Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

Ю. С. Боярович, Ю. Е. Дудовская

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Практическое руководство

для студентов специальностей
1-31 03 03-01 «Прикладная математика
(научно-производственная деятельность)»,
1-31 03 03-02 «Прикладная математика
(научно-педагогическая деятельность)»

Гомель
ГГУ им. Ф. Скорины
2012

УДК 519.22(076)
ББК 22.172я73
Б 869

Рецензенты:
доктор физико-математических наук, профессор В. Н. Семенчук;
кандидат физико-математических наук А. Н. Старовойтов

Рекомендовано к изданию научно-методическим советом учреждения образования «Гомельский государственный университет имени Франциска Скорины»

Боярович, Ю. С.
Б 869 Математическая статистика: практ. рук-во / Ю. С. Боярович, Ю. Е. Дудовская; М-во образования РБ, Гомельский гос. ун-т им. Ф. Скорины. – Гомель: ГГУ им. Ф. Скорины, 2012. – 44 с.
ISBN 978-985-439-720-7

В практическом руководстве изложены теоретические основы математической статистики. В издание включены следующие разделы: первичная обработка статистических данных, статистические оценки неизвестных параметров распределения, интервальные оценки неизвестных параметров распределения, проверка параметрических гипотез, гипотезы и критерии согласия, однофакторный дисперсионный анализ, корреляционный и регрессионный анализ. В каждом из разделов представлены решения типовых задач.

Предназначено для студентов специальностей 1-31 03 03-01 «Прикладная математика (научно-производственная деятельность)», 1-31 03 03-02 «Прикладная математика (научно-педагогическая деятельность)».

УДК 519.22(076)
ББК 22.172я73

ISBN 978-985-439-720-7

Содержание

Введение ………………………………………………………………. 4
1 Первичная обработка статистических данных …………………….. 5
2 Статистические оценки неизвестных параметров распределения ... 12
3 Интервальные оценки неизвестных параметров распределения …. 17
4 Проверка параметрических гипотез ……………………………... 21
5 Гипотезы и критерии согласия ………………………………... 28
6 Однофакторый дисперсионный анализ ……………………………. 33
7 Корреляционный и регрессионный анализ ……………………... 37
Литература ……………………………………………………….. 43

Введение

Практическое руководство «Математическая статистика» предназначено для студентов высших учебных заведений, изучающих дисциплину «Теория вероятностей и математическая статистика».

Практическое руководство состоит из семи разделов, посвященных основным понятиям и методам математической статистики: первичная обработка статистических данных, оценки параметров распределений и методы их построения, проверка статистических гипотез, дисперсионный анализ, корреляционный и линейный регрессионный анализ статистических данных. Разделы практического руководства имеют идентичную структуру: краткие теоретические сведения по теме раздела, решения и примеры оформления типовых задач.

В руководстве приводится список рекомендуемой литературы, который предлагается использовать при изучении дисциплины.

Данное издание адресовано преподавателям и студентам для проведения практических, лабораторных занятий и организации самостоятельной учебной работы.

1 Первичная обработка статистических данных

1. Абстрактная и конкретная выборки.
2. Основные числовые характеристики выборки.
3. Вариационные ряды выборки.
4. Гистограмма частот.
5. Эмпирическая функция распределения.

Пусть в одинаковых условиях и независимо друг от друга производится n измерений случайной величины ξ. Назовем случайную величину ξ *теоретической случайной величиной*, а ее функцию распределения $F(x)$ – *теоретической функцией распределения*. Пусть $x_1, x_2, ..., x_n$ – результаты измерений. Набор $X = (x_1, x_2, ..., x_n)$ называется *конкретной выборкой* объема n из распределения $F(x)$.

Абстрактной выборкой объема n называется совокупность n независимых одинаково распределенных случайных величин $\xi_1, ..., \xi_n$, распределение каждой из которых совпадает с распределением теоретической случайной величины ξ.

Если элементы выборки $X = (x_1, x_2, ..., x_n)$ упорядочить по возрастанию, получится новый набор, называемый *вариационным рядом*:

$$x_{(1)} \le x_{(2)} \le ... \le x_{(n)}.$$

Если среди элементов вариационного ряда есть повторяющиеся, то можно выделить $m \le n$ его различных значений, расположив их в порядке возрастания. Обозначим их $z_{(1)}, z_{(2)}, ..., z_{(m)}$ $(z_{(1)} < z_{(2)} < ... < z_{(m)})$. Число k_i, показывающее, сколько раз элемент $z_{(i)}$ встретился в выборке, называется *частотой*, а $\frac{k_i}{n}$ – *относительной частотой* (*частостью*) этого значения, $i = 1, ..., m$ $\left(\sum_{i=1}^{m} k_i = n\right)$.

Статистическим рядом называется таблица, содержащая в первой строке значения $z_{(1)}, z_{(2)}, ..., z_{(m)}$, а во второй строке – частоты значений.

$z_{(1)}$	$z_{(2)}$	...	$z_{(m)}$
k_1	k_2	...	k_m

Случайная величина τ с рядом распределения

τ	$z_{(1)}$	...	$z_{(m)}$
	$\dfrac{k_1}{n}$	...	$\dfrac{k_m}{n}$

называется *эмпирической случайной величиной*, а соответствующая ей функция распределения $F_n^*(x)$ – *выборочной или эмпирической функцией распределения:*

$$F_n^*(z) = \begin{cases} 0, & z \le z_{(1)} \\ \ldots & \\ \dfrac{k_1 + \ldots + k_i}{n}, & z_{(i)} < z \le z_{(i+1)} \\ \ldots & \\ 1, & z > z_{(m)} \end{cases}$$

Элементы выборки можно объединить в группы и построить *интервальный вариационный ряд*. Для этого отрезок $[x_{(1)}, x_{(n)}]$ разбивается на k равных промежутков $\Delta_1, \ldots, \Delta_k$. Определяются середины промежутков l_i, $i = 1, \ldots, k$. Количество промежутков k зависит от объема выборки n и может быть вычислено по формуле *Стерджесса*:

$$k \approx 1 + 3{,}32 \lg n.$$

Далее определяются *частоты интервального вариационного ряда* n_i – количество элементов выборки, попавших в i-й промежуток, $i = 1, \ldots, k$, $\sum\limits_{i=1}^{k} n_i = n$. *Относительные частоты (частости) интервального вариационного ряда* определяются как $\omega_i = \dfrac{n_i}{n}$, $i = 1, \ldots, k$. Результаты удобно представить в виде таблицы 1:

Таблица 1

Интервал	Δ_1	Δ_2	...	Δ_k
Середина интервала	l_1	l_2	...	l_k
Частота	n_1	n_2	...	n_k
Относительная частота	$\dfrac{n_1}{n}$	$\dfrac{n_2}{n}$	...	$\dfrac{n_k}{n}$

Заметим, что эмпирическая функция распределения может быть определена как функция распределения случайной величины, принимающей значения $l_1, \ldots, l_k$ с вероятностями $\dfrac{n_1}{n}, \ldots, \dfrac{n_k}{n}$ соответственно.

Статистические данные, представленные в виде статистического ряда или интервального вариационного ряда, называют *группированными*.

Гистограмма частот группированной выборки – это график кусочно-постоянной функции, принимающей на каждом из интервалов $\Delta_1,...,\Delta_k$ значение $\frac{n_i}{h}$ ($h=(x_{(n)}-x_{(1)})/k$ – длина интервала), $i=1,...,k$. Аналогично по значениям $\frac{n_i}{hn}$ строится *гистограмма относительных частот*, $i=1,...,k$.

Полигоном частот для данных, представленных в виде интервального вариационного ряда, называется график ломаной с вершинами в точках $\left(l_i, \frac{n_i}{h}\right)$, а полигоном относительных частот – в точках $\left(l_i, \frac{n_i}{hn}\right)$, $i=1,...,k$.

При увеличении объема выборки и уменьшении интервала группирования гистограмма и полигон относительных частот могут рассматриваться как статистические аналоги теоретической плотности распределения.

В таблице 2 приведены основные числовые характеристики выборки.

Таблица 2 – Основные выборочные характеристики

Выборочное среднее	$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$
Выборочная дисперсия	$\tilde{S}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-(\bar{x})^2$
Выборочное среднеквадратическое отклонение	$\tilde{S}=\sqrt{\tilde{S}^2}$
Выборочный начальный момент k-го порядка	$\overline{x^k}=\frac{1}{n}\sum_{i=1}^{n}x_i^k$
Выборочный центральный момент k-го порядка	$\tilde{S}^k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k$
Выборочная мода Mo	Элемент выборки, встречающийся с наибольшей частотой
Выборочная медиана Me	$Me=\begin{cases} x_{(l+1)}, n=2l+1; \\ \frac{x_{(l)}+x_{(l+1)}}{2}, \ n=2l. \end{cases}$

Окончание таблицы 2

Выборочный коэффициент асимметрии	$k_a = \dfrac{\tilde{S}^3}{(\tilde{S})^3}$
Выборочный коэффициент эксцесса	$k_e = \dfrac{\tilde{S}^4}{(\tilde{S})^4} - 3$

Пример 1.1 В результате наблюдений над случайной величиной ξ получена выборка X объема $n = 30$:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1,37	0,11	1,56	–0,11	0,23	–0,76	–0,13	–0,64	–0,46	–0,88

x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
–0,56	1,28	1,16	–0,3	–0,31	1,13	–0,17	0,6	–1,16	2,65

x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	x_{30}
1,55	0,29	–2,16	–0,77	0,93	0,01	–1,56	1,59	–1,13	–1,74

Произвести статистическую обработку результатов:

1) вычислить основные числовые характеристики выборки;

2) построить интервальный вариационный ряд выборки и гистограмму частот;

3) построить эмпирическую функцию распределения, взяв в качестве значений середины интервалов интервального вариационного ряда.

1) Основные числовые характеристики выборки.

Выборочное среднее:

$$\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i = \frac{1}{30}\left(1{,}37 + 0{,}11 + 1{,}56 + \ldots + (-1{,}74)\right) \approx 0{,}05.$$

Выборочная дисперсия:

$$\tilde{S}^2 = \frac{1}{30}\sum_{i=1}^{30} x_i^2 - (\bar{x})^2 \approx \frac{1}{30}\left(1{,}37^2 + 0{,}11^2 + 1{,}56^2 + \ldots + (-1{,}74)^2\right) - 0{,}05^2 \approx 1{,}26.$$

Выборочное среднеквадратическое отклонение:

$$\tilde{S} = \sqrt{\tilde{S}^2} \approx \sqrt{1{,}26} \approx 1{,}12;$$

Вариационный ряд выборки имеет вид:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
–2,16	–1,74	–1,56	–1,16	–1,13	–0,88	–0,77	–0,76	–0,64	–0,56

$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$	$x_{(16)}$	$x_{(17)}$	$x_{(18)}$	$x_{(19)}$	$x_{(20)}$
–0,46	–0,31	–0,3	–0,17	–0,13	–0,11	0,01	0,11	0,23	0,29

$x_{(21)}$	$x_{(22)}$	$x_{(23)}$	$x_{(24)}$	$x_{(25)}$	$x_{(26)}$	$x_{(27)}$	$x_{(28)}$	$x_{(29)}$	$x_{(30)}$
0,6	0,93	1,13	1,16	1,28	1,37	1,55	1,56	1,59	2,65

Размах выборки:

$$x_{(30)} - x_{(1)} = 2{,}65 - (-2{,}16) = 4{,}81.$$

Выборочная медиана. Объем выборки $n = 30$ – четное число, поэтому воспользуемся формулой

$$Me = \frac{x_{(15)} + x_{(15+1)}}{2} = \frac{-0{,}13 + (-0{,}11)}{2} = -0{,}12\,.$$

Перед вычислением выборочных коэффициентов асимметрии и эксцесса найдем выборочные центральные моменты третьего и четвертого порядков:

$$\tilde{S}^3 = \frac{1}{30}\sum_{i=1}^{30}(x_i - \bar{x})^3 \approx \frac{1}{30}\left((1{,}37-0{,}05)^3 + (0{,}11-0{,}05)^3 + \ldots + (-1{,}74-0{,}05)^3\right) \approx 0{,}29\,;$$

$$\tilde{S}^4 = \frac{1}{30}\sum_{i=1}^{30}(x_i - \bar{x})^4 \approx \frac{1}{30}\left((1{,}37-0{,}05)^4 + (0{,}11-0{,}05)^4 + \ldots + (-1{,}74-0{,}05)^4\right) \approx 3{,}92.$$

Выборочный коэффициент асимметрии:

$$k_a = \frac{\tilde{S}^3}{(\tilde{S})^3} \approx \frac{0{,}29}{(1{,}12)^3} \approx 0{,}21\,.$$

Выборочный коэффициент эксцесса:

$$k_e = \frac{\tilde{S}^4}{(\tilde{S})^4} - 3 \approx -0{,}52\,.$$

2) Построим интервальный вариационный ряд выборки. Число интервалов вычислим по формуле Стерджесса: $k = 1 + 3{,}32\lg 30 \approx 6$. Разобьем отрезок $\left[x_{(1)}, x_{(30)}\right] = \left[-2{,}16;\, 2{,}65\right]$ на 6 равных интервалов. Длина интервала $h = \dfrac{x_{(30)} - x_{(1)}}{k} = \dfrac{2{,}65 - (-2{,}16)}{6} \approx 0{,}8$. Результаты представим в виде таблицы 3:

Таблица 3

Интервал	Середина интервала z_i	Частота n_i	Относительная частота $\frac{n_i}{n}$
$[-2{,}16; -1{,}36)$	–1,76	3	$\frac{3}{30} = \frac{1}{10}$
$[-1{,}36; -0{,}56)$	–0,96	6	$\frac{6}{30} = \frac{1}{5}$
$[-0{,}56; 0{,}24)$	–0,16	10	$\frac{10}{30} = \frac{1}{3}$
$[0{,}24; 1{,}04)$	0,64	3	$\frac{3}{30} = \frac{1}{10}$
$[1{,}04; 1{,}84)$	1,44	7	$\frac{7}{30}$
$[1{,}84; 2{,}65]$	2,245	1	$\frac{1}{30}$

На рисунке 1 изображена гистограмма частот.

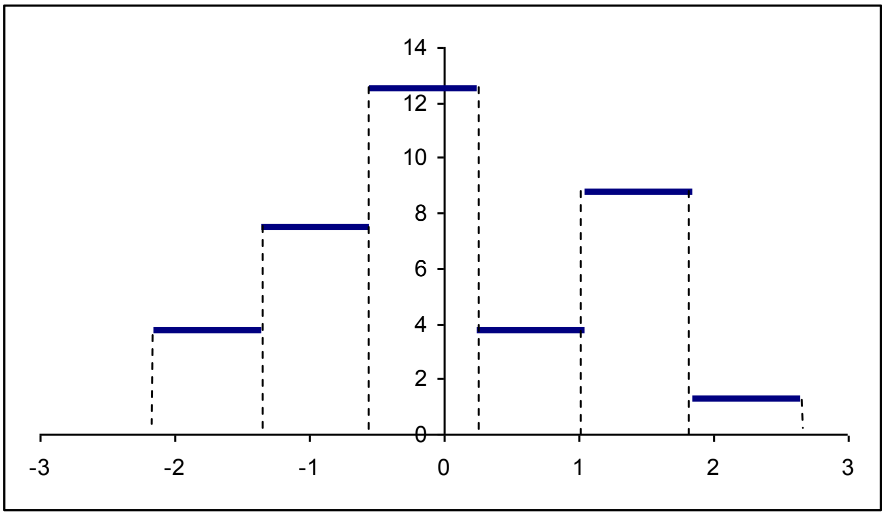


Рисунок 1

3) Построим эмпирическую функцию распределения, взяв в качестве значений середины интервалов интервального вариационного ряда

$$F^*(z)=\begin{cases}0, & z\le -1{,}76,\\ \dfrac{1}{10}, & -1{,}76<z\le -0{,}96,\\ \dfrac{3}{10}, & -0{,}96<z\le -0{,}16,\\ \dfrac{19}{30}, & -0{,}16<z\le 0{,}64,\\ \dfrac{11}{15}, & 0{,}64<z\le 1{,}44,\\ \dfrac{29}{30}, & 1{,}44<z\le 2{,}245,\\ 1, & z>2{,}245.\end{cases}$$

График эмпирической функции распределения изображен на рисунке 2.

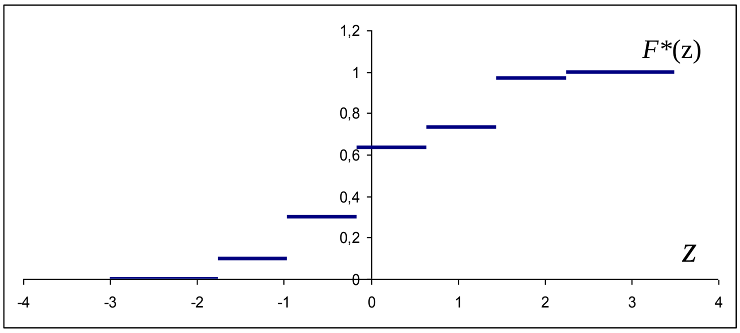


Рисунок 2

Вопросы для самоконтроля

1. Дайте определение абстрактной и конкретной выборок.
2. Укажите основные числовые характеристики выборки: размах выборки, выборочное среднее, выборочная дисперсия, выборочная медиана, выборочные коэффициенты асимметрии и эксцесса.
3. Как построить интервальный вариационный ряд выборки?
4. Как построить гистограмму частот?
5. Что называется эмпирической функцией распределения?

2 Статистические оценки неизвестных параметров распределения

1. Статистическая оценка неизвестного параметра теоретического распределения.
2. Виды статистических оценок.
3. Нахождение оценок неизвестных параметров теоретического распределения методами моментов и максимального правдоподобия.

Пусть $x_1, x_2, ..., x_n$ – выборка из теоретического распределения $F(x|\theta)$, зависящего от неизвестного параметра θ.

Оценкой (статистикой) $\hat{\theta}_n$ неизвестного параметра θ называется любая борелевская функция $\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, ..., x_n)$.

Оценка $\hat{\theta}_n$ неизвестного параметра θ называется *несмещенной*, если ее математическое ожидание равно оцениваемому параметру, т. е. $M(\hat{\theta}_n) = \theta$. В противном случае оценка называется *смещенной*.

Можно показать, что выборочное среднее $\bar{x}$ является несмещенной оценкой теоретического математического ожидания, а выборочная дисперсия $\tilde{S}^2$ является смещенной оценкой теоретической дисперсии. Исправленная выборочная дисперсия

$$S^2 = \frac{n}{n-1}\tilde{S}^2$$

является несмещенной оценкой теоретической дисперсии.

Оценка $\hat{\theta}_n$ неизвестного параметра θ называется *состоятельной*, если она сходится по вероятности к оцениваемому параметру:

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1.$$

Можно показать, что оценки $\bar{x}$, $\tilde{S}^2$ и S^2 являются состоятельными оценками теоретического математического ожидания и теоретической дисперсии соответственно.

Несмещенная оценка $\hat{\theta}_n$ параметра θ называется *эффективной*, если она имеет наименьшую дисперсию в классе несмещенных оценок, т. е. если $\tilde{\theta}_n$ – произвольная несмещенная оценка параметра θ, а $\hat{\theta}_n$ – эффективная, то $D(\hat{\theta}_n) \le D(\tilde{\theta}_n)$.

Рассмотрим основные методы получения оценок.

Метод моментов. Пусть имеется выборка $x_1, x_2, ..., x_n$ из распределения $F(x; \theta_1, \theta_2, ..., \theta_k)$, зависящего от неизвестных параметров $\theta_1, \theta_2, ..., \theta_k$, которые нужно оценить. Поскольку известен вид теоретической функции распределения, можем вычислить первые *k* теоретических моментов (начальных или центральных). Эти моменты будут зависеть от *k* неизвестных параметров $\theta_1, \theta_2, ..., \theta_k$:

$$\begin{cases} m_1 = m_1(\theta_1, \theta_2, ..., \theta_k), \\ m_2 = m_2(\theta_1, \theta_2, ..., \theta_k), \\ \quad ... \\ m_k = m_k(\theta_1, \theta_2, ..., \theta_k). \end{cases}$$

Суть метода моментов заключается в следующем: выборочные моменты являются оценками соответствующих теоретических моментов, поэтому теоретические моменты $m_1, m_2, ..., m_k$ приравнивают к соответствующим выборочным $m_1^*, m_2^*, ..., m_k^*$, а затем, решая систему относительно $\theta_1, \theta_2, ..., \theta_k$, находят оценки неизвестных параметров. Таким образом, в методе моментов оценки $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ неизвестных параметров $\theta_1, \theta_2, ..., \theta_k$ определяются как решение системы *k* уравнений с *k* неизвестными:

$$\begin{cases} m_1^* = m_1(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k), \\ m_2^* = m_2(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k), \\ \quad ... \\ m_k^* = m_k(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k). \end{cases}$$

В таблице 4 приведены формулы для вычисления выборочных и соответствующих им теоретических моментов порядка *k*:

Таблица 4 – Выборочные и теоретические моменты

Моменты	Теоретические	Выборочные
Начальные	$b_k = M\xi^k$	$b_k^* = \overline{x^k} = \frac{1}{n}\sum_i x_i^k$
Центральные	$c_k = M(\xi - M\xi)^k$	$c_k^* = \tilde{S}^k = \frac{1}{n}\sum_i (x_i - \bar{x})^k$

Пример 2.1 Предполагая, что выборка *X* из примера 1.1 получена из показательного распределения с параметром λ, методом моментов найти точечную оценку параметра.

Для показательного распределения плотность распределения вероятностей имеет вид

$$p(x) = \lambda e^{-\lambda x},\ x \geq 0.$$

Для данного закона распределения определим теоретический момент

$$b_1 = \int_{-\infty}^{+\infty} x p(x) dx = \int_{0}^{+\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda}.$$

По заданной выборке определим значение выборочного момента

$$b_1^* = \overline{x} = \frac{1}{30}\sum_{i=1}^{30} x_i \approx 0{,}05.$$

Составим уравнение

$$\frac{1}{\lambda} = \overline{x}.$$

Таким образом, получим оценку

$$\hat{\lambda} = 20.$$

Пример 2.2 Предполагая, что выборка *X* из примера 1.1 получена из равномерного распределения на отрезке $[a,b]$ методом моментов найти точечные оценки параметров a и b.

Для равномерного распределения плотность распределения вероятностей имеет вид

$$p(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a, b], \\ 0, & x \notin [a, b]. \end{cases}$$

Равномерное распределение имеет два параметра, поэтому необходимо составить два уравнения. Определим теоретические моменты

$$b_1 = \int_{-\infty}^{+\infty} x p(x) dx = \int_{a}^{b} \frac{x}{b-a} dx = \frac{a+b}{2};$$

$$b_2 = \int_{-\infty}^{+\infty} x^2 p(x) dx = \int_{a}^{b} \frac{x^2}{b-a} dx = \frac{a^2 + ab + b^2}{3}.$$

По выборке определим значения выборочных моментов

$$b_1^* = \overline{x} = \frac{1}{30}\sum_{i=1}^{30} x_i \approx 0{,}05;$$

$$b_2^* = \overline{x^2} = \frac{1}{30}\sum_{i=1}^{30} x_i^2 \approx 1{,}26.$$

Составим систему уравнений

$$\begin{cases} \dfrac{a+b}{2} = 0{,}05, \\ \dfrac{a^2 + ab + b^2}{3} = 1{,}26. \end{cases}$$

Решив ее относительно неизвестных параметров a и b, получим

$$a_1 = -1{,}89,\ b_1 = 1{,}99;$$
$$a_2 = 1{,}99,\ b_2 = -1{,}89.$$

Поскольку $a < b$,

$$\hat{a} = -1{,}89,$$
$$\hat{b} = 1{,}99.$$

Метод максимального правдоподобия. Пусть имеется выборка $x_1, x_2, ..., x_n$ из распределения $F(x; \theta_1, \theta_2, ..., \theta_k)$, зависящего от неизвестных параметров $\theta_1, \theta_2, ..., \theta_k$, которые нужно оценить. Основу метода составляет *функция правдоподобия:*

$$L(x_1, x_2, ..., x_n; \theta_1, \theta_2, ..., \theta_k) = p(x_1; \theta_1, \theta_2, ..., \theta_k)\, p(x_2; \theta_1, \theta_2, ..., \theta_k) ... p(x_n; \theta_1, \theta_2, ..., \theta_k).$$

В дискретном случае функция $p(x_i; \theta_1, \theta_2, ..., \theta_k)$ выражает вероятность того, что теоретическая случайная величина примет значение x_i, $i = 1, ..., n$. В абсолютно непрерывном случае функция $p(x_i; \theta_1, \theta_2, ..., \theta_k)$ – значение теоретической плотности распределения вероятностей в точке x_i, $i = 1, ..., n$.

Согласно методу максимального правдоподобия в качестве оценки неизвестных параметров $\theta_1, \theta_2, ..., \theta_k$ принимаются такие значение $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$, которые максимизируют функцию правдоподобия.

Нахождение оценок может упрощаться, если максимизировать не саму функцию L, а $\ln L$.

Пример 2.3 Предполагая, что выборка X из примера 1.1 получена из показательного распределения с параметром λ, методом максимального правдоподобия найти точечную оценку параметра.

Запишем функцию правдоподобия

$$L = \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} ... \lambda e^{-\lambda x_{30}} = \lambda^{30} e^{-\lambda \sum_{i=1}^{30} x_i}.$$

Логарифмируя, имеем

$$\ln L = 30\ln\lambda - \lambda\sum_{i=1}^{30} x_i .$$

Задача сводится к нахождению максимума функции одной переменной. Дифференцируя по параметру λ, находим

$$\frac{d\ln L}{d\lambda} = \frac{30}{\lambda} - \sum_{i=1}^{30} x_i .$$

Приравнивая производную к нулю, получим

$$\frac{30}{\lambda} - \sum_{i=1}^{30} x_i = 0 \Rightarrow \lambda = \frac{30}{\sum_{i=1}^{30} x_i} = \frac{1}{\overline{x}} .$$

Таким образом, получим оценку

$$\hat{\lambda} = 20 .$$

Вопросы для самоконтроля

1. Дайте определение статистической оценки неизвестного параметра теоретического распределения.
2. В чем заключается суть метода моментов?
3. В чем заключается суть метода максимального правдоподобия?

3 Интервальные оценки неизвестных параметров распределения

1. Доверительные пределы для неизвестного параметра теоретического распределения.
2. Доверительные интервалы для неизвестного параметра теоретического распределения.
3. Построение доверительных интервалов для неизвестных параметров нормального распределения.

Пусть $x_1, x_2,..., x_n$ – выборка из распределения $F(x|\theta)$, зависящего от неизвестного параметра θ. Статистика $\theta_B = \theta_B(x_1, x_2,..., x_n)$ называется *односторонним верхним доверительным пределом* для неизвестного параметра θ с надежностью (доверительной вероятностью) P или с уровнем значимости $\alpha = 1 - P$, если

$$P\{\theta < \theta_B\} = P, \quad (0{,}5 < P < 1).$$

Статистика $\theta_H = \theta_H(x_1, x_2,..., x_n)$ называется *односторонним нижним доверительным пределом* для неизвестного параметра θ с надежностью P или с уровнем значимости $\alpha = 1 - P$, если

$$P\{\theta > \theta_H\} = P.$$

При этом интервалы $(-\infty, \theta_B)$ и $(\theta_H, +\infty)$ называются соответственно *верхним* и *нижним односторонними (левосторонним и правосторонним) доверительными интервалами* для неизвестного параметра θ с надежностью P или с уровнем значимости $\alpha = 1 - P$.

Статистики $\theta_H = \theta_H(x_1, x_2,..., x_n)$ и $\theta_B = \theta_B(x_1, x_2,..., x_n)$ называются соответственно *двусторонними нижним и верхним доверительными пределами* для неизвестного параметра θ с надежностью P или с уровнем значимости $\alpha = 1 - P$, если

$$P\{\theta_H < \theta < \theta_B\} = P.$$

При этом интервал (θ_H, θ_B) называется *двусторонним доверительным интервалом* для неизвестного параметра θ с надежностью P или с уровнем значимости $\alpha = 1 - P$.

Оценка параметров нормального распределения. Пусть $X = (x_1, x_2,..., x_n)$ – выборка из нормального распределения $N(a, \sigma^2)$. Требуется построить интервальные оценки для параметров a и σ^2.

Пусть задан некоторый уровень значимости α (надежность $P = 1 - \alpha$). В таблице 5 приводятся доверительные интервалы параметров нормального распределения.

Таблица 5

	Доверительный интервал
Оценка математического ожидания a при известной дисперсии σ^2.	$\bar{x} - \frac{\sigma}{\sqrt{n}} u_\alpha < a < \bar{x} + \frac{\sigma}{\sqrt{n}} u_\alpha$, u_α – корень уравнения $\Phi(u_\alpha) = \frac{1-\alpha}{2}$, где $\Phi(x)$ – функция Лапласа.
Оценка математического ожидания a при неизвестной дисперсии σ^2.	$\bar{x} - \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1} < a < \bar{x} + \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1}$, $S = \sqrt{S^2}$, $t_{\alpha,n} = t\ \ 100\alpha\%; n$ – 100α-процентная точка распределения Стьюдента с n степенями свободы.
Оценка дисперсии σ^2 при известном математическом ожидании a.	$\frac{nS_1^2}{\chi^2_{\frac{\alpha}{2}, n}} < \sigma^2 < \frac{nS_1^2}{\chi^2_{1-\frac{\alpha}{2}, n}}$, $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$; $\chi^2_{\alpha,n} = \chi^2\ \ 100\alpha\%; n$ – 100α-процентная точка распределения χ^2 с n степенями свободы.
Оценка дисперсии σ^2 при неизвестном математическом ожидании a.	$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}, n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}}$

Пример 3.1 Предполагая, что выборка X из примера 1.1 получена из нормального распределения, построить доверительные интервалы для математического ожидания при неизвестной дисперсии и для дисперсии при неизвестном математическом ожидании. Уровень значимости $\alpha = 0{,}05$.

1) Построим доверительный интервал для математического ожидания a при неизвестной дисперсии σ^2. Уровень значимости $\alpha = 0{,}05$ (надежность $P = 0{,}95$).

Для математического ожидания доверительный интервал имеет вид:

$$\bar{x}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2},n-1}<a<\bar{x}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2},n-1}.$$

Находим значение $100\frac{\alpha}{2}$-процентной точки распределения Стьюдента с $n-1=29$ степенями свободы:

$$t_{\frac{\alpha}{2},n-1}=t(2{,}5\,\%;\,29)\approx 2{,}045;$$

$\bar{x}\approx 0{,}05$; $S^2=\frac{n}{n-1}\tilde{S}^2\approx 1{,}3$; $S\approx 1{,}14$.

Искомый доверительный интервал имеет вид:

$$0{,}05-\frac{1{,}14}{\sqrt{30}}\cdot 2{,}045<a<0{,}05+\frac{1{,}14}{\sqrt{30}}\cdot 2{,}045;$$
$$-0{,}37<a<0{,}48.$$

Таким образом, математическое ожидание a с вероятностью 0,95 принадлежит интервалу $(-0{,}37;\,0{,}48)$.

2) Построим доверительный интервал для дисперсии при неизвестном математическом ожидании.
Уровень значимости $\alpha=0{,}05$ (надежность $P=0{,}95$).

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

Находим значения $100\frac{\alpha}{2}$-процентной и $100\left(1-\frac{\alpha}{2}\right)$-процентной точек распределения χ^2 с $n-1=29$ степенями свободы:

$$\chi^2_{\frac{\alpha}{2},n-1}=\chi^2(2{,}5\,\%;\,29)\approx 45{,}72;$$
$$\chi^2_{1-\frac{\alpha}{2},n-1}=\chi^2(97{,}5\,\%;\,29)\approx 16{,}05.$$

Следовательно, искомый доверительный интервал имеет вид:

$$\frac{(30-1)1{,}3}{45{,}72}<\sigma^2<\frac{(30-1)1{,}3}{16{,}05};$$
$$0{,}82<\sigma^2<2{,}35.$$

Таким образом, дисперсия σ^2 с вероятностью 0,95 принадлежит интервалу $(0{,}82;\,2{,}35)$.

Вопросы для самоконтроля

1. Дайте определение доверительного предела для неизвестного параметра теоретического распределения.
2. Дайте определение доверительного интервала для неизвестного параметра теоретического распределения.
3. Как найти доверительный интервал для математического ожидания в случае нормального теоретического распределения?
4. Как построить доверительный интервал для дисперсии в случае нормального теоретического распределения?

4 Проверка параметрических гипотез

1. Статистическая гипотеза.
2. Параметрическая гипотеза.
3. Критерии проверки статистических гипотез.

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известного распределения.

Статистическая гипотеза называется *параметрической*, если в ней сформулированы предположения относительно значений параметров функции распределения известного вида.

Нулевой (*основной*) называют выдвинутую гипотезу H_0.

Конкурирующей (*альтернативной*) называют гипотезу H_1, которая противоречит нулевой.

Параметрическая гипотеза называется *простой*, если содержит только одно предположение относительно параметра (например, если a – математическое ожидание нормально распределенной случайной величины, то гипотеза $H_0 : a = 0$ – простая).

Параметрическая гипотеза называется *сложной*, если она состоит из конечного или бесконечного числа простых гипотез (например, если a – математическое ожидание нормально распределенной случайной величины, то гипотеза $H_0 : a > 0{,}5$ – сложная).

При проверке гипотезы могут быть допущены ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная нулевая гипотеза. Вероятность ошибки первого рода называют *уровнем значимости* и обозначают α.

Ошибка второго рода состоит в том, что будет принята неправильная нулевая гипотеза. Вероятность ошибки второго рода обозначают β.

Статистическим критерием (*статистикой критерия*) называют случайную величину, которая служит для проверки гипотезы.

Наблюдаемым (*эмпирическим*) значением называют то значение критерия, которое вычислено по выборке.

Критической областью называют совокупность значений критерия, при которых нулевую гипотезу отвергают.

Областью принятия гипотезы (*областью допустимых значений*) называют совокупность значений критерия, при которых нулевую гипотезу принимают.

Основной принцип проверки статистических гипотез: если наблюдаемое значение критерия принадлежит критической области, то нулевую гипотезу отвергают; если наблюдаемое значение критерия принадлежит области принятия гипотезы, то гипотеза согласуется с экспериментальными данными.

Правосторонней называют критическую область вида $K=[k_{кр};+\infty)$.

Левосторонней называют критическую область вида $K=(-\infty;k_{кр}]$.

Двусторонней называют критическую область вида $K=(-\infty;k_1]\cup[k_2;+\infty)$.

Проверка гипотез о равенстве математического ожидания случайной величины гипотетическому (предполагаемому) значению.

1) Пусть выборка *X* получена из нормального распределения с параметрами *a* и σ^2. Предположим, что дисперсия σ^2 известна.

При заданном уровне значимости α проверяется нулевая гипотеза

$$H_0: a=a_0,$$
$$H_1: a\neq a_0.$$

Критическая область имеет вид

$$K=(-\infty;-u_{кр}]\cup[u_{кр};+\infty),$$

где $u_{кр}$ – корень уравнения $\Phi(u_{кр})=\dfrac{1-\alpha}{2}$,

$\Phi(x)$ – функция Лапласа.

Для проверки нулевой гипотезы вычисляется наблюдаемое значение критерия

$$u_{набл}=\frac{\left(\bar{x}-a_0\right)\cdot\sqrt{n}}{\sigma}.$$

Если $u_{набл}\notin K$, т. е. $|u_{набл}|<u_{кр}$, то нет оснований отвергнуть нулевую гипотезу; если $u_{набл}\in K$, т. е. $|u_{набл}|\geq u_{кр}$, то нулевую гипотезу отвергают.

Пример 4.1 Пусть выборка *X* объема $n=30$ получена из нормального распределения с известным средним квадратическим отклонением $\sigma=1$, выборочное среднее $\bar{x}=0{,}05$. При уровне значимости $\alpha=0{,}01$ проверить нулевую гипотезу $H_0: a=0$ при альтернативной гипотезе $H_1: a\neq 0$.

Найдем наблюдаемое значение критерия

$$u_{набл}=\frac{\left(\bar{x}-a_0\right)\cdot\sqrt{n}}{\sigma}=\frac{\left(0{,}05-0\right)\cdot\sqrt{30}}{1}\approx 0{,}27.$$

По таблицам значений функции Лапласа найдем критическую точку $u_{\text{кр}} = 2{,}58$. Критическая область $K = (-\infty;\, -2{,}58] \cup [2{,}58;\, +\infty)$. Так как $u_{\text{набл}} \notin K$, то нет оснований отвергнуть нулевую гипотезу.

2) Пусть выборка X получена из нормального распределения с неизвестными параметрами a и σ^2.

При заданном уровне значимости α проверяется нулевая гипотеза

$$H_0 : a = a_0,$$

$$H_1 : a \neq a_0.$$

Критическая область имеет вид

$$K = \left(-\infty; -t_{\frac{\alpha}{2}, n-1} \right] \cup \left[t_{\frac{\alpha}{2}, n-1}; +\infty \right),$$

где $t_{\frac{\alpha}{2}, n-1} = t\left(100\frac{\alpha}{2}\%;\, n-1 \right)$ – $100\frac{\alpha}{2}$-процентная точка распределения Стьюдента с $(n-1)$ степенями свободы.

Для проверки нулевой гипотезы вычисляется наблюдаемое значение критерия

$$t_{\text{набл}} = \frac{\bar{x} - a_0}{S}\sqrt{n}.$$

Если $t_{\text{набл}} \notin K$, т. е. $\left| t_{\text{набл}} \right| < t_{\frac{\alpha}{2}, n-1}$, то нет оснований отвергнуть нулевую гипотезу; если $t_{\text{набл}} \in K$, т. е. $\left| t_{\text{набл}} \right| \geq t_{\frac{\alpha}{2}, n-1}$, то нулевую гипотезу отвергают.

Пример 4.2 Пусть выборка X из примера 1.1 получена из нормального распределения. При уровне значимости $\alpha = 0{,}01$ проверить нулевую гипотезу $H_0 : a = 0$ при альтернативной гипотезе $H_1 : a \neq 0$.

Найдем наблюдаемое значение критерия

$$t_{\text{набл}} = \frac{\bar{x} - a_0}{S}\sqrt{n} = \frac{0{,}05 - 0}{1{,}14}\sqrt{30} \approx 0{,}24.$$

По таблице процентных точек распределения Стьюдента найдем $t_{0{,}005;29} = 2{,}76$. Критическая область $K = (-\infty;\, -2{,}76] \cup [2{,}76;\, +\infty)$. Так как $t_{\text{набл}} \notin K$, то нет оснований отвергнуть нулевую гипотезу.

Проверка гипотез о равенстве математических ожиданий двух случайных величин, имеющих нормальное распределение.

1) Пусть исследуются две случайные величины X и Y, каждая из которых подчиняется нормальному закону: $X \sim N(a_1,\sigma_1^2), Y \sim N(a_2,\sigma_2^2)$, дисперсии σ_1^2,σ_2^2 известны.

При заданном уровне значимости α проверяется нулевая гипотеза

$$H_0: a_1 = a_2,$$
$$H_1: a_1 \neq a_2.$$

Критическая область имеет вид

$$K = (-\infty; -u_{\text{кр}}] \cup [u_{\text{кр}}; +\infty),$$

где $u_{\text{кр}}$ – корень уравнения $\Phi(u_{\text{кр}}) = \dfrac{1-\alpha}{2}$,

$\Phi(x)$ – функция Лапласа.

Для проверки нулевой гипотезы вычисляется наблюдаемое значение критерия

$$u_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}.$$

Если $u_{\text{набл}} \notin K$, т. е. $|u_{\text{набл}}| < u_{\text{кр}}$, то нет оснований отвергнуть нулевую гипотезу; если $u_{\text{набл}} \in K$, т. е. $|u_{\text{набл}}| \geq u_{\text{кр}}$, то нулевую гипотезу отвергают.

Пример 4.3 Пусть независимые выборки X и Y, объем которых $n_1 = 30$ и $n_2 = 31$, извлечены из нормальных распределений, выборочные средние $\bar{x} = 0{,}05$ и $\bar{y} = 0{,}78$, дисперсии $\sigma_1^2 = 1, \sigma_2^2 = 0{,}58$. При уровне значимости 0,01 проверить нулевую гипотезу $H_0: a_1 = a_2$ при альтернативной гипотезе $H_1: a_1 \neq a_2$ (a_1, a_2 – математические ожидания выборок X и Y соответственно).

Найдем наблюдаемое значение критерия

$$u_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{0{,}05 - 0{,}78}{\sqrt{\dfrac{1}{30} + \dfrac{0{,}58}{31}}} \approx -3{,}2.$$

По таблицам значений функции Лапласа найдем критическую точку $u_{\text{кр}} = 2{,}58$. Критическая область $K = (-\infty; -2{,}58] \cup [2{,}58; +\infty)$. Так как $u_{\text{набл}} \in K$, то нулевую гипотезу отвергают.

2) Пусть исследуются две случайные величины X и Y, каждая из которых подчиняется нормальному закону: $X \sim N(a_1, \sigma_1^2), Y \sim N(a_2, \sigma_2^2)$. Дисперсии σ_1^2, σ_2^2 неизвестны, но предполагается, что $\sigma_1^2 = \sigma_2^2$.

При заданном уровне значимости α проверяется нулевая гипотеза

$$H_0 : a_1 = a_2,$$
$$H_1 : a_1 \neq a_2.$$

Критическая область имеет вид

$$K = \left(-\infty; -t_{\frac{\alpha}{2}, n_1+n_2-2}\right] \cup \left[t_{\frac{\alpha}{2}, n_1+n_2-2}; +\infty\right),$$

где $t_{\frac{\alpha}{2}, n_1+n_2-2} = t\left(100\frac{\alpha}{2}\%; n_1+n_2-2\right)$ – $100\frac{\alpha}{2}$-процентная точка распределения Стьюдента с (n_1+n_2-2) степенями свободы.

Для проверки нулевой гипотезы вычисляется наблюдаемое значение критерия

$$t_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{n_1\tilde{S}_1^2 + n_2\tilde{S}_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}.$$

Если $t_{\text{набл}} \notin K$, т. е. $|t_{\text{набл}}| < t_{\frac{\alpha}{2}, n_1+n_2-2}$, то нет оснований отвергнуть нулевую гипотезу; если $t_{\text{набл}} \in K$, т. е. $|t_{\text{набл}}| \geq t_{\frac{\alpha}{2}, n_1+n_2-2}$, то нулевую гипотезу отвергают.

Пример 4.4 Пусть независимые выборки X и Y, объемы которых $n_1 = 30$ и $n_2 = 31$, извлечены из нормальных распределений, выборочные средние $\bar{x} = 0{,}05$, $\bar{y} = 0{,}78$, смещенные выборочные дисперсии $\tilde{S}_1^2 = 1{,}26$, $\tilde{S}_2^2 = 0{,}27$ соответственно. При уровне значимости 0,01 проверить нулевую гипотезу $H_0 : a_1 = a_2$ при альтернативной гипотезе $H_1 : a_1 \neq a_2$ (a_1, a_2 – математические ожидания выборок X и Y соответственно).

Найдем наблюдаемое значение критерия

$$t_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{n_1\tilde{S}_1^2 + n_2\tilde{S}_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \frac{0{,}05 - 0{,}78}{\sqrt{\dfrac{30 \cdot 1{,}26 + 31 \cdot 0{,}27}{30+31-2}\left(\dfrac{1}{30} + \dfrac{1}{31}\right)}} \approx -3{,}22.$$

По таблице процентных точек распределения Стьюдента найдем $t_{0,005;59}=2{,}66$. Критическая область $K=(-\infty;-2{,}66]\cup[2{,}66;+\infty)$. Так как $t_{\text{набл}}\in K$, то нулевую гипотезу отвергают.

Проверка гипотезы о дисперсиях двух случайных величин, распределенных по нормальному закону.

Пусть заданы две независимые выборки, извлеченные из нормальных распределений.

При заданном уровне значимости α проверяется нулевая гипотеза

$$H_0:\sigma_1^2=\sigma_2^2,$$
$$H_1:\sigma_1^2\neq\sigma_2^2.$$

Для проверки нулевой гипотезы вычисляется наблюдаемое значение критерия

$$F_{\text{набл}}=\frac{S_{\text{Б}}^2}{S_{\text{М}}^2}.$$

Критическая область имеет вид

$$K=\left[F_{\frac{\alpha}{2};n_1-1,n_2-1};+\infty\right),$$

где $F_{\frac{\alpha}{2};n_1-1,n_2-1}=F(100\frac{\alpha}{2}\%;n_1-1,n_2-1)$ – $100\frac{\alpha}{2}$-процентная точка распределения Фишера с (n_1-1) и (n_2-1) степенями свободы,

n_1, n_2 – объемы выборок с большей и меньшей дисперсиями соответственно.

Если $F_{\text{набл}}\notin K$, т. е. $F_{\text{набл}}<F_{\frac{\alpha}{2};n_1-1,n_2-1}$, то нет оснований отвергнуть нулевую гипотезу; если $F_{\text{набл}}\in K$, т. е. $F_{\text{набл}}\geq F_{\frac{\alpha}{2};n_1-1,n_2-1}$, то нулевую гипотезу отвергают.

*Пример 4.*5 Пусть независимые выборки X и Y, объем которых $n_1=30$ и $n_2=31$, извлечены из нормальных распределений, несмещенные выборочные дисперсии $S_1^2=1{,}3$, $S_2^2=0{,}27$. При уровне значимости 0,1 проверить нулевую гипотезу $H_0:\sigma_1^2=\sigma_2^2$ при альтернативной гипотезе $H_1:\sigma_1^2\neq\sigma_2^2$ (σ_1^2,σ_2^2 – дисперсии выборок X и Y соответственно).

Найдем наблюдаемое значение критерия

$$F_{\text{набл}} = \frac{S_{\text{Б}}^2}{S_{\text{М}}^2} = \frac{1,3}{0,27} \approx 4,81.$$

По таблице процентных точек распределения Фишера найдем $F_{0,05;29,30} = 1,62$. Критическая область $K = [1,62; +\infty)$. Так как $F_{\text{набл}} \in K$, то нулевую гипотезу отвергают.

Вопросы для самоконтроля

1. Дайте определение статистической гипотезы.
2. Что такое статистический критерий?
3. Какая гипотеза называется параметрической?

5 Гипотезы и критерии согласия

1. Гипотезы и критерии согласия.
2. Критерий согласия χ^2 - Пирсона.

Пусть $x_1, x_2,..., x_n$ – выборка из распределения теоретической случайной величины ξ с неизвестной функцией распределения $F(x)$. Проверяется гипотеза $H_0 : F(x) = F_0(x)$, где $F_0(x)$ – заданная функция распределения. Гипотезу такого вида называют *гипотезой согласия (непараметрической гипотезой)*, а критерии для ее проверки – *критериями согласия (непараметрическими критериями)*.

Для проверки гипотез согласия выбирают некоторую меру отклонения

$$\upsilon_n = \upsilon_n(F_n^*(x), F(x)) = \upsilon_n(x_1, x_2,..., x_n)$$

эмпирической функции распределения $F_n^*(x)$ от теоретической функции распределения $F(x)$. В зависимости от вида υ_n получаются разные критерии согласия.

Критическая область имеет вид

$$K = \{x : \upsilon \geq \upsilon_{\text{крит.}}\},$$

где случайная величина υ имеет предельное распределение υ_n при $n \to \infty$.

По заданному уровню значимости α $\upsilon_{\text{крит}}$ находится из уравнения

$$P\{\upsilon_0 \geq \upsilon_{\text{крит.}}\} = \alpha,$$

где $\upsilon_0 = \upsilon|_{H_0}$ – значение υ в предположении, что верна гипотеза H_0.

Критерий проверки гипотезы H_0 строится следующим образом:

– если $\upsilon_0 \in K$, то гипотеза H_0 отвергается;

– если $\upsilon_0 \notin K$, то гипотеза H_0 согласуется с экспериментальными данными.

Критерий согласия χ^2- Пирсона. При уровне значимости α проверяется гипотеза

$$H_0 : F(x) = F_0(x),$$

где $F_0(x)$ – заданная функция распределения.

1) Числовая ось разбивается на k непересекающихся интервалов $\Delta_1 = (-\infty, z_1)$, $\Delta_2 = [z_1, z_2)$, …, $\Delta_k = [z_{k-1}, +\infty)$. Обозначим $z_0 = -\infty, z_k = +\infty$.

2) Находятся частоты n_i – число выборочных значений, попавших в интервал Δ_i, $i = 1,...,k$.

3) Вычисляются $p_i^0 = F_0(z_i) - F_0(z_{i-1})$ – теоретические вероятности попадания в интервал $\Delta_i = [z_{i-1}, z_i)$, $i = 1,...,k$.

4) Статистика критерия имеет вид

$$\chi^2_{\text{набл}} = \sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0}.$$

5) Для выбранного уровня значимости α по таблице χ^2-распределения находим $\chi^2_{\alpha;l} = \chi^2(100\alpha\%; l)$ – 100α-процентную точку распределения χ^2 с $l = k - r - 1$ степенями свободы, где r – число неизвестных параметров теоретического распределения.

6) Схема принятия решения имеет вид:

– если $\chi^2_{\text{набл}} \ge \chi^2_{\alpha;l}$, то гипотеза H_0 отвергается;

– если $\chi^2_{\text{набл}} < \chi^2_{\alpha;l}$, то говорят, что гипотеза H_0 согласуется с экспериментальными данными.

Пример 5.1 Для выборки X из примера 1.1, используя критерий согласия χ^2-Пирсона, при уровне значимости 0,05 проверить гипотезу о том, что распределение наблюдаемой случайной величины ξ не противоречит нормальному закону с параметрами, вычисленными по выборке.

Поскольку параметры нормального распределения (математическое ожидание a и дисперсия σ^2 неизвестны), выдвигаем гипотезу $H_0 : \xi \sim N(\bar{x}, S^2)$, то есть

$$H_0 : F(x) = \frac{1}{S\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\bar{x})^2}{2S^2}} dt.$$

Числовая ось разбивается на $k = 6$ непересекающихся промежутков. Для вычисления теоретических вероятностей p_i^0 попадания случайной величины ξ в интервал Δ_i, $i = 1,...,k$, можно использовать функцию Лапласа:

$$p_i^0 = F_0(z_i) - F_0(z_{i-1}) = \Phi\left(\frac{z_i - \bar{x}}{S}\right) - \Phi\left(\frac{z_{i-1} - \bar{x}}{S}\right).$$

Вычисления приводятся в таблице 6.

Таблица 6

Интервал	Эмпирические частоты n_i	Теоретические вероятности попадания в интервал $p_i^0 = F_0(z_i) - F_0(z_{i-1})$	Теоретические частоты np_i^0	$\frac{(n_i - np_i^0)^2}{np_i^0}$
$(-\infty; -1{,}36)$	3	0,11	3,3	0,03
$[-1{,}36; -0{,}56)$	6	0,19	5,7	0,02
$[-0{,}56; 0{,}24)$	10	0,27	8,1	0,45
$[0{,}24; 1{,}04)$	3	0,24	7,2	2,45
$[1{,}04; 1{,}84)$	7	0,13	3,9	2,46
$[1{,}84; +\infty)$	1	0,06	1,8	0,36
	30	1,00		5,76

Таким образом, статистика критерия $\chi^2_{\text{набл}} = \sum_{i=1}^{6} \frac{(n_i - np_i^0)^2}{np_i^0} \approx 5{,}76$.

Число неизвестных параметров теоретического распределения $r = 2$ (математическое ожидание a и дисперсия σ^2).

По заданному уровню значимости $\alpha = 0{,}05$ находим значение процентной точки распределения χ^2 с $k - r - 1 = 3$ степенями свободы

$$\chi^2_{\alpha; k-r-1} = \chi^2(5\,\%; 3) \approx 7{,}8.$$

Поскольку $\chi^2_{\text{набл}} \approx 5{,}76 < \chi^2_{0,05;3} \approx 7{,}8$, делаем вывод, что гипотеза H_0 согласуется с экспериментальными данными.

Пример 5.2 Для выборки X из примера 1.1, используя критерий согласия χ^2- Пирсона, при уровне значимости 0,05 проверить гипотезу о том, что распределение наблюдаемой случайной величины ξ не противоречит равномерному закону.

Поскольку границы интервала предполагаемого равномерного распределения, на котором случайная величина ξ принимает свои значения, нам неизвестны, можно выдвинуть гипотезу о равномерном распределении случайной величины ξ на промежутке $[x_{(1)}; x_{(30)}] = [-2{,}16; 2{,}65]$. Оценки границ интервала предполагаемого равномерного распределения были получены методом максимального правдоподобия.

$$H_0 : F(y) = F_0(y) = \begin{cases} 0, & x < -2{,}16, \\ \dfrac{y+2{,}16}{4{,}81}, & -2{,}16 \le x \le 2{,}65, \\ 1, & x > 2{,}65. \end{cases}$$

Вычисления приводятся в таблице 7.

Таблица 7

Интервал	Эмпирические частоты n_i	Теоретические вероятности попадания в интервал $p_i^0 = F_0(z_i) - F_0(z_{i-1})$	Теоретические частоты np_i^0	$\dfrac{(n_i - np_i^0)^2}{np_i^0}$
$[-2{,}16; -1{,}36)$	3	$\frac{1}{6} \approx 0{,}16667$	5	0,8
$[-1{,}36; -0{,}56)$	6	$\frac{1}{6} \approx 0{,}16667$	5	0,2
$[-0{,}56; 0{,}24)$	10	$\frac{1}{6} \approx 0{,}16667$	5	5
$[0{,}24; 1{,}04)$	3	$\frac{1}{6} \approx 0{,}16667$	5	0,8
$[1{,}04; 1{,}84)$	7	$\frac{1}{6} \approx 0{,}16667$	5	0,8
$[1{,}84; 2{,}65]$	1	$\frac{1}{6} \approx 0{,}16667$	5	3,2
	30	1,00		10,8

Статистика критерия $\chi^2_{\text{набл}} = \sum\limits_{i=1}^{6} \dfrac{(n_i - np_i^0)^2}{np_i^0} \approx 10{,}8$.

Число неизвестных параметров теоретического распределения $r = 2$ (границы интервала предполагаемого равномерного распределения). По заданному уровню значимости $\alpha = 0{,}05$ находим значение процентной точки распределения χ^2 с $k - r - 1 = 3$ степенями свободы $\chi^2_{\alpha;k-r-1} = \chi^2(5\%; 3) \approx 7{,}8$.

Поскольку $\chi^2_{\text{набл}} \approx 10{,}8 > \chi^2_{0{,}05;3} \approx 7{,}8$, то гипотеза H_0 отвергается.

Заметим, что оценки границ интервала предполагаемого равномерного распределения могли быть получены и другими методами, например, методом моментов.

Вопросы для самоконтроля

1. Дайте определение критерия согласия.
2. Как найти статистику критерия согласия χ^2-Пирсона?
3. Как построить критическую область в критерии согласия χ^2-Пирсона?

6 Однофакторный дисперсионный анализ

1. Однофакторный дисперсионный анализ статистических данных.
2. Проверка гипотезы о значимости статистического влияния фактора на математические ожидания исследуемых случайных величин.

Дисперсионный анализ – статистический метод, предназначенный для оценки влияния различных факторов на результат эксперимента, а также для последующего планирования аналогичных экспериментов.

По числу факторов, влияние которых исследуется, различают *однофакторный* и *многофакторный* дисперсионный анализ.

Пусть на случайную величину о воздействует фактор F, который имеет k постоянных уровней $F_1, F_2, ..., F_k$. На каждом уровне произведено по n_i испытаний. Результаты наблюдений – числа x_{ij} $(i = \overline{1,k};\ j = \overline{1,n_i})$.

Будем рассматривать результаты измерений x_{ij} как выборки из нормальных распределений: $\xi_i \sim N(a_i, \sigma^2)$ $(i = \overline{1,k})$.

По выборочным данным вычисляются:

– групповые средние

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \quad i = \overline{1,k};$$

– общая средняя

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}, \quad n = \sum_{i=1}^{k} n_i .$$

Проверяется нулевая гипотеза

$$H_0 : a_1 = a_2 = ... = a_k .$$

Для проверки нулевой гипотезы вычисляется наблюдаемое значение критерия

$$F_{\text{набл}} = \frac{\frac{1}{k-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 n_i}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}.$$

Критическая область имеет вид

$$K = \left[F_{\alpha;k-1,n-k}; +\infty \right),$$

где $F_{\alpha;k-1,n-k} = F(100\alpha\%; k-1, n-k)$ – 100α-процентная точка распределения Фишера с $(k-1)$ и $(n-k)$ степенями свободы.

Если $F_{\text{набл}} \notin K$, т. е. $F_{\text{набл}} < F_{\alpha;k-1,n-k}$, то нет оснований отвергнуть нулевую гипотезу; если $F_{\text{набл}} \in K$, т. е. $F_{\text{набл}} \geq F_{\alpha;k-1,n-k}$, то нулевую гипотезу отвергают.

Формулы, используемые для расчета наблюдаемого значения критерия, приведены в таблице 8.

Таблица 8

Источник изменчивости	Суммы квадратов отклонений	Число степеней свободы	Дисперсия	Наблюдаемое значение критерия
Фактор F (между группами)	$\sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2 n_i$ (факторная)	$k-1$	$S_1^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 n_i$ (факторная)	$F_{\text{набл}} = \frac{S_1^2}{S_2^2}$
Остаточная (внутри групп)	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ (остаточная)	$n-k$	$S_2^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$ (остаточная)	
Общая изменчивость	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$ (общая)	$n-1$	$S^2 = S_1^2 + S_2^2$ (общая)	

Отклонение нулевой гипотезы является статистическим доказательством влияния фактора F на математические ожидания исследуемых случайных величин.

Пример 6.1 Произведено по шесть испытаний на каждом из пяти уровней фактора F. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты испытаний приведены в таблице 9.

Таблица 9

№ испытания	Уровни фактора				
i	F_1	F_2	F_3	F_4	F_5
1	1,37	–0,13	1,16	–1,16	0,93
2	0,11	–0,64	–0,3	2,65	0,01
3	1,56	–0,46	–0,31	1,55	–1,56
4	–0,11	–0,88	1,13	0,29	1,59
5	0,23	–0,56	–0,17	–2,16	–1,13
6	–0,76	1,28	0,6	–0,77	–1,74

Вычислим вспомогательные величины:

– групповые средние

$$\bar{x}_1 = \frac{1{,}37 + 0{,}11 + 1{,}56 - 0{,}11 + 0{,}23 - 0{,}76}{6} \approx 0{,}4;$$

$$\bar{x}_2 = \frac{-0{,}13 - 0{,}64 - 0{,}46 - 0{,}88 - 0{,}56 + 1{,}28}{6} \approx -0{,}23;$$

$$\bar{x}_3 = \frac{1{,}16 - 0{,}3 - 0{,}31 + 1{,}13 - 0{,}17 + 0{,}6}{6} \approx 0{,}35;$$

$$\bar{x}_4 = \frac{-1{,}16 + 2{,}65 + 1{,}55 + 0{,}29 - 2{,}16 - 0{,}77}{6} \approx 0{,}07;$$

$$\bar{x}_5 = \frac{0{,}93 + 0{,}01 - 1{,}56 + 1{,}59 - 1{,}13 - 1{,}74}{6} \approx -0{,}32;$$

– общая средняя

$$\bar{x} = \frac{1{,}62}{30} \approx 0{,}05.$$

$$\sum_{j=1}^{6} \left(x_{1j} - \bar{x}_1\right)^2 \approx 4; \quad \sum_{j=1}^{6} \left(x_{2j} - \bar{x}_2\right)^2 \approx 3{,}04; \quad \sum_{j=1}^{6} \left(x_{3j} - \bar{x}_3\right)^2 \approx 2{,}46;$$

$$\sum_{j=1}^{6} \left(x_{4j} - \bar{x}_4\right)^2 \approx 16{,}11; \quad \sum_{j=1}^{6} \left(x_{5j} - \bar{x}_5\right)^2 \approx 9{,}53; \quad \sum_{i=1}^{5}\sum_{j=1}^{6} \left(x_{ij} - \bar{x}_i\right)^2 \approx 35{,}14.$$

$$\sum_{i=1}^{5} \left(\bar{x}_i - \bar{x}\right)^2 n_i \approx 2{,}56.$$

Результаты расчетов приведены в таблице 10.

Таблица 10

Источник изменчивости	Суммы квадратов отклонений	Число степеней свободы	Дисперсия	Наблюдаемое значение критерия
Фактор F (между группами)	2,56	4	0,64	$F_{\text{набл}} = \frac{0{,}64}{1{,}41} \approx 0{,}45$
Остаточная (внутри групп)	35,14	25	1,41	
Общая изменчивость	37,70	29	1,3	

По таблице процентных точек распределения Фишера найдем $F_{0,05;4,25} = 2{,}76.$ Так как $F_{\text{набл}} \notin K$, то нет оснований отвергнуть нулевую гипотезу. Другими словами, групповые средние не различаются значимо.

Вопросы для самоконтроля

1. Сформулируйте основную задачу дисперсионного анализа.
2. Как найти факторную и остаточную дисперсии?
3. Как построить критическую область?

7 Корреляционный и регрессионный анализ

1. Корреляционный анализ статистических данных.
2. Регрессионный анализ статистических данных.

Статистические связи между переменными можно изучать методами дисперсионного, корреляционного и регрессионного анализа. Методами дисперсионного анализа устанавливается наличие влияния заданного фактора на изучаемый процесс. Корреляционный анализ позволяет оценить силу такой связи, а методами регрессионного анализа можно выбрать конкретную математическую модель и оценить ее адекватность.

Корреляционная связь – это согласованное изменение признаков, отражающее тот факт, что изменчивость одного признака находится в соответствии с изменчивостью другого. Парная корреляция изучает взаимосвязи между двумя случайными величинами, множественная – между большим числом величин.

Основная задача корреляционного анализа – выявление и оценка связи между случайными величинами, основная задача регрессионного анализа – установление формы и изучение зависимости между случайными величинами (рисунок 3).

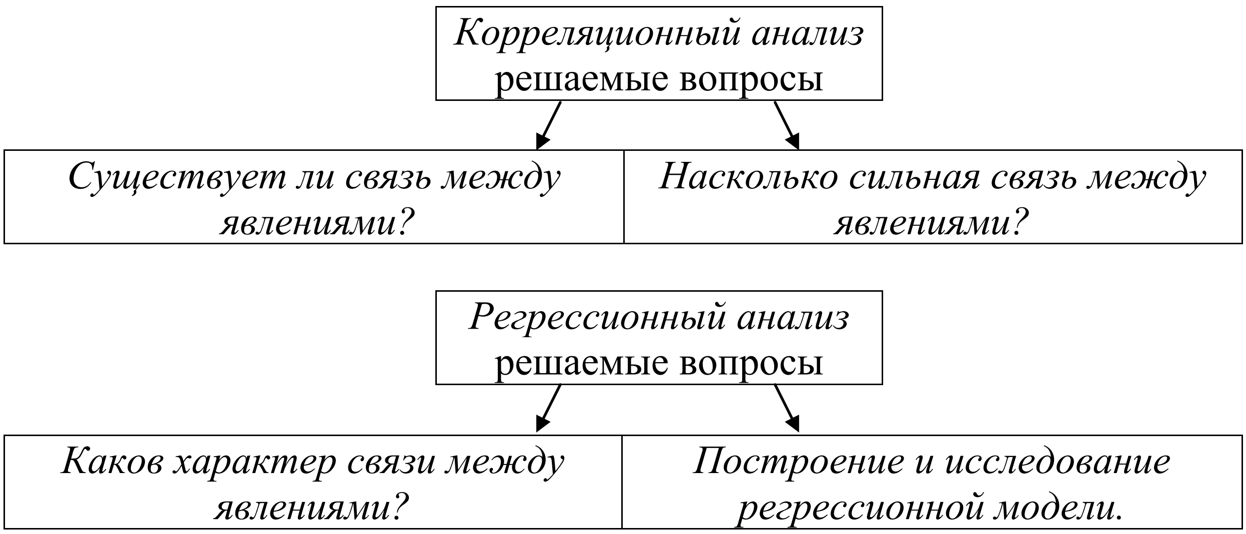


Рисунок 3

Элементы корреляционного анализа. Пусть $(x_1, y_1),...,(x_n, y_n)$ – выборка объема n из наблюдений случайной величины (ξ, η), имеющей двумерное нормальное распределение. Изображая элементы выборки точками в декартовой системе координат, получим *диаграмму* *рассеивания или корреляционное поле*. Иногда по виду корреляционного поля можно сделать предположение о наличии и характере связи между случайными величинами ξ и η.

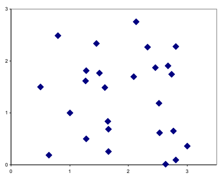
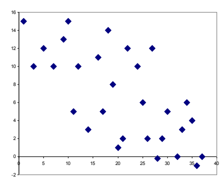
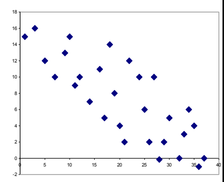
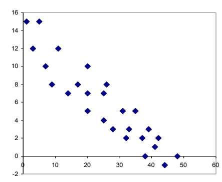
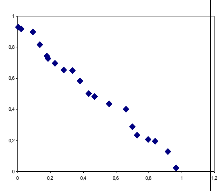
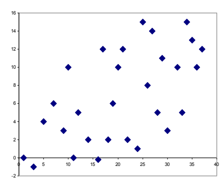
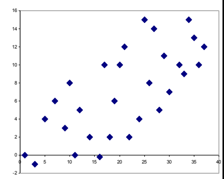
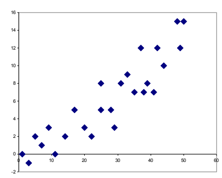
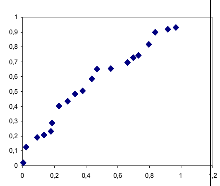
Выборочным *коэффициентом корреляции* называется число

$$r_{\text{выб.}} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\cdot\overline{y}}{\widetilde{S}_x \widetilde{S}_y}.$$

Можно показать, что $|r_{\text{выб.}}| \le 1$.

В таблице 11 приведены возможные формы корреляционного поля в зависимости от значения выборочного коэффициента корреляции.

Таблица 11

<table>
<tr><td>$-0{,}2 < r_{\text{выб.}} < 0{,}2$</td><td>$-0{,}5 < r_{\text{выб.}} \le -0{,}2$</td><td>$-0{,}75 < r_{\text{выб.}} \le -0{,}5$</td><td>$-0{,}95 < r_{\text{выб.}} \le -0{,}75$</td><td>$-1 \le r_{\text{выб.}} \le -0{,}95$</td></tr>
<tr><td rowspan="3"></td><td></td><td></td><td></td><td></td></tr>
<tr><td>$0{,}2 \le r_{\text{выб.}} < 0{,}5$</td><td>$0{,}5 \le r_{\text{выб.}} < 0{,}75$</td><td>$0{,}75 \le r_{\text{выб.}} < 0{,}95$</td><td>$0{,}95 \le r_{\text{выб.}} \le 1$</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
</table>

На практике большой интерес представляет задача проверки гипотезы о *значимости* корреляционной связи между случайными величинами, т. е. значимости отклонения коэффициента корреляции от нуля. Пусть $r_{\text{выб.}}$ – выборочный коэффициент корреляции. При заданном уровне значимости α проверяется гипотеза $H_0: r = 0$ о равенстве нулю теоретического коэффициента корреляции. Если нулевая гипотеза будет отвергнута, то говорят о значимости коэффициента корреляции, а значит о том, что случайные величины ξ и η коррелированы. Если нулевая гипотеза принимается, то коэффициент корреляции незначим, и случайные величины ξ и η некоррелированы.

Статистика критерия имеет вид

$$t_{\text{набл}} = r_{\text{выб.}}\sqrt{\frac{n-2}{1-r_{\text{выб.}}^2}}.$$

Находится $t_{\frac{\alpha}{2};n-2} = t(100\frac{\alpha}{2}\%, n-2)$ – значение процентной точки распределения Стьюдента с $(n-2)$ степенями свободы.

Схема принятия решения выглядит следующим образом:

– если $|t_{\text{набл}}| = \left| r_{\text{выб.}} \sqrt{\frac{n-2}{1-r^2_{\text{выб.}}}} \right| < t_{\frac{\alpha}{2};n-2}$, то нет оснований отвергать нулевую гипотезу, коэффициент корреляции не значим, а ξ и η некоррелированы;

– если $|t_{\text{набл}}| = \left| r_{\text{выб.}} \sqrt{\frac{n-2}{1-r^2_{\text{выб.}}}} \right| \geq t_{\frac{\alpha}{2};n-2}$, то гипотеза отвергается, и коэффициент корреляции значимо отличается от нуля, а ξ и η коррелированы.

Пример 7.1 Предполагая, что $(x_1, y_1),...,(x_n, y_n)$ – выборка из наблюдений случайной величины (ξ, η), имеющей двумерное нормальное распределение, вычислить выборочный коэффициент корреляции и при заданном уровне значимости $\alpha = 0{,}05$ проверить гипотезу о равенстве нулю теоретического коэффициента корреляции.

$x_{(i)}$	1,37	0,11	1,56	–0,11	0,23	–0,76	–0,13	–0,64	–0,46	–0,88
$y_{(i)}$	0,08	0,64	1,59	1,75	0,74	0,89	1,44	0,72	1,26	0,03

–0,56	1,28	1,16	–0,30	–0,31	1,13	–0,17	0,60	–1,16	2,65	1,55
0,92	0,51	0,52	0,43	1,06	0,47	0,11	1,27	1,33	0,32	1,48

0,29	–2,16	–0,77	0,93	0,01	–1,56	1,59	–1,13	–1,74
1,61	1,47	0,98	0,54	0,59	0,34	0,17	0,20	0,27

Выборочный коэффициент корреляции:

$$r_{\text{выб.}} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}}{\tilde{S}_x \tilde{S}_y} \approx \frac{\frac{1}{30}(1{,}37 \cdot 0{,}08 + ... + (-1{,}74) \cdot 0{,}27) - 0{,}05 \cdot 0{,}79}{1{,}12 \cdot 0{,}52} \approx -0{,}08.$$

Проверяется гипотеза $H_0 : r = 0$.

Статистика критерия имеет вид

$$t_{\text{набл}} = r_{\text{выб.}} \sqrt{\frac{n-2}{1-r_{\text{выб.}}^2}} \approx -0{,}08 \cdot \sqrt{\frac{30-2}{1-(-0{,}08)^2}} \approx -0{,}4 .$$

Находим значение процентной точки распределения Стьюдента $t_{\frac{0,05}{2};30-2} = t(2{,}5\,\%, 28) \approx 2{,}048$.

Поскольку $|t_{\text{набл}}| \approx |-0{,}4| < t_{0,025;28} \approx 2{,}048$, то нет оснований отвергать нулевую гипотезу H_0, и коэффициент корреляции не значим, а ξ и η некоррелированы.

Линейный регрессионный анализ. Часто требуется определить, как зависит наблюдаемая случайная величина от одной или нескольких других величин. Регрессионный анализ – раздел математической статистики, изучающий связь между зависимой переменной и одной или несколькими независимыми переменными.

Наблюдаются значения $(x_1, y_1),...,(x_n, y_n)$ двумерной случайной величины ξ, η . Исследуется зависимость случайной величины η от случайной величины ξ.

В общем случае регрессионная модель имеет вид

$$y = f(x, \beta_0, \beta_1, ..., \beta_k).$$

Параметры $\beta_0, \beta_1, ..., \beta_k$ называются *коэффициентами регрессии*.

Одна из задач регрессионного анализа – оценка коэффициентов регрессии. Для оценки коэффициентов регрессии, как правило, используется *метод наименьших квадратов*: в качестве оценок принимаются такие значения параметров, которые минимизируют сумму квадратов отклонений наблюдаемых значений y_i от $\tilde{y}_i = f(x_i, \beta_0, \beta_1, ..., \beta_k)$, $(i = 1, ..., n)$, т. е. метод наименьших квадратов основан на минимизации суммы квадратов:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i \right)^2 \to \min .$$

Если предположить, что связь между переменными линейна, то соответствующая регрессионная модель имеет вид

$$y=\beta_0+\beta_1 x,$$

где β_0 и β_1 – коэффициенты линейной регрессии.

Для линейной модели регрессии задача минимизации имеет вид:

$$\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2\rightarrow \min .$$

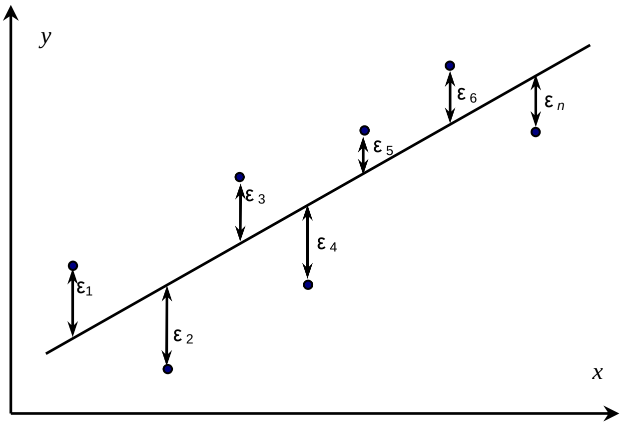


Рисунок 4

На рисунке 4 изображены отклонения $\varepsilon_i=y_i-\tilde{y}_i,\ i=1,...,n$.

Необходимым условием минимума функции двух переменных β_0 и β_1 является равенство ее частных производных по β_0 и β_1 нулю:

$$\begin{cases}\sum\limits_{i=1}^{n}x_i(y_i-\beta_1 x_i-\beta_0)=0,\\ \sum\limits_{i=1}^{n}(y_i-\beta_1 x_i-\beta_0)=0.\end{cases}$$

Решение системы дает искомые оценки коэффициентов линейной регрессии:

$$b_0=\bar{y}-r_{\text{выб.}}\frac{S_y}{S_x}\bar{x};$$

$$b_1=r_{\text{выб.}}\frac{S_y}{S_x}.$$

Здесь b_0 и b_1 – оценки β_0 и β_1 соответственно.

Пример 7.2 По данным наблюдений двумерной случайной величины ξ,η построить выборочное уравнение линейной регрессии η на ξ и выборочное уравнение линейной регрессии ξ на η.

1) Выборочное уравнение линейной регрессии η на ξ имеет вид:

$$y = b_1 x + b_0,$$

где $b_0 = \bar{y} - r_{\text{выб.}} \dfrac{S_y}{S_x} \bar{x} \approx 0{,}79 - (-0{,}08) \cdot \dfrac{0{,}51}{1{,}12} \cdot 0{,}05 \approx 0{,}79;$

$$b_1 = r_{\text{выб.}} \frac{S_y}{S_x} \approx (-0{,}08) \cdot \frac{0{,}51}{1{,}12} \approx -0{,}03.$$

Таким образом, искомое уравнение:

$$y = -0{,}03x + 0{,}79.$$

2) Выборочное уравнение линейной регрессии ξ на η имеет вид:

$$x = b_1^* y + b_0^*,$$

где $b_0^* = \bar{x} - r_{\text{выб.}} \dfrac{S_x}{S_y} \bar{y} \approx 0{,}05 - (-0{,}08) \cdot \dfrac{1{,}12}{0{,}51} \cdot 0{,}79 \approx 0{,}19;$

$$b_1^* = r_{\text{выб.}} \frac{S_x}{S_y} \approx (-0{,}08) \cdot \frac{1{,}12}{0{,}51} \approx -0{,}17.$$

Таким образом, искомое уравнение:

$$x = -0{,}17y + 0{,}19.$$

На рисунке 5 приводятся графики уравнений линейной регрессии.

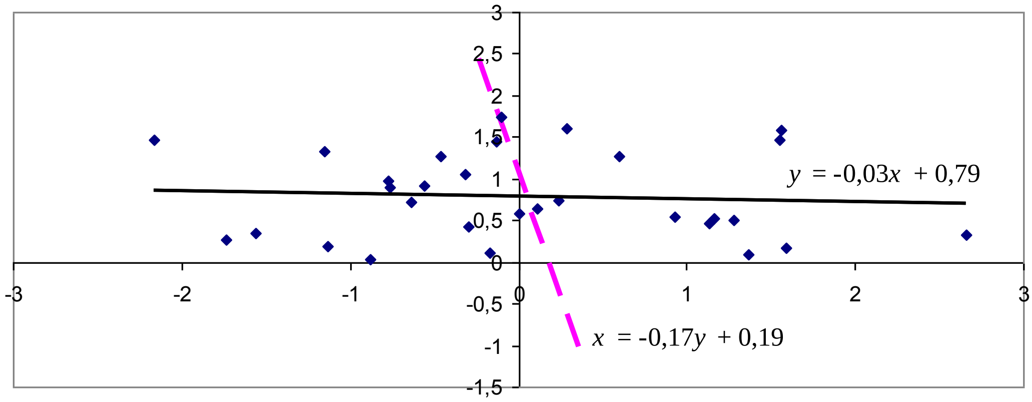


Рисунок 5

Вопросы для самоконтроля

1. Какие задачи решает корреляционный анализ?
2. Как найти выборочный коэффициент корреляции?
3. Какие задачи решает регрессионный анализ?
4. Как построить выборочные уравнения линейной регрессии?

Литература

1 Бочаров, П. П. Теория вероятностей и математическая статистика / П. П. Бочаров, А. В. Печинкин. – М.: ФИЗМАТЛИТ, 2005. – 296 с.

2 Валеев, С. Г. Практикум по прикладной статистике: учеб. пособие / С. Г. Валеев, В. Н. Клячкин. – Ульяновск: УлГТУ, 2008. – 129 с.

3 Герасимович, А. И. Математическая статистика / А. И. Герасимович, Я. И. Матвеева. – Мн.: «Вышэйшая школа», 1978. – 200 с.

4 Гмурман, В. Е. Теория вероятностей и математическая статистика: учеб. пособие для вузов / В. Е. Гмурман. – 9-е изд., стер. – М.: Высш. шк., 2003. – 479 с.

5 Горяинов, В. Б. Математическая статистика: учеб. для вузов / В. Б. Горяинов, И. В. Павлов, Г. М. Цветкова и др.; под ред. В. С. Зарубина, А. П. Крищенко. – М.: Изд-во МГТУ им. Н. Э. Баумана, 2001. – 424 с.

6 Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами: учебн. пособие / А. И. Кибзун [и др.]. – М.: ФИЗМАТЛИТ, 2002. – 224 с.

7 Кобзарь, А. И. Прикладная математическая статистика. Для инженеров и научных работников / А. И. Кобзарь. – М.: ФИЗМАТЛИТ, 2006. – 816 с.

8 Кремер, Н. Ш. Теория вероятностей и математическая статистика: учебник для вузов / Н. Ш. Кремер. – М.: ЮНИТИ–ДАНА, 2004. – 573 с.

9 Малинковский, Ю. В. Теория вероятностей и математическая статистика: учеб. пособие: в 2 ч. Ч. 1: Теория вероятностей / Ю. В. Малинковский. – Гомель: ГГУ им. Ф. Скорины, 2004. – 355 с.

10 Малинковский, Ю. В. Теория вероятностей и математическая статистика: учеб. пособие: в 2 ч. Ч. 2: Математическая статистика / Ю. В. Малинковский. – Гомель: ГГУ им. Ф. Скорины, 2004. – 146 с.

11 Степанова, М. Д. Проверка статистических гипотез: учебно-метод. пособие по курсу «Статистические основы индуктивного вывода» для студентов специальности «Искусственный интеллект» / М. Д. Степанова. – Мн.: БГУИР, 2000. – 36 с.

Производственно-практическое издание

Боярович Юлия Сигизмундовна
Дудовская Юлия Евгеньевна

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Практическое руководство

для студентов специальностей
1-31 03 03-01 «Прикладная математика
(научно-производственная деятельность)»,
1-31 03 03-02 «Прикладная математика
(научно-педагогическая деятельность)»

Редактор *В. И. Шкредова*
Корректор *В. В. Калугина*

Подписано в печать 27.12.2012. Формат 60x84 1/16.
Бумага офсетная. Ризография. Усл. печ. л. 2,6.
Уч.-изд. л. 2,8. Тираж 50 экз. Заказ № 694.

Издатель и полиграфическое исполнение :
учреждение образования
«Гомельский государственный университет
имени Франциска Скорины».
ЛИ № 02330/0549481 от 14.05.2009.
Ул. Советская, 104, 246019, г. Гомель.

Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

Ю. С. Боярович, Ю. Е. Дудовская

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Гомель
2012