МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное
образовательное учреждение
высшего профессионального образования

«КАЗАНСКИЙ ГОСУДАРСТВЕННЫЙ
ЭНЕРГЕТИЧЕСКИЙ УНИВЕРСИТЕТ»

Т.А. ГРИГОРЯН, Е.В. ЛИПАЧЕВА

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ЧАСТЬ II

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебно-методическое пособие

Казань 2013

УДК 517.1
ББК 22.1
Г83

Рецензенты:

кандидат физико-математических наук, доцент Казанского государственного энергетического университета

Г83 Григорян Т.А., Липачева Е.В. Теория вероятностей и математическая статистика. Часть II. Математическая статистика: Учебно-методическое пособие/Т.А. Григорян, Е.В. Липачева – Казань: Казан. гос. энерг. ун-т, 2012. – 154 с.

Учебно-методическое пособие охватывает классические разделы математической статистики. Каждая глава начинается с подробного изложения теоретического материала, затем приводятся примеры решения задач, и завершается глава набором задач для самостоятельного решения. Все задачи снабжены ответами.

Учебно-методическое пособие предназначено для студентов второго курса инженерно-экономических специальностей и может быть рекомендовано для использования при проведении практических занятий.

УДК 517.1
ББК 22.1

ПРЕДИСЛОВИЕ

Книга представляет собой вторую часть учебно-методического пособия по дисциплине «Теория вероятностей и математическая статистика», содержащую классические разделы математической статистики.

Математическая статистика – это наука, которая, основываясь на методах теории вероятностей, занимается систематизацией и обработкой статистических данных для получения научных и практических выводов.

Первая задача математической статистики – указать способы сбора и группировки статистических сведений.

Вторая задача – разработать методы анализа статистических данных, в зависимости от целей исследования.

В первых трех главах пособия описывается статистический метод исследования. И хотя черты статистического метода в применении к объектам различной природы весьма своеобразны, можно выделить общие черты статистического метода, формальную математическую сторону статистического исследования, включающую в себя рассмотрение распределения количественных признаков, применение выборочного метода, связь статистических распределений с вероятностными, оценку параметров распределений, проверку вероятностных гипотез.

В четвертой и пятой главах рассматриваются дальнейшие задачи математической статистики, такие ее разделы, как регрессионный анализ и метод математического моделирования.

Глава 1

МЕТОДЫ СТАТИСТИЧЕСКОГО ОПИСАНИЯ РЕЗУЛЬТАТОВ НАБЛЮДЕНИЙ

Генеральная и выборочная совокупности

Пусть требуется изучить совокупность однородных объектов относительно некоторого качественного или количественного признака, характеризующего этот объект. Иногда проводят сплошное обследование, т.е. исследуется каждый из объектов совокупности относительно признака, которым интересуются. Но на практике сплошное обследование применяется крайне редко, например, если совокупность содержит очень большое число элементов. В таких случаях из всей совокупности случайным образом отбирают ограниченное число объектов и подвергают их изучению.

Выборочной совокупностью (или *выборкой*) называют совокупность случайно отобранных объектов.

Генеральной совокупностью называют совокупность элементов, из которых производится выборка.

Объемом совокупности (выборочной или генеральной) называется число объектов этой совокупности. Объем генеральной совокупности принято обозначать буквой N, объем выборочной совокупности обозначают n.

Для того, чтобы по данным выборки можно было достаточно уверенно судить об интересующем нас признаке генеральной совокупности, необходимо, чтобы объекты выборки правильно ее представляли, т.е. выборка должна быть *репрезентативной*. В силу закона больших чисел можно утверждать, что выборка будет репрезентативной, если ее осуществлять случайно, т.е. все объекты генеральной совокупности должны иметь одинаковую вероятность попасть в выборку. Для этого существуют различные *виды отбора* выборки:

1. Простым случайным отбором называется отбор, при котором объекты извлекаются по одному из всей генеральной совокупности.

2. Механическим называется отбор, при котором генеральная совокупность делится на столько частей, сколько объектов должно войти в выборку, и из каждой части случайным образом отбирается один элемент.

3. Серийным называется отбор, при котором объекты из генеральной совокупности отбираются «сериями», которые подвергаются сплошному обследованию.

4. Типическим называется отбор, при котором объекты отбирают не из всей генеральной совокупности, а из каждой «типической» ее части.

На практике часто применяется комбинированный отбор, при котором сочетается сразу несколько видов отбора, образующих различные фазы выборочного обследования. Например, иногда разбивают генеральную совокупность на серии одинакового объема, затем простым случайным

отбором выбирают несколько серий и, наконец, из каждой серии простым случайным отбором извлекают отдельные объекты. Существуют и другие методы организации выборки. Выборки разделяют на повторные и бесповторные.

Повторной называют выборку, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность.

Бесповторной называют выборку, при которой отобранный объект в генеральную совокупность не возвращается.

На практике обычно имеют дело с бесповторными выборками.

Статистическое распределение выборки

Пусть из генеральной совокупности извлечена выборка, причем значение x_1 наблюдалось n_1 раз, x_2 – n_2 раз, x_k – n_k раз и $\sum_{i=1}^{k} n_i = n$ – объем выборки. Наблюдаемые значения x_i называются *вариантами*, числа n_i – частотами наблюдаемых значений, а отношения $\frac{n_i}{n} = w_i$ – относительными частотами вариант x_i. Очевидно, что $\sum_{i=1}^{k} w_i = 1$. Элементы выборки, расположенные в возрастающем порядке, называются *вариационным рядом*. Вариационный ряд называется *дискретным*, если его члены принимают конкретные изолированные значения. Если члены ряда могут заполнить некоторый интервал, то такой ряд называется *непрерывным*. Наименьшее и наибольшее значения вариационного ряда обозначают $x_{\min}$ и $x_{\max}$ и называют *крайними членами* вариационного ряда.

Статистическим распределением выборки называется перечень вариант и соответствующих им частот или относительных частот. В случае непрерывного вариационного ряда или когда объем выборки очень велик, статистическое распределение задается в виде последовательности интервалов и соответствующих им частот. В этом случае весь интервал наблюдаемых значений $[x_{\min}, x_{\max}]$ разбивают на k частичных интервалов $[c_0, c_1)$, $[c_1, c_2)$,…,$[c_{k-1}, c_k]$ одинаковой длины h. В качестве частоты, соответствующей интервалу, принимают сумму частот, попавших в этот интервал.

Согласно формуле Стерджеса рекомендуемое число интервалов разбиения

$$k \approx 1 + \log_2 n, \qquad (1.1)$$

а длины частичных интервалов

$$h = \frac{x_{\max} - x_{\min}}{k}. \qquad (1.2)$$

Понятно, что группировка связана с потерей части полезной информации, заключенной в выборке. Однако она имеет и свои преимущества. Например, в случае очень большого объема выборки при группировке значительно сокращается объем вычислений.

Отметим, что распределение выборки является некоторым приближенным распределением генеральной совокупности.

Эмпирическая функция распределения

Пусть известно статистическое распределение частот количественного признака X. Введем обозначения:

$n_x -$ число наблюдений, при которых наблюдалось значение признака, меньшее x,

$n -$ общее число наблюдений (объем выборки).

Ясно, что относительная частота события $X < x$ равна $\frac{n_x}{n}$. Если x будет меняться, то, вообще говоря, будет изменяться и относительная частота, т.е. относительная частота $\frac{n_x}{n}$ есть функция от x. Так как эта функция находится эмпирическим (опытным) путем, то ее называют эмпирической.

Эмпирической функцией распределения (функцией распределения выборки) называют функцию $F_n(x)$ (или $F^*(x)$), определяющую для каждого значения x относительную частоту события $X < x$, т.е.

$$F_n(x) = \frac{n_x}{n},$$

где $n_x -$ число вариант, меньших x; $n -$ объем выборки.

В отличие от эмпирической функции распределения, функцию распределения генеральной совокупности $F(x)$ называют теоретической функцией распределения. Различие между эмпирической и теоретической функциями состоит в том, что теоретическая функция $F(x)$ определяет вероятность события $X < x$, а эмпирическая функция $F_n(x)$ определяет относительную частоту этого события. Из теоремы Бернулли следует, что относительная частота события $X < x$, т.е. $F_n(x)$ *стремится по вероятности к вероятности* $F(x)$ *этого же события.* Поэтому эмпирическую функцию распределения используют для приближенного представления теоретической функции распределения генеральной совокупности.

Из определения эмпирической функции распределения вытекают следующие свойства:

1) $0 \le F_n(x) \le 1$;

2) $F_n(x)-$ неубывающая функция;

3) если x_1- наименьшая варианта, а x_k- наибольшая варианта, то $F_n(x)=0$ при $x \le x_1$ и $F_n(x)=1$ при $x > x_k$.

Итак, *эмпирическая функция распределения выборки служит для оценки теоретической функции распределения генеральной совокупности*.

Статистика $D_n = \sup|F_n(x)-F(x)|$ называется *отклонением* эмпирической функции распределения от теоретической.

Полигон и гистограмма

Графически статистическое распределение может быть представлено в виде полигона, гистограммы или графика накопленных частот.

Полигоном частот называют ломаную, отрезки которой соединяют точки $(x_1,n_1),(x_2,n_2),\ldots,(x_k,n_k)$. *Полигоном относительных частот* называют ломаную, отрезки которой соединяют точки $(x_1,w_1),(x_2,w_2),\ldots,(x_k,w_k)$ (рис. 1.1). Полигоны обычно служат для изображения в случае дискретного распределения.

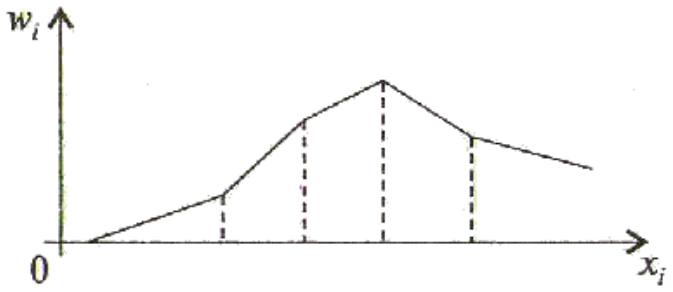


Рис. 1.1

Гистограммой частот называется ступенчатая фигура, состоящая из прямоугольников, основанием которых служат частичные интервалы длины h, а высоты равны $\frac{n_i}{h}$. Площадь гистограммы частот равна сумме всех частот, т.е. объему выборки.

Для построения *гистограммы относительных частот* за высоту прямоугольников берут величину $\frac{w_i}{h}$ (рис. 1.2). Площадь гистограммы относительных частот равна сумме всех относительных частот, т.е. единице.

Гистограммы обычно служат для изображения выборки в случае непрерывного распределения. Если на гистограмме частот соединить середины верхних сторон прямоугольников, то полученная ломаная образует

полигон частот. Аналогично получается полигон относительных частот из гистограммы относительных частот (рис. 1.3).

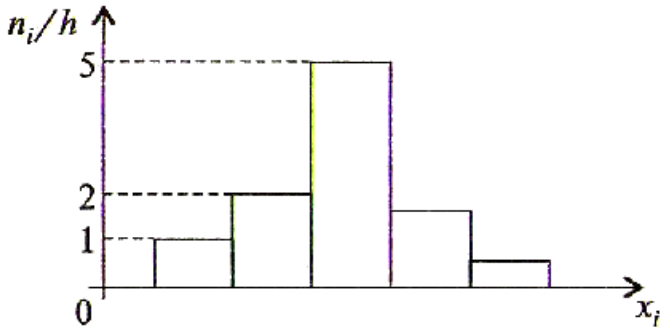


Рис. 1.2

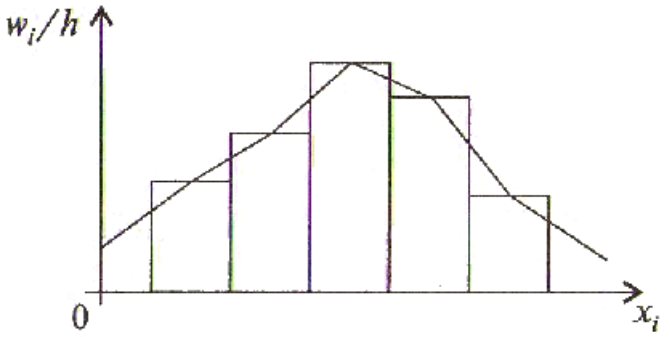


Рис. 1.3

При построении гистограммы в реальных исследованиях следует понимать, что формула Стерджеса для числа интервалов разбиения k дает лишь рекомендацию, а не строгое правило. Если взять слишком маленькое число k, то гистограмма получится грубой, плохо отражающей свойства распределения. При слишком больших k гистограмма становится «колючей», имеет игольчатый вид и может распасться на отдельные «иглы» и пустые интервалы. Оптимальное значение k в общем случае неизвестно – оно зависит как от типа распределения, так и от конкретной выборки.

Графиком накопленных частот называется фигура, строящаяся аналогично гистограмме относительных частот, с тем различием, что для расчета высот прямоугольников вместо относительных частот берутся накопленные относительные частоты, т.е. величины

$$w_i^c = \sum_{j=1}^{i} w_j \,. \qquad (1.3)$$

Эти величины не убывают и, таким образом, график накопленных частот имеет вид ступенчатой «лестницы». График эмпирической функции распределения проходит через правые верхние углы прямоугольников, т.е.

через точки вида (c_i, w_i^c). *График накопленных частот и эмпирическая функция распределения на практике используются для приближения теоретической функции распределения.*

Примеры решения задач к главе 1

1. По данным выборки составить дискретное статистическое распределение: 1, 10, 15, 13, 9, 13, 7, 6, 1, 2, 10, 9, 15, 1, 7, 7, 7, 8, 1, 13.

*Решение***.** Объем выборки $n = 20$. Построим дискретный вариационный ряд, для этого надо расположить все значения выборки в возрастающем порядке. Получим

1, 1, 1, 1, 2, 6, 7, 7, 7, 7, 8, 9, 9, 10, 10, 13, 13, 13, 15, 15.

Теперь, по вариационному ряду, составим таблицу. В первой строке записываем различные варианты ряда, во второй строке – частоты соответствующих вариант, т.е. число повторений каждой варианты. Например, варианта $x_1 = 1$ появляется четыре раза в выборке, значит частота этой варианты $n_1 = 4$. Получим следующую таблицу:

x_i	1	2	6	7	8	9	10	13	15
n_i	4	1	1	4	1	2	2	3	2

Для проверки можно сложить все частоты n_i, сумма должна получиться равной числу n, т.е. объему выборки. Проверим,

$$4+1+1+4+1+2+2+3+2=20.$$

2. По данным выборки объема $n = 60$ построить статистическое распределение.

1; 2,5; 1,5; 4; 6; 6,1; 3,7; 6,5; 2,5; 8,1; 7,8; 9,1; 5,1; 10; 13,4;
8,2; 4; 9,4; 11; 14,8; 13,2; 11,7; 15; 14,8; 12,8; 9,7; 13; 14; 3,4; 13,4;
8,4; 9,4; 9,1; 10,5; 4,8; 5; 1,3; 2,8; 8,6; 7; 14; 15; 7,3; 10; 12,7;
1,7; 5,6; 7,5; 8,9; 11,3; 13,7; 3; 6,3; 4; 3,2; 6,7; 15; 1,2; 7,9; 9,5.

*Решение***.** Так как объем выборки достаточно велик и выборка содержит большое количество различных элементов, то данные выборки лучше сгруппировать. Найдем сначала крайние элементы ряда: $x_{\min} = 1$, $x_{\max} = 15$. Значит, отрезок [1; 15] надо разбить на k интервалов одинаковой длины. Чтобы найти оптимальное число интервалов разбиения, воспользуемся формулой Стерджеса:

$$k \approx 1 + \log_2 n = 1 + \log_2 60 \approx 7.$$

То есть отрезок [1, 15] разбиваем на 7 интервалов. Длина этих интервалов одинакова и вычисляется по формуле

$$h = \frac{x_{\max} - x_{\min}}{k} = \frac{15-1}{7} = 2.$$

Получим следующие интервалы:

[1; 3), [3; 5), [5; 7), [7; 9), [9; 11), [11; 13), [13; 15].

Теперь надо найти частоты соответствующих интервалов. Частота каждого интервала – это количество вариант, вошедших в этот интервал. Например, в первый интервал [1;3) попадают элементы 1; 2,5; 1,5; 2,5; 1,3; 2,8; 1,7; 1,2. Всего 8 элементов, значит, частота интервала [1;3) равна 8. Проводя аналогичные рассуждения, получим следующую таблицу:

Номер интервала	Частичный интервал	Частоты n_i
1	1–3	8
2	3–5	8
3	5–7	8
4	7–9	10
5	9–11	9
6	11–13	5
7	13–15	12

Эта таблица задает статистическое распределение исходной выборки.

3. Построить полигон частот по данному распределению выборки:

x_i	2	3	5	6
n_i	10	15	5	20

Решение. Отложим на оси абсцисс варианты x_i, а на оси ординат – соответствующие им частоты n_i, затем соединим последовательно точки (x_i, n_i). Полигон частот изображен на рис. 1.4.

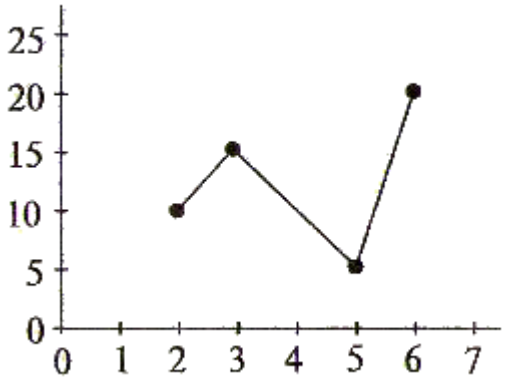


Рис. 1.4

4. Построить гистограмму частот по данному распределению выборки:

Номер интервала	Частичный интервал	Частоты n_i
1	2 – 7	5
2	7 – 12	10
3	12 – 17	25
4	17 – 22	6
5	22 – 27	4

*Решение***.** Вначале найдем плотности частот, т.е. величины $\frac{n_i}{h}$. Для данного примера $h = 5$.

Таким образом, получаем

Номер интервала	Плотность частоты
1	1
2	2
3	5
4	1,2
5	0,8

Отложим на оси абсцисс интервалы длиной $h=5$ каждый, а затем проведем над ними отрезки, параллельные оси x, на расстояниях от нее, равных соответствующим значениям плотности частоты (ось ординат), то есть гистограмма частот состоит из прямоугольников, ширина каждого прямоугольника равна длине интервала $h=5$, а высота равна плотности соответствующей частоты (см. рис. 1.5).

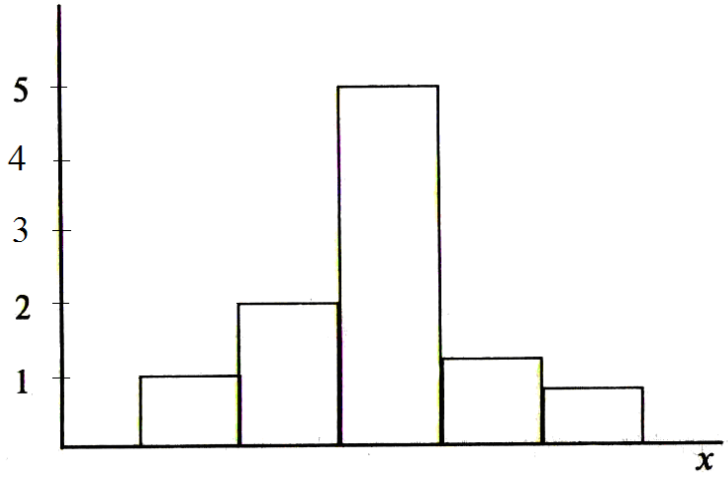


Рис. 1.5

5. Построить гистограмму относительных частот и график накопленных частот по данному распределению выборки:

Номер интервала	Частичный интервал	Частоты n_i
1	10–15	2
2	15–20	4
3	20–25	8
4	25–30	4
5	30–35	2

*Решение***.** Найдем относительные частоты и плотности относительных частот:

Частота w_i	Плотность $\frac{w_i}{h}$	Накопленные частоты w_i^c
$w_1 = \frac{n_1}{n} = \frac{2}{20} = 0,1$	$\frac{w_1}{h} = \frac{0,1}{5} = 0,02$	$w_1^c = 0,1$
$w_2 = \frac{n_2}{n} = \frac{4}{20} = 0,2$	$\frac{w_2}{h} = \frac{0,2}{5} = 0,04$	$w_2^c = 0,1 + 0,2 = 0,3$
$w_3 = \frac{n_3}{n} = \frac{8}{20} = 0,4$	$\frac{w_3}{h} = \frac{0,4}{5} = 0,08$	$w_3^c = 0,3 + 0,4 = 0,7$
$w_4 = \frac{n_4}{n} = \frac{4}{20} = 0,2$	$\frac{w_4}{h} = \frac{0,2}{5} = 0,04$	$w_4^c = 0,7 + 0,2 = 0,9$
$w_5 = \frac{n_5}{n} = \frac{2}{20} = 0,1$	$\frac{w_5}{h} = \frac{0,1}{5} = 0,02$	$w_5^c = 0,9 + 0,1 = 1$

Построим на оси абсцисс частичные интервалы $h = 5$, затем проведем параллельно им отрезки, отстоящие от оси x на соответствующие значения плотности относительной частоты (см. рис. 1.6).

График накопленных частот строится аналогично гистограмме относительных частот, но вместо значений плотностей относительных частот на оси ординат откладываются значения накопленных частот. То есть мы строим прямоугольники, ширина которых равна $h = 5$, а высоты прямоугольников равны накопленным частотам. Так как накопленные частоты возрастают с увеличением индекса, то и фигура должна получиться возрастающей.

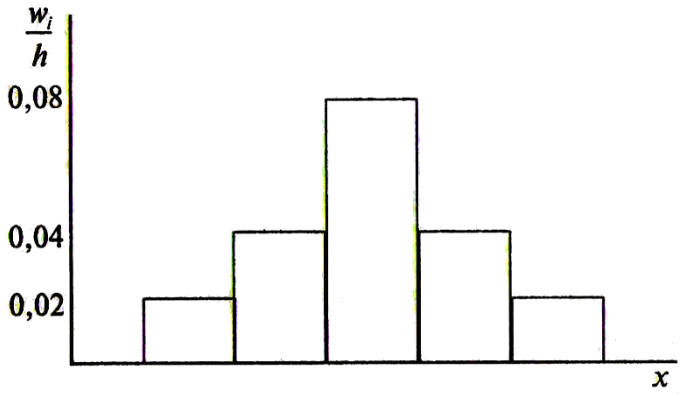


Рис. 1.6

6. Построить эмпирическую функцию по данному распределению выборки:

x_i	2	3	5
n_i	75	20	5
w_i	0,75	0,2	0,05

Решение. Найдем объем выборки n: $75+20+5=100$.
Наименьшая варианта равна 2, следовательно,

$$F_n(x)=0 \text{ при } x\le 2.$$

Значение $X<3$, а именно $x_1=2$ имеет частоту 75, т.е. это значение наблюдалось 75 раз. Следовательно,

$$F_n(x)=\frac{75}{100}=0,75 \text{ при } 2<x\le 3.$$

Значения $X<5$, а именно $x_1=2$ и $x_2=3$ наблюдались $75+20=95$ раз. Следовательно,

$$F_n(x)=\frac{95}{100}=0,95 \text{ при } 3<x\le 5.$$

Так как $x=5-$ наибольшая варианта, то

$$F_n(x)=1 \text{ при } x>5.$$

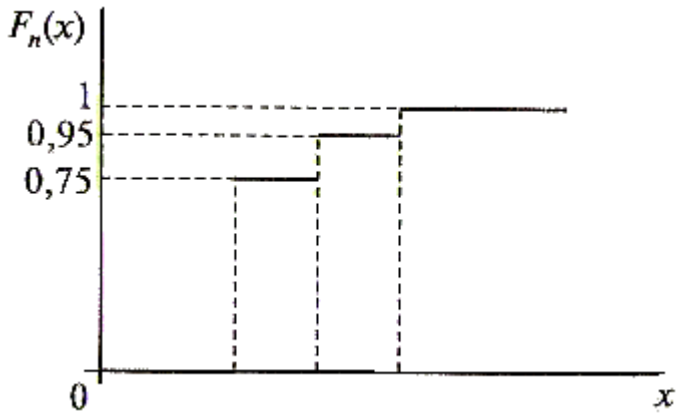


Рис. 1.7

Эмпирическая функция распределения имеет вид (рис. 1.7):

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

7. Анализируется выборка из 100 малых предприятий региона. Цель обследования – измерение коэффициента соотношения заемных и собственных средств (x_i) на каждом i-ом предприятии. Результаты анализа представлены ниже. Требуется построить гистограмму и график накопленных частот.

Коэффициенты соотношений собственных и заемных средств предприятий

5,56;	5,45;	5,48;	5,45;	5,39;	5,37;	5,46;	5,59;	5,61;	5,31;
5,46;	5,61;	5,11;	5,41;	5,31;	5,57;	5,33;	5,11;	5,54;	5,43;
5,34;	5,53;	5,46;	5,41;	5,48;	5,39;	5,11;	5,42;	5,48;	5,49;
5,36;	5,40;	5,45;	5,49;	5,68;	5,51;	5,50;	5,68;	5,21;	5,38;
5,58;	5,47;	5,46;	5,19;	5,60;	5,63;	5,48;	5,27;	5,22;	5,37;
5,33;	5,49;	5,50;	5,54;	5,40;	5,58;	5,42;	5,29;	5,05;	5,79;
5,79;	5,65;	5,70;	5,71;	5,85;	5,44;	5,47;	5,48;	5,47;	5,55;
5,67;	5,71;	5,73;	5,05;	5,35;	5,72;	5,49;	5,61;	5,57;	5,69;
5,54;	5,39;	5,32;	5,21;	5,73;	5,59;	5,38;	5,25;	5,26;	5,81;
5,27;	5,64;	5,20;	5,23;	5,33;	5,37;	5,24;	5,55;	5,60;	5,51.

Решение. 1) Определим по выборке $x_{\min}=5,05$ и $x_{\max}=5,85$.

2) Разобьем весь диапазон $[x_{\min}, x_{\max}]$ на k интервалов одинаковой длины:

$$k\approx 1+\log_2 100=7,62;\ k=8.$$

Отсюда получаем, что длина интервала

$$h=\frac{x_{\max}-x_{\min}}{k}=\frac{5,85-5,05}{8}=0,1.$$

Построим сгруппированный ряд наблюдений (табл. 1.1).

Таблица 1.1

Сгруппированный ряд наблюдений

Номер интервала	Интервал	Середина интервала x_i	w_i	w_i^c	$f_n(x)$
1	5,05 – 5,15	5,1	0,05	0,05	0,5
2	5,15 – 5,25	5,2	0,08	0,13	0,8
3	5,25 – 5,35	5,3	0,12	0,25	1,2
4	5,35 – 5,45	5,4	0,20	0,45	2,0
5	5,45 – 5,55	5,5	0,26	0,71	2,6
6	5,55 – 5,65	5,6	0,15	0,86	1,5
7	5,65 – 5,75	5,7	0,10	0,96	1,0
8	5,75 – 5,85	5,8	0,04	1,00	0,4

Гистограмма и график накопленных частот представлены на рис. 1.8 и 1.9.

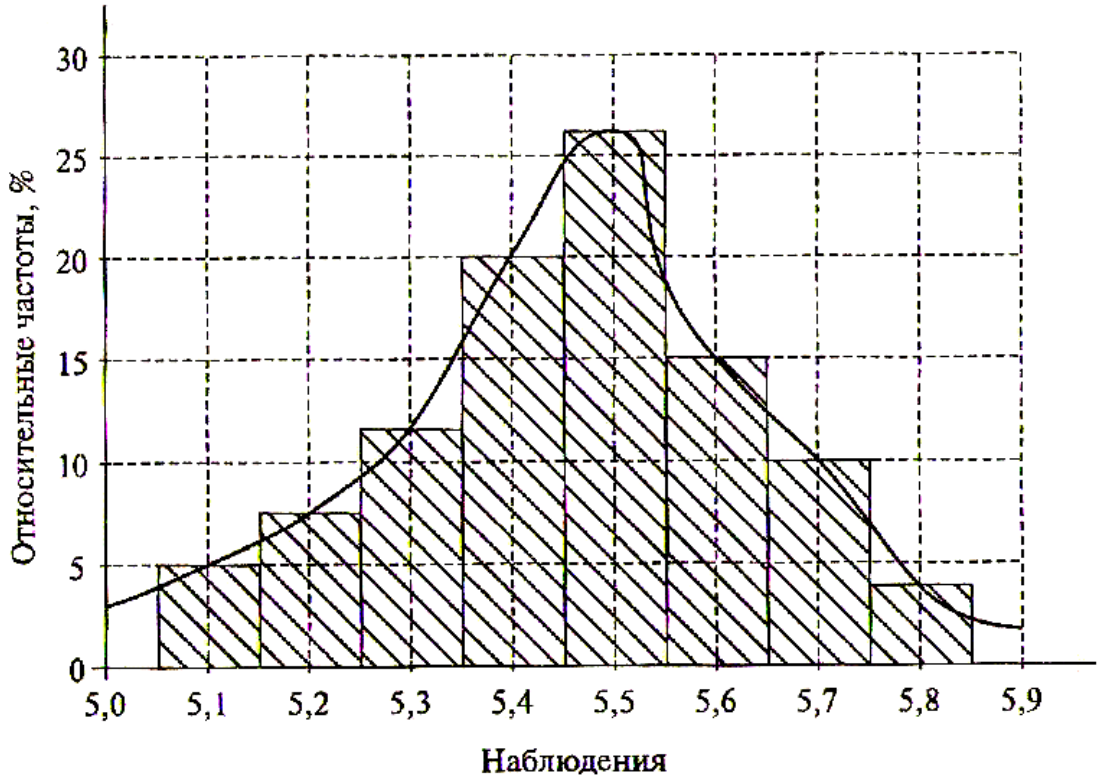


Рис. 1.8

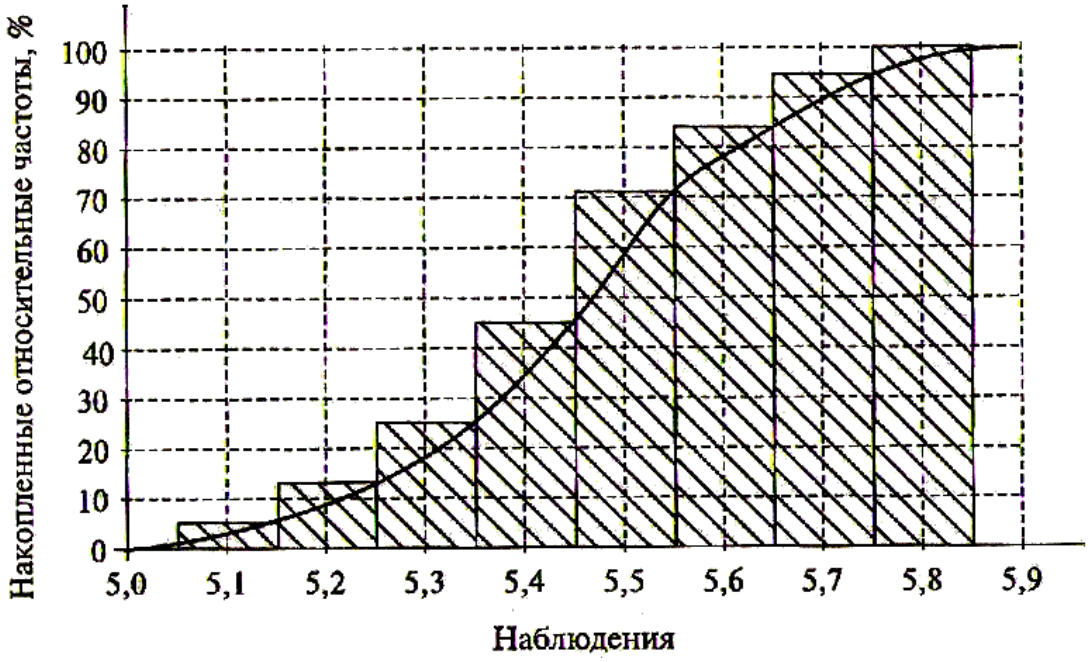


Рис. 1.9

8. Путем опроса получены следующие данные $(n=80)$:

2 4 2 4 3 3 3 2 0 6 1 2 3 2 2 4 3 3 5 1 0 2 4 3 2 2 3 3 1 3
3 3 1 1 2 3 1 4 3 1 7 4 3 4 2 3 2 3 3 1 4 3 1 4 5 3 4 2 4 5
3 6 4 1 3 2 4 1 3 1 0 0 4 6 4 7 4 1 3 5

а) Составить статистическое распределение выборки, предварительно записав дискретный вариационный ряд.

б) Построить полигон частот.

в) Составить ряд распределения относительных частот.

г) Составить эмпирическую функцию распределения.

д) Построить график эмпирической функции распределения.

Решение. а) Для составления дискретного вариационного ряда отсортируем данные опроса по величине и расположим их в порядке возрастания:

0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 6 6 6 7 7.

Примечание. Указанную процедуру, как и большинство расчетов по математической статистике, удобно выполнять, используя электронные таблицы, например, Microsoft Excel.

Более компактно эти данные можно представить в виде статистического распределения выборки (в виде таблицы, в которой первая

строка – варианты (наблюдаемые значения), вторая строка – частоты появления этих вариант):

x_i	0	1	2	3	4	5	6	7
n_i	4	13	14	24	16	4	3	2

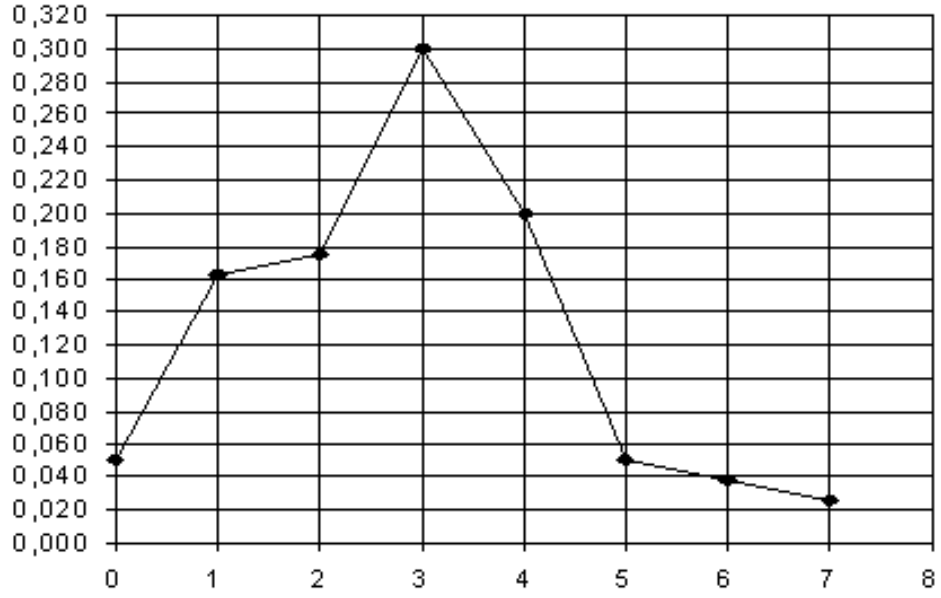
б) Для построения полигона частот найдем относительные частоты. Расчеты запишем в таблицу:

x_i	n_i	Относительные частоты $w_i = \frac{n_i}{n}$	Накопленные частоты w_i^c
0	4	0,050	0,050
1	13	0,163	0,213
2	14	0,175	0,388
3	24	0,300	0,688
4	16	0,200	0,888
5	4	0,050	0,938
6	3	0,038	0,975
7	2	0,025	1,000
	80	1	

Изобразим полигон частот вариационного ряда (рис. 1.10).

в) Запишем ряд распределения относительных частот в виде таблицы, в которой первая строка – варианты (изучаемый признак), вторая строка – относительные частоты.

x_i	0	1	2	3	4	5	6	7
w_i	0,05	0,163	0,175	0,3	0,2	0,05	0,038	0,025



Полигон частот вариационного ряда
Рис. 1.10

г) Эмпирическую функцию распределения найдем, используя накопленные частоты w_i^c, полученные в пункте б), и следующую формулу:

$$F^*(x)=\begin{cases}0, & x \le x_1,\\ \dfrac{n_1}{n}, & x_1 < x \le x_2,\\ \dfrac{n_1+n_2}{n}, & x_2 < x \le x_3,\\ \ldots\\ \sum\limits_{i=1}^{m-1}\dfrac{n_i}{n}, & x_{m-1} < x \le x_m,\\ 1, & x > x_m.\end{cases}$$

Таким образом, эмпирическая функция распределения примет вид

$$F^*(x)=\begin{cases}0;\ x\le 0,\\ 0,05;\ 0<x\le 1,\\ 0,213;\ 1<x\le 2,\\ 0,388;\ 2<x\le 3,\\ 0,688;\ 3<x\le 4,\\ 0,888;\ 4<x\le 5,\\ 0,938;\ 5<x\le 6,\\ 0,975;\ 6<x\le 7,\\ 1,\ x>7.\end{cases}$$

д) Построим график эмпирической функции распределения (рис. 1.11), используя значения, полученные в пункте г).

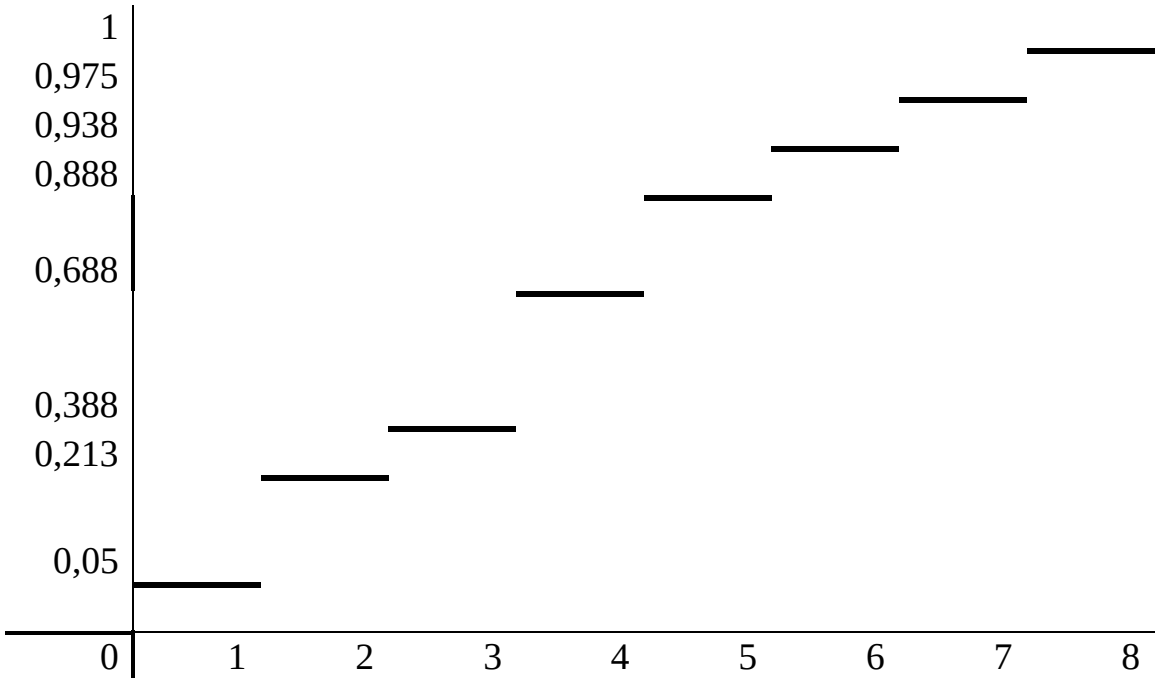


Рис. 1.11

Задачи для самостоятельного решения

1. Данные выборки приведены в таблице. Построить полигон частот.
а)

x_i	1	2	4	5	6
n_i	5	2	8	3	2

б)

x_i	-1	0	1	2	4
n_i	5	4	3	6	2

в)

x_i	0	1	2	3	4
n_i	3	6	4	5	2

2. По данным выборки построить полигон относительных частот.

а)

x_i	–2	–1	0	1	5
n_i	4	12	2	6	4

б)

x_i	–1	0	1	2	4
n_i	15	5	10	5	15

в)

x_i	–5	–2	0	3	4
n_i	1	8	4	5	2

3. По данным выборки составить дискретное статистическое распределение. Построить полигон частот.

4,1; 4,2; 3,8; 5; 5,1; 4,2; 3; 4,1; 3,8; 5,5; 3,6; 4,2; 4; 5,8; 6; 3; 5,1; 5,5; 3,6; 4.

4. По данным выборки составить статистическое распределение. Построить гистограмму частот, график накопленных частот.

1; 15; 13; 11; 9; 6; 6; 3,8; 9,1; 9,8;
2,7; 4,9; 3,5; 6,1; 9,9; 10,3; 2,8; 6,4; 2,9; 9,6;
13,5; 3,8; 8,6; 8,5; 7,6; 9,4; 4,8; 5,4; 2,3; 9;
10; 12; 11,8; 4,1; 5,3; 7,6; 3,8; 9,4; 4,5; 6;
6; 5,9; 9,1; 3,7; 11,7; 10,3; 4,2; 7,1; 6; 8.

5. Построить гистограмму частот и график накопленных частот по данному распределению выборки:

№ интервала	Интервал	Частоты
1	2 – 7	5
2	7 – 12	10
3	12 – 17	25
4	17 – 22	6
5	22 – 27	4

6. Построить гистограмму относительных частот по данному распределению выборки:

№ интервала	Интервал	Частоты
1	10 – 15	2
2	15 – 20	4
3	20 – 25	8
4	25 – 30	4
5	30 – 35	2

***Ответы*. 3)**

x_i	3	3,6	3,8	4	4,1	4,2	5	5,1	5,5	5,8	6
n_i	2	2	2	2	2	3	1	2	2	1	1

4) $[1,15]$; $n=50$; $k=7$; $h=2$.

№ интервала	Интервал	Частоты
1	1 – 3	5
2	3 – 5	10
3	5 – 7	10

4	7 – 9	6
5	9 – 11	12
6	11 – 13	4
7	13 – 15	3

Контрольные вопросы

1. Что такое генеральная и выборочная совокупности?
2. Какая выборка называется репрезентативной?
3. Какие существуют способы отбора?
4. Дайте определение вариационного ряда и статистического распределения выборки.
5. По какому правилу задается эмпирическая функция распределения?
6. Чему равна относительная частота варианты?
7. Как строятся полигон и гистограмма частот?
8. В чем отличие гистограммы частот от гистограммы относительных частот?

Глава 2

ТОЧЕЧНЫЕ И ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

Статистические оценки параметров распределения

Пусть требуется изучить некоторый количественный признак генеральной совокупности. Допустим, что из теоретических соображений установлено, какое именно распределение имеет признак. Возникает задача оценки параметров, которыми определяется данное распределение. Например, если известно, что изучаемый признак генеральной совокупности распределен нормально, то необходимо оценить, т.е. найти приближенно, математическое ожидание и среднее квадратическое отклонение, так как эти два параметра полностью характеризуют нормальное распределение. Если признак имеет распределение Пуассона, то необходимо оценить параметр λ, которым это распределение определяется.

Обычно имеются лишь данные выборки, например значения количественного признака $x_1, x_2, \ldots, x_n$, полученные в результате n наблюдений (мы предполагаем все наблюдения независимыми). Через эти данные и выражают оцениваемый параметр.

Рассматривая все значения генеральной совокупности как значения некоторой случайной величины X, можно сказать, что наблюдаемые значения $x_1, x_2, \ldots, x_n$ являются значениями системы независимых случайных величин $X_1, X_2, \ldots, X_n$. При этом случайные величины X_1, X_2,..., X_n имеют тот же самый закон распределения, что и X, так как все элементы x_i выбираются из генеральной совокупности. Поэтому можно сказать, что найти статистическую оценку неизвестного параметра – это значит найти функцию от наблюдаемых случайных величин, которая и дает приближенное значение оцениваемого параметра.

Оценкой числовой характеристики или параметра θ называется функция от наблюдаемых значений $\hat{\theta}_n = \hat{\theta}(x_1, x_2, \ldots, x_n)$, которая в определенном смысле «близка» к истинному значению θ. Оценки числовых характеристик, таких как математическое ожидание, дисперсия, начальные и центральные моменты, называются *выборочными характеристиками*.

Качество оценки устанавливают, проверяя, выполняются ли следующие свойства:

1) состоятельность;
2) несмещенность;
3) эффективность.

Оценка $\hat{\theta}$ называется *состоятельной*, если она сходится по вероятности к истинному значению (индекс n обычно опускается, но подразумевается по умолчанию), т.е.

$$\hat{\theta} \xrightarrow{p} \theta \text{ при } n \to \infty.$$

Оценка $\hat{\theta}$ называется *несмещенной*, если ее математическое ожидание равно истинному значению:

$$M\left(\hat{\theta}\right) = \theta.$$

Это свойство желательно, но не обязательно. Часто полученная оценка бывает смещенной, но ее можно поправить так, чтобы она стала несмещенной. Иногда оценка бывает смещенной, но *асимптотически несмещенной*, т.е. математическое ожидание оценки стремится к истинному значению:

$$M\left(\hat{\theta}\right) \to \theta \text{ при } n \to \infty.$$

Оценка θ^* называется *эффективной* в определенном классе оценок $\hat{\Theta}$, если она самая точная среди оценок этого класса, т.е. имеет минимальную дисперсию:

$$D\left(\theta^*\right) = \min_{\hat{\theta} \in \hat{\Theta}} D\left(\hat{\theta}\right).$$

Выборочные числовые характеристики

Приведем примеры основных числовых характеристик – моментов.

1) *Выборочное среднее*

$$\overline{x} = \hat{a} = \hat{\nu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i . \qquad (2.1)$$

Здесь суммируются все наблюдаемые значения x_i. Если же все значения x_i сгруппированы следующим образом: варианта x_1 имеет частоту n_1, варианта x_2 – частоту n_2, …, варианта x_k – частоту n_k, тогда выборочное среднее можно вычислить по формуле

$$\overline{x} = \hat{a} = \hat{\nu}_1 = \frac{\sum_{i=1}^{k} n_i \cdot x_i}{n}. \qquad (2.1a)$$

Согласно закону больших чисел (теорема Чебышева), среднее арифметическое независимых одинаково распределенных случайных величин, имеющих дисперсию σ^2, сходится по вероятности к математическому ожиданию. Это означает, что выборочное среднее есть состоятельная оценка математического ожидания генеральной совокупности.

Несмещенность $\bar{x}$ доказывается прямой проверкой:

$$M(\bar{x}) = M\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} M(x_i) = \frac{1}{n}\sum_{i=1}^{n} a = \frac{na}{n} = a.$$

Теперь докажем, что $\bar{x}$ эффективна в классе всех линейных несмещенных оценок. Произвольная линейная оценка имеет вид

$$\tilde{x} = \sum_{i=1}^{n} c_i x_i ,$$

линейная несмещенная оценка должна удовлетворять условию $M\tilde{x} = a$, поэтому

$$M(\tilde{x}) = M\left(\sum_{i=1}^{n} c_i x_i\right) = \sum_{i=1}^{n} c_i M(x_i) = \sum_{i=1}^{n} c_i a = a\sum_{i=1}^{n} c_i ,$$

т.е., для несмещенности $\tilde{x}$ необходимо, чтобы $\sum_{i=1}^{n} c_i = 1.$

Найдем дисперсию $\tilde{x}$

$$D(\tilde{x}) = D\left(\sum_{i=1}^{n} c_i x_i\right) = \sum_{i=1}^{n} c_i^2 D(x_i) = \sum_{i=1}^{n} c_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} c_i^2 .$$

Таким образом, для определения наиболее эффективной оценки в классе линейных несмещенных оценок надо исследовать дисперсию на минимум при условии несмещенности линейной оценки, т.е.

$$D(\tilde{x}) \to \min \text{ или (что то же самое) } \sum_{i=1}^{n} c_i^2 \to \min$$

при условии

$$\sum_{i=1}^{n} c_i = 1.$$

То есть, мы получили задачу на условный экстремум. Данная задача сводится к задаче на безусловный экстремум с помощью функции Лагранжа. Итак, составляем функцию Лагранжа:

$$L(c_1,...,c_n,\lambda) = \sum_{i=1}^{n} c_i^2 - \lambda\left(\sum_{i=1}^{n} c_i - 1\right).$$

Для отыскания точки экстремума находим частные производные и приравниваем их к нулю:

$$\frac{\partial L}{\partial c_i} = 2c_i - \lambda = 0,\ i = 1,...,n.$$

$$\frac{\partial L}{\partial \lambda} = \sum_{i=1}^{n} c_i - 1 = 0.$$

Подставляя значения $c_i = \lambda/2$, полученные из первых уравнений, в последнее уравнение, получаем

$$\frac{\lambda}{2} = \frac{1}{n} \text{ или } c_i = \frac{\lambda}{2} = \frac{1}{n}.$$

Итак, при $c_i = 1/n,\ i = 1,\ldots,n$, оценка $\tilde{x}$ имеет минимальную дисперсию, т.е. $\bar{x}$ является эффективной в классе всех линейных несмещенных оценок.

2) *Выборочная дисперсия*

$$\bar{D} = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2. \qquad (2.2)$$

Если варианта x_1 имеет частоту n_1, варианта x_k – частоту n_k, тогда выборочную дисперсию можно вычислить по формуле

$$\bar{D} = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{n}. \qquad (2.2a)$$

Теорема 2.1. Выборочную дисперсию можно вычислить по следующей формуле*:*

$$\bar{D} = \overline{x^2} - \left(\bar{x}\right)^2, \tag{2.3}$$

где $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

Доказательство

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}\left[x_i^2 - 2x_i\bar{x} + \left(\bar{x}\right)^2 \right] = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \left(\bar{x}\right)^2 =$$

$$= \overline{x^2} - 2\bar{x}\bar{x} + \left(\bar{x}\right)^2 = \overline{x^2} - 2\left(\bar{x}\right)^2 + \left(\bar{x}\right)^2 = \overline{x^2} - \left(\bar{x}\right)^2.$$

Итак, $\bar{D} = \overline{x^2} - \left(\bar{x}\right)^2$.

Покажем, что выборочная дисперсия является смещенной оценкой дисперсии σ^2 генеральной совокупности, и найдем несмещенную оценку.

Теорема 2.2. Для любого b справедливо равенство

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - b)^2 - (\bar{x} - b)^2 .$$

Эту теорему мы примем без доказательства.

Следствие. Если $M\left(x\right) = a$ есть среднее генеральной совокупности, то

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 - (\bar{x} - a)^2 .$$

Вычислим математическое ожидание выборочной дисперсии:

$$M\overline{D} = M\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - a\right)^2\right] - M\left(\bar{x} - a\right)^2 = \frac{1}{n}\sum_{i=1}^{n} M\left(x_i - a\right)^2 - D\bar{x} =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 - D\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{n\sigma^2}{n} - \frac{1}{n^2}\sum_{i=1}^{n} Dx_i = \frac{n\sigma^2}{n} - \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 =$$

$$= \frac{n\sigma^2}{n} - \frac{n\sigma^2}{n^2} = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Отсюда следует, что $\overline{D}$ является смещенной оценкой дисперсии. Эта оценка имеет систематическое смещение $\left(-\sigma^2/n\right)$, т.е. эта оценка занижает в среднем истинное значение дисперсии на величину σ^2/n. Правда, это смещение сходит на нет при $n \to \infty$, т.е. оценка асимптотически не смещена.

Чтобы устранить смещение, скорректируем оценку. Для этого введем величину

$$S^2 = \frac{n}{n-1}\overline{D} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2. \qquad (2.4)$$

Поскольку

$$MS^2 = M\left(\frac{n}{n-1}\bar{D}\right) = \frac{n}{n-1}M\bar{D} = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2,$$

статистика S^2 будет несмещенной оценкой дисперсии генеральной совокупности. Поэтому оценку $\overline{D}$ называют *смещенной*, или *неисправленной*, выборочной дисперсией, а оценку S^2 называют *несмещенной*, или *исправленной*, выборочной дисперсией. Выборочным средним квадратическим отклонением называют положительный корень из соответствующей дисперсии:

$\hat{\sigma} = \sqrt{\overline{D}}$ – *неисправленное* выборочное среднее квадратическое отклонение;

$s = \sqrt{S^2}$ – *исправленное* выборочное среднее квадратическое отклонение.

Эта оценка теоретического среднего квадратического отклонения $\sigma = \sqrt{D(X)}$ в обоих случаях является смещенной.

3) *Выборочные начальные моменты k-го порядка*

$$\overline{x^k} = \hat{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k. \qquad (2.5)$$

Аналогично выборочному среднему и выборочной дисперсии начальные моменты можно вычислить по двум формулам.

$$\overline{x^k} = \hat{\nu}_k = \frac{1}{n}\sum_{i=1}^{l} n_i \cdot x_i^k , \qquad (2.5а)$$

где $l -$ количество различных вариант x_i.

Выше было сказано, что оценка $\bar{x} = \hat{\nu}_1$ первого начального момента является состоятельной, несмещенной и эффективной (в классе линейных несмещенных оценок), если генеральная случайная величина имеет дисперсию σ^2.

Начальный момент и его оценка как функция от случайной выборки имеют вид

$$\nu_k = Mx^k, \ \hat{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k = \overline{x^k}.$$

Если ввести случайную величину $u = x^k$, то для нее рассматриваемый момент будет моментом первого порядка:

$$\nu_k = M\left(x^k\right) = M\left(u\right),$$

поэтому оценка $\hat{\nu}_k = \overline{x^k}$ как оценка математического ожидания $u = x^k$ обладает следующими свойствами: состоятельна, не смещена, эффективна в классе линейных относительно u_i (или x_i^k) оценок, если существует дисперсия

$$D(u) = M\left(u^2\right) - \left(Mu\right)^2 = \nu_{2k} - \nu_k^2,$$

т.е., если у генеральной случайной величины наряду с моментом ν_k, который оценивается, существует и момент ν_{2k}.

4) *Выборочные центральные моменты k-го порядка*

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k , \qquad (2.6)$$

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{l} n_i\left(x_i - \bar{x}\right)^k , \qquad (2.6а)$$

где $l -$ количество различных вариант x_i.

Отметим, что все выборочные центральные моменты $\hat{\mu}_k$ сходятся по вероятности, если существуют теоретические центральные моменты $\mu_k = M\left(x-a\right)^k$.

В качестве других используемых на практике выборочных характеристик можно назвать *выборочную моду* x_{mod}, равную значению варианты с наибольшей частотой, и *выборочную медиану* x_{med}, равную значению, стоящему в середине вариационного ряда (либо полусумме двух значений, стоящих рядом в середине, при четном числе наблюдений). *Коэффициентом вариации* вариационного ряда называется процентное отношение среднеквадратического отклонения к средней арифметической:

$$\hat{V} = \frac{\hat{\sigma}}{\bar{x}} 100\%,$$

коэффициентом асимметрии – число

$$\hat{\beta} = \frac{\hat{\mu}_3}{\hat{\sigma}^3},$$

и *эксцессом* вариационного ряда называется число

$$\hat{\nu} = \frac{\hat{\mu}_4}{\hat{\sigma}^4}.$$

Примеры

1. Найти несмещенные оценки математического ожидания и дисперсии, начальные моменты второго и третьего порядков и центральные моменты первого и второго порядков по выборке объема $n = 20$:

x_i	–1	1	2	3	5
n_i	2	3	10	4	1

Решение. Несмещенная оценка математического ожидания вычисляется по формуле (2.1а):

$$\bar{x} = \frac{2(-1) + 3\cdot 1 + 10\cdot 2 + 4\cdot 3 + 1\cdot 5}{20} = 1{,}9.$$

Вычислим начальные моменты второго и третьего порядков по формуле (2.5а). Берем $k = 2$ и $k = 3$ соответственно:

$$\overline{x^2} = \frac{2(-1)^2 + 3\cdot 1^2 + 10\cdot 2^2 + 4\cdot 3^2 + 1\cdot 5^2}{20} = \frac{2+3+40+36+25}{20} = 5,3;$$

$$\overline{x^3} = \frac{2(-1)^3 + 3\cdot 1^3 + 10\cdot 2^3 + 4\cdot 3^3 + 1\cdot 5^3}{20} = \frac{-2+3+80+108+125}{20} = 15,7.$$

Теперь вычислим центральный момент первого порядка по формуле (2.6а), в формуле надо взять $k=1$:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^1 = \frac{1}{n}\sum_{l=1}^{k} n_i (x_i - \bar{x}) = \frac{\sum_{i=1}^{k} n_i x_i}{n} - \bar{x} = \bar{x} - \bar{x} = 0,$$

таким образом, *центральный момент первого порядка всегда равен нулю*.

Центральный момент второго порядка – это неисправленная выборочная дисперсия, ее удобнее вычислить по формуле (2.3):

$$\hat{\mu}_2 = \overline{D} = \overline{x^2} - \left(\bar{x}\right)^2 = 5,3 - (1,9)^2 = 5,3 - 3,61 = 1,69.$$

Следует обратить внимание на то, что *выборочная дисперсия не может быть отрицательным числом*.

Осталось вычислить неисправленную выборочную дисперсию S^2, для этого используем формулу (2.4):

$$S^2 = \frac{20}{20-1}\overline{D} = \frac{20}{19}\cdot 1,69 = 1,78.$$

Ответ. Выборочное среднее $\bar{x} = 1,9$; начальный момент второго порядка $\hat{\nu}_2 = \overline{x^2} = 5,3$; начальный момент третьего порядка $\hat{\nu}_3 = \overline{x^3} = 15,7$; центральный момент первого порядка $\hat{\mu}_1 = 0$; центральный момент второго порядка $\hat{\mu}_2 = 1,69$ и несмещенная выборочная дисперсия $S^2 = 1,78$.

2. Найти выборочное среднее по данному распределению выборки объема $n=100$:

x_i	2702	2804	2903	3028
n_i	8	30	60	2

Решение. Для того чтобы упростить вычисления, перейдем к условным вариантам

$$u_i = x_i - 2844.$$

Получим следующее распределение:

u_i	–142	–40	59	184
n_i	8	30	60	2

Вычислим выборочное среднее условной величины по формуле (2.1а):

$$\bar{u} = \frac{8(-142) + 30(-40) + 60 \cdot 59 + 2 \cdot 184}{100} = \frac{-1136 - 1200 + 3540 + 368}{100} = 15,72.$$

Так как $u = x - 2844$, то $\bar{u} = \bar{x} - 2844$, следовательно,

$$\bar{x} = \bar{u} + 2844 = 15,72 + 2844 = 2859,72.$$

Отметим, что $\overline{D}(u) = \overline{D}(x - 2844) = \overline{D}(x)$. То есть при переходе к условным вариантам путем сдвига на некоторую константу выборочная дисперсия не меняется.

3. Найти выборочное среднее и выборочную дисперсию по данному распределению выборки объема $n = 10$:

x_i	0,01	0,04	0,08
n_i	5	3	2

Решение. Перейдем к условным вариантам

$$u_i = 100 \cdot x_i.$$

Получим следующее распределение:

u_i	1	4	8
n_i	5	3	2

Вычислим сначала выборочное среднее $\bar{u}$ и $\overline{u^2}$:

$$\bar{u} = \frac{5 + 12 + 16}{10} = 3,3,$$

$$\overline{u^2} = \frac{5 + 3 \cdot 16 + 2 \cdot 64}{10} = 18{,}1.$$

Выборочная дисперсия условной величины равна

$$\overline{D}(u) = \overline{u^2} - \left(\overline{u}\right)^2 = 18{,}1 - (3{,}3)^2 = 7{,}21.$$

Так как $u = 100 \cdot x$, то $\overline{u} = 100 \cdot \overline{x}$. Итак, выборочное среднее исходной варианты

$$\overline{x} = \frac{\overline{u}}{100} = \frac{3{,}3}{100} = 0{,}033.$$

Аналогично, используя свойства дисперсии, получаем

$$\overline{D}(u) = 100^2 \cdot \overline{D}(x).$$

Следовательно,

$$\overline{D}(x) = \frac{\overline{D}(u)}{100^2} = \frac{7{,}21}{10000} = 0{,}000721.$$

Основные распределения случайных величин, используемые в математической статистике

Квантили и процентные точки распределения

При использовании различных методов математической статистики, построении статистических критериев, интервальных оценок неизвестных параметров широко используются понятия квантилей, процентных точек односторонних и двусторонних критических границ распределений.

Квантилью уровня p, или p*-квантилью* непрерывной случайной величины X с функцией распределения $F_X(x)$, называется такое возможное значение x_p этой случайной величины, для которого вероятность события $X < x_p$ равна заданной величине p, т.е. $P\left(X < x_p\right) = p,\ 0 < p < 1.$

Из определения следует, что x_p есть решение, по предположению единственное, уравнения $F_X\left(x_p\right) = p,\ 0 < p < 1$. Геометрически x_p есть такое значение случайной величины X, при котором площадь криволинейной трапеции, ограниченная графиком плотности распределения и осью абсцисс и лежащая левее x_p, равна p.

На рис. 2.1 показана квантиль уровня 0,8 (20 %-ная точка) для стандартного нормального распределения (на графиках плотности и функции распределения).

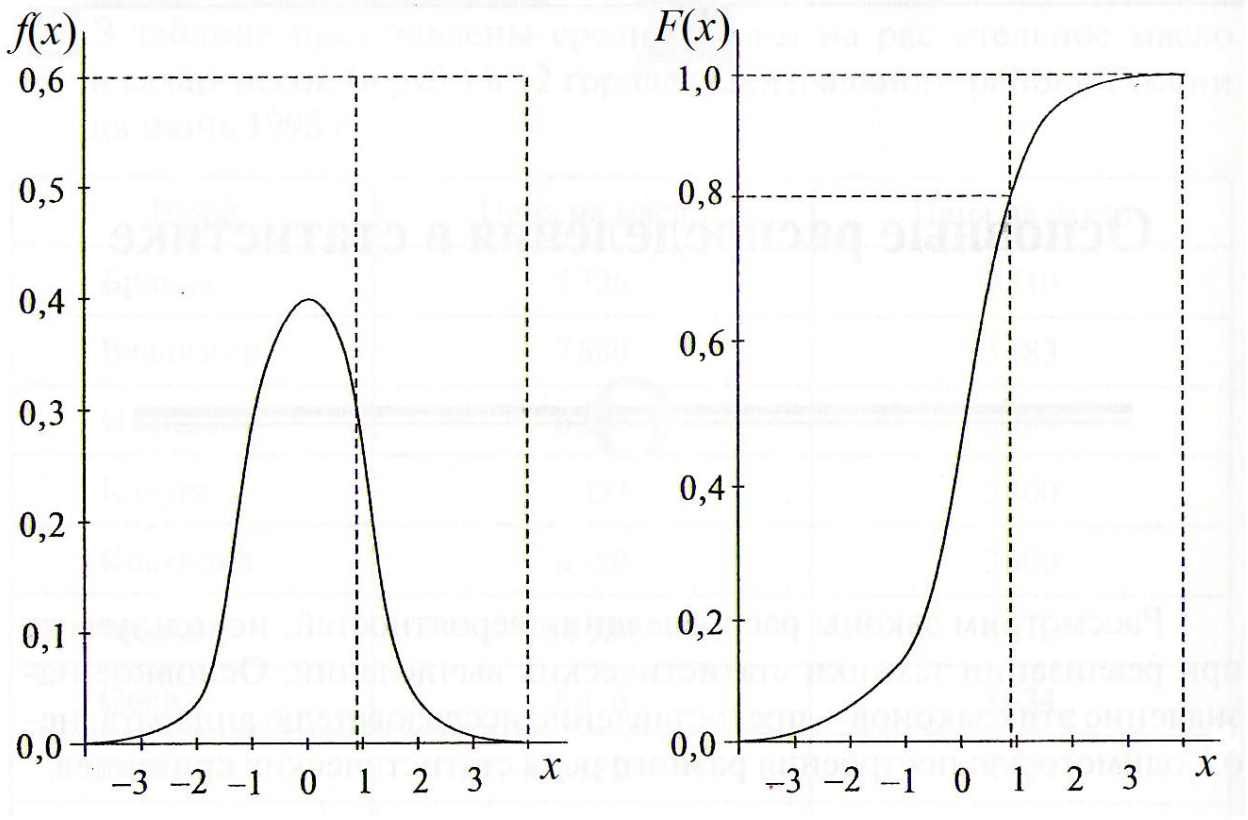


Рис. 2.1

Процентной точкой уровня q, или q*%-ной точкой* (при $0 \le q \le 100$) для непрерывной случайной величины X с функцией распределения $F_X(x)$, называется такое значение x_q этой случайной величины, что вероятность события $X \ge x_q$ равна $q/100$, т.е. $1 - F_X(x_q) = P(X \ge x_q) = q/100$.

Геометрически q%-ная точка – это значение случайной величины, при котором площадь криволинейной трапеции, ограниченная графиком плотности распределения, осью абсцисс и лежащая правее x_q, равна $q/100$ (см. рис. 2.2). Нижней критической границей $\underline{u}_\alpha$ и верхней критической границей $\overline{u}_\alpha$, соответствующей заданному уровню значимости α, называются значения случайной величины, для которых выполнены условия

$$P(X \le \underline{u}_\alpha) = F_X(\underline{u}_\alpha) = \frac{\alpha}{2};$$

$$P(\underline{u}_\alpha \le X < \overline{u}_\alpha) = F_X(\overline{u}_\alpha) - F_X(\underline{u}_\alpha) = 1 - \alpha;$$

$$P\left(X \geq \bar{u}_{\alpha}\right) = 1 - F_X\left(\bar{u}_{\alpha}\right) = \frac{\alpha}{2}.$$

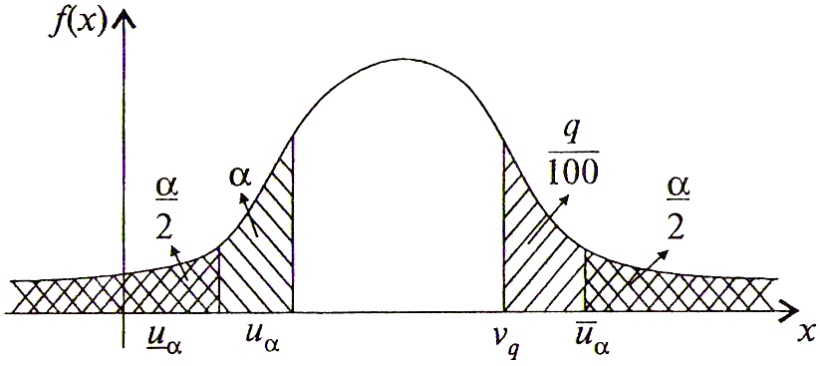


Рис. 2.2

Критические точки для заданного распределения определяют границы, за пределы которых случайная величина выходит достаточно редко. Например, для распределения Стьюдента с n степенями свободы критической точкой с уровнем значимости α (для двусторонней области) называется величина $t_{\text{кр}} = t_{\text{кр}}\left(\alpha; n\right)$, такая что для случайной величины X с данным распределением выполнено условие $P\left(\left|X\right| > t_{\text{кр}}\right) = \alpha$, где α обычно выбирается достаточно малым (близким к нулю), например, $\alpha = 0{,}05$. Более подробно это понятие будет рассмотрено далее.

Распределение Стьюдента

Пусть генеральная совокупность X имеет нормальное распределение $N_{a;\sigma}$, т.е. математическое ожидание случайной величины равно a, дисперсия – σ^2. Рассмотрим случайную выборку объема n из генеральной совокупности. Выборочное среднее

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

при больших n также имеет нормальное распределение с параметрами a и σ^2/n, т.е. $\bar{X} \sim N_{a;\sigma/\sqrt{n}}$. Центрируем и нормируем выборочное среднее, получим случайную величину

$$Z = \frac{\bar{X} - a}{\sigma/\sqrt{n}}, \tag{2.7}$$

которая будет распределена по стандартному нормальному закону $N_{0,1}$.

В большинстве случаев значение дисперсии генеральной совокупности неизвестно. Что изменится в формуле, если в знаменателе дроби (2.7) среднее квадратическое отклонение σ заменить на выборочное исправленное среднее квадратическое отклонение s?

Английский статистик В. Госсет решил эту задачу, имеющую исключительно важное значение для статистического анализа. Госсет работал исследователем на пивоваренном заводе Гиннеса в Дублине (Ирландия). Дирекция завода, где трудился Госсет, не позволяла своим работникам публиковать научные результаты. Поэтому Госсет опубликовал свое открытие под именем Стьюдент. В 1908 г. он открыл закон распределения случайной величины

$$t = \frac{\bar{X} - a}{s/\sqrt{n}} = \frac{\bar{X} - a}{s}\sqrt{n} \qquad (2.8)$$

и назвал его t-распределением. В его честь это распределение теперь называют *t-распределением Стьюдента* или *просто распределением Стьюдента*.

Сравним выражения (2.7) и (2.8). Из (2.7) ясно, что значения Z изменяются вследствие того, что каждая выборка имеет различные выборочные средние. В формуле (2.8) t имеет два источника вариации: $\bar{X}$ и s, которые меняются от выборки к выборке. Поэтому мы не можем утверждать, что распределение значений случайной величины t подчиняется нормальному закону распределения.

Распределение Стьюдента относится к семейству распределений, которые зависят от параметра, называемого числом степеней свободы (обозначается буквами k, ν или df – degree of freedom). Для значений t в выражении (2.8) число степеней свободы k равно $n-1$, где $n-$объем выборки. Оценка вариации зависит не только от объема выборки, но и от того, как много параметров должно оцениваться в выборке. Чем больше данных, тем больше мы можем доверять полученным результатам; чем больше параметров мы должны оценить, тем меньше мы им доверяем. Эти два момента в статистике учитываются при вычислении числа степеней свободы:

$$\begin{pmatrix}\text{число степеней}\\ \text{свободы}\end{pmatrix} = \begin{pmatrix}\text{число}\\ \text{наблюдений}\end{pmatrix} - \begin{pmatrix}\text{число параметров, которые}\\ \text{должны быть оценены}\end{pmatrix}$$

Смысл числа степеней свободы можно проиллюстрировать на следующем примере.

Пример. Менеджер компании имеет бюджет 150000 долл. на четыре различных проекта. Сколькими степенями свободы располагает менеджер?

Решение. Общий бюджет четырех проектов можно рассматривать как их среднюю арифметическую, умноженную на число проектов:

$$x_1 + x_2 + x_3 + x_4 = n \cdot \bar{X} .$$

Менеджер имеет три возможности (в пределах общего бюджета) любого распределения сумм на любые три из четырех проектов. Как только средства на первые три проекта распределены, у менеджера не остается выбора при распределении средств на четвертый проект. Он может выделить на этот проект средства, равные разности между 150000 (общей выделенной суммой) и суммой, выделенной на три предыдущие проекта. Следовательно, менеджер располагает тремя степенями свободы.

В распределении Стьюдента мы вычисляем s, используя n наблюдений и оценивая один параметр (среднюю арифметическую). Отсюда $k = n - 1$ степеней свободы.

Математическое ожидание и дисперсия случайной величины t равны соответственно:

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

дисперсия существует только при $k > 2$.

Мода и медиана равны математическому ожиданию и равны нулю. Таким образом, распределение Стьюдента симметрично относительно точки $x = 0$.

На рис. 2.3 представлен график плотности распределения Стьюдента с тремя степенями свободы.

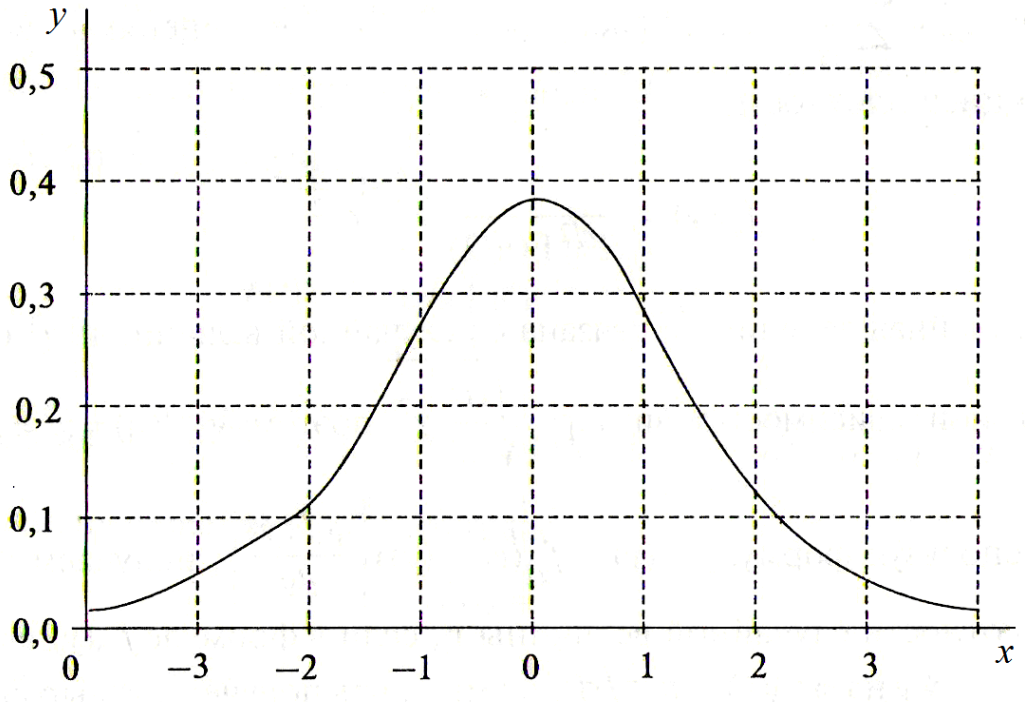


Рис. 2.3

График распределения Стьюдента отличается от графика стандартного нормального распределения, он более плосковершинный. Чем больше число степеней свободы, тем менее распределение Стьюдента отличается от нормального. В случае неограниченного возрастания числа степеней свободы распределение Стьюдента сходится к стандартному нормальному распределению. На практике формула (2.8) используется только тогда, когда мы имеем выборку, извлеченную из нормальной генеральной совокупности, объема $n \le 30$.

Аналитическое выражение функции плотности распределения Стьюдента имеет довольно сложную форму записи. Для определения вероятностей обычно пользуются готовыми таблицами (см. прил. 3).

Квантили распределения Стьюдента находятся из уравнения

$$P\left(|t_{n-1}| > t_{\alpha;n-1}\right) = \alpha$$

или

$$P\left(-t_{\alpha,n-1} < t_{n-1} < t_{\alpha,n-1}\right) = 1-\alpha\,.$$

Геометрически квантиль $t_{\alpha;n-1}$ есть такое значение случайной величины t, что суммарная площадь заштрихованных криволинейных трапеций равна α (рис. 2.4)

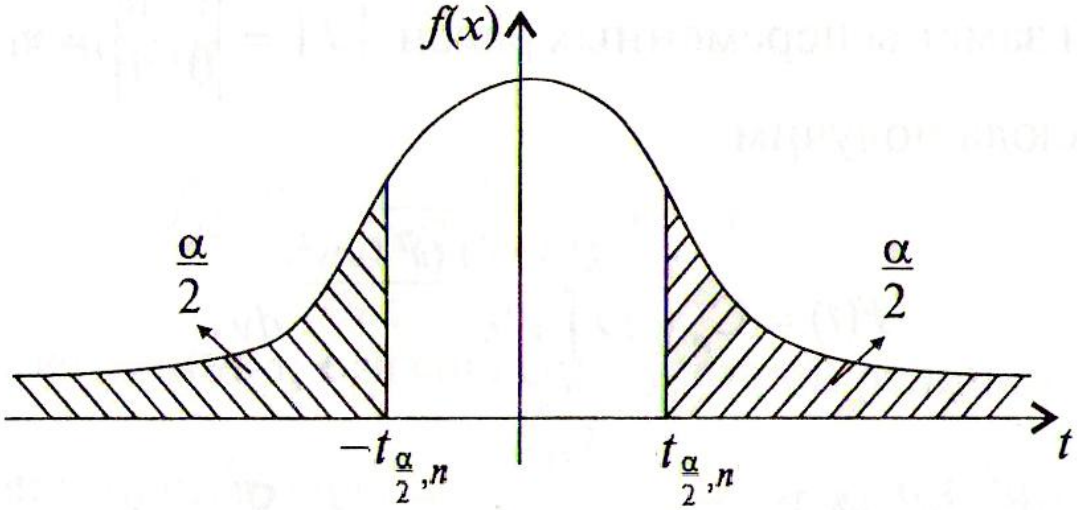


Рис. 2.4

Распределение хи-квадрат (закон Пирсона)

Распределением хи-квадрат $\left(\chi^2\right)$ с n степенями свободы, обозначается χ_n^2, называется распределение суммы квадратов n независимых случайных величин со стандартных нормальным распределением, т.е.

$$\chi_n^2 = \sum_{i=1}^{n} X_i^2 \text{, где } X_i \approx N_{0;1}.$$

Если независимые случайные величины $X_1, X_2, ..., X_n$ распределены по нормальному закону с параметрами a_i и σ_i, т.е. $X_i \sim N_{a_i;\sigma_i}$, их можно стандартизировать (центрируем и нормируем). Получим независимые случайные величины

$$Z_i = \frac{X_i - a_i}{\sigma_i},$$

которые будут иметь стандартное нормальное распределение $Z_i \sim N_{0;1}$, $i = 1,...,n$. Распределение суммы квадратов этих случайных величин

$$\chi_n^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - a_i}{\sigma_i} \right)^2, \text{ где } X_i \sim N_{a_i;\sigma_i},$$

по определению и является распределением χ^2 с n степенями свободы.

Распределение χ^2, так же как и распределение Стьюдента, зависит от одного параметра – числа степеней свободы.

Основные числовые характеристики распределения χ^2:

$$M\left(\chi^2\right)=k,\ D\left(\chi^2\right)=2k,$$

где $k-$ число степеней свободы.

На рис. 2.5 представлен график плотности распределения хи-квадрат с пятью степенями свободы.

Вид кривой распределения зависит от числа степеней свободы. Общий вид графика распределения хи-квадрат при различных значениях параметра n представлен на рис. 2.6. График плотности распределения χ^2 асимметричен, но, как видно, асимметрия уменьшается с увеличением числа степеней свободы. При неограниченном увеличении числа степеней свободы хи-квадрат распределение аппроксимируется нормальным распределением.

Значение квантили $\chi^2_{\alpha;n}$, соответствующей заданному уровню значимости α, определяется из уравнения $P\left(\chi^2>\chi^2_{\alpha;n}\right)=\alpha$. Геометрически квантиль $\chi^2_{\alpha;n}$ есть такое значение случайной величины, при котором площадь заштрихованной криволинейной трапеции равна α (рис. 2.6). Таблица квантилей χ^2 распределения содержит процентные точки только для $n\le 30$ (прил. 2). Для $n>30$ квантили $\chi^2_{\alpha;n}$ можно определять с помощью таблиц нормального закона распределения (прил. 1).

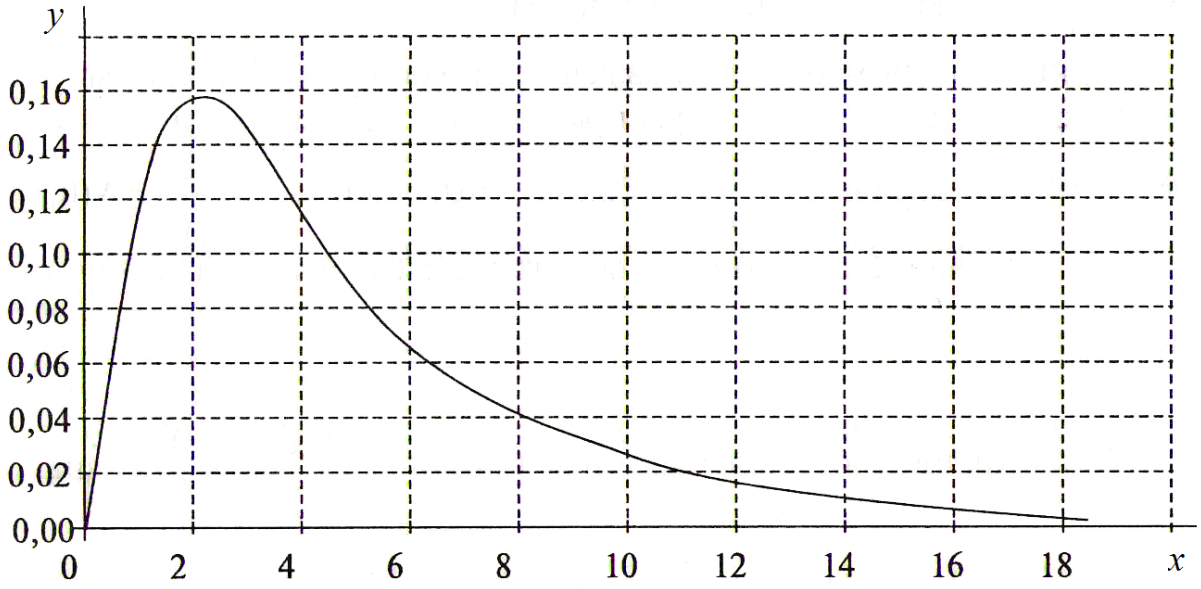


Рис. 2.5

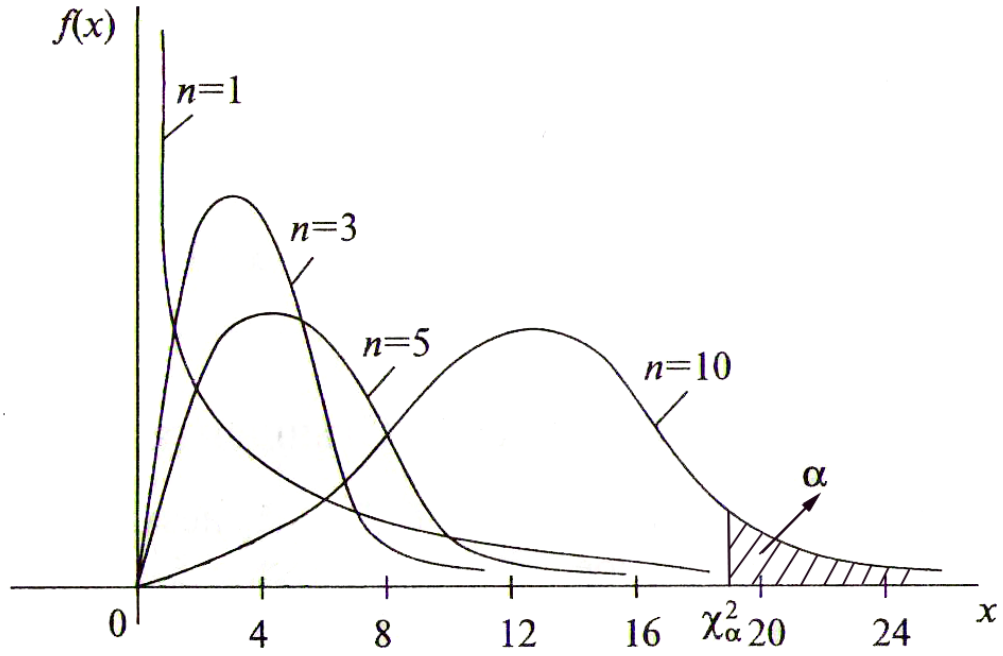


Рис. 2.6

Распределение Фишера

Зачастую мы сравниваем дисперсии. Например, необходимо знать, являются ли дисперсии двух распределений равными. Для ответа на этот вопрос и служит *F*-распределение.

Если χ_n^2 и χ_m^2 – независимые случайные величины, распределенные по закону χ^2 со степенями свободы n и m соответственно, то случайная величина

$$F(n,m)=\frac{\chi_n^2/n}{\chi_m^2/m}$$

имеет распределение, которое называют F-распределением Фишера (Фишера-Снедекора) со степенями свободы n и m, или распределением дисперсионного отношения.

Поскольку обе величины χ_n^2 и χ_m^2 неотрицательны, то случайная величина $F(n,m)\geq 0$ при $x>0$ и $F(n,m)=0$ при $x<0$.

На рис. 2.7 представлен график плотности распределения Фишера с числами степеней свободы $n=10$, $m=15$.

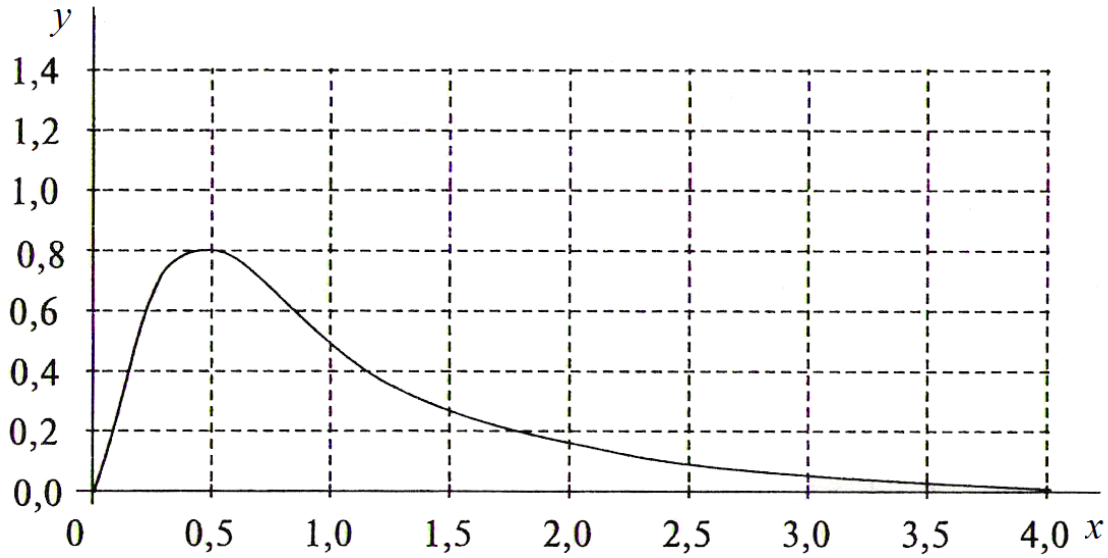


Рис. 2.7

Общий вид графика плотности распределения Фишера при различных значениях параметров приведен на рис. 2.8.

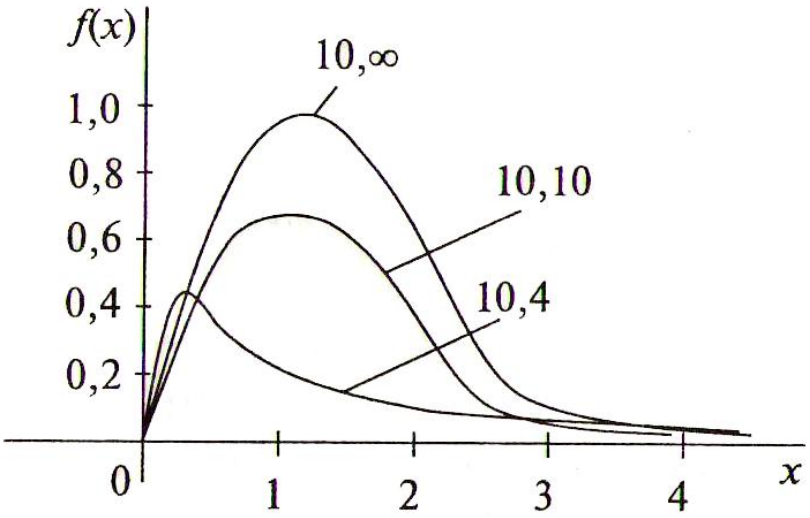


Рис. 2.8

Доверительные интервалы

Точечная оценка неизвестного параметра θ, найденная по выборке объемом n, не указывает на то, какую ошибку мы допускаем, принимая вместо точного значения параметра θ его приближенное значение $\hat{\theta}_n$. Поэтому вводят интервальную оценку, которая определяется двумя числами – концами интервала. Внутри этого интервала с определенной вероятностью находится неизвестное значение параметра θ, причем границы интервала не должны зависеть от искомого параметра.

Доверительным интервалом, или интервальной оценкой, называется интервал $\left(\hat{\theta}_1, \hat{\theta}_2\right)$, который покрывает неизвестный параметр θ с заданной вероятностью $0<\gamma<1$. Вероятность γ называют *надежностью доверительного интервала*, или *доверительной вероятностью*. Наименьшее число $\delta>0$, такое что для любой точки $\bar{\theta}\in\left(\hat{\theta}_1, \hat{\theta}_2\right)$ выполнено условие $\left|\bar{\theta}-\theta\right|<\delta$, называется *точностью* оценки (доверительного интервала). Ясно, что чем меньше длина интервала, тем точнее оценка. Число $\alpha=1-\gamma$ называется *уровнем значимости*.

Если границы доверительных интервалов, построенных формально, из каких-либо теоретических соображений, выходят за рамки возможного, их «округляют» до разумных пределов.

Доверительный интервал для математического ожидания нормально распределенной случайной величины при известной дисперсии

Пусть случайная величина X распределена по нормальному закону, т.е. $X \in N_{a,\sigma}$, причем значение a неизвестно, значение σ^2 – известно.

Задача заключается в том, чтобы построить доверительный интервал для неизвестного математического ожидания, соответствующий заданной надежности (доверительной вероятности) $\gamma = 1-\alpha$.

Для решения этой задачи из нормальной генеральной совокупности производится выборка $x_1, x_2, \ldots, x_n$ объема n. Элементы выборки независимы и имеют одинаковое распределение с дисперсией $Dx_i = \sigma^2$. На основании выборки найдена наилучшая несмещенная оценка математического ожидания $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Известно, что линейная комбинация нормально распределенных случайных величин является нормальной, и $\bar{x} \in N_{a,\sigma/\sqrt{n}}$ $\left(M\bar{x} = a, D\bar{x} = \sigma^2/n\right)$. Для нахождения интервальной оценки рассмотрим случайную величину $Z = \frac{\bar{x}-a}{\sigma/\sqrt{n}}$, распределенную по стандартному нормальному закону $N_{0,1}$. Найдем верхнюю и нижнюю критические границы для Z, соответствующие уровню значимости α:

$$P\left(|Z| < u_\alpha\right) = 1-\alpha.$$

Выше говорилось о том, что значение u_α находится из условия

$$\Phi_0\left(u_\alpha\right) = \frac{1-\alpha}{2} = \frac{\gamma}{2}, \qquad (2.9)$$

где $\Phi_0\left(x\right)$ – функция Лапласа, значения которой находятся по таблице (прил. 1).

Таким образом, с вероятностью $1-\alpha$ случайная величина Z попадает в интервал $\left(-u_\alpha, u_\alpha\right)$, т.е.

$$-u_\alpha < Z < u_\alpha, \text{ или } -u_\alpha < \frac{\bar{x}-a}{\sigma/\sqrt{n}} < u_\alpha.$$

Выразим математическое ожидание a из двойного неравенства и получим доверительный интервал для неизвестного математического ожидания

$$\bar{x} - \frac{\sigma}{\sqrt{n}} u_\alpha < a < \bar{x} + \frac{\sigma}{\sqrt{n}} u_\alpha . \qquad (2.10)$$

Как мы видим, получился симметричный интервал, который можно записать в виде

$$\left|\bar{x} - a\right| < \frac{\sigma}{\sqrt{n}} u_\alpha .$$

Обозначим за $\delta = \frac{\sigma}{\sqrt{n}} u_\alpha$, тогда неравенство примет вид $\left|\bar{x} - a\right| < \delta$. Величина δ показывает, насколько выборочное среднее $\bar{x}$ отличается от истинного значения математического ожидания a, поэтому δ называют *точностью вычисления* или *предельной ошибкой* вычисления.

Смысл полученного результата следующий: с надежностью $1-\alpha$ доверительный интервал $\left(\bar{x} - \frac{\sigma}{\sqrt{n}} u_\alpha ; \bar{x} + \frac{\sigma}{\sqrt{n}} u_\alpha\right)$ покрывает неизвестный параметр a с точностью, равной

$$\delta = \frac{\sigma}{\sqrt{n}} u_\alpha . \qquad (2.11)$$

Из этой формулы следует, что увеличение объема выборки приводит к уменьшению доверительного интервала, следовательно, к увеличению точности интервальной оценки. Из формулы (2.11) можно найти минимальный объем выборки, который обеспечит заданную точность δ:

$$n = \frac{\sigma^2}{\delta^2} u_\alpha^2 . \qquad (2.12)$$

Следствие. Рассмотрим схему испытаний Бернулли: проводится n независимых испытаний, в каждом из которых событие A либо происходит, либо нет. Пусть вероятность p того, что событие A произойдет в одном испытании, неизвестна. Требуется построить доверительный интервал с надежностью γ для неизвестной вероятности.

Пусть случайная величина X равна числу появлений события A в n испытаниях. Тогда, согласно центральной предельной теореме для схемы испытаний Бернулли, случайная величина

$$Z = \frac{X - np}{\sqrt{npq}}$$

имеет стандартное нормальное распределение при большом числе n.

Тогда с вероятностью $\gamma = 1 - \alpha$ выполняется неравенство

$$-u_\alpha < Z < u_\alpha, \text{ или } -u_\alpha < \frac{X - np}{\sqrt{npq}} < u_\alpha.$$

Сделаем преобразования

$$\frac{X - np}{\sqrt{npq}} = \frac{n\left(X/n - p\right)}{n\sqrt{pq/n}} = \frac{X/n - p}{\sqrt{pq/n}},$$

получим

$$-u_\alpha < \frac{X/n - p}{\sqrt{pq/n}} < u_\alpha, \frac{X}{n} - u_\alpha\sqrt{\frac{pq}{n}} < p < \frac{X}{n} + u_\alpha\sqrt{\frac{pq}{n}}.$$

Пусть в результате n независимых испытаний событие A произошло k раз, тогда $X = k$ и величину k/n можно принять за оценку $\hat{p}$ неизвестной вероятности p (при большом n). Заменяя значения p и $q = 1 - p$ в левой и правой частях неравенства их оценками $\hat{p} = k/n$, $\hat{q} = 1 - \hat{p}$, получим приближенный доверительный интервал для вероятности p

$$\hat{p} - u_\alpha\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + u_\alpha\sqrt{\frac{\hat{p}\hat{q}}{n}}. \qquad (2.13)$$

Примеры

1. Пусть случайная величина $X \in N_{a,2}$. Найти доверительный интервал для математического ожидания a, если $n = 16$; $\overline{x} = 20{,}01$; $\gamma = 1 - \alpha = 0{,}95$.

Решение. Сначала находим значение u_α, используя таблицу функции Лапласа (прил. 1):

$$\Phi_0\left(u_\alpha\right) = \frac{1 - \alpha}{2} = \frac{0{,}95}{2} = 0{,}475.$$

Находим по таблице $u_\alpha = 1{,}96$. Тогда точность оценки, согласно формуле (2.11), будет равна

$$\delta = \frac{2 \cdot 1{,}96}{\sqrt{16}} = 0{,}98.$$

Доверительный интервал будет

$$(\bar{x}-\delta;\bar{x}+\delta)=(20,01-0,98;20,01+0,98)=(19,03;\ 20,99).$$

2. Фирма коммунального хозяйства желает на основе выборки оценить среднюю квартплату за квартиры определенного типа с надежностью не менее 99 % и точностью, меньшей 10 ден.ед. Предполагая, что квартплата имеет нормальное распределение со средним квадратическим отклонением, равным 35 д.е., найдите минимальный объем выборки.

Решение. По условию требуется найти такое n, при котором

$$P(|\bar{x}-a|<10)\geq 0,99.$$

Приравняв $1-\alpha=0,99$, по таблице функции Лапласа находим u_α, при котором

$$\Phi_0(u_\alpha)=\frac{1-\alpha}{2}=0,495;\ u_{0,01}=2,6.$$

При $\delta=10$ и $\sigma=35$ из формулы (2.12) получим

$$n=\frac{u_{0,01}^2\cdot\sigma^2}{\delta^2}=\frac{2,6^2\cdot 35^2}{100}=82,81.$$

Но так как с ростом $1-\alpha$ и уменьшением δ растет n, то $n\geq 82,81$ и минимальный объем выборки, который обеспечивает точность вычисления, равную 10 д.е., $n_{\min}=83$.

Доверительный интервал для математического ожидания нормально распределенной случайной величины при неизвестной дисперсии

Пусть случайная величина $X\in N_{a,\sigma}$, и дисперсия для нее неизвестна.

Задача состоит в следующем: построить доверительный интервал для неизвестного математического ожидания, соответствующий доверительной вероятности $\gamma=1-\alpha$.

Для решения задачи из генеральной совокупности произведена выборка $x_1,\ldots,x_n$ объемом n. На основании выборки найдены наилучшие несмещенные оценки неизвестных параметров

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i;\ S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Известно, если $x_i \in N_{a,\sigma}$ для всех i, то $\bar{x} \in N_{a,\sigma/\sqrt{n}}$. Тогда $Z = \dfrac{\bar{x} - a}{\sigma/\sqrt{n}}$ имеет стандартное нормальное распределение. Так как дисперсия σ^2 неизвестна, среднее квадратическое отклонение σ в случайной величине Z заменим на исправленное среднее квадратическое отклонение s, при этом получим случайную величину

$$t = \frac{\bar{x} - a}{s/\sqrt{n}},$$

которая имеет распределение Стьюдента с $n-1$ степенями свободы. Введем критические границы распределения Стьюдента с $n-1$ степенями свободы, обеспечивающие надежность, равную $\gamma = 1 - \alpha$. Используя симметричность двусторонних критических границ распределения Стьюдента, имеем

$$P\left(|t| < t_{\alpha,n-1}\right) = 1 - \alpha.$$

Получаем, что с вероятностью $\gamma = 1 - \alpha$ выполняется неравенство

$$-t_{\alpha,n-1} < \frac{\bar{x} - a}{s/\sqrt{n}} < t_{\alpha,n-1}.$$

Решая неравенство относительно a, получим, что с вероятностью $1 - \alpha$ выполняется неравенство

$$\bar{x} - \frac{t_{\alpha,n-1} \cdot s}{\sqrt{n}} < a < \bar{x} + \frac{t_{\alpha,n-1} \cdot s}{\sqrt{n}}. \qquad (2.14)$$

Значения $t_{\alpha,n-1}$ находятся из таблиц распределения Стьюдента с уровнем значимости α и $n-1$ степенями свободы (прил. 3). Точность оценки вычисляется по формуле

$$\delta = \frac{s}{\sqrt{n}} t_{\alpha,n-1}. \qquad (2.15)$$

Наименьший объем выборки, который обеспечивает заданную точность δ, равен

$$n = \frac{S^2}{\delta^2} t^2_{\alpha,n-1}. \qquad (2.16)$$

Полученный интервал, как и в случае известной дисперсии, улучшить нельзя.

Таким образом, если дисперсия известна, то используют критическую границу нормального распределения, а если дисперсия неизвестна – критические границы распределения Стьюдента с $n-1$ степенями свободы. При малых выборках эти границы различаются значительно, но при больших выборках этой разницей уже можно пренебречь. Поэтому в случае неизвестной дисперсии и большого числа n с вероятностью $\gamma = 1-\alpha$ доверительный интервал для неизвестного математического ожидания можно брать

$$\bar{x} - \frac{u_\alpha \cdot s}{\sqrt{n}} < a < \bar{x} + \frac{u_\alpha \cdot s}{\sqrt{n}}.$$

Можно также по выборке $x_1, x_2, ..., x_n$ построить доверительный интервал для следующего, $(n+1)$-го наблюдения. А именно:

$$\bar{x} - s \cdot t_{\alpha, n-1}\sqrt{1+\frac{1}{n}} < x_{n+1} < \bar{x} + s \cdot t_{\alpha, n-1}\sqrt{1+\frac{1}{n}}. \qquad (2.17)$$

Данная формула может быть полезна в качестве прогноза на будущее.

Пример. Из генеральной совокупности извлечена выборка объемом $n = 12$:

x_i	–0,5	–0,4	–0,2	0	0,2	0,6	0,8	1	1,2	1,5
n_i	1	2	1	1	1	1	1	1	2	1

Оценить с надежностью 0,95 математическое ожидание a нормально распределенного признака генеральной совокупности с помощью доверительного интервала.

Решение. Найдем выборочное среднее $\bar{x}$ и исправленное выборочное среднее квадратическое отклонение s. Для этого перейдем к условным вариантам $u_i = 10 \cdot x_i$, тогда

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{10} n_i x_i = 4,2,$$

следовательно,

$$\bar{x} = \frac{\bar{u}}{10} = 0,42,$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{10} n_i x_i^2 - \frac{1}{n-1}\left(\frac{1}{n}\sum_{i=1}^{10} n_i x_i\right)^2 = 0,52;\ s = \sqrt{S^2} = \sqrt{0,52} = 0,72.$$

Находим для уровня значимости $\alpha = 1-\gamma = 1-0,95 = 0,05$ и числа степеней свободы $n-1=11$ по таблице распределения Стьюдента критическую точку $t_{0,05,11} = 2,2$. Определяем границы доверительного интервала:

$$\overline{x} - \frac{t_{\alpha,n-1}\cdot s}{\sqrt{n}} = 0,42 - \frac{2,2\cdot 0,72}{\sqrt{12}} = -0,04;\ \overline{x} + \frac{t_{\alpha,n-1}\cdot s}{\sqrt{n}} = 0,42 + \frac{2,2\cdot 0,72}{\sqrt{12}} = 0,88.$$

Таким образом, искомый доверительный интервал: $(-0,04;\ 0,88)$.

Доверительный интервал для дисперсии случайной величины, распределенной по нормальному закону при известном математическом ожидании

Пусть случайная величина $X \in N_{a,\sigma}$, причем параметр σ неизвестен.

Задача состоит в следующем: построить доверительный интервал для неизвестной дисперсии, соответствующий заданной надежности $\gamma = 1-\alpha$.

Наилучшей оценкой дисперсии, найденной по выборке $x_1, x_2, \ldots, x_n$ объема n при известном математическом ожидании a, будет величина $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$. Нормируем величину S_0^2, разделив ее на масштабный множитель σ^2:

$$\frac{S_0^2}{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-a)^2}{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - a}{\sigma}\right)^2,$$

так как $\frac{x_i - a}{\sigma} \in N_{0,1}$, то, по определению, случайная величина $\chi^2 = \frac{n\cdot S_0^2}{\sigma^2}$ имеет распределение χ^2 с n степенями свободы. Зададим доверительную вероятность $\gamma = 1-\alpha$. Найдем верхнюю и нижнюю критические границы $\overline{\chi}^2_{\alpha,n}$ и $\underline{\chi}^2_{\alpha,n}$, соответствующие уровню значимости α, т.е.

$$P\left(\underline{\chi}^2_{\alpha,n} < \chi^2 < \overline{\chi}^2_{\alpha,n}\right) = 1-\alpha.$$

Критические границы ищем из условия, что прямые $x=\underline{\chi}^2_{\alpha,n}$ и $x=\overline{\chi}^2_{\alpha,n}$ отсекают от криволинейной трапеции, ограниченной плотностью вероятностей случайной величины χ^2, области одинаковой площади $\alpha/2$. То есть

$$P\left(\chi^2>\overline{\chi}^2_{\alpha,n}\right)=\frac{\alpha}{2},\ P\left(\chi^2<\underline{\chi}^2_{\alpha,n}\right)=\frac{\alpha}{2}.$$

Из первого равенства, по определению квантили для распределения хи-квадрат, получаем $\overline{\chi}^2_{\alpha,n}=\chi^2_{\alpha/2,n}$. Из второго равенства, учитывая, что площадь всей криволинейной трапеции, ограниченной плотностью вероятностей, равна 1, имеем

$$P\left(\chi^2>\underline{\chi}^2_{\alpha,n}\right)=1-\frac{\alpha}{2},$$

следовательно, нижняя критическая граница

$$\underline{\chi}^2_{\alpha,n}=\chi^2_{1-\alpha/2,n}.$$

Таким образом, с вероятностью $\gamma=1-\alpha$ случайная величина χ^2 попадает в интервал $\left(\chi^2_{1-\alpha/2,n};\chi^2_{\alpha/2,\,n}\right)$, т.е.

$$\chi^2_{1-\alpha/2,n}<\chi^2<\chi^2_{\alpha/2,n}.$$

Разрешаем полученный интервал

$$\chi^2_{1-\alpha/2,n}<\frac{n\cdot S_0^2}{\sigma^2}<\chi^2_{\alpha/2,n}$$

относительно σ^2 и получаем доверительный интервал для неизвестной дисперсии:

$$\frac{n\cdot S_0^2}{\chi^2_{\alpha/2,n}}<\sigma^2<\frac{n\cdot S_0^2}{\chi^2_{1-\alpha/2,n}}. \qquad (2.18)$$

Значения $\chi^2_{\alpha/2,n}$ и $\chi^2_{1-\alpha/2,n}$ ищутся по таблицам распределения χ^2 с уровнями значимости $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$ соответственно и n степенями свободы (прил. 2).

Для того, чтобы получить доверительный интервал для неизвестного среднего квадратического отклонения σ, нужно извлечь квадратный корень из всех частей двойного неравенства (2.18):

$$\sqrt{\frac{n\cdot S_0^2}{\chi^2_{\alpha/2,n}}}<\sigma<\sqrt{\frac{n\cdot S_0^2}{\chi^2_{1-\alpha/2,n}}}. \qquad (2.18а)$$

Доверительный интервал для дисперсии случайной величины, распределенной по нормальному закону при неизвестном математическом ожидании

Пусть случайная величина $X\in N_{a,\sigma}$, причем параметр σ неизвестен.

Задача состоит в следующем: построить доверительный интервал для неизвестной дисперсии, соответствующий заданной надежности $\gamma=1-\alpha$.

Для решения задачи из генеральной совокупности произведена выборка $x_1,\ldots,x_n$ объемом n. На основании выборки найдем точечные несмещенные оценки неизвестных параметров

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i;\ S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

Случайная величина $\chi^2=\frac{(n-1)S^2}{\sigma^2}$ имеет распределение χ^2 с $n-1$ степенями свободы. Зададим доверительную вероятность $\gamma=1-\alpha$. Тогда

$$P\left(\underline{\chi}^2_{\alpha,n-1}<\chi^2<\overline{\chi}^2_{\alpha,n-1}\right)=1-\alpha,$$

где $\underline{\chi}^2_{\alpha,n-1}$ и $\overline{\chi}^2_{\alpha,n-1}$ – нижняя и верхняя критические границы, соответствующие уровню значимости α. Аналогично предыдущему случаю, получим

$$\underline{\chi}^2_{\alpha,n-1}=\chi^2_{1-\alpha/2,n-1},\ \overline{\chi}^2_{\alpha,n-1}=\chi^2_{\alpha/2,n-1}.$$

Таким образом, с вероятностью $\gamma = 1 - \alpha$ случайная величина χ^2 попадет в интервал

$$\chi^2_{1-\alpha/2,\,n-1} < \chi^2 < \chi^2_{\alpha/2,\,n-1}.$$

Разрешаем полученный интервал

$$\chi^2_{1-\alpha/2,\,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,\,n-1}$$

относительно σ^2 и получаем доверительный интервал для неизвестной дисперсии:

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,\,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\,n-1}}. \qquad (2.19)$$

Значения $\chi^2_{\alpha/2,\,n-1}$ и $\chi^2_{1-\alpha/2,\,n-1}$ ищутся по таблицам распределения χ^2 с уровнями значимости $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$ соответственно и числом степеней свободы $n-1$.

Доверительный интервал для среднего квадратического отклонения σ при неизвестном математическом ожидании a вычисляем по формуле

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,\,n-1}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\,n-1}}}. \qquad (2.19а)$$

Пример. Для отрасли, включающей 1200 фирм, составлена случайная выборка из 19 фирм. По выборке оказалось, что исправленное среднее квадратическое отклонение для числа работающих на фирме составляет $s = 25$ (человек). Пользуясь 90 %-ным доверительным интервалом, оценить среднее квадратическое отклонение для числа работающих на фирме по всей отрасли.

Решение. Так как среднее квадратическое отклонение есть корень квадратный из дисперсии, то необходимо составить доверительный интервал для неизвестной дисперсии и потом извлечь квадратный корень из концов интервала. Таким образом получим доверительный интервал для среднего квадратического отклонения.

Необходимо построить доверительный интервал для неизвестной дисперсии при неизвестном математическом ожидании с заданной точностью $1-\alpha=0,9$. Значит $\alpha=0,1;\ \alpha/2=0,05;\ 1-\alpha/2=0,95$. По таблице определяем для данной задачи $\chi^2_{\alpha/2,\,n-1}=\chi^2_{0,05;18}=28,9$ и $\chi^2_{1-\alpha/2,\,n-1}=\chi^2_{0,95;18}=9,39$. Подставив в формулу необходимые величины, получаем доверительный интервал:

$$\frac{18\cdot 25^2}{28,9}<\sigma^2<\frac{18\cdot 25^2}{9,39},$$

следовательно, искомый доверительный интервал имеет вид

$$19,74<\sigma<34,61 \text{ (человек)}.$$

Замечание. Поскольку при увеличении объема выборки распределение χ^2 приближается к нормальному, при достаточно большом объеме выборки доверительный интервал для дисперсии можно найти по формуле

$$P\left(\frac{s}{1+\frac{1}{\sqrt{2n}}u_\alpha}<\sigma<\frac{s}{1-\frac{1}{\sqrt{2n}}u_\alpha}\right)=1-\alpha,$$

где u_α находим по таблице функции Лапласа из условия $\Phi_0(u_\alpha)=\frac{1-\alpha}{2}$.

Примеры решения задач к главе 2

1. Найти несмещенные оценки математического ожидания и дисперсии, начальные моменты второго и третьего порядков и центральные моменты первого и второго порядков по выборке объема $n=20$:

x_i	–1	1	2	3	5
n_i	2	3	10	4	1

Решение. Несмещенная оценка математического ожидания вычисляется по формуле (2.1а):

$$\bar{x}=\frac{2(-1)+3\cdot 1+10\cdot 2+4\cdot 3+1\cdot 5}{20}=1,9.$$

Вычислим начальные моменты второго и третьего порядков по формуле (2.5а). Берем $k=2$ и $k=3$ соответственно:

$$\overline{x^2}=\frac{2(-1)^2+3\cdot 1^2+10\cdot 2^2+4\cdot 3^2+1\cdot 5^2}{20}=\frac{2+3+40+36+25}{20}=5,3;$$

$$\overline{x^3}=\frac{2(-1)^3+3\cdot 1^3+10\cdot 2^3+4\cdot 3^3+1\cdot 5^3}{20}=\frac{-2+3+80+108+125}{20}=15,7.$$

Теперь вычислим центральный момент первого порядка по формуле (2.6а), в формуле надо взять $k=1$:

$$\hat{\mu}_1=\frac{1}{n}\sum_{i=1}^{k}n_i(x_i-\bar{x})^1=\frac{1}{n}\sum_{i=1}^{k}n_i(x_i-\bar{x})=\frac{\sum_{i=1}^{k}n_ix_i}{n}-\bar{x}=\bar{x}-\bar{x}=0,$$

таким образом, *центральный момент первого порядка всегда равен нулю*.

Центральный момент второго порядка – это неисправленная выборочная дисперсия, ее удобнее вычислить по формуле

$$\hat{\mu}_2=\overline{D}=\overline{x^2}-\left(\bar{x}\right)^2=5,3-(1,9)^2=5,3-3,61=1,69.$$

Следует обратить внимание на то, что *выборочная дисперсия не может быть отрицательным числом*.

Осталось вычислить неисправленную выборочную дисперсию S^2, для этого используем формулу (2.4):

$$S^2=\frac{20}{20-1}\overline{D}=\frac{20}{19}\cdot 1,69=1,78.$$

Ответ: выборочное среднее $\bar{x}=1,9$; начальный момент второго порядка $\hat{\nu}_2=\overline{x^2}=5,3$; начальный момент третьего порядка $\hat{\nu}_3=\overline{x^3}=15,7$; центральный момент первого порядка $\hat{\mu}_1=0$; центральный момент второго порядка $\hat{\mu}_2=1,69$ и несмещенная выборочная дисперсия $S^2=1,78$.

2. Найти выборочное среднее по данному распределению выборки объема $n=100$:

x_i	2702	2804	2903	3028
n_i	8	30	60	2

Решение. Для того чтобы упростить вычисления, перейдем к условным вариантам

$$u_i = x_i - 2844.$$

Получим следующее распределение:

u_i	–142	–40	59	184
n_i	8	30	60	2

Вычислим выборочное среднее условной величины по формуле (2.1а):

$$\bar{u} = \frac{8(-142) + 30(-40) + 60 \cdot 59 + 2 \cdot 184}{100} = \frac{-1136 - 1200 + 3540 + 368}{100} = 15,72.$$

Так как $u = x - 2844$, то $\bar{u} = \bar{x} - 2844$, следовательно

$$\bar{x} = \bar{u} + 2844 = 15,72 + 2844 = 2859,72.$$

Отметим, что $\overline{D}(u) = \overline{D}(x - 2844) = \overline{D}(x)$. То есть при переходе к условным вариантам путем сдвига на некоторую константу выборочная дисперсия не меняется.

3. Найти выборочное среднее и выборочную дисперсию по данному распределению выборки объема $n = 10$:

x_i	0,01	0,04	0,08
n_i	5	3	2

Решение. Перейдем к условным вариантам

$$u_i = 100 \cdot x_i.$$

Получим следующее распределение:

u_i	1	4	8
n_i	5	3	2

Вычислим сначала выборочное среднее $\bar{u}$ и $\overline{u^2}$:

$$\overline{u} = \frac{5+12+16}{10} = 3{,}3,$$

$$\overline{u^2} = \frac{5+3\cdot 16+2\cdot 64}{10} = 18{,}1.$$

Выборочная дисперсия условной величины равна

$$\overline{D}(u) = \overline{u^2} - \left(\overline{u}\right)^2 = 18{,}1 - (3{,}3)^2 = 7{,}21.$$

Так как $u = 100 \cdot x$, то $\overline{u} = 100 \cdot \overline{x}$. Итак, выборочное среднее исходной варианты

$$\overline{x} = \frac{\overline{u}}{100} = \frac{3{,}3}{100} = 0{,}033.$$

Аналогично, используя свойства дисперсии, получаем

$$\overline{D}(u) = 100^2 \cdot \overline{D}(x).$$

Следовательно,

$$\overline{D}(x) = \frac{\overline{D}(u)}{100^2} = \frac{7{,}21}{10000} = 0{,}000721.$$

4. Путем опроса получены следующие данные $(n = 80)$:

2 4 2 4 3 3 3 2 0 6	1 2 3 2 2 4 3 3 5 1	0 2 4 3 2 2 3 3 1 3
3 3 1 1 2 3 1 4 3 1	7 4 3 4 2 3 2 3 3 1	4 3 1 4 5 3 4 2 4 5
3 6 4 1 3 2 4 1 3 1	0 0 4 6 4 7 4 1 3 5	

Найти основные числовые характеристики вариационного ряда (по возможности использовать упрощающие формулы для их нахождения):

1) выборочное среднее $\overline{x}$;

2) смещенную и несмещенную оценки дисперсии $\overline{D}$ и S^2;

3) выборочное среднее квадратическое отклонение $\hat{\sigma}$;

4) коэффициент вариации V.

Решение. Для составления дискретного вариационного ряда отсортируем данные опроса по величине и расположим их в порядке возрастания:

0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 6 6 6 7 7.

Более компактно эти данные можно представить в виде статистического распределения выборки (в виде таблицы, в которой первая строка – варианты (наблюдаемые значение), вторая строка – частоты появления этих вариант):

x_i	0	1	2	3	4	5	6	7
n_i	4	13	14	24	16	4	3	2

Найдем основные числовые характеристики вариационного ряда. Вычислим выборочное среднее $\bar{x}$ и выборочную дисперсию $\overline{D}$. Для этого составим расчетную таблицу:

x_i	n_i	$x_i \cdot n_i$	$(x_i - \bar{x})^2$	$n_i \cdot (x_i - \bar{x})^2$
0	4	0	8,1796	32,7184
1	13	13	3,4596	44,9748
2	14	28	0,7396	10,3544
3	24	72	0,0196	0,4704
4	16	64	1,2996	20,7936
5	4	20	4,5796	18,3184
6	3	18	9,8596	29,5788
7	2	14	17,1396	34,2792
Сумма	80	229		191,488

Используя суммы, полученные в расчетной таблице, найдем:
1) выборочное среднее

$$\overline{x} = \frac{\sum_{i=1}^{m} x_i n_i}{n} = \frac{229}{80} \approx 2{,}86;$$

2) выборочную дисперсию

$$\overline{D} = \frac{\sum_{i=1}^{m} n_i \left(x_i - \overline{x}\right)^2}{n} = \frac{191{,}488}{80} = 2{,}39.$$

Исправленную выборочную дисперсию вычисляем по формуле

$$S^2 = \frac{n}{n-1}\overline{D} = \frac{80}{79} \cdot 2{,}39 \approx 2{,}42.$$

Замечание. Если первоначальные варианты – большие числа или, наоборот, слишком малы и к тому же являются равноотстоящими, то удобно перейти к условным вариантам:

$$u_i = \frac{x_i - c}{k},$$

где c и k – произвольные числа.

В качестве c целесообразно выбирать одно из средних значений признака X, а в качестве k – разность между двумя соседними вариантами. В этом случае формулы для упрощенного вычисления принимают следующий вид:

– выборочное среднее

$$\overline{x} = \frac{\sum_{i=1}^{m} u_i n_i}{n} \cdot k + c;$$

– выборочная дисперсия

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

либо можно использовать такую формулу:

$$\overline{D} = \frac{\sum_{i=1}^{m} n_i \left(u_i - \overline{u}\right)^2}{n} \cdot k^2,$$

где $\overline{u}$ – выборочное среднее условной величины U.

Для вычисления основных числовых характеристик воспользуемся упрощающими формулами, для чего составим расчетную таблицу. Выберем $c = 3$ и $k = 1$.

Расчетная таблица.

x_i	n_i	$u_i = x_i - 3$	$u_i \cdot n_i$	$u_i^2 \cdot n_i$
0	4	–3	–12	36
1	13	–2	–26	52
2	14	–1	–14	14
3	24	0	0	0
4	16	1	16	16
5	4	2	8	16
6	3	3	9	27
7	2	4	8	32
Сумма	80		–11	193

Используя суммы, полученные в расчетной таблице, найдем:

1) выборочное среднее

$$\overline{x} = \frac{\sum_{i=1}^{m} u_i n_i}{n} \cdot k + c = \frac{-11}{80} \cdot 1 + 3 \approx 2{,}86;$$

2) выборочную дисперсию

$$\overline{D} = \frac{\sum_{i=1}^{m} n_i u_i^{\,2}}{n} \cdot k^2 - \left(\overline{x} - c\right)^2 = \frac{193}{80} \cdot 1^2 - \left(2{,}86 - 3\right)^2 \approx 2{,}39;$$

3) выборочное среднее квадратическое отклонение

$$\hat{\sigma} = \sqrt{\overline{D}} = \sqrt{2,39} \approx 1,55;$$

4) коэффициент вариации

$$V = \frac{\hat{\sigma}}{\overline{x}} \cdot 100\% = \frac{1,55}{2,86} \cdot 100\% \approx 54,2\%.$$

Замечание. Смысл полученных результатов заключается в том, что величина $\overline{x} \approx 2,86$ характеризует среднее значение признака X, т.е. среднее значение составило 2,86. Среднее квадратическое отклонение $\hat{\sigma}$ описывает абсолютный разброс значений показателя X и в данном случае составляет $1,55$. Коэффициент вариации V характеризует относительную изменчивость показателя X, т.е. относительный разброс вокруг его среднего значения $\overline{x}$, и в данном случае составляет $54,2\ \%$.

Ответ. $\overline{x} \approx 2,86$; $\overline{D} \approx 2,39$; $S^2 \approx 2,42$; $\hat{\sigma} \approx 1,55$; $V \approx 54,2\ \%$.

5. Найти доверительный интервал для оценки с надежностью 0,9 неизвестного математического ожидания a нормально распределенного признака X генеральной совокупности, если среднее квадратическое отклонение $\sigma = 5$, выборочная средняя $\overline{x} = 20$ и объем выборки $n = 100$.

Решение. Итак, нам известны выборочное среднее $\overline{x} = 20$, среднее квадратическое отклонение $\sigma = 5$, надежность $\gamma = 1 - \alpha = 0,9$ и объем выборки $n = 100$. Требуется найти доверительный интервал

$$\overline{x} - \frac{\sigma}{\sqrt{n}} u_\alpha < a < \overline{x} + \frac{\sigma}{\sqrt{n}} u_\alpha.$$

Все величины, кроме u_α, известны. Найдем u_α из формулы

$$\Phi_{0,1}(u_\alpha) = \frac{1-\alpha}{2} = \frac{0,9}{2} = 0,45.$$

По таблице значений функции Лапласа находим значение $u_\alpha = 1,65$, и получаем доверительный интервал $19,175 < a < 20,825$.

Ответ. $19,175 < a < 20,825$.

6. Используя условия задачи 5, найти доверительный интервал для математического ожидания, но в данной задаче считаем, что среднее квадратическое отклонение σ неизвестно, а известно исправленное выборочное среднее квадратическое отклонение $s = 5,67$.

Решение. Если среднее квадратическое отклонение неизвестно, то для оценки $M(X) = a$ служит доверительный интервал

$$\bar{x} - \frac{t_{\alpha, n-1} \cdot s}{\sqrt{n}} < a < \bar{x} + \frac{t_{\alpha, n-1} \cdot s}{\sqrt{n}}.$$

Значения $t_{\alpha, n-1}$ находятся из таблиц распределения Стьюдента с уровнем значимости α и $n-1$ степенями свободы. Итак, имеем $\bar{x} = 20$, $s = 5,67$, $n = 100$, $t_{\alpha, n-1} = t_{0,1;99} = 1,665$. Получим доверительный интервал

$$20 - \frac{1,665 \cdot 5,67}{10} < a < 20 + \frac{1,665 \cdot 5,67}{10};$$

$$19,056 < a < 20,944.$$

Ответ. (19,056; 20,944).

7. Найти минимальный объем выборки, при которой с надежностью 0,925 точность оценки математического ожидания нормально распределенной генеральной совокупности по выборочной средней равна 0,2, если известно среднее квадратическое отклонение генеральной совокупности $\sigma = 1,5$.

Решение. Для определения точности оценки используем формулу

$$\delta = \frac{\sigma}{\sqrt{n}} u_{\alpha}.$$

Из этой формулы выражаем n, получим

$$n = \frac{u_{\alpha}^2 \cdot \sigma^2}{\delta^2}.$$

При этом n обычно округляется в большую сторону для надежности. По таблице функции Лапласа находим u_{α} из условия

$$\Phi_{0,1}(u_{\alpha}) = \frac{0,975}{2} = 0,4625.$$

Таким образом, $u_{\alpha} = 1,78$. Подставляем эти данные и получаем искомый объем выборки n:

$$n = \frac{1,78^2 \cdot 1,5^2}{0,2^2} \approx 178,22.$$

Берем округленно $n = 179$.

Ответ. $n=179$.

8. Из генеральной совокупности извлечена выборка объема $n=50$:

x_i	–1	0	1	2	3
n_i	10	5	15	15	5

Оценить с надежностью 0,95 математическое ожидание нормально распределенного признака генеральной совокупности по выборочной средней.

Решение. Выборочную среднюю и исправленную выборочную дисперсию найдем соответственно по формулам (2.1а) и (2.4)

$$\overline{x}=\frac{-10+15+30+15}{50}=1;$$

$$S^2=\frac{(-2)^2\cdot 10+(-1)^2\cdot 5+1\cdot 15+2^2\cdot 5}{49}\approx 1{,}63.$$

Найдем исправленное среднее квадратическое отклонение

$$s=\sqrt{S^2}=\sqrt{1{,}63}\approx 1{,}28.$$

Пользуясь таблицей распределения Стьюдента, по $\alpha=0{,}95$ и $n=50$ находим $t_{\alpha,n-1}=2{,}01$.

Найдем искомый доверительный интервал:

$$\overline{x}-\frac{t_{\alpha,n-1}\cdot s}{\sqrt{n}}<a<\overline{x}+\frac{t_{\alpha,n-1}\cdot s}{\sqrt{n}},$$

подставляя $\overline{x}=1$, $t_{\alpha,n-1}=2{,}01$, $s\approx 1{,}28$, $n=50$, получим

$$0{,}64<a<1{,}36.$$

Ответ. (0,64; 1,36).

9. По данным выборки объема $n=16$ из генеральной совокупности найдено исправленное среднее квадратическое отклонение $s=1$ нормально распределенного количественного признака. Найти доверительный интервал, покрывающий генеральное среднее квадратическое отклонение σ с надежностью 0,95.

Решение. Задача сводится к отысканию доверительного интервала

$$\frac{(n-1)\cdot S^2}{\chi^2_{\alpha/2,\,n-1}} < \sigma^2 < \frac{(n-1)\cdot S^2}{\chi^2_{1-\alpha/2,\,n-1}}.$$

Значения $\chi^2_{\alpha/2,\,n-1}$ и $\chi^2_{1-\alpha/2,\,n-1}$ ищутся по таблицам распределения χ^2 с уровнями значимости $\alpha/2$ и $1-\alpha/2$ соответственно и $n-1$ степенями свободы. Подставляем данные значения в формулы, получим:

$$n-1=15;\ \frac{\alpha}{2}=\frac{1-0,95}{2}=\frac{0,05}{2}=0,025;\ 1-\frac{\alpha}{2}=1-0,025=0,975;$$

По таблице находим

$$\chi^2_{\frac{\alpha}{2};\,n-1}=\chi^2_{0,025;15}=27,5;\ \chi^2_{1-\frac{\alpha}{2};\,n-1}=\chi^2_{0,975;15}=6,26.$$

Доверительный интервал для дисперсии имеет вид

$$\frac{15\cdot 1}{27,5}<\sigma^2<\frac{15\cdot 1}{6,26};$$

$$0,55<\sigma^2<2,4.$$

Тогда, искомый доверительный интервал для среднего квадратического отклонения имеет вид

$$0,74<\sigma<1,55.$$

Ответ. (0,74; 1,55).

10. За последние пять лет годовой рост цены акции A составлял в среднем 20 % со средним квадратическим отклонением (исправленным) 5 %. Построить доверительный интервал с вероятностью 95 % для цены акции в конце следующего года, если в начале года она равна 100 ден. ед.

Решение. Рассмотрим величины относительно прироста цены акции за год. Таким образом, мы имеем пять наблюдений $x_1,\ldots,x_5$, где x_i – это прирост цены акции в i-м году (в процентах). Необходимо оценить прирост цены акции в следующем, шестом году. Используем формулу (2.17) при $n=5$:

$$\bar{x} - s \cdot t_{\alpha,4}\sqrt{1+\frac{1}{5}} < x_6 < \bar{x} + s \cdot t_{\alpha,4}\sqrt{1+\frac{1}{5}}.$$

Значение $t_{\alpha,n-1} = t_{0,05;4} = 2,78$ находим из таблицы распределения Стьюдента. Выборочное среднее дано в условии $\bar{x} = 20$, исправленное среднее квадратическое отклонение также дано $s = 5$. Получаем

$$20 - 5 \cdot 2,78\sqrt{1,2} < x_6 < 20 + 5 \cdot 2,78\sqrt{1,2},$$

откуда $5 < x_6 < 35$. Таким образом, в следующем, шестом году цена акции возрастет от 5 % до 35 %. Следовательно, цена составит от 105 до 135 ден. ед.

Ответ. $(105; 135)$.

11. По результатам социологического обследования при опросе 1500 респондентов рейтинг президента составил 35 %. Найти границы, в которых с надежностью 95 % заключен рейтинг президента. Сколько респондентов надо опросить, чтобы с надежностью 0,99 гарантировать предельную ошибку социологического обследования не более 1 %?

Решение. Для решения будем использовать формулу (2.13), которая применяется при схеме испытаний Бернулли. Пусть событие A заключается в том, что один опрошенный человек поддерживает президента. Тогда мы получаем схему испытаний Бернулли, число повторений опыта $n = 1500$, среди которых событие A произошло 525 раз (35 % от 1500 опрошенных), p – вероятность того, что один случайно взятый человек поддерживает президента, т.е. число $p \cdot 100$ % как раз и составляет рейтинг президента. Доверительный интервал для p строим по формуле (2.13), где $\hat{p} = \frac{525}{1500} = 0,35$, $\hat{q} = 1 - 0,35 = 0,65$, $u_{0,05} = 1,96$. Получим

$$0,35 - \frac{1,96 \cdot \sqrt{0,35 \cdot 0,65}}{\sqrt{1500}} < p < 0,35 + \frac{1,96 \cdot \sqrt{0,35 \cdot 0,65}}{\sqrt{1500}},$$

$$0,338 < p < 0,362.$$

Таким образом, рейтинг президента, с вероятностью 0,95, составляет от 33,8 % до 36,2 %.

Теперь решим вторую часть задачи. Предельная ошибка вычисления находится по формуле

$$\delta = \frac{u_\alpha \sqrt{\hat{p} \cdot \hat{q}}}{\sqrt{n}}.$$

Из этой формулы выражаем число испытаний n, получим:

$$n = \frac{u_\alpha^2 \cdot \hat{p} \cdot \hat{q}}{\delta^2}.$$

Значение u_α ищем из условия $\Phi_0(u_\alpha) = \frac{\gamma}{2} = \frac{0{,}99}{2} = 0{,}495$, получим $u_{0{,}01} = 2{,}58$. Подставляем данные:

$$n = \frac{2{,}58^2 \cdot 0{,}35 \cdot 0{,}65}{0{,}01^2} = 15143{,}31, \text{ значит } n = 15144.$$

Ответ. $(0{,}338;\ 0{,}362)$, $n = 15144$.

12. Урожайность зерновых культур в России в 1992–2001 гг. представлена в таблице.

Год	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Урожайность, ц/га	18	17,1	15,3	13,1	14,9	17,8	12,9	14,4	15,6	19,4

Зная, что генеральное среднее значение $a = 16{,}3$, построить доверительный интервал для среднего квадратического отклонения σ с надежностью $\gamma = 0{,}95$.

Решение. Так как известно генеральное среднее a, для построения доверительного интервала воспользуемся формулой (2.18а).

Сначала вычислим наилучшую оценку дисперсии при известном среднем $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$.

В таблице представлены данные за десять лет, с 1992 по 2001 гг., поэтому $n = 10$. Значения x_i, $i = 1,...,10$, берем из второй строки таблицы.

Составим расчетную таблицу:

x_i	$x_i - a$	$(x_i - a)^2$
18	1,7	2,89
17,1	0,8	0,64
15,3	–1	1

13,1	$-3,2$	10,24
14,9	$-1,4$	1,96
17,8	1,5	2,25
12,9	$-3,4$	11,56
14,4	$-1,9$	3,61
15,6	$-0,7$	0,49
19,4	3,1	9,61
		$\sum_{i=1}^{10}(x_i-a)^2=44,25$

Получим $S_0^2=\frac{44,25}{10}=4,425$. Уровень значимости $\alpha=1-0,95=0,05$,

$$\frac{\alpha}{2}=0,025,\ 1-\frac{\alpha}{2}=0,975.$$

Значения $\chi^2_{\alpha/2,n}$ и $\chi^2_{1-\alpha/2,n}$ берем из таблицы критических точек распределения хи-квадрат: $\chi^2_{0,025;10}=20,5;\ \chi^2_{0,975;10}=3,25$. Подставляя найденные значения в формулу (2.19а), получим искомый доверительный интервал

$$\sqrt{\frac{10\cdot 4,425}{20,5}}<\sigma<\sqrt{\frac{10\cdot 4,425}{3,25}},$$

$$1,47<\sigma<3,69.$$

Ответ. (1,47; 3,69).

Задачи для самостоятельного решения

1. Данные выборки приведены в таблице. Найти выборочное среднее, выборочные дисперсии. Построить эмпирическую функцию распределения.

а)

x_i	-4	-2	0	3	4

n_i	3	7	2	6	2

б)

x_i	–6	–3	1	2	4
n_i	3	1	4	5	2

в)

x_i	–1	0	1	3	5
n_i	4	2	2	5	2

2. Найти выборочную среднюю по данному распределению выборки:

а)

x_i	2560	2600	2620	2650	2700
n_i	2	3	10	4	1

б)

x_i	0,03	0,05	0,06	0,07
n_i	20	50	18	12

3. Найти выборочную дисперсию $\overline{D}$ по данному распределению выборки:

а)

x_i	340	360	375	380
n_i	20	50	18	12

б)

x_i	0,1	0,5	0,6	0,8
n_i	5	15	20	10

в)

x_i	18,4	18,9	19,3	19,6
n_i	5	10	20	15

4. Найти исправленную выборочную дисперсию и среднее квадратическое отклонение по данному распределению выборки:

а)

x_i	0,1	0,5	0,7	0,9
n_i	6	12	1	1

б)

x_i	23,5	26,1	28,2	30,4
n_i	2	3	4	1

5. Найти несмещенные оценки математического ожидания и дисперсии по данным выборки:

а)

№ интервала	Интервал	n_i
1	–5–0	25
2	0–5	5
3	5–10	15
4	10–15	5
5	15–20	20

б)

№ интервала	Интервал	Середины интервалов	n_i
1	5,05–5,15	5,1	5
2	5,15–5,25	5,2	8
3	5,25–5,35	5,3	12
4	5,35–5,45	5,4	20
5	5,45–5,55	5,5	26
6	5,55–5,65	5,6	15

7	5,65–5,75	5,7	10
8	5,75–5,85	5,8	4

6. По выборке объема $n=51$ найдена смещенная оценка $\overline{D}=5$ генеральной дисперсии. Найти несмещенную оценку дисперсии генеральной совокупности.

7. При измерении веса 20 шоколадных батончиков (с номинальным весом 50 г) получены следующие значения в граммах:

49,1; 50; 49,7; 50,5; 48,1; 50,3; 49,7; 51,6; 49,8; 50,1;
49,7; 48,8; 51,4; 49,1; 49,6; 50,9; 48,5; 52; 50,7; 50,6.

Найти выборочное среднее, выборочные дисперсии, средние квадратические отклонения, выборочную медиану, крайние члены вариационного ряда.

8. По данным выборки составить статистическое распределение. Построить гистограмму частот, график накопленных частот. Вычислить несмещенные точечные оценки параметров генеральной совокупности.

2,8; 1,3; 3,1; 4; 1; 2,3; 5,1; 4,8; 4; 2,4;
3,2; 4; 6,1; 8; 10,5; 11; 3,1; 4,5; 9; 12;
15; 2,1; 3,2; 2,8; 9,6; 9,5; 3,6; 14; 15; 14;
2,9; 9,1; 9,8; 4,7; 7,1; 6,2; 8,1; 8,2; 8,6; 13;
10,3; 11; 7,5; 4,7; 10,8; 11,6; 5,9; 6,9; 3,1; 10.

9. Найти доверительный интервал для оценки с надежностью 0,99 неизвестного математического ожидания a нормально распределенного признака X генеральной совокупности, если известны генеральное среднее квадратическое отклонение σ, выборочное среднее $\overline{x}$ и объем выборки n:

а) $\sigma=4, \overline{x}=10,2, n=16;$

б) $\sigma=5, \overline{x}=16,8, n=25.$

10. Найти минимальный объем выборки, при которой с надежностью 0,975 точность оценки математического ожидания нормально распределенной генеральной совокупности по выборочной средней равна 0,3, если известно среднее квадратическое отклонение генеральной совокупности $\sigma=1,2.$

11. Из генеральной совокупности извлечена выборка:

x_i	–2	1	2	3	4	5
n_i	2	1	2	2	2	1

Оценить с надежностью 0,95 математическое ожидание a нормально распределенного признака генеральной совокупности с помощью доверительного интервала.

12. При измерении веса 20 шоколадных батончиков (с номинальным весом 50 г) получены следующие значения в граммах:

49,1; 50; 49,7; 50,5; 48,1; 50,3; 49,7; 51,6; 49,8; 50,1;
49,7; 48,8; 51,4; 49,1; 49,6; 50,9; 48,5; 52; 50,7; 50,6.

а) Найти доверительный интервал для среднего веса с надежностью 95 %.

б) Найти доверительный интервал для среднего квадратического отклонения с вероятностью 90 %.

13. По данным выборки объема n из генеральной совокупности нормально распределенного количественного признака найдено исправленное среднее квадратическое отклонение s. Найти доверительный интервал, покрывающий генеральное среднее квадратическое отклонение σ с надежностью 0,95, если:

а) $n=10$, $s=5{,}1$;

б) $n=30$, $s=14$.

14. По данным выборки объема $n=25$ из генеральной совокупности найдено исправленное среднее квадратическое отклонение $s=1$ нормально распределенного признака. Найти доверительный интервал, покрывающий генеральное среднее квадратическое отклонение с надежностью 0,9.

15. По выборке из 25 упаковок товара средний вес составил 101 г с исправленным средним квадратическим отклонением 3 г. Построить доверительные интервалы для среднего и дисперсии с вероятностью 90 %.

16. По данным таблицы из задания 5б) построить доверительные интервалы для среднего и для следующего наблюдения с надежностью 95 %, используя нормальное приближение.

17. За последние пять лет годовой рост цены акции A составил в среднем 20 % со средним квадратическим отклонением (исправленным) 10 %. Построить доверительный интервал с вероятностью 95 % для средней цены акции в конце следующего года, если в начале года она была равна 1000 ден. ед.

18. Проведена случайная выборка личных заемных счетов в банке. Из $n=1000$ отобранных счетов 60 оказались с задолженностью по возврату ссуды сроком до трех месяцев. Найти доверительный интервал с вероятностью 90 % для числа счетов в генеральной совокупности, которые имеют задолженность до трех месяцев, если банк насчитывает 30000 личных заемных счетов.

19. По данным социологического опроса, среди 100 человек 20 % пользуются стиральным порошком фирмы A. Сколько еще людей следует

опросить, чтобы с вероятностью 99 % получить результат с точностью до 1 %?

20. В ходе аудиторской проверки фирмы была проведена случайная выборка записей по счетам. Из выборки $n=500$ записей 10 содержали некоторые ошибки в самой записи или в процедуре. Найти доверительный интервал для доли ошибок во всей генеральной совокупности с вероятностью 95 %. Определить объем выборки, которую следует произвести аудитору, если он хочет определить с точностью 0,005 генеральную долю с доверительной вероятностью 95 %.

Ответы.

1. а) $\bar{x}=0;\ \bar{D}=8{,}1;\ S^2=8{,}53$; **б)** $\bar{x}=0{,}067;\ \bar{D}=11{,}53;\ S^2=12{,}35$; **в)** $\bar{x}=1{,}53;\ \bar{D}=4{,}38;\ S^2=4{,}69$; **2. а)** 2621; **б)** 0,0502; **3. а)** 167,29; **б)** 0,0344; **в)** 0,1336; **4. а)** $S^2=0{,}0525;\ s=0{,}229$; **б)** $S^2=4{,}89;\ s=2{,}211$; **5. а)** $\bar{x}=6{,}7;\ S^2=68{,}33$; **б)** $\bar{x}=5{,}459;\ S^2=0{,}0297$; **6.** 5,1; **7.** $\bar{x}\approx 50;\ \bar{D}\approx 0{,}95;\ S^2\approx 1;\ \hat{\sigma}\approx 0{,}97;\ s=1;\ x_{\text{med}}=50{,}1;\ x_{\min}=48{,}1;\ x_{\max}=52$; **8.** [1,15]; $k=7;\ h=2;\ \bar{x}=7{,}08;\ S^2=15{,}38$;

№	Интервал	n_i	n_i/h	w_i	w_i^c	w_i^c/h	x_i
1	[1,3)	8	4	0,16	0,16	0,08	2
2	[3,5)	13	6,5	0,26	0,42	0,21	4
3	[5,7)	5	2,5	0,1	0,52	0,26	6
4	[7,9)	6	3	0,12	0,64	0,32	8
5	[9,11)	9	4,5	0,18	0,82	0,41	10
6	[11,13)	4	2	0,08	0,9	0,45	12
7	[13,15)	5	2,5	0,1	1	0,5	14

9. а) (7,63; 12,77); **б)** (14,23; 19,37); **10.** 81; **11.** (0,28; 3,72); **12. а)** (49,53; 50,47); **б)** (0,79; 1,37); **13. а)** (3,51; 9,31); **б)** (11,24; 18,85); **14.** (0,81; 1,32); **15.** $99{,}97<a<102{,}03;\ 5{,}95<\sigma^2<15{,}65$; **16.** $5{,}43<a<5{,}49,\ 5{,}12<x_{101}<5{,}8$; **17.** $895<A<1505$; **18.** $1440<K<2160$; **19.** $n=10651$, еще необходимо опросить 10551; **20.** $0{,}0072<p<0{,}0378;\ n=3012$.

Контрольные вопросы

1. Дайте определение оценки параметра распределения.

2. Какая оценка называется эффективной, несмещенной, состоятельной?

3. По каким формулам можно вычислить выборочное среднее?

4. Как определяются смещенная и несмещенная оценки дисперсии? Какой формулой эти выборочные характеристики связаны между собой?

5. Как определяются выборочные мода и медиана?

6. Дайте определение квантили уровня p.

7. Какими равенствами задаются верхняя и нижняя критические границы, соответствующие уровню значимости α?

8. Дайте определение доверительного интервала.

9. Что называется надежностью доверительного интервала?

10.По каким формулам вычисляется точность интервальной оценки для математического ожидания генеральной совокупности?

11. Критические точки какого распределения используются при построении доверительного интервала для математического ожидания генеральной совокупности, при неизвестной дисперсии?

12. Критические точки какого распределения используются при построении доверительного интервала для дисперсии генеральной совокупности?

Глава 3

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Статистическая проверка гипотез

Статистической гипотезой называется любое предположение относительно генеральной совокупности. Статистическая гипотеза называется *параметрической*, если в ней сформулированы предположения относительно значений параметров функции распределения, когда сам закон считается известным. Статистическая гипотеза называется *непараметрической*, если в ней сформулированы предположения относительно вида самой функции распределения, т.е. о виде распределения.

Метод использования выборки для проверки истинности (ложности) статистической гипотезы называется *статистическим доказательством* истинности выдвинутой гипотезы. Наряду с выдвинутой гипотезой рассматривают одну или несколько альтернативных гипотез. Если выдвинутая гипотеза отвергается, то вместо нее принимается альтернативная гипотеза. Поэтому статистические гипотезы подразделяются на нулевые и альтернативные.

Нулевой (*основной*) называют выдвинутую гипотезу H_0. *Конкурирующей* (*альтернативной*) называют гипотезу H_1, которая противоречит основной.

Различают гипотезы, которые содержат только одно или более одного предположений.

Простой называют гипотезу, содержащую только одно предположение. Например, если $a-$ математическое ожидание нормального распределения, то гипотеза $H_0 : a = 8$ – простая. *Сложной* называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез. Например, сложная гипотеза $H_0 : a > 8$ состоит из бесконечного множества простых гипотез вида $H_i : a = b_i$, где $b_i -$ любое число, большее 8.

Для проверки нулевой гипотезы используют специально подобранную случайную величину – выборочную статистику, точное или приближенное распределение которой известны.

Случайная величина K, построенная по результатам наблюдений для проверки нулевой гипотезы, называется *статистическим критерием*.

Схема построения критерия такова: все выборочное пространство делится на две взаимодополняющие области – область S отклонения основной гипотезы и область $\overline{S}$ принятия этой гипотезы. Область S, при попадании в которую выборочной точки основная гипотеза отвергается, называется критической.

В итоге статистической проверки гипотезы в двух случаях может быть принято неправильное решение, т.е. могут быть допущены ошибки двух видов.

Ошибка первого рода состоит в том, что отвергается основная гипотеза, хотя на самом деле она верна. Вероятность ошибки первого рода обозначают $P(S/H_0)=\alpha$.

Ошибка второго рода состоит в том, что будет принята неправильная гипотеза. Вероятность ошибки второго рода обозначают $P(\overline{S}/H_1)=\beta$.

Вероятность α называют также *уровнем значимости* (*размером*) критерия. Вероятность $1-\beta$ не совершить ошибку второго рода называют *мощностью* критерия.

Наблюдаемым значением $K_{\text{набл}}$ называют значение критерия, вычисленное по выборкам.

Поскольку критерий K – одномерная случайная величина, все ее возможные значения принадлежат некоторому интервалу. Поэтому критическая область и область принятия гипотезы являются интервалами и существуют точки, которые их разделяют.

Критическими точками (*границами*) $k_{\text{кр}}$ называют точки, отделяющие критическую область от области принятия гипотезы.

Различают одностороннюю (правостороннюю и левостороннюю) и двустороннюю критические области.

Правосторонней называют критическую область, определяемую неравенством $K>k_{\text{кр}}$, где $k_{\text{кр}}-$ положительное число. *Левосторонней* называют критическую область, определяемую неравенством $K<k_{\text{кр}}$, где $k_{\text{кр}}-$ отрицательное число. *Односторонней* называют правостороннюю или левостороннюю области.

Двусторонней называют критическую область, определяемую неравенствами $K<k_1, K>k_2$, где $k_2>k_1$. В частности, если критические точки симметричны относительно нуля, двусторонняя критическая область определяется неравенствами

$$K<-k_{\text{кр}}, K>k_{\text{кр}}$$

или равносильным неравенством

$$|K|>k_{\text{кр}}.$$

Критическую область следует строить так, чтобы мощность критерия была максимальной. В этом случае мы получим минимальную ошибку второго рода.

Проверка гипотез для одной выборки

Рассмотрим простые методы проверки параметрических гипотез в случае нормального распределения (которые являются формально точными), а также гипотезы о вероятности «успеха» в испытаниях Бернулли (на основе асимптотической нормальности). Следующие три типа гипотез проверяются для нормальных данных: $X \in N_{a,\sigma}$.

Гипотезы о неизвестном среднем a

Дисперсия генеральной совокупности σ^2 известна

Пусть генеральная совокупность распределена нормально, причем генеральная средняя a, хотя и неизвестна, но имеются основания полагать, что она равна гипотетическому значению a_0. Предположим, что дисперсия генеральной совокупности известна.

Итак, из нормальной генеральной совокупности извлечена выборка объема n и по ней найдена выборочная средняя $\bar{x}$, причем генеральная дисперсия σ^2 известна. Требуется по выборочной средней, при заданном уровне значимости α, проверить нулевую гипотезу $H_0 : a = a_0$ о равенстве генеральной средней a гипотетическому значению a_0.

В качестве критерия проверки нулевой гипотезы принимаем случайную величину

$$U = \frac{\bar{x} - a_0}{\sigma(\bar{x})} = \frac{\sqrt{n}\left(\bar{x} - a_0\right)}{\sigma},$$

которая распределена нормально, причем, при справедливости нулевой гипотезы, $M(U) = 0, \sigma(U) = 1.$

Критическая область также строится в зависимости от нулевой гипотезы. По данным наблюдений вычисляем значения критерия

$$U_{\text{набл}} = \frac{\sqrt{n}\left(\bar{x} - a_0\right)}{\sigma}. \qquad (3.1)$$

В первом случае, когда конкурирующая гипотеза имеет вид

$$H_1 : a \neq a_0,$$

строим двустороннюю критическую область и по таблице функции Лапласа (прил. 1) находим критическую точку по равенству

$$\Phi_{0,1}\left(u_{\text{кр}}\right)=\frac{1-\alpha}{2}.$$

Если $\left|U_{\text{набл}}\right|<u_{\text{кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $\left|U_{\text{набл}}\right|>u_{\text{кр}}$ – нулевую гипотезу отвергают.

Второй случай. При конкурирующей гипотезе $H_1:a>a_0$ получаем правостороннюю критическую область. Критическую точку находят по равенству

$$\Phi_{0,1}\left(u_{\text{кр}}\right)=\frac{1-2\alpha}{2}.$$

Если $U_{\text{набл}}<u_{\text{кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $U_{\text{набл}}>u_{\text{кр}}$ – нулевую гипотезу отвергают.

И, наконец, *в третьем случае*, при конкурирующей гипотезе $H_1:a<a_0$ строят левостороннюю критическую область. Ищется критическая точка по равенству

$$\Phi_{0,1}\left(u_{\text{кр}}\right)=\frac{1-2\alpha}{2}.$$

Если $U_{\text{набл}}>-u_{\text{кр}}$, тогда нет оснований отвергнуть нулевую гипотезу.

Если $U_{\text{набл}}<-u_{\text{кр}}$, тогда нулевую гипотезу отвергают.

Дисперсия генеральной совокупности σ^2 неизвестна

В этом случае в качестве критерия проверки нулевой гипотезы принимаем случайную величину

$$T=\frac{\sqrt{n}\left(\overline{x}-a_0\right)}{s},$$

где $s-$ исправленное выборочное среднее квадратическое отклонение. Величина T имеет распределение Стьюдента с $k=n-1$ степенями свободы.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай. Конкурирующая гипотеза

$$H_1:a\neq a_0.$$

Вычисляем наблюдаемое значение признака по данным выборки

$$T_{\text{набл}} = \frac{\sqrt{n}\left(\bar{x} - a_0\right)}{s} \tag{3.2}$$

и по таблице критических точек распределения Стьюдента (прил. 3), по заданному уровню значимости α, помещенному в верхней строке таблицы, и числу степеней свободы $k = n - 1$, находим критическую точку $t_{\text{двуст.кр}} = t_{\alpha; k}$.

Если $|T_{\text{набл}}| < t_{\text{двуст.кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $|T_{\text{набл}}| > t_{\text{двуст.кр}}$ – нулевую гипотезу отвергают.

Второй случай. При конкурирующей гипотезе

$$H_1 : a < a_0,$$

по уровню значимости α, помещенному в нижней строке таблицы, и числу степеней свободы $k = n - 1$, находим критическую точку $t_{\text{правост.кр}} = t_{\alpha; k}$ правосторонней критической области.

Если $T_{\text{набл}} < t_{\text{правост.кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $T_{\text{набл}} > t_{\text{правост.кр}}$ – нулевую гипотезу отвергают.

Третий случай. При конкурирующей гипотезе

$$H_1 : a < a_0$$

находят критическую точку $t_{\text{левост.кр}} = -t_{\text{правост.кр}}$ левосторонней критической области.

Если $T_{\text{набл}} > t_{\text{левост.кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $T_{\text{набл}} < t_{\text{левост.кр}}$ – нулевую гипотезу отвергают.

Пример

Менеджер кредитного отдела нефтяной компании хотел бы выяснить, является ли среднемесячный баланс владельцев кредитных карточек равным 75 у. е. Аудитор случайным образом отобрал 100 счетов и нашел, что среднемесячный баланс владельцев составил 83,4 у. е. с выборочным исправленным отклонением, равным 23,65 у. е. Определить на 5 % уровне значимости, может ли этот аудитор утверждать, что средний баланс отличен от 75 у. е.

Решение. Исходя из условия задачи, сформулируем нулевую и конкурирующую гипотезы:

$$H_0 : a = 75,\ \ H_1 : a \neq 75.$$

Причем дисперсия генеральной совокупности σ^2 неизвестна. Уровень значимости $\alpha = 0,05$. Для проверки нулевой гипотезы применим критерий T, имеющий распределение Стьюдента. Вычислим наблюдаемое значение критерия по формуле (3.2)

$$T_{\text{набл}} = \frac{\overline{x} - a_0}{s}\sqrt{n-1} = \frac{83,4-75}{23,65}\sqrt{100-1} = \frac{8,4}{23,65}\sqrt{99} \approx 3,53.$$

Далее, по таблице критических точек распределения Стьюдента для двусторонней критической области, находим критическое значение

$$t_{\text{кр}} = t_{0,05;99} \approx 2.$$

Так как $|T_{\text{набл}}| > t_{\text{кр}}$, то нулевая гипотеза отвергается и принимается конкурирующая гипотеза $H_1 : a \neq 75$, т.е. среднемесячный баланс владельцев отличен от 75 у. е.

Гипотезы о неизвестной дисперсии

Пусть генеральная совокупность распределена нормально, причем генеральная дисперсия, хотя и неизвестна, но имеются основания предполагать, что она равна гипотетическому значению σ_0^2.

Пусть из генеральной совокупности извлечена выборка объема n и по ней найдена исправленная выборочная дисперсия S^2. Требуется по исправленной дисперсии, при заданном уровне значимости, проверить нулевую гипотезу, состоящую в том, что генеральная дисперсия σ^2 равна гипотетическому значению σ_0^2. Таким образом, нулевая гипотеза имеет вид

$$H_0 : \sigma^2 = \sigma_0^2.$$

В качестве критерия проверки нулевой гипотезы примем случайную величину $\dfrac{(n-1)S^2}{\sigma_0^2}$. Эта величина случайная, потому что в разных опытах S^2 будет принимать различные, наперед неизвестные значения. Эта величина имеет распределение χ^2 с $k = n-1$ степенями свободы, поэтому будем обозначать ее

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}.$$

Критическая область строится в зависимости от конкурирующей гипотезы.

Первый случай, когда конкурирующая гипотеза имеет вид

$$H_1 : \sigma^2 > \sigma_0^2.$$

В этом случае строят правостороннюю критическую область. Критическую точку $\chi^2_{\text{кр}} = \chi^2_{\alpha;k}$ находят по таблице критических точек распределения χ^2 (прил. 2), и тогда правосторонняя критическая область определяется неравенством

$$\chi^2 > \chi^2_{\text{кр}},$$

а область принятия нулевой гипотезы неравенством

$$\chi^2 < \chi^2_{\text{кр}}.$$

Обозначим значение критерия, вычисленное по данным наблюдений, через $\chi^2_{\text{набл}}$. Сформулируем правило проверки нулевой гипотезы.

Для того, чтобы, при заданном уровне значимости, проверить нулевую гипотезу $H_0 : \sigma^2 = \sigma_0^2$ при конкурирующей гипотезе $H_1 : \sigma^2 > \sigma_0^2$, надо вычислить

$$\chi^2_{\text{набл}} = \frac{(n-1)S^2}{\sigma_0^2} \qquad (3.3)$$

и по таблице критических точек распределения χ^2, по заданному уровню значимости α и числу степеней свободы $k = n-1$, найти критическую точку $\chi^2_{\text{кр}} = \chi^2_{\alpha;k}$.

Если $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$ – нулевую гипотезу отвергают.

Второй случай, когда конкурирующая гипотеза имеет вид:

$$H_1 : \sigma^2 < \sigma_0^2.$$

В этом случае находят критическую точку $\chi^2_{\text{кр}} = \chi^2_{1-\alpha;k}$, т.е. в качестве уровня значимости берут величину $1-\alpha$, где $\alpha-$ заданный уровень значимости, $k = n-1$ – это число степеней свободы. Критическая точка также ищется по таблице критических точек распределения χ^2. Наблюдаемое значение $\chi^2_{\text{набл}}$ ищется по той же формуле, что и в первом случае.

Если $\chi^2_{\text{набл}} > \chi^2_{\text{кр}} -$ нет оснований отвергнуть нулевую гипотезу.

Если $\chi^2_{\text{набл}} < \chi^2_{\text{кр}} -$ нулевую гипотезу отвергают.

Таким образом, во втором случае получаем левостороннюю критическую область.

Третий случай: конкурирующая гипотеза имеет вид

$$H_1 : \sigma^2 \neq \sigma_0^2.$$

В этом случае строят двустороннюю критическую область.

Для того, чтобы при заданном уровне значимости α проверить нулевую гипотезу о равенстве неизвестной генеральной дисперсии σ^2 нормальной совокупности гипотетическому значению σ_0^2, при конкурирующей гипотезе $H_1 : \sigma^2 \neq \sigma_0^2$, надо вычислить наблюдаемое значение критерия $\chi^2_{\text{набл}} = \dfrac{(n-1)S^2}{\sigma_0^2}$ и по таблице критических точек распределения χ^2 найти левую критическую точку $\chi^2_{\text{лев.кр}} = \chi^2_{1-\alpha/2;k}$ и правую критическую точку $\chi^2_{\text{пр.кр}} = \chi^2_{\alpha/2;k}$, где $k = n-1$ – число степеней свободы, $(1-\alpha/2), \alpha/2 -$ соответствующие уровни значимости.

Если $\chi^2_{\text{лев.кр}} < \chi^2_{\text{набл}} < \chi^2_{\text{прав.кр}} -$ нет оснований отвергнуть нулевую гипотезу.

Если $\chi^2_{\text{набл}} < \chi^2_{\text{лев.кр}}$ или $\chi^2_{\text{набл}} > \chi^2_{\text{прав.кр}}$, то нулевую гипотезу отвергают.

Гипотеза о неизвестной вероятности «успеха» в испытаниях Бернулли

Основная гипотеза $H_0 : p = p_0$, альтернативная гипотеза H_1 может быть трех видов: а) $p \neq p_0$; б) $p > p_0$; в) $p < p_0$. Во всех трех случаях для проверки используется статистика критерия

$$U_{\text{набл}} = \frac{w - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}, \qquad (3.4)$$

где $w-$ относительная частота «успехов» в n наблюдениях.

Далее критические точки и области для проверки выбираются так же, как при проверке гипотезы о неизвестном среднем при известной дисперсии.

Замечание. Этим методом можно пользоваться только при больших объемах выборки (порядка сотен).

Пример. Партия изделий принимается, если доля брака составляет не более 2 %. Среди случайно отобранных 500 изделий оказалось 13 бракованных. Следует ли принять партию (на уровне значимости 0,05)?

Решение. Из условия задачи следует, что необходимо проверить гипотезу $H_0 : p = p_0$ при конкурирующей гипотезе

$$H_1 : p = p_1 > p_0,$$

где $p_0 = 0,02$ (2 %). Относительная частота брака составляет

$$w = \frac{13}{500} = 0,026.$$

Найдем наблюдаемое значение критерия:

$$U_{\text{набл}} = \frac{(0,026 - 0,02)}{\sqrt{0,02 \cdot 0,98}}\sqrt{500} \approx 0,96.$$

Из соотношения $\Phi_0(u_{\text{кр}}) = \frac{1}{2} - \alpha = \frac{1}{2} - 0,05 = 0,45$ находим $u_{\text{кр}} = 1,65$ и получаем $U_{\text{набл}} < u_{\text{кр}}$, так что основная гипотеза принимается. Таким образом, партию изделий можно принять.

Описанные выше критерии проверки гипотез можно представить в виде табл. 4.1.

Проверка гипотез для двух выборок

Пусть имеются две независимые выборки $x_1, x_2, ..., x_n$ и $y_1, y_2, ..., y_m$, имеющие нормальное распределение с параметрами (a_x, σ_x) и (a_y, σ_y) соответственно. Обычно строится задача проверки их однородности, т.е. равенства обоих параметров, либо необходимо проверить равенство параметров по отдельности.

Гипотеза о равенстве дисперсий двух выборок

На практике задача сравнения дисперсий возникает, если требуется сравнить точность приборов, методов измерений и т.д.

Пусть генеральные совокупности X и Y распределены нормально. По независимым выборкам объемов n_1 и n_2, извлеченным из этих совокупностей, найдены исправленные выборочные дисперсии S_x^2 и S_y^2. Требуется по исправленным дисперсиям, при заданном уровне значимости α, проверить нулевую гипотезу, состоящую в том, что генеральные дисперсии рассматриваемых совокупностей равны между собой, т.е.:

$$H_0 : \sigma_x^2 = \sigma_y^2.$$

В качестве критерия проверки нулевой гипотезы о равенстве генеральных дисперсий, примем отношение большей исправленной дисперсии к меньшей, т.е. случайную величину:

$$F = \frac{S_б^2}{S_м^2}.$$

Эта величина, при условии справедливости нулевой гипотезы, имеет распределение Фишера со степенями свободы $k_1 = n_1 - 1$, $k_2 = n_2 - 1$, где $n_1 -$ объем выборки, по которой вычислена большая исправленная дисперсия, $n_2 -$ объем выборки, по которой вычислена меньшая дисперсия.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай. Конкурирующая гипотеза имеет вид

$$H_1 : \sigma_x^2 > \sigma_y^2.$$

В этом случае строят одностороннюю, а именно правостороннюю, критическую область. Обозначим отношение большей исправленной дисперсии к меньшей, вычисленное по данным выборок, через $F_{\text{набл}}$. Получим следующее правило проверки нулевой гипотезы: для того, чтобы, при заданном уровне значимости, проверить нулевую гипотезу $H_0 : \sigma_x^2 = \sigma_y^2$ при конкурирующей гипотезе $H_1 : \sigma_x^2 > \sigma_y^2$, надо вычислить

$$F_{\text{набл}} = \frac{S_б^2}{S_м^2} \qquad (3.5)$$

Таблица 4.1 Алгоритм проверки гипотез о значениях параметров нормального распределения и вероятности успеха

H_0	Предположение	Статистика критерия	H_1	Область принятия H_0
$a = a_0$	σ^2 известно	$U = \dfrac{\bar{x} - a_0}{\sigma}\sqrt{n}$	$a = a_1 > a_0$	$U < u_{\text{кр}},\ \Phi_0\left(u_{\text{кр}}\right) = \dfrac{1}{2} - \alpha$
			$a = a_1 < a_0$	$U > -u_{\text{кр}},\ \Phi_0\left(u_{\text{кр}}\right) = \dfrac{1}{2} - \alpha$
			$a = a_1 \neq a_0$	$\lvert U\rvert < u_{\text{кр}},\ \Phi_0\left(u_{\text{кр}}\right) = \dfrac{1-\alpha}{2}$
$a = a_0$	σ^2 неизвестно	$T = \dfrac{\bar{x} - a_0}{s}\sqrt{n}$	$a = a_1 > a_0$	$T < t_{\alpha, n-1}$ (для односторонней области)
			$a = a_1 < a_0$	$T > -t_{\alpha, n-1}$ (для односторонней области)
			$a = a_1 \neq a_0$	$\lvert T\rvert < t_{\alpha/2, n-1}$ (для двусторонней области)
$\sigma^2 = \sigma_0^2$	a неизвестно	$\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 > \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha, n-1}$
			$\sigma^2 < \sigma_0^2$	$\chi^2 > \chi^2_{\alpha, n-1}$
			$\sigma^2 \neq \sigma_0^2$	$\chi^2_{\alpha/2, n-1} < \chi^2 < \chi^2_{1-\alpha/2, n-1}$
$p = p_0$	$N -$ порядка нескольких десятков $np_0 > 5$, $n(1-p_0) > 5$	$U = \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0}}\sqrt{n}$, $\hat{p} = \dfrac{k}{n}$, $q_0 = 1 - p_0$	$p = p_1 > p_0$	$U < u_{\text{кр}},\ \Phi_0\left(u_{\text{кр}}\right) = \dfrac{1}{2} - \alpha$
			$p = p_1 < p_0$	$U > -u_{\text{кр}},\ \Phi_0\left(u_{\text{кр}}\right) = \dfrac{1}{2} - \alpha$
			$p = p_1 \neq p_0$	$\lvert U\rvert < u_{\text{кр}},\ \Phi_0\left(u_{\text{кр}}\right) = \dfrac{1-\alpha}{2}$

и по таблице критических точек распределения Фишера (прил. 4), по заданному уровню значимости α и числам степеней свободы k_1 и k_2, найти критическую точку $F_{\text{кр}} = F_{\alpha;k_1;k_2}$.

Если $F_{\text{набл}} < F_{\text{кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $F_{\text{набл}} > F_{\text{кр}}$ – нулевую гипотезу отвергают.

Пример 1.

По двум независимым выборкам объемов $n_1 = 12$, $n_2 = 15$, извлеченным из нормальных генеральных совокупностей X и Y, найдены исправленные выборочные дисперсии $S_x^2 = 11,41$, $S_y^2 = 6,52$. При уровне значимости $\alpha = 0,05$ проверить нулевую гипотезу

$$H_0 : \sigma_x^2 = \sigma_y^2$$

о равенстве генеральных дисперсий, при конкурирующей гипотезе

$$H_1 : \sigma_x^2 > \sigma_y^2.$$

Решение. Найдем отношение большей исправленной дисперсии к меньшей:

$$F_{\text{набл}} = \frac{11,41}{6,52} = 1,75.$$

Как было отмечено выше, при данной конкурирующей гипотезе мы получаем правостороннюю критическую область.

По таблице критических точек распределения Фишера, по уровню значимости $\alpha = 0,05$ и числам степеней свободы $k_1 = 12 - 1 = 11$ и $k_2 = 15 - 1 = 14$, находим критическую точку $F_{\text{кр}} = F_{0,05;11;14} = 2,57$.

Так как $F_{\text{набл}} < F_{\text{кр}}$, то нет оснований отвергнуть нулевую гипотезу о равенстве генеральных дисперсий.

Замечание. Для случая, когда конкурирующая гипотеза имеет вид $H_1 : \sigma_x^2 < \sigma_y^2$, критическая область строится так же, как и в первом случае.

Второй случай. Конкурирующая гипотеза имеет вид

$$H_1 : \sigma_x^2 \neq \sigma_y^2.$$

В этом случае строят двустороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область, в

предположении справедливости нулевой гипотезы, была равна принятому уровню значимости α.

Обозначим отношение большей исправленной дисперсии к меньшей, вычисленное по данным выборок, через $F_{\text{набл}}$. Получим следующее правило проверки нулевой гипотезы: для того, чтобы, при заданном уровне значимости, проверить нулевую гипотезу $H_0 : \sigma_x^2 = \sigma_y^2$ при конкурирующей гипотезе $H_1 : \sigma_x^2 \neq \sigma_y^2$, надо вычислить

$$F_{\text{набл}} = \frac{S_{\text{б}}^2}{S_{\text{м}}^2}$$

и по таблице критических точек распределения Фишера, по уровню значимости $\frac{\alpha}{2}$ и числам степеней свободы k_1 и k_2, найти критическую точку $F_{\text{кр}} = F_{\alpha/2;k_1;k_2}$.

Если $F_{\text{набл}} < F_{\text{кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $F_{\text{набл}} > F_{\text{кр}}$ – нулевую гипотезу отвергают.

Пример 2

По двум независимым выборкам объемов $n_1 = 10$, $n_2 = 18$, извлеченным из генеральных совокупностей X и Y, найдены исправленные выборочные дисперсии $S_x^2 = 1,23$ и $S_y^2 = 0,41$. При уровне значимости $\alpha = 0,1$ проверить нулевую гипотезу о равенстве генеральных дисперсий при конкурирующей гипотезе $H_1 : \sigma_x^2 \neq \sigma_y^2$.

Решение. Найдем отношение большей исправленной дисперсии к меньшей:

$$F_{\text{набл}} = \frac{1,23}{0,41} = 3.$$

По таблице критических точек распределения Фишера, по уровню значимости, вдвое меньшим заданного, т.е. при $\alpha/2 = 0,05$, и числам степеней свободы $k_1 = 10 - 1 = 9$, $k_2 = 18 - 1 = 17$, находим критическую точку $F_{\text{кр}} = F_{0,05;9;17} = 2,5$.

Так как $F_{\text{набл}} > F_{\text{кр}}$, нулевую гипотезу о равенстве генеральных дисперсий отвергаем. Другими словами, выборочные исправленные дисперсии различаются значимо. Например, если бы рассматриваемые дисперсии характеризовали точность двух методов измерений, то следует предпочесть тот метод, который имеет меньшую дисперсию (0,41).

Гипотеза о равенстве средних при известных дисперсиях

Пусть генеральные совокупности X и Y распределены нормально, причем их дисперсии известны. По независимым выборкам объемов n и m, извлеченным из этих совокупностей, найдены выборочные средние $\bar{x}, \bar{y}$.

Требуется по выборочным средним, при заданном уровне значимости α, проверить нулевую гипотезу, состоящую в том, что генеральные средние (математические ожидания) рассматриваемых совокупностей равны между собой, т.е.

$$H_0 : a_x = a_y.$$

В качестве критерия проверки нулевой гипотезы принимают случайную величину

$$Z = \frac{\bar{x} - \bar{y}}{\sigma\left(\bar{x} - \bar{y}\right)}.$$

По определению среднего квадратического отклонения $\sigma\left(\bar{x} - \bar{y}\right) = \sqrt{D\left(\bar{x} - \bar{y}\right)}$.

Учитывая свойства дисперсии, получаем $D\left(\bar{x} - \bar{y}\right) = D\left(\bar{x}\right) + D\left(\bar{y}\right)$. Ранее, при доказательстве смещенности оценки $\overline{D}$, было получено следующее равенство: $D\left(\bar{x}\right) = \frac{D\left(x\right)}{n} = \frac{\sigma_x^2}{n}$. Аналогично, имеем $D\left(\bar{y}\right) = \frac{D\left(y\right)}{m} = \frac{\sigma_y^2}{m}$. Таким образом, критерий Z можно вычислить по следующей формуле:

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}.$$

Полученный критерий – нормированная нормальная величина, т.е. случайная величина Z имеет стандартное нормальное распределение: $M(Z) = 0$, при справедливости нулевой гипотезы, $\sigma(Z) = 1$, так как выборки независимы.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай, когда конкурирующая гипотеза имеет вид

$$H_1 : a_x \neq a_y .$$

В этом случае строят двустороннюю критическую область. Для проверки нулевой гипотезы о равенстве генеральных средних при известных дисперсиях надо вычислить наблюдаемое значение критерия

$$Z_{\text{набл}} = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \qquad (3.6)$$

и по таблице функции Лапласа найти критическую точку по равенству $\Phi_{0,1}\left(z_{\text{кр}}\right) = \frac{1-\alpha}{2}$.

Если $\left|Z_{\text{набл}}\right| < z_{\text{кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $\left|Z_{\text{набл}}\right| > z_{\text{кр}}$ – нулевую гипотезу отвергают.

Второй случай. Конкурирующая гипотеза имеет вид

$$H_1 : a_x > a_y .$$

В этом случае строят правостороннюю критическую область. Для этого надо вычислить наблюдаемое значение критерия $Z_{\text{набл}}$ по формуле (3.6) и по таблице функции Лапласа найти критическую точку из равенства

$$\Phi_{0,1}\left(z_{\text{кр}}\right) = \frac{1-2\alpha}{2}.$$

Если $Z_{\text{набл}} < z_{\text{кр}}$ – нет оснований отвергнуть нулевую гипотезу. Если $Z_{\text{набл}} > z_{\text{кр}}$ – нулевую гипотезу отвергают.

Третий случай, когда конкурирующая гипотеза имеет вид:

$$H_1 : a_x < a_y .$$

В этом случае строят левостороннюю критическую область. При конкурирующей гипотезе такого вида надо вычислить наблюдаемое значение $Z_{\text{набл}}$ и по таблице функции Лапласа найти критическую точку по равенству

$$\Phi_{0,1}\left(z_{\text{кр}}\right) = \frac{1-2\alpha}{2}.$$

Если $Z_{\text{набл}} > -z_{\text{кр}}$, тогда нет оснований отвергнуть нулевую гипотезу. Если $Z_{\text{набл}} < -z_{\text{кр}}$, тогда нулевую гипотезу отвергают.

Примеры

1. По двум независимым выборкам объемов $n = 60$, $m = 50$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x} = 1250$, $\bar{y} = 1275$. Генеральные дисперсии известны: $\sigma_x^2 = 120$, $\sigma_y^2 = 100$. При уровне значимости $\alpha = 0,01$ проверить нулевую гипотезу $H_0 : a_x = a_y$ при конкурирующей гипотезе $H_1 : a_x \neq a_y$.

Решение. Найдем наблюдаемое значение критерия по формуле (3.6)

$$Z_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} = \frac{1250 - 1275}{\sqrt{\frac{120}{60} + \frac{100}{50}}} = -12,5.$$

Критическая область двусторонняя. Найдем правую критическую точку по равенству

$$\Phi_{0,1}\left(z_{\text{кр}}\right) = \frac{1 - \alpha}{2} = \frac{1 - 0,01}{2} = 0,495.$$

По таблице функции Лапласа находим $z_{\text{кр}} = 2,58$. Так как $\left|Z_{\text{набл}}\right| < z_{\text{кр}}$ – нулевую гипотезу отвергаем.

2. По двум независимым выборкам объемов $n = 50$, $m = 50$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x} = 142$, $\bar{y} = 150$. Генеральные дисперсии известны: $\sigma_x^2 = 28,2$, $\sigma_y^2 = 22,8$. При уровне значимости $\alpha = 0,01$, проверить нулевую гипотезу $H_0 : a_x = a_y$ при конкурирующей гипотезе $H_1 : a_x < a_y$.

Решение. Подставив данные задачи в формулу (3.6), получим наблюдаемое значение критерия $Z_{\text{набл}} = -8$. Критическая область левосторонняя. Найдем левую критическую точку по равенству

$$\Phi_{0,1}\left(z_{\text{кр}}\right) = \frac{1 - 2\alpha}{2} = \frac{1 - 2 \cdot 0,01}{2} = 0,49.$$

По таблице функции Лапласа находим $z_{\text{кр}} = 2,33$. Так как $Z_{\text{набл}} < -z_{\text{кр}}$ $\left(-8 < -2,33\right)$ – нулевую гипотезу отвергаем.

Гипотеза о равенстве средних при неизвестных одинаковых дисперсиях

Пусть генеральные совокупности X и Y распределены нормально, причем их дисперсии неизвестны. В предположении, что генеральные дисперсии одинаковы, необходимо проверить нулевую гипотезу $H_0 : a_x = a_y$.

В качестве критерия проверки нулевой гипотезы берем случайную величину

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)S_x^2 + (m-1)S_y^2}} \sqrt{\frac{nm \cdot (n+m-2)}{n+m}}.$$

Эта величина, при справедливости нулевой гипотезы, имеет распределение Стьюдента с $k = n + m - 2$ степенями свободы, где $n-$ объем первой выборки, а $m-$ объем второй выборки, $S_x^2, S_y^2 -$ исправленные выборочные дисперсии.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай, когда конкурирующая гипотеза имеет вид

$$H_1 : a_x \neq a_y.$$

В этом случае строят двустороннюю критическую область. Вычисляем по данным наблюдений значение критерия

$$T_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)S_x^2 + (m-1)S_y^2}} \sqrt{\frac{nm \cdot (n+m-2)}{n+m}} \qquad (3.7)$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α (помещенному в верхней строке таблицы) и числу степеней свободы $k = n + m - 2$, находим критическую точку $t_{\text{двуст.кр}} = t_{\alpha; k}$.

Если $|T_{\text{набл}}| < t_{\text{двуст.кр}} -$ нет оснований отвергнуть нулевую гипотезу.

Если $|T_{\text{набл}}| > t_{\text{двуст.кр}} -$ нулевую гипотезу отвергаем.

Второй случай. Конкурирующая гипотеза имеет вид

$$H_1 : a_x > a_y.$$

В этом случае строится правосторонняя критическая область. Критическую точку $t_{\text{правост.кр}} = t_{\alpha; k}$ находят по таблице по уровню

значимости α, помещенному в нижней строке таблицы, и по числу степеней свободы $k = n + m - 2$.

Если $T_{\text{набл}} < t_{\text{правост.кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $T_{\text{набл}} > t_{\text{правост.кр}}$ – нулевую гипотезу отвергают.

Третий случай. Конкурирующая гипотеза

$$H_1 : a_x < a_y .$$

В этом случае строится левосторонняя критическая область. По таблице критических точек распределения Стьюдента находят левостороннюю критическую точку

$$t_{\text{левост.кр}} = -t_{\text{правост.кр}} .$$

Если $T_{\text{набл}} > t_{\text{левост.кр}}$ – нет оснований отвергнуть нулевую гипотезу.

Если $T_{\text{набл}} < t_{\text{левост.кр}}$ – нулевую гипотезу отвергают.

Замечание. Поскольку для проверки гипотезы требуется равенство дисперсий для двух выборок, то вначале необходимо проверить это равенство, в противном случае данный метод применять нельзя.

Гипотеза о равенстве вероятностей «успеха» в двух сериях испытаний Бернулли

Гипотеза проверяется на основе асимптотической нормальности относительных частот, так что данный метод может применяться только при больших объемах выборок (порядка сотен). Пусть в одной серии из n_1 испытаний получили m_1 «успехов», в другой серии из n_2 испытаний получили m_2 «успехов». Проверяем гипотезу $H_0 : p_1 = p_2$. Альтернативная гипотеза может быть трех видов: а) $p_1 \neq p_2$; б) $p_1 > p_2$; в) $p_1 < p_2$. Однако случай в) сводится к случаю б) перестановкой индексов, и мы не будем рассматривать его отдельно.

Во всех трех случаях вычисляют значение критерия

$$U_{\text{набл}} = \frac{w_1 - w_2}{\sqrt{w(1-w)\left(1/n_1 + 1/n_2\right)}}, \qquad (3.8)$$

где $w = \dfrac{m_1 + m_2}{n_1 + n_2}$.

В случае а) критическая точка $u_{\text{кр}}$ выбирается из условия $\Phi_0\left(u_{\text{кр}}\right)=\left(1-\alpha\right)/2$. Если $\left|U_{\text{набл}}\right|<u_{\text{кр}}$, то гипотеза H_0 принимается, в противном случае – отвергается.

В случае б) критическая точка $u_{\text{кр}}$ выбирается из условия $\Phi_0\left(u_{\text{кр}}\right)=1/2-\alpha$. Если $U_{\text{набл}}<u_{\text{кр}}$, то гипотеза H_0 принимается, если $U_{\text{набл}}>u_{\text{кр}}$ – отвергается.

Описанные выше критерии проверки гипотез можно представить виде табл. 4.2.

Таблица 4.2. Сравнение соответствующих параметров нормальных распределений

H_0	Предположение	Статистика критерия	H_1	Область принятия H_0
$a_1 = a_2$	σ_1^2, σ_2^2 известны	$U = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$a_1 > a_2$	$U < u_{\text{кр}}$, $\Phi_0\left(u_{\text{кр}}\right) = \dfrac{1}{2} - \alpha$
			$a_1 < a_2$	$U > -u_{\text{кр}}$, $\Phi_0\left(u_{\text{кр}}\right) = \dfrac{1}{2} - \alpha$
			$a_1 \neq a_2$	$\lvert U \rvert < u_{\text{кр}}$, $\Phi_0\left(u_{\text{кр}}\right) = \dfrac{1-\alpha}{2}$
$a_1 = a_2$	$\sigma_1^2 = \sigma_2^2 = \sigma^2$ не известны, но равны	$T_{n+m-2} = \dfrac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$, $S = \dfrac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$	$a_1 > a_2$	$T_{n+m-2} < t_{\text{кр}}$ (для односторонней области)
			$a_1 < a_2$	$T_{n+m-2} > -t_{\text{кр}}$ (для односторонней области)
			$a_1 \neq a_2$	$\lvert T_{n+m-2} \rvert < t_{\text{кр}}$ (для двусторонней области)
$\sigma_1^2 = \sigma_2^2$	a_1, a_2 неизвестны	$F_{n_1-1, n_2-1} = \dfrac{S_1^2}{S_2^2}$, $\left(S_1^2 > S_2^2\right)$	$\sigma_1^2 > \sigma_2^2$	$F_{n_1-1, n_2-1} < F_{\text{кр}}$
			$\sigma_1^2 \neq \sigma_2^2$	$F_{n_1-1, n_2-1} < F_{\text{кр}}$
$p_1 = p_2$	n_1, n_2 порядка нескольких десятков	$U = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, $\hat{p} = \dfrac{m_1 + m_2}{n_1 + n_2}$	$p_1 > p_2$	$U < u_{\text{кр}}$, $\Phi_0\left(u_{\text{кр}}\right) = \dfrac{1}{2} - \alpha$
			$p_1 < p_2$	$U > -u_{\text{кр}}$, $\Phi_0\left(u_{\text{кр}}\right) = \dfrac{1}{2} - \alpha$
			$p_1 \neq p_2$	$\lvert U \rvert < u_{\text{кр}}$, $\Phi_0\left(u_{\text{кр}}\right) = \dfrac{1-\alpha}{2}$

Критерии согласия

Эмпирические законы распределения вероятностей имеют дискретный характер независимо от того, является ли эта величина дискретной или непрерывной. Использование такого закона в различных расчетах оказывается не всегда удобным. Поэтому возникает задача замены его некоторым теоретическим законом распределения, который был бы в определенном смысле близким к эмпирическому закону. Эта задача решается следующим образом. По виду гистограммы, полигона или графика эмпирической функции распределения выбирается подходящий закон распределения и выдвигается гипотеза (H_0) о том, что именно этот выбранный закон является истинным законом распределения изучаемой случайной величины. Проверка гипотезы о предполагаемом законе неизвестного распределения производится так же, как и проверка гипотезы о параметрах распределения, т.е. при помощи специально подобранной случайной величины – критерия согласия.

Критерием согласия называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Имеется несколько критериев согласия: χ^2-Пирсона, Колмогорова, Смирнова и др.

Одним из наиболее распространенных является критерий χ^2-Пирсона.

Рассмотрим первый случай.

Пусть $x_1, x_2, ..., x_n -$ независимые наблюдения некоторой случайной величины X с неизвестной функцией распределения $F_X(x)$. Требуется по выборке $x_1, ..., x_n$ проверить нулевую гипотезу о том, что генеральная совокупность X имеет функцию распределения $F_0(x)$, если *известны параметры распределения*, а $F_0(x) -$ непрерывная или дискретная функция.

Для проверки этой гипотезы область значений $x_1, ..., x_n$ разбивают на r непересекающихся интервалов и полуинтервалов вида $(-\infty, c_1)$, $[c_1, c_2), \ldots,$ $[c_{r-1}, +\infty)$.

Если справедлива основная гипотеза, т.е. величины $x_1, ..., x_n$ своей функцией распределения имеют $F_0(x)$, то можно найти теоретические вероятности попадания случайной величины в частичные интервалы по формуле

$$p_i = P(c_i \le X \le c_{i+1}) = F_0(c_{i+1}) - F_0(c_i),$$

где $p_i > 0, \sum_{i=1}^{r-1} p_i = 1.$

Обозначим через $n_i -$ число вариант, среди $x_1,...,x_n$, попавших в интервал $\Delta_i = [c_i, c_{i+1})$. Тогда величину n_i/n можно взять за эмпирическую вероятность попадания X в тот же интервал Δ_i. В качестве меры отклонения эмпирического распределения от теоретического в i-ом интервале можно взять квадрат разности между эмпирической и теоретической вероятностью попадания изучаемой величины в этот интервал

$$\left(\frac{n_i}{n} - p_i\right)^2 = \frac{1}{n^2}(n_i - np_i)^2.$$

Если теперь суммировать квадраты отклонений по всем интервалам, то, согласно теории ошибок Гаусса, каждое слагаемое должно входить в сумму со своей точностью γ_i. Получим величину

$$\sum_{i=1}^{r} \gamma_i \left(\frac{n_i}{n} - p_i\right)^2.$$

Пирсон показал, что если полагать $\gamma_i = n/p_i$, т.е.

$$\sum_{i=1}^{r} \gamma_i \left(\frac{n_i}{n} - p_i\right)^2 = \sum_{i=1}^{r} \frac{n}{p_i} \frac{1}{n^2}(n_i - np_i)^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i},$$

величина $y_i = \dfrac{n_i - np_i}{\sqrt{np_i}}$ асимптотически нормальная величина и $y_i \in N_{0,1}$. Поэтому сумма квадратов y_i, по определению, будет иметь распределение χ^2 с $r-1$ степенями свободы, т.е.

$$\sum_{i=1}^{r} y_i^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} = \chi^2.$$

Следовательно, если $x_1,...,x_n -$ выборка из генеральной совокупности с функцией распределения $F_0(x)$, то статистика

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

имеет при достаточно больших n распределение хи-квадрат с $r-1$ степенями свободы, если основная гипотеза верна. В противном случае статистика стремится к бесконечности. Поэтому в качестве критической области выбирается область больших значений.

Поскольку односторонний критерий более «жестко» отвергает нулевую гипотезу, чем двусторонний, построим правостороннюю критическую область, исходя из требования

$$P\left(\chi^2 \geq \chi^2_{\alpha, r-1}\right) = \alpha,$$

где $\alpha -$ малая величина. Эта величина (уровень значимости) по смыслу является вероятностью отвергнуть истинную гипотезу, т.е. вероятностью ошибки первого рода. Величина $\chi^2_{\alpha, r-1} -$ критическое значение величины χ^2. Она является нижней границей односторонней критической области $\left[\chi^2_{\alpha, r-1}; +\infty\right)$ данного критерия. Правило принятия решения основывается на сравнении вычисленного значения χ^2 с критическим значением $\chi^2_{\alpha, r-1}$. Если $\chi^2 < \chi^2_{\alpha, r-1}$, то гипотеза о соответствии эмпирического и теоретического законов принимается, в противном случае, когда вычисленное χ^2 попадает в критическую область, эта гипотеза отвергается.

Идея этого критерия заключается в следующем. Величина $\chi^2_{\text{набл}} = \chi^2$ распределена по закону хи-квадрат только в том случае, когда гипотеза H_0 является истинной. Именно при этом условии было определено критическое значение $\chi^2_{\text{кр}} = \chi^2_{\alpha, r-1}$ таким образом, что вероятность выполнения неравенства $\chi^2 \geq \chi^2_{\alpha, r-1}$ очень мала (α). Предположим теперь, что это неравенство, тем не менее, выполнилось, т.е. вычисленное значение попало в критическую область. Это значит, что произошло маловероятное событие. Если все наши рассуждения верны, то такой результат является достаточно редким. Поэтому мы подвергаем сомнению наши рассуждения. Мы могли сделать ошибку лишь тогда, когда предположили истинность гипотезы H_0. Следовательно, эту гипотезу нужно отвергнуть. Принимая такое решение, мы можем совершить ошибку с вероятностью α, поскольку именно с этой вероятностью указанное маловероятное событие может произойти.

Таким образом, мы приходим к следующему *алгоритму проверки гипотезы*:

1. Из генеральной совокупности производится выборка объема n $(n > 50)$.

2. Весь диапазон полученных значений разбивается на r частичных интервалов одинаковой длины. Пусть в i-ом интервале будет n_i значений, так что $\sum_{i=1}^{r} n_i = n$, $n_i \approx 5-8$, иначе интервалы объединяются.

3. Составляется сгруппированный статистический ряд.

4. На основании гипотетической функции распределения $F_0(x)$ вычисляются вероятности попадания случайной величины в частичные интервалы

$$p_i = P(c_i \le X \le c_{i+1}) = F_0(c_{i+1}) - F_0(c_i).$$

5. Находим наблюдаемое значение критерия

$$\chi^2_{\text{набл}} = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}. \qquad (3.9)$$

6. Из таблицы критических точек распределения хи-квадрат по заданному уровню значимости α и числу степеней свободы $r-1$ находим критическую точку $\chi^2_{\text{кр}} = \chi^2_{\alpha, r-1}$.

7. Сравнивая наблюдаемое значение с критическим, принимаем одно из двух решений:

а) если $\chi^2_{\text{набл}} \ge \chi^2_{\text{кр}}$, то нулевая гипотеза отвергается в пользу альтернативной, т.е. считается, что гипотетическая функция распределения не согласуется с опытными данными;

б) если $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$, то для отклонения нулевой гипотезы нет оснований, т.е. гипотетическая функция распределения $F_0(x)$ согласуется с опытными данными.

Теперь рассмотрим второй случай.

Пусть $x_1, x_2, ..., x_n -$ независимые наблюдения некоторой случайной величины X с неизвестной функцией распределения $F_X(x)$. Требуется по выборке $x_1, ..., x_n$ проверить нулевую гипотезу о том, что генеральная совокупность X имеет функцию распределения $F_0(x)$, но при этом *параметры распределения неизвестны*.

В этом случае перед вычислением вероятностей p_i неизвестные параметры (например, математическое ожидание, дисперсия и т.п.) оцениваются по той же самой выборке $x_1, x_2, ..., x_n$. То есть сначала необходимо вычислить выборочные характеристики этих параметров. В таком случае число степеней свободы распределения χ^2 уменьшается и

становится равным $r-s-1$, где $s-$ число оцененных параметров. Число частичных интервалов r следует брать не менее 8, однако если для некоторых интервалов при вычислениях оказывается $np_i<5$, то их нужно объединять с соседними интервалами, чтобы было $np_i\geq 5$ (в ответственных случаях не допускается $np_i<10$). Общее число интервалов при этом, естественно, сокращается.

Сравнивая наблюдаемое значение критерия $\chi^2_{\text{набл}}$ с критическим значением $\chi^2_{\text{кр}}=\chi^2_{\alpha,\,r-s-1}$, делаем заключение об истинности нулевой гипотезы: гипотеза принимается, если $\chi^2_{\text{набл}}<\chi^2_{\text{кр}}$, и отвергается в противном случае.

Критерий хи-квадрат для простой гипотезы (в случае известных параметров распределения) называется критерием χ^2-Пирсона, а критерий хи-квадрат со сложной гипотезой (в случае неизвестных параметров) – критерием χ^2-Фишера.

Пример

В таблице представлены данные о числе сделок, заключенных на фондовой бирже за квартал, для 517 инвесторов:

i	0	1	2	3	4	5	6	7
n_i	112	168	130	68	32	5	1	1

В первой строке приведено число сделок, во второй – число инвесторов, заключивших указанное количество сделок за квартал.

Проверить, используя критерий Пирсона, что на уровне значимости $\alpha=0{,}05$ число сделок, заключенных одним инвестором за квартал, распределено по закону Пуассона с параметром $\lambda=1{,}5$.

Решение. Поскольку распределение Пуассона дискретное, в качестве различных исходов здесь можно принять сами значения случайной величины. Заметим, что два последних значения (6 и 7) встретились слишком мало раз, поэтому их следует объединить с предыдущим (5). Кроме того, распределение Пуассона не ограничено справа, и следует учесть все значения, превышающие число 7 (которые не встретились ни разу). Таким образом, вместо интервалов в качестве Δ_i берем $\{0\},\{1\},\{2\},\{3\},\{4\},[5;+\infty)$. Здесь $r=6$.

Найдем теоретические вероятности p_i по формуле распределения Пуассона:

$$p_i=P(X=i)=\frac{\lambda^i}{i!}e^{-\lambda}.$$

При $\lambda = 1{,}5$ получаем:

$$p_0 = P(X=0) \approx 0{,}2231;\ \ p_1 = P(X=1) \approx 0{,}3347;\ \ p_2 = P(X=2) \approx 0{,}251;$$

$$p_3 = P(X=3) \approx 0{,}1255;\ \ p_4 = P(X=4) \approx 0{,}0471;\ \ p_5 = P(X=5) \approx 0{,}0186.$$

Умножим их на число инвесторов $n = 517$ и составим расчетную таблицу.

Δ_i	n_i	np_i	$n_i - np_i$	$\dfrac{(n_i - np_i)^2}{np_i}$
0	112	115,34	–3,34	0,1
1	168	173,04	–5,04	0,15
2	130	129,77	0,23	0,00
3	68	64,88	3,12	0,15
4	32	24,35	7,65	2,40
≥ 5	7	9,62	–2,62	0,71

Суммируя значения в последнем столбце, получаем наблюдаемое значение статистики хи-квадрат $\chi^2_{\text{набл}} = 3{,}51$.

Из таблицы критических точек распределения хи-квадрат по уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $r - 1 = 5$ находим критическую точку $\chi^2_{\text{кр}} = 11{,}1$. Поскольку $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$, можно считать, что число сделок, заключенных одним инвестором за квартал, распределено по закону Пуассона с параметром $\lambda = 1{,}5$.

Замечание. Если бы значение параметра $\lambda = 1{,}5$ было оценено по самой выборке, следовало бы задать число степеней свободы $r - 2 = 4$. Тогда $\chi^2_{\text{кр}} = 9{,}5$, следовательно, гипотеза тоже принимается.

Примеры решения задач к главе 3

1. Из нормальной генеральной совокупности с известным средним квадратическим отклонением $\sigma = 0{,}2$ извлечена выборка объема $n = 25$ и по ней найдена выборочная средняя $\overline{x} = 21{,}04$. Проверить нулевую гипотезу $H_0 : a = a_0 = 21$ при конкурирующей гипотезе $H_1 : a \neq 21$ и уровне значимости 0,1 и 0,05.

Решение. Найдем наблюдаемое значение критерия:

$$U_{\text{набл}} = \frac{(21{,}04 - 21)\sqrt{25}}{0{,}2} = 1.$$

Найдем критическую точку двусторонней критической области:

$$\Phi\left(u_{\text{кр}}\right) = \frac{1 - 0{,}1}{2} = 0{,}45,$$

и по таблице функции Лапласа находим $u_{\text{кр}} \approx 1{,}65.$

Поскольку $u_{\text{кр}} = 1{,}65 > 1 = U_{\text{набл}}$, то нулевая гипотеза принимается.

Теперь решим вторую часть задачи, когда уровень значимости $\alpha = 0{,}05$. В этом случае критическая точка $u_{\text{кр}}$ ищется из выражения:

$$\Phi\left(u_{\text{кр}}\right) = \frac{1 - 0{,}05}{2} = 0{,}475.$$

По таблице функции Лапласа находим $u_{\text{кр}} \approx 2$.

Поскольку $\left|U_{\text{набл}}\right| = 1{,}18 < 2 = u_{\text{кр}}$, то нет оснований отвергнуть гипотезу о незначительном отличии наблюдаемой относительной частоты от гипотетической вероятности.

Ответ. Нет оснований отвергнуть нулевую гипотезу.

2. Точность работы станка-автомата проверяется по дисперсии размеров изделий, которая не должна превышать $\sigma_0^2 = 0{,}01$ (мм^2). По выборке из 25 изделий получена исправленная выборочная дисперсия $S^2 = 0{,}02$ (мм^2). На уровне значимости 0,05 проверить, обеспечивает ли станок необходимую точность.

Решение. Необходимо проверить гипотезу $H_0 : \sigma^2 = \sigma_0^2$. Найдем наблюдаемое значения критерия, по формуле

$$\chi^2_{\text{набл}} = \frac{(n-1) \cdot S^2}{\sigma_0^2} = \frac{24 \cdot 0{,}02}{0{,}01} = 48.$$

Исходя из условия задачи, берем конкурирующую гипотезу $H_1 : \sigma^2 > \sigma_0^2$, поэтому получаем правостороннюю критическую область. Находим критическую точку $\chi^2_{\text{кр}}$ по таблице критических точек

распределения χ^2 с $k = n - 1 = 24$ степенями свободы и уровнем значимости $\alpha = 0,05$, получаем $\chi^2_{\text{кр}} = \chi^2_{0,05;24} = 36,4$.

Поскольку $\chi^2_{\text{набл}} = 48 > 36,4 = \chi^2_{\text{кр}}$, то основная гипотеза отвергается, принимается конкурирующая гипотеза, т.е. станок не обеспечивает необходимой точности.

Ответ. Нулевая гипотеза отвергается.

3. Торговец утверждает, что он получает заказы в среднем по крайней мере от 30 % предполагаемых клиентов. Можно ли при 5 %-м уровне значимости считать это утверждение верным, если торговец получил заказы от 20 из 100 случайно отобранных потенциальных клиентов?

Решение. В данном случае нулевая гипотеза имеет вид $H_0 : p = p_0 = 0,3$, а конкурирующая гипотеза $H_1 : p < 0,3$.

Найдем значение статистики критерия, учитывая, что относительная частота равна $w = 20/100 = 0,2$:

$$U_{\text{набл}} = \frac{(w - p_0)\sqrt{n}}{\sqrt{p_0(1 - p_0)}} = \frac{(0,2 - 0,3)\cdot 10}{\sqrt{0,3 \cdot 0,7}} \approx -2,18.$$

Из соотношения $\Phi_0(u_{\text{кр}}) = 1/2 - \alpha = 0,45$ находим $u_{\text{кр}} = 1,65$. Так как $U_{\text{набл}} < -u_{\text{кр}}$, нулевая гипотеза отвергается, и с утверждением торговца мы не соглашаемся.

4. Исследование длительности оборотных средств двух групп предприятий (по 13 предприятий в каждой) дало следующие результаты: $\bar{x} = 23$ дня, $\bar{y} = 26$ дней, $\sigma_x^2 = 3$ дня, $\sigma_y^2 = 6$ дней. Можно ли считать, что отклонения в длительности оборота оборотных средств групп предприятий одинаковы для уровня значимости 0,1?

Решение. В этой задаче надо проверить нулевую гипотезу $H_0 : \sigma_x^2 = \sigma_y^2$ о равенстве генеральных дисперсий нормальных совокупностей при конкурирующей гипотезе $H_1 : \sigma_x^2 \neq \sigma_y^2$. Используем критерий Фишера со степенями свободы $k_1 = k_2 = 13 - 1 = 12$ и вычислим наблюдаемое значение критерия (отношение большей дисперсии к меньшей)

$$F_{\text{набл}} = \frac{\sigma_y^2}{\sigma_x^2} = \frac{6}{3} = 2.$$

По таблице критических точек распределения Фишера по уровню значимости для двусторонней критической области $\alpha/2 = 0,1/2 = 0,05$ и

числам степеней свободы $k_1 = k_2 = 12$ находим критическую точку $F_{\text{кр}} = F_{0,05;12;12} = 2,69.$

Так как $F_{\text{набл}} = 2 < 2,69 = F_{\text{кр}}$, то нет оснований отвергать нулевую гипотезу о равенстве отклонений в длительности оборота оборотных средств двух групп предприятий.

Ответ. Нет оснований отвергнуть нулевую гипотезу.

5. Школьникам давались обычные арифметические задачи, а потом одной случайно выбранной половине учащихся сообщалось, что они не выдержали испытания, а остальным – обратное. Затем у каждого из них спрашивали, сколько секунд ему потребуется для решения новой задачи. Экспериментатор, вычисляя разность между определенным временем решения задачи, которое называл школьник, и результатами ранее выполненного задания, получил следующие данные:

Группа 1 (учащиеся, которым сообщалось о положительном результате)	$n_1 = 13;\ S_1^2 = 4,06$
Группа 2 (учащиеся, которым сообщалось о неудаче)	$n_2 = 12,\ S_2^2 = 20,25$

Проверьте на уровне значимости 0,01 гипотезу о том, что дисперсия совокупности детских оценок, имеющих отношение к оценке их возможностей, не зависит от того, что сообщалось детям о плохих результатах испытаний или об удачном решении первой задачи.

Решение. Применим критерий Фишера для нулевой гипотезы $H_0 : \sigma_x^2 = \sigma_y^2$ и конкурирующей $H_1 : \sigma_x^2 > \sigma_y^2$.

Вычислим наблюдаемое значение критерия

$$F_{\text{набл}} = \frac{S_2^2}{S_1^2} = \frac{20,25}{4,06} \approx 4,99.$$

Критическую точку находим в приложении для уровня значимости $\alpha = 0,01$ и числам степеней свободы $k_1 = 12 - 1$ и $k_2 = 13 - 1$:

$$F_{\text{кр}} = F_{0,01;11;12} = 4,22.$$

Получили, что $F_{\text{набл}} = 4,99 > 4,22 = F_{\text{кр}}$ и нулевая гипотеза на уровне значимости 0,01 отвергается.

Ответ. Нулевая гипотеза отвергается.

6. По выборке объема $n=30$ найден средний вес изготовленных на первом станке изделий, равный 130 г; по выборке объемом $m=40$ найден средний вес изготовленных на втором станке изделий, равный 125 г. Генеральные дисперсии известны: $\sigma_x^2=60\ \text{г}^2$, $\sigma_y^2=80\ \text{г}^2$. На уровне значимости $\alpha=0,05$ требуется проверить нулевую гипотезу $H_0: a_x=a_y$ при конкурирующей гипотезе $H_1: a_x\neq a_y$. Предполагается, что случайные величины распределены нормально и выборки независимы.

Решение. Нулевая и конкурирующая гипотезы даны в условии задачи, поэтому сразу вычислим значение статистики критерия:

$$U=\frac{\overline{x}-\overline{y}}{\sqrt{\sigma_x^2/n+\sigma_y^2/m}}=\frac{130-125}{\sqrt{60/30+80/40}}=2,5.$$

По таблице функции Лапласа найдем критическую точку из равенства $\Phi\left(u_{\text{кр}}\right)=(1-\alpha)/2=0,475$, в результате получаем $u_{\text{кр}}=1,96$.

Поскольку $|U|>u_{\text{кр}}$, гипотеза H_0 отвергается.

Ответ. Нельзя утверждать, что средние значения веса изделий двух станков совпадают.

7. Реклама утверждает, что из двух типов пластиковых карточек «Русский экспресс» и «Super Card» богатые люди предпочитают первый. С целью проверки этого утверждения были обследованы ежемесячные платежи $n=16$ обладателей «Русского Экспресса» и $m=11$ обладателей «Super Card». Выяснилось, что платежи по карточкам «Русский Экспресс» составляют в среднем 563 ден. ед. с исправленным средним квадратическим отклонением 178 ден. ед., а по карточкам «Super Card» – в среднем 485 ден. ед. с исправленным средним квадратическим отклонением 196 ден. ед.

Предварительный анализ законов распределения ежемесячных расходов как среди обладателей «Русского Экспресса» так и среди обладателей «Super Card» показал, что они достаточно хорошо описываются нормальным распределением.

Проверить утверждение рекламы на уровне значимости $\alpha=10\ \%$.

Решение. В данном случае речь идет о проверке гипотезы о средних при неизвестных дисперсиях. Поэтому сначала необходимо проверить гипотезу о равенстве дисперсий, а лишь затем двигаться дальше. Имеем

$$F=\frac{S_{\text{б}}^2}{S_{\text{м}}^2}=\frac{196^2}{178^2}=\frac{38146}{31684}\approx 1,21.$$

Из таблицы критических точек распределения Фишера по уровню значимости $\alpha/2=0,05$ и числам степеней свободы $k_1=n_{\max}-1=m-1=10$ и

$k_2 = n_{\min} - 1 = n - 1 = 15$ найдем критическую точку $F_{\text{кр}} = 2,55$. Поскольку $1,21 < 2,55$, принимаем гипотезу о равенстве дисперсий двух выборок.

Теперь мы можем воспользоваться критерием Стьюдента для проверки гипотезы о равенстве средних. Имеем

$$T = \frac{\overline{x} - \overline{y}}{\sqrt{(n-1)S_x^2 + (m-1)S_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} =$$

$$= \frac{563 - 485}{\sqrt{15 \cdot 31684 + 10 \cdot 38416}} \cdot \sqrt{\frac{16 \cdot 11(16+11-2)}{16+11}} \approx 1,01.$$

Из таблиц критических точек распределения Стьюдента (для односторонней области) по уровню значимости $\alpha = 0,1$ и числу степеней свободы $n + m - 2 = 25$ находим $t_{\text{кр}} = 1,32$. Поскольку $t < t_{\text{кр}}$, принимается основная гипотеза (о равенстве средних).

Ответ. Утверждение рекламы не подтверждается имеющимися данными.

8. В партии из 500 деталей, изготовленных первым станком-автоматом, оказалось 60 нестандартных, а из 600 деталей второго станка – 42 нестандартных. На уровне значимости $\alpha = 0,01$ проверить нулевую гипотезу $H_0 : p_1 = p_2$ о равенстве вероятностей изготовления нестандартной детали обоими станками при конкурирующей гипотезе $H_1 : p_1 \neq p_2$.

Решение. Вычислим относительные частоты:

$$w_1 = \frac{60}{500} = 0,12;\ w_2 = \frac{42}{600} = 0,07;\ w = \frac{m_1 + m_2}{n_1 + n_2} = \frac{60 + 42}{500 + 600} = 0,09.$$

Найдем наблюдаемое значение критерия:

$$U = \frac{0,12 - 0,07}{\sqrt{0,09 \cdot 0,91\left(\frac{1}{500} + \frac{1}{600}\right)}} \approx 2,85.$$

Найдем критическую точку из соотношения:

$$\Phi_0\left(u_{\text{кр}}\right) = \frac{1 - \alpha}{2} = 0,495,$$

откуда $u_{\text{кр}} = 2,57$. Поскольку $|U| > u_{\text{кр}}$, нулевая гипотеза отвергается.

Ответ. Вероятности изготовления нестандартных деталей на двух станках различны.

9. В таблице приведены сгруппированные данные о коэффициентах соотношения заемных и собственных средств на 100 малых предприятиях.

№ интервала	Интервал	Середины интервалов	n_i
1	5,05–5,15	5,1	5
2	5,15–5,25	5,2	8
3	5,25–5,35	5,3	12
4	5,35–5,45	5,4	20
5	5,45–5,55	5,5	26
6	5,55–5,65	5,6	15
7	5,65–5,75	5,7	10
8	5,75–5,85	5,8	4

На уровне значимости $\alpha = 0,05$ проверить гипотезу о том, что коэффициенты можно описать нормальным распределением.

Решение. Необходимо проверить гипотезу о нормальном распределении, используя критерий согласия. В данном случае параметры распределения не заданы, и их следует оценить по сгруппированным данным. Находим выборочное среднее $\overline{x} = 5,46$ (середины интервалов умножаем на соответствующие частоты и сумму делим на $n = 100$) и выборочное среднее квадратическое отклонение $s = 0,03$.

Теоретические вероятности находим по формуле

$$P\left(c_i < X < c_{i+1}\right) = \Phi_0\left(\frac{c_{i+1} - \overline{x}}{s}\right) - \Phi_0\left(\frac{c_i - \overline{x}}{s}\right), \quad i = 0,1,2,...,7.$$

Следует продолжить крайние интервалы и считать $c_0 = -\infty, c_8 = +\infty$, поскольку нормальное распределение не ограничено с обеих сторон.

С учетом полученных значений построим таблицу:

Δ_i	n_i	np_i	$n_i - np_i$	$\frac{\left(n_i - np_i\right)^2}{np_i}$

$\le 5{,}15$	5	3,67	1,32	0,47
5,15–5,25	8	7,59	0,41	0,02
5,25–5,35	12	15,00	–3,00	0,60
5,35–5,45	20	21,43	–1,43	0,09
5,45–5,55	26	22,14	3,86	0,67
5,55–5,65	15	16,53	–1,53	0,14
5,65–5,75	10	8,93	1,07	0,13
$\ge 5{,}75$	4	4,70	–0,70	0,10

Суммируя значения в последнем столбце, получаем наблюдаемое значение критерия $\chi^2_{\text{набл}} = 2{,}22$.

Из таблицы критических точек распределения хи-квадрат по уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $r-1-s = 8-1-2 = 5$ находим критическую точку $\chi^2_{\text{кр}} = 11{,}1$. Поскольку $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$, можно считать, что коэффициенты хорошо описываются нормальным распределением.

Замечание. Здесь можно было объединить крайние интервалы с соседними. Вычисления показывают, что и в этом случае гипотеза принимается.

Ответ. Гипотеза о нормальном распределении принимается.

Задачи для самостоятельного решения

1. В результате длительного хронометража времени сборки узла различными сборщиками установлено, что дисперсия этого времени $\sigma_0^2 = 2$. Результаты 20 наблюдений за работой новичка таковы (x_i – время сборки одного узла в минутах (середины интервалов) n_i – частота):

x_i	56	58	60	62	64
n_i	1	4	10	3	2

Можно ли на уровне значимости 0,05 считать, что дисперсия затрачиваемого новичком времени существенно не отличается от дисперсии времени остальных сборщиков?

2. Независимому статистику поручено проверить информацию маркетинговой службы некоторого туристического бюро о том, что 70 % клиентов выбирают в качестве формы обслуживания полупансион. Статистик провел опрос 150 случайно выбранных туристов, из них

полупансион предпочли 84 человека. К какому выводу пришел статистик при проверке гипотезы $H_0: p = 0,7$ при альтернативе $H_1: p \neq 0,7$ на уровне значимости критерия $\alpha = 0,05$?

3. Статистику необходимо проверить экспертную оценку о том, что 75 % отечественных предприятий уклоняются (частично) от уплаты налогов. По результатам неофициального опроса руководителей предприятий 140 из 200 случайно отобранных директоров подтвердили, что используют различные схемы для ухода от уплаты налогов. Можно ли на уровне значимости 0,05 согласиться с приведенной экспертной оценкой?

4. Фирма разослала 1000 новых рекламных каталогов и получила 120 заказов. Можно ли утверждать (на уровне значимости 5 %), что эффективность рекламы повысилась, если ранее она составляла в среднем 10 %?

5. Средний доход фирмы в день составлял 1020 единиц. После реорганизации выборочный средний доход в день за 30 рабочих дней составил 1070 единиц с исправленным выборочным средним квадратическим отклонением 90 единиц. Можно ли утверждать (на уровне значимости 5 %), что реорганизация привела к увеличению среднего дохода?

6. Инвестор считает вложения в активы с дисперсией доходности более 0,04 слишком рискованными. За последние 10 лет исправленная выборочная дисперсия доходности актива составила 0,06. Следует ли делать вложения в этот актив, принимая решение на уровне значимости 5 %?

7. Рафинированный сахар упаковывается в пакеты с номинальным весом 1 кг со средним квадратическим отклонением, равным 0,01 кг. При 5 %-ном уровне значимости проверить нулевую гипотезу о том, что средний вес пакета соответствует номиналу.

8. В селе Петрово проведено выборочное обследование доходов жителей. По выборке из 25 человек получено среднее 2380 руб. и среднее квадратическое отклонение 90 руб. Можно ли утверждать на уровне значимости 5 %, что средний доход жителей составляет менее 2500 руб.?

9. Партия изделий принимается, если дисперсия размеров не превышает 0,2. Исправленная выборочная дисперсия для 30 изделий оказалась равной 0,3. Можно ли принять партию на уровне значимости 5 %?

10. Из нормальной генеральной совокупности извлечена выборка объема 31:

x_i	10,1	10,3	10,6	11,2	11,5	11,8	12,0
n_i	1	3	7	10	6	3	1

Требуется при уровне значимости 0,05 проверить нулевую гипотезу $H_0: \sigma^2 = \sigma_0^2 = 0,18$, приняв в качестве конкурирующей гипотезы $H_1: \sigma^2 > 0,18$.

11. По выборке объема 16, извлеченной из нормальной генеральной совокупности, найдены выборочное среднее $\bar{x} = 118,2$ и исправленное среднеквадратическое отклонение $s = 3,6$. Требуется при уровне значимости 0,05 проверить нулевую гипотезу $H_0 : a = a_0 = 120$ при конкурирующей гипотезе

а) $H_1 : a \neq 120$;

б) $H_1 : a < 120$.

12. За последние 5 лет выборочная дисперсия доходности актива A составила 0,04, актива $B - 0,05$. Есть ли основание утверждать, что вложения в актив A менее рискованны, чем в актив B? Уровень значимости 5 %.

13. По двум независимым извлеченным из нормальных генеральных совокупностей выборкам, объемы которых $n = 9, m = 16$, найдены исправленные выборочные дисперсии $S_x^2 = 34,02$ и $S_y^2 = 12,15$. На уровне значимости 0,01 проверить нулевую гипотезу $H_0 : \sigma_x^2 = \sigma_y^2$ против конкурирующей гипотезы $H_1 : \sigma_x^2 > \sigma_y^2$.

14. Двумя методами проведены измерения одной и той же физической величины. Получены следующие результаты:

а) в первом случае $x_1 = 9,6;\ x_2 = 10;\ x_3 = 9,8;\ x_4 = 10,2;\ x_5 = 10,6;$

б) во втором случае $y_1 = 10,4;\ y_2 = 9,7;\ y_3 = 10;\ y_4 = 10,3.$

Можно ли считать, что оба метода обеспечивают одинаковую точность измерений, если принять уровень значимости 0,1? Предполагается, что результаты измерений распределены нормально и выборки независимы.

15. Для сравнения точности двух станков-автоматов взяты две пробы, объемы которых $n_1 = 10,\ n_2 = 8$. В результате измерения контролируемого размера отобранных изделий получены следующие результаты:

x_i	1,08; 1,10; 1,12; 1,14; 1,15; 1,25; 1,36; 1,38; 1,40; 1,42;
y_i	1,11; 1,12; 1,18; 1,22; 1,33; 1,35; 1,36; 1,38.

Можно ли считать, что станки обладают одинаковой точностью ($H_0 : D(X) = D(Y)$), если принять уровень значимости 0,1 и в качестве конкурирующей гипотезы $H_1 : D(X) \neq D(Y)$?

16. Для оценки качества изделий, изготовленных двумя заводами, взяты выборки $n_1 = 200$ и $n_2 = 300$ (изделий). В этих выборках оказалось соответственно $m_1 = 20, m_2 = 15$ бракованных изделий. При уровне значимости 0,05 проверить нулевую гипотезу $H_0 : p_1 = p_2$ о равенстве вероятностей изготовления бракованного изделия обоими заводами при конкурирующей гипотезе $H_1 : p_1 > p_2$.

17. Из 100 выстрелов по цели каждым из двух орудий зарегистрировано $m_1 = 12$ и $m_2 = 8$ промахов соответственно. На уровне значимости 0,05 проверить нулевую гипотезу о равенстве вероятностей промаха обоих орудий при конкурирующей гипотезе $H_1 : p_1 \neq p_2$.

18. Аудиторы компании интересуются системой обработки счетов доходов. Они взяли случайную выборку объема $n_1 = 50$ законченных счетов, в которой 4 счета оказались дефектными. Тогда аудиторы предложили некоторые модификации в процедуре и через определенное время провели случайную выборку $n_2 = 60$ завершенных счетов и обнаружили 3 дефектных счета. Имеется ли основание предполагать на уровне значимости 5 %, что новые процедуры уменьшают ошибку?

19. Производство пшеницы в России в 1995–2002 гг. представлено в таблице:

Год	1995	1996	1997	1998	1999	2000	2001	2002
Урожайность	30,1	34,9	44,3	27,0	31,0	34,5	47,0	57,7

Можно ли утверждать, что производство пшеницы в 1995–1998 гг. и 1999–2002 гг. было в среднем одинаково (на уровне значимости 10 %)?

20. По выборке объема $n = 50$ найден средний размер диаметра валиков $\bar{x} = 20{,}1$ мм, изготовленных автоматом №1; по выборке объема $m = 50$ найден средний размер диаметра валиков $\bar{y} = 19{,}8$ мм, изготовленных автоматом № 2. Генеральные дисперсии известны: $D(X) = 1{,}75$, $D(Y) = 1{,}375$ мм^2. Требуется при уровне значимости 0,05 проверить нулевую гипотезу $H_0 : M(X) = M(Y)$ при конкурирующей гипотезе $H_1 : M(X) \neq M(Y)$. Предполагается, что случайные величины X и Y распределены нормально и выборки независимы.

21. По двум независимым, извлеченным из нормальных генеральных совокупностей выборкам, объемы которых $n = 10$, $m = 10$ соответственно, найдены выборочные средние, равные 14,3 и 12,2 соответственно. Генеральные дисперсии известны: $\sigma_x^2 = 22$, $\sigma_y^2 = 18$. На уровне значимости 0,05 проверить нулевую гипотезу $H_0 : a_x = a_y$ при конкурирующей гипотезе $H_1 : a_x > a_y$.

22. По двум независимым малым выборкам, объемы которых $n = 12$, $m = 18$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x} = 31{,}2$, $\bar{y} = 29{,}2$ и исправленные дисперсии $S_x^2 = 0{,}84$, $S_y^2 = 0{,}40$. Требуется при уровне значимости 0,05 проверить

нулевую гипотезу $H_0 : M(X) = M(Y)$ при конкурирующей гипотезе $H_1 : M(X) \neq M(Y)$.

23. В городе 17036 семей имеют двоих детей. В 4529 семьях – два мальчика, в 4019 – две девочки, в 8488 – мальчик и девочка. Можно ли на уровне значимости 0,05 считать, что количество мальчиков в семьях с двумя детьми имеет биномиальное распределение с вероятностью рождения мальчика 0,515?

24. В таблице представлены данные о числе сделок, заключенных на фондовой бирже за квартал для 400 инвесторов:

x_i	0	1	2	3	4	5	6	7	8	9	10
n_i	146	97	73	34	23	10	6	3	4	2	2

На уровне значимости $\alpha = 0,05$ проверить гипотезу о том, что число сделок, заключенных инвестором за квартал, имеет распределение Пуассона.

25. В таблице представлены данные о ежемесячных доходах жителей региона (в руб.) для 1000 жителей.

x_i	Менее 500	500–1000	1000–1500	1500–2000	2000–2500	Свыше 2500
n_i	58	96	239	328	147	132

На уровне значимости 0,05 проверить гипотезу о том, что доходы жителей региона можно описать нормальным распределением.

26. В таблице приведены сгруппированные данные по срокам службы (в часах) для 1000 изделий:

x_i	0–10	10–20	20–30	30–40	40–50	50–60	60–70
n_i	365	245	150	100	70	45	25

На уровне значимости 0,05 проверить гипотезу о том, что срок службы изделия имеет показательное распределение.

27. В таблице приведены сгруппированные данные о моментах прибытия машин к бензоколонке в течение 10 часов наблюдений (с 8 до 18 часов):

Интервал, ч	Число машин
8–9	12
9–10	40

10–11	22
11–12	16
12–13	28
13–14	6
14–15	11
15–16	33
16–17	18
17–18	14

На уровне значимости 0,01 проверить гипотезу о том, что моменты прибытия машин распределены равномерно.

Ответы. 1) нет; **2)** нулевая гипотеза отвергается; **3)** да; **4)** да; **5)** да; **6)** да; **7)** нулевая гипотеза отвергается; **8)** да; **9)** нет; **10)** нулевая гипотеза отвергается; **11) а)** принимаем нулевую гипотезу; **б)** отвергаем нулевую гипотезу; **12)** нет; **13)** нулевая гипотеза принимается; **14)** да; **15)** нет оснований считать точность станков различной; **16)** нулевая гипотеза отвергается; **17)** нулевая гипотеза принимается; **18)** нет; **19)** да; **20)** нет оснований отвергать нулевую гипотезу; **21)** нулевая гипотеза принимается; **22)** нулевая гипотеза отвергается; **23)** да; **24)** нет; **25)** да; **26)** нет; **27)** нет.

Контрольные вопросы

1. Что называется статистической гипотезой?
2. Дайте определение нулей гипотезы и конкурирующей гипотезы.
3. Что называется статистическим критерием?
4. Дайте определение ошибок первого и второго рода.
5. Как определяется мощность критерия?
6. Что такое критическая область?
7. Чем отличаются односторонняя и двусторонняя критические области?
8. Какие законы распределения можно применить при построении критической области в случае проверки гипотезы о математических ожиданиях?
9. Какой закон распределения применяется для построения критической области в случае проверки гипотезы о дисперсиях?
10. Какова основная идея критерия χ^2-Пирсона проверки гипотез о законах распределения?

Глава 4

ЭЛЕМЕНТЫ РЕГРЕССИОННОГО АНАЛИЗА

Функциональная, статистическая и корреляционная зависимости

Корреляционный анализ предназначен для изучения по выборочным данным статистической зависимости ряда величин. Мы ограничимся рассмотрением зависимости двух величин X и Y.

Две случайные величины могут быть связаны либо функциональной зависимостью, либо статистической, либо они могут быть независимыми.

Строгая функциональная зависимость между случайными величинами реализуется крайне редко, так как эти величины подвержены действию случайных факторов.

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. В частности, статистическая зависимость называется *корреляционной*, если изменение распределения одной из величин приводит к изменению среднего значения другой.

Условным средним $\bar{y}_x$ называется среднее арифметическое значений Y, соответствующих значению $X = x$.

Пример

Пусть изучается связь между величинами Y и X. Каждому значению X соответствует несколько значений Y. Пусть при $X = 2$ Y принимает три значения $y_1 = 5$, $y_2 = 6$, $y_3 = 10$. Тогда условное среднее

$$\bar{y}_2 = \frac{5+6+10}{3} = 7.$$

Если каждому значению x соответствует одно значение $\bar{y}_x$, то, очевидно, условная средняя величина есть функция от x. В этом случае говорят, что случайная величина Y зависит от X корреляционно.

Корреляционной зависимостью Y *от* X называется функциональная зависимость $\bar{y}_x$ от x:

$$\bar{y}_x = f(x) \qquad (4.1)$$

Уравнение (4.1) называется *уравнением регрессии* Y *на* X, функция $f(x)$ называется *регрессией* Y *на* X, а ее график – *линией регрессии* Y *на* X.

Аналогично определяется условная средняя величина $\bar{x}_y$ и корреляционная зависимость X от Y: $\bar{x}_y$ – среднее арифметическое

значений случайной величины X, соответствующих значению $Y = y$. Корреляционной зависимостью X от Y называется функциональная зависимость $\bar{x}_y$ от y:

$$\bar{x}_y = \phi(y) \tag{4.2}$$

Уравнение (4.2) – уравнение регрессии X на Y, функция $\phi(y)$ – регрессия X на Y, график функции $\phi(y)$ – линия регрессии.

Две основные задачи теории корреляции

Первая задача теории корреляции – установить форму корреляционной связи, т.е. вид функции регрессии (линейная, квадратичная, показательная и т.п.). Наиболее часто функции регрессии оказываются линейными. Если обе функции регрессии $f(x)$ и $\phi(x)$ линейны, то корреляцию называют *линейной*, в противном случае – нелинейной.

Вторая задача теории корреляции – оценить тесноту корреляционной связи. Теснота корреляционной зависимости Y от X оценивается по величине рассеяния значений Y вокруг условного среднего $\bar{y}_x$. Большее рассеяние свидетельствует о слабой зависимости либо об отсутствии зависимости. Малое рассеяние указывает наличие достаточно сильной зависимости; возможно даже, что Y и X связаны функционально, но под воздействием второстепенных случайных факторов эта связь оказалась размытой, в результате чего при одном и том же значении x величина Y принимает различные значения.

Аналогично, по величине рассеяния значений X вокруг условного среднего $\bar{x}_y$, оценивается теснота корреляционной связи X от Y.

Отыскание параметров выборочного уравнения прямой линии регрессии по несгруппированным данным

Пусть количественные признаки X и Y связаны линейной корреляционной зависимостью. В этом случае обе линии регрессии будут прямыми. Необходимо найти уравнения этих прямых.

Пусть проведено n независимых испытаний, и были получены n пар чисел:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

То есть это случайная выборка из генеральной совокупности случайной величины (X, Y). Рассмотрим простейший случай: различные значения x

признака X и соответствующие им значения y признака Y наблюдались по одному разу. В этом случае группировать данные нет необходимости. Также нет необходимости использовать понятие условной средней, поэтому искомое уравнение

$$\overline{y}_x = kx + b$$

можно записать в таком виде:

$$Y = kx + b.$$

Угловой коэффициент k прямой линии регрессии Y на X называют *выборочным коэффициентом регрессии Y на X* и обозначают ρ_{yx}. Таким образом, мы ищем уравнение

$$Y = \rho_{yx} \cdot x + b. \qquad (4.3)$$

Поставим своей задачей подобрать параметры ρ_{yx} и b так, чтобы точки (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$, построенные по данным наблюдений, как можно ближе лежали вблизи прямой (4.3).

Назовем отклонением разность

$$Y_i - y_i,\ i = 1,2,\ldots,n,$$

где $Y_i -$ вычисленная по уравнению (4.3) ордината, соответствующая наблюдаемому значению x_i; $y_i -$ наблюдаемая ордината, соответствующая x_i, т.е. из пары (x_i, y_i).

Подберем параметры ρ_{yx} и b так, чтобы сумма квадратов отклонений была минимальной, т.е. используем для отыскания параметров метод наименьших квадратов.

Сумма квадратов отклонений есть функция F, которая зависит от отыскиваемых параметров и имеет вид

$$F(\rho_{yx}, b) = \sum_{i=1}^{n} (Y_i - y_i)^2$$

или

$$F(\rho_{yx}, b) = \sum_{i=1}^{n} \left(\rho_{yx} x_i + b - y_i\right)^2.$$

Для отыскания минимума приравняем нулю соответствующие частные производные:

$$\frac{\partial F}{\partial \rho_{yx}} = 2\sum_{i=1}^{n}\left(\rho_{yx}x_i + b - y_i\right)x_i = 0;$$

$$\frac{\partial F}{\partial b} = 2\sum_{i=1}^{n}\left(\rho_{yx}x_i + b - y_i\right) = 0.$$

Выполнив элементарные преобразования, получим систему двух линейных уравнений относительно ρ_{yx} и b

$$\left(\sum_{i=1}^{n}x_i^2\right)\rho_{yx} + \left(\sum_{i=1}^{n}x_i\right)b = \sum_{i=1}^{n}x_i y_i;\ \left(\sum_{i=1}^{n}x_i\right)\rho_{yx} + nb = \sum_{i=1}^{n}y_i.$$

Решив эту систему, найдем искомые параметры:

$$\rho_{yx} = \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i \sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}; \qquad (4.4)$$

$$b = \frac{\sum_{i=1}^{n}x_i^2 \sum_{i=1}^{n}y_i - \sum_{i=1}^{n}x_i \sum_{i=1}^{n}x_i y_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}. \qquad (4.5)$$

Аналогично можно найти выборочное уравнение прямой линии регрессии X на Y:

$$\bar{x}_y = \rho_{xy}y + c,$$

где ρ_{xy} – *выборочный коэффициент регрессии* X *на* Y.

Пример

Затраты X на развитие производства и Y – величина годовой прибыли фирмы в течение 5 лет (в у. е.) представлены в таблице:

X	6	3	7	5	10
Y	33	27	32	28	42

На величину прибыли влияют случайные факторы: $Y = \rho_{yx} x + b$.

Оценить параметры ρ_{yx} и b. Оценить годовую прибыль в том случае, если на развитие производства будет затрачено 12 у. е.

Решение. Вычислим неизвестные параметры по формулам (4.4) и (4.5):

$$\sum_{i=1}^{5} x_i y_i = 6 \cdot 33 + 3 \cdot 27 + 7 \cdot 32 + 5 \cdot 28 + 10 \cdot 42 = 1063,$$

$$\sum_{i=1}^{5} x_i = 31,\ \sum_{i=1}^{5} y_i = 162,\ \sum_{i=1}^{5} x_i^2 = 36 + 9 + 49 + 25 + 100 = 219,$$

$$\rho_{yx} = \frac{5 \cdot 1063 - 31 \cdot 162}{5 \cdot 219 - 31^2} = \frac{5315 - 5022}{1095 - 961} = \frac{293}{134} = 2{,}186,$$

$$b = \frac{219 \cdot 162 - 31 \cdot 1063}{5 \cdot 219 - 31^2} = \frac{35478 - 32953}{134} = \frac{2525}{134} = 18{,}843.$$

Итак, выборочное уравнение прямой линии регрессии Y на X имеет вид:

$$Y = 2{,}186 \cdot x + 18{,}843.$$

Для того, чтобы решить вторую часть задачи, надо вычислить значение Y при $x = 12$: $Y|_{x=12} = 2{,}186 \cdot 12 + 18{,}843 = 45{,}072 \approx 45$ (у. е.).

Корреляционная таблица

При большом числе наблюдений одно и тоже значение x может появляться n_x раз, а значение y может появляться n_y раз, пара $(x, y) - n_{xy}$ раз. В таких случаях данные наблюдений группируют, т.е. подсчитывают частоты n_x, n_y, n_{xy}. Все сгруппированные данные записывают в виде таблицы, которую называют *корреляционной*.

Рассмотрим пример:

Y	X				n_y
	10	20	30	40	
0,4	5	–	7	14	26
0,6	–	2	6	4	12
0,8	3	19	–	–	22
n_x	8	21	13	18	$n=60$

В первой строке таблицы указаны наблюдаемые значения признака X, т.е. $(10;20;30;40)$, а в первом столбце – наблюдаемые значения признака Y: $(0,4;0,6;0,8)$. На пересечении строк и столбцов вписаны частоты n_{xy} наблюдаемых значений пар признаков. Например, частота 5 указывает, что пара чисел (10; 0,4) наблюдалась 5 раз. Черточка означает, что соответствующая пара чисел не наблюдалась, например пара (20; 0,4).

В последнем столбце записаны суммы частот строк. Например, сумма частот первой строки равна $n_y = 5+7+14=26$. Это число указывает, что значение признака Y, равное 0,4 (в сочетании с различными значениями признака X) наблюдалось 26 раз. В последней строке записаны суммы частот столбцов.

В клетке, расположенной в нижнем правом углу таблицы, помещена сумма всех частот (общее число наблюдений n). Ясно, что $\sum n_x = \sum n_y = n$. В нашем случае

$$\sum n_x = 8+21+13+18=60,\ \sum n_y = 26+12+22=60.$$

Теперь рассмотрим случай, когда имеется большое число данных, и среди них есть повторяющиеся, и они сгруппированы в виде корреляционной таблицы. Как в этом случае строится уравнение прямой линии регрессии?

Ранее, для случая различных значений признаков, нами была получена система из двух уравнений, по которой определялись неизвестные коэффициенты ρ_{yx}, b. Запишем еще раз эту систему:

$$\begin{cases} \left(\sum x_i^2\right)\rho_{yx} + \left(\sum x_i\right)b = \sum x_i y_i; \\ \left(\sum x_i\right)\rho_{yx} + nb = \sum y_i. \end{cases}$$

Теперь запишем эту систему так, чтобы она отражала данные корреляционной таблицы. Для этого используем тождества:

$\sum x_i = n \cdot \bar{x}$, следствие из формулы $\bar{x} = \dfrac{\sum x_i}{n}$;

$\sum y_i = n \cdot \overline{y}$, следствие из формулы $\overline{y} = \dfrac{\sum y_i}{n}$;

$\sum x_i^2 = n \cdot \overline{x^2}$, следствие из формулы $\overline{x^2} = \dfrac{\sum x_i^2}{n}$;

$\sum x_i y_i = \sum n_{xy} x_i y_i$, здесь мы учли, что пара чисел (x, y) наблюдалась n_{xy} раз.

Следует обратить внимание, что в последнем равенстве, в выражении слева суммирование ведется по всем значениям x_i, y_i, а в выражении справа сумма только по различным значениям пар.

Подставив правые части тождеств в систему уравнений и сократив обе части второго уравнения на n, получим:

$$\begin{cases} \left(n \cdot \overline{x^2}\right)\rho_{yx} + \left(n \cdot \overline{x}\right)b = \sum n_{xy} x_i y_i; \\ \left(\overline{x}\right)\rho_{yx} + b = \overline{y}. \end{cases}$$

Решив эту систему, найдем неизвестные параметры и искомое уравнение

$$\overline{y} = \rho_{yx} \cdot x + b.$$

Однако в этом случае целесообразно ввести новую величину – коэффициент корреляции. Тогда искомое уравнение запишется в ином виде.

Найдем b из второго уравнения системы:

$$b = \overline{y} - \rho_{yx} \cdot \overline{x}.$$

Подставив правую часть этого равенства в уравнение прямой линии регрессии, получим

$$\overline{y}_x - \overline{y} = \rho_{yx}\left(x - \overline{x}\right). \qquad (4.6)$$

Выразим из системы коэффициент регрессии ρ_{yx}, учитывая при этом, что $\overline{x^2} - \left(\overline{x}\right)^2 = \sigma_x^2$. Получим

$$\rho_{yx} = \frac{\sum n_{xy} x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{n\left(\overline{x^2} - \left(\overline{x}\right)^2\right)} = \frac{\sum n_{xy} x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{n \cdot \sigma_x^2}.$$

Умножим обе части равенства на дробь $\frac{\sigma_x}{\sigma_y}$:

$$\rho_{yx} \cdot \frac{\sigma_x}{\sigma_y} = \frac{\sum n_{xy} x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{n \cdot \sigma_x \cdot \sigma_y}.$$

Обозначим правую часть равенства через r_B, и назовем ее *выборочным коэффициентом корреляции*:

$$r_B = \rho_{yx} \cdot \frac{\sigma_x}{\sigma_y}$$

или

$$\rho_{yx} = r_B \cdot \frac{\sigma_y}{\sigma_x}.$$

Подставив правую часть этого равенства в (4.6), получим выборочное уравнение прямой линии регрессии Y на X вида

$$\bar{y}_x - \bar{y} = r_B \frac{\sigma_y}{\sigma_x}\left(x - \bar{x}\right). \qquad (4.7)$$

Аналогично находится выборочное уравнение прямой линии регрессии X на Y:

$$\bar{x}_y - \bar{x} = r_B \frac{\sigma_x}{\sigma_y}\left(y - \bar{y}\right), \qquad (4.8)$$

где

$$r_B \frac{\sigma_x}{\sigma_y} = \rho_{xy}.$$

Приведем свойства выборочного коэффициента корреляции, из которых следует, что он служит для оценки тесноты корреляционной зависимости.

1. Абсолютная величина выборочного коэффициента корреляции по модулю не превосходит единицы, т.е.

$$|r_B| \le 1.$$

2. Если выборочный коэффициент корреляции равен нулю и выборочные линии регрессии – прямые, то X и Y не связаны линейной корреляционной зависимостью.

3. Если абсолютная величина выборочного коэффициента корреляции равна единице, то наблюдаемые значения признаков связаны линейной функциональной зависимостью.

4. С возрастанием абсолютной величины выборочного коэффициента корреляции зависимость становится более тесной и при $|r_B|=1$ переходит в функциональную.

Замечание. Так как σ_x и σ_y положительные числа, то знак выборочного коэффициента регрессии всегда совпадает со знаком углового коэффициента в уравнении прямой линии регрессии.

Примеры решения задач к главе 4

1. Затраты X на развитие производства и $Y-$ величина годовой прибыли фирмы в течение 5 лет (в у. е.) представлены в таблице

X	6	3	7	5	10
Y	33	27	32	28	42

На величину прибыли влияют случайные факторы $Y=\rho_{yx}x+b$. Оценить параметры ρ_{yx} и b. Оценить годовую прибыль в том случае, если на развитие производства будет затрачено 12 у. е.

Решение. Для вычисления неизвестных параметров будем использовать формулы

$$\rho_{yx}=\frac{n\sum\limits_{i=1}^{n}x_iy_i-\sum\limits_{i=1}^{n}x_i\sum\limits_{i=1}^{n}y_i}{n\sum\limits_{i=1}^{n}x_i^2-\left(\sum\limits_{i=1}^{n}x_i\right)^2} \quad \text{и} \quad b=\frac{\sum\limits_{i=1}^{n}x_i^2\sum\limits_{i=1}^{n}y_i-\sum\limits_{i=1}^{n}x_i\sum\limits_{i=1}^{n}x_iy_i}{n\sum\limits_{i=1}^{n}x_i^2-\left(\sum\limits_{i=1}^{n}x_i\right)^2}.$$

Найдем

$$\sum_{i=1}^{5}x_iy_i=6\cdot33+3\cdot27+7\cdot32+5\cdot28+10\cdot42=1063,$$

$$\sum_{i=1}^{5}x_i=31,\ \sum_{i=1}^{5}y_i=162,\ \sum_{i=1}^{5}x_i^2=36+9+49+25+100=219,$$

$$\rho_{yx} = \frac{5 \cdot 1063 - 31 \cdot 162}{5 \cdot 219 - 31^2} = \frac{5315 - 5022}{1095 - 961} = \frac{293}{134} = 2{,}186,$$

$$b = \frac{219 \cdot 162 - 31 \cdot 1063}{5 \cdot 219 - 31^2} = \frac{35478 - 32953}{134} = \frac{2525}{134} = 18{,}843.$$

Итак, выборочное уравнение прямой линии регрессии Y на X имеет вид

$$Y = 2{,}186 \cdot x + 18{,}843.$$

Для того, чтобы решить вторую часть задачи, надо вычислить значение Y при $x = 12$:

$$Y|_{x=12} = 2{,}186 \cdot 12 + 18{,}843 = 45{,}072 \approx 45 \text{ (у. е.)}.$$

Ответ. Уравнение линии регрессии Y на X : $Y = 2{,}186 \cdot x + 18{,}843$.

2. В таблице приведены результаты $n = 11$ измерений значений x_i отклонений от номиналов длины моделей (признак X) и от ширины моделей y_i (признак Y):

№	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 $Y = \rho_{yx} x + b$.

Требуется

1) оценить параметры ρ_{yx} и b методом наименьших квадратов;

2) пользуясь эмпирическим уравнением регрессии, найти точечную оценку отклонения от номинального размера ширины модели, если длина модели отклоняется от номинального размера на величину $x^* = 1{,}1$ мм;

3) вычислить коэффициент корреляции. Объяснить смысл полученного результата.

Решение

1) Сначала вычислим вспомогательные суммы и выборочные средние. Находим

$$\sum_{i=1}^{n} x_i = 12,93,\ \sum_{i=1}^{n} y_i = 0,88,\ \sum_{i=1}^{n} x_i^2 = 15,6411,\ \sum_{i=1}^{n} y_i^2 = 1,8856;$$

$$\sum_{i=1}^{n} x_i y_i = 1,7193,\quad \bar{x} = \frac{1}{11}\cdot 12,93 = 1,175,\quad \bar{y} = \frac{1}{11}\cdot 0,88 = 0,08.$$

Найдем теперь оценки коэффициентов уравнения регрессии по формулам (4.4) и (4.5):

$$\rho_{yx} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{11\cdot 1,7193 - 12,93\cdot 0,88}{11\cdot 15,6411 - 12,93^2} = 1,5479,$$

$$b = \bar{y} - \rho_{yx}\bar{x} = 0,08 - 1,5479\cdot 1,175 = -1,7394.$$

Следовательно, эмпирическое уравнение регрессии имеет вид

$$Y = 1,5479x - 1,7394.$$

График построенного эмпирического уравнения линейной регрессии приведен на рис. 4.1.

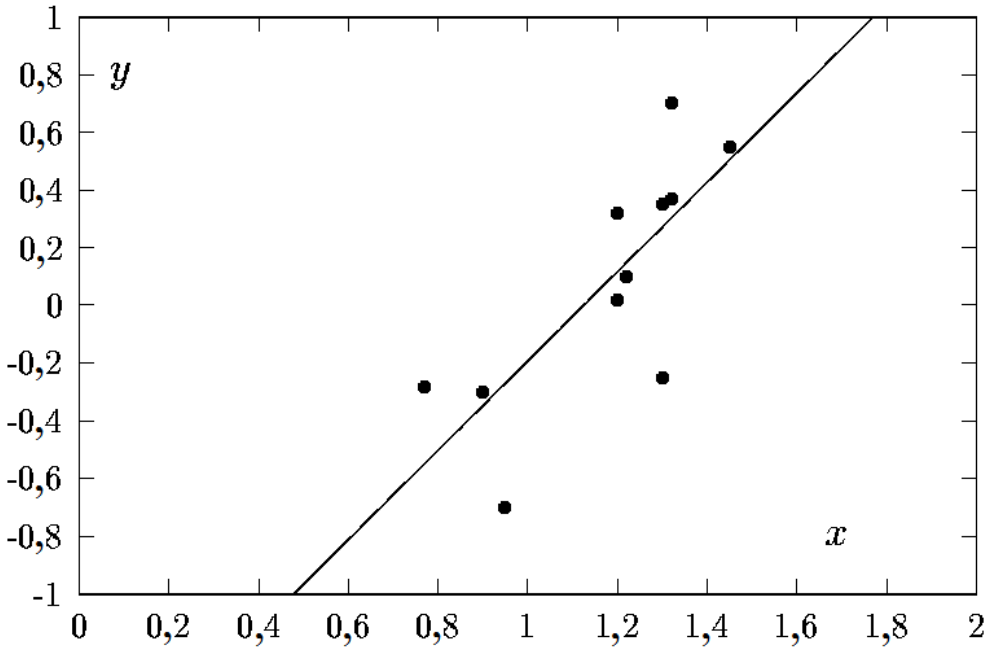


Рис. 4.1

2) Найдем точечную оценку отклонения от номинального размера ширины модели, если длина модели отклоняется от номинального размера на величину $x^* = 1{,}1$ мм:

$$\bar{y}^* = 1{,}5479 \cdot 1{,}1 - 1{,}7394 = -0{,}036 \text{ мм}.$$

3) Вычислим коэффициент корреляции по формуле

$$r_B = \rho_{yx} \cdot \frac{\sigma_x}{\sigma_y} = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{n \cdot \sigma_x \cdot \sigma_y} = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \cdot \sqrt{n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2}} =$$

$$= \frac{11 \cdot 1{,}7193 - 12{,}93 \cdot 0{,}88}{\sqrt{11 \cdot 15{,}6411 - 12{,}93^2} \sqrt{11 \cdot 1{,}8856 - 0{,}88^2}} = 0{,}7534 .$$

Из полученного результата мы видим, что связь между наблюдаемыми значениями признаков и линейной корреляционной зависимостью не достаточно высока. Это может быть вызвано либо случайными ошибками эксперимента, либо тем, что линейная регрессионная модель плохо согласуется с экспериментальными данными.

Ответ. $Y = 1{,}5479x - 1{,}7394$, $\bar{y}^* = -0{,}036$, $r_B = 0{,}7534$.

3. В таблице приведены результаты $n = 50$ измерений значений x_i признака X и значений y_i признака Y:

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

Используя значения, приведенные в данной таблице, требуется:

1) составить корреляционную таблицу;

2) найти по данным корреляционной таблицы числовые характеристики выборки $\bar{x}, \bar{y}, \sigma_x, \sigma_y, K_{xy}, r_B$;

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 20. После этого подсчитываем количество экспериментальных точек, попадающих в прямоугольники, образованные границами интервалов. В результате получаем корреляционную таблицу, в которой отмечены середины соответствующих интервалов:

Y	X						
	40	60	80	100	120	140	n_y
160	-	-	-	-	1	3	4
140	-	-	-	3	5	-	8
120	-	-	4	8	1	-	13
100	-	2	7	4	-	-	13
80	1	3	3	-	-	-	7
60	4	1	-	-	-	-	5
n_x	5	6	14	15	7	3	$n=50$

2) Для определения характеристик искомых эмпирических уравнений регрессии найдем:

средние арифметические

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 88{,}80, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 109{,}60,$$

выборочные дисперсии

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 690{,}56, \quad \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 771{,}84,$$

выборочные средние квадратические отклонения

$$\sigma_x = \sqrt{690{,}56} = 26{,}28, \quad \sigma_y = \sqrt{771{,}84} = 27{,}78,$$

выборочный корреляционный момент

$$K_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\cdot\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y} = 651{,}52,$$

выборочный коэффициент корреляции

$$r_B = \frac{\sum_{i=1}^{n} x_i y_i - n\cdot\bar{x}\cdot\bar{y}}{n\cdot\sigma_x\cdot\sigma_y} = \frac{K_{xy}}{\sigma_x\cdot\sigma_y} = \frac{651{,}52}{26{,}28\cdot 27{,}78} = 0{,}89.$$

Поскольку выборочный коэффициент корреляции получился близок к единице, то можно сделать вывод, что линейная регрессионная модель 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 $\{(x_i, y_i)\}$, $(i = 1,2,...,50)$ в виде точек в декартовой системе координат (рис. 4.2).

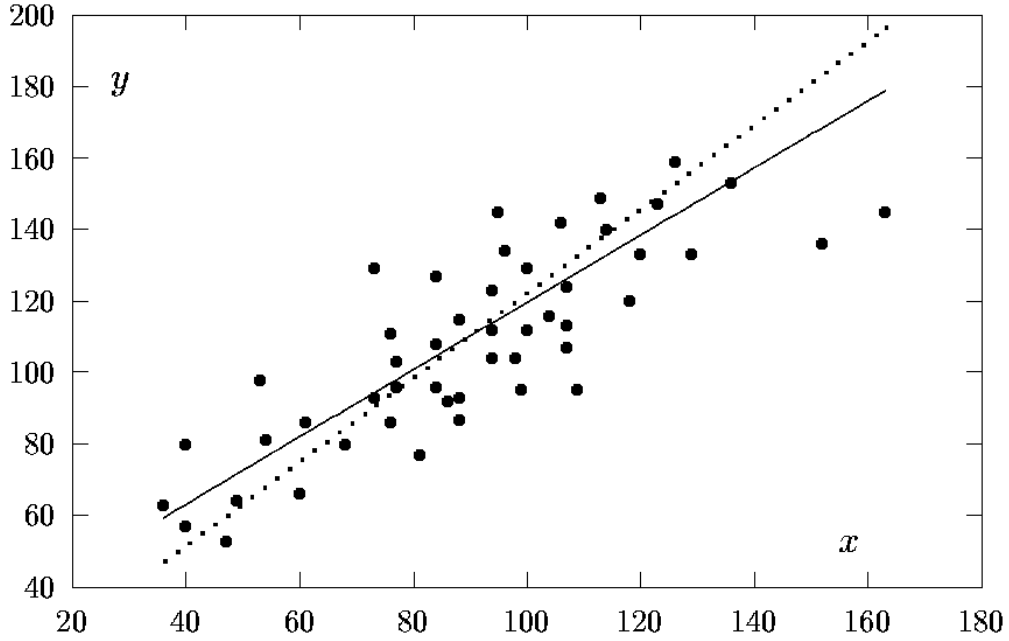


Рис. 4.2

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

$$\bar{y}_x = \bar{y} + r_B \frac{\sigma_y}{\sigma_x}\left(x - \bar{x}\right) = 109,60 + 0,89\frac{27,78}{26,28}(x - 88,80) = 25,68 + 0,94x,$$

значения параметров эмпирического уравнения регрессии признака X на признак Y

$$\bar{x}_y = \bar{x} + r_B \frac{\sigma_x}{\sigma_y}\left(y - \bar{y}\right) = 88,80 + 0,89\frac{26,28}{27,78}(y - 109,60) = -3,81 + 0,85y.$$

Графики найденных эмпирических функций линейной регрессии нанесены на рис. 4.2. Сплошной линией указано уравнение $\bar{y}_x = 25,68 + 0,94x$, пунктирной линией указано уравнение регрессии $\bar{x}_y = -3,81 + 0,85y$.

Ответ. $Y = \bar{y}_x = 25,68 + 0,94x$, $X = \bar{x}_y = -3,81 + 0,85y$.

Задачи для самостоятельного решения

1. Найти выборочные уравнения прямых линий регрессии Y на X и X на Y по данным, приведенным в следующих корреляционных таблицах:

а)

Y	X								n_y
	5	10	15	20	25	30	35	40	
100	2	1	–	–	–	–	–	–	3
120	3	4	3	–	–	–	–	–	10
140	–	–	5	10	8	–	–	–	23
160	–	–	–	–	–	6	1	1	9
180	–	–	–	–	–	–	4	1	5
n_x	5	5	8	11	8	6	5	2	$n=50$

б)

Y	X							n_y
	18	23	28	33	38	43	48	
125	–	1	–	–	–	–	–	1
150	1	2	5	–	–	–	–	8
175	–	3	2	12	–	–	–	17
200	–	–	1	8	7	–	–	16
225	–	–	–	–	3	3	–	6
250	–	–	–	–	–	1	1	2
n_x	1	6	8	20	10	4	1	$n=50$

2. Затраты x на развитие производства и y – величина годовой прибыли в течение пяти лет представлены (в у.е.) в таблице:

x	6	3	7	5	10
y	33	27	32	28	42

Предполагается, что имеет место линейная зависимость между затратами и прибылью. Найти уравнение зависимости и оценить годовую прибыль в том случае, если на развитие производства будет затрачено 12 у.е.

3. Распределение 50 гастрономических магазинов по уровню издержек обращения X (%) и годовому товарообороту Y (млн руб.) представлено в таблице:

X	Y					
	0,5–2	2–3,5	3,5–5	5–6,5	6,5–8	Итого
4–6	–	–	–	3	2	5
6–8	–	4	8	8	1	21
8–10	2	5	5	2	–	14
10–12	3	1	3	–	–	9
12–14	1	–	–	–	–	1
Итого	6	10	18	13	3	$n=50$

Построить уравнение линейной регрессии Y на X и оценить тесноту связи между переменными с помощью коэффициента корреляции r_{yx}.

4. Годовые прибыли фирмы (в тыс. долларов) за пять лет представлены в следующей таблице:

Год	1	2	3	4	5
Прибыль	99	112	120	135	144

Провести линейную регрессию и дать прогноз на следующий год.

5. 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.

6. Найти уравнения регрессии Y на X и X на Y по четырем парам наблюдаемых значений случайной величины (X,Y):

x_i	1	2	3	4
y_i	2	4	5	7

7. Составить корреляционную таблицу и найти линейные уравнения регрессии Y на X и X на Y по 10 парам наблюдаемых значений случайной величины (X,Y):

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

Ответы.

1) а) $\bar{y}_x = 1{,}92x + 100{,}9$; $\bar{x}_y = 0{,}42y - 38{,}3$; **б)** $\bar{y}_x = 3{,}69x + 66$; $\bar{x}_y = 0{,}19y - 3{,}1$; **3)** $\bar{y}_x = 8{,}465 - 0{,}525x$, $r_{yx} = -0{,}619$; **4)** $\bar{y}_x = 88{,}1 + 11{,}3 \cdot x$; $y(6) = 155{,}9$; **5)** $a = 0{,}87, b = 67{,}5$; **6)** $Y = 1{,}6x + 0{,}5$; $X = 0{,}615y - 0{,}269$; **7)** $Y = 1{,}25x + 1{,}75$; $X = 0{,}73y - 1{,}08$.

Контрольные вопросы

1. Какая зависимость называется корреляционной?
2. Дайте определение условного среднего.
3. Дайте определение уравнения регрессии.
4. Как составляется корреляционная таблица?
5. Как определяется выборочный коэффициент регрессии?
6. По какой формуле вычисляется выборочный коэффициент корреляции?
7. Напишите уравнения прямых линий регрессии Y на X и X на Y.
8. Что характеризует выборочный коэффициент корреляции?

Глава 5

МОДЕЛИРОВАНИЕ СЛУЧАЙНЫХ ВЕЛИЧИН

Датой рождения метода Монте – Карло принято считать 1949 г., когда американские ученые Николас Метрополис и Станислав Улам опубликовали статью «Метод Монте – Карло», в которой систематически его изложили. Название метода связано с названием города Монте – Карло, где в казино играют в рулетку. Рулетка – одно из простейших устройств для получения случайных чисел, на использовании которых основан данный метод.

Сущность метода Монте – Карло состоит в следующем: требуется найти значение a некоторой изучаемой величины. Для этого выбирают такую случайную величину X, математическое ожидание которой равно a.

Практически же поступают так: производят n испытаний, в результате которых получают n возможных значений X, вычисляют их среднее арифметическое $\overline{x}$ и принимают получившееся число в качестве оценки искомого числа a.

Поскольку метод Монте – Карло требует проведения большого числа испытаний, его часто называют *методом статистических испытаний*. Теория этого метода указывает, как наиболее целесообразно выбрать случайную величину X, как найти ее возможные значения.

Отыскание возможных значений случайной величины X (моделирование) называют *разыгрыванием случайной величины*.

Случайные числа

Ранее было указано, что метод Монте – Карло основан на применении случайных чисел. Обозначим через R непрерывную случайную величину, распределенную равномерно в интервале $(0,1)$.

Случайными числами называют возможные значения r_i непрерывной случайной величины R, распределенной равномерно в интервале $(0,1)$.

В действительности пользуются не равномерно распределенной случайной величиной R, возможные значения которой, вообще говоря, имеют бесконечное число знаков после запятой, а квазиравномерной случайной величиной $R*$, возможные значения которой имеют конечное число знаков после запятой. В результате замены R на $R*$ разыгрываемая величина имеет не точно, а приближенно заданное распределение.

Разыгрывание дискретной случайной величины

Пусть требуется разыграть дискретную случайную величину X, т.е. получить последовательность ее возможных значений $x_1, x_2, ..., x_n$, зная закон распределения X:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Обозначим через R непрерывную случайную величину, распределенную равномерно в интервале $(0,1)$, а через r_i – ее возможные значения, т.е. случайные числа.

Разобьем интервал $0 \le R < 1$ на оси Or точками с координатами

$$p_1, p_1 + p_2, p_1 + p_2 + p_3, ..., p_1 + ... + p_{n-1}$$

на n частичных интервалов $\Delta_1, \Delta_2, ..., \Delta_n$. Длины этих интервалов равны:

$$|\Delta_1| = p_1 - 0 = p_1, |\Delta_2| = (p_1 + p_2) - p_1 = p_2, ..., |\Delta_n| = 1 - (p_1 + ... + p_{n-1}) = p_n.$$

Видно, что длина каждого частичного интервала Δ_i равна соответствующей вероятности p_i.

Теорема. Если каждому случайному числу r_j $(0 \le r_j < 1)$, которое попало в интервал Δ_i, ставить в соответствие возможное значение x_i случайной величины X, то разыгрываемая величина будет иметь заданный закон распределения:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Доказательство. Так как при попадании случайного числа r_j в частичный интервал Δ_i разыгрываемая величина принимает значение x_i, а таких интервалов всего n, то разыгрываемая величина имеет те же возможные значения, что и X, а именно $x_1, x_2, ..., x_n$.

Вероятность попадания случайной величины R в интервал Δ_i равна его длине (вероятность равна отношению длины интервала Δ_i к длине интервала $(0,1)$, но так как длина всего интервала равна 1, получаем, что вероятность равна длине Δ_i). Мы вычислили, что длина каждого интервала равна соответствующей вероятности. Следовательно, вероятность того, что разыгрываемая величина примет значение x_i равна p_i (поскольку мы условились в случае попадания случайного числа r_j в интервал Δ_i считать, что разыгрываемая величина приняла значение x_i). Итак, разыгрываемая величина имеет заданный закон распределения.

Таким образом, мы приходим к **правилу 1**:

Для того чтобы разыграть дискретную случайную величину, заданную законом распределения

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

надо: 1) разбить интервал $(0,1)$ на n частичных интервалов

$$\Delta_1 = (0, p_1),\ \Delta_2 = (p_1, p_1 + p_2),\ldots,\ \Delta_n = (p_1 + p_2 + \ldots + p_{n-1}, 1);$$

2) выбрать из таблицы случайных чисел (прил. 5) случайное число r_j.

Если r_j попало в частичный интервал Δ_i, то разыгрываемая дискретная случайная величина приняла возможное значение x_i.

Разыгрывание противоположных событий

Разыгрывание случайных событий можно свести к разыгрыванию дискретной случайной величины.

Пусть требуется разыграть испытания, в каждом из которых событие A появляется с вероятностью p.

Введем в рассмотрение дискретную случайную величину X с двумя возможными значениями $(x_1 = 0, x_2 = 1)$ и соответствующими им вероятностями $p_1 = q = 1 - p,\ p_2 = p$. Условимся считать, что если в испытании величина X приняла значение $x_1 = 0$, то событие A не наступило, т.е. произошло противоположное событие $\bar{A}$, если же X приняла значение $x_2 = 1$, то произошло событие A.

Таким образом, разыгрывание противоположных событий A и $\bar{A}$ сведено к разыгрыванию дискретной случайной величины X, имеющей следующий закон распределения:

X	0	1
p	q	p

Для разыгрывания X надо интервал $(0,1)$ разбить точкой q на два частичных интервала: $\Delta_1 = (0, q),\ \Delta_2 = (q, 1)$. Затем выбирают случайное число r_j. Если r_j попадает в интервал Δ_1, то $X = x_1 = 0$ и, следовательно, наступило событие $\bar{A}$, если же $X = x_2 = 1$, то произошло событие A.

Разыгрывание полной группы событий

Разыгрывание полной группы n несовместных событий $A_1, A_2, ..., A_n$, вероятности которых $p_1, p_2, ..., p_n$ известны, можно свести к разыгрыванию дискретной случайной величины X. Пусть случайная величина X принимает значения $x_1 = 1,\ x_2 = 2, \ldots,\ x_n = n$ с вероятностями $p_1, p_2, ..., p_n$ соответственно. Случайная величина связывается с полной группой событий следующим образом: полагаем, если в испытании случайная величина приняла значение $x_i = i$, то наступило событие A_i. Справедливость этого утверждения следует из того, что число возможных значений случайной величины равно числу событий полной группы и при этом вероятности возможных значений x_i и соответствующих событий одинаковы: $P(X = x_i) = P(A_i) = p_i$. Таким образом, появление в испытании события A_i равносильно событию, состоящему в том, что дискретная случайная величина X приняла значение x_i.

Правило 2. Для того, чтобы разыграть испытания, в каждом из которых наступает одно из событий $A_1, A_2, ..., A_n$ полной группы, вероятности которых $p_1, p_2, ..., p_n$ известны, достаточно разыграть (по правилу 1) дискретную случайную величину X со следующим законом распределения:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Если в испытании величина X приняла возможное значение $x_i = i$, то наступило событие A_i.

Разыгрывание непрерывной случайной величины

Метод обратных функций

Пусть требуется разыграть непрерывную случайную величину X, т.е. получить последовательность ее возможных значений $x_i\ (i = 1, 2, ..., n)$, зная функцию распределения $F_X(x)$.

Теорема. Если $r_i -$ случайное число, то возможное значение x_i разыгрываемой непрерывной случайной величины X с заданной функцией распределения $F_X(x)$, соответствующее r_i, является корнем уравнения:

$$F_X(x_i) = r_i.$$

Эту теорему мы примем без доказательства.

Итак, получим **правило 3.**

Для того, чтобы найти возможное значение x_i непрерывной случайной величины X, зная ее функцию распределения $F_X(x)$, надо выбрать случайное число r_i, приравнять его к функции распределения и решить относительно x_i полученное уравнение: $F_X(x_i) = r_i$.

Пример. Непрерывная случайная величина X распределена по показательному закону, заданному функцией распределения

$$F_X(x) = 1 - e^{-6x} \ (x > 0).$$

Требуется найти явную формулу для разыгрывания возможных значений X.

Решение. Используя правило, выпишем уравнение:

$$1 - e^{-6x_i} = r_i.$$

Решим это уравнение относительно x_i: $e^{-6x_i} = 1 - r_i$, прологарифмируем обе части равенства, получим $-6x_i = \ln(1 - r_i)$. Отсюда

$$x_i = -\frac{1}{6}\ln(1 - r_i).$$

Правило 4. Для того, чтобы найти возможное значение x_i непрерывной случайной величины X, зная ее плотность вероятности $f(x)$, надо выбрать случайное число r_i и решить относительно x_i уравнение

$$\int_{-\infty}^{x_i} f(x)\,dx = r_i$$

или уравнение

$$\int_{a}^{x_i} f(x)\,dx = r_i,$$

где $a -$ наименьшее конечное возможное значение X.

Пример. Задана плотность вероятностей непрерывной случайной величины X

$$f(x)=\begin{cases}5\left(1-\dfrac{5x}{2}\right),\ x\in\left(0,\dfrac{2}{5}\right);\\ 0,\ x\notin\left(0,\dfrac{2}{5}\right).\end{cases}$$

Требуется найти явную формулу для разыгрывания возможных значений X.

Решение. Составим уравнение

$$5\int_0^{x_i}\left(1-\frac{5x}{2}\right)dx=r_i.$$

Вычислим определенный интеграл, получим равенство

$$5\left(x_i-\frac{5x_i^2}{4}\right)=5x_i-\frac{25x_i^2}{4}=r_i.$$

Чтобы получить явную формулу для x_i, необходимо решить квадратное уравнение $25x_i^2-20x_i+4r_i=0$. Решив уравнение, получим два корня:

$$x_{i,1}=\frac{2\left(1-\sqrt{1-r_i}\right)}{5},\ x_{i,2}=\frac{2\left(1+\sqrt{1-r_i}\right)}{5}.$$

Второй корень $x_{i,2}\notin(0,2/5)$, поэтому его исключаем. Таким образом, явная формула для разыгрывания значений X имеет вид

$$x_i=\frac{2\left(1-\sqrt{1-r_i}\right)}{5}.$$

Метод суперпозиции

Пусть функция распределения разыгрываемой случайной величины X может быть представлена в виде линейной комбинации двух функций распределения:

$$F_X(x)=C_1\cdot F_1(x)+C_2\cdot F_2(x)\ \ (C_1>0,C_2>0).$$

При $x \to +\infty$ каждая из функций распределения стремится к единице, поэтому $C_1 + C_2 = 1$.

Введем вспомогательную дискретную случайную величину Z со следующим законом распределения:

Z	1	2
p	C_1	C_2

Мы видим, что $P(Z=1) = C_1$, $P(Z=2) = C_2$.

Выберем два независимых случайных числа r_1 и r_2. По числу r_1 разыгрываем возможное значение Z (используя правило 1). Если окажется, что $Z = 1$, то возможное значение X ищут из уравнения $F_1(x) = r_2$; если $Z = 2$, то решают относительно x уравнение $F_2(x) = r_2$.

Докажем, что функция распределения разыгрываемой случайной величины равна заданной функции распределения $F_X(x)$. Для доказательства воспользуемся формулой полной вероятности

$$P(A) = P(H_1) \cdot P\left(A|_{H_1}\right) + P(H_2) \cdot P\left(A|_{H_2}\right).$$

Обозначим через A событие $\{X < x\}$, тогда $P(A) = P(X < x) = F_X(x)$. Рассмотрим гипотезы: $H_1 -$ случайная величина Z приняла значение $Z = 1$, $H_2 -$ случайная величина Z приняла значение $Z = 2$. По определению Z, вероятности этих гипотез равны:

$$P(H_1) = P(Z=1) = C_1,\ P(H_2) = P(Z=2) = C_2.$$

Условные вероятности появления события A соответственно равны:

$$P\left(A|_{H_1}\right) = P\left(\{X < x\}|_{H_1}\right) = F_1(x),\ P\left(A|_{H_2}\right) = P\left(\{X < x\}|_{H_2}\right) = F_2(x).$$

Подставив полученные вероятности в формулу полной вероятности, получим $F_X(x) = C_1 \cdot F_1(x) + C_2 \cdot F_2(x)$, что и требовалось доказать.

Правило 5. Для того, чтобы разыграть возможное значение случайной величины X, функция распределения которой представима в виде

$$F_X(x) = C_1 \cdot F_1(x) + C_2 \cdot F_2(x),$$

где $C_1 > 0, C_2 > 0$ и $C_1 + C_2 = 1$, надо выбрать два независимых случайных числа r_1, r_2 и по случайному числу r_1 разыграть возможное значение

вспомогательной дискретной случайной величины Z, имеющей закон распределения

Z	1	2
p	C_1	C_2

Если окажется, что $Z=1$, то решают относительно x уравнение $F_1(x)=r_2$; если $Z=2$, то решают уравнение $F_2(x)=r_2$.

Примеры решения задач к главе 5

1. Разыграть 6 возможных значений дискретной случайной величины X, закон распределения которой задан в виде таблицы:

X	2	10	18
p	0,22	0,17	0,61

Решение. Разобьем интервал $(0,1)$ оси Or точками с координатами $0{,}22;\ 0{,}22+0{,}17=0{,}39$ на три частичных интервала:

$$\Delta_1=(0;0{,}22),\ \Delta_2=(0{,}22;0{,}39),\ \Delta_3=(0{,}39;1).$$

Выпишем из таблицы приложения 5 шесть случайных чисел, например, $0{,}32;\ 0{,}17;\ 0{,}9;\ 0{,}05;\ 0{,}97;\ 0{,}87$ (пятая строка таблицы из прил. 5 снизу).

Случайное число $r_1=0{,}32$ принадлежит частичному интервалу Δ_2, поэтому разыгрываемая случайная величина приняла возможное значение $x_2=10$. Случайное число $r_2=0{,}17$ принадлежит частичному интервалу Δ_1, поэтому разыгрываемая величина приняла значение $x_1=2$.

Аналогично получаем остальные возможные значения:

$$r_3=0{,}9\in\Delta_3, x_3=18;\ r_4=0{,}05\in\Delta_1, x_1=2;$$

$$r_5=0{,}97\in\Delta_3, x_3=18; r_6=0{,}87\in\Delta_3, x_3=18.$$

Итак, разыгранные возможные значения таковы: 10, 2, 18, 2, 18, 18.

2. Разыграть шесть испытаний, в каждом из которых событие A появляется с вероятностью $p=0{,}35$.

Решение. Выберем из таблицы приложения 5 шесть случайных чисел, например, 0,1; 0,36; 0,08; 0,99; 0,12; 0,06. Считая, что при $r_j < 0{,}35$ событие A появилось, а при $r_j \geq 0{,}35$ наступило противоположное событие $\bar{A}$, получим искомую последовательность событий: $A, \bar{A}, A, \bar{A}, A, A$.

3. Разыграть пять опытов по схеме Бернулли: опыт состоит из трех независимых испытаний, в каждом из которых вероятность появления события A равна 0,4.

Решение. Сначала составим закон распределения дискретной случайной величины X – числа появлений события A в трех независимых испытаниях, если в каждом испытании вероятность появления события A равна 0,4.

Случайная величина X может принимать четыре значения: 0, 1, 2, 3. Найдем вероятности этих значений. Вероятность $P(X=0)$ – это вероятность того, что в трех опытах событие A ни разу не произойдет,

$$P(X=0) = (1-0{,}4)^3 = 0{,}6^3 = 0{,}216.$$

Вероятность $P(X=1)$ – вероятность того, что в трех испытаниях событие A произойдет один раз. Вычисляем $P(X=1)$ по формуле Бернулли:

$$P(X=1) = P_3(1) = C_3^1 \cdot 0{,}4 \cdot 0{,}6^2 = 0{,}432.$$

Аналогично, по формуле Бернулли, вычисляем вероятность $P(X=2)$ того, что событие A произойдет в двух испытаниях из трех:

$$P(X=2) = P_3(2) = C_3^2 \cdot 0{,}4^2 \cdot 0{,}6 = 0{,}288,$$

и вероятность $P(X=3)$ того, что событие A произойдет во всех трех испытаниях:

$$P(X=3) = 0{,}4^3 = 0{,}064.$$

Таким образом, получим распределение случайной величины X:

X	0	1	2	3
p	0,216	0,432	0,288	0,064

Теперь нам необходимо разыграть случайную величину X. Для этого интервал $(0,1)$ разбиваем на частичные интервалы точками 0,216; $0,216+0,432=0,648$; $0,648+0,288=0,936$. Получим четыре частичных интервала:

$$\Delta_1=(0;0,216),\ \Delta_2=(0,216;0,648),\ \Delta_3=(0,648;0,934),\ \Delta_4=(0,934;1).$$

Выпишем из таблицы приложения 5 пять случайных чисел: 0,945; 0,572; 0,857; 0,367; 0,897. Первое случайное число 0,945 попадает в интервал Δ_4, значит случайная величина X примет значение $x_4=3$. Случайное число 0,572 попадает в интервал Δ_2, следовательно, имеем значение $x_2=1$. Аналогично получим остальные значения величины X.

Итак, разыгранные возможные значения таковы: 3, 1, 2, 1, 2.

4. Заданы вероятности трех событий A_1, A_2, A_3, образующих полную группу: $p_1=P(A_1)=0,22$, $p_2=P(A_2)=0,31$, $p_3=P(A_3)=0,47$. Разыграть пять испытаний, в каждом из которых появляется одно из трех рассматриваемых событий.

Решение. Разыгрывание полной группы событий можно свести к разыгрыванию дискретной случайной величины X, которая имеет следующий закон распределения:

X	1	2	3
p	0,22	0,31	0,47

Полагаем, если величина X приняла значение $x_1=1$, то наступило событие A_1. Если величина приняла значение $x_2=2$, то наступило событие A_2, если $x_3=3$, то – A_3.

Теперь интервал $(0,1)$ разбиваем на три частичных интервала:

$$\Delta_1=(0;0,22),\ \Delta_2=(0,22;0,53),\ \Delta_3=(0,53;1).$$

Выберем из таблицы приложения 5 пять случайных чисел: 0,61; 0,19; 0,69; 0,04; 0,46.

Случайное число 0,61 принадлежит интервалу Δ_3, поэтому величина X приняла значение $x_3=3$ и, следовательно, наступило событие A_3. Аналогично находим остальные события. В итоге получим искомую последовательность событий: A_3, A_1, A_3, A_1, A_2.

5. События A и B независимы и совместны. Разыграть четыре испытания, в каждом из которых вероятность появления события A равна 0,7, а события B – 0,4.

Решение. Возможны четыре исхода испытания:

$A_1 = AB$, причем в силу независимости событий

$$P(AB) = P(A)P(B) = 0,7 \cdot 0,4 = 0,28;$$

$A_2 = A\bar{B}$, причем $P(A\bar{B}) = 0,7 \cdot 0,6 = 0,42$;

$A_3 = \bar{A}B$, причем $P(\bar{A}B) = 0,3 \cdot 0,4 = 0,12$;

$A_4 = \bar{A}\bar{B}$, причем $P(\bar{A}\bar{B}) = 0,3 \cdot 0,6 = 0,18$.

Таким образом, задача сведена к разыгрыванию полной группы четырех событий: A_1 с вероятностью $p_1 = 0,28$, A_2 с вероятностью $p_2 = 0,42$, A_3 с вероятностью $p_3 = 0,12$, A_4 с вероятностью $p_4 = 0,18$.

Эта задача, в свою очередь, сводится к разыгрыванию дискретной случайной величины X с законом распределения (см. пример 4)

X	1	2	3	4
p	0,28	0,42	0,12	0,18

Выберем из таблицы приложения 5 четыре случайных числа: 0,32; 0,17; 0,9; 0,05. Интервал $(0,1)$ разбиваем на четыре частичных интервала:

$$\Delta_1 = (0;0,28),\ \Delta_2 = (0,28;0,7),\ \Delta_3 = (0,7;0,82),\ \Delta_4 = (0,82;1).$$

Случайное число 0,32 принадлежит интервалу Δ_2, следовательно, $X = 2$ и произошло событие $A_2 = A\bar{B}$; $0,17 \in \Delta_1$, следовательно, $X = 1$ и произошло событие $A_1 = AB$. Аналогично находим остальные события. В итоге получим следующую последовательность: $A\bar{B}$, AB, $\bar{A}\bar{B}$, AB.

6. Найти явную формулу для разыгрывания непрерывной случайной величины X, распределенной равномерно в интервале $(2,9)$, зная ее функцию распределения $F_X(x) = (x-2)/(9-2) = (x-2)/7$ $(2 < x < 9)$.

Решение. Приравняем заданную функцию распределения к случайному числу r_i: $F_X(x_i) = r_i$, получим равенство

$$\frac{x_i - 2}{7} = r_i.$$

Решив это уравнение относительно x_i, получим явную формулу для разыгрывания возможных значений случайной величины X:

$$x_i = 7r_i + 2.$$

7. Разыграть три возможных значения непрерывной случайной величины X, распределенной равномерно в интервале $(2,10)$.

Решение. Напишем функцию распределения величины X, распределенной равномерно на интервале (a,b):

$$F_X(x) = \frac{x-a}{b-a}.$$

По условию $a = 2,\ b = 10$, следовательно,

$$F_X(x) = \frac{x-2}{8}.$$

Напишем уравнение для отыскания возможных значений x_i, для чего приравняем функцию распределения к случайному числу:

$$\frac{x_i - 2}{8} = r_i.$$

Отсюда $x_i = 8r_i + 2$.

Теперь выберем три случайных числа, например, $r_1 = 0{,}11$, $r_2 = 0{,}17$, $r_3 = 0{,}66$. Подставим эти числа в уравнение, разрешенное относительно x_i, в итоге получим соответствующие возможные значения величины X:

$$x_1 = 8 \cdot 0{,}11 + 2 = 2{,}88;\ x_2 = 8 \cdot 0{,}17 + 2 = 1{,}36;\ x_3 = 8 \cdot 0{,}66 + 2 = 7{,}28.$$

8. Найти явную формулу для разыгрывания непрерывной случайной величины, заданной плотностью вероятности $f(x) = b/(1+ax)^2$ в интервале $[0,1/(b-a)]$; вне интервала $f(x) = 0$.

Решение. Известно, что

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

В частности,

$$F_X(x_i) = \int_{-\infty}^{x_i} f(x)dx.$$

Следовательно, если известна плотность вероятности $f(x)$, то для разыгрывания величины X можно вместо уравнений $F_X(x_i) = r_i$ решить относительно x_i уравнение

$$\int_{-\infty}^{x_i} f(x)dx = r_i.$$

Составляем уравнение:

$$\int_{-\infty}^{x_i} f(x)dx = \int_{-\infty}^{0} 0dx + \int_{0}^{x_i} \frac{b}{(1+ax)^2}dx = b\int_{0}^{x_i} \frac{dx}{(1+ax)^2} = r_i.$$

Вычисляем определенный интеграл:

$$b\int_{0}^{x_i} \frac{dx}{(1+ax)^2} = -\frac{b}{a}\cdot\frac{1}{1+ax}\Bigg|_{0}^{x_i} = -\frac{b}{a}\left(\frac{1}{1+ax_i} - 1\right) = -\frac{b}{a}\cdot\frac{-ax_i}{1+ax_i} = \frac{bx_i}{1+ax_i}.$$

Таким образом, получим уравнение

$$\frac{bx_i}{1+ax_i} = r_i.$$

Разрешаем его относительно x_i:

$$bx_i = r_i(1+ax_i),\ bx_i = r_i + ar_ix_i,\ bx_i - ar_ix_i = r_i,\ x_i(b-ar_i) = r_i,\ x_i = \frac{r_i}{b-ar_i}.$$

Явная формула для разыгрывания возможных значений непрерывной случайной величины X имеет вид

$$x_i = \frac{r_i}{b-ar_i}.$$

9. Найти методом суперпозиции явные формулы для разыгрывания непрерывной случайной величины X, заданной функцией распределения

$$F_X(x)=1-\frac{1}{3}\left(2e^{-2x}+e^{-3x}\right),\ 0<x<\infty.$$

Решение. Представим заданную функцию в виде

$$F_X(x)=\frac{1}{3}\left(1-e^{-3x}\right)+\frac{2}{3}\left(1-e^{-2x}\right).$$

Функции, заключенные в скобках, являются функциями распределения показательного закона, поэтому можно принять $F_1(x)=1-e^{-3x}$, $C_1=1/3$, $F_2(x)=1-e^{-2x}$, $C_2=2/3$. Таким образом, исходную функцию $F_X(x)$ представили в виде

$$F_X(x)=C_1\cdot F_1(x)+C_2\cdot F_2(x).$$

Введем в рассмотрение вспомогательную дискретную случайную величину Z с законом распределения

Z	1	2
p	$1/3$	$2/3$

Выберем независимые случайные числа r_1 и r_2. Разыграем Z по случайному числу r_1, для чего построим частичные интервалы $\Delta_1=(0,1/3)$, $\Delta_2=(1/3,1)$. Если $r_1<1/3$, то $Z=1$; если $r_1\ge 1/3$, то $Z=2$.

Возможное значение X находят, решая относительно x уравнение

$$1-e^{-3x}=r_2,\ \text{если}\ r_1<1/3,$$

или

$$1-e^{-2x}=r_2,\ \text{если}\ r_1\ge 1/3.$$

Решив эти уравнения, получим

$$x = -\left[-\ln(1-r_2)\right]/3, \text{ если } r_1 < 1/3,$$

$$x = -\left[-\ln(1-r_2)\right]/2, \text{ если } r_1 \geq 1/3.$$

Задачи для самостоятельного решения

1. Разыграть восемь возможных значений дискретной случайной величины X, закон распределения которой задан в виде таблицы:

X	3	8	12	23
p	0,2	0,12	0,43	0,23

Указание. Для определенности принять случайные числа: 0,33; 0,18; 0,51; 0,62; 0,32; 0,41; 0,95; 0,15.

2. Разыграть шесть опытов по схеме Бернулли: опыт состоит из четырех испытаний, в каждом из которых вероятность появления события A равна 0,5.

Указание. Принять для определенности случайные числа: 0,1009; 0,7325; 0,3376; 0,5201; 0,3586; 0,3467.

3. Заданы вероятности четырех событий, образующих полную группу: $p_1 = P(A_1) = 0{,}15$, $p_2 = P(A_2) = 0{,}64$, $p_3 = P(A_3) = 0{,}05$, $p_4 = P(A_4) = 0{,}16$.

Разыграть десять испытаний, в каждом из которых появляется одно из рассматриваемых событий.

Указание. Принять для определенности случайные числа: 0,37; 0,54; 0,2; 0,48; 0,05; 0,64; 0,89; 0,47; 0,42; 0,96.

4. События A и B независимы и совместны. Разыграть 5 испытаний, в каждом из которых вероятность появления события A равна 0,6, а события B – 0,8.

Указание. Для определенности принять случайные числа: 0,69; 0,07; 0,49; 0,41; 0,38.

5. События A и B зависимы и совместны. Разыграть пять испытаний, в каждом из которых заданы вероятности: $P(A) = 0{,}5$, $P(B) = 0{,}6$, $P(AB) = 0{,}2$.

Указание. Для определенности принять следующие случайные числа: 0,66; 0,06; 0,57; 0,47; 0,17.

6. Разыграть пять возможных значений непрерывной случайной величины X, заданной плотностью вероятности $f(x) = 10/(1+2x)^2$ в интервале $(0, 1/8)$; вне этого интервала $f(x) = 0$.

Указание. Для определенности принять следующие случайные числа: 0,186; 0,333; 0,253; 0,798; 0,145.

7. Найти явную формулу для разыгрывания непрерывной случайной величины, распределенной по показательному закону, заданному плотностью вероятности $f(x)=\lambda e^{-\lambda x}$ в интервале $(0,\infty)$; вне этого интервала $f(x)=0$.

8. Разыграть четыре возможных значения непрерывной случайной величины X, заданной плотностью вероятности $f(x)=1-x/2$ в интервале $(0, 2)$; вне этого интервала $f(x)=0$.

Указание. Для определенности принять следующие случайные числа: 0,35; 0,96; 0,31; 0,53.

9. Найти методом суперпозиции явные формулы для разыгрывания непрерывной случайной величины X, заданной функцией распределения

$$F_X(x)=1-\frac{1}{5}\left(2e^{-3x}+3e^{-4x}\right),\ 0<x<\infty.$$

Указание. Принять $F_1(x)=1-e^{-3x}$, $F_2(x)=1-e^{-4x}$.

10. Найти методом суперпозиции явные формулы для разыгрывания непрерывной случайной величины X, заданной функцией распределения

$$F_X(x)=1-\frac{1}{7}\left(e^{-x}+2e^{-2x}+4e^{-3x}\right),\ 0<x<\infty.$$

Указание. Принять $F_1(x)=1-e^{-x}$, $F_2(x)=1-e^{-2x}$, $F_3(x)=1-e^{-3x}$.

Ответы. 1. 8, 3, 12, 12, 8, 12, 23, 3; **2.** 1, 3, 2, 2, 2, 2; **3.** A_2, A_2, A_2, A_2, A_1, A_2, A_4, A_2, A_4; **4.** A_3, A_1, A_2, A_1, A_1; **5.** A_3, A_1, A_3, A_2, A_1; **6.** 0,019; 0,036; 0,027; 0,095; 0,015; **7.** $x_i=(-\ln r_i)/\lambda$; **8.** 0,388; 1,6; 0,338; 0,628; **9.** $x=(-\ln r_2)/3$, если $r_1<2/5$; $x=(-\ln r_2)/4$, если $r_1\geq 2/5$; **10.** $x=-\ln r_2$, если $r_1<1/7$; $x=(-\ln r_2)/2$, если $1/7\leq r_1<3/7$; $x=(-\ln r_2)/3$, если $r_1\geq 3/7$.

Контрольные вопросы

1. В чем суть метода Монте – Карло?
2. Что такое случайное число?
3. Чем отличаются случайные числа от псевдослучайных чисел?
4. По какому правилу можно разыграть возможные значения дискретной случайной величины?
5. Каким образом разыгрывается полная группа событий?
6. По каким правилам разыгрывают возможные значения непрерывной случайной величины?
7. В чем заключается метод суперпозиций?

Приложение 1

Таблица значений функции Лапласа $\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\, dt$

x	$\Phi_0(x)$	x	$\Phi_0(x)$	x	$\Phi_0(x)$	x	$\Phi_0(x)$
0	0	0,01	0,004	0,02	0,008	0,03	0,012
0,04	0,016	0,05	0,0199	0,06	0,0239	0,07	0,0279
0,08	0,0319	0,09	0,0359	0,1	0,0398	0,11	0,0438
0,12	0,0478	0,13	0,0517	0,14	0,0557	0,15	0,0596
0,16	0,0636	0,17	0,0675	0,18	0,0714	0,19	0,0753
0,2	0,0793	0,21	0,0832	0,22	0,0871	0,23	0,091
0,24	0,0948	0,25	0,0987	0,26	0,1026	0,27	0,1064
0,28	0,1103	0,29	0,1141	0,3	0,1179	0,31	0,1217
0,32	0,1255	0,33	0,1293	0,34	0,1331	0,35	0,1368
0,36	0,1406	0,37	0,1443	0,38	0,148	0,39	0,1517
0,4	0,1554	0,41	0,1591	0,42	0,1628	0,43	0,1664
0,44	0,17	0,45	0,1736	0,46	0,1772	0,47	0,1808
0,48	0,1844	0,49	0,1879	0,5	0,1915	0,51	0,195
0,52	0,1985	0,53	0,2019	0,54	0,2054	0,55	0,2088
0,56	0,2123	0,57	0,2157	0,58	0,219	0,59	0,2224
0,6	0,2257	0,61	0,2291	0,62	0,2324	0,63	0,2357
0,64	0,2389	0,65	0,2422	0,66	0,2454	0,67	0,2486
0,68	0,2517	0,69	0,2549	0,7	0,258	0,71	0,2611
0,72	0,2642	0,73	0,2673	0,74	0,2703	0,75	0,2734
0,76	0,2764	0,77	0,2794	0,78	0,2823	0,79	0,2852
0,8	0,2881	0,81	0,291	0,82	0,2939	0,83	0,2967
0,84	0,2995	0,85	0,3023	0,86	0,3051	0,87	0,3078
0,88	0,3106	0,89	0,3133	0,9	0,3159	0,91	0,3186
0,92	0,3212	0,93	0,3238	0,94	0,3264	0,95	0,3289
0,96	0,3315	0,97	0,334	0,98	0,3365	0,99	0,3389
1	0,3413	1,01	0,3438	1,02	0,3461	1,03	0,3485
1,04	0,3508	1,05	0,3531	1,06	0,3554	1,07	0,3577
1,08	0,3599	1,09	0,3621	1,1	0,3643	1,11	0,3665
1,12	0,3686	1,13	0,3708	1,14	0,3729	1,15	0,3749
1,16	0,377	1,17	0,379	1,18	0,381	1,19	0,383
1,2	0,3849	1,21	0,3869	1,22	0,3883	1,23	0,3907

Окончание прил. 1

x	$\Phi_0(x)$	x	$\Phi_0(x)$	x	$\Phi_0(x)$	x	$\Phi_0(x)$
1,24	0,3925	1,25	0,3944	1,26	0,3962	1,27	0,398
1,28	0,3997	1,29	0,4015	1,3	0,4032	1,31	0,4049
1,32	0,4066	1,33	0,4082	1,34	0,4099	1,35	0,4115
1,36	0,4131	1,37	0,4147	1,38	0,4162	1,39	0,4177
1,4	0,4192	1,41	0,4207	1,42	0,4222	1,43	0,4236
1,44	0,4251	1,45	0,4265	1,46	0,4279	1,47	0,4292
1,48	0,4306	1,49	0,4319	1,5	0,4332	1,51	0,4345
1,52	0,4357	1,53	0,437	1,54	0,4382	1,55	0,4394
1,56	0,4406	1,57	0,4418	1,58	0,4429	1,59	0,4441
1,6	0,4452	1,61	0,4463	1,62	0,4474	1,63	0,4484
1,64	0,4495	1,65	0,4505	1,66	0,4515	1,67	0,4525
1,68	0,4535	1,69	0,4545	1,7	0,4554	1,71	0,4564
1,72	0,4573	1,73	0,4582	1,74	0,4591	1,75	0,4599
1,76	0,4608	1,77	0,4616	1,78	0,4625	1,79	0,4633
1,8	0,4641	1,81	0,4649	1,82	0,4656	1,83	0,4664
1,84	0,4671	1,85	0,4678	1,86	0,4686	1,87	0,4693
1,88	0,4699	1,89	0,4706	1,9	0,4713	1,91	0,4719
1,92	0,4726	1,93	0,4732	1,94	0,4738	1,95	0,4744
1,96	0,475	1,97	0,4756	1,98	0,4761	1,99	0,4767
2	0,4772	2,02	0,4783	2,04	0,4793	2,06	0,4803
2,08	0,4812	2,1	0,4821	2,12	0,483	2,14	0,4838
2,16	0,4846	2,18	0,4854	2,2	0,4861	2,22	0,4868
2,24	0,4875	2,26	0,4881	2,28	0,4887	2,3	0,4893
2,32	0,4898	2,34	0,4904	2,36	0,4909	2,38	0,4913
2,4	0,4918	2,42	0,4922	2,44	0,4927	2,46	0,4931
2,48	0,4934	2,5	0,4938	2,52	0,4941	2,54	0,4945
2,56	0,4948	2,58	0,4951	2,6	0,4953	2,62	0,4956
2,64	0,4959	2,66	0,4961	2,68	0,4963	2,7	0,4965
2,72	0,4967	2,74	0,4969	2,76	0,4971	2,78	0,4973
2,8	0,4974	2,82	0,4976	2,84	0,4977	2,86	0,4949
2,88	0,498	2,9	0,4981	2,92	0,4982	2,94	0,4984
2,96	0,4985	2,98	0,4986	3	0,4987	3,2	0,4993
3,4	0,4997	3,6	0,4998	3,8	0,4999	4	0,5
4,5	0,5	5	0,5				

Приложение 2

Критические точки распределения χ^2 (хи-квадрат)

Число степеней свободы k	Уровень значимости α					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5	3,8	0,0039	0,001	0,0002
2	9,2	7,4	6	0,103	0,051	0,02
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	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	16,9	3,33	2,7	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	5,23	4,4	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	7,26	6,26	5,23
16	32	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	8,9
22	40,3	36,8	33,9	12,3	11	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43	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	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	14,3
30	50,9	47	43,8	18,5	16,8	15

Приложение 3 Критические точки распределения Стьюдента

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)						
	0,2	0,1	0,05	0,02	0,01	0,002	0,001
1	3,08	6,31	12,7	31,82	63,7	318,3	637
2	1,89	2,92	4,3	6,97	9,92	22,33	31,6
3	1,64	2,35	3,18	4,54	5,84	10,22	12,9
4	1,53	2,13	2,78	3,75	4,6	7,17	8,61
5	1,48	2,01	2,57	3,37	4,03	5,89	6,86
6	1,44	1,94	2,45	3,14	3,71	5,21	5,96
7	1,42	1,89	2,37	3	3,5	4,79	5,41
8	1,4	1,86	2,31	2,9	3,36	4,5	5,04
9	1,38	1,83	2,26	2,82	3,25	4,3	4,78
10	1,37	1,81	2,23	2,76	3,17	4,14	4,59
11	1,36	1,8	2,2	2,72	3,11	4,03	4,44
12	1,36	1,78	2,18	2,68	3,06	3,93	4,32
13	1,35	1,77	2,16	2,65	3,01	3,85	4,22
14	1,34	1,76	2,15	2,62	2,98	3,79	4,14
15	1,34	1,75	2,13	2,6	2,95	3,73	4,07
16	1,34	1,75	2,12	2,58	2,92	3,69	4,02
17	1,33	1,74	2,11	2,57	2,9	3,65	3,97
18	1,33	1,73	2,1	2,55	2,88	3,61	3,92
19	1,33	1,73	2,093	2,54	2,861	3,58	3,883
20	1,33	1,73	2,09	2,53	2,85	3,55	3,85
21	1,32	1,72	2,08	2,52	2,83	3,53	3,82
22	1,32	1,72	2,07	2,51	2,82	3,51	3,79
23	1,32	1,71	2,07	2,5	2,81	3,49	3,77
24	1,32	1,71	2,06	2,49	2,8	3,47	3,74
25	1,32	1,71	2,064	2,49	2,797	3,45	3,72
26	1,32	1,71	2,06	2,48	2,78	3,44	3,71
27	1,31	1,71	2,05	2,47	2,77	3,42	3,69
28	1,31	1,7	2,05	2,46	2,76	3,4	3,66
29	1,31	1,7	2,05	2,46	2,76	3,4	3,66
30	1,31	1,7	2,045	2,46	2,756	3,39	3,659
40	1,3	1,68	2,023	2,42	2,708	3,31	3,558
60	1,3	1,67	2,001	2,39	2,662	3,23	3,464
120	1,29	1,66	1,98	2,36	2,617	3,17	3,374
∞	1,28	1,65	1,96	2,33	2,576	3,09	3,291
Число степеней свободы k	0,1	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)						

Приложение 4

Критические точки распределения Фишера (Фишера – Снедекора)

k_2	k_1										
	2	3	4	5	6	7	8	9	10	11	12
Уровень значимости α = 0,10											
2	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,40	9,41
3	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,22
4	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,91	3,90
5	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,28	3,27
6	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,92	2,90
7	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,68	2,67
8	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,52	2,50
9	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,40	2,38
10	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,30	2,28
11	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,23	2,21
12	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,17	2,15
13	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,12	2,10
14	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,07	2,05
Уровень значимости α = 0,05											
2	19,0	19,2	19,3	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,4
3	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74
4	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91
5	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68
6	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57

k_2	k_1										
	2	3	4	5	6	7	8	9	10	11	12
8	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28
9	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07
10	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91
11	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79
12	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69
13	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60
14	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53
Уровень значимости α = 0,01											
2	99,0	99,2	99,3	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4
3	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	27,1	27,1
4	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,6	14,5	14,4
5	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1	10,0	9,9
6	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
7	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47
8	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67
9	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11
10	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71
11	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40
12	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16
13	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96
14	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80

k_1 – число степеней свободы большей дисперсии;

k_2 – число степеней свободы меньшей дисперсии.

Приложение 5

Равномерно распределенные случайные числа

10 09 73 25 33 37 54 20 48 05 08 42 26 89 53 99 01 90 25 29 12 80 79 99 70	76 52 01 35 86 64 89 47 42 96 19 64 50 93 03 09 37 67 07 15 80 15 73 61 47	34 67 35 48 76 24 80 52 40 37 23 20 90 25 60 38 31 13 11 65 64 03 23 66 53	80 95 90 91 17 20 63 61 04 02 15 95 33 47 64 88 67 67 43 97 98 95 11 68 77
66 06 57 47 17 31 06 01 08 05 85 26 97 76 02 63 57 33 21 35 73 79 64 57 53	34 07 27 68 50 45 57 18 24 06 02 05 16 56 92 05 32 54 70 48 03 52 96 47 78	36 69 73 61 70 35 30 34 26 14 68 66 57 48 18 90 55 35 75 48 35 80 83 42 82	65 81 33 98 85 86 79 90 74 39 73 05 38 52 47 28 46 82 87 09 60 93 52 03 44
98 52 01 77 67 11 80 50 54 31 83 45 29 96 34 88 68 54 02 00 99 59 46 73 48	14 90 56 86 07 39 80 82 77 32 06 28 89 80 83 86 50 75 84 01 87 51 76 49 69	22 10 94 05 58 50 72 56 82 48 13 74 67 00 78 36 76 66 79 51 91 82 60 89 28	60 97 09 34 33 29 40 52 42 01 18 47 54 06 10 90 36 47 64 93 93 78 56 13 68
65 48 11 76 74 80 12 43 56 35 74 35 09 98 17 69 91 62 68 03 09 89 32 05 05	17 46 85 09 50 17 72 70 80 15 77 40 27 72 14 66 25 22 91 48 14 22 56 85 14	58 04 77 69 74 45 31 82 23 74 43 23 60 02 10 36 93 68 72 03 46 42 75 67 88	73 03 95 71 86 21 11 57 82 53 45 52 16 42 37 76 62 11 39 90 96 29 77 88 22
91 49 91 45 23 80 33 69 45 98 44 10 48 19 49 12 55 07 37 42 63 60 64 93 29	68 47 92 76 86 26 94 03 68 58 85 15 74 79 54 11 10 00 20 40 16 50 53 44 84	46 16 28 35 54 70 29 73 41 35 32 97 92 65 75 12 86 07 46 97 40 21 95 25 63	94 75 08 99 23 53 14 03 33 40 57 60 04 08 81 96 64 48 94 39 43 65 17 70 82
61 19 69 04 46 15 47 44 52 66 94 55 72 85 73 42 48 11 62 13 23 52 37 83 17	26 45 74 77 74 95 27 07 99 53 67 89 75 43 87 97 34 40 87 21 73 20 88 98 37	51 92 43 37 29 59 36 78 38 48 54 62 24 44 31 16 86 84 87 67 68 93 59 14 16	65 39 45 95 93 82 39 61 01 18 91 19 04 25 92 03 07 11 20 59 26 25 22 96 63
04 49 35 24 94 00 54 99 76 54 85 96 31 53 07 59 80 80 83 91 46 05 88 52 36	75 24 63 38 24 64 05 18 81 59 26 89 80 93 54 45 42 72 68 42 01 39 09 22 86	45 86 25 10 25 96 11 96 38 96 33 35 13 54 62 83 60 94 97 00 77 28 14 40 77	61 96 27 93 35 54 69 28 23 91 77 97 45 00 24 13 02 12 48 92 93 91 08 36 47
32 17 90 05 97 69 23 46 14 06 19 56 54 14 30 45 15 51 49 38 94 86 43 19 94	87 37 92 52 41 20 11 74 52 04 01 75 87 53 79 19 47 60 72 46 36 16 81 08 51	05 56 70 70 07 15 95 66 00 00 40 41 92 15 85 43 66 79 45 43 34 88 88 15 53	86 74 31 71 57 18 74 39 24 23 66 67 43 68 06 59 04 79 00 33 01 54 03 54 56

Окончание прил. 5

98 08 62 48 26 33 18 51 62 32 80 95 10 04 06 79 75 24 91 40 18 63 33 25 37	45 24 02 84 04 41 94 15 09 49 96 38 27 07 74 71 96 12 82 96 98 14 50 65 71	44 99 90 88 96 89 43 54 85 81 20 15 12 33 87 69 86 10 25 91 31 01 02 46 74	39 09 47 34 07 88 69 54 19 94 25 01 62 52 98 74 85 22 05 39 05 45 56 14 27
74 02 94 39 02 54 17 84 56 11 11 66 44 98 83 48 32 47 79 28 69 07 49 41 38	77 55 73 22 70 80 99 33 71 43 52 07 98 48 27 31 24 96 47 10 87 63 79 19 76	97 79 01 71 19 05 33 51 29 69 59 38 17 15 39 02 29 53 68 70 35 58 48 44 01	52 52 75 80 21 56 12 71 92 55 09 97 33 34 40 32 30 75 75 46 10 51 82 16 15

Библиографический список

1. Бусленко Н.П. Метод статистического моделирования /Н.П. Бусленко. – М.: Статистика, 1970.

2. Ван дер Варден Б. Математическая статистика /Ван Дер Варден Б.. – М.: Изд. ин. лит, 1960. – 436 с.

3. Вентцель Е.В. Теория вероятностей: – Учебник для вузов /Е.В. Вентцель – 7-е изд., стер. – М.: Высш. школа, 2001. – 575 с.

4. Вентцель Е.В. Теория вероятностей и ее инженерные приложения: – Учебное пособие для студ. втузов /Е.В. Вентцель, Л.А. Овчаров. – 3-е изд., перераб. и доп. – М.: Издательский центр «Академия», 2003. – 464 с.

5. Гмурман В.Е. Теория вероятностей и математическая статистика: – Учеб. пособие /В.Е. Гмурман. – 12-е изд., перераб. – М.: Высш. образование, 2008. – 479 с.

6. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике: – Учеб. пособие /В.Е. Гмурман. – 11-е изд., перераб. – М.: Высш. образование, 2008. – 404 с.

7. Крамер Г. Математические методы статистики. – М.: Мир, 1975.

8. Пугачев В.С. Теория вероятностей и математическая статистика: – Учеб. пособие /В.С. Пугачев. – 2-е изд., испр. и доп. – М.: Физматлит, 2002. – 496 с.

9. Соболь И.М. Метод Монте-Карло /И.М. Соболь. – М.: Наука, 1968. – 64 с.

10. Фадеева Л.Н. Математика для экономистов: Теория вероятностей и математическая статистика: Курс лекций /Л.Н. Фадеева. – М.: Эксмо, 2006. – 400 с. – (Высшее экономическое образование).

11. Фадеева Л.Н. Математика для экономистов: Теория вероятностей и математическая статистика. Задачи и упражнения /Л.Н. Фадеева, Ю.В. Жуокв, А.В. Лебедев. – М.: Эксмо, 2006. – 336 с.

ОГЛАВЛЕНИЕ

Учебное издание

Григорян Тамара Анатольевна,
Липачева Екатерина Владимировна

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
Часть 2
Математическая статистика
Учебно-методическое пособие

Кафедра высшей математики КГЭУ

Редактор издательского отдела Н.А. Артамонова
Компьютерная верстка Н.А. Артамонова

Подписано в печать 00.00.00.
Формат 60×84/16. Бумага «Business». Гарнитура «Times». Вид печати РОМ.
Усл. печ. л. . Уч.-изд. л. . Тираж 500 экз. Заказ № .

Издательство КГЭУ, 420066, Казань, Красносельская, 51
Типография КГЭУ, 420066, Казань, Красносельская, 51