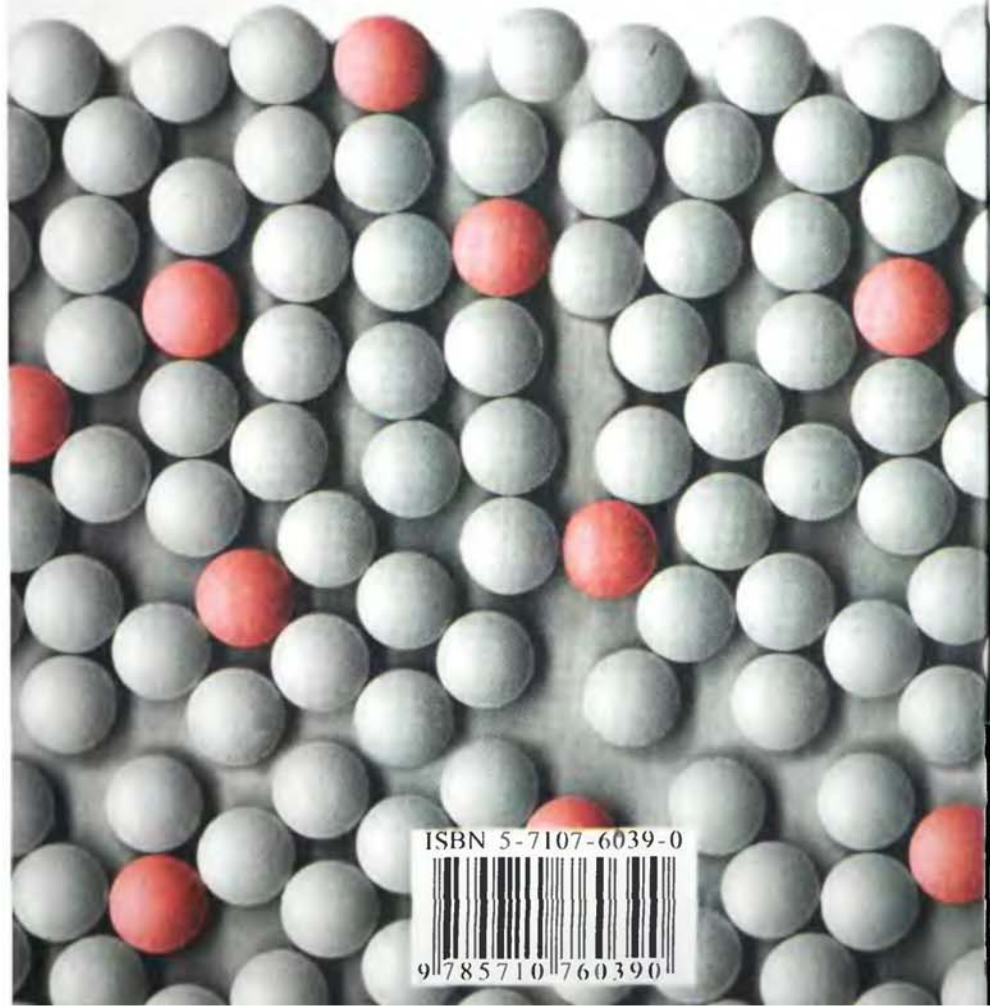
дрофа



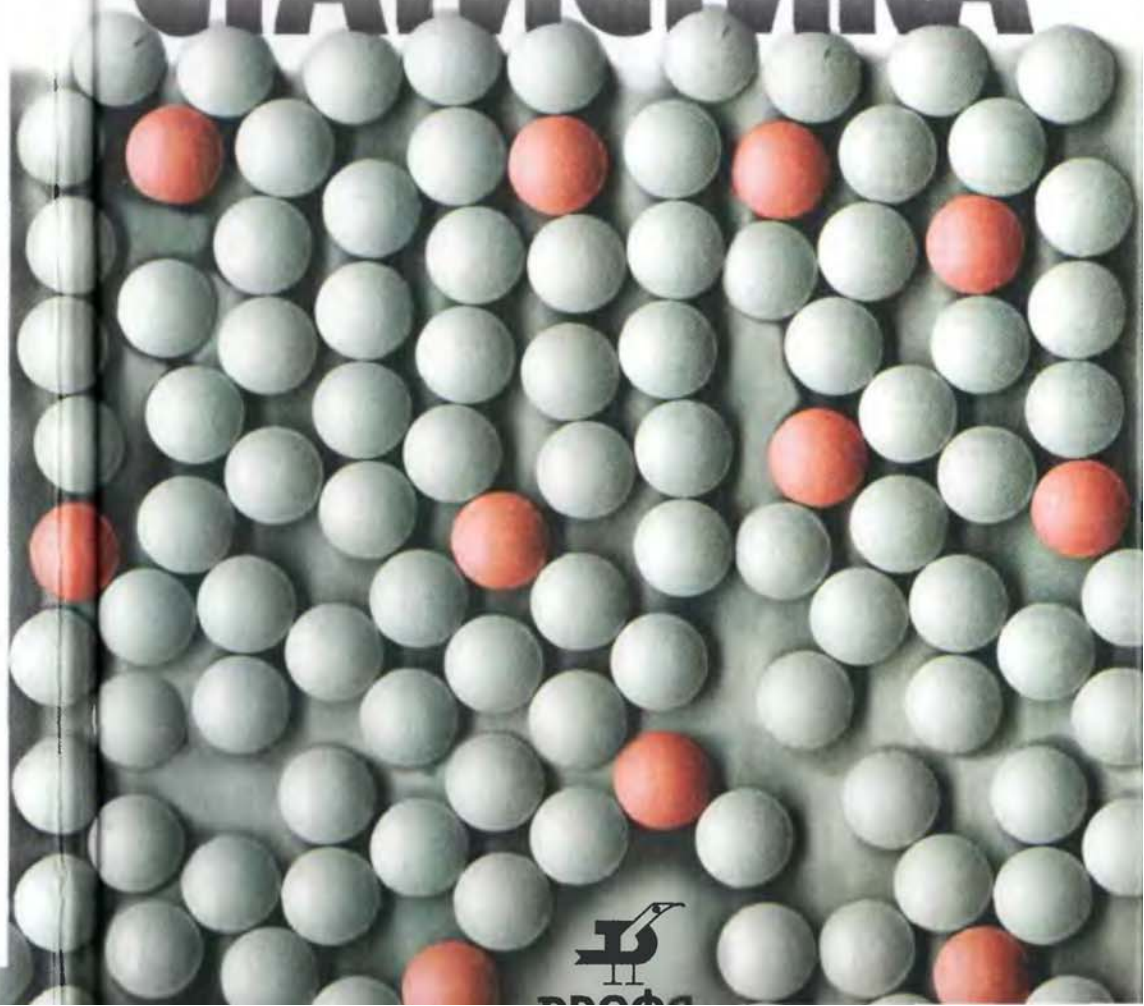
ISBN 5-7107-6039-0

В.Н. Калинина, В.Ф. Панкин МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

СРЕДНЕЕ ПРОФЕССИОНАЛЬНОЕ ОБРАЗОВАНИЕ

В.Н. Калинина, В.Ф. Панкин

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА



дрофа

СРЕДНЕЕ ПРОФЕССИОНАЛЬНОЕ ОБРАЗОВАНИЕ

В. Н. Калинина, В. Ф. Панкин

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Издание четвертое, исправленное

Допущено Министерством образования
Российской Федерации в качестве
учебника для студентов образовательных учреждений
среднего профессионального образования,
обучающихся по специальностям
информатики и вычислительной техники

дрофа

МОСКВА
2002

УДК 519.2(075.32)
ББК 22.172я722
К17

Рецензенты:
д-р физ.-мат. наук, проф. *Н. А. Холщевникова*
(зав. кафедрой математики МГТХ «Станкин»);
д-р экон. наук, проф. *М. Р. Ефимова*
(зав. кафедрой статистики ГУУ)

Калинина В. Н.

К17 Математическая статистика: Учеб. для студ. сред. спец. учеб. заведений / В. Н. Калинина, В. Ф. Панкин. — 4-е изд., испр. — М.: Дрофа, 2002. — 336 с.: ил.

ISBN 5—7107—6039—0

В учебнике (3-е изд. — 2001 г.) содержатся наиболее важные разделы математической статистики: оценивание числовых характеристик и закона распределения случайной величины, проверка гипотез, дисперсионный и корреляционно-регрессионный анализ, а также необходимые для понимания этих разделов сведения по теории вероятностей. Приведены примеры и упражнения, их разбор и решения, графические иллюстрации.

В учебник включены вопросы статистического моделирования случайных величин и систем массового обслуживания на ЭВМ, широко используемого специалистами, которые работают в области программирования и использования ЭВМ.

Для студентов средних специальных учебных заведений.

УДК 519.2(075.32)
ББК 22.172я722

ISBN 5—7107—6039—0

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Настоящий учебник написан в соответствии с программой курса «Математическая статистика», утвержденной для специальности «Программирование для быстродействующих математических машин».

Цель курса — изложить основы теории вероятностей и математической статистики, изучающей закономерности массовых случайных явлений. В книге рассматриваются важнейшие методы и приемы обработки результатов наблюдений, знание которых необходимо современному программисту при разработке алгоритмов решения практических задач. Уделяется внимание широко используемому на ЭВМ методу статистических испытаний; приводится пример моделирования системы массового обслуживания.

Данные в учебнике задачи не только разъясняют общетеоретические положения, но и наглядно иллюстрируют возможные области приложения математической статистики. Многие задачи основываются на решении конкретных инженерных, экономических и исследовательских вопросов, поэтому к ним следует отнестись с должным вниманием.

Введение, гл. 9—13 и § 6.5 написаны В. Н. Калининой, гл. 1—8 — В. Ф. Панкиным.

Авторы выражают благодарность В. Л. Тамбовцеву и Я. К. Колде за рецензирование рукописи книги и научному редактору проф. В. Н. Гудкову. Сделанные ими замечания способствовали улучшению книги. Авторы также будут признательны всем, кто пришлет свои критические замечания и пожелания, направленные на улучшение книги.

Авторы

ВВЕДЕНИЕ

Цель науки — описание, объяснение и предсказание явлений действительности на основе установленных законов, что позволяет находить решения в типичных ситуациях.

В основе научных знаний лежит наблюдение. Для обнаружения общей закономерности, которой подчиняется явление, необходимо многократно его наблюдать в одинаковых условиях. Чем это вызвано и что понимать под одинаковыми условиями?

Многие явления окружающего мира взаимно связаны и влияют одно на другое. Проследить все связи и определить влияние каждой из них на явление не всегда представляется возможным. Поэтому ограничиваются изучением влияния лишь основных факторов, определяющих течение явления. Под одинаковыми условиями наблюдений и понимается соблюдение во всех наблюдениях практически одинаковых значений основных факторов.

Рассмотрим пример. Станок, хорошо отлаженный в начале работы, со временем теряет настройку, режущий инструмент затупляется, что и приводит к ухудшению качества обработки изделий. Поставлена задача — определить момент, когда следует остановить станок и провести его подналадку или сменить инструмент. Для определения этого момента проверяют качество изготовленных деталей. Наблюдение, проведенное после двух часов работы станка, показало, что изделие не отвечает установленным требованиям. Причиной может быть качество заготовки, случайные изменения режима работы станка. По единичному замеру нельзя принимать решение об остановке станка. Нужны дополнительные замеры. Сколько должно быть проведено наблюдений? Как обработать результаты наблюдений и сделать обоснованные практические выводы? Получить ответы на эти вопросы позволяет математическая статистика.

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

Для широкого круга явлений при сохранении постоянными основных условий испытаиий отмечается иеоднозначность полученных результатов. Примером таких случайных явлений служат погрешности измерений. Измеряя один и тот же предмет, например взвешивая его на аиалитических весах много раз, получают близкие, но все же различные результаты. Это объясняется тем, что результат каждого измерения содержит случайную погрешность. Предвидеть эту погрешность, а следовательно, и результат каждого конкретного измерения нельзя. Однако если определенным образом систематизировать результаты измерений, то окажется, что в их изменении можно увидеть некоторую закономерность — статистическую устойчивость. Изучение этой закономерности позволяет, например, предвидеть в среднем результат серии измерений.

Математическая статистика — наука, изучающая методы обработки результатов наблюдений массовых случайных явлений, обладающих статистической устойчивостью, закономерностью, с целью выявления этой закономерности. Выводы о закономерностях, которым подчиняются явления, изучаемые методами математической статнстики, всегда основываются на ограниченном, выборочном числе наблюдений. При большем числе наблюдений эти выводы могут оказаться иными. Для вынесения более определенного заключения о закономерностях явления математическая статистика опирается на теорию вероятностей.

В отличие от математической статистики, имеющей дело с результатами наблюдений случайных явлений, теория вероятностей формально — логически изучает закономерности случайных явлений и имеет дело с математическими моделями случайных явлений. Обработав результаты наблюдений, исследователь выдвигает ряд гипотез, предположений о том, что рассматриваемое явление можио описать той или иной вероятностной теоретической моделью. Далее, используя математико-статистические методы, можно дать ответ на вопрос, какую из гипотез или моделей следует приият. Имеино эта модель и считается закономерностью изучаемого явления. Правомерен такой вывод или нет, покажет практика использования выбранной модели. Таков типичный путь математико-статистического исследования.

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

ЧАСТЬ I
ОСНОВНЫЕ ПОНЯТИЯ КОМБИНАТОРИКИ

ГЛАВА 1
РАЗМЕЩЕНИЯ, ПЕРЕСТАНОВКИ, СОЧЕТАНИЯ

§ 1.1. Правило умножения и сложения

Установим два важных правила, которые часто применяются при решении комбинаторных задач. Для этого рассмотрим решение следующей задачи.

Задача 1.1. Необходимо составить варианты контрольной работы, каждый из которых должен содержать три задачи. Одна задача выбирается из любого параграфа I гл. сборника задач, вторая — из любого параграфа II гл., а последняя — из любого параграфа III гл. При этом следует учесть, что I и III гл. содержат два параграфа, а II гл. — три параграфа. Сколько видов контрольной работы можно составить исходя из этих условий, если вид работы определяется только номерами параграфов, из которых выбраны задачи?

△ Задачу решим двумя способами.

Способ 1. Пусть каждой задаче соответствует двузначное число, где первая цифра соответствует номеру выбранной главы, а вторая — номеру параграфа. Чтобы не допустить ошибки при подсчете, воспользуемся специальным *графом*, который иногда называют *деревом* (рис. 1.1).

Начальную точку обозначим буквой O. Двигаясь всеми возможными путями по ребрам графа слева направо,

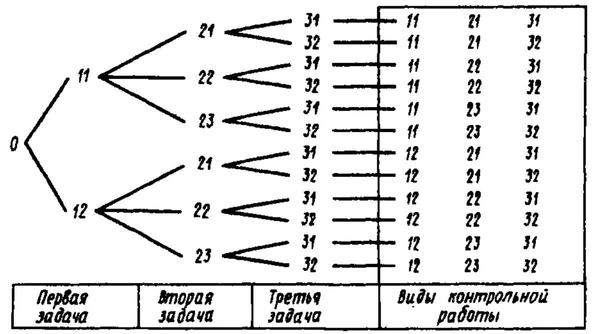


Рис. 1.1

начиная с точки O получим 12 различных видов контрольной работы.

Способ 2. В задаче требуется для каждого вида контрольной работы подобрать три параграфа по одному из указанных трех глав, т. е. следует заполнить три клетки на карточке для контрольной работы. Изобразим эти клетки следующим образом: $\boxed{}\boxed{}\boxed{}$

В первую клетку можно поместить либо 11, либо 12. Поэтому первую клетку можно заполнить двумя способами: $\boxed{2}\boxed{}\boxed{}$ На «дереве» это обстоятельство иллюстрируется двумя ветвями, исходящими из точки O и ведущими к столбцу «Первая задача». Для каждого из двух способов заполнения первой клетки имеется три варианта заполнения второй клетки, так как вторую задачу можно выбрать тремя способами, поскольку гл. II содержит три параграфа: $\boxed{2}\boxed{3}\boxed{}$

Первые две клетки можно заполнить шестью способами: $2 \cdot 3 = 6$. Заметим, что именно 6 ветвей заканчиваются в столбце «Вторая задача». Для каждого из этих шести способов существует два способа заполнения третьей клетки, так как третья задача может быть выбрана из двух параграфов гл. III: $\boxed{2}\boxed{3}\boxed{2}$

Тогда общее число способов заполнения трех клеток равно $2 \cdot 3 \cdot 2 = 12$. Именно столько ветвей заканчивается в столбце «Третья задача». Таким образом, исходя из условий задачи, можно составить 12 различных видов контрольной работы. ▲

Задача 1.2. В группе 30 человек. Необходимо выбрать старосту и профорга. Сколькими способами можно это сделать?

△ Старостой может быть выбран любой из 30 учащихся, т. е. существует 30 способов выбора старосты. После того как староста уже выбран, профоргом можно выбрать любого из оставшихся 29 учащихся. Таким образом, одному способу выбора старосты соответствуют 29 способов выбора профорга. Следовательно, общее число способов выбора старосты и профорга равно $30 \cdot 29 = 870$. ▲

Рассуждения, которые были проведены при решении предыдущих задач, подтверждают справедливость следующего простого утверждения.

Правило умножения. *Пусть требуется выполнить одно за другим какие-то k действий. Если первое действие можно выполнить n_1 способами, второе действие — n_2 способами, третье — n_3 способами и так до k-го действия, которое можно выполнить n_k способами, то все k действий вместе могут быть выполнены $n_1 n_2 n_3 \ldots n_k$ способами.*

Это правило дает удобный универсальный метод решения многих комбинаторных задач.

Задача 1.3. Четыре мальчика и четыре девочки садятся на 8 расположенных подряд стульев, причем мальчики садятся на места с четными номерами, а девочки — не места с нечетными номерами. Сколькими способами это можно сделать?

△ Первый мальчик может сесть на любое из четырех четных мест, второй — на любое из оставшихся трех мест, третий — на любое из оставшихся двух мест. Последнему мальчику предоставляется всего одна возможность. Согласно правилу умножения, мальчики могут занять четыре места $4 \cdot 3 \cdot 2 \cdot 1 = 24$ способами. Столько же возможностей имеют и девочки. Таким образом, согласно правилу умножения, мальчики и девочки могут занять все стулья $24 \cdot 24 = 576$ способами. ▲

Задача 1.4. Имеется 20 изделий 1-го сорта и 30 изделий 2-го сорта. Необходимо выбрать два изделия одного сорта. Сколько способов выбора двух изделий возможно в данной ситуации, если учитывается порядок выбора изделий?

△ Условимся первым действием считать выбор изделий 1-го, вторым — выбор изделий 2-го сорта. По правилу умножения два изделия 1-го сорта можно выбрать $20 \cdot 19 = 380$ способами. Аналогично, два изделия 2-го сорта можно выбрать $30 \cdot 29 = 870$ способами. Согласно условию задачи следует выбрать два изделия одного сорта, не важно какого. Это могут быть либо изделия 1-го сорта, либо изделия 2-го сорта. Таким образом, должно быть выполнено либо первое действие, либо второе, но не первое действие, а затем второе. Эти действия не могут быть выполнены одновременно, поскольку они взаимно исключают друг друга. Поэтому общее число способов выбора изделий одного сорта равно $380 + 870 = 1250$. ▲

Соображения, которые были приведены при решении последней задачи, позволяют сформулировать правило сложения.

Правило сложения. *Если два действия взаимно исключают друг друга, причем одно из них можно выполнить m способами, а другое — n способами, то выполнить одно любое из этих действий можно $n+m$ способами.*

Это правило легко распространить на любое конечное число действий.

§ 1.2. Размещения

Пусть имеется некоторое множество, содержащее конечное число членов. Например, множество учебных групп в техникуме, множество книг на полке, множество населенных пунктов в данной области или, например, множество целых положительных чисел, меньших 10, и т. д. Все элементы такого множества можно пронумеровать, т. е. каждому элементу множества поставить в соответствие одно из чисел: 1, 2, 3, 4, ..., n; в результате получается некоторая последовательность элементов данного множества, которые обычно записывают в виде a_1, a_2, a_3, ..., a_n. Такие «занумерованные» множества будем называть *упорядоченными*. Таким образом, упорядоченное множество представим в виде некоторой последовательности, что будет использовано в дальнейшем. Очевидно, если в упорядоченном множестве поменять местами хотя бы два его элемента, то получим новое упорядоченное множество, которому будет соответствовать новая последовательность элементов данного множества.

Определение. *Размещением из n элементов по m называется любое упорядоченное подмножество из m элементов множества, состоящего из n различных элементов.*

○ **Пример 1.1.** Пусть имеется множество, содержащее четыре буквы: $\{A; B; C; D\}$. Запишем все возможные размещения из четырех указанных букв по две. Таких размещений 12: AB; AC; AD; BC; BD; CD; BA; CA; DA; CB; DB; DC. ●

Заметим, что размещения отличаются порядком входящих в них элементов и их составом. Размещения AB и BA содержат одинаковые буквы, но порядок их расположения различен. Поэтому эти размещения считаются разными.

○ **Пример 1.2.** Следующие последовательности цифр являются размещениями по 4 элемента из 10 элементов множества: $\{0; 1; 2; 3; 4; 5; 6; 7; 8; 9\}$: 3021, 4590; 8612; 7485; 1234. ●

В последнем примере не выписаны все возможные размещения, так как таких размещений более 5000 (в этом можно убедиться решая задачу 1.6).

На практике чаще представляет интерес количество размещений, а не их конкретный вид. Число размещений из n элементов по m будем обозначать символом A_n^m, где $m \leqslant n$.

Задача 1.5. В некоторой газете 12 страниц. Необходимо на страницах этой газеты поместить четыре фотографии. Сколькими способами это можно сделать, если ни одна страница газеты не должна содержать более одной фотографии?

△ Первую фотографию можно поместить на любую из 12 страниц, т. е. 12 способами, вторую — на любую из оставшихся 11 страниц, т. е. 11 способами. Для размещения третьей фотографии имеется 10 способов, а для последней — 9 способов. По правилу умножения четыре фотографии можно разместить на 12 страницах $12 \cdot 11 \cdot 10 \cdot 9 = 11\,880$ способами. ▲

Найденное число размещений четырех фотографий на 12 страницах газеты — это число A_{12}^4. Действительно, для размещения фотографий следует отобрать 4 различных страницы газеты из 12 имеющихся. Затем необходимо отобранные страницы упорядочить, не обращая внимания на их номера, т. е. определить, на какую страницу поместить первую фотографию, на какую — вторую, и т. д.

Полученная упорядоченная совокупность страниц является согласно определению размещением из 12 элементов по 4, а число A_{12}^4 таких размещений является ответом.

Таким образом, $A_{12}^4 = 12 \cdot 11 \cdot 10 \cdot 9$. Последнее произведение можно представить в другом виде, если использовать символ факториала, для чего умножим и разделим его на 8! Имеем

$$12 \cdot 11 \cdot 10 \cdot 9 = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot (8!)}{8!} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{8!} = \frac{12!}{8!}$$

Итак, $A_{12}^4 = \frac{12!}{8!} = 12 \cdot 11 \cdot 10 \cdot 9$.

Следующая теорема дает общую формулу для вычисления размещений, которая позволяет значительно упростить решение подобных задач.

Теорема. *Число размещений из n элементов по m равно $\frac{n!}{(n-m)!}$, т. е.*

$$A_n^m = \frac{n!}{(n-m)!}. \tag{1.1}$$

□ Для доказательства воспользуемся правилом умножения. Чтобы составить какое-либо размещение, следует выбрать m элементов из множества, содержащего n элементов, и упорядочить полученную совокупность. Это означает, что надо заполнить m мест элементами рассматриваемого множества. На 1-е место можно поместить любой из n элементов. После этого останется $n-1$ элемент, каждый из которых может быть помещен на 2-е место. Поэтому 1-е место можно заполнить n способами, а 2-е — $n-1$ способами. Рассуждая аналогично, получаем, что 3-е место можно заполнить $n-2$ способами, 4-е — $n-3$ способами и т. д. Последнее m-е место можно заполнить $n-(m-1)$ способами. Все m мест по правилу умножения можно заполнить $n(n-1)(n-2)...(n-(m-1))$ способами. Таким образом,

$$A_n^m = n(n-1)(n-2)(n-3)...[n-(m-1)]. \tag{1.2}$$

Запишем выражение в правой части равенства (1.2) в более удобном виде, для чего умножим и разделим его на $(n-m)!$. Предполагая, что $m \neq n$, имеем

$$A_n^m = \frac{n(n-1)(n-2)...[n-(m-1)][(n-m)!]}{(n-m)!} =$$

$$= \frac{n(n-1)(n-2)...[n-(m-1)](n-m)(n-m-1)...3 \cdot 2 \cdot 1}{(m-n)!} = \frac{n!}{(n-m)!} \ ■.$$

Формулы (1.1) и (1.2) при $m<n$ совпадают, причем формула (1.2) имеет место для всех $m \leqslant n$. Если $n=m$, то

формула (1.1) принимает вид $A_n^m=n!/0!$, а формула (1.2) — следующий вид: $A_n^n=n!$. В дальнейшем будем считать $0!=1$, что позволяет использовать формулу (1.1) для случая $n=m$.

Задача 1.6. Сколько можно записать четырехзначных чисел, используя без повторения все десять цифр?

△ Так как в любом числе важную роль играет порядок входящих в него цифр, то для ответа на поставленный вопрос, очевидно, следует определить число размещений из 10 цифр по 4:

$$A_{10}^4=\frac{10!}{6!}=\frac{10!}{6!}=7\cdot 8\cdot 9\cdot 10=5040.$$

Однако не все последовательности из 4 цифр представляют собой четырехзначное число, поскольку среди них есть и такие, у которых на 1-м месте находится 0. Найдем число таких последовательностей. Так как у рассматриваемых последовательностей на 1-м месте уже стоит 0, то следует выбрать еще 3 цифры из оставшихся 9. Найдем число размещений из 9 по 3: $A_9^3=\frac{9!}{6!}=7\cdot 8\cdot 9=504$. Таким образом, искомое число четырехзначных чисел равно разности $A_{10}^4-A_9^3=5040-504=$ $=4536$. ▲

§ 1.3. Перестановки

Рассмотрим теперь отдельно случай, когда $m=n$. Соответствующие этому случаю размещения называются *перестановками.*

Определение. *Перестановкой из n элементов называется любое упорядоченное множество, в которое входят по одному разу все n различных элементов данного множества.*

○ **Пример 1.3.** Пусть имеются числа 3, 5, 7. Этому множеству чисел соответствует 6 перестановок: 357; 375; 537; 573; 753; 735. ●

○ **Пример 1.4.** Слова «барк» и «краб» образованы в результате перестановки букв, составляющих слово «брак». Число таких перестановок равно 24, так как

$$A_4^4=\frac{4!}{0!}=4!=1\cdot 2\cdot 3\cdot 4=24. \; ●$$

Отметим, что перестановки состоят из одних и тех же элементов, но отличающихся между собой порядком. Число перестановок n различных элементов будем обозначать символом P_n.

Теорема. *Число перестановок n различных элементов равно n!, т. е.*

$$P_n=n! \tag{1.3}$$

□ Так как перестановки являются частным случаем размещений, то при $n=m$ получаем $P_n=A_n^n=n!/0!=n!$ ■

Задача 1.7. Сколькими способами можно расставить девять различных книг на полке, чтобы определенные четыре книги стояли рядом?

△ *Способ* 1. Будем считать выделенные книги за одну книгу. Тогда для шести книг существует $P_6=6!=720$ перестановок. Однако четыре определенные книги можно переставить между собой $P_4=4!=24$ способами. По правилу умножения имеем $P_6\cdot P_4=720\cdot 24=17\ 280$.

Способ 2. Возможны следующие случаи: первая из четырех книг стоит на 1-м месте, тогда четвертая стоит на 4-м месте; первая книга стоит на 2-м месте, а четвертая стоит на 5-м, первая стоит на 6-м месте, тогда четвертая стоит на последнем, 9-м месте. Число таких случаев равно шести. Сами книги могут быть переставлены $P_4=4!=24$ способами. Значит, по правилу умножения выделенные четыре книги можно расставить $6\cdot P_4=144$ способами. Оставшиеся пять книг можно переставить $P_5=5!=120$ способами. Воспользовавшись снова правилом умножения, приходим к тому же результату, который был получен при первом способе решения: $P_4\cdot 6\cdot P_5=P_4\cdot P_6=17\ 280$. ▲

§ 1.4. Сочетания

Если два различных размещения состоят из одинаковых элементов некоторого множества, то они обязательно отличаются порядком входящих в них элементов. Иногда возникает необходимость не учитывать порядок элементов, входящих в размещение. В этом случае все $m!$ размещений, которые состоят из одних и тех же m элементов, считаются неразличимыми.

Предположим, что из чисел 3, 5, 7 необходимо составить различные произведения двух чисел. Таких произведений только три, а именно: $3\cdot 5=15$; $3\cdot 7=21$; $5\cdot 7=35$. Это объясняется тем, что произведения вида $3\cdot 5$ и $5\cdot 3$ совпадают, так как порядок сомножителей, входящих в произведение, не учитывается. Если требуется из указанных цифр составить двузначные числа, то таких чисел уже шесть. Запишем эти числа: 35, 53, 37, 73, 57, 75. Как видно, здесь уже пришлось учитывать порядок цифр.

При выборе делегации в составе 3 человек из 30 учащихся, очевидно, не надо учитывать порядок выбранных делегатов, так как все члены делегации равноправны. Однако, выбирая физорга, старосту, профорга из тех же учащихся, порядок уже приходится учитывать. Каждый конкретный результат выбора из 30 учащихся делегации в составе 3 человек — это сочетание из 30 по 3 в отличие от размещения из 30 по 3.

Определение. *Сочетанием из n элементов по m называется любое подмножество из m элементов, которые принадлежат множеству, состоящему из n различных элементов.*

○ **Пример 1.5.** Пусть имеется множество, содержащее четыре буквы: $\{A, B, C, D\}$. Запишем все возможные сочетания из указанных букв по три. Таких сочетаний будет четыре: ABC; ACD; ABD; BCD. Здесь в число сочетаний не включены, например, ACB, BCA, так как эти последовательности букв не отличаются от последовательности ABC, поскольку порядок элементов в сочетаниях не учитывается.

Число сочетаний из n различных элементов по m будем обозначать символом C_n^m, где $m \leqslant n$. Иногда это число обозначают символом $\binom{n}{m}$. ●

Прежде чем найти общую формулу для определения числа сочетаний, решим следующую задачу.

Задача 1.8. Необходимо выбрать в подарок 4 из 10 имеющихся различных книг. Сколькими способами можно это сделать?

△ Из смысла задачи следует, что порядок выбора книг не играет роли. Здесь важен только их состав. Как известно из предыдущего (см. задачу 1.6), число размещений из 10 по 4 равно $A_{10}^4 = 5040$. Пусть теперь выбраны 4 книги из 10. Число таких возможных выборов, где не учитывается порядок выбранных книг, равно C_{10}^4. Однако каждому из этих сочетаний (выборов) будут соответствовать $P_4 = 24$ перестановки выбранных книг. Тогда выбор 4 книг из 10 с учетом их порядка по правилу умножения возможен $C_{10}^4 \cdot P_4$ способами. С другой стороны, число указанных способов — это число размещений A_{10}^4. Таким образом, $A_{10}^4 = C_{10}^4 P_4$, откуда имеем $C_{10}^4 = A_{10}^4 / P_4 = 5040/24 = 210$. Таким образом, число возможных способов выбора подарка равно 210. ▲

Теорема. *Число сочетаний из n элементов по m равно* $\frac{n!}{m!(n-m)!}$, *т. е.*

$$C_n^m = \frac{n!}{m!(n-m)!}. \tag{1.4}$$

□ Чтобы получить размещение из n элементов по m, а их число равно A_n^m, надо выбрать m элементов из множества, содержащего n элементов, что можно сделать C_n^m способами, и организовать из них упорядоченное подмножество. Последнюю операцию можно выполнить P_m способами. Таким образом, чтобы получить A_n^m размещений, надо выполнить две операции, которые можно осуществить C_n^m и P_m способами соответственно. Поэтому, согласно правилу умножения, можно записать $A_n^m = C_n^m \cdot P_m$. Разделив последнее равенство почленно на P_m, что возможно, так как $P_m \neq 0$, получим

$$C_n^m = A_n^m / P_m = n!/m!(n-m)! \ ■$$

Следствие. *Число сочетаний из n элементов по $n-m$ равно числу сочетаний из n элементов по m, т. е.*

$$C_n^{n-m} = C_n^m. \tag{1.5}$$

□ В выражении для C_n^m поменяем местами сомножители $m!$ и $(n-m)!$; в результате имеем

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

Задача 1.9. Имеется 10 белых и 5 черных шаров. Сколькими способами можно выбрать 7 шаров, чтобы среди них были 3 черных?

△ Среди выбранных шаров 4 белых и 3 черных. Белые шары можно выбрать $C_{10}^4 = \frac{10!}{4!6!} = 210$ способами. Черные шары можно выбрать $C_5^3 = 5!/3!\,2! = 10$ способами. Тогда по правилу умножения искомое число способов равно $C_{10}^4 \cdot C_5^3 = 2100$. ▲

При решении последней задачи была использована формула (1.4), так как порядок выбора шаров не играл роли.

Еще раз подчеркнем разницу между размещениями и сочетаниями: в размещениях учитывается порядок входящих в них элементов, а в сочетаниях — не учитывается. При решении задач это не следует забывать. Кроме того, надо иметь в виду, что использование правила умножения приводит к необходимости учитывать порядок элементов при выборе их из какого-либо множества.

Задача 1.10. Сколькими способами можно группу из 12 человек разбить на две подгруппы, в одной из которых должно быть не более пяти, а во второй — не более девяти человек?

△ Первая подгруппа может состоять либо из трех, либо из четырех, либо из пяти человек. Подгруппу из трех человек можно выбрать $C_{12}^3 = 220$ способами. Подгруппу из четырех человек можно выбрать $C_{12}^4 = 495$ способами, а подгруппу из пяти человек — $C_{12}^5 = 792$ способами. Учитывая, что выбор первой подгруппы однозначно определяет вторую, найдем по правилу сложения искомое число способов:

$$C_{12}^3 + C_{12}^4 + C_{12}^5 = 1507. \ ▲$$

Задача 1.11. Десять команд участвуют в розыгрыше первенства по футболу, лучшие из которых занимают 1-е, 2-е и 3-е места. Две команды, занявшие последние места, не будут участвовать в следующем таком же первенстве. Сколько разных вариантов результата первенства может быть, если учитывать только положение первых трех и последних двух команд?

△ *Способ 1.* Первые три места могут быть распределены $A_{10}^3 = 10!/7! = 720$ способами. В результате останется семь команд, две из которых выбывают из следующего первенства. Так как в этом случае порядок выбывших команд не важен, то это может произойти $C_7^2 = 7!/2! \cdot 5! = 21$ способами. Согласно правилу умножения получаем, что число разных результатов первенства равно $A_{10}^3 \cdot C_7^2 = 15\,120$.

Способ 2. Выделим без учета порядка пять команд из общего числа команд. В эту группу входят три команды, занявшие призовые места, и две выбывающие команды. Такую операцию можно выполнить C_{10}^5 способами. Из этих пяти команд без учета порядка выделим две команды, которые выбывают, что можно сделать C_5^2 способами. Для оставшихся трех команд распределение призовых мест возможно P_3 способами. По правилу умножения все три операции можно выполнить $C_{10}^5 \cdot C_5^2 \cdot P_3 = 15\,120$ способами.

Способ 3. Распределение всех 10 мест в первенстве возможно $P_{10} = 10!$ способами. Однако перестановки команд, занявших места с 4-го по 8-е, и перестановки команд, занявшие 9-е и 10-е места, на результаты первенства не оказывают влияния. Число таких перестановок равно $5! \cdot 2!$, а число различных результатов первенства равно $10!/5! \cdot 2! = 15\,120$. ▲

Докажем два свойства числа сочетаний, которые могут быть полезными при решении комбинаторных задач.

Теорема. *Имеет место равенство (правило Паскаля)*

$$C_n^m = C_{n-1}^{m-1} + C_{n-1}^m, \quad 1 \leqslant m < n. \tag{1.6}$$

□ Имеем

$$C_{n-1}^{m-1} + C_{n-1}^m = \frac{(n-1)!}{(m-1)!\,(n-m)!} + \frac{(n-1)!}{m!\,(n-m-1)!} = \frac{(n-1)!}{(m-1)!\,(n-m-1)!} \times$$
$$\times \left[\frac{1}{n-m} + \frac{1}{m}\right] = \frac{(n-1)!}{(m-1)!\,(n-m-1)!}\,\frac{n}{(n-m)m} =$$
$$= \frac{(n-1)!\,n}{(m-1)!\,m\,(n-m-1)!\,(n-m)} = \frac{n!}{m!\,(n-m)!} = C_n^m. \ ■$$

Теорема. *Имеет место равенство*

$$C_n^0 + C_n^1 + C_n^2 + C_n^3 + \ldots + C_n^n = 2^n. \tag{1.7}$$

□ Предварительно покажем, что $C_n^0 = C_n^n = 1$ для любого $n \geqslant 1$. Действительно, учитывая, что $0! = 1$ (см. § 1.2), имеем $C_n^0 = n!/(0!\ n!) = 1$, $C_n^n = n!/n! \cdot 0! = 1$. Далее, используя равенство (1.6), получаем

$$C_n^0 + C_n^1 + C_n^2 + C_n^3 + \ldots + C_n^n = C_{n-1}^0 + (C_{n-1}^0 + C_{n-1}^1) +$$
$$+ (C_{n-1}^1 + C_{n-1}^2) + (C_{n-1}^2 + C_{n-1}^3) + \ldots + (C_{n-1}^{n-2} + C_{n-1}^{n-1}) + C_{n-1}^{n-1} =$$
$$= 2(C_{n-1}^0 + C_{n-1}^1 + C_{n-1}^2 + \ldots + C_{n-1}^{n-1}) = 2 \cdot 2(C_{n-2}^0 + C_{n-2}^1 +$$
$$+ C_{n-2}^2 + \ldots + C_{n-2}^{n-2}) = \ldots = 2^{n-1}(C_1^0 + C_1^1) = 2^{n-1}(1+1) = 2^n. \ ■$$

Формулу (1.7) проиллюстрируем следующей задачей.

Задача 1.12. Сколько существует вариантов опроса 11 учащихся на одном занятии, если ни один из них не будет подвергнут опросу дважды и на занятии может быть опрошено любое количество учащихся, причем порядок, в котором опрашиваются учащиеся, безразличен?

△ Преподаватель может не спросить ни одного из 11 учащихся, что является одним из вариантов. Этому случаю соответствует C_{11}^0. Преподаватель может опросить только одного из учащихся. Таких вариантов C_{11}^1. Если преподаватель будет опрашивать двух учащихся, то число вариантов опроса равно C_{11}^2. Для опроса трех учащихся существует C_{11}^3 вариантов и т. д. Наконец, могут быть опрошены все учащиеся. Число вариантов в этом случае равно C_{11}^{11}. Тогда по правилу сложения число всех возможных вариантов опроса равно

$$C_{11}^0 + C_{11}^1 + C_{11}^2 + C_{11}^3 + \ldots + C_{11}^{11}.$$

С другой стороны, для каждого из учащихся существует две возможности: он будет опрошен или не опрошен на данном занятии. Другими словами, каждую из 11 операций, заключающихся в том, что каждый ученик будет либо опрошен, либо не опрошен, можно выполнить по правилу умножения $2 \cdot 2 \cdot 2 \cdot 2 \ldots 2 = 2^{11}$ способами, что и следовало ожидать, так как

$$C_{11}^0 + C_{11}^1 + C_{11}^2 + \ldots + C_{11}^{11} = 2^{11}. \ ▲$$

ЧАСТЬ II
ЭЛЕМЕНТЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

ГЛАВА 2
ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ

§ 2.1. Случайные события

Каждая наука при изучении явлений материального мира оперирует теми или иными понятиями, среди которых обязательно имеются основополагающие. Если для геометрии это понятия точки, прямой, плоскости, а для математического анализа — функции и ее предела, то для теории вероятностей одним из основных является понятие события.

Под *событием* понимается явление, которое происходит в результате осуществления какого-либо определенного комплекса условий. Осуществление этого комплекса условий будем называть *опытом* или *испытанием*. Здесь предполагается, что комплекс условий, в результате которого наступает определенное событие, может быть воспроизведен сколь угодно большое число раз, т. е. имеется возможность проводить неоднократно испытание в неизменных условиях. Вообще говоря, полностью совпадения всех условий для каждого испытания добиться невозможно и поэтому можно говорить лишь о некотором приближенном равенстве условий испытаний. При проведении испытания необязательно должен присутствовать и участвовать в нем сам исследователь. Опыт можно поставить мысленно, или «окружающая природа поставит этот опыт сама». В последнем случае исследователь выступает в роли наблюдателя. Например, для появления радуги во время дождя должны иметь место определенные атмосферные условия. И каждый раз, как только возникают соответствующие условия, можно любоваться радугой. На практике часто в силу объективных причин бывает трудно воспроизвести весь комплекс условий, который необходим для появления определенного события. Это обычно связано с недостаточно подробно изученной природой явления или с технической невозможностью его воссоздания в данное время. Поэтому при выполнении неполного комплекса условий интересующее событие может не наступить, и будет иметь место какое-нибудь другое. В силу

изменяющихся независимо от воли исследователя неучтенных условий при повторении испытания будут наступать те или иные события, причем не известно заранее, какие из них. Такие события обычно называются *случайными*. Например, один выстрел из винтовки по мишени — испытание, причем его можно повторить сколько угодно раз. В результате такого испытания наступит одно из 11 следующих случайных событий: «в результате одного выстрела выбито 0 очков»; «в результате выстрела выбито 1 очко»; «в результате выстрела выбито 2 очка»; ...; «в результате выстрела выбито 10 очков». Здесь трудно сказать заранее, какое из названных случайных событий в результате выстрела произойдет наверняка.

Определение. *Случайным называется событие, которое может произойти или не произойти в результате некоторого испытания (опыта).*

В дальнейшем для простоты слово «случайный» будем опускать. События обычно обозначают заглавными буквами латинского алфавита A, B, C, ... и т. д.

○ **Пример 2.1.** Бросается симметричная монета. Бросание монеты — испытание. «Монета упала «гербом» вверх» и «монета упала решкой вверх» — возможные события. ●

○ **Пример 2.2.** Завтра днем — ясная погода. Здесь наступление дня является испытанием. «В течение дня наблюдается ясная погода» — событие. ●

Из приведенных примеров следует, что событие можно охарактеризовать, сформулировав его в виде предложения. Однако не всякое предложение выражает событие. Например, предложение «Миру нужен мир» не задает какого-либо события, которое могло бы произойти или не произойти.

На практике часто встречаются события, которые обязательно всякий раз происходят в результате определенного испытания. Также встречаются события, которые не могут произойти в результате испытания, сколько бы раз его ни проводили.

Определение. *Событие называется достоверным, если оно обязательно произойдет в результате данного испытания.*

○ **Пример 2.3.** Наступление ночи по прошествии дня — достоверное событие. ●

○ **Пример 2.4.** Кит — это млекопитающее. Это событие является достоверным. ●

Определение. *Событие называется невозможным, если оно не может произойти в результате данного испытания.*

○ **Пример 2.5.** Выпадение цифры 5 при бросании десятикопеечной монеты — невозможное событие. ●

○ **Пример 2.6.** Появление марсианина на улицах современного города — невозможное событие. ●

Ранее уже отмечалось, что в результате опыта (испытания) возможно появление того или иного события.

Некоторые из возможных в данном испытании событий могут наступить вместе, т. е. быть совместными. В результате испытания может возникнуть и другая ситуация, когда появление одного из возможных в данном случае событий обязательно исключает появление определенного другого события, т. е. события могут быть и несовместными. Пусть бросается игральный кубик, на гранях которого различное число очков от 1 до 6. В результате этого испытания могут наступить такие события, как «на верхней грани четное число очков»; «число очков на верхней грани равно 3»; «на верхней грани 6 очков». Появление на верхней грани кубика трех очков исключает появление шести очков. Поэтому последние два события являются несовместными при одном испытании. Такие события, как выпадение четного числа очков и выпадение числа очков, равного 6, уже являются совместными, поскольку число 6 является четным.

Определение. *События A и B называются несовместными, если в результате данного испытания появление одного из них исключает появление другого.*

○ **Пример 2.7.** Бросается монета. Появление решки исключает появление «герба», и наоборот. Поэтому события «появилась решка» и «появился «герб» — несовместные события. ●

Определение. *События A и B называются совместными, если в результате данного испытания появление одного из них не исключает появления другого.*

○ **Пример 2.8.** В аудиторию вошел человек. События «в аудиторию вошел человек старше 30 лет» и «в аудиторию вошел мужчина» — совместные, поскольку в аудиторию может войти мужчина старше 30 лет. ●

Пусть некто набирает номер телефона при исправной телефонной сети. В результате этого испытания возможны следующие события: «номер занят», «номер свободен». Эти два события взаимосвязаны, так как непоявление одного из них влечет за собой обязательно появление другого.

Выше понятие «влечет», употреблялось применительно к событиям. В дальнейшем будем говорить, что событие A влечет событие B, если всякий раз, когда наступает событие A, наступает и событие B.

Определение. *Два события A и $\bar{A}$* называются противоположными или взаимно дополнительными, если непоявление одного из них в результате данного испытания влечет появление другого ($\bar{A}$ читается «не A»).

Следует заметить, что события A и $\bar{A}$ несовместны.

○ **Пример 2.9.** Если при проверке оказалось, что некоторое изделие имеет дефекты, то это изделие не может быть стандартным, и наоборот. Поэтому события «изделие бракованное» и «изделие стандартное» — противоположные. ●

○ **Пример 2.10.** Событию «все спортсмены команды завоевали призовые места» противоположным является «хотя бы один из спортсменов команды не занял призовое место». ●

Рассмотрим теперь на примере множество события, которые могут произойти в результате некоторого испытания.

Пусть производится один выстрел по мишени. Выделим следующие события, которые могут произойти в результате этого испытания (опыта):

A_i — «выбито i очков, где i изменяется от 0 до 10»,
B — «выбито четное число очков»,
C — «выбито нечетное число очков»,
D — «выбито более 4 очков»,
E — «выбито менее 5 очков»,
F — «число выбитых очков делится на 11»,
Q — «число выбитых очков меньше 12».

Ограничимся этими событиями, хотя их список можно было продолжить. Нетрудно заметить, что между отдельными событиями наблюдается определенная связь. Например, события A_i попарно несовместны; то же самое можно сказать и про пары событий B и C, A_6 и E, A_3 и B. Пары событий B и C, D и E являются противоположными событиями, а пары событий A_8 и B, A_3 и C, A_2 и E, A_5 и Q — совместными. Если произошло событие A_8, т. е. выбито 8 очков, то обязательно произойдет событие B, так как число 8 четное. Заметим, что эти события совместны. Предположим, что выбито 2 очка, т. е. произошло событие B, однако событие D не наступило, хотя события B и D совместны. Особое место занимают события F и Q. Первое из них является невозможным событием, а второе — достоверным.

Обратим внимание на следующие группы события, которые обладают одинаковыми свойствами:

1) A_0, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9, A_{10};
2) B, C;
3) D, E.

В результате опыта обязательно произойдет одно и только одно событие, неважно какое, из всех событий, принадлежащих одной из рассматриваемых групп, причем любые два события, принадлежащие одной группе, несовместны. Теперь на основании рассмотренного примера сформулируем два определения.

Определение. *Событие A называется благоприятствующим событию B, если появление события A влечет за собой появление события B.*

В рассмотренном выше примере события A_1, A_2, A_3, A_4 являются благоприятствующими событию E, а события A_1, A_3, A_5, A_7, A_9 — благоприятствующими событию C.

Определение. *Если группа событий такова, что в результате испытания обязательно должно произойти хотя бы одно из них и любые два из них несовместны, то эта группа событий называется полной группой событий.*

Приведенные выше три группы событий, по определению, являются полными. При решении вероятностных задач большую роль играет выбор полной группы событий. Если полная группа выбрана удачно, то решение задачи значительно упрощается. В противном случае оно будет представлять определенные трудности или даже не будет найдено. С некоторым преувеличением можно сказать, что «выбор полной группы есть искусство». В дальнейшем каждое событие из полной группы попарно несовместных событий будем называть *исходом* данного опыта (испытания). Иногда исходы испытания называются *элементарными событиями*. Рассмотрим пример полной группы попарно несовместных событий.

○ **Пример 2.11.** Бросается игральный кубик. События, заключающиеся в том, что на верхней грани кубика появится 1, 2, 3, 4, 5, 6 очков, образуют полную группу событий, так как в результате опыта кубик обязательно упадет какой-нибудь гранью вверх, а значит, произойдет одно из указанных событий. Все эти события попарно несовместны, так как кубик не может упасть одновременно двумя гранями вверх. Группа событий, состоящая в том, что выпадет 1, 2, 4, 5, 6 очков, не является полной, поскольку в результате опыта может выпасть 3 очка, а такое событие не содержится в указанной группе. ●

На основании определений полной группы и достоверного события можно сделать вывод, что событие, заключающееся в появлении одного, неважно какого, из событий полной группы,— достоверное.

Понятие полной группы позволяет доказать утверждение, что два несовместных события, образующие полную группу, являются противоположными событиями. Действительно, если в результате опыта не произойдет одно из событий, то обязательно произойдет другое, так как, по определению полной группы, одно из них должно обязательно произойти. Такие события, по определению, называются *противоположными*.

Предположим, что имеется идеально симметричный кубик, изготовленный из однородного материала. В этом случае нет оснований считать выпадение 6 очков более возможным, чем выпадение, например, 2 очков. Поэтому правомочно считать, что возможность выпадения 6 очков та же, что и выпадения любого другого числа очков, т. е. все 6 событий, которые могут наступить в результате бросания игрального кубика, равновозможны.

Определение. *События называются равновозможными, если по условию испытания нет оснований считать какое-либо из них более возможным, чем любое другое.*

○ **Пример 2.12.** В урне находятся тщательно перемешанные 10 одинаковых на ощупь шаров. Среди них 5 белых и 5 черных. Наудачу вынимается один шар. Здесь события «появился белый шар» и «появился черный шар» равновозможны. ●

○ **Пример 2.13.** «Появление «герба» и «появление решки» при бросании симметричной монеты — равновозможные события. ●

В рассмотренных примерах вывод о равновозможности событий делался из соображений симметрии. Условия опытов

были симметричны относительно рассматриваемых событий, т. е. все события были «поставлены в равные условия». Высказывая утверждение о равновозможности событий, приходится делать определенные допущения. Например, при бросании монеты предполагается, что она сделана из однородного материала, представляет собой круг, симметрична (т. е. не погнута), что наличие чеканки не влияет на положение центра тяжести. Эти допущения, основанные на здравом смысле, позволяют обосновать вывод о равновозможности событий, однако это не всегда можно сделать исходя из условий испытания.

§ 2.2. Операции над событиями

В некоторых разделах математики вводятся операции над изучаемыми объектами. Так как свойства этих операций аналогичны свойствам арифметических действий, то и свое название они получили по аналогии с арифметическими действиями. К таким операциям относятся операции сложения и умножения чисел, векторов, матриц. Подобные операции над событиями вводятся и в теории вероятностей; они позволяют упростить форму записи, а иногда и логическое построение рассуждений.

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

Сумму событий будем обозначать знаком «+».

Если имеются два совместных события A и B, то сумма $A+B$ означает, что наступит событие A, или событие B, или оба события вместе. Если же события несовместны, то событие $A+B$ заключается в том, что наступит событие A или событие B, так как совместное наступление событий A и B невозможно. В этом случае знак «+» заменяет союз «или». Про событие $A+B+C$ можно сказать, что оно состоит в наступлении одного из событий A, B, C, или в совместном наступлении пары событий A и B, A и C, B и C, или в совместном наступлении всех трех событий.

○ **Пример 2.14.** В урне находятся красные, белые и черные шары. Вынимается один шар. Возможны следующие события: A — «вынут красный шар»; B — «вынут белый шар»; C — «вынут черный шар». Событие $A+B$ означает, что произошло событие «вынут не черный шар», а событие $B+C$ — «вынут не красный шар». ●

○ **Пример 2.15.** Турист хочет и имеет возможность посетить три города. Обозначим события: A — «турист посетил город A». B — «турист посетил город B»; C — «турист посетил город C». Событие $A+C$ заключается в том, что турист посетил только один из городов A и C или он посетил их оба. ●

○ **Пример 2.16.** Монета бросается четыре раза. Рассмотрим следующие события: A_i — «герб» появился i раз», $i=0, 1, 2, 3, 4$. Событие $B=A_0+A_1+A_2$ означает, что «герб» выпал не более двух раз, т. е. произошло событие

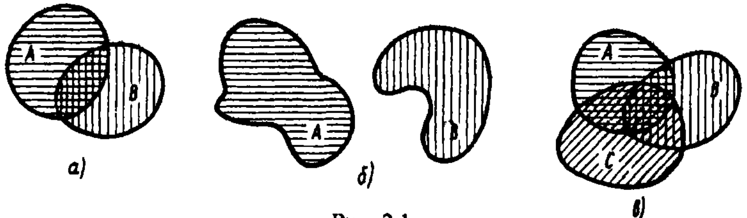


Рис. 2.1

«или «герб» не появился, или «герб» появился один раз, или «герб» появился два раза». Событие $A_1+A_2+A_3+A_4$ означает, что «герб» появился хотя бы один раз. Все указанные события попарио несовместны, поэтому знак «+» в данном случае означает союз «или». Заметим, что рассмотренное испытаиие можно замеиить одним бросаиием четырех монет сразу. ●

Операция сложения событий, как и другие операции, имеет полезную геометрическую интерпретацию. Пусть на плоскости имеется некоторая фигура A и на плоскость произвольным образом бросается точка. Если точка попала в фигуру A, то будем считать, что произошло событие A.

На рис. 2.1, *а* изображены фигуры A и B, которым соответствуют события A и B. Если точка попадет в область с двойной штриховкой, то события A и B произойдут одновременно. Этот рисунок иллюстрирует совместные события. Сумме событий соответствует вся заштрихованная область. Она обведена жирной линией. На рис. 2.1, *б* изображены две области, которым соответствуют два несовместных события, так как точка не может одновременно попасть и в область A, и в область B, поэтому события A и B не могут произойти вместе. Сумме $A+B$ соответствуют две непересекающиеся области, обведенные жирной линией. На рис. 2.1, *в* жирной линией отмечена область, соответствующая сумме $A+B+C$. Если в этом случае точка попадет в область с тройной штриховкой, то наступят совместно все три события. При попадании точки в область с двойной штриховкой произойдут совместно события A и B, или A и C, или B и C. Приведенные рисунки обычно называют *диаграммами Венна*.

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

Произведение событий будем обозначать знаком «·». В данном случае знак «·» заменяет союз «и». Например, если произошло ABC, то это означает, что наступило событие «A и B и C». В дальнейшем знак · будем опускать.

○ **Пример 2.17.** Пусть имеются следующие события: A — «из колоды карт вынута «дама»; B — «из колоды карт вынута карта пиковой масти». Очевидно, AB есть событие «вынута дама пик». ●

○ **Пример 2.18.** Бросается игральный кубик. Рассмотрим следующие события: A — «число выпавших очков меньше 5»; B — «число выпавших очков

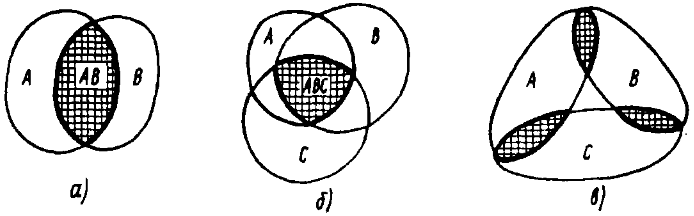


Рис. 2.2

больше 2»; C — «число выпавших очков четное». Тогда событие ABC заключается в том, что выпало 3 очка. ●

Рассмотрим теперь диаграммы Венна, соответствующие произведению событий. На рис. 2.2, *а* изображена диаграмма Венна, иллюстрирующая произведение двух событий, а на рис. 2.2, *б* — диаграмма, иллюстрирующая произведение трех событий. Любая точка, попавшая в заштрихованную и обведенную жирной линией область, обязательно попадет в области A и B в первом случае и в области A, B и C во втором случае. Это означает, что наступят события AB и ABC, так как в первом случае одновременно наступят события A и B, а во втором — события A, B и C.

Пусть имеются три события A, B, C, каждые два из которых совместны, но произведение этих событий является невозможным событием. Диаграмма Венна, иллюстрирующая этот случай, изображена на рис. 2.2, *в*. Как видно из рисунка, не найдется ни одной точки плоскости, которая принадлежала бы всем трем областям A, B, C одновременно. Следовательно, события A, B, C не могут наступить все вместе, т. е. они не являются совместными в совокупности, хотя они попарно совместны.

○ **Пример 2.19.** Бросается игральный кубик. Исходами этого опыта являются события A_i, $i=1, 2, ..., 6$, состоящие в том, что число выпавших очков равно i. Выделим следующие возможные события: B — «число выпавших очков меньше 4»; C — «число выпавших очков больше 2»; D — «число выпавших очков четно». Событня B благоприятствуют исходы A_1, A_2, A_3, событию C — исходы A_3, A_4, A_5, A_6, а событию D — исходы A_2, A_4, A_6. Произведеиием событий B и C является событие A_3, т. е. $BC=A_3$. События A_3 и D несовместны, так как число 3 не является четным. Поэтому событие $A_3D=BCD$ является невозможным. Этому событию не благоприятствует ни один из 6 исходов, несмотря на то что событню BC благоприятствует событие A_3, событию BD — событие A_2, событию CD — события A_4, A_6. ●

Определение. *События называются совместными в совокупности, если каждое из них и произведение остальных являются совместными событиями.*

Из определения произведения следует, что события, участвующие в произведении, должны быть совместными в совокупности. В противном случае произведение событий является событием невозможным. В дальнейшем для простоты слова «в совокупности» будем опускать.

Используя операции сложения и умножения, можно сложное событие разложить на более простые события, и наоборот.

○ **Пример 2.20.** Пусть в цехе некоторого завода имеется три станка. Рассмотрим следующие события: A_1 — «в течение дня произойдет поломка 1-го станка»; A_2 — «в течение дня произойдет поломка 2-го станка»; A_3 — «в течение дня произойдет поломка 3-го станка». Испытание будет заключаться в том, что станки работают с начала до конца рабочего дня и фиксируется поломка каждого из них. По определению противоположного события, события $\bar{A}_1$, $\bar{A}_2$, $\bar{A}_3$ означают бесперебойную работу соответствующих станков в течение дня. Событие $A_1\bar{A}_2\bar{A}_3$ заключается в том, что только 1-й станок потребует ремонта за время испытания. Событие $\bar{A}_1A_2\bar{A}_3$ означает поломку только 2-го станка, а событие $\bar{A}_1\bar{A}_2A_3$ — поломку только 3-го станка. Тогда сложное событие $A_1\bar{A}_2\bar{A}_3+\bar{A}_1A_2\bar{A}_3+\bar{A}_1\bar{A}_2A_3$ означает событие «в течение дня произошла поломка только одного станка». Событие, состоящее в том, что во время испытания произойдет поломка не менее одного станка (хотя бы одного), можно представить в виде $A_1\bar{A}_2\bar{A}_3+\bar{A}_1A_2\bar{A}_3+\bar{A}_1\bar{A}_2A_3+\bar{A}_1A_2A_3+A_1\bar{A}_2A_3+A_1A_2\bar{A}_3+A_1A_2A_3$. Противоположное ему событие $\bar{A}_1\bar{A}_2\bar{A}_3$ заключается в том, что ни один станок не потребует ремонта. Эти два события связаны равенством $\overline{\bar{A}_1\bar{A}_2\bar{A}_3}=A_1\bar{A}_2\bar{A}_3+\bar{A}_1A_2\bar{A}_3+\bar{A}_1\bar{A}_2A_3+\bar{A}_1A_2A_3+A_1\bar{A}_2A_3+A_1A_2\bar{A}_3+A_1A_2A_3$. ●

Рассмотрим примеры противоположных событий.

○ **Пример 2.21.** Необходимо выбрать делегата на конференцию на общем собрании четырех групп учащихся. Рассмотрим события A_i — «делегат выбран из i-й группы», $i=1, 2, 3, 4$. В этом случае событие $\bar{A}_1$ можно представить в виде $A_2+A_3+A_4$, т. е. событие «делегат выбран из 2-й, или 3-й, или 4-й группы» является противоположным событию «делегат выбран из 1-й группы». ●

○ **Пример 2.22.** Произведено три выстрела по мишени. Обозначим A_i — «в мишени i пробоин», $i=0, 1, 2, 3$, а B — «в мишени хотя бы одна пробоина». Очевидно, $B=A_1+A_2+A_3$. Событие A_0 — «в мишени ни одной пробоины» противоположно событию B. Символически это можно записать так: $\bar{B}=A_0$, или $A_1+A_2+A_3=\bar{A}_0$, или $\overline{A_1+A_2+A_3}=A_0$. ●

Рассмотренные приемы представления сложных событий, и особенно представление сложного события через ему противоположное, часто используют в теории вероятностей.

§ 2.3. Классическая формула вероятности

В повседневной жизни в разговоре часто используется слово «вероятный». Например, «к вечеру, вероятно, пойдет дождь», «это невероятный случай», «вероятнее всего он опоздает». При употреблении этого слова интуитивно оценивается возможность наступления того или иного события. Можно сказать, что одно событие наступит чаще, чем другое. В этом случае говорят, что оно более возможно, т. е. его наступление более вероятно. Естественно, при такой оценке человеку помогает здравый смысл и жизненный опыт.

Например, предполагая, что первым увиденным человеком в кинотеатре на детском киносеансе скорее всего будет школьник, а не взрослый, мы считаем, что на детском киносеансе заведомо больше детей, чем взрослых. Из опыта

известно, что при выполнении многих видов работ вредна торопливость. В спешке можно совершить такое действие, которое сведет на нет всю предыдущую работу. Иначе говоря, при спешке более вероятен брак в работе, т. е. вероятность (возможность) выхода брака выше.

Пусть, например, в ящике находятся 28 одинаковых по внешнему виду изделий, среди которых два изделия 3-го сорта и по тринадцать изделий 1-го и 2-го сорта. Наудачу вынимается одно изделие. В данном случае разумно считать событие «вынуто изделие 1-го сорта» более возможным, чем событие «вынуто изделие 3-го сорта», так как изделий 1-го сорта значительно больше, чем изделий 3-го сорта.

Очевидно, события «вынуто изделие 1-го сорта» и «вынуто изделие 2-го сорта» имеют одинаковую возможность появления, поскольку количество этих изделий одинаково.

Однако в жизни чаще встречаются события, сравнить или оценить возможности появления которых, основываясь на чисто интуитивных соображениях, трудно. Например, это можно сказать про события «герб» появился два раза при пятикратном бросании монеты», «во время решения задачи отказала ЭВМ» и т. д. Как видно из приведенных примеров, каждое событие обладает определенной степенью возможности наступления, т. е. определенной оценкой. Такую оценку события называют вероятностью события.

Определение. *Вероятность события — это численная мера объективной возможности его появления.*

По определению, событию можно поставить в соответствие определенное число — его вероятность. Однако приведенное определение не дает формулу для нахождения этого числа. Во многих случаях проблема решается, если применить классическую формулу вероятности, которая дается ниже.

Пусть проводится опыт, в результате которого могут наступить те или иные события. Если эти события образуют полную группу попарно несовместных и равновозможных событий, то говорят, что опыт «сводится к схеме случаев». Здесь *случаем* называют каждое из событий (исходов), принадлежащих выделенной полной группе. Для опытов, которые сводятся к схеме случаев, применима классическая формула вероятности.

Пусть имеется полная группа попарно несовместных и равновозможных событий. Вероятность $P(A)$ наступления события A вычисляется как отношение числа исходов, благоприятствующих наступлению события, к числу всех исходов испытания.

Если N — число всех исходов испытания, а M — число исходов, благоприятствующих событию A, то

$$P(A)=\frac{M}{N}. \tag{2.1}$$

Проиллюстрируем формулу решением следующих задач.

Задача 2.1. Какова вероятность появления четного числа очков при одном бросании игрального кубика?

Δ Обозначим через A событие «выпадет четное число». Рассмотрим события A_i — «выпадет i очков», $i=1, 2, ..., 6$. Очевидно, что эти события равновозможны и образуют полную группу (см. пример 2.11). Тогда число всех исходов $N=6$. Выпадению четного числа очков благоприятствуют события A_2, A_4, A_6, т. е. $M=3$. Искомая вероятность, по определению, равна отношению числа благоприятствующих исходов к числу всех исходов: $P(A)=M/N=3/6=1/2$. ▲

Задача 2.2. *Ошибка Д'Аламбера.* Бросаются две монеты. Какова вероятность, что обе монеты упадут «гербом» кверху?

Δ При решении этой задачи возможно рассуждать примерно так: в результате бросания двух монет возможны следующие три события: «выпали два «герба»; «выпали две решки»; «выпали «герб» и решка». Эти события находятся в равных условиях, поэтому их вероятности равны 1/3.

Решим эту задачу иначе. Возможные события, которые являются результатом опыта с двумя монетами, будем обозначать двумя буквами. Первая буква означает выпадание «герба» (Г) или решки (Р) на 1-й монете, а вторая — выпадание «герба» или решки на 2-й. Тогда 4 исхода бросания двух монет можно записать так: ГГ; ГР; РГ; РР. Все эти исходы несовместны, равновозможны и образуют полную группу. Пусть событие A — «выпали два «герба». Этому событию благоприятствует только один исход ГГ. Поэтому $M=1$, $N=4$, $P(A)=M/N=1/4$. Теперь нетрудно заметить ошибку Д'Аламбера. Он считал, что события «выпали два «герба», «выпали «герб» и решка» равновозможны, а это не так. Последнему событию благоприятствуют два исхода: ГР и РГ, поэтому вероятность события «выпали «герб» и решка» $P=M/N=2/4=1/2\neq 1/3$. ▲

Задача 2.3. Бросаются два игральных кубика. Какова вероятность того, что сумма выпавших очков равна 7?

Δ Каждый из кубиков может упасть шестью различными способами. Тогда два кубика по правилу умножения могут упасть $6\cdot 6=36$ различными способами. Каждому такому способу соответствует событие, которое является исходом испытания бросания двух кубиков. В силу симметричности кубиков все эти события равновозможны и образуют полную группу несовместных событий. Поэтому число всех исходов бросания двух кубиков $N=36$. Для подсчета числа исходов, благоприятствующих событию A, состоящему в выпадении суммы очков, равной 7, составим следующую таблицу:

Число выпавших очков на 1-м кубике	1	2	3	4	5	6
Число выпавших очков на 2-м кубике	6	5	4	3	2	1
Сумма очков	7	7	7	7	7	7

Как видно из таблицы, число благоприятствующих исходов $M=6$. Тогда, по определению, $P(A)=M/N=6/36=1/6$. ▲

Задача 2.4. В группе 30 учащихся. Из них 12 юношей, остальные — девушки. Известно, что к доске должны быть вызваны двое учащихся. Какова вероятность, что это девушки?

Δ Обозначим событие, вероятность которого следует найти, буквой A. Очевидно, что по условию задачи порядок вызова к доске учащихся не играет роли. Найдем число всех исходов испытания, состоящего в вызове двух учащихся. Это число равно количеству способов, которыми можно выбрать двух учащихся из 30. Порядок выбора не играет роли, поэтому $N=C_{30}^2$. Найдем теперь число M благоприятствующих исходов. Для этого следует определить число способов выбора двух девушек из 18. Оно равно C_{18}^2. По определению вероятности,

$$P(A)=M/N=C_{18}^2/C_{30}^2=51/145. \ ▲$$

Задача 2.5. Бросаются два игральных кубика. Какова вероятность того, что сумма выпавших очков меньше 13?

Δ Число N всех исходов испытания равно 36 (см. задачу 2.3) и любой из этих исходов благоприятствует наступлению события A, заключающегося в том, что сумма выпавших очков меньше 13. Действительно, максимальная сумма очков выпадет тогда, когда на каждом кубике выпадет по 6 очков. Но эта сумма равна 12, и она меньше 13. Следовательно, $M=36$. Тогда

$$P(A)=M/N=36/36=1. \ ▲$$

Задача 2.6. Цифры 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 написаны на карточках, которые тщательно перемешаны. Произвольным образом вынимаются подряд три карточки и кладутся в ряд. Какова вероятность того, что число, составленное из трех цифр, которые написаны на карточках, больше 987?

Δ Найдем число исходов, благоприятствующих событию A, вероятность которого следует определить. Однако это число равно нулю, так как из данных карточек нельзя составить число, большее 987. Действительно, любое число, большее 987, например 988, 991, предполагает повторение цифр, что невозможно по условию задачи. Таким образом, число благоприятствующих исходов $M=0$. Тогда

$$P(A)=M/N=0/N=0. \ ▲$$

Заметим, что, сколько бы раз ни проводился опыт, событие A в задаче 2.6 никогда не наступит, поэтому оно является невозможным, а в задаче 2.5 событие A наступит всегда; следовательно, оно является достоверным.

Обобщим результаты, полученные при решении задач, в виде свойств вероятности и докажем их.

1^0. *Вероятность достоверного события равна* 1.

□ Если событие A достоверное, то любой исход испытания благоприятствует этому событию, но тогда $M=N$. Следовательно,

$$P(A)=M/N=N/N=1. \blacksquare$$

2^0. *Вероятность невозможного события равна* 0.

□ Если событие A невозможное, то ни один из исходов испытания не благоприятствует ему. Следовательно, $M=0$, но тогда

$$P(A)=M/N=0/N=0. \blacksquare$$

3^0. *Вероятность события A удовлетворяет двойному неравенству* $0 \leqslant P(A) \leqslant 1$.

□ Число исходов, благоприятствующих наступлению события, либо равно 0, либо N, либо, по определению вероятности, является частью всех N исходов испытания. Тогда $0 \leqslant M \leqslant N$, а значит, $0 \leqslant M/N \leqslant 1$. Следовательно, $0 \leqslant P(A) \leqslant 1$. ■

В настоящее время свойства вероятности определяются в виде аксиом, сформулированных А. Н. Колмогоровым.

Одним из основных достоинств классического определения вероятности является возможность вычислить вероятность события непосредственно, т. е. не прибегая к опытам, которые заменяют логическими рассуждениями.

§ 2.4. Статистическая вероятность. Геометрические вероятности

В предыдущем параграфе для непосредственного нахождения вероятности события использовалась формула $P(A)=M/N$, которая предполагает выполнение определенных условий. Например, было нужно, чтобы опыт сводился к схеме случаев. Последнее означает, что среди всех возможных событий, появляющихся в результате данного опыта, можно выделить полную группу попарно несовместных и равновозможных событий. Однако на практике такую группу зачастую бывает выделить трудно. Известно много опытов, результаты которых оказались непредсказуемыми, хотя, казалось бы, были предусмотрены все его исходы. Например, любой полет в космос можно считать опытом (испытанием). Однако вряд ли кто возьмет на себя смелость представить результат этого опыта в виде полной совокупности исходов. Невозможность представить во многих случаях результат опыта в виде полной группы событий является одним из недостатков классической формулы вероятности.

Не менее серьезным недостатком является трудность обоснования равновозможности событий. Определяя некоторые события как равновозможные, обычно руководствуются соображениями симметрии. Но симметричность условий опыта на практике наблюдается только в искусственно организованных опытах. Так, например, при бросании игрального кубика или монеты считается, что эти объекты симметричны, изготовлены из однородного материала и т. д. Однако не всякая ситуация позволяет сделать подобные предположения и, следовательно, воспользоваться классической формулой нахождения вероятности.

Так как вероятность события существует объективно, независимо от того, можно или нельзя применить в данном конкретном случае классическую формулу, то возникает вопрос, что считать вероятностью события и как ее вычислить. На практике давно было замечено, что при многократном повторении опытов относительная частота появления события в этих опытах стремится к устойчивости. Другими словами, при малом количестве опытов относительная частота появления события подвержена резким колебаниям, а при увеличении их числа эти колебания уменьшаются, относительная частота выравнивается, приближаясь к некоторому постоянному числу. Под *относительной частотой* появления события понимается отношение m/n, где n — число опытов, m — число появлений события. Я. Бернулли доказал (см. § 7.4), что при неограниченном увеличении числа опытов относительная частота появления события будет практически сколь угодно мало отличаться от некоторого постоянного числа, которое и принимается за вероятность события в отдельном опыте. При этом должны выполняться некоторые условия, обеспечение которых обычно не представляет трудности. Поэтому естественно относительную частоту появления события при достаточно большом числе испытаний называть *статистической вероятностью* в отличие от ранее введенной «математической» вероятности. Многие исследователи проверяли закон Я. Бернулли. Проводились многократные опыты, сводящиеся к схеме случаев, так как в этом случае можно вычислить вероятность, используя ее классическое определение. Так, например, Дж. Керрих провел опыты с бросанием монеты. Им было осуществлено 10 серий, каждая из которых содержала по 1000 бросков монеты. Оказалось, что «герб» выпал 502, 497, 511, 529, 504, 476, 507, 528, 504, 529 раз. Как видно, ни в одной из серий относительная частота выпадания «герба» не равна 0,5, т. е. не совпадает с вероятностью выпадания «герба» при одном бросании монеты. Результаты этого и ряда других опытов приведены в табл. 2.1.

Таблица 2.1

	Число испытаний	Частота проявления «герба»	Относительная частота (статистическая вероятность)
Опыт Керриха	10 000	5 087	0,5087
Опыт Бюффона	4 040	2 048	0,5069
Первый опыт Пирсона	12 000	6 019	0,5016
Второй опыт Пирсона	24 000	12 012	0,5005

Нетрудно заметить при изучении данных таблицы стремление относительной частоты с возрастанием числа опытов к вероятности $P(\Gamma)=0{,}5$. Так, на практике можно установить, что относительная частота события (статистическая вероятность) стремится к вероятности события в отдельном испытании. Поэтому относительную частоту при достаточно большом числе опытов можно считать приближенным значением вероятности.

Классическая формула вероятности предполагает конечное число всех исходов испытания. Но часто встречаются такие испытания, для которых число возможных исходов бесконечно. В подобных случаях классическая формула вероятности также неприменима, и в этом заключается еще один ее недостаток. Например, при изготовлении на станке некоторой детали нужно выдержать определенный размер. Здесь точность изготовления детали зависит от мастерства рабочего, точности измерительного инструмента и т. д. Таким образом, можно получить деталь любого размера, как угодно близкого к требуемому. Если под опытом понимать изготовление детали, то в результате такого опыта возможно бесконечное множество исходов.

Для преодоления указанного недостатка классической формулы вероятности часто используют некоторые понятия геометрии, если позволяют обстоятельства опыта. Во всех этих случаях предполагается возможность проведения, пусть даже теоретически, любого числа испытаний, а понятию равновозможности отводится главная роль.

Пусть на плоскости имеется некоторая фигура F, которая содержит фигуру f. На фигуру F наугад бросается точка, которая может оказаться в любой точке фигуры F. Другими словами, в результате бросания точки (опыта) возможно бесчисленное множество исходов. В данном случае нет оснований считать неравновозможными хотя бы два каких-либо исхода из всего множества исходов. Понятно, что брошенная точка может оказаться в фигуре f, а может там и не оказаться, поэтому возможно говорить о вероятности попадания точки в фигуру f. В данном случае будет естественным

связать вероятность с площадями фигур f и F: чем больше площадь фигуры f, тем больше точка имеет возможностей попасть в нее. Обозначим через A событие, состоящее в попадании брошенной точки в фигуру f, а через S_f и S_F — площади фигур f и F соответственно. Тогда под вероятностью события A будем понимать отношение данных площадей, т. е. $P(A)=S_f/S_F$. По аналогии с понятием благоприятствующего исхода, фигуру f будем называть благоприятствующей появлению события A.

В приведенном выше примере рассматривались двумерные области, мерами которых были соответствующие площади. Но область может быть одномерной (кривая, прямая, отрезок), тогда ее мерой является длина. Область также может быть и трехмерной (некоторое тело в пространстве), мерой ее является объем. Исходя из этого, дадим определение геометрической вероятности.

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

Применение геометрической вероятности проиллюстрируем решением следующих задач.

Задача 2.7. Перед окопами вдоль прямой линии через каждые 10 м установлены противотанковые мины. Перпендикулярно этой линии движется танк, ширина которого 3 м. Какова вероятность того, что танк пересечет линию установки мин невредимым, т. е. что мина не взорвется?

△ Ось симметрии танка может пересечь линию установки мин в любой ее точке, т. е. исходы испытания (пересечения линии) образуют бесконечное множество, поэтому здесь классическое определение вероятности неприменимо. Пусть отрезок прямой, расположенный между двумя соседними минами, изображен на рис. 2.3. Танк при своем движении может попасть на один из таких отрезков. Расстояние $|AB|=10$ м, $|AC|=|DB|=1{,}5$ м. Если ось симметрии танка попадет на отрезок AC или DB, то произойдет взрыв, а если ось симметрии попадет на отрезок CD, то его не будет. Таким образом, областью, благоприятствующей наступлению события A, заключающегося в беспрепятственном пересечении линии установки мин, является отрезок CD, а множеству всех исходов соответствует отрезок AB. Тогда

$$P(A)=|CD|/|AB|=7/10. \ \blacktriangle$$

Задача 2.8. На плоскости нанесена сетка квадратов со стороной 8 см. Найти вероятность, что брошенный на плоскость круг радиуса 1 см не пересечет ни одной стороны квадрата.

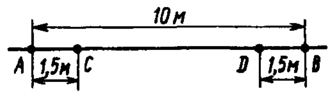


Рис. 2.3

△ При бросании круга его центр может попасть в любой из квадратов, начерченных на плоскости. Один из таких квадратов изображен на рис. 2.4. Сторона заштрихованного квадрата равна 6 см, а ширина «рамки» — 1 см. Если центр круга попадет в «рамку», то сам круг обязательно пересечет сторону квадрата со стороной 8 см. Если же центр круга попадет в заштрихованный квадрат со стороной 6 см, то он не пересечет границу квадрата со стороной 8 см. Поэтому заштрихованный квадрат является благоприятствующей областью наступления события A, вероятность которого следует найти. Таким образом,

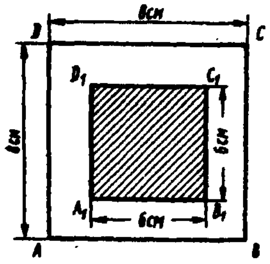


Рис. 2.4

$$P(A)=\frac{S_{A_1B_1C_1D_1}}{S_{ABCD}}=6^2/8^2=9/16. \ \blacktriangle$$

ГЛАВА 3. ТЕОРЕМЫ СЛОЖЕНИЯ И УМНОЖЕНИЯ ВЕРОЯТНОСТЕЙ

§ 3.1. Теорема сложения вероятностей

Теория вероятностей позволяет определять вероятность события по известным вероятностям других событий, если последние связаны с первым. В этом случае используют теоремы сложения и умножения вероятностей. Эти теоремы дают возможность найти вероятность появления одного из нескольких случайных событий или вероятность совместного поступления двух событий и более. Прежде чем рассматривать теорему сложения вероятностей, решим следующую задачу.

Задача 3.1. В ящике 12 белых, 7 черных и 11 синих одинаковых на ощупь шаров. Наудачу вынимается один шар. Какова вероятность того, что вынутый шар не белый?

△ Количество всех шаров в ящике равно 30, т. е. число N всех исходов испытания, заключающегося в вынимании одного шара, равно 30. Если вынутый шар не белый, то это означает, что он либо черный, либо синий. Вынуть либо черный, либо синий шар по правилу сложения можно $7+11=18$ способами. Следовательно, число M исходов, благоприятствующих событию C, которое состоит в вынимании не белого шара, равно 18. Тогда

$$P(C)=M/N=18/30=3/5. \ \blacktriangle$$

Пусть теперь событие A — «появится черный шар», а событие B — «появится синий шар». Найдем вероятности этих событий. Имеем $P(A)=7/30$, $P(B)=11/30$. Сложив полученные вероятности, получим $P(A)+P(B)=7/30+11/30=18/30=3/5=P(C)$. Заметим, что событие C является суммой событий A и B, т. е. $C=A+B$. Полученный результат обобщим в виде следующей теоремы.

Теорема. *Вероятность появления одного из двух несовместных событий, неважно какого, равна сумме вероятностей этих событий, т. е.*

$$P(A+B)=P(A)+P(B). \tag{3.1}$$

□ Введем следующие обозначения: N — общее число возможных исходов испытания; M — число исходов, благоприятствующих событию A; L — число исходов, благоприятствующих событию B. Тогда $P(A)=M/N$, $P(B)=L/N$. Так как события A и B несовместны, то нет таких исходов, которые были бы благоприятствующими и событию A, и событию B одновременно. Поэтому событию $A+B$ благоприятствуют $M+L$ исходов и

$$P(A+B)=(M+L)/N=M/N+L/N=P(A)+P(B). \ \blacksquare$$

Используя метод математической индукции, можно доказанную теорему обобщить на любое конечное число попарно несовместных событий.

Теорема. *Вероятность появления одного из нескольких попарно несовместных событий равна сумме вероятностей этих событий, т. е.*

$$P(A_1+A_2+A_3+\ldots+A_n)=P(A_1)+P(A_2)+P(A_3)+\ldots+P(A_n),$$

или

$$P\left(\sum_{i=1}^{n} A_i\right)=\sum_{i=1}^{n} P(A_i). \tag{3.2}$$

Задача 3.2. От коллектива бригады, которая состоит из 6 мужчин и 4 женщин, на профсоюзную конференцию выбирается два человека. Какова вероятность, что среди выбранных хотя бы одна женщина?

△ Обозначим через A событие, состоящее в том, что среди выбранных двух делегатов хотя бы одна женщина. Если произойдет событие A, то обязательно произойдет одно из следующих несовместных событий: B — «выбраны мужчина и женщина»; C — «выбраны две женщины». Поэтому можно записать: $A=B+C$. Найдем вероятности событий B и C. Два человека из 10 можно выбрать C_{10}^2 способами. Это и есть число всех исходов испытания (выбора двух человек). Двух женщин из 4 можно выбрать C_4^2 способами. Мужчину и женщину по правилу умножения можно выбрать $6\cdot 4$

способами. Тогда $P(B)=6\cdot 4/C_{10}^2$, $P(C)=C_4^2/C_{10}^2$. Так как события B и C несовместны, то, по теореме сложения,

$$P(A)=P(B+C)=P(B)+P(C)=8/15+2/15=2/3. \ \blacktriangle$$

При решении рассмотренной задачи не рассматривалось событие D — «выбраны двое мужчин», вероятность которого $P(D)$ равна $C_6^2/C_{10}^2=1/3$. Теперь можно заметить, что события B, C, D образуют полную группу, события D и A — противоположные, а $P(B)+P(C)+P(D)=1$.

Обобщим и докажем полученные результаты в виде следствий теоремы сложения. Эти действия помогают упростить решение многих вероятностных задач.

Следствие 1. *Сумма вероятностей попарно несовместных событий A_1, A_2, A_3, ..., A_n, образующих полную группу, равна 1.*

□ Рассмотрим событие $E=A_1+A_2+A_3+\ldots+A_n$. Оно заключается в том, что в результате опыта произойдет одно из событий $A_1+A_2+A_3+\ldots+A_n$, неважно какое. Так как указанные события образуют полную группу, то в результате опыта всегда наступит событие E, т. е. оно является достоверным (см. § 2.1). Тогда $P(E)=1$ (свойство 1^0 вероятности). События, образующие полную группу, попарно несовместны, поэтому можно применить теорему сложения. Имеем

$$P(E)=P(A_1+A_2+A_3+\ldots+A_n)=$$
$$P(A_1)+P(A_2)+P(A_3)+\ldots+P(A_n)=1. \ \blacksquare$$

Следствие 2. *Сумма вероятностей противоположных событий равна 1, т. е.*

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

□ Противоположные события несовместны и образуют полную группу, а сумма вероятностей таких событий равна 1 (следствие 1). ■

Для решения задач удобно использовать следствие 2, записанное в виде $P(A)=1-P(\bar{A})$. Это связано с тем, что часто бывает трудно вычислить вероятность некоторого события A, а вероятность противоположного события $\bar{A}$ вычисляется легко. В таких случаях сначала вычисляют вероятность противоположного события, а уже затем вероятность самого события, используя приведенную выше формулу. Проиллюстрируем применение следствий теоремы сложения следующими задачами.

Задача 3.3. Бросаются три игральных кубика (можно один кубик три раза). Какова вероятность того, что сумма выпавших очков меньше 17?

△ В результате бросания трех игральных кубиков могут появиться 16 различных сумм очков от 3 до 18, которые

образуют полную группу событий. Для решения задачи следует вычислить вероятности появления 14 сумм очков от 3 до 16, а затем сложить их. Это довольно трудоемкая операция. Поступим по-другому. Найдем вероятность выпадания 17 и 18 очков. При бросании трех кубиков возможно всего $N=6\cdot 6\cdot 6=216$ исходов; 18 очков могут выпасть только в одном случае, когда на всех кубиках 6 очков. Поэтому вероятность выпадания 18 очков равна 1/216. В трех случаях могут выпасть 17 очков. Это произойдет тогда, когда на одном из кубиков 5 очков, а на остальных по 6 очков. Следовательно, вероятность выпадания 17 очков равна 3/216. События «сумма выпавших очков меньше 17» и «сумма выпавших очков больше 16» являются противоположными. Обозначим их A и $\bar{A}$ соответственно. Тогда, по теореме сложения вероятностей, $P(\bar{A})=3/216+1/216=1/54$. Откуда

$$P(A)=1-P(\bar{A})=1-1/54=53/54. \ \blacktriangle$$

Задача 3.4. Среди одинаковых по внешнему виду 11 изделий находятся три бракованных. Произвольно вынимают три изделия. Найти вероятность того, что среди них хотя бы одно бракованное.

△ События «среди вынутых трех изделий хотя бы одно бракованное» и «среди вынутых трех изделий нет бракованных» являются противоположными. Обозначим их A и $\bar{A}$ соответственно. Найдем вероятность события $\bar{A}$. Число всех исходов испытания равно C_{11}^3, а число исходов, благоприятствующих событию $\bar{A}$, равно C_8^3 (из 11 изделий по условию 8 стандартных). Следовательно,

$$P(\bar{A})=C_8^3/C_{11}^3=56/165,$$

откуда

$$P(A)=1-P(\bar{A})=1-56/165=109/165. \ \blacktriangle$$

Рассмотренная выше теорема сложения вероятностей при вычислении вероятности суммы событий предусматривала их несовместность. Если же события совместны, то формула вероятности их суммы будет иной. Убедимся в этом при решении следующей задачи.

Задача 3.5. Отдел технического контроля проверяет на стандартность по двум параметрам серию изделий. Было установлено, что у 8 из 25 изделий не выдержан только первый параметр, у 6 изделий — только второй, а у 3 изделий не выдержаны оба параметра. Наудачу берется одно из изделий. Какова вероятность того, что оно не удовлетворяет стандарту?

△ Рассмотрим следующие события: A — «у изделия не выдержан первый параметр»; B — «у изделия не выдержан второй параметр»; C — «изделие не удовлетворяет стандарту». Событию AB, состоящему в том, что у взятой детали не выдержаны оба параметра, благоприятствуют три исхода. Событию A благоприятствуют $8+3=11$ исходов, а событию B благоприятствуют $6+3=9$ исходов. Число нестандартных изделий равно $8+6+3=11+9-3=17$. Следовательно, событию C благоприятствуют 17 исходов. Тогда

$$P(C)=17/25=(11+9-3)/25=11/25+9/25-3/25.$$

С другой стороны, $P(A)=11/25$, $P(B)=9/25$, $P(AB)=3/25$. Поэтому можно записать

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

Наступление события C означает, что у взятого наудачу изделия либо не выдержан первый параметр, либо второй, либо оба вместе, т. е. $C=A+B$ (см. определение суммы событий). Таким образом, можно записать

$$P(C)=P(A+B)=P(A)+P(B)-P(AB)=17/25. \ \blacktriangle$$

Докажем справедливость последней формулы для общего случая.

Теорема. *Вероятность появления хотя бы одного из двух совместных событий равна сумме вероятностей этих событий без вероятности их совместного наступления, т. е.*

$$P(A+B)=P(A)+P(B)-P(AB). \qquad (3.3)$$

□ Докажем теорему для схемы случаев. Пусть в результате опыта возможны N равновозможных исходов, которые для наглядности изобразим в виде последовательности кружков (рис. 3.1). Пусть, далее, событию A благоприятствуют M исходов (им соответствуют кружки с горизонтальной чертой), а событию B — K исходов (им соответствуют кружки с вертикальной чертой). События A и B совместны, поэтому часть указанных исходов благоприятствуют одновременно и событию A, и событию B (им соответствуют кружки с крестиками). Предположим, что количество таких исходов равно L. Тогда

$$P(A)=M/N, \quad P(B)=K/N, \quad P(AB)=L/N. \qquad (3.4)$$

Событие $A+B$ заключается в наступлении либо события A, либо события B, либо события AB. Поэтому ему будут

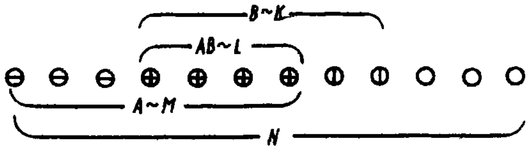


Рис. 3.1

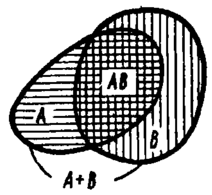


Рис. 3.2

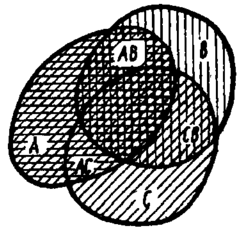


Рис. 3.3

благоприятствовать $M+K-L$ исходов (им соответствуют кружки либо с вертикальными, либо с горизонтальными линиями, либо с крестиками). Следовательно,

$$P(A+B)=(M+K-L)/N=M/N+K/N-L/N. \qquad (3.5)$$

Подставляя равенства (3.4) в равенство (3.5), получим равенство (3.3). ■

Доказанная теорема хорошо иллюстрируется диаграммой Венна. Здесь следует помнить, что вероятность события пропорциональна площади фигуры, которая соответствует данному событию. Событию $A+B$ на рис. 3.2 соответствует фигура $A+B$, которая обведена жирной линией. Событию A соответствует фигура A, а событию B — фигура B. Фигура, имеющая двойную штриховку, соответствует событию AB. Из рис. 3.2 следует, что площади этих фигур связаны соотношением $S_{A+B}=S_A+S_B-S_{AB}$. Данное соотношение соответствует формуле (3.3).

Рассуждая аналогично и используя рис. 3.3, можно доказать справедливость формулы трех совместных в совокупности событий:

$$P(A+B+C)=P(A)+P(B)+P(C)-P(AB)-P(AC)-\\ -P(CB)+P(ABC). \qquad (3.6)$$

При доказательстве последней теоремы предполагалась совместность событий. Если же события A и B несовместны, то событие AB будет невозможным, а значит, его вероятность равна нулю. Поэтому формулы (3.1) и (3.3) совпадут. Таким образом, формулы (3.3) и (3.6) верны как для совместных, так и для несовместных событий.

При вычислении вероятности суммы совместных событий требуется, как видно из формул, уметь находить вероятность произведений событий. В простейших случаях (см. задачу 3.5) это не представляет трудности. Однако при решении более сложных задач нахождение вероятности произведения нескольких событий затруднительно. Нахождению вероятности произведения событий по известным вероятностям этих событий и посвящен следующий параграф.

§ 3.2. Условная вероятность. Теорема умножения вероятностей

Вероятность произведения двух событий будем искать в виде произведения вероятностей, что вполне естественно, так как вероятность суммы событий выражается через сумму их вероятностей, а если эти события несовместны, то она равна сумме их вероятностей.

Пусть бросается игральный кубик. В результате этого испытания может наступить событие A, состоящее в выпадании четного числа очков, и может наступить событие B, состоящее в выпадании числа очков, меньшего 6. Эти события совместны, поэтому имеет смысл говорить о вероятности $P(AB)$ их совместного появления.

Найдем вероятности событий A, B и AB. Событию A благоприятствуют три исхода — A_2, A_4, A_6, событию B благоприятствуют пять исходов — A_1, A_2, A_3, A_4, A_5, а событию AB благоприятствуют два исхода — A_2, A_4 (см. пример 2.19). Поэтому $P(A)=3/6$, $P(B)=5/6$, $P(AB)=2/6$. Представим теперь вероятность $P(AB)$ в виде произведения, например вероятности $P(A)$ и неизвестной вероятности x, т. е. $P(AB)=P(A)x$. В данном случае нет оснований считать, что $x=P(B)$, так как $P(A)\cdot P(B)=5/12\neq P(AB)$. Найдем число x и выясним его вероятностный смысл. Имеем $x=P(AB)/P(A)=(1/3)/(1/2)=2/3$. Нетрудно заметить, что число 3 в знаменателе полученной дроби совпадает с числом исходов, благоприятствующих событию A, а два из них (это число стоит в числителе дроби) благоприятствуют событию B. Поэтому число $x=2/3$ соответствует вероятности события B, вычисленной с учетом только тех исходов испытания, которые благоприятствуют наступлению события A. Найденную при таких условиях вероятность события B назовем *условной* и обозначим $P_A(B)$. Таким образом, в данном случае $P_A(B)=2/3$, откуда $P(AB)=P(A)\cdot P_A(B)$.

С другой стороны, событию B благоприятствуют пять исходов — A_1, A_2, A_3, A_4, A_5. Из них событию A благоприятствуют два исхода — A_2, A_4. Тогда $P_B(A)=2/5$ и, следовательно, выполняется равенство $P(AB)=P(B)P_B(A)$.

Прежде чем обобщить полученные результаты в виде теоремы, дадим определение условной вероятности. В приведенном примере события A и B могут наступить в результате одного испытания, поэтому условная вероятность $P_A(B)$ в таких случаях показывает, как часто появляется событие B среди тех исходов, в которых появляется событие A. В других случаях события A и B могут наступать в разных испытаниях, условия которых не обязательно должны совпадать. Здесь вероятность одного события также может зависеть от наступления другого.

Определение. *Условной вероятностью $P_A(B)$ называется вероятность события B, вычисленная в предположении, что событие A уже наступило.*

Условную вероятность события B, вычисленную в предположении, что событие A уже наступило, часто обозначают так: $P(B/A)$.

○ **Пример 3.1.** В ящике 7 одинаковых на ощупь шаров с номерами от 1 до 7. Наудачу один за другим берут два шара, не возвращая их обратно.

Рассмотрим события: A — «первый вынутый шар имеет номер 3»; B — «второй вынутый шар имеет нечетный номер». Найдем условную вероятность $P_A(B)$. Если первый раз вынут шар под номером 3, то в ящике осталось 6 шаров, из которых 3 имеют нечетные номера (1, 5, 7). Следовательно, $P_A(B)=3/6=1/2$. ●

Если же вынутый шар в предыдущем примере возвращается назад в ящик, то условия второго испытания остаются неизменными после проведения первого испытания. Тогда $P(B)=P_A(B)=4/7$, т. е. в этом случае вероятность события B и его условная вероятность совпадают.

Рассмотрим теперь оба испытания как одно сложное. Запишем все исходы, благоприятствующие наступлению события A, в виде строки, составленной из двузначных чисел, где 1-я цифра означает номер шара, вынутого первым, а 2-я — номер шара, вынутого вторым. Имеем: 31; 32; 34; 35; 36; 37. Из этих шести исходов событию B благоприятствуют три исхода: 31; 35; 37. Поэтому $P_A(B)=3/6=1/2$. Таким образом, получен тот же самый результат.

Теорема. *Вероятность совместного появления двух событий равна произведению вероятности одного из них на условную вероятность другого, вычисленную в предположении, что первое событие уже наступило:*

$$P(AB)=P(A)\cdot P_A(B). \tag{3.7}$$

□ Докажем теорему для схемы случаев (см. рис. 3.1). Пусть в результате опыта возможны N исходов, M из которых благоприятствуют событию A, а K — событию B. Пусть, далее, L исходов благоприятствуют одновременному наступлению событий A и B. Тогда $P(A)=M/N$, $P(AB)=L/N$. Так как событию A благоприятствуют M исходов и только L из них благоприятствуют событию B, то условная вероятность $P_A(B)=L/M$. Полагая, что событие A может произойти в результате опыта, а это означает, что $M\neq 0$, получаем

$$P(AB)=L/N=(M/N)/(L/M)=P(A)P_A(B). \ ■$$

Задача 3.6. В коробке девять одинаковых радиоламп, три из которых были в употреблении. В течение рабочего дня мастеру для ремонта аппаратуры пришлось взять две радиолампы. Какова вероятность того, что обе взятые лампы были в употреблении?

△ Вероятность того, что первая взятая радиолампа была в употреблении (событие A), равна $P(A)=3/9$. После того как произошло событие A, в коробке осталось восемь радиоламп, из которых две были в употреблении. Поэтому для события B, состоящего в появлении второй раз радиолампы, бывшей в употреблении, условная вероятность $P_A(B)=2/8$. Следовательно, вероятность появления двух таких ламп

$$P(AB)=P(A)P_A(B)=3/9\cdot 2/8=1/12.$$

Заметим, что задачу можно решить, если воспользоваться комбинаторикой. В том случае ответом на вопрос задачи является число $C_3^2/C_9^2=1/12$. ▲

Задача 3.7. Бросается игральный кубик. Какова вероятность того, что выпадет четное и меньшее 5 число очков?

△ Подсчитаем искомую вероятность непосредственно. Она равна $1/3$. Получим этот же результат, используя теорему умножения. Под событием A будем понимать появление четного числа очков, а под событием B — появление числа очков, меньшего 5. Событию A благоприятствуют исходы A_2, A_4, A_6, а событию B — исходы A_1, A_2, A_3, A_4. Поэтому $P(A)=3/6$, $P(B)=4/6=2/3$. Из трех исходов, благоприятствующих событию A, исходы A_2, A_4 благоприятствуют событию B. Следовательно, $P_A(B)=2/3$. Тогда, по теореме умножения,

$$P(AB)=P(A)P_A(B)=3/6\cdot 2/3=1/3. \ ▲$$

В последней задаче наблюдается равенство вероятностей $P(B)$ и $P_A(B)$. Как видно из предыдущего (см. также пример 3.1), в одних случаях вероятность события и его условная вероятность совпадают, а в других различаются, причем совпадение или различие этих вероятностей наблюдается и для событий, которые могут наступать в одном испытании, и для событий, которые могут наступать в разных испытаниях. Другими словами, в одних случаях наступление события A изменяет вероятность наступления события B, а в других — нет. Если наступление события A изменяет вероятность наступления события B, то говорят о зависимости событий A и B.

Определение. *Два события называются независимыми, если появление любого из них не изменяет вероятность появления другого, т. е.*

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

Теорема. *Вероятность совместного появления двух независимых событий равна произведению их вероятностей, т. е.*

$$P(AB)=P(A)P(B). \qquad (3.8)$$

□ Пусть события A и B независимы; тогда должно выполняться равенство $P_A(B)=P(B)$. Подставляя это равенство в формулу $P(AB)=P(A)P_A(B)$, получаем

$$P(AB)=P(A)P(B). \ ■$$

Нетрудно доказать обратное утверждение.

Теорема. *Если для двух событий выполняется равенство* (3.8), *то эти события независимые.*

□ Пусть для событий A и B выполняется равенство (3.8). По теореме умножения вероятностей, $P(AB)=P(A)P_A(B)$. Тогда $P(AB)=P(A)P(B)=P(A)P_A(B)$, откуда $P(B)=P_A(B)$. Следовательно, события A и B независимые. ■

Если события наступают в разных испытаниях и условия этих испытаний не влияют друг на друга, т. е. наступление одного события не изменяет условий второго испытания, то события независимы.

Задача 3.8. Бросают два игральных кубика. Какова вероятность того, что на первом кубике выпадет четное число очков, а на втором — число, меньшее 6?

△ Рассмотрим события: A — «на первом кубике выпало четное число очков»; B — «на втором кубике выпало число очков, меньшее 6». Для решения задачи необходимо найти $P(AB)$. Исходя из условий опыта можно сделать вывод, что события A и B независимы. Действительно, событие A наступает в первом испытании, а событие B — во втором. Испытания проводятся в одинаковых условиях, и наступление одного события не изменяет условий испытания, в результате которого может произойти другое. Очевидно, что $P(A)=3/6$, а $P(B)=5/6$. Тогда по теореме о вероятности произведения независимых событий находим

$$P(AB)=P(A)P(B)=3/6\cdot 5/6=5/12. \ ▲$$

Проверим теперь предположение о независимости событий A и B. Будем бросание двух кубиков рассматривать как одно сложное испытание. При бросании двух кубиков возможны $N=6\cdot 6=36$ исходов сложного испытания. Поставим в соответствие каждому такому исходу двузначное число, где первая цифра — число выпавших очков на первом кубике, а вторая цифра — на втором. Событию A благоприятствуют 18 исходов, так как каждому четному числу (2, 4, 6) очков, выпавших на первом кубике, будут соответствовать шесть двузначных чисел (исходов), где вторая цифра принимает значения от 1 до 6, т. е. $M=3\cdot 6=18$. Из каждых таких шести двузначных чисел пять будут благоприятствовать наступлению события B (последняя цифра должна принимать значения от 1 до 5). Таким образом, из 18 исходов, благоприятствующих наступлению события A, $L=3\cdot 5=15$ исходов благоприятствуют наступлению события B. Выпишем для наглядности все эти исходы: 21; 22; 23; 24; 25; 41; 42; 43; 44; 45; 61; 62; 63; 64; 65. Итак, $P(A)=M/N=18/36$, $P_A(B)=L/M=15/18$, а $P(AB)=P(A)P_A(B)=5/12$. Результаты совпали, значит, предположение о независимости событий было верным.

Задача 3.9. Вероятность поломки первого станка в течение смены равна 0,2, а второго — 0,13. Чему равна вероятность того, что оба станка потребуют наладки в течение смены?

△ Станки работают независимо друг от друга, поэтому событие (поломка первого станка) и событие B (поломка второго станка) независимы. Тогда

$$P(AB)=P(A)P(B)=0{,}2\cdot 0{,}13=0{,}026. \ ▲$$

Задача 3.10. Два спортсмена независимо друг от друга стреляют по одной мишени. Вероятность попадания в мишень первого спортсмена равна 0,7, а второго — 0,8. Какова вероятность того, что мишень будет поражена?

△ Мишень будет поражена, если в нее попадет либо первый стрелок, либо второй, либо оба вместе, т. е. произойдет событие $A+B$, где событие A заключается в попадании в мишень первым спортсменом, а событие B — вторым. Тогда

$$P(A+B)=P(A)+P(B)-P(AB)=0{,}7+0{,}8-0{,}7\cdot 0{,}8=0{,}94. \ ▲$$

Независимые события обладают рядом свойств. Докажем два из них.

1°. *Если события A и B независимы, то независимы события A и $\bar{B}$.*

□ Пусть события A и B независимы. Воспользуемся диаграммой Венна. Из рис. 3.4 видно, что $A=A\bar{B}+AB$. Очевидно, что события $A\bar{B}$ и AB несовместны, поэтому к ним можно применить теорему сложения вероятностей для несовместных событий. Имеем

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

откуда

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

или

$$P(A\bar{B})=P(A)P(\bar{B}),$$

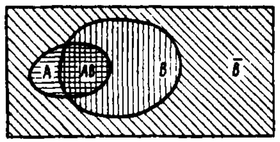


Рис. 3.4

т. е. события A и $\bar{B}$ независимы. ■

2°. *Если два события независимы, то независимы и противоположные им события.*

□ Это утверждение непосредственно следует из предыдущего. Приведем также другое доказательство. Пусть события A и B независимы. Наступление события $\bar{A}\bar{B}$ означает, что ни одно из событий A и B не наступило в результате опыта, а событие $A+B$ заключается в том, что наступило хотя бы одно из событий A и B. Следовательно, эти события противоположны. Тогда имеем

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

откуда в силу независимости событий A и B

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

Разложим на множители правую часть последнего равенства. В результате получим

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

что и означает независимость событий $\bar{A}$ и $\bar{B}$. ■

Теперь докажем теорему умножения вероятностей для произвольного числа событий.

Теорема. *Вероятность совместного наступления конечного числа событий равна произведению вероятности одного из них на условные вероятности всех остальных, причем условная вероятность каждого последующего события вычисляется в предположении, что все предыдущие уже наступили, т. е.*

$$P(A_1A_2A_3\ldots A_n)=P(A_1)P_{A_1}(A_2)P_{A_1A_2}(A_3)\ldots P_{A_1A_2A_3\ldots A_{n-1}}(A_n), \quad (3.9)$$

где $P_{A_1A_2A_3\ldots A_{n-1}}(A_n)$ — вероятность появления события A_n, вычисленная в предположении, что события A_1, A_2, A_3, ..., A_{n-1} произошли.

□ При доказательстве воспользуемся методом математической индукции. Для $n=2$ теорема уже доказана. Предположим, что она верна для $n-1$ событий, т. е.

$$P(A_1A_2A_3\ldots A_{n-1})=P(A_1)P_{A_1}(A_2)P_{A_1A_2}(A_3)\ldots P_{A_1A_2A_3\ldots A_{n-2}}(A_{n-1}) \quad (3.10)$$

Обозначим через B произведение событий A_1, A_2, A_3, ..., A_{n-1}, т. е. $B=A_1A_2A_3\ldots A_{n-1}$. Тогда, согласно теореме умножения вероятностей, для двух событий имеем

$$P(A_1A_2A_3\ldots A_{n-1}A_n)=P(B\cdot A_n)=P(B)P_B(A_n).$$

Так как, по предположению, верно равенство (3.10), то

$$P(A_1A_2A_3\ldots A_{n-1}A_n)=P(A_1)P_{A_1}(A_2)P_{A_1A_2}(A_3)\ldots P_{A_1A_2A_3\ldots A_{n-1}}(A_n). \ ■$$

Отметим, что порядок входящих в произведение событий не играет роли. В частности, для трех событий формула (3.9) принимает вид

$$P(ABC)=P(A)P_A(B)P_{AB}(C).$$

Задача 3.11. В гараж поступили 24 новые шины, предназначенные для определенной марки автомобиля. Шины имеют одинаковый внешний вид. Изготовлены они на двух различных заводах, причем 10 шин изготовлено на 1-м заводе, а остальные — на 2-м. Четырем водителям необходимо заменить по одной шине. Какова вероятность того, что первые три водителя воспользуются шинами 2-го завода, а четвертый — шиной 1-го завода?

△ Рассмотрим события: A — «первый водитель взял шину 2-го завода»; B — «второй водитель взял шину 2-го завода»; C — «третий водитель взял шину 2-го завода»; D — «четвертый водитель взял шину 1-го завода». Очевидно, вероятность события A равна $P(A)=14/24$. Если первый водитель взял шину 2-го завода, то осталось 23 шины, среди которых осталось 13 шин 2-го завода. Тогда $P_A(B)=13/23$. Если и второй водитель взял шину 2-го завода, то всех шин после этого осталось 22, из которых 12 шин 2-го завода. Следовательно, $P(C)=12/22$. Таким образом, после наступления событий A, B, C осталась 21 шина, среди которых 10 шин 1-го завода. Поэтому $P_{ABC}(D)=10/21$ и вероятность произведения событий A, B, C, D, по теореме умножения вероятностей для произвольного числа событий,

$$P(ABCD)=P(A)P_A(B)P_{AB}(C)P_{ABC}(D)=$$
$$=(14/24)(13/23)(12/22)(10/21)=65/759. \ \blacktriangle$$

Прежде чем доказать теорему умножения вероятностей для произвольного числа независимых событий, еще раз остановимся на понятии независимости событий. Из предыдущего изложения известно, что если появление одного из двух событий не изменяет вероятность появления другого, то такие события независимы. Если же вероятность появления одного события нарушается после появления другого, то такие события будем называть *зависимыми*. Другими словами, если для событий A и B справедливо неравенство $P(B)\neq P_A(B)$, а значит, и неравенство $P(AB)\neq P(A)P(B)$, то эти события зависимы.

Возникает вопрос: что понимать под независимостью трех (и более) событий? Рассмотрим для примера случай трех событий. Понятие независимости двух событий было связано с их произведением. По аналогии воспользуемся произведением событий и в этом случае. Произведение трех событий A, B, C можно представить в виде произведений двух событий AB и C, или AC и B, или BC и A, а для двух событий понятие независимости уже было определено. Естественно считать события независимыми или независимыми в совокупности, если независимы пары событий AB и C, AC и B, BC и A.

Достаточно ли для независимости в совокупности трех событий их попарной независимости? Приведенный ниже пример опровергает эту гипотезу.

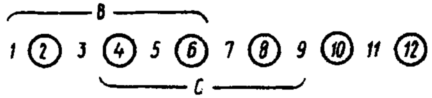


Рис. 3.5

○ **Пример 3.2.** Пусть на карточках, которые тщательно перемешаны, нанесены числа от 1 до 12. Вынимается наудачу одна карточка и рассматриваются следующие события: A — «появилось четное число»; B — «появилось число, меньшее 7»; C — «появилось число, большее 3 и меньшее 10». Для наглядности изобразим возможные исходы опыта на рис. 3.5. Нетрудно видеть, что $P(A)=P(B)=P(C)=1/2$. Событие B может наступить в шести случаях при появлении чисел 1, 2, 3, 4, 5, 6, из которых три случая благоприятствуют появлению события A, а именно появлению чисел 2, 4, 6. Таким образом, $P_B(A)=1/2=P(A)$. Последнее равенство означает, что события A и B независимы. Рассуждая аналогично, можно доказать, что события A и C или C и B также независимы, т. е. события A, B, C попарно независимы.

Пусть теперь произошло событие BC. Оно может наступить, если появится одно из трех чисел 4, 5, 6, но появление только двух из них (4 и 6) благоприятствует событию A. Поэтому $P_{BC}(A)=2/3\neq 1/2$. Следовательно, события A и CB зависимы. Нетрудно показать, что $P_{CA}(B)=2/3\neq 1/2$, $P_{BA}(C)=2/3\neq 1/2$. Из последних соотношений следует зависимость событий CA и B, BA и C. Итак, на примере было показано, что из попарной зависимости трех событий не следует их зависимость в совокупности, поскольку вероятность одного события может измениться после наступления остальных. ●

На основании предыдущих рассуждений введем следующее определение.

Определение. *События называются независимыми в совокупности (или просто независимыми), если наряду с их попарной независимостью независимы любое из них и произведение любого числа из остальных, в противном случае события называются зависимыми.*

Из определения следует, что события независимы в совокупности, если вероятность появления одного из них не изменяется при появлении каких-либо других из оставшихся событий.

Теорема. *Вероятность совместного появления нескольких событий, независимых в совокупности, равна произведению вероятностей этих событий, т. е.*

$$P(A_1A_2A_3\dots A_n)=P(A_1)P(A_2)P(A_3)\dots P(A_n). \qquad (3.11)$$

Доказательство теоремы предлагается провести самостоятельно.

Заметим, что для независимости в совокупности нескольких (больше двух) событий недостаточно выполнения равенства (3.11).

○ **Пример 3.3.** На карточках, которые тщательно перемешаны, нанесены числа от 1 до 8. Наудачу вынимается одна из карточек. Рассмотрим следующие события: A — «появится число, меньшее 5»; B — «появится число, большее 1, но меньшее 6»; C — «появится число, большее 3, но меньшее 8». Очевидно, что $P(A)=P(B)=P(C)=1/2$ и $P(A)P(B)P(C)=1/8$. Событию ABC благоприятствует только один исход — появление числа 4. Следовательно, $P(ABC)=1/8$ и $P(ABC)=P(A)P(B)P(C)$. Далее находим: $P_A(B)=3/4$, $P_{AB}(C)=1/3$, т. е. $P(B)\neq P_A(B)$, $P(C)\neq P_{AB}(C)$. Это говорит о том, что события A, B, C являются зависимыми, несмотря на выполнение равенства (3.11). ●

В дальнейшем понадобится ответ на вопрос: сохранится ли независимость в совокупности нескольких событий, если одно из них заменить ему противоположным? Для двух событий такая замена не изменяет их независимость, что было показано ранее (см. свойства 1^0 и 2^0 независимых событий). Убедимся теперь, что таким же свойством обладает и группа, состоящая из трех событий.

Пусть события A, B, C независимы в совокупности. Тогда такие пары событий, как, например, B и C, A и C, A и B, A и BC, будут независимыми, а значит, независимы и пары событий C и $\bar{A}$, B и $\bar{A}$, $\bar{A}$ и BC (см. свойство 1^0 независимых событий). Следовательно, имеют место равенства $P(C)=P_B(C)$; $P(\bar{A})=P_C(\bar{A})$; $P(B)=P_{\bar{A}}(B)$; $P(\bar{A})=P_{BC}(\bar{A})$. По теореме умножения вероятностей, для событий $\bar{A}$, B, C имеем

$$P(\bar{A}BC)=P(B)P_B(C)P_{BC}(\bar{A})=P(B)P(C)P(\bar{A});$$

$$P(\bar{A}BC)=P(C)P_C(\bar{A})P_{C\bar{A}}(B)=P(C)P(\bar{A})P_{C\bar{A}}(B);$$

$$P(\bar{A}BC)=P(\bar{A})P_{\bar{A}}(B)P_{\bar{A}B}(C)=P(\bar{A})P(B)P_{\bar{A}B}(C).$$

Сравнивая полученные равенства заключаем, что $P_{C\bar{A}}(B)=P(B)$, $P_{\bar{A}B}(C)=P(C)$. Последние равенства означают независимость событий B и $C\bar{A}$, C и $\bar{A}B$. Таким образом, все условия независимости событий $\bar{A}$, B, C выполнены. Из независимости событий $\bar{A}$, B, C сразу же следует независимость событий $\bar{A}$, $\bar{B}$, C и $\bar{A}$, $\bar{B}$, $\bar{C}$.

Итак, если три события независимы в совокупности, то эта независимость сохраняется и при замене любого числа из них на противоположные. Доказательство этого утверждения можно легко распространить на большее число событий. Рассмотренное свойство независимых событий часто используют на практике (см., например, задачу 3.12 и вывод формулы Бернулли в § 4.1).

При решении многих задач событие, вероятность которого следует найти, приходится представлять в виде суммы нескольких событий. Такое представление может быть громоздким и сопряжено с большими вычислениями. В том случае, когда противоположное событие представляется в виде суммы меньшего числа событий, имеет смысл воспользоваться вычислением вероятности противоположного события (см. задачи 3.3 и 3.4).

Задача 3.12. В трех залах кинотеатра идут три различных фильма. Вероятность того, что на определенный час в кассе 1-го зала есть билеты, равна 0,3, в кассе 2-го зала — 0,2,

а в кассе 3-го зала — 0,4. Какова вероятность того, что на данный час имеется возможность купить билет хотя бы на один фильм?

△ Обозначив A — «билеты есть хотя бы в одной кассе»; A_i — «билеты есть в i-й кассе, $i=1, 2, 3$», можно записать $A=A_1\bar{A}_2\bar{A}_3+\bar{A}_1A_2\bar{A}_3+\bar{A}_1\bar{A}_2A_3+\bar{A}_1A_2A_3+A_1\bar{A}_2A_3+A_1A_2\bar{A}_3+$ $+A_1A_2A_3$ (см. пример 2.20).

Выражение для A в данном случае довольно громоздкое, оно стало бы еще сложнее, если бы в кинотеатре было четыре зала. Такой путь решения из-за своей сложности нецелесообразен. Поэтому рассмотрим событие $\bar{A}$, противоположное A. Оно наступит тогда, когда наступит событие $\bar{A}_1\bar{A}_2\bar{A}_3$. Таким образом, $\bar{A}=\bar{A}_1\bar{A}_2\bar{A}_3$, откуда

$$P(A)=1-P(\bar{A})=1-P(\bar{A}_1\bar{A}_2\bar{A}_3)=1-P(\bar{A}_1)P(\bar{A}_2)P(\bar{A}_3)=$$
$$=1-0{,}7\cdot0{,}8\cdot0{,}6=0{,}664. \ ▲$$

При решении подобных задач следует использовать рассмотренный в задаче 3.12 способ, который обобщим в виде теоремы, приведенной ниже без доказательства.

Теорема. *Вероятность появления хотя бы одного из событий $A_1A_2A_3$, ..., A_n, независимых в совокупности, равна разности между единицей и произведением вероятностей противоположных событий $\bar{A}_1$, $\bar{A}_2$, $\bar{A}_3$, ..., $\bar{A}_n$, т. е.*

$$P(A_1+A_2+A_3+\ldots+A_n)=1-P(\bar{A}_1)P(\bar{A}_2)P(\bar{A}_3)\ldots P(\bar{A}_n). \quad (3.12)$$

§ 3.3. Формула полной вероятности. Формула Байеса

На практике часто возникают ситуации, когда требуется определить вероятность события, которое может произойти с одним из несовместных событий, образующих полную группу. Доказанная ниже теорема, являющаяся следствием теорем сложения и умножения вероятностей, допускает нахождение вероятности подобных событий и дает для этого формулу.

Теорема. *Вероятность события A, которое может наступить только при условии появления одного из событий H_1, H_2, H_3, ..., H_n, образующих полную группу попарно несовместных событий, равна сумме произведений вероятностей каждого из событий H_1, H_2, H_3, ..., H_n на соответствующую условную вероятность события A, т. е.*

$$P(A)=\sum_{i=1}^{n}P(H_i)P_{H_i}(A). \quad (3.13)$$

□ Согласно условию теоремы, событие A может наступить, если появится одно из событий H_1, H_2, H_3, ..., H_n. Это означает, что появление события A влечет появление одного из событий H_1A, H_2A, H_3A, ..., H_nA, неважно какого. Поэтому событие A можно представить в виде $A=H_1A+H_2A+$ $+H_3A+...+H_nA$. Так как события H_1, H_2, H_3, ..., H_n, по условию, попарно несовместны, то, очевидно, несовместны и события H_1A, H_2A, H_3A, ..., H_nA. Тогда, применяя теорему сложения для несовместных событий, получаем

$$P(A)=P(H_1A)+P(H_2A)+P(H_3A)+...+P(H_nA).$$

Далее, по теореме умножения вероятностей, можно записать

$$P(A)=P(H_1)P_{H_1}(A)+P(H_2)P_{H_2}(A)+P(H_3)P_{H_3}(A)+...$$
$$...+P(H_n)P_{H_n}(A)=\sum_{i=1}^{n}P(H_i)P_{H_i}(A). \blacksquare$$

Формула (3.13), в справедливости которой убеждает доказанная теорема, называется *формулой полной вероятности*, а события H_1, H_2, H_3, ..., H_n — *гипотезами.*

Задача 3.13. На трех станках различной марки изготовляется определенная деталь. Производительность 1-го станка за смену составляет 40 деталей, 2-го — 35 деталей, 3-го — 25 деталей. Установлено, что 2, 3 и 5% продукции этих станков соответственно имеют скрытые дефекты. В конце смены на контроль взята одна деталь. Какова вероятность, что она нестандартная?

Пусть A — событие, состоящее в том, что взятая наудачу деталь имеет дефект. Здесь возможны следующие три гипотезы:

1) деталь изготовлена на 1-м станке (гипотеза H_1);

2) деталь изготовлена на 2-м станке (гипотеза H_2);

3) деталь изготовлена на 3-м станке (гипотеза H_3).

Найдем вероятности гипотез: $P(H_1)=40/100$, $P(H_2)=35/100$, $P(H_3)=25/100$. Условные вероятности события A при этих гипотезах соответственно равны $P_{H_1}(A)=2/100$, $P_{H_2}(A)=3/100$, $P_{H_3}(A)=5/100$. По формуле полной вероятности имеем

$$P(A)=P(H_1)P_{H_1}(A)+P(H_2)P_{H_2}(A)+P(H_3)P_{H_3}(A)=$$
$$=40/100\cdot 2/100+35/100\cdot 3/100+25/100\cdot 5/100=0{,}031. \blacktriangle$$

Задача 3.14. Была проведена одна и та же контрольная работа в трех параллельных группах. В 1-й группе, где 30 учащихся, оказалось 8 работ, выполненных на «отлично»; во 2-й, где 28 учащихся,— 6 работ; в 3-й, где 27 учащихся,— 9 работ. Найти вероятность того, что первая взятая наудачу при повторной проверке работа из работ, принадлежащих группе, которая также выбрана наудачу, окажется выполненной на «отлично».

$\triangle$ Рассмотрим события: A — «взятая работа выполнена на «отлично»; H_i — «работа выполнена учащимся i-й группы, где $i=1, 2, 3$». Очевидно, $P(H_1)=P(H_2)=P(H_3)=1/3$ и $P_{H_1}(A)=8/30$, $P_{H_2}(A)=6/28$, $P_{H_3}(A)=9/27$. Применяя формулу полной вероятности, получаем

$$P(A)=P(H_1)P_{H_1}(A)+P(H_2)P_{H_2}(A)+P(H_3)P_{H_3}(A)=$$
$$=1/3\cdot 8/30+1/3\cdot 6/28+1/3\cdot 9/27=1/3\cdot(4/15+3/14+1/3)=$$
$$=57/210=19/70. \blacktriangle$$

В последней задаче искомая вероятность события равна среднему арифметическому условных вероятностей этих событий. Это имеет место потому, что равны вероятности гипотез.

При выводе формулы полной вероятности предполагалось, что событие A, вероятность которого следовало определить, могло произойти с одним из событий H_1, H_2, H_3, ..., H_n, образующих полную группу, при этом вероятности указанных событий (гипотез) были известны заранее. Предположим, что проведен опыт и событие A наступило. Установим, как изменятся после этого вероятности гипотез, т. е. найдем условную вероятность $P_A(H_i)$ для каждой гипотезы.

По теореме умножения имеем

$$P(AH_i)=P(A)P_A(H_i)=P(H_i)P_{H_i}(A),\quad i=1,\ 2,\ 3,\ ...,\ n,$$

или

$$P(A)P_A(H_i)=P(H_i)P_{H_i}(A).$$

Отсюда

$$P_A(H_i)=P(H_i)P_{H_i}(A)/P(A).$$

Воспользовавшись формулой (3.13), имеем

$$P_A(H_i)=P(H_i)P_{H_i}(A)\Big/\sum_{j=1}^{n}P(H_j)P_{H_j}(A),\quad j=1,\ 2,\ 3,\ ...,\ n. \qquad (3.14)$$

Полученная формула называется *формулой Байеса* или *теоремой гипотез*. Формула Байеса позволяет «пересмотреть» вероятности гипотез после того, как становится известным, что в результате опыта появилось событие A.

Задача 3.15. Две перфораторщицы набивают на одинаковых и исправных перфораторах перфокарты. Более опытная из них обрабатывает в среднем 60% перфокарт из предложенной партии, а менее опытная — 40%. Вероятность того, что опытная перфораторщица допустит ошибку при перфорировании одной перфокарты, равна 0,03, для менее опытной эта вероятность равна 0,05. Взятая на контроль перфокарта оказалась с ошибкой. Какова вероятность того, что ошиблась более опытная перфораторщица?

△ Пусть событие A состоит в том, что наудачу взятая перфокарта оказалась с ошибкой. Можно высказать два предположения: 1) перфокарту подготовила опытная перфораторщица (гипотеза H_1); 2) перфокарту подготовила менее опытная перфораторщица (гипотеза H_2). По условию задачи, $P(H_1)=0{,}6$, $P(H_2)=0{,}4$, $P_{H_1}(A)=0{,}03$, $P_{H_2}(A)=0{,}05$. Используя формулу Байеса, находим вероятность

$$P_A(H_1)=P(H_1)P_{H_1}(A)/(P(H_1)P_{H_1}(A)+P(H_2)P_{H_2}(A))=$$
$$=0{,}6\cdot 0{,}03/(0{,}6\cdot 0{,}03+0{,}4\cdot 0{,}05)=9/19. \ \blacktriangle$$

Формула Байеса имеет следующее применение. Пусть имеется несколько предположений (несовместных гипотез) для объяснения некоторого явления. Эти предложения проверяются с помощью опыта. Перед началом опыта (эксперимента) зачастую бывает трудно определить вероятности этих предположений — гипотез, которые обычно называют *доопытными (априорными)* вероятностями. Поэтому этим гипотезам приписывают из интуитивных или каких-либо других соображений определенные вероятности. Затем проводят эксперимент и получают первую информацию, на основании которой выполняют коррекцию доопытных вероятностей.

Таким образом, основываясь на результатах опыта, заменяют доопытные вероятности *послеопытными (апостериорными)* вероятностями. При этом вероятности гипотез после опыта могут измениться. Вероятности некоторых гипотез могут настолько уменьшиться, что в дальнейшем ими вообще можно пренебречь, что, например, имеет место при решении следующей задачи. Эксперимент можно продолжать далее (повторить опыт), в результате по мере получения новой информации будет укрепляться предположение о справедливости той или иной гипотезы.

В настоящее время с внедрением совершенной вычислительной техники практически во все сферы деятельности человека формула Байеса находит все более широкое применение при решении проблем управления в экономике и промышленности, связанных с недостаточной информацией. По мере поступления информации и ее накопления проводится корректировка различных решений и планов.

Задача 3.16. Электронный прибор содержит две микросхемы. Вероятность выхода первой микросхемы в течение определенного (достаточно большого) времени равна 0,2, а второй — 0,1. Известно, что из строя вышла одна микросхема. Какова вероятность того, что это первая микросхема?

△ Пусть событие A состоит в том, что по истечении определенного срока вышла из строя одна микросхема. Эксперимент (наблюдение над прибором в течение указанного

срока) может закончиться одним из следующих несовместных исходов: H_1 — «обе микросхемы выдержали испытание»; H_2 — «отказала только первая микросхема»; H_3 — «отказала только вторая микросхема»; H_4 — «отказали обе микросхемы». Других исходов, очевидно, быть не может. Следовательно, указанные исходы образуют полную группу событий.

Можно высказать предположение, что эксперимент закончится одним из этих исходов, поэтому исходы H_1, H_2, H_3, H_4 будем считать гипотезами. Используя теорему умножения вероятностей и понятие противоположных событий, имеем $P(H_1)=0{,}8\cdot 0{,}9=0{,}72$; $P(H_2)=0{,}2\cdot 0{,}9=0{,}18$; $P(H_3)=0{,}8\cdot 0{,}1=0{,}08$; $P(H_4)=0{,}2\cdot 0{,}1=0{,}02$. Условные вероятности события A при этих гипотезах соответственно равны: $P_{H_1}(A)=0$; $P_{H_2}(A)=1$; $P_{H_3}(A)=1$; $P_{H_4}(A)=0$.

Применяя теперь формулу Байеса, получим

$$P_A(H_2)=P(H_2)P_{H_2}(A)\Big/\sum_{i=1}^{4}P(H_i)P_{H_i}(A)=$$
$$=0{,}18\cdot 1/(0{,}18\cdot 1+0{,}08\cdot 1)=9/13.$$

Следовательно, вероятность того, что откажет только первая микросхема, равна 9/13. Рассуждая аналогично, получим $P_A(H_1)=0$, $P_A(H_3)=4/13$, $P_A(H_4)=0$. Оказалось, что две послеопытные вероятности равны нулю. Кроме того, можно заметить, что в задаче 3.15 вероятность гипотезы H_1 после опыта уменьшилась, так как $P(H_1)=0{,}6=57/95$, $P_A(H_1)=9/19=45/95$, а в задаче 3.16 вероятность гипотезы H_2 увеличилась, поскольку $P(H_2)=0{,}18=9/50$, $P_A(H_2)=9/13$. Возможность таких ситуаций уже предполагалась ранее. ▲

Не анализируя условие предыдущей задачи, можно было высказать только две гипотезы H_2 и H_3, не приняв во внимание гипотезы H_1 и H_4. Это было бы довольно правдоподобно, однако подобные «просчеты» могут привести к грубым ошибкам. Поэтому при решении задач следует помнить, что высказанные гипотезы должны образовывать полную группу, а значит, сумма их доопытных и послеопытных вероятностей должна быть равна единице.

ГЛАВА 4
ПОВТОРЕНИЕ ИСПЫТАНИЙ

§ 4.1. Формула Бернулли

При решении вероятностных задач часто приходится сталкиваться с ситуациями, в которых одно и то же испытание (опыт) повторяется многократно.

Например, стрелок, не сходя с места, каждый раз тщательно прицеливаясь, производит несколько выстрелов по мишени или несколько человек заполняют по одной карточке «Спортлото». В результате каждого такого испытания может наступить или не наступить некоторое событие A. В результате одного выстрела (испытания) мишень может быть поражена (событие A) или нет (событие $\bar{A}$). В результате заполнения одной карточки «Спортлото» (испытания) можно отгадать все шесть номеров (событие A) или не отгадать все номера (событие $\bar{A}$). Можно предположить, что в приведенных ситуациях вероятность появления события A в каждом испытании одна и та же (для каждой ситуации своя). Модель каждой из этих ситуаций выглядит следующим образом. Проводится n испытаний, в каждом из которых событие A может произойти или не произойти, причем вероятность события в каждом отдельном испытании постоянна, т. е. не меняется от испытания к испытанию. Вопрос о том, как находятся вероятность события в отдельном испытании, уже был рассмотрен. Поэтому представляет особый интерес появление любого определенного числа раз события A в n испытаниях, точнее, вероятность появления этого числа. Рассмотрению задач, в которых требуется определить вероятность m появлений события A в результате n испытаний, и посвящена настоящая глава. Подобные задачи решаются сравнительно легко, если испытания являются независимыми.

Определение. *Несколько испытаний называются независимыми относительно события A, если вероятность события A в каждом из них не зависит от исходов других испытаний.*

Примером независимых испытаний может служить несколько подбрасываний монеты. Несколько последовательных выниманий из урны одинаковых на ощупь, но разных по цвету шаров, также являются независимыми испытаниями, например относительно появления белого шара, если шары каждый раз возвращаются назад и тщательно перемешиваются.

Возникает вопрос: как связаны независимость испытаний и независимость событий, которые могут произойти в результате этих испытаний? Пусть проводятся n независимых испытаний, в которых событие A может наступить или не наступить. Это означает, что в результате каждого испытания может произойти событие A, причем его вероятность не изменяется от того, какие события произойдут в остальных испытаниях. Не исключена возможность, что во всех этих испытаниях произойдет событие A. Обозначим через A_i, $i=1$, 2, 3, ..., n, событие A, если оно произойдет в i-м испытании.

Из определений независимости испытаний и независимости событий в совокупности следует, что события A_1, A_2, A_3, ..., A_n независимы в совокупности. Однако совсем не обязательно, чтобы во всех испытаниях произошло событие A.

Пусть для определенности в пяти независимых испытаниях два раза произошло событие A, причем наступило оно во 2-м и 5-м испытаниях. Это означает, что в пяти испытаниях наступили события $\bar{A}_1$, A_2, $\bar{A}_3$, $\bar{A}_4$, A_5. А так как при замене

любого числа событий из группы независимых в совокупности событий на противоположные им события независимость событий сохраняется (см. § 3.2), то представленная группа событий также независима в совокупности.

Практически событие A может появиться в n независимых испытаниях любое число раз в разных последовательностях или комбинациях, чередуясь с противоположным событием $\bar{A}$. Такая группа событий независима в совокупности. Таким образом, из независимости n испытаний относительно события A следует независимость в совокупности группы n событий, представляющей собой произвольную комбинацию событий A и $\bar{A}$, одно из которых обязательно произойдет в каждом из рассматриваемых испытаний.

По теореме умножения вероятностей для независимых событий, вероятность совместного наступления таких событий равна произведению их вероятностей, т. е. выполняется, например, равенство

$$P(A_1\bar{A}_2\bar{A}_3\bar{A}_4\ldots\bar{A}_n)=P(A_1)P(\bar{A}_2)P(\bar{A}_3)P(\bar{A}_4)\ldots P(\bar{A}_n).$$

Рассмотрим одну из ситуаций, в которой событие может произойти в любом из нескольких независимых событий.

○ **Пример 4.1.** Монета в неизменных условиях бросается три раза. Пусть событие A состоит в выпадании «герба». Естественно считать эти подбрасывания независимыми испытаниями. Вероятность появления «герба» в каждом единичном испытании, очевидно, постоянна и равна 1/2.

Таким образом, в рассматриваемой ситуации налицо независимые испытания, в каждом из которых вероятность события A постоянна и от испытания к испытанию не меняется. Такие испытания обычно называют *испытаниями Бернулли* или *схемой Бернулли*. Найдем теперь вероятность того, что два раза появится «герб». Обозначим: A_i — «в i-м испытании появился «герб»; $\bar{A}_i$ — «в i-м испытании появилась решка».

При трехкратном бросании монеты возможны следующие восемь исходов: $A_1A_2A_3$; $\bar{A}_1A_2A_3$; $A_1\bar{A}_2A_3$; $A_1A_2\bar{A}_3$; $\bar{A}_1\bar{A}_2A_3$; $\bar{A}_1A_2\bar{A}_3$; $A_1\bar{A}_2\bar{A}_3$; $\bar{A}_1\bar{A}_2\bar{A}_3$ — и только три из них, а именно: $\bar{A}_1A_2A_3$, $A_1\bar{A}_2A_3$, $A_1A_2\bar{A}_3$ — благоприятствуют событию B, состоящему в двукратном появлении «герба» в трех испытаниях. Следовательно, можно записать $B=\bar{A}_1A_2A_3+A_1\bar{A}_2A_3+A_1A_2\bar{A}_3$. Таким образом, событие B наступит, если первый раз появится решка, а второй и третий — «герб»; первый и третий раз появится «герб», а второй — решка; первый и второй раз появится «герб», а последний — решка. Так как эти варианты несовместны и все события, входящие в произведения, независимы, то по теоремам умножения и сложения вероятностей имеем

$$P(B)=P(\bar{A}_1A_2A_3)+P(A_1\bar{A}_2A_3)+P(A_1A_2\bar{A}_3)=$$
$$=3\cdot 1/2\cdot 1/2\cdot 1/2=3/8.$$

Этот же результат можно получить, если воспользоваться классической формулой вероятности. Действительно, все перечисленные выше восемь исходов трехкратного бросания монеты равновозможны и только три из них благоприятствуют событию B. Поэтому $P(B)=M/N=3/8$. ●

При увеличении числа бросаний монеты количество исходов увеличивается, а следовательно, возрастает трудность вычисления необходимых вероятностей.

Выработаем теперь общий метод решения подобных задач, позволяющий с минимальными вычислительными затратами получить требуемый результат. Поставим задачу в общем виде. Пусть в результате испытания возможны два исхода: либо появится событие A, либо противоположное ему событие $\bar{A}$. Проведем n испытаний Бернулли. Напомним, это означает, что все n испытаний независимы; вероятность появления события A в каждом отдельно взятом, или единичном испытании постоянна и от испытания к испытанию не изменяется (т. е. испытания проводятся в одинаковых условиях). Обозначим вероятность $P(A)$ появления события A в единичном испытании буквой p, т. е. $P(A)=p$, а вероятность $P(\bar{A})$ — буквой q, т. е. $P(\bar{A})=1-P(A)=1-p=q$.

Найдем вероятность $P_n(m)$ наступления события A ровно m раз (ненаступления $n-m$ раз) в этих n испытаниях. Отметим, что здесь не требуется появление m раз события A в определенной последовательности.

Обозначим: A_i — появление события A в i-м опыте; $\bar{A}_i$ — непоявление события A в i-м опыте, где $i=1, 2, 3, ..., n$.

Для одного испытания возможны следующие два исхода: A, $\bar{A}$. Вероятности этих исходов выпишем в виде следующей таблицы:

События	A	$\bar{A}$
Вероятность	p	q

Очевидно, $P_1(1)=p$; $P_1(0)=q$ и $P_1(1)+P_1(0)=(p+q)^1=1$.

Для двух испытаний возможны следующие $4=2^2$ исхода: A_1A_2, $\bar{A}_1A_2$, $A_1\bar{A}_2$, $\bar{A}_1\bar{A}_2$. Вероятность этих исходов также выпишем в виде таблицы:

События	A_1A_2	$A_1\bar{A}_2$	$\bar{A}_1A_2$	$\bar{A}_1\bar{A}_2$
Вероятность	p^2	pq	pq	q^2

Очевидно, $P_2(2)=p^2$, $P_2(1)=P(A_1\bar{A}_2)+P(\bar{A}_1A_2)=2pq$, $P_2(0)=q^2$ и $P_2(2)+P_2(1)+P_2(0)=p^2+2pq+q^2=(p+q)^2=1$.

Для трех испытаний возможны следующие $8=2^3$ исходов: $A_1A_2A_3$, $\bar{A}_1A_2A_3$, $A_1\bar{A}_2A_3$, $A_1A_2\bar{A}_3$, $\bar{A}_1\bar{A}_2A_3$, $\bar{A}_1A_2\bar{A}_3$, $A_1\bar{A}_2\bar{A}_3$, $\bar{A}_1\bar{A}_2\bar{A}_3$. Вероятности этих исходов запишем в виде таблицы:

События	$A_1A_2A_3$	$\bar{A}_1A_2A_3$	$A_1\bar{A}_2A_3$	$A_1A_2\bar{A}_3$	$\bar{A}_1\bar{A}_2A_3$	$\bar{A}_1A_2\bar{A}_3$	$A_1\bar{A}_2\bar{A}_3$	$\bar{A}_1\bar{A}_2\bar{A}_3$
Вероятность	p^3	p^2q	p^2q	p^2q	pq^2	pq^2	pq^2	q^3

Очевидно, $P_3(3)=p^3$, $P_3(2)=P(\bar{A}_1A_2A_3)+P(A_1\bar{A}_2A_3)+$ $+P(A_1A_2\bar{A}_3)=3p^2q$, $P_3(1)=P(\bar{A}_1\bar{A}_2A_3)+P(\bar{A}_1A_2\bar{A}_3)+$ $+P(A_1\bar{A}_2\bar{A}_3)=3pq^2$, $P_3(0)=q^3$ и $P_3(3)+P_2(2)+P_3(1)+P_3(0)=$ $=p^3+3p^2q+3pq^2+q^3=(p+q)^3=1$. Анализируя эти случаи, можно сделать общий вывод: вероятность $P_n(m)$ пропорциональна произведению p^mq^{n-m}, причем коэффициент пропорциональности равен C_n^m, т. е.

$$P_n(m)=C_n^mp^mq^{n-m}.$$

Полученную формулу можно доказать используя метод математической индукции. Однако в качестве доказательства приведем следующие рассуждения.

Пусть проведено n независимых испытаний, в каждом из которых событие A может наступить и не наступить. Если в результате испытаний событие A произошло m раз (неважно, в каком порядке), то это означает, что совместно наступили m событий A и $n-m$ событий $\bar{A}$, вероятности которых в каждом отдельном опыте равны p и q соответственно. Так как все n событий независимы, то, по теореме умножения, вероятность появления m раз события A в определенной последовательности равна p^mq^{n-m}.

Однако событие A может появиться m раз в n опытах в совершенно другой последовательности и число таких последовательностей равно числу сочетаний из n по m, т. е. C_n^m (это число совпадает с числом способов, которыми можно выбрать m мест из имеющихся n, не учитывая их порядка). Все C_n^m вариантов появления события A m раз представляют собой несовместные события, вероятность каждого из которых равна p^mq^{n-m}. А так как наступление одного (любого) из этих событий означает наступление события, состоящего в том, что в n независимых испытаниях событие A появится m раз, то, по теореме сложения вероятностей для несовместных событий, искомая вероятность $P_n(m)$ равна сумме вероятностей всех указанных несовместных событий, т. е. произведению $C_n^mp^mq^{n-m}$.

Таким образом,

$$P_n(m)=C_n^mp^mq^{n-m}=\frac{n!}{m!(n-m)!}p^mq^{n-m}. \quad (4.1)$$

Полученную формулу называют *формулой Бернулли*.

Проиллюстрируем применение формулы (4.1).

Задача 4.1. Монету бросают 8 раз. Какова вероятность, что 4 раза выпадет «герб»?

△ Как правило, на вопрос задачи отвечают, что эта вероятность равна 1/2. Однако это ошибочный ответ. Действительно, здесь $n=8$, $m=4$, $p=q=1/2$. По формуле (4.1) получаем

$$P_8(4)=C_8^4(1/2)^4(1/2)^4=70/256<1/3. \ \blacktriangle$$

Задача 4.2. Вероятность изготовления на станке-автомате нестандартной детали равна 0,02. Какова вероятность того, что среди наудачу взятых шести деталей окажется более четырех стандартных?

△ Здесь $n=6$, $p=0{,}02$, $q=1-p=0{,}98$. Появление более четырех стандартных деталей означает, что среди взятых шести деталей пять или шесть стандартных, т. е. одна или ноль бракованных. По теореме сложения вероятностей,

$$P_6(0\leqslant m<1)=P_6(0)+P_6(1)=$$
$$=C_6^0(0{,}02)^0(0{,}98)^6+C_6^1(0{,}02)^1(0{,}098)^5\approx 0{,}9943. \ \blacktriangle$$

§ 4.2. Асимптотические формулы

При решении задач 4.1, 4.2 особых трудностей при вычислении искомых вероятностей не возникало, так как число испытаний n было невелико. Однако если число испытаний достаточно велико, то использование формулы (4.1) нецелесообразно в силу необходимости выполнения громоздких вычислений. Пусть, например, требуется вычислить P_{320} (285) (при $p=0{,}89$). В данном случае формула (4.1) принимает вид

$$P_{320}(285)=\frac{320!}{285!\,35!}(0{,}89)^{285}(0{,}11)^{35}.$$

Получить по указанной формуле более или менее точный результат практически невозможно.

В этом параграфе рассматриваются специальные методы, с помощью которых можно получить достаточно точные ответы в задачах, связанных с повторением испытаний, не прибегая к сложным вычислениям. Суть первого метода заключается в применении локальной теоремы Муавра — Лапласа, дающей асимптотическую формулу, которая позволяет вычислить вероятность приближенно.

Локальная теорема Муавра—Лапласа. *Если вероятность наступления события A в каждом из n независимых испытаниях равна p и отлична от нуля и единицы, а число испытаний достаточно велико, то вероятность $P_n(m)$ того, что в n испытаниях событие A наступит m раз, приближенно равна (чем больше n, тем точнее) значению функции* $y=\frac{1}{\sqrt{npq}}f(u)$, *где*

$$f(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2},\quad u=\frac{m-np}{\sqrt{npq}},$$

т. е.

$$P_n(m)\approx\frac{1}{\sqrt{npq}}f\left(u=\frac{m-np}{\sqrt{npq}}\right)^{*)}. \qquad (4.2)$$

Значение функции $f(u)$ для любого аргумента u можно получить разложив ее в степенной ряд, однако в этом нет необходимости, так как ее значения можно найти из специальной таблицы. При использовании этой таблицы следует помнить, что функция $f(u)$ четная, т. е. $f(-u)=f(u)$. Более подробно эта функция рассматривается в § 6.2.

Задача 4.3. Вероятность того, что сошедшая с конвейера деталь стандартная, равна 0,9. Найти вероятность того, что из 400 сошедших с конвейера деталей 356 окажутся стандартными.

△ Согласно условию задачи, $n=400$, $m=356$, $p=0{,}9$, $q=0{,}1$. По формуле (4.2) находим

$$p_{400}(356)\approx\frac{1}{\sqrt{400\cdot 0{,}9\cdot 0{,}1}}f(u)=\frac{1}{6}f(u).$$

Далее из условия задачи следует, что

$$u=\frac{m-np}{\sqrt{npq}}=\frac{356-400\cdot 0{,}9}{6}\approx -0{,}67.$$

По табл. П.1**), учитывая, что $f(-u)=f(u)$, находим $f(-0{,}67)=$ $=0{,}3188$. Искомая вероятность

$$P_{400}(356)\approx 0{,}3188/6\approx 0{,}0531. \ \blacktriangle$$

Такое небольшое значение полученной вероятности, несмотря на то что вероятность появления стандартной детали равна 0,9, объясняется тем, что была вычислена вероятность только одного из 401 исходов сложного испытания, состоящего в отборе 400 деталей.

Задача 4.4. Два спортсмена играют в настольный теннис. Вероятность выигрыша первого спортсмена равна 5/9. Какова вероятность того, что он выиграет две партии из пяти?

△ По условию, $p=5/9$, $q=4/9$, $n=5$, $m=2$. Воспользуемся формулой (4.2). Имеем

$$P_5(2)\approx\frac{1}{\sqrt{npq}}f(u)=\frac{1}{\sqrt{5\cdot(5/9)\cdot(4/9)}}f(u)=0{,}9f(u).$$

Найдем значение аргумента

*) «Грубое» правило для применения формулы (4.2) вместо формулы Бернулли (4.1) состоит в том, что n должно иметь порядок не менее нескольких десятков, а лучше нескольких сотен, а произведение $np>10$.

**) Запись П.1, П.2 и т. д. здесь и далее означает ссылку на табл. 1, табл. 2 и т. д. Приложений.

$$u=\frac{m-np}{\sqrt{npq}}=\frac{2-5\cdot 5/9}{10/9}=-0{,}7.$$

По табл. П.1 находим $f(-0{,}7)=0{,}3123$. Искомая вероятность

$$P_5(2)\approx 0{,}9\cdot 0{,}3123\approx 0{,}281.$$

Проверим полученный результат, воспользовавшись формулой Бернулли. Имеем

$$P_5(2)=C_5^2(5/9)^2(4/9)^3\approx 0{,}271. \ \blacktriangle$$

Расхождение ответов объясняется тем, что формула (4.2) дает хорошее приближение при больших значениях n, а в данном случае оно равно 5. Формула (4.2) позволяет получить более близкие к точному значению $P_n(m)$ результаты, чем больше значение $\sqrt{npq}$ и чем ближе значения p и q к 0,5.

Если вероятность события p (или q) в отдельном испытании близка к нулю (такие события называются *редкими*), то даже при большом числе испытаний n, но небольшой величине произведения np (меньше 10) вероятности $P_n(m)$, полученные по формуле (4.2), недостаточно близки к их истинным значениям. В таких случаях применяют другую асимптотическую формулу — формулу Пуассона, справедливость которой доказывает следующая теорема.

Теорема. *Если вероятность p наступления события A в каждом испытании постоянна, но близка к нулю, число независимых испытаний n достаточно велико, а произведение $np=\lambda$, то вероятность $P_n(m)$ того, что в n независимых испытаниях событие A наступит m раз, приближенно равна $\frac{\lambda^m e^{-\lambda}}{m!}$, т. е.*

$$P_n(m)\approx\frac{\lambda^m e^{-\lambda}}{m!}. \tag{4.3}$$

Формула (4.3) называется *формулой Пуассона.*

□ Для вычисления вероятности $P_n(m)$ воспользуемся формулой Бернулли. Имеем

$$P_n(m)=\frac{n!}{m!(n-m)!}p^m q^{n-m}=\frac{n(n-1)(n-2)\dots(n-m+1)}{m!}p^m q^{n-m}.$$

Так как, по условию, $np=\lambda$, то $p=\lambda/n$. Тогда

$$P_n(m)=\frac{n(n-1)(n-2)\dots(n-m+1)}{m!}\left(\frac{\lambda}{n}\right)^m\left(1-\frac{\lambda}{n}\right)^{n-m}.$$

Так как, по условию, n велико, то найдем предел правой части последнего равенства при $n\to\infty$. При этом будет получено приближенное значение вероятности $P_n(m)$. Итак,

$$P_n(m)\approx\lim_{n\to\infty}\frac{n(n-1)(n-2)\dots(n-m+1)}{m!}\left(\frac{\lambda}{n}\right)^m\left(1-\frac{\lambda}{n}\right)^{n-m}=$$

$$=\frac{\lambda^m}{m!}\lim_{n\to\infty}\left[\frac{n}{n}\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\dots\left(\frac{n-m+1}{n}\right)\right]\left(1-\frac{\lambda}{n}\right)^{-m}\left(1-\frac{\lambda}{n}\right)^{n}=$$

$$=\frac{\lambda^m}{m!}\lim_{n\to\infty}\left(1-\frac{1}{n}\right)\lim_{n\to\infty}\left(1-\frac{2}{n}\right)\dots\lim_{n\to\infty}\left(1-\frac{m-1}{n}\right)\times$$

$$\times\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-m}\lim_{n\to\infty}\left[\left(1-\frac{\lambda}{n}\right)^{-n/\lambda}\right]^{-\lambda}=\frac{\lambda^m}{m!}e^{-\lambda}.$$

Здесь пределы всех сомножителей, кроме последнего (второго замечательного предела), равны 1. Окончательно имеем

$$P_n(m)\approx\frac{\lambda^m}{m!}e^{-\lambda} \ \blacksquare$$

Условия теоремы требуют, чтобы вероятность события p была мала, а число испытаний n велико. Обычно указанную формулу используют, когда $n\geqslant 10$, лучше $n\geqslant 100$, а $np<10$.

Задача 4.5. Некоторое электронное устройство выходит из строя, если откажет определенная микросхема. Вероятность ее отказа в течение 1 ч работы устройства равна 0,004. Какова вероятность того, что за 100 ч работы устройства придется пять раз менять микросхему?

△ По условию, $n=1000$, $p=0{,}004$, а $\lambda=np=1000\cdot 0{,}004=4<10$. Для нахождения вероятности $P_{1000}(5)$ воспользуемся формулой Пуассона, так как условия ее применения выполнены. Имеем

$$P_{1000}(5)\approx\frac{4^5 e^{-4}}{5!}\approx 0{,}1563.$$

Найдем вероятность того же события по формуле (4.2). Значение u равно

$$u=\frac{m-np}{\sqrt{npq}}=\frac{5-4}{\sqrt{1000\cdot 0{,}004\cdot 0{,}966}}\approx\frac{1}{1{,}996}\approx 0{,}501.$$

Из табл. П.1, следует, что $f(u)\approx 0{,}3521$. Тогда

$$P_{1000}(5)\approx\frac{1}{\sqrt{npq}}f(u)\approx 0{,}3521/1{,}996\approx 0{,}1764.$$

Значение искомой вероятности по формуле Бернулли равно

$$C_{1000}(5)=C_{1000}^{5}(0{,}004)^{5}(0{,}996)^{995}\approx 0{,}1566.$$

Как видно из полученных результатов, формула (4.3) дает более точный результат, чем формула (4.2). ▲

ГЛАВА 5
СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

§ 5.1. Понятие случайной величины

В практической жизни часто приходится сталкиваться с различными величинами. Так, например, при покупке авиабилета нас интересуют величина его стоимости и продолжительность полета. Для выпечки торта необходимо знать его рецепт, т. е. количество муки, сахара, масла и т. д., необходимое на приготовление теста. Значения многих из встречающихся величин могут быть заранее известны. К таким величинам относятся: продолжительность земных суток в часах; процентное содержание компонентов выбранного торта; количество членов семьи. Значение других величин можно непосредственно найти из опыта или с помощью вычислений, получив предварительные данные с помощью измерений или пересчета, т. е. также из опыта.

В результате повторения некоторых опытов можно всегда получать одно и то же значение определенной величины, а в результате других значение величины изменяется, причем результат каждого отдельного опыта невозможно предугадать заранее.

Например, позвонив в справочное бюро (что является опытом), можно узнать стоимость авиабилета на выбранный рейс. В этом случае сообщенная конкретная стоимость билета является значением интересующей нас величины. Это значение (стоимость билета) неизменно, сколько бы раз мы ни звонили в справочное бюро. Если узнавать количество билетов, имеющихся на данный момент в кассе на конкретный рейс, то каждый раз в общем случае будут получены различные ответы, причем неизвестно заранее — какие. В данном опыте (звонок в справочное бюро) значение величины (количество билетов) меняется случайным образом от опыта к опыту. Величины, которые могут принять в результате опыта любое из возможных значений, неизвестно заранее — какое, заслуживают особого внимания и являются предметом дальнейшего изучения.

Определение. *Случайной называется величина, которая в результате опыта может принять то или иное возможное значение, неизвестное заранее, но обязательно одно.*

При многократном проведении опыта (испытания) в неизменных условиях в общем случае будут получены различные значения случайной величины. Это обусловлено случайными обстоятельствами, которые практически невозможно предусмотреть.

Например, при разовом бросании игрального кубика может появиться одно из чисел: 1; 2; 3; 4; 5; 6. Какое конкретное число появится, предугадать невозможно, так как появление любого из указанных чисел зависит от многих причин, которые нельзя учесть. Если при первом бросании может появиться 3, то при втором бросании возможно появление 1 или 5, а может вновь появиться 3. Таким образом, при повторении опыта число выпавших очков меняется случайным образом.

Итак, количество выпавших очков при бросании игрального кубика — величина переменная, характер ее изменения зависит от многих случайных причин. Такая величина является случайной.

Приведем примеры случайных величин.

○ **Пример 5.1.** Число выпавших «гербов» при пятикратном бросании монеты. ●

○ **Пример 5.2.** Число бракованных изделий в случайно отобранной партии из 20 изделий. ●

○ **Пример 5.3.** Дальность полета артиллерийского снаряда. ●

○ **Пример 5.4.** Наружный диаметр трубы. ●

○ **Пример 5.5.** Число мальчиков, родившихся в течение суток в определенной стране. ●

В примерах 5.1, 5.2 и 5.5 случайная величина может принимать отдельные изолированные значения, которые можно заранее перечислить. Так, в примере 5.1 такими значениями являются 0, 1, 2, 3, 4, 5, в примере 5.2—0, 1, 2, 3, ..., 20, в примере 5.5—0, 1, 2, 3, 4, 5, ... Подобные случайные величины называются *дискретными* (*прерывными*). Заметим, что значения трех указанных выше случайных величин отделены друг от друга промежутками, в которых нет других возможных значений соответствующих величин.

В примерах 5.3 и 5.4 возможные значения случайной величины не отделены друг от друга и заполняют некоторый интервал. В этом случае одно значение случайной величины нельзя отделить от другого промежутком, не содержащим возможного значения этой же случайной величины.

Предположим (см. пример 5.3), что расчетная дальность полета снаряда 7000 м. Пусть при первом выстреле снаряд пролетел 7020, а при втором — 7040 м. При последующих выстрелах снаряд может пролететь и 7030, и 6995 м. Другими словами, снаряд может попасть в любую точку некоторого промежутка и невозможно указать какие-либо два возможных значения дальности полета снаряда, между которыми не найдется хотя бы одного возможного значения рассматриваемой случайной величины. Такие случайные величины называются *непрерывными.*

Определение. *Дискретной случайной величиной называют такую случайную величину, множество возможных значений которой либо конечное, либо бесконечное, но счетное.*

Примерами дискретной случайной величины являются число учащихся, опрошенных на уроке; число солнечных дней в году.

Определение. *Непрерывной случайной величиной называют такую случайную величину, которая может принять любое значение из некоторого конечного или бесконечного интервала.*

Примерами непрерывной случайной величины служат: время безаварийной работы станка; расход горючего на единицу расстояния; количество осадков, выпавших в сутки.

Случайные величины будем обозначать заглавными буквами конца латинского алфавита — X, Y, Z, а их возможные значения — соответствующими малыми буквами — x, y, z. Например, X — число шахматных партий, окончившихся ничейным результатом, из трех сыгранных. В этом случае величина X может принять следующие значения: $x_1=0$, $x_2=1$, $x_3=2$, $x_4=3$.

Введем теперь операции над случайными величинами. Пусть имеются две случайные величины X и Y, возможными значениями которых являются соответственно $x_1, x_2, x_3, \ldots, x_n$ и $y_1, y_2, y_3, \ldots, y_n$.

Определение. *Суммой* $X+Y$ *случайных величин* X и Y *называется случайная величина* Z, *возможные значения которой есть* x_1+y_1, x_1+y_2, $x_1+y_3, \ldots, x_1+y_j, \ldots, x_2+y_1$, x_2+y_2, $x_2+y_3, \ldots, x_2+y_j, \ldots, x_i+y_1$, x_i+y_2, $x_i+y_3, \ldots, x_i+y_j, \ldots$ $\ldots x_n+y_n$.

Это определение следует понимать так: в результате опыта, в котором случайная величина X может принять то или иное значение, было получено число x_i (конкретное значение величины X), а в результате опыта, в котором уже случайная величина Y может принять то или иное значение, было получено число y_j (конкретное значение величины Y), после чего полученные числа складываются. Число x_i+y_j и является одним из возможных значений случайной величины $Z=X+Y$.

Определение. *Произведением* XY *случайных величин* X и Y *называется случайная величина* Z, *возможные значения которой есть* x_1y_1, x_1y_2, $x_1y_3, \ldots, x_2y_1$, x_2y_2, $x_2y_3, \ldots x_iy_j, \ldots$, x_ny_m.

Определение. *Произведением* CX случайной величины X на *постоянную* C *называется такая случайная величина* Z, *возможные значения которой есть* Cx_1, Cx_2, $Cx_3, \ldots, Cx_i, \ldots$, Cx_n.

Аналогично определяются разность $X-Y$ и частное X/Y двух случайных величин.

○ **Пример 5.6.** Пусть случайная величина X может принять следующие три значения: $x_1=1$, $x_2=2$, $x_3=3$, а случайная величина Y — значения $y_1=4$, $y_2=7$, $y_3=11$. Тогда возможные значения случайной величины $Z=X+Y$ таковы: $z_1=5$, $z_2=8$, $z_3=12$, $z_4=6$, $z_5=9$, $z_6=13$, $z_7=7$, $z_8=10$, $z_9=14$, а возможные значения случайной величины $Z=XY$ — следующие: $z_1=4$, $z_2=7$, $z_3=11$, $z_4=8$, $z_5=14$, $z_6=22$, $z_7=12$, $z_8=21$, $z_9=33$. Для величины $Z=5X$ возможными значениями являются $z_1=5$, $z_2=10$, $z_3=15$. ●

○ **Пример 5.7.** Пусть случайная величина X — число выпавших очков при бросании одного игрального кубика, а случайная величина Y — число выпавших очков при бросании второго игрального кубика. Рассмотрим случайную величину $Z=X+Y$ и обозначим $x_i+y_j=z_{ij}$, $i, j=1, 2, 3, 4, 5, 6$. Нетрудно заметить, что $z_{26}=z_{35}=z_{44}=z_{53}=z_{62}=8$. ●

Понятие суммы и произведения случайных величин можно по аналогии распространить на любое конечное число случайных величин.

§ 5.2. Ряд распределения случайной величины

Появление тех или иных значений случайной величины можно рассматривать как события, а различным событиям в общем случае, как известно из рассмотренного выше, соответствуют различные вероятности. Поэтому возможные значения случайной величины отличаются между собой с вероятностной точки зрения.

Так, например, при бросании двух игральных кубиков (см. пример 5.7) такие значения случайной величины $Z=X+Y$, как $z=2$ и $z=8$, находятся в «неодинаковых условиях». Значение $z=2$ может появиться только в одном случае, когда появятся значения $x_1=1$ и $y_1=1$, а значение $z=8$ может появиться в пяти случаях. Отсюда следует, что вероятность появления $z=2$ меньше, чем вероятность появления $z=8$.

Таким образом, перечисление всех возможных значений случайной величины не дает достаточно полного представления о ней. Кроме того, необходимо знать, как часто могут появляться те или иные ее значения в результате испытаний, проводящихся в одинаковых условиях, т. е. следует знать вероятности их появления.

Рассмотрим дискретную случайную величину X с возможными значениями $x_1, x_2, x_3, \ldots, x_n$. В результате опыта случайная величина примет одно и только одно из этих значений. Другими словами, произойдет одно из несовместных событий, образующих полную группу: $X=x_1$, $X=x_2$, $X=x_3, \ldots, X=x_n$. Обозначим вероятность этих событий буквами p с соответствующими индексами: $P(X=x_1)=p_1$, $P(X=x_2)=p_2$, $P(X=x_3)=p_3, \ldots, P(X=x_n)=p_n$. Так как указанные события образуют полную группу, то сумма вероятностей появления возможных значений случайной величины равна 1, т. е.

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

Если же множество значений случайной величины образует бесконечное, но счетное множество, то ряд $\sum_{i=1}^{\infty} p_i$ сходится и его сумма равна 1.

Таким образом, суммарная вероятность, равная 1, распределена между всеми значениями случайной величины.

Определив все возможные значения случайной величины X и правило, по которому каждому событию $X=x_i$, $i=1, 2, 3, \ldots, n$, ставится в соответствие вероятность, т. е. правило распределения вероятностей между значениями случайной величины, можно получить полное представление о случайной величине.

Определение. *Законом распределения дискретной случайной величины называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими вероятностями.*

Про случайную величину говорят, что она подчиняется данному закону распределения.

Закон распределения случайной величины можно задать так же, как в математическом анализе функцию одного аргумента, используя табличный, графический или аналитический способ задания. Рассмотрим первый из них.

При табличном способе задания закона распределения первая строка таблицы содержит возможные значения случайной величины (обычно в порядке возрастания), а вторая — соответствующие вероятности (см. табл. 5.1).

Таблица 5.1

X	x_1	x_2	x_3	…	x_i	…	x_n
P	p_1	p_2	p_3	…	p_i	…	p_n

Эта таблица называется *рядом распределения*.

Задача 5.1. В партии из восьми деталей пять стандартных. Наудачу взяты четыре детали. Построить ряд распределения числа стандартных деталей среди отобранных.

△ Пусть X — число стандартных деталей среди четырех отобранных. Оно может принять следующие четыре значения: $x_1=1$, $x_2=2$, $x_3=3$, $x_4=4$. Для определения вероятности появления конкретного числа стандартных деталей воспользуемся формулой

$$P(X=k)=\frac{C_m^k C_{n-m}^{l-k}}{C_n^l},$$

где n — число деталей в партии, l — число отобранных деталей, m — число стандартных деталей, k — число стандартных деталей среди отобранных. Далее имеем

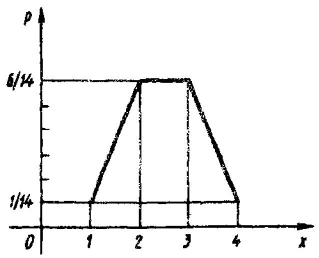


Рис. 5.1

$$P(X=1)=\frac{C_5^1 C_3^3}{C_8^4}=1/14,$$

$$P(X=2)=\frac{C_5^2 C_3^2}{C_8^4}=6/14,$$

$$P(X=3)=\frac{C_5^3 C_3^1}{C_8^4}=6/14,$$

$$P(X=4)=\frac{C_5^4 C_3^0}{C_8^4}=1/14.$$

Проверим вычисления. Складывая полученные вероятности, получаем $1/14+6/14+6/14+1/14=1$. Искомый ряд распределения имеет вид

X	1	2	3	4
P	1/14	6/14	6/14	1/14

▲

Ряд распределения можно задать графически, если по оси абсцисс отложить возможные значения случайной величины, а по оси ординат — вероятности этих значений. Соединив точки $(x_i;\ y_i)$ последовательно отрезками прямой линии, получим ломаную, которая называется *многоугольником распределения вероятностей.*

Воспользовавшись результатами задачи 5.1, построим многоугольник распределения (рис. 5.1). Многоугольник распределения, как и ряд распределения, полностью характеризует случайную величину и является одним из способов (графическим) задания закона распределения.

Заметим, что сумма ординат многоугольника равна единице. Это свойство многоугольника распределения является определяющим. Если в прямоугольной системе координат дана некоторая ломаная, удовлетворяющая определению функции и обладающая указанным выше свойством, то такая ломаная, очевидно, задает закон распределения некоторой случайной величины.

§ 5.3. Функция распределения вероятностей

В предыдущем параграфе были рассмотрены два способа задания распределения дискретной случайной величины. Напомним, что дискретная случайная величина может быть задана перечнем всех ее возможных значений и соответствующих вероятностей.

При задании закона распределения непрерывной случайной величины такой способ уже неприемлем хотя бы потому, что множество ее возможных значений бесконечно и сплошь заполняет некоторый промежуток. В этом случае не представляется возможности перечислить все значения случайной величины и их вероятности в виде таблицы (построить ряд распределения) или отметить их в системе координат (построить многоугольник распределения).

Кроме того, как будет показано в дальнейшем, каждое отдельное значение непрерывной случайной величины обладает нулевой вероятностью. Это означает, что все значения непрерывной случайной величины, образующие бесконечное множество, имеют одинаковую и равную нулю вероятность.

Однако, несмотря на равенство нулю вероятностей отдельных значений непрерывной случайной величины, нахождение ее возможных значений в различных интервалах обладает различными и отличными от нуля вероятностями. Таким образом, для непрерывной случайной величины, так же как и для дискретной, можно определить закон распределения, но несколько в ином виде, чем для дискретной.

Для характеристики поведения непрерывной случайной величины целесообразно использовать не вероятность события $X=x$, а вероятность события $X<x$, где x — некоторое действительное число. Другими словами, представляет интерес вероятность события, состоящего в том, что в результате опыта случайная величина X примет значение, которое окажется меньше некоторого фиксированного x. Если теперь x изменяется произвольно, то вероятность выполнения неравенства $X<x$ в общем случае будет изменяться. Следовательно, вероятность $P(X<x)$ является функцией аргумента x. Обозначим эту функцию $F(x)$.

Определение. *Функцией распределения случайной величины X называется функция $F(x)$, задающая вероятность того, что случайная величина X принимает значение, меньшее x, т. е.*

$$F(x)=P(X<x). \tag{5.1}$$

Иногда функцию $F(x)$ называют *интегральной функцией распределения.*

Функция распределения допускает простую геометрическую интерпретацию. Рассмотрим случайную величину X как случайную точку X на оси Ox (рис. 5.2), которая в результате опыта может занять то или иное положение. Пусть на оси выбрана конкретная точка x; тогда в результате опыта случайная точка X может оказаться левее или правее этой точки. Очевидно, вероятность того, что случайная точка X окажется левее точки x, будет зависеть от положения точки x, т. е. являться функцией аргумента x.

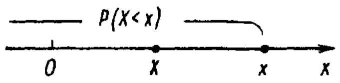


Рис. 5.2

Для дискретной случайной величины X, которая может принимать значения $x_1, x_2, x_3, \ldots, x_n$, функция распределения имеет вид

$$F(x)=\sum_{x_i<x} P(X=x_i), \tag{5.2}$$

где неравенство $x_i<x$ под знаком суммы означает, что суммирование касается всех тех значений x_i, величина которых меньше x. Поясним эту формулу исходя из определения функции $F(x)$. Предположим, что аргумент x принял какое-то определенное значение, но такое, что выполняется неравенство $x_i<x\leqslant x_{i+1}$. Тогда левее числа x на числовой оси окажутся только те значения случайной величины, которые имеют индекс 1, 2, 3, ..., i. Поэтому неравенство $X<x$ выполняется, если величина X примет значение x_k, где $k=1, 2, 3, \ldots, i$. Таким образом, событие $X<x$ наступит, если наступит любое, неважно какое, из событий $X=x_1$, $X=x_2$, $X=x_3$, ..., $X=x_i$. Так как эти события несовместны, то по теореме сложения вероятностей имеем

$$P(X<x)=P(X=x_1)+P(X=x_2)+$$
$$+P(X=x_3)+\ldots+P(X=x_i)=\sum_{x_i<x} P(X=x_i).$$

Прежде чем приступить к дальнейшему изучению функции $F(x)$, отметим большую значимость и глубокий смысл равенства (5.1). В левой части этого равенства находится «обыкновенная» функция действительного аргумента, а в правой — переменная вероятность. Эта формула связывает один из важных разделов математики, изучающий функции действительного аргумента, с теорией вероятностей, где изучаются, в частности, случайные события, которые могут произойти, а могут и не произойти в результате опыта.

В отличие от случайного события, наступление которого можно только предвидеть, так же как и появление определенного значения случайной величины, значение функции действительного аргумента всегда можно однозначно определить.

Формула (5.1) является своеобразным «мостом» между математическим анализом и теорией вероятностей, между функциями действительного переменного и случайными величинами. Эта формула дает возможность при исследовании в теории вероятностей использовать аппарат математического анализа, без которого было бы трудно получить хорошие результаты.

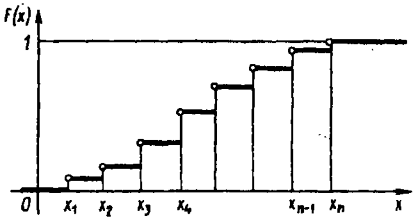


Рис. 5.3

Построим теперь функцию распределения случайной величины X, ряд распределения которой представлен в табл. 5.1:

при $x \leqslant x_1$ $\quad F(x)=P(X<x_1)=0;$

при $x_1<x\leqslant x_2$ $\quad F(x)=P(X<x_2)=P(X=x_1)=p_1;$

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

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

....

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

при $x>x_n$ $\quad F(x)=P(X=x_1)+P(X=x_2)+P(X=x_3)+\ldots$
$\ldots+P(X=x_n)=p_1+p_2+p_3+\ldots+p_n=1.$

График построенной функции изображен на рис. 5.3. Для дискретной случайной величины график функции распределения представляет собой разрывную ступенчатую линию. Когда переменная x проходит через какое-нибудь из возможных значений случайной величины, значение функции распределения меняется скачкообразно, т. е. функция имеет скачок в тех точках, в которых случайная величина принимает конкретное значение согласно ряду распределения, причем величина скачка равна вероятности этого значения. Сумма величин всех скачков функции распределения равна 1. В интервалах между значениями случайной величины функция $F(x)$ постоянна. Отметим, что по функции распределения дискретной случайной величины можно легко восстановить ее ряд распределения.

Задача 5.2. Используя результаты решения задачи 5.1, построить функцию распределения случайной величины X и ее график.

△ Имеем:

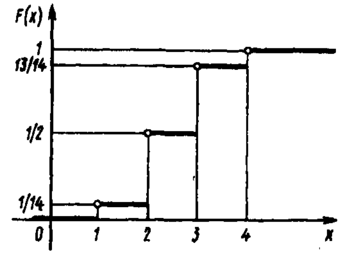


Рис. 5.4

при $x\leqslant 1$ $\quad F(x)=P(X<1)=0;$

при $1<x\leqslant 2$ $\quad F(x)=P(X<2)=1/14;$

при $2<x\leqslant 3$ $\quad F(x)=P(X<3)=1/14+6/14=1/2;$

при $3<x\leqslant 4$ $\quad F(x)=P(X<4)=1/14+6/14+6/14=13/14;$

при $x>4$ $\quad F(x)=1/14+6/14+6/14+1/14=1.$

Полученные результаты объединим в таблицу:

$F(x)$	0	1/14	1/2	13/14	1
Интервал	$]-\infty;\,1]$	$]1;\,2]$	$]2;\,3]$	$]3;\,4]$	$]4;\,\infty[$

График функции $F(x)$ изображен на рис. 5.4. ▲

Из рис. 5.4 видно, что $F(x)$ имеет четыре скачка по числу принимаемых случайной величиной X значений. По мере возрастания числа возможных значений случайной величины с одновременным уменьшением величины интервалов между ними число скачков становится больше, а сами скачки — меньше, вследствие чего ступенчатая кривая становится более плавной (рис. 5.5). В этом случае дискретная случайная величина постепенно приближается к непрерывной, а ее функция распределения — к непрерывной функции (рис. 5.6).

Рассмотрим свойства функции распределения.

1°. *Если $F(x)$ — функция распределения случайной величины X, то $0\leqslant F(x)\leqslant 1$ для любого x.*

□ Утверждение следует из формулы (5.1), так как веро-

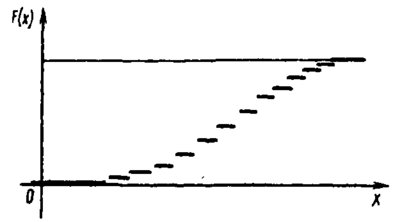


Рис. 5.5

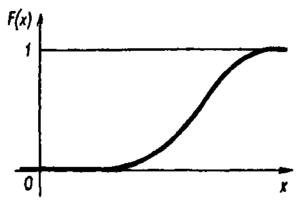


Рис. 5.6

ятность любого события есть неотрицательное число, не превышающее 1. ■

Прежде чем сформулировать и доказать второе свойство, приведем следующие соображения. Из геометрической интерпретации функции $F(x)$ следует, что для произвольных $x_1<x_2$ вероятность попадания значений случайной величины X в интервал $]-\infty;\ x_2[$ не меньше, чем вероятность попадания в интервал $]-\infty;\ x_1[$. Следовательно, функция $F(x)$ — неубывающая. Далее, при решении многих вероятностных задач возникает необходимость нахождения вероятности попадания значений случайной величины в заданный интервал. Следующее свойство, которое, в частности, подтверждает мысль о неубывании функции $F(x)$, позволяет определить значение указанной вероятности.

2^0. *Функция распределения $F(x)$ случайной величины X является неубывающей функцией, и для любых $\alpha<\beta$ выполняется равенство*

$$P(\alpha\leqslant X<\beta)=F(\beta)-F(\alpha). \tag{5.3}$$

□ Докажем предварительно равенство (5.3), для чего рассмотрим следующие события: событие A, состоящее в том, что $X<\alpha$; событие B, состоящее в том, что $X<\beta$; событие C, состоящее в том, что $\alpha\leqslant x<\beta$. Можно записать: $P(A)=P(X<\alpha)=F(\alpha)$; $P(B)=P(X<\beta)=F(\beta)$; $P(C)=P(\alpha\leqslant X<\beta)$. Нетрудно заметить, что событие B представляет собой сумму двух несовместных событий A и C. Геометрическая иллюстрация этого факта приведена на рис. 5.7. По теореме сложения вероятностей,

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

или

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

откуда и следует равенство (5.3). Так как вероятность любого события число неотрицательное, то, отбрасывая в последнем равенстве $P(\alpha\leqslant X<\beta)$, получаем $F(\beta)\geqslant F(\alpha)$. ■

Вернемся к геометрической интерпретации функции $F(x)$ (см. рис. 5.2). Пусть точка x неограниченно перемещается влево по оси, т. е. стремится к «минус бесконечности». Тогда попадание случайной точки X левее x в пределе становится невозможным событием. Другими словами, случайная точка X не может оказаться левее «минус бесконечности». Поэтому естественно предположить, что $\lim\limits_{x\to-\infty} F(x)=0$.

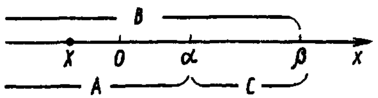


Рис. 5.7

При неограниченном перемещении точки x вправо, т. е. при

стремлении ее к «плюс бесконечности», попадание точки X левее x становится в пределе достоверным событием. Другими словами, случайная точка X никогда не может оказаться правее «плюс бесконечности». Куда бы эта точка ни попала, она всегда будет находиться левее «плюс бесконечности». Поэтому естественно предположить, что $\lim\limits_{x\to\infty} F(x)=1$.

3^0. *Если $F(x)$ — функция распределения, то*

$$\lim_{x\to-\infty} F(x)=0 \quad \textit{и} \quad \lim_{x\to\infty} F(x)=1.$$

□ Так как $F(x)$ — монотонная функция (см. свойство 2^0) и ограниченная (см. свойство 1^0), то, по известной из математического анализа теореме о монотонной и ограниченной функции, пределы $\lim\limits_{x\to\pm\infty} F(x)$ существуют. В силу предполагаемой непрерывности $F(x)$ можно записать

$$\lim_{x\to-\infty} F(x)=\lim_{x\to-\infty} P(X<x)=P(X<-\infty).$$

Так как событие $X<-\infty$ является невозможным, то $P(X<-\infty)=0$. Таким образом, $\lim\limits_{x\to-\infty} F(x)=0$. Рассуждая аналогично, получим

$$\lim_{x\to\infty} F(x)=\lim_{x\to\infty} P(X<x)=P(X<\infty).$$

Событие $X<\infty$ является достоверным. Тогда

$$P(X<\infty)=1 \quad \text{и} \quad \lim_{x\to\infty} F(x)=1. \ ■$$

4^0. *Вероятность любого отдельного значения непрерывной случайной величины равна нулю.*

□ Воспользуемся равенством (5.3) и устремим β к α. В результате получим

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

В левой части последнего равенства в пределе вместо вероятности попадания значения случайной величины в интервал $(\alpha;\ \beta)$ получим вероятность того, что случайная величина приняла отдельно взятое значение α, т. е. $P(X=\alpha)$. Значение предела в правой части равенства зависит от того, является ли функция $F(x)$ непрерывной в точке α или имеет в ней разрыв. В последнем случае указанный предел равен величине скачка функции $F(x)$ в точке α. Так как предполагается, что $F(x)$ всюду непрерывна, то

$$\lim_{\beta\to\alpha}[F(\beta)-F(\alpha)]=F(\alpha)-F(\alpha)=0.$$

Таким образом,

$$\lim_{\beta\to\alpha}P(\alpha\leqslant X<\beta)=\lim_{\beta\to\alpha}[F(\beta)-F(\alpha)]=P(X=\alpha)=0. \blacksquare$$

Поясним доказанное свойство. Ранее событие полагалось невозможным, если его вероятность равна нулю. Однако из свойства 4° функции $F(x)$ следует, что нулевой вероятностью могут обладать и возможные события, так как событие, заключающееся в том, что случайная величина примет конкретное значение, является возможным. Другими словами, появление любого отдельного значения случайной величины является возможным событием, несмотря на то что вероятность его появления равна нулю. Такие события (возможные, но с нулевой вероятностью) появляются при рассмотрении опытов, не сводящихся к схеме случаев, когда применение классического определения вероятности ограничено. Подобные ситуации выглядят парадоксально. Однако на самом деле ничего парадоксального нет в представлении о том, что определенное материальное тело, имеющее массу, состоит из материальных точек, каждая из которых не имеет массы, хотя сколь угодно малый объем этого тела имеет определенную, пусть очень малую, массу. Рассматриваемая ситуация аналогична приведенному примеру.

При непрерывном распределении вероятностей, т. е. когда функция распределения непрерывна, вероятность попадания значения непрерывной случайной величины на сколь угодно малый участок отлична от нуля, тогда как вероятность попадания в строго отдельную точку равна нулю. Очевидно, в данном случае нет смысла рассматривать вероятность принятия отдельного значения случайной величиной, но имеет смысл рассматривать вероятность попадания ее значения в интервал, пусть даже сколь угодно малый. Сказанное согласуется с практикой. Так, например, при изготовлении детали рабочий следит за тем, чтобы определенный размер не вышел за пределы допуска, зная, что вряд ли сможет точно выдержать размер, хотя последнее не исключается. Исследователя в данный момент будет интересовать вероятность попадания размера в определенные границы, а не вероятность совпадения размера изготовленной детали с требуемым по чертежу размером.

Теперь, воспользовавшись свойством 4°, докажем, что для непрерывной случайной величины выполняются следующие равенства:

$$P(\alpha\leqslant X<\beta)=P(\alpha<X<\beta)=P(\alpha\leqslant X\leqslant\beta)=P(\alpha<x\leqslant\beta).$$

Для этого достаточно доказать одно из них (остальные доказываются аналогично). Событие $\alpha\leqslant X<\beta$ представляет собой сумму двух несовместных событий $X=\alpha$ и $\alpha<X<\beta$. По теореме сложения имеем

$$P(\alpha\leqslant X<\beta)=P(X=\alpha)+P(\alpha<X<\beta).$$

Но согласно свойству 4°, $P(X=\alpha)=0$; тогда $P(\alpha\leqslant X<\beta)=$ $=P(\alpha<X<\beta)$ и формула (5.3) принимает вид

$$P(\alpha<X<\beta)=F(\beta)-F(\alpha). \qquad (5.4)$$

Эту формулу мы и будем использовать в дальнейшем.

Задача 5.3. Функция распределения случайной величины X задана выражением

$$F(x)=\begin{cases}0 \text{ при } x<-\pi/4,\\ a\sin(x-\pi/4)+1/2 \text{ при } -\pi/4\leqslant x\leqslant 3\pi/4,\\ 1 \text{ при } x\geqslant 3\pi/4.\end{cases}$$

Найти коэффициент a; вероятность попадания значения случайной величины X в результате опыта в интервал $(\pi/4;$ $3\pi/4)$; построить график функции $F(x)$.

△ При $x=3\pi/4$ функция $F(x)$ равна 1, т. е.

$$a\sin(x-\pi/4)+1/2=1$$

или

$$a\sin(\pi/2)+1/2=1,$$

откуда $a=1/2$.

Подставляя $\alpha=\pi/4$ и $\beta=3\pi/4$ в равенство (5.4), получаем

$$P(\pi/4<X<3\pi/4)=F(3\pi/4)-F(\pi/4)=(1/2)\sin(\pi/2)+1/2-$$
$$-(1/2)\sin 0-1/2=1/2.$$

График функции $y=\frac{1}{2}\sin(x-\pi/4)+1/2$ отличается от графика функции $y=\sin x$ тем, что он «сжат» по оси Oy в два раза, сдвинут вправо на $\pi/4$, поднят вверх на 1/2. Воспользовавшись этим замечанием, строим график $F(x)$ (рис. 5.8). ▲

§ 5.4. Плотность распределения вероятностей

Функция распределения вероятностей непрерывной случайной величины дает полную вероятностную характеристику ее поведения. Однако задание непрерывной случайной величины

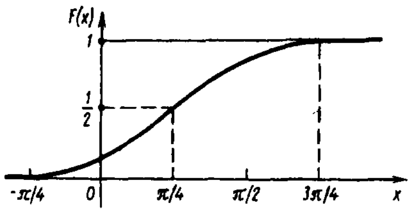


Рис. 5.8

с помощью функции распределения не является единственным. Ее можно задать с помощью другой функции, которая называется *дифференциальной функцией распределения* или *плотностью распределения вероятностей*. В некотором смысле эта функция «более удобная», чем интегральная функция $F(x)$. Используя функцию $F(x)$, трудно судить о характере распределения случайной величины в небольшой окрестности той или иной точки числовой оси. Решить эту задачу позволяет плотность распределения вероятностей.

Пусть имеем непрерывную случайную величину X с интегральной функцией распределения $F(x)$, относительно которой будем предполагать, что она непрерывна и дифференцируема в исследуемом интервале. Рассмотрим вероятность попадания значения случайной величины в элементарный участок $(x;\ x+\Delta x)$ длины Δx. Эту вероятность несложно вычислить, используя свойство 2^0 функции $F(x)$. Согласно формуле (5.4), искомая вероятность

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

т. е. равна приращению функции $F(x)$ на этом участке.

Определим теперь вероятность, которая приходится на единицу длины рассматриваемого участка, для чего разделим обе части последнего равенства на длину участка Δx. Имеем

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

Перейдем в полученном равенстве к пределу при $\Delta x\to 0$. В результате получим производную функции $F(x)$, которая существует, поскольку $F(x)$ предполагалась дифференцируемой:

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

Обозначим

$$F'(x)=f(x). \tag{5.5}$$

Определение. *Дифференциальной функцией распределения или плотностью распределения вероятностей называется первая производная интегральной функции распределения $F(x)$.*

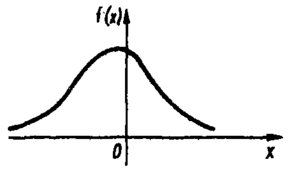


Рис. 5.9

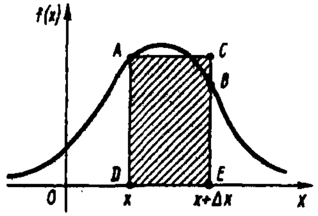


Рис. 5.10

Иногда дифференциальную функцию распределения $f(x)$ называют *плотностью* или *функцией плотности распределения вероятности*.

Заметим, что для характеристики распределения вероятностей значений дискретной случайной величины дифференциальная функция распределения неприменима хотя бы потому, что для существования $f(x)$ требуется непрерывность и дифференцируемость функции $F(x)$, а для дискретной случайной величины эти требования не выполняются.

График дифференциальной функции распределения $f(x)$ называется *кривой распределения* (рис. 5.9). Из математического анализа известно, что приращение функции $\Delta y=f(x+\Delta x)-f(x)$ приближенно равно дифференциалу функции dy. Запишем это приближенное равенство для функции $F(x)$. Имеем

$$F(x+\Delta x)-F(x)\approx dF(x)$$

или

$$F(x+\Delta x)-F(x)\approx F'(x)\,dx.$$

Так как $F'(x)=f(x)$ и $dx=\Delta x$, то

$$F(x+\Delta x)-F(x)\approx f(x)\Delta x.$$

Последнее равенство означает, что вероятность попадания значения случайной величины X в интервал $]x;\ x+\Delta x[$ приближенно равна произведению плотности вероятности в точке x на длину интервала Δx. С геометрической точки зрения это равенство можно истолковать так: вероятность попадания значения случайной величины в интервал $]x;\ x+\Delta x[$ приближенно равна площади прямоугольника с основанием Δx и высотой $f(x)$ (рис. 5.10). Из рис. 5.10 видно, что площадь заштрихованного прямоугольника, равная $f(x)\Delta x$, только приближенно равна площади криволинейной трапеции $DABE$. Как будет установлено ниже, площадь указанной криволинейной трапеции точно соответствует вероятности попадания значения случайной величины на элементарный участок $]x;\ x+\Delta x[$.

Величину $f(x)dx=f(x)\Delta x$ называют *элементом вероятности.*

Рассмотрим основные свойства дифференциальной функции распределения (функции плотности).

1°. *Для любых x дифференциальная функция распределения $f(x)$ неотрицательна, т. е. $f(x)\geq 0$.*

□ Доказательство непосредственно следует из определения функции $f(x)$ как первой производной функции $F(x)$, которая, в свою очередь, является неубывающей. Из математического анализа известно, что производная неубывающей функции неотрицательна, т. е.

$$F'(x)=f(x)\geq 0. \blacksquare$$

2°. *Для дифференциальной функции распределения имеет место равенство*

$$P(\alpha<X<\beta)=\int_{\alpha}^{\beta} f(x)\,dx. \quad (5.6)$$

□ Так как функция $F(x)$ является первообразной для функции $f(x)$ $(F'(x)=f(x))$, то из формулы (5.4) и формулы Ньютона—Лейбница следует, что

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

Формула (5.6) имеет простой геометрический смысл: вероятность $P(\alpha<X<\beta)$ численно равна площади криволинейной трапеции, ограниченной графиком функции $f(x)$, осью Ox и прямыми $x=\alpha$, $x=\beta$ (рис. 5.11).

Вернемся к рис. 5.10. Площадь криволинейной трапеции S_{DABE} на основании свойства 2° функции $f(x)$ равна

$$S_{DABE}=\int_{x}^{x+\Delta x} f(x)\,dx=F(x+\Delta x)-F(x)=P(x<X<x+\Delta x)\approx$$
$$\approx f(x)\Delta x.$$

3°. *Для дифференциальной функции распределения имеет место равенство*

$$\int_{-\infty}^{\infty} f(x)\,dx=1. \quad (5.7)$$

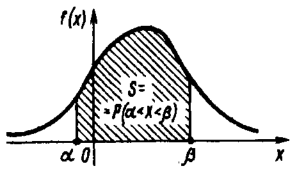


Рис. 5.11

□ Согласно определению несобственного интеграла по бесконечным пределам и свойству 3° функции $F(x)$, имеем

$$\int_{-\infty}^{\infty} f(x)\,dx=\lim_{\substack{\alpha\to-\infty\\ \beta\to\infty}}\int_{\alpha}^{\beta} f(x)\,dx=\lim_{\substack{\alpha\to-\infty\\ \beta\to\infty}}[F(\beta)-F(\alpha)]=$$

$$=\lim_{\beta\to\infty}F(\beta)-\lim_{\alpha\to-\infty}F(\alpha)=1-0=1. \blacksquare$$

Если интеграл $\int_{\alpha}^{\beta} f(x)dx$ выражает вероятность попадания значения случайной величины в интервал $]\alpha,\ \beta[$, то несобственный интеграл $\int_{-\infty}^{\infty} f(x)dx$ определяет вероятность такого попадания в интервал $]-\infty,\ \infty[$. С другой стороны, в результате опыта случайная величина обязательно примет какое-нибудь значение и это значение несомненно окажется в интервале $]-\infty,\ \infty[$, т. е. произойдет заведомо достоверное событие, вероятность которого равна единице.

Геометрически равенство (5.7) означает, что площадь, ограниченная осью Ox и кривой распределения, равна единице (рис. 5.12).

4°. *Для интегральной и дифференциальной функций распределения имеет место равенство*

$$F(x)=\int_{-\infty}^{x} f(t)\,dt. \quad (5.8)$$

□ Согласно определению несобственного интеграла и свойству 3° функции $F(x)$, имеем

$$\int_{-\infty}^{x} f(t)\,dt=\lim_{\alpha\to-\infty}\int_{\alpha}^{x} f(t)\,dt=\lim_{\alpha\to-\infty}[F(x)-F(\alpha)]=$$

$$=F(x)-\lim_{\alpha\to-\infty}F(\alpha)=F(x)-F(-\infty)=F(x)-0=F(x). \blacksquare$$

В заключение отметим, что если случайная величина принимает значения только в некотором интервале $]\alpha,\ \beta[$, то

$$\int_{\alpha}^{\beta} f(x)\,dx=1.$$

Задача 5.4. Используя условие задачи 5.3, найти дифференциальную функцию распределения и построить ее график.

△ Согласно определению дифференциальной функции распределения как первой производной интегральной функции распределения, имеем

$$F'(x)=f(x)=\begin{cases}0 \text{ при } x=-\pi/4,\\ (1/2)\cos(x-\pi/4) \text{ при } -\pi/4\leq x\leq 3\pi/4,\\ 0 \text{ при } x>3\pi/4.\end{cases}$$

Искомый график изображен на рис. 5.13. ▲

Задача 5.5. Случайная величина задана дифференциальной функцией распределения

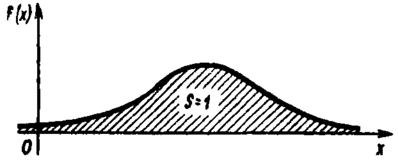


Рис. 5.12

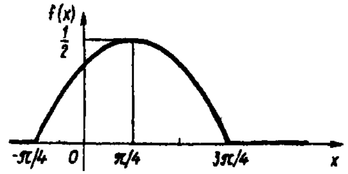


Рис. 5.13

$$f(x)=\begin{cases} 0 \text{ при } x<0, \\ (x-3)^2/9 \text{ при } 0\leqslant x\leqslant 3, \\ 0 \text{ при } x>3. \end{cases}$$

Найти интегральную функцию распределения $F(x)$ и построить графики функций $f(x)$ и $F(x)$.

△ Воспользуемся формулой (5.8). Имеем:
если $x<0$, то $f(x)=0$; следовательно,

$$F(x)=\int_{-\infty}^{x} 0\,dt=0;$$

если $0\leqslant x\leqslant 3$, то $f(x)=(x-3)^2/9$; следовательно,

$$F(x)=\int_{-\infty}^{0} 0\,dt+\int_{0}^{x}\frac{(t-3)^2}{9}dt=\left.\frac{(t-3)^3}{27}\right|_0^x=\frac{(x-3)^3}{27}+1;$$

если $x>3$, то $f(x)=0$; следовательно,

$$F(x)=\int_{-\infty}^{0} 0\,dt+\int_{0}^{3}\frac{(t-3)^2}{9}dt+\int_{3}^{x} 0\,dt=1.$$

Таким образом, искомая функция $F(x)$ имеет вид

$$F(x)=\begin{cases} 0 \text{ при } x<0, \\ (x-3)^3/27+1 \text{ при } 0\leqslant x\leqslant 3, \\ 1 \text{ при } x>3. \end{cases}$$

Графики функций $f(x)$ и $F(x)$ изображены на рис. 5.14 и 5.15 соответственно. ▲

§ 5.5. Числовые характеристики случайной величины

Для решения многих практических задач совсем необязательно знать все возможные значения случайной величины и соответствующие им вероятности, а достаточно указать отдельные числовые параметры, которые позволяют в удобной, компактной форме отразить существенные особенности случайной величины.

Эти характеристики случайной величины, являющиеся не функциями, а числами, называют *числовыми характеристиками*

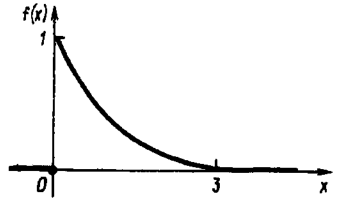


Рис. 5.14

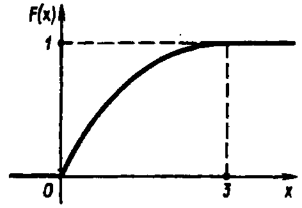


Рис. 5.15

случайной величины. Их назначение — в сжатой форме выразить наиболее важные черты распределения. К таким числовым характеристикам относятся математическое ожидание, дисперсия, моменты различных порядков и т. д. Рассмотрим некоторые наиболее важные числовые характеристики и изучим их свойства.

1. Математическое ожидание. Возможные значения случайной величины могут быть сосредоточены вокруг некоторого центра. Этот центр является некоторым средним значением случайной величины, вокруг которого группируются остальные ее значения. Для характеристики такой особенности распределения случайной величины служит математическое ожидание, которое иногда называют *центром распределения* или *средним значением* случайной величины.

Пусть имеется дискретная случайная величина X, принимающая значения $x_1, x_2, x_3, ..., x_n$ с вероятностями $p_1, p_2, p_3, ..., p_n$ соответственно.

Определение. *Математическим ожиданием MX дискретной случайной величины X называется сумма произведений всех возможных значений случайной величины на соответствующие вероятности появления этих значений, т. е.*

$$MX=\sum_{i=1}^{n} x_i p_i. \tag{5.9}$$

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

$$MX=\sum_{i=1}^{\infty} x_i p_i,$$

причем математическое ожидание существует, если ряд в правой части равенства сходится абсолютно.

Поясним формулу (5.9). Пусть проведено n испытаний, в которых случайная величина X приняла значения x_1, x_2, x_3, ..., x_k, причем значение x_1 появилось m_1 раз, значение x_2 появилось m_2 раз, ..., значение x_k появилось m_k раз. Очевидно, $m_1+m_2+m_3+...+m_k=n$. Найдем среднее арифметическое всех этих значений и обозначим его $\bar{X}$. Имеем

$$\bar{X}=\frac{x_1m_1+x_2m_2+x_3m_3+...+x_km_k}{m_1+m_2+m_3+...+m_k}$$

или

$$\bar{X}=\sum_{i=1}^{k} x_i\frac{m_i}{n}.$$

Заметим, что дробь m_i/n есть не что иное, как относительная частота появления значения x_i в n испытаниях, т. е. статистическая вероятность. Обозначим эту дробь $\hat{p}_i$. Учитывая, что $\hat{p}_i=m_i/n$, запишем теперь среднее арифметическое $\bar{X}$ в виде

$$\bar{X}=\sum_{i=1}^{k} x_i\hat{p}_i. \qquad (5.9')$$

Таким образом, формула (5.9) вполне объяснима, только следует помнить, что в этой формуле n — число всех возможных значений случайной величины; $x_1, x_2, x_3, ..., x_n$ — эти значения, а в формуле (5.9′) n — число испытаний или наблюдений над случайной величиной; $x_1, x_2, x_3, ..., x_k$ — наблюдаемые значения случайной величины.

Задача 5.6. Дискретная случайная величина X задана следующим рядом распределения:

X	2	5	8	19
P	0,2	0,3	0,4	0,1

Найти математическое ожидание этой случайной величины.

△ По формуле (5.9) имеем

$$MX=2\cdot 0{,}2+5\cdot 0{,}3+8\cdot 0{,}4+19\cdot 0{,}1=7. \ \blacktriangle$$

Задача 5.7. Найти математическое ожидание числа появлений события A в одном испытании, если $P(A)=p$.

△ Пусть X — число появлений события A в одном испытании. Эта величина может принять только два значения: $x_1=1$ (событие наступило) с вероятностью p и $x_2=0$ (событие не наступило) с вероятностью $1-p$. Искомое математическое ожидание

$$MX=1\cdot p+0(1-p)=p. \ \blacktriangle$$

Формулу (5.9) нельзя применять для непрерывной случайной величины. При введении понятия математического ожидания для этого случая воспользуемся понятием элемента вероятности $f(x)\Delta x$.

Предположим, что все возможные значения случайной величины X принадлежат отрезку $[a, b]$. Разобьем произвольным образом этот отрезок точками $a=x_0<x_1<x_2<x_3<...<x_n=b$ на n частичных отрезков. В каждом таком частичном отрезке, длина которого Δx_i, произвольно возьмем точку h_i, $i=1, 2, 3, ..., n$. Из § 5.4 известно, что вероятность попадания значения непрерывной случайной величины на отрезок Δx приближенно равна $f(x)\Delta x$ (элемент вероятности) и поэтому приближенно будем считать, что случайная величина X может принять n значений h_i на отрезке $[a, b]$ с вероятностями $f(h_i)\Delta x_i$. Теперь можно воспользоваться формулой (5.9), задающей математическое ожидание для дискретной случайной величины, заменив x_i на h_i, а p_i на $f(h_i)\Delta x_i$. В результате будем иметь

$$MX\approx\sum_{i=1}^{n} h_i f(h_i)\Delta x_i.$$

Переходя к пределу в правой части равенства при $n\to\infty$ и стремлении к нулю длины наибольшего из частичных отрезков, воспользовавшись понятием определенного интеграла, получаем

$$\int_a^b xf(x)\,dx.$$

О п р е д е л е н и е. *Математическим ожиданием MX непрерывной случайной величины X, возможные значения которой принадлежат отрезку $[a, b]$, называется определенный интеграл $\int_a^b xf(x)\,dx$, т. е.*

$$MX=\int_a^b xf(x)\,dx. \qquad (5.10)$$

Если возможные значения случайной величины распределены по всей оси Ox, то

$$MX=\int_{-\infty}^{\infty} xf(x)\,dx. \qquad (5.11)$$

Здесь предполагается, что несобственный интеграл $\int_{-\infty}^{\infty} xf(x)\,dx$ сходится абсолютно, т. е. существует.

Задача 5.8. Непрерывная случайная величина X задана дифференциальной функцией распределения

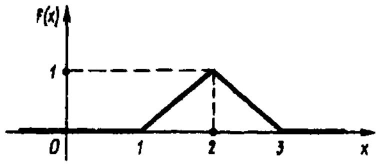


Рис. 5.16

$$f(x)=\begin{cases} 0 \text{ при } x<1, \\ x-1 \text{ при } 1\leqslant x<2, \\ -x+3 \text{ при } 2\leqslant x<3, \\ 0 \text{ при } x\geqslant 3. \end{cases}$$

Найти математическое ожидание этой случайной величины.

△ Так как все возможные значения случайной величины расположены на отрезке [1, 3], то воспользуемся формулой (5.10). Имеем

$$MX=\int_1^3 xf(x)\,dx=\int_1^2 x(x-1)\,dx+\int_2^3 x(3-x)\,dx=$$
$$=(x^3/3-x^2/2)\big|_1^2+(3x^2/2-x^3/3)\big|_2^3=2. \ ▲$$

Появление такого результата следовало ожидать, так как график функции $f(x)$ (рис. 5.16) симметричен относительно прямой $x=2$, т. е. значение случайной величины $X=2$ является ее средним значением.

Рассмотрим основные свойства математического ожидания, предварительно введя понятие независимых случайных величин.

Определение. *Две случайные величины называются независимыми, если закон распределения вероятностей одной из них не зависит от того, какие возможные значения приняла другая величина. В противном случае случайные величины называются зависимыми.*

Определение. *Несколько случайных величин называются взаимно независимыми, если закон распределения любой из них не зависит от того, какие значения приняли какие-либо другие из оставшихся величин.*

В дальнейшем слово «взаимно» будем опускать.

1^0. *Математическое ожидание алгебраической суммы двух случайных величин X и Y равно алгебраической сумме их математических ожиданий, т. е.*

$$M(X\pm Y)=M(X)\pm MY. \tag{5.12}$$

□ Пусть X и Y — дискретные случайные величины с математическими ожиданиями MX и MY соответственно, имеющие следующие ряды распределения:

X	x_1	x_2	x_3	...	x_i	...	x_n
P	p_1	p_2	p_3	...	p_i	...	p_n

Y	y_1	y_2	y_3	...	y_j	...	y_m
P	q_1	q_2	q_3	...	q_j	...	q_m

При этом $\sum_{i=1}^{n} p_i=1, \quad \sum_{j=1}^{m} q_j=1.$

Пусть случайные величины X и Y независимы и найдем вероятность появления значения $x_i\pm y_j$ случайной величины $Z=X\pm Y$. Для появления указанного значения необходимо, чтобы с вероятностью p_i появилось значение x_i случайной величины X и с вероятностью q_j появилось значение y_j случайной величины Y.

Таким образом, для появления значения $x_i\pm y_j$ необходимо одновременное наступление двух событий $X=x_i$ и $Y=y_j$. На основании теоремы умножения для независимых событий заключаем, что вероятность появления значения $x_i\pm y_j$ равна p_iq_j, т. е.

$$P(Z=x_i\pm y_j)=p_iq_j.$$

Ряд распределения случайной величины Z в этом случае имеет вид

$X\pm Y$	$x_1\pm y_1$	$x_2\pm y_1$	$x_1\pm y_2$	...	$x_i\pm y_j$	...	$x_n\pm y_m$
P	p_1q_1	p_2q_1	p_1q_2	...	p_iq_j	...	p_nq_m

Таким образом, математическое ожидание случайной величины $X\pm Y$ равно сумме произведений возможных значений этой величины на соответствующие вероятности:

$$M(X\pm Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i\pm y_j)p_iq_j=\sum_{i=1}^{n}\sum_{j=1}^{m}x_ip_iq_j\pm\sum_{i=1}^{n}\sum_{j=1}^{m}y_jp_iq_j=$$
$$=\sum_{i=1}^{n}x_ip_i\sum_{j=1}^{m}q_j\pm\sum_{j=1}^{m}y_jq_j\sum_{i=1}^{n}p_i=\sum_{i=1}^{n}x_ip_i\pm\sum_{j=1}^{m}y_jq_j=MX\pm MY. \ ■$$

Можно показать, что доказанное свойство верно и для зависимых случайных величин. Методом математической индукции рассмотренное свойство легко распространяется на произвольное конечное число слагаемых:

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

2^0. *Математическое ожидание произведения двух независимых случайных величин X и Y равно произведению их математических ожиданий, т. е.*

$$M(XY)=MX\cdot MY. \tag{5.13}$$

□ Пусть законы распределения дискретных случайных величин X и Y заданы рядами распределения, приведенными при доказательстве свойства 1^0. Очевидно, что ряд распределения случайной величины $Z=XY$ для независимых случайных величин имеет вид

XY	x_1y_1	x_2y_1	x_1y_2	...	x_iy_j	...	x_ny_m
P	p_1q_1	p_2q_1	p_1q_2	...	p_iq_j	...	p_nq_m

Согласно определению математического ожидания для дискретной случайной величины, имеем

$$M(XY)=\sum_{i=1}^{n}\sum_{j=1}^{m}x_iy_jp_iq_j=\sum_{i=1}^{n}x_ip_i\sum_{j=1}^{m}y_jq_j=MX\cdot MY. \blacksquare$$

Это свойство легко распространить на любое конечное число независимых случайных величин:

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

Следует отметить, что оба свойства доказаны здесь только для дискретных случайных величин. Однако они справедливы и для непрерывных случайных величин.

3^0. *Математическое ожидание постоянной величины равно самой постоянной, т. е.*

$$MC=C. \tag{5.14}$$

□ Постоянную C можно рассматривать как дискретную величину, которая с вероятностью $p=1$ принимает единственное значение C. Тогда по определению математического ожидания имеем

$$MC=C\cdot 1=C. \blacksquare$$

4^0. *Постоянный множитель случайной величины может быть вынесен за знак математического ожидания, т. е.*

$$M(CX)=CMX. \blacksquare \tag{5.15}$$

□ Рассматривая произведение CX как произведение двух независимых случайных величин и воспользовавшись свойствами 2^0 и 3^0 математического ожидания, получаем

$$M(CX)=MC\cdot MX=CMX. \blacksquare$$

5^0. *Математическое ожидание отклонения $X-MX$ случайной величины X от ее математического ожидания MX равно нулю, т. е.*

$$M(X-MX)=0. \tag{5.16}$$

□ Используя свойства 1^0 и 3^0 и учитывая, что математическое ожидание — величина постоянная, получаем

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

Поясним понятие отклонения случайной величины от ее математического ожидания. Математическое ожидание MX случайной величины X, как известно, вполне определенная для данной случайной величины постоянная. В результате опыта величина X принимает одно из ее возможных значений x. Разность $x-MX$ показывает, насколько отклонилось это значение случайной величины в данном опыте от MX. Очевидно, что разность $x-MX$ является случайной величиной. Эту величину будем обозначать $X-MX$ и называть *отклонением* случайной величины X от ее математического ожидания MX.

2. Дисперсия. На практике встречаются случайные величины, имеющие одинаковые математические ожидания, однако принимающие резко отличающиеся значения. У одних из этих величин отклонения значений от математического ожидания небольшие, а у других, наоборот, значительны, т. е. для одних рассеивание значений случайной величины вокруг математического ожидания мало, а для других оно велико.

Таким образом, математическое ожидание характеризует поведение случайной величины далеко не полностью. Приведем следующий пример.

Пример 5.8. Пусть дискретные случайные величины X и Y заданы следующими рядами распределения:

X	2	3	4	5
P	0,1	0,2	0,3	0,4

Y	−1	3	8	11
P	0,2	0,5	0,2	0,1

Найдем математическое ожидание этих величин. Имеем

$$MX=2\cdot 0{,}1+3\cdot 0{,}2+4\cdot 0{,}3+5\cdot 0{,}4=4;$$
$$MY=-1\cdot 0{,}2+3\cdot 0{,}5+8\cdot 0{,}2+11\cdot 0{,}1=4.$$

Отложим значения этих величин на числовых осях с одинаковым масштабом (рис. 5.17, *а*, *б*). Рассматриваемые случайные величины имеют одинаковые математические ожидания, равные 4. Однако рассеивание значений случайной величины X вокруг математического ожидания значительно меньше, чем у величины Y. ●

Приведенные рассуждения и пример свидетельствуют о целесообразности введения такой характеристики случайной величины, которая оценивала бы меру рассеивания значений

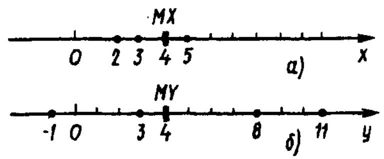


Рис. 5.17

случайной величины вокруг ее математического ожидания, тем более что на практике часто приходится оценивать такое рассеивание. Например, артиллеристам необходимо знать, как кучно лягут снаряды вблизи цели, по которой ведется стрельба.

Ранее уже отмечалось, что случайная величина $X-MX$ характеризует отклонение значений случайной величины от ее математического ожидания. Однако брать среднее значение (математическое ожидание) этой величины в качестве характеристики рассеивания не имеет смысла, так как оно равно нулю (свойство 5^0 математического ожидания). Последнее объясняется тем, что возможные значения $X-MX$ могут иметь как положительные, так и отрицательные знаки.

Избежать изменения знаков отклонений x_i-MX можно, если заменить их абсолютными значениями или возвести в квадрат. Замена отклонений их абсолютными величинами нецелесообразна, так как действия с абсолютными величинами, как правило, вызывают затруднения. Поэтому следует использовать величину $(X-MX)^2$ (точнее, ее среднее значение) для характеристики рассеивания значений случайной величины.

Определение. *Математическое ожидание квадрата отклонения случайной величины X от ее математического ожидания MX называют дисперсией случайной величины X и обозначают DX, т. е.*

$$DX=M(X-MX)^2. \tag{5.17}$$

В пользу использования дисперсии DX в качестве характеристики рассеяния значений случайной величины говорит следующий факт: DX обладает свойством минимальности. Это означает, что

$$DX=\min_{c}[M(X-c)^2]. \tag{5.18}$$

Действительно, найдем минимум функции $F(c)=M(X-c)^2$. Представим эту функцию в виде

$$F(c)=M(X-c)^2=M[(X-MX)-(c-MX)]^2=M[(X-MX)^2+(c-MX)^2-$$
$$-2(X-MX)(c-MX)]=M(X-MX)^2+M(c-MX)^2+M[-2(X-MX)(c-MX)]=$$
$$=DX+M(c-MX)^2-2(c-MX)M(X-MX)=DX+M(c-MX)^2.$$

Итак,

$$F(c)=DX+M(c-MX)^2.$$

Минимум функции $F(c)$ достигается при $c=MX$, но тогда математическое ожидание

$$M(X-c)^2=M(X-MX)^2=DX,$$

что и требовалось доказать.

Вернемся к формуле (5.17). Очевидно, что законы распределения вероятностей случайных величин X и $(X-MX)^2$ одинаковы. Если X является дискретной случайной величиной,

то по определению математического ожидания для случайной величины $(X-MX)^2$ имеем

$$DX=\sum_{i=1}^{n}(x_i-MX)^2 p_i. \tag{5.19}$$

Для непрерывной случайной величины с функцией плотности $f(x)$

$$DX=\int_{-\infty}^{\infty}(x-MX)^2 f(x)\,dx. \tag{5.19'}$$

Если случайная величина и ее математическое ожидание имеют одну и ту же размерность, то дисперсия имеет размерность квадрата случайной величины. Этого недостатка можно избежать, если воспользоваться средним квадратическим отклонением σ_X случайной величины, которым является арифметический корень из дисперсии. Таким образом, $\sigma_X=\sqrt{DX}$ — среднее квадратическое отклонение.

Задача 5.9. Случайная величина задана следующим рядом распределения:

X	2	4	7	10	12
P	0,1	0,2	0,4	0,2	0,1

Найти математическое ожидание и дисперсию этой величины.

Так как случайная величина является дискретной, то для вычисления MX воспользуемся формулой (5.9), а для нахождения DX — формулой (5.19). Для удобства все результаты вычислений сведем в табл. 5.2.

Таблица 5.2

x_i	p_i	$x_i p_i$	x_i-MX (x_i-7)	$(x-MX)^2$ $((x_i-7)^2)$	$(x_i-MX)^2 p_i$ $((xi-7)^2 p_i)$	x_i^2	$x_i^2 p_i$
1	2	3	4	5	6	7	8
2	0,1	0,2	−5	25	2,5	4	0,4
4	0,2	0,8	−3	9	1,8	16	3,2
7	0,4	2,8	0	0	0	49	19,6
10	0,2	2,0	3	9	1,8	100	20,0
12	0,1	1,2	5	25	2,5	144	14,4
Σ	1	7,0			8,6		57,6

Очевидно, что сумма чисел, находящихся в третьем столбце, равна MX, а в шестом — DX. Тогда $MX=7$, $DX=8{,}6$. ▲

Задача 5.10. Случайная величина задана дифференциальной функцией распределения

$$f(x)=\begin{cases} 0 \text{ при } x<-1, \\ 3x^2 \text{ при } -1\leqslant x\leqslant 0, \\ 0 \text{ при } x>0. \end{cases}$$

Найти дисперсию этой случайной величины.

△ Предварительно найдем математическое ожидание, воспользовавшись формулой (5.11). Имеем

$$MX=\int_{-\infty}^{\infty} xf(x)\,dx=\int_{-1}^{0} 3x^3\,dx=3x^4/4\,|_{-1}^{0}=-3/4.$$

Далее по формуле (5.19′) получаем

$$DX=\int_{-\infty}^{\infty}(x-MX)^2 f(x)\,dx=\int_{-1}^{0}(x+3/4)^2\,3x^2\,dx=3\int_{-1}^{0}(x^4+(3/2)x^3+$$
$$+(9/16)\,x^2)\,dx=3(x^5/5+3x^4/8+3x^3/16)|_{-1}^{0}=0{,}0375.\ ▲$$

Рассмотрим основные свойства дисперсии.

1^0. *Дисперсия алгебраической суммы двух независимых случайных величин X и Y равна сумме дисперсий этих величин, т. е.*

$$D(X\pm Y)=DX+DY. \qquad (5.20)$$

□ Запишем для случайной величины $X\pm Y$ равенство (5.17). Имеем

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

Воспользовавшись свойством 1^0 математического ожидания, получаем

$$D(X\pm Y)=M\left[X\pm Y-(MX\pm MY)\right]^2=M\left[(X-MX)\pm(Y-MY)\right]^2=$$
$$=M\left[(X-MX)^2\pm2(X-MX)(Y-MX)+(Y-MY)^2\right]=$$
$$=M(X-MX)^2\pm M\left[2(X-MX)(Y-MY)\right]+M(Y-MY)^2.$$

Рассмотрим отдельно слагаемые в правой части полученного равенства. Используя свойства 2^0, 4^0 и 5^0 математического ожидания, для второго из слагаемых имеем

$$M\left[2(X-MX)(Y-MY)\right]=2M(X-MX)\,M(Y-MY)=0.$$

Первое и третье слагаемые, по определению, представляют собой соответствующие дисперсии DX и DX. Тогда можно записать

$$D\,(X\pm Y)=DX+DY.$$

Заметим, что дисперсия алгебраической суммы величин равна не алгебраической сумме, а просто сумме соответствующих дисперсий.

Доказанное свойство легко распространить на любое конечное число независимых случайных величин.

2°. *Дисперсия постоянной величины C равна нулю, т. е.* $DC=0$.

□ Так как $MC=C$ (см. свойство 3°), то, согласно определению, имеем

$$DC=M(C-MC)^2=M(C-C)^2=M(0)=0\ ■.$$

Этот результат следоаало ожидать, так как постоянная величина принимает всего одно значение, что говорит об отсутствии рассеивания значений велнчины.

3^0. *Постоянный множитель C случайной величины X можно выносить за знак дисперсии, предварительно возведя его в квадрат, т. е.*

$$D(CX)=C^2DX. \qquad (5.21)$$

□ Так как постоянный множитель можно выносить за знак математического ожидания (см. свойство 4^0), то по определению дисперсии имеем

$$D(CX)=M\left[CX-M(CX)\right]^2=M\left[CX-CMX\right]^2=$$
$$=M\left[C(X-MX)\right]^2=M\left[C^2(X-MX)^2\right]=C^2M\,(X-MX)^2=$$
$$=C^2DX.\ ■$$

4^0. *Дисперсия случайной величины X равна разности между математическим ожиданием квадрата случайной величины и квадратом ее математического ожидания, т. е.*

$$D(X)=M(X^2)-[MX]^2. \qquad (5.22)$$

□ По определению дисперсии имеем

$$DX=M(X-MX)^2=M\left[X^2-2XMX+(MX)^2\right]=$$
$$=M(X^2)-M\left[2XMX\right]+M(MX)^2.$$

Так как $2MX$ и $(MX)^2$ — постоянные величины, то, используя свойства 3^0 и 4^0 математического ожидания, имеем

$$DX=M(X^2)-2MX\,MX+(MX)^2=M(X^2)-(MX)^2.\ ■$$

При расчетах формула (5.22) более удобна, чем формула (5.17).

Задача 5.11. Используя формулу (5.22), вычислить дисперсию по данным задачи 5.9.

△ Результаты вычислений представлены в табл. 5.2 (столбцы 1, 2, 3, 7, 8). Для вычисления $M(X^2)$ следует воспользоваться формулой

$$M(X^2)=\sum_{i=1}^{n} x_i^2 p_i.$$

Тогда сумма чисел, находящихся в восьмом столбце, совпадает с $M(X^2)$. По формуле (5.22) находим

$$DX=57{,}6-7^2=8{,}6.\ ▲$$

Далее нам потребуются некоторые соотношения. Докажем их в качестве следствий свойств математического ожидания и дисперсии.

Ранее было введено понятие отклонения случайной величины от ее математического ожидания. Это отклонение

$X-MX$ называют *центрированной случайной величиной*. Было показано (см. свойство 5°), что математическое ожидание центрированной случайной величины равно нулю. Найдем дисперсию этой случайной величины. На основании свойства 1° дисперсии имеем

$$D(X-MX)=DX+D(MX),$$

а так как $D(MX)=0$ (см. свойство 2° дисперсии), то

$$D(X-MX)=DX.$$

Следствие 1. *Дисперсии случайных величин X и $X-MX$ равны между собой.*

Иногда бывает удобно использовать безразмерные центрированные случайные величины. Величину $X-MX$ разделим на среднее квадратическое отклонение σ_X, имеющее ту же размерность. Вновь полученную случайную величину будем называть *стандартной случайной величиной*.

Обозначим эту величину Z. Таким образом,

$$Z=\frac{X-MX}{\sigma_X}.$$

Найдем математическое ожидание и дисперсию величины Z. Имеем

$$MZ=M\left[\frac{X-MX}{\sigma_X}\right]=\frac{1}{\sigma_X}M(X-MX)=0,$$

$$DZ=D\left[\frac{X-MX}{\sigma_X}\right]=\frac{1}{\sigma_X^2}D(X-MX)=\frac{DX}{\sigma_X^2}.$$

Так как, по определению, $\sigma_X=\sqrt{DX}$, то

$$DZ=\frac{DX}{DX}=1.$$

Следствие 2. *Математическое ожидание стандартной случайной величины $Z=\frac{X-MX}{\sigma_X}$ равно* 0.

Следствие 3. *Дисперсия стандартной случайной величины $Z=\frac{X-MX}{\sigma_X}$ равна* 1.

В заключение отметим, что центрирование случайной величины геометрически означает перенос начала координат в точку с абсциссой, равной математическому ожиданию.

3. Начальные и центральные моменты. Прежде чем перейти к изучению указанных характеристик, рассмотрим пример.

Пример 5.9. Пусть случайная величина X задана следующим рядом распределения:

X	-150	3	4	5
P	0,02	0,18	0,6	0,2

Найдем математическое ожидание MX. Имеем

$$MX=-150\cdot 0{,}02+3\cdot 0{,}18+4\cdot 0{,}6+5\cdot 0{,}2=0{,}94.$$

Составим ряд распределения случайной величины X^2:

X^2	22500	9	16	25
P	0,02	0,18	0,6	0,2

Найдем математическое ожидание X^2. Имеем

$$M(X^2)=22500\cdot 0{,}2+9\cdot 0{,}18+16\cdot 0{,}6+25\cdot 0{,}2=466{,}22.$$

Нетрудно заметить, что $M(X^2)$ значительно больше MX. Это можно объяснить тем, что значение $x=-150$, намного отличающееся от остальных значений, при возведении в квадрат резко возросло. Однако вероятность этого значения (0,02) мала по сравнению с вероятностями других значений. ●

Таким образом, переходя от MX к $M(X^2)$, можно учесть наличие таких значений случайной величины, которые велики по абсолютной величине, но вероятность их появления мала, т. е. учесть их влияние на математическое ожидание. При переходе к более высоким степеням X это влияние только усиливается. Приведенный пример подтверждает целесообразность использования для этой цели математического ожидания целой положительной степени случайной величины.

Определение. *Начальным моментом k-го порядка ν_k случайной величины X называется математическое ожидание величины X^k, т. е.*

$$\nu_k=M(X^k). \tag{5.23}$$

Для дискретной случайной величины

$$\nu_k=\sum_{i=1}^{n}x_i^k p_i. \tag{5.24}$$

Для непрерывной случайной величины

$$\nu_k=\int_{-\infty}^{\infty}x^k f(x)\,dx. \tag{5.25}$$

Ранее уже отмечалось, что величина $X-M(X)$ связана с рассеиванием значений случайной величины вокруг ее математического ожидания. Рассмотрим характеристики, в определении которых используется эта величина.

Определение. *Центральным моментом* k-го *порядка* μ_k *случайной величины* X *называется математическое ожидание величины* $(X-MX)^k$, *т. е.*

$$\mu_k = M(X-MX)^k. \tag{5.26}$$

Для дискретной случайной величины

$$\mu_k = \sum_{i=1}^{n} (x_i - MX)^k p_i. \tag{5.27}$$

Для непрерывной случайной величины

$$\mu_k = \int_{-\infty}^{\infty} (x_i - MX)^k f(x)\, dx. \tag{5.28}$$

Воспользовавшись определениями и свойствами математического ожидания и дисперсии, можно показать, что

$$\nu_1 = MX;\quad \nu_2 = M(X^2);$$
$$\mu_1 = 0;\quad \mu_2 = DX = \nu_2 - \nu_1^2,$$
$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3,\quad \mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4. \tag{5.29}$$

В определении центральных моментов используются отклонения случайной величины от ее математического ожидания (центра). Поэтому моменты называются *центральными*. В определении начальных моментов также используются отклонения случайной величины. Однако в этом случае отклонение берется не от математического ожидания, а от точки, абсцисса которой равна нулю, являющейся началом координат. Поэтому моменты называют *начальными*. Моменты могут рассматриваться не только относительно начала координат или математического ожидания, но и относительно произвольной точки $x=a$:

$$\gamma_k = M(X-a)^k.$$

Предложим, что распределение случайной величины симметрично относительно математического ожидания. Тогда все центральные моменты нечетного порядка равны нулю. Это можно объяснить тем, что для каждого положительного значения величины $X-MX$ найдется (в силу симметричности распределения относительно MX) равное ему по абсолютной величине отрицательное значение этой величины, причем их вероятности будут одинаковыми. Поэтому при $k=2n+1$ сумма $\mu_k = \sum_{i=1}^{n} (x_i - MX)^k p_i$ будет равна нулю. То же самое можно сказать и про интеграл

$$\mu_k = \int_{-\infty}^{\infty} (x - MX)^k f(x)\, dx,$$

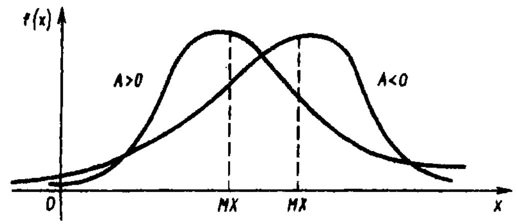


Рис. 5.18

который равен нулю как интеграл в симметричных пределах от нечетной функции, в чем можно убедиться предварительно выполнив замену переменной по формуле $t=x-MX$. Если центральный момент нечетного порядка не равен нулю, то это говорит об асимметричности распределения и чем больше момент, тем больше асимметрия. Поэтому в качестве характеристики асимметрии распределения разумнее всего взять какой-нибудь нечетный центральный момент. Так как центральный момент первого порядка всегда равен нулю, то целесообразно для этой цели использовать центральный момент третьего порядка.

Определение. *Коэффициентом асимметрии* A *называется величина* μ_3/σ_X^3, *т. е.*

$$A = \mu_3/\sigma_X^3. \tag{5.30}$$

Если коэффициент асимметрии отрицательный, то это говорит о большем влиянии на величину μ_3 отрицательных отклонений. В этом случае кривая распределения (рис. 5.18) более полога слева от MX. Если коэффициент A положительный, а значит, преобладает влияние положительных отклонений, то кривая распределения (рис. 5.18) более полога справа.

Как известно, второй центральный момент (дисперсия) служит для характеристики рассеивания значений случайной величины вокруг ее математического ожидания. Если этот момент для некоторой случайной величины достаточно большой, т. е. рассеивание велико, то соответствующая кривая распределения более пологая, чем кривая распределения случайной величины, имеющей меньший момент второго порядка. Однако нормированный момент $\mu_2 = DX$ не может служить характеристикой «плосковершинности» или «островершинности» распределения потому, что для любого распределения $\mu_2/\sigma_X^2 = DX/(\sqrt{DX})^2 = 1$. В этом случае следует использовать центральный момент четвертого порядка.

Определение. *Эксцессом* E *называется величина*

$$E = \mu_4/\sigma_X^4 - 3. \tag{5.31}$$

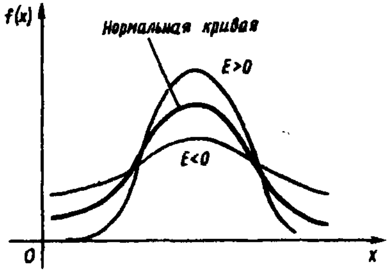


Рис. 5.19

Можно показать, что для наиболее распространенного в природе нормального закона распределения, который будет рассматриваться в следующей главе, отношение $\mu_4/\sigma_X^4=3$. Поэтому эксцесс, заданный формулой (5.31), служит для сравнения данного распределения с нормальным, у которого эксцесс равен нулю. Если для данного распределения эксцесс положительный, то это означает, что соответствующая кривая распределения более островершинная по сравнению с кривой нормального распределения. Распределения с отрицательным эксцессом имеют более «плосковершинные» кривые распределения по сравнению с нормальным (рис. 5.19).

○ **Задача 5.12.** Случайная величина задана функцией

$$f(x)=\begin{cases} 0 & \text{при } x<-2, \\ (-1/4)x^3 & \text{при } -2\leqslant x\leqslant 0, \\ 0 & \text{при } x>0. \end{cases}$$

Найти коэффициент асимметрии и эксцесс.

Предварительно вычислим начальные моменты до четвертого порядка. Имеем:

$$\nu_1=MX=-\frac{1}{4}\int_{-2}^{0}x^4dx=-\frac{1}{4}\frac{x^5}{5}\bigg|_{-2}^{0}=-1{,}6;$$

$$\nu_2=M(X^2)=-\frac{1}{4}\int_{-2}^{0}x^5dx=-\frac{1}{4}\frac{x^6}{6}\bigg|_{-2}^{0}=8/3\approx 2{,}67;$$

$$\nu_3=M(X^3)=-\frac{1}{4}\int_{-2}^{0}x^6dx=-\frac{1}{4}\frac{x^7}{7}\bigg|_{-2}^{0}=-32/7\approx 4{,}57;$$

$$\nu_4=M(X^4)=-\frac{1}{4}\int_{-2}^{0}x^7dx=-\frac{1}{4}\frac{x^8}{8}\bigg|_{-2}^{0}=8.$$

Теперь, воспользовавшись формулами (5.29), найдем центральные моменты:

$$\mu_2=DX=\nu_2-\nu_1^2\approx 2{,}67-(1{,}6)^2\approx 0{,}11;$$

$$\mu_3=\nu_3-3\nu_1\nu_2+2\nu_1^3\approx -4{,}57+3\cdot 1{,}6\cdot 2{,}67-2\cdot(1{,}6)^3\approx 0{,}054;$$

$$\mu_4=\nu_4-4\nu_1\nu_3+6\nu_1^2\nu_2-3\nu_1^4\approx 8-4\cdot 1{,}6\cdot 4{,}57+6(1{,}6)^2\cdot 2{,}67-3(1{,}6)^4\approx 0{,}1024.$$

Отсюда следует, что

$$\sigma_X=\sqrt{DX}\approx 0{,}33;\ \ \sigma_X^3\approx 0{,}036;\ \ \sigma_X^4\approx 0{,}0121.$$

Далее имеем

$$A=\mu_3/\sigma_X^3\approx 1{,}5;\ \ E=\mu_4/\sigma_X^4-3\approx 5{,}46.$$

ГЛАВА 6. ВИДЫ РАСПРЕДЕЛЕНИЙ

§ 6.1. Равномерное распределение

На практике встречаются случайные величины, о которых заранее известно, что они могут принять какое-либо значение в строго определенных границах, причем в этих границах все значения случайной величины имеют одинаковую вероятность (обладают одной и той же плотностью вероятностей).

Например, при поломке часов остановившаяся минутная стрелка будет с одинаковой вероятностью (плотностью вероятности) показывать время, прошедшее от начала данного часа до поломки часов. Это время является случайной величиной, принимающей с одинаковой плотностью вероятности значения, которые не выходят за границы, определенные продолжительностью одного часа. К подобным случайным величинам относится также и погрешность округления. Про такие величины говорят, что они распределены равномерно, т. е. имеют равномерное распределение.

Определение. *Непрерывная случайная величина X имеет равномерное распределение на отрезке $[a,b]$, если на этом отрезке плотность распределения вероятности случайной величины постоянна, т. е. если дифференциальная функция распределения $f(x)$ имеет следующий вид*:

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

Иногда это распределение называют *законом равномерной плотности*. Про величину, которая имеет равномерное распределение на некотором отрезке, будем говорить, что она распределена равномерно на этом отрезке. График плотности вероятности $f(x)$ равномерного распределения изображен на рис. 6.1. Найдем значение постоянной c. Так как площадь, ограниченная кривой распределения и осью Ox, равна 1, то

$$\int_{-\infty}^{\infty} f(x)\,dx=\int_a^b c\,dx=c(b-a)=1,$$

откуда $c=1/(b-a)$.

Теперь функцию $f(x)$ можно представить в виде

$$f(x)=\begin{cases} 0 & \text{при } x<a, \\ 1/(b-a) & \text{при } a\leqslant x\leqslant b, \\ 0 & \text{при } x>b. \end{cases} \qquad (6.1)$$

Построим функцию распределения $F(x)$, для чего найдем выражение $F(x)$ на интервале $[a, b]$. По формуле (5.8) имеем

$$F(x)=\int_{-\infty}^{x} f(t)\,dt=\int_a^x \frac{1}{b-a}\,dt=\left.\frac{t}{b-a}\right|_a^x=\frac{x-a}{b-a}.$$

При $x<a$ функция $F(x)=0$ и $F(x)=1$ при $x>b$. Таким образом,

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

График функции $F(x)$ приведен на рис. 6.2.

Найдем основные числовые характеристики рассматриваемой случайной величины.

По формуле (5.10) находим математическое ожидание:

$$MX=\int_a^b x f(x)\,dx=\int_a^b \frac{x}{b-a}\,dx=\frac{b+a}{2}.$$

Таким образом, математическое ожидание случайной величины, равномерно распределенной на отрезке $[a, b]$, совпадает с серединой этого отрезка.

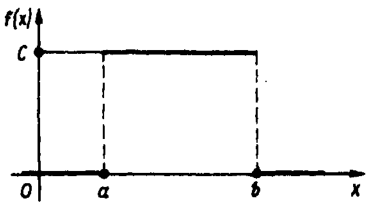


Рис. 6.1

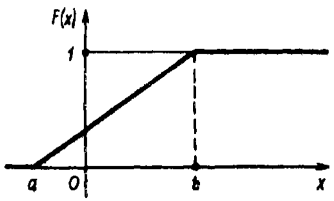


Рис. 6.2

По формуле (5.19′) находим дисперсию:

$$DX=\int_{-\infty}^{\infty}(x-MX)^2 f(x)\,dx=\int_a^b\left(x-\frac{b+a}{2}\right)^2\frac{1}{b-a}\,dx=\frac{(b-a)^2}{12},$$

откуда сразу же следует, что среднее квадратическое отклонение

$$\sigma_X=\frac{b-a}{2\sqrt{3}}.$$

Так как исследуемое распределение симметрично относительно математического ожидания, то для него центральные моменты, имеющие нечетный порядок, равны нулю, а значит, коэффициент асимметрии тоже равен нулю, т. е.

$$A=\mu_3/\sigma_X^3=0.$$

Для определения эксцесса находим центральный момент четвертого порядка:

$$\mu_4=\int_{-\infty}^{\infty}(x-MX)^4 f(x)\,dx=\int_a^b\left(x-\frac{a+b}{2}\right)^4\frac{1}{b-a}\,dx=\frac{(b-a)^4}{80}.$$

Тогда

$$E=\mu_4/\sigma_X^4-3=\frac{(b-a)^4\cdot 144}{80(b-a)^4}-3=-1{,}2.$$

Найдем теперь вероятность попадания значения случайной величины, имеющей равномерное распределение, на интервал (α, β), принадлежащий целиком отрезку $[a, b]$. По формуле (5.6) имеем

$$P(\alpha<X<\beta)=\int_\alpha^\beta f(x)\,dx=\int_\alpha^\beta\frac{dx}{b-a}=\frac{\beta-\alpha}{b-a},$$

т. е. вычисленная вероятность равна отношению длины отрезка $[\alpha, \beta]$ к длине отрезка $[a, b]$. Геометрически эта вероятность представляет собой площадь прямоугольника, заштрихованного на рис. 6.3.

В заключение отметим, что числа a и b, которые называются *параметрами распределения*, однозначно определяют равномерное распределение.

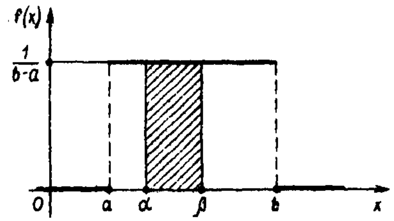


Рис. 6.3

§ 6.2. Нормальное распределение

Одним из наиболее часто встречающихся распределений является нормальное распределение. Оно играет большую роль в теории вероятностей и занимает среди других распределений особое положение. Нормальный закон распределения является предельным законом, к которому приближаются другие законы распределения при часто встречающихся аналогичных условиях.

Если предоставляется возможность рассматривать некоторую случайную величину как сумму достаточно большого числа других случайных величин, то данная случайная величина обычно подчиняется нормальному закону распределения. Суммируемые случайные величины могут подчиняться каким угодно распределениям, но при этом должно выполняться условие их независимости (или слабой зависимости). При соблюдении некоторых не очень жестких условий указанная сумма случайных величин подчиняется приближенно нормальному закону распределения и тем точнее, чем большее количество величин суммируется.

Из теоремы Ляпунова (см. § 7.5) следует, что ни одна из суммируемых случайных величин не должна резко отличаться от других, т. е. каждая из них должна играть в общей сумме примерно одинаковую роль и не иметь исключительно большую по сравнению с другими величинами дисперсию.

Для примера рассмотрим изготовление некоторой детали на станке-автомате. Размеры изготовленных деталей несколько отличаются от требуемых. Это отклонение размеров от стандарта вызывается различными причинами, которые более или менее независимы друг от друга. К ним могут относиться: неравномерный режим обработки детали; неоднородность обрабатываемого материала; неточность установки заготовки в станке; износ режущего инструмента и деталей станков; упругие деформации узлов станка; состояние микроклимата в цехе; колебание напряжения в электросети и т. д. Каждая из перечисленных и подобных им причин влияет на отклонение размера изготовляемой детали от стандарта. Таким образом, общее отклонение размера, фиксируемое измерительным прибором, является суммой большего числа отклонений, обусловленных различными причинами. Если ни одна из этих причин не является доминирующей, то суммарное отклонение является случайной величиной, имеющей нормальный закон распределения.

Так как нормальному закону подчиняются только непрерывные случайные величины, то это распределение можно задать в виде плотности распределения вероятности.

Определение. *Непрерывная случайная величина X имеет нормальное распределение (распределена по нормальному закону), если плотность распределения вероятности $f(x)$ имеет вид*

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}, \tag{6.2}$$

где a и σ — некоторые постоянные, называемые параметрами нормального распределения.

На основании равенства (5.8) функция распределения $F(x)$ в рассматриваемом случае принимает вид

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(t-a)^2}{2\sigma^2}}dt. \tag{6.3}$$

Как видно из формул (6.2) и (6.3), нормальное распределение определяется двумя параметрами: a и σ.

Определим основные числовые характеристики нормального распределения.

Воспользовавшись формулой (5.11), имеем

$$MX=\int_{-\infty}^{\infty}xf(x)\,dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{(x-a)^2}{2\sigma^2}}dx.$$

Введем новую переменную u по формуле $u=\frac{x-a}{\sigma}$. Тогда $x=\sigma u+a$, а $dx=\sigma du$, причем пределы интегрирования остаются теми же, в чем нетрудно убедиться. Итак, получаем

$$MX=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma u+a)e^{-u^2/2}du=\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}ue^{-u^2/2}du+\frac{a}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-u^2/2}du.$$

Первый из интегралов в правой части полученного равенства равен нулю как интеграл от нечетной функции в симметричных пределах, а второй интеграл (так называемый интеграл Пуассона) равен $\sqrt{2\pi}$, т. е.

$$\int_{-\infty}^{\infty}e^{-u^2/2}du=\sqrt{2\pi}.$$

Тогда

$$MX=\frac{a}{\sqrt{2\pi}}\sqrt{2\pi}=a.$$

Таким образом, параметр a нормального распределения равен математическому ожиданию соответствующей случайной величины, т. е.

$$MX=a. \tag{6.4}$$

Для вычисления дисперсии воспользуемся формулой (5.19), учитывая, что $MX=a$. Имеем

$$DX=\int_{-\infty}^{\infty}(x-MX)^2 f(x)\,dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-a)^2 e^{-\frac{(x-a)^2}{2\sigma^2}}\,dx.$$

Введем переменную u по формуле $u=\frac{x-a}{\sigma}$. В результате получаем

$$DX=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\sigma^2u^2e^{-u^2/2}\sigma\,du=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}u^2e^{-u^2/2}\,du.$$

Далее интегрируем по частям, полагая $u=t$, а $ue^{-u^2/2}du=dv$, откуда $dt=du$, $v=-e^{-u^2/2}$. Тогда

$$DX=\frac{\sigma^2}{\sqrt{2\pi}}\left(-ue^{-u^2/2}\Big|_{-\infty}^{\infty}+\int_{-\infty}^{\infty}e^{-u^2/2}\,du\right).$$

Первое из слагаемых в скобках равно нулю, так как $\lim\limits_{u\to\pm\infty}u/e^{+u^2/2}=0$ (по правилу Лопиталя), а второе равно $\sqrt{2\pi}$. Следовательно,

$$DX=\frac{\sigma^2}{\sqrt{2\pi}}\sqrt{2\pi}=\sigma^2.$$

Итак, дисперсия нормально распределенной случайной величины X равна σ^2, т. е.

$$DX=\sigma^2, \tag{6.5}$$

а σ не что иное, как среднее квадратическое отклонение, поскольку

$$\sigma_X=\sqrt{DX}=\sqrt{\sigma^2}=\sigma.$$

После вычисления математического ожидания и дисперсии становится ясным вероятностный смысл параметров a и σ нормального распределения. В дальнейшем дисперсию случайной величины, распределенной нормально, будем обозначать σ^2, а математическое ожидание — a.

Параметры a и σ нормального распределения имеют простую геометрическую интерпретацию, для выяснения которой исследуем поведение функции $f(x)$. В дальнейшем кривую нормального распределения (график функции $f(x)$) будем называть *нормальной кривой*.

1^0. *Областью определения функции $f(x)$ является вся числовая ось.*

□ Доказательство следует из формулы (6.2). ■

2^0. *Функция $f(x)$ может принимать только положительные значения, т. е. $f(x)>0$.*

Доказательство следует из того, что показательная функция $e^{-\frac{(x-a)^2}{2\sigma^2}}$ не может быть отрицательной, а параметр $\sigma>0$. ■

3^0. *Предел функции $f(x)$ при неограниченном возрастании $|x|$ равен нулю, т. е. ось OX является горизонтальной асимптотой графика функции.*

□ Действительно,

$$\lim_{x\to\pm\infty}f(x)=\lim_{x\to\pm\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}=\frac{1}{\sigma\sqrt{2\pi}}\lim_{x\to\pm\infty}\frac{1}{e^{\frac{(x-a)^2}{2\sigma^2}}}=0. \ ■$$

4^0. *Функция $f(x)$ имеет в точке $x=a$ максимум, равный $1/(\sigma\sqrt{2\pi})$.*

□ Найдем первую производную функции и приравняем ее нулю:

$$f'(x)=-\frac{(x-a)}{\sigma^3\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}=0.$$

Из уравнения следует, что точка $x=a$ является критической точкой. Найдем вторую производную функции. Имеем

$$f''(x)=-\frac{1}{\sigma^3\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}+\frac{(x-a)^2}{\sigma^5\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}=\frac{1}{\sigma^3\sqrt{2\pi}}\left(\frac{(x-a)^2}{\sigma^2}-1\right)e^{-\frac{(x-a)^2}{2\sigma^2}}$$

Определим знак второй производной в точке $x=a$:

$$f''(a)=\frac{1}{\sigma^3\sqrt{2\pi}}(-1)<0.$$

Следовательно, по второму правилу нахождения экстремума в точке $x=a$ имеет место максимум, величина которого определяется равенством

$$f(a)=\frac{1}{\sigma\sqrt{2\pi}}e^0=\frac{1}{\sigma\sqrt{2\pi}}. \ ■$$

5^0. *График функции $f(x)$ симметричен относительно прямой $x=a$.*

□ Выделим два произвольных значения x_1 и x_2, которые расположены на оси Ox симметрично относительно точки $x=a$. Пусть $x_1=a-\gamma$, $x_2=a+\gamma$, где γ — произвольное число. Тогда $(x_1-a)^2=(a-\gamma-a)^2=\gamma^2$, $(x_2-a)^2=(a+\gamma-a)^2=\gamma^2$, откуда

$$f(x_1)=f(x_2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\gamma^2}{2\sigma^2}}. \blacksquare$$

6⁰. ***Нормальная кривая в точках $x=a\pm\sigma$ имеет перегиб.***

□ Приравняем нулю вторую производную функции $f(x)$. В результате имеем

$$\frac{1}{\sigma^3\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}\left(\frac{(x-a)^2}{\sigma^2}-1\right)=0,$$

откуда $\frac{(x-a)^2}{\sigma^2}=1$, или $(x-a)^2=\sigma^2$. Из последнего равенства получаем $x=a\pm\sigma$. Нетрудно установить, что при переходе через точки $x=a\pm\sigma$ вторая производная меняет знак. Следовательно, в этих точках кривая претерпевает перегиб, причем

$$f(a\pm\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-1/2}=\frac{1}{\sigma\sqrt{2\pi e}}. \blacksquare$$

На основании доказанных свойств построим график плотности нормального распределения $f(x)$ (рис. 6.4). Как видно из рисунка, нормальная кривая имеет колоколообразную форму. Эта форма является отличительной чертой нормального распределения. Иногда нормальную кривую называют *кривой Гаусса.*

Используя определение центральных моментов, можно получить следующее рекуррентное соотношение:

$$\mu_k=(k-1)\sigma^2\mu_{k-2}. \qquad (6.6)$$

Эта формула позволяет убедиться, что все нечетные центральные моменты равны нулю, так как $\mu_1=0$. Найдем центральный момент μ_4. Так как $\mu_2=DX=\sigma^2$, то по формуле (6.6) получаем $\mu_4=3\sigma^2\sigma^2=3\sigma^4$. Следовательно, коэффициент асимметрии $A=\mu_3/\sigma^3=0/\sigma^3=0$, а эксцесс $E=\mu_4/\sigma^4-3=3\sigma^4/\sigma^4-3=0$. Таким образом, для нормального распределения

$$A=E=0.$$

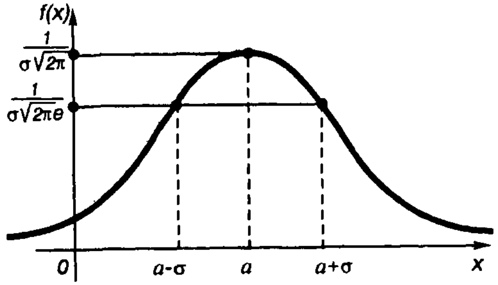


Рис. 6.4

Установим теперь, как влияет изменение параметров a и σ на вид нормальной кривой. При

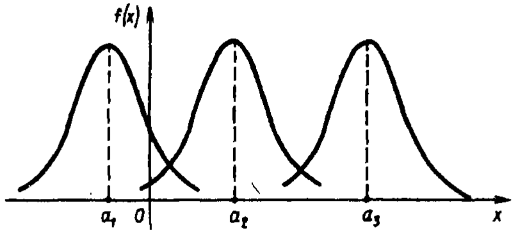


Рис. 6.5

изменении параметра a форма нормальной кривой не изменяется. В этом случае, если математическое ожидание (параметр a) уменьшилось или увеличилось, график нормальной кривой сдвигается влево или вправо (рис. 6.5).

При изменении параметра σ изменяется форма нормальной кривой. Если этот параметр увеличивается, то максимальное значение $1/(\sigma\sqrt{2\pi})$ функции $f(x)$ убывает, и наоборот. Так как площадь, ограниченная кривой распределения и осью Ox, должна быть постоянной и равной 1, то с увеличением параметра σ кривая приближается к оси Ox и растягивается вдоль нее, а с уменьшением σ кривая стягивается к прямой $x=a$ (рис. 6.6).

Использование формул (6.2) и (6.3) для практических расчетов затруднительно, так как функции $f(x)$ и $F(x)$ в данном случае являются трансцендентными, а интеграл $\int e^{-\frac{(x-a)^2}{2\sigma^2}}dx$ относится к «неберущимся». Однако решение задач по формулам (6.2) и (6.3) можно упростить, если от нормального распределения с произвольными параметрами a и σ перейти также к нормальному распределению с параметрами $a=0$, $\sigma=1$.

Функция плотности нормального распределения $f(x)$ с параметрами $a=0$, $\sigma=1$ называется *плотностью стандартной*

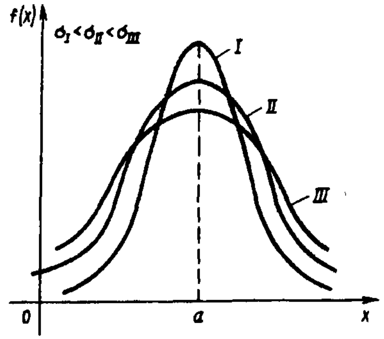


Рис. 6.6

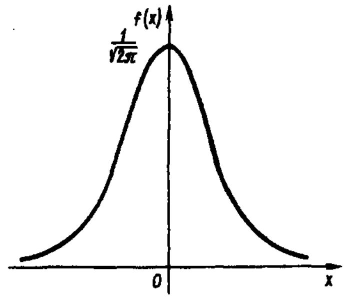


Рис. 6.7

нормальной случайной величины, а ее график — *стандартной кривой Гаусса*.

Функция плотности нормальной стандартной величины определяется формулой

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \qquad (6.7)$$

а ее график изображен на рис. 6.7.

В предыдущей главе было показано, что для величины $Z=\frac{X-MX}{\sigma_X}$ $DZ=1$, $MZ=0$. Поэтому стандартную нормальную кривую можно рассматривать как кривую распределения случайной величины $U=\frac{X-a}{\sigma}$, где X — случайная величина, подчиненная нормальному закону распределения с параметрами a и σ. Значения функции $f(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$ приведены в табл. П.1 (см. также § 4.2).

Для вычисления вероятности попадания значения случайной величины, распределенной нормально, в заданный интервал (α, β), можно воспользоваться специальной функцией

$$\Phi(u)=\frac{2}{\sqrt{2\pi}}\int_0^u e^{-t^2/2}\,dt. \qquad (6.8)$$

Эта функция называется *функцией Лапласа* или *интегралом вероятности*. Ее значения приведены в табл. П.2. При использовании указанной таблицы следует помнить, что $\Phi(u)$ — нечетная функция и поэтому в таблице приведены ее значения только для положительного аргумента.

Используя формулу (5.4), найдем вероятность попадания значения случайной величины X в интервал (α, β), для чего перейдем к стандартной случайной величине $U=\frac{X-a}{\sigma}$. Неравенство $\alpha<X<\beta$ равносильно неравенству

$$\frac{\alpha-a}{\sigma}<\frac{X-a}{\sigma}<\frac{\beta-a}{\sigma},$$

поэтому вероятности выполнения этих неравенств равны, т. е.

$$P(\alpha<X<\beta)=P\left(\frac{\alpha-a}{\sigma}<\frac{X-a}{\sigma}<\frac{\beta-a}{\sigma}\right).$$

Теперь можно записать

$$P(\alpha<X<\beta)=P(u_1<U<u_2)=F(u_2)-F(u_1), \qquad (6.9)$$

где

$$u_1=\frac{\alpha-a}{\sigma}, \quad u_2=\frac{\beta-a}{\sigma}.$$

Далее, используя определение интегральной функции распределения, получаем

$$F(u)=P(U<u)=\int_{-\infty}^{u} f(t)\,dt=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-t^2/2}\,dt=$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{-t^2/2}dt+\frac{1}{\sqrt{2\pi}}\int_0^u e^{-t^2/2}dt.$$

Первое из слагаемых в правой части полученного равенства численно равно половине площади, заключенной между стандартной кривой Гаусса и осью Ox (рис. 6.7), а второе слагаемое похоже на выражение функции Лапласа. Поэтому равенство можно переписать в виде

$$F(u)=1/2+(1/2)\Phi(u). \qquad (6.10)$$

Используя формулу (6.10) и выражение (6.9), получим

$$P(\alpha<X<\beta)=(1/2)\Phi(u_2)-(1/2)\Phi(u_1). \qquad (6.11)$$

Выражение (6.11) можно также получить, воспользовавшись понятием дифференциальной функцией распределения $f(x)$.

Итак, для случайной величины, распределенной нормально, с параметрами a и σ имеем

$$P(\alpha<X<\beta)=(1/2)\Phi(u_2)-(1/2)\Phi(u_1), \qquad (6.12)$$

где $u_1=(\alpha-a)/\sigma$, $u_2=(\beta-a)/\sigma$.

Найдем вероятность того, что случайная величина X, распределенная нормально с параметрами a и σ, отклонится от своего математического ожидания на величину, меньшую ε. Другими словами, найдем вероятность выполнения неравенства $|X-a|<\varepsilon$. Эта вероятность равна вероятности выполнения равносильного двойного неравенства $a-\varepsilon<X<a+\varepsilon$. Воспользовавшись равенством (6.12), имеем

$$u_1=(a-\varepsilon-a)/\sigma=-\varepsilon/\sigma, \quad u_2=(a+\varepsilon-a)/\sigma=\varepsilon/\sigma.$$

Учитывая нечетность функции $\Phi(u)$, получаем

$$P(|X-a|<\varepsilon)=P(a-\varepsilon<X<a+\varepsilon)=$$
$$=(1/2)\Phi(\varepsilon/\sigma)-(1/2)\Phi(-\varepsilon/\sigma)=\Phi(\varepsilon/\sigma).$$

Таким образом,

$$P(|X-a|<\varepsilon)=\Phi(\varepsilon/\sigma). \qquad (6.13)$$

Для стандартной случайной величины X, когда $a=0$, $\sigma=1$, последнее равенство принимает вид

$$P(|X|<\varepsilon)=\Phi(\varepsilon). \tag{6.14}$$

Обозначим $\varepsilon/\sigma=u$. Тогда из равенства (6.13) следует, что

$$P(|X-a|<\sigma u)=\Phi(u). \tag{6.15}$$

Из равенства (6.15) при различных значениях u получаем различные значения вероятности $P(|X-a|<\varepsilon)$:

$$P(|X-a|<\sigma)=\Phi(1)=0{,}6837 \text{ при } u=1,$$

$$P(|X-a|<2\sigma)=\Phi(2)=0{,}9545 \text{ при } u=2,$$

$$P(|X-a|<3\sigma)=\Phi(3)=0{,}9973 \text{ при } u=3.$$

Последнее из равенств показывает, что вероятность отклонения случайной величины X, распределенной нормально, от своего математического ожидания a меньше чем на 3σ, близка к 1, т. е. такое отклонение является практически достоверным событием. Другими словами, вероятность того, что то или иное значение величины X не попадает в интервал с границами $a\pm3\sigma$, равна 0,0027, т. е. близка к нулю. Такое событие можно считать практически невозможным. Сформулируем теперь правило «трех сигм».

Правило «трех сигм». *Если случайная величина имеет нормальное распределение, то отклонение этой величины от ее математического ожидания по абсолютной величине практически не превышает утроенного среднего квадратического отклонения.*

На этом правиле основана приближенная оценка среднего квадратического отклонения. Из полученных данных наблюдения над случайной величиной выбирают максимальное и минимальное значения и их разность делят на 6. Полученное число является грубой оценкой среднего квадратического отклонения при условии, что распределение случайной величины является нормальным.

Задача 6.1. Случайная величина распределена нормально с параметрами $a=8$, $\sigma=3$. Найти вероятность того, что случайная величина в результате опыта примет значение, заключенное в интервале (12,5, 14).

△ Воспользуемся формулой (6.12). Имеем $u_1=(12{,}5-8)/3=1{,}5$; $u_2=(14-8)/3=2$. Тогда

$$P(12{,}5<X<14)=(1/2)\Phi(2)-(1/2)\Phi(1{,}5).$$

По табл. П.2 находим $(1/2)\Phi(1{,}5)=0{,}4332$, $(1/2)\Phi(2)=0{,}4773$. Искомая вероятность

$$P(12{,}5<X<14)=0{,}0441. \ \blacktriangle$$

Задача 6.2. Случайная погрешность измерения подчинена нормальному закону распределения с параметрами $a=0$, $\sigma=9$ мм. Проводятся три независимых измерения. Найти вероятность того, что погрешность хотя бы одного измерения не превосходит по абсолютной величине 3 мм.

△ По формуле (6.13) для $a=0$, $\sigma=9$, $\varepsilon=3$ находим вероятность того, что погрешность измерения в одном испытании не превышает 3 мм. Имеем

$$P(|X|<3)=\Phi(3/9)\approx\Phi(0{,}33)=0{,}2586.$$

Вероятность того, что эта погрешность превышает 3 мм, равна

$$P(|X|>3)=1-P(|X|<3)=0{,}7414.$$

Вероятность того, что во всех трех испытаниях погрешность измерения превышает 3 мм, по теореме умножения вероятностей равна произведению вероятностей $[P(|X|)>3]^3\approx0{,}4075$. Искомая вероятность равна

$$1-[P(|X|>3)]^3\approx0{,}5925. \ \blacktriangle$$

В заключение приведем теорему, которая будет использована в гл. 9 и 10.

Теорема. *Алгебраическая сумма независимых нормально распределенных случайных величин имеет нормальный закон распределения с математическим ожиданием, равным алгебраической сумме математических ожиданий слагаемых, и дисперсией, равной сумме дисперсий слагаемых.*

§ 6.3. Биномиальное распределение

Пусть проводятся n независимых испытаний, в каждом из которых событие A может появиться либо не появиться. Пусть, далее, вероятность p появления события A в единичном испытании постоянна и не меняется от испытания к испытанию. Рассмотрим в качестве дискретной случайной величины X число появлений события A в этих испытаниях. Формула, позволяющая найти вероятность появления m раз события A в n независимых испытаниях, была приведена в § 4.1. Используя эту формулу (формулу Бернулли), приведем следующее определение.

Определение. *Дискретная случайная величина X, которая может принимать только целые неотрицательные значения с вероятностью*

$$P_n(m)=P(X=m)=C_n^m p^m q^{n-m}, \tag{6.16}$$

где $p+q=1$, $p>0$, $q>0$, $m=0, 1, 2, \ldots, n$, называется распределенной по биномиальному закону, а p — параметром биномиального распределения.

Ряд распределения случайной величины, подчиненной биномиальному закону, можно представить в следующем виде:

$X=m$	0	1	...	k	...	n
$P_n(m)$	$C_n^0 p^0 q^n$	$C_n^1 p^1 q^{n-1}$	...	$C_n^k p^k q^{n-k}$	...	$C_n^n p^n q^0$

Функция распределения в этом случае определяется формулой

$$F(x)=\begin{cases} 0 & \text{при } x\leqslant 0, \\ \sum\limits_{m<x} C_n^m p^m q^{n-m} & \text{при } 0<x\leqslant n, \\ 1 & \text{при } x>n. \end{cases}$$

Найдем числовые характеристики этого распределения.

По определению математического ожидания для дискретной случайной величины имеем

$$MX=\sum_{i=1}^{n} x_i p_i=\sum_{m=0}^{n} mC_n^m p^m q^{n-m}. \tag{6.17}$$

Запишем равенство, являющееся биномом Ньютона,

$$(p+q)^n=\sum_{m=0}^{n} C_n^m p^m q^{n-m} \tag{6.18}$$

и продифференцируем его по p. В результате получим

$$n(p+q)^{n-1}=\sum_{m=0}^{n} mC_n^m p^{m-1} q^{n-m}.$$

Умножая полученное равенство на p, имеем

$$np(p+q)^{n-1}=\sum_{m=0}^{n} mC_n^m p^m q^{n-m}.$$

Сравнивая последнее равенство с равенством (6.17), заключаем, что $np(p+q)^{n-1}=MX$, откуда, так как $p+q=1$, следует

$$MX=np, \text{ или } Mm=np. \tag{6.19}$$

Для вычисления дисперсии воспользуемся ее свойством 4°. Предварительно находим

$$M(X^2)=\sum_{m=0}^{n} m^2 C_n^m p^m q^{n-m}.$$

Дифференцируя равенство (6.18) дважды по p, получаем

$$n(n-1)(p+q)^{n-2}=\sum_{m=0}^{n} m(m-1)C_n^m p^{m-2} q^{n-m}$$

Умножим последнее равенство на p^2 и затем преобразуем его правую часть. Имеем

$$n(n-1)p^2(p+q)^{n-2}=\sum_{m=0}^{n} m^2 C_n^m p^m q^{n-m}-\sum_{m=0}^{n} mC_n^m p^m q^{n-m},$$

откуда при $p+q=1$ следует, что

$$n^2p^2-np^2=M(X^2)-MX=M(X^2)-np.$$

Тогда $M(X^2)=n^2p^2-np^2+np$, а значит,

$$DX=M(X^2)-(MX)^2=n^2p^2-np^2+np-n^2p^2=$$
$$=np-np^2=np(1-p)=npq.$$

Таким образом,

$$DX=npq, \text{ или } Dm=npq, \tag{6.20}$$

а среднее квадратическое

$$\sigma_X=\sqrt{npq}, \text{ или } \sigma_m=\sqrt{npq}. \tag{6.21}$$

Воспользовавшись понятием начальных и центральных моментов, можно получить следующие соотношения:

$$A=\frac{q-p}{\sqrt{npq}}; \tag{6.22}$$

$$E=\frac{1-6pq}{npq}. \tag{6.23}$$

Исследуем форму многоугольника биномиального распределения. На рис. 6.8 изображены многоугольники распределения для $n=6$ и $p=0{,}3$; 0,5; 0,7. Для этих распределений вероятность $P_n(m)$ сначала возрастает при увеличении m и достигает наибольшего значения при некотором наивероятнейшем значении $m=m_0$. При дальнейшем увеличении числа m вероятность $P_n(m)$ начинает убывать.

Биномиальное распределение асимметрично, за исключением случая, когда $p=0{,}5$, причем при $p<0{,}5$ асимметрия положительна, а при $p>0{,}5$ — отрицательна.

На рис. 6.9 изображены многоугольники распределения для $p=0{,}4$ и $n=3$, 5, 7. Из рисунка видно, что при возрастании n многоугольник распределения сдвигается вправо и становится более «плосковершинным», приближаясь к симметричной форме. Последнее утверждение следует также из формул (6.22) и (6.23), так как с ростом n коэффициент асимметрии A и эксцесс E стремятся к нулю.

Рассмотрим теперь вероятность попадания случайной величины X, распределенной по биномиальному закону, в заданный интервал $[k_1, k_2]$, где k_1 и k_2 — целые числа. По теореме

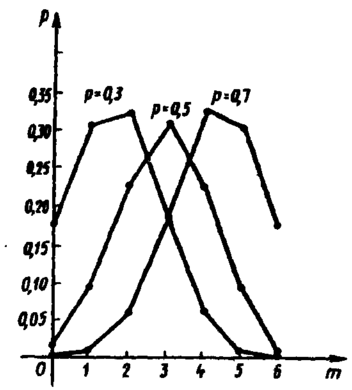


Рис. 6.8

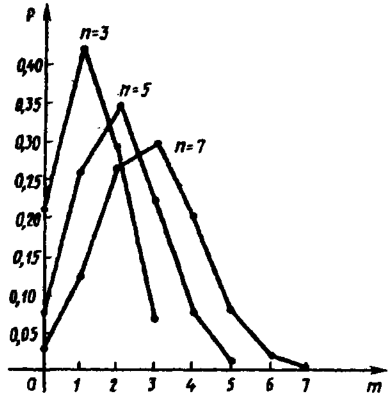


Рис. 6.9

сложения для несовместных событий, вероятность $P_n(k_1, k_2)$ того, что событие A появилось в n испытаниях от k_1 до k_2 раз, равна

$$P(k_1 \leqslant X \leqslant k_2) = P_n(k_1;\, k_2) = \sum_{k_1 \leqslant m \leqslant k_2} C_n^m p^m q^{n-m}.$$

При больших n эта формула приводит к громоздким вычислениям. В этих случаях следует использовать приближенную формулу, справедливость которой следует из приведенной ниже без доказательства теоремы.

Интегральная теорема Муавра — Лапласа. *Если вероятность p наступления события A в каждом испытании отлична от нуля и единицы, то для вероятности $P_n(k_1;\, k_2)$ того, что событие A появится в n независимых испытаниях от k_1 до k_2 раз, справедливо следующее соотношение:*

$$P_n(k_1 \leqslant m \leqslant k_2) = P_n(k_1;\, k_2) \approx \frac{1}{\sqrt{2\pi}} \int_{u_1}^{u_2} e^{-t^2/2}\, dt, \tag{6.24}$$

где $u_1 = \dfrac{k_1 - np}{\sqrt{npq}}$, $u_2 = \dfrac{k_2 - np}{\sqrt{npq}}$ *).

Преобразуем интеграл в правой части (6.24) так, чтобы было можно использовать функцию $\Phi(u)$. Имеем

$$\frac{1}{\sqrt{2\pi}} \int_{u_1}^{u_2} e^{-t^2/2}\, dt = \frac{1}{\sqrt{2\pi}} \int_{u_1}^{0} e^{-t^2/2}\, dt + \frac{1}{\sqrt{2\pi}} \int_{0}^{u_2} e^{-t^2/2}\, dt =$$

*) «Грубое» правило для применения формулы (6.24) таково: n должно иметь порядок не менее нескольких десятков (лучше сотен), а $np > 10$.

$$= \frac{1}{\sqrt{2\pi}} \int_{0}^{u_2} e^{-t^2/2}\, dt - \frac{1}{\sqrt{2\pi}} \int_{0}^{u_1} e^{-t^2/2}\, dt = \frac{1}{2}\Phi(u_2) - \frac{1}{2}\Phi(u_1).$$

Следовательно,

$$P_n(k_1 \leqslant m \leqslant k_2) = P_n(k_1;\, k_2) \approx \frac{1}{2}\Phi(u_2) - \frac{1}{2}\Phi(u_1), \tag{6.25}$$

где

$$u_1 = \frac{k_1 - np}{\sqrt{npq}}, \quad u_2 = \frac{k_2 - np}{\sqrt{npq}}.$$

Используя равенство (6.25), можно получить некоторые полезные соотношения. Очевидно, что приведенные ниже неравенства равносильны:

$$k_1 \leqslant m \leqslant k_2;$$
$$\frac{k_1 - np}{\sqrt{npq}} \leqslant \frac{m - np}{\sqrt{npq}} \leqslant \frac{k_2 - np}{\sqrt{npq}};$$
$$u_1\sqrt{npq} \leqslant m - np \leqslant u_2\sqrt{npq};$$
$$np + u_1\sqrt{npq} \leqslant m \leqslant np + u_2\sqrt{npq};$$
$$u_1\sqrt{\frac{pq}{n}} \leqslant \frac{m}{n} - p \leqslant u_2\sqrt{\frac{pq}{n}},$$

где

$$u_1 = \frac{k_1 - np}{\sqrt{npq}}, \quad u_2 = \frac{k_2 - np}{\sqrt{npq}}.$$

В силу равносильности этих неравенств вероятности их выполнения совпадают и приближенно равны $(1/2)\Phi(u_2) - (1/2)\Phi(u_1)$. Таким образом,

$$P(u_1\sqrt{npq} \leqslant m - np \leqslant u_2\sqrt{npq}) \approx \frac{1}{2}\Phi(u_2) - \frac{1}{2}\Phi(u_1); \tag{6.26}$$

$$P(np + u_1\sqrt{npq} \leqslant m \leqslant np + u_2\sqrt{npq}) \approx \frac{1}{2}\Phi(u_2) - \frac{1}{2}\Phi(u_1); \tag{6.27}$$

$$P\left(u_1\sqrt{\frac{pq}{n}} \leqslant \frac{m}{n} - p \leqslant u_2\sqrt{\frac{pq}{n}}\right) \approx \frac{1}{2}\Phi(u_2) - \frac{1}{2}\Phi(u_1). \tag{6.28}$$

Если $u_2 = -u_1 = u$, то формулы (6.26) и (6.28) принимают соответственно вид

$$P(|m - np| \leqslant u\sqrt{npq}) \approx \Phi(u); \tag{6.29}$$

$$P\left(\left|\frac{m}{n} - p\right| \leqslant u\sqrt{\frac{pq}{n}}\right) \approx \Phi(u). \tag{6.30}$$

Обозначая $u\sqrt{\dfrac{qp}{n}} = \varepsilon$, запишем соотношение (6.30) в виде

$$P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) \approx \Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right). \tag{6.31}$$

Задача 6.3. Известно, что при контроле бракуется 10% изделий. На контроль отобрано 625 изделий. Какова вероятность того, что среди отобранных не менее 550 и не более 575 стандартных изделий?

$\triangle$ По условию, $n=625$; $p=0{,}9$; $q=0{,}1$; $k_1=550$, $k_2=575$. Искомую вероятность находим по формуле (6.25). Имеем

$$P_{625}(550\leqslant m\leqslant 575)\approx\frac{1}{2}\Phi\left(\frac{575-562{,}5}{\sqrt{625\cdot 0{,}9\cdot 0{,}1}}\right)-\frac{1}{2}\Phi\left(\frac{550-562{,}5}{\sqrt{625\cdot 0{,}9\cdot 0{,}1}}\right)\approx$$

$$\approx\frac{1}{2}\Phi(1{,}67)-\frac{1}{2}\Phi(-1{,}67)\approx 0{,}4526+0{,}4526=0{,}9052.\ \blacktriangle$$

Задача 6.4. Вероятность того, что с конвейера сойдет бракованный прибор, равна 0,02. За смену было изготовлено 3600 приборов. Найти максимальное отклонение относительной частоты появления бракованных приборов от вероятности 0,02, если вероятность такого отклонения равна 0,95.

$\triangle$ По условию, $p=0{,}02$; $q=0{,}98$; $P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=\Phi(u)=0{,}95$. По табл. П.2 находим $u=1{,}96$, откуда на основании формулы (6.31) следует, что $u=\varepsilon\sqrt{\frac{n}{pq}}=1{,}96$. Тогда

$$\varepsilon=1{,}96\sqrt{pq/n}=1{,}96\sqrt{0{,}02\cdot 0{,}98/3600}\approx 0{,}0046.$$

Итак, искомое максимальное отклонение равно 0,0046. $\blacktriangle$

Задача 6.5. Вероятность того, что на станке-автомате будет отштампован корпус некоторого механического устройства, не удовлетворяющий допуску, равна 0,01. Сколько надо изготовить корпусов, чтобы с вероятностью 0,99 ожидать не превосходящее 0,03 по абсолютной величине отклонение относительной частоты появления нестандартного корпуса от вероятности его появления?

$\triangle$ По условию, $p=0{,}01$; $q=0{,}99$, $\varepsilon=0{,}03$; следовательно, $P\left(\left|\frac{m}{n}-0{,}01\right|<0{,}03\right)=0{,}99$. Воспользовавшись равенством (6.31), можно записать

$$\Phi(0{,}03\sqrt{n/(0{,}01\cdot 0{,}99)})=0{,}99.$$

По табл. П.2 находим $\Phi(2{,}58)=0{,}99$. Далее имеем $2{,}58=$ $=0{,}03\cdot\sqrt{n/(0{,}02\cdot 0{,}99)}$, или $n=(2{,}58/0{,}03)^2\cdot 0{,}01\cdot 0{,}99\approx 74$. $\blacktriangle$

В дальнейшем для простоты случайную величину X, подчиненную нормальному закону распределения с параметрами a и σ, будем обозначать $N(a,\sigma)$. Стандартную случайную величину, имеющую нормальное распределение, будем обозначать U и называть *стандартной нормальной величиной.* Очевидно, $U=N(0;\,1)$.

Если случайная величина X имеет биномиальное распределение, то есть смысл обозначать ее буквой m или $m_n(p)$, где n — число испытаний Бернулли, и называть *биномиальной.*

На основании локальной (см. § 4.2) и интегральной теорем Муавра — Лапласа можно утверждать, что для достаточно большого n и не очень малой вероятности p биномиальная величина $m_n(p)$ имеет нормальное распределение с параметрами $a=np$, $\sigma=\sqrt{npq}$, т. е.

$$m=m_n(p)\approx N(np;\,\sqrt{npq}).$$

§ 6.4. Распределение Пуассона

В гл. 4 было отмечено, что в силу возникающих вычислительных трудностей нецелесообразно использовать формулу Бернулли. Этих трудностей можно избежать, если воспользоваться асимптотической формулой Лапласа. Однако и эта формула малопригодна, если вероятность p события очень мала. Было доказано, что при большом n и малом p имеет место асимптотическая формула Пуассона, т. е. при $n\to\infty$, $p\to 0$

$$P_m=P(X=m)\approx\frac{\lambda^m e^{-\lambda}}{m!},$$

где $\lambda=np$.

Определение. *Дискретная случайная величина X, которая может принимать только целые неотрицательные значения с вероятностями*

$$P_m=P(X=m)\approx\frac{\lambda^m e^{-\lambda}}{m!},\tag{6.32}$$

называется распределенной по закону Пуассона с параметром λ.

В отличие от биномиального распределения здесь случайная величина может принимать бесконечное множество значений, представляющее собой бесконечную последовательность целых чисел 0, 1, 2, 3, ...

Закон Пуассона описывает число событий m, происходящих за одинаковые промежутки времени. При этом полагается, что события появляются независимо друг от друга с постоянной средней интенсивностью, которая характеризуется параметром $\lambda=np$. Так как для распределения Пуассона вероятность p появления события в каждом испытании мала, то это распределение называют *законом распределения редких явлений.* На рис. 6.10 изображены многоугольники

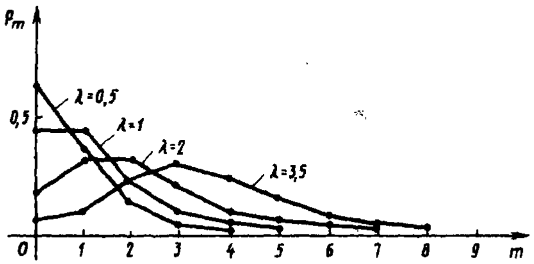


Рис. 6.10

распределения Пуассона, соответствующие различным значениям параметра λ.

Примерами ситуаций, в которых возникает распределение Пуассона, могут служить распределения числа определенных микробов в единице объема, числа вылетевших электронов с накаленного катода за единицу времени, числа α-частиц, испускаемых радиоактивным источником за определенный промежуток времени, числа вызовов, поступающих на телефонную станцию за определенное время суток. Возникновение распределения Пуассона в подобных случаях предполагает выполнение определенных условий (см. § 4.2).

Случайная величина X, подчиненная закону распределения Пуассона, имеет следующий ряд распределения:

$X=m$	0	1	2	3	...	m	...
P_m	$e^{-\lambda}$	$\frac{\lambda e^{-\lambda}}{1!}$	$\frac{\lambda^2 e^{-\lambda}}{2!}$	$\frac{\lambda^3 e^{-\lambda}}{3!}$	...	$\frac{\lambda^m e^{-\lambda}}{m!}$	...

Нетрудно убедиться, что сумма вероятностей P_m равна 1. В этом заключается необходимое требование для дискретного ряда распределения. Действительно, функцию e^x, как известно из математического анализа, можно разложить в ряд Маклорена, который сходится для любого x. В данном случае имеем

$$e^{\lambda}=\sum_{m=0}^{\infty}\frac{\lambda^m}{m!}=1+\lambda+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\ldots+\frac{\lambda^m}{m!}+\ldots.$$

Тогда

$$\sum_{m=0}^{\infty}P_m=\sum_{m=0}^{\infty}\frac{\lambda^m e^{-\lambda}}{m!}=e^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^m}{m!}=e^{-\lambda}e^{\lambda}=1.$$

Найдем математическое ожидание и дисперсию случайной величины, распределенной по закону Пуассона. Имеем

$$MX=\sum_{m=0}^{\infty}mP_m=\sum_{m=0}^{\infty}\frac{m\lambda^m e^{-\lambda}}{m!}=\lambda e^{-\lambda}\sum_{m=1}^{\infty}\frac{\lambda^{m-1}}{(m-1)!}=\lambda e^{-\lambda}e^{\lambda}=\lambda=np.$$

В последней сумме суммирование начинается с $m=1$, так как первый член суммы, соответствующий $m=0$, равен нулю.

Для вычисления дисперсии найдем предварительно математическое ожидание квадрата случайной величины:

$$M(X^2)=\sum_{m=0}^{\infty}m^2\frac{\lambda^m e^{-\lambda}}{m!}=\lambda e^{-\lambda}\sum_{m=1}^{\infty}m\frac{\lambda^{m-1}}{(m-1)!}=$$

$$=\lambda e^{-\lambda}\sum_{m=1}^{\infty}[(m-1)+1]\frac{\lambda^{m-1}}{(m-1)!}=\lambda e^{-\lambda}\sum_{m=1}^{\infty}(m-1)\frac{\lambda^{m-1}}{(m-1)!}+$$

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

$$=\lambda^2 e^{-\lambda}e^{\lambda}+\lambda e^{-\lambda}e^{\lambda}=\lambda^2+\lambda.$$

Тогда

$$DX=M(X^2)-(MX)^2=\lambda^2+\lambda-\lambda^2=\lambda=np.$$

Таким образом, математическое ожидание и дисперсия случайной величины, распределенной по закону Пуассона, совпадают и равны параметру этого распределения λ. В этом и состоит отличительная особенность изучаемого распределения, которая используется на практике.

Предположим, что в результате опытов получено несколько значений случайной величины, распределение которой неизвестно. На основании полученных данных находят оценки для математического ожидания и дисперсии этой величины (способы получения таких оценок рассмотрены в дальнейшем). Если полученные оценки близки между собой, то имеется основание предположить, что случайная величина подвержена распределению Пуассона. В противном случае таких оснований нет.

Воспользовавшись понятием начальных и центральных моментов, можно показать, что для распределения Пуассона коэффициент асимметрии $A=1/\sqrt{\lambda}$, а эксцесс $E=1/\lambda$. Поскольку параметр λ всегда положителен, у распределения Пуассона всегда положительная асимметрия и эксцесс.

Задача 6.6. Вероятность попадания в цель при одном выстреле равна 0,015. Сделано 600 выстрелов. Какова вероятность того, что число попаданий в цель не меньше 7 и не больше 10?

$\triangle$ По условию следует найти вероятность того, что значение случайной величины X — число попаданий в цель — попадает в интервал $[7; 10]$, т. е. найти $P(7\leqslant X\leqslant 10)$. По теореме сложения вероятностей эта вероятность равна $P_7+P_8+P_9+P_{10}$. Суммируемые вероятности находятся по формуле (6.32), так как вероятность попадания при одном выстреле мала. В данном случае параметр $\lambda=np=600\cdot 0{,}015=9$. Тогда $P_7\approx 0{,}1171$; $P_8\approx 0{,}1318$; $P_9\approx 0{,}1318$; $P_{10}\approx 0{,}1186$. Следовательно, $P(7\leqslant X\leqslant 10)\approx 0{,}1171+0{,}1318+0{,}1318+0{,}1186=0{,}4993$. ▲

§ 6.5. Распределения, связанные с нормальным распределением

Распределение χ^2 (распределение К. Пирсона*)).

Определение. *Пусть независимые случайные величины U_1, U_2, ..., U_k являются стандартными нормально распределенными величинами, т. е. $U_i=N(0, 1)$, $i=1, 2, ..., k$. Распределение случайной величины*

$$\chi^2(k)=U_1^2+U_2^2+\ldots+U_k^2 \tag{6.33}$$

называется распределением хи-квадрат с k степенями свободы, а сама величина (6.33) — *величиной хи-квадрат с k степенями свободы.*

Далее вместо «хи-квадрат с k степенями свободы» будем писать $\chi^2(k)$. Обратим внимание, что $\chi^2(k)\geqslant 0$.

Подобно тому как математическое ожидание a и среднее квадратическое σ являются параметрами нормального закона, так и число k является параметром $\chi^2(k)$-распределения. Почему это число называют «k степенями свободы»? Вообще число степеней свободы определяют как разность между числом суммируемых случайных величин и числом линейных связей, ограничивающих свободу изменения этих величин. Так как в сумме (6.33) слагаемые независимы, то число степеней свободы равно числу слагаемых, т. е. k.

Формула функции плотности $\chi^2(k)$-распределения имеет сложный вид, поэтому здесь не приводится. Графики этой функции — кривые Пирсона — при $k=1$, 2, 6 изображены на рис. 6.11.

В табл. П.4 и П.6 содержатся определенные значения величины $\chi^2(k)$. Назначение этих таблиц выясняется в § 9.7 и гл. 10.

Приведем без доказательства два утверждения, которые будут использоваться в дальнейшем:

дисперсия величины $\chi^2(k)$ равна $2k$, т. е.

$$D[\chi^2(k)]=2k; \tag{6.34}$$

если случайные величины $\chi^2(k_1)$ и $\chi^2(k_2)$ независимы, то их сумма имеет хи-квадрат-распределение с числом степеней свободы k_1+k_2, т. е.

---

*) Пирсон Карл (1857—1936) — английский статистик.

$$\chi^2(k_1)+\chi^2(k_2)=\chi^2(k_1+k_2). \tag{6.35}$$

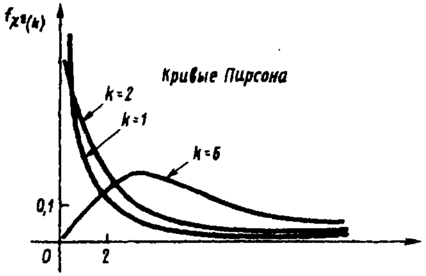


Рис. 6.11

Распределение Стьюдента*) (t-распределение).

Определение. *Пусть U — стандартная нормально распределенная случайная величина, т. е. $U=N(0,1)$, а $\chi^2(k)$ — случайная величина, имеющая хи-квадрат-распределение с k степенями свободы, причем U и $\chi^2(k)$ — независимые величины. Распределение случайной величины*

$$t(k)=\frac{U}{\sqrt{\chi^2(k)/k}} \tag{6.36}$$

называется t-распределением с k степенями свободы или $t(k)$-распределением, а сама величина (6.36) — *t-величиной с k степенями свободы или $t(k)$-величиной.*

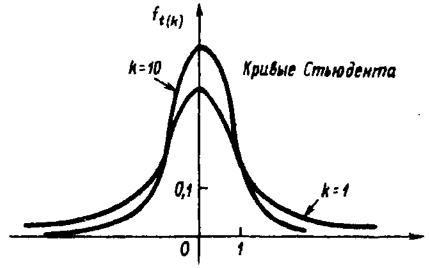


Рис. 6.12

Графики функции плотности распределения величины $t(k)$ — кривые Стьюдента — при $k=1$, 10 изображены на рис. 6.12. Обратим внимание на то, что функция плотности $f_{t(k)}$ симметрична относительно оси ординат.

В табл. П.3 содержатся такие значения t_γ величины $t(k)$, при которых вероятность $P(|t(k)|<t_\gamma)=\gamma$. Эта таблица используется в § 9.7 и 10.2.

Распределение Фишера) (F-распределение)**

Определение. *Пусть $\chi^2(l)$ и $\chi^2(k)$ — независимые случайные величины, имеющие χ^2-распределение соответственно с l и k степенями свободы. Распределение случайной величины*

$$F(l, k)=\frac{\chi^2(l)/l}{\chi^2(k)/k} \tag{6.37}$$

называется F-распределением с l и k степенями свободы или $F(l, k)$-распределением, а сама величина (6.37) *$F(l, k)$-величиной.*

Так как случайные величины $\chi^2(l)\geqslant 0$ и $\chi^2(k)\geqslant 0$, то и $F(l, k)\geqslant 0$.

---

*) Стьюдент — псевдоним английского статистика В. Госсета, исследовавшего это распределение.

**) Фишер Рональд — английский статистик.

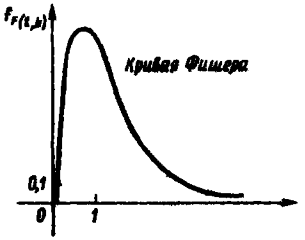


Рис. 6.13

График функции плотности распределения величины $F(l, k)$ — кривая Фишера — при $l=4$ и $k=40$ изображен на рис. 6.13.

В табл. П.7 содержатся такие значения f_γ величины $F(l, k)$, при которых вероятность $P(F(l, k)<f_\gamma)=\gamma$. Эта таблица используется в гл. 10—12.

§ 6.6. Показательное распределение

Определение. *Непрерывная случайная величина X, функция плотности которой задается выражением*

$$f(x)=\begin{cases}\mu e^{-\mu x} & \text{при } x>0,\\ 0 & \text{при } x\leqslant 0,\end{cases} \qquad (6.38)$$

называется случайной величиной, имеющей показательное, или экспоненциальное, распределение.

Величина срока службы различных устройств и времени безотказной работы отдельных элементов этих устройств при выполнении определенных условий обычно подчиняется показательному распределению. Другими словами, величина промежутка времени между появлениями двух последовательных редких событий подчиняется зачастую показательному распределению.

Как видно из формулы (6.38), показательное распределение определяется только одним параметром μ.

Найдем функцию распределения показательного закона, используя свойства дифференциальной функции распределения:

$$F(x)=\begin{cases}0 \text{ при } x\leqslant 0,\\ \int\limits_{-\infty}^{x} f(t)\,dt=\int\limits_{0}^{x}\mu e^{-\mu t}\,dt=1-e^{-\mu x} \text{ при } x>0.\end{cases} \qquad (6.39)$$

Графики дифференциальной и интегральной функций показательного распределения изображены соответственно на

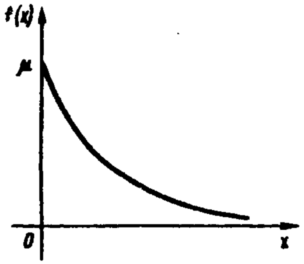


Рис. 6.14

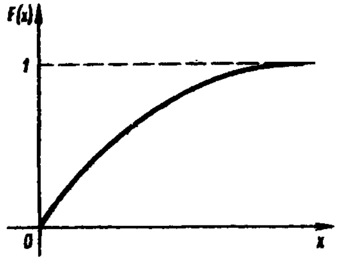


Рис. 6.15

рис. 6.14, 6.15. Предлагаем самостоятельно убедиться, что для показательного распределения

$$MX=1/\mu, \quad DX=1/\mu^2, \quad \sigma_X=1/\mu,$$

Таким образом, для показательного распределения характерно, что среднее квадратическое отклонение численно равно математическому ожиданию. Нетрудно убедиться, что коэффициент асимметрии и эксцесс для показательного распределения являются постоянными величинами: $A=2$, $E=6$.

ГЛАВА 7. ПРЕДЕЛЬНЫЕ ТЕОРЕМЫ

§ 7.1. Предварительные замечания

При изучении теории вероятностей приходится использовать понятия случайного события и случайной величины. При этом предсказать заранее результат испытания, в котором может появиться или не появиться то или иное событие или какое-либо определенное значение случайной величины, невозможно, так как исход испытания зависит от многих случайных причин, не поддающихся учету.

Однако при неоднократном повторении испытаний могут наблюдаться определенные закономерности. Эти закономерности, свойственные массовым случайным явлениям, и изучает теория вероятностей. Следует отметить, что математические законы теории вероятностей получены в результате абстрагирования реальных ситуаций, в которых наблюдаются случайные массовые явления.

При изучении результатов наблюдений над реальными случайными массовыми явлениями также имеют место некоторые закономерности. Следует обратить внимание на то, что они обладают свойством устойчивости. Суть этого свойства состоит в том, что конкретные особенности каждого отдельного случайного явления почти не сказываются на среднем результате большой массы подобных явлений, а характеристики случайных событий и случайных величин, наблюдаемых в испытаниях, при неограниченном увеличении числа испытаний становятся практически не случайными.

Предельные теоремы вероятностей устанавливают зависимость между случайностью и необходимостью. По смыслу их можно разбить на две группы, одна из которых называется *законом больших чисел*, а другая — *центральной предельной теоремой.*

Под законом больших чисел не следует понимать какой-то один общий закон, связанный с большими числами. Закон больших чисел — это обобщенное название нескольких теорем, из которых следует, что при неограниченном увеличении числа испытаний средние величины стремятся к некоторым постоянным.

§ 7.2. Неравенство Чебышева

Лемма Чебышева. *Если случайная величина X имеет конечные математическое ожидание MX и дисперсию DX, то для любого положительного ε справедливо неравенство Чебышева*

$$P(|X-MX|\leqslant\varepsilon)>1-DX/\varepsilon^2. \tag{7.1}$$

□ Для простоты ограничимся доказательством неравенства для дискретной случайной величины. Сумма вероятностей двух событий, заключающихся в выполнении неравенств $|X-MX|\leqslant\varepsilon$ и $|X-MX|>\varepsilon$, равна 1, так как эти события противоположны, т. е.

$$P(|X-MX|\leqslant\varepsilon)+P(|X-MX|>\varepsilon)=1.$$

Отсюда следует, что

$$P(|X-MX|\leqslant\varepsilon)=1-P(|X-MX|>\varepsilon). \tag{7.2}$$

Так как мы предполагаем, что случайная величина X является дискретной, то ее дисперсия

$$DX=\sum_{i=1}^{n}(x_i-MX)^2p_i.$$

Все слагаемые в правой части последнего равенства неотрицательные, поэтому при отбрасывании хотя бы одного слагаемого значение суммы может только уменьшиться. Отбросим те слагаемые, у которых $|x_i-MX|\leqslant\varepsilon$. В результате сумма может только уменьшиться. Предположим для определенности, что отброшены первые k слагаемых. В противном случае все значения x_i можно перенумеровать так, что отброшенными окажутся первые k слагаемых. Таким образом,

$$DX\geqslant\sum_{i=k+1}^{n}(x_i-MX)^2p_i.$$

Так как для оставшихся слагаемых выполняется неравенство $|x_i-MX|>\varepsilon$, то будут выполняться и неравенства $|x_i-MX|^2>\varepsilon^2$. Воспользовавшись последним соотношением, для дисперсии DX имеем

$$DX>\sum_{i=k+1}^{n}\varepsilon^2p_i=\varepsilon^2\sum_{i=k+1}^{n}p_i. \tag{7.3}$$

Сумма $\sum\limits_{i=k+1}^{n}p_i=p_{k+1}+p_{k+2}+\ldots+p_n$, по теореме сложения вероятностей, является вероятностью того, что отклонение $|X-MX|$ больше ε, т. е.

$$P(|X-MX|>\varepsilon)=\sum_{i=k+1}^{n}p_i.$$

Тогда неравенство (7.3) можно переписать в виде

$$DX>\varepsilon^2P(|X-MX|>\varepsilon),$$

откуда

$$P(|X-MX|>\varepsilon)<\frac{DX}{\varepsilon^2}. \tag{7.4}$$

Подставляя неравенство (7.4) в равенство (7.2), получаем

$$P(|X-MX|\leqslant\varepsilon)>1-\frac{DX}{\varepsilon^2}. \blacksquare$$

Неравенство Чебышева зачастую дает грубую, а иногда тривиальную, не представляющую интереса оценку. Пусть, например, $\varepsilon=\sqrt{DX}/2$. Тогда неравенство Чебышева принимает вид

$$P(|X-MX|\leqslant\varepsilon)>1-4DX/DX=-3.$$

Получена заведомо известная оценка вероятности $P(|X-MX|\leqslant\varepsilon)$, так как вероятность любого события всегда неотрицательна. Тем не менее неравенство Чебышева имеет большое теоретическое значение. С его помощью доказываются теоремы и делаются теоретические выводы.

§ 7.3. Теорема Чебышева

Теорема Чебышева. *При достаточно большом числе независимых случайных величин $X_1, X_2, X_3, \ldots, X_n$, дисперсия каждой из которых не превышает одного и того же постоянного числа B, для произвольного сколько угодно малого числа ε справедливо неравенство*

$$P\left(\left|\frac{\sum\limits_{i=1}^{n}X_i}{n}-\frac{\sum\limits_{i=1}^{n}MX_i}{n}\right|\leqslant\varepsilon\right)>1-h, \tag{7.5}$$

где h — положительное, близкое к нулю число.

□ Пусть имеется n независимых случайных величин X_1, X_2, X_3, ..., X_n. Очевидно, среднее арифметическое этих величин является, в свою очередь, также случайной величиной, которую обозначим $\bar{X}$: $\bar{X}=\sum_{i=1}^{n} X_i/n$. Используя свойства 1^0 и 4^0 математического ожидания, получаем

$$M\bar{X}=M\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)=\frac{\sum_{i=1}^{n} MX_i}{n}. \tag{7.6}$$

Аналогично, используя свойства 1^0 и 3^0 дисперсии, имеем

$$D\bar{X}=D\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)=\frac{\sum_{i=1}^{n} DX_i}{n^2}. \tag{7.7}$$

Теперь, применяя к случайной величине $\bar{X}$ неравенство Чебышева, имеем

$$P(|\bar{X}-M\bar{X}|\leqslant\varepsilon)>1-\frac{D\bar{X}}{\varepsilon^2}.$$

Заменяя в этом неравенстве $M\bar{X}$ и $D\bar{X}$ их значениями по формулам (7.6) и (7.7), получаем

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n}-\frac{\sum_{i=1}^{n} MX_i}{n}\right|\leqslant\varepsilon\right)>1-\frac{\sum_{i=1}^{n} DX_i}{\varepsilon^2 n^2}. \tag{7.7'}$$

Так как дисперсии DX_i не превосходят, по условию, числа B. то при замене каждой дисперсии числом B последнее неравенство только усиливается. Можно записать

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n}-\frac{\sum_{i=1}^{n} MX_i}{n}\right|\leqslant\varepsilon\right)>1-\frac{B}{n\varepsilon^2}. \tag{7.8}$$

Очевидно, что для любого, сколь угодно малого $h>0$ можно подобрать такое n, при котором выполняется неравенство $B/n\varepsilon^2<h$, а следовательно, и неравенство (7.5) ■.

Теорему Чебышева можно сформулировать следующим образом.

Теорема. *Если дисперсии независимых случайных величин X_1, X_2, X_3, ..., X_n не превышают постоянного числа B, то для произвольного, сколь угодно малого числа $\varepsilon>0$ справедливо равенство*

$$\lim_{n\to\infty} P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n}-\frac{\sum_{i=1}^{n} MX_i}{n}\right|\leqslant\varepsilon\right)=1.$$

□ Доказательство этой теоремы аналогично доказательству предыдущей теоремы. Здесь достаточно в неравенстве (7.8)

перейти к пределу при $n\to\infty$, в результате чего получаем требуемое равенство. ■

Как следует из доказанной теоремы, среднее арифметическое случайных величин при возрастании их числа проявляет свойство устойчивости, т. е. стремится по вероятности к неслучайной величине, которой является среднее арифметическое математических ожиданий этих величин.

Не следует считать, что предел величины $\sum_{i=1}^{n} X_i/n$ при $n\to\infty$ равен $\sum_{i=1}^{n} MX_i/n$. Это равенство означает, что вероятность отклонения по абсолютной величине $\sum_{i=1}^{n} X_i/n$ от $\sum_{i=1}^{n} MX_i/n$ меньше чем на ε при неограниченном возрастании n стремится к 1.

Рассмотрим частный случай теоремы Чебышева. Пусть при n испытаниях наблюдаются n значений случайной величины X, имеющей математическое ожидание MX и дисперсию DX. Полученные значения можно рассматривать как случайные величины X_1, X_2, X_3, ..., X_n. Это следует понимать так. Серия из n испытаний проводится неоднократно. Поэтому в результате i-го испытания, $i=1, 2, 3, ..., n$, в каждой серии испытаний появится то или иное значение случайной величины X, не известное заранее. Следовательно, i-е значение x_i случайной величины, полученное в i-м испытании, изменяется случайным образом, если переходить от одной серии испытаний к другой. Таким образом, каждое значение x_i можно считать случайной величиной X_i.

Предположим, что испытания удовлетворяют следующим требованиям:

1) испытания независимы. Это означает, что результаты X_1, X_2, ..., X_n испытаний — независимые случайные величины;

2) испытания проводятся в одинаковых условиях — это означает, с точки зрения теории вероятностей, что каждая из случайных величин X_1, X_2, ..., X_n имеет такой же закон распределения, что и исходная величина X, поэтому $MX_i=MX$ и $DX_i=DX$, $i=1, 2, ..., n$.

Для среднего арифметического $\bar{X}=\sum_{i=1}^{n} x_i/n$, считая, что каждое x_i — случайная величина X_i, найдем математическое ожидание и дисперсию подобно тому, как это делалось при доказательстве теоремы Чебышева. Имеем

$$M\bar{X}=M\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)=M\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)=\frac{1}{n}\sum_{i=1}^{n} MX_i=\frac{MX\cdot n}{n}=MX;$$

$$D\bar{X}=D\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)=D\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)=\frac{1}{n^2}\sum_{i=1}^{n} DX_i=\frac{DX\cdot n}{n^2}=\frac{DX}{n}.$$

Далее, применяя неравенство Чебышева для $\bar{X}$, получаем

$$P(|\bar{X}-MX|\leqslant\varepsilon)>1-\frac{DX}{n\varepsilon^2}. \qquad (7.9)$$

Переходя в последнем неравенстве к пределу, имеем

$$\lim_{n\to\infty} P(|\bar{X}-MX|\leqslant\varepsilon)=1. \qquad (7.9')$$

Таким образом, в условиях теоремы Чебышева имеют место равенства (7.9) и (7.9′). Из равенства (7.9′) следует, что среднее арифметическое наблюдений случайной величины X также обладает свойством устойчивости.

Теорема Чебышева имеет большое практическое применение. Она позволяет, используя среднее арифметическое, получить представление о величине математического ожидания, и наоборот. Так, измеряя какой-либо параметр с помощью прибора, не дающего систематической погрешности, можно получить достаточно большое число результатов измерений, среднее арифметическое которых по теореме Чебышева будет практически мало отличаться от истинного значения параметра.

§ 7.4. Теорема Бернулли

Теорема Бернулли. *Если вероятность события A в каждом из n независимых испытаний постоянна и равна p*), то при достаточно большом n для произвольного $\varepsilon>0$ справедливо неравенство*

$$P\left(\left|\frac{m}{n}-p\right|\leqslant\varepsilon\right)>1-h, \qquad (7.10)$$

где m — число появлений события A в n испытаниях, h — число, близкое к нулю.

Рассмотрим в качестве случайных величин X_i, $i=1, 2, 3, \ldots, n$, число появлений события A в i-м испытании. Каждая из этих величин может принимать только два значения: 1 с вероятностью p; 0 с вероятностью $q=1-p$. Математическое ожидание каждой из этих величин

$$MX_i=1\cdot p+0\cdot q=p,$$

*) Так как вероятность события A от испытания к испытанию не изменяется (она остается равной p), то это с вероятностной точки зрения означает, что испытания проводятся в одинаковых условиях.

а дисперсия

$$DX_i=(1-p)^2p+(0-p)^2q=pq.$$

Следовательно, все случайные величины имеют одно и то же математическое ожидание p и одну и ту же дисперсию pq.

Используя понятие производной, можно показать, что произведение двух сомножителей, сумма которых равна 1, не превышает 1/4. Таким образом, $DX_i=pq\leqslant 1/4$, т. е. дисперсии всех величин X_i ограничены одним числом. Кроме того, случайные величины независимы, так как проводятся независимые испытания. Необходимые условия выполнены, поэтому, применяя к величине $\bar{X}=\sum_{i=1}^{n} X_i/n$ теорему Чебышева (ее частный случай) и учитывая, что $X_1+X_2+X_3+\ldots+X_n=m$, получим

$$P\left(\left|\frac{m}{n}-p\right|\leqslant\varepsilon\right)>1-\frac{pq}{n\varepsilon^2}. \qquad (7.10')$$

Так как для любого h можно найти такое n, для которого выполняется неравенство $\frac{pq}{n\varepsilon^2}<h$, то справедливость формулы (7.10) доказана.

Если в неравенстве (7.10′) перейти к пределу, то получим

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n}-p\right|\leqslant\varepsilon\right)=1. \qquad (7.10'')$$

Теорема Бернулли устанавливает связь между вероятностью появления события и его относительной частотой появления и позволяет при этом предсказать, какой примерно будет эта частота в n испытаниях. Из теоремы видно, что отношение m/n обладает свойством устойчивости при неограниченном росте числа испытаний.

Справедливость последнего утверждения подтверждают опыты (см. табл. 2.1). С ростом n относительная частота m/n стремится к неслучайной величине $P(\Gamma)=1/2$.

В заключение подчеркнем, что для доказательства теоремы Бернулли и частного случая теоремы Чебышева мы потребовали неизменности условий испытаний наряду с их независимостью. Однако в теореме Чебышева рассматривалась случайная величина X, а в теореме Бернулли — событие A.

Иногда (при решении практических задач) требуется оценить вероятность того, что отклонение числа m появления события в n испытаниях от ожидаемого результата np не превысит определенного числа ε. В этом случае роль случайной величины играет число m. Пусть $m=X_1+X_2+X_3+\ldots+X_n$. Тогда

$$Mm=M\left(\sum_{i=1}^{n} X_i\right)=\sum_{i=1}^{n} MX_i=np,$$

$$Dm = D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} DX_i = npq.$$

Воспользовавшись неравенством Чебышева, имеем

$$P(|m-np| \leqslant \varepsilon) > 1 - \frac{npq}{\varepsilon^2}. \tag{7.11}$$

§ 7.5. Центральная предельная теорема

Вспоминая доказанные выше теоремы, можно сделать вывод, что при выполнении довольно «нежестких» требований некоторые случайные величины с увеличением числа испытаний приближаются к определенным предельным значениям, не зависящим от вида распределения самих величин. Каждая из этих теорем является одной из форм закона больших чисел. В рассмотренных теоремах, а значит, и в законе больших чисел ничего не говорится о виде распределения рассматриваемой случайной величины.

Другая группа теорем теории вероятностей, которые устанавливают связь между законом распределения суммы случайных величин и его предельной формой — нормальным законом распределения, называется *центральной предельной теоремой*. Различные формы центральной предельной теоремы отличаются между собой условиями, накладываемыми на сумму рассматриваемых случайных величин. Поскольку несложные условия на практике выполняются очень часто, нормальный закон является самым распространенным среди законов распределения, наиболее часто используемым при объяснении случайных явлений природы.

Одной из теорем, относящихся к центральной предельной теореме, является теорема Ляпунова.

Теорема Ляпунова. *Распределение суммы независимых случайных величин X_i, $i=1, 2, 3, \ldots, n$, приближается к нормальному закону распределения при неограниченном увеличении n, если выполняются следующие условия:*

1) *все величины имеют конечные математические ожидания и дисперсии;*

2) *ни одна из величин по своему значению резко не отличается от всех остальных, т. е. оказывает ничтожное влияние на их сумму.*

Теорема Ляпунова имеет большое практическое применение. На опыте было установлено, что распределение суммы независимых случайных величин, у которых дисперсии не отличаются резко друг от друга, довольно быстро приближается к нормальному. Уже при числе слагаемых, большем 10, распределение суммы можно заменить нормальным. В заключение отметим, что теорема Ляпунова справедлива не только для непрерывных, но и для дискретных случайных величин. Рассмотренные ранее локальная (§ 4.2) и интегральная (§ 6.3) теоремы являются частным случаем центральной предельной теоремы.

При решении многих практических задач, связанных со случайной величиной $\bar{X}$, являющейся средним арифметическим наблюдаемых значений случайной величины X, применяется теорема Ляпунова в следующей формулировке:

если случайная величина X имеет конечные математическое ожидание MX и дисперсию DX, то распределение среднего арифметического $\bar{X} = \sum_{i=1}^{n} X_i / n$, вычисленного по наблюдавшимся значениям случайной величины в n независимых испытаниях, проведенных в одинаковых условиях, при $n \to \infty$ приближается к нормальному с математическим ожиданием MX и дисперсией DX/n, т. е.

$$P(\bar{X} < x) \to \frac{1}{\sqrt{2\pi DX/n}} \int_{-\infty}^{x} e^{-\frac{(t-MX)^2}{2DX/n}} dt, \tag{7.12}$$

или

$$\bar{X} \approx N\left(MX, \sqrt{\frac{DX}{n}}\right).$$

Воспользовавшись соотношением (7.12), нетрудно показать, что

$$P(\alpha < \bar{X} < \beta) \approx \frac{1}{\sqrt{2\pi DX/n}} \int_{\alpha}^{\beta} e^{-\frac{(x-MX)^2}{2DX/n}} dx,$$

или

$$P(\alpha < \bar{X} < \beta) \approx \frac{1}{2}\Phi(u_2) - \frac{1}{2}\Phi(u_1), \tag{7.13}$$

где

$$u_1 = \frac{\alpha - MX}{\sqrt{DX/n}}, \quad u_2 = \frac{\beta - MX}{\sqrt{DX/n}}.$$

Из формулы (6.13) следует, что

$$P(|\bar{X} - MX| < \varepsilon) \approx \Phi\left(\frac{\varepsilon}{\sqrt{DX/n}}\right). \tag{7.14}$$

ЧАСТЬ III
МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ГЛАВА 8
ВЫБОРОЧНЫЕ АНАЛОГИ ЗАКОНА РАСПРЕДЕЛЕНИЯ И ЧИСЛОВЫХ ХАРАКТЕРИСТИК СЛУЧАЙНОЙ ВЕЛИЧИНЫ

§ 8.1. Генеральная совокупность и выборка

Напомним, что предметом математической статистики является изучение случайных величин (или случайных событий) по результатам наблюдений. Для получения опытных данных необходимо провести обследование соответствующих объектов. Например, если исследователя интересует вероятность того, что диаметр валика определенного типоразмера после шлифовки окажется за пределами технического допуска, то надо знать закон распределения этого диаметра, а для этого прежде всего надо располагать набором возможных значений диаметра. Однако обследовать все валики зачастую трудно, поскольку их количество может быть велико. Поэтому приходится из всей совокупности объектов для обследования отбирать только часть, т. е. проводить выборочное обследование. В некоторых случаях обследование объектов всей совокупности практически не имеет смысла, поскольку они разрушаются в результате обследования.

○ **Пример 8.1.** Пусть на некотором комбинате выпускаются рыбные консервы. Для проверки на качество каждую банку приходится вскрывать, тем самым портить продукт. Как же в этом случае проверить качество консервного производства, если сплошное обследование всех банок невозможно?

Допустим, что комбинату к определенному сроку требуется отправить в торговую сеть определенное количество качественной продукции. Чтобы иметь представление о качестве всей отправляемой партии консервов, берут небольшую часть продукции и проверяют на качество. По полученным результатам можно судить о качестве всей продукции, не приводя в негодность всю партию консервов. ●

○ **Пример 8.2.** При проверке качества производства электролампочек последние должны находиться под напряжением довольно большое время, что, естественно, невозможно в условиях массового производства. Поэтому для проверки на стандартность подвергают контролю только небольшую часть изготовленных лампочек. Практика подтверждает, что выводы о всей совокупности объектов, сделанные на основании анализа данных наблюдения только над заведомо меньшей частью этой совокупности, бывают достаточно надежными. ●

Зачастую реально существующую совокупность объектов можно мысленно дополнить любым количеством таких же однородных объектов. Например, совокупность электромоторов определенной марки, изготовленных на данном заводе в течение квартала, можно дополнить гипотетической совокупностью таких же электромоторов, которые могут быть изготовлены во II, в III и т. д. кварталах. В соответствии с этим наблюдения над объектами такой совокупности, в результате которых «снимаются» конкретные значения случайной величины (значения изучаемого признака объекта), можно мысленно продолжать в неизменных условиях как угодно долго.

Такие совокупности объектов или совокупности значений определенной случайной величины, соответствующие каждому из этих объектов, будем называть *генеральными*.

Определение. *Совокупность всех мысленно возможных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений определенной случайной величины, или совокупность результатов всех мыслимых наблюдений, проводимых в неизменных условиях*) над одной из случайных величин, связанных с данным видом объектов, называется генеральной совокупностью.*

Как видно из определения, генеральная совокупность объектов данного вида и соответствующая совокупность значений случайной величины не различаются. Так как понятия генеральной совокупности и случайной величины связаны с наблюдениями (испытаниями) в неизменных условиях, то для простоты в дальнейшем эти понятия не будем различать. На самом деле понятие генеральной совокупности несколько шире понятия случайной величины, так как любое значение случайной величины может быть результатом нескольких наблюдений.

Генеральную совокупность будем называть *конечной* или *бесконечной* в зависимости от того, конечна или бесконечна совокупность составляющих ее элементов. Если множество значений случайной величины X бесконечно, то генеральная совокупность бесконечна. Если случайная величина дискретна и ее множество значений конечно, то генеральная совокупность может быть как конечной (например, по статистическим данным оценивается доля мальчиков среди детей, родившихся за год; здесь генеральная совокупность — это все родившиеся за год дети), так и бесконечной (если рассматривать до бесконечности непрерывное воспроизводство населения).

*) Напомним, что «неизменность условий» означает неизменность только всех тех условий, которые можно проконтролировать при проведении наблюдений. Прочие неконтролируемые условия изменяются, что приводит к случайности результатов наблюдений.

В заключение отметим, что не следует смешивать понятие генеральной совокупности с реально существующими совокупностями. Например, на склад поступила продукция некоторого цеха за месяц, что является реально существующей совокупностью, которую нельзя назвать генеральной, поскольку выпуск этой продукции можно мысленно продолжить сколь угодно долго.

Определение. *Часть отобранных объектов из генеральной совокупности (результаты наблюдений над ограниченным числом объектов из этой совокупности) называется выборочной совокупностью или выборкой.*

Число N объектов генеральной совокупности и число n объектов выборочной совокупности будем называть объемами генеральной и выборочной совокупностей соответственно. При этом будем предполагать, что $N \gg n$ (N значительно больше n). Как уже отмечалось выше, о свойствах генеральной совокупности (случайной величины X) можно судить по данным наблюдений над отобранными объектами, т. е. по выборке. Однако не всякая выборка может дать действительное представление о генеральной совокупности.

○ **Пример 8.3.** В цехе по производству специальных втулок на токарных станках работают квалифицированные токари и только начинающие. Для проверки качества продукции на контроль взята партия втулок. Если эти втулки изготовлены квалифицированным токарем, то, очевидно, представление о качестве всей продукции цеха будет «завышенным», а если втулки изготовлены начинающим токарем, то это представление будет «заниженным». ●

Для того чтобы по выборке можно было достаточно уверенно судить о случайной величине, выборка должна быть *представительной* (*репрезентативной*). Репрезентативность выборки означает, что объекты выборки достаточно хорошо представляют генеральную совокупность. Заметим, что при отборе объектов могут сыграть роль личные мотивы или психологические факторы, о которых исследователь, проводящий выборку, и не подозревает. При этом, как правило, выборка не будет репрезентативной.

Репрезентативность выборки обеспечивается случайностью отбора. Последнее означает, что любой объект выборки отобран случайно, при этом все объекты имеют одинаковую вероятность попасть в выборку. Существует несколько способов отбора, обеспечивающих репрезентативность выборки. Рассмотрим некоторые из них.

Пусть небольшие по размеру объекты генеральной совокупности находятся, например, в ящике. Каждый раз после тщательного перемешивания, если оно не является причиной деформации объектов, из ящика наудачу берут один объект. Эту операцию повторяют до тех пор, пока не образуется выборочная совокупность. Такой отбор невозможен, если

генеральная совокупность состоит из достаточно больших по размерам объектов, например из мощных электромоторов, или из таких объектов, которые при перемешивании разрушаются, например из электролампочек. Тогда поступают следующим образом. Все объекты генеральной совокупности нумеруют, а затем каждый номер записывают на отдельную карточку. После этого карточки с номерами тщательно перемешивают и из полученной пачки карточек выбирают одну наудачу. Объект, номер которого совпал с номером на карточке, считается попавшим в выборку. Такую операцию повторяют до тех пор, пока не образуется необходимая выборка. При этом можно осуществить два различных варианта выборки.

1) Каждая вынутая карточка возвращается назад в пачку, и карточки снова тщательно перемешиваются. Повторяя эту операцию необходимое число раз, можно получить выборочную совокупность, которая называется *случайной выборкой с возвратом.*

2) Каждая вынутая карточка не возвращается назад в пачку. Образованная таким способом выборка называется *случайной выборкой без возврата.*

Так как при выборке с возвратом одну и ту же карточку можно выбрать дважды, а значит, соответствующий объект придется обследовать также дважды, то эту выборку называют также *случайной повторной*. Аналогично, выборку без возврата называют *случайной бесповторной.*

При большом объеме генеральной совокупности применение карточек для организации случайной выборки затруднительно, что связано с необходимостью написания большого числа номеров, при этом хорошее перемешивание карточек трудно обеспечить. В таких случаях прибегают к помощи таблицы случайных чисел. В табл. П.8 представлены эти числа. Предположим, например, что требуется сделать для контроля выборку из генеральной совокупности большого объема, представляющей собой изготовленные заводом в течение квартала электромоторы, каждый из которых имеет четырехзначный заводской номер. Если выборка должна содержать 20 моторов, то из таблицы произвольным образом берут 20 четырехзначных чисел (можно подряд) и моторы с соответствующими номерами отправляют на контроль. В выборку могут попасть моторы с номерами 1534, 7106, 2836 и т. д. Если не обращать внимание на то, что некоторые номера могут повторяться и, следовательно, некоторые моторы должны обследоваться дважды, то выборка является, очевидно, выборкой с возвратом. Если же необходимо организовать случайную выборку без возврата, то при отборе случайных чисел из таблицы следует вновь встретившееся число пропустить.

Пусть требуется организовать выборку без возврата из

100 объектов (они все пронумерованы), содержащую семь объектов. Для этого достаточно выбрать в таблице любой столбец, а в каждом числе этого столбца — две определенные цифры, которые будут означать двузначный номер объекта. Выберем, например, третий столбец и две последние цифры чисел этого столбца. Для определенности возьмем первые семь чисел этого столбца. Они дадут следующие семь номеров объекта: 36; 02; 44; 05; 25; 41; 88.

Если объем генеральной совокупности велик, то различие между выборками с возвратом и без возврата, которые составляют ее небольшую часть, незначительно и практически не сказывается на окончательных результатах. В таких случаях, как правило, используют выборку без возврата. Если генеральная совокупность имеет не очень большой объем, то различие между указанными выборками будет существенным.

При любой выборке предполагается, что все объекты генеральной совокупности имеют в одном испытании одинаковую вероятность попасть в выборку. Убедимся на примере в том, что эта вероятность и для выборки с возвратом и для выборки без возврата не изменяется при переходе от одного испытания к другому.

○ **Пример 8.4.** В урне a белых и b черных шаров. Шары отличаются только цветом. Из урны наугад вынули два шара. Найдем вероятности двух событий: A_1 — первый шар белый, A_2 — второй шар также белый — для следующих двух случаев: выборка с возвратом и выборка без возврата.

Очевидно, для выборки с возвратом

$$P(A_1)=P(A_2)=\frac{a}{a+b}.$$

Для выборки без возврата

$$P(A_1)=\frac{a}{a+b}.$$

Найдем $P(A_2)$. Событие A_2 может наступить лишь при условии появления одного из двух следующих событий: A_1 — первый шар белый (гипотеза H_1), B_1 — первый шар черный (гипотеза H_2). Тогда по формуле полной вероятности (3.13) получим

$$P(A_2)=P(H_1)P_{H_1}(A_2)+P(H_2)P_{H_2}(A_2)=P(A_1)P_{A_1}(A_2)+P(B_1)P_{B_1}(A_2)=$$

$$=\frac{a}{a+b}\cdot\frac{a-1}{a+b-1}+\frac{b}{a+b}\cdot\frac{a}{a+b-1}=\frac{a}{a+b}=P(A_1). \; \bullet$$

Таким образом, и для выборки с возвратом, и для выборки без возврата вероятность того, что объект попадет в выборку, не изменяется при переходе от одного испытания к другому, или, иными словами, с вероятностной точки зрения условия испытаний не изменяются. Однако если в выборке с возвратом испытания независимы, то в выборке без возврата испытания таким свойством не обладают: здесь испытания зависимы. Это следует из примера 8.4. При выборке с возвратом условная вероятность $P_{A_1}(A_2)$ вытащить второй шар

белый при условии, что первый — белый, совпадает с безусловной вероятностью $P(A_2)$:

$$P_{A_1}(A_2)=\frac{a}{a+b}, \quad P(A_2)=\frac{a}{a+b}.$$

Для выборки без возврата

$$P_{A_1}(A_2)=\frac{a-1}{a+b-1}, \quad P(A_2)=\frac{a}{a+b};$$

таким образом,

$$P_{A_1}(A_2)\neq P(A_2).$$

Условие независимости является одним из основных используемых в теоремах теории вероятностей (см., например, гл. 7), поэтому в дальнейшем будем предполагать, что имеет место случайная выборка с возвратом, и при этом иметь в виду, что выражение «случайная выборка с возвратом» тождественно выражению «испытания независимы и проведены в одинаковых условиях».

После того как сделана выборка, т. е. получена выборочная совокупность объектов, все объекты этой совокупности обследуют по отношению к определенной случайной величине (или случайному событию) и в результате этого получают наблюдаемые данные.

Следующая задача математической статистики заключается в обработке результатов наблюдений.

§ 8.2. Вариационные ряды

Обычно полученные наблюдаемые данные представляют собой множество расположенных в беспорядке чисел. Просматривая это множество чисел, зачастую бывает трудно выявить какую-либо закономерность их варьирования (изменения). Для изучения закономерностей (если таковые вообще имеются) варьирования значений случайной величины опытные данные подвергают обработке.

○ **Пример 8.5.** На телефонной станции проводились наблюдения над числом X неправильных соединений в минуту. Наблюдения в течение часа дали следующие результаты: 3; 1; 3; 1; 4; 2; 2; 4; 0; 3; 0; 2; 2; 0; 2; 1; 4; 3; 3; 1; 4; 2; 2; 1; 1; 2; 1; 0; 3; 4; 1; 3; 2; 7; 2; 0; 0; 1; 3; 3; 1; 2; 4; 2; 0; 2; 3; 1; 2; 5; 1; 1; 0; 1; 1; 2; 2; 1; 1; 5. Здесь, очевидно, число X является дискретной случайной величиной (см. § 5.1), а полученные о ней сведения представляют собой статистические (наблюдаемые) данные. ●

О п р е д е л е н и е. *Операция, заключающаяся в том, что результаты наблюдений над случайной величиной, т. е. наблюдаемые значения случайной величины, располагают в порядке неубывания, называется ранжированием опытных данных.*

После проведения операции ранжирования опытные данные нетрудно объединить в группы, т. е. сгруппировать так, что в каждой отдельной группе значения случайной величины будут одинаковы. Расположив приведенные выше данные в порядке неубывания и сгруппировав их, получаем следующий ранжированный ряд данных наблюдения: 0; 0; 0; 0; 0; 0; 0; 0; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 4; 4; 4; 5; 5; 7. Из полученного ряда чисел видно, что все 60 значений случайной величины разбиты на семь групп, в пределах каждой из которых все значения случайной величины одинаковы. Таким образом, имеется семь различных значений случайной величины: 0; 1; 2; 3; 4; 5; 7. Каждое такое значение обычно называют *вариантом.*

Определение. *Значение случайной величины, соответствующее отдельной группе сгруппированного ряда наблюдаемых данных, называется вариантом, а изменение этого значения — варьированием.*

Варианты будем обозначать малыми буквами конца латинского алфавита с соответствующими порядковому номеру группы индексами x_i, y_j, z_k, ... В приведенном примере четвертая группа данных содержит 10 одинаковых значений случайной величины, равных 3, т. е. $x_4=3$. Просматривая ряд полученных данных, нетрудно убедиться, что значение случайной величины варьирует (изменяется) от 0 до 7, причем наиболее часто встречается вариант $x_2=1$.

Для каждой группы сгруппированного ряда данных можно подсчитать их численность, т. е. определить число, которое показывает, сколько раз встречается соответствующий вариант в ряде наблюдений. Такие числа называют *частотой варианта.*

Определение. *Численность отдельной группы сгруппированного ряда наблюдаемых данных называется частотой или весом соответствующего варианта и обозначается m_i, где i — индекс варианта.*

Например, для варианта x_4 частота $m_4=10$. В ряде случаев представляет практический интерес относительная частота того или иного варианта, называемая *частостью.*

Определение. *Отношение частоты данного варианта к общей сумме частот всех вариантов называется частостью или долей этого варианта и обозначается $\hat{p}_i$, где i — индекс варианта.*

Так как в приведенном примере общая сумма всех частот равна 60, то нетрудно подсчитать, что $\hat{p}_5=0{,}1$. Таким образом, частость выражает долю (удельный вес) данного варианта во всей совокупности наблюдаемых значений случайной величины.

По определению частости имеем

$$\hat{p}_i=\frac{m_i}{\sum\limits_{i=1}^{\nu} m_i}, \tag{8.1}$$

где ν — число вариантов. Предполагая, что $n=\sum\limits_{i=1}^{\nu} m_i$, где n — объем выборки, последнюю формулу перепишем в виде

$$\hat{p}_i=\frac{m_i}{n}. \tag{8.2}$$

Теперь нетрудно заметить, что частость $\hat{p}_i$ является статистической вероятностью появления варианта x_i (см. § 2.4). Естественно считать частость $\hat{p}_i$ выборочным аналогом (вычисленной по выборочным данным) вероятности p_i появления значения x_i случайной величины X, так как частость (статистическая вероятность) $\hat{p}_i$ обладает свойством устойчивости, или, иначе, при выполнении определенных условий (см. § 7.4) стремится по вероятности к вероятности p_i.

Подсчитав частоты и частости для каждого варианта, представим наблюдаемые данные в виде табл. 8.1, где в первой строке расположены индексы вариантов i, во второй — варианты x_i, в третьей — соответствующие частоты m_i, в четвертой — соответствующие частости $\hat{p}_i$.

Определение. *Дискретным вариационным рядом распределения называется ранжированная совокупность вариантов x_i с соответствующими им частотами m_i или частостями $\hat{p}_i$.*

Как следует из определения, в табл. 8.1 представлен дискретный вариационный ряд распределения 60 неправильных соединений по числу этих соединений в минуту.

Пусть для дискретной случайной величины X по выборочным данным построен дискретный вариационный ряд, в котором каждому варианту x_i ставится в соответствие его частость $\hat{p}_i$. Тогда, учитывая отмеченную связь между $\hat{p}_i$ и p_i, естественно считать данный ряд выборочным аналогом ряда распределения (см. § 5.2) упомянутой случайной величины, тем более что Σp_i и $\Sigma \hat{p}_i$ равны 1.

Если изучаемая случайная величина является непрерывной, то ранжирование и группировка наблюдаемых значений зачастую не позволяют выделить характерные черты варьирования ее значений. Это объясняется тем, что отдельные значения

Таблица 8.1

Индекс	i	1	2	3	4	5	6	7
Число неправильных соединений в минуту	x_i	0	1	2	3	4	5	7
Частота	m_i	8	17	16	10	6	2	1
Частость	$\hat{p}_i$	8/60	17/60	16/60	10/60	6/60	2/60	1/60

случайной величины могут как угодно мало отличаться друг от друга и поэтому в совокупности наблюдаемых данных одинаковые значения величины могут встречаться редко, а частоты вариантов мало отличаются друг от друга.

Нецелесообразно также построение дискретного ряда для дискретной случайной величины, число возможных значений которой велико. В подобных случаях следует построить интервальный (вариационный) ряд распределения. Для построения такого ряда весь интервал варьирования наблюдаемых значений случайной величины разбивают на ряд частичных интервалов и подсчитывают частоту попадания значений величины в каждый частичный интервал.

Определение. *Интервальным вариационным рядом называется упорядоченная совокупность интервалов варьирования значений случайной величины с соответствующими частотами или частостями попаданий в каждый из них значений величины.*

Построение интервального вариационного ряда рассмотрим на примере.

Пример 8.6. При измерении диаметра валиков после шлифовки получены следующие результаты:

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

Для построения интервального ряда необходимо определить величину частичных интервалов, на которые разбивается весь интервал варьирования наблюдаемых значений случайной величины. Считая, что все частичные интервалы имеют одну и ту же длину, для каждого интервала следует установить его верхнюю и нижнюю границы, а затем в соответствии с полученной упорядоченной совокупностью частичных интервалов сгруппировать результаты наблюдений. Длину частичного интервала h следует выбрать так, чтобы построенный ряд не был громоздким и в то же время позволял выявить характерные черты изменения значений случайной величины, т. е. характерные черты изучаемого явления.

Просматривая приведенные выше результаты наблюдений, находим, что наибольшим значением случайной величины

($x_{\text{наиб}}$) является 6,83, а наименьшим ($x_{\text{наим}}$) — 6,68. Найдем размах варьирования R. Имеем

$$R = x_{\text{наиб}} - x_{\text{наим}} = 6{,}83 - 6{,}68 = 0{,}15.$$

Теперь следует выбрать число интервалов ν. Для того чтобы вариационный ряд не был слишком громоздким, обычно число интервалов берут от 7 до 11. Положим предварительно $\nu = 7$, тогда длина частичного интервала $h = R/\nu = 0{,}15/7 \approx 0{,}0214$. Однако в данном случае более удобно взять $h = 0{,}02$. Для более точного определения величины частичного интервала можно воспользоваться формулой Стерджеса

$$h = (x_{\text{наиб}} - x_{\text{наим}})/(1 + 3{,}322 \lg n). \qquad (8.2)$$

Если окажется, что h — дробное число, то за длину частичного интервала следует брать либо ближайшее целое число, либо ближайшую простую дробь. Имеем

$$h = 0{,}15/(1 + 3{,}322 \lg 200) \approx 0{,}0174 \approx 0{,}02.$$

За начало первого интервала рекомендуется брать величину $x_{\text{нач}} = x_{\text{наим}} - 0{,}5h$. В данном случае

$$x_{\text{нач}} = 6{,}68 - 0{,}5 \cdot 0{,}02 = 6{,}67.$$

Конец последнего интервала ($x_{\text{кон}}$) должен удовлетворять условию

$$x_{\text{кон}} - h \leqslant x_{\text{наиб}} < x_{\text{кон}}.$$

Промежуточные интервалы получают прибавляя к концу предыдущего интервала длину частичного интервала h (в рассматриваемом случае $h = 0{,}02$).

Теперь, просматривая результаты наблюдений, определяем, сколько значений признака попало в каждый конкретный интервал. При этом в интервал включают значения случайной величины, большие или равные нижней границе и меньшие верхней границы. Это можно выполнить следующим образом. Границы полученной последовательности интервалов записывают в столбец, а затем, просматривая данные в том порядке, в котором они были получены, проставляют справа от соответствующего интервала точки или черточки. Подсчет частот для каждого интервала удобно проводить методом «конвертиков». Суть этого метода состоит в том, что попадание значения случайной величины в тот или иной интервал отмечается точкой, а также и черточкой. В результате каждому десятку будет соответствовать фигура, похожая на конверт:

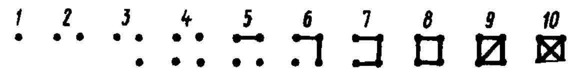


Таблица 8.2

№ п/п	Диаметр валика после шлифовки (интервалы), мм	Рабочее поле	Частота m_i	Частость $\hat{p}_i$
1	6,67—6,69		2	0,010
2	6,69—6,71		15	0,075
3	6,71—6,73		17	0,085
4	6,73—6,75		44	0,220
5	6,75—6,77		52	0,260
6	6,77—6,79		44	0,220
7	6,79—6,81		14	0,070
8	6,81—6,83		11	0,055
9	6,83—6,85		1	0,005
Σ			200	1

Шкала интервалов и группировка результатов наблюдений методом «конвертиков» для рассматриваемого случая приведены в табл. 8.2, которая задает искомый интервальный ряд распределения. В табл. 8.2 частота m_i показывает, в скольких наблюдениях случайная величина приняла значения, принадлежащие тому или иному интервалу, причем нижний конец интервала входит в него, а верхний — нет. Например, из таблицы видно, что в четвертый интервал попало 44 значения случайной величины. Однако определенно нельзя сказать, каковы эти значения, зато можно утверждать, что все эти значения принадлежат интервалу $[6{,}73;\ 6{,}75[$. Такие частоты обычно называют интервальными, а их отношение к общему числу наблюдений — интервальными частостями.

При вычислении интервальных частостей округление результатов следует проводить таким образом, чтобы общая сумма частостей была равна 1.

Иногда интервальный вариационный ряд для простоты исследований условно заменяют дискретным. В этом случае серединное значение i-го интервала принимают за вариант x_i, а соответствующую интервальную частоту m_i — за частоту этого интервала.

§ 8.3. Выборочные аналоги интегральной и дифференциальной функций распределения. Полигон и гистограмма

Как известно (см. гл. 5), закон распределения (или просто распределение) случайной величины можно задать различными способами. Например, дискретную случайную величину можно задать с помощью или ряда распределения, или интегральной функции, а непрерывную случайную величину — с помощью или интегральной, или дифференциальной функции. Рассмотрим выборочные аналоги этих двух функций.

В теории вероятностей для характеристики распределения случайной величины X служит интегральная функция распределения $F(x)=P(X<x)$. В дальнейшем, если величина X распределена по некоторому закону $F(x)$, будем говорить, что и генеральная совокупность распределена по закону $F(x)$. Введем выборочный аналог функции $F(x)$.

Пусть имеется выборочная совокупность значений некоторой случайной величины X объема n и каждому варианту из этой совокупности поставлена в соответствие его частость. Пусть, далее, x — некоторое действительное число, а m_x — число выборочных значений случайной величины X, меньших x. Тогда число m_x/n является частостью наблюдаемых в выборке значений величины X, меньших x, т. е. частостью появления события $X<x$. При изменении x в общем случае будет изменяться и величина m_x/n. Это означает, что относительная частота m_x/n является функцией аргумента x. А так как эта функция находится по выборочным данным, полученным в результате опытов, то ее называют *выборочной* или *эмпирической*.

Определение. *Выборочной функцией распределения (или функцией распределения выборки) называется функция* $\hat{F}(x)$, *задающая для каждого значения x относительную частоту события* $X<x$.

Итак, по определению, $\hat{F}(x)=m_x/n$, где n — объем выборки, m_x — число выборочных значений величины X, меньших x. В отличие от выборочной функции $\hat{F}(x)$ интегральную функцию $F(x)$ генеральной совокупности называют *теоретической функцией распределения*. Главное различие функций $F(x)$ и $\hat{F}(x)$ состоит в том, что теоретическая функция распределения $F(x)$ определяет вероятность события $X<x$, а выборочная функция — относительную частоту этого события.

Свойство статистической устойчивости частоты, обоснованное теоремой Бернулли, оправдывает целесообразность использования функции $\hat{F}(x)$ при больших n в качестве приближенного значения неизвестной функции $F(x)$.

В заключение отметим, что функция $F(x)$ и ее выборочный аналог $\hat{F}(x)$ обладают одинаковыми свойствами. Действительно, из определения функции $\hat{F}(x)$ имеем следующие свойства:

Таблица 8.3

x	$\hat{F}(x)$
$x \leqslant 0$	0
$0 < x \leqslant 1$	$\hat{p}_1 = 8/60$
$1 < x \leqslant 2$	$\hat{p}_1 + \hat{p}_2 = 25/60$
$2 < x \leqslant 3$	$\hat{p}_1 + \hat{p}_2 + \hat{p}_3 = 41/60$
$3 < x \leqslant 4$	$\hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4 = 51/60$
$4 < x \leqslant 5$	$\hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4 + \hat{p}_5 = 57/60$
$5 < x \leqslant 7$	$\hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4 + \hat{p}_5 + \hat{p}_6 = 59/60$
$x > 7$	$\hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4 + \hat{p}_5 + \hat{p}_6 + \hat{p}_7 = 60/60 = 1$

1°. $0 \leqslant \hat{F}(x) \leqslant 1$.

2°. $\hat{F}(x)$ — *неубывающая функция.*

3°. $\hat{F}(-\infty) = 0$, $\hat{F}(\infty) = 1$.

Такими же свойствами обладает и функция $F(x)$ (см. § 5.3).

○ **Пример 8.7.** Построим выборочную функцию распределения по данным табл. 8.1.

Объем выборки по условию равен 60, т. е. $n = 60$. Наименьший вариант равен 0, значит, $m_x = 0$ при $x \leqslant 0$. Тогда $m_x/n = 0/60 = 0$, т. е. $\hat{F}(x) = 0$ при $x \leqslant 0$. Если $0 < x \leqslant 1$, то неравенство $X < x$ выполняется при условии, что $X = 0$. Так как этот вариант встречается в выборке 8 раз, то $m_x/n = 8/60 = \hat{p}_1$, т. е. $\hat{F}(x) = 8/60$. Если $1 < x \leqslant 2$, то неравенство $X < x$ выполняется при условии, что $X = 0$ или $X = 1$. Так как вариант $x_1 = 0$ встречается 8 раз, а вариант $x_2 = 1$ — 17 раз, то $m_x/n = (8+17)/60 = 25/60$, т. е. $\hat{F}(x) = \hat{p}_1 + \hat{p}_2 = 25/60$ и т. д. В результате получаем искомую функцию распределения, значения которой представим в виде табл. 8.3. ●

График этой функции изображен на рис. 8.1. На этом графике видны все особенности выборочной функции распределения. Она не убывает, а ее значения находятся в интервале [0; 1]. Резкие «скачки» графика функции $\hat{F}(x)$, придающие ей ступенчатый вид, имеют место в тех точках, которым соответствуют наблюдаемые значения вариантов, при этом величина скачка равна частости варианта.

Функцию $\hat{F}(x)$ наряду с табличным способом задания (см. табл. 8.3) можно задать аналитически. В этом случае $\hat{F}(x)$ определяется так:

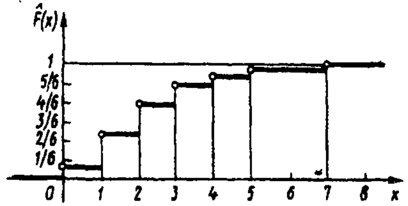


Рис. 8.1

$$\hat{F}(x) = \begin{cases} 0 & \text{при } x \leqslant x_1, \\ \sum\limits_{l=1}^{i-1} \hat{p}_l & \text{при } x_{i-1} < x \leqslant x_i, \ i = 1, 2, 3, \ldots, \nu, \\ 1 & \text{при } x > x_\nu. \end{cases} \qquad (8.4)$$

Здесь x_ν совпадает с $x_{\text{наиб}}$. Частости $\sum\limits_{l=1}^{i-1} \hat{p}_l$ обычно называются *накопленными частостями.*

В рассматриваемом примере функция $\hat{F}(x)$ построена по дискретному вариационному ряду и для дискретной случайной величины. Если результаты наблюдений представлены в виде интервального вариационного ряда, то выборочную функцию распределения построить в том виде, как это было сделано в примере 8.7, уже не представляется возможным. Рассмотрим на примере построение функции $\hat{F}(x)$ по интервальному вариационному ряду для непрерывной случайной величины.

○ **Пример 8.8.** Используя данные табл. 8.2, построим выборочную функцию распределения. Очевидно, что всех $x \in]-\infty; 6{,}67]$ функция распределения равна нулю. Пусть теперь $x \in]6{,}67; 6{,}69]$. В этом случае число m_x/n не определено, так как неизвестно, сколько выборочных значений случайной величины, принадлежащих этому интервалу, меньше x. Если $x = 6{,}69$, то $m_x = 2$. Следовательно, $\hat{F}(6{,}69) = 2/200 = 0{,}01$. Рассуждая аналогично, убеждаемся, что точками, в которых значение функции $\hat{F}(x)$ можно определить, являются правые концы интервалов и все точки интервала $[6{,}85; \infty[$. Определяем теперь значение функции $\hat{F}(x)$ в указанных точках и запишем в виде табл. 8.4. ●

Таблица 8.4

x	6,67	6,69	6,71	6,73	6,75	6,77	6,79	6,81	6,83	6,85
$\hat{F}(x)$	0	0,010	0,085	0,170	0,390	0,650	0,870	0,940	0,995	1

Так как эта таблица определяет функцию $\hat{F}(x)$ не полностью (не для всех x известны ее значения), то при графическом изображении данной функции целесообразно ее доопределить, соединив точки графика, соответствующие концам интервалов, отрезками прямой (рис. 8.2). В результате

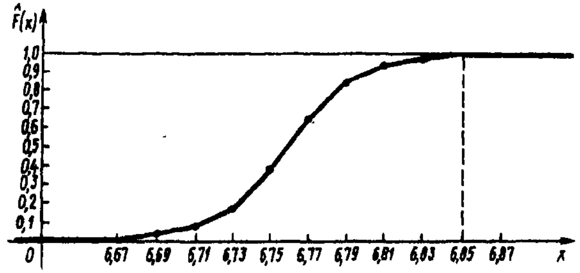


Рис. 8.2

график функции $\hat{F}(x)$ будет представлять собой непрерывную линию. Отметим, что подобный график выборочной функции $\hat{F}(x)$, дающий приближенное представление о графике теоретической функции $F(x)$, часто называют *кумулятивной кривой* (от англ. accumulation — накопление).

Как уже известно (см. § 5.4), для интегральной функции распределения $F(x)$ справедливо приближенное равенство $F(x+\Delta x)-F(x)\approx f(x)\Delta x$, где $f(x)$ — дифференциальная функция распределения или функция плотности вероятности. Из этого равенства следует, что $f(x)\approx(F(x+\Delta x)-F(x))/\Delta x$. Поэтому естественно выборочным аналогом функции $f(x)$ считать функцию

$$\hat{f}(x)=\frac{\hat{F}(x+\Delta x)-\hat{F}(x)}{\Delta x}, \qquad (8.5)$$

где $\hat{F}(x+\Delta x)-\hat{F}(x)$ — частость попадания наблюдаемых значений случайной величины X в интервал $[x, x+\Delta x[$. Таким образом, значение $\hat{f}(x)$ характеризует плотность частости на этом интервале.

Пусть наблюдаемые над непрерывной случайной величиной данные представлены в виде интервального вариационного ряда. Полагая, что $\hat{p}_i$ — частость попадания наблюдаемых значений случайной величины в интервал $[a_i;\ a_i+h[$, где h — длина частичного интервала, и учитывая равенство (8.5), для $x\in[a_i,\ a_i+h[$ запишем $\hat{f}(x)=\hat{p}_i/h$. Тогда выборочную функцию плотности $\hat{f}(x)$ можно задать соотношением

$$\hat{f}(x)=\begin{cases} 0 & \text{при } x<a_1, \\ \hat{p}_i/h & \text{при } a_i\leqslant x<a_{i+1},\ i=1, 2, 3, \dots, \nu, \\ 0 & \text{при } x\geqslant a_{\nu+1}, \end{cases} \qquad (8.6)$$

где $a_{\nu+1}$ — конец последнего ν-го интервала. В табл. 8.5 приведены значения функции $\hat{f}(x)$, построенной по наблюдаемым данным, взятым из примера 8.5.

Таблица 8.5

i	Диаметр валика после шлифовок (интервалы, мм)	Частость $\hat{p}_i$	$\hat{f}(x)$
1	6,67—6,69	0,010	0,50
2	6,69—6,71	0,075	3,75
3	6.71—6,73	0,085	4,25
4	6,73—6,75	0,220	11,0
5	6,75—6.77	0,260	13,0
6	6,77—6,79	0,220	11,0
7	6,79—6,81	0,070	3,50
8	6,81—6,83	0,055	2,75
9	6,83—6,85	0,005	0,25

Наблюдаемые данные, представленные в виде вариационного ряда, можно изобразить графически, используя не только функцию $\hat{F}(x)$. К наиболее распространенным видам графического изображения вариационных рядов относятся *полигон* и *гистограмма*. Графическое изображение рядов с помощью полигона или гистограммы позволяет получить наглядное представление о закономерности варьирования наблюдаемых значений случайной величины.

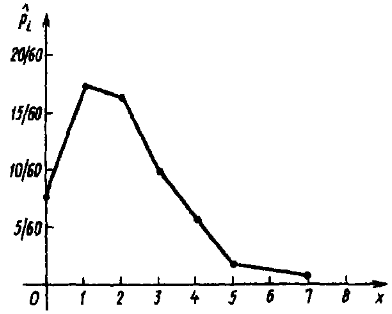


Рис. 8.3

Полигон обычно используют для изображения дискретного вариационного ряда. Для его построения в прямоугольной системе координат наносят точки с координатами $(x_i;\ m_i)$ или $(x_i;\ \hat{p}_i)$, где x_i — значение i-го варианта, а $m_i(\hat{p}_i)$ — соответствующие частоты (частости). Затем отмеченные точки соединяют отрезками прямой линии. Полученная ломаная называется *полигоном*. На рис. 8.3 изображен полигон частостей, при построении которого использованы данные табл. 8.1.

Если полигон частостей построен по дискретному вариационному ряду дискретной случайной величины, то его называют *многоугольником распределения частостей*, который является выборочным аналогом многоугольника распределения вероятностей (см. § 5.2). Заметим, что сумма ординат многоугольника распределения частостей, как и у многоугольника распределения вероятностей, равна 1, так как $\sum\hat{p}_i=1$.

Гистограмма служит только для изображения интервальных вариационных рядов. Для ее построения в прямоугольной системе координат на оси Ox откладывают отрезки, изображающие частичные интервалы варьирования, и на этих отрезках, как на основаниях, строят прямоугольники с высотами, равными частотам или частостям соответствующих интервалов. В результате такой операции получают ступенчатую фигуру, состоящую из прямоугольников, которую называют гистограммой.

На рис. 8.4 изображена гистограмма частостей, при построении которой были использованы данные табл. 8.2.

Для графического изображения интервального вариационного ряда можно использовать полигон, если этот ряд преобразовать в дискретный. В этом случае интервалы заменяют их серединными значениями и ставят им в соответствие интервальные частоты (частости). Для полученного дискретного ряда строят полигон (см. рис. 8.4, пунктирная линия).

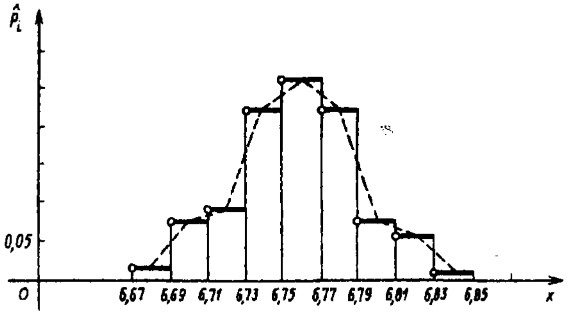


Рис. 8.4

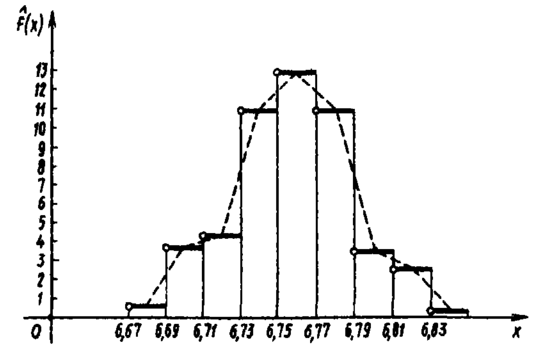


Рис. 8.5

Изобразим функцию $\hat{f}(x)$, значения которой приведены в табл. 8.5, графически (рис. 8.5). Полученная гистограмма внешне мало чем отличается от гистограммы, изображенной на рис. 8.4. Полигон, представленный на рис. 8.5 (пунктирная линия), дает первоначальное представление о дифференциальной функции распределения.

§ 8.4. Статистические характеристики вариационных рядов

Построив вариационный ряд и изобразив его графически, можно получить первоначальное представление о закономерностях, имеющих место в ряду наблюдений. Однако на практике зачастую этого недостаточно. Такая ситуация возникает, когда следует уточнить те или иные сведения о ряде распределения или когда имеется необходимость сравнить два ряда и более. При этом следует сравнивать однотипные вариационные ряды, т. е. такие ряды, которые получены при обработке сравнимых статистических данных.

Например, можно сравнить распределения длины втулок, изготовленных на двух однотипных станках-автоматах, или

распределения количества отказов определенных электронных устройств, изготовленных на разных заводах. Обычно графики таких распределений имеют почти одинаковый вид. Особенно это относится к кумулятивной кривой.

Сравниваемые распределения могут существенно отличаться друг от друга. Они могут иметь различные средние значения случайной величины, вокруг которых группируются в основном остальные значения, или различаться рассеиванием данных наблюдений вокруг указанных значений и т. д. Поэтому для дальнейшего изучения изменения значений случайной величины используют числовые характеристики вариационных рядов. Поскольку эти характеристики вычисляются по статистическим данным (по данным, полученным в результате наблюдений), их обычно называют *статистическими характеристиками* или *оценками.*

Пусть собранный и обработанный статистический материал представлен в виде вариационного ряда. Теперь результаты наблюдений над случайной величиной следует подвергнуть анализу и выявить характерные особенности поведения случайной величины. Для этого удобнее всего выделить некоторые постоянные, которые представляли бы вариационный ряд в целом и отражали присущие изучаемой совокупности закономерности.

Некоторые из этих постоянных отличаются тем, что вокруг них концентрируются остальные результаты наблюдений. Такие величины называются *средними величинами.* К ним относятся среднее арифметическое, среднее геометрическое, среднее гармоническое и т. д. Однако эти характеристики не отражают «величину изменчивости» наблюдаемых данных, например величину разброса значений признака вокруг среднего арифметического. Другими словами, упомянутые средние величины не отражают вариацию.

Для характеристики изменчивости случайной величины, т. е. вариации, служат показатели вариации. К ним относятся размах варьирования R, среднее квадратическое отклонение, дисперсия и т. д.

§ 8.5. Среднее арифметическое и его свойства

Простейшей из средних величин является среднее арифметическое, которое уже упоминалось в гл. 5 и 7. Оно проще других и по смыслу, и по свойствам, и по способу получения. Пусть $x_1, x_2, \ldots, x_n$ — данные наблюдений над случайной величиной X.

Определение. *Средним арифметическим $\bar{X}$ наблюдаемых значений случайной величины X называется частное от деления суммы всех этих значений на их число, т. е.*

$$\bar{X}=\frac{x_1+x_2+\ldots+x_n}{n}=\frac{\sum_{i=1}^{n} x_i}{n}. \tag{8.7}$$

Если данные наблюдений представлены в виде дискретного ряда, где x_1, x_2, $x_3, \ldots, x_\nu$ — наблюдаемые варианты, а $m_1, m_2, m_3, \ldots, m_\nu$ — соответствующие им частоты, причем $\sum_{i=1}^{\nu} m_i = n$, то, по определению,

$$\bar{X} = \frac{x_1 m_1 + x_2 m_2 + x_3 m_3 + \ldots + x_\nu m_\nu}{m_1 + m_2 + m_3 + \ldots + m_\nu} = \frac{\sum_{i=1}^{\nu} x_i m_i}{n}. \tag{8.8}$$

Вычисленное по формуле (8.8) среднее арифметическое называется *взвешенным*, так как частоты m_i называются *весами*, а операция умножения x_i на m_i — взвешиванием.

Для интервального вариационного ряда за x_i принимают середину i-го интервала, а за m_i — соответствующую интервальную частоту. В этом случае значения средних арифметических, вычисленных по формулам (8.7) и (8.8), могут не совпадать, так как в формуле (8.8) значения случайной величины внутри каждого интервала принимаются равными серединам интервалов, в то время как эти значения могут быть произвольно расположены в интервале. Возникающая при этом погрешность тем меньше, чем равномернее распределены значения признака внутри каждого интервала, т. е. если они не группируются вокруг произвольных точек и не сосредоточиваются у нижних или верхних границ интервалов.

Преобразуем формулу (8.8) (см. § 5.5). Имеем

$$\bar{X} = \frac{\sum_{i=1}^{\nu} x_i m_i}{n} = \sum_{i=1}^{\nu} x_i \frac{m_i}{n} = \sum_{i=1}^{\nu} x_i \hat{p}_i. \tag{8.9}$$

Заметим, что среднее арифметическое — величина той же размерности, что значения случайной величины.

○ **Пример 8.9.** По данным, приведенным в табл. 8.1, вычислим среднее арифметическое числа неправильных соединений в минуту. Среднее арифметическое вычислим по формуле (8.8). Имеем

$$\bar{X} = (0 \cdot 8 + 1 \cdot 17 + 2 \cdot 16 + 3 \cdot 10 + 4 \cdot 6 + 5 \cdot 2 + 7 \cdot 1)/60 = 2. \ ●$$

○ **Пример 8.10.** По данным, приведенным в табл. 8.2, вычислим среднее арифметическое диаметра валика. Для вычислении воспользуемся формулой (8.9) среднего арифметического:

$$\bar{X} = 6{,}68 \cdot 0{,}010 + 6{,}70 \cdot 0{,}075 + 6{,}72 \cdot 0{,}085 + 6{,}74 \cdot 0{,}220 + 6{,}76 \cdot 0{,}260 + \\ + 6{,}78 \cdot 0{,}220 + 6{,}80 \cdot 0{,}070 + 6{,}82 \cdot 0{,}055 + 6{,}84 \cdot 0{,}005 = 6{,}7578.$$

Используя несгруппированные данные из примера 8.6 и формулу (8.7), имеем

$$\bar{X} = (6{,}75 + 6{,}77 + 6{,}77 + \ldots + 6{,}77 + 6{,}75 + 6{,}78)/200 = 6{,}754. \ ●$$

Полученные результаты говорят о том, что точное значение $\bar{X}$, вычисленное по формуле (8.7), равно 6,754, а значение

взвешенного $\bar{X}$, найденного по формуле (8.9) (частный случай формулы (8.8)), равно 6,7578. Расхождение, возникшее в результате замены данных наблюдений серединами соответствующих интервалов, составляет 0,0038.

Рассмотрим основные свойства среднего арифметического.

1°. *Среднее арифметическое алгебраической суммы соответствующих друг другу значений, принадлежащих двум группам наблюдений, равна алгебраической сумме средних арифметических этих групп, т. е.*

$$\overline{X \pm Y} = \bar{X} \pm \bar{Y}.$$

□ Пусть наблюдаемые значения над случайными величинами X и Y представлены в виде дискретных вариационных рядов. Так как каждому значению x_i случайной величины X соответствует одно определенное значение y_i случайной величины Y, то численности соответствующих вариантов x_i, y_i и варианта $x_i \pm y_i$ совпадают. Тогда

$$\overline{X \pm Y} = \frac{\sum_{i=1}^{\nu} (x_i \pm y_i) m_i}{\sum_{i=1}^{\nu} m_i} = \frac{\sum_{i=1}^{\nu} x_i m_i \pm \sum_{i=1}^{\nu} y_i m_i}{n} =$$

$$= \frac{\sum_{i=1}^{\nu} x_i m_i}{n} \pm \frac{\sum_{i=1}^{\nu} y_i m_i}{n} = \bar{X} \pm \bar{Y}. \ ■$$

Доказанное свойство можно распространить на любое конечное число групп.

2°. *Если ряд наблюдений состоит из двух непересекающихся групп наблюдений, то среднее арифметическое $\bar{Z}$ всего ряда наблюдений равно взвешенному среднему арифметическому групповых средних $\bar{X}$ и $\bar{Y}$, причем весами являются объемы групп n_1 и n_2 соответственно, т. е.*

$$\bar{Z} = \frac{\bar{X} n_1 + \bar{Y} n_2}{n_1 + n_2}.$$

□ Пусть имеются две группы результатов наблюдений, причем первая группа, состоящая из значений x_1, x_2, $x_3, \ldots, x_{n_1}$, имеет численность n_1, а вторая, состоящая из y_1, y_2, $y_3, \ldots, y_{n_2}$, — численность n_2. Пусть, далее, среднее арифметическое первой группы равно $\bar{X}$, а второй — $\bar{Y}$. Тогда по определению среднего арифметического имеем

$$\bar{X} = \frac{\sum_{i=1}^{n_1} x_i}{n_1}; \quad \bar{Y} = \frac{\sum_{j=1}^{n_2} y_j}{n_2},$$

откуда

$$\sum_{i=1}^{n_1} x_i=\bar{X}n_1, \quad \sum_{j=1}^{n_2} y_j=\bar{Y}n_2.$$

Окончательно имеем

$$\bar{Z}=\frac{x_1+x_2+x_3+\ldots+x_{n_1}+y_1+y_2+y_3+\ldots+y_{n_2}}{n_1+n_2}=$$

$$=\frac{\sum_{i=1}^{n_1} x_i+\sum_{j=1}^{n_2} y_j}{n_1+n_2}=\frac{\bar{X}n_1+\bar{Y}n_2}{n_1+n_2}. \blacksquare$$

3⁰. *Среднее арифметическое постоянной равно самой постоянной, т. е.* $\bar{C}=C$.

Рассмотрим случайную величину $Z=CX$, наблюдаемые значения которой получаются, если каждое наблюдаемое значение величины X умножить на одно и то же число C, т. е. $z_i=Cx_i$.

4⁰. *Если все результаты наблюдений умножить на одно и то же число, то имеет место равенство* $\bar{Z}=\overline{CX}=C\bar{X}$, *т. е. постоянную можно выносить за знак среднего арифметического.*

□ Если данные наблюдений представлены в виде дискретного вариационного ряда, то при умножении каждого варианта x_i на число C его частота m_i не изменится, т. е. произведение Cx_i и вариант x_i имеют одну и ту же частоту. Тогда

$$\bar{Z}=\overline{CX}=\frac{\sum_{i=1}^{\nu} Cx_i m_i}{n}=\frac{C\sum_{i=1}^{\nu} x_i m_i}{n}=C\bar{X}. \blacksquare$$

5⁰. *Сумма отклонений* $x_i-\bar{X}$ *результатов наблюдений от их среднего арифметического равна нулю.*

□ Пусть результаты наблюдений над случайной величиной X представлены в виде дискретного вариационного ряда, причем каждый вариант x_i имеет частоту m_i. Тогда, очевидно, $x_i-\bar{X}$ также будет иметь частоту m_i. Далее имеем

$$\sum_{i=1}^{\nu}(x_i-\bar{X})m_i=\sum_{i=1}^{\nu} x_i m_i-\sum_{i=1}^{\nu}\bar{X}m_i=\sum_{i=1}^{\nu} x_i m_i-\bar{X}\sum_{i=1}^{\nu} m_i=$$

$$=\sum_{i=1}^{\nu} x_i m_i-\frac{\sum_{i=1}^{\nu} x_i m_i}{\sum_{i=1}^{\nu} m_i}\sum_{i=1}^{\nu} m_i=\sum_{i=1}^{\nu} x_i m_i-\sum_{i=1}^{\nu} x_i m_i=0. \blacksquare$$

Рассмотрим случайную величину $Z=X\pm C$, наблюдаемые значения которой получаются, если каждое наблюдаемое

значение величины X увеличить (уменьшить) на одно и то же число $C>0$, т. е. $z_i=x_i\pm C$.

6⁰. *Если все результаты наблюдений увеличить (уменьшить) на одно и то же число, то среднее арифметическое увеличится (уменьшится) на то же число, т. е.*

$$\bar{Z}=\overline{X\pm C}=\bar{X}\pm C.$$

$$\square\ \overline{X\pm C}=\bar{X}\pm\bar{C}=\bar{X}\pm C. \blacksquare$$

7⁰. *Если все частоты вариантов умножить на одно и то же число, то среднее арифметическое не изменится.*

□ Рассмотрим случайную величину Z, все варианты z_i которой совпадают с соответствующими вариантами x_i случайной величины X, при этом частота каждого варианта z_i равна Cm_i, где m_i — частота соответствующего варианта x_i. Тогда для Z имеем

$$\bar{Z}=\frac{\sum_{i=1}^{\nu} z_i Cm_i}{\sum_{i=1}^{\nu} Cm_i}=\frac{C\sum_{i=1}^{\nu} x_i m_i}{C\sum_{i=1}^{\nu} m_i}=\frac{\sum_{i=1}^{\nu} x_i m_i}{n}=\bar{X}. \blacksquare$$

Сравним теперь математическое ожидание, являющееся числовой характеристикой случайной величины, со средним арифметическим ее наблюдаемых значений, которое, в свою очередь, является статистической характеристикой вариационного ряда распределения этих значений.

Как известно (см. § 5.5), математическое ожидание дискретной случайной величины определяется формулой $MX=\sum_{i=1}^{n} x_i p_i$, а среднее арифметическое — формулой $\bar{X}=\sum_{i=1}^{\nu} x_i\hat{p}_i$. При сравнении этих формул очевидна их внешняя схожесть. Однако в формуле математического ожидания x_i — возможные значения случайной величины, которые могут появиться в результате наблюдений (опытов), а p_i — их вероятности. В формуле среднего арифметического x_i — конкретные значения — варианты случайной величины, полученные в результате наблюдений (выборки), а $\hat{p}_i$ — относительная частота этих значений.

Математическое ожидание, как следует из определения, является постоянной величиной в отличие от среднего арифметического, которое не обладает этим свойством. Действительно, проводя различные серии из n испытаний, можно получить различные серии конкретных значений случайной величины, которые в общем случае будут отличаться вычисленными по этим сериям средними арифметическими. Здесь следует отметить, что при выполнении определенных условий среднее арифметическое наблюдаемых значений случайной величины обладает свойством устойчивости и при увеличении числа наблюдений n сходится по вероятности к математическому ожиданию этой величины (см. теорему Чебышева).

Математическое ожидание и среднее арифметическое обладают аналогичными свойствами. Это можно сказать про свойства 1^0, 3^0. 4^0 и 5^0. Таким образом, несмотря на принципиальные различия, математическое ожидание и среднее арифметическое имеют между собой много общего и поэтому естественно считать последнее выборочным аналогом математического ожидания.

§ 8.6. Выборочная дисперсия и ее свойства

В гл. 5 одной из рассмотренных характеристик была дисперсия. Как известно, дисперсия случайной величины X является мерой рассеивания ее возможных значений вокруг математического ожидания. Так как среднее арифметическое является выборочным аналогом математического ожидания, то имеет смысл ввести подобную характеристику и для вариационных рядов, которая оценивала бы величину рассеивания наблюдаемых данных $x_1, x_2, x_3, \ldots, x_n$ вокруг их среднего арифметического и была бы связана с отклонениями $x_i - \bar{X}$.

На величину рассеивания наблюдаемых значений случайной величины X вокруг их среднего арифметического $\bar{X}$ будет влиять каждое отклонение $x_i - \bar{X}$, однако сумма всех этих отклонений не может быть мерой рассеивания, так как эта сумма равна нулю (свойство 5^0 среднего арифметического). Чтобы на оценку меры рассеивания влияли все указанные отклонения и их знаки при этом не играли роли, целесообразно использовать не сами отклонения, а их квадраты. Лучше всего для этой цели взять среднее арифметическое величин $(x_i - \bar{X})^2$.

Определение. *Выборочной дисперсией значений случайной величины X называется среднее арифметическое квадратов отклонений наблюдаемых значений этой величины от их среднего арифметического (обозначение $\hat{D}X$ или $\hat{\sigma}_X^2$):*

$$\hat{D}X = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}. \tag{8.10}$$

Если данные наблюдений представлены в виде дискретного вариационного ряда, причем $x_1, x_2, x_3, \ldots, x_\nu$ — наблюдаемые варианты, а $m_1, m_2, m_3, \ldots, m_\nu$ — соответствующие им частоты, то выборочная дисперсия определяется формулой

$$\hat{D}X = \frac{\sum_{i=1}^{\nu}(x_i - \bar{X})^2 m_i}{n}, \tag{8.11}$$

где $n = \sum_{i=1}^{\nu} m_i$ — объем выборки.

Используя равенство $\hat{p}_i = m_i/n$, последнюю формулу перепишем в виде

$$\hat{D}X = \sum_{i=1}^{\nu}(x_i - \bar{X})^2 \hat{p}_i. \tag{8.12}$$

Вычисленная по формулам (8.11) и (8.12) дисперсия называется *взвешенной выборочной дисперсией*. Если данные представлены в виде интервального ряда, то вычислить дисперсию по формулам (8.11) и (8.12) не представляется возможным, так как неизвестны конкретные значения случайной величины, попавшие в тот или иной интервал. В этом случае интервальный ряд заменяют дискретным и затем используют указанные формулы. Подобная операция уже была применена при вычислении среднего арифметического.

○ **Пример 8.11.** Используя данные табл. 8.1, вычислим выборочную дисперсию по формуле (8.11), учитывая при этом, что $\bar{X} = 2$ (см. пример 8.9). Имеем

$\hat{D}X = [(0-2)^2 \cdot 8 + (1-2)^2 \cdot 17 + (2-2)^2 \cdot 16 + (3-2)^2 \cdot 10 + (4-2)^2 \cdot 6 + (5-2)^2 \cdot 2 + (7-2)^2]/60 = 2{,}1$. ●

○ **Пример 8.12.** Используя данные табл. 8.2, вычислим выборочную дисперсию, используя формулу (8.12) и учитывая, что взвешенное среднее арифметическое равно 6,7578 (см. пример 8.10). Находим

$\hat{D}X = (6{,}68 - 6{,}7578)^2 \cdot 0{,}010 + (6{,}70 - 6{,}7578)^2 \cdot 0{,}075 + (6{,}72 - 6{,}7578)^2 \cdot 0{,}085 + (6{,}74 - 6{,}7578)^2 \cdot 0{,}220 + (6{,}76 - 6{,}7578)^2 \cdot 0{,}260 + (6{,}78 - 6{,}7578)^2 \cdot 0{,}220 + (6{,}80 - 6{,}7578)^2 \cdot 0{,}070 + (6{,}82 - 6{,}7578)^2 \cdot 0{,}055 + (6{,}84 - 6{,}7578)^2 \cdot 0{,}005 = 0{,}00098316 \approx 0{,}001$. ●

Выборочная дисперсия обладает одним существенным недостатком: если среднее арифметическое выражается в тех же единицах, что и значения случайной величины, то, как следует из формул, задающих дисперсию, последняя выражается уже в квадратных единицах. Этого недостатка можно избежать, взяв в качестве меры рассеивания арифметический квадратный корень из дисперсии.

Определение. *Выборочным средним квадратическим отклонением называется арифметический квадратный корень из выборочной дисперсии (обозначение $\hat{\sigma}_X$).*

Среднее квадратическое отклонение можно выразить следующей формулой:

$$\hat{\sigma}_X = \sqrt{\hat{D}X} = \sqrt{\frac{\sum_{i=1}^{\nu}(x_i - \bar{X})^2}{n}} \tag{8.13}$$

или

$$\hat{\sigma}_X = \sqrt{\hat{D}X} = \sqrt{\frac{\sum_{i=1}^{\nu}(x_i - \bar{X})^2 m_i}{n}}. \tag{8.14}$$

Рассмотрим основные свойства выборочной дисперсии, считая при этом, что наблюдаемые данные представлены в виде дискретного вариационного ряда.

1^0. *Дисперсия постоянной величины равна нулю.*

□ Так как постоянная C принимает только одно значение, равное C и $\bar{C}=C$ (см. свойство 3^0), то, по определению,

$$\hat{D}C=\frac{\sum_{i=1}^{n}(C-\bar{C})^2}{n}=(C-C)^2=0. \blacksquare$$

Этого следовало ожидать, поскольку дисперсия является показателем рассеивания данных вокруг их среднего арифметического, а постоянная принимает только одно значение и поэтому никакой речи о рассеивании быть не может.

2^0. *Если все результаты наблюдений увеличить (уменьшить) на одно и то же число C, то дисперсия и среднее квадратическое отклонение не изменятся, т. е.*

$$\hat{D}(X\pm C)=\hat{D}X \text{ и } \hat{\sigma}_{X\pm C}=\hat{\sigma}_X.$$

□ Для случайной величины $X\pm C$ по определению имеем

$$\hat{D}(X\pm C)=\frac{\sum_{i=1}^{\nu}[(x_i\pm C)-(\overline{X\pm C})]^2 m_i}{n}=$$

$$=\frac{\sum_{i=1}^{\nu}[(x_i\pm C)-(\bar{X}\pm C)]^2 m_i}{n}=\frac{\sum_{i=1}^{\nu}(x_i-\bar{X})^2 m_i}{n}=\hat{D}X. \blacksquare$$

3^0. *Если все результаты наблюдений умножить на одно и то же число, то имеет место равенство*

$$\hat{D}(CX)=C^2\hat{D}X \text{ или } \hat{\sigma}_{CX}=|C|\,\hat{\sigma}_X.$$

□ Для случайной величины CX по определению выборочной дисперсии имеем

$$\hat{D}(CX)=\frac{\sum_{i=1}^{\nu}(Cx_i-\overline{CX})^2 m_i}{n}=\frac{\sum_{i=1}^{\nu}(Cx_i-C\bar{X})^2 m_i}{n}=$$

$$=\frac{\sum_{i=1}^{\nu}C^2(x_i-\bar{X})^2 m_i}{n}=C^2\hat{D}X.$$

Тогда

$$\hat{\sigma}_{CX}=\sqrt{\hat{D}(CX)}=\sqrt{C^2\hat{D}X}=|C|\,\hat{\sigma}_X. \blacksquare$$

4^0. *Если все частоты вариантов умножить на одно и то же число, то выборочные дисперсия и среднее квадратическое отклонение не изменятся.*

□ Для случайной величины Z, у которой все варианты совпадают с соответствующими вариантами случайной величины X, а частотами являются частоты вариантов величины X, умноженные на число C, имеем

$$\hat{D}Z=\frac{\sum_{i=1}^{\nu}(z_i-\bar{Z})^2 Cm_i}{\sum_{i=1}^{\nu}Cm_i}=\frac{C\sum_{i=1}^{\nu}(x_i-\bar{X})^2 m_i}{C\sum_{i=1}^{\nu}m_i}=\hat{D}X.$$

Тогда

$$\hat{\sigma}_Z=\sqrt{\hat{D}Z}=\sqrt{\hat{D}X}=\hat{\sigma}_X. \blacksquare$$

5^0. *Выборочная дисперсия равна разности между средним арифметическим квадратов наблюдений над случайной величиной X и квадратом ее среднего арифметического, т. е.*

$$\hat{D}X=\overline{X^2}-(\bar{X})^2. \tag{8.15}$$

□ По определению $\bar{X}$ для случайной величины X^2 имеем

$$\overline{X^2}=\sum_{i=1}^{\nu}(x_i^2\cdot m_i)\Big/\sum_{i=1}^{\nu}m_i.$$

Тогда

$$\hat{D}X=\frac{\sum_{i=1}^{\nu}(x_i-\bar{X})^2 m_i}{\sum_{i=1}^{\nu}m_i}=\frac{\sum_{i=1}^{\nu}x_i^2 m_i-2\bar{X}\sum_{i=1}^{\nu}x_i m_i+(\bar{X})^2\sum_{i=1}^{\nu}m_i}{\sum_{i=1}^{\nu}m_i}=$$

$$=\frac{\sum_{i=1}^{\nu}x_i^2 m_i}{\sum_{i=1}^{\nu}m_i}-2\bar{X}\frac{\sum_{i=1}^{\nu}x_i m_i}{\sum_{i=1}^{\nu}m_i}+(\bar{X})^2=\overline{X^2}-2(\bar{X})^2+(\bar{X})^2=\overline{X^2}-(\bar{X})^2. \blacksquare$$

В этом параграфе понятие выборочной дисперсии было введено по аналогии с понятием дисперсии дискретной случайной величины. Сравним эти дисперсии. Для этого еще раз запишем формулы, их определяющие:

$$\sigma_X^2=\sum_{i=1}^{n}(x_i-MX)^2 p_i;\quad \hat{\sigma}_X^2=\sum_{i=1}^{\nu}(x_i-\bar{X})^2\hat{p}_i.$$

Различие этих формул состоит в том, что во вторую формулу вместо математического ожидания MX входит среднее арифметическое, а на месте вероятности p_i значения случайной величины, равного x_i,— частость $\hat{p}_i$ (статистическая вероятность) наблюдавшегося варианта x_i. Несмотря на это различие, между дисперсиями много общего. Во-первых, обе они являются мерой рассеивания. Во-вторых, кроме того, что формулы внешне похожи, соответствующие дисперсии обладают схожими свойствами. В-третьих, при возрастании объема выборки и выполнении некоторых нежестких условий частость $\hat{p}_i$ стремится по вероятности к вероятности p_i, а среднее арифметическое $\bar{X}$ — к математическому ожиданию MX. Все это позволяет вполне обоснованно называть число $\hat{\sigma}_X^2$ *выборочной дисперсией* и брать его в качестве приближенного значения дисперсии σ_X^2.

§ 8.7. Выборочные начальные и центральные моменты. Асимметрия. Эксцесс

Приведем краткий обзор характеристик, которые наряду с уже рассмотренными применяются для анализа вариационных рядов и являются аналогами соответствующих числовых характеристик случайной величины.

Среднее арифметическое и выборочная дисперсия являются частным случаем более общего понятия — момента вариационного ряда.

Определение. *Начальным выборочным моментом порядка k называется среднее арифметическое k-х степеней наблюдаемых значений случайной величины и обозначаются $\hat{\nu}_k$, т. е.*

$$\hat{\nu}_k=\frac{\sum_{i=1}^{l} x_i^k m_i}{\sum_{i=1}^{l} m_i}, \tag{8.16}$$

где $\sum_{i=1}^{l} m_i=n$.

Из определения следует, что начальный выборочный момент нулевого порядка

$$\hat{\nu}_0=\sum_{i=1}^{l} x_i^0 m_i \Big/ \sum_{i=1}^{l} m_i=1,$$

а начальный выборочный момент первого порядка

$$\hat{\nu}_1=\sum_{i=1}^{l} x_i m_i \Big/ \sum_{i=1}^{l} m_i=\bar{X}.$$

Определение. *Центральным выборочным моментом порядка k называется среднее арифметическое k-х степеней отклонений наблюдаемых значений случайной величины от их среднего арифметического и обозначается $\hat{\mu}_k$, т. е.*

$$\hat{\mu}_k=\frac{\sum_{i=1}^{l} (x_i-\bar{X})^k m_i}{\sum_{i=1}^{l} m_i}. \tag{8.17}$$

Из определения следует, что центральный выборочный момент нулевого порядка

$$\hat{\mu}_0=\sum_{i=1}^{l} (x_i-\bar{X})^0 m_i \Big/ \sum_{i=1}^{l} m_i=1,$$

центральный выборочный момент первого порядка

$$\hat{\mu}_1=\sum_{i=1}^{l} (x_i-\bar{X}) m_i \Big/ \sum_{i=1}^{l} m_i=0$$

(в силу свойства 5^0 среднего арифметического), а центральный выборочный момент второго порядка

$$\hat{\mu}_2=\sum_{i=1}^{l} (x_i-X)^2 m_i \Big/ \sum_{i=1}^{l} m_i=\hat{\sigma}_X^2.$$

Можно доказать следующие формулы, выражающие центральные выборочные моменты различных порядков через начальные:

$$\begin{gathered}\hat{\mu}_2=\hat{\nu}_2-\nu_1^2;\\ \hat{\mu}_3=\hat{\nu}_3-3\hat{\nu}_1\hat{\nu}_2+2\hat{\nu}_1^3;\\ \hat{\mu}_4=\hat{\nu}_4-4\hat{\nu}_1\hat{\nu}_3+6\hat{\nu}_1^2\hat{\nu}_2-3\hat{\nu}_1^4 \text{ и т. д.}\end{gathered} \tag{8.18}$$

Доказательства следующих свойств моментов аналогичны доказательствам соответствующих свойств дисперсии и среднего арифметического и не представляют трудностей.

1^0. *Если все наблюдаемые значения случайной величины увеличить (уменьшить) на одно и то же число, то центральный выборочный момент k-го порядка не изменится.*

2^0. *Если все наблюдаемые значения случайной величины умножить на одно и то же число C, то начальный и центральный выборочные моменты k-го порядка изменятся в C^k раз.*

Если воспользоваться указанными свойствами, то вычисление выборочных моментов можно значительно упростить (что и показано в следующем параграфе).

Определение. *Выборочным коэффициентом асимметрии называется число $\hat{A}$, определяемое формулой*

$$\hat{A}=\frac{\hat{\mu}_3}{\hat{\sigma}_X^3}=\frac{\sum_{i=1}^{\nu} (x_i-\bar{X})^3 m_i}{\hat{\sigma}_X^3 \sum_{i=1}^{\nu} m_i}. \tag{8.19}$$

Выборочный коэффициент асимметрии служит для характеристики асимметрии полигона вариационного ряда. Если полигон асимметричен, то одна из ветвей его начиная с вершины имеет более пологий «спуск», чем другая.

В случае отрицательного коэффициента асимметрии более пологий «спуск» полигона наблюдается слева, в противном случае — справа. В первом случае асимметрию называют *левосторонней*, а во втором — *правосторонней*.

Определение. *Выборочным эксцессом или коэффициентом крутости называется число $\hat{E}$, определяемое формулой*

$$\hat{E}=\frac{\hat{\mu}_4}{\hat{\sigma}_X^4}-3. \qquad (8.20)$$

Выборочный эксцесс служит для сравнения на «крутость» выборочного распределения с нормальным распределением. В гл. 6 подчеркивалось, что эксцесс для случайной величины, распределенной нормально, равен нулю. Поэтому за стандартное значение выборочного эксцесса принимают $\hat{E}=0$. Если выборочному распределению соответствует отрицательный эксцесс, то соответствующий полигон имеет более пологую вершину по сравнению с нормальной кривой. В случае положительного эксцесса полигон более крутой по сравнению с нормальной кривой.

§ 8.8. Упрощенный способ вычисления статистических характеристик вариационных рядов

Вычисление среднего арифметического, дисперсии и выборочных моментов по приведенным выше формулам приводит к громоздким вычислениям, если числовые значения вариантов и соответствующие им частоты велики. Использование свойств указанных характеристик позволяет эти вычисления значительно упростить.

Если все частоты вариантов имеют общее кратное q, то частоты следует разделить на q, т. е. умножить на $1/q$. Для вновь полученного ряда значения упомянутых характеристик не изменятся. После проведенной операции (если в этом была необходимость) целесообразно все значения вариантов преобразовать по формуле

$$\dot{x}_i=\frac{x_i-c}{h}. \qquad (8.21)$$

Для полученной совокупности преобразованных вариантов вычисляют среднее арифметическое $\bar{X}_*$ и центральные моменты $\hat{\mu}_{k*}$. Истинные значения $\bar{X}$ и $\hat{\mu}_k$ находятся затем по следующим формулам:

$$\bar{X}=\bar{X}_*h+c; \qquad (8.22)$$

$$\hat{\mu}_k=h^k\hat{\mu}_{k*}. \qquad (8.23)$$

Последние два равенства предлагается доказать самостоятельно.

Таблица 8.6

Диаметр валика, мм	Середина интервала x_i	Частота m_i	x_i-c	$\dot{x}_i=\frac{x_i-c}{h}$	$\dot{x}_i m_i$	$\dot{x}_i^2 m_i$	$\dot{x}_i^3 m_i$	$\dot{x}_i^4 m_i$
1	2	3	4	5	6	7	8	9
6,67—6,69	6,68	2	−0,08	−4	−8	32	−128	512
6,69—6,71	6,70	15	−0,06	−3	−45	135	−405	1215
6,71—6,73	6,72	17	−0,04	−2	−34	68	−136	272
6,73—6,75	6,74	44	−0,02	−1	−44	44	−44	44
6,75—6,77	6,76	52	0,00	0	0	0	0	0
6,77—6,79	6,78	44	0,02	1	44	44	44	44
6,79—6,81	6,80	14	0,04	2	28	56	112	224
6,81—6,83	6,82	11	0,06	3	33	99	297	291
6,83—6,85	6,84	1	0,08	4	4	16	64	256
$\sum$		200			−22	494	−196	2858

Постоянные h и c выбирают произвольно. Однако их следует подбирать так, чтобы было можно максимально упростить вычисления. Обычно в качестве c выбирают вариант, который имеет наибольшую частоту или занимает среднее положение в ряду данных, при этом стремятся, чтобы разности x_i-c были возможно «проще». В качестве h можно взять наибольший делитель разностей x_i-c или такое число, которое позволило бы избавиться от дробей.

○ **Пример 8.13.** Используя данные табл. 8.2, найдем $\bar{X}$, $\hat{\sigma}_X^2$, $\hat{A}$, $\hat{E}$, преобразуя прежде всего данный интервальный ряд в дискретный. Для этого найдем середину каждого интервала и заполним столбец 2 табл. 8.6, где будем располагать промежуточные результаты вычислений.

Из табл. 8.6 видно, что наибольшую частоту имеет 5-й интервал (этому серединному интервалу соответствует частота 52). В качестве c возьмем середину этого интервала. Таким образом, $c=6{,}76$. Заполним теперь 4-й столбец таблицы. Данные, записанные в этом столбце, целесообразно разделить на 0,02 (умножить на 50), что позволяет получить целые числа. Поэтому полагаем $h=0{,}02$. Отметим, что h равно длине интервала. Так как числа c и h выбраны, то, выполняя дальнейшие вычисления, заполняем остальные столбцы таблицы.

По определению начальных моментов имеем

$$\hat{\nu}_{1*}=\sum\dot{x}_i m_i/\sum m_i=-22/200=-0{,}11;\quad \hat{\nu}_{2*}=\sum\dot{x}_i^2 m_i/\sum m_i=494/200=2{,}47;$$

$$\hat{\nu}_{3*}=\sum\dot{x}_i^3 m_i/\sum m_i=-196/200=-0{,}98;\quad \hat{\nu}_{4*}=\sum\dot{x}_i^4 m_i/\sum m_i=2858/200=14{,}29.$$

Используя формулы (8.18), получим следующие значения центральных моментов:

$$\hat{\mu}_{2*}=\hat{\nu}_{2*}-\hat{\nu}_{1*}^2=2{,}47-0{,}0121=2{,}4579;$$

$$\hat{\mu}_{3*}=\hat{\nu}_{3*}-3\hat{\nu}_{1*}\hat{\nu}_{2*}+2\hat{\nu}_{1*}^3=-0{,}98+0{,}8151-0{,}002662=-0{,}167562;$$

$$\hat{\mu}_{4*}=\hat{\nu}_{4*}-4\hat{\nu}_{1*}\hat{\nu}_{3*}+6\hat{\nu}_{1*}^2\hat{\nu}_{2*}-3\hat{\nu}_{1*}^4=14{,}29-0{,}4312+0{,}179322-0{,}00043923=14{,}037683.$$

На основании формулы (8.23) получаем следующие значения центральных моментов для первоначального вариационного ряда:

$$\hat{\mu}_2=2{,}4579\cdot(0{,}02)^2=0{,}00098316\approx0{,}001;$$

$$\hat{\mu}_3=-0{,}167562\cdot(0{,}02)^3=-1{,}340496\cdot10^{-6}\approx2{,}246\cdot10^{-6}.$$

$$\hat{\mu}_4=14{,}037683\cdot(0{,}02)^4=2{,}2460292\cdot10^{-6}\approx2{,}246\cdot10^{-6}.$$

Так как $\bar{X}_*=\hat{\nu}_{1*}$, то $\bar{X}_*=-0{,}11$. Тогда по формуле (8.22) получаем $\bar{X}=-0{,}11\cdot0{,}02+6{,}76=6{,}7578$ (мм).

Так как $\hat{\sigma}_X^2=\hat{\mu}_2$, то $\hat{\sigma}_X^2=0{,}00098316\approx0{,}001$ (мм^2).

Среднее квадратическое отклонение

$$\hat{\sigma}_X=\sqrt{0{,}00098316}\approx0{,}0314 \text{ (мм)}.$$

На основании формул (8.19) и (8.20) получаем значение коэффициента асимметрии и эксцесса:

$$\hat{A}=\hat{\mu}_3/\hat{\sigma}_X^3\approx-0{,}0435;\quad \hat{E}=\hat{\mu}_4/\hat{\sigma}_X^4-3\approx-0{,}676. \ \bullet$$

ГЛАВА 9
СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ЧИСЛОВЫХ ХАРАКТЕРИСТИК СЛУЧАЙНОЙ ВЕЛИЧИНЫ И ЗАКОНА РАСПРЕДЕЛЕНИЯ

§ 9.1. Понятие о точечной оценке числовой характеристики случайной величины; свойства точечной оценки

В гл. 8 были рассмотрены различные выборочные характеристики случайной величины X; среднее арифметическое $\bar{X}$, выборочная дисперсия $\hat{D}X$ и др. Эти характеристики используются 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.

Определение. *Выборочная характеристика, используемая в качестве приближенного значения неизвестной генеральной характеристики, называется ее точечной статистической оценкой.*

Среднее арифметическое $\bar{X}$ — это точечная статистическая оценка математического ожидания MX; $\hat{D}X$ — оценка дисперсии DX.

«Точечная» означает, что оценка представляет собой число или точку на числовой оси. «Статистическая» означает, что оценка рассчитывается по результатам наблюдений, или, иначе, по собранной исследователем статистике. Далее слово «статистическая» будем опускать.

Обозначим через Θ («тэта») некоторую генеральную характеристику (ею может быть и MX, и любая другая числовая характеристика случайной величины X). 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

$$\hat{\Theta}_{(n)}=f(X_1, X_2, \ldots, X_n). \tag{9.1}$$

Подставив в (9.1) вместо $X_1, X_2, \ldots, X_n$ конкретные результаты наблюдений (конкретные числа), получим число, 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, хорошо или нет найденное приближение, рассмотрим оценку $\hat{\Theta}_{(n)}$ с других позиций.

Пусть в формуле (9.1) $X_1, X_2, \ldots, X_n$ — не конкретные числа, а лишь обозначения тех результатов наблюдений, которые мы могли бы получить. Но результат каждого отдельного наблюдения случайной величины случаен, т. е. $X_1, X_2, \ldots, X_n$ — это 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. Это свойства состоятельности, несмещенности и эффективности.

Состоятельность. *Оценка $\hat{\Theta}_{(n)}$ генеральной характеристики Θ называется состоятельной, если для любого $\varepsilon>0$ выполняется равенство*

$$\lim_{n\to\infty} P(|\hat{\Theta}_{(n)}-\Theta|<\varepsilon)=1. \tag{9.2}$$

Поясним смысл равенства (9.2). Пусть ε — очень малое положительное число. Тогда равенство (9.2) 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.

Несмещенность. *Оценка $\hat{\Theta}_{(n)}$ генеральной характеристики Θ называется несмещенной, если для любого фиксированного числа наблюдений n выполняется равенство*

$$M\hat{\Theta}_{(n)}=\Theta, \tag{9.3}$$

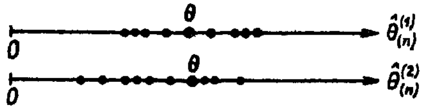


Рис. 9.1

т. е. математическое ожидание оценки равно неизвестной характеристике.

Поясним смысл равенства (9.3) 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 (9.3) означает: если оценка несмещенная, то при любом фиксированном n среднее из значений оценки, вычисленных 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:

$$\hat{\Theta}^{(1)}_{(n)}=f_1(X_1, X_2, ..., X_n) \text{ и } \hat{\Theta}^{(2)}_{(n)}=f_2(X_1, X_2, ..., X_n).$$

Значения этих оценок, вычисленные по выборкам объема n, изображены точками на рис. 9.1. 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.

Задача 9.1. Пусть генеральную совокупность образуют пять чисел: -2, -1, 0, 6, 2. Вычислить генеральное среднее MX и генеральную дисперсию DX. Составить всевозможные выборки с возвратом объема $n=2$; для каждой из них вычислить значения средней $\bar{X}_{(n=2)}$ и дисперсии $\hat{D}X_{n=2}$. Установить, выполняется ли при $n=2$ равенство

$$MX_{(n)}=MX. \tag{9.4}$$

Является ли выборочная дисперсия несмещенной оценкой генеральной дисперсии?

△ Генеральной совокупности -2, -1, 0, 6, 2 соответствует случайная величина X с рядом распределения

x	-2	-1	0	2	6	
P	1/5	1/5	1/5	1/5	1/5	$\sum=1$

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 всевозможные выборки с возвратом объема $n=2$ из данной генеральной совокупности. Они приведены в столбцах 1 и 2 табл. 9.1.

Например, $X_1=-2$ и $X_2=-2$ означает, что первое число, попавшее в выборку, равно -2 и второе число, попавшее в выборку, также равно -2. Вероятность появления такой выборки равна вероятности произведения события $X_1=-2$ и события $X_2=-2$:

$$P[(X_1=-2)(X_2=-2)]=P(X_1=-2)P_{X_1=-2}(X_2=-2)=\frac{1}{5}\cdot\frac{1}{5}=\frac{1}{25}.$$

При нахождении вероятности учитывалось, что для выборки с возвратом условная вероятность

$$P_{X_1=-2}(X_2=-2)=P(X_2=-2)=1/5.$$

Таблица 9.1

Выборка		Вероятность появления выборки	$\bar{X}=\frac{X_1+X_2}{2}$	$\overline{X^2}=\frac{X_1^2+X_2^2}{2}$	$\hat{D}X=\overline{X^2}-(\bar{X})^2$	$s^2=\frac{n\hat{D}X}{n-1}$ $(n=2)$
X_1	X_2					
1	2	3	4	5	6	7*)
-2	-2	1/25	-2	4	0	0
-2 -1	-1 -2	1/25 +1/25=2/25	$-3/2$	5/2	1/4	1/2
-2 0	0 -2	2/25	-1	2	1	2
-2 6	6 -2	2/25	2	20	16	32
-2 2	2 -2	2/25	0	4	4	8
-1	-1	1/25	-1	1	0	0
-1 0	0 -1	2/25	$-1/2$	1/2	1/4	1/2
-1 6	6 -1	2/25	5/2	37/2	49/4	49/2
-1 2	2 -1	2/25	1/2	5/2	9/4	9/2
0	0	1/25	0	0	0	0
0 6	6 0	2/25	3	18	9	18
0 2	2 0	2/25	1	2	1	2
6	6	1/25	6	36	0	0
6 2	2 6	2/25	4	20	4	8
2	2	1/25	2	4	0	0
		$\sum=1$				

*) Содержание столбца 7 выясняется в задаче 9.2.

Нетрудно убедиться в том, что вероятность появления любой другой из перечисленных в табл. 9.1 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 чисел -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 табл. 9.1.

Вычислим для каждой выборки среднее $\bar{X}_{(n=2)}$ и дисперсию $\hat{D}X_{(n=2)}$ (см. столбцы 4—6 табл. 9.1). Используя столбцы 4 и 3 табл. 9.1, построим ряд распределения случайной величины $\bar{X}_{(n=2)}$ (этот ряд приведен в табл. 9.2) и найдем математическое ожидание $M\bar{X}_{(n=2)}$.

Таблица 9.2

Значение средней $\bar{X}_{(n=2)}$	-2	$-3/2$	-1	$-1/2$	0	1/2	1	2	5/2	3	4	6	
Вероятность P	1/25	2/25	3/25	2/25	3/25	2/25	2/25	3/25	2/25	2/25	2/25	1/25	$\sum=1$

Математическое ожидание

$$M\bar{X}_{(n=2)}=-2\cdot 1/25+(-3/2)2/25+\ldots+6\cdot 1/25=1=MX.$$

Таким образом, при $n=2$ равенство (9.4) выполняется.

Аналогично можно убедиться в том, что равенство (9.4) выполняется и для выборок с возвратом объема $n=3$, и для выборок с возвратом объема $n=4$ и т. д. Выполнение равенства (9.4) при любом n означает, что $\bar{X}_{(n)}$— несмещенная оценка математического ожидания MX. Строгое доказательство этого утверждения приводится в § 9.2.

Обратим внимание на следующее. Если рассматривать выборки без возврата, то и в этом случае при любом n будет выполняться равенство (9.4). Так, при $n=2$ число выборок без возврата равно 20 — это все выборки, приведенные в табл. 9.1, кроме тех, где $X_1=X_2$; вероятность появления каждой из них равна 1/20 (например, $P[(X_1=-2)(X_2=-1)]=P(X_1=-2)\times \times P_{X_1=-2}(X_2=-1)=\frac{1}{5}\cdot\frac{1}{4}=\frac{1}{20}\Big)$. Если же не акцентировать внимание на порядке следования чисел в выборке, то число выборок без возврата объема $n=2$ равно 10 и вероятность появления каждой из них равна 2/20. Ряд распределения средней $\bar{X}_{(n=2)}$ для таких выборок приведен в табл. 9.3.

Таблица 9.3

Значение средней $\bar{X}_{(n=2)}$	$-3/2$	-1	$-1/2$	0	1/2	1	2	5/2	3	4	
Вероятность P	2/20	2/20	2/20	2/20	2/20	2/20	2/20	2/20	2/20	2/20	$\sum=1$

Математическое ожидание

$$M\bar{X}_{(n=2)}=-3/2\cdot 2/20+(-1)\cdot 2/20+\ldots+4\cdot 2/20=1=MX.$$

Таким образом, и для выборок с возвратом равенство (9.4) выполняется.

Чтобы установить, является ли выборочная дисперсия $\hat{D}X_{(n)}$ несмещенной оценкой генеральной дисперсии DX, обратимся к равенству (9.3), 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 табл. 9.1, построим ряд распределения случайной величины $\hat{D}X_{(n)}$. Этот ряд приведен в табл. 9.4.

Таблица 9.4

Значение дисперсии $\hat{D}X_{(n=2)}$	0	1/4	1	9/4	4	49/4	9	16	
Вероятность P	5/25	4/25	4/25	2/25	4/25	2/25	2/25	2/25	$\sum=1$

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 при $n=2$ $M(\hat{D}X_{(n)})\neq DX$ (напомним, $DX=8$), то $\hat{D}X$ не обладает свойством несмещенности, т. е. $\hat{D}X$ — смещенная оценка дисперсии DX. ▲

Эффективность. *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 $\hat{\Theta}^{(1)}_{(n)}$ и $\hat{\Theta}^{(2)}_{(n)}$ одной и той же характеристики Θ

$$\hat{\Theta}^{(1)}_{(n)}=\varphi_1(X_1, X_2, \ldots, X_n) \text{ и } \hat{\Theta}^{(2)}_{(n)}=\varphi_2(X_1, X_2, \ldots, X_n),$$

причем обе оценки несмещенные, т. е. при любом n

$$M\hat{\Theta}^{(1)}_{(n)}=\Theta \text{ и } M\hat{\Theta}^{(2)}_{(n)}=\Theta.$$

Какая же из оценок «лучше»? Оценка — это 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 рис. 9.2 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.

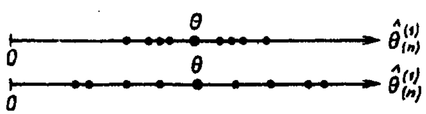


Рис. 9.2

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,

нижняя граница для дисперсий различных несмещенных оценок математического ожидания равна σ^2/n, (9.5) нижняя граница для дисперсий различных несмещенных оценок дисперсии равна $2\sigma^4/n$. (9.6)

Для случайной величины X с законом распределения

x	1	0
P	p	$1-p$

(X — это число появлений события в одном испытании при вероятности его появления в этом испытании равной p)

нижняя граница для дисперсий различных несмещенных оценок вероятности p равна $p(1-p)/n$. (9.7)

В выражениях (9.5)—(9.7) n — число независимых, проводимых в одинаковых условиях наблюдений случайной величины X. Напомним, что с вероятностной точки зрения:

независимость наблюдений величины X означает, что результаты $X_1, X_2, ..., X_n$ наблюдений — независимые случайные величины; (9.8)

одинаковость условий наблюдений означает, что каждая из величин $X_1, X_2, ..., X_n$ имеет такой же закон распределения, как и исходная величина X. (9.9)

Обратим внимание на то, что из (9.9) следует равенство математических ожиданий

$$MX_1 = MX_2 = ... = MX_n = MX \qquad (9.10)$$

и равенство дисперсий

$$DX_1 = DX_2 = ... = DX_n = DX. \qquad (9.11)$$

Заметим, что если использовать термин выборки, то выполнение условий (9.8) и (9.9) означает, что $X_1, X_2, ..., X_n$ — случайная выборка с возвратом.

§ 9.2. Точечные оценки математического ожидания и дисперсии

В качестве приближенных значений математического ожидания MX и дисперсий DX в § 9.1 было предложено использовать выборочное среднее $\bar{X} = \sum_{i=1}^{n} X_i/n$ и выборочную дисперсию $\hat{D}X = \sum_{i=1}^{n} (X_i - \bar{X})^2/n$. Выясним, какими из трех свойств (состоятельностью, несмещенностью, эффективностью) обладают средняя $\bar{X}$ и дисперсия $\hat{D}X$.

Выборочное среднее как точечная оценка математического ожидания.

1°. *Состоятельность.*

Теорема. *Пусть результаты $X_1, X_2, ..., X_n$ наблюдений величины X независимы, т. е. выполняется условие* (9.8) *и $MX_1 = MX_2 = ... = MX_n = MX$, т. е. выполняется условие* (9.10), *а дисперсии DX_1, $DX_2, ..., DX_n$ конечны. Тогда среднее $\bar{X}_{(n)} = \sum_{i=1}^{n} X_i/n$ — состоятельная оценка математического ожидания MX.*

□ Если выполняется условие (9.8) и дисперсии DX_1, $DX_2, ..., DX_n$ конечны, то, согласно теореме Чебышева, для любого $\varepsilon > 0$ имеет место равенство

$$\lim_{n\to\infty} P\left(\left|\sum_{i=1}^{n} X_i/n - \sum_{i=1}^{n} MX_i/n\right| < \varepsilon\right) = 1, \qquad (9.12)$$

которое, учитывая (9.10), можно записать так:

$$\lim_{n\to\infty} P(|\bar{X}_{(n)} - MX| < \varepsilon) = 1. \qquad (9.13)$$

Сравнивая (9.13) с определением (9.2) свойства состоятельности ($\bar{X}_{(n)}$ — это $\hat{\Theta}_{(n)}$, MX — это Θ), заключаем, что выборочное среднее $\bar{X}_{(n)}$ является состоятельной оценкой математического ожидания MX. ■

Заметим, что для случайной выборки с возвратом условия (9.8) и (9.10) выполняются, поэтому среднее для этой выборки (при конечности дисперсии DX) является состоятельной оценкой математического ожидания.

2°. *Несмещенность.*

Теорема. *Пусть $MX_1 = MX_2 = ... = MX_n = MX$, т. е. выполняется условие* (9.10). *Тогда среднее $\bar{X}_{(n)} = \sum_{i=1}^{n} X_i/n$ — несмещенная оценка математического ожидания MX.*

□ Если выполняется условие (9.10), то при любом фиксированном n имеет место следующая цепочка равенств:

$$M\bar{X}_{(n)} = M\left(\sum_{i=1}^{n} X_i/n\right) = \frac{1}{n}\sum_{i=1}^{n} MX_i \underset{(9.10)}{=} \frac{1}{n}\sum_{i=1}^{n} MX = MX.$$

Таким образом, при выполнении условия (9.10) справедливо равенство $M\bar{X}_{(n)} = MX$. Сравнивая его с определением (9.3) свойства несмещенности ($\bar{X}_{(n)}$ — это $\hat{\Theta}_{(n)}$, MX — это Θ), заключаем, что $\bar{X}_{(n)}$ — несмещенная оценка математического ожидания MX. ■

Замечание. При доказательстве несмещенности среднего арифметического не использовалось условие независимости наблюдений. Поэтому при выполнении условия (9.10) $\bar{X}$ является несмещенной оценкой математического ожидания как при независимых наблюдениях (это наблюдения в выборке с возвратом), так и при зависимых наблюдениях (это наблюдения в выборке без возврата). Этот факт был подтвержден при решении задачи 9.1.

3°. *Эффективность.*

Напомним, что свойство эффективности рассматривалось только для таких оценок, которые являются несмещенными.

Как было доказано выше, выборочное среднее $\bar{X}=\sum_{i=1}^{n} X_i/n$ — несмещенная оценка математического ожидания MX, если выполняется условие (9.10). Допустим, что оно выполняется; приведем доказательство эффективности выборочного среднего для случая, когда величина X имеет нормальный закон распределения, т. е. $X=N(a, \sigma)$, где $a=MX$, $\sigma=\sqrt{DX}$.

Теорема. *Пусть случайная величина X имеет нормальный закон распределения и при этом результаты $X_1, X_2, \ldots, X_n$ ее наблюдений независимы, т. е. выполняется условие (9.8), а также выполняются условия (9.11): $DX_1=DX_2=\ldots=DX_n=DX$ и (9.10): $MX_1=MX_2=\ldots=MX_n=MX$. Тогда выборочное среднее $\bar{X}=\sum_{i=1}^{n} X_i/n$ — несмещенная эффективная оценка математического ожидания MX.*

□ Для того чтобы доказать эффективность среднего $\bar{X}$, надо доказать, что его дисперсия совпадает с минимальной границей, равной в случае нормального распределения σ^2/n (см. (9.5)). Найдем дисперсию $D\bar{X}$:

$$D\bar{X}=D\left(\sum_{i=1}^{n} X_i/n\right)=\frac{1}{n^2}D\left(\sum_{i=1}^{n} X_i\right)\underset{(9.8)}{=}\frac{1}{n^2}\sum_{i=1}^{n} DX_i=$$

$$\underset{(9.11)}{=}\frac{1}{n^2}\sum_{i=1}^{n} DX=\frac{nDX}{n^2}=\frac{\sigma^2}{n}.$$

Таким образом, в условиях теоремы средняя $\bar{X}$ — несмещенная эффективная оценка математического ожидания. ■

Точечные оценки генеральной дисперсии. Наряду с выборочной дисперсией $\hat{D}X=\sum_{i=1}^{n}(X_i-\bar{X})^2/n$ в качестве приближенного значения генеральной дисперсии DX используют величину

$$s^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1), \tag{9.14}$$

которая связана с $\hat{D}X$ соотношением

$$s^2=n\hat{D}X/(n-1). \tag{9.15}$$

Выясним, какими свойствами обладают $\hat{D}X$ и s^2, как точечные оценки дисперсии.

1°. *Состоятельность.*

Теорема. *Пусть результаты $X_1, X_2, \ldots, X_n$ наблюдений случайной величины X независимы и удовлетворяют условиям $MX_1=MX_2=\ldots=MX_n=MX$ и $DX_1=DX_2=\ldots=DX_n=DX$, а центральные моменты второго и четвертого порядков величины*

X конечны. Тогда при любом $\varepsilon>0$ 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

$$\lim_{n\to\infty} P(|s^2-DX|<\varepsilon)=1,$$

т. е. s^2 также состоятельная оценка генеральной дисперсии DX.

2°. *Несмещенность.* В задаче 9.1 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 $M(\hat{D}X)$ в общем случае; при этом будем предполагать, что *выполняются условия* (9.8), (9.10) и (9.11).

□ Предварительно обратим внимание на следующие два факта:

так как для любого $i=1, 2, \ldots, n$ дисперсия $DX_i=MX_i^2-(MX_i)^2$, то при выполнении условий (9.10) и (9.11) справедлива цепочка равеиств

$$MX_1^2=MX_2^2=\ldots=MX_n^2=MX^2; \tag{9.16}$$

при выполнении условия (9.8) справедливы соотношения

$$M(X_iX_j)=MX_iMX_j,\quad i,\ j=1, 2, \ldots, n,\quad i\neq j. \tag{9.17}$$

Найдем $M(\hat{D}X)$. Имеем

$$M(\hat{D}X)=M\left[\sum_{i=1}^{n}(X_i-\bar{X})^2/n\right]=M\left[\sum_{i=1}^{n}(X_i^2-2X_i\bar{X}+\bar{X}^2)/n\right]=$$

$$=M\left(\sum_{i=1}^{n}X_i^2/n\right)-M\left(\sum_{i=1}^{n}2X_i\bar{X}/n\right)+M\left(\sum_{i=1}^{n}\bar{X}^2/n\right)=$$

$$=\sum_{i=1}^{n}MX_i^2/n-2M\left(\bar{X}\sum_{i=1}^{n}X_i/n\right)+M(n\bar{X}^2/n)\underset{(9.10)}{=}$$

$$=\sum_{i=1}^{n}MX^2/n-2M(\bar{X}^2)+M(\bar{X}^2)=MX^2-M(\bar{X}^2),$$

но

$$M(\bar{X}^2)=M\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)^2=\frac{1}{n^2}M\left(\sum_{i=1}^{n}X_i^2+\sum_{\substack{i,j=1\\ i\neq j}}^{n}X_iX_j\right)=$$

$$=\sum_{i=1}^{n}MX_i^2/n^2+\sum_{\substack{i,j=1\\ i\neq j}}^{n}M(X_iX_j)/n^2\underset{\substack{(9.16)\\(9.17)}}{=}$$

$$=\sum_{i=1}^{n}MX^2/n^2+\sum_{\substack{i,j=1\\ i\neq j}}^{n}(MX_iMX_j)/n^2\underset{(9.10)}{=}$$

$$=nMX^2/n^2+\sum_{\substack{i,j=1\\ i\neq j}}^{n}(MX\,MX)/n^2=MX^2/n+n(n-1)(MX)^2/n^2=$$

$$=MX^2/n+(n-1)(MX)^2/n.$$

Поэтому

$$M(\hat{D}X)=MX^2-[MX^2/n+(n-1)M^2X/n]=$$

$$=\frac{n-1}{n}MX^2-\frac{n-1}{n}M^2X=\frac{n-1}{n}(MX^2-M^2X)=\frac{n-1}{n}DX. \blacksquare$$

Итак, окончательно получаем

$$M(\hat{D}X)=\frac{n-1}{n}DX<DX, \tag{9.18}$$

т. е. ни при каком объеме выборки n математическое ожидание $M(\hat{D}X)$ не равно DX. Это означает, что оценка $\hat{D}X$ — смещенная оценка дисперсии DX, при этом смещение равно

$$|M(\hat{D}X)-DX|=\left|\frac{n-1}{n}DX-DX\right|=\frac{DX}{n}.$$

Несмещенной оценкой дисперсии DX является оценка s^2, определяемая по формуле (9.15).

Теорема. *Пусть результаты $X_1, X_2, \dots, X_n$ наблюдений величины X независимы, т. е. выполняется условие* (9.8), *а* $MX_1=MX_2=\dots=MX_n=MX$ *и* $DX_1=DX_2=\dots=DX_n=DX$, *т. е. выполняются условия* (9.10) *и* (9.11). *Тогда* $s^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)$ — *несмещенная оценка генеральной дисперсии* DX.

□ Находим Ms^2:

$$Ms^2 \underset{(9.15)}{=} M\left(\frac{n\hat{D}X}{n-1}\right)=\frac{n}{n-1}M(\hat{D}X) \underset{(9.18)}{=} \frac{n}{n-1}\frac{n-1}{n}DX=DX.$$

Итак, при любом n

$$Ms^2=DX. \tag{9.19}$$

Сравнивая равенство (9.19) с определением (9.3) свойства несмещенности (s^2 — это $\Theta_{(n)}$, DX — это Θ), заключаем, что s^2 — несмещенная оценка дисперсии DX. ■

Именно несмещенностью оценки s^2 объясняется более частое, по сравнению с $\hat{D}X$, ее использование в качестве приближенного значения генеральной дисперсии DX.

Замечание. Доказательство соотношения (9.18), а поэтому и (9.19) проводилось для независимых наблюдений (это наблюдения в выборке с возвратом). Если наблюдения зависимы (выборка без возврата), то s^2 не обладает свойством несмещенности.

Задача 9.2. В условиях задачи (9.1) для всевозможных выборок с возвратом объема $n=2$ найти значение величины s^2, построить ее ряд распределения и вычислить Ms^2. Аналогичные значения найти для выборок без возврата объема $n=2$; убедиться в том, что в этом случае оценка s^2 — смещенная.

△ Используя столбцы 7 и 3 табл. 9.1, построим ряд распределения случайной величины s^2, значения которой были найдены для всевозможных выборок с возвратом объема $n=2$. Он приведен в табл. 9.5.

Таблица 9.5

Значение дисперсии s^2	0	1/2	2	9/2	8	18	49/2	32	
Вероятность P	5/25	4/25	4/25	2/25	4/25	2/25	2/25	2/25	$\Sigma=1$

Найдем

$$Ms^2=0\cdot 5/25+1/2\cdot 4/25+\dots+32\cdot 2/25=8=DX$$

(напомним, что в задаче 9.1 генеральная дисперсия $DX=8$).

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 $Ms^2=DX$. Таким образом, если выборка с возвратом (наблюдения независимы), то s^2 — несмещенная оценка дисперсии DX.

Теперь из табл. 9.1 возьмем выборки без возврата (это выборки, для которых $X_1\neq X_2$). Ряд распределения случайной величины s^2 в этом случае задан табл. 9.6.

Таблица 9.6

Значение дисперсии s^2	1/2	2	9/2	8	18	49/2	32	
Вероятность P	4/20	4/20	2/20	4/20	2/20	2/20	2/20	$\Sigma=1$

Далее получаем

$$Ms^2=1/2\cdot 4/20+2\cdot 4/20+\dots+32\cdot 2/20=10\neq DX.$$

Таким образом, если выборка без возврата (наблюдения зависимы), то оценка s^2 — смещенная. ▲

Еще раз обратим внимание на соотношения (9.18) и (9.19). Из (9.18) следует, что при любом фиксированном n среднее значений оценки

$$\hat{D}X=\sum_{i=1}^{n}(X_i-\bar{X})^2/n,$$

вычисленных для всевозможных выборок с возвратом объема n, т. е. $M(\hat{D}X)$, всегда меньше генеральной дисперсии DX. Если же использовать не $\hat{D}X$, а

$$s^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1),$$

то при любом фиксированном n среднее значений величины s^2, вычисленных для всевозможных выборок с возвратом объема n, т. е. Ms^2, совпадает с генеральной дисперсией. Это вытекает из равенства (9.19).

3°. *Эффективность.* Напомним, что свойство эффективности рассматривалось только для таких оценок, которые являются несмещенными. Будет ли несмещенная оценка s^2 дисперсии обладать свойством эффективности?

Пусть случайная величина X имеет нормальный закон распределения: $X=N(a,\ \sigma)$, где $a=MX$, а $\sigma=\sqrt{DX}$.

Для того чтобы убедиться в эффективности оценки s^2, надо доказать, что ее дисперсия совпадает с минимальной границей, равной в случае нормального распределения $2\sigma^4/n$ (см. (9.6)). Найдем дисперсию величины s^2.

□ Будем предлагать, что выполняются условие (9.8), т. е. результаты $X_1, X_2, \ldots, X_n$ наблюдений величины X — независимые случайные величины, и условие (9.9), которое в случае нормального распределения величины X имеет вид

$$X_1=N(a,\ \sigma),\ X_2=N(a,\ \sigma), \ldots, X_n=N(a,\ \sigma). \tag{9.20}$$

Соотношения (9.20) означают, что результат каждого наблюдения имеет такой же закон распределения, что и закон распределения величины X, а именно нормальный закон с параметрами a и σ.

При выполнении условий (9.8) и (9.20) выполняется соотношение (9.64), из которого получаем

$$s^2=\chi^2(k=n-1)\,\sigma^2/n, \tag{9.21}$$

где $\chi^2(k=n-1)$ — случайная величина, имеющая χ^2-распределение с числом степеней свободы $k=n-1$.

Поэтому

$$Ds^2=D\left[\frac{\chi^2(k=n-1)\,\sigma^2}{n}\right] \underset{\substack{\sigma^2=\mathrm{const}\\ n=\mathrm{const}}}{=} \frac{\sigma^4}{n^2}D\left[\chi^2(k=n-1)\right] \underset{(6.34)}{=} \frac{\sigma^4 2(n-1)}{n^2}\neq\frac{2\sigma^4}{n}. \ \blacksquare$$

Так как Ds^2 не совпадает с нижней границей, то s^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.

§ 9.3. Частость как точечная оценка вероятности события

Обозначим через p неизвестную вероятность появления случайного события A в единичном испытании. Найдем приближенное значение $\hat{p}$ вероятности p. Проведем n независимых испытаний, в каждом из которых событие A может произойти 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 произошло событие A. Тогда $\hat{p}=m/n$ — это опытная вероятность или частость появления события A. Выясним, какими свойствами обладает $\hat{p}$, как точечная оценка вероятности p.

Теорема. *Пусть m — число наступлений события A в n независимых испытаниях, p — вероятность наступления события A в каждом из испытаний. Тогда $\hat{p}=m/n$ — состоятельная, несмещенная и эффективная оценка вероятности p.*

□ 1°. *Состоятельность.* Для испытаний Бернулли (именно о таких испытаниях идет речь в условиях теоремы) справедлива теорема Бернулли, согласно которой для любого $\varepsilon>0$ выполняется равенство

$$\lim_{n\to\infty}P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1 \tag{9.22}$$

Сравнивая (9.22) с определением (9.2), свойства состоятельности (m/n — это $\Theta_{(n)}$, p — это Θ), заключаем, что $\hat{p}=m/n$ — состоятельная оценка вероятности p.

2°. *Несмещенность.* Зафиксируем число n испытаний Бернулли. Поскольку A — случайное событие, количество m испытаний, в которых произойдет событие A, — величина случайная, поэтому частость $\hat{p}=m/n$ также случайная величина. Найдем математическое ожидание частости, при этом будем иметь в виду, что в условиях испытаний Бернулли величина m имеет биномиальный закон распределения. Следовательно, математическое ожидание $Mm=np$, а дисперсия $Dm=npq$. Имеем

$$M\hat{p}=M\left(\frac{m}{n}\right) \underset{n=\mathrm{const}}{=} \frac{1}{n}M(m)=\frac{1}{n}np=p.$$

Таким образом, $M\hat{p}=p$, т. е. при любом фиксированном числе n испытаний Бернулли математическое ожидание частости $\hat{p}$ равно неизвестной вероятности p. Это означает, что $\hat{p}=m/n$ — несмещенная оценка вероятности p.

3°. *Эффективность.* Сопоставим с результатом единичного испытания случайную величину

$$X=\begin{cases}1, & \text{если событие } A \text{ произойдет в испытании,}\\ 0, & \text{если событие } A \text{ не произойдет.}\end{cases}$$

Ряд распределения этой величины имеет вид

x	1	0
P	p	$1-p$

Тогда последовательность n испытаний Бернулли — это n независимых наблюдений случайной величины X, проводимых в одинаковых условиях. А согласно (9.7), минимум среди дисперсий различных несмещенных оценок вероятности p равен $p(1-p)/n$.

Теперь найдем дисперсию частости:

$$D\hat{p}=D\left(\frac{m}{n}\right)\underset{n=\text{const}}{=\!=}\frac{1}{n^2}Dm=\frac{1}{n^2}npq=\frac{pq}{n}=\frac{p(1-p)}{n}.$$

Так как $D\hat{p}$ совпадает с минимальной границей, то частость $\hat{p}$, будучи несмещенной оценкой, является также и эффективной оценкой вероятности p. ■

§ 9.4. Методы получения точечных оценок

Метод максимального правдоподобия. Мы рассмотрели различные случаи использования функций от результатов наблюдений в качестве точечных оценок. Однако осталось неясным, каким образом, например, установлено, что именно комбинацию $\bar{X}=\sum\limits_{i=1}^{n} X_i/n$ лучше всего использовать в качестве оценки математического ожидания. Рассмотрим метод, позволяющий получить ответ на этот вопрос.

Допустим, что известна модель закона распределения случайной величины, однако неизвестно значение входящего в эту модель параметра. Найдем его оценку, если известны результаты независимых проводимых в одинаковых условиях наблюдений случайной величины. Используем *метод максимального правдоподобия*. Рассмотрим этот метод на двух примерах: в первом примере случайная величина дискретна, во втором — непрерывна.

○ **Пример 9.1.** Случайная величина X распределена по закону Пуассона. Это означает, что вероятность

$$P(X=x)=\frac{\lambda^x}{x!}e^{-\lambda},\quad x=0,\ 1,\ 2,\ \dots\ . \tag{9.23}$$

Параметр этого закона — λ, его значение неизвестно.

Как и раньше, обозначим через $X_1,\ X_2,\ \dots,\ X_n$ возможные результаты независимых наблюдений случайной величины X, проводимых в одинаковых условиях. Напомним, что эти предположения означают следующее:

$X_1,\ X_2,\ \dots,\ X_n$ — независимые случайные величины; (9.24)

Закон распределения любой из величин $X_1,\ X_2,\ \dots,$ X_n совпадает с законом распределения величины X, т. е.

$$\begin{aligned} &P(X_1=x_1)=P(X=x_1),\ x_1=0,\ 1,\ 2,\ \dots,\\ &P(X_2=x_2)=P(X=x_2),\ x_2=0,\ 1,\ 2,\ \dots,\\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\ &P(X_n=x_n)=P(X=x_n),\ x_n=0,\ 1,\ 2,\ \dots. \end{aligned} \tag{9.25}$$

Введем функцию $L(X_1,\ X_2,\ \dots,\ X_n)$, которую назовем *функцией правдоподобия*. Ее значение

$$L(x_1,\ x_2,\ \dots,\ x_n)=P[(X_1=x_1)(X_2=x_2)\ \dots\ (X_n=x_n)]$$

— это вероятность произведения событий $X_1=x_1$, $X_2=x_2$, ..., $X_n=x_n$, или, иначе, совместная вероятность появления чисел $x_1,\ x_2,\ \dots,\ x_n$. Чем больше значение $L(x_1,\ x_2,\ \dots,\ x_n)$, тем правдоподобнее или более вероятно появление в результате наблюдений чисел $x_1,\ x_2,\ \dots,\ x_n$. Отсюда и название функции $L(X_1,\ X_2,\ \dots,\ X_n)$ — функция правдоподобия результатов наблюдений.

Учитывая (9.24) и (9.25), получаем, что совместная вероятность появления чисел $x_1,\ x_2,\ \dots,\ x_n$

$$P[(X_1=x_1)(X_2=x_2)\dots(X_n=x_n)]\underset{(9.24)}{=}$$

$$=P(X_1=x_1)P(X_2=x_2)\dots P(X_n=x_n)\underset{(9.25)}{=}$$

$$=P(X=x_1)P(X=x_2)\dots P(X=x_n).$$

Таким образом, функция правдоподобия для независимых, проводимых в одинаковых условиях наблюдений, имеет вид

$$L(X_1,\ X_2,\ \dots,\ X_n)=P(X=X_1)P(X=X_2)\dots P(X=X_n). \tag{9.26}$$

Подставляя в (9.26) значения вероятностей, вычисленные по формуле (9.23), получаем

$$L(X_1,\ X_2,\ \dots\ X_n)=\frac{\lambda^{X_1}}{X_1!}e^{-\lambda}\,\frac{\lambda^{X_2}}{X_2!}e^{-\lambda}\ \dots\ \frac{\lambda^{X_n}}{X_n!}e^{-\lambda}=\frac{\lambda^{\sum\limits_{i=1}^{n}X_i}e^{-n\lambda}}{X_1!\,X_2!\dots X_n!}.$$

Обратим внимание на то, что функция L зависит не только от $X_1,\ X_2,\ \dots,\ X_n$, но и от неизвестного параметра λ. Поэтому в число аргументов функции L введем λ. Получаем

$$L(X_1,\ X_2,\ \dots\ X_n;\ \lambda)=\frac{\lambda^{\sum\limits_{i=1}^{n}X_i}e^{-n\lambda}}{X_1!\,X_2!\dots X_n!}. \tag{9.27}$$

За оценку неизвестного параметра λ примем такое число $\hat{\lambda}$, при котором функция правдоподобия $L(X_1,\ X_2,\ \dots,\ X_n;\ \lambda)$, рассматриваемая как функция от λ, достигает максимума, т. е. $\hat{\lambda}$ определяется из условия

$$L(X_1,\ X_2,\ \dots,\ X_n;\ \hat{\lambda})=\max_{\lambda} L(X_1,\ X_2,\ \dots\ X_n;\ \lambda). \tag{9.28}$$

Естественность такого подхода к определению оценки вытекает из смысла функции правдоподобия. Действительно, $L(X_1,\ X_2,\ \dots,\ X_n;\ \lambda)$ при каждом зафиксированном значении λ является мерой правдоподобия (вероятностью) получения результатов наблюдений $X_1,\ X_2,\ \dots,\ X_n$. Поэтому, изменяя значения λ, можно проследить, при каких значениях появление результатов $X_1,\ X_2,\ \dots,\ X_n$

более правдоподобно (более вероятно), и выбрать такое значение λ, при котором результаты X_1, X_2, ..., X_n будут наиболее правдоподобными.

Метод нахождения оценки неизвестного параметра, основанный на требовании максимизации функции правдоподобия, называется *методом максимального правдоподобия*, а найденная этим методом оценка — *оценкой максимального правдоподобия*.

Функции L и $\ln L$, рассматриваемые как функции параметра λ, достигают максимума при одном и том же значении λ, так как $\ln L$ — монотонно возрастающая функция. Поэтому вместо отыскания максимума функции L находят (что удобнее) максимум функции $\ln L$. Функция $\ln L$ называется *логарифмической функцией правдоподобия*.

Для (9.27) логарифмическая функция правдоподобия

$$\ln L(X_1, X_2, ..., X_n; \lambda) = \ln \frac{\lambda^{\sum_{i=1}^{n} X_i} e^{-n\lambda}}{X_1!\, X_2! \dots X_n!} =$$

$$= \left(\sum_{i=1}^{n} X_i\right) \ln \lambda - n\lambda - \ln(X_1!) - \ln(X_2!) - \dots - \ln(X_n!).$$

Найдем точку максимума этой функции, рассматривая ее как функцию параметра λ. Для этого:

найдем производную функции $\ln L$ по λ:

$$\frac{\partial \ln L}{\partial \lambda} = \frac{\sum_{i=1}^{n} X_i}{\lambda} - n;$$

приравняв производную нулю, определим критическую точку — корень полученного уравнения — *уравнения правдоподобия*:

$$\frac{\sum_{i=1}^{n} X_i}{\lambda} - n = 0 \rightarrow \lambda_{кр} = \sum_{i=1}^{n} X_i / n;$$

найдем вторую производную функции $\ln L$ и ее значение в точке $\lambda_{кр}$:

$$\frac{\partial^2 \ln L}{\partial \lambda^2} = -\frac{\sum_{i=1}^{n} X_i}{\lambda^2}; \quad \frac{\partial^2 \ln L(\lambda_{кр})}{\partial \lambda^2} = -\frac{\sum_{i=1}^{n} X_i}{\lambda_{кр}^2} = -\frac{n^2}{\sum_{i=1}^{n} X_i};$$

Докажем, что $\lambda_{кр}$ — это точка максимума.

Согласно (9.25), $X_i \geqslant 0$, $i = 1, 2, ..., n$, поэтому $\sum_{i=1}^{n} X_i \geqslant 0$.

Если $\sum_{i=1}^{n} X_i > 0$, то $\frac{\partial^2 \ln L(\lambda_{кр})}{\partial \lambda^2} < 0$ и, следовательно, $\lambda_{кр}$ — точка максимума функции $\ln L$.

Если же $\sum_{i=1}^{n} X_i = 0$ (а это возможно, когда все $X_i = 0$), то из (9.27) получаем $L(0, 0, ..., 0; \lambda) = e^{-n\lambda}$; а так как $L_{max} = 1$ (L — это вероятность), то максимальное значение функции $L(\lambda) = e^{-n\lambda}$ достигается при $\lambda = 0$, т. е. при $\lambda = \sum_{i=1}^{n} X_i / n = \sum_{i=1}^{n} 0/n$.

Итак, всегда $\lambda_{кр} = \sum_{i=1}^{n} X_i / n$ — это точка максимума функции $\ln L$ (или L), поэтому она и является оценкой $\hat{\lambda}_{мп}$ максимального правдоподобия для неизвестного параметра λ, т. е.

$$\hat{\lambda}_{мп} = \sum_{i=1}^{n} X_i / n = \bar{X}. \tag{9.29}$$

Напомним, что в формуле (9.29) X_1, X_2, ..., X_n — результаты независимых, проведенных в одинаковых условиях наблюдений пуассоновской случайной величины. Если они сгруппированы в вариационный ряд с числом групп, равным ν, и m_i — частота появления варианта x_i, то

$$\hat{\lambda}_{мп} = \sum_{i=1}^{n} x_i m_i / n. \tag{9.30}$$

Формула (9.30) тождественна формуле (9.29). ●

○ **Пример 9.2.** Случайная величина X распределена по показательному закону, т. е. ее функция плотности

$$f_X(x) = \mu e^{-\mu x}, \quad x > 0. \tag{9.31}$$

Параметр этого закона — μ, его значение неизвестно. Найдем оценку этого параметра методом максимального правдоподобия.

Если наблюдения над величиной X независимы и проводятся в одинаковых условиях, то функция правдоподобия имеет вид, аналогичный (9.26), с той лишь разницей, что вместо вероятностей $P(X = X_1)$, $P(X = X_2)$, ..., $P(X = X_n)$ фигурируют значения функции плотности $f_X(X_1)$, $f_X(X_2)$, ..., $f_X(X_n)$; это различие объясняется тем, что в примере 9.1 рассматривалась дискретная случайная величина, а здесь — непрерывная. Итак, для непрерывной величины функция правдоподобия

$$L(X_1, X_2, ..., X_n) = f_X(X_1) f_X(X_2) \dots f_X(X_n). \tag{9.32}$$

Подставив в (9.32) значения функции плотности, вычисленные по формуле (9.31), и учитывая, что параметр μ неизвестен, получим

$$L(X_1, X_2, ..., X_n; \mu) = \mu e^{-\mu X_1} \mu e^{-\mu X_2} \dots \mu e^{-\mu X_n} = \mu^n e^{-\mu \sum_{i=1}^{n} X_i},$$

где $X_1, X_2, ..., X_n$ — результаты независимых, проведенных в одинаковых условиях наблюдений случайной величины с показательным законом распределения. Перейдем к логарифмической функции правдоподобия

$$\ln L(X_1, X_2, ..., X_n; \mu) = \ln\left(\mu^n e^{-\mu\sum_{i=1}^{n} X_i}\right) = n\ln\mu - \mu\sum_{i=1}^{n} X_i$$

и найдем точку максимума этой функции, рассматривая ее как функцию параметра μ. Для этого:

найдем производную функции $\ln L$ по μ:

$$\frac{\partial \ln L}{\partial \mu} = \frac{n}{\mu} - \sum_{i=1}^{n} X_i;$$

приравняв производную нулю, определим критическую точку — корень полученного уравнения правдоподобия:

$$\frac{n}{\mu} - \sum_{i=1}^{n} X_i = 0 \to \mu_{кр} = n / \sum_{i=1}^{n} X_i;$$

найдем вторую производную функции $\ln L$ и ее значение в точке $\mu_{кр}$:

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\mu^2}; \quad \frac{\partial^2 \ln L(\mu_{кр})}{\partial \mu^2} = -\frac{n}{\mu_{кр}^2} = -\frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}.$$

Так как X_i — это значение случайной величины с показательным законом распределения (см. (9.31)), то $X_i > 0$, $i = 1, 2, ..., n$, поэтому $\sum_{i=1}^{n} X_i > 0$ и вторая производная в точке $\mu_{кр}$ отрицательна; следовательно, $\mu_{кр}$ — точка максимума функции $\ln L$ и ее надо принять в качестве оценки максимального правдоподобия параметра μ, т. е.

$$\hat{\mu}_{мп} = n / \sum_{i=1}^{n} X_i = 1/\bar{X}. \ \bullet$$

Оценки, получаемые методом максимального правдоподобия, обычно, особенно при большом числе наблюдений n, обладают свойствами «хороших» оценок, а именно свойствами состоятельности, несмещенности и эффективности.

Метод наименьших квадратов как частный случай метода максимального правдоподобия. Пусть случайная величина X имеет нормальный закон распределения с параметрами $a = MX$ и $\sigma = \sqrt{DX}$, т. е. функция плотности

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/(2\sigma^2)}, \tag{9.33}$$

причем числовое значение параметра a неизвестно, а значение параметра σ известно. Для нахождения оценки параметра a воспользуемся методом максимального правдоподобия. Функция правдоподобия

$$L(X_1, X_2, ..., X_n; a) = f_X(X_1) f_X(X_2) \dots f_X(X_n) =$$

$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-(X_1-a)^2/(2\sigma^2)} \frac{1}{\sigma\sqrt{2\pi}} e^{-(X_2-a)^2/(2\sigma^2)} \dots \frac{1}{\sigma\sqrt{2\pi}} e^{-(X_n-a)^2/(2\sigma^2)} =$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n e^{-\sum_{i=1}^{n}(X_i-a)^2/(2\sigma^2)},$$

где $X_1, X_2, ..., X_n$ — результаты независимых, проведенных в одинаковых условиях наблюдений нормально распределенной случайной величины. Логарифмическая функция правдоподобия

$$\ln L = n\ln\frac{1}{\sqrt{2\pi}} + n\ln\frac{1}{\sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - a)^2. \tag{9.34}$$

Найдем точку максимума функции (9.34), рассматривая ее как функцию параметра a. Однако для этого не будем использовать первую и вторую производные функции $\ln L$, а поступим следующим образом. Из (9.34) видно, что функция $\ln L$, рассматриваемая как функция от a, достигает максимального значения в том и только том случае, когда сумма $\sum_{i=1}^{n}(X_i - a)^2$ достигает минимального значения. Таким образом, в случае нормального распределения оценку $\hat{a}_{мп}$ максимального правдоподобия можно определить как точку минимума функции:

$$F(a) = \sum_{i=1}^{n}(X_i - a)^2, \tag{9.35}$$

т. е. из условия

$$F(a) = \sum_{i=1}^{n}(X_i - a)^2 \to \min. \tag{9.36}$$

Метод нахождения оценки параметра a в соответствии с требованием (9.36) называется *методом наименьших квадратов*.

Найдем точку минимума функции (9.35). Первая производная

$$\frac{dF(a)}{da} = \left[\sum_{i=1}^{n}(X_i - a)^2\right]'_a = \sum_{i=1}^{n}\left[(X_i - a)^2\right]'_a = \sum_{i=1}^{n} 2(X_i - a)(-1) = -2\sum_{i=1}^{n}(X_i - a).$$

Приравняв производную нулю, найдем критическую точку:

$$-2\sum_{i=1}^{n}(X_i - a) = 0 \to \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} a \to \sum_{i=1}^{n} X_i = na \to a_{кр} = \sum_{i=1}^{n} X_i / n.$$

Так как вторая производная

$$\frac{d^2F(a)}{da^2}=\left[\frac{dF(a)}{da}\right]'_a=\left[-2\sum_{i=1}^{n}(X_i-a)\right]'_a=-2\sum_{i=1}^{n}(X_i-a)'_a=-2\sum_{i=1}^{n}(-1)=2n>0$$

больше нуля при любом значении a, в том числе и при $a=a_{\text{кр}}$, то $a_{\text{кр}}$ — точка минимума функции (9.35) и ее надо принять в качестве оценки $\hat{a}_{\text{мп}}$ максимального правдоподобия параметра a.

Итак,

$$\hat{a}_{\text{мп}}=\sum_{i=1}^{n} X_i/n.$$

В заключение обратим внимание на следующее.

Во-первых, оценки, получаемые методом наименьших квадратов и методом максимального правдоподобия, не всегда совпадают.

Во-вторых, при оценивании параметра a нормального распределения (9.33) предполагалось, что значение дисперсии σ^2 известно. Если неизвестны ни a, ни σ^2, то функцию (9.34) следует рассматривать как функцию от a и σ^2 и для нахождения оценок $\hat{a}_{\text{мп}}$ и $\hat{\sigma}^2_{\text{мп}}$ максимального правдоподобия следует решить следующую *систему уравнений правдоподобия*:

$$\begin{cases}\dfrac{\partial \ln L(X_1, X_2, \dots, X_n; a, \sigma^2)}{\partial a}=0,\\[2ex] \dfrac{\partial \ln L(X_1, X_2, \dots, X_n; a, \sigma^2)}{\partial(\sigma^2)}=0.\end{cases}$$

§ 9.5. Параметрическое оценивание закона распределения

Результаты предварительной обработки наблюдений случайной величины, дополненные сведениями о сущности изучаемого явления, зачастую оказываются достаточными для того, чтобы сформулировать гипотезу о модели закона распределения изучаемой случайной величины, нормальный ли этот закон, биномиальный или какой-либо другой. Используя наблюдения, можно найти оценки параметров предполагаемой модели, т. е. оценки входящих в модель числовых характеристик (см. § 9.4). Подставив в модель вместо параметров найденные оценки, получим оценку предполагаемой модели закона распределения, которая называется *параметрической*. Оценивание закона распределения, не требующее предварительного выбора его модели и оценивания входящих в нее параметров, называется *непараметрическим*. Примерами непараметрических оценок неизвестного закона распределения являются вариационный ряд, выборочная функция распределения и выборочная плотность распределения, которые рассматривались в § 8.3.

Рассмотрим два примера параметрического оценивания закона распределения: в первом примере случайная величина дискретна, во втором — непрерывна.

○ **Пример 9.3.** Дано следующее распределение успеваемости 100 студентов-заочников, сдававших четыре экзамена:

Число сданных экзаменов	0	1	2	3	4
Число студентов	1	1	3	35	60

Здесь случайной величиной является число сданных экзаменов среди четырех. Обозначим ее X. Установим закон распределения этой величины.

Построим сначала его непараметрическую оценку. Величина X — дискретная. Дискретный вариационный ряд, заданный столбцами 2 и 4 табл. 9.7, дает непараметрическую оценку закона распределения числа сданных экзаменов среди четырех сдаваемых.

Теперь попробуем сформулировать гипотезу о модели закона распределения случайной величины X — числе сданных экзаменов среди четырех сдаваемых. Процесс сдачи четырех экзаменов представим как четыре испытания, относительно которых сделаем следующие допущения:

— эти испытания независимы, т. е. вероятность сдачи любым студентом любого экзамена не зависит от того, будет сдано или нет любое количество других экзаменов;

*Таблица 9.7**)*

i	Число сданных экзаменов x_i	Число студентов m_i	Частость $\hat{p}_i=\frac{m_i}{n}$	$p_i^{\text{теор}}=C_4^{x_i}\times 0{,}88^{x_i}\cdot 0{,}12^{4-x_i}$	$m_i^{\text{теор}}=np_i^{\text{теор}}$	$(m_i-m_i^{\text{теор}})^2$	$(m_i-m_i^{\text{теор}})^2:m_i^{\text{теор}}$
1	2	3	4	5	6	7	8
1	0	1	0,01	0,00021	0,021		
2	1	1 } 5	0,01	0,00608	0,608 } 7,32	5,382	0,735
3	2	3	0,03	0,06691	6,691		
4	3	35	0,35	0,32711	32,711	5,239	0,160
5	4	60	0,60	0,59969	59,969	0,001	0,000
Итого		$n=100$	1,00	1,00000			0,895

*) Содержание столбцов 5 и 6 выясняется далее в этом же примере. Содержание столбцов 7 и 8 выясняется в задаче 10.11.

— вероятность сдачи студентом любого отдельно взятого экзамена одна и та же и равна p, а вероятность «несдачи» равна $(1-p)$.

Конечно, эти допущения могут вызвать некоторые сомнения, но возможно, что они не будут противоречить результатам наблюдений. При этих допущениях мы имеем дело с испытаниями Бернулли и число сданных экзаменов среди четырех сдаваемых будет иметь биномиальный закон распределения, т. е. вероятность того, что студент сдаст x экзаменов, равна

$$P(X=x)=C_4^x p^x (1-p)^{4-x}, \quad x=0,\ 1,\ 2,\ 3,\ 4. \tag{9.37}$$

Найдем оценку параметра p, входящего в модель (9.37). Вспомним, что в условиях испытаний Бернулли состоятельной несмещенной и эффективной оценкой вероятности является частость. В рассматриваемом примере p — вероятность того, что студент сдаст экзамент, поэтому частость $\hat{p}$ этого события, учитывая, что имеются сведения об успеваемости 100 студентов, вычисляем следующим образом:

$$\hat{p}=\frac{\text{число экзаменов, сданных 100 студентами}}{\text{число экзаментов, сдаваемых 100 студентами}}=\frac{\sum\limits_{i=1}^{5} x_i m_i}{4\cdot 100}=$$

$$=\frac{0\cdot 1+1\cdot 1+2\cdot 3+3\cdot 35+4\cdot 60}{100\cdot 4}=0{,}88.$$

Так как $\sum\limits_{i=1}^{5} x_i m_i/100=\bar{X}$ — это среднее число экзаменов, сданных одним студентом, то $\hat{p}$ можно было бы определить и так:

$$\hat{p}=\frac{\text{среднее число экзаменов, сданных одним студентом}}{\text{число экзаменов, сдаваемых одним студентом}}=\frac{\bar{X}}{4}=0{,}88.$$

Заметим, что если находить оценку параметра p в модели (9.37) методом максимального правдоподобия и при этом учесть, что число x_i наблюдалось m_i раз, то мы получили бы для $\hat{p}$ такую же формулу, а именно

$$\hat{p}_{\text{мп}}=\sum_{i=1}^{5} x_i m_i/(4n).$$

Подставив в модель (9.37) вместо параметра p его оценку $\hat{p}$, получим параметрическую оценку неизвестного закона распределения числа сданных экзаменов, построенную в предположении, что допустима биномиальная модель

$$P(X=x)=C_4^x 0{,}88^x 0{,}12^{4-x}; \quad x=0,\ 1,\ 2,\ 3,\ 4. \tag{9.38}$$

Теоретические вероятности $p_i^{\text{теор}}$ и частоты $m_i^{\text{теор}}$, вычисленные в предположении, что имеет место модель (9.38), содержатся в столбцах 5 и 6 табл. 9.7. Поскольку различия между соответствующими числами столбцов 4 и 5 или между числами столбцов 3 и 6 небольшие, можно сделать предварительное заключение о приемлемости биномиальной модели. Графически это заключение подтверждается рис. 9.3, на котором кривая вероятностей $p_i^{\text{теор}}$ близка к кривой частостей $\hat{p}_i$. ●

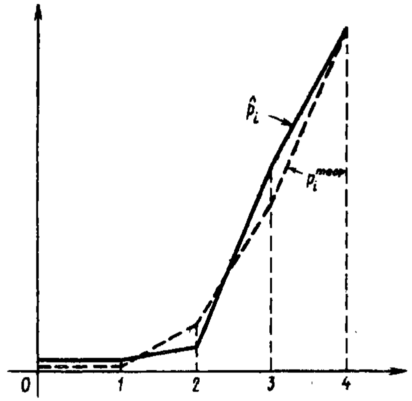


Рис. 9.3

Метод более глубокого обоснования приемлемости той или иной модели называется *критерием согласия* (он рассматривается в § 10.8).

○ **Пример 9.4.** По данным примера 8.12 найдем параметрические оценки функции плотности и функции распределения случайной величины X — размера диаметра валика, предполагая, что она имеет нормальный закон т. е. $X=N(a, \sigma)$.

Неизвестные математическое ожидание a и дисперсию σ^2 нормального закона заменим их точечными оценками: a заменим средней $\bar{X}$, ее значение было найдено в примере 8.10, $\bar{X}=6{,}7578$; σ^2 заменим оценкой, равной s^2. В примере 8.12 была вычислена выборочная дисперсия $\hat{D}X=0{,}9832\cdot 10^{-3}$; используя соотношение (9.15), получим $s^2=0{,}9881\cdot 10^{-3}$; $s=0{,}0314$.

Тогда параметрическая оценка неизвестной функции плотности размера диаметра валика, построенная в предположении, что допустима нормальная модель, имеет вид

$$f_N(x)=\frac{1}{0{,}0314\sqrt{2\pi}}\,e^{-\frac{(x-6{,}7578)^2}{2\cdot 0{,}0314^2}}, \tag{9.39}$$

а параметрическая оценка функции распределения такова:

$$F_N(x)=\frac{1}{0{,}0314\sqrt{2\pi}}\int\limits_{-\infty}^{x} e^{-\frac{(t-6{,}7578)^2}{2\cdot 0{,}0314^2}}\,dt. \tag{9.40}$$

Напомним алгоритм вычисления значения функции (9.39) в точке x с помощью функции $f_U(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$, затабулированной в табл. П. 1. Алгоритм следующий:

$$x \to u = \frac{x-6{,}7578}{0{,}0314} \xrightarrow{\text{П.1}} f_U(u) \to f_N(x) = \frac{f_U(u)}{0{,}0314}.$$

Например, для $x = 6{,}68$ получим

$$x = 6{,}68 \to u = \frac{6{,}68-6{,}7578}{0{,}0314} = -2{,}48 \xrightarrow{\text{П.1}} f_U(-2{,}48) =$$

$$= 0{,}0184 \to f_N(6{,}68) = \frac{0{,}0184}{0{,}0314} = 0{,}58599.$$

Значения функции (9.39), вычисленные для середин интервалов вариационного ряда, заданного табл. 8.5, приведены в строке 6 табл. 9.8. В этой же таблице в строке 3 приведены $\hat{f}(x_i)$ — значения выборочной функции плотности в точках x_i, взятые из табл. 8.5. Ряд чисел $\hat{f}(x_i)$ дает непараметрическую оценку неизвестной функции плотности.

Графики функции $\hat{f}$ и f_N изображены на рис. 9.4, *а*.

Теперь напомним алгоритм вычисления значения функции (9.40) в точке x с помощью функции Лапласа

$$\Phi(u) = \frac{2}{\sqrt{2\pi}} \int_0^u e^{-t^2/2}\, dt,$$

затабулированной в табл. П.2. Алгоритм имеет вид

$$x \to u = \frac{x-6{,}7578}{0{,}0314} \xrightarrow{\text{П.2}} \Phi(u) \to F_N(x) = 0{,}5 + \frac{\Phi(u)}{2}.$$

Например, для $x = 6{,}67$ получаем

$$x = 6{,}67 \to u = \frac{6{,}67-6{,}7578}{0{,}0314} = -2{,}8 \xrightarrow{\text{П.2}} \Phi(-2{,}80) =$$

$$= -0{,}9948 \to F_N(6{,}67) = \frac{1}{2} + \frac{-0{,}9948}{2} = 0{,}0026.$$

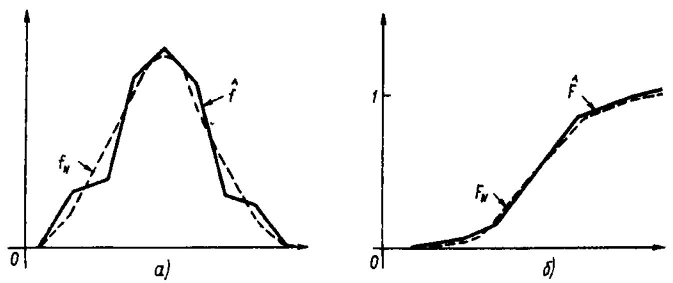


Рис. 9.4

Таблица 9.8

Номер интервала i	1	1	2	3	4	5	6	7	8	9
Середина интервала x_i	2	6,68	6,70	6,72	6,74	6,76	6,78	6,80	6,82	6,84
$\hat{f}(x_i)$	3	0,5	3,75	4,25	11,0	13,0	11,0	3,5	2,75	0,25
$u_i = (x_i - \bar{X})/s$	4	−2,48	−1,84	−1,20	−0,57	0,07	0,71	1,34	1,98	2,62
$f_U(u_i)$ (см. табл. П.1)	5	0,0184	0,0734	0,1942	0,3391	0,3980	0,3101	0,1626	0,0562	0,0129
$f_N(x_i) = f_U(u_i)/s$	6	0,59	2,34	6,18	10,80	12,68	9,88	5,18	1,79	0,41

Таблица 9.9

i	Граница интервала a_i	$\hat{F}(a_i)$	$u_i=\frac{a_i-\bar{X}}{s}$	$\Phi(u_i)$	$F_N(a_i)=\frac{1}{2}+\frac{\Phi(u_i)}{2}$
1	2	3	4	5	6
1	6,67	0,000	−2,80	−0,9948	0,0026
2	6,69	0,010	−2,16	−0,9692	0,0154
3	6,71	0,085	−1,52	−0,8714	0,0643
4	6,73	0,170	−0,88	−0,6212	0,1894
5	6,75	0,390	−0,25	−0,1974	0,4013
6	6,77	0,650	0,39	0,3034	0,6517
7	6,79	0,870	1,03	0,6970	0,8485
8	6,81	0,940	1,66	0,9030	0,9515
9	6,83	0,995	2,30	0,9784	0,9892
10	6,85	1,000	2,94	0,9966	0,9983

Значения функции (9.40), вычисленные для концов интервалов вариационного ряда, заданного табл. 8.4, приведены в столбце 6 табл. 9.9. В этой же таблице в столбце 3 приведены $\hat{F}(a_i)$—значения выборочной функции распределения в точках a_i, взятые из табл. 8.4. Ряд чисел $\hat{F}(a_i)$ дает непараметрическую оценку неизвестной функции распределения.

Графики функций $\hat{F}$ и F_N изображены на рис. 9.4, *б*.

Ответ на вопрос, можно ли считать, что случайная величина (размер диаметра валика) имеет нормальный закон распределения, будет дан в примере 10.10.

§ 9.6. Понятие об интервальной оценке числовой характеристики случайной величины

Вычисляя на основании результатов наблюдений точечную оценку $\hat{\Theta}$ неизвестной числовой характеристики Θ, мы понимаем, что величина $\hat{\Theta}$ является лишь приближенным значением характеристики Θ. Если для большого числа наблюдений точность приближения бывает достаточной для практических выводов (в силу несмещенности, состоятельности и эффективности «хороших» оценок), то для выборок небольшого объема вопрос о точности оценок очень важен. В математической статистике он решается следующим образом. По сделанной выборке находится точечная оценка $\hat{\Theta}$ неизвестной характеристики Θ, затем задаются вероятностью γ и по определенным правилам находят такое число $\varepsilon>0$, чтобы выполнялось соотношение

$$P(\underbrace{\hat{\Theta}-\varepsilon}_{\Theta_1}<\Theta<\underbrace{\hat{\Theta}+\varepsilon}_{\Theta_2})=\gamma. \qquad (9.41)$$

Соотношению (9.41) тождественно соотношение

$$P(|\hat{\Theta}-\Theta|<\varepsilon)=\gamma, \qquad (9.42)$$

из которого видно, что абсолютная погрешность оценки $\hat{\Theta}$ не превосходит числа ε. Это высказывание верно с вероятностью, равной γ. Число ε называется *точностью оценки* $\hat{\Theta}$ (чем меньше ε, тем выше точность оценки); числа Θ_1 и Θ_2 называются *доверительными границами*, интервал (Θ_1, Θ_2)—*доверительным интервалом* или *интервальной оценкой* характеристики Θ, вероятность γ называется *доверительной вероятностью* или *надежностью* интервальной оценки.

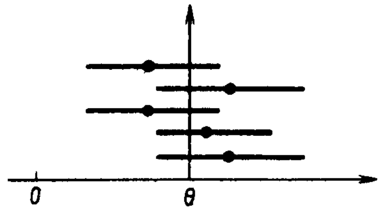


Рис. 9.5

В соотношении (9.41) случайными величинами являются доверительные границы Θ_1 и Θ_2: во-первых, эти границы могут изменяться при переходе от одной выборки к другой хотя бы потому, что при этом изменяется значение оценки $\hat{\Theta}$; во-вторых, при фиксированной выборке границы Θ_1 и Θ_2 изменяются при изменении вероятности γ, поскольку ε выбирается в зависимости от γ. Генеральная же характеристика Θ—постоянная величина. Поэтому соотношение (9.41) следует читать так: «вероятность того, что интервал (Θ_1, Θ_2) накроет характеристику Θ, равна γ»; именно «интервал накроет характеристику», а не «характеристика попадет в интервал». На рис. 9.5 друг над другом изображены доверительные интервалы характеристики Θ, построенные для разных выборок; центры интервалов—это выборочные значения оценки $\hat{\Theta}$.

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 является событие, состоящее в том, что погрешность оценки $\hat{\Theta}$ меньше ε, или, иначе, точность оценки $\hat{\Theta}$ больше ε.

В соотношении (9.41) границы Θ_1 и Θ_2 симметричны относительно точечной оценки $\hat{\Theta}$. 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.

§ 9.7. 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 одинаковых условиях, т. е. возможные результаты X_1, X_2 ..., X_n этих наблюдений обладают следующими свойствами:

X_1, X_2, ..., X_n — независимые случайные величины; (9.43)

закон распределения любой из величин X_1, X_2, ..., X_n совпадает с законом распределения величины X, т. е.

$$X_1=N(a, \sigma),\ X_2=N(a, \sigma),\ \ldots,\ X_n=N(a, \sigma). \qquad (9.44)$$

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 $\bar{X}=\sum_{i=1}^{n} X_i/n$ математического ожидания a. Зададимся вероятностью γ и попробуем найти такое число ε, чтобы выполнялось соотношение

$$P(\bar{X}-\varepsilon<a<\bar{X}+\varepsilon)=\gamma. \qquad (9.45)$$

Обратим внимание на то, что равенство (9.45) аналогично (9.41): параметр a — это Θ, а средняя $\bar{X}$ — это оценка $\hat{\Theta}$.

Нахождение ε основано на следующей теореме.

Теорема о распределении случайной величины $\frac{\bar{X}-a}{\sigma/\sqrt{n}}$. *Пусть X_1, X_2, ..., X_n — результаты n независимых наблюдений нормально распределенной случайной величины, проводимых в одинаковых условиях, т. е. X_1, X_2, ..., X_n обладают свойствами (9.43) и (9.44). Тогда величина $\frac{\bar{X}-a}{\sigma/\sqrt{n}}$ имеет нормальное распределение с нулевым математическим ожиданием и единичной дисперсией, т. е.*

$$\frac{\bar{X}-a}{\sigma/\sqrt{n}}=N(0, 1). \qquad (9.46)$$

□ В условиях теоремы результаты X_1, X_2, ..., X_n наблюдений обладают свойствами (9.43) и (9.44), поэтому их сумма $\sum_{i=1}^{n} X_i$ в силу теоремы о суммировании нормально распределенных величин имеет нормальный закон распределения, при этом математическое ожидание этой суммы

$$M\left(\sum_{i=1}^{n} X_i\right)=\sum_{i=1}^{n} MX_i\underset{(9.44)}{=}\sum_{i=1}^{n} a=na,$$

дисперсия

$$D\left(\sum_{i=1}^{n} X_i\right)\underset{(9.43)}{=}\sum_{i=1}^{n} DX_i\underset{(9.44)}{=}\sum_{i=1}^{n}\sigma^2=n\sigma^2,$$

т. е.

$$\sum_{i=1}^{n} X_i=N(na, \sqrt{n\sigma^2}).$$

Теперь найдем закон распределения величины $\bar{X}=\sum_{i=1}^{n} X_i/n$.

При делении $\sum_{i=1}^{n} X_i$ на постоянную величину n вид закона распределения не изменится, изменятся только его параметры. Поэтому средняя $\bar{X}$, так же как и сумма $\sum_{i=1}^{n} X_i$, имеет нормальный закон распределения, но с математическим ожиданием

$$M\bar{X}=M\left(\sum_{i=1}^{n} X_i/n\right)=\frac{1}{n}M\left(\sum_{i=1}^{n} X_i\right)=\frac{1}{n}na=a$$

и дисперсией

$$D\bar{X}=D\left(\sum_{i=1}^{n} X_i/n\right)=\frac{1}{n^2}D\left(\sum_{i=1}^{n} X_i\right)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n},$$

т. е.

$$\bar{X}=N(a,\ \sigma/\sqrt{n}). \qquad (9.47)$$

Следовательно, величина $\frac{\bar{X}-a}{\sigma/\sqrt{n}}$ будет иметь нормальный закон с нулевым математическим ожиданием и единичной дисперсией, т. е. выполняется соотношение (9.46), что и требовалось доказать. ■

Вернемся к нахождению числа ε в соотношении (9.45).

□ Воспользуемся табл. П.2 функции Лапласа и найдем для заданной вероятности γ число u_γ такое, при котором $\Phi(u_\gamma)=\gamma$, или, иначе, такое, при котором

$$P(|N(0, 1)|<u_\gamma)=\gamma.$$

Учитывая (9.46), получим

$$P\left(\left|\frac{\bar{X}-a}{\sigma/\sqrt{n}}\right|<u_\gamma\right)=\gamma. \qquad (9.48)$$

Выполним следующие тождественные преобразования:

$$\left|\frac{\bar{X}-a}{\sigma/\sqrt{n}}\right|<u_\gamma \to |\bar{X}-a|<\frac{u_\gamma\sigma}{\sqrt{n}} \to -\frac{u_\gamma\sigma}{\sqrt{n}}<\bar{X}-a<\frac{u_\gamma\sigma}{\sqrt{n}} \to -\bar{X}-\frac{u_\gamma\sigma}{\sqrt{n}}<-\bar{X}+\bar{X}-$$

$$-a<-\bar{X}+\frac{u_\gamma\sigma}{\sqrt{n}} \to \bar{X}-\frac{u_\gamma\sigma}{\sqrt{n}}<a<\bar{X}+\frac{u_\gamma\sigma}{\sqrt{n}}.$$

Таким образом, вероятность

$$P\left(\bar{X}-\frac{u_\gamma\sigma}{\sqrt{n}}<a<\bar{X}+\frac{u_\gamma\sigma}{\sqrt{n}}\right)=\gamma. \qquad (9.49)$$

Сравнивая (9.49) и (9.45), заключаем, что

$$\varepsilon = u_\gamma \sigma/\sqrt{n}. \blacksquare \tag{9.50}$$

Интервальная оценка математического ожидания такова:

$$\left(\bar{X} - u_\gamma \sigma/\sqrt{n},\ \ \bar{X} + u_\gamma \sigma/\sqrt{n}\right). \tag{9.51}$$

Полученный результат имеет следующий смысл: с вероятностью γ можно быть уверенным в том, что вычисленное по выборке среднее $\bar{X}$ дает значение математического ожидания с точностью (9.50), или, иначе, с вероятностью γ можно быть уверенным в том, что интервал (9.51) накроет неизвестное математическое ожидание.

В заключение приведем схему нахождения точности ε и доверительных границ, отвечающих надежности γ:

$$\gamma \xrightarrow{\text{П.2}} u_\gamma \to \varepsilon = \frac{u_\gamma \sigma}{\sqrt{n}} \begin{cases} \to \bar{X} - \varepsilon \text{ (нижняя граница)}, \\ \to \bar{X} + \varepsilon \text{ (верхняя граница)}. \end{cases} \tag{9.52}$$

Графическая иллюстрация схемы приведена на рис. 9.6.

Задача 9.3. Случайная величина распределена по нормальному закону с параметром $\sigma = 2$. Сделана случайная выборка с возвратом объема $n = 25$. Найти с надежностью $\gamma = 0{,}95$ точность выборочной средней и интервальную оценку для неизвестного математического ожидания a.

$\triangle$ По схеме (9.52) находим

$$\gamma = 0{,}95 \xrightarrow{\text{П.2}} u_\gamma = 1{,}96 \to \varepsilon = \frac{1{,}96 \cdot 2}{\sqrt{25}} = 0{,}784 \begin{cases} \to \bar{X} - 0{,}784 \\ \to \bar{X} + 0{,}784 \end{cases}.$$

Таким образом,

$$P(\bar{X} - 0{,}784 < a < \bar{X} + 0{,}784) = 0{,}95, \text{ или}$$
$$P(|\bar{X} - a| < 0{,}784) = 0{,}95.$$

Это означает, что с вероятностью 95% можно быть уверенным в том, что интервал $(\bar{X} - 0{,}784;\ \bar{X} + 0{,}784)$ накроет параметр a, или с вероятностью 95% быть уверенным в том, что вычисленное по выборке среднее $\bar{X}$ дает значение параметра a с точностью 0,784. $\blacktriangle$

Задача 9.4. Найти минимальный объем выборки из нормальной генеральной совокупности, при котором с надежностью, не меньшей $\gamma = 0{,}9$, погрешность средней, найденной по этой выборке, будет меньше $\varepsilon = 0{,}3$, если среднее квадратическое отклонение $\sigma = 2$.

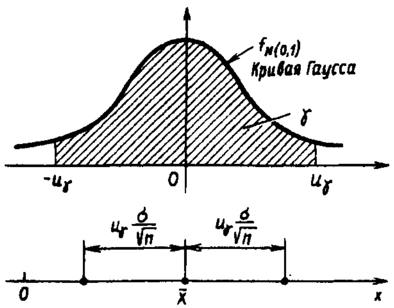


Рис. 9.6

$\triangle$ Обратим внимание на то, что при заданных ε и σ с ростом объема выборки n увеличивается вероятность $P(|\bar{X} - a| < \varepsilon) = \gamma$. В этом нетрудно убедиться, проанализировав формулу (9.50). Поэтому неравенство

$$P(|\bar{X} - a| < \varepsilon) \geqslant \gamma$$

выполняется для

$$n \geqslant u_\gamma^2 \sigma^2/\varepsilon^2, \tag{9.53}$$

при этом n должно быть целым числом.

В данном случае $\sigma = 2$, $\varepsilon = 0{,}3$, а u_γ — это число, при котором значение

функции Лапласа $\Phi(u_\gamma) = \gamma$. По табл. П.2 этой функции найдем, что при $\gamma = 0{,}9$ число $u_\gamma = 1{,}65$. Тогда $n \geqslant \frac{1{,}65^2 \cdot 4}{0{,}3^2} = 121$. Таким образом, минимальный объем выборки, при котором $P(|\bar{X} - a| < 0{,}3) \geqslant 0{,}9$, равен 121. $\blacktriangle$

Задача 9.5. На контрольных испытаниях $n = 15$ ламп была определена средняя продолжительность горения лампы, $\bar{X} = 3000$ ч. Считая, что срок службы лампы распределен нормально с $\sigma = 16$ ч, определить доверительную вероятность того, что точность средней равна 10 ч.

$\triangle$ По формуле (9.50) получаем

$$u_\gamma = \varepsilon \sqrt{n}/\sigma. \tag{9.54}$$

Подставляя числовые данные, находим $u_\gamma = 2{,}4$. Теперь, зная u_γ, по табл. П.2 функции Лапласа найдем вероятность $\Phi(2{,}4) = 0{,}9836$; это и есть доверительная вероятность γ. $\blacktriangle$

Интервальная оценка математического ожидания нормального распределения при неизвестной дисперсии. Выше была решена задача построения интервальной оценки для математического ожидания нормального распределения, когда его дисперсия известна. Теперь решим эту же задачу, но уже в условиях, когда дисперсия изучаемого нормального распределения неизвестна.

Итак, случайная величина $X = N(a, \sigma)$, причем неизвестны ни a, ни σ^2. По наблюдениям $X_1, X_2, \ldots, X_n$ вычислим среднее $\bar{X} = \sum\limits_{i=1}^{n} X_i/n$ и оценку $s^2 = \sum\limits_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$ дисперсии σ^2. Зададимся доверительной вероятностью γ и найдем такое число ε, чтобы выполнялось соотношение

$$P(\bar{X} - \varepsilon < a < \bar{X} + \varepsilon) = \gamma. \tag{9.55}$$

$\square$ Нахождение ε основано на следующей теореме, которую приведем без доказательства.

Теорема о распределении случайной величины $\frac{\bar{X} - a}{s/\sqrt{n}}$. *Пусть $X_1, X_2, \ldots, X_n$ — результаты n независимых наблюдений нормально распределенной случайной величины, проводимых в одинаковых условиях, т. е. $X_1, X_2, \ldots, X_n$ обладают свойствами* (9.43) *и* (9.44). *Тогда величина $\frac{\bar{X} - a}{s/\sqrt{n}}$ имеет распределение Стьюдента (или t-распределение) с $(n-1)$-й степенью свободы* (см. § 6.5), *или, иначе, величина $\frac{\bar{X} - a}{s/\sqrt{n}}$ является t-величиной с $(n-1)$-й степенью свободы:*

$$\frac{\bar{X} - a}{s/\sqrt{n}} = t(k = n - 1). \tag{9.56}$$

Воспользуемся табл. П.3 и найдем для заданных вероятности γ и числа $k=n-1$ число t_γ такое, при котором вероятность

$$P(|t(k)|<t_\gamma)=\gamma.$$

Учитывая (9.56), получаем

$$P\left(\left|\frac{\bar{X}-a}{s/\sqrt{n}}\right|<t_\gamma\right)=\gamma. \tag{9.57}$$

Выполним следующие тождественные преобразования:

$$\left|\frac{\bar{X}-a}{s/\sqrt{n}}\right|<t_\gamma \to |\bar{X}-a|<\frac{t_\gamma s}{\sqrt{n}} \to -\frac{t_\gamma s}{\sqrt{n}}<\bar{X}-a<\frac{t_\gamma s}{\sqrt{n}} \to -\bar{X}-\frac{t_\gamma s}{\sqrt{n}}<-\bar{X}+$$

$$+\bar{X}-a<-\bar{X}+\frac{t_\gamma s}{\sqrt{n}} \to \bar{X}-\frac{t_\gamma s}{\sqrt{n}}<a<\bar{X}+\frac{t_\gamma s}{\sqrt{n}}.$$

Таким образом,

$$P\left(\bar{X}-\frac{t_\gamma s}{\sqrt{n}}<a<\bar{X}+\frac{t_\gamma s}{\sqrt{n}}\right)=\gamma. \tag{9.58}$$

Сравнивая (9.58) и (9.55), заключаем, что

$$\varepsilon=t_\gamma s/\sqrt{n}. \;\blacksquare \tag{9.59}$$

Таким образом, с вероятностью γ можно утверждать, что среднее $\bar{X}$ дает значение неизвестного математического ожидания с точностью $\varepsilon=t_\gamma s/\sqrt{n}$, а интервальная оценка математического ожидания такова:

$$(\bar{X}-t_\gamma s/\sqrt{n},\ \ \bar{X}+t_\gamma s/\sqrt{n}). \tag{9.60}$$

В заключение приведем схему нахождения точности ε и доверительных границ, отвечающих надежности γ:

$$\left.\begin{matrix}\gamma \\ k=n-1\end{matrix}\right\} \xrightarrow{\text{П.3}} t_\gamma \to \varepsilon=\frac{t_\gamma s}{\sqrt{n}} \to \begin{cases}\bar{X}-\varepsilon\ (\text{нижняя граница}),\\ \bar{X}+\varepsilon\ (\text{верхняя граница}),\end{cases} \tag{9.61}$$

где

$$\bar{X}=\sum_{i=1}^{n} X_i/n,\quad s=\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)}.$$

Графическая иллюстрация схемы приведена на рис. 9.7.

Задача 9.6. На контрольных испытаниях 16 осветительных ламп были определены: средняя продолжительность работы лампы $\bar{X}=3000$ ч и среднее квадратическое отклонение $s=20$ ч. Считая, что срок службы каждой лампы является нормально распределенной случайной величиной, определить с надежностью 0,9 доверительный интервал для математического ожидания.

$\triangle$ В соответствии с (9.61) находим

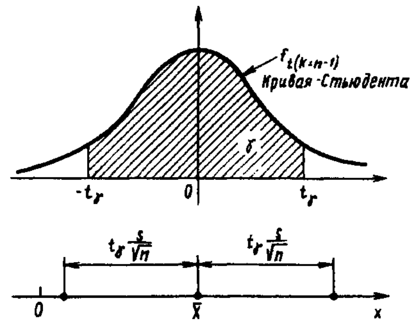


Рис. 9.7

$$\left.\begin{matrix}\gamma=0{,}9 \\ k=16-1\end{matrix}\right\} \xrightarrow{\text{П.3}} t_\gamma=1{,}753 \to \varepsilon=\frac{1{,}753\cdot 20}{4}=8{,}765 \to \begin{cases}3000-8{,}765=2991{,}235,\\ 3000+8{,}765=3008{,}765.\end{cases}$$

Таким образом, с вероятностью 0,9 можно быть уверенным в том, что доверительный интервал (2991,235; 3008,765) накроет неизвестное математическое ожидание, а среднее $\bar{X}=3000$ ч определяет значение математического ожидания с точностью 8,765. ▲

Задача 9.7. По результатам измерения диаметра 25 корпусов электродвигателей было получено, что $\bar{X}=100$ мм, $s=16$ мм. Предполагая нормальное распределение результата измерения, найти вероятность того, что $(0{,}9\bar{X};\ 1{,}1\bar{X})$ накроет математическое ожидание.

$\triangle$ Приравняем нижнюю границу, равную $0{,}9\ \bar{X}$, нижней границе в схеме (9.61), получаем $0{,}9\bar{X}=\bar{X}-\varepsilon$. Отсюда $\varepsilon=0{,}1\bar{X}=0{,}1\cdot 100=10$. Такой же результат будем иметь, если приравнять $1{,}1\bar{X}$ верхней границе, равной $\bar{X}+\varepsilon$.

Из (9.59) получаем $t_\gamma=\varepsilon\sqrt{n}/s$; в условиях примера $t_\gamma=10\cdot 5/16=3{,}125$. Зная $t_\gamma=3{,}125$ и число степеней свободы $k=n-1=24$, по табл. П.3 находим, что $\gamma\approx 0{,}99$. Таким образом, вероятность $P(0{,}9\bar{X}<MX<1{,}1\bar{X})\approx 0{,}99$. ▲

Интервальная оценка среднего квадратического отклонения и дисперсии нормального распределения. Итак, мы рассматриваем нормально распределенную случайную величину X, дисперсия σ^2 которой неизвестна. Проведено n наблюдений этой величины, результаты которых $X_1, X_2, \ldots, X_n$ обладают свойствами (9.43) и (9.44). По этим результатам вычислим среднюю $\bar{X}=\sum_{i=1}^{n} X_i/n$ и оценку $s^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)$ неизвестной дисперсии σ^2. За оценку среднего квадратического отклонения σ примем $s=+\sqrt{s^2}$.

Теперь зададимся надежностью γ интервальной оценки и найдем такое число ε, чтобы выполнялось соотношение

$$P(s-\varepsilon<\sigma<s+\varepsilon)=\gamma. \tag{9.62}$$

$\square$ Среднее квадратическое отклонение σ случайной величины всегда положительно $(\sigma>0)$, поэтому ε разумнее находить не из условия (9.62), а из условия

$$P[\max(0;\ s-\varepsilon)<\sigma<s+\varepsilon]=\gamma. \tag{9.63}$$

Действительно, если $s-\varepsilon\leqslant 0$, то максимум из чисел 0 и $(s-\varepsilon)$ равен нулю и нижней границей для σ является ноль; если $s-\varepsilon>0$, то нижней границей для σ является $s-\varepsilon$.

Нахождение числа ε основано на следующей теореме, которую приведем без доказательства.

Теорема о распределении случайной величины $\frac{(n-1)s^2}{\sigma^2}$. *Пусть $X_1, X_2, \dots, X_n$ — результаты n независимых наблюдений нормально распределенной случайной величины, проводимых в одинаковых условиях, т. е. $X_1, X_2, \dots, X_n$ обладают свойствами (9.43) и (9.44). Тогда величина $\frac{(n-1)s^2}{\sigma^2}$ имеет χ^2-распределение с числом степеней свободы, равным $(n-1)$* (см. § 6.5), *или, иначе, величина $\frac{(n-1)s^2}{\sigma^2}$ является χ^2-величиной с $(n-1)$-й степенью свободы:*

$$\frac{(n-1)s^2}{\sigma^2}=\chi^2(k=n-1). \tag{9.64}$$

Воспользуемся табл. П.4 и найдем для заданных вероятности γ и числа $k=n-1$ число q_γ такое, при котором вероятность

$$P\left[\frac{k}{(1+q_\gamma)^2}<\chi^2(k)<\frac{k}{\max^2(0;\,1-q_\gamma)}\right]=\gamma.$$

Учитывая (9.64), получаем

$$P\left[\frac{n-1}{(1+q_\gamma)^2}<\frac{(n-1)s^2}{\sigma^2}<\frac{n-1}{\max^2(0;\,1-q_\gamma)}\right]=\gamma. \tag{9.65}$$

Двойное неравенство, стоящее в скобках выражения (9.65), разделим на $(n-1)$; единицу разделим на каждую часть неравенства, все части умножим на s^2, а затем извлечем квадратный корень. В результате получим

$$P\left[s\max(0;\,1-q_\gamma)<\sigma<s(1+q_\gamma)\right]=\gamma,$$

или

$$P\left[\max(0;\,s(1-q_\gamma))<\sigma<s(1+q_\gamma)\right]=\gamma. \tag{9.66}$$

Сравнивая (9.66) с (9.63), заключаем, что

$$\varepsilon=sq_\gamma. \blacksquare \tag{9.67}$$

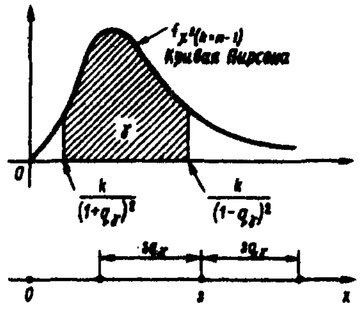


Рис. 9.8

Таким образом, с вероятностью γ можно утверждать, что интервал

$$(\max(0;\,s-sq_\gamma);\ s+sq_\gamma) \tag{9.68}$$

накроет неизвестное среднее квадратическое отклонение σ; с такой же вероятностью можно утверждать, что интервал

$$(\max^2(0;\ s-sq_\gamma),\ (s+sq_\gamma)^2) \tag{9.69}$$

накроет неизвестную дисперсию σ^2.

В заключение приведем схему нахождения ε и границ интервала (9.68), соответствующего надежности γ:

$$\left.\begin{matrix}\gamma\\ k=n-1\end{matrix}\right\}\xrightarrow{\text{П.4}} q_\gamma\longrightarrow\varepsilon=sq_\gamma\begin{cases}\to\max(0;\,s-\varepsilon)\text{(нижняя граница)},\\ \to s+\varepsilon\text{ (верхняя граница)}.\end{cases} \tag{9.70}$$

Графическая иллюстрация схемы приведена на рис. 9.8.

Задача 9.8. Вычислить с надежностью 0,98 интервальную оценку для дисперсии нормального распределения, если по выборке объема $n=17$ вычислена оценка $s^2=25$.

Δ В соответствии с (9.70) имеем

$$\left.\begin{matrix}\gamma=0{,}98\\ k=17-1\end{matrix}\right\}\xrightarrow{\text{П.4}} q_\gamma=0{,}564\longrightarrow\varepsilon=\sqrt{25}\cdot 0{,}564=$$

$$=2{,}82\begin{cases}\to\max(0;\,5-2{,}82)=2{,}18,\\ \to 5+2{,}82=7{,}82.\end{cases}$$

Таким образом, интервальная оценка для среднего квадратического отклонения имеет вид (2,18; 7,82), а для дисперсии — следующий вид: (4,7524; 61,1524). Смысл полученного результата такой: с вероятностью 98% можно ожидать, что интервал (4,7524; 61,1524) накроет неизвестную дисперсию. $\blacktriangle$

Задача 9.9. По результатам измерения диаметра 51 корпуса электродвигателей было получено, что $s^2=2{,}56$. Предполагая, что распределение результата измерения нормальное, найти вероятность того, что интервал $(s-0{,}16;\ s+0{,}16)$ накроет среднее квадратическое отклонение σ.

△ Приравняем верхнюю границу, равную $s+0{,}16$, верхней границе в схеме (9.70), получаем $s+0{,}16=s+\varepsilon$. Отсюда $\varepsilon=0{,}16$. Из (9.67) будем иметь $q_\gamma=\varepsilon/s$; в условиях примера $q_\gamma=0{,}1$.

Зная $q_\gamma=0{,}1$ и число степеней свободы $k=n-1=50$, по табл. П.4 найдем $\gamma\approx0{,}7$. Таким образом, вероятность $P(s-0{,}16<\sigma<s+0{,}16)\approx0{,}7$. ▲

§ 9.8. Интервальная оценка вероятности события

Выше было показано, что «хорошей» точечной оценкой вероятности p события A является частость $\hat{p}=m/n$, где n — общее число независимых испытаний, в каждом из которых событие A может произойти с вероятностью p или не произойти с вероятностью $q=1-p$ (напомним, что серия испытаний подобного типа называется последовательностью испытаний Бернулли), а m — число испытаний, в которых произошло событие A.

Зададимся вероятностью γ и найдем числа p_1 и p_2 такие, чтобы выполнялось соотношение

$$P(p_1<p<p_2)=\gamma. \qquad (9.71)$$

Интервал (p_1, p_2) и является интервальной оценкой вероятности p, отвечающей надежности γ.

Интервальную оценку построим для двух случаев: когда число ипытаний Бернулли велико и для малого 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, причем при большом n и немалой вероятности p распределение величины m в силу локальной и интегральной теорем Муавра — Лапласа близко к нормальному распределению с математическим ожиданием, равным np, и дисперсией, равной npq, т. е.

$$m\approx N(np,\ \sqrt{npq}). \qquad (9.72)$$

Напомним, что «грубое» правило использования формулы (9.72) состоит в том, что n должно быть порядка нескольких десятков, лучше сотен, а $np>10$.

При делении m на постоянную величину n вид закона распределения не изменится, изменятся только его параметры. Поэтому при большом n распределение частости $\hat{p}=m/n$, так же как и распределение частоты m, близко к нормальному, но с математическим ожиданием

$$M\hat{p}=M\left(\frac{m}{n}\right)\underset{n=\mathrm{const}}{=\!=\!=}\frac{1}{n}M(m)=\frac{1}{n}np=p$$

и дисперсией

$$D\hat{p}=D\left(\frac{m}{n}\right)\underset{n=\mathrm{const}}{=\!=\!=}\frac{1}{n^2}D(m)=\frac{1}{n^2}npq=\frac{pq}{n}.$$

Таким образом, при большом числе n испытаний Бернулли

$$\hat{p}\approx N\left(p,\ \sqrt{pq/n}\right). \qquad (9.73)$$

Следовательно, при большом n распределение величины $\frac{\hat{p}-p}{\sqrt{pq/n}}$ близко к нормальному с нулевым математическим ожиданием и единичной дисперсией, т. е.

$$\frac{\hat{p}-p}{\sqrt{pq/n}}\approx N(0,\ 1). \qquad (9.74)$$

Теперь, используя табл. П.2 функции Лапласа, найдем для заданной вероятности γ число u_γ такое, при котором $\Phi(u_\gamma)=\gamma$, или, иначе, такое, при котором

$$P(|N(0,\ 1)|<u_\gamma)=\gamma.$$

Учитывая (9.74), получим

$$P\left(\left|\frac{\hat{p}-p}{\sqrt{pq/n}}\right|<u_\gamma\right)=\gamma. \qquad (9.75)$$

Неравенство, стоящее в скобках выражения (9.75), решим относительно p.

Для этого:

возведем его в квадрат, в результате чего получим

$$(\hat{p}-p)^2<\frac{p(1-p)}{n}u_\gamma^2;$$

возведем $(\hat{p}-p)$ в квадрат и перенесем все члены влево; имеем

$$\left(1+\frac{u_\gamma^2}{n}\right)p^2-\left(2\hat{p}+\frac{u_\gamma^2}{n}\right)p+\hat{p}^2<0; \qquad (9.76)$$

найдем корни p_1 и p_2 квадратного трехчлена, стоящего в правой части неравенства (9.76); имеем

$$p_1=\frac{\hat{p}+u_\gamma^2/(2n)-u_\gamma\sqrt{\hat{p}(1-\hat{p})/n+u_\gamma^2/(4n^2)}}{1+u_\gamma^2/n}, \qquad (9.77)$$

$$p_2=\frac{\hat{p}+u_\gamma^2/(2n)+u_\gamma\sqrt{\hat{p}(1-\hat{p})/n+u_\gamma^2/(4n^2)}}{1+u_\gamma^2/n}. \qquad (9.78)$$

Так как коэффициент $\left(1+\frac{u_\gamma^2}{n}\right)$ квадратного трехчлена в неравенстве (9.76) положителен, то решением этого неравенства является интервал (p_1, p_2).

Приведенному решению можно дать геометрическую иллюстрацию. Рассмотрим декартову систему координат, у которой по оси абсцисс откладывается частость $\hat{p}$, а по оси ординат — вероятность p (рис. 9.9). Точки $(\hat{p}, p)$, координаты которых удовлетворяют неравенству (9.76), располагаются внутри эллипса. Чтобы построить интервальную оценку вероятности p при известной частости $\hat{p}$, надо рассмотреть множество точек внутри эллипса

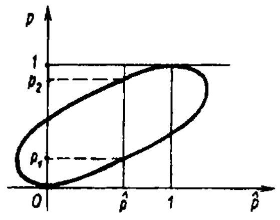


Рис. 9.9

с абсциссой, равной $\hat{p}$,— этот интервал и является искомой интервальной оценкой (p_1, p_2).

Итак, при вычислении p_1 и p_2 по формулам (9.77) и (9.78) вероятность $P(p_1<p<p_2)=\gamma$.

Приведем схему построения интервальной оценки, отвечающей надежности γ:

$$\gamma \xrightarrow{\text{П.2}} u_\gamma \begin{cases} \to p_1 \text{ (вычисляется по формуле (9.77))}, \\ \to p_2 \text{ (вычисляется по формуле (9.78))}, \end{cases} \qquad (9.79)$$

при этом в формулах (9.77) и (9.78) $\hat{p}=m/n$.

Из формул (9.77) и (9.78) при $n \gg 100$ (n значительно больше 100) можно получить следующие приближенные формулы:

$$p_1 \approx \hat{p} - u_\gamma \sqrt{\hat{p}(1-\hat{p})/n},$$
$$p_2 \approx \hat{p} + u_\gamma \sqrt{\hat{p}(1-\hat{p})/n}.$$

В этом случае с вероятностью γ можно утверждать, что интервал

$$\left(\hat{p} - u_\gamma\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p} + u_\gamma\sqrt{\hat{p}(1-\hat{p})/n}\right) \qquad (9.79')$$

накроет неизвестную вероятность p, или, иначе, с вероятностью γ можно быть уверенным в том, что вычисленная по результатам n испытаний Бернулли частость $\hat{p}=m/n$ определяет значение неизвестной вероятности p с точностью

$$\varepsilon = u_\gamma\sqrt{\hat{p}(1-\hat{p})/n}. \qquad (9.80)$$

При $n \gg 100$ схема построения интервальной оценки, отвечающей надежности γ, такова:

$$\gamma \xrightarrow{\text{П.2}} u_\gamma \to \varepsilon = u_\gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \begin{cases} \to \hat{p}-\varepsilon \text{ (нижняя граница)}, \\ \to \hat{p}+\varepsilon \text{ (верхняя граница)}, \end{cases} \qquad (9.81)$$

где $\hat{p}=m/n$.

Задача 9.10. Событие A в серии из $n=100$ испытаний Бернулли произошло $m=78$ раз. Найти интервальную оценку для вероятности p события A с надежностью $\gamma=0{,}9$.

△ В условиях примера точечная оценка вероятности p равна $\hat{p}=78/100=0{,}78$. Поскольку $n=100$, построение интервальной оценки проведем по схеме (9.79):

$$\gamma=0{,}9 \xrightarrow{\text{П.2}} u_\gamma = 1{,}643 <^{p_1}_{p_2};$$

$$p_1 = \frac{0{,}78+\dfrac{1{,}643^2}{200} - 1{,}643\sqrt{\dfrac{0{,}78\cdot 0{,}22}{100}+\dfrac{1{,}643^2}{4\cdot 100^2}}}{1+\dfrac{1{,}643^2}{100}} = 0{,}705,$$

$$p_2 = \frac{0{,}78+\dfrac{1{,}643^2}{200} + 1{,}643\sqrt{\dfrac{0{,}78\cdot 0{,}22}{100}+\dfrac{1{,}643^2}{4\cdot 100^2}}}{1+1{,}643^2/100} = 0{,}848.$$

Следовательно, с надежностью 0,9 интервал (0,705; 0,848) накроет неизвестную вероятность p. ▲

Задача 9.11. 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.

△ Здесь $n=1000$, что значительно больше 100. Поэтому для нахождения вероятности γ воспользуемся схемой (9.81).

В условиях примера точечная оценка вероятности p равна $\hat{p}=50/1000=0{,}05$, а следовательно, в (9.81) 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 получить ответ на вопрос задачи используемым здесь способом было бы нельзя).

Теперь по формуле (9.80) найдем

$$u_\gamma = \frac{\varepsilon}{\sqrt{\hat{p}(1-\hat{p})/n}} = \frac{0{,}01}{\sqrt{0{,}1\cdot 0{,}9/1000}} = 1{,}16.$$

Зная $u_\gamma=1{,}16$, по табл. П.2 найдем $\gamma=0{,}7540$. Таким образом, вероятность $P(0{,}04<p<0{,}06)=0{,}7540$. ▲

Интервальная оценка вероятности при малом числе n испытаний Бернулли. В этом случае соотношение (9.72) не будет иметь места и следует использовать точный, а не приближенный закон распределения величины m. В § 6.3 было установлено, что 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

$$P(m=x) = C_n^x p^x (1-p)^{n-x}, \quad x=0,\ 1,\ 2,\ \ldots,\ n, \qquad (9.82)$$

— это точный закон распределения величины m.

Зададимся вероятностью γ и найдем числа p_1 и p_2 такие, чтобы

$$P(p_1<p<p_2)=\gamma.$$

Можно показать, что при использовании закона (9.82) число p_1 является решением уравнения

$$\sum_{x=0}^{m-1} C_n^x p_1^x (1-p_1)^{n-x} = \frac{1+\gamma}{2}, \qquad (9.83)$$

а число p_2 — решением уравнения

$$\sum_{x=0}^{m} C_n^x p_2^x (1-p_2)^{n-x} = \frac{1-\gamma}{2}. \qquad (9.84)$$

Обратим внимание на то, что и в (9.83), и в (9.84) m — конкретное число (именно число) испытаний среди n испытаний Бернулли, в которых произошло событие A.

Существуют специальные таблицы для нахождения чисел p_1 и p_2, удовлетворяющих уравнениям (9.83) и (9.84), по заданным n, $n-m$ и γ. Фрагмент их дан в табл. П.5.

Задача 9.12. В пяти испытаниях Бернулли событие A произошло три раза. Найти с надежностью 0,95 интервальную оценку для вероятности p появления события A в единичном испытании.

△ Имеем $n=5$, $m=3$, $\gamma=0{,}95$. По табл. П.5 находим $p_1=0{,}147$, $p_2=0{,}947$. Таким образом, интервальная оценка вероятности p, отвечающая надежности 0,95, такова: (0,147; 0,947). ▲

§ 9.9. Понятие доверительной области

Выше были рассмотрены схемы построения интервальной оценки, или доверительного интервала для какой-то одной, отдельно взятой числовой характеристики.

В ряде задач рассматриваются одновременно несколько характеристик. Обозначим их Θ_1, Θ_2, ..., Θ_ν и построим доверительные интервалы сразу для всех этих характеристик, т. е. найдем такие интервалы Δ_1, Δ_2, ..., Δ_ν, чтобы вероятность произведения следующих событий:

события A_1, состоящего в том, что интервал Δ_1 накроет неизвестную характеристику Θ_1;

события A_2, состоящего в том, что интервал Δ_2 накроет неизвестную характеристику Θ_2;

....

события A_ν, состоящего в том, что интервал Δ_ν накроет неизвестную характеристику Θ_ν,

была равна γ. Иначе говоря,

$$P(A_1A_2\dots A_\nu)=P[(\Theta_1\in\Delta_1)(\Theta_2\in\Delta_2)\dots(\Theta_\nu\in\Delta_\nu)]=\gamma. \quad (9.85)$$

Область, образованная интервалами Δ_1, Δ_2, ..., Δ_ν, называется *доверительной областью* для совокупности характеристик Θ_1, Θ_2, ..., Θ_ν, а вероятность γ — *надежностью* доверительной области.

Вообще говоря, задача построения доверительной области довольно сложная. Ее решение упрощается в том случае, если события A_1, A_2, ..., A_m независимы. При независимости этих событий цепочка равенств (9.85) тождественна следующей цепочке:

$$P(A_1)P(A_2)\dots P(A_\nu)=P(\Theta_1\in\Delta_1)P(\Theta_2\in\Delta_2)\dots P(\Theta_\nu\in\Delta_\nu)=\gamma. \quad (9.86)$$

Полагая, что $P(A_1)=P(A_2)=\dots=P(A_\nu)$, из (9.86) получаем

$$P(\Theta_1\in\Delta_1)=\sqrt[\nu]{\gamma},$$

$$P(\Theta_2\in\Delta_2)=\sqrt[\nu]{\gamma},$$

....

$$P(\Theta_\nu\in\Delta_\nu)=\sqrt[\nu]{\gamma},$$

т. е. построение доверительной области свелось к построению для каждой из характеристик Θ_1, Θ_2, ... Θ_ν доверительного интервала с надежностью, равной $\sqrt[\nu]{\gamma}$.

ГЛАВА 10
ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

§ 10.1. 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 генеральная средняя (математическое ожидание случайной величины) равна 5. 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 этапы проверки гипотезы и используемые при этом понятия.

Этап 1. Располагая выборочными данными X_1, X_2, ..., 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 будет принято решение о справедливости для генеральной совокупности гипотезы H_0;

по выборке будет принято решение о справедливости для генеральной совокупности гипотезы H_1.

Гипотезу H_1 называют также *альтернативной*.

Например, если нулевая гипотеза такова: математическое ожидание равно 5,— то альтернативная гипотеза может быть следующей: математическое ожидание меньше 5, что записывается следующим образом:

$$H_0: MX=5;\ H_1: MX<5.$$

Этап 2. Задаются вероятностью α («альфа»), 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 ошибка двух родов:

отвергают гипотезу H_0, или, иначе, 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 высказывание H_0 неверно, т. е. верной является гипотеза H_1; это *ошибка второго рода*.

Так вот уровень значимости α — это вероятность ошибки первого рода, т. е.

$$\alpha = P_{H_0}(H_1), \tag{10.1}$$

где $P_{H_0}(H_1)$ — вероятность того, что будет принята гипотеза H_1, если на самом деле в генеральной совокупности верна гипотеза H_0. Вероятность α задается заранее, разумеется, 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 5 случаях из 100 мы совершим ошибку первого рода.

Вероятность ошибки второго рода обозначают β, т. е.

$$\beta = P_{H_1}(H_0), \tag{10.2}$$

где $P_{H_1}(H_0)$ — вероятность того, что будет принята гипотеза H_0, если на самом деле верна гипотеза H_1. В задаче 10.1 будет показано, что, зная α, можно найти вероятность β.

Сказанное иллюстрирует табл. 10.1.

Таблица 10.1

Решение, принимаемое о гипотезе H_0 по выборке / Верна ли гипотеза H_0 или нет?	Гипотеза отвергается, т. е. принимается гипотеза H_1	Гипотеза H_0 принимается
Гипотеза H_0 верна	Ошибка первого рода, ее вероятность $P_{H_0}(H_1)=\alpha$	Правильное решение, его вероятность $P_{H_0}(H_0)=1-\alpha$
Гипотеза H_0 неверна, т. е. верна гипотеза H_1	Правильное решение, его вероятность $P_{H_1}(H_1)=1-\beta$	Ошибка второго рода, ее вероятность $P_{H_1}(H_0)=\beta$

Обратим внимание на то, что в результате проверки гипотезы относительно гипотезы H_0 может быть принято и правильное решение. Существует правильное решение двух следующих видов:

принимают гипотезу H_0, тогда как и в действительности, в генеральной совокупности, она имеет место; вероятность этого решения $P_{H_0}(H_0)=1-\alpha$;

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 $P_{H_1}(H_1)=1-\beta$.

Этап 3. Находят величину φ такую, что:

ее значения зависят от выборочных данных $X_1, X_2, \ldots, X_n$, т. е. для которой справедливо равенство $\varphi=\varphi(X_1, X_2, \ldots, X_n)$;

ее значения позволяют судить о «расхождении выборки с гипотезой H_0»;

и которая, будучи величиной случайной в силу случайности выборки $X_1, X_2, \ldots, X_n$, подчиняется при выполнении гипотезы H_0 некоторому известному, затабулированному закону распределения.

Величину φ называют *критерием*.

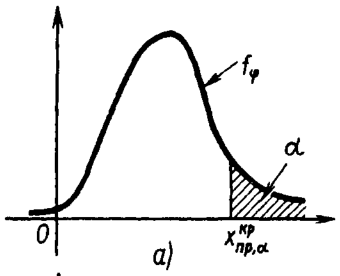
Отметим, что в основе метода построения критерия лежит понятие функции правдоподобия.

Этап 4. Далее рассуждают так. Так как значения критерия позволяют судить о «расхождении выборки с гипотезой H_0», то из области допустимых значений критерия φ следует выделить подобласть ω таких значений, которые свидетельствовали бы о существенном расхождении выборки с гипотезой H_0 и, следовательно, о невозможности принять гипотезу H_0. Подобласть ω называют *критической областью*. Допустим, что критическая область выделена. Тогда руководствуются следующим правилом: если вычисленное по выборке значение критерия φ попадает в критическую область, то гипотеза H_0 отвергается и принимается гипотеза 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 критерий φ примет значение из критической области ω, должна быть равна заданному числу α, т. е.

$$P(\varphi \in \omega)=\alpha. \tag{10.3}$$

Однако критическая область равенством (10.3) определяется неоднозначно. Действительно, представив себе график функции плотности $f_\varphi(x)$ критерия φ, нетрудно понять, что на оси



абсцисс существует бесчисленное множество областей-интервалов таких, что площади построенных на них криволинейных трапеций равны α, т. е. областей, удовлетворяющих требованию (10.3). Поэтому кроме тре-

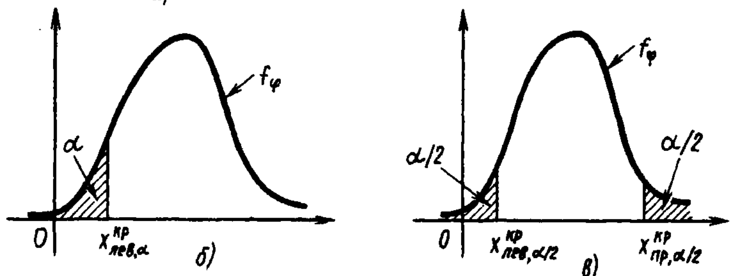


Рис. 10.1

бования (10.3) 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.

Возможны три вида расположения критической области (в зависимости от вида нулевой и альтернативной гипотез, вида и распределения критерия φ):

правосторонняя критическая область (рис. 10.1, *а*), состоящая из интервала $(x^{\text{кр}}_{\text{пр},\alpha}, +\infty)$, где точка $x^{\text{кр}}_{\text{пр},\alpha}$ определяется из условия

$$P(\varphi > x^{\text{кр}}_{\text{пр},\alpha}) = \alpha \tag{10.4}$$

и называется *правосторонней критической точкой*, отвечающей уровню значимости α;

левосторонняя критическая область (рис. 10.1, *б*), состоящая из интервала $(-\infty, x^{\text{кр}}_{\text{лев},\alpha})$, где точка $x^{\text{кр}}_{\text{лев},\alpha}$ определяется из условия

$$P(\varphi < x^{\text{кр}}_{\text{лев},\alpha}) = \alpha \tag{10.5}$$

и называется *левосторонней критической точкой*, отвечающей уровню значимости α;

двусторонняя критическая область (рис. 10.1, *в*), состоящая из следующих двух интервалов: $(-\infty, x^{\text{кр}}_{\text{лев},\alpha/2})$ и $(x^{\text{кр}}_{\text{пр},\alpha/2}, +\infty)$, где точки $x^{\text{кр}}_{\text{лев},\alpha/2}$ и $x^{\text{кр}}_{\text{пр},\alpha/2}$ определяются из условий

$$P(\varphi < x^{\text{кр}}_{\text{лев},\alpha/2}) = \alpha/2 \quad \text{и} \quad P(\varphi > x^{\text{кр}}_{\text{пр},\alpha/2}) = \alpha/2 \tag{10.6}$$

и называются *двусторонними критическими точками*.

Замечание. По значению критерия φ 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 критерия φ (бóльших критической точки).

Включение же в ряде случаев в критическую область малых значений критерия φ (меньших критической точки) на первый взгляд противоречит смыслу этой величины. Однако не следует забывать, что φ — случайная величина (она зависит от результатов наблюдений $X_1, X_2, \ldots, X_n$, которые случайны), поэтому маловероятно появление не только слишком больших, но и слишком малых ее значений и их следует включить в критическую область.

Этап 5. В формулу критерия $\varphi = \varphi(X_1, X_2, \ldots, X_n)$ вместо $X_1, X_2, \ldots, X_n$ подставляют конкретные числа, полученные в результате n наблюдений, и подсчитывают числовое значение $\varphi_{\text{чис}}$ критерия.

Если $\varphi_{\text{чис}}$ попадает в критическую область ω, то гипотеза H_0 отвергается и принимается гипотеза H_1. Поступая таким образом, следует понимать, что можно допустить ошибку с вероятностью α.

Если $\varphi_{\text{чис}}$ не попадает в критическую область, 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.

§ 10.2. Проверка гипотез о числовых значениях параметров нормального распределения

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 выборочные наблюдения, при этом предполагается, что они независимы и условия их проведения одинаковы, т. е. возможные результаты $X_1, X_2, \ldots, X_n$ наблюдений обладают свойствами (9.43) и (9.44). По этим наблюдениям вычисляют $\bar{X} = \sum_{i=1}^{n} X_i/n$ и $s^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$. Полученные 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.

Проверка гипотезы о числовом значении математического ожидания нормального распределения при известной 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 неизвестный параметр a равен числу a_0. Относительно альтернативной гипотезы H_1 возможны три случая: 1) параметр a равен числу a_1, которое больше числа a_0; 2) параметр a равен числу a_1, которое меньше a_0; 3) параметр a равен 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

$$H_0: a=a_0 \tag{10.7}$$

и альтернативную

$$H_1: a=a_1>a_0. \tag{10.8}$$

Этап 2. Зададимся уровнем значимости α.

Этап 3. В качестве критерия возьмем величину

$$\varphi=\frac{\bar{X}-a_0}{\sigma/\sqrt{n}}, \tag{10.9}$$

которая удовлетворяет следующим требованиям, предъявляемым к критериям:

величина (10.9) зависит от выборочных данных $X_1, X_2, ..., X_n$, поскольку $\bar{X}=(X_1+X_2+...+X_n)/n$;

по значениям величины (10.9) можно судить о «расхождении выборки с гипотезой H_0»; в данном случае это следует понимать так: чем ближе выборочная средняя $\bar{X}$ к предполагаемому гипотезой H_0 значению a_0 математического ожидания, тем меньше по модулю значение величины (10.9);

величина (10.9) при выполнении гипотезы (10.7) (а также при соблюдении требований независимости наблюдений величины X и одинаковости условий этих наблюдений) подчиняется нормальному закону распределения с нулевым математическим ожиданием и единичной дисперсией (см. (9.46)), т. е.

$$\varphi=\frac{\bar{X}-a_0}{\sigma/\sqrt{n}}=N(0, 1). \tag{10.10}$$

Этап 4. Выделим критическую область ω — область таких значений критерия φ, при которых гипотеза H_0 отвергается.

Если нулевая и альтернативная гипотезы имеют вид соответственно (10.7) и (10.8), а критерий — вид (10.10), то критическая область будет правосторонней: ее образует интервал $(x^{кр}_{пр,\alpha}, +\infty)$, где точка $x^{кр}_{пр,\alpha}$ определяется из условия (10.4), которое, учитывая (10.10), запишем в виде

$$P(N(0, 1)>x^{кр}_{пр,\alpha})=\alpha. \tag{10.11}$$

Исходя из этого равенства найдем критическую точку $x^{кр}_{пр,\alpha}$.

Воспользуемся табл. П.2. Полагая $\gamma=1-2\alpha$, найдем число u_γ такое, при котором

$$P(|N(0, 1)|<u_\gamma)=\gamma,$$

или

$$P(-u_\gamma<N(0, 1)<u_\gamma)=1-2\alpha. \tag{10.12}$$

Но тогда вероятность противоположного события такова:

$$P(N(0, 1)<-u_\gamma \text{ или } N(0, 1)>u_\gamma)=2\alpha, \tag{10.13}$$

а так как для случайной величины $N(0, 1)$

$$P(N(0, 1)<-u_\gamma)=P(N(0, 1)>u_\gamma),$$

то из (10.13) получим

$$P(N(0, 1)>u_\gamma)=\alpha. \tag{10.14}$$

Сопоставив (10.14) с (10.11), заключаем, что $x^{кр}_{пр,\alpha}=u_{\gamma=1-2\alpha}$.

Итак, схема нахождения точки $x^{кр}_{пр,\alpha}$ следующая:

$$\alpha\to\gamma=1-2\alpha\xrightarrow{\text{П.2}}u_\gamma\to x^{кр}_{пр,\alpha}=u_\gamma. \tag{10.15}$$

Критическая область изображена на рис. 10.2, *а*.

Этап 5. Подставим вместо $X_1, X_2, ..., X_n$ в формулу $\bar{X}=\sum_{i=1}^{n}X_i/n$ конкретные числа, полученные в результате наблюдений, а затем найдем числовое значение $\varphi_{чис}$ критерия (10.9).

Если $\varphi_{чис}>x^{кр}_{пр,\alpha}$, то гипотезу H_0 (10.7) отвергаем и принимаем гипотезу H_1 (10.8). Поступая таким образом, можно допустить ошибку (на самом деле гипотеза (10.7) будет верной,

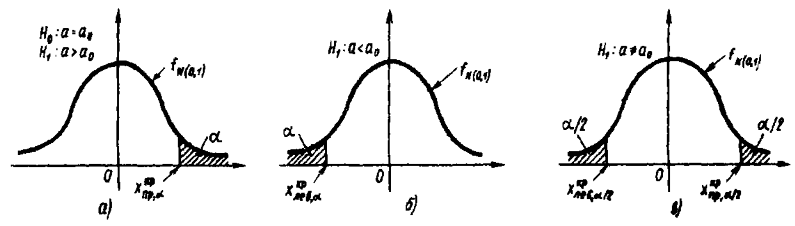


Рис. 10.2

а мы ее отвергли); вероятность появления такой ошибки равна α.

Если $\varphi_{\text{чис}} < x^{\text{кр}}_{\text{пр},\alpha}$, то гипотеза H_0 (10.7) не отвергается.

Случай 2.

Этап 1. Сформулируем нулевую гипотезу

$$H_0: a = a_0 \tag{10.16}$$

и альтернативную

$$H_1: a = a_1 < a_0. \tag{10.17}$$

Этап 2. Зададимся уровнем значимости α.

Этап 3. В качестве критерия, как и в случае 1, возьмем величину (10.9), которая при выполнении гипотезы (10.16) (а также при соблюдении требований независимости наблюдений величины X и одинаковости условий их проведения) удовлетворяет равенству (10.10).

Этап 4. Если нулевая и альтернативная гипотезы имеют соответственно вид (10.16) и (10.17), а критерий — вид (10.10), то критическая область будет левосторонней: ее образует интервал $(-\infty, x^{\text{кр}}_{\text{лев},\alpha})$, где точка $x^{\text{кр}}_{\text{лев},\alpha}$ определяется из условия (10.5), которое, учитывая (10.10), запишем в виде

$$P(N(0,\,1) < x^{\text{кр}}_{\text{лев},\alpha}) = \alpha. \tag{10.18}$$

Исходя из этого равенства найдем критическую точку $x^{\text{кр}}_{\text{лев},\alpha}$.

Воспользуемся табл. П.2. Полагая $\gamma = 1 - 2\alpha$, найдем число u_γ такое, при котором

$$P(|N(0,\,1)| < u_\gamma) = \gamma,$$

или

$$P(-u_\gamma < N(0,\,1) < u_\gamma) = 1 - 2\alpha.$$

Тогда

$$P(N(0,\,1) < -u_\gamma \text{ или } N(0,\,1) > u_\gamma) = 2\alpha. \tag{10.19}$$

Так как $P(N(0,\,1) < -u_\gamma) = P(N(0,\,1) > u_\gamma)$, то из (10.19) получим

$$P(N(0,\,1) < -u_\gamma) = \alpha. \tag{10.20}$$

Сопоставив (10.20) и (10.18), заключаем, что $x^{\text{кр}}_{\text{лев}.\alpha} = -u_{\gamma = 1 - 2\alpha}$.

Итак, схема нахождения точки $x^{\text{кр}}_{\text{лев}.\alpha}$ следующая:

$$\alpha \to \gamma = 1 - 2\alpha \xrightarrow{\text{П.2}} u_\gamma \to x^{\text{кр}}_{\text{лев}.\alpha} = -u_\gamma. \tag{10.21}$$

Критическая область изображена на рис. 10.2, *б*.

Этап 5. Находим числовое значение $\varphi_{\text{чис}}$ критерия (10.9). Если $\varphi_{\text{чис}} < x^{\text{кр}}_{\text{лев}.\alpha}$, то гипотезу H_0 (10.16) отвергаем и принимаем гипотезу H_1 (10.17). Поступая таким образом, можно с вероятностью α допустить ошибку первого рода.

Если $\varphi_{\text{чис}} > x^{\text{кр}}_{\text{лев}.\alpha}$, то гипотеза H_0 (10.16) не отвергается.

Случай 3.

Этап 1. Сформулируем нулевую гипотезу

$$H_0: a = a_0 \tag{10.22}$$

и альтернативную

$$H_1: a = a_1 \neq a_0. \tag{10.23}$$

Этап 2. Зададимся уровнем значимости α.

Этап 3. В качестве критерия, как и в случаях 1 и 2, возьмем величину (10.9), которая при выполнении гипотезы (10.22) удовлетворяет равенству (10.10).

Этап 4. Если нулевая и альтернативная гипотезы имеют соответственно вид (10.22) и (10.23), а критерий — вид (10.10), то критическая область будет двусторонней: ее образуют интервалы $(-\infty, x^{\text{кр}}_{\text{лев}.\alpha/2})$ и $(x^{\text{кр}}_{\text{пр}.\alpha/2}, +\infty)$, где критические точки определяются из условия (10.6), которое, учитывая (10.10), запишем в виде

$$P(N(0,\,1) < x^{\text{кр}}_{\text{лев}.\alpha/2}) = \frac{\alpha}{2} \text{ и } P(N(0,\,1) > x^{\text{кр}}_{\text{пр}.\alpha/2}) = \frac{\alpha}{2}. \tag{10.24}$$

Для нахождения критических точек воспользуемся табл. П.2. Полагая $\gamma = 1 - \alpha$, найдем число u_γ такое, при котором

$$P(|N(0,\,1)| < u_\gamma) = \gamma,$$

или

$$P(-u_\gamma < N(0,\,1) < u_\gamma) = 1 - \alpha.$$

Но тогда

$$P(N(0,\,1) < -u_\gamma \text{ или } N(0,\,1) > u_\gamma) = \alpha. \tag{10.25}$$

Так как $P(N(0,\,1) < -u_\gamma) = P(N(0,\,1) > u_\gamma)$, то из (10.25) получим

$$P(N(0,\,1) < -u_\gamma) = \alpha/2 \text{ и } P(N(0,\,1) > u_\gamma) = \alpha/2. \tag{10.26}$$

Сопоставляя (10.26) с (10.24), заключаем, что

$x^{\text{кр}}_{\text{лев}.\alpha/2} = -u_{\gamma = 1 - \alpha}$, $x^{\text{кр}}_{\text{пр}.\alpha/2} = u_{\gamma = 1 - \alpha}$.

Итак, схема нахождения двусторонних критических точек для заданного уровня значимости α следующая:

$$\alpha \to \gamma = 1 - \alpha \xrightarrow{\text{П.2}} u_\gamma \begin{cases} \to x^{\text{кр}}_{\text{лев}.\alpha/2} = -u_\gamma, \\ \to x^{\text{кр}}_{\text{пр}.\alpha/2} = u_\gamma. \end{cases} \tag{10.27}$$

Критическая область изображена на рис. 10.2, *в*.

Этап 5. Находим числовое значение $\varphi_{\text{чис}}$ критерия (10.9). Если $\varphi_{\text{чис}}$ попадет в интервал $(-\infty, x^{\text{кр}}_{\text{лев}.\alpha/2})$ или в интервал $(x^{\text{кр}}_{\text{пр}.\alpha/2}, +\infty)$, то гипотезу H_0 (10.22) отвергаем и принимаем гипотезу H_1 (10.23). Поступая таким образом, можно с вероятностью α допустить ошибку первого рода (на самом деле гипотеза H_0 (10.22) верна, а мы ее отвергли). Если

$\varphi_{\text{чнс}}$ попадает в интервал $\left(x^{\text{кр}}_{\text{лев.}\alpha/2},\ x^{\text{кр}}_{\text{пр.}\alpha/2}\right)$, гипотезу H_0 (10.22) принимаем.

Задача 10.1. По результатам $n=9$ замеров установлено, что среднее время изготовления детали $\bar{X}=48$ с. Предполагая, что время изготовления — нормально распределенная случайная величина с дисперсией $\sigma^2=9$ с^2, на уровне значимости $\alpha=0{,}05$ решить:

а) можно ли принять 50 с в качестве нормативного времени (математического ожидания) изготовления детали?

б) можно ли принять за норматив 49 с?

Каковы вероятности того, что принятые решения будут ошибочными?

△ а) По условию, нулевая гипотеза $H_0{:}a=50$ с. Так как в результате выборочных наблюдений получено $\bar{X}=48$ с, то в качестве альтернативной возьмем гипотезу $H_1{:}a=48$ с<50 с.

Таким образом, имеет место случай 2 (см. (10.16), (10.17)), причем $a_0=50$ с, а $a_1=48$ с. По схеме (10.21) получаем $\alpha=0{,}05\rightarrow 1-2\alpha=$
$=0{,}9\xrightarrow{\text{П.2}} u_\gamma=1{,}65\rightarrow x^{\text{кр}}_{\text{лев.}\alpha}=-1{,}65.$

Подставляя в (10.9) исходные данные $\bar{X}=48$, $a_0=50$, $\sigma=3$, $n=9$, получим $\varphi_{\text{чнс}}=\dfrac{48-50}{3/\sqrt{9}}=-2.$

Так как число -2 попадает в критическую область $(-\infty,\ -1{,}65)$, то гипотезу $H_0{:}a=50$ с отвергаем и принимаем гипотезу $H_1{:}a=48$ с. Или, иначе говоря, мы не можем принять 50 с в качестве нормативного времени изготовления детали; за норматив принимаем 48 с. Поступая таким образом, можно допустить ошибку первого рода (на самом деле в генеральной совокупности среднее время изготовления детали будет равно 50 с, а не 48 с). Вероятность этой ошибки равна $\alpha=0{,}05$.

б) Здесь нулевая гипотеза $H_0{:}a=49$ с; альтернативная $H_1{:}a=48$ с<49 с. Таким образом, опять имеет место случай 2, при этом $a_0=49$, а $a_1=48$.

Так как $\varphi_{\text{чнс}}=\dfrac{49-50}{3/\sqrt{9}}=-1$ и -1 не попадает в критическую область $(-\infty,\ -1{,}65)$, то гипотезу $H_0{:}a=49$ не отвергаем, т. е. за норматив времени изготовления детали берем 49 с. Поступая таким образом, можно допустить ошибку второго рода (на самом деле нормативное время равно 48 с, а не 49 с). Вычислим вероятность $\beta=P_{H_1}(H_0)$ этой ошибки.

Рассуждаем так: гипотезу $H_0{:}a=a_0$ при альтернативной гипотезе $H_1{:}a=a_1<a_0$ мы не отвергаем, только если значение величины $\varphi=\dfrac{\bar{X}-a_0}{\sigma/\sqrt{n}}$ не попадает в интервал $(-\infty,\ x^{\text{кр}}_{\text{лев.}\alpha})$ или (см. (10.21) в интервал $(-\infty,\ -u_{\gamma=1-2\alpha})$, т. е. если $\dfrac{\bar{X}-a_0}{\sigma/\sqrt{n}}>-u_{\gamma=1-2\alpha}$ или $\bar{X}>-u_{\gamma=1-2\alpha}\dfrac{\sigma}{\sqrt{n}}+a_0$. Таким образом, равносильны два следующих события:

1) принята гипотеза H_0;

2) $$\bar{X}>-u_{\gamma=1-2\alpha}\frac{\sigma}{\sqrt{n}}+a_0. \qquad (10.28)$$

С другой стороны, в § 9.7 было доказано, что при проведении в одинаковых условиях независимых наблюдений нормально распределенной величины с параметрами a и σ выборочное среднее $\bar{X}=N(a,\ \sigma/\sqrt{n})$ (см. (9.47)). Поэтому если для генеральной совокупности верна гипотеза $H_1{:}a=a_1$, то $\bar{X}=N(a_1,\ \sigma/\sqrt{n})$. Таким образом, равносильны следующие два события:

1) гипотеза H_1 верна;

2) $$\bar{X}=N(a_1,\ \sigma/\sqrt{n}). \qquad (10.29)$$

Учитывая (10.28) и (10.29), получаем

$$\beta=P_{H_1}(H_0)=P_{\bar{X}=N(a_1,\,\sigma/\sqrt{n})}\left(\bar{X}>-u_{\gamma=1-2\alpha}\frac{\sigma}{\sqrt{n}}+a_0\right)$$

— это вероятность того, что случайная величина $\bar{X}$ больше числа

$$C=-u_{\gamma=1-2\alpha}\frac{\sigma}{\sqrt{n}}+a_0$$

при условии, что $\bar{X}$ является нормально распределенной величиной с параметрами a_1 и $\sigma/\sqrt{n}$. Найдем эту вероятность, используя функцию Лапласа:

$$\beta=P_{\bar{X}=N(a_1,\,\sigma/\sqrt{n})}(\bar{X}>C)=P\left[N(a_1,\ \sigma/\sqrt{n})>C\right]=1-P\left[N(a_1,\ \sigma/\sqrt{n})<C\right]=$$
$$=1-\left[\frac{1}{2}+\frac{1}{2}\Phi\left(u=\frac{C-a_1}{\sigma/\sqrt{n}}\right)\right]=\frac{1}{2}-\frac{1}{2}\Phi\left(u=\frac{C-a_1}{\sigma/\sqrt{n}}\right)=\frac{1}{2}+\frac{1}{2}\Phi\left(u=\frac{a_1-C}{\sigma/\sqrt{n}}\right)=$$
$$=\frac{1}{2}+\frac{1}{2}\Phi\left(u=\frac{a_1+u_{\gamma=1-2\alpha}\sigma/\sqrt{n}-a_0}{\sigma/\sqrt{n}}\right)=\frac{1}{2}+\frac{1}{2}\Phi\left(u=u_{\gamma=1-2\alpha}-\frac{a_0-a_1}{\sigma/\sqrt{n}}\right).$$

Итак,

$$\beta=\frac{1}{2}+\frac{1}{2}\Phi\left(u=u_{\gamma=1-2\alpha}-\frac{a_0-a_1}{\sigma/\sqrt{n}}\right). \qquad (10.30)$$

В полученную формулу подставим исходные данные: $a_0=49$, $a_1=48$, $\sigma=3$, $n=9$ и найденное выше при $\alpha=0{,}05$ значение $u_{\gamma=1-2\alpha}=1{,}65$. В результате получим

$$\beta=P_{H_1}(H_0)=\frac{1}{2}+\frac{1}{2}\Phi\left(u=1{,}65-\frac{49-48}{3/\sqrt{9}}\right)=$$
$$=\frac{1}{2}+\frac{1}{2}\Phi(u=0{,}65)\underset{(\text{П.2})}{=\!=}\frac{1}{2}+\frac{1}{2}\cdot 0{,}4843=0{,}7422.$$

Итак, принимая 49 с за нормативное время изготовления детали (принимая гипотезу $H_0{:}a=49$), можно допустить ошибку второго рода (в действительности нормативное время равно 48 с, т. е. верна гипотеза $H_1{:}a=48$). Вероятность появления такой ошибки

$$\beta=P_{H_1}(H_0)=0{,}7422.\ ▲$$

В заключение проанализируем формулу (10.30):

— уменьшим α, не изменяя значении остальных переменных; тогда $\gamma=1-2\alpha$ и $u_{\gamma=1-2\alpha}$ увеличатся (см. табл. П.2); увеличатся аргумент u и значение $\Phi(u)$; следовательно, и вероятность $\beta=P_{H_1}(H_0)$ также увеличится;

— увеличим объем выборки n, не изменяя значения остальных переменных; тогда u и $\Phi(u)$ уменьшатся; следовательно, и вероятность ошибки второго рода также уменьшится.

Таким образом, можно сделать следующие выводы, справедливость которых не ограничивается рамками рассмотренного примера:

при уменьшении вероятности α ошибки первого рода (при фиксированном объеме выборки n) вероятность β ошибки второго рода увеличивается;

при увеличении объема выборки n (при фиксированной вероятности α ошибки первого рода) вероятность β ошибки второго рода уменьшается.

Проверка гипотезы о числовом значении математического ожидания нормального распределения при неизвестной

дисперсии. Выше подробно рассматривался 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 дисперсия σ^2 неизвестна.

По результатам $X_1, X_2, \ldots, X_n$ независимых проведенных в одинаковых условиях наблюдений случайной величины X найдем $\bar{X}=\sum\limits_{i=1}^{n} X_i/n$ и $s^2=\sum\limits_{i=1}^{n}(X_i-\bar{X})^2/(n-1)$.

При неизвестном числовом значении дисперсии σ^2 в основу проверки гипотезы

$$H_0: a=a_0, \tag{10.31}$$

где a_0 — заранее заданное число, положен критерий

$$\varphi=\frac{\bar{X}-a_0}{s/\sqrt{n}}, \tag{10.32}$$

который при выполнении гипотезы (10.31) имеет t-распределение с числом степеней свободы $k=n-1$ (см. (9.56)), т. е.

$$\varphi=\frac{\bar{X}-a_0}{s/\sqrt{n}}=t(k=n-1). \tag{10.33}$$

Теперь зададимся уровнем значимости α, построим критическую область и проверим гипотезу (10.31) для альтернативной гипотезы следующих трех видов:

1) $H_1: a>a_0$. (10.34)

Критическая область является правосторонней: ее образует интервал $(x^{кр}_{пр.\alpha};\ +\infty)$, где точка $x^{кр}_{пр.\alpha}$ определяется из условия $P(\varphi>x^{кр}_{пр.\alpha})=\alpha$, которое, учитывая (10.33), запишем в виде

$$P(t(k=n-1)>x^{кр}_{пр.\alpha})=\alpha. \tag{10.35}$$

Исходя из этого равенства найдем точку $x^{кр}_{пр.\alpha}$.

Воспользуемся табл. П.3. Полагая $\gamma=1-2\alpha$, найдем при $k=n-1$ число t_γ такое, при котором

$$P(|t(k=n-1)|<t_\gamma)=\gamma,$$

или

$$P(-t_\gamma<t(k=n-1)<t_\gamma)=1-2\alpha.$$

Но тогда

$$P(t(k=n-1)<-t_\gamma \text{ или } t(k=n-1)>t_\gamma)=2\alpha. \tag{10.36}$$

Так как кривая функции плотности t-распределения симметрична, т. е.

$$P(t(k=n-1)<-t_\gamma)=P(t(k=n-1)>t_\gamma),$$

то из равенства (10.36) следует, что

$$P(t(k=n-1)>t_\gamma)=\alpha. \tag{10.37}$$

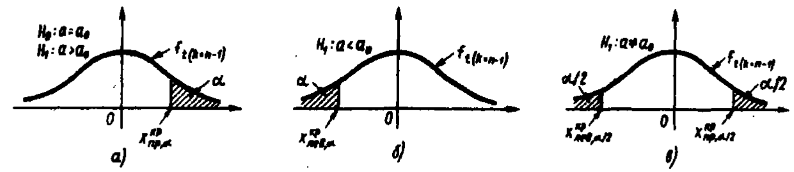


Рис. 10.3

Сопоставляя (10.37) с (10.35), заключаем, что $x^{кр}_{пр.\alpha}=t_{\gamma=1-2\alpha}$.

Итак, схема нахождения точки $x^{кр}_{пр.\alpha}$ такая:

$$\left.\begin{array}{l}\alpha\to\gamma=1-2\alpha\\ n\to k=n-1\end{array}\right\}\xrightarrow{\text{П.3}} t_\gamma \to x^{кр}_{пр.\alpha}=t_\gamma. \tag{10.38}$$

Критическая область изображена на рис. 10.3, *а*.

Теперь подставим в формулу (10.32) числовые значения средней $\bar{X}$, оценки s, a_0 и n. В результате находим число $\varphi_{чис}$. Если $\varphi_{чис}>x^{кр}_{пр.\alpha}$, то гипотезу H_0 (10.31) отвергаем и принимаем гипотезу H_1 (10.34). Если $\varphi_{чис}<x^{кр}_{пр.\alpha}$, то принимаем гипотезу H_0 (10.31).

2) $H_1: a<a_0$. (10.39)

Критическая область является левосторонней: ее образует интервал $(-\infty,\ x^{кр}_{лев.\alpha})$, где точка $x^{кр}_{лев.\alpha}$ определяется из условия $P(\varphi<x^{кр}_{лев.\alpha})=\alpha$, которое, учитывая (10.33), запишем в виде

$$P(t(k=n-1)<x^{кр}_{лев.\alpha})=\alpha. \tag{10.40}$$

Исходя из этого равенства найдем точку $x^{кр}_{лев.\alpha}$.

Воспользуемся табл. П.3. Полагая $\gamma=1-2\alpha$, найдем при $k=n-1$ число t_γ такое, при котором

$$P(|t(k=n-1)|<t_\gamma)=\gamma,$$

или

$$P(-t_\gamma<t(k=n-1)<t_\gamma)=1-2\alpha.$$

Но тогда

$$P(t(k=n-1)<-t_\gamma \text{ или } t(k=n-1)>t_\gamma)=2\alpha. \tag{10.41}$$

Так как кривая функции плотности t-распределения симметрична, т. е.

$$P(t(k=n-1)<-t_\gamma)=P(t(k=n-1)>t_\gamma),$$

то из равенства (10.41) получаем

$$P(t(k=n-1)<-t_\gamma)=\alpha. \tag{10.42}$$

Сопоставляя (10.42) с (10.40), заключаем, что $x^{кр}_{лев.\alpha}=-t_{\gamma=1-2\alpha}$.

Итак, схема нахождения точки $x^{кр}_{лев.\alpha}$ такая:

$$\left.\begin{array}{l}\alpha\to\gamma=1-2\alpha\\ n\to k=n-1\end{array}\right\}\xrightarrow{\text{П.3}} t_\gamma \to x^{кр}_{лев.\alpha}=-t_\gamma. \tag{10.43}$$

Критическая область изображена на рис. 10.3, *б*.

Теперь находим числовое значение $\varphi_{\text{чис}}$ критерия (10.32). Если $\varphi_{\text{чис}} < x^{\text{кр}}_{\text{лев.}\,\alpha}$, то принимаем гипотезу H_1 (10.39); в противном случае принимаем гипотезу H_0 (10.31).

3) $H_1: a \neq a_0$. (10.44)

В этом случае критическая область состоит из двух интервалов: $\left(-\infty,\ x^{\text{кр}}_{\text{лев.}\,\alpha/2}\right)$, и $\left(x^{\text{кр}}_{\text{пр.}\,\alpha/2},\ +\infty\right)$, где критические точки определяются из условия

$$P\left(\varphi < x^{\text{кр}}_{\text{лев.}\,\alpha/2}\right) = \alpha/2 \quad \text{и} \quad P\left(\varphi > x^{\text{кр}}_{\text{пр.}\,\alpha/2}\right) = \alpha/2,$$

которое, учитывая (10.33), запишется в виде

$$\begin{aligned} &P\left(t(k=n-1) < x^{\text{кр}}_{\text{лев.}\,\alpha/2}\right) = \alpha/2 \quad \text{и} \\ &P\left(t(k=n-1) > x^{\text{кр}}_{\text{пр.}\,\alpha/2}\right) = \alpha/2. \end{aligned} \tag{10.45}$$

Для нахождения критических точек воспользуется табл. П. 3. Полагая $\gamma = 1 - \alpha$, при $k = n - 1$ найдем число t_γ такое, при котором

$$P(|t(k=n-1)| < t_\gamma) = \gamma,$$

или

$$P(-t_\gamma < t(k=n-1) < t_\gamma) = 1 - \alpha.$$

Но тогда

$$P\left(t(k=n-1) < -t_\gamma \quad \text{или} \quad t(k=n-1) > t_\gamma\right) = \alpha. \tag{10.46}$$

Так как в силу симметричности функции плотности t-распределения

$$P\left(t(k=n-1) < -t_\gamma\right) = P\left(t(k=n-1) > t_\gamma\right),$$

то из (10.46) получим

$$P\left(t(k=n-1) < -t_\gamma\right) = \alpha/2 \quad \text{и} \quad P\left(t(k=n-1) > t_\gamma\right) = \alpha/2. \tag{10.47}$$

Сопоставляя (10.47) с (10.45), заключаем, что

$$x^{\text{кр}}_{\text{лев.}\,\alpha/2} = -t_{\gamma = 1-\alpha}, \quad x^{\text{кр}}_{\text{пр.}\,\alpha/2} = t_{\gamma = 1-\alpha}.$$

Итак, схема нахождения двусторонних критических точек такая:

$$\left.\begin{array}{l} \alpha \to \gamma = 1 - \alpha \\ n \to k = n - 1 \end{array}\right\} \xrightarrow{\text{П.3}} t_\gamma \begin{array}{l} \to x^{\text{кр}}_{\text{лев.}\,\alpha/2} = -t_\gamma \\ \to x^{\text{кр}}_{\text{пр.}\,\alpha/2} = t_\gamma \end{array} \tag{10.48}$$

Критическая область изображена на рис. 10.3, *в*.

Теперь находим числовое значение $\varphi_{\text{чис}}$ критерия (10.32). Если $\varphi_{\text{чис}}$ попадет в интервал $(-\infty,\ x^{\text{кр}}_{\text{лев.}\,\alpha/2})$ или в интервал $(x^{\text{кр}}_{\text{пр.}\,\alpha/2},\ +\infty)$, то гипотезу H_0 (10.31) отвергаем и принимаем гипотезу H_1 (10.44). Если $\varphi_{\text{чис}}$ попадет в интервал $(x^{\text{кр}}_{\text{лев.}\,\alpha/2},\ x^{\text{кр}}_{\text{пр.}\,\alpha/2})$, то гипотезу (10.31) не отвергаем.

Задача 10.2. Хронометраж затрат времени на сборку узла машины $n = 20$ слесарей показал, что среднее время сборки $\bar{X} = 77$ мин, а $s^2 = 4$ мин. В предположении о нормальности распределения решить вопрос о том, можно ли на уровне значимости $\alpha = 0{,}01$ считать 80 мин нормативом (математическим ожиданием) трудоемкости.

△ Итак, гипотеза $H_0: a = 80$ мин. За альтернативную примем гипотезу $H_1: a \neq 80$ мин, т. е. имеет место случай 3, при этом $a_0 = 80$. По схеме (10.48) найдем

$$\left.\begin{array}{l} \alpha = 0{,}01 \to \gamma = 1 - \alpha = 0{,}99 \\ n = 20 \to k = n - 1 = 19 \end{array}\right\} \xrightarrow{\text{П.3}} t_\gamma = 2{,}861 \begin{array}{l} \to x^{\text{кр}}_{\text{лев.}\,\alpha/2} = -2{,}861, \\ \to x^{\text{кр}}_{\text{пр.}\,\alpha/2} = 2{,}861. \end{array}$$

Таким образом, критическая область состоит из интервалов $(-\infty;\ -2{,}861)$ и $(2{,}861;\ +\infty)$. Теперь по формуле (10.32) найдем $\varphi_{\text{чис}} = \dfrac{77 - 80}{2/\sqrt{20}} = -6{,}708$. Так как число $(-6{,}708)$ попадает в критическую область, гипотезу H_0 отвергаем. ▲

Проверка гипотезы о числовом значении дисперсии нормального распределения. Итак, известно, что случайная величина X имеет нормальный закон распределения, т. е. $X = N(a,\ \sigma)$, но числовое значение дисперсии σ^2 неизвестно.

Пусть $X_1, X_2, \ldots, X_n$ — результаты независимых наблюдений величины X, проводимых в одинаковых условиях. Найдем точечную оценку $s^2 = \sum\limits_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$ неизвестной дисперсии σ^2, где $\bar{X} = \sum\limits_{i=1}^{n} X_i/n$.

Проверим гипотезу:

$$H_0: \sigma^2 = \sigma_0^2, \tag{10.49}$$

где σ_0^2 — заранее заданное число.

В основе проверки этой гипотезы лежит сравнение σ_0^2 и s^2. Критерий имеет вид

$$\varphi = \frac{(n-1)s^2}{\sigma_0^2}. \tag{10.50}$$

При выполнении гипотезы (10.49) величина (10.50) имеет χ^2-распределение с числом степеней свободы $k = n - 1$ (см. (9.64)), т. е.

$$\varphi = \frac{(n-1)s^2}{\sigma_0^2} = \chi^2(k = n - 1). \tag{10.51}$$

Теперь зададимся уровнем значимости α и перейдем к построению критической области и проверке гипотезы H_0 (10.49) для трех видов альтернативной гипотезы H_1.

1) $H_1: \sigma^2 > \sigma_0^2$. (10.52)

В этом случае критическая область имеет вид $(x^{\text{кр}}_{\text{пр.}\,\alpha},\ +\infty)$, где точка $x^{\text{кр}}_{\text{пр.}\,\alpha}$ определяется из условия $P(\varphi > x^{\text{кр}}_{\text{пр.}\,\alpha}) = \alpha$, которое, учитывая (10.51), запишем в виде

$$P\left(\chi^2(k = n - 1) > x^{\text{кр}}_{\text{пр.}\,\alpha}\right) = \alpha. \tag{10.53}$$

Исходя из этого равенства найдем $x^{\text{кр}}_{\text{пр.}\,\alpha}$.

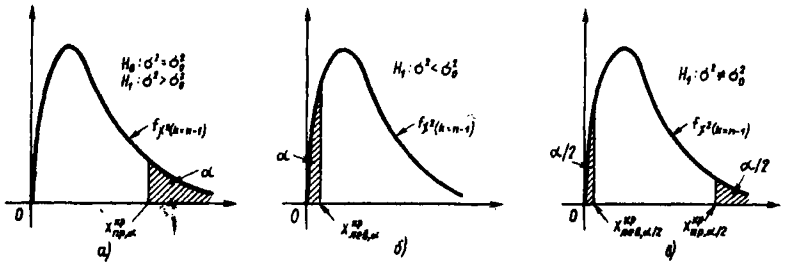


Рис. 10.4

Воспользуемся табл. П.6.

Полагая $\gamma=1-\alpha$, найдем при $k=n-1$ число χ_γ^2 такое, при котором

$$P(\chi^2(k=n-1)<\chi_\gamma^2)=\gamma=1-\alpha.$$

Тогда

$$P(\chi^2(k=n-1)>\chi_\gamma^2)=\alpha. \qquad (10.54)$$

Сопоставляя (10.54) с (10.53), заключаем, что $x^{\text{кр}}_{\text{пр},\alpha}=\chi^2_{\gamma=1-\alpha}$.

Итак, схема нахождения точки $x^{\text{кр}}_{\text{пр},\alpha}$ такая:

$$\left.\begin{array}{l}\alpha\to\gamma=1-\alpha\\ n\to k=n-1\end{array}\right\}\xrightarrow{\text{П.6}}\chi_\gamma^2\to x^{\text{кр}}_{\text{пр},\alpha}=\chi_\gamma^2. \qquad (10.55)$$

Критическая область изображена на рис. 10.4, *а*.

Подставив в (10.50) вместо n, s^2 и σ_0^2 числа, найдем $\varphi_{\text{чис}}$. Если $\varphi_{\text{чис}}>x^{\text{кр}}_{\text{пр},\alpha}$, то гипотезу H_0 (10.49) отвергаем и принимаем гипотезу H_1 (10.52); если $\varphi_{\text{чис}}<x^{\text{кр}}_{\text{пр},\alpha}$, то принимаем гипотезу H_0 (10.49).

2) $H_1:\sigma^2<\sigma_0^2$. (10.56)

В этом случае критическая область является левосторонней, причем, поскольку критерий (10.50) неотрицателен, она имеет вид $(0,\ x^{\text{кр}}_{\text{лев},\alpha})$, где точка $x^{\text{кр}}_{\text{лев},\alpha}$ находится из условия $P(\varphi<x^{\text{кр}}_{\text{лев},\alpha})=\alpha$, которое, учитывая (10.51), запишется в виде

$$P(\chi^2(k=n-1)<x^{\text{кр}}_{\text{лев},\alpha})=\alpha. \qquad (10.57)$$

Воспользуемся табл. П.6. Полагая $\gamma=\alpha$, найдем при $k=n-1$ число χ_γ^2 такое, при котором

$$P(\chi^2(k=n-1)<\chi_\gamma^2)=\gamma=\alpha. \qquad (10.58)$$

Сопоставляя (10.58) с (10.57), заключаем, что $x^{\text{кр}}_{\text{лев},\alpha}=\chi^2_{\gamma=\alpha}$.

Итак, схема нахождения точки $x^{\text{кр}}_{\text{лев},\alpha}$ такая:

$$\left.\begin{array}{l}\alpha\to\gamma=\alpha\\ n\to k=n-1\end{array}\right\}\xrightarrow{\text{П.6}}\chi_\gamma^2\to x^{\text{кр}}_{\text{лев},\alpha}=\chi_\gamma^2. \qquad (10.59)$$

Критическая область изображена на рис. 10.4, *б*.

Если $\varphi_{\text{чис}}$, найденное по формуле (10.50), попадет в интервал $(0,\ x^{\text{кр}}_{\text{лев},\alpha})$, то принимаем гипотезу H_1 (10.56); в противном случае принимаем гипотезу H_0 (10.49).

3) $H_1:\sigma^2\neq\sigma_0^2$. (10.60)

В этом случае критическая область состоит из двух интервалов $(0,\ x^{\text{кр}}_{\text{лев},\alpha/2})$ и $(x^{\text{кр}}_{\text{пр},\alpha/2},\ +\infty)$, где критические точки определяются из условий

$$P(\varphi<x^{\text{кр}}_{\text{лев},\alpha/2})=\alpha/2 \ \text{ и } \ P(\varphi>x^{\text{кр}}_{\text{пр},\alpha/2})=\alpha/2,$$

которые, учитывая (10.51), запишем в виде

$$P(\chi^2(k=n-1)<x^{\text{кр}}_{\text{лев},\alpha/2})=\alpha/2, \qquad (10.61)$$

$$P(\chi^2(k=n-1)>x^{\text{кр}}_{\text{пр},\alpha/2})=\alpha/2. \qquad (10.62)$$

Сравним равенства (10.61) и (10.58); (10.62) и (10.53). Их отличие состоит в замене вероятности α и $\alpha/2$. Поэтому схема нахождения точки $x^{\text{кр}}_{\text{лев},\alpha/2}$ аналогична схеме (10.59), а именно:

$$\left.\begin{array}{l}\alpha\to\gamma=\alpha/2\\ n\to k=n-1\end{array}\right\}\xrightarrow{\text{П.6}}\chi_\gamma^2\to x^{\text{кр}}_{\text{лев},\alpha/2}=\chi_\gamma^2, \qquad (10.63)$$

а схема нахождения точки $x^{\text{кр}}_{\text{пр},\alpha/2}$ подобна схеме (10.55), если в ней α заменить на $\alpha/2$:

$$\left.\begin{array}{l}\alpha\to\gamma=1-\alpha/2\\ n\to k=n-1\end{array}\right\}\xrightarrow{\text{П.6}}\chi_\gamma^2\to x^{\text{кр}}_{\text{пр},\alpha/2}=\chi_\gamma^2. \qquad (10.64)$$

Критическая область изображена на рис. 10.4, *в*.

Если $\varphi_{\text{чис}}$, найденное по формуле (10.50), попадет в один из интервалов $(0,\ x^{\text{кр}}_{\text{лев},\alpha/2})$ или $(x^{\text{кр}}_{\text{пр},\alpha/2},\ +\infty)$, то гипотезу $H_0:\sigma^2=\sigma_0^2$ отвергаем и принимаем гипотезы (10.60). В противном случае нет оснований отвергнуть гипотезу $H_0:\sigma^2=\sigma_0^2$.

Задача 10.3. 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 $\alpha=0{,}05$ дать ответ на вопрос, обеспечивает ли станок требуемую точность. Предполагается, что размер изделия — нормально распределенная случайная величина.

△ Станок будет обеспечивать 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. имеем случай 1, при этом $\sigma_0^2=0{,}15$. По схеме (10.55) найдем

$$\left.\begin{array}{l}\alpha=0{,}05\to\gamma=1-\alpha=0{,}95\\ n=25\to k=n-1=24\end{array}\right\}\xrightarrow{\text{П.6}}\chi_\gamma^2=36{,}4\to x^{\text{кр}}_{\text{пр},\alpha}=36{,}4.$$

Таким образом, критическая область $(36{,}4;\ +\infty)$. Теперь по формуле (10.51) найдем $\varphi_{\text{чис}}=\dfrac{(25-1)\,0{,}25}{0{,}15}=40$. Так как $\varphi_{\text{чис}}$ попадает в критическую область, гипотезу H_0 отвергаем, т. е. считаем, что станок не обеспечивает требуемой точности. ▲

§ 10.3. Проверка гипотезы о равенстве математических ожиданий двух нормальных распределений с известными дисперсиями

Проверка гипотезы о равенстве математических ожиданий двух совокупностей имеет важное практическое значение. Действительно, иногда оказывается, что средний результат $\bar{X}$ одной серии наблюдений отличается от среднего результата $\bar{Y}$ другой серии. Возникает вопрос: можно ли это различие объяснить случайной ошибкой экспериментов или это отличие не случайно? Иначе говоря, можно ли считать, что результаты экспериментов представляют собой выборки из двух генеральных совокупностей с одинаковыми средними или средние этих совокупностей не равны? Подобная задача возникает, например, при сравнении качества изделий, изготовленных на разных установках.

Приведем точную формулировку задачи.

Изучаются две случайные величины: $X=N(a_X,\ \sigma_X)$ и $Y=N(a_Y,\ \sigma_Y)$, числовые значения дисперсий σ_X^2 и σ_Y^2 которых известны; числовые значения математических ожиданий a_X и a_Y неизвестны;

Пусть $X_1,\ X_2,\ ...,\ X_{n_X}$ — результаты независимых, проводимых в одинаковых условиях наблюдений величины X. Число наблюдений равно n_X. Средний результат этих наблюдений $\bar{X}=(X_1+X_2+\ ...\ +X_{n_X})/n_X$. Сразу отметим, что для $\bar{X}$ будет выполняться равенство (9.47):

$$\bar{X}=N(a_X,\ \sigma_X/\sqrt{n_X}).$$

Пусть, далее, $Y_1,\ Y_2,\ ...,\ Y_{n_Y}$ — результаты независимых, проводимых в одинаковых условиях наблюдений величины Y. Число наблюдений равно n_Y. Средний результат этих наблюдений $\bar{Y}=(Y_1+Y_2+...+Y_{n_Y})/n_Y$. Для него выполняется равенство $\bar{Y}=N(a_Y,\ \sigma_Y/\sqrt{n_Y})$.

Кроме того, наблюдения организованы таким образом, что результаты $X_1,\ X_2,\ ...,\ X_{n_X}$ и $Y_1,\ Y_2,\ ...$ $...,\ Y_{n_Y}$ независимы. Из этого условия следует, что $\bar{X}$ и $\bar{Y}$ — независимые случайные величины.

При сформулированных условиях требуется проверить гипотезу о равенстве математических ожиданий случайных величин X и Y, т. е. гипотезу

$$H_0: a_X=a_Y. \tag{10.65}$$

Построим критерий проверки этой гипотезы.

□ Так как приближенное представление о математическом ожидании дает выборочная средняя, то в основе проверки гипотезы (10.65) должно лежать сравнение средних $\bar{X}$ и $\bar{Y}$. Найдем закон распределения разности $(\bar{X}-\bar{Y})$.

Величины $\bar{X}$ и $\bar{Y}$ независимы и имеют нормальный закон распределения, поэтому разность $(\bar{X}-\bar{Y})$ в силу теоремы о суммировании нормально распределенных величин также будет иметь нормальный закон, при этом математическое ожидание

$$M(\bar{X}-\bar{Y})=M\bar{X}-M\bar{Y}=a_X-a_Y,$$

а дисперсия

$$D(\bar{X}-\bar{Y})=D\bar{X}+D\bar{Y}=\sigma_X^2/n_X+\sigma_Y^2/n_Y,$$

т. е.

$$\bar{X}-\bar{Y}=N(a_X-a_Y;\ \sqrt{\sigma_X^2/n_X+\sigma_Y^2/n_Y}).$$

А раз так, то

$$\frac{(\bar{X}-\bar{Y})-(a_X-a_Y)}{\sqrt{\sigma_X^2/n_X+\sigma_Y^2/n_Y}}=N(0,1).$$

Если гипотеза (10.65) выполняется, т. е. $a_X=a_Y$, то из последнего равенства получаем, что величина

$$\varphi=\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_X^2/n_X+\sigma_Y^2/n_Y}} \tag{10.66}$$

имеет нормальный закон распределения с нулевым математическим ожиданием и единичной дисперсией, т. е.

$$\varphi=\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_X^2/n_X+\sigma_Y^2/n_Y}}=N(0,\ 1). \tag{10.67}$$

Величину (10.66) и используют в качестве критерия при проверке гипотезы (10.65). ■

Теперь зададимся уровнем значимости α и перейдем к построению критических областей и проверке гипотезы (10.65) для трех видов альтернативной гипотезы H_1.

1) $H_1: a_X>a_Y$. (10.68)

В этом случае критическая область имеет вид $(x^{кр}_{пр,\,\alpha},\ +\infty)$, где критическая точка $x^{кр}_{пр,\,\alpha}$ определяется из условия $P(\varphi>x^{кр}_{пр,\,\alpha})=\alpha$, которое, учитывая (10.67), запишем в виде $P(N(0,\ 1)>x^{кр}_{пр,\,\alpha})=\alpha$. Обратим внимание на то, что полученное равенство полностью совпадает с равенством (10.11). Поэтому все выкладки, связанные с нахождением критической точки, аналогичны выкладкам (10.11)—(10.15).

Итак, схема нахождения точки $x^{кр}_{пр,\,\alpha}$ имеет вид (10.15); критическая область изображена на рис. 10.2, *а*.

Теперь подставим в формулу (10.66) числовые значения выборочных средних $\bar{X}$ и $\bar{Y}$, дисперсий σ_X^2 и σ_Y^2 и объемов n_X и n_Y выборок. В результате получим $\varphi_{чис}$. Если $\varphi_{чис}>x^{кр}_{пр,\,\alpha}$,

то гипотезу H_0 (10.65) отвергаем и принимаем гипотезу H_1 (10.68). Поступая таким образом, можно допустить ошибку первого рода (на самом деле гипотеза H_0 (10.65) верна, а мы ее отвергли). Вероятность появления этой ошибки равна α.

Если $\varphi_{\text{чис}} < x^{\text{кр}}_{\text{пр},\,\alpha}$, то гипотеза H_0 (10.65) не отвергается.

2) $H_1 : a_X < a_Y$. (10.69)

В этом случае критическая область имеет вид $(-\infty, x^{\text{кр}}_{\text{лев},\,\alpha})$, где критическая точка определяется из условия $P(\varphi < x^{\text{кр}}_{\text{лев},\,\alpha}) = \alpha$, которое, учитывая (10.67), запишем в таком же виде, как (10.18). Поэтому все выкладки, связанные с нахождением точки $x^{\text{кр}}_{\text{лев},\,\alpha}$, аналогичны выкладкам (10.18)—(10.21).

Итак, схема нахождения точки $x^{\text{кр}}_{\text{лев},\,\alpha}$ имеет вид (10.21); критическая область изображена на рис. 10.2, *б*.

Теперь найдем числовое значение $\varphi_{\text{чис}}$ критерия (10.66). Если оно попадет в критическую область $(-\infty, x^{\text{кр}}_{\text{лев},\,\alpha})$, то принимаем гипотезу H_1 (10.69); в противном случае—гипотезу H_0 (10.65).

3) $H_1 : a_X \neq a_Y$. (10.70)

В этом случае критическая область состоит из двух интервалов $(-\infty, x^{\text{кр}}_{\text{лев},\,\alpha/2})$ и $(x^{\text{кр}}_{\text{пр},\,\alpha/2}, +\infty)$, где критические точки определяются соответственно из условий

$$P(\varphi < x^{\text{кр}}_{\text{лев},\,\alpha/2}) = \alpha/2 \quad \text{и} \quad P(\varphi > x^{\text{кр}}_{\text{пр},\,\alpha/2}) = \alpha/2,$$

которые с учетом (10.67) запишем в таком же виде, как (10.24). Поэтому все выкладки, связанные с нахождением критических точек, аналогичны выкладкам (10.24)—(10.27).

Итак, схема нахождения критических точек $x^{\text{кр}}_{\text{лев},\,\alpha/2}$ и $x^{\text{кр}}_{\text{пр},\,\alpha/2}$ имеет вид (10.27). Критическая область изображена на рис. 10.2, *в*.

Теперь найдем числовое значение $\varphi_{\text{чис}}$ критерия (10.66). Если $\varphi_{\text{чис}}$ попадает в интервал $(-\infty, x^{\text{кр}}_{\text{лев},\,\alpha/2})$ или в интервал $(x^{\text{кр}}_{\text{пр},\,\alpha/2}, +\infty)$, то принимаем гипотезу H_1 (10.70). Если $\varphi_{\text{чис}} \in (x^{\text{кр}}_{\text{лев},\,\alpha/2}, x^{\text{кр}}_{\text{пр},\,\alpha/2})$, то принимаем гипотезу $H_0 : a_X = a_Y$.

Задача 10.4. 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 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?

$\triangle$ Обозначим через a_X и a_Y математические ожидания размера диаметра валиков, изготовленных соответственно на автоматах 1 и 2.

Напомним, что ошибка ε называется случайной, если математическое ожидание $M\varepsilon = 0$. Поэтому выражение «различие выборочных средних вызвано случайной ошибкой» означает, что $\bar{X} - \bar{Y} = \varepsilon$. Применяя к обеим частям этого равенства операцию математического ожидания, получаем $M\bar{X} - M\bar{Y} = 0$ или, учитывая, что $M\bar{X} = a_X$, а $M\bar{Y} = a_Y$, имеем $a_X = a_Y$. Таким образом, в примере требуется решить вопрос о том, можно принять гипотезу $H_0 : a_X = a_Y$ или нет. В качестве конкурирующей возьмем гипотезу $H_1 : a_X \neq a_Y$, т. е. имеет место случай 3. Двусторонние критические точки найдем по схеме (10.27):

$$\alpha = 0{,}05 \to \gamma = 1 - \alpha = 0{,}95 \xrightarrow{\text{п.2}} u_\gamma = 1{,}96 \to \begin{cases} x^{\text{кр}}_{\text{лев},\,\alpha/2} = -1{,}96, \\ x^{\text{кр}}_{\text{пр},\,\alpha/2} = 1{,}96. \end{cases}$$

Итак, критическая область состоит из интервалов $(-\infty; -1{,}96)$ и $(1{,}96; +\infty)$.

Теперь по формуле (10.66) найдем $\varphi_{\text{чис}} = \dfrac{182 - 185}{\sqrt{5/14 + 7/9}} = -2{,}57$.

Так как $\varphi_{\text{чис}} < -1{,}96$ и попадает в критическую область, то гипотезу $H_0 : a_X = a_Y$ отвергаем, т. е. считаем, что различие выборочных средних неслучайно. $\blacktriangle$

§ 10.4. Проверка гипотезы о равенстве математических ожиданий двух нормальных распределений с неизвестными, но равными дисперсиями

Сформулируем задачу:

Изучаются две случайные величины $X = N(a_X, \sigma)$ и $Y = N(a_Y, \sigma)$ с одинаковыми дисперсиями, равными σ^2, однако числовое значение σ^2 неизвестно; неизвестны также и числовые значения математических ожиданий a_X и a_Y.

Пусть $X_1, X_2, \ldots, X_{n_X}$—результаты независимых, проводимых в одинаковых условиях наблюдений величины X. Число наблюдений равно n_X. Средний результат этих наблюдений $\bar{X} = \sum_{i=1}^{n_X} X_i / n_X$; за приближенное значение дисперсии σ^2 величины X возьмем оценку $s_X^2 = \sum_{i=1}^{n_X} (X_i - \bar{X})^2 / (n_X - 1)$. Напомним, что для $\bar{X}$ выполняется равенство

$$\bar{X} = N(a_X, \sigma/\sqrt{n_X}), \tag{10.71}$$

а величина $\dfrac{(n_X-1)s_X^2}{\sigma^2}$ имеет χ^2-распределение с $k = n_X - 1$ степенями свободы, т. е.

$$\frac{(n_X-1)s_X^2}{\sigma^2} = \chi^2(k = n_X - 1). \tag{10.72}$$

Пусть, далее, $Y_1, Y_2, \ldots, Y_{n_Y}$—результаты независимых, проводимых в одинаковых условиях наблюдений величины Y. Число наблюдений равно n_Y. Средний результат наблюдений $\bar{Y} = \sum_{i=1}^{n_Y} Y_i / n_Y$; приближенное значение дисперсии σ^2 величины Y равно $s_Y^2 = \sum_{i=1}^{n_Y} (Y_i - \bar{Y})^2 / (n_Y -$

$-1)$. Аналогично (10.71) и (10.72), имеют место следующие соотношения:

$$\bar{Y}=N\left(a_Y, \sigma/\sqrt{n_Y}\right); \tag{10.73}$$

$$\frac{(n_Y-1)s_Y^2}{\sigma^2}=\chi^2\left(k=n_Y-1\right). \tag{10.74}$$

Наблюдения организованы так, что результаты $X_1, X_2, ..., X_{n_X}$ и $Y_1, Y_2, ..., Y_{n_Y}$ независимы. Из этого условия следует, что $\bar{X}$ и $\bar{Y}$ — независимые величины, s_X^2 и s_Y^2 также независимые величины.

При сфóрмулированных условиях требуется проверить гипотезу о равенстве математических ожиданий случайных величин X и Y, т. е. гипотезу

$$H_0: a_X=a_Y. \tag{10.75}$$

Построим критерий проверки этой гипотезы.

□ В основе проверки гипотезы лежит сравнение средних $\bar{X}$ и $\bar{Y}$. Найдем закон распределения разности $(\bar{X}-\bar{Y})$.

Поскольку величины $\bar{X}$ и $\bar{Y}$ независимы и имеют нормальный закон, разность $(\bar{X}-\bar{Y})$ также имеет нормальный закон, при этом

$$M(\bar{X}-\bar{Y})=M\bar{X}-M\bar{Y}=a_X-a_Y,$$

а дисперсия

$$D(\bar{X}-\bar{Y})=D\bar{X}+D\bar{Y}=\sigma^2/n_X+\sigma^2/n_Y,$$

т. е.

$$\bar{X}-\bar{Y}=N\left(a_X-a_Y;\ \sqrt{\sigma^2/n_X+\sigma^2/n_Y}\right).$$

В этом случае

$$\frac{(\bar{X}-\bar{Y})-(a_X-a_Y)}{\sqrt{\sigma^2/n_X+\sigma^2/n_Y}}=N(0,1)=U. \tag{10.76}$$

Далее в силу независимости величин s_X^2 и s_Y^2 величины $\chi^2(k=n_X-1)$ и $\chi^2(k=n_Y-1)$, заданные формулами (10.72) и (10.74), также будут независимыми. Поэтому, учитывая (6.35), имеем, что сумма $\chi^2(k=n_X-1)+\chi^2(k=n_Y-1)$ имеет χ^2-распределение с числом степеней свободы $k=n_X+n_Y-2$, т. е.

$$\frac{(n_X-1)s_X^2}{\sigma^2}+\frac{(n_Y-1)s_Y^2}{\sigma^2}=\chi^2\left(k=n_X+n_Y-2\right). \tag{10.77}$$

Убедимся в независимости величин (10.76) и (10.77).

Приведем без доказательства следующую теорему.

Теорема о независимости случайных величин $\bar{X}$ и s_X^2. *Пусть $X_1, X_2, ..., X_n$ — результаты независимых наблюдений нормально распределенной случайной величины, проводимых в одинаковых условиях. Тогда случайные величины $\bar{X}=\sum_{i=1}^{n} X_i/n$ и $s_X^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)$ независимы.*

Используя эту теорему, докажем независимость величин (10.76) и (10.77). В выражении (10.76) случайными величинами являются $\bar{X}$ и $\bar{Y}$, а в выражении (10.77) — величины s_X^2 и s_Y^2. В силу теоремы $\bar{X}$ и s_X^2 независимы, также независимы $\bar{Y}$ и s_Y^2. С другой стороны, независимыми являются $\bar{X}$ и s_Y^2, так как $\bar{X}$ вычисляется по ряду $X_1, X_2, ..., X_{n_X}$, а s_Y^2 — по ряду $Y_1, Y_2, ..., Y_{n_Y}$, а по условию эти ряды независимы; также независимы $\bar{Y}$ и s_X^2.

Итак, $\bar{X}$ и s_X^2, $\bar{X}$ и s_Y^2 независимы, поэтому независимы $\bar{X}$ и величина (10.77). Аналогично, независимы $\bar{Y}$ и s_Y^2 и $\bar{Y}$ и s_X^2, поэтому независимы $\bar{Y}$ и величина (10.77). Отсюда и следует независимость величины (10.76), в которую входит разность $(\bar{X}-\bar{Y})$, и величины (10.77).

Вспоминая способ получения распределения Стьюдента и подставляя в выражение (6.36) величины (10.76) и (10.77), получим

$$t\left(k=n_X+n_Y-2\right)=\frac{N(0,1)}{\sqrt{\frac{1}{n_X+n_Y-2}\chi^2(k=n_X+n_Y-2)}}=$$

$$=\frac{(\bar{X}-\bar{Y})-(a_X-a_Y)}{\sqrt{\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)\frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{n_X+n_Y-2}}}.$$

Если гипотеза (10.75) выполняется, т. е. $a_X=a_Y$, то из последнего равенства следует, что величина

$$\varphi=\frac{\bar{X}-\bar{Y}}{\sqrt{\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)\frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{n_X+n_Y-2}}} \tag{10.78}$$

имеет t-распределение с $k=n_X+n_Y-2$ степенями свободы, т. е.

$$\varphi=\frac{\bar{X}-\bar{Y}}{\sqrt{\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)\frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{n_X+n_Y-2}}}=t\left(k=n_X+n_Y-2\right). \tag{10.79}$$

Величину (10.78) и используют в качестве критерия при проверке гипотезы (10.75). ■

Зададимся уровнем значимости α и перейдем к построению критических областей и проверке гипотезы (10.75) для трех видов альтернативной гипотезы H_1. При этом обратим внимание на то, что уже ранее рассматривался критерий, имеющий t-распределение — это критерий (10.33), который имеет t-распределение с $k=n-1$ степенями свободы. Сейчас мы рассматриваем критерий (10.78), имеющий t-распределение с $k=n_X+n_Y-2$ степенями свободы. Никаких принципиальных отличий в алгоритмы построения критических областей это

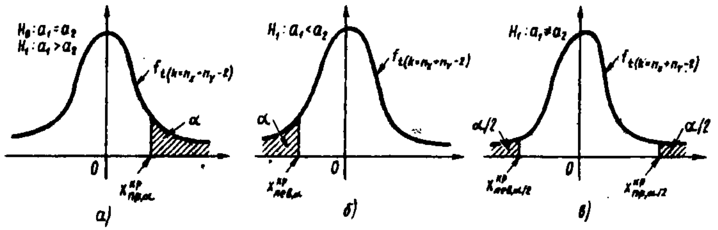


Рис. 10.5

не вносит (кроме изменения числа степеней свободы). Поэтому приведем лишь схемы нахождения критических точек.

1) $H_1: a_X > a_Y$. (10.80)

Критическая область имеет вид $(x^{кр}_{пр,\alpha}, +\infty)$, где точка $x^{кр}_{пр,\alpha}$ находится по схеме, аналогичной (10.38):

$$\left.\begin{array}{l}\alpha \to \gamma = 1-2\alpha \\ n_X,\ n_Y \to k = n_X + n_Y - 2\end{array}\right\} \xrightarrow{\text{п.3}} t_\gamma \to x^{кр}_{пр,\alpha} = t_\gamma. \qquad (10.81)$$

Критическая область изображена на рис. 10.5, *а*.

Подставляя в (10.78) числовые значения, найдем $\varphi_{чис}$. Если $\varphi_{чис} > x^{кр}_{пр,\alpha}$, то принимаем гипотезу H_1 (10.80); в противном случае принимаем гипотезу $H_0: a_X = a_Y$.

2) $H_1: a_X < a_Y$.

Критическая область имеет вид $(-\infty,\ x^{кр}_{лев,\alpha})$, где точка $x^{кр}_{лев,\alpha}$ находится по схеме, аналогичной (10.43):

$$\left.\begin{array}{l}\alpha \to \gamma = 1-2\alpha \\ n_X,\ n_Y \to k = n_X + n_Y - 2\end{array}\right\} \xrightarrow{\text{п.3}} t_\gamma \to x^{кр}_{лев,\alpha} = -t_\gamma. \qquad (10.82)$$

Критическая область изображена на рис. 10.5, *б*.

Если числовое значение $\varphi_{чис}$ критерия (10.78) принадлежит интервалу $(-\infty,\ x^{кр}_{лев,\alpha})$, то принимаем гипотезу $H_1: a_X < a_Y$; в противном случае принимаем гипотезу $H_0: a_X = a_Y$.

3) $H_1: a_X \neq a_Y$.

Критическая область состоит из двух интервалов $(-\infty,\ x^{кр}_{лев,\alpha/2})$ и $(x^{кр}_{пр,\alpha/2},\ +\infty)$, где критические точки находятся по схеме, аналогичной (10.48):

$$\left.\begin{array}{l}\alpha \to \gamma = 1-\alpha \\ n_X,\ n_Y \to k = n_X + n_Y - 2\end{array}\right\} \xrightarrow{\text{п.3}} t_\gamma \begin{array}{l}\to x^{кр}_{лев,\alpha/2} = -t_\gamma \\ \to x^{кр}_{пр,\alpha/2} = t_\gamma.\end{array} \qquad (10.83)$$

Критическая область изображена на рис. 10.5, *в*.

Если числовое значение $\varphi_{чис}$ критерия (10.78) попадает в интервал $(-\infty,\ x^{кр}_{лев,\alpha/2})$ или в интервал $(x^{кр}_{пр,\alpha/2},\ +\infty)$, то принимаем гипотезу $H_1: a_X \neq a_Y$. Если $\varphi_{чис}$ попадает в интервал $(x^{кр}_{лев,\alpha/2},\ x^{кр}_{пр,\alpha/2})$, то принимаем гипотезу $H_0: a_X = a_Y$.

Замечание. Если гипотеза $H_0: a_X = a_Y$ принимается, то говорят, что различие выборочных средних $\bar{X}$ и $\bar{Y}$ *статистически незначимо*. Напомним, что мы рассматривали случайные величины $X = N(a_X, \sigma)$ и $Y = N(a_Y, \sigma)$ с одинаковыми дисперсиями. Принятие гипотезы $H_0: a_X = a_Y$ означает, что величины X и Y полностью совпадают, или образуют одну генеральную совокупность. Приближенное значение средней этой совокупности, или математического ожидания, в этом случае равно $\dfrac{\bar{X}n_X + \bar{Y}n_Y}{n_X + n_Y}$.

Задача 10.5. Расход сырья на единицу продукции по старой технологии составил:

Расходы 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	$n_Y = 13$

Предположив, что соответствующие генеральные совокупности X и Y имеют нормальные распределения с одинаковыми дисперсиями и математическими ожиданиями a_X и a_Y, проверить при уровне значимости 0,1 гипотезу $H_0: a_X = a_Y$ при альтернативной гипотезе $H_1: a_X \neq a_Y$ (случай 3).

△ По схеме (10.83) найдем

$$\left.\begin{array}{l}\alpha = 0{,}1 \to \gamma = 1-\alpha = 0{,}9 \\ n_X = 9,\ n_Y = 13 \to k = n_X + n_Y - 2 = 20\end{array}\right\} \xrightarrow{\text{п.3}} t_\gamma = 1{,}725 \begin{array}{l}\to x^{кр}_{лев,\alpha/2} = -1{,}725, \\ \to x^{кр}_{пр,\alpha/2} = 1{,}725.\end{array}$$

Критическую область образуют интервалы $(-\infty;\ -1{,}725)$, $(1{,}725;\ +\infty)$. Теперь найдем числовое значение $\varphi_{чис}$ критерия (10.78).

При старой технологии средний расход сырья

$$\bar{X} = \frac{\sum_{i=1}^{3} x_i m_i}{n_X} = \frac{304 \cdot 1 + 307 \cdot 4 + 308 \cdot 4}{9} = 307{,}11;$$

выборочная дисперсия

$$\hat{D}X = \bar{X^2} - (\bar{X})^2 = \frac{\sum_{i=1}^{3} x_i^2 m_i}{n_X} - (307{,}11)^2 = 2{,}114;$$

$$s_X^2 = \frac{\hat{D}X \cdot n_X}{n_X - 1} = 2{,}378.$$

При новой технологии

$$\bar{Y} = \frac{\sum_{i=1}^{4} y_i n_i}{n_Y} = \frac{303 \cdot 2 + 304 \cdot 6 + 306 \cdot 4 + 308 \cdot 1}{13} = 304{,}77;$$

$$\hat{D}Y = \bar{Y^2} - (\bar{Y})^2 = \frac{\sum_{i=1}^{4} y_i^2 n_i}{n_Y} - (304{,}77)^2 = 1{,}555,$$

$$s_Y^2=\frac{\hat{D}Y\cdot n_Y}{n_Y-1}=1{,}685.$$

Тогда

$$\varphi_{\text{чнс}}=\frac{307{,}11-304{,}77}{\sqrt{\frac{(1/9+1/13)(8\cdot 2{,}378+12\cdot 1{,}685)}{9+13-2}}}=3{,}696.$$

Так как $\varphi_{\text{чнс}}$ попадает в критическую область, то гипотезу $H_0: a_X=a_Y$ отвергаем, т. е. при переходе на новую технологию происходит изменение среднего расхода сырья на одно изделие. При этом не надо забывать, что это заключение может оказаться ошибочным (на самом деле $a_X=a_Y$), т. е. имеет место ошибка первого рода, вероятность которой равна $P_{H_0}(H_1)=0{,}1$. ▲

§ 10.5. Проверка гипотезы о равенстве дисперсий двух нормальных распределений

При проверке гипотезы о равенстве математических ожиданий двух совокупностей предполагалось, что дисперсии этих совокупностей одинаковы. Как, не располагая всеми сведениями о генеральных совокупностях, а имея лишь выборки из них, убедиться в приемлемости гипотезы о равенстве генеральных дисперсий? Этот вопрос и рассматривается в настоящем параграфе.

Отметим, что задача проверки гипотезы о равенстве дисперсий имеет и самостоятельное значение. Дисперсия характеризует точность работы приборов, технологических процессов и т. д.; убедившись в равенстве двух дисперсий, мы тем самым убеждаемся, например, в том, что два прибора, два технологических процесса обеспечивают одинаковую точность.

Приведем точную формулировку задачи.

Изучаются две случайные величины $X=N(a_X, \sigma_X)$ и $Y=N(a_Y, \sigma_Y)$ с неизвестными дисперсиями.

Пусть $X_1, X_2, ..., X_{n_X}$ — результаты независимых, проводимых в одинаковых условиях наблюдений величины X. За приближенное значение дисперсии σ_X^2 величины X возьмем $s_X^2=\sum_{i=1}^{n_X}(X_i-\bar{X})^2/(n_X-1)$, где $\bar{X}=\sum_{i=1}^{n_X}X_i/n_X$. Напомним, что имеет место соотношение

$$\frac{(n_X-1)s_X^2}{\sigma_X^2}=\chi^2(l=n_X-1). \qquad (10.84)$$

Пусть, далее, $Y_1, Y_2, ..., Y_{n_Y}$ — результаты независимых, проводимых в одинаковых условиях наблюдений величины Y. За приближенное значение дисперсии σ_Y^2 возьмем $s_Y^2=\sum_{i=1}^{n_Y}(Y_i-\bar{Y})^2/(n_Y-1)$, где $\bar{Y}=\sum_{i=1}^{n}Y_i/n_Y$. Аналогично (10.84) имеет место соотношение

$$\frac{(n_Y-1)s_Y^2}{\sigma_Y^2}=\chi^2(k=n_Y-1). \qquad (10.85)$$

Наблюдения организованы так, что результаты $X_1, X_2, ..., X_{n_X}$ и $Y_1, Y_2, ..., Y_{n_Y}$ независимы.

При сформулированных условиях требуется проверить гипотезу

$$H_0: \sigma_X^2=\sigma_Y^2. \qquad (10.86)$$

Построим критерий для проверки этой гипотезы.

□ Из независимости рядов $X_1, X_2, ..., X_{n_X}$ и $Y_1, Y_2, ..., Y_{n_Y}$ следует независимость величин s_X^2 и s_Y^2, а стало быть, и величин $\chi^2(l=n_X-1)$ и $\chi^2(k=n_Y-1)$, определяемых соотношениями (10.84) и (10.85). Поэтому в соответствии с определением F-распределения (см. (6.37)) отношение $\frac{\chi^2(l)/l}{\chi^2(k)/k}$, или отношение $\frac{\frac{(n_X-1)s_X^2}{\sigma_X^2}/(n_X-1)}{\frac{(n_Y-1)s_Y^2}{\sigma_Y^2}/(n_Y-1)}$ будем иметь F-распределение с $l=n_X-1$ и $k=n_Y-1$ степенями свободы, т. е.

$$\frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}=F(l=n_X-1,\ k=n_Y-1). \qquad (10.87)$$

Если гипотеза (10.86) верна, то из (10.87) получаем, что для

$$\varphi=s_X^2/s_Y^2 \qquad (10.88)$$

справедливо соотношение

$$\varphi=s_X^2/s_Y^2=F(l=n_X-1,\ k=n_Y-1). \qquad (10.89)$$

Величину (10.88) и используют в качестве критерия при проверке гипотезы (10.86). ■

Теперь зададимся уровнем значимости α и перейдем к построению критических областей и проверке гипотезы (10.86) для трех видов альтернативной гипотезы.

1) $H_1: \sigma_X^2>\sigma_Y^2$. (10.90)

В этом случае критическая область будет иметь вид $(x_{\text{пр},\alpha}^{\text{кр}}, +\infty)$, где точка $x_{\text{пр},\alpha}^{\text{кр}}$ определяется из условия $P(\varphi>x_{\text{пр},\alpha}^{\text{кр}})=\alpha$, которое, учитывая (10.89), запишем в виде

$$P(F(l=n_X-1,\ k=n_Y-1)>x_{\text{пр},\alpha}^{\text{кр}})=\alpha. \qquad (10.91)$$

Исходя из этого равенства найдем $x_{\text{пр},\alpha}^{\text{кр}}$.

Воспользуемся табл. П.7. Полагая $\gamma=1-\alpha$, найдем при $l=n_X-1$ и $k=n_Y-1$ число f_γ такое, при котором

$$P(F(l=n_X-1,\ k=n_Y-1)<f_\gamma)=\gamma=1-\alpha.$$

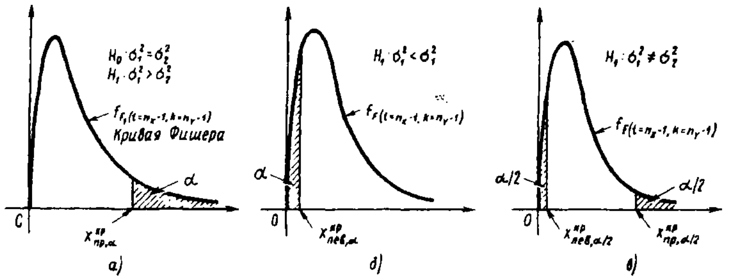


Рис. 10.6

Тогда

$$P\left(F(l=n_X-1,\ k=n_Y-1)>f_\gamma\right)=\alpha. \qquad (10.92)$$

Сравнивая (10.92) с (10.91), заключаем, что $x^{\text{кр}}_{\text{пр},\alpha}=f_{\gamma=1-\alpha}$. Следовательно, схема нахождения точки $x^{\text{кр}}_{\text{пр},\alpha}$ имеет вид

$$\left.\begin{array}{l}\alpha\ \to\gamma=1-\alpha\\ n_X\to l=n_X-1\\ n_Y\to k=n_Y-1\end{array}\right\}\xrightarrow{\text{п.7}} f_\gamma\to x^{\text{кр}}_{\text{пр},\alpha}=f_\gamma. \qquad (10.93)$$

Критическая область изображена на рис. 10.6, *а*.

Теперь подставим в (10.88) найденные по выборке числовые значения величин s_X^2 и s_Y^2. В результате получим $\varphi_{\text{чис}}$. Если $\varphi_{\text{чис}}$ попадает в интервал $(x^{\text{кр}}_{\text{пр},\alpha},\ +\infty)$, то гипотезу H_0 (10.86) отвергаем и принимаем гипотезу H_1 (10.90). Если $\varphi_{\text{чис}}<x^{\text{кр}}_{\text{пр},\alpha}$, то принимаем гипотезу H_0 (10.86).

2) $H_1:\sigma_X^2<\sigma_Y^2$. (10.94)

В этом случае критическая область является левосторонней, причем, поскольку величина (10.88) неотрицательна, область будет иметь вид $(0,\ x^{\text{кр}}_{\text{лев},\alpha})$, где точка $x^{\text{кр}}_{\text{лев},\alpha}$ находится из условия $P(\varphi<x^{\text{кр}}_{\text{лев},\alpha})=\alpha$, которое, учитывая (10.89), запишем в виде

$$P\left(F(l=n_X-1,\ k=n_Y-1)<x^{\text{кр}}_{\text{лев},\alpha}\right)=\alpha.$$

Далее находим $x^{\text{кр}}_{\text{лев},\alpha}$ по следующей схеме:

$$\left.\begin{array}{l}\alpha\to\gamma=1-\alpha\\ n_Y\to l=n_Y-1\\ n_X\to k=n_X-1\end{array}\right\}\xrightarrow{\text{п. 7}} f_\gamma\to x^{\text{кр}}_{\text{лев},\alpha}=\frac{1}{f_\gamma}. \qquad (10.95)$$

Критическая область изображена на рис. 10.6, *б*.

Если $\varphi_{\text{чис}}$, найденное по формуле (10.88), попадет в интервал $(0,\ x^{\text{кр}}_{\text{лев},\alpha})$, то принимаем гипотезу H_1 (10.94), в противном случае принимаем гипотезу $H_0:\sigma_X^2=\sigma_Y^2$.

3) $H_1:\sigma_X^2\neq\sigma_Y^2$.

В этом случае критическая область состоит из двух интервалов: $(0,\ x^{\text{кр}}_{\text{лев},\alpha/2})$ и $(x^{\text{кр}}_{\text{пр},\alpha/2},\ +\infty)$, где точки $x^{\text{кр}}_{\text{лев},\alpha/2}$

и $x^{\text{кр}}_{\text{пр},\alpha/2}$ находятся по схемам (10.93) и (10.95), в которых вероятность α следует заменить на $\alpha/2$. Критическая область изображена на рис. 10.6, *в*.

При попадании числового значения $\varphi_{\text{чис}}$ критерия (10.88) в критическую область принимаем гипотезу $H_1:\sigma_X^2\neq\sigma_Y^2$; в противном случае принимаем гипотезу $H_0:\sigma_X^2=\sigma_Y^2$.

Замечание. Если гипотеза $H_0:\sigma_X^2=\sigma_Y^2$ принимается, то говорят, что различие выборочных оценок s_X^2 и s_Y^2 статистически незначимо и за оценку общего значения дисперсии принимают величину $\dfrac{s_X^2(n_X-1)+s_Y^2(n_Y-1)}{n_X+n_Y-2}$.

Задача 10.6. Можно ли при уровне значимости 0,1 считать статистически незначимым различие между оценками $s_X^2=2{,}378$ и $s_Y^2=1{,}685$, вычисленными по данным задачи 10.5?

△ Итак, требуется проверить гипотезу $H_0:\sigma_X^2=\sigma_Y^2$. В качестве альтернативной возьмем гипотезу $H_1:\sigma_X^2\neq\sigma_Y^2$, т. е. имеет место случай 3.

Критические точки $x^{\text{кр}}_{\text{лев},\alpha/2}$ и $x^{\text{кр}}_{\text{пр},\alpha/2}$ найдем по схемам (10.93) и (10.95), заменив в них α на $\alpha/2$. Имеем

$$\left.\begin{array}{l}\alpha/2=0{,}05\to\gamma=1-\alpha/2=0{,}95\\ n_X=9\to l=n_X-1=8\\ n_Y=13\to k=n_Y-1=12\end{array}\right\}\xrightarrow{\text{п. 7}} f_\gamma=2{,}85\to x^{\text{кр}}_{\text{пр},\alpha/2}=2{,}85;$$

$$\left.\begin{array}{l}\alpha/2=0{,}05\to\gamma=1-\alpha/2=0{,}95\\ n_Y=13\to l=n_Y-1=12\\ n_X=9\to k=n_X-1=8\end{array}\right\}\xrightarrow{\text{п. 7}} f_\gamma=3{,}28\to x^{\text{кр}}_{\text{лев},\alpha/2}=\frac{1}{3{,}28}=0{,}30.$$

Таким образом, критическая область такова: (0; 0,3); (2,85; $+\infty$). Теперь найдем числовое значение критерия (10.88): $\varphi_{\text{чис}}=\dfrac{2{,}378}{1{,}685}=1{,}41$. Так как это число не попадает в критическую область, то гипотезу о равенстве дисперсий расхода сырья при старой и новой технологиях принимаем. ▲

В заключение отметим следующее.

Выше при проверке гипотез предполагалась нормальность распределения исследуемых случайных величин. Однако специальные исследования показали, что предложенные алгоритмы весьма устойчивы (особенно при больших объемах выборок) по отношению к отклонению от нормального распределения.

Сравнение параметров проводилось только для двух случайных величин. Осуществляя последовательно попарные сравнения, можно делать заключения о соотношении между параметрами любого числа случайных величин. Однако существуют критерии, позволяющие сравнивать сразу параметры более чем двух случайных величин (см. гл. 11).

§ 10.6. Проверка гипотезы о числовом значении вероятности события

Пусть A — случайное событие, вероятность p появления которого в единичном испытании неизвестна.

Выдвинем гипотезу

$$H_0: p = p_0 \tag{10.96}$$

о том, что вероятность p равна числу p_0. В основе проверки этой гипотезы должно лежать сравнение числа p_0 с приближенным значением вероятности p, найденным из опыта. Хорошим приближением к p является частость $\hat{p}=m/n$, где n — число испытаний Бернулли (это независимые испытания, проводимые в одинаковых условиях), а m — число испытаний среди n испытаний, в которых произошло событие A. Напомним, что, поскольку A — случайное событие, число m случайно, поэтому и частость $\hat{p}$ — случайная величина. Напомним также, что при большом n, согласно (9.74), имеем

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \approx N(0,1).$$

Если гипотеза (10.96) выполняется, то из последнего равенства получаем, что

$$\varphi = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \tag{10.97}$$

имеет нормальное распределение с нулевым математическим ожиданием и единичной дисперсией, т. е.

$$\varphi = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \approx N(0,1). \tag{10.98}$$

Величину (10.97) и используют в качестве критерия при проверке гипотезы (10.96).

Обратим внимание на то, что мы уже сталкивались с критерием, имеющим распределение $N(0, 1)$,— это критерий (10.10), и подробно рассмотрели алгоритмы построения критических областей для этого случая. Поэтому здесь лишь сошлемся на соответствующие схемы для трех видов альтернативной гипотезы H_1.

1) $H_1: p > p_0$, (10.99)

т. е. неизвестная вероятность p больше числа p_0. В этом случае критическая область имеет вид $(x^{\text{кр}}_{\text{пр},\alpha}, +\infty)$, где точка $x^{\text{кр}}_{\text{пр},\alpha}$ определяется по схеме (10.15). Критическая область изображена на рис. 10.2, *а*.

Теперь в формулу (10.97) подставим числовое значение частости $\hat{p}$, найденное из опыта, и заданные числа p_0 и n. В результате получим $\varphi_{\text{чис}}$. Если $\varphi_{\text{чис}}$ попадает в критическую область, то принимаем гипотезу H_1 (10.99), в противном случае принимаем гипотезу H_0 (10.96).

2) $H_1: p < p_0$, (10.100)

т. е. неизвестная вероятность p меньше p_0.

В этом случае критическая область имеет вид $(-\infty, x^{\text{кр}}_{\text{лев},\alpha})$, где точка $x^{\text{кр}}_{\text{лев},\alpha}$ определяется по схеме (10.21). Критическая область изображена на рис. 10.2, *б*.

Если числовое значение $\varphi_{\text{чис}}$ попадает в интервал $(-\infty, x^{\text{кр}}_{\text{лев},\alpha})$, то принимаем гипотезу H_1 (10.100); в противном случае принимаем гипотезу H_0 (10.96).

3) $H_1: p \neq p_0$. (10.101)

В этом случае критическая область состоит из двух интервалов $(-\infty, x^{\text{кр}}_{\text{лев},\alpha/2})$ и $(x^{\text{кр}}_{\text{пр},\alpha/2}, +\infty)$, где критические точки определяются по схеме (10.27). Критическая область изображена на рис. 10.2, *в*.

Если числовое значение $\varphi_{\text{чис}}$ критерия (10.97) попадает в критическую область, то принимаем гипотезу H_1 (10.101); в противном случае принимаем гипотезу H_0 (10.96).

Задача 10.7. Партия изделий принимается, если вероятность p того, что изделие окажется бракованным, не превышает $p_0=0{,}02$. Среди случайно отобранных $n=1000$ изделий оказалось $m=40$ бракованных. Можно ли при уровне значимости $\alpha=0{,}01$ принять партию?

△ Нулевая гипотеза $H_0: p=0{,}02$. Найдем относительную частоту брака:

$$\hat{p}=40/1000=0{,}04.$$

В качестве конкурирующей примем гипотезу $H_1: p>0{,}02$, т. е. имеет место случай 1, при этом $p_0=0{,}02$.

Критическую точку $x^{\text{кр}}_{\text{пр},\alpha}$ найдем по схеме (10.15):

$$\alpha=0{,}01 \to \gamma=1-2\alpha=0{,}98 \xrightarrow{\text{П. 2}} u_\gamma=2{,}33 \to x^{\text{кр}}_{\text{пр},\alpha}=2{,}33.$$

Критическую область образует интервал $(2{,}33; +\infty)$. Числовое значение критерия (10.97) таково: $\varphi_{\text{чис}}=\dfrac{0{,}04-0{,}02}{\sqrt{\dfrac{0{,}02\cdot 0{,}98}{1000}}}=4{,}5$. Так как это число попадает в критическую область, то гипотезу $H_0: p=0{,}02$ отвергаем и делаем заключение, что на уровне значимости $\alpha=0{,}01$ партию изделий принять нельзя. ▲

При малом числе наблюдений соотношение (10.98) не имеет места. В этом случае проверка гипотезы $H_0: p=p_0$ проводится следующим образом.

1) $H_1: p > p_0$.

Задаемся уровнем значимости α. Полагая $\gamma=1-2\alpha$ и зная число n испытаний Бернулли и число m таких испытаний, в которых произошло событие A, по табл. П. 5 найдем p_1 (это нижнее число).

Если $p_0 < p_1$, то принимаем гипотезу $H_1: p > p_0$; в противном случае принимаем гипотезу $H_0: p = p_0$.

2) $H_1: p < p_0$.

Полагая $\gamma=1-2\alpha$ и зная n и m, по табл. П. 5 найдем p_2 (это верхнее число).

Если $p_0>p_2$, то принимаем гипотезу $H_1: p<p_0$; в противном случае принимаем гипотезу $H_0: p=p_0$.

3) $H_1: p\neq p_0$.

Полагая $\gamma=1-\alpha$ и зная n и m, по табл. П. 5 находим p_1 и p_2.

Если $p_0<p_1$ или $p_0>p_2$, то принимаем гипотезу $H_1: p\neq p_0$; если $p_1<p_0<p_2$, то принимаем гипотезу $H_0: p=p_0$.

Задача 10.8. В $n=5$ опытах событие A произошло $m=4$ раза. Можно ли принять вероятность p события A (в генеральной совокупности) равной 0,2 при уровне значимости $\alpha=0{,}025$?

△ Итак, $H_0: p=p_0$, где $p_0=0{,}2$. Рассмотрим три случая для альтернативной гипотезы.

1) $H_1: p>p_0$. Пусть $\alpha=0{,}025$. Принимая $\gamma=1-2\alpha=0{,}95$, по табл. П. 5 найдем $p_1=0{,}284$. Так как $p_0<p_1$, то принимаем гипотезу H_1, т. е. считаем вероятность события большей чем 0,2.

2) $H_1: p<p_0$. Пусть $\alpha=0{,}025$. Принимая $\gamma=1-2\alpha=0{,}95$, по табл. П. 5 найдем $p_2=0{,}995$. Так как $p_0<p_2$, то принимаем гипотезу H_0, т. е. считаем вероятность события равной 0,2.

3) $H_1: p\neq p_0$. Пусть $\alpha=0{,}05$. Принимая $\gamma=1-\alpha=0{,}95$, по табл. П. 5 найдем $p_1=0{,}284$ и $p_2=0{,}995$. Так как p_0 не попадает в интервал (0,284; 0,995), то принимаем гипотезу $H_1: p\neq 0{,}2$. ▲

§ 10.7. Проверка гипотезы о равенстве вероятностей

Допустим, что в цехе работают ν станков. Пусть p_1, p_2, ..., p_ν — вероятности изготовления бракованного изделия на 1-м, 2-м, ... ν-м станке, числовые значения которых неизвестны. Возникает вопрос: можно ли считать эти вероятности одинаковыми, т. е. можно ли принять гипотезу

$$H_0: p_1=p_2=\ldots=p_\nu? \qquad (10.102)$$

В основе проверки этой гипотезы должно лежать сравнение приближенных значений этих вероятностей, вычисленных по выборочным наблюдениям. Организуем выборочные наблюдения следующим образом.

Из продукции, изготовленной i-м станком, отберем n_i изделий, подсчитаем среди них число m_i бракованных и за приближенное значение вероятности примем частость $\hat{p}_i=m_i/n$; отбор случаен и с возвратом, т. е. мы имеем дело с испытаниями Бернулли (напомним, что только при испытаниях Бернулли частость $\hat{p}_i$ является хорошим приближением к неизвестной вероятности p_i (см. § 9.3)). К выборкам с различных станков предъявим такое требование: появление или непоявление бракованных изделий на любом фиксированном количестве станков не должно изменять вероятности появления брака

на любом количестве других станков, или, иначе говоря, выборки должны быть независимы между собой.

Поступая таким образом, мы получим следующие три ряда чисел:

$$\begin{array}{lll} n_1 & n_2 & \ldots\ n_\nu, \\ m_1 & m_2 & \ldots\ m_\nu, \\ \hat{p}_1=m_1/n_1 & \hat{p}_2=m_2/n_2 & \ldots\ \hat{p}_\nu=m_\nu/n_\nu. \end{array}$$

Общее число проведенных испытаний $n=n_1+n_2+\ldots+n_\nu$, причем все n испытаний независимы.

Допустим, что гипотеза (10.102) выполняется: вероятность появления брака на каждом из ν станков одинакова (обозначим ее буквой p); это означает неизменность условий проведения n испытаний. Если учесть, что n испытаний независимы, то принятие гипотезы (10.102) означает, что мы располагаем n испытаниями Бернулли. И тогда наилучшее приближение к неизвестной вероятности p дает частость

$$\hat{p}=\frac{m_1+m_2+\ldots+m_\nu}{n_1+n_2+\ldots+n_\nu}. \qquad (10.103)$$

Поэтому при выполнении гипотезы (10.102) можно ожидать, что среди n_1 изделий, произведенных первым станком, число бракованных изделий $m_1^{\text{теор}}=n_1\hat{p}$; среди n_2 изделий, произведенных вторым станком, число бракованных изделий $m_2^{\text{теор}}=n_2\hat{p}$ и т. д.

Числа

$$m_1^{\text{теор}}=n_1\hat{p},\ m_2^{\text{теор}}=n_2\hat{p},\ \ldots,\ m_\nu^{\text{теор}}=n_\nu\hat{p} \qquad (10.104)$$

называются в отличие от опытных частот m_1, m_2, ..., m_ν *теоретическими* частотами.

Результаты поместим в табл. 10.2.

Таблица 10.2

Исход испытания	Номер исхода i	Номер станка j			
		1	2	...	ν
Изделие бракованное	1	$m_{11}=m_1$ / $m_{11}^{\text{теор}}=m_1^{\text{теор}}$	$m_{12}=m_2$ / $m_{12}^{\text{теор}}=m_2^{\text{теор}}$	...	$m_{1\nu}=m_\nu$ / $m_{1\nu}^{\text{теор}}=m_\nu^{\text{теор}}$
Изделие годное	2	$m_{21}=n_1-m_1$ / $m_{21}^{\text{теор}}=n_1-m_1^{\text{теор}}$	$m_{22}=n_2-m_2$ / $m_{22}^{\text{теор}}=n_2-m_2^{\text{теор}}$	...	$m_{2\nu}=n_\nu-m_\nu$ / $m_{2\nu}^{\text{теор}}=n_\nu-m_\nu^{\text{теор}}$
Число испытаний		n_1	n_2		n_ν

Теперь в каждой клетке проделаем следующие действия: от опытной частоты m_{ij} отнимем теоретическую $m_{ij}^{\text{теор}}$, возведем эту разность в квадрат, а затем разделим на $m_{ij}^{\text{теор}}$. Сложив найденные отношения, получим

$$\varphi=\sum_{i=1}^{2}\sum_{j=1}^{\nu}\frac{(m_{ij}-m_{ij}^{\text{теор}})^2}{m_{ij}^{\text{теор}}}. \qquad (10.105)$$

Оказывается, что эта величина имеет χ^2-распределение с числом степеней свободы $k=(2-1)(\nu-1)$, т. е.

$$\varphi\doteq\sum_{i=1}^{2}\sum_{j=1}^{\nu}\frac{(m_{ij}-m_{ij}^{\text{теор}})^2}{m_{ij}^{\text{теор}}}=\chi^2(k=\nu-1). \qquad (10.106)$$

Замечание. Используя соотношение (10.106), следует иметь в виду, что теоретические частоты в любой клетке табл. 10.2 должны быть не меньше пяти: $m_{ij}^{\text{теор}}\geqslant 5$.

Величину (10.106) и используют в качестве критерия при проверке гипотезы (10.102). При этом обычно критическая область является правосторонней, т. е. $(x_{\text{пр},\alpha}^{\text{кр}},\ +\infty)$. Точка $x_{\text{пр},\alpha}^{\text{кр}}$ находится из условия $P(\varphi>x_{\text{пр},\alpha}^{\text{кр}})=\alpha$, которое, учитывая (10.106), запишем в виде

$$P\left(\chi^2(k=\nu-1)>x_{\text{пр},\alpha}^{\text{кр}}\right)=\alpha. \qquad (10.107)$$

Для нахождения $x_{\text{пр},\alpha}^{\text{кр}}$ воспользуемся табл. П. 6. Полагая $\gamma=1-\alpha$ при $k=\nu-1$, найдем число χ_γ^2 такое, при котором

$$P\left(\chi^2(k=\nu-1)<\chi_\gamma^2\right)=\gamma=1-\alpha.$$

Тогда

$$P(\chi^2(k=\nu-1)>\chi_\gamma^2)=\alpha. \qquad (10.108)$$

Сопоставив (10.108) с (10.107), заключаем, что $x_{\text{пр},\alpha}^{\text{кр}}=\chi_{\gamma=1-\alpha}^2$.

Итак, схема нахождения критической точки такая:

$$\left.\begin{matrix}\alpha\to\gamma=1-\alpha\\ \nu\to k=\nu-1\end{matrix}\right\}\xrightarrow{\text{П. 6}}\chi_\gamma^2\to x_{\text{пр},\alpha}^{\text{кр}}=\chi_\gamma^2. \qquad (10.109)$$

Критическая область изображена на рис. 10.7.

Теперь, подставив в (10.105) конкретные числа, получим $\varphi_{\text{чис}}$. Если $\varphi_{\text{чис}}>x_{\text{пр},\alpha}^{\text{кр}}$, то гипотезу H_0 (10.102) о равенстве

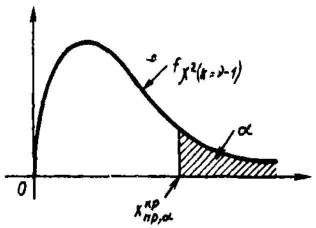


Рис. 10.7

вероятностей отвергаем; в противном случае эту гипотезу не отвергаем.

Замечание. Если гипотеза (10.102) принимается, то говорят, что различие частостей $\hat p_1=m_1/n_1$, $\hat p_2=m_2/n_2$, ..., $\hat p_\nu=m_\nu/n_\nu$ статистически незначимо и приближенное значение единой вероятности в этом случае равно $\hat p=\dfrac{m_1+m_2+\ldots+m_\nu}{n_1+n_2+\ldots+n_\nu}$.

Задача 10.9. Сравнить при уровне значимости $\alpha=0{,}1$ два участка цеха, выпускающих одну и ту же продукцию, с точки зрения доли рабочих, не выполнивших сменного задания. Исходные данные приведены в табл. 10.3.

Таблица 10.3

Исход испытания	Номер участка	
	1	2
Задание не выполнено	$m_1=15$ (рабочих)	$m_2=35$
Задание выполнено	$n_1-m_1=20$ (рабочих)	$n_2-m_2=50$
Число рабочих	$n_1=35$	$n_2=85$

△ Обозначим p_1 вероятность невыполнения сменного задания рабочим 1-го участка, p_2 — вероятность невыполнения сменного задания рабочим 2-го участка. Требуется проверить гипотезу $H_0: p_1=p_2$. Будем предполагать, что при сборе сведений относительно выполнения (невыполнения) сменного задания у $n_1+n_2=120$ рабочих соблюдались условия независимости испытаний.

Согласно (10.103), найдем

$$\hat p=\frac{m_1+m_2}{n_1+n_2}=\frac{15+35}{35+85}=\frac{50}{120}$$

— это оценка вероятности невыполнения задания рабочим, вычисленная в предположении, что вероятности p_1 и p_2 невыполнения сменного задания рабочим 1-го и 2-го участков одинаковы, и найдем теоретические частоты (10.104):

$$m_1^{\text{теор}}=n_1\hat p=35\cdot\frac{50}{120}=14{,}6;\quad m_2^{\text{теор}}=n_2\hat p=85\cdot\frac{50}{120}=35{,}4.$$

Заполним табл. 10.4.

Таблица 10.4

Исход испытания	Номер участка	
	1	2
Задание не выполнено	$m_{11}=m_1=15$	$m_{12}=m_2=35$
	$m_{11}^{\text{теор}}=m_1^{\text{теор}}=14{,}6$	$m_{12}^{\text{теор}}=m_2^{\text{теор}}=35{,}4$
Задание выполнено	$m_{21}=n_1-m_1=20$	$m_{22}=n_2-m_2=50$
	$m_{21}^{\text{теор}}=n_1-m_1^{\text{теор}}=20{,}4$	$m_{21}^{\text{теор}}=n_2-m_2^{\text{теор}}=49{,}6$
	$n_1=35$	$n_2=85$

Числовое значение критерия (10.105)

$$\varphi_{\text{чис}}=\frac{(m_{11}-m_{11}^{\text{теор}})^2}{m_{11}^{\text{теор}}}+\frac{(m_{12}-m_{12}^{\text{теор}})^2}{m_{12}^{\text{теор}}}+\frac{(m_{21}-m_{21}^{\text{теор}})^2}{m_{21}^{\text{теор}}}+$$
$$+\frac{(m_{22}-m_{22}^{\text{теор}})^2}{m_{22}^{\text{теор}}}=\frac{0,4^2}{14,6}+\frac{0,4^2}{35,4}+\frac{0,4^2}{35}+\frac{0,4^2}{49,6}=0,03.$$

Теперь по схеме (10.109) найдем критическую точку:

$$\left.\begin{array}{l}\alpha=0,1\to\gamma=1-\alpha=0,9\\ \nu=2\to k=\nu-1=1\end{array}\right\}\xrightarrow{\text{П. 6}}\chi_\gamma^2=2,71\to x_{\text{пр},\alpha}^{\text{кр}}=2,71.$$

Так как $\varphi_{\text{чис}}=0,03$ не попадает в критическую область $(2,71;\ +\infty)$, то гипотезу $H_0:p_1=p_2$ принимаем, т. е. считаем, что оба участка одинаковы с точки зрения доли рабочих, не выполнивших сменное задание, и за приближенное значение вероятности невыполнения сменного задания принимаем $\hat{p}=\frac{50}{120}$. ▲

Сравнить *две* вероятности p_1 и p_2, имея результаты достаточно большого числа n_1 и n_2 испытаний Бернулли, можно более простым способом. Построим критерий проверки гипотезы $H_0:p_1=p_2$.

□ Известно, что при большом числе испытаний для частостей $\hat{p}_1=m_1/n_1$ и $\hat{p}_2=m_2/n_2$ выполняется соотношение (9.73):

$$\hat{p}_1\approx N\left(p_1,\ \sqrt{p_1(1-p_1)/n_1}\right),$$
$$\hat{p}_2\approx N\left(p_2,\ \sqrt{p_2(1-p_2)/n_2}\right).$$

Как и прежде, будем полагать, что испытания независимы не только в пределах каждой выборки, но и между выборками. Отсюда следует, что величины m_1 и m_2 независимы. Следовательно, частости $\hat{p}_1$ и $\hat{p}_2$ также независимы; так как, кроме того, распределение величин $\hat{p}_1$ и $\hat{p}_2$ близко к нормальному, то и закон распределения разности $(\hat{p}_1-\hat{p}_2)$ будет близок к нормальному, при этом математическое ожидание

$$M(\hat{p}_1-\hat{p}_2)=M\hat{p}_1-M\hat{p}_2=p_1-p_2,$$

а дисперсия

$$D(\hat{p}_1-\hat{p}_2)=D\hat{p}_1+D\hat{p}_1=\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2},$$

т. е.
$$\hat{p}_1-\hat{p}_2\approx N\left(p_1-p_2,\ \sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}\right).$$

Таким образом,

$$\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}\approx N(0,1).$$

Если гипотеза $H_0:p_1=p_2$ верна (общее значение вероятности обозначим символом p), то из последнего равенства получаем

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{p(1-p)(1/n_1+1/n_2)}}\approx N(0,1). \qquad (10.110)$$

Учитывая, что n_1+n_2 велико, в (10.110) выражение $p(1-p)$ заменим его приближенным значением $\hat{p}(1-\hat{p})$, где $\hat{p}=\frac{m_1+m_2}{n_1+n_2}$. В результате имеем

$$\varphi=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{m_1+m_2}{n_1+n_2}\left(1-\frac{m_1+m_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\approx N(0,1). \qquad (10.111)$$

Эту величину и используют в качестве критерия при проверке гипотезы $H_0:p_1=p_2$. ■

Поскольку мы уже неоднократно сталкивались с критерием, имеющим распределение $N(0,1)$, не будем рассматривать дальнейшие этапы проверки гипотезы H_0. Отметим лишь, что схемы нахождения критических точек следующие:

(10.15) — для гипотезы $H_1:p_1>p_2$ (см. рис. 10.2, *а*);
(10.21) — для гипотезы $H_1:p_1<p_2$ (см. рис. 10.2, *б*);
(10.27) — для гипотезы $H_1:p_1\neq p_2$ (см. рис. 10.2, *в*).

Задача 10.10. Сравнить с точки зрения доли брака при уровне значимости $\alpha=0,1$ две партии изделий, располагая следующими данными:

Номер партии	Объем выборки	Число дефектных изделий в выборке	Частость появления дефектного изделия
1	$n_1=200$	$m_1=5$	$\hat{p}_1=m_1/n_1=5/200$
2	$n_2=300$	$m_2=10$	$\hat{p}_2=m_2/n_2=10/300$

△ Итак, проверим гипотезу $H_0:p_1=p_2$, где p_1 и p_2 — вероятности появления бракованного изделия соответственно в 1-й и 2-й партиях. Поскольку в выборках $\hat{p}_1<\hat{p}_2$, в качестве альтернативной возьмем гипотезу $H_1:p_1<p_2$.

Найдем числовое значение критерия (10.111). Имеем

$$\varphi_{\text{чис}}=\frac{5/200-10/300}{\sqrt{(15/500)(1-15/500)(1/200+1/300)}}=-0,54.$$

Так как альтернативная гипотеза имеет вид $H_1:p_1<p_2$, то критическая область является левосторонней (см. рис. 10.2, *б*) и точку $x_{\text{лев},\alpha}^{\text{кр}}$ следует находить по схеме (10.21):

$$\alpha=0,1\to\gamma=1-2\alpha=0,8\xrightarrow{\text{П.2}}u_\gamma=1,28\to x_{\text{лев},\alpha}^{\text{кр}}=-1,28.$$

Значение $\varphi_{\text{чис}}=-0,54$ не попадает в критическую область $(-\infty;\ -1,28)$, поэтому гипотеза об идентичности партий с точки зрения доли брака не отвергается. ▲

§ 10.8. Проверка гипотезы о модели закона распределения. Критерий согласия Пирсона

В предыдущих параграфах настоящей главы рассматривались гипотезы, относящиеся к отдельным параметрам распределения случайных величин, причем модели законов распределения этих величин предполагались известными. Однако во многих практических задачах модель закона распределения заранее не известна и возникает задача выбора модели, согласующейся с результатами наблюдений над случайной величиной.

Пусть высказано предположение, что неизвестная функция распределения $F_X(x)$ исследуемой случайной величины X имеет вполне определенную модель $F_{\text{теор}}(x)$, т. е. высказана гипотеза

$$H_0: F_X(x) = F_{\text{теор}}(x). \tag{10.112}$$

В качестве теоретической модели $F_{\text{теор}}(x)$ может быть рассмотрена нормальная, биномиальная или какая-либо другая модель. Это определяется сущностью изучаемого явления, а также результатами предварительной обработки наблюдений над случайной величиной (формой графика вариационного ряда, соотношениями между выборочными характеристиками и т. д.).

Критерии, с помощью которых проверяется гипотеза (10.112), называются *критериями согласия*. Рассмотрим лишь один из них, использующий χ^2-распределение и получивший название *критерия согласия Пирсона*.

Критерий предполагает, что результаты наблюдений сгруппированы в вариационный ряд. Для определенности положим, что это дискретный вариационный ряд с числом групп, равным ν (см. строки 1 и 2 табл. 10.5).

Однако, прежде чем рассматривать сам критерий Пирсона, вспомним параметрическое оценивание закона распределения (см. § 9.5). Последовательность оценивания такая: формулируют гипотезу о модели закона распределения случайной величины; по результатам наблюдений находят оценки неизвестных параметров этой модели (допустим, что число неизвестных параметров равно l); вместо неизвестных параметров подставляют в модель закона найденные оценки. В результате предполагаемая модель закона оказывается полностью определенной и, используя ее, рассчитывают вероятности $p_i^{\text{теор}} = P(X = x_i)$ того, что случайная величина X примет зафиксированные в наблюдениях значения x_i, $i = 1, 2, \ldots, \nu - 1$; эти вероятности называют *теоретическими*. Обратим внимание на следующее обстоятельство: так как сумма вероятностей ряда распределения должна быть равна единице, т. е.

$$\sum_i p_i^{\text{теор}} = 1, \tag{10.113}$$

Таблица 10.5

x_i	x_1	...	$x_{\nu-1}$	x_ν
m_i	m_1	...	$m_{\nu-1}$	m_ν
$p_i^{\text{теор}} = P(X = x_i)$	$p_1^{\text{теор}} = P(X = x_1)$	...	$p_{\nu-1}^{\text{теор}} = P(X = x_{\nu-1})$	$p_\nu^{\text{теор}} = 1 - p_1^{\text{теор}} - \ldots - p_{\nu-1}^{\text{теор}}$
$m_i^{\text{теор}} = np_i^{\text{теор}}$	$m_1^{\text{теор}} = np_1^{\text{теор}}$	...	$m_{\nu-1}^{\text{теор}} = np_{\nu-1}^{\text{теор}}$	$m_\nu^{\text{теор}} = np_\nu^{\text{теор}}$

то полагаем вероятность $p_\nu^{\text{теор}} = 1 - p_1^{\text{теор}} - p_2^{\text{теор}} - \ldots - p_{\nu-1}^{\text{теор}}$. Теоретические вероятности записаны в строке 3 табл. 10.5. Теперь найдем теоретические частоты $m_i^{\text{теор}} = np_i^{\text{теор}}$; они записаны в строке 4 табл. 10.5.

Обратим внимание на следующее: критерий согласия Пирсона можно использовать только в том случае, когда

$$m_i^{\text{теор}} \geqslant 5, \quad i = 1, 2, \ldots, \nu. \tag{10.114}$$

Поэтому ту группу вариационного ряда, для которой это условие не выполняется, объединяют с соседней и соответственно уменьшают число групп; так поступают до тех пор, пока для каждой новой группы $m_i^{\text{теор}}$ будет не меньше 5. Новое число групп, как и прежде обозначим символом ν.

Оказывается, что если предполагаемая модель закона распределения действительно имеет мето, т. е. верна гипотеза (10.112), и если к тому же выполняются условия (10.113) и (10.114), то величина

$$\varphi = \sum_{i=1}^{\nu} \frac{(m_i - m_i^{\text{теор}})^2}{m_i^{\text{теор}}} \tag{10.115}$$

будет иметь χ^2-распределение с числом степеней свободы $k = \nu - l - 1$, т. е.

$$\varphi = \sum_{i=1}^{\nu} \frac{(m_i - m_i^{\text{теор}})^2}{m_i^{\text{теор}}} = \chi^2(k = \nu - l - 1),$$

где ν — *число (новое) групп вариационного ряда;* l — *число неизвестных параметров предполагаемой модели, оцениваемых по результатам наблюдений (если все параметры предполагаемого закона известны точно, то* $l = 0$). Величину (10.115) и называют *критерием согласия* χ^2 или *критерием согласия Пирсона*.

Далее поступаем так же, как обычно при проверке гипотез. Задаемся уровнем значимости α. Зная распределение критерия φ, находим критическую область; как правило, это область правосторонняя, т. е. она имеет вид $(x_{\text{пр}, \alpha}^{\text{кр}}, +\infty)$; найдем числовое значение $\varphi_{\text{чис}}$ критерия (10.115). Если $\varphi_{\text{чис}}$ попадает в интервал $(x^{\text{кр}}_{\text{пр},\,\alpha},\ +\infty)$, то делаем вывод о неправомерности гипотезы H_0 (10.112); при этом не следует забывать, что этот вывод может быть ошибочным (на самом деле в генеральной совокупности гипотеза H_0 (10.112) имеет место) и вероятность того, что вывод ошибочен, равна α.

Если $\varphi_{\text{чис}}$ не попадает в интервал $(x^{\text{кр}}_{\text{пр},\,\alpha},\ +\infty)$, то гипотеза H_0 (10.112) не отвергается.

В заключение приведем схему определения точки $x^{\text{кр}}_{\text{пр},\,\alpha}$:

$$\left.\begin{aligned}&\alpha\to\gamma = 1-\alpha\\ &l,\ \nu\to k=\nu-l-1\end{aligned}\right\}\xrightarrow{\text{П.6}}\chi^2_\gamma\to x^{\text{кр}}_{\text{пр},\,\alpha}=\chi^2_\gamma. \qquad (10.116)$$

Задача 10.11. По распределению успеваемости 100 студентов-заочников, которые сдавали четыре экзамена, заданному в примере 9.3, выяснить, можно ли при уровне значимости $\alpha=0{,}1$ считать, что число сданных экзаменов среди четырех имеет биномиальный закон распределения.

△ Обратимся к примеру 9.3. В нем была приведена формула биномиального закона (9.37); этот закон имеет один параметр — это вероятность p. Числовое значение этого параметра было неизвестно, поэтому по результатам наблюдений находилась оценка $\hat p$ этого параметра: $\hat p=0{,}88$, которая использовалась в формуле биномиального закона. Теоретические вероятности $p_i^{\text{теор}}$, вычисленные по формуле (9.38), и теоретические частоты $m_i^{\text{теор}}$ приведены в столбцах 5 и 6 табл. 9.7. Обратим внимание на то, что в этой таблице условие (10.113) выполняется. Условие же (10.114) не выполняется для первых двух групп. Присоединим эти группы к третьей группе. В результате получим новый ряд с числом групп $\nu=3$, при этом $m_1=1+1+3=5$, $m_1^{\text{теор}}=0{,}021+0{,}608+6{,}691=7{,}320$; $m_2=35$, а $m_2^{\text{теор}}=32{,}711$; $m_3=60$, $m_3^{\text{теор}}=59{,}969$.

Числовое значение критерия (10.115) приведено в столбце 8 табл. 9.4. Оно равно $\varphi_{\text{чис}}=0{,}895$. Теперь по схеме (10.116) найдем критическую точку:

$$\left.\begin{aligned}&\alpha=0{,}1 \quad\to\gamma=1-\alpha=0{,}9\\ &l=1,\ \nu=3\to k=\nu-l-1=1\end{aligned}\right\}\xrightarrow{\text{П.6}}\chi^2_\gamma=2{,}71\to x^{\text{кр}}_{\text{пр},\,\alpha}=2{,}71.$$

Так как $\varphi_{\text{чис}}=0{,}895$ не попадет в критическую область $(2{,}71;\ +\infty)$, то гипотезу о том, что число сданных экзаменов среди четырех имеет биномиальный закон распределения, не отвергаем. ▲

Выше предполагалось, что результаты наблюдений сгруппированы в дискретный вариационный ряд. Если они сгруппированы в интервальный ряд, то при использовании критерия согласия Пирсона теоретические вероятности $p_i^{\text{теор}}$ находятся так, как указано в табл. 10.6.

Обратим внимание на то, что в табл. 10.6 сумма теоретических вероятностей $\sum_{i=1}^{\nu} p_i^{\text{теор}}=1$. Обратим также внимание на то, что критерий согласия можно использовать только в том случае, когда для каждого интервала $m_i^{\text{теор}}\geqslant 5$. Интервалы, для которых это условие не выполняется, объединим с соседними. Новое число интервалов обозначим, как и прежде, буквой ν.

Таблица 10.6

Интервал $[a_i;\ a_{i+1})$ Интервальная частота m_i	$[-\infty,\ a_2)$ m_1	$[a_2,\ a_3)$ m_2	…	$[a_{\nu-1},\ a_\nu)$ $m_{\nu-1}$	$[a_\nu,\ +\infty)$ m_ν
$p_i^{\text{теор}}$	$p_1^{\text{теор}}=P(X<a_2)$	$p_2^{\text{теор}}= =P(a_2\leqslant X<a_3)$	…	$p_{\nu-1}^{\text{теор}}= =P(a_{\nu-1}\leqslant X<a_\nu)$	$p_\nu^{\text{теор}}= =P(X\geqslant a_\nu)= =1-p_1^{\text{теор}}-\ldots \ldots-p_{\nu-1}^{\text{теор}}$
$m_1^{\text{теор}}$	$m_l^{\text{теор}}=np_1^{\text{теор}}$	$m_2^{\text{теор}}=np_2^{\text{теор}}$	…	$m_{\nu-1}^{\text{теор}}=np_{\upsilon-1}^{\text{теор}}$	$m_\nu^{\text{теор}}=np_\nu^{\text{теор}}$

Задача 10.12. По распределению валиков, заданному табл. 8.5, выяснить, можно ли при уровне значимости $\alpha=0{,}1$ считать, что размер диаметра валика имеет нормальный закон распределения.

△ Обратимся к примеру 9.4. В этом примере в предположении, что нормальный закон имеет место, были оценены два параметра данного закона, а затем в табл. 9.9 в столбце 6 были приведены значения функции нормального распределения в точках, совпадающих с границами интервала. Покажем, как, зная эти значения, найти теоретические вероятности из табл. 10.6 ($\nu=9$). Имеем:

$$p_1^{\text{теор}}=P(X<a_2)=F_N(a_2)-0=0{,}0154,$$
$$p_2^{\text{теор}}=P(a_2\leqslant X<a_3)=F_N(a_3)-F_N(a_2)=0{,}0643-0{,}0154,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$p_8^{\text{теор}}=P(a_8\leqslant X<a_9)=F_N(a_9)-F_N(a_8)=0{,}9892-0{,}9515,$$
$$p_9^{\text{теор}}=P(X\geqslant a_9)=1-P(X<a_9)=1-F_N(a_9)=1-0{,}9892.$$

Эти результаты, а также числовое значение критерия (10.115) приведены в табл. 10.7; $\varphi_{\text{чис}}=6{,}755$.

Таблица 10.7

i	a_i-a_{i+1}	m_i	$F_N(a_i)$	$F_N(a_{i+1})$	$p_i^{\text{теор}}= =F_N(a_{i+1})- -F_N(a_i)$	$m_i^{\text{теор}}=np_i^{\text{теор}}$	$\dfrac{(m_i^{\text{теор}}-m_i)^2}{m_i^{\text{теор}}}$
1	6,67—6,69	2 } 17	0	0,0154	0,0154	3,08 } 12,86	1,333
2	6,69—6,71	15	0,0154	0,0643	0,0489	9,78	
3	6,71—6,73	17	0,0643	0,1894	0,1251	25,02	2,571
4	6,73—6,75	44	0,1894	0,4013	0,2119	42,38	0,062
5	6,75—6,77	52	0,4013	0,6517	0,2504	50,08	0,074
6	6,77—6,79	44	0,6517	0,8485	0,1968	39,36	0,055
7	6,79—6,81	14	0,8485	0,9515	0,1030	20,6	2,115
8	6,81—6,83	11 } 12	0,9515	0,9892	0,0377	7,54 } 9,70	0,545
9	6,83—6,85	1	0,9892	1	0,0108	2,16	
Итого		200			1,0000		6,755

Теперь по схеме (10.116) найдем критическую точку, при этом отметим, что:

у предполагаемой нормальной модели были неизвестны оба параметра и они оценивались по результатам наблюдений, поэтому $l=2$;

при расчете критерия в табл. 10.7 использовалось семь интервалов, т. е. $\nu=7$.

Тогда

$$\left.\begin{array}{l}\alpha=0{,}1 \quad \to\gamma=1-\alpha=0{,}9\\ l=2,\ \nu=7\to k=\nu-l-1=4\end{array}\right\}\xrightarrow{\text{П.6}}\chi^2_\gamma=7{,}78\to x^{\text{кр}}_{\text{пр.}\,\alpha}=7{,}78.$$

Так как $\varphi_{\text{чис}}=6{,}755$ не попадает в критическую область $(7{,}78;\ +\infty)$, то гипотезу о том, что размер диаметра валика имеет нормальный закон распределения, не отвергаем. ▲

ГЛАВА 11
ОСНОВЫ ДИСПЕРСИОННОГО АНАЛИЗА

§ 11.1. Однофакторный дисперсионный анализ

Задача дисперсионного анализа и предварительная обработка результатов наблюдений. Допустим, что экономиста строительно-монтажного управления интересует зависимость объема выполненных на стройке работ за смену от работающей на стройке бригады. Предположим, что на стройке могут работать ν бригад. Будем называть объем выполненных работ *результативным признаком*, обозначать Y и полагать, что Y — случайная величина; работающую бригаду назовем *фактором* A, а номер работающей бригады — *уровнем или группой фактора* A и через $A^{(i)}$ обозначать i-й уровень или группу фактора A (i-ю бригаду, $i=1,\ 2,\ ...,\ \nu$).

Приступая к выяснению интересующей нас зависимости, необходимо над каждой бригадой провести наблюдения. Обратим внимание на то, что объем выполненных работ зависит не только от работающей бригады, но и от ряда случайных факторов. Поэтому по каждой бригаде будет наблюдаться вариация, изменчивость ежедневного объема выполненных работ. Результаты наблюдений и результаты их предварительной обработки расположим в табл. 11.1.

В табл. 11.1 $Y_j^{(i)}$ — объем работ, зафиксированный в j-м наблюдении над i-й бригадой, или, иначе говоря, значение результативного признака, зафиксированное в j-м наблюдении при i-м уровне фактора A, при этом $i=1,\ 2,\ ...,\ \nu$, а $j=1,\ 2,\ ...,\ n_i$;

$\bar{Y}^{(i)}$ — среднее, вычисленное по результатам наблюдений при i-м уровне фактора, или в i-й группе, $i=1,\ 2,\ ...,\ \nu$. Эти средние вычисляют по формуле

$$\bar{Y}^{(i)}=\frac{Y_1^{(i)}+Y_2^{(i)}+...+Y_{n_i}^{(i)}}{n_i}=\sum_{j=1}^{n_i}Y_j^{(i)}/n_i \qquad (11.1)$$

и называют *групповыми средними*.

Таблица 11.1

i	1	2	...	ν
Уровень фактора A / Номер (j) наблюдения в столбце	$A^{(1)}$	$A^{(2)}$	...	$A^{(\nu)}$
1	$Y_1^{(1)}$	$Y_1^{(2)}$	...	$Y_1^{(\nu)}$
2	$Y_2^{(1)}$	$Y_2^{(2)}$	...	$Y_2^{(\nu)}$
...	...	...	...	...
n_1	$Y_{n_1}^{(1)}$			$Y_{n_\nu}^{(\nu)}$
...		$Y_{n_2}^{(2)}$		
Число наблюдений в группе	n_1	n_2	...	n_ν
Групповое среднее	$\bar{Y}^{(1)}$	$\bar{Y}^{(2)}$		$\bar{Y}^{(\nu)}$
Групповая выборочная дисперсия	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$		$\hat{\sigma}_\nu^2$

$\hat{\sigma}_i^2$ — выборочная дисперсия, вычисленная по результатам наблюдений при i-м уровне фактора, $i=1,\ 2,\ ...,\ \nu$; выборочные дисперсии вычисляют по формуле

$$\hat{\sigma}_i^2=\frac{(Y_1^{(i)}-\bar{Y}^{(i)})^2+(Y_2^{(i)}-\bar{Y}^{(i)})^2+...+(Y_{n_i}^{(i)}-\bar{Y}^{(i)})^2}{n_i}=\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y}^{(i)})^2/n_i \qquad (11.2)$$

и называют *групповыми выборочными дисперсиями*.

Предварительное суждение о том, зависит ли объем выполненных работ от работающей бригады, можно вынести сравнив групповые средние: если различие между ними существенно, то, по-видимому, такая зависимость имеется. Однако не надо забывать, что мы имеем дело с выборочными наблюдениями, а ответ на вопрос, существует ли зависимость результата от фактора, должен быть дан применительно к генеральной совокупности. Поэтому логика рассуждений должна быть следующей.

Обозначим: a_1 — математическое ожидание результативного признака при уровне $A^{(1)}$; a_2 — математическое ожидание результативного признака при уровне $A^{(2)}$; ...; a_ν — математическое ожидание при уровне $A^{(\nu)}$.

Если при изменении уровня фактора групповые математические ожидания не изменяются, т. е. $a_1 = a_2 = \ldots = a_\nu$, то считаем, что результативный признак не зависит от фактора A, в противном случае такая зависимость имеется. Но поскольку числовые значения математических ожиданий неизвестны, возникает задача проверки гипотезы

$$H_0: a_1 = a_2 = \ldots = a_\nu. \tag{11.3}$$

Условия проведения дисперсионного анализа. Критерий Бартлетта. Проверить гипотезу о равенстве групповых математических ожиданий можно при соблюдении следующих требований:

при каждом уровне фактора наблюдения независимы и проводятся в одинаковых условиях; наблюдения независимы при различных уровнях фактора; (11.4)

при каждом уровне фактора результативный признак имеет нормальный закон распределения с постоянной для различных уровней генеральной дисперсией (обозначим ее σ_0^2). (11.5)

Если соблюдение требования (11.4) определяется организацией эксперимента, то выполнение требования (11.5) от этого не зависит.

При решении вопроса о виде распределения результативного признака при каждом уровне фактора, во-первых, надо учитывать природу изучаемого явления: возможно, что механизм формирования значений результативного признака удовлетворяет условиям центральной предельной теоремы и тогда этот признак будет иметь 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.

Как установить, одинаковы ли генеральные дисперсии результативного признака при различных уровнях фактора или нет? 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 ответить на этот вопрос, можно лишь проверить гипотезу

$$H_0: \sigma_i^1 = \sigma_2^2 = \ldots = \sigma_\nu^2. \tag{11.6}$$

Эта гипотеза проверяется следующим образом.

1. Находим несмещенные оценки s_i^2 групповых дисперсий σ_i^2 по формуле

$$s_i^2 = \frac{\hat{\sigma}_i^2 n_i}{n_i - 1}, \quad i = 1, 2, \ldots, \nu, \tag{11.7}$$

где $\hat{\sigma}_i^2$ — выборочные групповые дисперсии, найденные по формуле (11.2), n_i — численность наблюдений в группах.

2. Находим

$$s_0^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_\nu-1)s_\nu^2}{(n_1-1)+(n_2-1)+\ldots+(n_\nu-1),}. \tag{11.8}$$

3. Далее находим

$$q = \left[1 + \frac{1}{3(\nu-1)}\left(\frac{1}{n_1-1} + \frac{1}{n_2-1} + \ldots + \frac{1}{n_\nu-1} - \frac{1}{(n_1-1)+\ldots+(n_\nu-1)}\right)\right]^{-1} \tag{11.9}$$

и

$$\varphi = q\left[(n_1-1)\ln\frac{s_0^2}{s_1^2} + (n_2-1)\ln\frac{s_0^2}{s_2^2} + \ldots + (n_\nu-1)\ln\frac{s_0^2}{s_\nu^2}\right]. \tag{11.10}$$

Величиниа (11.10) называется *критерием Бартлетта*; при выполнеиии условия

$$n_i > 3, \quad i = 1, 2, \ldots, \upsilon, \tag{11.11}$$

и гнпотезы (11.6) величина φ имеет распределеиие, близкое к χ^2-распределению с $k = \nu - 1$ степенями свободы.

4. Задавшись уровием значимости α, находим правостороииюю критическую точку $x_{\text{пр},\alpha}^{\text{кр}}$ по следующей схеме:

$$\left.\begin{array}{l} \alpha \to \gamma = 1 - \alpha \\ \nu \to k = \nu - 1 \end{array}\right\} \xrightarrow{\text{П.6}} \chi_\gamma^2 \to x_{\text{пр},\alpha}^{\text{кр}} = \chi_\gamma^2. \tag{11.12}$$

5. Находим числовое значенне $\varphi_{\text{чис}}$ критерия (11.10). Если $\varphi_{\text{чис}}$ попадает в интервал $(x_{\text{пр},\alpha}^{\text{кр}}, +\infty)$, то гипотезу H_0 (11.6) отвергаем. В противиом случае считаем, что гипотеза (11.6) не противоречит результатам иаблюдений.

Итак, допустим, что требования (11.4) и (11.5) выполняются. Прежде чем приступить к проверке гипотезы (11.3), введем ряд новых понятий.

Разложение дисперсии наблюдаемых «игреков». Обратимся к табл. 11.1. Изменчивость, или вариация наблюдаемых значений результативного признака Y может быть вызвана изменчивостью уровней фактора A и изменчивостью значений случайных неконтролируемых факторов, влияющих на Y, которые называют *остаточными**⁾.

1) Вариация наблюдаемых «игреков», вызванная влиянием на Y фактора A и остаточных факторов,— *общая вариация величины* Y — измеряется выборочной дисперсией $\hat{\sigma}_Y^2$. Прежде чем привести формулу для $\hat{\sigma}_Y^2$, вычислим среднее $\bar{Y}$ всех наблюдаемых значений результативного признака, которое называют *общим средним признака* Y. Его можно найти по одной из следующих тождественных формул:

$$\bar{Y} = \frac{n_1 \bar{Y}^{(1)} + n_2 \bar{Y}^{(2)} + \ldots + n_\nu \bar{Y}^{(\nu)}}{n_1 + n_2 + \ldots + n_\nu}, \tag{11.13}$$

*⁾ Следует обратить внимаиие на то, что мы считаем уровень фактора A фиксированной величиной; случаен же признак Y. Поэтому речь может идти о случайных факторах, влияющих только на Y.

где $\bar{Y}^{(1)}, \bar{Y}^{(2)}, \ldots, \bar{Y}^{(\nu)}$ — групповые средние, вычисляемые по формуле (11.1), или

$$\bar{Y}=\frac{(Y_1^{(1)}+Y_2^{(1)}+\ldots+Y_{n_1}^{(1)})+(Y_1^{(2)}+Y_2^{(2)}+\ldots+Y_{n_2}^{(2)})+\ldots+(Y_1^{(\nu)}+Y_2^{(\nu)}+\ldots+Y_{n_\nu}^{(\nu)})}{n}. \tag{11.14}$$

Далее находим

$$\begin{aligned}S_Y^2=&(Y_1^{(1)}-\bar{Y})^2+(Y_2^{(1)}-\bar{Y})^2+\ldots+(Y_{n_1}^{(1)}-\bar{Y})^2+\\&+(Y_1^{(2)}-\bar{Y})^2+(Y_2^{(2)}-\bar{Y})^2+\ldots+(Y_{n_2}^{(2)}-\bar{Y})^2+\\&+\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots+\\&+(Y_1^{(\nu)}-\bar{Y})^2+(Y_2^{(\nu)}-\bar{Y})^2+\ldots+(Y_{n_\nu}^{(\nu)}-\bar{Y})^2=\\&=\sum_{i=1}^{\nu}\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y})^2.\end{aligned} \tag{11.15}$$

Обратим внимание на то, что S_Y^2 получена как сумма $(n_1+n_2+\ldots+n_\nu)=n$ слагаемых. Тогда

$$\hat{\sigma}_Y^2=S_Y^2/n=\sum_{i=1}^{\nu}\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y})^2/n. \tag{11.16}$$

2) Построим показатель вариации наблюдаемых «игреков», вызванной изменчивостью уровней фактора A. Очевидно, что чем сильнее зависимость Y от фактора A, тем сильнее изменчивость групповых средних $\bar{Y}^{(1)}, \bar{Y}^{(2)}, \ldots, \bar{Y}^{(\nu)}$ и тем больше их разброс около общего среднего $\bar{Y}$. Показателем этого разброса, или вариации, является *дисперсия* $\hat{\sigma}^2_{\bar{Y}^{(i)}}$ *групповых средних*, которую обозначим $\hat{\sigma}_\Phi^2$. Она вычисляется следующим образом:

находят

$$\begin{aligned}S_\Phi^2=&(\bar{Y}^{(1)}-\bar{Y})^2n_1+(\bar{Y}^{(2)}-\bar{Y})^2n_2+\ldots+(\bar{Y}^{(\nu)}-\bar{Y})^2n_\nu=\\&=\sum_{i=1}^{\nu}(\bar{Y}^{(i)}-\bar{Y})^2n_i.\end{aligned} \tag{11.17}$$

Тогда

$$\hat{\sigma}_\Phi^2=S_\Phi^2/n=\sum_{i=1}^{\nu}(\bar{Y}^{(i)}-\bar{Y})^2n_i/n. \tag{11.18}$$

Обратим внимание на следующее: групповое среднее $\bar{Y}^{(i)}$ является «представителем» n_i наблюдений в группе, поэтому в формулу (11.17) введены веса $n_1, n_2, \ldots, n_\nu$, т. е. S_Φ^2 является взвешенной суммой, а $\hat{\sigma}_\Phi^2$ — выборочной взвешенной дисперсией групповых средних.

3) Построим показатель вариации наблюдаемых «игреков», вызванной изменчивостью остаточных случайных факторов.

Зафиксируем какой-либо уровень фактора, например $A^{(i)}$. Вариация наблюдений $Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_{n_i}^{(i)}$ внутри i-й группы относительно группового среднего $\bar{Y}^{(i)}$ вызывается влиянием остаточных факторов и измеряется групповой дисперсией $\hat{\sigma}_i^2$, вычисляемой по формуле (11.2). Тогда по всем группам вариация «игреков», вызванная влиянием на Y остаточных факторов, будет измеряться *средней* $\overline{\hat{\sigma}_i^2}$ *групповых дисперсий*, которую обозначим $\hat{\sigma}_0^2$, причем, поскольку i-я групповая дисперсия вычисляется по n_i наблюдениям, средняя будет взвешенной. Таким образом,

$$\hat{\sigma}_0^2=\frac{\hat{\sigma}_1^2n_1+\hat{\sigma}_2^2n_2+\ldots+\hat{\sigma}_\nu^2n_\nu}{n_1+n_2+\ldots+n_\nu}=\frac{S_0^2}{n}, \tag{11.19}$$

где $S_0^2=\hat{\sigma}_1^2n_1+\hat{\sigma}_2^2n_2+\ldots+\hat{\sigma}_\nu^2n_\nu$; учитывая (11.2), имеем

$$\begin{aligned}S_0^2=&(Y_1^{(1)}-\bar{Y}^{(1)})^2+(Y_2^{(1)}-\bar{Y}^{(1)})^2+\ldots+(Y_{n_1}^{(1)}-\bar{Y}^{(1)})^2+\\&+(Y_1^{(2)}-\bar{Y}^{(2)})^2+(Y_2^{(2)}-\bar{Y}^{(2)})^2+\ldots+(Y_{n_2}^{(2)}-\bar{Y}^{(2)})^2+\\&+\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots+\\&+(Y_1^{(\nu)}-\bar{Y}^{(\nu)})^2+(Y_2^{(\nu)}-\bar{Y}^{(\nu)})^2+\ldots+(Y_{n_\nu}^{(\nu)}-\bar{Y}^{(\nu)})^2=\\&=\sum_{i=1}^{\nu}\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y}^{(i)})^2.\end{aligned} \tag{11.20}$$

Докажем справедливость *формулы разложения дисперсии*

$$\hat{\sigma}_Y^2=\hat{\sigma}^2_{\bar{Y}^{(i)}}+\overline{\hat{\sigma}_i^2}\quad(\text{или}\ \ \hat{\sigma}_Y^2=\hat{\sigma}_\Phi^2+\hat{\sigma}_0^2), \tag{11.21}$$

которая читается так: *общая выборочная дисперсия равна сумме дисперсии групповых средних и средней из групповых дисперсий.*

Учитывая формулы (11.16), (11.18) и (11.19), перейдем от (11.21) к соотношению

$$S_Y^2=S_\Phi^2+S_0^2 \tag{11.22}$$

или к соотношению

$$\sum_{i=1}^{\nu}\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y})^2=\sum_{i=1}^{\nu}(\bar{Y}^{(i)}-\bar{Y})^2n_i+\sum_{i=1}^{\nu}\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y}^{(i)})^2. \tag{11.23}$$

Докажем последнюю формулу.

Имеем очевидное тождество

$$Y_j^{(i)}-\bar{Y}=(\bar{Y}^{(i)}-\bar{Y})+(Y_j^{(i)}-\bar{Y}^{(i)}), \tag{11.24}$$

которое называется *дисперсионным тождеством*. Возведем обе его части в квадрат. Имеем

$$(Y_j^{(i)}-\bar{Y})^2=(\bar{Y}^{(i)}-\bar{Y})^2+(Y_j^{(i)}-\bar{Y}^{(i)})^2+2(\bar{Y}^{(i)}-\bar{Y})(Y_j^{(i)}-\bar{Y}^{(i)});$$

просуммируем обе части по j от 1 до n_i. В результате получаем

$$\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y})^2=\sum_{j=1}^{n_i}(\bar{Y}^{(i)}-\bar{Y})^2+\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y}^{(i)})^2+2\sum_{j=1}^{n_i}(\bar{Y}^{(i)}-\bar{Y})(Y_j^{(i)}-\bar{Y}^{(i)}). \tag{11.25}$$

Первое из слагаемых в (11.25)

$$\sum_{j=1}^{n_i}(\bar{Y}^{(i)}-\bar{Y})^2=(\bar{Y}^{(i)}-\bar{Y})^2\sum_{j=1}^{n_i}1=(\bar{Y}^{(i)}-\bar{Y})^2 n_i,$$

так как выражение $(\bar{Y}^{(i)}-\bar{Y})^2$, стоящее под знаком суммы, от j не зависит; последнее слагаемое

$$2\sum_{j=1}^{n_i}(\bar{Y}^{(i)}-\bar{Y})(Y_j^{(i)}-\bar{Y}^{(i)})=2(\bar{Y}^{(i)}-\bar{Y})\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y}^{(i)})=$$

$$=2(\bar{Y}^{(i)}-\bar{Y})\left(\sum_{j=1}^{n_i}Y_j^{(i)}-\sum_{j=1}^{n_i}\bar{Y}^{(i)}\right)=2(\bar{Y}^{(i)}-\bar{Y})\left(\sum_{j=1}^{n_i}Y_j^{(i)}-\bar{Y}^{(i)}n_i\right)=$$

$$\underset{(11.1)}{=}2(\bar{Y}^{(i)}-\bar{Y})(n_i\bar{Y}^{(i)}-\bar{Y}^{(i)}n_i)=0,$$

поэтому равенство (11.25) принимает вид

$$\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y})^2=(\bar{Y}^{(i)}-\bar{Y})^2 n_i+\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y}^{(i)})^2.$$

Просуммировав обе части этого равенства по i от 1 до ν, получим равенство (11.23).

Дисперсионная таблица и проверка гипотезы (11.3) о равенстве групповых математических ожиданий. Введенные показатели вариации запишем в табл. 11.2, которую назовем *таблицей однофакторного дисперсионного анализа.*

Поясним *третий столбец* этой таблицы.

В формулу для S_Φ^2 входят ν слагаемых: $(\bar{Y}^{(1)}-\bar{Y})^2 n_1$, $(\bar{Y}^{(2)}-\bar{Y})^2 n_2$, ..., $(\bar{Y}^{(\nu)}-\bar{Y})^2 n_\nu$, «свобода» изменения которых ограничена одним соотношением, а именно $(\bar{Y}^{(1)}-\bar{Y})n_1+$ $+(\bar{Y}^{(2)}-\bar{Y})n_2+...+(\bar{Y}^{(\nu)}-\bar{Y})n_\nu=0$ (это соотношение вытекает из формулы (11.13)). Поэтому говорят, что величина S_Φ^2 имеет $(\nu-1)$ степень свободы.

Таблица 11.2

Источник вариации результативного признака Y	Показатель вариации	Число степеней свободы	Несмещенная оценка дисперсии σ_o^2
Фактор A	$\hat{\sigma}_\Phi^2=S_\Phi^2/n=\sum_{i=1}^{\nu}(\bar{Y}^{(i)}-\bar{Y})^2 n_i/n$	$\nu-1$	$s_\Phi^2=S_\Phi^2/(\nu-1)$ (при выполнении гипотезы H_0 (11.3))
Остаточные факторы	$\hat{\sigma}_o^2=S_o^2/n=\sum_{i=1}^{\nu}\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y}^{(i)})^2/n$	$n-\nu$	$s_o^2=S_o^2/(n-\nu)$
Общая вариация	$\hat{\sigma}_Y^2=S_Y^2/n=$ $=\sum_{i=1}^{\nu}\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y})^2/n=$ $=\hat{\sigma}_\Phi^2+\hat{\sigma}_o^2$	$n-1=$ $=(\nu-1)+$ $+(n-\nu)$	$s_Y^2=S_Y^2/(n-1)$ (при выполнении гипотезы H_0 (11.3))

В формулу расчета S_o^2 входит $n_1+n_2+...+n_\nu=n$ слагаемых (см. (11.20)). Свобода первых n_1 слагаемых ограничена одним соотношением: $(Y_1^{(1)}-\bar{Y}^{(1)})+(Y_2^{(1)}-\bar{Y}^{(1)})+...+(Y_{n_1}^{(1)}-\bar{Y}^{(1)})=0$, свобода следующих n_2 слагаемых также ограничена одним соотношением: $(Y_1^{(2)}-\bar{Y}^{(2)})+(Y_2^{(2)}-\bar{Y}^{(2)})+...+(Y_{n_2}^{(2)}-\bar{Y}^{(2)})=0$ и т. д. (эти соотношения вытекают из формул групповых средних (11.1)). Таким образом, «свобода» изменения n слагаемых ограничена ν соотношениями. Это означает, что величина S_o^2 имеет $(n-\nu)$ степеней свободы.

И наконец, в формулу для S_Y^2 входят $n_1+n_2+...+n_\nu=n$ слагаемых (см. (11.15)), «свобода» изменения которых ограничена одним соотношением:

$$(Y_1^{(1)}-\bar{Y})+(Y_2^{(1)}-\bar{Y})+...+(Y_{n_1}^{(1)}-\bar{Y})+$$
$$+(Y_1^{(2)}-\bar{Y})+(Y_2^{(2)}-\bar{Y})+...+(Y_{n_2}^{(2)}-\bar{Y})+$$
$$+\cdots\cdots\cdots\cdots\cdots+$$
$$+(Y_1^{(\nu)}-\bar{Y})+(Y_2^{(\nu)}-\bar{Y})+...+(Y_{n_\nu}^{(\nu)}-\bar{Y})=0.$$

Это соотношение вытекает из формулы (11.14). Поэтому S_Y^2 имеет $(n-1)$ степень свободы.

Теперь поясним *четвертый столбец* табл. 11.2. Напомним, что нашей целью была проверка гипотезы (11.3) о равенстве групповых математических ожиданий. Напомним также, что эту гипотезу можно проверить только при соблюдении требований (11.4) и (11.5), включающих в том числе и требование о равенстве генеральных групповых дисперсий.

Так как генеральные групповые дисперсии должны быть равными, т. е. $\sigma_1^2=\sigma_2^2=...=\sigma_\nu^2$, и общее их значение обозначено через σ_o^2 (см. (11.5)), то

$$\sigma_1^2=\sigma_2^2=...=\sigma_\nu^2=\sigma_o^2. \qquad (11.26)$$

Во-первых, убедимся в том, что несмещенная оценка дисперсии σ_o^2 равна отношению $S_o^2/(n-\nu)$, которое обозначим через s_o^2. Напомним, что оценка генеральной характеристики называется несмещенной, если математическое ожидание этой оценки равно генеральной характеристике. Докажем, что

$$Ms_o^2=\sigma_o^2. \qquad (11.27)$$

Находим

$$Ms_o^2=M(S_o^2/(n-\nu))=M\left(\sum_{i=1}^{\nu}\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y}^{(i)})^2/(n-\nu)\right)=$$

$$=\frac{1}{n-\nu}\sum_{i=1}^{\nu}M\left(\sum_{j=1}^{n_i}(Y_j^{(i)}-\bar{Y}^{(i)})^2\right)\underset{(11.2)}{=}\frac{1}{n-\nu}\sum_{i=1}^{\nu}M(n_i\hat{\sigma}_i^2)=$$

$$=\underset{(11.7)}{}\frac{1}{n-\nu}\sum_{i=1}^{\nu} M\left((n_i-1)s_i^2\right)=\frac{1}{n-\nu}\sum_{i=1}^{\nu}(n_i-1)Ms_i^2 \underset{Ms_i^2=\sigma_i^2}{=} \frac{1}{n-\nu}\times$$

$$\times\sum_{i=1}^{\nu}(n_i-1)\sigma_i^2 \underset{(11.26)}{=} \frac{1}{n-\nu}\sum_{i=1}^{\nu}(n_i-1)\sigma_o^2=\frac{1}{n-\nu}\sigma_o^2\sum_{i=1}^{\nu}(n_i-1)=\frac{1}{n-\nu}\sigma_o^2(n-\nu)=\sigma_o^2.$$

Итак, $Ms_o^2=\sigma_o^2$.

Обратим внимание на то, что при доказательстве использовалось равенство $Ms_i^2=\sigma_i^2$, которое читается так: математическое ожидание выборочной дисперсии s_i^2 i-й группы равно генеральной дисперсии σ_i^2 этой группы. Это равенство было доказано в § 9.2 (см. (9.19)); оно верно только в том случае, когда наблюдения в группе независимы и проводятся в одинаковых условиях; оба эти требования, согласно (11.4), выполняются.

Во-вторых, убедимся в том, что при выполнении гипотезы (11.3) оценка s_Y^2 (см. табл. 11.2) является несмещенной оценкой дисперсии σ_o^2.

Итак, допустим, что наряду с условиями (11.4) и (11.5), необходимыми для проведения дисперсионного анализа, *выполняется гипотеза* H_0 (11.3) о равенстве групповых математических ожиданий. Тогда при каждом уровне фактора величина Y будет иметь нормальный закон с одним и тем же математическим ожиданием и одной и той же дисперсией, равной σ_o^2, т. е. переход от уровня к уровню не вносит никаких изменений: имеется одна генеральная совокупность, и результаты наблюдений, приведенные в табл. 11.1,—это выборка объема n из этой генеральной совокупности. А так как наблюдения независимы и проведены в одинаковых условиях, то несмещенная оценка генеральной дисперсии σ_o^2 получается как частное от деления суммы квадратов отклонений n наблюдений от общей средней $\bar{Y}$ на $(n-1)$, т. е. эта оценка равна отношению $S_Y^2/(n-1)$, которое обозначим через s_Y^2. Таким образом,

$$Ms_Y^2=\sigma_o^2. \tag{11.28}$$

И наконец, убедимся в том, что при *выполнении гипотезы* (11.3) и соблюдении требований (11.4) и (11.5) величина $s_\Phi^2=S_\Phi^2/(\nu-1)$ (см. табл. 11.2) является несмещенной оценкой генеральной дисперсии σ_o^2. Для этого докажем, что

$$Ms_\Phi^2=\sigma_o^2. \tag{11.29}$$

Имеем

$$Ms_\Phi^2=M\left(S_\Phi^2/(\nu-1)\right)=\frac{1}{\nu-1}MS_\Phi^2 \underset{(11.22)}{=} \frac{1}{\nu-1}M\left(S_Y^2-S_o^2\right)=$$

$$=\frac{1}{\nu-1}\left(MS_Y^2-MS_o^2\right) \underset{\text{табл. 11.2}}{=} \frac{1}{\nu-1}\left[M\left((n-1)s_Y^2\right)-M\left((n-\nu)s_o^2\right)\right]=$$

$$=\frac{1}{\nu-1}\left[(n-1)Ms_Y^2-(n-\nu)Ms_o^2\right] \underset{\substack{(11.28)\\(11.27)}}{=} \frac{1}{\nu-1}\left[(n-1)\sigma_o^2-(n-\nu)\sigma_o^2\right]=$$

$$=\frac{1}{\nu-1}(\nu-1)\sigma_o^2=\sigma_o^2.$$

Таким образом, $Ms_\Phi^2=\sigma_o^2$.

Итак, имеются три несмещенные оценки одной и той же дисперсии σ_o^2, причем s_o^2 является несмещенной оценкой

в любом случае, а s_Φ^2 и s_Y^2—только при выполнении гипотезы (11.3), т. е. только в том случае, когда фактор A не влияет на результативный признак.

Проверка гипотезы H_0 (11.3) о равенстве групповых математических ожиданий основывается на сравнении дисперсии s_Φ^2 и s_o^2 (см. § 10.5). Оказывается, что если гипотеза (11.3) верна, то величина

$$F=s_\Phi^2/s_o^2 \tag{11.30}$$

имеет F-распределение с числом степеней свободы $l=\nu-1$ и $k=n-\nu$, т. е.

$$s_\Phi^2/s_o^2=F\left(l=\nu-1;\ k=n-\nu\right). \tag{11.31}$$

Используя соотношение (11.31), найдем для заданного уровня значимости α правостороннюю критическую точку $x_{\text{пр},\alpha}^{\text{кр}}$ по следующей схеме:

$$\left.\begin{array}{l}\alpha\to\gamma=1-\alpha\\ n,\ \nu\begin{array}{l}\to l=\nu-1\\ \to k=n-\nu\end{array}\end{array}\right\}\xrightarrow{\text{П.7}} f_\gamma\to x_{\text{пр},\alpha}^{\text{кр}}=f_\gamma. \tag{11.32}$$

Обратим внимание на то, что (11.30), (11.31) и (11.32) аналогичны соответственно (10.88), (10.89) и (10.93).

Если числовое значение $F_{\text{чис}}$ величины (11.30) попадает в интервал $(x_{\text{пр},\alpha}^{\text{кр}};\ +\infty)$, то гипотезу H_0 о равенстве групповых математических ожиданий отвергаем: считаем, что фактор A влияет на результативный признак Y. Если же $F_{\text{чис}}<x_{\text{пр},\alpha}^{\text{кр}}$, то гипотезу о равенстве групповых средних не отвергаем; в этом случае говорят, что влияние фактора A на признак Y не подтвердилось выборочными наблюдениями*).

Замечание. При проверке гипотезы (11.3) используются именно величины s_Φ^2 и s_o^2. Это объясняется тем, что s_Φ^2 и s_o^2 независимы, тогда как величины s_Φ^2 и s_Y^2 (см. табл. 11.2) зависимы, также зависимы и величины s_o^2 и s_Y^2.

Независимость же величин s_Φ^2 и s_o^2 используется при доказательстве соотношения (11.31). Действительно, так как $s_\Phi^2=S_\Phi^2/(\nu-1)$ и $s_o^2=S_o^2/(n-\nu)$ являются несмещенными оценками дисперсии σ_o^2, то имеют место следующие соотношения, аналогичные соотношению (9.64):

$$\frac{(\nu-1)s_\Phi^2}{\sigma_o^2}=\chi^2(l=\nu-1)\ \text{ и }\ \frac{(n-\nu)s_o^2}{\sigma_o^2}=\chi^2(k=n-\nu). \tag{11.33}$$

Поскольку s_Φ^2 и s_o^2 независимы, независимы и величины $\chi^2(l=\nu-1)$ и $\chi^2(k=n-\nu)$. Следовательно, согласно (6.37), отношение $\dfrac{\chi^2(l=\nu-1)/(\nu-1)}{\chi^2(k=n-\nu)/(n-\nu)}$

*) Здесь и далее при использовании критерия, имеющего F-распределение, строится правосторонняя критическая область. Это объясняется тем, что в дисперсионном анализе, как правило, числитель критерия больше знаменателя ($s_\Phi^2>s_0$). Если это не так, то считают, что наблюдения не подтверждают влияние фактора на признак.

или равное ему отношение s_Φ^2/s_o^2 имеют F-распределение с числом степеней свободы $l=\nu-1$ и $k=n-\nu$.

Поскольку вывод о том, влияет или нет фактор на результативный признак, основан на сопоставлении дисперсий, рассмотренный метод называют *дисперсионным анализом*.

Допустим, что фактор A влияет на результативный признак. Для измерения степени влияния используют *выборочный коэффициент детерминации*

$$\hat{\rho}^2=\frac{\hat{\sigma}_\Phi^2}{\hat{\sigma}_Y^2}=\frac{\sum\limits_{i=1}^{\nu}(\bar{Y}^{(i)}-\bar{Y})^2 n_i/n}{\hat{\sigma}_Y^2}, \tag{11.34}$$

который показывает, какую долю выборочной дисперсии $\hat{\sigma}_Y^2$ «игреков» составляет дисперсия групповых средних «игреков», или, иначе говоря, какая доля дисперсии $\hat{\sigma}_Y^2$ объясняется зависимостью «игрека» от фактора A.

Обратим внимание на то, что в силу (11.21): $\hat{\rho}^2\leqslant 1$.

Задача 11.1. Выяснить при уровне значимости $\alpha=0{,}05$, зависит ли объем работ, выполненных на стройке за смену, от работающей бригады. Данные по четырем бригадам приведены в табл. 11.3.

Таблица 11.3

Номер бригады i	Объем выполненной работы				Число смен n_i	Групповое среднее $\bar{Y}^{(i)}$	Выборочная дисперсия $\hat{\sigma}_i^2$
1	140	144	142	145	4	142,75	3,69
2	150	149	152	150	4	150,25	1,19
3	148	149	146	147	4	147,5	1,25
$4=\nu$	150	155	154	152	4	152,75	3,69

$\triangle$ Прежде чем проводить дисперсионный анализ, следует убедиться в том, что выполняются требования (11.4) и (11.5). Будем предполагать, что требование (11.4) выполняется, т. е. по каждой отдельно взятой бригаде сведения собирались в одинаковых условиях (конечно, примерно одинаковых, поскольку все условия зафиксировать нельзя), и что результаты любых выработок не зависят друг от друга. Что касается требования (11.5), то предположим, что объем работ, выполненных любой отдельно взятой бригадой,— величина случайная с нормальным законом распределения, и убедимся в том, что гипотеза

$$H_0: \sigma_1^2=\sigma_2^2=\sigma_3^2=\sigma_4^2 \tag{11.35}$$

о равенстве групповых дисперсий не противоречит результатам наблюдений. Для этого:

1. По формуле (11.7) найдем

$$s_1^2=\frac{\hat{\sigma}_1^2 n_1}{n_1-1}=\frac{3{,}69\cdot 4}{3}=4{,}92;\quad s_2^2=\frac{\hat{\sigma}_2^2 n_2}{n_2-1}=\frac{1{,}19\cdot 4}{3}=1{,}58;$$

$$s_3^2=1{,}67;\quad s_4^2=4{,}92;$$

2. С помощью формулы (11.8) вычислим

$$s_o^2=\frac{(n_1-1)s_1^2+\ldots+(n_4-1)s_4^2}{(n_1-1)+\ldots+(n_4-1)}=\frac{3\cdot 4{,}92+3\cdot 1{,}58+3\cdot 1{,}67+3\cdot 4{,}92}{3+3+3+3}=3{,}27;$$

3. По формулам (11.9) и (11.10) определим

$$q=\left[1+\frac{1}{3(4-1)}\left(\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}-\frac{1}{3+3+3+3}\right)\right]^{-1}=0{,}878;$$

$$\varphi=0{,}878\left(3\ln\frac{3{,}27}{4{,}92}+3\ln\frac{3{,}27}{1{,}58}+3\ln\frac{3{,}27}{1{,}67}+3\ln\frac{3{,}27}{4{,}92}\right)=1{,}538.$$

Поскольку $n_i>3$, $i=1, 2, 3, 4$, можно перейти к п. 4.

4. Пусть $\alpha=0{,}05$. По схеме (11.12) найдем

$$\left.\begin{aligned}\alpha=0{,}05\to\gamma=1-\alpha=0{,}95\\ \nu=4\to k=\nu-1=3\end{aligned}\right\}\xrightarrow{\text{П.6}}\chi_\gamma^2=7{,}82\to x_{\text{кр},\alpha}^{\text{кр}}=7{,}82.$$

5. Так как $\varphi_{\text{числ}}=1{,}538$ не попадает в критическую область $(7{,}82;\ \infty)$, то гипотезу H_0 (11.35) о равенстве групповых генеральных дисперсий не отвергаем.

Теперь запишем гипотезу, проверяемую дисперсионным анализом; это гипотеза

$$H_0: a_1=a_2=a_3=a_4, \tag{11.36}$$

где a_i — математическое ожидание ежедневного объема работы, который может быть выполнен i-й бригадой, и приступим к собственно дисперсионному анализу.

Вычислим:

средний объем выполненных работ по всем 16 наблюдениям:

$$\bar{Y}=\frac{140+144+142+\ldots+154+152}{16}=148{,}31;$$

общую вариацию признака Y, используя формулы (11.15) и (11.16):

$$S_Y^2=(140-148{,}31)^2+(144-148{,}31)^2+\ldots+(154-148{,}31)^2+(152-148{,}31)^2=259{,}46;$$

$$\hat{\sigma}_Y^2=\frac{259{,}46}{16};$$

факторную вариацию, используя формулы (11.17) и (11.18):

$$S_\Phi^2=(142{,}75-148{,}31)^2\cdot 4+(150{,}25-148{,}31)^2\cdot 4+(147{,}5-148{,}31)^2\cdot 4+$$
$$+(152{,}75-148{,}31)^2\cdot 4=220{,}19;\quad \hat{\sigma}_\Phi^2=220{,}19/16;$$

остаточную вариацию, используя формулу (11.19):

$$S_o^2=3{,}69\cdot 4+1{,}19\cdot 4+1{,}25\cdot 4+3{,}69\cdot 4=39{,}27;\quad \hat{\sigma}_o^2=39{,}27/16.$$

Составим для этого примера дисперсионную табл. 11.4 типа табл. 11.2.

Таблица 11.4

Источник вариации объема работы	Показатель вариации	Число степеней свободы	Несмещенная оценка дисперсии σ_o^2
Работающая бригада	$\hat{\sigma}_\Phi^2=S_\Phi^2/n=220{,}19/16$	$4-1=3$	$s_\Phi^2=220{,}19/3=73{,}4$ (при выполнении гипотезы H_0 (11.36))
Остаточные факторы	$\hat{\sigma}_o^2=S_o^2/n=39{,}27/16$	$16-4=12$	$s_o^2=39{,}27/12=3{,}27$
Общая вариация	$\hat{\sigma}_Y^2=S_Y^2/n=259{,}46/16$	$16-1=15$	$s_Y^2=259{,}46/15=17{,}30$ (при выполнении гипотезы H_0 (11.36))

Найдем числовое значение критерия (11.30). Имеем

$$F_{\text{чис}} = s_{\phi}^2/s_o^2 = 73{,}4/3{,}27 = 22{,}45.$$

Теперь по схеме (11.32) определим правостороннюю критическую точку:

$$\begin{array}{l} \alpha=0{,}05 \quad \rightarrow \quad \gamma = 1-\alpha = 0{,}95 \\ n=16,\ \nu=4 \rightarrow \begin{cases} l=\nu-1=3 \\ k=n-\nu=12 \end{cases} \end{array} \Bigg\} \rightarrow f_{\gamma} = 3{,}49 \rightarrow x^{\text{кр}}_{\text{пр},\alpha} = 3{,}49.$$

Так как $F_{\text{чис}} = 22{,}45$ попадает в критическую область $(3{,}49;\ \infty)$, то гипотезу H_o (11.36) отвергаем, т. е. считаем, что объем ежедневной выработки зависит от работающей бригады. Оценим степень этой зависимости. Найдем выборочный коэффициент детерминации (11.34):

$$\hat{\rho}^2 = \hat{\sigma}_{\phi}^2/\hat{\sigma}_Y^2 = 0{,}849.$$

Таким образом, 84,9% общей вариации ежедневного объема выработки связано с работающей сменой. ▲

Итак, однофакторный дисперсионный анализ позволяет по выборочным данным выяснить, влияет ли контролируемый фактор на результативный признак, и при наличии такого влияния оценить его степень.

Модель однофакторного дисперсионного анализа. Очень часто при проведении дисперсионного анализа строят модель наблюдаемых значений результативного признака.

При исследовании влияния одного фактора модель имеет следующий вид:

$$Y_j^{(i)} = a + \lambda^{(i)} + \varepsilon_j^{(i)}, \quad \begin{array}{l} i=1,\ 2,\ \ldots,\ \nu, \\ j=1,\ 2,\ \ldots,\ n_i, \end{array} \tag{11.37}$$

где $Y_j^{(i)}$ — значение результативного признака Y, которое может быть зафиксировано в j-м наблюдении при i-м уровне фактора A (напомним, что случайные величины $Y_j^{(i)}$ должны удовлетворять требованиям (11.4) и (11.5)); a — генеральное среднее всех мыслимых результатов наблюдений, или, иначе говоря, математическое ожидание результативного признака Y; $\lambda^{(i)}$ — генеральный эффект влияния на Y, вызванный i-м уровнем фактора A, или, иначе, отклонение математического ожидания a_i результативного признака при i-м уровне фактора от общего математического ожидания a; т. е. $\lambda^{(i)} = a_i - a$; $\varepsilon_j^{(i)}$ — случайный остаток, отражающий влияние на величину $Y_j^{(i)}$ всех прочих неконтролируемых факторов.

Можно доказать, что при выполнении требований (11.4) и (11.5); остаток

$$\varepsilon_j^{(i)} = N(0,\ \sigma_o), \quad \begin{array}{l} i=1,\ 2,\ \ldots,\ \nu, \\ j=1,\ 2,\ \ldots,\ n_i, \end{array} \tag{11.38}$$

где σ_o^2 — общее значение групповых генеральных дисперсий (см. (11.26)), и все n величин $\varepsilon_j^{(i)}$ независимы между собой.

Обратим внимание на то, что параметрами модели (11.37) являются математическое ожидание a, эффекты $\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(\nu)}$

влияния уровней фактора и дисперсия σ_o^2 случайного остатка.

Если гипотеза H_o (11.3) о равенстве групповых математических ожиданий не отвергается, т. е. влияние фактора A не подтверждается, то в модели (11.37) эффекты влияния уровней фактора будут нулевыми, т. е. $\lambda^{(1)} = \lambda^{(2)} = \ldots = \lambda^{(\nu)} = 0$; остальные параметры по результатам наблюдений, приведенным в табл. 11.1, оцениваются следующим образом:

оценка параметра a равна общему среднему $\bar{Y}$, вычисляемому по формуле (11.14);

оценка параметра σ_o^2 равна s_o^2 (см. табл. 11.2).

Если же гипотеза H_o (11.3) отвергается, т. е. влияние фактора A подтверждается по результатам наблюдений, то оценки параметров в модели (11.37) таковы:

оценка параметра a равна общему среднему $\bar{Y}$, вычисляемому по формуле (11.14);

оценка эффекта $\lambda^{(i)}$ влияния i-го уровня фактора равна $(\bar{Y}^{(i)} - \bar{Y})$, где $\bar{Y}^{(i)}$ — групповое среднее при i-м уровне фактора A, вычисляемое по формуле (11.1);

оценка параметра σ_o^2 равна s_o^2 (см. табл. 11.2).

Задача 11.2. По данным задачи 11.1 найти оценки параметров модели (11.37) однофакторного дисперсионного анализа.

△ В задаче 11.1 было выяснено, что объем ежедневной выработки зависит от работающей смены, поэтому:

оценка параметра a равна $\bar{Y}$, т. е. $a \approx 148{,}31$;

оценка эффекта $\lambda^{(i)}$ равна $(\bar{Y}^{(i)} - \bar{Y})$, $i = 1, 2, 3, 4$, т. е.

$$\begin{aligned} \lambda^{(1)} &\approx 142{,}75 - 148{,}31 = -5{,}56, \\ \lambda^{(2)} &\approx 150{,}25 - 148{,}31 = 1{,}94, \\ \lambda^{(3)} &\approx 147{,}5 - 148{,}31 = -0{,}81, \\ \lambda^{(4)} &\approx 152{,}75 - 148{,}31 = 4{,}44; \end{aligned}$$

оценка параметра σ_o^2 равна s_o^2, т. е. $\sigma_o^2 \approx 3{,}27$.

Отсюда, учитывая (11.37) и (11.38), получаем, что для ежедневного объема выработки $Y^{(1)}$ 1-й бригады имеет место следующее приближенное равенство:

$$Y^{(1)} \approx 148{,}31 - 5{,}56 + N(0;\ \sqrt{3,\ 27}), \tag{11.39}$$

где $N(0;\ \sqrt{3,\ 27})$ — нормально распределенная случайная величина с нулевым математическим ожиданием и дисперсией, равной 3,27. Вопрос о том, как получить числовые значения этой величины, рассматривается в § 13.6.

Для ежедневного объема выработки $Y^{(2)}$ 2-й бригады имеет место следующее равенство:

$$Y^{(2)} \approx 148{,}31 + 1{,}94 + N(0,\ \sqrt{3,\ 27}) \text{ и т. д. } \blacktriangle$$

§ 11.2. Двухфакторный дисперсионный анализ с одним наблюдением в клетке

Задача двухфакторного дисперсионного анализа и предварительная обработка результатов наблюдений. Допустим, что исследователя интересует зависимость результативного призна-

Таблица 11.5

Уровни фактора B		B_1	B_2	....	B_{ν_B}	Среднее групповое (по строке)
Уровни фактора A	j / i	1	2	...	ν_B	
A_1	1	Y_{11}	Y_{12}	...	$Y_{1\nu_B}$	$\bar{Y}_{1\cdot}$
A_2	2	Y_{21}	Y_{22}	...	$Y_{2\nu_B}$	$\bar{Y}_{2\cdot}$
....	....	....	....	....	....	....
A_{ν_A}	ν_A	$Y_{\nu_A 1}$	$Y_{\nu_A 2}$	...	$Y_{\nu_A \nu_B}$	$\bar{Y}_{\nu_A\cdot}$
Среднее групповое (по столбцу)		$\bar{Y}_{\cdot 1}$	$\bar{Y}_{\cdot 2}$		$\bar{Y}_{\cdot \nu_B}$	

ка Y, который является случайной величиной, от двух факторов A и B. Например, требуется выяснить, зависит ли качество обрабатываемых деталей от типа станка и вида сырья, из которого она изготовляется. Это типичная задача двухфакторного дисперсионного анализа.

Обозначим через $A_1, A_2, \ldots, A_{\nu_A}$ уровни фактора A (типы станков); всего этих уровней ν_A. Через $B_1, B_2, \ldots, B_{\nu_B}$ обозначим уровни фактора B (виды сырья); всего этих уровней ν_B. Как и в однофакторном анализе, будем считать уровни фиксированными величинами. Количество различных комбинаций уровней факторов A и B равно $\nu_A \nu_B$.

Решение задачи двухфакторного дисперсионного анализа зависит от количества проведенных наблюдений при каждой комбинации уровней факторов, или, иначе говоря, в каждой клетке двухфакторного комплекса. Рассмотрим случай, когда в каждой клетке по одному наблюдению.

Результаты наблюдений и результаты их предварительной обработки расположим в табл. 11.5.

В табл. 11.5 Y_{11} — результат наблюдения, зафиксированный на 1-м уровне фактора A и 1-м уровне фактора B; Y_{12} — результат на 1-м уровне фактора A и 2-м уровне фактора B и т. д. Очевидно, что общее число наблюдений

$$n = \nu_A \nu_B.$$

Групповое среднее (по строке) получено как частное от деления суммы результатов наблюдений в строке на число наблюдений в строке, которое равно ν_B. Например, $\bar{Y}_{1\cdot} = (Y_{11} + Y_{12} + \ldots + Y_{1\nu_B})/\nu_B$ (точка, стоящая на месте второго индекса в обозначении средней, означает, что у складываемых в числителе «игреков» изменяется второй индекс).

Групповое среднее (по столбцу) получено как частное от деления суммы результатов наблюдений в столбце на число

наблюдений в столбце, которое равно ν_A. Например, $\bar{Y}_{\cdot 1} = (Y_{11} + Y_{21} + \ldots + Y_{\nu_A 1})/\nu_A$ (здесь у складываемых «игреков» изменяется первый индекс, поэтому при обозначении средней точка стоит на месте первого индекса).

Предварительное суждение о том, зависит ли признак Y от фактора A, можно сделать, сравнив групповые средние $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{\nu_A\cdot}$, расположенные в строках. Если различие между ними существенно, то, по-видимому, такая зависимость имеется. Аналогично, предварительное суждение о том, влияет или нет фактор B, можно сделать сравнив групповые средние $\bar{Y}_{\cdot 1}, \bar{Y}_{\cdot 2}, \ldots, \bar{Y}_{\cdot \nu_B}$, расположенные в столбцах. Однако не следует забывать, что в табл. 11.5 представлены выборочные наблюдения, а ответ на вопрос, существует или нет зависимость Y от фактора A и B, должен быть дан применительно к генеральной совокупности. Поэтому логика рассуждений должна быть следующей.

Обозначим через a_i математическое ожидание величины Y при уровне A_i, $i = 1, 2, \ldots, \nu_A$; через b_j — математическое ожидание величины Y при уровне B_j, $j = 1, 2, \ldots, \nu_B$. Если при изменении уровня фактора A сохраняется равенство $a_1 = a_2 = \ldots = a_{\nu_A}$, то считаем, что Y не зависит от фактора A. В противном случае Y зависит от фактора A. Аналогично, если при изменении уровня фактора B сохраняется равенство $b_1 = b_2 = \ldots = b_{\nu_B}$, то считаем, что Y не зависит от фактора B. Но так как числовые значения математических ожиданий неизвестны, то встает задача проверки следующих гипотез:

$$H_A: a_1 = a_2 = \ldots = a_{\nu_A}; \tag{11.40}$$

$$H_B: b_1 = b_2 = \ldots = b_{\nu_B}. \tag{11.41}$$

Обратим внимание на то, что проверить эти гипотезы можно только в том случае, если:

при различных сочетаниях уровней факторов A и B наблюдения независимы,

при каждом сочетании уровней факторов A и B результативный признак имеет нормальный закон с постоянной для различных сочетаний дисперсий (эту дисперсию обозначим σ_0^2). (11.42)

Разложение дисперсии наблюдаемых «игреков». Обратимся снова к табл. 11.5. Изменчивость, или вариация, наблюдаемых «игреков» может быть вызвана изменчивостью уровней фактора A, фактора B и изменчивостью случайных неконтролируемых факторов, называемых *остаточными*.

1) Вариация «игреков», вызванная влиянием и фактора A, и фактора B и влиянием остаточных факторов, измеряется выборочной дисперсией $\hat{\sigma}_Y^2$ величины Y. Прежде чем привести формулу для $\hat{\sigma}_Y^2$, вычислим среднее $\bar{Y}$ всех наблюдений,

которое называют *общим средним* признака Y. Его можно вычислить по одной из следующих формул:

$$\bar{Y}=(Y_{11}+Y_{12}+\dots+Y_{\nu_A\nu_B})/n=(\bar{Y}_{1\cdot}\nu_B+\bar{Y}_{\cdot 2}\nu_B+\dots+\bar{Y}_{\nu_A\cdot}\nu_B)/n=$$
$$=(\bar{Y}_{\cdot 1}\nu_A+\bar{Y}_{\cdot 2}\nu_A+\dots+\bar{Y}_{\cdot\nu_B}\nu_A)/n. \qquad (11.43)$$

Тогда

$$\hat{\sigma}_Y^2=S_Y^2/n, \qquad (11.44)$$

где

$$S_Y^2=(Y_{11}-\bar{Y})^2+(Y_{12}-\bar{Y})^2+\dots+(Y_{\nu_A\nu_B}-\bar{Y})^2=\sum_{i=1}^{\nu_A}\sum_{j=1}^{\nu_B}(Y_{ij}-\bar{Y})^2. \qquad (11.45)$$

2) Вариация наблюдаемых «игреков», вызванная изменчивостью уровней фактора A, измеряется *выборочной дисперсией* $\hat{\sigma}^2_{\bar{Y}_{i\cdot}}$ *групповых средних* $\bar{Y}_{i\cdot}$, расположенных в строках; обозначим ее $\hat{\sigma}_A^2$. Имеем

$$\hat{\sigma}_A^2=S_A^2/n, \qquad (11.46)$$

где

$$S_A^2=(\bar{Y}_{1\cdot}-\bar{Y})^2\nu_B+(\bar{Y}_{2\cdot}-\bar{Y})^2\nu_B+\dots+(\bar{Y}_{\nu_A\cdot}-\bar{Y})^2\nu_B=$$
$$=\sum_{i=1}^{\nu_A}(\bar{Y}_{i\cdot}-\bar{Y})^2\nu_B. \qquad (11.47)$$

Обратим внимание на то, что в формулу (11.47) введены «веса», равные ν_B, так как каждая групповая средняя в строке является «представителем» ν_B наблюдений.

3) Аналогично (11.46) и (11.47), вариация «игреков», вызванная изменчивостью уровней фактора B, будет измеряться *выборочной дисперсией* $\hat{\sigma}^2_{\bar{Y}_{\cdot j}}$ *групповых средних* $\bar{Y}_{\cdot j}$, расположенных в столбцах; обозначим ее $\hat{\sigma}_B^2$. Имеем

$$\hat{\sigma}_B^2=S_B^2/n, \qquad (11.48)$$

где $$S_B^2=(\bar{Y}_{\cdot 1}-\bar{Y})^2\nu_A+(\bar{Y}_{\cdot 2}-\bar{Y})^2\nu_A+\dots+(\bar{Y}_{\cdot\nu_B}-\bar{Y})^2\nu_A=$$
$$=\sum_{j=1}^{\nu_B}(\bar{Y}_{\cdot j}-\bar{Y})^2\nu_A. \qquad (11.49)$$

4) Кроме контролируемых факторов A и B на результативный признак Y влияют случайные неконтролируемые факторы — остаточные факторы. Чтобы ответить на вопрос, как измерить вариацию, вызванную этими факторами, составим следующее тождество:

$$Y_{ij}-\bar{Y}=(\bar{Y}_{i\cdot}-\bar{Y})+(\bar{Y}_{\cdot j}-\bar{Y})+(Y_{ij}-\bar{Y}_{\cdot i}-\bar{Y}_{\cdot j}+\bar{Y}). \qquad (11.50)$$

Стоящая в левой части равенства разность $(Y_{ij}-\bar{Y})$ входит в формулу (11.45) расчета общей вариации; первое из слагаемых $(\bar{Y}_{i\cdot}-\bar{Y})$ в правой части тождества (11.50) входит в формулу (11.47) расчета влияния фактора A; слагаемое $(\bar{Y}_{\cdot j}-\bar{Y})$ входит в формулу (11.49), измеряющую влияние фактора B. Поэтому третье слагаемое $(Y_{ij}-\bar{Y}_{i\cdot}-\bar{Y}_{\cdot j}+\bar{Y})$, введенное в тождество (11.50), отражает влияние на Y остаточных факторов. Это влияние измеряется величиной

$$\hat{\sigma}_o^2=S_o^2/n, \qquad (11.51)$$

где

$$S_o^2=\sum_{i=1}^{\nu_A}\sum_{j=1}^{\nu_B}(Y_{ij}-\bar{Y}_{i\cdot}-\bar{Y}_{\cdot j}+\bar{Y})^2. \qquad (11.52)$$

Можно доказать, что

$$\hat{\sigma}_Y^2=\hat{\sigma}_A^2+\hat{\sigma}_B^2+\hat{\sigma}_o^2. \qquad (11.53)$$

Для этого следует:
возвести части тождества (11.50) в квадрат;
просуммировать сначала по j от 1 до ν_B, а затем по i от 1 до ν_A;
убедиться, что все суммы с удвоенными произведениями равны нулю;
обе части полученного тождества разделить на n.

Дисперсионная таблица и проверка гипотез (11.40) и (11.41). Введенные показатели вариации запишем в дисперсионную табл. 11.6.

Таблица 11.6

Источник вариации результативного признака Y	Показатель вариации	Число степеней свободы	Несмещенная оценка дисперсии σ_o^2
Фактор A	$\hat{\sigma}_A^2=S_A^2/n$ (см. (11.46))	ν_A-1	$s_A^2=S_A^2/(\nu_A-1)$ (при выполнении гипотезы H_A)
Фактор B	$\hat{\sigma}_B^2=S_B^2/n$ (см. (11.48))	ν_B-1	$s_B^2=S_B^2/(\nu_B-1)$ (при выполнении гипотезы H_B)
Остаточные факторы	$\hat{\sigma}_o^2=S_o^2/n$ (см. (11.51))	$(\nu_A-1)\times$ $\times(\nu_B-1)$	$s_o^2=S_o^2/((\nu_A-1)(\nu_B-1))$
Общая вариация	$\hat{\sigma}_Y^2=S_Y^2/n$ (см. (11.44))	$n-1$	$s_Y^2=S_Y^2/(n-1)$ (при выполнении гипотез H_A и H_B)

Проверка гипотезы H_A (11.40) основывается на сравнении величин s_A^2 и s_o^2. Если гипотеза H_A (11.40) верна, то величина

$$F^{(A)}=s_A^2/s_o^2 \tag{11.54}$$

имеет F-распределение с числами степеней свободы $l=\nu_A-1$ и $k=(\nu_A-1)(\nu_B-1)$, т. е.

$$s_A^2/s_o^2=F(l=\nu_A-1,\ k=(\nu_A-1)(\nu_B-1)).$$

Используя последнее соотношение, найдем для заданного уровня значимости α, правостороннюю критическую точку $x_{\text{пр},\alpha}^{\text{кр}}$ по следующей схеме:

$$\left.\begin{array}{l}\alpha\to\gamma=1-\alpha\\ \nu_A,\ \nu_B - \begin{cases} l=\nu_A-1\\ k=(\nu_A-1)(\nu_B-1)\end{cases}\end{array}\right\} \xrightarrow{\text{п.7}} f_\gamma\to x_{\text{пр},\alpha}^{\text{кр}}=f_\gamma. \tag{11.55}$$

Если числовое значение $F_{\text{чис}}^{(A)}$ величины (11.54) попадает в интервал $(x_{\text{пр},\alpha}^{\text{кр}}, +\infty)$, то гипотезу (11.40) отвергаем и считаем, что фактор A влияет на результативный признак. Степень этого влияния по результатам наблюдений измеряется выборочным коэффициентом детерминации

$$\hat\rho_A^2=\hat\sigma_A^2/\hat\sigma_Y^2, \tag{11.56}$$

который показывает, какая доля дисперсии результативного признака в выборке обусловлена влиянием на него фактора A. Если же $F_{\text{чис}}^{(A)}<x_{\text{пр},\alpha}^{\text{кр}}$, то гипотезу (11.40) не отвергают и считают, что влияние фактора A не подтвердилось.

Аналогично проверяется гипотеза H_B (11.41) о влиянии фактора B. Если эта гипотеза выполняется, то величина

$$F^{(B)}=s_B^2/s_o^2 \tag{11.57}$$

имеет F-распределение с числами степеней свободы $l=\nu_B-1$ и $k=(\nu_A-1)(\nu_B-1)$. Найдем критическую точку $x_{\text{пр},\alpha}^{\text{кр}}$ для заданного уровня значимости α:

$$\left.\begin{array}{l}\alpha\to\gamma=1-\alpha\\ \nu_A,\ \nu_B - \begin{cases} l=\nu_B-1\\ k=(\nu_A-1)(\nu_B-1)\end{cases}\end{array}\right\} \xrightarrow{\text{п.7}} f_\gamma\to x_{\text{пр},\alpha}^{\text{кр}}=f_\gamma. \tag{11.58}$$

Если значение $F_{\text{чис}}^{(B)}$ величины (11.57) попадает в интервал $(x_{\text{пр},\alpha}^{\text{кр}}, +\infty)$, то гипотезу H_B (11.41) отвергают и считают, что фактор B влияет на результативный признак. Степень этого влияния по результатам наблюдений измеряется выборочным коэффициентом детерминации

$$\hat\rho_B^2=\hat\sigma_B^2/\hat\sigma_Y^2, \tag{11.59}$$

который показывает, какая доля дисперсии результативного признака в выборке обусловлена влиянием на него фактора B. Если же $F_{\text{чис}}^{(B)}<x_{\text{пр},\alpha}^{\text{кр}}$, то гипотезу H_B (11.41) не отвергают и считают, что влияние фактора B не подтвердилось.

Модель двухфакторного дисперсионного анализа с одним наблюдением в клетке. Как отмечалось в предыдущем параграфе, очень часто при проведении дисперсионного анализа строят модель наблюдаемых значений результативного признака. Для двухфакторного комплекса с одним наблюдением в ячейке модель имеет следующий вид:

$$Y_{ij}=a+\alpha_i+\beta_j+\varepsilon_{ij},\quad i=1, 2, ..., \nu_A,\quad j=1,\ 2, ..., \nu_B, \tag{11.60}$$

где Y_{ij} — значение результативного признака Y, которое может быть зафиксировано при i-м уровне фактора A и j-м уровне фактора B; a — генеральное среднее всех мыслимых результатов наблюдений, или, иначе говоря, математическое ожидание результативного признака Y; α_i — генеральный эффект влияния на Y, вызванный i-м уровнем фактора A, или, иначе говоря, отклонение математического ожидания a_i результативного признака при i-м уровне фактора A от общего математического ожидания, т. е. $\alpha_i=a_i-a$, $i=1, 2, ..., \nu_A$; β_j — генеральный эффект влияния на Y, вызванный j-м уровнем фактора B, или, иначе говоря, отклонение математического ожидания b_j результативного признака при j-м уровне фактора B от общего математического ожидания, т. е. $\beta_j=b_j-a$, $j=1, 2, ..., \nu_B$; ε_{ij} — случайный остаток, отражающий влияние на Y всех прочих неконтролируемых факторов.

Можно показать, что при выполнении требований (11.42) остаток

$$\varepsilon_{ij}=N(0, \sigma_o^2),\quad i=1, 2, ..., \nu_A,\quad j=1, 2, ..., \nu_B \tag{11.61}$$

и все $n=\nu_A\nu_B$ величин ε_{ij} независимы.

Параметрами модели (11.60) являются: математическое ожидание a; эффекты $\alpha_1, \alpha_2, ..., \alpha_{\nu_A}$ влияния уровней фактора A; эффекты $\beta_1, \beta_2, ..., \beta_{\nu_B}$ влияния уровней фактора B и дисперсия σ_o^2.

Если гипотезы H_A (11.40) и H_B (11.41) не отвергаются, т. е. влияние фактора A и B не подтверждается, то в модели (11.60) эффекты влияния уровней факторов будут нулевыми, т. е. $\alpha_1=\alpha_2=...=\alpha_{\nu_A}=0$ и $\beta_1=\beta_2=...=\beta_{\nu_B}=0$, а остальные параметры по результатам наблюдений, представленным в табл. 11.5, оцениваются следующим образом:

оценка параметра a равна среднему $\bar Y$, вычисляемому по формулам (11.43);

оценка параметра σ_o^2 равна s_o^2 (см. табл. 11.6).

Если гипотезы H_A (11.40) и H_B (11.41) отвергаются, т. е. влияние и фактора A и фактора B подтверждается, то в модели (11.60):

Таблица 11.7

Тип машины (уровень фактора A) i	Виды сырья (уровень фактора B) j	B_1 1	B_2 $2=\nu_B$	Среднее групповое по строке
A_1	1	$10=Y_{11}$	$50=Y_{12}$	$30=\bar{Y}_{1\cdot}$
A_2	2	$20=Y_{21}$	$60=Y_{22}$	$40=\bar{Y}_{2\cdot}$
A_3	$3=\nu_A$	$30=Y_{31}$	$100=Y_{32}$	$65=\bar{Y}_{3\cdot}$
Среднее групповое по столбцу		$20=\bar{Y}_{\cdot 1}$	$70=\bar{Y}_{\cdot 2}$	

оценка параметра a равна $\bar{Y}$;

оценка эффекта α_i равна $\bar{Y}_{i\cdot}-\bar{Y}$, где $\bar{Y}_{i\cdot}$ — групповое среднее по строке при i-м уровне фактора A, $i=1, 2, ..., \nu_A$;

оценка эффекта β_j равна $\bar{Y}_{\cdot j}-\bar{Y}$, где $\bar{Y}_{\cdot j}$ — групповое среднее по столбцу при j-м уровне фактора B, $j=1, 2, ..., \nu_B$;

оценка параметра σ_o^2 равна s_o^2 (см. табл. 11.6).

Задача 11.3. Выяснить при уровне значимости $\alpha=0{,}05$, влияют ли на качество пряжи, измеряемое величиной разрывной нагрузки, тип машины и вид сырья, из которого пряжа производится. Необходимые данные помещены в табл. 11.7.

В табл. 11.7 для каждого сочетания типа станка и виды сырья указана нагрузка, при которой пряжа разрывается. Число уровней фактора A равно $\nu_A=3$. Число уровней фактора B равно $\nu_B=2$. Общее число наблюдений $n=6$.

△ Будем предполагать, что требования (11.42), необходимые для проведения дисперсионного анализа, выполняется.

Вычислим:

общее среднее:

$$\bar{Y}=\frac{10+50+20+60+30+100}{6}=45;$$

общую вариацию (используя формулы (11.45) и (11.44)):

$$S_Y^2=(10-45)^2+(50-45)^2+(20-45)^2+(60-45)^2+$$
$$+(30-45)^2+(100-45)^2=5350;\quad \hat{\sigma}_Y^2=5350/6;$$

вариацию, обусловленную фактором A (с помощью формул (11.47) и (11.46)):

$$S_A^2=(30-45)^2\cdot 2+(40-45)^2\cdot 2+(65-45)^2\cdot 2=1300;\quad \hat{\sigma}_A^2=1300/6;$$

вариацию, обусловленную фактором B (на основании (11.49) и (11.48)):

$$S_B^2=(20-45)^2\cdot 3+(70-45)^2\cdot 3=3750;\quad \hat{\sigma}_B^2=3750/6;$$

остаточную вариацию (по формулам (11.52) и (11.51)):

$$S_o^2=(10-30-20+45)^2+(50-30-70+45)^2+(20-40-20+45)^2+$$
$$+(60-40-70+45)^2+(30-65-20+45)^2+(100-65-70+45)^2=300;\quad \hat{\sigma}_o^2=300/6.$$

Таблица 11.8

Источник вариации качества пряжи	Показатель вариации	Число степеней свободы	Несмещенная оценка дисперсии σ_o^2
Тип машины (фактор A)	$\hat{\sigma}_A^2=1300/6$	2	$s_A^2=1300/2=650$
Вид сырья (фактор B)	$\hat{\sigma}_B^2=3750/6$	1	$s_B^2=3750/1=3750$
Остаточные факторы	$\hat{\sigma}_o^2=300/6$	$(3-1)\times(2-1)==2$	$s_o^2=300/2=150$
Общая вариации	$\hat{\sigma}_Y^2=5350/6$	5	$s_Y^2=5350/5=1070$

Убедимся в том, что тождество (11.53) действительно выполняется. Имеем $5350/6=1300/6+3750/6+300/6$.

Составим дисперсионную табл. 11.8 типа табл. 11.6.

Найдем числовое значение $F^{(A)}$-критерия (11.54):

$$F_{\text{чис}}^{(A)}=\frac{s_A^2}{s_o^2}=\frac{650}{150}=4{,}3.$$

Теперь по схеме (11.55) определим критическую точку:

$$\left.\begin{array}{l}\alpha=0{,}05\to\gamma=1-\alpha=0{,}95\\ \nu_A=3,\ \nu_B=2 \to \begin{cases} l=\nu_A-1=2\\ k=(\nu_A-1)(\nu_B-1)=2\end{cases}\end{array}\right\}\xrightarrow{\text{п.7}} f_\gamma=19\to x_{\text{пр},\alpha}^{\text{кр}}=19.$$

Так как $F_{\text{чис}}^{(A)}=4{,}3$ не попадает в критическую область $(19, +\infty)$, то считаем, что влияние типа машины на качество пряжи не подтвердилось.

Найдем числовое значение критерия (11.57):

$$F_{\text{чис}}^{(B)}=\frac{s_B^2}{s_o^2}=\frac{3750}{150}=25.$$

Теперь по схеме (11.58) определим критическую точку:

$$\left.\begin{array}{l}\alpha=0{,}05\to\gamma=1-\alpha=0{,}95\\ \nu_A=3,\ \nu_B=2 \to \begin{cases} l=\nu_B-1=1\\ k=(\nu_A-1)(\nu_B-1)=2\end{cases}\end{array}\right\}\xrightarrow{\text{п.7}} f_\gamma=18{,}51\to x_{\text{пр},\alpha}^{\text{кр}}=18{,}51.$$

Так как $F_{\text{чис}}^{(B)}=25$ попадет в критическую область $(18.51; +\infty)$, то считаем, что вид сырья влияет на качество пряжи. А раз это так, то следует оценить степень этого влияния. Рассчитаем выборочный коэффициент детерминации (11.59):

$$\hat{\rho}_B^2=\frac{\hat{\sigma}_B^2}{\hat{\sigma}_Y^2}=\frac{3750/6}{5350/6}=0{,}7.$$

Таким образом, 70% общей выборочной вариации качества пряжи связано с влиянием на нее вида сырья.

В условиях задачи оценим параметры модели (11.60). Так как тип машины (фактор A) не влияет на качество пряжи, то эффекты влиянии этого фактора будут равны нулю, т. е. $\alpha_1=\alpha_2=\alpha_3=0$. Оценки остальных параметров таковы:

оценка параметра a равна $\bar{Y}$, т. е. $a \approx 45$;

оценка эффекта β_j влияния вида сырья равна $(\bar{Y}_{.j} - \bar{Y})$, $j=1, 2$, т. е. $\beta_1 \approx 20-45=-25$; $\beta_2 \approx 70-45=25$;

оценка дисперсии σ_0^2 равна $s_0^2=150$, т. е. $\sigma_0^2 \approx 150$.

Отсюда, принимая во внимание (11.60) и (11.61), получаем, что для качества пряжи, или разрывной нагрузки Y_{i1} нити, изготовленной на любом типе машины из сырья первого вида, справедливо следующее приближенное равенство:

$$Y_{i1} \approx 45-25+N(0; \sqrt{150}), \quad i=1, 2, 3. \tag{11.62}$$

а для нагрузки Y_{i2} нити, изготовленной на любом типе машин из сырья второго вида, имеет место равенство

$$Y_{i2} \approx 45+25+N(0; \sqrt{150}), \quad i=1, 2, 3. \ \blacktriangle \tag{11.63}$$

ГЛАВА 12
КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

§ 12.1. Понятие функциональной, стохастической и корреляционной зависимости. Функция регрессии

Условимся обозначать 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. принимающая вполне определенные значения), то и функционально зависящая от нее величина Y тоже является детерминированной; если же X — случайная величина, то и Y также случайная величина.

Однако гораздо чаще в окружающем нас мире имеет место не функциональная, а *стохастическая*, *или вероятностная*, *зависимость*, 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 результирующую переменную не только контролируемого или контролируемых факторов (в данном случае таким контролируемым фактором является переменная X), а и многочисленных неконтролируемых случайных факторов. В этой ситуации переменная Y является случайной величиной. Переменная же X может быть как детерминированной, так и случайной величиной. Следует заметить, что со стохастической зависимостью мы уже сталкивались в дисперсионном анализе. Вспомним табл. 11.1. В этой таблице в роли переменной X выступает фактор A; уровни

фактора A — это значения переменной X; каждому такому значению соответствует не одно, а множество непредсказуемых значений переменной Y.

Допустим, что существует стохастическая зависимость случайной переменной Y от X. Зафиксируем некоторое значение x переменной X. При $X=x$ переменная 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 *групповым генеральным средним* переменной Y при $X=x$ или **математическим ожиданием случайной величины Y, вычисленным при условии, что $X=x$; это условное математическое ожидание обозначают так: $M(Y/X=x)$.** Если существует стохастическая зависимость Y от X, то прежде всего стараются выяснить, изменяются или нет при изменении x условные математические ожидания $M(Y/X=x)$. Если при изменении x условные математические ожидания $M(Y/X=x)$ изменяются, то говорят, что имеет место *корреляционная зависимость* величины Y от X; если же условные математические ожидания остаются неизменными, то говорят, что корреляционная зависимость величины Y от X отсутствует.

Функция $\varphi(x)=M(Y/X=x)$, описывающая изменение условного математического ожидания случайной переменной Y при изменении значений x переменной X, называется *функцией регрессии*.

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. Зафиксируем значение x уровня квалификации; ему соответствует некоторое множество значений выработки Y. Тогда $M(Y/X=x)$ — средняя выработка рабочего при условии, что его уровень квалификации равен x, или, иначе говоря, $M(Y/X=x)$ — 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 12.1.** Допустим, что одновременно изучаются две случайные величины X и Y, или, иначе говоря, двумерная случайная величина (X, Y), которая задана табл. 12.1.

Таблица 12.1

j	i / y_j \ x_i	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

Табл. 12.1 называют *таблицей распределения 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 $X=2$ и одновременно $Y=0{,}4$, или, иначе говоря, вероятность произведения двух событий: события, состоящего в том, что $X=2$, и события, состоящего в том, что $Y=0{,}4$, т. е. $P((X=2)(Y=0{,}4))=0{,}15$. Аналогично, вероятность $P((X=2)(Y=0{,}8))=0{,}05$ и т. д. Обратим внимание на следующее: поскольку в табл. 12.1 указаны все возможные значения величин X и Y, сумма вероятностей, стоящих в таблице, должна быть равна единице: $0{,}15+0{,}05+0{,}12+0{,}30+0{,}03+0{,}35=1$.

Прежде чем выяснить тип зависимости величины Y от X, найдем:

а) Закон распределения величины X. Он определяется табл. 12.2.

Таблица 12.2

x	$x_1=2$	$x_2=5$	$x_3=8$		$MX=5{,}54$,
$P(X=x)$	$0{,}15+0{,}05=0{,}2$	$0{,}12+0{,}30=0{,}42$	$0{,}35+0{,}03=0{,}38$	$\Sigma=1$	$DX=4{,}9284$

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.

$$P(X=2)=P((X=2)(Y=0{,}4))+P((X=2)(Y=0{,}08))=$$
$$=0{,}15+0{,}05=0{,}2.$$

Справа от ряда распределения величины X находятся ее математическое ожидание и дисперсия.

б) Закон распределения величины Y. Он имеет вид табл. 12.3.

Таблица 12.3

y	$y_1=0{,}4$	$y_2=0{,}8$		$MY=0{,}68$,
$P(Y=y)$	$0{,}15+0{,}12+0{,}03=0{,}30$	$0{,}05+0{,}30+0{,}35=0{,}7$	$\Sigma=1$	$DY=0{,}0336$

в) Условные законы распределения величины Y, а именно закон распределения величины Y сначала при условии, что $X=2$, затем при условии, что $X=5$, и наконец, при условии, что $X=8$.

Итак, пусть $X=2$. Тогда условная вероятность

$$P(Y=0{,}4/X=2)=\frac{P((Y=0{,}4)(X=2))}{P(X=2)}=\frac{0{,}15}{0{,}2}=0{,}75,$$

а условная вероятность

$$P(Y=0{,}8/X=2)=\frac{P((Y=0{,}8)(X=2))}{P(X=2)}=\frac{0{,}05}{0{,}2}=0{,}25.$$

Таким образом, закон распределения величины Y при условии, что $X=2$, задан табл. 12.4.

Таблица 12.4

y	$y_1=0{,}4$	$y_2=0{,}8$	
$P(Y=y/X=2)$	0,75	0,25	$\Sigma=1$

$M(Y/X=2)=0{,}4\cdot 0{,}75+0{,}8\times$ $\times 0{,}25=0{,}5$, $D(Y/X=2)=0{,}03$

Справа помещено условное математическое ожидание и значение условной дисперсии. Покажем, как вычисляется условная дисперсия. Общая формула условной дисперсии имеет вид

$$D(Y/X=x)=M[(Y/X=x)-M(Y/X=x)]^2. \qquad (12.1)$$

Для табл. 12.4 получаем

$$D(Y/X=2)=M[(Y/X=2)-M(Y/X=2)]^2=M[Y/X=2)-0{,}5]^2=$$
$$=\sum_{i=1}^{2}(y_i-0{,}5)^2\cdot P(Y=y_i/X=2)=$$
$$=(0{,}4-0{,}5)^2\cdot 0{,}75+(0{,}8-0{,}5)^2\cdot 0{,}25=0{,}03.$$

Пусть $X=5$. Тогда $P(Y=0{,}4/X=5)=\frac{P((Y=0{,}4)(X=5))}{P(X=5)}=\frac{0{,}12}{0{,}42}=\frac{2}{7}$;

$$P(Y=0{,}8/X=5)=\frac{P((Y=0{,}8)(X=5))}{P(X=5)}=\frac{0{,}30}{0{,}42}=\frac{5}{7}.$$

Таким образом, закон распределения величины Y при условии, что $X=5$, имеет вид табл. 12.5.

Таблица 12.5

y	0,4	0,8	
$P(Y=y/X=5)$	2/7	5/7	$\Sigma=1$

$M(Y/X=5)=\frac{24}{35}\approx 0{,}686$, $D(Y/X=5)=0{,}03265$

И наконец, при $X=8$ ряд распределения задан табл. 12.6.

Таблица 12.6

y	0,4	0,8	
$P(Y=y/X=8)$	$\frac{3}{38}$	$\frac{35}{38}$	$\Sigma=1$

$$M(Y/X=8)=\frac{73}{95}\approx 0{,}768,\quad D(Y/X=8)=0{,}01163$$

Из табл. 12.4—12.6 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 именно из этих чисел она будет равна — сказать заранее нельзя. Ясно прослеживается и корреляционная зависимость величины Y от X, поскольку с изменением значений x величины X меняются и условные математические ожидания $M(Y/X=x)$. Функция регрессии, т. е. зависимость условного математического ожидания $M(Y/X=x)$ от x, задается в виде табл. 12.7.

Таблица 12.7

x	2	5	8
$M(Y/X=x)$	0,5	$24/35\approx 0{,}686$	$73/95\approx 0{,}768$

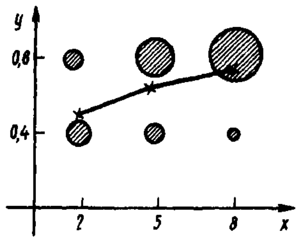


Рис. 12.1

На рис. 12.1 приведено графическое изображение табл. 12.1, которое называется *диаграммой рассеяния*. Координатами точек являются значения x_i и y_i величин X и Y, а «размер» точки пропорционален вероятности $P((X=x_i)(Y=y_i))$. «Крестиками» на рис. 12.1 отмечены условные математические ожидания. ○

§ 12.2. Генеральное корреляционное отношение. Его свойства

Можно ли измерить степень корреляционной и стохастической зависимости величины Y от X?

Ответ на этот вопрос проиллюстрируем примером 12.1. Все полученные в примере результаты объединим в табл. 12.8.

Таблица 12.8

x_i	$x_1=2$	$x_2=5$	$x_3=8$	
$P(X=x_i)$	0,2	0,42	0,38	
$M(Y/X=x_i)$	0,5	0,686	0,768	$MY=0{,}68$ (см. табл. 12.3),
$D(Y/X=x_i)$	0,03	0,03265	0,01163	$DY=0{,}0336$ (см. табл. 12.3)

Замечание. Так как X — случайная величина, принимающая значения 2, 5 и 8 с вероятностью 0,2; 0,42 и 0,38, то такими же будут вероятности и условных математических ожиданий, и дисперсий. Таким образом, условное математическое ожидание $M(Y/X)$, так же как и условная дисперсия $D(Y/X)$, — случайные величины.

Обратим также внимание на то, что MY, найденное в табл. 12.3, можно вычислить и по табл. 12.8 следующим образом:

$$MY=M[M(Y/X)]=\sum_{i=1}^{3}M(Y/X=x_i)P(X=x_i)=$$
$$=0{,}5\cdot 0{,}2+0{,}686\cdot 0{,}42+0{,}768\cdot 0{,}38=0{,}68.$$

Разброс значений величины Y вокруг математического ожидания MY измеряется дисперсией DY, или σ_Y^2:

$$\sigma_Y^2=DY=M(Y-MY)^2. \qquad (12.2)$$

(По табл. 12.3 $\sigma_Y^2=0{,}0336$.) Этот разброс может быть вызван:

зависимостью величины Y от X (эта зависимость может быть обусловлена не только непосредственным влиянием X на Y, но и наличием случайных факторов, действующих на Y через переменную X);

зависимостью Y от случайных факторов, влияющих только на Y и не влияющих на X; эти факторы называют *остаточными*.

1) Построим показатель разброса значений величины Y, связанного с ее зависимостью от фактора X.

□ Условное математическое ожидание $M(Y/X=x)$ является «представителем игреков», которые имеют место при $X=x$. Характеристикой разброса условных математических ожиданий $M(Y/X=x)$ относительно MY является дисперсия $D[M(Y/X)]$, или

$$\sigma_\Phi^2=D[M(Y/X)]=M[M(Y/X)-MY]^2 \qquad (12.3)$$

— эта величина и будет показателем разброса значений величины Y, связанного с ее зависимостью от фактора X. (По табл. 12.8 найдем:

$$\sigma_\Phi^2=M[M(Y/X)-MY]^2=$$
$$=(0{,}5-0{,}68)^2\cdot 0{,}2+(0{,}686-0{,}68)^2\cdot 0{,}42+(0{,}768-0{,}68)^2\cdot 0{,}38=$$
$$=0{,}0095.)\ ■$$

2) Теперь построим показатель разброса «игреков», связанного с влиянием остаточных факторов.

□ Зафиксируем какое-либо значение x величины X. Тогда причиной вариации величины Y при $X=x$ будут остаточные факторы, влияющие только на Y и не влияющие на X. Измерителем этой вариации является условная дисперсия $D(Y/X=x)$. При различных же «иксах» характеристикой разброса «игреков», вызванного влиянием на Y остаточных факторов, будет генеральное среднее из условных дисперсий, или, иначе, математическое ожидание условной дисперсии. Эту величину обозначим σ_o^2. Имеем

$$\sigma_o^2 = M[D(Y/X)], \tag{12.4}$$

где при $X=x$ дисперсия $D(Y/X=x)$ вычисляется по формуле (12.1). (По табл. 12.8 найдем

$$\sigma_o^2 = M[D(Y/X)] = \sum_{i=1}^{3} D(Y/X=x_i)P(X=x_i) =$$

$$= 0{,}03 \cdot 0{,}2 + 0{,}03265 \cdot 0{,}42 + 0{,}01163 \cdot 0{,}38 = 0{,}0241.) \ \blacksquare$$

Для вычисленных дисперсий справедливо тождество

$$DY = D[M(Y/X)] + M[D(Y/X)]$$

или (12.5)

$$\sigma_Y^2 = \sigma_\Phi^2 + \sigma_o^2.$$

Степень стохастической зависимости величины Y от X измеряется *генеральным корреляционным отношением*

$$\rho_{Y/X} = +\sqrt{\frac{D[M(Y/X)]}{DY}} = +\sqrt{\frac{\sigma_\Phi^2}{\sigma_Y^2}} \underset{(12.5)}{=} +\sqrt{1-\frac{\sigma_o^2}{\sigma_Y^2}} = +\sqrt{1-\frac{M[D(Y/X)]}{DY}}. \tag{12.6}$$

Квадрат корреляционного отношения

$$\rho_{Y/X}^2 = \frac{\sigma_\Phi^2}{\sigma_Y^2} = \frac{D[M(Y/X)]}{DY} \underset{\substack{(12.3)\\(12.2)}}{=} \frac{M[M(Y/X)-MY]^2}{M(Y-MY)^2} \tag{12.7}$$

называется *генеральным коэффициентом детерминации*; он показывает, какую долю дисперсии величины Y составляет дисперсия условных математических ожиданий, или, иначе говоря, какая доля дисперсии DY объясняется корреляционной зависимостью Y от X. (В примере 12.1 $\sigma_\Phi^2 = 0{,}0095$, $\sigma_Y^2 = 0{,}0336$, поэтому $\rho_{Y/X}^2 = \frac{0{,}0095}{0{,}0336} = 0{,}28$, т. е. 28% дисперсии величины Y объясняется ее корреляционной зависимостью от X; $\rho_{Y/X} = +\sqrt{0{,}28} = 0{,}53$.)

Свойства генерального корреляционного отношения как измерителя степени корреляционной и стохастической зависимости.

1^0. $0 \leqslant \rho_{Y/X} \leqslant 1$.

Действительно, согласно (12.6), $\rho_{Y/X} \geqslant 0$; с другой стороны, из (12.5) следует, что $\sigma_\Phi^2 \leqslant \sigma_Y^2$, поэтому $\rho_{Y/X} \leqslant 1$.

2^0. *Условие $\rho_{Y/X}=0$ является необходимым и достаточным для отсутствия корреляционной зависимости Y от X, т. е. для того, чтобы $M(Y/X=x)=\text{const}$ при любом значении x величины X.*

Достаточность. Пусть $\rho_{Y/X}=0$. Тогда по формуле (12.6) имеем $D[M(Y/X)]=0$. Из этого равенства следует, что $M(Y/X)=\text{const}$, т. е. условное математическое ожидание $M(Y/X=x)$ «не реагирует» на изменение значения x величины X. Это и означает, что корреляционная зависимость Y от X отсутствует.

Необходимость. Пусть отсутствует корреляционная зависимость величины Y от X; это означает, что $M(Y/X=x)=\text{const}$ при любом допустимом значении x, поэтому $D[M(Y/X)]=0$. Поскольку это так, постольку, согласно (12.6), $\rho_{Y/X}=0$.

Следствие. Чем ближе $\rho_{Y/X}$ к нулю, тем в соответствии с (12.6) ближе к нулю $D[M(Y/X)]$, а это, учитывая (12.3), означает, что уменьшается разброс условных математических ожиданий $M(Y/X=x)$ относительно MY. Таким образом, *чем ближе $\rho_{Y/X}$ к нулю, тем меньше «реакция условного математического ожидания $M(Y/X=x)$ на изменение x», или, иначе говоря, «тем меньше степень корреляционной зависимости Y от X».*

И, наоборот, чем «меньше степень корреляционной зависимости Y от X», тем ближе $\rho_{Y/X}$ к нулю.

3^0. *Условие $\rho_{Y/X}=1$ является необходимым и достаточным для функциональной зависимости величины Y от X.*

Достаточность. Пусть $\rho_{Y/X}=1$. Это в силу (12.6) означает, что $\sigma_o^2=0$ или $M[D(Y/X)]=0$. Но так как дисперсия любой величины неотрицательна, то из последнего равенства следует, что $D(Y/X=x)=0$ при любом x, а это означает, что при $X=x$ величина Y остается постоянной (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 $D(Y/X=x)=0$, поэтому и $\sigma_o^2 = M[D(Y/X)] = M0 = 0$. Но тогда из (12.6) следует, что $\rho_{Y/X}=1$.

Следствие. Чем ближе $\rho_{Y/X}$ к единице, тем в силу (12.6) ближе к нулю $M[D(Y/X)]$, а следовательно, и условные дисперсии $D(Y/X=x)$. Это означает, что при каждом допустимом значении x уменьшается разброс «игреков» относительно $M(Y/X=x)$. Таким образом, *чем ближе $\rho_{Y/X}$ к единице, тем меньше при каждом x отличие «игреков» от постоянного числа, равного $M(Y/X=x)$, или, иначе говоря, тем выше степень стохастической зависимости Y от X. И, наоборот, чем выше степень стохастической зависимости Y от X, тем ближе $\rho_{Y/X}$ к единице.*

Проиллюстрируем сформулированные свойства диаграммами рассеивания, приведенными на рис. 12.2. Поясним этот рисунок.

а) корреляционная зависимость отсутствует: условные математические ожидания не изменяются, $\rho_{Y/X}=0$;

б) зависимость Y от X функциональная: $\rho_{Y/X}=1$;

в) стохастическая зависимость «довольно сильная»: точки сконцентрированы в относительно узкой «параболической полосе»;

г) стохастическая зависимость по сравнению с зависимостью, изображенной на рис. 12.2, *в*, менее сильная: точки концентрируются в более широкой «параболической полосе».

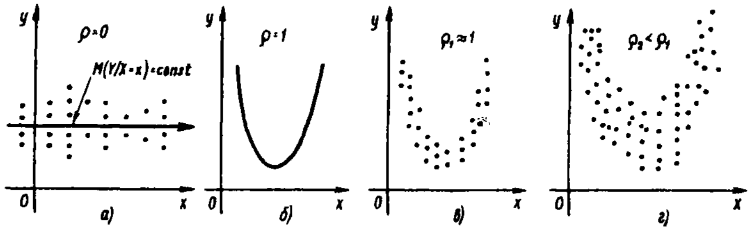


Рис. 12.2

В практических задачах наибольший интерес представляют следующие вопросы:

существует корреляционная зависимость Y от X или нет, или, иначе говоря, отлично ли генеральное корреляционное отношение $\rho_{Y/X}$ от нуля или равно нулю;

если корреляционная зависимость существует, то какой вид имеет функция регрессии (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 12.1 этот закон задан табл. 12.1, в которой даны 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.

§ 12.3. Выборочное корреляционное отношение. Его значимость

Пусть имеется n наблюдений двумерной величины (X, Y). Наблюдавшиеся «иксы» и «игреки» поместим в табл. 12.9, которая называется *корреляционной таблицей* и строится следующим образом:

«иксы» группируются в вариационный ряд, число групп которого обозначим ν; если это дискретный ряд, то x_1, x_2, ..., x_ν — различающиеся между собой результаты наблюдений, или варианты; если это интервальный ряд, то x_1, x_2, ..., x_ν — центры интервалов;

«игреки» группируют в вариационный ряд, число групп которого обозначим q; y_1, y_2, ..., y_q — это либо варианты, если ряд дискретный, либо середины 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, $i = 1, 2, ..., \nu$, $j = 1, 2, ..., q$; 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}, ..., $m_{q\nu}$ называются *частотами*.

Таблица 12.9

j \ i		1	2	...	ν	
	Y \ X	x_1	x_2	...	x_ν	$\sum$
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
$\sum$		n_1	n_2	...	n_ν	$n=\sum_{i=1}^{\nu} n_i=\sum_{j=1}^{q} m_j$
Групповое среднее		$\bar{Y}^{(1)}$	$\bar{Y}^{(2)}$	...	$\bar{Y}^{(\nu)}$	
Групповая выборочная дисперсия		$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	...	$\hat{\sigma}_\nu^2$	

Прежде чем пояснить остальные элементы этой таблицы, сделаем следующее замечание по поводу схемы построения выборочного корреляционного отношения.

Замечание. От табл. 12.9, содержащей частоты, можно перейти к таблице частостей (табл. 12.10).

Таблица 12.10

j \ i		1	2	...	ν
	Y \ X	x_1	x_2	...	x_ν
1	y_1	$\hat{p}_{11}$	$\hat{p}_{12}$	...	$\hat{p}_{1\nu}$
2	y_2	$\hat{p}_{21}$	$\hat{p}_{22}$	...	$\hat{p}_{2\nu}$
...	...	...	...	...	...
q	y_q	$\hat{p}_{q1}$	$\hat{p}_{q2}$	...	$\hat{p}_{q\nu}$

$$\hat{p}_{ji} = m_{ji}/n, \quad i = 1, 2, ..., \nu; \quad j = 1, 2, ..., q.$$

Сравним табл. 12.10 и 12.1. Их различие состоит в следующем: табл. 12.1 относилась к генеральной совокупности, поэтому в ней были указаны все мыслимые значения величин X и Y и вероятности комбинаций этих значений; табл. 12.10 относится к выборочной совокупности, и в ней приведены наблюдаемые значения величин X и Y и частости, или опытные вероятности комбинаций наблюдаемых значений. Поэтому выборочное корреляционное отношение можно строить по той же схеме, что и генеральное корреляционное отношение (см. § 12.2), если заменить возможные значения величин X и Y на наблюдаемые, вероятности на частости, математические ожидания на средние, дисперсии на выборочные дисперсии.

Однако чаще выборочное корреляционное отношение строят, используя непосредственно табл. 12.9, а не табл. 12.10. Так мы и поступим.

В табл. 12.9 кроме сгруппированных наблюдений и частот содержатся следующие данные:

суммы частот по каждой строке

$$m_1=\sum_{i=1}^{\nu} m_{1i},\quad m_2=\sum_{i=1}^{\nu} m_{2i},\quad \ldots,\quad m_q=\sum_{i=1}^{\nu} m_{qi}, \tag{12.8}$$

суммы частот по каждому столбцу

$$n_1=\sum_{j=1}^{q} m_{j1},\quad n_2=\sum_{j=1}^{q} m_{j2},\quad \ldots,\quad n_\nu=\sum_{j=1}^{q} m_{j\nu}, \tag{12.9}$$

групповые средние значения «игреков»

$$\begin{aligned}
&\bar Y^{(1)}=(y_1m_{11}+y_2m_{21}+\ldots+y_qm_{q1})/n_1,\\
&\bar Y^{(2)}=(y_1m_{12}+y_2m_{22}+\ldots+y_qm_{q2})/n_2,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\bar Y^{(\nu)}=(y_1m_{1\nu}+y_2m_{2\nu}+\ldots+y_qm_{q\nu})/n_\nu.
\end{aligned} \tag{12.10}$$

Эти средние являются выборочными аналогами соответствующих условных математических ожиданий: $\bar Y^{(1)}$ — выборочный аналог математического ожидания величины Y при условии, что $X=x_1$, т. е. аналог величины $M(Y/X=x_1)$; $\bar Y^{(2)}$ — аналог $M(Y/X=x_2)$ и т. д.

Групповые выборочные дисперсии:

$$\begin{aligned}
&\hat\sigma_1^2=\left[(y_1-\bar Y^{(1)})^2m_{11}+(y_2-\bar Y^{(1)})^2m_{21}+\ldots+(y_q-\bar Y^{(1)})^2m_{q1}\right]/n_1,\\
&\hat\sigma_2^2=\left[(y_1-\bar Y^{(2)})^2m_{12}+(y_2-\bar Y^{(2)})^2m_{22}+\ldots+(y_q-\bar Y^{(2)})^2m_{q2}\right]/n_2,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\hat\sigma_\nu^2=\left[(y_1-\bar Y^{(\nu)})^2m_{1\nu}+(y_2-\bar Y^{(\nu)})^2m_{2\nu}+\ldots+(y_q-\bar Y^{(\nu)})^2m_{q\nu}\right]/n_\nu.
\end{aligned} \tag{12.11}$$

Эти дисперсии являются выборочными аналогами соответствующих условных дисперсий: $\hat\sigma_1^2$ — выборочный аналог условной дисперсии $D(Y/X=x_1)$, $\hat\sigma_2^2$ — аналог $D(Y/X=x_2)$ и т. д.

Теперь приступим к построению выборочного аналога генерального корреляционного отношения (12.6).

□ 1) Выборочным аналогом генеральной дисперсии $\sigma_Y^2=DY$ является величина $\hat\sigma_Y^2$. Для того чтобы вычислить $\hat\sigma_Y^2$, найдем сначала среднее $\bar Y$ по данным табл. 12.9. Это можно сделать по одной из следующих тождественных формул:

$$\bar Y=\frac{1}{n}(y_1m_1+y_2m_2+\ldots+y_qm_q)=\frac{1}{n}(\bar Y^{(1)}n_1+\bar Y^{(2)}n_2+\ldots+\bar Y^{(\nu)}\cdot n_\nu) \tag{12.12}$$

Напомним, что $\bar Y$ — это выборочный аналог MY. Теперь найдем

$$\begin{aligned}
S_Y^2&=(y_1-\bar Y)^2m_1+(y_2-\bar Y)^2m_2+\ldots+(y_q-\bar Y)^2m_q=\\
&=\sum_{j=1}^{q}(y_j-\bar Y)^2m_j.
\end{aligned} \tag{12.13}$$

Тогда

$$\hat\sigma_Y^2=S_Y^2/n. \tag{12.14}$$

2) Выборочным аналогом генеральной дисперсии $\sigma_\Phi^2=D\left[M(Y/X)\right]$ (см. 12.6)) является выборочная дисперсия $\hat\sigma^2_{\bar Y^{(i)}}$ групповых средних; обозначим ее $\hat\sigma_\Phi^2$. Имеем

$$\begin{aligned}
\hat\sigma_\Phi^2=\hat\sigma^2_{\bar Y^{(i)}}&=\frac{1}{n}\left[(\bar Y^{(1)}-\bar Y)^2n_1+(\bar Y^{(2)}-\right.\\
&\left.-\bar Y)^2n_2+\ldots+(\bar Y^{(\nu)}-\bar Y)^2n_\nu\right]=S_\Phi^2/n,
\end{aligned} \tag{12.15}$$

где

$$\begin{aligned}
S_\Phi^2&=(\bar Y^{(1)}-\bar Y)^2n_1+(\bar Y^{(2)}-\bar Y)^2n_2+\ldots+(\bar Y^{(\nu)}-\bar Y)^2n_\nu=\\
&=\sum_{i=1}^{\nu}(\bar Y^{(i)}-\bar Y)^2n_i.
\end{aligned} \tag{12.16}$$

3) Выборочным аналогом дисперсии $\sigma_o^2=M\left[D(Y/X)\right]$ является средняя $\overline{\hat\sigma_i^2}$ групповых выборочных дисперсий (см. табл. 12.9). Обозначим $\hat\sigma_o^2$ эту среднюю

$$\hat\sigma_o^2=\overline{\hat\sigma_i^2}=\frac{1}{n}(\hat\sigma_1^2n_1+\hat\sigma_2^2n_2+\ldots+\hat\sigma_\nu^2n_\nu)=S_o^2/n, \tag{12.17}$$

где $S_o^2=\hat\sigma_1^2n_1+\hat\sigma_2^2n_2+\ldots+\hat\sigma_\nu^2n_\nu$. Учитывая (12.11), получаем

$$\begin{aligned}
S_o^2&=(y_1-\bar Y^{(1)})^2m_{11}+(y_2-\bar Y^{(1)})^2m_{21}+\ldots+(y_q-\bar Y^{(1)})^2m_{q1}+\\
&+(y_1-\bar Y^{(2)})^2m_{12}+(y_2-\bar Y^{(2)})^2m_{22}+\ldots+(y_q-\bar Y^{(2)})^2m_{q2}+\\
&+\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots+\\
&+(y_1-\bar Y^{(\nu)})^2m_{1\nu}+(y_2-\bar Y^{(\nu)})^2m_{2\nu}+\ldots+(y_q-\bar Y^{(\nu)})^2m_{q\nu}=\\
&=\sum_{i=1}^{\nu}\sum_{j=1}^{q}(y_j-\bar Y^{(i)})^2m_{ji}.
\end{aligned} \tag{12.18}$$

Для вычисленных дисперсий справедливо тождество

$$\hat\sigma_Y^2=\hat\sigma^2_{\bar Y^{(i)}}+\overline{\hat\sigma_i^2},$$

или

$$\hat\sigma_Y^2=\hat\sigma_\Phi^2+\hat\sigma_o^2, \tag{12.19}$$

аналогичное тождеству (12.5), имеющему место в генеральной совокупности.

Замечание. Тождество (12.19) аналогично тождеству (11.21), доказанному в дисперсионном анализе. Убедимся в том, что нет различия в дисперсиях, используемых в (12.19) и в (11.21). Сравним (12.13) и (11.15), где $Y_j^{(i)}$ — результат j-го наблюдения на i-м уровне фактора, или, если X рассматривать как фактор, при $X=x_i$. Эти формулы отличаются только по форме: в (12.13) используются сгруппированные наблюдения, а в (11.15) — непосредственные наблюдения; результаты же вычислений по формулам (12.13) и (11.15) будут совпадать. Аналогичное замечание относится и к формулам (12.18) и (11.20). Формулы же (12.15) и (11.18) полностью совпадают. Поэтому дисперсии, используемые в тождестве (12.19), совпадают с дисперсиями в формуле (11.21).

Выборочный аналог генерального корреляционного отношения (12.6) вычисляется следующим образом:

$$\hat{\rho}_{Y/X}=+\sqrt{\frac{\hat{\sigma}^2_{\bar{Y}^{(i)}}}{\hat{\sigma}^2_Y}}=+\sqrt{\frac{\hat{\sigma}^2_\Phi}{\hat{\sigma}^2_Y}}\underset{(12.19)}{=\!=}+\sqrt{1-\frac{\hat{\sigma}^2_o}{\hat{\sigma}^2_Y}}=+\sqrt{1-\frac{\overline{\hat{\sigma}^2_i}}{\hat{\sigma}^2_Y}}. \quad (12.20)$$

Величина $\hat{\rho}^2_{Y/X}$ называется *выборочным коэффициентом детерминации. Этот коэффициент показывает, какую долю дисперсии* $\hat{\sigma}^2_Y$ *составляет выборочная дисперсия групповых средних «игреков» или, иначе говоря, какая доля дисперсии* $\hat{\sigma}^2_Y$ *объясняется зависимостью* Y от X. ■

Обратим внимание на следующее: из (12.19) имеем, что $\sigma^2_\Phi \leqslant \sigma^2_Y$, поэтому, учитывая (12.20), получаем

$$0\leqslant \hat{\rho}_{Y/X}\leqslant 1.$$

Замечание. Как правило, дисперсии $\hat{\sigma}^2_Y$ и $\hat{\sigma}^2_\Phi$ находят не по рассмотренным выше формулам, а по следующим, более удобным для вычислений:

$$\hat{\sigma}^2_Y=\frac{1}{n}\sum_{j=1}^{q} y_j^2 m_j-(\bar{Y})^2, \quad (12.21)$$

$$S^2_Y=\hat{\sigma}^2_Y n.$$

Эти формулы тождественны (12.14) и (12.13);

$$\hat{\sigma}^2_i=\frac{1}{n_i}\sum_{j=1}^{q} y_j^2 m_{ji}-\left(\bar{Y}^{(i)}\right)^2, \quad i=1, 2, \ldots, \nu. \quad (12.22)$$

Эти формулы тождественны формулам (12.11):

$$\hat{\sigma}^2_\Phi=\hat{\sigma}^2_{\bar{Y}^{(i)}}=\frac{1}{n}\sum_{i=1}^{\nu}\left(\bar{Y}^{(i)}\right)^2 n_i-(\bar{Y})^2, \quad (12.23)$$

$$S^2_\Phi=\hat{\sigma}^2_\Phi n.$$

Эти формулы тождественны (12.15) и (12.16).

Задача 12.1. Данные об объеме выпуска продукции (Y) и стоимости основных промышленно-производственных фондов (X) по 60 предприятиям сгруппированы в табл. 12.11. Вычислить корреляционное отношение $\hat{\rho}_{Y/X}$.

△ Необходимые для вычисления корреляционного отношения промежуточные данные содержатся в табл. 12.11.

Поясним табл. 12.11.

В этой таблице $\nu=5$, $q=5$. 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. Числа m_j в столбце (1) получены по формулам (12.8), а числа n_i в строке (1) — по формулам (12.9).

Групповые средние вычисляются по формулам (12.10); они приведены в строке (3). Например, $\bar{Y}^{(1)}=0{,}8/4$. Числители дробей (12.10) содержатся в строке (2) — они получены как суммы произведений «игреков» на частоты соответствующего столбца. Например, $0{,}8=0{,}1\cdot 2+0{,}3\cdot 2$.

Групповые дисперсии содержатся в строке (7). Они вычисляются по формулам (12.22).

Вычисление $\bar{Y}$ производится по первой из формул (12.12) в столбце (2). Дисперсия $\hat{\sigma}^2_Y$ и значение S^2_Y находятся по формулам (12.21).

Вычисление $\hat{\sigma}^2_\Phi$ и S^2_Φ производится в строке (8) по формулам (12.23). Дисперсия $\hat{\sigma}^2_o$ вычисляется в строке (9) в соответствии с (12.17).

Назначение строк (10)—(12) выяснено далее.

Обратим внимание на то, что $\hat{\sigma}^2_Y=\hat{\sigma}^2_\Phi+\hat{\sigma}^2_o$. Действительно, $0{,}035=0{,}021+0{,}014$. Теперь в соответствии с (12.20) найдем $\hat{\rho}_{Y/X}=+\sqrt{0{,}021/0{,}035}=0{,}78$; $\hat{\rho}^2_{Y/X}=0{,}60$. Это означает, что 60% выборочной дисперсии объема выпуска продукции объясняется его зависимостью от стоимости основных промышленно-производственных фондов. ▲

Но ведь этот результат получен по выборочным данным. Возможно ли, что в генеральной совокупности $\rho_{Y/X}=0$?

Проверим гипотезу

$$H_0\colon\ \rho_{Y/X}=0. \quad (12.24)$$

Предварительно отметим, что в силу свойства 2^0 корреляционного отношения при выполнении гипотезы (12.24) имеет место равенство условных, или групповых математических ожиданий величины Y:

$$M(Y/X=x_1)=M(Y/X=x_2)=\ldots=M(Y/X=x_\nu).$$

Поэтому проверка гипотезы (12.24) сводится к проверке гипотезы о равенстве групповых математических ожиданий — это задача дисперсионного анализа. Ее можно решить (см. § 11.1), если выполняются требования (11.4) и (11.5). Применительно к нашим условиям эти требования формулируются следующим образом:

при каждом наблюдаемом значении x_i величины X наблюдения величины Y должны быть независимыми и проводиться в одинаковых условиях; наблюдения должны быть независимы и при различных «иксах»; (12.25)

при каждом значении x_i величина Y должна иметь нормальный закон с постоянной для различных «иксов» генеральной дисперсией (обозначим эту дисперсию σ^2_o; $\sigma^2_o=D(Y/X=x_1)=D(Y/X=x_2)=\ldots=D(Y/X=x_\nu)$). (12.26)

Допустим, что эти требования выполняются. Тогда для проверки гипотезы (12.24) следует заполнить табл. 11.2 и использовать критерий (11.31) и схему (11.32).

Задача 12.2. Предположив, что в условиях задачи 12.1 требования (12.25) и (12.26) выполняются, проверить гипотезу (12.24).

△ Использовав данные табл. 12.11, заполним табл. 11.2. В результате получим табл. 12.2.

Таблица 12.11

j	Объем продукции (интервал)	Стоимость фондов (интервал) / центр интервала x_i / центр интервала y_j	i	1	2	3	4	$5=\nu$	m_j		
		Стоимость фондов (интервал)		0—2	2—4	4—6	6—8	8—10			
		центр интервала x_i		1	3	5	7	9			
		центр интервала y_j							(1)	$(2)=\sum_{j=1}^{4} y_j m_j$	$(3)=\sum_{j=1}^{4} y_j^2 m_j$
1	0—0,2	0,1		2	2				4	0,4	0,04
2	0,2—0,4	0,3		2	7	10			19	5,7	1,71
3	0,4—0,6	0,5			2	17	7		26	13,0	6,50
4	0,6—0,8	0,7				4	3	2	9	6,3	4,41
$5=q$	0,8—1,0	0,9						2	2	1,8	1,62
	n_i		(1)	4	11	31	10	4	$n=60$	$\sum_2=27{,}2$	$\sum_3=14{,}28$

$\sum_{j=1}^{5} y_i m_{ji}$	(2)	0,8	3,3	14,3	5,6	3,2	$\bar{Y}=\sum_2/n=0{,}45$
$\bar{Y}^{(i)}=(2)/(1)$	(3)	0,2	0,3	0,46	0,56	0,8	$\hat{\sigma}_Y^2=\sum_3/n-(\bar{Y})^2=0{,}035$
$\sum_{j=1}^{5} y_j^2 m_{ji}$	(4)	0,2	1,15	7,11	3,22	2,60	$S_Y^2=\hat{\sigma}_Y^2 n=2{,}1$
(4)/(1)	(5)	0,05	0,104	0,229	0,322	0,65	
$(\bar{Y}^{(i)})^2=(3)^2$	(6)	0,04	0,09	0,212	0,314	0,64	
$\hat{\sigma}_i^2=(5)-(6)$	(7)	0,01	0,014	0,017	0,008	0,01	
(6)·(1)	(8)	0,16	0,99	6,572	3,14	2,56	$\sum_8=13{,}422$ $\hat{\sigma}_\Phi^2=\sum_8/n-(\bar{Y})^2=0{,}021$; $S_\Phi^2=\hat{\sigma}_\Phi^2 n=1{,}26$
(7)·(1)	(9)	0,04	0,154	0,527	0,08	0,04	$\sum_9=0{,}841$ $\hat{\sigma}_o^2=\sum_9/n=0{,}014$; $S_o^2=\hat{\sigma}_o^2 n=0{,}841$
$x_i n_i$	(10)	4	33	155	70	36	$\sum_{10}=298$ $\bar{X}=\sum_{10}/n=4{,}97$
$x_i^2 n_i$	(11)	4	99	775	490	324	$\sum_{11}=1692$ $\hat{\sigma}_X^2=\sum_{11}/n-(\bar{X})^2=3{,}499$
$x_i\cdot(2)$	(12)	0,8	9,9	71,5	39,2	28,8	$\sum_{12}=150{,}2$ $\overline{XY}=\sum_{12}/n=2{,}503$

Таблица 12.12

Источник вариации выпуска продукции	Показатель вариации	Число степеней свободы	Несмещенная оценка дисперсии σ_o^2
Основные фонды	$\hat{\sigma}_\Phi^2=\hat{\sigma}_{\bar{Y}^{(i)}}^2=S_\Phi^2/n= =1{,}26/60=0{,}021$	$\nu-1= =5-1=4$	$s_\Phi^2=S_\Phi^2/(\nu-1)=0{,}315$ (при выполнении гипотезы H_0 (12.24))
Остаточные факторы	$\hat{\sigma}_o^2=\overline{\hat{\sigma}_i^2}=S_o^2/n= =0{,}841/60=0{,}014$	$n-\nu= =60-5=55$	$s_o^2=S_o^2/(n-\nu)=0{,}015$
Общая вариация	$\hat{\sigma}_Y^2=S_Y^2/n=2{,}1/60= =0{,}035$	$n-1= =60-1=59$	$s_Y^2=S_Y^2/(n-1)=0{,}036$ (при выполнении гипотезы H_0 (12.24))

Найдем числовое значение критерия (11.31):

$$F_{\text{чис}}=\frac{s_\Phi^2}{s_o^2}=\frac{0{,}315}{0{,}015}=21.$$

Теперь по схеме (11.32) определим правостороннюю критическую точку, например при уровне значимости $\alpha=0{,}05$. Имеем

$$\left.\begin{array}{l}\alpha=0{,}05\to\gamma=1-\alpha=0{,}95\\ n=60,\ \nu=5 \to \begin{cases} l=\nu-1=4\\ k=n-\nu=55\end{cases}\end{array}\right\}\xrightarrow{\text{п. 7}} 2{,}61>f_\gamma>2{,}52\to x_{\text{пр.}\alpha}^{\text{кр}}=f_\gamma.$$

Так как $F_{\text{чис}}>x_{\text{пр.}\alpha}^{\text{кр}}$, то гипотезу H_0 (12.24) отвергаем, т. е. считаем, что $\rho_{Y/X}\neq 0$; это означает, что при изменеиии x изменяются условные математические ожидания $M(Y/X=x)$, или, иначе говоря, что существует корреляционная зависимость выпуска продукции от стоимости основных фондов. Напомним, что такой вывод может оказаться ошибочным: на самом деле гипотеза (12.24) верна; вероятность появлеиия такой ошибки равиа 0,05. ▲

В заключение заметим, что если гипотеза (12.24) отвергается, то говорят, что *выборочное корреляционное отношение статистически значимо*. Если гипотеза (12.24) не отвергается, то говорят, что выборочное корреляционное отношение незначимо.

Обратим внимание на то, что вычислить выборочное корреляционное отношение, а также проверить его значимость можно только в том случае, когда результаты наблюдений сгруппированы в таблицу типа табл. 12.11.

Итак, допустим, что, располагая выборочными данными, мы пришли к выводу, что корреляционная зависимоть Y от X существует, т. е. при изменении x изменяются условные математические ожидания $M(Y/X=x)$. Тогда возникает вопрос: каков вид функции регрессии, т. е. функции $\varphi(x)=M(Y/X=x)$? Конечно, располагая только выборочными данными, нельзя дать точный ответ на поставленный вопрос, но высказать гипотезу о виде функции $\varphi(x)$ можно; также можно провести

статистическую проверку этой гипотезы, т. е. выяснить, противоречит или нет эта гипотеза имеющимся выборочным данным. Мы рассмотрим только схему проверки гипотезы о том, что функция регрессии линейная, т. е. $\varphi(x)=a+bx$ (это сделано в § 12.7). В § 12.4 покажем, что при линейной функции регрессии алгоритм вычисления корреляционного отношения значительно упрощается.

§ 12.4. Линейная функция регрессии. Генеральный коэффициент корреляции

Итак, допустим, что при изменении x условное математическое ожидание $M(Y/X=x)$ изменяется по линейному закону, т. е. функция регрессии $\varphi(x)=M(Y/X=x)$ линейная:

$$M^{\text{лин}}(Y/X=x)=a+bx.$$

Найдем для этого случая сначала выражение для *параметров* a и b и линейной функции регрессии, а затем выражение для корреляционного отношения (12.6). При этом договоримся используемые обозначения снабжать индексом «лин», что означает «при условии линейной функции регрессии».

Выражения для параметров a и b и линейной функции регрессии. Обратимся к формуле (12.1) условной дисперсии. В случае линейной функции регрессии формула принимает вид

$$D^{\text{лин}}(Y/X=x)=M\left[(Y/X=x)-M^{\text{лин}}(Y/X=x)\right]^2= =M\left[(Y/X=x)-a-bx\right]^2. \tag{12.27}$$

Напомним, в общем случае при изменении x условная дисперсия $D^{\text{лин}}(Y/X=x)$ изменяется. Найдем характеристику (12.4) разброса «игреков», вызванного влиянием на Y остаточных факторов, она примет вид

$$\sigma_o^{2\,\text{лин}}=M\left[D^{\text{лин}}(Y/X)\right]=M(M(Y/X)-a-bX)^2 \tag{12.28}$$

—эта величина при фиксированных значениях параметров a и b является постоянной.

Принимая во внимание свойство (см. (5.18)) минимальности дисперсии, значения параметрв a и b находят из условия

$$F(a,\,b)=M(M(Y/X)-a-bX)^2\to\min. \tag{12.29}$$

Окончательный результат такой:

$$b=r_{XY}\sigma_Y/\sigma_X, \tag{12.30}$$

$$a=m_Y-r_{XY}\frac{\sigma_Y}{\sigma_X}m_X, \tag{12.31}$$

где $m_X=MX,\ m_y=MY,\ r_{XY}=\dfrac{K_{XY}}{\sigma_X\sigma_Y}=\dfrac{M[(X-MX)(Y-MY)]}{\sigma_X\sigma_Y}$.

Отметим, что выражение

$$K_{XY}=M[(X-MX)(Y-MY)] \qquad (12.32)$$

называют *генеральным корреляционным моментом*, а

$$r_{XY}=\frac{K_{XY}}{\sigma_X\sigma_Y}=\frac{M[(X-MX)(Y-MY)]}{\sigma_X\sigma_Y} \qquad (12.33)$$

генеральным коэффициентом корреляции.

Подставив найденные значения параметров a и b в линейную функцию регрессии $M^{\text{лин}}\ (Y/X=x)=a+bx$, получим

$$M^{\text{лин}}(Y/X=x)=m_Y+r_{XY}\frac{\sigma_Y}{\sigma_X}(x-m_X). \qquad (12.34)$$

Выражение для корреляционного отношения. Напомним вид генерального корреляционного отношения

$$\rho_{Y/X}=+\sqrt{\sigma_\Phi^2/\sigma_Y^2}=\sqrt{1-\sigma_o^2/\sigma_Y^2}.$$

Установим, как преобразуются выражения для σ_o^2, σ_Φ^2 и $\rho_{Y/X}$, если учесть, что функция регрессии имеет вид (12.34).

Согласно (12.3),

$$\sigma_\Phi^2=D[M(Y/X)],$$

поэтому

$$\sigma_\Phi^{2\,\text{лин}}=D[M^{\text{лин}}(Y/X)]\underset{(12.34)}{=\!=\!=}D\left[m_Y+r_{XY}\frac{\sigma_Y}{\sigma_X}(X-m_X)\right]=$$

$$\underset{m_Y=\text{const}}{=\!=\!=}D\left[r_{XY}\frac{\sigma_Y}{\sigma_X}(X-m_X)\right]=$$

$$\underset{r,\sigma_Y,\sigma_X=\text{const}}{=\!=\!=}r_{XY}^2\frac{\sigma_Y^2}{\sigma_X^2}D(X-m_X)=$$

$$\underset{m_X=\text{const}}{=\!=\!=}r_{XY}^2\frac{\sigma_Y^2}{\sigma_X^2}DX=r_{XY}^2\sigma_Y^2.$$

Итак,

$$\sigma_\Phi^{2\,\text{лин}}=D[M^{\text{лин}}(Y/X)]=r_{XY}^2\sigma_Y^2. \qquad (12.35)$$

Теперь, используя соотношение (12.5), определим вид выражения $\sigma_o^{2\,\text{лин}}=M\ [D^{\text{лин}}(Y/X)]$ при линейной функции регрессии.

Имеем

$$\sigma_o^{2\,\text{лин}}=\sigma_Y^2-\sigma_\Phi^{2\,\text{лин}}\underset{(12.35)}{=\!=\!=}\sigma_Y^2-r_{XY}^2\sigma_Y^2=\sigma_Y^2(1-r_{XY}^2).$$

Итак,

$$\sigma_o^{2\,\text{лин}}=M[D^{\text{лин}}(Y/X)]=(1-r_{XY}^2)\sigma_Y^2. \qquad (12.36)$$

Для генерального корреляционного отношения получим следующее выражение:

$$\rho_{Y/X}^{\text{лин}}=+\sqrt{\frac{\sigma_\Phi^{2\,\text{лин}}}{\sigma_Y^2}}=+\sqrt{\frac{r_{XY}^2\sigma_Y^2}{\sigma_Y^2}}=|r_{XY}|, \qquad (12.37)$$

где r_{XY} — коэффициент корреляции, вычисляемый по формуле (12.30).

Свойства коэффициента корреляции как измерителя степени линейности стохастической зависимости. Итак, согласно (12.33), генеральный коэффициент корреляции

$$r_{XY}=\frac{M[(X-MX)(Y-MY)]}{\sigma_X\sigma_Y},$$

при этом если функция регрессии линейна и имеет вид (12.34), то $|r_{XY}|$ совпадает с генеральным корреляционным отношением $\rho_{Y/X}$. Таким образом, $|r_{XY}|$ — это частный случай $\rho_{Y/X}$, имеющий место при линейной функции регрессии, поэтому

$$|r_{XY}|\leqslant\rho_{Y/X} \qquad (12.38)$$

и $|r_{XY}|$ будет обладать свойствами корреляционного отношения, «подправленными» с учетом линейности функции регрессии.

Сформулируем эти свойства.

1^0. $|r_{XY}|\leqslant 1$ *или* $-1<r_{XY}<1$.

Замечание. В отличие от корреляционного отношения, которое не может быть отрицательным, r_{XY} может иметь как знак «+», так и «−»; как видно из (12.34), знак «+» означает, что с ростом значений X увеличивается условное математическое ожидание $M(Y/X=x)$ (рис. 12.3, *а*); отрицательное значение r_{XY} говорит о противоположной тенденции (рис. 12.3, *б*).

2^0. *Если корреляционная зависимость Y от X отсутствует, то* $r_{XY}=0$.

Действительно, при отсутствии корреляционной зависимости Y от X коэффициент $\rho_{Y/X}=0$, поэтому в силу (12.38) и $r_{XY}=0$.

Замечание. Обратное же утверждение верно не всегда: из равенства $r_{XY}=0$ не всегда следует, что $\rho_{Y/X}=0$, или, иначе говоря, не всегда следует отсутствие корреляционной зависимости. Только в том случае, когда функция регрессии линейна и имеет вид (12.34), из равенства $r_{XY}=0$ следует отсутствие корреляционной зависимости Y от X. Действительно, подставив $r_{XY}=0$ в (12.34), получим, что $M(Y/X=x)=m_Y=\text{const}$ при любом x.

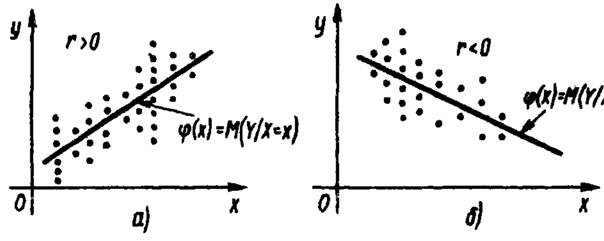


Рис. 12.3

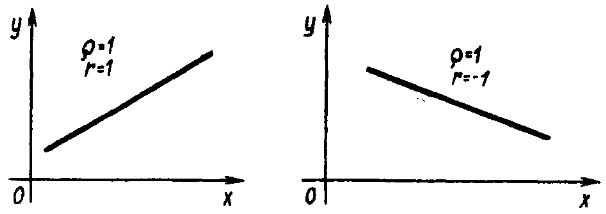


Рис. 12.4

3^0. *Условие $|r_{XY}|=1$ является необходимым и достаточным для существования линейной функциональной зависимости между Y и X.*

Замечание. Коэффициент корреляции симметричен относительно X и Y, т. е. $r_{XY}=r_{YX}$. Это следует из формулы (12.33). Если $|r_{XY}|=1$, то $|r_{YX}|=1$, поэтому вместо выражения «линейная функциональная зависимость Y от X» мы употребим «линейная функциональная зависимость между Y и X».

Графически свойство 3^0 проиллюстрировано на рис. 12.4.

Следствие. *Чем ближе $|r_{XY}|$ к единице, тем ближе стохастическая зависимость между величинами X и Y к линейной функциональной, или, иначе говоря, выше степень линейности стохастической зависимости. И, наоборот, чем выше степень линейности стохастической зависимости между величинами X и Y, тем ближе $|r_{XY}|$ к единице.*

Проиллюстрируем следствие диаграммами рассеивания, приведенными на рис. 12.5. Поясним этот рисунок:

а) линейная стохастическая зависимость довольно сильная; точки сконцентрированы в относительно узкой линейной полосе; $r_1 \approx 1$;

б) линейная стохастическая зависимость по сравнению с зависимостью, изображенной с рис. 12.5, *а*, менее сильная: точки концентрируются в более широкой линейной полосе; $r_2 < r_1$.

○ **Пример 12.2.** Вычислим коэффициент корреляции между случайными величинами X и Y в условиях примера 12.1.

Прежде чем вычислять коэффициент корреляции r_{XY}, преобразуем числитель формулы (12.33):

$$M[(X-m_X)(Y-m_Y)]=M(XY-m_XY-m_YX+m_Xm_Y)\underset{m_X,m_Y=\text{const}}{=\!=\!=\!=}M(XY)-$$

$$-m_XMY-m_YMX+m_Xm_Y=M(XY)-MXMY.$$

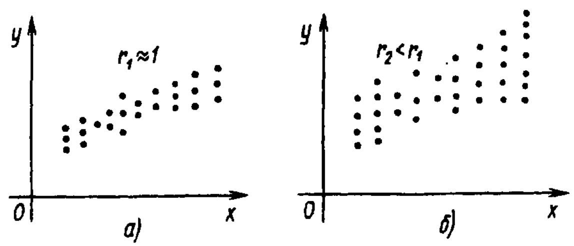


Рис. 12.5

Таким образом,

$$r_{XY}=\frac{M(XY)-MXMY}{\sigma_X\sigma_Y}. \tag{12.39}$$

По данным табл. 12.1 были найдены следующие характеристики:

$MX=5{,}54$, $DX=4{,}9284$ (см. табл. 12.2);
$MY=0{,}68$, $DY=0{,}0336$ (см. табл. 12.3).

По табл. 12.1 находим

$$M(XY)=\sum_{i=1}^{3}\sum_{j=1}^{2}x_iy_jP[(X=x_i)(Y=y_j)]=$$

$$=2\cdot0{,}4\cdot0{,}15+2\cdot0{,}8\cdot0{,}05+5\cdot0{,}4\cdot0{,}12+5\cdot0{,}8\cdot0{,}30+8\cdot0{,}4\cdot0{,}03+8\cdot0{,}8\cdot0{,}35=3{,}92.$$

Тогда

$$r_{XY}=\frac{3{,}92-5{,}54\cdot0{,}68}{\sqrt{4{,}9284}\sqrt{0{,}0336}}=0{,}375.$$

Напомним, что для табл. 12.1 корреляционное отношение $\rho_{Y/X}=0{,}53$. Мы получили, что $|r_{XY}|\neq\rho_{Y/X}$; это объясняется тем, что в примере 12.1 функция регрессии, заданная табл. 12.7, не является линейной (прирост «иксов» постоянен — равен трем, а прирост условных математических ожиданий изменяется). ●

§ 12.5. Поле корреляции. Выборочный коэффициент корреляции

Как уже отмечалось, при решении практических задач обычно не располагают сведениями о законе распределения двумерной случайной величины (X, Y), поэтому нельзя знать генеральное корреляционное отношение и коэффициент корреляции, а также вид функции регрессии.

В § 12.3 было показано, как вычисляется корреляционное отношение по результатам наблюдений величины (X, Y) и как, зная это отношение, проверить гипотезу об отсутствии корреляционной зависимости в генеральной совокупности, т. е. о том, что $M(Y/X=x)=\text{const}$ при любом допустимом x.

Если эта гипотеза не принимается, то возникает вопрос о виде функции регрессии $\varphi(x)=M(Y/X=x)$. Конечно, располагая только результатами наблюдений, точный ответ на поставленный вопрос дать нельзя. Можно лишь сформулировать гипотезу о виде функции регрессии. Для того чтобы сформулировать эту гипотезу, поступают следующим образом.

Результаты наблюдений группируют в табл. 12.9, по которой строят *поле корреляции* — это прямоугольная сетка, в каждом прямоугольнике (j, i) которого проставляют m_{ji} точек. Затем на поле корреляции наносят точки с координатами $(x_1, \bar{Y}^{(1)})$, $(x_2, \bar{Y}^{(2)})$, ..., $(x_\nu, \bar{Y}^{(\nu)})$ и соединяют их отрезками. Полученная линия дает представление об изменении групповых средних при изменении x. Так как групповая средняя — это выборочный аналог условного математического ожидания, то по поведению *линии групповых средних* можно составить некоторое представление о виде функции регрессии.

Дополнительно вычисляют *выборочный коэффициент корреляции*

$$\hat{r}_{XY}=\frac{\overline{(X-\bar{X})(Y-\bar{Y})}}{\hat{\sigma}_X\hat{\sigma}_Y} \qquad (12.40)$$

(выборочный аналог генерального коэффициента корреляции (12.33): математические ожидания заменены на средние, а генеральные средние квадратические отклонения — выборочными средними квадратическими отклонениями).

Выполним тождественные преобразования числителя выражения (12.40):

$$\overline{(X-\bar{X})(Y-\bar{Y})}=\overline{XY-\bar{X}Y-\bar{Y}X+\bar{X}\bar{Y}}=\overline{XY}-\overline{\bar{X}Y}-\overline{\bar{Y}X}+\overline{\bar{X}\bar{Y}}=$$

$$\underset{\substack{\bar{X},\,\bar{Y}=\text{const (для}\\ \text{фиксированной}\\ \text{выборки)}}}{=\!=\!=\!=\!=\!=\!=}\ \overline{XY}-\bar{X}\bar{Y}-\bar{Y}\bar{X}+\bar{X}\bar{Y}=\overline{XY}-\bar{X}\bar{Y}.$$

Получаем следующую, более удобную для вычислений выборочного коэффициента корреляции, формулу:

$$\hat{r}_{XY}=\frac{\overline{XY}-\bar{X}\bar{Y}}{\hat{\sigma}_X\hat{\sigma}_Y}. \qquad (12.41)$$

Если результаты $(X_i,\ Y_i)$ наблюдений величины $(X,\ Y)$ не сгруппированы в корреляционную таблицу, то

$$\overline{XY}=\frac{1}{n}\sum_{i=1}^{n}X_iY_i,\quad \bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad \bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i,$$

$$\hat{\sigma}_X=\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2-(\bar{X})^2},\quad \hat{\sigma}_Y=\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2-(\bar{Y})^2};$$

Если наблюдения сгруппированы в табл. 12.9, то

$$\overline{XY}=\frac{1}{n}\sum_{i=1}^{v}\sum_{j=1}^{q}x_iy_jm_{ji}=\frac{1}{n}\sum_{i=1}^{v}x_i\sum_{j=1}^{q}y_jm_{ji}; \qquad (12.42)$$

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{v}x_in_i,\quad \bar{Y}=\frac{1}{n}\sum_{j=1}^{q}y_jm_j,$$

$$\hat{\sigma}_X=\sqrt{\frac{1}{n}\sum_{i=1}^{v}x_i^2n_i-(\bar{X})^2},\quad \hat{\sigma}_Y=\sqrt{\frac{1}{n}\sum_{j=1}^{q}y_j^2m_j-(\bar{Y})^2}.$$

Вычислив $\hat{r}_{XY}$, сравнивают $|\hat{r}_{XY}|$ с выборочным корреляционным отношением $\hat{\rho}_{Y/X}$. Если их значения близки, то это дает основание выдвинуть гипотезу о том, что в генеральной совокупности функция регрессии линейная, т. е. $M(Y/X=x)=$ $=a+bx$.

Задача 12.3. По результатам наблюдений, сгруппированным в табл. 12.11, изобразить поле корреляции и линию групповых средних объема выпускаемой продукции. Вычислить коэффициент корреляции. Сформулировать гипотезу о виде функции регрессии.

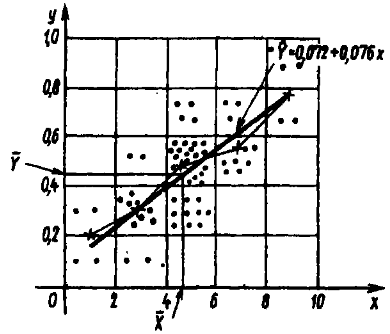


Рис. 12.6

△ Поле корреляции и ломаная линия групповых средних (эти средние вычислены в строке (3) табл. 12.11) изображены на рис. 12.6.

Из рис. 12.6 делаем следующие выводы.

С ростом стоимости основных фондов (x_i) средний объем выпуска продукции $(\bar{Y}^{(i)})$ увеличивается.

Концентрация точек около линии групповых средних довольно тесная. Это говорит о том, что, по-видимому, в генеральной совокупности высока степень стохастической зависимости величины Y от X, т. е. значение генерального корреляционного отношения $\rho_{Y/X}$ довольно большое (напомним, что выборочное корреляционное отношение $\hat{\rho}_{Y/X}=0{,}783$).

Линия групповых средних испытывает небольшие колебания относительно прямой, что позволяет выдвинуть гипотезу о том, что функция регрессии $\varphi(x)=M(Y/X=x)$ имеет линейный вид.

Концентрация точек около прямой, «выравнивающей» линию групповых средних, довольно тесная. Это говорит о том, что, по-видимому, в генеральной совокупности высока степень линейной стохастической зависимости, т. е. значение генерального коэффициента корреляции r_{XY} довольно большое.

Вычислим выборочный коэффициент корреляции по формуле (12.41). Среднее $\overline{XY}=2{,}503$; оно рассчитано по формуле (12.42) в строке (12) табл. 12.11. Среднее $\bar{X}$ и дисперсия $\hat{\sigma}_X^2$ величины X вычислены в строках (10) и (11): $\bar{X}=4{,}95$; $\hat{\sigma}_X^2=3{,}499$. Среднее $\bar{Y}$ и дисперсия $\hat{\sigma}_Y^2$ величины Y вычислены в столбцах (2) и (3): $\bar{Y}=0{,}45$; $\hat{\sigma}_Y^2=0{,}035$. Тогда

$$\hat{r}_{XY}=\frac{2{,}503-4{,}97\cdot 0{,}45}{\sqrt{3{,}499}\sqrt{0{,}035}}=0{,}76.$$

Так как значение $|\hat{r}_{XY}|$ близко к $\hat{\rho}_{Y/X}=0{,}78$, то это, так же как и поведение линии групповых средних, дает основание выдвинуть гипотезу о том, что функция регрессии линейная (напомним, что если функция регрессии линейная, то модуль генерального коэффициента корреляции равен генеральному корреляционному отношению, т. е. $|r_{XY}|=\rho_{Y/X}$). ▲

Однако прежде чем рассматривать, как проводится проверка гипотезы о том, что функция регрессии линейная, т. е. $M(Y/X=x)=a+bx$, покажем, как найти по выборочным данным оценки $\hat{a}$ и $\hat{b}$ параметров a и b линейной функции регрессии.

§ 12.6. Метод наименьших квадратов. Линейное уравнение регрессии

Итак, пусть функция регрессии линейная, т. е. $M(Y/X=x)=$ $=a+bx$. Найдем оценки $\hat{a}$ и $\hat{b}$ параметров a и b. Критерием нахождения оценок $\hat{a}$ и $\hat{b}$ является следующее требование: средняя квадратов отклонений наблюдаемых «игреков» от «игреков», рассчитанных по уравнению $\hat{Y}=\hat{a}+\hat{b}x$, должна быть минимальной. Запишем это требование в виде формулы

$$\hat{F}(\hat{a}, \hat{b})=\overline{(Y-\hat{Y})^2} \to \min. \qquad (12.43)$$

Метод нахождения значений оценок $\hat{a}$ и $\hat{b}$ (в соответствии с требованием (12.43)) называется *методом наименьших квадратов.*

Для результатов (X_i, Y_i) наблюдений величины (X, Y), не сгруппированных в корреляционную табл. 12.9, критерий (12.43) имеет вид

$$\hat{F}(\hat{a}, \hat{b})=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{a}-\hat{b}X_i)^2 \to \min. \qquad (12.44)$$

Если наблюдения сгруппированы в табл. 12.9, то критерий (12.43) принимает следующий вид:

$$\hat{F}(\hat{a}, \hat{b})=\frac{1}{n}\sum_{i=1}^{\upsilon}\sum_{j=1}^{q}(y_j-\hat{a}-\hat{b}x_i)^2 m_{ji} \to \min. \qquad (12.45)$$

Разумеется, значения $\hat{a}$ и $\hat{b}$, найденные в соответствии с критериями (12.44) и (12.45), будут одинаковыми. Критерий (12.44) имеет более простую форму, поэтому значения $\hat{a}$ и $\hat{b}$ найдем исходя из него.

Необходимые условия минимума функции $\hat{F}(\hat{a}, \hat{b})$ образуют систему

$$\begin{cases} \dfrac{\partial \hat{F}}{\partial a}=\dfrac{1}{n}\sum\limits_{i=1}^{n}2(Y_i-\hat{a}-\hat{b}X_i)(-1)=0, \\ \dfrac{\partial \hat{F}}{\partial b}=\dfrac{1}{n}\sum\limits_{i=1}^{n}2(Y_i-\hat{a}-\hat{b}X_i)(-X_i)=0, \end{cases}$$

которая в результате тождественных преобразований принимает вид

$$\begin{cases} \hat{a}+\hat{b}\bar{X}=\bar{Y}, \\ \hat{a}\bar{X}+\hat{b}\overline{X^2}=\overline{YX}, \end{cases} \qquad (12.46)$$

где

$$\bar{X}=\sum_{i=1}^{n}X_i/n,\ \bar{Y}=\sum_{i=1}^{n}Y_i/n,\ \overline{X^2}=\sum_{i=1}^{n}X_i^2/n,\ \overline{YX}=\sum_{i=1}^{n}Y_iX_i/n.$$

Система (12.46) называется *нормальной системой уравнений.* Решим ее относительно $\hat{a}$ и $\hat{b}$. Из первого уравнения находим $\hat{a}=\bar{Y}-\hat{b}\bar{X}$. Подставив это выражение во второе из уравнений, получим

$$(\bar{Y}-\hat{b}\bar{X})\bar{X}+\hat{b}\overline{X^2}=\overline{YX},$$

откуда находим

$$\hat{b}=\frac{\overline{YX}-\bar{Y}\bar{X}}{\overline{X^2}-(\bar{X})^2},$$

или, учитывая (12.41),

$$\hat{b}=\hat{r}_{XY}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}. \qquad (12.47)$$

Тогда

$$\hat{a}=\bar{Y}-\hat{r}_{XY}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\bar{X}. \qquad (12.48)$$

Подставив выражения для $\hat{a}$ и $\hat{b}$ в уравнение $\hat{Y}=\hat{a}+\hat{b}x$, получим

$$\hat{Y}=\bar{Y}+\hat{r}_{XY}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}(x-\bar{X}). \qquad (12.49)$$

Уравнение (12.49) называется *выборочным линейным уравнением регрессии.*

Пусть $x=x_i$. Тогда

$$\hat{Y}_i=\bar{Y}+\hat{r}_{XY}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}(x_i-\bar{X})$$

—это оценка условного математического ожидания, вычисляемого по формуле (12.34)

$$M^{\text{лин}}(Y/X=x_i)=MY+r_{XY}\frac{\sigma_Y}{\sigma_X}(x_i-MX).$$

Замечание. Выше в качестве оценок условных математических ожиданий $M(Y/X=x_i)$ использовались групповые средние $\bar{Y}^{(i)}$—они находились по результатам n_i наблюдений в i-й, $i=1, 2, \ldots, \upsilon$, группе (см. табл. 12.9). Теперь в качестве оценки предлагается использовать $\hat{Y}_i$. Подсчитаем средний квадрат разности между $\bar{Y}^{(i)}$ и $\hat{Y}_i$, учитывая при этом, что в i-й группе число наблюдений равно n_i:

$$\overline{(\bar{Y}^{(i)}-\hat{Y}_i)^2}=\frac{1}{n}\sum_{i=1}^{\upsilon}(\bar{Y}^{(i)}-\hat{Y}_i)^2 n_i=\frac{1}{n}\sum_{i=1}^{\upsilon}\left[\left(\bar{Y}^{(i)}-\left(\bar{Y}+\hat{r}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}(x_i-\bar{X})\right)\right)\right]^2 n_i=$$

$$=\frac{1}{n}\sum_{i=1}^{\upsilon}\left[(\bar{Y}^{(i)}-\bar{Y})-\hat{r}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}(x_i-\bar{X})\right]^2 n_i=\underbrace{\frac{1}{n}\sum_{i=1}^{\upsilon}(\bar{Y}^{(i)}-\bar{Y})^2 n_i}_{\hat{\sigma}_\Phi^2}+\hat{r}^2\frac{\hat{\sigma}_Y^2}{\hat{\sigma}_X^2}\underbrace{\frac{1}{n}\sum_{i=1}^{\upsilon}(x_i-\bar{X})^2 n_i}_{\hat{\sigma}_X^2}-$$

$$-2\hat{r}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\frac{1}{n}\sum_{i=1}^{\upsilon}(\bar{Y}^{(i)}-\bar{Y})(x_i-\bar{X})n_i=\hat{\sigma}_\Phi^2+\hat{r}^2\frac{\hat{\sigma}_Y^2}{\hat{\sigma}_X^2}\hat{\sigma}_X^2-2\hat{r}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\frac{1}{n}\sum_{i=1}^{\upsilon}(\underbrace{\bar{Y}^{(i)}n_i}_{\sum_{j=1}^{q}y_j m_{ji}}x_i-\bar{Y}x_i n_i-$$

$$-\bar{Y}^{(i)}\bar{X}n_i+\bar{Y}\bar{X}n_i)=\hat{\sigma}_\Phi^2+\hat{r}^2\hat{\sigma}_Y^2-2\hat{r}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\Bigg(\underbrace{\frac{1}{n}\sum_{i=1}^{\upsilon}\sum_{j=1}^{q}y_j m_{ji}x_i}_{\overline{XY}}-\bar{Y}\underbrace{\frac{1}{n}\sum_{i=1}^{\upsilon}x_i n_i}_{\bar{X}}-$$

$$-\bar X\underbrace{\frac{1}{n}\sum_{i=1}^{l}\bar Y^{(i)}n_i}_{\bar Y}+\bar Y\bar X\underbrace{\frac{1}{n}\sum_{i=1}^{v}n_i}_{1}\Bigg)\underset{(12.20)}{=}\hat\rho^2_{Y/X}\hat\sigma^2_Y+\hat r^2\hat\sigma^2_Y-2\hat r\frac{\hat\sigma_Y}{\hat\sigma_X}\underbrace{(\overline{XY}-\bar X\bar Y)}_{\hat r\hat\sigma_Y\hat\sigma_X\ (\text{см. }(12.41))}=$$

$$=\hat\rho^2_{Y/X}\hat\sigma^2_Y-\hat r^2\hat\sigma^2_Y=\hat\sigma^2_Y(\hat\rho^2_{Y/X}-\hat r^2).$$

Итак,

$$\overline{(\bar Y^{(i)}-\hat Y_i)^2}=\hat\sigma^2_Y(\hat\rho^2_{Y/X}-\hat r^2_{XY}). \qquad (12.50)$$

Обратим внимание на следующее: так как левая часть равенства (12.50) неотрицательна, то и правая его часть $\hat\sigma^2_Y(\hat\rho^2_{Y/X}-\hat r^2_{XY})\geqslant 0$. Отсюда следует, что

$$\hat\rho^2_{Y/X}\geqslant\hat r^2_{XY}. \qquad (12.51)$$

§ 12.7. Погрешность выборочного линейного уравнении регрессии. Смысл выборочного коэффициента корреляции, его значимость

Располагая результатами (X_i, Y_i) n наблюдений величины (X, Y) мы, используя метод наименьших квадратов, получили линейное уравнение регрессии

$$\hat Y=\bar Y+\hat r_{XY}\frac{\hat\sigma_Y}{\hat\sigma_X}(x-\bar X). \qquad (12.52)$$

Приравняв $x=X_i$, получим

$$\hat Y_i=\bar Y+\hat r_{XY}\frac{\hat\sigma_Y}{\hat\sigma_X}(X_i-\bar X). \qquad (12.53)$$

Вычислим средний квадрат отклонения наблюдаемых «игреков» (Y_i) от «игреков» $(\hat Y_i)$, рассчитанных по уравнению регрессии:

$$\hat\sigma_o^{2\,\text{лин}}=\overline{(Y-\hat Y)^2}=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat Y_i)^2.$$

Средней квадратической погрешностью или ошибкой уравнения регрессии называется величина

$$\hat\sigma_o^{\text{лин}}=+\sqrt{\sum_{i=1}^{n}(Y_i-\hat Y_i)^2/n}. \qquad (12.54)$$

Замечание. Заменяя наблюдаемые значения Y_i на $\hat Y_i$, мы совершаем ошибку

$$\hat\sigma_o^{\text{лин}}=\sqrt{\sum_{i=1}^{n}(Y_i-\hat Y_i)^2/n};$$

заменяя Y_i на $\bar Y$, мы совершаем ошибку

$$\hat\sigma_Y=\sqrt{\sum_{i=1}^{n}(Y_i-\bar Y_i)^2/n}.$$

Таким образом, отношение $\hat\sigma_Y/\hat\sigma_o^{\text{лин}}$ показывает, во сколько раз погрешность модели $Y_i\approx\hat Y_i$ меньше погрешности модели $Y_i\approx\bar Y$.

Введем также величину

$$\hat\sigma_{\hat Y}^{2\,\text{лин}}=\overline{(\hat Y-\overline{\hat Y})^2}=\sum_{i=1}^{n}(\hat Y_i-\overline{\hat Y})^2/n,$$

где $\overline{\hat Y}=\sum_{i=1}^{n}\hat Y_i/n$, которая называется *выборочной дисперсией «игреков», рассчитанных по уравнению регрессии* (12.52).

Докажем следующее тождество:

$$\hat\sigma^2_Y=\hat\sigma_{\hat Y}^{2\,\text{лин}}+\hat\sigma_o^{2\,\text{лин}}. \qquad (12.55)$$

□ Найдем более простое выражение для $\hat\sigma_{\hat Y}^{2\,\text{лин}}=\overline{(\hat Y-\overline{\hat Y})^2}$. Убедимся сначала, что $\overline{\hat Y}=\bar Y$. Действительно,

$$\overline{\hat Y}=\frac{1}{n}\sum_{i=1}^{n}\hat Y_i\underset{(12.53)}{=}\frac{1}{n}\sum_{i=1}^{n}\left[\bar Y+\hat r_{XY}\frac{\hat\sigma_Y}{\hat\sigma_X}(X_i-\bar X)\right]=\frac{1}{n}\underbrace{\sum_{i=1}^{n}\bar Y}_{n\bar Y}+$$

$$+\hat r_{XY}\frac{\hat\sigma_Y}{\hat\sigma_X}\Big(\underbrace{\frac{1}{n}\sum_{i=1}^{n}X_i}_{\bar X}-\frac{1}{n}\underbrace{\sum_{i=1}^{n}\bar X}_{n\bar X}\Big)=\bar Y,$$

поэтому

$$\hat\sigma_{\hat Y}^{2\,\text{лин}}=\overline{(\hat Y-\bar Y)^2},$$

или

$$\hat\sigma_{\hat Y}^{2\,\text{лин}}=\frac{1}{n}\sum_{i=1}^{n}(\hat Y_i-\bar Y)^2\underset{(12.53)}{=}\frac{1}{n}\sum_{i=1}^{n}(\bar Y+\hat r\frac{\hat\sigma_Y}{\hat\sigma_X}(X_i-\bar X)-\bar Y)^2=$$

$$=\hat r^2\frac{\hat\sigma^2_Y}{\hat\sigma^2_X}\underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar X)^2}_{\hat\sigma^2_X}=\hat r^2\hat\sigma^2_Y.$$

Итак,

$$\hat\sigma_{\hat Y}^{2\,\text{лин}}=\hat r^2_{XY}\hat\sigma^2_Y. \qquad (12.56)$$

Найдем теперь более простое выражение для $\hat\sigma_o^{2\,\text{лин}}$. Имеем

$$\hat\sigma_o^{2\,\text{лин}}=\sum_{i=1}^{n}(Y_i-\hat Y_i)^2/n\underset{(12.53)}{=}\frac{1}{n}\sum_{i=1}^{n}[Y_i-(\bar Y+\hat r\frac{\hat\sigma_Y}{\hat\sigma_X}(X_i-\bar X))]^2=$$

$$=\frac{1}{n}\sum_{i=1}^{n}[(Y_i-\bar Y)-\hat r\frac{\hat\sigma_Y}{\hat\sigma_X}(X_i-\bar X)]^2=\underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar Y)^2}_{\hat\sigma^2_Y}+$$

$$+\hat r^2\frac{\hat\sigma_Y^2}{\hat\sigma_X^2}\underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar X)^2}_{\hat\sigma_X^2}-2\hat r\frac{\hat\sigma_Y}{\hat\sigma_X}\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar Y)(X_i-\bar X)=\hat\sigma_Y^2+\hat r^2\hat\sigma_Y^2-$$

$$-2\hat r\frac{\hat\sigma_Y}{\hat\sigma_X}\underbrace{\overline{(Y-\bar Y)(X-\bar X)}}_{\hat r\hat\sigma_X\hat\sigma_Y\ (\text{см. } (12.40))}=\hat\sigma_Y^2+\hat r^2\hat\sigma_Y^2-2\hat r^2\hat\sigma_Y^2=\hat\sigma_Y^2(1-\hat r^2).$$

Итак,

$$\hat\sigma_{\text{о}}^{2\,\text{лин}}=\overline{(\hat Y-Y)^2}=(1-\hat r_{XY}^2)\,\hat\sigma_Y^2. \tag{12.57}$$

Сложив (12.56) и (12.57), получим

$$\hat\sigma_{\hat Y}^{2\,\text{лин}}+\hat\sigma_{\text{о}}^{2\,\text{лин}}=\hat r_{XY}^2\hat\sigma_Y^2+(1-\hat r_{XY}^2)\hat\sigma_Y^2=\hat\sigma_Y^2. \ \blacksquare$$

Из соотношения (12.56) имеем

$$\hat r_{XY}^2=\frac{\hat\sigma_{\hat Y}^{2\,\text{лин}}}{\hat\sigma_Y^2}=\frac{\overline{(\hat Y-\bar Y)^2}}{\overline{(Y-\bar Y)^2}} \tag{12.58}$$

(так как $\hat\sigma_{\hat Y}^{2\,\text{лин}}\leqslant\hat\sigma_Y^2$, то $\hat r_{XY}^2\leqslant 1$).

Таким образом, $\hat r_{XY}^2$ *показывает, какую долю выборочной дисперсии* $\hat\sigma_Y^2$ *«игреков» составляет выборочная дисперсия* $\hat\sigma_{\hat Y}^{2\,\text{лин}}$ *«игреков», вычисленных по линейному уравнению регрессии, или, иначе говоря, какая доля выборочной дисперсии* $\hat\sigma_Y^2$ *объясняется линейной в среднем зависимостью «игреков» от «иксов».*

Тождество (12.55) является основным при проверке гипотезы

$$H_0: r_{XY}=0 \tag{12.59}$$

о том, что генеральный коэффициент корреляции равен нулю.

Гипотезу (12.59) проверяют методом однофакторного дисперсионного анализа, в котором *фактор — это линейная функция регрессии.* Обратим внимание на то, что если в линейной функции регрессии

$$M^{\text{лин}}(Y/X=x)=MY+r_{XY}\frac{\sigma_Y}{\sigma_X}(x-MX)$$

коэффициент корреляции $r_{XY}=0$, то $M^{\text{лин}}(Y/X=x)=MY$ — это означает, что при любом x математическое ожидание $M^{\text{лин}}(Y/X=x)$ остается постоянным, равным MY. Таким образом, проверка гипотезы (12.59) *при линейной функции регрессии (только в этом случае!)* равносильна проверке гипотезы о постоянстве условных (или групповых) математических ожиданий, а это задача дисперсионного анализа.

Допустим, что выполняются условия (12.25) и (12.26), необходимые для проведения дисперсионного анализа. Дисперсионная табл. 11.2 имеет вид табл. 12.13.

Таблица 12.13

Источник вариации результативного признака	Показатель вариации	Число степеней свободы	Насмещенная оценка дисперсии σ_0^2
Фактор — линейная функция регрессии	$\hat\sigma_\Phi^2=\hat\sigma_{\hat Y}^{2\,\text{лин}}=\frac{1}{n}\sum_{i=1}^{n}(\hat Y_i-\bar Y)^2=\hat r^2\hat\sigma_Y^2$	$2^{*)}-1=1$	$s_\Phi^2=\hat\sigma_\Phi^2 n/1$ (при выполнении гипотезы H_0 (12.59))
Остаточные факторы	$\hat\sigma_{\text{о}}^2=\hat\sigma_{\text{о}}^{2\,\text{лин}}=\frac{1}{n}\times\sum_{i=1}^{n}(Y_i-\hat Y_i)^2=(1-\hat r^2)\hat\sigma_Y^2$	$n-2$	$s_{\text{о}}^2=\frac{\hat\sigma_{\text{о}}^2 n}{n-2}$
Общая вариация	$\hat\sigma_Y^2=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar Y_i)^2=\hat\sigma_\Phi^2+\hat\sigma_{\text{о}}^2$	$n-1=1+(n-2)$	$s_Y^2=\frac{\hat\sigma_Y^2 n}{n-1}$ (при выполнении гипотезы H_0 (12.59))

*) Число 2 — это количество параметров a и b в линейной функции регрессии $M^{\text{лин}}(Y/X=x)=a+bx$.

Если гипотеза (12.59) верна, то величина

$$F=s_\Phi^2/s_{\text{о}}^2 \tag{12.60}$$

имеет F-распределение с $l=2-1$ и $k=n-2$ степенями свободы. Используя этот факт, найдем для заданного уровня значимости α правостороннюю критическую точку $x_{\text{пр},\alpha}^{\text{кр}}$ по следующей схеме:

$$\left.\begin{array}{r}\alpha\to\gamma=1-\alpha\\ l=2-1\\ n\to k=n-2\end{array}\right\}\xrightarrow{\text{п.7}} f_\gamma\to x_{\text{пр},\alpha}^{\text{кр}}=f_\gamma. \tag{12.61}$$

Если значение $F_{\text{чнс}}$ критерия (12.60) больше $x_{\text{пр},\alpha}^{\text{кр}}$, то гипотеза H_0(12.59) отвергается; в этом случае говорят, что выборочный коэффициент корреляции $\hat r_{XY}$ *статистически значим.* Если $F_{\text{чнс}}<x_{\text{пр},\alpha}^{\text{кр}}$, то гипотезу H_0(12.59) не отвергают; в этом случае говорят, что выборочный коэффициент корреляции незначим.

Задача 12.4. По результатам наблюдений, сгруппированным в табл. 12.11, построить линейное уравнение регрессии; рассчитать его среднюю квадратическую погрешность. Установить, значим или нет выборочный коэффициент корреляции.

△ В табл. 12.11 было вычислено: $\bar Y=0{,}45$; $\hat\sigma_Y^2=0{,}035$; $\bar X=4{,}97$; $\hat\sigma_X^2=3{,}499$. Уравнение регрессии (12.52) принимает вид

$$\hat Y=0{,}45+0{,}76\frac{0{,}187}{1{,}871}(x-4{,}97),$$

Таблица 12.14

k	X_k	Y_k	m_k	$\hat{Y}_k$	$(Y_k-\hat{Y}_k)^2$	$(Y_k-\hat{Y}_k)^2 m_k$
1	1*)	0,1*)	2*)	0,15	0,0025	0,005
2	1*)	0,3*)	2*)	0,15	0,0225	0,045
3	3	0,1	2	0,30	0,0400	0,080
4	3	0,3	7	0,30	0,0000	0,000
5	3	0,5	2	0,30	0,0400	0,080
6	5	0,3	10	0,45	0,0225	0,225
7	5	0,5	17	0,45	0,0025	0,042
8	5	0,7	4	0,45	0,0625	0,25
9	7	0,5	7	0,60	0,01	0,07
10	7	0,7	3	0,60	0,01	0,03
11	9	0,7	2	0,76	0,0036	0,007
12	9	0,9	2	0,76	0,0196	0,039
	Итого		$n=60$			0,873

*) Эти числа взяты из табл. 12.11: комбинация чисел $x=1$ и $y=0,1$ была зафиксирована $m=2$ раза; комбинация чисел $x=1$ и $y=0,3$ была зафиксирована $m=2$ раза и т. д.

или

$$\hat{Y}=0,072+0,076x. \tag{12.62}$$

Среднюю квадратическую погрешность этого уравнения найдем двумя способами:

1. По формуле (12.54) (см. табл. 12.14).

Отсюда погрешность уравнения (12.62)

$$\hat{\sigma}_0^{\text{лин}}=\sqrt{\frac{1}{n}\sum_{k=1}^{12}(Y_k-\hat{Y}_k)^2 m_k}=\sqrt{\frac{0,873}{60}}=0,121.$$

2. По формуле (12.57). В этом случае погрешность

$$\hat{\sigma}_0^{\text{лин}}=\sqrt{\hat{\sigma}_Y^2(1-\hat{r}_{XY}^2)}=\sqrt{0,035(1-0,76^2)}=0,121.$$

Итак, в рассматриваемой задаче $\hat{\sigma}_0^{\text{лин}}=0,121$, $\hat{\sigma}_Y^2=0,035$ и $\dfrac{\hat{\sigma}_Y}{\hat{\sigma}_0^{\text{лин}}}=\dfrac{0,187}{0,121}=1,5$. Это говорит о том, что погрешность модели $Y_i\approx\hat{Y}_i$ в 1,5 раза меньше погрешности модели $Y_i\approx\bar{Y}$.

Прямая, соответствующая уравнению регрессии (12.62), изображена на рис. 12.6. Для ее построения достаточно иметь координаты двух точек, лежащих на этой прямой. Это могут быть любые две точки с координатами $(X_k, \hat{Y}_k)$ из табл. 12.14. Отметим, что прямая обязательно должна проходить через точку с координатами $(\bar{X}, \bar{Y})$.

Допустим, что *функция регрессии линейная*. Проверим гипотезу H_0: $r_{XY}=0$. Построим дисперсионную таблицу (табл. 12.15) типа табл. 12.13. Числовое значение критерия (12.60) таково:

$$F_{\text{чис}}=1,2/0,016=75,0.$$

Пусть уровень значимости $\alpha=0,05$. По схеме (12.61) найдем точку $x^{\text{кр}}_{\text{пр},\alpha}$:

$$\left.\begin{array}{l}\alpha=0,05\rightarrow\gamma=1-\alpha=0,95\\ l=2-1=1\\ n=60\ \rightarrow k=n-2=58\end{array}\right\}\xrightarrow{\text{П.7}}4,00\rightarrow<f_\gamma<4,08\rightarrow x^{\text{кр}}_{\text{пр},\alpha}=f_\gamma.$$

Таблица 12.15

Источник вариации выпуска продукции	Показатель вариации	Число степеней свободы	Несмещенная оценка дисперсии $\hat{\sigma}_0^2$
Фактор — линейная в среднем зависимость выпуска от стоимости фондов	$\hat{\sigma}_\Phi^2=\hat{\sigma}_{\hat{Y}}^{2\,\text{лин}}=0,020$	$2-1=1$	$s_\Phi^2=0,020\cdot 60=1,2$
Остаточные факторы	$\hat{\sigma}_0^2=\hat{\sigma}_0^{2\,\text{лин}}=0,015$	$n-2=58$	$s_0^2=\dfrac{0,015\cdot 60}{58}=0,016$
Общая вариация	$\hat{\sigma}_Y^2=0,035$	$n-1=59$	$s_Y^2=\dfrac{0,035\cdot 60}{59}=0,036$

Так как $75,0>x^{\text{кр}}_{\text{пр},\alpha}$, то гипотезу $H_0: r=0$ отвергаем. Считаем, что выборочный коэффициент корреляции статистически значим. ▲

§ 12.8. Проверка гипотезы о линейности функции регрессии

Выше мы считали известным, что функция регрессии $\varphi(x)=M(Y/X=x)$ линейная. Однако, располагая результатами наблюдений, можно не знать вида функции регрессии; можно только сформулировать гипотезу о ее виде.

Допустим, что высказана гипотеза о том, что функция регрессии линейная и имеет вид (12.34):

$$H_0: M(Y/X=x)=MY+r_{XY}\frac{\sigma_Y}{\sigma_X}(x-MX). \tag{12.63}$$

Эту гипотезу можно проверить только в том случае, когда наблюдения сгруппированы в корреляционную табл. 12.9. При выполнении условий (12.25) и (12.26) гипотезу (12.63) проверяют методом дисперсионного анализа. Основным при этом является тождество

$$\hat{\sigma}_Y^2=\hat{\sigma}_{\hat{Y}}^{2\,\text{лин}}+\overline{(\bar{Y}^{(i)}-\hat{Y}_i)^2}+\overline{\hat{\sigma}_i^2}, \tag{12.64}$$

где $\hat{\sigma}_{\hat{Y}}^{2\,\text{лин}}$ — дисперсия «игреков», рассчитанных по линейному уравнению регрессии (12.49); согласно (12.56),

$$\hat{\sigma}_{\hat{Y}}^{2\,\text{лин}}=\hat{r}_{XY}^2\hat{\sigma}_Y^2;$$

$\overline{(\bar{Y}^{(i)}-\hat{Y}_i)^2}$ — средний квадрат отклонений групповых средних «игреков» (см. табл. 12.9) от «игреков», рассчитанных по линейному уравнению регрессии; согласно (12.50), он равен

$$\overline{(\bar{Y}^{(i)}-\hat{Y}_i)^2}=\hat{\sigma}_Y^2(\hat{\rho}_{Y/X}^2-\hat{r}_{XY}^2);$$

Таблица 12.16

Источник вариации признака	Показатель вариации	Число степеней свободы	Несмещенная оценка дисперсии σ_0^2
Фактор A — линейная составляющая зависимости Y от X	$\hat\sigma_A^2=\hat\sigma_{\bar Y}^{2\,\text{лин}}=\hat r_{XY}^2\hat\sigma_Y^2$	$2-1=1$	$s_A^2=\hat\sigma_A^2 n/1$ (при выполнении гипотез (12.63), (12.59))
Фактор B — нелинейная составляющая зависимости Y от X	$\hat\sigma_B^2=\overline{(\bar Y^{(i)}-\hat Y_i)^2}=$ $=\hat\sigma_Y^2(\hat\rho_{Y/X}^2-\hat r_{XY}^2)$	$\nu-2$ (ν — число групп в табл. 12.9)	$s_B^2=\hat\sigma_B^2 n/(\nu-2)$ (при выполнении гипотезы (12.63))
Остаточные факторы	$\hat\sigma_o^2=\overline{\hat\sigma_i^2}=\sigma_Y^2(1-$ $-\hat\rho_{Y/X}^2)$	$n-\nu$	$s_o^2=\hat\sigma_o^2 n/(n-\nu)$
Общая вариация	$\hat\sigma_Y^2$	$n-1=1+$ $+(\nu-2)+$ $+(n-\nu)$	$s_Y^2=\hat\sigma_Y^2 n/(n-1)$ (прн выполиении гипотез (12.63) и (12.59))

$\overline{\hat\sigma_i^2}$ — среднее групповых дисперсий «игреков» (см. табл. 12.9);

$$\overline{\hat\sigma_i^2}\underset{(12.19)}{=\!=}\hat\sigma_Y^2-\hat\sigma_{\bar Y^{(i)}}^2\underset{(12.20)}{=\!=}\hat\sigma_Y^2-\hat\rho_{Y/X}^2\hat\sigma_Y^2=\hat\sigma_Y^2(1-\hat\rho_{Y/X}^2).$$

Дисперсионная таблица будет иметь вид табл. 12.16.

Если гипотеза (12.63) верна, то величиил

$$F=s_B^2/s_o^2 \qquad (12.65)$$

имеет F-распределение с $l=\nu-2$ и $k=n-\nu$ степенями свободы. Используя этот факт, найдем для заданного уровня значимости α правостороннюю критическую точку $x_{\text{пр},\alpha}^{\text{кр}}$ по следующей схеме:

$$\left.\begin{matrix}\alpha\to\gamma=1-\alpha\\ n,\ \nu\ \begin{cases}\to l=\nu-2\\ \to k=n-\nu\end{cases}\end{matrix}\right\}\xrightarrow{\text{П.7}} f_\gamma\to x_{\text{пр},\alpha}^{\text{кр}}=f_\gamma. \qquad (12.66)$$

Если значение $F_{\text{чис}}$ критерия (12.65) больше $x_{\text{пр},\alpha}^{\text{кр}}$, то гипотеза H_0 (12.63) отвергается; если $F_{\text{чис}}<x_{\text{пр},\alpha}^{\text{кр}}$, то гипотеза H_0 (12.63) принимается.

Замечание. Построение правостороиней критической области объясиятся тем, что в дисперсионном анализе, как правило, числитель F-критерия больше знаменателя ($s_B^2>s_o^2$). Если это не так, считают, что наблюдения не подтверждают влияние фактора на признак и проверяемая гипотеза принимается.

Принимая гипотезу (12.63), мы только подтверждаем непротиворечивость линейного вида функции регрессии имеющимся результатам наблюдений, но вовсе не утверждаем, что этот вид зависимости является единственно возможным.

Задача 12.5. По результатам наблюдений, сгруппированным в табл. 12.11, выяснить, можно ли на уровне значимости $\alpha=0{,}05$ принять гнпотезу о лннейности функции регресснн.

△ В табл. 12.11 $n=60$, $\nu=5$ (это число групп по «нксу»). В задачах 12.1 и 12.3 было вычислено $\hat\sigma_Y^2=0{,}035$; $\hat\rho_{Y/X}^2=0{,}60$, $\hat r_{XY}^2=0{,}76^2=0{,}578$, а поэтому в табл. 12.16

$$s_B^2=\frac{0{,}035\,(0{,}60-0{,}578)\,60}{5-2}=0{,}0154;$$

$$s_o^2=\frac{0{,}035\,(1-0{,}60)\,60}{60-5}=0{,}0152.$$

Числовое значение критерия (12.65) таково:

$$F_{\text{чис}}=\frac{0{,}0154}{0{,}0152}=1{,}013.$$

По схеме (12.66) найдем точку $x_{\text{пр},\alpha}^{\text{кр}}$:

$$\left.\begin{matrix}\alpha=0{,}05\to\gamma=1-\alpha=0{,}95\\ n=60,\ \nu=5\ \begin{cases}\to l=\nu-2=3\\ \to k=n-\nu=55\end{cases}\end{matrix}\right\}\xrightarrow{\text{П.7}} 2{,}84>f_\gamma>2{,}76\to x_{\text{пр},\alpha}^{\text{кр}}=f_\gamma.$$

Так как $F_{\text{чис}}<x_{\text{пр},\alpha}^{\text{кр}}$, то гипотезу о линейности функции регрессии не отвергаем. ▲

§ 12.9. Пример нелинейной функции регрессии

Допустим, что изучается зависимость стоимости (Y, руб.) одного экземпляра книги от тиража (X, тыс. экз.) по следующим данным, приведенным в табл. 12.17.

В табл. 12.17 каждому наблюдаемому «иксу» соответствует одно значение «игрека». Однако следует иметь в виду, что стоимость книги зависит не только от тиража, но и от ряда других незафиксированных 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

Таблица 12.17

i	1	2	3	4	5	6	7	$8=n$
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

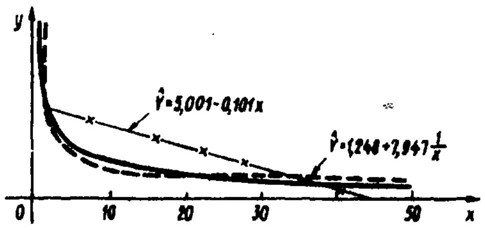


Рис. 12.7

заранее, какое именно значение примет величина Y, нельзя. Как и прежде, изучим корреляционную зависимость величины Y от X, т. е. зависимость условного математического ожидания $M(Y/X=x)$ от значения x переменной X.

Допустим, вид функции регрессии $\varphi(x)=M(Y/X=x)$ неизвестен. Ранее, чтобы составить по результатам наблюдений представление о виде этой функции, мы строили линию групповых средних (см. § 12.4). В данном примере каждому значению тиража X соответствует только одно значение стоимости Y (см. табл. 12.17), поэтому говорить о групповых средних не имеет смысла. Поступим следующим образом.

По данным табл. 12.17 построим график зависимости Y_i от X_i. Он изображен на рис. 12.7 ломаной линией. «Выровним» эту ломаную гиперболой $\hat{Y}=\hat{a}+\hat{b}\frac{1}{x}$, при этом коэффициенты $\hat{a}$ и $\hat{b}$ найдем из требования наименьших квадратов:

$$\hat{F}(\hat{a},\ \hat{b})=\overline{(Y-\hat{Y})^2}=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{a}-\hat{b}\frac{1}{X_i}\right)^2\to\min. \qquad (12.67)$$

Необходимые условия минимума функции $\hat{F}(\hat{a},\hat{b})$ образуют систему

$$\begin{cases}\dfrac{\partial\hat{F}}{\partial a}=\dfrac{1}{n}\sum\limits_{i=1}^{n}2\left(Y_i-\hat{a}-\hat{b}\dfrac{1}{X_i}\right)^2(-1)=0,\\ \dfrac{\partial\hat{F}}{\partial b}=\dfrac{1}{n}\sum\limits_{i=1}^{n}2\left(Y_i-\hat{a}-\hat{b}\dfrac{1}{X_i}\right)\left(-\dfrac{1}{X_i}\right)=0,\end{cases}$$

которая после тождественных преобразований принимает вид

$$\begin{cases}\hat{a}+\hat{b}\overline{\left(\dfrac{1}{X}\right)}=\overline{Y},\\ \hat{a}\overline{\left(\dfrac{1}{X}\right)}+\hat{b}\overline{\left(\dfrac{1}{X}\right)^2}=\overline{\left(Y\dfrac{1}{X}\right)}.\end{cases} \qquad (12.68)$$

По данным табл. 12.17 вычислим:

$$\overline{\left(\frac{1}{X}\right)}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i}=\frac{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{5}+\frac{1}{10}+\frac{1}{20}+\frac{1}{30}+\frac{1}{50}}{8}=0{,}280;$$

$$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i=3{,}473;\quad \overline{\left(\frac{1}{X}\right)^2}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{X_i}\right)^2=0{,}177;$$

$$\overline{\left(Y\cdot\frac{1}{X}\right)}=\frac{1}{n}\sum_{i=1}^{n}Y_i\frac{1}{X_i}=1{,}756.$$

Систему (12.68) запишем в виде

$$\begin{cases}\hat{a}+0{,}280\hat{b}=3{,}473,\\ 0{,}280\hat{a}+0{,}177\hat{b}=1{,}756.\end{cases}$$

Решив эту систему, найдем $\hat{b}=7{,}847$, $\hat{a}=1{,}248$. Итак,

$$\hat{Y}=1{,}248+7{,}947\frac{1}{x}. \qquad (12.69)$$

Эта гипербола изображена на рис. 12.7 пунктирной линией.

Рассчитаем погрешность

$$\hat{\sigma}_0^{\text{гип}}=+\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i-Y_i)^2}$$

уравнения (12.69). Заполним табл. 12.18.

Таблица 12.18

X_i	1	2	3	5	10	20	30	50	
Y_i	9,1	5,3	4,11	2,83	2,11	1,62	1,41	1,30	
$\hat{Y}_i=1{,}248+$ $+7{,}947\,{}^1/X_i$	9,195	5,221	3,897	2,837	2,043	1,645	1,513	1,407	
$(\hat{Y}_i-Y_i)^2$	0,0090	0,0062	0,0454	0,0001	0,0045	0,0006	0,0106	0,0114	$\sum=0{,}0878$

Итак, $\hat{\sigma}_0^{\text{гип}}=\sqrt{0{,}0878/8}=0{,}105$.

Выровним ломаную линию, изображающую на рис. 12.7 зависимость наблюдаемых «игреков» от «иксов», прямой линией $\hat{Y}=\hat{a}+\hat{b}x$, руководствуясь при этом требованием метода наименьших квадратов. С такой задачей мы уже сталкивались в § 12.5. Система уравнений для определения $\hat{a}$ и $\hat{b}$ имеет вид (12.46). Для данных, приведенных в табл. 12.17, получаем $\overline{X}=15{,}125$; $\overline{X^2}=492{,}125$; $\overline{XY}=25{,}873$. Система (12.46) принимает вид

$$\begin{cases} \hat{a}+15{,}125\hat{b}=3{,}473, \\ 15{,}125\hat{a}+492{,}375\hat{b}=25{,}873. \end{cases}$$

Решив ее, получим $\hat{b}=-0{,}101$; $\hat{a}=5{,}001$. Линейное уравнение таково:

$$\bar{Y}=5{,}001-0{,}101x. \qquad (12.70)$$

Эта прямая изображена на рис. 12.7 пунктирной линией.

Найдем погрешность уравнения (12.70). Заполним табл. 12.19.

Таблица 12.19

X_i	1	2	3	5	10	20	30	50	
Y_i	9,1	5,3	4,11	2,83	2,11	1,62	1,41	1,30	
$\hat{Y}_i=5{,}001-$ $-0{,}101X_i$	4,900	4,799	4,698	4,496	3,991	2,981	1,971	−0,049	
$(\hat{Y}_i-Y_i)^2$	17,64	0,251	0,346	2,776	3,538	1,852	0,315	1,820	$\sum=28{,}538$

Таким образом, $\hat{\sigma}_0^{\text{лин}}=\sqrt{28{,}538/8}=1{,}889$, что примерно в 18 раз больше погрешности уравнения (12.69). Поэтому из рассмотренных двух видов функций предпочтение следует отдать гиперболе.

Если допустить, что в генеральной совокупности функция регрессии имеет вид

$$M(Y/X=x)=a+b\frac{1}{x},$$

то уравнение (12.69) является выборочным уравнением регрессии, а значение

$$\hat{Y}_i=1{,}248+7{,}947\frac{1}{X_i}$$

дает выборочную оценку условного математического ожидания $M(Y/X=X_i)$.

В заключение обратим внимание на следующее. Если в гиперболической функции регрессии приравнять $\frac{1}{x}$ новой переменной z, то эта функция сводится к линейной функции $M(Y/Z=z)=a+bz$, которая была подробно рассмотрена выше.

§ 12.10. Множественная регрессия

Допустим, что изучается зависимость температуры (Z, °C) объекта от процентного содержания (X, %) компонента A в теплоносителе и температуры (Y, °C) окружающей среды по $n=11$ наблюдениям. Результаты наблюдений даны в табл. 12.20.

Таблица 12.20

i	1	2	3	4	5	6	7	8	9	10	11
X, %	1	4	9	11	3	8	5	10	2	7	6
Y, °C	8	2	−8	−10	6	−6	0	−12	4	−2	−4
Z, °C	6	8	1	0	5	3	2	−4	10	−3	5
$\hat{Z}$, °C	7,90	4,96	0,06	−2,96	4,88	1,04	3,98	0,14	7,98	0,96	4,06

В табл. 12.20 каждой наблюдаемой паре значений величин X и Y соответствует единственное значение величины Z. Однако вряд ли таким же свойством будет обладать генеральная совокупность: ведь на Z помимо X и Y влияет и ряд других случайных факторов. Поэтому обратим внимание на изучение корреляционной зависимости величины Z от X и Y, т. е. зависимости условного математического ожидания $M(Z/X=x,\ Y=y)$ (математического ожидания величины Z, вычисленного при условии $X=x$ и $Y=y$) от значений x и y.

Предположим, что функция регрессии линейная, т. е. что $M(Z/X=x,\ Y=y)=a+bx+cy$. Коэффициенты $\hat{a}$, $\hat{b}$ и $\hat{c}$ уравнения регрессии $\hat{Z}=\hat{a}+\hat{b}x+\hat{c}y$ найдем из требования метода наименьших квадратов:

$$\hat{F}(\hat{a},\hat{b},\hat{c})=\frac{1}{n}\sum_{i=1}^{n}(Z_i-\hat{Z})^2=\frac{1}{n}\sum_{i=1}^{n}(Z_i-\hat{a}-\hat{b}X_i-\hat{c}Y_i)^2\to\min.$$

Необходимые условия минимума функции $\hat{F}$ образуют систему

$$\begin{cases} \dfrac{\partial\hat{F}}{\partial\hat{a}}=\dfrac{1}{n}\sum\limits_{i=1}^{n}2(Z_i-\hat{a}-\hat{b}X_i-\hat{c}Y_i)(-1)=0, \\ \dfrac{\partial\hat{F}}{\partial\hat{b}}=\dfrac{1}{n}\sum\limits_{i=1}^{n}2(Z_i-\hat{a}-\hat{b}X_i-\hat{c}\hat{Y}_i)(-X_i)=0, \\ \dfrac{\partial\hat{F}}{\partial\hat{c}}=\dfrac{1}{n}\sum\limits_{i=1}^{n}2(Z_i-\hat{a}-\hat{b}X_i-\hat{c}\hat{Y}_i)(-Y_i)=0, \end{cases}$$

которая в результате тождественных преобразований принимает вид

$$\begin{cases} \hat{a}+\hat{b}\overline{X}+\hat{c}\overline{Y}=\overline{Z}, \\ \hat{a}\overline{X}+\hat{b}\overline{X^2}+\hat{c}\overline{YX}=\overline{ZX}, \\ \hat{a}\overline{Y}+\hat{b}\overline{XY}+\hat{c}\overline{Y^2}=\overline{ZY}. \end{cases} \tag{12.71}$$

По данным табл. 12.20 вычисляем: $\overline{Z}=3$, $\overline{X}=6$, $\overline{Y}=-2$, $\overline{X^2}=46$, $\overline{Y^2}=44$, $\overline{XY}=-31{,}45$, $\overline{ZX}=7{,}73$, $\overline{ZY}=12{,}9$. Подставив эти значения в систему (12.71), найдем $\hat{a}=14{,}186$, $\hat{b}=-2{,}041$, $\hat{c}=-0{,}53$. Итак,

$$\hat{Z}=14{,}186-2{,}041x-0{,}53y. \tag{12.72}$$

Уравнение регрессии, содержащее более одной независимой переменной, называется *множественным.*

Значения величины $\hat{Z}$, вычисленные для наблюдаемых значений величин X и Y, приведены в последней строке табл. 12.20.

Найдем погрешность уравнения (12.72):

$$\hat{\sigma}_o^{\text{лин}}=+\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z_i-\hat{Z}_i)^2}=\sqrt{6{,}19}=2{,}49,$$

а также выборочное среднее квадратическое отклонение величины Z:

$$\hat{\sigma}_Z=+\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z_i-\overline{Z})^2}=\sqrt{17{,}27}=4{,}16.$$

Таким образом, погрешность модели $Z_i\approx\hat{Z}_i$ в 1,7 раза меньше погрешности модели $Z_i\approx\overline{Z}$.

Степень линейной корреляционной зависимости величины Z от X и Y по выборочным данным измеряется *выборочным множественным коэффициентом корреляции* $\hat{R}_{Z/X,Y}$, квадрат которого вычисляется по формуле, аналогичной (12.58),

$$\hat{R}^2_{Z/X,Y}=\frac{\hat{\sigma}_Z^{2\,\text{лин}}}{\hat{\sigma}_Z^2}=\frac{\frac{1}{n}\sum\limits_{i=1}^{n}(\hat{Z}_i-\overline{Z})^2}{\frac{1}{n}\sum\limits_{i=1}^{n}(Z_i-\overline{Z})^2}. \tag{12.73}$$

Коэффициент $\hat{R}^2_{Z/X,Y}$ показывает, какую долю от дисперсии $\hat{\sigma}_Z^2$ результативного признака Z составляет дисперсия $\hat{\sigma}_Z^{2\,\text{лин}}$ «зетов», вычисленных по линейному уравнению регрессии, или, иначе говоря, какая доля дисперсии $\hat{\sigma}_Z^2$ объясняется линейной в среднем зависимостью величины Z от X и Y. Обратим внимание, что $\hat{R}^2_{Z/X,Y}\leqslant 1$ ($\hat{R}^2_{Z/X,Y}=1$ только в случае линейной функциональной зависимости наблюдаемых «зетов» от наблюдаемых «иксов» и «игреков»).

По данным табл. 12.20, $\hat{\sigma}_Z^{2\,\text{лин}}=\frac{1}{n}\sum\limits_{i=1}^{n}(\hat{Z}_i-\overline{Z})^2=10{,}78$ (заметим, что с точностью до ошибок округления должно выполняться равенство $\hat{\sigma}_Z^2=\hat{\sigma}_Z^2+\hat{\sigma}_o^{2\,\text{лин}}$) и $\hat{R}^2_{Z/X,Y}=\frac{10{,}78}{17{,}27}=0{,}62$. Таким образом, судя по выборочным данным, 62% дисперсии величины Z (температуры объекта) объясняется ее линейной в среднем зависимостью от X (процентного содержания компонента A в теплоносителе) и Y (температуры окружающей среды).

Можно ли считать, что в генеральной совокупности множественный коэффициент корреляции $R_{Z/X,Y}$ отличен от нуля? Для проверки гипотезы $H_0: R_{Z/X,Y}=0$ используется критерий

$$F=\frac{\hat{\sigma}_Z^{2\,\text{лин}}\cdot n/2}{\hat{\sigma}_o^{2\,\text{лин}}n/(n-3)},$$

который при условии справедливости гипотезы H_0 имеет F-распределение с $l=2$ и $k=n-3$ степенями свободы. Обычно числитель F-критерия больше знаменателя, поэтому для проверки гипотезы H_0 строим правостороннюю критическую область. Критическую точку $x^{\text{кр}}_{\text{пр},\alpha}$ находим по схеме (12.61), учитывая при этом, что $l=2$, а $k=n-3$.

Проверим гипотезу $H_0: R_{Z/X,Y}=0$ по данным табл. 12.20. Числовое значение $F_{\text{чис}}=\frac{10{,}78\cdot 11/2}{6{,}19\cdot 11/8}=6{,}97$. Пусть $\alpha=0{,}05$. Найдем критическую точку:

$$\left.\begin{array}{l}\alpha=0{,}05\to\gamma=1-\alpha=0{,}95\\ n=11\to\begin{array}{l}l=2\\k=n-3=8\end{array}\end{array}\right\}\xrightarrow{\text{П.7}} f_\gamma=4{,}46\to x^{\text{кр}}_{\text{пр},\alpha}=4{,}46.$$

Так как $F_{\text{чис}}=6{,}97>x^{\text{кр}}_{\text{пр},\alpha}$, то гипотезу H_0 отвергаем, т. е. с вероятностью ошибки, равной 0,05, считаем, что в генеральной совокупности изменчивость «зетов» связана с линейным влиянием на них «икса» и «игрека». Судя по выборке, степень этого влияния составляет 62%, на долю неконтролируемых факторов приходится 38%.

В заключение отметим, что формуле (12.73) тождественна следующая формула:

$$\hat{R}^2_{Z/X,Y}=\frac{\hat{r}^2_{ZX}+\hat{r}^2_{ZY}-2\hat{r}_{ZX}\hat{r}_{ZY}\hat{r}_{XY}}{1-\hat{r}^2_{XY}},$$

где $\hat{r}$ — выборочный парный коэффициент корреляции. Отметим также, что при изучении множественных зависимостей наряду с множественным коэффициентом корреляции используют *частные коэффициенты корреляции. Выборочный частный коэффициент корреляции*

$$\hat{r}_{ZX/Y}=\frac{\hat{r}_{ZX}-\hat{r}_{ZY}\hat{r}_{XY}}{\sqrt{(1-\hat{r}^2_{ZY})(1-\hat{r}^2_{XY})}} \qquad (12.74)$$

показывает степень линейной зависимости между наблюдаемыми значениями величин Z и X при условии, что линейное в среднем влияние на них наблюдаемых значений величины Y устранено (напомним, что $\hat{r}_{ZX}$ показывает степень линейной зависимости между наблюдаемыми значениями величин Z и X без какого-либо дополнительного условия).

Замечание. Устранить линейное в среднем влияние наблюдаемых «игреков» на «зеты» и «иксы» — это значит, *во-первых*, построить выборочное линейное уравнение регрессии Z на $Y: \hat{Z}=a_1+b_1y$ — и от наблюдаемых «зетов» отнять соответствующие значения, рассчитанные по этому уравнению, т. е. перейти к ряду чисел

$$Z_1-\hat{Z}_1,\ Z_2-\hat{Z}_2,\ \ldots,\ Z_n-\hat{Z}_n; \qquad (12.75)$$

во-вторых, построить линейное уравнение регрессии X на Y: $\hat{X}=a_2+b_2y$ и от наблюдаемых «иксов» отнять соответствующие значения, рассчитанные по этому уравнению, т.е. перейти к ряду чисел

$$X_1-\hat{X}_1,\ X_2-\hat{X}_2,\ \ldots,\ X_n-\hat{X}_n. \qquad (12.76)$$

Частный коэффициент корреляции $\hat{r}_{ZX/Y}$ равен парному коэффициенту $\hat{r}_{Z-\hat{Z},\,X-\hat{X}}$ между рядами (12.75) и (12.76), т. е.

$$\hat{r}_{ZX/Y}=\hat{r}_{Z-\hat{Z},\,X-\hat{X}}.$$

Частный коэффициент корреляции $\hat{r}_{ZY/X}$ между наблюдаемыми значениями величин Z и Y при условии, что влияние на них величины X устранено, вычисляется по формуле, аналогичной формуле (12.74):

$$\hat{r}_{ZY/X}=\frac{\hat{r}_{ZY}-\hat{r}_{ZX}\hat{r}_{YX}}{\sqrt{(1-\hat{r}^2_{ZX})(1-\hat{r}^2_{YX})}}.$$

ГЛАВА 13. МЕТОД СТАТИСТИЧЕСКИХ ИСПЫТАНИЙ

§ 13.1. Общая идея метода статистических испытаний

Во всех рассмотренных выше вероятностных задачах (см. гл. 2—6) удавалось установить формульную зависимость конечного результата от исходных данных, т. е. получить аналитическое решение задачи. Если этого сделать нельзя, то используют метод статистических испытаний.

Основная идея метода состоит в следующем: вместо аналитического решения задачи либо проводят эксперименты

испытания, непосредственно рассматриваемые в задаче, либо эти испытания заменяют другими, имеющими с исходными одинаковую вероятностную структуру, или, иначе говоря, рассматриваемые в задаче случайные явления имитируют, моделируют другими случайными явлениями*).

Определенные по результатам достаточно большого числа испытаний характеристики случайных явлений (относительные частоты, средние арифметические) используют в качестве приближенного решения задачи (в качестве оценок вероятностей, математических ожиданий). Допустимость этого приближения основывается на законе больших чисел.

Метод статистических испытаний применяют для решения не только тех задач, в которых в явном виде имеются случайные явления, но также и для решения многих математических задач, не содержащих таких явлений. В этом случае искусственно подбирают такое случайное явление, характеристики которого связаны с результатом решения исходной задачи. Для определения числовых значений этих характеристик используется метод статистических испытаний.

Так как достаточно высокая точность решения при использовании метода статистических испытаний гарантируется, как правило, только при проведении большого числа испытаний, этот метод практически можно реализовать только на быстродействующих ЭВМ. По этой причине метод статистических испытаний называют иногда «машинным».

Для иллюстрации метода статистических испытаний рассмотрим несколько задач.

Задача 1. Система контроля качества продукции состоит из трех приборов. Вероятность безотказной работы каждого из них в течение времени T равна $\frac{5}{6}$. Приборы выходят из строя независимо друг от друга. При отказе хотя бы одного прибора вся система перестает работать. Найти вероятность $p_{\text{отк}}$ того, что система откажет за время T.

Решим задачу аналитически и методом статистических испытаний.

$\triangle$ *Аналитическое решение*. Событие A (выход из строя хотя бы одного из трех приборов за время T) и событие $\bar{A}$ (ни один из трех приборов не выйдет из строя за время T) — противоположные. Вероятность $P(\bar{A})=(5/6)^3$. Искомая вероятность

$$p_{\text{отк}}=P(A)=1-P(\bar{A})=1-(5/6)^3=0{,}42.$$

*) Одним из возможных способов имитации случайных явлений является рулетка. Игрой в рулетку знаменит город Монте-Карло. Именно этим объясняется другое часто встречающееся название метода статистических испытаний метод Монте-Карло.

Теперь решим задачу *методом статистических испытаний*. Напомним, что при использовании данного метода возможны два подхода: либо непосредственно проводят эксперименты, либо имитируют их другими экспериментами, имеющими с исходными одинаковую вероятностную структуру. В условиях данной задачи «натуральный» эксперимент — наблюдение за работой системы в течение времени T. Многократное повторение этого эксперимента может оказаться трудноосуществимым или просто невозможным. Заменим этот эксперимент другим.

Для определения того, выйдет или не выйдет из строя за время T отдельный прибор, будем подбрасывать игральную кость. Если выпадет одно очко, то будем считать, что прибор вышел из строя; если два, три, ..., шесть очков, то будем считать, что прибор работал безотказно. Вероятность того, что выпадет одно очко, так же как и вероятность выхода прибора из строя, равна $1/6$, а вероятность того, что выпадет любое другое число очков, как и вероятность безотказной работы прибора, равна $5/6$.

Чтобы определить, откажет или нет вся система за время T, будем подбрасывать три игральные кости (или одну кость три раза). Если хотя бы на одной из трех костей выпадет одно очко, то это будет означать, что система отказала.

Повторим испытание, состоящее в подбрасывании трех игральных костей, много раз подряд и найдем отношение числа m «отказов» системы к общему числу n проведенных испытаний. Вероятность отказа

$$p_{\text{отк}} \approx m/n. \ \blacktriangle$$

Теперь рассмотрим математическую задачу, не являющуюся по своей природе вероятностной, эффективным методом решения которой является метод статистических испытаний. Это задача вычисления определенного интеграла.

Задача 2 (вычисление определенного интеграла). Пусть требуется вычислить интеграл $\int_0^1 f(x)\,dx$. Предположим, что $0 \leqslant f(x) \leqslant 1$ при $0 \leqslant x \leqslant 1$.

△ Значение этого интеграла равно площади области G, ограниченной кривой $y=f(x)$, осью Ox и прямыми $x=0$ и $x=1$ (рис. 13.1).

Будем многократно бросать на единичный квадрат случайным образом точку (маленький шарик). Подсчитаем отношение числа m бросаний, при которых точка попадает в область G, к общему числу n бросаний. Это отношение является оценкой вероятности $p_{\text{п}}$ попадания точки в область G:

$$p_{\text{п}} \approx m/n.$$

С другой стороны, так как точка бросается в единичный квадрат наудачу, то можно считать, что вероятность попадания ее в область G пропорциональна площади области и не зависит от расположения области внутри единичного квадрата. Тогда

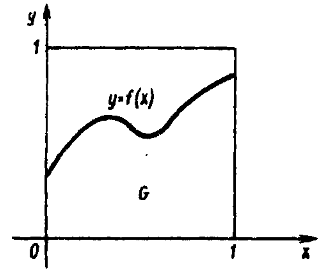


Рис. 13.1

$$p_{\text{п}} = \frac{S_G}{S_{\text{кв}}} = \frac{\int_0^1 f(x)\,dx}{1} = \int_0^1 f(x)\,dx.$$

Таким образом,

$$\int_0^1 f(x)\,dx \approx m/n. \ \blacktriangle$$

В рассмотренных задачах непосредственной целью проведения испытаний было определение вероятности p наступления того или иного события. При этом использовалось приближенное равенство

$$p \approx m/n, \tag{13.1}$$

где m — число наступлений события в n испытаниях.

Будем предполагать, что это испытания Бернулли, т. е. что они независимы и проводятся в одинаковых условиях. Напомним, что при выполнении этих предположений частость m/n при достаточно большом числе испытаний является хорошим приближением вероятности p (см. § 9.3).

Рассмотрим задачу, в которой метод статистических испытаний используется для определения математического ожидания.

Задача 3. Некоторое тело с равной вероятностью перемещается на единичное расстояние либо вправо, либо влево, либо вверх, либо вниз. Требуется оценить математическое ожидание MX расстояния тела от начального положения после k перемещений (расстояние от начального положения — величина случайная в силу случайности перемещений; обозначим его X):

△ Предположим, что в начальном положении тело имеет координаты $x=0$ и $y=0$. Будем одно перемещение имитировать двукратным подбрасыванием монеты. Условимся, что появление двух «гербов» означает движение тела вправо, что, в свою очередь, приводит к увеличению ее абсциссы x на единицу. Появление двух решек означает движение влево и, следовательно, абсциссу x частицы надо уменьшить на единицу. Появление при первом подбрасывании монеты «герба», а при втором — решки означает движение тела вверх, что приводит к увеличению его ординаты y на единицу. При появлении же сначала решки, а затем «герба» тело «движется» вниз и его ордината y уменьшается на единицу. Вероятности исходов, возможных при двукратном подбрасывании монеты, так же как и вероятности движения тела по любому из четырех направлений, равны $1/4$.

Имитировать k перемещений будем подбрасыванием монеты $2k$ раз. При этом после каждых двух подбрасываний либо абсциссу x пересчитаем, либо ординату y тела. Смещение тела относительно начального положения после k перемещений равно $X=\sqrt{x^2+y^2}$.

Случайное испытание, состоящее в подбрасывании монеты $2k$ раз, повторим достаточно большое число n раз. Результатом i-го испытания ($i=1, 2, ..., n$) является «смещение» тела, равное X_i. Вычислим среднее арифметическое этих смещений и примем его за приближенное значение математического ожидания MX, т. е.

$$MX\approx (X_1+X_2+...+X_n)/n. \qquad (13.2)$$

Напомним, что при соблюдении достаточно общих требований (испытания должны быть независимыми и проводиться в одинаковых условиях) средняя $\bar{X}=(X_1+X_2+...+X_n)/n$ при достаточно большом числе испытаний является хорошим приближением математического ожидания MX (см. § 9.2). ▲

Выясним, сколько следует провести испытаний n, чтобы с достаточно высокой вероятностью (не меньшей γ) быть уверенным в том, что погрешность ε приближенного равенства (13.1) или (13.2) будет незначительной.

Определение числа испытаний при оценивании математического ожидания. Напомним, что для средней $\bar{X}$ имеет место неравенство Чебышева (7.9):

$$P(|\bar{X}-MX|\leqslant\varepsilon)>1-\frac{DX}{n\varepsilon^2}.$$

Дисперсия DX, как правило, заранее не известна. Однако всегда, хотя бы ориентировочно, можно указать ее верхнюю границу C. Тогда будет иметь место следующая цепочка неравенств:

$$P(|\bar{X}-MX|\leqslant\varepsilon)>1-\frac{DX}{n\varepsilon^2}\geqslant 1-\frac{C}{n\varepsilon^2}.$$

Зададимся малым числом ε и вероятностью γ, с которой мы хотим гарантировать погрешность ε. Кроме того, положим $1-\frac{C}{n\varepsilon^2}=\gamma$. Решение этого равенства $n_0=\frac{C}{(1-\gamma)\varepsilon^2}$. Отсюда при числе испытаний

$$n\geqslant\frac{C}{(1-\gamma)\varepsilon^2},\quad DX\leqslant C, \qquad (13.3)$$

с вероятностью, не меньшей γ, можно быть уверенным в том, что абсолютная погрешность приближенного равенства $MX\approx\bar{X}$ не превысит ε.

Если предположить, что $\bar{X}=N\left(MX, \sqrt{\frac{DX}{n}}\right)$, т. е. что $\bar{X}$ имеет нормальный закон распределения с математическим ожиданием, равным MX, и дисперсией, равной DX/n (основание для этого предположения при большом n дает теорема Ляпунова), то, воспользовавшись (7.14), получим

$$P(|\bar{X}-MX|<\varepsilon)=\Phi\left(\frac{\varepsilon}{\sqrt{DX/n}}\right)\geqslant\Phi\left(\frac{\varepsilon}{\sqrt{C/n}}\right),\quad DX\leqslant C.$$

Положим, что $\Phi\left(\frac{\varepsilon}{\sqrt{C/n}}\right)=\gamma$. Решение этого равенства $n_0=u_\gamma^2\,C/\varepsilon^2$, где u_γ найдено по табл. П.2 при заданной вероятности γ. Итак, в этом случае при числе испытаний

$$n\geqslant u_\gamma^2\,C/\varepsilon^2 \qquad (13.4)$$

с вероятностью, не меньшей γ, можно быть уверенным в том, что абсолютная погрешность приближенного равенства $MX\approx\bar{X}$ не превысит ε.

Определение числа испытаний при оценивании вероятности p. Для частости m/n имеет место неравенство Чебышева (7.10):

$$P\left(\left|\frac{m}{n}-p\right|\leqslant\varepsilon\right)>1-\frac{p(1-p)}{n\varepsilon^2}.$$

Нетрудно убедиться в том, что $p(1-p)\leqslant\frac{1}{4}$.

Замечание. Найдем максимум функции $f(p)=p(1-p)$; $\frac{df}{dp}=1-2p=0\to p=1/2$. Так как $\frac{d^2f}{dp^2}<0$, то $p=1/2$ — это точка максимума и $f_{max}=f(p=1/2)=1/4$. Итак, $p(1-p)\leqslant 1/4$.

Поэтому справедлива следующая цепочка неравенств:

$$P\left(\left|\frac{m}{n}-p\right|\leqslant\varepsilon\right)>1-\frac{p(1-p)}{n\varepsilon^2}\geqslant 1-\frac{1}{4n\varepsilon^2}.$$

Зададимся малой погрешностью ε и вероятностью γ, гарантирующей эту погрешность. Положим $1-\frac{1}{4n\varepsilon^2}=\gamma$. Решение этого равенства $n_0=\frac{1}{4(1-\gamma)\varepsilon^2}$. Итак, при числе испытаний

$$n\geqslant\frac{1}{4(1-\gamma)\varepsilon^2} \qquad (13.5)$$

можно с вероятностью γ быть уверенным в том, что абсолютная погрешность приближенного равенства $p \approx m/n$ не превысит ε.

Если предположить, что частость $\frac{m}{n} = N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, т. е. что m/n имеет нормальный закон распределения с математическим ожиданием p и дисперсией $p(1-p)/n$ (основание для этого предположения при большом числе n дает локальная и интегральная теорема Муавра—Лапласа), то, учитывая (6.31), будем иметь

$$P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=\Phi\left(\frac{\varepsilon}{\sqrt{p(1-p)/n}}\right) \geqslant \Phi\left(\frac{\varepsilon}{\sqrt{1/(4n)}}\right).$$

Приравняв $\Phi\left(\frac{\varepsilon}{\sqrt{1/(4n)}}\right)=\gamma$, получим решение $n_0 = u_\gamma^2/(4\varepsilon^2)$. Итак, при числе испытаний

$$n \geqslant \frac{u_\gamma^2}{4\varepsilon^2} \tag{13.6}$$

с вероятностью, не меньшей γ, можно быть уверенным, что абсолютная погрешность приближенного равенства $p \approx m/n$ не превысит ε.

Сравним результаты вычислений числа испытаний по формулам (13.5) и (13.6). Пусть $\varepsilon=0{,}01$, $\gamma=0{,}99$. Тогда, согласно (13.5),

$$n \geqslant \frac{1}{4(1-0{,}99)0{,}01^2}=250\ 000,$$

а по формуле (13.6)

$$n \geqslant \frac{2{,}58^2}{4 \cdot 0{,}01^2}=16\ 641,$$

где 2,58 — число u_γ, найденное по табл. П.2 при $\gamma=0{,}99$. Число испытаний значительно уменьшилось, но осталось достаточно большим.

§ 13.2. Моделирование случайной величины R с равномерным распределением на отрезке [0, 1]

Случайная величина с равномерным распределением на отрезке [0, 1]. Оказывается, что для имитации на ЭВМ случайных явлений самой различной природы достаточно получить на ЭВМ последовательность значений случайной величины, равномерно распределенной на отрезке [0, 1]. Процесс получения значений случайной величины называют ее *моделированием*.

Напомним, что непрерывная случайная величина X имеет равномерное распределение на отрезке $[a, b]$, если ее функция плотности

$$f_X(x)=\begin{cases}\frac{1}{b-a} \text{ при } a \leqslant x \leqslant b,\\ 0 \text{ при } x<a \text{ и } x>b.\end{cases}$$

Функция распределения случайной величины X имеет вид

$$F_X(x)=\begin{cases}0 \text{ при } x \leqslant a,\\ \frac{x-a}{b-a} \text{ при } 0 \leqslant x \leqslant b,\\ 1 \text{ при } x \geqslant b.\end{cases}$$

Математическое ожидание и дисперсия случайной величины X соответственно равны

$$MX=\frac{a+b}{2}, \quad DX=\frac{(b-a)^2}{12}.$$

Случайную величину с равномерным распределением на отрезке [0, 1] будем обозначать R. Имеем

$$f_R(x)=\begin{cases}1 \text{ при } 0 \leqslant x \leqslant 1,\\ 0 \text{ при } x<0 \text{ и } x>1;\end{cases}$$

$$F_R(x)=\begin{cases}0 \text{ при } x \leqslant 0,\\ x \text{ при } 0 \leqslant x \leqslant 1,\\ 1 \text{ при } x \geqslant 1.\end{cases}$$

Математическое ожидание $MR=\frac{1}{2}$, дисперсия $DR=\frac{1}{12}$. Графики функции плотности $f_R(x)$ и функции распределения $F_R(x)$ изображены на рис. 13.2.

Случайная величина R обладает следующим свойством: вероятность попадания ее в любой подотрезок отрезка [0, 1] равна длине этого подотрезка. Действительно, если $[r_1, r_2]$ — некоторый подотрезок отрезка [0, 1], то вероятность

$$P(r_1 \leqslant R < r_2)=F_R(r_2)-F_R(r_1)=r_2-r_1. \tag{13.7}$$

Моделирование случайной величины R с помощью бросания монеты. Образуем бесконечную двоичную дробь $Y=0, X_1X_2...X_k...$, используя для определения значения X_i, $i=1, 2,$

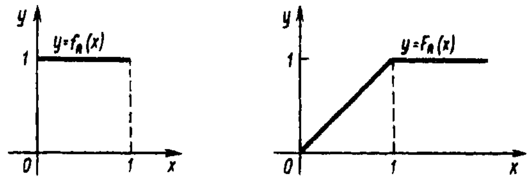


Рис. 13.2

..., следующий физический эксперимент: бросаем монету и, если выпадет «герб», то полагаем $X_i=1$, если решка, то $X_i=0$.

Таким образом, $X_1, X_2, ..., X_k, ...$ — последовательность значений случайной величины X, которая имеет следующий ряд распределения:

$$\begin{array}{c|cc} X & 0 & 1 \\ \hline P & 0{,}5 & 0{,}5 \end{array} \qquad (13.8)$$

Дробь Y представим в таком виде:

$$Y=X_1 2^{-1}+X_2 2^{-2}+...+X_k 2^{-k}+... . \qquad (13.9)$$

Очевидно, что числовые значения этой дроби лежат в отрезке [0, 1], а сама дробь — непрерывная случайная величина, и, более того, можно доказать, что она имеет равномерное распределение, т. е. $Y=R$.

Числовое значение дроби Y формировалось с помощью бросания монеты бесконечное число раз. Реализовать этот процесс в принципе нельзя, да и не следует стремиться к его реализации, поскольку все ЭВМ работают с числами, содержащими конечное число цифр.

В бесконечной двоичной дроби Y, определяемой соотношениями (13.9) и (13.8), сохраним только старшие k разрядов, т. е. перейдем к конечной двоичной дроби

$$\hat{Y}=X_1 2^{-1}+X_2 2^{-2}+...+X_k 2^{-k}. \qquad (13.10)$$

Эта дробь по-прежнему случайна, но ее распределение уже будет дискретным. Однако при достаточно большом числе k разрядов распределение двоичной дроби $\hat{Y}$, определяемой соотношениями (13.10) и (13.8), близко к распределению дроби Y, определяемой соотношениями (13.9) и (13.8), т. е. близко к равномерному распределению на отрезке [0, 1]. Поэтому в дальнейшем случайные величины $\hat{Y}$ и R различать не будем.

Физические генераторы. *Физическим генератором* (датчиком случайного двоичного разряда X_i) называется специальное приспособление к ЭВМ, в котором результаты некоторого случайного физического процесса преобразуются в последовательность значений случайной величины X с распределением (13.8). Это, например, генераторы, использующие случайный процесс радиоактивного распада.

Пусть в ЭВМ имеется какой-либо источник излучения радиоактивных частиц. Счетчик подсчитывает количество радиоактивных частиц за некоторое время Δt. Если число частиц четное, то в разряд посылается единица, если нечетное, то нуль. При параллельной работе k генераторов будет получено значение k-разрядной двоичной дроби (13.10). Время Δt выбирается таким, чтобы вероятность получения в разряде единицы, так же как и вероятность получения нуля, была равна 0,5.

Таблицы случайных чисел. Чтобы в процессе решения задач на ЭВМ методом статистических испытаний не тратить время на получение значений случайной величины с равномерным распределением на отрезке [0, 1], можно, вообще говоря, заранее составить таблицы этих значений, используя, например, физические генераторы. Существуют готовые таблицы случайных чисел: фрагмент их приведен в табл. П.8. Если ЭВМ работает с k-разрядными десятичными числами, то, разделив очередную совокупность k цифр, взятую из таблицы, на 10^k, получим достаточно хорошее приближение к значению величины R. Однако при расчетах на ЭВМ таблицы случайных чисел, как правило, не используют. Дело в том, что хранение таблиц во внутренней памяти ЭВМ обычно невозможно вследствие ее загруженности информацией, относящейся непосредственно к решаемой задаче. Хранение же таблицы во внешней памяти и постоянное обращение к ней замедляют счет.

Псевдослучайные последовательности чисел. Возможен и следующий подход к моделированию случайной величины R, имеющей равномерное распределение на отрезке [0, 1]. Вместо последовательности значений случайной величины R получают с помощью некоторой рекуррентной формулы последовательность чисел, обладающих статистическими свойствами, близкими к свойствам равномерного распределения на отрезке [0, 1]. Эту последовательность чисел называют *псевдослучайной*.

Один из первых методов получения псевдослучайной последовательности чисел был предложен Дж. Нейманом. Он называется *методом середины квадратов*. Поясним его на примере.

Возьмем некоторое число r_0. Пусть $r_0=0{,}9876$. Возведем его в квадрат ($r_0^2=0{,}97535376$). Выберем четыре средние цифры этого числа и положим $r_1=0{,}5353$. Затем возводим r_1 в квадрат ($r_1^2=0{,}28654609$) и снова выбираем четыре средние цифры. Получаем $r_2=0{,}6546$. Далее находим $r_2^2=0{,}42850116$, $r_3=0{,}8501$, $r_3^2=0{,}72267001$, $r_4=0{,}2670$ и т. д. Последовательность чисел r_1, r_2, ... принимают за последовательность значений случайной величины R, имеющей равномерное распределение на отрезке [0, 1].

Существуют и другие алгоритмы образования псевдослучайной последовательности.

Метод псевдослучайных последовательностей прост и экономичен: он не требует разработки специальных приспособлений к ЭВМ, и при использовании его на получение каждого числа затрачивается всего несколько простых операций. Однако этот метод имеет ряд существенных недостатков, главным из которых является трудность теоретической оценки статистических свойств псевдослучайной последовательности. Обычно для оценки степени приближения этой последовательности к последовательности случайных чисел с равномерным распределением используют систему статистических критериев.

При этом исследуют лишь сравнительно «небольшие участки» последовательности псевдослучайных чисел. В практических же задачах, решаемых методом статистических испытаний, количество используемых случайных чисел чрезвычайно велико.

Помимо этого, числа, входящие в вырабатываемую программным способом псевдослучайную последовательность, зависимы между собой, а сама последовательность является периодической, так как в ЭВМ может быть представлено только конечное число различных чисел и повторное появление какого-либо числа (а значит, и всех последующих за ним чисел) неизбежно. Поэтому с практической точки зрения даже очень длинные последовательности нельзя считать случайными.

§ 13.3. Имитация случайных испытаний на ЭВМ

Нами рассмотрены способы моделирования на ЭВМ случайной величины R, равномерно распределенной на отрезке $[0, 1]$. Выясним, как, зная последовательность значений величины R, можно имитировать на ЭВМ случайные процессы, используемые в решении задач из § 13.1. Задачи будем рассматривать в той же последовательности, что и в § 13.1.

Задача 1. В задаче любой из трех приборов, входящих в систему контроля, мог выйти из строя за время T (вероятность этого события равна $1/6$) и мог в течение времени T безотказно работать (вероятность этого события равна $5/6$). Для определения того, выйдет или нет прибор из строя, использовался случайный эксперимент — подбрасывалась игральная кость. Выпадание одного очка считалось «выходом из строя»; любое другое число очков трактовалось как «безотказная работа».

Для случайной величины R с равномерным распределением на отрезке $[0, 1]$ справедливы (см. (13.7)) следующие равенства:

$$P(0 \leqslant R \leqslant 1/6) = 1/6 - 0 = 1/6, \quad P(1/6 \leqslant R < 1) = 1 - 1/6 = 5/6.$$

Поэтому на ЭВМ «выход из строя» или «безотказность работы» будем определять таким образом:

1) получают значение r_i случайной величины R;

2) если $r_i < 1/6$, то фиксируют «выход из строя», в противном случае — «безотказность работы».

Чтобы определить, откажет или нет вся система, нужно получить три значения случайной величины R, и если хотя бы одно из них меньше $1/6$, то считаем, что система «отказала»; в противном случае — система «работает безотказно».

Задача 2 (вычисление определенного интеграла). В задаче вычислялся интеграл $\int_0^1 f(x)\,dx$, где $0 \leqslant f(x) \leqslant 1$ при $0 \leqslant x \leqslant 1$. Для этого многократно повторялся случайный эксперимент: в единичный квадрат наудачу бросали точку. Бросание наудачу означает, что координаты точки — независимые случайные величины, равномерно распределенные на отрезке $[0, 1]$. Алгоритм вычисления интеграла на ЭВМ следующий:

1) получают значение r_i случайной величины R и принимают его за абсциссу точки, $x_i = r_i$;

2) вычисляют значение функции $f(x)$ в точке $x = x_i$, $f(x_i)$;

3) получают следующее значение r_{i+1}, которое принимают за ординату точки с абсциссой x_i, т. е. $y_i = r_{i+1}$ (определены координаты точки, брошенной случайным образом, в единичный квадрат);

4) к содержимому n счетчика выработанных точек прибавляют единицу;

5) если $y_i \leqslant f(x_i)$, то точка (x_i, y_i) попала внутрь области G (см. рис. 13.1), в этом случае полагают, что $\omega = 1$. Если $y_i > f(x_i)$, то полагают, что $\omega = 0$;

6) к содержимому «m» счетчика точек, попавших в область G, прибавляют ω;

7) если $n < n_0$, где n_0 — заданное число испытаний (бросаний точки), то переходят к п. 1) алгоритма. Если $n = n_0$, то вычисляют значение интеграла

$$\int_0^1 f(x)\,dx \approx m/n.$$

Задача 3. В этой задаче случайное перемещение тела, результатом которого могло быть четыре равновероятных исхода, имитировалось подбрасыванием двух монет. Если выпадало два «герба», то тело «перемещалось» вправо; две решки — влево; «герб и цифра» — «двигалась» вверх; «решка и герб» — вниз.

Для случайной величины R, имеющей равномерное распределение на отрезке $[0, 1]$, справедливо соотношение (см. (13.7)). Поэтому

$$P(0 < R < 0{,}25) = P(0{,}25 < R < 0{,}5) = P(0{,}5 < R < 0{,}75) =$$
$$= P(0{,}75 < R < 1) = 1/4.$$

На ЭВМ результат перемещения определяем так:

1) получаем значение r_i случайной величины R;

2) если $r_i < 0{,}25$, то фиксируем перемещение вправо, если $0{,}25 < r_i < 0{,}5$, фиксируем перемещение влево; если $0{,}5 < r_i < 0{,}75$, то тело смещается вверх, при $r_i > 0{,}75$ — вниз.

Для имитации k перемещений нужно взять k значений случайной величины R.

При использовании метода статистических испытаний для решения различных задач необходимо уметь моделировать испытания с различным числом исходов, при этом испытания могут быть как независимыми, так и зависимыми. Необходимо также уметь моделировать случайные величины с самыми разнообразными законами распределения. Эти вопросы рассматриваются в следующих параграфах.

§ 13.4. Моделирование последовательности случайных испытаний

Последовательность независимых испытаний. Пусть проводится последовательность k независимых испытаний. В результате каждого испытания может произойти одно из двух противоположных событий A и B. Известна вероятность наступления события A, $P(A)=p$. Так как события A и B противоположные, то $P(B)=1-p$.

Моделирование последовательности испытаний осуществляется таким образом.

Получаем последовательность значений $r_1, r_2, ..., r_k$ случайной величины R, имеющей равномерное распределение на отрезке $[0, 1]$. Если $r_i<p$, $i=1, 2, ..., k$, то считаем, что в i-м испытании наступило событие A; если $r_i>p$, то считаем, что в i-м испытании наступило событие B. Эти допущения правомерны.

□ Действительно, так как R — случайная величина, имеющая равномерное распределение на отрезке $[0, 1]$, то $P(0<R<p)\underset{(13.7)}{=}p$, т. е. $P(R<p)=P(A)$. Справедливо также следующее равенство: $P(p<R<1)\underset{(13.7)}{=}1-p$, т. е. $P(R>p)=P(B)$. ■

Теперь предположим, что результатом каждого из k независимых испытаний может быть появление одного из n несовместных событий $A_1, A_2, ..., A_n$, образующих полную группу. Известна вероятность появления каждого события $P(A_i)=p_i$, $i=1, 2, ..., n$, которая не изменяется при переходе от одного испытания к другому. Заметим, что так как события $A_1, A_2, ..., A_n$ несовместны и образуют полную группу, то $p_1+p_2+...+p_n=1$.

Моделирование такой последовательности испытаний осуществляется следующим образом.

Разделим отрезок $[0, 1]$ на n участков $\Delta_1, \Delta_2, ..., \Delta_n$, длины которых соответственно равны $p_1, p_2, ..., p_n$ (рис. 13.3). Получаем последовательность значений $r_1, r_2, ..., r_k$ случайной величины R. Если $r_i \in \Delta_m$, то считаем, что в i-м испытании наступило событие A_m. Это допущение правомерно, так как $P(R\in\Delta_m)=P(A_m)$.

□ Действительно, $P(R\in\Delta_m)\underset{(13.7)}{=}$Длине отрезка $\Delta_m=p_m=$ $=P(A_m)$. ■

○ **Пример 13.1.** Проводится последовательность независимых испытаний. В результате каждого испытания может произойти одно из несовместимых событий A_1, A_2, A_3, образующих полную группу; $P(A_1)=0{,}35$; $P(A_2)=0{,}25$; $P(A_3)=0{,}4$.

Моделирование такой последовательности испытаний осуществляется следующим образом. Разделим, как показано на рис. 13.4, отрезок $[0, 1]$ на три

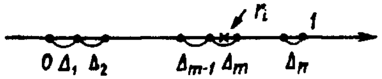


Рис. 13.3

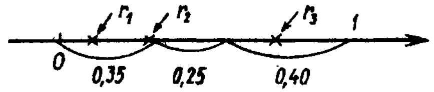


Рис. 13.4

участка. Из табл. П.8 возьмем двухразрядные числа: 15, т. е. $r_1=0{,}15$; это число попало на первый участок, значит, в первом испытании произойдет событие A_1. Значение $r_2=0{,}34$, т. е. во втором испытании опять произойдет событие A_1; далее, $r_3=0{,}71$ — в третьем испытании произойдет событие A_3 и т. д. ●

Последовательность зависимых испытаний. Проводится последовательность зависимых испытаний. В результате каждого испытания может произойти одно из противоположных событий A и B.

Моделирование этой последовательности испытаний осуществляется так:

1) получаем значение r_1 случайной величины R. Если $r_1<P_1(A)$, где $P_1(A)$ — вероятность наступления события A в 1-м испытании, то считаем, что в 1-м испытании произошло событие A. Если $r_1 \geqslant P_1(A)$, то фиксируем появление события B. Допустим, что в первом испытании появилось событие A;

2) получаем следующее значение r_2. Если $r_2<P_2(A/A)$ где $P_2(A/A)$ — условная вероятность появления во 2-м испытании события A при условии, что в первом испытании произошло событие A, то фиксируем появление во 2-м испытании события A; если $r_2 \geqslant P_2(A/A)$, считаем, что произошло событие B. Допустим, что произошло B;

3) получаем следующее значение r_3. Если $r_3<P_3(A/AB)$, где $P_3(A/AB)$ — вероятность появления в 3-м испытании события A при условии наступления в 1-м и 2-м испытаниях событий A и B, то считаем, что в 3-м испытании появилось событие A; в противном случае — событие B и т. д.

Этот алгоритм моделирования зависимых испытаний легко обобщить на случай, когда результатом каждого испытания может быть появление не только двух, а и большего числа событий.

§ 13.5. Моделирование дискретной случайной величины

Общий алгоритм моделирования. Если случайная величина дискретная, то ее моделирование (получение последовательности ее значений) можно свести к моделированию независимых испытаний. Действительно, пусть имеет место следующий ряд распределения:

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

Обозначим A_i событие, состоящее в том, что случайная величина X примет значение x_i. Тогда нахождение значения, принятого случайной величиной X в результате испытания, сводится к определению того, какое из событий $A_1, A_2, \ldots, A_n$ появится. Так как события $A_1, A_2, \ldots, A_n$ несовместны, образуют полную группу и вероятность появления каждого из них не изменяется от испытания к испытанию, то для определения последовательности значений, принятых случайной величиной X, можно использовать процедуру моделирования последовательности независимых испытаний (см. § 13.4).

Помимо рассмотренного выше общего алгоритма моделирования дискретной случайной величины для многих законов распределения существуют специальные алгоритмы. Рассмотрим специальные алгоритмы моделирования случайных величин с биномиальным распределением и распределением Пуассона.

Моделирование случайной величины с биномиальным распределением. Напомним, что в соответствии с биномиальным распределением вероятность $P_n(m)$ того, что определенное событие появится m раз в n независимых испытаниях:

$$P_n(m) = C_n^m p^m (1-p)^{n-m},$$

где p — вероятность появления события в каждом отдельно взятом испытании.

Введем случайную величину X_i — число появлений события в i-м испытании, $i = 1, 2, \ldots, n$. Очевидно, что эта величина может принимать только два значения: либо 1 с вероятностью p, либо 0 с вероятностью $(1-p)$, т. е. ряд распределения величины X_i такой:

$$\begin{array}{c|cc} X_i & 1 & 0 \\ \hline P & p & 1-p \end{array} \qquad (13.11)$$

Тогда случайное число m появлений события в n испытаниях

$$m = X_1 + X_2 + \ldots + X_n. \qquad (13.12)$$

Исходя из соотношения (13.12) и распределения (13.11), определение значения случайной величины m сводится к следующей процедуре:

1) получают последовательность значений $r_1, r_2, \ldots, r_n$ случайной величины R;

2) для каждого числа r_i, $i = 1, 2, \ldots, n$, проверяют, выполняется ли неравенство $r_i < p$. Если неравенство выполняется, то полагают $X_i = 1$; в противном случае считают $X_i = 0$;

3) находят сумму значений n случайных величин X_i (это и будет значение случайной величины m).

Повторяя эту процедуру, получают последовательность значений m_1, $m_2, \ldots$ случайной величины с биномиальным законом распределения.

Пример 13.2. Найдем последовательность значений случайной величины m с биномиальным законом распределения, если $n = 7$, $p = 0{,}3$.

Из табл. П.8 берем семь значений случайной величины R. Пусть $r_1 = 0{,}15$, $r_2 = 0{,}34$, $r_3 = 0{,}71$, $r_4 = 0{,}06$, $r_5 = 0{,}28$, $r_6 = 0{,}36$, $r_7 = 0{,}78$. Три числа не превосходят $p = 0{,}3$. Следовательно, m принимает значение 3. Из той же табл. П.8 берем очередные семь чисел: 0,73; 0,55; 0,74; 0,75; 0,45; 0,61; 0,28. Одно из них не превосходит $p = 0{,}3$. Поэтому следующее значение величины m равно 1 и т. д.

Моделирование случайной величины, распределенной по закону Пуассона. Распределение Пуассона используют в том случае, когда число n независимых испытаний велико, а вероятность p появления события в каждом отдельно взятом испытании мала; при этих условиях вероятность появления события m раз в n испытаниях

$$P_m = \frac{\lambda^m}{m!} e^{-\lambda},$$

где $\lambda = np$ — математическое ожидание, или, иначе говоря, среднее число появлений события в n испытаниях.

Напомним «грубое» правило для применения распределения Пуассона (вместо биномиального распределения), которое состоит в том, что n должно иметь порядок не менее нескольких десятков, лучше нескольких сотен, а $np < 10$.

При решении практических задач, связанных с законом Пуассона, обычно задается параметр λ, а ни n, ни p неизвестны. Алгоритм моделирования случайной величины, распределенной по закону Пуассона, при заданном λ следующий:

1) выбирают n такое, чтобы вероятность $p = \lambda/n$ была достаточно малой ($p < 0{,}01$);

2) получают последовательность значений $r_1, r_2, \ldots, r_n$ случайной величины R, равномерно распределенной на отрезке $[0, 1]$;

3) для каждого числа r_i, $i = 1, 2, \ldots, n$, проверяют, выполняется ли неравенство $r_i < p$; если это неравенство выполняется, то полагают $X_i = 1$, в противном случае считают $X_i = 0$;

4) вычисляют $\sum_{i=1}^{n} X_i$ (это и есть значение случайной величины, распределенной по закону Пуассона).

§ 13.6. Моделирование непрерывной случайной величины

Метод обратных функций. Допустим, что нужно найти последовательность значений случайной величины X, имеющей монотонно возрастающую функцию распределения $F_X(x)$ (рис. 13.5).

Введем случайную величину $Y = F_X(X)$; обратим внимание на то, что так как $0 \leqslant F_X \leqslant 1$, то и $0 \leqslant Y \leqslant 1$. Найдем функцию распределения случайной величины Y:

$$F_Y(y) = P(Y < y) = \begin{cases} 0 \text{ при } y \leqslant 0, \\ P\left[F_X(X) < y\right] = P\left[X < F_X^{-1}(y)\right] = F_X\left[F_X^{-1}(y)\right] = y \text{ при } 0 \leqslant y \leqslant 1, \\ 1 \text{ при } y \geqslant 1, \end{cases}$$

т. е. Y — случайная величина, имеющая равномерное распределение на отрезке $[0, 1]$. Напомним, что такую случайную величину мы обозначали R.

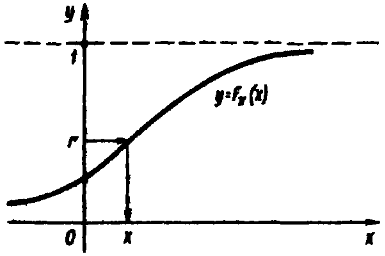


Рис. 13.5

Таким образом, случайная величина X с монотонно возрастающей функцией распределения $F_X(x)$ связана со случайной величиной R соотношением $F_X(X)=R$. Отсюда следует, что значение x случайной величины X является решением уравнения

$$F_X(x)=r, \tag{13.13}$$

где r — значение случайной величины R, т. е.

$$x=F_X^{-1}(r).$$

Графическое решение уравнения (13.13) изображено на рис. 13.5.

Последовательности значений $r_1, r_2, \ldots$ случайной величины R соответствует последовательность $F_X^{-1}(r_1), F_X^{-1}(r_2), \ldots$ значений величины X с функцией распределения $F_X(x)$.

Метод получения значений величины X, использующий функцию, обратную функции $F_X(x)$, называют *методом обратных функций.*

Моделирование случайной величины с равномерным распределением на отрезке [a, b]. Пусть случайная величина X имеет равномерное распределение на отрезке $[a, b]$. Тогда для $x\in[a, b]$ функция распределения

$$F_X(x)=\frac{x-a}{b-a}.$$

Составим уравнение (13.13). Имеем

$$\frac{x-a}{b-a}=r.$$

Отсюда

$$x=a+r(b-a). \tag{13.14}$$

Последовательности значений $r_1, r_2, \ldots$ случайной величины R соответствует последовательность значений $x_1=a+r_1(b-a)$, $x_2=a+r_2(b-a), \ldots$ величины X, равномерно распределенной на отрезке $[a, b]$.

Моделирование случайной величины с показательным распределением. Пусть случайная величина X имеет показательное распределение с параметром μ. Тогда функция распределения $F_X(x)=1-e^{-\mu x}$ для $x\geqslant 0$.

Составим уравнение (13.13). Имеем

$$1-e^{-\mu x}=r.$$

Отсюда

$$x=-\frac{1}{\mu}\ln(1-r). \tag{13.15}$$

Так как R — случайная величина, равномерно распределенная на отрезке $[0, 1]$, то и $(1-R)$ также случайная величина, равномерно распределенная на отрезке $[0, 1]$. Поэтому вместо формулы (13.15) для моделирования случайной величины X используют следующее соотношение:

$$x=-\frac{1}{\mu}\ln r. \tag{13.16}$$

Моделирование случайной величины с нормальным распределением. Для моделирования случайной величины с нормальным законом распределения можно использовать как метод обратных функций, так и методы, специально разработанные для нормального закона.

При использовании ЭВМ обычно применяют метод, основанный на центральной предельной теореме: *если случайные величины $X_1, X_2, \ldots, X_n$ независимы, одинаково распределены и их математические ожидания и дисперсии конечны, то при увеличении n закон распределения суммы $(X_1+X_2+\ldots+X_n)$ приближается к нормальному.*

Оказывается, что для получения хорошего приближения к нормальному распределению достаточно сравнительно небольшого числа слагаемых.

Допустим, что требуется получать значения нормально распределенной случайной величины X с математическим ожиданием MX и дисперсией σ_X^2, т. е. $X=N(MX, \sigma_X)$.

□ Пусть $R_1, R_2, \ldots, R_n$ — независимые случайные величины, равномерно распределенные на отрезке $[0, 1]$. Обозначим через Y сумму этих величин

$$Y=R_1+R_2+\ldots+R_n. \tag{13.17}$$

Учитывая, что $MR_i=0{,}5$, $DR_i=1/12$, $i=1, 2, \ldots, n$, найдем $MY=0{,}5n$, $DY=n/12$.

При достаточно большом n (практически при $n\geqslant 12$) можно считать, что Y имеет нормальный закон распределения с математическим ожиданием $MY=0{,}5n$ и дисперсией $DY=n/12$, т. е.

$$Y=N(0{,}5n;\ \sqrt{n/12}).$$

Перейдем от величины Y к стандартной нормально распределенной случайной величине

$$U=\frac{Y-MY}{\sqrt{DY}}=\frac{Y-0{,}5n}{\sqrt{n/12}}=(Y-0{,}5n)\frac{6}{\sqrt{3n}}, \tag{13.18}$$

для которой $MU=0$, а $DU=1$.

Перейдем от величины X к стандартной нормально распределенной величине

$$U=\frac{X-MX}{\sigma_X}.$$

Тогда

$$X=MX+\sigma_X U.$$

Учитывая (13.18) и (13.17), получаем

$$X=MX+\sigma_X\left(\sum_{i=1}^{n} R_i-0,5n\right)\frac{6}{\sqrt{3n}}.$$

Например, при $n=12$

$$X=MX+\sigma_X\left(\sum_{i=1}^{12} R_i-6\right).$$

Отсюда значение x случайной величины X

$$x=MX+\sigma_X\left(\sum_{i=1}^{12} r_i-6\right), \tag{13.19}$$

где $r_1, r_2, ..., r_{12}$ — значения случайной величины R, равномерно распределенной на отрезке [0, 1].

Таким образом, имея 12 значений случайной величины R и подставив их в формулу (13.19), получают значение случайной величины $X=N(MX, \sigma_X)$; имея следующие 12 значений величины R и подставив их в (13.19), получают следующее значение величины $X=N(MX, \sigma_X)$ и т. д. ■

§ 13.7. Применение метода статистических испытаний к моделированию системы массового обслуживания

Общие сведения о системах массового обслуживания. За последние десятилетия в самых разных областях народного хозяйства возникла необходимость решения вероятностных задач, связанных с работой систем массового обслуживания. Примерами таких систем служат телефонные станции, ремонтные мастерские, торговые предприятия, билетные кассы и т. д. Работа любой системы массового обслуживания состоит в обслуживании поступающего в нее потока требований (вызовы абонентов, приход покупателей в магазин, требования на выполнение работы в мастерской и т. д.).

Система массового обслуживания считается заданной, если определены:

1) входящий поток требований, или, иначе говоря, моменты поступления требований в систему. Первопричину требований называют *источником*. В дальнейшем условимся считать, что источник располагает неограниченным числом требований и что требования однородны, т. е. различаются только моментами появления в системе;

2) система обслуживания, состоящая из накопителя и узла обслуживания. Последний представляет собой одно или несколько обслуживающих устройств, которые в дальнейшем будем называть *приборами*. Каждое требование должно поступить на один из приборов, чтобы пройти обслуживание. Может оказаться, что требованиям придется ожидать, пока приборы освободятся. В этом случае требования находятся в накопителе, образуя одну или несколько *очередей*. Положим, что переход требования из накопителя в узел обслуживания происходит мгновенно;

3) время обслуживания требования каждым прибором;

4) дисциплина ожидания, т. е. совокупность правил, регламентирующих количество требований, находящихся в один и тот же момент времени в системе. Система, в которой поступившее требование получает отказ, когда все приборы заняты, называется *системой без ожидания*. Если требование, заставшее все приборы занятыми, становится в очередь и ожидает до тех пор, пока освободится один из приборов, то такая система называется *чистой системой с ожиданием*. Система, в которой требование, заставшее все приборы занятыми, становится в очередь только в том случае, когда число требований, находящихся в системе, не превышает определенного уровня (в противном случае происходит потеря требования), называется *смешанной системой обслуживания*;

5) дисциплина обслуживания, т. е. совокупность правил, в соответствии с которыми требование выбирает прибор, которым оно будет обслужено. В дальнейшем условимся считать, что все приборы одинаковы и требование не отдает предпочтения ни одному из приборов, когда некоторые из них свободны;

6) дисциплина очереди, т. е. совокупность правил, в соответствии с которыми требование отдает предпочтение той или иной очереди (если их несколько) и располагается в выбранной очереди. Например, поступившее требование может занять место в самой короткой очереди; в этой очереди оно может расположиться последним (такая очередь называется *упорядоченной*), а может пойти на обслуживание вне очереди. Возможны и другие варианты.

На рис. 13.6 приведена схема системы с ожиданием.

Математическая дисциплина, изучающая модели реальных систем массового обслуживания, получила название *теории массового обслуживания*. Задача теории массового обслуживания — установить зависимость результирующих показателей

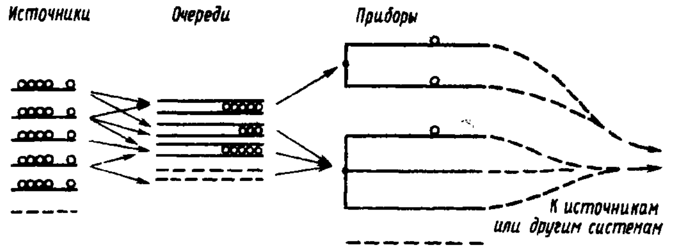


Рис. 13.6

работы системы массового обслуживания (вероятности того, что требование будет обслужено; математического ожидания числа обслуженных требований и т. д.) от входных показателей (количества приборов в системе, параметров входящего потока требований и т. д.). Установить такие зависимости в формульном виде можно только для простых систем массового обслуживания. Изучение же реальных систем проводится путем имитации, или моделирования их работы на ЭВМ с привлечением метода статистических испытаний. Рассмотрим пример.

Моделирование системы массового обслуживания. Рассмотрим следующую систему:

1. Требования поступают в случайные моменты времени, при этом промежуток времени Θ между любыми двумя последовательными требованиями имеет показательный закон с параметром μ, т. е. функция распределения

$$F_{\Theta}(\tau)=1-e^{-\mu\tau},\quad \tau\geqslant 0. \qquad (13.20)$$

2. Система обслуживания состоит из s одинаковых, пронумерованных приборов.

3. Время $T_{обсл}$ — случайная величина с равномерным законом распределения на отрезке $[a,\ b]$.

4. Система без ожидания, т. е. требование, заставшее все приборы занятыми, покидает систему.

5. Дисциплина обслуживания такова: если в момент поступления k-го требования первый прибор свободен, то он приступает к обслуживанию требования; если этот прибор занят, а второй свободен, то требование обслуживается вторым прибором, и т. д.

Требуется оценить математические ожидания числа требований, обслуженных системой за время T и получивших отказ.

За начальный момент расчета выберем момент поступления первого требования $T_1=0$. Введем следующие обозначения: T_k — момент поступления k-го требования; t_i — момент окончания обслуживания требования i-м прибором, $i=1,\ 2,\ 3,\dots,s$.

Предположим, что в момент T_1 все приборы свободны.

Первое требование поступает на прибор 1. Время обслуживания этим прибором имеет равномерное распределение на отрезке $[a,\ b]$. Поэтому конкретное значение $t_{обсл}$ этого времени находим по формуле (13.14), т. е.

$$t_{обсл}=a+r(b-a),$$

где r — значение случайной величины R, равномерно распределенной на отрезке $[0,\ 1]$. Прибор 1 будет занят в течение времени $t_{обсл}$. Поэтому момент t_1 окончания обслуживания требования прибором 1 следует считать равным

$$t_1=T_1+t_{обсл}.$$

Затем следует добавить единицу в счетчик обслуженных требований и перейти к рассмотрению следующего требования.

Предположим, что k требований уже рассмотрено. Определим момент T_{k+1} поступления $(k+1)$-го требования. Для этого найдем значение τ промежутка времени между последовательными требованиями. Так как этот промежуток имеет показательный закон (13.20), то в соответствии с (13.16) будем иметь

$$\tau=-\frac{1}{\mu}\ln r,$$

где r — очередное значение случайной величины R. Тогда момент поступления $(k+1)$-го требования

$$T_{k+1}=T_k+\tau.$$

Свободен ли в этот момент первый прибор? Для ответа на этот вопрос необходимо проверить условие $t_1\leqslant T_{k+1}$. Если это условие выполнено, то к моменту T_{k+1} первый прибор освободился и может обслуживать требование. В этом случае t_1 заменяем на $(T_{k+1}+t_{обсл})$, добавляем единицу в счетчик обслуженных требований и переходим к следующему требованию. Если $t_1>T_{k+1}$, то первый прибор в момент T_{k+1} занят. В этом случае проверяем, свободен ли второй прибор. Если условие $t_2\leqslant T_{k+1}$ выполнено, заменяем t_2 на $(T_{k+1}+t_{обсл})$, добавляем единицу в счетчик обслуженных требований и переходим к следующему требованию. Если $t_2>T_{k+1}$, то проверяем условие $t_3\leqslant T_{k+1}$ и т. д. Если при всех i от 1 до s имеет $t_i>T_{k+1}$, то в момент T_{k+1} все приборы заняты. В этом случае прибавляем единицу в счетчик отказов и переходим к рассмотрению следующего требования. Каждый раз, вычислив T_{k+1}, надо проверить еще условие окончания реализации: $T_{k+1}\geqslant T$. Если это условие выполнено, то одна реализация процесса функционирования системы воспроизведена и испытание заканчивается. В счетчике обслуженных требований и в счетчике отказов находятся числа $n_{обсл}$ и $n_{отк}$.

Повторив такое испытание n раз (с использованием различных r) и осреднив результаты опытов, определим оценки математических ожиданий числа обслуженных требований и числа требований, получивших отказ:

$$Mn_{\text{обсл}} \approx \frac{1}{n}\sum_{j=1}^{n}\left(n_{\text{обсл}}\right)_j,$$

$$Mn_{\text{отк}} \approx \frac{1}{n}\sum_{j=1}^{n}\left(n_{\text{отк}}\right)_j,$$

где $\left(n_{\text{обсл}}\right)_j$ и $\left(n_{\text{отк}}\right)_j$ — значения величин $n_{\text{обсл}}$ и $n_{\text{отк}}$ в j-м опыте. На рис. 13.7 приведена схема рассмотренного алгоритма.

В заключение заметим, что для получения результатов, имеющих практическую ценность, надо хорошо изучить систему обслуживания, знать вероятностные законы функционирования отдельных ее частей. Тогда, используя метод статистических испытаний, можно оценить вероятностные характеристики любой системы, как бы сложна она ни была.

Моделирование систем массового обслуживания находит широкое применение при проектировании.

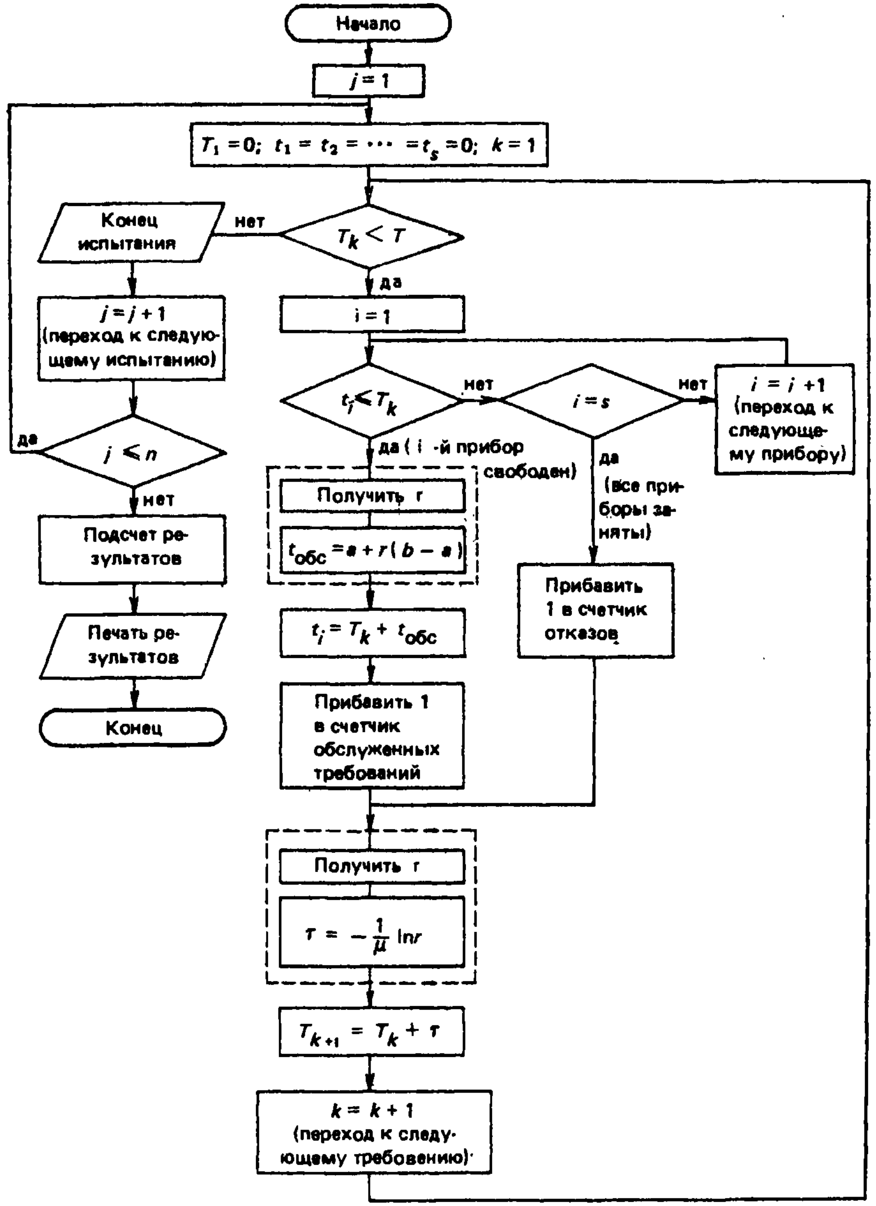
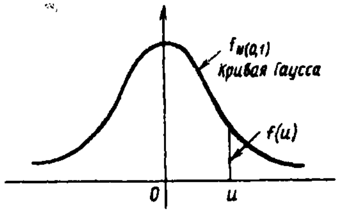


Рис. 13.7

Таблица 1

Значения функции плотности стандартной нормальной величины $U=N(0,1)$:

$$f_U(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$$

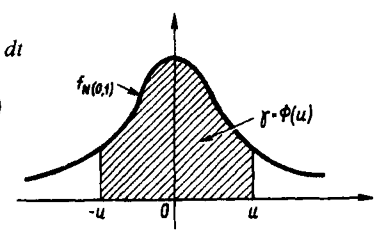


u	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3639	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3188	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Таблица 2

Значение функции Лапласа $\Phi(u)=\frac{2}{\sqrt{\pi}}\int_0^u e^{-t^2/2}\,dt$

или значение вероятности $\gamma=P(|N(0,1)|<u)$

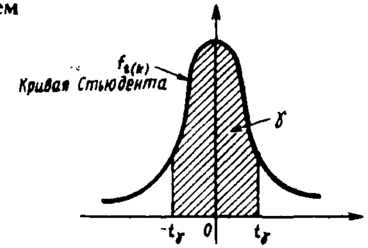


u	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0080	0,0160	0,0238	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,1	0797	0876	0955	1034	1113	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2960	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6679	6729	6778
1,0	0,6827	0,6875	0,6923	0,6970	0,7017	0,7063	0,7109	0,7154	0,7199	0,7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7994	8029
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8789	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9090
1,7	9109	9127	9146	9164	9181	9189	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9426	9439	9451	9464	9476	8488	9500	9512	9523	9534
2,0	0,9545	0,9556	0,9566	0,9576	0,9586	0,9596	0,9606	0,9616	0,9625	0,9634
2,1	9643	9651	9660	9668	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9756	9762	9768	9774	9780
2,3	9786	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9841	9845	9849	9853	9857	9861	9865	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9910	9912	99915	99917	99920	9922	9924	9926	9928
2,7	9931	9933	9935	9937	9938	9940	9942	9944	9946	9947
2,8	9949	9951	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972
3,0	0,9973	0,9974	0,9975	0,9976	0,9976	0,9977	0,9978	0,9979	0,9979	0,9980
3,1	9981	9981	9982	9983	9983	9984	9984	9985	9985	9986
3,5	9995	9996	9996	9996	9996	9996	9996	9996	9997	9997
3,6	9997	9997	9997	9997	9997	9997	9997	9998	9998	9998
3,7	9998	9998	9998	9998	9998	9998	9998	9998	9998	9998
3,8	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
3,9	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
4,0	0,999936	9999	9999	9999	9999	9999	9999	9999	9999	9999
4,5	0,999994	—	—	—	—	—	—	—	—	—
5,0	0,99999994	—	—	—	—	—	—	—	—	—

Таблица 3

Значения t_γ, определяемые уравнением

$$P=(|t(k)|<t_\gamma)=\gamma$$

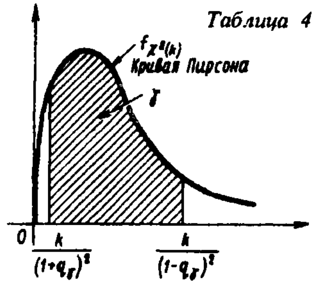


k \ γ	0,5	0,6	0,7	0,8	0,9	0,95	0,98	0,99	0,999
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,816	1,061	1,336	1,886	2,920	4,303	6,965	9,925	31,598
3	765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941
4	741	941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	727	920	1,156	1,476	2,015	2,571	3,365	4,032	6,859
6	718	906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	711	896	1,119	1,415	1,895	2,365	2,998	3,499	5,405
8	706	889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	703	883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	700	879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	697	876	1,088	1,363	1,796	2,201	2,718	3,106	4,487
12	695	873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	694	870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	692	868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	691	866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	690	865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	689	863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	688	862	1,067	1,330	1,734	2,103	2,552	2,878	3,922
19	688	861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	687	860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	686	859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	686	858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	685	858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	685	857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	684	856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	684	856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	684	855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	683	855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	683	854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	683	854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	681	851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	679	848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	677	845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	674	842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Таблица 4

Значения q_γ, определяемые уравнением

$$P\left(\frac{k}{(1+q_\gamma)^2}<\chi^2(k)<\frac{k}{\max^2(0,(1-q_\gamma))}\right)=\gamma$$



k \ γ	0,5	0,6	0,7	0,8	0,9	0,95	0,98	0,99	0,999
1	0,568	906	1,602	2,946	6,923				
2	367	473	0,678	1,125	2,086	3,400	5,857	8,500	
3	290	370	482	0,730	1,270	1,932	3,000	4,200	9,00
4	248	306	398	563	0,941	1,382	2,056	2,700	5,00
5	221	277	348	475	738	1,104	1,594	2,00	3,80
6	200	251	308	416	623	0,918	1,306	1,650	3,00
7	185	232	290	380	576	800	1,143	1,393	2,50
8	173	216	269	354	516	713	0,986	1,225	2,05
9	162	202	252	329	476	650	889	1,094	1,75
10	153	192	239	304	442	596	814	0,980	1,50
12	140	176	218	276	388	527	700	840	1,30
14	130	162	290	252	357	468	620	740	1,14
16	122	150	188	236	325	422	564	671	1,02
18	115	143	177	223	297	390	500	600	0,92
20	108	136	168	210	282	370	480	567	85
25	096	122	148	187	247	317	408	485	70
30	0,088	111	137	172	226	281	369	425	60
35	085	101	127	156	207	261	347	400	56
40	076	095	119	146	193	342	312	375	52
50	068	084	105	133	174	212	270	311	45
60	062	077	095	122	155	193	242	283	40
70	057	072	088	112	145	180	222	250	37
80	054	067	082	103	138	167	200	236	35
90	051	063	078	096	131	151	192	220	32
100	048	060	074	092	125	146	184	200	30

Таблица 5

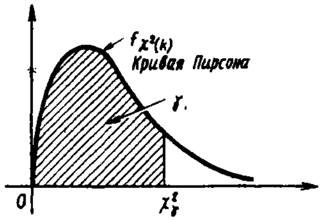
Доверительные границы p_2 и p_1 для параметра p при $\gamma=0,95$

m \ $n-m$	1	2	3	4	5	6	7	8	9	10
0	0,975	842	708	602	522	459	410	369	336	308
	000	000	000	000	000	000	000	000	000	000
1	987	906	806	716	641	579	527	483	445	413
	013	008	006	005	004	004	003	003	003	002
2	992	932	853	777	710	651	600	556	518	484
	094	068	053	043	037	032	028	025	023	021
3	994	947	882	816	755	701	652	610	572	538
	194	147	118	099	085	075	067	060	055	050
4	995	957	901	843	788	738	692	651	614	581
	284	223	184	157	137	122	109	099	091	084

Таблица 6

Значения χ^2_γ, определяемые уравнением

$P(\chi^2(k) < \chi^2_\gamma) = \gamma$

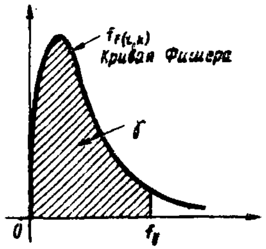


k \ γ	0,50	0,70	0,80	0,90	0,95	0,98	0,99	0,999
1	0,455	1,074	1,642	2,71	3,84	5,41	6,64	10,83
2	1,386	2,41	3,22	4,60	5,99	7,82	9,21	13,82
3	2,37	3,66	4,64	6,25	7,82	9,84	11,34	16,27
4	3,36	4,88	5,99	7,78	9,49	11,67	13,28	18,46
5	4,35	6,06	7,29	9,24	11,07	13,39	15,09	20,5
6	5,35	7,23	8,56	10,64	12,59	15,03	16,81	22,5
7	6,35	8,38	9,80	12,02	14,07	16,62	18,48	24,3
8	7,34	9,52	11,03	13,36	15,51	18,17	20,1	26,1
9	8,34	10,66	12,24	14,68	16,92	19,68	21,7	27,9
10	9,34	11,78	13,44	15,99	18,31	21,2	23,2	29,6
11	10,34	12,90	14,63	17,28	19,68	22,6	24,7	31,3
12	11,34	14,01	15,81	18,55	21,0	24,1	26,2	32,9
13	12,34	15,12	16,98	19,81	22,4	25,5	27,7	34,5
14	13,34	16,22	18,15	21,1	23,7	26,9	29,1	36,1
15	14,34	17,32	19,31	22,3	25,0	28,3	30,6	37,7
16	15,34	18,42	20,5	23,5	26,3	29,6	32,0	39,3
17	16,34	19,51	21,6	24,8	27,6	31,0	33,4	40,8
18	17,34	20,6	22,8	26,0	28,9	32,3	34,8	42,3
19	18,34	21,7	23,9	27,2	30,1	33,7	36,2	43,8
20	19,34	22,8	25,0	28,4	31,4	35,0	37,6	45,3
21	20,3	23,9	26,2	29,6	32,7	36,3	38,9	46,8
22	21,3	24,9	27,3	30,8	33,9	37,7	40,3	48,3
23	22,3	26,0	28,4	32,0	35,2	39,0	41,6	49,7
24	23,3	27,1	29,6	33,2	36,4	40,3	43,0	51,2
25	24,3	28,2	30,7	34,4	37,7	41,6	44,3	52,6
26	25,3	29,2	31,8	35,6	38,9	42,9	45,6	54,1
27	26,3	30,3	32,9	36,7	40,1	44,1	47,0	55,5
28	27,3	31,4	34,0	37,9	41,3	45,4	48,3	56,9
29	28,3	32,5	35,1	39,1	42,6	46,7	49,6	58,3
30	29,3	33,5	36,2	40,3	43,8	48,0	50,9	59,7

Таблица 7

Значения f_γ, определяемые уравнением

$P(F(l, k) < f_\gamma) = \gamma = 0,95$



k	l									
	1	2	3	4	5	6	8	12	24	∞
1	161,40	199,50	215,70	224,60	230,20	234,00	238,90	243,90	249,00	254,30
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,33	3,51	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,02	1,83	1,61	1,25
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Таблица 8

Таблица случайных чисел (фрагмент)

1534	7106	2836	7873	5574	7545
6128	8993	4102	2551	0330	2358
6047	8566	8644	9343	9297	6751
0806	5201	5705	7355	1448	9562
9915	8274	4525	5695	5752	9630
2882	7158	4341	3463	1178	5786
9213	1223	4388	9760	6691	6861
8410	9836	3899	3683	1253	1683
9974	2362	2103	4326	3825	9079
3402	8162	8226	0782	3364	7871

Учебное издание

Калинина Вера Николаевна
Панкин Виктор Федорович

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебник для студентов
средних специальных учебных заведений

Зав. редакцией *Б. В. Понкратов*
Ответственный редактор *Ж. И. Яковлева*
Художественное оформление *Е. А. Адамов*

Изд. лиц. № 061622 от 07.10.97.
Подписано к печати 30.05.02. Формат 60х90 $^1/_{16}$.
Бумага типографская. Гарнитура «Таймс». Печать офсетная.
Усл. печ. л. 20,58. Тираж 10 000 экз. Заказ № 4210170.

ООО «Дрофа». 127018, Москва, Сущевский вал, 49.

По вопросам приобретения продукции издательства «Дрофа» обращаться по адресу: 127018, Москва, Сущевский вал, 49.
Тел.: (095) 795-05-50, 795-05-51. Факс: (095) 795-05-52.
Торговый дом «Школьник».
109172, Москва, ул. Малые Каменщики, д. 6, стр. 1А.
Тел.: (095) 911-70-24, 912-15-16, 912-45-76.
Магазин «Переплетные птицы».
127018, Москва, ул. Октябрьская, д. 89, стр. 1.
Тел.: (095) 912-45-76.

Отпечатано с готовых диапозитивов
в ФГУИПП «Нижполиграф»
603006, г. Нижний Новгород, ул. Варварская, 32.