

**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ имени академика С. П. КОРОЛЕВА»**

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ

**МЕТОДИЧЕСКИЕ УКАЗАНИЯ
к расчетной работе**

**САМАРА
2004**

УДК 519.2(075)

Статистический анализ данных: Методические указания к расчетной работе / Самар. гос. аэрокосм. ун-т; Сост. Е. А. Денискина, П. Э. Коломиец. Самара, 2004. 64 с.

Методические указания составлены в соответствии с действующей программой по курсу математики для инженерно – технических специальностей вузов.

Методические указания представляют полное методическое обеспечение расчетно–графической работы «Статистический анализ данных».

Методические указания предназначены для студентов второго курса радиотехнического факультета СГАУ, а также могут быть использованы студентам других факультетов СГАУ.

Печатаются по решению редакционно-издательского совета Самарского государственного аэрокосмического университета имени академика С. П. Королева

Рецензент: Жданов А. И.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	4
ЗАДАНИЕ НА РАСЧЕТНУЮ РАБОТУ	4
1. СТАТИСТИЧЕСКИЙ АНАЛИЗ ОДНОМЕРНЫХ ДАННЫХ	6
1.1. ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И ВЫБОРКА	6
1.2. ТОЧЕЧНЫЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ.....	7
1.3. СТАТИСТИЧЕСКИЕ РЯДЫ.....	10
1.4. ГИСТОГРАММА И ПОЛИГОН ЧАСТОТ.....	16
1.5. ЭМПИРИЧЕСКАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ	22
1.6. РАСПРЕДЕЛЕНИЯ χ^2 И СТЬЮДЕНТА	26
1.7. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ	29
1.8. КРИТЕРИЙ СОГЛАСИЯ χ^2 ПИРСОНА.....	31
1.9. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ ..	37
2. СТАТИСТИЧЕСКИЙ АНАЛИЗ ДВУМЕРНЫХ ДАННЫХ	46
2.1. ФУНКЦИОНАЛЬНАЯ, СТАТИСТИЧЕСКАЯ И КОРРЕЛЯЦИОННАЯ ЗАВИСИМОСТИ.....	47
2.2. ЛИНЕЙНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ	47
2.3. УРАВНЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ	51
ЛИТЕРАТУРА	57
ПРИЛОЖЕНИЕ	58

ВВЕДЕНИЕ

Математическая статистика – это прикладная математическая дисциплина, примыкающая к теории вероятностей. Она базируется на понятиях и методах теории вероятностей, но решает свои специфические задачи специальными методами.

Основная задача математической статистики – получить обоснованные выводы о параметрах, видах распределений и других свойствах случайных величин по конечной совокупности наблюдений над ними.

В расчетной работе рассматриваются основные методы анализа одномерных статистических данных: определение точечных и интервальных оценок параметров распределения, проверка гипотез о виде распределения. Рассматриваются также элементы корреляционного и регрессионного анализа двумерных статистических данных.

ЗАДАНИЕ НА РАСЧЕТНУЮ РАБОТУ

Часть 1: «СТАТИСТИЧЕСКИЙ АНАЛИЗ ОДНОМЕРНЫХ ДАННЫХ»

Дана выборка значений случайной величины X (выборка объема $n = 100$ из генеральной совокупности).

1. Найти выборочную оценку математического ожидания случайной величины X , указать свойства этой оценки.
2. Найти выборочные оценки дисперсии и среднеквадратического отклонения случайной величины X , указать свойства этих оценок.
3. Составить группированный вариационный ряд, разбив выборку на $N = 10$ равных интервалов.
4. Построить гистограмму и полигон относительных частот. На их основе выдвинуть нулевую гипотезу H_0 о виде распределения (нормальное распределение).

5. На одном чертеже с гистограммой построить график теоретической плотности вероятностей. Сделать вывод об их визуальном совпадении.
6. Составить эмпирическую функцию распределения $F_n(x)$ и построить ее график.
7. На одном чертеже с эмпирической функцией распределения построить график теоретической функции распределения. Сделать вывод об их визуальном совпадении.
8. С помощью критерия согласия χ^2 Пирсона проверить гипотезу H_0 о виде распределения генеральной совокупности для уровня значимости $\alpha = 0,1$. Сделать статистический вывод.
9. Построить доверительные интервалы для неизвестных математического ожидания и дисперсии нормально распределенной генеральной совокупности с параметрами $m = \bar{x}$ и $\sigma = S_0$ для уровней значимости $\alpha = 0,1$, $\alpha = 0,05$ и $\alpha = 0,01$. Сделать вывод о ширине доверительного интервала, в зависимости от уровня значимости α .

У к а з а н и е: все вычисления проводить с точностью до 0,0001

Часть 2. «СТАТИСТИЧЕСКИЙ АНАЛИЗ ДВУМЕРНЫХ ДАННЫХ»

Дана выборка из n наблюдений случайного вектора (X, Y) . При этом $X = \{x_1, x_2, \dots, x_n\}$ и $Y = \{y_1, y_2, \dots, y_n\}$.

1. Определить выборочный коэффициент корреляции величин X и Y .
2. Составить уравнение линейной регрессии Y на X . Построить график уравнения линейной регрессии на одном чертеже с опытными данными.
3. Оценить качество линейной модели регрессии по коэффициенту детерминации R^2 .
4. На уровне значимости $\alpha = 0,1$ найти доверительный интервал, в который попадает прогнозное значение фактора y для $x^* = x_{\max} + 1$.

У к а з а н и е: все вычисления проводить с точностью до 0,0001.

1. СТАТИСТИЧЕСКИЙ АНАЛИЗ ОДНОМЕРНЫХ ДАННЫХ

1.1. ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И ВЫБОРКА

Предположим, что изучается некоторая случайная величина X , закон распределения которой неизвестен. Требуется приближенно определить этот закон из опыта и проверить гипотезу о том, что случайная величина X подчинена этому закону.

Генеральной совокупностью называют всю совокупность реализации случайной величины X , все возможные наблюдения некоторого показателя, все возможные исходы некоторого испытания.

Выборкой называют часть генеральной совокупности $X_n = \{x_1, x_2, \dots, x_n\}$, то есть конечное подмножество значений случайной величины из множества элементов генеральной совокупности.

Объемом выборки n называют количество содержащихся в ней значений случайной величины X .

Задача математической статистики состоит в исследовании свойств выборки и обобщении этих свойств на всю генеральную совокупность.

Выборка является исходной информацией для статистического анализа и принятия решений о неизвестных вероятностных характеристиках случайной величины X . Для этих целей на выборку следует смотреть как на набор реализаций n независимых одинаково распределенных случайных величин (X_1, X_2, \dots, X_n) .

Для того чтобы по выборке можно было достаточно уверенно судить о генеральной совокупности, выборка должна быть **представительной (репрезентативной)**, то есть достаточно полно представлять признаки и параметры генеральной совокупности. Репрезентативность выборки улучшается при увеличении ее объема.

1.2. ТОЧЕЧНЫЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

Пусть $X_n = \{ x_1, x_2, \dots, x_n \}$ – выборка объема n из генеральной совокупности значений случайной величины X с математическим ожиданием $M[X]$, дисперсией $D[X]$ и среднеквадратическим отклонением $\sigma = \sqrt{D[X]}$.

Выборочным средним выборки называется среднее арифметическое

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i.$$

Согласно закону больших чисел, при увеличении объема выборки среднее арифметическое выборки сходится по вероятности к математическому ожиданию генеральной совокупности, то есть

$$\lim_{n \rightarrow \infty} P \left(\left| \frac{1}{n} \cdot \sum_{i=1}^n x_i - M[X] \right| \geq \varepsilon \right) = 0.$$

Таким образом, среднее арифметическое может служить приближением (оценкой) математического ожидания генеральной совокупности.

Выборочной дисперсией называется

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

Модифицированной выборочной дисперсией называется

$$S_0^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n \cdot S^2}{n-1}.$$

Все эти выборочные величины зависят от выборки и сами являются случайными величинами. Их значения лишь приблизительно равны соответствующим числовым характеристикам генеральной совокупности.

Статистикой называется любая функция, зависящая от выборки и сама являющаяся случайной величиной. Таким образом, выборочное среднее \bar{x} , выборочная дисперсия S^2 и модифицированная выборочная дисперсия S_0^2 – это статистики.

Точечной оценкой $\tilde{\theta}$ неизвестного параметра θ распределения случайной величины X называется такая функция от выборки (статистика) $\tilde{\theta}(X_n) = \tilde{\theta}(x_1, x_2, \dots, x_n)$, что ее значение от любой выборки приближенно равно истинному значению параметра, то есть $\tilde{\theta}(X_n) \approx \theta$.

Оценки параметров принято обозначать символом с тильдой наверху: $\tilde{\theta}$.

Существует несколько методов нахождения точечных оценок: метод наименьших квадратов, метод моментов, метод максимального правдоподобия и другие. Таким образом, для каждого независимого параметра может быть несколько оценок, полученных различными методами. Для того, чтобы точечная оценка давала хорошее приближение оцениваемому параметру, она должна обладать следующими свойствами:

1. Оценка $\tilde{\theta}$ параметра называется **несмещенной**, если ее математическое ожидание равно оцениваемому параметру θ :

$$M[\tilde{\theta}] = \theta.$$

Известно, что \bar{x} – несмещенная оценка математического ожидания, S^2 – смещенная оценка дисперсии и S_0^2 – несмещенная оценка дисперсии.

2. Оценка $\tilde{\theta}$ параметра называется **состоятельной**, если она сходится по вероятности к точному значению оцениваемого параметра θ , то есть

$$\lim_{n \rightarrow \infty} P(|\tilde{\theta} - \theta| \geq \varepsilon) = 0 \quad (\forall \varepsilon > 0).$$

Состоятельной оценкой математического ожидания является выборочное среднее \bar{x} , а состоятельными оценками дисперсии – выборочная дисперсия S^2 и модифицированная выборочная дисперсия S_0^2 .

3. Несмещенная оценка $\tilde{\theta}$ параметра называется **эффективной**, если она имеет минимальную дисперсию среди всех несмещенных оценок этого параметра. Доказано, что \bar{x} и S_0^2 являются эффективными оценками

математического ожидания и дисперсии соответственно, а так как S^2 – смещенная оценка дисперсии, то это и неэффективная оценка.

ПРИМЕР 1 (пункты 1 и 2 части 1 Задания):

Пусть дана выборка значений случайной величины X (выборка объема $n = 100$ из генеральной совокупности) (таблица 1).

Таблица 1

2,88	2,78	4,90	4,41	4,86	4,46	4,76	4,48	4,71	4,70
2,94	5,37	7,48	-3,32	5,79	8,55	8,27	5,65	7,23	7,95
2,95	2,44	7,89	2,45	5,90	2,45	2,67	2,50	2,67	2,51
5,16	4,40	9,12	5,52	1,56	8,46	1,34	5,69	9,57	-1,07
5,20	4,99	9,00	8,47	6,55	2,88	6,78	5,72	6,10	0,13
4,23	5,15	6,39	4,39	6,56	5,78	6,85	4,40	6,23	0,56
4,23	2,99	6,46	6,88	9,63	4,22	3,58	6,57	5,83	9,35
4,33	3,24	9,97	6,99	5,22	8,93	3,69	6,58	7,09	5,68
4,38	3,27	7,19	1,73	5,29	1,96	3,71	1,99	2,31	2,30
5,67	3,90	7,38	3,94	5,33	3,98	3,79	4,08	4,12	4,12

Требуется найти выборочные оценки математического ожидания, дисперсии и среднеквадратического отклонения случайной величины X . Указать свойства этих оценок.

Оценкой математического ожидания случайной величины X служит выборочное среднее $\tilde{m}_X = \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i = \frac{1}{100} \cdot 491,2900 = 4,9129$. Данная

оценка $\tilde{m}_X = \bar{x}$ является несмещенной, эффективной и состоятельной.

Оценкой дисперсии случайной величины X служат выборочная дисперсия и модифицированная выборочная дисперсия, вычисляемые по формулам:

$$S^2 = \frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1}{100} \cdot 2985,1739 - 4,9129^2 = 5,7152,$$

$$S_0^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n \cdot S^2}{n-1} = \frac{100 \cdot 5,7152}{99} = 5,7729.$$

Оценка S_0^2 является несмещенной, эффективной, состоятельной, а S^2 – смещенная, неэффективная, но состоятельная. Следовательно, S_0^2 дает лучшее приближение оцениваемой дисперсии, поэтому в дальнейших расчетах в качестве оценки дисперсии используется S_0^2 : $\tilde{D}_X = S_0^2$.

Оценка среднеквадратического отклонения, являющаяся несмещенной, эффективной, состоятельной:

$$\tilde{\sigma}_X = \sqrt{\tilde{D}_X} = \sqrt{S_0^2} = S_0 = 2,4027.$$

Найденные оценки параметров распределения можно найти с помощью статистических функций СРЗНАЧ, ДИСП, ДИСПР и СТАНДОТКЛОН пакета прикладных программ EXCEL.

1.3. СТАТИСТИЧЕСКИЕ РЯДЫ

Пусть $X_n = \{ x_1, x_2, \dots, x_k \}$ – выборка объема n , содержащая k различных вариантов, из генеральной совокупности случайной величины X .

Статистическим рядом называется совокупность пар (i, x_i) , полученных в результате эксперимента. Обычно статистические ряды оформляются в виде таблицы (таблица 2), в первом столбце которой стоит номер опыта, а во втором – наблюдаемое значение случайной величины, которое называется *вариантой*.

Размахом выборки называют разность между наибольшей и наименьшей вариантами выборки:

$$R = x_{\max} - x_{\min}.$$

Если одна и та же варианта встречается в выборке несколько раз, то статистический ряд удобнее записывать в виде таблицы 3.

Таблица 2

Индекс i	Варианта x_i
1	x_1
2	x_2
...	...
n	x_n

Таблица 3

Индекс i	Варианта x_i	Частота n_i	Относит. частота \bar{n}_i
1	x_1	n_1	\bar{n}_1
2	x_2	n_2	\bar{n}_2
...
k	x_k	n_k	\bar{n}_k

Частотой n_i ($i = \overline{1, k}$) варианты x_i называется число повторений варианты x_i в выборке, причем $\sum_{i=1}^k n_i = n$.

Относительной частотой или **весом** \bar{n}_i ($i = \overline{1, k}$) варианты x_i называется отношение частоты варианты x_i к объему выборки n , то есть

$$\bar{n}_i = \frac{n_i}{n}, \text{ причем } \sum_{i=1}^k \bar{n}_i = 1.$$

При большом числе наблюдений простой статистический ряд перестает быть удобной формой записи статистических данных. Для придания ему большей компактности и наглядности статистический материал подвергают дополнительной обработке – строят вариационные ряды или группированные вариационные ряды.

Вариационным рядом называется **упорядоченная** совокупность вариант x_i ($i = \overline{1, k}$) с соответствующими им частотами n_i или относительными частотами \bar{n}_i .

Для построения группированного вариационного ряда интервал изменения наблюдаемых значений случайной величины $[x_{\min}; x_{\max}]$ разбивают на N непересекающихся интервалов $[u_1 = x_{\min}; u_2]$, $(u_2; u_3]$,

..., $(u_N; u_{N+1} = x_{\max}]$, называемых **частичными интервалами** или **разрядами**. Число интервалов группировки зависит от объема выборки и определяется по правилу:

$$N \geq [1 + 3,32 \cdot \lg n] + 1,$$

где n – объем выборки, а квадратные скобки обозначают целую часть числа. Разбиение на малое число интервалов может привести к неверным статистическим выводам. Согласно этой формуле, необходимо брать не менее 8 интервалов на 100 наблюдений.

Интервалы могут быть как одинаковой длины, так и различной. Для упрощения дальнейшей обработки статистических данных интервалы желательно делать одинаковой длины:

$$\Delta = \frac{x_{\max} - x_{\min}}{N} = \frac{R}{N}.$$

Частотой n_i ($i = \overline{1, N}$) интервала $(u_i; u_{i+1}]$ называется число вариант x_i , попавших в этот интервал, причем $\sum_{i=1}^N n_i = n$. При группировке

наблюденных значений по разрядам возникает вопрос о том, к какому интервалу отнести значение, находящееся на границе двух разрядов. В этих случаях считают данное значение принадлежащим к левому интервалу.

Относительной частотой или **весом** \bar{n}_i ($i = \overline{1, N}$) интервала $(u_i; u_{i+1}]$ называется отношение частоты интервала к объему выборки n :

$$\bar{n}_i = \frac{n_i}{n}, \text{ причем } \sum_{i=1}^N \bar{n}_i = 1.$$

Накопленной относительной частотой w_i ($i = \overline{1, N}$) интервала $(u_i; u_{i+1}]$ называется сумма относительных частот первых i интервалов, то

$$\text{есть } w_i = \sum_{j=1}^i \bar{n}_j.$$

Группированным вариационным рядом называется упорядоченная совокупность непересекающихся интервалов с соответствующими им частотами n_i , относительными частотами \bar{n}_i и накопленными относительными частотами w_i (таблица 4).

Таблица 4

Индекс i	Интервал $(u_i; u_{i+1}]$	Частота n_i	Относит. частота \bar{n}_i	Накопл. относит. частота w_i
1	$[u_1; u_2]$	n_1	\bar{n}_1	$w_1 = \bar{n}_1$
2	$(u_2; u_3]$	n_2	\bar{n}_2	$w_2 = \bar{n}_1 + \bar{n}_2$
...
N	$(u_N; u_{N+1}]$	n_N	\bar{n}_N	$w_N = 1$
$\sum_{i=1}^N$		n	1	

ПРИМЕР 2 (пункт 3 части 1 Задания):

Требуется составить группированный вариационный ряд для выборки из генеральной совокупности значений случайной величины X (таблица 1), разбив выборку на $N = 10$ равных интервалов.

Данная выборка имеет объем $n = 100$.

Определим интервал изменения случайной величины X . Для этого в таблице 1 находим максимальный и минимальный элементы:

$$x_{\max} = 9,97, \quad x_{\min} = -3,32.$$

Определим размах выборки:

$$R = x_{\max} - x_{\min} = 13,29.$$

Для удобства дальнейшей обработки статистических данных округляем x_{\max} и x_{\min} до ближайших целых чисел таких, что x_{\max} и x_{\min} вошли бы в новый интервал:

$$x_{\max}^o = 10, \quad x_{\min}^o = -4.$$

Тогда новый размах выборки: $R^o = x_{\max}^o - x_{\min}^o = 14$.

Разбиваем выборку на $N=10$ равных интервалов. Длина каждого частичного интервала равна $\Delta = \frac{R^o}{N} = \frac{14}{10} = 1,4$.

Частичные интервалы приведены во втором столбце таблицы 5.

Найдем количество вариантов, попавших в каждый частичный интервал разбиения, и заполним столбец три таблицы 5. Сумма всех частот должна быть равна $n = 100$.

Найдем относительные частоты $\bar{n}_i = \frac{n_i}{n}$ и накопленные относительные

частоты $w_i = \sum_{j=1}^i \bar{n}_j$ (четвертый и пятый столбцы таблицы 5).

Группированный вариационный ряд оформим в таблицу 5.

Из таблицы 5 видно, что данная выборка имеет одну изолированную точку $x_{\min} = -3,32$, удаленную от группы других экспериментальных точек. В таком случае можно считать эту изолированную точку аномальным наблюдением, грубой ошибкой измерения и удалить ее из выборки. Тогда объем выборки уменьшится и будет равен $n = 99$. Изменяются также и

выборочные характеристики: $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i = 4,9961$,

$$S_0^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = 5,1332, \quad S_0 = \sqrt{S_0^2} = \sqrt{5,1332} = 2,2657. \quad \text{В}$$

дальнейших расчетах будут использоваться именно эти значения.

Таблица 5

Индекс i	Интервал ($u_i; u_{i+1}$]	Частота n_i	Относит. частота \bar{n}_i	Накопл. относит. частота w_i
1	[-4,0; -2,6]	1	0,01	0,01
2	(-2,6; -1,2]	0	0	0,01
3	(-1,2; 0,2]	2	0,02	0,03
4	(0,2; 1,6]	3	0,03	0,06
5	(1,6; 3,0]	18	0,18	0,24
6	(3,0; 4,4]	20	0,2	0,44
7	(4,4; 5,8]	24	0,24	0,68
8	(5,8; 7,2]	16	0,16	0,84
9	(7,2; 8,6]	9	0,09	0,93
10	(8,6; 10,0]	7	0,07	1
Сумма		100	1	

З а м е ч а н и е: Проверка гипотезы об аномальности наблюдения проводится следующим образом: значение x_m признается аномальным и выбрасывается из выборки объема n , если $|x_m - \bar{x}| \geq C_p \cdot S_0 \cdot \sqrt{\frac{n+1}{n}}$, где значение квантили C_p определяется для данной доверительной вероятности p по таблице нормального распределения (таблица П 3 Приложения). Выберем доверительную вероятность $p = 0,95$ и по таблице П 3 Приложения найдем $C_{0,95} = 1,645$. Значения \bar{x} и S_0 определяются по выборке уменьшенного объема, то есть $\bar{x} = 4,9961$ и $S_0 = 2,2657$.

Проверим гипотезу об аномальности $x_{\min} = -3,32$:

$$|x_m - \bar{x}| = |-3,32 - 4,9961| = 8,3161,$$

$$C_p \cdot S_0 \cdot \sqrt{\frac{n+1}{n}} = 1,645 \cdot 2,2657 \cdot \sqrt{\frac{101}{100}} = 3,7457.$$

Так как условие $|x_m - \bar{x}| \geq C_p \cdot S_0 \cdot \sqrt{\frac{n+1}{n}}$ выполняется, точку

$x = -3,32$ можно из выборки исключить. Соответственно в таблице 5 можно исключить два первых интервала. Заметим, что число оставшихся интервалов группировки оказалось равно 8, что соответствует условию:

$$N \geq [1 + 3,32 \cdot \lg n] + 1 = [1 + 3,32 \cdot \lg 99] + 1 = 8.$$

В противном случае число интервалов пришлось бы увеличить.

Новое разбиение на интервалы оформим в таблицу 6.

Таблица 6

Индекс i	Интервал ($u_i; u_{i+1}$]	Частота n_i	Относит. частота \bar{n}_i	Накопл. относит. частота w_i
1	(-1,2; 0,2]	2	0,0202	0,0202
2	(0,2; 1,6]	3	0,0303	0,0505
3	(1,6; 3,0]	18	0,1818	0,2323
4	(3,0; 4,4]	20	0,2020	0,4343
5	(4,4; 5,8]	24	0,2424	0,6768
6	(5,8; 7,2]	16	0,1616	0,8384
7	(7,2; 8,6]	9	0,0909	0,9293
8	(8,6; 10,0]	7	0,0707	1,0000
Сумма		99	1,0000	

1.4. ГИСТОГРАММА И ПОЛИГОН ЧАСТОТ

Пусть $X_n = \{x_1, x_2, \dots, x_k\}$ – выборка объема n , содержащая k различных вариантов, из генеральной совокупности случайной величины X с неизвестной плотностью вероятностей $f(x)$. Приближением (оценкой) неизвестной плотности вероятностей могут служить **гистограмма** или **полигон относительных частот**. Гистограмма и полигон относительных частот служат для геометрического изображения группированного вариационного ряда.

Гистограмма относительных частот представляется в виде примыкающих друг к другу прямоугольников с основаниями $\Delta = \frac{R}{N}$, равными ширине интервалов группировок, и высотами $h_i = \frac{\bar{n}_i}{\Delta} = \frac{n_i}{n \cdot \Delta}$ (рис. 1). Для гистограммы относительных частот площадь ступенчатой фигуры соответствует сумме вероятностей и равна 1. Площадь любого прямоугольника гистограммы равна вероятности попадания значений рассматриваемой случайной величины в интервал, соответствующий основанию прямоугольника.

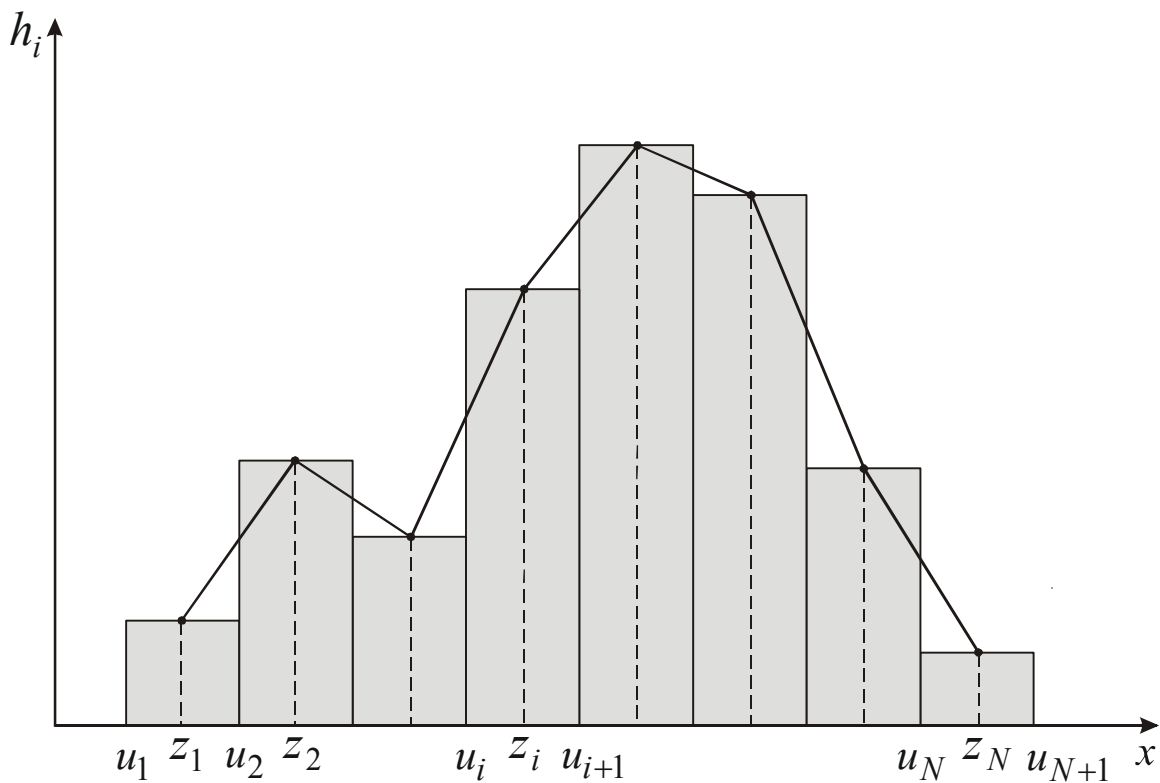


Рис. 1
Гистограмма и полигон относительных частот

Полигоном относительных частот называется ломаная, соединяющая точки (z_1, h_1) , (z_2, h_2) , ..., (z_N, h_N) (рис. 1), где $z_i = \frac{u_i + u_{i+1}}{2}$ –

середины интервалов группировки; $h_i = \frac{n_i}{n \cdot \Delta}$ – высоты прямоугольников гистограммы.

При увеличении объема выборки и уменьшении длин интервалов гистограмма и полигон относительных частот приближаются к графику неизвестной функции $f(x)$ – плотности вероятности генеральной совокупности.

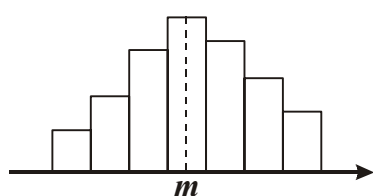
По виду гистограммы или полигона частот можно выдвинуть гипотезу о виде распределения генеральной совокупности. Например, если гистограмма имеет вид, представленный на рис. 2а, то можно предположить, что генеральная совокупность имеет нормальный закон распределения с

плотностью вероятностей $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$; рис. 2б – равномерное

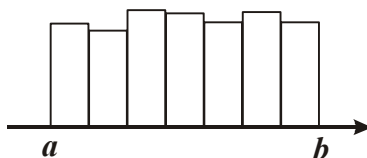
распределения с плотностью вероятностей $f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a;b] \\ 0, & x \notin [a;b] \end{cases}$; рис. 2в

– показательное распределение с плотностью вероятностей

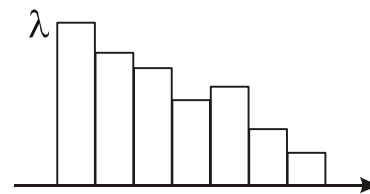
$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0 \end{cases}.$$



2а



2б



2в

Рис. 2
Виды гистограмм

ПРИМЕР 3 (пункты 4 и 5 части 1 Задания):

Требуется построить гистограмму и полигон относительных частот для известного группированного вариационного ряда (таблица 6). На их основе выдвинуть нулевую гипотезу H_0 о виде распределения генеральной совокупности. В данном случае это нормальное распределение. На одном чертеже с гистограммой построить график теоретической плотности вероятностей. Сделать вывод об их визуальном совпадении.

Для удобства заполним таблицу 7. В третий столбец таблицы 7 занесены середины интервалов $z_i = \frac{u_i + u_{i+1}}{2}$, в четвертый – относительные частоты интервалов $\bar{n}_i = \frac{n_i}{n}$, в пятый – высоты прямоугольников гистограммы относительных частот $h_i = \frac{\bar{n}_i}{\Delta} = \frac{n_i}{n \cdot \Delta}$.

Таблица 7

Индекс i	Интервал ($u_i; u_{i+1}$]	Середина интервала z_i	Относит. частота \bar{n}_i	Высота прямоуг. h_i
1	(-1,2; 0,2]	-0,5	0,0202	0,0144
2	(0,2; 1,6]	0,9	0,0303	0,0216
3	(1,6; 3,0]	2,3	0,1818	0,1299
4	(3,0; 4,4]	3,7	0,2020	0,1443
5	(4,4; 5,8]	5,1	0,2424	0,1732
6	(5,8; 7,2]	6,5	0,1616	0,1154
7	(7,2; 8,6]	7,9	0,0909	0,0649
8	(8,6; 10,0]	9,3	0,0707	0,0505
Сумма			1,0000	1,0000

По данным таблицы 7 построим гистограмму. Для этого в прямоугольной системе координат на оси абсцисс откладываем значения границ интервалов разбиения и на каждом из интервалов с номером i строим прямоугольник с высотой h_i (рис. 3).

Для такой гистограммы площадь ступенчатой фигуры соответствует сумме вероятностей и равна 1. Площадь каждого прямоугольника гистограммы равна вероятности попадания случайной величины в интервал, соответствующий основанию прямоугольника.

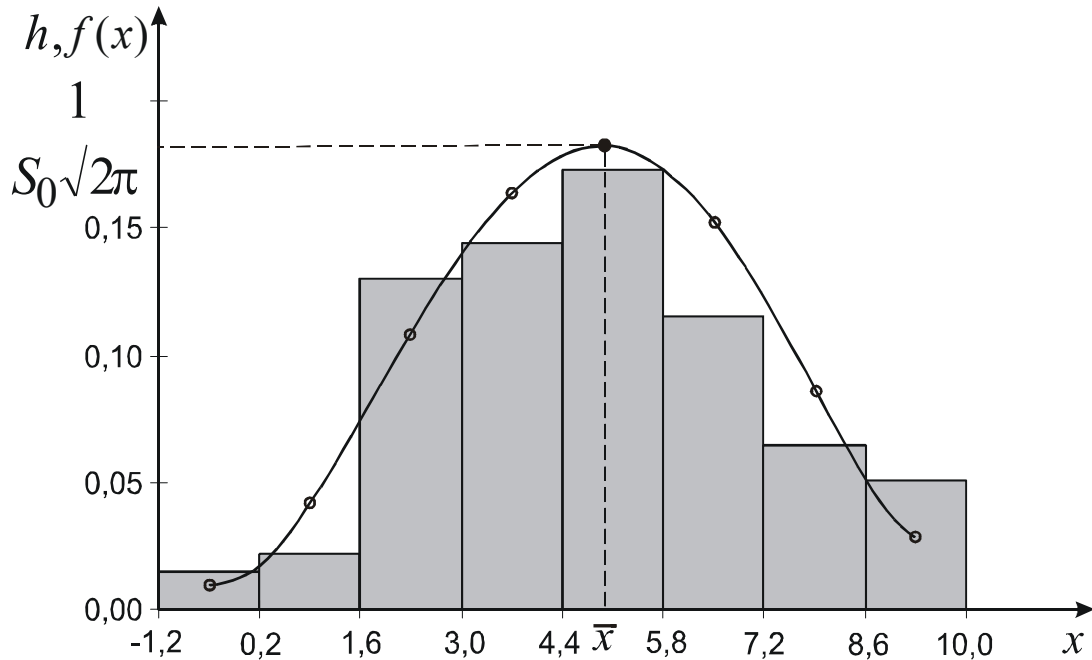


Рис. 3
Гистограмма относительных частот и
кривая теоретической плотности вероятностей

Полигон относительных частот – ломаная, соединяющая точки (z_i, h_i) , $i = \overline{1, N}$ (рис. 4).

Гистограмма и полигон относительных частот, являющиеся статистическими оценками плотности вероятностей генеральной совокупности, схожи с кривой плотности вероятностей нормального закона. На основании этого выдвигаем нулевую гипотезу H_0 : Генеральная совокупность, из которой взята выборка, распределена по нормальному закону с параметрами $m = \bar{x} = 4,9961$, $\sigma = S_0 = 2,2657$, то есть теоретическая плотность

вероятностей имеет вид:
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-m)^2}{2\sigma^2}}$$

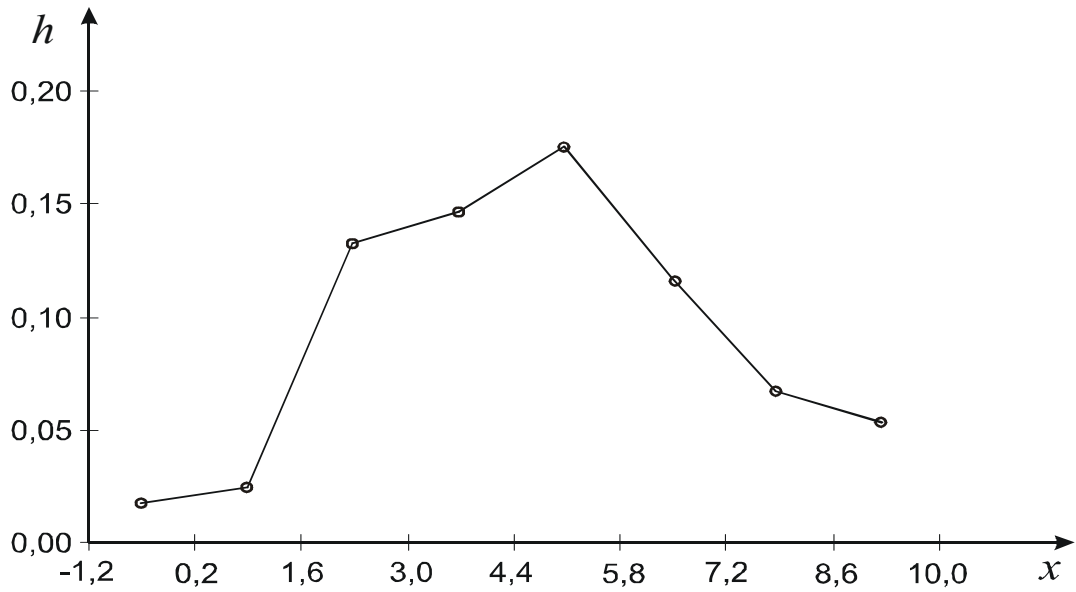


Рис. 4
Полигон относительных частот

Вычислим значения теоретической плотности вероятностей в точках z_i – середины интервалов по таблице П 2 Приложения. Результаты вычислений занесем в таблицу 8. Заметим, что $f_{\max}(x) = f(\bar{x}) = \frac{1}{S_0 \sqrt{2\pi}} = 0,1761$.

Таблица 8

i	z_i	$t_i = \frac{z_i - \bar{x}}{S_0}$	$f_0(t_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t_i^2}{2}}$	$f(z_i) = \frac{f_0(t_i)}{S_0}$
1	-0,5	-2,4258	0,0210	0,0093
2	0,9	-1,8079	0,0778	0,0344
3	2,3	-1,1900	0,1965	0,0867
4	3,7	-0,5721	0,3387	0,1495
5	5,1	0,0459	0,3985	0,1759
6	6,5	0,6638	0,3201	0,1413
7	7,9	1,2817	0,1755	0,0774
8	9,3	1,8996	0,0657	0,0290
	$\bar{x} = 4,9961$	0,0000	0,3989	0,1761

Последний столбец таблицы 8 можно вычислить сразу по серединам интервалов z_i с помощью статистической функции НОРМРАСП пакета EXCEL с логическим значением ЛОЖЬ.

Для построения теоретической плотности вероятностей на рисунке 3 поставим точки $(z_i; f(z_i))$, $i = \overline{1, N}$ и $(\bar{x}, f(\bar{x}))$ и соединим их плавной линией. Из рисунка 3 видно, что график теоретической плотности вероятностей и гистограмма достаточно хорошо совпадают.

1.5. ЭМПИРИЧЕСКАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

Пусть $X_n = \{x_1, x_2, \dots, x_k\}$ – выборка объема n , содержащая k различных вариантов, из генеральной совокупности случайной величины X , имеющая функцию распределения $F(x)$, $x \in \mathfrak{R}$.

Неизвестную функцию распределения генеральной совокупности $F(x)$ называют *теоретической функцией распределения*.

Эмпирической функцией распределения группированной выборки X_n называется функция $F_n(x)$, определяющая для любого $x \in \mathfrak{R}$ относительную частоту события $(X < x)$, то есть $F_n(x) = \sum_{z_i < x} \bar{n}_i$, где $z_i = \frac{u_i + u_{i+1}}{2}$ –

середина интервалов группировки; \bar{n}_i – относительные частоты тех интервалов, середины которых меньше x .

По определению $F_n(x)$ зависит от выборки и обладает свойствами функции распределения случайной величины. В частности $F_n(x)$:

По определению $F_n(x)$ зависит от выборки и обладает свойствами функции распределения случайной величины. В частности $F_n(x)$:

1. неубывающая функция;
2. непрерывная слева;
3. имеет значения, принадлежащие отрезку $[0, 1]$;
4. при $x \leq z_1$ $F_n(x) = 0$, а при $x > z_N$ $F_n(x) = 1$.

Различие между эмпирической и теоретической функциями состоит в том, что теоретическая функция $F(x)$ определяет вероятность события $(X < x)$, а эмпирическая функция $F_n(x)$ определяет относительную частоту этого же события, найденную по данной выборке.

Значение эмпирической функции распределения для статистики определяется следующим утверждением.

Теорема (Гливенко): Пусть $F_n(x)$ – эмпирическая функция распределения, построенная по выборке объема n из генеральной совокупности с функцией распределения $F(x)$. Тогда для любого $x \in \mathfrak{R}$ и $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|F_n(x) - F(x)| < \varepsilon) = 1.$$

Таким образом, при каждом x $F_n(x)$ сходится по вероятности к $F(x)$ и, следовательно, при большом объеме выборки может служить приближенным значением (оценкой) функции распределения генеральной совокупности в каждой точке x .

Обычно эмпирическую функцию распределения $F_n(x)$ группированной выборки записывают в виде:

$$F_n(x) = \begin{cases} 0, & x \leq z_1, \\ w_1, & z_1 < x \leq z_2, \\ w_2, & z_2 < x \leq z_3, \\ \dots & \\ w_N = 1, & z_N < x. \end{cases}$$

где w_i ($i = \overline{1, N}$) – накопленные относительные частоты (таблица 4).

График эмпирической функции распределения $F_n(x)$ имеет ступенчатый вид (рис. 5).

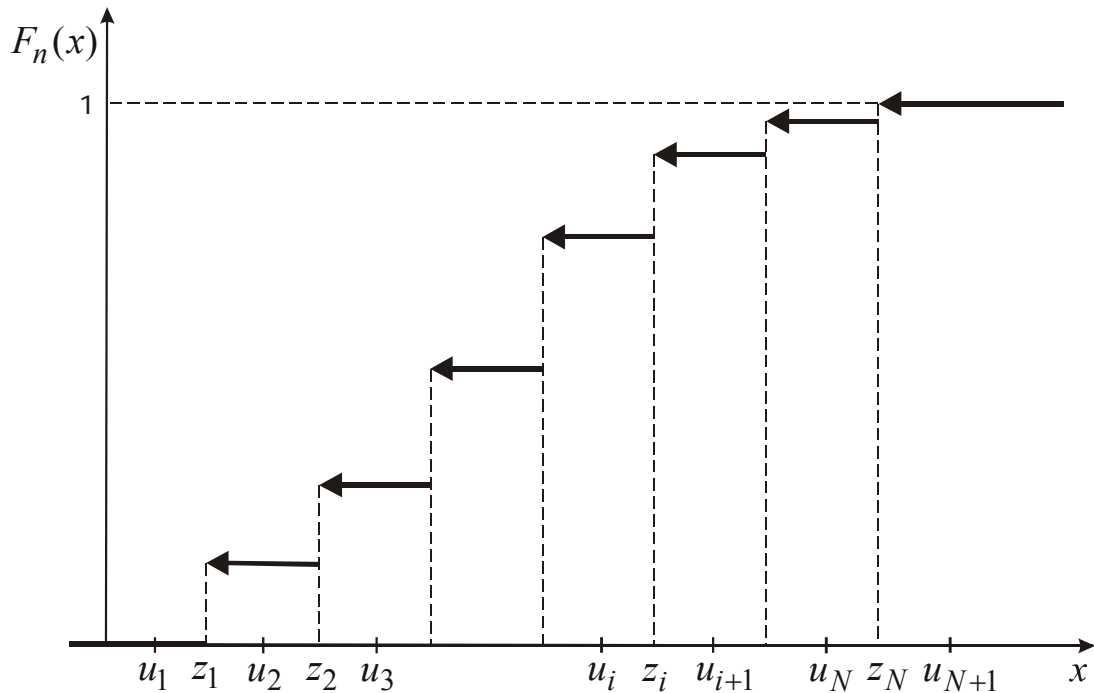


Рис. 5
Эмпирическая функция распределения

П Р И М Е Р 4 (пункты 6 и 7 части 1 Задания):

Требуется составить эмпирическую функцию распределения $F_n(x)$ группированной выборки (таблица 6) и построить ее график. На одном чертеже с эмпирической функцией распределения построить график теоретической функции распределения. Сделать вывод об их визуальном совпадении.

Взяв значения накопленных относительных частот из таблицы 6, а значения середин интервалов из таблицы 7, составим эмпирическую функцию распределения и построим ее график (рис. 6):

$$F_n(x) = \begin{cases} 0,0000, & x \leq -0,5, \\ 0,0202, & -0,5 < x \leq 0,9, \\ 0,0505, & 0,9 < x \leq 2,3, \\ 0,2323, & 2,3 < x \leq 3,7, \\ 0,4343, & 3,7 < x \leq 5,1, \\ 0,6768, & 5,1 < x \leq 6,5, \\ 0,8384, & 6,5 < x \leq 7,9, \\ 0,9293, & 7,9 < x \leq 9,3, \\ 1,0000, & x > 9,3. \end{cases}$$

Согласно выдвинутой гипотезе о виде распределения генеральной совокупности, теоретическая функция распределения генеральной совокупности является функцией распределения нормального закона:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt = \Phi_0\left(\frac{x-m}{\sigma}\right) + \frac{1}{2},$$

где $\Phi_0(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{u^2}{2}} du$ – функция Лапласа. Здесь, как и ранее,

$$m = \bar{x} = 4,9961, \sigma = S_0 = 2,2657.$$

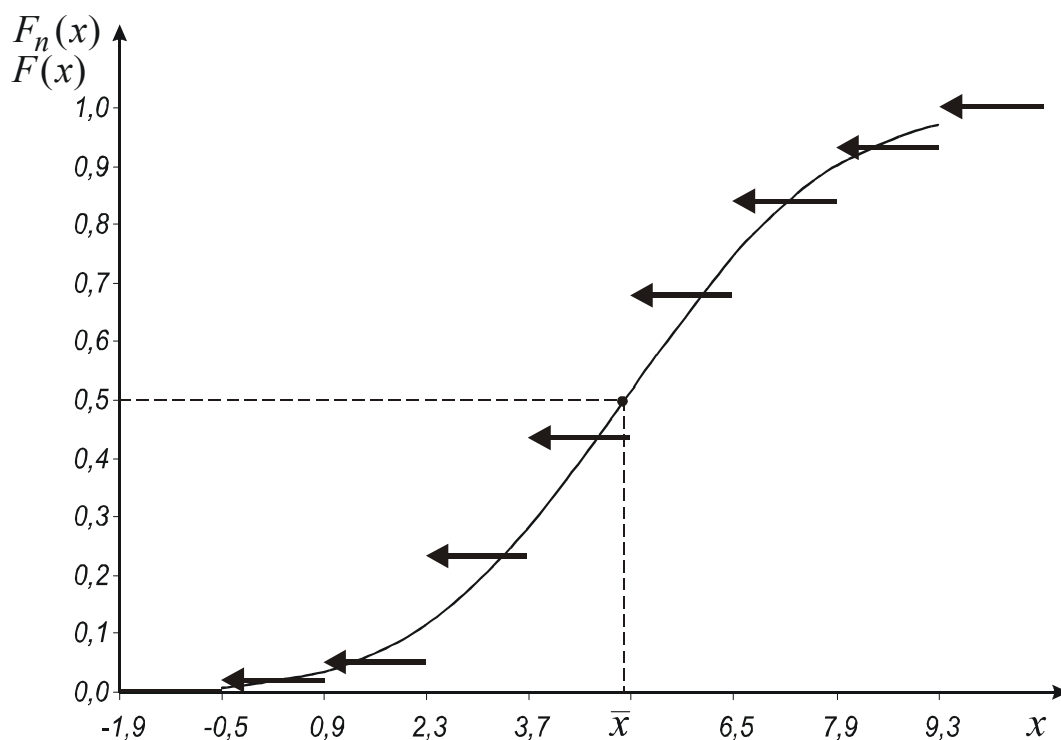


Рис. 6
Эмпирическая и теоретическая функции распределения

На одном чертеже с эмпирической функцией распределения построим график теоретической функции распределения. Для этого найдем значения теоретической функции распределения в точках z_i . Для удобства вычислений значений теоретической функции распределения заполним таблицу 9.

Значения функции Лапласа $\Phi_0(t_i)$, по которой вычисляются значения функции распределения $F(z_i)$, приведены в таблице П 1 Приложения.

Значения функции распределения $F(z_i)$ в точках z_i можно также найти с помощью пакета прикладных программ EXCEL, используя статистическую функцию НОРМРАСП с логическим значением ИСТИНА.

Таблица 9

i	z_i	$t_i = \frac{z_i - \bar{x}}{S_0}$	$\Phi_0(t_i)$	$F(z_i) = \Phi_0(t_i) + \frac{1}{2}$
1	-0,50	-2,4258	-0,4924	0,0076
2	0,90	-1,8079	-0,4649	0,0351
3	2,30	-1,1900	-0,3830	0,1170
4	3,70	-0,5720	-0,2157	0,2843
5	5,10	0,0459	0,0160	0,5160
6	6,50	0,6638	0,2454	0,7454
7	7,90	1,2817	0,3997	0,8997
8	9,30	1,8997	0,4713	0,9713
	$\bar{x} = 4,9961$	0,0000	0,0000	0,5000

График теоретической функции распределения строим на рисунке 6 по второму и пятому столбцам таблицы 9, соединив точки плавной линией. Заметим, что точка перегиба кривой теоретической функции распределения имеет координаты $(\bar{x}; 0,5)$.

Сравнивая графики $F_n(x)$ и $F(x)$ (рис. 6) можно сделать вывод, что $F_n(x)$ является статистическим аналогом $F(x)$.

1.6. РАСПРЕДЕЛЕНИЯ χ^2 И СТЬЮДЕНТА

Рассмотрим некоторые виды специальных распределений, используемых в математической статистике. Сначала введем определение:

Квантилью, соответствующей вероятности p , называется такое значение x_p , при котором выполняется соотношение:

$$P(X < x_p) = \int_{-\infty}^{x_p} f(x) dx = p,$$

где $f(x)$ – плотность вероятностей соответствующего закона распределения (слово квантиль – женского рода). Геометрическое пояснение смысла квантили, отвечающей вероятности p , приведено на рисунке 8.

РАСПРЕДЕЛЕНИЕ χ^2

Пусть X_1, X_2, \dots, X_k – нормально распределенные независимые случайные величины, причем математическое ожидание каждой из них равно нулю, а среднеквадратическое отклонение – единице, то есть $X_i \sim N(0,1)$.

Тогда сумма квадратов этих величин $\chi^2(k) = X_1^2 + X_2^2 + \dots + X_k^2 = \sum_{i=1}^k X_i^2$

распределена по закону χ^2 («хи квадрат») с k степенями свободы.

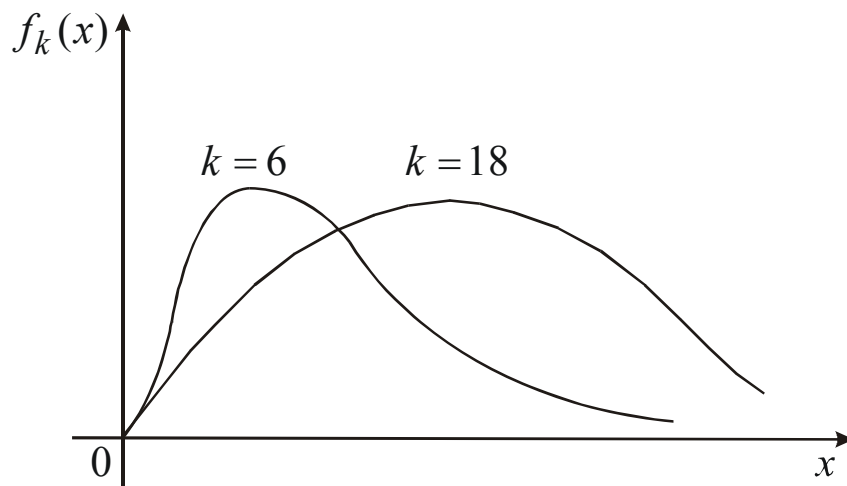


Рис. 7

Графики плотности вероятностей распределения χ^2

Плотность вероятностей этого распределения имеет вид:

$$f_k(x) = \frac{1}{2^{k/2} \Gamma(k/2)} \cdot x^{\frac{k}{2}-1} \cdot e^{-\frac{x}{2}}, \quad (x > 0),$$

где $\Gamma(k/2) = \int_0^{\infty} e^{-t} \cdot t^{k/2-1} dt$ - гамма- функция.

График плотности вероятностей $f_k(x)$ при малых k имеет длинный правый «хвост», а с ростом k становится почти симметричным (рис. 7).

Квантили распределения χ^2 обозначаются $x_p = \chi_p^2(k)$ (рис. 8) и находятся по таблицам (таблица П 5 Приложения).

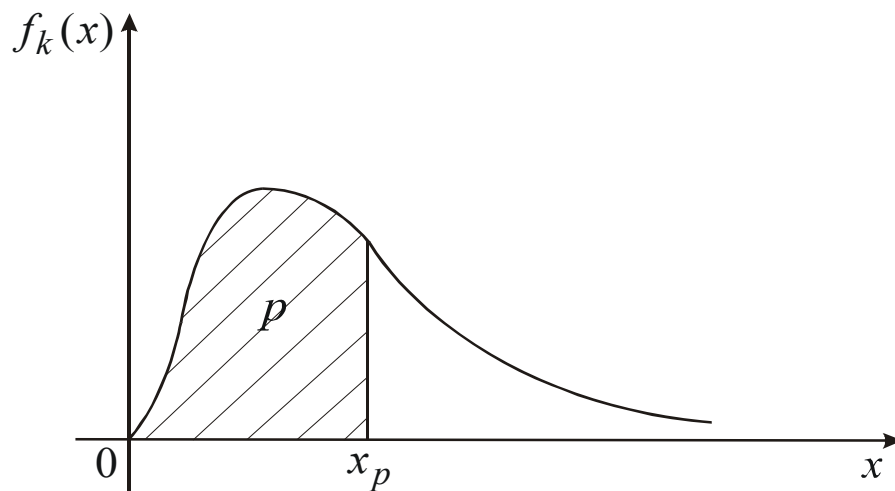


Рис. 8

Геометрическое пояснение смысла квантили x_p , отвечающей вероятности p

РАСПРЕДЕЛЕНИЕ СТЬЮДЕНТА

Пусть U – нормально распределенная случайная величина, причем $U \sim N(0,1)$, а V – независимая от U случайная величина, распределенная по закону χ^2 с k степенями свободы. Тогда известно, что случайная величина

$T = \frac{U\sqrt{k}}{\sqrt{V}}$ имеет t -распределение или распределение Стьюдента с k

степенями свободы. Плотность вероятностей этого распределения имеет вид:

$$f_k(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma(k/2)\sqrt{\pi k}} \cdot \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} \quad (\text{рис. 9}).$$

При $k \rightarrow \infty$ распределение Стьюдента стремится к нормальному и при $k \geq 30$ практически не отличается от нормального $N(0,1)$.

Квантили распределения Стьюдента t_p находят по таблицам (таблица П 4 Приложения) в зависимости от вероятности p и числа степеней свободы k . Так как график плотности вероятностей распределения Стьюдента симметричен относительно $t = 0$, то $t_p = -t_{1-p}$ (рис. 9).

Квантили распределений Стьюдента и χ^2 можно найти с помощью статистических функций СТЬЮДРАСПОБР и ХИ2ОБР пакета прикладных программ EXCEL.

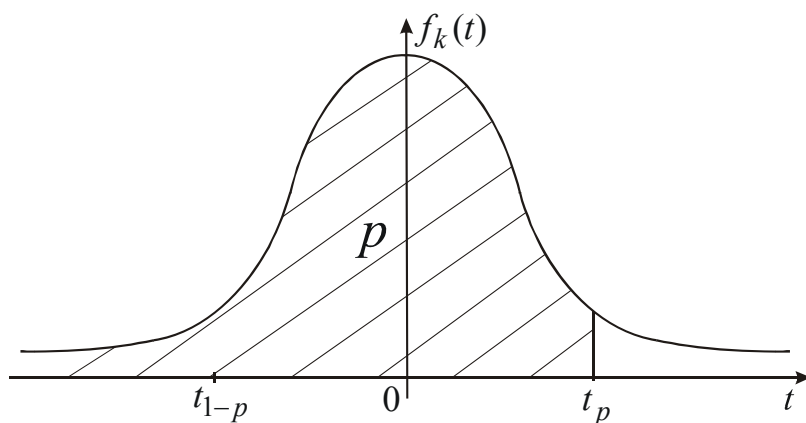


Рис. 9

Плотность вероятностей и квантили распределения Стьюдента

1.7. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Для получения обоснованных выводов о параметрах, виде распределения и других свойствах случайных величин необходимо проверить гипотезу о соответствии эмпирической функции распределения одному из известных теоретических законов.

Статистической гипотезой называют любое утверждение о виде или о параметрах распределения генеральной совокупности. Например, статистическими являются гипотезы:

1. генеральная совокупность распределена по нормальному закону или любому другому конкретно заданному закону (гипотеза о виде распределения);

2. если известно, что генеральная совокупность распределена по нормальному закону, то параметры нормального закона равны выборочным характеристикам: $m = \tilde{m}_X = \bar{x}$, $\sigma = \tilde{\sigma}_X = S_0$ (параметрическая гипотеза).

Гипотезу о виде распределения выдвигают на основе схожести гистограммы или полигона частот с соответствующей кривой одного из теоретических законов (нормального, равномерного, Пуассона и т. п.).

Когда предположение о виде распределения генеральной совокупности принято, следует проверить гипотезу о параметрах этого распределения.

Нулевой (основной) называют выдвинутую гипотезу H_0 .

Альтернативными называют гипотезы, которые противоречат нулевой. Если отвергается H_0 , то принимается одна из альтернативных гипотез. При проверке статистических гипотез могут быть допущены ошибки двух родов с вероятностями:

1. α – вероятность отклонить гипотезу H_0 , при условии, что она верна (ошибка первого рода);
2. β – вероятность принять гипотезу H_0 , при условии, что она неверна (ошибка второго рода).

Например, в радиолокации α – вероятность пропуска сигнала, β – вероятность ложной тревоги.

Ясно, что чем меньше будут ошибки первого и второго рода, тем точнее статистический вывод. Однако при заданном объеме выборке одновременно уменьшить α и β невозможно. Единственный способ одновременного уменьшения α и β состоит в увеличении объема выборки.

Если формулируется только одна гипотеза H_0 и требуется проверить, согласуются ли статистические данные с этой гипотезой или они ее опровергают, то критерии, используемые для этого, называют **критериями согласия**. В таких критериях не выставляется конкретная альтернативная гипотеза.

Прежде, чем привести схему статистической проверки гипотез, дадим используемые ниже определения новых понятий.

Статистикой критерия называется специально подобранная функция выборки $K = K(x_1, x_2, \dots, x_n)$, которая служит для проверки гипотезы H_0 . Статистика K является мерой расхождения экспериментальных данных с гипотетическим распределением.

Как правило, перед анализом выборки задается **уровень значимости** α – вероятность ошибочного отклонения нулевой гипотезы. Обычно полагают $\alpha = 0,05$, $\alpha = 0,01$, $\alpha = 0,001$.

Критической областью называется совокупность значений статистики, при которых нулевая гипотеза отвергается. Обычно критическую область выбирают из условия $P(K > K_{кр}) = \alpha$. Критическую точку критерия $K_{кр}$ находят по соответствующим таблицам.

Схема статистической проверки гипотезы по критерию согласия:

- 1) формулировка нулевой H_0 гипотезы;
- 2) выбор уровня значимости α ;
- 3) выбор статистики K и соответствующего критерия;
- 4) определение критической области и области принятия гипотезы;
- 5) вычисление выборочной статистики $K_{выб}$ и проверка гипотезы;
- 6) принятие статистического решения.

1.8. КРИТЕРИЙ СОГЛАСИЯ χ^2 ПИРСОНА

Для проверки гипотез о виде распределения применяются различные критерии согласия: χ^2 («хи- квадрат») К. Пирсона, критерий Колмогорова, критерий Смирнова и др. Наиболее удобным и универсальным критерием является критерий χ^2 Пирсона. Он совершенно не зависит ни от вида распределения случайной величины, ни от ее размерности.

Ограничимся описанием применения критерия Пирсона для проверки гипотезы о нормальном распределении генеральной совокупности (критерий аналогично применяется и для других распределений).

Схема применения критерия согласия χ^2 :

1). Выдвигается гипотеза H_0 : генеральная совокупность имеет нормальное распределение с плотностью вероятностей :

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

с параметрами $m = \tilde{m}_X = \bar{x}$, $\sigma = \tilde{\sigma}_X = S_0$, то есть выборочное среднее \bar{x} и модифицированная выборочная дисперсия S_0^2 принимаются соответственно за математическое ожидание m и дисперсию σ^2 нормально распределенной случайной величины.

2). По выборке наблюдений случайной величины X составляется группированный вариационный ряд (таблица 4).

3). Вычисляются вероятности p_i ($i = \overline{1, N}$) попадания значений случайной величины X в i -тый интервал.

Для нормального закона

$$p_i = P(u_i < X \leq u_{i+1}) = \Phi\left(\frac{u_{i+1} - m}{\sigma}\right) - \Phi\left(\frac{u_i - m}{\sigma}\right).$$

Здесь $\Phi(x)$ – функция распределения нормального закона $N(0;1)$, значения которой находят по таблицам.

4). Вычисляется выборочное значение статистики критерия χ^2 :

$$\chi_{выб}^2 = \sum_{i=1}^N \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i},$$

где N – число интервалов разбиения выборки; n – объем выборки; n_i – частота i -того интервала; p_i – теоретическая вероятность попадания значений случайной величины X в i -тый интервал.

К. Пирсон доказал, что эта статистика независимо от вида распределения генеральной совокупности при $n \rightarrow \infty$ имеет χ^2 - распределение с $q = N - s - 1$ степенями свободы, где N – число интервалов разбиения, s – число оцениваемых параметров гипотетического закона распределения. Для нормального закона $s = 2$ (параметры m и σ).

5). Областью отклонения G (критической областью) гипотезы H_0 называется такая область, при попадании в которую статистики $\chi_{выб}^2$ гипотеза H_0 отклоняется. Область отклонения G выбирается так, чтобы вероятность попадания в нее величины $\chi_{выб}^2$, когда гипотеза H_0 верна, была равна уровню значимости α . Тогда критическая точка $\chi_{кр}^2$, ограничивающая область G , определяется из уравнения:

$$P\left(\chi_{выб}^2 \in G/H_0\right) = P\left(K > \chi_{кр}^2\right) = \int_{\chi_{кр}^2}^{+\infty} f_k(x) dx = \alpha.$$

Из этой формулы следует, что критическая точка $\chi_{кр}^2$ равна с квантили распределения Пирсона χ_p^2 , отвечающей вероятности $p = 1 - \alpha$ с числом степеней свободы $q = N - s - 1$ (таблица П 5 Приложения).

Таким образом, если вычисленная выборочная статистика $\chi_{выб}^2 \in [0; \chi_{кр}^2)$, то гипотеза H_0 принимается. Если $\chi_{выб}^2 \geq \chi_{кр}^2$, то гипотеза H_0 отвергается.

Область принятия критерия имеет вид, представленный на рис. 10.

Выбор области принятия гипотезы можно объяснить следующим образом: значения теоретических вероятностей p_i и относительных частот интервалов $\bar{n}_i = \frac{n_i}{n}$ должны быть достаточно близки, поэтому разности $(n_i - n \cdot p_i)$ не должны быть слишком велики.

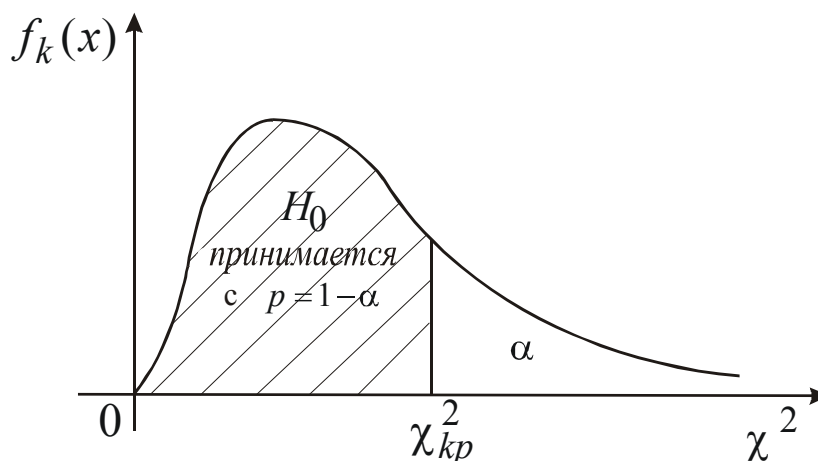


Рис. 10
Область принятия критерия

Статистический вывод неверно формулировать в виде: генеральная совокупность имеет нормальный закон распределения. Можно лишь утверждать, что данная выборка **согласуется** с гипотезой о нормальном распределении генеральной совокупности с параметрами $m = \tilde{m}_X = \bar{x}$, $\sigma = \tilde{\sigma}_X = S_0$ на уровне значимости α .

З а м е ч а н и е: критерий χ^2 использует тот факт, что случайная величина $\frac{n_i - n \cdot p_i}{\sqrt{n \cdot p_i}}$ имеет распределение, близкое к нормальному. Чтобы это утверждение было достаточно точным, необходимо выполнение условия $n \cdot p_i \geq 5$ для всех интервалов. Интервалы, для которых это условие не выполняется, следует объединить с соседними.

ПРИМЕР 5 (пункт 8 части 1 Задания):

Требуется для выборки (таблица 1) с помощью критерия согласия Пирсона χ^2 проверить гипотезу H_0 о виде распределения генеральной совокупности (нормальное распределение) на уровне значимости $\alpha = 0,1$. Сделать статистический вывод.

Для данной выборки объема $n = 99$ ранее были вычислены выборочное среднее $\bar{x} = 4,9961$ и модифицированная выборочная дисперсия $S_0^2 = 5,1332$, составлен группированный вариационный ряд (таблица 6), а также выдвинута гипотеза H_0 о нормальном распределении генеральной совокупности.

Вычислим теперь вероятности p_i ($i = \overline{1, N}$) попадания значений случайной величины X в i -тый интервал и выборочное значение статистики

критерия χ^2 :
$$\chi_{выб}^2 = \sum_{i=1}^N \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}.$$

Результаты вычислений занесем в таблицу 10.

Таблица 10

i	u_i	n_i	$t_i = \frac{u_i - \bar{x}}{S_0}$	$\Phi_0(t_i)$	p_i	$n \cdot p_i$
1	-1,2		-2,7348	-0,4968		
2	0,20	2	-2,1169	-0,4830	0,0138	1,3662
3	1,60	3	-1,4989	-0,4332	0,0498	4,9302
4	3,00	18	-0,8810	-0,3106	0,1226	12,1374
5	4,40	20	-0,2631	-0,1026	0,2080	20,5920
6	5,80	24	0,3548	0,1368	0,2394	23,7006
7	7,20	16	0,9728	0,3340	0,1972	19,5228
8	8,60	9	1,5907	0,4441	0,1101	10,8999
9	10,00	7	2,2086	0,4864	0,0423	4,1877

Пятый столбец таблицы 10 можно вычислить с помощью статистической функции НОРМРАСП пакета EXCEL. Шестой столбец представляет собой разности значений из пятого столбца: $p_i = \Phi_0(t_{i+1}) - \Phi_0(t_i)$.

Заметим, что в таблице 9 вычислялись значения функции Лапласа в серединах интервалов, а в таблице 10 для проверки критерия χ^2 – именно в концах интервалов разбиения.

Так как в двух первых и в последнем интервалах не выполняется условие $n \cdot p_i \geq 5$, то объединим эти интервалы с соседними. При объединении интервалов значения n_i и $n \cdot p_i$ суммируются (таблица 11).

Таблица 11

i	n_i	$n \cdot p_i$	$\chi_i^2 = \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$
1	5	6,2964	0,2669
2	18	12,1374	2,8317
3	20	20,5920	0,0170
4	24	23,7006	0,0038
5	16	19,5228	0,6357
6	16	15,0876	0,0552
Сумма	99		3,8103

Суммируя элементы последнего столбца таблицы 11, получим $\chi_{выб}^2 = 3,8103$. Число степеней свободы после укрупнения таблицы 10 равно $q = N - s - 1 = 6 - 2 - 1 = 3$ ($N = 6$, так как в укрупненной таблице 6 интервалов). **Область принятия** гипотезы можно записать в виде

$$P(K \leq \chi_{кр}^2) = \int_0^{\chi_{кр}^2} f_k(x) dx = 1 - \alpha = p,$$

откуда следует, что критическое значение $\chi_{кр}^2$ совпадает с квантилем $\chi_p^2(k)$ распределения хи-квадрат с доверительной вероятностью $p = 1 - \alpha$.

В нашем случае $\alpha = 0,1$ и $p = 1 - \alpha = 0,9$, число степеней свободы $q = 3$. По таблице П 5 Приложения (или функции ХИ2ОБР) находим значение критической точки распределения (квантили) $\chi_{кр}^2 = \chi_{0,9}^2(3) = 6,251$. Так как

$\chi_{выб}^2 = 3,8103 \in [0 ; 6,251)$, то на данном уровне значимости гипотеза H_0 принимается.

Статистический вывод: данная выборка согласуется с гипотезой о нормальном распределении с параметрами $m = \bar{x} = 4,9961$, $\sigma = S_0 = 2,2657$ на уровне значимости $\alpha = 0,1$, то есть вероятность отвергнуть гипотезу H_0 , при условии, что она верна, равна $0,1$.

1.9. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

Интервальное оценивание параметров распределения генеральной совокупности состоит в построении доверительных интервалов.

Доверительным интервалом для параметра θ называется интервал (θ_1, θ_2) , содержащий истинное значение параметра с заданной вероятностью $p = 1 - \alpha$. Таким образом, $P(\theta_1 < \theta < \theta_2) = 1 - \alpha$. Число $p = 1 - \alpha$ называется *доверительной вероятностью*, а значение α – *уровнем значимости*.

При построении доверительных интервалов вводят в рассмотрение специально подобранную статистику K , распределение которой известно. Наиболее распространенными являются статистики, имеющие нормальное, Стьюдента и χ^2 распределения.

Методика построения доверительных интервалов для отдельных параметров распределения генеральной совокупности зависит как от вида распределения, так и от знания значений остальных параметров закона распределения.

1.9.1. Рассмотрим задачу построения **доверительного интервала для математического ожидания m** нормально распределенной генеральной совокупности при неизвестной дисперсии.

Пусть случайная величина X имеет нормальное распределение с параметрами m и σ . Найдем доверительный интервал для математического ожидания m в предположении, что дисперсия σ^2 неизвестна и задан уровень значимости α .

Английский математик Госсет (псевдоним Стьюдент) доказал, что статистика $T = \frac{\bar{x} - m}{S_0} \sqrt{n}$ имеет распределение Стьюдента с $k = n - 1$

степенями свободы. Так как кривая плотности вероятностей распределения Стьюдента симметрична относительно $t = 0$, будем искать доверительную

область в виде:
$$P(|T| < \tau) = P(-\tau < T < \tau) = \int_{-\tau}^{\tau} f_k(t) dt = p = 1 - \alpha.$$

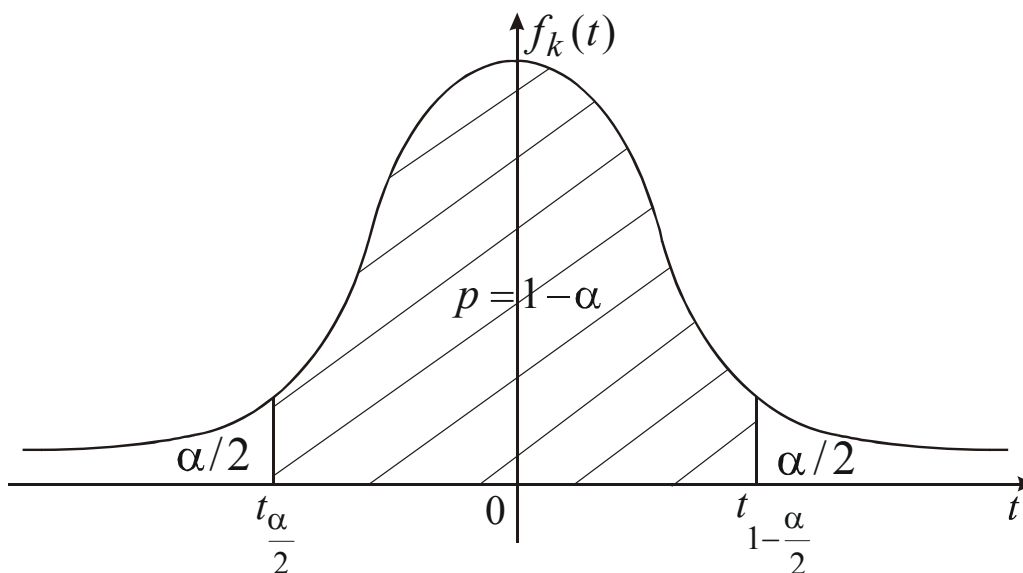


Рис. 11
Геометрическое пояснение смысла квантилей
распределения Стьюдента

Из рисунка 11 видно, что площадь под графиком каждого из симметричных «хвостов» будет равна $\frac{\alpha}{2}$, тогда значения границ интервала

совпадут с квантилями $\tau = t_{1-\frac{\alpha}{2}}$ и $-\tau = t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$.

В таблице П 4 Приложения приведены значения $t_p(k)$ в зависимости от доверительной вероятности p и числа степеней свободы $k = n - 1$. Можно также использовать функцию СТЬЮДРАСПОБР пакета прикладных программ EXCEL.

Таким образом, получаем: $t_{\frac{\alpha}{2}} < T < t_{1-\frac{\alpha}{2}}$ или

$$-t_{1-\frac{\alpha}{2}} < \frac{\bar{x} - m}{S_0} \sqrt{n} < t_{1-\frac{\alpha}{2}}.$$

Подставив в полученное неравенство значения $t_{1-\frac{\alpha}{2}}$, \bar{x} , S_0 , n и разрешив это неравенство относительно m , получим доверительный интервал для неизвестного математического ожидания m нормально распределенной случайной величины X с неизвестной дисперсией σ^2 и заданным уровнем

значимости α :

$$\bar{x} - \frac{S_0}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} < m < \bar{x} + \frac{S_0}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}.$$

П Р И М Е Р 6 (пункт 9 части 1 Задания):

Требуется построить доверительный интервал для математического ожидания нормально распределенной генеральной совокупности с параметрами $m = \bar{x}$ и $\sigma = S_0$ для уровней значимости $\alpha = 0,1$, $\alpha = 0,05$ и $\alpha = 0,01$ при неизвестной дисперсии.

При построении доверительного интервала для неизвестного математического ожидания m нормально распределенной генеральной совокупности используется статистика $T = \frac{\bar{x} - m}{S_0} \sqrt{n}$, имеющая распределение

Стьюдента с $k = n - 1$ степенями свободы. Общее уравнение доверительного интервала $P(T \in (\theta_1; \theta_2)) = 1 - \alpha$ в данном случае имеет вид:

$$t_{\frac{\alpha}{2}} < \frac{\bar{x} - m}{S_0} \sqrt{n} < t_{1-\frac{\alpha}{2}}.$$

Вычислим этот интервал для различных уровней значимости.

$$\underline{\alpha = 0,1}: \quad \frac{\alpha}{2} = 0,05, \quad 1 - \frac{\alpha}{2} = 0,95,$$

$k = n - 1 = 99 - 1 = 98$ – число степеней свободы.

Так как в таблице П 4 Приложения нет числа степеней свободы $k = 98$, то для вычисления $t_p(k)$ можно воспользоваться одним из трех методов.

1. Известно, что при $k \geq 30$ $t_p(k) \approx C_p$, где C_p – квантиль нормального распределения (таблица П 3 Приложения). Тогда

$$t_{1-\frac{\alpha}{2}}(98) = t_{0,95}(98) \approx C_{0,95} = 1,645,$$

$$t_{\frac{\alpha}{2}}(98) = t_{0,05}(98) \approx -C_{0,95} = -1,645.$$

2. Можно использовать линейную интерполяцию между точками таблицы П 4 Приложения $(k_1; t_1) = (60; 1,671)$ и $(k_2; t_2) = (120; 1,658)$. Значение квантили при $k = 98$ найдем по формуле линейной интерполяции:

$$t_{0,95}(98) = \frac{t_2 - t_1}{k_2 - k_1} \cdot (k - k_1) + t_1 = \frac{1,658 - 1,671}{120 - 60} \cdot (98 - 60) = 1,663.$$

Тогда $t_{\frac{\alpha}{2}}(98) = t_{0,05}(98) = -1,663$.

3. Статистическая функция СТЬЮДРАСПОБР пакета EXCEL дает значение квантили $t_{0,95}(98) = 1,661$. Нужно иметь в виду, что в EXCEL вычисляются значения двусторонних «антиквантилей» $P(|T| > x_p) = p$. Поэтому чтобы получить значение односторонней квантили $t_{0,95}(98)$, нужно в этой функции задать вероятность $p = 2 \cdot (1 - 0,95) = 0,1$ (см. справку к функции СТЬЮДРАСПОБР).

В дальнейших расчетах используем значения, даваемые EXCEL.

$$t_{1-\frac{\alpha}{2}}(98) = t_{0,95}(98) = 1,661, \quad t_{\frac{\alpha}{2}}(98) = t_{0,05}(98) = -1,661,$$

$$-1,661 < \frac{4,9961 - m}{2,2657} \sqrt{99} < 1,661$$

Выражая из неравенства неизвестный параметр m , получим доверительный интервал для математического ожидания для уровня значимости $\alpha = 0,1$:

$$4,6178 < m < 5,3743.$$

Таким образом, неизвестное математическое ожидание $m \in (4,6178 ; 5,3743)$ с вероятностью $p = 0,9$.

Аналогично найдем доверительные интервалы для математического ожидания для уровней значимости $\alpha = 0,05$ и $\alpha = 0,01$.

$$\underline{\alpha = 0,05}: \quad \frac{\alpha}{2} = 0,025, \quad 1 - \frac{\alpha}{2} = 0,975, \quad k = n - 1 = 99 - 1 = 98,$$

$$t_{1-\frac{\alpha}{2}}(98) = t_{0,975}(98) = 1,984, \quad t_{\frac{\alpha}{2}}(98) = t_{0,025}(98) = -1,984,$$

$$-1,984 < \frac{4,9961 - m}{2,2657} \sqrt{99} < 1,984.$$

Выражая из неравенства неизвестный параметр m , получим доверительный интервал для математического ожидания для уровня значимости $\alpha = 0,5$:

$$4,5442 < m < 5,4479.$$

Таким образом, неизвестное математическое ожидание $m \in (4,5442 ; 5,4479)$ с вероятностью $p = 0,95$.

$$\underline{\alpha = 0,01}: \quad \frac{\alpha}{2} = 0,005, \quad 1 - \frac{\alpha}{2} = 0,995, \quad k = n - 1 = 99 - 1 = 98,$$

$$t_{1-\frac{\alpha}{2}}(98) = t_{0,995}(98) = 2,627, \quad t_{\frac{\alpha}{2}}(98) = t_{0,005}(98) = -2,627,$$

$$-2,627 < \frac{4,9961 - m}{2,2657} \sqrt{99} < 2,627.$$

Выражая из неравенства неизвестный параметр m , получим доверительный интервал для математического ожидания для уровня значимости $\alpha = 0,01$:

$$4,3978 < m < 5,5948.$$

Таким образом, неизвестное математическое ожидание $m \in (4,3978 ; 5,5948)$ с вероятностью $p = 0,99$.

1.9.2. Определим теперь доверительный интервал для неизвестной дисперсии σ^2 нормально распределенной случайной величины X с неизвестным математическим ожиданием и заданным уровнем значимости α .

В этом случае рассматривается статистика $B = \frac{S_0^2}{\sigma^2}(n-1)$, имеющая распределение χ^2 с $k = n-1$ степенями свободы, где n – объем выборки.

Будем искать доверительную область в виде:

$$P(\theta_1 < \chi_p^2 < \theta_2) = \int_{\theta_1}^{\theta_2} f_k(x) dx = p = 1 - \alpha.$$

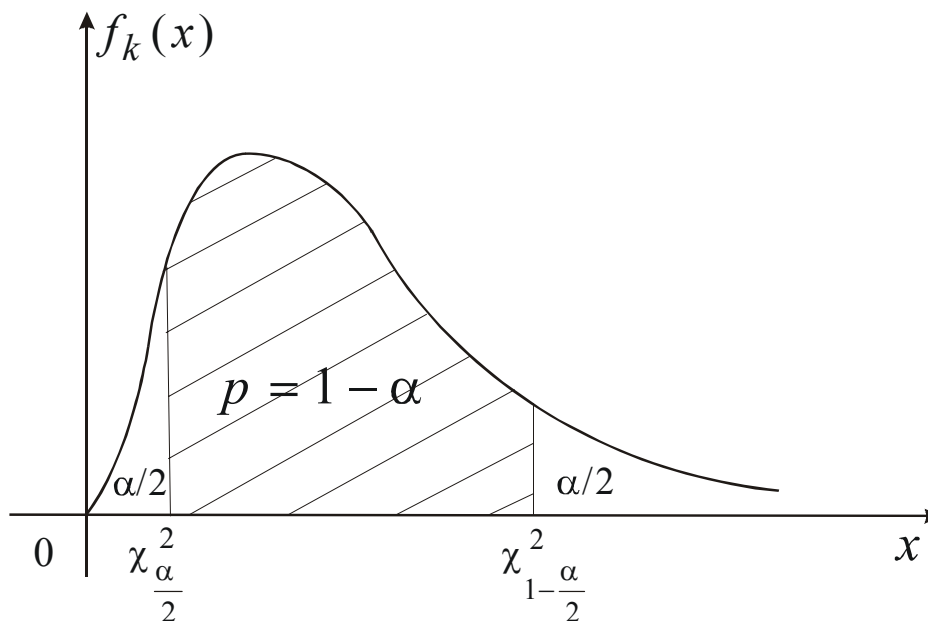


Рис. 12
Квантили распределения χ^2

Как и в предыдущем случае, будем считать площади под «хвостами» кривой распределения равными по $\frac{\alpha}{2}$ каждая (рис. 12).

Тогда границы интервала совпадут с квантилями:

$$\theta_1 = \chi_{\frac{\alpha}{2}}^2(n-1), \quad \theta_2 = \chi_{1-\frac{\alpha}{2}}^2(n-1).$$

В таблице П 5 Приложения приведены значения $\chi_p^2(k)$ в зависимости от доверительной вероятности p и числа степеней свободы k . Можно также использовать функцию ХИ2ОБР пакета прикладных программ EXCEL.

Таким образом, получаем

$$\chi_{\frac{\alpha}{2}}^2(n-1) < \frac{S_0^2}{\sigma^2}(n-1) < \chi_{1-\frac{\alpha}{2}}^2(n-1).$$

Подставив в полученное неравенство значения $\chi_{\frac{\alpha}{2}}^2$, $\chi_{1-\frac{\alpha}{2}}^2$, S_0^2 , n и

разрешив это неравенство относительно σ^2 , получим доверительный интервал для неизвестной дисперсии σ^2 нормально распределенной случайной величины X с неизвестным математическим ожиданием и заданным уровнем значимости α :

$$\frac{(n-1) \cdot S_0^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)} < \sigma^2 < \frac{(n-1) \cdot S_0^2}{\chi_{\frac{\alpha}{2}}^2(n-1)}.$$

Следует отметить, что если математическое ожидание генеральной совокупности известно, то доверительный интервал для дисперсии будет иметь другой вид.

Длина доверительного интервала характеризует точность оценивания и зависит от объема выборки n и доверительной вероятности $p = 1 - \alpha$. Чем меньше длина доверительного интервала, тем надежнее оценка. При увеличении объема выборки длина доверительного интервала уменьшается.

ПРИМЕР 7 (пункт 9 части 1 Задания):

Требуется построить доверительный интервал для неизвестной дисперсии нормально распределенной генеральной совокупности с параметрами $m = \bar{x}$ и $\sigma = S_0$ для уровней значимости $\alpha = 0,1$, $\alpha = 0,05$ и $\alpha = 0,01$.

Для построения доверительного интервала для неизвестной дисперсии σ^2 нормально распределенной генеральной совокупности используется

статистика $B = \frac{S_0^2}{\sigma^2} \cdot (n-1)$, имеющая распределение χ^2 с $k = n-1$

степенями свободы:

$$\chi_{\frac{\alpha}{2}}^2 < \frac{S_0^2}{\sigma^2} \cdot (n-1) < \chi_{1-\frac{\alpha}{2}}^2.$$

Вычислим этот интервал для различных уровней значимости.

$$\underline{\alpha = 0,1}: \quad \frac{\alpha}{2} = 0,05, \quad 1 - \frac{\alpha}{2} = 0,95, \quad k = n - 1 = 99 - 1 = 98.$$

Так как в таблице П 5 Приложения нет числа степеней свободы $k = 98$, то для вычисления $\chi_p^2(k)$ можно воспользоваться одним из способов:

1. Известно, что при $k \geq 30$ $\chi_p^2(k) \approx \frac{(C_p + \sqrt{2k-1})^2}{2}$, где C_p –

квантиль нормального распределения (таблица П 3 Приложения).

По этой формуле получим:

$$\chi_{\frac{\alpha}{2}}^2(98) = \chi_{0,05}^2(98) \approx \frac{(C_{0,05} + \sqrt{2 \cdot 98 - 1})^2}{2} = 75,882,$$

$$\chi_{1-\frac{\alpha}{2}}^2(98) = \chi_{0,95}^2(98) \approx \frac{(C_{0,95} + \sqrt{2 \cdot 98 - 1})^2}{2} = 121,824.$$

2. Статистическая функция ХИ2ОБР пакета EXCEL дает следующие значения квантилей распределения хи- квадрат:

$$\chi_{0,05}^2(98) = 76,164, \quad \chi_{0,95}^2(98) = 122,108.$$

Следует иметь в виду, что в функции ХИ2ОБР вычисляются «антиквантили» $P(B > \chi_p^2) = p$. Чтобы получить значение квантили $\chi_{0,05}^2(98)$, нужно ввести обратную вероятность $p = 0,95$.

В дальнейших расчетах используются значения квантилей, вычисленные

в EXCEL:
$$76,164 < \frac{5,1332}{\sigma^2} \cdot (99 - 1) < 122,108$$

Выражая из неравенства неизвестный параметр σ^2 , получим доверительный интервал для дисперсии для уровня значимости $\alpha = 0,1$:

$$4,1197 < \sigma^2 < 6,6049.$$

Таким образом, неизвестная дисперсия $\sigma^2 \in (4,1197 ; 6,6049)$ с вероятностью $p = 0,9$.

Аналогично найдем доверительные интервалы для дисперсии для уровней значимости $\alpha = 0,05$ и $\alpha = 0,01$.

$\alpha = 0,05$: $\frac{\alpha}{2} = 0,025$, $1 - \frac{\alpha}{2} = 0,975$, $k = n - 1 = 99 - 1 = 98$,

$$\chi_{\frac{\alpha}{2}}^2(98) = \chi_{0,025}^2(98) = 72,501, \quad \chi_{1-\frac{\alpha}{2}}^2(98) = \chi_{0,975}^2(98) = 127,282,$$

$$72,501 < \frac{5,1332}{\sigma^2} \cdot (99 - 1) < 127,282.$$

Выражая из неравенства неизвестный параметр σ^2 , получим доверительный интервал для дисперсии для уровня значимости $\alpha = 0,05$:

$$3,9523 < \sigma^2 < 6,9386.$$

Таким образом, неизвестная дисперсия $\sigma^2 \in (3,9523 ; 6,9386)$ с вероятностью $p = 0,95$.

$\alpha = 0,01$: $\frac{\alpha}{2} = 0,005$, $1 - \frac{\alpha}{2} = 0,995$, $k = n - 1 = 99 - 1 = 98$,

$$\chi_{\frac{\alpha}{2}}^2(98) = \chi_{0,005}^2(98) = 65,693, \quad \chi_{1-\frac{\alpha}{2}}^2(98) = \chi_{0,995}^2(98) = 137,803,$$

$$65,693 < \frac{5,1332}{\sigma^2} \cdot (99 - 1) < 137,803.$$

Выражая из неравенства неизвестный параметр σ^2 , получим доверительный интервал для дисперсии для уровня значимости $\alpha = 0,01$:

$$3,6505 < \sigma^2 < 7,6576.$$

Таким образом, неизвестная дисперсия $\sigma^2 \in (3,6505 ; 7,6576)$ с вероятностью $p = 0,99$.

Заметим, что полученные ранее выборочное среднее $\bar{x} = 4,9961$ и выборочная дисперсия $S_0^2 = 5,1332$ попадают во все найденные доверительные интервалы соответственно, причем, чем меньше уровень значимости α , то есть больше вероятность $p = 1 - \alpha$, тем больше длина соответствующего доверительного интервала.

2. СТАТИСТИЧЕСКИЙ АНАЛИЗ ДВУМЕРНЫХ ДАННЫХ

В практических применениях теории вероятностей очень часто приходится сталкиваться с задачами, в которых результат опыта описывается не одной случайной величиной, а двумя или более случайными величинами. Изучение каждой из этих случайных величин отдельно от другой может привести к недопустимому упрощению вероятностной модели явления. В данном разделе рассматриваются такие методы статистического анализа двумерных данных, как корреляционный и регрессионный анализ.

2.1. ФУНКЦИОНАЛЬНАЯ, СТАТИСТИЧЕСКАЯ И КОРРЕЛЯЦИОННАЯ ЗАВИСИМОСТИ

Рассмотрим зависимость случайной величины Y от одной величины X (случайной или детерминированной).

Если каждому возможному значению X соответствует единственное возможное значение Y , то Y называют *функцией* аргумента X . Строгая функциональная зависимость между двумя случайными величинами реализуется редко, так как обе величины могут быть подвержены воздействию случайных факторов.

Статистической называется зависимость, при которой изменение одной из величин влечет за собой изменение закона распределения другой.

Если статистическая зависимость проявляется в том, что при изменении одной величины изменяется среднее значение (математическое ожидание) другой, то зависимость называется *корреляционной*.

Между функциональной и корреляционной зависимостями случайных величин существует связь.

1. Если случайные величины X и Y функционально зависимы, то они коррелированы. Обратное утверждение, в общем, не верно.

2. Если X и Y независимы, то они некоррелированы. Обратное утверждение, в общем, не верно.

Таким образом, корреляционная зависимость занимает промежуточное значение между функциональной зависимостью и независимостью. Поэтому корреляционную зависимость считают «слабой» зависимостью между случайными величинами.

2.2. ЛИНЕЙНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Пусть имеется n наблюдений случайного вектора (X, Y) . При этом $X = \{x_1, x_2, \dots, x_n\}$ и $Y = \{y_1, y_2, \dots, y_n\}$.

По данным наблюдениям можно вычислить следующие статистики:

$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$ – выборочное среднее случайной величины X ;

$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^n y_i$ – выборочное среднее случайной величины Y ;

$\overline{x \cdot y} = \frac{1}{n} \cdot \sum_{i=1}^n x_i \cdot y_i$ – выборочное среднее произведений случайных величин;

$\tilde{D}_X = \tilde{\sigma}_X^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$ – выборочная дисперсия случайной величины X ;

$\tilde{D}_Y = \tilde{\sigma}_Y^2 = \frac{1}{n} \cdot \sum_{i=1}^n (y_i - \bar{y})^2$ – выборочная дисперсия случайной величины Y ;

$\tilde{K}_{XY} = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \overline{x \cdot y} - \bar{x} \cdot \bar{y}$ – выборочный корреляционный

момент.

Все эти статистики являются оценками соответствующих параметров генеральной совокупности, вычисленными по данной выборке наблюдений.

Выборочным коэффициентом корреляции \tilde{r}_{XY} называется отношение выборочного корреляционного момента \tilde{K}_{XY} к произведению выборочных среднеквадратических отклонений величин X и Y :

$$\tilde{r}_{XY} = \frac{\tilde{K}_{XY}}{\tilde{\sigma}_X \cdot \tilde{\sigma}_Y} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\tilde{\sigma}_X \cdot \tilde{\sigma}_Y}.$$

Выборочный коэффициент корреляции можно вычислить также с помощью функции КОРРЕЛ пакета EXCEL.

Если выборочный коэффициент корреляции равен нулю, то случайные величины называются **некоррелированными**. Из независимости случайных величин следует некоррелированность, то есть для независимых случайных величин коэффициент корреляции равен нулю. Обратное утверждение в общем

случае неверно, из коррелированности случайных величин не всегда следует их независимость.

Выборочный коэффициент корреляции характеризует линейную зависимость. Линейная вероятностная зависимость случайных величин заключается в том, что при возрастании одной случайной величины другая имеет тенденцию возрасть (или убывать) по линейному закону. Если случайные величины X и Y связаны точной линейной зависимостью:

$$Y = aX + b,$$

то $\tilde{r}_{XY} = \pm 1$, причем знак «плюс» или «минус» берется в зависимости от того, положителен или отрицателен коэффициент a . В общем случае, когда величины X и Y связаны произвольной вероятностной зависимостью, выборочный коэффициент корреляции может иметь значения в пределах:

$$-1 \leq \tilde{r}_{XY} \leq 1.$$

В случае $\tilde{r}_{XY} > 0$ говорят о *положительной корреляции* величин X и Y , в случае $\tilde{r}_{XY} < 0$ – об *отрицательной корреляции*. Положительная корреляция между случайными величинами означает, что при возрастании одной из них другая имеет тенденцию в среднем возрасть; отрицательная корреляция означает, что при возрастании одной из случайных величин другая имеет тенденцию в среднем убывать.

По величине коэффициента корреляции можно судить о тесноте связи между случайными величинами X и Y : если его значение по модулю близко к единице, то связь достаточно тесная; если $\tilde{r}_{XY} \approx 0$, то связь между случайными величинами слабая.

ПРИМЕР 8 (пункт 1 части 2 Задания):

В таблице 12 представлены результаты испытаний двух случайных величин X и Y .

Таблица 12

i	x_i	y_i	i	x_i	y_i
1	4,08	2,14	11	4,31	6,29
2	6,91	3,00	12	2,34	5,52
3	7,42	1,73	13	3,82	3,11
4	3,58	4,24	14	3,98	5,70
5	5,16	3,27	15	3,24	2,60
6	5,19	2,83	16	2,88	5,13
7	4,10	4,22	17	6,19	1,44
8	5,37	4,40	18	5,86	2,20
9	5,02	2,19	19	2,67	3,58
10	6,19	3,20	20	4,36	3,90

Требуется определить выборочный коэффициент корреляции и проанализировать результаты.

Вычислим необходимые для нахождения выборочного коэффициента корреляции оценки параметров распределения генеральной совокупности:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i = \frac{1}{20} \cdot 92,67 = 4,6335,$$

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^n y_i = \frac{1}{20} \cdot 70,69 = 3,5345,$$

$$\overline{x \cdot y} = \frac{1}{n} \cdot \sum_{i=1}^n x_i \cdot y_i = \frac{1}{20} \cdot 306,9292 = 15,3465,$$

$$\tilde{\sigma}_X = \sqrt{\tilde{D}_X} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{20} \cdot 37,6767} = 1,3725,$$

$$\tilde{\sigma}_Y = \sqrt{\tilde{D}_Y} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (y_i - \bar{y})^2} = \sqrt{\frac{1}{20} \cdot 35,8161} = 1,3381,$$

$$\tilde{K}_{XY} = \overline{x \cdot y} - \bar{x} \cdot \bar{y} = 15,3465 - 4,6335 \cdot 3,5345 = -1,0306.$$

Найдем теперь выборочный коэффициент корреляции:

$$\tilde{r}_{XY} = \frac{\tilde{K}_{XY}}{\tilde{\sigma}_X \cdot \tilde{\sigma}_Y} = \frac{-1,0306}{1,3725 \cdot 1,3381} = -0,5611.$$

Анализ полученного выборочного коэффициента корреляции позволяет выдвинуть следующую гипотезу: связь между случайными величинами не очень тесная (значение выборочного коэффициента корреляции по модулю значительно отличается от единицы). Так как выборочный коэффициент корреляции отрицателен, то при возрастании одной случайной величины другая имеет тенденцию в среднем убывать.

2.3. УРАВНЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ

Пусть имеется n наблюдений случайного вектора (X, Y) . При этом $X = \{x_1, x_2, \dots, x_n\}$ и $Y = \{y_1, y_2, \dots, y_n\}$.

Будем считать одну из наблюдаемых величин X неслучайной (детерминированной). Действительно, при сопоставлении этих величин можно отнести все случайные ошибки лишь к величине Y . Тогда ошибка наблюдения будет складываться из собственной случайной ошибки величины Y и из «ошибки сопоставления», возникающей из-за того, что с величиной Y сопоставляется не совсем то значение X , которое имело место на самом деле.

Требуется по данным наблюдений подобрать функцию $\tilde{y} = \tilde{y}(x)$, наилучшим образом описывающую истинную зависимость между случайными величинами. Класс математических функций для описания связи двух случайных величин достаточно широк (линейная, квадратичная, степенная, показательная, логарифмическая). На практике предпочтение отдается простым видам функций, так как они требуют меньшего объема наблюдений. Результаты многих исследований подтверждают, что число наблюдений должно как минимум в 7 раз превышать число коэффициентов при переменной x в модели.

В модели линейной регрессии зависимость между величинами X и Y представляется в виде:

$$y_i = a \cdot x_i + b, \quad i = \overline{1, n},$$

где x_i – детерминированная (неслучайная) величина.

В качестве оценок параметров a и b принимаются величины \tilde{a} и \tilde{b} , минимизирующие сумму квадратов отклонений наблюдаемых значений y_i от расчетных (теоретических) значений $\tilde{y}_i = \tilde{a} \cdot x_i + \tilde{b}$:

$$S(\tilde{a}, \tilde{b}) = \sum_{i=1}^n \left[y_i - (\tilde{a} \cdot x_i + \tilde{b}) \right]^2 \rightarrow \min.$$

Для нахождения минимума функции двух переменных, требуется вычислить частные производные по этим переменным и приравнять их к нулю:

$$\frac{\partial S}{\partial \tilde{a}} = 0 \quad \text{и} \quad \frac{\partial S}{\partial \tilde{b}} = 0.$$

В результате получаем систему двух линейных уравнений:

$$\begin{cases} \sum_{i=1}^n y_i = \tilde{b} \cdot n + \tilde{a} \cdot \sum_{i=1}^n x_i, \\ \sum_{i=1}^n x_i \cdot y_i = \tilde{b} \cdot \sum_{i=1}^n x_i + \tilde{a} \cdot \sum_{i=1}^n x_i^2. \end{cases}$$

Решая данную систему, найдем оценки коэффициентов уравнения линейной регрессии:

$$\tilde{a} = \frac{n \cdot \sum_{i=1}^n x_i \cdot y_i - \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n \cdot \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\tilde{\sigma}_X^2} = \tilde{r}_{XY} \cdot \frac{\tilde{\sigma}_Y}{\tilde{\sigma}_X},$$

$$\tilde{b} = \frac{1}{n} \cdot \sum_{i=1}^n y_i - \tilde{a} \cdot \frac{1}{n} \cdot \sum_{i=1}^n x_i = \bar{y} - \tilde{r}_{XY} \cdot \frac{\tilde{\sigma}_Y}{\tilde{\sigma}_X} \cdot \bar{x}.$$

Коэффициенты уравнения были найдены с помощью метода наименьших квадратов, однако существуют и другие методы нахождения оценок линейной регрессии, например, метод максимального правдоподобия. Критериями

лучшего способа оценивания являются требования состоятельности, несмещенности и эффективности оценок, найденных данным способом. Оценки, полученные по методу наименьших квадратов, удовлетворяют всем этим требованиям, то есть являются «наилучшими».

Коэффициент $\tilde{a} = \tilde{r}_{XY} \cdot \frac{\tilde{\sigma}_Y}{\tilde{\sigma}_X}$ называется *коэффициентом регрессии* Y

на X , а уравнение $y - \bar{y} = \tilde{r}_{XY} \cdot \frac{\tilde{\sigma}_Y}{\tilde{\sigma}_X} (x - \bar{x})$ называется *уравнением линейной регрессии* Y на X .

Аналогично получается уравнение линейной регрессии X на Y :

$$x - \bar{x} = \tilde{r}_{XY} \cdot \frac{\tilde{\sigma}_X}{\tilde{\sigma}_Y} (y - \bar{y}).$$

Обе прямые регрессии проходят через точку (\bar{x}, \bar{y}) , которая называется *центром совместного распределения* величин X и Y . Если $\tilde{r}_{XY} = \pm 1$, то обе прямые регрессии совпадают.

Уравнения линейной регрессии можно получить и с помощью пакета прикладных программ EXCEL, воспользовавшись статистической функцией ЛИНЕЙН.

Главным показателем качества регрессионной модели является **коэффициент детерминации:**

$$R^2 = \frac{1}{\tilde{D}_Y} \cdot \frac{1}{n} \cdot \sum_{i=1}^n (\tilde{y}(x_i) - \bar{y})^2 = 1 - \frac{1}{\tilde{D}_Y} \cdot \frac{1}{n} \cdot \sum_{i=1}^n (\tilde{y}(x_i) - y_i)^2.$$

Значения коэффициента детерминации изменяются от нуля до единицы ($0 \leq R^2 \leq 1$). Если значение R^2 близко к единице, то уравнение регрессии хорошо описывает фактические данные. Если значение R^2 близко к нулю, то связь между величинами отсутствует и уравнение регрессии плохо описывает данные. В случае линейной регрессии $R^2 = \tilde{r}_{XY}^2$.

По уравнению регрессии можно оценить прогнозное значение y^* для заданного значения x^* и получить **точечный прогноз**: $\tilde{y}(x^*) = \tilde{a} \cdot x^* + \tilde{b}$. Однако точечный прогноз дает «плохие», неточные результаты. Поэтому он дополняется интервальной оценкой прогнозного значения – доверительным интервалом.

Известно, что прогнозируемое значение y^* с доверительной вероятностью $p = 1 - \alpha$ принадлежит интервалу прогноза:

$$\left(\tilde{y}(x^*) - t_p(k) \cdot \mu ; \tilde{y}(x^*) + t_p(k) \cdot \mu \right),$$

где $\tilde{y}(x^*)$ – точечный прогноз; $t_p(k)$ – квантиль распределения Стьюдента, определяемый по таблице П 4 Приложения в зависимости от доверительной вероятности и числа степеней свободы $k = n - 2$; μ – средняя ошибка прогноза, вычисляемая по формуле:

$$\mu = \frac{\tilde{\sigma}_Y}{\tilde{\sigma}_X} \cdot \sqrt{\frac{1 - R^2}{n - 2} \cdot \left[(n + 1) \cdot \tilde{D}_X + (x^* - \bar{x})^2 \right]}.$$

ПРИМЕР 9 (пункты 2, 3 и 4 части 2 Задания):

Требуется составить уравнение линейной регрессии Y на X для данных из таблицы 12; оценить качество модели по коэффициенту детерминации; найти доверительный интервал для уровня значимости $\alpha = 0,1$, в который попадает прогнозное значение y^* для $x^* = x_{\max} + 1$.

Уравнение линейной регрессии Y на X имеет вид:

$$y - \bar{y} = \tilde{r}_{XY} \cdot \frac{\tilde{\sigma}_Y}{\tilde{\sigma}_X} (x - \bar{x}).$$

Подставляя найденные ранее значения, получим следующее уравнение:

$$y - 3,5345 = -0,5611 \cdot \frac{1,3381}{1,3725} (x - 4,6335).$$

Разрешив данное уравнение относительно y , получим уравнение линейной регрессии Y на X :

$$y = -0,5470 \cdot x + 6,0692.$$

График прямой регрессии и опытных данных приведены на рисунке 13. Из этого рисунка видно, что при возрастании одной случайной величины другая имеет тенденцию в среднем убывать, о чем свидетельствует и отрицательный выборочный коэффициент корреляции.

В нашем случае коэффициент детерминации имеет достаточно малое, далекое от 1 значение: $R^2 = \tilde{r}_{XY}^2 = 0,3148$, что свидетельствует о том, что линейная регрессия не соответствует опытным данным.

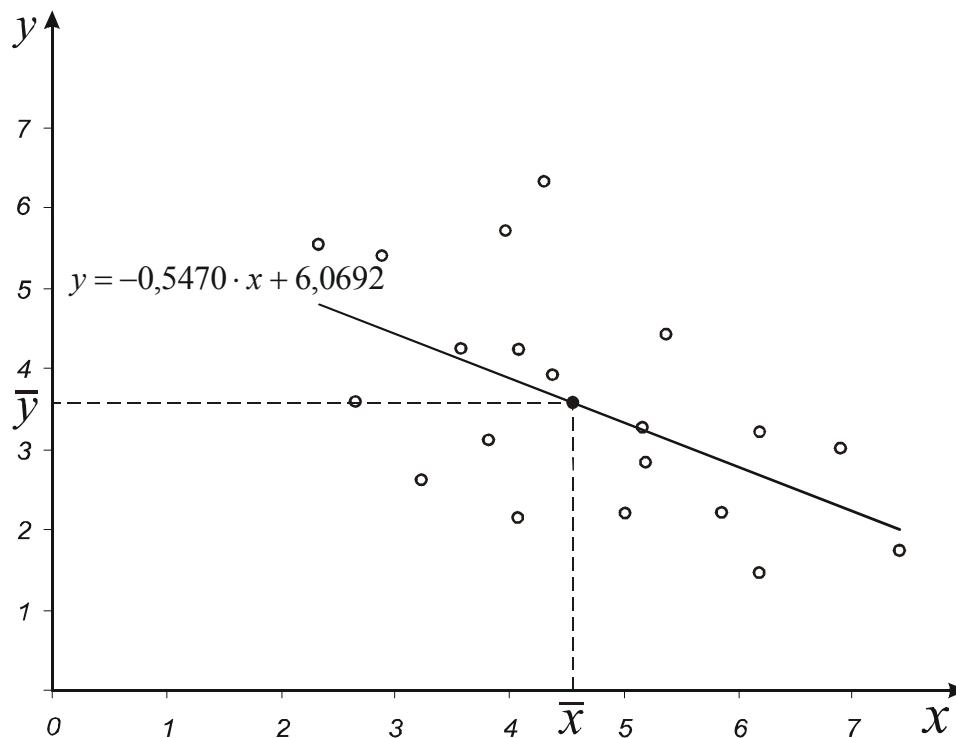


Рис. 13
Опытные данные и прямая регрессии

Найдем теперь доверительный интервал, в который попадает прогнозное значение y^* для $x^* = x_{\max} + 1 = 8,42$. Точечный прогноз:

$$\tilde{y}(x^*) = -0,5470 \cdot 8,42 + 6,0692 = 1,4635,$$

ошибка прогноза:

$$\mu = \frac{1,3381}{1,3725} \cdot \sqrt{\frac{1-0,3148}{18} \cdot [21 \cdot 1,8837 + (8,42 - 4,6335)^2]} = 1,3964.$$

При $\alpha = 0,1$ и $k = 20 - 2 = 18$ квантиль распределения Стьюдента $t_p(k) = t_{0,9}(18) = 1,33$ (таблица П 4 Приложения). Подставив все найденные значения в формулу доверительного интервала прогнозного значения, получим, что прогнозное значение y^* для $x^* = x_{\max} + 1 = 8,42$ с доверительной вероятностью $p = 1 - \alpha$ принадлежит интервалу:

$$(1,4635 - 1,33 \cdot 1,3964 \ ; \ 1,4635 + 1,33 \cdot 1,3964), \text{ или} \\ (-0,3937 \ ; \ 3,3207).$$

Вследствие того, что между случайными величинами Y и X существует слабая зависимость, далекая от линейной, по линейному уравнению регрессии получен слишком широкий интервал изменений прогнозного значения. Таким образом, в данном случае прогноз по линейной модели не соответствует характеру изменения опытных данных.

ЛИТЕРАТУРА

1. Айвазян С.А. Теория вероятностей и прикладная статистика. – М.: «Издательство ЮНИТИ-ДАНА», 2001.
2. Вентцель Е.С., Овчаров Л.А. Теория вероятностей и ее инженерные приложения. - М.: Наука, 1988.
3. Вентцель Е.С. Теория вероятностей. – М.: Высшая школа, 2002.
4. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 2004.
5. Сборник задач по математике для вузов «Теория вероятностей и математическая статистика» под ред. А.В. Ефимова, Ч. 3. – М.: Наука, 1990.

ПРИЛОЖЕНИЕ

Таблица П 1

$$\text{Функция Лапласа } \Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$$

<i>x</i>	Сотые доли <i>x</i>									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0800	0,0120	0,0160	0,0199	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	0,3261	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	0,4484	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	0,4641	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	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998

Функция распределения нормального закона $N(0,1)$ $\Phi(x) = \Phi_0(x) + \frac{1}{2}$.

Таблица П 2

Значения функции плотности нормального закона $N(0,1)$ $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

x	Сотые доли x									
	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
0,1	0,3970	0,3965	0,3961	0,3956	0,3951	0,3945	0,3939	0,3932	0,3925	0,3918
0,2	0,3910	0,3902	0,3894	0,3885	0,3876	0,3867	0,3857	0,3847	0,3836	0,3825
0,3	0,3814	0,3802	0,3790	0,3778	0,3765	0,3752	0,3739	0,3725	0,3712	0,3697
0,4	0,3683	0,3668	0,3653	0,3637	0,3621	0,3605	0,3589	0,3572	0,3555	0,3538
0,5	0,3521	0,3503	0,3485	0,3467	0,3448	0,3429	0,3410	0,3391	0,3372	0,3352
0,6	0,3332	0,3312	0,3292	0,3271	0,3251	0,3230	0,3209	0,3187	0,3166	0,3144
0,7	0,3123	0,3101	0,3079	0,3056	0,3034	0,3011	0,2989	0,2966	0,2943	0,2920
0,8	0,2897	0,2874	0,2850	0,2827	0,2803	0,2780	0,2756	0,2732	0,2709	0,2685
0,9	0,2661	0,2637	0,2613	0,2589	0,2565	0,2541	0,2516	0,2492	0,2468	0,2444
1,0	0,2420	0,2396	0,2371	0,2347	0,2323	0,2299	0,2275	0,2251	0,2227	0,2203
1,1	0,2179	0,2155	0,2131	0,2107	0,2083	0,2059	0,2036	0,2012	0,1989	0,1965
1,2	0,1942	0,1919	0,1895	0,1872	0,1849	0,1826	0,1804	0,1781	0,1758	0,1736
1,3	0,1714	0,1691	0,1669	0,1647	0,1626	0,1604	0,1582	0,1561	0,1539	0,1518
1,4	0,1497	0,1476	0,1456	0,1435	0,1415	0,1394	0,1374	0,1354	0,1334	0,1315
1,5	0,1295	0,1276	0,1257	0,1238	0,1219	0,1200	0,1182	0,1163	0,1145	0,1127
1,6	0,1109	0,1092	0,1074	0,1057	0,1040	0,1023	0,1006	0,0989	0,0973	0,0957
1,7	0,0940	0,0925	0,0909	0,0893	0,0878	0,0863	0,0848	0,0833	0,0818	0,0804
1,8	0,0790	0,0775	0,0761	0,0748	0,0734	0,0721	0,0707	0,0694	0,0681	0,0669
1,9	0,0656	0,0644	0,0632	0,0620	0,0608	0,0596	0,0584	0,0573	0,0562	0,0551
2,0	0,0540	0,0529	0,0519	0,0508	0,0498	0,0488	0,0478	0,0468	0,0459	0,0449
2,1	0,0440	0,0431	0,0422	0,0413	0,0404	0,0396	0,0387	0,0379	0,0371	0,0363
2,2	0,0355	0,0347	0,0339	0,0332	0,0325	0,0317	0,0310	0,0303	0,0297	0,0290
2,3	0,0283	0,0277	0,0270	0,0264	0,0258	0,0252	0,0246	0,0241	0,0235	0,0229
2,4	0,0224	0,0219	0,0213	0,0208	0,0203	0,0198	0,0194	0,0189	0,0184	0,0180
2,5	0,0175	0,0171	0,0167	0,0163	0,0158	0,0154	0,0151	0,0147	0,0143	0,0139
2,6	0,0136	0,0132	0,0129	0,0126	0,0122	0,0119	0,0116	0,0113	0,0110	0,0107
2,7	0,0104	0,0101	0,0099	0,0096	0,0093	0,0091	0,0088	0,0086	0,0084	0,0081
2,8	0,0079	0,0077	0,0075	0,0073	0,0071	0,0069	0,0067	0,0065	0,0063	0,0061
2,9	0,0060	0,0058	0,0056	0,0055	0,0053	0,0051	0,0050	0,0048	0,0047	0,0046
3,0	0,0044	0,0043	0,0042	0,0040	0,0039	0,0038	0,0037	0,0036	0,0035	0,0034
3,1	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026	0,0025	0,0025
3,2	0,0024	0,0023	0,0022	0,0022	0,0021	0,0020	0,0020	0,0019	0,0018	0,0018
3,3	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014	0,0013	0,0013
3,4	0,0012	0,0012	0,0012	0,0011	0,0011	0,0010	0,0010	0,0010	0,0009	0,0009
3,5	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007	0,0007	0,0007	0,0006
3,6	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0004
3,7	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003	0,0003	0,0003	0,0003
3,8	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002
3,9	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001
4,0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

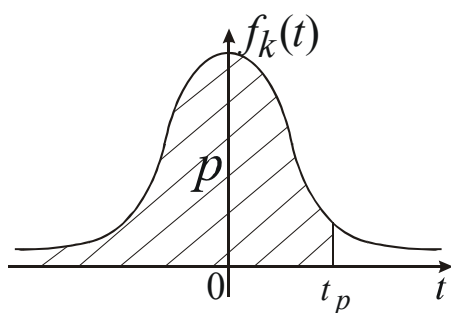
Нормальное распределение

Квантили распределения C_p :
$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{C_p} e^{-\frac{x^2}{2}} dx = p$$

p	C_p	p	C_p	p	C_p
0,50	0,000	0,860	1,080	0,9910	2,366
0,55	0,126	0,870	1,126	0,9920	2,409
0,60	0,253	0,880	1,175	0,9930	2,457
0,65	0,385	0,890	1,227	0,9940	2,512
0,70	0,524	0,900	1,282	0,9950	2,576
0,75	0,674	0,910	1,341	0,9955	2,612
0,76	0,706	0,920	1,405	0,9960	2,652
0,77	0,739	0,930	1,476	0,9965	2,697
0,78	0,772	0,940	1,555	0,9970	2,748
0,79	0,806	0,950	1,645	0,9975	2,807
0,80	0,842	0,960	1,751	0,9980	2,878
0,81	0,878	0,970	1,881	0,9985	2,968
0,82	0,915	0,975	1,960	0,9990	3,090
0,83	0,954	0,980	2,051	0,9995	3,291
0,84	0,994	0,985	2,170	0,9999	3,720
0,85	1,036	0,990	2,326	0,99999	4,265

Примечание. Если $0 < p < 0,5$, то $C_p = -C_{1-p}$.

Таблица П 4

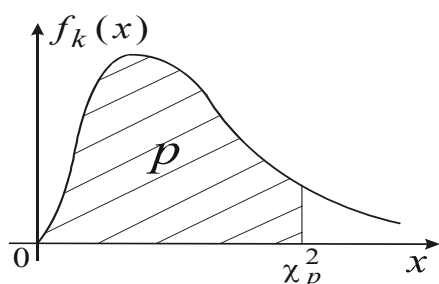
Распределение Стьюдента $f_k(t)$

$$\text{Квантили распределения } t_p: \int_{-\infty}^{t_p} f_k(t) dt = p$$

Число степеней свободы k	Вероятность, p					
	0,9	0,95	0,975	0,99	0,995	0,9995
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,965	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,941
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,869
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,405
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
12	1,356	1,782	2,179	2,681	3,055	4,318
14	1,345	1,761	2,145	2,625	2,977	4,140
16	1,337	1,746	2,120	2,584	2,921	4,015
18	1,330	1,734	2,101	2,552	2,878	3,922
20	1,325	1,725	2,086	2,528	2,845	3,850
22	1,321	1,717	2,074	2,508	2,819	3,792
24	1,318	1,711	2,064	2,492	2,797	3,745
26	1,315	1,706	2,056	2,479	2,779	3,707
28	1,313	1,701	2,048	2,467	2,763	3,674
30	1,310	1,697	2,042	2,457	2,750	3,646
60	1,296	1,671	2,000	2,390	2,660	3,460
120	1,289	1,658	1,980	2,358	2,617	3,373
∞	1,282	1,645	1,960	2,326	2,576	3,291

Примечание. При $k \geq 30$ $t_p(k) \approx C_p$, где C_p – квантиль нормального распределения (таблица П 3 Приложения).

Таблица П 5

Распределение хи – квадрат χ^2 Квантили распределения χ_p^2 : $\int_0^{\chi_p^2} f_k(x) dx = p$

Число степеней свободы k	Вероятность, p						
	0,001	0,005	0,01	0,02	0,025	0,05	0,01
1	0,0000016	0,000039	0,00016	0,00063	0,00098	0,004	0,016
2	0,002	0,010	0,020	0,040	0,051	0,103	0,211
3	0,024	0,072	0,115	0,185	0,216	0,352	0,584
4	0,091	0,207	0,297	0,429	0,484	0,711	1,064
5	0,210	0,412	0,554	0,752	0,831	1,145	1,610
6	0,381	0,676	0,872	1,134	1,237	1,635	2,204
7	0,598	0,989	1,239	1,564	1,690	2,167	2,833
8	0,857	1,344	1,646	2,032	2,180	2,733	3,490
9	1,152	1,735	2,008	2,532	2,700	3,325	4,168
10	1,479	2,156	2,558	3,059	3,247	3,940	4,865
11	1,834	2,603	3,053	3,609	3,816	4,575	5,578
12	2,214	3,074	3,571	4,178	4,404	5,226	6,304
13	2,617	3,565	4,107	4,765	5,009	5,892	7,042
14	3,041	4,075	4,660	5,368	5,629	6,571	7,790
15	3,483	4,601	5,229	5,985	6,262	7,261	8,547
16	3,942	5,142	5,812	6,614	6,908	7,962	9,312
17	4,416	5,697	6,408	7,255	7,564	8,672	10,085
18	4,905	6,265	7,015	7,096	8,231	9,390	10,865
19	5,407	6,844	7,633	8,567	8,907	10,117	11,651
20	5,921	7,434	8,260	9,237	9,591	10,851	12,443
21	6,447	8,034	8,897	9,915	10,283	11,591	13,240
22	6,983	8,643	9,542	10,600	10,982	12,338	14,041
23	7,529	9,260	10,196	11,293	11,688	13,091	14,848
24	8,085	9,886	10,856	11,992	12,401	13,848	15,659
25	8,649	10,520	11,524	12,697	13,120	14,611	16,473
26	9,222	11,160	12,198	13,409	13,844	15,379	17,292
27	9,803	11,808	12,879	14,125	14,573	16,151	18,114
28	10,391	12,461	13,565	14,847	15,308	16,928	18,939
29	10,986	13,121	14,256	15,574	16,047	17,708	19,768
30	11,588	13,787	14,953	16,306	16,791	18,493	20,599

Продолжение таблицы П 5

Распределение хи – квадрат χ^2

Число степеней свободы k	Вероятность, p						
	0,9	0,95	0,975	0,98	0,99	0,995	0,999
1	2,706	3,841	5,024	5,412	6,635	7,879	10,827
2	4,605	5,991	7,378	7,824	9,210	10,597	13,815
3	6,251	7,815	9,348	9,837	11,345	12,838	16,268
4	7,779	9,488	11,143	11,668	13,277	14,860	18,465
5	9,236	11,070	12,832	13,388	15,086	16,750	20,517
6	10,645	12,592	14,449	15,033	16,812	18,548	22,457
7	12,017	14,067	16,013	16,622	18,475	20,278	24,322
8	13,362	15,507	17,535	18,168	20,090	21,955	26,125
9	14,684	16,919	19,023	19,679	21,666	23,589	27,877
10	15,987	18,307	20,483	21,161	23,209	25,188	29,588
11	17,275	19,675	21,920	22,618	24,725	26,757	31,264
12	18,549	21,026	23,337	24,054	26,217	28,300	32,909
13	19,812	22,362	24,736	25,472	27,688	29,819	34,528
14	21,064	23,685	26,119	26,873	29,141	31,319	36,123
15	22,307	24,996	27,488	28,259	30,578	32,801	37,697
16	23,542	26,296	28,845	29,633	32,000	34,267	39,252
17	24,769	27,587	30,191	30,995	33,409	35,718	40,790
18	25,989	28,869	31,526	32,346	34,805	37,156	42,312
19	27,204	30,144	32,852	33,687	36,191	38,582	43,820
20	28,412	31,410	34,170	35,020	37,566	39,997	45,315
21	29,615	32,671	35,479	36,343	38,932	41,401	46,797
22	30,813	33,924	36,781	37,659	40,289	42,796	48,268
23	32,007	35,172	38,076	38,968	41,638	44,181	49,728
24	33,196	36,415	39,364	40,270	42,980	45,558	51,179
25	34,382	37,652	40,646	41,566	44,314	46,928	52,620
26	35,563	38,885	41,923	42,856	45,642	48,290	54,052
27	36,741	40,113	43,194	44,140	46,963	49,645	55,476
28	37,916	41,337	44,461	45,419	48,278	50,993	56,893
29	39,087	42,557	45,722	46,693	49,588	52,336	58,302
30	40,256	43,773	46,979	47,962	50,892	53,672	59,703

Примечание. При $k \geq 30$ $\chi_p^2(k) \approx \frac{(C_p + \sqrt{2k-1})^2}{2}$, где C_p – квантиль нормального распределения (таблица П 3 Приложения).

Учебное издание

Денискина Екатерина Александровна

Коломиец Павел Эдуардович

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ

**МЕТОДИЧЕСКИЕ УКАЗАНИЯ
к расчетной работе**

Редактор ■■■
Корректор ■■■

Подписано в печать ■■■ Формат 60x84 1/16.
Бумага офсетная. Печать офсетная.
Усл. печ. л. ■■■. Усл. Кр. – отт. ... Уч. – изд. л. ...
Тираж ... экз. Заказ ...