Математическая статистика. Основные понятия.

Математическая статистика – это раздел математики, который изучает методы сбора, систематизации, обработки результатов наблюдений массовых случайных явлений.

Любое множество, подлежащее изучению в статистике, называется *генеральной совокупностью*. Любое подмножество генеральной совокупности называется *выё боркой*. Количество элементов в генеральной совокупности или в выборке называется *объемом*. Элементы выборки могут характеризоваться числами, отражающими какой-либо признак изучаемого объекта. Эти числа называются *вариантами,* так как от выборки к выборке эти значения меняются.

Первым шагом в обработке полученных данных является составление статистического или вариационного ряда.

Статистический ряд – это таблица, в которой перечислены варианты в порядке возрастания и указаны соответствующие им частоты.

Для графического изображения статистического ряда частот служит ломаная в прямоугольной декартовой системе координат с вершинами в точках (x_i, n_i) - называемая *полигоном частот,* или ломаная с вершинами в точках $\left(x_i, \frac{n_i}{n}\right)$ - называемая *полигоном относительных частот.* Здесь x_i - возможные значения вариант, n_i - частота, т. е. количество появления i варианты, n - объем выборки. При большом объеме выборки ее элементы объединяются в группы (разряды), представляя результаты опытов в виде сгруппированного статистического ряда. Для этого интервал, содержащий все элементы выборки, разбивается на k непересекающихся интервалов, обычно одинаковой длины l.

Для графического изображения сгруппированной выборки служит ступенчатая фигура из прямоугольников, называемая *гистограммой*. Для построения гистограммы на оси ox откладываются интервалы длины l, которые служат основаниями прямоугольников, а их высоты определяются отношением $\frac{n_i}{l}$, если мы строим гистограмму частот, или $\frac{n_i}{n \cdot l}$, если мы строим гистограмму относительных частот.

Пример 1. а) Дан статистический ряд. Требуется построить полигон относительных частот. б) Дан сгруппированный статистический ряд. Требуется построить гистограмму относительных частот.

а)

x_i значения вариант	15	16	17	18	19
n_i частоты	1	5	6	5	3

б)

границы интервалов	10-20	20-30	30-40	40-50	50-60
частоты	1	2	7	18	12

Решение. а) Для построения полигона частот найдем относительные частоты по формуле $\frac{n_i}{n}$, где $n=\sum_{i=1}^{5} n_i = 1+5+6+5+3=20$.

Результат запишем в таблицу

x_i	15	16	17	18	19	
n_i	1	5	6	5	3	$\sum = 20$
$\frac{n_i}{n}$	1/20=0,05	5/20=0,25	6/20=0,3	5/20=0,25	3/20=0,15	$\sum = 1$

Строим ломаную с координатами $\left(x_i, \frac{n_i}{n}\right)$ (рис. 1).

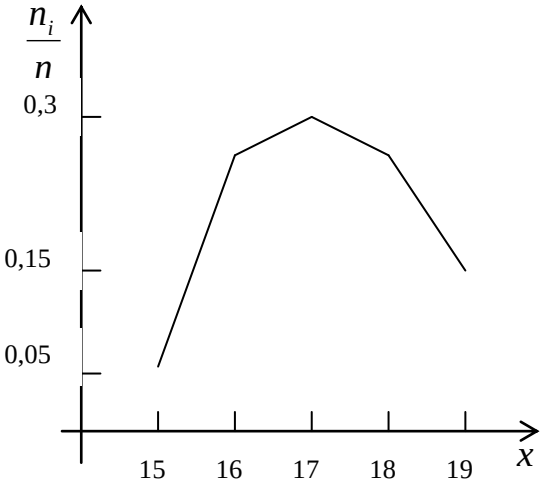


Рис. 1

Замечание. Обычно при построении полигона масштаб по осям берется неодинаковым.

б) Для построения гистограммы относительных частот найдем относительные частоты по формуле $\frac{n_i}{n}$, высоты прямоугольников — по формуле $h=\frac{n_i}{nl}$, где $n=\sum_{i=1}^{n} n_i = 1+2+7+18+12=40$, $l=10$. Величина h характеризует плотность попадания вариант в i-ый интервал. Результаты удобно записать в таблицу.

$(x_i - x_{i-1})$	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	
n_i	1	2	7	18	12	$\sum = 40$
$\frac{n_i}{n}$	1/40 = 0,025	2/40 = 0,05	7/40 = 0,175	18/40 = 0,45	12/40 = 0,3	$\sum = 1$
$\frac{n_i}{nl}$	0,025/10 = 0,0025	0,05/10 = 0,005	0,175/10 = 0,0175	0,45/10 = 0,045	0,3/10 = 0,03	

Строим гистограмму (рис. 2).

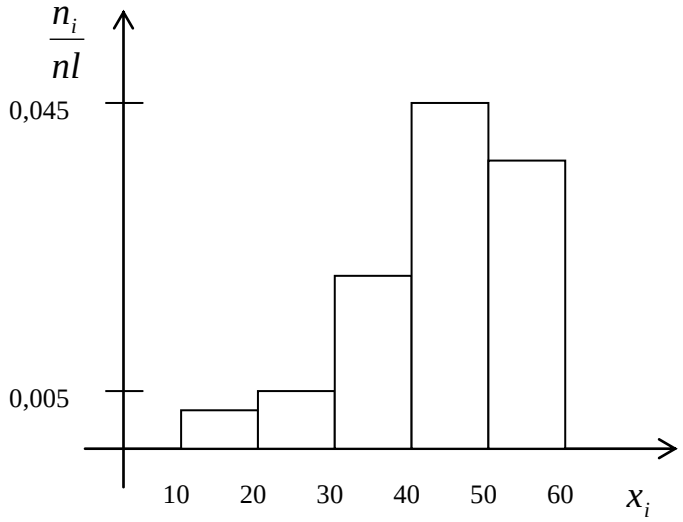


Рис. 2

Статистические гипотезы

Во многих случаях результаты наблюдений используются для проверки предположений (гипотез) относительно тех или иных свойств распределения генеральной совокупности. В частности, такого рода задачи возникают при сравнении различных технологических процессов или методов обработки по определенным измеряемым признакам, например, по точности, производительности и т. д.

Пусть X — наблюдаемая дискретная или непрерывная случайная величина.

Статистической гипотезой называется предположение относительно параметров или вида распределения случайной величины X.

Основной или *нулевой* гипотезой H_0 называют выдвинутую гипотезу, а гипотезу H_1, ей противоречащую — *конкурирующей* или *альтернативной.*

Правило, по которому принимается решение принять или отклонить гипотезу H_0 называют *статистическим критерием* K. Обычно статистические критерии выражаются числами, которые вычисляются по вариантам выборки, или находятся теоретически. Значение критерия, найденное на основе выборки наблюдений случайной величины X, называют

выборочным и обозначают $K_в$. Значение критерия, которое находится по таблице, называется *теоретическим* и обозначается K_T.

Проверка статистической гипотезы основывается на принципе, в соответствии с которым маловероятные события считаются невозможными, а события, имеющие большую вероятность, считаются достоверными. Этот принцип можно реализовать следующим образом. Перед анализом выборки фиксируется некоторая малая вероятность α, называемая *уровнем значимости,* и равная вероятности отвергнуть правильную H_0 гипотезу. Таким образом, вероятность принять правильную H_0 гипотезу будет равна $1-\alpha$. Уровень значимости α определяет размер «*критической* области».

Критическая область V_k — те значения критерия K, при которых гипотезу H_0 отвергают. Критерий, основанный на использовании заранее заданного уровня значимости, называется критерием *значимости.*

Таким образом, проверка значимости статистической гипотезы при помощи критерия значимости может быть разбита на следующие этапы:

1) сформулировать проверяемую (H_0) и альтернативную (H_1) гипотезы;
2) назначить уровень значимости α;
3) выбрать статистический критерий;
4) определить теоретическое (K_T) и выборочное ($K_в$) значения критерия;
5) определить критическую область V_k;
6) принять статистическое решение: если $K_в \notin V_k$, то гипотезу H_0 принять, т. е. считать, что гипотеза H_0 не противоречит результатам наблюдений; если $K_в \in V_k$, то отклонить гипотезу H_0 как не согласующуюся с результатами наблюдений.

Критерий Пирсона χ^2 (хи-квадрат)

Этот критерий был введен английским математиком К. Пирсоном (1857 – 1936). Критерий служит для проверки гипотезы о виде распределения случайной величины X.

Итак, пусть имеется сгруппированный статистический ряд, разбитый на k интервалов, где k - заранее выбранное число, n_i - число вариант, попадающих в i интервал, n - объем выборки, $p_i = P\left(x_{i-1} \le X \le x_i\right)$ - вероятность попадания случайной величины X в i - ый интервал при выбранном законе распределения случайной величины.

При этих условиях Пирсон предложил в качестве критерия K рассмотреть случайную величину

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i\right)^2}{np_i}, \text{ (n_i - случайные величины).} \qquad (1)$$

Он доказал, что χ^2 при больших k практически не зависит от гипотетического распределения и определяется функцией плотности

$$\psi_r(u)=\frac{1}{2^{r/2}\,\Gamma\left(\frac{r}{2}\right)u}\,u^{\frac{r}{2}-1}\cdot e^{-\frac{r}{2}},\ u\geq 0 \qquad (2)$$

где r - число степеней свободы, определяемое по формуле $r=k-m-1$, здесь m- число параметров гипотетического закона распределения, подлежащих определению по опытным данным.

График функции плотности $\psi_r(u)$ имеет вид (рис. 3):

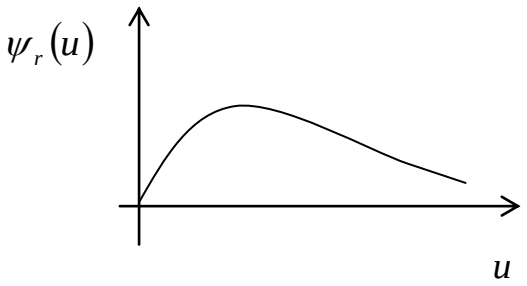


Рис. 3

Критерий χ^2 заключается в следующем. По опытным данным считают выборочное значение критерия Пирсона

$$\chi^2_в=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i},\ (n_i\text{ - выборочные частоты}).$$

По таблице критических точек распределения χ^2 (прил. 1) по заданному уровню значимости α и числу степеней свободы r находят теоретическое значение критерия Пирсона χ^2_T.

Если значение $\chi^2_в$ окажется больше или равно χ^2_T, то гипотезу отвергают. Если же $\chi^2_в$ меньше χ^2_T, то гипотезу принимают и считают ее не противоречащей опытным данным.

При использовании критерия хи-квадрат рекомендуем промежуточные результаты заносить в таблицу:

$(x_{i-1},x_i]$	n_i	p_i	np_i	n_i-np_i	$(n_i-np_i)^2$	$\frac{(n_i-np_i)^2}{np_i}$
$(x_0,x_1]$	n_1	p_1	np_1	n_1-np_1	$(n_1-np_1)^2$	$\frac{(n_1-np_1)^2}{np_1}$
-	-	-	-	-	-	
$(x_{k-1},x_k]$	n_k	p_k	np_k	n_k-np_k	$(n_k-np_k)^2$	$\frac{(n_k-np_k)^2}{np_k}$

Замечание. Разбивку на интервалы надо производить так, чтобы в каждом из них было 5-10 наблюдений. Интервалы, содержащие мало наблюдений, рекомендуется объединять с соседними.

Пример 2. Даны результаты наблюдений некоторой случайной величины X. Проверить гипотезу о ее нормальном распределении.

интервалы	3,5-4,5	4,5-5,5	5,5-6,5	6,5-7,5	7,5-8,5	8,5-9,5
число вариант	6	13	25	16	11	9

Решение. 1. Построим гистограмму относительных частот (рис. 4), данные для ее построения занесем в таблицу ($n = \sum n_i = 80$, длина интервалов $l = 1$).

$(x_{i-1}, x_i]$	(4) 3,5-4,5	(5) 4,5-5,5	(6) 5,5-6,5	(7) 6,5-7,5	(8) 7,5-8,5	(9) 8,5-9,5
n_i	6	13	25	16	11	9
$\frac{n_i}{n}$	$\frac{6}{80} = 0{,}075$	$\frac{13}{80} = 0{,}1625$	$\frac{25}{80} = 0{,}3125$	$\frac{16}{80} = 0{,}2$	$\frac{11}{80} = 0{,}1375$	$\frac{9}{80} = 0{,}11125$
$h = \frac{n_i}{nl}$	0,075	0,1625	0,3125	0,2	0,1375	0,1125

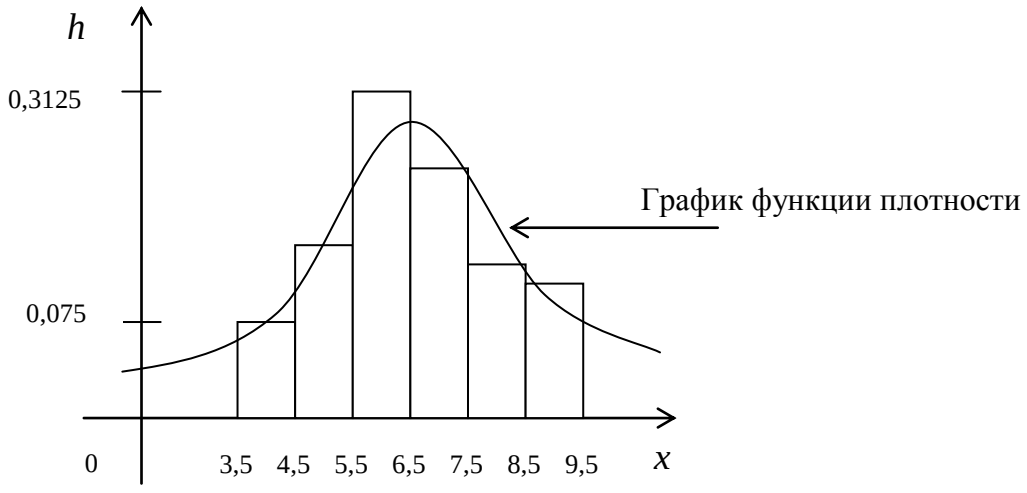


Рис. 4

2. По виду гистограммы можно предположить, что наблюдаемая случайная величина имеет нормальное распределение - $N(a, \sigma^2)$. Функция плотности

вероятности нормального распределения имеет вид $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$, где параметры a и σ неизвестны.

В качестве значений параметров распределения возьмем их оценки, полученные на основе опытных данных. Оценкой параметра a является величина

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k}x_i n_i=\frac{(x_1n_1+x_2n_2+\ldots+x_kn_k)}{n}, \quad (3)$$

оценкой параметра σ^2 является величина

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 n_i. \quad (4)$$

В обеих формулах x_i - середина i-го интервала.

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{6}x_i n_i=\frac{1}{80}(4\cdot6+5\cdot13+6\cdot25+7\cdot16+8\cdot11+9\cdot9)=6{,}5$$

$$s^2=\frac{1}{n-1}\sum_{i=1}^{6}\left(x_i-\bar{x}\right)^2 n_i=\frac{1}{79}((4-6{,}5)^2\cdot6+(5-6{,}5)^2\cdot13+(6-6{,}5)^2\cdot25+$$

$$+(7-6{,}5)^2\cdot11+(8-6{,}5)^2\cdot9)=1{,}97\Rightarrow s=\sqrt{1{,}97}=1{,}4.$$

Итак, выдвигаем гипотезу о том, что изучаемая случайная величина имеет функцию плотности вероятности

$$f(x)=\frac{1}{1{,}4\cdot\sqrt{2\pi}}e^{-\frac{(x-6{,}5)^2}{2\cdot1{,}97}} \quad (5)$$

Ее график построим на том же чертеже, что и гистограмму (рис. 4). Для построения достаточно найти точки максимума $x_{\max}=\bar{x}=6{,}5$,

$y_{\max}=\frac{0{,}4}{s}=\frac{0{,}4}{\sqrt{1{,}97}}\approx0{,}28$ и точки перегиба $x_{пер}=\bar{x}\pm s=6{,}5\pm1{,}4$,

$y_{пер}=\frac{0{,}24}{s}=\frac{0{,}24}{\sqrt{1{,}97}}\approx0{,}17$. Затем эти точки следует соединить плавной линией, учитывая форму кривой нормального распределения. (рис. 4).

3. Зададимся уровнем значимости, например, $\alpha=0{,}05$. Для получения надежных выводов на основе критерия хи-квадрат нужно объединить первый интервал, содержащий мало наблюдений, со вторым интервалом. Тогда имеем всего $k=5$ интервалов. Определим $\chi^2_T(\alpha,r)$, $r=k-m-1=5-3=2$ (r – число степеней свободы, m – число неизвестных параметров). Итак, $\chi^2_T(0{,}05;2)=5{,}99$ (прил. 1).

4. Вычислим $\chi^2_в=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}$. Для этого сначала вычислим вероятности, попадания исследуемой случайной величины в каждый интервал, согласно

гипотезе. В случае нормального распределения они вычисляются по формуле:

$$p_i = P\left(x_{i-1} < X < x_i\right) = \Phi\left(\frac{x_i - \bar{x}}{s}\right) - \Phi\left(\frac{x_{i-1} - \bar{x}}{s}\right).$$

Тогда $P\left(3{,}5 < X < 5{,}5\right) = \Phi\left(\frac{5{,}5 - 6{,}5}{\sqrt{1{,}97}}\right) - \Phi\left(\frac{3{,}5 - 6{,}5}{\sqrt{1{,}97}}\right) = 0{,}22$,

$$P\left(5{,}5 < X < 6{,}5\right) = \Phi\left(\frac{6{,}5 - 6{,}5}{\sqrt{1{,}97}}\right) - \Phi\left(\frac{5{,}5 - 6{,}5}{\sqrt{1{,}97}}\right) = 0{,}26,$$

где $\Phi(x)$ – функция Лапласа, значения которой приведены в прил. 2.

Аналогично $P\left(6{,}5 < x < 7{,}5\right) = 0{,}16$, $P\left(7{,}5 < x < 8{,}5\right) = 0{,}16$,

$$P\left(8{,}5 < x < 9{,}5\right) = 0{,}06.$$

Вычисления χ^2 удобно вести, фиксируя промежуточные результаты в таблице.

n_i	p_i	np_i	$n_i - np_i$	$\left(n_i - np_i\right)^2$	$\frac{\left(n_i - np_i\right)^2}{np_i}$
19	0,22	17,6	1,4	1,96	0,11
25	0,26	20,8	4,2	17,64	0,85
16	0,26	20,8	4,8	23,06	1,11
11	0,16	12,8	1,8	3,24	0,25
9	0,08	4,8	4,2	17,64	3,89

$\chi^2_в = 6{,}21$. Величина $\chi^2_в$ равна сумме значений в последнем столбце таблицы.

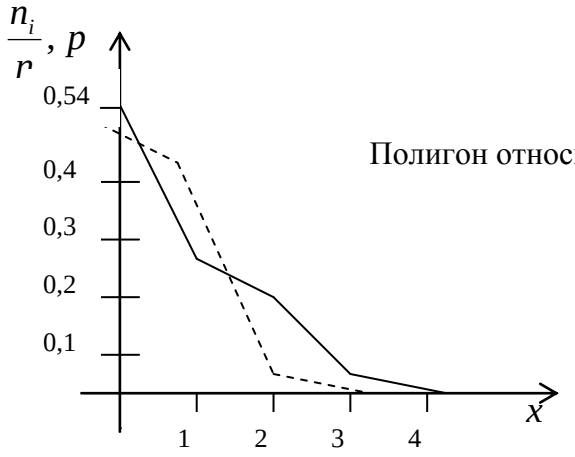
5. Сравним $\chi^2_в$ и χ^2_T : $\chi^2_в = 6{,}21 > \chi^2_T = 5{,}99$. Таким образом, при выбранном уровне значимости $\chi^2_в$ принадлежит критической области V_k, а значит гипотезу о нормальном распределении следует отвергнуть. Следует отметить, что вероятность того, что мы ошибаемся, меньше 0,05.

Пример 3. Результаты наблюдений случайной величины представлены в виде статистического ряда.

x_i	0	1	2	3	4 и более
n_i	54	27	14	5	0

$$n = \sum_{n=1}^{n} n_i = 100$$

Решение. 1. Построим полигон относительных частот $\left(\frac{n_i}{n}\right)$ - ломаную линию с вершинами в точках $\left(x_i, \frac{n_i}{n}\right)$, рис. 5 (на рис. сплошная линия).



Полигон относительных частот

Рис. 5

2. По виду полигона частот можно выдвинуть предположение, что изучаемая случайная величина имеет пуассоновский закон распределения, т. е. $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$ Так как в законе Пуассона параметр равен математическому ожиданию, а его оценкой является величина $\overline{x}$, то

$$\lambda=\overline{x}=\sum_{n=1}^{k}\frac{x_i n_i}{n},\quad \overline{x}=\frac{0{,}54+1\cdot 27+2\cdot 14+3\cdot 3+4\cdot 0}{100}=0{,}7,$$

и изучаемая случайная величина имеет закон распределения

$$P(X=k)=p_k=\frac{(0{,}7)^k e^{-0{,}7}}{k!}, \qquad (6)$$

где $k=0,1,2,3.$

3. Зададимся уровнем значимости, например, $\alpha=0{,}05$. Последние 2 разряда, содержащие мало наблюдений (нужно 5-10), можно объединить. Определим $\chi^2_T(\alpha,r)\quad r=k-m-1=4-1-1=2$, итак $\chi^2_T(0{,}05;2)=5{,}99$ (прил. 1).

4. Вычислим $\chi^2_в=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}$. Для этого сначала вычислим вероятности p_k для каждого из четырех интервалов: $p_0=\frac{(0{,}7)^0 e^{-0{,}5}}{0!}=0{,}5$, $p_1=\frac{(0{,}7)^1 e^{-0{,}7}}{1!}=0{,}35$,

$p_2=\frac{(0{,}7)^2 e^{-0{,}7}}{2!}=0{,}12$, $p_3=1-p_1-p_2=1-0{,}5-0{,}35-0{,}12=0{,}03.$

Используя полученные вероятности, построим ломаную с вершинами в точках (x_i, p_i). На рис. 5 эта ломаная показана пунктирной линией. Вычисление $\chi^2_в$ оформляем в виде таблицы.

n_i	p_i	np_i	$n_i - np_i$	$(n_i - np_i)^2$	$\frac{(n_i - np_i)^2}{np_i}$
54	0,5	$100 \cdot 0{,}5 = 50$	54-50=4	$4^2 = 16$	$\frac{16}{50} = 0{,}32$
27	0,35	35	-8	64	1,83
14	0,12	12	2	4	0,33
5	0,03	3	2	4	1,33

Величина $\chi^2_в$ равна сумме величин в последнем столбце таблицы, т. е. $\chi^2_в$=3,18.

5. Сравним $\chi^2_в$ и χ^2_T. $\chi^2_в$=3,18< χ^2_T=5,99. Таким образом, $\chi^2_в$ в критическую область не входит. Делаем вывод: гипотеза опытным данным не противоречит.

Линейная корреляция

Две случайные величины X и Y могут быть функционально зависимы, статистически зависимы или независимы. Наиболее простой формой зависимости между величинами является функциональная зависимость, при которой каждому значению одной величины соответствует определенное значение другой. Однако на практике связь между величинами носит случайный характер.

Статистической называется зависимость, при которой изменение одной из случайных величин ведет к изменению закона распределения другой величины. В частности, если при изменении одной из величин изменяется среднее значение другой, то статистическая зависимость называется корреляционной. Статистическая зависимость более сложна, чем функциональная. Она возникает, если одна величина зависит не только от другой, но и от ряда прочих случайных факторов. Примерами статистической зависимости являются связи между ростом ребенка и его возрастом, между урожайностью ягодных культур и их рыночными ценами, между температурой закалки и твердостью стали и т. д.

Пусть произведено n независимых опытов, в которых наблюдались случайные величины X и Y. В результате опытов получены пары чисел (x_i, y_j) $(i = \overline{1,l};\ j = \overline{1,k})$. Данные сводят в корреляционную таблицу:

X / Y	y_1	y_2	…	y_k	n_x	$\overline{y}_x$
x_1	n_{11}	n_{12}	…	n_{1k}	n_{x_1}	$\overline{y}_{x_1}$
x_2	n_{21}	n_{22}	…		n_{x_2}	$\overline{y}_{x_2}$
x_l	n_{p_1}	n_{p_2}		n_{lk}	n_{x_l}	$\overline{y}_{x_l}$
n_y	n_{y_1}	n_{y_2}		n_{y_k}	n	

В первой строке таблицы указаны наблюденные значения случайной величины $Y: y_1, y_2, ..., y_k$; в первом столбце – величины $X: x_1, x_2, ..., x_l$. На пересечении строк и столбцов вписаны частоты n_{ij} наблюдаемых пар значений случайных величин. Пустая клетка означает, что соответствующая пара чисел в результате опытов не наблюдалась. В столбце n_x записаны суммы частот строк, в строке n_y - суммы частот столбцов, причем $\sum n_{x_i} = \sum n_{y_j} = n$ — объем выборки.

Назовем условным средним $\overline{y}_x$ среднее арифметическое значений случайной величины Y, соответствующих значению $X = x$.

Уравнение $\overline{y}_x = f(x)$ называют уравнением регрессии Y на X; функцию $f(x)$ называют регрессией Y на X, а ее график - линией регрессии.

Если функция регрессии $f(x)$ известна, то можно по значению одной случайной величины прогнозировать значение другой случайной величины. Корреляция называется линейной, если линия регрессии является прямой, т. е. $\overline{y}_x = ax + b$.

Ломаная, соединяющая точки $M_i\left(x_i, \overline{y}_{x_i}\right)$, называется эмпирической (опытной) линией регрессии. Если точки $M_i\left(x_i, \overline{y}_{x_i}\right)$ располагаются около некоторой прямой, то в качестве уравнения теоретической линии регрессии берется $f(x) = ax + b$, где коэффициенты находятся по формулам:

$$a = r_{xy}\frac{\sigma_y}{\sigma_x};\ b = \overline{y} - a\,\overline{x}, \quad (r_{xy} \text{ определен ниже}). \tag{7}$$

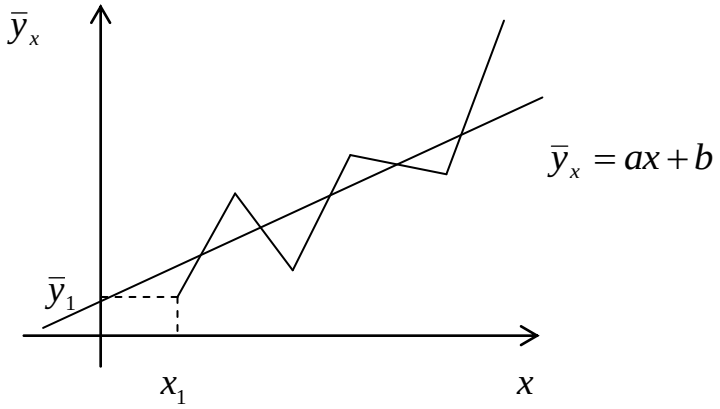


Рис. 6

Ковариацией двух случайных величин X и Y называется числовая характеристика

$$cov(X, Y) = M(X \cdot Y) - M(X) \cdot M(Y).$$

Коэффициентом корреляции между случайными величинами X и Y называется безразмерная величина

$$r_{xy} = \frac{cov(X,Y)}{\sigma_x \cdot \sigma_y}; \qquad (8)$$

где σ_x и σ_y - средние квадратические отклонения величин X и Y.

Коэффициент корреляции r_{xy} характеризует степень тесноты линейной зависимости между случайными величинами X и Y, при этом связь тем теснее, чем ближе $|r_{xy}|$ к единице ($-1 \le r_{xy} \le 1$). Применяется таблица Чеддока для характеристики тесноты связи между случайными величинами X и Y:

Диапазон измерения выборочного $\|r_{xy}\|$	Характер тесноты
0,1-0,3	слабая
0,3-0,5	умеренная
0,5-0,7	заметная
0,7-0,9	высокая
0,9-0,99	линейная

Если $r_{xy} > 0$, то при возрастании одной случайной величины другая имеет тенденцию в среднем возрастать. Если $r_{xy} < 0$, то при возрастании одной случайной величины другая имеет тенденцию в среднем убывать.

Если $r_{xy} = 0$, то линейная корреляционная связь отсутствует, и случайные величины называются некоррелированными. Если $|r_{xy}|\sqrt{n-1} \ge 3$, то связь между случайными величинами X и Y достаточно вероятна.

Чтобы сделать обоснованные выводы о тесноте зависимости между случайными величинами X и Y по опытным данным, нужно установить значимость коэффициента корреляции, т. е. проверить нулевую гипотезу H_0 о том, что $r_{xy} = 0$.

По опытным данным вычисляют критерий проверки

$$T_{набл.} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}. \qquad (9)$$

При заданном уровне значимости α и числу степеней свободы $r = n-2$ находят критическое значение $t_{крит}$ для двусторонней критической области по таблице Стьюдента (смотрите таблицу прил. 3).

Если $|T_{набл}| < t_{крит}$, то выдвинутую гипотезу H_0 принимают, т. е. выборочный коэффициент незначим, а случайные величины X и Y некоррелированы.

Если $|T_{набл}| > t_{крит}$ - гипотезу H_0 отвергают, т. е. выборочный коэффициент корреляции значимо отличается от нуля, а случайные величины коррелированны.

Пример 4. Вычислить выборочный коэффициент корреляции r_{xy}, проверить его значимость и найти уравнение линии регрессии.

X	Y						
	16,5-19,5	19,5-22,5	22,5-25,5	25,5-28,5	28,5-21,5	31,5-34,5	34,5-37,5
97,5-102,5	6	3	1				
102,5-107,5				4	3	2	
107,5-112,5			6	5	2		
112,5-117,5			1	6	3		
117,5-122,5			2	3	9	2	1
122,5-127,5				5	7	3	
127,5-132,5			1		4	4	
132,5-137,5				1	5	1	
137,5-142,5					2	4	4

Решение. Найдем условные средние, соответствующие значению $X = x_i$, по формуле $\overline{y}_{x_i} = \frac{1}{n_{x_i}} \sum_{j=1}^{7} y_j n_{ij}$. Тогда $\overline{y}_{x_1} = \frac{1}{10}(18 \cdot 6 + 21 \cdot 3 + 24 \cdot 1) = 19{,}5$;

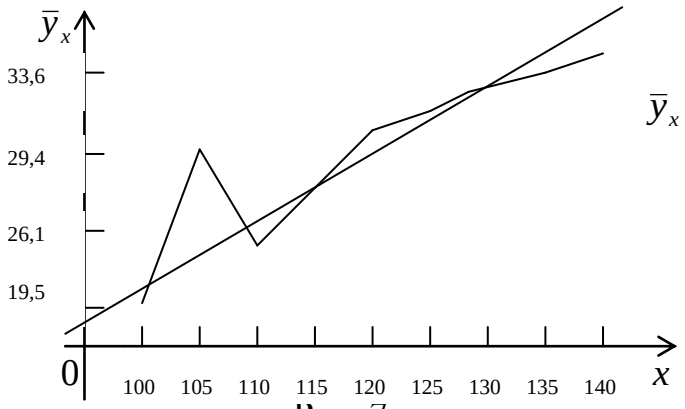
$\overline{y}_{x_2} = \frac{1}{9}(27 \cdot 4 + 30 \cdot 3 + 33 \cdot 2) = 29{,}4$ и т. д.

Составим корреляционную таблицу

X / Y	18	21	24	27	30	33	36	n_{x_i}	$\overline{y}_{x_i}$
100	6	3	1					10	19,5
105				4	3	2		9	29,4
110			6	5	2			13	26,1
115			1	6	3			10	27,6
120			2	3	9	2	1	17	29,5
125				5	7	3		15	29,6
130			1		4	4		9	30,7
135				1	5	1		7	30,0
140					2	4	4	10	33,6
n_{y_j}	6	3	11	24	35	16	5	100	

Контроль расчетов: $n = \sum n_{x_i} = \sum n_{y_j} = 100$ - объем выборки.

Для построения эмпирической линии регрессии точки $M_1(100; 19{,}5)$, $M_2(105; 29{,}4), \ldots, M_9(140; 33{,}6)$ соединим ломаной линией.



$$\bar{y}_x = 0,21x + 3,54$$

Рис. 7

Для нахождения выборочного коэффициента линейной корреляции r_{xy} найдем

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{9} x_i n_{x_i} = \frac{1}{100}\big(100\cdot 10 + 105\cdot 9 + 110\cdot 13 + 115\cdot 10 + 120\cdot 17 + 125\cdot 15 + 130\cdot 9 + 135\cdot 7 + 140\cdot 10\big) = 119,55;$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{7} y_j n_{y_j} = \frac{1}{100}\big(18\cdot 6 + 21\cdot 3 + 24\cdot 11 + 27\cdot 24 + 30\cdot 35 + 33\cdot 16 + 36\cdot 5\big) = 28,41.$$

Вспомогательно найдем

$$\sum_{i=1}^{9}(x_i)^2 n_{x_i} = (100)^2\cdot 10 + (105)^2\cdot 9 + (110)^2\cdot 13 + \ldots + (140)^2\cdot 10 = 1443625;$$

$$\sum_{j=1}^{7}(y_j)^2 n_{y_j} = (18)^2\cdot 18 + (21)^2\cdot 3 + (24)^2\cdot 11 + \ldots + (36)^2\cdot 5 = 82503;$$

$$\sum_{i,j} x_i y_j n_{ij} = 100\cdot 18\cdot 6 + 100\cdot 21\cdot 3 + 100\cdot 24\cdot 1 + 105\cdot 27\cdot 4 + 105\cdot 30\cdot 3 + 105\cdot 33\cdot 2 + 110\cdot 24\cdot 6 + 110\cdot 27\cdot 5 + 110\cdot 30\cdot 2 + 115\cdot 24\cdot 1 + 115\cdot 27\cdot 6 + 115\cdot 30\cdot 3 + 120\cdot 24\cdot 2 + 120\cdot 27\cdot 3 + 120\cdot 30\cdot 9 + 120\cdot 33\cdot 2 + 120\cdot 36\cdot 1 + 125\cdot 27\cdot 5 + 125\cdot 30\cdot 7 + 125\cdot 33\cdot 3 + 130\cdot 24\cdot 1 + 130\cdot 30\cdot 4 + 130\cdot 33\cdot 4 + 135\cdot 27\cdot 1 + 135\cdot 30\cdot 5 + 135\cdot 33\cdot 1 + 140\cdot 30\cdot 2 + 140\cdot 33\cdot 4 + 140\cdot 36\cdot 4 = 342600.$$

Тогда $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{9}(x_i)^2 n_{x_i} - (\bar{x})^2 = \frac{1}{100}1443625 - (119,55)^2 = 144,05 \Rightarrow$

$$\Rightarrow \sigma_x = \sqrt{144,05} = 12,002.$$

$$\sigma_y^2 = \frac{1}{n}\sum_{j=1}^{7}(y_j)^2 n_{y_j} - (\bar{y})^2 = \frac{1}{100}82503 - (28,41)^2 = 17,9 \Rightarrow$$

$$\Rightarrow \sigma_y = \sqrt{17,9} = 4,23.$$

Определим ковариацию между X и Y по формуле

$$cov(X,Y) = \frac{1}{n}\sum_{i,j} x_i y_j n_{ij} - \bar{x}\,\bar{y} = \frac{1}{100}342600 - 119,55\cdot 28,41 = 29,585.$$

Находим коэффициент корреляции по формуле (8):

$$r_{xy} = \frac{29{,}585}{12{,}002 \cdot 4{,}29} = 0{,}59.$$

Имеем $|r_{xy}|\sqrt{n-1} = 0{,}59\sqrt{99} = 5{,}87 > 3$, следовательно, связь между случайными величинами X и Y достаточно вероятна.

Для проверки значимости коэффициента корреляции проверим нулевую гипотезу $H_0 : r_{xy} = 0$; конкурирующая гипотеза $H_1 : r_{xy} \neq 0$.

Найдем по опытным данным величину

$$T_{набл} = \frac{0{,}59\sqrt{98}}{\sqrt{1-(0{,}59)^2}} = 8{,}99.$$

Найдем критическое значение $t_{крит}$ по таблице критерия Стьюдента (прил. 3) при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $r = n-2 = 98 \Rightarrow t_{крит} = 1{,}98$. Тогда $|T_{набл}| > t_{крит}$, поэтому гипотезу H_0 отвергаем и принимаем гипотезу H_1, т. е. случайные величины X и Y коррелированы.

По виду эмпирической линии регрессии можно предположить, что между случайными величинами существует линейная корреляция, т. е. $\overline{y}_x = ax + b$. Находим коэффициенты a и b по формулам (7):

$$a = 0{,}59\frac{4{,}23}{12{,}002} = 0{,}21; \quad b = 28{,}41 - 0{,}21 \cdot 119{,}55 = 3{,}54.$$

Тогда уравнение линейной регрессии

$$\overline{y}_x = 0{,}21x + 3{,}54.$$

Для построения полученной прямой возьмем две точки

x	110	140
$\overline{y}_x$	26,4	32,7

График прямой $\overline{y}_x$ достаточно близко расположен по отношению к опытной линии регрессии. Коэффициент корреляции $r_{xy} = 0{,}59$ показывает, что зависимость между случайными величинами X и Y заметная и с увеличением значений одной случайной величины значения другой случайной величины имеют тенденцию в среднем увеличиваться.

Задачи для контрольной работы № 12

В задачах 1-10 произведены измерения отклонений размера деталей от стандарта. Результаты сведены в таблицу. Построить гистограмму, выдвинуть гипотезу о законе распределения исследуемой случайной величины и с помощью критерия согласия Пирсона при заданном уровне значимости α проверить данную гипотезу.

1. $\alpha = 0{,}01$

Границы отклонений	7-9	9-11	11-13	13-15	15-17
Число деталей	5	23	41	20	11

2. $\alpha = 0{,}01$

Границы отклонений	2-6	6-10	10-14	14-18	18-22
Число деталей	7	15	29	18	11

3. $\alpha = 0{,}05$

Границы отклонений	5-11	11-17	17-23	23-29	29-35
Число деталей	7	12	18	15	8

4. $\alpha = 0{,}01$

Границы отклонений	8-10	10-12	12-14	14-16	16-18
Число деталей	7	17	33	14	7

5. $\alpha = 0{,}01$

Границы отклонений	20-24	24-28	28-32	32-36	36-40
Число деталей	10	21	30	17	12

6. $\alpha = 0{,}05$

Границы отклонений	0-6	6-12	12-18	18-24	24-30
Число деталей	5	11	23	13	8

7. $\alpha = 0{,}01$

Границы отклонений	4-8	8-12	12-16	16-20	20-24
Число деталей	7	25	38	21	9

8. $\alpha = 0{,}05$

Границы отклонений	2-14	14-26	26-38	38-50	50-62
Число деталей	6	13	19	15	7

9. $\alpha = 0{,}01$

Границы отклонений	8-12	12-16	16-20	20-24	24-28	28-32
Число деталей	6	11	25	13	4	2

10. $\alpha = 0{,}05$

Границы отклонений	1-5	5-9	9-13	13-17	17-21	21-25
Число деталей	6	10	17	12	4	1

В задачах 11-20 требуется:

а) Найти условные средние $\overline{y}_x$ и построить эмпирическую линию регрессии Y на X.

б) Вычислить выборочный коэффициент корреляции, проверить его значимость и сделать вывод о связи случайных величин X и Y.

в) Определить линейную модель регрессии и построить ее график.

11. В таблице дано распределение 100 проб руды по содержанию окиси железа X (%) и закиси железа Y (%):

Y	X					
	40-50	50-60	60-70	70-80	80-90	90-100
0-6					4	6
6-12			6	6	8	
12-18	1	2	14	3		
18-24	6	18	2			
24-30	4	10	2			
30-36	6	2				

12. В таблице дано распределение 60 предприятий по величине основных фондов X (млрд. руб.) и себестоимости продукции Y (млн. руб.):

Y	X				
	15-30	30-45	45-60	60-75	75-90
16-24			1	4	1
24-32			7	7	2
32-40		4	12	2	
40-48		8	6		
48-56	2	4			

13. В таблице дано распределение 200 растений по весу X (г.) каждого из них и по весу Y (г.) его семян:

Y	X				
	40	50	60	70	80
15	5				
20	7	4	8		
25		16	20	11	
30		23	32	29	9
35			27	2	7

14. При обследовании детей четырехлетнего возраста получено распределение их по росту X (см) и весу Y (кг):

Y	X					
	98-100	100-102	104-104	104-106	106-108	108-110
15,5-16,5	2	3	1			
16,5-17,5	3	6	4	1		
17,5-18,5		4	13	14	10	
18,5-19,5			5	10	8	6
19,5-20,5				2	5	3

15. В таблице дано распределение 100 прямоугольных чугунных плиток по длине X (см) и весу Y (кг):

Y	X					
	30	40	50	60	70	80
30	3	6	12	7	2	
36		2	8	10	2	1
42			1	4	16	6
48				2	3	5
54					4	6

16. В таблице дано распределение 100 заводов по производственным средствам X (млрд. руб.) и суточной выработке Y (т.):

Y	X							
	10	15	20	25	30	35	40	45
15	2	4						
20	1	6	5	8		3		
25		3	13	4	2	1		
30			4	11	5	7		
35				2	1	4	3	1
40				1	2	5	1	1

17. В таблице дано распределение 100 проб руды по глубине залегания X (см) и содержанию окиси железа Y (%):

Y	X				
	30	40	50	60	70
12	8	8			4
18	7	16	7		
24		15	10	1	
30		4	9	5	3
36				2	1

18. В таблице дано распределение 50 магазинов края по уровню издержек обращения X (%) и годовому объему товарооборота Y (млн. руб.):

Y	X				
	4-6	6-8	8-10	10-12	12-14
0,5-2,0			2	3	1
2,0-3,5		4	5	1	2
3,5-5,0		8	5	5	
5,0-6,5	3	8			
6,5-8,0	2	1			

19. В таблице дано распределение 100 заводов по объему валовой продукции X (млн. руб.) и среднесписочной численности работающих Y (тыс. человек):

Y	X				
	20	30	40	50	60
1	8	2			
3	12	20	8		
5			10	1	
7			9	6	2
9			10	4	8

20. В таблице дано распределение 100 магазинов по величине товарооборота X (млн. руб.) и размеру торговой площади магазина Y (кв. м.):

Y	X				
	1,0-1,5	1,5-2,0	2,0-2,5	2,5-3,0	3,0-3,5
100-150	4				
150-200	12	4	2		
200-250	2	9	10	4	
250-300		9	18	9	3
300-350				11	3

Приложение 1

Критические точки распределения Пирсона

Число степеней свободы r	Уровень значимости α										
	0,99	0,95	0,90	0,50	0,25	0,10	0,05	0,025	0,01	0,005	0,001
1	0,0002	0,004	0,02	0,46	1,32	2,71	3,84	5,20	6,63	7,88	10,8
2	0,02	0,10	0,21	1,39	2,77	4,61	5,99	7,38	9,21	10,6	13,8
3	0,12	0,35	0,58	2,37	4,11	6,25	7,81	9,35	11,3	12,8	16,3
4	0,30	0,71	1,06	3,36	5,39	7,78	9,49	11,1	13,3	14,9	18,5
5	0,55	1,15	1,61	4,35	6,63	9,24	11,1	12,8	15,1	16,7	20,5
6	0,87	1,64	2,20	5,35	7,84	10,6	12,6	14,4	16,8	18,5	22,5
7	1,24	2,17	2,83	6,35	9,04	12,0	14,1	16,0	18,5	20,3	24,3
8	1,65	2,73	3,49	7,34	10,2	13,4	15,5	17,5	20,1	22,0	26,1
9	2,09	3,33	4,17	8,34	11,4	14,7	16,9	19,0	21,7	23,6	27,9
10	2,56	3,94	4,87	9,34	12,5	16,0	18,3	20,5	23,2	25,2	29,6
11	3,05	4,57	5,58	10,3	13,7	17,3	19,7	21,9	24,7	26,8	31,3
12	3,57	5,23	6,30	11,3	14,8	18,5	21,0	23,3	26,2	28,3	32,9
13	4,11	5,89	7,04	12,3	16,0	19,8	22,4	24,7	27,7	29,8	34,5
14	4,66	6,57	7,79	13,3	17,1	21,1	23,7	26,1	29,1	31,3	36,1
15	5,23	7,26	8,55	14,3	18,2	22,3	25,0	27,5	30,6	32,8	37,7
16	5,81	7,96	9,31	15,3	19,4	23,5	26,3	28,8	32,0	34,3	39,3
17	6,41	8,67	10,1	16,3	20,5	24,8	27,6	30,2	33,4	35,7	40,8
18	7,01	9,39	10,9	17,3	21,6	26,0	28,9	31,5	34,8	37,2	42,3
19	7,63	10,1	11,7	18,3	22,7	27,2	30,1	32,9	36,2	38,6	43,8
20	8,26	10,9	12,4	19,3	23,8	28,4	31,4	34,2	37,6	40,0	45,3
21	8,90	11,6	13,2	20,3	24,9	29,6	32,7	35,5	38,9	41,4	46,8
22	9,54	12,3	14,0	21,3	26,0	30,8	33,9	36,8	40,3	42,8	48,3
23	10,2	13,1	14,8	22,3	27,1	32,0	35,2	38,1	41,6	44,2	49,7
24	10,9	13,8	15,7	23,3	28,2	33,2	36,4	39,4	43,0	45,6	51,2
25	11,5	14,6	16,5	24,3	29,3	34,4	37,7	40,6	44,3	46,9	52,6
26	12,2	15,4	17,3	25,3	30,4	35,6	38,9	41,9	45,6	48,3	54,1
27	12,9	16,2	18,1	26,3	31,5	36,7	40,1	43,2	47,0	49,6	55,5
28	13,6	16,9	18,9	27,3	32,6	37,9	41,3	44,5	48,3	51,0	56,9
29	14,3	17,7	19,8	28,3	33,7	39,1	42,6	45,7	49,6	52,3	58,3
30	15,0	18,5	20,6	29,3	34,8	43	43,8	47,0	50,9	53,7	59,7

Приложение 2

Таблица значений функции $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}d\,z$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0,1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0754
0,2	0793	0832	0871	0909	0948	0987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1363	1406	1443	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0,5	1915	1950	1986	2019	2054	2088	2123	2157	2190	2224
0,6	2258	2291	2324	2357	2389	2421	2454	2486	2518	2549
0,7	2580	2612	2642	2673	2704	2734	2764	2794	2823	2852
0,8	2881	2910	2939	2967	2996	3023	3051	3078	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1,0	0,3413	3438	3461	3485	3508	3531	3554	3577	3599	3627
1,1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1,2	3849	3869	3888	3906	3925	3940	3962	3980	3897	4016
1,3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1,4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1,5	4332	4345	4357	4370	4382	4394	4406	4418	4430	4441
1,6	4452	4463	4474	4485	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4572	4582	4591	4599	4608	4616	4625	4633
1,8	4641	4648	4656	4664	4671	4678	4686	4693	4699	4706
1,9	4713	4719	4726	4732	4738	4744	4750	4756	4762	4767
2,0	0,4772	4778	4383	4788	4793	4798	4803	4808	4812	4817
2,1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4858
2,2	4861	4864	4868	4881	4875	4878	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2,4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2,5	4938	4940	4941	4943	4945	4946	4941	4949	4951	4952
2,6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2,8	4974	4975	4976	4977	4977	4978	4979	4980	4980	4981
2,9	0,49813	2,91		0,49819	2,92	0,49825	0,49825	2,93	2,93	0,49831
2,94	0,49836		2,95	0,49841		2,96	0,49846		2,97	0,49851
2,98	49856		2,99	49861		3,0	0,49865		3,1	49903
3,2	49931		3,3	49952		3,4	0,49966		3,5	49977
3,6	49984		3,7	49989		3,8	0,49993		3,9	49995
4,0	0,499968		4,5	0,499997		5,0	0,49999997			

Если $x>5$, то $\Phi(x)$ полагают равной 0,5.

$\Phi(-x)=-\Phi(x)$ - нечетная функция.

$\Phi(2{,}54)=0{,}4945.$

$\Phi(-2{,}54)=-\Phi(2{,}54)=-0{,}4945.$

Приложение 3

Критические точки распределения Стьюдента

Число степеней свободы r	Уровень значимости α (двусторонняя критическая область)			
	0,10	0,05	0,02	0,01
1	6,31	12,7	31,82	63,7
2	2,92	4,30	6,97	9,92
3	2,35	3,18	4,54	5,84
4	2,13	2,78	3,75	4,60
5	2,01	2,57	3,37	4,03
6	1,94	2,45	3,14	3,71
7	1,89	2,36	3,00	3,50
8	1,86	2,31	2,90	3,36
9	1,83	2,26	2,82	3,25
10	1,81	2,23	2,76	3,17
11	1,80	2,20	2,72	3,11
12	1,78	2,18	2,68	3,05
13	1,77	2,16	2,65	3,01
14	1,76	2,14	2,62	2,98
15	1,75	2,13	2,60	2,95
16	1,75	2,12	2,58	2,92
17	1,74	2,11	2,57	2,90
18	1,73	2,10	2,55	2,88
19	1,73	2,09	2,54	2,86
20	1,73	2,09	2,53	2,85
21	1,72	2,08	2,52	2,83
22	1,72	2,07	2,51	2,82
23	1,71	2,07	2,50	2,81
25	1,71	2,06	2,49	2,79
27	1,71	2,05	2,47	2,77
29	1,70	2,05	2,46	2,76
30	1,70	2,04	2,46	2,75
40	1,68	2,02	2,42	2,70
60	1,67	2,00	2,39	2,66
120	1,66	1,98	2,36	2,62
∞	1,64	1,96	2,33	2,58

Список литературы

1. *Гмурман В. Е.* Теория вероятностей и математическая статистика. – М.: Высшая школа, 1977. – с. 286.
2. *Гмурман В. Е.* Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 1979. – с. 400.
3. *Демиденко Е. З.* Линейная и нелинейная регрессия. – М.: Финансы и статистика, 1981.
4. *Колмогоров А. Н. и др.* Введение в теорию вероятностей. – М.: Наука, 1982. – с. 64.
5. *Чистяков В. П.* Курс теории вероятностей. – М.: Наука, 1982. – с. 207.