МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ

НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
"ХАРЬКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ"

А.С. МАЗМАНИШВИЛИ

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

УЧЕБНОЕ ПОСОБИЕ К ПРАКТИЧЕСКИМ ЗАНЯТИЯМ

Рекомендовано Министерством образования и науки Украины
как учебное пособие для студентов специальностей
7.080201 "Информатика" и 7.080202 "Прикладная математика"

Утверждено Редакционно-издательским
Советом НТУ "ХПИ"

Харьков НТУ "ХПИ" 2010

ББК 22.171
М12
УДК 519.2

Рецензенты:
Е.В. Бодянский, д-р техн. наук, проф.,
Харьковский национальный университет радиоэлектроники
Г.И. Загарий, д-р техн. наук, проф., Харьковская государственная академия железнодорожного транспорта

Гриф присвоен Министерством образования и науки Украины, письмо № @@@@ от """"

М12 Математическая статистика: Учебн. пособие к практическим занятиям / Мазманишвили А.С. — Харьков: НТУ "ХПИ", 2010. — 232 с. — Русск. яз.

ISBN 966–593–270–5

Подготовлено для выполнения практикума по курсу "Математическая статистика". В пособии систематизированы материалы по основным темам дисциплины: теоретические сведения, необходимые для решения задач, примеры таких решений.

Предназначено для студентов, обучающихся по специальности "Прикладная математика". Будет полезно студентам физико-математических, инженерно-технических и экономических специальностей университетов, а также специалистам.

Підготовлений для виконання практикуму з курсу "Математична статистика". У посібнику систематизовано матеріали з основних тем дисципліни: теоретичні відомості, необхідні для розв'язування задач, приклади таких розв'язків.

Призначений для студентів, що навчаються за спеціальністю "Прикладна математика". Буде корисний студентам фізико-математичних, інженерно-технічних і економічних спеціальностей університетів, а також фахівцям.

This textbook contains tasks on Mathematical Statistics sorted according to the topics of the course: theoretical information that is necessary for solving the tasks, examples of the solutions.

This textbook is created for students on Applied Mathematics. It can also be useful for students on Mathematics, Physics, Engineering and Economy.

Ил. 55. Табл. 21. Библиогр. 42 назв.

ББК 22.171

ISBN 966–593–270–5

Оглавление

Введение

Математическая и прикладная статистика широко используется в науке, экономике, технике, медицине и других областях благодаря её постоянному развитию, в том числе программному.

Сведения на перспективу изучает теория вероятностей, а ретроспективные сведения – статистика.

Теория вероятностей – это один из разделов чистой математики. Строится эта теория дедуктивно, исходя из некоторых аксиом и предположений. Наиболее строгий подход связан с использованием теории множеств, теории меры и интеграла Лебега. Обычно начинают с построения "элементарной теории вероятностей", в которой рассматриваются случайные события с конечным числом исходов. Затем теория распространяется на случай, когда число возможных исходов бесконечно. Применение теорем к решению различных задач теории вероятностей связано с использованием сочетаний, перестановок, операций суммирования и интегрирования. Применяемые в теории вероятностей методы, такие, например, как преобразование Лапласа, используются и в других разделах математики.

В противоположность теории вероятностей статистика – это раздел прикладной математики. Для неё характерно главным образом индуктивное построение, поскольку в этом случае мы идем в обратном направлении – от наблюдения события к гипотезе. При этом наша аргументация основывается на выводах теории вероятностей, всестороннее знание которой, таким образом, оказывается совершенно необходимым.

Пример 1. *Типичная задача теории вероятностей.* Когда подбрасывается монета, то имеется известная вероятность p, что выпадет "орел", и вероятность $1-p$, что выпадет "решка". Какова вероятность того, что в результате N бросаний "орел" выпадет n раз?

Используя биномиальное распределение, мы получим следующий результат:

$$\Pr(n) \;=\; C_N^n p^n (1-p)^{N-n}.$$

Пример 2. *Типичная задача математической статистики.* Монета подбрасывается N раз, при этом "орел" выпадает n раз. Что можно сказать о неизвестном параметре p?

Очевидно, нельзя надеяться получить на этот вопрос столь же определенный ответ, что и в предыдущем случае. С самого начала мы знаем, что $0 \le p \le 1$. Кроме того, $p \neq 0$, если $n \neq 0$, и $p \neq 1$, если $n \neq N$.

Рассмотрим понятие *наиболее правдоподобного значения* параметра. В данном случае мы могли бы сказать, что *наиболее правдоподобное значение p равно n/N*. Затем следовало бы рассмотреть и другие столь же правдоподобные значения p. В

результате получим малый интервал

$$p_1 < n/N < p_2,$$

который, как мы надеемся, будет содержать истинное значение p.

Пусть $\delta p = p_2 - p_1$. Чем больше δp, тем с большей достоверностью p попадет в указанный интервал. С другой стороны, более широкий интервал дает нам меньшую информацию относительно самой величины p. Таким образом, в статистическом анализе всегда присутствует принципиальная неопределенность. Мы можем во всяком случае рассчитывать, что оценим эту неопределенность.

К статистическим методам прибегают в тех случаях, когда приходится рассматривать не единичные, а массовые явления. Первичной обработкой сведений занимается общая статистика, а обработкой сведений на основе применения математических методов – математическая статистика. Последняя является наукой о методах количественного анализа массовых явлений.

Первые шаги по математической статистике были сделаны в XVIII в., они были связаны со статистикой народонаселения и с вопросами страхования. В конце XVIII в. началась серьёзная работа по теории ошибок измерений, приведшая в начале XIX в. к созданию далеко продвинутых её основ. Биологические исследования послужили в XIX в. толчком для постановки многочисленных вопросов, приведших в начале XX в. к выделению математической статистики в особую науку. Сейчас математическая статистика применяется буквально во всех сферах человеческой деятельности. Выпущено огромное количество литературы, освещающей методы математической статистики как общего, так и специализированного направления.

Издание по своему замыслу преследует основную цель: дать студентам удобный для работы и практически апробированный материал, который обучает практическим методам и технике решения различных задач математической статистики.

Пособие включает материал традиционных разделов, образующих в совокупности содержание дисциплины "Математическая статистика" как второй части всей дисциплины "Теория вероятностей и математическая статистика". Материал сгруппирован по следующим основным темам:

1. Основные понятия и задачи математической статистики.
2. Специальные законы распределения математической статистики.
3. Статистическая теория оценивания параметров распределения.
4. Статистическая проверка параметрических гипотез.
5. Статистическая проверка непараметрических гипотез.
6. Элементы линейного регрессионного анализа.

Каждый из перечисленных разделов состоит из теоретической части, объем которой достаточен для усвоения соответствующего материала, и частей, содержащих как развернутые решения примеров, так и задачи для решения, приложения содержат необходимые статистические таблицы. Таким образом, настоящее пособие может служить решебником по дисциплине "Математическая статистика". Учебный материал приводится в таком объеме, что овладение им даёт доступ к использованию современных пакетов прикладных программ по статистической обработке данных.

Для удобства работы, а также возможности самостоятельного углубленного обучения и контроля освоения материала, издание скомпоновано из отдельных само-

стоятельных тем, которые адаптированы к соответствующим разделам дисциплины "Математическая статистика".

В каждом разделе приведены необходимые исходные и справочные данные, принципы построения методов решения соответствующих задач, типовые алгоритмы, сформулированы задания для самостоятельной работы. Практикум основан на знаниях, получаемых студентвами в стандартном объёме специальности "Прикладная математика".

Пособие является решебником задач по курсу математической статистики, он содержит как традиционные примеры и задачи по курсу, так и новые, которые составитель взял из научной практики.

При составлении материала пособия имелось в виду также возможность его адаптации к учебным программам различного объема и длительности.

Настоящее учебно-методическое пособие может рассматриваться как вторая (односеместровая) часть пособия для двухсеместровой дисциплины "Теория вероятностей и математическая статистика", первая (также односеместровая) часть которой – "Теория вероятностей". Изложение материала и все обозначения в этих двух частях согласованы.

1. Основные понятия и задачи математической статистики

1.1. Предмет и задачи математической статистики

Теория вероятностей и математическая статистика занимаются количественным и качественным анализом закономерностей случайных массовых явлений. При рассмотрении задач теории вероятностей исходят из предположения, что вероятности наступления отдельных событий известны и заданы. Считались известными законы распределения случайных величин или их числовые характеристики. Оперируя этими понятиями, находят вероятности, законы распределения и числовые характеристики других более сложных событий и случайных величин (СВ). На практике вероятности наступления событий, законы распределения случайных величин или параметры этих законов распределения неизвестны. Для их определения (оценивания) необходимо производить эксперимент, специальные исследования.

Математическая статистика разрабатывает методы математической обработки результатов испытаний с целью получения сведений о вероятностях наступления отдельных событий, о законах распределения случайных величин или о параметрах этих законов.

При обработке результатов эксперимента статистическими методами основные понятия теории вероятностей – вероятность наступления случайного события, законы распределения случайных величин, параметры законов распределения случайных величин и т.д. – выступают как некоторые математические модели реальных закономерностей.

Таким образом, теория вероятностей разрабатывает математические модели для описания реальных закономерностей случайных массовых явлений, формирует систему взглядов на статистическую обработку результатов эксперимента.

Основой статистических методов являются экспериментальные данные, часто называемые статистическими данными.

Статистическими данными называют сведения о числе объектов, обладающих теми или иными признаками. Например, статистическими данными являются данные об отклонении размеров детали от номинальных размеров; данные о числе вызовов на телефонную станцию между 8 и 9 часами утра и т.д. Перечисленные данные являются числовыми характеристиками массовых случайных явлений (сортности деталей, нагрузки АТС, производительности труда), поэтому предметом математической статистики являются случайные явления, а её основной задачей – количественный и качественный анализы этих явлений.

Основные задачи математической статистики состоят в разработке методов:

1) организации и планирования статистических наблюдений;

2) сбора статистических данных;

3) "свертки информации", т.е. методов группировки и сокращения статистических данных с целью сведения большого числа таких данных к небольшому числу параметров, которые в сжатом виде характеризуют всю исследуемую совокупность;

4) анализа статистических данных;

5) принятия решений, рекомендаций и выводов на основе анализа статистических данных;

6) прогнозирования случайных явлений.

1.2. Генеральная и выборочная совокупности

Одним из основных методов статистического наблюдения является *выборочный метод.* Рассмотрим основные понятия этого метода.

Пусть для исследования закономерностей случайного явления произведено n опытов, в результате которых получен ряд наблюдений $x_1, x_2, \ldots, x_n$. Требуется обработать этот ряд статистически. Для любой статистической обработки необходимо вначале построить *математическую модель* ряда наблюдений, т.е. указать, какие величины случайны, какие не случайны, какие зависимы, какие независимы и т.д.

Для результатов наблюдений $x_1, x_2, \ldots, x_n$ можно построить различные математические модели. Рассмотрим модель, в которой ряд наблюдений дается формулой

$$x_i \;=\; f(t_i) + \varepsilon_i, \tag{1.1}$$

где t_i – значение некоторой детерминированной функции, характеризующей i-й опыт; $f(t_i)$ – некоторая функция определенного или неизвестного вида; ε_i – случайная величина, обычно называемая ошибкой i-го эксперимента.

Относительно ошибок ε_i в модели ряда наблюдений, задаваемого формулой (1.1), можно также высказать различные предположения. Например, можно считать, что измерения x_i сопровождаются систематическими ошибками, т.е. $\mathrm{M}[\varepsilon_i] \neq 0$. Причем можно предполагать, что либо эти систематические ошибки не зависят от i (постоянны): $\mathrm{M}[\varepsilon_i] = const$, либо изменяются по определенному закону: $\mathrm{M}[\varepsilon_i] = \psi(t_i)$.

Предположения о виде функции $f(t)$ и о характере ошибок ε_i в модели ряда наблюдений, задаваемого формулой (1.1), определяют методику обработки этого ряда. Чем сложнее предположения будут выдвинуты относительно $f(t)$ и ошибок ε_i в модели (1.1), тем сложнее будут методы его статистической обработки.

Относительно последовательности, образованной из случайных ошибок ε_i, где $i = 1, 2, \ldots, n$, как правило, предполагают, что они независимы, а проведенные измерения выполнены в одинаковых и стабильных условиях (однородные измерения), т.е. $\mathrm{D}[\varepsilon_1] = \mathrm{D}[\varepsilon_2] = \ldots = \mathrm{D}[\varepsilon_n] = \sigma^2$. В этом случае говорят, что измерения $x_1, x_2, \ldots, x_n$ *равноточны.*

Чаще всего полагают, что ошибки измерения ε_i имеют нормальный закон распределения с параметрами $\mathrm{M}[\varepsilon_i] = 0$ и $\mathrm{M}[\varepsilon_i^2] = \sigma^2$ или в более краткой записи:

$\varepsilon_i \to N(0;\sigma)$. При этом предположении плотность распределения вероятностей исследуемой случайной величины X имеет вид

$$f_{\mathrm{X}} = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right). \tag{1.2}$$

Построив нормальную модель ряда наблюдений, производят оценку параметров a и σ вероятностной модели ряда наблюдений.

Наиболее точные сведения о СВ X можно получить, производя максимально возможное количество измерений этой случайной величины.

Определение 1. *Генеральной совокупностью* называется совокупность всех мыслимых наблюдений, которые могли бы быть сделаны при данном реальном комплексе условий измерений. Число членов, образующих генеральную совокупность, называется *объемом генеральной совокупности*.

Можно выделить три основных вида генеральной совокупности:

1) конечная и реально существующая, например, число бракованных изделий в некоторой партии;

2) бесконечная и реально существующая, например, множество действительных чисел, лежащих между нулем и единицей;

3) воображаемая (гипотетическая) конечная или бесконечная. Например, результаты $x_1, x_2, \ldots, x_n$ измерений некоторой постоянной физической величины являются элементами воображаемой бесконечной совокупности.

Генеральная совокупность является также понятием модельным. Возможно строить различные предположения (строить модели) о функции распределения F_{X} случайной величины X или о параметрах этой модели.

Определение 2. *Выборочной совокупностью* или просто *выборкой объема* n называется совокупность n объектов, отобранных из исследуемой генеральной совокупности.

Ряд распределений $x_1, x_2, \ldots, x_n$ принято рассматривать как выборку объема n из конечной или бесконечной генеральной совокупности.

Определение 3. Метод, состоящий в том, что на основании характеристик и свойств выборки $x_1, x_2, \ldots, x_n$ делаются заключения о числовых характеристиках и законе распределения случайной величины X, называется *выборочным методом*.

Для того, чтобы суждения о законах распределения случайной величины X или о их числовых характеристиках были объективными, необходимо, чтобы выборка была представительная (*репрезентативная*), т.е. чтобы достаточно хорошо представляла исследуемую случайную величину. Важно, чтобы при извлечении выборки каждый элемент генеральной совокупности имел одинаковую с другими элементами вероятность быть включенным в выборку. Другими словами, выбор элементов из генеральной совокупности должен быть случайным и однородным.

По технике отбора элементов из генеральной совокупности в выборочную выборки делятся на повторные и бесповторные. Если каждый обследованный элемент возвращается обратно в генеральную совокупность и, следовательно, может участвовать в дальнейшем отборе, то выборка называется *повторной*. *Бесповторная* выборка состоит в том, что отобранные элементы обратно в генеральную совокупность не возвращаются.

Среднее арифметическое изучаемого признака X в генеральной совокупности называют *генеральным средним* $\overline{x}$, а его дисперсию – *генеральной дисперсией* σ_0^2.

Среднее арифметическое и дисперсия X в выборке называются *выборочным средним* x^* и *выборочной дисперсией* σ^{*2} (звездочкой часто указывают на то, что рассматриваются средние, отнесенные к выборке).

1.3. Статистический ряд

Предположим, что изучается некоторая дискретная или непрерывная СВ, закон распределения которой неизвестен. Для оценки закона распределения этой СВ или его числовых характеристик производится ряд независимых наблюдений $x_1, x_2, \ldots, x_n$. Статистический материал, полученный в результате измерений, представляют в виде таблицы, состоящей из двух строк, в первой из которых даны номера измерений, а во второй – результаты измерений.

i – номер измерения	1	2	3	$\cdots$	n
x_i – результат измерения	x_1	x_2	x_3	$\cdots$	x_n

Таблицу такого вида называют *статистическим рядом*. Она являет собой первичную форму представления статистического материала. Если информация в виде простого статистического ряда при большом числе измерений трудно обозрима, то по нему затруднительно оценить закон распределения исследуемой СВ. Для визуальной оценки закона распределения исследуемой СВ X производят *группировку данных*.

Если изучается дискретная СВ, то наблюденные значения располагаются в порядке возрастания и подсчитываются *частоты* m_i или *частости* $p_i = m_i/n$ появления одинаковых значений СВ X. В результате получаем сгруппированные статистические ряды следующего вида:

x_i – результат измерения	x_1	x_2	x_3	$\cdots$	x_n
m_i – частоты	m_1	m_2	m_3	$\cdots$	m_n

Контроль: $\sum_{i=1}^{n} m_i = n$.

x_i – результат измерения	x_1	x_2	x_3	$\cdots$	x_n
$p_i = m_i/n$ – частости	m_1/n	m_2/n	m_3/n	$\cdots$	m_n/n

Контроль: $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} m_i/n = 1$.

Если изучается непрерывная СВ, то группировка заключается в разбиении интервала наблюденных значений СВ на k частичных интервалов равной длины $[x_0; x_1]$, $[x_1; x_2]$, $[x_2; x_3]$, $\ldots$, $[x_{k-1}; x_k]$ и подсчете частоты m_i или частости $p_i = m_i/n$ попадания наблюденных значений в частичные интервалы. Количество интервалов выбирается произвольно, но обычно не меньше 5 и не больше 15.

В результате составляется интервальный статистический ряд следующего вида:

Интервалы наблюденных значений	$[x_0; x_1]$	$[x_1; x_2]$	$\ldots$	$[x_{k-1}; x_k]$
Частости $p_i = m_i/n$	m_1/n	m_2/n	$\ldots$	m_k/n

Контроль: $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} m_i/n = 1$.

Определение. Перечень наблюденных значений случайной величины X (или интервалов наблюденных значений и соответствующих им частостей $p_i = m_i/n$) называется *статистическим законом распределения случайной величины* X.

В теории вероятностей под законом распределения СВ понимают соответствие между возможными значениями (или интервалами возможных значений СВ) и их вероятностями, а в математической статистике статистический закон распределения устанавливает соответствие между наблюденными значениями (или интервалами наблюденных значений) СВ и соответствующими им частостями.

Статистические законы распределения случайных величин и их графическое изображение позволяют визуально произвести оценку закона распределения исследуемой случайной величины.

1.4. Эмпирическая функция распределения

Эмпирической функцией распределения случайной величины X называют функцию $F_n^*(x)$, определяющую для каждого значения x частость события $\{\mathrm{X} < x\}$:

$$F_n^*(x) = \frac{n_x}{n}, \tag{1.3}$$

где n_x – число x_i, меньших x; n – объем выборки.

Значение эмпирической функции распределения для статистики определяется следующим утверждением.

Теорема Бернулли

Пусть $F_n^*(x)$ – эмпирическая функция распределения, построенная по однородной выборке объёма n из генеральной совокупности с функцией распределения $F_{\mathrm{X}}(x)$. Тогда для любого $x \in (-\infty, \infty)$ и любого $\varepsilon > 0$ справедливо

$$\lim_{n\to\infty} \Pr\left(|F_n^*(x) - F_{\mathrm{X}}(x)| < \varepsilon \right) = 1. \tag{1.4}$$

Из теоремы Бернулли следует, что при достаточно большом объеме выборки функции $F_n^*(x)$ и $F_{\mathrm{X}}(x) = \Pr(\mathrm{X} < x)$ мало отличаются друг от друга.

Отличие эмпирической функции распределения от теоретической состоит в том, что теоретическая функция распределения определяет вероятность события $\{\mathrm{X} < x\}$, а эмпирическая функция определяет относительную частоту этого события.

Таким образом, при каждом x эмпирическая функция $F_n^*(x)$ сходится по вероятности к $F_{\mathrm{X}}(x)$ и, следовательно, при большом объёме выборки может служить приближенным значением (оценкой) функции распределения генеральной совокупности в каждой точке x. Её краткое название – *кумулята*. Она представляет собой долю тех результатов эксперимента, которые не превосходят данное текущее значение.

Эмпирическая функция распределения обладает всеми свойствами интегральной функции распределения.

Из определения эмпирической функции распределения следует, что:

1) значения эмпирической функции распределения $F_n^*(x)$ принадлежат отрезку $[0;\ 1]$;

2) $F_n^*(x)$ – неубывающая функция;

3) если $x_{\min}$ – наименьшее, а $x_{\max}$ – наибольшее наблюденное значение, то $F_n^*(x) = 0$ при $x \le x_{\min}$ и $F_n^*(x) = 1$ при $x > x_{\max}$.

Основное значение эмпирической функции распределения состоит в том, что она используется в качестве оценки функции распределения $F_{\mathrm{X}}(x) = \Pr(\mathrm{X} < x)$.

С увеличением объёма выборки ступеньки эмпирической функции распределения будут уменьшаться и в пределе, при $n \to \infty$, эта функция, согласно закону больших чисел, превратится в функцию распределения вероятностей значений СВ.

Пример

Построить эмпирическую функцию распределения по статистическому распределению случайной величины X.

Наблюденные значения случайной величины X	2	3	5
Частости, $p_i = m_i/n$	0,75	0,20	0,05

Решение

Относительная частота события $\{\mathrm{X} < x\}$ равна $F^*(x)$. Следовательно,

$$F_n^*(x) = \begin{cases} 0 & \text{при} \quad x \le 2; \\ 0,75 & \text{при} \quad 2 < x \le 3; \\ 0,95 & \text{при} \quad 3 < x \le 5; \\ 1 & \text{при} \quad x > 5. \end{cases}$$

График эмпирической функции распределения приведен на рис. 1.1.

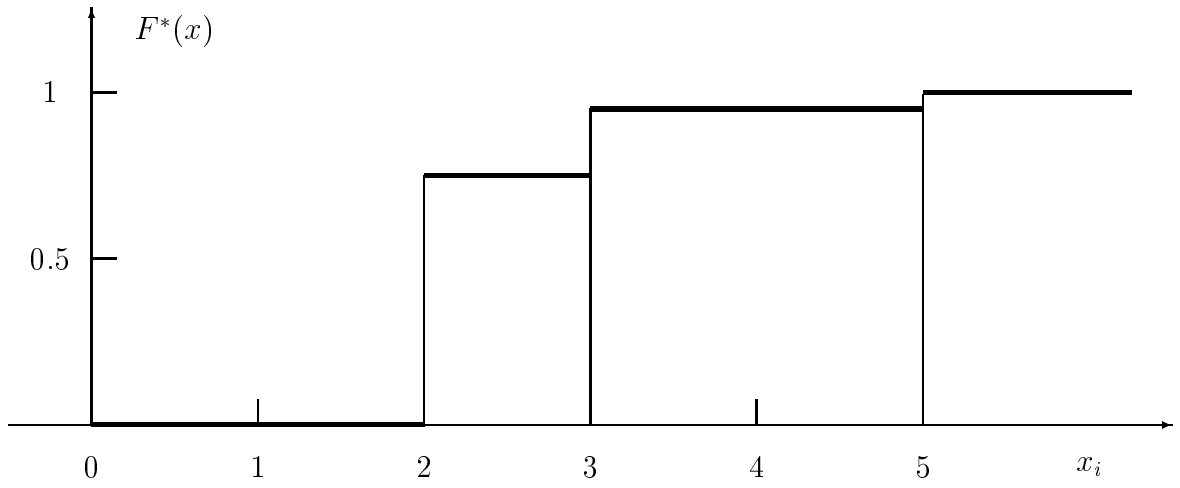


Рисунок 1.1 — Эмпирическая функция распределения

1.5. Графическое изображение статистических рядов

Для наглядности сгруппированные статистические ряды представляют графиками и диаграммами.

Наиболее распространенными являются *полигон, гистограмма, кумулята и огива*.

Полигон, кумулята и огива применяются для изображения как дискретных, так и интервальных статистических рядов, гистограмма – для изображения только интервальных рядов.

В результате построения полигона или гистограммы, можно получить первое представление о *форме распределения*, под которой подразумевается форма его графика в пределе бесконечной выборки, т. е. форма *кривой распределения*. Различают одновершинные (унимодальные) и многовершинные (многомодальные) распределения.

Для построения гистограммы относительных частот (частостей) на оси абсцисс откладываем частичные интервалы наблюденных значений случайной величины X, на каждом из которых строим прямоугольник, площадь которого равна частости данного частичного интервала.

Если на гистограмме частостей соединить середины верхних сторон элементарных прямоугольников, то полученная замкнутая ломаная образует *полигон* распределения частостей. Из принципа построения гистограммы и полигона распределения частостей следует, что площадь под гистограммой и полигоном частостей $S = 1$ (*единиц*^{2}).

Если в гистограмме вместо частот (частостей) записать соответственно накопленные частоты (частости), то получится *кумулянтный ряд*.

Для построения кумуляты на оси абсцисс откладываем наблюденные значения случайной величины X, на оси ординат – накопленные частости.

Накопленной частостью в точке x называется суммарная частость членов статистического ряда, значения которых меньше x, т.е. значения накопленных частостей являются значениями эмпирической функции распределения $F^*(x)$.

Если для построения кумуляты оси координат поменять местами, т.е. на горизонтальной оси откладывать значения эмпирической функции распределения $F^*(x)$, а на вертикальной – наблюденные значения случайной величины X, то полученная ломаная линия называется *огивой*.

Пример

Результаты исследования прочности 200 образцов бетона на сжатие представлены в виде интервального статистического ряда.

Интервалы прочности, *кг/см*^{2}	Частоты, m_i	Частости, m_i/n
190–200	10	0,05
200–210	26	0,13
210–220	56	0,28
220–230	64	0,32
230–240	30	0,15
240–250	14	0,07

При этом

$$n = \sum_{i=1}^{6} m_i = 200; \qquad \sum_{i=1}^{6} \frac{m_i}{n} = 1.$$

Требуется построить гистограмму, полигон распределения частостей и огиву данного статистического распределения.

Решение

Используем значения эмпирической функции распределения $F^*(x)$:

$$F^*(x) = \begin{cases} 0 & \text{при} \quad x \le 190; \\ 0,05 & \text{при} \quad 190 < x \le 200; \\ 0,18 & \text{при} \quad 200 < x \le 210; \\ 0,46 & \text{при} \quad 210 < x \le 220; \\ 0,78 & \text{при} \quad 220 < x \le 230; \\ 0,93 & \text{при} \quad 230 < x \le 240; \\ 1 & \text{при} \quad 240 < x \le 250. \\ 1 & \text{при} \quad 240 < x \le 250; \\ 1 & \text{при} \quad x > 250. \end{cases}$$

На рис. 1.2 изображена гистограмма частостей данного статистического ряда.

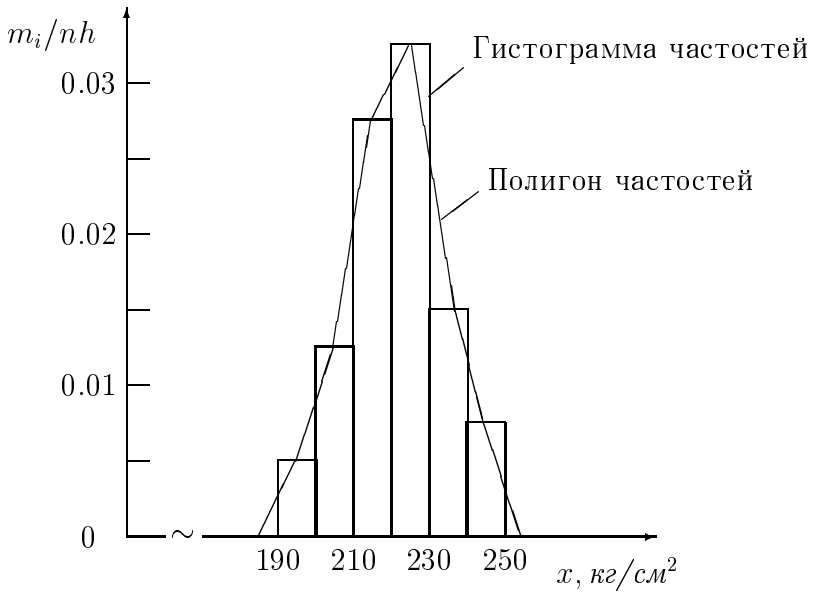


Рисунок 1.2 — Гистограмма и полигон распределения частостей

На рис. 1.3 изображены графики кумуляты и огивы данного интервального ряда.

В теории вероятностей гистограмме и полигону частостей соответствует график плотности распределения, соответственно, кумуляте отвечает график функции распределения $F_{\mathrm{X}}(x) = \Pr(\mathrm{X} < x)$.

Установлению и изучению форм кривых генеральных совокупностей по выборочным данным в математической статистике уделяется значительное внимание.

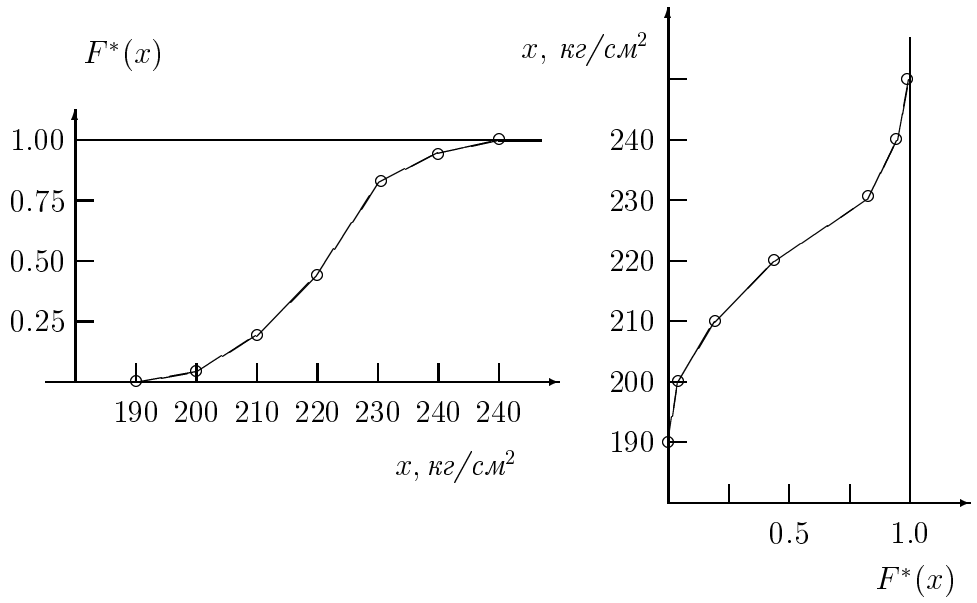


Рисунок 1.3 — Кумулята (слева) и огива (справа)

Характеристика рядов распределения предполагает выяснение условий, под влиянием которых сформировалось изучаемое распределение, выражение его основных особенностей, числовых характеристик.

1.6. Пример графической обработки выборочной информации

Измерения случайных величин, получаемые по ходу проведения эксперимента и регистрируемые в порядке их поступления, как правило, носят хаотичный характер и трудно обозримы. Поэтому в начале производится первичная обработка выборочной информации, которую продемонстрируем в следующем примере.

Предположим, что районный военный комиссар (военком) получил данные о контингенте очередного призыва, в частности, рост молодых людей.

Сведения о росте молодых людей (без указания фамилий) сведены в табл. 1.1.

Таблица 1.1 — Рост призывников в списочном порядке

Номер по списку	1	2	3	4	5	6	7	8
Рост, *см*	165	171	182	165	183	180	183	166
Номер по списку	9	10	11	12	13	14	15	
Рост, *см*	173	184	168	164	170	174	172	

Военком просит своего заместителя представить ему информацию о росте призывников в таком виде, чтобы он без расчетов и больших затрат времени смог бы ответить на следующие вопросы:

1) Какой вид имеет шеренга призывников, построенных по росту от меньшего к большему?

2) Какова величина роста призывников наиболее распространенная, сколько процентов призывников самого малого роста?

3) Сколько призывников следует определить в пехотные войска, какой процент это составляет от всего контингента, насколько плотно заполняется интервал роста, отводимого для пехотинцев?

4) Сколько комплектов обмундирования каждого роста следует заказать?

Решение

1. Ознакомившись с заданием, заместитель военкома для ответа на первый вопрос упорядочил сведения, построив *ранжированный ряд* распределения призывников – ряд, в котором все значения вариант (*вариантой* в статистике называют измеренное значение признака) располагаются по ранжиру (по порядку).

В результате переформирования исходных данных, содержащихся в табл. 1.1, получается табл. 1.2.

Таблица 1.2 — Ранжированный ряд распределения роста призывников

Номер по ранжиру	1	2	3	4	5	6	7	8
Рост, *см*	164	165	165	166	168	170	171	172
Номер по ранжиру	9	10	11	12	13	14	15	
Рост, *см*	173	174	180	182	183	183	184	

Информация, содержащаяся в ранжированном ряде распределения, обычно иллюстрируется графиком (рис. 1.4).

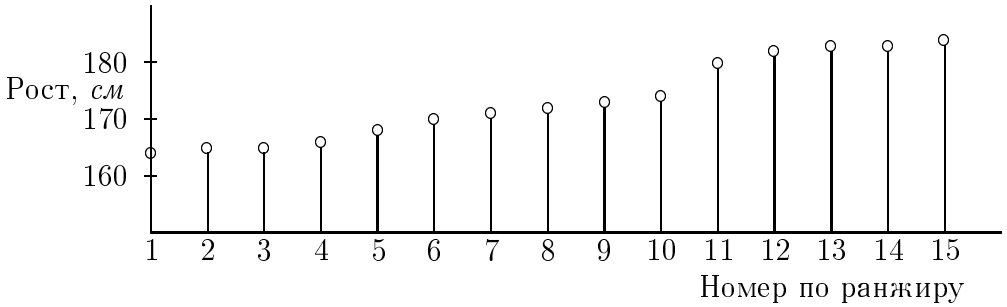


Рисунок 1.4 — График ранжированного ряда распределения роста призывников

2. Для ответа на второй вопрос заместитель военкома концентрирует информацию, содержащуюся в ранжированном ряде распределения, и строит зависимость частоты (числа объектов, обладающих одинаковым значением признака, т.е. числа объектов с одинаковым значением варианты) от роста. Кроме того, ему потребуется также значения *частостей* – отношений частот к общему числу объектов, т.е. объему выборки.

Все эти сведения сводятся в табл. 1.3. Для её построения обозначим значение варианты x_i, частоты – m_i, частости – p_i^*.

Зависимость частоты (или частости) от значений варьируемого признака называется *дискретным вариационным рядом распределения*.

Таблица 1.3 — Дискретный вариационный ряд распределения роста призывников

x_i	164	165	166	168	170	171	172
m_i	1	2	1	1	1	1	1
p_i^*	1/15	2/15	1/15	1/15	1/15	1/15	1/15
x_i	173	174	180	182	183	184	185
m_i	1	1	1	1	2	1	0
p_i^*	1/15	1/15	1/15	1/15	2/15	1/15	0

Тогда

$$p_i^* = m_i/n, \tag{1.5}$$

где n — объем выборки.

График, иллюстрирующий дискретный вариационный ряд, называется *полигоном распределения*. Он представлен на рис. 1.5.

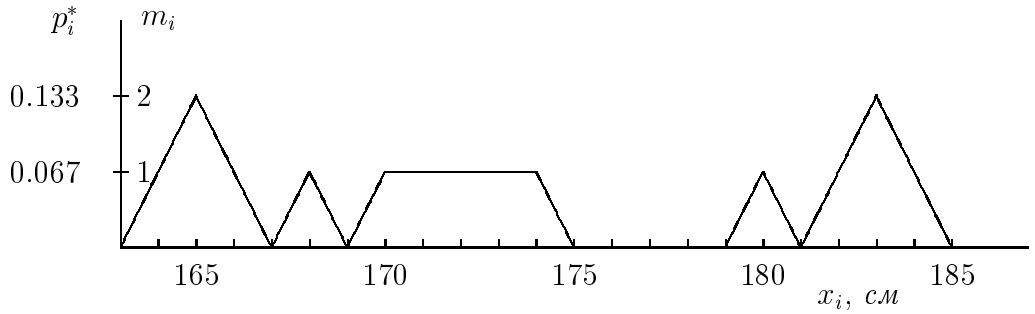


Рисунок 1.5 — Полигон распределения призывников по росту

Рассматривая этот график, военкому легко обнаружить, что самым распространенным ростом в данной выборке является рост 165 *см* и 183 *см*, при этом доля призывников (частость) самого малого роста (164 *см*) составляет 0,067, или 6,7 %.

3. Для того чтобы облегчить военкому решение третьего вопроса, его заместитель группирует призывников по их принадлежности к нормативам роста, принятым для различных родов войск, условно к группе низкого (до 170 *см* включительно), среднего (171–180 *см*) и высокого (свыше 180 *см*) роста.

В качестве характеристик распределения объектов по интервалам признака могут применяться частоты m_i (в единицах или штуках), частости p_i^* (в количестве объектов, приходящихся на единицу изменения признака).

Соответствующими аналогами в теории вероятностей являются число событий, вероятность события и плотность распределения вероятностей.

Зависимость перечисленных характеристик от интервалов значений признака называется *интервальным рядом распределения*, а её графическая интерпретация — *гистограммой распределения*.

Интервальный ряд и гистограмма распределения призывников для нашего примера представлены в табл. 1.4 и на рис. 1.6. Масштаб: m_i – 2 *человека на см*, p_i^* – 0,133 *единиц на см*, f_i^* – 0,02 $см^{-1}$ на *см*.

Табл. 1.4 и рис. 1.6 наглядно показывают, что в пехоту следует определить пятерых призывников, что составляет 33,3% (частость 0,333).

Таблица 1.4 — Интервальный ряд распределения роста призывников

Интервал, *см*	до 171	171–180	свыше 180
Частота, m_i	6	5	4
Частость, p_i^*	0,40	0,333	0,267
Плотность частостей, f_i^*, $см^{-1}$	0,0571	0,0333	0,0667

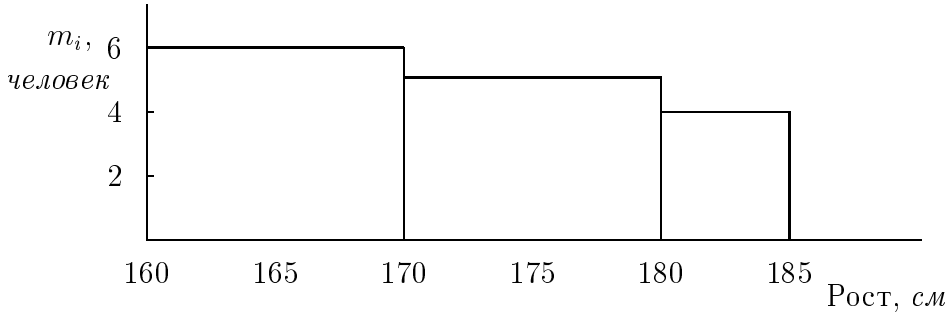


Рисунок 1.6 — Гистограмма распределения призывников по росту

Тот же самый приём можно применять и для решения последнего вопроса, только интервалы изменения признака (роста) следует выбрать по-другому. В применении к нашему примеру интервальный ряд распределения и гистограмма (условно) будут выглядеть так, как это представлено в табл. 1.5 и на рис. 1.7.

Таблица 1.5 — Интервальный ряд распределения роста призывников

Интервал, *см*	161–165	166–170	171–175	176–180	181–184
Частота, *чел.*	3	3	4	1	4

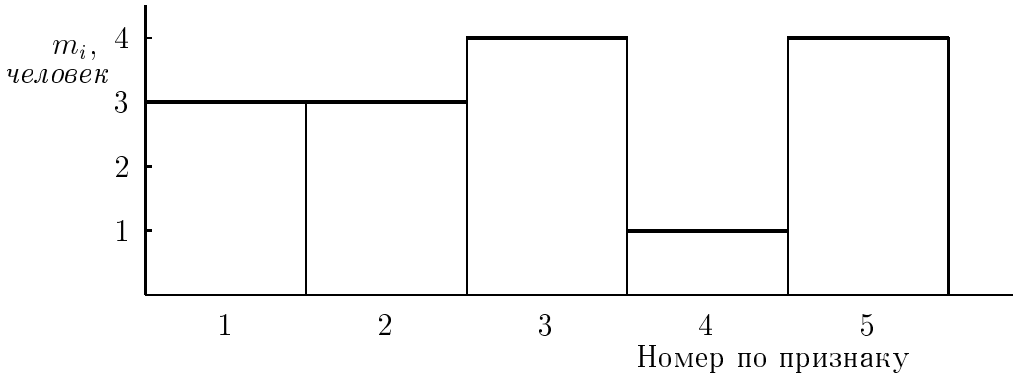


Рисунок 1.7 — Гистограмма распределения призывников по росту как признаку

4. С помощью табл. 1.5 и рис. 1.7 последнее задание военкома решается его заместителем автоматически — число комплектов обмундирования по росту составляет 3, 3, 4, 1 и 4.

Наряду с интервальным рядом и гистограммой употребляется кумулятивная (или эмпирическая) функция распределения $F^*(x)$ (табл. 1.6 и рис. 1.8).

Как и в теории вероятностей, где описание случайной величины рассматривается на нескольких смысловых уровнях, например на уровне плотности распределения вероятностей (или в дискретном варианте – на уровне дискретного ряда распределения) и функции распределения, в математической статистике приняты аналоги этих уровней.

Всесторонняя характеристика рядов распределения предполагает выяснение условий, под влиянием которых сформировалось изучаемое распределение, выражение его основных особенностей числовыми характеристиками.

Таблица 1.6 — Эмпирическая (кумулятивная) функция распределения призывников

Рост, *см*	163	164	165	166	167	168
Частость	0	1/15	3/15	4/15	4/15	5/15
Рост, *см*	169	170	171	172	173	174
Частость	5/15	6/15	7/15	8/15	9/15	10/15
Рост, *см*	175	176	177	178	179	180
Частость	10/15	10/15	10/15	10/15	10/15	11/15
Рост, *см*	181	182	183	184	185	
Частость	11/15	12/15	14/15	15/15	1	

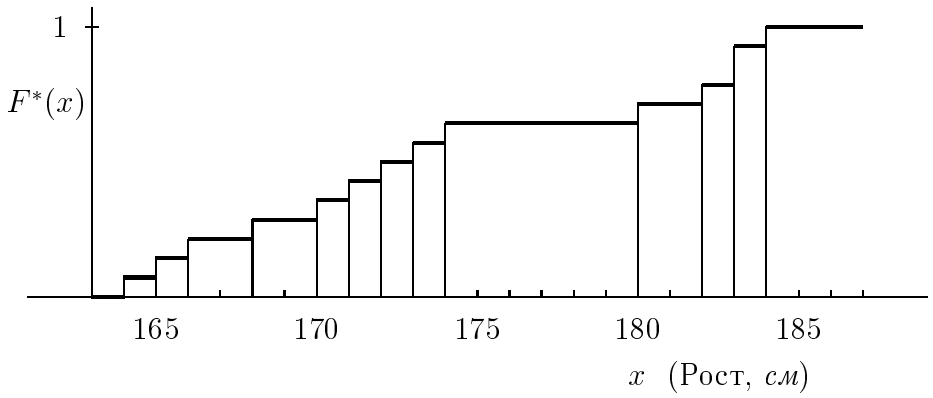


Рисунок 1.8 — Эмпирическая функция распределения призывников

1.7. Примеры

Пример 1.1

Для следующего распределения 45 пар мужской обуви, проданной в магазине за день, построить дискретный вариационный ряд.

40 41 42 40 40 39 40 41 39 41 40 42 41 40 42
41 42 41 40 38 44 39 42 41 42 39 41 37 43 41
44 40 42 41 39 41 42 43 42 41 43 43 38 38 43

Решение

Для построения вариационного ряда различные значения признака располагаем в порядке их возрастания. Окончательно вариационный ряд принимает вид:

Размер	37	38	39	40	41	42	43	44
Частота	1	3	5	8	12	9	5	2

Пример 1.2

Из партии деталей было отобрано 400, распределение которых по размеру дается следующей таблицей:

Размер детали, *мм*	7,95–8,00	8,00–8,05	8,05–8,10
Число деталей	12	28	132
Размер детали, *мм*	8,10–8,15	8,15–8,20	8,20–8,25
Число деталей	150	62	16

Найти ошибку выборки μ при определении среднего.

Решение

Сначала находим выборочную дисперсию σ^{*2}: $\sigma^{*2} = 0,00272$. Это дает для ошибки выборки при определении среднего

$$\mu = \sigma_0/\sqrt{n-1} = \sqrt{0,00272/(400-1)} = 0,002611.$$

Так как $\sqrt{0,00272/400} = 0{,}002608$, то при таких больших объемах выборки n можно ошибку выборки μ находить практически по формуле $s = \sigma_0/\sqrt{n}$.

Пример 1.3

Представить выборку из n наблюдений в виде таблицы частот, используя 7 интервалов группировки. Выборка:

20,3	15,4	17,2	19,2	23,3	18,1	21,9	15,3	16,8	13,2	20,4
19,1	21,3	16,5	19,7	20,5	14,0	20,1	16,8	14,7	20,8	19,5
19,3	17,8	11,8	17,8	17,8	16,2	15,7	22,8	21,9	12,5	10,1
18,3	14,7	14,5	17,8	13,5	19,6	18,1	18,4	13,9	19,1	18,5
23,6	16,7	20,4	19,5	17,2	17,5	19,4	18,6	15,3	21,1	20,2

Решение

Размах выборки $w = 23,6 - 10,1 = 13,5$. Длина интервала группировки $h = 13,5/7 \approx 2$. В качестве первого интервала удобно взять 10–12 и так далее для следующих интервалов.

Обозначим:

i — номер интервала;

Ω_i – границы i-го интервала;

x_i – середины интервалов;

n_i – частоты;

$m_i = \sum_{j=1}^{i} n_j$ – накопленные частоты;

$p_i = n_i/n$ – относительные частоты (частости);

$F_i = \sum_{j=1}^{i} p_j$ – накопленные частости.

В результате получаем следующую таблицу:

i	Ω_i	x_i	n_i	$quadm_i$	p_i	F_i
1	10–12	11	2	2	0,0364	0,0364
2	12–14	13	4	6	0,0727	0,1091
3	14–16	15	8	14	0,1455	0,2545
4	16–18	17	12	26	0,2182	0,4727
5	18–20	19	15	41	0,2727	0,7455
6	20–22	21	11	52	0,2000	0,9455
7	22–24	23	3	55	0,0545	1,0000

Пример 1.4

Для определения крепости нити проведены испытания $n = 1000$ проб, давшие следующие результаты:

Крепость нити, $г$	180–190	190–200	200–210	210–220
Число проб, n_i	50	90	150	280
Крепость нити, $г$	220–230	230–240	240–250	
Число проб, n_i	220	120	90	

Построить кумулянтный ряд.

Решение

Сначала находим накопленные частоты для каждого из интервалов данного интервального вариационного ряда:

$$m_1 = n_1 = 50; \quad m_2 = n_1 + n_2 = 50 + 90 = 140;$$
$$m_3 = 290; \quad m_4 = 570; \quad m_5 = 790; \quad m_6 = 910; \quad m_7 = 1000.$$

Таким образом, после нормировки на $n = 1000$ кумулянтный ряд для данного распределения имеет вид:

Крепость нити, $г$	180–190	190–200	200–210	210–220
Число проб, n_i	50	90	150	280
$F_i = m_i/n$	0,050	0,140	0,290	0,570
Крепость нити, $г$	220–230	230–240	240–250	
Число проб, n_i	220	120	90	
$F_i = m_i/n$	0,790	0,910	1,000	

Пример 1.5

Записать в виде вариационного и статистического рядов выборку

$$\{5 \;\; 3 \;\; 7 \;\; 10 \;\; 5 \;\; 5 \;\; 2 \;\; 10 \;\; 7 \;\; 2 \;\; 7 \;\; 7 \;\; 4 \;\; 2 \;\; 4\}$$

Составить статистический ряд и определить размах выборки.

Решение

Объем выборки $n = 15$. Упорядочивая элементы выборки по величине, получаем вариационный (ранжированный) ряд

$$\{2 \;\; 2 \;\; 2 \;\; 3 \;\; 4 \;\; 4 \;\; 5 \;\; 5 \;\; 5 \;\; 7 \;\; 7 \;\; 7 \;\; 7 \;\; 10 \;\; 10\}$$

Размах выборки составляет $W = 10 - 2 = 8$.

Различными в заданной выборке являются элементы $x_1 = 2$, $x_2 = 3$, $x_3 = 4$, $x_4 = 5$, $x_5 = 7$, $x_6 = 10$.

Их частоты соответственно равны $n_1 = 3$, $n_2 = 1$, $n_3 = 2$, $n_4 = 3$, $n_5 = 4$, $n_6 = 2$.

Следовательно, статистический ряд исходной выборки можно записать в виде следующей таблицы:

x_i	2	3	4	5	7	10
n_i	3	1	2	3	4	2

Для контроля правильности находим $\sum_i n_i = 15$.

Пример 1.6

С целью определения средней суммы вкладов в сберегательной кассе, имеющей $N = 2200$ вкладчиков, произведено выборочное обследование (бесповторный отбор) $n = 111$ вкладов, которое дало следующие результаты:

Сумма вклада	10–30	30–50	50–70	70–90	90–110	110–130
Число вкладов	1	3	10	30	60	7

Пользуясь этими данными, найти доверительные границы для генерального среднего, которое можно было бы гарантировать с вероятностью $p = 0,96$.

Решение

Шаг интервала составляет $h = 20$. Обозначим середину i-го интервала через x_i, $i = 1, \ldots, 6$. Для упрощения расчетов введем вспомогательную случайную величину

$$\mathrm{U} = (\mathrm{X} - x_4)/h = (\mathrm{X} - 80)/20.$$

Сначала найдем выборочное среднее и выборочную дисперсию, для чего сведем вычисления в таблицу:

i	Границы интервала	x_i	m_i	$u_i = (x_i - 80)/20$	$m_i u_i$	$m_i u_i^2$
1	10–30	20	1	–3	–3	9
2	30–50	40	3	–2	–6	12
3	50–70	60	10	–1	–10	10
4	70–90	80	30	0	0	0
5	90–110	100	60	1	60	60
6	110–130	120	7	2	14	28
	Сумма		111		55	119

Из таблицы имеем

$$\overline{\mathrm{U}} = \frac{\sum_i m_i u_i}{\sum_i m_i} = \frac{55}{111} = 0,4955;$$

$$\overline{\mathrm{U}^2} = \frac{\sum_i m_i u_i^2}{\sum_i m_i} = \frac{119}{111} = 1,0721.$$

Отсюда

$$\sigma_{\mathrm{U}}^2 = \overline{\mathrm{U}^2} - \left(\overline{\mathrm{U}}\right)^2 = 1,0721 - 0,2455 = 0,8266,$$

что дает

$$\overline{\mathrm{X}} \;=\; 80 + 20 \cdot 0,4955 \;=\; 89,91; \qquad \sigma_{\mathrm{x}}^2 \;=\; 20^2 \cdot 0,8266 \;=\; 330,64.$$

Поэтому ошибка определения среднего μ составляет

$$\mu \;\approx\; \sqrt{\frac{\sigma_{\mathrm{x}}^2}{n-1}\left(1-\frac{n}{N}\right)} \;=\; \sqrt{\frac{330,64}{111-1}\left(1-\frac{111}{2200}\right)} \;=\; 1,69.$$

Так как $\Phi(t) = 0{,}96$ при $t = 2{,}05$ (см. таблицы функции Лапласа), то предельная погрешность $\varepsilon = t\mu = 2{,}05 \cdot 1{,}69 = 3{,}46$. Итак, с вероятностью $p = 0,96$ доверительными границами для генерального среднего будут

$$\overline{\mathrm{X}} - \varepsilon = 89,91 - 3,46 = 86,45 \quad \text{и} \quad \overline{\mathrm{X}} + \varepsilon = 89,91 + 3,46 = 93,37.$$

Пример 1.7

На каждую сотню изготовленных деталей в среднем бывают две, не удовлетворяющих стандарту (брак). Было проверено 10 партий по 100 изделий в каждой. Отклонения числа обнаруженных бракованных изделий от среднего приведены в таблице:

Номер партии	1	2	3	4	5	6	7	8	9	10
Отклонение от среднего	–1	0	1	1	–1	1	0	–2	2	1

Построить вариационный ряд. Найти выборочное среднее уклонения числа бракованных изделий от установленного и его выборочную дисперсию.

Решение

Выборочный ряд {–2, –1, –1, 0, 0, 1, 1, 1, 1, 2} содержит пять различных значений: {–2, –1, 0, 1, 2}. Их частости соответственно равны {0,1; 0,2; 0,2; 0,4; 0,1}.

Таблица, представляющая свойства выборочной случайной величины, следующая:

Номер, i	1	2	3	4	5
Значение, x_i	–2	–1	0	1	2
Частость, f_i	0,1	0,2	0,2	0,4	0,1

Выборочное среднее составляет $\overline{\mathrm{X}} = 0,2$.

Выборочная дисперсия равна $\sigma^{*2} = 1,36$.

Пример 1.8

Вычислить среднее $\overline{\mathrm{X}}$ и дисперсию $\sigma_{\mathrm{x}}^{*\,2}$ группированной выборки

Номер интервала i	1	2	3
Граница интервала Ω_i	134–138	138–142	142–146
Частота n_i	1	3	15
Номер интервала i	4	5	6
Граница интервала Ω_i	146–150	150–154	154–158
Частота n_i	18	14	2

Решение

У случайной величины X, заданной выборкой, длина каждого из интервалов группировки составляет $h = 4$, а значение середины интервала, встречающегося с наибольшей частотой, равно $d^* = 148$. Обозначим набор $\{z_i\}$ – значения середин i-х интервалов и преобразуем группированную выборку следующим образом:

$$u_i = \frac{x_i - d^*}{h} = \frac{x_i - 148}{4}, \qquad i = 1, 2, \ldots, 6,$$

т.е. используем приведенную случайную величину U.

Вычисления сведем в таблицу:

i	$quadz_i$	u_i	n_i	$n_i u_i$	$n_i u_i^2$
1	136	–3	1	–3	9
2	140	–2	3	–6	12
3	144	–1	15	–15	15
4	148	0	18	0	0
5	152	1	14	14	14
6	156	2	2	4	8
Сумма	–	–	53	–6	58

Теперь находим

$$\overline{\mathrm{X}} = h\overline{\mathrm{U}} + d^* = 4 \cdot \frac{-6}{53} + 148 \approx 147,55,$$

$$\sigma^{*\,2}_{\mathrm{X}} = h^2 \sigma^{*\,2}_{\mathrm{U}} = 4^2 \cdot \frac{58 - (-6)^2/53}{53} \approx 17,30.$$

Пример 1.9

Для проверки работы программы построения гистограммы был обработан массив $\{x_n\}$, $n = 1, 2, \ldots, N$, где объем выборки $N = 10000$. Значения аргумента x были получены по правилу $x_n = n\pi/N$, а значения функции $y(x)$ – по правилу $y_n = \sin(x_n)$. Таким образом, исходный массив представляет собой выборку значений *неслучайной функции*.

Построить гистограмму массива $\{y_n\}$, выбрав число каналов анализа $M = 24$.

Решение

Значения функции ограничены, очевидно, интервалом $0 \le y \le 1$. Поэтому в этом интервале и будем задавать каналы анализа гистограммы шириной $h = 1/M$ каждый.

Гистограмма приведена на рис 1.9. Такой её вид соответствует графику плотности некоторой случайной величины, подчиняющейся закону арксинуса.

Замечание. Если бы переменная x была *случайной величиной*, равномерно распределенной на интервале $(0 \le x \le \pi)$, то тип гистограммы был бы таким же.

Пример 1.10

Для определения среднего процента сырого белка в зернах пшеницы было отобрано 626 зерен, обследование которых показало, что выборочное среднее равно 16,8, а выборочная дисперсия составила 4.

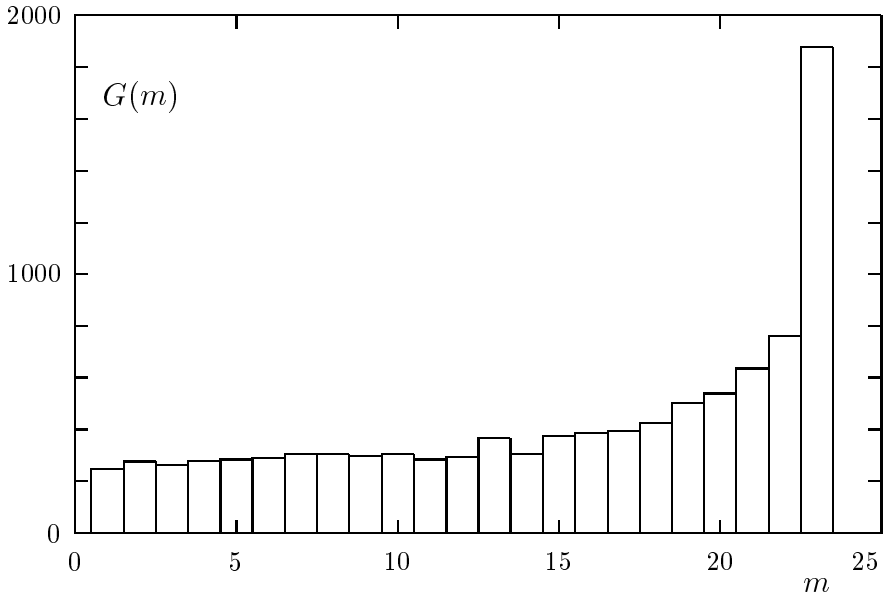


Рисунок 1.9 — Гистограмма $G(m)$ функции $y = \sin(x)$ (неслучайная величина x равномерно заполняет интервал $(0;\, \pi/2)$; объем выборки $N = 10000$; число каналов анализа гистограммы $M = 24$; значения ограничены величиной $\max y = 1$)

Чему равна вероятность p того, что средний процент сырого белка отличается от 16,8 по абсолютной величине меньше чем на 0,2%?

Решение

Объем генеральной совокупности неизвестен. Поэтому ошибку выборки находим по формуле $\mu \approx \sigma/\sqrt{n-1}$.

В нашем случае $n = 626$, $\sigma^2 = 4$ и, следовательно,

$$\mu \;\approx\; 2/\sqrt{625} \;=\; 2/25 \;=\; 0,08.$$

Так как $\Pr(|\overline{\mathrm{X}} - \mathrm{X}| < \varepsilon) \;\approx\; 2\Phi(\varepsilon/\mu)$, то искомая вероятность

$$p \;=\; \Pr(|\overline{\mathrm{X}} - 16,8| < 0,2) \;\approx\; 2\Phi\,(0,2/0,08) \;=\; 2\Phi(2,5) \;=\; 0,98758.$$

Пример 1.11

Значения n независимых случайных величин $\{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n\}$, имеющих одну и ту же функцию распределения $G(x)$, расположены в порядке возрастания. Пусть ν и μ – заданные целые числа $(1 \le \nu,\, \mu \le n)$.

Найти распределение ν-го значения снизу U и μ-го значения сверху V в этом ранжированном ряде.

Решение

Для того, чтобы величина U попала в интервал $[u,\, u + du]$, необходимо и достаточно, чтобы какие-нибудь $\nu - 1$ из n величин $\{\mathrm{X}_1,\, \mathrm{X}_2,\, \ldots,\, \mathrm{X}_n\}$ приняли значения, меньшие чем u, одно значение попало в интервал $[u, u + du]$, а остальные $n - \nu$ приняли значения, не меньшие чем $u + du$.

Таким образом, мы приходим к схеме повторения опытов с тремя несовместными событиями, образующими полную группу:

$$A_1 = \{\mathrm{X} < u\}, \qquad A_2 = \{\mathrm{X} \geq u + du\}, \qquad A_3 = \{u \leq \mathrm{X} < u + du\}.$$

Если обозначить плотность распределения $g(u) = dG(u)/du$, то вероятности приведенных событий при одном опыте равны соответственно (с точностью до бесконечно малых высших порядков)

$$\Pr(A_1) = G(u), \qquad \Pr(A_2) = 1 - G(u), \qquad \Pr(A_3) = g(u)du.$$

Тогда получим

$$f_{\mathrm{U}}(u) \;=\; \frac{n\,!}{(\nu-1)\,!(n-\nu)}\, G^{\nu-1}(u)\,[1-G(u)]^{n-\nu}\, g(u)$$

– эта формула определяет плотность распределения ν-го значения ряда.

Аналогично находим плотность μ-го значения V

$$f_{\mathrm{V}}(v) \;=\; \frac{n\,!}{(\mu-1)\,!(n-\mu)}\, G^{\mu-1}(v)\,[1-G(v)]^{n-\mu}\, g(v).$$

Для ранжированного ряда рассмотрим случайные величины:

наименьшее значение ряда $\mathrm{U} = \min\limits_{1\leq i\leq n}\{\mathrm{X}_i\}$;

наибольшее значение ряда $\mathrm{V} = \max\limits_{1\leq i\leq n}\{\mathrm{X}_i\}$.

При $\nu = 1$ и $\mu = 1$ получаем для плотности распределения наименьшего U и наибольшего V значение рассматриваемого ряда

$$f_{\mathrm{U}}(u) \;=\; n[1-G(u)]^{n-1}\, g(u), \qquad f_{\mathrm{V}}(v) \;=\; nG^{n-1}(u)\, g(v).$$

Приведем также выражение для плотности распределения разности между ν-м значением сверху и ν-м значением снизу R = V - U

$$f_{\mathrm{R}}(r) \;=\; \frac{n!}{[(\nu-1)!]^2\,(n-2\nu)!}\times$$

$$\times \int\limits_{-\infty}^{\infty} g(x)g(r+x)\; G^{\nu-1}(x)[G(r+x)-G(x)]^{n-2\nu}\,[1-G(r+x)]^{\nu-1}\,dx.$$

В частном случае при $\nu = 1$ отсюда получаем плотность распределения *широты разброса*

$$\mathrm{S} \;=\; \max_{1\leq i\leq n}\{\mathrm{X}_i\} - \min_{1\leq i\leq n}\{\mathrm{X}_i\}$$

в выборке из n независимых случайных величин (*размах выборки*)

$$f_{\mathrm{S}}(s) \;=\; n(n-1)\int\limits_{-\infty}^{\infty}[G(x+s)-G(x)]^{n-2}g(x)g(s+x)\,dx.$$

1.8. Задачи для решения

Задача 1.1

Испытывалась чувствительность 40 приемников. Данные приведены в таблице, в которой в первой строке даны интервалы чувствительности в микровольтах, во второй – средние точки этих интервалов $f_{\text{ср}}$, в третьей – число приемников n_i, чувствительность которых оказалась в этом интервале.

Интервал	25–75	75–125	125–175	175–225	225–275
$f_{\text{ср}}$	50	100	150	200	250
n_i	0	0	1	5	8
Интервал	275–325	325–375	375–425	425–475	475–525
$f_{\text{ср}}$	300	350	400	450	500
n_i	6	8	6	2	2
Интервал	525–575	575–625	625–675	675–725	725–775
$f_{\text{ср}}$	550	600	650	700	750
n_i	0	1	1	0	0

Построить эмпирическую функцию распределения и гистограмму выборки, найти среднюю чувствительность приемников из этой партии и её средневыборочное уклонение.

Задача 1.2

Наблюдения за толщиной (в *мм*) медных образцов дали следующие результаты:

0,031	0,040	0,049	0,041	0,052	0,044	0,046	0,040	0,038	0,040
0,043	0,034	0,041	0,050	0,041	0,043	0,041	0,037	0,041	0,055
0,041	0,044	0,047	0,040	0,058	0,040	0,038	0,040	0,043	0,056
0,053	0,040	0,043	0,038	0,041	0,037	0,041	0,046	0,061	0,044
0,041	0,046	0,044	0,047	0,038	0,040	0,049	0,041	0,042	0,037

Построить кумулянтный ряд, начертить кумуляту и огиву статистического распределения приведенных данных.

Задача 1.3

Обследование показало следующее распределение роста группы юношей:

Рост, *см*	Число юношей
143–146	1
146–149	2
149–152	8
152–155	26
155–158	65
158–161	120
161–164	181
164–167	201

Рост, *см*	Число юношей
167–170	170
170–173	120
173–176	64
176–179	28
179–182	10
182–185	3
185–188	1
188–191	1

Построить кумулянтный ряд, начертить кумуляту и огиву.

Задача 1.4

Средняя температура воздуха в сентябре в двух городах (X) и (Y) измерялась в течение 40 лет. Данные приведены в таблице.

X	Y	X	Y	X	Y	X	Y	X	Y
12,0	10,8	13,9	10,1	14,9	13,0	16,0	16,0	18,0	14,0
12,0	11,3	14,2	10,0	14,9	14,2	16,9	12,9	18,0	14,9
12,0	12,0	14,0	10,0	15,1	13,8	17,2	13,9	18,1	16,0
12,0	13,0	14,0	12,0	15,0	16,0	16,9	14,8	18,4	17,8
12,8	10,9	13,9	12,4	15,5	13,9	16,9	15,0	19,2	15,0
13,8	10,0	15,0	11,0	15,9	14,7	17,0	16,0	19,3	16,1
13,1	13,0	14,0	14,8	16,0	13,0	16,8	17,0	20,0	17,0
13,0	13,0	14,0	15,2	15,9	15,0	17,5	16,0	20,1	17,7

Найти выборочные среднемесячные температуры в обоих населенных пунктах и их среднеквадратичные уклонения.

Задача 1.5

Группа абитуриентов получили следующие баллы на приёмных экзаменах:

20	19	22	24	21	18	23	17	20	16	15	23	20	21	24	21
24	21	20	18	19	17	22	20	16	22	18	20	17	21	17	19
22	23	21	25	22	20	19	21	24	22	23	21	29	22	21	19
21	19	20	20	21	18	22	25	21	18	23	20	16	20	23	22

Найти моду, медиану и среднее арифметическое этого распределения.

Задача 1.6

Обследование показало следующее распределение роста группы девушек:

Рост, *см*	Число девушек	Рост, *см*	Число девушек
133–136	1	157–160	150
136–139	2	160–163	109
139–142	18	163–166	54
142–145	36	166–169	18
145–148	85	169–172	8
148–151	140	172–175	1
151–154	211	175–178	0
154–157	181	178–181	0

Найти моду, медиану и среднее арифметическое этого распределения.

Задача 1.7

Случайная величина X равномерно распределена на интервале $[0, 2\pi]$, а случайная величина Y связана с X соотношением $Y = \mathrm{tg}(X)$. Получена выборка объемом n значений случайной величины Y.

Привести дифференциальный и интегральный законы распределения случайной величины Y.

Построить гистограмму и кумуляту выборки.

Задача 1.8

На молочной ферме зарегистрировали сведения о величине удоя коров:

Удой, *кг*	Количество коров
400–600	1
600–800	3
800–1000	6
1000–1200	11
1200–1400	15
1400–1600	20

Удой, *кг*	Количество коров
1600–1800	14
1800–2000	12
2000–2200	10
2200–2400	6
2400–2600	2
2600 и выше	2

Найти дисперсию, коэффициент вариации и размах вариации распределения.

Задача 1.9

Наблюдения за толщиной (в *мм*) слюдяных образцов дали результаты:

0,030	0,039	0,021	0,030	0,039	0,031	0,042	0,034	0,036	0,030	0,028
0,031	0,042	0,030	0,033	0,024	0,031	0,040	0,031	0,033	0,031	0,027
0,037	0,031	0,031	0,045	0,031	0,034	0,027	0,030	0,048	0,030	0,028
0,036	0,028	0,030	0,033	0,046	0,043	0,030	0,033	0,028	0,031	0,027
0,026	0,031	0,036	0,051	0,034	0,031	0,036	0,034	0,037	0,028	0,030

Построить по этим данным интервальный вариационный ряд с равными интервалами (первый интервал 0,020–0,024, второй 0,024–0,028 и т.д.) и начертить гистограмму.

Задача 1.10

Через каждый час измерялось напряжение тока в электросети. При этом были получены следующие значения (в V):

227	219	215	230	232	223	220	222	218	219	222
221	227	226	226	209	211	215	218	220	216	220
220	221	225	224	212	217	219	220	220	222	218
218	219	222	218	225	224	212	217	219	220	220

Построить статистическое распределение и начертить полигон.

Задача 1.11

Случайная величина X равномерно распределена на интервале $(0, 2\pi)$, а случайная величина Y связана с X соотношением $\mathrm{Y} = \sin(\mathrm{X})$. Получена выборка объемом n значений случайной величины Y.

Привести дифференциальный и интегральный законы распределения случайной величины Y. Построить гистограмму и кумуляту выборки.

Задача 1.12

Построить дискретный вариационный ряд и начертить полигон распределения группы абитуриентов по числу баллов, полученных ими на приёмных экзаменах:

20	19	22	24	21	18	23	17	20	16	15	23	21	24	21
24	21	20	18	17	22	20	16	22	18	20	17	21	17	19
20	19	21	24	22	23	21	23	22	21	19	20	23	22	25
22	23	21	25	22	20	16	20	21	18	18	23	21	19	20

Задача 1.13

Данные об урожайности ржи на различных участках поля приведены в следующей таблице:

Урожайность, *ц/га*	9–12	12–15	15–18	18–21	21–24	24–27
Доля участка от общей посевной площади, %	6	12	33	22	19	8

Построить кумулянтный ряд, начертить кумуляту и огиву.

Задача 1.14

Наблюдения за процентом жира в молоке 50 коров дали результаты:

3,86	4,06	3,67	3,97	3,76	3,61	3,96	4,04	3,84	3,94
3,98	3,57	3,87	4,07	3,99	3,69	3,76	3,71	3,94	3,82
4,16	3,76	4,00	3,46	4,08	3,88	4,01	3,93	3,71	3,81
4,02	4,17	3,72	4,09	3,78	4,02	3,73	3,52	3,89	3,92
4,18	4,26	4,03	4,14	3,72	4,33	3,82	4,03	3,62	3,91

Построить по этим данным интервальный вариационный ряд с равными интервалами (первый интервал 3,45–3,55 %, второй интервал 3,55–3,65 % и т.д.) и изобразить его графически.

1.9. Задание на практическую работу

Настоящая практическая работа рассчитана на два часа и содержит одно задание. Задание должно выполняться в выбранной программной среде.

Задание 1

Напишите программу, которая создает выборку объемом N случайных величин, подчиняющихся заданному закону распределения. В каждом из законов распределения предусмотрите возможность варьирования параметров. В случае использования математических пакетов пользоваться встроенными функциями можно лишь для сравнения. Результат работы программы – массив, содержащий значения искомой выборки. Необходимо предусмотреть визуализацию данных (построить программно гистограмму и кумуляту). Результаты оформите графически.

Вариант 1

Равномерное распределение.

Исходные данные для программы:

n_1 – левая граница возможных значений;

n_2 – правая граница возможных значений;

N – объем выборки;

x_n – последовательность выборочных значений, $1 \leq n \leq N$;

M – число каналов анализа гистограммы и кумуляты.

Вариант 2

Нормальный закон Гаусса.

Исходные данные для программы :

m_{x} – математической ожидание;

σ^2_{x} – дисперсия;

N – объем выборки;

x_n – последовательность выборочных значений, $1 \leq n \leq N$;

M – число каналов анализа гистограммы и кумуляты.

Вариант 3

Распределение Пуассона.

Исходные данные для программы :

m_{x} – математической ожидание;

N – объем выборки;

x_n – последовательность выборочных значений, $1 \leq n \leq N$;

M – число каналов анализа гистограммы и кумуляты.

1.10. Задания для проверки

1. Какие задачи рассматривает математическая статистика?

2. Что называется генеральной совокупностью, выборочной совокупностью?

3. В чем заключается сущность выборочного метода?

4. Что называется статистическим законом распределения случайной величины?

5. Что называется эмпирической функцией распределения случайной величины?

6. В чем состоит различие между эмпирической функцией распределения и теоретической (интегральной) функцией распределения случайной величины?

7. Какие свойства имеет эмпирическая функция распределения случайной величины?

8. Перечислите основные виды графиков, служащих для изображения статистических рядов.

2. Специальные законы распределения математической статистики

Наиболее часто встречающиеся распределения основных статистик, вычисляемых по выборке из нормально распределенной генеральной совокупности, связаны с нормальным законом Гаусса, гамма-распределением, распределением хи-квадрат $f(\chi^2)$, распределением Стьюдента и распределением Фишера-Снедекора. При их рассмотрении мы также встретимся с гаммой-функцией Эйлера $\Gamma(\alpha)$. В задачах математической статистики также часто используются распределения Колмогорова и Колмогорова-Смирнова $K(\lambda)$, распределения Бартлетта, Крамера и другие.

2.1. Нормальный закон Гаусса

Случайная величина X имеет *нормальное распределение* с параметрами m и σ^2, если её плотность распределения следующая:

$$f_{\mathrm{X}}(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \tag{2.1}$$

областью определения X является вся числовая ось, $-\infty < x < \infty$.

Про такую случайную величину ещё говорят, что *она распределена по закону Гаусса*. Эта случайная величина наиболее широко применяется при построении статистических моделей.

На практике для краткости нормальную случайную величину X с параметрами m и σ часто обозначают как $\mathcal{N}(m;\sigma)$.

Кривая распределения, описывающая плотность (2.1), симметрична относительно точки m, в которой плотность достигает максимума. Из этой симметрии непосредственно вытекает, что математическое ожидание случайной величины X равно

$$\mathrm{M}[\mathrm{X}] \;=\; m. \tag{2.2}$$

С изменением значения математического ожидания кривая, как целое, смещается с сторону изменения, как это показано на рис. 2.1–2.2. В точке $x = m$ достигается максимум, равный $\left(\sqrt{2\pi}\sigma\right)^{-1} = 0,3989/\sigma$.

Дисперсия нормальной случайной величины X равна

$$\mathrm{D}[\mathrm{X}] \;=\; \sigma^2. \tag{2.3}$$

Таким образом, параметр σ можно интерпретировать как меру рассеяния случайной величины X вокруг своего математического ожидания (см. рис. 2.3–2.4).

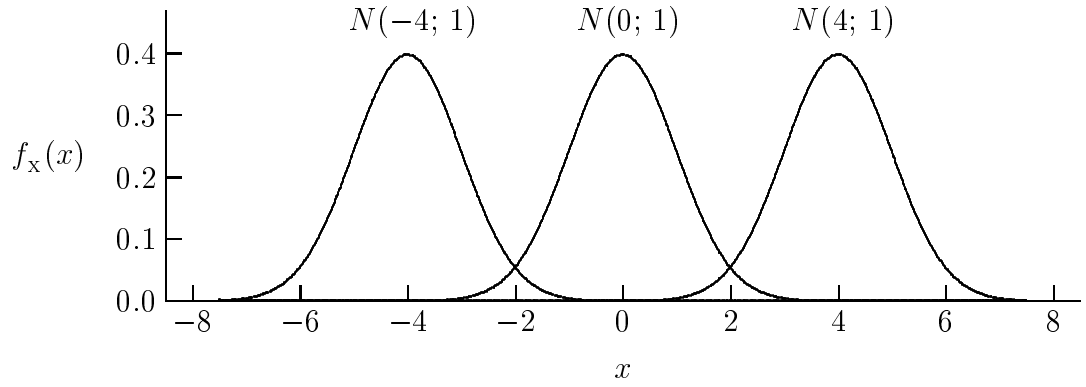


Рисунок 2.1 — Плотность $f_{\mathrm{X}}(x)$ нормального распределения Гаусса для случаев $N(-4;\,1)$, $N(0;\,1)$ и $N(4;\,1)$ с параметрами $\sigma = 1$ и $m = -4$, 0 и 4 соответственно

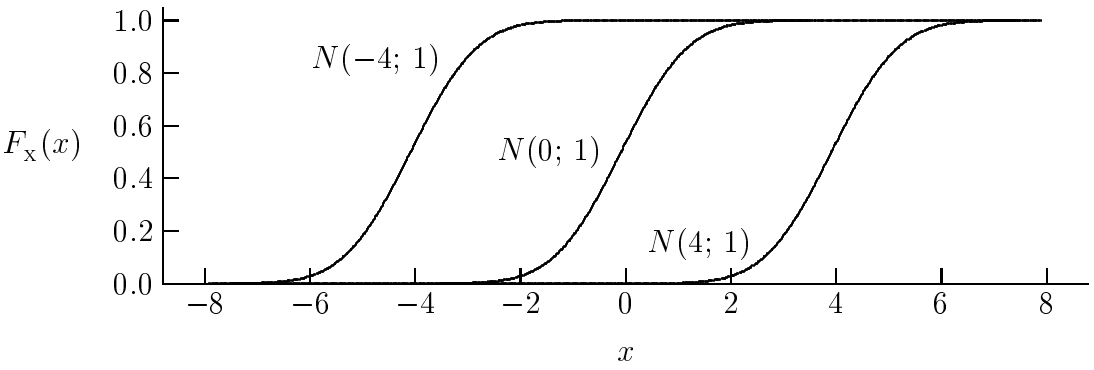


Рисунок 2.2 — Интегральный закон $F_{\mathrm{X}}(x)$ распределения Гаусса для случаев $N(-4;\,1)$, $N(0;\,1)$ и $N(4;\,1)$ с параметрами $\sigma = 1$ и $m = -4$, 0 и 4 соответственно

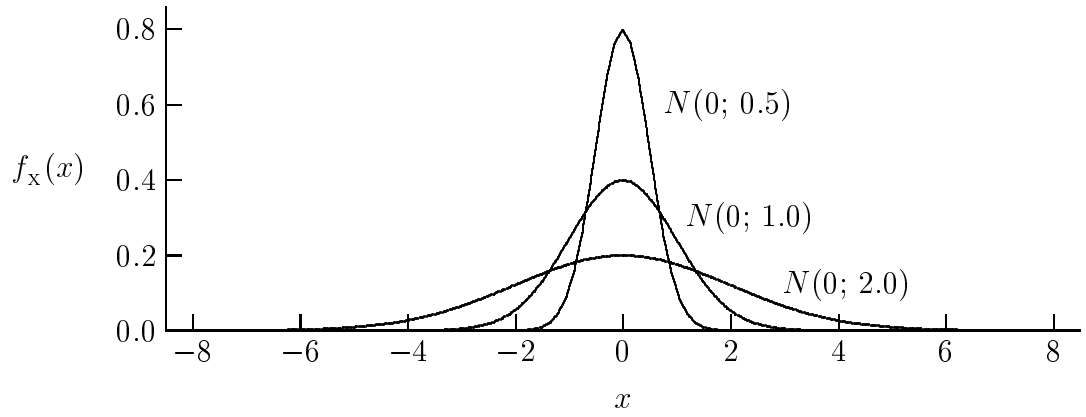


Рисунок 2.3 — Плотность нормального распределения $f_{\mathrm{X}}(x)$ Гаусса для случаев $N(0;\,0,5)$, $N(0;\,1,0)$, $N(0;\,2,0)$ с параметрами $m = 0$ и $\sigma = 0{,}5$, 1,0 и 2,0

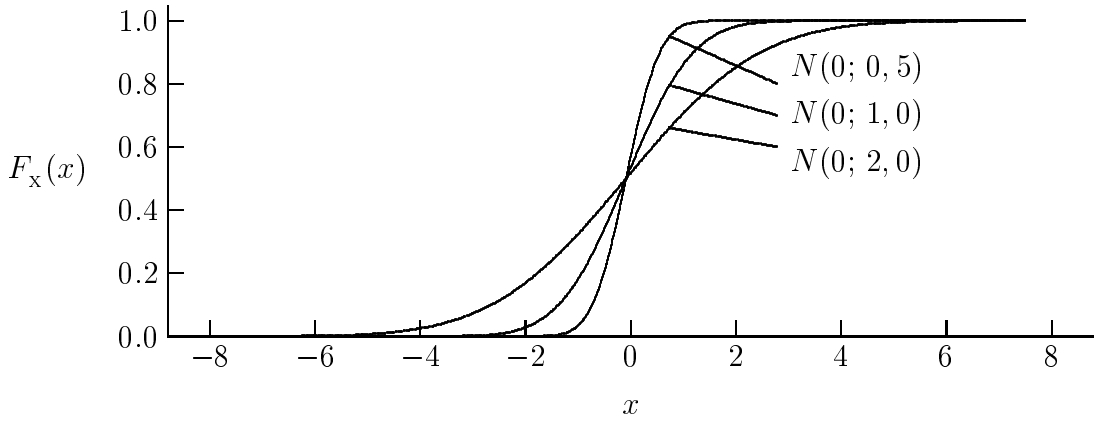


Рисунок 2.4 — Интегральный закон $F_{\mathrm{X}}(x)$ распределения Гаусса для случаев $N(0;\,0,5)$, $N(0;\,1,0)$, $N(0;\,2,0)$ с параметрами $m = 0$ и $\sigma = 0{,}5,\ 1{,}0$ и $2{,}0$

Из определения (2.1) следует, что нормальное распределение полностью определяется своими двумя параметрами – математическим ожиданием m и дисперсией σ^2.

Интегральная функция распределения $F_{\mathrm{X}}(x)$ нормальной случайной величины имеет вид

$$F_{\mathrm{X}}(x) \;=\; \Pr(\mathrm{X} < x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma} \int\limits_{-\infty}^{x} \exp\left(-\frac{(x'-m)^2}{2\sigma^2}\right) dx'. \tag{2.4}$$

С помощью функции $F_{\mathrm{X}}(x)$ может быть определена вероятность попадания нормальной случайной величины в заданный интервал $(\alpha,\,\beta)$:

$$\Pr(\alpha \le \mathrm{X} < \beta) \;=\; F_{\mathrm{X}}(\beta) - F_{\mathrm{X}}(\alpha). \tag{2.5}$$

Характеристическая функция $q(\lambda)$ нормальной случайной величины X следующая $(\sqrt{-1})$:

$$q(\lambda) \;=\; \mathrm{M}[\exp(i\lambda \mathrm{X})] \;=\; \exp\left(im\lambda - \frac{\lambda^2\sigma^2}{2}\right). \tag{2.6}$$

Одним из важнейших свойств нормальной случайной величины является её устойчивость при композиции. А именно, пусть X_1 и X_2 – нормальные случайные величины с параметрами m_1, m_2 и σ_1^2, σ_2^2. Тогда их сумма $\mathrm{X} = \mathrm{X}_1 + \mathrm{X}_2$ будет также нормальной случайной величиной с параметрами $m = m_1 + m_2$ и $\sigma^2 = \sigma_1^2 + \sigma_2^2$.

Для произвольных значений параметров m и σ табулировать функции $f_{\mathrm{X}}(x)$ и $F_{\mathrm{X}}(x)$ достаточно сложно, поэтому пользуются *стандартной нормальной величиной* (*стандартом*)

$$\mathrm{Z} \;=\; \frac{\mathrm{X} - m}{\sigma}, \tag{2.7}$$

для которой плотность распределения $f_{\mathrm{Z}}(z)$ следующая:

$$f_{\mathrm{Z}}(z) \;=\; \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \tag{2.8}$$

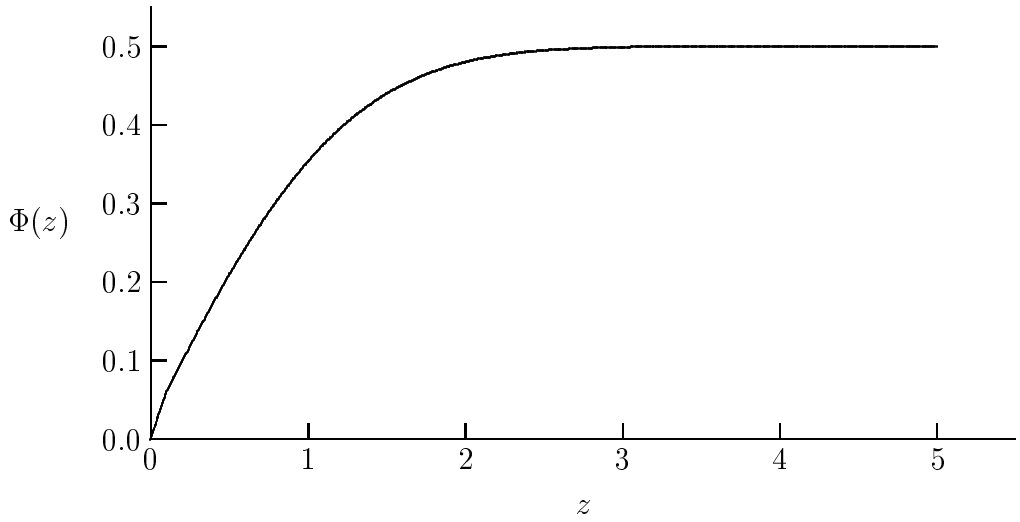


Рисунок 2.5 — Функция Лапласа $\Phi(z)$

т.е. её параметры равны $m_{\text{Z}} = 0$, $\sigma_{\text{Z}} = 1$ и, таким образом, $\text{Z} \to \mathcal{N}(0;\, 1)$. График $f_{\text{Z}}(z)$, приведенный на рис. 2.1, называют *кривой Гаусса*.

Интегральная функция распределения $F_{\text{Z}}(z)$ для стандартной нормальной величины определяется из выражения

$$F_{\text{Z}}(z) \;=\; \Pr(\text{Z} < z) \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{z} \exp\left(-\frac{u^2}{2}\right) du. \tag{2.9}$$

В силу симметрии события $\text{Z}< 0$ и $\text{Z}> 0$ равновероятны, поэтому $F_{\text{Z}}(0) = 0,5$ и

$$F_{\text{Z}}(z) \;=\; \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{z} \exp\left(-\frac{u^2}{2}\right) du.$$

Нормальное распределение часто используется в статистических вычислениях.

Функция распределения $F_{\text{Z}}(z)$ зависит только от одной переменной z. Поскольку интеграл в (2.9) не выражается через элементарные функции, в практических применениях пользуются *функцией Лапласа*

$$\Phi(z) \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{z} \exp\left(-\frac{u^2}{2}\right) du, \tag{2.10}$$

и поэтому

$$F_{\text{Z}}(z) \;=\; 0,5 + \Phi(z). \tag{2.11}$$

Функция Лапласа является нечетной, $\Phi(-z) = -\Phi(z)$, поэтому при табулировании приводят только неотрицательные значения аргумента z.

Для нормальной случайной величины X можно записать при $x_1 \le x_2$

$$\Pr(x_1 \le \text{X} < x_2) \;=\; \Phi\left(\frac{x_2 - m}{\sigma}\right) - \Phi\left(\frac{x_1 - m}{\sigma}\right). \tag{2.12}$$

В том случае, когда границы x_1 и x_2 располагаются симметрично относительно m, т.е. $x_1 = m - \varepsilon$ и $x_2 = m + \varepsilon$, то можно записать в более компактной форме

$$\Pr(|\mathrm{X} - m| < \varepsilon) \;=\; 2\Phi\left(\frac{\varepsilon}{\sigma}\right) \tag{2.13a}$$

или, если подставить $\varepsilon = z\sigma$,

$$\Pr(|\mathrm{X} - m| < z\sigma) \;=\; 2\Phi(z). \tag{2.13b}$$

Полагая в приведенной формуле $z = 1,\ 2,\ 3,\ 4,$ из таблиц функции Лапласа найдем

$$\begin{aligned}
\Pr(|\mathrm{X} - m| < \sigma) &= 2\Phi(1) = 0,683;\\
\Pr(|\mathrm{X} - m| < 2\sigma) &= 2\Phi(2) = 0,954;\\
\Pr(|\mathrm{X} - m| < 3\sigma) &= 2\Phi(3) = 0,9973;\\
\Pr(|\mathrm{X} - m| < 4\sigma) &= 2\Phi(4) = 0,99994.
\end{aligned} \tag{2.14}$$

Это равенство обычно для $z = 3$ формулируют в виде *правила трех сигм*: практически достоверно (т.е. с вероятностью 0,9973), что отклонение нормально распределенной случайной величины от её математического ожидания не превысит по абсолютной величине трех сигм. Вероятность противоположного события $\Pr(|\mathrm{X} - m| > 3\sigma)$ при этом составит 0,0027.

Определение. *Квантилем* $u_{1-\alpha/2}$ стандартного нормального распределения с плотностью $f_{\mathrm{Z}}(z)$, отвечающим заданному уровню значимости α, называется такое значение, при котором выполняется равенство

$$\Pr\left(|\mathrm{Z}| < u_{1-\alpha/2}\right) \;=\; \int_{-u_{1-\alpha/2}}^{u_{1-\alpha/2}} f_{\mathrm{Z}}(z)\,dz \;=$$

$$= \; F_{\mathrm{Z}}(u_{1-\alpha/2}) - F_{\mathrm{Z}}(-u_{1-\alpha/2}) \;=\; 2\Phi(u_{1-\alpha/2}) \;=\; 1 - \alpha. \tag{2.15}$$

С геометрической точки зрения нахождение квантиля $u_{1-\alpha/2}$ заключается в таком выборе двух граничных значений z, при которых площадь, ограниченная сверху кривой плотности $f_{\mathrm{Z}}(z)$, осью абсцисс снизу и вертикальными линиями, проходящими через точки $z = -u_{1-\alpha/2}$ и $z = u_{1-\alpha/2}$, была бы равной α. Другими словами, для нахождения квантиля $u_{1-\alpha/2}$ при заданном уровне α необходимо решить уравнение

$$2\Phi(u_{1-\alpha/2}) \;=\; 1 - \alpha. \tag{2.16}$$

Такой квантиль называется *двусторонним.*

Возможно также использование *левостороннего* и *правостороннего* квантилей.

Левосторонний квантиль $-u_{1-\alpha}$ ищут из условия

$$\Pr\left(\mathrm{Z} < -u_{1-\alpha}\right) \;=\; \int_{-\infty}^{-u_{1-\alpha}} f_{\mathrm{Z}}(z)\,dz \;=\; 0,5 - \Phi(u_{1-\alpha}) \;=\; \alpha, \tag{2.17a}$$

соответственно правосторонний квантиль $u_{1-\alpha}$ — из условия

$$\Pr\left(\mathrm{Z} > u_{1-\alpha}\right) \;=\; \int_{u_{1-\alpha}}^{\infty} f_{\mathrm{Z}}(z)\,dz \;=\; 0,5 + \Phi(u_{1-\alpha}) \;=\; 1 - \alpha. \tag{2.17b}$$

В случае нормального распределения левосторонний и правосторонний квантили равны по модулю, это же относится к обоим двусторонним квантилям. Конкретные их значения можно найти на основании формул (2.16) и (2.17) из таблиц функции Лапласа (см. приложение).

Подробно техника применения нормальной случайной величины при решении задач математической статистики изложена ниже.

2.2. Системы нормальных случайных величин

Нормальный закон распределения для системы из двух случайных величин (X,Y) (нормальный закон на плоскости) имеет плотность вида

$$f(x,y) \;=\; \frac{1}{2\pi\,\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}\,\sqrt{1-r_{\mathrm{XY}}^2}}\exp\left\{-\;\frac{1}{2(1-r_{\mathrm{XY}}^2)}\,Q(x,y)\right\}, \tag{2.18}$$

$$Q(x,y) \;=\; \frac{(x-m_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2} - 2r_{\mathrm{XY}}\frac{(x-m_{\mathrm{X}})(y-m_{\mathrm{Y}})}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} + \frac{(y-m_{\mathrm{Y}})^2}{\sigma_{\mathrm{Y}}^2},$$

где $m_{\mathrm{X}}, m_{\mathrm{Y}}$ – математические ожидания случайных величин X и Y; $\sigma_{\mathrm{X}}, \sigma_{\mathrm{Y}}$ – их средние квадратические отклонения; r_{XY} – их коэффициент корреляции.

Для случайных величин, распределенных по нормальному закону, некоррелированность равносильна независимости. Если случайные величины X и Y некоррелированы (независимы), то $r_{\mathrm{XY}} = 0$ и

$$f(x,y) \;=\; \frac{1}{2\pi\,\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}}\exp\left\{-\;\frac{(x-m_{\mathrm{X}})^2}{2\sigma_{\mathrm{X}}^2} - \frac{(y-m_{\mathrm{Y}})^2}{2\sigma_{\mathrm{Y}}^2}\right\}. \tag{2.19}$$

Вероятность попадания случайной точки, распределенной по нормальному закону, в прямоугольник R со сторонами, параллельными главным осям рассеивания, выражается формулой

$$\Pr\{(\mathrm{X},\mathrm{Y})\in R\} \;= \tag{2.20}$$

$$= \;\left[\Phi\left(\frac{\beta-m_{\mathrm{X}}}{\sigma_{\mathrm{X}}}\right) - \Phi\left(\frac{\alpha-m_{\mathrm{X}}}{\sigma_{\mathrm{X}}}\right)\right]\left[\Phi\left(\frac{\delta-m_{\mathrm{Y}}}{\sigma_{\mathrm{Y}}}\right) - \Phi\left(\frac{\gamma-m_{\mathrm{Y}}}{\sigma_{\mathrm{Y}}}\right)\right], \tag{2.20}$$

здесь $x\in[\alpha,\,\beta),\;\; y\in[\gamma,\,\delta)$.

Эллипсом равной плотности (эллипсом рассеивания) называется эллипс, во всех точках которого совместная плотность нормального закона постоянна: $f(x,y) = const$. Полуоси эллипса пропорциональны σ_{X}, σ_{Y}: $a = k\sigma_{\mathrm{X}}$, $b = k\sigma_{\mathrm{Y}}$.

Вероятность попадания случайной точки, распределенной по нормальному закону, в область E_k, ограниченную эллипсом рассеивания с полуосями a и b,

$$\Pr\{(\mathrm{X},\mathrm{Y})\in \mathrm{E}_k\} \;=\; 1-\exp(-k^2/2), \tag{2.21}$$

где k – размеры полуосей эллипса, выраженные в средних квадратичных отклонениях: $k = a/\sigma_{\mathrm{X}} = b/\sigma_{\mathrm{Y}}$.

Если $\sigma_{\rm X} = \sigma_{\rm Y} = \sigma$, рассеивание по нормальному закону называется *круговым*. При круговом нормальном рассеивании с $m_{\rm X} = m_{\rm Y} = 0$ расстояние R от точки $({\rm X};{\rm Y})$ до начала координат (центра рассеивания) распределено *по закону Релея*

$$f(r) \;=\; \left(r/\sigma^2\right) \exp\left(-r^2/2\sigma^2\right), \qquad r \geq 0. \tag{2.22}$$

Распределению Релея подчиняется модуль вектора на плоскости, если его ортогональные составляющие (проекции на координатные оси) независимы и распределены нормально с нулевыми математическими ожиданиями и равными дисперсиями.

Обобщением закона Релея является *распределение Релея-Райса*. При заданной константе s плотность этого закона следующая:

$$f(r) \;=\; \left(r/\sigma^2\right) \exp\left(-r^2/2\sigma^2\right)\, I_0\left(rs/\sigma^2\right), \qquad r \geq 0, \tag{2.23}$$

где

$$I_0(x) \;=\; (2\pi)^{-1} \int\limits_0^{2\pi} \exp[x\cos(\varphi)]\, d\varphi$$

– *модифицированная функция Бесселя нулевого индекса.*

Нормальный закон в пространстве трех измерений для независимых случайных величин X, Y, Z

$$f(x,y,z) \;= \tag{2.24}$$

$$= \;\frac{1}{(2\,\pi)^{3/2}\,\sigma_{\rm X}\,\sigma_{\rm Y}\,\sigma_{\rm Z}} \exp\left(-\frac{(x-m_{\rm X})^2}{2\,\sigma_{\rm X}^2} - \frac{(y-m_{\rm Y})^2}{2\,\sigma_{\rm Y}^2} - \frac{(z-m_{\rm Z})^2}{2\,\sigma_{\rm Z}^2}\right).$$

Вероятность попадания случайной точки $({\rm X},{\rm Y},{\rm Z})$ в область E_k, ограниченную эллипсоидом равной плотности с полуосями $a = k\sigma_{\rm X}$, $b = k\sigma_{\rm Y}$, $c = k\sigma_{\rm Z}$, равна

$$\Pr\{({\rm X},{\rm Y},{\rm Z}) \in E_k\} \;=\; 2\,\Phi(k) - 1 - \sqrt{2/\pi}k\exp(-k^2/2). \tag{2.25}$$

Нормальный закон распределения для системы из n случайных величин ${\rm X}=({\rm X}_1,{\rm X}_2,\ldots,{\rm X}_n)$ (нормальный закон в евклидовом пространстве размерности n) имеет плотность вида

$$f(X) \;=\; \frac{1}{\sqrt{(2\pi)^n\,\det(K)}} \exp\left(-\frac{1}{2}(X-X_c)^{\rm T} K^{-1} (X-X_c)\right). \tag{2.26}$$

Вектор X образован из n-компонентного набора и распределен по многомерному нормальному закону с корреляционной матрицей K и вектором математических ожиданий $X_c = (X_{c1}, X_{c2}, \ldots, X_{cn})$.

Из выражения (2.26) следует, что математическое ожидание и дисперсия ${\rm D[X]}$ случайного вектора X следующие:

$${\rm M[X]} \;=\; X_c, \tag{2.27a}$$

$${\rm D[X]} \;=\; {\rm M}[({\rm X}-X_c)^{\rm T}({\rm X}-X_c)] \;=\; {\rm Sp}(K). \tag{2.27b}$$

Итак, искомая дисперсия случайной векторной величины X равна сумме диагональных элементов ${\rm Sp}(K)$ корреляционной матрицы K.

2.3. Гамма-функция Эйлера и её свойства. Гамма-распределение

I. Гамма-функцией или интегралом Эйлера называется функция вида

$$\Gamma(\alpha) \;=\; \int\limits_0^\infty x^{\alpha-1}\exp(-x)\,dx. \tag{2.28}$$

Область определения гамма-функции – вся числовая ось, кроме нуля и отрицательных целых чисел, $\alpha = 0, -1, -2, \ldots$.

Гамма-функция является интегралом, зависящим от параметра α. Она удовлетворяет следующим свойствам:

$$1)\ \Gamma(\alpha+1) \;=\; \alpha\Gamma(\alpha) \quad \text{при} \quad \alpha > 0. \tag{2.29}$$

$$2)\ \Gamma(1) \;=\; \Gamma(2) \;=\; 1. \tag{2.30}$$

$$3)\ \Gamma\left(\tfrac{1}{2}\right) \;=\; \sqrt{\pi}. \tag{2.31}$$

Таким образом, если α – положительное целое число, $\alpha = n$, то

$$\Gamma(n) \;=\; (n-1)! \qquad \text{или} \qquad \Gamma(n+1) \;=\; n! \tag{2.32}$$

Следовательно, гамма-функция может рассматриваться как обобщение факториала. Если же аргумент α равен нулю или отрицательному целому числу, то значение гамма-функции расходится.

Если аргумент α пропорционален $\frac{1}{2}$, то $\Gamma(\alpha)$ может быть легко вычислена. Например, $\Gamma(-1/2) = -2\sqrt{\pi}$.

При больших значениях аргумента α гамма-функция вычисляется по формуле $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) = (\alpha-1)(\alpha-2)\Gamma(\alpha-2) = \ldots$ и так далее, пока аргумент гамма-функции не окажется в интервале $(1;,2)$. В этом интервале гамма-функцию подробно табулируют. График гамма-функции $\Gamma(\alpha)$ изображен на рис. 2.6.

II. Моделью образования *гамма-распределения* является поток событий с постоянной интенсивностью λ. При этом в качестве случайной величины T рассматривается время, необходимое для появления данного числа β событий. Таким образом, как и в показательном законе, величина T меняется в интервале $(0, \infty)$. Можно показать, что это распределение будет задаваться плотностью распределения вероятностей

$$f_{\text{т}}(t;\beta,\lambda) \;=\; \frac{\lambda^\beta}{\Gamma(\beta)}\,t^{\beta-1}\mathrm{e}^{-\lambda t}, \qquad t \geq 0, \tag{2.33}$$

или функцией распределения

$$\Pr(\mathrm{T} \leq t) \;=\; F_{\text{т}}(t;\beta,\lambda) \;=\; \frac{\lambda^\beta}{\Gamma(\beta)}\int\limits_0^t u^{\beta-1}\mathrm{e}^{-\lambda u}\,du. \tag{2.34}$$

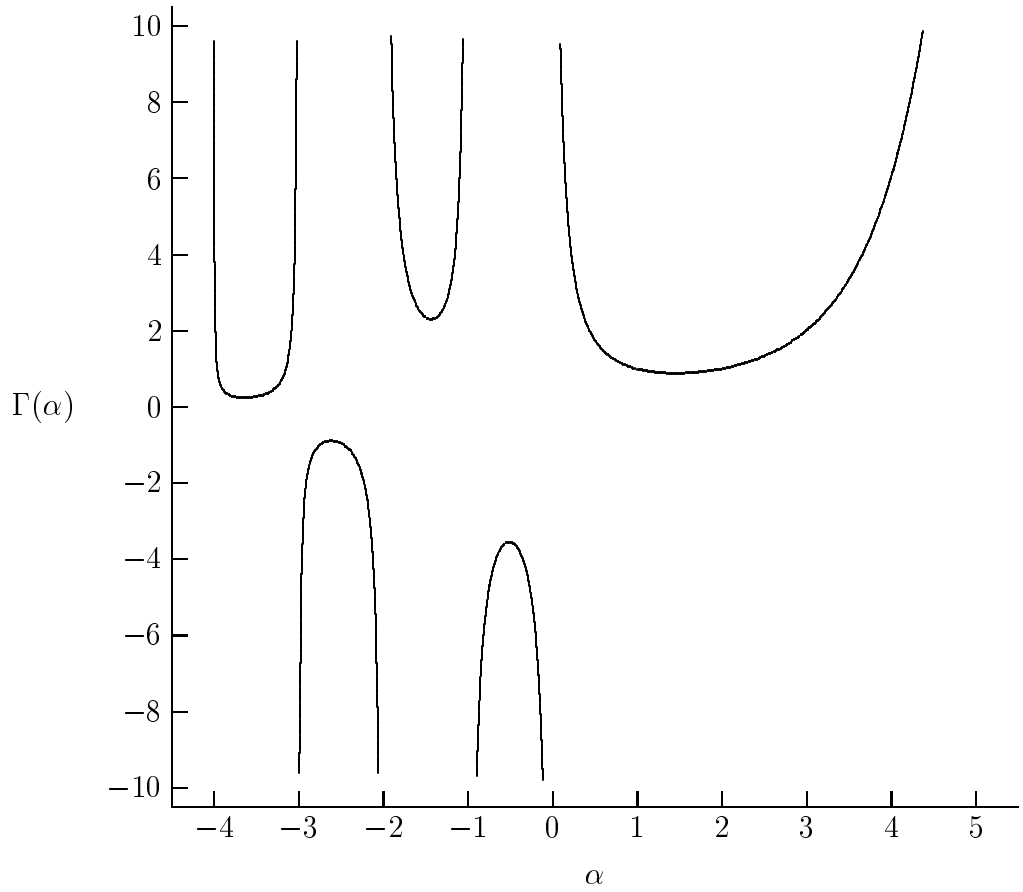


Рисунок 2.6 — График гамма-функции Эйлера $\Gamma(\alpha)$

Распределение (2.33) или (2.34) называется *гамма-распределением* (рис. 2.7). Оно зависит от своего аргумента t и параметров β, λ.

Моменты первого и второго порядков определяются из равенств

$$\mathrm{M[T]} \;=\; \frac{\beta}{\lambda}, \qquad \mathrm{D[T]} \;=\; \frac{\beta}{\lambda^2}. \tag{2.35}$$

Из выражений (2.35) можно получить оценки параметров гамма-распределения по опытным данным:

$$\lambda^* \;=\; \mathrm{M[T]}^* \,/\, \mathrm{D[T]}^*; \tag{2.36a}$$

$$\beta^* \;=\; \mathrm{M[T]}^{*2} \,/\, \mathrm{D[T]}^* \;=\; \lambda^* \mathrm{M[T]}^*. \tag{2.36b}$$

При некоторых определенных значениях параметра β получаем частные случаи гамма-распределения. Так, при $\beta = 1$ гамма-распределение совпадает с показательным. При β – положительном целом числе получаем *распределение Эрланга*, широко используемое в теории массового обслуживания.

Характеристическая функция гамма-распределения следующая:

$$q_{_T}(z) \;=\; \mathrm{M}\left[\exp(iz\mathrm{T})\right] \;=\; (1 - iz/\lambda)^{-\beta}\,. \tag{2.37}$$

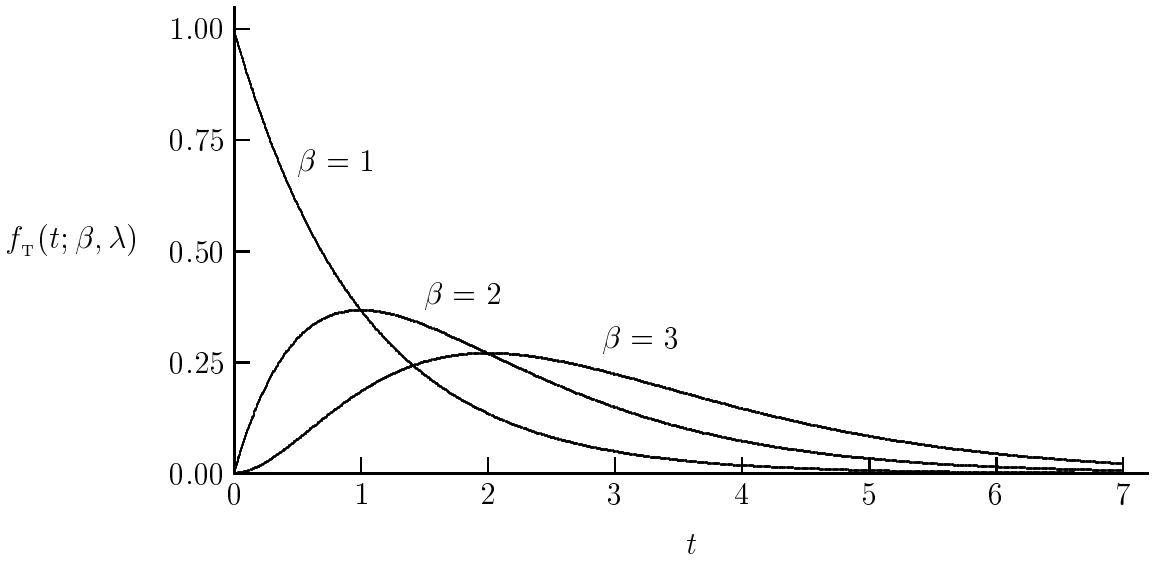


Рисунок 2.7 — Плотность распределения вероятностей $f_{\text{Т}}(t;\beta,\lambda)$ гамма-распределения при $\lambda = 1{,}0$ и трех значениях параметра β

2.4. Распределение χ^2 (хи-квадрат)

Это распределение связано с нормальным и широко используется при решении различных задач статистического анализа. В основе образования этого распределения лежит рассмотрение выборки из нормальной совокупности.

Рассмотрим случайную величину Y, распределенную по нормальному закону с математическим ожиданием $\text{M[Y]} = a$ и средним квадратическим отклонением σ, или, более кратко, пусть $\text{Y} \to \mathcal{N}(a;\sigma)$.

Тогда случайная величина $\text{U} = (\text{Y} - a)/\sigma$, называемая *стандартизованной случайной величиной*, распределена по нормальному закону с параметрами $\text{M[U]} = 0$ и $\sigma_{\text{U}} = 1$, т.е. $\text{U} \to \mathcal{N}(0;\,1)$.

Квадрат стандартизованной случайной величины

$$\text{U}^2 \;=\; \left(\frac{\text{Y} - a}{\sigma}\right)^2 \;\equiv\; \chi^2 \tag{2.38}$$

называется *случайной величиной χ^2 (хи-квадрат) с одной степенью свободы.*

Рассмотрим теперь n независимых случайных величин $\text{Y}_1,\ \text{Y}_2,\ \ldots,\ \text{Y}_n$, распределенных по нормальному закону с математическими ожиданиями $a_1, a_2, \ldots, a_n$ и средними квадратичными отклонениями $\sigma_1, \sigma_2, \ldots, \sigma_n$. Образуем для каждой из этих случайных величин *стандартизованную* нормальную случайную величину

$$\text{U}_i \;=\; \frac{\text{Y}_i - a_i}{\sigma_i}, \qquad i = 1, 2, \ldots, n.$$

Сумма квадратов стандартизованных переменных

$$\chi^2 \;=\; \text{U}_1^2 + \text{U}_2^2 \ldots + \text{U}_n^2 \;= \tag{2.39}$$

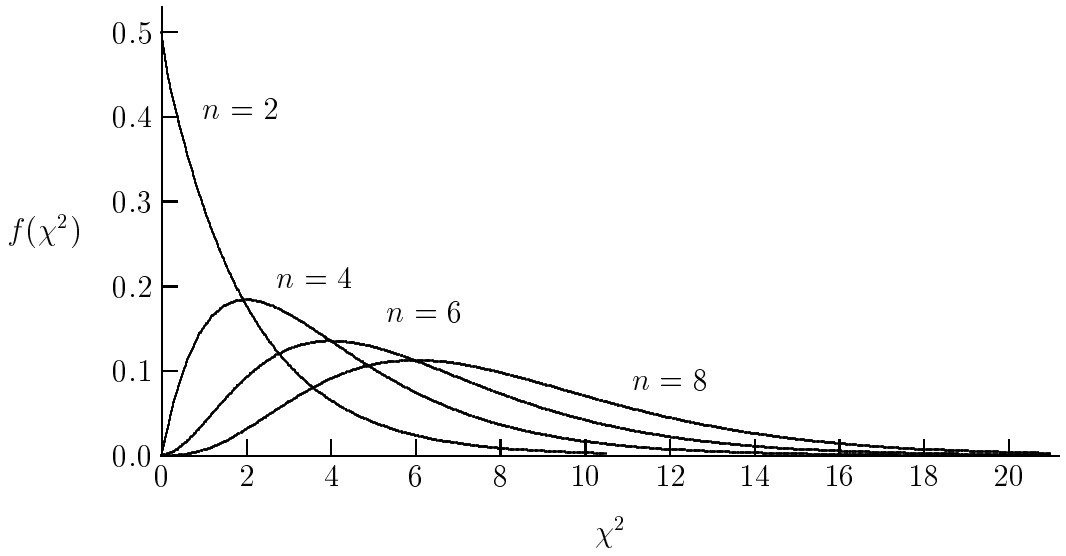


Рисунок 2.8 — Плотность распределения вероятностей $f(\chi^2)$ распределения χ^2 для числа степеней свободы $n = 2,\ 4,\ 6,\ 8$

$$= \left(\frac{\mathrm{Y}_1 - a_1}{\sigma_1}\right)^2 + \left(\frac{\mathrm{Y}_2 - a_2}{\sigma_2}\right)^2 + \ldots + \left(\frac{\mathrm{Y}_n - a_n}{\sigma_n}\right)^2$$

называется *случайной величиной* χ^2 с $\nu = n$ *степенями свободы*. В статистических таблицах и при выполнении расчетов число степеней свободы принято обозначать буквой ν.

Плотность распределения случайной величины χ^2 имеет вид

$$f(t) \;=\; \left[2^{\nu/2}\,\Gamma(\nu/2)\right]^{-1}\, t^{\nu/2-1}\, \exp(-t/2), \tag{2.40}$$

при этом $t \geq 0$.

Иногда удобно плотность распределения указывать непосредственно в терминах χ^2:

$$f(\chi^2)\,d(\chi^2) \;=\; \left[2^{\nu/2}\,\Gamma(\nu/2)\right]^{-1}\, (\chi^2)^{\nu/2-1}\, \exp\left(-\chi^2/2\right) d(\chi^2). \tag{2.40$'$}$$

Сравнивая плотность распределения (2.40) с ранее приведенной плотностью (2.33), заключаем, что распределение χ^2 представляет собой частный случай гамма-распределения при $\beta = \dfrac{n}{2}$ и $\lambda = \dfrac{1}{2}$.

Интегральная функция χ^2-распределения имеет вид

$$F(\chi^2) \;=\; \Pr(\mathrm{X} \leq \chi^2) \;=\; \left[2^{\nu/2}\,\Gamma(\nu/2)\right]^{-1} \int_0^{\chi^2} x^{\nu/2-1}\, \exp(-x/2)\, dx. \tag{2.41}$$

Таким образом, распределение χ^2 зависит от одного параметра ν – числа степеней свободы.

На рис. 2.8–2.9 изображены примеры графиков плотности распределения вероятностей $f(\chi^2)$ и интегральной функции χ^2-распределения $F(\chi^2)$ соответственно.

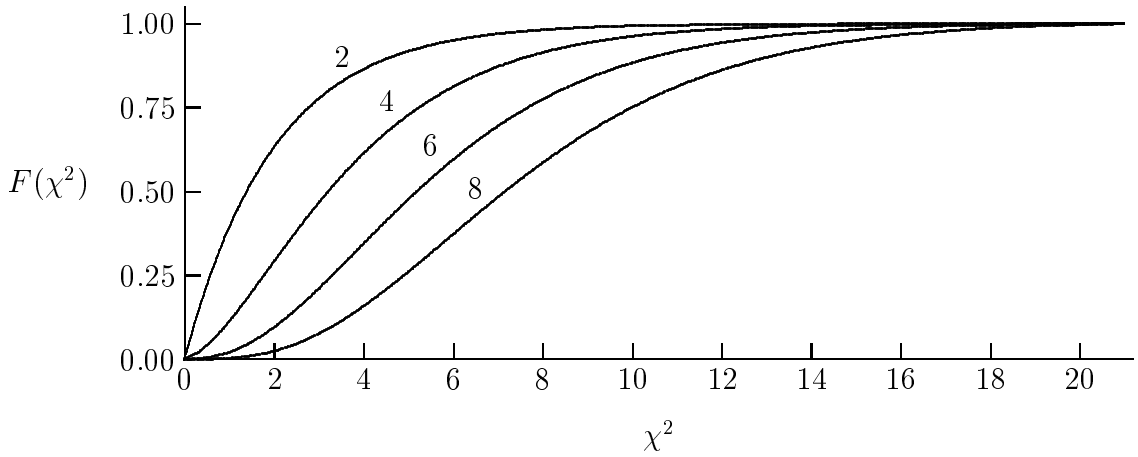


Рисунок 2.9 — Интегральный закон распределения $F(\chi^2)$ (числа степеней свободы $n = 2,\ 4,\ 6,\ 8$ указаны цифрами у кривых)

Как это видно из графиков для $f(\chi^2)$, плотность χ^2-распределения асимметрична, она имеет правый удлиненный "хвост" (в периферийной области больших уклонений χ^2). С ростом n (или ν) асимметричность плотности уменьшается, при этом закон распределения стремится к нормальному.

Рассчитав первые два момента χ^2, получим

$$\mathrm{M}[\chi^2] \;=\; \nu, \qquad \mathrm{D}[\chi^2] \;=\; 2\nu. \tag{2.42}$$

Распределение χ^2 часто используется в статистических вычислениях, в частности, в связи со следующей теоремой.

Т е о р е м а. *Пусть $x_1, x_2, \ldots, x_n$ – заданная выборка из нормально распределенной генеральной совокупности $\mathcal{N}(m, \sigma)$ объема n, при этом $x^* = \frac{1}{n}\sum_{i=1}^{n} x_i$ и $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x^*)^2$ – соответственно выборочное арифметическое среднее и выборочная дисперсия. Тогда статистики X^* и S^2 – независимые случайные величины, причем статистика $(n-1)\sigma^{-2}\mathrm{S}^2$ имеет распределение χ^2_{n-1}.*

Характеристическая функция распределения χ^2 с числом степеней свободы $n = \nu$ следующая:

$$q_\nu(z) \;=\; \mathrm{M}\left[\exp(iz\chi^2_\nu)\right] \;=\; (1-2iz)^{-\nu/2}\,. \tag{2.43}$$

Пусть рассматривается случайная величина χ^2_n с n степенями свободы и случайная величина χ^2_m с m степенями свободы, а также аддитивная случайная величина $\chi^2_{n+m} = \chi^2_n + \chi^2_m$ с $n+m$ степенями свободы. Из вида характеристической функции (2.43) вытекает, что

$$q_{n+m}(z) \;=\; q_n(z)\,q_m(z) \;=\; (1-2iz)^{-(n+m)/2}\,, \tag{2.44}$$

а плотность распределения вероятностей композиции $\chi^2_{n+m} = \chi^2_n + \chi^2_m$ следующая:

$$f_{n+m}(t) \;=\; \left[2^{(n+m)/2}\,\Gamma\left(\frac{n+m}{2}\right)\right]^{-1} t^{(n+m)/2-1}\,\exp(-t/2). \tag{2.45}$$

Таким образом, аддитивная случайная величина $\chi^2_{n+m} = \chi^2_n + \chi^2_m$ также описывается законом χ^2 с $n+m$ степенями свободы, т.е. случайная величина χ^2 обладает свойством устойчивости при композиции.

Пусть имеется n квадратов стандартизованных переменных U_1^2, U_2^2, ..., U_n^2, на которые, в свою очередь, наложены s линейных зависимостей (связей). Тогда распределение их суммы будет подчиняться закону (2.40), но с числом степеней свободы, равным $\nu = n - s$.

В практике статистических задач чаще используют не саму плотность распределения вероятностей или функцию распределения, а *квантили* χ^2*-распределения*, отвечающие *уровню значимости* α, которые при заданном ν обозначают $\chi^2_{\alpha,\nu}$.

Определение. *Квантилем* $\chi^2_{\alpha,\nu}$, отвечающим заданному уровню значимости α, называется такое значение $\chi^2 = \chi^2_{\alpha,\nu}$, при котором выполняется равенство

$$\Pr\left(\chi^2 > \chi^2_{\alpha,\nu}\right) \;=\; 1 - F(\chi^2_{\alpha,\nu}) \;=\; \int_{\chi^2_{\alpha,\nu}}^{\infty} f(\chi^2)\, d(\chi^2) \;=\; \alpha. \qquad (2.46)$$

С геометрической точки зрения нахождение квантиля $\chi^2_{\alpha,\nu}$ заключается в таком выборе значения $\chi^2 = \chi^2_{\alpha,\nu}$, при котором площадь, ограниченная сверху кривой плотности $f(\chi^2)$, осью абсцисс снизу и вертикальной линией, проходящей через точку $\chi^2 = \chi^2_{\alpha,\nu}$, была бы равной α. Другими словами, для нахождения квантиля $\chi^2_{\alpha,\nu}$ при заданных α и ν необходимо решить уравнение (2.46).

На рис. 2.10 (сверху), на котором приведена плотность $f(\chi^2)$ для случая $\alpha = 0,20$ и $\nu = 10$, эта площадь, численно равная уровню значимости α, заштрихована вертикальными линиями. На том же рисунке (снизу) приведена интегральная функция распределения $F(\chi^2)$, рассмотрен тот же случай ($\alpha = 0,20$ и $\nu = 10$). Пунктиром на рисунке отмечен уровень значимости α и отвечающий ему квантиль $\chi^2_{\alpha,\nu}$.

В приложении приведены значения квантилей $\chi^2_{\alpha,\nu}$ для различных значений числа ν степеней свободы и уровня значимости α.

2.5. Распределение Стьюдента

Распределение Стьюдента (t-распределение) имеет большое значение при статистических вычислениях, связанных с нормальным законом, а именно тогда, когда среднее квадратическое отклонение σ неизвестно и еще подлежит определению по опытным данным.

Пусть Y и Y_1, Y_2, ..., Y_n – независимые случайные величины, имеющие нормальное распределение с параметрами

$$\mathrm{M}[\mathrm{Y}] = \mathrm{M}[\mathrm{Y}_1] = \mathrm{M}[\mathrm{Y}_2] = \ldots = \mathrm{M}[\mathrm{Y}_n] = 0$$

и

$$\sigma[\mathrm{Y}] = \sigma[\mathrm{Y}_1] = \sigma[\mathrm{Y}_2] = \ldots = \sigma[\mathrm{Y}_n] = 1.$$

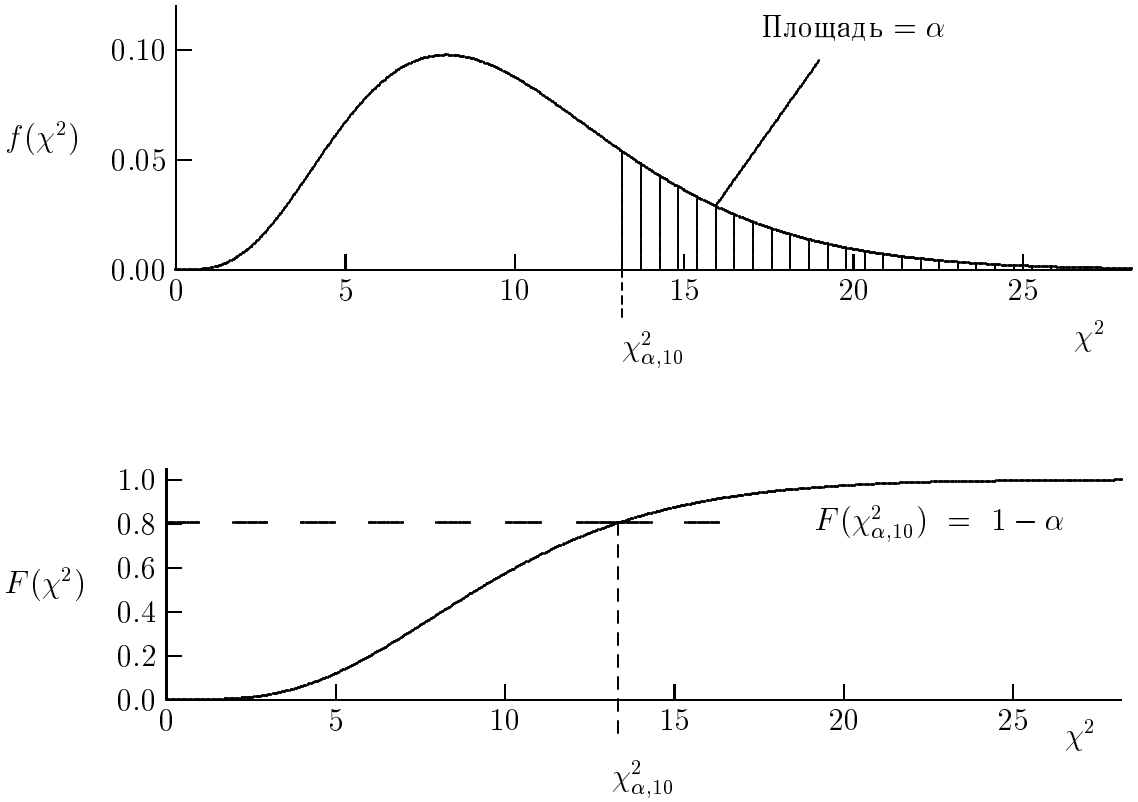


Рисунок 2.10 — Плотность распределения вероятностей $f(\chi^2)$ (сверху) и интегральный закон распределения $F(\chi^2)$ (снизу) для числа степеней свободы $\nu = 10$ ($\alpha = 0,2$; квантиль $\chi^2_{\alpha,\nu} = \chi^2_{0,2,\,10}$ равен 13,20)

Случайная величина

$$\mathrm{T} \;=\; \mathrm{Y}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{Y}_i^2\right)^{-1/2} \;=\; \mathrm{Y}\left(\frac{1}{n}\chi_n^2\right)^{-1/2}, \qquad (2.47)$$

являющаяся функцией нормально распределенных случайных величин, называется *безразмерной дробью Стьюдента*.

В курсе теории вероятностей показывается, что плотность распределения вероятностей случайной величины T имеет вид (см. примеры к разделу)

$$f_{\mathrm{T}}(t) \;=\; S(t,\nu) \;=\; \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+t^2/\nu\right)^{-(\nu+1)/2}, \qquad -\infty < t < \infty. \qquad (2.48)$$

Это распределение непараметрическое, т.е. оно не зависит от параметров исходных случайных величин, а зависит лишь от их числа.

В этой формуле буквой ν обозначено число слагаемых в подкоренном выражении дроби Стьюдента, т.е. $\nu = n$. Такое обозначение числа степеней свободы общепринято в математической статистике, так как облегчает пользование статистическими таблицами.

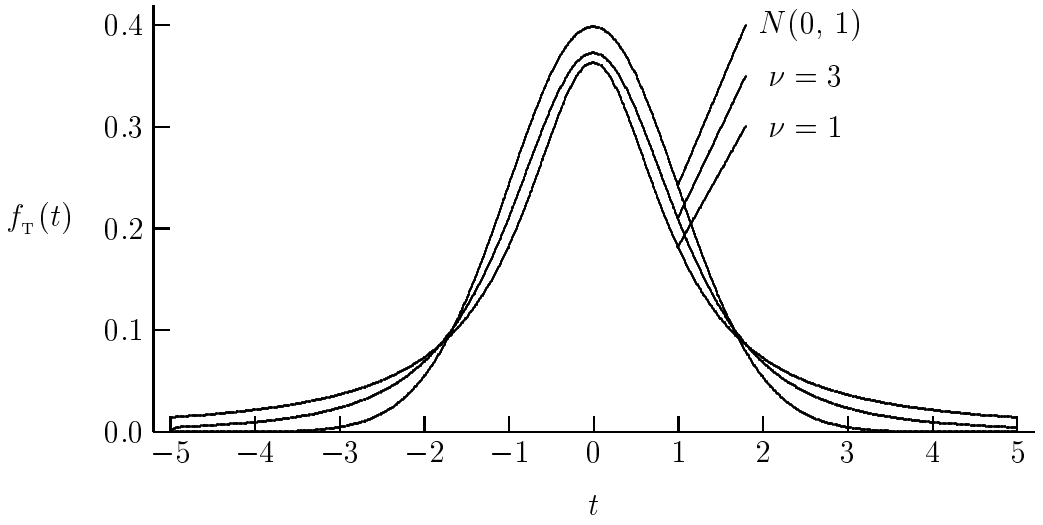


Рисунок 2.11 — Плотности распределения Стьюдента $f_{\mathrm{T}}(t)$ при степенях свободы $\nu = 1$, $\nu = 3$ и $\nu = \infty$ – кривая нормального распределения $N(0;\,1)$

Из формулы (2.48) следует, что распределение случайной величины T не зависит от параметров распределения независимых случайных величин Y и $\mathrm{Y}_1,\ \mathrm{Y}_2,\ \ldots,\ \mathrm{Y}_n$, а зависит только от одного параметра – числа степеней свободы ν, равного числу слагаемых в подкоренном выражении дроби Стьюдента (2.47).

Математическое ожидание и дисперсия случайной величины T соответственно равны

$$\mathrm{M[T]} \;=\; 0, \qquad \mathrm{D[T]} \;=\; \frac{\nu}{\nu - 2}, \qquad \nu > 2. \tag{2.49}$$

При неограниченном увеличении числа степеней свободы распределение Стьюдента асимптотически переходит в нормальное распределение Гаусса с параметрами $\mathrm{M[T]} = 0,\ \mathrm{D[T]} = 1$.

На рис. 2.11 изображен график плотности распределения Стьюдента при различных степенях свободы. Анализируя этот график, замечаем, что с увеличением числа степеней свободы ν он приближается к кривой Гаусса. Если же число степеней свободы ν мало, то вероятности больших отклонений несколько больше по сравнению с нормальным законом (при $\mathrm{T} > 2$ кривая t-распределения располагается выше нормальной кривой).

В математической статистике достаточно часто используются квантили $t_{\alpha/2;\,\nu}$ распределения Стьюдента в зависимости от числа ν степеней свободы и заданного уровня вероятности α. Значения квантилей распределения Стьюдента $t_{\alpha/2;\,\nu}$ могут быть найдены из решения уравнения

$$\Pr\left(|\mathrm{T}| > t_{\alpha/2;\,\nu}\right) \;=\; 2 \int\limits_{t_{\alpha/2;\,\nu}}^{\infty} f_{\mathrm{T}}(t')\,dt' \;=\; \alpha. \tag{2.50}$$

С геометрической точки зрения нахождение квантилей распределения Стьюдента $t_{\alpha/2;\,\nu}$ заключается в таком выборе значения $t_{\alpha/2;\,\nu}$, при котором суммарная площадь под кривой плотности $f_{\mathrm{T}}(t)$ на участках $\left(-\infty,\ -t_{\alpha/2;\,\nu}\right)$ и $\left(t_{\alpha/2;\,\nu}, \infty\right)$

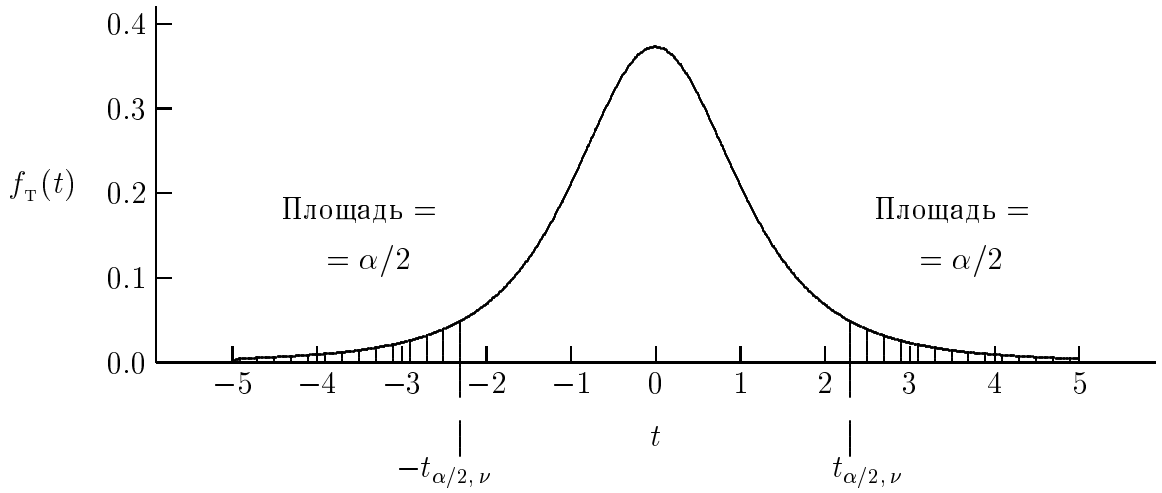


Рисунок 2.12 — Плотность вероятностей $f_{\mathrm{T}}(t)$ распределения Стьюдента для числа степеней свободы $\nu = 3$ (заштрихованная площадь равна α)

была бы равной α. На рис. 2.12 суммарная площадь заштрихованных двух участков составляет α.

2.6. Распределение Фишера

Распределение Фишера (F-распределение) употребляется при сравнении дисперсий нормальных распределений, вычисленных на основании опытных данных. (Этот закон еще часто называют распределением Фишера–Снедекора).

Пусть случайные величины $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_m$ и $\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_n$ независимы и имеют нормальное распределение с параметрами

$$\mathrm{M}[\mathrm{X}_i] = 0; \qquad \mathrm{D}[\mathrm{X}_i] = 1; \qquad i = 1, 2 \ldots m;$$
$$\mathrm{M}[\mathrm{Y}_j] = 0; \qquad \mathrm{D}[\mathrm{Y}_j] = 1; \qquad j = 1, 2 \ldots n.$$

Безразмерная случайная величина

$$\mathrm{F} \;=\; \left(\frac{1}{m}\sum_{i=1}^{m}\mathrm{X}_i^2\right)\left(\frac{1}{n}\sum_{j=1}^{n}\mathrm{Y}_j^2\right)^{-1} \tag{2.51}$$

распределена по закону распределения Фишера, т.е. имеет плотность распределения вероятностей $(x \geq 0)$

$$f_{\mathrm{F}}(x) \;=\; \frac{\Gamma\left(\dfrac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\dfrac{\nu_1}{2}\right)\Gamma\left(\dfrac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{-1+\nu_1/2}\left(1+\frac{\nu_1}{\nu_2}x\right)^{-(\nu_1+\nu_2)/2}, \tag{2.52}$$

где $\nu_1 = m$ – число степеней свободы числителя; $\nu_2 = n$ – число степеней свободы знаменателя в (2.51).

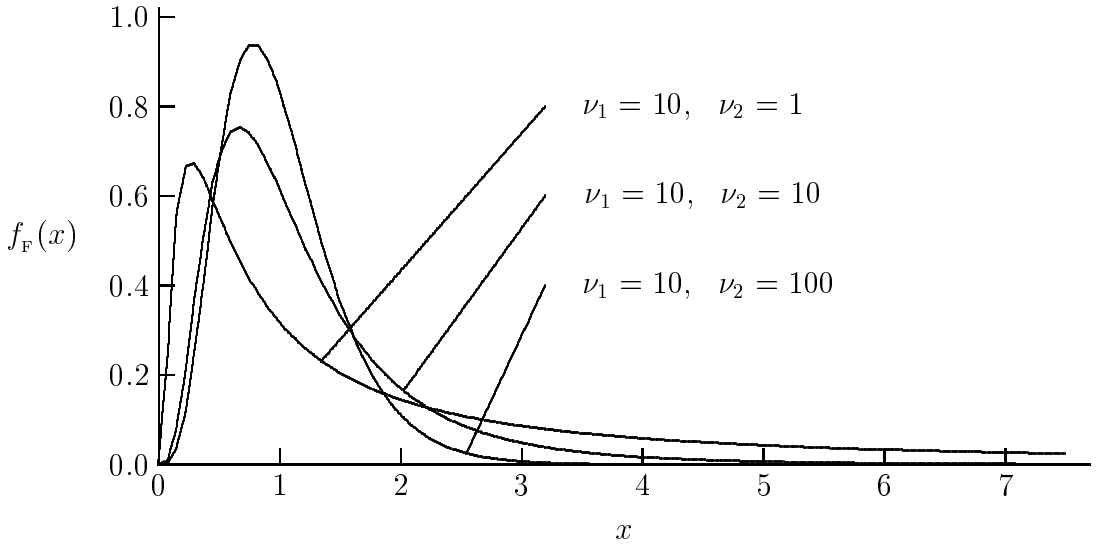


Рисунок 2.13 — Плотность распределения Фишера $f_{\mathrm{F}}(x)$ для различных чисел степеней свободы ν_1 и ν_2

Из формулы (2.52) следует, что распределение случайной величины F зависит от двух параметров – *чисел степеней свободы* $\nu_1 = m$ и $\nu_2 = n$. Для удобства пользователей статистическими таблицами числа степеней свободы по-прежнему обозначаются буквами $\nu_1 = m$ и $\nu_2 = n$. График плотности вероятностей F-распределения изображен на рис. 2.13.

В математической статистике достаточно часто используются квантили F-распределения $f_{\alpha;\nu_1,\nu_2}$ в зависимости от числа степеней свободы ν_1 и ν_2 и заданного уровня вероятности α. Из решения уравнения

$$\Pr\left(|\mathrm{F}| > f_{\alpha;\nu_1,\nu_2}\right) = 2\int\limits_{f_{\alpha;\nu_1,\nu_2}}^{\infty} f_{\mathrm{F}}(x)\,dx = \alpha \tag{2.53}$$

могут быть найдены значения квантилей $f_{\alpha;\nu_1,\nu_2}$ распределения Фишера.

2.7. Распределение Колмогорова

В практических задачах проверки гипотез о согласии данных выборки с конкретным теоретическим законом распределения для любой непрерывной случайной величины применяется λ-критерий А.Н. Колмогорова.

Возникающая при проверке такой гипотезы случайная величина (выборочная статистика) Λ имеет интегральную функцию распределения $K(\lambda)$ вида (рис. 2.14)

$$K(\lambda) = \Pr\left(\Lambda < \lambda\right) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2\lambda^2) \tag{2.54}$$

и плотность распределения вероятностей

$$f(\lambda) = \frac{d}{d\lambda} K(\lambda) = 4\lambda \sum_{k=-\infty}^{\infty} (-1)^k k^2 \exp(-2k^2\lambda^2). \tag{2.55}$$

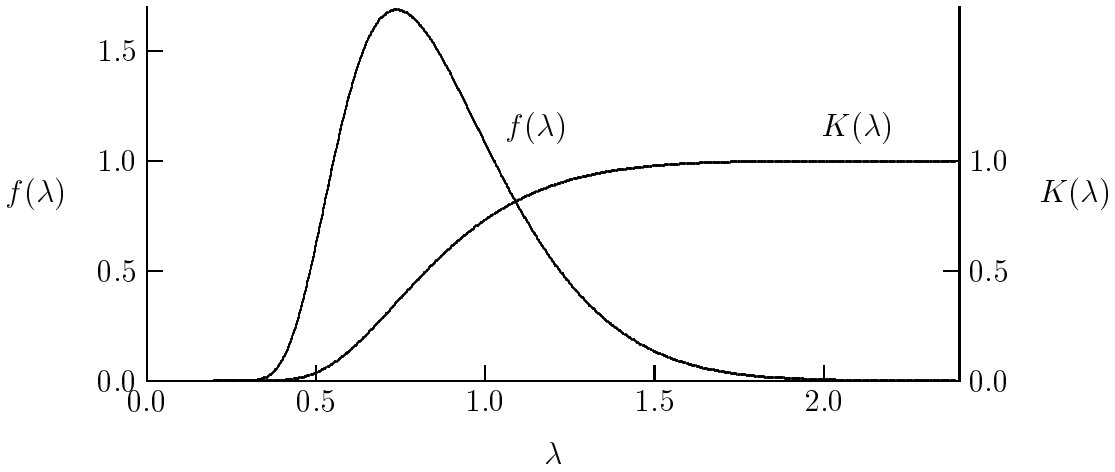


Рисунок 2.14 — Плотность распределения вероятностей $f(\lambda)$ и интегральный закон распределения $K(\lambda)$ случайной величины, распределенной по закону Колмогорова

Так же, как и распределение Стьюдента, это распределение непараметрическое, т.е. оно не зависит от какого-либо параметра.

В односторонних критериях согласия выборочной статистики Λ с заданным теоретическим распределением также используется распределение Смирнова-Колмогорова, у которого интегральная функция распределения $K(\lambda)$ следующая

$$K(\lambda) \;=\; \Pr\left(\Lambda < \lambda\right) \;=\; 1 - \exp(-2\lambda^2), \tag{2.56}$$

с отвечающей ей плотностью распределения вероятностей

$$f(\lambda) \;=\; \frac{d}{d\lambda}\,K(\lambda) \;=\; 4\lambda \exp(-2\lambda^2). \tag{2.57}$$

Это распределение также применяется для проверки гипотезы о том, что две выборки извлечены из одной и той же генеральной совокупности.

2.8. Распределение Бернулли

Случайная величина X называется *распределенной по биномиальному закону* (по закону Бернулли), если она является числом m появлений случайного события в n испытаниях, при условии, что в одном (любом) опыте это событие появляется с заданной вероятностью p.

Закон распределения случайной величины X

$$\mathrm{P}_{m,n} \;=\; \Pr\{\mathrm{X} = m\} \;=\; C_n^m p^m q^{n-m}, \tag{2.58a}$$

или

$$\mathrm{P}_{m,n} \;=\; \frac{n\,!}{m\,!\,(n-m)\,!}\,p^m q^{n-m}, \tag{2.58b}$$

где $0 < p < 1,\;\; q = 1 - p;\;\; m = 0, 1, \ldots, n.$

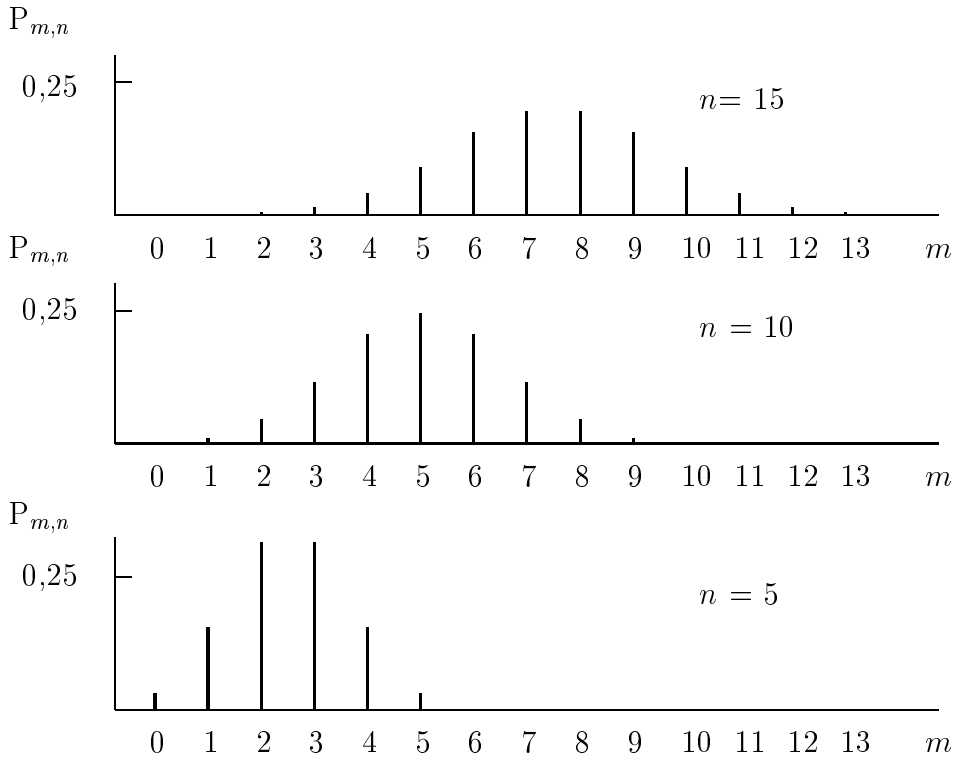


Рисунок 2.15 — Распределение Бернулли; $n = 5; 10; 15$; $p = 0{,}5$

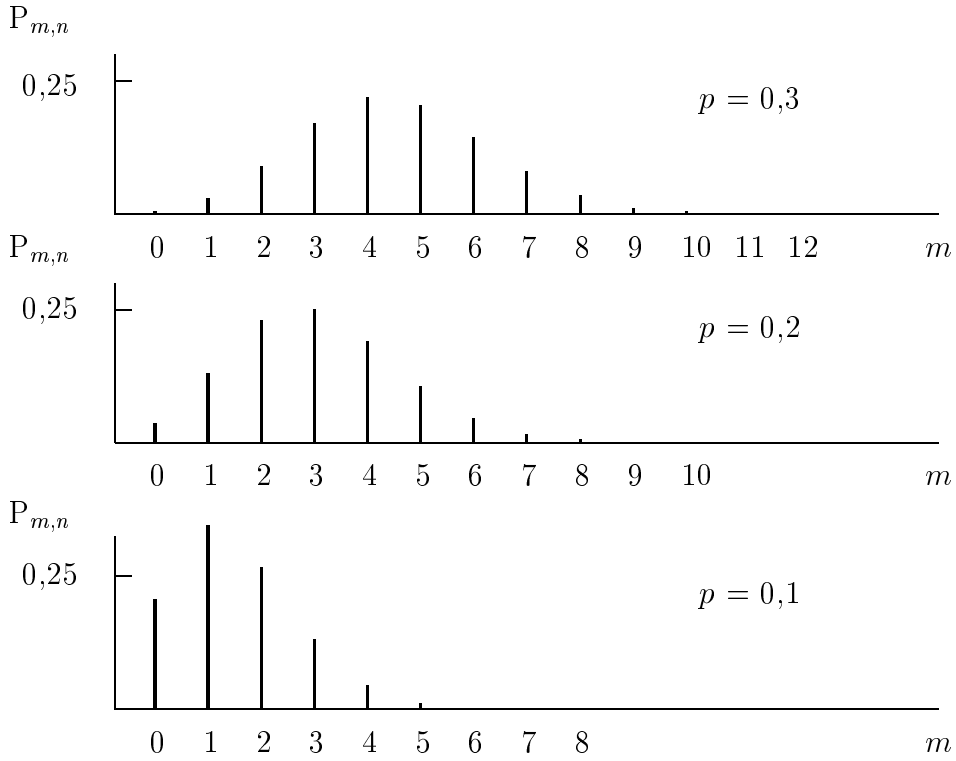


Рисунок 2.16 — Распределение Бернулли; $p = 0{,}1; 0{,}2; 0{,}3$; $n = 15$

Распределение (2.58) зависит от двух параметров: n (рис. 2.16) и p (рис. 2.17).

Из теоремы о повторении опытов следует, что *число* X *появлений события при* n *независимых опытах имеет биномиальное распределение.* Таким образом, вероятность одного сложного события, состоящего в том, что в n испытаниях событие A наступит m раз и не наступит $n-m$ раз, равна $\mathrm{P}_{m,n}$. Для случайной величины X, имеющей биномиальное распределение с параметрами p и n, выполняется

$$\mathrm{M[X]} \;=\; np, \qquad \mathrm{D[X]} \;=\; npq, \tag{2.59}$$

где $q = 1 - p$.

Пользоваться формулой Бернулли (2.58) при больших значениях n достаточно трудно, так как она требует выполнения действий над большими числами.

Имеют место асимптотические формулы, которые позволяют приближенно найти вероятность появления события ровно k раз в n испытаниях, если число испытаний достаточно велико. Соответствующие выражения формулируются в виде локальной и интегральной теорем Муавра–Лапласа.

Эта теорема записывается в локальной и в интегральной формах.

Локальная форма теоремы Муавра-Лапласа:

Если вероятность p *появления события* A *в каждом испытании постоянна и отлична от нуля и единицы, то вероятность* $\mathrm{P}_n(k)$ *того, что событие* A *появится в* n *испытаниях ровно* k *раз, приближенно равна*

$$\mathrm{P}_n(k) \;=\; \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{(k-np)^2}{2npq}\right), \tag{2.60}$$

тем точнее, чем больше число испытаний n.

Интегральная форма теоремы Муавра–Лапласа:

Если вероятность p *наступления события* A *в каждом испытании постоянна и отлична от нуля и единицы, то вероятность* $P_n(k_1, k_2)$ *того, что событие* A *появится в* n *испытаниях в диапазоне от* k_1 *до* k_2 *раз, приближенно равна*

$$\mathrm{P}_n(k_1, k_2) \;=\; \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} \exp(-z^2/2)\, dz, \tag{2.61}$$

где $\alpha = (k_1 - np)/\sqrt{npq}$ *и* $\beta = (k_2 - np)/\sqrt{npq}$.

2.9. Распределение Пуассона

Дискретная случайная величина X называется *распределенной по закону Пуассона*, если её возможные значения $0, 1, 2, \ldots, m, \ldots$, а вероятность события $\{\mathrm{X} = m\}$ выражается формулой

$$\mathrm{P}_m \;=\; \Pr\{\mathrm{X} = m\} \;=\; \frac{\lambda^m}{m\,!} \exp(-\lambda), \tag{2.62}$$

где $\lambda > 0$.

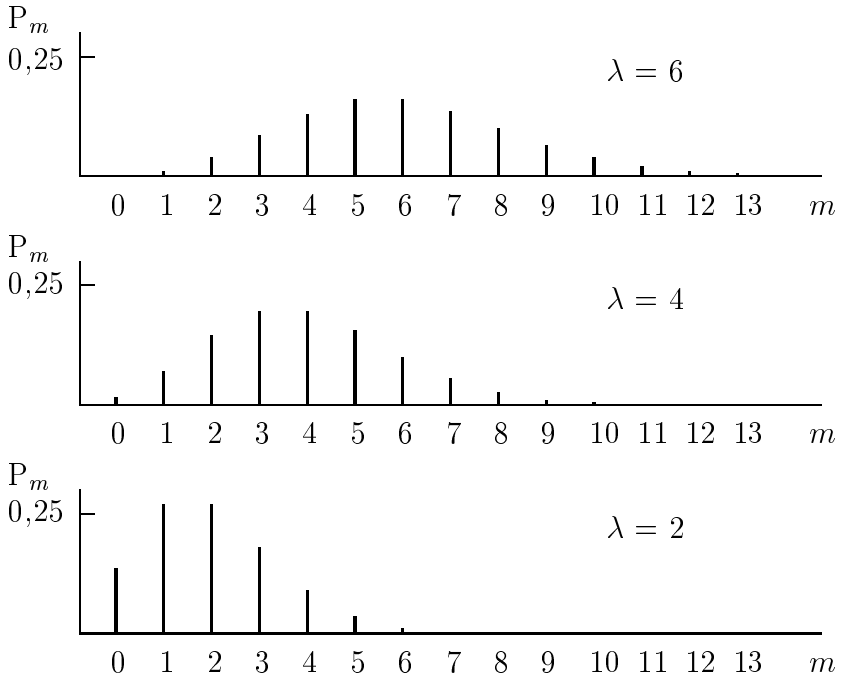


Рисунок 2.17 — Распределение Пуассона; $\lambda = 2;\ 4;\ 6$

Распределение Пуассона зависит от одного параметра λ (рис. 2.17).
Для случайной величины X, распределенной по закону Пуассона,

$$\mathrm{M[X]} = \lambda, \qquad \mathrm{D[X]} = \lambda. \tag{2.63}$$

Пуассоновское распределение является предельным для биномиального (2.58) при $p \to 0,\ \ n \to \infty$, если $np = \lambda = \mathrm{const}$. Этим распределением можно пользоваться приближенно, если производится большое число независимых опытов, в каждом из которых событие A происходит с малой вероятностью.

Пуассоновскому закону распределения подчиняется также количество точек, попадающих в заданную область пространства (одномерного, двумерного или трехмерного), если случайное расположение точек в этом пространстве удовлетворяет некоторым ограничениям.

Одномерный вариант встречается при рассмотрении "потоков событий". *Потоком событий* называется последовательность однородных событий, наступающих одно за другим в случайные моменты времени.

Среднее число событий λ, приходящихся на единицу времени, называется *интенсивностью потока*. Величина λ может быть как постоянной, так и переменной : $\lambda = \lambda(t)$.

Поток событий называется *потоком без последействий*, если вероятность попадания того или иного числа событий на какой-то участок времени не зависит от того, сколько событий попало на любой другой непересекающийся с ним участок.

Поток событий называется *ординарным*, если вероятность появления на элементарном участке двух или более событий пренебрежимо мала по сравнению с вероятностью появления одного события.

Ординарный поток событий без последействий называется *пуассоновским.* Если события образуют пуассоновский поток, то число X событий, попадающих на любой участок времени $(t,\, t+\tau)$, распределено по закону Пуассона:

$$\mathrm{P}_m \;=\; \frac{a^m}{m\,!}\exp(-a), \tag{2.64}$$

где a – математическое ожидание числа точек, попадающих на участок:

$$a \;=\; \int\limits_{t}^{t+\tau} \lambda(t)\,dt. \tag{2.65}$$

Если $\lambda(t) = \lambda = \mathrm{const}$, пуассоновский поток называется стационарным пуассоновским или *простейшим.* Для простейшего потока число событий, попадающих на любой участок времени продолжительностью τ, распределено по закону Пуассона с параметром $a = \lambda\tau$.

Случайным полем точек называется совокупность точек, случайным образом разбросанных на плоскости (или в пространстве).

Интенсивностью (или плотностью) поля λ называется среднее число точек, попадающих в единицу площади (объема).

Поле точек называется *пуассоновским*, если оно обладает следующими свойствами:

1) вероятность попадания того или иного числа точек в любую область плоскости (пространства) не зависит от того, сколько их попало в любую область, не пересекающуюся с данной;

2) вероятность попадания в элементарную область $\Delta x\,\Delta y$ двух или более точек пренебрежимо мала по сравнению с вероятностью попадания одной точки (свойство ординарности).

Число X точек пуассоновского поля, попадающих в любую область S плоскости (пространства), распределено по закону Пуассона:

$$\mathrm{P}_m \;=\; \Pr\{\mathrm{X} = m\} \;=\; \frac{a^m}{m\,!}\exp(-a), \qquad\qquad m = 0, 1, 2, \ldots, \tag{2.66}$$

где a – математическое ожидание числа точек, попадающих в область S. Если интенсивность поля постоянна, $\lambda(x, y) = \lambda = \mathrm{const}$, поле называется *однородным* (свойство, аналогичное стационарности потока событий). При однородном поле с интенсивностью λ имеем $a = S\lambda$, где S – площадь области, или $a = V\lambda$, где V – объем области.

Если поле неоднородно, то

$$a \;=\; \iint\limits_{(S)} \lambda(x, y)\,dx\,dy \qquad\qquad \text{– для плоскости;} \tag{2.67a}$$

$$a \;=\; \iiint\limits_{(V)} \lambda(x, y, z)\,dx\,dy\,dz \qquad\qquad \text{– для объема.} \tag{2.67b}$$

В задачах математической статистики (при оценивании или проверке статистических гипотез) используются также и другие законы распределения.

2.10. Примеры

Пример 2.1

Пусть на производстве расход материала носит случайный характер со средней интенсивностью 20 единиц в день. Для покрытия расхода производятся ежемесячные поставки объемом в 640 единиц.

Определить:

а) вероятность образования дефицита (нехватки) материалов;

б) объем (величину) поставок, при котором вероятность β возникновения дефицита не превысит 0,01.

Решение

Обозначим через T промежуток времени, в течение которого суммарный расход окажется равным объему поставки β. Величина T является случайной, подчиняющейся гамма–распределению с параметрами β и λ. Дефицит образуется, если T окажется меньше заданного интервала между поставками, т.е. при $\mathrm{T} \le 30$. Из определения функции распределения имеем $\Pr(\mathrm{T} \le 30) = F_{\mathrm{T}}(30;\, \beta,\, \lambda)$.

а) Задаваясь в (2.34) $\beta = 640$, $\lambda = 20$ и $t = 30$, получим

$$\Pr(\mathrm{T} \le 30) \;=\; F_{\mathrm{T}}(30;\, 640,\, 20) \;=\; \frac{20^{640}}{\Gamma(640)} \int\limits_0^{30} t^{639} \mathrm{e}^{-20t}\, dt \;=\; 0,0569.$$

Итак, дефицит будет иметь место с вероятностью 0,0569.

б) По заданной величине $p = \Pr(\mathrm{T} \le 30) = 0{,}01 = F_{\mathrm{T}}(30;\, \beta,\, 20)$ находим $\beta = 660$ (единиц). При таком объеме поставок вероятность дефицита не превысит 0,01. Соответственно, с вероятностью 0,99 дефицита за рассматриваемый временной интервал не будет.

Пример 2.3

Найти математические ожидания, дисперсии и коэффициент корреляции двумерной нормальной случайной величины.

Решение

Запишем выражение для плотности распределения $f_{\mathrm{XY}}(x, y)$ нормальной системы из двух случайных величин (X, Y)

$$f(x,y) \;=\; \frac{1}{2\pi\, \sigma_{\mathrm{X}}\, \sigma_{\mathrm{Y}}\, \sqrt{1 - r_{\mathrm{XY}}^2}} \exp\left\{ -\; \frac{1}{2(1 - r_{\mathrm{XY}}^2)}\, Q(x,y) \right\}, \qquad (*)$$

$$Q(x,y) \;=\; \frac{(x - m_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2} - \; 2r_{\mathrm{XY}} \frac{(x - m_{\mathrm{X}})(y - m_{\mathrm{Y}})}{\sigma_{\mathrm{X}} \sigma_{\mathrm{Y}}} + \frac{(y - m_{\mathrm{Y}})^2}{\sigma_{\mathrm{Y}}^2},$$

где m_{X}, m_{Y}, σ_{X}, σ_{Y}, r_{XY} – параметры закона.

1) Очевидно, что

$$\mathrm{M}_{\mathrm{XY}}[\mathrm{X}] \;=\; \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} x\, f_{\mathrm{XY}}(x,y)\, dx\, dy \;=\; \int\limits_{-\infty}^{\infty} x\, f_{\mathrm{X}}(x)\, dx \;=\; \mathrm{M}_{\mathrm{X}}[\mathrm{X}],$$

где $f_{\mathrm{X}}(x)$ – парциальная плотность X-компоненты:

$$f_{\mathrm{X}}(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_{\mathrm{X}}}\exp\left\{-\frac{1}{2}\,\frac{(x-m_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2}\right\},$$

получающаяся из выражения $(*)$ интегрированием по переменной y.

Интегрируя, получим для первых двух моментов X-компоненты

$$\mathrm{M}_{\mathrm{XY}}[\mathrm{X}] \;=\; m_{\mathrm{X}}, \qquad \mathrm{D}_{\mathrm{XY}}[\mathrm{X}] \;=\; \mathrm{D}_{\mathrm{X}}[\mathrm{X}] \;=\; \sigma_{\mathrm{X}}^2.$$

2) Аналогично поступая для Y-компоненты, найдем

$$\mathrm{M}_{\mathrm{XY}}[\mathrm{Y}] \;=\; m_{\mathrm{Y}}, \qquad \mathrm{D}_{\mathrm{XY}}[\mathrm{Y}] \;=\; \mathrm{D}_{\mathrm{Y}}[\mathrm{Y}] \;=\; \sigma_{\mathrm{Y}}^2.$$

3) Запишем выражение для смешанного второго момента

$$k_{\mathrm{XY}} \;=\; \mathrm{M}_{\mathrm{XY}}[(\mathrm{X}-m_{\mathrm{X}})(\mathrm{Y}-m_{\mathrm{Y}})] \;=\; \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}(x-m_{\mathrm{X}})(y-m_{\mathrm{Y}})\,f_{\mathrm{XY}}(x,y)\,dx\,dy.$$

В возникшем двойном интеграле перейдем к новым переменным интегрирования (u,v) по правилу: $u=(x-m_{\mathrm{X}})/\sigma_{\mathrm{X}}$, $v=(y-m_{\mathrm{Y}})/\sigma_{\mathrm{Y}}$, что дает

$$k_{\mathrm{XY}} \;=\; \frac{1}{2\pi\sqrt{1-r_{\mathrm{XY}}^2}}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}uv\,\exp\left\{-\frac{u^2-2r_{\mathrm{XY}}uv+v^2}{2(1-r_{\mathrm{XY}}^2)}\right\}du\,dv.$$

Теперь удобно вместо переменной u ввести ещё одну переменную интегрирования по правилу $t=u-r_{\mathrm{XY}}v$:

$$k_{\mathrm{XY}} \;=\; \frac{1}{2\pi\sqrt{1-r_{\mathrm{XY}}^2}}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}(t+r_{\mathrm{XY}}v)v\exp\left\{-\frac{t^2+(1-r_{\mathrm{XY}}^2)v^2}{2(1-r_{\mathrm{XY}}^2)}\right\}dt\,dv.$$

В этом интеграле вклад от первого слагаемого в круглых скобках $(t+r_{\mathrm{XY}}v)$ тождественно равен нулю из-за нечетности подинтегрального выражения и четности пределов интегрирования относительно переменной t. Оставшийся двойной интеграл становится произведением двух однократных интегралов:

$$k_{\mathrm{XY}} \;=\; \frac{1}{\sqrt{2\pi(1-r_{\mathrm{XY}}^2)}}\int\limits_{-\infty}^{\infty}\exp\left\{-\frac{t^2}{2(1-r_{\mathrm{XY}}^2)}\right\}dt\;\frac{r_{\mathrm{XY}}}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}v^2\,\exp\left\{-\frac{v^2}{2}\right\}dv.$$

С учетом нормировочного множителя первый из этих интегралов тождественно равен 1.

Второй интеграл (без сомножителя r_{XY}) совпадает со значением дисперсии переменной v, которая также равна 1. Итак,

$$k_{\mathrm{XY}} \;=\; r_{\mathrm{XY}}.$$

Таким образом, закон двумерного распределения $(*)$ полностью определяется заданием его числовых характеристик m_{X}, m_{Y}, σ_{X}, σ_{Y} и r_{XY}.

Пример 2.3

Для случайной величины X, подчиняющейся: а) биномиальному распределению, б) распределению Пуассона, найти её характеристическую функцию $q(\lambda)$.

Решение

Воспользуемся определением характеристической функции

$$q(\lambda) \;=\; \mathrm{M}[\exp(i\lambda \mathrm{X})] \;=\; \sum_i \mathrm{P}_i \exp(i\lambda x_i).$$

Для биномиального распределения с параметрами p и n имеем

$$q(\lambda) \;=\; \left(p\mathrm{e}^{i\lambda} + 1 - p\right)^n.$$

Для распределения Пуассона с параметром a найдем

$$q(\lambda) \;=\; \exp\left[a\left(\mathrm{e}^{i\lambda} - 1\right)\right].$$

Пример 2.4 (Распределение χ^2 с n степенями свободы)

Задан набор независимых нормальных величин $\{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n\}$ с параметрами $\{m_1 = 0, m_2 = 0, \ldots, m_n = 0\}$ и $\{\sigma_1 = 1, \sigma_2 = 1, \ldots, \sigma_n = 1\}$. Рассмотрим случайную величину

$$\mathrm{Y} \;=\; \sum_{i=1}^{n} \mathrm{X}_i^2.$$

Получить закон распределения случайной величины Y.

Решение

Пусть сначала $n = 1$. В этом случае характеристическая функция $q_{\mathrm{Y}_1}(\lambda)$ равна

$$q_{\mathrm{Y}_1}(\lambda) \;=\; \mathrm{M}[\exp(i\lambda \mathrm{Y}_1)] \;=\; \mathrm{M}[\exp(i\lambda \mathrm{X}_1^2)] \;=\; \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-x_1^2/2 + i\lambda x_1^2\right) dx_1.$$

Интегрируя, найдем

$$q_{\mathrm{Y}_1}(\lambda) \;=\; (1 - 2i\lambda)^{-1/2}.$$

Поскольку по условию компоненты СВ Y независимы, то в случае, когда она образована из n слагаемых, получим

$$q_{\mathrm{Y}}(\lambda) \;=\; \prod_{i=1}^{n} q_{\mathrm{Y}_i}(\lambda) \;=\; \left[q_{\mathrm{Y}_1}(\lambda)\right]^n \;=\; (1 - 2i\lambda)^{-n/2}.$$

Плотность распределения $f_{\mathrm{Y}}(t)$ СВ Y можно определить с помощью обратного преобразования Фурье, что дает

$$f_{\mathrm{Y}}(t) \;=\; Ct^{\frac{n}{2}-1}\mathrm{e}^{-t/2}, \qquad t \geq 0,$$

где C — постоянная. Её определим из условия нормировки $\int_0^\infty f_{\mathrm{Y}}(t)\,dt \;=\; 1$, откуда $C^{-1} = 2^{n/2}\,\Gamma(n/2)$. Итак,

$$f_{\mathrm{Y}}(t) \;=\; \frac{1}{2^{n/2}\,\Gamma(n/2)}\, t^{n/2-1}\mathrm{e}^{-t/2}.$$

Найденная плотность отвечает случайной величине, распределенной по закону χ^2 с n степенями свободы.

Пример 2.5 (Распределение Стьюдента)

Задан набор независимых нормальных величин $\{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n\}$ с параметрами $\{m_1 = 0, m_2 = 0, \ldots, m_n = 0\}$ и $\{\sigma_1 = 1, \sigma_2 = 1, \ldots, \sigma_n = 1\}$. Кроме того, имеется еще одна независимая нормальная величина X_0 с параметрами $m_0 = 0$ и $\sigma_0 = 1$. Рассмотрим случайную величину

$$\mathrm{T} \;=\; \mathrm{X}_0 \left(\frac{1}{n} \sum_{i=1}^{n} \mathrm{X}_i^2 \right)^{-1/2}.$$

Получить закон распределения случайной величины T.

Решение

Рассматриваемая случайная величина T является стандартизованным отношением Стьюдента

$$\mathrm{T} \;=\; \frac{\mathrm{X}_0}{\sqrt{\chi^2/n}}.$$

Для нахождения плотности распределения $f_{\mathrm{T}}(t)$ воспользуемся техникой δ-функции Дирака

$$f_{\mathrm{T}}(t) \;=\; \mathrm{M}\left[\delta\left(\frac{\mathrm{X}_0}{\sqrt{\chi^2/n}} - t \right) \right].$$

Здесь математическое ожидание находят относительно СВ χ^2 и СВ X_0 с плотностями распределений

$$f_{\mathrm{X}_0}(x_0)\, dx_0 \;=\; \frac{1}{2\pi} \exp\left(-x_0^2/2\right) dx_0,$$

$$f_{\chi^2}(\chi^2)\, d\chi^2 \;=\; \frac{1}{2^{n/2}\,\Gamma(n/2)} (\chi^2)^{\frac{n}{2}-1} \mathrm{e}^{-\chi^2/2} \chi^2\, d\chi^2.$$

Подставим эти плотности в выражение для $f_{\mathrm{T}}(t)$ и учтем фильтрующее свойство δ-функции, тогда получим

$$f_{\mathrm{T}}(t) \;=\; \frac{1}{2^{n/2}\,\Gamma(n/2)} \frac{1}{n} \int_0^\infty (\chi^2)^{n-1} \exp\left(-\frac{1}{2} \frac{t^2\chi^2}{n} - \frac{\chi^2}{2} \right) d\chi^2.$$

Используя теперь новую переменную интегрирования u по правилу

$$\frac{1}{2} \frac{t^2\chi^2}{n} + \frac{\chi^2}{2} \;=\; u,$$

получим

$$f_{\mathrm{T}}(t) \;=\; \left[2^{n/2} \sqrt{n} \left(\frac{1 + t^2/n}{2} \right) \right]^{-1} \int_0^\infty u^{(n-1)/2} \mathrm{e}^{-u}\, du.$$

Пользуясь определением Г-функции, окончательно найдем

$$f_{\mathrm{T}}(t) \;=\; \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\,\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}.$$

Найденная плотность отвечает случайной величине, распределенной по закону Стьюдента.

Пример 2.6 (Распределение Фишера–Снедекора)

Заданы две случайные величины S и T, подчиняющиеся закону χ^2 соответственно с n_1 и n_2 степенями свободы и имеющие один и тот же параметр σ.

Получить закон распределения случайной величины $\mathrm{U} = \mathrm{S}/\mathrm{T}$.

Решение

Закон распределения исходных величин следующий (здесь C_1 и C_2 – нормировочные константы):

$$f_{\mathrm{S}}(s)\,ds \;=\; C_1\,\left(\frac{s}{\sigma}\right)^{n_1-1}\exp\left(-\frac{s^2}{2\sigma^2}\right)\frac{ds}{\sigma}, \qquad s\ge 0,$$

$$f_{\mathrm{T}}(t)\,dt \;=\; C_2\,\left(\frac{t}{\sigma}\right)^{n_2-1}\exp\left(-\frac{t^2}{2\sigma^2}\right)\frac{dt}{\sigma}, \qquad t\ge 0.$$

Чтобы определить форму закона величины U, рассмотрим распределение пары S и T. Введем случайные величины R, U и сделаем замену переменных

$$s^2+t^2 \;=\; r^2, \qquad \frac{t}{s} \;=\; u, \qquad u,r\ge 0,$$

или

$$s \;=\; \frac{r}{\sqrt{1+u^2}}, \qquad t \;=\; \frac{ru}{\sqrt{1+u^2}}.$$

Отсюда

$$\left|\frac{D(s,t)}{D(r,u)}\right| \;=\; \frac{r}{\sqrt{1+u^2}}.$$

Тогда элемент дифференциальной вероятности $f_{\mathrm{R,U}}(r,u)\,dr\,du$ пары (R,U) запишется в виде (C_3 – нормировочная константа)

$$f_{\mathrm{R,U}}(r,u)\,dr\,du \;=\; C_3\,\exp\left(-\frac{r^2}{2\sigma^2}\right)\left(\frac{r}{\sigma}\right)^{n_1+n_2-1}\frac{dr}{\sigma}\,\frac{u^{n_2-1}\,du}{(1+u^2)^{(n_1+n_2)/2}}.$$

Отсюда следует, что величины R и U – независимы, при этом величина U подчиняется закону χ^2 с n_1+n_2 степенями свободы.

Проинтегрировав последнее выражение по r, получим элемент дифференциальной вероятности СВ U в форме

$$f_{\mathrm{U}}(u)\,du \;=\; C_4\,\frac{u^{n_2-1}}{(1+u^2)^{(n_1+n_2)/2}}\,du,$$

где C_4 – нормировочная константа. Эту постоянную найдем из условия нормировки $\int\limits_0^\infty f_{\mathrm{U}}(u)\,du = 1$. Итак, плотность распределения СВ U следующая:

$$f_{\mathrm{U}}(u) \;=\; \frac{2\,\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\,\Gamma\left(\frac{n_2}{2}\right)}\,\frac{u^{n_2-1}}{(1+u^2)^{(n_1+n_2)/2}}.$$

Найденная плотность отвечает случайной величине, распределенной по закону Фишера-Снедекора.

Пример 2.7

В задачах оценивания характеристик по ограниченному числу наблюдений n, идентификации процессов и др. представляется целесообразным использование статистики ($\mathrm{X}^{(n)}_{\max}$ и $\mathrm{X}^{(n)}_{\min}$ – крайние члены вариационного ряда)

$$\mathrm{R} \;=\; \mathrm{X}^{(n)}_{\max} - \mathrm{X}^{(n)}_{\min},$$

называемой *размахом* или *широтой* выборки. В общем виде распределение размаха определяется следующим образом:

$$\Pr\left((X^{(n)}_{\max} - X^{(n)}_{\min}) < t\right) \;=\; n\int\limits_{-\infty}^{\infty}[F_{\mathrm{X}}(x+t) - F_{\mathrm{X}}(x)]^{n-1}\,dF_{\mathrm{X}}(x).$$

Используя это результат, приведем (без доказательства) несколько утверждений, которые оказываются полезными в теории оценивания.

Утверждение 1. Для выборки из равномерно распределенной генеральной совокупности с плотностью распределения

$$f_{\mathrm{X}}(x) \;=\; \frac{1}{b-a}, \qquad\qquad x \in (a,\,b),$$

плотность распределения размаха R имеет вид

$$f_{\mathrm{R}}(t) \;=\; n(n-1)\,\frac{t^{n-2}}{(b-a)^{n-1}}.$$

Утверждение 2. Пусть Y – центрированная и равномерно распределенная случайная величина $\mathrm{Y} \in (-a,\,a)$. Тогда плотность распределения её размаха R имеет вид

$$f_{\mathrm{R}}(t) \;=\; \frac{1}{2a}.$$

Утверждение 3. Пусть случайная величина X имеет нормальное распределение с нулевым математическим ожиданием и дисперсией σ^2_{X}, а случайная величина Y является центрированной равномерно распределенной случайной величиной, $\mathrm{Y} \in$

$(-a, a)$. Введем в рассмотрение статистику $\mathrm{T} = \mathrm{X}/\mathrm{Y}$. Тогда плотность распределения статистики T определяется зависимостью

$$f_{\mathrm{T}}(t) \;=\; \frac{4\sigma_{\mathrm{X}}^2}{a\sqrt{\pi}}\, t^{-2}\, \left[1 - \exp\left(-\frac{a^2 t^2}{4\sigma_{\mathrm{X}}^2}\right)\right].$$

Утверждение 4. Пусть имеется выборка объемом n из нормальной генеральной совокупности с параметрами m_{X}, σ_{X}. Тогда закон распределения размаха R выборки определяется следующим образом:

$$F_{\mathrm{R}}(t) \;=\; \Pr\left((X_{\max}^{(n)} - X_{\min}^{(n)}) < t\right) \;=\; n \int\limits_{-\infty}^{\infty} [F(x+t) - F(x)]^{n-1}\, dF(x),$$

где $F(x)$ – интегральный закон распределения нормальной величины.

Дифференцируя под знаком интеграла, получим для плотности распределения размаха этой выборки

$$f_{\mathrm{R}}(t) \;=\; \frac{n(n-1)}{2\pi\sigma_{\mathrm{X}}^2} \int\limits_{-\infty}^{\infty} [F(x+t) - F(x)]^{n-2} \exp\left(-\frac{(x - m_{\mathrm{X}})^2}{2\sigma_{\mathrm{X}}^2}\right) dx.$$

2.11. Задачи для решения

Задача 2.1

В некотором районе в личном владении находится 3500 коров. В порядке случайной выборки обследовали 800 коров и установили, что у этой группы средний годовой удой равен 2800 *кг*, а среднее квадратическое отклонение $\sigma = 250$ *кг*.

С какой вероятностью можно гарантировать, что средний годовой удой всех коров отличается от 2800 *кг* по абсолютной величине меньше чем 10 *кг*?

Задача 2.2

Партия изделий считается годной к выпуску, если брак в ней не превышает 3 %. Из партии в 2000 изделий было отобрано и проверено 400. При этом бракованных изделий оказалось 6.

Какова вероятность того, что вся партия удовлетворяет техническим условиям и может быть принята?

Задача 2.3

Даны две выборки: (0,79; 0,75; 0,86; 0,05; 1,29; 0,42; 1,12; 0,70; 1,54; 1,43) и (2,05; 1,38; 1,45; 0,35; 0,64; 0,58; 1,03; 0,12; 1,30; 1,09; 0,33) из двух нормальных распределений со средними a и b и одинаковой дисперсией σ^2, все параметры неизвестны.

При уровне доверия $p = 0,95$ построить доверительный интервал для разности средних $\gamma = a - b$.

Задача 2.4

Сколько лиц в возрасте от 20 до 25 лет надо опросить выборочно, чтобы установить среди них процент студентов с точностью до 0,5 %, гарантируемой с вероятностью 0,999 ?

Задача 2.5

Найти математическое ожидание и дисперсию случайной величины $\mathrm{Z} = |\mathrm{X}|$, если СВ X распределена по нормальному закону с параметрами m, σ^2.

Ответ: $\mathrm{M[Z]} = \dfrac{2\sigma}{\sqrt{2\pi}} \exp\left(-\dfrac{m^2}{2\sigma^2}\right) + m\Phi\left(\dfrac{m}{\sigma}\right)$, $\mathrm{D[Z]} = \sigma^2 + m^2 - (\mathrm{M[Z]})^2$.

Задача 2.6

Случайная точка (X, Y) распределена равномерно в круге радиусом 1 с центром в точке $(0; 0)$.

Найти законы распределения случайных величин U=X/Y и V=Y/X.

Ответ: $f_{\mathrm{U}}(u) = [\pi(1+u^2)]^{-1}$, $f_{\mathrm{V}}(v) = [\pi(1+v^2)]^{-1}$.

Задача 2.7

Непрерывная случайная величина X имеет закон распределения $F_{\mathrm{X}}(x)$, а связанная с ней случайная величина Y определяется законом распределения

$$F_{\mathrm{Y}}(y) \;=\; \Pr(\mathrm{Y} < y) \;=\; n \int\limits_{-\infty}^{\infty} [F_{\mathrm{X}}(x+y) - F_{\mathrm{X}}(x)]^{n-1} f_{\mathrm{X}}(x)\, dx,$$

где $n = 2, 3, \ldots$ – параметр.

Для различных случайных величин X найти математическое ожидание и дисперсию СВ Y.

Задача 2.8

Система случайных величин (X, Y) подчинена нормальному закону распределения

$$f_{\mathrm{XY}}(x, y) \;=\; \frac{1}{2\pi\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} \exp\left(-\frac{x^2}{2\sigma_{\mathrm{X}}^2} - \frac{y^2}{2\sigma_{\mathrm{Y}}^2}\right).$$

Найти закон распределения случайных величин $\mathrm{Z}_1 = \mathrm{X} + \mathrm{Y}$ и $\mathrm{Z}_2 = \mathrm{X} - \mathrm{Y}$.

Ответ: Случайные величины Z_1 и Z_2 имеют одинаковый закон распределения с плотностью

$$f_{\mathrm{Z}}(z) \;=\; \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{Z}}} \exp\left(-\frac{z^2}{2\sigma_{\mathrm{Z}}^2}\right), \qquad \sigma_{\mathrm{Z}}^2 = \sigma_{\mathrm{X}}^2 + \sigma_{\mathrm{Y}}^2.$$

Задача 2.9

Система случайных величин (X, Y) подчинена нормальному закону распределения

$$f_{\mathrm{XY}}(x, y) \;=\; \frac{1}{2\pi\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\left[\frac{x^2}{\sigma_{\mathrm{X}}^2} - 2\rho\,\frac{xy}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} + \frac{y^2}{\sigma_{\mathrm{Y}}^2}\right]\right).$$

Определить закон распределения случайной величины $\mathrm{Z} = \mathrm{X}/\mathrm{Y}$. Рассмотреть случай, когда $\rho = 0$.

2.12. Задание на практическую работу

Настоящая практическая работа рассчитана на два часа и содержит два задания. Задания должны выполняться в выбранной программной среде.

Задание 1

Напишите программу визуализации плотности распределения и интегрального закона распределения для указанных случаев.

Результаты оформите графически.

Вариант 1

Равномерный закон.

Исходные данные для программы:

a – левая граница возможных значений;

b – правая граница возможных значений;

$f_{\mathrm{X}}(x)$ – плотность распределения вероятностей;

$F_{\mathrm{X}}(x)$ – интегральный закон распределения вероятностей.

Результат работы программы – массив, содержащий значения плотности распределения вероятностей $f_{\mathrm{X}}(x)$ и интегрального закона распределения вероятностей $F_{\mathrm{X}}(x)$ в выбранных узлах аргумента x случайной величины X.

Вариант 2

Нормальный закон Гаусса.

Исходные данные для программы:

m_{X} – математическое ожидание;

σ^2_{X} – дисперсия;

$f_{\mathrm{X}}(x)$ – плотность распределения вероятностей;

$F_{\mathrm{X}}(x)$ – интегральный закон распределения вероятностей.

Результат работы программы – массив, содержащий значения плотности распределения вероятностей $f_{\mathrm{X}}(x)$ и интегрального закона распределения вероятностей $F_{\mathrm{X}}(x)$ в выбранных узлах аргумента x случайной величины X.

Вариант 3

Распределение χ^2.

Исходные данные для программы:

n – число степеней свободы;

$f(\chi^2)$ – плотность распределения вероятностей;

$F(\chi^2)$ – интегральный закон распределения вероятностей.

Результат работы программы – массив, содержащий значения плотности распределения вероятностей $f(\chi^2)$ и интегрального закона распределения вероятностей $F(\chi^2)$ в выбранных узлах аргумента x случайной величины χ^2_n.

Задание 2

Напишите программу визуализации последовательности амплитуд распределения вероятностей P_m для указанных случаев.

Результат работы программы – массив, содержащий значения искомой функции. Результаты оформите графически.

Вариант 1

Равномерный дискретный закон.

Исходные данные для программы:

n_1 – левая граница возможных значений;

n_2 – правая граница возможных значений;

P_m – последовательность амплитуд распределения вероятностей.

Результат работы программы – массив, содержащий значения последовательности амплитуд равномерного распределения вероятностей P_m в узлах дискретного аргумента $n_1 \leq m \leq n_2$ случайной величины X.

Вариант 2

Распределение Бернулли.

Исходные данные для программы:

n – общее число испытаний;

p – вероятность появления благоприятного события за одно испытание.

P_m – последовательность амплитуд распределения вероятностей.

Результат работы программы – массив, содержащий значения последовательности амплитуд распределения вероятностей Бернулли P_m в узлах дискретного аргумента $0 \leq m \leq n$ случайной величины X.

Вариант 3

Распределение Пуассона.

Исходные данные для программы:

a – математическое ожидание.

P_m – последовательность амплитуд распределения вероятностей.

Результат работы программы – массив, содержащий значения последовательности амплитуд распределения вероятностей Пуассона P_m в узлах дискретного аргумента $m \geq 0$ случайной величины X.

2.13. Задания для проверки

1. Перечислите основные характеристики одномерного нормального закона, двумерного нормального закона, многомерного нормального закона.

2. Какая функция называется гамма-функцией?

3. Перечислите основные свойства гамма-функции.

4. По какому закону распределена сумма квадратов случайных величин, каждая из которых имеет стандартизованное нормальное распределение?

5. Что означает число степеней свободы случайной величины χ^2 ?

6. Какая из рассмотренных случайных величин называется безразмерной дробью Стьюдента?

7. Какой закон распределения имеет отношение суммы квадратов нормальных стандартизованных случайных величин?

8. Какой смысл имеет число степеней свободы t-распределения? Раскройте смысл числа степеней свободы случайной величины, имеющей распределение Фишера.

9. Как определяется квантиль χ^2-распределения, квантиль t-распределения, квантиль F-распределения?

10. Какой смысл имеет параметр корреляции двумерного нормального закона?

11. Какой вид имеет характеристическая функция одномерного нормального закона, двумерного нормального закона, многомерного нормального закона?

12. Приведите выражение для характеристической функции χ^2-распределения, t-распределения Стьюдента, гамма-распределения.

13. Какое распределение является предельным (при $n \to \infty$) для χ^2-распределения, для t-распределения, для F-распределения?

3. Статистическая теория оценивания параметров распределения

3.1. Постановка задачи оценивания

Предположим, что для оценки закона распределения исследуемой случайной величины X из генеральной совокупности с неизвестной функцией распределения $F_{\text{X}}(x)$ извлечена выборка $x_1, x_2, \ldots, x_n$. Предположим также, что экспериментатор визуально по виду гистограммы или полигона частостей или на основании каких-либо других соображений выбрал класс Ω функций определенного вида (нормальных, показательных, биномиальных и т.д.), к которому, по его мнению, может принадлежать функция распределения исследуемой случайной величины X.

Наиболее часто при исследованиях непрерывных случайных величин экспериментатор старается выбрать класс нормальных функций, т.е. построить нормальную модель совокупности, так как эта модель наиболее разработана в аналитическом отношении. Кроме того, нормальный закон является устойчивым при композиции и является предельным законом многих законов распределения. Поэтому класс нормальных функций часто принимают за приближенную модель генеральной совокупности.

После того как класс функций Ω выбран, производится *оценка* (подгонка) параметров внутри выбранного класса функций. Например, если выбран нормальный класс функций для описания исследуемой случайной величины X, то по выборке $x_1, x_2, \ldots, x_n$ требуется оценить два параметра — математическое ожидание m_{X} и среднее квадратическое отклонение σ_{X}, от которых зависит нормальное распределение.

Если выбран класс функций Пуассона, то на основании выборки $x_1, x_2, \ldots, x_n$ требуется оценить только один параметр, которым определяется закон Пуассона – интенсивность λ.

Пусть из генеральной совокупности с функцией распределения $F_{\text{X}}(x;\theta)$, где θ – неизвестный параметр, произведена выборка объема n и получены результаты $x_1, x_2, \ldots, x_n$. Вообще говоря, по результатам выборки, какого бы большого объема она ни была, нельзя определить точное значение неизвестного параметра θ, а можно лишь найти его приближенное значение $\hat{\theta}$, которое и называется оценкой.

Для нахождения приближенных значений (оценки) неизвестного параметра θ будем рассматривать функции вида

$$\hat{\theta} \;=\; u\left(x_1, x_2, \ldots, x_n\right), \tag{3.1}$$

которые называются *выборочными функциями* или *статистиками*. Задача оценки неизвестного параметра θ сводится к нахождению таких выборочных функций

$\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, которые можно использовать в качестве оценки неизвестного параметра θ.

Любая выборка является ограниченной и случайной. Следовательно, все выборочные функции $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ также являются случайными. Например, оценкой математического ожидания является среднее арифметическое $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Если повторить опыт k раз, извлекая из генеральной совокупности случайные выборки одного и того же объема n, то по их данным найдем ряд значений $\{\overline{x}_k\}$, которые, вообще говоря, будут отличаться друг от друга.

Таким образом, будем рассматривать оценку $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ неизвестного параметра θ как случайную величину, а её значение, вычисленное на основании данной выборки объема n, – как одну реализацию случайной величины, т.е. как одно из множества возможных значений этой случайной величины.

Возможные оценки параметров подразделяются на точечные и интервальные. *Точечная оценка* параметра θ определяется одним числом $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$. *Интервальной оценкой* параметра θ называют оценку, которая определяется двумя числами $\hat{\theta}_1$ и $\hat{\theta}_2$ – концами интервала, накрывающего оцениваемый параметр.

3.2. Непараметрическое и параметрическое оценивание. Статистические оценки и их свойства

Из вышеизложенного следует, что оценкой неизвестного параметра θ является функция $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, зависящая от наблюденных значений случайной величины, которая может быть использована для нахождения приближенного значения неизвестного параметра θ. Таким образом, далее будем рассматривать только определенные классы функций, близкие в определенном смысле к оцениваемому параметру θ. Имеется специальный раздел математической статистики — *теория оценивания*, в которой занимаются выработкой правил конструирования функции $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ для нахождения точечных оценок неизвестных параметров.

Перейдем к основным свойствам, которые должны иметь "хорошие" оценки неизвестного параметра $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$.

Прежде всего, с точки зрения точности и надежности оценок желательно, чтобы найденные на основании выборочных функций $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ оценки неизвестных параметров по возможности были тесно сконцентрированы около значений оцениваемых параметров, другими словами, чтобы рассеивание случайной величины $\hat{\theta}$ около θ было по возможности наименьшим.

Определение 1. Оценка $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ называется *состоятельной*, если при увеличении числа измерений оценка сходится по вероятности к оцениваемому параметру, т.е. если

$$\lim_{n\to\infty} \Pr\left(|\theta - \hat{\theta}| < \varepsilon\right) = 1. \tag{3.2}$$

Требование состоятельности гарантирует от грубых ошибок ε в определении $\hat{\theta}$ при достаточно больших n.

Определение 2. Оценка $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ называется *несмещенной* (оценкой

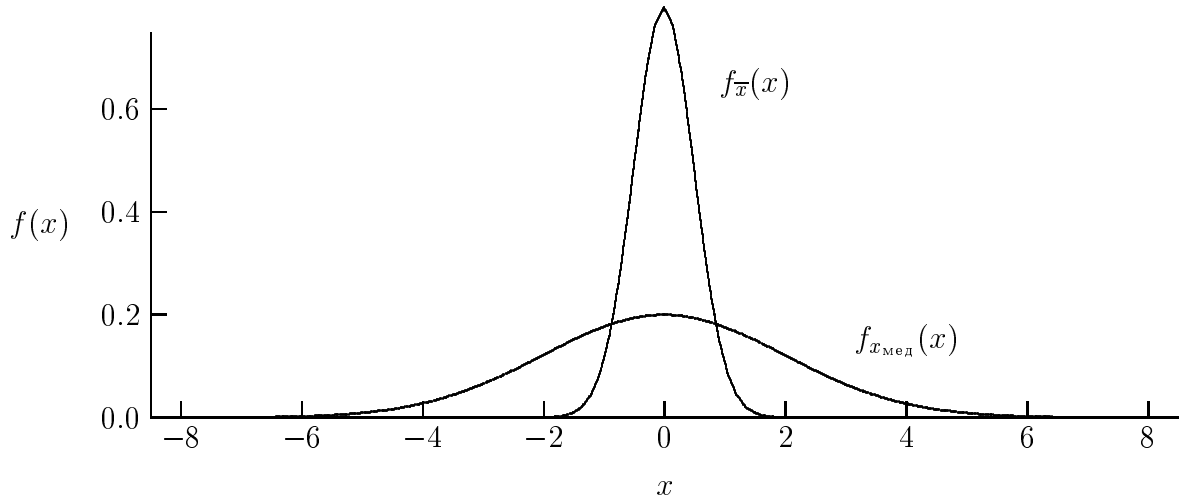


Рисунок 3.1 — Плотности распределения $f_{\overline{x}}(x)$ и $f_{x_{\text{мед}}}(x)$

без систематической ошибки), если её математическое ожидание равно оцениваемому параметру, т.е. если

$$\mathrm{M}[\hat{\theta}] \;=\; \theta. \tag{3.3}$$

Если условие (3.3) не выполняется, то оценка называется *смещенной* (содержащей систематическую ошибку). Часто наряду с несмещенными оценками $\hat{\theta} = u\,(x_1, x_2, \ldots, x_n)$ применяются *асимптотически несмещенные оценки*, т.е. такие оценки, для которых $\mathrm{M}[\hat{\theta}] \to \theta$ при увеличении объема выборки.

Состоятельные, несмещенные или асимптотически несмещенные оценки могут быть получены различными методами. Например, две оценки математического ожидания — среднее арифметическое $\overline{x}$ и выборочная медиана $x_{\text{мед}}$ – являются несмещенными и состоятельными оценками.

Примеры распределений этих оценок изображены на рис. 3.1.

Из этих двух оценок целесообразнее выбрать $\overline{x}$, так как дисперсия этой оценки меньше, чем дисперсия выборочной медианы. В строгих курсах математической статистики доказывается, что дисперсия любой несмещенной оценки одного параметра θ удовлетворяет *неравенству Крамера–Рао*

$$\mathrm{D}[\hat{\theta}] \;\geq\; \frac{1}{I_n(\theta)}, \tag{3.4}$$

где $I_n(\theta)$ – *информация Фишера*, содержащаяся в выборке объема n относительно неизвестного параметра θ. Для непрерывной случайной величины X с плотностью распределения $f_{\mathrm{X}}(x;\theta)$ справедливо неравенство

$$\mathrm{D}[\hat{\theta}] \;\geq\; \frac{1}{n\,I_1}, \qquad I_1 \;=\; -\mathrm{M}\left[\,\frac{\partial^2}{\partial\theta^2}\ln f_{\mathrm{X}}(x;\theta)\right], \tag{3.5}$$

где I_1 – количество информации о параметре θ, содержащееся в одном наблюдении, n – число произведенных испытаний, оценка параметра может быть получена по каждому из n испытаний.

Следовательно, скорость сходимости выборочной дисперсии $\mathrm{D}[\hat{\theta}]$ к нулю не может быть быстрее чем $1/n$.

Определение 3. Несмещенная оценка $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, для которой в неравенстве Крамера-Рао (3.4) достигается знак равенства, называется *эффективной*.

В математической статистике применяются также асимптотически эффективные оценки. Дисперсия асимптотически эффективных оценок стремится к нижней границе неравенства Крамера-Рао при неограниченном увеличении объема выборки, т.е. при $n \to \infty$.

Кроме перечисленных трех основных свойств "хороших" оценок (состоятельность, несмещенность, эффективность), существует еще понятие достаточной оценки.

Определение 4. Оценка $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ называется *достаточной*, если она использует всю информацию относительно оцениваемого параметра, содержащуюся в выборке. Достаточные оценки построены таким образом, что никакие другие оценки не могут дать какой-либо дополнительной информации об оцениваемых параметрах.

Кроме указанных свойств, имеются и другие, которые должны иметь хорошие оценки. Например, желательно, чтобы оценки параметров имели простой линейный вид. К сожалению, не всегда возможно найти функции $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, которые имели бы все указанные свойства.

Ниже приведем основные методы нахождения точечных оценок неизвестных параметров и укажем свойства таких оценок.

3.3. Метод моментов

Пусть по результатам выборки $x_1, x_2, \ldots, x_n$ объема n, извлеченной из генеральной совокупности с функцией распределения $F(x;\theta)$, требуется оценить неизвестный параметр θ этого распределения.

По аналогии с моментами случайной величины X введем понятие эмпирических моментов. *Эмпирические* и соответствующие им *теоретические начальные моменты* порядка k определяются следующими формулами:

т е о р е т и ч е с к и е:

$$\nu_k = \sum_{i=1}^{n} x_i^k \, p_i(x;\theta) \text{ – для дискретных СВ X;} \tag{3.6}$$

$$\nu_k = \int\limits_{-\infty}^{\infty} x^k f(x;\theta)\, dx \text{ – для непрерывных СВ X;} \tag{3.7}$$

э м п и р и ч е с к и е:

$$\nu_k^* = \overline{x^k} = \frac{1}{n} \sum_{i=1}^{n} x_i^k. \tag{3.8}$$

В частности, заметим, что начальный теоретический момент первого порядка является математическим ожиданием СВ X, а эмпирический начальный момент первого порядка является средним арифметическим значением наблюденных значений СВ X, т.е.

$$\nu_1^* = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.9}$$

Аналогично эмпирические и соответствующие им теоретические *центральные моменты* порядка k определяются формулами:

т е о р е т и ч е с к и е:

$$\mu_k = \sum_{i=1}^{n} (x_i - m_x)^k \, p_i(x;\theta) \text{ – для дискретных СВ X;} \tag{3.10a}$$

$$\mu_k = \int\limits_{-\infty}^{\infty} (x - m_x)^k \, f(x;\theta)\, dx \text{ – для непрерывных СВ X;} \tag{3.10b}$$

э м п и р и ч е с к и е:

$$\mu_k^* = \overline{(x - \overline{x})^k} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^k. \tag{3.11}$$

При этом теоретический центральный момент второго порядка является дисперсией случайной величины X.

Следует иметь в виду, что эмпирические моменты являются случайными величинами, в то время как теоретические моменты являются фиксированными постоянными величинами.

Наиболее простым является метод оценивания, называемый *методом моментов*, предложенный английским статистиком Карлом Пирсоном. Метод моментов основывается на том, что эмпирические моменты (или их функции) принимаются за оценки соответствующих теоретических моментов (или их функций) и параметры выражаются через эти моменты. Например, для нахождения оценок параметров функции распределения $F(x;\theta_1,\theta_2)$, содержащей два неизвестных параметра θ_1 и θ_2, составляется система двух уравнений

$$\begin{cases} \nu_1(\hat{\theta}_1,\hat{\theta}_2) = \nu_1^*(\hat{\theta}_1,\hat{\theta}_2), \\ \mu_2(\hat{\theta}_1,\hat{\theta}_2) = \mu_2^*(\hat{\theta}_1,\hat{\theta}_2). \end{cases} \tag{3.12}$$

Решая эту систему, находят оценки $\hat{\theta}_1$ и $\hat{\theta}_2$ функции распределения $F(x;\hat{\theta}_1,\hat{\theta}_2)$.

Таким образом, приравнивая теоретический начальный момент первого порядка к эмпирическому начальному моменту первого порядка, приходим к выводу, что оценкой математического ожидания случайной величины X, распределенной по любому закону, является среднее арифметическое наблюденных значений случайной величины X, т. е.

$$\mathrm{M}[\hat{\mathrm{X}}] = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.13}$$

Приравнивая теоретический и эмпирический центральные моменты второго порядка, приходим к выводу, что оценка дисперсии случайной величины X, распределенной по любому закону, находится по формуле

$$\mathrm{D}[\hat{\mathrm{X}}] = \hat{\sigma}_x^2 = s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2. \tag{3.14}$$

Поступая аналогичным образом, можно найти оценки теоретических моментов любого порядка.

Метод моментов отличается простотой, однако оценки, найденные этим методом, как правило, являются смещенными и малоэффективными, т. е. не являются наилучшими из возможных. Исключение представляет лишь нормальное распределение, при котором метод моментов дает эффективные и состоятельные оценки $\overline{x}$ и s параметров m_{X} и σ_{X}. Строгое исследование свойств оценок, найденных методом моментов, можно найти в курсах по математической статистике и теории оценивания.

3.4. Метод наибольшего правдоподобия

Метод наибольшего правдоподобия является широко распространенным методом точечной оценки. Он предложен английским статистиком Робертом Фишером.

Пусть из генеральной совокупности с плотностью распределения вероятностей $f(x;\theta)$ произведена выборка объема n и получены результаты $x_1, x_2, \ldots, x_n$. Предположим вначале, что X – дискретная случайная величина, закон распределения которой зависит от неизвестного параметра θ. Например, можно предположить, что случайная величина X распределена по закону Пуассона $\mathrm{P}_k = \frac{1}{k!}\lambda^k \exp(-\lambda)$, где λ – неизвестный параметр, который надо оценить по данным выборки. Будем рассматривать результаты выборки как реализацию n-мерной случайной величины (X_1, X_2, …, X_n). Предположим далее, что составляющие этой случайной величины независимы и получены в однородных условиях.

В этом случае вероятность (её называют *функцией правдоподобия*), того, что составляющие примут значения, равные наблюденным значениям, равна

$$L \;=\; \Pr\left(\mathrm{X}_1 = x_1, \mathrm{X}_2 = x_2, \ldots, \mathrm{X}_n = x_n\right) \;= \tag{3.15}$$

$$= \Pr\left(x_1;\theta\right) \cdot \Pr\left(x_2;\theta\right) \cdot \ldots \cdot \Pr\left(x_n;\theta\right) \;=\; \prod_{i=1}^{n} \Pr\left(x_i;\theta\right).$$

В случае непрерывной случайной величины функция правдоподобия имеет вид

$$L \;=\; f(x_1, x_2, \ldots, x_n) \;=\; f(x_1;\theta)\, f(x_2;\theta) \ldots f(x_n;\theta) \;=\; \prod_{i=1}^{n} f\left(x_i;\theta\right). \tag{3.16}$$

Формула (3.16) определяет плотность распределения вероятностей непрерывной случайной величины (X_1, X_2, …, X_n) или плотность распределения выборки.

В качестве искомой оценки $\hat{\theta}$ неизвестного параметра θ, найденной по методу наибольшего правдоподобия, выбирается такая функция $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, которая *максимизирует функцию правдоподобия*. Следовательно, на основании известных правил дифференциального исчисления для нахождения оценок наибольшего правдоподобия составляется система m уравнений (здесь m – число оцениваемых параметров)

$$\frac{\partial}{\partial\theta_i} L \;=\; 0, \qquad i = 1, 2, \ldots, m, \tag{3.17}$$

и выбирается то решение, которое обращает функцию правдоподобия в максимум.

Найденные таким методом оценки называют *оценками максимального (наибольшего) правдоподобия* или ММП–оценками.

Поскольку экстремум функций L и $\ln L$ достигается при одних и тех же значениях выборочной функции $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, то иногда для упрощения расчетов пользуются *логарифмической функцией правдоподобия*. В этом случае оценки наибольшего правдоподобия находятся из системы уравнений:

$$\frac{\partial}{\partial \theta_i} \ln L = 0, \qquad i = 1, 2, \ldots, m. \tag{3.18}$$

Метод наибольшего правдоподобия обладает рядом преимуществ по сравнению с методом моментов.

Укажем некоторые важные свойства оценок наибольшего правдоподобия.

1. Метод наибольшего правдоподобия дает состоятельные оценки.

2. Если существует эффективная оценка, то метод наибольшего правдоподобия дает именно эту оценку и другой ММП-оценки не существует.

3. Оценки наибольшего правдоподобия асимптотически эффективны.

4. Оценки наибольшего правдоподобия имеют асимптотически нормальное распределение с параметрами:

$$\mathrm{M}[\hat{\theta}] = \theta, \qquad \mathrm{D}[\hat{\theta}] = -\frac{1}{n\,\mathrm{M}\left[\partial^2 \ln f(x;\theta)/\partial\theta^2\right]}. \tag{3.19}$$

5. Если существуют достаточные оценки, то метод наибольшего правдоподобия дает эти оценки.

Недостатком метода является то, что иногда оценки наибольшего правдоподобия являются смещенными. Они имеют место лишь при выполнении некоторых условий регулярности. Смещение оценок можно компенсировать введением поправок (с ростом n смещение уменьшается, т.е. оценки наибольшего правдоподобия асимптотически несмещенные). Кроме того, для нахождения оценок методом наибольшего правдоподобия часто приходится решать сложные системы уравнений.

3.5. Точечные оценки неизвестных параметров распределения

Кроме метода моментов и метода максимального правдоподобия, оценки среднего положения могут находиться и другими методами, например, методом квантилей или методом наименьших квадратов, которые будут рассмотрены ниже. Кроме того, в практике применяется и ряд других оценок, предложенных интуитивно, без полного теоретического обоснования.

Для оценки математического ожидания по выборке объема n применяются, например:

а) средняя гармоническая

$$\overline{x}_{\text{гарм}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} = n\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}; \tag{3.20а}$$

б) средняя геометрическая

$$\overline{x}_{\text{геом}} \;=\; \left(\prod_{i=1}^{n} x_i\right)^{1/n}; \tag{3.20b}$$

в) степенные средние с постоянной γ

$$\overline{x}_{\text{степ}} \;=\; \left(\prod_{i=1}^{n} x_i^{\gamma}\right)^{1/\gamma}; \tag{3.20c}$$

г) медиана $x_{\text{мед}}$ – значение случайной величины X, приходящееся на середину упорядоченного (ранжированного) статистического ряда;

д) мода $x_{\text{мод}}$ – наблюденное значение случайной величины X, встречающееся с наибольшей частотой.

Для оценки дисперсии наблюденных значений случайной величины иногда применяются:

а) размах

$$R \;=\; x_{\max} - x_{\min}, \tag{3.20d}$$

где $x_{\max}$ – наибольшее из значений случайной величины X; $x_{\min}$ – наименьшее из значений случайной величины X;

б) коэффициент вариации

$$var \;=\; s/\overline{x}\;100\,(\%), \tag{3.20e}$$

где $s = (\mu_2^*)^{1/2}$.

В большинстве практических случаев эти оценки можно считать оценками "хуже по качеству", чем оценки, полученные методом наибольшего правдоподобия или методом моментов. Их применение объясняется либо простотой их вычисления, либо исторически сложившейся традицией. Поэтому эти оценки можно применять как дополнительные характеристики, описывающие центр распределения или рассеивание наблюденных значений случайной величины X.

3.6. Интервальные оценки параметров

Точечная оценка неизвестного параметра θ, найденная по выборке объема n из генеральной совокупности с функцией распределения $F(x;\theta)$, не позволяет непосредственно ответить на вопрос, какую ошибку мы совершаем, принимая вместо точного значения параметра θ некоторое его приближенное значение (оценку) $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$. В связи с этим во многих случаях более выгодно пользоваться интервальной оценкой, основанной на определении некоторого интервала, внутри которого с определенной вероятностью находится значение параметра θ.

Пусть найденная по результатам выборки объема n статистическая характеристика $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ является точечной оценкой неизвестного параметра θ. Чем меньше разность $|\theta - \hat{\theta}|$, тем лучше качество оценки, тем точнее оценка. Таким образом, положительное число ε характеризует точность оценки

$$|\theta - \hat{\theta}| \;<\; \varepsilon. \tag{3.21}$$

Понятно, что точность ε зависит от объема выборки n. Каков должен быть объем выборки n, чтобы обеспечить заданную точность ε, или как определить точность ε при данном объеме выборки? На эти вопросы нельзя непосредственно ответить, используя приведенное неравенство, так как статистические методы не позволяют категорически утверждать, что оценка удовлетворяет этому неравенству в смысле математического анализа, т.е. из некоторого объема выборки n. Можно только говорить о вероятности $\mathrm{P} = 1-\alpha$, с которой оно выполняется.

Определение 1. *Доверительной вероятностью* оценки называют вероятность $\mathrm{P} = 1-\alpha$ выполнения неравенства $|\theta - \hat{\theta}| < \varepsilon$.

Обычно доверительная вероятность оценки задается заранее. Наиболее часто полагают: $1-\alpha = 0,95$; 0,99; 0,9973; 0,999.

Доверительная вероятность точечной оценки показывает, что при извлечении выборок объема n из одной и той же генеральной совокупности с функцией распределения $F(x;\theta)$ в $(1-\alpha)\cdot 100\%$ случаях параметр θ будет накрываться данным интервалом. Если доверительная вероятность $(1-\alpha)$ выбрана достаточно близкой к единице $(1-\alpha) \geq 0,90$, то число ε определяет предельную погрешность точечной оценки неизвестного параметра θ.

Пусть вероятность того, что $|\theta - \hat{\theta}| < \varepsilon$, равна $(1-\alpha)$:

$$\Pr\left(|\theta - \hat{\theta}| < \varepsilon\right) = 1-\alpha. \tag{3.22}$$

Преобразуем формулу:

$$\Pr\left(\hat{\theta} - \varepsilon < \theta < \hat{\theta} + \varepsilon\right) = 1-\alpha. \tag{3.23}$$

Формула (3.23) показывает, что неизвестный параметр θ заключен внутри интервала $[\hat{\theta} - \varepsilon;\ \hat{\theta} + \varepsilon]$ (рис. 3.2).

Определение 2. *Доверительным интервалом* называется интервал $[\hat{\theta}-\varepsilon;\ \hat{\theta}+\varepsilon]$, накрывающий неизвестный оцениваемый параметр θ с заданной доверительной вероятностью $\mathrm{P}= 1-\alpha$.

Длина доверительного интервала играет важную роль в практических приложениях. Чем меньше длина доверительного интервала $[\hat{\theta} - \varepsilon;\ \hat{\theta} + \varepsilon]$, тем точнее оценка. Если же длина доверительного интервала велика, то оценка малопригодна для практики.

Из формулы (3.23) следует, что длина доверительного интервала равна 2ε. Анализируя формулу (3.23), приходим к заключению, что длина доверительного интервала 2ε определяется двумя величинами: доверительной вероятностью $\mathrm{P}= 1-\alpha$ и объемом выборки n. Таким образом, величины ε, $1-\alpha$ и n тесно взаимосвязаны. Задавая определенные значения двум из них, можно определить величину третьей.

Общая схема построения доверительных интервалов для параметров нормального закона распределения вероятностей состоит в следующем.

1. Из генеральной совокупности с функцией распределения $F(x;\theta)$ извлекается выборка объема n. На основании этих данных методом наибольшего правдоподобия или методом моментов находится точечная оценка $\hat{\theta}$ оцениваемого параметра θ.

2. Составляется случайная величина, например, $\mathrm{Y}(\theta)$, связанная с параметром θ и имеющая известную плотность распределения вероятностей $f_{\mathrm{Y}}(y;\theta)$.

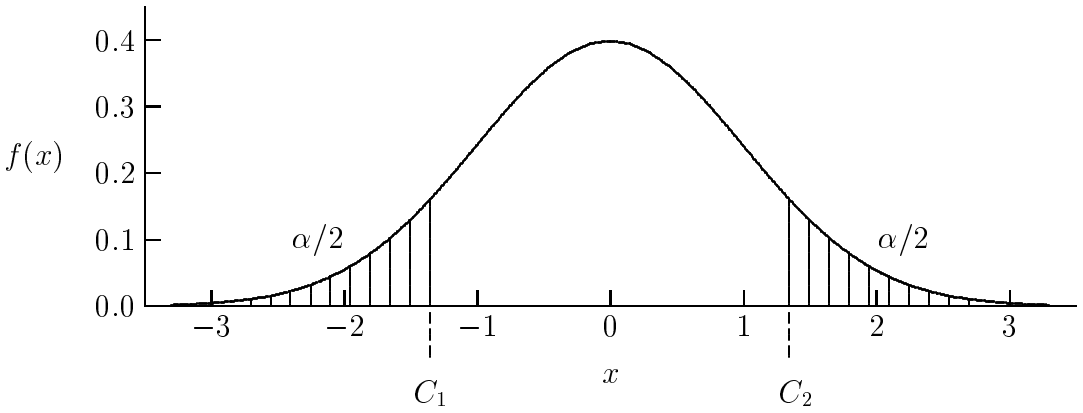


Рисунок 3.2 — К построению доверительного интервала

3. Задают доверительную вероятность равной $1-\alpha$. Обычно принимают доверительную вероятность $1-\alpha$ равной 0,90; 0,95; 0,99; 0,9973; 0,999.

4. Используя плотность вероятностей распределения случайной величины Y, находят такие два числа C_1 и C_2, что

$$\Pr\left(C_1 < \mathrm{Y} < C_2\right) \;=\; \int_{C_1}^{C_2} f_{\mathrm{Y}}(y;\theta)\,dy \;=\; 1-\alpha. \tag{3.24}$$

Значения C_1 и C_2 выбираются, как правило, при симметричных условиях:

$$\Pr\left(\mathrm{Y}(\theta) < C_1\right) \;=\; \frac{\alpha}{2}, \qquad \Pr\left(\mathrm{Y}(\theta) > C_2\right) \;=\; \frac{\alpha}{2}, \tag{3.25}$$

т. е. чтобы суммарная площадь заштрихованных на рис. 3.2 площадей была равна α.

5. Неравенство, стоящее в круглых скобках уравнения (3.24), преобразуется в равносильное неравенство

$$\Pr\left(\hat{\theta}-\varepsilon \;<\; \theta \;<\; \hat{\theta}+\varepsilon\right) \;=\; 1-\alpha, \tag{3.26}$$

накрывающее с заданной вероятностью $1-\alpha$ неизвестный параметр θ.

3.7. Доверительные интервалы для математического ожидания нормальной случайной величины с известной дисперсией

Задачу анализа доверительных интервалов при оценивании математического ожидания нормальной случайной величины обычно рассматривают в двух вариантах:

– случай, когда дисперсия известна;

– случай, когда дисперсия неизвестна.

Используем указанную выше схему для нахождения доверительных интервалов параметров нормального закона a и σ.

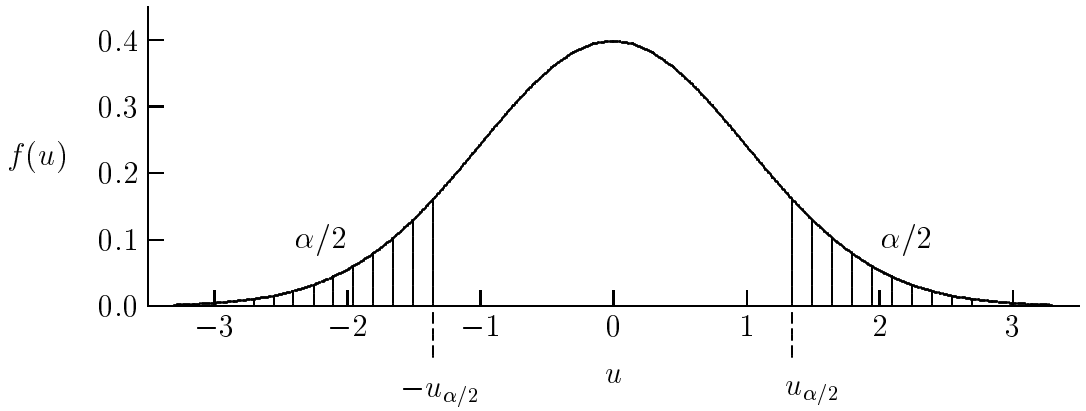


Рисунок 3.3 — К построению доверительного интервала для оценки математического ожидания нормальной случайной величины при известной σ

Рассмотрим нормальную модель генеральной совокупности $\mathcal{N}(a, \sigma)$, в которой параметр σ будем считать фиксированным (известным), а параметр a – неизвестным. Для нахождения точечной оценки параметра a из генеральной нормальной совокупности извлечена выборка объема n. На основании этой выборки найдена точечная оценка математического ожидания

$$\hat{a} \;=\; \overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{3.27}$$

Ранее было показано, что если $\mathrm{X} \to\mathcal{N}(a;\, \sigma)$, то случайная величина $\overline{\mathrm{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{X}_i$, линейно связанная с $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$, распределена также по нормальному закону, но с математическим ожиданием a и дисперсией σ^2/n, т.е. $\overline{\mathrm{X}} \to\mathcal{N}(a;\, \sigma/\sqrt{n})$.

С целью построить доверительный интервал для параметра a составим стандартизованную случайную величину (выборочную статистику)

$$u \;=\; \frac{\overline{x} - a}{\sigma/\sqrt{n}}. \tag{3.28}$$

Случайная величина u имеет стандартизованное нормальное распределение $u \to\mathcal{N}(0;\, 1)$.

Вероятность того, что стандартизованная случайная величина u отклонится от своего математического ожидания на величину $u_{\alpha/2}$, находится по формуле

$$\Pr\left(-u_{\alpha/2} < \frac{\overline{x} - a}{\sigma/\sqrt{n}} < u_{\alpha/2}\right) \;=\; \frac{1}{\sqrt{2\pi}}\int\limits_{-u_{\alpha/2}}^{u_{\alpha/2}} \exp\left(-t^2/2\right)\, dt. \tag{3.29}$$

Схема, отвечающая этой вероятности, приведена на рис. 3.3.

Зададим эту вероятность равной $\mathrm{P} = 1 - \alpha$, тогда получим

$$\Pr\left(-u_{\alpha/2} < \frac{\overline{x} - a}{\sigma/\sqrt{n}} < u_{\alpha/2}\right) \;=\; \frac{2}{\sqrt{2\pi}}\int\limits_{0}^{u_{\alpha/2}} \exp\left(-t^2/2\right)\, dt \;=\; 1 - \alpha. \tag{3.30}$$

Решая уравнение

$$\frac{2}{\sqrt{2\pi}} \int\limits_0^{u_{\alpha/2}} \exp\left(-t^2/2\right)\, dt \;=\; 2\,\Phi\left(u_{\alpha/2}\right) \;=\; 1-\alpha, \tag{3.31}$$

находят *квантили* стандартизованного нормального распределения $u_{\alpha/2}$. Обычно квантили нормального распределения $u_{\alpha/2}$ находят по таблицам Лапласа (см. приложение) из условия $2\,\Phi(u_{\alpha/2}) = 1-\alpha$.

Для наиболее употребительных значений доверительной вероятности $\mathrm{P}= 1-\alpha$ квантили стандартизованного нормального распределения приведены в табл. 3.1.

Таблица 3.1 — Квантили стандартизованного нормального распределения

Доверительная вероятность $\mathrm{P} = 1-\alpha$	Квантиль $u_{\alpha/2}$
0,90	1,64
0,95	1,96
0,99	2,58
0,9973	3,00
0,999	3,37

Будем считать квантиль $u_{\alpha/2}$, соответствующий заданной доверительной вероятности $\mathrm{P}= 1-\alpha$, известным. Преобразуем неравенство, стоящее в левой части уравнения (3.29),

$$\Pr\left(-\frac{\sigma}{\sqrt{n}}\,u_{\alpha/2} \;<\; (\overline{x}-a) \;<\; \frac{\sigma}{\sqrt{n}}\,u_{\alpha/2}\right) \;=\; 1-\alpha,$$

или

$$\Pr\left(\overline{x}-\frac{\sigma}{\sqrt{n}}\,u_{\alpha/2} \;<\; a \;<\; \overline{x}+\frac{\sigma}{\sqrt{n}}\,u_{\alpha/2}\right) \;=\; 1-\alpha. \tag{3.32}$$

Следовательно, доверительный интервал $[\overline{x}-\sigma\,u_{\alpha/2}/\sqrt{n};\ \overline{x}+\sigma\,u_{\alpha/2}/\sqrt{n}]$ накрывает неизвестное математическое ожидание a с заданной вероятностью $1-\alpha$. Точность оценки математического ожидания (предельная погрешность): $\varepsilon = \sigma\,u_{\alpha/2}/\sqrt{n}$.

Замечание 1. Анализируя формулу (3.32), замечаем, что:

а) увеличение объема выборки n приводит к уменьшению длины доверительного интервала, т.е. к улучшению точности ε;

б) увеличение доверительной вероятности $1-\alpha$ приводит к увеличению длины доверительного интервала, т.е. к уменьшению точности ε;

в) если задать точность (предельную погрешность интервальной оценки) ε и доверительную вероятность $1-\alpha$, то из соотношения

$$\varepsilon \;=\; \frac{\sigma}{\sqrt{n}}\,u_{\alpha/2} \tag{3.33}$$

можно найти минимальный объем выборки, который обеспечит заданную точность

$$n \;=\; \frac{1}{\varepsilon^2}\,u_{\alpha/2}^2\sigma^2. \tag{3.34}$$

Замечание 2. Пусть СВ X имеет произвольную функцию распределения вероятностей $F(x)$. Пусть также для СВ X выполняется центральная предельная теорема теории вероятностей. Тогда, в силу этой теоремы, при достаточно большом объеме выборки закон распределения выборочной статистики $u = (\overline{x} - \mathrm{M}[\overline{x}])/\sigma_{\overline{x}}$ будет приближенно нормальным. В этом случае доверительный интервал для M[X] случайной величины X, определенный по формуле (3.32), будет приближенным.

С целью построения точного доверительного интервала для M[X] случайной величины X, имеющей произвольный закон распределения, необходимо знать либо закон распределения вероятностей средней арифметической $\overline{x} = n^{-1}\sum_i x_i$, либо закон распределения выборочной статистики $u = (\overline{x} - \mathrm{M}[\overline{x}])/\sigma_{\overline{x}}$.

Законы распределения этих величин зависят от закона распределения случайной величины X.

Пример

Мальчики и девочки рождаются не одинаково часто. Этот факт был замечен уже давно. Для создания математических моделей, способных проследить процессы, наблюдающееся в составе народонаселения страны, необходим большой статистический материал.

С 1874 года по 1900 год в Швейцарии родилось 2644757 детей, в том числе 1359671 мальчик и 1285086 девочек.

Какова вероятность рождения мальчиков?

Решение

Статистическая частота рождения мальчиков за указанные годы составила

$$\nu = 1359671/2644757 = 0,5141.$$

Здесь мы имеем дело с последовательностью испытаний с двумя возможными исходами в каждом, причем число испытаний n очень велико. Поэтому можно использовать как теорему Бернулли, так и теорему Муавра–Лапласа. Первая из них указывает на то, что вероятность появления мальчика близка к 1/2 (как и вероятность рождения девочки).

В математической модели частота рождения мальчика w_n, наблюдаемая среди n родившихся детей, имеет дисперсию $\mathrm{D}[w_n] = p(1-p)/n$. Поскольку объем выборки n очень велик (он равен 2644757) и вероятность p близка к 1/2, то с высокой степенью точности $\mathrm{D}[w_n] = (0,0003)^2$.

Если практически достоверными условиться считать события с вероятностью не менее 0,95, то из оценки (3.32) следует, что вероятность p рождения мальчика заключена между следующими границами:

$$p > \nu - 1,96\sqrt{\mathrm{D}[w_n]} = 0,5135, \qquad p < \nu + 1,96\sqrt{\mathrm{D}[w_n]} = 0,5146.$$

Значения со столь высокой точностью вероятности p позволяет принять за практически достоверное, что по крайней мере в течение нескольких лет, последующих за 1900 годом, число рождавшихся мальчиков на каждые 10000 детей составляло от 5135 до 5146.

Как показывают многочисленные и продолжительные наблюдения, вероятность p является довольно устойчивой характеристикой.

3.8. Доверительные интервалы для математического ожидания нормальной случайной величины при неизвестной дисперсии

Пусть $\mathrm{X} \to \mathcal{N}(a;\,\sigma)$, причем a и σ неизвестны. Для нахождения точечных оценок параметров a и σ из генеральной нормальной совокупности извлечена выборка объема n. На основании этой выборки найдены точечные несмещенные оценки неизвестных параметров

$$\hat{a} \;=\; \overline{x}; \qquad \hat{\sigma} \;=\; s \;=\; \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}. \tag{3.35}$$

Для построения доверительного интервала для математического ожидания составим вспомогательную случайную величину (выборочную статистику)

$$t \;=\; \frac{\overline{x}-a}{s/\sqrt{n-1}}. \tag{3.36}$$

Умножая числитель и знаменатель (3.36) на положительную величину $\sqrt{n}/\sigma$, получим

$$t \;=\; \frac{\overline{x}-a}{\sigma/\sqrt{n}}\cdot\left(\frac{1}{n-1}\cdot\frac{ns^2}{\sigma^2}\right)^{-1/2}. \tag{3.37}$$

В курсе теории вероятностей доказывается теорема о том, что величина ns^2/σ^2 распределена по закону χ^2 с $\nu = n-1$ степенями свободы. Введем обозначение $\mathrm{Y}= (\overline{x}-a)\sqrt{n}/\sigma$, тогда

$$t \;=\; \mathrm{Y}\left(\frac{1}{n-1}\chi^2_{n-1}\right)^{-1/2}. \tag{3.38}$$

Нетрудно видеть, что $\mathrm{Y} \to \mathcal{N}(0;\,1)$. Анализируя случайную величину t, определяемую равенством (3.38), приходим к выводу, что эта случайная величина распределена по закону Стьюдента с $\nu = n-1$ степенями свободы. Вероятность того, что случайная величина t попадет в интервал $[-t_{\alpha/2,\,n-1};\,t_{\alpha/2,\,n-1}]$, находится по формуле (рис. 3.4)

$$\Pr\left(-t_{\alpha/2,\,n-1} \;<\; \frac{\overline{x}-a}{s/\sqrt{n}} \;<\; t_{\alpha/2,\,n-1}\right) \;=\; 2\int_0^{t_{\alpha/2,\,n-1}} f(t)\,dt. \tag{3.39}$$

Зададим эту вероятность равной $1-\alpha$. Тогда из решения уравнения

$$2\int_0^{t_{\alpha/2,\,n-1}} f(t)\,dt \;=\; 1-\alpha \tag{3.40}$$

можно найти квантиль распределения Стьюдента $t_{\alpha/2,\,n-1}$. Имеются специальные таблицы (см. приложение), в которых помещены квантили t–распределения, соответствующие доверительной вероятности $1-\alpha$ и объему выборки n.

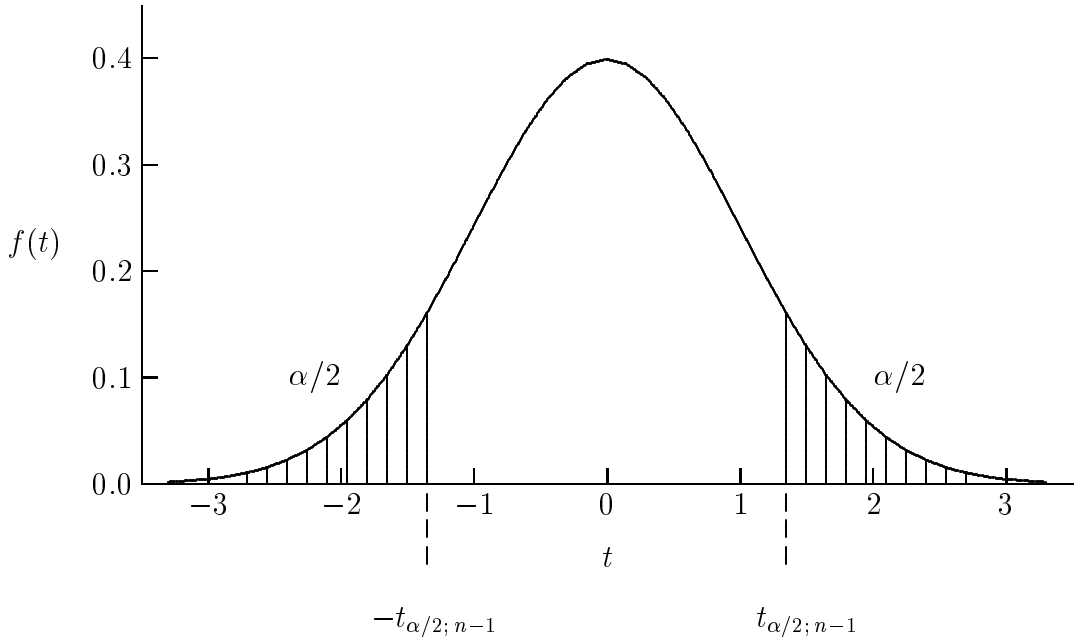


Рисунок 3.4 — К построению доверительного интервала для оценки математического ожидания случайной величины, распределенной по закону Стьюдента

Будем считать квантили $t_{\alpha/2,\,n-1}$, соответствующие заданной доверительной вероятности $1-\alpha$ и объему выборки n, известными. Преобразуем неравенство, стоящее в левой части уравнения (3.39):

$$\Pr\left(-t_{\alpha/2,\,n-1}\,s/\sqrt{n} \;<\; (\overline{x}-a) \;<\; t_{\alpha/2,\,n-1}\,s/\sqrt{n}\right) \;=\; 1-\alpha \tag{3.41a}$$

или

$$\Pr\left(\overline{x}-t_{\alpha/2,\,n-1}\,s/\sqrt{n} \;<\; a \;<\; \overline{x}+t_{\alpha/2,\,n-1}\,s/\sqrt{n}\right) \;=\; 1-\alpha. \tag{3.41b}$$

Из этого выражения для вероятности вытекает, что доверительный интервал $(\overline{x}-t_{\alpha/2,\,n-1}\,s/\sqrt{n},\ \overline{x}+t_{\alpha/2,\,n-1}\,s/\sqrt{n})$ накрывает неизвестное математическое ожидание с заданной вероятностью $1-\alpha$. Точность (предельная погрешность) оценки математического ожидания следующая:

$$\varepsilon \;=\; \frac{s}{\sqrt{n}}\,t_{\alpha/2,\,n-1}. \tag{3.42}$$

Полученный доверительный интервал обладает теми же свойствами, что и доверительный интервал для математического ожидания при известном σ.

Замечание. Выше отмечалось, что при неограниченном возрастании объема выборки распределение Стьюдента стремится к нормальному. Поэтому уже при $n \geq 30$ с целью построения доверительных интервалов для математического ожидания можно вместо распределения Стьюдента использовать нормальное распределение. В этом случае приближенные доверительные интервалы для математического ожидания находятся по формуле (3.32), в которой следует принять $\sigma = s$.

3.9. Доверительный интервал для среднего квадратического отклонения нормальной случайной величины

Пусть случайная величина $X \to \mathcal{N}(a, \sigma)$, причем параметры a и σ неизвестны. Для нахождения точечных оценок параметров a и σ из генеральной нормальной совокупности извлечена выборка объема n. Пусть на основании этой выборки найдены точечные несмещенные оценки неизвестных параметров

$$\hat{a} = \overline{x}; \qquad \hat{\sigma} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{3.43}$$

Для построения доверительного интервала для среднего квадратического отклонения σ составим вспомогательную случайную величину

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}. \tag{3.44}$$

Эта случайная величина имеет распределение χ^2 с $n-1$ степенями свободы. Вероятность того, что случайная величина χ^2 попадет в интервал $[C_1; C_2]$, равна (рис. 3.5)

$$\Pr\left(C_1 < \chi^2 < C_2\right) = \int_{C_1}^{C_2} f(\chi^2)\, d\chi^2. \tag{3.45}$$

Зададим эту вероятность равной $1-\alpha$. Выберем значения C_1 и C_2 из условий

$$\Pr\left(\chi^2 > C_2\right) = \int_{C_2}^{\infty} f(\chi^2)\, d\chi^2 = \frac{\alpha}{2}, \qquad \Pr\left(\chi^2 < C_1\right) = \int_{0}^{C_1} f(\chi^2)\, d\chi^2 = \frac{\alpha}{2}. \tag{3.46}$$

На рис. 3.5 соответствующие две криволинейные трапеции, каждая с площадью $\alpha/2$, заштрихованы.

Решая (например, численными методами) уравнения (3.46), находят квантили χ^2-распределения $C_1 = \chi^2_{1-\alpha/2;\, n-1}$ и $C_2 = \chi^2_{\alpha/2;\, n-1}$. Будем в дальнейшем считать квантили χ^2-распределения известными.

Преобразуя формулу

$$\Pr\left(\chi^2_{1-\alpha/2;\, n-1} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2;\, n-1}\right) = 1-\alpha,$$

имеем

$$\Pr\left(s\sqrt{\frac{n-1}{\chi^2_{\alpha/2;\, n-1}}} < \sigma < s\sqrt{\frac{n-1}{\chi^2_{1-\alpha/2;\, n-1}}}\right) = 1-\alpha, \tag{3.47}$$

или, более кратко:

$$\Pr\left(s\,\gamma_1 < \sigma < s\,\gamma_2\right) = 1-\alpha, \tag{3.48}$$

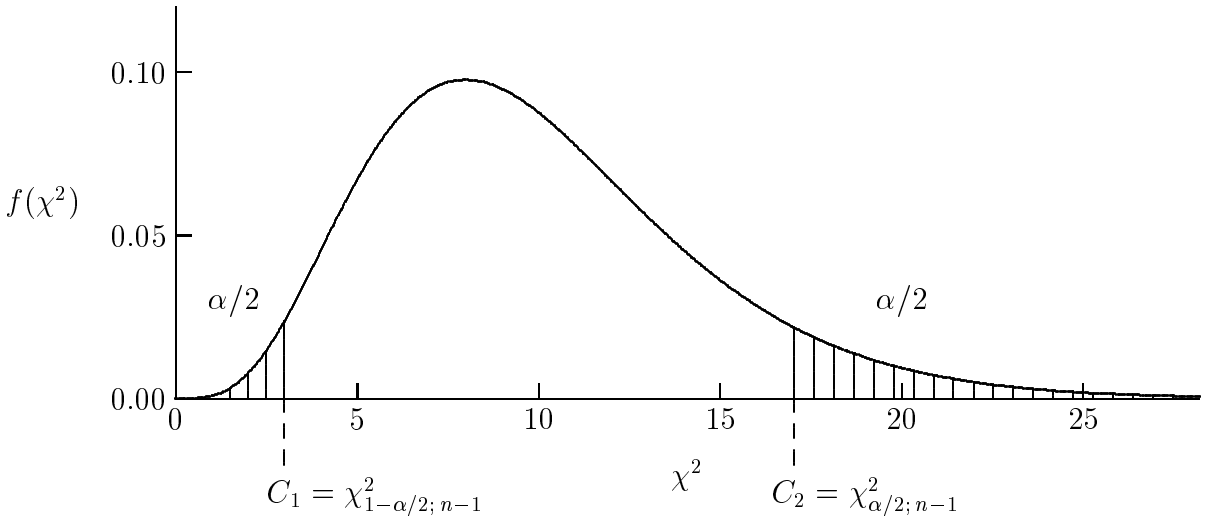


Рисунок 3.5 — К доверительному интервалу для среднего квадратического отклонения σ нормальной случайной величины

где

$$\gamma_1 \;=\; \sqrt{\frac{n-1}{\chi^2_{\alpha/2;\,n-1}}}, \qquad \gamma_2 \;=\; \sqrt{\frac{n-1}{\chi^2_{1-\alpha/2;\,n-1}}}. \tag{3.49}$$

Коэффициенты γ_1 и γ_2, соответствующие доверительной вероятности $1-\alpha$ и числу степеней свободы n, помещены в приложении.

Замечание. Так как при $n \to \infty$ распределение χ^2 приближается к нормальному, то при достаточно большом объеме выборки ($n \geq 50$) доверительный интервал можно найти по формуле

$$\Pr\left(\frac{s}{1+u_{\alpha/2}/\sqrt{2n}} < \sigma < \frac{s}{1-u_{\alpha/2}/\sqrt{2n}}\right) \;=\; 1-\alpha, \tag{3.50}$$

где $u_{\alpha/2}$ – квантиль стандартизованного нормального распределения, соответствующий доверительной вероятности $1-\alpha$.

3.10. Примеры

Пример 3.1

Производится изучение случайной величины X с математическим ожиданием M[X]= a и дисперсией D[X]= σ^2. Из генеральной совокупности извлечена выборка $x_1, x_2, \ldots, x_N$ объемом N, на основании которой построена оценка среднего

$$X^*_N \;=\; \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{$*$}$$

Требуется изучить зависимость дисперсии оценки среднего D[X^*_N] от объема выборки N.

Решение

Оценка $(*)$ является несмещенной

$$\mathrm{M}\left[X_N^*\right] \;=\; \frac{1}{N}\sum_{n=1}^{N}\mathrm{M}\left[x_n\right] \;=\; \frac{1}{N}\sum_{n=1}^{N}a \;=\; a.$$

Введем новую случайную величину $\mathrm{Y} = X_N^* - a$. Поскольку при постоянном значении a имеем $\mathrm{D}[\mathrm{X}_N^*] = \mathrm{D}[\mathrm{X}_N^*$ - a$] = \mathrm{D}[\mathrm{Y}]$, то дальше будем рассматривать $\mathrm{D}[\mathrm{Y}]$. Так как $\mathrm{M}[\mathrm{Y}] = 0$, то для дисперсии оценки $(*)$ имеем

$$\mathrm{D}\left[X_N^*\right] \;=\; \mathrm{D}\left[\mathrm{Y}\right] \;=\; \mathrm{M}\left[\mathrm{Y}^2\right] \;=\; \mathrm{M}\left[\left(\frac{1}{N}\sum_{n=1}^{N}y_n\right)^2\right],$$

где $y_n = x_n - a,\;\; n = 1, 2, \ldots, N$. Это дает

$$\mathrm{D}\left[X_N^*\right] \;=\; \frac{1}{N^2}\mathrm{M}\left[\left(\sum_{n=1}^{N}y_n\right)^2\right] \;=\; \frac{1}{N^2}\mathrm{M}\left[\sum_{n=1}^{N}\sum_{m=1}^{N}y_n y_m\right].$$

В возникшей двойной сумме выделим N совпадающих слагаемых и $N(N-1)$ несовпадающих слагаемых

$$\sum_{n=1,\;m=1}^{N} y_n y_m \;=\; \sum_{n=1}^{N} y_n^2 \;+\; \sum_{n=1,\;m=1,\;n\neq m}^{N} y_n y_m.$$

Математическое ожидание от каждого из $N(N-1)$ несовпадающих слагаемых равно нулю, поэтому

$$\mathrm{D}\left[X_N^*\right] \;=\; \frac{1}{N^2}\mathrm{M}\left[\sum_{n=1}^{N}y_n^2\right] \;=\; \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{M}\left[y_n^2\right] \;=\; \frac{N}{N^2}\sigma^2 \;=\; \frac{1}{N}\mathrm{D}\left[\mathrm{X}\right].$$

Итак, с ростом объема выборки дисперсия оценки среднего $(*)$ уменьшается обратно пропорционально N.

Пример 3.2

Пусть случайная величина X распределена по нормальному закону с параметрами $\mathrm{M}[\mathrm{X}] = a$ и $\sigma = \sqrt{\mathrm{D}[\mathrm{X}]}$, или, более кратко, пусть $\mathrm{X} \to N(a, \sigma)$.

Требуется по результатам наблюдений $x_1, x_2, \ldots, x_n$ оценить параметры a и σ и найти оценки асимметрии и эксцесса.

Решение

Применяя метод моментов, имеем

$$\left\{\begin{array}{l} \nu_1 = \nu_1^* \Rightarrow \mathrm{M}[\bar{\mathrm{X}}] \;=\; \hat{a} \;=\; \bar{x} \;=\; \frac{1}{n}\sum\limits_{i=1}^{n} x_i, \\ \mu_2 = \mu_2^* \Rightarrow \mathrm{D}[\bar{\mathrm{X}}] \;=\; \sigma_x^2 = s^2 \;=\; \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2, \end{array}\right. \qquad (*)$$

где X_1, X_2, ..., X_n представляют как бы n экземпляров случайной величины X с одним и тем же математическим ожиданием a.

Покажем, что среднее арифметическое наблюдаемых значений является состоятельной, несмещенной и эффективной оценкой математического ожидания. Состоятельность оценки следует из теоремы Бернулли

$$\lim_{n\to\infty} \Pr\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) \;=\; 1, \qquad (\varepsilon > 0).$$

Из теоремы Бернулли следует, что среднее арифметическое является состоятельной оценкой математического ожидания при любом законе распределения.

Найдем математическое ожидание средней арифметической

$$\mathrm{M}[\overline{x}] \;=\; \mathrm{M}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{X}_i\right] \;=\; \frac{1}{n}\mathrm{M}\left[\sum_{i=1}^{n}\mathrm{X}_i\right] \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}[\mathrm{X}_i] \;=\; \frac{1}{n}\,na \;=\; a.$$

Следовательно, среднее арифметическое является несмещенной оценкой математического ожидания.

Исследуем эффективность оценки $\hat{a} \;=\; \bar{x},$ считая σ известным. Для этого вычислим дисперсию средней арифметической

$$\mathrm{D}[\overline{x}] \;=\; \mathrm{D}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{X}_i\right] \;=\; \frac{1}{n^2}\mathrm{D}\left[\sum_{i=1}^{n}\mathrm{X}_i\right] \;=\; \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{D}[\mathrm{X}_i] \;=\; \frac{1}{n^2}\cdot n\sigma^2 \;=\; \frac{\sigma^2}{n}.$$

Найдем нижнюю границу неравенства Рао-Крамера. Для этого вычислим

$$\mathrm{M}\left[\frac{\partial^2}{\partial a^2}\,\ln\left\{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)\right\}\right] \;=\; -\mathrm{M}\left[\frac{\partial^2}{\partial a^2}\frac{(x-a)^2}{2\sigma^2}\right] \;=\; -\frac{1}{\sigma^2}.$$

Так как нижняя граница неравенства Рао-Крамера при $N \;=\; n$ совпадает с дисперсией средней арифметической

$$-\,\frac{1}{n\,\mathrm{M}\left[\,\partial^2 f(x,\theta)/\partial\theta^2\right]} \;=\; \frac{\sigma^2}{n},$$

то среднее арифметическое при известном σ является эффективной оценкой математического ожидания.

Покажем, что оценка дисперсии, вычисляемая по формуле $(*)$, является смещенной. Найдем математическое ожидание эмпирической дисперсии, предварительно преобразовав формулу $(*)$:

$$s^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - a) - (\overline{x} - a)\right]^2 \;=$$

$$=\; \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - a)^2 \;-\; 2\,(\overline{x} - a)\sum_{i=1}^{n}(x_i - a) \;+\; n\,(\overline{x} - a)^2\right] \;=$$

$$=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 \;-\; (\overline{x} - a)^2\,;$$

$$\mathrm{M}[s^2] \;=\; \frac{1}{n}\sum_{i}\mathrm{M}\left[(\mathrm{X}_i - a)^2\right] - \mathrm{M}\left[(\overline{x} - a)^2\right] \;=$$

$$= \sigma^2 - \mathrm{M}\left[\left(\frac{1}{n}\sum_i(\mathrm{X}_i - x)\right)^2\right] = \sigma^2 - \frac{1}{n^2}\mathrm{M}\left[\sum_i(\mathrm{X}_i - a)^2\right] = \sigma^2 - \frac{n\,\sigma^2}{n^2} = \frac{n-1}{n}\sigma^2.$$

Так как $\mathrm{M}[s^2] \neq \sigma^2$, то оценка дисперсии, определяемая по формуле $(*)$, является смещенной. Смещение оценки устраняется умножением её на множитель $n/(n-1)$. При достаточно больших n $(n > 30)$ смещение оценки незначительно. При малом объеме выборки $(n \leq 30)$ несмещенная оценка вычисляется по формуле

$$\mathrm{D}[\mathrm{X}] = \hat{\sigma}_X^2 = s^2_{\text{несм}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Приведем без доказательства некоторые результаты относительно свойств оценок дисперсии.

Если параметры нормального закона неизвестны, то оценка дисперсии, определяемая по формуле $(*)$, не обладает свойством эффективности. Однако оценка $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$, если известно математическое ожидание a, обладает свойством несмещенности, состоятельности и эффективности.

Оценки асимметрии $\hat{A}$ и эксцесса $\hat{E}$, являющиеся функциями центральных моментов, находятся по формулам:

$$\frac{\mu_3}{\sigma^3} = \left(\frac{\mu_3}{\sigma^3}\right)^* \Rightarrow \hat{A} = \left(\frac{\mu_3}{\sigma^3}\right)^* = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \overline{x})^3;$$

$$\frac{\mu_4}{\sigma^4} - 3 = \left(\frac{\mu_4}{\sigma^4} - 3\right)^* \Rightarrow \hat{E} = \left(\frac{\mu_4}{\sigma^4}\right)^* - 3 = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \overline{x})^4 - 3.$$

Пример 3.3

Показать, что относительная частота m/n появления события A при n испытаниях Бернулли является состоятельной, несмещенной и эффективной оценкой вероятности p.

Решение

Состоятельность оценки следует из теоремы Бернулли:

$$\lim_{n\to\infty}\Pr\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1, \qquad (\varepsilon > 0).$$

Определим математическое ожидание относительной частоты. Будем полагать, что $m = \mathrm{X}$, т. е. частоту появления события A в n испытаниях схемы Бернулли рассматриваем как случайную величину X, распределенную по биномиальному закону

$$\mathrm{M}\left[\frac{\mathrm{X}}{n}\right] = \frac{1}{n}\mathrm{M}[\mathrm{X}] = \frac{1}{n}np = p.$$

Так как $\mathrm{M}[m/n] = p$, то относительная частота m/n является несмещенной оценкой вероятности p.

Определим дисперсию относительной частоты

$$\mathrm{D}\left[\frac{m}{n}\right] = \mathrm{D}\left[\frac{\mathrm{X}}{n}\right] = \frac{1}{n^2}\mathrm{D}[\mathrm{X}] = \frac{1}{n^2}npq = \frac{pq}{n}.$$

Найдем нижнюю границу неравенства Рао-Крамера, учитывая, что

$$f(x,\theta) \;=\; f(\mathrm{X}=x;\,p) \;=\; C_n^{\mathrm{X}}\,p^{\mathrm{X}}(1-p)^{n-\mathrm{X}}.$$

Тогда

$$\mathrm{M}\left[\frac{\partial^2\,\ln f(x;p)}{\partial p^2}\right] \;=\; \mathrm{M}\left[\frac{\partial^2}{\partial p^2}\Big(\ln C_n^{\mathrm{X}} + \mathrm{X}\ln p + (n-\mathrm{X})\ln(1-p)\Big)\right] \;=$$

$$=\; \mathrm{M}\left[-\frac{\mathrm{X}}{p^2} - \frac{n-\mathrm{X}}{(1-p)^2}\right] \;=\; -\frac{np}{p^2} - \frac{n-np}{(1-p)^2} \;=\; -\frac{n}{pq}.$$

Следовательно, полагая, что для оценки вероятности сделан один эксперимент ($n=1$), найдем, что нижняя граница неравенства Рао-Крамера равна

$$-\,\frac{1}{n\,\mathrm{M}\,[\,\partial^2 f(x;p)/\partial p^2\,]} \;=\; \frac{pq}{n}.$$

Так как нижняя граница неравенства Рао-Крамера совпадает с дисперсией относительной частоты, то относительная частота является эффективной оценкой вероятности p.

Пример 3.4

Случайная величина X распределена по показательному закону с плотностью вероятностей

$$f(x;\theta) \;=\; \theta\exp(-\theta x)$$

с параметром θ.

Требуется по результатам наблюденных значений $x_1, x_2, \ldots, x_n$ этой случайной величины найти оценку $\hat{\theta}$ параметра θ.

Решение

Запишем функцию правдоподобия

$$L \;=\; \theta\exp(-\theta x_1)\;\theta\exp(-\theta x_2)\ldots\;\theta\exp(-\theta x_n) \;=\; \theta^n\exp\left(-\theta\sum_{i=1}^{n}x_i\right).$$

Логарифмируя, имеем

$$\ln L \;=\; n\ln\theta - \theta\sum_{i=1}^{n}x_i.$$

Дифференцируя по параметру θ, находим

$$\frac{\partial\ln L}{\partial\theta} \;=\; \frac{n}{\theta} - \sum_{i=1}^{n}x_i.$$

Приравнивая производную к нулю, получаем

$$\frac{n}{\theta} - \sum_{i=1}^{n}x_i \;=0 \quad\Rightarrow\quad \theta = \frac{n}{\sum_{i=1}^{n}x_i} \;=\; 1/\overline{x}.$$

Пример 3.5

Случайная величина X распределена по нормальному закону.

Требуется по результатам наблюденных значений $x_1, x_2, \ldots, x_n$ этой случайной величины оценить параметры a и σ нормального закона.

Решение

Плотность распределения вероятностей нормально распределенной случайной величины равна

$$f(x; a, \sigma) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right),$$

следовательно, функция правдоподобия имеет вид

$$L \;=\; \frac{1}{(\sqrt{2\pi}\,\sigma)^n}\,\exp\left(-\frac{1}{2\sigma^2}\,\sum_{i=1}^{n}(x_i-a)^2\right).$$

Запишем логарифмическую функцию правдоподобия

$$\ln L \;=\; -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2.$$

Дифференцируя логарифмическую функцию правдоподобия по a и σ, имеем

$$\frac{\partial}{\partial a}\ln L \;=\; \frac{1}{\sigma^2}\,\sum_{i=1}^{n}(x_i-a) \;=\; 0,$$

$$\frac{\partial}{\partial \sigma}\ln L \;=\; -\frac{n}{\sigma} + \frac{1}{\sigma^3}\,\sum_{i=1}^{n}(x_i-a)^2 \;=\; 0,$$

откуда находим

$$\hat{a} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \;=\; \overline{x}; \qquad\qquad \hat{\sigma}^2 \;=\; s^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

Эти оценки совпали с оценками, полученными методом моментов.

Пример 3.6

Пусть случайная величина X распределена по закону Пуассона

$$\Pr(\mathrm{X}=k) \;=\; \frac{\lambda^k}{k!}\,\exp(-\lambda), \qquad k = 1, 2, \ldots.$$

Требуется на основании наблюденных значений $x_1 = k_1$, $x_2 = k_2$, ..., $x_n = k_n$, оценить неизвестный параметр λ этого распределения.

Решение

Запишем функцию правдоподобия

$$L \;=\; \frac{\lambda^{x_1}}{x_1!}\,\exp(-\lambda)\frac{\lambda^{x_2}}{x_2!}\,\exp(-\lambda)\ldots\frac{\lambda^{x_n}}{x_n!}\,\exp(-\lambda) \;=$$

$$= \frac{\lambda^S}{x_1!x_2!\ldots x_n!}\exp(-n\lambda), \qquad S = \sum_{i=1}^{n} x_i.$$

Для нахождения оценки λ, т.е. такого значения $\lambda = \lambda(x_1, x_2, \ldots, x_n)$, при котором функция правдоподобия обращается в максимум, удобно перейти к *логарифмической функции правдоподобия*

$$\ln L = -n\lambda + \sum_{i=1}^{n} x_i \ln\lambda - \sum_{i=1}^{n} \ln(x_i!),$$

следовательно,

$$\frac{\partial}{\partial\lambda}\ln L = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i.$$

Приравнивая производную к нулю, имеем

$$-n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0,$$

что дает

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}.$$

Пример 3.7

Двумерная случайная величина (X, Y) распределена по нормальному закону с плотностью вероятностей

$$f(x,y) = \frac{1}{2\pi\,\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}\sqrt{1-\rho^2}} \times$$

$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-m_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2} - \frac{2\rho\,(x-m_{\mathrm{X}})(y-m_{\mathrm{Y}})}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}} + \frac{(y-m_{\mathrm{Y}})^2}{\sigma_{\mathrm{Y}}^2}\right)\right\}.$$

Требуется по результатам наблюдений двумерной случайной величины $(\mathrm{X}_i, \mathrm{Y}_i)$, где $i = 1, 2, \ldots, n$, оценить параметры m_{X}, m_{Y}, σ_{X}, σ_{Y} и ρ.

Решение

Применяя метод моментов, имеем

$$\nu_{10} = \nu_{10}^* \;\Rightarrow\; \hat{m}_{\mathrm{X}} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad \nu_{01} = \nu_{01}^* \;\Rightarrow\; \hat{m}_{\mathrm{Y}} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i;$$

$$\mu_{20} = \nu_{20}^* \;\Rightarrow\; \hat{\sigma}_{\mathrm{X}}^2 = s_{\mathrm{X}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2;$$

$$\mu_{02} = \nu_{02}^* \;\Rightarrow\; \hat{\sigma}_{\mathrm{Y}}^2 = s_{\mathrm{Y}}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2;$$

$$\frac{\mu_{11}}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} \;\Rightarrow\; \hat{\rho} = \frac{1}{n}\frac{1}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Пример 3.8

Пусть СВ X имеет нормальное распределение с известным средним квадратическим отклонением $\sigma = 2$.

Требуется найти доверительный интервал для математического ожидания a, если среднее арифметическое значение результатов $n = 16$ прямых равноточных измерений $\overline{x} = 20,09$. Доверительную вероятность $(1 - \alpha)$ принять равной 0,90.

Решение

Пользуясь таблицами функции Лапласа, по заданной вероятности $(1 - \alpha)$ находим $u_{\alpha/2} = u_{0,05} = 1,64$. Найдем точность (предельную погрешность) оценки

$$\varepsilon \;=\; 1,64\,\frac{2}{\sqrt{16}} \;=\; 0,82.$$

Следовательно, искомый доверительный интервал составляет $[20,09 - 0,82;\ 20,09 + 0,82]$, или $[19,27;\ 20,91]$.

Смысл полученного результата: если будет произведено достаточно большое число выборок данного объема, то в 90 % из них доверительные интервалы накроют математическое ожидание и только в 10 % случаев оцениваемое математическое ожидание может выйти за границы доверительных интервалов.

Пример 3.9

Пусть $x_1 = 2,015;\ x_2 = 2,020;\ x_3 = 2,025;\ x_4 = 2,020;\ x_5 = 2,015$ — результаты независимых равноточных измерений толщины металлической пластинки.

Требуется:

1. Оценить с помощью доверительного интервала истинную толщину (математическое ожидание) пластинки. Доверительную вероятность $1 - \alpha$ принять равной 0,95.

2. Найти минимальное число измерений, которые надо выполнить, чтобы с надежностью $1 - \alpha = 0,95$ можно было утверждать, что предельная погрешность точечной оценки истинной толщины металлической пластинки не превышает 0,95.

Решение

Будем считать результаты измерения наблюденными значениями случайной величины X, которая распределена по нормальному закону с неизвестными параметрами a и σ. Найдем точечные оценки этих параметров.

Номер наблюдения	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
1	2,015	–0,004	0,000016
2	2,020	+0,001	0,000001
3	2,025	+0,006	0,000036
4	2,020	+0,001	0,000001
5	2,015	–0,004	0,000016

Следовательно,

$$\hat{a} \;=\; \overline{x} \;=\; \frac{10,095}{5} \;=\; 2,019;$$

$$\hat{\sigma} \;=\; s_{\text{несм}} \;=\; \sqrt{\frac{0,00007}{4}} \;=\; 0,004183.$$

По таблицам распределения Стьюдента (см. приложение) по доверительной вероятности $1-\alpha = 0,95$ и числу степеней свободы $\nu = n-1 = 4$ находим квантиль распределения $t_{0,025;\,4} \;=\; 2,776$. Следовательно, предельная погрешность точечной оценки

$$\varepsilon \;=\; \frac{s}{\sqrt{n}}\, t_{\alpha/2;\,n-1} \;=\; \frac{0,004183}{\sqrt{5}} \cdot 2,776 \;=\; 0,0052.$$

Доверительный интервал равен $[\overline{x}-\varepsilon;\ \overline{x}+\varepsilon] \;=\; [2,0138;\ 2,0242]$.

Для расчета минимального числа измерений, необходимых для определения истинной толщины пластинки с погрешностью, не превышающей 0,003, применим формулу $\varepsilon \;=\; st_{\alpha/2,\,n-1}/\sqrt{n}$.

Следовательно,

$$n \;\geq\; \varepsilon^{-2} t^2_{\alpha/2,\,n-1} s^2 \;=\; \frac{2,776^2 \cdot 0,004183^2}{0,000009} \;\approx\; 15 \text{ измерений.}$$

Таким образом, для того чтобы предельная погрешность ε истинного значения толщины пластинки не превышала $\varepsilon \;=\; 0,003$, следует, помимо 5 проделанных измерений, выполнить еще $(n-5) = 15-5 = 10$ измерений.

Пример 3.10 (Увеличение выборок)

Имеется выборка объемом N. Для неё найдены выборочное среднее x^*_N и выборочная дисперсия D^*_N. Получено еще одно значение x_{N+1}. Рассматривается объединенная выборка объемом $N+1$.

Требуется выразить выборочные оценки x^*_{N+1} и D^*_{N+1} через уже имеющиеся оценки x^*_N, D^*_N и новое значение x_{N+1}.

Решение

Для выборочного среднего имеем

$$x^*_{N+1} \;=\; \frac{1}{N+1} \sum_{i=1}^{N+1} x_i \;=\; \frac{1}{N+1} \left(\sum_{i=1}^{N} x_i + x_{N+1} \right) \;=\; \frac{1}{N+1} \left(N x^*_N + x_{N+1} \right). \qquad (1)$$

Найденное выборочное среднее x^*_{N+1} используем для нахождения выборочной дисперсии

$$D^*_{N+1} \;=\; \frac{1}{N+1} \sum_{i=1}^{N+1} \left(x_i - x^*_{N+1} \right)^2. \qquad (2)$$

В круглых скобках под знаком суммирования вычтем и добавим выборочное среднее x^*_N. Это дает

$$D^*_{N+1} \;=\; \frac{1}{N+1} \sum_{i=1}^{N+1} \left[(x_i - x^*_N)^2 + 2\,(x_i - x^*_N) \left(x^*_N - x^*_{N+1} \right) + \left(x^*_N - x^*_{N+1} \right)^2 \right]. \qquad (3)$$

Первое слагаемое в (3) равно

$$\frac{1}{N+1} \left[\sum_{i=1}^{N} (x_i - x^*_N)^2 + (x_{N+1} - x^*_N)^2 \right] \;=\; \frac{1}{N+1} \left[N D^*_N + (x_{N+1} - x^*_N)^2 \right]. \qquad (4)$$

Аналогично преобразовывая второе и третье слагаемые в (3), получим

$$D^*_{N+1} \;=\; \frac{1}{N+1}\left[ND^*_N + N\left(x^*_{N+1} - x^*_N\right)^2 + \left(x_{N+1} - x^*_{N+1}\right)^2\right]. \qquad (5)$$

Пример 3.11 (Объединение выборок)

Имеются две выборки объемом N и M соответственно. Для них найдены выборочные средние x^*_N, x^*_M и выборочные дисперсии D^*_N, D^*_M. Рассматривается объединенная выборка объемом $N+M$.

Требуется выразить выборочные оценки x^*_{N+M} и D^*_{N+M} через имеющиеся оценки x^*_N, D^*_N и x^*_M, D^*_M.

Решение

Имеем

$$x^*_{N+M} \;=\; \frac{1}{N+M}\left(Nx^*_N + Mx^*_M\right).$$

Найденное выборочное среднее x^*_{N+M} используем для нахождения выборочной дисперсии

$$D^*_{N+M} \;=\; \frac{1}{N+M}\left[ND^*_N + N\left(x^*_{N+M} - x^*_N\right)^2 + MD^*_M + M\left(x^*_{N+M} - x^*_M\right)^2\right].$$

Пример 3.12

Найти минимальный объем выборки, на основании которой можно было бы оценить математическое ожидание параметра некоторой технической операции с ошибкой, не превышающей 10, и надежностью $(1-\alpha) = 0,95$, если предположить, что параметр этой технической операции X является случайной величиной, имеющей нормальное распределение $\mathrm{X} \to N(a;\, 50)$.

Решение

Из условия следует, что дисперсия случайной величины X известна: $\sigma^2 = 50^2$. Воспользуемся формулой, связывающей предельную погрешность ε оценки математического ожидания по средней арифметической:

$$\varepsilon \;=\; u_{\alpha/2}\,\frac{\sigma}{\sqrt{n}},$$

отсюда

$$n \;=\; \frac{1}{\varepsilon^2}\,u^2_{\alpha/2}\sigma^2.$$

Пользуясь таблицей функции Лапласа, по доверительной вероятности $1-\alpha = 0,95$ находим квантиль $u_{0,025} = 1,96$. Следовательно,

$$n \;=\; \frac{1}{100}\,1,96^2 \cdot 2500 \;=\; 96 \text{ измерений.}$$

Пример 3.13 (Неравенство Крамера-Рао)

Пусть $\{x_1, x_2, \ldots, x_n\}$ – независимая выборка, в которой каждая компонента x_k $(k = 1, 2, \ldots, n)$ имеет плотность распределения $f(x;\theta)$, где θ – неизвестный

параметр. Пусть $\theta^* = \theta^*(x_1, x_2, \ldots, x_n) \equiv \theta^*(x)$ – несмещенная оценка параметра θ. Тогда $\mathrm{M}[\theta^*] = \theta$. Это равенство можно записать в виде

$$\theta \;=\; \int_{\theta^*} \theta^* f_\theta(x)\, dx, \tag{1}$$

где

$$f_\theta(x) \;=\; \prod_{i=1}^{n} f(x_i;\theta)$$

– совместная плотность распределения выборки $\{x_1, x_2, \ldots, x_n\}$.

Требуется получить выражение, ограничивающее дисперсию $\mathrm{D}[\theta^*]$ снизу.

Решение

Выражение в правой части (1) дифференцируемо по θ так же, как и выражение

$$\int_{\theta^*} f_\theta(x)\, dx \;=\; 1. \tag{2}$$

Легко видеть, что следующее выражение положительно:

$$I(\theta) \;=\; \int \left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2 f_\theta(x)\, dx \;=\; \mathrm{M}\left[\left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2\right]. \tag{3}$$

Так как $f_\theta(x)$ – плотность распределения, то справедливо

$$\int f_\theta(x)\, dx \;=\; 1,$$

где $x = (x_1, x_2, \ldots, x_n)$. Продифференцировав (1) и (3) по θ, найдем

$$1 \;=\; \int_{\theta^*} \theta^*(x)\, \frac{\partial f_\theta(x)}{\partial\theta}\, dx, \qquad 0 \;=\; \int_{\theta^*} \frac{\partial f_\theta(x)}{\partial\theta}\, dx. \tag{4}$$

Умножим теперь второе равенство в (4) на θ и вычтем из первого, тогда получим

$$1 \;=\; \int_{\theta^*} [\theta^*(x) - \theta]\, \frac{\partial f_\theta(x)}{\partial\theta}\, dx.$$

Так как $f_\theta(x) > 0$, то $\partial f_\theta(x)/\partial\theta = f_\theta(x)\, \partial \ln f_\theta(x)/\partial\theta$. Подставим это выражение в (4), которое возведем в квадрат, далее, используя неравенство Коши-Буняковского, находим

$$1 \;=\; \left(\int [\theta^*(x) - \theta]\, \frac{\partial f_\theta(x)}{\partial\theta}\, f_\theta(x)\, dx\right)^2 \;\le$$

$$\le\; \int [\theta^*(x) - \theta]^2\; f_\theta(x)\, dx \cdot \int \left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2 \; f_\theta(x)\, dx.$$

Иначе,

$$1 \;\le\; \mathrm{D}[\theta^*] \cdot \mathrm{M}\left[\left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2\right]. \tag{5}$$

Выразим теперь второй сомножитель в (5) через функцию $I(\theta)$, определенную в (3). Из (4) при $n = 1$ следует, что

$$\mathrm{M}\left[\frac{\partial \ln f(x_1;\theta)}{\partial\theta}\right] \;=\; 0. \tag{6}$$

Поскольку

$$\ln f_\theta(x) \;=\; \sum_{i=1}^{n} \ln f(x_i;\theta),$$

то

$$\left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2 \;=\; \sum_{i=1}^{n}\sum_{k=1}^{n} \frac{\partial \ln f_\theta(x_i;\theta)}{\partial\theta}\,\frac{\partial \ln f_\theta(x_k;\theta)}{\partial\theta}.$$

Отсюда, учитывая независимость сомножителей при $i \neq k$, находим

$$\mathrm{M}\left[\left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2\right] \;=\; n\,\mathrm{M}\left[\left(\frac{\partial \ln f(x_1;\theta)}{\partial\theta}\right)^2\right] \;=\; nI_1(\theta),$$

где $I_1(\theta)$ – количество информации о параметре θ, содержащееся в одном наблюдении.

Итак, приходим к выражению

$$\mathrm{D}[\theta^*] \;\geq\; \frac{1}{nI_1(\theta)} \tag{7}$$

– *неравенство Крамера-Рао.*

3.11. Задачи для решения

Задача 3.1

В результате проведения n независимых экспериментов в одних и тех же условиях случайное событие A произошло m раз.

а) Показать, что относительная частота $f = m/n$ появления события A будет несмещенной и состоятельной оценкой события A: $\Pr(\mathrm{A}) = p$ в одном эксперименте.

б) Определить такое значение p, при котором дисперсия относительной частоты f будет максимальной.

Задача 3.2

Отказ прибора произошел при k-м испытании.

Найти оценку максимального правдоподобия вероятности отказа p при одном испытании и вычислить её математическое ожидание.

Задача 3.3

Построить функцию правдоподобия выборки из n независимых величин, распределенных по экспоненциальному закону с одинаковым параметром.

Задача 3.4

Имеется выборка объема n, извлеченная из распределения Лапласа с плотностью $f(x) = \alpha^{-1}\exp(-|x-\mu|/\alpha); \quad -\infty < x < \infty$.

Найти оценки максимального правдоподобия для параметров α и μ.

Задача 3.5

Для того чтобы событие A произошло ровно m раз, было проведено n испытаний $(n \geq m)$. Найти оценку максимального правдоподобия вероятности p появления события A в одном испытании.

Задача 3.6

Из равномерного распределения извлечена выборка $x_1, x_2, \ldots, x_n$ объема n. В качестве статистической оценки $\hat{l}$ длины интервала $(a; b)$ предлагается широта выборки $\hat{l} = \eta - \xi$, где $\xi = \min\limits_i\{x_i\}$, $\eta = \max\limits_i\{x_i\}$.

Будет ли эта оценка обладать свойствами эффективности и состоятельности? Привести численный пример.

Задача 3.7

Из генеральной совокупности, распределенной по нормальному закону с параметрами $(m; 1)$, извлечена выборка объема $n = 3$ и построен вариационный ряд: $x_1 \leq x_2 \leq x_3$. Для неизвестного математического ожидания m рассматривается оценка: $\theta = \gamma x_1 + (1-2\gamma)x_2 + \gamma x_3$, где γ – параметр.

Изучить свойства предлагаемой оценки. При каком значении параметра γ эта оценка будет обладать свойствами несмещенности и эффективности?

Задача 3.8

Из распределения с плотностью $f(x) = \exp(\alpha - x)$, где $x \geq \alpha$, извлечена выборка $x_1, x_2, \ldots, x_n$ объема n. Для статистической оценки неизвестного параметра α предлагается $\hat{\alpha} = \min\limits_i\{x_i\}$.

Будет ли эта оценка обладать свойствами эффективности и состоятельности? Привести численный пример.

Задача 3.9

При помощи n различных приборов получены n независимых измерений случайной величины X.

В предположении, что X имеет нормальное распределение, а дисперсия i-го измерения известна и равна σ_i^2 $(i = 1, 2, \ldots, n)$, найти оценку максимального правдоподобия математического ожидания случайной величины X. Показать, что полученная оценка является несмещенной, и вычислить её дисперсию.

Задача 3.10

Пусть $(x_1, x_2, \ldots, x_n)$ – выборка из генеральной совокупности, имеющей равномерное распределение на интервале с фиксированными границами.

Найти оценки максимального правдоподобия для границ a и b интервала.

Задача 3.11

Из генеральной совокупности, имеющей плотность распределения вероятностей

$$f(x) \;=\; \frac{1}{2}\exp(-|x|)/[1-\exp(-\beta)],$$

где $-\beta < x < \beta$, извлечена выборка $x_1, x_2, \ldots, x_n$ объема n.

Построить оценку $\hat{\beta}$ для параметра β. Изучить свойства предлагаемой оценки.

Задача 3.12

Оценка величины сопротивления для большой партии резисторов, определенная по результатам 100 случайно отобранных экземпляров, равна $\overline{x} = 10$ *кОм*.

а) Считая, что дисперсия измерений известна: $\sigma^2 = 1$ *кОм*^{2}, найти вероятность того, что для резисторов всей партии величина сопротивления лежит в пределах $10 \pm 0,1$ *кОм*.

б) Сколько измерений нужно произвести, чтобы с вероятностью 0,95 утверждать, что для всей партии резисторов величина сопротивления лежит в пределах $10 \pm 0,1$ *кОм*?

Задача 3.13

Выборка $x_1, x_2, \ldots, x_n$ объема n извлечена из равномерного распределения на отрезке $[a, b]$. Известна длина этого отрезка $l = b - a$, но неизвестна середина интервала $c = (a+b)/2$. В качестве статистической оценки середины интервала предлагается среднее арифметическое экстремальных значений выборки $\overline{c} = (\xi + \eta)/2$, где $\xi = \min\limits_i\{x_i\}$, $\eta = \max\limits_i\{x_i\}$.

Будет ли эта оценка обладать свойствами эффективности и состоятельности? Привести численный пример.

Задача 3.14

Осуществлены две серии из n_1 и n_2 независимых испытаний, причем в первой серии событие A произошло m_1 раз, а во второй серии – m_2 раз.

Найти оценку максимального правдоподобия для неизвестной вероятности p события A в каждом испытании (считая эту вероятность постоянной).

Задача 3.15

Найти оценку максимального правдоподобия $\hat{\sigma}$ параметра σ по выборке объема n из нормально распределенной генеральной совокупности с известным математическим ожиданием m.

Показать, что полученная оценка максимального правдоподобия является несмещенной.

Задача 3.16

Построить функцию правдоподобия выборки из n независимых величин, распределенных по геометрическому закону с одинаковым параметром.

3.12. Задание на практическую работу

Настоящая практическая работа рассчитана на два часа и содержит два задания. Задания должны выполняться в выбранной программной среде.

Задание 1

Напишите программу, с помощью которой исследуйте зависимость выборочного среднего x^*_N от объема выборки N. Результаты оформите графически.

Вариант 1

Равномерный закон.

Исходные данные для программы:

a – левая граница возможных значений;

b – правая граница возможных значений;

N – объем выборки.

Результат работы программы – массив, содержащий значения выборочного среднего x^*_N от объема выборки N.

Вариант 2

Нормальный закон Гаусса.

Исходные данные для программы:

m_{x} – математическое ожидание;

σ^2_{x} – дисперсия;

N – объем выборки.

Результат работы программы – массив, содержащий значения выборочного среднего x^*_N от объема выборки N.

Задание 2

Напишите программу, с помощью которой исследуйте зависимость выборочной дисперсии D^*_N от объема выборки N. Результаты оформите графически.

Вариант 1

Равномерный закон.

Исходные данные для программы:

a – левая граница возможных значений;

b – правая граница возможных значений;

N – объем выборки.

Результат работы программы – массив, содержащий значения выборочной дисперсии D^*_N от объема выборки N.

Вариант 2

Нормальный закон Гаусса.

Исходные данные для программы:

m_{x} – математическое ожидание;

σ^2_{x} – дисперсия;

N – объем выборки.

Результат работы программы – массив, содержащий значения выборочной дисперсии D^*_N от объема выборки N.

3.13. Задания для проверки

1. В чем заключается сущность задачи нахождения точечных оценок неизвестных параметров распределения?

2. Какая оценка параметра называется состоятельной? Почему желательно, чтобы оценка была состоятельной?

3. Какая оценка параметра называется несмещенной? Почему желательно, чтобы оценка была несмещенной?

4. Какие оценки параметров называются эффективными, достаточными?

5. В чем заключается сущность метода моментов точечной оценки параметров?

6. В чем заключается сущность метода наибольшего правдоподобия точечной оценки параметров?

7. Сравните свойства, которыми обладают точечные оценки параметров, найденные по методу моментов, по методу наибольшего правдоподобия.

8. Какая оценка вероятности наступления события в схеме Бернулли обладает свойством состоятельности, несмещенности и эффективности?

9. Какая оценка математического ожидания случайной величины, подчиненной нормальному распределению, обладает свойствами состоятельности, несмещенности и эффективности?

10. Приведите пример оценок математического ожидания нормальной случайной величины, не обладающих указанными свойствами.

11. Какая оценка среднего квадратического отклонения нормально распределенной случайной величины не обладает свойством несмещенности? не обладает свойством эффективности, обладает свойством состоятельности?

12. Что называется доверительным интервалом, доверительной вероятностью?

13. Что называется предельной погрешностью точечной оценки параметра?

14. Что происходит с длиной доверительного интервала при увеличении объема выборки? увеличении доверительной вероятности?

15. Являются ли концы доверительных интервалов постоянными величинами, случайными величинами?

16. Сформулируйте общую схему построения доверительных интервалов.

17. Как строится доверительный интервал для математического ожидания случайной величины X, распределенной по нормальному закону?

18. Как строится доверительный интервал для среднего квадратического отклонения σ случайной величины X, распределенной по нормальному закону?

4. Статистическая проверка параметрических гипотез

4.1. Постановка задачи. Основные определения

Ранее были рассмотрены методы получения оценок неизвестных параметров распределения. Нахождение точечных или интервальных оценок является, как правило, некоторой предварительной стадией статистических исследований. Конечная цель исследования может, например, заключаться в сравнительной оценке различных технологических процессов по их производительности, точности или экономичности, в сравнении характеристик приборов, изделий и т. п. Задачи такого рода носят название – *задачи сравнения*.

На математическом языке задачи сравнения формулируются как задачи статистической проверки гипотез относительно параметров законов распределения, так как изменение параметров характеризует различия технологических процессов, конструкций, приборов и т. п.

Предположим, что для решения задачи сравнения из генеральной совокупности извлечена выборка $(x_1, x_2, \ldots, x_n)$ объема n. Пусть, далее, экспериментатор визуально – по виду гистограммы или полигона относительных частот или из каких-либо других соображений – выдвинул гипотезу о законе распределения исследуемой случайной величины (СВ) X и, оценив параметры этого закона, построил теоретико-вероятностную модель распределения вероятностей этой случайной величины, которая, по его мнению, отражает в себе основные особенности статистического ряда.

Определение 1. Статистическая гипотеза называется *непараметрической*, если в ней сформулированы предположения относительно вида функции распределения.

Дальнейшая задача экспериментатора состоит в проверке выдвинутой гипотезы, т. е. в выяснении, насколько хорошо подобрана вероятностная модель. Для проверки этой гипотезы используются различные статистические критерии, методика применения которых будет дана несколько позже.

Предположим, что с помощью этих критериев экспериментатор убедился, что модель "хорошая", опытные данные не противоречат этой модели, т. е. он "погрузил" опытные данные $(x_1, x_2, \ldots, x_n)$ в модель удачно.

Если теперь, например, оценивать параметр θ вероятностной модели $F(x; \theta)$ статистического ряда наблюдений по двум независимым выборкам, взятым, скажем, до введения некоторого усовершенствования в технологический процесс и после введения, то получим две оценки $\hat{\theta}_1$ и $\hat{\theta}_2$, которые в силу их случайного характера не будут равны между собой. Спрашивается (*выдвигается гипотеза*), являются ли эти оценки оценками одного и того же параметра θ вероятностной модели или

в связи с введением усовершенствования в технологический процесс этот параметр изменился ?

Задачи такого рода являются типичными задачами сравнения.

Определение 2. Статистическая гипотеза называется *параметрической*, если в ней сформулированы предположения относительно значений параметров функции распределения известного вида.

Наиболее точные и безошибочные суждения относительно истинности такого рода гипотез можно было бы сделать при исследовании всей генеральной совокупности. Однако в большинстве практических случаев сплошное исследование по ряду причин провести нельзя. Например, объем генеральной совокупности часто бывает бесконечным, сплошное обследование всей генеральной совокупности требует больших материальных затрат. Кроме того, такое обследование может привести к порче предмета, а результаты в момент окончания исследования могут оказаться неактуальными.

Таким образом, суждения об истинности (ложности) статистических гипотез относительно вида функции распределения генеральной совокупности $F(x;\theta_i)$ или о значениях параметров функции распределения известного вида принимаются на основании выборки объема n.

Процесс использования выборки для проверки истинности (ложности) статистических гипотез называется *статистическим доказательством истинности (ложности) выдвинутой гипотезы.*

Наряду с выдвинутой гипотезой рассматривают одну или несколько альтернативных (конкурирующих) гипотез. Если выдвинутая гипотеза будет отвергнута, то её место занимает альтернативная гипотеза. С этой точки зрения статистические гипотезы подразделяются на *нулевые* и *альтернативные.*

Определение 3. *Нулевой гипотезой* называют основную (выдвинутую) гипотезу.

Нулевую гипотезу обозначают символом H_0.

Обычно нулевые гипотезы утверждают, что различие между сравниваемыми величинами (параметрами или функциями распределения) отсутствует, а наблюдаемые отклонения объясняются лишь случайными колебаниями выборки.

Определение 4. *Альтернативной гипотезой* называется гипотеза, конкурирующая с нулевой гипотезой в том смысле, что если нулевая гипотеза отвергается, то принимается альтернативная.

Альтернативную гипотезу обозначают символом H_a или H_1.

Приведем примеры нулевых и альтернативных статистических гипотез параметрического вида. Предположим, что по выборке $(x_1, x_2, \ldots, x_n)$ построена нормальная модель с параметрами a и σ. Тогда относительно параметров генеральной совокупности a и σ можно выдвинуть следующие гипотезы :

Нулевые	Альтернативные
1. $\{H_0 :\ a = a_0\}$.	1. $\{H_a :\ a \neq a_0\}$; $\quad a < a_0;\ a > a_0$.
2. $\{H_0 :\ \sigma = \sigma_0\}$.	2. $\{H_a :\ \sigma \neq \sigma_0\}$; $\quad \sigma < \sigma_0;\ \sigma > \sigma_0$.
3. $H_0 :\ \begin{cases} a = a_0; \\ \sigma = \sigma_0. \end{cases}$	3. $H_a :\ \begin{cases} a \neq a_0; \\ \sigma \neq \sigma_0. \end{cases}$

Содержательная сущность этих гипотез может принимать различный вид в зависимости от конкретной задачи исследования.

Рассмотрим, например, следующую задачу: будет ли усовершенствованный способ изготовления электроламп увеличивать срок их службы по сравнению с существующим способом, при котором средний срок службы $a_0 = 4500$ *час*? Испытания небольшой партии электроламп дали $x = 4800$ *час*.

Можно ли на основании этих данных считать, что новый способ производства электроламп лучше старого?

Можно выдвинуть нулевую и альтернативную гипотезы:

$\{H_0 : \ a = a_0\}$ – *при новом способе производства электроламп срок службы остался прежним*;

$\{H_a : \ a > a_0\}$ – *при новом способе производства электроламп срок их службы увеличился*.

Если экспериментатор приходит к выводу, что нормальная модель удачно отражает в себе закономерности двух статистических рядов $(x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_1}^{(1)})$ и $(x_1^{(2)}, x_2^{(2)}, \ldots, x_{n_2}^{(2)})$, то относительно параметров этой модели могут быть выдвинуты следующие нулевые и соответствующие им альтернативные гипотезы:

Нулевые	Альтернативные
1. $\{H_0 : \ a_1 = a_2\}$.	1. $\{H_a : \ a_1 \neq a_2\}$; $a_1 < a_2$; $a_1 > a_2$.
2. $\{H_0 : \ \sigma_1 = \sigma_2\}$.	2. $\{H_a : \ \sigma_1 \neq \sigma_2\}$; $\sigma_1 < \sigma_2$; $\sigma_1 > \sigma_2$.
3. $H_0 : \begin{cases} a_1 = a_2, \\ \sigma_1 = \sigma_2. \end{cases}$	3. $H_a : \begin{cases} a_1 \neq a_2; \\ \sigma_1 \neq \sigma_2. \end{cases}$

Нулевую гипотезу

$$H_0 : \begin{cases} a_1 = a_2, \\ \sigma_1 = \sigma_2 \end{cases}$$

словесно можно сформулировать следующим образом: *две выборки извлечены из одной и той же генеральной совокупности*; альтернативную гипотезу

$$H_a : \begin{cases} a_1 \neq a_2, \\ \sigma_1 \neq \sigma_2 \end{cases}$$

– *две выборки извлечены из разных генеральных совокупностей*.

Статистические параметрические гипотезы могут содержать одно или несколько предположений относительно параметров распределения функции.

Определение 5. Параметрическая гипотеза называется *простой*, если содержит только одно предположение относительно параметра.

Например, если a – математическое ожидание нормально распределенной случайной величины, то гипотеза $\{H_0 : \ a = 0\}$ – простая.

Определение 6. Параметрическая гипотеза называется *сложной*, если она состоит из конечного или бесконечного числа простых гипотез.

Например, если a – математическое ожидание нормально распределенной случайной величины, то гипотеза $\{H_a : a > 0,5\}$ является сложной, так как она состоит из бесконечного множества простых гипотез вида $\{H_a : \ a = a_i\}$, где a_i – любое заданное число, большее 0,5.

4.2. Статистический критерий значимости проверки нулевой гипотезы

Проверка статистических гипотез осуществляется на основе данных выборки. Для этого используют специальным образом подобранную случайную величину (выборочную статистику), являющуюся функцией наблюденных значений, точное или приближенное распределение которой известно. Эту выборочную статистику обозначают различными буквами в зависимости от закона её распределения, например, u, если она распределена по нормальному закону, F – если она имеет распределение Фишера, t – если она имеет распределение Стьюдента. В данном параграфе в целях общности будем обозначать её через K.

Определение 1. *Статистическим критерием (тестом)* K называют случайную величину K, с помощью которой принимаются решения о принятии или отклонении выдвинутой нулевой гипотезы.

Для проверки нулевых гипотез по выборочным данным вычисляют частные значения входящих в критерий величин и, таким образом, получают частное (наблюдаемое) значение критерия.

Пусть для проверки некоторой нулевой гипотезы H_0 относительно параметров распределения служит выборочная статистика (критерий) K. Предположим, что плотность распределения вероятностей выборочной статистики K при условии справедливости проверяемой гипотезы H_0 равна $f(K|H_0)$, а математическое ожидание статистики K равно K_0.

Тогда вероятность того, что случайная величина K попадет в произвольный интервал $[K_{1-\alpha/2};\, K_{\alpha/2}]$, можно найти по формуле (рис. 4.1)

$$\Pr\left(K_{1-\alpha/2} < \mathrm{K} < K_{\alpha/2}\right) = \int\limits_{K_{1-\alpha/2}}^{K_{\alpha/2}} f(K|H_0)\, dK. \tag{4.1}$$

Зададим эту вероятность равной $1-\alpha$ и вычислим квантили $K_{1-\alpha/2}$ и $K_{\alpha/2}$ плотности распределения $f(K|H_0)$ при условиях:

$$\Pr\left(\mathrm{K} \le K_{1-\alpha/2}\right) = \int\limits_{-\infty}^{K_{1-\alpha/2}} f(K|H_0)\, dK = \frac{\alpha}{2}, \tag{4.2a}$$

$$\Pr\left(\mathrm{K} \ge K_{\alpha/2}\right) = \int\limits_{K_{\alpha/2}}^{\infty} f(K|H_0)\, dK = \frac{\alpha}{2}. \tag{4.2b}$$

Следовательно, вероятность того, что случайная величина K будет находиться внутри интервала $[K_{1-\alpha/2};\, K_{\alpha/2}]$, равна $1-\alpha$. Соответственно вероятность того, что случайная величина K будет находиться вне этого интервала, равна α.

Зададим вероятность α настолько малой, чтобы попадание случайной величины K за пределы интервала $[K_{1-\alpha/2};\, K_{\alpha/2}]$ можно было считать маловероятным событием. Тогда, исходя из принципа практической невозможности маловероятных

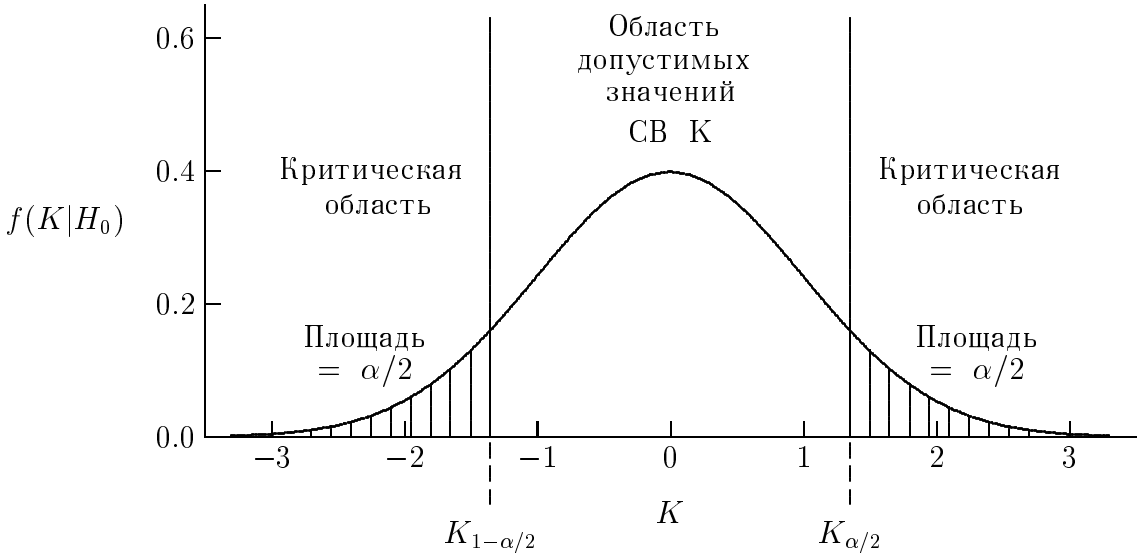


Рисунок 4.1 — Формирование критических областей и области допустимых значений случайного критерия K при заданном уровне значимости α (двусторонний случай)

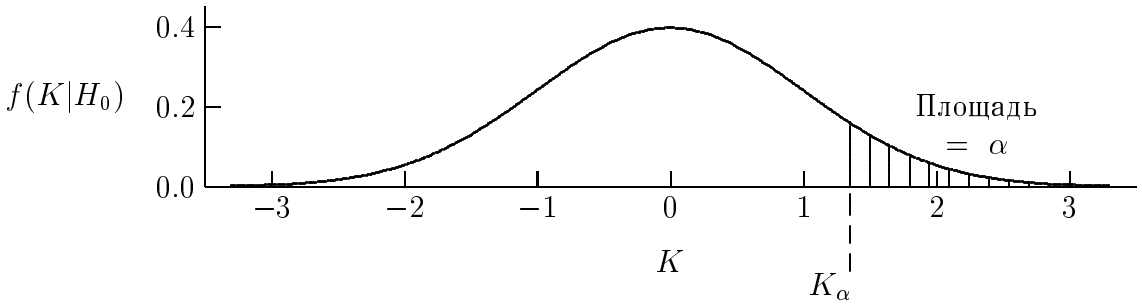


Рисунок 4.2 — К формированию правосторонней критической области критерия K при заданном уровне значимости α

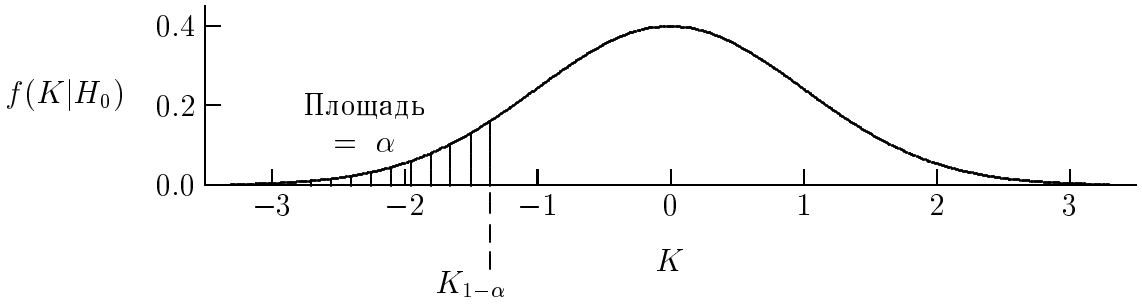


Рисунок 4.3 — К формированию левосторонней критической области критерия K при заданном уровне значимости α

событий, можно считать, что если нулевая гипотеза справедлива, то при её проверке с помощью критерия K по данным одной выборки наблюдаемое (частное) значение критерия K должно обязательно попасть в интервал $[K_{1-\alpha/2};\, K_{\alpha/2}]$.

Если же наблюдаемое (частное) значение критерия K попадает за пределы интервала $[K_{1-\alpha/2};\, K_{\alpha/2}]$, то произойдет маловероятное, практически невозможное событие, т.е. считается, что с вероятностью $(1-\alpha)$ проверяемая нулевая гипотеза не справедлива.

Поэтому область $[K_{1-\alpha/2};\, K_{\alpha/2}]$ называют *областью допустимых значений* случайной величины K, при которых нулевая гипотеза не отклоняется, области $(-\infty;\, K_{1-\alpha/2})$ и $(K_{\alpha/2};\, \infty)$ – *областями отклонения проверяемой нулевой гипотезы* или критической областью критерия K.

Если критические области располагаются справа и слева от математического ожидания случайной величины K, так, как это изображено на рис. 4.1, то критическая область называется *двусторонней*, а критерий K *двусторонним критерием значимости.*

В некоторых случаях экспериментатор бывает твердо убежден, что $\mathrm{K} > K_0$ или $\mathrm{K} < K_0$. В этом случае критические области являются односторонними. Возможны *правосторонняя* и *левосторонняя* критические области.

На рис. 4.2 и 4.3 изображены правосторонняя и левосторонняя критические области соответственно.

Замечание.

Критическая точка K_α (*квантиль K-распределения*), отделяющая области принятия или отклонения проверяемой нулевой гипотезы, для критерия с правосторонней критической областью находится из условия

$$\Pr\left(\mathrm{K} \geq K_\alpha\right) \;=\; \int\limits_{K_\alpha}^{\infty} f(K|H_0)\, dK \;=\; \alpha. \tag{4.3}$$

Эта же критическая точка, отделяющая области принятия или отклонения проверяемой нулевой гипотезы, для критерия с левосторонней критической областью находится из условия

$$\Pr\left(\mathrm{K} \leq K_{1-\alpha}\right) \;=\; \int\limits_{-\infty}^{K_{1-\alpha}} f(K|H_0)\, dK \;=\; \alpha. \tag{4.4}$$

Итак, можно сформулировать процедуру проверки параметрических гипотез с помощью критериев значимости следующим образом.

Определение 2.

Проверка гипотезы с помощью статистического критерия значимости есть правило отклонения нулевой гипотезы. Это правило заключается в разбиении области возможных значений случайной величины K на две непересекающиеся подобласти. При этом нулевая гипотеза отвергается, если наблюдаемое (частное) значение критерия K принадлежит критической подобласти, и считается согласующейся с опытом, если наблюдаемое значение критерия K не принадлежит критической подобласти.

4.3. Ошибки, допускаемые при проверке гипотез. Уровень значимости статистического критерия

При проверке статистических гипотез по выборочным данным всегда существует риск принятия ложного решения. Это объясняется тем, что объем выборки конечен, и поэтому нельзя точно определить ни вид функции $F(x;\theta)$, ни значения её параметров. Однако при многократном применении критериев теория статистической проверки гипотез позволяет оценить вероятности принятия ложных решений, и если вероятности малы, то можно считать, что данный статистический критерий обеспечивает малый риск ошибки.

Возможные ошибки могут быть двоякого рода.

Определение 1. *Ошибкой первого рода* называется ошибка отклонения верной нулевой гипотезы H_0.

В предыдущем параграфе было показано, что нулевая гипотеза отклоняется, если наблюдаемое (частное) значение критерия K попадет в критическую область.

Определение 2. *Уровнем значимости статистического критерия* называется вероятность α совершения ошибки первого рода.

Отклонение нулевой гипотезы H_0 на уровне значимости $\alpha = 0,05$ означает, что, отклоняя эту гипотезу, мы или не ошибаемся (т. е. гипотеза H_0 действительно ложная) или все-таки совершаем ошибку первого рода, считая правильную гипотезу H_0 ложной. В последнем случае ($\alpha = 0,05$) частота принятия ошибочного решения равна в среднем 5 на 100 случаев применения данного статистического критерия значимости.

Определение 3. *Ошибкой второго рода* называется ошибка принятия ложной гипотезы H_0.

Вероятность совершения ошибки второго рода принято обозначать β. Если $f(K|H_a)$ – плотность распределения выборочной статистики K при условии, что альтернативная гипотеза H_a является верной, то вероятность совершения ошибки второго рода для критерия с левосторонней критической областью можно вычислить по формуле

$$\beta \;=\; \Pr\left(\mathrm{K} \geq K_{1-\alpha}\right) \;=\; \int\limits_{K_{1-\alpha}}^{\infty} f(K|H_a)\,dK. \tag{4.5}$$

Определение 4. *Мощностью M критерия* K называется вероятность $(1-\beta)$ несовершения ошибки второго рода (мощность критерия K – это вероятность отклонения неверной гипотезы H_a), т. е. $M = 1-\beta$.

На рис. 4.4 дана геометрическая интерпретация вероятностей ошибок первого рода, второго рода и мощности критерия K, имеющего левостороннюю критическую область.

Из рис. 4.4 видно, что, передвигая квантиль $K_{1-\alpha}$ влево (уменьшая ошибку первого рода), мы тем самым увеличиваем ошибку второго рода. Можно показать, что вероятность совершения ошибки второго рода является функцией нескольких переменных: числа измерений n, уровня значимости α, характера альтернативной

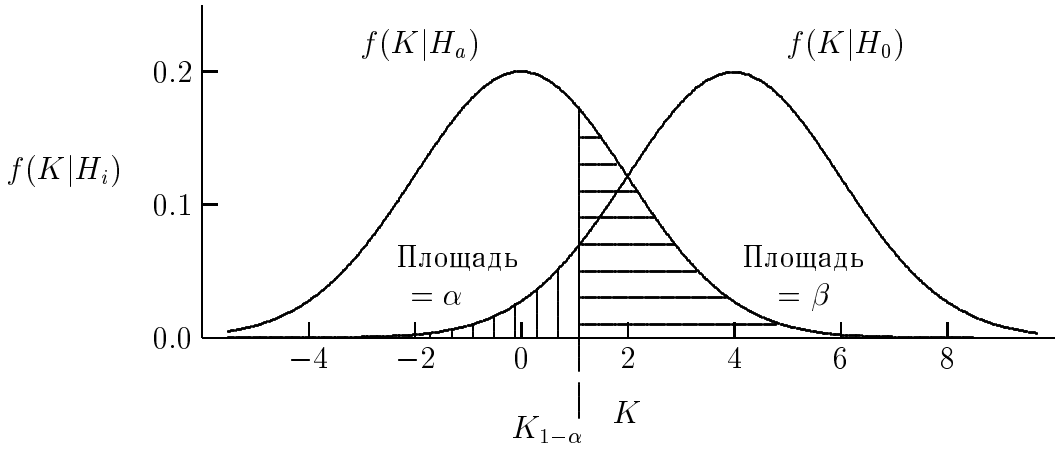


Рисунок 4.4 — Ошибки первого и второго рода критерия K, имеющего левостороннюю критическую область

гипотезы H_a, применяемого критерия K, т. е.

$$\beta = f(n, \alpha, H_a, K).$$

При этом выполняются следующие предельные соотношения:

$$\lim_{n\to\infty} \alpha = 0; \tag{4.6}$$

$$\lim_{n\to\infty} \beta = 0; \tag{4.7}$$

$$\lim_{\alpha\to 0} \beta = 1. \tag{4.8}$$

Равенства (4.6) и (4.7) говорят о том, что статистические доказательства истинности гипотез H_0 и H_a становятся достоверными только при бесконечно большом объеме выборки. Кроме того, из этих предельных соотношений следует, что единственным способом одновременного уменьшения вероятностей ошибок первого и второго рода является увеличение объема выборки.

Равенство (4.8) говорит о том, что, уменьшая вероятность ошибки первого рода до нуля, мы при фиксированном объеме выборки допускаем неограниченный риск сделать ошибку второго рода.

Как же следует выбирать уровень значимости α статистических критериев? Ответ на этот вопрос зависит от потерь (тяжести последствий), вызываемых ошибками первого и второго рода. Например, если совершение ошибки первого рода приведет к большим потерям по сравнению с потерями, совершаемыми при ошибке второго рода, то следует принять по возможности меньшее значение α. Конечно, нельзя положить $\alpha = 0$, так как из $\alpha = 0$ следует $\beta = 1$, следовательно, в этом случае будут приниматься все нулевые гипотезы, в том числе и неправильные.

В данном пособии мы будем рассматривать только один вид статистических критериев – статистические критерии значимости, при применении которых вероятность совершения ошибки первого рода (уровень значимости α) фиксируется заранее.

Статистические критерии значимости – это односторонне действующие критерии, т. е. на основании их применения принимается с заранее фиксированным риском только одно решение: "*Отклонить проверяемую нулевую гипотезу*".

Если же для проверяемой нулевой гипотезы нет оснований отклонить её данным критерием, то утверждается: "*Результаты выборки не дают основания для отклонения выдвинутой нулевой гипотезы*".

Таким образом, статистические критерии значимости не позволяют принимать решение: "*Нулевая гипотеза H_0 является правильной*", так как при применении указанных критериев вероятность ошибки второго рода (вероятность принятия ложной нулевой гипотезы) остается неизвестной.

В большинстве случаев достаточно применения статистических критериев значимости. Действительно, практик-экспериментатор, как правило, хочет проверить, дают ли результаты эксперимента право отклонить нулевую гипотезу с тем, чтобы принять вместо неё некоторую альтернативную гипотезу, которую он отстаивает (новая технология изготовления изделий, усовершенствование некоторого узла автомобиля и т.д.). Доказательством истинности нулевой гипотезы, например, подтверждением эффективности старой технологии изготовления изделия, он не занимается.

Проиллюстрируем достаточность применения только статистических критериев значимости на примере, приведенном в начале настоящего раздела. Допустим, что предложена новая технология изготовления электроламп, предположительно увеличивающая срок их службы. Для доказательства увеличения срока службы электроламп произведено испытание по определению длительности их горения.

Выдвинем нулевую и альтернативную гипотезы :

$\{H_0 : \ a_1 = a_2\}$ – *при новой технологии срок службы электроламп остался прежним*;

$\{H_a : \ a_1 > a_2\}$ – *в результате новой технологии фактический срок службы электроламп увеличился*.

Если статистический критерий значимости отклонит выдвинутую нулевую гипотезу, то повышение срока службы электроламп, изготовленных по новой технологии, считается доказанным (с соответствующим малым риском совершения ошибки первого рода, измеряемым уровнем значимости α).

Если же на основании критерия значимости придем к выводу: "*Нет оснований для отклонения нулевой гипотезы*", то это означает, что имеющиеся статистические данные, которые должны свидетельствовать об эффективности новой технологии, не могут служить доказательством повышения срока службы электроламп.

Рассмотренный пример показывает, что в большинстве случаев для практических приложений достаточно статистических критериев значимости, позволяющих только отклонять выдвинутую нулевую гипотезу H_0 с фиксированной малой вероятностью ошибки первого рода (уровнем значимости α).

При использовании статистических критериев значимости выбор уровня значимости α до некоторой степени произволен, поскольку в большинстве случаев нет точной границы "разрешенной" вероятности ошибки первого рода α.

Стало практически обычным выбирать для α одно из стандартных значений: $\alpha = 0{,}005;\ 0{,}01;\ 0{,}05;\ 0{,}10$, хотя это не означает, что нельзя выбирать $\alpha = 0,03$. Принятая стандартизация имеет некоторое преимущество, так как она позволяет

сократить объем таблиц критических значений статистических критериев.

Следует учитывать, что чем меньше уровень значимости α, тем труднее отклонить нулевую гипотезу. Поэтому не следует стремиться выбирать уровень значимости α слишком малым. При проведении технических исследований наиболее часто принимают $\alpha = 0,05$ и только в исключительно важных исследованиях (например, медицинских) полагают $\alpha = 0,01$.

Замечания:

1. Если строгие критерии для проверки гипотез относительно параметров распределения СВ X, имеющей произвольный закон распределения, отличный от нормального, отсутствуют, то описываемые далее статистические критерии значимости можно считать приближенными (отклонение от нормального закона приводит к увеличению уровня значимости).

2. Существуют методы и критерии проверки статистических гипотез, учитывающие вероятности совершения ошибок как первого, так и второго рода, например, критерии, построенные по типу последовательного анализа, критерий Неймана–Пирсона, критерии, использующие теорию статистических решений. Такие критерии нашли, например, широкое применение в радиотехнике при выделении сигналов на фоне шумов.

3. Если в обычном критерии значимости зафиксировать заранее ошибки первого и второго рода, то, пользуясь кривыми мощностей, можно определить минимальный объем выборки, необходимый для различения гипотез H_0 и H_a с фиксированными вероятностями совершения ошибок первого и второго рода.

Определение наилучшей критической области для проверки простых гипотез.

На множестве значений статистики критерия можно выбрать сколько угодно критических областей V_k для заданного уровня значимости α, однако соответствующие им критерии будут иметь, вообще говоря, различные вероятности ошибок второго рода. *Наилучшей критической областью (НКО)* называют критическую область, которая при заданном уровне значимости α обеспечивает минимальную вероятность ошибки второго рода. Критерий, использующий НКО, имеет максимальную мощность.

При проверке простой гипотезы H_0 против альтернативы H_a НКО определяется *леммой Неймана-Пирсона*: НКО критерия заданного уровня значимости α состоит из точек выборочного пространства (выборок объема n), для которых удовлетворяется неравенство

$$\frac{L(x_1, x_2, \ldots, x_n \,|\, H_0)}{L(x_1, x_2, \ldots, x_n \,|\, H_a)} \;<\; C_\alpha, \tag{4.9}$$

где C_α – константа, зависящая от заданного уровня значимости, $x_1, x_2, \ldots, x_n$ – элементы выборки, а $L(x_1, x_2, \ldots, x_n|H_i)$ – функция правдоподобия, вычисленная при условии, что верна гипотеза H_i.

Перейдем к изложению конкретных статистических критериев значимости для проверки гипотез о параметрах нормального закона распределения. Необходимо отметить, что, к сожалению, разработка теории проверки статистических гипотез относительно параметров законов распределений, отличных от нормального, связана с большими трудностями.

4.4. Проверка гипотез о математическом ожидании

Ситуация, когда случайная величина X обладает свойствами нормальной случайной величины, имеет большое практическое значение. Итак, пусть случайная величина $X \to \mathcal{N}(a;\, \sigma)$. Требуется проверить нулевую гипотезу, что математическое ожидание случайной величины X равно некоторому гипотетическому значению a_0, т. е. гипотезу $\{H_0 : a = a_0\}$. Другими словами, надо установить, значимо или незначимо различаются среднее арифметическое $\overline{x}$ и гипотетическое математическое ожидание a генеральной совокупности.

В зависимости от имеющейся информации о параметрах генеральной совокупности можно сформулировать две основные модели и построить для них соответствующие критерии значимости.

Модель 1. Пусть генеральная совокупность имеет нормальное распределение: $X \to \mathcal{N}(a;\, \sigma)$. Предположим, что σ известно. На основании случайной выборки $(x_1, x_2, \ldots, x_n)$ из этой генеральной совокупности требуется проверить гипотезу $\{H_0 :\ a = a_0\}$ против альтернативной гипотезы $\{H_a :\ a \neq a_0\}$.

Критерий значимости для проверки указанной гипотезы основывается на вычислении выборочной статистики

$$u \;=\; \frac{\overline{x} - a}{\sigma}\sqrt{n}. \tag{4.10}$$

Выше было показано, что если $X \to \mathcal{N}(a;\, \sigma)$, то $\overline{x} \to \mathcal{N}(a;\, \sigma/\sqrt{n})$. Следовательно, если нулевая гипотеза $\{H_0 :\ a = a_0\}$ справедлива, то нормированная случайная величина $u \to \mathcal{N}(0; 1)$.

Зададим уровень значимости данного критерия равным α. По таблицам стандартизованного нормального распределения по заданному уровню значимости α находят две критические точки (квантили) $u_{1-\alpha/2} = -u_{\alpha/2}$ и $u_{\alpha/2}$. Множество значений u, определяемое неравенством $|u| \geq u_{\alpha/2}$, является критической областью критерия $u = (\overline{x} - a)\sqrt{n}/\sigma$ (см. рис. 4.5).

Следовательно, если вычисленное по результатам выборки объема n наблюдаемое значение критерия таково, что

$$|u_{\text{набл}}| \;\geq\; u_{\alpha/2}, \tag{4.11a}$$

то гипотеза H_0 отклоняется.

Если

$$|u_{\text{набл}}| \;<\; u_{\alpha/2}, \tag{4.11b}$$

то нет оснований для отклонения нулевой гипотезы.

Замечание. Если альтернативная гипотеза имеет вид $\{H_a : a < a_0\}$, то используют критерий u с левосторонней критической областью (рис. 4.6). В этом одностороннем случае критические точки (квантили) u_α стандартизованного нормального распределения находятся из условия $\Pr(u \leq -u_\alpha) = \alpha$. При таком выборе альтернативной гипотезы нулевая гипотеза отклоняется только тогда, когда наблюдаемое значение выборочной статистики $u_{\text{набл}} \leq -u_\alpha$.

Если альтернативная гипотеза имеет вид $\{H_a : a > a_0\}$, то применяют критерий u с правосторонней критической областью В этом случае критические точки (квантили) u_α стандартизованного нормального распределения находятся из условия $\Pr(u \geq u_\alpha) = \alpha$ (рис. 4.7). При таком виде альтернативной гипотезы нулевая гипотеза отклоняется только тогда, когда наблюдаемое значение выборочной статистики $u_{\text{набл}} \geq u_\alpha$.

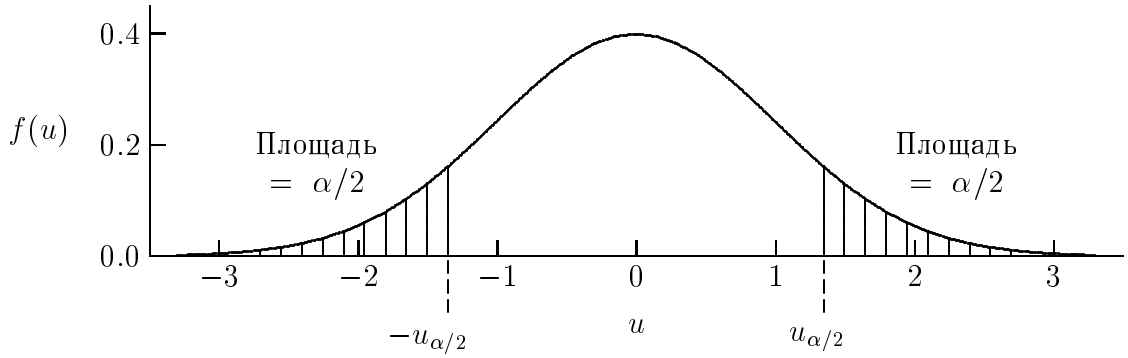


Рисунок 4.5 — Двусторонняя критическая область критерия u

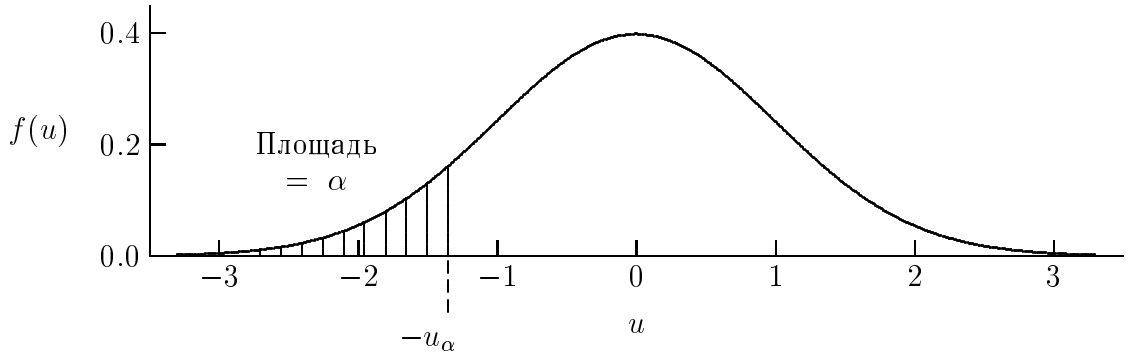


Рисунок 4.6 — Левосторонняя критическая область критерия u

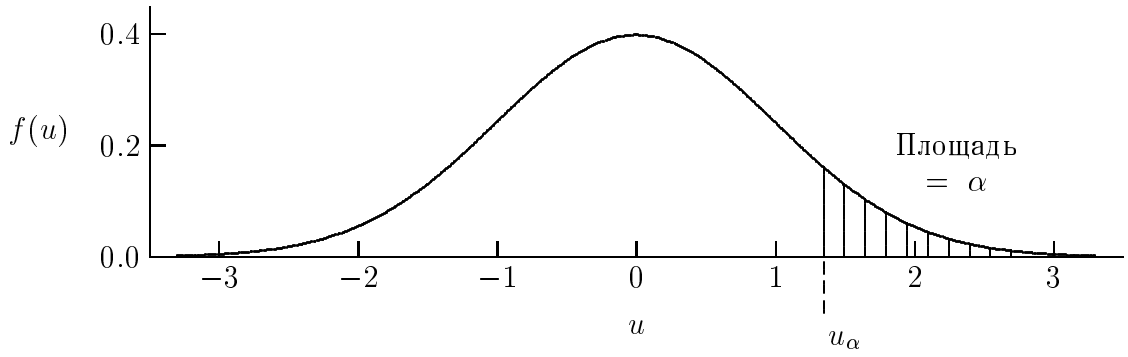


Рисунок 4.7 — Правосторонняя критическая область критерия u

Модель 2. Пусть генеральная совокупность имеет нормальное распределение: $X \to \mathcal{N}(a, \sigma)$. Параметры a и σ неизвестны. По результатам случайной выборки объема n найдены точечные оценки параметров

$$\hat{a} = \overline{x}, \qquad \hat{\sigma} = s_{\text{несм}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{4.12}$$

Требуется проверить нулевую гипотезу $\{H_0 : a = a_0\}$ против альтернативы $\{H_a : a \neq a_0\}$. Тогда критерий значимости для проверки нулевой гипотезы $\{H_0 : a = a_0\}$ основывается на вычислении тестовой статистики

$$t = \frac{\overline{x} - a_0}{s}\sqrt{n} = \frac{\overline{x} - a_0}{s_{\text{несм}}}\sqrt{n-1}. \tag{4.13}$$

Из курса теории вероятностей следует, что если гипотеза H_0 верна, то случайная величина t имеет распределение Стьюдента с $\nu = n - 1$ степенями свободы. По таблицам распределения Стьюдента по заданному уровню значимости α и числу степеней свободы $\nu = n - 1$ находят критические точки (квантили) $t_{\alpha/2;\,n-1}$ распределения Стьюдента.

Далее вычисляется значение критерия $t_{\text{набл}} = (\overline{x} - a_0)\sqrt{n}/s$, полученное на основании наблюдений.

Если

$$t_{\text{набл}} > t_{\alpha/2;\,n-1}, \tag{4.14a}$$

то нулевая гипотеза отклоняется в пользу альтернативной гипотезы.

Если

$$t_{\text{набл}} < t_{\alpha/2;\,n-1}, \tag{4.14b}$$

то считается, что нет оснований для отклонения нулевой гипотезы.

Замечание 1. Выше приведенное в модели 1 замечание остается справедливым и для выборочной статистики t. Это значит, что если рассматриваемая альтернативная гипотеза имеет вид $\{H_a : a > a_0\}$, то применяется правосторонний t-критерий, т.е. критические точки (квантили) $t_{\alpha/2;\,n-1}$ находятся из условия $\Pr(t > t_{\alpha/2;\,n-1}) = \alpha$. Если же рассматриваемая альтернативная гипотеза имеет вид $\{H_a : a < a_0\}$, то применяется левосторонний t-критерий, критические точки которого (квантили) $-t_{\alpha;\,n-1}$ находятся из условия $\Pr(t \leq -t_{\alpha;\,n-1}) = \alpha$.

Замечание 2. Если объем выборки n достаточно велик ($n > 50$), то для проверки гипотезы $\{H_0 : a = a_0\}$ можно применять критерий u (модель 1), в котором следует положить

$$\sigma = s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{4.15}$$

В курсах математической статистики показано, что критерии, основанные на вычислении тестовых статистик по формулам (4.10) и (4.13), являются наилучшими, так как обеспечивают максимальную мощность. Напомним, что мощность критерия

зависит от вида альтернативной гипотезы, принятого уровня значимости α и объема выборки n.

Например, если мы хотим проверить нулевую гипотезу $\{H_0 : \ a = a_0\}$ при простой альтернативной гипотезе $\{H_a : \ a = a_1\}$, причем $a_1 \neq a_0$, то мощность критерия M выражается формулой

$$M \ = \ 1 - \beta \ = \ 1 + \frac{1}{2}\,\Phi\left(-u_\alpha + \frac{a_1 - a_0}{\sigma/\sqrt{n}}\right) - \frac{1}{2}\,\Phi\left(-u_\alpha - \frac{a_1 - a_0}{\sigma/\sqrt{n}}\right). \qquad (4.16)$$

Зависимость мощности критерия от вида альтернативной гипотезы и уровня значимости используется при планировании эксперимента для расчета объема выборки, необходимой для получения заданной мощности критерия.

4.5. Проверка гипотез равенства математических ожиданий двух нормальных случайных величин

В практике при обработке статистических данных нередко возникает потребность в решении задач "сравнения". Например, часто приходится сравнивать новый и старый технологические методы изготовления некоторых изделий, успеваемость в двух группах, применяющих различные методы обучения, производительность труда на двух заводах, результаты двух серий экспериментов и т.д. Задачи такого типа можно решать, произведя построение теоретико-вероятностной модели (см. постановку задачи в начале данного раздела) генеральной совокупности $F(x;\theta)$.

В большинстве случаев законы распределения этих совокупностей предполагают нормальными, а изменения в "технологиях" сказываются на изменении математических ожиданий моделируемой нормальной совокупности. Таким образом, большинство задач сравнения сводится к проверке гипотез относительно математических ожиданий двух случайных величин, распределенных по нормальному закону.

В зависимости от имеющейся в распоряжении экспериментатора информации относительно параметров исследуемых нормальных совокупностей можно сформулировать две основные модели, в каждой из которых применяется определенный критерий значимости.

Модель 1. Пусть исследуются две случайные величины X и Y, каждая из которых подчиняется нормальному закону: $\mathrm{X} \to \mathcal{N}(a_1;\, \sigma_1)$ и $\mathrm{Y} \to \mathcal{N}(a_2;\, \sigma_2)$. Предположим, что средние квадратические отклонения σ_1 и σ_2 известны, а значения a_1 и a_2 неизвестны.

Требуется на основании двух независимых выборок объемом n_1 и n_2, соответственно извлеченных из исследуемых генеральных совокупностей, проверить нулевую гипотезу $\{H_0 : \ a_1 = a_2\}$ против альтернативной гипотезы $\{H_a : \ a_1 \neq a_2\}$.

Вычислим по выборкам средние арифметические $\overline{x}$ и $\overline{y}$. Известно, что если $\mathrm{X} \to \mathcal{N}(a_1;\, \sigma_1)$ и $\mathrm{Y} \to \mathcal{N}(a_2;\, \sigma_2)$, то $\overline{x} \to \mathcal{N}(a_1;\, \sigma_1/\sqrt{n_1})$ и $\overline{y} \to \mathcal{N}(a_2;\, \sigma_2/\sqrt{n_2})$. Так как выборки независимы, то независимы и средние арифметические $\overline{x}$ и $\overline{y}$, а

следовательно,

$$\mathrm{D}[\overline{x}-\overline{y}] \;=\; \mathrm{D}[\overline{x}]+\mathrm{D}[\overline{y}] \;=\; \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}. \tag{4.17}$$

Если проверяемая гипотеза $\{H_0 : a_1 = a_2\}$ справедлива, то

$$\mathrm{M}[\overline{x}-\overline{y}] \;=\; \mathrm{M}[\overline{x}]-\mathrm{M}[\overline{y}] \;=\; 0. \tag{4.18}$$

Следовательно, нормированная разность (выборочная статистика)

$$u \;=\; \frac{\overline{x}-\overline{y}}{\sigma_1^2/n_1+\sigma_2^2/n_2} \tag{4.19}$$

имеет стандартизованное нормальное распределение $u \to \mathcal{N}(0;\,1)$. Эта выборочная статистика и применяется в качестве статистического критерия значимости для проверки нулевой гипотезы $\{H_0 :\; a_1 = a_2\}$.

Для проверки этой нулевой гипотезы необходимо задать уровень значимости α. Затем по таблицам стандартизованного нормального распределения по заданному уровню значимости α найти критическое значение (квантиль) $u_{\alpha/2}$, удовлетворяющее условию $\Pr(|u| \geq u_{\alpha/2})$, которое определяет двустороннюю критическую область u-критерия (см. приложение).

Таким образом, если вычислить согласно (4.19) наблюдаемое (частное) значение критерия $u_{\text{набл}}$ и при этом окажется, что

$$|u_{\text{набл}}| \;\geq\; u_{\alpha/2}, \tag{4.20a}$$

то нулевая гипотеза отклоняется в пользу альтернативной.

Если же

$$|u_{\text{набл}}| \;<\; u_{\alpha/2}, \tag{4.20b}$$

то считается, что нет оснований для отклонения нулевой гипотезы.

Замечание. В том случае, когда рассматриваемая альтернативная гипотеза имеет вид $\{H_a : a_1 < a_2\}$, то применяется левосторонний u–критерий. При его применении по таблицам стандартизованного нормального распределения находится такое критическое значение $-u_\alpha$, чтобы $\Pr(u < -u_\alpha) = \alpha$. В случае же, когда альтернативная гипотеза имеет вид $\{H_a : a_1 > a_2\}$, то применяется правосторонний u–критерий. В этом случае по таблицам стандартизованного нормального распределения находится такое критическое значение u_α, чтобы $\Pr(u \geq u_\alpha) = \alpha$.

Модель 2. Пусть исследуются две случайные величины X и Y, каждая из которых подчиняется нормальному закону $\mathrm{X} \to \mathcal{N}(a_1;\,\sigma_1)$ и $\mathrm{Y} \to \mathcal{N}(a_2;\,\sigma_2)$, причем средние квадратические отклонения σ_1 и σ_2 хотя и неизвестны, но предполагается, что $\sigma_1 = \sigma_2$, а параметры a_1 и a_2 неизвестны.

Требуется на основании двух независимых выборок объемом n_1 и n_2 ($n_1 \geq 30$ и $n_2 \geq 30$), извлеченных из исследуемых генеральных нормальных совокупностей, проверить нулевую гипотезу $\{H_0 : a_1 = a_2\}$ против альтернативной гипотезы $\{H_a : a_1 \neq a_2\}$.

Критерий значимости для проверки данной нулевой гипотезы основывается на вычислении выборочной статистики

$$t \;=\; (\overline{x}-\overline{y})\left[\frac{n_1\hat{\sigma}_1^2+n_2\hat{\sigma}_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right]^{-1/2}, \tag{4.21}$$

где

$$\overline{x} \;=\; \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, \qquad \overline{y} \;=\; \frac{1}{n_2}\sum_{i=1}^{n_2} y_i \tag{4.22a}$$

— средние арифметические;

$$\hat{\sigma}_1^2 \;=\; \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i-\overline{x})^2, \qquad \hat{\sigma}_2^2 \;=\; \frac{1}{n_2}\sum_{i=1}^{n_2}(y_i-\overline{y})^2 \tag{4.22b}$$

– оценки средних квадратических отклонений первой и второй совокупностей соответственно.

Можно доказать, что если нулевая гипотеза H_0 справедлива, то выборочная статистика t имеет распределение Стьюдента с $\nu = n_1 + n_2 - 2$ степенями свободы. Для проверки нулевой гипотезы необходимо вычислить согласно (4.21) наблюдаемое значение t-критерия $t_{\text{набл}}$ и задать уровень значимости α. По таблицам квантилей распределения Стьюдента по заданной вероятности α и числу степеней свободы ν найти критические точки (квантили) $t_{\alpha/2;\,n_1+n_2-2}$.

Если при этом окажется

$$|t_{\text{набл}}| \;\geq\; t_{\alpha/2;\,n_1+n_2-2}, \tag{4.23a}$$

то нулевая гипотеза отклоняется в пользу альтернативной.

Если же

$$|t_{\text{набл}}| \;<\; t_{\alpha/2;\,n_1+n_2-2}, \tag{4.23b}$$

то считается, что нет оснований для отклонения нулевой гипотезы.

Критерии, изложенные в данном параграфе, являются наиболее мощными.

Замечание 1. Если альтернативная гипотеза имеет вид $\{H_a:\ a_1 < a_2\}$, то, как и в модели 1, применяется левосторонний t-критерий. Если же альтернативная гипотеза имеет вид $\{H_a:\ a_1 > a_2\}$, то применяется правосторонний t-критерий. Условия нахождения критических точек $t_{\alpha;\,n_1+n_2-2}$ этих критериев остаются теми же, что и в модели 1.

Замечание 2. Иногда в практике случается, что результаты двух выборок рассматриваются как измерения одной и той же случайной величины до и после проведения некоторой технологической операции. Пусть имеются упорядоченные пары таких чисел $(x_i,\, y_i)$, i – номер измерения.

Будем рассматривать разности $z_i = x_i - y_i$ как компоненты одной выборки и вычислим их среднее арифметическое $\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ и среднее квадратическое $s_z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i-\overline{z})^2$ отклонения. Будем вместо гипотезы $\{H_0: a_1 = a_2\}$ проверять эквивалентную ей гипотезу $\{H_0: a_1 - a_2 = 0\}$. Проверка этой гипотезы производится с помощью критерия $t = (\overline{z}/s_z)\sqrt{n-1}$ методами, описанными в предыдущем параграфе.

Замечание 3. Если объемы выборок n_1 и n_2 достаточно велики ($n_1 \geq 30$ и $n_2 \geq 30$), то вместо t-критерия (модель 2) можно применять u-критерии (модель 1). В этом случае при вычислении выборочной статистики u по формуле (4.13) в ней следует заменить средние квадратические отклонения σ_1 и σ_2 их точечными оценками s_1 и s_2.

4.6. Проверка гипотез о дисперсии нормальной случайной величины

В практических задачах проверка гипотез о дисперсиях играет большую роль, так как именно дисперсия характеризует такие важные технологические и конструкторские показатели, как точность работы машин, погрешности показаний измерительных приборов, ритмичность производства, устойчивость работы автоматических линий и т. д.

Критерий значимости, применяемый для проверки равенства неизвестной дисперсии генеральной нормальной совокупности некоторому гипотетическому (предполагаемому) значению σ_0^2, основывается на ряде исходных вероятностных предположений относительно исследуемой генеральной совокупности. Будем рассматривать эти предположения как некоторую вероятностную модель.

Модель. Пусть случайная величина X $\to \mathcal{N}(a;\,\sigma)$, причем параметры a и σ – неизвестны. Из генеральной совокупности $\mathcal{N}(a;\,\sigma)$ извлечена случайная выборка объема n и найдены точечные оценки параметров нормального закона:

$$\hat{a} \;=\; \overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad \hat{\sigma}^2 \;=\; s^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \,. \tag{4.24}$$

Требуется на основании имеющейся информации проверить нулевую гипотезу $\{H_0 : \sigma^2 = \sigma_0^2\}$ против альтернативной гипотезы $\{H_a : \sigma^2 \neq \sigma_0^2\}$.

Критерий значимости для проверки данной нулевой гипотезы основывается на вычислении выборочной статистики

$$\chi^2 \;=\; \frac{n\,s^2}{\sigma_0^2} \;=\; \frac{1}{\sigma_0^2}\sum_{i=1}^{n} (x_i - \overline{x})^2 \,. \tag{4.25}$$

Если предположить, что нулевая гипотеза справедлива, то можно доказать, что выборочная статистика имеет χ^2 (хи-квадрат) распределение с $\nu = n - 1$ степенями свободы. Зададим уровень значимости данного критерия равным α. Тогда по таблицам χ^2-распределения по уровню значимости α и числу степеней свободы $\nu = n - 1$ можно найти критические точки (квантили) χ^2-распределения $\chi^2_{1-\alpha/2;\,\nu}$ и $\chi^2_{\alpha/2;\,\nu}$ (см. рис. 4.8). Теперь необходимо вычислить наблюдаемое (частное) значение критерия

$$\chi^2_{\text{набл}} \;=\; \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \,. \tag{4.26}$$

Если при этом окажется, что

$$\chi^2_{\text{набл}} \;\geq\; \chi^2_{\alpha/2;\,\nu} \qquad \text{или} \qquad \chi^2_{\text{набл}} \;\leq\; \chi^2_{1-\alpha/2;\,\nu}, \tag{4.27a}$$

то нулевая гипотеза отклоняется в пользу альтернативной.

Если же окажется, что

$$\chi^2_{1-\alpha/2;\,\nu} \;<\; \chi^2_{\text{набл}} \;<\; \chi^2_{\alpha/2;\,\nu}, \tag{4.27b}$$

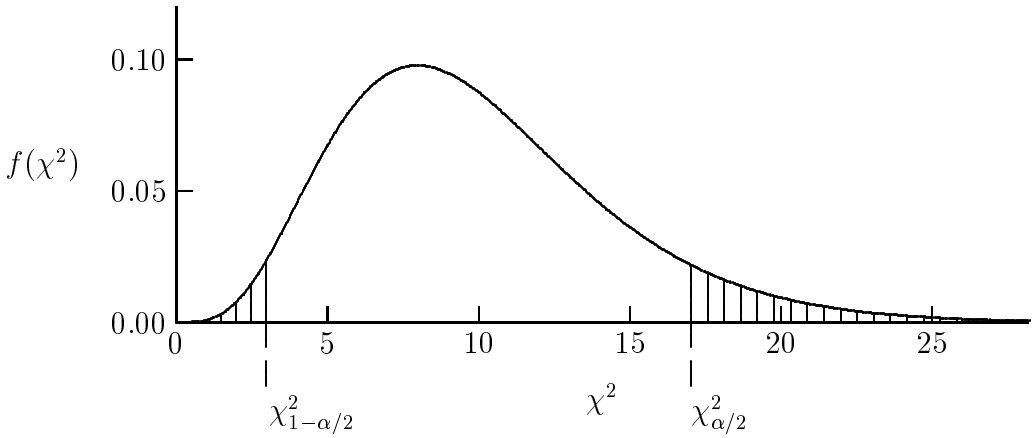


Рисунок 4.8 — Двусторонняя критическая область критерия χ^2

то считается, что нет оснований для отклонения нулевой гипотезы.

Замечание 1. Если альтернативная гипотеза имеет вид $\{H_a : \sigma^2 < \sigma_0^2\}$, то применяется критерий χ^2 с левосторонней критической областью. В этом случае критические точки (квантили) $\chi^2_{1-\alpha/2;\,\nu}$ находятся из условия $\Pr\left(\chi^2 \le \chi^2_{1-\alpha;\,\nu}\right)$, а нулевая гипотеза отклоняется, если $\chi^2_{\text{набл}} \le \chi^2_{1-\alpha;\,\nu}$.

В случае, когда альтернативная гипотеза имеет вид $\{H_a : \sigma^2 > \sigma_0^2\}$, то применяется критерий χ^2, отвечающий правосторонней критической области. В этом случае критические точки $\chi^2_{\alpha;\,\nu}$ находятся из условия $\Pr\left(\chi^2 \ge \chi^2_{\alpha;\,\nu}\right)$, а нулевая гипотеза отклоняется, если $\chi^2_{\text{набл}} \ge \chi^2_{\alpha;\,\nu}$.

Замечание 2. В случае, когда число степеней свободы велико, $\nu = n - 1 > 30$, проверку нулевой гипотезы можно производить, вычисляя следующую выборочную статистику $u = \sqrt{2\chi^2_{\text{набл}}} - \sqrt{2\nu}$. Можно показать, что эта статистика имеет асимптотически нормальное стандартизованное распределение $u \to \mathcal{N}(0;\,1)$.

Критерий, основанный на вычислении тестовой статистики, определяемой формулой (4.25), является наиболее мощным. Исследование мощности данного критерия можно найти в учебниках по математической статистике.

4.7. Проверка гипотез о дисперсиях двух нормальных случайных величин

В случае, когда статистические исследования некоторого количественного признака производятся в двух генеральных совокупностях, часто появляется потребность в проверке гипотезы о равенстве степени рассеивания исследуемого признака в этих совокупностях. Пусть имеются две выборки, дисперсии которых соответственно равны s_1^2 и s_2^2. Можно ли считать при наличии некоторых различий между величинами s_1^2 и s_2^2, что данные выборки принадлежат одной и той же генеральной совокупности?

Можно сформулировать довольно распространенную типовую задачу: произведено две серии опытов, из которых один опыт производится с учетом фактора A,

а другой — без учета. Оказывает ли фактор А влияние на рассеивание исследуемого признака?

Для ответа на поставленные вопросы необходимо произвести проверку нулевой гипотезы $\{H_0 : \sigma_1^2 = \sigma_2^2\}$. Критерий значимости, применяемый для проверки равенства неизвестных дисперсий нормальных совокупностей, т.е. для проверки указанной нулевой гипотезы, основывается на ряде исходных вероятностных предположений относительно этих совокупностей. Будем рассматривать эти предположения как некоторую вероятностную модель.

Модель. Пусть СВ $\mathrm{X}_1 \to \mathcal{N}(a_1; \sigma_1)$ и $\mathrm{X}_2 \to \mathcal{N}(a_2; \sigma_2)$. Параметры нормальных законов распределения a_1 и a_2 – неизвестны. Из этих двух генеральных нормальных совокупностей извлечены выборки объемом n_1 и n_2. На основе этих выборок вычислены точечные оценки параметров нормального закона:

$$\overline{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{i1}, \qquad s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{i1} - \overline{x}_1)^2$$

и

$$\overline{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{i2}, \qquad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{i2} - \overline{x}_2)^2 .$$

Требуется на основании имеющейся информации проверить нулевую гипотезу $\{H_0 : \sigma_1^2 = \sigma_2^2\}$ против альтернативной гипотезы $\{H_a : \sigma_1^2 > \sigma_2^2\}$.

Критерий значимости для проверки данной нулевой гипотезы основывается на вычислении выборочной статистики

$$F = s_1^2/s_2^2, \tag{4.28}$$

где s_1^2 и s_2^2 – соответственно наибольшая и наименьшая дисперсии. Если предположить, что нулевая гипотеза верна, то выборочная статистика F имеет распределение Фишера с $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$ степенями свободы. Зададим уровень значимости данного критерия равным α. Тогда по таблицам квантилей F-распределения по уровню значимости α и числу степеней свободы $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$ можно найти критическую точку (квантиль) $F_{\alpha;\nu_1;\nu_2}$, удовлетворяющую условию (рис. 4.9)

$$\Pr\left(F > F_{\alpha;\nu_1;\nu_2}\right) = \alpha. \tag{4.29}$$

Вычислим теперь согласно (4.28) наблюдаемое значение критерия $F_{\text{набл}}$.

Если при этом окажется, что

$$F_{\text{набл}} \geq F_{\alpha;\nu_1;\nu_2}, \tag{4.30a}$$

то нулевая гипотеза отклоняется в пользу альтернативной.

Если же

$$F_{\text{набл}} < F_{\alpha;\nu_1;\nu_2}, \tag{4.30b}$$

то считается, что нет оснований для отклонения нулевой гипотезы.

Замечание 1. Если альтернативная гипотеза имеет вид $\{H_a : \sigma_1^2 \neq \sigma_2^2\}$, то применяется F-критерий с двусторонней критической областью. В этом случае критические точки $F_{\alpha/2;\nu_1;\nu_2}$ находятся по уровню значимости $\alpha/2$.

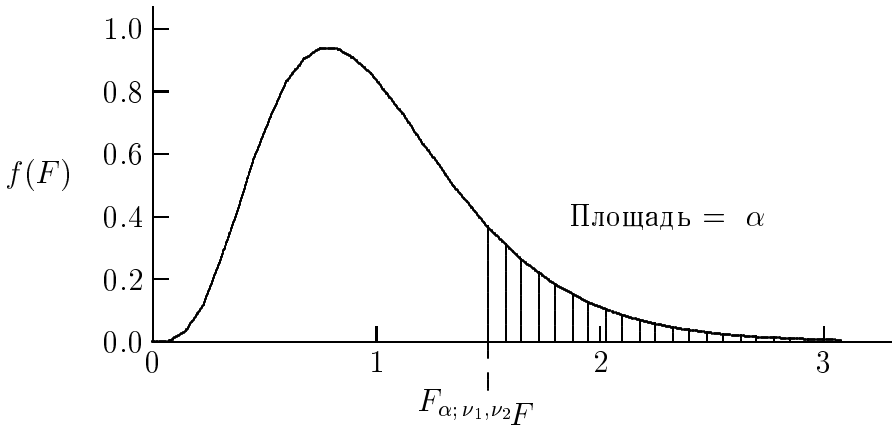


Рисунок 4.9 — К формированию критической области критерия F при заданном уровне значимости α

Замечание 2. Очень часто F-критерий применяется для проверки предположения о равенстве дисперсий, заложенного в критерии Стьюдента.

Критерий, основанный на вычислении тестовой статистики, определяемой формулой (4.28), является наиболее мощным.

4.8. Проверка гипотез о дисперсиях нескольких нормальных величин

Проверку гипотез о дисперсиях нескольких нормальных совокупностей по выборкам одинакового объема можно проводить методом, изложенным в предыдущем параграфе, т. е. сравнивая по критерию Фишера наибольшую и наименьшую из k рассматриваемых эмпирических дисперсий. Если при этом окажется, что различие между ними незначимо, то тем более незначимо и различие между остальными дисперсиями.

В случае, если объемы выборок, извлеченных из исследуемых нормальных совокупностей, различны, то можно применять специальный критерий значимости – критерий Бартлетта. Применение критерия Бартлетта основывается на ряде исходных предположений относительно исследуемых генеральных совокупностей. Будем рассматривать эти исходные предположения как некоторую вероятностную модель.

Модель. Пусть имеется k нормальных совокупностей $\mathrm{X}_i \to \mathcal{N}(a_i;\, \sigma_i)$. Из этих совокупностей извлечены независимые выборки объема n_i каждая. Результаты каждой выборки обозначим x_{ij}, где $i = 1,\ 2,\ \ldots,\ k$, $j = 1,\ 2,\ \ldots,\ n_i$. По результатам этих выборок вычислены точечные оценки параметров нормальных законов распределения:

а) средние арифметические

$$\overline{x}_i \;=\; \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \qquad i = 1, 2, \ldots, k; \tag{4.31}$$

б) несмещенные оценки дисперсий

$$s_i^2 \;=\; \frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x}_i\right)^2; \tag{4.32}$$

в) взвешенную по числам степеней свободы среднюю арифметическую несмещенных эмпирических дисперсий

$$\overline{s^2} \;=\; \frac{1}{n-k}\sum_{i=1}^{k}\left(x_{ij}-\overline{x}_i\right)^2. \tag{4.33}$$

Требуется на основании имеющейся информации проверить нулевую гипотезу $\{H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2\}$ против альтернативной гипотезы $\{H_a :$ *не все эти дисперсии равны между собой*$\}$.

Проверка нулевой гипотезы "однородности" дисперсий основывается на вычислении выборочной статистики

$$\chi^2 \;=\; \frac{2,303}{C}\left[(n-k)\,\lg(\overline{s^2}) - \sum_{i=1}^{k}(n_i-1)\lg(s_i^2)\right], \tag{4.34}$$

где

$$C \;=\; 1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1}-\frac{1}{n-k}\right).$$

Выборочная статистика χ^2 при условии справедливости нулевой гипотезы имеет приближенное χ^2-распределение с $\nu = k-1$ степенями свободы. Более точно, распределение выборочной статистики (4.34) асимптотически сходится к χ^2-распределению, причем сходимость эта является очень быстрой, так что критерий Бартлетта можно применять даже при очень малых объемах выборок.

Для проверки нулевой гипотезы по формуле (4.34) вычисляется наблюдаемое значение $\chi^2_{\text{набл}}$ выборочной статистики χ^2.

Затем по таблицам квантилей χ^2-распределения по заданному уровню значимости α и числу степеней свободы $\nu = k-1$ находят критическую точку $\chi^2_{\alpha;\,\nu}$, удовлетворяющую условию

$$\Pr\left(\chi^2 \geq \chi^2_{\alpha;\,\nu}\right) \;=\; \alpha. \tag{4.35}$$

Если окажется, что

$$\chi^2_{\text{набл}} \;\geq\; \chi^2_{\alpha;\,\nu}, \tag{4.36a}$$

то нулевая гипотеза отклоняется в пользу альтернативной гипотезы.

Если

$$\chi^2_{\text{набл}} \;<\; \chi^2_{\alpha;\,\nu}, \tag{4.36b}$$

то считается, что нет основания для отклонения нулевой гипотезы.

Если проверяется гипотеза о дисперсиях двух случайных величин, то критерий Бартлетта имеет меньшую мощность, чем критерий Фишера, основанный на вычислении тестовой статистики (4.28).

4.9. Проверка гипотез о параметре биномиального закона распределения

Очень часто при обработке статистической информации можно встретиться с признаками, не поддающимися количественной оценке. Например, невозможно дать количественную оценку математическим способностям студентов и т. д. В этих случаях принято подсчитывать долю или процент элементов генеральной совокупности, обладающих тем или иным качественным признаком. Например, можно подсчитать:

а) долю или процент студентов данного университета, занимающихся в библиотеке более 6 часов в неделю;

б) долю или процент этих же студентов, знающих два иностранных языка;

в) долю или процент бракованной продукции в некоторой партии;

г) долю или процент мужчин ростом от 168 *см* до 172 *см* и т. д.

Долю элементов генеральной совокупности, обладающих тем или иным качественным признаком, будем обозначать p $(0 \le p \le 1)$. Выборочной оценкой доли является частость (относительная частота) m/n.

Проверка гипотез о доле основывается на модели биномиального распределения, поскольку доля представляет параметр p в этом распределении. Существует много методов проверки гипотез о доле.

Ниже мы рассмотрим только один тип критериев, который можно применять при достаточно большом объеме выборки $(n \ge 100)$. Эти критерии основаны на том, что выборочная оценка доли генеральной совокупности p имеет асимптотически нормальный закон распределения с параметрами

$$\mathrm{M}[m/n] \;=\; p, \qquad \sigma[m/n] \;=\; \sqrt{p(1-p)/n}. \tag{4.37}$$

Ниже приводятся две модели, первая из которых отражает вероятностные предпосылки, необходимые для проверки нулевой гипотезы равенства доли генеральной совокупности p некоторому гипотетическому числу p_0, т. е. гипотезы $\{H_0 :\; p = p_0\}$.

Модель 2 отражает вероятностные предпосылки, необходимые для проверки нулевой гипотезы равенства долей двух генеральных совокупностей.

Модель 1. Пусть число элементов x генеральной совокупности, обладающих некоторым качественным признаком, распределено по биномиальному закону с параметром p, т. е.

$$\mathrm{P}_x \;=\; C_n^x p^x (1-p)^{n-x}, \qquad (x = 0, 1, 2, \ldots, n), \tag{4.38}$$

где p – доля элементов генеральной совокупности, обладающих некоторым качественным признаком. Из этой генеральной совокупности извлечена независимая выборка объема n $(n \ge 100)$ и по ней вычислена точечная оценка параметра p: $\hat{p} = m/n$.

Требуется на основании имеющейся информации проверить нулевую гипотезу $\{H_0 : p = p_0\}$, где p_0 – некоторое гипотетическое число, против альтернативной гипотезы $\{H_a : p \ne p_0\}$.

В курсе теории вероятностей доказывается, что частость m/n имеет асимптотически нормальное распределение с математическим ожиданием $\mathrm{M}[m/n] = p$ и средним квадратическим отклонением $\sigma[m/n] = \sqrt{p(1-p)/n}$.

Следовательно, если нулевая гипотеза верна, то нормированная выборочная статистика

$$u \;=\; \frac{m/n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \tag{4.39}$$

имеет стандартизованное нормальное распределение: $u \to N(0;\,1)$. Эта выборочная статистика применяется для проверки нулевой гипотезы $\{H_0 : p = p_0\}$.

Для проверки нулевой гипотезы по таблицам квантилей стандартизованного нормального распределения по заданному уровню значимости α находят критическое значение $u_{\alpha/2}$, удовлетворяющее условию

$$\Pr\left(|u| \geq u_{\alpha/2}\right) \;=\; \alpha. \tag{4.40}$$

Затем вычисляется наблюдаемое значение критерия $u_{\text{набл}}$.

Если окажется, что

$$|u_{\text{набл}}| \;\geq\; u_{\alpha/2}, \tag{4.41a}$$

то нулевая гипотеза отвергается в пользу альтернативной.

Если же

$$|u_{\text{набл}}| \;<\; u_{\alpha/2}, \tag{4.41b}$$

то считается, что нет оснований для отклонения нулевой гипотезы.

Замечание. Если альтернативная гипотеза имеет вид $\{H_a :\; p < p_0\}$, то применяется u-критерий с левосторонней критической областью. Если же альтернативная гипотеза имеет вид $\{H_a :\; p > p_0\}$, то применяется u-критерий с правосторонней критической областью. Критические значения $-u_\alpha$ для критерия с левосторонней критической областью находятся из условия $\Pr(u \leq -u_\alpha) = \alpha$, для критерия с правосторонней критической областью – из условия $\Pr(u \geq u_\alpha) = \alpha$.

Модель 2. Пусть даны две генеральные совокупности, имеющие биномиальный закон распределения с параметрами p_1 и p_2 (здесь p_1, p_2 – неизвестные доли элементов генеральных совокупностей, обладающие заданным качественным признаком). Из этих генеральных совокупностей извлечены выборки объема n_1 и n_2 ($n_1 \geq 100$ и $n_2 \geq 100$), после чего вычислены точечные оценки параметров p_1 и p_2:

$$\hat{p}_1 \;=\; m_1/n_1, \qquad \hat{p}_2 \;=\; m_2/n_2.$$

Требуется на основе имеющейся информации проверить нулевую гипотезу $\{H_0 : p_1 = p_2\}$ против альтернативной гипотезы $\{H_a : p_1 \neq p_2\}$.

Проверка нулевой гипотезы основывается на вычислении выборочной статистики

$$u \;=\; \frac{m_1/n_1 - m_2/n_2}{\sqrt{\overline{p}\,\overline{q}}}\sqrt{n}, \tag{4.42}$$

где

$$n \;=\; \frac{n_1 \cdot n_2}{n_1 + n_2}, \qquad \overline{p} \;=\; \frac{m_1 + m_2}{n_1 + n_2}, \qquad \overline{q} \;=\; 1 - \overline{p}. \tag{4.43}$$

Если нулевая гипотеза верна, то эта выборочная статистика имеет асимптотически нормальное стандартизованное распределение $\mathcal{N}(0;\,1)$.

Правило проверки гипотезы остается тем же, что и в модели 1.

Если

$$|u_{\text{набл}}| \;\geq\; u_{\alpha/2}, \tag{4.44a}$$

то нулевая гипотеза отклоняется в пользу альтернативной.

Если

$$|u_{\text{набл}}| \;<\; u_{\alpha/2}, \tag{4.44b}$$

то считается, что нет оснований для отклонения нулевой гипотезы.

Замечание. Если альтернативная гипотеза имеет вид $\{H_a : p_1 < p_2\}$, то применяется критерий с левосторонней критической областью. Если же альтернативная гипотеза имеет вид $\{H_a : p_1 > p_2\}$, то применяется критерий с правосторонней критической областью.

4.10. Проверка гипотез о математических ожиданиях нескольких нормальных величин методом однофакторного дисперсионного анализа

Критерий дисперсионного анализа является одним из основных понятий быстро развивающейся ветви математической статистики – теории планирования эксперимента. Метод дисперсионного анализа позволяет проверить, оказывают ли влияние на математические ожидания случайных величин определенные факторы, которые можно произвольно изменять в ходе эксперимента, выбрать наиболее важные факторы и оценить степень их влияния.

Если на математические ожидания оказывает влияние только один фактор, то соответствующий критерий значимости называется *однофакторным дисперсионным анализом*, если же несколько – *многофакторным дисперсионным анализом*. Ниже ограничимся рассмотрением только однофакторного дисперсионного анализа.

Идея однофакторного дисперсионного анализа заключается в разбиении общей дисперсии случайной величины X на два независимых слагаемых – *факторную дисперсию*, порождаемую воздействием исследуемого фактора, и *остаточную дисперсию*, обусловленную различными другими неучтенными и случайными факторами, т.е. $s^2_{\text{общ}} = s^2_{\text{факт}} + s^2_{\text{ост}}$.

Замечание. В процессе применения однофакторного дисперсионного анализа результаты измерений случайной величины X разбиваются в зависимости от степени действия (уровня) фактора A на группы. С этой точки зрения факторную дисперсию называют иногда междугрупповой, а остаточную – внутригрупповой (внутри групп фактор A не действует). В результате сравнения факторной и остаточной дисперсий по критерию Фишера $F = s^2_{\text{факт}}/s^2_{\text{ост}}$ приходят к выводу о значимости расхождения средних значений в группах.

Сформулируем основные предположения и ограничения, лежащие в обосновании дисперсионного анализа в виде вероятностной модели.

Модель. Предположим, что в эксперименте с целью изучения влияния фактора A на некоторый результативный признак были разбиты в зависимости от вариации признака A результаты измерений на k групп по n_i измерений в каждой группе. Будем рассматривать результаты измерений x_{ij} $(i = 1, 2, \ldots, k,\ j = 1, 2, \ldots, n_i,$ где i – номер уровня фактора A, j – номер результата измерения на данном уровне) как выборки из генеральных нормальных совокупностей: $\mathrm{X}_i \to N(a_i;\, \sigma_i)$ $(i = 1, 2, \ldots, k)$. Параметры (a_i, σ_i) хотя и неизвестны, но предполагается, что $\sigma_1 = \sigma_2 = \ldots = \sigma_k$.

Выполнение последнего равенства можно проверить с помощью критерия Бартлетта. Представим результаты измерений x_{ij} в виде суммы двух слагаемых:

$$x_{ij} \;=\; a_i + \varepsilon_{ij}, \tag{4.45}$$

где a_i – математическое ожидание случайной величины X_i; ε_{ij} – случайная ошибка (остаток), характеризующая влияние на результаты X_i неучтенных и случайных факторов. Относительно ошибки предполагается, что $\varepsilon_{ij} \to N(0;\, \sigma)$.

Предположим, что по выборочным данным вычислены:

а) групповые средние арифметические

$$\overline{x}_i \;=\; \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \qquad i = 1, 2, \ldots, k; \tag{4.46}$$

б) общая средняя арифметическая

$$\langle x \rangle \;=\; \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}, \qquad n = \sum_{i=1}^{k} n_i. \tag{4.47}$$

На основе имеющейся информации требуется проверить нулевую гипотезу $\{H_0 : a_1 = a_2 = \ldots = a_k\}$ против альтернативной гипотезы $\{H_a :$ *не все математические ожидания равны между собой*$\}$.

Проверка нулевой гипотезы основывается на вычислении выборочной статистики

$$F \;=\; \left[\frac{1}{k-1}\sum_{i=1}^{k}\left(\overline{x}_i - \langle x \rangle\right)^2 n_i\right]\left[\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}_i\right)^2 n_i\right]^{-1}. \tag{4.48}$$

Если нулевая гипотеза верна, то эта выборочная статистика имеет F–распределение с $\nu_1 = k - 1$ и $\nu_2 = n - k$ степенями свободы. Далее вычисляется наблюдаемое значение критерия. Для удобства все вычисления располагаются в таблице дисперсионного анализа.

Вычисленное наблюдаемое значение критерия F сравнивается с критическим значением $F_{\alpha;\nu_1;\nu_2}$, найденным по таблицам квантилей F-распределения по заданному уровню значимости α и числу степеней свободы ν_1 и ν_2.

Если

$$F_{\text{набл}} \;\geq\; F_{\alpha;\nu_1;\nu_2}, \tag{4.49a}$$

то нулевая гипотеза отклоняется в пользу альтернативной гипотезы.

Если

$$F_{\text{набл}} \;<\; F_{\alpha;\nu_1;\nu_2}, \tag{4.49b}$$

то считается, что нет оснований для отклонения нулевой гипотезы.

Отклонение нулевой гипотезы является статистическим доказательством влияния фактора A на математические ожидания исследуемых случайных величин X.

4.11. Примеры

Пример 4.1

Техническая норма предусматривает в среднем 40 c на выполнение определенной технической операции на часовом конвейере. От работниц, занятых на этой операции, поступили жалобы, что они в действительности затрачивают на эту операцию больше времени. Для проверки жалобы произведены хронометрические измерения времени выполнения этой технической операции у 16 работниц и получены следующие результаты:

$$\overline{x} \;=\; 42\ c \qquad \text{(среднее время выполнения операции)},$$

$$s_{\text{несм}} \;=\; \left[\frac{1}{n-1}\sum_{i=1}^{16}(x_i-\overline{x})^2\right]^{1/2} \;\approx\; 3,5\ c.$$

Можно ли по имеющимся хронометрическим данным на уровне значимости $\alpha = 0,01$ отклонить гипотезу, что действительное среднее время исполнения этой технической операции соответствует норме?

Решение

Из условия примера следует, что нам надо проверить нулевую гипотезу $\{H_0 : a = 40\ c\}$ (*техническая норма установлена верно*) против альтернативной гипотезы $\{H_a : a \neq 40\ c\}$ (*техническая норма установлена неверно*).

Для проверки данной нулевой гипотезы применим t–критерий значимости (модель 2) с правосторонней критической областью.

Вычислим наблюдаемое значение t-критерия

$$t_{\text{набл}} \;=\; \frac{\overline{x}-a_0}{s_{\text{несм}}}\sqrt{n-1} \;=\; \frac{42-40}{3,5}\sqrt{15} \;=\; 2,21.$$

Пользуясь таблицей квантилей распределения Стьюдента (см. приложение), по уровню значимости $\alpha = 0,01$ и числу степеней свободы $\nu = n-1 = 15$ находим значение квантиля $t_{0,01;\,15}$, удовлетворяющее условию $\Pr(t \geq t_{0,01;\,15}) = 0,01$. Это значение $t_{\alpha;\,n-1} = t_{0,01;\,15} = 2,602$.

Так как $t_{\text{набл}} = 2,21 < 2,602$, то нет оснований для отклонения нулевой гипотезы (пересмотра технической нормы времени исполнения данной операции).

Таким образом, мы доказали, что при $\alpha = 0,01$ разность между средним временем (по хронометражу), затрачиваемым на данную техническую операцию, и нормой времени существенно незначима (случайна).

Пример 4.2

Выдвинута гипотеза, что применение нового типа резца сокращает время обработки некоторой детали. Проведено 10 измерений времени, затрачиваемого на обработку этой детали старым и новым резцом. Получены следующие результаты (в минутах):

старый тип резца – 58, 58, 56, 38, 70, 38, 42, 75, 68, 67;

новый тип резца – 57, 55, 63, 24, 67, 43, 33, 68, 56, 54.

Проверить гипотезу равенства среднего времени, затрачиваемого на изготовление этой детали с помощью двух типов резцов. Уровень значимости принять равным $\alpha = 0,05$.

Решение

Предположим, что время, необходимое для обработки детали старым и новым типом резца, является случайной величиной, распределенной по нормальному закону $X \to \mathcal{N}(a_1; \sigma_1)$ и $Y \to \mathcal{N}(a_2; \sigma_2)$, причем a_1 и a_2 неизвестны, а σ_1 и σ_2 хотя и неизвестны, но предполагается, что $\sigma_1 = \sigma_2$.

Согласно условию, нам необходимо проверить нулевую гипотезу $\{H_0 : a_1 = a_2\}$ (*среднее время, затрачиваемое на изготовление детали старым и новым типом резца, одинаково*) против альтернативной гипотезы $\{H_a : a_1 > a_2\}$ (*новый тип резца сокращает среднее время обработки данной детали*).

Так как объемы выборок $n_1 = n_2 = 10$ малы и условия, заложенные в модели 2, выполняются, то для проверки нулевой гипотезы применим правосторонний t-критерий. Вычислим наблюдаемое значение статистики. Для этого проведем (см. ниже таблицу) вспомогательные вычисления данных для старого типа резца (переменные $\{x_i\}$) и нового типа резца (переменные $\{y_i\}$).

i	x_i	$(x_i - \overline{x})^2$	y_i	$(y_i - \overline{y})^2$
1	58	1	57	25
2	58	1	55	9
3	56	1	63	121
4	38	361	24	784
5	70	169	67	225
6	38	361	43	81
7	42	225	33	361
8	75	324	68	256
9	68	121	56	16
10	67	100	54	4
	$\sum_i x_i = 570$ $\overline{x} = 57,0$	$ns_1^2 = \sum_i (x_i - \overline{x})^2$ $= 1664$	$\sum_i y_i = 520$ $\overline{y} = 52,0$	$ns_2^2 = \sum_i (y_i - \overline{y})^2$ $= 1882$

Из данных, приведенных в таблице, вытекает: $s_1^2 = 166,4$ и $s_2^2 = 188,2$. Следовательно,

$$t_{\text{набл}} = (\overline{x} - \overline{y}) \left[\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{-1/2} =$$

$$= (57 - 52) \left[\frac{1664 + 1882}{10 + 10 - 2} \left(\frac{1}{10} + \frac{1}{10} \right) \right]^{-1/2} = 0,797.$$

Число степеней свободы $\nu = n_1 + n_2 - 2$ составляет $\nu = 18$. По таблицам квантилей распределения Стьюдента по заданному уровню значимости $\alpha = 0,05$ находим квантиль $t_{0,05;\,18} = 1,734$.

Так как $t_{\text{набл}} = 0,797 < 1,734$, то нет оснований для отклонения нулевой гипотезы равенства среднего времени, затрачиваемого на изготовление детали двумя типами резцов.

Это означает, что разность в средних арифметических $(\overline{x} - \overline{y})$ в пользу нового типа резца является статистически незначимой (случайной). Другими словами, преимущество нового типа резца осталось недоказанным, хотя это не означает, что этого преимущества нет. При большем объеме выборок это преимущество, если оно действительно имеется, может быть доказано.

Пример 4.3

Контролер автопарка определил, что расход топлива на одной машине в среднем составил $m = 10,0$ *л* на 100 *км*. С целью уменьшения расхода топлива была проведена модернизация двигателей $n = 25$ автомашин. После модернизации оказалось, что расход топлива у этих 25 автомашин составил $X^*_{25} = 9,3$ *л* на 100 *км*. Известно, что рассмотренная выборка является нормальной с $\overline{\mathrm{X}} = m$ и $\sigma^2 = 4\,л^2$.

Требуется проверить гипотезу: *модернизация не повлияла на расход топлива.*

Решение

1) Рассмотрим две гипотезы:

а) $\{H_0 : \ m = 10,0\}$ (*модернизация не повлияла на расход*);

б) $\{H_1 : \ m < 10,0\}$ (*модернизация привела к уменьшению расхода*).

2) Примем уровень значимости $\alpha = 0,05$.

3) Используем в качестве статистики критерия оценку математического ожидания X^*_{25}.

4) Поскольку выборка из нормальной совокупности, то выборочное значение также нормальное с дисперсией, удовлетворяющей соотношению $\sigma^{*2}/n = 4/25$, т. е. $\sigma^* = 0,4\,(л)$.

Если справедлива гипотеза H_0, то $m = 10,0$. Перейдем к стандартизованной переменной $U = (X^*_{25} - m)\,/0,4$, являющейся нормальной случайной величиной $\mathcal{N}(0,\,1)$.

5) Альтернативная гипотеза $\{H_1 : m < 10,0\}$. Поэтому следует использовать односторонний критерий. В данном случае этот критерий является левосторонним.

Критическая область определяется из неравенства $U \le u_\alpha$, т.е.

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u_\alpha} \exp\left(-t^2/2\right)\, dt \ = \ \alpha \ = \ 0,05.$$

По таблицам функции Лапласа по уровню значимости $\alpha = 0,05$ находим квантиль распределения $u_{0,05}$. Это значение (см. приложение) равно $u_{0,05} = -1,64$.

6) Выборочное значение стандартизованной статистики критерия составляет

$$u_{\text{набл}} \ = \ \frac{X^*_{25} - 10,0}{0,4} \ = \ \frac{9,3 - 10,0}{0,4} \ = \ -1,75.$$

7) Примем статистическое решение.

Поскольку $u_{\text{набл}} < u_{0,05}$, т.е. выборочное значение статистики принадлежит критической области, то гипотеза H_0 отклоняется.

Ответ: Принимается решение $\{H_1$: *модернизация двигателей привела к уменьшению расхода топлива*$\}$.

Примечание. Для принятого уровня значимости α граница критической области составляет $X_{\text{кр}} = 10,0 - 0,4 \cdot 1,645 = 9,342$.

Пример 4.4

Выборка в 50 электроламп завода А показала среднюю продолжительность работы $\overline{x} = 1282$ *час* со средним квадратическим отклонением $s_1 = 80$ *час*, а такая же по объему выборка того же типа ламп с завода Б показала $\overline{y} = 1208$ *час* со средним квадратическим отклонением $s_2 = 94$ *час*.

Проверить гипотезу о том, что эти заводы выпускают лампы одинакового качества (средний срок службы ламп обоих заводов одинаков). Уровень значимости принять равным $\alpha = 0,05$.

Решение

Так как объемы выборок достаточно велики, то применим модель 1. При этом дополнительно предположим, что продолжительности работы электроламп, выпускаемых заводами А и Б, являются случайными величинами, распределенными по нормальному закону $X \to \mathcal{N}(a_1; s_1)$ и $Y \to \mathcal{N}(a_2; s_2)$, причем $s_1 = 80$, $s_2 = 94$, а величины a_1 и a_2 неизвестны. Согласно условию, необходимо проверить нулевую гипотезу $\{H_0 : a_1 = a_2\}$ (*средний срок службы ламп, выпускаемых заводами* А *и* Б, *одинаков*) против альтернативной гипотезы $\{H_a : a_1 > a_2\}$ (*лампы, выпускаемые заводом* А, *имеют больший срок службы*).

Для проверки нулевой гипотезы H_0 применим правосторонний u–критерий. Вычислим наблюдаемое значение статистики

$$u_{\text{набл}} = \frac{\overline{x} - \overline{y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{1282 - 1208}{\sqrt{80^2/50 + 94^2/50}} = 4,24.$$

По таблице функции Лапласа найдем критическую точку (квантиль) $u_{0,05}$, удовлетворяющую условию $\Pr(u \geq u_{0,05})$. Это значение равно 1,64.

Так как $u_{\text{набл}} = 4,24 > 1,64$, то нулевая гипотеза отклоняется в пользу альтернативной. Другими словами, считается "статистически доказанным", что срок службы ламп, выпускаемых заводом А, больше срока службы ламп, выпускаемых заводом Б.

Пример 4.5

Точность работы станка-автомата проверяется по дисперсии контролируемого размера деталей, которая не должна превышать $\sigma_0^2 = 0,04$. Взята проба из 11 случайно отобранных деталей и получены следующие результаты (в миллиметрах):

100,6 99,6 100,0 100,1 100,3 100,0 99,9 100,2 100,4 100,6 100,5

На основании имеющихся данных проверить, обеспечивает ли станок заданную точность. Уровень значимости принять равным 0,05.

Решение

Из условия примера следует, что необходимо проверить нулевую гипотезу $\{H_0 : \sigma = 0,04\}$ (*станок обеспечивает заданную точность*) против альтернативной гипотезы $\{H_a : \sigma > 0,04\}$ (*станок не обеспечивает заданную точность*).

Альтернативная гипотеза сформулирована в виде $\{H_a : \sigma > 0,04\}$, поэтому случай, когда $\sigma < 0,04$, не является существенным. Если в действительности и

окажется, что $\sigma < 0,04$, то это означает, что станок хорошо налажен и выпускает детали более высокого качества, чем предполагалось.

Найдем точечные оценки параметров нормального закона:

$$\hat{a} = \overline{x} = 100,2;$$

$$\hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{i=1}^{11}(x_i - \overline{x})^2 = \frac{1,00}{11} = 0,091.$$

Для проверки нулевой гипотезы применим критерий χ^2 с правосторонней критической областью.

Вычислим наблюдаемое значение тестовой статистики

$$\chi^2_{\text{набл}} = \frac{1}{\sigma_0^2}\sum_{i=1}^{11}(x_i - \overline{x})^2 = \frac{1}{0,04} = 25.$$

По таблицам квантилей χ^2-распределения (см. приложение) по заданному уровню значимости $\alpha = 0,05$ и числу степеней свободы $\nu = 10$ находим критическую точку $\chi^2_{\alpha;\,n-1} = \chi^2_{0,05;\,10}$, удовлетворяющую условию $\Pr\left(\chi^2 \geq \chi^2_{0,05;\,10}\right)$. Это значение равно 18,307.

Так как $\chi^2_{\text{набл}} = 25 > 18,307$, нулевая гипотеза отклоняется в пользу альтернативной. Это значит, что станок не обеспечивает заданную точность и требует подналадки.

Пример 4.6

На станке-автомате изготавливается деталь с номинальным контролируемым размером $a = 12$ *мм*. Известно, что распределение контролируемого размера является нормальным $\mathrm{X} \to \mathcal{N}(a;\,0,5)$. Отдел технического контроля в течение смены произвел измерение 36 случайно отобранных деталей и подсчитал средний размер контролируемого параметра $\overline{x} = 11,7$ *мм*.

Можно ли утверждать, что станок-автомат изготавливает детали уменьшенного размера и поэтому требуется произвести подналадку станка?

Решение

Из условия примера следует, что необходимо проверить нулевую гипотезу $\{H_0 :\ a = 12$ *мм* $\}$ (*автомат изготавливает детали номинального размера*) против альтернативной гипотезы $\{H_a :\ a < 12$ *мм* $\}$ (*автомат изготавливает детали, размер которых меньше номинального*).

Так как отдел технического контроля подозревает, что автомат разрегулировался и производит детали уменьшенного размера, то для проверки нулевой гипотезы применим критерий, соответствующий модели 1 с левосторонней критической областью.

Вычислим наблюдаемое значение критерия

$$u_{\text{набл}} = \frac{\overline{x} - a}{\sigma}\sqrt{n} = \frac{11,7 - 12,0}{0,5}\sqrt{36} = -3,6.$$

По таблицам функции Лапласа по уровню значимости $\alpha = 0,05$ находим квантиль распределения $-u_{0,05}$, удовлетворяющий условию $\Pr(u \leq -u_{0,05})$. Это значение (см. приложение) равно –1,64.

Так как наблюдаемое значение критерия u находится в критической области $u_{\text{набл}} = -3,6 < -1,64$, то гипотезу $\{H_0 :\ a = 12\ мм\}$ следует отклонить в пользу альтернативной гипотезы. Это означает, что с вероятностью ошибки, меньшей чем 0,05, можно утверждать, что контролируемый размер деталей, изготавливаемых на станке-автомате, является заниженным по сравнению с номинальным размером и поэтому необходимо произвести подналадку станка.

Пример 4.7

Следует проверить, что три марки строительного бетона имеют одинаковое рассеивание прочности на сжатие. Для проверки этой гипотезы произведено измерение прочности на сжатие и получены следующие результаты ($кг/см^2$).

Бетон марки № 1	Бетон марки № 2	Бетон марки № 3
195	215	201
200	201	204
204	202	221
205	198	210
201	—	199

Уровень значимости принять $\alpha = 0,05$.

Решение

Предположим, что результаты измерений прочности на сжатие трех марок бетона подчиняются нормальным или приближенно нормальным распределениям. Согласно условию, нам необходимо проверить нулевую гипотезу $\{H_0 : \sigma_1 = \sigma_2 = \sigma_3\}$ против альтернативной гипотезы $\{H_a :\ не\ все\ эти\ дисперсии\ равны\ между\ собой\}$.

Объединим вспомогательные вычисления, необходимые для расчета выборочной статистики χ^2 по формуле (4.34), в таблицу.

№	x_{1i}	x_{2i}	x_{3i}	$(x_{1i} - \overline{x}_1)^2$	$(x_{2i} - \overline{x}_2)^2$	$(x_{3i} - \overline{x}_3)^2$
1	195	215	201	36	121	36
2	200	201	204	1	9	9
3	204	202	221	9	4	196
4	205	198	210	16	36	9
5	201	—	199	0	—	64
Суммы	1005	816	1035	62	170	314

Отсюда:

$$\overline{x}_1 = 201; \qquad \overline{x}_2 = 204; \qquad \overline{x}_3 = 207;$$

$$\overline{s_1^2} = 15,5; \qquad \overline{s_2^2} = 56,7; \qquad \overline{s_3^2} = 78,7.$$

Вычислим взвешенную среднюю арифметическую эмпирических дисперсий:

$$\overline{s^2} = \frac{1}{14-3} \cdot 546 = 49,64;$$

$$\lg \overline{s^2} = 1,696; \qquad (n-k) \lg \overline{s^2} = (14-3) \cdot 1,696 = 18,656.$$

Далее вычислим

i	s_i^2	$\lg s_i^2$	$n_i - 1$	$(n_i - 1)\ \lg s_i^2$
1	15,5	1,190	4	4,760
2	56,7	1,754	3	5,262
3	78,7	1,895	4	7,580

Отсюда найдем сумму

$$\sum_{i=1}^{3}(n_i - 1)\lg s_i^2 = 17,602.$$

Постоянная C равна

$$C \;=\; 1 + \frac{1}{3 \cdot 2}\left[\left(\frac{1}{4} + \frac{1}{3} + \frac{1}{4}\right) - \frac{1}{11}\right] \;=\; 1,124.$$

Рассчитаем наблюдаемое значение выборочной статистики χ^2:

$$\chi^2_{\text{набл}} \;=\; \frac{2,303}{1,124} \cdot (18,656 - 17,602) \;=\; 2,049 \cdot 1,053 = 2,158.$$

Найдем по таблице квантилей χ^2-распределения по уровню значимости $\alpha = 0,05$ и числу степеней свободы $\nu = k - 1 = 2$ значение критической точки $\chi^2_{0,05;\,2} = 5,99$.

Так как $\chi^2_{\text{набл}} = 2,158 < 5,99,$ то нет оснований для отклонения нулевой гипотезы. Это значит, что имеющаяся информация о рассеивании прочности на сжатие трех марок бетона не дает оснований считать, что их уровень рассеивания является различным.

Пример 4.8

Двумя методами произведены измерения одной и той же физической величины. Первым методом эта величина измерялась 10 раз. Получены следующие результаты:

$$\overline{x}_1 \;=\; 10,28, \qquad s_1^2 \;=\; \frac{1}{n_1 - 1}\sum_{i=1}^{10}(x_{i1} - \overline{x}_1)^2 = 0,00084.$$

Вторым методом эта же величина измерялась 8 раз, что дало

$$\overline{x}_2 \;=\; 10,30, \qquad s_2^2 \;=\; \frac{1}{n_2 - 1}\sum_{i=1}^{8}(x_{i2} - \overline{x}_2)^2 = 0,00041.$$

Можно ли считать, что оба метода обеспечивают одинаковую точность? Уровень значимости принять $\alpha = 0,05$. Предполагается, что результаты измерений распределены нормально и выборки независимы.

Решение

Из условия примера следует, что нам необходимо проверить нулевую гипотезу $\{H_0 : \sigma_1^2 = \sigma_2^2\}$ (*оба метода обеспечивают одинаковую точность*) против альтернативной гипотезы $\{H_a : \sigma_1^2 > \sigma_2^2\}$ (*второй метод измерений обеспечивает более высокую точность*).

Вычислим наблюдаемые значения F-критерия:

$$F_{\text{набл}} = \frac{0,00084}{0,00041} = 2,05.$$

По таблице квантилей F-распределения (см. приложение) по уровню значимости $\alpha = 0,05$ и числу степеней свободы $\nu_1 = 10 - 1 = 9$ и $\nu_2 = 8 - 1 = 7$ находим критическую точку $F_{0,05;\,9;\,7} = 3,68$.

Так как $F_{\text{набл}} = 2,05 < 3,68$, то нет основания для отклонения нулевой гипотезы. Другими словами, имеющаяся информация о точности этих методов не дает основания считать, что второй метод измерения лучше первого.

Пример 4.9

Группа социологов исследовала влияние стажа работы по профессии на производительность труда рабочих механического цеха некоторого завода. Получены следующие результаты:

	Стаж работы		
Результативный признак	до 10 лет	от 10 лет до 15 лет	от 15 лет до 25 лет
Количество деталей, вырабатываемых за смену одним рабочим, штук	135 156 165 — —	176 196 204 180 —	155 160 149 171 140

Предполагая, что производительность труда рабочих, имеющих различный стаж работы, подчиняется нормальному закону: $\mathrm{X}_i \to \mathcal{N}(a_i;\, \sigma_i)$ $(i = 1,\, 2,\, 3)$, причем $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, требуется проверить методом дисперсионного анализа нулевую гипотезу $\{H_0 : a_1 = a_2 = a_3\}$ (*средняя производительность труда не зависит от стажа работы*). Уровень значимости $\alpha = 0,05$.

Решение

Согласно условию примера, нулевая и альтернативная гипотезы имеют вид $\{H_0 : a_1 = a_2 = a_3\}$ – *производительность труда не зависит от стажа работы*, $\{H_a : a_1 \neq a_2 \neq a_3\}$ – *производительность труда зависит от стажа работы*.

Вычислим вспомогательные величины, необходимые для составления таблицы дисперсионного анализа:

а) групповые средние арифметические

$$n = n_1 + n_2 + n_3 = 3 + 4 + 5 = 12;$$

$$\overline{x}_1 = \frac{456}{3} = 152; \qquad \overline{x}_2 = \frac{756}{4} = 189; \qquad \overline{x}_3 = \frac{775}{5} = 155;$$

б) общую среднюю арифметическую

$$< x > = \frac{1}{n} \sum_{i=1}^{3} \sum_{j=1}^{n_i} x_{ij} = \frac{1987}{12} = 165,58.$$

На основе имеющейся информации требуется проверить нулевую гипотезу $\{H_0 : a_1 = a_2 = a_3$ (*средняя производительность труда не зависит от стажа работы*)$\}$ против альтернативной гипотезы $\{H_1$: *средняя производительность труда зависит от стажа работы*$\}$.

Согласно (4.48) проверка нулевой гипотезы основывается на вычислении выборочной статистики ($k = 3$)

$$F_{\text{набл}} = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}} = \left[\frac{1}{k-1}\sum_{i=1}^{k}(\overline{x}_i - \langle x\rangle)^2 n_i\right]\left[\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 n_i\right]^{-1}.$$

Если нулевая гипотеза верна, то эта статистика имеет F–распределение с $\nu_1 = k - 1 = 2$ и $\nu_2 = n - k = 9$ степенями свободы.

Составим таблицу дисперсионного анализа:

Источник изменчивости	Суммы квадратов	Число степеней свободы	Дисперсия
Стаж работы (между группами)	3306,92	2	1653,5
Остаточная (внутри групп)	6228	9	692,0
Полная изменчивость	9534,92	11	

Имеем $s^2_{\text{факт}} = 1653,56$ и $s^2_{\text{ост}} = 692$, откуда наблюденное значение F-критерия составляет

$$F_{\text{набл}} = 2,389.$$

По таблице квантилей F-распределения (см. приложение) по уровню значимости $\alpha = 0,05$ и числу степеней свободы $\nu_1 = 2$ и $\nu_2 = 9$ находим критическую точку $F_{0,05;\,2;\,9} = 4,26$. Так как $F_{\text{набл}} = 2,389 < 4,26 = F_{0,05;\,2;\,9}$, то принимаем решение, что нет оснований для отклонения нулевой гипотезы, т.е. считается статистически доказанным, что средняя производительность труда не зависит от стажа работы.

Пример 4.10

Анкетным обследованием установлено, что 32% студентов университета являются слушателями телевизионных лекций по математической статистике. Кафедра прикладной математики напечатала конспекты телевизионных лекций и разослала их студентам. После этого вновь было произведено анкетирование и установлено, что 80 студентов из 200 опрошенных оказались слушателями телевизионных лекций, $m/n = 80/200 = 0,4$.

Можно ли считать, что издание телевизионного конспекта лекции способствовало увеличению контингента студентов, слушающих телевизионные лекции по математической статистике (уровень значимости $\alpha = 0,05$)?

Решение

Согласно условию, необходимо проверить нулевую гипотезу $\{H_0 : p = 0,32\}$ против альтернативной гипотезы $\{H_a : p > 0,32\}$. Для проверки данной нулевой

гипотезы применим u-критерий с правосторонней критической областью. Вычислим наблюдаемое значение критерия

$$u_{\text{набл}} = \frac{m/n - p_0}{\sqrt{\overline{p}_0\,\overline{q}_0}}\sqrt{n} = \frac{0,40 - 0,32}{\sqrt{0,32 \cdot 0,68}}\sqrt{200} = 2,43.$$

По таблицам функции Лапласа (см. приложение) по заданному уровню значимости $\alpha = 0,05$ найдем критическое значение u_α, удовлетворяющее условию $\Pr(u \geq u_\alpha)$. Это значение равно 1,64.

Так как $u_{\text{набл}} = 2,43 > 1,64$, то нулевая гипотеза отвергается в пользу альтернативной, т. е. считается, что процент студентов, слушающих телевизионные лекции по математической статистике, значительно увеличился.

Пример 4.11

На экзамене по данному предмету экзаменатор задает студенту только один вопрос по одной из четырёх частей курса. Из 100 студентов 26 получили вопрос из первой части, 32 – из второй, 17 – из третьей и остальные по четвертой.

Можно ли по этим результатам принять гипотезу, что для пришедшего на экзамен студента имеется одинаковая вероятность получить вопрос по любой из четырех частей? Принять $\alpha = 0{,}05$.

Решение

В данном случае: $m_1 = 26$, $m_2 = 32$, $m_3 = 17$, $m_4 = 25$, откуда следует, что $p_i = 0{,}25$, $n = 100$, $np_i = 25$ $(i = 1, 2, 3, 4)$.

Находим

$$\chi_0^2 = \frac{(26-25)^2}{25} + \frac{(32-25)^2}{25} + \frac{(17-25)^2}{25} + \frac{(25-25)^2}{25} = 4,56.$$

Поскольку ни один из параметров предполагаемого распределения нами не находился по выборке, то для параметра s имеем $s = 0$. Поэтому число степеней свободы равно $k - s - 1 = 4 - 0 - 1 = 3$. По таблицам находим границу критической области. Для $\alpha = 0{,}05$ она равна 7,815.

Так как $4,56 < 7,815$, то гипотеза подтвердилась.

Пример 4.12

Два завода изготавливают однотипные детали. Для оценки их качества взяты выборки из продукции этих заводов и получены следующие результаты:

завод № 1

– объем выборки $n_1 = 200$, число бракованных деталей $m_1 = 20$;

завод № 2

– объем выборки $n_2 = 300$, число бракованных деталей $m_2 = 15$.

Определить, имеется ли существенное различие в качестве деталей, изготавливаемых этими заводами. Уровень значимости $\alpha = 0,05$.

Решение

Согласно условию, требуется проверить нулевою гипотезу $\{H_0 : p_1 = p_2\}$ (*доли бракованных деталей, изготавливаемых заводами* № 1 *и* 2, *равны*) против альтернативной гипотезы $\{H_a : p_1 \neq p_2\}$.

Вычислим наблюдаемое значение u-критерия. Так как

$$m_1/n_1 = \frac{20}{200} = 0,10; \quad m_2/n_2 = \frac{15}{300} = 0,05; \quad n = \frac{n_1 \cdot n_2}{n_1 + n_2} = 120;$$

$$\overline{p} = \frac{m_1 + m_2}{n_1 + n_2} = 0,07; \qquad \overline{q} = 1 - \overline{p} = 0,93,$$

то

$$u_{\text{набл}} = \frac{m_1/n_1 - m_2/n_2}{\sqrt{\overline{p}\,\overline{q}}}\sqrt{n} = \frac{0,10 - 0,05}{\sqrt{\overline{p}\,\overline{q}}}\sqrt{120} = 2,15.$$

По таблицам функции Лапласа (см. приложение) по заданному уровню значимости $\alpha = 0,05$ найдем критическое значение $u_{0,025} = 1,96$.

Поскольку $u_{\text{набл}} > 1,96$, то нулевая гипотеза отклоняется в пользу альтернативной, т. е. считается, что качество деталей, изготавливаемых этими заводами, различно.

Пример 4.13 (Проверка гипотезы о независимости случайных величин)

Вследствие универсальности χ^2-критерия и его применимости к многомерным распределениям он может служить и для проверки гипотезы о независимости случайных величин. Предположим, что область значений величины X разбита на r_1 интервалов, а область значений величины Y – на r_2 интервалов. Пусть $\hat{P}_{ij}$ и p_{ij} – случайная частота и вероятность попадания вектора $[X^{\mathrm{T}}, Y^{\mathrm{T}}]^{\mathrm{T}}$ в пересечение i-го интервала значений X и j-го интервала значений Y $(i = 1, 2, \ldots, r_1;\ j = 1, 2, \ldots, r_2)$. Если X и Y независимы, то $p_{ij} = p_{i.} \cdot p_{.j}$, где $p_{i.}$ и $p_{.j}$ – вероятности попадания X в i-й интервал и Y в j-й $(i = 1, 2, \ldots, r_1;\ j = 1, 2, \ldots, r_2)$. Вероятности $p_{i.}$ и $p_{.j}$ можно рассматривать как $r_1 + r_2 - 2$ неизвестных параметра распределения вектора $[X^{\mathrm{T}}, Y^{\mathrm{T}}]^{\mathrm{T}}$, при этом нужно иметь в виду соотношения $\sum_i p_{i.} = 1$ и $\sum_j p_{.j} = 1$.

Эффективными асимптотически нормальными оценками вероятностей $p_{i.}$ и $p_{.j}$ могут служить соответствующие частоты

$$\hat{P}_{i.} = \sum_{j=1}^{r_2} \hat{P}_{ij}, \qquad \hat{P}_{.j} = \sum_{i=1}^{r_1} \hat{P}_{ij}.$$

Поэтому величина

$$Z = n \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \frac{\left(\hat{P}_{ij} - \hat{P}_{i.}\,\hat{P}_{.j}\right)^2}{\hat{P}_{i.}\,\hat{P}_{.j}}$$

имеет асимптотическое χ^2-распределение с $k = r_1 r_2 - (r_1 + r_2 - 2) - 1 = (r_1 - 1)(r_2 - 1)$ степенями свободы.

Это дает возможность проверять гипотезу о независимости двух величин (как скалярных, так и векторных).

Пример 4.14 (Проверка гипотез о совпадении распределений)

Предположим, что в результате N независимых последовательностей опытов, содержащих $n_1, n_2, \ldots, n_N$ наблюдений, $n_1 + n_2 + \ldots + n_N = n$, получены частоты попаданий наблюдаемой величины X в интервалы Δ_ν, $\nu = 1, 2, \ldots, r$, на которые

разбита область её возможных значений. Требуется проверить гипотезу о совпадении распределений величины X (или N различных наблюдаемых величин) в этих N последовательностях опытов.

Введем обозначения: $\hat{P}_{\mu\nu}$ – случайная частота попадания наблюдаемой величины в ν-й интервал в μ-й последовательности опытов, p_ν — вероятность попадания в ν-й интервал ($\nu = 1, 2, \ldots, r$ и $\mu = 1, 2, \ldots, N$).

Так как сумма независимых величин с χ^2-распределением каждая имеет χ^2-распределение с суммарным числом степеней свободы, то случайная величина

$$Z = \sum_{\mu=1}^{N} n_\mu \sum_{\nu=1}^{r} \frac{\left(\hat{P}_{\nu\mu} - p_\nu\right)^2}{p_\nu}$$

при данных известных вероятностях $p_1, p_2, \ldots, p_r$ описывается асимптотическим χ^2-распределением с $N(r-1)$ степенями свободы, если распределение наблюдаемой величины одно и то же во всех сериях опытов.

Это дает возможность проверять гипотезу о том, что во всех N последовательностях опытов наблюдаемая величина имеет одно и то же распределение, для которого вероятности попадания в интервалы имеют данные значения $p_1, p_2, \ldots, p_r$.

4.12. Задачи для решения

Задача 4.1

Рассматривается случайная величина Z = X – Y, где X и Y – независимые случайные величины. Выборочные оценки для X и Y определялись по результатам $n_1 = 16$ и $n_2 = 36$ наблюдений соответственно.

Найти 95 %–й доверительный интервал для математического ожидания Z, если $\overline{x} = 10$, $\overline{y} = 4$, σ_{X} и σ_{Y} известны и таковы: $\sigma_{\text{X}} = 1$ и $\sigma_{\text{Y}} = 4$.

Задача 4.2

В контейнере содержатся готовые болты с номинальным значением контролируемого размера $m_0 = 40$ *мм*. Была взята выборка болтов объема $n = 36$. Выборочное среднее контролируемого размера болтов оказалось равным $\overline{x} = 40,2$ *мм*. Результаты предыдущих измерений дают основание предполагать, что действительные размеры болтов образуют нормальную совокупность с дисперсией $\sigma^2 = 1$ *мм*^{2}.

Можно ли по результатам проведенного выборочного обследования утверждать, что контролируемый размер болтов не имеет положительного смещения по отношению к номинальному размеру? Принять $\alpha = 0,10$. Какова критическая область в этом случае?

Задача 4.3

С автоматической линии, производящей подшипники, было отобрано 400 штук, причем 10 оказались бракованными.

Найти 90%-й доверительный интервал для вероятности появления бракованного подшипника. Сколько подшипников надо проверить, чтобы с вероятностью P = 0,9973

можно было бы утверждать, что вероятность появления бракованного подшипника не отличается от частоты более чем на 5% ?

Задача 4.4

Выборочно обследовали качество кирпича. Из 1600 проб в 32 случаях кирпич оказался бракованным.

Требуется определить, в каких пределах заключается доля брака для всей продукции, если результат необходимо гарантировать с вероятностью P = 0,945.

Задача 4.5

Из урны, содержащей неотличимые на ощупь черные и белые шары в неизвестной пропорции, случайным образом извлекается 100 шаров (с возвращением). Среди них оказалось 39 черных шаров.

Найти: 90%-е и 95%-е доверительные интервалы для доли черных шаров.

Задача 4.6

В 10000 сеансах игры с автоматом выигрыш появился 4000 раз.

Найти 95%-й доверительный интервал для вероятности выигрыша. Сколько сеансов игры следует провести, чтобы с вероятностью P = 0,99 вероятность выигрыша отличалась от частоты не более чем на 1 % ?

Задача 4.7

Из большой партии транзисторов одного типа были случайным образом отобраны и проверены 100 штук. У 36 транзисторов один номинальный параметр оказался ниже допустимого.

Найти 95%-й доверительный интервал для доли транзисторов с таким дефектом из всей партии.

Задача 4.8

По схеме повторной выборки произведено выборочное измерение выработки на земляных работах у 145 рабочих. В результате этого обследования средняя выработка определена в 4,95 $м^3$ на одного рабочего, а среднее квадратическое отклонение оказалось равным 1,5 $м^3$.

Найти доверительные границы для генерального среднего, отвечающие вероятности P = 0,9973.

Задача 4.9

При просмотре 10000 волокон из партии льна обнаружено 1200 недозрелых.

Сколько надо просмотреть волокон льна из этой партии, чтобы с вероятностью P = 0,997 можно было ручаться за точность определения доли недозрелых волокон из всей партии в пределах 1% ? Отбор бесповторный.

Задача 4.10

Для проверки утверждения о том, что вероятность отказа прибора p равна 0,01, было проведено испытание 100 приборов. При этом один прибор отказал.

Построить 95%-ю доверительную границу одностороннего доверительного интервала для p по этим данным.

4.13. Задание на практическую работу

Настоящая практическая работа рассчитана на два часа и содержит два задания. Задания должны выполняться в выбранной программной среде.

Задание 1

Сахар, получаемый на склад, упакован в мешки, при этом каждый мешок должен содержать в среднем 50 *кг* полезного веса. Появились основания предполагать, что в действительности в мешках содержится сахар меньшего веса. Для проверки предположения произведены измерения веса сахара в $n = 16$ мешках (данные о выборочных средних приводятся).

Требуется выяснить, можно ли по имеющимся измерениям отклонить нулевую гипотезу, о том, что действительный полезный вес соответствует норме.

Уровень значимости принять равным $\alpha = 0,01$.

Результат работы – принятие решения относительно нулевой гипотезы.

При каком значении уровня значимости α нулевая гипотеза будет отклонена?

Вариант 1

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 49,2\,\text{кг}, \qquad s_{\text{несм}} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{1/2} = 0,3\,\text{кг}.$$

Вариант 2

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 49,5\,\text{кг}, \qquad s_{\text{несм}} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{1/2} = 0,2\,\text{кг}.$$

Вариант 3

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 49,8\,\text{кг}, \qquad s_{\text{несм}} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{1/2} = 0,1\,\text{кг}.$$

Задание 2

На первом и втором станках производятся детали. Выдвинута гипотеза, что второй станок (нового типа) сокращает время обработки некоторой детали. Проведено 10 измерений времени, затрачиваемого на обработку этой детали первым и вторым станком (данные приводятся в таблице).

Требуется проверить нулевую гипотезу о равенстве среднего времени, затрачиваемого на изготовление этой детали на каждом из этих станков. Уровень значимости принять равным $\alpha = 0,05$.

Результаты оформите графически.

Результат работы – принятие решения относительно нулевой гипотезы.

При каком значении уровня значимости α нулевая гипотеза будет отклонена?

Вариант 1

первый станок – 38, 38, 36, 18, 50, 18, 22, 55, 48, 47;

второй станок – 27, 35, 43, 14, 47, 23, 13, 48, 36, 34.

Вариант 2

первый станок – 35, 37, 38, 21, 52, 19, 24, 54, 49, 47;
второй станок – 26, 34, 43, 15, 46, 22, 14, 43, 33, 34.

Вариант 3

первый станок – 48, 48, 46, 28, 60, 28, 32, 75, 58, 57;
второй станок – 37, 45, 53, 24, 57, 33, 23, 58, 46, 44.

4.14. Задания для проверки

1. Дайте определение статистической гипотезы.

2. Дайте определение параметрической статистической гипотезы, непараметрической статистической гипотезы.

3. Приведите примеры нулевой, альтернативной, простой и сложной гипотез. Объясните принцип проверки нулевых гипотез с помощью статистических критериев значимости.

4. Что называется ошибкой первою рода, ошибкой второго рода? Дайте геометрическую интерпретацию вероятностям совершения ошибок первого и второго рода.

5. Как изменяются вероятности совершения ошибок первого и второго рода при увеличении объема выборки?

6. Как изменяется вероятность совершения ошибок второго рода при $\alpha \to 0$?

7. Зависят ли вероятности совершения ошибок первого и второго рода от вида альтернативной гипотезы, от применяемого критерия?

8. В чем состоит односторонность действия статистических критериев значимости?

9. Принимается ли во внимание вероятность совершения ошибки второго рода при проверке нулевых гипотез с помощью статистических критериев значимости?

10. Можно ли, применяя статистический критерий значимости, сделать вывод: "Проверяемая нулевая гипотеза верна"?

11. В чем состоит различие между построением двусторонней критической области и построением доверительного интервала для одного и того же параметра?

12. Как находятся критические точки (квантили) статистических критериев значимости (u, t, χ^2, F) в случае двусторонней критической области, в случае левосторонней критической области, в случае правосторонней критической области?

13. С помощью каких выборочных статистик производится проверка гипотез о математических ожиданиях одной случайной величины двух случайных величин, нескольких случайных величин?

14. С помощью каких выборочных статистик производится проверка гипотез о дисперсиях одной случайной величины, двух случайных величин, нескольких случайных величин?

15. С помощью каких выборочных статистик производится проверка гипотез о параметре p биномиального закона распределения?

5. Статистическая проверка непараметрических гипотез

5.1. Основные понятия

В предыдущих разделах было уделено внимание проверке гипотез относительно параметров законов распределения, вид которых предполагался известным (нормальный, биномиальный и т. д.).

При обработке опытных статистических данных для характеристики частотных свойств ряда наблюдений $x_1, x_2, \ldots, x_n$ экспериментатор подбирает теоретико-вероятностную модель (нормальную, показательную, биномиальную и т. д.) этого ряда. Предположим, что экспериментатор визуально по виду гистограммы (полигона частостей) или из каких-либо других соображений выдвинул гипотезу о множестве функций Ω определенного вида (нормальных, показательных, биномиальных и т. д.), к которому может принадлежать функция распределения исследуемой случайной величины X. Предположения такого рода называются *непараметрическими гипотезами.*

Определение. *Нулевой непараметрической гипотезой* называется гипотеза относительно общего вида функции распределения случайной величины X, т. е. гипотеза вида $\{H_0 : F(x) = F_0(x)\}$.

Гипотетическая (предполагаемая) функция распределения случайной величины X может быть определена полностью либо с точностью до её параметров, т. е. нулевая гипотеза может иметь вид $\{H_0 : F(x) \in \Omega\}$, где Ω означает совокупность функций определенного вида (нормальных, показательных, биномиальных и т. д.).

Предположим, что класс таких функций выбран и произведена точечная оценка параметров этих функций внутри выбранного класса. Дальнейшая задача экспериментатора состоит в проверке выдвинутой гипотезы о классе функций Ω, т.е. в выяснении, насколько хорошо подобрана вероятностная модель ряда наблюдений.

Проверка гипотезы о предполагаемом законе распределения производится с помощью непараметрических критериев значимости. Принципы построения таких критериев и методика проверки остаются практически теми же, что и при проверке параметрических гипотез, т. е. проверка непараметрических гипотез производится на основании вычисления некоторой выборочной статистики (критерия), закон распределения которой получен в предположении истинности нулевой гипотезы, и сравнения наблюдаемого значения этой выборочной статистики с критическим значением.

Непараметрические критерии значимости условно можно подразделить на две группы.

К первой группе относятся *критерии согласия*, с помощью которых проверяются нулевые гипотезы относительно общего вида функции распределения. Наиболее распространенными критериями согласия являются критерий согласия χ^2 Пирсона и λ-критерий Колмогорова.

К другой многочисленной *группе непараметрических критериев* относятся критерии, с помощью которых проверяется нулевая гипотеза о принадлежности двух выборок одной и той же генеральной совокупности (или о том, что две генеральные совокупности имеют одну и ту же функцию распределения).

В практике с помощью непараметрических критериев значимости особенно часто проверяется нулевая гипотеза о том, что исследуемая генеральная совокупность имеет нормальный закон распределения.

Например, при применении параметрических критериев для проверки гипотез относительно параметров нормального закона предполагалось, что генеральная совокупность нормальна. Это предположение предварительно следует проверить с помощью критериев согласия, а затем, если нулевая гипотеза не отклонена, применять параметрические критерии значимости.

5.2. Критерий согласия χ^2 Пирсона

Критерий χ^2 Пирсона позволяет производить проверку согласия эмпирической функции распределения с гипотетической функцией $F(x)$, принадлежащей к некоторому множеству Ω функций определенного вида (нормальных, показательных, биномиальных и т.д.). Сформулируем основные вероятностные предпосылки и ограничения, которые должны быть выполнены при применении критерия χ^2 в виде модели.

Модель. Пусть генеральная совокупность имеет функцию распределения $F(x)$, принадлежащую некоторому классу функций Ω. Из генеральной совокупности извлечена выборка объема n $(n \geq 50)$.

Разобьем весь диапазон полученных результатов на k частичных интервалов равной длины и пусть в каждом частичном интервале оказалось m_i измерений, причем

$$\sum_{i=1}^{k} m_i \;=\; n. \tag{5.1}$$

Составим сгруппированный статистический ряд :

Интервалы наблюденных значений СВ X	$[x_0; x_1]$	$[x_1; x_2]$	$\ldots$	$[x_{i-1}; x_i]$	$\ldots$	$[x_{k-1}; x_k]$
Частоты	m_1	m_2	$\ldots$	m_i	$\ldots$	m_k

Требуется на основе имеющейся информации проверить нулевую гипотезу о том, что *гипотетическая функция распределения $F(x)$ значимо представляет данную выборку*, т. е. гипотезу $\{H_0 : F(x) \in \Omega\}$.

При проверке нулевой гипотезы с помощью критерия согласия χ^2 придерживаются следующей последовательности действий :

1) На основании гипотетической функции распределения $F(x)$ вычисляют вероятности попадания случайной величины X в частичные интервалы (разряды) группирования $[x_{i-1};\, x_i]$:

$$p_i \;=\; \Pr\left(x_{i-1} \le \mathrm{X} < x_i\right) \;=\; \int\limits_{x_{i-1}}^{x_i} f(x)\,dx \;=\; F(x_i) - F(x_{i-1}), \tag{5.2}$$

где $i = 1, 2, \ldots, k$.

2) Умножают полученные вероятности p_i на объем выборки n и получают теоретические частоты np_i частичных интервалов $[x_{i-1},\, x_i]$, т.е. частоты, которые следует ожидать, если нулевая гипотеза справедлива.

3) Вычисляют выборочную статистику (критерий) χ^2:

$$\chi^2_{\text{набл}} \;=\; \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i}. \tag{5.3}$$

Можно показать, что если нулевая гипотеза верна, то при $n \to \infty$ закон распределения выборочной статистики (5.3), независимо от вида функции $F(x)$, стремится к закону распределения χ^2 с $\nu = k - r - 1$ степенями свободы (здесь k – число частичных интервалов; r – число параметров гипотетической функции $F(x)$, оцениваемых по данным выборки);

4) По заданному уровню значимости α и числу степеней свободы $\nu = k - r - 1$ находят по таблицам квантилей χ^2-распределения критическое значение $\chi^2_{\alpha;\,\nu}$, удовлетворяющее условию

$$\Pr\left(\chi^2 \ge \chi^2_{\alpha;\,\nu}\right) \;=\; \alpha. \tag{5.4}$$

Критерий χ^2 сконструирован таким образом, что чем ближе к нулю наблюдаемое значение критерия χ^2, тем вероятнее, что нулевая гипотеза справедлива. Поэтому для проверки нулевой гипотезы применяется критерий χ^2 с правосторонней критической областью.

5) Если окажется, что

$$\chi^2_{\text{набл}} \;\ge\; \chi^2_{\alpha;\,\nu}, \tag{5.5a}$$

то рассматриваемая нулевая гипотеза $\{H_0 : F(x) \in \Omega\}$ отвергается в пользу альтернативной $\{H_1 : F(x) \notin \Omega\}$, т. е. считается, что гипотетическая функция не согласуется с опытными данными.

Если же

$$\chi^2_{\text{набл}} \;<\; \chi^2_{\alpha;\,\nu}, \tag{5.5b}$$

то считается, что нет оснований для отклонения нулевой гипотезы, т. е. гипотетическая функция $F(x)$ согласуется с опытными данными.

Замечание. При применении критерия χ^2 необходимо, чтобы в каждом частичном интервале было не менее 5 элементов. Если число элементов (частота) меньше 5, то рекомендуется объединять такие частичные интервалы с соседними.

5.3. Критерий согласия λ Колмогорова

Выше был изложен критерий согласия χ^2, позволяющий проверить гипотезу о согласии данных выборки с конкретным теоретическим законом распределения для любой случайной величины, как непрерывной, так и дискретной.

Критерий согласия λ Колмогорова применяется для проверки гипотез о законах распределения только непрерывных величин.

Его отличие от критерия согласия χ^2 Пирсона состоит в том, что при применении критерия согласия χ^2 сравнивались эмпирические и теоретические частоты распределения; при применении λ-критерия Колмогорова сравниваются эмпирическая $F^*(x)$ и гипотетическая $F(x)$ функции распределения. Кроме того, при применении λ-критерия Колмогорова предполагается, что теоретические значения параметров гипотетической функции известны (в критерии согласия χ^2 они могут определяться по данным выборки). Эти ограничения сужают область практического применения λ-критерия Колмогорова. Тем не менее этот критерий широко применяется на практике.

При его использовании неизвестные теоретические параметры гипотетического распределения оцениваются по данным выборок большого объема, параллельных исследуемой, либо по данным исследуемой выборки. В последнем случае λ-критерий Колмогорова становится приближенным в том смысле, что действительный уровень значимости α приближенно равен заданному уровню α ($\alpha_{\text{факт}} < \alpha_{\text{задан}}$). В случае, когда параметры гипотетического закона распределения оцениваются по данным исследуемой выборки, λ-критерий Колмогорова показывает лучшее согласие с эмпирическими данными, чем критерий согласия χ^2 Пирсона. Поэтому при его применении рекомендуется использовать несколько больший уровень значимости $\alpha = 0{,}10 - 0{,}20$.

Ниже приводятся две вероятностные модели.

В первой из них указаны вероятностные предпосылки и правило проверки нулевой гипотезы о виде функции распределения непрерывной случайной величины с помощью λ-критерия Колмогорова.

Во второй модели даны вероятностные предпосылки и правило проверки нулевой гипотезы о принадлежности двух выборок к одной и той же генеральной совокупности (или две генеральные совокупности имеют одну и ту же функцию распределения). Используемый при этом критерий носит название *λ–критерий Смирнова–Колмогорова*.

Модель 1. Пусть известно, что исследуемая случайная величина X имеет непрерывную функцию распределения $F(x)$. Из генеральной совокупности с функцией распределения $F(x)$ извлечена случайная выборка объема n ($n \geq 50$). На основе имеющейся информации требуется проверить нулевую гипотезу $\{H_0$: *опытные данные согласуются с гипотетической функцией распределения*$\}$.

Проверку нулевой гипотезы с помощью критерия согласия Колмогорова производят по следующей схеме:

1) Располагают результаты наблюдений в возрастающем порядке либо представляют их в виде интервального статистического ряда.

2) Находят эмпирическую функцию распределения

$$F^*(x) \;=\; \frac{n_x}{n}. \tag{5.6}$$

3) Вычисляют, пользуясь гипотетической функцией распределения, значения теоретической функции распределения $F(x)$, соответствующие наблюденным значениям случайной величины X.

4) Находят для каждого текущего значения x_i модуль разности между эмпирической и теоретической функциями распределения, т. е.

$$|F^*(x) - F(x)|. \tag{5.7}$$

5) Вычисляют наблюдаемое значение выборочной λ–статистики:

$$\lambda \;=\; D\sqrt{n}, \tag{5.8}$$

где

$$D \;=\; \max_x |F^*(x) - F(x)|. \tag{5.9}$$

Академик А. Н. Колмогоров показал, что если нулевая гипотеза верна, то выборочная статистика $\lambda = D\sqrt{n}$ при $n \to \infty$ имеет функцию распределения, которую принято обозначать $K(\lambda)$, следующего вида:

$$K(\lambda) \;=\; \Pr(D\sqrt{n} < \lambda) \;=\; \sum_{k=-\infty}^{\infty} (-1)^k \exp\left(-2k^2\lambda^2\right). \tag{5.10}$$

Зададим уровень значимости α. Тогда из соотношения

$$\Pr(\lambda \geq \lambda_\alpha) \;=\; \Pr(D\sqrt{n} \geq \lambda_\alpha) \;=\; 1 - \sum_{k=-\infty}^{\infty} (-1)^k \exp\left(-2k^2\lambda_\alpha^2\right) \;=\; \alpha \tag{5.11}$$

можно найти квантили λ–распределения Колмогорова.

Таблица квантилей (критических значений) распределения Колмогорова приведена в приложении.

Сравним наблюдаемое значение выборочной статистики $\lambda_{\text{набл}} = D\sqrt{n}$ с критическим значением λ_α, определяемым по таблицам квантилей распределения Колмогорова по заданному уровню значимости α.

Если при этом окажется, что

$$D\sqrt{n} \;\geq\; \lambda_\alpha, \tag{5.12a}$$

то проверяемая гипотеза отклоняется.

Если же

$$D\sqrt{n} \;<\; \lambda_\alpha, \tag{5.12b}$$

то считается, что нет оснований для отклонения нулевой гипотезы, т. е. гипотетическая функция распределения считается согласующейся с опытными данными.

Модель 2. Пусть в результате наблюдений получены две случайные выборки объема n_1 и n_2 ($n_1 \geq 50$ и $n_2 \geq 50$). Требуется на основании имеющейся

информации проверить нулевую гипотезу $\{H_0:$ *две выборки извлечены из одной и той же генеральной совокупности с гипотетической функцией распределения* $F(x)\}$.

Проверка нулевой гипотезы основывается на вычислении выборочной статистики λ-критерия Смирнова-Колмогорова:

$$\lambda \;=\; D^*\sqrt{n} \;=\; \max_x |F_1^*(x) - F_2^*(x)|\,\sqrt{n}, \tag{5.13}$$

где $n \;=\; n_1 n_2/(n_1+n_2)$, а $F_1^*(x)$ и $F_2^*(x)$ – эмпирические функции распределения, построенные по данным первой и второй выборок соответственно.

Если нулевая гипотеза верна, то при $n \;\to\; \infty$ распределение выборочной статистики λ асимптотически сходится к распределению Смирнова-Колмогорова *независимо от вида функции* $F(x)$.

По таблице квантилей распределения Смирнова-Колмогорова по заданному уровню значимости α находят критические значения λ_α (см. приложение), удовлетворяющие условию

$$\Pr\,(\lambda \geq \lambda_\alpha) \;=\; \alpha. \tag{5.14}$$

Если окажется, что

$$D^*\sqrt{n} \;\geq\; \lambda_\alpha, \tag{5.15a}$$

то нулевая гипотеза отклоняется.

Если же

$$D^*\sqrt{n} \;<\; \lambda_\alpha, \tag{5.15b}$$

то считается, что нет оснований для отклонения гипотезы о том, что две исследуемые совокупности имеют одну и ту же функцию распределения.

5.4. Критерий знаков

На практике часто приходится иметь дело со случайными величинами, закон распределения которых неизвестен. В этом случае вместо параметрических критериев значимости можно использовать *непараметрические критерии значимости*, при применении которых не требуется делать каких-либо предположений о законе распределения исследуемой случайной величины.

Непараметрические критерии по сравнению с параметрическими имеют несколько меньшую мощность, т. е. они менее эффективны. Однако этот недостаток компенсируется более простым построением выборочных статистик, на вычисление затрачивается гораздо меньше времени.

Критерий знаков – один из непараметрических критериев, с помощью которого проверяется нулевая гипотеза о том, что две выборки извлечены из одной и той же генеральной совокупности. Его главное преимущество состоит в том, что при применении не нужно ставить никаких ограничений относительно вида функции распределения, кроме её непрерывности.

Критерий знаков применяется как критерий сравнения "спаренных" наблюдений. Обычно сравниваются результаты двух выборок одинакового объема. Критерий

знаков, например, часто используют для проверки устойчивости генеральной совокупности во времени, т. е. для сравнения двух спаренных наблюдений, соответствующих двум моментам времени. Его можно применять при обработке экспериментов для сравнения измеряемых величин в экспериментальной и контрольной группе.

Модель. Пусть имеются две случайные выборки $(x_1, x_2, \ldots, x_n)$ и $(y_1, y_2, \ldots, y_n)$ одинакового объема n. Требуется проверить нулевую гипотезу о том, что они извлечены из одной и той же генеральной совокупности.

Для проверки данной нулевой гипотезы с помощью критерия знаков исследуют знаки разностей спаренных результатов обеих выборок и находят число тех знаков, которых меньше. Обозначим их число r.

Если нулевая гипотеза верна, то

$$\Pr(x - y > 0) \;=\; \Pr(x - y < 0) \;=\; \frac{1}{2}, \tag{5.16}$$

число r является дискретной случайной величиной, распределенной по биномиальному закону с параметром $p = \frac{1}{2}$, т. е.

$$\mathrm{P}_n(r) \;=\; C_n^r \left(\frac{1}{2}\right)^n. \tag{5.17}$$

Пусть теперь r_α – наименьшее значение r, для которого выполняется неравенство $\mathrm{P}_n(r) \le \alpha$. Таблица критических значений числа знаков r_α, соответствующих заданному уровню значимости α и объему выборки n, приведена в приложении.

Если

$$r_{\text{набл}} \;\le\; r_\alpha, \tag{5.18a}$$

то нулевая гипотеза отклоняется, т. е. считается, что выборки извлечены из генеральных совокупностей с различными функциями распределения.

Если же

$$r_{\text{набл}} \;>\; r_\alpha, \tag{5.18b}$$

то считается, что нет оснований для отклонения нулевой гипотезы о том, что две выборки извлечены из одной и той же генеральной совокупности.

5.5. Методические указания по применению критериев согласия

Критерии согласия, как и все непараметрические критерии, являются статистическими критериями значимости. Это означает, что с их помощью нулевая гипотеза о виде функции распределения либо отклоняется, либо считается, что имеющаяся информация не дает повода для отклонения выдвинутой гипотезы о виде функции распределения.

Если объем выборок n невелик ($n \le 50$) или результаты измерений располагаются в достаточно узком интервале изменений случайной величины X, то экспериментальные данные могут достаточно хорошо согласовываться с рядом различных вероятностных моделей, т. е. с различными законами распределения.

Поэтому не следует придавать слишком большого значения положительному результату (*"Нулевая гипотеза не отклоняется"*) проверки нулевых гипотез о виде функции распределения с помощью функции согласия.

Здесь еще раз подчеркнем, что никакими обработками экспериментального материала нельзя извлечь из него информации больше, чем этот материал содержит.

Есть только один реальный способ уточнить сведения об исследуемом объекте – это увеличивать объем выборки экспериментальных данных.

На практике все шире применяют критерии согласия не столько для проверки согласия экспериментальных данных с некоторой гипотетической функцией распределения, сколько для подбора наилучшей функции распределения (вероятностной модели) из ряда рассматриваемых функций (моделей).

Выбор подходящего закона распределения должен базироваться прежде всего на понимании механизма изучаемого явления. Однако если механизм изучаемого явления неизвестен, то предварительный выбор закона распределения может быть сделан исходя из следующих соображений:

1) По виду гистограммы частостей статистического ряда распределения. Вид гистограммы дает ориентировку на возможный закон распределения. Например, если гистограмма имеет многомодовый (многогорбый) вид, такое её свойство, возможно, следует объяснить смешением разнородных по своим качествам объектов наблюдения.

Достоинства – простота применения, наглядность. Недостатки – гистограмма может одновременно напоминать несколько законов распределения. Например, по гистограмме практически нельзя различить логарифмически нормальный закон и закон распределения Вейбулла даже при большом объеме выборки.

2) С помощью графического представления эмпирической функции распределения на вероятностных бумагах (особенно это удобно реализовать с помощью ЭВМ).

Если закон распределения выбран правильно, то при нанесении эмпирической функции распределения на вероятностную бумагу со шкалами, соответствующими гипотетическому закону распределения, эти значения будут располагаться на прямой линии либо вблизи прямой линии.

Достоинства метода – простота, наглядность. Недостатки – необходимо иметь специальные бумаги (или математические пакеты, позволяющие выполнить необходимые расчеты); отсутствие количественного критерия возможного отклонения значений эмпирической функции распределения от прямой линии; неоднозначность выбора закона, вызванная тем, что иногда на таких вероятностных бумагах со шкалами, соответствующими различным законам распределения, значения эмпирической функции распределения располагаются примерно по прямой линии.

Закон распределения исследуемой случайной величины X	Пределы изменения	Среднее значение
Нормальный закон	[0,08; 0,40]	0,25
Закон Вейбулла	[0,40; 0,85]	0,71
Логарифмический закон	[0,35; 0,80]	0,68
Экспоненциальный закон	[0,60; 1,30]	0,92

3) Предварительный выбор закона может производиться по величине эмпири-

ческого коэффициента вариации $Var = s/\overline{x}$. Известно, что каждому закону распределения соответствует определенный приближенный диапазон значений коэффициента вариации (см. сводку–таблицу).

Достоинство метода – простота. Недостатки – коэффициент вариации не отражает степень симметрии эмпирической кривой распределения; неоднозначность выбора.

4) По опытным данным ранее проведенных исследований.

Недостаток метода – могут быть значительные расхождения в механизмах изучаемых явлений, описываемых случайными величинами, т. е. в конструкциях, технологиях изготовления, условиях эксплуатации изделий, отличных от ранее описанных.

5) В качестве приближенного критерия для предварительного выбора закона распределения могут быть использованы выборочные коэффициенты асимметрии и эксцесса.

Если по данным выборки объема n найдены точечные оценки асимметрии и эксцесса

$$\hat{A} = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \overline{x})^3, \qquad \hat{E} = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \overline{x})^4 - 3 \tag{5.19}$$

и их средних арифметических отклонений

$$s_{_A} = \left(\frac{6(n-1)}{(n+1)(n+3)}\right)^{1/2}, \qquad s_{_E} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+4)}}, \tag{5.20}$$

то эмпирическая функция считается согласующейся с гипотетической функцией при условии, что выборочные коэффициенты асимметрии и эксцесса отличаются по абсолютной величине от своих математических ожиданий не более чем на утроенные средние квадратические отклонения.

Таким образом, если

$$|\hat{A} - m_{_A}| < 3\sigma_{_A} \qquad \text{и} \qquad |\hat{E} - m_{_E}| < 3, \tag{5.21}$$

то считается, что нулевая гипотеза согласуется с экспериментальными данными.

Если хотя бы одно из этих неравенств не выполняется, то выдвинутая нулевая гипотеза отклоняется.

Замечание. Для нормального закона распределения математические ожидания выборочных коэффициентов асимметрии и эксцесса равны нулю. Поэтому гипотеза нормальности принимается, если $|\hat{A}| < 3\sigma_{_A}$ и $|\hat{E}| < 3$.

Достоинство метода – учет симметрии и крутости, т.е. формы кривой. Недостаток – нет строгой количественной оценки допустимого расхождения между выборочными коэффициентами асимметрии и эксцесса и их математическими ожиданиями, так как правило "трех сигм" является эмпирическим.

После предварительного выбора закона распределения по одному или нескольким перечисленным методам рекомендуется применять строгие критерии согласия.

Здесь следует учесть, что кроме изложенных в данном пособии основных критериев согласия χ^2 и критерия λ Колмогорова, имеется ряд других специфических

критериев согласия, например, критерий ω^2 Мизеса-Смирнова, который в противоположность критерию χ^2 не требует объединения числовых данных в разряды, т. е. более полно использует информацию, содержащуюся в выборке. Также имеется, например, специальный критерий для проверки гипотез нормальности по совокупности достаточно большого числа выборок малого объема.

5.6. Развернутый пример обработки данных для нормального закона распределения

Приведем пример развернутого решения, в котором используется нормальный закон распределения и применяются основные понятия математической статистики, основанные на применении нормального закона распределения.

Пример

При сверлении 80 отверстий одним и тем же сверлом и последующим измерением диаметров отверстий получены следующие данные (в *мм*):

40,26	40,35	40,44	40,35	40,39	40,40	40,42	40,32	40,37	40,35
40,44	40,35	40,30	40,34	40,31	40,32	40,33	40,41	40,35	40,30
40,33	40,38	40,33	40,33	40,28	40,30	40,40	40,36	40,32	40,32
40,42	40,35	40,29	40,33	40,31	40,33	40,36	40,34	40,30	40,30
40,41	40,40	40,33	40,37	40,34	40,30	40,43	40,34	40,35	40,34
40,34	40,31	40,43	40,36	40,34	40,34	40,28	40,44	40,32	40,34
40,31	40,31	40,36	40,34	40,29	40,39	40,39	40,37	40,37	40,38
40,36	40,41	40,27	40,38	40,37	40,37	40,36	40,35	40,32	40,36

Требуется:

1. Составить интервальные статистические ряды распределения частот и частостей наблюденных значений непрерывной случайной величины X – диаметров отверстий x_i.

2. Построить гистограмму и полигон частостей диаметров отверстий.

3. Найти эмпирическую функцию распределения $F^*(x)$ и построить ее график.

4. Вычислить числовые характеристики выборки:

– среднее арифметическое;

– выборочную дисперсию;

– выборочное среднее квадратическое отклонение;

– выборочный коэффициент асимметрии;

– выборочный коэффициент эксцесса;

– выборочный коэффициент вариации.

5. По виду гистограммы и полигона частостей, а также по значениям выборочных коэффициентов асимметрии и эксцесса, и, исходя из механизма образования исследуемой случайной величины X, сделать предварительный выбор вида закона распределения этой случайной величины.

6. Найти точечные оценки параметров нормального закона распределения

(предполагается, что исследуемая случайная величина распределена по нормальному закону), записать плотность вероятности и функцию распределения случайной величины X.

7. Найти теоретические частоты нормального закона распределения, проверить согласие эмпирической функции распределения с нормальным законом с помощью основных критериев согласия — критерия χ^2 Пирсона и λ–критерия Колмогорова.

8. Найти интервальные оценки параметров нормального закона распределения. Доверительную вероятность принять $\mathrm{P}= 1-\alpha = 0,95$.

Решение

1. Изучение непрерывных случайных величин начинается с группировки статистического материала, т. е. с разбиения интервала наблюденных значений случайной величины X на k частичных интервалов равной длины и подсчета частот попадания наблюденных значений СВ X в частичные интервалы группирования. Количество интервалов выбирается произвольно. Обычно число интервалов бывает не менее 5 и не более 15.

Разобьем весь диапазон наблюденных значений на 5 интервалов (разрядов).

Длину частичного интервала определим по формуле

$$h \;=\; \frac{1}{5}\,(x_{\max} - x_{\min}) \;=\; \frac{1}{5}\,(40,44 - 40,26) \;\approx\; 0,04.$$

Границы интервалов выберем следующим образом. За начало первого интервала принимаем величину a_0, равную $a_0 = 40,26 - 0,02 = 40,24$, тогда первый интервал будет $[40,24;\ 40,28]$, второй – $[40,28;\ 40,32]$ и т.д. Шкала интервалов и группировка исходных статистических данных сведены в таблицу.

В результате получаем статистический ряд распределения частот:

Интервалы наблюденных значений СВ X	[40,24; 40,28]	[40,28; 40,32]	[40,32; 40,36]	[40,36; 40,40]	[40,40; 40,44]
Частота m_i	4	19	32	15	10

Контроль: $n = \sum\limits_{i=1}^{5} m_i \;=\; 80.$

Для получения статистического ряда частостей разделим частоты m_i на объем выборки n. В результате получаем интервальный статистический ряд распределений частостей.

Интервалы наблюденных значений СВ X	[40,24; 40,28]	[40,28; 40,32]	[40,32; 40,36]	[40,36; 40,40]	[40,40; 40,44]
Частости m_i/n	0,0500	0,2379	0,4000	0,1875	0,1250
Накопленные частости $F^*(x)$	0,0500	0,2875	0,6875	0,8750	1,0000

Контроль: $\sum\limits_{i=1}^{5} m_i/n = \;1.$

2. Для построения гистограммы частостей на оси $0x$ откладываются частичные интервалы, на каждом из них строится прямоугольник, площадь которого равна частости данного частичного интервала. Если частости отнести к серединам частичных интервалов, то полученная замкнутая линия образует *полигон* частостей. На рис. 5.1 изображена гистограмма и полигон частостей.

3. Значения эмпирической функции распределения выписаны в последней строке статистического ряда распределения частостей. Запишем значения эмпирической функции распределения в аналитическом виде:

$$F^*(x) = \begin{cases} 0,0000, & \text{если} \quad x \le 40,24; \\ 0,0500, & \text{если} \quad 40{,}24 < x \le 40,28; \\ 0,2875, & \text{если} \quad 40{,}28 < x \le 40,32; \\ 0,6875, & \text{если} \quad 40{,}32 < x \le 40,36; \\ 0,8750, & \text{если} \quad 40{,}36 < x \le 40,40; \\ 1,0000, & \text{если} \quad 40{,}40 < x \le 40,44; \\ 1,0000, & \text{если} \quad x > 40,44. \end{cases}$$

Замечание. Значения эмпирической функции распределения отнесены к верхней границе частичного интервала. График эмпирической функции изображен на рис. 5.2.

4. В тех случаях, когда наблюденные значения случайной величины задаются многозначными числами и объем выборки достаточно велик ($n > 25$), вычисления достаточно объемны. Поэтому вначале обозначим x_i – середина i-го интервала ($i = 1 \div 5$) и найдем среднюю арифметическую $\overline{x}$ по формуле

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{5} x_i m_i.$$

Обозначим $q_{1i} = (x_i - \overline{x})\, m_i$, $q_{2i} = (x_i - \overline{x})^2\, m_i$, $q_{3i} = (x_i - \overline{x})^3\, m_i$, $q_{4i} = (x_i - \overline{x})^4\, m_i$ и затем перейдем к вычислению центральных моментов порядка k ($k = 2, 3, 4$):

i	Интервалы значений СВ X	x_i	m_i	q_{1i}	q_{2i}	q_{3i}	q_{4i}
1	40,24 – 40,28	40,26	4	–0,3360	0,0282	–0,0024	0,0002
2	40,28 – 40,32	40,30	19	–0,8360	0,0368	–0,0016	0,0001
3	40,32 – 40,36	40,34	32	–0,1280	0,0005	–0,0000	0,0000
4	40,36 – 40,40	40,38	15	+0,5400	0,0194	+0,0007	0,0000
5	40,40 – 40,44	40,42	10	+0,7600	0,0578	+0,0044	0,0003
	Сумма		80	0,0000	0,1427	+0,0011	0,0006

Следовательно,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{5} m_i x_i =$$

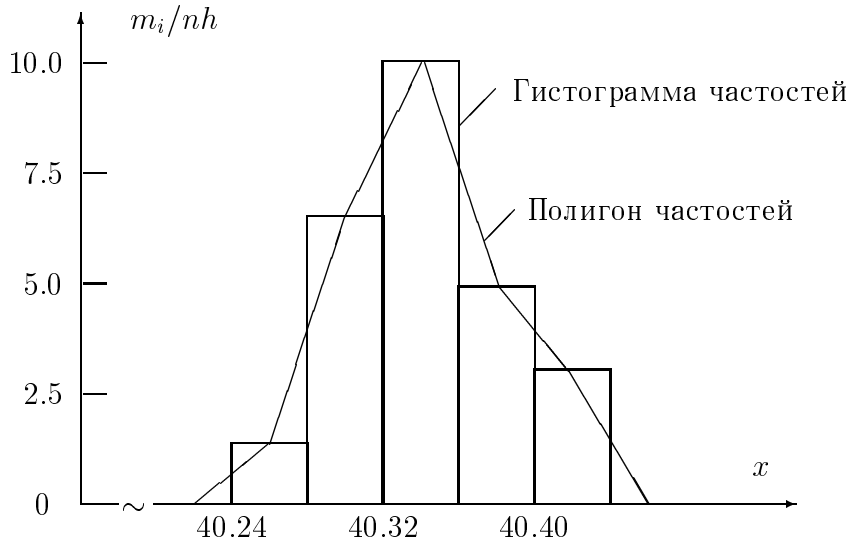


Рисунок 5.1 — Гистограмма частостей и полигон частостей

$$= \frac{40,26 \cdot 4 + 40,30 \cdot 19 + 40,34 \cdot 32 + 40,38 \cdot 15 + 40,42 \cdot 10}{80} = 40,344;$$

$$\hat{\mathrm{D}}[\mathrm{X}] = s_x^2 = \frac{1}{n}\sum_{i=1}^{5}(x_i - \overline{x})^2 m_i = \frac{0,1427}{80} = 0,001784;$$

$$s_x = \sqrt{0,001784} = 0,0422;$$

$$\hat{A} = \frac{1}{n s_x^3}\sum_{i=1}^{5}(x_i - \overline{x})^3 m_i = \frac{0,001}{80 \cdot 0,042^3} = 0,1822;$$

$$\hat{E} = \frac{1}{n s_x^4}\sum_{i=1}^{5}(x_i - \overline{x})^4 m_i - 3 = \frac{0,0006}{80 \cdot 0,042^4} - 3 = -0,5288;$$

$$\mathrm{Var} = \frac{s_x}{\overline{x}}\,100\,\% = \frac{0,0422}{40,344}\,100\,\% = 0,1047\,\%.$$

5. Для предварительного выбора искомого закона распределения вычислим вначале средние квадратические ошибки определения асимметрии

$$s_{_A} = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}} = \sqrt{\frac{6 \cdot 79}{81 \cdot 83}} = 0,2655$$

и эксцесса

$$s_{_E} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+4)}} = \sqrt{\frac{24 \cdot 80 \cdot 78 \cdot 77}{79^2 \cdot 83 \cdot 84}} = 0,5148.$$

Критерием распределения диаметров отверстий по нормальному закону является равенство нулю асимметрии и эксцесса. Из приведенных расчетов видно, что выборочные коэффициенты асимметрии $\hat{A}$ и эксцесса $\hat{E}$ отличаются от нуля не более

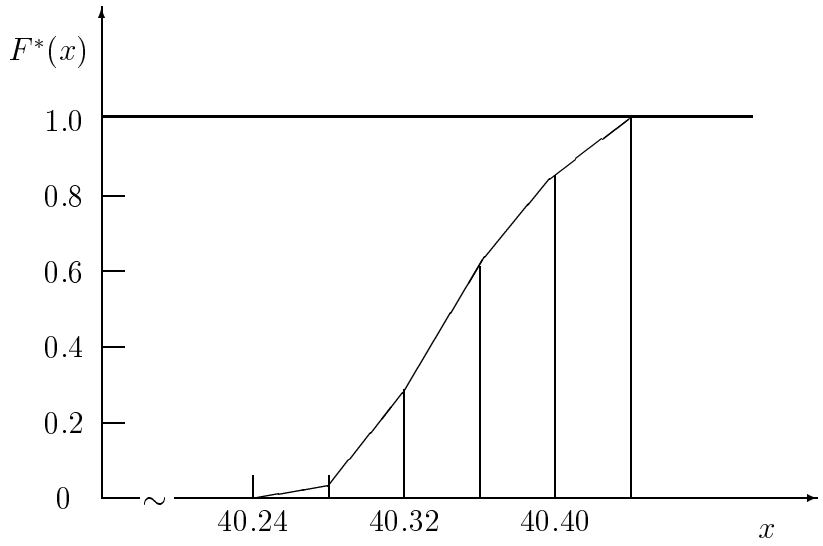


Рисунок 5.2 — Эмпирическая функция распределения (кумулята)

чем на удвоенные средние квадратические ошибки их определения, что соответствует нормальному закону распределения.

Вид полигона и гистограммы частостей также напоминает нормальную кривую (кривую Гаусса).

Можно предположить, что диаметр отверстия (СВ X) изменяется под влиянием большого числа факторов, примерно равнозначных по силе (изменение температуры сверла или заготовки, вибрации заготовки, вибрации сверла, изменение механических или химических свойств заготовки и т. д.). Поэтому, исходя из "технологии" формирования СВ X, т.е. механизма образования отклонений диаметров отверстий от некоторого номинального значения, можно предположить, что закон распределения диаметров отверстий является нормальным.

6. Плотность вероятности нормального закона имеет вид

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right).$$

Найдем точечные оценки $\hat{a}$ и $\hat{\sigma}$ параметров a и σ нормального закона распределения методом моментов:

$$\hat{a} \;=\; \overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{5} x_i m_i \;=\; 40,344\,(\text{мм});$$

$$\hat{\sigma} \;=\; s \;=\; \sqrt{\frac{1}{n}\sum_{i=1}^{5}(x_i-\overline{x})^2 m_i} \;=\; 0,0422\,(\text{мм}).$$

Следовательно, плотность вероятности предполагаемого нормального закона распределения имеет вид

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}\,0,0422}\,\exp\left(-\frac{(x-40,344)^2}{2\cdot 0,0422^2}\right).$$

Функция распределения предполагаемого нормального закона

$$F(x) \;=\; \frac{1}{\sqrt{2\,\pi}\,0,0422}\int\limits_{-\infty}^{x}\exp\left(-\frac{(x'-40,344)^2}{0,003567}\right)\,dx'.$$

Используя функцию Лапласа $\Phi(x)$, функцию распределения нормального закона можно записать в виде

$$F(x) \;=\; \frac{1}{2}+\Phi\left(\frac{x-40,344}{0,0422}\right).$$

7. Проведем детальную проверку гипотезы о распределении СВ X (диаметра отверстий) по нормальному закону с помощью критерия согласия χ^2.

Для этого интервалы наблюденных значений отнормируем, т.е. выразим их в единицах среднего квадратического отклонения s:

$$u_i \;=\; \frac{x_i-\overline{x}}{s},$$

причем наименьшее значение u_i примем равным $-\infty$, а наибольшее равным $+\infty$.

Далее вычислим вероятности попадания СВ X, распределенной по нормальному закону с параметрами $a=40,344$ и $\sigma=0,042$, в частичные интервалы $[x_{i-1};\,x_i]$ по формуле

$$p_i \;=\; \Pr\left(x_{i-1}<\mathrm{X}<\,x_i\right) \;=\; \Phi(u_i)-\Phi(u_{i-1}),$$

где $\Phi(z)$ – функция Лапласа

$$\Phi(z) \;=\; \frac{1}{\sqrt{2\pi}}\int_0^z \exp(-t^2/2)\,dt.$$

Например, вероятность того, что СВ X (диаметр отверстий) попадет в первый частичный интервал $(-\infty;\,40,28]$, равна

$$p_1 \;=\; \Pr(-\infty<\mathrm{X}<40,28) \;=$$

$$=\; \Phi\left(\frac{40,28-40,344}{0,0422}\right)-\Phi\left(\frac{-\infty-40,344}{0,0422}\right) \;=$$

$$=\; \Phi(\infty)-\Phi(1,5166) \;=\; \frac{1}{2}-0,4352 \;=\; 0,0648.$$

Аналогично

$$p_2 \;=\; \Pr(40,28<\mathrm{X}<40,32) \;=$$

$$=\; \Phi\left(\frac{40,32-40,344}{0,042}\right)-\Phi\left(\frac{40,28-40,344}{0,042}\right) \;=$$

$$=\; \Phi(-0,5685)-\Phi(-1,5155) \;=\; 0,4351-0,2151 \;=\; 0,2200$$

и так далее для p_3, p_4, p_5.

После этого вычислим теоретические частоты нормального закона распределения $n_{\text{теор}}=np_i$ и наблюдаемое значение критерия χ^2

$$\chi^2_{\text{набл}} \;=\; \sum_{i=1}^{5}\frac{(m_i-np_i)^2}{np_i}.$$

Затем по таблицам квантилей χ^2-распределения по уровню значимости $\alpha = 0,05$ и числу степеней свободы $\nu = k - r - 1$ ($k = 5$ – число интервалов, $r = 2$ – число параметров предполагаемого закона распределения СВ X) найдем (см. приложение) критическое значение $\chi^2_{0,05;\nu}$.

Если $\chi^2_{\text{набл}} \le \chi^2_{0,05;\nu}$, то примем, что нет оснований для отклонения гипотезы о распределении диаметров отверстий по нормальному закону.

В противном случае, т. е. если $\chi^2_{\text{набл}} > \chi^2_{0,05;\nu}$, считается, что гипотеза распределения диаметров отверстий по нормальному закону не согласуется с экспериментальными данными.

Вычисления, необходимые для определения наблюдаемого значения $\chi^2_{\text{набл}}$ выборочной статистики χ^2, сведены в таблицу, в которой обозначено: $Q_{2i} = (m_i - np_i)^2$ и $R_{2i} = (m_i - np_i)^2/(np_i)$.

Замечание. При построении таблицы наименьшее значение стандартизованной переменной $(40,24-40,344)/0,042 = -2,48$ заменено на $-\infty$, а наибольшее значение $(40,44 - 40,344)/0,042 = +2,29$ заменено на ∞. Эта замена произведена для того, чтобы сумма теоретических частот np_i была равна объему выборки.

i	Интервалы наблюденных значений X	Частоты m_i	Нормированные интервалы $u_i; u_{i+1}$	p_i	np_i	Q_{2i}	R_{2i}
1	40,24–40,28	4	$-\infty; -1,5152$	0,0648	5,1860	1,4067	0,2712
2	40,28–40,32	19	$-1,5152; -0,5682$	0,2200	17,6031	1,9514	0,1109
3	40,32–40,36	32	$-0,5682; 0,3788$	0,3626	29,0107	8,9356	0,3080
4	40,36–40,40	15	$0,3788; 1,3258$	0,2600	20,8013	33,6550	1,6179
5	40,40–40,44	10	$1,3258; \infty$	0,0925	7,3989	6,7659	0,9144
	Сумма	80		1,0000	80,0000		3,2225

Итак, в результате вычислений получим

$$\chi^2_{\text{набл}} = \sum_{i=1}^{5} R_{2i} = 3,2225.$$

Найдем теперь по таблице квантилей χ^2-распределения по уровню значимости $\alpha = 0,05$ и числу степеней свободы $\nu = k-r-1 = 5-2-1 = 2$ критическое значение $\chi^2_{0,05;2} = 5,99$ (см. приложение).

Так как $\chi^2_{\text{набл}} = 3,2225 < 5,99$, то нет оснований для отклонения гипотезы о нормальном законе распределения диаметров отверстий.

Далее проверим гипотезу распределения диаметров отверстий по нормальному закону с помощью λ-критерия Колмогорова. С этой целью для каждого значения x_i найдем модуль разности между эмпирической и теоретической функциями распределений $|F^*(x) - F(x)|$ и вычислим наблюдаемое значение выборочной статистики λ Колмогорова:

$$\lambda_{\text{набл}} = D\sqrt{n} = \max_x |F^*(x) - F(x)| \sqrt{n}.$$

Наблюдаемое значение статистики λ Колмогорова сравним с критическим значением, определяемым по уровню значимости $\alpha = 0,05$ (см. приложение).

Если $\lambda_{\text{набл}} \le \lambda_{0,05}$, то считается, что гипотеза нормального распределения исследуемой случайной величины согласуется с экспериментальными данными, если же $\lambda_{\text{набл}} > \lambda_{0,05}$, – не согласуется с экспериментальными данными.

Пользуясь λ-критерием согласия Колмогорова, проверим гипотезу нормального распределения диаметров отверстий. Все вспомогательные расчеты, которые необходимы для вычисления выборочной статистики $\lambda = D\sqrt{n}$, сведем в таблицу, в которой обозначено $m_{\text{н.э.ч}}$ – накопленные эмпирические частоты; $p_{\text{н.в}}$ – вероятности; $D(x) = |F^*(x) - F(x)|$ – текущий модуль разности.

i	Интервалы значений СВ X	Частоты m_i	$m_{\text{н.э.ч}}$	$p_{\text{н.в}}$	$F^*(x)$	$F(x)$	$D(x)$
1	40,24–40,28	4	4	0,064	0,0500	0,0648	0,0148
2	40,28–40,32	19	23	0,220	0,2875	0,2849	0,0026
3	40,32–40,36	32	55	0,364	0,6875	0,6475	0,0400
4	40,36–40,40	15	70	0,260	0,8750	0,9075	0,0325
5	40,40–40,44	10	80	0,092	1,0000	1,0000	0,0000
	Сумма	80					

Просматривая последний столбец таблицы, замечаем, что наибольший модуль разности между эмпирической и теоретической функциями распределения составляет

$$D = \max_x |F^*(x) - F(x)| = 0,040.$$

Вычислим наблюдаемое значение выборочной статистики λ Колмогорова:

$$\lambda_{\text{набл}} = D\sqrt{n} = 0,040\sqrt{80} = 0,358.$$

Примем уровень значимости $\alpha = 0,05$. По таблицам квантилей λ-распределения Колмогорова (см. приложение) по уровню значимости $\alpha = 0,05$ находим величину – критическое значение $\lambda_{0,05} = 1,358$. Так как $\lambda_{\text{набл}} = 0,358 < 1,358$, то нет основания для отклонения гипотезы о нормальном законе распределения диаметров отверстий.

Для построения нормальной кривой из середин частичных интервалов восстановим перпендикуляры высотой p_i/h (p_i – вероятность попадания СВ X в частичный интервал, h – длина интервала). На рис. 5.3 концы этих перпендикуляров отмечены кружками. Полученные точки соединены плавной кривой.

Сравнение гистограммы и нормальной кривой наглядно показывает, что нормальная кривая хорошо сглаживает гистограмму относительных частот.

8. Найдем интервальные оценки параметров нормального закона распределения.

Для нахождения доверительного интервала, который содержит математическое ожидание диаметров отверстия (X), найдем по таблицам квантилей распределения Стьюдента по заданной доверительной вероятности P$= 1-\alpha = 0,95$ и числу степеней свободы $\nu = n - 1 = 80 - 1 = 79$ квантиль $t_{\alpha/2;\,\nu} = t_{0,025;\,79} = 1,99$.

Вычислим предельную погрешность интервального оценивания

$$\varepsilon = t_{\alpha/2;\,\nu}\frac{s}{\sqrt{n}} = 1,99 \cdot \frac{0,042}{\sqrt{80}} = 0,009.$$

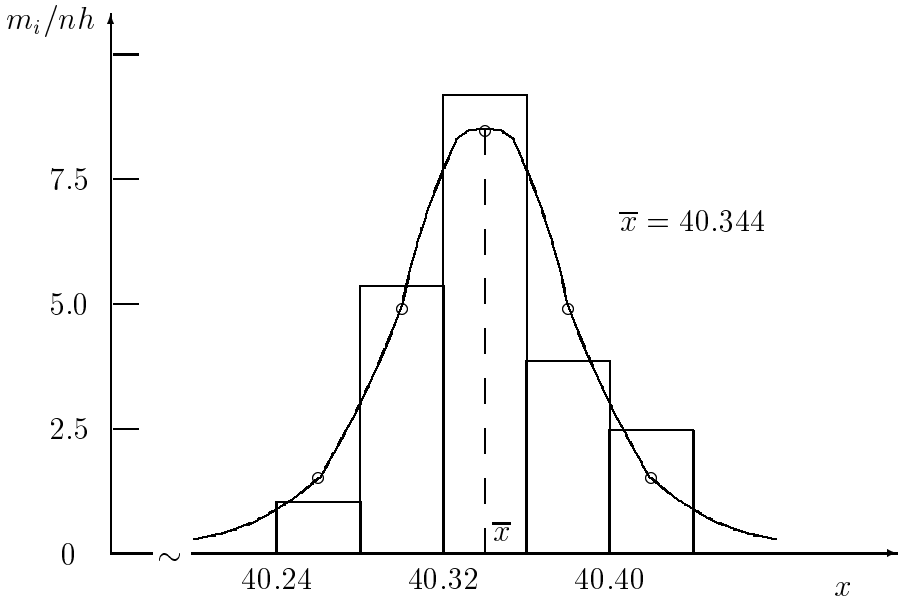


Рисунок 5.3 — Гистограмма и нормальная кривая

На основании формулы $\overline{x} - \varepsilon < a < \overline{x} + \varepsilon$ найдем, что искомый доверительный интервал для математического ожидания равен

$$40,344 - 0,009 < a < 40,344 + 0,009.$$

Итак, получаем доверительный интервал для математического ожидания

$$40,335 < a < 40,353.$$

Смысл полученного результата: если будет произведено достаточно большое число выборок по 80 сверлений отверстий, то в 95 % из них доверительный интервал накроет математическое ожидание диаметра отверстия и только в 5 % случаев математическое ожидание может выйти за границы доверительного интервала.

Для нахождения доверительного интервала, содержащего неизвестное среднее квадратическое отклонение σ с заданной вероятностью $\mathrm{P} = 1 - \alpha = 0,95$, найдем по таблицам квантилей распределения Стьюдента по доверительной вероятности $\mathrm{P} = 1 - \alpha = 0,95$ и числу степеней свободы $\nu = n - 1 = 80 - 1 = 79$ два числа: $\gamma_1 = 0,87$ и $\gamma_2 = 1,18$.

На основании формулы $\gamma_1 s < \sigma < \gamma_2 s$ найдем, что искомый доверительный интервал равен

$$0,87 \cdot 0,042 < \sigma < 1,18 \cdot 0,042.$$

Итак, получаем доверительный интервал для среднего квадратического отклонения

$$0,037 < \sigma < 0,050.$$

Полученный результат означает, что если будет произведено достаточно большое число выборок по 80 сверлений отверстий, то в 95 % из них доверительный интервал накроет среднее квадратическое отклонение σ и только в 5 % среднее квадратическое отклонение σ может выйти за границы доверительного интервала.

5.7. Примеры

Пример 5.1

Имеются данные о распределении толщины X 12000 бобов (в *мм*):

№ интервала	1	2	3	4	5	6	7	8
частота	32	103	239	624	1187	1650	1883	1930
№ интервала	9	10	11	12	13	14	15	16
частота	1638	1130	737	427	221	110	57	32

Здесь первый интервал — значения X, меньшие 7,00 *мм*, второй — 7,00–7,25, третий — 7,25–7,50 и т.д.

Требуется проверить, согласуется ли толщина бобов в выборке с предположением, что этот признак в генеральной совокупности распределен по нормальному закону?

Решение

Так как в исходных данных к примеру отсутствуют сведения о параметрах нормального распределения, то в качестве параметра $\overline{\mathrm{X}}$ используем $\tilde{\mathrm{X}}$, значение которого в данном случае равно 8,562, а за генеральную дисперсию – её несмещенную оценку

$$s^2 \;=\; \frac{n}{n-1}\,\sigma^2 \;=\; 0,3833,$$

откуда $s = 0{,}6191$.

Находим величины np_i для каждого из интервалов.

Таким образом, получим следующий ряд теоретических частот:

$$69{,}9;\;\; 134{,}8;\;\; 313{,}5;\;\; 620{,}8;\;\; 1046{,}6;\;\; 1502{,}3;\;\; 1836{,}0;\;\; 1910{,}3;$$
$$1692{,}3;\;\; 1276{,}3;\;\; 819{,}6;\;\; 448{,}1;\;\; 208{,}6;\;\; 82{,}6;\;\; 27{,}8;\;\; 10{,}5.$$

Покажем, как производится расчет на примере второго интервала:

$$np_2 \;=\; 12000\cdot \Pr(7{,}00 \le \mathrm{X} \le 7{,}25) \;=$$

$$=\; 12000\cdot\left[\Phi\left(\frac{7{,}25-8{,}562}{0{,}619}\right) - \Phi\left(\frac{7{,}00-8{,}562}{0{,}619}\right)\right] \;=\; 134{,}8.$$

Поэтому

$$\chi_0^2 \;=\; \frac{(32-69{,}9)^2}{49{,}3} + \frac{(103-134{,}8)^2}{134{,}8} + \cdots + \frac{(57-27{,}8)^2}{27{,}8} + \frac{(32-10{,}5)^2}{8{,}1},$$

что дает $\chi_0^2 \;=\; 192{,}99$.

Так как два параметра генерального распределения были найдены по выборке, то $s = 2$, а число степеней свободы $k-s-1 = 16-2-1 = 13$. При $\alpha = 0{,}01$ границей критической области является $\chi^2 = 27{,}688$.

Так как $\chi_0^2 = 192{,}99 > 27{,}688$, то гипотеза о том, что случайная величина X (толщина бобов в выборке) в генеральной распределена по нормальному закону, не подтвердилась.

Пример 5.2

В первых двух столбцах таблицы приведены данные об отказах аппаратуры за 10000 часов работы. Общее число обследованных экземпляров аппаратуры $n = 757$; при этом наблюдался $0 \cdot 427 + 1 \cdot 235 + 2 \cdot 72 + 3 \cdot 21 + 4 \cdot 1 + 5 \cdot 1 \;=\; 451$ отказ.

Приняв уровень значимости $\alpha = 0,01$, проверить гипотезу о том, что число отказов имеет распределение Пуассона:

$$p_k \;=\; \Pr[\mathrm{X} = k] \;=\; \frac{\lambda^k}{k!}\exp(-\lambda), \qquad k = 0, 1, 2, \ldots .$$

Решение

В качестве оценки $\hat{\lambda}$ параметра λ используем среднее число отказов:

$$\hat{\lambda} \;=\; 451/757 \;=\; 0,596.$$

По таблице распределения Пуассона с $\lambda = 0,596$ находим вероятности p_k и ожидаемое число случаев с k отказами (третий и четвертый столбцы таблицы).

Число отказов k	Количество случаев с k отказами n_k	Вероятность $p_k = \frac{0,596^k}{k!}\exp(-0,596)$	Ожидаемое число случаев с k отказами np_k
0	427	0,5511	417
1	235	0,3284	249
2	72	0,0978	74
3	21	0,0194	15
4	1	0,0029	2
5	1	0,0003	0

Для $k = 4$ и 5 значения $np_k < 5$, поэтому объединяем эти строки со строкой для $k = 3$.

В результате получаем значения, приведенные в таблице:

k	n_k	np_k	$(n_k - np_k)^2/np_k$
0	427	417	0,230
1	235	249	0,740
2	72	74	0,056
≥ 3	23	17	1,991

Имеем

$$\chi^2_{\text{набл}} \;=\; \sum_{k=0}^{3} \frac{(n_k - np_k)^2}{np_k} \;=\; 3,017.$$

Так как по выборке оценивался один параметр λ генеральной совокупности, то $l = 1$, число степеней свободы равно $4 - 1 - 1 = 2$. По таблице квантилей χ^2-распределения находим $\chi^2_{0,99;\,2} = 9,21$.

Итак, имеем $\chi^2_{\text{набл}} \;<\; \chi^2_{0,99;\,2}$. Следовательно, при данном уровне значимости гипотеза о распределении числа отказов по закону Пуассона принимается.

Пример 5.3

Результаты исследования прочности на сжатие (случайная величина X) 200 образцов бетона представлены в виде сгруппированного статистического ряда.

i	Интервалы прочности, *кг/см*^{2}	Частоты m_i
1	190–200	10
2	200–210	26
3	210–220	56
4	220–230	64
5	230–240	30
6	240–250	14

Требуется проверить нулевую гипотезу о нормальном законе распределения прочности на сжатие. Уровень значимости принять $\alpha = 0,05$.

Решение

Из условия следует, что точные параметры гипотетического нормального закона нам неизвестны, поэтому нулевую гипотезу словесно можно сформулировать следующим образом: $\{H_0$: *$F(x)$ является функцией нормального распределения*$\}$ с параметрами $\mathrm{M[X]} = \hat{a}$ и $\mathrm{D[X]} = \hat{\sigma}^2 = s^2$.

Для проверки этой нулевой гипотезы определим значение x_i^* середин интервалов и найдем точечные оценки математического ожидания и среднего квадратического отклонения нормально распределенной случайной величины по формулам ($n = 200$):

$$\hat{a} \;=\; \overline{x} \;=\; \frac{1}{200}\sum_{i=1}^{6} x_i^* m_i \;=$$

$$= \frac{195\cdot 10 + 205\cdot 26 + 215\cdot 56 + 225\cdot 64 + 235\cdot 30 + 245\cdot 14}{200} = 221 \text{ кг/см}^2;$$

$$\hat{\sigma}^2 \;=\; s^2 \;=\; \frac{1}{200}\sum_{i=1}^{6} \left(x_i^* - \overline{x}\right)^2 m_i \;=$$

$$= \;\frac{1}{200}[(-26)^2\cdot 10 + (-16)^2\cdot 26 + (-6)^2\cdot 56 + 4^2\cdot 64 + 14^2\cdot 30 + 24^2\cdot 14] \;=\; 152;$$

$$\hat{\sigma} \;=\; s \;=\; \sqrt{152} \;=\; 12,33 \text{ кг/см}^2.$$

Вычислим теоретические вероятности p_i попадания случайной величины X в частичные интервалы $[x_{i-1}; x_i]$ по формуле

$$p_i \;=\; \Pr\left(x_{i-1} \le \mathrm{X} < x_i\right) \;=\; \Phi(u_i) - \Phi(u_{i-1}); \qquad i = 1, 2, \ldots, k,$$

где

$$u_i \;=\; \frac{x_i^* - \overline{x}}{s}; \qquad \Phi(u) \;=\; \frac{1}{\sqrt{2\pi}}\int_0^u \exp(-t^2/2)\,dt.$$

Дальнейшие вычисления, необходимые для определения наблюдаемого значения выборочной статистики χ^2, сведены в таблице, в которой обозначено $Q_i = (m_i - np_i)^2$ и $R_i = (m_i - np_i)^2/np_i$.

i	Интервалы изменения	m_i	p_i	np_i	Q_i	R_i
1	190–200	10	0,0443	8,8507	1,3208	0,1492
2	200–210	26	0,1419	28,3769	5,6496	0,1991
3	210–220	56	0,2815	56,3078	0,0947	0,0017
4	220–230	64	0,2996	59,9254	16,6026	0,2771
5	230–240	30	0,1711	34,2101	17,7248	0,5181
6	240–250	14	0,0617	12,3292	2,7917	0,2264
	Суммы:	200	1,0000	200,00	44,1842	1,3716

В результате вычислений находим: $\chi^2_{\text{набл}} = 1,37$.

Замечание. Так как случайная величина X, распределенная по нормальному закону, определена на $(-\infty; \infty)$, то наименьшее значение стандартизованной переменной $(190 - 221)/12,33 = -2,51$ заменено на $-\infty$, наибольшее значение $(250 - 221)/12,33 = 2,35$ заменено на $+\infty$.

По таблице квантилей χ^2-распределения по заданному уровню значимости $\alpha = 0,05$ и числу степеней свободы $\nu = k - r - 1 = 6 - 2 - 1 = 3$ найдем критическое значение $\chi^2_{0,05;3} = 7,82$.

Так как $\chi^2_{\text{набл}} = 1,37 < 7,82$, то нет оснований для отклонения нулевой гипотезы о нормальном законе распределения предела прочности на сжатие с параметрами $a = 221$ и $\sigma^2 = 152$.

Пример 5.4

В течение 100 дней фиксировалось количество аварий водопроводной сети в некотором районе города. Получены следующие числовые данные:

Число аварий (СВ X)	0	1	2	3	4	5
Частоты m_i	8	28	31	18	9	6

$$n = \textstyle\sum_i m_i = 100$$

Требуется проверить гипотезу о том, что распределение числа аварий водопроводной сети города подчиняется закону Пуассона. Уровень значимости принять $\alpha = 0,05$.

Решение

Из условия следует, что необходимо проверить нулевую гипотезу $\{H_0$: *функция распределения $F(x)$ числа аварий имеет вид*$\}$

$$F(x;\lambda) = \sum_{i=1}^{x} \frac{\lambda^i}{i!} \exp(-\lambda)$$

с параметром

$$\hat{\lambda} = \overline{x} = \frac{1}{n} \sum_{i=0}^{5} m_i x_i = \frac{8 \cdot 0 + 28 \cdot 1 + 31 \cdot 2 + 18 \cdot 3 + 9 \cdot 4 + 6 \cdot 5}{100} = 2,1.$$

Вычислим теоретические вероятности p_i появления ровно x_i аварий в течение n дней по формуле Пуассона:

$$p_i = \mathrm{P}_n(x_i) = \frac{1}{x_i!} \lambda^{x_i} \exp(-\lambda); \qquad x_i = 0, 1, 2, 3, 4, 5.$$

Дальнейшие вычисления сводим в таблицу.

Число аварий x_i	m_i	p_i	np_i	$(m_i - np_i)^2$	$(m_i - np_i)^2/np_i$
0	8	0,1225	12,25	18,03	1,47
1	28	0,2572	25,72	5,22	0,20
2	31	0,2700	27,00	15,99	0,59
3	18	0,1890	18,90	0,81	0,04
4	9	0,0992	9,92	0,85	0,09
≥ 5	6	0,0621	6,21	0,05	0,01
Суммы:	100	1,0000	100,0	44,25	2,40

В результате вычислений находим $\chi^2_{\text{набл}} = 2,40$.

По таблице квантилей χ^2-распределения по заданному уровню значимости $\alpha = 0,05$ и числу степеней свободы $\nu = k - r - 1 = 6 - 1 - 1 = 4$ найдем критическое значение $\chi^2_{0,05;\,4} = 9,49$.

Так как $\chi^2_{\text{набл}} = 2,40 < 9,49$, то нет оснований для отклонения гипотезы о том, что закон распределения числа аварий водопроводной сети является законом Пуассона с параметром $\lambda = 2,1$.

Пример 5.5

При исследовании предела пластичности 15 образцов определенного сорта стали получены следующие результаты (в *кг/см*^{2}):

3540, 3580, 3570, 3560, 3500, 3610, 3720,
3640, 3600, 3650, 3750, 3590, 3600, 3550, 3770.

При дополнительном технологическом процессе, который предположительно приведет к увеличению предела пластичности, получены для тех же образцов следующие результаты:

3580, 3570, 3680, 3880, 3530, 3680, 3730,
3720, 3670, 3710, 3810, 3660, 3770, 3640, 3670.

Проверить с помощью критерия знаков гипотезу, что предел пластичности стали при проведении дополнительного технологического процесса увеличился. Уровень значимости принять в данной задаче $\alpha = 0,05$.

Решение

Сформулируем нулевую гипотезу $\{H_0 : F(x) = F(y)\}$ — предел пластичности исследуемого сорта стали не меняется при проведении дополнительного технологического процесса. Обозначим знаком "+" возрастание предела пластичности, а знаком "–" — уменьшение предела пластичности, вызванное дополнительным процессом. Получаем следующую последовательность знаков:

$$+ \; - \; + \; + \; + \; + \; + \; + \; + \; + \; + \; + \; + \; + \; -$$

Число знаков "–" составляет $r = 2$.

Найдем по таблице критических значений числа знаков (см. приложение) по заданному уровню значимости $\alpha = 0,05$ и объему выборки $n = 15$ критическое значение $r_{\alpha,\,n} = r_{0,05;\,15} = 3$.

Так как $r_{\text{набл}} = 2 < 3$, то нулевая гипотеза отклоняется. Другими словами,

считается статистически установленным, что дополнительный процесс приводит к увеличению предела пластичности данного сорта стали.

Пример 5.6

Из текущей продукции токарного станка-автомата, настроенного на обработку заданной детали, взяты две выборки объема $n_1 = 150$ и $n_2 = 100$. Первая выборка произведена в начале смены, а вторая – после двух часов работы станка. Результаты измерений отклонений контролируемого размера (СВ X) от номинала в микрометрах приведены в таблице.

Интервалы изменения СВ X, *мкм*	Частота в выборке № 1, m_{1j}	Частота в выборке № 2, m_{2j}
–15; –10	10	0
–10; –5	27	7
–5; 0	43	17
0; 5	38	30
5; 10	23	29
10; 15	8	15
15; 20	1	1
20; 25	0	1
Сумма:	150	100

Требуется с помощью λ-критерия Смирнова-Колмогорова (модель 2) проверить нулевую гипотезу о том, что распределение погрешностей обработки станка-автомата в течение исследуемого промежутка времени описывается одной и той же функцией распределения. Принять уровень значимости $\alpha = 0,05$.

Решение

Согласно условию примера, необходимо проверить нулевую гипотезу $\{H_0 : F_1(x) = F_2(x)\}$ (*процесс изготовления деталей на данном станке-автомате является устойчивым во времени*). Вычисление выборочной статистики $D^* = \max\limits_x |F_1^*(x) - F_2^*(x)|$ приведено в таблице, в которой m_{1j} и m_{2j} – частоты, n_{1j} и n_{2j} – накопленные частоты, $F_1^*(x) = n_1(x)/n_1$ и $F_2^*(x) = n_2(x)/n_2$.

x_{i+1}	m_{1i}	m_{2i}	$n_1(x)$	$n_2(x)$	$F_1^*(x)$	$F_2^*(x)$	$\lvert F_1^*(x) - F_2^*(x)\rvert$
–10	10	0	10	0	0,067	0,000	0,067
–5	27	7	37	7	0,247	0,070	0,177
0	43	17	80	24	0,533	0,240	0,293
5	38	30	118	54	0,787	0,540	0,247
10	23	29	141	83	0,940	0,830	0,110
15	8	15	149	98	0,993	0,980	0,013
20	1	1	150	99	1,000	0,990	0,010
25	0	1	150	100	1,000	1,000	0,000

Анализируя последний столбец данной таблицы, замечаем, что наибольший модуль разности между эмпирическими функциями распределения $F_1^*(x)$ и $F_2^*(x)$ равен $D^* = \max\limits_x |F_1^*(x) - F_2^*(x)| = 0,293$.

Так как

$$n = \frac{n_1 \cdot n_2}{n_1 + n_2} = \frac{150 \cdot 100}{150 + 100} = 60,$$

то наблюдаемое значение выборочной статистики

$$\lambda_{\text{набл}} = D^* \sqrt{n} = 0,293 \sqrt{60} = 2,272.$$

По таблице квантилей λ-распределения Смирнова–Колмогорова для заданного уровня значимости $\alpha = 0,05$ найдем критическое значение $\lambda_{0,05} = 1,358$.

Так как $\lambda_{\text{набл}} = 2,272 > 1,358$, нулевую гипотезу следует отклонить.

Таким образом, нельзя утверждать, что погрешности обработки в течение исследуемого интервала времени описываются одной и той же функцией распределения. Другими словами, полученный результат говорит о том, что процесс изготовления деталей на станке-автомате не является устойчивым во времени.

Пример 5.7

Комплектующие изделия одного наименования поступают с трех предприятий: A, B и C. Результаты проверки изделий приведены в следующей таблице (для поставщиков A, B и C):

Результаты проверки	A	B	C	Всего (v_i)
Годные	29	38	53	120
Негодные	1	2	7	10
Всего (u_j)	30	40	60	130

Приняв уровень значимости $\alpha = 0,10$, выяснить, можно ли считать, что качество изделий не зависит от поставщика?

Решение

Проверим гипотезу H_0 о независимости двух признаков: качества изделия X и места его изготовления Y. Для этого по критерию χ^2 используем статистику

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}. \tag{1}$$

Здесь n_{ij} – число исходов, в которых реализовалось очередное событие $\{\text{X} = x_i \text{ и } \text{Y} = y_j\}$, $\tilde{n}_{ij} = np_{ij}$ – ожидаемые частоты, p_{ij} – ожидаемые частости. В рамках гипотезы о независимости признаков X и Y имеем $p_{ij} = p_i p_j$, где p_i – вероятности попадания X в i-й интервал $(i = 1, \ldots, k)$, p_j – вероятности попадания Y в j-й интервал $(j = 1, \ldots, l)$, соответственно.

Из формулы (1) имеем

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij}^2}{\tilde{n}_{ij}} - 2 \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} + \sum_{i=1}^{k} \sum_{j=1}^{l} \tilde{n}_{ij}.$$

В этом выражении второе слагаемое равно $2n$, третье слагаемое равно n, поэтому

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij}^2}{np_{ij}} - n = n \left(\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij}^2}{n^2 p_i p_j} - 1 \right). \tag{2}$$

Примем для ожидаемых вероятностей, что $p_i = v_i/n$ и $p_j = u_j/n$, тогда приводя выражение (2) к удобному для вычислений виду, запишем

$$\chi^2_{\text{набл}} \;=\; n\left(\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}^2}{v_i u_j} - 1\right). \qquad (3)$$

При условии, что гипотеза H_0 верна и для всех ожидаемых частот выполняется $\tilde{n}_{ij} \geq 4$, статистика (1) имеет распределение χ^2 с $(k-1)(l-1)$ степенями свободы.

В нашем примере параметры k и l равны $k = 3$ и $l = 2$, так что полное число степеней свободы составляет $(k-1)(l-1) = 2$.

Пользуясь формулой (3), найдем

$$\chi^2_{\text{набл}} \;=\; 130 \cdot \left(\frac{29^2}{30\cdot 120} + \frac{38^2}{40\cdot 120} + \frac{53^2}{60\cdot 120} + \frac{1^2}{30\cdot 10} + \frac{2^2}{40\cdot 10} + \frac{7^2}{60\cdot 10} - 1\right),$$

откуда

$$\chi^2_{\text{набл}} \;=\; 2,546.$$

По таблице квантилей χ^2-распределения находим $\chi^2_{0,90;\,2} = 4{,}605$.

Так как $\chi^2_{\text{набл}} < \chi^2_{0,90;\,2}$, то при данном уровне значимости следует считать, что качество изделий не зависит от поставщика.

Пример 5.8

При помощи измерительного прибора было проведено 200 измерений заданного расстояния. Случайная погрешность измерений записана в микрометрах. Действительная ось была разделена на 9 промежутков, результаты (случайная погрешность измерений) сведены в таблицу (дробные значения m_i в ней появились из-за того, что значения, попавшие на границу интервала, принято записывать поровну как одному, так и другому интервалу).

Номер интервала	Интервал, *мкм*	Частота m_i
1	меньше –15	6
2	от –15 до –10	11,5
3	от –10 до –5	15,5
4	от –5 до 0	22
5	от 0 до 5	47,5
6	от 5 до 10	42
7	от 10 до 15	28
8	от 15 до 20	17
9	больше 20	10,5

Требуется проверить гипотезу H_0 о том, что случайная погрешность измерения X распределена нормально.

Решение

Поскольку ширина каналов составляет 5 *мкм*, примем для граничных 1-го и 9-го каналов значения $x_1 = -17,5$ *мкм* и $x_9 = 22,5$ *мкм* соответственно.

Необходимые действия выполним в следующем порядке:

а) рассмотрим гипотетический (нормальный) закон распределения;

б) по заданной выборке получим наиболее правдоподобные значения параметров распределения;

в) построим критерий χ^2, который используем для проверки гипотезы о нормальности случайной погрешности измерения X.

Гипотетическое распределение имеет плотность

$$f_{\mathrm{X}}(x;a,\sigma) \;=\; \frac{1}{\sqrt{2\pi\,\sigma^2}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right),$$

которая содержит два неизвестных параметра a и σ. Построим вспомогательную таблицу статистического распределение случайной погрешности измерений. Для этого по имеющейся выборке из $n = 200$ измерений вычислим оценки для a и σ, что дает

$$a \;\Rightarrow\; \overline{x}^* \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i m_i \;=\; 4,49\ \textit{мкм},$$

$$\sigma^2 \;\Rightarrow\; \sigma^{*2} \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}^*)^2 m_i \;=\; 90,86\ \textit{мкм}^2.$$

Таким образом, функция $f^*(x)$, используемая для проверки, имеет вид

$$f^*(x) \;=\; \frac{1}{\sqrt{2\pi\cdot 90,86}}\exp\left(-\frac{(x-4,49)^2}{2\cdot 90,86}\right).$$

Отсюда по таблице функции Лапласа можно найти вероятности p_i попадания в i-й интервал. Например, при $i = 1$ и $i = 2$

$$p_1 \;=\; \Pr(\mathrm{X}<-15) \;=\; \Pr\left(\frac{\mathrm{X}-4,49}{\sqrt{90,86}} < \frac{-15-4,49}{\sqrt{90,86}}\right) \;=$$

$$=\; \frac{1}{2}-\Phi(2,0444) \;=\; \frac{1}{2}-0,4795 \;=\; 0,0204,$$

$$p_2 \;=\; \Pr(-15<\mathrm{X}<-10) \;=\; \Pr\left(-2,0447 < \frac{\mathrm{X}-4,49}{\sqrt{90,86}} < -1,5201\right) \;=$$

$$=\; \Phi(-1,5199)-\Phi(-2,0444) \;=\; \Phi(2,0444)-\Phi(1,5199) \;=\; 0,0439$$

и так далее ($i = 3,\ldots,9$).

В результате строим таблицу данных, необходимых для нахождения значения критерия χ^2 (см. таблицу расчетных данных).

i	Частота m_i	p_i	np_i	$(m_i-np_i)^2/np_i$
1	6	0,0204	4,088	0,894
2	11,5	0,0438	8,763	0,855
3	15,5	0,0955	19,104	0,680
4	22	0,1591	31,822	3,032
5	47,5	0,2025	40,508	1,207
6	42	0,1970	39,406	0,171
7	28	0,1465	29,296	0,057
8	17	0,0832	16,644	0,008
9	10,5	0,0518	10,383	0,002
Сумма	200,0	1,0000	200,00	$\chi^2 = 6,905$

Таким образом, вычисление дает $\chi^2 = 6,905$.

Поскольку производится оценка двух параметров, то имеем здесь $k = 9 - 2 - 1 = 6$ степеней свободы. Примем, что уровень значимости составляет $\alpha = 0,05$, тогда из таблицы квантилей χ^2-распределения (см. приложение) получаем $\chi^2_{0,05;\,6} = 12,6$.

Так как вычисленное значение χ^2-критерия оказалось меньше квантиля $\chi^2_{0,05;\,6}$, то гипотеза о нормальном распределении случайной погрешности измерения прибора не противоречит результатам наблюдений на уровне значимости 0,05.

5.8. Задачи для решения

Задача 5.1

Часы, выставленные в витринах часовых магазинов, показывают случайное время. Предлагается гипотеза, что показания этих часов в витринах большого числа магазинов распределены равномерно в интервале (0; 12). Наблюдения 500 витрин 500 магазинов дали следующую выборку (весь интервал (0; 12) разбит на 12 часовых интервалов):

Час	0	1	2	3	4	5	6	7	8	9	10	11
Число выборочных значений	41	34	54	39	49	45	41	33	37	41	47	39

Согласуются ли эти данные с гипотезой о равномерности? Выбрать $\alpha = 0,10$ и $\alpha = 0,05$.

Задача 5.2

Амплитуда колебаний определялась двумя лаборантами. Первый лаборант по 10 наблюдениям получил среднее значение амплитуды $\overline{x}_1 = 81\,мм$, а второй лаборант по 15 наблюдениям получил $\overline{x}_2 = 84\,мм$.

В предположении, что дисперсии выполненных измерений известны и равны $\sigma_1^2 = 64\,мм^2$ и $\sigma_2^2 = 64\,мм^2$ для первого и второго лаборанта соответственно, найти 99 %-й доверительный интервал для разности средних $\overline{X}_1$ и $\overline{X}_2$. Можно ли считать, что результаты лаборантов действительно различаются?

Задача 5.3

Ниже приводится время (в секундах) решения контрольных задач учащимися до и после специальных упражнений по устному счету.

До упражнений	87	61	98	90	74	83	72	81	75	83	85
После упражнений	50	45	79	88	65	52	79	84	61	52	85

Можно ли считать, что эти упражнения улучшили способность учащихся к решению задач? Принять $\alpha = 0,10$.

Задача 5.4

За год на опытный район упало 537 небольших метеоритов. Вся территория района была разделена на 576 участков площадью по $0,25\,км^2$ каждый. Ниже

приведены числа участков n_k, на которые упало k метеоритов:

k	0	1	2	3	4	5 и больше
n_k	229	211	93	35	7	1

Согласуются ли эти данные с гипотезой о том, что число метеоритов, упавших на каждый их участков, имеет распределение Пуассона? Принять $\alpha = 0{,}10$.

Задача 5.5

Получены следующие две группы данных (см. таблицы).

40,25	40,37	40,33	40,28	40,29	40,41	40,35	40,28	40,29	40,27
40,35	40,35	40,41	40,30	40,33	40,40	40,34	40,46	40,39	40,38
40,45	40,44	40,35	40,40	40,31	40,33	40,34	40,32	40,39	40,37
40,39	40,30	40,33	40,32	40,36	40,34	40,43	40,31	40,37	40,36

40,40	40,34	40,38	40,32	40,34	40,30	40,36	40,31	40,38	40,35
40,42	40,31	40,33	40,42	40,30	40,43	40,34	40,36	40,36	40,32
40,35	40,35	40,30	40,36	40,33	40,37	40,31	40,34	40,37	40,37
40,32	40,32	40,33	40,35	40,30	40,34	40,34	40,34	40,41	40,36

С помощью критерия Колмогорова-Смирнова проверить гипотезу о том, что обе эти приведенные выборки извлечены из одной и той же генеральной совокупности.

Задача 5.6

Получены следующие данные (см. таблицу).

40,25	40,37	40,33	40,28	40,29	40,41	40,35	40,28	40,29	40,27
40,35	40,35	40,41	40,30	40,33	40,40	40,34	40,46	40,39	40,38
40,45	40,44	40,35	40,40	40,31	40,33	40,34	40,32	40,39	40,37
40,39	40,30	40,33	40,32	40,36	40,34	40,43	40,31	40,37	40,36
40,40	40,34	40,38	40,32	40,34	40,30	40,36	40,31	40,38	40,35
40,42	40,31	40,33	40,42	40,30	40,43	40,34	40,36	40,36	40,32
40,35	40,35	40,30	40,36	40,33	40,37	40,31	40,34	40,37	40,37
40,32	40,32	40,33	40,35	40,30	40,34	40,34	40,34	40,41	40,36

С помощью λ-критерия Колмогорова проверить гипотезу о том, что приведенная выборка извлечена из генеральной совокупности, равномерно распределенной на интервале (40,238; 40,462).

Задача 5.7

Предполагается, что один из двух приборов, определяющих скорость автомобиля, имеет систематическую ошибку (завышение). Для проверки этого предположения определили скорость 10 автомобилей, причем скорость каждого из них фиксировалась одновременно двумя приборами. В результате получены следующие данные:

v_1, *км/час*	70	85	63	54	65	80	75	95	52	55
v_2, *км/час*	72	86	62	55	63	80	78	90	53	57

Позволяют ли эти данные утверждать, что второй прибор действительно дает завышенные значения скорости? Принять $\alpha = 0{,}10$.

Задача 5.8

Утверждается, что результат действия лекарства зависит от способа его применения.

Проверить это утверждение при $\alpha = 0{,}05$ по следующим данным

Результат	Способ 1	Способ 2	Способ 3
Неблагоприятный	11	17	16
Благоприятный	20	23	19

Задача 5.9

В соответствии с техническими условиями среднее время безотказной работы для приборов из большой партии должно составлять не менее 1000 *час* с СКО $\sigma = 100$ *час*. Выборочное среднее времени безотказной работы для случайно отобранных 25 приборов оказалось равным 970 *час*. Предположим, что СКО времени безотказной работы для приборов в выборке совпадает с СКО времени безотказной работы всей партии.

Можно ли считать, что вся партия приборов не удовлетворяет техническим условиям, если: а) $\alpha = 0{,}10$; б) $\alpha = 0{,}05$?

Задача 5.10

Технология производства некоторого вещества дает в среднем 1000 *кг* вещества в сутки с СКО среднего, равным $\sigma^* = 80$ *кг*. Новая технология производства в среднем дает 1100 *кг* вещества в сутки с тем же СКО среднего.

Можно ли считать, что новая технология обеспечивает повышение производительности, если: а) $\alpha = 0{,}05$; б) $\alpha = 0{,}10$?

Задача 5.11

При измерении производительности двух агрегатов получены следующие результаты (в *кг* вещества за час работы):

№ замера	1	2	3	4	5
Агрегат A	14,1	10,1	14,7	13,7	14,0
Агрегат B	14,0	14,5	13,7	12,7	14,1

Можно ли считать, что производительности агрегатов A и B одинаковы, в предположении, что обе выборки получены из нормально распределенных генеральных совокупностей? Принять $\alpha = 0{,}05$.

Задача 5.12

Отношение зрителей к включению данной передачи в программу выразилось следующими данными:

	Положительное	Отрицательное	Безразличное
Мужчины	14	24	2
Женщины	29	36	15

Можно ли считать, что отношение к включению данной передачи в программу не зависит от пола зрителя? Принять $\alpha = 0{,}10$.

Задача 5.13

В больнице наблюдалось распределение красных кровяных шариков по 169 отделениям прибора (гемацитометра). Числа ν_i отделений, содержащих по n_i красных кровяных шариков, указаны в таблице.

ν_i	4	5	6	7	8	9	10	11	12
n_i	1	3	5	8	13	14	15	15	21
ν_i	13	14	15	16	17	18	19	20	21
n_i	18	17	16	9	6	3	2	2	1

Найти среднее значение числа красных кровяных шариков в одном отделении. Приняв найденное значение за параметр λ распределения Пуассона, проверить с помощью χ^2-критерия гипотезу о том, что выборка согласуется с распределением Пуассона.

Задача 5.14

Экспериментатор при $n = 4040$ бросаниях монеты получил $m = 2048$ выпадений герба.

Совместимо ли это с гипотезой о том, что существует постоянная вероятность $p = 0{,}5$ выпадения герба?

Задача 5.15

Испытывалась чувствительность 40 приемников. Данные приведены в таблице, в которой в первой строке даны интервалы чувствительности в микровольтах, во второй – средние точки этих интервалов $f_{\text{ср}}$, в третьей – число приемников n_i, чувствительность которых оказалась в этом интервале.

Интервал	25–75	75–125	125–175	175–225	225–275
$f_{\text{ср}}$	50	100	150	200	250
n_i	0	0	1	5	9
Интервал	275–325	325–375	375–425	425–475	475–525
$f_{\text{ср}}$	300	350	400	450	500
n_i	6	8	6	2	2
Интервал	525–575	575–625	625–675	675–725	725–775
$f_{\text{ср}}$	550	600	650	700	750
n_i	0	1	1	0	0

С помощью χ^2-критерия проверить гипотезу о том, что выборка извлечена из нормальной совокупности.

Задача 5.16

Предполагается, что применение новой технологии в производстве микросхем приведет к увеличению выхода годной продукции. Результаты контроля двух партий продукции, изготовленных по старой и новой технологии, приведены ниже:

Изделия	Старая технология	Новая технология
Годные	140	185
Негодные	10	15
Всего	150	200

Подтверждают ли эти результаты предположение об увеличении выхода годной продукции? Принять $\alpha = 0{,}10$.

Задача 5.17

При испытании радиоэлектронной аппаратуры фиксировалось число отказов. Результаты 59 испытаний приводятся ниже:

Число отказов	0	1	2	3
Число испытаний	42	10	4	3

Проверить гипотезу H_0 о том, что число отказов имеет распределение Пуассона, приняв $\alpha = 0{,}10$.

Задача 5.18

До наладки станка была проверена точность изготовления 10 втулок и найдено значение оценки дисперсии диаметра $s^{*2} = 9,6\,мкм^2$. После наладки подверглось контролю еще 15 втулок и получено новое значение дисперсии $s^{*2} = 5,7\,мкм^2$.

Можно ли считать, что в результате наладки станка точность изготовления втулок увеличилась? Принять $\alpha = 0,05$.

Задача 5.19

При 600 подбрасываниях шестигранной игральной кости шестерка появилась 75 раз.

Верна ли гипотеза о том, что вероятность появления шестерки меньше чем $1/6$, если $\alpha = 0{,}01$?

Задача 5.20

При 50 подбрасываниях монеты "герб" появился 20 раз.

Можно ли считать монету симметричной? Принять $\alpha = 0,10$.

Задача 5.21

При 120 подбрасываниях игральной шестигранной кости на верхней грани единица выпала $m_1 = 25$ раз, двойка $m_2 = 19$, тройка $m_3 = 15$, четверка $m_4 = 22$, пятерка $m_5 = 15$ и, наконец, шестерка $m_6 = 21$ раз.

Согласуется ли это с тем, что игральная кость правильной формы?

Задача 5.22

Метод получения случайных чисел был применен 250 раз. При этом были получены следующие результаты:

Цифра	0	1	2	3	4	5	6	7	8	9
Частота появления	27	18	23	31	21	23	28	25	22	32

Можно ли считать, что примененный метод действительно дает случайные числа? Принять $\alpha = 0{,}05$.

5.9. Задание на практическую работу

Настоящая практическая работа рассчитана на два часа и содержит два задания. Задания должны выполняться в выбранной программной среде.

Задание 1

В течение 100 дней фиксировалось число посетителей (X) некоторого учреждения (данные приводятся в таблице).

Требуется с помощью критерия χ^2 Пирсона проверить нулевую гипотезу о том, что распределение числа посетителей подчиняется закону Пуассона. Уровень значимости принять $\alpha = 0,05$.

Результаты оформите графически.

Результат работы – оценка параметра λ закона Пуассона, массив амплитуд вероятностей P_m, принятие решения относительно нулевой гипотезы.

Вариант 1

Число посетителей (X)	0	1	2	3	4	5
Частоты m_i	7	27	32	19	11	4

Вариант 2

Число посетителей (X)	0	1	2	3	4	5
Частоты m_i	16	21	23	17	13	10

Вариант 3

Число посетителей (X)	0	1	2	3	4	5
Частоты m_i	3	27	39	24	6	1

Задание 2

Имеется выборка (данные приводятся в таблице).

Требуется с помощью λ–критерия Колмогорова проверить нулевую гипотезу о том, что приведенная выборка извлечена из генеральной совокупности, равномерно распределенной на интервале $(0,0;\ 2,0)$. Уровень значимости принять $\alpha = 0,05$.

Результаты оформите графически.

Результат работы – принятие решения относительно нулевой гипотезы.

При каком значении уровня значимости α нулевая гипотеза будет отклонена?

Вариант 1

0,5	1,7	1,3	0,8	0,9	0,1	1,5	1,8	0,9	0,7	1,5	0,5	0,1	0,0	1,3
1,0	0,4	1,6	1,9	1,8	0,5	1,4	1,5	1,0	1,1	0,3	1,4	1,2	1,9	1,7
1,9	0,0	0,3	1,2	0,6	1,4	0,3	0,1	1,7	0,6	0,0	1,4	1,8	0,2	0,4
0,0	1,6	0,1	0,8	0,5	1,2	0,1	0,3	0,2	1,0	1,3	0,4	1,6	0,6	1,2

Вариант 2

0,5	1,7	1,3	0,8	0,9	0,1	1,5	1,8	0,9	0,7	1,5	0,5	0,1	0,0	1,3
0,5	1,5	1,0	1,6	1,3	0,7	1,1	0,4	0,7	1,7	1,2	0,2	0,3	1,5	0,0
1,9	0,0	0,3	1,2	0,6	1,4	0,3	0,1	1,7	0,6	0,0	1,4	1,8	0,2	0,4
0,0	1,6	0,1	0,8	0,5	1,2	0,1	0,3	0,2	1,0	1,3	0,4	1,6	0,6	1,2

Вариант 3

1,5	0,5	0,0	0,6	0,3	1,7	0,1	1,4	1,7	0,7	0,2	1,2	0,3	0,5	1,0
0,5	1,5	1,0	1,6	1,3	0,7	1,1	0,4	0,7	1,7	1,2	0,2	0,3	1,5	0,0
1,9	0,0	0,3	1,2	0,6	1,4	0,3	0,1	1,7	0,6	0,0	1,4	1,8	0,2	0,4
0,0	1,6	0,1	0,8	0,5	1,2	0,1	0,3	0,2	1,0	1,3	0,4	1,6	0,6	1,2

5.10. Задания для проверки

1. Что называется критерием согласия?

2. Являются ли критерии согласия статистическими критериями значимости?

3. Какими характерными особенностями обладают статистические критерии значимости и, в частности, критерии согласия?

4. Дайте общую схему проверки гипотез о виде функции распределения с помощью критерия согласия χ^2 Пирсона.

5. Дайте общую схему проверки гипотез о виде функции распределения с помощью λ-критерия Колмогорова.

6. Укажите достоинства и недостатки критерия согласия χ^2 Пирсона; λ-критерия Колмогорова.

7. На основании каких признаков или критериев можно произвести предварительный выбор закона распределения?

8. Могут ли имеющиеся опытные данные одновременно согласовываться с несколькими законами распределения?

9. Укажите достоинства и недостатки непараметрических критериев по сравнению с параметрическими критериями значимости.

10. Для проверки каких гипотез применяется критерий знаков?

11. Является ли критерий знаков статистическим критерием значимости?

12. Аналогом какого параметрического критерия является критерий знаков?

6. Линейный регрессионный анализ

6.1. Задачи регрессионного и корреляционного анализа

Одной из основных задач математической статистики является исследование зависимости между двумя или несколькими параметрами. Две переменные X и Y могут быть либо независимыми, либо связанными функциональной или иной статистической зависимостью.

Определение 1. *Функциональной зависимостью* между переменными X и Y называется правило f, которое каждому элементу X из произвольного множества Ω ставит в соответствие определенный элемент Y множества F, т.е. $y = f(x)$. Например, площадь круга ($S = \pi R^2$) и длина окружности ($L = 2\pi R$) полностью определяются радиусом R, площадь треугольника – его сторонами и т. д.

Определение 2. *Статистической зависимостью* между случайными величинами X и Y – составляющими двумерной случайной величины (X, Y) – называется правило f, которое каждому числу x из числового множества R ставит в соответствие условный закон распределения составляющей Y, т.е. каждому x соответствует $f(x, y)$.

На практике функциональные связи между признаками встречаются редко. Чаще имеют место такие связи между переменными величинами, при которых численному значению одной из них соответствует несколько значений других. Например, известно, что урожайность зависит от количества внесенных удобрений, но на неё влияют и другие факторы (качество почвы, осадки и т. д.). Кроме того, одни и те же дозы удобрений, даже при очень выровненных условиях, часто по-разному влияют на урожайность.

Определение 3. *Случайные величины* X *и* Y *называются независимыми*, если условный закон распределения одной из составляющих не зависит от того, какие значения приняла другая составляющая, т.е. если выполняется $f(y|x) = f(y)$ или $f(x|y) = f(x)$.

В курсе теории вероятностей показывается, что для того, чтобы составляющие X и Y двумерной случайной величины были независимыми, необходимо и достаточно, чтобы плотность распределения была равна произведению парциальных плотностей распределения составляющих: $f(x, y) = f(x) \cdot f(y)$.

Из определения статистической зависимости следует, что для изучения изменения случайной величины Y по значениям случайной величины X на основании имеющихся статистических данных, т.е. наблюденным значениям двумерной случайной величины (X, Y) $(x_i, y_i,\ i = 1, 2, \ldots, n)$, необходимо:

1) подобрать теоретико-вероятностную модель, характеризующую частотные

закономерности рассматриваемого двумерного статистического ряда $(x_i, y_i,\ i = 1, 2, \ldots, n)$, т.е. выбрать функцию $F(x, y)$ или $f(x, y)$;

2) оценить параметры этой функции;

3) найти условные законы распределения

$$f(y|x) = \frac{f(x,y)}{f(x)} \quad \text{или} \quad f(x|y) = \frac{f(x,y)}{f(y)}; \tag{6.1}$$

4) задать вероятность $\mathrm{P}= 1 - \alpha$ и по значению случайной величины $\mathrm{X}= x$ определить интервал $[a,\ b]$ изменения случайной величины Y:

$$\Pr(a < \mathrm{Y} < b) \ = \ \int\limits_a^b f(y|x)\, dy \ = \ 1 - \alpha. \tag{6.2}$$

На практике при исследовании зависимости между случайными величинами X и Y часто ограничиваются исследованием зависимости между X и условным математическим ожиданием

$$\mathrm{M}[\mathrm{Y}|\mathrm{X} = x] \ = \ \int\limits_{-\infty}^{\infty} y\, f(y|x)\, dy.$$

Зависимости такого рода называются *регрессионными зависимостями*.

Условное математическое ожидание $\mathrm{M}[\mathrm{Y}|\mathrm{X}= x]$ зависит от выбранной теоретико-вероятностной модели $f(x, y)$, т.е. оно является понятием модельным.

Определение 4. *Уравнением регрессии первого рода* Y *на* X или *модельным уравнением регрессии* называется математическое ожидание составляющей Y двумерной случайной величины (X, Y), рассматриваемое как функция x, вычисленное при условии, что составляющая X приняла некоторое фиксированное значение $\mathrm{X}= x$:

$$\mathrm{M}[\mathrm{Y}|\mathrm{X} = x] \ = \ \int\limits_{-\infty}^{\infty} y\, f(y|x)\, dy \ = \ \varphi(x). \tag{6.3}$$

Функцию $\varphi(x)$ называют функцией регрессии первого рода или модельной функцией регрессии Y на X.

Определение 5. *Уравнением регрессии* X *на* Y *второго рода* называется условное математическое ожидание составляющей X двумерной случайной величины (X, Y), рассматриваемое как функция y:

$$\mathrm{M}[\mathrm{X}|\mathrm{Y} = y] \ = \ \int\limits_{-\infty}^{\infty} x\, f(x|y)\, dx \ = \ \psi(y). \tag{6.4}$$

Функцию $\psi(y)$ называют функцией регрессии X на Y.

Модельное уравнение регрессии вида $\mathrm{M}[\mathrm{Y}|\mathrm{X}= x] = \varphi(x)$ позволяет делать "точечное" предсказание значений условных математических ожиданий составляющей Y двумерной случайной величины (X, Y) по значениям составляющей $\mathrm{X}= x$. Однако, для такого прогноза необходимо знать закон распределения двумерной случайной величины (X, Y).

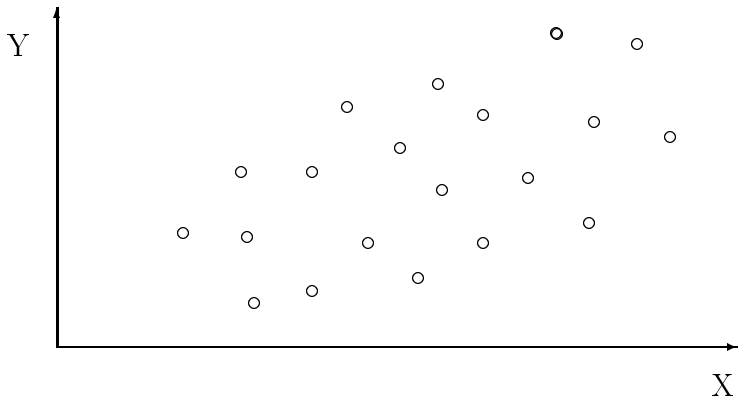


Рисунок 6.1 — Пример корреляционного поля

На практике при обработке экспериментальных данных закон распределения двумерной случайной величины (X,Y), как правило, неизвестен. В распоряжении экспериментатора имеются только наблюденные значения двумерной величины – точки (x_i, y_i) $(i = 1, 2, 3, \ldots, n)$ или, более кратко, выборка объема n. Если результаты выборки (x_i, y_i) $(i = 1, 2, 3, \ldots, n)$ изобразить в виде точек в декартовой системе координат, то получим точечную диаграмму, называемую *корреляционным полем* (рис. 6.1).

Определение 6. *Эмпирической функцией регрессии* Y на X называется функция $\overline{y}_x = f(x; a, b, \ldots, d)$ определенного класса, совокупность параметров которой a, b, ..., d находится по наблюденным значениям двумерной случайной величины (x_i, y_i) $(i = 1, 2, \ldots, n)$, т.е. по результатам выборки объема n.

Для решения задачи нахождения параметров эмпирических уравнений регрессии $\mathrm{M}[\mathrm{Y}|\mathrm{X}= x] = f(x; a, b, \ldots, d)$ применяется метод наименьших квадратов (МНК). Этот метод позволяет при заданном виде выбранной зависимости $\mathrm{M}[\mathrm{Y}|x] = f(x; a, b, \ldots, d)$ так выбрать совокупность параметров $a, b, \ldots, d$, что эмпирическая функция регрессии $\overline{y}_x = f(x; a, b, \ldots, d)$ будет наилучшей оценкой истинной функции регрессии. "Наилучшая оценка" здесь понимается в том смысле, что сумма квадратов (*невязка*) отклонений ε наблюденных значений переменной Y от соответствующих ординат эмпирической функции регрессии $\overline{y}_x = f(x; a, b, \ldots, d)$ будет минимальной в пространстве указанных параметров.

Искомые параметры a, b, ..., d заданной функции $f(x; a, b, \ldots, d)$ по МНК находятся из условия

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - f(x_i; a, b, \ldots, d)]^2 \Rightarrow \min. \tag{6.5}$$

Из условия экстремальности невязки S следует система уравнений:

$$\partial S/\partial a = 0; \qquad \partial S/\partial b = 0; \qquad \ldots \qquad \partial S/\partial d = 0. \tag{6.6}$$

Выбор класса эмпирической функции регрессии может быть сделан исходя из следующего:

а) визуальной оценки характера расположения точек (x_i, y_i) на корреляционном поле;

б) опыта предыдущих исследований;

в) соображений теоретического характера, основанных на знании сущности решаемой задачи.

Итак, эмпирическое уравнение регрессии $\overline{y}_x = f(x; a, b, \ldots, d)$ интерпретируется как оценка (приближенное выражение) модельного уравнения регрессии Y на X. Аналогично интерпретируется эмпирическое уравнение регрессии X на Y.

Регрессионный анализ – это анализ функций регрессии первого и второго рода. С его помощью решаются следующие задачи:

1) находятся точечные и интервальные оценки параметров эмпирической функции регрессии;

2) производятся точечное и интервальное оценивания условных математических ожиданий, необходимое для предсказания средних значений одной случайной величины, соответствующих определенным фиксированным значениям другой случайной величины;

3) проверяется согласованность найденной эмпирической функции регрессии с экспериментальными данными и др.

Корреляционный анализ – это анализ свойств оценок $\hat{\rho}$ коэффициента корреляции

$$\rho \;=\; \frac{\mathrm{M}\left[(\mathrm{X} - m_{\mathrm{X}})\,(\mathrm{Y} - m_{\mathrm{Y}})\right]}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}}. \tag{6.7}$$

Он позволяет дать ответ на вопрос о существовании линейной функциональной зависимости между математическими ожиданиями случайных величин X и Y. В случае положительного ответа метод корреляционного анализа позволяет измерить степень тесноты статистической зависимости (степень близости статистической зависимости к функциональной).

6.2. Вероятностное введение в регрессионный анализ

Если известен закон распределения двумерной случайной величины (X, Y), то можно найти условный закон распределения составляющей Y при условии, что составляющая X приняла некоторое фиксированное значение x, по формуле

$$f(y|x) \;=\; \frac{f(x,y)}{f_1(x)}, \tag{6.8}$$

где

$$f_1(x) \;=\; \int\limits_{-\infty}^{\infty} f(x,y)\,dy \tag{6.9}$$

– плотность распределения составляющей X двумерной случайной величины (X, Y).

Условное математическое ожидание составляющей Y (модельное уравнение регрессии Y на X) находится по формуле

$$\mathrm{M}[\mathrm{Y}|\mathrm{X} = x] \;=\; \int\limits_{-\infty}^{\infty} y\,f(y|x)\,dy. \tag{6.10}$$

Рассмотрим двумерную случайную величину (X,Y), распределенную по нормальному закону с плотностью распределения вероятностей (рис. 6.2 и 6.3)

$$f(x,y) \;=\; \frac{1}{2\pi\,\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}\,\sqrt{1-\rho^2}}\;\times \tag{6.11}$$

$$\times\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-m_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2}-\frac{2\rho\,(x-m_{\mathrm{X}})(y-m_{\mathrm{Y}})}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}}+\frac{(y-m_{\mathrm{Y}})^2}{\sigma_{\mathrm{Y}}^2}\right)\right\}.$$

Найдем плотности распределения вероятностей составляющих X и Y:

$$f_1(x) \;=\; \int\limits_{-\infty}^{\infty} f(x,y)\,dy \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_{\mathrm{X}}}\,\exp\left(-\frac{(x-m_{\mathrm{X}})^2}{2\sigma_{\mathrm{X}}^2}\right), \tag{6.12}$$

поэтому из симметрии плотности нормального двумерного распределения следует, что

$$f_1(y) \;=\; \int\limits_{-\infty}^{\infty} f(x,y)\,dx \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_{\mathrm{Y}}}\,\exp\left(-\frac{(y-m_{\mathrm{Y}})^2}{2\sigma_{\mathrm{Y}}^2}\right). \tag{6.13}$$

Найдем условные распределения составляющих X и Y:

$$f(x|y) \;=\; \frac{f(x,y)}{f_1(y)} \;= \tag{6.14}$$

$$=\;\frac{1}{\sqrt{2\pi(1-\rho^2)}\,\sigma_{\mathrm{X}}}\,\exp\left\{-\frac{1}{2(1-\rho^2)\sigma_{\mathrm{X}}^2}\left[x-m_{\mathrm{X}}-\rho\,\frac{\sigma_{\mathrm{X}}}{\sigma_{\mathrm{Y}}}\,(y-m_{\mathrm{Y}})\right]^2\right\};$$

$$f(y|x) \;=\; \frac{f(x,y)}{f_1(x)} \;= \tag{6.15}$$

$$=\;\frac{1}{\sqrt{2\pi(1-\rho^2)}\,\sigma_{\mathrm{Y}}}\,\exp\left\{-\frac{1}{2(1-\rho^2)\sigma_{\mathrm{Y}}^2}\left[y-m_{\mathrm{Y}}-\rho\,\frac{\sigma_{\mathrm{Y}}}{\sigma_{\mathrm{X}}}\,(x-m_{\mathrm{X}})\right]^2\right\}.$$

Условные математические ожидания следующие:

$$\begin{aligned} m_{\mathrm{X|Y}} \;&=\; \mathrm{M}[\mathrm{X}|\mathrm{Y}=y] \;=\; m_{\mathrm{X}}+\rho\,\frac{\sigma_{\mathrm{X}}}{\sigma_{\mathrm{Y}}}\,(y-m_{\mathrm{Y}})\,;\\ m_{\mathrm{Y|X}} \;&=\; \mathrm{M}[\mathrm{Y}|\mathrm{X}=x] \;=\; m_{\mathrm{Y}}+\rho\,\frac{\sigma_{\mathrm{Y}}}{\sigma_{\mathrm{X}}}\,(x-m_{\mathrm{X}})\,; \end{aligned} \tag{6.16}$$

$$\begin{aligned} \sigma[\mathrm{X}|\mathrm{Y}=y] \;&=\; \sigma_{\mathrm{X}}\sqrt{1-\rho^2};\\ \sigma[\mathrm{Y}|\mathrm{X}=x] \;&=\; \sigma_{\mathrm{Y}}\sqrt{1-\rho^2}. \end{aligned} \tag{6.17}$$

В приведенном условном распределении составляющей Y при условии, что составляющая X приняла некоторое значение x, только условное математическое ожидание $m_{\mathrm{Y|X}}=\mathrm{M}[\mathrm{Y}|\mathrm{X}=x]$ зависит от значения x, в то время как условное среднее квадратическое отклонение не зависит от значения x величины X.

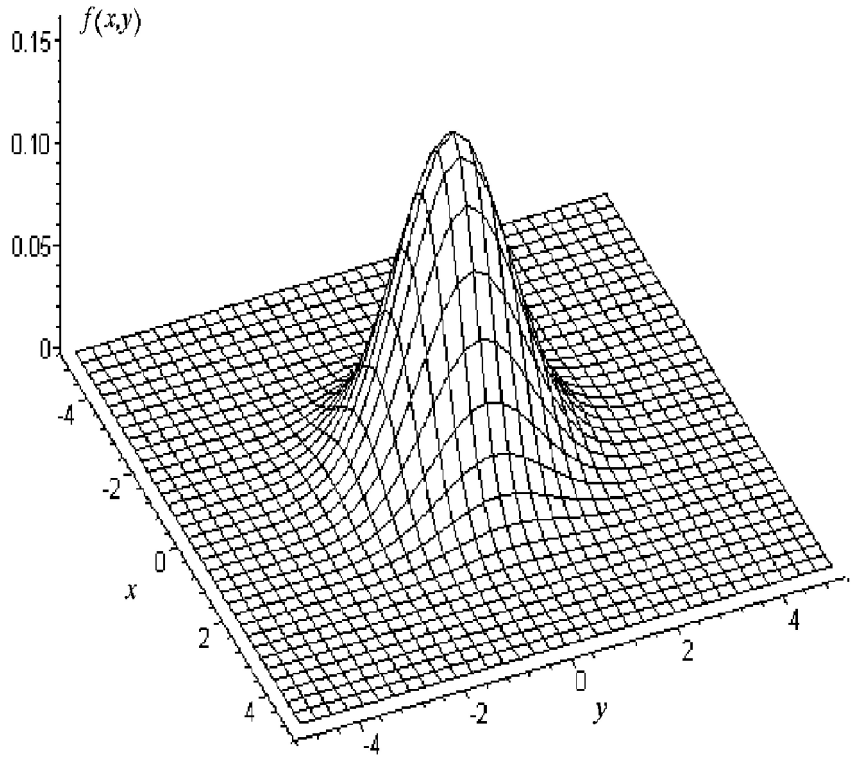


Рисунок 6.2 — Плотность распределения вероятностей $f(x,y)$ системы из двух случайных величин X, Y (параметры: $m_{\mathrm{X}} = m_{\mathrm{Y}} = 0$; $\sigma_{\mathrm{X}} = \sigma_{\mathrm{Y}} = 1$; коэффициент корреляции $\rho = 0,0$)

Таким образом, модельное уравнение регрессии Y на X двумерной случайной величины (X, Y), распределенной по нормальному закону, имеет вид

$$m_{\mathrm{Y|X}} - m_{\mathrm{Y}} \;=\; \rho\,\frac{\sigma_{\mathrm{Y}}}{\sigma_{\mathrm{X}}}\,(x - m_{\mathrm{X}})\,. \tag{6.18}$$

График этой функции, называемой *модельной линией регрессии* Y *на* X, является прямой линией, проходящей через центр распределения – точку $(m_{\mathrm{X}}, m_{\mathrm{Y}})$.

Аналогичные результаты получаем, анализируя условную плотность распределения вероятностей $f(x|y)$. В этом случае уравнение регрессии X на Y имеет вид

$$m_{\mathrm{X|Y}} - m_{\mathrm{X}} \;=\; \rho\,\frac{\sigma_{\mathrm{X}}}{\sigma_{\mathrm{Y}}}\,(y - m_{\mathrm{Y}})\,, \tag{6.19}$$

при этом соответствующая модельная линия регрессии X на Y также проходит через центр распределения – точку $(m_{\mathrm{X}}, m_{\mathrm{Y}})$.

6.3. Линейная регрессия

Рассмотрим методику расчета эмпирических линейных уравнений регрессии по *несгруппированным* и *сгруппированным данным.*

Несгруппированные данные.

Предположим, что произведен эксперимент, в результате которого зафиксировано n значений исследуемых переменных X и Y: (x_i, y_i) $(i = 1, 2, \ldots, n)$. Нанося

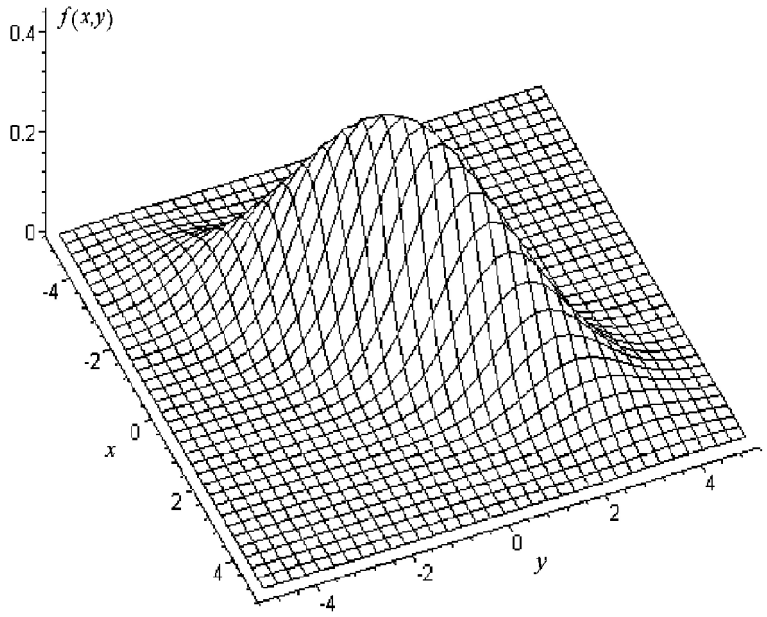


Рисунок 6.3 — Плотность распределения вероятностей $f(x,y)$ системы из двух случайных величин X, Y (параметры: $m_{\rm X} = m_{\rm Y} = 0$; $\sigma_{\rm X} = \sigma_{\rm Y} = 1$; коэффициент корреляции $\rho = 0,8$)

экспериментальные данные в виде точек в декартовой системе координат, получаем корреляционное поле (рис. 6.1). Если есть основания полагать, что двумерная случайная величина (X,Y) распределена по нормальному закону или если точки на корреляционном поле группируются вокруг прямой линии (рис. 6.4), то эмпирическое уравнение регрессии подбирается в виде

$$\overline{y}_x \;=\; b_0 + b_1 x. \tag{6.20}$$

Следующая задача – нахождение коэффициентов (параметров) b_0 и b_1 линейной эмпирической функции регрессии Y на X. Будем искать эти параметры методом наименьших квадратов, т.е. потребуем выполнения условия минимума невязки

$$S \;=\; \sum_{i=1}^{n} \varepsilon_i^2 \;=\; \sum_{i=1}^{n} \left(y_i - \overline{y}_x\right)^2 \;=\; \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right)^2 \;\Rightarrow\; \min. \tag{6.21}$$

Поскольку функционал невязки S в пространстве параметров b_0 и b_1 имеет минимум, то для определения координат этого экстремума найдем частные производные $\partial S/\partial b_0$, $\partial S/\partial b_1$ и приравняем их нулю:

$$\partial S/\partial b_0 \;=\; 0 \;\Rightarrow\; 2\sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right) \;=\; 0, \tag{6.22a}$$

$$\partial S/\partial b_1 \;=\; 0 \;\Rightarrow\; 2\sum_{i=1}^{n} x_i \left(y_i - b_0 - b_1 x_i\right) \;=\; 0. \tag{6.22b}$$

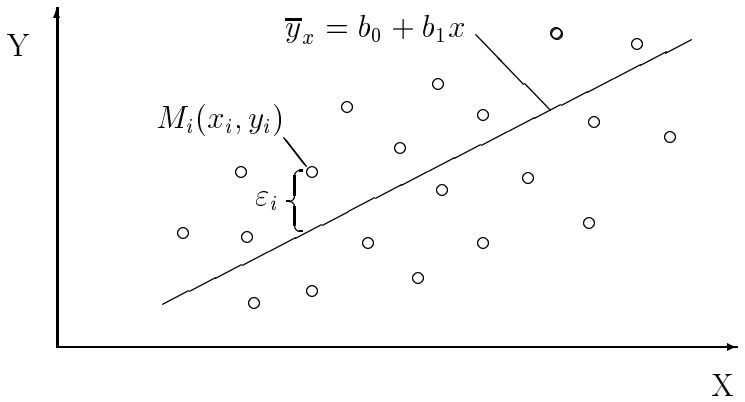


Рисунок 6.4 — Пример корреляционного поля и линейного уравнения регрессии $\overline{y}_x = b_0 + b_1 x$

Получаем систему из двух линейных уравнений, которые называют *нормальными уравнениями*:

$$\begin{cases} b_0 n + b_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i, \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases} \tag{6.23}$$

Решая систему, находим искомые коэффициенты b_0 и b_1.

Если требуется по экспериментальным данным рассчитать линейное уравнение регрессии X на Y вида

$$\overline{x}_y = a_0 + a_1 y, \tag{6.24}$$

то её коэффициенты (параметры) a_0 и a_1 находят из решения системы нормальных уравнений:

$$\begin{cases} a_0 n + a_1 \sum\limits_{i=1}^{n} y_i = \sum\limits_{i=1}^{n} x_i, \\ a_0 \sum\limits_{i=1}^{n} y_i + a_1 \sum\limits_{i=1}^{n} y_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases} \tag{6.25}$$

Сгруппированные данные.

Если число измерений велико, то для упрощения расчетов экспериментальные данные принято группировать, т.е. объединять в сводную таблицу, называемую *корреляционной таблицей* (табл. 6.1).

В этой таблице:

$x_1, x_2, \ldots, x_k$ – середины интервалов или значения признаков X;

$y_1, y_2, \ldots, y_l$ – середины интервалов или значения признаков Y;

$m_{x1}, m_{x2}, \ldots, m_{xi}, \ldots, m_{xk}$ и $m_{y1}, m_{y2}, \ldots, m_{yj}, \ldots, m_{yl}$ – соответствующие частоты;

m_{ij} – частота, с которой встречается пара (x_i, y_j);

$m_x = \sum_{j=1}^{k} m_{xj}$; $m_y = \sum_{i=1}^{l} m_{yi}$;

$n = \sum_{i=1}^{k} \sum_{j=1}^{l} m_{ij}$ – объем выборки.

Подробное изложение вычислений с помощью корреляционной таблицы приведено ниже и в примерах к разделу.

Таблица 6.1 — Корреляционная таблица

X	Y						
	y_1	y_2	$\cdots$	y_j	$\cdots$	y_l	m_x
x_1	m_{11}	m_{12}	$\cdots$	m_{1j}	$\cdots$	m_{1l}	m_{x1}
x_2	m_{21}	m_{22}	$\cdots$	m_{2j}	$\cdots$	m_{2l}	m_{x2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_i	m_{i1}	m_{i2}	$\cdots$	m_{ij}	$\cdots$	m_{il}	m_{xi}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_k	m_{k1}	m_{k2}	$\cdots$	m_{kj}	$\cdots$	m_{kl}	m_{xk}
m_y	m_{y1}	m_{y2}	$\cdots$	m_{yj}	$\cdots$	m_{yl}	n

Пусть требуется на основании выборки объема n "оценить" значение модельной функции регрессии и сделать предсказание условных математических ожиданий случайной величины Y, соответствующих определенным значениям случайной величины X= x. Для этого используются уравнения регрессии второго рода (эмпирические уравнения регрессии). Приближенное выражение (оценку) модельной функции регрессии называют *эмпирической функцией регрессии* $\overline{y}_x = f(x, a, b, \ldots, d)$.

6.4. Коэффициент корреляции

Из теории вероятностей известно, что основными характеристиками, описывающими степень связи между составляющими X и Y двумерной случайной величины (X,Y), являются *ковариация* μ_{XY} (корреляционный момент)

$$\mu_{11} = \mu_{\mathrm{XY}} = \mathrm{M}[(\mathrm{X} - m_{\mathrm{X}})(\mathrm{Y} - m_{\mathrm{Y}})] = \tag{6.26}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_{\mathrm{X}})(y - m_{\mathrm{Y}}) f(x, y)\, dx\, dy$$

и *коэффициент корреляции*

$$\rho = \frac{\mu_{\mathrm{XY}}}{\sigma_{\mathrm{X}} \sigma_{\mathrm{Y}}} = \frac{\mathrm{M}[(\mathrm{X} - m_{\mathrm{X}})(\mathrm{Y} - m_{\mathrm{Y}})]}{\sigma_{\mathrm{X}} \sigma_{\mathrm{Y}}}. \tag{6.27}$$

Из выражений (6.26) и (6.27) следует, что для нахождения μ_{XY} и ρ необходимо знать закон распределения двумерной случайной величины. В большинстве случаев при обработке экспериментальных данных закон распределения двумерной случайной величины неизвестен. Поэтому для оценки тесноты связи (рис. 6.5, 6.6 и 6.7) применяются точечные оценки $\hat{\mu}_{\mathrm{XY}}$ и $\hat{\rho}$ величин μ_{XY} и ρ:

эмпирический корреляционный момент

$$\hat{\mu}_{\mathrm{XY}} = K_{\mathrm{XY}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}); \tag{6.28}$$

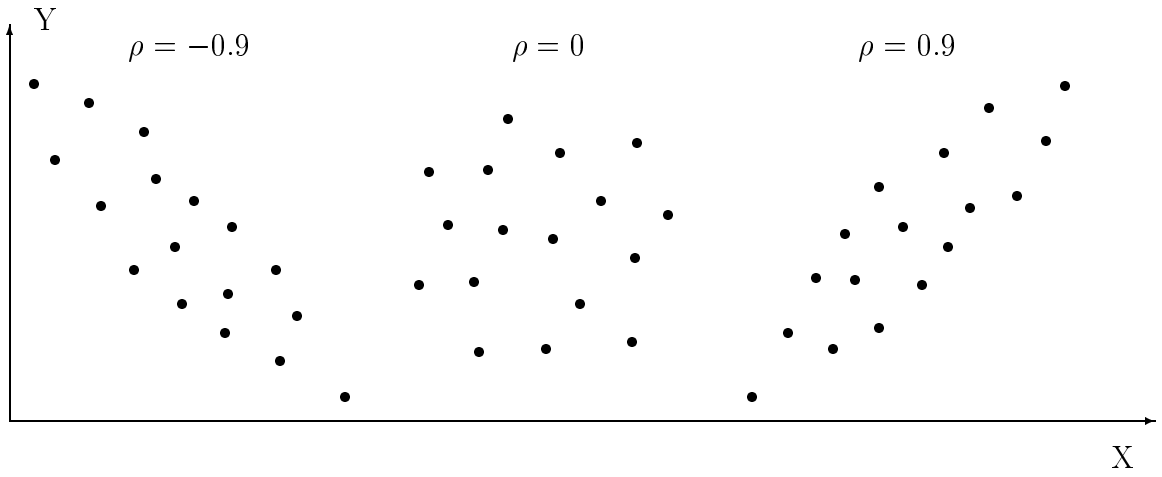


Рисунок 6.5 — Три примера корреляционного поля ($\rho = -0,9;\ 0{,}0;\ 0{,}9$; линии регрессии отсутствуют)

эмпирический коэффициент корреляции

$$\hat{\rho} \;=\; r \;=\; \frac{K_{\mathrm{XY}}}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} \;= \tag{6.29}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\right)\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\;\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2\right)^{-1/2}.$$

В дальнейшем будет полезной несколько иная форма записи эмпирического корреляционного момента:

$$K_{\mathrm{XY}} \;=\; \overline{xy} - \overline{x}\cdot\overline{y}. \tag{6.30}$$

Аналогично можно записать для выборочных дисперсий s_x^2 и s_y^2, что

$$s_x^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \;=\; \overline{x^2} - \overline{x}^2, \tag{6.31a}$$

$$s_y^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 \;=\; \overline{y^2} - \overline{y}^2. \tag{6.31b}$$

С учетом этих выражений получаем еще одну формулу для вычисления эмпирического коэффициента корреляции:

$$\hat{\rho} \;=\; r \;=\; \frac{K_{\mathrm{XY}}}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} \;=\; (\overline{xy} - \overline{x}\cdot\overline{y})\left(\left[\,\overline{x^2} - \overline{x}^2\,\right]\left[\,\overline{y^2} - \overline{y}^2\,\right]\right)^{-1/2}. \tag{6.32}$$

Система нормальных уравнений в этих обозначениях принимает вид (регрессия Y на X)

$$\begin{cases} b_0 + b_1\overline{x} \;=\; \overline{y}, \\ \\ b_0\overline{x}^2 + b_1\overline{x^2} \;=\; \overline{xy}. \end{cases} \tag{6.33}$$

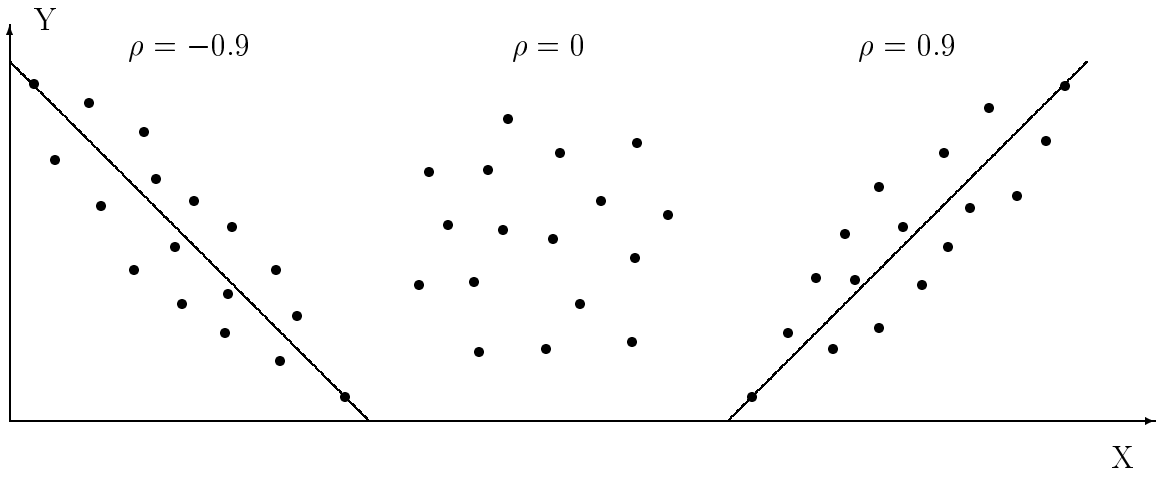


Рисунок 6.6 — Три примера корреляционного поля ($\rho = -0,9$; 0,0; 0,9; приведены линии регрессии)

Решая эту систему, находим

$$b_0 \;=\; \overline{y} - b_1\overline{x}, \qquad b_1 \;=\; \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2} \;=\; r\,\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}. \tag{6.34}$$

Из системы (6.33) следует, что прямая регрессии Y на X проходит через точку $C = (\overline{x}, \overline{y})$. Следовательно, уравнение прямой регрессии Y на X можно записать в виде

$$\overline{y}_x - \overline{y} \;=\; r\,\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}\,(x - \overline{x}). \tag{6.35}$$

Поступая аналогичным образом, можно показать, что система нормальных уравнений регрессии X на Y имеет вид

$$\begin{cases} a_0 + a_1\overline{y} \;=\; \overline{x}, \\ a_0\overline{y}^2 + a_1\overline{y^2} \;=\; \overline{xy}. \end{cases} \tag{6.36}$$

Решая эту систему, находим

$$a_0 \;=\; \overline{x} - a_1\overline{y}, \qquad a_1 \;=\; \frac{\overline{xy} - \overline{y}\cdot\overline{x}}{\overline{y^2} - \overline{y}^2} \;=\; r\,\frac{s_{\mathrm{X}}}{s_{\mathrm{Y}}}. \tag{6.37}$$

Следовательно, уравнение прямой регрессии X на Y можно записать в виде

$$\overline{x}_y - \overline{x} \;=\; r\,\frac{s_{\mathrm{X}}}{s_{\mathrm{Y}}}\,(y - \overline{y}). \tag{6.38}$$

Найденные формулы удобны с вычислительной точки зрения. Действительно, чтобы записать уравнение регрессии Y на X или X на Y, достаточно найти точечные оценки нормальной двумерной случайной величины (X, Y): $\overline{x}$, $\overline{y}$, s_x, s_y, r.

Из формул (6.35) и (6.38) вытекает важное соотношение, связывающее коэффициенты регрессии a_1, b_1 и коэффициент корреляции:

$$r^2 \;=\; a_1 b_1 \qquad \text{или} \qquad r \;=\; \pm\sqrt{a_1 b_1}.$$

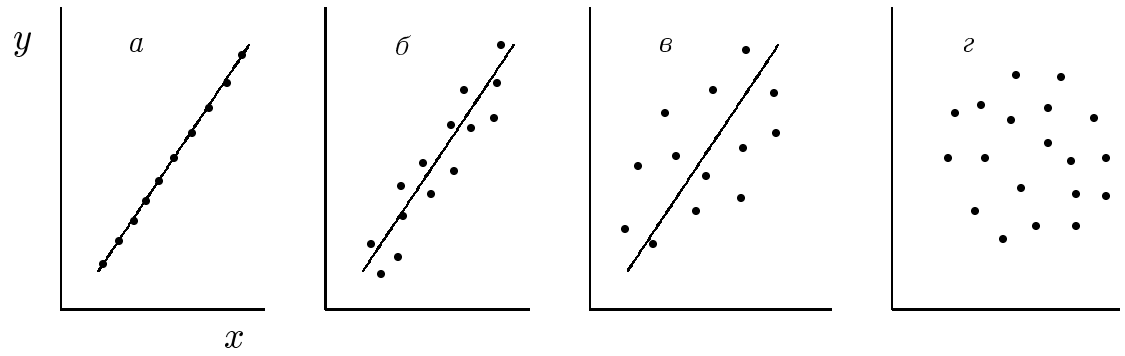


Рисунок 6.7 — Четыре примера положительной корреляционной зависимости: *а*) практически полная корреляция ($\rho = 1$); *б*) высокая степень корреляции ($\rho \approx 0,8$); *в*) умеренная корреляция ($\rho \approx 0,4$); *г*) отсутствие корреляции ($\rho \approx 0,0$)

Знак коэффициента корреляции совпадает со знаком коэффициентов регрессии.

Если в формулах (6.35) и (6.38) коэффициент регрессии Y на X равен $b_1 = rs_{\mathrm{Y}}/s_{\mathrm{X}}$ и, соответственно, коэффициент регрессии X на Y равен $a_1 = rs_{\mathrm{X}}/s_{\mathrm{Y}}$ и они *положительны* (*отрицательны*), то говорят, что направление зависимости Y на X *положительно* (*отрицательно*). Это значит, что переменные Y и X одновременно *возрастают* (*убывают*).

Коэффициенты a_1 и b_1 не позволяют судить о степени связи между случайными величинами Y и X. Степень связи зависит от угла, образованного прямыми регрессии. Чем меньше угол между прямыми регрессии, тем теснее связь между случайными величинами Y и X. При слиянии двух прямых регрессии в одну имеет место линейная функциональная зависимость между Y и X.

Угол между линиями регрессии определяется уравнениями (6.35) и (6.38). Так как $\mathrm{tg}\alpha = b_1 = rs_{\mathrm{Y}}/s_{\mathrm{X}}$ и $\mathrm{tg}\beta = a_1 = rs_{\mathrm{X}}/s_{\mathrm{Y}}$, то из рис. 6.8 вытекает, что $\mathrm{tg}\theta = \mathrm{tg}\left(\frac{\pi}{2} - \alpha - \beta\right) = \mathrm{ctg}\,(\alpha + \beta)$, и поэтому

$$\mathrm{tg}\theta = \frac{1 - \mathrm{tg}\alpha \cdot \mathrm{tg}\beta}{\mathrm{tg}\alpha + \mathrm{tg}\beta} = \frac{1 - r^2}{r} \frac{s_{\mathrm{X}} s_{\mathrm{Y}}}{s_{\mathrm{X}}^2 + s_{\mathrm{Y}}^2}. \tag{6.39}$$

В вычислительной практике принято пользоваться коэффициентом корреляции $\hat{\rho} = r = \pm\sqrt{a_1 b_1} = K_{\mathrm{XY}}/(s_{\mathrm{X}} s_{\mathrm{Y}})$ в качестве показателя степени связи.

Возможные значения коэффициента корреляции $\hat{\rho}$ ограничены интервалом $[-1;\, 1]$. Внутри этого интервала величина $\hat{\rho}$ определяется из следующих условий:

1) если между переменными X и Y существует линейная положительная связь, то коэффициент корреляции $\hat{\rho} = 1$;

2) если между переменными X и Y существует линейная отрицательная связь, то коэффициент корреляции $\hat{\rho} = -1$;

3) при отсутствии зависимости между переменными X и Y коэффициент корреляции $\hat{\rho} = 0$;

4) во всех остальных случаях $-1 < \hat{\rho} < 1$.

Справедливо и обратное утверждение: чем ближе по модулю коэффициент корреляции $\hat{\rho}$ к единице, тем сильнее *линейная зависимость* между случайными величинами X и Y, а чем ближе ρ к нулю, тем эта зависимость слабее.

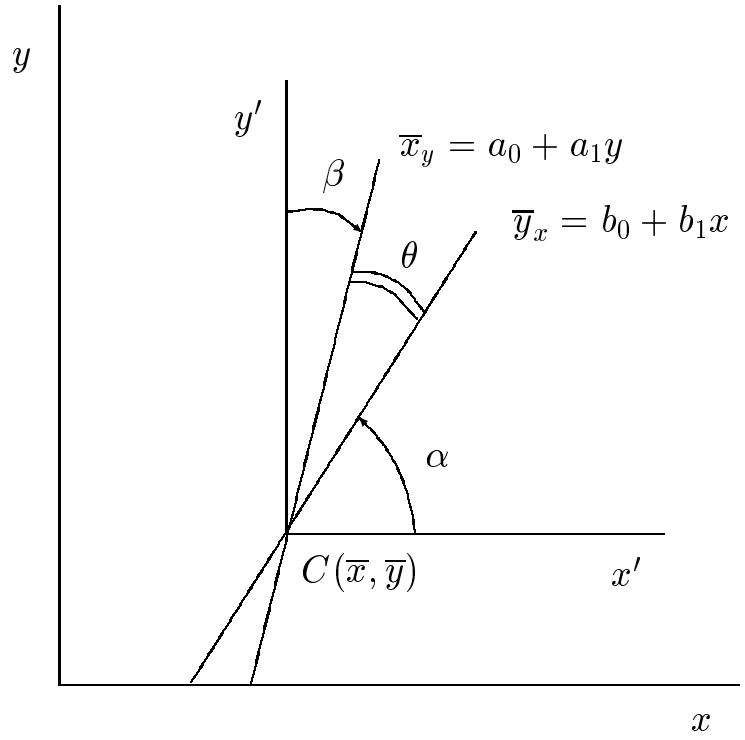


Рисунок 6.8 — К вычислению угла θ между линиями регрессии

Напомним, что коэффициент корреляции $\hat{\rho}$ характеризует степень линейной зависимости, поэтому даже при $\hat{\rho} = 0$ может оказаться, что между X и Y существует функциональная связь нелинейного вида.

Угловой коэффициент прямой регрессии Y на X называют *коэффициентом регрессии* Y на X и обозначают через $\rho_{\mathrm{Y|X}}$. Аналогично угловой коэффициент прямой регрессии X на Y называют *коэффициентом регрессии* X на Y и обозначают через $\rho_{\mathrm{X|Y}}$. Эти коэффициенты могут быть вычислены по формулам

$$\rho_{\mathrm{Y|X}} = \frac{1}{\sigma_{\mathrm{X}}^2}\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\right] = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sigma_{\mathrm{X}}^2}, \tag{6.40a}$$

$$\rho_{\mathrm{X|Y}} = \frac{1}{\sigma_{\mathrm{Y}}^2}\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\right] = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sigma_{\mathrm{Y}}^2}. \tag{6.40b}$$

В этих обозначениях уравнения прямых регрессий принимают вид

$$y - \overline{y} = \rho_{\mathrm{Y|X}}(x - \overline{x}), \qquad x - \overline{x} = \rho_{\mathrm{X|Y}}(y - \overline{y}). \tag{6.41}$$

Коэффициентом линейной корреляции признаков X и Y называют величину

$$r = r(\mathrm{X},\mathrm{Y}) = \pm\sqrt{\rho_{\mathrm{X|Y}}\,\rho_{\mathrm{Y|X}}} = \tag{6.42}$$

$$= \frac{1}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\right].$$

Квадрат коэффициента линейной корреляции дает *коэффициент детерминации* k_{det}, который измеряет долю вариации компоненты X, объясняемую влиянием Y, и наоборот.

$$k_{\text{det}} \;=\; r^2 \;=\; \rho_{\text{X|Y}}\,\rho_{\text{Y|X}}. \tag{6.43}$$

Если линии регрессии отличны от прямых, то коэффициент линейной корреляции не дает полного представления о силе связи между соответствующими признаками X и Y.

В этом случае за меру зависимости признака Y от признака X принимают *корреляционное отношение* $\eta_{\text{Y|X}}$, которое является отношением среднего квадратического отклонения условных средних $\overline{y}_x$ к среднему квадратическому отклонению признака Y

$$\eta_{\text{Y|X}} \;=\; \frac{\sigma\left(\overline{\text{Y}}_x\right)}{\sigma_{\text{Y}}} \;=\; \frac{\left(\sum\limits_{j=1}^{k} m_{x_j}\left(\overline{\text{Y}}_{x_j}-\overline{\text{Y}}\right)^2\right)^{1/2}}{\left(\sum\limits_{j=1}^{n} m_{y_j}\left(y_j-\overline{\text{Y}}\right)^2\right)^{1/2}}. \tag{6.44a}$$

Аналогично вводится *корреляционное отношение* $\eta_{\text{X|Y}}$ признака X от признака Y

$$\eta_{\text{X|Y}} \;=\; \frac{\sigma\left(\overline{\text{X}}_y\right)}{\sigma_{\text{X}}} \;=\; \frac{\left(\sum\limits_{j=1}^{n} m_{y_j}\left(\overline{\text{X}}_{y_j}-\overline{\text{X}}\right)^2\right)^{1/2}}{\left(\sum\limits_{j=1}^{k} m_{x_j}\left(x_j-\overline{\text{X}}\right)^2\right)^{1/2}}. \tag{6.44b}$$

Возможные значения корреляционных отношений $\eta_{\text{Y|X}}$ и $\eta_{\text{X|Y}}$ ограничены интервалом $[0;\,1]$. Внутри этого интервала эти величины определяются из следующих условий:

1) если $\eta_{\text{Y|X}} = 0$ (или $\eta_{\text{X|Y}} = 0$), то признаки Y и X не коррелируют;

2) если $\eta_{\text{Y|X}} = 1$, то признак Y связан с признаком X функциональной связью, $y = f(x)$;

3) выполняются неравенства $\eta_{\text{Y|X}} \geq |r(\text{X},\text{Y})|$ и $\eta_{\text{X|Y}} \geq |r(\text{X},\text{Y})|$.

6.5. Проверка гипотез о значимости коэффициента корреляции

Основная цель корреляционного анализа состоит в выявлении связи между случайными величинами X и Y и, если окажется, что эта связь существует, – определение степени близости этой связи к функциональной. На практике при обработке экспериментальных данных коэффициент корреляции генеральной совокупности ρ неизвестен. По результатам эксперимента на основании выборки находится его точечная оценка (приближенное значение) – *выборочный коэффициент корреляции* $r = \hat{\rho}$.

Если выборочный коэффициент корреляции r оказался равным нулю, это еще не значит, что соответствующий коэффициент корреляции ρ равен нулю, а случайные величины X и Y независимы (в случае их нормальности), и наоборот, если $r \neq 0$, то это еще не значит, что $\rho \neq 0$, а случайные величины могут оказаться независимыми.

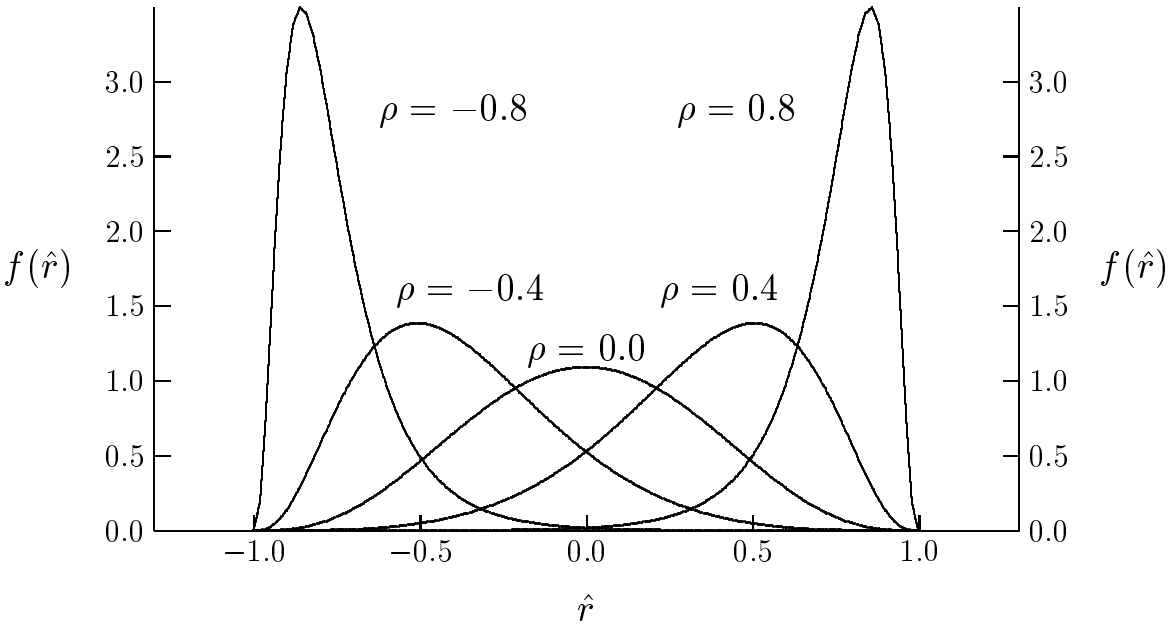


Рисунок 6.9 — Семейство из 5 плотностей распределения вероятностей $f(\hat{r}) = f(\hat{r}; \rho, n)$; (объем выборки равен $n = 10$; зависимости приведены для идеальных коэффициентов корреляции, равных $\rho =$ –0,8; –0,4; 0,0; 0,4; 0,8)

Для того чтобы на основе статистического анализа ответить на вопрос, находятся ли случайные величины в корреляционной зависимости, необходимо проверить нулевую гипотезу $\{H_0 : \rho = 0\}$ против одной из альтернативных гипотез $\{H_a : \rho \neq 0\}$, или $\{H_a : \rho > 0\}$, или $\{H_a : \rho < 0\}$.

Ясно, что для проверки нулевой гипотезы необходимо знать закон распределения выборочного коэффициента корреляции.

Плотность распределения вероятностей эмпирического (выборочного) коэффициента корреляции r в случае, если выборка получена из совокупности с двумерным нормальным распределением, зависит от объема выборки n и коэффициента корреляции ρ генеральной совокупности и имеет вид:

$$f(\hat{r}) = \frac{n-2}{\pi}(1-\rho^2)^{(n-1)/2}(1-\hat{r}^2)^{(n-4)/2}\int\limits_0^1 \frac{x^{n-2}}{(1-\rho\hat{r}x)^{n-1}}\frac{dx}{\sqrt{1-x^2}}. \qquad (6.45)$$

Для выборок из генеральной совокупности, в которой величины X и Y нормальны и независимы, т.е. $\rho = 0$, плотность распределения вероятностей эмпирического коэффициента корреляции определяется выражением

$$f(\hat{r}) = \frac{1}{\pi}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}(1-\hat{r}^2)^{(n-4)/2}. \qquad (6.46)$$

Формулы (6.45) и (6.46) называют *распределением Крамера*.

На рис. 6.9 и 6.10 приведены графики дифференциальной и интегральной функций распределения выборочного коэффициента корреляции при $n = 10$ и $\rho = -0.8$, –0,4; 0,0; 0,4; 0,8.

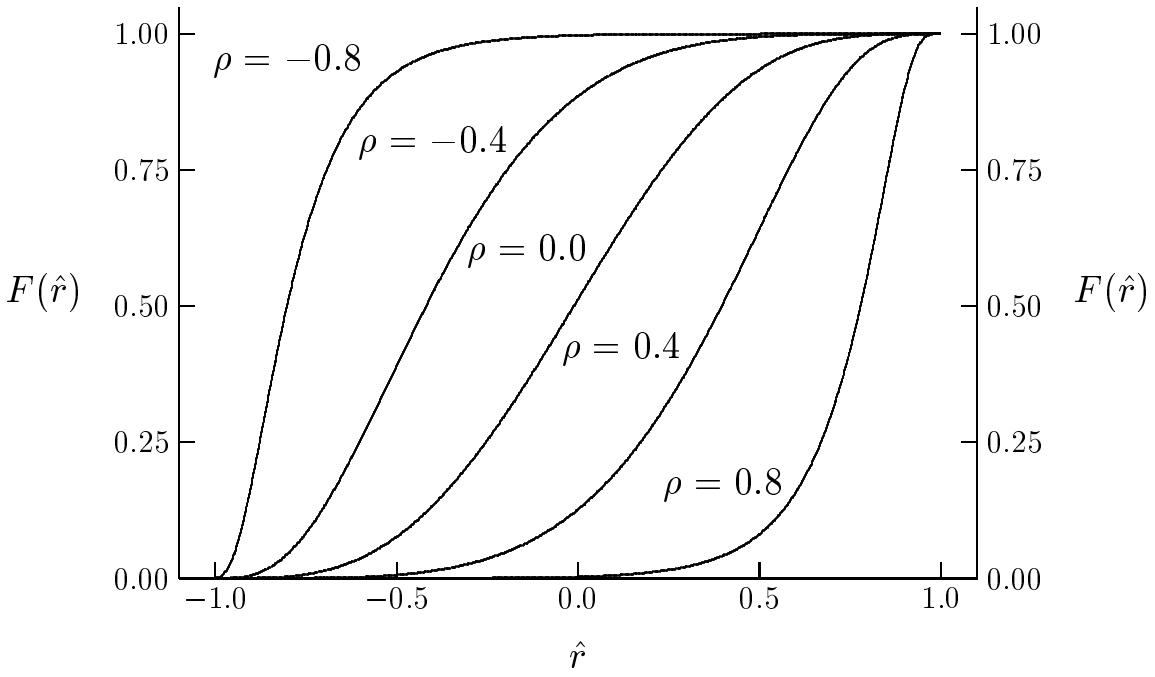


Рисунок 6.10 — Семейство из 5 интегральных распределений $F(\hat{r}) = F(\hat{r}; \rho, n)$ (объем выборки равен $n = 10$; зависимости приведены для идеальных коэффициентов корреляции, равных $\rho =$ –0,8; –0,4; 0,0; 0,4; 0,8)

Построение интервальных оценок для коэффициента корреляции генеральной совокупности методами, непосредственно использующими формулы (6.45)–(6.46), применяется достаточно редко. В практике для построения интервальных оценок для ρ пользуются либо номограммами (в специализированных математических или статистических пакетах), основанными на этих формулах, либо применяют специальные преобразования эмпирического коэффициента корреляции, которые позволяют свести распределение некоторой функции от выборочного коэффициента корреляции к хорошо изученным распределениям, например, нормальному или распределению Стьюдента.

Проверка значимости выборочного коэффициента корреляции r, т.е. проверка, какую величину выборочного коэффициента корреляции следует считать достаточной для статистически обоснованного вывода о наличии корреляционной связи между исследуемыми переменными X и Y, основывается на следующих трех математических моделях, которые являются обобщением исходных предпосылок о двумерной генеральной совокупности (X, Y).

Модель 1.

Применяется для проверки гипотезы об отсутствии корреляционной связи между исследуемыми переменными, т.е. для проверки гипотезы $\{H_0 : \rho = 0\}$.

Исходные предположения и ограничения:

а) двумерный закон распределения исследуемых переменных (X, Y) в генеральной совокупности предполагается нормальным;

б) объем выборки n может быть любым.

Для проверки значимости выборочного коэффициента корреляции вычисляется статистика

$$t \;=\; r\sqrt{(n-2)(1-r^2)}, \tag{6.47}$$

имеющая распределение Стьюдента с $\nu = n-2$ степеней свободы.

Для проверки нулевой гипотезы $\{H_0 : \rho = 0\}$ находят по таблицам распределения Стьюдента по фиксированному уровню значимости α и числу степеней свободы $\nu = n-2$ критическое значение $t_{\alpha/2;\,n-2}$, удовлетворяющее условию $\Pr(|t| \geq t_{\alpha/2;\,n-2}) = \alpha$.

Если

$$|t_{\text{набл}}| \;\geq\; t_{\alpha/2;\,n-2}, \tag{6.48a}$$

то нулевую гипотезу об отсутствии линейной зависимости между переменными X и Y следует отвергнуть.

Если же

$$|t_{\text{набл}}| \;<\; t_{\alpha/2;\,n-2}, \tag{6.48b}$$

то нет оснований отвергать нулевую гипотезу о некоррелированности переменных X и Y.

Модель 2.

Применяется для проверки гипотезы о силе корреляционной связи между переменными X и Y, иначе говоря, для проверки нулевой гипотезы о том, что коэффициент корреляции ρ генеральной совокупности равен некоторому фиксированному числу, т.е. $\{H_0 : \rho = \rho_0\}$. (Если $\rho_0 = 0$, то проверяется гипотеза о некоррелированности переменных X и Y.)

Исходные предположения и ограничения:

а) двумерный закон распределения исследуемых переменных (X, Y) в генеральной совокупности предполагается нормальным;

б) объем выборки n достаточно велик $(n \geq 50)$;

в) вычисленное значение выборочного коэффициента корреляции невелико $(|\rho| \leq 0,5)$.

Если эти предположения выполнены, то выборочный коэффициент корреляции r имеет приближенно нормальное распределение с математическим ожиданием, равным коэффициенту корреляции генеральной совокупности ρ и дисперсией, равной $\sigma_r = (1-r^2)/\sqrt{n}$. Это непосредственно можно заметить, анализируя график плотности распределения выборочного коэффициента корреляции (рис. 6.9). Отсюда следует, что если нулевая гипотеза $\{H_0 : \rho = \rho_0\}$ верна, то статистика

$$u \;=\; \frac{r-\rho_0}{\sqrt{(1-\rho_0^2)/\sqrt{n}}} \tag{6.49}$$

имеет приближенно нормальное распределение с нулевым математическим ожиданием и дисперсией, равной единице. Выберем $u_{\alpha/2}$ – критическое значение стандартизованной нормальной случайной величины, удовлетворяющей условию $\Pr(|u| \geq u_{\alpha/2}) = \alpha$.

Если

$$|u_{\text{набл}}| \;\geq\; u_{\alpha/2}, \tag{6.50a}$$

то гипотеза $\{H_0 : \rho = \rho_0\}$ отвергается.

Если же

$$|u_{\text{набл}}| \;<\; u_{\alpha/2}, \tag{6.50b}$$

то нет основания отвергать нулевую гипотезу.

Применение этой модели позволяет также находить приближенные доверительные интервалы для коэффициента корреляции ρ генеральной совокупности по формуле

$$\left(r - u_{\alpha/2}\sqrt{(1-\rho_0^2)/\sqrt{n}}\right) \;<\; \rho \;<\; \left(r + u_{\alpha/2}\sqrt{(1-\rho_0^2)/\sqrt{n}}\right). \tag{6.51}$$

Если окажется, что вычисленный доверительный интервал не накрывает значение $\rho = \rho_0$, то гипотеза $\{H_0 : \rho = \rho_0\}$ отвергается, в противном случае данные эксперимента не позволяют отвергнуть нулевую гипотезу.

Модель 3.

Исходные предположения и ограничения:

а) двумерный закон распределения исследуемых переменных (X, Y) в генеральной совокупности предполагается нормальным;

б) объем выборки n достаточно велик $(n \geq 50)$.

Для проверки гипотезы о силе корреляционной связи вычисляется статистика (преобразование Фишера)

$$z \;=\; \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right), \qquad \text{т.е.} \qquad r \;=\; \mathrm{th}z \;=\; \frac{\mathrm{e}^{2z}-1}{\mathrm{e}^{2z}+1}. \tag{6.52}$$

Можно показать, что распределение статистики z достаточно хорошо аппроксимируется нормальным распределением с математическим ожиданием

$$\mathrm{M[Z]} \;=\; \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) + \frac{\rho_0}{2(n-1)} \tag{6.53}$$

и не зависящей от ρ дисперсией

$$\mathrm{D[Z]} \;=\; \frac{1}{n-3}. \tag{6.54}$$

Статистика

$$u \;=\; \left(1,1513\lg\frac{1+r}{1-r} - 1,1513\lg\frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{2(n-1)}\right)\sqrt{n-3} \tag{6.55}$$

имеет асимптотически нормальное распределение с нулевым математическим ожиданием и дисперсией, равной единице. Проверка гипотезы $\{H_0 : \rho = \rho_0\}$ проводится по той же схеме, что и в Модели 2, т.е. по формуле (6.55) вычисляется эмпирическое значение $u_{\text{набл}}$ статистики u. Пользуясь таблицей функции Лапласа, по фиксированному уровню значимости α находят критическое значение $u_{\alpha/2}$, удовлетворяющее условию $\Pr(|u| \geq u_{\alpha/2}) = \alpha$.

Если вычисленное значение статистики удовлетворяет неравенству

$$|u_{\text{набл}}| \;\geq\; u_{\alpha/2}, \tag{6.56a}$$

то гипотезу $\{H_0 :\; \rho = \rho_0\}$ следует отвергнуть.

Если же

$$|u_{\text{набл}}| \;<\; u_{\alpha/2}, \tag{6.56b}$$

то нет основания отвергать нулевую гипотезу.

Замечание. Модель 3 является наиболее общей, так как может применяться при любых значениях n, ρ и r. Но когда проверяется нулевая гипотеза $\{H_0 : \rho = 0\}$, то целесообразнее использовать Модель 1. Если же вычисленное значение выборочного коэффициента корреляции не является очень большим ($|r| < 0,5$) и объем выборки достаточно велик ($n \geq 50$), то следует выбрать Модель 2.

6.6. Оценка точности нахождения точечных оценок коэффициентов линейного уравнения регрессии

После нахождения эмпирического уравнения регрессии Y на X

$$\overline{y}_x \;=\; b_0 + b_1 x$$

подсчитывают среднюю квадратичную ошибку $s_e \equiv s_{y.x}$, характеризующую степень рассеивания экспериментальных точек вокруг линии регрессии, т.е. степень рассеивания зависимой переменной Y, очищенной от влияния переменной X:

$$s_e \;\equiv\; s_{y.x} \;=\; \left(\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \overline{y}_x)^2\right)^{1/2}. \tag{6.57}$$

Средние квадратичные ошибки σ_{b_0} и σ_{b_1} определения коэффициентов b_0 и b_1 определяются выражениями

$$\sigma^2_{b_0} \;=\; \sigma^2_{y.x}\left(\frac{1}{n} + \overline{x^2}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{-1}\right), \tag{6.58a}$$

$$\sigma^2_{b_1} \;=\; \sigma^2_{y.x}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{-1}. \tag{6.58b}$$

Заменяя в этих формулах неизвестную дисперсию $\sigma^2_{y.x}$ её несмещенной оценкой $s^2_{y.x}$, после преобразований получим эмпирические дисперсии коэффициентов β_0 и β_1 линейного уравнения регрессии $s_e \equiv s_{y.x}$

$$\sigma^2_{b_0} \;=\; \left(s^2_{y.x}\sum_{i=1}^{n}x_i^2\right)\left[n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right]^{-1}, \tag{6.59}$$

$$\sigma^2_{b_1} \;=\; \left(ns^2_{y.x}\right)\left[n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right]^{-1}. \tag{6.60}$$

Квадратный корень из этих значений называется *средней квадратичной ошибкой нахождения оценок* b_0 *и* b_1. Средняя квадратичная ошибка указывает, насколько в среднем оценки коэффициентов b_0 и b_1 отличаются от модельных коэффициентов регрессии β_0 и β_1.

Методы регрессионного анализа оказываются содержательными и корректными лишь при соблюдении некоторых требований, предъявляемых к условиям сбора статистического материала в эксперименте.

Если гипотеза о нормальном законе распределения системы случайных величин (X, Y) справедлива, то:

а) математическое ожидание $\mathrm{M}[\mathrm{Y}|\mathrm{X}= x] = \beta_0 + \beta_1 x$;

б) условная дисперсия $\mathrm{D}[\mathrm{Y}|\mathrm{X}= x]$ постоянна для всех значений x и равна $\sigma^2_{y.x} = \sigma^2_y(1-\rho^2)$;

в) распределение $f_{\mathrm{Y}}(y|x)$ нормальное;

г) наблюдения $\{(x_i, y_i)\}$, $(i = 1, 2, \ldots, n)$ независимы.

Если оказывается возможным принять, что нормальная модель ряда наблюдений $\{(x_i, y_i)\}$ хорошо отражает закономерности исследуемой двумерной величины (X, Y), т.е. перечисленные требования а) – г) выполняются, то в этом случае оценки коэффициентов регрессии b_0 и b_1 в практических расчетах описывают (приближенно) нормально распределенными случайными величинами с математическими ожиданиями β_0, β_1 и дисперсиями $\sigma^2_{b_0}$, $\sigma^2_{b_1}$. Тогда если $\sigma^2_{y.x}$ известна, то статистики $u_0 = (b_0 - \beta_0)/\sigma^2_{y.x}$ и $u_1 = (b_1 - \beta_1)/\sigma^2_{y.x}$ распределены по нормальному закону с нулевым математическим ожиданием и дисперсией, равной единице.

Статистики u_j, $j = 0, 1$, можно применять для построения интервальных оценок коэффициентов. В этом случае искомые доверительные интервалы для коэффициентов истинного уравнения регрессии будут находиться по заданной доверительной вероятности $p = 1 - \alpha$ по формуле

$$\Pr\left(-u_{\alpha/2} < \frac{b_j - \beta_j}{\sigma_{y.x}\sqrt{c_{jj}}} < u_{\alpha/2}\right) = 1 - \alpha, \qquad j = 0, 1, \tag{6.61}$$

где $u_{\alpha/2}$ – квантиль нормального распределения $N(0;\, 1)$;

$$c_{00} = \frac{\sum\limits_{i=1}^{n} x_i^2}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}, \qquad c_{11} = \frac{n}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}. \tag{6.62}$$

Преобразуя выражение (6.61), получим $(j = 0, 1)$

$$\Pr\left(b_j - u_{\alpha/2}\sigma_{y.x}\sqrt{c_{jj}} < \beta_j < b_j + u_{\alpha/2}\sigma_{y.x}\sqrt{c_{jj}}\right) = 1 - \alpha. \tag{6.63}$$

При проведении экспериментальных работ величина $\sigma_{y.x}$ неизвестна. Предположим, что с помощью метода наименьших квадратов была найдена несмещенная оценка $\hat{\sigma}_{y.x}$ величины $\sigma_{y.x}$ по формуле

$$\hat{\sigma}_{y.x} = s_{y.x} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - b_0 - b_1 x)^2. \tag{6.64}$$

Тогда статистика ($j = 0, 1$)

$$t \;=\; \frac{b_j - \beta_j}{s_{b_j}} \;=\; \frac{b_j - \beta_j}{s_{y.x}\sqrt{c_{jj}}} \tag{6.65}$$

имеет распределение Стьюдента с $\nu = n - 2$ степенями свободы, которую можно использовать для построения интервальных оценок коэффициентов линейного уравнения регрессии.

Рассмотрим $100(1-\alpha)\,\%$-й доверительный интервал для коэффициентов β_0 и β_1. Для его построения по таблицам t-распределения Стьюдента по числу степеней свободы $\nu = n - 2$ и доверительной вероятности $p = 1 - \alpha$ находим значение $t_{\alpha/2;\,n-2}$, удовлетворяющее условию

$$\Pr\left(|t| < t_{\alpha/2;\,n-2}\right) \;=\; 1 - \alpha. \tag{6.66}$$

После простых преобразований получим

$$\Pr\left(-t_{\alpha/2;\,n-2} < \frac{b_0 - \beta_0}{s_{y.x}\sqrt{c_{00}}} < t_{\alpha/2;\,n-2}\right) \;=\; 1 - \alpha, \tag{6.67}$$

$$\Pr\left(-t_{\alpha/2;\,n-2} < \frac{b_1 - \beta_1}{s_{y.x}\sqrt{c_{11}}} < t_{\alpha/2;\,n-2}\right) \;=\; 1 - \alpha. \tag{6.68}$$

В результате преобразования обоих неравенств в круглых скобках найдем для $100(1-\alpha)\,\%$-го доверительного интервала для коэффициентов линейной регрессии β_0 и β_1:

$$\left(b_0 - t_{\alpha/2;\,n-2}s_{y.x}\sqrt{c_{00}}\right) \;<\; \beta_0 \;<\; \left(b_0 + t_{\alpha/2;\,n-2}s_{y.x}\sqrt{c_{00}}\right), \tag{6.69}$$

$$\left(b_1 - t_{\alpha/2;\,n-2}s_{y.x}\sqrt{c_{11}}\right) \;<\; \beta_1 \;<\; \left(b_1 + t_{\alpha/2;\,n-2}s_{y.x}\sqrt{c_{11}}\right). \tag{6.70}$$

Найденными выражениями часто пользуются при рассмотрении задач прогнозирования временных зависимостей (вместе с предъявлением оценки прогнозных погрешностей).

6.7. Линейный регрессионный анализ между двумя переменными

В процессе выполнения практических работ и решения задач линейного регрессионного анализа между двумя переменными следует помнить о соответствии между характеристиками, которые относятся к теоретико-вероятностным моделям и отвечающим им статистическим величинам, к их выборочным аналогам.

Обрабатывая статистические данные, необходимо иметь в виду соответствие между характеристиками генеральной совокупности и выборочной совокупности.

Основные из указанных величин приведены в таблице-сводке.

Сводка числовых характеристик генеральной и выборочной совокупностей

Генеральная совокупность	Выборочная совокупность
$\mathrm{M[X]} = m_{\mathrm{X}} = \int\int x f(x,y)\,dx\,dy = \int x f(x)\,dx$	$\overline{x} = \frac{1}{n}\sum\limits_i x_i$
$\mathrm{M[Y]} = m_{\mathrm{Y}} = \int\int y f(x,y)\,dx\,dy = \int y f(y)\,dy$	$\overline{y} = \frac{1}{n}\sum\limits_j y_j$
$\sigma_{\mathrm{X}}^2 = \int\int (x-m_{\mathrm{X}})^2 f(x,y)\,dx\,dy = \int (x-m_{\mathrm{X}})^2 f(x)\,dx$	$s_{\mathrm{X}}^2 = \frac{1}{n}\sum\limits_i (x_i-\overline{x})^2 = \overline{x^2}-\overline{x}^2$
$\sigma_{\mathrm{Y}}^2 = \int\int (y-m_{\mathrm{Y}})^2 f(x,y)\,dx\,dy = \int (y-m_{\mathrm{Y}})^2 f(y)\,dy$	$s_{\mathrm{Y}}^2 = \frac{1}{n}\sum\limits_j (y_j-\overline{y})^2 = \overline{y^2}-\overline{y}^2$
$\mu_{\mathrm{XY}} = \int\int xy f(x,y)\,dx\,dy - \mathrm{M[X]M[Y]}$	$K_{\mathrm{XY}} = \frac{1}{n}\sum\limits_i\sum\limits_j (x_i-\overline{x})(y_j-\overline{y})$
$\rho = \frac{\mu_{\mathrm{XY}}}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}$	$r = \frac{K_{\mathrm{XY}}}{s_{\mathrm{X}}s_{\mathrm{Y}}}$
Модельное уравнение регрессии Y на X $\mathrm{M[Y\|X}=x] = m_{\mathrm{Y}} + \rho\frac{\sigma_{\mathrm{Y}}}{\sigma_{\mathrm{X}}}(x-m_{\mathrm{X}})$	Эмпирическое уравнение регрессии Y на X $\overline{y}_x = \overline{y} + r\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}(x-\overline{x})$
Модельное уравнение регрессии X на Y $\mathrm{M[X\|Y}=y] = m_{\mathrm{X}} + \rho\frac{\sigma_{\mathrm{X}}}{\sigma_{\mathrm{Y}}}(y-m_{\mathrm{Y}})$	Эмпирическое уравнение регрессии X на Y $\overline{x}_y = \overline{x} + r\frac{s_{\mathrm{X}}}{s_{\mathrm{Y}}}(y-\overline{y})$

Пример 1

В таблице приведены результаты $n = 11$ измерений значений x_i номиналов длин моделей (признак X) и ширин моделей y_i (признак Y):

i	1	2	3	4	5	6	7	8	9	10	11
x_i, *мм*	0,90	1,22	1,32	0,77	1,30	1,20	1,32	0,95	1,45	1,30	1,20
y_i, *мм*	–0,30	0,10	0,70	–0,28	–0,25	0,02	0,37	–0,70	0,55	0,35	0,32

Для этих данных предполагается, что между признаками X и Y существует линейная регрессионная зависимость $\mathrm{M}[\mathrm{Y}|\mathrm{X}=x]=\beta_0+\beta_1 x$.

Требуется:

1) оценить параметры β_0 и β_1 модельного уравнения регрессии методом наименьших квадратов;

2) найти средние квадратические ошибки коэффициентов найденного эмпирического уравнения регрессии;

3) построить 95%-е доверительные интервалы для коэффициентов линейного уравнения регрессии β_0 и β_1;

4) пользуясь эмпирическим уравнением регрессии, найти точечную оценку отклонения от номинального размера длины модели, если ширина модели отклоняется от номинального размера на величину $x_*=1,1$ *мм*;

5) вычислить коэффициент корреляции и коэффициент детерминации. Объяснить смысл коэффициента детерминации.

Решение

1. Сначала вычислим вспомогательные суммы.

Последовательно находим:

$$\sum_{i=1}^{n} x_i = 12,93; \qquad \sum_{i=1}^{n} y_i = 0,88;$$

$$\sum_{i=1}^{n} x_i y_i = 1,7193; \qquad \sum_{i=1}^{n} x_i^2 = 15,6411;$$

Найдем выборочные средние:

$$\overline{x} \;=\; \frac{1}{11}\cdot 12,93 \;=\; 1,175; \qquad \overline{y} \;=\; \frac{1}{11}\cdot 0,88 \;=\; 0,080.$$

Найдем оценки b_0 и b_1 коэффициентов истинного уравнения регрессии по формулам:

$$b_1 \;=\; \frac{n\left(\sum_{i=1}^{n} x_i y_i\right)-\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right)-\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{11\cdot 1,7193-12,93\cdot 0,88}{11\cdot 15,6411-12,93^2} \;=\; 1,5479;$$

$$b_0 \;=\; \overline{y}-b_1\overline{x} \;=\; 0,08-1,5479\cdot 1,175 \;=\; -1,7388.$$

Следовательно, эмпирическое уравнение регрессии имеет следующий вид:

$$\overline{y}_x \;=\; -1,7388+1,5479\overline{x}.$$

График построенного эмпирического уравнения линейной регрессии приведен на рис. 6.11.

2. Для нахождения средних квадратических ошибок σ_{b_0} и σ_{b_1}, характеризующих точность найденных коэффициентов b_0 и b_1 эмпирического уравнения регрессии,

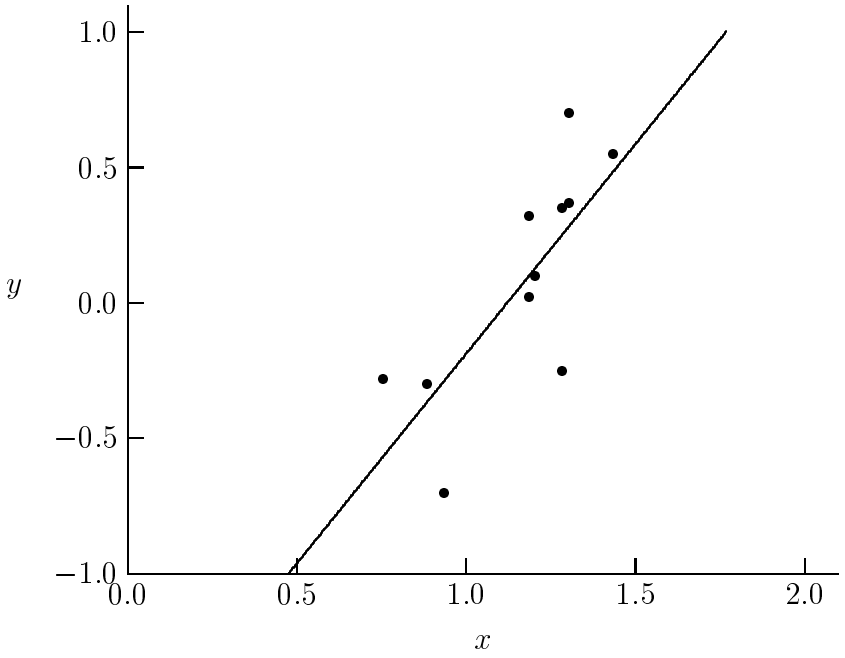


Рисунок 6.11 — Диаграмма рассеяния экспериментальных точек; указано эмпирическое уравнение регрессии $\overline{y}_x = -1,7388 + 1,5479\overline{x}$

вычислим вначале среднюю квадратическую ошибку, характеризующую рассеивание эмпирических точек вокруг линии регрессии. Для этого произведем оценку значений зависимой переменной по формуле $\overline{y}_x = -1,7388 + 1,5479x$ и вычислим отклонения $e_i = \overline{y}_x + 1,7388 - 1,5479x_i$. В результате найдем $\sum_{i=1}^{11} e_i^2 = 0,7561$.

Определим несмещенную оценку дисперсии зависимой переменной Y, очищенной от влияния переменной X, по формуле

$$\hat{\sigma}^2_{y.x} = s^2_{y.x} = \frac{1}{n-2}\sum_{i=1}^{n}(\overline{y}_x + 1,7388 - 1,5479x_i)^2 = \frac{0,7561}{9} = 0,08401.$$

Вычислим эмпирические дисперсии точечных оценок коэффициентов регрессии:

$$s^2_{b_0} = \left(s^2_{y.x}\sum_{i=1}^{n} x_i^2\right)\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]^{-1} =$$

$$= \frac{15,6411 \cdot 0,08401}{11 \cdot 15,6411 - 12,93^2} = 0,26997;$$

$$s^2_{b_1} = \left(ns^2_{y.x}\right)\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]^{-1} =$$

$$= \frac{11 \cdot 0,08401}{11 \cdot 15,6411 - 12,93^2} = 0,18986.$$

Следовательно, $s_{b_0} = 0,520$; $s_{b_1} = 0,436$.

3. Для построения 95%-х доверительных интервалов для коэффициентов линейного уравнения регрессии β_0 и β_1 по таблицам t–распределения Стьюдента по

числу степеней свободы $\nu = n - 2 = 9$ и доверительной вероятности $p = 1 - \alpha = 0,95$ находим критическое значение статистики (квантиль) $t_{\alpha/2;\,n-2} = t_{0,025;\,9} = 2,262$.

Используя формулу

$$\left(b_0 - t_{\alpha/2;\,n-2} s_{b_0}\right) \;<\; \beta_0 \;<\; \left(b_0 + t_{\alpha/2;\,n-2} s_{b_0}\right),$$

находим 95%-й доверительный интервал для коэффициента b_0

$$b_0 \pm t_{0,025;\,9}\, s_{b_0} \;=\; -1,7388 \pm (2,262 \cdot 0,520) \;=\; [-2,9150;\; -0,5626].$$

Следовательно, $-2,9150 < b_0 < -0,5626$.

Аналогично, используя формулу

$$\left(b_1 - t_{\alpha/2;\,n-2} s_{b_1}\right) \;<\; \beta_1 \;<\; \left(b_1 + t_{\alpha/2;\,n-2} s_{b_1}\right),$$

находим 95%–й доверительный интервал для коэффициента b_1

$$b_1 \pm t_{0,025;\,9}\, s_{b_1} \;=\; 1,5479 \pm (2,262 \cdot 0,436) \;=\; 1,5479 \pm 0,9885.$$

Следовательно, $0,5594 < b_1 < 2,5364$.

4. Найдем точечную оценку отклонения от номинального размера длины модели, если ширина модели отклоняется от номинального размера на величину $x_* = 1,1$ *мм*:

$$\overline{y}_{x^*} \;=\; 1,7394 - 1,5479 \cdot 1,10 \;=\; 0,036 \text{ мм}.$$

5. Вычислим коэффициент корреляции

$$r \;=\; \frac{n\left(\sum_{i=1}^n x_i y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{\sqrt{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2}\,\sqrt{n\left(\sum_{i=1}^n y_i^2\right) - \left(\sum_{i=1}^n y_i\right)^2}} \;=$$

$$= \;\frac{11 \cdot 1,7193 - 12,93 \cdot 0,88}{\sqrt{11 \cdot 15,6411 - 12,93^2}\,\sqrt{11 \cdot 1,8856 - 0,88^2}} \;=\; 0,7644.$$

Следовательно, коэффициент детерминации следующий:

$$k_{\text{det}} \;=\; r^2 \;=\; 0,584.$$

Полученный результат означает, что 58,4 % рассеивания зависимой переменной объясняется линейной регрессией Y на X, а 41,6 % рассеивания Y могут быть вызваны либо случайными ошибками эксперимента, либо тем, что линейная регрессионная модель плохо согласуется с экспериментальными данными.

Пример 2

В таблице приведены результаты $n = 50$ измерений значений x_i признака X и значений y_i признака Y.

Используя значения, приведенные в данной таблице, требуется:

1) составить корреляционную таблицу;

Таблица 6.2 — Данные к примеру 2

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
81	77	54	81	100	129	94	104	84	96
77	96	40	57	95	145	84	108	94	112
76	86	61	86	106	142	73	93	152	136
86	92	68	87	118	120	107	124	98	104
53	98	53	98	109	95	94	112	77	103
47	53	88	87	107	107	107	113	88	115
36	63	136	153	120	133	99	95	94	123
40	80	129	133	114	140	100	112	76	111
49	64	126	159	113	149	104	116	84	127
60	66	96	134	123	147	88	93	73	129

2) найти по данным корреляционной таблицы числовые характеристики выборки $\overline{x}$, $\overline{y}$, s_{X}, s_{Y}, K_{XY}, r;

3) построить корреляционное поле; по характеру расположения точек на корреляционном поле подобрать общий вид функции регрессии;

4) найти параметры эмпирической линейной функции регрессии Y на X и X на Y и построить их графики.

Решение

1. Составим корреляционную таблицу.

Примем для признака X следующие границы интервалов: (30–50), (50–70), ..., (130–150), а для признака Y – (50–70), (70–90), ..., (150–170). Таким образом, длины интервалов составляют $h_x = h_y = 20$. После этого подсчитываем количество экспериментальных точек, попадающих в прямоугольники, образованные границами интервалов. В результате получаем (см. ниже) корреляционную таблицу, в которой отмечены середины соответствующих интервалов.

Таблица 6.3 — Корреляционная таблица

Y	X						
	40	60	80	100	120	140	n_y
160					1	1	2
140				3	5	1	9
120			4	8	1		13
100		2	7	5			14
80	1	3	3				7
60	4	1					5
n_x	5	6	14	16	7	2	$n = 50$

2. Для определения характеристик искомых эмпирических уравнений регрессии найдем:

средние арифметические

$$\overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \;=\; 88,62; \qquad \overline{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i \;=\; 107,66;$$

выборочные дисперсии

$$s_x^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \;=\; 678,796; \qquad s_y^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 \;=\; 668,184;$$

выборочные средние квадратические отклонения

$$s_x \;=\; \sqrt{678,796} \;=\; 26,054; \qquad s_y \;=\; \sqrt{668,184} \;=\; 25,849;$$

выборочный корреляционный момент

$$K_{xy} \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \;=\; \overline{xy}-\overline{x}\,\overline{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y} \;=\; 546,191;$$

выборочный коэффициент корреляции

$$r \;=\; \frac{K_{xy}}{s_x s_y} \;=\; \frac{546,191}{26,054\cdot 25,849} \;=\; 0,811.$$

Проверим значимость полученного выборочного коэффициента корреляции. Для этого вычислим статистику:

$$t_{\text{набл}} \;=\; \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \;=\; \frac{0,811\sqrt{50-2}}{\sqrt{1-0,811^2}} \;=\; 9,604.$$

Найдем по таблицам квантилей распределения Стьюдента по часто употребляемому в практике статистических расчетов уровню значимости $\alpha = 0,05$ и числу степеней свободы $\nu = n-2 = 48$ квантиль $t_{\alpha/2;\,n-2} = t_{0,025;\,48} = 2,02$.

Поскольку оказалось, что $t_{\text{набл}} > t_{\alpha/2;\,n-2}$, то можно сделать вывод, что линейная регрессионная модель $\mathrm{M}[\mathrm{Y}|\mathrm{X}=x] = \beta_0 + \beta_1 x$ выбрана удачно, т.е. она согласуется с экспериментальными данными.

3. Для подтверждения существования линейной регрессионной зависимости между исследуемыми переменными X и Y построим корреляционное поле.

Изобразим результаты измерений $\{(x_i, y_i)\}$, $(i = 1, 2, \ldots, 50)$ в виде точек в декартовой системе координат (рис. 6.12).

Визуальная оценка расположения точек на корреляционном поле позволяет принять гипотезу линейной регрессионной зависимости между признаками X и Y.

4. Найдем значения параметров эмпирического уравнения регрессии признака Y на признак X

$$\overline{y}_x \;=\; \overline{y} + r\frac{s_y}{s_x}\,(x-\overline{x}) \;=\; 107,66 + 0,811\,\frac{25,849}{26,054}\,(x-88,62) \;=\; 36,352 + 0,805x$$

и значения параметров эмпирического уравнения регрессии признака X на признак Y

$$\overline{x}_y = \overline{x} + r\frac{s_x}{s_y}(y - \overline{y}) = 88,62 + 0,811\,\frac{26,054}{25,849}(y - 107,66) = 0,616 + 0,817y.$$

Контроль вычислений: $a_1 b_1 = 0,805 \cdot 0,817 \approx 0,65 \approx r^2$.

Графики найденных эмпирических функций линейной регрессии нанесены на рис. 6.12.

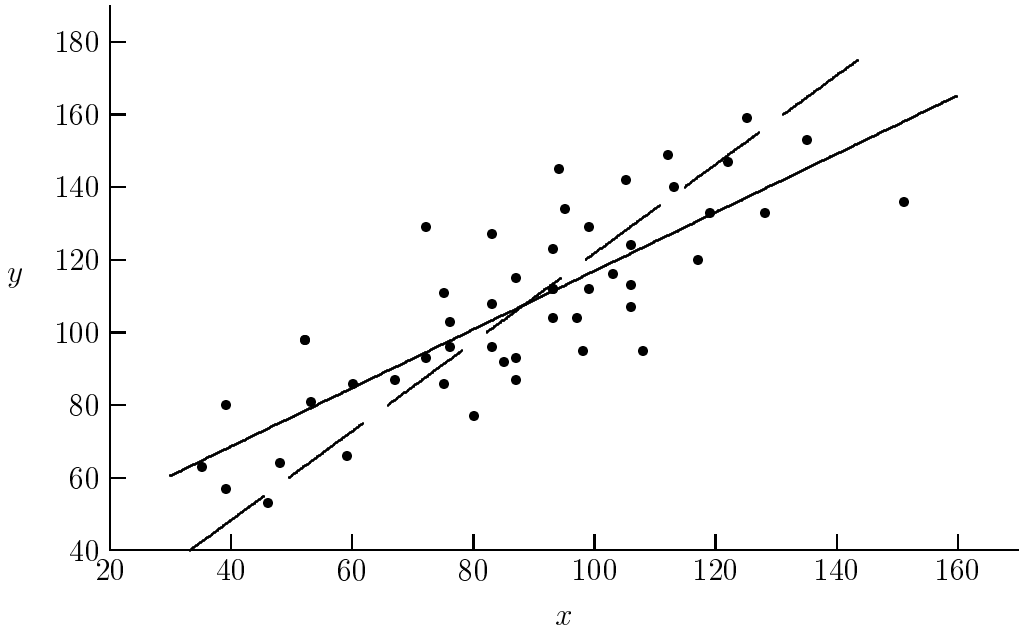


Рисунок 6.12 — Корреляционное поле экспериментальных точек; сплошной линией указано уравнение регрессии $\overline{y}_x = 36,352 + 0,805x$; пунктирной линией указано уравнение регрессии $\overline{x}_y = 0,616 + 0,817y$

6.8. Примеры

Пример 6.1

Имеются данные о растворимости азотнокислого натрия $NaNO_3$ в зависимости от температуры воды. В 100 частях воды растворяется следующее число условных частей $NaNO_3$ при соответствующих температурах.

$t°C$	0	4	10	15	21	29	36	51	68
$NaNO_3$	66,7	71,0	76,3	80,6	85,7	92,9	99,4	113,6	125,1

Предполагая, что количество $NaNO_3$ (случайная величина Y), которое растворяется в 100 частях воды, зависит линейно от температуры (случайная величина X) раствора, найти параметры a и b в формуле $y = ax + b$ по методу наименьших квадратов.

Решение

Для нахождения параметров a и b по методу наименьших квадратов необходимо решить систему

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i y_i, \\ a\sum\limits_{i=1}^{n} x_i + bn = \sum\limits_{i=1}^{n} y_i, \end{cases}$$

в которой в этом случае x_i – температура раствора; y_i – количество $\mathrm{Na\,N\,O_3}$, которое растворяется в 100 частях воды при данной температуре.

После расчета коэффициентов системы получим

$$\begin{cases} 10144a + 234b = 24628,6, \\ 234a + 9b = 811,3. \end{cases}$$

Отсюда $a = 0,87$; $b = 67,5$. Следовательно, зависимость Y от X имеет вид $y = 0,87\,x + 67,5$.

Пример 6.2

Найти уравнения регрессии Y на X и X на Y по четырем парам наблюденных значений случайной величины (X, Y):

x_i	1	2	3	4
y_i	2	4	5	7

Решение

На рис. 6.13 нанесены точки $\{(x_i, y_i)\}$ для $i = 1,\ 2,\ 3,\ 4$.

Анализируя их расположение, замечаем, что они группируются вокруг прямой линии. Поэтому будем подбирать уравнения регрессии линейного вида. Необходимые вычисления расположим в таблице.

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	1	2	2	1	4
2	2	4	8	4	16
3	3	5	15	9	25
4	4	7	28	16	49
Суммы	10	18	53	30	94

Подставляя суммы в систему нормальных уравнений (6.23), имеем

$$\begin{cases} 4b_0 + 10b_1 = 18, \\ 10b_0 + 30b_1 = 53. \end{cases}$$

Решая эту систему, находим: $b_0 = 0,5$; $b_1 = 1,6$.

Следовательно, уравнение регрессии Y на X имеет вид

$$\overline{y}_x = 0,5 + 1,6x.$$

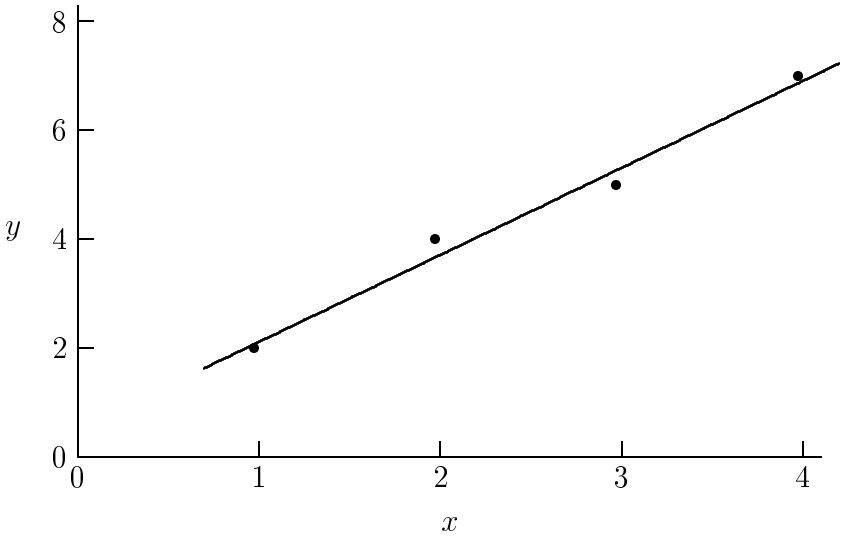


Рисунок 6.13 — К построению линейной регрессии Y на X

Найдем теперь уравнение регрессии X на Y вида (6.25). Используя значения уже найденных сумм, имеем

$$\begin{cases} 4a_0 + 18a_1 = 10, \\ 18a_0 + 94a_1 = 53. \end{cases}$$

Решая систему, находим: $a_0 = -7/26$; $a_1 = 8/13$.

Следовательно, уравнение регрессии X на Y имеет вид

$$\overline{x}_y = -0,269 + 0,615y.$$

Пример 6.3

Пусть требуется найти линейные уравнения регрессии Y на X и X на Y по 10 парам наблюденных значений случайной величины (X, Y) — точкам (x_i, y_i) $(i = 1, 2, \ldots, 10)$.

Номер измерения i	1	2	3	4	5	6	7	8	9	10
x_i	1	1	1	2	2	2	3	3	3	4
y_i	3	3	3	4	4	5	5	5	6	7

Решение

Если построить корреляционное поле, то визуально можно согласиться с тем, что точки (x_i, y_i) $(i = 1, 2, \ldots, 10)$ группируются около прямой линии. Поэтому будем подбирать уравнения регрессии линейного вида.

Вычислим суммы, входящие в системы нормальных уравнений (6.23) и (6.25), не производя группировки экспериментальных данных.

i	x_i	y_i	x_iy_i	x_i^2	y_i^2
1	1	3	3	1	9
2	1	3	3	1	9
3	1	3	3	1	9
4	2	4	8	4	16
5	2	4	8	4	16
6	2	5	10	4	25
7	3	5	15	9	25
8	3	5	15	9	25
9	3	6	18	9	36
10	4	7	28	16	49
Суммы	22	45	111	58	219

Представим эти же данные в виде корреляционной таблицы.

x_i	y_i					n_x
1	3					3
2		2	1			3
3			2	1		3
4					1	1
n_y	3	2	3	1	1	$n = 10$

Найдем уравнение регрессии Y на X. Подставляя найденные суммы в систему (6.23), имеем

$$\begin{cases} 10b_0 + 22b_1 = 45, \\ 22b_0 + 58b_1 = 111. \end{cases}$$

Решая систему, находим $b_0 = 1,75$ и $b_1 = 1,25$. Следовательно, уравнение регрессии Y на X имеет вид

$$\overline{y}_x = 1,75 + 1,25x.$$

Найдем уравнение регрессии X на Y. Подставляя найденные суммы в систему (6.25), имеем

$$\begin{cases} 10a_0 + 45a_1 = 22, \\ 45a_0 + 219a_1 = 111. \end{cases}$$

Решая систему, находим $a_0 = -1,073$ и $a_1 = 0,727$. Следовательно, уравнение регрессии X на Y имеет вид

$$\overline{x}_y = -1,073 + 0,727y.$$

Пример 6.4

Распределение признаков X и Y приводится в следующей корреляционной таблице :

X	Y									
	0	10	20	30	40	50	60	70	80	m_x
–2				1	2	1				4
–1			1	3		3	1			8
0		2	4				4	2		12
1	1	5						5	1	12
2	3								3	6
m_y	4	7	5	4	2	4	5	7	4	42

Найти корреляционные отношения $\eta_{\mathrm{Y|X}}$ и $\eta_{\mathrm{X|Y}}$ и сравнить их с соответствующим коэффициентом линейной регрессии.

Решение

Расчет корреляционных отношений будем производить по следующей схеме:

$$\overline{\mathrm{X}} = \frac{\sum_x m_x x}{\sum_x m_x} = \frac{8}{42} = 0,190; \qquad \overline{\mathrm{X}^2} = \frac{\sum_x m_x x^2}{\sum_x m_x} = \frac{60}{42} = 1,429;$$

$$\sigma^2_{\mathrm{X}} = \overline{\mathrm{X}^2} - (\overline{\mathrm{X}})^2 = 1,392.$$

Из исходных данных следует, что $y_0 = 40$ и $\beta = 0{,}1$. Воспользуемся центрированной величиной $\mathrm{V} = (\mathrm{Y} - 40)/10$. Для неё

$$\overline{\mathrm{V}} = \frac{\sum_y m_y v_y}{\sum_y m_y} = \frac{0}{42} = 0, \qquad \overline{\mathrm{V}^2} = \frac{\sum_y m_y v_y^2}{\sum_y m_y} = \frac{302}{42} = 7,190,$$

$$\sigma^2_{\mathrm{V}} = \overline{\mathrm{V}^2} - (\overline{\mathrm{V}})^2 = 7,190.$$

Далее

$$\overline{(\overline{\mathrm{V}_x})^2} = \frac{\sum_x \frac{1}{m_x}\left(\sum_y m_{xy} v_y\right)^2}{\sum_x m_x} = \frac{0}{42} = 0,$$

$$\sigma^2(\overline{\mathrm{V}_x}) = \overline{(\overline{\mathrm{V}_x})^2} - (\overline{\mathrm{V}})^2 = 0$$

и

$$\overline{(\overline{\mathrm{X}_v})^2} = \frac{\sum_y \frac{1}{m_y}\left(\sum_x m_{xy} x\right)^2}{\sum_y m_y} = \frac{52,542}{42} = 1,251,$$

$$\sigma^2(\overline{\mathrm{X}_v}) = \overline{(\overline{\mathrm{X}_v})^2} - (\overline{\mathrm{X}})^2 = 1,215.$$

Отсюда

$$\eta_{\mathrm{Y|X}} = \eta_{\mathrm{V|X}} = \sqrt{\frac{\sigma^2\left(\overline{\mathrm{V}_x}\right)}{\sigma^2_{\mathrm{Y}}}} = \sqrt{\frac{0}{7,190}} = 0;$$

$$\eta_{\mathrm{X|Y}} = \eta_{\mathrm{X|V}} = \sqrt{\frac{\sigma^2\left(\overline{\mathrm{X}_v}\right)}{\sigma^2_{\mathrm{X}}}} = \sqrt{\frac{1,215}{1,393}} = \sqrt{0,872} = 0,934.$$

Очевидно, что $\rho_{\mathrm{Y|X}} = \rho_{\mathrm{X|Y}} = 0$ и $r(\mathrm{X},\mathrm{Y}) = 0$.

Пример 6.5

Имеется две выборки $(x_1, x_2, \ldots, x_n)$ и $(y_1, y_2 \ldots, y_n)$ объемом n каждая. Построить систему нормальных уравнений, считая, что корреляционная зависимость признака Y на признак X имеет криволинейный параболический вид $\overline{y}_x = a + bx + cx^2$ (рис. 6.14).

Решение

На основе метода наименьших квадратов параметры a, b, c будем определять исходя из минимума функционала невязки

$$\sum_{i=1}^{n} (y_i - \overline{y}_x)^2 \;\Rightarrow\; \min .$$

Подставляя в это соотношение параболическую зависимость $\overline{y}_x = a + bx + cx^2$, после дифференцирования по параметрам a, b, c получим искомую систему (в ней для общности величина n обозначена через $\sum\limits_{i=1}^{n} 1$)

$$\begin{cases} a \sum\limits_{i=1}^{n} 1 \; + b \sum\limits_{i=1}^{n} x_i \; + c \sum\limits_{i=1}^{n} x_i^2 \; = \; \sum\limits_{i=1}^{n} y_i, \\ a \sum\limits_{i=1}^{n} x_i + b \sum\limits_{i=1}^{n} x_i^2 + c \sum\limits_{i=1}^{n} x_i^3 \; = \; \sum\limits_{i=1}^{n} y_i x_i, \\ a \sum\limits_{i=1}^{n} x_i^2 + b \sum\limits_{i=1}^{n} x_i^3 + c \sum\limits_{i=1}^{n} x_i^4 \; = \; \sum\limits_{i=1}^{n} y_i x_i^2. \end{cases}$$

Аналогично можно построить систему нормальных уравнений и в том случае, когда предполагаемая корреляционная зависимость Y на X имеет криволинейный параболический вид степени, более высокой чем 2.

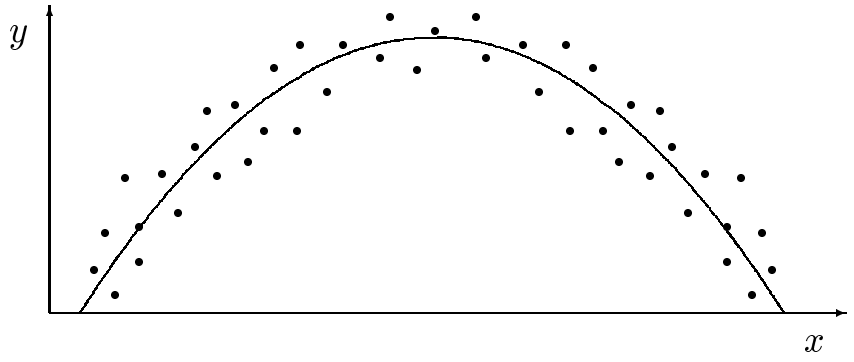


Рисунок 6.14 — Пример параболической регрессии

Пример 6.6

Из генеральной совокупности, распределение признаков X и Y в которой нормальное, произведена выборка объемом в $N = 530$ единиц. Результаты измерения признаков X и Y у компонент выборки приводятся в следующей таблице:

X	Y							
	15–25	25–35	35–45	45–55	55–65	65–75	75–85	m_y
200–300	19	5						24
300–400	23	116	11					150
400–500	1	41	98	9				149
500–600		4	32	65	7			108
600–700		1	4	21	36	3		65
700–800			1	2	11	13	1	28
800–900					1	3	2	6
m_x	43	167	146	97	55	19	3	530

Найти выборочный коэффициент линейной корреляции и его среднее квадратическое отклонение. Написать уравнения прямых регрессий Y на X и X на Y.

Решение

Сначала интервалы значений признаков X и Y заменим их серединами.

Так как коэффициент линейной корреляции не изменяется от изменения начал отсчетов и масштабов признаков (свойство 2), то для упрощения расчетов заменим значения признаков X и Y на значения признаков U и V, связанных с первыми следующими соотношениями:

$$\mathrm{U} = \alpha(\mathrm{X} - x_0), \qquad \mathrm{V} = \beta(\mathrm{Y} - y_0),$$

в которых положим $x_0 = 550$; $y_0 = 50$; $\alpha = 0{,}01$; $\beta = 0{,}1$. Полученные значения α и β запишем над соответствующими интервалами.

В результате получим корреляционную таблицу:

U	V							
	–3	–2	–1	0	1	2	3	m_v
–3	19	5						24
–2	23	116	11					150
–1	1	41	98	9				149
0		4	32	65	7			108
1		1	4	21	36	3		65
2			1	2	11	13	1	28
3					1	3	2	6
m_u	43	167	146	97	55	19	3	530

Все дальнейшие вычисления производятся над значениями признаков U и V:

$$\overline{\mathrm{U}} = \frac{1}{N}\sum_i m_{x_i} u_i = \frac{-382}{530} = -0,721;$$

$$\overline{\mathrm{V}} = \frac{1}{N}\sum_j m_{y_j} v_j = \frac{-507}{530} = -0,957;$$

$$\overline{\mathrm{UV}} = \frac{1}{N}\sum_i \sum_j m_{ij} u_i v_j = \frac{1161}{530} = 2,191.$$

Далее

$$\overline{\mathrm{V}^2} \;=\; \frac{1}{N}\sum_j m_{y_j} v_j^2 \;=\; \frac{1359}{530} \;=\; 2,564;$$

$$\overline{\mathrm{U}^2} \;=\; \frac{1}{N}\sum_i m_{x_i} u_i^2 \;=\; \frac{1196}{530} \;=\; 2,257,$$

что дает

$$\sigma_{\mathrm{U}}^2 \;=\overline{\mathrm{U}^2} - \overline{\mathrm{U}}^2 \;=\; 2,257 - 0,520 \;=\; 1,737;$$

$$\sigma_{\mathrm{V}}^2 \;=\overline{\mathrm{V}^2} - \overline{\mathrm{V}}^2 \;=\; 2,564 - 0,916 \;=\; 1,648,$$

откуда имеем

$$\sigma_{\mathrm{U}} \;=\; 1,318;$$

$$\sigma_{\mathrm{V}} \;=\; 1,284.$$

Для коэффициента корреляции получим

$$r \;=\; \frac{\overline{\mathrm{UV}} - \overline{\mathrm{U}}\cdot\overline{\mathrm{V}}}{\sigma_{\mathrm{U}}\,\sigma_{\mathrm{V}}} \;=\; \frac{2,191 - 0,721\cdot 0.957}{1,318\cdot 1,284} \;=\; \frac{1,501}{1,692} \;=\; 0,887.$$

Далее

$$\overline{\mathrm{X}} \;=\; x_0 + \frac{\overline{\mathrm{U}}}{\alpha} \;=\; 550 + \frac{-0,721}{0,01} \;=\; 550 - 72,1 \;=\; 477,9;$$

$$\overline{\mathrm{Y}} \;=\; y_0 + \frac{\overline{\mathrm{V}}}{\beta} \;=\; 50 + \frac{-0,957}{0,1} \;=\; 50 - 9,57 \;=\; 40,43.$$

Для коэффициентов регрессии найдем

$$\rho_{\mathrm{Y|X}} \;=\; r\,\frac{\sigma_{\mathrm{Y}}}{\sigma_{\mathrm{X}}} \;=\; 0,887\cdot\frac{(1,318/0,1)}{(1,284/0,01)} \;=\; 0,887\cdot\frac{13,18}{128,4} \;=\; 0,086;$$

$$\rho_{\mathrm{X|Y}} \;=\; r\,\frac{\sigma_{\mathrm{X}}}{\sigma_{\mathrm{Y}}} \;=\; 0,887\cdot\frac{(1,284/0,01)}{(1,318/0,1)} \;=\; 0,887\cdot\frac{128,4}{13,18} \;=\; 9,103.$$

Поэтому уравнения прямых регрессий Y на X и X на Y следующие:

$$y - 40,43 \;=\; 0,086\,(x - 477,9);$$

$$x - 477,9 \;=\; 9,103\,(y - 40,43),$$

а для среднего квадратического отклонения выборочного коэффициента линейной корреляции получим:

$$\sigma_r \;\approx\; \frac{1-r^2}{\sqrt{N}} \;=\; \frac{1-0,887^2}{\sqrt{530}} \;=\; 0,009.$$

Пример 6.7

Имеется выборка объема $n = 11$, извлеченная из двумерной нормальной совокупности. Для неё вычислен выборочный (эмпирический) коэффициент корреляции $r = 0,76$.

Требуется:

а) построить 95%-й доверительный интервал для коэффициента корреляции выборочной совокупности;

б) проверить нулевую гипотезу $\{H_0 : \rho = 0\}$ против альтернативной гипотезы $\{H_a : \rho \neq 0\}$.

Уровень значимости α принять 0,01.

Решение

Найдем 95%–й доверительный интервал для коэффициента корреляции выборочной совокупности ρ. Выполнив расчеты, пользуясь распределением Крамера, получим для 95%–го доверительного интервала $+0,27 < \rho < +0,92$.

Так как данный доверительный интервал не накрывает значение $\rho = 0$, то на уровне значимости $\alpha = 0,01$ отвергается нулевая гипотеза $\{H_0 : \rho = 0\}$ некоррелированности переменных X и Y.

Проверку этой же нулевой гипотезы можно провести, используя Модель 1. Для этого вычислим статистику

$$t_{\text{набл}} = r\sqrt{\frac{n-2}{1-r^2}} = 0,76\sqrt{\frac{11-2}{1-0,76^2}} = 3,508.$$

По таблицам стандартизированного нормального распределения по $\alpha = 0,01$ находим $u_{\alpha/2} = u_{0,005} = 2,58$. Следовательно, применяя Модель 3, также приходим к выводу о том, что выборочный коэффициент корреляции отличен от нуля.

По таблицам распределения Стьюдента по $\alpha = 0,01$ и числу степеней свободы $\nu = n - 2 = 9$ находим $t_{\alpha/2;\,n-2} = t_{0,005;\,9} = 3,25$. Так как $t_{\text{набл}} = 3,50 > 3,25$, то выборочный коэффициент корреляции значимо отличается от нуля, т.е. переменные X и Y являются коррелированными.

Для проверки этой же нулевой гипотезы H_0, т.е. $\rho = 0$, применим Модель 3. Для этого вычислим статистику u по формуле

$$u = \left(1,1513\lg\frac{1+r}{1-r} - 1,1513\lg\frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{2(n-1)}\right)\sqrt{n-3}.$$

Полагая $\rho_0 = 0$, получим

$$u = 1,1513\lg\frac{1,76}{0,24}\sqrt{11-3} = 2,82.$$

Сравнивая с $u_{\alpha/2} = u_{0,005} = 2,58$, приходим к выводу о том, что выборочный коэффициент корреляции отличен от нуля.

Пример 6.8

Выпуск некоторым предприятием промышленной продукции (Y) за семь лет (X) характеризуется следующими данными:

X, *год*	1	2	3	4	5	6	7
Y, *усл. ед.*	0,5	0,5	1,5	3,5	6,5	10,5	15,5

Выровнять зависимость Y от X по параболе $y = ax^2 + bx + c$.

Решение

Для нахождения параметров a, b и c по методу наименьших квадратов необходимо решить систему

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^4 + b\sum\limits_{i=1}^{n} x_i^3 + c\sum\limits_{i=1}^{n} x_i^2 &= \sum\limits_{i=1}^{n} x_i^2\, y_i, \\ a\sum\limits_{i=1}^{n} x_i^3 + b\sum\limits_{i=1}^{n} x_i^2 + c\sum\limits_{i=1}^{n} x_i &= \sum\limits_{i=1}^{n} x_i\, y_i, \\ a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i + c\sum\limits_{i=1}^{n} 1 &= \sum\limits_{i=1}^{n} y_i, \end{cases}$$

в которой в данном случае $n = 7$, x_i — это текущий год семилетки, y_i — соответствующая продукция, а $\sum\limits_{i=1}^{n} 1 = n$.

Составляем систему для определения параметров a, b, c:

$$\begin{cases} 4676\,a + 784\,b + 140\,c &= 1372, \\ 784\,a + 140\,b + 28\,c &= 224, \\ 140\,a + 28\,b + 7\,c &= 38,5. \end{cases}$$

Численное решение этой системы можно искать, например, методом Крамера или методом исключения Гаусса.

В результате находим $a = 0{,}5$; $b = -1{,}5$; $c = 1{,}5$.

Пример 6.9 (Распределение выборочного коэффициента корреляции)

Рассмотрим нормальное распределение системы двух величин (X, Y). Без ограничения общности можно предполагать, что моменты первого порядка равны нулю, так что плотность вероятности имеет вид

$$f_{\mathrm{XY}}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)\right] \equiv$$

$$\equiv \frac{1}{2\pi\sqrt{\det G}} \exp\left[-\frac{1}{2D}\left(\mu_{02}x^2 + 2\mu_{11}xy + \mu_{20}y^2\right)\right], \qquad (1)$$

где $D = \mu_{20}\mu_{02} - \mu_{11}^2 = \sigma_1^2\sigma_2^2(1-\rho^2)$ — детерминант матрицы вторых моментов;

$$G = \begin{pmatrix} \mu_{20} & \mu_{11} \\ \mu_{11} & \mu_{02} \end{pmatrix} = \begin{pmatrix} \sigma_2^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}.$$

По выборке из n наблюденных пар значений $(x_1, y_1), \ldots, (x_n, y_n)$ вычислим моментные характеристики первого и второго порядка:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad (2a)$$

$$\overline{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{2b}$$

$$m_{20} \;=\; s_1^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2, \tag{2c}$$

$$m_{02} \;=\; s_2^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2, \tag{2d}$$

$$m_{11} \;=\; r s_1 s_2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\cdot\overline{y}. \tag{2e}$$

Коэффициент корреляции

$$r \;=\; \frac{m_{11}}{s_1 s_2}.$$

Теперь рассмотрим характеристическую функцию распределения пяти случайных величин $\overline{x}$, $\overline{y}$, m_{20}, m_{11}, m_{02}, которая есть функция от пяти аргументов t_1, t_2, t_{20}, t_{11}, t_{02} :

$$Q(t_1,t_2,t_{20},t_{11},t_{02}) \;=\; \mathrm{M}\left[\mathrm{e}^{i(t_1\overline{x}+t_2\overline{y}+t_{20}m_{20}+t_{11}m_{11}+t_{02}m_{02})}\right] \;=$$

$$=\; \frac{1}{(2\pi)^n D^{n/2}}\int\ldots\int \mathrm{e}^{\Omega}\,dx_1\ldots dx_n\,dy_1\ldots dy_n, \tag{3}$$

где

$$\Omega \;=\; -\frac{1}{2D}\sum_{i=1}^{n}\left(\mu_{02}x_i^2 - 2\mu_{11}x_iy_i + \mu_{20}y_i^2\right) +$$

$$+i(t_1\overline{x}+t_2\overline{y}+t_{20}m_{20}+t_{11}m_{11}+t_{02}m_{02}),$$

а интегрирование распространяется на $2n$-мерное пространство переменных x_1, x_2, $\ldots$, x_n, y_1, y_2, $\ldots$, y_n.

Заменим $x_1,\ldots,x_n$ новыми переменными $\xi_1,\ldots,\xi_n$ с помощью ортогонального преобразования, при котором $\xi_1 = \sqrt{n}\overline{x}$, и применим преобразование с такой же матрицей к величинам $y_1,\ldots,y_n$, которые при этом заменяются новыми величинами $\eta_1,\ldots,\eta_n$, причём $\eta_1 = \sqrt{n}\overline{y}$. Тогда имеем

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}\xi_i^2, \qquad \sum_{i=1}^{n} x_iy_i = \sum_{i=1}^{n}\xi_i\eta_i, \qquad \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n}\eta_i^2,$$

$$nm_{20} = \sum_{i=2}^{n}\eta_i^2, \qquad nm_{11} = \sum_{i=2}^{n}\xi_i\eta_i, \qquad nm_{02} = \sum_{i=2}^{n}\eta_i^2$$

и отсюда

$$\Omega \;=\; i\frac{t_1\xi_1+t_2\eta_1}{\sqrt{n}} - \frac{1}{2D}\left(\mu_{02}\xi_1^2 - 2\mu_{11}\xi_1\eta_1 + \mu_{20}\eta_1^2\right) -$$

$$-\frac{1}{n}\sum_{i=2}^{n}\left[\left(\frac{n\mu_{02}}{2M} - it_{20}\right)\xi_i^2 + 2\left(-\frac{n\mu_{11}}{2D} - \frac{1}{2}it_{11}\right)\xi_i\eta_i + \left(\frac{n\mu_{20}}{2D} - it_{02}\right)\eta_i^2\right].$$

Подставляя это выражение для Ω в формулу (3), приведём $2n$-кратный интеграл к произведению n двойных интегралов. Тогда совместная характеристическая функция $Q(t_1,t_2,t_{20},t_{11},t_{02})$ примет вид

$$Q(t_1,t_2,t_{20},t_{11},t_{02}) = \exp\left[-\frac{1}{2n}\left(\mu_{20}t_1^2 + 2\mu_{11}t_1t_2 + \mu_{02}t_2^2\right)\right](A/A^*)^{(n-1)/2}, \tag{4}$$

где

$$A = \det \begin{pmatrix} (n\mu_{02})/2D & -(n\mu_{11})/2D \\ -(n\mu_{11})/2D & (n\mu_{20})/2D \end{pmatrix},$$

$$A^* = \det \begin{pmatrix} (n\mu_{02})/2m - it_{20} & -(n\mu_{11})/2D - it_{11}/2 \\ -(n\mu_{11})/2D - it_{11}/2 & (n\mu_{20})/2D - it_{02} \end{pmatrix}.$$

Совместная характеристическая функция есть произведение двух множителей, первый из которых содержит лишь переменные t_1 и t_2, а второй – лишь t_{20}, t_{11} и t_{02}. Первый множитель является характеристической функцией некоторого нормального распределения с нулевым средним значением и матрицей вторых моментов $n^{-1}G$. Второй множитель есть частный случай характеристической функции

$$q_n\left(t_{20}, t_{11}, t_{02}\right) = \left(A/A^*\right)^{(n-1)/2}.$$

Соответствующее распределение есть частный случай распределения

$$f_n(x_{11}, x_{12}, x_{22}) = C_{2n}\left(a_{11}a_{22} - a_{12}^2\right)^{(n-1)/2} \times \tag{5}$$

$$\times \left(x_{11}x_{22} - x_{12}^2\right)^{(n-4)/2} \exp\left[-a_{11}x_{11} - a_{22}x_{22} - 2a_{12}x_{12}\right],$$

где

$$C_{2n} = \frac{1}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)} = \frac{2^{n-3}}{\pi\Gamma(n-2)}$$

с заменой x_{11}, x_{12}, x_{22} на m_{20}, m_{11}, m_{02}.

Таким образом, составные случайные величины $(\overline{x},\ \overline{y})$ и $(m_{20},\ m_{11},\ m_{02})$ независимы. Совместное распределение величин $\overline{x}$ и $\overline{y}$ нормально и имеет те же моменты первого порядка, что и распределение генеральной совокупности, и матрицу вторых моментов $n^{-1}G$.

Совместное распределение величин m_{20}, m_{11}, m_{02} имеет плотность распределения вероятности f_n, задаваемую формулой

$$f_n\left(m_{20}, m_{11}, m_{02}\right) = \frac{n^{n-1}}{4\pi\Gamma(n-2)} \frac{\left(m_{20}m_{02} - m_{11}^2\right)^{(n-4)/2}}{M^{(n-1)/2}} \times \tag{6}$$

$$\times \exp\left[-\frac{n}{2M}\left(\mu_{02}m_{20} - 2\mu_{11}m_{11} + \mu_{20}m_{02}\right)\right]$$

в области $m_{20} > 0$, $m_{02} > 0$ и $m_{11}^2 < m_{20}m_{02}$, причем $f_n = 0$ вне этой области.

Введем новую переменную r в совместное распределение (6) величин m_{20}, m_{11} и m_{02}, полагая $m_{11} = r\sqrt{m_{20}m_{02}}$, так что r – коэффициент корреляции выборки. Тогда получим следующее выражение для совместной плотности распределения вероятностей величин m_{20}, m_{02} и r:

$$\sqrt{m_{20}m_{02}}\, f_n(m_{20}, r\sqrt{m_{20}m_{02}}, m_{02}) = \tag{7}$$

$$= \frac{n^{n-1}}{4\pi\Gamma(n-2)D^{(n-1)/2}} \left(m_{20}m_{02}\right)^{(n-3)/2} \left(1 - r^2\right)^{(n-4)/2} \times$$

$$\times \exp\left[-\frac{n}{2D}(\mu_{02}m_{20} - 2\mu_{11}r\sqrt{m_{20}m_{02}} + \mu_{20}m_{02})\right],$$

где $m_{20} > 0$, $m_{02} > 0$, $r^2 < 1$.

Частная плотность распределения вероятности для r получится интегрированием совместной плотности распределения вероятности по m_{20} и m_{02} в пределах от 0 до ∞. Если разложить в степенной ряд множитель $\exp\left(-n\mu_{11}r\sqrt{m_{20}m_{02}}/D\right)$, то можно получить плотность распределения вероятностей для выборочного коэффициента корреляции r:

$$f_n(r) = \frac{2^{n-3}}{\pi(n-3)!}(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}\sum_{\nu=0}^{\infty}\Gamma^2\left(\frac{n+\nu-1}{2}\right)\frac{(2\rho r)^\nu}{\nu!}, \qquad (8)$$

при этом $-1 < r < 1$.

Входящий в это выражение степенной ряд можно преобразовать. Поскольку справедливо

$$\frac{2^{n-3}}{(n-2)!}\sum_{\nu=0}^{\infty}\Gamma^2\left(\frac{n+\nu-1}{2}\right)\frac{(2\rho r)^\nu}{\nu!} = \int_0^1 \frac{x^{n-2}}{(1-\rho r x)^{n-1}}\frac{dx}{\sqrt{1-x^2}}, \qquad (9)$$

то отсюда получается следующее выражение для плотности распределения вероятностей $f_n(r)$ величины r:

$$f_n(r) = \frac{n-2}{\pi}(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}\int_0^1 \frac{x^{n-2}}{(1-\rho r x)^{n-1}}\frac{dx}{\sqrt{1-x^2}}. \qquad (10)$$

Найденное распределение зависит только от объема выборки n и от коэффициента корреляции генеральной совокупности ρ.

При $n = 2$ плотность распределение вероятностей $f_n(r)$ обращается в нуль в соответствии с тем фактом, что коэффициент корреляции, вычисленный по выборке, состоящей только из двух значений, необходимо равен ± 1, так что в этом случае распределение принадлежит дискретному типу.

При $n = 3$ плотность распределение вероятностей U-образна, с бесконечными ординатами в точках $r = \pm 1$.

При $n = 4$ получается прямоугольное распределение, если $\rho = 0$ и J-образное распределение в случае $\rho \neq 0$.

При $n > 4$ распределение вероятностей унимодально, её мода сосредоточена в точке $r = 0$, если $\rho = 0$, и около точки $r = \rho$, если $\rho \neq 0$.

6.9. Задачи для решения

Задача 6.1

Пусть зависимость признака Y от признака X характеризуется следующей таблицей:

X	–2	–1	0	1	2	3
Y	–2	–3	–3	–1	3	7

Предполагая, что $y = ax + b$, найти параметры этой зависимости, пользуясь методом наименьших квадратов.

Задача 6.2

Пусть зависимость признака Y от признака X характеризуется следующей таблицей:

X	−2	−1	0	1	2	3
Y	2	3	3	1	−3	−7

Предполагая, что $y = ax + b$, найти параметры этой зависимости, пользуясь методом наименьших квадратов.

Задача 6.3

Следующая корреляционная таблица дает распределение двух признаков X и Y:

X	Y					
	10	20	30	40	50	m_x
5	2				2	4
6	1	1		1	1	4
7		2	1	2		5
8			1			1
m_y	3	3	2	3	3	14

Построить зависимости линейной регрессии X на Y и Y на X.

Задача 6.4

В десяти городах имеются аптеки. Количество аптек и численность населения (в десятках тысяч человек) в этих 10 городах приведено в таблице.

Нанести данные на диаграмму рассеяния, вычислить коэффициенты корреляции: а) для первых девяти городов, б) для всех десяти городов. Сравнить результаты.

№ города	1	2	3	4	5	6	7	8	9	10
Население	45	45	47	48	51	58	59	65	67	80
Число аптек	12	12	29	25	38	35	16	43	22	34

Задача 6.5

Из 125 опытных участков 55 находились на неудобренном массиве и 70 на удобренном. При этом на всех 125 участках имела место следующая урожайность:

Средний урожай, $г/м^2$	95	105	115	125	135	145	155	165	175
Число участков на I массиве	2	5	12	15	10	7	3	1	0
Число участков на II массиве	0	0	1	2	8	24	19	11	5

Найти корреляционное отношение урожайности от удобрений.

Задача 6.6

Вычислить коэффициент линейной корреляции и получить уравнение прямой регрессии X на Y по данным таблицы распределения 100 растений ржи по общему весу всего растения Y и по весу семян X:

X, *г*	Y, *г*										
	20--30	30--40	40--50	50--60	60--70	70--80	80--90	90--100	100--110	110--120	120--130
10–15	3										
15–20	2	6	1	1							
20–25		4	13	2	1						
25–30			5	4							
30–35				8	4	2					
35–40				1	4	6					
40–45					2	6	1				
45–50		1				1	5	1			
50–55								4	2		
55–60								1	4	1	
60–65									1		
65–70									1	1	1

Задача 6.7

Средняя температура воздуха в сентябре в двух городах (X) и (Y) измерялась в течение 40 лет. Данные приведены в таблице.

X	Y	X	Y	X	Y	X	Y	X	Y
12,0	10,8	13,9	10,1	14,9	13,0	16,0	16,0	18,0	14,0
12,0	11,3	14,2	10,0	14,9	14,2	16,9	12,9	18,0	14,9
12,0	12,0	14,0	10,0	15,1	13,8	17,2	13,9	18,1	16,0
12,0	13,0	14,0	12,0	15,0	16,0	16,9	14,8	18,4	17,8
12,8	10,9	13,9	12,4	15,5	13,9	16,9	15,0	19,2	15,0
13,8	10,0	15,0	11,0	15,9	14,7	17,0	16,0	19,3	16,1
13,1	13,0	14,0	14,8	16,0	13,0	16,8	17,0	20,0	17,0
13,0	13,0	14,0	15,2	15,9	15,0	17,5	16,0	20,1	17,7

Найти выборочные среднемесячные температуры в обоих населенных пунктах и их среднеквадратичные уклонения. Найти выборочный коэффициент корреляции X и Y, написать выборочное уравнение линейной регрессии Y на X.

Задача 6.8

Ниже приведена таблица значений признака Y при различных значениях признака X:

X	12	13	14	16	17	18	20	21	21	24	25	26	28
Y	54	59	67	76	85	97	107	118	127	139	153	160	178

Выровнять зависимость Y от X вдоль прямой $y = ax + b$, пользуясь методом наименьших квадратов. Выровнять зависимость X от Y вдоль прямой $x = cy + d$, пользуясь методом наименьших квадратов. Полученные зависимости сравнить.

Задача 6.9

Пусть зависимость признака Y от признака X характеризуется следующей таблицей:

X	−2	−1	0	1	2	3
Y	−2	−3	−3	−1	3	7

Предполагая, что $y = ax^2 + bx + c$, найти параметры этой зависимости, пользуясь методом наименьших квадратов.

Задача 6.10

В опытах были получены следующие результаты при измерениях диаметра (Y) пыльцы шаровидной фуксии в зависимости от числа спор (X), расположенных в экваториальной плоскости пылинки:

X	Y, *мк*					
	0	1	2	3	4	m_y
10	3					3
15	7	3				10
20		6				6
25		6	1			7
30			4			4
35			5			5
40			1	3		4
45				4		4
50				3	3	6
55					4	4
60					3	3

Найти коэффициент линейной корреляции между признаками X и Y и уравнение линейной регрессии Y на X.

Задача 6.11

Зависимость продолжительности t решения систем линейных уравнений одинаковой степени трудности от порядка системы n приведена в следующей таблице:

n	2	3	4	5	6	7	8	9	10
t, *минут*	12	35	75	130	210	315	445	600	800

Предполагая справедливость того, что $t = An^{\alpha}$, найти значения параметров A и α методом наименьших квадратов.

Задача 6.12

Восемь раз при различных значениях признака X было измерено значение признака Y. Полученные результаты приводятся в следующей таблице:

X	0,30	0,91	1,50	2,00	2,20	2,62	3,00	3,30
Y	0,20	0,43	0,31	0,52	0,81	0,68	1,15	0,85

Предполагая теоретически, что $y = ax + b$, найти a и b.

6.10. Задание на практическую работу

Настоящая практическая работа рассчитана на два часа и содержит два задания. Задания должны выполняться в выбранной программной среде.

Задание 1

Напишите программу нахождения значений параметров a и b линейной зависимости $y = ax+b$, а также значений параметров c и d линейной зависимости $x = cy+d$. Воспользуйтесь методом наименьших квадратов.

Вариант 1

Зависимость признака Y от признака X характеризуется следующей таблицей:

X	–3	–2	–1	0	1	2	3
Y	–5	–3	–3	–1	–1	3	3

Результат работы – массив, содержащий значения искомых параметров. Необходимо предусмотреть визуализацию данных (построить программно точки заданных значений, а также линии прямых регрессий $y = ax + b$ и $x = cy + d$).

Вариант 2

Зависимость признака Y от признака X характеризуется следующей таблицей:

X	–3	–2	–1	0	1	2	3
Y	–4	–2	–3	–1	–1	3	2

Результат работы – массив, содержащий значения искомых параметров. Необходимо предусмотреть визуализацию данных (построить программно точки заданных значений, а также линии прямых регрессий $y = ax + b$ и $x = cy + d$).

Вариант 3

Зависимость признака Y от признака X характеризуется следующей таблицей:

X	–3	–2	–1	0	1	2	3
Y	–4	–2	–2	–1	–1	2	3

Результат работы – массив, содержащий значения искомых параметров. Необходимо предусмотреть визуализацию данных (построить программно точки заданных значений, а также линии прямых регрессий $y = ax + b$ и $x = cy + d$).

Задание 2

Напишите программу нахождения значений параметров a, b, c параболической зависимости $y = ax^2 + bx + c$. Воспользуйтесь методом наименьших квадратов.

Вариант 1

Зависимость признака Y от признака X характеризуется следующей таблицей:

X	–3	–2	–1	0	1	2	3
Y	7	5	3	1	–1	3	7

Результат работы – массив, содержащий значения искомых параметров a, b, c параболической аппроксимации. Необходимо предусмотреть визуализацию данных (построить программно точки заданных значений, а также параболу $y = ax^2 + bx + c$).

Вариант 2

Зависимость признака Y от признака X характеризуется следующей таблицей:

X	–3	–2	–1	0	1	2	3
Y	7	5	3	0	3	5	7

Результат работы – массив, содержащий значения искомых параметров a, b, c параболической аппроксимации. Необходимо предусмотреть визуализацию данных (построить программно точки заданных значений, а также параболу $y = ax^2 + bx + c$).

Вариант 3

Зависимость признака Y от признака X характеризуется следующей таблицей:

X	–3	–2	–1	0	1	2	3
Y	–6	–5	–3	–1	–1	–3	–7

Результат работы – массив, содержащий значения искомых параметров a, b, c параболической аппроксимации. Необходимо предусмотреть визуализацию данных (построить программно точки заданных значений, а также параболу $y = ax^2 + bx + c$).

6.11. Задания для проверки

1. В чем состоит различие между функциональной и статистической зависимостью между двумя переменными?

2. Что называется модельным уравнением регрессии Y на X?

3. Что называется эмпирическим уравнением регрессии Y на X?

4. Объясните общую идею подбора эмпирических уравнений методом наименьших квадратов.

5. Сформулируйте основные задачи корреляционного анализа, регрессионного анализа.

6. Запишите общий вид модельных функций регрессии Y на X и X на Y, если известно, что двумерная случайная величина (X, Y) распределена по нормальному закону.

7. Выведите формулы для вычисления коэффициентов эмпирического уравнения регрессии по сгруппированным и несгруппированным опытным данным.

8. Что называется ковариацией (ковариационным моментом) генеральной совокупности?

9. Что называется ковариацией (ковариационным моментом) выборочной совокупности?

12. Сформулируйте основные исходные предположения, лежащие в обосновании линейной модели регрессии.

13. Какой вид имеет график эмпирической функция регрессии, если известно, что случайная выборка извлечена из генеральной нормальной совокупности?

14. Запишите систему уравнений для определения коэффициентов b_0 и b_1 линейного уравнения регрессии Y на X вида $\overline{y}_x = b_0 + b_1 x$ по методу наименьших квадратов.

15. Запишите систему уравнений для определения коэффициентов a_0 и a_1 линейного уравнения регрессии X на Y вида $\overline{x}_y = a_0 + a_1 y$ по методу наименьших квадратов.

16. В чем состоит свойство минимальности модельных функций регрессии?

17. Сопоставьте два уравнения линейной регрессии : Y на X и X на Y. Поясните связь между коэффициентами a_0, a_1 и b_0, b_1.

Приложение

П.1. Справочные таблицы

В настоящем приложении к пособию приведены таблицы величин, часто используемых при решении задач математической статистики:

1. Нормальная функция Гаусса $\exp(-x^2)$, производная нормальной плотности $(2\pi)^{-1/2}\exp(-x^2/2)$, функция ошибок $\mathrm{erf}(x)$ и дополнительная функция ошибок $\mathrm{erfc}(x)$;

2. Функция Лапласа $\Phi(x)$ и функции $d\Phi(x)/dx$, $2\Phi(x)$, $1-2\Phi(x)$;

3. Квантили u_α нормального распределения Гаусса, отвечающие равенству $\alpha = 1-\Phi(u_\alpha)$;

4. Границы γ_1 и γ_2 доверительного интервала для СКО σ нормальной случайной величины: $p = \Pr(\gamma_1 s < \sigma < \gamma_2 s) = 1-\alpha$;

5. Квантили $\chi^2_{\alpha;\,\nu}$ распределения χ^2 с ν степенями свободы, отвечающие равенству $\Pr\left(\chi^2 > \chi^2_{\alpha,\nu}\right) = \alpha$;

6. Правосторонние квантили $\chi^2_{1-\alpha,\nu}$ распределения χ^2 с ν степенями свободы, отвечающие равенству $\Pr\left(\chi^2 > \chi^2_{1-\alpha,\nu}\right) = 1-\alpha$;

7. Квантили $t_{\alpha;\,\nu}$ распределения Стьюдента с ν степенями свободы, определяемые вероятностью $\Pr(|t| > t_{\alpha;\,\nu}) = \alpha$ (двусторонняя критическая область);

8. Левосторонние квантили распределения Стьюдента $t_p(k)$, с k степенями свободы, отвечающие равенству $\Pr(t < t_p(k)) = p$;

9. Квантили распределения $F_{\alpha;\,k_1,k_2}$ Фишера-Снедекора с числами степеней свободы k_1 и k_2;

10. Предельное распределение Колмогорова $K(\lambda)$;

11. Квантили λ_α распределения Колмогорова: $\Pr(\lambda \geq \lambda_\alpha) = \alpha$;

12. Квантили λ_α распределения Смирнова-Колмогорова, отвечающие равенству $\Pr(\lambda \geq \lambda_\alpha) = \alpha$;

13. Квантили r_α числа знаков: $\Pr(r \geq r_\alpha) = \alpha$.

1. Функция Гаусса $\exp(-x^2)$, производная нормальной плотности $(2\pi)^{-1/2}\exp(-x^2/2)$, функция ошибок $\operatorname{erf}(x)$ и дополнительная функция ошибок $\operatorname{erfc}(x)$

x	$\exp(-x^2)$	$(2\pi)^{-1/2}\exp(-x^2/2)$	$\operatorname{erf}(x)$	$\operatorname{erfc}(x)$
0,00	1,00000	0,39894	0,00000	1,00000
0,10	0,99005	0,39695	0,11246	0,88754
0,20	0,96079	0,39104	0,22270	0,77730
0,30	0,91393	0,38139	0,32863	0,67137
0,40	0,85214	0,36827	0,42839	0,57161
0,50	0,77880	0,35207	0,52050	0,47950
0,60	0,69768	0,33322	0,60386	0,39614
0,70	0,61263	0,31225	0,67780	0,32220
0,80	0,52729	0,28969	0,74210	0,25790
0,90	0,44486	0,26609	0,79691	0,20309
1,00	0,36788	0,24197	0,84270	0,15730
1,10	0,29820	0,21785	0,88021	0,11979
1,20	0,23693	0,19419	0,91031	0,08969
1,30	0,18452	0,17137	0,93401	0,06599
1,40	0,14086	0,14973	0,95229	0,04771
1,50	0,10540	0,12952	0,96611	0,03389
1,60	0,07730	0,11092	0,97635	0,02365
1,70	0,05558	0,09405	0,98379	0,01621
1,80	0,03916	0,07895	0,98909	0,01091
1,90	0,02705	0,06562	0,99279	0,00721
2,00	0,01832	0,05399	0,99532	0,00468
2,10	0,01216	0,04398	0,99702	0,00298
2,20	0,00791	0,03547	0,99814	0,00186
2,30	0,00504	0,02833	0,99886	0,00114
2,40	0,00315	0,02239	0,99931	0,00069
2,50	0,00193	0,01753	0,99959	0,00041
2,60	0,00116	0,01358	0,99976	0,00024
2,70	0,00068	0,01042	0,99987	0,00013
2,80	0,00039	0,00792	0,99992	0,00008
2,90	0,00022	0,00595	0,99996	0,00004
3,00	0,00012	0,00443	0,99998	0,00002

Здесь

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int\limits_0^x \exp(-t^2)\,dt, \qquad \operatorname{erfc}(x) = 1 - \operatorname{erf}(x).$$

x	$d\Phi(x)/dx$	$\Phi(x)$	$2\,\Phi(x)$	$1-\Phi(x)$
0,00	0,39894	0,00000	0,00000	1,00000
0,10	0,39695	0,03983	0,07966	0,96017
0,20	0,39104	0,07926	0,15852	0,92074
0,30	0,38139	0,11791	0,23582	0,88209
0,40	0,36827	0,15542	0,31084	0,84458
0,50	0,35207	0,19146	0,38292	0,80854
0,60	0,33322	0,22575	0,45149	0,77425
0,70	0,31225	0,25804	0,51607	0,74196
0,80	0,28969	0,28814	0,57629	0,71186
0,90	0,26609	0,31594	0,63188	0,68406
1,00	0,24197	0,34134	0,68269	0,65866
1,10	0,21785	0,36433	0,72867	0,63567
1,20	0,19419	0,38493	0,76986	0,61507
1,30	0,17137	0,40320	0,80640	0,59680
1,40	0,14973	0,41924	0,83849	0,58076
1,50	0,12952	0,43319	0,86639	0,56681
1,60	0,11092	0,44520	0,89040	0,55480
1,70	0,09405	0,45543	0,91087	0,54457
1,80	0,07895	0,46407	0,92814	0,53593
1,90	0,06562	0,47128	0,94257	0,52872
2,00	0,05399	0,47725	0,95450	0,52275
2,10	0,04398	0,48214	0,96427	0,51786
2,20	0,03547	0,48610	0,97219	0,51390
2,30	0,02833	0,48928	0,97855	0,51072
2,40	0,02239	0,49180	0,98360	0,50820
2,50	0,01753	0,49379	0,98758	0,50621
2,60	0,01358	0,49534	0,99068	0,50466
2,70	0,01042	0,49653	0,99307	0,50347
2,80	0,00792	0,49744	0,99489	0,50256
2,90	0,00595	0,49813	0,99627	0,50187
3,00	0,00443	0,49865	0,99730	0,50135

Здесь

$$\Phi(x) \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_0^x \exp(-t^2/2)\,dt.$$

$$\Phi(x) = \frac{1}{2}\,\mathrm{erf}(x/\sqrt{2}), \qquad\qquad \mathrm{erf}(x) = 2\,\Phi(\sqrt{2}\,x).$$

3. Квантили u_α нормального распределения

При данном уровне значимости α правосторонний квантиль u_α определяется равенством:

$$\alpha \;=\; \frac{1}{\sqrt{2\pi}} \int_{u_\alpha}^{\infty} \exp\left(-t^2/2\right)\, dt \;=\; \frac{1}{2} - \Phi(u_\alpha)\,.$$

Уровень значимости α	Квантиль u_α
0,0001	3,0902
0,005	2,5758
0,010	2,3263
0,015	2,1701
0,020	2,0537
0,025	1,9600
0,030	1,8808
0,035	1,8119
0,040	1,7507
0,045	1,6954
0,050	1,6449

При численном нахождении значений функции Лапласа $\Phi(x)$ и правостороннего квантиля u_α можно также пользоваться следующими приближенными выражениями:

$$q_x \;=\; 2[1-\Phi(x)] \;=\; \frac{2}{\sqrt{2\pi}} \int_{x}^{\infty} \exp\left(-t^2/2\right)\, dt \;=$$

$$= \exp\left(-\frac{(83x+35)x+562}{165+703/x}\right), \qquad 0 < x \le 5,5;$$

$$x \;=\; \left[\frac{((4y+100)y+205)y}{((2y+50)y+192)y+13}\right]^{-1/2}, \qquad 2\cdot 10^{-7} \le q_x \le 1,$$

где $y = -\ln q_x$.

4. Нижние γ_1 и верхние γ_2 границы доверительного интервала, отвечающие равенству

$$p \;=\; \Pr\left(\gamma_1 s < \sigma < \gamma_2 s\right) \;=\; 1-\alpha$$

$$s \;=\; \left[\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right]^{1/2}, \qquad \overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i$$

Число степеней свободы ν	Уровень значимости					
	$p=1-\alpha=0,99$		$p=1-\alpha=0,95$		$p=1-\alpha=0,90$	
	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2
1	0,356	159	0,446	31,9	0,510	15,9
2	0,434	14,1	0,521	6,28	0,578	4,40
3	0,483	6,47	0,566	3,73	0,620	2,92
4	0,519	4,39	0,599	2,87	0,649	2,37
5	0,546	3,48	0,624	2,45	0,672	2,090
6	0,569	2,98	0,644	2,202	0,690	1,916
7	0,588	2,66	0,661	2,035	0,705	1,797
8	0,604	2,440	0,675	1,916	0,718	1,711
9	0,618	2,277	0,688	1,826	0,729	1,645
10	0,630	2,154	0,699	1,755	0,739	1,593
11	0,641	2,056	0,708	1,698	0,748	1,550
12	0,651	1,976	0,717	1,651	0,755	1,515
13	0,660	1,910	0,725	1,611	0,762	1,485
14	0,669	1,854	0,732	1,577	0,769	1,460
15	0,676	1,806	0,739	1,548	0,775	1,437
16	0,683	1,764	0,745	1,522	0,780	1,418
17	0,690	1,727	0,750	1,499	0,785	1,400
18	0,696	1,695	0,756	1,479	0,790	1,385
19	0,702	1,666	0,760	1,460	0,794	1,370
20	0,707	1,640	0,765	1,444	0,798	1,358
21	0,712	1,617	0,769	1,429	0,802	1,346
22	0,717	1,595	0,773	1,416	0,805	1,335
23	0,722	1,576	0,777	1,402	0,809	1,326
24	0,726	1,558	0,781	1,391	0,812	1,316
25	0,730	1,541	0,784	1,380	0,815	1,308
26	0,734	1,526	0,788	1,371	0,818	1,300
27	0,737	1,512	0,791	1,361	0,820	1,293
28	0,741	1,499	0,794	1,352	0,823	1,286
29	0,744	1,487	0,796	1,344	0,825	1,279
30	0,748	1,475	0,799	1,337	0,828	1,274

5. Критические точки $\chi^2_{\alpha,\nu}$ распределения χ^2, отвечающие равенству $\Pr\left(\chi^2 > \chi^2_{\alpha,\nu}\right) = \alpha$

Число степеней свободы ν	Уровень значимости α (двусторонняя критическая область)					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,001	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

6. Правосторонние квантили $\chi^2_{1-\alpha,\nu}$ закона χ^2 с ν степенями свободы, отвечающие равенству $\Pr\left(\chi^2 > \chi^2_{1-\alpha,\nu}\right) = 1-\alpha$

ν	$1-\alpha$							
	0,70	0,80	0,90	0,95	0,975	0,990	0,995	0,999
1	1,07	1,64	2,71	3,84	5,02	6,63	7,88	10,8
2	2,41	3,22	4,61	5,99	7,38	9,21	10,6	13,8
3	3,67	4,64	6,25	7,81	9,35	11,3	12,8	16,3
4	4,88	5,99	7,78	9,49	11,1	13,3	14,9	18,5
5	6,06	7,29	9,24	11,1	12,8	15,1	16,7	20,5
6	7,23	8,56	10,6	12,6	14,4	16,8	18,5	22,5
7	8,38	9,80	12,0	14,1	16,0	18,5	20,3	24,3
8	9,52	11,0	13,4	15,5	17,5	20,1	22,0	26,1
9	10,7	12,2	14,7	16,9	19,0	21,7	23,6	27,9
10	11,8	13,4	16,0	18,3	20,5	23,2	25,2	29,6
11	12,9	14,6	17,3	19,7	21,9	24,7	26,8	31,3
12	14,0	15,8	18,5	21,0	23,3	26,2	28,3	32,9
13	15,1	17,0	19,8	22,4	24,7	27,7	29,8	34,5
14	16,2	18,2	21,1	23,7	26,1	29,1	31,3	36,1
15	17,3	19,3	22,3	25,0	27,5	30,6	32,8	37,7
16	18,4	20,5	23,5	26,3	28,8	32,0	34,3	39,3
17	19,5	21,6	24,8	27,6	30,2	33,4	35,7	40,8
18	20,6	22,8	26,0	28,9	31,5	34,8	37,2	42,3
19	21,7	23,9	27,2	30,1	32,9	36,2	38,6	43,8
20	22,8	25,0	28,4	31,4	34,2	37,6	40,0	45,3
21	23,9	26,9	29,6	32,7	35,5	38,9	41,4	46,8
22	24,9	27,3	30,8	33,9	36,8	40,3	42,8	48,3
23	26,0	28,4	32,0	35,2	38,1	41,6	44,2	49,7
24	27,1	29,6	33,2	36,4	39,4	43,0	45,6	51,2
25	28,2	30,7	34,4	37,7	40,6	44,3	46,9	52,6
26	29,2	31,8	35,6	38,9	41,9	45,6	48,3	54,1
27	30,3	32,9	36,7	40,1	43,2	47,0	49,6	55,5
28	31,4	34,0	37,9	41,3	44,5	48,3	51,0	56,9
29	32,5	35,1	39,1	42,6	45,7	49,6	52,3	58,3
30	33,5	36,3	40,3	43,8	47,0	50,9	53,7	59,7

7. Квантили $t_{\alpha,\nu}$ распределения Стьюдента с ν степенями свободы, определяемые вероятностью $\Pr(|t| > t_{\alpha,\nu}) = \alpha$ (двусторонняя критическая область)

Число степеней свободы ν	Уровень значимости α					
	0,050	0,025	0,020	0,010	0,005	0,002
1	12,706	25,452	31,821	63,657	127,3	318,3
2	4,303	6,205	6,965	9,925	14,09	22,33
3	3,182	4,177	4,541	5,841	7,453	10,21
4	2,776	3,495	3,747	4,604	5,598	7,173
5	2,571	3,163	3,365	4,032	4,773	5,893
6	2,447	2,969	3,143	3,707	4,317	5,208
7	2,365	2,841	2,998	3,499	4,029	4,785
8	2,306	2,752	2,896	3,355	3,833	4,501
9	2,262	2,685	2,821	3,250	3,690	4,297
10	2,228	2,634	2,764	3,169	3,581	4,144
11	2,201	2,593	2,718	3,106	3,497	4,025
12	2,179	2,560	2,681	3,055	3,428	3,930
13	2,160	2,533	2,650	3,012	3,372	3,852
14	2,145	2,510	2,624	2,977	3,326	3,787
15	2,131	2,490	2,602	2,947	3,286	3,733
16	2,120	2,473	2,583	2,921	3,252	3,686
17	2,110	2,458	2,567	2,898	3,222	3,646
18	2,101	2,445	2,552	2,878	3,197	3,610
19	2,093	2,433	2,539	2,861	3,174	3,579
20	2,086	2,423	2,528	2,845	3,153	3,552
21	2,080	2,414	2,518	2,831	3,135	3,527
22	2,074	2,405	2,508	2,819	3,119	3,505
23	2,069	2,398	2,500	2,807	3,104	3,485
24	2,064	2,391	2,492	2,797	3,091	3,467
25	2,060	2,385	2,485	2,787	3,078	3,450
26	2,056	2,379	2,479	2,779	3,067	3,435
27	2,052	2,373	2,473	2,771	3,057	3,421
28	2,048	2,368	2,467	2,763	3,047	3,408
29	2,045	2,364	2,462	2,756	3,038	3,396
30	2,042	2,360	2,457	2,750	3,030	3,385
40	2,021	2,329	2,423	2,705	2,971	3,307
60	2,000	2,300	2,390	2,660	2,915	3,232
120	1,980	2,980	2,358	2,617	2,860	3,107
∞	1,960	2,241	2,326	2,576	2,807	3,090

8. Левосторонние квантили распределения Стьюдента $t_p(k)$ с k степенями свободы, отвечающие равенству $\Pr(t < t_p(k)) = p$

	p						
k	0,750	0,900	0,950	0,975	0,990	0,995	0,999
1	1,000	3,078	6,314	12,706	31,821	63,657	318,3
2	0,817	1,886	2,920	4,303	6,965	9,925	22,33
3	0,765	1,638	2,353	3,182	4,541	5,841	10,21
4	0,741	1,533	2,132	2,776	3,747	4,604	7,173
5	0,727	1,476	2,015	2,571	3,365	4,032	5,893
6	0,718	1,440	1,943	2,447	3,143	3,707	5,208
7	0,711	1,415	1,895	2,365	2,998	3,499	4,785
8	0,706	1,397	1,860	2,306	2,896	3,355	4,501
9	0,703	1,383	1,833	2,262	2,821	3,250	4,297
10	0,700	1,372	1,812	2,228	2,764	3,169	4,144
11	0,697	1,363	1,796	2,201	2,718	3,106	4,025
12	0,696	1,356	1,782	2,179	2,681	3,055	3,930
13	0,694	1,350	1,771	2,160	2,650	3,012	3,852
14	0,692	1,345	1,761	2,145	2,624	2,977	3,787
15	0,691	1,341	1,753	2,131	2,602	2,947	3,733
16	0,690	1,337	1,746	2,120	2,583	2,921	3,686
17	0,689	1,333	1,740	2,110	2,567	2,898	3,646
18	0,688	1,330	1,734	2,101	2,552	2,878	3,610
19	0,688	1,328	1,729	2,093	2,539	2,861	3,579
20	0,687	1,325	1,725	2,086	2,528	2,845	3,552
21	0,686	1,323	1,721	2,080	2,518	2,831	3,527
22	0,686	1,321	1,717	2,074	2,508	2,819	3,505
23	0,685	1,319	1,714	2,069	2,500	2,807	3,485
24	0,686	1,318	1,711	2,064	2,492	2,797	3,467
25	0,684	1,316	1,708	2,060	2,485	2,787	3,450
26	0,684	1,315	1,706	2,056	2,479	2,779	3,435
27	0,686	1,314	1,703	2,052	2,473	2,771	3,421
28	0,683	1,313	1,701	2,048	2,467	2,763	3,408
29	0,683	1,311	1,699	2,045	2,462	2,756	3,390
30	0,686	1,310	1,697	2,042	2,457	2,750	3,385
40	0,681	1,303	1,684	2,021	2,423	2,704	3,307
60	0,679	1,296	1,671	2,000	2,390	2,660	3,232
120	0,677	1,289	1,658	1,980	2,358	2,617	3,160
∞	0,675	1,282	1,645	1,960	2,326	2,576	3,070

9. Квантили распределения $F_{\alpha;\,k_1,k_2}$ Фишера-Снедекора; k_1 – число степеней свободы большей дисперсии; k_2 – число степеней свободы меньшей дисперсии

Уровень значимости $\alpha = 0{,}05$						
	k_1					
k_2	1	2	3	4	5	6
1	161	200	216	225	230	234
2	18,51	19,00	19,16	19,25	19,30	19,33
3	10,13	9,55	9,28	9,12	9,01	8,94
4	7,71	6,94	6,59	6,39	6,26	6,16
5	6,61	5,79	5,41	5,19	5,05	4,95
6	5,99	5,14	4,76	4,53	4,39	4,28
7	5,59	4,74	4,35	4,12	3,97	3,87
8	5,32	4,46	4,07	3,84	3,69	3,58
9	5,12	4,26	3,86	3,63	3,48	3,37
10	4,96	4,10	3,71	3,48	3,33	3,22
11	4,84	3,98	3,59	3,36	3,20	3,09
12	4,75	3,88	3,49	3,26	3,11	3,00
13	4,67	3,80	3,41	3,18	3,02	2,92
14	4,60	3,74	3,34	3,11	2,96	2,85
15	4,54	3,68	3,29	3,06	2,90	2,79
16	4,49	3,63	3,24	3,01	2,85	2,74

Уровень значимости $\alpha = 0{,}05$						
	k_1					
k_2	7	8	9	10	11	12
1	237	239	241	242	243	244
2	19,36	19,37	19,38	19,39	19,40	19,41
3	8,88	8,84	8,81	8,78	8,76	8,74
4	6,09	6,04	6,00	5,96	5,93	5,91
5	4,88	4,82	4,78	4,74	4,70	4,68
6	4,21	4,15	4,10	4,06	4,03	4,00
7	3,79	3,73	3,68	3,63	3,60	3,57
8	3,50	3,44	3,39	3,34	3,31	3,28
9	3,29	3,23	3,18	3,13	3,10	3,07
10	3,14	3,07	3,02	2,97	2,94	2,91
11	3,01	2,95	2,90	2,86	2,82	2,79
12	2,92	2,85	2,80	2,76	2,72	2,69
13	2,84	2,77	2,72	2,67	2,63	2,60
14	2,77	2,70	2,65	2,60	2,56	2,53
15	2,70	2,64	2,59	2,55	2,51	2,48
16	2,66	2,59	2,54	2,49	2,45	2,42

10. Предельное распределение Колмогорова

$$K(\lambda) \;=\; \sum_{\nu=-\infty}^{\infty} (-1)^{\nu} \exp\left(-2\nu^2\lambda^2\right)$$

λ	$K(\lambda)$		λ	$K(\lambda)$		λ	$K(\lambda)$
0,30	0,0000		0,90	0,6073		1,50	0,9778
0,35	0,0003		0,95	0,6725		1,55	0,9836
0,40	0,0028		1,00	0,7300		1,60	0,9880
0,45	0,0126		1,05	0,7798		1,65	0,9914
0,50	0,0361		1,10	0,8223		1,70	0,9938
0,55	0,0772		1,15	0,8580		1,75	0,9956
0,60	0,1357		1,20	0,8878		1,80	0,9969
0,65	0,2080		1,25	0,9121		1,85	0,9979
0,70	0,2888		1,30	0,9319		1,90	0,9985
0,75	0,3728		1,35	0,9478		1,95	0,9990
0,80	0,4559		1,40	0,9603		2,00	0,99933
0,85	0,5347		1,45	0,9702		2,05	0,99999

11. Квантили λ_α распределения Колмогорова

$$\Pr\left(\lambda \geq \lambda_\alpha\right) \;=\; \alpha$$

Уровень значимости α	0,20	0,10	0,05	0,02	0,01	0,001
λ_α	1,070	1,224	1,358	1,520	1,627	1,950

12. Квантили λ_α распределения Смирнова-Колмогорова

$$\Pr(\lambda \geq \lambda_\alpha) = \alpha$$

Уровень значимости α	0,20	0,10	0,05	0,02	0,01	0,001
λ_α	0,892	1,075	1,223	1,395	1,518	1,860

13. Квантили r_α числа знаков

$$\Pr(r \geq r_\alpha) = \alpha$$

n	Уровень значимости α				n	Уровень значимости α		
	0,01	0,05	0,10			0,01	0,05	0,10
6		0	0		21	4	5	6
7		0	0		22	4	5	6
8	0	0	1		23	4	6	7
9	0	1	1		23	4	6	7
10	0	1	1		24	5	6	7
11	0	1	2		25	5	7	7
12	1	2	2		26	6	7	8
13	1	2	2		27	6	7	8
14	1	2	3		28	6	8	9
15	2	3	3		29	7	8	9
16	2	3	4		30	7	9	10
17	2	4	4		31	7	9	10
18	3	4	4		32	8	9	10
19	3	4	5		33	8	10	11
20	3	5	5		34	8	10	11

Список литературы

Учебники

1. Вентцель Е.С. *Теория вероятностей.* — М.: Наука, 1964.

2. Гнеденко Б.В. *Курс теории вероятностей.* — М.: Наука, 1961.

3. Коваленко И.М., Филиппова А.А. *Теория вероятностей и математическая статистика.* — М.: Высшая школа, 1973.

4. Пугачев В.С. *Теория вероятностей и математическая статистика.* — М.: Наука, 1979.

5. Гихман И.И., Скороход А.В., Ядренко М.И. *Теория вероятностей и математическая статистика.* — Киев: Вищ. шк., 1979.

6. Ивченко Г.И., Медведев Ю.И. *Математическая статистика: Учебное пособие для втузов.* — М.: Высш. шк., 1984.

7. Четыркин Е.М., Калихман И.Л. *Вероятность и статистика.* — М.: Финансы и статистика, 1982.

8. Гмурман В.Е. *Теория вероятностей и математическая статистика.* — М.: Высш. шк., 2000.

Задачники и пособия

9. Емельянов Г.В., Скитович В.П. *Задачник по теории вероятностей и математической статистике.* — Л.: Изд–во ЛГУ, 1967.

10. Герасимович А.И., Матвеева Я.И. *Математическая статистика.* — Минск: Вышэйшая шк., 1978.

11. Гмурман В.Е. *Руководство к решению задач по теории вероятностей и математической статистике.* — М.: Высш. шк., 2000.

12. Зубков А.М., Севастьянов Б.А., Чистяков В.П. *Сборник задач по теории вероятностей.* — М.: Наука, 1989.

13. Сборник задач по математике. Теория вероятностей и математическая статистика / Под ред. А.В. Ефимова. — М.: Наука, 1990.

14. Харин Ю.С., Степанова М.Д. *Практикум на ЭВМ по математической статистике.* — Минск: "Университетское", 1987.

15. Турчин В.М. *Математична статистика в прикладах і задачах: Навчальний посібник.* — К.: НМК ВО, 1993.

16. Мазманишвили А.С. *Теория вероятностей: Учебное пособие.* — Харьков: ХГПУ, 1994.

17. Боровиков В.П., Боровиков. *STATISTIKA. — Статистический анализ и обработка данных в среде Windows.* — М.: "Филин", 1997.

18. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. *Справочник по теории вероятностей и математической статистике.* — М.: Наука, 1985.

19. Корн Г., Корн Т. *Справочник по математике.* — М.: Наука, 1983.

20. Абрамович М., Стиган И. *Справочник по специальным функциям.* — М.: Наука, 1979.

21. Большев Л.Н., Смирнов Н.В. *Таблицы математической статистики.* — М.: Наука, 1983.

Дополнительная литература

22. Чебышев П.Л. *Теория вероятностей.* — М.–Л.: Изд–во АН СССР, 1936.

23. Дюге Д. *Теоретическая и прикладная статистика.* — М.: Наука, 1972.

24. Крамер Г. *Математические методы статистики.* — М.: Мир, 1975.

25. Боровков А.А. *Математическая статистика. Оценка параметров. Проверка гипотез.* — М.: Наука, 1984.

26. Феллер В. *Введение в теорию вероятностей и ее применение.* — М.: Мир, 1984. – Т.1; 1984. – Т.2.

27. Митропольский А.К. *Техника статистических вычислений.* — М.: Наука, 1971.

28. Смирнов Н.В., Дунин–Барковский И.В. *Курс теории вероятностей и математической статистики для технических приложений.* — М.: Наука, 1969.

29. Яноши Л. *Теория и практика обработки результатов измерений.* — М.: Мир, 1968.

30. Анго А. *Математика для электро– и радиоинженеров.* — М.: Наука, 1965.

31. Воинов В.Г., Никулин М.С. *Несмещенные оценки и их применения.* — М.: Наука, 1989.

32. Розанов Ю.А. *Теория вероятностей, случайные процессы и математическая статистика.* — М.: Наука, 1985.

33. Мартынов Г.В. *Критерий омега–квадрат.* — М.: Наука, 1978.

34. Шеффе Г. *Дисперсионный анализ.* — М.: Наука, 1980.

35. Кокс Д., Хинкли Д. *Теоретическая статистика.* — М.: Мир, 1984.

36. Бикел П., Доксам К. *Математическая статистика.* — М.: Финансы и статистика, 1983.

37. Ван дер Варден Б.Л. *Математическая статистика.* — М.: Изд–во иностр. лит., 1960.

38. Себер Дж. *Линейный регрессионный анализ.* — М.: Мир, 1982.

39. Хьюбер П. *Робастность в статистике.* — М.: Мир, 1984.

40. Налимов В.В. *Теория эксперимента.* — М.: Наука, 1971.

41. Секей Г. *Парадоксы в теории вероятностей и математической статистике.* — М.: Мир, 1990.

42. Уилкс С.С. *Математическая статистика.* — М.: Наука, 1967.

Навчальне видання

МАЗМАНІШВІЛІ Олександр Сергійович

МАТЕМАТИЧНА СТАТИСТИКА

Навчальний посібник до практичних занять

Російською мовою

Роботу до видання рекомендував О.В. Горілий

Редактор О.С. Самініна

План 2010, поз. 68

Підп. до друку 20.10.2009 Формат 60×94 1/16 Папір офсетн.
Друк – ризографія. Гарнітура Times New Roman. Ум. друк. арк. 12,0.
Обл.–вид. арк. 15,0. Наклад 200 прим. Зам. № Ціна договірна

Видавничий центр НТУ "ХПІ", 61002, Харків, вул. Фрунзе, 21
Свідоцтво про реєстрацію ДК № 116 від 10.07.2000

Друкарня НТУ "ХПІ", 61002, Харків, вул. Фрунзе, 21

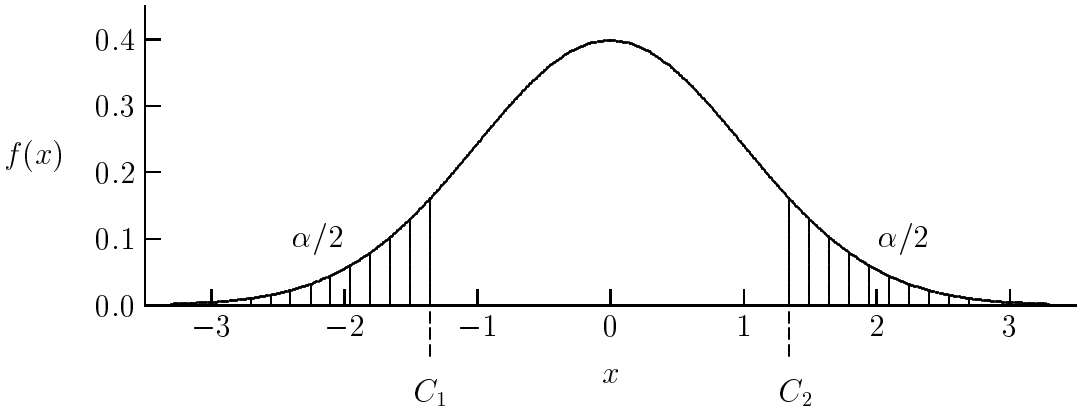
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ

НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
"ХАРЬКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ"

А.С. МАЗМАНИШВИЛИ

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

УЧЕБНОЕ ПОСОБИЕ К ПРАКТИЧЕСКИМ ЗАНЯТИЯМ



Харьков НТУ "ХПИ" 2010