И. С. Астахова
В. К. Кошмак
А. В. Лисенков

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Псков
Псковский государственный университет
2016

Министерство образования и науки Российской Федерации
Псковский государственный университет

И. С. Астахова
В. К. Кошмак
А. В. Лисенков

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие

Рекомендовано к изданию редакционно-издательским советом
Псковского государственного университета

Псков
Псковский государственный университет
2016

УДК 330.4
ББК 65В6
А91

Рекомендовано к изданию редакционно-издательским советом Псковского государственного университета

Рецензенты:
— В. Н. Мельник, доцент кафедры прикладной информатики в образовании ПГУ, канд. физ.-мат. наук;
— В. Г. Дегтярёв, доктор технических наук, академик МАН ВШ, профессор Санкт-Петербургского государственного университета путей сообщения, Заслуженный деятель науки и техники РФ.

Астахова, И. С., Кошмак, В. К., Лисенков, А. В.
А91 Математическая статистика : Учебное пособие для заочного отделения — Псков : Псковский государственный университет, 2016. — 56 с.
ISBN 978-5-91116-483-6

Методическое пособие по математической статистике предназначено для самостоятельной работы студентов заочной формы обучения над практической частью курса и выполнения контрольных работ.

Для облегчения работы при выполнении задания приведены формулировки определений, основные теоремы, таблицы и разобраны типовые задачи. Приведён список литературы, необходимый для самостоятельного изучения курса.

Пособие разработано в полном соответствии с государственным образовательным стандартом по дисциплине математическая статистика.

УДК 330.4
ББК 65В6

ISBN 978-5-91116-483-6

Введение

Математическая статистика — прикладная математическая дисциплина, родственная теории вероятностей, базируется на её понятиях и методах, но решает свои специфические задачи своими методами — методами количественного анализа массовых случайных явлений.

Пособие предназначено для самостоятельной работы студентов заочного отделения над практической частью курса. Цель контрольной работы — детальная и более тщательная проработка лекционного и практического материала, с целью проверки и контроля степени его усвоения, формирование у студентов предусмотренных рабочей программой навыков.

В приложении приведены краткие статистические таблицы необходимые для решения задач. Номер варианта выбирается по последним двум цифрам зачетной книжки.

Программа курса математическая статистика

1. Генеральная совокупность и выборка. Статистические оценки выборочной совокупности и их свойства. Несмещенные и состоятельные оценки центра распределения и дисперсии. Точность оценок. Доверительная вероятность, доверительный интервал. Доверительные оценки параметров нормального распределения при случайном отборе. Доверительные оценки вероятности.
2. Проверка статистических гипотез. Понятие и виды статистических гипотез. Простые и сложные гипотезы. Критерий и критическая область. Ошибки первого и второго рода. Критерий согласия Пирсона. Проверка гипотез о равенстве средних и долей.
3. Дисперсионный анализ. Модели дисперсионного анализа. Однофакторный дисперсионный анализ. Многофакторный дисперсионный анализ. Оценка влияния одновременно действующих факторов.
4. Корреляционно-регрессионный анализ. Функциональная и корреляционная зависимость. Определение параметров линейного уравнения регрессии методом наименьших квадратов. Коэффициент корреляции и его свойства. Определение параметров нелинейных уравнений регрессии методом наименьших квадратов. Корреляционное отношение и его свойства. Понятие о множественной корреляции.
5. Временные ряды. Анализ составляющих. Методы наименьших квадратов и скользящей средней.
6. Основные понятия многомерного анализа. Методы факторного анализа, область их применения. Метод главных компонент. Классификация объектов, описываемых количественными и качественными признаками.

Выборка. Эмпирическая функция распределения

На практике во многих случаях функция распределения рассматриваемой случайной величины ξ неизвестна; ее определяют по результатам наблюдений, или, как говорят по выборке. Общая совокупность объектов, подвергающаяся изучению, называется ***генеральной совокупностью***.

Выборкой объема n для данной случайной величины ξ называют последовательность $X_1, X_2, ..., X_n$ n независимых наблюдений этой величины, т. е. часть элементов, отобранная на удачу из генеральной совокупности. Таким образом, $X_1, X_2, ..., X_n$ — это совокупность значений, принятых n независимыми случайными величинами $\xi_1, \xi_2, ..., \xi_n$, имеющими тот же закон распределения $F_\xi(X)$, что и рассматриваемая величина ξ. В этом случае говорят, что выборка $X_1, X_2, ..., X_n$ взята из генеральной совокупности величины ξ, а под законом распределения генеральной совокупности понимают закон распределения случайной величины ξ.

Значения $X_1, X_2, ..., X_n$ называются выборочными значениями. По выборке будем:

а) определять вид закона распределения;
б) определять параметры распределения;
в) проверять статистические гипотезы о законе распределения и о его параметрах.

При образовании выборки следует добиваться того, чтобы данные выборочного наблюдения как можно точнее (полнее) воспроизводили характерные свойства генеральной совокупности. Выборка, удовлетворяющая этому условию, называется ***репрезентативной***, т. е. представительной. Чтобы репрезентативность была достигнута надо, чтобы при организации выборки были выполнены условия:

1) генеральная совокупность должна состоять по возможности из однородных объектов;
2) выборка должна быть случайной, должна быть произведена по возможности по определенной методике.

Кроме того, необходимо добиваться, чтобы объем выборки был достаточно большим.

Пусть $X_1, X_2, ..., X_n$ — выборка, x — произвольное число. Обозначим через ν_x количество выборочных значений, меньших x.

Тогда ν_x / n является ***относительной частотой*** попадания выборочных значений левее точки x в данной выборке. Эта частота является функцией от x и называется ***эмпирической функцией распределения*** случайной величины ξ, полученной по данной выборке и обозначается $F_n^*(x)$:

$$F_n^*(x) = \frac{\nu_x}{n}. \qquad (1)$$

Эмпирическая функция распределения обладает всеми свойствами функции распределения: изменяется от 0 до 1, не убывает и непрерывна слева, при этом она растет скачками в каждой из точек $X_1, X_2, ..., X_n$.

Типичный график функции $F_n^*(x)$ имеет ступенчатый вид:

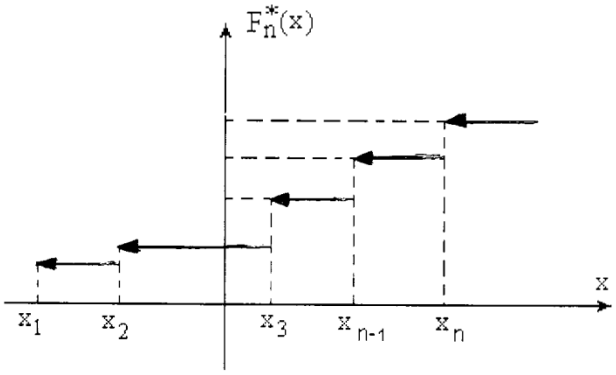


Рис. 1

Эмпирическая функция распределения играет фундаментальную роль в математической статистике. Важнейшее ее свойство состоит в том, что при увеличении числа испытаний над случайной величиной ξ происходит сближение этой функции с теоретической. Согласно закону больших чисел Бернулли при каждом фиксированном x имеем

$$F_n^*(x) \xrightarrow[n \to \infty]{\text{по вероятности}} F_\xi(x).$$

Таким образом, если объем выборки большой, то значение эмпирической функции распределения в каждой точке x может служить приближенным значением (оценкой) теоретической функции распределения $F_\xi(x)$ в этой точке.

Пусть из некоторой генеральной совокупности производится выборка объема n и при этом случайная величина ξ приняла n_1-раз

значение X_1, n_2-раз — значение X_2, …, n_k-раз — значение X_k, $(n_1 + n_2 + ... + n_k = n)$.

Величину n_i называют ***частотой***, соответствующей варианте X_i, а величину $W_i = n_i / n$ — ***относительной частотой***, $\sum_{i=1}^{k} W_i = 1$. Совокупность значений случайной величины, записанных в порядке их возрастания, называется ***вариационным рядом***: $X_1, X_2, ..., X_k$; $X_i < X_{i+1}$, X_i — варианты.

ЭМПИРИЧЕСКИЙ РЯД: $X_1 < X_2 < X_3 < .... < X_k$. $\sum_{i=1}^{k} \frac{n_i}{n} = 1$.

x_1	x_2	x_3	….	x_k
n_1/n	n_2/n	n_3/n	….	n_k/n

Эмпирический ряд — аналог теоретического закона распределения.

Гистограмма и полигон

Эмпирическая функция распределения — удобный способ представления статистических данных. Существуют и другие способы представления статистических данных. Одним из них является построение гистограммы.

При построении гистограммы область задач наблюдаемой случайной величины ξ разбивают на интервалы (разряды), например, на равные, подсчитывают число выборочных значений n_i, попавших в соответствующий интервал и определяют относительные частоты $W_i = n_i / n$.

	$[X_0, X_1[$	$[X_0, X_2[$	…..	$[X_{k-1}, X_k[$
W_i	$\frac{n_0}{n} = W_0$	$\frac{n_1}{n} = W_1$	…..	$\frac{n_{k-1}}{n} = W_{k-1}$

На каждом интервале, как на основании, строят прямоугольники с высотой $\frac{n_i}{nh_i}$, где h_i — длина интервала (в частности $h_i = h$ для всех i).

Полученную при этом фигуру называют ***гистограммой*** (относительных частот) случайной величины ξ.

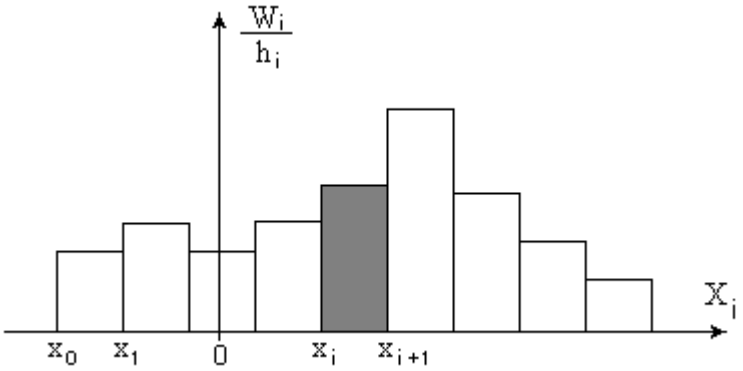


Рис. 2

Таким образом, площадь каждого прямоугольника равна $n_i / n = W_i$, т. е. относительной частоте попадания выборочных значений в соответствующий интервал. Площадь гистограммы относительных частот равна сумме всех относительных частот, т. е. единице (аналогично можно построить гистограмму частот).

Если длина интервала h достаточно мала, а объем выборки n большой, верхнюю границу гистограммы можно рассматривать статистическим аналогом теоретической плотности распределения вероятности — эмпирической плотности вероятности для непрерывной случайной величины (см. рис. 3):

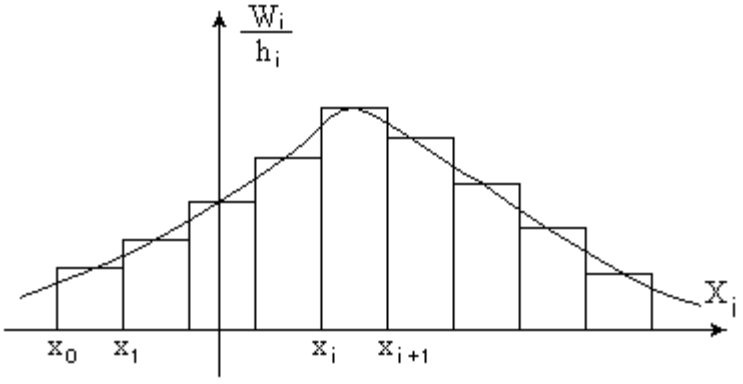


Рис. 3

Если высоту прямоугольников взять равной n_i / h_i или n_i / h, то получим гистограмму частот. Построенная гистограмма позволяет выдвинуть гипотезу о том или ином законе распределения случайной величины по результатам испытаний.

Рассмотренный способ представления статистических данных рекомендуется применять только для непрерывных случайных величин; кроме того, он обладает очевидными недостатками, например, неопределенностью в способе построения интервалов.

Полигон частот — это ломаная, которую можно построить так: если построена гистограмма, то ординаты, соответствующие средним точкам интервалов, последовательно соединяют отрезками прямых. Построенный таким образом кусочно-линейный график является статистическим аналогом теоретической плотности вероятности а) для непрерывной случайной величины; б) многоугольника распределения для дискретной случайной величины.

Если гистограмма не построена, то полигон частот (или относительных частот) можно построить как ломаную, отрезки которой соединяют точки (x_i, n_i) или (x_i, W_i), где x_i — варианты выборки, n_i — соответствующие им частоты, W_i — относительные частоты.

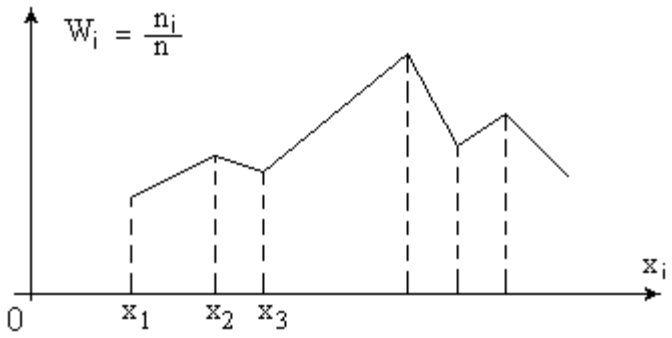


Рис. 4

Пример 1. Пусть (–0,8; 2,9; 4,4; –5,6; 1,1; –3,2) — наблюдавшиеся значения случайной величины ξ. Построить соответствующую функцию распределения $F_n^*(x)$.

Решение: Согласно формулы (1) $F_n^*(x) = \frac{\nu_x}{n}$, где n — объем выборки; ν_x — количество выборочных значений, меньших x (x — произвольное число).

По условию задачи $n = 6$. Построим вариационный ряд, для чего расположим значения случайной величины x_i в возрастающем порядке (–5,6; –3,2; –0,8; 1,1; 2,9; 4,4). Каждое значение случайная величина ξ приняла по одному разу. Тогда частота появления любого из значений x_i будет $n_i = 1$, а относительная частота $W_i = \frac{n_i}{n} = \frac{1}{6}$.

Составим таблицу (эмпирический ряд).

x_i	–5,6	–3,2	–0,8	1,1	2,9	4,4
W_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Тогда график эмпирической функции $F_n^*(x)$ будет

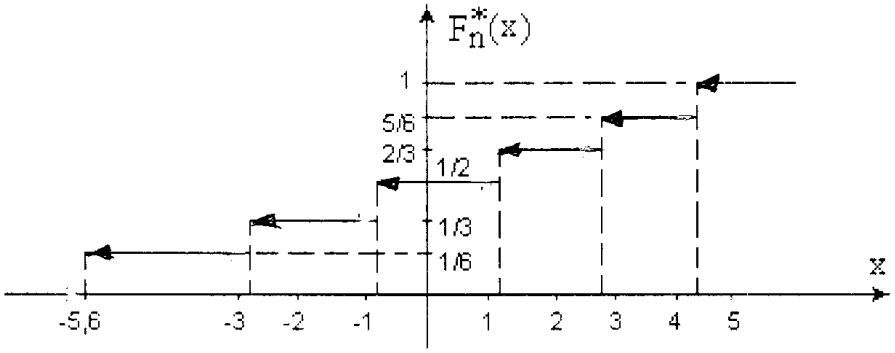


Рис. 5

Пример 2. Через каждый час измерялось напряжение тока в электросети. При этом были получены следующие значения:

227 219 215 230 231 223 220 222 218 219 222 221 227 226 226

209 211 215 218 220 217 220 221 225 224 212 217 219 220 226

Построить полигон относительных частот. Объем выборки равен 30.

Решение: Расположим значения напряжения в порядке их возрастания, т. е. построим вариационный ряд. Под каждым из этих значений запишем соответствующую частоту появления $W_i = \frac{n_i}{n}$.

x_i	209	211	212	215	217	218	219	220	221
W_i	$\frac{1}{30}$	$\frac{1}{30}$	$\frac{1}{30}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{15}$

x_i	222	223	224	225	226	227	230	231
W_i	$\frac{1}{15}$	$\frac{1}{30}$	$\frac{1}{30}$	$\frac{1}{30}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{30}$	$\frac{1}{30}$

$\sum W_i = 1$. Полигон этого распределения изображен на рисунке:

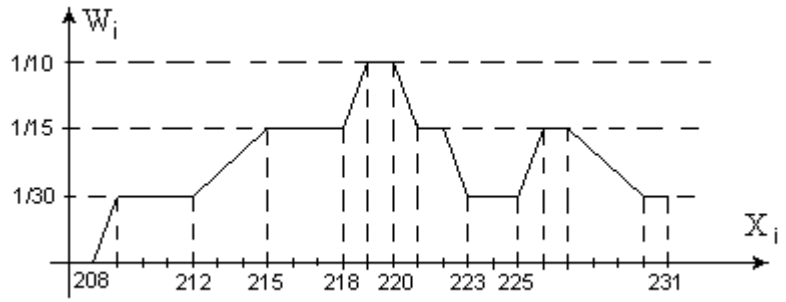


Рис. 6

Точечные оценки параметров распределения

Первая и основная задача математической статистики — получение по данным выборки наиболее рационально построенных статистических характеристик распределения. На практике во многих задачах форма теоретического закона распределения может считаться определенной. Например, при обработке деталей на металлорежущих станках по методу автоматического получения размера (при устойчивом технологическом процессе) можно считать, что распределение погрешностей деталей подчиняется нормальному закону распределения.

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Если бы точные значения параметров m и σ при нормальном законе были бы известны, то закон распределения для данного случая был бы полностью определен. Таким образом, задача сводится к определению значений параметров.

Пусть $X_1, X_2, ..., X_n$ — выборка для данной случайной величины ξ; $F_\xi(x)$ и $F_n^*(x)$ — соответствующие теоретическая и эмпирическая функции распределения. Точно так же, как функции $F_\xi(x)$ ставят в соответствие $F_n^*(x)$, так и любой теоретической характеристике (параметру) можно поставить в соответствие ее статистический аналог.

Всякое приближенное значение параметра закона распределения, определенное каким-либо способом по элементам выборки, называется ***оценкой*** параметра закона распределения.

Следует обратить внимание на то, что ни при каком n, вообще говоря, нельзя определить по выборке точные значения неизвестного параметра, а можно лишь найти его приближенное значение, и оценка параметра является случайной величиной.

Если α — параметр, α^* — его оценка, то $\varepsilon=\alpha-\alpha^*$ — полная ошибка, допущенная при вычислении оценки.

$$\varepsilon=\alpha-\alpha^*=\alpha-M(\alpha^*)+M(\alpha^*)-\alpha^*=\varepsilon_1+\varepsilon_2,$$

где $\varepsilon_1=M(\alpha^*)-\alpha^*$, $\varepsilon_2=\alpha-M(\alpha^*)$, ε_1 — случайная ошибка, ε_2 — систематическая ошибка, неслучайная величина.

Определение: ***Точечными*** оценками параметров распределения называются такие оценки, которые характеризуются одним числом.

В общем виде задача формулируется так: используя статистическую информацию, доставляемую выборкой $X_1, X_2, ..., X_n$ сделать статистические выводы об истинном значении α неизвестного параметра.

Рассмотрим точечную оценку теоретических начальных моментов k-го порядка $m_k = \sum_{i=1}^{n} \alpha_i^{(k)} p_i$. Выборочный момент k-го порядка будем обозначать m_k^*:

$$m_k^* = m_k^*(\xi) = \frac{1}{n}\sum_{i=1}^{n} X_i^{(k)} . \qquad (1)$$

При $k=1$ величина m_1^* называется ***выборочным средним*** (или средним арифметическим) и обозначается символом $\overline{X}$.

$$\overline{X} = m_1^* = \frac{1}{n}\sum_{i=1}^{n} X_i . \qquad (2)$$

Аналогично, ***выборочным центральным моментом*** k-го порядка называют случайную величину

$$\mu_k^* = \mu_k^*(\xi) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k . \qquad (3)$$

При $k=2$ величину μ_k^* называют ***выборочной дисперсией*** и обозначают

$$S^2 = S^2(\xi) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 . \qquad (4)$$

Замечание: Если среди X_i встречаются повторяющиеся, например, варианта X_i встречается n_i раз, то средняя выборочная и выборочная дисперсия определяются по формулам

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} X_i n_i , \qquad (2^*)$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{k}\left(X_i - \overline{X}\right)^2 n_i . \qquad (4^*)$$

Пример 3. Дан эмпирический ряд частот:

X_i	1	2	3
n_i	2	1	4

Найти $\overline{X}$ и S^2.

Решение:

$$\overline{X} = \frac{1}{7}\left[1\cdot 2 + 2\cdot 1 + 3\cdot 4\right] = \frac{16}{7};$$

$$S^2 = \frac{1}{7}\left[\left(1-\frac{16}{7}\right)^2\cdot 2 + \left(2-\frac{16}{7}\right)^2\cdot 1 + \left(\frac{3}{4}-\frac{16}{7}\right)^2\cdot 4\right] \approx 1{,}06 .$$

Между выборочными начальными и центральными моментами сохраняются те же соотношения, что и между теоретическими начальными m_k и центральными μ_k моментами, в частности:

$$S^2 = m_2^* - \left(\overline{X}\right)^2 ; \mu_3^* = m_3^* - 3\overline{X}m_2^* + 2\left(\overline{X}\right)^3 ,$$

$$\mu_4^* = m_4^* - 4\overline{X}m_3^* + 6\left(\overline{X}\right)^2 m_2^* - 3\left(\overline{X}\right)^4 ,$$

$$S_k^* = \frac{\mu_3^*}{\left(\mu_2^*\right)^{3/2}}; E_x^* = \frac{\mu_4^*}{\left(\mu_2^*\right)^2} - 3.$$

Таким образом, за оценку математического ожидания $M\xi$ и дисперсии $D\xi$ можно принять соответствующие выборочные характеристики, т. е.:

$$M\xi \approx \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i ,$$

$$D\xi \approx S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2 .$$

В статистических задачах используются самые различные выборочные характеристики. Например, выборочная ***медиана*** η^* — это среднее значение вариационного ряда, т. е. значение $\eta^* = X_m$, если $n = 2m-1$ и $\eta^* = \frac{\left(X_m + X_{m+1}\right)}{2}$, если $n = 2m$.

Напомним, что медианой Me непрерывного распределения F называется решение уравнения $F(x) = \frac{1}{2}$.

Приведенный способ оценок параметров не всегда приводит к наилучшим оценкам. В связи с этим возник вопрос о том, какую оценку считать наилучшей, т. е. какие требования предъявлять к оценкам.

Пример 4. Пусть необходимо составить ряд распределения 50 хозяйств по среднегодовой численности работников на 100 га сельскохозяйственных угодий. Рассчитаны значения данного признака по каждому предприятию:

2,57	2,33	6,28	7,16	12,68	4,34	3,20	5,46	4,38	6,65
2,89	4,14	4,03	3,31	5,61	3,48	3,94	5,40	4,77	4,81
3,29	4,42	5,29	1,25	5,56	2,91	3,66	2,72	5,88	6,87
3,87	3,79	4,50	2,24	5,34	4,28	3,05	3,52	6,59	6,33
6,13	3,62	4,31	4,85	5,71	7,65	4,29	4,10	4,74	4,74

Так как значения признака заполняют промежуток значений, то стоится интервальный ряд распределения с равными интервалами. Число групп, на которые разбивается вариационный ряд, определяется по следующей формуле:

$$k = 1 + 3{,}322 \lg n; \quad k = 1 + 3{,}322 \lg 50 = 6{,}6 .$$

Учитывая небольшой объём совокупности предприятий, примем число групп равным 6, значит $k = 6$.

Величина интервала находится по формуле

$$h = \frac{x_{\max} - x_{\min}}{k},$$

где $x_{\max}$ и $x_{\min}$, соответственно, наименьшее и наибольшее значения признака.

Величина интервала округляется обычно в сторону увеличения до принятой точности измерения признака. Если крайние значения значительно отличаются от рядом расположенных значений, то в приведенной формуле они не учитываются, тогда строится ряд распределения с открытыми крайними интервалами. Например, значение 1,25 существенно отличается от следующего за ним 2,24, а также 12,68 существенно отличается от предыдущего значения 7,65, тогда:

$$h = \frac{7{,}65 - 2{,}24}{6} = 0{,}902.$$

Округлив величину интервала, получим $h = 0{,}9$. Границы интервалов составят: $2{,}2 + 0{,}9 = 3{,}1$; $3{,}1 + 0{,}9 = 4{,}0$; $4{,}0 + 0{,}9 = 4{,}9$; $4{,}9 + 0{,}9 = 5{,}8$; $5{,}8 + 0{,}9 = 6{,}7$; $6{,}7 + 0{,}9 = 7{,}6$.

Так как наименьшее значение (1,25) и наибольшее (12,68) в формуле были отброшены, а величина h округлена в сторону уменьшения, то чтобы всё учесть все значения вариационного ряда, крайние интервалы берутся открытыми. Подсчитав число хозяйств, попавших в соответствующий интервал, составляем вариационный ряд распределения (таблица 1).

Таблица 1

Распределение сельскохозяйственных предприятий
по численности работников на 100 га сельскохозяйственных угодий

Группы предприятий по численности работников на 100 га сельскохозяйственных угодий, чел.	Число предприятий в группе (n_i)	В % к итогу (w_i)	Накопленное число предприятий (S_i)	В % к итогу
до 3,1	7	14,0	7	14,0
3,1–4,0	11	22,0	18	36,0
4,0–4,9	15	30,0	33	66,0
4,9–5,8	7	14,0	40	80,0
5,8–6,7	6	12,0	46	92,2
Свыше 6,7	4	8,0	50	100,0
ИТОГО	50	100,0		

Графически ряд распределения изображается в виде полигона, гистограммы или кумуляты распределения. На оси абсцисс откладываются значения изучаемого признака (границы интервалов), а на оси ординат число хозяйств или накопленное число хозяйств.

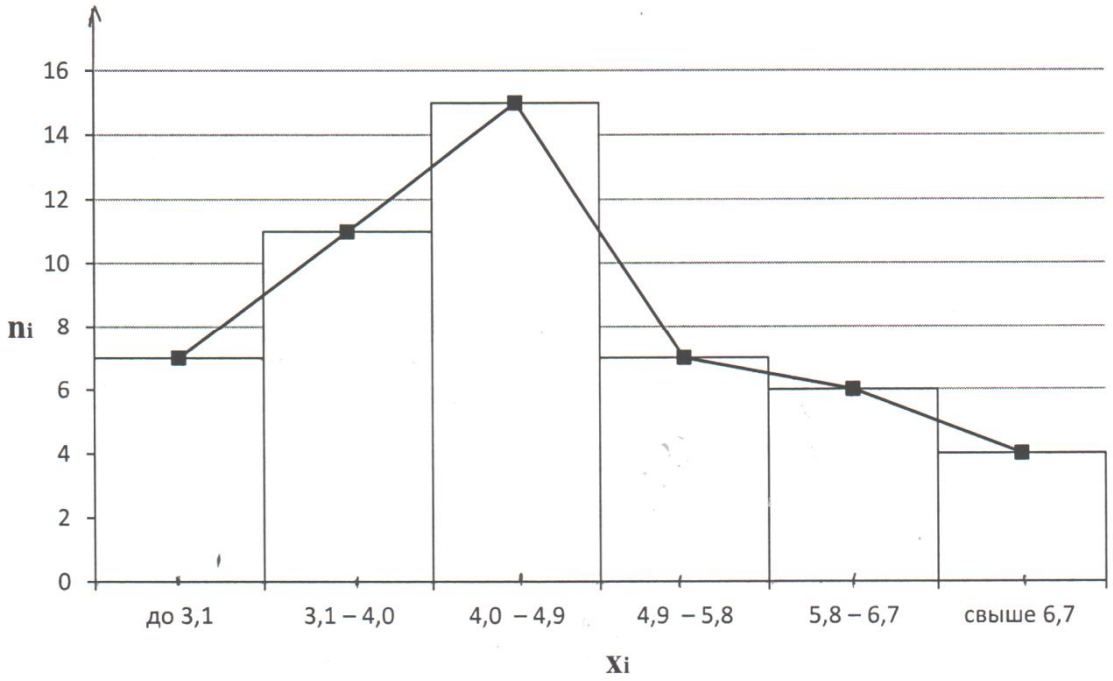


Рис. 7. Полигон и гистограмма распределения предприятий по численности работников на 100 га сельскохозяйственных угодий, чел.

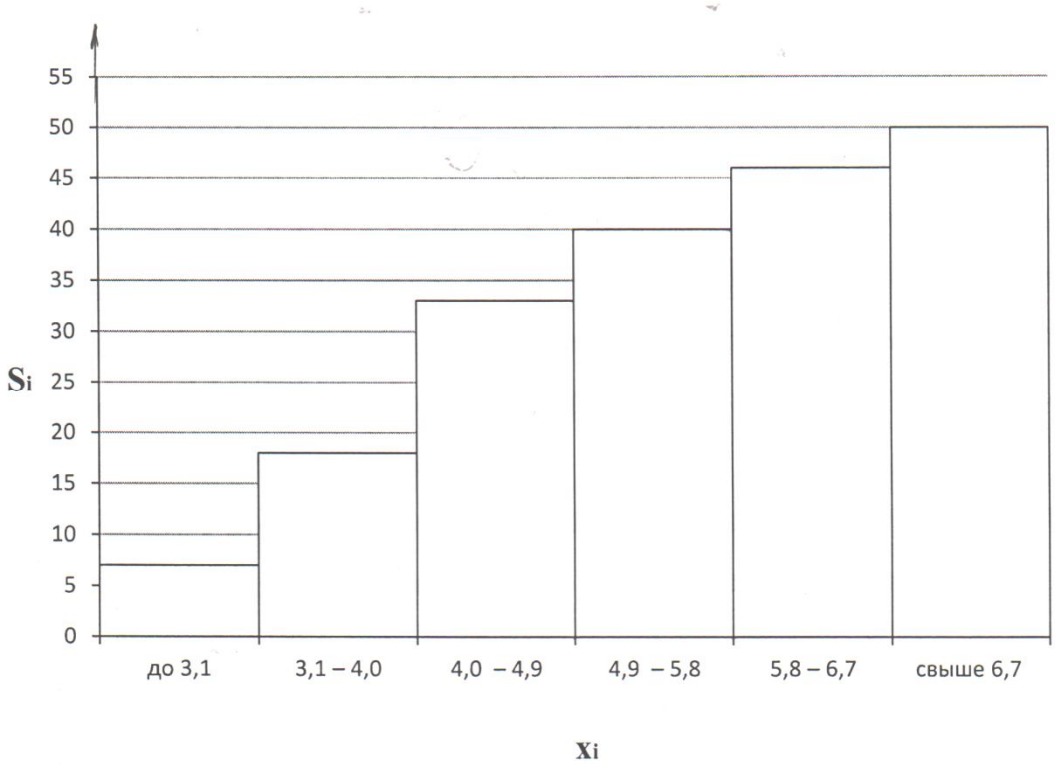


Рис. 8. Кумулята распределения предприятий по численности работающих на 100 га сельхозугодий

Рассчитаем основные характеристики вариационного ряда, к которым относятся мода, медиана, среднее значение, дисперсия и среднее квадратическое отклонение.

Модой называется значение признака, имеющее наибольшую частоту в ряду распределения. Вариационные ряды могут иметь одну или несколько модальных значений. Так как в примере распределение одномодальное, то мода находится в интервале с самой большой частотой (4,0–4,9).

В рядах с равными интервалами мода внутри модального интервала определятся по формуле.

$$M_0 = X_{M_0} + h\frac{n_{M_0} - n_{M_{0-1}}}{\left(n_{M_0} - n_{M_{0-1}}\right) + \left(n_{M_0} - n_{M_{0+1}}\right)},$$

где X_{M_0} — нижняя граница модального интервала;

$n_{M_0}, n_{M_{0-1}}, n_{M_{0+1}}$ — частоты модального, предмодального и послемодального интервалов, соответственно.

$$M_0 = 4 + 0{,}9 \cdot \frac{15-11}{(15-11)+(15-7)} = 4{,}3\,.$$

Значит, наиболее часто встречаются предприятия с численностью работников на 100 га сельскохозяйственных угодий 4,3 человека.

Медианой называется значение признака, находящегося в середине ряда распределения. В интервальном ряду она находится по формуле:

$$M_e = X_{M_e} + h\frac{0{,}5n - s_{M_{e-1}}}{s_{M_e}},$$

где X_{M_e} — нижняя граница медианного интервала;

$n_{M_{e-1}}$ — накопленная частота интервала, предшествующего медианному;

n_{M_e} — частота медианного интервала.

В примере $0{,}5n = 0{,}5 \cdot 50 = 25$. По накопленным частотам видно, что медиана находится в интервале $(4{,}0 - 4{,}9)$, поэтому $X_{M_e} = 4{,}0$, тогда

$$X_{M_e} = 4{,}0 + 0{,}9\frac{25-18}{15} = 4{,}42$$

Значит, половина сельскохозяйственных предприятий имеет численность работников на 100 га сельхозугодий до 4,42 чел., а половина хозяйств более 4,42 чел.

Для расчета средней величины признака, дисперсии, среднего квадратического отклонения составляется вспомогательная таблица 2. Так как в примере представлен вариационный ряд с открытыми крайними интервалами, то до расчета обобщающих характеристик их необходимо закрыть. Первый интервал до 3,1: 3,1 – 0,9 = 2,2, его границы 2,2–3,1. Последний интервал 6,7; 6,7 + 0,9 = 7,6, его границы 6,7–7,6.

Таблица 2

Вспомогательная таблица для расчета
средней дисперсии ряда распределения

Группы предприятий по численности работников на 100 га сельхозугодий, чел.	Число хозяйств в группе (n_i)	Среднее значение интервала (x_i)	$x_i n_i$	$\lvert x_i - \bar{x}\rvert n_i$	$(x_i - \bar{x})^2 n_i$
2,2–3,1	7	2,65	18,55	13,37	25,5367
3,1–4,0	11	3,55	39,05	11,11	11,2211
4,0–4,9	15	4,45	66,75	1,65	0,1815
4,9–5,8	7	5,35	37,45	5,53	4,3687
5,8–6,7	6	6,25	37,5	10,14	17,1366
6,7–7,6	4	7,15	28,6	10,36	26,8324
ИТОГО	50		227,9	52,16	85,276

Для определения среднего значения признака вначале находят среднее значение каждого интервала, как полусумму границ интервала.

Среднее значение признака составит:

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{227,9}{50} = 4,56 .$$

Найдём показатели вариации.

Размах вариации $R = X_{\max} - X_{\min} = 12,68 - 1,25 = 11,43$.

Среднее линейное отклонение $L = \dfrac{\sum \lvert x_i - \bar{x}\rvert}{n} = \dfrac{52,16}{50} = 1,04$.

Дисперсия и среднее квадратическое отклонение

$$\sigma^2 - \frac{\sum n_i (x_i - \bar{x})^2}{n} = \frac{85,276}{50} = 1,70552 ; \ \sigma = \sqrt{1,70552} = 1,306 \approx 1,31 .$$

Коэффициент вариации

$$V = \frac{\sigma}{x} \cdot 100 = \frac{1,31}{4,56} \cdot 100 = 28,7\% .$$

Таким образом, средняя численность работников на 100 га сельскохозяйственных угодий по совокупности предприятий составила 4,56 чел. Плотность работников в среднем колебалась в промежутке $\bar{x} \pm \sigma = 4,56 \pm 1,31$, т. е. от 3,25 до 5,87 чел. на 100 га сельхозугодий. Этот интервал, а также коэффициент вариации показывают, что имеются большие различия в обеспеченности предприятий рабочей силой.

Построение доверительного интервала для математического ожидания при известной дисперсии нормально распределенной генеральной совокупности

Для математического ожидания m при известной дисперсии σ^2 нормально распределенной генеральной совокупности доверительный интервал строится следующим образом.

Пусть выборка объема n из генеральной совокупности состоит из независимых нормально распределенных с параметрами m и σ случайных величин, причем среднее квадратичное σ известно, а величину m оцениваем по выборке $m \approx \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\overline{X}$ — состоятельная, несмещенная оценка математического ожидания.

Поставим задачу оценить точность этого приближенного равенства, т. е. указать границы (доверительные пределы), в которых практически достоверно лежит неизвестное значение m. Так как X_i распределены нормально с $MX_i = m$ и $DX_i = \sigma^2$, то случайная величина $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ распределена так же нормально как линейная функция с параметрами $M\overline{X} = m$ и $D\overline{X} = \frac{\sigma^2}{n}$. Поэтому вероятность γ того, что $\left(\overline{X} - m\right)$ (отклонение оценки от оцениваемого параметра) не превзойдет по абсолютной величине некоторого наперед заданного числа $\varepsilon > 0$, равна

$$P\left(\left|\overline{X} - m\right| \le \varepsilon\right) = P\left(\overline{X} - \varepsilon \le m \le \overline{X} + \varepsilon\right) = \gamma .$$

Согласно известной в теории вероятностей формуле

$$P\left(a \le x < b\right) = F\left(b\right) - F\left(a\right),$$

где $F(x)$ — функция распределения.

Тогда

$$P\left(\overline{X} - \varepsilon < m < \overline{X} + \varepsilon\right) = \Phi\left(\frac{\overline{X} + \varepsilon - M\overline{X}}{\sigma_{\overline{X}}}\right) - \Phi\left(\frac{\overline{X} - \varepsilon - M\overline{X}}{\sigma_{\overline{X}}}\right),$$

т. к. для нормального распределения

$$F\left(x\right) = \frac{1}{2} + \Phi\left(\frac{X - MX}{\sigma}\right). \qquad (5)$$

Таким образом,

$$\Phi\left(\frac{\overline{X} + \varepsilon - M\overline{X}}{\sigma_{\overline{X}}}\right) - \Phi\left(\frac{\overline{X} - \varepsilon - M\overline{X}}{\sigma_{\overline{X}}}\right) = \gamma ,$$

где

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-t^2/2}\,dt \ \text{— функция Лапласа.} \qquad (6)$$

Тогда

$$\Phi\left(\frac{\overline{X}+\varepsilon-\overline{X}}{\sigma/\sqrt{n}}\right)-\Phi\left(\frac{\overline{X}-\varepsilon-\overline{X}}{\sigma/\sqrt{n}}\right)=\gamma\cdot$$

$$\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)+\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)=\gamma,\ 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)=\gamma. \qquad (7)$$

Зададим надежность γ такой, чтобы событие с вероятностью γ можно считать практически достоверным, и пусть t_γ — корень уравнения $2\Phi(t)=\gamma$, который можно найти по таблицам функции Лапласа (см. приложение).

Определим из условия $\frac{\varepsilon\sqrt{n}}{\sigma}=t_\gamma$ число ε, характеризующее точность оценки:

$$\varepsilon=\frac{t_\gamma\sigma}{\sqrt{n}}, \qquad (*)$$

которое иногда называют ***предельной ошибкой***.

Для данного ε имеем

$$P\left(\left|\overline{X}-m\right|\le\frac{t_\gamma\sigma}{\sqrt{n}}\right)=2\Phi(t)=\gamma. \qquad (8)$$

Таким образом, практически достоверно (точнее с вероятностью γ), выполняется

$$\left|\overline{X}-m\right|\le\frac{t_\gamma\sigma}{\sqrt{n}},\ \text{где}\ 2\Phi(t)=\gamma.$$

Следовательно, интервал со случайными концами $\overline{X}-t_\gamma\cdot\frac{\sigma}{\sqrt{n}}$, $\overline{X}+t_\gamma\cdot\frac{\sigma}{\sqrt{n}}$ с вероятностью γ покрывает неизвестное значение m (при выборке объема n в $(100\cdot\gamma)\%$ случаев случайный интервал будет содержать неизвестное значение параметра m). Оценка (доверительный интервал)

$$\overline{X}-t_\gamma\cdot\frac{\sigma}{\sqrt{n}}\le m\le\overline{X}+t_\gamma\cdot\frac{\sigma}{\sqrt{n}} \qquad (9)$$

предполагает известным среднее квадратичное отклонение σ, которое на практике чаще всего неизвестно. Если величину σ в (9) заменить приближенным значением (несмещенной оценкой дисперсии $\tilde{S}^2$):

$$\sigma \approx \tilde{S},\ \sigma \approx \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}\ , \qquad (10)$$

то надежность оценки (9) уменьшится (особенно, если объем выборки не велик). Поэтому, если σ неизвестно, пользуются другим способом построения доверительного интервала для математического ожидания.

Из (9) видно, что при увеличении объема выборки длина доверительного интервала I_γ уменьшается, а точность оценки увеличивается, т. к. в этом случае согласно (*) величина ε уменьшается.

Пример 5. Определить доверительный интервал для математического ожидания m при заданных величинах: $\bar{X} = 3$; $n = 16$; $\sigma = 2$; $\gamma = 0,9;\ \ 0,95;\ \ 0,99$.

Решение: Вычисления произведем по формуле (8)

$$P\left(\left|\overline{X} - m\right| \le \frac{t_\gamma \sigma}{\sqrt{n}}\right) = 2\Phi(t) = \gamma\ .$$

Результаты вычислений сведены в таблицу (t_γ находим, пользуясь таблицей П. 4 приложения):

γ	t_γ	$t_\gamma \frac{\sigma}{\sqrt{n}} = \varepsilon$	I_γ	l
0,9	1,64	0,82	(2,18; 3,82)	1,64
0,95	1,96	0,98	(2,02; 3,98)	1,96
0.99	2,58	1,29	(1,71; 4,29)	2,58

Как видно из полученной таблицы, чем больше доверительная вероятность γ, тем шире границы доверительного интервала, т. е. чем больше степень уверенности нам требуется, тем более широкие границы придется указывать, т. е. соглашаться на меньшую точность оценки.

Если требуется оценить математическое ожидание с наперед заданной точностью $\varepsilon = \frac{t_\gamma \sigma}{\sqrt{n}}$ и надежностью γ, то минимальный объем выборки, который обеспечит эту точность, находят по формуле

$$n = \frac{t_\gamma^2 \sigma^2}{\varepsilon^2}\ . \qquad (11)$$

При неизменном заданном объеме n увеличение доверительной вероятности γ влечет за собой увеличение ε, а это означает уменьшение точности оценки.

Построение доверительного интервала для математического ожидания при неизвестной дисперсии нормально распределенной генеральной совокупности

По данной выборке $X_1, X_2, ..., X_n$ случайной величины ξ, распределенной по нормальному закону с неизвестными математическим ожиданием m и дисперсией σ, требуется построить доверительный интервал для математического ожидания m.

Лемма Фишера: В выборке $X_1, X_2, ..., X_n$ из нормально распределенной генеральной совокупности выборочное среднее $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ и выборочная дисперсия $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ взаимно независимы. Величина $\frac{nS^2}{\sigma^2}$ имеет распределение χ^2 с $(n-1)$ степенями свободы.

Рассмотрим две величины $Z = \sqrt{n}\,\frac{\overline{X} - m}{\sigma}$ и $\upsilon = \frac{nS^2}{\sigma^2}$, которые согласно лемме независимы, причем величина Z распределена нормально с $MZ = 0$ и $DZ = 1$, а величина υ — по закону χ^2 с $(n-1)$ степенями свободы.

Тогда величина

$$\zeta = \frac{z}{\sqrt{\upsilon}}\sqrt{n-1} = \frac{\overline{X} - m}{S}\sqrt{n-1}. \qquad (12)$$

имеет распределение Стьюдента с $(n-1)$ степенями свободы. Зададим надежность γ и предположим, что t_γ — корень уравнения

$$\int_{-t}^{t} f_{n-1,t}(x)\,dx = \gamma,$$

где $f_{n-1,t}(x)$ — плотность распределения вероятностей закона Стьюдента с $(n-1)$ степенями свободы. Для значения t_γ, которое находится из таблицы (имеющей два входа — по числу степеней свободы и надежности γ), имеем

$$P\left(\left|\overline{X} - m\right| \le \varepsilon\right) = \gamma \;\Rightarrow\; P\left(\left|\overline{X} - m\right| \cdot \frac{\sqrt{n-1}}{S} \le \varepsilon \cdot \frac{\sqrt{n-1}}{S}\right) = \gamma.$$

Обозначим

$$t = \frac{\sqrt{n-1}}{S} \cdot \varepsilon. \qquad (13)$$

Тогда

$$P\left(|\zeta| \le t_\gamma\right) = \int_{-t_\gamma}^{t_\gamma} f_{n-1,t}(x)\,dx = \gamma,$$

или т. к. $f_{n-1,t}(x)$ — четная функция, то

$$\gamma = P\left(|\zeta| < t_\gamma\right) = 2\int_0^{t_\gamma} f_{n-1,t}\left(x\right)dx = 2F_{n-1,t}\left(t_\gamma\right), \text{ т. е. } 2F_{n-1,t}\left(t_\gamma\right) = \gamma .$$

Таким образом, с надежностью γ (доверительной вероятностью) выполняется неравенство

$$|\zeta| \le t_\gamma \text{ или } \left|\frac{\overline{X} - m}{S}\cdot\sqrt{n-1}\right| \le t_\gamma .$$

Преобразуем последнее неравенство, заменив его равносильным ему двойным неравенством

$$\overline{X} - t_\gamma \cdot \frac{S}{\sqrt{n-1}} \le m \le \overline{X} + t_\gamma \cdot \frac{S}{\sqrt{n-1}}. \qquad (14)$$

Итак, случайный интервал с концами в точках $\overline{X} - t_\gamma \cdot \frac{S}{\sqrt{n-1}}$ и $\overline{X} + t_\gamma \cdot \frac{S}{\sqrt{n-1}}$, с вероятностью γ содержит внутри себя неизвестное значение m. Таким образом, пользуясь распределением Стьюдента, построен доверительный интервал для величины m, соответствующий надежности γ. По таблицам распределения Стьюдента по заданным $(n-1)$ и γ можно найти t_γ (см. приложение, табл. П. 5).

Пример 6. По выборке объема $n = 10$ построить доверительный интервал для m при доверительной вероятности $\gamma = 0,95$, если $\overline{X} = 3$, $S = 2,7$.

Решение: Число степеней свободы $k = n - 1 = 10 - 1 = 9$. По k и $\gamma = 0,95$ в таблице П. 5 находим $t_\gamma = 2,26$. Далее по формуле (13) найдем доверительный интервал

$$I_\gamma = \left[3 - 2,26\frac{2,7}{\sqrt{10-1}};\quad 3 + 2,26\frac{2,7}{\sqrt{10-1}}\right].$$

$$I_\gamma = \left[0,97; 5,03\right].$$

Замечание: Как отмечалось выше при неограниченном возрастании объема выборки n распределение Стьюдента стремиться к нормальному. Поэтому практически при $n \ge 50$ можно вместо распределения Стьюдента пользоваться нормальным распределением. При этом интервалы (9) и (14) практически совпадают. Но при $n < 50$ интервал (9) может быть существенно уже интервала (14), т. к. при его построении использовалась дополнительная информация о значении σ.

Построение доверительного интервала для математического ожидания в случае ненормально распределенной генеральной совокупности

Каков бы ни был закон распределения независимых одинаково распределенных случайных величин $\xi_1, \xi_2, ..., \xi_n$, имеющих конечную дисперсию, их сумма $\sum_{i=1}^{n} \xi_i$ распределена приближенно нормально при достаточно больших n (согласно центральной предельной теореме). Значит и величина, оценивающая математическое ожидание $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, при большом n тоже имеет распределение близкое к нормальному. Таким образом, оценка математического ожидания при известном σ

$$\bar{X} - t_\gamma \frac{\sigma}{\sqrt{n}} \le m \le \bar{X} + t_\gamma \frac{\sigma}{\sqrt{n}}$$

имеет место с вероятностью, близкой к γ при достаточно больших n в случае, когда закон распределения генеральной совокупности не является нормальным, т. е.

$$P\left(\bar{X} - t_\gamma \frac{\sigma}{\sqrt{n}} \le m \le \bar{X} + t_\gamma \frac{\sigma}{\sqrt{n}} \right) = \gamma,$$

где t_γ — корень уравнения $2\Phi(t) = \gamma$. Если же σ неизвестно, то при больших n можно использовать состоятельную оценку величины σ по выборке с высокой степенью надежности

$$\sigma \approx \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2} = \tilde{S}$$

и заменить в равенстве неизвестную величину σ величиной $\tilde{S}$. Тогда

$$P\left(\bar{X} - t_\gamma \frac{\tilde{S}}{\sqrt{n}} \le m \le \bar{X} + t_\gamma \frac{\tilde{S}}{\sqrt{n}} \right) = \gamma,$$

т. е. интервал $\left(\bar{X} - t_\gamma \frac{\tilde{S}}{\sqrt{n}}, \ \bar{X} + t_\gamma \frac{\tilde{S}}{\sqrt{n}} \right)$ является доверительным интервалом для m с достоверной вероятностью γ.

Постановка задачи о проверке статистических гипотез

Часто функция распределения случайной величины бывает заранее неизвестна, и возникает необходимость ее определения по эмпирическим данным, однако решение задачи в такой общей постановке вызывает значительные трудности и в большинстве случаев не является необходимым.

Во многих случаях из некоторых дополнительных соображений могут быть сделаны предположения о виде функции распределения

$F_\xi(x)$. Простейшим и в то же время наиболее сильным предположением такого рода является предположение, что функция $F_\xi(x)$ есть вполне определенная функция $F_\xi(x)=F(x)$. В тех случаях, когда нет оснований сделать такое предположение, часто оказывается возможным предположить, что $F_\xi(x)$ принадлежит некоторому классу функций, зависящих от одного или нескольких параметров $\alpha_1,\alpha_2,...,\alpha_k$, $F_\xi(x)=F(x,\alpha_1,\alpha_2,...,\alpha_k)$. Параметры $\alpha_1,\alpha_2,...,\alpha_k$ неизвестны, их значения следует получить из опытных данных, т. е. оценить по выборке. Возможны и другие предположения о виде $F_\xi(x)$. То есть может быть выдвинута какая-либо гипотеза, которую необходимо проверить по выборке. Таким образом, необходимы правила, которые позволяли бы судить, согласуются ли наблюдаемые значения $X_1,X_2,...,X_n$ величины ξ с гипотезой относительно ее функции распределения.

Статистической гипотезой называют любое утверждение, предположение о виде или свойствах распределения наблюдаемых в эксперименте случайных величин.

Если гипотеза H однозначно фиксирует распределение наблюдений, то ее называют ***простой***. ***Сложной*** называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез.

Бывают ситуации, когда проверяемая гипотеза состоит в том, что некоторый параметр семейства распределений соответствующей совокупности (например, среднее значение, дисперсия и т. п.) имеет наперед заданное множество значений. Такие гипотезы называют ***параметрическими***.

Если для исследуемого явления сформулирована та или иная гипотеза (обычно ее называют ***основной*** или ***нулевой*** гипотезой и обозначают H_0), то задача состоит в том, чтобы сформулировать такое правило, которое позволяло бы по результатам соответствующих наблюдений принять или отклонить эту гипотезу. Гипотеза H, которая противоречит основной, называется ***альтернативной*** (конкурирующей). Правило, согласно которому проверяемая гипотеза H_0 принимается (не отвергается), называется ***статистическим критерием*** гипотезы H_0. Разработка таких правил (критериев) и их обоснование с точки зрения требования оптимальности и составляет предмет теории проверки статистических гипотез.

Критерий (правило) позволяет через проверку числового неравенства принять или отвергнуть гипотезу. Числовые неравенства получают на основе статистики и ее функции распределения (которая должна быть известна).

Например, если проверяют гипотезу H_0 о равенстве дисперсии двух нормальных генеральных совокупностей, то в качестве критерия

принимают отношение выборочных дисперсий $F = S_1^2 / S_2^2$. Эта величина случайная (потому что в различных опытах дисперсии принимают различные значения) и распределена по закону Фишера — Снедекора.

После выбора определенного критерия множество всех его возможных значений разбивают на два непересекающихся подмножества. Одно из них содержит значения критерия, при котором выбранная гипотеза отвергается, а другое — при которых она принимается.

Критической областью называют совокупность значений критерия, при которых гипотезу отвергают. По своему смыслу критическая область должна включать все маловероятные при гипотезе H_0 значения критерия. Обычно используют области вида $t \geq t_q$, или $t \leq t_q$, или $|t| \geq t_q$ (первые две — односторонние, третья — двусторонняя область).

Областью принятия гипотезы (областью допустимых значений или доверительной областью) называют совокупность значений критерия, при которых гипотезу принимают. ***Критическими точками*** t_q называются точки, отделяющие критическую область от области принятия гипотезы.

Принимая решение по результатам проверки какой-либо гипотезы, мы можем допустить ошибки различные по своему характеру. ***Ошибки 1 рода*** состоят в том, что отвергается гипотеза, которая на самом деле верна. ***Ошибки 2 рода*** состоят в том, что гипотеза принимается, когда на самом деле гипотеза неверна. Будем обозначать вероятности этих ошибок q и β соответственно.

Проверить гипотезу желательно таким образом, чтобы минимизировать вероятности ошибок обоих типов. Единственным способом одновременного уменьшения этих ошибок при фиксированном критерии является увеличение объема выборки (что нежелательно).

Величину q выбирают малой, такой, что событие, происходящее с вероятностью q, можно считать практически невозможным. Ее называют ***уровнем значимости критерия***. Обычно в качестве практически невозможных событий принимают такие, вероятность которых не превышает 0,05; 0,005; 0,001 и т. п.

Мощностью критерия называется вероятность попадания критерия в критическую область, при условии, что справедлива конкурирующая гипотеза. Другими словами, мощность критерия есть вероятность того, что выдвинутая гипотеза будет отвергнута, если верна конкурирующая гипотеза (не будет допущена ошибка 2 рода).

Пусть для проверки гипотезы принят определенный уровень значимости, q и выборка имеет фиксированный объем. Остается произвол в выборке критической области. Если вероятность ошибки 2 ро-

да равна β, т. е. вероятность события «принята выдвинутая гипотеза, а справедлива конкурирующая», то мощность критерия равна $1-\beta$.

Если мощность критерия $1-\beta$ возрастает, то уменьшается вероятность β совершить ошибку 2 рода. Таким образом, чем больше мощность, тем вероятность ошибки 2 рода меньше.

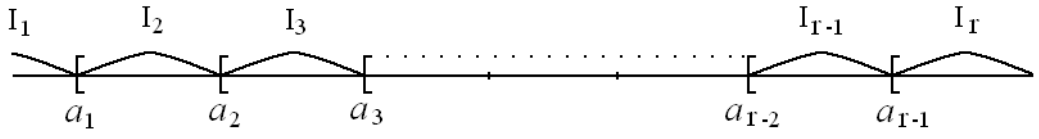
Итак, если уровень значимости q выбран, то критическую область следует строить так, чтобы мощность критерия была максимальной.

Выбор величины q до некоторой степени произволен. Если q увеличить, то критическая область будет расширяться и тем самым гипотеза будет чаще опровергаться, но эти опровержения могут стать ненадежными, т. е. могут относиться к верной гипотезе. Если уровень значимости q уменьшается, критическая область суживается, и проверяемая гипотеза все реже будет опровергаться, когда она верна, но с уменьшением уровня значимости понижается чувствительность критерия, т. к. расширяется область допустимых значений и увеличивается вероятность совершения ошибки 2 рода, т. е. вероятность принятия проверяемой гипотезы, когда она неверна.

Критерий согласия χ^2 ***К. Пирсона***

Наиболее употребительной мерой расхождения между $F_n^*(x)$ и $F(x)$ является критерий согласия, введенный английским биологом К. Пирсоном. Этот критерий можно использовать для любых распределений, в том числе и многомерных. Чтобы воспользоваться им, выборочные данные предварительно группируют.

Разобьем множество значений величины ξ на конечное число r интервалов $I_1, I_2, ..., I_r$ без общих точек (правый конец исключается из соответствующего интервала, а левый включается).



Пусть $p_i\ (i=1,2,...,r)$ вероятность того, что величина X принадлежит интервалу I_i; $p_i = F(x_i) - F(x_{i-1})$; $\left(\sum_{i=1}^{n} p_i = 1\right)$. Пусть $\nu_i\ (i=1,2,...)$ количество величин из числа наблюдаемых величин $X_1, X_2, ..., X_n$, принадлежащих интервалу I_i. Тогда $\frac{\nu_i}{n}$ — относительная частота попа-

дания величины ξ в интервал I_i при n наблюдениях. Очевидно, что $\sum_{i=1}^{r} \nu_i = n$, $\sum_{i=1}^{r} \frac{\nu_i}{n} = 1$ (таким образом строится интервальный (для непрерывной случайной величины) вариационный ряд). Для разбиения, приведенного на рисунке, p_i есть приращение теоретической функции распределения на интервале I_i, а $\frac{\nu_i}{n}$ — приращение эмпирической функции распределения $F_n^*(x)$ выборки на том же интервале. Если проверяемая гипотеза верна, то ν_i представляет частоту появления события, имеющего в каждом из произведенных испытаний вероятность p_i, следовательно мы можем рассматривать ν_i как случайную величину, подчиняющуюся биномиальному закону распределения с центром в точке np_i и средним квадратическим $\sigma_i = \sqrt{np_i q_i} = \sqrt{np_i (1-p_i)}$ (p_i — теоретическое значение вероятности попадания в i-ый интервал).

Когда n велико, можно считать согласно теореме Лапласа, что частота распределена асимптотически нормально с теми же параметрами. При правильности нашей гипотезы мы можем ожидать, что будут асимптотически нормально распределены также величины

$$\eta_i = \frac{\nu_i - np_i}{\sqrt{np_i}}, \ (i = 1, 2, ..., r).$$

В качестве меры расхождения μ данных выборки $\nu_1, \nu_2, ..., \nu_r$ с «теоретическими» данными $np_1, np_2, ..., np_r$ рассмотрим величину (хи-квадрат)

$$\chi^2 = \sum_{i=1}^{r} \eta_i^2 .$$

Другими словами, в качестве меры отклонения μ эмпирической функции распределения от теоретической, т. е. критерия проверки, принимается величина

$$\chi^2_{\text{ðàñ÷.}} = \chi^2_n = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} = \sum_{i=1}^{r} \left(\frac{\nu_i}{n} - p_i \right)^2 \frac{n}{p_i} \qquad (1)$$

— взвешенная сумма квадратов отклонения ν_i от их математических ожиданий np_i.

Величина χ^2_n — случайная, т. к. в различных опытах она принимает различные значения, и нас интересует ее распределение, вычисленное в предположении, что наша гипотеза H_0 верна, т. е. $F_\xi(x) = F(x)$. Ясно, что чем меньше различаются эмпирические величины ν_i и теоретические np_i, тем меньше величина критерия (1).

Если распределение χ_n^2 известно, то по заданному уровню значимости q можно найти предел значимости μ_0 для проверки принятой гипотезы.

Теорема Пирсона: Какова бы ни была функция распределения $F(x)$ случайной величины ξ, при $n \to \infty$ распределение величины χ_n^2 стремится к χ^2 распределению с $(r-m-t)$ степенями свободы, т. е.

$$p\left(\chi_n^2 < x\right) \xrightarrow[n\to\infty]{} \int_{-\infty}^{x} k(t)\,dt$$

в каждой точке x, где $k(x)$ — плотность распределения χ^2 с $(r-m-t)$ степенями свободы, где $t=1$ — одно условие, налагаемое на частоты $\sum \frac{\nu_i}{n} = 1$ для простой гипотезы, m — число оцениваемых параметров.

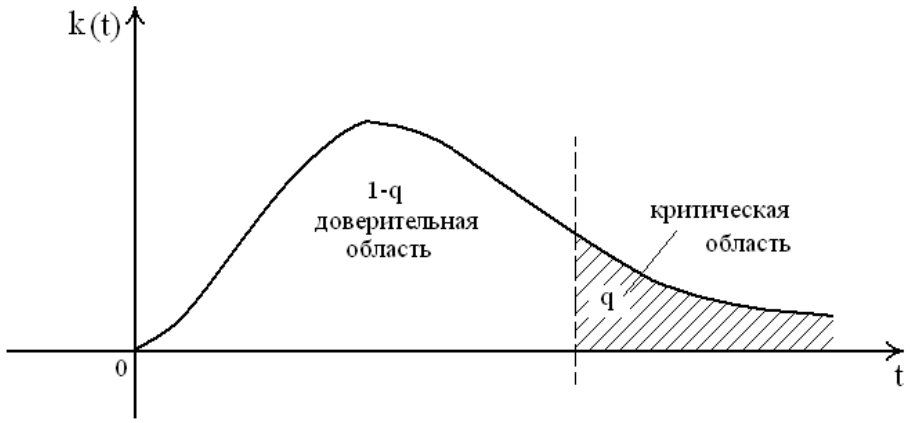


Рис. 9

Схема применения критерия χ^2 сводится к следующему:

1. Исходя из теоретического (предполагаемого) закона распределения, находим вероятности p_i попадания случайной величины в каждый из построенных r интервалов.
2. Вычисляем по формуле (3) χ_n^2 — значение критерия χ^2, соответствующее опытным данным.
3. Определяем число k степеней свободы распределения по формуле $k = r - m - t$ (если число оцениваемых параметров = m; если параметры известны, то $m = 0$).
4. Зная k и χ_n^2, с помощью таблицы 2 (см. приложение) определяем вероятность того, что величина, имеющая распределение χ^2 с k степенями свободы, превзойдет данное значение χ_n^2. Если эта вероятность мала, гипотеза отбрасывается как неправдоподобная, иначе гипотезу можно признать не противоречащей опытным данным (обычно считают вероятности, не превосходящие 0,01 — 0,05 достаточно малыми).

5. Если же задан некоторый уровень значимости q, то по таблице 2 по двум входам q и k находим $\chi^2_{q,k} = \chi^2_{крит}$. Сравниваем полученное по формуле (3) значение $\chi^2_n = \chi^2_{расч}$ с табличной величиной $\chi^2_{q,k}$. Если $\chi^2_{расч} > \chi^2_{q,k}$, то гипотеза считается опровергнутой опытными данными. Если $\chi^2_{расч} \le \chi^2_{q,k}$, то опытные данные можно считать совместными с принятой гипотезой (однако это еще недостаточно для установления истинности гипотезы, рекомендуется проверить эксперимент, используя другие критерии согласия).

Недостатком метода Пирсона является то, что группировка данных по интервалам I_i приводит к некоторой потери информации. Кроме того, остается еще вопрос о выборе числа интервалов и длине самих интервалов.

Следует иметь в виду, что критерий Пирсона применим только в том случае, когда число n наблюдений достаточно велико $(n \ge 60)$ и числа ν_i не менее 5–10. Если частота ν_i в отдельных разрядах (интервалах) мала, то их объединяют в 1 разряд.

Пример 7. При $n = 4040$ бросаниях монеты Бюффон получил $\nu_i = 2048$ выпадений «герба» и $\nu_i = 1992$ выпадений «решки». Проверим, используя критерий χ^2, совместимы ли эти данные с гипотезой H_0 о том, что монета была симметричной, т. е. что вероятность выпадения «герба» $p = \frac{1}{2}$. Здесь $r = 2;\ p_1 = \frac{1}{2};\ p_2 = 1 - p_1 = \frac{1}{2}$. Согласно (1)

$$\chi^2_n = \sum_{i=1}^{2} \left(\frac{\nu_i}{n} - p_i \right)^2 \cdot \frac{n}{p_i} = \left(\frac{\nu_1}{n} - p_1 \right)^2 \cdot \frac{n}{p_1} + \left(\frac{\nu_2}{n} - p_2 \right)^2 \cdot \frac{n}{p_2} = 0,776 .$$

Пусть уровень значимости $q = p\left\{\chi^2_n > \chi^2_{q,k}\right\}$ задан равным 0,05. По таблице П. 6 распределения χ^2 находим по двум входам k и q ($k = r - 1 = 1$ числу степеней свободы и уровню значимости q=0,05)

$$\chi^2_{q,k} = \chi^2_{0,05;1} = 3,841 .$$

Сравниваем полученное значение χ^2_n с табличной величиной $\chi^2_{0,05;1}$. Так как $\chi^2_n \le \chi^2_{q,k}$, то делаем вывод, что данные не противоречат гипотезе.

Пример 8. Наблюдались показания 500 наугад выбранных часов выставленных в витринах часовщиков. Интервалы группировки $[0,1),[1,2),...,[11,12)$, так что $p_i = \frac{1}{12}(i = 1,2,...,12)$.

Взяты две выборки объемом $n_1 = n_2 = 500$, ν_i — число часов, показания которых принадлежат i-тому интервалу. Результаты наблюдений оказались следующие:

i	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)
1-ая выборка ν_i	41	34	54	39	49	45	41
2-ая выборка ν_i	36	47	41	47	49	45	32

i	[7,8)	[8,9)	[9,10)	[10,11)	[11,12)
1-ая выборка ν_i	33	37	41	47	39
2-ая выборка ν_i	37	40	41	37	48

$r = 12$. Проверяемая гипотеза H_0: показания часов равномерно распределены на интервале $(0,12)$. Согласно гипотезе $p_1 = p_2 = ... = p_{12} = \frac{1}{12}$. Пусть уровень значимости q=0,05. Для первой выборки имеем

$$\chi^2_{n_1} = \sum_{i=1}^{12}\left(\frac{\nu_i}{n} - p_i\right)^2 \cdot \frac{n}{p_i} = 10,000;$$

для второй $\chi^2_{n_2} = 8,032$.

По таблице распределения χ^2 находим по 12 – 1 = 11 степеням свободы и q=0,05 $\chi^2_{0,05;1} = 19,675$. Так как $\chi^2_n < \chi^2$, то следует признать, что согласие хорошее и в первом и во втором случаях. Можно рассмотреть величину $\chi^2_{n_1} + \chi^2_{n_2} = 18,032$, имеющую 22 степени свободы для выборки объема $n = 1000$. Так как $\chi^2_{0,05;22} = 33,9$, то согласие получается снова хорошим.

Пример 9. Дан интервальный статистический ряд частот непрерывной случайной величины x

Частичные интервалы	$[40{,}24;40{,}28)$	$[40{,}28;40{,}32)$	$[40{,}32;40{,}36)$	$[40{,}36;40{,}40)$	$[40{,}40;40{,}44)$
Середины интервалов, x_i	40,26	40,30	40,34	40,38	40,42
Частоты, m_i	4	14	30	19	13

$$n = \sum m_i = 80$$

Требуется: 1) проверить гипотезу о нормальном распределении генеральной совокупности с помощью критерия согласия χ^2 (Пиреона); 2) Найти интервальную оценку для математического ожидания.

1) Выборочное среднее вычислим по формуле $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i m_i$:

$$\bar{x} = \frac{40{,}26 \cdot 4 + 40{,}30 \cdot 14 + 40{,}34 \cdot 30 + 40{,}38 \cdot 19 + 40{,}42 \cdot 13}{80} = 40{,}352 .$$

Вычислим центральные эмпирические моменты 2 порядка:

$$\mu_2 = \frac{\sum m_i (x_i - \bar{x})^2}{n}$$

Результаты представлены в табл. 3.

Таблица 3

Частичные интервалы	**Середины интервалов** x_i	**Частоты**, m_i	$(\bar{x} - x)^2 m_i$
$[40{,}24;40{,}28)$	40,26	4	0,0339
$[40{,}28;40{,}32)$	40,30	14	0,0379
$[40{,}32;40{,}36)$	40,34	30	0,0043
$[40{,}36;40{,}40)$	40,38	19	0,0149
$[40{,}40;40{,}44)$	40,42	13	0,0601
$\sum$		80	0,1511
μ_2			0,0019

Получим выборочную дисперсию $s^2 = \mu_2 = 0{,}0019$.

2. Дифференциальная функция нормального закона распределения $N(a,\sigma)$ с параметрами a,σ имеет вид

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Точечными оценками параметров a,σ нормального распределения являются выборочное среднее и выборочное среднее квадратичное отклонение соответственно:

$$a=\bar{x}=40{,}352;\ \sigma=s=0{,}044.$$

Следовательно, дифференциальная функция предполагаемого нормального закона распределения имеет вид

$$f(x)=\frac{1}{0{,}044\sqrt{2\pi}}e^{-\frac{(x-40{,}352)^2}{0{,}0038}},$$

интегральная функция предполагаемого нормального распределения имеет вид

$$F(x)=\frac{1}{0{,}044\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(x-40{,}352)^2}{0{,}0038}}dt.$$

Используя нормированную функцию Лапласа $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$, интегральную функцию распределения нормального закона можно записать в виде

$$F(x)=\frac{1}{2}+\Phi\left(\frac{x-40{,}352}{0{,}044}\right).$$

3. Проведём детальную проверку гипотезы о распределении СВ X (диаметра отверстий) по нормальному закону с помощью критерия согласия χ^2. Для этого прономеруем частичные интервалы, выразив их в единицах среднего квадратического отклонения s:

$$u_i=\frac{x_i^*-\bar{x}}{s},$$

причем наименьшее значение u_i положим равным $-\infty$, наибольшее — $+\infty$ (см. столбец 3 табл. 4). Заметим, что так определённая СВ U является случайной величиной, распределённой по нормальному закону с параметрами $a=0$; $\sigma=1$.

Таблица 4

Частичные интервалы	**Частоты,** m_i	**Нормированные интервалы,** (u_i, u_{i+1})	**Теоретические вероятности,** p_i	**Теоретические частоты,** $n \cdot p_i$	$(m_i - np_i)^2$	$\frac{(m_i - n\,p_i)^2}{n\,p_i}$
$[40{,}24; 40{,}28)$	4	$(-\infty; -1{,}64)$	0,0505	4,0	0,00	0,00
$[40{,}28; 40{,}32)$	14	$(-1{,}64; -0{,}73)$	0,1822	14,6	0,36	0,02
$[40{,}32; 40{,}36)$	30	$(-0{,}73; 0{,}18)$	0,3387	27,1	8,41	0,31
$[40{,}36; 40{,}40)$	19	$(0{,}18; 1{,}09)$	0,2907	23,3	18,49	0,79
$[40{,}40; 40{,}44)$	13	$(1{,}09; \infty)$	0,1379	11,0	4,00	0,36
Суммы	80		1,000	80,0		$\chi^2_H = 1{,}48$

Далее вычисляем теоретические вероятности — вероятности попадания СВ X, распределённой по нормальному закону с параметрами $a = 40{,}352$; $\sigma = 0{,}044$, в частичные интервалы $\left(x_i^*, x_{i+1}^*\right)$ по формуле

$$p_i = P(x_i^* < X < x_{i+1}^*) = \Phi(u_{i+1}) - \Phi(u_i),$$

где

$$u_i = \frac{x_i^* - \bar{x}}{s},$$

$$\Phi(u_i) = \frac{1}{\sqrt{2\pi}} \int_0^{u_i} e^{-\frac{t^2}{2}}\, dt$$

(значения функции Лапласа приведены в таблице П3).

Например, вероятность того, что СВ X попадает в первый частичный нормированный интервал $(-\infty < x < 40{,}28)$ равна

$$p_1 = P(-\infty < x < 40{,}28) = \Phi\left(\frac{40{,}28 - 40{,}352}{0{,}044}\right) - \Phi\left(\frac{-\infty - 40{,}352}{0{,}044}\right) =$$
$$= \Phi(-1{,}64) - \Phi(-\infty) = -\Phi(1{,}64) + \Phi(\infty) = -0{,}4495 + 0{,}5 = 0{,}0505\,.$$

Аналогично

$$p_2 = P(40{,}28 < X < 40{,}32) = \Phi\left(\frac{40{,}32 - 40{,}352}{0{,}044}\right) - \Phi\left(\frac{40{,}28 - 40{,}352}{0{,}044}\right) =$$
$$= \Phi(-0{,}73) - \Phi(-1{,}64) = -\Phi(0{,}73) + \Phi(1{,}64) = -0{,}2673 + 0{,}4495 = 0{,}1822$$

и т. д. (см. столбец 4 таблицы 4).

После этого вычисляют теоретические частоты нормального закона распределения $n_i' = np_i$ (см. столбец 5 таблица 4) и наблюдаемое значение критерия χ^2:

$$\chi^2_H = \sum \frac{(m_i - n\,p_i)^2}{n\;p_i},$$

где m_i — эмпирические частоты, np_i — теоретические частоты.

В результате вычислений получили $\chi^2_H = 1{,}48$ (см. столбец 7 таблица 4).

Затем по таблицам квантилей распределения χ^2 (см. таблицу П. 6) находят критическое значение $\chi^2_{кр} = \chi^2_{\alpha;v}$, где $\alpha = 1 - \gamma$ — уровень значимости и $v = k - r - 1$ — число степеней свободы (здесь k — число интервалов, r — число параметров предполагаемого закона распределения СВ *X)*. В нашей задаче $\alpha = 1 - \gamma = 0{,}05$, $k = 5$, $r = 2$ и, следовательно, $v = 2$. Таким образом, критическое значение $\chi^2_{кр} = \chi^2_{0{,}05;2} = 5{,}99$.

Если $\chi^2_H \le \chi^2_{кр}$, то гипотеза о нормальном распределении СВ X (диаметра отверстий) принимается; в противном случае, т. е. если $\chi^2_H > \chi^2_{кр}$, гипотеза о распределении СВ X по нормальному закону отвергается. Поскольку в нашей задаче $\chi^2_H = 1{,}48 < \chi^2_{кр} = 5{,}99$, то нет оснований отвергнуть гипотезу о нормальном распределении СВ X — диаметра отверстий.

3. Найдем интервальную оценку математического ожидания нормального распределения.

Интервальная оценка для математического ожидания нормального распределения имеет вид

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

где $\bar{x}$ — выборочное среднее;

n — объём выборки;

$t_{\alpha;v}$ — квантиль распределения Стьюдента для уровня значимости $\alpha = 1 - \gamma$ и числа степеней свободы $v = n - 1$, отыскиваемое по таблице П. 5.

Так как в нашей задаче $\bar{x} = 40{,}352$; $s = 0{,}044$; $n = 80$; $\alpha = 0{,}05$; $v = 79$, то $t_{\alpha;v} = t_{0{,}05;79} = 1{,}99$.

Таким образом, имеем следующую интервальную оценку математического ожидания:

$$40{,}352 - 1{,}99 \cdot \frac{0{,}044}{\sqrt{80}} < a < 40{,}352 + 1{,}99 \cdot \frac{0{,}044}{\sqrt{80}}.$$

Парная линейная регрессия

В этой части работы исследуются такие связи между переменными, при которых значения одной переменной (признака-результата, зависимой переменной) в среднем изменяется в зависимости от того, какие значения принимает другая переменная (независимая, признак-фактор). Причем рассматривается простая (парная) линейная регрессия между 2 переменными: $\hat{y} = a_0 + b_0 x$ или $\hat{y} = a_0 + b_0 x + e$, e — отражает случайную составляющую вариации результативного признака, включает влияние не учтенных в модели факторов, случайных ошибок.

Теоретическая линия регрессии — это линия, вокруг которой группируются точки корреляционного поля и которая указывает основное направление, тенденцию связи. Построение линейной регрессии сводится к определению таких параметров a и b, которые наилучшим образом соответствуют эмпирическим данным. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК). Считаем, что сумма квадратов отклонений эмпирических (фактических) точек от точек теоретической линии регрессии (расчетных $\hat{y}_i$) минимальна. Эмпирическое уравнение регрессии $\hat{y} = a + bx$, a и b найдем по выборке объема n. Функция $S(a,b) = \sum\left(y - \hat{y}\right)^2$ минимизируется при использовании МНК.

Функция S 2-х переменных a и b может достигать экстремума в тех точках, в которых $\begin{cases} S'_a = 0, \\ S'_b = 0 \end{cases}$ или $\begin{cases} -2\sum y + 2na + 2b\sum x = 0, \\ -2\sum yx + 2a\sum x + 2b\sum x^2 = 0. \end{cases}$

Отсюда получаем систему нормальных уравнений:

$$\begin{cases} na + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum yx. \end{cases}$$

Решение системы: $a = \overline{y} - b\overline{x}$;

$$b = \frac{\overline{xy} - \overline{y} \cdot \overline{x}}{\overline{x}^2 - (\overline{x})^2}.$$

b — коэффициент регрессии, показывает на сколько в среднем изменилась величина признака-результата y при изменении факторного признака на 1. Для расчета оценок параметров a и b удобно построить корреляционную таблицу 5. (см. пример 10).

Уравнение регрессии всегда дополняется показателем r_{xy} — коэффициентом линейной корреляции:

$$r_{xy} = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\sigma_x \sigma_y} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}};$$

$$r = r_{xy} = b\frac{\sigma_x}{\sigma_y}\ ; \quad \left|r_{xy}\right| \le 1,\ b = r_{xy}\frac{\sigma_y}{\sigma_x}.$$

Величина коэффициента корреляции не является доказательством наличия причинно-следственной связи между исследуемыми признаками, а являются оценкой степени взаимной согласованности в изменениях признаков. Следует осторожно подходить к истолкованию полученного коэффициента корреляции при незначительных объемах выборочной совокупности. Возникает необходимость оценки существенности r_{xy}. Значимость r_{xy} проверяется на основе t — критерия Стьюдента с(n-2) степенями свободы:

$$t_{расч.} = \sqrt{\frac{r^2(n-2)}{1-r^2}} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n-2}\,.$$

Если $t_{н.} \rangle t_{кр.}$, то гипотеза H_0: $r_{xy}=0$ отвергается при заданном уровне значимости α, что свидетельствует о статистической значимости линейного коэффициента корреляции b ($b = \frac{r_{xy}\sigma_y}{\sigma_x}$). Если $t_{расч.} \langle t_{кр.}$, то эмпирический коэффициент корреляции r_{xy} приблизительно равен 0.

Пример 10. Изучается зависимость материалоемкости продукции от выпуска продукции по 10 однородным заводам: x — выпуск продукции, тыс. ед.; y — потребление материалов на единицу продукции, кг (x и y заданы в табл. 5).

Требуется:

I) определить спецификацию функции регрессии, для этого построить поле корреляции, сформулировать гипотезу о форме связи;

II) рассчитать параметры линейной парной регрессии $\hat{y} = a + bx + \varepsilon$ по методу наименьших квадратов.

Определить тесноту и направление связи между признаком-фактором и признаком-результатом, для этого вычислить линейный коэффициент парной корреляции:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y}.$$

Количественные оценки тесноты связи x и y

Величина модуля коэфф. корреляции	Характер связи
0	отсутствует
до 0,3	практически отсутствует
от 0,3 до 0,5	слабая
от 0,5 до 0,7	умеренная
от 0,7 до 1	сильная
1	функциональная, каждому значению факторного признака строго соответствует 1 значение результативного признака
при: $0 < r < 1$	прямая (с увеличением x увеличивается y)
$-1 < r < 0$	обратная (с увеличением x уменьшается y)

Решение.

I) Строим поле корреляции.

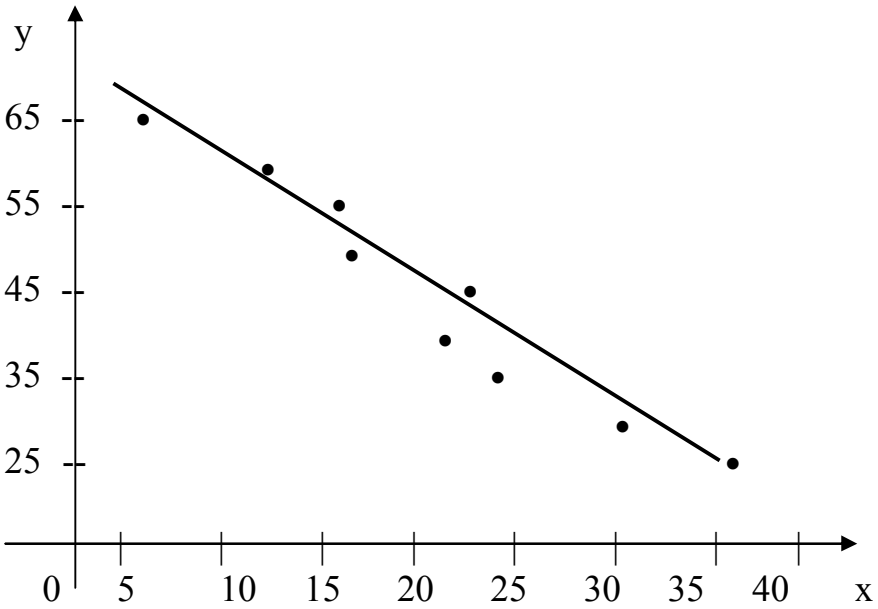


Рис. 10

В системе координат Oxy откладываем точки $M_i(x_i,y_i)$. Из рисунка видно, что точки располагаются вдоль прямой линии. Полагаем, что зависимость между x и y линейная.

II) Для расчета оценок параметров a и b линейной регрессии $y = a + bx$ решаем систему нормальных уравнений:

$$\begin{cases} na + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum xy. \end{cases} \qquad (**)$$

Для наглядности вычислений по МНК составляем вспомогательную (корреляционную) таблицу 5.

Таблица 5

	x	y	xy	x^2	y^2	$\hat{y}$	$y-\hat{y}$	$\left\|\frac{y-\hat{y}}{y}\right\|$	$(y-\hat{y})^2$
1	5	70	350	25	4900	71,8	–1,8	0,0257	3,24
2	11	65	715	121	4225	62,02	2,98	0,045	8,88
3	15	55	825	225	3025	55,5	-0,5	0,009	0,25
4	17	60	1020	289	3600	52,24	7,76	0,129	60,2
5	20	50	1000	400	2500	47,35	2,65	0,053	7,02
6	22	35	770	484	1225	44,09	–9,09	0,259	82,63
7	25	40	1000	625	1600	39,2	0,8	0,02	0,64
8	27	30	810	729	900	35,94	–5,94	0,198	35,28
9	30	25	750	900	625	31,05	–6,05	0,242	36,6
10	35	32	1120	1225	1024	22,9	9,1	0,284	82,81
Итого	207	462	8360	5023	23624	–	–	1,2647	317,55
Среднее значение	20,7	46,2	836	502,3	2362,4	–	–	–	31,76

Замечание. Объем выборки n зависит от числа факторов x_i , включаемых в модель. В случае линейной регрессии для получения статистически значимой модели требуется не менее 7 наблюдений.

Результаты расчетов подставим в (**):

$$\begin{cases} 10a + 207b = 462, \\ 207a + 5023b = 8360; \end{cases}$$

отсюда $b = \dfrac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} = \dfrac{836 - 20{,}7 \cdot 46{,}2}{502{,}3 - (20{,}7)^2} = -1{,}63$,

$a = \bar{y} - b\bar{x} = 46{,}2 + 1{,}63 \cdot 20{,}7 = 79{,}95$.

Таким образом, уравнение парной линейной регрессии имеет вид:

$$\hat{y} = -1{,}63x + 79{,}95. \qquad (1)$$

Изобразим прямую регрессии (1) на корреляционном поле (рис. 10)

Для анализа силы линейной зависимости вычислим коэффициент корреляции:

$$r_{xy} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\sqrt{\overline{y^2} - \bar{y}^2}} = \frac{836 - 20{,}7 \cdot 46{,}2}{\sqrt{502{,}3 - (20{,}7)^2}\sqrt{236{,}2 - (46{,}2)^2}} = -0{,}928.$$

Данное значение коэффициента корреляции позволяет сделать вывод о сильной (обратной) линейной зависимости между x и y.

По таблице распределения Стьюдента для $\alpha = 0{,}05$ и $k = 10 - 2 = 8$ находим $t_{табл.} = t_{0{,}05;8} = 2{,}3$, коэффициента корреляции r_{xy}.

$$S_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n -}} = 0{,}13$$

Сопоставим значение r_{xy} с соответствующей величиной случайной ошибки S_r

$$t_r = \frac{|r_{xy}|}{S_r} = \frac{0{,}928}{0{,}13} = 7{,}198.$$

В таблице распределений Стьюдента $\alpha = 0{,}05$ и числу степеней свободы $k = 8$ находим $t_{табл.} = 2{,}3$ т. к. $t_r > t_{табл.}$, то гипотеза H_0 отклоняется и r_{xy} не случайно отличается от 0, т. е. статистически значимы.

Пример 11. Экономист, изучая зависимость выработки Y (*ден. ед.*) на одного работника торговли от величины товарооборота X (*ден. ед.*) магазина за определённый период, получил данные по $n = 15$ магазинам одинакового профиля (см. табл. 5).

Таблица 6

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	150	38	85	28	146	34	95	50	134	120	74	140	110	60	86
Y	7,2	5,8	7,5	4,4	8,4	4,5	7,0	5,0	6,4	8,0	6,0	7,8	6,2	5,8	6,0

Требуется:

1. По данным, приведённым в таблице, вычислить числовые характеристики величин X и Y: средние $\bar{x}, \bar{y}$; средние квадратические отклонения s_x, s_y, корреляционный момент K_{xy}, коэффициент корреляции r.

2. Проверить значимость коэффициента корреляции.

3. Построить диаграмму рассеяния и по характеру расположения точек на диаграмме подобрать общий вид функции регрессии.

4. Найти эмпирические функции регрессии Y на X и X на Y и построить их графики.

Решение.

1. Прежде чем вычислять основные статистические характеристики, составим таблицу, содержащую исходные данные и промежуточные вычисления (табл. 7).

Таблица 7

№ п/п	X	Y	X^2	Y^2	$X\ Y$
1	**2**	**3**	**4**	**5**	**6**
1	150	7,2	22500	51,84	1080,0
2	38	5,8	1444	33,64	220,4
3	85	7,5	7225	56,25	637,5
4	28	4,4	784	19,36	123,2
5	146	8,4	21316	70,56	1226,4
6	34	4,5	1156	20,25	153,0
7	95	7,0	9025	49,00	665,0
8	50	5,0	2500	25,00	250,0
9	134	6,4	17956	40,96	857,6
10	120	8,0	14400	64,00	960,0
11	74	6,0	5476	36,00	444,0
12	140	7,8	19600	60,84	1092,0
13	110	6,2	12100	38,44	682,0
14	60	5,8	3600	33,64	348,0
15	86	6,0	7396	36,00	516,0
$\sum_{i=1}^{15}$	1350	96,0	146482	640,78	9261,1

Используя полученные суммы по столбцам, вычислим статистические характеристики:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1350}{15} = 90;\ \ \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i = \frac{96}{15} = 6{,}4;$$

$$s_x^2 = \bar{x}^2 - (\bar{x})^2 = \frac{146478}{15} - 90^2 = 1665{,}20;$$

$$s_x^2 = \bar{y}^2 - (\bar{y})^2 = \frac{640{,}78}{15} - (6{,}4)^2 = 1{,}425;$$

$$s_x = \sqrt{s_x^2} = \sqrt{1665{,}20} \approx 40{,}81; \quad s_y = \sqrt{s_y^2} = \sqrt{1{,}425} \approx 1{,}19;$$

$$K_{xy} = \overline{xy} - \bar{x} \cdot \bar{y} = \frac{9261{,}1}{15} - 90 \cdot 6{,}4 \approx 41{,}01;$$

$$r = \frac{K_{xy}}{s_x \cdot s_y} = \frac{41{,}01}{40{,}81 \cdot 1{,}19} \approx 0{,}84.$$

2. Проверим значимость полученного выборочного коэффициента корреляции по t-критерию Стьюдента. Наблюдаемое значение критерия

$$t_{набл} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}84\sqrt{15-2}}{\sqrt{1-(0{,}84)^2}} \approx 5{,}58.$$

По таблицам квантилей распределения Стьюдента по наиболее употребляемому в технике уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $v = n - 2 = 15 - 2 = 13$ находим критическое значение критерия Стьюдента

$$t_{кр} = t_{\alpha;v} = t_{0{,}05;16} = 2{,}12.$$

Так как $t_{набл} = 5{,}58 > t_{кр} = 2{,}12$, то выборочный коэффициент корреляции значимо отличается от нуля.

Значение коэффициента корреляции попало в интервал от 0,7 до 0,9 таблицы Чеддока (табл. П. 7), т. е. связь между признаками Y и X высокая.

3. По приведённым данным строим диаграмму рассеяния.

Для этого на плоскость xOy наносим 15 точек, координаты которых заданы в условии: $(x_1; y_1) = (150; 7{,}2)$, $(x_2; y_2) = (38; 5{,}8)$, $(x_3; y_3) = (85; 7{,}5)$, …, $(x_{15}; y_{15}) = (86; 6{,}0)$ (см. рис. 5. На этом же рисунке начертим, после получения уравнений, и линии регрессии.)

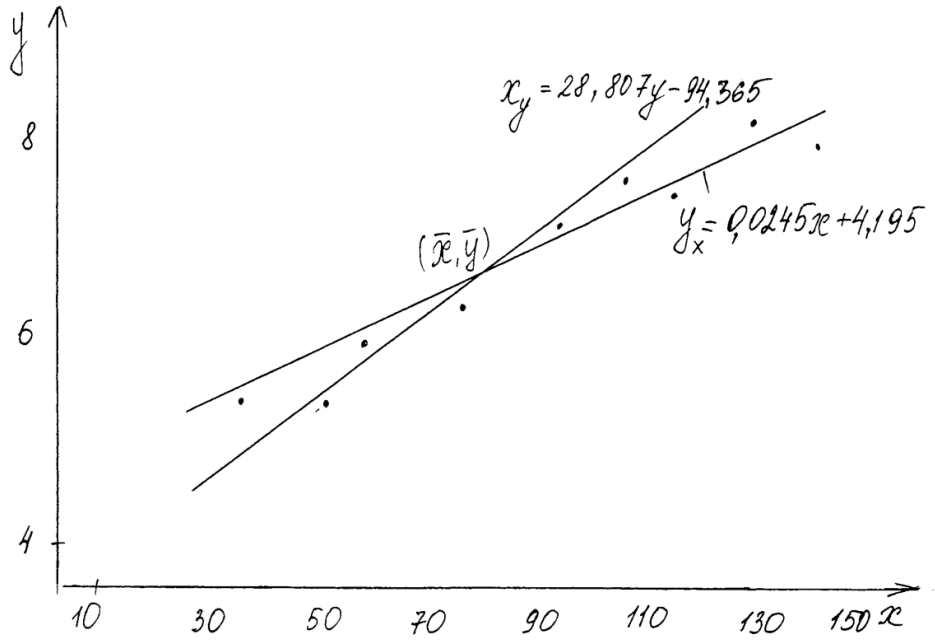


Рис. 11

4. Из диаграммы рассеяния видно, что связь между признаками X и Y можно принять линейной. Находить уравнения прямых линий регрессии Y на X и X на Y будем в виде

$$\bar{y}_x = \rho_{xy} x + a \text{ и } \bar{x}_y = \rho_{yx} y + b$$

соответственно. Параметры этих уравнений вычисляются по формулам

$$\rho_{xy} = r_B \frac{s_y}{s_x},\ a = \bar{y} - \rho_{xy}\bar{x};$$

$$\rho_{yx} = r_B \frac{s_x}{s_y},\ b = \bar{x} - \rho_{xy}\bar{y}.$$

Таким образом, имеем

$$\rho_{xy} = 0{,}84 \cdot \frac{1{,}19}{40{,}81} \approx 0{,}0245,\ a = 6{,}4 - 0{,}0245 \cdot 90 \approx 4{,}186;$$

$$\rho_{yx} = 0{,}84 \cdot \frac{40{,}81}{1{,}19} \approx 28{,}807,\ b = 90 - 28{,}807 \cdot 6{,}4 \approx -94{,}365.$$

Контроль вычислений по формуле (7):

$$\sqrt{\rho_{xy} \cdot \rho_{yx}} = \sqrt{0{,}0245 \cdot 28{,}807} \approx ,84 = r_B.$$

Тогда линейные уравнения регрессии имеют вид:

Y на X

$$\bar{y}_x = 0{,}0245x + 4{,}195.$$

ЗАДАНИЯ К ТИПОВОМУ РАСЧЕТУ (КОНТРОЛЬНОЙ РАБОТЕ) ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

При решении задач типового расчета по математической статистике используются основные понятия математической статистики: способы группировки статистического материала, числовые характеристики выборки, теория точечных и интервальных оценок параметров нормального распределения, теория проверки статистических гипотез, элементы регрессионного анализа и теории корреляции. Соответствующий теоретический материал должен быть изучен студентами самостоятельно.

ЗАДАЧА 1

В таблице 8 приведён статистический ряд распределения случайной величины X.

Требуется:

а) Вычислить числовые характеристики выборки: выборочное среднее $\bar{x}$; выборочное среднее квадратическое отклонение s; выборочные коэффициенты асимметрии и эксцесса A^* и E^*; выборочный коэффициент вариации V.

б) Найти интервальные оценки параметров нормального закона распределения (доверительную вероятность принять равной $\gamma = 0{,}95$).

Таблица 8

1	x_i	4,0–4,2	4,2–4,4	4,4–4,6	4,6–4,8	4,5–5,0
	n_i	6	20	46	23	11
2	x_i	3,4–3,6	3,6–3,8	3,8–4,0	4,0–4,2	4,2–4,4
	n_i	9	24	119	43	5
3	x_i	4,0–4,02	4,02–4,04	4,04–4,06	4,06–4,08	4,08–4,10
	n_i	6	20	46	23	11
4	x_i	2,0–2,2	2,2–2,4	2,4–2,6	2,6–2,8	2,8–3,0
	n_i	6	23	38	25	8
5	x_i	9,7–9,8	9,8–9,9	9,9–10,0	10,0–10,01	10,1–10,2
	n_i	4	11	17	13	5
6	x_i	-0,4– -0,2	-0,2– 0,0	0,0 – 0,2	0,2 – 0,4	0,4 – 0,6
	n_i	13	18	45	19	5
7	x_i	19,80–19,82	19,82–19,84	19,84–19,86	19,86–19,88	19,88–19,90
	n_i	6	13	15	11	5
8	x_i	20,00–20,04	20,04–20,08	20,08–20,12	20,12–20,16	20,16–20,20
	n_i	7	19	45	20	9
9	x_i	2,0–2,2	2,2–2,4	2,4–2,6	2,6–2,8	2,8–3,0
	n_i	7	20	44	21	8
10	x_i	1,2–1,6	1,6–2,0	2,0–2,4	2,4–2,8	2,8–3,2
	n_i	7	20	48	19	6

ЗАДАЧА 2

В условиях задачи 1 требуется:

а) Проверить гипотезу о нормальном распределении генеральной совокупности с помощью критерия согласия χ^2 (Пирсона).

б) Найти интервальные оценки параметров нормального закона распределения (с доверительной вероятностью $\gamma = 0{,}95$).

ЗАДАЧА 3

В таблице 9 приведены наблюдаемые значения признаков X и Y. Требуется:

1. По данным, приведённым в таблице, вычислить числовые характеристики величин X и Y: средние $\bar{x}$, $\bar{y}$; средние квадратические отклонения s_x, s_y, корреляционный момент K_{xy}, коэффициент корреляции r_B.

2. Проверить значимость коэффициента корреляции.

3. Построить диаграмму рассеяния и по характеру расположения точек на диаграмме подобрать общий вид функции регрессии.

4. Найти эмпирические функции регрессии Y на X и X на Y и построить их графики.

Таблица 9

1	*X*	110	85	70	120	150	90	60	140	100	115
	Y	6,1	4,2	2,9	5,8	8,3	5,2	3,4	7,5	4,9	5,4
2	*X*	80	60	100	130	120	50	90	150	70	125
	Y	4,2	4,9	7,2	9,1	6,4	3,9	5,1	8,4	3,5	8,7
3	*X*	160	120	110	80	90	130	150	70	100	60
	Y	12,5	9,3	9,2	6,4	7,5	11,6	13,1	5,2	7,9	4,4
4	*X*	50	130	100	80	90	70	150	60	140	110
	Y	4,2	10,8	9,6	5,1	7,4	6,2	11,4	3,3	12,2	10,5
5	*X*	60	90	150	80	110	120	70	130	100	140
	Y	2,9	7,1	11,8	6,3	7,2	8,4	4,8	11,2	6,7	10,6
6	*X*	70	110	85	65	100	90	120	80	130	110
	Y	2,8	3,5	2,4	2,1	3,4	3,2	3,6	2,5	4,1	3,3
7	*X*	80	60	100	70	50	110	90	40	75	105
	Y	4,2	4,0	4,5	3,6	3,4	5,2	3,9	3,1	3,3	4,9
8	*X*	100	110	60	120	70	80	130	75	105	50
	Y	3,8	4,4	3,2	4,8	3,0	3,5	4,5	3,3	4,1	3,1
9	*X*	120	85	110	70	115	90	60	55	100	130
	Y	4,0	3,6	4,0	2,6	4,3	3,4	2,9	2,5	3,0	4,5
10	*X*	140	110	120	90	130	80	100	75	135	60
	Y	5,4	4,1	5,6	3,3	4,2	2,9	3,6	2,5	4,9	3,0

ПРИЛОЖЕНИЯ

Таблица П. 1

Значение функции $P(x=m)=\frac{a^m}{m!}e^{-a}$

m \ a	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488
1	0,0905	0,1638	0,2222	0,2681	0,3033	0,3293
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198
4		0,0001	0,0002	0,0007	0,0016	0,0030
5				0,0001	0,0002	0,0004
m \ a	**0,7**	**0,8**	**0,9**	**1,0**	**2,0**	**3,0**
0	0,4966	0,4493	0,4066	0,3679	0,1353	0,0498
1	0,3476	0,3595	0,3659	0,3679	0,2707	0,1494
2	0,1217	0,1438	0,1647	0,1839	0,2707	0,2240
3	0,0284	0,0383	0,0494	0,0613	0,1804	0,2240
4	0,0050	0,0077	0,0111	0,0153	0,0902	0,1680
5	0,0007	0,0012	0,0020	0,0031	0,0361	0,1008
6	0,0001	0,0002	0,0003	0,0005	0,0120	0,0504
7				0,0001	0,0034	0,0216
8					0,0009	0,0081
9					0,0002	0,0027
10						0,0008
11						0,0002
12						0,0001
m \ a	**4,0**	**5,0**	**6,0**	**7,0**	**8,0**	**9,0**
0	0,0183	0,0067	0,0025	0,0009	0,0003	0,0001
1	0,0733	0,0337	0,0149	0,0064	0,0027	0,0011
2	0,1465	0,0842	0,0446	0,0223	0,0107	0,0050
3	0,1954	0,1404	0,0892	0,0521	0,0286	0,0150
4	0,1954	0,1755	0,1339	0,0912	0,0572	0,0337
5	0,1563	0,1555	0,1606	0,1277	0,0916	0,0607
6	0,1042	0,1462	0,1606	0,1490	0,1221	0,0911
7	0,0595	0,1044	0,1377	0,1490	0,1396	0,1171
8	0,0298	0,0653	0,1033	0,1304	0,1396	0,1318
9	0,0132	0,0363	0,0688	0,1014	0,1241	0,1318
10	0,0053	0,0181	0,0413	0,0710	0,0993	0,1186
11	0,0019	0,0082	0,0225	0,0452	0,0722	0,0970
12	0,0006	0,0034	0,0113	0,0264	0,0481	0,0728
13	0,0002	0,0013	0,0052	0,0142	0,0296	0,0504
14	0,0001	0,0005	0,0022	0,0071	0,0169	0,0324
15		0,0022	0,0009	0,0033	0,0090	0,0194
16		0,0001	0,0003	0,0015	0,0045	0,0109
17			0,0001	0,0006	0,0021	0,0058
18				0,0002	0,0009	0,0029
19					0,0004	0,0014
20					0,0001	0,0006
21						0,0001

Таблица П. 2

Значение функции $P(m \le k) \sum_{m=0}^{k} \frac{a^m}{m!} e^{-a}$

m \ a	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488
1	0,9953	0,9825	0,9631	0,9384	0,9098	0,8781
2	0,9998	0,9989	0,9964	0,9921	0,9856	0,9769
3	1,0000	0,9999	0,9997	0,9992	0,9983	0,9966
4		1,0000	1,0000	0,9999	0,9998	0,9996
5				1,0000	1,0000	1,0000
m \ a	**0,7**	**0,8**	**0,9**	**1,0**	**2,0**	**3,0**
0	0,4966	0,4493	0,4066	0,3679	0,1353	0,0498
1	0,8442	0,8088	0,7725	0,7358	0,4060	0,1991
2	0,9659	0,9526	0,9371	0,9197	0,6767	0,4232
3	0,9942	0,9909	0,9865	0,9810	0,8571	0,6472
4	0,9992	0,9986	0,9977	0,9963	0,9473	0,8153
5	0,9999	0,9998	0,9997	0,9994	0,9834	0,9161
6	1,000	1,0000	1,000	0,9999	0,9955	0,9665
7				1,0000	0,9989	0,9881
8					0,9998	0,9962
9					1,0000	0,9989
10						0,9997
11						0,9999
12						1,0000
m \ a	**4,0**	**5,0**	**6,0**	**7,0**	**8,0**	**9,0**
0	0,0183	0,0067	0,0025	0,0009	0,0003	0,0001
1	0,0916	0,0404	0,0174	0,0073	0,0030	0,0012
2	0,2381	0,1247	0,0620	0,0296	0,0138	0,0062
3	0,4335	0,2650	0,1512	0,0818	0,0424	0,0212
4	0,6288	0,4405	0,2851	0,1730	0,0996	0,0550
5	0,7851	0,6160	0,4457	0,3008	0,1912	0,1157
6	0,8893	0,7622	0,6063	0,4497	0,3133	0,2068
7	0,9489	0,8666	0,7440	0,5987	0,4530	0,3239
8	0,9786	0,9318	0,8472	0,7291	0,5925	0,4557
9	0,9919	0,9682	0,9161	0,8305	0,7166	0,5874
10	0,9972	0,9863	0,9574	0,9015	0,8159	0,7060
11	0,9991	0,9945	0,9799	0,9467	0,8881	0,8030
12	0,9997	0,9980	0,9912	0,9730	0,9362	0,8758
13	1,0000	0,9992	0,9964	0,9872	0,9658	0,9261
14		0,9998	0,9986	0,9943	0,9837	0,9585
15		1,0000	0,9995	0,9976	0,9918	0,9780
16			0,9998	0,9990	0,9963	0,9889
17			0,9999	0,9996	0,9984	0,9947
18			1,0000	0,9999	0,9994	0,9976
19				1,0000	0,9997	0,9989
20					0,9999	0,9996
21					1,0000	0,9998

Таблица П. 3

Значение функции $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$

x	$\varphi(x)$	x	$\varphi(x)$	x	$\varphi(x)$
0,00	0,3989	**0,40**	0,3683	**0,80**	0,2897
0,01	0,3989	0,41	0,3668	0,81	0,2874
0,02	0,3989	0,42	0,3653	0,82	0,2850
0,03	0,3988	0,43	0,3637	0,83	0,2827
0,04	0,3986	0,44	0,3621	0,84	0,2803
0,05	0,3984	0,45	0,3605	0,85	0,2780
0,06	0,3982	0,46	0,3589	0,86	0,2756
0,07	0,3980	0,47	0,3572	0,87	0,2732
0,08	0,3977	0,48	0,3555	0,88	0,2709
0,09	0,3973	0,49	0,3538	0,89	0,2685
0,10	0,3970	**0,50**	0,3521	**0,90**	0,2661
0,11	0,3965	0,51	0,3503	0,91	0,2637
0,12	0,3961	0,52	0,3485	0,92	0,2613
0,13	0,3956	0,53	0,3467	0,93	0,2589
0,14	0,3951	0,54	0,3448	0,94	0,2565
0,15	0,3945	0,55	0,3429	0,95	0,2541
0,16	0,3939	0,56	0,3410	0,96	0,2516
0,17	0,3932	0,57	0,3391	0,97	0,2492
0,18	0,3925	0,58	0,3372	0,98	0,2468
0,19	0,3918	0,59	0,3352	0,99	0,2444
0,20	0,3910	**0,60**	0,3332	**1,00**	0,2420
0,21	0,3902	0,61	0,3312	1,01	0,2396
0,22	0,3894	0,62	0,3292	1,02	0,2371
0,23	0,3885	0,63	0,3271	1,03	0,2347
0,24	0,3876	0,64	0,3251	1,04	0,2323
0,25	0,3867	0,65	0,3230	1,05	0,2299
0,26	0,3857	0,66	0,3209	1,06	0,2275
0,27	0,3847	0,67	0,3187	1,07	0,2251
0,28	0,3836	0,68	0,3166	1,08	0,2227
0,29	0,3825	0,69	0,3144	1,09	0,2203
0,30	0,3814	**0,70**	0,3123	**1,10**	0,2179
0,31	0,3802	0,71	0,3101	1,11	0,2155
0,32	0,3790	0,72	0,3079	1,12	0,2131
0,33	0,3778	0,73	0,3056	1,13	0,2107
0,34	0,3765	0,74	0,3034	1,14	0,2083
0,35	0,3752	0,75	0,3011	1,15	0,2059
0,36	0,3739	0,76	0,2989	1,16	0,2036
0,37	0,3725	0,77	0,2966	1,17	0,2012
0,38	0,3712	0,78	0,2943	1,18	0,1989
0,39	0,3697	0,79	0,2920	1,19	0,1965

Продолжение табл. П. 3

x	$\varphi(x)$	x	$\varphi(x)$	x	$\varphi(x)$	x	$\varphi(x)$
1,20	0,1942	**1,60**	0,1109	**2,00**	0,0540	**2,80**	0,0079
1,21	0,1919	1,61	0,1092	2,02	0,0519	2,82	0,0075
1,22	0,1895	1,62	0,1074	2,04	0,0498	2,84	0,0071
1,23	0,1872	1,63	0,1057	2,06	0,0478	2,86	0,0067
1,24	0,1849	1,64	0,1040	2,08	0,0459	2,88	0,0063
1,25	0,1825	1,65	0,1023	**2,10**	0,0440	**2,90**	0,0060
1,26	0,1804	1,66	0,1006	2,12	0,0422	2,92	0,0056
1,27	0,1781	1,67	0,0989	2,14	0,0404	2,94	0,0053
1,28	0,1758	1,68	0,0973	2,16	0,0387	2,96	0,0050
1,29	0,1736	1,69	0,0957	2,18	0,0371	2,98	0,0047
1,30	0,1714	**1,70**	0,0940	**2,20**	0,0355	**3,00**	0,00443
1,31	0,1691	1,71	0,0925	2,22	0,0339	3,10	0,00327
1,32	0,1669	1,72	0,0909	2,24	0,0325	3,20	0,00238
1,33	0,1647	1,73	0,0898	2,26	0,0310	3,30	0,00172
1,34	0,1626	1,74	0,0878	2,28	0,0297	3,40	0,00123
1,35	0,1604	1,75	0,0863	**2,30**	0,0283	3,50	0,00087
1,36	0,1582	1,76	0,0848	2,32	0,0270	3,60	0,00061
1,37	0,1561	1,77	0,0833	2,34	0,0258	3,70	0,00042
1,38	0,1539	1,78	0,0818	2,36	0,0246	3,80	0,00029
1,39	0,1518	1,79	0,0804	2,38	0,0235	3,90	0,00020
1,40	0,1497	**1,80**	0,0790	**2,40**	0,0224	**4,00**	0,0001338
1,41	0,1476	1,81	0,0775	2,42	0,0213	4,50	0,0000160
1,42	0,1456	1,82	0,0761	2,44	0,0203	5,00	0,0000015
1,43	0,1435	1,83	0,0748	2,46	0,0194		
1,44	0,1415	1,84	0,0734	2,48	0,0184		
1,45	0,1394	1,85	0,0721	**2,50**	0,0175		
1,46	0,1374	1,86	0,0707	2,52	0,0167		
1,47	0,1354	1,87	0,0694	2,54	0,0158		
1,48	0,1334	1,88	0,0681	2,26	0,0151		
1,49	0,1315	1,89	0,0669	2,58	0,0143		
1,50	0,1295	**1,90**	0,0656	**2,60**	0,0135		
1,51	0,1276	1,91	0,0644	2,62	0,0129		
1,52	0,1257	1,92	0,0632	2,64	0,0122		
1,53	0,1238	1,93	0,0620	2,66	0,0116		
1,54	0,1219	1,94	0,0608	2,68	0,0110		
1,55	0,1200	1,95	0,0596	**2,70**	0,0104		
1,56	0,1182	1,96	0,0584	2,72	0,0099		
1,57	0,1163	1,97	0,0573	2,74	0,0093		
1,58	0,1145	1,98	0,0562	2,76	0,0088		
1,59	0,1127	1,99	0,0551	2,78	0,0084		

Таблица П. 4

Значения функции Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$.

x	Сотые доли x									
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0200	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3437	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4485	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4642	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4985	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Таблица П. 5

Распределение Стьюдента.

Таблица значений t_γ, удовлетворяющих равенству $\gamma = \int_{-t_\gamma}^{t_\gamma} S(t,n)dt$ в зависимости от γ и $n-1$, где $S(t,n) = \frac{\tilde{A}\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\cdot\tilde{A}\left(\frac{n-1}{2}\right)}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}$,

$\alpha = q = 1-\gamma$ — уровень значимости.

γ — надежность							
$k = n-1$ **Число Степеней свободы**	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963
2	142	289	445	617	0,816	1,061	1,336
3	137	277	424	584	765	0,978	1,250
4	134	271	414	569	741	941	1,190
5	132	267	408	559	727	920	1,156
6	131	265	404	553	718	906	1,134
7	130	263	402	549	711	896	1,119
8	130	262	399	546	706	889	1,108
9	129	261	398	543	703	883	1,100
10	129	260	397	542	700	879	1,093
11	129	260	396	540	697	876	1,088
12	128	259	395	539	695	873	1.083
13	128	259	394	538	694	870	1,079
14	128	258	393	537	692	868	1,076
15	128	258	393	536	691	866	1,074
16	128	258	392	535	690	865	1,071
17	128	257	392	534	689	863	1,069
18	127	257	392	534	688	862	1,067
19	127	257	391	533	688	861	1,066
20	127	257	391	533	687	860	1,064
21	127	257	391	532	686	859	1,063
22	127	256	390	532	686	858	1,061
23	127	256	390	532	685	858	1,060
24	127	256	390	531	685	857	1.059
25	127	256	390	531	684	856	1,058
26	127	256	390	531	684	856	1,058
27	127	256	389	531	684	855	1,057
28	127	256	389	530	683	855	1,056
29	127	256	389	530	683	854	1,055
30	127	256	389	530	683	854	1,055
40	126	255	388	529	681	851	1.050
60	126	254	387	527	679	848	1,046
120	126	254	386	526	677	845	1,041
	0,126	0,253	0,385	0,524	0,674	0,842	1,036

γ — надежность						
$k = n - 1$ **Число степеней свободы**	**0,8**	**0,9**	**0,95**	**0,98**	**0,99**	**0,999**
1	3,08	6,31	12,71	31,8	63,7	63,70
2	1,886	2,92	4,30	6,96	9,92	31,6
3	1,638	2,35	3,18	4,54	5,84	12,94
4	1,533	2,13	2,77	3,75	4,60	8,61
5	1,476	2,02	2,57	3,36	4,03	6,86
6	1,440	1,943	2,45	3,14	4,71	5,96
7	1,415	1,895	2,36	3,00	3,50	5,40
8	1,397	1,860	2,31	2,90	3,36	5,04
9	1,383	1,833	2,26	2,82	3,25	4,78
10	1,372	1,812	2,23	2,76	3,17	4,59
11	1,363	1,796	2,20	2,72	3,11	4,49
12	1,356	1,782	2,18	2,68	3,06	4,32
13	1,350	1,771	2,16	2,65	3,01	4,22
14	1,345	1,761	2,14	2,62	2,98	4,14
15	1,341	1,753	2,13	2,60	2,95	4,07
16	1,337	1,746	2,12	2,58	2,92	4,02
17	1,333	1,740	2,11	2,57	2,90	3,96
18	1,330	1,734	2,10	2,55	2,88	3,92
19	1,328	1,729	2,09	2,54	2,86	3,88
20	1,325	1,725	2,09	2,53	2,84	3,85
21	1,323	1,721	2,08	2,52	2,83	3,82
22	1,321	1,717	2,07	2,51	2,82	3,79
23	1,319	1,714	2,07	2,50	2,81	3,77
24	1,318	1,711	2,06	2,49	2,80	3,74
25	1,316	1,708	2,06	2,48	2,79	3,72
26	1,315	1,706	2,06	2,48	2,78	3,71
27	1,314	1,703	2,05	2,47	2,77	3,69
28	1,313	1,701	2,05	2,47	2,76	3,67
29	1,311	1,699	2,04	2,46	2,76	3,66
30	1,310	1,697	2,04	2,46	2,75	3,65
40	1,303	1,684	2,02	2,42	2,70	3,55
60	1,296	1,671	2,00	2,39	2,66	3,46
120	1,289	1,658	1,980	2,36	2.62	3,37

Замечание: при больших k $(k > 30)$ можно пользоваться нормальным распределением.

Таблица П. 6

χ^2 — распределение.

Значения функции $\chi^2_{q,k}$.

Функция $\chi^2_{q,k}$ определяется равенством $P\left(\chi^2_k > \chi^2_{q,k}\right) = q$, где случайная величина χ^2_k имеет χ^2 — распределение с k степенями свободы.

$q = \alpha = 1 - \gamma$ — уровень значимости

γ — надежность

k	$q = \alpha$					
	0,99	**0,10**	**0,05**	**0,02**	**0,01**	**0,001**
1	0,00016	2,7	3,8	5,4	6,6	7,9
2	0,020	4,6	6,0	7,8	9,2	11,6
3	0,115	6,3	7,8	9,8	11,3	12,8
4	0,30	7,8	9,5	11,7	13,3	14,9
5	0,55	9,2	11,1	13,4	15,1	16,3
6	0,87	10,6	12,6	15,0	16,8	18,6
7	1,24	12,0	14,1	16,6	18,5	20,3
8	1,65	13,4	15,5	18,2	20,1	21,9
9	2,09	14,7	16,9	19,7	21,7	23,6
10	2,56	16,0	18,3	21,2	23,2	25,2
11	3,1	17,3	19,7	22,6	24,7	26,8
12	3,6	18,5	21,0	24,1	26,2	28,3
13	4,1	19,8	22,4	25,5	27,7	28,8
14	4,7	21,1	23,7	26,9	29,1	31,0
15	5,2	22,3	25,0	28,3	30,6	32,5
16	5,8	23,5	26,3	29,6	32,0	34,0
17	6,4	24,8	27,6	31,0	33,4	35,5
18	7,0	26,0	28,9	32,3	34,8	37,0
19	7,6	27,2	30,1	33,7	36,2	38,5
20	8,3	28,4	31,4	35,0	37,6	40,0
21	8,9	29,6	32,7	36,3	38,9	41,5
22	9,5	30,8	33,9	37,7	40,3	42,5
23	10,2	32,0	35,2	39,0	41,6	44,0
24	10,9	33,2	36,4	40,3	43,0	45,5
25	11,5	34,4	37,7	41,6	44,3	47,0

Плотность распределения χ^2_k равна

$$f_{\chi^2_k}(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right) \cdot 2^{\frac{k}{2}}} \cdot x^{\frac{k}{2}-1} \cdot e^{-\frac{x}{2}}, \; x > 0$$

Таблица П. 7

Таблица Чеддока

Диапазон изменения $\lvert r_B \rvert$:	< 0,3	0,3–0,5	0,5–0,7	0,7–0,9	> 0,9
Характер тесноты связи	Слабая	Умеренная	Заметная	Высокая	Весьма высокая

Литература

1. Вентцель Е. С. Теория вероятностей. М.: Физматгиз, 1963.
2. Гмурман В. Е. Введение в теорию вероятностей и математическую статистику. М.: Высшая школа, 1966.
3. Гурский В. И. Сборник задач по теории вероятностей и математической статистике. М.; Н.: Высшая школа, 2008.
4. Космачева И. М. Задачи теории вероятностей и математической статистики. Астрахань: Гостехуниверситет, 2002.
5. Студенецкая В. Н. Решение задач по статистике, комбинаторике и теории вероятностей. Волгоград: Учитель, 2005.
6. Бондаренко П. С., Кацко И. А. Методические указания по теории вероятностей и математической статистики. Краснодар: КГАУ, 2013.
7. Синяев Н. И. Теория вероятностей и математическая статистика. Учебное пособие. М.: Юрайт, 2011.

Содержание

Для заметок

Учебное издание

Астахова Ина Сергеевна
Кошмак Виктор Константинович
Лисенков Андрей Викторович

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие

Технический редактор: И. С. Астахова
Компьютерная верстка: Т. М. Терентьева
Корректор: С. Н. Емельянова

Подписано в печать 12.09.2016. Формат 60×90/16.
Гарнитура Times New Roman. Усл. п. л. 3,5.
Тираж 100 экз. Заказ № 5211.

Изготовлено на Versant 2100.

Адрес издательства:
Россия, 180000, г. Псков, ул. Л. Толстого, д. 4а, корп. 3а
Издательство Псковского государственного университета

ISBN 978-5-91116-483-6