) определяется неоднозначно. Действительно, представив себе график функции плотности $f_V(x)$ критической статистики V, заметим, что на оси абсцисс существует бесчисленное множество областей-интервалов таких, что площади построенных на них криволинейных трапеций равны α, т. е. областей, удовлетворяющих требованию (9.3). Поэтому кроме требования (9.3) выдвигается следующее требование: критическая область ω должна быть расположена так, чтобы при заданной вероятности α ошибки первого рода вероятность β ошибки второго рода была минимальной. Различают три вида расположения критической области:

правосторонняя критическая область (рис. 9.1, *а*), состоящая из интервала $(x_{\text{пр}}^{\text{кр}}, +\infty)$, где точка $x_{\text{пр}}^{\text{кр}}$ определяется из условия

$$P(V > x_{\text{пр}}^{\text{кр}}) = \alpha \qquad (9.4)$$

и называется ***правосторонней критической точкой,*** отвечающей уровню значимости α;

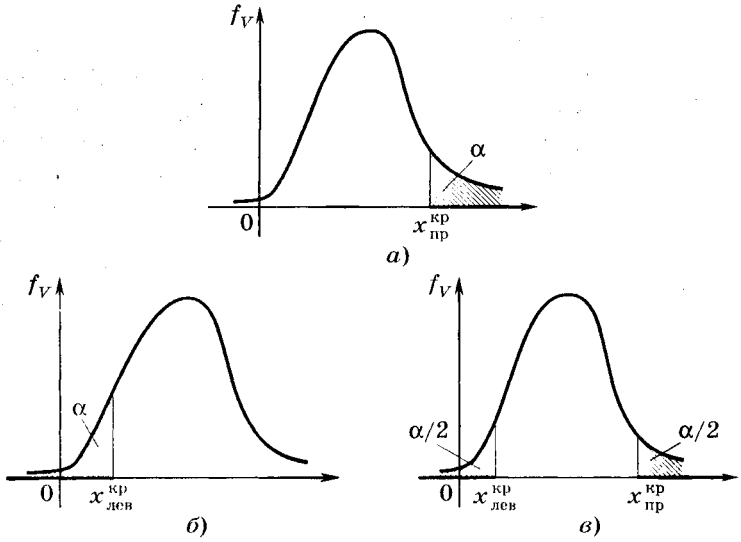


Рис. 9.1

левосторонняя критическая область (рис. 9.1, *б*), состоящая из интервала $(-\infty,\ x^{кр}_{лев})$, где точка $x^{кр}_{лев}$ определяется из условия

$$P(V < x^{кр}_{лев}) = \alpha \qquad (9.5)$$

и называется ***левосторонней критической точкой***, отвечающей уровню значимости α;

двусторонняя критическая область (рис. 9.1, *в*), состоящая из следующих двух интервалов: $(-\infty,\ x^{кр}_{лев})$ и $(x^{кр}_{пр},\ +\infty)$, где точки $x^{кр}_{лев}$ и $x^{кр}_{пр}$ определяются из условий

$$P(V < x^{кр}_{лев}) = \alpha/2 \text{ и } P(V > x^{кр}_{пр}) = \alpha/2 \qquad (9.6)$$

и называются ***двусторонними критическими точками***.

Замечание. По значению критической статистики V судят о «расхождении выборочных данных с гипотезой H_0». Естественно, что гипотеза H_0 должна быть отвергнута, если расхождения велики; именно этим объясняется включение в критическую область больших значений критической статистики V (бóльших критической точки).

Включение в ряде случаев в критическую область малых значений критической статистики V (меньших критической точки) на первый взгляд противоречит смыслу этой величины. Однако не следует забывать, что V — случайная величина, поэтому маловероятно появление не только слишком больших, но и слишком малых ее значений и их следует включить в критическую область.

Э т а п 5. Для конкретных результатов x_1, x_2, ... , x_n наблюдений находят значение статистики, число $v = V(x_1, x_2, \ldots, x_n)$.

Если v попадает в критическую область ω, то гипотезу H_0 забраковывают, так как при ее справедливости вероятность попадания критической статистики V в область ω чрезвычайно мала, а именно $P(V \in \omega) = \alpha$, и попадание v в область ω говорит о том, что произошло то, что вряд ли могло произойти при выполнении гипотезы H_0.

Если v не попадает в критическую область, то гипотеза H_0 не отвергается. Но это вовсе не означает, что H_0 является единственно подходящей гипотезой: просто расхождение между выборочными данными и гипотезой H_0 невелико, или иначе H_0 не противоречит результатам наблюдений; однако таким же свойством наряду с H_0 могут обладать и другие гипотезы.

§ 9.2. Проверка гипотез о числовых значениях параметров нормально распределенной совокупности

Обозначим через X случайную величину, имеющую нормальный закон распределения с параметрами a и σ, т. е. $X = N(a, \sigma)$, причем числовые значения либо одного, либо обоих параметров неизвестны.

Напомним, что $a = MX$, а $\sigma = \sqrt{DX}$.

Дать точный ответ на вопрос, каково числовое значение неизвестного параметра, можно обследовав всю генеральную совокупность, что сделать, как правило, нельзя. В этом случае поступают следующим образом. Проводят наблюдения величины X, при этом предполагают, что они независимы и условия их проведения типичны, т. е. возможные результаты X_1, X_2, ... , X_n наблюдений обладают свойствами (8.45) и (8.46). По конкретным результатам x_1, x_2, ... , x_n наблюдений вычисляют $\bar{x} = \sum_{i=1}^{n} x_i/n$ и $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$. Эти числа дают приближенное представление соответственно об a и σ^2 и помогают сформулировать гипотезы о том, каковы их числовые значения (при формулировке этих гипотез следует также учитывать конкретные условия задачи). Затем приступают к проверке гипотез.

9.2.1. Гипотеза о значении математического ожидания при известном значении дисперсии. Итак, $X = N(a, \sigma)$, причем

числовое значение математического ожидания a неизвестно, а числовое значение дисперсии σ^2 известно.

Выдвинем гипотезу H_0 о том, что неизвестное значение параметра a равно числу a_0. Относительно альтернативной гипотезы H_1 возможны три случая: 1) значение параметра a равно числу a_1, которое больше числа a_0; 2) значение параметра a равно числу a_1, которое меньше a_0; 3) значение параметра a равно числу a_1, которое не равно a_0. Для каждого из этих случаев рассмотрим этапы проверки гипотезы H_0, приведенные в предыдущем параграфе.

Случай 1. Этап 1. Сформулируем нулевую гипотезу

$$H_0: a = a_0 \tag{9.7}$$

и альтернативную

$$H_1: a = a_1 > a_0. \tag{9.8}$$

Этап 2. Зададимся уровнем значимости α.

Этап 3. В качестве критической статистики возьмем величину

$$Z = \frac{\overline{X}_{(n)} - a_0}{\sigma/\sqrt{n}}, \tag{9.9}$$

которая удовлетворяет требованиям, предъявляемым к критической статистике:

— величина (9.9) зависит от результатов наблюдений $X_1, X_2, \ldots, X_n$, поскольку $\overline{X}_{(n)} = (X_1 + X_2 + \ldots + X_n)/n$;

— по значениям величины (9.9) можно судить о «расхождении выборки с гипотезой H_0»; в данном случае это следует понимать так: чем ближе значение среднего $\overline{X}_{(n)}$ к предполагаемому гипотезой H_0 значению a_0 математического ожидания, тем меньше по модулю значение величины (9.9);

— величина (9.9) при выполнении гипотезы (9.7) (а также при соблюдении требований независимости наблюдений величины X и типичности условий этих наблюдений) подчиняется нормальному закону распределения с нулевым математическим ожиданием и единичной дисперсией

$$Z = \frac{\overline{X}_{(n)} - a_0}{\sigma/\sqrt{n}} = N(0, 1). \tag{9.10}$$

≫ Действительно, при соблюдении только что перечисленных требований к наблюдениям нормально распределенной величины $X = N(a, \sigma)$

имеет место соотношение (8.48): $\overline{X}_{(n)} = N(a, \sigma/\sqrt{n})$. Тогда соответствующая величине $\overline{X}_{(n)}$ стандартная величина

$$\frac{\overline{X}_{(n)} - a}{\sigma/\sqrt{n}} = N(0; 1).$$

Подставив в последнее равенство число a_0 — предполагаемое гипотезой H_0 значение параметра a, получим соотношение (9.10). ≪

Этап 4. Выделим критическую область ω — область таких значений статистики Z, при которых гипотеза H_0 отвергается.

В рассматриваемом случае доказано, что наименьшее значение вероятности β ошибки второго рода обеспечивается при правосторонней критической области $(x_{\text{пр}}^{\text{кр}}, +\infty)$, где точка $x_{\text{пр}}^{\text{кр}}$ определяется из условия (9.4), которое, учитывая, что критическая статистика $V = Z$, можно записать в виде

$$P(Z > x_{\text{пр}}^{\text{кр}}) = \alpha, \quad \text{или} \quad P(Z < x_{\text{пр}}^{\text{кр}}) = 1 - \alpha,$$

где Z — стандартная нормально распределенная случайная величина. Но, согласно (5.39),

$$P(Z < x_{\text{пр}}^{\text{кр}}) = 1/2 + \Phi(x_{\text{пр}}^{\text{кр}}),$$

тогда

$$\Phi(x_{\text{пр}}^{\text{кр}}) = P(Z < x_{\text{пр}}^{\text{кр}}) - 1/2 = (1 - \alpha) - 1/2 = 1/2 - \alpha.$$

Итак, $\Phi(x_{\text{пр}}^{\text{кр}}) = 1/2 - \alpha$, следовательно, $x_{\text{пр}}^{\text{кр}}$ можно найти по таблице П. 1 как число $z_{0,5-\alpha}$, при котором $\Phi(z_{0,5-\alpha}) = 0,5 - \alpha$.

Приведем схему нахождения точки $x_{\text{пр}}^{\text{кр}}$:

$$\alpha \to \Phi(z) = 0,5 - \alpha \xrightarrow[\text{П.1}]{} z_{0,5-\alpha} \to x_{\text{пр}}^{\text{кр}} = z_{0,5-\alpha}. \quad (9.11)$$

Таким образом, критическая область определяется неравенством $Z > z_{0,5-\alpha}$; она изображена на рисунке 9.2, *а*.

Этап 5. По конкретным результатам $x_1, x_2, \ldots, x_n$ наблюдений вычислим среднее $\bar{x}$ и, подставив его в формулу (9.10) Z-статистики, найдем ее значение

$$z = \frac{\bar{x} - a_0}{\sigma/\sqrt{n}}.$$

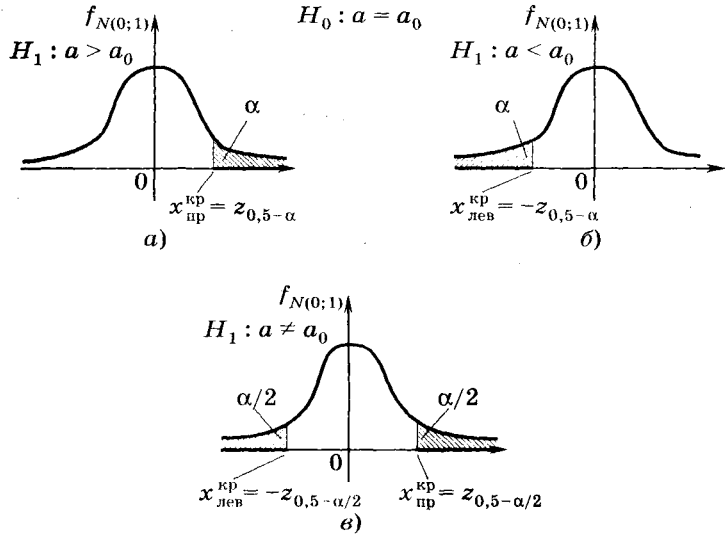


Рис. 9.2

Критерий проверки гипотезы H_0: $a = a_0$ при **альтернативной** гипотезе H_1: $a = a_1 > a_0$ такой:

$$\left.\begin{array}{l} \text{если } \dfrac{\bar{x} - a_0}{\sigma/\sqrt{n}} > z_{0,5-\alpha}, \text{ то } H_0 \text{ отклоняют} \\ (H_1 \text{ принимают}); \\ \text{если } \dfrac{\bar{x} - a_0}{\sigma/\sqrt{n}} < z_{0,5-\alpha}, \textbf{ то } H_0 \textbf{ принимают} \\ (H_1 \text{ отклоняют}). \end{array}\right\} \quad \textbf{(9.12)}$$

Проведя тождественные преобразования неравенств критерия (9.12), получим:

$$\left.\begin{array}{l} \text{если } \bar{x} > a_0 + z_{0,5-\alpha}\sigma/\sqrt{n}, \text{ то } H_0 \text{ отклоняют} \\ (H_1 \text{ принимают}); \\ \text{если } \bar{x} < a_0 + z_{0,5-\alpha}\sigma/\sqrt{n}, \text{ то } H_0 \text{ принимают} \\ (H_1 \text{ отклоняют}); \end{array}\right\} \quad \textbf{(9.13)}$$

При использовании критерия (9.12), или (9.13), вероятность ошибки первого рода (гипотеза H_0 отклоняется, тогда как на самом деле она верна) равна заданному числу α. Рассчитаем вероятность ошибки второго рода (гипотеза H_0 принимается, тогда как на самом деле она неверна, а верна гипотеза H_1).

≫ Из (9.13) следует, что область принятия гипотезы H_0 определяется неравенством

$$\overline{X}_{(n)} < a_0 + z_{0,5-\alpha}\sigma/\sqrt{n},$$

где $\overline{X}_{(n)}$ — среднее, понимаемое как случайная величина, т. е. события «будет принята гипотеза H_0» и «будет иметь место неравенство $\overline{X}_{(n)} < a_0 + z_{0,5-\alpha}\sigma/\sqrt{n}$» равносильны. Учитывая это, получим

$$\beta = P(H_0 \mid H_1) = P((\overline{X}_{(n)} < \underbrace{a_0 + z_{0,5-\alpha}\sigma/\sqrt{n}}_{c}) \mid H_1: a = a_1) =$$

$$= P((\overline{X}_{(n)} < c) \mid X = N(a_1, \sigma)) \underset{(*)}{=} P(N(a_1, \sigma/\sqrt{n}) < c) =$$

$$= 1/2 + \Phi\left(\frac{c - a_1}{\sigma/\sqrt{n}}\right) = 1/2 + \Phi\left(\frac{a_0 + z_{0,5-\alpha}\sigma/\sqrt{n} - a_1}{\sigma/\sqrt{n}}\right) =$$

$$= 1/2 + \Phi\left(z_{0,5-\alpha} - \frac{a_1 - a_0}{\sigma/\sqrt{n}}\right).$$

(При переходе (*) учитывалось следующее: поскольку предполагается верной гипотеза H_1: $a = a_1$, наблюдаемая нормально распределенная случайная величина X имеет математическое ожидание, равное a_1, т. е. $X = N(a_1, \sigma)$, отсюда следует, что среднее $\overline{X}_{(n)} = N(a_1, \sigma/\sqrt{n})$. ≪

Итак,

$$\beta = 1/2 + \Phi\left(z_{0,5-\alpha} - \frac{a_1 - a_0}{\sigma/\sqrt{n}}\right). \qquad (9.14)$$

Из соотношения (9.14) следует, что:

— с ростом вероятности α ошибки первого рода вероятность β ошибки второго рода уменьшается (с ростом α уменьшается разность $0,5 - \alpha$, что ведет к уменьшению рассчитанного по схеме (9.11) числа $z_{0,5-\alpha}$, следовательно, к уменьшению аргумента функции Ф в (9.14) и к уменьшению значения этой функции);

— с ростом числа n наблюдений вероятность β ошибки второго рода уменьшается (с ростом n увеличивается дробь $\frac{a_1 - a_0}{\sigma/\sqrt{n}}$, что при $a_1 > a_0$ ведет к уменьшению аргумента функции Ф в (9.14) и к уменьшению значения этой функции).

Справедливость этих выводов не ограничивается рамками рассмотренного случая. Из (9.14) можно получить, что заданные вероятности α и β ошибок первого и второго рода обеспечиваются при числе наблюдений

$$n = (z_{0,5-\alpha} + z_{0,5-\beta})^2\sigma^2/(a_1 - a_0)^2, \qquad (9.15)$$

где $\Phi(z_{0,5-\alpha}) = 0,5 - \alpha$; $\Phi(z_{0,5-\beta}) = 0,5 - \beta$.

Случай 2. Проверяется гипотеза H_0: $a = a_0$, утверждающая, что неизвестное значение математического ожидания a нормально распределенной величины X, $X = N(a, \sigma)$, равно числу a_0, при альтернативной гипотезе H_1: $a = a_1 < a_0$, утверждающей, что a равно числу a_1, которое меньше a_0.

Не рассматривая всех этапов проверки гипотезы H_0: $a = a_0$ при альтернативе H_1: $a = a_1 < a_0$, приведем формулировку критерия:

$$\left.\begin{array}{l} \text{если } \dfrac{\bar{x} - a_0}{\sigma/\sqrt{n}} < -z_{0,5-\alpha}, \text{ или } \bar{x} < a_0 - z_{0,5-\alpha}\sigma/\sqrt{n}, \\ \text{то } H_0 \text{ отклоняют } (H_1 \text{ принимают}); \\ \text{если } \dfrac{\bar{x} - a_0}{\sigma/\sqrt{n}} > -z_{0,5-\alpha}, \text{ или } \bar{x} > a_0 - z_{0,5-\alpha}\sigma/\sqrt{n}, \\ \text{то } H_0 \text{ принимают } (H_1 \text{ отклоняют}). \end{array}\right\} \quad (9.16)$$

В (9.16) $\bar{x}$ — значение случайной среднего $\overline{X}_{(n)}$, вычисленное по конкретным результатам $x_1, x_2, \ldots, x_n$ наблюдений величины X, число $z_{0,5-\alpha}$ находится по схеме (9.11).

В рассматриваемом случае критическая область — область отклонения гипотезы H_0 определяется неравенством $Z < -z_{0,5-\alpha}$, где $Z = \dfrac{\overline{X}_{(n)} - a_0}{\sigma/\sqrt{n}}$ (она изображена на рисунке 9.2, *б*), или неравенством $\overline{X}_{(n)} < a_0 - z_{0,5-\alpha}\sigma/\sqrt{n}$; область принятия гипотезы H_0 определяется неравенством $Z > -z_{0,5-\alpha}$, или $\overline{X}_{(n)} > a_0 - z_{0,5-\alpha}\sigma/\sqrt{n}$.

Случай 3. Проверяется гипотеза H_0: $a = a_0$ при альтернативе H_1: $a = a_1 \neq a_0$ — математическое ожидание наблюдаемой нормально распределенной случайной величины равно числу a_1, не равному a_0. Здесь также используется Z-статистика [см. (9.9)]; критерий таков:

$$\left.\begin{array}{l} \text{если } \left|\dfrac{\bar{x} - a_0}{\sigma/\sqrt{n}}\right| > z_{0,5-\alpha/2}, \text{ или если } \bar{x} < a_0 - z_{0,5-\alpha/2}\sigma/\sqrt{n} \\ \text{или } \bar{x} > a_0 + z_{0,5-\alpha/2}\sigma/\sqrt{n}, \text{ то } H_0 \text{ отклоняют} \\ (H_1 \text{ принимают}); \\ \text{если } \left|\dfrac{\bar{x} - a_0}{\sigma/\sqrt{n}}\right| < z_{0,5-\alpha/2}, \text{ или если } a_0 - z_{0,5-\alpha/2}\sigma/\sqrt{n} < \\ < \bar{x} < a_0 + z_{0,5-\alpha/2}\sigma/\sqrt{n}, \text{ то } H_0 \text{ принимают} \\ (H_1 \text{ отклоняют}). \end{array}\right\} \quad (9.17)$$

В рассматриваемом случае критическая область — область отклонения гипотезы H_0 — определяется неравенством $|Z| > z_{0,5-\alpha/2}$, где $Z = \dfrac{\overline{X}_{(n)} - a_0}{\sigma/\sqrt{n}}$ (рис. 9.2, *в*).

Проведем параллель между критерием (9.17) и интервальной оценкой (8.52) параметра a при известной дисперсии σ^2. Нетрудно убедиться в том, что тождественные преобразования неравенства

$$a_0 - z_{0,5-\alpha/2}\sigma/\sqrt{n} < \bar{x} < a_0 + z_{0,5-\alpha/2}\sigma/\sqrt{n}$$

(именно, от всех частей неравенства следует отнять a_0, затем $\bar{x}$, а затем умножить все части вновь полученного неравенства на -1) приводят к следующему результату:

$$\bar{x} - z_{0,5-\alpha/2}\sigma/\sqrt{n} < a_0 < \bar{x} + z_{0,5-\alpha/2}\sigma/\sqrt{n}.$$

Сравнив левую и правую части этого неравенства с границами интервала (8.52), заключаем, что интервал

$$(\bar{x} - z_{0,5-\alpha/2}\sigma/\sqrt{n};\ \bar{x} + z_{0,5-\alpha/2}\sigma/\sqrt{n}) \qquad (9.18)$$

представляет собой полученную по конкретным результатам $x_1, x_2, \ldots, x_n$ наблюдений интервальную оценку математического ожидания a, при этом ее надежность $\gamma = 1 - \alpha$.

В результате получаем, что проверить гипотезу H_0: $a = a_0$ при альтернативе H_1: $a \neq a_0$ и уровне значимости, равном α, можно следующим образом: построить интервальную оценку (9.18) математического ожидания a, соответствующую надежности, равной $1 - \alpha$;

$$\left.\begin{array}{l}\text{если } a_0 \notin (\bar{x} - z_{0,5-\alpha/2}\sigma/\sqrt{n};\ \bar{x} + z_{0,5-\alpha/2}\sigma/\sqrt{n}),\\ \text{то } H_0 \text{ отклоняют } (H_1 \text{ принимают});\\ \text{если } a_0 \in (\bar{x} - z_{0,5-\alpha/2}\sigma/\sqrt{n};\ \bar{x} + z_{0,5-\alpha/2}\sigma/\sqrt{n}),\\ \text{то } H_0 \text{ принимают } (H_1 \text{ отклоняют}).\end{array}\right\} \qquad (9.19)$$

Ясно, что критерий (9.19) тождественен критерию (9.17). Используемая при проверке гипотезы H_0: $a = a_0$ критическая статистика Z (9.9), закон ее распределения при выполнении H_0, а также указанные в Z-критериях (9.12), (9.16) и (9.17) области отклонения гипотезы H_0 при трех видах альтернативной гипотезы H_1 приведены в первой строке таблицы 9.2.

▶ **ЗАДАЧА 9.1.** Крупная торговая фирма предполагает открыть в новом районе города филиал. Из опыта работы известно, что фирма будет работать прибыльно, если еже-

Таблица 9.2

Критерии проверки гипотез о числовых значениях параметров нормального распределения $N(a, \sigma)$, вероятности p успеха в испытании Бернулли и коэффициента корреляции

Гипотеза H_0	Предположение	Критическая статистика	Распределение критической статистики при выполнении гипотезы H_0	Гипотеза H_1	Область отклонения гипотезы H_0
1	2	3	4	5	6
$a = a_0$	σ^2 известно	$Z = \dfrac{\overline{X} - a_0}{\sigma/\sqrt{n}}$	$N(0; 1)$	$a = a_1 > a_0$	$Z > z_{0,5-\alpha}$
				$a = a_1 < a_0$	$Z < -z_{0,5-\alpha}$
				$a = a_1 \neq a_0$	$\lvert Z\rvert > z_{0,5-\alpha/2}$
$a = a_0$	σ^2 неизвестно	$T = \dfrac{\overline{X} - a_0}{s/\sqrt{n}}$	$T(n-1)$	$a = a_1 > a_0$	$T > t_{n-1,2\alpha}$
				$a = a_1 < a_0$	$T < -t_{n-1,2\alpha}$
				$a = a_1 \neq a_0$	$\lvert T\rvert > t_{n-1,\alpha}$
$\sigma^2 = b_0$ $(b_0 > 0)$	a неизвестно	$\chi^2 = \dfrac{(n-1)s^2}{b_0}$	$\chi^2(n-1)$	$\sigma^2 = b_1 > b_0$	$\chi^2 > \chi^2_{n-1,\alpha}$
				$\sigma^2 = b_1 < b_0$	$\chi^2 < \chi^2_{n-1,1-\alpha}$
				$\sigma^2 = b_1 \neq b_0$	$\chi^2 < \chi^2_{n-1,1-\alpha/2}$ $\chi^2 > \chi^2_{n-1,\alpha/2}$

Окончание табл.

Гипотеза H_0	Предположение	Критическая статистика	Распределение критической статистики при выполнении гипотезы H_0	Гипотеза H_1	Область отклонения гипотезы H_0
1	2	3	4	5	6
$p = p_0$	$n >> 100$	$Z = \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$, $\hat{p} = m/n$, $q_0 = 1 - p_0$	$N(0; 1)$	$p = p_1 > p_0$	$Z > z_{0,5-\alpha}$
				$p = p_1 < p_0$	$Z < -z_{0,5-\alpha}$
				$p = p_1 \neq p_0$	$\lvert Z\rvert > z_{0,5-\alpha/2}$
$r = 0$		$T = \dfrac{\hat{r}}{\sqrt{(1-\hat{r}^2)/(n-2)}}$	$T(n-2)$	$r > 0$	$T > t_{n-2,2\alpha}$
				$r < 0$	$T < -t_{n-2,2\alpha}$
				$r \neq 0$	$\lvert T\rvert > t_{n-2,\alpha}$

недельный средний доход жителей района больше 400 ден. ед. Также известно, что дисперсия дохода $\sigma^2 = 400$.

а) Определите правило принятия решения, с помощью которого, основываясь на выборке из $n = 100$ жителей и уровне значимости $\alpha = 0{,}05$, может быть установлено, что филиал будет работать прибыльно.

б) Допустим, что средний доход жителей за неделю достигает 406 ден. ед. Рассчитайте вероятность того, что при применении правила принятия решения из п. а) будет совершена ошибка второго рода.

в) Считая альтернативное значение генерального среднего дохода равным 406 ден. ед., рассчитайте объем выборки, при котором риск ошибки первого и второго рода равен 0,05.

Р е ш е н и е. а) Фирма не откроет филиал, если неизвестный генеральный средний доход жителей не превысит 400 ден. ед. Примем за предполагаемое значение генерального среднего a число $a_0 = 400$ и рассмотрим гипотезы H_0: $a = 400$ и H_1: $a > 400$. Если будет принята гипотеза H_1, то это означает прибыльность будущей работы филиала фирмы в новом районе. Поскольку альтернативная гипотеза имеет вид H_1: $a > 400$ (случай 1), для проверки гипотезы H_0: $a = 400$ следует воспользоваться критерием (9.12) или (9.13). Согласно (9.13), гипотезу H_1: $a > 400$ принимают, если недельный среднедушевой доход 100 жителей $\bar{x} > a_0 + z_{0,5-\alpha}\sigma/\sqrt{n}$, или, учитывая, что $z_{0,5-\alpha} = z_{0,45} = 1{,}65$ (см. табл. П.1), если $\bar{x} > 400 + 1{,}65 \cdot 20/\sqrt{100} = 403{,}3$. Итак, если недельный среднедушевой доход 100 жителей нового района $\bar{x} > 403{,}3$, то имеет смысл открыть филиал.

б) По условию 406 ден. ед. — предполагаемое альтернативное (к 400 ден. ед.) значение a_1 генерального среднего a дохода жителей. Таким образом, здесь $a_0 = 400$, $a_1 = 406$ и H_0: $a = a_0$, а H_1: $a = a_1 > a_0$. Прежде чем рассчитывать вероятность ошибки второго рода отметим следующее. В случае, если средний доход 100 жителей равен $\bar{x} = 406$ ден. ед., то была бы принята гипотеза H_1: $a > 400$: ведь $406 > 403{,}3$ (филиал фирмы был бы открыт), а гипотеза H_0: $a = 400$ отклонена. Поступив таким образом, можно допустить ошибку первого рода: на самом деле генеральное среднее $a = 400$. Вероятность этой ошибки $P(H_1 \mid H_0) = \alpha = 0{,}05$.

Рассчитать вероятность ошибки второго рода — ошибки, состоящей в том, что будет принята гипотеза H_0: $a =$

$= 400$ (филиал фирмы не будет открыт), тогда как на самом деле верной является гипотеза H_1: $a > 400$, можно лишь, когда задано альтернативное значение a_1 генерального среднего a. В задаче $a_1 = 406$. Согласно (9.14), вероятность ошибки второго рода

$$P(H_0 \mid H_1) = \beta = 1/2 + \Phi\Big(z_{0,5-\alpha} - \frac{a_1 - a_0}{\sigma/\sqrt{n}}\Big) =$$

$$= 1/2 + \Phi\Big(z_{0,45} - \frac{406 - 400}{20/\sqrt{100}}\Big) = 1/2 + \Phi(1{,}65 - 3) =$$

$$= 0{,}5 - \Phi(1{,}35) = 0{,}5 - 0{,}4115 = 0{,}09.$$

в) По условию H_0: $a = 400$ ($a_0 = 400$), а H_1: $a = 406$ ($a_1 = 406 > a_0$); $\alpha = \beta = 0{,}05$. Согласно формуле (9.15), объем выборки

$$n = (z_{0,5-\alpha} + z_{0,5-\beta})^2\sigma^2/(a_1 - a_0)^2 =$$

$$= (z_{0,5-0,05} + z_{0,5-0,05})^2 400/(406 - 400)^2 =$$

$$= (1{,}65 + 1{,}65)^2 400/36 \approx 121.$$

Как и следовало ожидать, объем выборки по сравнению с прежним объемом, равным 100, увеличился: вероятность ошибки первого рода осталась прежней, $\alpha = 0{,}05$, а вероятность ошибки второго рода по условию $\beta = 0{,}05$, что меньше вероятности $\beta = 0{,}09$ этой же ошибки, рассчитанной при $n = 100$. ◀

9.2.2. Гипотеза о значении математического ожидания при неизвестном значении дисперсии. Выше был подробно рассмотрен вопрос, можно ли считать неизвестное математическое ожидание a нормального распределения равным числу a_0, при этом предполагалось, что дисперсия σ^2 распределения известна. Изучим теперь этот вопрос при условии, что значение дисперсии σ^2 неизвестно.

По результатам $X_1, X_2, \ldots, X_n$ независимых проведенных в типичных условиях наблюдений случайной величины X найдем $\overline{X} = \sum_{i=1}^{n} X_i/n$ и $s^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$.

При неизвестном числовом значении дисперсии σ^2 для проверки гипотезы

$$H_0: a = a_0,$$

где a_0 — заранее заданное число, используют критическую статистику — случайную величину, равную

$$T = \frac{\overline{X} - a_0}{s/\sqrt{n}}, \qquad (9.20)$$

которая при выполнении гипотезы H_0 имеет T-распределение с числом степеней свободы $n-1$ [см. (8.58)]:

$$T = \frac{\overline{X} - a_0}{s/\sqrt{n}} = T(n-1). \qquad (9.21)$$

Пусть α — заданный уровень значимости (заданная вероятность ошибки первого рода). Приведем критерии проверки гипотезы H_0: $a = a_0$ для трех видов альтернативной гипотезы.

1. H_1: $a = a_1 > a_0$ — неизвестное значение генерального среднего предполагается равным числу a_1, большему числа a_0. В этом случае критическая область — область отклонения гипотезы H_0 правосторонняя. Она определяется неравенством $T > t_{n-1,\,2\alpha}$, где $T = \dfrac{\overline{X} - a_0}{s/\sqrt{n}}$; $t_{n-1,\,2\alpha}$ — число, найденное по таблице П. 4, связанной с T-распределением (распределением Стьюдента) при $k = n-1$ и $p = 2\alpha$. Критическая область $T > t_{n-1,\,2\alpha}$ изображена на рисунке 9.3, *а*.

Критерий проверки гипотезы H_0: $a = a_0$ по конкретным результатам $x_1, x_2, \ldots, x_n$ наблюдений величины X таков:

$$\left.\begin{array}{l} \text{если } \dfrac{\bar{x} - a_0}{s/\sqrt{n}} > t_{n-1,\,2\alpha}, \text{ то } H_0 \text{ отклоняют } (H_1 \text{ принимают}); \\ \text{если } \dfrac{\bar{x} - a_0}{s/\sqrt{n}} < t_{n-1,\,2\alpha}, \text{ то } H_0 \text{ принимают } (H_1 \text{ отклоняют}). \end{array}\right\} \qquad (9.22)$$

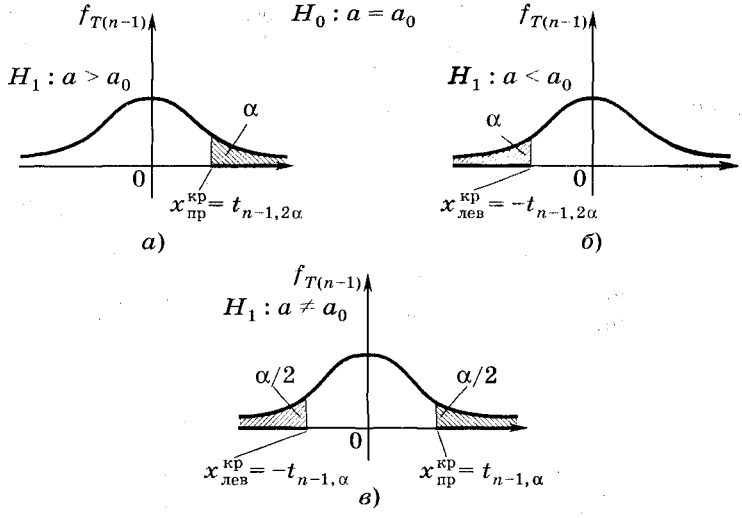


Рис. 9.3

2. H_1: $a = a_1 < a_0$. В этом случае критическая область левосторонняя; она определяется неравенством $T < -t_{n-1,\,2\alpha}$, где $t_{n-1,\,2\alpha}$ — число, найденное по таблице П. 4 при $k = n - 1$ и $p = 2\alpha$. Критическая область: $T < -t_{n-1,\,2\alpha}$ изображена на рисунке 9.3, *б*.

Критерий проверки гипотезы H_0: $a = a_0$ таков:

$$\left.\begin{array}{l} \text{если } \dfrac{\bar{x} - a_0}{s/\sqrt{n}} < -t_{n-1,\,2\alpha}, \text{ то } H_0 \text{ отклоняют} \\ (H_1 \text{ принимают}); \\ \text{если } \dfrac{\bar{x} - a_0}{s/\sqrt{n}} > -t_{n-1,\,2\alpha}, \text{ то } H_0 \text{ принимают} \\ (H_1 \text{ отклоняют}). \end{array}\right\} \qquad (9.23)$$

3. H_1: $a = a_1 \neq a_0$. В этом случае критическая область двухсторонняя; она определяется неравенством $|T| > t_{n-1,\,\alpha}$, или неравенствами: $T < -t_{n-1,\,\alpha}$ и $T > t_{n-1,\,\alpha}$, где $t_{n-1,\,\alpha}$ — число, найденное по таблице П. 4 при $k = n - 1$ и $p = \alpha$. Критическая область $|T| > t_{n-1,\,\alpha}$ изображена на рисунке 9.3, *в*.

Критерий проверки гипотезы H_0: $a = a_0$ будет таким:

$$\left.\begin{array}{l} \text{если } \left|\dfrac{\bar{x} - a_0}{s/\sqrt{n}}\right| > t_{n-1,\,\alpha}, \text{ то } H_0 \text{ отклоняют} \\ (H_1 \text{ принимают}); \\ \text{если } \left|\dfrac{\bar{x} - a_0}{s/\sqrt{n}}\right| < t_{n-1,\,\alpha}, \text{ то } H_0 \text{ принимают} \\ (H_1 \text{ отклоняют}). \end{array}\right\} \qquad (9.24)$$

Нетрудно убедиться в том, что неравенство $\left|\dfrac{\bar{x} - a_0}{s/\sqrt{n}}\right| < t_{n-1,\,\alpha}$ тождественно неравенству $\bar{x} - t_{n-1,\,\alpha}s/\sqrt{n} < a_0 < \bar{x} + t_{n-1,\,\alpha}s/\sqrt{n}$.

Сравнив левую и правую части этого неравенства с границами интервала (8.62), заключаем, что интервал

$$(\bar{x} - t_{n-1,\,\alpha}s/\sqrt{n};\ \bar{x} + t_{n-1,\,\alpha}s/\sqrt{n}) \qquad (9.25)$$

представляет собой полученную по конкретным результатам $x_1, x_2, \ldots, x_n$ наблюдений интервальную оценку математического ожидания a (имеющую место при неизвестном значении дисперсии σ^2), при этом надежность этой интервальной оценки $\gamma = 1 - \alpha$. Учитывая сказанное, проверить гипотезу H_0: $a = a_0$ при альтернативе H_1: $a = a_1 \neq a_0$ можно следующим образом: построить интервальную оценку (9.25) математического ожидания a, соответствующую надежности, равной $1 - \alpha$;

$$\left.\begin{array}{l}\text{если } a_0 \notin (\bar{x} - t_{n-1,\alpha}s/\sqrt{n};\ \bar{x} + t_{n-1,\alpha}s/\sqrt{n}),\\ \text{то } H_0 \text{ отклоняют } (H_1 \text{ принимают});\\ \text{если } a_0 \in (\bar{x} - t_{n-1,\alpha}s/\sqrt{n};\ \bar{x} + t_{n-1,\alpha}s/\sqrt{n}),\\ \text{то } H_0 \text{ принимают } (H_1 \text{ отклоняют}).\end{array}\right\} \quad (9.26)$$

Ясно, что критерий (9.26) тождествен критерию (9.24).

При проверке гипотезы H_0: $a = a_0$ (в случае неизвестного значения дисперсии σ^2) используется T-статистика (9.20). Закон ее распределения при выполнении H_0, а также указанные в T-критериях (9.22)—(9.24) области отклонения гипотезы H_0 при трех видах альтернативной гипотезы H_1 приведены во второй строке таблицы 9.2.

▶ **ЗАДАЧА 9.2.** Хронометраж затрат времени на сборку узла машины $n = 20$ слесарей дал следующие результаты:

Затраты времени, мин	102—104	104—106	106—108	108—110	110—112	
Середина интервала, x_i	103	105	107	109	111	
Количество слесарей, m_i	2	4	6	5	3	$n = 20$

В предположении о нормальности распределения затрат времени на сборку узла решите вопрос о том, можно ли при уровне значимости $\alpha = 0{,}05$ считать 105 мин нормативом трудоемкости, т. е. полагать, что генеральное среднее затрат времени на сборку узла a равно $a_0 = 105$?

Решение. По собранным данным находим $\bar{x} = \Sigma x_i m_i/n = (103 \cdot 2 + 105 \cdot 4 + \ldots + 111 \cdot 3)/20 = 107{,}3$; $\hat{\sigma}^2 = \Sigma x_i^2 m_i/n - (\bar{x})^2 = (103^2 \cdot 2 + 105^2 \cdot 4 + \ldots + 111^2 \cdot 3)/20 - 107{,}3^2 = 5{,}71$; $s^2 = \hat{\sigma}^2 n/(n-1) = 5{,}71 \cdot 20/19 = 6{,}011$; $s = 2{,}45$.

Проверим гипотезу H_0: $a = 105$ при трех видах альтернативной гипотезы, предварительно вычислив значение критической статистики T [см. (9.19)]. Имеем

$$t = \frac{\bar{x} - a_0}{s/\sqrt{n}} = \frac{107{,}3 - 105}{2{,}45/\sqrt{20}} = 4{,}198.$$

1. H_1: $a > 105$. По таблице П. 4 найдем критическую точку $t_{n-1,\,2\alpha} = t_{19;\,0,1} = 1{,}729$. Так как число $4{,}198 > 1{,}73$, то, согласно критерию (9.22), гипотезу H_0: $a = 105$ отклоняем; принимаем гипотезу H_1: $a > 105$, т. е. допущение о равенстве норматива трудоемкости 105 мин не согласуется с результатами наблюдений; норматив трудоемкости больше 105 мин.

2. H_1: $a < 105$. В этом случае критическая точка левосторонняя — это число $-t_{19;\,0,1} = -1{,}729$. И так как значение $t = 4{,}198$ критической статистики больше критической точки: $4{,}198 > -1{,}729$, то, в соответствии с критерием (9.22), будет принята гипотеза H_0: $a = 105$.

Можно было ожидать, что гипотеза H_0: $a = a_0$, где $a_0 = 105$, будет в условиях задачи принята, если альтернативной является гипотеза H_1: $a < a_0$, поскольку выборочное среднее $\bar{x} = 107{,}3 > a_0$ и, как следствие этого, значение критической статистики, равное $\dfrac{\bar{x} - a_0}{s/\sqrt{n}}$, положительно, тогда как критическая точка отрицательна. При выборочном среднем $\bar{x}$, превышающем предполагаемое значение a_0 генерального среднего a ($\bar{x} > a_0$), при любом $\alpha < 0{,}5$, будет отдано предпочтение гипотезе H_0: $a = a_0$, а не гипотезе H_1: $a < a_0$.

Если альтернативной к гипотезе H_0: $a = a_0$ является гипотеза H_1: $a > a_0$ или гипотеза H_1: $a \neq a_0$ и при этом $\bar{x} > a_0$, то однозначно предсказать результат проверки гипотезы H_0 нельзя.

3. H_1: $a \neq 105$. В этом случае критическая область двухсторонняя — это интервал $(-\infty;\ -t_{n-1,\,\alpha} = -t_{19;\,0,05}) = (-\infty;\ -2{,}09)$ и интервал $(t_{n-1,\,\alpha} = t_{19;\,0,05};\ +\infty) = (2{,}09;\ +\infty)$. Значение $t = 4{,}198$ критической статистики попало в критическую область, поэтому гипотезу H_0: $a = 105$ отклоняют в пользу гипотезы H_1: $a \neq 105$; или, иначе, модуль значения критической статистики больше критической точки $t_{n-1,\,\alpha}$: $|4{,}198| > 2{,}093$, поэтому, согласно критерию (9.24), гипотезу H_0 отклоняют в пользу гипотезы H_1. ◄

9.2.3. Гипотеза о значении дисперсии при неизвестном значении математического ожидания. Случайная величина X имеет нормальный закон распределения, т. е. $X = N(a, \sigma)$, но числовые значения математического ожидания a и дисперсии σ^2 неизвестны. Пусть $X_1, X_2, \ldots, X_n$ — возможные результаты независимых, проведенных в типичных усло-

виях наблюдений величины X. Проверим при уровне значимости α гипотезу

$$H_0\colon \sigma^2 = b_0,$$

где b_0 — заранее заданное число (предполагаемое значение дисперсии σ^2, $b_0 > 0$).

Для проверки гипотезы $H_0\colon \sigma^2 = b_0$ используют критическую статистику — случайную величину

$$\chi^2 = (n-1)s^2/b_0, \qquad (9.27)$$

где $s^2 = \sum\limits_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$, а $\overline{X} = \sum\limits_{i=1}^{n} X_i/n$, которая при выполнении гипотезы H_0 имеет χ^2 — распределение с числом степеней свободы $n-1$ (см. (8.65)), т. е.

$$\chi^2 = (n-1)s^2/b_0 = \chi^2(n-1).$$

Критерии проверки гипотезы $H_0\colon \sigma^2 = b_0$ при трех видах альтернативной гипотезы H_1 приведены в третьей строке таблицы 9.2. Кратко поясним их.

1. $H_1\colon \sigma^2 > b_0$. В этом случае критическая область правосторонняя (рис. 9.4, *а*) — это интервал $(\chi^2_{n-1,\alpha}; +\infty)$, где $\chi^2_{n-1,\alpha}$ — число, найдено по таблице П. 2, связанной с χ^2-рас-

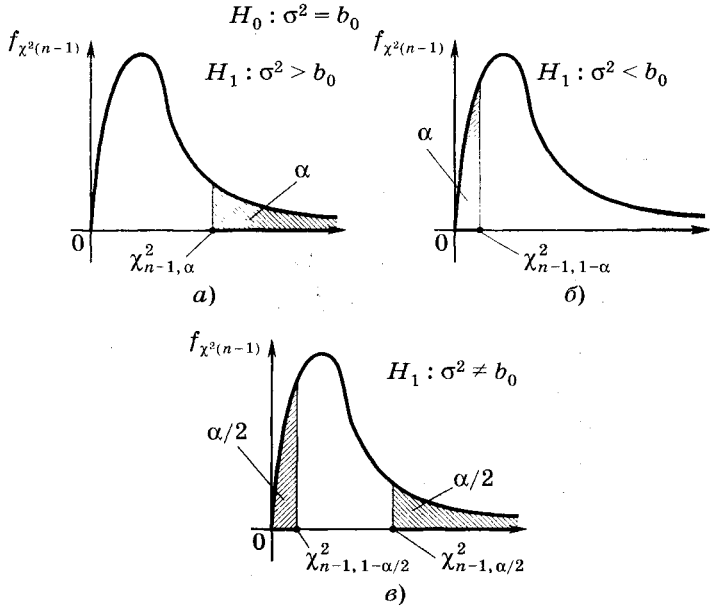


Рис. 9.4

пределением, при $k = n - 1$ и $p = \alpha$. Если вычисленное по конкретным результатам $x_1, x_2, \ldots, x_n$ наблюдений величины X значение $\chi^2_{\text{набл}}$ критической статистики $(n-1)s^2/b_0$ попадает в критическую область, то гипотезу H_0: $\sigma^2 = b_0$ отклоняют (принимают гипотезу H_1: $\sigma^2 > b_0$); в противном случае гипотезу H_0 принимают. Итак, область отклонения гипотезы H_0: $\sigma^2 = b_0$ задается неравенством $(n-1)s^2/b_0 > \chi^2_{n-1,\alpha}$, а область ее принятия — неравенством $(n-1)s^2/b_0 < \chi^2_{n-1,\alpha}$, при этом вероятность

$$P((n-1)s^2/b_0 > \chi^2_{n-1,\alpha}) = P(\chi^2(n-1) > \chi^2_{n-1,\alpha}) = \alpha.$$

2. H_1: $\sigma^2 < b_0$. Критическая область левосторонняя (рис. 9.4, *б*) — это интервал $(0; \chi^2_{n-1,1-\alpha})$, где $\chi^2_{n-1,1-\alpha}$ найдено по таблице П. 2 при $k = n - 1$ и $p = 1 - \alpha$. Область отклонения гипотезы H_0: $\sigma^2 = b_0$ задается неравенством $(n-1)s^2/b_0 < \chi^2_{n-1,1-\alpha}$, а область ее принятия — неравенством $(n-1)s^2/b_0 > \chi^2_{n-1,1-\alpha}$, при этом вероятность

$$P((n-1)s^2/b_0 < \chi^2_{n-1,1-\alpha}) = P(\chi^2(n-1) < \chi^2_{n-1,1-\alpha}) = \alpha.$$

3. H_1: $\sigma^2 \neq b_0$. Критическая область двухсторонняя (рис. 9.4, *в*) — это интервал $(0; \chi^2_{n-1,1-\alpha/2})$ и интервал $(\chi^2_{n-1,\alpha/2}; +\infty)$. Область отклонения гипотезы H_0: $\sigma^2 = b_0$ в пользу гипотезы H_1: $\sigma^2 \neq b_0$ задается неравенствами $(n-1)s^2/b_0 < \chi^2_{n-1,1-\alpha/2}$ и $(n-1)s^2/b_0 > \chi^2_{n-1,\alpha/2}$, а область принятия гипотезы H_0: $\sigma^2 = b_0$ — неравенством $\chi^2_{n-1,1-\alpha/2} < (n-1)s^2/b_0 < \chi^2_{n-1,\alpha/2}$, при этом

$$P((n-1)s^2/b_0 < \chi^2_{n-1,1-\alpha/2}) = P((n-1)\,s^2/b_0 > \chi^2_{n-1,\alpha/2}) = \alpha/2$$

и вероятность того, что случайная величина $(n-1)s^2/b_0$ примет значение, расположенное в критической области, равна α.

ЗАДАЧА 9.3. Точность работы станка-автомата проверяется по дисперсии контролируемого размера изделий, которая не должна превышать 0,15. По пробе из 25 случайно отобранных изделий вычислена оценка дисперсии $s^2 = 0{,}25$. При уровне значимости $\alpha = 0{,}05$ ответить на вопрос, обеспечивает ли станок требуемую точность. Предполагается,

что размер изделия — нормально распределенная случайная величина.

Р е ш е н и е. Станок обеспечивает требуемую точность, если генеральная дисперсия $\sigma^2 \leqslant 0{,}15$. Поскольку выборочная дисперсия $s^2 = 0{,}25 > 0{,}15$, примем за нулевую гипотезу H_0: $\sigma^2 = 0{,}15$, а за альтернативную H_1: $\sigma^2 > 0{,}15$, т. е. имеет место случай 1, при этом $b_0 = 0{,}15$. По таблице П. 2 найдем значение критической точки $\chi^2_{n-1,\alpha}$:

$$\left.\begin{array}{l} n = 25 \to k = n - 1 = 24, \\ \alpha = 0{,}05 \end{array}\right\} \xrightarrow[\text{п.2}]{} \chi^2_{24;0{,}05} = 37{,}65.$$

Таким образом, критическая область $(37{,}65; +\infty)$. Теперь найдем значение критической статистики (9.27): $\chi^2_{\text{набл}} = (25 - 1)0{,}25/0{,}15 = 40$. Так как это число попадает в критическую область, то гипотезу H_0 отвергаем, т. е. считаем, что станок не обеспечивает требуемой точности. ◀

§ 9.3. Проверка гипотезы о числовом значении вероятности события

Пусть A — случайное событие, вероятность p появления которого в единичном испытании неизвестна. Выдвинем гипотезу H_0: $p = p_0$ о том, что вероятность p равна заданному числу p_0; проверим эту гипотезу при уровне значимости α.

В основе проверки гипотезы лежит сравнение p_0 с приближенным значением $\hat{p} = m/n$ — вероятности p, где n — достаточно большое число испытаний Бернулли (независимые испытания, проводимые в типичных условиях); m — число испытаний среди n испытаний, в которых произойдет событие A. Зафиксируем число испытаний n. Напомним, что m и $\hat{p}$ могут иметь двоякую интерпретацию: если речь идет о конкретных результатах n испытаний, то m и $\hat{p}$ — конкретные числа; если же речь идет о возможных результатах испытаний, то и m и $\hat{p}$ — случайные величины. В дальнейшем вариант интерпретации не всегда уточняется: читатель уже имеет достаточный опыт, чтобы понять из контекста, какой вариант используется.

Для проверки гипотезы H_0: $p = p_0$ при числе $n \gg 100$ испытаний Бернулли используют критическую статистику — случайную величину, равную $(\hat{p} - p_0)/\sqrt{p_0(1 - p_0)/n}$, закон распределения которой при выполнении гипотезы H_0: $p = p_0$ и большом n близок к стандартному нормальному (нормаль-

ный закон с нулевым математическим ожиданием и единичной дисперсией; стандартную нормально распределенную величину обозначают символами $N(0; 1)$ или Z), т. е.

$$(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n} \underset{n \text{ велико}}{\approx} N(0; 1) = Z. \quad (9.28)$$

≫ Действительно, при большом числе n испытаний Бернулли имеет место приближенное равенство (6.41), которое, учитывая, что $\hat{p} = m/n$, можно записать в виде

$$\hat{p} \approx N(p, \sqrt{p(1-p)/n}).$$

Тогда соответствующая случайной величине $\hat{p}$ стандартная величина $(\hat{p} - p)/\sqrt{p(1-p)/n} \approx N(0; 1)$. Подставив в последнее приближенное равенство число p_0 — предполагаемое гипотезой H_0 значение вероятности p, получим соотношение (9.28). ≪

Критическая статистика, имеющая стандартное нормальное распределение [см. (9.10)], использовалась при проверке гипотезы H_0: $a = a_0$ (неизвестное значение математического ожидания a равно числу a_0) в том случае, когда значение дисперсии известно. Критические области, соответствующие трем видам альтернативной гипотезы: H_1: $a > a_0$, H_1: $a < a_0$, H_1: $a \neq a_0$, изображены на рисунке 9.2. Такими же будут критические области и при проверке гипотезы H_0: $p = p_0$. Следовательно, критерии ее проверки имеют такой вид:

1. H_1: $p > p_0$. Если

$$(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n} > z_{0,5-\alpha},$$

где $\hat{p}$ — конкретное значение опытной вероятности, рассчитанное после проведения n испытаний, $z_{0,5-\alpha}$ — найденное по таблице П. 1 значение аргумента функции $\Phi(z)$, при котором $\Phi(z) = 0{,}5 - \alpha$, то гипотезу H_0: $p = p_0$ отклоняют в пользу гипотезы H_1: $p > p_0$; в противном случае гипотезу H_0 принимают.

2. H_1: $p < p_0$. Если

$$(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n} < -z_{0,5-\alpha},$$

то гипотезу H_0: $p = p_0$ отклоняют в пользу гипотезы H_1: $p < p_0$, в противном случае принимают гипотезу H_0.

3. H_1: $p \neq p_0$. Если

$$(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n} < -z_{0,5-\alpha/2}$$

или

$$(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n} > z_{0,5-\alpha/2},$$

т. е., иначе, если

$$|\hat{p} - p_0|/\sqrt{p_0(1-p_0)/n} > z_{0,5-\alpha/2},$$

то гипотезу H_0: $p = p_0$ отклоняют в пользу гипотезы H_1: $p \neq p_0$; в противном случае принимают гипотезу H_0.

Эти критерии приведены в четвертой строке таблицы 9.2.

ЗАДАЧА 9.4. Согласно статистике, по городу раскрывают примерно 45 на каждые 100 преступлений. ГУВД одного из районов утверждает, что ее сотрудники за последний год раскрыли 150 преступлений из 300. Случайны ли результаты работы этого ГУВД или они говорят о высоком профессионализме его работников? Принять $\alpha = 0{,}05$.

Р е ш е н и е. Пусть p — вероятность раскрытия преступления районным ГУВД, ее числовое значение неизвестно. Известно лишь, что из $n = 300$ преступлений ГУВД раскрыло $m = 150$, т. е. $\hat{p} = m/n = 0{,}5$.

Случайность результатов работы ГУВД означает, что разность $\varepsilon = \hat{p} - 0{,}45$, где $\hat{p}$ — опытная вероятность раскрытия преступления районным ГУВД, рассматриваемая как случайная величина, является случайной ошибкой — это ошибка, не содержащая систематических составляющих. Математическое ожидание такой ошибки $M\varepsilon = 0$. Но в этом случае и $M(\hat{p} - 0{,}45) = M\hat{p} - 0{,}45 = 0$ или, учитывая, что при любом фиксированном числе n испытаний Бернулли $M\hat{p} = p$, получим $p = 0{,}45$. Таким образом, выяснение вопроса о случайности результатов работы ГУВД сводится к проверке гипотезы H_0: $p = 0{,}45$, альтернативой которой является гипотеза H_1: $p > 0{,}45$ — это случай 1, при этом $p_0 = 0{,}45$. Значение критической статистики $(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n}$ равно

$$z = (0{,}5 - 0{,}45)/\sqrt{0{,}45(1-0{,}45)/300} = 1{,}74.$$

По таблице П. 1 найдем критическую точку $z_{0,5-\alpha}$ — это число, при котором $\Phi(z) = 0{,}5 - \alpha = 0{,}5 - 0{,}05 = 0{,}45$; $z_{0,45} = 1{,}65$. Так как $1{,}74 > 1{,}65$ (значение z критической статистики попадает в критическую область), то принимаем гипотезу H_1, согласно которой вероятность раскрытия преступления районным ГУВД больше, чем вероятность в целом по городу. Это свидетельствует о высоком профессионализме его работников.

Допустим, что вопрос задачи следующий: случайно или нет отличие результатов районного ГУВД от результатов по городу? По-прежнему H_0: $p = 0{,}45$, но альтернативная гипотеза принимает вид H_1: $p \neq 0{,}45$ — это случай 3. Число $z_{0,5-\alpha/2} = z_{0,5-0,025} = z_{0,475} = 1{,}95$. Так как $|z| < z_{0,5-\alpha/2}$ ($|1{,}74| < 1{,}95$), то принимаем гипотезу H_0: $p = 0{,}45$. Считаем, что вероятность раскрытия преступления районным ГУВД такая же, как и в целом по городу.

Кажущаяся противоречивость этого и ранее полученного выводов объясняется различием альтернативных гипотез: здесь H_1: $p \neq 0{,}45$, а ранее H_1: $p > 0{,}45$. ◀

Проверяя гипотезу H_0: $p = p_0$, было принято допущение, что число n испытаний Бернулли достаточно велико, $n \gg 100$. Изложим критерий проверки гипотезы H_0: $p = p_0$ при альтернативе H_1: $p \neq p_0$, который обычно используют при n, близком к 100 (критерий можно использовать и при $n \gg 100$, в этом случае его результаты практически не отличаются от результатов рассмотренного выше критерия, хотя являются более точными).

Критерий проверки гипотезы H_0: $p = p_0$ при H_1: $p \neq p_0$ такой: найти границы p_1 и p_2 интервальной оценки вероятности p, соответствующей надежности $\gamma = 1 - \alpha$, используя схему (8.85) и формулы (8.84) (эти формулы применяют для нахождения p_1 и p_2 при n, близком к 100);

$$\left.\begin{array}{l}\text{если } p_0 \notin (p_1; p_2), \text{ то } H_0\text{: } p = p_0 \text{ отклоняют}\\ (H_1\text{: } p \neq p_0 \text{ принимают});\\ \text{если } p_0 \in (p_1; p_2), \text{ то } H_0\text{: } p = p_0 \text{ принимают}\\ (H_1\text{: } p \neq p_0 \text{ отклоняют}).\end{array}\right\} \qquad (9.29)$$

Критерий (9.29) можно использовать для проверки гипотезы H_0: $p = p_0$ при альтернативе H_1: $p \neq p_0$ и когда число n испытаний Бернулли мало (n существенно меньше 100), но в этом случае границы p_1 и p_2 находят соответственно как решения уравнений (8.91) и (8.92), или по специальным таблицам при заданных n, $n - m$ и $\gamma = 1 - \alpha$ (фрагмент их дан в таблице П. 6).

Критерии, подобные (9.29), использующие идею интервальной оценки, рассматривались и раньше: при проверке гипотезы H_0: $a = a_0$, если альтернатива H_1: $a \neq a_0$ [см. (9.19) и (9.26)].

Заметим, что в таблице 9.2 также приведен алгоритм проверки гипотезы H_0: $r_{X,Y} = 0$ о том, что неизвестное значение коэффициента корреляции $r_{X,Y}$ между случайными

величинами X и Y равно нулю. Примеры задач, в которых возникает необходимость проверки такой гипотезы, рассматриваются в гл. 11.

§ 9.4. Проверка гипотез о равенстве неизвестных значений соответствующих параметров двух нормально распределенных совокупностей

Допустим, что требуется сравнить средний возраст и среднюю вариацию (дисперсию) возраста гражданина крупного города на момент нарушения им впервые уголовного законодательства с соответствующими характеристиками после проведения в городе определенных профилактических мероприятий. Переведем задачу на язык математической статистики. Введем следующие обозначения:

X, MX, DX — соответственно возраст случайно выбранного нарушителя, генеральные средний возраст и дисперсия возраста до проведения профилактических мероприятий;

Y, MY, DY — аналогичные характеристики после проведения профилактических мероприятий.

Не имея возможности собрать сведения о возрасте всех нарушителей на момент совершения ими преступления (следовательно, не зная значений генеральных средних и дисперсий), а располагая лишь такими выборочными обследованиями: собраны сведения о возрасте n_X нарушителей до проведения профилактических мероприятий и о возрасте n_Y нарушителей после их проведения, требуется проверить гипотезы H_0: $MX = MY$ и H_0': $DX = DY$ о том, что профилактические мероприятия не изменили ни среднего возраста нарушителей, ни дисперсии возраста, или, иначе, различия, которые могут быть между выборочными средними $\bar{x}$ и $\bar{y}$ и выборочными дисперсиями $\hat{\sigma}_X^2$ и $\hat{\sigma}_Y^2$, объясняются тем, что собраны сведения о n_X и n_Y, а не о всех нарушителях. Если же, например, в результате проверки гипотезы H_0: $MX = MY$ при альтернативе H_1: $MX < MY$ гипотеза H_0 будет отклонена, то это означает, что после проведения профилактических мероприятий средний возраст нарушителя увеличился, т. е. мероприятия способствовали предотвращению нарушений молодыми людьми.

Обратим внимание на то, что сравнивать математические ожидания, так же как и дисперсии, имеет смысл только в том случае, когда качественное содержание величин X и Y одинаково (X и Y — возраст нарушителя до и после

мероприятия, X и Y — объем продаж до и после рекламы и т. д.)

В таблице 9.3 (строки 1—4) приведены критерии проверки гипотез о равенстве математических ожиданий и о равенстве дисперсий в предположении, что X и Y имеют нормальный закон распределения, $X = N(a_X, \sigma_X)$ и $Y = N(a_Y, \sigma_Y)$, имеется n_X независимых, проводимых в типичных условиях наблюдений величины X и n_Y независимых, проводимых в типичных условиях наблюдений величины Y и, кроме того, наблюдения организованы таким образом, что n_X наблюдений величины X и n_Y наблюдений величины Y независимы между собой.

Отметим следующее:

— Если гипотезу H_0: $a_X = a_Y$ принимают, то говорят, что *различие выборочных средних* $\bar{x}$ *и* $\bar{y}$, которое может иметь место, *случайно, статистически незначимо* или *несущественно*. В этом случае оценка математического ожидания a ($a = a_X = a_Y$), рассчитанная по $n_X + n_Y$ наблюдениям величин X и Y, такова: $(\bar{x}\, n_X + \bar{y}\, n_Y)/(n_X + n_Y)$.

— Если принимают гипотезу H_0: $\sigma_X^2 = \sigma_Y^2$, то говорят, что *различие оценок* s_X^2 *и* s_Y^2 *дисперсий* σ_X^2 и σ_Y^2, которое может иметь место, *случайно, статистически незначимо* или *несущественно*. В этом случае оценка дисперсии σ^2 ($\sigma^2 = \sigma_X^2 = \sigma_Y^2$), рассчитанная по $n_X + n_Y$ наблюдениям величин X и Y, такова:

$$s^2 = \frac{s_X^2(n_X - 1) + s_Y^2(n_Y - 1)}{(n_X - 1) + (n_Y - 1)}.$$

9.4.1. Гипотеза о равенстве математических ожиданий при известных значениях дисперсий. Заметим, что используемая здесь критическая статистика Z (см. таблицу 9.3, строка 1) при выполнении гипотезы H_0: $a_X = a_Y$ является стандартной нормально распределенной величиной, функция плотности которой с указанием критической области изображена на рисунке 9.2 (при H_1: $a_X > a_Y$ — область правосторонняя, при H_1: $a_X < a_Y$ — левосторонняя, при H_1: $a_X \neq a_Y$ — область двусторонняя). Критическая точка находится по таблице П.1.

ЗАДАЧА 9.5. По выборке объема $n_X = 14$ найден средний размер $\bar{x} = 182$ мм диаметра валиков, изготовленных автоматом *1*; по выборке объемом $n_Y = 9$ найден средний размер $\bar{y} = 185$ мм диаметра валиков, изготовленных

Таблица 9.3

Сравнение соответствующих параметров двух нормальных распределений $X = N(a_X, \sigma_X)$ и $Y = N(a, \sigma_Y)$; сравнение двух вероятностей

Гипотеза H_0	Предположение	Критическая статистика	Распределение критической статистики при выполнении гипотезы H_0	Гипотеза H_1	Область отклонения гипотезы H_0
1	2	3	4	5	6
$a_X = a_Y$	σ_X^2, σ_Y^2 известны	$Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}}$	$N(0; 1)$	$a_X > a_Y$ $a_X < a_Y$ $a_X \neq a_Y$	$Z > z_{0,5-\alpha}$ $Z < -z_{0,5-\alpha}$ $\lvert Z\rvert > z_{0,5-\alpha/2}$
$\sigma_X^2 = \sigma_Y^2$	a_X, a_Y неизвестны	$F = s_X^2/s_Y^2$, $s_X^2 > s_Y^2$	$F(n_X - 1, n_Y - 1)$	$\sigma_X^2 > \sigma_Y^2$ $\sigma_X^2 \neq \sigma_Y^2$	$F > f_{n_X-1, n_Y-1, \alpha}$ $F > f_{n_X-1, n_Y-1, \alpha/2}$
$a_X = a_Y$	σ_X^2, σ_Y^2 неизвестны, но $\sigma_X^2 = \sigma_Y^2$	$T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{(1/n_X + 1/n_Y)s^2}}$, где $s^2 = \dfrac{s_X^2(n_X - 1) + s_Y^2(n_Y - 1)}{n_X + n_Y - 2}$	$T(n_X + n_Y - 2)$	$a_X > a_Y$ $a_X < a_Y$ $a_X \neq a_Y$	$T > t_{n_X+n_Y-2, 2\alpha}$ $T < -t_{n_X+n_Y-2, 2\alpha}$ $\lvert T\rvert > t_{n_X+n_Y-2, \alpha}$

Окончание табл.

Гипотеза H_0	Предположение	Критическая статистика	Распределение критической статистики при выполнении гипотезы H_0	Гипотеза H_1	Область отклонения гипотезы H_0
1	2	3	4	5	6
$a_X = a_Y$	σ_X^2, σ_Y^2 неизвестны, но $\sigma_X^2 \neq \sigma_Y^2$	$T = \dfrac{(\overline{X} - \overline{Y})}{\sqrt{s_X^2/n_X + s_Y^2/n_Y}}$	$T(k_1)$, где k_1 — целое, ближайшее к числу $\dfrac{(s_X^2/n_X + s_Y^2/n_Y)^2}{\dfrac{(s_X^2/n_X)^2}{n_X - 1} + \dfrac{(s_Y^2/n_Y)^2}{n_Y - 1}}$	$a_X > a_Y$ $a_X < a_Y$ $a_X \neq a_Y$	$T > t_{k_1, 2\alpha}$ $T < -t_{k_1, 2\alpha}$ $\lvert T \rvert > t_{k_1, \alpha}$
$p_1 = p_2$	$n_1 \gg 100$ $n_2 \gg 100$	$Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1 - \hat{p})}}$, где $\hat{p}_1 = m_1/n_1$, $\hat{p}_2 = m_2/n_2$, $\hat{p} = (m_1 + m_2)/(n_1 + n_2)$	$N(0; 1)$	$p_1 > p_2$ $p_1 < p_2$ $p_1 \neq p_2$	$Z > z_{0,5-\alpha}$ $Z < -z_{0,5-\alpha}$ $\lvert Z \rvert > z_{0,5-\alpha/2}$

автоматом *2*. Предварительным анализом установлено, что размер диаметра валиков, изготовленных каждым автоматом, имеет нормальный закон распределения с дисперсией $\sigma_X^2 = 5$ мм^2 для автомата *1* и $\sigma_Y^2 = 7$ мм^2 — для автомата *2*. Можно ли при уровне значимости $\alpha = 0{,}05$ объяснить различие выборочных средних случайной ошибкой?

Р е ш е н и е. Обозначим через a_X и a_Y математическое ожидание размера диаметра валиков, изготовленных соответственно на автоматах *1* и *2*.

Ошибка ε называется *случайной*, если математическое ожидание $M\varepsilon = 0$. Поэтому выражение «различие выборочных средних вызвано случайной ошибкой» означает, что $\overline{X} - \overline{Y} = \varepsilon$. Применяя к обеим частям этого равенства операцию математического ожидания, получаем $M\overline{X} - M\overline{Y} = 0$ или, учитывая, что $M\overline{X} = a_X$, а $M\overline{Y} = a_Y$, имеем $a_X = a_Y$. Таким образом, в задаче требуется решить вопрос о том, можно принять гипотезу H_0: $a_X = a_Y$ или нет, т. е. в качестве конкурирующей гипотезы предлагается рассмотреть гипотезу H_1: $a_X \neq a_Y$ (см. таблицу 9.3, строка 1).

Значение z критической статистики Z равно $z = (182 - 185)/\sqrt{5/14 + 7/9} = -2{,}82$; критическая область двусторонняя (см. рис. 9.2, *в*), критическая точка $z_{0,5-\alpha/2} = z_{0,475} = 1{,}95$. Поскольку $|-2{,}82| > 1{,}95$, гипотезу H_0: $a_X = a_Y$ отклоняем — различие выборочных средних неслучайно. ◀

9.4.2. Гипотеза о равенстве дисперсий при неизвестных значениях математических ожиданий. При проверке гипотезы о равенстве математических ожиданий двух нормально распределенных совокупностей предполагалось, что дисперсии этих совокупностей известны — редкий случай в практических задачах. Обычно числовые значения генеральных дисперсий неизвестны. Для того чтобы проверить гипотезу о равенстве математических ожиданий в случае, когда генеральные дисперсии неизвестны, надо знать, равны эти генеральные дисперсии или неравны. Но так как значения этих дисперсий неизвестны, то выяснение того, равны они или нет, может сводиться лишь к проверке гипотезы о равенстве дисперсий.

Отметим, что задача проверки гипотезы о равенстве дисперсий имеет и самостоятельное значение. Дисперсия характеризует точность работы приборов, технологических процессов, потенциальные возможности фирмы и т. д. Убедившись в равенстве двух дисперсий, мы тем самым убеждаемся, например, в том, что два прибора, два технологических процесса обеспечивают одинаковую точность.

Используемая здесь критическая статистика F (см. табл. 9.3, строка 2) при выполнении гипотезы H_0: $\sigma_X^2 = \sigma_Y^2$ имеет распределение Фишера, график функции плотности которого с указанием критических областей изображен на рисунке 9.5. При этом, учитывая особенности алгоритма проверки гипотезы H_0 (предполагается, что $s_X^2 > s_Y^2$), разумными альтернативами к H_0 могут быть либо H_1: $\sigma_X^2 > \sigma_Y^2$, либо H_0: $\sigma_X^2 \neq \sigma_Y^2$. В обоих случаях критическая область правосторонняя (см. рис. 9.5), но значения критических точек разные; их находят по таблице П.5.

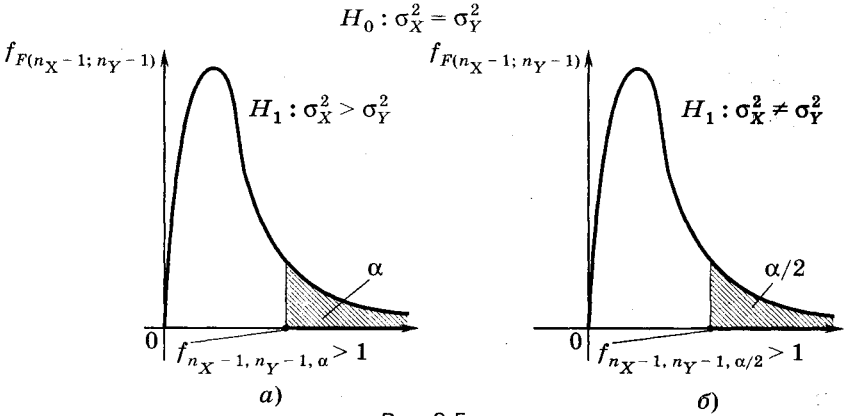


Рис. 9.5

ЗАДАЧА 9.6. По двум фирмам A и B, производящим однотипный продукт, собраны сведения о ежемесячном объеме производства: по фирме A за $n_A = 9$ месяцев, по фирме B — за $n_B = 7$ месяцев. Рассчитанные по этим данным исправленные выборочные дисперсии таковы: $s_A^2 = 5{,}80$ и $s_B^2 = 22{,}81$. 1) Можно ли при уровне значимости $\alpha = 0{,}05$ считать, что потенциальные возможности фирмы B больше, чем фирмы A? 2) Существенно ли при $\alpha = 0{,}1$ различие зафиксированных значений потенциальных возможностей фирм или не существенно? Предполагается, что ежемесячный объем производства каждой фирмы — нормально распределенная случайная величина.

Р е ш е н и е. Напомним, что дисперсия — это характеристика среднего разброса значений случайной величины вокруг среднего этих значений. Фирмы производят однотипный продукт и если у одной из них разброс ежемесячного объема производства, или дисперсия, больше, чем у другой, то и потенциальные возможности этой фирмы

больше: она за месяц может выпустить и «мало», и «много» продукции.

Проверим гипотезу H_0: $\sigma_A^2 = \sigma_B^2$ о равенстве дисперсий потенциальных возможностей фирм, используя изложенные выше два подхода. При этом, без ограничения общности, договоримся за X принимать величину с бóльшим значением исправленной выборочной дисперсии, а за Y — величину с меньшим значением исправленной выборочной дисперсии. По условию $s_B^2 > s_A^2$, поэтому X и Y — ежемесячные объемы производства соответственно фирм B и A; $s_X^2 = s_B^2 = 22{,}81$; $s_Y^2 = s_A^2 = 5{,}80$; $n_X = n_B = 7$; $n_Y = n_A = 9$; $\sigma_X^2 = \sigma_B^2$; $\sigma_Y^2 = \sigma_A^2$.

1) Судя по первому вопросу задачи, альтернативой к гипотезе H_0: $\sigma_X^2 = \sigma_Y^2$ является гипотеза H_1: $\sigma_X^2 > \sigma_Y^2$.

Значение критической статистики F (см. табл. 9.3, строка 2) равно $f = s_X^2 / s_Y^2 = 22{,}81/5{,}80 = 3{,}93$; критическая точка

$$x_{\text{пр},\alpha}^{\text{кр}} = f_{k_1 = n_X - 1,\, k_2 = n_Y - 1,\, \alpha} = f_{6;\,8;\,0{,}05} \underset{\text{П. 5}}{=} 3{,}58.$$

Так как 3,93 попадает в критическую область ($3{,}93 > 3{,}58$), то гипотезу H_0: $\sigma_X^2 = \sigma_Y^2$ отклоняем и принимаем гипотезу H_1: $\sigma_X^2 > \sigma_Y^2$ — имеется основание считать потенциальные возможности фирмы B бóльшими по сравнению с возможностями фирмы A.

2) Судя по второму вопросу задачи, альтернативой гипотезе H_0: $\sigma_X^2 = \sigma_Y^2$ является гипотеза H_1: $\sigma_X^2 \neq \sigma_Y^2$; $\alpha = 0{,}1$ (рис. 9.5, *б*).

В этом случае значение критической статистики F попрежнему равно $s_X^2 / s_Y^2 = 3{,}93$, что больше правосторонней критической точки $f_{k_1 = n_X - 1,\, k_2 = n_Y - 1,\, \alpha/2} = f_{6;\,8;\,0{,}05} = 3{,}58$, найденной по таблице П.5. Гипотезу H_0: $\sigma_X^2 = \sigma_Y^2$ отклоняем и принимаем гипотезу H_1: $\sigma_X^2 \neq \sigma_Y^2$. ◄

9.4.3. Гипотеза о равенстве математических ожиданий при неизвестных, но равных значениях дисперсий и неравных значениях дисперсий. Заметим, что используемая здесь критическая статистика T (см. табл. 9.3, строки 3, 4) имеет при выполнении гипотезы H_0: $a_X = a_Y$ распределение Стьюдента с числом степеней свободы $n_X + n_Y - 2$ (для случая неизвестных, но равных дисперсий, $\sigma_X^2 = \sigma_Y^2$) и числом степеней свободы k_1 (для случая неизвестных, но не равных дисперсий, $\sigma_X^2 \neq \sigma_Y^2$).

С критической статистикой, имеющей T-распределение, и соответствующими критическими областями мы уже сталкивались (см. рисунок 9.3). При проверке гипотезы H_0: $a_X = a_Y$ критические области расположены так же, как на рисунке 9.3: при H_1: $a_X > a_Y$ область правосторонняя, при H_1: $a_X < a_Y$ — левосторонняя, при H_1: $a_X \neq a_Y$ — двухсторонняя, но T-величина здесь имеет либо $n_X + n_Y - 2$ степеней свободы, либо k_1, а не $n - 1$, как имеет место на рисунке 9.3. Критические точки находят по таблице П.4.

ЗАДАЧА 9.6 (продолжение). Дополним условия задачи 9.6. Допустим, что среднемесячный объем производства фирмы A, вычисленный за $n_A = 9$ месяцев, равен 151,2 у. е., а среднемесячный объем производства фирмы B, вычисленный за $n_B = 7$ месяцев, равен 165 у. е. Исправленные выборочные дисперсии объема производства, рассчитанные за эти же месяцы, остались прежними: $s_A^2 = 5{,}80$, а $s_B^2 = 22{,}81$. Можно ли при уровне значимости $\alpha = 0{,}1$ считать различие выборочных средних случайным или нет?

Р е ш е н и е. В задаче 9.6 ежемесячный объем производства фирмы B был обозначен через X, а фирмы A — через Y; напомним, что эти объемы производства — нормально распределенные величины: $X = N(a_X, \sigma_X)$, $Y = N(a_Y, \sigma_Y)$.

В рассматриваемой задаче надо ответить на вопрос: можно ли при уровне значимости $\alpha = 0{,}1$ принять гипотезу H_0: $a_X = a_Y$ о равенстве генеральных среднемесячных объемов производства фирм, если альтернативной является гипотеза H_1: $a_X \neq a_Y$.

Значения генеральных дисперсий σ_X^2 и σ_Y^2 неизвестны, но в задаче 9.6 при $\alpha = 0{,}1$ гипотеза H_0: $\sigma_X^2 = \sigma_Y^2$ была отклонена в пользу гипотезы H_1: $\sigma_X^2 \neq \sigma_Y^2$, поэтому для проверки гипотезы H_0: $a_X = a_Y$ используем критическую статистику T, приведенную в таблице 9.3 (строка 4); ее значение

$$t = (\bar{x} - \bar{y}) / \sqrt{s_X^2/n_X + s_Y^2/n_Y} =$$

$$= (165 - 151{,}2) / \sqrt{22{,}81/7 + 5{,}80/9} = 6{,}99.$$

Число степеней свободы k_1 статистики T равно целому числу, ближайшему к

$$(s_X^2/n_X + s_Y^2/n_Y)^2 / [(s_X^2/n_X)^2/(n_X - 1) + (s_Y^2/n_Y)^2/(n_Y - 1)] =$$

$$= (22{,}81/7 + 5{,}80/9)^2 / [(22{,}81/7)^2/6 + (5{,}80/9)^2/8] = 8{,}36,$$

т. е. $k_1 = 8$. При альтернативе H_1: $a_X \neq a_Y$ гипотезу H_0: $a_X = a_Y$ принимают, если $|t| < t_{k_1, p=\alpha}$. Используя таблицу П.4, найдем $t_{8,\, p=0,1} = 1{,}86$. Так как $|6{,}99| > 1{,}86$, то гипотезу H_0: $a_X = a_Y$ отклоняем, принимаем гипотезу H_1: $a_X \neq a_Y$. Это означает, что при $\alpha = 0{,}1$ различие выборочных средних $\bar{x} = 165$ и $\bar{y} = 151{,}2$ не может быть объяснено случайными ошибками выборок. ◀

9.4.4. Excel-программы, реализующие проверку гипотез о равенстве параметров двух нормально распределенных совокупностей. Имеются две нормально распределенные совокупности — две случайные величины $X = N(a_X, \sigma_X)$ и $Y = N(a_Y, \sigma_Y)$ и два ряда результатов наблюдений этих величин $x_1, x_2, \ldots, x_{n_X}$ и $y_1, y_2, \ldots, y_{n_Y}$ (предполагается, что наблюдения каждой величины независимы и проведены в типичных условиях и, более того, наблюдения величины X и наблюдения величины Y независимы между собой).

Критерии проверки гипотез о равенстве соответствующих параметров двух нормальных совокупностей, представленные в первых четырех строках таблицы 9.3, в Microsoft Excel реализованы в следующих программах.

Программа **«Двухвыборочный *F*-тест для дисперсий»** используется для проверки гипотезы H_0: $\sigma_X^2 = \sigma_Y^2$ (генеральные дисперсии одинаковы). *Исходные данные* — введенные в рабочий лист наблюдения переменной 1 (величины X) и наблюдения переменной 2 (величины Y), введенный в диалоговое окно уровень значимости. По этим данным программа рассчитывает: $\bar{x}$ и $\bar{y}$, дисперсии s_X^2 и s_Y^2 и ряд других величин, необходимых для проверки гипотезы.

Замечания. 1. Для обеспечения соответствия результатов проверки гипотезы H_0: $\sigma_X^2 = \sigma_Y^2$ с помощью этой программы и с помощью критериев, приведенных во второй строке таблицы 9.3, будем считать, что выполняется соотношение $s_X^2 > s_Y^2$. Если $s_X^2 < s_Y^2$, то за наблюдения переменной 1 следует принять числа $y_1, y_2, \ldots, y_{n_Y}$, а за наблюдения переменной 2 числа $x_1, x_2, \ldots, x_{n_X}$.

2. Поскольку $s_X^2 > s_Y^2$, разумными альтернативами гипотезы H_0: $\sigma_X^2 = \sigma_Y^2$ являются:

1) гипотеза H_1: $\sigma_X^2 > \sigma_Y^2$; в этом случае в *«уровень значимости»* диалогового окна вводится α — уровень значимости, при котором проверяется гипотеза H_0: $\sigma_X^2 = \sigma_Y^2$;

2) гипотеза H_1: $\sigma_X^2 \neq \sigma_Y^2$; в этом случае в *«уровень значимости»* диалогового окна вводится вероятность $\alpha_д = \alpha/2$ (см. вторую строку таблицы 9.3).

Рассмотрим результаты работы программы на примере следующей задачи.

▶ **ЗАДАЧА 9.7.** Расход сырья на единицу продукции по старой технологии составил:

Расход сырья, x_i	304	307	308
Число изделий, m_i	1	4	4

$n_X = 9$

По новой технологии:

Расход сырья, y_i	303	304	306	308
Число изделий, n_i	2	6	4	1

$n_Y = 13$

Выяснить можно ли на уровне значимости $\alpha = 0{,}1$ принять гипотезу H_0: $\sigma_X^2 = \sigma_Y^2$ при альтернативе H_1: $\sigma_X^2 \neq \sigma_Y^2$.

Р е ш е н и е. В рабочее поле введем $n_X = 9$ наблюдений величины X (столбец A) и $n_Y = 13$ наблюдений величины Y (столбец B) (рис. 9.6, *а*) и, обратившись к **Статистической функции** ДИСП, вычислим исправленные оценки дисперсий: $s_A^2 = 1{,}61(1)$ и $s_B^2 = 2{,}192$. Так как $s_B^2 > s_A^2$, то при обращении к программе **«Двухвыборочный *F*-тест для дисперсий»** второй столбец чисел (столбец B) примем за наблюдения переменной 1 — величины X, а первый (столбец A) за наблюдения переменной 2 — величины Y. В диалоговое окно введем уровень значимости $\alpha_д = \alpha/2$, поскольку альтернативная гипотеза H_1: $\sigma_X^2 \neq \sigma_Y^2$. Результаты работы программы представлены на рисунке 9.6, *б*; это:

— средний расход сырья на единицу продукции по новой технологии $\bar{x} = 304{,}77$ и по старой $\bar{y} = 307{,}1(1)$;

— дисперсии $s_X^2 = 2{,}192$ и $s_Y^2 = 1{,}61(1)$;

— числа наблюдений $n_X = 13$ и $n_Y = 9$;

— числа степеней свободы k_1 и k_2 (*df* — от англ. *degree of freedom* — степень свободы) критической статистики — случайной величины $F(k_1 = n_X - 1,\ k_2 = n_Y - 1)$ (см. табл. 9.3, строка 2); $k_1 = 13 - 1 = 12$, $k_2 = 9 - 1 = 8$;

A (П2, Y)	B(П1, x)
304	303
307	303
307	304
307	304
307	304
308	304
308	304
308	304
308	306
	306
	306
	306
	308

а)

Двухвыборочный *F*-тест для дисперсий

	Переменная 1	Переменная 2
Среднее	304,7692	307,1111
Дисперсия	2,192308	1,611111
Наблюдения	13	9
df	12	8
F	1,360743	
P(*F*<=*f*) одностороннее	0,338654	
F критическое одностороннее	3,283944	

б)

Двухвыборочный *t*-тест с одинаковыми дисперсиями

	Переменная 1	Переменная 2
Среднее	304,7692	307,1111
Дисперсия	2,192308	1,611111
Наблюдения	13	9
Объединенная дисперсия	1,959829	
Гипотетическая разность средних	0	
df	20	
t-статистика	–3,85778	
P(*T*<=*t*) одностороннее	0,00049	
t критическое одностороннее	1,325341	
P(*T*<=*t*) двухстороннее	0,000981	
t критическое двухстороннее	1,724718	

в)

Рис. 9.6

— числовое значение названной критической **статистики** F, равное $f = s_X^2 / s_Y^2 = 2{,}192/1{,}61(1) = 1{,}361$;

— «P одностороннее» — это вероятность того, что критическая статистика — случайная величина $F(k_1, k_2)$ превзойдет число f, т. е. $P(F(k_1, k_2) > f)$, которая находится с помощью **Статистической функции** FРАСП(f; k_1; k_2). В условиях задачи $P(F(12; 8) > 1{,}361) =$ FРАСП(1,361; 12; 8) $=$ $= 0{,}339$;

— «F критическое» — это критическая точка $f_{k_1, k_2, \alpha_д}$, которая находится с помощью **Статистической функции** FРАСПОБР($\alpha_д$; k_1; k_2). В условиях задачи $\alpha_д = \alpha/2 = 0{,}05$ и $f_{12;\,8;\,0{,}05}$ = FРАСПОБР(0,05; 12; 8) = 3,284. (Сравните это число с найденным по таблице П. 5.)

Ответим на вопрос, принять гипотезу H_0: $\sigma_X^2 = \sigma_Y^2$ или отклонить в пользу гипотезы H_1: $\sigma_X^2 \neq \sigma_Y^2$, двумя способами:

1) по таблице 9.3 область отклонения гипотезы H_0: $\sigma_X^2 = \sigma_Y^2$ определяется неравенством $F > f_{k_1 = n_X - 1,\, k_2 = n_Y - 1,\, \alpha/2}$. Это неравенство, судя по рисунку 9.6, *б*, не выполняется: 1,361 < 3,284, поэтому гипотезу H_0: $\sigma_X^2 = \sigma_Y^2$ принимаем;

2) сравнить «P односторонее» с уровнем значимости $\alpha_д$, указанным в диалоговом окне. Очевидно, что при «P односторонее» > $\alpha_д$ значение f критической статистики меньше $f_{k_1, k_2, \alpha_д}$ и гипотезу H_0 принимают; при «P односторонее» < < $\alpha_д$ гипотезу H_0 отклоняют. В условиях задачи «P односторонее» = 0,339 > 0,05 = α_g (f = 1,361 меньше «критического одностороннего», равного 3,284), поэтому гипотезу H_0: $\sigma_X^2 = \sigma_Y^2$ принимаем. ◄

Программа **«Двухвыборочный *t*-тест с одинаковыми дисперсиями»** используется для проверки гипотезы H_0: $a_X - a_Y = a_0$ (разность генеральных средних равна числу a_0) в том случае, когда значения генеральных дисперсий σ_X^2 и σ_Y^2 неизвестны, но есть основание считать эти значения равными. Исходными данными являются введенные в рабочий лист наблюдения переменной 1 (величины X) и переменной 2 (величины Y) и введенные в диалоговое окно уровень значимости α, при котором проверяется гипотеза H_0: $a_X - a_Y = a_0$, и число a_0 — «гипотетическая разность средних». По умолчанию $a_0 = 0$ и тогда проверяемая гипотеза принимает вид H_0: $a_X = a_Y$. В дальнейшем будем полагать, что $a_0 = 0$. Программа реализует проверку гипотезы H_0: $a_X = a_Y$ сразу при следующих двух альтернативах:

1) H_1: $a_X > a_Y$, если $\bar{x} > \bar{y}$ (при $\bar{x} > \bar{y}$ альтернативная гипотеза H_1: $a_X < a_Y$ не рассматривается, поскольку она всегда будет отклонена в пользу гипотезы H_0: $a_X = a_Y$), и H_1: $a_X < a_Y$, если $\bar{x} < \bar{y}$;

2) H_1: $a_X \neq a_Y$ при любом соотношении между $\bar{x}$ и $\bar{y}$.

Проанализируем результаты работы программы на примере задачи 9.7.

▶▶ **ЗАДАЧА 9.7** (продолжение). В условиях задачи 9.7 при уровне значимости $\alpha = 0{,}1$ выяснить: а) уменьшается ли при переходе на новую технологию средний расход сырья на единицу продукции; б) можно ли объяснить различие выборочных средних расходов сырья на единицу продукции при старой и новой технологиях случайной ошибкой.

Р е ш е н и е. Напомним, что при решении задачи 9.7 за переменную 1 (величину X) был принят расход сырья на единицу продукции при новой технологии, а за переменную 2 (величину Y) — расход сырья при старой технологии (рис. 9.6, *а*); напомним также, что была принята гипотеза H_0: $\sigma_X^2 = \sigma_Y^2$ об отсутствии различия в значениях дисперсий расхода сырья при новой (X) и старой (Y) технологиях.

С введенными в рабочее поле двумя столбцами данными (см. рис. 9.6, *а*) обратимся к программе «**Двухвыборочный *t*-тест с одинаковыми дисперсиями**», в диалоговом окне «уровень значимости» приравняем числу $\alpha = 0{,}1$, а «гипотетическую разность средних» — числу нуль. Результаты работы программы приведены на рисунке 9.6, *в*, это:

— средние $\bar{x} = 304{,}77$ (новая технология) и $\bar{y} = 307{,}11$ (старая технология);

— дисперсии $s_X^2 = 2{,}192$ и $s_Y^2 = 1{,}61(1)$;

— числа наблюдений $n_X = 13$ и $n_Y = 9$;

— объединенная дисперсия

$$s^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{(n_X - 1) + (n_Y - 1)} = \frac{12 \cdot 2{,}192 + 8 \cdot 1{,}61(1)}{13 + 9 - 2} = 1{,}96$$

(«объединить дисперсии s_X^2 и s_Y^2 можно, поскольку в данной программе предполагается, что $\sigma_X^2 = \sigma_Y^2$);

— гипотетическая разность средних, равная нулю;

— число степеней свободы критической статистики — случайной величины $T(k = n_X + n_Y - 2)$ (см. табл. 9.3, строка 3), $k = 13 + 9 - 2 = 20$;

— числовое значение названной критической статистики

$$t = (\bar{x} - \bar{y})/\sqrt{(1/n_X + 1/n_Y)s^2} =$$

$$= (304{,}769 - 307{,}11(1))/\sqrt{(1/13 + 1/9)1{,}96} = -3{,}858;$$

— результаты, необходимые для проверки гипотезы H_0: $a_X = a_Y$ при альтернативе H_1: $a_X < a_Y$ (так как $\bar{x} < \bar{y}$) и альтернативе H_1: $a_X \neq a_Y$. Поясним эти результаты.

При H_1: $a_X < a_Y$ (так же, как и при H_1: $a_X > a_Y$) программа вычисляет:

— «P одностороннее» — это вероятность того, что критическая статистика — случайная величина $T(k)$ превзойдет число $|t|$, т. е. «P одностороннее» = $P(T(k) > |t|)$, которая находится с помощью **Статистической функции** СТЬЮДРАСП($|t|$; k; 1). В условиях задачи «P одностороннее» = $P(T(20) > |-3{,}85778|)$ = СТЬЮДРАСП(3,85778; 20; 1) = 0,00049;

— «t критическое одностороннее» — это правосторонняя критическая точка $t_{k,\,2\alpha}$, которая находится с помощью **Статистической функции** СТЬЮДРАСПОБР(2α; k). В условиях задачи α = 0,1 и $t_{20;\,0{,}2}$ = СТЬЮДРАСПОБР(0,2; 20) = = 1,325341.

Ответ на вопрос, принять гипотезу H_0: $a_X = a_Y$ или отклонить в пользу гипотезы H_1: $a_X < a_Y$, дадим двумя способами.

1) Согласно таблице 9.3, область отклонения гипотезы H_0: $a_X = a_Y$ определяется неравенством $T < -\,t_{k=n_X+n_Y-2,\;2\alpha}$. Это неравенство, судя по результатам, представленным на рисунке 9.6, *в*, выполняется: t = –3,858, $t_{20;\,0{,}2}$ = 1,325 и –3,858 < –1,325, поэтому гипотезу H_0: $a_X = a_Y$ отклоняем и принимаем H_1: $a_X < a_Y$.

2) Сравнить «P одностороннее» с заданным уровнем значимости α. Если «P одностороннее» < α, гипотезу H_0 отклоняем. В задаче «P одностороннее» = 0,00049, α = 0,1, и поэтому гипотезу H_0: $a_X = a_Y$ отклоняем и принимаем H_1: $a_X < a_Y$.

При H_1: $a_X \neq a_Y$ программа вычисляет:

— «P двухстороннее», равное $2P(T(k = (n_X + n_Y - 2) > |t|)$ = СТЬЮДРАСП($|t|$; k; 2). По условию «P двухстороннее» = $2P(T(20) > |-3{,}85778|)$ = СТЬЮДРАСП(3,85778; 20; 2) = 0,000981;

— «t критическое двухстороннее» — правостороннюю критическую точку $t_{k,\,\alpha}$ = СТЬЮДРАСПОБР(α; k). В условиях задачи $t_{20;\,0{,}1}$ = СТЬЮДРАСПОБР(0,1; 20) = 1,724718.

Ответ на вопрос, принять гипотезу H_0: $a_X = a_Y$ или отклонить в пользу гипотезы H_1: $a_X \neq a_Y$, можно дать двумя способами:

1) согласно таблице 9.3, область отклонения гипотезы H_0: $a_X = a_Y$ определяется неравенством $|T| > t_{k\,=\,n_X\,+\,n_Y\,-\,2,\,\alpha}$. В задаче $|-3{,}858| > t_{20;\,0{,}1}$ = 1,725, поэтому гипотезу H_0: $a_X = a_Y$ отклоняем в пользу гипотезы H_1: $a_X \neq a_Y$;

2) сравнить «P двухстороннее» с α; если «P двухстороннее» $< \alpha$, H_0 отклоняют. В задаче «P двухстороннее» $=$ $= 0{,}000981 < 0{,}1$, поэтому H_0: $a_X = a_Y$ отклоняем в пользу гипотезы H_1: $a_X \neq a_Y$.

Напомним, в задаче 9.7 требовалось при $\alpha = 0{,}1$ ответить на два вопроса: а) уменьшается ли при переходе на новую технологию расход сырья на единицу продукции; б) можно ли объяснить различие средних расходов на единицу продукции при старой и новой технологиях случайной ошибкой?

Для ответа на вопрос а) надо было проверить гипотезу H_0: $a_X = a_Y$, где a_X (a_Y) — математическое ожидание расхода сырья на единицу продукции при новой (старой) технологии, при альтернативе H_1: $a_X < a_Y$; в соответствии с изложенным выше была принята гипотеза H_1: $a_X < a_Y$ — уменьшение расхода сырья при новой технологии скорее всего имеет место.

Для ответа на вопрос б) следовало проверить гипотезу H_0: $a_X = a_Y$ при альтернативе H_1: $a_X \neq a_Y$; в соответствии с изложенным была принята гипотеза H_1: $a_X \neq a_Y$ — различие средних расходов сырья при новой и старой технологиях неслучайно. ◀

Программа «**Двухвыборочный t-тест с разными дисперсиями**» используется для проверки гипотезы H_0: $a_X - a_Y = a_0$ в том случае, когда значения генеральных дисперсий σ_X^2 и σ_Y^2 неизвестны, но есть основание считать эти значения неравными.

В этом случае критическая статистика, как и в случае «**Двухвыборочного t-теста с равными дисперсиями**», — случайная величина Стьюдента, но ее вид и число степеней свободы отличаются от вида и числа степеней свободы величины Стьюдента, используемой «при равных дисперсиях» (см. таблицу 9.3, третью и четвертую строки). По форме распечатки результатов этих программ не различаются.

Программа «**Двухвыборочный Z-тест для средних**» используется для проверки гипотезы H_0: $a_X - a_Y = a_0$, когда числовые значения генеральных дисперсий σ_X^2 и σ_Y^2 известны. Здесь при $a_0 = 0$ выдаются результаты, необходимые для проверки гипотезы H_0: $a_X = a_Y$ с помощью критической статистики Z (см. табл. 9.3, первую строку), сразу при двух альтернативах:

1) H_1: $a_X > a_Y$, если $\bar{x} > \bar{y}$, и H_1: $a_X < a_Y$, если $\bar{x} < \bar{y}$ (при «P одностороннее» $< \alpha$ гипотезу H_0: $a_X = a_Y$ отклоняют);

2) H_1: $a_X \neq a_Y$ при любом соотношении между $\bar{x}$ и $\bar{y}$ (если «P двухстороннее» $< \alpha$, гипотезу H_0: $a_X = a_Y$ отклоняют).

Рассмотренные программы работают лишь в случае, когда *известны непосредственные результаты наблюдений двух случайных величин*. Последовательность подключения этих программ при проверке гипотезы H_0: $a_X = a_Y$ изображена на рисунке 9.7.

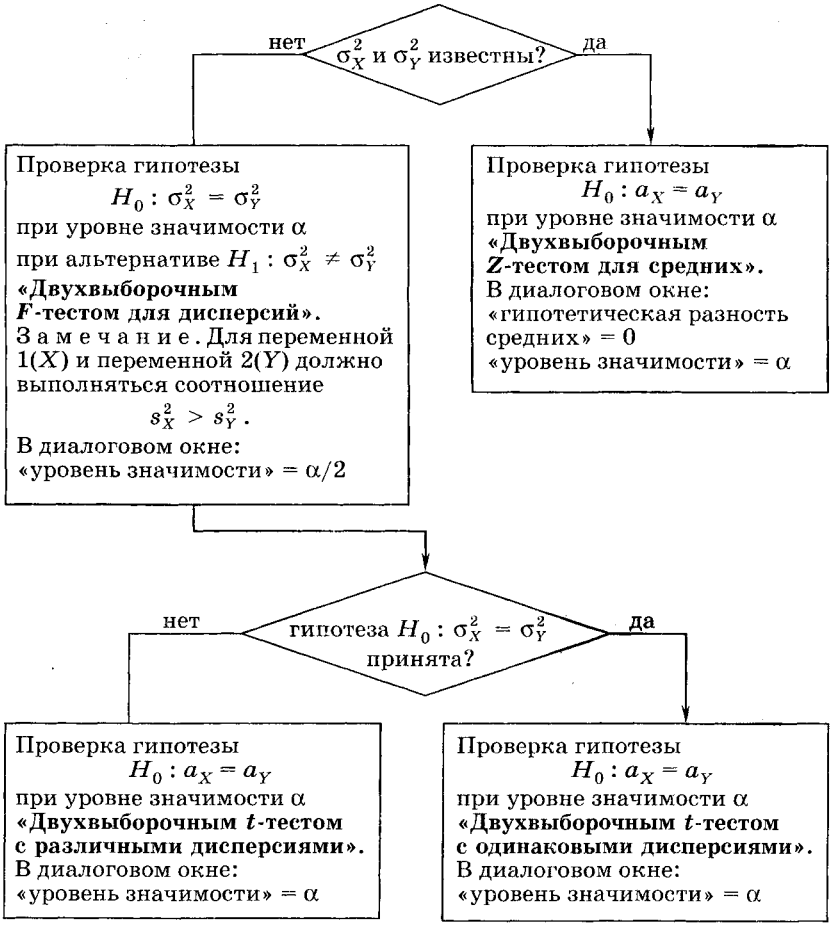


Рис. 9.7

§ 9.5. Проверка гипотезы о равенстве неизвестных значений вероятностей

А. Сравнение двух вероятностей. Пусть p_1 и p_2 — вероятности появления соответственно событий A_1 и A_2, числовые значения которых неизвестны (например, p_1 и p_2 — соответственно вероятности того, что случайно выбранные мужчина и женщина — пацифисты). Требуется проверить гипотезу H_0: $p_1 = p_2$ при заданном уровне значимости α, если известно, что в n_1 испытаниях Бернулли (независимых испытаниях, проведенных в типичных условиях) событие A_1 появилось m_1 раз (среди опрошенных n_1 мужчин оказалось m_1 пацифистов), а в n_2 испытаниях Бернулли событие A_2 появилось m_2 раз (среди опрошенных n_2 женщин оказалось m_2 пацифисток).

Критическая статистика и критерий, используемые при проверке гипотезы H_0: $p_1 = p_2$, когда n_1 и n_2 — достаточно большие числа ($n_1 \gg 100$, $n_2 \gg 100$) приведенные в таблице 9.3 (строка 5). При выполнении гипотезы H_0: $p_1 = p_2$ критическая статистика Z является стандартной нормально распределенной величиной, функция плотности которой изображена на рисунке 9.2 (при H_1: $p_1 > p_2$ — критическая область правосторонняя, при H_1: $p_1 < p_2$ — левосторонняя, при H_1: $p_1 \neq p_2$ — двухсторонняя). Критическую точку находят по таблице П.1.

Если гипотезу H_0: $p_1 = p_2$ принимают, то говорят, что *различие опытных вероятностей* $\hat{p}_1 = m_1/n_1$ и $\hat{p}_2 = m_2/n_2$, которое может иметь место, *случайно*, *статистически незначимо* или *несущественно*, и за оценку общей вероятности p ($p_1 = p_2 = p$) принимают

$$\hat{p} = (m_1 + m_2)/(n_1 + n_2).$$

ЗАДАЧА 9.8. Существует общее мнение, что среди женщин больше пацифистов, чем среди мужчин. При опросе $n_1 = 400$ мужчин оказалось, что 40% из них — пацифисты. При опросе $n_2 = 300$ женщин 50% из них оказались пацифистками. Можно ли считать, что женщины более склонны к пацифизму? Принять $\alpha = 0{,}05$.

Р е ш е н и е. Обозначим через p_1 и p_2 соответственно вероятности того, что случайно выбранные мужчина и женщина — пацифисты. По условию требуется проверить гипотезу H_0: $p_1 = p_2$ при альтернативе H_1: $p_1 < p_2$, при этом $\hat{p}_1 = 0{,}4$; $m_1 = 160$, $\hat{p}_2 = 0{,}5$; $m_2 = 150$.

Выполним необходимые вычисления согласно алгоритму, изложенному в табл. 9.3 (см. пятую строку, $H_1: p_1 < p_2$):

— предположив, что гипотеза $H_0: p_1 = p_2$ выполняется, найдем оценку $\hat{p}$ общей вероятности $p_1 = p_2 = p$

$$\hat{p} = (160 + 150)/(400 + 300) = 0{,}443;$$

— вычислим значение z критической статистики Z

$$z = \frac{0{,}4 - 0{,}5}{\sqrt{(1/400 + 1/300)0{,}443(1 - 0{,}443)}} = -2{,}636;$$

— используя таблицу П. 1, найдем число $z_{0,5-\alpha} = z_{0,5-0,05} = z_{0,45} = 1{,}65$ (в таблице П. 1 ближайшим к $\Phi = 0{,}45$ будет число 0,4505, ему соответствует $z = 1{,}65$); левосторонняя критическая точка $-z_{0,5-\alpha} = -z_{0,45} = -1{,}65$;

— так как число $z < -z_{0,5-\alpha}$ $(-2{,}636 < -1{,}65)$, гипотезу $H_0: p_1 = p_2$ отклоняем, принимаем гипотезу $H_1: p_1 < p_2$. Таким образом, мнение о том, что среди женщин пацифистов больше, чем среди мужчин, не противоречит результатам наблюдений.

Если за альтернативу гипотезе $H_0: p_1 = p_2$ принять гипотезу $H_1: p_1 \neq p_2$, то при $\alpha = 0{,}05$ гипотеза $H_0: p_1 = p_2$ также будет отклонена, поскольку $|z| = |-2{,}636|$ больше $z_{0,5-\alpha/2} = z_{0,5-0,05/2} = z_{0,475} = 1{,}95$ (см. табл. 9.3, пятую строку, $H_1: p_1 \neq p_2$). ◀

Б. Сравнение двух и более вероятностей с помощью критерия χ^2. Допустим, что фирма имеет в четырех ($q = 4$) городах отделения, производящие однотипные изделия трех ($\nu = 3$) сортов. Обозначим через p_{ji} вероятность того, что j-е отделение выпустит изделие i-го сорта ($j = 1, 2, 3, 4$; $i = 1, 2, 3$); эти вероятности находятся в левом верхнем углу каждой клетки таблицы 9.4. Числовые значения вероятностей неизвестны, но требуется выяснить, можно ли считать, что вероятность изготовления изделия того или иного сорта постоянна, не зависит от номера отделения, т. е. требуется проверить гипотезу

$$\left.\begin{aligned} H_0: p_{11} = p_{21} = p_{31} = p_{41} (= p_1),^{(*)} \\ p_{12} = p_{22} = p_{32} = p_{42} (= p_2),^{(**)} \\ p_{13} = p_{23} = p_{33} = p_{43} (= p_3)^{(***)} \end{aligned}\right\} \qquad (9.30)$$

о том, что вероятность изготовления изделия первого сорта на каждом из отделений фирмы одинакова (равна p_1), вероятность изготовления изделия второго сорта на каждом из отделений фирмы одинакова (равна p_2) и т. д.

Таблица 9.4

j \ i	1	2	$\nu = 3$	m_{j*}
1	p_{11} $m_{11} = 60$ 56	p_{12} $m_{12} = 25$ 31	p_{13} $m_{13} = 15$ 13	$m_{1*} = 100$
2	p_{21} $m_{21} = 100$ 85	p_{22} $m_{22} = 25$ 46	p_{23} $m_{23} = 25$ 19	$m_{2*} = 150$
3	p_{31} $m_{31} = 50$ 56	p_{32} $m_{32} = 40$ 31	p_{33} $m_{33} = 10$ 13	$m_{3*} = 100$
$q = 4$	p_{41} $m_{41} = 100$ 113	p_{42} $m_{42} = 80$ 62	p_{43} $m_{43} = 20$ 25	$m_{4*} = 200$
m_{*i}	$m_{*1} = 310$	$m_{*2} = 170$	$m_{*3} = 70$	$m_{**} = n = 550$

Замечание. Так как изделия могут быть только трех сортов, то по каждому отделению (по каждой строке таблицы 9.4) сумма вероятностей равна единице, следовательно, $p_{13} = 1 - (p_{11} + p_{12})$, $p_{23} = 1 - (p_{21} + p_{22})$, $p_{33} = 1 - (p_{31} + p_{32})$, $p_{43} = 1 - (p_{41} + p_{42})$. И если в (9.30) выполняются соотношения (*) и (**), то будет выполняться и соотношение (***). Поэтому гипотеза (9.30) равносильна гипотезе

$$\left.\begin{aligned} H_0\colon\ & p_{11} = p_{21} = p_{31} = p_{41}\ (= p_1),^{(*)} \\ & p_{12} = p_{22} = p_{32} = p_{42}\ (= p_2).^{(**)} \end{aligned}\right\} \qquad (9.31)$$

Проведем выборочные наблюдения. Пусть первое отделение произвело $m_{1*} = 100$ изделий, из которых первого сорта оказалось $m_{11} = 60$, второго сорта $m_{12} = 25$ и третьего сорта $m_{13} = 15$. Эти данные, а также данные о количестве изделий, произведенных другими отделениями фирмы, и их распределении по сортам приведены в таблице 9.4. Частоты $m_{11}, m_{12}, \ldots, m_{43}$ называют клеточными, частоты $m_{1*}, m_{2*}, m_{3*}, m_{4*}$, равные суммам клеточных частот по каждой строке, — маргинальными (краевыми) строчными, частоты m_{*1}, m_{*2}, m_{*3}, равные суммам клеточных частот по каждому столбцу, — маргинальными по столбцам. Общее число наблюдений $n = m_{**} = 550$.

Допустим, что гипотеза (9.30) выполняется. Это означает следующее.

1) Вероятность изготовления изделия первого сорта одинакова для всех отделений, поэтому оценка общей вероятности p_1 равна отношению числа изделий первого сорта, произведенных всеми отделениями, к общему числу произведенных изделий, т. е.

$$\hat{p}_1 = m_{*1}/m_{**}. \tag{9.32}$$

Тогда число изделий первого сорта, выпущенных j-м отделением, которое обозначим $m_{j1}^{\text{теор}}$, должно быть равно произведению общего числа m_{j*}, изделий, выпущенных этим подразделением, на $\hat{p}_1$, т. е.

$$m_{j1}^{\text{теор}} = m_{j*}\hat{p}_1 \underset{(9.32)}{=} m_{j*}m_{*1}/m_{**}. \tag{9.33}$$

В соответствии с формулой (9.33) получим $m_{11}^{\text{теор}} = m_{1*}m_{*1}/m_{**} = 100 \cdot 310/550 \approx 56$, $m_{21}^{\text{теор}} = m_{2*}m_{*1}/m_{**} = 150 \cdot 310/550 \approx 85$, $m_{31}^{\text{теор}} = m_{3*}m_{*1}/m_{**} = 100 \cdot 310/550 \approx 56$, $m_{41}^{\text{теор}} = m_{4*}m_{*1}/m_{**} = 200 \cdot 310/550 \approx 113$ — эти частоты проставлены в правом нижнем углу первого столбца таблицы 9.4.

2) Вероятность изготовления изделия второго сорта одинакова для всех отделений, поэтому оценка общей вероятности p_2 равна

$$\hat{p}_2 = m_{*2}/m_{**}. \tag{9.34}$$

Тогда число изделий второго сорта, произведенных j-м отделением,

$$m_{j2}^{\text{теор}} = m_{j*}\hat{p}_2 \underset{(9.34)}{=} m_{j*}m_{*2}/m_{**}. \tag{9.35}$$

Воспользовавшись формулой (9.35), найдем

$$m_{12}^{\text{теор}} = 100 \cdot 170/550 \approx 31,\ m_{22}^{\text{теор}} = 150 \cdot 170/550 \approx 46,$$

$$m_{32}^{\text{теор}} = 100 \cdot 170/550 \approx 31,\ m_{42}^{\text{теор}} = 200 \cdot 170/550 \approx 62.$$

Обобщим формулы (9.33) и (9.35):

$$m_{ji}^{\text{теор}} = m_{j*}m_{*i}/m_{**}. \tag{9.36}$$

Таким образом, *теоретическая частота в клетке* (j, i), т. е. частота, рассчитанная в предположении равенства вероятностей, стоящих в q клетках i-го столбца, *равна частному от деления произведения маргинальных частот j-й строки и i-го столбца на общее число наблюдений.*

Клеточные теоретические частоты третьего столбца рассчитаем по формуле

$$m_{j3}^{\text{теор}} = m_{j*}m_{*3}/m_{**}, \qquad j = 1, 2, 3, 4,$$

вытекающей из (9.36) при $i = 3$. Эти частоты находятся в правом нижнем углу клеток третьего столбца таблицы 9.4.

Замечание. Обратим внимание на то, что сумма клеточных теоретических частот $m_{ji}^{\text{теор}}$ по каждой строке (каждому столбцу) должна быть равной маргинальной частоте соответствующей строки (соответствующего столбца).

В качестве критической статистики при проверке гипотезы (9.30) используют величину

$$\chi^2 = \sum_{j=1}^{4} \sum_{i=1}^{3} \frac{(m_{ji} - m_{ji}^{\text{теор}})^2}{m_{ji}^{\text{теор}}}, \qquad (9.37)$$

обобщение которой на любое число q строк ($q > 1$) и любое число ν столбцов ($\nu > 1$) имеет вид

$$\chi^2 = \sum_{j=1}^{q} \sum_{i=1}^{\nu} \frac{(m_{ji} - m_{ji}^{\text{теор}})^2}{m_{ji}^{\text{теор}}}. \qquad (9.38)$$

Предположим, что число строк q и столбцов ν, а также все маргинальные частоты — фиксированные числа, следовательно, и общее число наблюдений m_{**} — фиксированное число. В этих условиях при переходе от одной серии m_{**} наблюдений к другой теоретические частоты, рассчитываемые по формуле (9.36), — вполне определенные числа, а значения клеточных частот m_{ji} заранее не предсказуемы. Поэтому в формуле (9.38) m_{ji} могут иметь двоякую интерпретацию: это случайные величины и тогда χ^2 — случайная величина, или m_{ji} — числа, полученные в конкретном выборочном обследовании, и тогда χ^2 — число.

Доказано, что распределение величины (9.38), рассматриваемой как случайной, при выполнении гипотезы о равенстве q клеточных вероятностей в каждом из ν столбцов [см. таблицу 9.4 и гипотезу (9.30)] имеет, при достаточно большом общем числе наблюдений m_{**} и $m_{ji}^{\text{теор}} > 5$, распределение, близкое к χ^2-распределению с числом степеней свободы $k = (q - 1)(\nu - 1)$. Это дает основание, располагая результатами m_{**} наблюдений, сгруппированными в таблицу типа таблицы 9.4, построить процедуру проверки гипотезы о равенстве q клеточных вероятностей в каждом из ν столбцов (см. (9.30)) при уровне значимости α следующим образом:

— в соответствии с формулой (9.36) рассчитать теоретические частоты $m_{ji}^{\text{теор}}$;

— рассчитать значение χ^2-критической статистики (9.38); в условиях таблицы 9.4 $\chi^2 = (60 - 56)^2/56 + (25 - 31)^2/31 + (15 - 13)^2/13 + (100 - 85)^2/85 + (25 - 46)^2/46 +$

$+(25-19)^2/19+(50-56)^2/56+(40-31)^2/31+(10-13)^2/13+$
$+(100-113)^2/113+(80-62)^2/62+(20-25)^2/25=27{,}6;$

— если $\chi^2 < \chi^2_{(q-1)(\nu-1),1-\alpha/2}$ или $\chi^2 > \chi^2_{(q-1)(\nu-1),\alpha/2}$, где $\chi^2_{(q-1)(\nu-1),1-\alpha/2}$ и $\chi^2_{(q-1)(\nu-1),\alpha/2}$ — соответственно левосторонняя и правосторонняя критические точки, найденные по таблице П.2 при $k=(q-1)(\nu-1)$ и вероятностях $p=1-\alpha/2$ и $p=\alpha/2$, то гипотезу о равенстве q внутриклеточных вероятностей в каждом из ν столбцов отклоняют[1]. Если $\chi^2_{(q-1)\cdot(\nu-1),1-\alpha/2} < \chi^2 < \chi^2_{(q-1)\cdot(\nu-1),\alpha/2}$, то эту гипотезу не отклоняют. В условиях таблицы 9.4 $(q-1)(\nu-1)=(4-1)(3-1)=6$. Приняв $\alpha=0{,}05$, по таблице П. 2 найдем $\chi^2_{6;1-0{,}05/2}=\chi^2_{6;0{,}975}=1{,}24$ и $\chi^2_{6;0{,}05/2}=14{,}45$. Так как число $\chi^2=27{,}6>\chi^2_{6;0{,}05/2}$, то гипотезу о том, что вероятность изготовления изделия фиксированного сорта одинакова для всех отделений (и не зависит от номера отделения), отклоняем.

В заключение заметим, что рассмотренный χ^2-критерий может быть использован и для проверки гипотезы о равенстве двух вероятностей.

§ 9.6. Проверка гипотезы о законе распределения случайной величины. Критерий согласия Пирсона

Для того чтобы рассчитать те или иные вероятности, надо знать закон распределения случайной величины. Однако во многих практических задачах известны лишь результаты наблюдений величины X, а ее закон распределения неизвестен и возникает задача проверки гипотезы о виде, или о модели этого закона, т. е. необходимо выяснить, является ли неизвестный закон нормальным, или биномиальным, или каким-нибудь другим.

Универсальной формой задания закона распределения дискретной и непрерывной случайной величины X является функция распределения $F_X(x)$, поэтому упомянутую выше гипотезу можно в общем виде записать так:

$$H_0: F_X(x)=F_{\text{теор}}(x), \qquad (9.39)$$

[1] Отклонение гипотезы в случае слишком малых значений критической статистики (9.38), а именно при $\chi^2 < \chi^2_{(q-1)(\nu-1),1-\alpha/2}$, только на первый взгляд противоречит здравому смыслу. Не следует забывать, что величина (9.38) — случайная, поэтому маловероятно появление не только слишком больших ее значений ($\chi^2 > \chi^2_{(q-1)(\nu-1),\alpha/2}$), но и слишком малых.

неизвестный закон распределения $F_X(x)$ изучаемой величины X предполагается совпадающим с законом $F_{\text{теор}}(x)$ конкретного вида, конкретной модели, которую называют «теоретической» — предполагаемой моделью закона распределения величины X.

В качестве теоретической модели, как отмечалось ранее, может быть рассмотрена нормальная, биномиальная или какая-то другая модель. Это определяется сущностью изучаемого явления, а также результатами предварительной обработки наблюдений случайной величины (формой многоугольника распределения частостей, полигона, соотношениями между выборочными характеристиками и т. д.).

Критерии, с помощью которых проверяется гипотеза (9.39), называются ***критериями согласия***. Рассмотрим один из них, использующий χ^2-распределение и получивший название *критерия согласия Пирсона*.

Процедура проверки гипотезы (9.39) с помощью χ^2-критерия такова.

Предположив, что изучаемая случайная величина X подчиняется предполагаемому закону распределения:

1) *весь диапазон* значений случайной величины X, а не только наблюдавшихся, разбивают на ν непересекающихся групп (интервалов) и, зная результаты n наблюдений величины X, подсчитывают числа m_i, $i = 1, 2, \ldots, \nu$, наблюдений, попавших в соответствующие группы;

2) находят выборочные оценки тех входящих в модель предполагаемого закона распределения параметров, числовые значения которых неизвестны (количество таких параметров обозначим через l), и в предполагаемой модели эти параметры заменяют их выборочными оценками;

3) рассчитывают теоретические вероятности $p_i^{\text{теор}}$ того, что случайная величина X примет значения из i-й группы, $i = 1, 2, \ldots, \nu$, и теоретические частоты $m_i^{\text{теор}} = np_i^{\text{теор}}$, $i = 1, 2, \ldots, \nu$; при этом, если для некоторых из групп $m_i^{\text{теор}} \leqslant 5$, то их объединяют с соседними так, чтобы в результате для каждой группы «теоретическая» частота была больше 5 (новое число групп обозначим через ν^*);

4) за меру расхождения результатов наблюдений величины X с предполагаемой моделью закона распределения принимают величину

$$\chi^2 = \sum_{i=1}^{\nu^*} \frac{(m_i - m_i^{\text{теор}})^2}{m_i^{\text{теор}}}, \qquad (9.40)$$

которая, при ее интерпретации как случайной величины, имеет (в предположении справедливости гипотезы (9.39) и

достаточно большом n) распределение, близкое к χ^2-распределению с числом степеней свободы $k = \nu^* - l - 1$, т. е.

$$\chi^2 = \sum_{i=1}^{\nu^*} \frac{(m_i - m_i^{\text{теор}})^2}{m_i^{\text{теор}}} = \chi^2(\nu^* - l - 1), \qquad (9.41)$$

где ν^* — новое число групп разбиения диапазона значений величины X (оставшееся после объединения некоторых из первоначальных ν групп, если упомянутая выше причина объединения имела место); l — число параметров предполагаемой модели закона распределения величины X, значения которых неизвестны и которые были заменены их выборочными оценками. Если значения всех параметров модели известны, то $l = 0$.

Замечание. Случайность величины (9.40) объясняется непредсказуемостью результатов n наблюдений случайной величины X, следовательно, непредсказуемостью групповых частот m_i при заданном разбиении на группы диапазона значений величины X.

Гипотезу (9.39) отклоняют, если вычисленная по формуле (9.40) величина $\chi^2 < \chi^2_{\nu^*-l-1,1-\alpha/2}$ или $\chi^2 > \chi^2_{\nu^*-l-1,\alpha/2}$, где $\chi^2_{\nu^*-l-1,1-\alpha/2}$ и $\chi^2_{\nu^*-l-1,\alpha/2}$ — соответственно левосторонняя и правосторонняя критические точки, определяемые по таблице П. 2 при $k = \nu^* - l - 1$ и вероятностях $p = 1 - \alpha/2$ и $p = \alpha/2$ (α — заданный уровень значимости). Гипотезу (9.39) не отклоняют, если $\chi^2_{\nu^*-l-1,1-\alpha/2} < \chi^2 < \chi^2_{\nu^*-l-1,\alpha/2}$.

▶ **ПРИМЕР 9.1** (продолжение примеров 7.1 и 7.11). В примере 7.1 рассматривалась случайная величина X — количество сданных экзаменов (из четырех сдаваемых) случайно выбранным студентом, и по сведениям о количестве сданных экзаменов каждым из $n = 100$ случайно выбранных студентов построен статистический ряд распределения (см. таблицу 7.4). Обратим внимание, что в этих сведениях — результатах наблюдений величины X — присутствовали все возможные ее значения: 0, 1, 2, 3, 4. Затем в предположении, что X — биномиально распределенная случайная величина, т. е. что выполняется гипотеза

$$H_0: P(X = x) = C_4^x p^x (1-p)^{4-x},\ x = 0, 1, 2, 3, 4, \qquad (9.42)$$

были в примере 7.11 рассчитаны вероятности сдачи студентом x экзаменов (эти вероятности приведены в последней строке таблицы 7.4), при этом неизвестное значение параметра p биномиальной модели было заменено его выбороч-

ной оценкой $\hat{p} = 0{,}88$, т. е. расчеты проводились по формуле $P(X = x) = C_4^x 0{,}88^x 0{,}12^{4-x}$.

Возникает вопрос, можно ли принять гипотезу (9.42), например, при уровне значимости $\alpha = 0{,}1$? Конкретизируем применительно к гипотезе (9.42) рассмотренные выше этапы использования χ^2-критерия.

1) Согласно гипотезе (9.42), величина X может принимать пять значений: 0, 1, 2, 3, 4. Поскольку в наблюдениях все эти значения были зафиксированы, их следует принять за «группы»: группа «$x = 0$», группа «$x = 1$», ... , группа «$x = 4$». Далее, «разбросав» по этим группам результаты 100 наблюдений, найти групповые частоты m_i.

Подчеркнем еще раз: в наблюдениях присутствовали все значения величины X (и 0, и 1, и 2, и 3, и 4), поэтому только что упомянутые групповые частоты m_i совпадают с частотами, приведенными в таблице 7.4 (если, например, в наблюдениях отсутствовало бы число 1, то частота группы «$x = 1$» была бы равна нулю).

2) Теоретической моделью является биномиальная с неизвестным числовым значением параметра p (поэтому $l = 1$). Заменим p его выборочной оценкой $\hat{p} = 0{,}88$ (см. пример 7.11).

3) Рассчитаем вероятности

$$p_i^{\text{теор}} = P(X = x_i) = C_4^{x_i}\, \hat{p}^{x_i}(1 - \hat{p})^{4 - x_i},\ x_i = 0, 1, 2, 3, 4,$$

и частоты $m_i^{\text{теор}} = np_i^{\text{теор}} = 100p_i^{\text{теор}}$; они приведены в таблице 9.5.

Таблица 9.5

Номер группы, i	Группы значений величины X	Частоты, m_i (см. табл. 7.4)	$p_i^{\text{теор}}$ (см. табл. 7.4)	$m_i^{\text{теор}} = 100p_i^{\text{теор}}$	$\dfrac{(m_i - m_i^{\text{теор}})^2}{m_i^{\text{теор}}}$
1	0	1	0,00021	0,021	
2	1	1 } 5	0,00608	0,608 } 7,32	0,735
3	2	3	0,06691	6,691	
4	3	35	0,32711	32,711	0,160
5	4	60	0,59969	59,969	0,000
	Итого	100	1,00000	100	$0{,}895 = \chi^2$

В таблице 9.5 первые три группы объединены в одну. Это вызвано тем, что и $m_1^{\text{теор}} < 5$, и $m_2^{\text{теор}} < 5$, и $m_1^{\text{теор}} + m_2^{\text{теор}} < 5$,

а $m_1^{\text{теор}} + m_2^{\text{теор}} + m_3^{\text{теор}} = 7{,}32 > 5$. Поэтому новое число групп $\nu^* = 3$.

4) В итоговой строке последнего столбца таблицы 9.5 найдем значение критической статистики (9.40): $\chi^2 = 0{,}895$. Затем по таблице П. 2 найдем левостороннюю и правостороннюю критические точки:

$$\chi^2_{\nu^*-l-1,1-\alpha/2} = \chi^2_{3-1-1,1-0,1/2} = \chi^2_{1;0,95} = 0{,}004$$

и

$$\chi^2_{\nu^*-l-1,\alpha/2} = \chi^2_{1;0,05} = 3{,}84.$$

Так как $\chi^2 = 0{,}895 \in (0{,}004;\ 3{,}84)$, то делаем вывод, что гипотеза (9.42) о биномиальном законе распределения числа экзаменов, сданных из четырех экзаменов случайно выбранным студентом, не противоречит результатам наблюдений. Более того, эта гипотеза хорошо согласуется с результатами наблюдений: число $\chi^2 = 0{,}895$, попадая внутрь названного интервала, достаточно удалено от его концов — критических точек.

ПРИМЕР 9.2 (продолжение примера 7.12 — задачи Борткевича). В задаче рассматривалась случайная величина X — число убитых ударом копыта в армейском корпусе за год. В таблице 7.13 представлен статистический ряд распределения величины X, построенный польским исследователем Борткевичем по данным о числе убитых лиц в каждом из десяти корпусов за каждый год в течение 20 лет (1875—1894); общее число наблюдений $n = 10 \cdot 20 = 200$.

Обратим внимание на то, что в 200 наблюдениях зафиксированы не все возможные значения величины X (максимальное из которых равно численности корпуса, т. е. практически неограниченно), а лишь значения 0, 1, 2, 3, 4.

В примере 7.12 в предположении, что X — величина, распределенная по закону Пуассона, т. е. что выполняется гипотеза

$$H_0\colon P(X = x) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, \qquad x = 0, 1, 2, \ldots, \tag{9.43}$$

были рассчитаны «пуассоновские» вероятности для $x = 0, 1, 2, 3, 4$ (эти вероятности приведены в последней строке таблицы 7.13), при этом неизвестное значение параметра λ (напомним, что $\lambda = MX$) в пуассоновской модели заменено его выборочной оценкой $\bar{x} = 0{,}61$, т. е. расчеты проводились по формуле

$$P(X = x) = \frac{0{,}61^x}{x!}\mathrm{e}^{-0{,}61}, \qquad x = 0, 1, 2, 3, 4.$$

Ответим на вопрос, можно ли принять гипотезу (9.43), например при уровне значимости 0,05? Конкретизируем применительно к гипотезе (9.43) этапы использования χ^2-критерия.

1) Согласно гипотезе (9.43), число значений величины X не ограничено. В наблюдениях были зафиксированы значения 0, 1, 2, 3, 4; значений, больших 4, не было, а они возможны, поэтому группы будут такими: «$x = 0$», «$x = 1$», «$x = 2$», «$x = 3$» и «$x \geqslant 4$». Эти группы приведены в таблице 9.6. «Разбросав» результаты 200 наблюдений по группам, найдем групповые частоты m_i; они совпадут с частотами, приведенными в таблице 7.13.

2) Параметр λ пуассоновского закона, числовое значение которого неизвестно, заменим его выборочной оценкой $\bar{x} = 0{,}61$.

3) Найдем теоретические — пуассоновские вероятности: $p_1^{\text{теор}} = P(X = 0) = \frac{0{,}61^0}{0!} e^{-0{,}61}$, $p_2^{\text{теор}} = P(X = 1) = \frac{0{,}61^1}{1!} e^{-0{,}61}$, $p_3^{\text{теор}} = P(X = 2) = \frac{0{,}61^2}{2!} e^{-0{,}61}$, $p_4^{\text{теор}} = P(X = 3) = \frac{0{,}61^3}{3!} e^{-0{,}61}$ (их значения приведены в последней строке таблицы 7.13) и вероятность $p_5^{\text{теор}} = P(X \geqslant 4)$, значение которой в таблице 7.13 не указано (в ней дана $P(X = 4)$). Найдем эту вероятность:

$$P(X \geqslant 4) = 1 - [P(X = 0) + \ldots + P(X = 3)] =$$
$$= 1 - (0{,}543 + 0{,}331 + 0{,}101 + 0{,}020) = 0{,}005.$$

Вероятности $p_i^{\text{теор}}$ и теоретические частоты $m_i^{\text{теор}} = 200 p_i^{\text{теор}}$ приведены в таблице 9.6.

Таблица 9.6

Номер группы, i	Группы значений величины X	Частоты, m_i (см. табл. 7.13)	$p_i^{\text{теор}}$	$m_i^{\text{теор}} = 200 p_i^{\text{теор}}$	$\frac{(m_i - m_i^{\text{теор}})^2}{m_i^{\text{теор}}}$
1	0	109	0,543	108,6	0,0015
2	1	65	0,331	66,2	0,0218
3	2	22	0,101	20,2	
4	3	3 } 26	0,020	4,0 } 25,2	0,0254
5	$x \geqslant 4$	1	0,005	1,0	
	Итого	200	1,000	200	$0{,}0487 = \chi^2$

В таблице 9.6 последние три группы объединены в одну: только таким образом можно получить теоретические частоты, большие 5.

4) В итоговой строке последнего столбца таблицы 9.6 найдем значение $\chi^2 = 0{,}0487$. Затем по таблице П. 2 найдем критические точки, при этом будем иметь в виду, что число групп $v^* = 3$, а $l = 1$ (в распределении Пуассона значение параметра λ не было известно; оно было заменено выборочным средним $\bar{x} = 0{,}61$). Критические точки: $\chi^2_{v^*-l-1,1-\alpha/2} = \chi^2_{1;0{,}975} = 0{,}001$, $\chi^2_{v^*-l-1,\alpha/2} = \chi^2_{1;0{,}025} = 5{,}02$. Число $\chi^2 = 0{,}0487 \in (0{,}001;\ 5{,}02)$, поэтому гипотезу (9.43), состоящую в том, что число убитых ударом копыта в корпусе за год имеет пуассоновское распределение, не отклоняем.

ПРИМЕР 9.3 (продолжение примеров 7.2 и 7.13). В примере 7.2 рассматривалась случайная величина X — ежедневный объем продаж дилером товара за день, и по ежедневным сведениям об объеме продаж за 100 дней построен интервальный ряд (см. таблицу 7.9, столбцы 2 и 5). Затем в примере 7.13 в предположении, что X — нормально распределенная случайная величина, т. е. что выполняется гипотеза

$$H_0: F_X(x) = F_{N(a,\sigma)}(x)^1, \qquad -\infty < x < +\infty \qquad (9.44)$$

(функция распределения величины X совпадает с функцией распределения «нормальной» величины), рассчитаны значения $F_{N(a,\sigma)}(a_{i+1})$ функции $F_{N(a,\sigma)}(x)$ в концах каждого интервала (см. табл. 7.9, столбец 13). При этом неизвестные значения параметров a и σ нормального закона были заменены их оценками, рассчитанными по интервальному ряду: параметр a заменен на $\bar{x} = 49{,}485 \approx 49{,}5$; σ заменено на $\hat{\sigma}_{(\text{Ш})} = \sqrt{\hat{D}X_{(\text{Ш})}} = \sqrt{117{,}2} = 10{,}83$ (напомним, что $\hat{D}X_{(\text{Ш})} = \hat{D}X - h^2/12$ — рассчитанная по интервальному ряду дисперсия $\hat{D}X$, скорректированная на поправку Шеппарда $h^2/12$; h — длина интервала).

[1] Объяснение допустимости предположения о том, что неотрицательная величина X (объем продаж неотрицателен) является нормально распределенной величиной, множеством значений которой является интервал $(-\infty;\ +\infty)$, было дано в примере 7.13; оно базируется на анализе имеющихся результатов наблюдений величины X. Вместе с тем при более широком толковании объема продаж он может иметь и отрицательные значения: например, в какой-то день по вине дилера товар был испорчен или утерян.

Замечание. Доказано, что при использовании χ^2-критерия для проверки согласия результатов наблюдений с нормальным законом, значения параметров a и σ которого неизвестны, более правдоподобный вывод будет получен при замене a и σ^2 соответственно средним $\bar{x}$ и дисперсией $\widehat{D}X_{(\text{Ш})}$, рассчитанными по интервальному ряду.

Чтобы ответить на вопрос, можно ли принять гипотезу (9.44), например, при уровне значимости 0,1, выполним следующее.

1) Предположив, что X — нормально распределенная величина, диапазон всех ее значений разобьем на группы. Так как в интервальном ряду (см. таблицу 7.9) первый интервал [20,95; 27,45), а последний девятый интервал [72,95; 79,45) и в этих интервалах были зафиксированы наблюдения, и поскольку значения величины X могут быть и меньше, чем 20,95, и больше, чем 79,45, то за первую группу значений величины X примем интервал $(-\infty; 27{,}45)$, а за последнюю девятую — интервал $[72{,}95; +\infty)$, остальные же группы (2-я, 3-я, ... , 8-я) совпадут с соответствующими интервалами интервального ряда. Группы значений величины X приведены в таблице 9.7.

2) Параметры нормального закона a и σ заменим их выборочными оценками $\bar{x} = 49{,}5$ и $\hat{\sigma}_{(\text{Ш})} = 10{,}83$.

3) Найдем теоретические — «нормальные» вероятности

$$p_i^{\text{теор}} = P(a_i < N(a, \sigma) < a_{i+1}) = F_N(a_{i+1}) - F_N(a_i),$$

$$i = 1, 2, \ldots, 9,$$

используя значения функции нормального распределения, приведенные в таблице 7.9 (столбец 13), и учитывая, что, согласно свойствам функции распределения, $F_N(-\infty) = 0$, а $F_N(+\infty) = 1$. Вероятности $p_i^{\text{теор}}$ и теоретические частоты $m_i^{\text{теор}} = 100 p_i^{\text{теор}}$ приведены в таблице 9.7.

В таблице 9.7 объединены первые две группы и последние две группы; в результате все теоретические частоты стали бóльшими пяти.

4) В итоговой строке последнего столбца таблицы 9.7 найдем значение критической статистики (9.40): $\chi^2 = 3{,}469$. Затем по таблице П. 2 найдем критические точки, при этом имеем в виду, что число группы $\nu^* = 7$, а $l = 2$ (в нормальном распределении значения параметров a и σ не были известны). Критические точки: $\chi^2_{\nu^*-l-1,1-\alpha/2} = \chi^2_{4;0,95} = 0{,}71$; $\chi^2_{\nu^*-l-1,\alpha/2} = \chi^2_{4;0,05} = 9{,}49$. Число $\chi^2 = 3{,}459 \in (0{,}71; 9{,}49)$, оно не попадает ни в левостороннюю, ни в правостороннюю

Таблица 9.7

Номер интервала, i	Группы значений величины X $[a_i, a_{i+1})$	Частоты m_i (см. табл. 7.9)	$F_N(a_i)$	$F_N(a_{i+1})$	$p_i^{\text{теор}} = F_N(a_{i+1}) - F_N(a_i)$	$m_i^{\text{теор}} = 100\ p_i^{\text{теор}}$	$\frac{(m_i - m_i^{\text{теор}})^2}{m_i^{\text{теор}}}$
1	$(-\infty; 27{,}45)$	1 } 8	0	0,0202	0,0202	2,02 } 7,35	0,057
2	[27,45; 33,95)	7	0,0202	0,0735	0,0533	5,33	
3	[33,95; 40,45)	12	0,0735	0,1977	0,1242	12,42	0,014
4	[40,45; 46,95)	25	0,1977	0,4013	0,2036	20,36	1,057
5	[46,95; 53,45)	18	0,4013	0,6368	0,2355	23,55	1,308
6	[53,45; 59,95)	20	0,6368	0,8289	0,1921	19,21	0,032
7	[59,95; 66,45)	9	0,8289	0,9394	0,1105	11,05	0,380
8	[66,45; 72,95)	7 } 8	0,9394	0,9842	0,0448	4,48 } 6,06	0,621
9	$[72{,}95; +\infty)$	1	0,9842	1	0,0158	1,58	
	Итого	$n = 100$			1,00000	100	$3{,}469 = \chi^2$

критические области. Поэтому гипотезу (9.44) о нормальности распределения ежедневного объема продаж товара не отклоняем. ◄

УПРАЖНЕНИЯ

1. По выборке объема n из нормальной совокупности проверяется гипотеза H_0: $a = 2$ при альтернативе H_1: $a \neq 2$. Если при уровне значимости $\alpha = 0{,}01$ гипотеза H_0 отклоняется, будет ли она принята при уровне значимости $\alpha = 0{,}05$?

2. По девяти отделениям банка вычислен средний по отделению суммарный вклад населения за месяц 48 тыс. ден. ед. и среднее квадратическое отклонение 3 тыс. ден. ед. Можно ли при уровне значимости 0,01 принять 50 тыс. ден. ед. в качестве нормативного месячного суммарного вклада?

3. Фирма рассылает рекламные каталоги заказчикам. Вероятность того, что организация, получившая каталог, закажет изделие фирмы равна 0,08. Было разослано 100 каталогов новой фирмы и получено 10 заказов. Можно ли считать, что новая форма каталога эффективнее ранее используемой? Принять $\alpha = 0{,}1$.

4. При измерении производительности двух агрегатов получены следующие результаты:

Агрегат 1	14,1	10,1	14,7	13,7	14,0
Агрегат 2	14,0	14,5	13,7	12,7	14,1

Можно ли при уровне значимости 0,1, считать, что в среднем производительность агрегатов одинакова? Используйте «Двухвыборочные тесты», имеющиеся в пакете **«Анализ данных»** Microsoft Excel.

5. Для изучения эффективности профилактического лекарства против аллергии были обследованы две группы людей: 120 и 180 человек, предрасположенных к этому заболеванию. В первой группе, принимавшей лекарство, заболело четыре человека; во второй, не принимавшей лекарство, заболело семь человек. Свидетельствуют ли при уровне значимости 0,05 эти результаты об эффективности лекарства?

6. Число телезрителей (мужчин и женщин) положительно (+), безразлично (±) и отрицательно (–), относящихся к рекламной телевизионной передаче, таково:

Отношение к передаче / Пол	+	±	–
мужчины	28	24	48
женщины	42	26	52

Предположив, что пол не влияет на отношение к рекламной передаче, найдите ожидаемые (теоретические) частоты. Влияет ли пол на отношение к рекламе? Примите $\alpha = 0{,}1$.

7. Проверьте гипотезу о том, что неизвестный закон распределения случайной величины X совпадает с некоторым определенным законом, имеющим два параметра, причем значение одного из них известно, а значение другого оценивается по выборочным данным. Опытные и теоретические частоты таковы:

m_i	2	5	15	14	15	21
$m_i^{\text{теор}}$	1	3	11	15	18	24

Принять $\alpha = 0{,}05$.

Изучение зависимостей

ГЛАВА 10

Основы дисперсионного анализа и реализация его моделей в Microsoft Excel

Дисперсионный анализ используют при изучении влияния на случайную величину Y некоторых интересующих исследователя факторов, обычно не поддающихся количественному измерению, а также оценке степени влияния (если оно существует). Суть метода состоит в выделении из общей вариации наблюдаемых значений величины Y (относительно среднего этих значений) частей, соответствующих раздельному и совместному влиянию факторов, и статистическом изучении таких частей для выяснения приемлемости гипотез о существовании этих влияний. Модели дисперсионного анализа в зависимости от числа факторов классифицируют на однофакторные, двухфакторные и т. д. В главе будут рассмотрены однофакторные и двухфакторные модели (с повторениями и без повторений) и их реализации в Microsoft Excel. При этом мы ограничимся *детерминированными* вариантами моделей, в которых значения (уровни) факторов, влияние которых на случайную величину Y изучается, определены исследователем, детерминированы, а не представляют собой случайную выборку из множества всех возможных значений того или иного фактора.

§ 10.1. Однофакторный дисперсионный анализ

1. Задача однофакторного дисперсионного анализа и предварительная обработка результатов наблюдений. Допустим, что экономиста строительно-монтажного управления интересует, зависит или не зависит объем выполненных на стройке работ за смену от работающей на стройке бригады. Предположим, что на стройке могут работать ν бригад. Назовем объем выполненных работ *результативным признаком*, обозначим его через Y и, поскольку объем выполненных работ зависит от ряда случайных обстоятельств, будем полагать, что Y — случайная величина. Работающую бригаду назовем *фактором* A, а номер работающей бригады — *уровнем, группой* или *значением фактора* A и обозначим его через $A^{(i)}$, $i = 1, 2, \ldots, \nu$.

Приступая к выяснению интересующей нас зависимости, необходимо над каждой бригадой провести наблюдения. Обратим внимание на то, что объем работ, выполненных бригадой, зависит не только от бригады, но и от других причин. Поэтому по каждой бригаде будет наблюдаться вариация, изменчивость ежедневного объема выполненных работ. Результаты наблюдений и результаты их предварительной обработки поместим в таблицу 10.1. В этой таблице: ν — число уровней фактора A; i — текущий номер уровня фактора A, $i = 1, 2, \ldots, \nu$; n_i — число наблюдений признака Y при i-м уровне фактора A; n — общее число наблюдений, $n = n_1 + n_2 + \ldots + n_\nu$; $y_k^{(i)}$ — значение признака Y, зафиксированное в k-м наблюдении при уровне $A^{(i)}$, $k = 1, 2, \ldots, n_i$; $\bar{y}^{(i)}$ — среднее результатов наблюдений признака Y при уровне $A^{(i)}$ (*групповое среднее*),

$$\bar{y}^{(i)} = (y_1^{(i)} + y_2^{(i)} + \ldots + y_{n_i}^{(i)})/n_i = \sum_{k=1}^{n_i} y_k^{(i)}/n_i; \qquad (10.1)$$

$\bar{y}$ — *общее среднее*, вычисляемое либо непосредственно по результатам наблюдений

$$\bar{y} = (y_1^{(1)} + y_2^{(1)} + \ldots + y_{n_1}^{(1)} + y_1^{(2)} + \ldots + y_{n_\nu}^{(\nu)})/n,$$

либо с использованием групповых средних

$$\bar{y} = (\bar{y}^{(1)} n_1 + \bar{y}^{(2)} n_2 + \ldots + \bar{y}^{(\nu)} n_\nu)/n = \sum_{i=1}^{\nu} \bar{y}^{(i)} n_i/n; \qquad (10.2)$$

$\hat{\sigma}_i^2$ — дисперсия результатов наблюдений признака Y при уровне $A^{(i)}$ (*выборочная групповая дисперсия*)

$$\hat{\sigma}_i^2 = \frac{(y_1^{(i)} - \bar{y}^{(i)})^2 + (y_2^{(i)} - \bar{y}^{(i)})^2 + \ldots + (y_{n_i}^{(i)} - \bar{y}^{(i)})^2}{n_i} =$$

$$= \sum_{k=1}^{n_i} (y_k^{(i)} - \bar{y}^{(i)})^2/n_i. \qquad (10.3)$$

Таблица 10.1

i	1	2	...	ν
Уровень фактора A ($A^{(i)}$)	$A^{(1)}$	$A^{(2)}$	...	$A^{(\nu)}$
$y_k^{(i)}$	$y_1^{(1)}$ $y_2^{(1)}$ ... $y_{n_1}^{(1)}$	$y_1^{(2)}$ $y_2^{(2)}$ ... $y_{n_2}^{(2)}$	... ...	$y_1^{(\nu)}$ $y_2^{(\nu)}$ ... $y_{n_\nu}^{(\nu)}$

Окончание табл.

i	1	2	...	ν
Число наблюдений в группе, n_i	n_1	n_2	...	n_ν
Групповое среднее, $\bar{y}_{(i)}$	$\bar{y}^{(1)}$	$\bar{y}^{(2)}$	...	$\bar{y}^{(\nu)}$
Выборочная групповая дисперсия, $\hat{\sigma}_i^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	...	$\hat{\sigma}_\nu^2$

Предварительный вывод о том, зависит ли Y от фактора A (объем выполненных работ от работающей бригады), можно сделать, сравнив групповые средние: если различие между ними существенно, то, по-видимому, такая зависимость имеется. Однако не надо забывать, что Y — случайная величина, следовательно, результаты ее наблюдений и групповые средние этих результатов случайны. Поэтому ответ на вопрос, существует ли зависимость результата от фактора, можно дать только сравнив *генеральные групповые средние*, а не выборочные, или, иначе, сравнив *групповые математические ожидания* $M(Y \mid A^{(i)})$, $i = 1, 2, \ldots, \nu$, — математические ожидания случайной величины Y, вычисленные при условии, что фактор A зафиксирован на уровне $A^{(1)}, A^{(2)}, \ldots, A^{(\nu)}$. Поскольку значения этих математических ожиданий неизвестны, возникает задача проверки гипотезы

$$H_0: M(Y \mid A^{(1)}) = M(Y \mid A^{(2)}) = \ldots = M(Y \mid A^{(\nu)}). \quad (10.4)$$

Если эту гипотезу не отклоняют, то считают, что фактор A (в рамках заданных исследователем уровней фактора) не влияет на Y; в противном случае влияние, скорее всего, имеет место.

Прежде чем привести критерии проверки гипотезы (10.4), рассмотрим вероятностную модель формирования случайных результатов наблюдений величины Y, составляющую основу однофакторного дисперсионного анализа.

2. Модель однофакторного дисперсионного анализа. Предположим, что случайный результат $Y_k^{(i)}$ k-го наблюдения величины Y при уровне $A^{(i)}$ фактора A формируется следующим образом:

$$Y_k^{(i)} = MY + \underbrace{(M(Y \mid A^{(i)}) - MY)}_{\theta^{(i)}} + \varepsilon_k^{(i)},$$

$$i = 1, 2, \ldots, \nu; \; k = 1, 2, \ldots, n_i, \quad (10.5)$$

где MY — генеральное среднее величины Y, постоянная величина (не зависящая ни от $A^{(i)}$, ни от k); $\theta^{(i)} = (M(Y \mid A^{(i)}) - MY)$ — постоянная при каждом $i = 1, 2, \ldots, \nu$ величина (зависящая от $A^{(i)}$), называемая *эффектом влияния на* $Y_k^{(i)}$ *уровня* $A^{(i)}$ (чем больше отличие генерального группового среднего $M(Y \mid A^{(i)})$ от общего среднего MY, тем сильнее, эффективнее влияние соответствующего уровня $A^{(i)}$ фактора A на результаты наблюдений); $\varepsilon_k^{(i)}$ — случайная при каждом $i = 1, 2, \ldots, \nu$ и $k = 1, 2, \ldots, n_i$ величина, отражающая влияние на результат наблюдения неконтролируемых случайных, или *остаточных факторов* (число таких величин равно $n = n_1 + n_2 + \ldots + n_\nu$). При этом предполагается, что:

$$\left.\begin{array}{l} \text{случайные величины } \varepsilon_k^{(i)} \text{ независимы;} \\ \text{каждая из величин } \varepsilon_k^{(i)} \text{ — нормально распределенная величина с нулевым математическим ожиданием и дисперсией } \sigma^2_{\text{ост}}, \text{ не зависящей ни от } A^{(i)}, \text{ ни от } k, \text{ т. е. одинаковой при всех } A^{(i)} \text{ и } k: \\ \qquad \varepsilon_k^{(i)} = N(M\varepsilon_k^{(i)} = 0;\ D\varepsilon_k^{(i)} = \sigma^2_{\text{ост}}), \\ \qquad i = 1, 2, \ldots, \nu;\ k = 1, 2, \ldots, n_i. \end{array}\right\} \quad (10.6)$$

Входящие в модель (10.6) постоянные величины MY, $\theta^{(i)} = M(Y \mid A^{(i)}) - MY$, $i = 1, 2, \ldots, \nu$, и $\sigma^2_{\text{ост}}$ называют параметрами модели однофакторного дисперсионного анализа.

Замечание. Выборочными аналогами MY и $M(Y \mid A^{(i)})$ являются соответственно общее среднее $\bar{y}$ и групповое среднее $\bar{y}^{(i)}$, поэтому при фиксированных числах n_i, $i = 1, 2, \ldots, \nu$, наблюдений в группах между MY и групповыми математическими ожиданиями $M(Y \mid A^{(i)})$ имеет место соотношение, подобное (10.2),

$$MY = \sum_{i=1}^{\nu} M(Y \mid A^{(i)}) n_i / n. \quad (10.7)$$

Учитывая, что $n = \sum_{i=1}^{\nu} n_i$, проведем тождественные преобразования равенства (10.7):

$$MY \sum_{i=1}^{\nu} n_i = \sum_{i=1}^{\nu} M(Y \mid A^{(i)}) n_i,\ \sum_{i=1}^{\nu} MY n_i = \sum_{i=1}^{\nu} M(Y \mid A^{(i)}) n_i,$$

$$\sum_{i=1}^{\nu} (M(Y \mid A^{(i)}) - MY) n_i = 0,$$

или

$$\sum_{i=1}^{v} \theta^{(i)} n_i = 0. \qquad (10.8)$$

Допустим, что все групповые математические ожидания одинаковы и равны числу c: $M(Y|A^{(1)}) = M(Y|A^{(2)}) = \ldots = M(Y|A^{(v)}) = c$, т. е. выполняется гипотеза (10.4). Тогда, учитывая (10.7), получим, что

$$MY = \sum_{i=1}^{v} cn_i/n = c \sum_{i=1}^{v} n_i/n = cn/n = c$$

и

$$\theta^{(i)} = M(Y|A^{(i)}) - MY = c - c = 0, \; i = 1, 2, \ldots, v.$$

Верно и обратное утверждение: если все $\Theta^{(i)} = 0$, то все групповые математические ожидания одинаковы и равны MY. Таким образом, гипотеза (10.4) тождественна гипотезе

$$H_0: M(Y|A^{(1)}) = M(Y|A^{(2)}) = \ldots = M(Y|A^{(v)}) = MY \qquad (10.9)$$

и гипотезе

$$H_0: \theta^{(1)} = \theta^{(2)} = \ldots = \theta^{(v)} = 0. \qquad (10.10)$$

Если учесть, что в модели (10.5) MY и (при каждом i) $M(Y|A^{(i)})$ — постоянные величины, то нетрудно убедиться, что условия (10.6) тождественны следующим условиям:

— случайные результаты $Y_k^{(i)}$, $i = 1, 2, \ldots, v$; $k = 1, 2, \ldots, n_i$ наблюдений — независимые случайные величины; } (10.11)

— для каждой из величин $Y_k^{(i)}$ имеет место соотношение

$$Y_k^{(i)} = N(MY_k^{(i)} = M(Y|A^{(i)}); DY_k^{(i)} = \sigma^2_{\text{ост}}), \qquad (10.12)$$

$$i = 1, 2, .., v; \quad k = 1, 2, \ldots, n_i.$$

Условие (10.12) означает, что: в пределах каждой из групп закон распределения случайных результатов $Y_k^{(i)}$ наблюдений не меняется, т. е. *в пределах каждой из групп наблюдения проводятся в типичных условиях*, и, более того, *закон распределения величин* $Y_k^{(i)}$ — *нормальный*; и при переходе от одной группы к другой *дисперсии величин* $Y_k^{(i)}$ *остаются постоянными*, равными $\sigma^2_{\text{ост}}$.

Выполнение требований независимости наблюдений и типичности (в пределах каждой из групп) условий их проведения определяется организацией эксперимента. При решении вопроса о виде закона распределения результатив-

ного признака при каждом уровне фактора, во-первых, учитывается природа изучаемого явления: возможно, что механизм формирования значений результативного признака удовлетворяет условиям центральной предельной теоремы и тогда этот признак имеет нормальный закон распределения, и, во-вторых, если в группе достаточно большое число данных, то, используя критерий согласия Пирсона, можно выяснить, приемлема гипотеза о нормальности распределения или нет.

Покажем, как установить, одинаковы ли генеральные групповые дисперсии результативного признака. Обозначим через σ_i^2 генеральную дисперсию результативного признака при i-м уровне фактора, $i = 1, 2, \ldots, \nu$. Не зная числовых значений этих дисперсий, нельзя однозначно сказать, равны дисперсии или нет, можно лишь проверить гипотезу

$$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_\nu^2. \qquad (10.13)$$

Эта гипотезу проверяют с помощью критерия Бартлетта.

1) Находят несмещенные оценки s_i^2 групповых дисперсий σ_i^2 по формуле

$$s_i^2 = \hat{\sigma}_i^2 n_i/(n_i - 1), \qquad i = 1, 2, \ldots, \nu, \qquad (10.14)$$

где $\hat{\sigma}_i^2$ — выборочные групповые дисперсии, определенные по формуле (10.3); n_i — численность наблюдений в группах.

2) Вычисляют величину

$$s_{\text{ост}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_\nu - 1)s_\nu^2}{(n_1 - 1) + (n_2 - 1) + \ldots + (n_\nu - 1)}, \qquad (10.15)$$

которая при выполнении гипотезы (10.13) является несмещенной оценкой любой из равных друг другу генеральных групповых дисперсий, или, иначе, несмещенной оценкой дисперсии $\sigma_{\text{ост}}^2$ в равенстве (10.12).

3) Определяют

$$q = \left[1 + \frac{1}{3(\nu - 1)}\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} + \ldots \right.\right.$$
$$\left.\left. \ldots \frac{1}{n_\nu - 1} - \frac{1}{(n_1 - 1) + \ldots + (n_\nu - 1)}\right)\right]^{-1} \qquad (10.16)$$

и величину

$$\chi^2 = q[(n_1 - 1)\ln(s_{\text{ост}}^2/s_1^2) + (n_2 - 1)\ln(s_{\text{ост}}^2/s_2^2) + \ldots$$
$$\ldots + (n_\nu - 1)\ln(s_{\text{ост}}^2/s_\nu^2)]. \qquad (10.17)$$

При выполнении гипотезы (10.13) и условия $n_i > 3$, $i = 1, 2, \ldots, \nu$, величина (10.17), рассматриваемая (при фиксированных числах ν, n_1, n_2, ..., n_ν) как случайная величина (ее случайность объясняется случайностью результатов наблюдений и как следствие случайностью оценок $s_1^2, s_2^2, \ldots, s_\nu^2$), имеет при достаточно большом числе наблюдений $n = n_1 + n_2 + \ldots + n_\nu$ распределение, близкое к χ^2-распределению с $k = \nu - 1$ степенями свободы.

4) Находят (по таблице П. 2) левостороннюю и правостороннюю критические точки: $\chi^2_{\nu-1,1-\alpha/2}$ и $\chi^2_{\nu-1,\alpha/2}$. Если найденное по формуле (10.17) число χ^2 попадает в критическую область, т. е. $\chi^2 < \chi^2_{\nu-1,1-\alpha/2}$ или $\chi^2 > \chi^2_{\nu-1,\alpha/2}$, то гипотезу (10.13) о равенстве генеральных групповых дисперсий отклоняют; если $\chi^2_{\nu-1,1-\alpha/2} < \chi^2 < \chi^2_{\nu-1,\alpha/2}$, то гипотезу (10.13) не отклоняют.

Допустим, что условия (10.6), или тождественные им условия (10.11) и (10.12) выполняются. Тогда можно приступить к проверке гипотезы (10.4), или тождественной ей гипотезы (10.9), или (10.10) однофакторного дисперсионного анализа. Проверка основана на разложении общей вариации наблюдаемых значений величины Y относительно их среднего на две составляющие: вариацию, причиной которой является изменение уровней фактора A, и вариацию, обусловленную влиянием на наблюдения неконтролируемых случайных, остаточных факторов.

3. Разложение вариации наблюдаемых значений величины *Y*. Выборочный, статистический аналог вероятностной модели (10.5) имеет следующий вид:

$$y_k^{(i)} = \bar{y} + \underbrace{(\bar{y}^{(i)} - \bar{y})}_{\hat{\theta}^{(i)}} + (y_k^{(i)} - \bar{y}^{(i)}), \qquad i = 1, 2, \ldots, \nu; \; k = 1, 2, \ldots, n_i. \tag{10.18}$$

Здесь $y_k^{(i)}$ — конкретное значение случайной величины $Y_k^{(i)}$, зафиксированное в наблюдении; $\bar{y}$ — выборочный аналог MY, общее среднее, которое находят с помощью формулы (10.2); $\bar{y}^{(i)}$ — выборочный аналог $M(Y \mid A^{(i)})$, групповое среднее, вычисляемое по формуле(10.1); $\hat{\theta}^{(i)} = \bar{y}^{(i)} - \bar{y}$ — выборочный аналог вероятностного (генерального) эффекта влияния на наблюдения уровня $A^{(i)}$, равного $\theta^{(i)} = M(Y \mid A^{(i)}) - MY$ (обратим внимание, что выборочным аналогом равенства

(10.8) является равенство $\sum_{i=1}^{v} \hat{\theta}^{(i)} n_i = 0$); $(y_k^{(i)} - \bar{y}^{(i)})$ — отклонение k-го результата наблюдения в i-й группе от среднего этой группы, объясняемое, при каждом фиксированном i, влиянием на результат наблюдения неконтролируемых случайных, или остаточных факторов.

Соотношение (10.18) представляет собой тождество (после приведения подобных получим $y_k^{(i)} = y_k^{(i)}$); запишем его в виде

$$y_k^{(i)} - \bar{y} = (\bar{y}^{(i)} - \bar{y}) + (y_k^{(i)} - \bar{y}^{(i)}). \qquad (10.19)$$

Докажем следующее тождество:

$$\underbrace{\sum_{i=1}^{v} \sum_{k=1}^{n_i} (y_k^{(i)} - \bar{y})^2}_{SS_{\text{итого}}} =$$

$$= \underbrace{\sum_{i=1}^{v} (\bar{y}^{(i)} - \bar{y})^2 n_i}_{SS_{\text{между группами}}} + \underbrace{\sum_{i=1}^{v} \sum_{k=1}^{n_i} (y_k^{(i)} - \bar{y}^{(i)})^2}_{SS_{\text{внутри групп}}}. \qquad (10.20)$$

Замечание. Под составляющими (10.20) проставлены обозначения, используемые в программе **«Однофакторный дисперсионный анализ»** пакета **«Анализ данных»** Microsoft Excel (*SS* от англ. *sum of squares* — сумма квадратов). При использовании этой программы в рабочий лист надо ввести из таблицы 10.1 v столбцов результатов $\bar{y}_k^{(i)}$ наблюдений величины Y при различных уровнях фактора A.

≫ Возведем в квадрат обе части тождества (10.19). Имеем

$$(y_k^{(i)} - \bar{y})^2 =$$

$$= (\bar{y}^{(i)} - \bar{y})^2 + (y_k^{(i)} - \bar{y}^{(i)})^2 + 2(\bar{y}^{(i)} - \bar{y})(y_k^{(i)} - \bar{y}^{(i)})$$

и просуммируем обе части этого тождества по k от 1 до n_i. В результате получаем

$$\sum_{k=1}^{n_i} (y_k^{(i)} - \bar{y})^2 =$$

$$= \sum_{k=1}^{n_i} (\bar{y}^{(i)} - \bar{y})^2 + \sum_{k=1}^{n_i} (y_k^{(i)} - \bar{y}^{(i)})^2 + 2 \sum_{k=1}^{n_i} (\bar{y}^{(i)} - \bar{y})(y_k^{(i)} - \bar{y}^{(i)}). \qquad (10.21)$$

Первое из слагаемых в (10.21) имеет вид

$$\sum_{k=1}^{n_i} (\bar{y}^{(i)} - \bar{y})^2 = (\bar{y}^{(i)} - \bar{y})^2 \sum_{k=1}^{n_i} 1 = (\bar{y}^{(i)} - \bar{y})^2 n_i,$$

так как выражение $(\bar{y}^{(i)} - \bar{y})^2$, стоящее под знаком суммы, от k **не зависит**. Вычислим последнее слагаемое

$$2\sum_{k=1}^{n_i}(\bar{y}^{(i)}-\bar{y})(y_k^{(i)}-\bar{y}^{(i)})=2(\bar{y}^{(i)}-\bar{y})\sum_{k=1}^{n_i}(y_k^{(i)}-\bar{y}^{(i)})=$$

$$=2(\bar{y}^{(i)}-\bar{y})\Big(\sum_{k=1}^{n_i}y_k^{(i)}-\sum_{k=1}^{n_i}\bar{y}^{(i)}\Big)=2(\bar{y}^{(i)}-\bar{y})\Big(\sum_{k=1}^{n_i}y_k^{(i)}-\bar{y}^{(i)}n_i\Big)=$$

$$=2(\bar{y}^{(i)}-\bar{y})(n_i\bar{y}^{(i)}-\bar{y}^{(i)}n_i)=0.$$

Таким образом, тождество (10.21) принимает вид

$$\sum_{k=1}^{n_i}(y_k^{(i)}-\bar{y})^2=(\bar{y}^{(i)}-\bar{y})^2n_i+\sum_{k=1}^{n_i}(y_k^{(i)}-\bar{y}^{(i)})^2.$$

Просуммировав обе части этого тождества по i от 1 до ν, получим (10.20). ≪

Из тождества (10.20) видно, что общая вариация наблюдаемых значений $y_k^{(i)}$ величины Y относительно общего среднего (она измеряется величиной $SS_{\text{итого}}$) раскладывается на две составляющие: на вариацию групповых средних $\bar{y}^{(i)}$ около $\bar{y}$, обусловленную изменением уровней фактора A (она измеряется величиной $SS_{\text{между группами}}$) и вариацию наблюдаемых значений $y_k^{(i)}$ около соответствующих или групповых средних $\bar{y}^{(i)}$, обусловленную влиянием на наблюдения внутри каждой группы остаточных факторов (она измеряется величиной $SS_{\text{внутри групп}}$).

4. Таблица однофакторного дисперсионного анализа и проверка гипотезы об отсутствии влияния фактора *A* на результативный признак *Y*. В общем случае таблица однофакторного дисперсионного анализа имеет вид таблицы 10.2, в которой наряду с принятыми в отечественной литературе обозначениями и пояснениями приведены обозначения, используемые в программе **«Однофакторный дисперсионный анализ»** пакета **«Анализ данных»** Microsoft Excel.

Поясним таблицу 10.2, считая выполненными условия (10.6), или (10.11) и (10.12).

≫ В столбце (3) приведены степени свободы (*df* от англ. *degree of freedom* — степени свободы); формальный способ их подсчета такой:

— $SS_{\text{между группами}}=\sum_{i=1}^{\nu}(\bar{y}^{(i)}-\bar{y})^2n_i$ вычисляют, исходя из отклонений ν независимых групповых средних $\bar{y}^{(i)}$ от общего среднего $\bar{y}$, поэтому $df=\nu-1$;

— $SS_{\text{внутри групп}}=\sum_{i=1}^{\nu}\sum_{k=1}^{n_i}(y_k^{(i)}-\bar{y}^{(i)})^2$ вычисляют, исходя из отклонений n независимых результатов $y_k^{(i)}$ наблюдений от ν групповых средних, $df=n-\nu$;

Таблица 10.2

Источник вариации значений величины Y	Измеритель вариации значений величины Y, (SS) [см. (10.20)]	Степень свободы, df	Несмещенная оценка дисперсии $\sigma^2_{\text{ост}}$, $(MS = SS/df)$	F	P-значение	$f^{\text{кр}}$, $(F_{\text{критическое}})$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Изменение уровней фактора A (между группами)	SS_A $(SS_{\text{между группами}})$	$\nu - 1$	$s^2_A = SS_A/(\nu - 1)$ (при выполнении гипотезы (10.4))	$s^2_A / s^2_{\text{ост}}$	$P(F(\nu - 1, n - \nu) > s^2_A / s^2_{\text{ост}})$	$f_{\nu - 1,\, n - \nu,\, \alpha}$
Изменение значений прочих (помимо A), остаточных факторов (внутри групп)	$SS_{\text{ост}}$ $(SS_{\text{внутри групп}})$	$n - \nu$	$s^2_{\text{ост}} = SS_{\text{ост}}/(n - \nu)$			
Общая вариация (итого)	$SS_{\text{итого}}$	$n - 1$				

— $SS_{\text{итого}} = \sum_{i=1}^{\nu} \sum_{k=1}^{n_i} (y_k^{(i)} - \bar{y})^2$ вычисляют исходя из отклонений n независимых результатов $y_k^{(i)}$ наблюдений от общего среднего $\bar{y}$, $df = n - 1$.

Обратим внимание на то, что $(\nu - 1) + (n - \nu) = n - 1$.

В столбце (4) даны значения MS (MS от англ. *mean of squares* — среднее квадратов), а именно:

— $s_A^2 = SS_A/(\nu - 1)$ — несмещенная оценка дисперсии $\sigma_{\text{ост}}^2$ в случае, когда гипотеза (10.4), или (10.10) дисперсионного анализа выполняется;

— $s_{\text{ост}}^2 = SS_{\text{ост}}/(n - \nu)$ — несмещенная оценка дисперсии $\sigma_{\text{ост}}^2$ вне зависимости от того, выполняется или нет гипотеза (10.4), или (10.10) (можно доказать, что приведенная формула вычисления $s_{\text{ост}}^2$ и формула (10.15), тождественны).

В столбце (5) приведено F-отношение, равное $s_A^2 / s_{\text{ост}}^2$, которое, если его интерпретировать как случайную (при фиксированных числах ν, n_1, n_2, ... , n_ν) величину, имеет при выполнении гипотезы (10.4), или (10.9) дисперсионного анализа F-распределение с числами степеней свободы $k_1 = \nu - 1$ и $k_2 = n - \nu$, т. е. $s_A^2 / s_{\text{ост}}^2 = F(\nu - 1, n - \nu)$.

В столбце (6) приведен рассчитанный уровень значимости — вероятность того, что случайная величина $F(\nu - 1, n - \nu)$ превысит число, равное $s_A^2 / s_{\text{ост}}^2$.

В столбце (7) приведена правосторонняя критическая точка $f^{\text{кр}} = f_{\nu-1, n-\nu, \alpha}$, где α — заданный уровень значимости (при ручном счете $f^{\text{кр}}$ находят по таблице П. 5 при $k_1 = \nu - 1$ и $k_2 = n - \nu$ и $p = \alpha$). ≪

Используя числа, стоящие в столбцах (6), (7) таблицы 10.2, можно решить вопрос о том, отклоняется гипотеза (10.4) о равенстве генеральных групповых средних или не отклоняется.

Если P-значение $< \alpha$, или если число $s_A^2 / s_{\text{ост}}^2 > f^{\text{кр}}$, гипотезу (10.4) о равенстве генеральных групповых средних, или, иначе, об отсутствии влияния фактора (в рамках заданных его уровней) на признак Y, отклоняют. В этом случае вычисляют *коэффициент детерминации* результатов наблюдений случайной величины Y фактором A, уровни которого заданы исследователем,

$$\hat{\eta}^2_{Y|A} = SS_A/SS_{\text{итого}},$$

показывающий, какую долю от общей вариации наблюдаемых значений результативного признака Y составляет вариация этих значений, обусловленная изменением уровней фактора A, или, короче, долю вариации наблюдаемых «игреков», обусловленную фактором A.

Если P-значение $> \alpha$, или если число $s_A^2 / s_{\text{ост}}^2 < f^{\text{кр}}$, гипотезу (10.4) не отклоняют: фактор A не влияет на признак Y; в этом случае оценивание силы влияния лишено смысла.

▶ **ПРИМЕР 10.1.** При уровне значимости $\alpha = 0{,}05$ выясним, зависит или нет объем выполненных на стройке работ за смену от работающей бригады. Данные по четырем бригадам приведены в таблице 10.3 (там же содержатся результаты предварительных вычислений).

Таблица 10.3

Номер бригады, i	1	2	3	4 = v	
Объем выполненных за смену работ	140 144 142 145	150 149 152 150	148 149 146 147	150 155 154 152	
Число смен, n_i	4	4	4	4	$n = 16$
Групповое среднее, $\bar{y}^{(i)}$	142,75	150,25	147,50	152,75	
Групповая дисперсия, $\hat{\sigma}_i^2$	3,69	1,19	1,25	3,69	
$\hat{s}_i^2 = (\hat{\sigma}_i^2 n_i)/(n_i - 1)$	4,92	1,58	1,67	4,92	

Групповые средние $\bar{y}^{(i)}$ и исправленные групповые дисперсии s_i^2 содержатся в результатах работы программы **«Однофакторный дисперсионный анализ»**, приведенных на рисунке 10.1.

Однофакторный дисперсионный анализ

Итоги

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
Столбец 1	4	571	142,75	4,916667
Столбец 2	4	601	150,25	1,583333
Столбец 3	4	590	147,5	1,666667
Столбец 4	4	611	152,75	4,916667

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	220,1875	3	73,39583	22,43949	3,28E-05	3,4903
Внутри групп	39,25	12	3,270833			
Итого	259,4375	15				

Рис. 10.1

В рассматриваемом примере результативный признак Y — объем работ, выполненных за смену; фактор A — работающая бригада, зафиксировано $v = 4$ уровня этого фактора, и на каждом уровне проведено четыре наблюдения: $n_1 = n_2 = n_3 = n_4 = 4$, всего наблюдений $n = 16$.

Модель (10.5) принимает вид

$$Y_k^{(i)} = MY + \underbrace{(M(Y \mid A^{(i)}) - MY)}_{\theta^{(i)}} + \varepsilon_k^{(i)}, \qquad i = k = 1, 2, 3, 4,$$

где $Y_k^{(i)}$ — случайная величина, объем работ, выполненных i-й бригадой в k-ю смену; $\theta^{(i)}$ — эффект влияния на этот объем работ i-й бригады, $\sum_{i=1}^{4} n_i \theta^{(i)} = 0$; $\varepsilon_k^{(i)}$ — случайная величина, отражающая влияние на объем выполненных работ неконтролируемых случайных, или остаточных, факторов.

Выясним, выполняются ли для величин $Y_k^{(i)}$ условия (10.11) и (10.12). Допустим, что:

— независимость величин $Y_k^{(i)}$ гарантируется организацией наблюдений;

— условия работы каждой из бригад в каждую из смен типичны, т. е. в основном одинаковы (имеется строительный материал, не меняются погодные условия, нет чрезвычайных происшествий и т. д.);

— каждая из величин $Y_k^{(i)}$ имеет нормальный закон распределения.

Проверим гипотезу H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ о равенстве групповых генеральных дисперсий при $\alpha = 0{,}05$. Для этого:

1) По формуле (10.14) найдем оценки s_i^2, $i = 1, 2, 3, 4$ (они приведены в последней строке таблицы 10.3).

2) По формуле (10.15) рассчитаем

$$s_{\text{ост}}^2 = (3 \cdot 4{,}92 + 3 \cdot 1{,}58 + 3 \cdot 1{,}67 + 3 \cdot 4{,}92)/(3 + 3 + 3 + 3) = 3{,}27.$$

3) По формулам (10.16) и (10.17) определим

$$q = \left[1 + \frac{1}{3(4-1)}\left(\frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} - \frac{1}{3+3+3+3}\right)\right]^{-1} = 0{,}878$$

и

$$\chi^2 = 0{,}878\,[(3 \ln 3{,}27/4{,}92 + 3 \ln (3{,}27/1{,}58) + 3 \ln (3{,}27/1{,}67) + 3 \ln (3{,}27/4{,}92)] = 1{,}538.$$

4) По таблице П. 2 при $k = \nu - 1 = 3$ и вероятностях $p = 1 - \alpha/2 = 0{,}975$ и $p = \alpha/2 = 0{,}025$ найдем критические точки: $\chi^2_{3;0,975} = 0{,}22$ и $\chi^2_{3;0,025} = 9{,}35$. Так как $\chi^2 = 1{,}538 \in (0{,}22;\ 9{,}35)$, то гипотезу о равенстве групповых генеральных дисперсий не отклоняем.

Теперь приступим к собственно дисперсионному анализу, т. е. к проверке гипотезы (10.4) о равенстве четырех генеральных средних объемов работ, выполненных за смену каждой из бригад, или, иначе, к проверке гипотезы (10.10) о равенстве нулю четырех эффектов влияния на объем работ работающей бригады. Используя формулу (10.2), вычислим

$$\bar{y} = (142{,}75 \cdot 4 + 150{,}25 \cdot 4 + 147{,}50 \cdot 4 + 152{,}75 \cdot 4)/16 = 148{,}31.$$

Далее в соответствии с формулами, приведенными в тождестве (10.20), найдем

$$SS_{\text{итого}} = \underbrace{(140 - 148{,}31)^2 + (144 - 148{,}31)^2 + \ldots + (152 - 148{,}31)^2}_{\text{16 слагаемых}} = 259{,}44;$$

$$SS_{\text{между группами}} = SS_A = \underbrace{4(142{,}75 - 148{,}31)^2 + \ldots + 4(152{,}75 - 148{,}31)^2}_{\text{4 слагаемых}} = 220{,}19;$$

$$SS_{\text{внутри групп}} = SS_{\text{ост}} = \underbrace{\begin{matrix}(140 - 142{,}75)^2 + (144 - 142{,}75)^2 + .. + (145 - 142{,}75)^2 + \\ + (150 - 150{,}25)^2 + \ldots + (152 - 152{,}75)^2\end{matrix}}_{\text{16 слагаемых}} = 39{,}25.$$

Полученные результаты приведены в столбце «SS» таблицы «Дисперсионный анализ» (см. рисунок 10.1). Сопоставив ее с таблицей 10.2 делаем вывод, что:

— в столбце «df» приведены степени свободы $\nu - 1 = 4 - 1 = 3$, $n - \nu = 16 - 4 = 12$ и $n - 1 = 16 - 1 = 15$;

— в столбце «MS» приведены $s_A^2 = SS_A/(\nu - 1) = 220{,}19/3 = 73{,}39$ и $s^2_{\text{ост}} = SS_{\text{ост}}/(n - \nu) = 39{,}25/12 = 3{,}27$;

— в столбце «F» стоит число $s_A^2/s^2_{\text{ост}} = 22{,}44$;

— в столбце «P-значение» приведена вероятность $P(F(\nu - 1;\ n - \nu) > s_A^2/s^2_{\text{ост}}) = P(F(3;\ 12) > 22{,}44) = 3{,}28 \cdot 10^{-5}$ (най-

дите эту вероятность, используя **Статистическую функцию** FРАСП(22,44; 3; 12));

— в столбце «$F_{\text{критическое}}$» приведена критическая точка $f^{\text{кр}} = f_{\nu-1,\, n-\nu,\, \alpha} = f_{3;\,12;\,0{,}05} = 3{,}4903$ (найдите ее, используя **Функцию** FРАСПОБР(0,05; 3; 12) или таблицу П. 5).

Так как P-значение $= 3{,}28 \cdot 10^{-5} < \alpha = 0{,}05$ (или так как $F = 22{,}44 > F_{\text{критического}} = 3{,}4903$, то гипотезу H_0: $M(Y \mid A^{(1)}) = M(Y \mid A^{(2)}) = M(Y \mid A^{(3)}) = M(Y \mid A^{(4)})$ о равенстве генеральных средних объемов работ, выполненных каждой из бригад, или, иначе, об отсутствии влияния работающей бригады на объем выполненных работ, отклоняем. Силу этого влияния измерим коэффициентом детерминации:

$$\hat{\eta}^2_{Y|A} = SS_A/SS_{\text{итого}} = 220{,}19/259{,}44 = 0{,}849,$$

84,9% вариации наблюдаемых значений объема выполненных работ связано с изменением состава работающей бригады, а 15,1% этой вариации связано с действием неконтролируемых случайных, или остаточных, факторов. ◄

§ 10.2. Двухфакторный дисперсионный анализ с повторениями

Двухфакторный дисперсионный анализ с повторениями предназначен для изучения влияния на результативный признак — случайную величину Y — двух факторов: A, имеющего ν_A уровней $A^{(1)}, A^{(2)}, \ldots, A^{(\nu_A)}$ и B, имеющего ν_B уровней $B^{(1)}, B^{(2)}, \ldots, B^{(\nu_B)}$, а также взаимодействия этих факторов.

Исходными данными метода являются результаты наблюдений случайной величины Y, проведенные при различных комбинациях уровней факторов (число таких комбинаций равно $\nu_A\nu_B$), причем при каждой комбинации проводится $m > 1$ повторных наблюдений величины Y. Общее число наблюдений $n = \nu_A\nu_B m$. Результаты наблюдений записывают в виде таблицы 10.4.

При использовании программы **«Двухфакторный дисперсионный анализ с повторениями»** пакета **«Анализ данных»** Microsoft Excel надо иметь в виду следующее:

1) при каждой комбинации уровней факторов m результатов наблюдений величины Y записываются в рабочий лист Microsoft Excel как столбцы, содержащие m строк; поэтому числовая информация в рабочем листе занимает $\nu_A m$ строк и ν_B столбцов;

2) имеющаяся в таблице 10.4 строка с именами уровней фактора B и столбец с именами уровней фактора A должны быть сохранены и в рабочем листе, но ввод имен не обязателен;

3) в результате вся информация, необходимая для работы программы, в рабочем листе занимает $(1 + \nu_A m)$ строк и $(1 + \nu_B)$ столбцов и «входной интервал» — это «интервал от * до *» (см. таблицу 10.4); «число строк для выборки» — это число m — количество наблюдений при каждой комбинации уровней.

Таблица 10.4

*	$B^{(1)}$	$B^{(2)}$	...	$B^{(\nu_B)}$
$A^{(1)}$	Результаты m наблюдений величины Y	Результаты m наблюдений величины Y	...	Результаты m наблюдений величины Y
$A^{(2)}$	Результаты m наблюдений величины Y	Результаты m наблюдений величины Y	...	Результаты m наблюдений величины Y
...	...	...	...	...
$A^{(\nu_A)}$	Результаты m наблюдений величины Y	Результаты m наблюдений величины Y	...	Результаты m наблюдений величины Y*

Не следует забывать, что Y — случайная величина, поэтому результаты m ее наблюдений, помещенные в каждой клетке таблицы 10.4, можно интерпретировать как случайные величины, которые при проведении конкретных наблюдений принимают конкретные числовые значения.

Прежде чем приводить модель двухфакторного дисперсионного анализа с повторениями, введем обозначения, которые будут использованы в дальнейшем (см. таблицу 10.5).

Модель двухфакторного дисперсионного анализа с повторениями имеет следующий вид:

$$Y_k^{(i,j)} = MY + \theta_A^{(i)} + \theta_B^{(j)} + \theta_{AB}^{(i,j)} + \varepsilon_k^{(i,j)},$$

$$i = 1, 2, \ldots, \nu_A, \quad j = 1, 2, \ldots, \nu_B, \quad k = 1, 2, \ldots, m, \quad (10.22)$$

где $Y_k^{(i,j)}$ — случайный результат k-го наблюдения величины Y при уровнях $A^{(i)}$ и $B^{(j)}$ факторов A и B;

$$\theta_A^{(i)} = M(Y \mid A^{(i)}) - MY$$

Таблица 10.5

Характеристика	Обозначение в генеральной совокупности	Обозначение в выборочной совокупности
Среднее результативного признака Y	MY	$\bar{y}$
Среднее признака Y при уровне $A^{(i)}$	$M(Y \mid A^{(i)})$	$\bar{y}_A^{(i)}$
Среднее признака Y при уровне $B^{(j)}$	$M(Y \mid B^{(j)})$	$\bar{y}_B^{(j)}$
Среднее признака Y при комбинации уровней $A^{(i)}$ и $B^{(j)}$	$M(Y \mid A^{(i)} \cap B^{(j)})$	$\bar{y}_{AB}^{(i,j)}$

— эффект влияния на $Y_k^{(i,j)}$ уровня $A^{(i)}$, при этом $\sum_{i=1}^{\nu_A} \theta_A^{(i)} = 0$;

$$\theta_B^{(j)} = M(Y \mid B^{(j)}) - MY$$

— эффект влияния на $Y_k^{(i,j)}$ уровня $B^{(j)}$, при этом $\sum_{j=1}^{\nu_B} \theta_B^{(j)} = 0$;

$$\theta_{AB}^{(i,j)} = [M(Y \mid A^{(i)} \cap B^{(j)}) - MY] - (\theta_A^{(i)} + \theta_B^{(j)}) =$$

$$= M(Y \mid A^{(i)} \cap B^{(j)}) - M(Y \mid A^{(i)}) - M(Y \mid B^{(j)}) + MY$$

— эффект влияния на $Y_k^{(i,j)}$ взаимодействия уровней $A^{(i)}$ и $B^{(j)}$, при этом при каждом j сумма $\sum_{i=1}^{\nu_A} \theta_{AB}^{(i,j)} = 0$ и при каждом i сумма $\sum_{j=1}^{\nu_B} \theta_{AB}^{(i,j)} = 0$;

$\varepsilon_k^{(i,j)}$ — случайная величина, отражающая влияние на $Y_k^{(i,j)}$ неконтролируемых случайных, или остаточных, факторов (число таких величин $n = \nu_A \nu_B m$). При этом предполагается, что:

$$\left.\begin{array}{l} \text{случайные величины } \varepsilon_k^{(i,j)} \text{ независимы;} \\ \text{каждая из величин } \varepsilon_k^{(i,j)} = N(M\varepsilon_k^{(i,j)} = 0, \\ D\varepsilon_k^{(i,j)} = \sigma_{\text{ост}}^2). \end{array}\right\} \qquad (10.23)$$

Если учесть, что в модели (10.22) MY, $\theta_A^{(i)}$, $\theta_B^{(j)}$ и $\theta_{AB}^{(i,j)}$ — постоянные (при каждом i и j) величины, то можно убедиться в том, что условия (10.23) тождественны следующим:

случайные величины $Y_k^{(i,j)}$ независимы; (10.24)

каждая из величин

$$Y_k^{(i,j)} = N(MY_k^{(i,j)} = M(Y \mid A^{(i)} \cap B^{(j)}), DY_k^{(i,j)} = \sigma^2_{\text{ост}}). \quad (10.25)$$

Условие (10.25) означает, что при каждой комбинации уровней факторов A и B закон распределения случайных результатов наблюдений не меняется, т. е. наблюдения проводятся в типичных условиях, и, более того, этот закон — нормальный; и при переходе от одной комбинации уровней к другой дисперсии случайных результатов наблюдений остаются постоянными (равными $\sigma^2_{\text{ост}}$).

Для модели (10.22), в предположении, что требования (10.23), или (10.24) и (10.25), выполнены, проверяются следующие три гипотезы:

$$H_A: M(Y \mid A^{(1)}) = M(Y \mid A^{(2)}) = \ldots = M(Y \mid A^{(\nu_A)}); \quad (10.26)$$

$$H_B: M(Y \mid B^{(1)}) = M(Y \mid B^{(2)}) = \ldots = M(Y \mid B^{(\nu_B)}); \quad (10.27)$$

$$H_{AB}: M(Y \mid A^{(1)} \cap B^{(1)}) = M(Y \mid A^{(1)} \cap B^{(2)}) = \ldots$$

$$\ldots = M(Y \mid A^{(\nu_A)} \cap B^{(\nu_B)}), \quad (10.28)$$

которые соответственно тождественны следующим гипотезам:

$$H_A: \theta_A^{(1)} = \theta_A^{(2)} = \ldots = \theta_A^{(\nu_A)} = 0; \quad (10.29)$$

$$H_B: \theta_B^{(1)} = \theta_B^{(2)} = \ldots = \theta_B^{(\nu_B)} = 0; \quad (10.30)$$

$$H_{AB}: \theta_{AB}^{(1,1)} = \theta_{AB}^{(1,2)} = \ldots = \theta_{AB}^{(\nu_A, \nu_B)} = 0, \quad (10.31)$$

утверждающим, что нулевыми являются эффекты влияния на признак Y всех уровней фактора A, всех уровней фактора B и всех взаимодействий этих уровней.

Проверка гипотез основана на разложении общей вариации $SS_{\text{итого}}$ наблюдаемых значений величины Y на четыре составляющие:

— вариацию SS_A, причиной которой является изменение уровней фактора A;

— вариацию SS_B, обусловленную изменениями уровней фактора B;

— вариацию SS_{AB}, связанную с изменением уровней фактора A и B в их взаимодействии;

— вариацию $SS_{\text{ост}}$, обусловленную влиянием на результат наблюдения неконтролируемых случайных, или остаточных, факторов.

Это разложение имеет вид

$$SS_{\text{итого}} = SS_A + SS_B + SS_{AB} + SS_{\text{ост}}, \qquad (10.32)$$

или, если использовать обозначения, принятые в программе «**Двухфакторный дисперсионный анализ с повторениями**», следующий вид:

$$SS_{\text{итого}} = SS_{\text{выборка}} + SS_{\text{столбцы}} + SS_{\text{взаимодействие}} + SS_{\text{внутри}}.$$

В этих тождествах

$$SS_{\text{итого}} = \sum_{i=1}^{\nu_A} \sum_{j=1}^{\nu_B} \sum_{k=1}^{m} (y_k^{(i,j)} - \bar{y})^2, \qquad (10.33)$$

$$SS_A = SS_{\text{выборка}} = \nu_B m \sum_{i=1}^{\nu_A} (\bar{y}_A^{(i)} - \bar{y})^2, \qquad (10.34)$$

$$SS_B = SS_{\text{столбцы}} = \nu_A m \sum_{j=1}^{\nu_B} (\bar{y}_B^{(j)} - \bar{y})^2, \qquad (10.35)$$

$$SS_{AB} = SS_{\text{взаимодействие}} = m \sum_{i=1}^{\nu_A} \sum_{j=1}^{\nu_B} (\bar{y}_{A,B}^{(i,j)} - \bar{y}_A^{(i)} - \bar{y}_B^{(j)} + \bar{y})^2, \qquad (10.36)$$

$$SS_{\text{ост}} = SS_{\text{внутри}} = \sum_{i=1}^{\nu_A} \sum_{j=1}^{\nu_B} \sum_{k=1}^{m} (y_k^{(i,j)} - \bar{y}_{A,B}^{(i,j)})^2. \qquad (10.37)$$

Приведем дисперсионную таблицу двухфакторного дисперсионного анализа с повторениями в общем случае (табл. 10.6). Поясним таблицу 10.6, считая условия (10.23), или (10.24) и (10.25) выполненными.

≫ Из приведенных в столбце (4) оценок дисперсии $\sigma^2_{\text{ост}}$ только $s^2_{\text{ост}}$ всегда несмещенная; остальные являются несмещенными оценками только при условиях, указанных в скобках. Каждая из величин, указанных в столбце (5), если ее интерпретировать как случайную, имеет при выполнении соответствующей гипотезы F-распределение, а именно:

$s_A^2 / s^2_{\text{ост}} = F(\nu_A - 1,\ n - \nu_A\nu_B)$, если выполняется гипотеза H_A [см. (10.26)];

$s_B^2 / s^2_{\text{ост}} = F(\nu_B - 1,\ n - \nu_A\nu_B)$, если выполняется гипотеза H_B [см. (10.27)];

$s_{AB}^2 / s^2_{\text{ост}} = F((\nu_A - 1)(\nu_B - 1),\ n - \nu_A\nu_B)$, если выполняется гипотеза H_{AB} [см. (10.28)].

Таблица 10.6

Источник вариации значений величины Y	Измеритель вариации значений величины Y, (SS) (см. (10.33) — (10.37)	Степень свободы, df	Несмещенная оценка дисперсии $\sigma^2_{\text{ост}}$, $(MS = SS/df)$	F	P-значение	$f^{\text{кр}}$, $(F_{\text{критическое}})$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Изменение уровней фактора A (выборка)	SS_A $(SS_{\text{выборка}})$	$\nu_A - 1$	$s^2_A = SS_A/(\nu_A - 1)$ (при выполнении гипотезы (10.26))	$F_A = s^2_A / s^2_{\text{ост}}$	P_A	$f^{\text{кр}}_A = f^{\text{кр}}_{\nu_A - 1,\, n - \nu_A\nu_B,\, \alpha}$
Изменение уровней фактора B (столбцы)	SS_B $(SS_{\text{столбцы}})$	$\nu_B - 1$	$s^2_B = SS_B/(\nu_B - 1)$ (при выполнении гипотезы (10.27))	$F_B = s^2_B / s^2_{\text{ост}}$	P_B	$f^{\text{кр}}_B = f^{\text{кр}}_{\nu_B - 1,\, n - \nu_A\nu_B,\, \alpha}$
Изменение комбинаций уровней факторов A и B (взаимодействие)	SS_{AB} $(SS_{\text{взаимодействие}})$	$\nu_A\nu_B - \nu_A - \nu_B + 1 = (\nu_A - 1)(\nu_B - 1)$	$s^2_{AB} = SS_{AB}/[(\nu_A - 1)(\nu_B - 1)]$ (при выполнении гипотезы (10.28))	$F_{AB} = s^2_{AB} / s^2_{\text{ост}}$	P_{AB}	$f^{\text{кр}}_{AB} = f^{\text{кр}}_{(\nu_A - 1)(\nu_B - 1),\, n - \nu_A\nu_B,\, \alpha}$
Изменение значений остаточных факторов (внутри)	$SS_{\text{ост}}$ $(SS_{\text{внутри}})$	$n - \nu_A\nu_B$	$s^2_{\text{ост}} = \dfrac{SS_{\text{ост}}}{n - \nu_A\nu_B}$			
Общая вариация (итого)	$SS_{\text{итого}}$	$n - 1$				

В столбце (6) приведены рассчитанные (при выполнении соответствующей гипотезы) уровни значимости — P-значения, а в столбце (7) — правосторонние критические точки. ≪

Проверка гипотез (10.26)—(10.28), или гипотез (10.29)—(10.31), проводится следующим образом.

Если $P_A < \alpha$, где α — заданный уровень значимости, или $F_A > f_A^{\text{кр}}$, то гипотезу H_A об отсутствии влияния фактора A на Y отклоняют и силу влияния измеряют выборочным коэффициентом детерминации

$$\hat{\eta}^2_{Y|A} = SS_A/SS_{\text{итого}}; \qquad (10.38)$$

при $P_A > \alpha$, или $F_A < f_A^{\text{кр}}$, гипотезу H_A не отклоняют: влияние фактора A на Y не подтверждается наблюдениями.

Если $P_B < \alpha$, или $F_B > f_B^{\text{кр}}$, то гипотезу H_B об отсутствии влияния фактора B на Y отклоняют и силу влияния фактора B на Y измеряют коэффициентом

$$\hat{\eta}^2_{Y|B} = SS_B/SS_{\text{итого}}; \qquad (10.39)$$

при $P_B > \alpha$, или $F_B < f_B^{\text{кр}}$, гипотезу H_B не отклоняют: влияние фактора B на Y не подтверждается наблюдениями.

Если $P_{AB} < \alpha$, или $F_{AB} > f_{AB}^{\text{кр}}$, то гипотезу H_{AB} об отсутствии влияния взаимодействия факторов A и B на Y отклоняют и силу влияния взаимодействия факторов A и B измеряют коэффициентом

$$\hat{\eta}^2_{Y|AB} = SS_{AB}/SS_{\text{итого}}; \qquad (10.40)$$

при $P_{AB} > \alpha$, или $F_{AB} < f_{AB}^{\text{кр}}$, гипотезу H_{AB} не отклоняют: влияние взаимодействия факторов A и B на Y не подтверждается.

▶ **ПРИМЕР 10.2.** При исследовании зависимости средней оценки Y по предмету «Математическая статистика» в группе студентов от фактора A — метод обучения (A имеет три уровня: $A^{(1)}$ — традиционный классический метод, $A^{(2)}$ — компьютерный, $A^{(3)}$ — комбинированный), от фактора B — будущая специальность (B имеет два уровня: $B^{(1)}$ — специальность «Социология», $B^{(2)}$ — «Информатика и вычислительная техника») и взаимодействия факторов A и B было выделено случайным образом 18 групп, которые приписывались в равных количествах шести комбинациям методов и специальностей. Знания оценивались тестом, состоящим

из 120 вопросов. Сведения о среднем числе правильных ответов в группах приведены в таблице 10.7.

Таблица 10.7

*	$B^{(1)}$	$B^{(2)}$
$A^{(1)}$	63 63 64	72 73 75
$A^{(2)}$	65 68 69	79 79 80
$A^{(3)}$	79 79 80	105 104 104*

Используя программу **«Двухфакторный дисперсионный анализ с повторениями»** пакета **«Анализ данных»**, выясним, существует или нет зависимость средней оценки по математической статистике в группе от метода обучения (A), от выбранной специальности (B) и от взаимодействия метода обучения и специальности (A B). Примем $\alpha = 0{,}05$.

В рабочий лист введем данные, содержащиеся в таблице 10.7 «от * до *». В диалоговом окне, появившемся после вызова программы, укажем в том числе «число строк для выборки» — это число m наблюдений при каждой комбинации уровней факторов A и B, равное трем.

Таблица дисперсионного анализа, полученная в результате работы программы, приведена на рисунке 10.2. Сопоставляя эту таблицу с таблицей 10.6, делаем выводы.

1) $F_A = 689{,}625$, $P_A = 4{,}12 \cdot 10^{-13}$, $f_A^{\text{кр}} = 3{,}885$. Так как $P_A < \alpha = 0{,}05$, или $F_A > f_A^{\text{кр}}$, то гипотезу об отсутствии влияния метода обучения на среднюю оценку в группе от-

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Выборка	1839	2	919,5	689,625	4,12E-13	3,88529
Столбцы	1104,5	1	1104,5	828,375	1,92E-12	4,747221
Взаимодействие	199	2	99,5	74,625	1,7E-07	3,88529
Внутри	16	12	1,333333			
Итого	3158,5	17				

Рис. 10.2

клоняем. Согласно (10.38), $\hat{\eta}^2_{Y|A} = 1839/3158{,}5 = 0{,}58$, т. е. 58% общей вариации наблюдаемых значений средней оценки в группе связано с изменением метода обучения.

2) $F_B = 828{,}375$, $P_B = 1{,}92 \cdot 10^{-12}$, $f_B^{\text{кр}} = 4{,}747$. Так как $P_B < \alpha$, или $F_B > F_B^{\text{кр}}$, то гипотезу об отсутствии зависимости средней оценки в группе от будущей специальности отклоняем. Согласно (10.39), $\hat{\eta}^2_{Y|B} = 1104{,}5/3158{,}5 = 0{,}35$, т. е. 35% вариации наблюдаемых значений средней оценки в группе объясняется разными способностями студентов групп усвоить математическую дисциплину, определившими выбор будущей специальности («гуманитарии» выбрали «Социологию», а «негуманитарии» — «Информатику и вычислительную технику»).

3) $F_{AB} = 74{,}625$, $P_{AB} = 1{,}7 \cdot 10^{-7}$, $f_{AB}^{\text{кр}} = 3{,}8859$. Судя по этим данным, взаимодействие обоих факторов также влияет на среднюю оценку в группе; взаимодействием факторов A и B объясняется, согласно (10.40), $199/3158{,}5 \cdot 100\% = 6\%$ общей вариации наблюдаемых значений средней оценки.

В результате лишь $(100 - 58 - 35 - 6)\% = 1\%$ вариации средних оценок связан с влиянием на среднюю оценку неконтролируемых случайных, или остаточных, факторов. ◀

§ 10.3. Двухфакторный дисперсионный анализ без повторений

Исходными данными двухфакторного дисперсионного анализа без повторений являются наблюдения результативного признака — случайной величины Y, проведенные при различных комбинациях уровней двух факторов A и B (A принимает ν_A уровней $A^{(1)}, A^{(2)}, \ldots, A^{(\nu_A)}$; фактор B принимает ν_B уровней $B^{(1)}, B^{(2)}, \ldots, B^{(\nu_B)}$), причем при каждой комбинации уровней проводится только одно наблюдение. В результате общее число наблюдений $n = \nu_A \nu_B$.

Отметим, что здесь можно проверить только две гипотезы H_A и H_B об отсутствии влияния на Y соответственно фактора A и фактора B, имеющие вид (10.26) и (10.27) или (10.29) и (10.30).

Проверка гипотез основана на разложении общей вариации ($SS_{\text{итого}}$) наблюдаемых значений величины Y на три составляющие:

— вариацию SS_A, причиной которой является изменение уровней фактора A;

— вариацию SS_B, обусловленную изменениями уровней фактора B;

— вариацию $SS_{ост}$, обусловленную влиянием на результат наблюдения случайных, остаточных факторов.

Это разложение имеет вид

$$SS_{итого} = SS_A + SS_B + SS_{ост}, \qquad (10.41)$$

или, если использовать обозначения, принятые в программе **«Двухфакторный дисперсионный анализ без повторений»**, следующий вид:

$$SS_{итого} = SS_{строки} + SS_{столбцы} + SS_{погрешность}.$$

В этих тождествах

$$SS_{итого} = \sum_{i=1}^{\nu_A} \sum_{j=1}^{\nu_B} (y^{(i,j)} - \bar{y})^2, \qquad (10.42)$$

$$SS_A = SS_{строки} = \nu_B \sum_{i=1}^{\nu_A} (\bar{y}_A^{(i)} - \bar{y})^2, \qquad (10.43)$$

$$SS_B = SS_{столбцы} = \nu_A \sum_{j=1}^{\nu_B} (\bar{y}_B^{(j)} - \bar{y})^2, \qquad (10.44)$$

$$SS_{ост} = SS_{погрешность} = \sum_{i=1}^{\nu_A} \sum_{j=1}^{\nu_B} (y^{(i,j)} - \bar{y}_A^{(i)} - \bar{y}_B^{(j)} + y)^2. \qquad (10.45)$$

Приведем дисперсионную таблицу двухфакторного дисперсионного анализа без повторений в общем случае (табл. 10.8).

Проверка гипотез H_A и H_B проводится так же, как и в дисперсионном анализе с повторениями, либо сравнивая «P-значение» с α, либо сравнивая «F» с критической точкой $f^{кр}$ (см. табл. 10.8).

ПРИМЕР 10.3. При уровне значимости $\alpha = 0{,}05$ выясним, влияют ли на качество пряжи (Y), измеряемое величиной разрывной нагрузки нити, тип станка (A) и вид сырья (B), из которого пряжа производится. Исходные данные приведены в таблице 10.9.

Таблица 10.9

A \ B	$B^{(1)}$	$B^{(2)}$
$A^{(1)}$	10	50
$A^{(2)}$	20	60
$A^{(3)}$	30	100

Таблица 10.8

Источник вариации значений величины Y	Измеритель вариации значений величины Y, (SS)	Степень свободы (df)	Несмещенная оценка дисперсии $\sigma^2_{\text{ост}}$, $(MS = SS/df)$	F	P-значение	$f^{\text{кр}}$, $(F_{\text{критическое}})$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Изменение уровней фактора A (строки)	SS_A $(SS_{\text{строки}})$	$\nu_A - 1$	$s^2_A = SS_A/(\nu_A - 1)$ (при выполнении гипотезы (10.26))	$F_A = s^2_A / s^2_{\text{ост}}$	P_A	$f^{\text{кр}}_A = f_{\nu_A - 1, (\nu_A - 1)(\nu_B - 1), \alpha}$
Изменение уровней фактора B (столбцы)	SS_B $(SS_{\text{столбцы}})$	$\nu_B - 1$	$s^2_B = SS_B/(\nu_B - 1)$ (при выполнении гипотезы (10.27))	$F_B = s^2_B / s^2_{\text{ост}}$	P_B	$f^{\text{кр}}_B = f_{\nu_B - 1, (\nu_A - 1)(\nu_B - 1), \alpha}$
Изменение значений остаточных факторов (погрешность)	$SS_{\text{ост}}$ $(SS_{\text{погрешность}})$	$n - \nu_A - \nu_B + 1 = (\nu_A - 1) \times (\nu_B - 1)$	$s^2_{\text{ост}} = SS_{\text{ост}}/[(\nu_A - 1)(\nu_B - 1)]$			
Общая вариация (итого)	$SS_{\text{итого}}$	$n - 1$				

Таблица дисперсионного анализа, полученная в результате работы программы «**Двухфакторный дисперсионный анализ без повторений**», приведены на рисунке 10.3. Сопоставляя ее с таблицей 10.8 делаем вывод:

— тип станка не влияет на качество пряжи ($F_A = 4{,}3(3) < f_A^{\text{кр}} = 19{,}00$);

— вид сырья влияет на качество пряжи ($F_B = 25 > f_B^{\text{кр}} = 18{,}51$); $\hat{\eta}^2_{Y|B} = 3750/5350 = 0{,}701$, т. е. 70% вариации замеров качества пряжи связано с изменением вида сырья.

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Строки	1300	2	650	4,333333	0,1875	19,00003
Столбцы	3750	1	3750	25	0,03775	18,51276
Погрешность	300	2	150			
Итого	5350	5				

Рис. 10.3

◀

УПРАЖНЕНИЯ

1. Владелец трех юридических контор пытается выяснить, отличается ли средний объем выполняемой ими за неделю работы. Для этого в каждой из контор собраны следующие сведения о еженедельном объеме выполненных работ:

Контора	1	2	3
Еженедельный объем работ, ден. ед.	280	300	350
	250	250	240
	200	210	170
	290	310	200
		270	150
		300	

а) Запишите модель дисперсионного анализа, сформулируйте проверяемую им гипотезу и условия, обеспечивающие возможность ее проверки.

б) Выясните, различаются ли конторы по среднему (генеральному) объему выполненных работ при 5% уровне значимости. Необходимые вычисления проведите «ручным способом» и используя соответствующую программу дисперсионного анализа пакета «Анализ данных».

в) Заполните таблицу дисперсионного анализа.

2. Используя соответствующую программу дисперсионного анализа пакета «**Анализ данных**», проведите дисперсионный анализ по следующим данным:

A \ B	$B^{(1)}$	$B^{(2)}$
$A^{(1)}$	6,3 6,3 6,4	7,2 7,3 7,5
$A^{(2)}$	6,5 6,8 6,9	7,9 7,9 8,0
$A^{(3)}$	7,9 7,9 8,0	9,0 8,9 8,9

Примите $\alpha = 0{,}01$. Предложите содержательную интерпретацию этих данных (признака Y, факторов A, B и их уровней); запишите модель формирования значений величины Y и проверяемые гипотезы.

3. В чем состоит отличие двухфакторного дисперсионного анализа с повторениями и без повторений? Почему в двухфакторном дисперсионном анализе без повторений не проверяется гипотеза об отсутствии влияния на результативный признак Y взаимодействия факторов?

4. Проведите двухфакторный дисперсионный анализ при следующих данных:

а) $\nu_A = 3$, $\nu_B = 5$, $m = 2$, $SS_A = 31$, $SS_B = 17$, $SS_{\text{ост}} = 2$, $SS_{\text{итого}} = 53$;

б) $\nu_A = 3$, $\nu_B = 5$, $m = 1$, $SS_A = 31$, $SS_B = 17$, $SS_{\text{итого}} = 53$.

Примите $\alpha = 0{,}01$.

ГЛАВА 11

Основы корреляционного и регрессионного анализа

В главе вводится понятие стохастической зависимости случайных величин и на примере дискретной двумерной случайной величины рассматриваются основные направления ее изучения (корреляционная зависимость, функция регрессии). Основное внимание уделяется линейной корреляционной взаимозависимости двух величин и показателю этой взаимозависимости — коэффициенту парной корреляции; также рассмотрены показатели корреляционной зависимости одной величины от другой и от группы других. Вначале излагаются вероятностные аспекты изучения зависимостей, а затем статистические — изучение зависимостей по результатам наблюдений.

Поясним на примерах различия между понятиями взаимозависимости и зависимости и между корреляционным и регрессионным анализом. Рас-

смотрим, например, вопрос о существовании связи, зависимости между ростом и весом человека; при такой постановке это задача о взаимозависимости. Но если мы хотим, используя информацию о росте человека, получить представление о его весе, то это задача о зависимости веса от роста. Это пример ситуации, в которой может представлять интерес как взаимозависимость, так и зависимость. С другой стороны, имеются ситуации, в которых важна только зависимость одной величины от другой или группы других величин. Например, интересна зависимость величины урожая от количества выпавших осадков; здесь из «вневероятностных» соображений понятно, что урожайность зависит от количества дождей, и определенно, что количество дождей не зависит от урожайности.

Рост и вес случайно выбранного человека, так же как величина урожая и количество выпавших осадков в случайно выбранный год, — случайные величины, и при изучении их связи исследователя интересуют ответы на следующие вопросы.

— Какова сила взаимозависимости этих случайных величин или какова сила зависимости одной случайной величины от другой? На этот вопрос отвечает корреляционный анализ.

— Как, зная значение x величины X, можно сделать прогноз Y_x величины Y, имея при этом в виду, что Y зависит не только от X, но и от других величин, включая и случайные? В такой постановке при каждом конкретном x величина Y_x — случайная, однако нет необходимости учитывать вероятность того, что случайная величина X примет значение x, т. е. X можно рассматривать как «неслучайную» величину. Ответ на поставленный вопрос дает регрессионный анализ.

Приведем еще один пример. Пусть Y_t — объем продукции, произведенной фирмой в году t, $t = 1, 2, \ldots, T$. Так как значение величины Y_t формируется под влиянием различных экономических факторов, включая и случайные, то Y_t — случайная величина; год t не случаен. Изучение зависимости Y_t от t — задача регрессионного анализа.

§ 11.1. Понятие функциональной, стохастической и корреляционной зависимости. Функция регрессии

Обозначим через X независимую переменную, а через Y — зависимую переменную.

Зависимость величины Y от X называется *функциональной*, если каждому значению величины X соответствует единственное значение величины Y. С функциональной зависимостью мы встречаемся, например, в математике, при изучении физических законов. Обратим внимание на то, что если X — детерминированная величина (т. е. величина, значения которой не зависят от случая), то и функционально зависящая от нее величина Y тоже является детерминированной; если же X — случайная величина, то и Y также случайная величина.

Гораздо чаще в окружающем нас мире имеет место не функциональная, а ***стохастическая***, или ***вероятност-***

ная, зависимость Y от X, когда каждому фиксированному значению независимой переменной X соответствует не одно, а множество значений переменной Y, причем сказать заранее, какое именно значение примет величина Y, нельзя. Более частое появление такой зависимости объясняется действием на Y не только величины X, но и других величин, включая и случайные.

В этой ситуации переменная Y является случайной величиной. Переменная X может быть как детерминированной, так и случайной величиной. Заметим, что со стохастической зависимостью мы уже сталкивались в дисперсионном анализе. В таблице 10.1 в роли переменной X выступает фактор A; уровни фактора A — это значения переменной X; каждому такому значению соответствует не одно, а множество непредсказуемых заранее значений величины Y.

Допустим, что существует стохастическая зависимость случайной величины Y от X. Зафиксируем некоторое значение x переменной X, тогда величина Y в силу ее стохастической зависимости от X может принять любое значение из некоторого множества, причем какое именно — заранее не известно. Среднее этого множества называют ***генеральным групповым средним*** величины Y, или математическим ожиданием случайной величины Y, вычисленным при условии, что X примет значение x; это условное математическое ожидание обозначают так: $M(Y|x)$. Если существует стохастическая зависимость Y от X, то прежде всего стараются выяснить, изменяются или нет при изменении x условные математические ожидания $M(Y|x)$. Если при изменении x условные математические ожидания $M(Y|x)$ изменяются, то говорят, что имеет место ***корреляционная зависимость*** величины Y от X; если условные математические ожидания остаются неизменными, то говорят, что корреляционная зависимость величины Y от X отсутствует.

Функция $M(Y|x) = \varphi(x)$, описывающая изменение условного математического ожидания случайной величины Y при изменении значений x величины X, называется ***функцией регрессии***, или ***регрессией Y на x***[1].

[1] Термин «регрессия» (от англ. *regression* — упадок, движение назад) был введен английским ученым Ф. Гальтоном (1822—1911) для обозначения «возврата» в среднем особенностей детей к показателям родителей. В современной науке регрессией называют зависимость условного математического ожидания $M(Y|x)$, или условного среднего случайной величины Y от x.

Выясним, почему именно при наличии стохастической зависимости интересуются поведением условного математического ожидания. Рассмотрим пример. Пусть X — уровень квалификации рабочего, Y — его выработка за смену. Ясно, что зависимость Y от X не функциональная, а стохастическая: на выработку помимо квалификации влияет множество других факторов, включая и случайные. Зафиксируем значение x уровня квалификации; ему соответствует некоторое множество значений выработки Y. Тогда $M(Y \mid x)$ — средняя выработка рабочего при условии, что его уровень квалификации равен x, или, иначе говоря, $M(Y \mid x)$ — это норматив выработки при уровне квалификации, равной x. Зная зависимость этого норматива от уровня квалификации, можно для любого уровня квалификации рассчитать норматив выработки и, сравнив его с реальной выработкой, оценить работу рабочего.

Обратим внимание на то, что введенные понятия стохастической и корреляционной зависимости относились к генеральной совокупности. Поясним эти понятия числовым примером.

▶ **ПРИМЕР 11.1.** Допустим, что одновременно изучаются две дискретные случайные величины X и Y, или, иначе говоря, двумерная дискретная случайная величина (X, Y), которая задана таблицей 11.1.

Таблица 11.1

j	i	1	2	3
	x_i / y_j	$x_1 = 2$	$x_2 = 5$	$x_3 = 8$
1	$y_1 = 0{,}4$	0,15	0,12	0,03
2	$y_2 = 0{,}8$	0,05	0,30	0,35

Таблицу 11.1 называют ***таблицей распределения вероятностей*** двумерной величины (X, Y); ее следует понимать так. Случайная величина X может принять одно из следующих значений: 2, 5 и 8. Случайная величина Y — значения 0,4 и 0,8. Число 0,15 — это вероятность того, что X примет значение 2 и одновременно Y примет значение 0,4, или, иначе говоря, вероятность пересечения двух событий, $P((X = 2) \cap (Y = 0{,}4)) = 0{,}15$. Аналогично, вероятность $P((X = 2) \cap (Y = 0{,}8)) = 0{,}05$ и т. д. Поскольку в таблице 11.1 указаны все возможные значения величин X и Y, сумма вероятностей, стоящих в таблице, должна быть равна единице: $0{,}15 + 0{,}05 + 0{,}12 + 0{,}30 + 0{,}03 + 0{,}35 = 1$.

Из таблицы 11.1 видно, что зависимость Y от X — стохастическая, поскольку при каждом фиксированном значении величины X величина Y может быть равной либо 0,4, либо 0,8, причем, какому из этих чисел она будет равна, сказать заранее нельзя.

Установим законы распределения и характеристики величин X и Y.

а) Закон распределения величины X. Он определяется таблицей 11.2.

Таблица 11.2

x	$x_1 = 2$	$x_2 = 5$	$x_3 = 8$	$MX = 5{,}54$
$P(X = x)$	$0{,}15 + 0{,}05 = 0{,}2$	$0{,}12 + 0{,}30 = 0{,}42$	$0{,}03 + 0{,}35 = 0{,}38$	$DX = 4{,}9284$

Действительно, например, величина X примет значение, равное 2, только в том случае, когда одновременно с этим величина Y примет значение 0,4 или 0,8, т. е.

$$P(X = 2) = P((X = 2) \cap (Y = 0{,}4)) + P((X = 2) \cap (Y = 0{,}08)) = 0{,}15 + 0{,}05 = 0{,}2.$$

Справа от ряда распределения величины X находятся ее математическое ожидание и дисперсия.

б) Закон распределения величины Y. Он определяется таблицей 11.3.

Таблица 11.3

y	$y_1 = 0{,}4$	$y_2 = 0{,}8$	$MY = 0{,}68$
$P(Y = y)$	$0{,}15 + 0{,}12 + 0{,}03 = 0{,}30$	$0{,}05 + 0{,}30 + 0{,}35 = 0{,}7$	$DY = 0{,}0336$

Для того чтобы выяснить, существует или нет корреляционная зависимость Y от X, найдем условные законы распределения величины Y, а именно закон распределения величины Y сначала при условии, что случайная величина X примет значение 2, затем при условии, что она примет значение 5, и наконец, при условии, что она примет значение 8.

1. Пусть $x = 2$. Тогда условная вероятность

$$P((Y = 0{,}4) \mid (x = 2)) = \frac{P((Y = 0{,}4) \cap (X = 2))}{P(X = 2)} = \frac{0{,}15}{0{,}2} = 0{,}75,$$

а условная вероятность

$$P((Y = 0{,}8) \mid (x = 2)) = \frac{P((Y = 0{,}8) \cap (X = 2))}{P(X = 2)} = \frac{0{,}05}{0{,}2} = 0{,}25.$$

Таким образом, закон распределения величины Y при условии, что $x = 2$, задан таблицей 11.4.

Таблица 11.4

y	$y_1 = 0{,}4$	$y_2 = 0{,}8$
$P((Y = y) \mid (x = 2))$	0,75	0,25

$M(Y \mid (x = 2)) = 0{,}4 \cdot 0{,}75 + 0{,}8 \cdot 0{,}25 = 0{,}5$

$D(Y \mid (x = 2)) = 0{,}03$

Справа от таблицы помещено условное математическое ожидание и значение условной дисперсии. Покажем, как вычисляется условная дисперсия. Общая формула условной дисперсии имеет вид

$$D(Y \mid x) = M[(Y \mid x) - M(Y \mid x)]^2. \qquad (11.1)$$

Для таблицы 11.4 получаем

$$D(Y \mid (x = 2)) = M[(Y \mid (x = 2)) - M(Y \mid (x = 2))]^2 =$$

$$= M[(Y \mid (x = 2)) - 0{,}5]^2 = \sum_{i=1}^{2} (y_i - 0{,}5)^2 P((Y = y_i) \mid (x = 2)) =$$

$$= (0{,}4 - 0{,}5)^2 \cdot 0{,}75 + (0{,}8 - 0{,}5)^2 \cdot 0{,}25 = 0{,}03.$$

2. Пусть $x = 5$. Тогда $P((Y = 0{,}4) \mid (x = 5)) = \frac{P((Y = 0{,}4) \cap (X = 5))}{P(X = 5)} = \frac{0{,}12}{0{,}42} = \frac{2}{7}$; $P((Y = 0{,}8) \mid (x = 5)) = \frac{P((Y = 0{,}8) \cap (X = 5))}{P(X = 5)} = \frac{0{,}30}{0{,}42} = \frac{5}{7}$. Таким образом, закон распределения величины Y при условии, что $x = 5$, имеет вид таблицы 11.5.

Таблица 11.5

y	0,4	0,8
$P((Y = y) \mid (x = 5))$	2/7	5/7

$M(Y \mid (x = 5)) = 24/35 = 0{,}686$

$D(Y \mid (x = 5)) = 0{,}03265$

3. И наконец, при $x = 8$ ряд распределения величины Y задан таблицей 11.6.

Таблица 11.6

y	0,4	0,8
$P((Y = y) \mid (x = 8))$	3/38	35/38

$M(Y \mid (x = 8)) = 73/95 = 0{,}768$

$D(Y \mid (x = 8)) = 0{,}01163$

Проведенные расчеты показывают, что с изменением значений x величины X меняются условные математические ожидания $M(Y \mid x)$ случайной величины Y. Следовательно, корреляционная зависимость Y от X существует. Функция регрессии Y на x задается таблицей 11.7.

Таблица 11.7

x	$x_1 = 2$	$x_2 = 5$	$x_3 = 3$
$M(Y \mid x)$	0,5	$24/35 = 0{,}686$	$73/95 = 0{,}768$

На рисунке 11.1 дано графическое изображение заданного таблицей 11.1 распределения вероятностей $P[(X = x_i) \cap (Y = y_j)]$ между всевозможными парами (x_i, y_j) значений величин X и Y: центры окружностей — это точки $(x_i; y_j)$, а «массы точек» (площади кругов) равны соответствующим вероятностям. Такое изображение распределения вероятностей называют ***диаграммой разброса*** (смысл ломаной и прямой линий выясняется в 11.2.1).

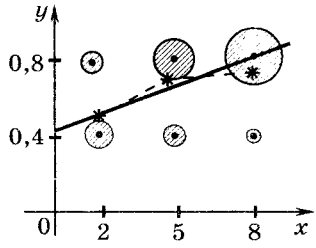


Рис. 11.1

Дополним таблицу 11.1 полученными результатами (см. таблицу 11.8).

Таблица 11.8

j	i / x_i / y_j	1	2	3
		$x_1 = 2$	$x_2 = 5$	$x_3 = 8$
1	$y_1 = 0{,}4$	0,15	0,12	0,03
2	$y_2 = 0{,}8$	0,05	0,30	0,35
	$P(X = x_i)$	0,20	0,42	0,38
	$M(Y \mid x_i)$	0,5	0,686	0,768
	$D(Y \mid x_i)$	0,03	0,03265	0,01163
	$M^{\text{лин}}(Y \mid x_i)$[1]	0,54	0,66	0,78

[1] Смысл этой величины выясняется далее в п. 11.2.1.

Обратим внимание на следующее. Величина X — случайная (вероятности ее значений приведены в таблице 11.8); сказать заранее, какое значение она примет, нельзя. Поэтому и нельзя заранее сказать, какое значение примет условное математическое ожидание величины Y, т. е. условное математическое ожидание — это случайная величина, которую обозначим $M(Y \mid X)$. Очевидно, вероятность

того, что эта величина примет значение $M(Y \mid x_i)$ равна $P(X = x_i)$, т. е.

$$P[M(Y \mid X) = M(Y \mid x_i)] = P(X = x_i).$$

Также нельзя заранее сказать, какое значение примет и условная дисперсия, т. е. условная дисперсия $D(Y \mid X)$ — это случайная величина, и вероятность того, что она примет значение $D(Y \mid x_i)$ равна $P(X = x_i)$.

Так как $M(Y \mid X)$ и $D(Y \mid X)$ — случайные величины, то каждая из них имеет математическое ожидание и дисперсию. Найдем, например, математическое ожидание случайной величины $M(Y \mid X)$. Используя таблицу 11.8, получим

$$M[M(Y \mid X)] = \sum_{i=1}^{3} M(Y \mid x_i)P(X = x_i) =$$

$$= 0{,}5 \cdot 0{,}2 + 0{,}686 \cdot 0{,}42 + 0{,}768 \cdot 0{,}38 = 0{,}68$$

— результат совпал с $MY = 0{,}68$, найденным в таблице 11.3. Это совпадение не случайно. ◀

Всегда имеет место следующее равенство

$$M[M(Y \mid X)] = MY. \qquad (11.2)$$

≫ Доказательство проведем на примере дискретной двумерной величины. Имеем

$$M[M(Y \mid X)] = \sum_i M(Y \mid x_i)P(X = x_i) =$$

$$= \sum_i [\sum_j y_j P((Y = y_j) \mid x_i)]P(X = x_i) = \sum_i \sum_j y_j P((Y = y_j) \mid x_i)P(X = x_i) =$$

$$= \sum_i \sum_j y_j P[(Y = y_j) \cap (X = x_i)] = \sum_j \sum_i y_j P[(Y = y_j) \cap (X = x_i)] =$$

$$= \sum_j y_j \sum_i P[(Y = y_j) \cap (X = x_i)] = \sum_j y_j P(Y = y_j) = MY. \quad ≪$$

Рассмотрим теперь основы корреляционного анализа — изучим показатели корреляционной зависимости. Основным показателем является коэффициент корреляции, характеризующий линейную корреляционную зависимость.

§ 11.2. Основы корреляционного анализа

11.2.1. Коэффициент парной корреляции, линейная регрессия и свойства коэффициента парной корреляции

Коэффициент парной корреляции

Коэффициент парной корреляции (или просто *коэффициент корреляции*) случайных величин X и Y вычисляется по одной из двух следующих тождественных формул:

$$r_{X,Y} = \frac{M[(X - MX)(Y - MY)]}{\sigma_X \sigma_Y} \tag{11.3}$$

или

$$r_{X,Y} = \frac{M(XY) - MX \cdot MY}{\sigma_X \sigma_Y}, \tag{11.4}$$

где $\sigma_X \neq 0$ и $\sigma_Y \neq 0$.

Тождественность формул (11.3) и (11.4) была доказана в п. 4.3.2. Постоянная величина [числитель дроби (11.3), или дроби (11.4)]

$$K(X, Y) = M[(X - MX)(Y - MY)] \equiv M(XY) - MX \cdot MY \tag{11.5}$$

называется ***корреляционным моментом*** или ***ковариацией*** случайных величин X и Y. Корреляционный момент связан с коэффициентом корреляции соотношением

$$K(X, Y) = r_{X,Y}\sigma_X\sigma_Y. \tag{11.6}$$

Обратим внимание на следующее.

1. $r_{X,Y} = r_{Y,X}$, т. е. коэффициент корреляции симметричен относительно величин X и Y.

2. Положив $Y = X$, имеем

$$K(X, X) \underset{(11.5)}{=} M(X - MX)^2 \equiv M(X^2) - (MX)^2 = DX$$

и

$$r_{X,X} \underset{(11.4)}{=} \frac{M(X^2) - (MX)^2}{\sigma_X \sigma_X} = \frac{DX}{DX} = 1.$$

Таким образом, корреляционный момент и коэффициент корреляции величины «самой с собой» соответственно равны

$$K(X, X) = DX; \tag{11.7}$$

$$r_{X,X} = 1, (\sigma_X \neq 0). \tag{11.8}$$

3. Имеем $r_{b+cX,\,d+eY} = \begin{cases} r_{X,Y}, & \text{если } ce > 0, \\ -r_{X,Y}, & \text{если } ce < 0. \end{cases}$ (11.9)

(В справедливости равенств (11.9) убедитесь самостоятельно.)

Вычисление коэффициента корреляции по формуле (11.4) менее трудоемко, чем по формуле (11.3), поэтому в дальнейшем чаще будем использовать формулу (11.4).

В этой формуле «присутствует» математическое ожидание произведения двух случайных величин. Чтобы его найти (ограничимся случаем когда X и Y — дискретные слу-

чайные величины), надо знать вероятности $P[(X = x_i) \cap (Y = y_j)]$, или, иначе, знать закон распределения вероятностей двумерной случайной величины (X, Y). Основной формой его задания является таблица распределения вероятностей между всевозможными парами значений величин X и Y (примером является таблица 11.1).

▶ **ПРИМЕР 11.1** (продолжение). Вычислим коэффициент корреляции между компонентами двумерной случайной величины (X, Y), распределение которой задано таблицей 11.1.

Воспользуемся формулой (11.4). В § 11.1 получено $MX = 5{,}54$, $DX = 4{,}9284$, $MY = 0{,}68$, $DY = 0{,}0336$ (см. таблицы 11.2 и 11.3); найдем $M(XY)$. Имеем

$$M(XY) = \sum_{i=1}^{3} \sum_{j=1}^{2} x_i y_j P((X = x_i) \cap (Y = y_j)) = 2 \cdot 0{,}4 \cdot 0{,}15 +$$

$$+ 2 \cdot 0{,}8 \cdot 0{,}05 + 5 \cdot 0{,}4 \cdot 0{,}12 + 5 \cdot 0{,}8 \cdot 0{,}30 + 8 \cdot 0{,}4 \cdot 0{,}03 +$$

$$+ 8 \cdot 0{,}8 \cdot 0{,}35 = 3{,}976.$$

Коэффициент корреляции

$$r_{X,Y} \underset{(11.4)}{=} \frac{3{,}976 - 5{,}54 \cdot 0{,}68}{\sqrt{4{,}9284}\sqrt{0{,}0336}} = 0{,}513. \quad ◀$$

Итак, в примере 11.1 $r_{X,Y} = 0{,}513$. Чтобы ответить на вопрос, много это или мало, а также каков смысл выражения «коэффициент корреляции равен 0,513», введем понятие линейной регрессии (или прямой регрессии) и рассмотрим свойства коэффициента корреляции.

Линейная регрессия

Обратимся к примеру 11.1, в котором изучалась дискретная двумерная случайная величина (X, Y), заданная таблицей 11.1. Напомним, при каждом значении x_i случайной величины X было вычислено условное математическое ожидание $M(Y \mid x_i)$ (см. таблицу 11.8). Точки с координатами $(x_i, M(Y \mid x_i))$, $i = 1, 2, 3$, отмечены крестиками на рисунке 11.1; линия, соединяющая эти точки, называется ***линией регрессии*** Y на x. На рисунке 11.1 линия регрессии — ломаная линия.

Если при изменении значений x величины X условное математическое ожидание $M(Y \mid x)$ изменяется по линейному закону, т. е.

$$M(Y \mid x) = a_0 + a_1 x, \; a_1 \neq 0, \qquad (11.10)$$

то говорят, что ***регрессия Y на x линейная*** или ***прямая, а корреляционная зависимость Y от X линейная.***

Числа a_0 и a_1 называют *параметрами линейной регрессии*; чтобы подчеркнуть, что изучается регрессия Y на x, параметр a_1 часто обозначают $a_{Y|x}$.

Далее уравнение линейной регрессии (11.10) будем записывать так:

$$M^{\text{лин}}(Y \mid x) = a_0 + a_1 x,\ a_1 \neq 0. \qquad (11.11)$$

Найдем выражения параметров a_0 и a_1 через характеристики случайных величин X и Y.

≫ Так как X — случайная величина, то и величина $M^{\text{лин}}(Y \mid X)$ — случайная, ее математическое ожидание найдем двумя способами.

1) Используя (11.11), получим

$$M[M^{\text{лин}}(Y \mid X)] = M(a_0 + a_1 X) = a_0 + a_1 MX,$$

или

$$M[M^{\text{лин}}(Y \mid X)] = a_0 + a_1 MX. \qquad (11.12)$$

2) Так как равенство (11.2) имеет место при любом виде зависимости условного математического ожидания $M(Y \mid X)$ от X, оно выполняется и при линейной зависимости, т. е.

$$M[M^{\text{лин}}(Y \mid X)] = MY. \qquad (11.13)$$

Сравнив это равенство с (11.12), делаем вывод

$$MY = a_0 + a_1 MX. \qquad (11.14)$$

Учитывая (11.11) и (11.14), найдем разность

$$M^{\text{лин}}(Y \mid X) - MY = a_1(X - MX),$$

а затем левую и правую части этого равенства умножим на $(X - MX)$ и применим к ним операцию математического ожидания:

$$M[(M^{\text{лин}}(Y \mid X) - MY)(X - MX)] = M[a_1(X - MX)^2]. \qquad (11.15)$$

Докажем, что

$$M[(M(Y \mid X) - MY)(X - MX)] = K(X, Y), \qquad (11.16)$$

где $K(X, Y)$ — корреляционный момент случайных величин X и Y [см. (11.5)].

≫ $M[(M(Y \mid X) - MY)(X - MX)] = M[X \cdot M(Y \mid X) - X \cdot MY -$
$- MX \cdot M(Y \mid X) + MX \cdot MY] = M[X \cdot M(Y \mid X)] - M(X \cdot \underbrace{MY}_{\text{const}}) -$
$- M[\underbrace{MX}_{\text{const}} \cdot M(Y \mid X)] + M(\underbrace{MX \cdot MY}_{\text{const}}) = M[X \cdot M(Y \mid X)] -$
$- MY \cdot MX - MX \cdot \underbrace{M[M(Y \mid X)]}_{MY\ (\text{см.}(11.2))} + MX \cdot MY =$
$= M[X \cdot M(Y \mid X)] - MX \cdot MY \underset{(*)}{=} M(XY) - MX \cdot MY \underset{(11.5)}{=}$
$\underset{(11.5)}{=} K(X, Y).$

На примере дискретной случайной величины убедимся в обоснованности перехода (*), а именно в том, что

$$M[X \cdot M(Y \mid X)] = M(XY). \tag{11.17}$$

Действительно,

$$M[X \cdot M(Y \mid X)] = \sum_i x_i M(Y \mid (X = x_i)) P(X = x_i) =$$

$$= \sum_i x_i [\sum_j y_j P((Y = y_j) \mid (X = x_i))] P(X = x_i) =$$

$$= \sum_i \sum_j x_i y_j P((Y = y_j) \mid (X = x_i)) P(X = x_i) =$$

$$= \sum_i \sum_j x_i y_j P[(Y = y_j) \cap (X = x_i)] = M(XY).$$

Учитывая соотношение (11.16), которое имеет место при любом виде зависимости условного математического ожидания $M(Y \mid X)$ от X, включая и линейную зависимость, заменим левую часть равенства (11.15) корреляционным моментом $K(X, Y)$. Тогда равенство (11.15) принимает вид

$$K(X, Y) = a_1 M(X - MX)^2, \text{ или } K(X, Y) = a_1 \sigma_X^2 .$$

Далее, учитывая, что $K(X, Y) \underset{(11.6)}{=} r_{X,Y}\sigma_X\sigma_Y$, получим $r_{X,Y}\sigma_X\sigma_Y = a_1\sigma_X^2$. Отсюда

$$a_1 = a_{Y|x} = r_{X,Y}\sigma_Y/\sigma_X, \tag{11.18}$$

и тогда из равенства (11.14) имеем

$$a_0 = MY - r_{X,Y}\frac{\sigma_Y}{\sigma_X} MX. \quad \ll \tag{11.19}$$

Подставив найденные значения параметров a_0 и a_1 в выражение (11.11), получим

$$M^{\text{лин}}(Y \mid x) = \underbrace{MY - r_{X,Y}\frac{\sigma_Y}{\sigma_X} MX}_{a_0} + \underbrace{r_{X,Y}\frac{\sigma_Y}{\sigma_X}}_{a_1 = a_{Y|x}} x, \; a_1 \neq 0, \tag{11.20}$$

или

$$M^{\text{лин}}(Y \mid x) = MY + r_{X,Y}\frac{\sigma_Y}{\sigma_X}(x - MX). \tag{11.21}$$

Итак, линейная регрессия Y на x, или прямая регрессии Y на x имеет вид (11.20), или (11.21).

Обратим внимание на следующее.

1) При $r_{X,Y} > 0$ угловой коэффициент прямой регрессии $a_1 > 0$ и увеличение значений x величины X сопровождается увеличением значений $M^{\text{лин}}(Y \mid x)$ величины $M^{\text{лин}}(Y \mid X)$. При $r_{X,Y} < 0$ угловой коэффициент $a_1 < 0$ и уве-

личение значений x сопровождается уменьшением значений $M^{\text{лин}}(Y \mid x)$.

2) Прямая регрессии Y на x проходит через точку $(MX; MY)$. Действительно, если в уравнении (11.21) положить $x = MX$, то его правая часть $M^{\text{лин}}(Y \mid x)$ принимает значение MY.

▶ **ПРИМЕР 11.1** (продолжение). Напомним, что в условиях примера $MX = 5{,}54$, $DX = 4{,}9284$, $MY = 0{,}68$, $DY = 0{,}0336$, $r_{X,Y} = 0{,}513$. Линейная регрессия Y на x (11.21) принимает вид $M^{\text{лин}}(Y \mid x) = 0{,}68 + 0{,}042(x - 5{,}54)$. Эта прямая изображена на рисунке 11.1; она проходит через точку $(MX = 5{,}54; MY = 0{,}68)$ и «выравнивает»(«спрямляет») фактическую (ломаную) линию регрессии. Значения $M^{\text{лин}}(Y \mid x)$, найденные при $x = 2$, $x = 5$ и $x = 8$, приведены в последней строке таблицы 11.8: $0{,}53 = 0{,}68 + 0{,}042(2 - 5{,}54)$; $0{,}66 = 0{,}68 + 0{,}042(5 - 5{,}54)$; $0{,}78 = 0{,}68 + 0{,}042(8 - 5{,}54)$. ◀

Докажем три теоремы, связанные с величиной $M^{\text{лин}}(Y \mid X)$, которая определяется равенством (11.20), или (11.21) при $x = X$. Эти теоремы содержат утверждения относительно математических ожиданий и дисперсий величин $M^{\text{лин}}(Y \mid X)$ и $M^{\text{лин}}(Y \mid X) - Y$. Предварительно еще раз подчеркнем следующее. Так как X — случайная величина, то и зависящая от нее величина $M^{\text{лин}}(Y \mid X)$ является случайной; также случайной будет и величина $M^{\text{лин}}(Y \mid X) - Y$, которая, по сути, представляет собой случайную ошибку, возникающую при использовании величины $M^{\text{лин}}(Y \mid X)$ вместо Y.

Теорема 1. *Математическое ожидание, или среднее значений случайной величины* $M^{\text{лин}}(Y \mid X)$, *равно* MY:

$$M[M^{\text{лин}}(Y \mid X)] = MY, \qquad (11.22)$$

а ее дисперсия $D[M^{\text{лин}}(Y \mid X)]$ (*которую обозначим* $\sigma^2_{LR\,Y|X}$ *и назовем дисперсией линейной регрессии* Y *на* X), *равна*

$$\sigma^2_{LR\,Y|X} = D[M^{\text{лин}}(Y \mid X)] \underset{(*)}{=}$$

$$= M[M^{\text{лин}}(Y \mid X) - MY]^2 \underset{(**)}{=} \sigma^2_Y r^2_{X,Y}{}^{1}. \qquad (11.22)$$

Замечание. Равенство (*) читается так: «дисперсия линейной регрессии Y на X равна математическому ожиданию, или среднему значению квадрата отклонения линейной регрессии от среднего значения величин Y».

≫ Равенство $M[M^{\text{лин}}(Y \mid X)] = MY$ в доказательстве не нуждается: его справедливость была обоснована выше [см. (11.13)].

[1] Индекс «LR» от англ. *linear regression* — линейная регрессия.

Докажем равенства (*) и (**) в (11.23). Соответственно имеем

$$D\underbrace{[M^{\text{лин}}(Y \mid X)]}_{Z} = M(Z - MZ)^2 = M[M^{\text{лин}}(Y \mid X) -$$

$$- M(M^{\text{лин}}(Y \mid X))]^2 \underset{(11.13)}{=} M[M^{\text{лин}}(Y \mid X) - MY]^2;$$

$$D[M^{\text{лин}}(Y \mid X)] \underset{(11.20)}{=} D(a_0 + a_1 X) = a_1^2 DX \underset{(11.18)}{=}$$

$$\underset{(11.18)}{=} r_{X,Y}^2 \sigma_Y^2 / \sigma_X^2 DX = \sigma_Y^2 r_{X,Y}^2 . \quad \ll$$

Подтвердим утверждения теоремы на данных примера 11.1.

▶ **ПРИМЕР 11.1** (продолжение). Обратимся к таблице 11.8 и найдем:

$$M[M^{\text{лин}}(Y \mid X)] = \sum_{i=1}^{3} M^{\text{лин}}(Y \mid x_i) P(X = x_i) =$$

$$= (0{,}53 \cdot 0{,}20 + 0{,}66 \cdot 0{,}42 + 0{,}78 \cdot 0{,}38 = 0{,}68 = MY;$$

$$D[M^{\text{лин}}(Y \mid X)] = \sum_{i=1}^{3} [M^{\text{лин}}(Y \mid x_i) - MY]^2 P(X = x_i) =$$

$$= (0{,}53 - 0{,}68)^2 0{,}20 + (0{,}66 - 0{,}68)^2 0{,}42 + (0{,}78 -$$

$$- 0{,}68)^2 0{,}38 = 0{,}008,$$

что совпадает с числовым значением выражения $\sigma_Y^2 r_{X,Y}^2 =$ $= 0{,}0336 \cdot 0{,}513^2 = 0{,}008$. ◀

Теорема 2. *Математическое ожидание, или среднее значение случайной ошибки, которая равна* $Y - M^{\text{лин}}(Y \mid X)$, *возникающей при замене величины* Y *величиной* $M^{\text{лин}}(Y \mid X)$, *равно нулю*:

$$M[Y - M^{\text{лин}}(Y \mid X)] = 0, \qquad (11.24)$$

а ее дисперсия $D[Y - M^{\text{лин}}(Y \mid X)]$ (*которую обозначим* $\sigma^2_{ELR\,Y|X}$ *и назовем дисперсией ошибки линейной регрессии* Y *на* X) *имеет следующий вид*:

$$\sigma^2_{ELR\,Y|X} = D[Y - M^{\text{лин}}(Y \mid X)] \underset{(*)}{=}$$

$$\underset{(*)}{=} M[Y - M^{\text{лин}}(Y \mid X)]^2 \underset{(**)}{=} \sigma_Y^2 (1 - r_{X,Y}^2)^1. \qquad (11.25)$$

Замечания. 1. Замена величины Y величиной $M^{\text{лин}}(Y \mid X)$ означает, что при любом значении x величины X соответствующие ему значе-

[1] Индекс «*ELR*» от англ. *error of linear regression* — ошибка линейной регрессии.

ния y_x величины Y заменяют значениями $M^{\text{лин}}(Y \mid x)$ случайной величины $M^{\text{лин}}(Y \mid X)$, рассчитанными по уравнению (11.20).

2. Равенство (*) в (11.25) читается так: «дисперсия ошибки линейной регрессии Y на X равна математическому ожиданию квадрата отклонения величины Y от линейной регрессии Y на X, или среднему квадрата ошибки, возникающей при замене Y величиной $M^{\text{лин}}(Y \mid X)$».

3. Можно доказать, что среднее квадрата ошибки, возникающей при замене Y любой величиной, линейно зависящей от X, но отличной от $M^{\text{лин}}(Y \mid X) = a_0 + a_1 X$, где a_0 и a_1 определяются по формулам (11.18) и (11.19), больше $M[Y - M^{\text{лин}}(Y \mid X)]^2$; говорят, что равная этому математическому ожиданию дисперсия $D[Y - M^{\text{лин}}(Y \mid X)]$ ошибки линейной регрессии обладает свойством минимальности.

≫ Докажем (11.24). Имеем

$$M[Y - M^{\text{лин}}(Y \mid X)] = MY - M[M^{\text{лин}}(Y \mid X)] \underset{(11.13)}{=} MY - MY = 0.$$

Докажем равенства (*) и (**) в (11.25). Соответственно имеем:

$$D[\underbrace{Y - M^{\text{лин}}(Y \mid X)}_{V}] = M(V - MV)^2 = M[V - M(Y - M^{\text{лин}}(Y \mid X)]^2 \underset{(11.24)}{=}$$

$$\underset{(11.24)}{=} M(V - 0)^2 = MV^2 = M[Y - M^{\text{лин}}(Y \mid X)]^2;$$

$$D[Y - M^{\text{лин}}(Y \mid X)] \underset{(11.20)}{=} D(Y - a_0 - a_1 X) \underset{a_0 = \text{const}}{=} D(Y - a_1 X) \underset{(4.61)}{=}$$

$$\underset{(4.61)}{=} DY + D(a_1 X) - 2 r_{Y, a_1 X} \sigma_Y \sigma_{a_1 X} \underset{(11.6)}{=} DY + D(a_1 X) - 2K(Y, a_1 X) \underset{(11.5)}{=}$$

$$\underset{(11.5)}{=} DY + D(a_1 X) - 2[M(Y a_1 X) - MY \cdot M(a_1 X)] \underset{a_1 = \text{const}}{=}$$

$$\underset{a_1 = \text{const}}{=} DY + a_1^2 DX - 2a_1[M(YX) - MY \cdot MX] \underset{(11.5)}{=}$$

$$\underset{(11.5)}{=} DY + a_1^2 DX - 2a_1 K(Y, X) \underset{(11.6)}{=} \sigma_Y^2 + a_1^2 \sigma_X^2 - 2a_1 r_{Y, X} \sigma_Y \sigma_X \underset{(11.18)}{=}$$

$$\underset{(11.18)}{=} \sigma_Y^2 + r_{X, Y}^2 \frac{\sigma_Y^2}{\sigma_X^2} \sigma_X^2 - 2 r_{X, Y} \frac{\sigma_Y}{\sigma_X} r_{Y, X} \sigma_Y \sigma_X =$$

$$= \sigma_Y^2 + r_{X, Y}^2 \sigma_Y^2 - 2 r_{X, Y}^2 \sigma_Y^2 = \sigma_Y^2 (1 - r_{X, Y}^2). \quad ≪$$

Подтвердим утверждения теоремы 2 на данных примера 11.1.

▶ **ПРИМЕР 11.1** (продолжение). Обратимся к таблице 11.8 и найдем:

$$M[Y - M^{\text{лин}}(Y \mid X)] = \sum_{i=1}^{3} \sum_{j=1}^{2} [y_j - M^{\text{лин}}(Y \mid x_i)] \times$$

$$\times P[(X = x_i) \quad (Y = y_j)] = (0{,}4 - 0{,}53) \cdot 0{,}15 +$$

$$+ (0{,}8 - 0{,}53) \cdot 0{,}05 + (0{,}4 - 0{,}66) \cdot 0{,}12 + (0{,}8 - 0{,}66) \cdot 0{,}30 +$$
$$+ (0{,}4 - 0{,}78) \cdot 0{,}03 + (0{,}8 - 0{,}78) \cdot 0{,}35 = 0{,}000;$$
$$D[Y - M^{\text{лин}}(Y \mid X)] = M[Y - M^{\text{лин}}(Y \mid X)]^2 =$$
$$= \sum_{i=1}^{3} \sum_{j=1}^{2} [y_j - M^{\text{лин}}(Y \mid x_i)]^2 \cdot P[(X = x_i) \quad (Y = y_j)] =$$
$$= (-0{,}13)^2 \cdot 0{,}15 + 0{,}27^2 \cdot 0{,}05 + (-0{,}26)^2 \cdot 0{,}12 + 0{,}14^2 \cdot 0{,}30 +$$
$$+ (-0{,}38)^2 \cdot 0{,}03 + 0{,}02^2 \cdot 0{,}35 = 0{,}025.$$

Полученное значение совпадает с числовым значением выражения $\sigma_Y^2(1 - r_{X,Y}^2) = 0{,}0336(1 - 0{,}513^2) = 0{,}025$. ◂

Теорема 3. *Сумма дисперсии линейной регрессии Y на X и дисперсии ошибки этой регрессии равна дисперсии величины Y:*

$$\sigma_{LR\,Y|X}^2 + \sigma_{ELR\,Y|X}^2 = \sigma_Y^2, \tag{11.26}$$

или

$$D[M^{\text{лин}}(Y \mid X)] + D[Y - M^{\text{лин}}(Y \mid X)] = DY.$$

≫ Учитывая соотношения (11.23) и (11.25), получим

$$\sigma_{LR\,Y|X}^2 + \sigma_{ELR\,Y|X}^2 = D[M^{\text{лин}}(Y \mid X)] + D[Y - M^{\text{лин}}(Y \mid X)] =$$
$$= \sigma_Y^2 r_{X,Y}^2 + \sigma_Y^2 (1 - r_{X,Y}^2) = \sigma_Y^2 = DY,$$

что и требовалось доказать. ≪

Выше была рассмотрена линейная регрессия Y на x (Y — зависимая переменная, x — независимая переменная). Аналогично можно получить следующие результаты:

— линейная регрессия X на y определяется формулой

$$M^{\text{лин}}(X \mid y) = \underbrace{MX - r_{X,Y}\frac{\sigma_X}{\sigma_Y} MY}_{b_0} + \underbrace{r_{X,Y}\frac{\sigma_X}{\sigma_Y}}_{b_1 = a_{X|y}} y, \; b_1 \neq 0, \tag{11.27}$$

или

$$M^{\text{лин}}(X \mid y) = MX + r_{X,Y}\frac{\sigma_X}{\sigma_Y}(y - MY);$$

— **дисперсия** линейной регрессии X на Y имеет вид

$$\sigma_{LR\,X|Y}^2 = D[M^{\text{лин}}(X \mid Y)] =$$
$$= M[M^{\text{лин}}(X \mid Y) - MX]^2 = \sigma_X^2 r_{X,Y}^2; \tag{11.28}$$

— дисперсия ошибки линейной регрессии X на Y, возникающей при использовании случайной величины $M^{\text{лин}}(X \mid Y)$ вместо X, такова:

$$\sigma^2_{ELR\,X|Y} = D[X - M^{\text{лин}}(X \mid Y)] =$$

$$= M[X - M^{\text{лин}}(X \mid Y)]^2 = \sigma^2_X(1 - r^2_{X,Y}); \qquad (11.29)$$

— дисперсионное тождество для линейной регрессии X на Y имеет вид

$$\sigma^2_{LR\,X|Y} + \sigma^2_{ELR\,X|Y} = \sigma^2_X; \qquad (11.30)$$

— случайная величина $M^{\text{лин}}(X \mid Y)$, определяемая равенством (11.27) при $y = Y$, обладает «свойством минимальности» (аналогичным рассмотреному выше), позволяющим использовать значение $M^{\text{лин}}(X \mid y)$ в качестве приближения к значениям x_y величины X при Y, равном y.

Свойства коэффициента парной корреляции

Прежде чем рассматривать свойства коэффициента корреляции, обратим еще раз внимание на следующее:

— $M(Y \mid X)$ — случайная величина, которая равна числу $M(Y \mid x)$ — математическому ожиданию случайной величины Y при условии, что X примет значение x; функцию $M(Y \mid x) = \varphi(x)$ называют регрессией Y на x;

— $M^{\text{лин}}(Y \mid X)$ — случайная величина, линейно зависящая от X, числовое значение $M^{\text{лин}}(Y \mid x)$ которой при X, равном x, рассчитывается по уравнению (11.20), называемому линейной регрессией Y на x. Для этой величины всегда выполняются соотношения (11.22)—(11.26) независимо от того, является ли в действительности регрессия линейной, т. е. имеет ли место равенство $M(Y \mid x) = M^{\text{лин}}(Y \mid x)$, где $M^{\text{лин}}(Y \mid x) = a_0 + a_1 x$, $a_1 \neq 0$, или, иначе, является ли в действительности корреляционная зависимость Y от X линейной или нет.

Аналогичные замечания относятся к величинам $M(X \mid Y)$ и $M^{\text{лин}}(X \mid Y)$, связанной с Y соотношением (11.27) при $y = Y$.

Сказанное позволяет без ограничения общности использовать соотношения (11.23), (11.25), (11.26), как и соотношения (11.28)—(11.30), при доказательстве свойств коэффициента корреляции.

Приведем свойства коэффициента парной корреляции.

1^0. *Абсолютная величина коэффициента корреляции*

$$|r_{X,Y}| \leqslant 1, \text{ или } -1 \leqslant r_{X,Y} \leqslant 1. \qquad (11.31)$$

≫ Из (11.23) получим

$$r^2_{X,Y} = \sigma^2_{LR\,Y|X} / \sigma^2_Y ,$$

или, учитывая (11.26), имеем,

$$r^2_{X,Y} = \frac{\sigma^2_{LR\,Y|X}}{\sigma^2_{LR\,Y|X} + \sigma^2_{\bar{E}LR\,Y|X}} ;$$

но так как в этой дроби числитель не больше знаменателя, то $r^2_{X,Y} \leqslant 1$. Поэтому $|r_{X,Y}| \leqslant 1$, или $-1 \leqslant r_{X,Y} \leqslant 1$. ≪

2^0. *Условие* $r_{X,Y} = 0$ *является достаточным, но не необходимым условием отсутствия линейной корреляционной зависимости между случайными величинами* X *и* Y, *т. е. из того, что* $r_{X,Y} = 0$ *следует отсутствие линейной корреляционной зависимости между* X *и* Y, *но из отсутствия линейной корреляционной зависимости не следует, что* $r_{X,Y} = 0$.

≫ Пусть $r_{X,Y} = 0$. Докажем, что линейная корреляционная зависимость между X и Y отсутствует. Допустим, что корреляционная зависимость между X и Y существует и является линейной. Это означает, что линейный характер имеет корреляционная зависимость Y от X, т. е. $M(Y \mid x) = M^{\text{лин}}(Y \mid x)$, где $M^{\text{лин}}(Y \mid x) = a_0 + a_1 x$ и $a_1 \neq 0$ [см. (11.20)], и корреляционная зависимость X от Y, т. е. $M(X \mid y) = M^{\text{лин}}(X \mid y)$, где $M^{\text{лин}}(X \mid y) = b_0 + b_1 y$ и $b_1 \neq 0$ [см. (11.27)]. Но из того, что $a_1 \neq 0$, так же как из того, что $b_1 \neq 0$, следует, что $r_{X,Y} \neq 0$, а по условию $r_{X,Y} = 0$.

Итак, если $r_{X,Y} = 0$, то линейной корреляционной зависимости между X и Y быть не может. Отсутствие линейной корреляционной зависимости Y от X означает, что:

— либо корреляционная зависимость Y от X как таковая отсутствует, т. е. условное среднее $M(Y \mid x)$ при изменении значений x величины X неизменно: $M(Y \mid x) = \text{const}$; в этом случае график функции регрессии Y на x представляет собой прямую линию, параллельную оси Ox, все точки с координатами $(x; M(Y \mid x))$ лежат на этой прямой;

— либо корреляционная зависимость Y от X имеет место, т. е. условное среднее $M(Y \mid x)$ изменяется при изменении x, но эта зависимость является нелинейной функцией от x; в этом случае график функции регрессии $M(Y \mid x) = \varphi(x)$ не является прямой линией.

Какой из этих случаев имеет место при $r_{X,Y} = 0$, выясняется ниже при рассмотрении свойств корреляционного отношения.

Аналогично толкование выражения «отсутствие линейной корреляционной зависимости X от Y».

Подтверждением того, что из отсутствия линейной корреляционной зависимости между X и Y не следует, что $r_{X,Y} = 0$, является пример 11.1, в котором корреляционная зависимость Y от X нелинейная, или, иначе, функция регрессии Y на x нелинейна (ее график изображен на рисунке 11.1, ломаная линия), но при этом $r_{X,Y} = 0{,}513 \neq 0$. ≪

Итак, если $r_{X,Y} = 0$, то корреляционная зависимость между X и Y не может быть линейной. Однако из равенства $r_{X,Y} = 0$ не следует, что случайные величины X и Y независимы, хотя из независимости величин X и Y следует, что $r_{X,Y} = 0$. Последнее утверждение и составляет следующее свойство коэффициента корреляции.

3^0. *Условие* $r_{X,Y} = 0$ *является необходимым, но не достаточным условием независимости случайных величин* X *и* Y, *т. е. из независимости этих величин следует, что* $r_{X,Y} = 0$, *но из того, что* $r_{X,Y} = 0$ *не всегда следует независимость* X *и* Y.

≫ Пусть X и Y — независимые случайные величины. В соответствии со свойством (4.20) математического ожидания, для независимых величин X и Y имеет место равенство $M(XY) = MX \cdot MY$, но тогда $M(XY) - MX \cdot MY = 0$ и коэффициент корреляции

$$r_{X,Y} \underset{(11.4)}{=} \frac{M(XY) - MX \cdot MY}{\sigma_X \sigma_Y} = 0.$$

Подтверждением того, что из условия $r_{X,Y} = 0$ не всегда следует независимость величин X и Y, является следующий пример.

ПРИМЕР 11.2. Рассмотрим случайную величину X, имеющую ряд распределения

x	-1	0	1	
$P(X = x)$	p_1	$1 - 2p_1$	p_1	$\Sigma = 1$

и величину $Y = X^2$. Таблица распределения двумерной величины (X, Y) такова:

$y = x^2$ \ x	-1	0	1
$(-1)^2$	p_1	0	0
0	0	$1 - 2p_1$	0
1^2	0	0	p_1

Действительно, если $x = -1$, то

$$y = (-1)^2 = 1 \text{ и } P[(X = -1) \cap (Y = (-1)^2)] = P(X = -1) = p_1;$$

если $x = 0$, то

$$y = 0^2 = 0 \text{ и } P[(X = 0) \cap (Y = 0)] = P(X = 0) = 1 - 2p_1;$$

и наконец, если $x = 1$, то

$$y = 1^2 = 1 \text{ и } P[(X = 1) \cap (Y = 1^2)] = P(X = 1) = p_1.$$

Из таблицы распределения получим

$$M(XY) = (-1)\cdot(-1)^2 p_1 + 0\cdot 0(1-2p_1) + 1\cdot 1^2 p_1 = 0,$$

а так как $MX = 0$, то

$$r_{X,Y} = \frac{M(XY) - MX\cdot MY}{\sigma_X \sigma_Y} = 0. \quad \blacktriangleleft$$

Итак, $r_{X,Y} = 0$, а зависимость Y от X функциональная: $Y = X^2$, т. е. величины X и Y не являются независимыми. $\ll$

Величины X и Y, коэффициент корреляции между которыми $r_{X,Y} = 0$, называются ***некоррелированными***. Из доказанного свойства вытекает, что для независимых величин X и Y коэффициент корреляции $r_{X,Y} = 0$ (независимость влечет некоррелированность), однако из равенства $r_{X,Y} = 0$ не следует независимость величин X и Y (некоррелированность не влечет независимость). Поэтому интерпретировать коэффициент корреляции $r_{X,Y}$ как меру зависимости случайных величин X и Y, вообще говоря, нельзя.

Согласно свойству 1^0, $0 \leqslant |r_{X,Y}| \leqslant 1$, или $-1 \leqslant r_{X,Y} \leqslant 1$; согласно свойству 2^0, условие $r_{X,Y} = 0$ влечет отсутствие линейной корреляционной зависимости между X и Y. Ответ на вопрос, что означает условие $|r_{X,Y}| = 1$, а также каков смысл знака коэффициента корреляции, дает следующее свойство коэффициента корреляции.

4^0. Условие $|r_{X,Y}| = 1$ *является достаточным и необходимым условием линейкой зависимости между величинами X и Y.*

Замечание. Выражение «линейная зависимость между величинами X и Y» означает, что Y линейно зависит от X и X линейно зависит от Y. Но поскольку из линейной зависимости Y от X следует линейная зависимость X от Y, и наоборот, утверждение, сформулированное в свойстве 4^0, эквивалентно утверждению: «условие $|r_{X,Y}| = 1$ является достаточным и необходимым для линейной зависимости величины Y от величины X».

$\gg$ Достаточность. Пусть $|r_{X,Y}| = 1$. Докажем, что Y линейно зависит от X.

Действительно, если $|r_{X,Y}| = 1$, то $1 - r^2_{X,Y} = 0$ и из равенства (11.25) получаем $M[Y - M^{\text{лин}}(Y|X)]^2 = 0$, что возможно только при $Y = M^{\text{лин}}(Y|X)$, или, если принять во внимание (11.20), при $Y = a_0 + a_1 X$, где $a_1 \neq 0$, что и требовалось доказать.

Обратим внимание на следующее:

— если $r_{X,Y} = 1$, то $a_1 = \sigma_Y/\sigma_X$ [см. (11.18)], и

$$Y = a_0 + \frac{\sigma_Y}{\sigma_X} X, \text{ где } a_0 = MY - \frac{\sigma_Y}{\sigma_X} MX. \qquad (11.32)$$

— если $r_{X,Y} = -1$, то $a_1 = -\sigma_Y/\sigma_X$, и

$$Y = a_0 - \frac{\sigma_Y}{\sigma_X} X, \text{ где } a_0 = MY + \frac{\sigma_Y}{\sigma_X} MX. \qquad (11.33)$$

Графики прямых (11.32) и (11.33) изображены на рисунке 11.2, *а* (самостоятельно докажите перпендикулярность этих прямых).

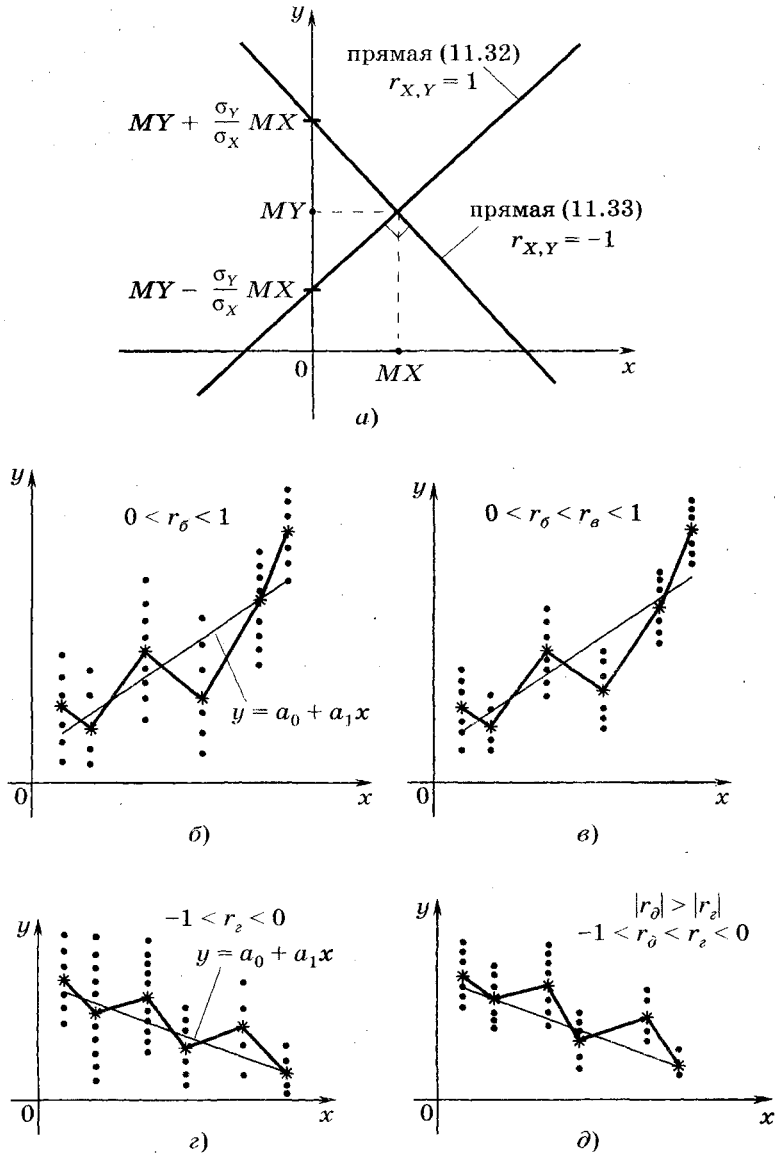


Рис. 11.2

Необходимость. Пусть Y линейно зависит от X, т. е. $Y = c_0 + c_1X$, $c_1 \neq 0$. Докажем, что $|r_{X,Y}| = 1$.

Действительно, $MY = M(c_0 + c_1X) = c_0 + c_1MX$; $DY = D(c_0 + c_1X) = c_1^2 DX$; $\sigma_Y = |c_1|\sigma_X$. Тогда

$$r_{X,Y} \underset{(11.3)}{=} \frac{M[(X - MX)(Y - MY)]}{\sigma_X\sigma_Y} =$$

$$= \frac{M[(X - MX)(c_0 + c_1X - c_0 - c_1MX)]}{\sigma_X|c_1|\sigma_X} = \frac{M(c_1[X - MX]^2)}{|c_1|\sigma_X^2} =$$

$$= \frac{c_1\sigma_X^2}{|c_1|\sigma_X^2} = \frac{c_1}{|c_1|} = \begin{cases} 1, & \text{если } c_1 > 0, \\ -1, & \text{если } c_1 < 0, \end{cases}$$

что и требовалось доказать. Итак, если $Y = c_0 + c_1X$, где $c_1 \neq 0$, то $|r_{X,Y}| = 1$.

Убедимся в том, что прямая $Y = c_0 + c_1X$, $c_1 \neq 0$, совпадает либо с прямой, определяемой (11.32), либо с прямой, определяемой (11.33). Действительно, если $Y = c_0 + c_1X$, $c_1 \neq 0$, то $|r_{X,Y}| = 1$, но при $|r_{X,Y}| = 1$ уравнение линейной зависимости Y от X имеет вид (11.32) или вид (11.33), что определяется знаком коэффициента $r_{X,Y}$; следовательно, прямая $Y = c_0 + c_1X$ совпадает с прямой (11.32), если $r_{X,Y} = 1$, или с прямой (11.33), если $r_{X,Y} = -1$. ≪

Таким образом, согласно свойству 1^0, $0 \leqslant |r_{X,Y}| \leqslant 1$, при этом:

если $r_{X,Y} = 0$, то, согласно свойству 2^0, линейная корреляционная зависимость между X и Y невозможна, т. е. условные математические ожидания — условные средние $M(Y \mid x)$ и $M(X \mid y)$ — не могут линейно зависеть соответственно от x и от y, или, иначе, стохастическая зависимость Y от X, так же как и стохастическая зависимость X от Y, даже в среднем не является линейной. Естественно «линейность такой стохастической зависимости» оценить числом ноль, что совпадает со значением коэффициента корреляции $r_{X,Y} = 0$;

если же $|r_{X,Y}| = 1$, и только в этом случае, зависимость между X и Y линейная функциональная, т. е. стохастическая зависимость между X и Y (тем более корреляционная зависимость — зависимость в среднем) «трансформируется» в абсолютно линейную зависимость между X и Y. Естественно «линейность такой трансформированной стохастической зависимости» оценить числом единица, что совпадает со значением $|r_{X,Y}| = 1$.

Таким образом, $|r_{X,Y}|$ можно *интерпретировать как меру линейности зависимости между X и Y*: чем отчетливее проявляется линейность в зависимости между X и Y,

тем больше $|r_{X,Y}|$, и наоборот, чем больше $|r_{X,Y}|$, тем отчетливее проявление линейности в зависимости между X и Y. При этом, если эта «линейность» в среднем возрастающая, т. е. при увеличении значений одной величины значения другой увеличиваются или имеют тенденцию к увеличению, и только в этом случае, то $r_{X,Y} > 0$. Если эта «линейность» в среднем убывающая, т. е. при увеличении значений одной величины значения другой уменьшаются или имеют тенденцию к уменьшению, и только в этом случае, то $r_{X,Y} < 0$.

Графическая иллюстрация $|r_{X,Y}|$ как меры линейности зависимости между X и Y дана на рисунке 11.2, *б*—*д*. При этом для наглядности будем изучать проявление линейности лишь в зависимости Y от X и предполагать, что на каждой диаграмме разброса точки (x, y) имеют одинаковые «массы», или, иначе, вероятности $P[(X = x) \cap (Y = y)]$ значений (x, y) величины (X, Y) одинаковы. В этом случае линией регрессии Y на x — линией, соединяющей точки с координатами $(x, M(Y \mid x))$, будет линия, соединяющая «серединные значения игреков». На графиках линии регрессии изображены ломаными линиями.

Сравним рисунки 11.2, *б* и 11.2, *в*. На обоих рисунках при увеличении значений величины X значения величины Y имеют тенденцию к увеличению, поэтому коэффициенты корреляции $r_б > 0$ и $r_в > 0$, но на втором рисунке зависимость Y от X ближе к линейной: концентрация точек около прямой

$$M^{\text{лин}}(Y \mid x) = a_0 + a_1 x,\ a_1 > 0;$$

выше, чем на первом рисунке, или, иначе, на рисунке 11.2, *в* линейность проявляется в большей мере, чем на рис. 11.2, *б*, поэтому $|r_в| > |r_б|$, или, учитывая, что $r_б > 0$ и $r_в > 0$, $r_в > r_б$.

Сравним рисунки 11.2, *г* и *д*. На обоих рисунках при увеличении значений величины X значения величины Y имеют тенденцию к уменьшению, поэтому коэффициенты корреляции $r_г < 0$ и $r_д < 0$, но на втором рисунке концентрация точек около прямой $M^{\text{лин}}(Y \mid x) = a_0 + a_1 x$, $a_1 < 0$, выше, чем на первом, или, иначе, на рисунке 11.2, *д* линейность проявляется в большей мере, чем на рисунке 11.2, *г*, поэтому $|r_д| > |r_г|$, или, учитывая, что $r_г < 0$ и $r_д < 0$, $r_д < r_г$.

Проведенные сравнения позволяют дать графическое толкование выражения «более четкое проявление линейности в зависимости между X и Y», или «проявление линейности в зависимости между X и Y в большей мере»:

они означают, что диаграмма разброса принимает форму более узкого облака точек, вытянутого вдоль прямой, угловой коэффициент которой отличен от нуля.

В примере 11.1 (см. табл. 11.8 и рис. 11.1) $r_{X,Y} = 0{,}513$. Судя по этому числу, можно лишь отметить умеренное проявление линейности в зависимости между X и Y (хотя при этом корреляционная зависимость Y от X близка к линейной: ломаная линия, изображающая график функции регрессии Y на x, близка к прямой), а также то, что с ростом значений одной величины значения другой имеют тенденцию к увеличению.

Более четкое числовое содержание, по сравнению с $r_{X,Y}$, имеет квадрат коэффициента корреляции $r^2_{X,Y}$ — его называют ***коэффициентом линейной детерминации*** случайной величины Y случайной величиной X или, в силу того, что $r_{X,Y} = r_{X,Y}$, величины X величиной Y. Или короче $r^2_{X,Y}$ — коэффициент линейной детерминации одной из случайных величин, все равно какой, другой величиной. Из соотношений (11.23) и (11.28) соответственно получим

$$r^2_{X,Y} = \sigma^2_{LR\,Y|X} / \sigma^2_Y = D[M^{\text{лин}}(Y \mid X)] / \sigma^2_Y \qquad (11.34)$$

и

$$r^2_{X,Y} = \sigma^2_{LR\,X|Y} / \sigma^2_X = D[M^{\text{лин}}(X \mid Y)] / \sigma^2_X, \qquad (11.35)$$

т. е. коэффициент линейной детерминации $r^2_{X,Y}$ показывает, какую долю дисперсии одной случайной величины составляет дисперсия математического ожидания этой величины, при условии его линейной зависимости от другой случайной величины, или, иначе, какая доля средней колеблемости значений одной случайной величины около ее математического ожидания объясняется корреляционной зависимостью этой величины от другой при условии, что эта зависимость линейна (т. е. функция регрессии линейная).

▶ **ПРИМЕР 11.1** (продолжение). В примере коэффициент линейной детерминации $r^2_{X,Y} = 0{,}513^2 = 0{,}263$, т. е. 26,3% колеблемости значений величины Y (величины X) объясняется ее корреляционной зависимостью от величины X (величины Y) при условии линейности этой зависимости. ◀

11.2.2. Корреляционное отношение и его свойства. На рисунке 11.3, *е*, *ж* изображены диаграммы разброса двух двумерных случайных величин: точки с координатами (x, y) — значения величины (X, Y), при этом для наглядности «массы точек», или вероятности $P[(X = x) \cap (Y = y)]$

приняты одинаковыми. И на первом, и на втором рисунке зависимость между X и Y далека от линейной, поэтому мера линейности зависимости, или $|r_{X,Y}|$, близко к нулю. Однако на рисунке 11.3, *ж* точки сконцентрированы в более узкой «параболической» полосе, чем на рисунке 11.3, *е*; на рисунке 11.3, *ж* зависимость Y от X ближе к функциональной, а именно к параболической, чем на рисунке 11.3, *е*. Мерой близости зависимости Y от X к функциональной зависимости, или, иначе, мерой «функциональности зависимости Y от X» (при этом ограничения на вид функции не вводятся) является *корреляционное отношение*, которое обозначают символом $\rho_{Y|X}$. Мерой близости зависимости X от Y к функциональной является $\rho_{X|Y}$ — корреляционное отношение X на Y.

Построим формулу корреляционного отношения Y на X и рассмотрим его свойства, показывающие, что приближение зависимости Y от X к функциональной сопровождается увеличением $\rho_{Y|X}$. Пусть каждому значению x величины X соответствует множество значений случайной величины Y, т. е. зависимость Y от X стохастическая, и с изменением x меняется условное математическое ожидание $M(Y|x)$, т. е. имеет место и корреляционная зависимость Y от X. Как и прежде, функцию $M(Y|x) = \varphi(x)$, описывающую изменение условного математического ожидания величины Y при изменении значений величины X, назовем функцией регрессии Y на x. Так как X — случайная величина, то и зависящая от нее величина $M(Y|X) = \varphi(X)$ тоже случайная. Можно доказать, что при любом виде функции регрессии, или любом виде корреляционной зависимости Y от X выполняется дисперсионное тождество, аналогичное дисперсионному тождеству (11.26) для линейной регрессии Y на X, а именно

$$\sigma^2_{RY|X} + \sigma^2_{ERY|X} = \sigma^2_Y, \qquad (11.36)$$

или

$$D[M(Y|X)] + D[Y - M(Y|X)] = DY,$$

где:

$\sigma^2_Y = DY = M(Y - MY)^2$ — дисперсия величины Y, характеризующая средний разброс ее значений около MY (причины этого разброса — влияние на Y величины X и других величин, помимо X);

$\sigma^2_{RY|X} = D[M(Y|X)] = M[M(Y|X) - MY]^2$ — дисперсия регрессии Y на X, характеризующая средний разброс значений условного математического ожидания $M(Y|X)$ около MY (причина этого разброса — корреляционная зависи-

мость Y от X, т. е. изменение $M(Y \mid x)$ при изменении значений x величины X);

$\sigma^2_{ER\,Y|X} = D[Y - M(Y \mid X)] = M[Y - M(Y \mid X)]^2$ — дисперсия ошибки регрессии Y на X, равная среднему квадрату отклонения величины Y от $M(Y \mid X)$ (причина этого отклонения — влияние на Y других, помимо X, величин)[1].

Для дисперсии ошибки регрессии Y на X имеет место следующее соотношение:

$$D[Y - M(Y \mid X)] = M[D(Y \mid X)],$$

или

$$\sigma^2_{ER\,Y|X} = M[D(Y \mid X)], \qquad (11.37)$$

т. е. дисперсия ошибки регрессии равна среднему значению условной дисперсии.

≫ Приведем доказательство равенства (11.37) на примере дискретной двумерной величины:

$$\sigma^2_{ER\,Y|X} = D[Y - M(Y \mid X)] = M[Y - M(Y \mid X)]^2 =$$

$$= \sum_i \sum_j [y_j - M(Y \mid x_i)]^2 P[(X = x_i) \quad (Y = y_j)] =$$

$$= \sum_i \sum_j [y_j - M(Y \mid x_i)]^2 P(X = x_i) P[(Y = y_j) \mid (X = x_i)] =$$

$$= \sum_i P(X = x_i) \sum_j [y_j - M(Y \mid x_i)]^2 P[(Y = y_j) \mid (X = x_i)] =$$

$$= \sum_i P(X = x_i) D(Y \mid (X = x_i)) = M[D(Y \mid X)]. \quad \ll$$

Учитывая (11.37), дисперсионное тождество (11.36)

$$\sigma^2_{R\,Y|X} + \sigma^2_{ER\,Y|X} = \sigma^2_Y,$$

можно записать так:

$$D[M(Y \mid X)] + M[D(Y \mid X)] = DY. \qquad (11.38)$$

Корреляционное отношение Y на X — это неотрицательное число

$$\rho_{Y|X} = \sqrt{\frac{\sigma^2_{R\,Y|X}}{\sigma^2_Y}} = \sqrt{\frac{D[M(Y \mid X)]}{DY}}, \qquad \rho_{Y|X} \geqslant 0. \qquad (11.39)$$

Учитывая, что, согласно (11.36), $\sigma^2_{R\,Y|X} = \sigma^2_Y - \sigma^2_{ER\,Y|X}$, или $D[M(Y \mid X)] = DY - M[D(Y \mid X)]$, получим

$$\rho_{Y|X} = \sqrt{1 - \frac{\sigma^2_{ER\,Y|X}}{\sigma^2_Y}} = \sqrt{1 - \frac{M[D(Y \mid X)]}{DY}}. \qquad (11.40)$$

[1] Индекс «R» от англ. *regression* — регрессия; индекс «ER» от англ. *error of regression* — ошибка регрессии.

▶ **ПРИМЕР 11.1** (продолжение). Проверим тождество (11.38). В условиях примера $\sigma_Y^2 = 0{,}0336$ (см. таблицу 11.3). Используя данные таблицы 11.8, найдем:

$$\sigma^2_{R\,Y|X} = D[M(Y \mid X)] = M[M(Y \mid X) - MY]^2 =$$

$$= (0{,}5 - 0{,}68)^2 \cdot 0{,}2 + (0{,}686 - 0{,}68)^2 \cdot 0{,}42 +$$

$$+ (0{,}768 - 0{,}68)^2 \cdot 0{,}38 = 0{,}0095;$$

$$\sigma^2_{ER\,Y|X} \underset{(11.37)}{=} M[D(Y \mid X)] = 0{,}03 \cdot 0{,}2 + 0{,}03265 \cdot 0{,}42 +$$

$$+ 0{,}01163 \cdot 0{,}38 = 0{,}0241.$$

В результате: $0{,}0095 + 0{,}0241 = 0{,}0336$.

Корреляционное отношение

$$\rho_{Y|X} = \sqrt{\sigma^2_{R\,Y|X}/\sigma^2_Y} = \sqrt{0{,}0095/0{,}0336} = 0{,}532.$$

Ответы на вопросы, много это или мало и каков смысл выражения «корреляционное отношение $\rho_{Y|X} = 0{,}532$», дадим после рассмотрения свойств корреляционного отношения. Здесь обратим внимание на то, что в условиях примера $r_{X,Y} = 0{,}513$ и что $\rho_{Y|X} > |r_{X,Y}|$. ◀

Докажем, что всегда $\rho_{Y|X} \geqslant |r_{X,Y}|$. Это неравенство вытекает из теоремы 4.

Теорема 4. *Математическое ожидание случайной ошибки, равной* $M(Y \mid X) - M^{\text{лин}}(Y \mid X)$, *возникающей при замене случайной величины* $M(Y \mid X)$ *величиной* $M^{\text{лин}}(Y \mid X)$ *равно нулю*:

$$M[M(Y \mid X) - M^{\text{лин}}(Y \mid X)] = 0, \qquad (11.41)$$

а ее дисперсия

$$D[M(Y \mid X) - M^{\text{лин}}(Y \mid X)] = M[M(Y \mid X) -$$

$$- M^{\text{лин}}(Y \mid X)]^2 = \sigma_Y^2(\rho^2_{Y|X} - r^2_{X,Y}). \qquad (11.42)$$

Замечания. 1. Замена случайной величины $M(Y \mid X)$ величиной $M^{\text{лин}}(Y \mid X)$ означает, что при любом значении x величины X соответствующее ему значение $M(Y \mid x)$ условного математического ожидания заменяют значением $M^{\text{лин}}(Y \mid x)$, рассчитанным по уравнению (11.20).

2. Можно доказать, что среднее квадрата ошибки, возникающей при замене $M(Y \mid X)$ любой величиной, линейно зависящей от X, но отличной от $M^{\text{лин}}(Y \mid X) = a_0 + a_1 X$, где a_0 и a_1 определяются по формулам (11.19) и (11.18), больше $M[(Y \mid X) - M^{\text{лин}}(Y \mid x)]^2$; говорят, что равная этому математическому ожиданию дисперсия $D[M(Y \mid X) - M^{\text{лин}}(Y \mid X)]$ обладает свойством минимальности.

≫ Докажем (11.41). Имеем

$$M[M(Y\mid X)-M^{\text{лин}}(Y\mid X)]=$$

$$=M[M(Y\mid X)]-M[M^{\text{лин}}(Y\mid X)]\underset{(11.2)\ (11.13)}{=}MY-MY=0.$$

Докажем (11.42). Учитывая (11.41), получим

$$D[M(Y\mid X)-M^{\text{лин}}(Y\mid X)]=M[M(Y\mid X)-M^{\text{лин}}(Y\mid X)]^2.$$

Найдем эту дисперсию. Имеем

$$D[\underbrace{M(Y\mid X)}_{Z}-M^{\text{лин}}(Y\mid X)]\underset{(11.20)}{=}D(Z-a_0-a_1X)\underset{a_0=\text{const}}{=}$$

$$\underset{a_0=\text{const}}{=}D(Z-a_1X)\underset{(4.61)}{=}DZ+D(a_1X)-2r_{Z,a_1X}\sigma_Z\sigma_{a_1X}\underset{(11.6)}{=}$$

$$\underset{(11.6)}{=}DZ+D(a_1X)-2K(Z,a_1X)\underset{(11.5)}{=}$$

$$\underset{(11.5)}{=}DZ+D(a_1X)-2[M(Za_1X)-MZ\cdot M(a_1X)]\underset{a_1=\text{const}}{=}$$

$$\underset{a_1=\text{const}}{=}DZ+a_1^2DX-2a_1[M(ZX)-MZ\cdot MX]=$$

$$=D[M(Y\mid X)]+a_1^2\sigma_X^2-2a_1\Big[\underbrace{M(M(Y\mid X)\cdot X)}_{M(XY)(\text{см.}(11.17))}-\underbrace{M(M(Y\mid X))}_{MY(\text{см.}(11.2))}MX\Big]\underset{(11.39)}{=}$$

$$\underset{(11.39)}{=}\sigma_Y^2\rho_{Y\mid X}^2+a_1^2\sigma_X^2-2a_1[M(XY)-MY\cdot MX]\underset{(11.6)}{=}$$

$$\underset{(11.6)}{=}\sigma_Y^2\rho_{Y\mid X}^2+a_1^2\sigma_X^2-2a_1r_{X,Y}\sigma_X\sigma_Y\underset{(11.18)}{=}$$

$$\underset{(11.18)}{=}\sigma_Y^2\rho_{Y\mid X}^2+r_{X,Y}^2\frac{\sigma_Y^2}{\sigma_X^2}\sigma_X^2-2r_{X,Y}\frac{\sigma_Y}{\sigma_X}r_{X,Y}\sigma_X\sigma_Y=$$

$$=\sigma_Y^2(\rho_{Y\mid X}^2-r_{X,Y}^2).\quad\ll$$

Следствие. *Корреляционное отношение не меньше абсолютной величины коэффициента корреляции*:

$$\rho_{Y\mid X}\geqslant|r_{X,Y}|,\quad\rho_{Y\mid X}^2\geqslant r_{X,Y}^2.\qquad(11.43)$$

≫ Рассмотрим равенство (11.42). Левая часть этого равенства неотрицательна, поэтому и его правая часть неотрицательна, т. е. $\sigma_Y^2(\rho_{Y\mid X}^2-r_{X,Y}^2)\geqslant 0$, следовательно, $\rho_{Y\mid X}^2\geqslant r_{X,Y}^2$ (дисперсия σ_Y^2 случайной величины всегда положительна). Учитывая, что, согласно (11.39), $\rho_{Y\mid X}$ — неотрицательное число, а $r_{X,Y}$ может быть и отрицательным, из последнего неравенства получим $\rho_{Y\mid X}\geqslant|r_{X,Y}|$. ≪

Свойства корреляционного отношения

$\underline{1}^0$. $0 \leqslant \rho_{Y|X} \leqslant 1$.

≫ Согласно (11.39), $\rho_{Y|X} \geqslant 0$; из (11.36) следует, что $\sigma^2_{R\,Y|X} \leqslant \sigma^2_Y$, поэтому $\rho_{Y|X} = \sqrt{\sigma^2_{R\,Y|X}/\sigma^2_Y} \leqslant 1$. В результате получаем $0 \leqslant \rho_{Y|X} \leqslant 1$. ≪

$\underline{2}^0$. *Условие* $\rho_{Y|X} = 0$ *является достаточным и необходимым условием отсутствия корреляционной зависимости* Y *от* X, *т. е. условием равенства* $M(Y \mid x) = \text{const}$ *при любом допустимом значении* x *величины* X, *а именно условием равенства* $M(Y \mid x) = MY$.

≫ Достаточность. Пусть $\rho_{Y|X} = 0$. Тогда по формуле (11.39) имеем $D[M(Y \mid X)] = 0$, или $M[M(Y \mid X) - MY]^2 = 0$. Из этого равенства следует, что $M(Y \mid X) = MY$, т. е. условное математическое ожидание $M(Y \mid x)$ «не реагирует» на изменение значения x величины X. Это и означает, что корреляционная зависимость Y от X отсутствует.

Необходимость. Если отсутствует корреляционная зависимость величины Y от X, то $M(Y \mid x) = \text{const}$ при любом допустимом значении x величины X, поэтому $D[M(Y \mid X)] = 0$, тогда, согласно (11.39), $\rho_{Y|X} = 0$. ≪

Графическая иллюстрация свойства дана на рисунке 11.3, *а*. Линией регрессии Y на x (линией, соединяющей при каждом x «серединные значения игреков») является прямая $y = MY$.

Обратим внимание на следующее: если $\rho_{Y|X} = 0$, то и $r_{X,Y} = 0$ (так как всегда $\rho_{Y|X} \geqslant |r_{X,Y}|$). Поэтому доказанное свойство можно сформулировать так: условие $\rho_{Y|X} = r_{X,Y} = 0$ является достаточным и необходимым условием отсутствия корреляционной зависимости Y от X, т. е. условием выполнения равенства $M(Y \mid x) = MY$ при любом допустимом значении x величины X.

Следующее свойство позволяет ответить на вопрос о виде корреляционной зависимости, если $\rho_{Y|X} = |r_{X,Y}| > 0$.

$\underline{3}^0$. *Условие* $\rho_{Y|X} = |r_{X,Y}| > 0$ *является достаточным и необходимым условием линейной корреляционной зависимости* Y *от* X, *т. е. условием выполнения равенства*

$$M(Y \mid x) = a_0 + a_1 x,\ a_1 \neq 0,$$

при любом допустимом значении x *величины* X.

≫ Достаточность. Пусть $\rho_{Y|X} = |r_{X,Y}| > 0$, или $\rho^2_{Y|X} = r^2_{X,Y} > 0$. Если $\rho^2_{Y|X} = r^2_{X,Y}$, то правая часть равенства (11.42) равна нулю, поэтому и его левая часть

$$M[M(Y \mid X) - M^{\text{лин}}(Y \mid X)]^2 = 0,$$

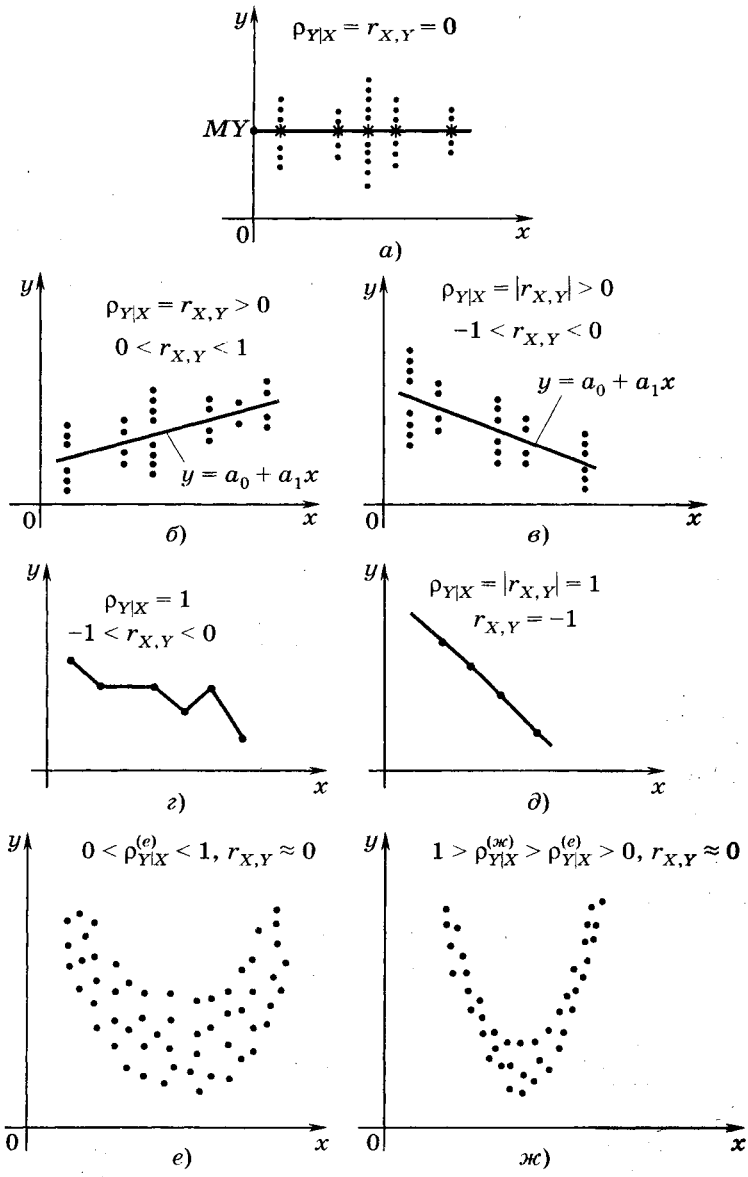


Рис. 11.3

что возможно только если при любом допустимом значении x величины X $M(Y \mid x) = M^{\text{лин}}(Y \mid x)$, где, согласно (11.20), $M^{\text{лин}}(Y \mid x) = a_0 + a_1 x$. Так как по условию $r_{X,Y} \neq 0$, то и $a_1 \neq 0$. Окончательно получаем

$$M(Y \mid x) = M^{\text{лин}}(Y \mid x) = a_0 + a_1 x,\ a_1 \neq 0,$$

т. е. корреляционная зависимость Y от X является линейной.

Необходимость. Пусть корреляционная зависимость Y от X линейная, т. е.

$$M(Y \mid X) = M^{\text{лин}}(Y \mid X) = a_0 + a_1 X,$$

где $a_1 \neq 0$ [см. (11.20)]. Так как $a_1 \neq 0$, то и $r_{X,Y} \neq 0$. Поскольку $M(Y \mid X) = M^{\text{лин}}(Y \mid X)$, левая часть равенства (11.42) равна нулю, но тогда и его правая часть $\sigma_Y^2 (\rho_{Y|X}^2 - r_{X,Y}^2) = 0$, что возможно только при $\rho_{Y|X}^2 = r_{X,Y}^2$. В результате получаем $\rho_{Y|X} = |r_{X,Y}| > 0$. ≪

Графическая иллюстрация свойства дана на рисунке 11.3, *б*, *в*. На каждом из рисунков линия регрессии Y на x — прямая линия, совпадающая с прямой $y = a_0 + a_1 x$, угловой коэффициент которой $a_1 \neq 0$.

Ответим на вопрос о виде корреляционной зависимости Y от X, если $\rho_{Y|X} > |r_{X,Y}|$.

$\underline{4}^0$. *Условие* $\rho_{Y|X} > |r_{X,Y}|$ *является достаточным и необходимым условием нелинейной корреляционной зависимости* Y *от* X.

≫ Достаточность. Пусть $\rho_{Y|X} > |r_{X,Y}|$. Покажем, что корреляционная зависимость Y от X нелинейная. Доказательство проведем методом от противного. Отрицанием утверждения «корреляционная зависимость нелинейная» является утверждение «корреляционная зависимость отсутствует или она имеет линейный характер». При этом:

— если корреляционная зависимость Y от X отсутствует, то, согласно свойству $\underline{2}^0$ корреляционного отношения, $\rho_{Y|X} = 0$, что противоречит неравенству $\rho_{Y|X} > 0$, которое вытекает из условия $\rho_{Y|X} > |r_{X,Y}|$;

— если корреляционная зависимость Y от X имеет линейный характер, то, согласно свойству $\underline{3}^0$ корреляционного отношения, $\rho_{Y|X} = |r_{X,Y}| > 0$, что противоречит условию $\rho_{Y|X} > |r_{X,Y}|$.

В результате получаем: если $\rho_{Y|X} > |r_{X,Y}|$, то корреляционная зависимость Y от X имеет нелинейный характер.

Необходимость. Пусть корреляционная зависимость Y от X нелинейна. Докажем, используя метод от противного, что $\rho_{Y|X} > |r_{X,Y}|$.

Так как всегда $\rho_{Y|X} \geqslant |r_{X,Y}|$, то отрицанием неравенства $\rho_{Y|X} > |r_{X,Y}|$ является равенство $\rho_{Y|X} = |r_{X,Y}|$ и тогда возможны два варианта:

— либо $\rho_{Y|X} = |r_{X,Y}| = 0$, но если это равенство имеет место, то, согласно свойству $\underline{2}^0$ корреляционного отношения, корреляционная зависимость Y от X должна отсутствовать;

— либо $\rho_{Y|X} = |r_{X,Y}| > 0$, но тогда, согласно свойству $\underline{3}^0$ корреляционного отношения, корреляционная зависимость Y от X должна иметь линейный характер.

Оба варианта противоречат тому, что по условию корреляционная зависимость Y от X нелинейна. ≪

$\underline{5}^0$. *Условие* $\rho_{Y|X} = 1$ *является достаточным и необходимым для функциональной зависимости величины* Y *от* X.

≫ Достаточность. Пусть $\rho_{Y|X} = 1$. Это, в силу (11.40), означает, что $M[D(Y|X)] = 0$. Но так как дисперсия любой величины неотрицательна, то из последнего равенства следует, что $D(Y|x) = 0$ при любом допустимом значении x величины X, а это значит, что при любом допустимом значении x величины X величина Y принимает единственное значение, т. е. зависимость Y от X — функциональная.

Необходимость. Пусть любому фиксированному значению x величины X соответствует только одно значение величины Y. Это означает, что при любом x дисперсия $D(Y|x) = 0$, поэтому и $M[D(Y|X)] = M0 = 0$. Но тогда из (11.40) следует, что $\rho_{Y|X} = 1$. ≪

Графическая иллюстрация свойства дана на рисунках 11.3, *г*, *д*. На каждом из этих рисунков зависимость Y от X функциональная, так как каждому x соответствует единственное y, и поэтому корреляционное отношение $\rho_{Y|X} = 1$. Но на рисунке 11.3, *г* функциональная зависимость Y от X не является линейной и с ростом x значения величины Y имеют тенденцию к уменьшению, поэтому $-1 < r_{X,Y} < 0$; а на рисунке 11.3, *д* функциональная зависимость линейная и с ростом x значения величины Y уменьшаются, поэтому $r_{X,Y} = -1$.

Таким образом, согласно свойству $\underline{1}^0$, $0 \leqslant \rho_{Y|X} \leqslant 1$, при этом:

— если $\rho_{Y|X} = 0$, и только в этом случае, согласно свойству $\underline{2}^0$, корреляционная зависимость Y от X отсутствует, или, иначе, при любом допустимом значении x величины X условное среднее $M(Y|x) = MY$, т. е. даже в среднем Y функционально не зависит от X. В этом случае естественно «проявление функциональной зависимости в стохастической зависимости Y от X» оценить числом 0, что совпадает с минимальным значением корреляционного отношения $\rho_{Y|X} = 0$;

— если $\rho_{Y|X} = 1$, и только в этом случае, согласно свойству $\underline{5}^0$, зависимость Y от X функциональная, т. е. стохастическая зависимость Y от X (тем более, корреляционная зависимость — зависимость в среднем) как бы трансформируется в абсолютно функциональную зависимость Y от X. В этом случае естественно «проявление функциональности в зависимости Y от X» оценить числом 1 — самым большим значением корреляционного отношения $\rho_{Y|X}$.

Таким образом, $\rho_{Y|X}$ можно интерпретировать как *меру проявления функциональной зависимости в зависимости Y от X или меру близости зависимости Y от X к функциональной зависимости*: чем отчетливее проявляется функциональная зависимость в зависимости Y от X, или чем ближе зависимость Y от X к функциональной, тем больше

$\rho_{Y|X}$, и наоборот, чем больше $\rho_{Y|X}$, тем ближе зависимость Y от X к функциональной.

Графическая иллюстрация корреляционного отношения $\rho_{Y|X}$, как меры близости зависимости Y от X к функциональной, дана на рисунках 11.3, *е*, *ж*. На рисунке 11.3, *ж* зависимость Y от X ближе к функциональной (именно к параболической), чем на рисунке 11.3, *е*: на рисунке 11.3, *ж* точки сконцентрированы в более узкой «параболической» полосе, чем на рисунке 11.3, *е*, поэтому корреляционное отношение Y на X на рисунке 11.3, *ж* больше корреляционного отношения на рисунке 11.3, *е*. Значение абсолютной величины коэффициента корреляции, которая является мерой линейности зависимости, на обоих рисунках близко к нулю.

Более четкое числовое содержание, по сравнению с $\rho_{Y|X}$, имеет квадрат корреляционного отношения $\rho^2_{Y|X}$ — его называют ***коэффициентом детерминации*** случайной величины Y случайной величиной X. Из соотношения (11.39) получим

$$\rho^2_{Y|X} = \sigma^2_{RY|X}/\sigma^2_Y = D[M(Y \mid X)]/DY, \qquad (11.44)$$

т. е. коэффициент детерминации $\rho^2_{Y|X}$ показывает, какую долю дисперсии величины Y составляет дисперсия условного математического ожидания $M(Y \mid X)$, или, иначе, какая доля средней колеблимости значений величины Y около MY объясняется корреляционной зависимостью Y от X.

Свойства корреляционного отношения $\rho_{Y|X}$ и коэффициента корреляции $r_{X,Y}$ позволяют отвечать на вопросы, существует ли корреляционная зависимость Y от X или нет, и если такая зависимость существует, то линейная она или нелинейная. Последовательность ответов изображена на рисунке 11.4, где предполагается, что значения $\rho_{Y|X}$ и $r_{X,Y}$ известны. Если известно только числовое значение коэффициента корреляции $r_{X,Y}$, то получить однозначные ответы на поставленные вопросы нельзя; варианты ответов изображены на рисунке 11.5.

Замечание. Если $\rho_{Y|X}$ неизвестно, но $|r_{X,Y}|$ близок к единице, то, в силу (11.43) и свойства $\underline{1}^0$ корреляционного отношения, $\rho_{Y|X}$ близко к единице, и разность $(\rho^2_{Y|X} - r^2_{X,Y})$ близка к нулю. Тогда, согласно (11.42), близко к нулю средее квадратов ошибок, возникающих при замене фактических значений $M(Y \mid x)$ значениями $M^{лин}(Y \mid x)$, рассчитанными по уравнению (11.20).

При $|r_{X,Y}|$, близком к единице, разность $(1 - r^2_{X,Y})$ близка к нулю, и тогда, в силу (11.25), близко к нулю среднее квадратов ошибок, возникающих при замене значений y_x величины Y значениями $M^{лин}(Y \mid x)$.

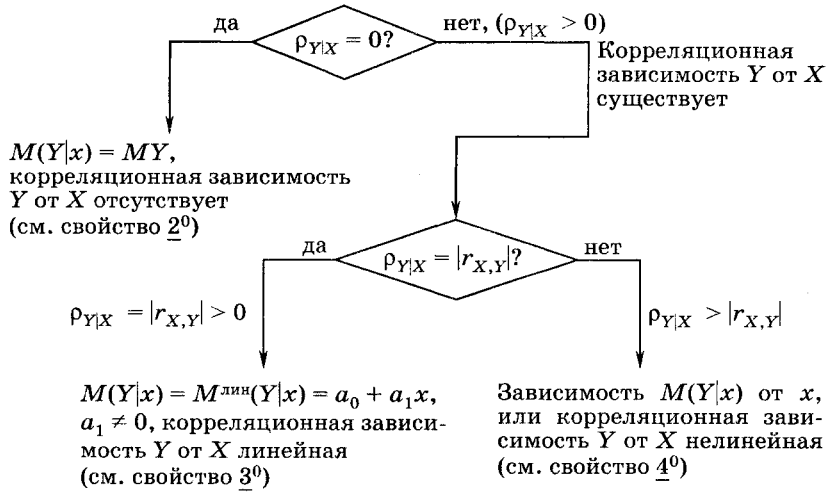


Рис. 11.4

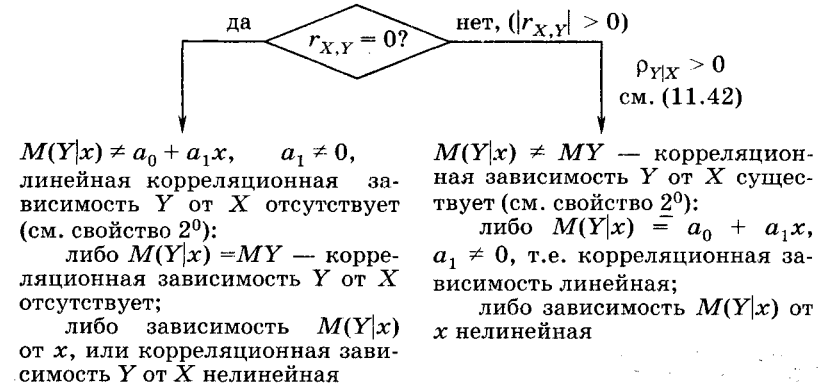


Рис. 11.5

▶ **ПРИМЕР 11.1** (продолжение). В условиях примера $\rho_{Y|X} = 0{,}532 > r_{X,Y} = 0{,}513$. Согласно схеме, изображенной на рисунке 11.4, корреляционная зависимость Y от X нелинейная, что соответствует действительности. На рисунке 11.1 функция регрессии $M(Y \mid x) = \varphi(x)$ изображена ломаной линией.

Судя по значению корреляционного отношения ($\rho_{Y|X} = 0{,}532$), можно отметить лишь очень умеренное проявление в стохастической зависимости Y от X функциональной зависимости; изображенные на диаграмме разброса (см. рисунок 11.1) круги (напомним, площадь круга с центром в точке $(x_i; y_j)$ равна вероятности $P[(X = x_i) \cap (Y = y_j)]$) да-

ют представление об умеренной линейности облака этих точек — кругов.

Далее имеем $\rho^2_{Y|X} = 0{,}532^2 = 0{,}283$, т. е. 28,3% колеблемости значений величины Y объясняется ее корреляционной зависимостью от X. Напомним, $r^2_{X,Y} = 0{,}263$, т. е. 26,3% колеблемости значений величины Y объяснялось бы ее линейной корреляционной зависимостью от X, если таковая имела бы место. ◄

Подобно тому, как было рассмотрено $\rho_{Y|X}$ — корреляционное отношение Y на X, можно рассмотреть $\rho_{X|Y}$ — корреляционное отношение X на Y. Отметим, что, вообще говоря, $\rho_{Y|X} \neq \rho_{X|Y}$, тогда как $r_{X,Y} = r_{Y,X}$. Используя $r_{X,Y}$, мы имеем в виду *меру линейности стохастической зависимости между X и Y*, а используя $\rho_{Y|X}$, — *меру проявления функциональной зависимости в стохастической зависимости Y от X*.

В практических задачах наибольший интерес представляют следующие вопросы:

существует или нет корреляционная зависимость Y от X [изменяются или нет с изменением x условные математические ожидания $M(Y \mid x)$], или, иначе говоря, отлично ли корреляционное отношение $\rho_{Y|X}$ от нуля или равно нулю;

если корреляционная зависимость существует, то какой вид имеет функция регрессии $M(Y \mid x) = \varphi(x)$ (линейный, параболический или какой-либо другой).

Точно ответить на поставленные вопросы можно лишь только в том случае, когда известен закон распределения двумерной величины (X, Y). В примере 11.1 этот закон задан таблицей 11.1, в которой приведены все возможные значения случайных величин X и Y и вероятности совместного появления этих значений. Обычно такими сведениями не располагают; как правило, имеются лишь наблюдавшиеся значения двумерной величины (X, Y). Покажем, как, имея наблюдавшиеся значения, ответить на поставленные выше вопросы.

11.2.3. Первичная обработка результатов наблюдений двумерной случайной величины. Выборочная функция регрессии. Пусть имеется n парных наблюдений двумерной случайной величины (X, Y):

x_k	x_1	x_2	...	x_n
y_k	y_1	y_2	...	y_n

Наблюдавшиеся «иксы» и «игреки» сгруппируем в таблицу 11.9, которую называют ***корреляционной таблицей*** и строят следующим образом:

«иксы» группируют в ряд, число групп которого обозначим ν; если это статистический ряд, то $x'_1, x'_2, \ldots, x'_\nu$ — различающиеся между собой результаты наблюдений, или варианты; если это интервальный ряд, то $x'_1, x'_2, \ldots, x'_\nu$ — центры интервалов;

«игреки» группируют в ряд, число групп которого обозначим q; $y'_1, y'_2, \ldots, y'_q$ — это либо варианты, если ряд статистический, либо центры интервалов, если ряд интервальный;

подсчитывают числа m_{ji} таких наблюдавшихся пар чисел (x, y), у которых x попадает в группу x'_i, а y — в группу y'_j, $i = 1, 2, \ldots, \nu$, $j = 1, 2, \ldots, q$; например, m_{12} — число пар чисел (x, y), у которых x попало в группу x'_2, а y — в группу y'_1. Числа $m_{11}, m_{12}, \ldots, m_{q\nu}$ называются ***клеточными частотами***.

Таблица 11.9

j	i / y'_j \ x'_i	1	2	…	ν	
		x'_1	x'_2	…	x'_ν	m_{j*}
1	y'_1	m_{11}	m_{12}	…	$m_{1\nu}$	m_{1*}
2	y'_2	m_{21}	m_{22}	…	$m_{2\nu}$	m_{2*}
…	…	…	…	…	…	…
q	y'_q	m_{q1}	m_{q2}	…	$m_{q\nu}$	m_{q*}
m_{*i}		m_{*1}	m_{*2}	…	$m_{*\nu}$	$m_{**} = n$
Групповое среднее $\bar{y}_{(i)}$		$\bar{y}_{(1)}$	$\bar{y}_{(2)}$	…	$\bar{y}_{(\nu)}$	
Выборочная групповая дисперсия $\hat{\sigma}^2_{(i)}$		$\hat{\sigma}^2_{(1)}$	$\hat{\sigma}^2_{(2)}$	…	$\hat{\sigma}^2_{(\nu)}$	
$\bar{y}^{\text{лин}}_{(i)}$		$\bar{y}^{\text{лин}}_{(1)}$	$\bar{y}^{\text{лин}}_{(2)}$	…	$\bar{y}^{\text{лин}}_{(\nu)}$	

В таблице 11.9 наряду с клеточными частотами приведены ***маргинальные (краевые) частоты***:

суммы частот по каждой строке

$$m_{1*} = \sum_{i=1}^{\nu} m_{1i}, \qquad m_{2*} = \sum_{i=1}^{\nu} m_{2i}, \ \ldots, \ m_{q*} = \sum_{i=1}^{\nu} m_{qi},$$

суммы частот по каждому столбцу

$$m_{*1} = \sum_{j=1}^{q} m_{j1}, \qquad m_{*2} = \sum_{j=1}^{q} m_{j2}, \ \ldots, \ m_{*\nu} = \sum_{j=1}^{q} m_{j\nu};$$

общее число наблюдений

$$m_{**} = m_{1*} + m_{2*} + \ldots + m_{q*} =$$

$$= m_{*1} + m_{*2} + \ldots + m_{*\nu} = \sum_{j=1}^{q} \sum_{i=1}^{\nu} m_{ji} = n.$$

Используя данные корреляционной таблицы, приведем формулы вычисления характеристик (в том числе и групповых средних и выборочных групповых дисперсий), необходимых для построения выборочных аналогов показателей связи величин X и Y, и произведем расчеты этих характеристик, используя данные примера 11.3 (формулы и расчеты содержатся в таблице 11.11). Смысл $\overline{y}_{(i)}^{\text{лин}}$ выясняется в п. 11.2.4.

ПРИМЕР 11.3. Исследуется зависимость между стоимостью основных промышленно-производственных фондов (X, тыс. ден. ед.) и объемом выпущенной продукции (Y, тыс. ден. ед.). Данные по 60 случайно выбранным однотипным фирмам сгруппированы в корреляционную таблицу 11.10.

В таблице 11.10 $\nu = 5$, $q = 5$. Число 2, стоящее на пересечении строки $j = 1$ и столбца $i = 1$, — это m_{11}; оно означает, что у двух предприятий стоимость основных фондов лежит в интервале от 0 до 2, а выпуск продукции — в интервале 0—0,2. Маргинальные частоты: $4 = 2 + 2$, $19 = 2 + 7 + 10$, ..., $2 = 2$; $4 = 2 + 2$, $11 = 2 + 7 + 2$, ..., $4 = 2 + 2$. Общее число наблюдений $n = 60 = 4 + 19 + \ldots + 2 = 4 + 11 + \ldots + 4 = 2 + 2 + 2 + 7 + 10 + \ldots + 2$.

Графическое изображение корреляционной таблицы 11.10, называемое ***полем корреляции***, дано на рисунке 11.6. Это прямоугольная сетка, в каждой клетке которой столько точек, какова соответствующая клеточная частота (смысл изображенных на рис. 11.6 ломаных и прямой линий выясняется дальше). Расчеты выборочных характеристик приведены в таблице 11.11 (смысл $\overline{y}_{(i)}^{\text{лин}}$ выясняется в п. 11.2.4.).

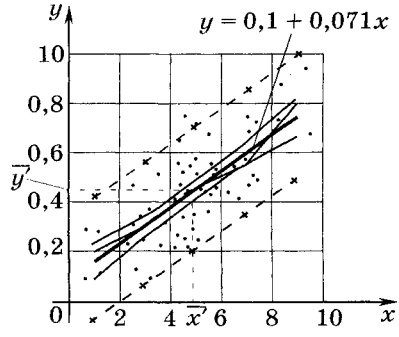


Рис. 11.6

Таблица 11.10

i			1	2	3	4	$5 = \nu$	
Стоимость фондов (интервал)			0—2	2—4	4—6	6—8	8—10	
j	объем продукции (интервал)	центр интервала x_i' / центр интервала y_j'	$x_1' = 1$	$x_2' = 3$	$x_3' = 5$	$x_4' = 7$	$x_5' = 9$	m_{j*}
1	0—0,2	$y_1' = 0{,}1$	$m_{11} = 2$	$m_{12} = 2$				$m_{1*} = 4$
2	0,2—0,4	$y_2' = 0{,}3$	$m_{21} = 2$	$m_{22} = 7$	$m_{23} = 10$			$m_{2*} = 19$
3	0,4—0,6	$y_3' = 0{,}5$		$m_{32} = 2$	$m_{33} = 17$	$m_{34} = 7$		$m_{3*} = 26$
4	0,6—0,8	$y_4' = 0{,}7$			$m_{43} = 4$	$m_{44} = 3$	$m_{45} = 2$	$m_{4*} = 9$
$5 = q$	0,8—1,0	$y_5' = 0{,}9$					$m_{55} = 2$	$m_{5*} = 2$
m_{*i}			$m_{*1} = 4$	$m_{*2} = 11$	$m_{*3} = 31$	$m_{*4} = 10$	$m_{*5} = 4$	$n = 60$
Групповое среднее $\bar{y}_{(i)}$			0,20	0,30	0,46	0,56	0,80	$\bar{x} = 4{,}967$; $\hat{\sigma}_X^2 = 3{,}5322$
Выборочная групповая дисперсия $\hat{\sigma}_{(i)}^2$			0,010	0,014	0,016	0,008	0,010	$\bar{y} = 0{,}453$; $\hat{\sigma}_Y^2 = 0{,}0325$
$\bar{y}_{(i)}^{\text{лин}}$			0,171	0,313	0,455	0,597	0,739	$\overline{xy} = 2{,}502$

Таблица 11.11

Генеральная характеристика	Соответствующая выборочная характеристика и формула ее вычисления по данным таблицы 11.9	Расчет выборочной характеристики по данным таблицы 11.10
MX	$\bar{x} = (x_1' m_{*1} + x_2' m_{*2} + \ldots + x_\nu' m_{*\nu})/n$	$\bar{x} = (1 \cdot 4 + 3 \cdot 11 + \ldots + 9 \cdot 4)/60 = 4{,}967$
DX, σ_X^2	$\hat{\sigma}_X^2 = (x_1'^2 m_{*1} + x_2'^2 m_{*2} + \ldots + x_\nu'^2 m_{*\nu})/n - \bar{x}^2$	$\hat{\sigma}_X^2 = (1^2 \cdot 4 + 3^2 \cdot 11 + \ldots + 9^2 \cdot 4)/60 - 4{,}967^2 = 3{,}5322$
MY	$\bar{y} = (y_1' m_{1*} + y_2' m_{2*} + \ldots + y_q' m_{q*})/n$	$\bar{y} = (0{,}1 \cdot 4 + 0{,}3 \cdot 19 + \ldots + 0{,}9 \cdot 2)/60 = 0{,}453$
DY, σ_Y^2	$\hat{\sigma}_Y^2 = (y_1'^2 m_{1*} + y_2'^2 m_{2*} + \ldots + y_q'^2 m_{q*})/n - \bar{y}^2$	$\hat{\sigma}_Y^2 = (0{,}1^2 \cdot 4 + 0{,}3^2 \cdot 19 + \ldots + 0{,}9^2 \cdot 2)/60 - 0{,}453^2 = 0{,}0325$
$M(XY)$	$\overline{xy} = (x_1' y_1' m_{11} + x_1' y_2' m_{21} + \ldots + x_\nu' y_q' m_{q\nu})/n$	$\overline{xy} = (1 \cdot 0{,}1 \cdot 2 + 1 \cdot 0{,}3 \cdot 2 + \ldots + 9 \cdot 0{,}9 \cdot 2)/60 = 2{,}502$
$M(Y \mid x)$ — среднее значение величины Y, если X примет значение x	$\bar{y}_{(i)}$ — среднее из «игреков», зафиксированных при x_i' (групповое среднее)	
	$\bar{y}_{(1)} = (y_1' m_{11} + y_2' m_{21} + \ldots + y_q' m_{q1})/m_{*1}$	$\bar{y}_{(1)} = (0{,}1 \cdot 2 + 0{,}3 \cdot 2)/4 = 0{,}20$
	$\bar{y}_{(2)} = (y_1' m_{12} + y_2' m_{22} + \ldots + y_q' m_{q2})/m_{*2}$	$\bar{y}_{(2)} = (0{,}1 \cdot 2 + 0{,}3 \cdot 7 + 0{,}5 \cdot 2)/11 = 0{,}30$
	.....	.....

Окончание табл.

Генеральная характеристика	Соответствующая выборочная характеристика и формула ее вычисления по данным таблицы 11.9	Расчет выборочной характеристики по данным таблицы 11.10
	$\bar{y}_{(v)} = (y'_1 m_{1v} + y'_2 m_{2v} + \ldots + y'_q m_{qv})/m_{*v}$	$\bar{y}_{(5)} = (0{,}7 \cdot 2 + 0{,}9 \cdot 2)/4 = 0{,}80$
$D(Y \mid x)$ — дисперсия величины Y, если X примет значение x	$\hat{\sigma}^2_{(i)}$ — дисперсия «игреков», зафиксированных при x'_i, выборочная групповая дисперсия[1]	
	$\hat{\sigma}^2_{(1)} = (y'^2_1 m_{11} + y'^2_2 m_{21} + \ldots + y'^2_q m_{q1})/m_{*1} - \bar{y}^2_{(1)}$	$\hat{\sigma}^2_{(1)} = (0{,}1^2 \cdot 2 + 0{,}3^2 \cdot 2)/4 - 0{,}2^2 = 0{,}010$
	$\hat{\sigma}^2_{(2)} = (y'^2_1 m_{12} + y'^2_2 m_{22} + \ldots + y'^2_q m_{q2})/m_{*2} - \bar{y}^2_{(2)}$	$\hat{\sigma}^2_{(2)} = (0{,}1^2 \cdot 2 + 0{,}3^2 \cdot 7 + 0{,}5^2 \cdot 2)/11 - 0{,}3^2 = 0{,}014$
	.....	.....
	$\hat{\sigma}^2_{(v)} = (y'^2_1 m_{1v} + y'^2_2 m_{2v} + \ldots + y'^2_q m_{qv})/m_{*v} - \bar{y}^2_{(v)}$	$\hat{\sigma}^2_{(5)} = (0{,}7^2 \cdot 2 + 0{,}9^2 \cdot 2)/4 - 0{,}8^2 = 0{,}010$

[1] При каждом x'_i должно быть зафиксировано по крайней мере два различных «игрека»; это гарантирует отличие групповых выборочных дисперсий от нуля.

В таблице 11.11 вычисление характеристик проводилось с использованием **Статистических и Математических функций** Microsoft Excel.

Данные корреляционной таблицы 11.10 были введены в рабочее поле в виде двух строчек (x'_k и y'_k) (табл. 11.12).

Таблица 11.12

x'_k	1	1	1	1	3	3	3	...	3
y'_k	0,1	0,1	0,3	0,3	0,1	0,1	0,3	...	0,3
	$m_{11}=2$		$m_{21}=2$		$m_{12}=2$		$m_{22}=7$		

x'_k	3	3	5	...	5	5	...	5	
y'_k	0,5	0,5	0,3	...	0,3	0,5	...	0,5	
	$m_{32}=2$		$m_{23}=10$			$m_{33}=17$			

x'_k	5	...	5	7	...	7			
y'_k	0,7	...	0,7	0,5	...	0,5			
	$m_{43}=4$			$m_{34}=7$					

x'_k	7	7	7	9	9	9	9		
y'_k	0,7	0,7	0,7	0,7	0,7	0,9	0,9		
	$m_{44}=3$			$m_{45}=2$		$m_{55}=2$			

Затем с помощью **Статистической функции** СРЗНАЧ найдены $\bar{x} = 4{,}967$ и $\bar{y} = 0{,}453$, а с помощью **Функции** ДИСПР $\hat{\sigma}_X^2 = 3{,}5322$ и $\hat{\sigma}_Y^2 = 0{,}0325$. Используя **Математическую функцию** СУММПРОИЗВ, нашли сумму произведений $\sum_{k=1}^{60} x'_k y'_k$, разделив которую на 60, получим $\overline{xy} = 2{,}502$.

Для нахождения групповых средних и групповых дисперсий, приведенных в таблице 11.11, для каждого $x'_i = 1, 3, 5, 7, 9$ в рабочее поле введем столбик зафиксированных «игреков» (табл. 11.13).

Таблица 11.13

x_i'	$x_1' = 1$	$x_2' = 3$	$x_3' = 5$	$x_4' = 7$	$x_5' = 9$
y_j'	0,1 0,1 } 2 0,3 0,3 } 2	0,1 0,1 } 2 0,3 … } 7 0,3 0,5 0,5 } 2	0,3 … } 10 0,3 0,5 … } 17 0,5 0,7 … } 4 0,7	0,5 … } 7 0,5 0,7 … } 3 0,7	0,7 0,7 } 2 0,9 0,9 } 2
m_{*i}	$m_{*1} = 4$	$m_{*2} = 11$	$m_{*3} = 31$	$m_{*4} = 10$	$m_{*5} = 4$

и к каждому столбцу «игреков» применим **Статистические функции** СРЗНАЧ и ДИСПР. ◀

Напомним, что функция регрессии (генеральная) Y на x — это функция $M(Y \mid x) = \varphi(x)$, описывающая изменение условного математического ожидания $M(Y \mid x)$ при изменении значений x величины X. Поскольку выборочным аналогом значения $M(Y \mid x_i')$ условного математического ожидания является групповое среднее $\bar{y}_{(i)}$, выборочным аналогом функции регрессии Y на x является функция, задаваемая таблицей 11.14.

Таблица 11.14

x_i'	x_1'	x_2'	…	x_v'
$\bar{y}_{(i)}$	$\bar{y}_{(1)}$	$\bar{y}_{(2)}$	…	$\bar{y}_{(v)}$

Функцию, заданную этой таблицей, будем называть ***выборочной функцией регрессии Y на x.***

Среднее групповых средних $\bar{y}_{(i)}$, вычисленное с учетом маргинальных частот m_{*i} появления вариантов x_i' (см. таблицу 11.9), равно $\bar{y}$:

$$(\bar{y}_{(1)} m_{*1} + \bar{y}_{(2)} m_{*2} + \ldots + \bar{y}_{(v)} m_{*v})/n = \bar{y}. \qquad (11.45)$$

Равенство (11.45) является выборочным аналогом равенства (11.2).

▶ **ПРИМЕР 11.3** (продолжение). В условиях примера (данных таблицы 11.10) выборочная функция регрессии задана таблицей 11.15.

Таблица 11.15

x_i'	1	3	5	7	9
$\bar{y}_{(i)}$	0,20	0,30	0,46	0,56	0,80

Графическое изображение функции, заданной таблицей 11.15, дано на рисунке 11.6 (ломаная линия).

В подтверждение равенства (11.45) вычислим среднее групповых средних, учитывая при этом значения маргинальных частот: $m_{*1} = 4$, $m_{*2} = 11$, $m_{*3} = 31$, $m_{*4} = 10$, $m_{*5} = 4$; получим $(0{,}20 \cdot 4 + 0{,}30 \cdot 11 + 0{,}46 \cdot 31 + 0{,}56 \cdot 10 + 0{,}80 \cdot 4)/60 = 0{,}453$, что совпадает с $\bar{y} = 0{,}453$. ◂

11.2.4. Выборочный коэффициент парной корреляции, выборочная линейная регрессия и метод наименьших квадратов. Свойства выборочного коэффициента парной корреляции.

Выборочный коэффициент парной корреляции (или просто *выборочный коэффициент корреляции*) вычисляют по результатам (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) парных наблюдений двумерной случайной величины (X, Y) по одной из двух следующих тождественных формул:

$$\hat{r}_{X,Y} = \frac{\overline{(x - \bar{x})(y - \bar{y})}}{\hat{\sigma}_X \hat{\sigma}_Y}, \qquad (11.46)$$

или

$$\hat{r}_{X,Y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\hat{\sigma}_X \hat{\sigma}_Y}, \qquad (11.47)$$

в которых $\bar{x} = \sum_{k=1}^{n} x_k/n$, $\hat{\sigma}_X = \sqrt{\sum_{i=1}^{n}(x_k - \bar{x})^2/n}$, $\overline{xy} = \sum_{k=1}^{n} x_k y_k/n$, $\hat{\sigma}_X \neq 0$, $\hat{\sigma}_Y \neq 0$.

Тождественность формул (11.46) и (11.47) доказана в § 7.2. Числитель дроби (11.46) или дроби (11.47)]

$$\hat{K}(X, Y) = \overline{(x - \bar{x})(y - \bar{y})} \equiv \overline{xy} - \bar{x} \cdot \bar{y} \qquad (11.48)$$

называется ***выборочным корреляционным моментом или выборочной ковариацией*** случайных величин X и Y.

Вычисление $\hat{r}_{X,Y}$ по формуле (11.47) менее трудоемко, чем по формуле (11.46); поэтому формулу (11.47) используют чаще. Для вычисления $\hat{r}_{X,Y}$, можно использовать **Статистическую функцию** КОРРЕЛ.

▶ **ПРИМЕР 11.3** (продолжение). В условиях примера (данных таблицы 11.10).

$$\hat{r}_{X,Y} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\hat{\sigma}_X\hat{\sigma}_Y} \underset{\text{(см. таблицу 11.11)}}{=} \frac{2{,}502 - 4{,}967\cdot 0{,}453}{\sqrt{3{,}5322}\sqrt{0{,}0325}} = 0{,}74.$$

Статистическая функция КОРРЕЛ, аргументы которой — числовые данные строк таблицы 11.12 возвратила $\hat{r}_{X,Y} = 0{,}74$. ◀

Коэффициент $\hat{r}_{X,Y}$ является выборочным аналогом (генерального) коэффициента $r_{X,Y}$: если в формулах (11.3) и (11.4) заменить математические ожидания — генеральные средние и генеральные средние квадратические отклонения соответственно выборочными средними и выборочными средними квадратическими отклонениями, то получим формулы (11.46) и (11.47). Поэтому для $\hat{r}_{X,Y}$ и $\hat{K}(X, Y)$ имеют место следующие соотношения:

$$\hat{K}(X, Y) = \hat{r}_{X,Y}\hat{\sigma}_X\hat{\sigma}_Y; \tag{11.49}$$

$$\hat{K}(X, X) = \hat{D}X; \tag{11.50}$$

$$\hat{r}_{X,X} = 1, \ (\hat{\sigma}_X \neq 0); \tag{11.51}$$

$$\hat{r}_{b+cX,d+eY} = \begin{cases} \hat{r}_{X,Y}, & \text{если } ce > 0, \\ -\hat{r}_{X,Y}, & \text{если } ce < 0, \end{cases} \tag{11.52}$$

которые являются выборочными аналогами соотношений (11.6)—(11.9).

Введем понятие выборочной линейной регрессии. Напомним, линейная (генеральная) регрессия Y на x имеет вид (11.20):

$$M^{\text{лин}}(Y \mid x) = a_0 + a_1 x, \ a_1 \neq 0,$$

где

$$a_1 = a_{Y|x} = r_{X,Y}\frac{\sigma_Y}{\sigma_X}, \qquad a_0 = MY - r_{X,Y}\frac{\sigma_Y}{\sigma_X}MX.$$

Оцененная по выборочным данным линейная регрессия, или *выборочная линейная регрессия Y на x*, имеет вид

$$\bar{y}_x^{\text{лин}} = \hat{a}_0 + \hat{a}_1 x, \ \hat{a}_1 \neq 0, \tag{11.53}$$

где $\hat{a}_0$ и $\hat{a}_1$ — выборочные оценки параметров a_0 и a_1:

$$\hat{a}_1 = \hat{a}_{Y|x} = \hat{r}_{X,Y}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}, \qquad \hat{a}_0 = \bar{y} - \hat{r}_{X,Y}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\bar{x}. \tag{11.54}$$

Прямая (11.20) проходит через точку $(MX; MY)$; аналогично и прямая (11.53) проходит через точку $(\bar{x}; \bar{y})$. Действительно, если в уравнении (11.53) положить $x = \bar{x}$, то, учитывая формулы (11.54), получим

$$\bar{y}_{\bar{x}}^{\text{лин}} = \hat{a}_0 + \hat{a}_1 \bar{x} = \bar{y} - \hat{r}_{X,Y} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \bar{x} + \hat{r}_{X,Y} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \bar{x} = \bar{y}.$$

▶ **ПРИМЕР 11.3** (продолжение). По условию (см. табл. 11.10) был вычислен выборочный коэффициент корреляции $\hat{r}_{X,Y} = 0{,}74$. Тогда, согласно формулам (11.54) и (11.53) и учитывая результаты вычислений, приведенные в таблице 11.11, найдем $\hat{a}_1 = 0{,}74\sqrt{0{,}0325} / \sqrt{3{,}5322} = 0{,}071$; $\hat{a}_0 = 0{,}453 - 0{,}071 \cdot 4{,}967 = 0{,}1$, выборочную линейную регрессию Y на x

$$\bar{y}_x^{\text{лин}} = 0{,}1 + 0{,}071x. \qquad (11.55)$$

В Microsoft Excel $\hat{a}_1$ и $\hat{a}_0$ можно рассчитать по результатам (x_k, y_k) парных наблюдений двумерной величины $(X; Y)$, используя **Статистическую функцию** ЛИНЕЙН.

Рассчитанные по уравнению (11.55) значения $\bar{y}_{(i)}^{\text{лин}} = 0{,}1 + 0{,}071\, x_i'$, $i = 1, 2, \ldots, 5$, приведены в таблице 11.10. Прямая (11.55) изображена на рисунке 11.6 (она «выравнивает», «спрямляет» ломаную линию — график выборочной функции регрессии, заданной таблицей 11.15). Эта прямая, как и следовало ожидать, проходит через следующую точку $(\bar{x} = 4{,}967; \bar{y} = 0{,}453)$. ◀

Учитывая, что реализуемые в Microsoft Excel математико-статистические методы в качестве исходной информации используют результаты непосредственных (а не сгруппированных) наблюдений, а также для упрощения последующих формул, будем предполагать, что n парных наблюдений (x_k, y_k) двумерной случайной величины (X, Y) не сгруппированы. Исходную информацию, заданную в форме корреляционной таблицы 11.9, нетрудно свести к n парным наблюдениям (x_k, y_k) (такое представление корреляционной табл. 11.10 дано в табл. 11.12[1]).

Для наглядности рассуждений построим таблицу 11.16, в которой приведем:

— результаты (x_k, y_k) парных наблюдений величины (X, Y);

[1] К представлению исходных данных в форме корреляционной таблицы 11.9 будем обращаться только при изучении групповых средних и групповых выборочных дисперсий.

— значения $\bar{y}_k^{\text{лин}} = \hat{a}_0 + \hat{a}_1 x_k$, где a_0 и a_1 рассчитываются по формулам (11.54); обратим внимание на то, что число $\bar{y}_k^{\text{лин}}$ можно интерпретировать как выборочную оценку условного математического ожидания $M^{\text{лин}}(Y \mid x_k) = a_0 + a_1 x_k$ и как наблюдавшееся значение случайной величины $M^{\text{лин}}(Y \mid X) = a_0 + a_1 X$, если X примет значение x_k;

— ошибки $(y_k - \bar{y}_k^{\text{лин}})$, возникающие при использовании $\bar{y}_k^{\text{лин}}$ вместо y_k, при этом разность $(y_k - \bar{y}_k^{\text{лин}})$ будем интерпретировать как наблюдавшееся значение случайной ошибки, равной $Y - M^{\text{лин}}(Y \mid X)$, если X примет значение x_k.

Таблица 11.16

k	1	2	...	n
x_k	x_1	x_2	...	x_n
y_k	y_1	y_2	...	y_n
$\bar{y}_k^{\text{лин}} = \hat{a}_0 + \hat{a}_1 x_k;\ \hat{a}_1 \neq 0$	$\bar{y}_1^{\text{лин}}$	$\bar{y}_2^{\text{лин}}$	...	$\bar{y}_n^{\text{лин}}$
$y_k - \bar{y}_k^{\text{лин}}$	$y_1 - \bar{y}_1^{\text{лин}}$	$y_2 - \bar{y}_2^{\text{лин}}$	...	$y_n - \bar{y}_n^{\text{лин}}$

Сформулируем теоремы для набора чисел $\bar{y}_k^{\text{лин}}$ и набора чисел $y_k - \bar{y}_k^{\text{лин}}$, $k = 1, 2, \ldots, n$, которые являются выборочными аналогами теорем 1—3, приведенных в п. 11.2.1 и имеющих место для случайных величин $M^{\text{лин}}(Y \mid X)$ и $Y - M^{\text{лин}}(Y \mid X)$.

Теорема 1′. *Среднее чисел* $\bar{y}_k^{\text{лин}}$, $k = 1, 2, \ldots, n$, *равно* $\bar{y}$, *т. е.*

$$\sum_{k=1}^{n} \bar{y}_k^{\text{лин}} / n = \bar{y}, \qquad (11.56)$$

а среднее квадратов их отклонений от $\bar{y}$ (*которое обозначим через* $\hat{\sigma}^2_{LR\,Y|X}$ *и назовем выборочной дисперсией линейной регрессии* Y *на* X) *таково*:

$$\hat{\sigma}^2_{LR\,Y|X} = \sum_{k=1}^{n} (\bar{y}_k^{\text{лин}} - \bar{y})^2 / n = \hat{\sigma}^2_Y \hat{r}^2_{X,Y} \qquad (11.57)$$

[(11.56) и (11.57) — *выборочные аналоги соотношений* (11.22) и (11.23)].

Теорема 2′. *Среднее чисел* $(y_k - \bar{y}_k^{\text{лин}})$, $k = 1, 2, \ldots, n$, *равно нулю*:

$$\sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})/n = 0, \tag{11.58}$$

а *среднее квадратов их отклонений от нуля* (*которое обозначим через* $\hat{\sigma}^2_{ELR\,Y|X}$ *и назовем выборочной дисперсией ошибки линейной регрессии* Y *на* X) *таково*:

$$\hat{\sigma}^2_{ELR\,Y|X} = \sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2/n = \hat{\sigma}^2_Y(1 - \hat{r}^2_{X,Y}) \tag{11.59}$$

[(11.58) и (11.59) — *выборочные аналоги соотношений* (11.24) и (11.25)].

Теорема 3′. *Сумма выборочной дисперсии линейной регрессии* Y *на* X *и выборочной дисперсии ошибки этой регрессии равна выборочной дисперсии величины* Y:

$$\hat{\sigma}^2_{LR\,Y|X} + \hat{\sigma}^2_{ELR\,Y|X} = \hat{\sigma}^2_Y, \tag{11.60}$$

или

$$\sum_{k=1}^{n} (\bar{y}_k^{\text{лин}} - \bar{y})^2/n + \sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2/n = \sum_{k=1}^{n} (y_k - \bar{y})^2/n \tag{11.61}$$

[*тождество* (11.60) — *выборочный аналог дисперсионного тождества* (11.26) *для линейной регрессии* Y *на* X].

Предоставляем самостоятельно убедиться в правомерности равенств (11.56) — (11.60), используя данные таблицы 11.12 и считая, что $\bar{y}_k^{\text{лин}} = 0{,}1 + 0{,}071x'_k$ (см. (11.55)), а $y_k = y'_k$.

Убедимся в том, что выборочная дисперсия $\hat{\sigma}^2_{ELR\,Y|X}$ ошибки линейной регрессии (см. (11.59)), подобно генеральной дисперсии $\hat{\sigma}^2_{ELR\,Y|X}$ (см. (11.25)), обладает свойством минимальности, которое в данном случае означает, что равное выборочной дисперсии $\hat{\sigma}^2_{ELR\,Y|X}$ *среднее квадратов ошибок, возникающих при замене чисел* y_k *числами* $\bar{y}_k^{\text{лин}}$ (табл. 11.16), *т. е.* $\sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2/n$, *меньше среднего квадратов ошибок, возникающих при замене* y_k *любой величиной, линейно зависящей от* x_k, *но отличной от*

$$\bar{y}_k^{\text{лин}} = \underbrace{\bar{y} - \hat{r}_{X,Y}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\bar{x}}_{\hat{a}_0} + \underbrace{\hat{r}_{X,Y}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}}_{\hat{a}_1 \neq 0} x_k, \quad k = 1, 2, \ldots, n. \tag{11.62}$$

≫ Найдем такие коэффициенты b_0 и b_1 в уравнении прямой $\hat{y} = b_0 + b_1 x$, при которых величина $\sum\limits_{k=1}^{n} (y_k - \hat{y}_k)^2/n$, где $\hat{y}_k = b_0 + b_1 x_k$, а (x_k, y_k) — результаты наблюдения двумерной величины (X, Y), имеет минимальное значение, или, иначе, значение функции

$$F(b_0, b_1) = \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k)^2/n \qquad (11.63)$$

минимально (требование минимизации функции (11.63) называется ***требованием метода наименьших квадратов***).

Найдем такие значения b_0 и b_1, при которых частные производные функции (11.63) по b_0 и b_1 равны нулю ($F'_{b_0} = 0$ и $F'_{b_1} = 0$), затем убедимся, что при найденных значениях функция (11.63) достигает минимума. Имеем систему

$$\begin{cases} F'_{b_0} = -\sum\limits_{k=1}^{n} 2(y_k - b_0 - b_1 x_k)/n = 0, \\ F'_{b_1} = -\sum\limits_{k=1}^{n} 2(y_k - b_0 - b_1 x_k)x_k/n = 0, \end{cases}$$

которая в результате тождественных преобразований принимает вид

$$\begin{cases} b_0 n + b_1 \sum\limits_{k=1}^{n} x_k = \sum\limits_{k=1}^{n} y_k, \\ b_0 \sum\limits_{k=1}^{n} x_k + b_1 \sum\limits_{k=1}^{n} x_k^2 = \sum\limits_{k=1}^{n} y_k x_k. \end{cases} \qquad (11.64)$$

Система (11.64) называется ***системой нормальных уравнений***. Обратим внимание на «формальное» правило составления этой системы: первое уравнение $b_0 \sum\limits_{k=1}^{n} 1 + b_1 \sum\limits_{k=1}^{n} x_k = \sum\limits_{k=1}^{n} y_k$ «соответствует» уравнению $b_0 \cdot 1 + b_1 \cdot x_k = y_k$, второе уравнение получается из первого «введением под знаки сумм множителя x_k».

Решим систему (11.64). Из первого уравнения находим: $b_0 = \bar{y} - b_1 \bar{x}$. Подставляя это выражение во второе уравнение, предварительно поделенное на n, получаем

$$(\bar{y} - b_1 \bar{x})\bar{x} + b_1 \overline{x^2} = \overline{yx};$$

отсюда находим

$$\left.\begin{aligned} b_1 &= \frac{\overline{yx} - \bar{y}\bar{x}}{\overline{x^2} - (\bar{x})^2} \underset{(11.47)}{=} \frac{\hat{r}_{X,Y}\hat{\sigma}_X\hat{\sigma}_Y}{\hat{\sigma}_X^2} = \frac{\hat{r}_{X,Y}\hat{\sigma}_Y}{\hat{\sigma}_X} \underset{(11.54)}{=} \hat{a}_1, \\ b_0 &= \bar{y} - b_1\bar{x} = \bar{y} - \frac{\hat{r}_{X,Y}\hat{\sigma}_Y}{\hat{\sigma}_X}\bar{x} \underset{(11.54)}{=} \hat{a}_0. \end{aligned}\right\} \qquad (11.65)$$

Таким образом, при $b_0 = \hat{a}_0$ и $b_1 = \hat{a}_1$ частные производные функции (11.63) равны нулю.

Убедимся в том, что эти значения b_0 и b_1 дают минимум функции (11.63). Найдем для функции (11.63) частные производные второго порядка:

$$F''_{b_0b_0} = (F'_{b_0})'_{b_0} = \sum_{k=1}^{n} 2/n = 2;\ F''_{b_0b_1} = (F'_{b_0})'_{b_1} = \sum_{k=1}^{n} 2x_k/n = 2\bar{x};$$

$$F''_{b_1b_1} = (F'_{b_1})'_{b_1} = \sum_{k=1}^{n} 2x_k^2/n = 2\overline{x^2}.$$

Обратим внимание на то, что найденные производные не зависят от b_0 и b_1, поэтому и при $b_0 = \hat{a}_0$, $b_1 = \hat{a}_1$ значения этих производных

$$A = F''_{b_0b_0}(\hat{a}_0, \hat{a}_1) = 2;\ B = F''_{b_0b_1}(\hat{a}_0, \hat{a}_1) = 2\bar{x},$$

$$C = F''_{b_1b_1}(\hat{a}_0, \hat{a}_1) = 2\overline{x^2}.$$

Напомним, что если $\Delta = AC - B^2 > 0$, то в точке $(\hat{a}_0; \hat{a}_1)$ функция $F(b_0, b_1)$ имеет экстремум: при $A < 0$ — максимум, при $A > 0$ — минимум; если $\Delta < 0$, то функция экстремума не имеет; если $\Delta = 0$, вопрос о наличии экстремума остается открытым.

В рассматриваемом случае

$$\Delta = AC - B^2 = 2 \cdot 2\overline{x^2} - (2\bar{x})^2 = 4(\overline{x^2} - (\bar{x})^2) = 4\hat{\sigma}_X^2 > 0,$$

при этом $A = 2 > 0$. Следовательно, точка $(\hat{a}_0, \hat{a}_1)$ — точка минимума (единственная) функции $F(b_0, b_1)$.

Таким образом, $\sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2/n$, где $\bar{y}_k^{\text{лин}}$ рассчитывается по уравнению (11.62), или $\hat{\sigma}^2_{EL\,RY|X}$ (см. (11.59)), действительно обладает свойством минимальности. Прямую (11.62) называют прямой, построенной по методу наименьших квадратов.

Замечание. В данном случае метод наименьших квадратов — это метод нахождения неизвестных коэффициентов b_0 и b_1 прямой $\hat{y} = b_0 + b_1x$, исходя из требования

$$\sum_{k=1}^{n} (y_k - \hat{y}_k)^2/n \to \min, \qquad \text{или} \qquad \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \to \min, \quad (11.66)$$

т. е. из требования минимизации суммы квадратов отклонений значений $\hat{y}_k = b_0 + b_1x_k$ от наблюдаемых значений y_k (напомним (x_k, y_k), $k = 1, 2, \ldots, n$, — результаты наблюдения двумерной величины (X, Y)).

В общем случае метод наименьших квадратов является методом нахождения неизвестных коэффициентов функции $\hat{y} = f(x)$ заданного вида (линейного, параболического или любого другого), исходя из требования

$$\sum_{k=1}^{n} \delta_k^2 = \sum_{k=1}^{n} (y_k - f(x_k))^2 \to \min. \qquad (11.67)$$

Графическая иллюстрация «невязок» δ_k — отклонений фактических значений y_k от значений $\hat{y}_k = f(x_k)$ дана на рисунке 11.7.

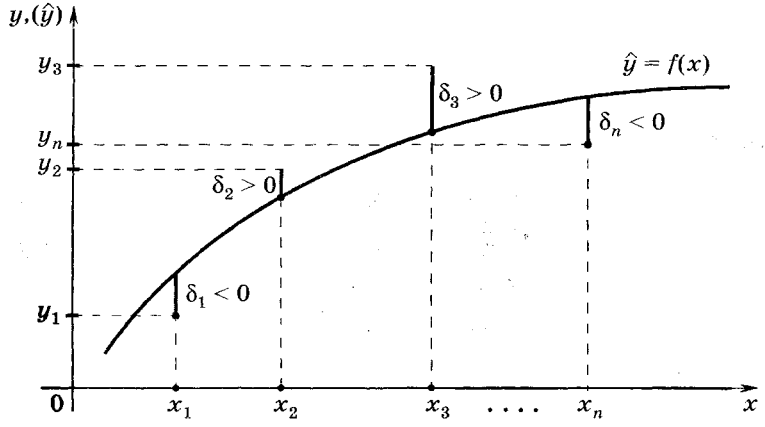


Рис. 11.7 ≪

Выше рассматривалась выборочная линейная регрессии Y на x. Аналогично можно рассмотреть выборочную линейную регрессию X на y и получить следующие результаты:

— выборочная линейная регрессия X на y имеет вид

$$\bar{x}_y^{\text{лин}} = \underbrace{\bar{x} - \hat{r}_{X,Y}\frac{\hat{\sigma}_X}{\hat{\sigma}_Y}\bar{y}}_{\hat{b}_0} + \underbrace{\hat{r}_{X,Y}\frac{\hat{\sigma}_X}{\hat{\sigma}_Y}}_{\hat{b}_1 \neq 0} y; \qquad (11.68)$$

— выборочная дисперсия линейной регрессии X на Y определяется формулой

$$\hat{\sigma}^2_{LR\ X|Y} = \sum_{k=1}^{n} (\bar{x}_k^{\text{лин}} - \bar{x})^2/n = \hat{\sigma}_X^2 r^2_{X,Y}, \qquad (11.69)$$

где $\bar{x}_k^{\text{лин}} = \hat{b}_0 + \hat{b}_1 y_k$, $k = 1, 2, \ldots, n$;

— выборочная дисперсия ошибки линейной регрессии X на Y такова:

$$\hat{\sigma}^2_{ELR\ X|Y} = \sum_{k=1}^{n} (x_k - \bar{x}_k^{\text{лин}})^2/n = \hat{\sigma}_X^2(1 - \hat{r}^2_{X,Y}); \qquad (11.70)$$

— дисперсионное тождество для выборочной линейной регрессии X на Y

$$\hat{\sigma}^2_{LR\ X|Y} + \hat{\sigma}^2_{ELR\ X|Y} = \hat{\sigma}^2_X. \qquad (11.71)$$

Соотношения (11.68)—(11.71) являются выборочными аналогами соотношений (11.27)—(11.30).

Приведем свойства выборочного коэффициента парной корреляции $\hat{r}_{X,Y}$, которые являются выборочными аналогами свойств 1^0 и 4^0 генерального коэффициента корреля-

ции $r_{X,Y}$. Свойства 2^0 и 3^0 коэффициента $r_{X,Y}$ выборочных аналогов не имеют, поскольку используемые в этих свойствах понятия корреляционной зависимости случайных величин и независимости случайных величин относятся только к генеральной совокупности: эти понятия формулируются в терминах математических ожиданий и истинных вероятностей и не могут быть сформулированы в терминах выборочных средних и опытных вероятностей.

$1^{0'}$. $|\hat{r}_{X,Y}| \leqslant 1$, *или* $-1 \leqslant \hat{r}_{X,Y} \leqslant 1$.

$4^{0'}$. *Условие* $|\hat{r}_{X,Y}| = 1$ *является достаточным и необходимым условием для того, чтобы результаты* n *парных наблюдений* (x_k, y_k), $k = 1, 2, \ldots, n$ — *точки* (x_k, y_k) *принадлежали одной прямой.*

Доказательства этих свойств совпадают с доказательствами свойств 1^0 и 4^0 генерального коэффициента корреляции, если в последних используемые генеральные характеристики заменить их выборочными аналогами.

Если $|r_{X,Y}|$ интерпретируют как меру линейности зависимости между случайными величинами X и Y, то $|\hat{r}_{X,Y}|$ — это мера линейности зависимости между x_k и y_k, $k = 1, 2, \ldots, n$: чем четче проявляется линейность, т. е. чем ближе зависимость между x_k и y_k к линейной, тем больше $|\hat{r}_{X,Y}|$, и наоборот, чем больше $|\hat{r}_{X,Y}|$, тем ближе зависимость между x_k и y_k к линейной. При этом, если при расположении чисел x_k в неубывающем порядке числа y_k увеличиваются или имеют тенденцию к увеличению, и только в этом случае, $\hat{r}_{X,Y} > 0$; если же при расположении чисел x_k в неубывающем порядке числа y_k уменьшаются или имеют тенденцию к уменьшению, и только в этом случае, $\hat{r}_{X,Y} < 0$.

Поскольку $\hat{r}_{X,Y}$ — выборочная оценка коэффициента $r_{X,Y}$, в дальнейшем, используя $|\hat{r}_{X,Y}|$, будем считать, что $|\hat{r}_{X,Y}|$ — выборочная оценка меры линейности зависимости между X и Y.

▶ **ПРИМЕР 11.3** (продолжение). По условию $\hat{r}_{X,Y} = 0{,}74$ — проявление линейности в существующей зависимости между x'_k и y'_k (значения x'_k и y'_k приведены в таблице 11.12, которая получена из корреляционной таблицы 11.10) довольно высокое, что подтверждается изображенным на ри-

сунке 11.6 полем корреляции. Точки образуют неширокое «облако» точек, вытянутое вдоль прямой с угловым коэффициентом, не равным нулю; $\hat{r}_{X,Y} = 0{,}74$ — полученная по 60 фирмам оценка меры линейности зависимости между стоимостью фондов и объемом произведенной продукции. ◀

Более четкое числовое содержание по сравнению с $\hat{r}_{X,Y}$ имеет $\hat{r}^2_{X,Y}$ — ***выборочный коэффициент линейной детерминации*** одной из случайных величин, все равно какой другой величиной. Из соотношений (11.57) и (11.69) соответственно получим

$$\hat{r}^2_{X,Y} = \frac{\hat{\sigma}^2_{LR\,Y|X}}{\hat{\sigma}^2_Y} = \frac{\sum\limits_{k=1}^{n}(\bar{y}_k^{\text{лин}} - \bar{y})^2/n}{\sum\limits_{k=1}^{n}(y_k - \bar{y})^2/n} = \frac{\sum\limits_{k=1}^{n}(\bar{y}_k^{\text{лин}} - \bar{y})^2}{\sum\limits_{k=1}^{n}(y_k - \bar{y})^2}; \qquad (11.72)$$

$$\hat{r}^2_{X,Y} = \frac{\hat{\sigma}^2_{LR\,X|Y}}{\hat{\sigma}^2_X} = \frac{\sum\limits_{k=1}^{n}(\bar{x}_k^{\text{лин}} - \bar{x})^2/n}{\sum\limits_{k=1}^{n}(x_k - \bar{x})^2/n} = \frac{\sum\limits_{k=1}^{n}(\bar{x}_k^{\text{лин}} - \bar{x})^2}{\sum\limits_{k=1}^{n}(x_k - \bar{x})^2}. \qquad (11.73)$$

Эти формулы являются выборочными аналогами формул (11.34) и (11.35), $\hat{r}^2_{X,Y}$ — выборочная оценка генерального коэффициента линейной детерминации $r^2_{X,Y}$. Так как $r^2_{X,Y}$ — доля дисперсии одной из величин, все равно какой, объясняемая корреляционной зависимостью этой величины от другой при условии, что эта зависимость линейная, то выборочный коэффициент линейной детерминации $\hat{r}^2_{X,Y}$ — выборочная оценка названной доли дисперсии.

11.2.5. Выборочное корреляционное отношение и его свойства. В п. 11.2.2 введено понятие корреляционного отношения $\rho_{Y|X}$ — меры проявления функциональной зависимости в стохастической зависимости Y от X: чем четче проявляется функциональная зависимость, или чем ближе зависимость Y от X к функциональной, тем больше $\rho_{Y|X}$.

Выборочное корреляционное отношение $\hat{\rho}_{Y|X}$ — выборочный аналог генерального корреляционного отношения $\rho_{Y|X}$. Вычислить $\hat{\rho}_{Y|X}$ можно лишь когда результаты парных наблюдений (x_k, y_k) двумерной величины (X, Y) сгруппированы в корреляционную таблицу (см. табл. 11.9).

Выборочное корреляционное отношение $\hat{\rho}_{Y|X}$ находят по одной из двух тождественных формул, являющихся выборочными аналогами формул (11.39) и (11.40):

$$\hat{\rho}_{Y|X} = \sqrt{\hat{\sigma}^2_{R\,Y|X}/\hat{\sigma}^2_Y}, \qquad (11.74)$$

или

$$\hat{\rho}_{Y|X} = \sqrt{1 - \hat{\sigma}^2_{ER\,Y|X}/\hat{\sigma}^2_Y}, \ \hat{\rho}_Y \geqslant 0. \qquad (11.75)$$

Тождественность этих формул вытекает из дисперсионного тождества

$$\hat{\sigma}^2_{R\,Y|X} + \hat{\sigma}^2_{ER\,Y|X} = \hat{\sigma}^2_Y, \qquad (11.76)$$

являющегося выборочным аналогом тождества (11.36).

▶ **ПРИМЕР 11.3** (продолжение). Вычислим корреляционное отношение $\hat{\rho}_{Y|X}$ по данным таблицы 11.10 (см. табл. 11.17). Напомним, что выборочной регрессией Y на x называется зависимость групповых средних $\bar{y}_{(i)}$ от вариантов x'_i (см. табл. 11.14) и что для данных таблицы 11.10 выборочная регрессия задана таблицей 11.15.

Таблица 11.17

Выборочная характеристика и формула ее вычисления по данным таблицы 11.9	Расчет выборочной характеристики по данным таблицы 11.10
$\hat{\sigma}^2_Y = \sum_{j=1}^{q} (y'_j - \bar{y})^2 m_{j*}/n =$ $= \sum_{j=1}^{q} (y'_j)^2 m_{j*}/n - (\bar{y})^2$ — выборочная дисперсия величины Y	$\hat{\sigma}^2_Y = 0{,}0325$ (см. табл. 11.11)
$\hat{\sigma}^2_{R\,Y|X} = \sum_{i=1}^{v} (\bar{y}_{(i)} - \bar{y})^2 m_{*i}/n$ — выборочная дисперсия регрессии Y на X	$\hat{\sigma}^2_{R\,Y|X} = [(0{,}2 - 0{,}453)^2 \cdot 4 +$ $+ (0{,}3 - 0{,}453)^2 \cdot 11 + \ldots$ $\ldots + (0{,}8 - 0{,}453)^2 \cdot 4]/60 = 0{,}0185$
$\hat{\sigma}^2_{ER\,Y|X} = \sum_{i=1}^{v} \sum_{j=1}^{q} (y'_j - \bar{y}_{(i)})^2 m_{ji}/n$ — выборочная дисперсия ошибки регрессии Y на X	$\hat{\sigma}^2_{ER\,Y|X} = \sum_{i=1}^{5} \sum_{j=1}^{5} (y'_j - \bar{y}_{(i)})^2 m_{ji}/60 =$ $= [(0{,}1 - 0{,}2)^2 \cdot 2 + (0{,}3 - 0{,}2)^2 \cdot 2 +$ $+ (0{,}1 - 0{,}3)^2 \cdot 2 + (0{,}3 - 0{,}3)^2 \cdot 7 +$ $+ (0{,}5 - 0{,}3)^2 \cdot 2 + \ldots$ $\ldots + (0{,}9 - 0{,}8)^2 \cdot 2]/60 = 0{,}014$
$\hat{\rho}^2_{Y|X} = \sqrt{\hat{\sigma}^2_{R\,Y|X}/\hat{\sigma}^2_Y} =$ $= \sqrt{1 - \hat{\sigma}^2_{ER\,Y|X}/\hat{\sigma}^2_Y}$	$\hat{\rho}^2_{Y|X} = \sqrt{0{,}0185/0{,}0325} =$ $= \sqrt{1 - 0{,}014/0{,}0325} = 0{,}755$

Замечания. 1. Для дисперсии $\hat{\sigma}^2_{ER\,Y|X}$ имеет место соотношение

$$\hat{\sigma}^2_{ER\,Y|X} = \sum_{i=1}^{v} \hat{\sigma}^2_{(i)} m_{*i}/n, \qquad (11.77)$$

являющееся выборочным аналогом соотношения (11.37), т. е. $\hat{\sigma}^2_{ER\,Y|X}$ равна среднему выборочных групповых дисперсий (для данных табл. 11.10 $\hat{\sigma}^2_{ER\,Y|X} = (0{,}01 \cdot 4 + 0{,}014 \cdot 11 + 0{,}016 \cdot 31 + 0{,}008 \cdot 10 + 0{,}01 \cdot 4)/60 = 0{,}014$).

2. Для дисперсий $\hat{\sigma}^2_Y$, $\hat{\sigma}^2_{R\,Y|X}$ и $\hat{\sigma}^2_{ER\,Y|X}$ имеет место тождество (11.76) (для данных табл. 11.10 $\hat{\sigma}^2_{R\,Y|X} = 0{,}0185$; $\hat{\sigma}^2_{ER\,Y|X} = 0{,}014$; $\sigma^2_Y = 0{,}0325$ и $0{,}0185 + 0{,}014 = 0{,}0325$). ◀

При вычислении $\hat{\rho}_{Y|X}$ можно воспользоваться программой **«Однофакторный дисперсионный анализ»** пакета **«Анализ данных»** Microsoft Excel.

▶ **ПРИМЕР 11.3** (продолжение). Вычислим корреляционное отношение $\hat{\rho}_{Y|X}$, используя программу **«Однофакторный дисперсионный анализ»**. Введем в рабочее поле пять столбцов «игреков» из таблицы 11.13, сформированной на базе данных корреляционной таблицы 11.10. Результаты работы программы представлены на рисунке 11.8. Из таблицы дисперсионного анализа получим

$$\hat{\rho}_{Y|X} = \sqrt{\frac{SS_{\text{между группами}}}{SS_{\text{итого}}}} = \sqrt{\frac{1{,}112}{1{,}949}} = 0{,}755.$$

Однофакторный дисперсионный анализ

итоги

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
Столбец 1	4	0,8	0,2	0,013333333
Столбец 2	11	3,3	0,3	0,016
Столбец 3	31	14,3	0,461290323	0,01711828
Столбец 4	10	5.6	0,56	0,009333333
Столбец 5	4	3,2	0,8	0,013333333

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*Критическое*
Между группами	1,111734946	4	0,277946237	18,25213116	1,35898E-09	2,539585795
Внутри групп	0,837548387	55	0,015228152			
Итого	1,949333333	59				

Рис. 11.8

Замечание. Результатами работы программы (см. рисунок 11.8) также являются:

— групповые средние, которые совпадают со средними $\bar{y}_{(i)}$, указанными в таблице 11.10;

— исправленные выборочные групповые дисперсии $s^2_{(i)}$, которые связаны с групповыми дисперсиями $\hat{\sigma}^2_{(i)}$, приведенными в таблице 11.10, соотношением

$$s^2_{(i)} = \hat{\sigma}^2_{(i)} m_{*i}/(m_{*i} - 1), \quad i = 1, 2, \ldots, 5.$$

Например, $s^2_{(1)} = 0{,}013(3)$ (см. рис. 11.8), $\hat{\sigma}^2_{(1)} = 0{,}01$ (см. табл. 11.10) и $s^2_{(1)} = \hat{\sigma}^2_{(1)} m_{*1}/(m_{*1} - 1) = 0{,}01 \cdot 4/3 = 0{,}013(3)$. ◀

Приведем теорему, являющуюся выборочным аналогом теоремы 4.

Теорема 4′. *Среднее чисел* $(\bar{y}_{(i)} - \bar{y}_i^{\text{лин}})$, $i = 1, 2, \ldots,$ *равно нулю*

$$\sum_{i=1}^{\nu} (\bar{y}_{(i)} - \bar{y}_i^{\text{лин}}) m_{*i}/n = 0, \qquad (11.78)$$

а среднее их квадратов

$$\sum_{i=1}^{\nu} (\bar{y}_{(i)} - \bar{y}_i^{\text{лин}})^2 m_{*i}/n = \hat{\sigma}^2_Y (\hat{\rho}^2_{Y|X} - \hat{r}^2_{X,Y}). \qquad (11.79)$$

Здесь $\bar{y}_{(i)}$ — групповые средние, найденные по корреляционной таблице 11.9, а $\bar{y}_{(i)}^{\text{лин}} = \hat{a}_0 + \hat{a}_1 x'_i$, где $\hat{a}_0$ и $\hat{a}_1$ вычисляются по формулам (11.54).

Замечание. Ранее было доказано, что $\hat{\sigma}^2_{ELR\,Y|X}$, или среднее $\sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2/n$ (см. (11.59)), рассчитанное по результатам n наблюдений величины (X,Y), обладает свойством минимальности. Таким же свойством обладает и величина (11.79), рассчитанная по данным, сгруппированным в корреляционную таблицу 11.9: среднее квадратов ошибок, возникающая при замене групповых средних $\bar{y}_{(i)}$ числами $\bar{y}_i^{\text{лин}}$ меньше среднего квадратов ошибок, возникающих при замене $\bar{y}_{(i)}$ любой величиной, линейно зависящей от x'_i, но отличной от $\bar{y}_i^{\text{лин}} = \hat{a}_0 + \hat{a}_1 x'_i$. Это свойство является выборочным аналогом свойства минимальности дисперсии $D[M(Y|X) - M^{\text{лин}}(Y|X)]$, равной $M[M(Y|X) - M^{\text{лин}}(Y|X)]^2$ (см. (11.42)).

Предоставляем самостоятельно убедиться в правомерности равенств (11.78) и (11.79), используя данные табли-

цы 11.10, содержащей в том числе и $\bar{y}_{(i)}$, и $\bar{y}_i^{\text{лин}}$, $i = 1, 2, \ldots$ $\ldots, \nu = 5$.

Для выборочного корреляционного отношения имеет место следующее утверждение:

$$\hat{\rho}_{Y|X} \geqslant |\hat{r}_{X,Y}|, \tag{11.80}$$

вытекающее из равенства (11.79) и являющееся выборочным аналогом соотношения (11.43), и выполняются свойства, являющиеся выборочными аналогами свойств $\underline{1}^0$ и $\underline{5}^0$ генерального корреляционного отношения $\rho_{Y|X}$.

$\underline{1}^{0'}$. $0 \leqslant \hat{\rho}_{Y|X} \leqslant 1$. (11.81)

$\underline{5}^{0'}$. *Условие* $\hat{\rho}_{Y|X} = 1$ *является достаточным и необходимым условием функциональной зависимости наблюдаемых «игреков» от соответствующих им наблюдаемых «иксов», т. е. для того чтобы каждому «иксу» соответствовал единственный «игрек» и равным «иксам» соответствовали равные «игреки»* (в рамках табл. 11.9 это означает, что в каждом из столбцов «x_1'», «x_2'», $\ldots$ $\ldots$, «x_ν'» присутствует только одна клеточная частота; следствием является равенство всех выборочных групповых дисперсий нулю: $\hat{\sigma}^2_{(1)} = \hat{\sigma}^2_{(2)} = \ldots = \hat{\sigma}^2_{(\nu)} = 0$).

Выборочные аналоги свойств 2^0—4^0 генерального корреляционного отношения, связанных с понятием корреляционной зависимости, относящейся к генеральной совокупности, не рассматриваются.

Если $\rho_{Y|X}$ интерпретируют как меру близости зависимости Y от X к функциональной, то $\hat{\rho}_{Y|X}$ — мера близости зависимости наблюдаемых «игреков» от соответствующих «иксов» к функциональной зависимости.

Так как $\hat{\rho}_{Y|X}$ — выборочная оценка корреляционного отношения $\rho_{Y|X}$, то в дальнейшем, используя $\hat{\rho}_{Y|X}$, будем говорить, что $\hat{\rho}_{Y|X}$ — выборочная оценка меры близости зависимости Y от X к функциональной зависимости.

Более четкое числовое содержание по сравнению с $\hat{\rho}_{Y|X}$ имеет $\hat{\rho}^2_{Y|X}$ — ***выборочный коэффициент детерминации***. Из соотношения (11.74) получим

$$\hat{\rho}^2_{Y|X} = \hat{\sigma}^2_{\hat{R}\,Y|X} / \hat{\sigma}^2_Y, \tag{11.82}$$

эта формула является выборочным аналогом формулы (11.44); $\hat{\rho}^2_{Y|X}$ — выборочная оценка генерального коэффи-

циента детерминации, оценка доли дисперсии величины Y, объясняемой корреляционной зависимостью Y от X.

▶ **ПРИМЕР 11.3** (продолжение). $\hat{\rho}^2_{Y|X} = 0{,}755^2 = 0{,}570$ — такова, судя по 60 фирмам, оценка доли дисперсии объема произведенной продукции Y, объясняемой его корреляционной зависимостью от стоимости фондов X. В условиях этого же примера $\hat{r}^2_{X,Y} = 0{,}74^2 = 0{,}548$ — оценка доли дисперсии объема произведенной продукции, объясняемой его линейной корреляционной зависимостью от стоимости фондов. ◀

11.2.6. Выяснение по выборочным наблюдениям существования корреляционной зависимости и ее линейности. Точный ответ на вопрос, существует или нет корреляционная зависимость Y от X, т. е. изменяется с изменением x условное среднее $M(Y \mid x)$ или нет, и каков вид этой зависимости (линейный, параболический или другой), можно дать лишь когда известен закон распределения двумерной случайной величины (X, Y). Это подтверждается примером 11.1, в котором рассматривалась дискретная двумерная величина с известным законом распределения, заданным таблицей распределения вероятностей (см. табл. 11.1).

Ответить на более «узкий» вопрос, существует или нет корреляционная зависимость Y от X и является ли она линейной, можно и не зная закона распределения величины (X, Y). Достаточно знать генеральные коэффициент корреляции $r_{X,Y}$ и корреляционное отношение $\rho_{Y|X}$. Последовательность ответа на этот вопрос изображена на рисунке 11.4.

Наконец, если известен только генеральный коэффициент корреляции $r_{X,Y}$ и $r_{X,Y} \neq 0$, то можно лишь утверждать, что корреляционная зависимость Y от X существует, но линейная она или нет — сказать нельзя (см. рис. 11.5).

Располагая лишь выборочными наблюдениями, дать точный ответ ни на один из поставленных выше вопросов нельзя. Однако проверка определенных статистических гипотез может внести достаточную ясность.

Рассмотрим два случая:

1) результаты (x_k, y_k) n наблюдений двумерной величины (X, Y) сгруппированы в корреляционную таблицу 11.9, в которой число вариантов «иксов» (число столбцов) равно ν, и по этой таблице рассчитаны коэффициент корреляции $\hat{r}_{X|Y}$ и корреляционное отношение $\hat{\rho}_{Y|X}$;

2) по наблюдениям вычислен только коэффициент $\hat{r}_{X,Y}$, а группировка наблюдений в корреляционную таблицу не проводилась или ее нельзя было провести (например, из-за малости n), следовательно, значение $\hat{\rho}_{Y|X}$, вычисляемое по корреляционной таблице, неизвестно.

Замечание. Приступая к проверке статистических гипотез, следует иметь в виду, что результат наблюдения двумерной случайной величины (X, Y), так же как и одномерной, можно интерпретировать следующим образом: либо это фактически полученная пара чисел (x_k, y_k), либо мыслимый, возможный случайный результат (тогда его обозначают (X_k, Y_k)). В силу этого, любую выборочную характеристику двумерной случайной величины (X, Y), в том числе и $\hat{\rho}_{Y|X}$ и $r_{X,Y}$, можно интерпретировать по-разному: это число, если речь идет о конкретных числовых результатах наблюдений, или случайная величина, если речь идет о мыслимых результатах наблюдений. Какой вариант используется, ясно из контекста.

Случай 1. Результаты наблюдений двумерной случайной величины (X, Y) сгруппированы в корреляционную таблицу 11.9.

Чтобы выяснить, существует ли корреляционная зависимость Y от X и является ли она линейной, следует (согласно схеме, изображенной на рис. 11.4) проверить гипотезу H_0: $\rho_{Y|X} = 0$ и при ее отклонении проверить гипотезу H_0: $\rho_{Y|X} = |r_{Y|X}|$. Рассмотрим способы проверки этих гипотез.

1) Проверим гипотезу

$$H_0: \rho_{Y|X} = 0 \qquad (11.83)$$

об отсутствии корреляционной зависимости Y от X при альтернативе H_1: $\rho_{Y|X} > 0$ (напомним, что всегда $\rho_{Y|X} \geqslant 0$). Прежде чем излагать способ проверки гипотезы (11.83), отметим, что она, в силу свойства $\underline{2}^0$ генерального корреляционного отношения, эквивалентна гипотезе

$$H_0: M(Y|x) = MY, \qquad (11.84)$$

где x — любое допустимое значение случайной величины X.

Проверка гипотезы (11.83) [следовательно, и гипотезы (11.84)] проводится с помощью критической статистики

$$F = \frac{\hat{\rho}^2_{Y|X}/(\nu - 1)}{(1 - \hat{\rho}^2_{Y|X})/(n - \nu)}, \qquad (11.85)$$

которая при выполнении гипотезы имеет F-распределение с числами степеней свободы $\nu - 1$ и $n - \nu$. Для ответа на вопрос, отклонить или принять гипотезу H_0: $\rho_{Y|X} = 0$, следует вычислить значение f статистики (11.85) и сравнить его с

критической точкой $f_{\nu-1,\,n-\nu,\,\alpha}$, найденной в таблице П. 5 при $p=\alpha$, $k_1=\nu-1$ и $k_2=n-\nu$. При $f>f_{\nu-1,\,n-\nu,\,\alpha}$ гипотезу H_0 отклоняют; в противном случае — принимают.

Принятие гипотезы H_0: $\rho_{Y|X}=0$ является аргументом в пользу отсутствия корреляционной зависимости Y от X; отклонение гипотезы H_0: $\rho_{Y|X}=0$ (т. е. принятие гипотезы H_1: $\rho_{Y|X}>0$) — аргумент в пользу существования корреляционной зависимости Y от X; в этом случае, т. е. при $\rho_{Y|X}>0$, надо выяснить, линейная эта зависимость или нелинейная (см. рис. 11.4) — перейти к проверке гипотезы H_0: $\rho_{Y|X}=|r_{X,\,Y}|$.

2) Проверим гипотезу

$$H_0\colon \rho_{Y|X}=|r_{X,\,Y}| \qquad (11.86)$$

о том, что корреляционная зависимость Y от X линейная, при альтернативе H_1: $\rho_{Y|X}>|r_{X,\,Y}|$ (всегда $\rho_{Y|X}\geqslant|r_{X,\,Y}|$).

Прежде чем излагать способ проверки гипотезы (11.86), обратим внимание на то, что ее проверяют только в том случае, когда корреляционное отношение $\rho_{Y|X}>0$ (корреляционная зависимость Y от X существует); также напомним, что, согласно свойству 3^0 генерального корреляционного отношения, условие $\rho_{Y|X}=|r_{X,\,Y}|>0$ является достаточным и необходимым условием линейной корреляционной зависимости Y от X, т. е. условием выполнения равенства $M(Y\,|\,x)=a_0+a_1x$, где x — любое допустимое значение случайной величины X, а a_0 и $a_1\neq 0$ рассчитывают по формулам (11.19) и (11.18). Так как предполагается, что $\rho_{Y|X}>0$, то при выполнении гипотезы (11.86) $\rho_{Y|X}=|r_{X,\,Y}|>0$. Поэтому гипотеза (11.86) эквивалента гипотезе

$$H_0\colon M(Y\,|\,x)=a_0+a_1x, \qquad (11.87)$$

где x — любое допустимое значение X.

Проверка гипотезы (11.86) (следовательно, и гипотезы (11.87)) проводится с помощью критической статистики

$$F=\frac{(\widehat{\rho}^{\,2}_{Y|X}-\widehat{r}^{\,2}_{Y|X})/(\nu-2)}{(1-\widehat{\rho}^{\,2}_{Y|X})/(n-\nu)}, \qquad (11.88)$$

которая при выполнении гипотезы имеет F-распределение с $k_1=\nu-2$ и $k_2=n-\nu$ степенями свободы. Поэтому гипотезу о линейности корреляционной зависимости Y от X [гипотезу (11.86) или (11.87)] не принимают, если числовое значение критической статистики F больше критической точки $f_{\nu-2,\,n-\nu,\,\alpha}$, найденной в таблице П. 5 при $p=\alpha$,

$k_1 = \nu - 2$, $k_2 = n - \nu$; в противном случае гипотезу о линейности корреляционной зависимости не отклоняют.

▶ **ПРИМЕР 11.3** (продолжение). По данным корреляционной таблицы 11.10 выясним, существует или нет корреляционная зависимость Y от X, т. е. проверим гипотезу (11.83). Если зависимость подтвердится, то выясним, линейная она или нелинейная, т. е. проверим гипотезу (11.86). Примем $\alpha = 0{,}05$. Напомним, что в примере $n = 60$, $\nu = 5$, $\hat{r}_{X,Y} = 0{,}74$, $\rho_{Y|X} = 0{,}755$.

1) Поверим гипотезу H_0: $\rho_{Y|X} = 0$ при альтернативе H_1: $\rho_{Y|X} > 0$. Значение критической статистики (11.85) равно $\frac{0{,}755^2/4}{(1-0{,}755^2)/55} = 18{,}23$. Найденная по таблице П.5 критическая точка $f_{0{,}05;\,4;\,55} \approx 2{,}53$. Так как $18{,}23 > 2{,}53$, то гипотезу H_0 не принимаем; принимаем гипотезу H_1: $\rho_{Y|X} > 0$ о существовании корреляционной зависимости произведенной на фирме продукции Y от стоимости основных фондов X.

2) Проверим гипотезу H_0: $\rho_{Y|X} = |r_{X,Y}|$ о линейности корреляционной зависимости. Значение критической статистики (11.88) равно:

$$\frac{(0{,}755^2 - 0{,}74^2)/(5-2)}{(1-0{,}755^2)/(60-5)} = 0{,}956$$

— это число меньше критической точки $f_{\nu-2,\,n-\nu,\,\alpha} = f_{3;\,55;\,0{,}05} \approx 2{,}76$, поэтому гипотезу о линейности корреляционной зависимости произведенной фирмой продукции Y от стоимости основных фондов X принимаем, т. е. при любом допустимом значении x величины X

$$M(Y \mid x) \underset{(11.20)}{=} a_0 + a_1 x.$$

Делаем вывод: оценкой среднего объема продукции, произведенной фирмами, на каждой из которых стоимость основных фондов равна x, т. е. оценкой $M(Y \mid x)$, является

$$\bar{y}_x \underset{(11.53)}{=} \hat{a}_0 + \hat{a}_1 x \underset{(11.55)}{=} 0{,}1 + 0{,}071x. \quad ◀$$

Случай 2. По наблюдениям (x_k, y_k), $k = 1, 2, \ldots, n$, двумерной случайной величины (X, Y) рассчитан коэффициент корреляции $\hat{r}_{X,Y}$, но группировка наблюдений в корреляционную таблицу не производилась, следовательно, корреляционное отношение $\hat{\rho}_{Y|X}$, которое рассчитывается по данным корреляционной таблицы, неизвестно.

Проверим гипотезу

$$H_0: r_{X,Y} = 0 \qquad (11.89)$$

о том, что генеральный коэффициент корреляции равен нулю, при альтернативе $H_1: r_{X,Y} \neq 0$, или $H_1: |r_{X,Y}| > 0$.

Если гипотеза $H_0: r_{X,Y} = 0$ не будет отклонена, то это аргумент в пользу отсутствия линейной корреляционной зависимости Y от X (рис. 11.5). Последнее означает:

— либо отсутствие корреляционной зависимости Y от X, т. е. выполнение при любом допустимом значении x величины X равенства $M(Y \mid x) = MY$ (с изменением x условное среднее $M(Y \mid x)$ не меняется);

— либо наличие нелинейной корреляционной зависимости Y от X, т. е. нелинейность функции $M(Y \mid x) = \varphi(x)$ регрессии Y на x.

Вопрос о том, какая из этих двух ситуаций имеет место, в случае принятия гипотезы $H_0: r_{X,Y} = 0$ остается открытым.

Отклонение гипотезы $H_0: r_{X,Y} = 0$ (принятие гипотезы $H_1: |r_{X,Y}| > 0$) — аргумент в пользу существования корреляционной зависимости Y от X (см. рис. 11.5), т. е. в пользу того, что ожидаемое в среднем значение $M(Y \mid x)$ величины Y изменяется с изменением x. Вопрос о характере этого изменения, т. е. о том, является ли функция регрессии линейной или нелинейной, в случае принятия гипотезы $H_1: |r_{X,Y}| > 0$ остается открытым.

Аргументом в пользу линейной функции регрессии (в пользу того, что она имеет вид (11.20), или, по крайней мере, в пользу того, что ошибка, возникающая при замене нелинейной функции регрессии линейной, мала) является близость абсолютной величины выборочного коэффициента корреляции, рассчитанного по достаточно большому числу наблюдений n, к единице: $|\hat{r}_{X,Y}| \approx 1$.

Для проверки гипотезы $H_0: r_{X,Y} = 0$ используют критическую статистику

$$T = \frac{\hat{r}_{X,Y}}{\sqrt{(1 - \hat{r}^2_{X,Y})/(n-2)}}, \qquad (11.90)$$

имеющую при выполнении гипотезы $H_0: r_{X,Y} = 0$ T-распределение с $(n-2)$-мя степенями свободы. Область принятия гипотезы $H_0: r_{X,Y} = 0$ при альтернативе $H_1: r_{X,Y} \neq 0$ задается неравенством $|T| < t_{n-2,\alpha}$; область ее отклонения (принятия гипотезы $H_1: r_{X,Y} \neq 0$) — неравенством $|T| > t_{n-2,\alpha}$,

где число $t_{n-2,\alpha}$ определяют по таблице П. 4 при $k = n - 2$ и $p = \alpha$.

Критерии проверки гипотезы H_0: $r_{X,Y} = 0$, основанные на использовании критической статистики (11.90), приведены в последней строке таблицы 9.2.

▶ **ПРИМЕР 11.3** (продолжение). Ранее по данным корреляционной таблицы 11.10 было выяснено, что корреляционная зависимость объема Y произведенной фирмой продукции от стоимости X имеющихся у фирмы фондов существует (гипотеза H_0: $\rho_{Y|X} = 0$ была отклонена) и что эта зависимость линейная (гипотеза H_0: $\rho_{Y|X} = |r_{X,Y}|$ принята).

Проверим гипотезу H_0: $r_{X,Y} = 0$, используя критическую статистику (11.90). Вспомнив, что $\hat{r}_{X,Y} = 0{,}74$, найдем значение критической статистики

$$t = 0{,}74/\sqrt{(1 - 0{,}74^2)/(60 - 2)} = 8{,}4$$

— это значение больше числа $t_{n-2,\alpha} = t_{58;\,0{,}05} \approx 2{,}0$, найденного по таблице П. 4. Поэтому гипотезу H_0: $r_{X,Y} = 0$ отклоняем, что является аргументом в пользу существования корреляционной зависимости Y от X. Однако, отклонив гипотезу H_0, сделать заключение о линейности этой зависимости нельзя, кроме того $|\hat{r}_{X,Y}| = 0{,}74$, что, вообще говоря, не близко к единице (доля дисперсии величины Y, объясняемая линейной корреляционной зависимостью ее от величины X, оценивается только в $\hat{r}^2_{X,Y} \cdot 100\% = 55\%$; на долю прочих причин вариации значений величины Y приходится 45%). Только зная выборочное корреляционное отношение $\hat{\rho}_{Y|X}$, можно проверить гипотезу о линейности корреляционной зависимости Y от X. В условиях данного примера $\hat{\rho}_{Y|X} = 0{,}755$, и гипотеза о линейности корреляционной зависимости Y от X была принята. ◀

11.2.7. Понятие о частном и множественном коэффициентах корреляции. Выше в качестве меры линейности зависимости между случайными величинами Y и X был использован коэффициент корреляции $r_{X,Y}$ (точнее $|r_{X,Y}|$), который часто называют *парным коэффициентом корреляции*. Зависимость между случайными величинами X и Y может быть обусловлена не только их взаимовлиянием, но и влиянием и на X, и на Y других случайных величин, поэтому, наряду с парным коэффициентом корреляции $r_{X,Y}$, рассматривают частные коэффициенты корреляции — *меры линейности зависимости между X и Y при условии иск-*

лючения (элиминирования) линейного влияния на каждую из величин X и Y одной или нескольких случайных величин.

Рассмотрим *частный коэффициент корреляции первого порядка* $r_{(X,Y)|Z}$, являющийся измерителем линейности зависимости между X и Y при условии исключения линейного влияния на X и Y величины Z.

Приведем формулу вычисления частного коэффициента корреляции $r_{(X,Y)|Z}$

$$r_{(X,Y)|Z} = \frac{r_{X,Y} - r_{X,Z}r_{Y,Z}}{\sqrt{(1 - r_{X,Z}^2)(1 - r_{Y,Z}^2)}}. \qquad (11.91)$$

Выборочной оценкой частного коэффициента корреляции является выборочный частный коэффициент корреляции

$$\hat{r}_{(X,Y)|Z} = \frac{\hat{r}_{X,Y} - \hat{r}_{X,Z}\hat{r}_{Y,Z}}{\sqrt{(1 - \hat{r}_{X,Z}^2)(1 - \hat{r}_{Y,Z}^2)}}. \qquad (11.92)$$

Чтобы определить $\hat{r}_{(X,Y)|Z}$ по формуле (11.91), надо знать три парных коэффициента корреляции: $\hat{r}_{X,Y}$, $\hat{r}_{X,Z}$, $\hat{r}_{Y,Z}$. Для их вычисления можно использовать программу **«Корреляция»** пакета **«Анализ данных** Microsoft Excel. Исходными данными для этой программы являются введенные в рабочий лист три столбца (или три строки) «X», «Y», «Z» результатов (x_k, y_k, z_k), $k = 1, 2, \ldots, n$, наблюдений трехмерной случайной величины (X, Y, Z). Результаты работы программы выдаются в следующем виде:

1,000		
$\hat{r}_{Y,X}$,	1,000	
$\hat{r}_{Z,X}$,	$\hat{r}_{Z,Y}$,	1,000

Напомним, что коэффициент корреляции любых случайных величин U и V симметричен относительно этих величин, т. е. $\hat{r}_{U,V} = \hat{r}_{V,U}$, и что $\hat{r}_{U,U} = 1$.

Приведем алгоритм расчета выборочного коэффициента $\hat{r}_{(X,Y)|Z}$, вытекающий из данного выше определения частного коэффициента.

По результатам (x_k, y_k, z_k), $k = 1, 2, \ldots, n$, наблюдений трехмерной случайной величины (X, Y, Z) следует:

— рассчитать линейную регрессию X на Z и найти значения $\bar{x}_k^{\text{лин}} = \hat{b}_0 + \hat{b}_1 z_k$, где $\hat{b}_1 = \hat{r}_{X,Z}\frac{\hat{\sigma}_X}{\hat{\sigma}_Z}$, $\hat{b}_0 = \bar{x} - \hat{b}_1\bar{z}$;

— элиминировать линейное влияние Z на X, т. е. найти значения $x'_k = x_k - \bar{x}_k^{\text{лин}}$;

— рассчитать линейную регрессию Y на Z и найти значения $y_k^{\text{лин}} = \hat{c}_0 + \hat{c}_1 z_k$;

— элиминировать линейное влияние Z на Y, т. е. найти значения $y'_k = y_k - \bar{y}_k^{\text{лин}}$;

— рассчитать парный коэффициент корреляции $\hat{r}_{X', Y'} = \frac{\overline{x'y'} - \overline{x'}\,\overline{y'}}{\hat{\sigma}_{X'}\hat{\sigma}_{Y'}}$, где $\overline{x'y'} = \sum_{k=1}^{n} x'_k y'_k/n$, $\bar{x}' = \sum_{k=1}^{n} x'_k/n$, $\hat{\sigma}_{X'} = \sqrt{\sum_{k=1}^{n}(x'_k - \bar{x}')^2/n}$, и т. д.

Парный коэффициент $\hat{r}_{X', Y'}$ и является частным коэффициентом корреляции между X и Y при устранении линейного влияния на них величины Z, т. е. $\hat{r}_{X', Y'} = \hat{r}_{(X, Y)|Z}$.

Можно доказать, что рассчитанный таким способом выборочный частный коэффициент корреляции равен коэффициенту, найденному по формуле (11.91).

Отметим одно из свойств (генерального) частного коэффициента корреляции:

$$|r_{(X, Y)|Z}| \leqslant 1,$$

а также следующее:

если $|r_{(X, Y)|Z}| < |r_{X, Y}|$, то это свидетельствует об «уменьшении линейной взаимосвязи» между величинами X и Y после исключения линейного влияния на них величины Z, или, иначе, что взаимосвязь между X и Y возникает частично (или полностью) из-за их взаимосвязи с величиной Z (из этого не следует, что Z — одна из причин (или причина) взаимосвязи между X и Y; предположение о причинности должно всегда иметь внестатистические обоснования); если $|r_{(X, Y)|Z}| > |r_{X, Y}|$, то это говорит о том, что влияние величины Z на X и Y «маскирует» взаимосвязь X и Y.

Аналогичные свойства и утверждения (в интерпретации «выборки») имеют место и для соответствующих выборочных коэффициентов корреляции. Прежде чем на основании выборочного частного коэффициента корреляции $\hat{r}_{(X, Y)|Z}$ делать какие-либо выводы о влиянии Z на X и Y, следует проверить гипотезу H_0: $r_{(X, Y)|Z} = 0$ о том, что генеральный коэффициент равен нулю. При ее проверке используют критическую статистику, аналогичную статистике (11.90),

$$T = \frac{\hat{r}_{(X, Y)|Z}}{\sqrt{(1 - \hat{r}^2_{(X, Y)|Z})/(n-3)}}, \qquad (11.93)$$

имеющую при выполнении гипотезы H_0: $r_{(X,Y)|Z} = 0$ T-распределение с числом степеней свободы $k = n - 3$. Область принятия гипотезы H_0 при альтернативе H_1: $r_{(X,Y)|Z} \neq 0$ определяется неравенством $|T| < t_{n-3,\alpha}$, где число $t_{n-3,\alpha}$ находят по таблице П. 4 при $k = n - 3$ и $p = \alpha$; область ее отклонения определяется неравенством $|T| > t_{n-3,\alpha}$. При отклонении гипотезы H_0: $r_{(X,Y)|Z} = 0$ говорят, что *выборочный коэффициент* $\hat{r}_{(X,Y)|Z}$ *статистически значим.*

Приведем следующий пример, подтверждающий необходимость знания и парного, и частного коэффициентов корреляции и, вместе с тем, показывающий абсурдность их толкования как показателей причинных зависимостей. «При анализе большого числа наблюдений, относящихся к отливке труб на сталелитейных заводах, была установлена значимая положительная корреляционная связь между временем плавки X и процентом забракованных труб Y. Дать какое-либо причинное толкование этой связи было невозможно, поэтому рекомендации ограничить продолжительность плавки для снижения процента забракованных труб малосостоятельны... Позже выяснилось, что объяснение значимой положительной корреляции следует искать в одновременном влиянии на оба показателя: и на продолжительность плавки, и на процент брака третьего показателя — типа используемого сырья Z ... Если влияние типа сырья Z исключить, то никакой значимой корреляционной связи между временем плавки X и процентом забракованных труб Y мы не обнаружим» [2].

Для «увязки» изложенной ситуации с введенными выше понятиями запишем ее на «языке формул»:

— значимая положительная корреляция между X и Y означает, что $\hat{r}_{X,Y} > 0$ и что в результате проверки гипотезы H_0: $r_{X,Y} = 0$ при альтернативе H_1: $r_{X,Y} > 0$ гипотеза H_0 была отклонена;

— отсутствие значимой корреляции между X и Y после исключения влияния на них величины Z означает, что в результате проверки гипотезы H_0: $r_{(X,Y)|Z} = 0$ при альтернативе H_1: $r_{(X,Y)|Z} \neq 0$ была принята гипотеза H_0 — это, в терминах рассматриваемого примера, означает, что взаимосвязь между временем плавки и процентом бракованных труб определяется типом используемого сырья.

В заключение еще раз заметим, что наряду с частным коэффициентом корреляции первого порядка $r_{(X,Y)|Z}$ могут быть рассмотрены частные коэффициенты корреляции больши́х порядков. Например, частный коэффициент кор-

реляции второго порядка $r_{(X,Y)|(Z,U)}$, являющийся мерой линейности зависимости между X и Y при условии исключения линейного влияния на каждую из этих величин одновременно двух величин: и Z, и U.

Множественный коэффициент корреляции является мерой линейности зависимости одной случайной величины от двух и более случайных величин. Приведем только формулу множественного коэффициента корреляции $R_{Y|(X,Y)}$, измеряющего линейность зависимости случайной величины Y от случайных величин X и Z:

$$R_{Y|(X,Z)} = \sqrt{\frac{r_{Y,X}^2 + r_{Y,Z}^2 - 2r_{Y,X}r_{Y,Z}r_{X,Z}}{1 - r_{X,Z}^2}}. \qquad (11.94)$$

Для множественного коэффициента корреляции $R_{Y|(X,Y)}$ справедливы следующие утверждения:

$$0 \leqslant R_{Y|(X,Z)} \leqslant 1;$$

$$R_{Y|(X,Z)} \geqslant \max\{|r_{Y,X}|, |r_{Y,Z}|, |r_{(Y,X)|Z}|, |r_{(Y,Z)|X}|\}.$$

Выборочный множественный коэффициент корреляции

$$\widehat{R}_{Y|(X,Z)} = \sqrt{\frac{\widehat{r}_{Y,X}^2 + \widehat{r}_{Y,Z}^2 - 2\widehat{r}_{Y,X}\widehat{r}_{Y,Z}\widehat{r}_{X,Z}}{1 - \widehat{r}_{X,Z}^2}}. \qquad (11.95)$$

Для него справедливы утверждения, аналогичные приведенным выше утверждениям относительно генерального коэффициента $R_{Y|(X,Z)}$.

Более общая формула выборочного множественного коэффициента корреляции, которая может быть использована как при вычислении $\widehat{R}_{Y|(X,Z)}$ [дающая такой же результат, что и формула (11.95)], так и при нахождении коэффициентов, измеряющих зависимость Y от более чем двух случайных величин, будет приведена в п. 11.3.3. Там же рассмотрен критерий проверки гипотезы о равенстве нулю (генерального) множественного коэффициента корреляции.

§ 11.3. Задачи регрессии

11.3.1. Постановка вопроса. В § 11.2 с привлечением понятия линейной функции регрессии были подробно изучены свойства и смысл коэффициента корреляции — основного показателя взаимозависимости компонент X и Y двумерной случайной величины (X, Y). Там же с привлечением понятия функции регрессии (не обязательно линейной) рассмотрены свойства и смысл корреляционного отноше-

ния — показателя зависимости одной компоненты от другой. И коэффициент корреляции, и корреляционное отношение являются характеристиками двумерной случайной величины (X, Y) — величины, обе компоненты которой случайны. Поэтому в функциях регрессии, используемых при изучении этих показателей, и X, и Y интерпретировались как случайные величины.

Вместе с тем в практических задачах, связанных с изучением стохастических зависимостей (зависимостей, в которых каждому набору значений независимых переменных соответствует множество значений зависимой переменной, причем сказать заранее, какое именно значение примет зависимая переменная, нельзя), в число независимых переменных могут быть включены и детерминированные (а не только случайные) величины, а зависимая переменная всегда случайная величина.

Пусть, например, изучается зависимость ежегодных вкладов Y_t, $t = 1, 2, \ldots, T$, населения в сберегательные банки за T лет от t и годового уровня инфляции X_t. Здесь независимая переменная t — детерминированная величина, X_t — случайная; случайность зависимой переменной Y_t обусловлена не столько ее зависимостью от случайной переменной X_t, сколько тем, что значение величины Y_t формируется под влиянием различных случайных обстоятельств, не учтенных в рассматриваемой зависимости и заранее не предсказуемых. Практический интерес представляет решение следующих задач: при заданном значении x_t инфляции в году t дать точечный и интервальный прогноз вкладов населения в этом году; или дать прогноз изменения этих вкладов при увеличении инфляции и т. д. При такой постановке задач нет необходимости учитывать вероятность того, что инфляция примет значение, равное x_t, т. е. X_t можно рассматривать как «неслучайную» величину.

Введем следующие обозначения: Y — изучаемая случайная величина (зависимая переменная); $x_1, x_2, \ldots, x_m$ — поддающиеся контролю факторы (их называют ***регрессорами***), влияющие на Y, но не исчерпывающие совокупности факторов, которые определяют значение величины Y; при проведении наблюдений (экспериментов) регрессоры могут принимать определенные (а не зависящие от случая) значения, в определенных пределах их изменения, предписываемые постановкой экспериментов; ε — эффект влияния на Y неконтролируемых случайных факторов, в результате которого Y будет испытывать случайные колебания.

Функция

$$M(Y \mid x_1, x_2, \ldots, x_m) = \varphi(x_1, x_2, \ldots, x_m), \qquad (11.96)$$

описывающая изменение среднего значения величины Y при изменении значений регрессоров $x_1, x_2, \ldots, x_m$, называется ***функцией регрессии Y на $x_1, x_2, \ldots, x_m$*** или ***множественной регрессией.***

Если регрессор один — x, то функцию регрессии

$$M(Y \mid x) = \varphi(x) \qquad (11.97)$$

называют ***парной регрессией.***

Так как регрессоры принимают значения, не зависящие от случая, то и значения функции (11.96), как и функции (11.97), неслучайны и вполне определены, т. е. условное математическое ожидание $M(Y \mid x_1, x_2, \ldots, x_m)$, как и $M(Y \mid x)$, — детерминированная величина.

Замечание. В § 11.1 и 11.2 рассмотрена двумерная случайная величина (X, Y). В силу случайности величины X функция регрессии $M(Y \mid X) = \varphi(X)$ принимала значения, зависящие от случая, т. е. условное математическое ожидание $M(Y \mid X)$ — это случайная величина, имеющая (как и любая случайная величина) математическое ожидание $M[M(Y \mid X)]$ и дисперсию $D[M(Y \mid X)]$, для обозначения которой в общем случае использовался символ $\sigma^2_{R\,Y|X}$ [см. (11.36)], а в случае линейной функции $\varphi(X)$ — символ $\sigma^2_{LR\,Y|X}$ [см. (11.23)].

Говорить о математическом ожидании и дисперсии детерминированной величины (каковой является и $M(Y \mid x)$, и $M(Y \mid x_1, x_2, \ldots, x_n)$), не имеет смысла, поскольку, как известно, они равны соответственно самой величине и нулю.

Проведем n наблюдений — экспериментов. Пусть в k-м эксперименте, $k = 1, 2, \ldots, n$, регрессоры $x_1, x_2, \ldots, x_m$ фиксировались на уровнях $x_{k1}, x_{k2}, \ldots, x_{km}$, а случайная величина Y приняла значение, равное y_k (табл. 11.18).

Таблица 11.18

Номер наблюдения, k	x_1	x_2	...	x_m	Y
1	x_{11}	x_{12}	...	x_{1m}	y_1
2	x_{21}	x_{22}	...	x_{2m}	y_2
...	...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{nm}	y_n

Напомним, что результат k-го наблюдения величины Y можно интерпретировать и как число y_k, и как случайную величину, которую обозначают через Y_k; значения регрессоров — всегда числа.

Для описания механизма формирования случайного результата Y_k наблюдения величины Y используют ***модель множественного регрессионного анализа***:

$$Y_k = \varphi(x_{k1}, x_{k2}, \ldots, x_{km}) + \varepsilon_k,$$
$$k = 1, 2, \ldots, n, \qquad (11.98)$$

где $\varphi(x_{k1}, x_{k2}, \ldots, x_{km}) = M(Y \mid x_{k1}, x_{k2}, \ldots, x_{km})$ — значение функции регрессии (11.96) при значениях $x_{k1}, x_{k2}, \ldots, x_{km}$ регрессоров $x_1, x_2, \ldots, x_m$; ε_k — эффект влияния на Y в k-м наблюдении неконтролируемых случайных факторов, или, иначе, случайная ошибка регрессии (*error of regression*), возникающая при замене Y_k значением функции регрессии $\varphi(x_{k1}, x_{k2}, \ldots, x_{km})$.

Предполагают, что:

— каждая из ошибок ε_k имеет нормальный закон распределения с нулевым математическим ожиданием и не зависящей от значений регрессоров и номера наблюдения дисперсией, которую, учитывая, что ε_k — ошибка регрессии, будем обозначать через $\sigma^2_{ER\,Y|x_1, x_2, \ldots, x_m}$; т. е.

$$\varepsilon_k = N(M\varepsilon_k = 0;\ D\varepsilon_k = \sigma^2_{ER\,Y|x_1, x_2, \ldots, x_m}), \qquad (11.99)$$
$$k = 1, 2, \ldots, n;$$

— ошибки $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ — независимые случайные величины. (11.100)

Используя соотношения (11.98), в которых случайными являются только величины Y_k и ε_k, нетрудно убедиться в том, что требования (11.99) и (11.100) тождественны следующим требованиям к случайным результатам Y_k наблюдений величины Y:

— $Y_k = N(MY_k = M(Y \mid x_{k1}, x_{k2}, \ldots, x_{km};$
$DY_k = \sigma^2_{ER\,Y|x_1, x_2, \ldots, x_m}), k = 1, 2, \ldots, n; \qquad (11.101)$

— случайные величины $Y_1, Y_2, \ldots, Y_n$ независимы. (11.102)

В дальнейшем будем предполагать, что вид функции регрессии $M(Y \mid x_1, x_2, \ldots, x_m) = \varphi(x_1, x_2, \ldots, x_m)$ известен.

Основные задачи регрессионного анализа состоят в том, чтобы, используя результаты наблюдений, представленные в таблице 11.18, оценить:

— параметры модели (11.98), включающие постоянные величины, входящие в $\varphi(x_1, x_2, \ldots, x_m)$, числовые значения которых не известны, и дисперсию $\sigma^2_{ER\,Y|x_1, x_2, \ldots, x_m}$;

— среднее и «индивидуальное» значение величины Y при заданных значениях $x_1^{(0)}, x_2^{(0)}, \ldots, x_m^{(0)}$ регрессоров $x_1, x_2, \ldots, x_m$.

11.3.2. Парная линейная регрессия. Изучается влияние на случайную величину Y одной (неслучайной) переменной x и функция регрессии (11.97) имеет линейный вид относительно параметров a_0 и a_1 и переменной x, т. е.

$$M(Y \mid x) = M^{\text{лин}}(Y \mid x),$$

где

$$M^{\text{лин}}(Y \mid x) = a_0 + a_1 x. \qquad (11.103)$$

Проведем n наблюдений — экспериментов. Пусть в k-м эксперименте ($k = 1, 2, \ldots, n$) регрессор x фиксировался на уровне x_k, а величина Y принимала значение, равное y_k (табл. 11.19).

Таблица 11.19

Номер наблюдения, k	1	2	...	n
x	x_1	x_2	...	x_n
Y	y_1	y_2	...	y_n

Модель (11.98) формирования случайного результата Y_k наблюдения величины Y, при значении регрессора x, равном x_k, запишем в виде $Y_k = M^{\text{лин}}(Y \mid x_k) + \varepsilon_k$, или, учитывая соотношение (11.103), в виде

$$Y_k = a_0 + a_1 x_k + \varepsilon_k,\ k = 1, 2, \ldots, n, \qquad (11.104)$$

где ε_k, $k = 1, 2, \ldots, n$, — случайные ошибки линейной регрессии Y на x (*error of linear regression*), каждая из которых имеет нормальный закон распределения с нулевым

математическим ожиданием и не зависящей от значения регрессора x и номера наблюдения дисперсией (которую обозначим $\sigma^2_{ELR\ Y|x}$), и при этом $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ — независимые величины.

Сформулированные требования к величинам ε_k тождественны следующим требованиям:

$$— Y_k = N(MY_k = a_0 + a_1 x_k,\ DY_k = \sigma^2_{ELR\ Y|x}),\ k = 1, 2, \ldots, n; \tag{11.105}$$

$$— Y_1, Y_2, \ldots, Y_n \text{ — независимые случайные величины.} \tag{11.106}$$

В этом нетрудно убедиться, если использовать соотношение (11.104) и учесть, что $(a_0 + a_1 x_k)$ — постоянная при каждом x_k величина.

Задачи парного линейного регрессионного анализа состоят в том, чтобы, используя результаты наблюдений, представленные в таблице 11.19, оценить:

— параметры модели (11.104), включающие постоянные величины a_0, a_1 и дисперсию $\sigma^2_{ELR\ Y|x}$;

— среднее и «индивидуальное» значение величины Y при заданном значении x_0 регрессора x.

Оценивание параметров a_0, a_1 *и* $\sigma^2_{ELR\ Y|x}$ *модели* (11.104) *по данным таблицы* 11.19. Согласно методу максимального правдоподобия, оценки $\hat{a}_0$, $\hat{a}_1$ параметров a_0 и a_1 при выполнении условий (11.105) и (11.106) являются решением системы нормальных уравнений [см. (11.64)]:

$$\begin{cases} a_0 n + a_1 \sum\limits_{k=1}^{n} x_k = \sum\limits_{k=1}^{n} y_k, \\ a_0 \sum\limits_{k=1}^{n} x_k + a_1 \sum\limits_{k=1}^{n} x_k^2 = \sum\limits_{k=1}^{n} y_k x_k, \end{cases} \tag{11.107}$$

равным

$$\hat{a}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}, \qquad \hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x}, \tag{11.108}$$

где $\overline{xy} = \sum\limits_{k=1}^{n} x_k y_k / n$, $\bar{x} = \sum\limits_{k=1}^{n} x_k / n$, $\overline{x^2} = \sum\limits_{k=1}^{n} x_k^2 / n$ [аналогичное решение было получено для системы (11.64)], а оценка параметра $\sigma^2_{ELR\ Y|x}$ имеет вид

$$\hat{\sigma}^2_{ELR\ Y|x} = \sum_{k=1}^{n} (y_k - \hat{a}_0 - \hat{a}_1 x_k)^2 / n. \tag{11.109}$$

≫ Действительно, так как, в силу (11.105) и (11.106), величины Y_k нормально распределены и независимы, то логарифмическая функция правдоподобия, по аналогии с (8.36), имеет следующий вид:

$$\ln L = n \ln (1/\sqrt{2\pi}) + n \ln (1/\sigma_{ELR\,Y|x}) - \\ - \sum_{i=1}^{n} (y_k - a_0 - a_1 x_k)^2 / (2\sigma^2_{ELR\,Y|x}),$$

являясь функцией переменных a_0, a_1 и $\sigma^2_{ELR\,Y|x}$. Нетрудно догадаться, что в точке максимума [обозначим ее через ($\hat{a}_0$, $\hat{a}_1$, $\hat{\sigma}^2_{ELR\,Y|x}$)] функции $\ln L$:

- функция

$$F(a_0, a_1) = \sum_{i=1}^{n} (y_k - a_0 - a_1 x_k)^2$$

должна достигать минимального значения (напомним, что требование минимизации функции $F(a_0, a_1)$ называется требованием метода наименьших квадратов (см. (11.63)), приводящим к решению системы нормальных уравнений (см. (11.64)); в данном случае система нормальных уравнений имеет вид (11.107), а ее решением являются выражения (11.108);

- $\dfrac{\partial \ln L}{\partial \sigma^2_{ELR\,Y|x}} = 0$, или, иначе,

$$-n\hat{\sigma}_{ELR\,Y|x} / (2\hat{\sigma}^3_{ELR\,Y|x}) + \sum_{i=1}^{n} (y_k - \hat{a}_0 - \hat{a}_1 x_k)^2 / 2\hat{\sigma}^4_{ELR\,Y|x} = 0,$$

откуда и следует выражение (11.109). ≪

Выборочным аналогом линейной функции регрессии (11.103) является функция

$$\bar{y}^{\text{лин}}_x = \hat{a}_0 + \hat{a}_1 x;$$

ее значение при $x = x_k$ равно

$$\bar{y}^{\text{лин}}_k = \hat{a}_0 + \hat{a}_1 x_k.$$

При каждом фиксированном значении x_k регрессора x результат наблюдения величины Y, в силу ее случайности, случаен, поэтому оценки $\hat{a}_0$, $\hat{a}_1$ и $\hat{\sigma}^2_{ELR\,Y|x}$, рассчитанные по формулам (11.108) и (11.109), могут иметь два варианта интерпретации. Это числа, если речь идет о конкретных числовых результатах наблюдения величины Y, и это случайные величины, когда речь идет о случайных, возможных результатах наблюдения величины Y. С подобными вариантами интерпретации результатов наблюдений случайной величины и рассчитанным по наблюдениям характеристикам мы сталкивались неоднократно. Обратим еще раз внимание на то, что случайность оценок $\hat{a}_0$, $\hat{a}_1$ и

$\hat{\sigma}^2_{ELR\,Y|x}$ связана со случайностью величины Y, но не x, которая является детерминированной величиной. При интерпретации оценок $\hat{a}_0$, $\hat{a}_1$ и $\hat{\sigma}^2_{ELR\,Y|x}$ как случайных величин для их обозначения не вводят специальных символов.

Можно доказать, что:

• оценки $\hat{a}_0$ и $\hat{a}_1$ являются несмещенными оценками параметров a_0 и a_1, т. е. при любом n

$$M\hat{a}_0 = a_0 \text{ и } M\hat{a}_1 = a_1, \qquad (11.110)$$

а оценка $\hat{\sigma}^2_{ELR\,Y|x}$ смещена влево относительно $\sigma^2_{ELR\,Y|x}$, т. е. $M\hat{\sigma}^2_{ELR\,Y|x} < \sigma^2_{ELR\,Y|x}$; несмещенной оценкой дисперсии $\sigma^2_{ELR\,Y|x}$ является

$$s^2_{ELR\,Y|x} = \sum_{k=1}^{n} (y_k - \hat{a}_0 - \hat{a}_1 x_k)^2/(n-2), \qquad (11.111)$$

т. е. $Ms^2_{ELR\,Y|x} = \sigma^2_{ELR\,Y|x}$ (обратим внимание на то, что в знаменателе выражения (11.111) из n вычитается число 2 — это количество параметров a_0 и a_1 линейной функции регрессии);

• каждая из оценок $\hat{a}_0$ и $\hat{a}_1$ (интерпретируемая как случайная величина) имеет нормальный закон распределения:

$$\begin{aligned} \hat{a}_0 &= N(M\hat{a}_0 = a_0;\ D\hat{a}_0 = \sigma^2_{\hat{a}_0}), \\ \hat{a}_1 &= N(M\hat{a}_1 = a_1;\ D\hat{a}_1 = \sigma^2_{\hat{a}_1}), \end{aligned} \qquad (11.112)$$

при этом несмещенные оценки дисперсий $\sigma^2_{\hat{a}_0}$ и $\sigma^2_{\hat{a}_1}$ таковы:

$$\begin{aligned} s^2_{\hat{a}_0} &= s^2_{ELR\,Y|x} \sum_{k=1}^{n} x_k^2/(n\sum_{k=1}^{n}(x_k - \bar{x})^2), \\ s^2_{\hat{a}_1} &= s^2_{ELR\,Y|x} / \sum_{k=1}^{n}(x_k - \bar{x})^2; \end{aligned} \qquad (11.113)$$

• случайная величина $(\hat{a}_j - a_j)/s_{\hat{a}_j}$, $j = 0;\ 1$, имеет распределение Стьюдента с числом степеней свободы, равным $n-2$, т. е.

$$(\hat{a}_j - a_j)/s_{\hat{a}_j} = T(n-2), \qquad j = 0;\ 1. \qquad (11.114)$$

Из соотношения (11.114) получим:

— гарантируемую с надежностью γ интервальную оценку параметра a_j

$$(\hat{a}_j - t_{n-2,\,1-\gamma}s_{\hat{a}_j};\ \hat{a}_j + t_{n-2,\,1-\gamma}s_{\hat{a}_j}), \qquad j = 0,\ 1, \qquad (11.115)$$

где число $t_{n-2,\,1-\gamma}$ находим по таблице П. 4 при $k = n - 2$ и $p = 1 - \gamma$;

— следующие три тождественных варианта проверки на уровне значимости α гипотезы H_0: $a_j = 0$, при альтернативе H_1: $a_j \neq 0$, $j = 0; 1$:

1) если число 0 не принадлежит интервалу (11.115), где $\gamma = 1 - \alpha$, то гипотезу H_0: $a_j = 0$ не принимаем; в противном случае — гипотезу H_0: $a_j = 0$ принимаем;

2) если $|\hat{a}_j / s_{\hat{a}_j}| > t_{n-2,\,\alpha}$, где число $t_{n-2,\,\alpha}$ находим по таблице П. 4 при $k = n - 2$ и $p = \alpha$, то гипотезу H_0: $a_j = 0$ не принимаем, в противном случае гипотезу H_0 принимаем;

3) если рассчитанный уровень значимости

$$P_j = 2P(T(n-2) > |\hat{a}_j / s_j|),\ j = 0; 1, \qquad (11.116)$$

меньше α, $P_j < \alpha$, то гипотезу H_0: $a_j = 0$ не принимаем; при $P_j > \alpha$ — гипотезу H_0 принимаем.

В случае отклонения гипотезы H_0: $a_j = 0$, $j = 0, 1$, говорят, что число $\hat{a}_j$ значимо отличается от нуля, или оценка $\hat{a}_j$ статистически значима; принимая гипотезу H_0: $a_j = 0$, говорят о статистической незначимости оценки $\hat{a}_j$.

Техника проведения регрессионного анализа реализована программой **«Регрессия»** пакета **«Анализ данных»** Microsoft Excel, предполагающей задание «Входного интервала Y» и «Входного интервала X» (соответственно диапазона введенных в рабочее поле значений «игрека» и «икса» — см. табл. 11.19), «Уровня надежности» (доверительной вероятности γ интервальных оценок, по умолчанию 95%; уровень значимости α принимается равным $1 - \gamma$), при этом «константу — нуль» не активизируется, поскольку в модели (11.104) константа присутствует — это параметр a_0.

▶ **ПРИМЕР 11.3** (продолжение). Проанализируем представленные на рисунке 11.9 результаты обработки программой **«Регрессия»** Microsoft Excel $n = 60$ пар чисел $(x'_k,\ y'_k)$, содержащихся в таблице 11.12:

— рассчитанные по формулам (11.108) значения оценок $\hat{a}_0$ и $\hat{a}_1$ помещены в столбец «Коэффициенты»: $\hat{a}_0 = 0{,}099$, $\hat{a}_1 = 0{,}071$ ($\bar{y}_x^{\text{лин}} = 0{,}099 + 0{,}071x$, что совпадает с (11.55));

— оценка $s_{ELRY|x} = \sqrt{s^2_{ELRY|x}}$, где $s_{ELRY|x}$, рассчитанная по формуле (11.111), помещена в «Регрессионной статистике» под именем «Стандартная ошибка»: $s_{ELRY|x} = 0{,}123$;

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный *R*	0,743234665
R-квадрат	0,552397768
Нормированный *R*-квадрат	0,544680488
Стандартная ошибка	0,122652154
Наблюдения	60

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	1,076807382	1,07680738	71,579335	1,04092E-11
Остаток	58	0,872525952	0,01504355		
Итого	59	1,949333333			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	0,099307958	0,044740424	2,219647216	0,03036673	0,009750336	0,188865581
Переменная *X* 1	0,071280277	0,008425109	8,46045719	1,04092E-11	0,054415601	0,088144953

Рис. 11.9

— оценки $s_{\hat{a}_0}$ и $s_{\hat{a}_1}$, рассчитанные с использованием формул (11.113), приведены в столбце «Стандартная ошибка»: $s_{\hat{a}_0} = 0{,}045$, $s_{\hat{a}_1} = 0{,}008$;

— границы интервальных оценок параметров a_0 и a_1, рассчитанные согласно (11.115) при заданной надежности γ (по умолчанию $\gamma = 0{,}95$), приведены в столбцах «Нижние 95%» и «Верхние 95%»: (0,01; 0,189), (0,054; 0,088);

— гипотезы H_0: $a_0 = 0$ (при альтернативе H_1: $a_0 \neq 0$) и H_0: $a_1 = 0$ (при альтернативе H_1: $a_1 \neq 0$) не принимаются на уровне значимости $\alpha = 0{,}05$, что подтверждается любым из трех вышеприведенных тождественных вариантов проверки этих гипотез:

1) число «0» не попадает в интервальную оценку ни параметра a_0, ни параметра a_1, построенную с надежностью $\gamma = 1 - \alpha = 1 - 0{,}05 = 0{,}95$;

2) модули чисел $t_j = \hat{a}_j / s_{\hat{a}_j}$, $j = 0; 1$, приведенных в столбце «t-статистика» ($t_0 = 2{,}219$, $t_1 = 8{,}460$), больше найденного по таблице П. 4 числа $t_{n-2;\alpha} = t_{58;\,0{,}05} \approx 2{,}000$ (которое не содержится в результатах работы программы);

3) рассчитанные по формуле (11.116) уровни значимости P_j, $j = 0, 1$, приведенные в столбце «P-значение» ($P_0 = 0{,}030$, $P_1 = 1{,}041 \cdot 10^{-11}$), меньше $\alpha = 0{,}05$. ◀

Смысл характеристик «R-квадрат», «Множественный R», «Нормированный R-квадрат», приведенных на рисунке 11.9, включенных в «Регрессионную статистику», и таблицы «Дисперсионный анализ» пояснен в п. 11.3.3.

Прогнозирование среднего и индивидуального значения случайной величины Y при заданном значении x_0 регрессора.

а) Найдем точечную и интервальную оценку числовой характеристики $M^{\text{лин}}(Y \mid x_0)$ — среднего значения зависимой переменной (случайной величины Y) при заданном значении x_0 независимой неслучайной переменной (регрессора x), предположив, что имеет место линейная регрессия Y на x. Точечную оценку этой числовой характеристики называют ее точечным прогнозом при $x = x_0$, а интервальную — интервальным прогнозом.

Точечная оценка числовой характеристики $M^{\text{лин}}(Y \mid x_0)$ такова:

$$\bar{y}_{x_0}^{\text{лин}} = \hat{a}_0 + \hat{a}_1 x_0. \qquad (11.117)$$

Эта оценка, как и любая точечная оценка, имеет два варианта интерпретации: либо это число $\bar{y}_{x_0}^{\text{лин}}$, либо это случайная величина и тогда будем ее обозначать $\bar{Y}_{x_0}^{\text{лин}}$. Возможность второго варианта интерпретации объясняется тем, что коэффициенты $\hat{a}_0$ и $\hat{a}_1$, вычисляемые по формулам (11.108), можно трактовать не только как числа, но (в силу случайности величины Y, следовательно, случайности и результатов Y_k, $k = 1, 2, \ldots, n$, ее наблюдений при $x = x_k$) и как случайные величины, для обозначения которых не вводят специальных символов.

Приведем следующее утверждение, касающееся закона распределения точечной оценки математического ожидания $M^{\text{лин}}(Y \mid x_0)$, интерпретируемой как случайная величина: случайная величина $\bar{Y}_{x_0}^{\text{лин}}$, определяемая соотношением (11.117), имеет нормальный закон распределения

$$\bar{Y}_{x_0}^{\text{лин}} = N(M\bar{Y}_{x_0}^{\text{лин}} = M^{\text{лин}}(Y \mid x_0); D\bar{Y}_{x_0}^{\text{лин}}). \qquad (11.118)$$

При этом несмещенная оценка дисперсии $D\bar{Y}_{x_0}^{\text{лин}}$ такова:

$$s^2_{\bar{Y}_{x_0}^{\text{лин}}} = \hat{s}^2_{ELR\ Y|x}\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{n}(x_k - \bar{x})^2}\right), \qquad (11.119)$$

где $\hat{s}^2_{ELR\,Y|x}$ рассчитывается по формуле (11.111), $\bar{x} = \sum_{k=1}^{n} x_k/n$.

Интервальная оценка математического ожидания $M^{\text{лин}}(Y \mid x)$, или, иначе, интервал, который с вероятностью, равной γ, накроет $M^{\text{лин}}(Y \mid x)$, имеет следующий вид:

$$\bar{y}^{\text{лин}}_{x_0} - t_{n-2,\,1-\gamma} s_{\bar{Y}^{\text{лин}}_{x_0}} < M^{\text{лин}}(Y \mid x_0) < \bar{y}^{\text{лин}}_{x_0} + t_{n-2,\,1-\gamma}\, s_{\bar{Y}^{\text{лин}}_{x_0}}, \qquad (11.120)$$

где $\bar{y}^{\text{лин}}_{x_0} = \hat{a}_0 + \hat{a}_1 x_0$; $t_{n-2,\,1-\gamma}$ — число, найденное по таблице П. 4 при $k = n - 2$ и $p = 1 - \gamma$, $s^2_{\bar{Y}^{\text{лин}}_{x_0}}$ рассчитывается по формуле (11.119).

Итак, точечный прогноз среднего значения величины Y при $x = x_0$ при условии, что имеет место линейная регрессия Y на x, т. е. прогноз характеристики $M^{\text{лин}}(Y \mid x)$ определяется равенством (11.117), а интервальный прогноз — неравенством (11.120).

б) Приведем формулы нахождения точечного и интервального прогноза значения зависимой переменной (случайной величины Y) при заданном значении x_0 независимой неслучайной переменной (регрессора x) при условии линейности регрессии Y на x.

Обозначим через Y_{x_0} случайную величину Y при $x = x_0$, а ее числовое значение при $x = x_0$ — через y_{x_0}. При условии линейности функции регрессии Y на x точечный прогноз значения y_{x_0} определяется следующей формулой:

$$y_{x_0} \approx \bar{y}^{\text{лин}}_{x_0} = \hat{a}_0 + \hat{a}_1 x_0, \qquad (11.121)$$

а интервальный прогноз задается интервалом

$$(\bar{y}^{\text{лин}}_{x_0} - t_{n-2,\,1-\gamma}\delta_{x_0};\ \bar{y}^{\text{лин}}_{x_0} + t_{n-2,\,1-\gamma}\delta_{x_0}), \qquad (11.122)$$

где

$$\delta^2_{x_0} = s^2_{\bar{Y}^{\text{лин}}_{x_0}} + s^2_{ELR\,Y|x} \qquad (11.123)$$

($s^2_{\bar{Y}^{\text{лин}}_{x_0}}$ рассчитывается по формуле (11.119)), число $t_{n-2,\,1-\gamma}$ находят по таблице П. 4 при $k = n - 2$ и $p = 1 - \gamma$).

▶ **ПРИМЕР 11.3** (продолжение). Напомним, что в примере по результатам обследования 60-ти фирм, сгруппированным в

корреляционную таблицу 11.10, изучалась зависимость объема продукции Y, произведенной фирмой, от стоимости X имеющихся у нее фондов. Было выяснено, что корреляционная зависимость Y от X линейная, т. е. что функция регрессии Y на x имеет вид $M(Y|x) = a_0 + a_1 x$, и найдена выборочная оценка этой функции $\bar{y}_x^{\text{лин}} = 0{,}1 + 0{,}071x$, при этом $s_{ELR\,Y|x} = 0{,}123$ (см. рис. 11.9).

а) Рассчитаем точечные и 95%-е интервальные оценки ожидаемого в среднем объема выпуска продукции при условии, что стоимость фондов фирмы зафиксирована уровнях x_i' (см. табл. 11.10). Расчеты приведены в таблице 11.20, при этом использованы следующие ранее полученные результаты, фигурирующие в формуле (11.120): $s^2_{ELR\,Y|x} = 0{,}015$, $\bar{x} = 4{,}967$, $\sum_{k=1}^{n} (x_k - \bar{x})^2 = \hat{\sigma}_X^2 \cdot n = 3{,}5322 \cdot 60 = 211{,}93$ (см. табл. 11.10); число $t_{n-2,\,1-\gamma} = t_{58;\,0{,}05} = 2{,}000$ (см. табл. П. 4).

Таблица 11.20

$x_0\,(= x_i')$ (см. табл. 11.10)	(1)	1	3	5	7	9
$\bar{y}_{x_0}^{\text{лин}} = 0{,}1 + 0{,}071x_0$	(2)	0,171	0,313	0,455	0,597	0,739
$s_{\bar{Y}_{x_0}^{\text{лин}}} = \sqrt{s^2_{\bar{Y}_{x_0}^{\text{лин}}}}$ [см. (11.119)]	(3)	0,0369	0,0229	0,0158	0,0233	0,0374
$\varepsilon = t_{n-2,\,1-\gamma} s_{\bar{Y}_{x_0}^{\text{лин}}}$	(4)	0,0738	0,0458	0,0316	0,0466	0,0748
$\bar{y}_{x_0}^{\text{лин}} - \varepsilon$ [нижняя граница интервала (11.120)]	(5)	0,097	0,267	0,423	0,550	0,664
$\bar{y}_{x_0}^{\text{лин}} + \varepsilon$ [верхняя граница интервала (11.120)]	(6)	0,245	0,359	0,487	0,643	0,814

Находящиеся в строках (5) и (6) таблицы 11.20 границы интервальных оценок изображены на рисунке 11.6: ломаные линии, проходящие через «нижние и верхние границы», симметричны относительно прямой регрессии и «удаляются» от этой прямой при увеличении $|x_i' - \bar{x}|$ (обратите внимание на то, что минимальное значение величина ε

имеет при $x_i' = 5$, которое, по сравнению с остальными x_i', меньше всего отличается от $\bar{x} = 4{,}967$).

Дадим в терминах рассматриваемого примера интерпретацию точечной и интервальной оценок характеристики $M(Y \mid (x = 5))$:

— $\bar{y}^{\text{лин}}_{x_0=5} = 0{,}455$ тыс. ден. ед. — такова оценка среднего объема продукции, произведенной фирмами, каждая из которых располагает фондами стоимостью 5 тыс. ден. ед.;

— с вероятностью, равной 0,95, можно утверждать, что интервал (0,423; 0,487) накроет $M(Y \mid (x = 5))$ — значение среднего объема продукции, произведенной фирмами, на каждой из которых стоимость фондов равна 5 тыс. ден. ед.

б) Рассчитаем границы интервалов, в которые с вероятностью 0,95 попадет значение объема продукции, произведенной фирмой, располагающей фондами стоимостью в x_i' тыс. ден. ед. Расчеты приведены в таблице 11.21 (напомним, что $s^2_{ELR\,Y|x} = 0{,}015$).

Таблица 11.21

$x_0 (= x_i')$	(1)	1	3	5	7	9
$\bar{y}^{\text{лин}}_{x_0}$	(2)	0,171	0,313	0,455	0,597	0,739
$\delta_{x_0} = \sqrt{\delta^2_{x_0}}$ [см. (11.123) и табл. 11.20]	(3)	0,1279	0,1246	0,1235	0,1247	0,1281
$\Delta = t_{n-2,\,1-\gamma}\delta_{x_0}$	(4)	0,2558	0,2492	0,2470	0,2494	0,2562
$\bar{y}^{\text{лин}}_{x_0} - \Delta$ [нижняя граница интервала (11.122)]	(5)	−0,085	0,064	0,208	0,348	0,483
$\bar{y}^{\text{лин}}_{x_0} + \Delta$ (верхняя граница интервала (11.122))	(6)	0,427	0,562	0,702	0,846	0,995

Находящиеся в строках (5) и (6) таблицы 11.21 границы интервалов изображены на рисунке 11.6; ломаные линии, проходящие через «нижние и верхние границы» (пунктирные линии), симметричны относительно прямой регрессии.

Из таблицы 11.21 делаем вывод: если, например, фирма располагает фондами стоимостью 5 тыс. ден. ед., то с вероятностью 95% можно утверждать, что объем произведенной ею продукции попадет в интервал (0,208; 0,702); при стоимости фондов в 1 тыс. ден. ед. 95%-й интервал для

объема продукции Y таков: (–0,085; 0,427), или, если учесть, что $Y > 0$, — следующий (0; 0,427). ◂

11.3.3. Множественная линейная регрессия. Функция регрессии в этом случае является линейной функцией не одного, а нескольких регрессоров ($m > 1$), т. е.

$$M(Y \mid x_1, x_2, \ldots, x_m) = M^{\text{лин}}(Y \mid x_1, x_2, \ldots, x_m),$$

где

$$M^{\text{лин}}(Y \mid x_1, x_2, \ldots, x_m) = \\ = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m. \qquad (11.124)$$

Зафиксированные в n экспериментах значения регрессоров $x_1, x_2, \ldots, x_m$ и соответствующие значения случайной величины Y задаются в форме таблицы 11.18.

Модель формирования случайного результата Y_k наблюдения величины Y при значениях регрессоров $x_1, x_2, \ldots, x_m$, равных соответственно $x_{k1}, x_{k2}, \ldots, x_{km}$, запишем в виде

$$Y_k = M^{\text{лин}}(Y \mid x_{k1}, x_{k2}, \ldots, x_{km}) + \varepsilon_k,$$

или, учитывая соотношение (11.124), в виде

$$Y_k = a_0 + a_1 x_{k1} + a_2 x_{k2} + \ldots + a_m x_{km} + \varepsilon_k, \\ k = 1, 2, \ldots, n, \qquad (11.125)$$

где ε_k, $k = 1, 2, \ldots, n$, — случайные ошибки линейной регрессии Y на $x_1, x_2, \ldots, x_m$, каждая из которых имеет нормальный закон распределения с нулевым математическим ожиданием и не зависящей от значения регрессоров $x_1, x_2, \ldots, x_m$ и номера наблюдения дисперсией (обозначим ее через $\sigma^2_{ELR\,Y|x_1, x_2, \ldots, x_m}$), т. е.

$$\varepsilon_k = N(M\varepsilon_k = 0;\ D\varepsilon_k = \sigma^2_{ELR\,Y|x_1, x_2, \ldots, x_m}), \qquad (11.126)$$

при этом

$$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \text{ — независимые величины.} \qquad (11.127)$$

Сформулированные требования к величинам ε_k тождественны следующим требованиям к случайным результатам Y_k наблюдений величины Y:

— $Y_k = N(MY_k = a_0 + a_1 x_{k1} + \ldots + a_m x_{km};$

$$DY_k = \sigma^2_{ELR\,Y|x_1, x_2, \ldots, x_m});$$

$$k = 1, 2, \ldots, n,$$

— $Y_1, Y_2, \ldots, Y_n$ — независимые величины.

В этом нетрудно убедиться, если использовать соотношение (11.125) и учесть, что $a_0 + a_1x_{k1} + \ldots + a_mx_{km}$ — постоянная при каждом наборе значений регрессоров величина.

Основная задача множественного линейного регрессионного анализа состоит в том, чтобы, используя результаты наблюдений, представленные в таблице 11.18, оценить параметры $a_0, a_1, \ldots, a_m$ и $\sigma^2_{ELR\,Y|x_1, x_2, \ldots, x_m}$ модели (11.125) и сделать заключение, приемлема эта модель или нет.

Проиллюстрируем решение этой задачи (с привлечением программы **«Регрессия»** пакета **«Анализ данных»** Microsoft Excel) на следующем примере.

ПРИМЕР 11.4. Изучим зависимость числа правонарушений Y в районе за год от двух регрессоров ($m = 2$): численности населения (x_1, тыс. чел.) района и размера среднедушевого дохода его жителей (x_2, ден. ед.) по данным таблицы 11.22.

Таблица 11.22

Номер района, k	1	2	3	4	5	6	7	8
x_1	750	367	267	500	233	700	317	600
x_2	18	20	33	18	31	18	31	20
Y	367	133	100	200	120	270	120	260

Примем следующую модель формирования количества правонарушений за год в k-м районе

$$Y_k = a_0 + a_1x_{k1} + a_2x_{k2} + \varepsilon_k,\ k = 1, 2, \ldots, 8. \quad (11.128)$$

Здесь Y_k — количество правонарушений, которое может иметь место в районе с численностью населения x_{k1}, размером среднедушевого дохода x_{k2}, и, с учетом влияния прочих неконтролируемых факторов, представленного случайной величиной ε_k. При этом случайные величины ε_k удовлетворяют требованиям (11.126) и (11.127), а именно:

$$\varepsilon_k = N(M\varepsilon_k = 0;\ D\varepsilon_k = \sigma^2_{ELR\,Y|x_1, x_2,}),\ k = 1, 2, \ldots, 8;$$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_8$ — независимые величины.

В рабочее поле Microsoft Excel введем взятые из таблицы 11.22 значения величины Y (столбец) и значения величин x_1 и x_2 (два столбца). Активизируем программу **«Рег-**

рессия» пакета «Анализ данных». В окне ввода исходных данных программы **«Регрессия»** помимо «Входного интервала Y» и «Входного интервала X». Укажем «Уровень надежности», равный 95%, активизируем «Остатки».

Результаты работы программы **«Регрессия»** приведены на рисунке 11.10.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0.970
R-квадрат	0.941
Нормированный R-квадрат	0.917
Стандартная ошибка	27.444
Наблюдения	8.000

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2.000	59799.742	29899.871	39.700	0.001
Остаток	5.000	3765.758	753.152		
Итого	7.000	63565.500			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	−102.080	107.119	−0.953	0.384	−377.438	173.278
Переменная X 1	0.524	0.095	5.513	0.003	0.280	0.768
Переменная X 2	2.274	2.823	0.805	0.457	−4.984	9.531

ВЫВОДЫ ОСТАТКА

Наблюдение	*Предсказанное Y*	*Остатки*
1	331.907	35.093
2	135.731	−2.731
3	112.880	−12.880
4	200.887	−0.887
5	90.514	29.486
6	305.703	−35.703
7	134.537	−14.537
8	257.842	2.158

Рис. 11.10 ◀|

Покажем, как находят оценки параметров $a_0, a_1, \ldots, a_m$ и $\sigma^2_{ELR\,Y|x_1, x_2, \ldots, x_m}$ модели (11.125).

Оценки $\hat{a}_0$, $\hat{a}_1$, ... , $\hat{a}_m$ параметров a_0, a_1, ... , a_m находят, исходя из требования метода наименьших квадратов, сводящегося к минимизации функции

$$F(a_0, a_1, \ldots, a_m) =$$

$$= \sum_{k=1}^{n} (y_k - a_0 - a_1 x_{k1} - \ldots - a_m x_{km})^2, \qquad (11.129)$$

в которой y_k, $k = 1, 2, \ldots, n$, — значение зависимой переменной (случайной величины Y), зафиксированное при значениях x_{k1}, x_{k2}, ... , x_{km} независимых переменных (регрессоров x_1, x_2, ... , x_m) (см. табл. 11.18).

Минимум функции (11.129) достигается при таких значениях $\hat{a}_0$, $\hat{a}_1$, ... ,$\hat{a}_m$ ее аргументов a_0, a_1, ... , a_m, которые являются решением системы m нормальных уравнений. При числе регрессоров $m = 2$ система имеет следующий вид:

$$\begin{cases} a_0 n + a_1 \sum\limits_{k=1}^{n} x_{k1} + a_2 \sum\limits_{k=1}^{n} x_{k2} = \sum\limits_{k=1}^{n} y_k, \\ a_0 \sum\limits_{k=1}^{n} x_{k1} + a_1 \sum\limits_{k=1}^{n} x_{k1} x_{k1} + a_2 \sum\limits_{k=1}^{n} x_{k2} x_{k1} = \sum\limits_{k=1}^{n} y_k x_{k1}, \\ a_0 \sum\limits_{k=1}^{n} x_{k2} + a_1 \sum\limits_{k=1}^{n} x_{k1} x_{k2} + a_2 \sum\limits_{k=1}^{n} x_{k2} x_{k2} = \sum\limits_{k=1}^{n} y_k x_{k2}. \end{cases} \qquad (11.130)$$

Напомним, что при одном регрессоре система нормальных уравнений имеет вид (11.107).

Нетрудно догадаться, как выглядит система нормальных уравнений при произвольном числе $m > 2$ регрессоров.

Выборочным аналогом линейной функции регрессии (11.124) является функция

$$\bar{y}^{\text{лин}}_{x_1, x_2, \ldots, x_m} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \ldots + \hat{a}_m x_m; \qquad (11.131)$$

ее значение при $x_1 = x_{k1}$, $x_2 = x_{k2}$, ... , $x_m = x_{km}$ равно

$$\bar{y}^{\text{лин}}_k = \hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} + \ldots + \hat{a}_m x_{km}. \qquad (11.132)$$

Можно доказать, что оценки $\hat{a}_j$, $j = 0, 1, \ldots, m$, интерпретируемые как случайные величины, — несмещенные оценки соответствующих параметров a_j, $j = 0, 1, \ldots, m$, т. е. математическое ожидание $M\hat{a}_j = a_j$, $j = 0, 1, \ldots, m$; не-

смещенная оценка дисперсии $\sigma^2_{ELR\,Y|x_1,x_2,\ldots,x_m}$ ошибки линейной регрессии

$$s^2_{ELR\,Y|x_1,x_2,\ldots,x_m} = \sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2/(n-m-1), \tag{11.133}$$

где y_k — наблюдаемые значения зависимой переменной Y при значениях x_{k1}, x_{k2}, ..., x_{km} независимых переменных $x_1, x_2, \ldots, x_m$; $\bar{y}_k^{\text{лин}}$ рассчитывается по формуле (11.132).

На рисунке 11.10 числовые значения оценок $\hat{a}_j$, $j = 0, 1, 2$, помещены в столбце «Коэффициенты», а значение оценки

$$s_{ELR\,Y|x_1,x_2,\ldots,x_m} = +\sqrt{s^2_{ELR\,Y|x_1,x_2,\ldots,x_m}} \tag{11.134}$$

имеет название «Стандартная ошибка».

▶ **ПРИМЕР 11.4** (продолжение). В рассматриваемом примере два регрессора x_1 и x_2. Система нормальных уравнений (11.130) с учетом данных, приведенных в таблице 11.22, принимает следующий вид:

$$\begin{cases} 8a_0 + 3734a_1 + 189a_2 = 1570, \\ 3734a_0 + 2\,023\,256a_1 + 80\,301a_2 = 861\,761, \\ 189a_0 + 80\,301a_1 + 4783a_2 = 33\,666. \end{cases}$$

Решение этой системы (с точностью до ошибок округления) таково: $\hat{a}_0 = -102{,}080$, $\hat{a}_1 = 0{,}524$, $\hat{a}_2 = 2{,}274$ (см. столбец «Коэффициенты» на рис. 11.10).

Выборочным аналогом функции регрессии $M^{\text{лин}}(Y \mid x_1, x_2) = a_0 + a_1x_1 + a_2x_2$ является функция

$$\bar{y}^{\text{лин}}_{x_1,x_2} = -102{,}080 + 0{,}524x_1 + 2{,}274x_2.$$

Ее значения при значениях регрессоров x_1 и x_2, взятых из таблицы 11.22, приведены в столбце «Предсказанное Y» (см. рис. 11.10). Например, для первого района $x_1 = 750$, $x_2 = 18$, поэтому

$$\bar{y}_1^{\text{лин}} = -102{,}080 + 0{,}524 \cdot 750 + 2{,}274 \cdot 18 = 331{,}907.$$

Согласно формуле (11.133), оценка дисперсии $\sigma^2_{ELR\,Y|x_1,x_2}$

$$s^2_{ELR\,Y|x_1,x_2} = \sum_{k=1}^{8} (y_k - \bar{y}_k^{\text{лин}})^2/(8-2-1) =$$
$$= [(367 - 331{,}907)^2 + (133 - 135{,}731)^2 + \ldots$$
$$\ldots + (260 - 257{,}842)^2]/5 = 753{,}173.$$

«Стандартная ошибка» $s^2_{ELRY|x_1,x_2} = \sqrt{753{,}173} = 27{,}444$ (см. рис. 11.10). ◀|

Теперь ответим на вопрос, приемлема модель (11.125) или нет (в условиях примера 11.4 ставится вопрос о приемлемости модели (11.128)), т. е. можно ли, используя эту модель, прогнозировать среднее значение зависимой переменной (случайной величины Y) при заданных значениях независимых переменных.

1. Введем следующие «суммы квадратов»: $SS_{\text{итого}} = \sum_{k=1}^{n} (y_k - \bar{y})^2$, $SS_{\text{регрессия}} = \sum_{k=1}^{n} (\bar{y}_k^{\text{лин}} - \bar{y})^2$ и $SS_{\text{остаток}} = \sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2$, где y_k — наблюдаемые значения зависимой переменной Y при значениях $x_{k1}, x_{k2}, \ldots, x_{km}$ независимых переменных $x_1, x_2, \ldots, x_m$, $\bar{y} = \sum_{k=1}^{n} y_k/n$; $\bar{y}_k^{\text{лин}}$ рассчитывается по формуле (11.132). Из равенства (11.61) следует, что

$$SS_{\text{итого}} = SS_{\text{регрессия}} + SS_{\text{остаток}}. \qquad (11.135)$$

Вычислим коэффициент

$$\text{«}R\text{-квадрат»} = \widehat{R}^2_{Y|x_1,x_2,\ldots,x_m} = \begin{cases} \dfrac{SS_{\text{регрессия}}}{SS_{\text{итого}}} = \dfrac{\sum_{k=1}^{n} (\bar{y}_k^{\text{лин}} - \bar{y})^2}{\sum_{k=1}^{n} (y_k - \bar{y})^2}, & (11.136) \\ 1 - \dfrac{SS_{\text{остаток}}}{SS_{\text{итого}}} = 1 - \dfrac{\sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2}{\sum_{k=1}^{n} (y_k - \bar{y})^2}. & (11.137) \end{cases}$$

Эквивалентность формул (11.136) и (11.137) вытекает из тождества (11.135); $\widehat{R}^2_{Y|x_1,x_2,\ldots,x_m}$ называют коэффициентом линейной детерминации наблюдаемых значений зависимой переменной (случайной величины Y) соответствующими им значениями независимых переменных (регрессоров $x_1, x_2, \ldots, x_m$), или ***выборочным коэффициентом линейной детерминации*** Y на $x_1, x_2, \ldots, x_m$. Из формулы (11.136) видно, что $\widehat{R}^2_{Y|x_1,x_2,\ldots,x_m}$ измеряет долю вариации наблюдаемых значений y_k, $k = 1, 2, \ldots, n$, зависимой переменной Y, обусловленную линейным влиянием на них соответствующих значений $x_{k1}, x_{k2}, \ldots, x_{km}$ независимых переменных — регрессоров $x_1, x_2, \ldots, x_m$.

$\hat{R}^2_{Y|x_1,x_2,\ldots,x_m}$ является выборочной оценкой генерального коэффициента линейной детерминации $R^2_{Y|x_1,x_2,\ldots,x_m}$. Всегда

$$0 \leqslant \hat{R}^2_{Y|x_1,x_2,\ldots,x_m} \leqslant 1; \qquad (11.138)$$

$\hat{R}^2_{Y|x_1,x_2,\ldots,x_m}$ тем ближе к единице, чем меньше средний квадрат ошибки, возникающей при замене фактических значений y_k, $k = 1, 2, \ldots, n$, значениями $\bar{y}_k^{\text{лин}} = \hat{a}_0 + \hat{a}_1 x_{k1} + \ldots + \hat{a}_m x_{km}$, т. е. чем меньше величина $SS_{\text{остаток}}/n = \sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2/n$. Если эта ошибка равна нулю, т. е. если $y_k = \bar{y}_k^{\text{лин}}$, то $\hat{R}^2_{Y|x_1,x_2,\ldots,x_m} = 1$. Таким образом, $\hat{R}^2_{Y|x_1,x_2,\ldots,x_m}$ — мера близости точек с координатами $(x_{k1}, x_{k2}, \ldots, x_{km}, y_k)$, $k = 1, 2, \ldots, n$, к прямой $y = \hat{a}_0 + \hat{a}_1 x_1 + \ldots + \hat{a}_m x_m$. При $\hat{R}^2_{Y|x_1,x_2,\ldots,x_m} = 0$ приближение $y_k \approx \hat{a}_0 + \hat{a}_1 x_{k1} + \ldots + \hat{a}_2 x_{km}$, $k = 1, 2, \ldots, n$ не имеет «преимущества» перед приближением $y_k \approx \bar{y}$, где $\bar{y} = \sum_{k=1}^{n} x_k/n$.

2. После нахождения выборочного коэффициента линейной детерминации $\hat{R}^2_{Y|x_1,x_2,\ldots,x_n}$ проверяют гипотезу

$$H_0\colon R^2_{Y|x_1,x_2,\ldots,x_m} = 0 \qquad (11.139)$$

о том, что генеральный коэффициент множественной детерминации равен нулю при альтернативе $H_1\colon R^2_{Y|x_1,x_2,\ldots,x_m} > 0$.

Гипотеза (11.139) эквивалентна гипотезе о том, что все параметры при регрессорах в модели (11.125) одновременно равны нулю:

$$H_0\colon a_1 = a_2 = \ldots = a_m = 0. \qquad (11.140)$$

Проверка гипотезы (11.139), или (11.140) проводится в таблице «Дисперсионный анализ» (см. рис. 11.10). В общем случае эта таблица имеет вид таблицы 11.23. Поясним эту таблицу.

≫ Общая вариация наблюдаемых «игреков» измеряется величиной

$$SS_{\text{итого}} = \sum_{k=1}^{n} (y_k - \bar{y})^2 = \hat{\sigma}_Y^2 n, \qquad (11.141)$$

число степеней свободы которой равно $n - 1$.

Таблица 11.23

Источник вариации значений величины Y	Степень свободы, (df)	Измеритель вариации значений величины Y, (SS)	Несмещенная оценка дисперсии $\hat{\sigma}^2_{ELR\ Y\mid x_1, x_2, \ldots, x_m}$, $(MS = SS/df)$	F	Значимость F (рассчитанный уровень значимости, P-значение)	$f^{кр}$, $(F_{критическое})$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Линейная регрессия Y на m регрессоров (Регрессия)	$(m+1)-1=m$	$SS_{LR\,(11.142)} \underset{(11.142)}{=} \hat{\sigma}^2_Y \hat{R}^2 n$[1]	$s^2_{LR} = SS_{LR}/m = \hat{\sigma}^2_Y \hat{R}^2 n/m$ (при выполнении гипотезы (11.139), или (11.40))	$F = \dfrac{s^2_{LR}}{s^2_{ELR}} = \dfrac{\hat{R}^2/m}{(1-\hat{R}^2)/(n-m-1)}$	$P(F(m, n-m-1) > F)$	$f_{m,\,l,\,\alpha}$, где $l = n-m-1$
Изменение значений прочих (помимо регрессоров $x_1, x_2, \ldots, x_m$) остаточных факторов (Остаток)	$n-(m+1)=n-m-1$	$SS_{ELR\,(11.143)} \underset{(11.143)}{=} \hat{\sigma}^2_Y (1-\hat{R}^2) n$	$s^2_{ELR} = SS_{ELR}/(n-m-1) = \dfrac{\hat{\sigma}^2_Y (1-\hat{R}^2) n}{n-m-1}$			
Итого	$n-1$	$SS_{итого\ (11.141)} \underset{(11.141)}{=} \hat{\sigma}^2_Y n$				

[1] Здесь и далее в таблице индекс «$Y \mid x_1, x_2, \ldots, x_m$» у $\hat{R}^2$ опущен.

Вариация «игреков», обусловленная линейным влиянием регрессоров $x_1, x_2, \ldots, x_m$, измеряется величиной

$$SS_{\text{регрессия}} = SS_{LR} = \sum_{k=1}^{n} (\bar{y}_k^{\text{лин}} - \bar{y})^2 \underset{(11.136)}{=}$$

$$= \hat{R}^2_{Y|x_1, x_2, \ldots, x_m} SS_{\text{итого}} = \hat{\sigma}_Y^2 \hat{R}^2_{Y|x_1, x_2, \ldots, x_m} n, \qquad (11.142)$$

число степеней свободы которой равно $(m + 1) - 1$ ($m + 1$ — число коэффициентов в уравнении $\bar{y}_k^{\text{лин}} = \hat{a}_0 + \hat{a}_1 x_1 + \ldots + \hat{a}_m x_m$, 1 — это «вычитаемая средняя $\bar{y}$» в выражении $\sum_{k=1}^{n} (\bar{y}_k^{\text{лин}} - \bar{y})^2$).

Вариация «игреков», обусловленная влиянием прочих (помимо регрессоров $x_1, x_2, \ldots, x_m$) остаточных факторов, измеряется величиной

$$SS_{\text{остаток}} = SS_{ELR} = \sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2 \underset{(11.137)}{=}$$

$$\underset{(11.137)}{=} (1 - \hat{R}^2_{Y|x_1, x_2, \ldots, x_m}) SS_{\text{итого}} = \hat{\sigma}_Y^2 (1 - \hat{R}^2_{Y|x_1, x_2, \ldots, x_m}) n, \qquad (11.143)$$

число степеней свободы которой равно $n - m - 1$ (n — число «игреков» в выражении $\sum_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2$, $m + 1$ — число коэффициентов в уравнении $\bar{y}_k^{\text{лин}} = \hat{a}_0 + \hat{a}_1 x_1 + \ldots + \hat{a}_m x_m$). ≪

Проверка гипотезы H_0: $R^2_{Y|x_1, x_2, \ldots, x_m} = 0$, или гипотезы H_0: $a_1 = a_2 = \ldots = a_m = 0$ основана на приведенной в таблице 11.23 (столбец 5) критической статистики

$$F = \frac{\hat{R}^2 / m}{(1 - \hat{R}^2)/(n - m - 1)}, \qquad (11.144)$$

которая при выполнении гипотезы H_0 имеет F-распределение с числами степеней свободы m и $(n - m - 1)$. Поэтому, если значение этой статистики «число F» $< f_{m, n-m-1, \alpha}$, где $f_{m, n-m-1, \alpha}$ — критическая точка, найденная по таблице П. 5 при $k_1 = m$, $k_2 = n - m - 1$ и $p = \alpha$, или если указанная в столбце 6 таблицы 11.23 «Значимость F» $> \alpha$, то гипотезу H_0: $R^2_{Y|x_1, x_2, \ldots, x_m} = 0$ (или H_0: $a_1 = a_2 = \ldots = a_m = 0$) не отклоняют. В этом случае говорят, что выборочный коэффициент $\hat{R}^2_{Y|x_1, x_2, \ldots, x_m}$ линейной детерминации *статистически незначим*. Неотклонение гипотезы H_0 — это довод в пользу отсутствия влияния (по крайней мере линейного) совокупности регрессоров $x_1, x_2, \ldots, x_m$ на ожидаемое

в среднем значение величины Y и бессмысленности дальнейшего изучения регрессионной модели (11.125) по результатам наблюдений.

Если же число $F > f_{m,\,n-m-1,\,\alpha}$ или «Значимость F» $< \alpha$, то гипотезу H_0: $R^2_{Y|x_1,x_2,\ldots,x_m} = 0$ (или H_0: $a_1 = a_2 = \ldots = a_m = 0$) отклоняют, принимают гипотезу H_1: $R^2_{Y|x_1,x_2,\ldots,x_m} > 0$ и приступают к проверке гипотез

$$H_0^{(j)}: a_j = 0,\ j = 1, 2, \ldots, m, \qquad (11.145)$$

о равенстве нулю отдельно каждого параметра a_j при альтернативе

$$H_1^{(j)}: a_j \neq 0,\ j = 1, 2, \ldots, m.$$

Как и в случае парной линейной регрессии, существуют три тождественных варианта проверки гипотезы $H_0^{(j)}: a_j = 0$ на уровне значимости α при альтернативе $H_1^{(j)}: a_j \neq 0$.

— Находят гарантируемую с надежностью $\gamma = 1 - \alpha$ интервальную оценку параметра a_j. Эта оценка имеет вид

$$(\hat{a}_j - t_{n-m-1,\,\alpha} s_{\hat{a}_j};\ \hat{a}_j + t_{n-m-1,\,\alpha} s_{\hat{a}_j}), \qquad (11.146)$$

где $\hat{a}_j$ — оценка параметра a_j (см. рис. 11.10, столбец «Коэффициенты»); $t_{n-m-1,\,\alpha}$ — число, найденное в П. 4 при $k = n - m - 1$ и $p = \alpha$; $s_{\hat{a}_j}$ — оценка среднего квадратического отклонения $\sigma_{\hat{a}_j}$ оценки $\hat{a}_j$, интерпретируемой как случайная величина (значение оценки $s_{\hat{a}_j}$ приведено на рис. 11.10 в столбце «Стандартная ошибка»). Нижняя и верхняя границы интервала (11.146) при $\gamma = 1 - \alpha$ приведены соответственно в столбцах «Нижние γ100%» и «Верхние γ100%».

Если число 0 не принадлежит интервалу (11.146), то гипотезу $H_0^{(j)}$: $a_j = 0$ не принимают; принимают гипотезу $H_1^{(j)}: a_j \neq 0$. Если число 0 принадлежит интервалу (11.146), то гипотезу $H_0^{(j)}: a_j = 0$ принимают.

— Сравнивают значение «t-статистики», равное $\hat{a}_j / s_{\hat{a}_j}$, где $\hat{a}_j$ и $s_{\hat{a}_j}$ приведены соответственно в столбцах «Коэффициенты» и «Стандартная ошибка» (см. рис. 11.10), с чис-

лом $t_{n-m-1,\alpha}$, найденным по таблице П. 4 при $k = n - m - 1$ и $p = \alpha$. Если $|\hat{a}_j / s_{\hat{a}_j}| > t_{n-m-1,\alpha}$, то гипотезу $H_0^{(j)}: a_j = 0$ не принимают, в противном случае гипотезу $H_0^{(j)}$ принимают.

— Сравнивают приведенный в столбце «*P*-значение» (см. рис. 11.10), рассчитанный уровень значимости $P_j = 2P\,(T(n-m-1) > |\hat{a}_j / s_{\hat{a}_j}|)$, где $T(n-m-1)$ — случайная величина Стьюдента с числом степеней свободы $k = n - m - 1$, с заданным уровнем значимости α. Если $P_j < \alpha$, то гипотезу $H_0^{(j)}: a_j = 0$ не принимают; при $P_j > \alpha$ гипотезу $H_0^{(j)}$ принимают.

В случае отклонения гипотезы $H_0^{(j)}: a_j = 0$ говорят, что число $\hat{a}_j$ значимо отличается от нуля или оценка $\hat{a}_j$ статистически значима; принимая гипотезу $H_0^{(j)}$, говорят о статистической незначимости оценки $\hat{a}_j$. Если в результате проверок гипотез H_0: $a_j = 0$, $j = 1, 2, \ldots, m$, какие-то из них будут приняты, то это довод в пользу отсутствия влияния (по крайней мере линейного) соответствующих регрессоров на ожидаемое в среднем значение величины Y. Такие регрессоры следует удалить из регрессионной модели. Удаление производится согласно *алгоритму многошагового регрессионного анализа с удалением*:

а) из незначимых оценок $\hat{a}_j$ при регрессорах — переменных (оценок, для которых помещенные в столбце «*P*-значение» значения $P_j > \alpha$) находят самую незначимую (с наибольшим P_j);

б) соответствующий ей регрессор удаляют из регрессионной модели (11.125) и, используя программу «**Регрессия**», проводят расчеты с числом регрессоров на единицу меньшим;

в) анализируют вновь полученную выборочную регрессию, выясняют, не ухудшилось ли ее «качество» по сравнению с прежней регрессией. Если «качество» ухудшилось, то возвращаются к прежней регрессии и удаляют из регрессионной модели регрессор, соответствующий «следующей по незначимости» оценке $\hat{a}_j$, если таковая имеется. Если «качество» вновь полученной выборочной регрессии не ухудшилось по сравнению с «качеством» выборочной регрессии, рассчитанной на предыдущем шаге, к ней применяют изложенный алгоритм, начиная с п. а).

В результатах работы программы **«Регрессия»** содержатся три показателя «качества» регрессии.

1) «R-квадрат» [см. формулы (11.136), или (11.137)], измеряющий долю вариации наблюдаемых «игреков», обусловленную линейным влиянием на них соответствующих значений регрессоров x_1, x_2, ... , x_m. Чем ближе «R-квадрат» к единице, тем больше оснований считать приемлемой линейную модель (11.125).

Однако «R-квадрат» нельзя использовать для сравнения «качества» выборочных регрессий в алгоритме многошагового регрессионного анализа, поскольку при удалении регрессора из предыдущей модели «R-квадрат» всегда уменьшается (вне зависимости, значим или незначим коэффициент при этом регрессоре).

2) «Нормированный R-квадрат» =

$$= 1 - \frac{SS_{\text{остаток}}/(n-m-1)}{SS_{\text{итого}}/(n-1)} = 1 - \frac{s^2_{ELRY|x_1, x_2, \ldots, x_m}}{s^2_Y}, \qquad (11.147)$$

где $SS_{\text{итого}} = \sum\limits_{k=1}^{n} (y_k - \bar{y})^2$, $SS_{\text{остаток}} = \sum\limits_{k=1}^{n} (y_k - \bar{y}_k^{\text{лин}})^2$, $\bar{y}_k^{\text{лин}} = \hat{a}_0 + \hat{a}_1 x_{k1} + \ldots + \hat{a}_m x_{km}$ (значения величин $SS_{\text{итого}}$ и $SS_{\text{остаток}}$ приведены в таблице «Дисперсионный анализ» на рис. 11.10); $s^2_{ELRY|x_1, x_2, \ldots, x_m}$ вычисляют по формуле (11.133) (значение величины $s_{ELRY|x_1, x_2, \ldots, x_m}$, называемой «Стандартная ошибка», приведено в «Регрессионной статистике» на рис. 11.10).

Выразим «Нормированный R-квадрат» через «R-квадрат»:

$$\text{«Нормированный } R\text{-квадрат»} = 1 - \frac{SS_{\text{остаток}}(n-1)}{SS_{\text{итого}}(n-m-1)} \overset{(11.137)}{=}$$

$$\overset{(11.137)}{=} 1 - (1 - \text{«}R\text{-квадрат»})\frac{n-1}{n-m-1}. \qquad (11.148)$$

И «R-квадрат», и «Нормированный R-квадрат» — две выборочные оценки генерального «R-квадрата», но смещение второй оценки (интерпретируемой как случайная величина) относительно генерального «R-квадрата» меньше смещения первой оценки. Поэтому со статистической точки зрения более обоснованным для оценивания проявления линейности во влиянии на Y регрессоров x_1, x_2, ... , x_m является использование «Нормированного R-квадрата».

Этот показатель используют и при сравнении качества двух последовательных регрессий в алгоритме много-

шагового регрессионного анализа, тем более, что предсказать, как изменяются его значения при удалении регрессора (в отличие от «R-квадрата», который всегда уменьшается), нельзя, так как удаление регрессора приводит в выражении (11.148) к увеличению (1 – «R-квадрат») и вместе с тем (в связи с уменьшением m) к уменьшению дроби $(n-1)/(n-m-1)$. Чем ближе «Нормированный R-квадрат» к единице, тем больше оснований считать приемлемой линейную регрессионную модель.

3) «Стандартная ошибка», $s_{ELR\,Y|x_1,x_2,\ldots,x_m} = \sqrt{s^2_{ELR\,Y|x_1,x_2,\ldots,x_m}}$ [см. формулу (11.133)].

Из формулы (11.147) видно, что при неизменяющемся значении s_Y^2 с увеличением «Нормированного R-квадрата» «Стандартная ошибка» уменьшается. Поэтому уменьшение «Стандартной ошибки», так же как и увеличение «Нормированного R-квадрата», свидетельствует об улучшении «качества» регрессии, полученной после очередного удаления регрессора.

Наконец, есть еще один вспомогательный показатель «качества» выборочной регрессии, не связанный с изложенной теорией регрессионного анализа, но широко используемый на практике, — это средняя относительная погрешность выборочной регрессии

$$\bar{\varepsilon} = \sum_{k=1}^{n} \frac{|y_k - \bar{y}_k^{\text{лин}}|}{y_k} / n, \qquad (11.149)$$

которая не должна превышать 0,05 ÷ 0,1. В формуле (11.149) y_k — наблюдаемые значения зависимой переменной Y при значениях x_{k1}, x_{k2}, ... , x_{km} регрессоров x_1, x_2, ... , x_m; $\bar{y}_k^{\text{лин}} = \hat{a}_0 + \hat{a}_1 x_{k1} + \ldots + \hat{a}_m x_{km}$.

ПРИМЕР 11.4 (продолжение). В примере по данным таблицы 11.22 изучается зависимость числа правонарушений (Y) в районе за год от численности населения (x_1, тыс. чел.) района и размера среднедушевого дохода его жителей (x_2 ден. ед.). Предполагается, что модель формирования количества правонарушений в k-м районе имеет вид (11.128). С использованием программы **«Регрессия»** построен выборочный аналог функции регрессии $M^{\text{лин}}(Y \mid x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2$, имеющий вид

$$\bar{y}^{\text{лин}}_{x_1,x_2} = -102{,}80 + 0{,}524 x_1 + 2{,}274 x_2 \qquad (11.150)$$

(см. рис. 11.10 столбец «Коэффициенты»).

Выясним, приемлема модель (11.128) или нет, используя результаты, приведенные на рисунке 11.10.

1. Выборочный коэффициент линейной детерминации Y на x_1 и x_2, согласно формуле (11.137),

$$«R\text{-квадрат}» = \hat{R}^2_{Y|x_1,x_2} = 0{,}941$$

(«R-квадрат» $= SS_{\text{регрессия}}/SS_{\text{итого}} =$
$= 59\,799{,}742/63\,565{,}500 = 0{,}941$).

2. Проверим гипотезу H_0: $R^2_{Y|x_1,x_2} = 0$ о том, что генеральный коэффициент линейной детерминации равен нулю, или эквивалентную ей гипотезу H_0: $a_1 = a_2 = 0$ о том, что параметры a_1 и a_2 одновременно равны нулю, двумя способами, приняв $\alpha = 0{,}05$:

— значение критической статистики (11.144)

$$F = \frac{0{,}941/2}{(1 - 0{,}941)/(8 - 2 - 1)} = 39{,}700,$$

приведенное в таблице «Дисперсионный анализ», больше критической точки $f_{2;\,8-2-1;\,0{,}05} = 5{,}79$, найденной по таблице П. 5 при $k_1 = 2$, $k_2 = 5$ и $p = 0{,}05$. Поэтому гипотезу H_0 не принимаем;

— «Значимость F» $= 0{,}001$, что меньше $\alpha = 0{,}05$, гипотезу H_0 не принимаем.

Таким образом, выборочный коэффициент $\hat{R}^2_{Y|x_1,x_2} =$ $= 0{,}941$ статистически значим (существенно отличается от нуля). Его содержание в терминах примера таково: «судя по наблюдениям в восьми районах доля вариации количества преступлений, обусловленная линейным влиянием на него числа жителей района и размера среднедушевого дохода, оценивается в 94% ».

3. Проверим гипотезы $H_0^{(1)}$: $a_1 = 0$ и $H_0^{(2)}$: $a_2 = 0$ при альтернативах $H_1^{(1)}$: $a_1 \neq 0$ и $H_1^{(2)}$: $a_2 \neq 0$, приняв $\alpha = 0{,}05$, тремя способами:

— гарантируемые с надежностью $\gamma = 1 - \alpha = 0{,}95$ интервальные оценки параметров a_1 и a_2 соответственно таковы: (0,280; 0,768) и (–4,984; 9,531) (см. рис. 11.10). Так как $0 \notin (0{,}280;\ 0{,}768)$, то гипотезу $H_0^{(1)}$ не принимаем, принимаем гипотезу $H_1^{(1)}$: $a_1 \neq 0$; так как $0 \in (-4{,}984;\ 9{,}531)$, то гипотезу $H_0^{(2)}$: $a_2 = 0$ принимаем;

— сравниваем значения «t-статистик», равных $t_1 = \hat{a}_1 / s_{\hat{a}_1} = 0{,}524/0{,}095 = 5{,}513$ и $t_2 = \hat{a}_2 / s_{\hat{a}_2} = 2{,}274/2{,}823 = 0{,}805$ с числом $t_{n-m-1,\alpha} = t_{8-2-1,0{,}05} = 2{,}571$, найденным по таблице П. 4 при $k = 5$ и $p = 0{,}05$. Так как $|t_1| > 2{,}571$, а $|t_2| < 2{,}571$, то принимаем гипотезы $H_1^{(1)}$: $a_1 \neq 0$ и $H_0^{(2)}$: $a_2 = 0$;

— сравним «P-значения» с $\alpha = 0{,}05$. Так как $P_1 = 0{,}003 < 0{,}05$, а $P_2 = 0{,}457 > 0{,}05$, то принимаем гипотезы $H_1^{(1)}$ и $H_0^{(2)}$.

Таким образом, оценка $\hat{a}_1 = 0{,}524$ статистически значима (гипотеза $H_0^{(1)}$: $a_1 = 0$ отклонена), а оценка $\hat{a}_2 = 2{,}274$ незначима (гипотеза $H_0^{(2)}$: $a_2 = 0$ принята).

Согласно алгоритму многошагового регрессионного анализа, из модели (11.128) удалим регрессор x_2 и проведем расчеты для модели

$$Y_k = b_0 + b_1 x_{k1} + \varepsilon_k, \quad k = 1, 2, \ldots, 8. \qquad (11.151)$$

Результаты работы программы **«Регрессия»** (в рабочий лист были введены значения регрессора x_1 и величины Y, приведенные в табл. 11.22) даны на рисунке 11.11. В этом случае:

— выборочный аналог функции регрессии $M^{\text{лин}}(Y \mid x_1) = b_0 + b_1 x_1$ имеет вид

$$\bar{y}_{x_1}^{\text{лин}} = -18{,}412 + 0{,}460 x_1; \qquad (11.152)$$

— выборочный коэффициент линейной детерминации Y на x_1

$$\hat{R}^2_{Y|x_1} = 0{,}933;$$

— гипотеза H_0: $R^2_{Y|x_1} = 0$ о равенстве нулю генерального коэффициента линейной детерминации, или равносильная ей гипотеза H_0: $b_1 = 0$ отклоняется на уровне значимости $\alpha = 0{,}05$, поскольку «Значимость F» $= 0{,}000$, так же как и совпадающее с ней (только для модели с одним регрессором) «P_1-значение» $= 0{,}000$, меньше $\alpha = 0{,}05$.

Сравним «качество» уравнения (11.150), включающего два регрессора x_1 и x_2, и уравнения (11.152) с одним регрессором x_1.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный *R*	0,966
R-квадрат	0,933
Нормированный *R*-квадрат	0,922
Стандартная ошибка	26,628
Наблюдения	8,000

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1,000	59311,349	59311,349	83,652	0,000
Остаток	6,000	4254,151	709,025		
Итого	7,000	63565,500			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	−18,412	25,288	−0,728	0,494	−80.289	43,465
Переменная *X* 1	0,460	0,050	9,146	0,000	0,337	0,583

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное Y*	*Остатки*
1	326,519	40,481
2	150,374	−17,374
3	104,383	−4,383
4	211,542	−11,542
5	88,746	31,254
6	303,524	−33,524
7	127,379	−7,379
8	257,533	2,467

Рис. 11.11

«*R*-квадрат», как и следовало ожидать, при переходе от уравнения (11.151) к уравнению (11.152) уменьшился, $\hat{R}^2_{Y|x_1} = 0{,}933 < \hat{R}^2_{Y|x_1,x_2} = 0{,}941$ [ранее отмечалось, что при удалении регрессора из предыдущей модели «*R*-квадрат» всегда уменьшается; именно поэтому «*R*-квадрат» нельзя использовать для сравнения «качества» уравнения (11.151) и (11.152)].

«Нормированный *R*-квадрат» для уравнения (11.150) равен 0,917, а для уравнения (11.152) он равен 0,922 (соответственно «Стандартная ошибка» для первого уравнения больше, чем для второго, 27,444 > 26,628). Хотя различие

значений «Нормированного R-квадрата» невелико, отдадим предпочтение модели (11.151) и уравнению (11.152) с одним регрессором. Наконец, близость и «R-квадрата», и «Нормированного R-квадрата» к единице является основанием считать приемлемой линейную регрессионную модель (11.151).

Рассчитаем значение вспомогательного показателя «качества» уравнения (11.152), а именно значение средней относительной погрешности (11.149). В условиях примера значения $\bar{y}_k^{\text{лин}}$, рассчитанные по выборочной регрессии (11.152) при значениях регрессора x_1, приведенных в таблице 11.22, даны на рисунке 11.11 в таблице «Вывод остатка» в столбце «Предсказанное Y», а разности $(y_k - \bar{y}_k^{\text{лин}})$ — в столбце «Остатки». Используя эти «Остатки» и фактические значения y_k, приведенные в таблице 11.22, вычислим среднюю относительную погрешность $\bar{\varepsilon} = 0{,}099$; погрешность, меньшая 10%, не считается большой (для регрессии (11.150) относительная погрешность $\bar{\varepsilon} = 0{,}094$).

Проинтерпретируем в терминах примера некоторые числовые результаты, полученные при расчете регрессии (11.150) и приведенные на рисунке 11.11:

— значимая оценка $\hat{b}_1 = 0{,}46$ при регрессоре x_1 (численности населения района, тыс. чел.) показывает, что при росте численности на одну тысячу можно ожидать увеличения числа преступлений в среднем (по всем районам с одинаковой численностью) на 0,46 ед.; при росте численности на две тысячи можно ожидать увеличения числа преступлений в среднем на 0,92 ед.;

— гарантируемая с надежностью 95% интервальная оценка (0,337; 0,583) параметра b_1 показывает, что при росте численности района на одну тысячу верхний предел увеличения числа преступлений, гарантируемый с надежностью 95%, в среднем составит 0,583 ед., а нижний 0,337 ед.;

— пусть численность населения района составляет 300 тыс. чел. Найдем точечную и интервальную оценки числовой характеристики $M^{\text{лин}}(Y \mid x_1 = 300)$ — среднего числа преступлений в районах с численностью $x_0 = 300$ тыс. чел., предположив, что имеет место линейная регрессия Y на x_1.

Из уравнения (11.152) получим $\bar{y}_{300}^{\text{лин}} = -18{,}412 + 0{,}460 \cdot 300 = 119{,}59$ — такова точечная оценка среднего числа преступлений в районах с численностью 300 тыс. чел. Приняв надежность γ интервальной оценки, равной

0,95, построим эту оценку, используя выражение (11.120). На рисунке 11.11 найдем $s^2_{ELR\,Y|x_1}$ = «Стандартная ошибка»^{2} = 26,628^2 = 709,050. Используя значения регрессора x_1, приведенные в таблице 11.22, и **Статистические функции** Microsoft Excel СРЗНАЧ и КВАДРОТКЛ, найдем $\bar{x}_1 =$ $= 466{,}75$ и $\sum_{k=1}^{8} (x_{k1} - \bar{x}_1)^2 = 280\,411{,}5$, а затем по формуле (11.119) рассчитаем

$$s^2_{\bar{Y}^{\text{лин}}_{300}} = 709{,}050\left(\frac{1}{8} + \frac{(466{,}75 - 300)^2}{280411{,}5}\right)^2 = 158{,}94.$$

Тогда, согласно (11.120),

$$\bar{y}^{\text{лин}}_{300} - t_{8-2;\,1-0{,}95}\sqrt{158{,}94} < M^{\text{лин}}(Y \mid 300) <$$

$$< \bar{y}^{\text{лин}}_{300} + t_{6;\,0{,}05}\sqrt{158{,}94},$$

где число $t_{6;\,0{,}05} = 2{,}447$ (см. табл. П. 4), или

$$119{,}59 - 2{,}447 \cdot 12{,}61 < M^{\text{лин}}(Y \mid 300) <$$

$$< 119{,}59 + 2{,}447 \cdot 12{,}61,$$

или

$$88{,}73 < M^{\text{лин}}(Y \mid 300) < 150{,}45. \qquad (11.153)$$

Таким образом, с вероятностью 0,95 можно ожидать, что интервал (88,73; 150,45) накроет среднее количество преступлений в районах с численностью населения 300 тыс. чел.

Наконец, используя формулу (11.122), найдем интервал, в который с вероятностью 0,95 попадет количество преступлений в районе с численностью $x_0 = 300$ тыс. чел. Согласно (11.123), $\delta^2_{300} = s^2_{\bar{Y}^{\text{лин}}_{300}} + s^2_{ELR\,Y|x_1} = 158{,}94 +$ $+ 709{,}050 = 867{,}99$; интервал (11.122) принимает вид $(\bar{y}^{\text{лин}}_{300} - t_{8-2;\,1-0{,}95}\delta_{300};\ \bar{y}^{\text{лин}}_{300} + t_{6;\,0{,}05}\delta_{300})$, или $(119{,}59 -$ $- 2{,}447\sqrt{867{,}99};\ 119{,}59 + 2{,}447\sqrt{867{,}99})$, или (47,50; 191,68). Таким образом, с вероятностью 0,95 можно утверждать, что число преступлений в районе с численностью населения 300 тыс. чел. попадет в интервал (47,50; 191,68). Этот интервал используют для прогнозирования числа преступлений в районе с численностью 300 тыс. чел.; он значительно шире интервала (11.153), используемого для прогнозирования среднего числа преступлений по районам с численностью 300 тыс. чел. ◄

11.3.4. Нелинейные модели регрессии. До сих пор мы рассматривали линейные регрессионные модели $Y = a_0 + a_1x_1 + \ldots + a_mx_m + \varepsilon$ (модели, линейные и по регрессорам $x_1, x_2, \ldots, x_m$, и по параметрам $a_1, a_2, \ldots, a_m$). Однако далеко не всегда регрессионная модель линейная. Рассмотрим только такие нелинейные модели, которые можно *линеаризовать*, т. е. с помощью подходящих преобразований зависимой и независимых переменных представить в виде линейной модели.

1. **Модели, нелинейные по регрессорам.** Примерами таких моделей являются:

— *параболическая модель*

$$Y = a_0 + a_1x + a_2x^2 + \varepsilon,$$

которая введением новых регрессоров $z_1 = x$ и $z_2 = x^2$ сводится к множественной линейной модели $Y = a_0 + a_1z_1 + a_2z_2 + \varepsilon$. Следовательно, изучение параболической модели по результатам (x_k, y_k), $k = 1, 2, \ldots, n$, наблюдений регрессора x и величины Y сводится к изучению линейной регрессии по следующим данным:

Номер наблюдения	1	2	...	n
z_1	x_1	x_2	...	x_n
z_2	x_1^2	x_2^2	...	x_n^2
Y	y_1	y_2	...	y_n

— *гиперболическая модель*

$$Y = a_0 + a_1/x + \varepsilon,$$

которая введением регрессора $z = 1/x$ сводится к линейной модели $Y = a_0 + a_1z + \varepsilon$. В этом случае исходная информация задается в таком виде:

z	$1/x_1$	$1/x_2$	...	$1/x_n$
Y	y_1	y_2	...	y_n

где (x_k, y_k), $k = 1, 2, \ldots, n$, — значения регрессора x и величины Y, зафиксированные в n наблюдениях.

2. **Модели нелинейные по параметрам и регрессорам.** Примерами таких моделей являются:

— *экспоненциальная модель*

$$Y = a_0 e^{a_1 x + \varepsilon},$$

которая после логарифмирования сводится к модели $\ln Y = \ln a_0 + a_1 x + \varepsilon$, а затем, введением новой переменной $U = \ln Y$ и обозначения $c_0 = \ln a_0$, к линейной модели $U = c_0 + a_1 x + \varepsilon$, которая изучается по следующим данным:

x	x_1	x_2	...	x_n
U	$\ln y_1$	$\ln y_2$	...	$\ln y_n$

— *степенная модель*

$$Y = a_0 x^{a_1} e^{\varepsilon},$$

которая после логарифмирования сводится к модели $\ln Y = \ln a_0 + a_1 \ln x + \varepsilon$, а затем, введением новых переменных $U = \ln Y$ и $z = \ln x$ и обозначения $c_0 = \ln a_0$, к линейной модели $U = c_0 + a_1 z + \varepsilon$, изучение которой проводится по следующим данным:

z	$\ln x_1$	$\ln x_2$	...	$\ln x_n$
U	$\ln y_1$	$\ln y_2$	...	$\ln y_n$

Любая линеаризованная модель может быть изучена с привлечением программы **«Регрессия»**, основу алгоритма которой составляет метод наименьших квадратов. В соответствии с этим методом оценки параметров линеаризованной функции регрессии (функции, остающейся в правой части линеаризованной модели, если положить $\varepsilon = 0$) находятся из требования минимизации суммы квадратов отклонений значений этой функции от соответствующих значений или преобразованных значений зависимой переменной Y. Переход от оценок параметров линеаризованной модели к оценкам параметров исходной модели не всегда прост, особенно если учесть, что полученные таким образом оценки параметров исходной модели могут и не совпадать с оценками, рассчитанными исходя из требования минимизации суммы квадратов отклонений значений исходной функции регрессии (функции, остающейся в правой части исходной модели, если положить $\varepsilon = 0$) от значений зависимой переменной Y.

Рассмотрим пример изучения гиперболической модели.

▶ **ПРИМЕР 11.5.** Изучается зависимость стоимости (Y, ден. ед.) одного экземпляра книги от тиража (x, тыс. экз.) по данным, приведенным в таблице 11.24. В этой таблице каждому «иксу» соответствует одно значение «игрека». Следует иметь в виду, что стоимость книги зависит не только от тиража, но и от ряда других неконтролируемых случайных факторов. Поэтому в генеральной совокупности зависимость стоимости книги Y от тиража x является стохастической, а не функциональной. Это означает, что в генеральной совокупности каждому фиксированному значению переменной x соответствует не одно, а множество значений переменной Y, причем сказать заранее, какое именно значение примет величина Y, нельзя.

Таблица 11.24

k	1	2	3	4	5	6	7	$n=8$
x	1	2	3	5	10	20	30	50
Y	9,10	5,30	4,11	2,83	2,11	1,62	1,41	1,30

Ориентиром в установлении вида функции регрессии $M(Y|x) = \varphi(x)$ — функции, описывающей зависимость средней стоимости экземпляра книги от тиража, служит график зависимости приведенных в таблице 11.24 значе-

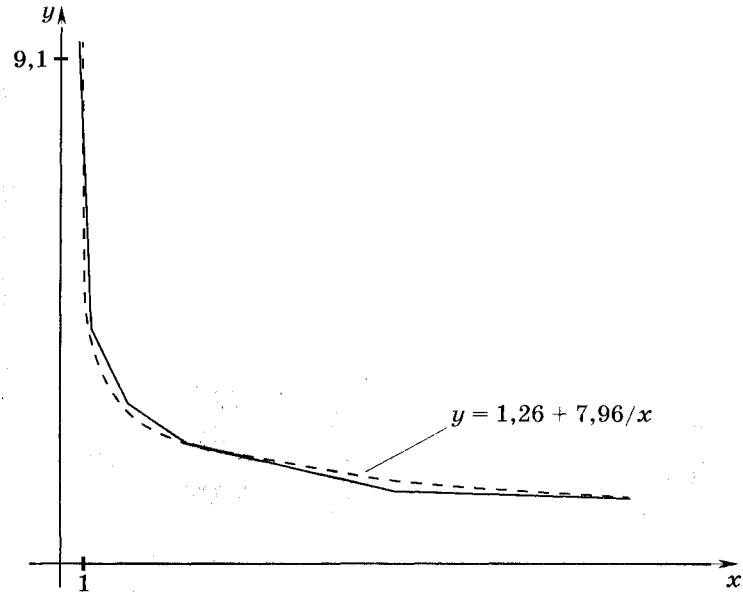


Рис. 11.12

ний y_k от x_k, который изображен на рисунке 11.12, — ломаная сплошная линия. Эта линия близка к гиперболе, что позволяет сделать допущение, что функция регрессии $\varphi(x) = a_0 + a_1/x$, а модель формирования случайного значения Y_k стоимости экземпляра книги при фиксированном тираже такова:

$$Y_k = a_0 + a_1/x_k + \varepsilon_k, \quad k = 1, 2, \ldots, 8. \qquad (11.154)$$

В результате преобразования $z = 1/x$ сведем гиперболическую модель к линейной:

$$Y_k = a_0 + a_1 z_k + \varepsilon_k, \quad k = 1, 2, \ldots, 8, \qquad (11.155)$$

и предположив, что ошибки ε_k удовлетворяют традиционно предъявляемым к ним требованиям, а именно,

$$\varepsilon_k = N(M\varepsilon_k = 0, D\varepsilon_k = \sigma^2_{ELR\,Y|z}), \quad k = 1, 2, \ldots, 8,$$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_8$ — независимые величины,

оценим параметры модели (11.155), используя в качестве исходной информации данные, содержащиеся в первых двух строках таблицы 11.25.

Таблица 11.25

$z_k = 1/x_k$	1,00	0,50	0,33(3)	0,20	0,10	0,05	0,03(3)	0,02
y_k	9,10	5,30	4,11	2,83	2,11	1,62	1,41	1,30
$\bar{y}_k^{\text{лин}} = 1{,}26 + 7{,}96 z_k$	9,21	5,23	3,90	2,84	2,04	1,65	1,51	1,41

Оценки $\hat{a}_0$, $\hat{a}_1$ параметров модели (11.155) вычислим «ручным способом» (расчеты по программе **«Регрессия»** приведены на рис. 11.13); они являются решением системы нормальных уравнений (11.107), в которой x_k заменим на z_k и которая для данных таблицы 11.25 имеет следующий вид:

$$\begin{cases} 8a_0 + 2{,}23a_1 = 27{,}78, \\ 2{,}23a_0 + 1{,}42a_1 = 14{,}05. \end{cases}$$

Решив эту систему, найдем: $\hat{a}_0 = 1{,}25$, $\hat{a}_1 = 7{,}96$. Тогда оценки $\bar{y}_k^{\text{лин}}$ значений линейной функции регрессии $M(Y \mid z_k) = a_0 + a_1 z_k$

$$\bar{y}_k^{\text{лин}} = 1{,}25 + 7{,}96 z_k, \quad k = 1, 2, \ldots, 8.$$

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный *R*	0,999124423
R-квадрат	0,998249613
Нормированный *R*-квадрат	0,997957882
Стандартная ошибка	0,120882932
Наблюдения	8

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	50,0018739	50,0018739	3421,813324	1,67702E-09
Остаток	6	0,087676099	0,014612683		
Итого	7	50,08955			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	1,247914377	0,057208659	21,81338293	6,06287E-07	1,10792973	1,387899025
Переменная X 1	7,956791484	0,136022208	58,49626761	1,67702E-09	7,623956887	8,28962608

Рис. 11.13

Эти оценки приведены в последней строке таблицы 11.25. Гипербола $y = 1{,}25 + 7{,}96/x$ изображена на рисунке 11.12 пунктирной линией.

Рассчитаем оценку $s^2_{ELR\,Y|z}$ дисперсии $\sigma^2_{ELR\,Y|z}$. Согласно (11.111),

$$s^2_{ELR\,Y|z} = \sum_{k=1}^{8} (y_k - \bar{y}_k^{\text{лин}})^2/(8-2) = 0{,}015,$$

и стандартная ошибка ошибки линейной регрессии Y на z $s_{ELR\,Y|z} = 0{,}12$.

Рассчитаем несмещенную оценку $s^2_{\hat{a}_1}$ дисперсии $\sigma^2_{\hat{a}_1}$. Согласно формуле (11.113), в которой x заменим на z,

$$s^2_{\hat{a}_1} = s^2_{ELR\,Y|z} \Big/ \sum_{k=1}^{8} (z_k - \bar{z})^2 = 0{,}015/0{,}79 = 0{,}019,$$

и стандартная ошибка оценки $\hat{a}_1$ равна $s_{\hat{a}_1} = 0{,}14$.

Далее, согласно (11.115), 95%-я интервальная оценка параметра a_1 такова:

$$(\hat{a}_1 - t_{6;\,0{,}05}s_{\hat{a}_1};\ \hat{a}_1 + t_{6;\,0{,}05}s_{\hat{a}_1}) = (7{,}62;\ 8{,}29),$$

где $\hat{a}_1 = 7{,}96$, найденное по таблице П. 4 число $t_{6;\,0{,}05} = 2{,}447$, а $s_{\hat{a}_1} = 0{,}14$.

Число 0 не попадает в этот интервал, поэтому гипотезу H_0: $a_1 = 0$ отклоняем на уровне значимости $\alpha = 1 - 0{,}95 = 0{,}05$ в пользу гипотезы H_1: $a_1 \neq 0$. Это дает основание говорить о существовании линейного влияния регрессора z (величины, обратной размеру тиража) на стоимость одного экземпляра книги. Причем это влияние, судя по выборочным данным, велико, поскольку вычисленный по формуле (11.136) «R-квадрат»

$$\widehat{R}^2_{Y|z} = \sum_{k=1}^{8} (\bar{y}_k^{\text{лин}} - \bar{y})^2 / \sum_{k=1}^{8} (y_k - \bar{y})^2 = 0{,}998,$$

т. е. практически вся вариация зафиксированных в таблице 11.25 значений стоимости экземпляра книги обусловлена линейным влиянием на стоимость величины, обратной размеру тиража.

Учитывая сказанное, дадим содержательную интерпретацию точечной и 95%-й интервальной оценки параметра a_1:

$\hat{a}_1 = 7{,}95$ — на столько ден. ед. можно ожидать увеличения в среднем стоимости одного экземпляра книги при фиксированном размере тиража, если величину, обратную размеру тиража, увеличить на единицу;

95%-й интервал (7,62; 8,29) параметра a_1 означает, что при увеличении величины, обратной размеру тиража, на единицу можно с вероятностью 0,95 ожидать, что максимальное увеличение средней стоимости одного экземпляра при фиксированном размере тиража составит 8,29 ден. ед., а минимальное увеличение составит 7,62 ден. ед.

Дадим прогноз средней стоимости экземпляра книги при тираже $x_0 = 25$ тыс. экз.

При $x_0 = 25$ $z_0 = 1/x_0 = 0{,}04$. Точечный прогноз таков: $\bar{y}_{z_0}^{\text{лин}} = 1{,}25 + 7{,}95 z_0 = 1{,}57$ (ден. ед.); 95%-й интервальный прогноз найдем, используя формулу (11.120), в которой x заменим на z:

$$\bar{y}_{z_0}^{\text{лин}} - t_{n-2,\,1-\gamma} s_{\bar{Y}_{z_0}^{\text{лин}}} < M^{\text{лин}}(Y \mid z_0) < \bar{y}_{z_0}^{\text{лин}} + t_{n-2,\,1-\gamma} s_{\bar{Y}_{z_0}^{\text{лин}}},$$

в которой $\bar{y}_{z_0}^{\text{лин}} = 1{,}57$, $\gamma = 0{,}95$, $t_{n-2,\,1-\gamma} = t_{6;\,0{,}05} = 2{,}447$, и, согласно (11.119),

$$s^2_{\bar{Y}_{z_0}^{\text{лин}}} = s^2_{ELR\,Y|z}\left(\frac{1}{8} + \frac{(\bar{z} - z_0)^2}{\sum_{k=1}^{8}(z_k - \bar{z})^2}\right) =$$

$$= 0{,}015\left(\frac{1}{8} + \frac{(0{,}28 - 0{,}04)^2}{0{,}79}\right) = 0{,}003.$$

В результате получаем

$$1{,}57 - 2{,}447 \cdot \sqrt{0{,}003} < M^{\text{лин}}(Y \mid 0{,}04) < \\ < 1{,}57 + 2{,}447 \cdot \sqrt{0{,}003},$$

или

$$1{,}44 < M^{\text{лин}}(Y \mid 0{,}04) < 1{,}70$$

т. е. с вероятностью, равной 0,95, можно утверждать, что интервал (1,44; 1,70) накроет значение средней стоимости одного экземпляра книги при тираже 25 тыс. экз.

При этом же тираже $x_0 = 25$ тыс. экз. с вероятностью, равной 0,95, можно утверждать, что случайное значение $Y_{x_0=25}$ стоимости одного экземпляра книги попадает в интервал (1,23; 1,91), рассчитанный по формуле (11.122), в которой «x» следует заменить на «z». ◀

УПРАЖНЕНИЯ

1. Двумерная случайная величина (X, Y) задана таблицей распределения вероятностей:

y \ x	0	1	3
−1	0,15	0,05	0,1
1	0,25	0,35	0,1

а) Постройте диаграмму разброса. Найдите $M(Y \mid x)$ при $x = 0, 1, 3$; выясните, существует ли корреляционная зависимость Y от X. Задайте в табличной форме функцию регрессии и постройте линию регрессии.

б) Рассчитайте коэффициент корреляции. Каков его смысл? Запишите уравнение прямой регрессии Y на x и постройте ее график. Рассчитайте дисперсию линейной регрессии Y на X и дисперсию ошибки этой регрессии и убедитесь в справедливости равенства (11.26).

в) Рассчитайте корреляционное отношение $\rho_{Y|X}$. Каков его смысл? Рассчитайте дисперсию регрессии Y на X и дисперсию ошибки этой регрессии и убедитесь в справедливости равенства (11.36).

г) Рассчитайте коэффициент линейной детерминации и коэффициент детерминации $\rho^2_{Y|X}$. Каков смысл этих коэффициентов?

2. Докажите справедливость соотношения (11.9). Используя его, выясните, изменится ли коэффициент корреляции $r_{X,Y}$ между случайными величинами X — ростом (см) бегуна и Y — скоростью его бега (м/с), если рост измерять в метрах, а скорость бега — в прежних единицах.

3. Известно, что $\sigma_X = 5$, а $\sigma_Y = 4$. Каково наибольшее и наименьшее значение ковариации $K(X, Y)$?

4. Исследуется зависимость производительности труда (Y), измеренной в тыс. ден. ед. на человека, от фондовооруженности (X), измеренной в тыс. ден. ед. на человека, по сгруппированным в корреляционную таблицу данным по 100 случайно выбранным однотипным фирмам:

j	i / x / y	1	2	3
		4—5	5—7	7—9
1	3,0—3,5	20	15	5
2	3,5—4,0	5	20	10
3	4,0—4,5		5	20

а) Выясните, существует ли корреляционная зависимость производительности труда Y от фондовооруженности X. Для этого:

— постройте поле корреляции; рассчитайте групповые средние $\bar{y}_{(i)}$ для указанных в корреляционной таблице интервалов фондовооруженности; задайте в табличной форме выборочную функцию регрессии Y на x и постройте ее график;

— рассчитайте корреляционное отношение $\hat{\rho}_{Y|X}$ и при уровне значимости $\alpha = 0{,}05$ проверьте гипотезу H_0: $\rho_{Y|X} = 0$. Поясните смысл выборочного коэффициента детерминации $\hat{\rho}^2_{Y|X}$.

б) Рассчитайте коэффициент парной корреляции между производительностью труда и фондовооруженностью, запишите уравнение выборочной линейной регрессии Y на x и постройте ее график. При $\alpha = 0{,}05$ проверьте гипотезу о линейности функции регрессии Y на x. Поясните смысл выборочного коэффициента линейной детерминации.

в) Рассчитайте точечную и 95%-ю интервальную оценку среднего значения производительности труда при фондовооруженности, равной 9 тыс. ден. ед. Рассчитайте интервал, в который с вероятностью 0,95 попадает значение производительности труда при фондовооруженности, равной 9 тыс. ден. ед. Предполагается, что функция регрессии Y на x линейная.

При выполнении заданий а), б) и в) используйте табличный процессор Microsoft Excel.

5. По результатам $n = 20$ наблюдений трехмерной случайной величины (X, Y, Z) вычислены парные коэффициенты корреляции $\hat{r}_{X,Y} = -0{,}6$, $\hat{r}_{X,Z} = 0{,}8$, $\hat{r}_{Y,Z} = -0{,}6$. Рассчитайте все частные коэффициенты корреляции и множественный коэффициент корреляции $\hat{R}_{Y|(X,Z)}$. Значимы ли парные, частные и множественный коэффициент корреляции при $\alpha = 0{,}05$?

6. Данные о динамике процента хронических болезней на тысячу жителей приведены в следующей таблице:

Год, t	0	1	2	3	4
Y (%)	10	8	5	3	4

В предположении линейной регрессии Y на t на основании этих данных:

а) Запишите регрессионную модель процента хронических болезней и рассчитайте точечные оценки параметров этой модели.

б) При $\alpha = 0{,}1$ проверьте гипотезу H_0: $a_1 = 0$.

в) С надежностью 0,9 найдите интервальную оценку параметра a_1. Дайте содержательную интерпретацию точечной и интервальной оценок параметра a_1.

г) С надежностью 0,9 установите при $t = 5$ интервальную оценку $M^{\text{лин}}(Y \mid t = 5)$ и интервал, в котором с вероятностью 0,9 будет находиться процент хронических болезней.

7. Используя программу **«Регрессия»**, исследуйте зависимость валового дохода Y торговых предприятий за год от стоимости основных (x_1, млн руб.) и оборотных (x_2, млн руб.) средств по данным 12 торговых предприятий:

x_1	118	28	17	50	56	102	116	124	114	154	115	98
x_2	105	56	54	63	28	50	54	42	36	106	88	46
Y	203	63	45	113	121	88	110	56	80	237	160	75

При необходимости используйте алгоритм многошагового регрессионного анализа с удалением.

8. Данные о зависимости темпа прироста валового национального продукта (Y, %) от темпа прироста индекса цен (x, %) по данным 10 развитых стран мира за один и тот же год приведены в следующей таблице:

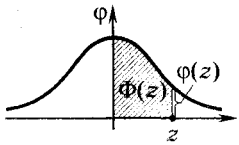
x	4,3	4,6	2,0	3,1	3,0	1,4	3,4	2,6	2,6	2,4
Y	3,5	3,1	2,2	2,7	2,7	1,6	3,1	1,8	2,3	2,3

По этим данным изучите экспоненциальную ($Y = a_0 e^{a_1 x + \varepsilon}$) и степенную ($Y = b_0 x^{b_1} e^{\varepsilon}$) модели, предварительно преобразовав их к линейному виду. Какая модель предпочтительнее: экспоненциальная или степенная?

ПРИЛОЖЕНИЯ

П. 1. Значения функции $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ и функции

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-t^2/2} dt$$



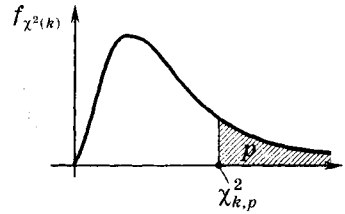
z	$\varphi(z)$	$\Phi(z)$	z	$\varphi(z)$	$\Phi(z)$	z	$\varphi(z)$	$\Phi(z)$
0,00	0,3989	0,0000	1,25	0,1826	0,3944	2,50	0,0175	0,4938
0,05	0,3984	0,0199	1,30	0,1714	0,4032	2,55	0,0154	0,4946
0,10	0,3970	0,0398	1,35	0,1604	0,4115	2,60	0,0136	0,4953
0,15	0,3945	0,0596	1,40	0,1497	0,4192	2,65	0,0119	0,4960
0,20	0,3910	0,0793	1,45	0,1394	0,4265	2,70	0,0104	0,4965
0,25	0,3867	0,0987	1,50	0,1295	0,4332	2,75	0,0091	0,4970
0,30	0,3814	0,1179	1,55	0,1200	0,4394	2,80	0,0079	0,4974
0,35	0,3752	0,1368	1,60	0,1109	0,4452	2,85	0,0069	0,4978
0,40	0,3683	0,1554	1,65	0,1023	0,4505	2,90	0,0060	0,4981
0,45	0,3605	0,1736	1,70	0,0940	0,4554	2,95	0,0051	0,4984
0,50	0,3521	0,1915	1,75	0,0863	0,4599	3,00	0,0044	0,4986
0,55	0,3429	0,2088	1,80	0,0790	0,4641	3,05	0,0038	0,4989
0,60	0,3332	0,2257	1,85	0,0721	0,4678	3,10	0,0033	0,4990
0,65	0,3230	0,2422	1,90	0,0656	0,4713	3,15	0,0028	0,4992

Окончание табл.

z	$\varphi(z)$	$\Phi(z)$	z	$\varphi(z)$	$\Phi(z)$	z	$\varphi(z)$	$\Phi(z)$
0,70	0,3123	0,2580	1,95	0,0596	0,4744	3,20	0,0024	0,4993
0,75	0,3011	0,2734	2,00	0,0540	0,4773	3,25	0,0020	0,4994
0,80	0,2897	0,2881	2,05	0,0488	0,4798	3,30	0,0017	0,4995
0,85	0,2780	0,3023	2,10	0,0440	0,4821	3,35	0,0015	0,4996
0,90	0,2661	0,3159	2,15	0,0396	0,4842	3,40	0,0012	0,4997
0,95	0,2541	0,3289	2,20	0,0355	0,4861	3,45	0,0010	0,4997
1,00	0,2420	0,3413	2,25	0,0317	0,4878	3,50	0,0009	0,4998
1,05	0,2299	0,3531	2,30	0,0283	0,4893	3,55	0,0007	0,4998
1,10	0,2179	0,3643	2,35	0,0252	0,4906	3,60	0,0006	0,4998
1,15	0,2059	0,3749	2,40	0,0224	0,4918	3,70	0,0004	0,4998
1,20	0,1942	0,3849	2,45	0,0198	0,4929	3,80	0,0003	0,4999

П. 2. Значения $\chi^2_{k,p}$, соответствующие вероятности

$$p = P(\chi^2(k) > \chi^2_{k,p})$$



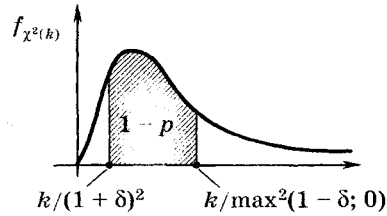
k \ p	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
1	0,000	0,000	0,001	0,004	0,02	2,71	3,84	5,02	6,64	7,88
2	0,01	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	10,60
3	0,07	0,12	0,22	0,35	0,58	6,25	7,82	9,35	11,34	12,84
4	0,21	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	10,65	12,59	14,45	16,81	18,55

Окончание табл.

k \ p	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
7	0,99	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28
8	1,44	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,96
9	1,54	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	22,59
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
11	2,31	3,05	3,82	4,58	5,58	17,28	19,68	21,92	24,72	26,76
12	3,47	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	28,30
13	3,56	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82
14	4,57	4,66	5,63	6,57	7,79	21,06	23,69	26,12	29,14	31,32
15	4,11	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	32,80
16	5,24	5,81	6,91	7,96	9,31	23,54	26,30	28,84	32,00	34,27
17	5,80	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41	35,72
18	6,56	7,02	8,23	9,39	10,86	25,99	28,87	31,53	34,81	37,16
19	6,45	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
25	10,02	11,52	13,12	14,61	16,47	34,38	37,65	40,64	44,31	46,93
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	63,17	67,51	71,42	76,15	79,49
100	67,33	70,07	74,22	77,93	82,36	118,50	124,34	129,56	135,81	140,17

П. 3. Значения $\delta_{k,p}$, определяемые уравнением

$$P\left(\frac{k}{(1+\delta_{k,p})^2} < \chi^2(k) < \frac{k}{\max^2(1-\delta_{k,p};\,0)}\right) = 1-p$$

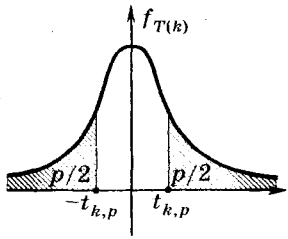


k \ p	0,1	0,05	0,02	0,01	0,001
1	6,923				
2	2,086	3,400	5,857	8,500	
3	1,270	1,932	3,000	4,200	9,00
4	0,941	1,382	2,056	2,700	5,00
5	0,738	1,104	1,594	2,000	3,80
6	0,623	0,918	1,306	1,650	3,00
7	0,576	0,800	1,143	1,393	2,50
8	0,516	0,713	0,986	1,225	2,05
9	0,476	0,650	0,889	1,094	1,75
10	0,442	0,596	0,814	0,980	1,50
12	0,388	0,527	0,700	0,840	1,30
14	0,357	0,468	0,620	0,740	1,14
16	0,325	0,422	0,564	0,671	1,02
18	0,297	0,390	0,500	0,600	0,92
20	0,282	0,370	0,480	0,567	0,85
25	0,247	0,317	0,408	0,485	0,70
30	0,226	0,281	0,369	0,425	0,60
40	0,193	0,242	0,312	0,375	0,52
50	0,174	0,212	0,270	0,311	0,45
60	0,155	0,193	0,242	0,283	0,40

Окончание табл.

k \ p	0,1	0,05	0,02	0,001	0,0001
100	0,125	0,146	0,184	0,200	0,30
1000	0,044	0,047	0,056	0,059	0,080

П. 4. Значения $t_{k,p}$, соответствующие вероятности $p = P(|T(k)| > t_{k,p})$

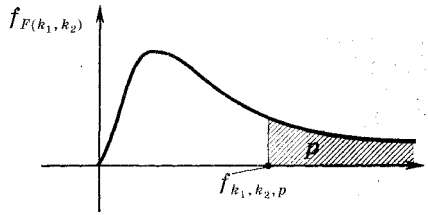


k \ p	0,1	0,05	0,01	0,005
1	6,314	12,706	63,657	127,32
2	2,920	4,303	9,925	14,089
3	2,353	3,182	5,841	7,453
4	2,132	2,776	4,604	5,598
5	2,015	2,571	4,032	4,773
6	1,943	2,447	3,707	4,317
7	1,895	2,365	3,499	4,029
8	1,860	2,306	3,355	3,833
9	1,833	2,262	3,250	3,690
10	1,812	2,228	3,169	3,581
11	1,796	2,201	3,106	3,497
12	1,782	2,179	3,055	3,428
13	1,771	2,160	3,012	3,372
14	1,761	2,145	2,977	3,326
15	1,753	2,131	2,947	3,286

Окончание табл.

k \ p	0,1	0,05	0,01	0,005
16	1,746	2,120	2,921	3,252
17	1,740	2,110	2,898	3,222
18	1,734	2,101	2,878	3,197
19	1,729	2,093	2,861	3,174
20	1,725	2,086	2,845	3,153
24	1,711	2,064	2,797	3,091
25	1,708	2,060	2,787	3,078
30	1,697	2,042	2,750	3,030
40	1,684	2,021	2,704	2,971
60	1,671	2,000	2,660	2,915
120	1,658	1,980	2,617	2,860
500	1,648	1,965	2,586	2,820
$p/2$	0,05	0,025	0,005	0,0025

П. 5. Значения $f_{k_1, k_2, p}$, соответствующие вероятности $p = P(F(k_1, k_2) > f_{k_1, k_2, p})$ при $p = 0{,}05$



k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	12
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28

Окончание табл.

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	12
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75

П. 6. Границы p_2 и p_1 интервальной оценки вероятности p при $\gamma = 0{,}95$

m \ $n-m$	1	2	3	4	5	6	7	8	9	10	...	50
0	0,975 0,000	0,842 0,000	0,708 0,000	0,602 0,000	0,522 0,000	0,459 0,000	0,410 0,000	0,369 0,000	0,336 0,000	0,308 0,000	... ...	0,071 0,000
1	0,987 0,013	0,906 0,008	0,806 0,006	0,716 0,005	0,641 0,004	0,579 0,004	0,527 0,003	0,483 0,003	0,445 0,003	0,413 0,002	... ...	0,104 0,001
2	0,992 0,094	0,932 0,068	0,853 0,053	0,777 0,043	0,710 0,037	0,651 0,032	0,600 0,028	0,556 0,025	0,518 0,023	0,484 0,021	... ...	0,132 0,005
3	0,994 0,194	0,947 0,147	0,882 0,118	0,816 0,099	0,755 0,085	0,701 0,075	0,652 0,067	0,610 0,060	0,572 0,055	0,538 0,050	... ...	0,157 0,012
4	0,995 0,284	0,957 0,223	0,901 0,184	0,843 0,157	0,788 0,137	0,738 0,122	0,692 0,109	0,651 0,099	0,614 0,091	0,581 0,084	... ...	0,179 0,021
...	...	...	...	...	...	...	...	...	...	...	...	...
100	1,000 0,946	0,998 0,931	0,994 0,917	0,989 0,904	0,984 0,892	0,979 0,881	0,973 0,870	0,967 0,859	0,962 0,849	0,955 0,838	... ...	0,741 0,585

ЛИТЕРАТУРА

1. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Основы моделирования и первичная обработка данных. — М.: Финансы и статистика, 1983.

2. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Исследование зависимостей. — М.: Финансы и статистика, 1985.

3. *Айвазян С. А., Мхитарян В. С.* Прикладная статистика. Основы эконометрики: В 2-х т. Т. 1. Теория вероятностей и прикладная статистика: Учеб. — М.: ЮНИТИ-ДАНА, 2001.

4. *Айвазян С. А., Мхитарян В. С.* Прикладная статистика в задачах и упражнениях. — М.: ЮНИТИ-ДАНА, 2001.

5. *Андреев В. Д.* Случайные события и дискретные случайные величины. — М.: МИУ, 1982.

6. *Боровков А. А.* Математическая статистика. — Новосибирск: Наука, 1997.

7. *Боровков А. А.* Теория вероятностей. — М.: Эдиториал УРСС, 1999.

8. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. — М.: Наука, 1983.

9. *Вентцель Е. С.* Теория вероятностей. — М.: Высшая школа, 1998.

10. *Гмурман В. Е.* Теория вероятностей и математическая статистика. — М.: Высшая школа, 1998.

11. *Гнеденко Б. В.* Курс теории вероятностей. — М.: Эдиториал УРСС, 2001.

12. *Калинина В. Н., Панкин В. Ф.* Математическая статистика. — М.: Дрофа, 2002.

13. *Колемаев В. А., Калинина В. Н.* Теория вероятностей и математическая статистика.—М.: ИНФРА —М, 1997, 1999, 2000.

14. *Колемаев В. А., Калинина В. Н., Соловьев В. И.* и др. Теория вероятностей в примерах и задачах. — М.: ГУУ, 2001.

15. *Колемаев В. А., Калинина В. Н., Соловьев В. И.* Математическая статистика в примерах и задачах. — М.: ГУУ, 2001.

16. *Крамер Г.* Математические методы статистики. — Ижевск: Регулярная и хаотическая динамика, 2003.

17. *Кремер Н. Ш.* Теория вероятностей и математическая статистика. — М.: ЮНИТИ, 2004.

18. *Малыхин В. И.* Финансовая математика. — М.: ЮНИТИ, 1999.

19. *Мешалкин Л. Д.* Сборник задач по теории вероятностей. — М.: Изд-во Моск. ун-та, 1963.

20. *Ниворожкина Л. И., Морозова З. А.* Математическая статистика с элементами теории вероятностей. — Ростов-на-Дону: РГЭУ, 2001.

21. Практикум по эконометрике / Под ред. Елисеевой И. И. — М.: Финансы и статистика, 2001.

22. *Пугачев В. С.* Теория вероятностей и математическая статистика. — М.: Наука, 1979.

23. *Смирнов Н. В., Дунин-Барковский И. В.* Курс теории вероятностей и математической статистики для технических приложений. — М.: Наука, 1969.

24. *Тутубалин В. Н.* Теория вероятностей. — М.: МГУ, 1972.

25. *Четыркин Е. М., Калихман И. Л.* Вероятность и статистика. — М.: Финансы и статистика, 1982.

26. *Чистяков В. П.* Курс теории вероятностей. — СПб.: Лань, 2003.

ОГЛАВЛЕНИЕ

Учебное издание

Калинина Вера Николаевна

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА. КОМПЬЮТЕРНО-ОРИЕНТИРОВАННЫЙ КУРС

Учебное пособие для вузов

Зав. редакцией *Т. Д. Гамбурцева*
Ответственный редактор *Ж. И. Яковлева*
Художественный редактор *А. В. Пряхин*
Технический редактор *И. В. Грибкова*
Компьютерная верстка *Г. А. Фетисова*
Корректор *Г. И. Мосякина*

Санитарно-эпидемиологическое заключение
№ 77.99.02.953.Д.006315.08.03 от 28.08.2003.

Подписано к печати 21.01.08. Формат 60×90 $^1/_{16}$.
Бумага типографская. Гарнитура «Школьная». Печать офсетная.
Усл. печ. л. 30,0. Тираж 3000 экз. Заказ № 7511.

ООО «Дрофа». 127018, Москва, Сущевский вал, 49.

По вопросам приобретения продукции издательства «Дрофа» обращаться по адресу:
127018, Москва, Сущевский вал, 49.
Тел.: (495) 795-05-50, 795-05-51. Факс: (495) 795-05-52.

Торговый дом «Школьник».
109172, Москва, ул. Малые Каменщики, д. 6, стр. 1А.
Тел.: (495) 911-70-24, 912-15-16, 912-45-76.

Магазины «Переплетные птицы»:
127018, Москва, ул. Октябрьская, д. 89, стр. 1.
Тел.: (495) 912-45-76;
140408, Московская обл., г. Коломна, Голутвин,
ул. Октябрьской революции, 366/2.
Тел.: (495) 741-59-76.

Интернет-магазин: http://www.drofa.ru

Отпечатано с предоставленных диапозитивов
в ОАО «Тульская типография». 300600, г. Тула, пр. Ленина, 109.

В. Н. Калинина

Теория вероятностей и математическая статистика

Компьютерно-ориентированный курс