Министерство образования Республики Беларусь
БЕЛОРУССКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра «Инженерная математика»

О.В. Дубровина
Н.К. Прихач
В.М. Романчак

ПРИКЛАДНАЯ МАТЕМАТИКА

Методическое пособие
по выполнению практических и лабораторных работ
для студентов специальности 1-54 01 01 «Метрология,
стандартизация и сертификация» заочной формы обучения

Учебное электронное издание

М и н с к 2 0 1 0

УДК 51-7(075.8)
ББК 22.1я7
Д 80

А в т о р ы :

О.В. Дубровина, Н.К. Прихач, В.М. Романчак

Р е ц е н з е н т ы :

П.С. Серенков, заведующий кафедрой «Стандартизация, метрология и информационные системы» БНТУ, доктор технических наук;

П.М. Лаппо, доцент кафедры «Теория вероятностей и математическая статистика» БГУ.

Данное методическое пособие предназначено для обучения студентов заочной формы обучения специальности 1-54 01 01 «Метрология, стандартизация и сертификация». В пособии в компактной форме предоставлены основные теоретические сведения по статистическим методам обработки эксперимента. Методические материалы содержат материалы по обработке результатов эксперимента, а также необходимые теоретические сведения по каждой теме и задания для самостоятельного изучения. Более сложные задания выполняются в пакете Statistica.

Белорусский национальный технический университет
пр-т Независимости, 65, г. Минск, Республика Беларусь
Тел.(017)292-77-52 факс (017)292-91-37
E-mail: emd@bntu.by
http://www.bntu.by/ru/struktura/facult/psf/chairs/im/
Регистрационный № БНТУ/ПСФ85-8.2010

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В методическом пособии разбираются наиболее важные темы курса «Прикладная математика» для студентов заочной формы обучения специальности 1-54 01 01 «Метрология, стандартизация и сертификация».

С помощью пособия студенты могут самостоятельно освоить практические занятия и подготовиться к лабораторным работам.

Работа включает два практических занятия и три лабораторных работы.

Практические занятия содержат материалы по обработке результатов эксперимента в случае нормального распределения, разобраны темы корреляционный и регрессионный анализ.

Лабораторные работы содержат необходимые сведения по работе с пакетом STATISTICA, Рассматриваются вопросы вычисление выборочных характеристик, критерий согласия Пирсона, линейная регрессия, дисперсионный анализ.

Каждая лабораторная работа содержит теоретическую часть, примеры выполнения заданий и задания для самостоятельной работы.

ПРАКТИЧЕСКИЕ ЗАНЯТИЯ

Практическое занятие №1
ОБРАБОТКА РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА В СЛУЧАЕ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

1.1. Основные понятия

При обработке экспериментальных данных, при решении многих практических задач для характеристики свойств наблюдаемых случайных величин (СВ) и для проведения теоретических выкладок приходится делать предположение о виде законов распределения этих величин (нормальном, показательном, Пуассона, биномиальном и т.д.) или о соотношении между параметрами распределений. Такие предположения называют *гипотезами*. Приняв гипотезу, из нее получают определенные теоретические данные и проверяют, насколько они согласуются с результатами опыта.

Выбор распределения по опытным данным может быть сделан из следующих соображений:

- исходя из физической природы исследуемого объекта;
- по виду гистограммы или полигона относительных частот;
- по опытным данным ранее проведенных исследований;
- с помощью графического представления эмпирической функции;
- с помощью критериев согласия и т.д.

Нормальное распределение задается функцией, называемой плотностью распределения вероятности

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2} \right), \qquad (1.1)$$

где m – математическое ожидание СВ X;

σ – среднее квадратичное отклонение СВ X.

Таким образом, нормальное распределение СВ X определяется двумя параметрами: $M(X) = m$ и $\sigma = \sqrt{D(X)}$, где $D(X)$ – дисперсия СВ X ($X \in N(m, \sigma)$).

График плотности вероятности называется нормальной кривой (кривой Гаусса). СВ $X \in N(0,1)$ называют *стандартизированной нормальной величиной*.

Пусть при проведении n опытов некоторая СВ X принимает значения $x_1, x_2, \ldots, x_n$. Выдвинуто предположение, что СВ $X \in N(m,\sigma)$, причем m и σ неизвестны. Для построения теоретико-вероятностной модели необходимо на основании выборки оценить математическое ожидание m и среднее квадратичное отклонение σ.

Изучение случайных величин обычно начинают с группировки статистических данных, и.е. с разбиения интервала наблюдаемых значений СВ X на N подынтервалов равной длины $h = \dfrac{x_{\max} - x_{\min}}{N}$ и подсчета эмпирических частот r_k, $k = \overline{1, N}$ попадания значений СВ X в соответствующие подынтервалы. Обычно $5 \le N \le 20$.

1.2. Оценки случайных величин

Различают точечные и интервальные оценки. *Точечная* оценка $\hat{\theta}$ некоторого параметра θ определяется по результатам выборки одним числом. Для того, чтобы точечная оценка была «хорошей» необходимо, чтобы она была состоятельной, несмещенной, эффективной. Задача оценивания параметров m и σ сводится к нахождению таких функций от выборки $\hat{m}$ и $\hat{\sigma}$, которые могут быть использованы для приближенного определения параметров m и σ. В качестве точечных оценок для m и σ нормально распределенной СВ X ($x \in N(m,\sigma)$) принимаются:

$$\hat{m} = \bar{x} = \frac{1}{n}\sum_{k=1}^{n} n_k x_k \,, \tag{1.2}$$

$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n} n_k \cdot (x_k - \bar{x})^2} \,. \tag{1.3}$$

Точечные оценки не указывают величины ошибки, которая совершается при замене m и σ их приближенными значениями $\hat{m}$ и $\hat{\sigma}$. Поэтому иногда выгоднее пользоваться интервальной оценкой, которая определяется двумя числами $\bar{\theta}_1$ и $\bar{\theta}_2$ – концами интервала, накрывающего оцениваемый параметр θ с заданной вероятностью (надежностью).

Пусть $\hat{\theta}$ – точечная оценка параметра θ. Она тем лучше, чем меньше разность $\left|\theta-\hat{\theta}\right|$. Тогда в качестве характеристики точности оценки можно взять некоторое $\varepsilon>0$ такое, что $\left|\theta-\hat{\theta}\right|<\varepsilon$.

Доверительной вероятностью оценки называется вероятность $\gamma=1-\alpha$ выполнения неравенства $\left|\theta-\hat{\theta}\right|<\varepsilon$. *Доверительный интервал* – это интервал, который накрывает неизвестный параметр θ с заданной надежностью $\gamma=1-\alpha$. Чем меньше длина доверительного интервала, тем точнее оценка.

При неизвестном σ доверительный интервал для математического ожидания m СВ $X\in N(m,\sigma)$ имеет вид:

$$\bar{x}-t_\gamma\cdot\frac{s}{\sqrt{n}}<m<\bar{x}+t_\gamma\cdot\frac{s}{\sqrt{n}}, \qquad (1.4)$$

где величина t_γ определяется по таблицам по заданному уровню значимости α (либо надежности $\gamma=1-\alpha$) и объему выборки n.

Доверительный интервал для σ задается неравенствами

$$s\cdot(1-q)<\sigma<s\cdot(1+q), \text{ если } q<1, \qquad (1.5)$$

либо

$$0<\sigma<s\cdot(1+q), \text{ если } q>1. \qquad (1.6)$$

Величина q определяется по таблице доверительных интервалов для σ по доверительной вероятности $\gamma=1-\alpha$ и объему выборки n.

Медианой M_e называется вариант, который приходится на середину ряда распределения. При вычислении медианы дискретного ряда рассматриваются два случая: объем совокупности четный и нечетный. В первом случае применяется формула $M_e=x_m$, если $n=2m-1$ (n – объем совокупности). Если $n=2m$, то медиана: $M_e=\frac{1}{2}(x_m+x_{m+1})$.

Модой M_o называется вариант, который наиболее часто встречается. Мода – это вариант, которому соответствует наибольшая частота или частоты.

Эмпирической функцией распределения СВ X называют функцию

$$F^*(x) = \frac{n_x}{n},$$

где n_x – число значений x_i меньших, чем x;

n – объем выборки.

Эмпирическая функция распределения используется в качестве оценки функции распределения.

Для наглядности данные выборки можно представить графически в виде гистограммы, а также полигона относительных частот. Для построения гистограммы интервал наблюдаемых значений СВ X разбивается над подынтервалы равной длины $h > 0$, на каждом из которых строится прямоугольник с высотой $\frac{n_i}{n \cdot h}$, где n_i – число значений СВ X из выборки, попадающих в рассматриваемый подынтервал. Ломаная, соединяющая точки пересечения середин подынтервалов с соответствующими высотами $\frac{n_i}{n \cdot h}$, образуют полигон относительных частот.

Если форма гистограммы или полигона относительных частот напоминает кривую Гаусса, то можно выдвинуть гипотезу о нормальном распределении СВ X. Для проверки того, что СВ $X \in N(m, \sigma)$ можно использовать следующие характеристики: *асимметрию* $A_s = \frac{\mu_3}{\sigma^3}$ и *эксцесс* $E_k = \frac{\mu_4}{\sigma^4} - 3$, где $\mu_k = M\left((X - M(X))^k\right)$.

Для нормального распределения $A_s = 0$, $E_k = 0$. По данным выборки объема n можно найти точечные оценки A_s и E_k:

$$\hat{A}_s = \frac{\sum_{k=1}^{N} r_k \cdot (\bar{x}_k - \bar{x})^3}{n \cdot s^3}, \quad \hat{E}_k = \frac{\sum_{k=1}^{N} r_k \cdot (\bar{x}_k - \bar{x})^4}{n \cdot s^4} - 3, \qquad (1.7)$$

где $\bar{x}_k = x_{\min} + \frac{2 \cdot k - 1}{2} \cdot h$, а также средние квадратичные ошибки и их определения

$$S_A = \sqrt{\frac{6 \cdot (n-1)}{(n+1) \cdot (n+3)}};\ S_E = \sqrt{\frac{24 \cdot n \cdot (n-2) \cdot (n-3)}{(n-1)^2 (n+3)(n+5)}}. \qquad (1.8)$$

Гипотеза о нормальности закона распределения СВ X выдвигается, если $\left|\hat{A}_s\right| < 3 \cdot S_A$ и $\left|\hat{E}_k\right| < 3 \cdot S_E$. В противном случае она отвергается.

После предварительного выбора закона распределения рекомендуется применять строгие критерии согласия.

1.3. Критерий χ^2-Пирсона

При проверке гипотезы о нормальном распределения СВ X с помощью критерия χ^2-Пирсона поступают следующим образом:

1) вычисляют вероятности p_i попадания СВ X в подынтервалы $[x_{i-1}, x_i[$, $i = 1,2,\ldots, N$;
2) вычисляют выборочную статистику

$$\chi^2_{набл} = \sum_{i=1}^{N} \frac{(r_i - np_i)^2}{np_i}; \qquad (1.9)$$

3) сравнивают $\chi^2_{набл}$ с квантилем $\chi^2_{\alpha,\nu}$, определяемым по таблицам по заданному уровню значимости α и числу степеней свободы $\nu = N - r - 1$, где r – число параметров предполагаемого распределения СВ X. Если $\chi^2_{набл} \le \chi^2_{\alpha,\nu}$, то считают, что нет оснований для отклонения проверяемой гипотезы. В противном случае гипотеза отклоняется.

1.4. Контрольный пример

Из генеральной совокупности извлечена выборка, представленная в виде статистического ряда. Требуется:

1) вычислить выборочное среднее $\bar{x}$, выборочную дисперсию σ_B^2, исправленную выборочную дисперсию s^2 и среднее квадратичное отклонение s;

2) найти с доверительной вероятностью $\gamma = 0{,}99$ доверительный интервал для математического ожидания, а также доверительный интервал для $\sigma(x)$;

3) найти размах варьирования и среднее абсолютное отклонение;

4) вычислить моду и медиану;

5) построить эмпирическую функцию распределения;

6) проверить, согласуются ли выборочные данные с гипотезой о нормальном распределении СВ X графически, с помощью асимметрии и эксцесса и с помощью критерия согласия Пирсона при уровне значимости $\alpha = 0.01$, разбив отрезок $[x_{\min}, x_{\max}]$ на $N = 5$ интервалов одинаковой длины $[x'_i, x'_{i+1}]$ с границами $x'_i = x_{\min} + \dfrac{x_{\max} - x_{\min}}{N} \cdot i, \qquad i = \overline{1, N}$.

x_i	4	6	7	8	10	11	12
n_i	3	7	10	11	9	6	4

Решение.

1) Объем выборки равен $n = \sum n_i = 3+7+10+11+9+6+4 = 50$. Выборочное среднее и дисперсия определяются по формулам (1.2), (1.3)

$$\bar{x} = \frac{1}{n}\sum x_i n_i = \frac{1}{50}\left(4\cdot 3 + 6\cdot 7 + 7\cdot 10 + 8\cdot 11 + 10\cdot 9 + 11\cdot 6 + 12\cdot 4\right) = 8{,}32;$$

$$\sigma_B^2 = \frac{1}{n}\sum x_i^2 \cdot n_i - \bar{x}^2 = \frac{1}{50}\left(4^2\cdot 3 + 6^2\cdot 7 + 7^2\cdot 10 + 8^2\cdot 11 + 10^2\cdot 9 + 11^2\cdot 6 + 12^2\cdot 4\right) - 8{,}32^2 = 4{,}7$$

Исправленная выборочная дисперсия равна $s^2 = \dfrac{n}{n-1}\cdot\sigma_B^2 = \dfrac{50}{49}\cdot 4{,}7 = 4{,}79$.

Исправленное среднее квадратичное отклонение будет $s = \sqrt{4{,}79} = 2{,}19$.

2) Доверительный интервал для математического ожидания найдем по формуле (1.4). Значение t_γ определим из таблицы по доверительной вероятности $\gamma = 0{,}99$ и объему выборки $n = 50$: $t_\gamma = 2{,}679$. Тогда доверительный интервал имеет вид:

$$7{,}49 < m < 9{,}15.$$

Доверительный интервал для дисперсии определим по формуле (1.5): $q = q(50, 0.99) = 0.30$ ($q < 1$). Тогда границы интервала принимают вид:

$s(1-q)=2.19\cdot(1-0.3)=1,53$; $s(1+q)=2.19\cdot(1+0.3)=2,85$, т.е.

$$1,53<\sigma<2,85.$$

3) Размах варьирования находится по формуле $\omega=x_{\max}-x_{\min}=12-4=8$. Среднее абсолютное отклонение

$$\theta=\frac{\sum_{i=1}^{n}n_i(x_i-\bar{x})}{\sum_{i=1}^{n}n_i};$$

$$\theta=\frac{1}{50}\left(3\cdot(4-8,32)+7\cdot(6-8,32)+10\cdot(7-8,32)+11\cdot(8-8,32)+9\cdot(10-8,32)+6\cdot(11-8,32)+11\cdot(12-8,32)\right)$$

$$\theta=0.$$

4) Вычислим медиану и моду. Так как $n=50=2\cdot25$, значит

$$M_e=\frac{1}{2}(x_{25}+x_{26})=\frac{1}{2}(8+8)=8.$$

Мода $M_o=8$.

5) Согласно определению эмпирической функции распределения, ее значение при любом x равно $F^*(x)=\frac{n_x}{\sum_i n_i}$, где n_x – количество элементов x_i выборки, меньших чем x.

Например, при $x=3$ имеем $n_x=0,\ F^*(3)=0$;

при $x=5$ $\quad n_x=3,\ F^*(x)=3/50=0,06$;

при $x=6,5$ $\quad n_x=3+7=10,\ F^*(x)=10/50=0,2$;

при $x=7,5$ $\quad F^*(x)=20/50=0,4$;

при $x=9$ $\quad F^*(x)=31/50=0,62$;

при $x=10,5$ $\quad F^*(x)=40/50=0,8$;

при $x=11,5$ $\quad F^*(x)=46/50=0,92$;

при $x>12$ $\quad F^*(x)=1$.

Итак, эмпирическая функция распределения $F^*(x)$ имеет вид:

$$F^*(x)=\begin{cases}0, & x\le 4,\\ 0{,}06 & 4<x\le 6,\\ 0{,}2 & 6<x\le 7,\\ 0{,}4 & 7<x\le 8,\\ 0{,}62 & 8<x\le 10,\\ 0{,}8 & 10<x\le 11,\\ 0{,}92 & 11<x\le 12,\\ 1 & x>12.\end{cases}$$

6) Из статистического ряда видно, что $x_{\min}=4$, $x_{\max}=12$, поэтому $\frac{x_{\max}-x_{\min}}{5}=\frac{12-4}{5}=1.6$. Границы интервалов будут $x'_0=4; x'_1=4+1.6=5.6;$ $x'_2=5.6+1.6=7.2; x'_3=7.2+1.6=8.8; x'_4=8.8+1.6=10.4;\ x'_5=10.4+1.6=12$. Частота r_i интервала $[x'_i, x'_{i+1}]$ $\left(i=\overline{0,4}\right)$ подсчитывается с помощью ряда как число наблюдений, попавших в интервал. Так в первый $(i=0)$ интервал $[4;5.6]$ попало 3 значения, во второй $[5.6;\ 7.2]$ - 7+10=17 значений. Аналогично, $r_2=11, r_3=9, r_4=10$. Сведем полученные данные в таблицу:

$x'_i-x'_{i+1}$	4 – 5.6	5.6 – 7.2	7.2 – 8.8	8.8 – 10.4	10.4 - 12
r_i	3	17	11	9	10

Найдем точечные оценки асимметрии и эксцесса. Применим формулы (7), предварительно вычислив величины $\bar{x}_k$: $\bar{x}_1=4+\frac{2-1}{2}\cdot 1.6=4.8$, $\bar{x}_2=4+\frac{2\cdot 2-1}{2}\cdot 1.6=6.4$, $\bar{x}_3=4+\frac{2\cdot 3-1}{2}\cdot 1.6=8$, $\bar{x}_4=4+\frac{2\cdot 4-1}{2}\cdot 1.6=9.6$, $\bar{x}_5=4+\frac{2\cdot 5-1}{2}\cdot 1.6=11.2$.

Отсюда

$$\hat{A}_s=\frac{3\cdot(4.8-8.32)^3+17\cdot(6.4-8.32)^3+11\cdot(8-8.32)^3+9\cdot(9.6-8.32)^3+10\cdot(11.2-8.32)^3}{50\cdot 2.19^3}=0.012;$$

$$\hat{E}_k = \frac{3\cdot(4.8-8.32)^4 + 17\cdot(6.4-8.32)^4 + 11\cdot(8-8.32)^4 + 9\cdot(9.6-8.32)^4 + 10\cdot(11.2-8.32)^4}{50\cdot 2.19^4} - 3 = -1.78$$

Теперь по формулам (1.8) вычислим их средние квадратичные ошибки:

$$S_A = \sqrt{\frac{6\cdot(50-1)}{(50+1)(50+3)}} = 0.33, \qquad S_E = \sqrt{\frac{24\cdot 50\cdot(50-2)(50-3)}{(50-1)^2(50+3)(50+5)}} = 0.62.$$

Так как $\left|\hat{A}_s\right| < 3\cdot S_A$ (0.012<0.99) и $\left|\hat{E}_k\right| < 3\cdot S_E$ (1.78<1.86), то можно сделать предположение, что гипотеза о нормальном распределении СВ X может быть принята.

Проверим данное утверждение с помощью критерия согласия Пирсона. Найдем теоретические вероятности p_i по формуле

$$p_i = P(z_i < z < z_{i+1}) = \Phi\left(\frac{x'_{i+1} - \bar{x}}{\sigma_B}\right) - \Phi\left(\frac{x'_i - \bar{x}}{\sigma_B}\right),$$

где $\Phi(x)$ – функция Лапласа, значения которой взяты из приложения (табл. П1). Результаты вычислений сведем в таблицу:

i	x'_i	x'_{i+1}	$x'_i - \bar{x}$	$x'_{i+1} - \bar{x}$	z_i	z_{i+1}	$\Phi(z_i)$	$\Phi(z_{i+1})$	p_i
1	4	5.6	–	-2,72	–	-1,26	-0,5	-0,3962	0,1038
2	5.6	7.2	-2,72	-1,12	-1,26	-0,53	-0,3962	-0,2019	0,1943
3	7.2	8.8	-1,12	0,48	-0,52	0,22	-0,2019	0,0871	0,289
4	8.8	10.4	0,48	2,08	0,22	0,96	0,0871	0,3315	0,2444
5	10.4	12	2,08	–	0,96	–	0,3315	0,5	0,1685

Найдем теоретические частоты $n'_i = n\cdot p_i$. Получим столбец:

n'_i
5,19
9,715
14,45
12,22
8,425

Вычислим наблюдаемое значение критерия Пирсона. Для этого составим следующую расчетную таблицу:

i	n_i	n_i'	$n_i - n_i'$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
1	3	5,19	-2,19	4,7961	0,924
2	17	9,715	7,285	53,07123	5,463
3	11	14,45	-3,45	11,9025	0,824
4	9	12,22	-3,22	10,3684	0,848
5	10	8,425	1,575	2,480625	0,294
$\sum_i$	**50**	**50**			**8,354**

По таблице критических точек распределения χ^2, уровню значимости $\alpha = 0.01$ и числу степеней свободы $\nu = 5 - 2 - 1 = 2$ находим $\chi^2_{\alpha,\nu} = 9.2$. Так как $\chi^2_{набл} = 8.354 < \chi^2_{\alpha,\nu}$, то гипотеза о нормальном распределении принимается.

1.5. Контрольное задание

Из генеральной совокупности извлечена выборка, представленная в виде статистического ряда. Требуется для выборки, соответствующей номеру варианта:

1) вычислить выборочное среднее $\bar{x}$, выборочную дисперсию σ_B^2, исправленную выборочную дисперсию s^2 и среднее квадратичное отклонение s;
2) найти с доверительной вероятностью $\gamma = 0{,}99$ доверительный интервал для математического ожидания, а также доверительный интервал для $\sigma(x)$;
3) найти размах варьирования и среднее абсолютное отклонение;
4) вычислить моду и медиану;
5) построить эмпирическую функцию распределения;
6) проверить, согласуются ли выборочные данные с гипотезой о нормальном распределении СВ X графически, с помощью асимметрии и эксцесса и с помощью критерия

согласия Пирсона при уровне значимости $\alpha = 0.01$, разбив отрезок $[x_{\min}, x_{\max}]$ на $N = 5$ интервалов одинаковой длины $[x'_i, x'_{i+1}]$ с границами $x'_i = x_{\min} + \frac{x_{\max} - x_{\min}}{N} \cdot i, \qquad i = \overline{1, N}$.

1.6. Варианты заданий для самостоятельного решения

1	x_i	80	90	100	110	120	130	140
	n_i	4	8	14	40	16	12	6
2	x_i	13	14	15	16	17	18	19
	n_i	7	16	40	25	7	5	3
3	x_i	21	28	35	42	49	56	63
	n_i	7	11	22	50	5	3	2
4	x_i	2	3	4	5	6	7	8
	n_i	4	11	25	30	15	10	5
5	x_i	20	26	32	38	44	50	56
	n_i	2	3	15	50	12	11	7
6	x_i	13	23	33	43	53	63	73
	n_i	3	17	25	40	8	4	3
7	x_i	30	35	40	45	50	55	60
	n_i	4	16	20	40	13	4	3
8	x_i	33	38	46	54	62	70	78
	n_i	7	11	12	60	5	3	2
9	x_i	12	15	22	25	30	35	40
	n_i	3	7	12	40	18	12	8
10	x_i	10	20	30	40	50	60	70
	n_i	4	11	25	30	15	10	5

Практическое занятие № 2
КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

2.1. Линейная регрессия

Пусть изучается система количественных признаков (X, Y). В результате n независимых опытов получены n пар чисел $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Найдем по данным наблюдений выборочное уравнение прямой линии среднеквадратичной регрессии. Для определенности будем искать уравнение регрессии Y на X:

$$\bar{y}_x = kx + b.$$

Поскольку различные значения x признака X и соответствующие им значения y признака Y наблюдались по одному разу, то группировать данные нет необходимости. Также нет надобности использовать понятие условной средней, поэтому искомое уравнение можно записать так:

$$y = kx + b.$$

Угловой коэффициент прямой линии регрессии Y на X называют *выборочным коэффициентом регрессии* Y на X и обозначают через ρ_{yx}; он является оценкой коэффициента регрессии β.

Итак, будем искать выборочное уравнение прямой линии регрессии Y на X вида

$$Y = \rho_{yx} x + b. \qquad (2.1)$$

Подберем параметры ρ_{yx} и b так, чтобы точки $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$, построенные по данным наблюдений, на плоскости XOY лежали как можно ближе к прямой. Уточним смысл этого требования. Назовем отклонением разность

$$Y_i - y_i, \ (i = 1,2\ldots, n),$$

где Y_i – вычисленная по уравнению (2.1) ордината, соответствующая наблюдаемому значению x_i, y_i – наблюдаемая ордината, соответствующая x_i.

Подберем параметры ρ_{yx} и b так, чтобы сумма квадратов отклонений была минимальной (в этом состоит сущность метода наименьших квадратов). Так как каждое отклонение зависит от отыскиваемых параметров, то и сумма квадратов отклонений есть функция F этих параметров (временно вместо ρ_{yx} будем писать ρ):

$$F(\rho, b) = \sum_{i=1}^{n} (\mathrm{Y}_i - y_i)^2 \text{, или } F(\rho, b) = \sum_{i=1}^{n} (\rho x_i + b - y_i)^2 .$$

Для отыскивания минимума приравняем нулю соответствующие частные производные:

$$\frac{\partial F}{\partial b} = 2\sum_{i=1}^{n} (\rho x_i + b - y_i) x_i = 0 , \frac{\partial F}{\partial b} = 2\sum_{i=1}^{n} (\rho x_i + b - y_i) = 0 .$$

Выполнив элементарные преобразования, получим систему двух линейных уравнений относительно ρ и b :

$$\left(\sum x^2\right)\rho + \left(\sum x\right) b = \sum xy ; \left(\sum x\right)\rho + nb = \sum y . \qquad (2.2)$$

Решив эту систему, найдем искомые параметры:

$$\rho_{xy} = \left(n \times \sum xy - \sum x \times \sum y\right) / \left(n \sum x^2 - \left(\sum x\right)^2\right);$$

$$b = \left(\sum x^2 \times \sum y - \sum x \times \sum xy\right) / \left(n \sum x^2 - \left(\sum x\right)^2\right). \qquad (2.3)$$

Аналогично можно найти выборочное уравнение прямой линии регрессии X на Y :

$$\bar{x}_y = \rho_{xy} y + C ,$$

где ρ_{xy} – выборочный коэффициент регрессии X на Y .

Пример 1. Найти выборочное уравнение прямой линии регрессии Y на X .

Опытные данные представлены в таблице:

x	-2	-1	0	1	2	3
y	-0,4	0,2	0,7	1,6	2,0	3,5

Проверить адекватность полученной модели.

Решение.

1) Множество точек, заданных таблицей, построим на плоскости (рис. 2.1).

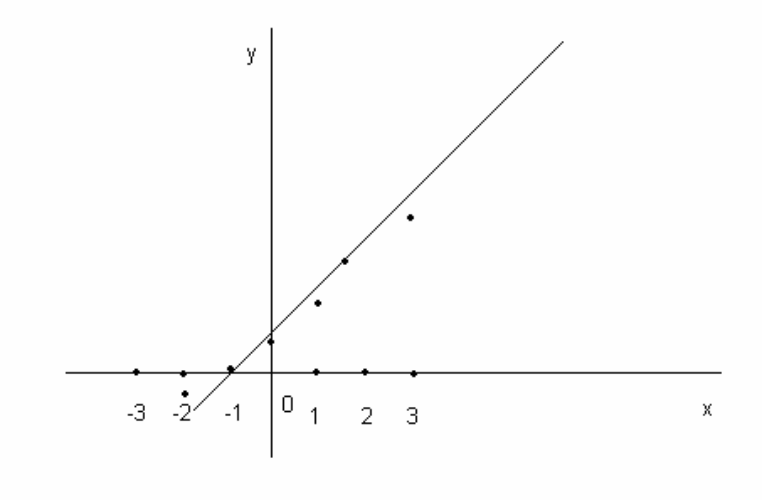

Рис. 2.1. Графическая иллюстрация исходных данных.

Из рисунка видно, что точки группируются около некоторой прямой. Следовательно, зависимость между переменными x и y близка к линейной. Найдем методом наименьших квадратов эмпирическую формулу вида $y = ax + b$.

2) Определим модель. Для вычисления коэффициентов a и b воспользуемся таблицей:

№	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	-2	-0,4	4	0,16	0,8
2	-1	0,2	1	0,04	-0,2
3	0	0,7	0	0,49	0
4	1	1,6	1	2,56	1,6
5	2	2,0	4	4	4,0
6	3	3,5	9	12,25	10,5
СУММА		**7,6**	**19**	**19,5**	**16,7**

Напишем нормальную систему уравнений (2.2): $\begin{cases} 19a + 3b = 16{,}7 \\ 3a + 6b = 7{,}6 \end{cases}$. Из этой системы уравнений найдем a=0,74 и b=0,90. Следовательно, модель имеет вид $y = 0{,}74x + 0{,}9$.

2.2. Выборочный коэффициент корреляции

Выборочный коэффициент корреляции определяется равенством

$$r_B = \frac{\sum n_{xy} xy - n\overline{x}\overline{y}}{n\sigma_x \sigma_y}, \qquad (2.4)$$

где x, y – варианты (наблюдавшиеся значения) признаков X и Y; n_{xy} – частота пары варианта (x, y); n – объем выборки (сумма всех частот); σ_x, σ_y – выборочные средние квадратические отклонения; $\overline{x}, \overline{y}$ – выборочные средние.

Известно, что если величины Y и X независимы, то коэффициент корреляции $r = 0$; если $r = \pm 1$, то Y и X связаны линейной функциональной зависимостью. Следовательно, что коэффициент корреляции r измеряет силу линейной связи между Y и X.

Выборочный коэффициент корреляции r_b является оценкой коэффициента корреляции r генеральной совокупности и поэтому также служит для измерения линейной связи между величинами – количественными признаками Y и X. Допустим, что выборочный коэффициент корреляции, найденный по выборке, оказался отличным от нуля. Так как выборка отобрана случайно, то отсюда еще нельзя заключить, что коэффициент корреляции генеральной совокупности также отличен от нуля. Возникает необходимость проверить гипотезу о значимости (существенности) выборочного коэффициента корреляции (или о равенстве нулю коэффициента корреляции генеральной совокупности). Если гипотеза о равенстве нулю генерального коэффициента корреляции будет отвергнута, то выборочный коэффициент корреляции значим, а величины X и Y коррелированны; если гипотеза принята, то выборочный коэффициент корреляции незначим, а величины X и Y не коррелированны.

Для того, чтобы при уровне значимости α проверить гипотезу о равенстве нулю генерального коэффициента корреляции нормальной двумерной случайной величины, надо вычислить наблюдаемое значение критерия

$$T_{набл} = r_b \cdot \sqrt{n-2} \Big/ \sqrt{1 - r_b^2}, \qquad (2.5)$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α и числу степеней свободы $k = n - 2$ найти критическую точку $t_{крит}(\alpha; k)$ двусторонней критической области. Если $|T_{набл}| < t_{крит}$, тогда нет оснований отвергать гипотезу. Если $|T_{набл}| > t_{крит}$, то гипотезу отвергают.

Пример 2. По данным примера 1 вычислить выборочный коэффициент корреляции и при уровне значимости 0.05 проверить гипотезу о равенстве нулю генерального коэффициента корреляции.

x	-2	-1	0	1	2	3
y	-0,4	0,2	0,7	1,6	2,0	3,5

По данным примера $n=6$, $n_{xy}=1$ для всех x и y. Найдем выборочные средние $\bar{x}$ и $\bar{y}$ и средние квадратические отклонения σ_x, σ_y.

$$\bar{x}=\frac{1}{6}(-2-1+0+1+2+3)=0{,}5,\ \bar{y}=\frac{1}{6}(-0{,}4+0{,}2+0{,}7+1{,}6+2+3{,}5)=1{,}27;$$

$$\sigma_x=\sqrt{\overline{x^2}-(\bar{x})^2},\ \sigma_y=\sqrt{\overline{y^2}-(\bar{y})^2}.$$

Найдем $\overline{x^2}$ и $\overline{y^2}$.

$$\overline{x^2}=\frac{1}{6}\left((-2)^2+(-1)^2+0^2+1^2+2^2+3^2\right)=\frac{4+1+1+4+9}{6}=3{,}17,$$

$$\overline{y^2}=\frac{1}{6}\left((-0{,}4)^2+0{,}2^2+0{,}7^2+1{,}6^2+2^2+3{,}5^2\right)=4{,}03.$$

Тогда $\sigma_x=\sqrt{3{,}17-0{,}5^2}=1{,}71$, $\sigma_y=\sqrt{4{,}03-1{,}27^2}=1{,}55$.

Выборочное значение коэффициента корреляции вычислим по формуле (2.4):

$$r_B=\frac{16{,}7-6\cdot 0{,}5\cdot 1{,}37}{6\cdot 1{,}71\cdot 1{,}55}=0{,}29.$$

Проверим значимость полученного выборочного коэффициента корреляции. Найдем наблюдаемое значение критерия

$$T_{набл}=r_B\cdot\sqrt{n-2}\Big/\sqrt{1-r_B^2}=0{,}29\sqrt{6-2}\Big/\sqrt{1-0.29^2}=0{,}6.$$

По таблице критических точек распределения Стьюдента по уровню значимости 0.05 и числу степеней свободы 4 находим критическую точку двусторонней критической области $t_{крит}=2.78$.

Так как $T_{набл}>t_{крит}$, то отвергаем гипотезу о равенстве нулю генерального коэффициента корреляцию, значит, X и Y коррелированны.

2.3. Контрольное задание

По выборочным данным своего варианта построить линию регрессии Y на X, отобразить графически выборочные данные. Найти выборочный коэффициент корреляции и проверить его значимость.

2.4. Варианты заданий для самостоятельного решения

1	X	2.6	5.4	4.0	0.7	5.8	**2**	X	0.7	2.0	3.7	6.2	7.0
	Y	-1.2	-0.1	-1.0	-3.2	0.4		Y	-2.5	-2.5	-1.3	0.1	0.5
3	X	7.0	2.3	9.2	3.3	9.0	**4**	X	5.1	8.8	8.9	8.7	0.2
	Y	0.2	-2.7	1.7	-0.8	1.4		Y	0.1	1.5	1.4	1.7	-3.0
5	X	1.2	7.9	4.4	2.4	2.1	**6**	X	9.5	5.8	4.0	1.3	3.4
	Y	-3.4	1.8	-0.2	-1.9	-1.1		Y	1.5	0.1	-1.3	-2.1	-1.6
7	X	5.6	4.8	9.6	5.0	5.3	**8**	X	8.6	4.3	5.2	9.2	4.8
	Y	-0.4	-1.6	1.3	-0.2	-0.1		Y	1.2	-0.4	-0.6	1.1	-0.9
9	X	0.0	1.7	4.7	7.5	8.5	**10**	X	6.1	0.8	0.3	1.2	0.4
	Y	-3.2	-2.7	-1.0	0.2	1.8		Y	0.0	-2.0	-3.3	-1.8	-2.9
11	X	4.2	7.4	10	6.6	1.6	**12**	X	0.5	7.7	5.4	3.7	3.6
	Y	-1.0	1.4	1.3	1.5	-2.8		Y	-3.2	1.2	0.2	-0.6	-1.1
13	X	8.6	4.3	5.2	9.2	4.8	**14**	X	4.7	3.3	0.7	2.9	6.7
	Y	1.2	-0.4	-0.6	1.1	-0.9		Y	-0.6	-1.8	-2.5	-1.6	0.6
15	X	3.0	1.2	0.3	1.7	8.0	**16**	X	9.1	0.2	5.9	5.4	9.9
	Y	-1.2	-1.8	-2.9	-1.9	0.1		Y	2.0	-2.8	0.1	0.0	1.6
17	X	3.5	7.1	1.0	7.9	3.1	**18**	X	7.2	6.9	5.6	0.9	1.1
	Y	-1.2	0.4	-2.2	1.2	-1.1		Y	0.2	-0.2	0.6	-2.1	-2.4
19	X	1.6	3.8	0.7	3.4	4.3	**20**	X	7.7	2.8	6.3	7.1	6.2
	Y	-2.2	-1.6	-2.5	-0.9	-0.2		Y	1.4	-2.1	0.1	0.5	-0.3

ЛАБОРАТОРНЫЕ РАБОТЫ

Лабораторная работа №1

ЗНАКОМСТВО С ПАКЕТОМ STATISTICA. ВЫБОРОЧНЫЕ ХАРАКТЕРИСТИКИ. ТОЧЕЧНЫЕ И ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ КРИТЕРИЙ СОГЛАСИЯ ПИРСОНА

Цель: ознакомиться с пакетом *STATISTICA*, процедурой получения выборочные характеристик, ознакомиться с основными видами распределений случайных величин, получить практический навык построения графиков плотности и функции распределения, подбора распределения по экспериментальным данным. Получить практические навыки расчета точечных оценок и интервальных характеристик для математического ожидания и дисперсии, проверке статистической гипотезы о нормальном распределении выборки с использованием пакета *STATISTICA*.

3.1. Введение в пакет STATISTICA

Пакет *STATISTICA* состоит из статистических модулей для анализа и обработки данных, которые в свою очередь состоят из статистических процедур. Выбор необходимого модуля осуществляется с помощью меню **Statistics**, либо с помощью кнопки в левом нижнем углу окна (рис. 3.1).

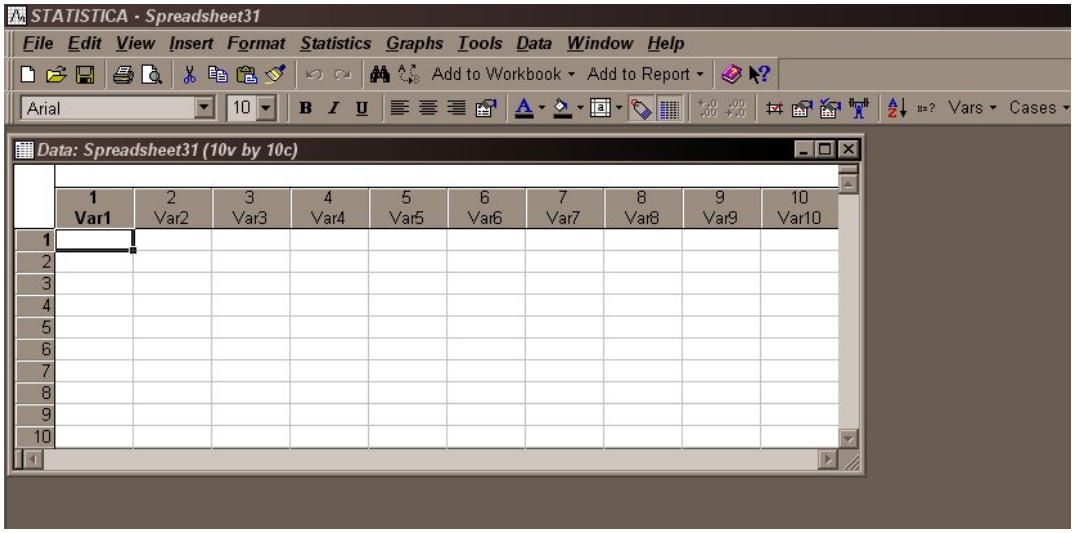

Рис. 3.1. Окно пакета *STATISTICA*

Основным является рабочее окно, в котором вводятся исходные данные и выводятся результаты их статистической обработки в табличном или графическом виде. Ввод данных осуществляется в табличном виде.

Набор данных в пакете STATISTICA – это прямоугольная таблица, столбцам которой соответствуют обрабатываемые *переменные* (**Variables**), а строкам отвечают *наблюдения* (**Cases**) значений переменных. Для создания нового набора данных нужно, прежде всего, завести файл с *трафаретом* таблицы нужных размеров.

Сделать это можно так (см. рис. 3.1). По меню *File –New Data*... через раскрывшееся диалоговое окно нужно завести новый файл с расширением **.sta**. В строке для заголовка можно дать комментарий к содержимому набора данных (для входа в строку заголовка достаточно дважды кликнуть на ней левой кнопкой мыши). В результате открытия нового файла в окне пакета появляется (как на рис. 3.1) начальный трафарет создаваемого набора данных с исходными размерами в 10 переменных на 10 наблюдений. Реально нужное количество переменных и наблюдений выставляется после этого у трафарета по меню инструментальных кнопок *Vars* и *Cases*. Как наблюдениям, так и переменным в трафарете создаваемого набора данных можно дать содержательные названия по меню *Cases – Case Name Manager.*

Названные действия по определению переменных могут быть проделаны из основного окна с трафаретом набора данных по меню *Vars –All Specs*. В результате появляется окно со списком установленных по умолчанию атрибутов переменных, которые можно поправить и дополнить с клавиатуры. При этом особенно тщательно нужно определить формат каждой переменной. По умолчанию он есть числовой с размерами “8.3” (т.е. с фиксированной точкой, где под все значащие цифры, знак числа и десятичную точку отведено 8 символов, 3 из которых предназначены для дробной части). Сменить и детализировать формат отдельной переменной можно в диалоговом окне, которое раскрывается, если дважды кликнуть левой кнопкой мыши на нужной переменной в трафарете. Это же окно раскрывается и по меню *Vars – Specs*.

Что касается имён переменных, то их лучше всегда давать содержательными (а не абстрактными VAR1, VAR2 и т.д.). Кроме имени (**Name**) для каждой переменной надо указать так называемый *код пропущенного значения* (**MD Code**). По умолчанию этот код есть “–9999”, и он отмечает в памяти для процедур обработки пакета, что, на самом деле, на его месте (в определенной клетке трафарета) реального значения нет. А изображается пропущенное значение на экране в наборе данных пробелом. Из обязательных атрибутов переменной надо указать тип и формат её значений. Тип определяет, будет ли переменная числовой, текстовой, датой, временем и проч., а формат

(**Format**) описывает размеры значений переменной. Значениям переменной можно также дать развернутый содержательный комментарий (**Long Name**).

Файлы проекта. При работе с пакетом *STATISTICA*:образуется пять различных типов документов: рабочая книга (*Workbook*), рабочий лист – мультимедийная таблица (*Spreadsheet*), отчет (*Report*), графическую область (*Graph)* и макрокоманда (*Macros*) для языка *STATISTICA* Visual Basic. Рабочая книга представляет собой упорядоченный вывод данных, объединяя в себе рабочие листы и графики. Каждый документ представляет собой таблицу. Файл рабочей книги имеет расширение **.stw*. Рабочие листы пакета *STATISTICA* предназначены для ввода данных в числовом или текстовом форме имеет расширение **.sta*. Форматом рабочего листа является двумерная таблица с неограниченным количеством наблюдений (строк) и переменных (столбцов), каждый их которых содержит неограниченное количество символов. В рабочий лист могут быть внедрены звук, видеодокументы, графика, анимация. Отчеты *STATISTICA* позволяют организовать вывод данных в текстовом формате, более удобном для вывода документов на печать. По умолчанию файл отчета имеет расширение **.str*, но существует возможность преобразования отчета в стандартный файл формата RTF. Вся графическая обработка данных сохраняется в отдельных файлы с расширением **.stg*. При этом поддерживается внедрение графических объектов из других программ. Макрокоманда представляют собой программный код, написанный на языке Visual Basic. Каждый из описанных компонентов проекта отображается в отдельном окне и имеет свою пиктограмму в дереве проекта на панели экрана слева.

3.2. Решение задач описательной статистики. Визуализация результатов

3.2.1. Создание таблицы исходных данных

Для ввода с клавиатуры нужно сделать активным окно с трафаретом, клавишами управления курсором выделить нужную переменную и наблюдение, набрать требуемое значение в выделенной клетке трафарета, нажать клавишу Enter для сохранения значения, перейти к следующей клетке трафарета и т.д. При этом для ввода повторяющихся значений можно воспользоваться операциями копирования/вставки через буфер с помощью подходящих инструментальных кнопок основного окна. Если при вводе возникла необходимость быстро подправить форматные размеры под значения

переменных, то это также можно сделать по соответствующим инструментальным кнопкам.

Для ввода в создаваемый набор данных значений переменным за счёт их вычисления можно пойти двумя путями. При несложных вычислениях можно воспользоваться простым *редактором формул*, который запускается по кнопке *Functions* из диалогового окна переменной (последнее окно раскрывается, если дважды кликнуть левой кнопкой мыши на нужной переменной в трафарете).

Новый набор данных необходимо сохранить в файл с данными в своей личной папке по меню *File – Save As* основного окна.

Создание таблицы требуемых размеров. Из пункта меню *File* выберем команду *New;* укажем имя файла для сохранения будущей информации и в окне задания размеров таблицы укажем 20×10 (20 выборок по 10 наблюдений). Другой способ изменения размеров таблицы заключается в использовании кнопок *Vars* (переменные) и *Cases* (наблюдения), или через меню *Edit,* команда *Variables*. В этом случае во всплывшем меню выберем команду *Add* (добавить). На экране запрос о числе добавляемых переменных (столбцов) и о том, куда их поместить. Добавим 10 переменных, проставив *10* в поле *Number... to add* (набором на клавиатуре или кнопками справа от поля); в поле *Insert after* укажем имя *Var10*, после которой будут вставлены в матрицу новые столбцы; затем *OK.* Теперь можно убедиться (просмотром таблицы), что в ней 20 столбцов; кроме того, размеры таблицы (в данном случае, *20v*10c*) всегда указаны и ее заголовке. Количество строк не изменяем: оно равно 10. Заметим, что если число строк (C*ase*) или столбцов (*variable*) в таблице превышает необходимое, можно таблицу не уменьшать.

Генерация выборок. Последовательность действий следующая: нажмем клавишу *Vars* и в появившемся окне выберем пункт *All specs*... (спецификация всех). В результате этих действий окно-таблица, в первом столбце которой находятся названия переменных (*var1, var2,..., var20*), во втором – тип этих переменных (по умолчанию – *Double*), в третьем – код переменной, а в четвертом длина, в последнем (*Long Name*) – функция расчета переменных. Выделим первую клетку пятого столбца и введем формулу = *rnd(10)*. Это означает, что будут сгенерированы случайные числа, равномерно распределенные на отрезке [0, 10]. Скопируем эту запись в буфер обмена, выбрав в меню *Edit* или в контекстном меню (правая кнопка мыши) пункт *Copy*, а затем пере-

несем ее в остальные клетки (со 2 по 20): выделим очередную клетку, выберем в меню *Edit* или в контекстном меню пункт *Past*. Закроем окно. Другой способ быстрого копирования заключается в следующем: необходимо установить курсор на первую клетку пятого столбца (он будет заключен в рамку), и затем поместить его в нижний правый угол клетки. При этом изменится внешний вид курсора. Далее перемещаем мышь при нажатой левой кнопке на необходимое количество клеток (в данном случае – на весь столбец) и отпускаем кнопку мыши. В результате формулой будут заполнены все строки последнего столбца.

Рассчитаем значения переменных по заданной нами формуле. Нажмем на панели инструментов кнопку  . В появившемся окне выберем пункт *All variables*, нажмем кнопку *OK* (рис. 3.2).

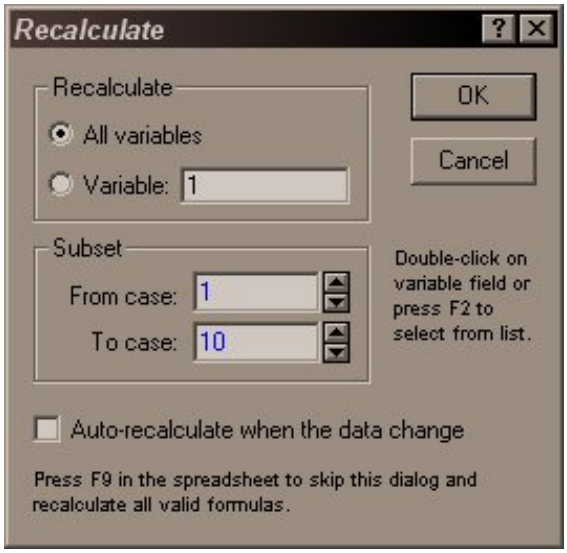

Рис. 3.2. Окно расчета переменных

Отформатируем полученные данные. Изменим имена трех последних столбцов: выделим столбец, нажмем правую кнопку мыши, выберем *Variable Spec*, в открывшемся окне зададим новое имя столбца в поле *Name*, формат данных в поле *Display Format* – числовой (*Number*) и количество цифр после десятичной запятой – *Decimal places* –3). Скорректируем размеры таблицы: в меню *Format* выберем *Variables*, далее *AutoFit* (*Автоподбор*).

3.2.2. Вычисление выборочных характеристик

Рассмотрим пример. Зададим исходные данные (рис. 3.3).

	1 X
1	60
2	100
3	80
4	45
5	45
6	40
7	45
8	60
9	80
10	40
11	20
12	10
13	60
14	20
15	10
16	6
17	10
18	60
19	20
20	80

Рис. 3.3. Исходные данные

Запустим модуль *Basic Statistics and Tables*. В окне *Descriptive Statistic* выберем вкладку *Advanced* (рис. 3.4), в результате появится окно, содержащее список числовых характеристик, которые могут быть вычислены. Отметим *Valid N* (объем выборки), *Mean* (среднее), *Median* (медиана), *Standart Devaition* (среднее квадратическое отклонение), *Variance* (дисперсия), *Mimimum&Maximum* (минимум и максимум), *Range* (размах варьирования), *Qartile range* (размах квартилей) и активизируем кнопку *Summary*. В появившемся окне выделим переменную, для которой нужно произвести расчеты (переменная *X)*.

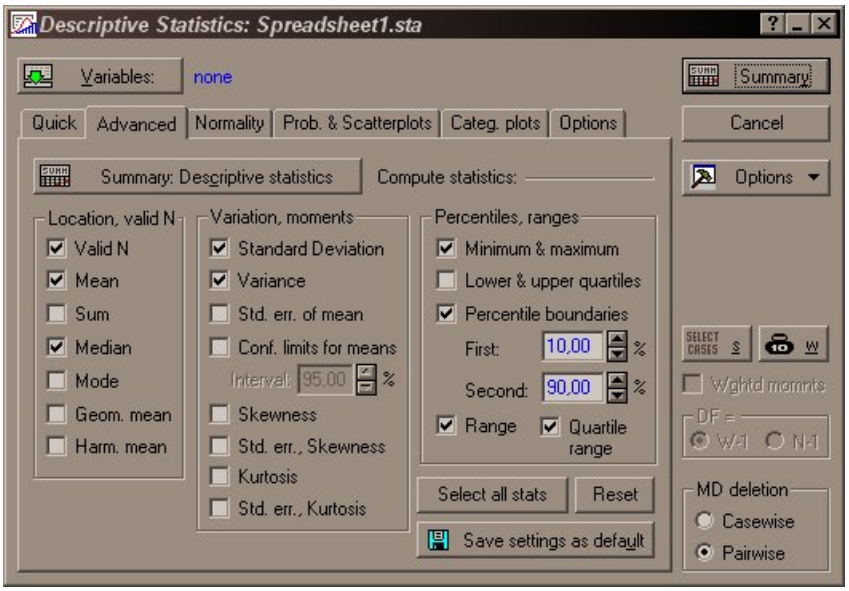

Рис. 3.4. Вкладка *Advanced* окна *Descriptive Statistic*

Результаты вычислений размещаются в активном окне внизу (при анализе данных столбца) или слева (при анализе данных строке) от исходных данных (рис. 3.5).

Descriptive Statistics (Spreadsheet1.sta)

Variable	Valid N	Mean	Median	Minimum	Maximum	Percentile 10,00000	Percentile 90,00000	Range	Quartile Range	Variance	Std.Dev.
X	20	44,55000	45,00000	6,000000	100,0000	10,00000	80,00000	94,00000	40,00000	774,5763	27,83121

Рис. 3.5. Вычисленные параметры описательной статистики

3.2.3. Построение таблицы и графиков частот, диаграммы размаха

В окне *Descriptive Statistic* во вкладке *Quick* нажмем на кнопку *Frequency Tables*, в появившемся окне выделим необходимый нам столбец переменных Var1, нажмем OK. В результате получим таблицу частот (см. рис. 3.6). В первом столбце заданы интервалы для переменной x, причем последняя строка содержит пропущенные значения. Второй столбец содержит число попаданий переменной в интервалы (*Count*), третий столбец – кумулятивное число попаданий (*Cumulative Count*), четвертый и шестой – частоты в процентных соотношениях для имеющихся в наличии (не пропущенных) наблюдений (*Percent of Valid*) и для всех наблюдений (*% of Cases*), пятый и седьмой столбцы – кумулятивные частоты в процентах соответственно для (не пропущенных) наблюдений (*Cumul. % of Valid*) и для всех наблюдений (*Cumul. % of All*).

Frequency table: X (Spreadsheet1.sta)
K-S d=,16114, p> .20; Lilliefors p<,20

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
-20,0000<x<=0,000000	0	0	0,00000	0,0000	0,00000	0,0000
0,000000<x<=20,00000	7	7	35,00000	35,0000	35,00000	35,0000
20,00000<x<=40,00000	2	9	10,00000	45,0000	10,00000	45,0000
40,00000<x<=60,00000	7	16	35,00000	80,0000	35,00000	80,0000
60,00000<x<=80,00000	3	19	15,00000	95,0000	15,00000	95,0000
80,00000<x<=100,0000	1	20	5,00000	100,0000	5,00000	100,0000
Missing	0	20	0,00000		0,00000	100,0000

Рис. 3.6. Таблица частот

Для построения графиков частот и кумулятивных частот выделим столбцы *Percent of valid*, нажмем правую кнопку мыши и в контекстном меню выберем команду *Graph of Block Data – Line Plot: Entire Columns* (так как данные для построения графика расположены в столбцах).

Для построения гистограммы частот вернемся в окно *Descriptive Statistic* (оно располагается внизу основного окна в свернутом состоянии). С помощью функциональной кнопки *Histograms* получим гистограмму. Кроме гистограммы частот на рисунке также будет показана теоретическая кривая плотности распределения наблюдаемой случайной величины в случае нормального распределения.

Для построения диаграммы размаха варьирования, в окне *Descriptive Statistic* во вкладке *Quick* нажмем на функциональную кнопку *Box&Wisker Plot*. В результате получим диаграмму размаха (график “ящик с усами”), которая показывает, как отклоняются данные от среднего значения. Если необходимо зажать дополнительные параметры диаграммы размаха варьирования, то необходимо выбрать вкладку *Options* в окне *Descriptive Statistic*.

3.2.4. Виды распределений случайных величин. Процедура Probability Calculator. Расчет квантилей. Построение графиков плотности и функции распределения

Задача формулируется следующим образом: определить вероятность, что случайная величина не превысит заданного значения; найти значение случайной величины, для которого функция распределения равна заданному значению p. Искомое значение случайной величины называется *квантилью*, соответствующей вероятности p.

Для работы с распределениями непрерывных случайных величин в пакете *STATISTICA* используется калькулятор вероятностных распределений. Для его вызова выполняется процедура *Probability Calculator*, которая входит в модуль *Basic Statistics and Tables*. С помощью вероятностного калькулятора могут решаться разнообразные вероятностные задачи, например, построение графиков плотностей и функций распределения, определение квантили для заданной вероятности и пр.

При запуске процедуры *Probability Calculator* открывается окно *Probability Distribution Calculator*. В левой части данного окна содержится список распределений непрерывных случайных величин, доступных пользователю. В средней части в строке p содержится вероятность $P(X \le x) = F(x)$, в стоке X – значение случайной величины или квантиль, соответствующая этой вероятности. Далее содержатся поля для ввода параметров распределения. В верней части окна содержатся опции, предназначенные для настройки режимов работы процедуры. Назначение этих опций следующее: *Inverse* – работа с обратной функцией распределения, которая используется при

вычислении квантили; *Two-tailed* – построение двустороннего интервала для плотности распределения; *1-Cumulative p* – использование вместо вероятности p значения вероятности $1 - p$; *Create Graph* – создание графика; *Print* – вывод результатов на печать. Пользователь может выбрать закон распределения и для заданного значения случайной величины вычислить значение функции распределения или, наоборот, задать вероятность p и определить соответствующую ей квантиль $F^{-1}(x)$. В последнем случае следует переключить опцию *Inverse*.

В случае работы с дискретными распределениями для расчета вероятностей и функций распределения применяют встроенные функции. Например, для биномиального распределения имеется две встроенные функции: *Binom(x, p, n) IBinom(x, p, n)*, вычисляющие соответственно вероятность принятия значения $x = i$ и значение распределения в точке x для биномиального распределения с параметрами *p* и *n*, определяемого по формуле $P(X = i) = C_n^i p^i (1 - p)^{n-i}$, $i = \overline{0,n}$. В пакете *STATISTICA* имеются встроенные функции для всех основных законов распределения, причем функции пакета, имена которых начинаются с буквы *I*, вычисляют значения функций распределения.

Для вычисления встроенной функции в пакете *STATISTICA* следует выделить незаполненный столбец таблицы **Spreadsheet** исходных данных и затем выполнить команду *Vars–Current Spec*, нажав правую кнопку мыши. В нижней части отрывшегося окна спецификации переменной находится рабочая область *Long name* (*label, link or formula with function*), которая предназначена для ввода выражений и комментариев. Набор формулы следует производить, начиная со знака равно (=), далее встроенные функции могут быть набраны непосредственно с клавиатуры либо с помощью конструктора (кнопка *Function*).

В последнем случае открывается окно *Spreadsheet Formulas*, в котором содержится список встроенных функций. Вставки необходимой функций производится с помощью двойного щелчка мыши по имени функции. Далее задаются фактические параметры распределения и значения аргумента, в качестве последних могут использоваться имена переменных (столбцов) или константы.

3.2.5. Нормальное распределение

Запустим процедуру *Probability Calculator*, для чего в меню модуля *Basic Statistics and Tables* выделим строку *Probability Calculator* и нажмем кнопку *OK*. В результате откроется окно калькулятора *Probability Distribution Calculator*.

Выполним расчеты для нормального распределения со средним значением $\mu = 0.5$ и $\sigma = 1$. Первая задача будет заключаться в поиске квантили для вероятности $p = 0.8$ и построении графиков плотности и функции распределения.

В окне калькулятора в соответствующие поля введем параметры распределения, вероятность p и отметим опции *Inverse* и *Create Graph*.

После нажатия кнопки *Compute* получим окна с результатами, представленные на рис. 3.7. Значение квантили для заданной вероятности равно 1.341612.

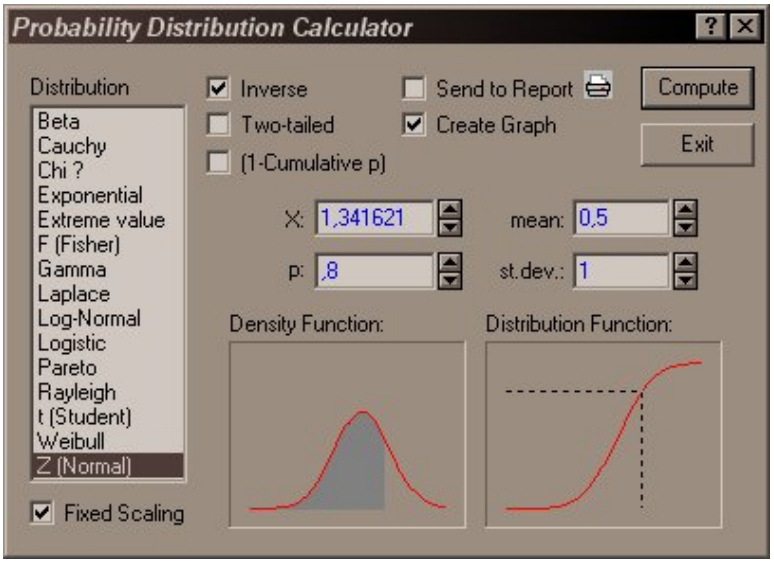

Рис. 3.7. Результаты расчета квантили для нормального распределения

Следующая задача заключается в определении значения функции распределения для заданного значения случайной величины $x = 1$. Введя в поле *X* значение, равное 1, и нажав кнопку *Compute*, в поле p получим значение функции распределения, равное 0.691462.

Расчет квантилей для распределений χ^2, Фишера и Стьюдента определяется аналогично путем выбора соответствующего распределения в разделе *Distribution*.

3.2.6. Биномиальное распределение. Построение полигона вероятностей

Выполним расчеты для биномиального распределения с параметрами $n = 10$ и $p = 0.7$ в точке $x = 9$. Введем в таблицу с исходными данными (рис. 3.8) заданные значения (первый и второй столбец).

	1 N	2 P	3 X	4 P_X	5 F_X
1	10	0,7	9	0,121061	0,971752
2					

Рис. 3.8. Таблица с исходными данными и результатами расчета

Далее в окне спецификации четверного столбца, названного нами *P_X*, в поле *Long name* введем формулу для биномиального распределения.

Аналогично в окне спецификации пятого столбца, названного нами *F_X*, в поле *Long name* введем формулу для функции распределения биномиального распределения вида: = *IBinom(x, p, n)*. В результате получим следующие ответы: $P(X = 9) = 0.121$ и $F(9) = 0.972$ (рис. 3.8).

Для биномиального распределения с параметрами рассчитаем распределение вероятностей и функцию распределения. Для этого выполним описанный выше пример для множества точек $x = 0, 1, 2, \ldots 10$ путем формирования 11 строк таблицы. Тогда таблица с исходными данными и полученными результатами будет иметь вид, представленный на рис. 3.9.

	1 N	2 P	3 X	4 P_X	5 F_X
1	10	0,7	0	0,00001	0,00001
2	10	0,7	1	0,00014	0,00014
3	10	0,7	2	0,00145	0,00159
4	10	0,7	3	0,00900	0,01059
5	10	0,7	4	0,03676	0,04735
6	10	0,7	5	0,10292	0,15027
7	10	0,7	6	0,20012	0,35039
8	10	0,7	7	0,26683	0,61722
9	10	0,7	8	0,23347	0,85069
10	10	0,7	9	0,12106	0,97175
11	10	0,7	10	0,25348	1,00000

Рис. 3.9. Таблица с исходными данными и результаты расчета полигона вероятностей и функции распределения

Используя полученную таблицу с результатами, построим полигон вероятностей и функцию распределения для заданного биномиального распределения. Для построения соответствующих графиков необходимо: выделить столбцы *P_X* и *F_X*; выполнить команду меню *Graphs – Sequential/Stacked Plots*; в открывшемся окне выбирать тип графиков – *Mixed Step* (ступенчатый график), нажав на кнопку *Variables*, задать в левом окне *F_X* , в правом – *P_X*, и задать имя по переменной абсцисс – *X*.

После нажатия кнопки ОК получим полигон вероятностей *P_X* и график функции распределения *F_X*, представленные на рис. 3.10.

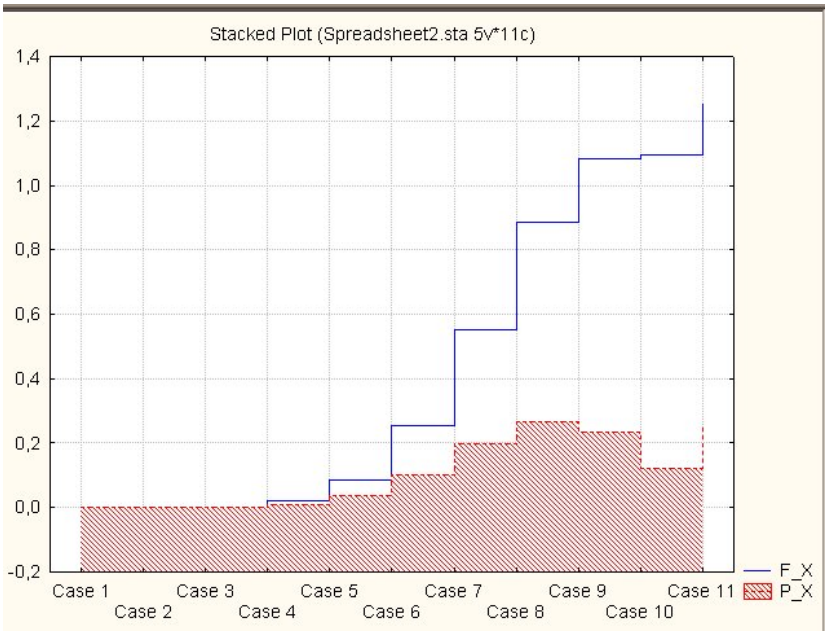

Рис. 3.10. Полигон вероятностей и график функции распределения

3.3. Вычисление доверительных интервалов для параметров нормально распределенной случайной величины

3.3.1. Доверительный интервал для математического ожидания

Для вычислений будем использовать сгенерированную выше выборку (столбец) из 20 наблюдений над нормальной случайной величиной со средним a = 10 и дисперсией σ^2 = 4 и определим доверительные интервалы для a с уровнем доверия $P_Д$ =0.8.

В меню *Statistics* запустим модуль *Basic Statistics and Tables*, в котором выберем процедуру *Descriptive Statistics*. Во вкладке *Advanced* установим *Conf. Limits for means* и укажем значение *Interval: 80%*, нажав на кнопку *Variables*, зададим диапазон переменных *Var1- Var10*. Результаты вычислений будут представлены в отдельном окне.

Первый столбец содержит список рассмотренных переменных, два других – левую и правую границы доверительных интервалов, последний столбец – стандартную ошибку. Построим диаграмму рассеяния, наглядно иллюстрирующую выполненные расчеты. Во вкладке *Quick* нажмем кнопку *Box&Wisker plot for all variables*.

3.3.2. Доверительный интервал для дисперсии

Для построения доверительно интервал для дисперсии вычислим с помощью процедуры *Descriptive Statistics* модуля *Basic Statistics and Tables* значения количества измерений (*Valid N*) и стандартного отклонения по каждой переменной. В получившейся таблице добавим три столбца, назовем их *D*, *D1*, *D2*. В первом вычислим значение эмпирической дисперсии как квадрат стандартного отклонения, в двух других будут находиться значения левой и правой границы доверительного интервала. Значения левой и правой границ доверительного интервала будем вычислять по формуле $\frac{N-1}{\chi_2^2} D < D_X < \frac{N-1}{\chi_1^2}$. Величины χ_1^2 и χ_2^2 – это значения распределения χ^2 в точках $\frac{1+\beta}{2}$ и $\frac{1-\beta}{2}$ с количеством степеней свободы *N-1*, β – уровень доверия. В данной задаче $N = 50$, выберем $\beta = 0{,}8$. Зададим расчетные формулы для переменных *D1* и *D2*. Пересчитав переменные, получим требуемый результат.

3.4. Проверка гипотезы о нормальном распределении генеральной совокупности. Критерий Пирсона

Для проверки гипотезы о нормальном распределении случайных величин в пакете *STATISTICA* используется модуль *Descriptive Statistics* (*Описательные статистики*). При этом используется критерий согласия Колмогорова-Смирнова, предполагающий, что параметры нормального распределения известны. Проверим гипотезу о нормальном законе распределения размеров головок заклепок, сделанных на одном станке, по выборке объема *n* = 200; измерения приведены в табл. 3.1. Оценками для *a* (среднего) и σ (стандартного отклонения) являются

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{и} \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\,.$$

Таблица 3.1

Диаметры 200 головок заклепок, мм									
13.39	13.33	13.56	13.38	13.43	13.37	13.53	13.40	13.25	13.39
13.28	13.34	13.50	13.38	13.38	13.45	13.47	13.62	13.45	13.39
13.53	13.58	13.32	13.27	13.42	13.40	13.57	13.46	13.33	13.40
13.57	13.36	13.43	13.38	13.26	13.52	13.35	13.29	13.48	13.43
13.40	13.39	13.50	13.52	13.39	13.39	13.46	13.29	13.55	13.31
13.29	13.33	13.38	13.61	13.55	13.40	13.20	13.31	13.46	13.13
13.43	13.51	13.50	13.38	13.44	13.62	13.42	13.54	13.31	13.58
13.41	13.49	13.42	13.45	13.34	13.47	13.48	13.59	13.20	14.56
13.55	13.44	13.50	13.40	13.48	13.29	13.31	13.42	13.32	13.48
13.43	13.26	13.58	13.38	13.48	13.45	13.29	13.32	13.24	13.38
13.34	13.14	13.31	13.51	13.59	13.32	13.52	13.57	13.62	13.29
13.23	13.37	13.64	13.30	13.40	13.58	13.24	13.32	13.52	13.50
13.43	13.58	13.63	13.48	13.34	13.37	13.18	13.50	13.45	13.60
13.38	13.33	13.57	13.28	13.32	13.40	13.40	13.33	13.20	13.44
13.34	13.54	13.40	13.47	13.28	13.41	13.39	13.48	13.42	13.46
13.28	13.46	13.37	13.53	13.43	13.30	13.45	13.40	13.45	13.40
13.33	13.39	13.56	13.46	13.26	13.35	13.42	13.36	13.44	13.41
13.43	13.51	13.51	13.24	13.34	13.28	13.37	13.54	13.43	13.35
13.52	13.23	13.48	13.48	13.54	13.41	13.51	13.44	13.36	13.36
13.53	13.44	13.69	13.66	13.32	13.26	13.51	13.38	13.46	13.34

Результаты измерения диаметров заклепок занесем в таблицу с одним столбцом (*d*) и 200 строками.

Для проверки гипотезы о нормальном распределении исходных данных будем использовать процедуру *Distribution Fitting* (подбор распределения), которая находится в меню *Statistics*.

Зададим диапазон исходных данных, нажав на кнопку *Variable* и выбрав там единственно возможную*: d* (рис. 3.11). Далее нажмем кнопку *OK*.

Выберем во вкладке *Continuous Distributions:*требуемый тип распределения *Normal* (нормальное). Это можно также осуществить с помощью вкладки *Quick*.

Во вкладке *Options* откажемся от теста Колмогорова-Смирнова. Для этого установим соответствующий переключатель в положение None. При этом выключатель теста χ^2 (*Chi-Square*) должен быть активизирован. Для построения графика отметим опцию *Frequency* d*istribution* (частоты распределения).

Во вкладке *Parameters* установим количество интегралов разбиения равное 19. В этом же окне наблюдаем значения нижней и верхней границы значений исходных данных, наблюдаемые значение математического ожидания и дисперсии. Нажав кнопку *Summary* во вкладке *Quick*, получаем таблицу частот. Первый столбец данных

содержит левые границы интервалов группирования данных, второй – наблюдаемые частоты попадания данных в интервал, третий – накопленные частоты.

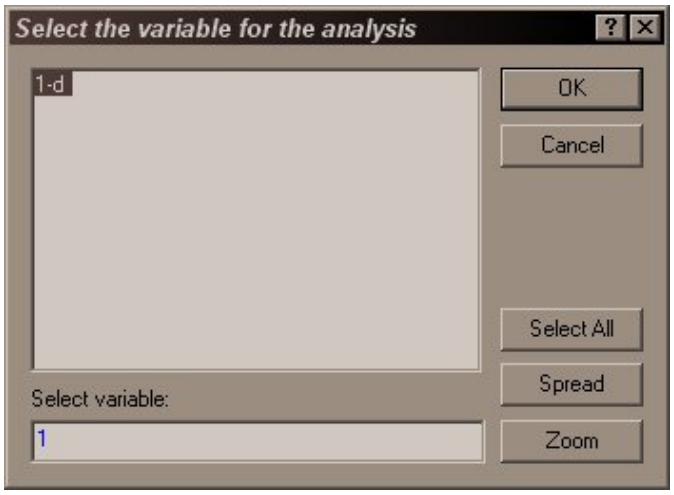

Рис. 3.11. Выделение переменной для анализа

В таблице на рис. 3.12 приведено значение статистики $\chi^2 = 11.99951$, количество степеней свободы $df = 3$, которое получилось при объединении интервалов и приведено значение вероятности $P\left(\chi_3^2 \geq 12\right) = p = 0{,}00738$. Последнее означает, что если гипотеза верна, вероятность получить 12.00 или больше равна 0.00738 – слишком мала, чтобы поверить в нормальность распределения исходных данных. Следовательно, гипотезу о нормальности отклоняем.

Variable: d, Distribution: Normal (DZ.sta)
Chi-Square = 11,99951, df = 3 (adjusted) , p = 0,00738

Upper Boundary	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= 13,00000	0	0	0,00000	0,0000	0,17182	0,1718	0,08591	0,0859	-0,17182
13,10000	0	0	0,00000	0,0000	1,50721	1,6790	0,75360	0,8395	-1,50721
13,20000	6	6	3,00000	3,0000	8,26685	9,9459	4,13342	4,9729	-2,26685
13,30000	24	30	12,00000	15,0000	26,66965	36,6155	13,33483	18,3078	-2,66965
13,40000	67	97	33,50000	48,5000	50,67937	87,2949	25,33969	43,6475	16,32063
13,50000	58	155	29,00000	77,5000	56,77439	144,0693	28,38719	72,0346	1,22561
13,60000	36	191	18,00000	95,5000	37,50152	181,5708	18,75076	90,7854	-1,50152
13,70000	8	199	4,00000	99,5000	14,59742	196,1682	7,29871	98,0841	-6,59742
13,80000	0	199	0,00000	99,5000	3,34433	199,5126	1,67217	99,7563	-3,34433
13,90000	0	199	0,00000	99,5000	0,45020	199,9628	0,22510	99,9814	-0,45020
14,00000	0	199	0,00000	99,5000	0,03554	199,9983	0,01777	99,9992	-0,03554
14,10000	0	199	0,00000	99,5000	0,00164	200,0000	0,00082	100,0000	-0,00164
14,20000	0	199	0,00000	99,5000	0,00004	200,0000	0,00002	100,0000	-0,00004
14,30000	0	199	0,00000	99,5000	0,00000	200,0000	0,00000	100,0000	-0,00000
14,40000	0	199	0,00000	99,5000	0,00000	200,0000	0,00000	100,0000	-0,00000
14,50000	0	199	0,00000	99,5000	0,00000	200,0000	0,00000	100,0000	0,00000
14,60000	1	200	0,50000	100,0000	0,00000	200,0000	0,00000	100,0000	1,00000
14,70000	0	200	0,00000	100,0000	0,00000	200,0000	0,00000	100,0000	0,00000
< Infinity	0	200	0,00000	100,0000	0,00000	200,0000	0,00000	100,0000	0,00000

Рис.3.12. Результаты расчетов

Построим гистограмму наблюдений и сравним распределение наблюдаемых и ожидаемых частоты с помощью графиков. Для этого во вкладке *Quick* нажмем кнопку *Plot of observed and expected distribution*. Очевидно, что гистограмма, построенная по исходным данным, значительно отличается от кривой плотности нормального распределения, которая на полученном графике отмечена красной линией (рис.3.13).

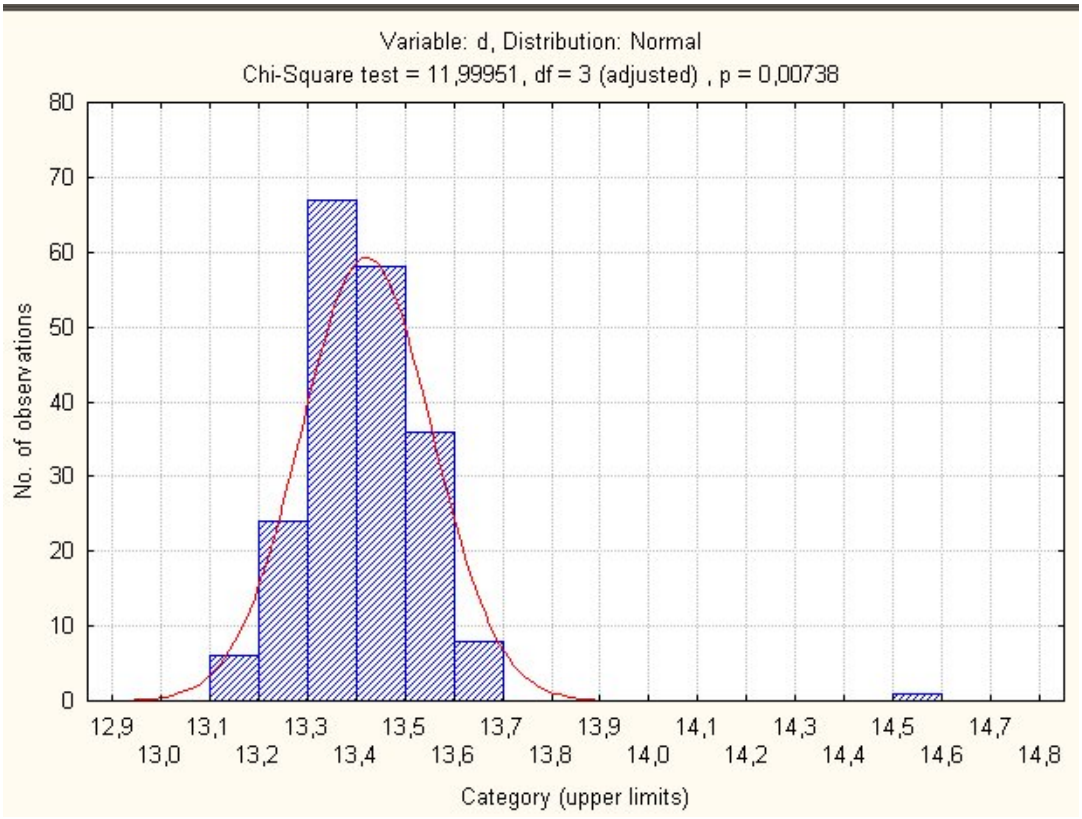

Рис. 3.13. Гистограмма частот и график плотности нормального распределения

Если посмотреть гистограмму наблюдений, видно, что в выборке имеется одно аномальное значение 14.56 (№ 188), которое могло появиться в результате какой-либо ошибки (при записи наблюдений, при перепечатке или попалась деталь с другого станка и т.д.). Удалим его и снова проверим гипотезу. Удаление одного наблюдения, если оно типично, не может изменить характеристики совокупности из 200 элементов; если же изменение происходит, следовательно, это наблюдение типичным не является и должно быть удалено.

Чтобы не искажать исходные данные, продублируем их в новый столбец, который назовем, например, *dc*, и удалим аномальное наблюдение.

Повторим проверку гипотезы для полученной выборки и убедимся в том, что наблюдения не противоречат гипотезе о нормальности, так как в данном случае значение вероятности $P\left(\chi_3^2 \geq 12\right) = p = 0{,}27927$ (рис. 3.14). Этот же вывод подтверждается с помощью графиков. В данном случае гистограмма, построенная по исходным данным, лучше подходит к кривой плотности нормального распределения.

Variable: dc, Distribution: Normal (DZ.sta)
Chi-Square = 2,55113, df = 2 (adjusted) , p = 0,27927

Upper Boundary	**Observed Frequency**	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= 13,10000	0	0	0,00000	0,0000	0,33356	0,3336	0,16762	0,1676	-0,33356
13,20000	6	6	3,01508	3,0151	4,14579	4,4793	2,08331	2,2509	1,85421
13,30000	24	30	12,06030	15,0754	23,59336	28,0727	11,85596	14,1069	0,40664
13,40000	67	97	33,66834	48,7437	59,83206	87,9048	30,06636	44,1732	7,16794
13,50000	58	155	29,14573	77,8894	67,91270	155,8175	34,12698	78,3002	-9,91270
13,60000	36	191	18,09045	95,9799	34,52392	190,3414	17,34870	95,6489	1,47608
13,70000	8	199	4,02010	100,0000	7,83437	198,1758	3,93687	99,5858	0,16563
< Infinity	0	199	0,00000	100,0000	0,82424	199,0000	0,41419	100,0000	-0,82424

Рис. 3.14. Расчеты по скорректированным данным

3.5. Контрольное задание

По выборочным данным своего варианта практического задания № 1:

1) рассчитать выборочные характеристики: среднее, медиану, среднее квадратическое отклонение, минимальное и максимальное значения выборки;

2) построить график функции распределения;

3) вычислить доверительные интервалы для среднего и для дисперсии;

4) пользуясь критерием Пирсона, проверить гипотезу о нормальном распределении выборки.

Лабораторная работа № 2
ЛИНЕЙНАЯ РЕГРЕССИЯ

Цель: получить практические навыки построения линейной регрессии по исходным данным с использованием пакета *STATISTICA* и оценки регрессионных параметров.

4.1. Построение линейной регрессионной модели по выборочным данным

Рассмотрим построение линейной регрессионной модели по выборочным данным следующего примера.

Пример. В табл. 4.1 приведены данные по 45 предприятиям по статистической связи между стоимостью основных фондов (*fonds,* млн. денежных единиц) и средней выработкой на 1 работника (*product,* тыс. денежных единиц); z – вспомогательный признак: z = 1 – федеральное подчинение, z = 2 – муниципальное.

Таблица 4.1

fonds	*product*	*z*	*fonds*	*product*	*z*	*fonds*	*product*	*z*
6,5	18,3	1	9,3	17,2	2	10,4	21,4	2
10,3	31,1	1	5,7	19,0	2	10,2	23,5	2
7,7	27,0	1	12,9	24,8	2	18,0	31,1	2
15,8	37,9	1	5,1	21,5	2	13,8	43,2	2
7,4	20,3	1	3,8	14,5	2	6,0	19,5	2
14,3	32,4	1	17,1	33,7	2	11,9	42,1	2
15,4	31,2	1	8,2	19,3	2	9,4	18,1	2
21,1	39,7	1	8,1	23,9	2	13,7	31,6	2
22,1	46,6	1	11,7	28,0	2	12,0	21,3	2
12,0	33,1	1	13,0	30,9	2	11,6	26,5	2
9,5	26,9	1	15,3	27,2	2	9,1	31,6	2
8,1	24,0	1	13,5	29,9	2	6,6	12,6	2
8,4	24,2	1	10,5	34,9	2	7,6	28,4	2
15,3	33,7	1	7,3	24,4	2	9,9	22,4	2
4,3	18,5	1	13,8	37,4	2	14,7	27,7	2

Предварительно построим диаграмму рассеяния, чтобы убедиться, что предположение линейности регрессионной зависимости не лишено смысла. Для этого в меню *Graphs* выберем команду *Scatter plots*. В полученном окне нажмем кнопку *Variables*., и установим зависимые данные – *X: fonds, Y: product* и опции графика – *Graphs Type: Regular, Fit* (подбор): *Linear.*

Наблюдаем диаграмму рассеяния с подобранной прямой регрессии, параметры которой отражены в ее заголовке (рис. 4.1). Это означает, что уравнение линейной регрессии имеет вид $y = 1{,}4344x + 11{,}5021$.

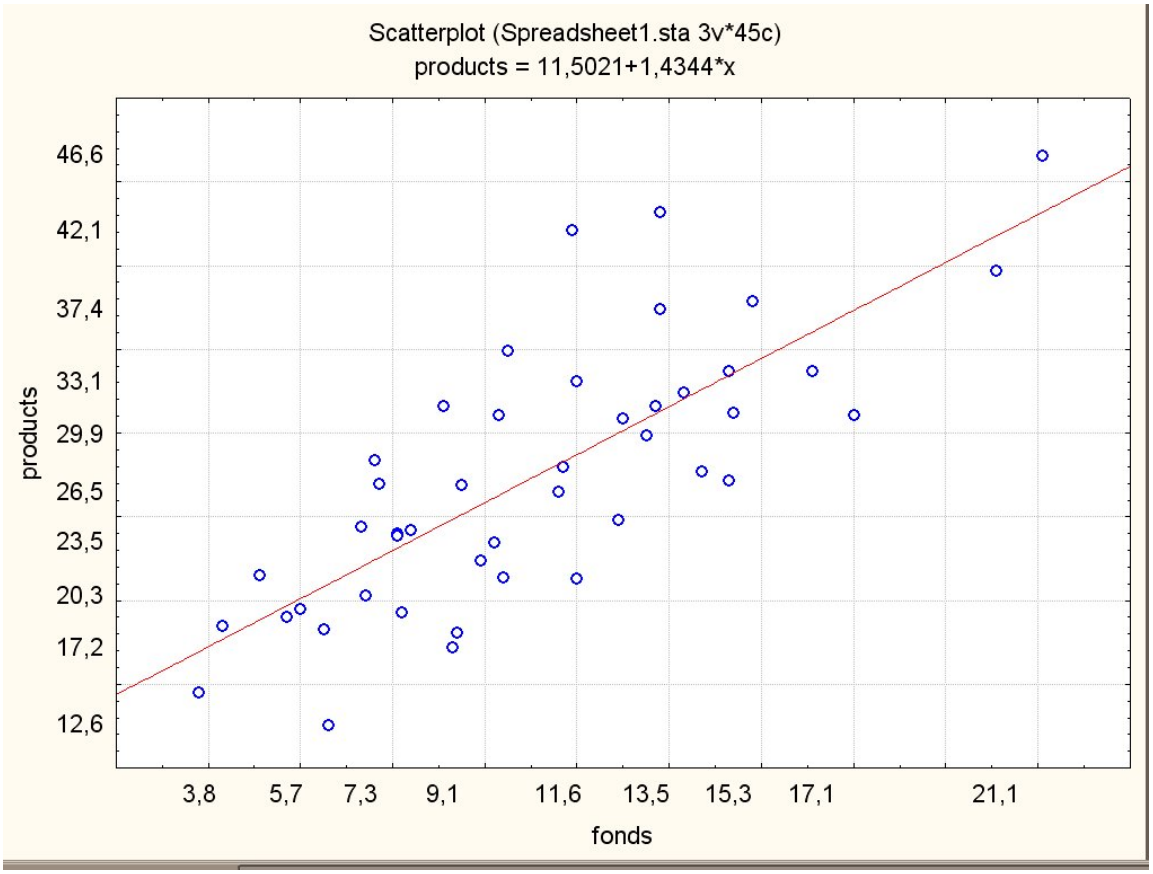

Рис. 4.1. Диаграмма рассеяния

Чтобы получить обратную зависимость, в окне задания опций следует поменять местами переменные *X* и *Y*, то есть переменной *X* назначить колонку *products*, а переменной *Y* – *fonds*. В этом случае уравнение регрессии задается уравнением $y = 0{,}4158x - 0{,}3125$, а прямая имеет вид, представленный на рис. 4.2.

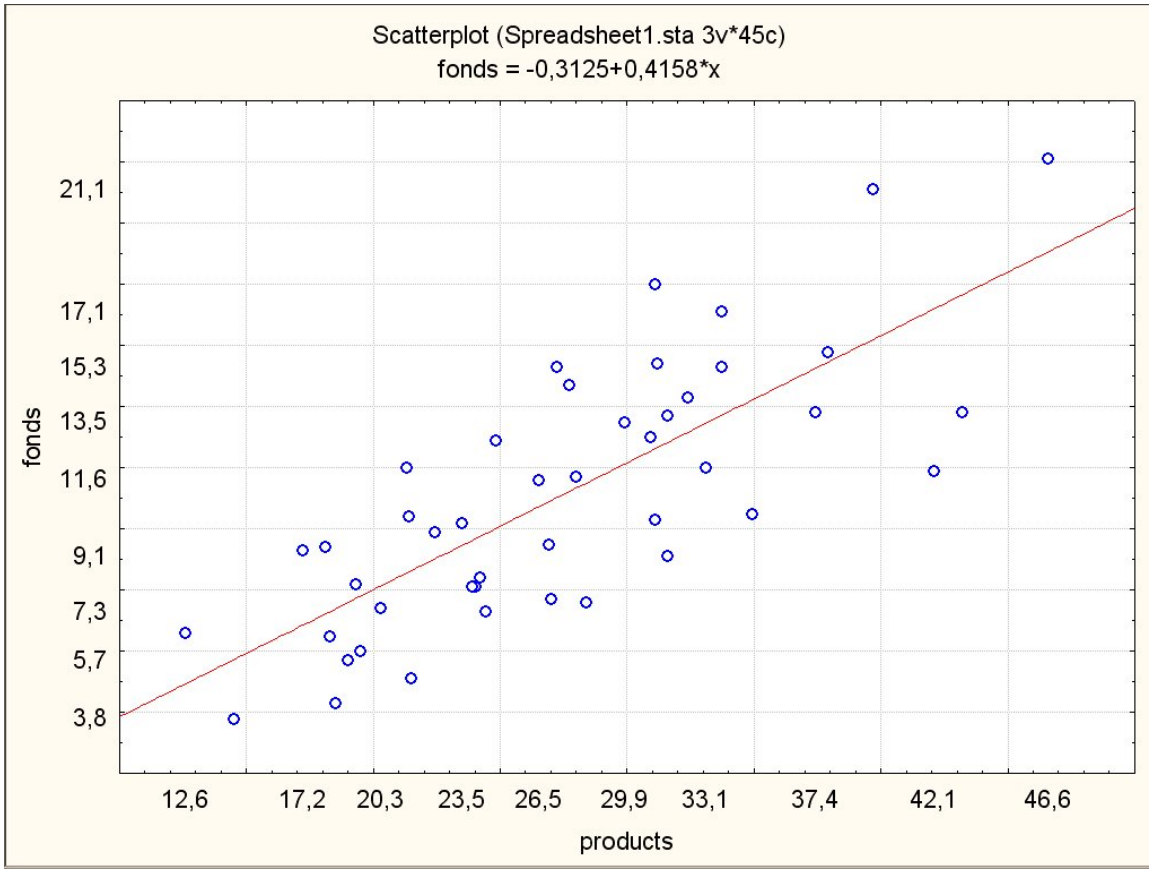

Рис 4.2. Обратная диаграмма рассеяния

По полученным графикам делаем вывод, что имеет смысл проводить регрессионный анализ по имеющимся исходным данным.

Будем работать в модуле *Multiple Regression* (множественная регрессия); меню *Statistics – Multiple Regression*. В качестве зависимой переменной выберем колонку *fonds,* в качестве независимой – колонку *products*, во вкладке *Advanced* установим опцию *Input file* (входной файл): *Raw Data* (необработанные данные).

Нажав кнопку *OK*, получаем основные результаты анализа (рис. 4.3) коэффициент детерминации R^2: 0.597; гипотеза о нулевом значении наклона отклоняется с высоким уровнем значимости $p = 0.000000$ (т.е. $p < 10^{-6}$).

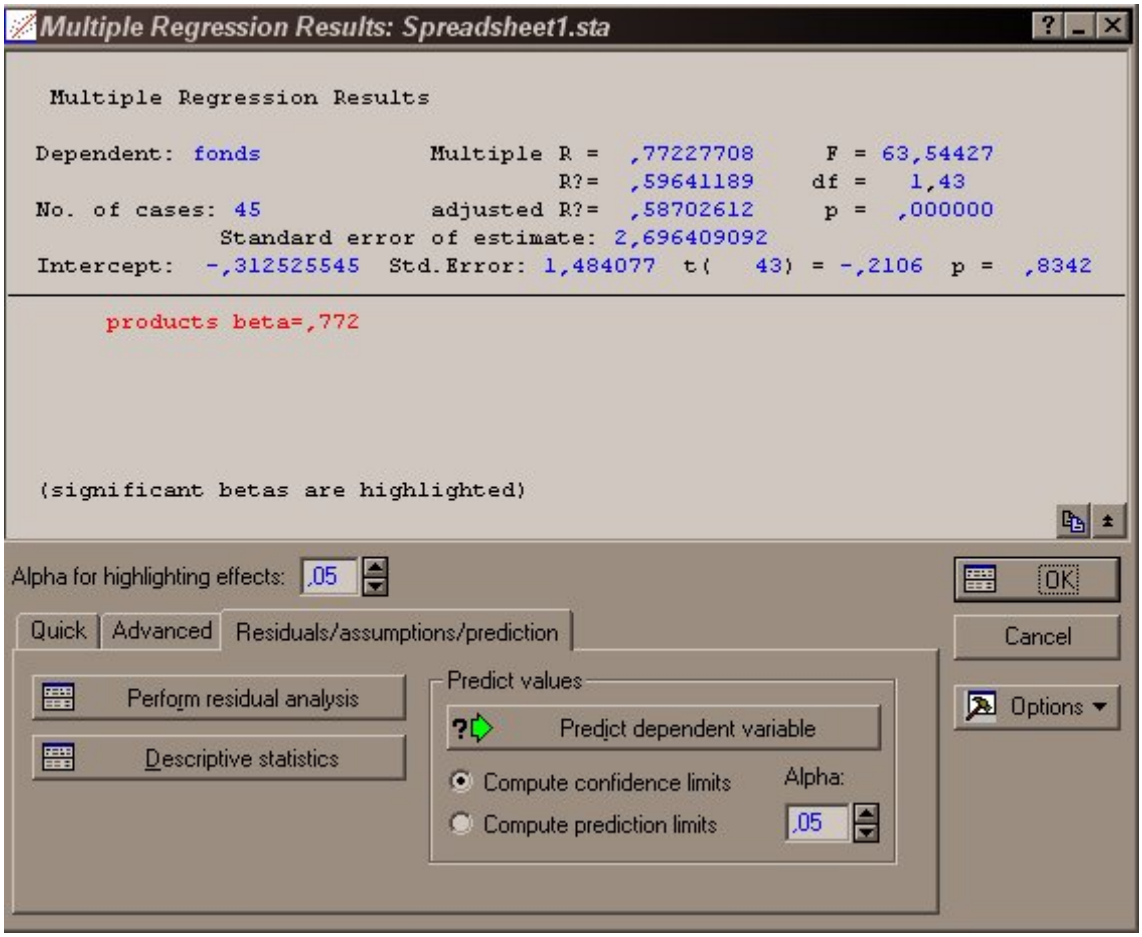

Рис. 4.3. Окно результатов регрессионного анализа

Поясним значения характеристик:

Dependent – имя зависимой переменной (в примере – fonds);

Multiple R – множественный коэффициент корреляции;

F – значение критерия Фишера, F=63, 54427;

R? (R^2) – множественный коэффициент детерминации;

df – количество степеней свободы *F*-критерия;

No. of cases – количество наблюдений;

adjusted R? (R^2) – скорректированный коэффициент детерминации, определяемый по формуле $1-\left(1-R^2\right)\frac{n}{n-2}$;

p – критический уровень значимости модели;

Standard error of estimate – среднеквадратическая ошибка;

Intercept – оценка свободного члена модели регрессии;

Std. Error – стандартная ошибка оценки свободного члена модели регрессии;

t(43) = -0,2106 и p = 0,8342 – значения критерия и критического уровня значимости, используемые для проверки гипотезу о равенстве нулю свободного члена регрессии. В данном случае гипотеза должна быть принята, если уровень значимости равен 0,8342 или ниже.

На вкладке *Quick* нажмем кнопку *Summary Regression Results* и получим таблицу результатов (рис. 4.4).

Regression Summary for Dependent Variable: fonds (Spreadsheet1.sta)
R= ,77227708 R?= ,59641189 Adjusted R?= ,58702612
F(1,43)=63,544 p<,00000 Std.Error of estimate: 2,6964

N=45	Beta	Std.Err. of Beta	B	Std.Err. of B	t(43)	p-level
Intercept			-0,312526	1,484077	-0,210586	0,834205
products	0,772277	0,096880	0,415792	0,052160	7,971466	0,000000

Рис. 4.4. Таблица результатов регрессионного анализа

В заголовке полученной таблицы повторены результаты предыдущего окна; в столбцах приведены: *B* – значения оценок параметров модели регрессии $\beta_0^* = -0{,}312526$ и $\beta_1^* = -0{,}415792$; столбец *St. Err. of B* – параметры стандартных ошибок параметров модели регрессии, соответственно $D^*(\beta_0^*) = 1{,}484077$ и $D^*(\beta_1^*) = 0{,}052160$; столбец *t(43)* – значение статистики Стьюдента (*t*-критерия) для проверки гипотезы о нулевом значении коэффициента (т.е. $\beta_0 = 0$ и $\beta_1 = 1$); столбец *p-level* – минимальный уровень значимости отклонения этой гипотезы. В данном случае, поскольку значения *p-level* очень малы (меньше 10^{-4}), гипотезы о нулевых значениях коэффициентов отклоняются с высокой значимостью. Итак, имеем регрессию:

$$product = 11.5 + 1.43\, fonds,$$

соответствующие стандартные ошибки коэффициентов: 2.1 и 0.18; значение s = 5.01 (*Std Error of estimate* – ошибка прогноза выработки по фондам с помощью этой функции). Значение коэффициента детерминации $R^2 = RI = 0.597$ достаточно велико (доля $R = 0.77$ всей изменчивости объясняется вариацией фондов). Уравнение регрессии показывает, что увеличение основных фондов на 1 млн. денежных единиц приводит к увеличению выработки 1 работника в среднем на $\beta_1 = 1.43$ тыс. денежных единиц.

Многочисленные дополнительные опции модуля регрессии позволяют, например, вычислить результаты описательной статистики (среднее значение и среднее квадратическое отклонение), а также коэффициент корреляции между данными. Для этого можно воспользоваться вкладкой *Advanced*, нажав на ней кнопку *Descriptive Statistics* и выбрав необходимые кнопки. Результат будет отображен в отдельном окне. Нажав на кнопку во вкладке *Matrix*, получим общее окно, приведенное на рис. 4.5.

	Spreadsheet1.sta	
	1 products	2 fonds
products	1,00000	0,77228
fonds	0,77228	1,00000
Means	27,38889	11,07556
Std.Dev.	7,79330	4,19589
No.Cases	45,00000	
Matrix	1,00000	

Рис. 4.5. Описательная статистика и коэффициент корреляции

4.2. Анализ остатков

В окне *Multiple Regression* выберем вкладку *Residuals/assumptions/prediction*, позволяющую оценить остатки и нажмем на кнопку *Perform Residual analysis*. Далее кнопкой активизируем окно (рис. 4.6).

	Predicted & Residual Values (Spreadsheet1.sta) Dependent variable: fonds								
Case No.	**Observed Value**	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1	6,50000	4,82004	1,67996	-1,40632	0,95233	0,804407	1,977737	2,12099	0,150295
2	10,30000	12,77493	-2,47492	0,17827	-1,40297	0,463168	0,031779	-2,65817	0,078263
3	7,70000	10,22687	-2,52688	-0,32930	-1,43242	0,481211	0,108436	-2,73002	0,089108
4	15,80000	17,00096	-1,20096	1,02008	-0,68079	0,662386	1,040558	-1,39808	0,044279
5	7,40000	6,06299	1,33701	-1,15873	0,75792	0,711270	1,342652	1,59657	0,066582
6	14,30000	13,58284	0,71716	0,33920	0,40654	0,482737	0,115057	0,77521	0,007231
7	15,40000	12,83707	2,56293	0,19065	1,45286	0,464263	0,036346	2,75365	0,084384
8	21,10000	18,11962	2,98039	1,24291	1,68950	0,742188	1,544825	3,62142	0,372993
9	22,10000	22,40780	-0,30780	2,09710	-0,17448	1,088581	4,397832	-0,49709	0,015118
10	12,00000	14,01788	-2,01788	0,42586	-1,14388	0,497767	0,181355	-2,19244	0,061493
11	9,50000	10,16473	-0,66473	-0,34168	-0,37682	0,483125	0,116743	-0,71863	0,006224
12	8,10000	8,36245	-0,26245	-0,70068	-0,14877	0,562664	0,490959	-0,29217	0,001395
13	8,40000	8,48674	-0,08674	-0,67593	-0,04917	0,555891	0,456875	-0,09631	0,000148
14	15,30000	14,39076	0,90924	0,50014	0,51542	0,512894	0,250135	0,99320	0,013398
15	4,30000	4,94433	-0,64433	-1,38156	-0,36525	0,794813	1,908711	-0,80845	0,021318
Minimum	4,30000	4,82004	-2,52688	-1,40632	-1,43242	0,463168	0,031779	-2,73002	0,000148
Maximum	22,10000	22,40780	2,98039	2,09710	1,68950	1,088581	4,397832	3,62142	0,372993
Mean	11,88000	11,88000	0,00000	-0,00000	0,00000	0,620491	0,933333	0,03131	0,067482
Median	10,30000	12,77493	-0,26245	0,17827	-0,14877	0,555891	0,456875	-0,29217	0,044279

Рис. 4.6. Наблюдаемые и предсказанные значения остатков

Первые четыре столбца этой таблицы определяют: номера наблюдений (названия областей), фактические (*Observed Value*) и расчетные значения (*Predicted Value*) количества продукции, отклонения фактических данных от расчетных (*Residual*). Четыре последних строки содержат минимальное, максимальное, среднее и медианное значения показателей. Равенство нулю среднего значения остатков свидетельствует о корректности расчетов.

Построим регрессию выработки по фондам для более однородной совокупности – для предприятий федерального подчинения (при $z = 1$). Можно ожидать, что качество подгонки улучшится. Предварительно визуально оценим данные процедурой *Scatterplot.* При отборе наблюдений будем использовать кнопку *Select cases* во вкладке Advanced. Зададим условие отбора в окне *By expression*: $z = 1$. Полученный график отображен на рис. 4.7.

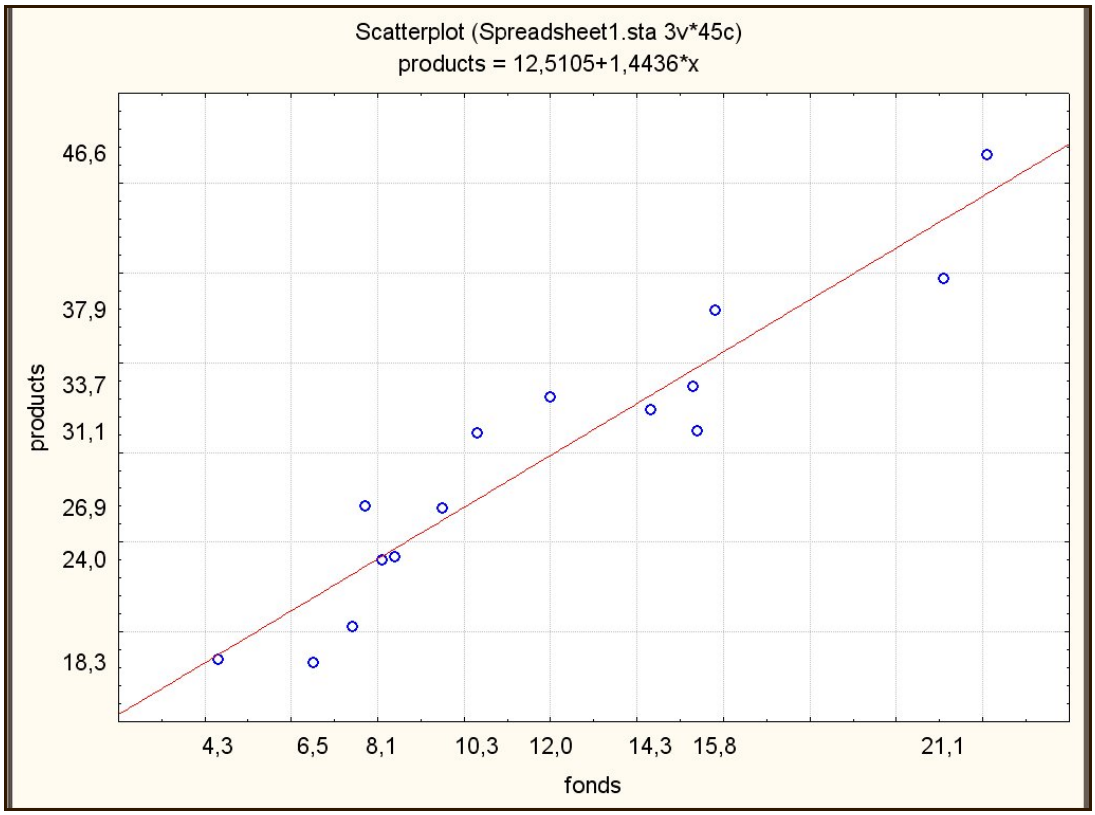

Рис. 4.7. Диаграмма рассеяния по отобранным данным

Возвращаемся в окно *Multiple Regression*. Нажав на кнопку *Select cases*, убеждаемся, что там также автоматически установлено условие отбора $z = 1$ (если нет – устанавливаем это условие).

Получаем результаты анализа (рис. 4.8).

Regression Summary for Dependent Variable: fonds (Spreadsheet1.sta)
R= ,94717253 R?= ,89713581 Adjusted R?= ,88922318
F(1,13)=113,38 p<,00000 Std.Error of estimate: 1,7641

N=15	Beta	Std.Err. of Beta	B	Std.Err. of B	t(13)	p-level
Intercept			-6,55297	1,790036	-3,66080	0,002877
products	0,947173	0,088953	0,62148	0,058365	10,64802	0,000000

Рис. 4.8. Таблица результатов регрессионного анализа по отобранным данным

$$Product = 12.55 + 1.44\, fonds,$$

$$R^2 = RI = 0.897,\ S = 2.68.$$

Коэффициент детерминации увеличился с 0.597 до 0.897, значение s уменьшилось с 5.01 до 2.68; действительно, подгонка улучшилась.

4.3. Контрольное задание

По заданной таблице зависимости признаков X и Y, соответствующей номеру варианта, провести регрессионный анализ:

1) Найти выборочные уравнения прямых линий регрессии X на Y и Y на X.

2) Найти значение выборочного коэффициента корреляции, провести анализ результатов.

3) Провести анализ остатков.

Варианты заданий находятся в практическом задании № 2.

Лабораторная работа №3
ДИСПЕРСИОННЫЙ АНАЛИЗ

Цель: изучение методики применения дисперсионного анализа при проверке гипотезы о равенстве математических ожиданий, либо при установлении того, оказывает ли качественный фактор F существенное влияние на исследуемую величину X.

5.1. Теоретическая часть

Задачей дисперсионного анализа является изучение влияния одного или нескольких факторов на рассматриваемый признак.

Однофакторный дисперсионный анализ используется в тех случаях, когда есть в распоряжении три или более независимые выборки, полученные из одной генеральной совокупности путем изменения какого-либо независимого фактора, для которого по каким-либо причинам нет количественных измерений.

Для этих выборок предполагают, что они имеют разные выборочные средние и одинаковые выборочные дисперсии. Поэтому необходимо ответить на вопрос, оказал ли этот фактор существенное влияние на разброс выборочных средних или разброс является следствием случайностей, вызванных небольшими объемами выборок. Другими словами если выборки принадлежат одной и той же генеральной совокупности, то разброс данных между выборками (между группами) должен быть не больше, чем разброс данных внутри этих выборок (внутри групп).

Пусть x_{ik} – i-й элемент $\left(i=\overline{1,n_k}\right)$ k-выборки $\left(k=\overline{1,m}\right)$, где m – число выборок, n_k – число данных в k-выборке. Тогда $\bar{x}_k$ – выборочное среднее k-выборки определяется по формуле $\bar{x}_k=\frac{1}{n_k}\sum_{i=1}^{m}x_{ik}$. Общее среднее вычисляется по формуле $\bar{x}=\frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{k}x_{ik}$, где $n=\sum_{k=1}^{m}n_k$.

Основное тождество дисперсионного анализа имеет следующий вид:

$$Q=Q_1+Q_2,$$

где Q_1 – сумма квадратов отклонений выборочных средних $\bar{x}_k$ от общего среднего $\bar{x}$ (сумма квадратов отклонений между группами);

Q_2 – сумма квадратов отклонений наблюдаемых значений x_{ik} от выборочной средней $\bar{x}_k$ (сумма квадратов отклонений внутри групп);

Q – общая сумма квадратов отклонений наблюдаемых значений x_{ik} от общего среднего $\bar{x}$.

Расчет этих сумм квадратов отклонений осуществляется по следующим формулам:

$$Q = \sum_{k=1}^{m}\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}\right)^2 = \sum_{k=1}^{m}\sum_{i=1}^{n_k} x_{ik}^2 - n\bar{x}^2,$$

$$Q_1 = \sum_{k=1}^{m} n_k\left(\bar{x}_k - \bar{x}\right)^2 = \sum_{k=1}^{m} n_k \bar{x}_k^2 - n\bar{x}^2,\ Q_2 = \sum_{k=1}^{m}\sum_{i=1}^{n_k}\left(x_{ik} - \bar{x}_k\right)^2 = \sum_{k=1}^{m}\sum_{i=1}^{n_k} x_{ik}^2 - \sum_{k=1}^{m} n_k \bar{x}_k^2 .$$

В качестве критерия необходимо воспользоваться критерием Фишера:

$$F = \frac{Q_1 \big/ (m-1)}{Q_2 \big/ (n-m)}.$$

Если расчетное значение критерия Фишера будет меньше, чем табличное значение $F_{\lambda;m-1;n-m}$ – нет оснований считать, что независимый фактор оказывает влияние на разброс средних значений, в противном случае, независимый фактор оказывает существенное влияние на разброс средних значений (λ – уровень значимости, уровень риска, обычно для экономических задач $\lambda = 0.05$).

5.2. Практическая часть

Задача 1.

Три группы продавцов продавали штучный товар, расфасованный в различные упаковки. После окончания срока распродажи был произведен тестовый контроль над случайно отобранными продавцами из каждой группы. Были получены следующие результаты:

Номер испытания	Уровни фактора		
	F1	F2	F3
1	38	20	21
2	36	24	22
3	35	26	31
4	31	30	34

Решение в пакете STATISTICA.

1. Запустить пакет STATISTICA.

2. Ввести исходные данные для переменных в столбцы VAR1 и VAR2 в следующем виде (нужно добавить 2 Cases: выполним команду *Insert/Add Cases*):

3. Выполнить команду *Statistics/ANOVA*. Появится *меню General ANOVA/MANOVA* (рис. 5.1).

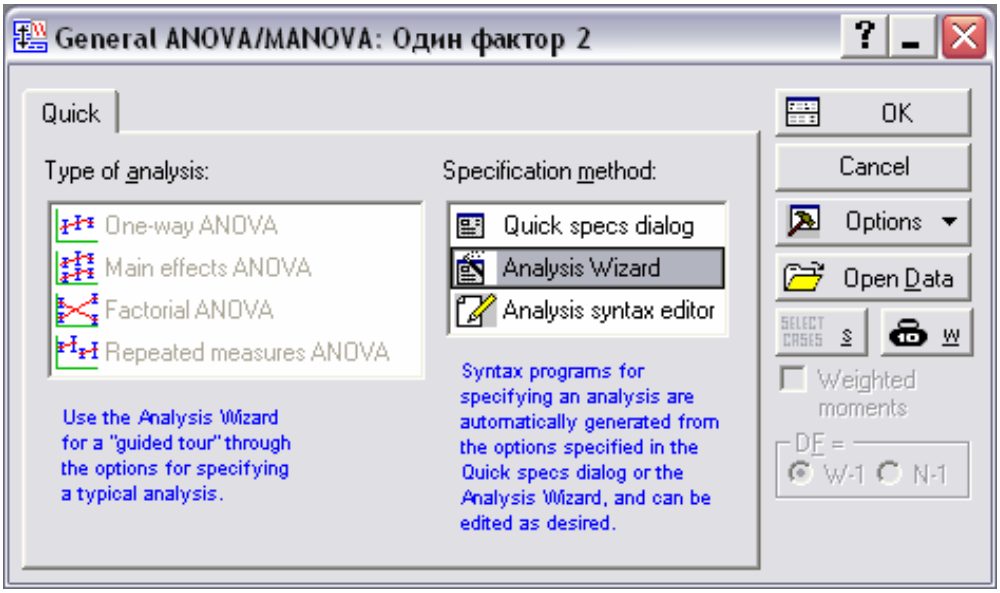

Рис. 5.1. Меню *General ANOVA/MANOVA*

4. В нем выбрать пункт *Analysis Wizard* в колонке *Specification Method*. Нажать **ОК.** Откроется окно *Variables* (рис 5.2).

Определить независимую (VAR1) и зависимую (VAR2) переменные. Нажать ОК.

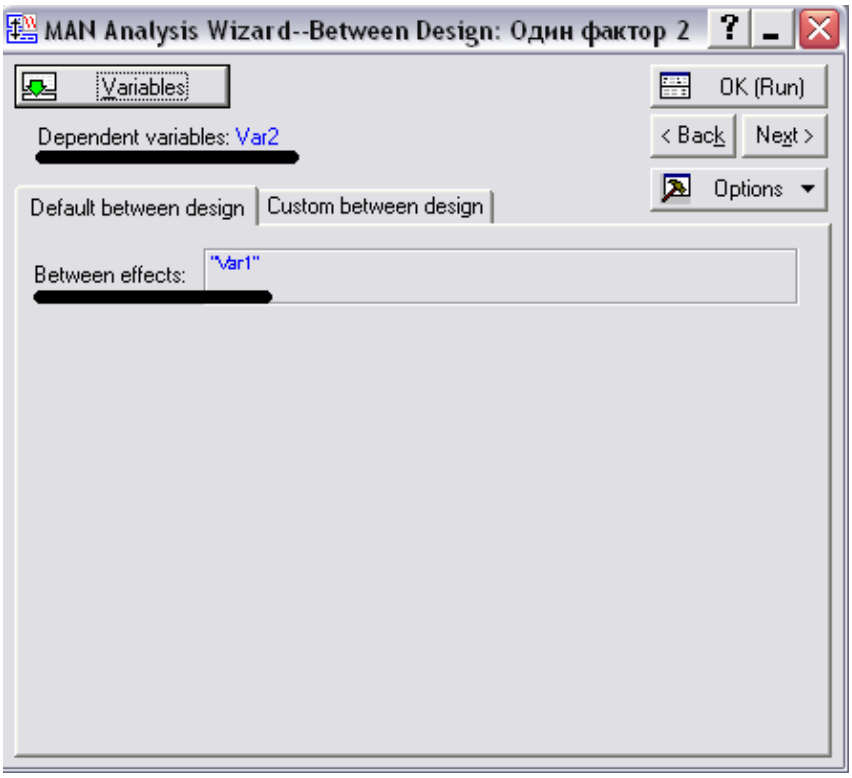

Рис. 5.2. Окно *Variables*

5. Появится панель ***ANOVA Results*.**

6. Для решения данной задачи достаточно нажать кнопку *All effects/Graphs***,** и в открывшемся окне поставить галочку возле *SpreadSheet* (рис.5.3). Нажать **ОК:**

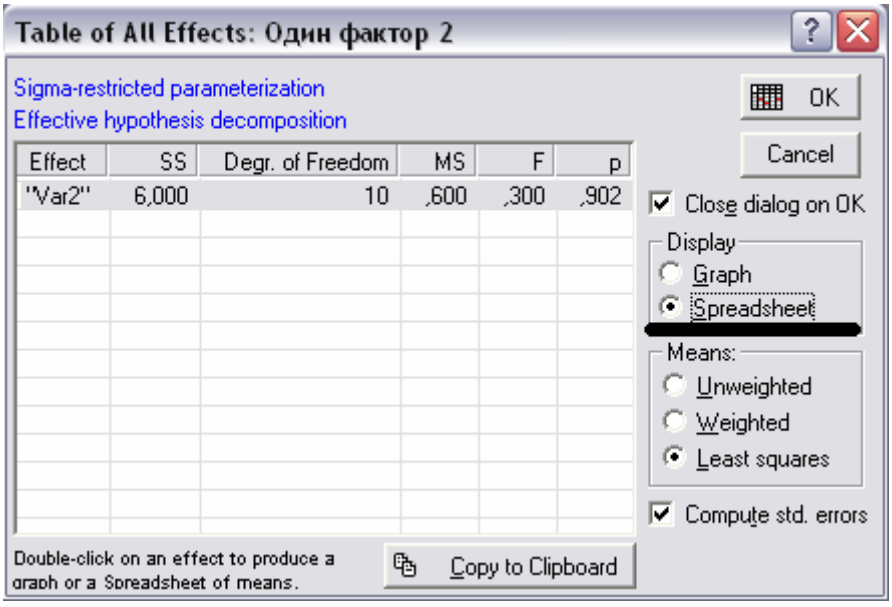

Рис. 5.3. Таблица однофакторного дисперсионного анализа.

7. В открывшемся окне появится результат решения задачи (рис. 5.4).

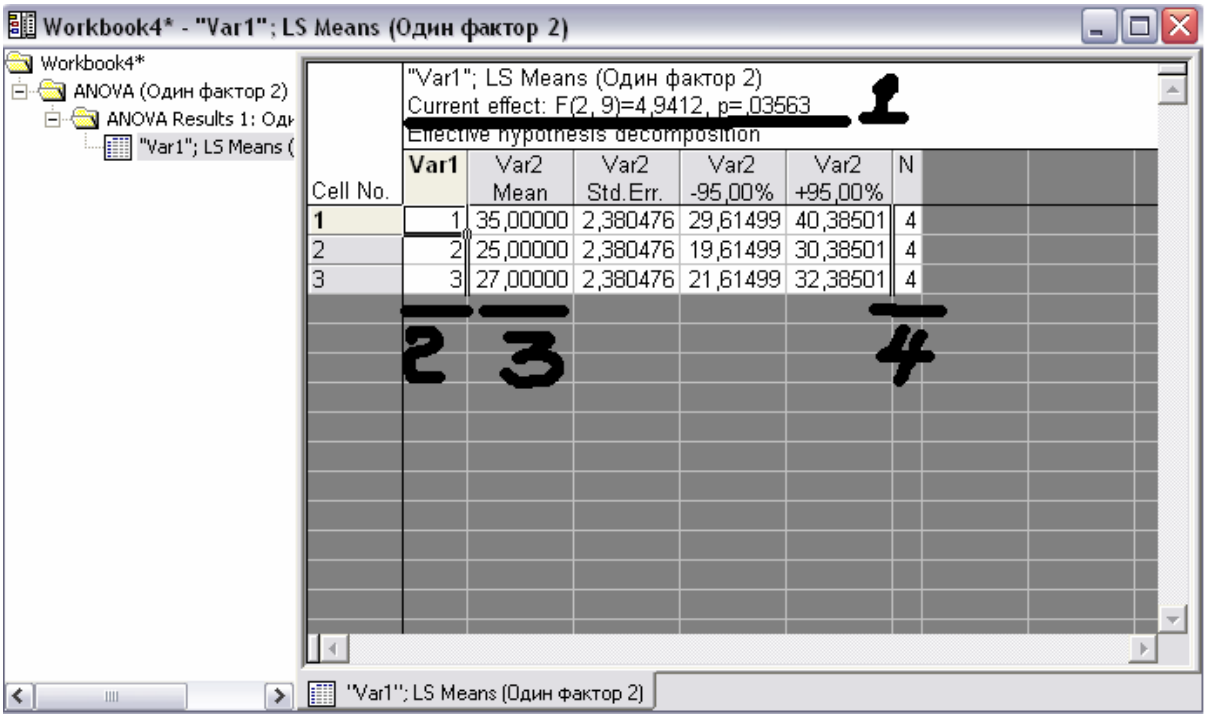

Рис. 5.4. Результат решения задачи

Итак, подчеркнутое предложение **1 –** это ключ-решение: тут показан критерий Фишера-Снедекора, полученный в ходе решения задачи. Этот критерий надо сравнить с табличным F_{kp} Фишера-Снедекора.

В столбце **2** показаны уровни фактора (1,2,3). Следующий столбец **3** показывает групповую среднюю для каждого уровня фактора (X_{cp}). А столбец **4** – показывает количество испытаний на каждом уровне фактора. Сравнив критерий Фишера-Снедекора полученный и табличный, делаем вывод о различии групповых переменных «в целом». Неотмеченные столбцы показывают различного рода погрешности расчетов, поэтому существенной роли не играют.

5.3. Варианты заданий

Вариант 1

1. Проведено по пять испытаний на каждом из трех уровней фактора. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора		
	F1	**F2**	**F3**
1	36	56	52
2	47	61	57
3	50	64	59
4	58	66	58
5	67	66	79

2. Произведено 12 испытаний, из них 5 – на первом уровне фактора, 4 – на втором и 3 – на третьем. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора		
	F1	**F2**	**F3**
1	6	14	12
2	5	11	4
3	12	5	7
4	9	6	
5	10		

Вариант 2

1. Проведено по четыре испытания на каждом из трех уровней фактора. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора		
	F1	**F2**	**F3**
1	2.9	3.5	3.3
2	3.7	3.1	3.3
3	3.4	3.7	3.4
4	3.1	3.0	3.2

2. Произведено 11 испытаний, из них 3 – на первом уровне фактора, 3 – на втором, 3 – на третьем и 2 – на четвертом. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора			
	F1	**F2**	**F3**	**F4**
1	8	9	16	9
2	11	10	9	8
3	8	7	12	

Вариант 3

1. Проведено по пять испытаний на каждом из трех уровней фактора. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора		
	F1	**F2**	**F3**
1	39	100	92
2	57	101	102
3	63	126	104
4	61	128	115
5	65	133	119

2. Произведено 12 испытаний, из них 4 – на первом уровне фактора, 5 – на втором, 3 – на третьем. Методом дисперсионного анализа при уровне значимости $\alpha = 0.1$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора		
	F1	**F2**	**F3**
1	2.8	2.9	3.7
2	3.1	3.3	3.4
3	3.6	3.0	3.7
4	3.2	3.1	
5		3.2	

Вариант 4

1. Проведено по четыре испытания на каждом из трех уровней фактора. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора		
	F1	**F2**	**F3**
1	141	122	101
2	147	128	102
3	148	127	105
4	146	111	106

2. Произведено 13 испытаний, из них 3 – на первом уровне фактора, 2 – на втором, 4 – на третьем и 4 – на четвертом. Методом дисперсионного анализа при уровне значимости $\alpha = 0.01$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора			
	F1	**F2**	**F3**	**F4**
1	83.5	91.5	82.5	91.5
2	85.0	93.0	94.0	95.0
3	87.0		83.5	90.5
4			85.5	89.0

Вариант 5

1. Проведено по четыре испытания на каждом из трех уровней фактора. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора		
i	**F1**	**F2**	**F3**
1	35	30	21
2	32	24	22
3	31	26	34
4	30	20	31

2. Произведено 17 испытаний, из них 5 – на первом уровне фактора, 5 – на втором, 4 – на третьем и 3 – на четвертом. Методом дисперсионного анализа при уровне значимости $\alpha = 0.01$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора			
	F1	**F2**	**F3**	**F4**
1	21	25	28	20
2	23	30	29	22
3	22	32	34	25
4	27	30	30	
5	20	33		

Вариант 6

1. Проведено по четыре испытания на каждом из трех уровней фактора. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора		
	F1	**F2**	**F3**
1	27	24	22
2	23	20	21
3	29	26	36
4	29	30	37

2. Произведено 14 испытаний, из них 5 – на первом уровне фактора, 5 – на втором, 4 – на третьем. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями

Номер испытания	Уровни фактора		
	F1	**F2**	**F3**
1	20	40	33
2	32	42	37
3	27	35	32
4	24	30	35
5	28	34	

Вариант 7

1. Проведено по четыре испытания на каждом из трех уровней фактора. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора		
	F1	**F2**	**F3**
1	51	52	56
2	59	58	56
3	53	66	58
4	59	69	58

2. Произведено 15 испытаний, из них 4 – на первом уровне фактора, 3 – на втором, 4 – на третьем и 4 – на четвертом. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями

Номер испытания	Уровни фактора			
	F1	**F2**	**F3**	**F4**
1	25.7	25.01	25.9	24.6
2	25.75	25.03	25.8	24.8
3	25.8	25.05	25.75	25.0
4	25.95		25.6	25.1

Вариант 8

1. Проведено по четыре испытания на каждом из трех уровней фактора. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Номер испытания	Уровни фактора		
	F1	**F2**	**F3**
1	63	70	70
2	69	72	74
3	72	74	78
4	73	76	82

2. Произведено 14 испытаний, из них 4 – на первом уровне фактора, 3 – на втором, 3 – на третьем и 4 – на четвертом. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями

Номер испытания	Уровни фактора			
	F1	**F2**	**F3**	**F4**
1	7.3	5.4	6.4	7.9
2	8.3	7.1	8.1	9.5
3	8.3	7.4	6.3	9.6
4	8.4			9.1

ПРИЛОЖЕНИЕ

Таблица П1

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

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,31	0,1217	0,62	0,2324	0,93	0,3238	1,25	0,3944
0,01	0,0040	0,32	0,1255	0,63	0,2357	0,94	0,3264	1,26	0,3962
0,02	0,0080	0,33	0,1293	0,64	0,2389	0,95	0,3289	1,27	0,3980
0,03	0,0120	0,34	0,1331	0,65	0,2422	0,96	0,3315	1,28	0,3997
0,04	0,0160	0,35	0,1368	0,66	0,2454	0,97	0,3340	1,29	0,4015
0,05	0,0199	0,36	0,1406	0,67	0,2486	0,98	0,3365	1,30	0,4032
0,06	0,0239	0,37	0,1443	0,68	0,2517	0,99	0,3389	1,31	0,4049
0,07	0,0279	0,38	0,1480	0,69	0,2549	1,00	0,3413	1,32	0,4066
0,08	0,0319	0,39	0,1517	0,70	0,2580	1,01	0,3438	1,33	0,4082
0,09	0,0359	0,40	0,1554	0,71	0,2611	1,02	0,3461	1,34	0,4099
0,10	0,0398	0,41	0,1591	0,72	0,2642	1,03	0,3485	1,35	0,4115
0,11	0,0438	0,42	0,1628	0,73	0,2673	1,04	0,3508	1,36	0,4131
0,12	0,0478	0,43	0,1664	0,74	0,2703	1,05	0,3531	1,37	0,4147
0,13	0,0517	0,44	0,1700	0,75	0,2734	1,06	0,3554	1,38	0,4162
0,14	0,0557	0,45	0,1736	0,76	0,2764	1,07	0,3577	1,39	0,4177
0,15	0,0596	0,46	0,1772	0,77	0,2794	1,08	0,3599	1,40	0,4192
0,16	0,0636	0,47	0,1808	0,78	0,2823	1,09	0,3621	1,41	0,4207
0,17	0,0675	0,48	0,1844	0,79	0,2852	1,11	0,3643	1,42	0,4222
0,18	0,0714	0,49	0,1879	0,80	0,2881	1,12	0,3665	1,43	0,4236
0,19	0,0753	0,50	0,1915	0,81	0,2910	1,13	0,3686	1,44	0,4251
0,20	0,0793	0,51	0,1950	0,82	0,2939	1,14	0,3708	1,45	0,4265
0,21	0,0832	0,52	0,1985	0,83	0,2967	1,15	0,3749	1,46	0,4279
0,22	0,0871	0,53	0,2019	0,84	0,2995	1,16	0,3770	1,47	0,4292
0,23	0,0910	0,54	0,2054	0,85	0,3023	1,17	0,3790	1,48	0,4306
0,24	0,0948	0,55	0,2088	0,86	0,3051	1,18	0,3810	1,49	0,4319
0,25	0,0987	0,56	0,2123	0,87	0,3078	1,19	0,3830	1,50	0,4332
0,26	0,1026	0,57	0,2157	0,88	0,3106	1,20	0,3849	1,51	0,4345
0,27	0,1064	0,58	0,2190	0,89	0,3133	1,21	0,3869	1,52	0,4357
0,28	0,1103	0,59	0,2224	0,90	0,3159	1,22	0,3883	1,53	0,4370
0,29	0,1141	0,60	0,2257	0,91	0,3186	1,23	0,3907	1,54	0,4382

Окончание табл. П1

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,30	0,1179	0,61	0,2291	0,92	0,3212	1,24	0,3925	1,55	0,4394
1,56	0,4406	1,77	0,4616	1,98	0,4761	2,38	0,4913	2,80	0,4974
1,57	0,4418	1,78	0,4625	1,99	0,4767	2,40	0,4918	2,82	0,4976
1,58	0,4429	1,79	0,4633	2,00	0,4772	2,42	0,4922	2,84	0,4977
1,59	0,4441	1,80	0,4641	2,02	0,4783	2,44	0,4927	2,86	0,4979
1,60	0,4452	1,81	0,4649	2,04	0,4793	2,46	0,4931	2,88	0,4980
1,61	0,4463	1,82	0,4656	2,06	0,4803	2,48	0,4934	2,90	0,4981
1,62	0,4474	1,83	0,4664	2,08	0,4812	2,50	0,4938	2,92	0,4982
1,63	0,4484	1,84	0,4671	2,10	0,4821	2,52	0,4941	2,94	0,4984
1,64	0,4495	1,85	0,4678	2,12	0,4830	2,54	0,4945	2,96	0,4985
1,65	0,4505	1,86	0,4686	2,14	0,4838	2,56	0,4948	2,98	0,4986
1,66	0,4515	1,87	0,4693	2,16	0,4846	2,58	0,4951	3,00	0,49865
1,67	0,4525	1,88	0,4699	2,18	0,4854	2,60	0,4953	3,20	0,49931
1,68	0,4535	1,89	0,4706	2,20	0,4861	2,62	0,1956	3,40	0,49966
1,69	0,4545	1,90	0,4713	2,22	0,4868	2,64	0,4959	3,60	0,499841
1,70	0,4554	1,91	0,4719	2,24	0,4875	2,66	0,4961	3,80	0,499928
1,71	0,4564	1,92	0,4726	2,26	0,4881	2,68	0,4963	4,00	0,499968
1,72	0,4573	1,93	0,4732	2,28	0,4887	2,70	0,4965	4,50	0,499997
1,73	0,4582	1,94	0,4738	2,30	0,4893	2,72	0,4967	5,00	0,499999
1,74	0,4591	1,95	0,4744	2,32	0,4898	2,74	0,4969		
1,75	0,4599	1,96	0,4750	2,34	0,4904	2,76	0,4971		
1,76	0,4608	1,97	0,4756	2,36	0,4909	2,78	0,4973		

Таблица П2

Таблица значений $t_\gamma = t(\gamma, n)$

n	γ			n	γ		
	0,95	**0,99**	**0,999**		**0,95**	**0,99**	**0,999**
5	2,78	4,60	8,61	**20**	2,093	2,861	3,883
6	2,58	4,03	6,86	**25**	2,064	2,797	3,745
7	2,45,	3,71	5,96	**30**	2,045	2,756	3,659
8	2,37	3,50	5,41	**35**	2,032	2,720	3,600
9	2,31	3,36	5,04	**40**	2,023	2,708	3,558
10	2,26	3,25	4,78	**45**	2,016	2,692	3,527
11	2,23	3,17	4,59	**50**	2,009	2,679	3,502
12	2,20	3,11	4,44	**60**	2,001	2,662	3,464
13	2,18	3,06	4,32	**70**	1,996	2,649	3,439
14	2,16	3,01	4,22	**80**	1,991	2,640	3,418
15	2,15	2,98	4,14	**90**	1,987	2,633	3,403
16	2,13	2,95	4,07	**100**	1,984	2,627	3,392
17	2,12	2,92	4,02	**120**	1,980	2,617	3,374
18	2,11	2,90	3,97	∞	1,960	2,576	3,291
19	2,10	2,88	3,92				

Таблица П3

Критические точки распределения χ^2

число степеней свободы k	Уровень значимости α					
	0,01	**0,025**	**0,05**	**0,95**	**0,975**	**0,99**
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	1,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	4,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Таблица П4

Таблица значений $q = q(\gamma, n)$

n	γ			n	γ		
	0,95	**0,99**	**0,999**		**0,95**	**0,99**	**0,999**
5	1,37	2,67	5,64	**20**	0,37	0,58	0,88
6	1,09	2,01	3,88	**25**	0,32	0,49	0,73
7	0,92	1,62	2,98	**30**	0,28	0,43	0,63
8	0,80	1,38	2,42	**35**	0,26	0,38	0,56
9	0,71	1,20	2,06	**40**	0,24	0,35	0,50
10	0,65	1,08	1,80	**45**	0,22	0,32	0,46
11	0,59	0,98	1,60	**50**	0,21	0,30	0,43
12	0,55	0,90	1,45	**60**	0,188	0,269	0,38
13	0,52	0,83	1,33	**70**	0,174	0,245	0,34
14	0,48	0,78	1,23	**80**	0,161	0,226	0,31
15	0,46	0,73	1,15	**90**	0,151	0,211	0,29
16	0,44	0,70	1,07	**100**	0,143	0,198	0,27
17	0,42	0,66	1,01	**150**	0,115	0,160	0,211
18	0,40	0,63	0,96	**200**	0,099	0,136	0,185
19	0,39	0,60	0,92	**250**	0,089	0,120	0,162

Таблица П5

Критические точки распределения Стьюдента

число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0,10	**0,05**	**0,02**	**0,01**	**0,002**	**0,001**
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	5,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	5,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	2,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,54	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	**0,025**	**0,01**	**0,005**	**0,001**	**0,0005**
	Уровень значимости α (односторонняя критическая область)					

Таблица П6

Критические точки распределения F Фишера-Снедекора

k_1 – число степеней свободы большей дисперсии;

k_2 – число степеней свободы меньшей дисперсии;

уровень значимости $\alpha = 0{,}05$

k_2 \ k_1	**1**	**2**	**3**	**4**	**5**	**6**	**8**	**12**	**24**
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04
2	18,512	18,999	19,163	19,248	19,298	19,329	19,371	19,414	19,453
3	10,129	9,552	9,276	9,118	9,014	8,941	8,844	8,744	8,638
4	7,710	6,945	6,591	6,388	6,257	6,164	6,041	5,912	5,774
5	6,607	5,786	5,410	5,192	5,050	4,950	4,818	4,678	4,527
6	5,987	5,143	4,756	4,388	4,284	4,147	4,000	3,841	3,669
7	5,591	4,737	4,347	4,121	3,972	3,866	3,725	3,574	3,410
8	5,317	4,459	4,067	3,838	3,688	3,580	3,438	3,284	3,116
9	5,117	4,256	3,863	3,633	3,482	3,374	3,230	3,073	2,900
10	4,965	4,103	3,708	3,478	3,326	3,217	3,072	2,913	2,737
11	4,844	3,982	3,587	3,357	3,204	3,094	2,948	2,778	2,609
12	4,747	3,885	3,490	3,259	3,106	2,999	2,848	2,686	2,505
13	4,667	3,805	3,410	3,179	3,025	2,915	2,767	2,604	2,420
14	4,600	3,739	3,344	3,112	2,958	2,848	2,699	2,534	2,349
15	4,543	3,683	3,287	3,056	2,901	2,790	2,641	2,475	2,288
16	4,494	3,634	3,239	3,007	2,853	2,741	2,591	2,424	2,235
17	4,451	3,592	3,197	2,965	2,810	2,699	2,548	2,381	2,190
18	4,414	3,555	3,160	2,928	2,773	2,661	2,510	2,342	2,150
19	4,381	3,522	3,127	2,895	2,740	2,629	2,477	2,308	2,114
20	4,351	3,493	3,098	2,866	2,711	2,599	2,447	2,278	2,083
21	4,325	3,467	3,072	2,840	2,685	2,573	2,421	2,250	2,054
22	4,301	3,443	3,049	2,817	2,661	2,549	2,397	2,226	2,028
23	4,279	3,422	3,028	2,795	2,640	2,528	2,375	2,203	2,005
24	4,260	3,403	3,009	2,777	2,621	2,508	2,355	2,183	1,984
25	4,242	3,385	2,991	2,759	2,603	2,490	2,337	2,165	1,965
26	4,225	3,369	2,975	2,743	2,587	2,474	2,321	2,148	1,947
27	4,210	3,354	2,961	2,728	2,572	2,459	2,305	2,132	1,930
28	4,196	3,340	2,947	2,714	2,558	2,445	2,292	2,118	1,915
29	4,183	3,328	2,934	2,702	2,545	2,432	2,278	2,104	1,901
30	4,171	3,316	2,922	2,690	2,534	2,421	2,266	2,092	1,887
60	4,001	3,151	2,758	2,525	2,368	2,254	2,097	1,918	1,700
120	3,920	3,072	2,680	2,447	2,290	2,175	2,016	1,834	1,608

ЛИТЕРАТУРА

1. Боровиков, В. STATISTICA: искусство анализа данных на компьютере / В. Боровиков. – СПб.: Питер, 2003. – 688 с.
2. Корн, Г. Статистические методы построения эмпирических формул / Г. Корн, Т. Корн. – М.: Высшая школа, 1988.
3. Айвазян, С.А. Прикладная статистика / С.А. Айвазян. – М.: Финансы и статистика, 1989.
4. Тарасевич, Ю.Ю. Математическое и компьютерное моделирование. Вводный курс / Ю.Ю. Тарасевич. – М.: Едиториал-УРСС, 2001. – 144 с.
5. Гмурман, В.Е. Теория вероятностей и математическая статистика. Учебное пособие для вузов / В.Е. Гмурман. – М.: Высшая школа, 1998. – 479 с.