

С. Г. Валеев, В. Н. Клячкин

---

# **ПРАКТИКУМ ПО ПРИКЛАДНОЙ СТАТИСТИКЕ**

*Допущено УМО по образованию в области  
Прикладной математики и управления качеством  
в качестве учебного пособия для студентов  
высших учебных заведений, обучающихся по  
направлению подготовки 230400 «Прикладная математика»  
специальности 230401 «Прикладная математика»*

Ульяновск  
2008

УДК 519.24 (075)  
ББК 22.172  
В 11

*РЕЦЕНЗЕНТЫ:*

Кафедра «Прикладная математика»  
Ульяновского государственного университета  
(зав. кафедрой д-р физ.-мат. наук, профессор А. А. Бутов);

А. Г. Варжапетян,  
Засл. деятель науки РФ, д-р техн. наук, профессор  
(Санкт-Петербургский государственный университет  
аэрокосмического приборостроения - ГУАП)

**Валеев С. Г.**

В 11 Практикум по прикладной статистике : учебное пособие / С. Г. Валеев,  
В. Н. Клячкин. – Ульяновск : УлГТУ, 2008. – 129 с.: ил.  
ISBN 978-5-9795-0318-9

В пособии содержатся краткие сведения об алгоритмах прикладной математической статистики, примеры расчетов в среде электронных таблиц Excel и системе Statistica, а также варианты для выполнения индивидуальных заданий.

Для студентов технических и экономических специальностей вузов, изучающих курс «Теория вероятностей и математическая статистика».

**УДК 519.24 (075)**  
**ББК 22.172**

Учебное издание  
ВАЛЕЕВ Султан Галимзянович  
КЛЯЧКИН Владимир Николаевич

**Практикум по прикладной статистике**  
Учебное пособие

Редактор М. Штаева

Подписано в печать 11.12.2008. Формат 60×84/16.  
Усл. печ. л. 7,67. Тираж 150 экз.

Ульяновский государственный технический университет  
432027, Ульяновск, ул. Северный Венец, д. 32.

Типография УлГТУ, 432027, Ульяновск, ул. Северный Венец, д. 32.

ISBN 978-5-9795-0318-9

© С. Г. Валеев, В. Н. Клячкин, 2008  
© Оформление. УлГТУ, 2008

# ОПИСАТЕЛЬНАЯ СТАТИСТИКА

## 1.1.

### Способы представления выборки

Рассмотрим совокупность объектов, однородную относительно некоторого признака. Например, если этой совокупностью является партия деталей, то представляет интерес соответствие параметров этих деталей техническим требованиям. Чтобы сделать какие-то выводы об этой партии деталей, можно провести сплошное обследование, то есть изучить каждую деталь. Однако гораздо чаще из всей совокупности отбирают ограниченное количество деталей и по результатам его изучения делают заключение обо всей партии.

*Генеральной совокупностью* называется вероятностное пространство  $(\Omega, F, P)$ , то есть пространство элементарных событий  $\Omega$  с заданным на нем полем событий  $F$  и вероятностями  $P$ , – и определенная на этом пространстве случайная величина  $X$ . Эта случайная величина  $X$  имеет определенную функцию распределения  $F(x)$  и соответствующие числовые характеристики.

*Выборкой* объема  $n$  называется последовательность  $n$  независимых одинаково распределенных случайных величин  $X_1, X_2, \dots, X_n$ , распределение каждой из которых совпадает с распределением исследуемой случайной величины  $X$ . Выборка – это результат  $n$  независимых последовательных наблюдений за случайной величиной  $X$  из рассматриваемой генеральной совокупности. Результат наблюдений  $x_1, x_2, \dots, x_n$  – одна из многих реализаций многомерной случайной величины  $X_1, X_2, \dots, X_n$ .

Основная задача статистики – по результатам исследования выборки дать заключение о характеристиках генеральной совокупности.

Для получения достоверных результатов выборка должна правильно отражать пропорции генеральной совокупности, то есть быть *репрезентативной*. Очевидно, если партия деталей изготовлена рабочими разной

квалификации, а в выборку попали лишь детали, изготовленные рабочим с более высокой квалификацией, вряд ли можно ожидать правильные данные для всей партии деталей. Можно показать, что выборка репрезентативна, если она отобрана из генеральной совокупности случайным образом. На практике такой отбор не всегда легко осуществим, поэтому используют различные способы отбора, обеспечивающие случайность в большей или меньшей степени.

*Вариационным рядом* называется последовательность упорядоченных элементов выборки  $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ , где

$$x^{(1)} \leq x^{(2)} \leq \dots \leq x^{(n)}, .$$

Если объем выборки достаточно велик, то ее обработка оказывается громоздкой; в этом случае элементы выборки объединяют в группы. Для этого интервал  $[x^{(1)}, x^{(n)}]$  разбивается на  $k$  равных интервалов. Количество интервалов  $k$  в зависимости от объема выборки  $n$  обычно принимают от 8 до 20, или вычисляют по эмпирической формуле

$$k = 1 + 3,32 \lg n.$$

Далее определяются частоты  $n_i$  — количество элементов выборки, попавших в  $i$ -ый интервал. Получающийся группированный статистический ряд содержит середины интервалов  $z_i$  и частоты  $n_i$  ( $i = 1, \dots, k$ ). Кроме того,

подсчитываются накопленные частоты  $\sum_{j=1}^i n_j$ , относительные частоты  $\frac{n_i}{n}$ ,

накопленные относительные частоты  $\sum_{j=1}^i \frac{n_j}{n}$ ;  $i = 1, \dots, k$ .

Пусть  $x_1, x_2, \dots, x_n$  — выборка из генеральной совокупности с функцией распределения  $F(x)$ . *Выборочным распределением* называется распределение дискретной случайной величины, принимающей значения  $x_1, x_2, \dots, x_n$  с вероятностями  $1/n$ . Соответствующая функция распределения  $F^*(x)$  называется выборочной или эмпирической функцией распределения и определяется по значениям накопленных частот. При  $x \leq x^{(1)}$   $F(x) = 0$ ; при  $x > x^{(n)}$   $F(x) = 1$ . На промежутке  $[x^{(1)}, x^{(n)}]$   $F^*(x)$  — неубывающая кусочно-постоянная функция.

Можно показать, что при большом объеме выборки эмпирическая функция распределения стремится к функции распределения генеральной совокупности.

*Гистограмма* частот группированной выборки – это график кусочно-постоянной функции, принимающей на каждом из интервалов значение  $n_i/w$  ( $w = (x^{(n)} - x^{(1)})/k$  – ширина интервала). Аналогично по значениям  $n_i/nw$  строится гистограмма относительных частот. Нетрудно показать, что площадь фигуры под гистограммой частот равна объему выборки  $n$ , а под гистограммой относительных частот – единице.

*Полигоном* частот называется график ломаной с вершинами в точках  $(z_i, n_i)$ , а полигоном относительных частот – в точках  $(z_i, n_i/n)$ .

При увеличении объема выборки и уменьшении интервала группирования гистограмма и полигон относительных частот могут рассматриваться как статистические аналоги плотности распределения генеральной совокупности  $f(x)$ .

## 1.2.

### Числовые характеристики выборки

Числовые характеристики выборочного распределения определяются по соответствующим формулам для дискретных случайных величин с учетом того, что вероятности  $p_i = 1/n_i$ .

Основными характеристиками выборки являются:

– математическое ожидание (выборочное среднее):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i; \quad (1.1)$$

для группированного ряда

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k z_i n_i; \quad (1.2)$$

– выборочная дисперсия

$$D_X^* = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (1.3)$$

или, учитывая, что  $D_X = m_{X^2} - m_X^2$ ,

$$D_X^* = \frac{1}{n} \sum x_i^2 - \bar{x}^2, \quad (1.4)$$

для группированного ряда

$$D_X^* = \frac{1}{n} \sum z_i^2 n_i - \bar{x}^2; \quad (1.5)$$

– выборочное среднеквадратическое (стандартное) отклонение

$$\sigma_X = \sqrt{D_X^*}; \quad (1.6)$$

– выборочная мода: для унимодального (одновершинного) распределения это элемент выборки  $Mo_X$ , встречающийся с наибольшей частотой;

– выборочная медиана – число  $Me_X$ , которое делит вариационный ряд на две части, содержащие одинаковое число элементов. Если объем выборки  $n = 2l+1$  (нечетен), то

$$Me_X = x^{(l+1)}.$$

Если же  $n = 2l$ , то

$$Me_X = (x^{(l+1)} + x^{(l+1)})/2;$$

– выборочный коэффициент асимметрии

$$a_X^* = \frac{\mu_3^*}{(\sigma_X^*)^3}, \quad (1.7)$$

где  $\mu_k = \frac{1}{n} \sum (x_i - \bar{x})^k$  – центральный момент  $k$ -го порядка ( $k = 3$ );

– выборочный коэффициент эксцесса

$$e_X^* = \frac{\mu_4^*}{(\sigma_X^*)^4} - 3. \quad (1.8)$$

### 1.3.

#### Пример расчета

Стоимость книги по математической статистике в тридцати различных интернет-магазинах оказалась (в рублях):

200, 198, 201, 203, 203, 204, 196, 200, 203, 198, 199, 197, 197, 199, 199, 196, 199, 200, 201, 200, 200, 200, 203, 200, 200, 199, 204, 202, 205, 199.

Построить таблицу частот, разбив данные на 6 интервалов, график выборочной функции распределения и гистограмму частот. Вычислить числовые характеристики выборки.

Объем выборки – количество ее элементов  $n = 30$ .

Строим вариационный ряд:

196, 196, 197, 197, 198, 198, 199, 199, 199, 199, 199, 199, 200, 200, 200, 200, 200, 200, 200, 200, 201, 201, 202, 203, 203, 203, 203, 204, 204, 205.

Минимальное значение ряда 196, максимальное – 205, размах выборки –  $R = 205 - 196 = 9$ , длина интервала –  $w = 9/6 = 1,5$ .

При построении таблицы частот в качестве нижней границы первого интервала принято минимальное значение выборки. При подсчете частот в случае совпадения элемента выборки с верхней границей соответствующий элемент учитывался в данном интервале.

Таблица частот имеет вид:

№	Границы	$z_i$	$n_i$	$n_i/n$	$\sum n_i/n$	$n_i/wn$
1	196 – 197,5	196,75	4	0,133	0,133	0,089
2	197,5 – 199	198,25	8	0,267	0,400	0,178
3	199 – 200,5	199,75	8	0,267	0,667	0,178
4	200,5 – 202	201,25	3	0,100	0,767	0,067
5	202 – 203,5	202,75	4	0,133	0,900	0,089
6	203,5 – 205	204,25	3	0,100	1	0,067

На рис. 1.1 показана соответствующая гистограмма частот, а на рис. 1.2 – график выборочной функции распределения.

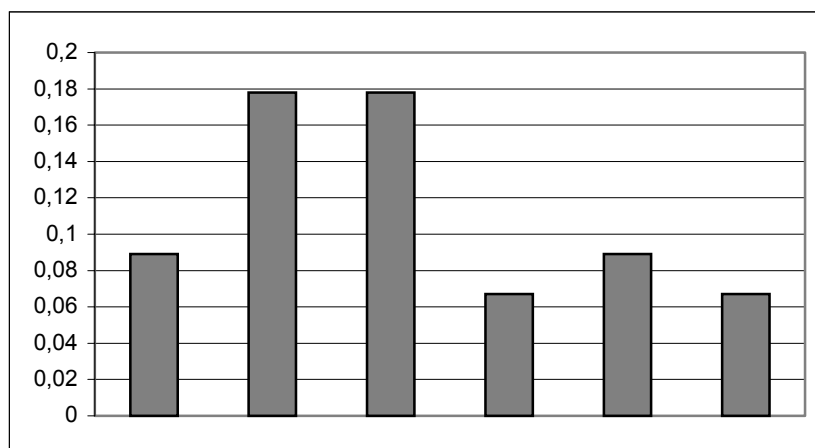


Рис. 1.1.



Рис. 1.2.

Выборочная средняя – это средняя стоимость книги по всем тридцати магазинам:



$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{30} (196 + 196 + 197 + \dots + 204 + 205) = 200,17;$$

выборочная медиана

$$Me^* = 200$$

(среднее между 15-м и 16-м элементами вариационного ряда);

выборочная дисперсия (характеристика рассеяния цен)

$$D_X^* = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1}{30} (196^2 + \dots + 204^2 + 205^2) - 200,17^2 = 5,41;$$

выборочное стандартное отклонение

$$\sigma_X^* = \sqrt{D_X^*} = \sqrt{5,41} = 2,32.$$

## 1.4. Описательная статистика в Excel

Для использования электронных таблиц Excel при работе со статистическими методами могут применяться как обычные средства, такие, как вставка функций (в первую очередь статистических), мастер диаграмм, так и специальные, в частности, надстройка «Пакет анализа» (рис. 1.3).

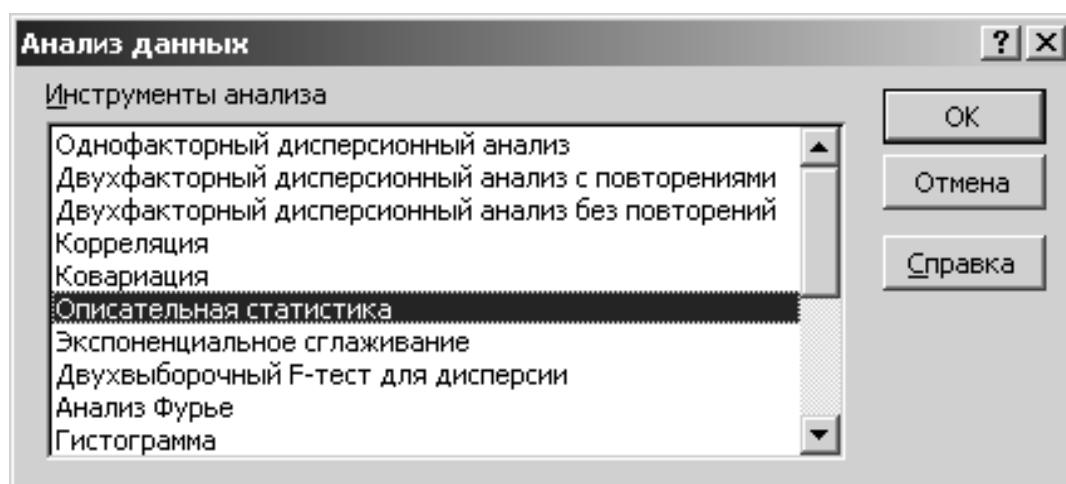


Рис. 1.3

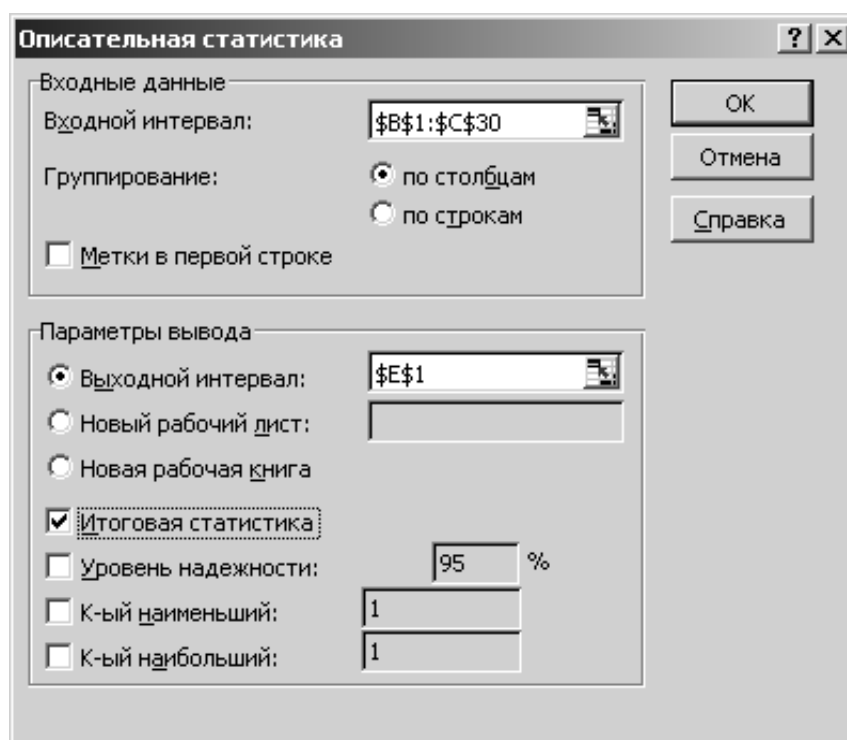


Рис. 1.4

Для определения числовых характеристик выборки можно воспользоваться статистическими функциями, однако большинство характеристик можно

получить проще, используя инструмент *Описательная статистика* пакета анализа. На рис. 1.4 показано заполнение соответствующего диалогового окна; результаты расчета см. на рис. 1.8.

При необходимости расчета других числовых характеристик используется кнопка *Вставка функций*. Например, для расчета среднего геометрического значения (рис. 1.5) необходимо ввести =СРГЕОМ(B1:B30) (*Вставка функций / Категория – статистические / Функция: СРГЕОМ / ОК / Число1: B1:B30 – протаскиванием мышью / ОК – рис. 1.6*).

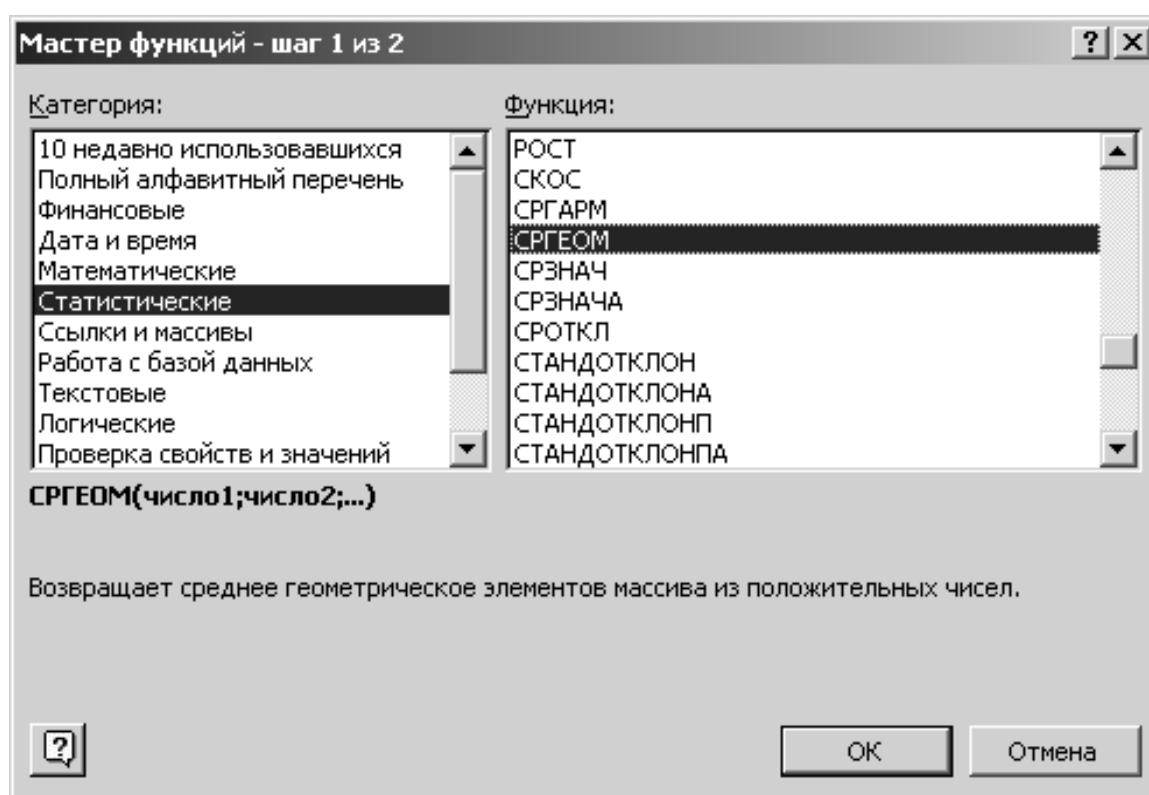


Рис. 1.5

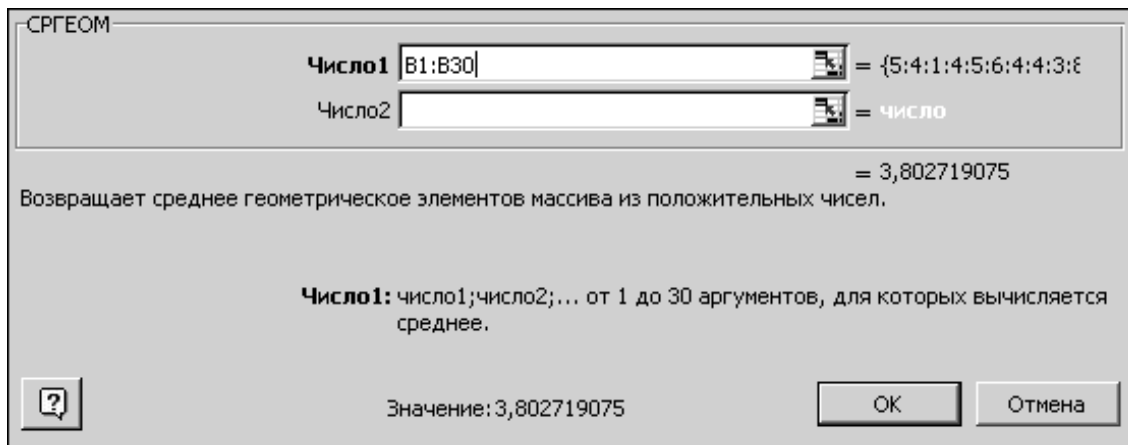


Рис. 1.6

Наиболее простой способ построения гистограммы частот в Excel – использование инструмента *Гистограмма* (рис. 1.7). Построим гистограмму частот и график выборочной функции распределения (в терминологии Excel – интегральный процент: значения накопленных относительных частот вычисляются в процентах) для следующей выборки.

Замерялись отклонения толщины бетонных блоков от номинала. Результаты измерений представлены в таблице:

5	4	1	4	5	6	4	4	3	8	3	5	5	2	7
5	7	2	4	9	2	3	3	3	2	2	2	6	4	10

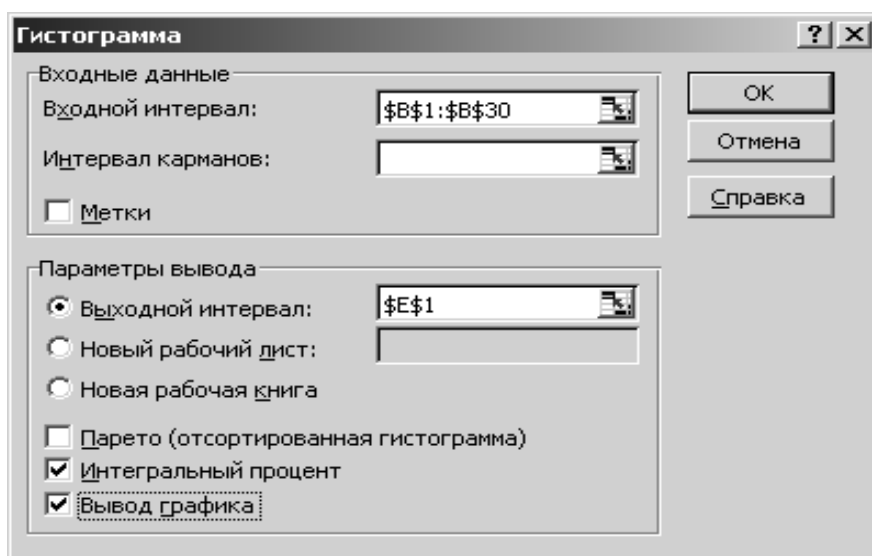


Рис. 1.7

Если поле *Интервал карманов* (границы интервалов) не заполнять, границы будут определены автоматически. Результат представлен на рис. 1.8.

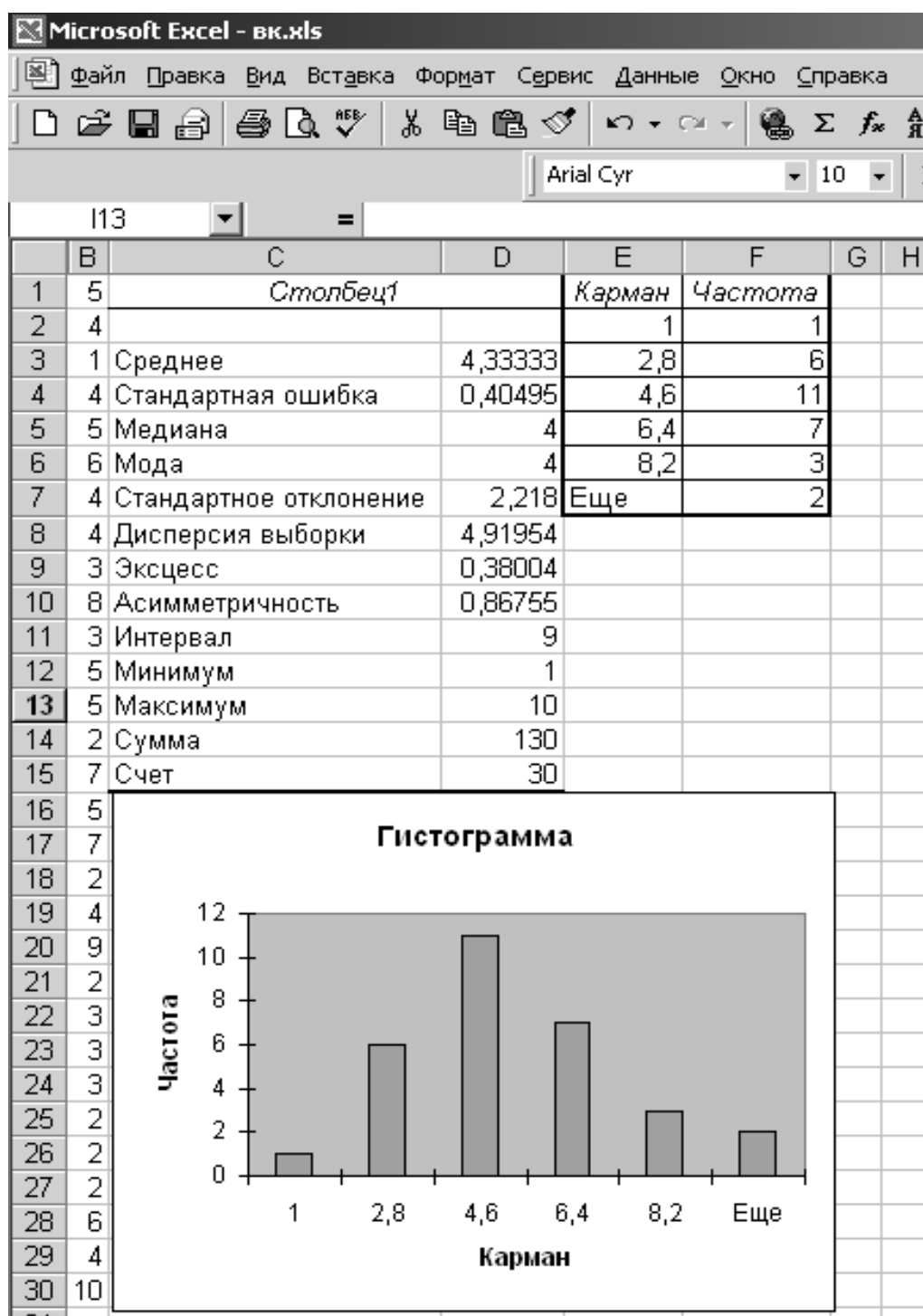


Рис. 1.8

Для изменения числа интервалов или границ интервалов необходимо подготовить границы интервалов (карманы) вручную: на рис. 1.9 показано заполнение диалогового окна *Гистограмма*.

Полученная гистограмма показана на рис. 1.10 (флажок *Интегральный процент* при вводе данных снят).

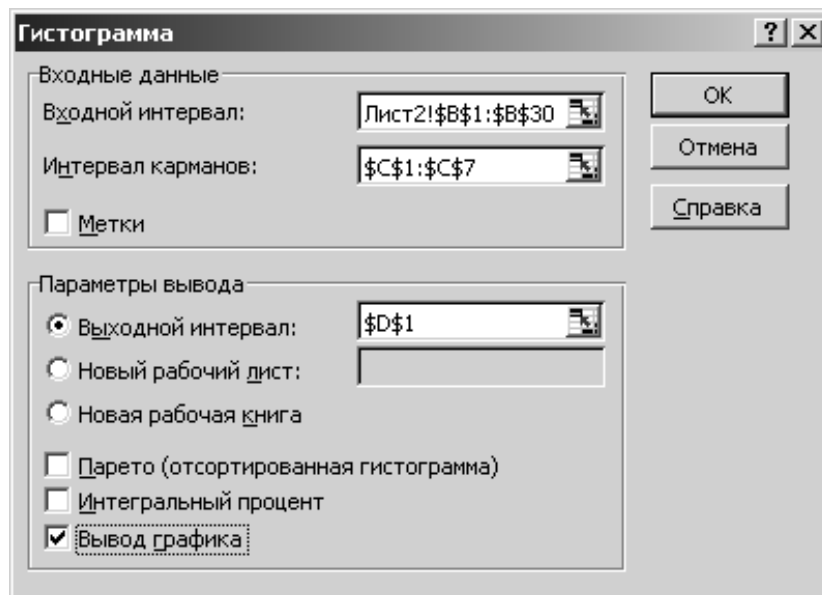


Рис. 1.9

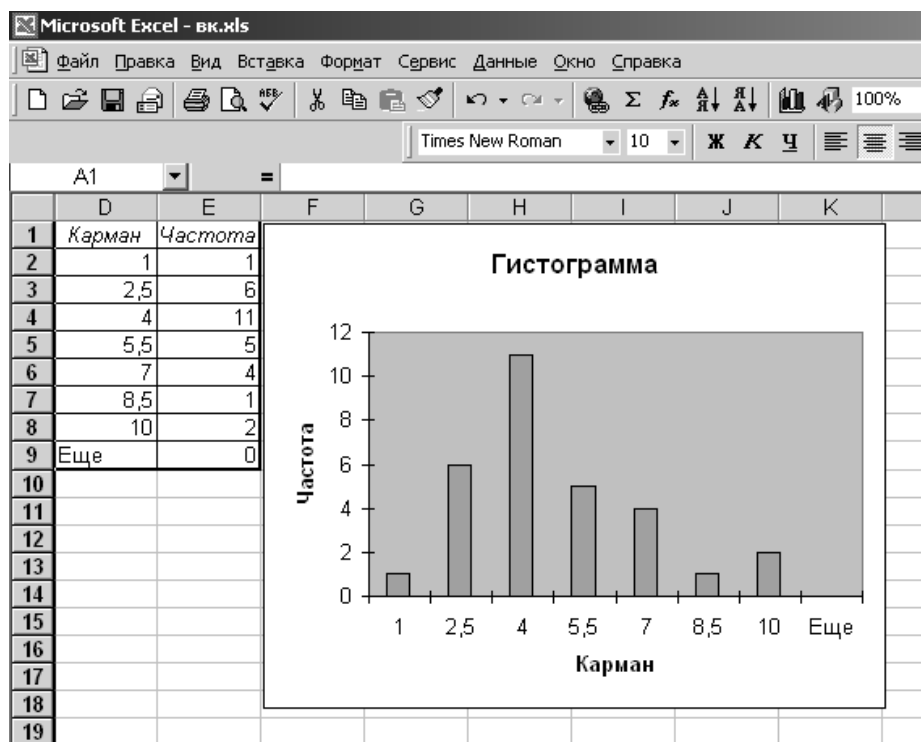


Рис. 1.10

## 1.5.

### Описательная статистика в Statistica

#### Подготовка исходных данных

Загрузите систему Statistica: на экране появляется окно с переключателем модулей (в английской версии – *Module switcher*). С его помощью выбирается необходимый для работы модуль (рис. 1.11). Выберите модуль *Основные статистики и таблицы*.

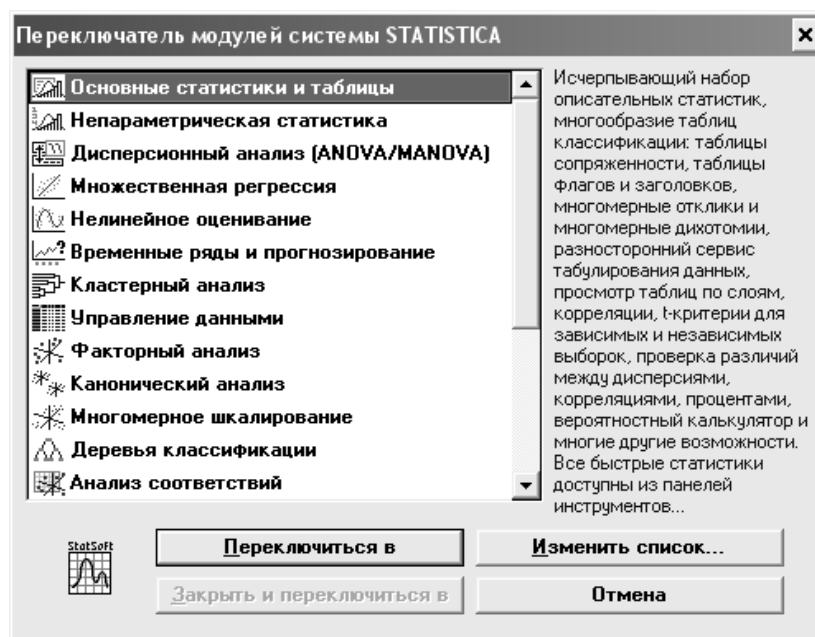


Рис. 1.11

На экране открываются два окна: окно с таблицей исходных данных и стартовая панель. В стартовой панели выбранного модуля (рис. 1.12) – перечень методов этого модуля. С помощью кнопки *Данные (Open Data)* можно ввести файл данных для обработки.

Загрузите данные любого примера из папки с примерами Examples. Просмотрите структуру данных. Данные представляют электронную таблицу, состоящую из столбцов – переменных (*Variables*) и строк – значений, которые эти переменные принимают – случаев (*Cases*).

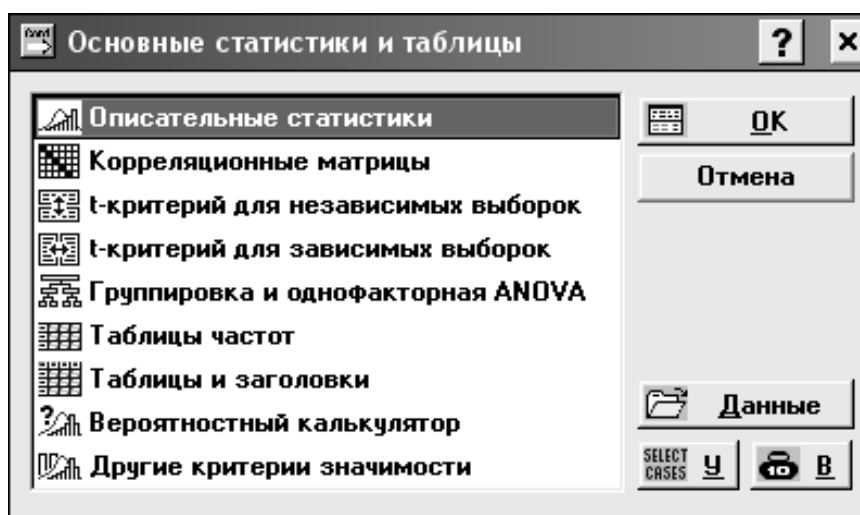


Рис. 1.12.

При активизации таблицы исходных данных стартовая панель сворачивается в кнопку. При необходимости ее можно открыть через меню *Анализ / Стартовая панель*.

Создайте новую таблицу исходных данных: *Файл / Создать (File / New Data)*; выберите нужную папку на диске и введите имя файла. Расширение *sta* будет присвоено файлу по умолчанию: это стандартное расширение файлов исходных данных в системе Statistica.

Новая таблица имеет 10 строк и 10 столбцов. В таблицу надо ввести данные о результатах исследования качества пряжи на двух прядильных машинах: в 15 выборках фиксировалось количество обрывов нити за определенное время. Для изменения размеров таблицы (необходимы два столбца по 15 строк) можно использовать контекстное меню (щелчок по таблице правой кнопкой). Выберите команду *Изменить столбцы (Modify Variables)*, *Удалить (Delete)*. Удалите столбцы с 3-его по 10-ый. По аналогии добавьте строки: *Изменить строки (Modify Cases)*, / *Добавить (Add)* и вставьте 5 строк после 10-ой.

Двойным щелчком по первому столбцу откройте окно для задания спецификации первой переменной. Введите имя переменной M1 (данные по первой машине), установите категорию данных (число), количество десятичных



знаков (ноль, так как данные – целые). По аналогии установите спецификации второй переменной. Введите данные в два *столбца*:

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
M1	12	5	14	10	7	10	4	8	5	12	8	14	3	5	9
M2	18	21	15	16	10	24	23	18	14	9	14	12	22	18	14

Иногда данные необходимо преобразовать с использованием формул или функций. Добавьте в таблицу с данными третий столбец и в окне спецификации в поле *Длинное имя (Long Name)* введите формулу: = LOG(M1+M2). В общем случае формула начинается со знака равенства, в ней могут использоваться знаки арифметических и логических операций, встроенные функции (вводятся соответствующей кнопкой), в качестве переменных – имена или номера столбцов. Сохраните полученную таблицу данных.

### Определение числовых характеристик

Для определения числовых характеристик переменных M1 и M2 выберите в стартовой панели команду *Описательные статистики (Descriptive statistics)*; с помощью кнопки *Переменные* выберите из списка переменных нужные для анализа (рис. 1.13), и нажмите кнопку *Подробные описательные статистики (Detailed descriptive statistics)*. В появившемся на экране окне выведены количество наблюдений, среднее значение, стандартное отклонение, минимальное и максимальное значения выборки.

Для возврата в диалоговое окно нажмите кнопку *Далее (Continue)*. С помощью кнопки *Другие статистики (More statistics)* можно получить и другие статистики, поставив соответствующие флажки: дисперсию (*Variance*), размах (*Range*), коэффициенты асимметрии (*Skewness*) и эксцесса (*Kurtosis*) и другие. Для вывода всех статистик используется кнопка *Все (All)*.

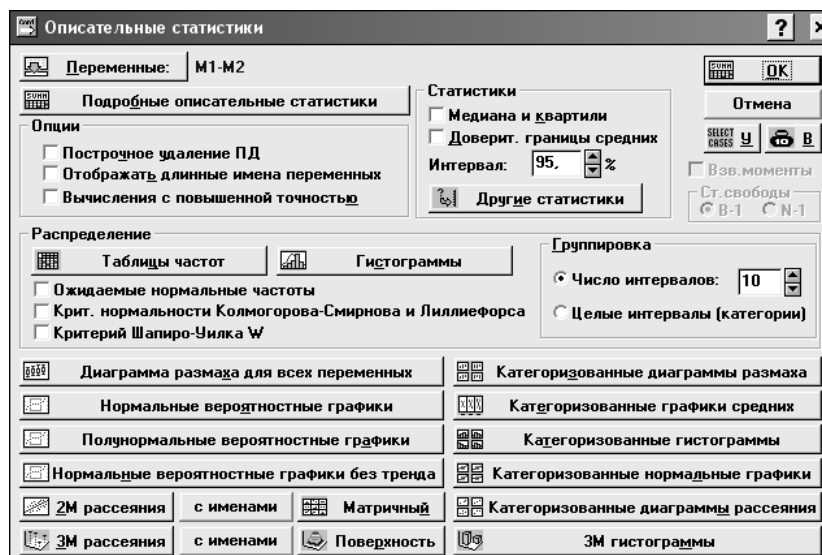


Рис. 1.13.

STATISTICA: Basic Statistics and Tables							
Descriptive Statistics (new1.sta)							
Variable	Valid N	Mean	Confid. -95.000%	Confid. +95.000%	Variance	Std.Dev.	Skew
M1	15	8.40000	6.41652	10.38348	12.82857	3.581700	.16
M2	15	16.53333	13.98812	19.07855	21.12381	4.596065	.08

Рис. 1.14

На рис. 1.14 для переменных M1 и M2 показаны объемы выборок, средние, границы 95%-го доверительного интервала, дисперсии, стандартные отклонения, коэффициенты асимметрии и эксцесса.

## Построение гистограммы

Для построения таблицы частот и гистограммы можно использовать соответствующие кнопки диалогового окна, показанного на рис. 1.13. Большие возможности предоставляет команда *Таблица частот* в стартовой панели.

В диалоговом окне *Таблицы частот* укажите переменные M1 и M2, для которых надо построить таблицы частот; в группе *Методы группировки для таблиц и графиков (Categorization methods for table & graph)* пометьте *Число равных интервалов (No of exact intervals)*, укажите 6 интервалов разбиения данных (рис. 1.15).

После нажатия кнопки *Таблица частот* будет выведено две таблицы для каждой из указанных переменных. В таблицах подсчитаны абсолютные частоты, накопленные значения и соответствующие проценты.

Для построения гистограмм нажмите соответствующую кнопку, и на экран будут выведены две гистограммы вместе с наложенными на них кривыми нормального распределения (рис. 1.16).

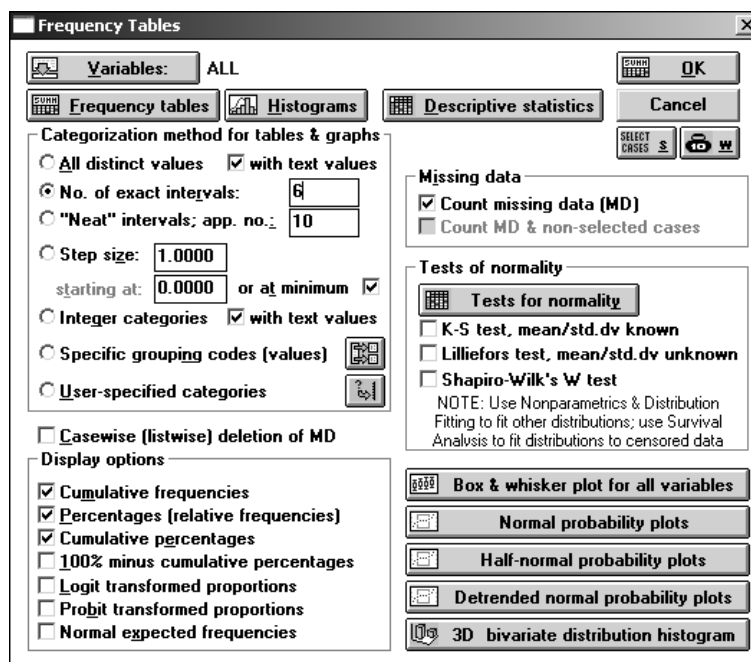


Рис. 1.15

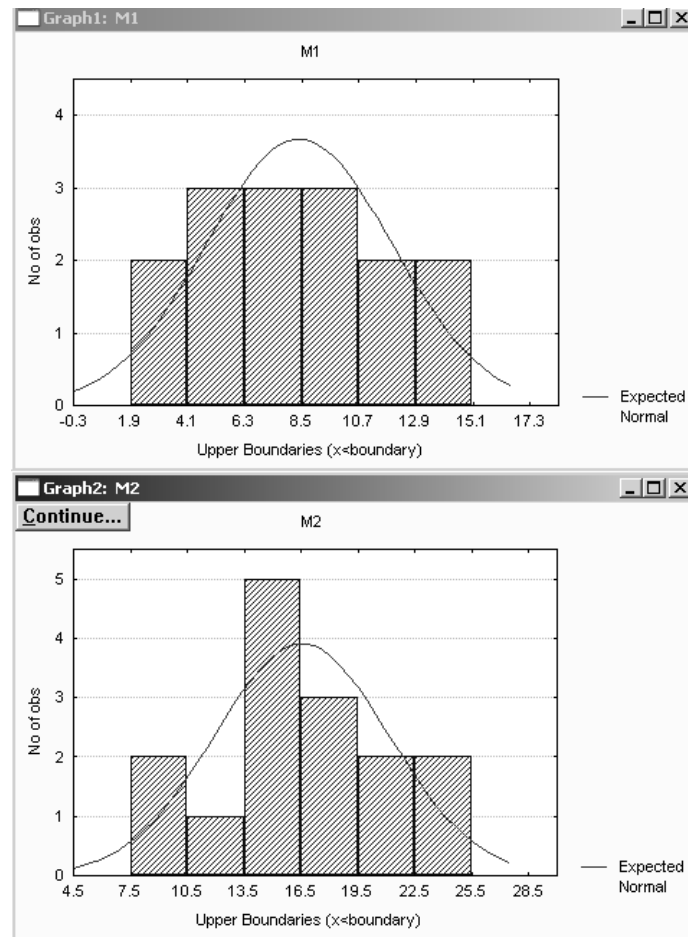


Рис. 1.16

## Контрольные вопросы

1. Что называется генеральной совокупностью?
2. Что называется выборкой? В чем состоит репрезентативность выборки?
3. Как строится вариационный ряд?
4. Какое распределение называется выборочным?
5. Как строится гистограмма? Полигон? График выборочной функции распределения?
6. Как вычисляется выборочное среднее? Выборочная дисперсия? Выборочное стандартное отклонение?
7. В чем состоят особенности вычислений числовых характеристик для группированного ряда?
8. Как определяется выборочная мода? Медиана?
9. Как вычисляется выборочный центральный момент?

10. Как вычисляется и что характеризует коэффициент асимметрии выборки? Коэффициент эксцесса?

# ОЦЕНКА ПАРАМЕТРОВ И ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

## 2.1.

### Точечные оценки параметров

Предположим, что вид распределения генеральной совокупности известен (нормальное, экспоненциальное и т.д.), тогда задача статистики сводится к оценке параметров этого распределения по результатам выборочных наблюдений. *Точечной оценкой*  $\tilde{\theta}$  неизвестного параметра распределения  $\theta$  называется приближенное значение этого параметра, полученное по данным выборки  $X_1, X_2, \dots, X_n$ .

Качество точечных оценок характеризуется следующими свойствами.

1. *Состоятельность*: оценка  $\tilde{\theta}$  называется состоятельной оценкой параметра  $\theta$ , если  $\tilde{\theta}$  сходится по вероятности к  $\theta$  при  $n \rightarrow \infty$ , то есть

$$\lim_{n \rightarrow \infty} P\left[|\tilde{\theta} - \theta| < \varepsilon\right] = 1 \quad (2.1)$$

при любом сколь угодно малом  $\varepsilon$ . Можно показать, что это условие соответствует двум условиям:

$$\lim_{n \rightarrow \infty} M(\tilde{\theta}) = \theta, \quad (2.2)$$

$$\lim_{n \rightarrow \infty} D(\tilde{\theta}) = 0.$$

2. *Несмещенность*: оценка  $\tilde{\Theta}$  называется несмещенной оценкой параметра  $\Theta$ , если ее математическое ожидание равно оцениваемому параметру, то есть

$$M(\tilde{\Theta}) = \Theta.$$

Разность  $M(\tilde{\Theta}) - \Theta$  называют смещением.

3. *Эффективность*: оценка  $\tilde{\Theta}$  называется эффективной оценкой параметра  $\Theta$ , если при заданном объеме выборки она имеет наименьшую возможную дисперсию.

Простейшим методом точечного оценивания является метод подстановки, когда в качестве оценки параметра используют соответствующую выборочную характеристику. Например, в качестве оценки  $\tilde{m}$  математического ожидания  $m$  генеральной совокупности принимается выборочная средняя  $\tilde{m} = \bar{x}$ .

Можно показать, что эта оценка является состоятельной и несмещенной, а если выборка взята из нормального распределения, то и эффективной.

Подобным образом в качестве оценки дисперсии  $D_x = \sigma^2$  генеральной совокупности можно принять выборочную дисперсию  $D_x^*$ . Эта оценка является состоятельной, но смещенной, так как  $M(D_x^*) \neq \sigma^2$ , и равно

$$M(D_x^*) = \frac{n-1}{n} \sigma^2 = \sigma^2 - \frac{1}{n} \sigma^2,$$

то есть смещена на  $(\sigma^2/n)$ . Можно исправить выборочную дисперсию так, чтобы ее математическое ожидание было равно дисперсии генеральной совокупности – умножить на дробь  $n/(n-1)$ . Полученная исправленная дисперсия является несмещенной оценкой дисперсии генеральной совокупности; будем называть ее *несмещенной дисперсией*:

$$S^2 = \frac{n}{n-1} D_x^* = \frac{n}{n-1} \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2. \quad (2.4)$$

Одним из наиболее распространенных методов оценивания параметров распределения является *метод максимального правдоподобия*. Для непрерывной случайной величины с известной плотностью  $f(x, \theta)$ , зависящей от некоторого неизвестного параметра  $\theta$ , вводится функция правдоподобия

$$L(\theta) = \prod_{i=1}^n f(x_i, \theta),$$

где  $x_i$  – фиксированные выборочные данные. В качестве оценки параметра  $\theta$  принимается такое значение, которое обеспечивает максимум функции правдоподобия. На практике, как правило, используется  $\ln L(\theta)$  – логарифмическая функция правдоподобия. Приравнявая нулю производную  $\left( \frac{d \ln L(\theta)}{d\theta} = 0 \right)$ , находят оценку максимального правдоподобия.

## 2.2. Интервальные оценки

Иногда в статистических расчетах важно не только найти оценку параметра распределения, но и охарактеризовать ее точность. Для этого вводится понятие о *доверительном интервале* для параметра  $\theta$  – это интервал  $(\theta_1, \theta_2)$ , содержащий (накрывающий) истинное значение  $\theta$  с заданной вероятностью  $p = 1 - \alpha$ , то есть

$$P\{\theta_1 < \theta < \theta_2\} = 1 - \alpha \quad (2.5)$$

Число  $p = 1 - \alpha$  называют *доверительной вероятностью* (или надежностью), а значение  $\alpha$  – *уровнем значимости*.

Для определения доверительного интервала необходимо знать закон распределения функции  $\tilde{\theta} = \tilde{\theta}(x_1, x_2, \dots, x_n)$ . Любая функция элементов выборки называется *статистикой*. Наиболее распространенными распределениями статистик являются нормальное, хи-квадрат, Стьюдента и Фишера.

Как известно из теории вероятностей, нормальным называется распределение случайной величины  $X$ , плотность которого

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (2.6)$$



где  $m$  – математическое ожидание,  $\sigma$  – среднее квадратичное отклонение; в общем случае используется обозначение  $N(m, \sigma)$ . При  $m = 0$ ,  $\sigma = 1$  имеем  $N(0,1)$  – стандартное нормальное распределение.

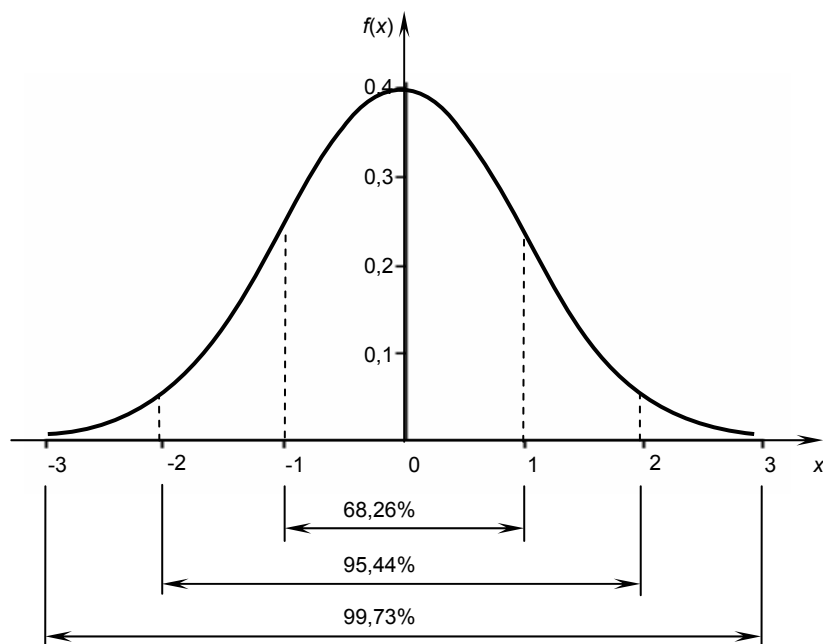


Рис. 2.1

Для этого случая функция нормального распределения

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \quad (2.7)$$

табулирована.

*Квантилью* нормального распределения, как известно, называется число  $u_p$ , для которого  $\Phi(u_p) = p$ . Квантили  $u_p$  табулированы и определяются в зависимости от вероятности  $p$ , причем на основании свойств нормального распределения

$$u_{1-p} = -u_p. \quad (2.8)$$

Пусть  $X_i$  ( $i = 1, \dots, k$ ) – независимые случайные величины, каждая из которых распределена по закону  $N(0,1)$ . Тогда сумма квадратов этих величин

$$\chi^2(k) = \sum_{i=1}^k X_i^2 \quad (2.9)$$

распределены по закону  $\chi^2$  с  $k$  степенями свободы. Распределение  $\chi^2$  определяется одним параметром  $k$ : его математическое ожидание  $m_{\chi} = k$ , а дисперсия  $D_{\chi} = 2k$ . Квантили распределения  $\chi_p^2(k)$  табулированы и определяются в зависимости от вероятности  $p$  и числа степеней свободы  $k$ .

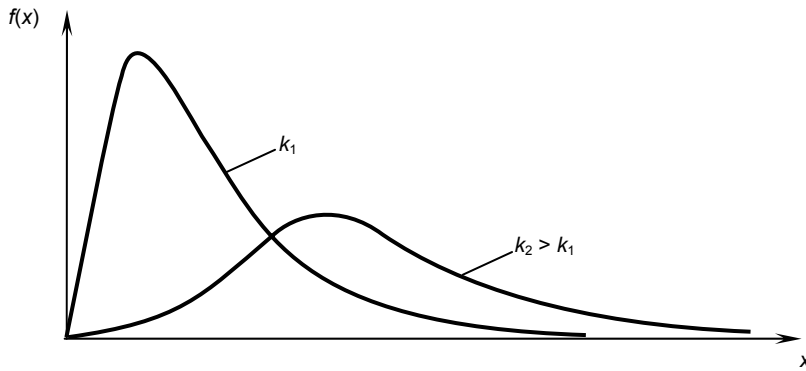


Рис. 2.2

Пусть случайная величина  $X$  распределена по закону  $N(0,1)$ , а независимая от нее случайная величина  $Y$  имеет распределение  $\chi^2$  с  $k$  степенями свободы. Тогда величина

$$t(k) = \frac{X}{\sqrt{Y/k}} \quad (2.10)$$

имеет распределение Стьюдента (или  $t$ -распределение) с  $k$  степенями свободы. Квантили  $t_p(k)$  распределения Стьюдента табулированы; вследствие симметрии распределения справедливо равенство

$$t_{p-1}(k) = -t_p(k). \quad (2.11)$$

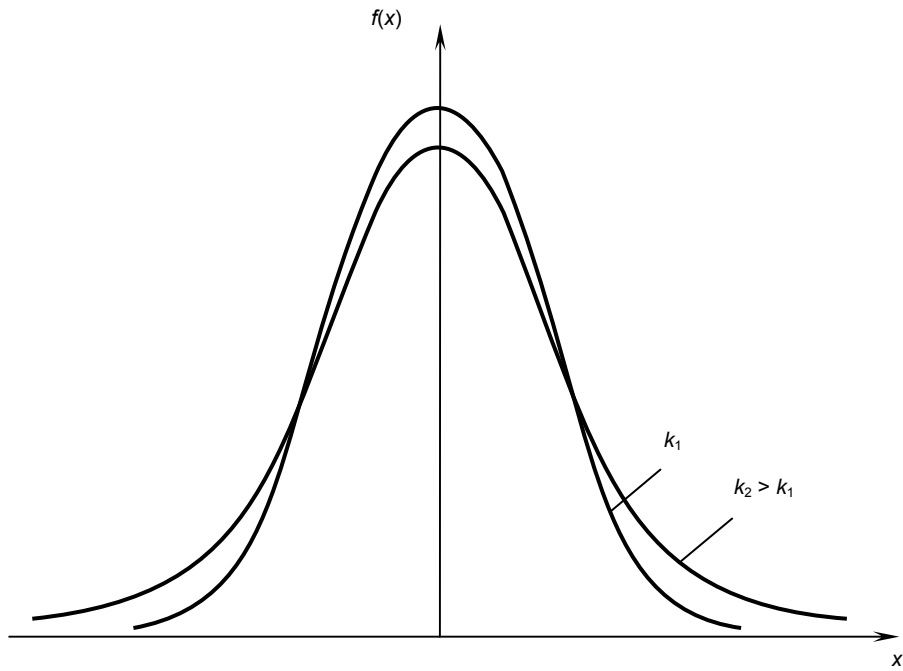


Рис. 2.3

Пусть  $X_1$  и  $X_2$  – независимые случайные величины, распределенные по закону  $\chi^2$  с  $k_1$  и  $k_2$  степенями свободы соответственно. Тогда величина

$$F(k_1, k_2) = \frac{X_1 / k_1}{X_2 / k_2} \quad (2.12)$$

имеет распределение Фишера (или  $F$ -распределение) с числом степеней свободы  $k_1$  и  $k_2$ .

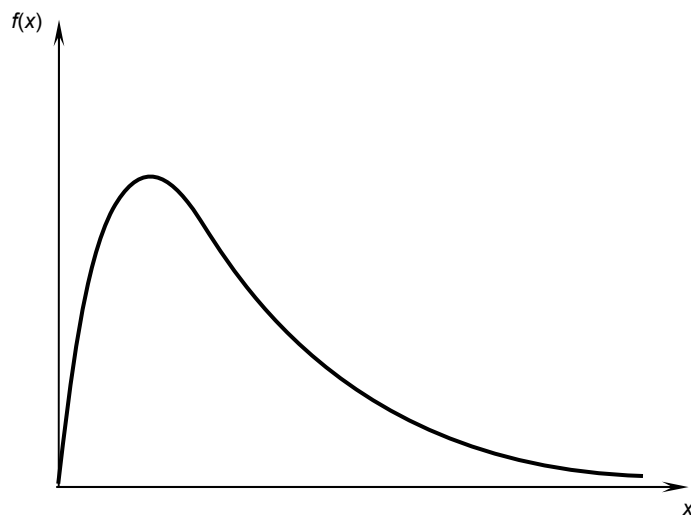


Рис. 2.4

Квантили  $F_p(k_1, k_2)$  распределения Фишера табулированы, причем справедливо равенство

$$F_{1-p}(k_1, k_2) = \frac{1}{F_p(k_1, k_2)} . \quad (2.13)$$

Рассмотренные распределения широко используются при решении задач статистики, в частности – при определении доверительных интервалов.

Пусть, например, случайная величина  $X$  распределена по нормальному закону с известной дисперсией  $\sigma^2$ . Тогда доверительный интервал для математического ожидания имеет вид

$$\bar{x} - \frac{\sigma}{\sqrt{n}} U_{1-\alpha/2} < m < \bar{x} + \frac{\sigma}{\sqrt{n}} U_{1-\alpha/2} , \quad (2.14)$$

где  $\bar{x}$  – выборочная средняя,  $n$  – объем выборки,  $U_{1-\alpha/2}$  – квантиль нормального распределения порядка  $(1 - \alpha/2)$ , а  $(1 - \alpha)$  – доверительная вероятность.

Если же дисперсия  $\sigma^2$  генеральной совокупности неизвестна, то в качестве оценки дисперсии используют несмещенную дисперсию; в этом случае

$$\bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) < m < \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) , \quad (2.15)$$

где  $t_{1-\alpha/2}(n-1)$  – квантиль распределения Стьюдента порядка  $(1-\alpha/2)$  с  $(n - 1)$  степенями свободы.

## 2.3.

### Проверка параметрических гипотез

*Статистическими* называются гипотезы о виде неизвестного распределения или о параметрах известного распределения. Проверяемая гипотеза называется *нулевой* и обозначается  $H_0$ .

*Конкурирующая* (или альтернативная) гипотеза  $H_1$  – это гипотеза, противоречащая нулевой. При проверке возможна ошибка, состоящая в том, что будет отвергнута правильная нулевая гипотеза – вероятность такой ошибки обозначается  $\alpha$  и называется *уровнем значимости*. Например,  $\alpha = 0.05$  означает, что в 5 случаях из 100 мы рискуем отвергнуть правильную гипотезу  $H_0$ .

Решение – принять или отвергнуть гипотезу  $H_0$  – принимается на основании некоторого правила или критерия по выборочным данным. При этом выбирается подходящая функция элементов выборки, или статистика критерия, которую в общем случае будем обозначать  $Z$ . Если распределение этой статистики известно (а это обычно  $N(0,1)$ , или  $\chi^2$ , или распределение Стьюдента или Фишера), то для обозначения будет использоваться та же буква, что и для обозначения соответствующей квантили.

Множество значений статистики  $Z$ , при которых принимается решение отклонить гипотезу  $H_0$ , называется *критической областью*. Графически эта область определяется по кривой распределения. Пусть, например, проверяется гипотеза о том, что параметр  $\Theta$  распределения генеральной совокупности равен некоторому значению  $\Theta_0$ , то есть  $H_0 : \Theta = \Theta_0$ . При этом возможны различные варианты альтернативных гипотез. Если  $H_0 : \Theta < \Theta_0$ , то критическая область расположена в левом «хвосте» соответствующего распределения, причем граница критической области определяется квантилью  $z_\alpha$  ( $\alpha$  – уровень значимости). Если  $H_0 : \Theta > \Theta_0$ , то критическая область – в правом «хвосте»; ее граница определяется квантилью  $z_{1-\alpha}$ . В этих двух случаях критическая область называется *односторонней*. Если же альтернативная гипотеза имеет вид  $H_0 : \Theta \neq \Theta_0$ , то имеем *двухстороннюю* критическую область, границы которой

определяются соответственно квантилями  $z_{\alpha/2}$  и  $z_{1-\alpha/2}$ . Множество значений статистики  $Z$ , при которых гипотеза  $H_0$  принимается, называется *областью принятия решения*.

Общая последовательность проверки гипотезы о параметрах распределения такова:

- формулируются гипотезы  $H_0$  и  $H_1$ ;
- задается уровень значимости  $\alpha$ ;
- выбирается статистика  $Z$  для проверки  $H_0$ ;
- определяется выборочное распределение статистики  $Z$ ;
- в зависимости от  $H_1$  определяется критическая область;
- вычисляется выборочное значение статистики  $z$ ;
- принимается статистическое решение: если выборочное значение статистики  $z$  оказывается в области принятия решения, гипотеза  $H_0$  принимается; в противном случае гипотеза  $H_0$  отклоняется, как несогласующаяся с результатами наблюдений.

Рассмотрим некоторые наиболее важные для практики случаи. Предположим, что проверяется гипотеза о средней нормально распределенной генеральной совокупности при известной дисперсии  $\sigma^2$ , то есть  $H_0 : m = m_0$ . Нетрудно показать, что статистикой критерия может служить величина

$$u = \frac{\bar{x} - m_0}{\sigma / \sqrt{n}}, \quad (2.16)$$

распределенная по закону  $N(0,1)$ . Если же дисперсия неизвестна, то используется статистика

$$t = \frac{\bar{x} - m_0}{s / \sqrt{n}}, \quad (2.17)$$

имеющая распределение Стьюдента с  $(n - 1)$  степенью свободы.

Часто на практике возникает задача о сравнении средних двух нормально распределенных совокупностей, то есть о проверке гипотезы  $H_0 : m_1 = m_2$ . Если

соответствующие дисперсии  $\sigma_1^2$  и  $\sigma_2^2$  известны, то в качестве статистики принимается величина

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad (2.18)$$

распределенная по закону  $N(0,1)$ . Здесь  $\bar{x}_1$  и  $\bar{x}_2$  – соответствующие выборочные средние,  $n_1$  и  $n_2$  – объемы выборок.

Аналогичным образом решаются вопросы о проверке гипотез, связанных с дисперсиями. Если проверяется гипотеза о равенстве дисперсий двух нормально распределенных совокупностей, то есть  $H_0 : \sigma_1^2 = \sigma_2^2$  при неизвестных математических ожиданиях  $m_1$  и  $m_2$ , то используется статистика

$$F = \frac{S_1^2}{S_2^2}, \quad (2.19)$$

имеющая распределение Фишера с числом степеней свободы  $(n_1 - 1)$  и  $(n_2 - 1)$ ; здесь  $S_1^2$  и  $S_2^2$  – соответствующие несмещенные дисперсии; предполагается, что  $S_1^2 > S_2^2$ .

Данные о статистиках критериев и их распределениях для различных гипотез приводятся в справочной литературе.

## 2.4. Критерии согласия

Другой группой статистических гипотез являются гипотезы о проверке вида распределения: неизвестен вид распределения генеральной совокупности, и в частности, неизвестна функция распределения  $F(x)$ .

Пусть  $x_1, x_2, \dots, x_n$  – выборка наблюдений случайной величины  $X$ . Проверяется гипотеза  $H_0$  о том, что случайная величина  $X$  имеет функцию распределения  $F(x)$ .

Разобьем область возможных значений  $X$  на  $r$  интервалов  $\Delta_1, \Delta_2, \dots, \Delta_r$ . Пусть  $n_i$  – число элементов выборки, принадлежащих интервалу  $\Delta_i$  ( $i = 1, \dots, r$ ). Используя предполагаемый закон распределения – с функцией  $F(x)$ , с учетом оценок параметров этого закона, найденных по выборке, находят вероятности того, что значения  $X$  принадлежат интервалу  $\Delta_i$ , то есть

$$p_i = P\{X \in \Delta_i\}, i = \overline{1, r}.$$

Очевидно, что  $\sum_{i=1}^r p_i = 1$ .

Результаты представляют в виде таблицы:

Интервалы	Число наблюдений фактическое	Число наблюдений расчетное
$\Delta_1$	$n_1$	$np_1$
$\Delta_2$	$n_2$	$np_2$
...	...	...
$\Delta_r$	$n_r$	$np_r$

Можно показать, что статистика

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - np_i)^2}{np_i} \quad (2.20)$$

имеет распределение  $\chi^2$  с числом степеней свободы  $(r - l - 1)$ , где  $r$  – число интервалов,  $l$  – число неизвестных параметров распределения. Например, для нормального распределения  $l = 2$  (неизвестные параметры  $m$  и  $\sigma$ ). Считается, что гипотеза  $H_0$  согласуется с опытом, если

$$\chi^2 < \chi_{1-\alpha}^2(r-l-1),$$

где  $\chi^2$  – выборочное значение статистики,  $\chi_{1-\alpha}^2(r-l-1)$  – квантиль порядка  $(1 - \alpha)$  распределения  $\chi^2$  с числом степеней свободы  $(r - l - 1)$ .

Рассмотренный метод проверки гипотезы вида распределения называется *критерием хи-квадрат* или *критерием согласия* Пирсона.



## 2.5.

### Примеры расчета

Пример 2.1. Найти 95%-ный доверительный интервал для математического ожидания твердости сплава (в условных единицах), если по результатам измерений получены следующие значения: 14,2; 14,8; 14,0; 14,7; 13,9; 14,8; 15,1; 15,0; 14,5.

Объем выборки  $n = 9$ . Выборочное среднее

$$\bar{x} = (14,2 + 14,8 + \dots + 14,5) / 9 = 14,56;$$

выборочная дисперсия

$$D_X^* = (14,2^2 + 14,8^2 + \dots + 14,5^2) / 9 - 14,56^2 = 0,17;$$

несмещенная дисперсия

$$s^2 = 9 \cdot 0,17 / 8 = 0,19; \quad s = 0,43;$$

доверительная вероятность  $p = 0,95$ ;

уровень значимости  $\alpha = 0,05$ ;  $1 - \alpha/2 = 0,975$ ;

квантиль распределения Стьюдента  $t_{0,975}(8) = 2,306$  (по таблице).

Тогда, используя формулу (2.15), получим:

$$14,56 - 0,33 < m < 14,56 + 0,33.$$

С вероятностью 0,95 математическое ожидание твердости сплава лежит в пределах от 14,23 до 14,89.

Пример 2.2. Проверить гипотезу о том, что средний диаметр валиков, изготавливаемых на станке-автомате, равен  $m_0 = 12$  мм, если по выборке из  $n = 16$  валиков найдены среднее значение  $\bar{x} = 11,7$  мм и несмещенная дисперсия  $s^2 = 0,25$  мм<sup>2</sup>. Распределение диаметра валика предполагается нормальным.

Проверяется нулевая гипотеза  $H_0: m = m_0$  при альтернативной гипотезе  $H_1: m < m_0$  (поскольку среднее значение оказалось меньше, чем  $m_0$ ). Принимаем уровень значимости  $\alpha = 0,05$ . Выборочное значение статистики Стьюдента  $t_b = (11,7 - 12) \cdot 4 / 0,5 = -2,4$ .

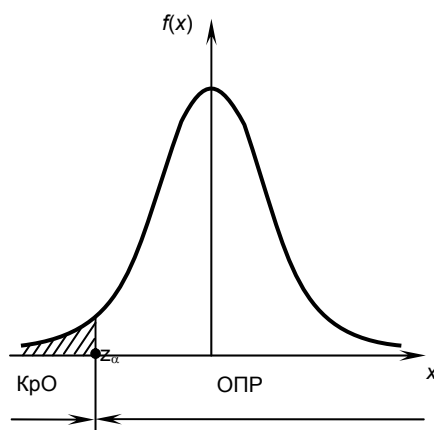


Рис. 2.5 Для левосторонней критической области положение границы

$$z_{кр} = z_{\alpha} = t_{0,05}(15) = -t_{0,95}(15) = -1,753.$$

Выборочное значение статистики  $-2,4$  попало в критическую область (рис. 2.5), нулевая гипотеза о том, что средний диаметр валиков равен 12 мм, отвергается.

Пример 2.3. Используя двусторонний критерий, проверить гипотезу о равенстве внутренних диаметров втулок, изготавливаемых на двух станках по одному чертежу. Из деталей, изготовленных на первом станке, отобрано  $n_1 = 12$  втулок; при этом средний диаметр  $\bar{x}_1 = 8,5$  мм, на втором станке –  $n_2 = 14$ ,  $\bar{x}_2 = 8,3$  мм. Распределение диаметров предполагается нормальным, дисперсии известны и равны соответственно  $\sigma_1^2 = 0,2 \text{ мм}^2$ ,  $\sigma_2^2 = 0,25 \text{ мм}^2$ .

Нулевая гипотеза  $H_0: m_1 = m_2$  при альтернативе  $H_1: m_1 \neq m_2$  (двусторонний критерий). Принимаем уровень значимости  $\alpha = 0,05$ .

Выборочное значение статистики по формуле (2.18)

$$u_B = (8,5 - 8,3) / (0,2/12 + 0,25/14)^{1/2} = 1,08.$$

Для двусторонней критической области положение границ

$$z_{кр1} = z_{\alpha/2} = u_{0,025} = -u_{0,975} = -1,96; \quad z_{кр2} = z_{1-\alpha/2} = u_{0,975} = 1,96.$$

Выборочное значение статистики 1,08 попало в область принятия решения (рис. 2.6); нулевая гипотеза о том, что диаметры втулок одинаковы, принимается.

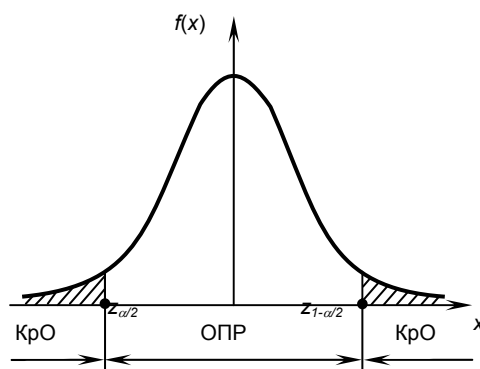


Рис. 2.6

## 2.6. Проверка гипотез в Excel

### Гипотеза о равенстве дисперсий

Исследуются результаты обработки деталей на двух станках. Предполагается, что точность обработки одинакова, то есть, что дисперсии равны. Для проверки этой гипотезы проведены замеры 22 деталей на первом станке и 24 деталей на втором. Результаты представлены в первых трех столбцах на рис. 2.8.

Для проверки гипотезы о равенстве дисперсий выберем *Сервис / Анализ данных / Двухвыборочный F-тест*. Введем в качестве значений переменной 1 результаты измерений на первом станке, переменной 2 – на втором; уровень значимости 0,05 (рис. 2.7).

**Двухвыборочный F-тест для дисперсии** [?] [X]

Входные данные

Интервал переменной 1:  [...]

Интервал переменной 2:  [...]

☐ Метки

Альфа:

Параметры вывода

☒ Выходной интервал:  [...]

☐ Новый рабочий лист:

☐ Новая рабочая книга

OK Отмена Справка

Рис. 2.7

Результаты замеров			Двухвыборочный F-тест для дисперсии		
№	Станок 1	Станок 2		Станок 1	Станок 2
1	12,05	12,36			
2	12,08	12,45	Среднее	12,249	12,449
3	12,33	12,48	Дисперсия	0,04476	0,01712
4	12,34	12,56	Наблюдения	22	24
5	12,75	12,63	df	21	23
6	12,32	12,25	F	2,6136	
7	12,12	12,54	P(F<=f) одностороннее	0,0136	
8	12,05	12,35	F критическое одностороннее	2,0356	
9	12,08	12,54			
10	12,33	12,33			
11	12,08	12,85			
12	12,75	12,42			
13	12,05	12,47			
14	12,08	12,41			
15	12,33	12,34			
16	12,05	12,51			
17	12,08	12,45			
18	12,31	12,24			
19	12,34	12,55			
20	12,42	12,32			
21	12,42	12,44			
22	12,12	12,41			
23		12,38			
24		12,51			

Рис. 2.8.

В полученной таблице с результатами, показанной на рис. 2.8 справа, приводятся средние значения, дисперсии, количество наблюдений и степени свободы для каждой выборки, значение статистики Фишера (определяется как отношение дисперсий) и критическое значение (квантиль распределения Фишера) при заданном уровне значимости.

Гипотеза о равенстве дисперсий принимается, если выборочное значение статистики Фишера попало в область принятия решения, в противном случае гипотеза отклоняется.

В условиях рассматриваемой задачи выборочное значение статистики Фишера 2,61 больше критического значения 2,04, то есть попало в критическую область. Гипотеза о равенстве дисперсий отклоняется.

### Гипотеза о равенстве средних

Проверка этой гипотезы проводится по-разному в зависимости от того, принята или отклонена гипотеза о значимости дисперсий: используются двухвыборочные  $t$ -тесты с одинаковыми или неодинаковыми дисперсиями.

Проверьте гипотезу о равенстве средних для рассмотренного примера (*Сервис / Анализ данных / Двухвыборочный  $t$ -тест с неодинаковыми дисперсиями*).

Введите данные по аналогии с двухвыборочным  $F$ -тестом (рис. 2.9).

Рис. 2.9

В таблице с результатами расчета приводятся статистика Стьюдента и критические значения для одностороннего и двухстороннего критериев (рис. 2.10).

Гипотеза о равенстве средних принимается, если выборочное значение статистики Стьюдента попало в область принятия решения, в противном случае гипотеза отклоняется.

Двухвыборочный t-тест с различными дисперсиями		
	<i>Переменная 1</i>	<i>Переменная 2</i>
Среднее	12,24909091	12,44958333
Дисперсия	0,044761039	0,017125906
Наблюдения	22	24
Гипотетическая разность средних	0	
df	34	
t-статистика	-3,824511797	
P(T<=t) одностороннее	0,000266932	
t критическое одностороннее	1,690923455	
P(T<=t) двухстороннее	0,000533864	
t критическое двухстороннее	2,032243174	

Рис. 2.10

В условиях рассматриваемого примера как для одностороннего, так и двухстороннего критериев выборочное значение статистики Стьюдента – 3,82 – оказалось больше (по модулю), чем критическое, то есть попало в критическую область: гипотеза о равенстве средних отклоняется.

### **Гипотеза о виде распределения**

Смоделируйте нормально распределенную совокупность (рис. 2.11) из 1000 элементов с средним значением 12 и стандартным отклонением 0,25 (рис. 2.12). Сформируйте случайную выборку из 200 элементов для этой совокупности (рис. 2.13). Используя критерий хи-квадрат, проверим, действительно ли выборка сделана из нормально распределенной генеральной совокупности.

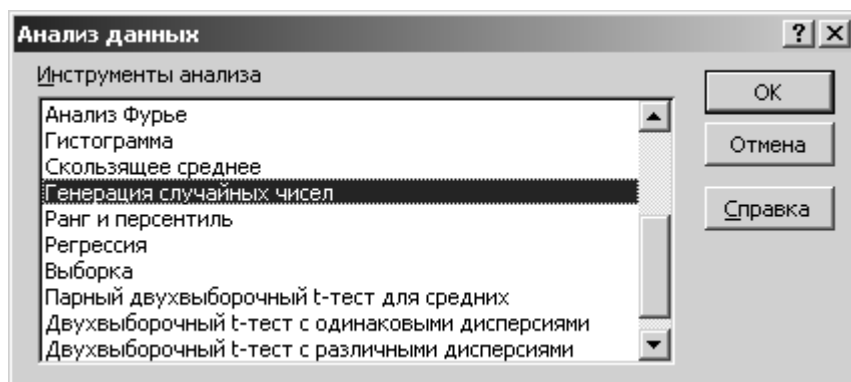


Рис. 2.11

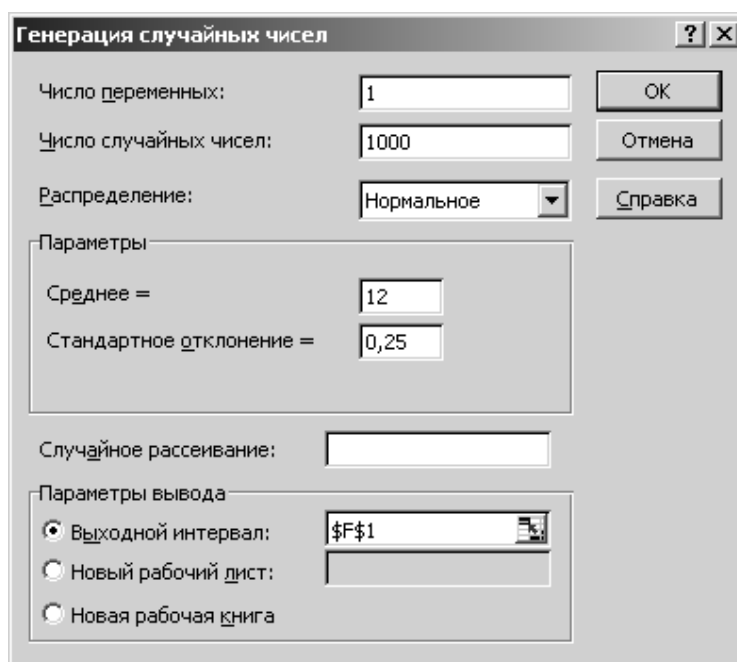


Рис. 2.12

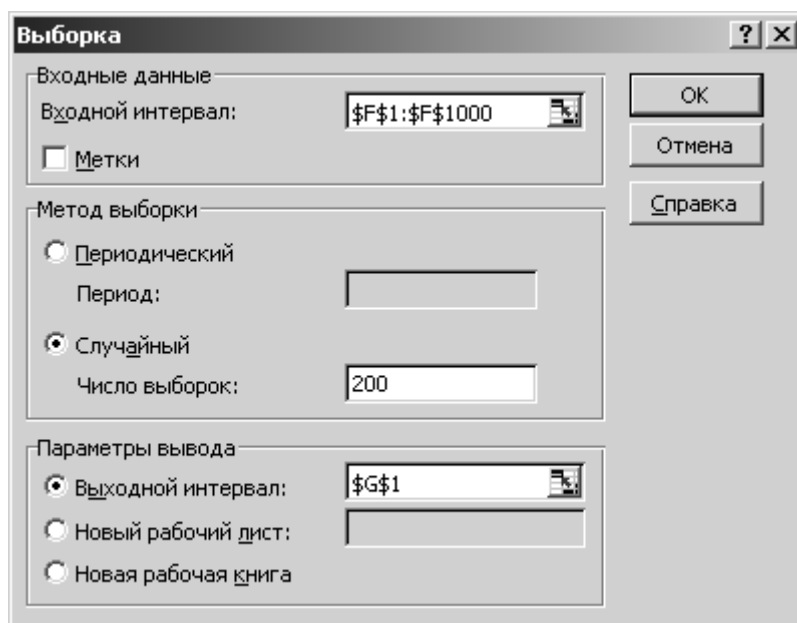


Рис. 2.13

В качестве точечных оценок математического ожидания и дисперсии примите соответствующие выборочные характеристики. Найдите их, используя инструмент *Описательная статистика* пакета *Анализ данных*.

С помощью инструмента *Гистограмма* найдите опытные частоты  $n_i$ . При использовании критерия хи-квадрат количество опытных значений в каждом интервале должно быть не менее пяти. Если в каком-то интервале их меньше, то интервалы объединяют. Например, если в промежутке от 4 до 6 оказалось три значения, а в промежутке от 6 до 8 – четыре, то вводится новый интервал от 4 до 8 с семью значениями. С учетом этого перестройте таблицу частот вручную. На рис. 2.14 в колонках *Карман – Частота* показаны данные, полученные автоматически, в колонках *Границы – Опытные частоты* данные пересчитаны частично вручную.

F	G	H	I	J	K	L	M
№	Карман	Частота	Границы	Опытные частоты	НОРМ РАСП	Вероятности	Расчетные частоты
1	11,350	1	11,350				
2	11,446	0	11,639	17	0,0972	0,0972	19,45
3	11,543	3	11,736	18	0,1788	0,0815	16,31
4	11,637	13	11,832	33	0,2938	0,1149	22,99
5	11,736	18	11,929	28	0,4345	0,1407	28,14
6	11,832	33	12,025	18	0,5842	0,1496	29,93
7	11,929	28	12,122	29	0,7224	0,1382	27,64
8	12,025	18	12,218	26	0,8333	0,1109	22,18
9	12,122	29	12,315	11	0,9107	0,0773	15,46
10	12,218	26	12,412	10	0,9575	0,0468	9,362
11	12,315	11	12,508	5	0,9821	0,0246	4,923
12	12,412	10	12,701	5	0,9978	0,0157	3,141
13	12,508	5					
14	12,605	2					
15	Еще	3					

ХИ2ТЕСТ	0,238
ХИ2ОБР	15,5

Рис. 2.14.



Расчетные частоты  $np_i$  вычисляются через вероятности попадания нормально распределенной величины в соответствующий интервал:

$$p_i = \Phi\left(\frac{x_{i+1} - m}{\sigma}\right) - \Phi\left(\frac{x_i - m}{\sigma}\right),$$

где функция стандартного нормального распределения  $\Phi(\cdot)$  вычисляется с помощью встроенной статистической функции НОРМРАСП ( $x$ , среднее значение  $m$ , стандартное отклонение  $\sigma$ , интегральный). Аргументы этой функции (рис. 2.15):  $x$  - граница интервала, вводится адрес соответствующей ячейки;  $m$  и  $\sigma$  - вводятся абсолютные адреса характеристик, полученных с помощью Описательной статистики; значение интегральный = 1 (истина), в противном случае (ложь) вычисляется не функция распределения, а его плотность. На рис. 2.14 вычисленные значения этой функции рассчитаны в колонке НОРМРАСП. Вероятности  $p_i$  (колонка Вероятности) вычисляются как разности между значениями НОРМРАСП в последующей и предыдущей строках. В последней колонке подсчитаны расчетные частоты  $np_i$  ( $n = 200$ ).

Для вычисления статистики хи-квадрат в Excel встроена функция ХИ2ТЕСТ (фактический интервал, ожидаемый интервал).

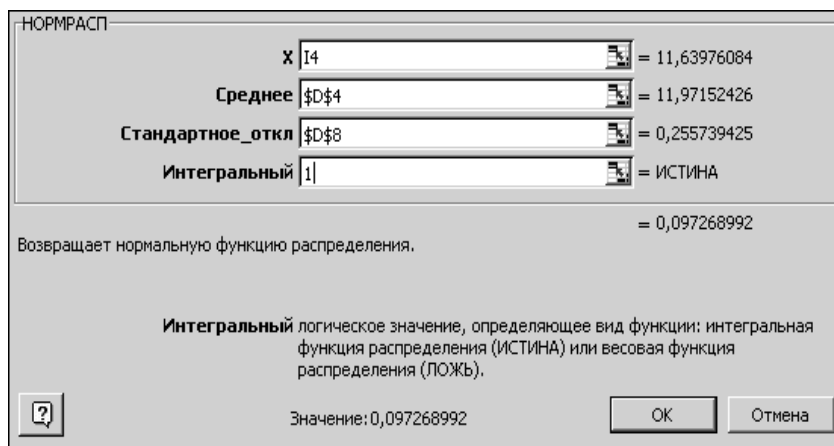


Рис. 2.15.

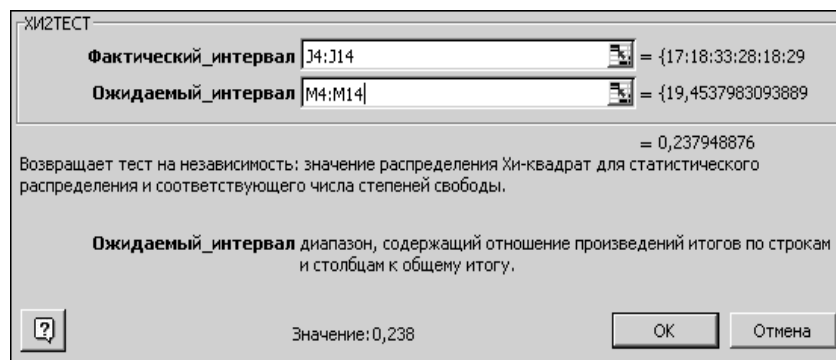


Рис. 2.16.

В качестве фактического интервала вводятся опытные частоты, в качестве ожидаемого – расчетные (рис. 2.16).

Граница критической области – квантиль распределения хи-квадрат, может быть найдена с помощью встроенной функции ХИ2ОБР (вероятность, степени свободы). Аргумент вероятность – это уровень значимости ( $\alpha = 0,05$ ), а степени свободы  $k - l - 1$  определяются как количество интервалов (на рис. 2.14  $k = 11$ ) за вычетом количества оцениваемых параметров (здесь два –  $m$  и  $\sigma$ ) минус единица.

Гипотеза о нормальности распределения принимается, если выборочное значение статистики ХИ2ТЕСТ окажется меньше критического ХИ2ОБР.

Подобным образом может быть проверена гипотеза о виде любого распределения.

## 2.7.

### Оценка параметров и проверка гипотез в Statistica

Доверительный интервал для математического ожидания строится одновременно с расчетом числовых характеристик. Доверительная вероятность по умолчанию 0,95; при необходимости можно установить нужный уровень: на рис. 2.17 показаны 99% границы доверительных интервалов.

Descriptive Statistics (new1.sta)				
Continue...	Mean	Confid. -99.000%	Confid. +99.000%	Std.Dev.
M1	8.40000	5.64704	11.15296	3.581700
M2	16.53333	13.00072	20.06595	4.596065

Рис. 2.17

В модуль *Основные статистики и таблицы* встроены *t*-критерии для проверки равенства средних в зависимых и независимых выборках (см. рис. 1.12). *F* - критерий для проверки равенства дисперсий в этих выборках выводится автоматически.

Для проверки нормальности распределения используются несколько критериев согласия: критерии Колмогорова – Смирнова (K-S), Лиллиефорса, Шапиро-Уилка. Значения критериев и соответствующие доверительные вероятности приводятся одновременно с гистограммами (рис. 2.18).

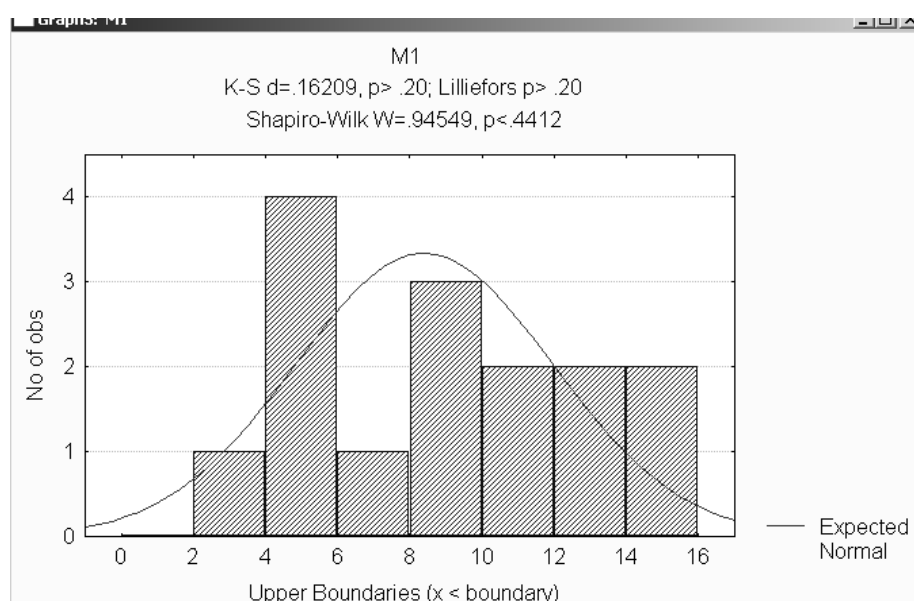


Рис. 2.18

Кроме того, нормальность распределения приближенно можно оценить графически по нормальным вероятностным графикам: чем ближе опытные точки к прямой линии, тем ближе распределение к нормальному (рис. 2.19).

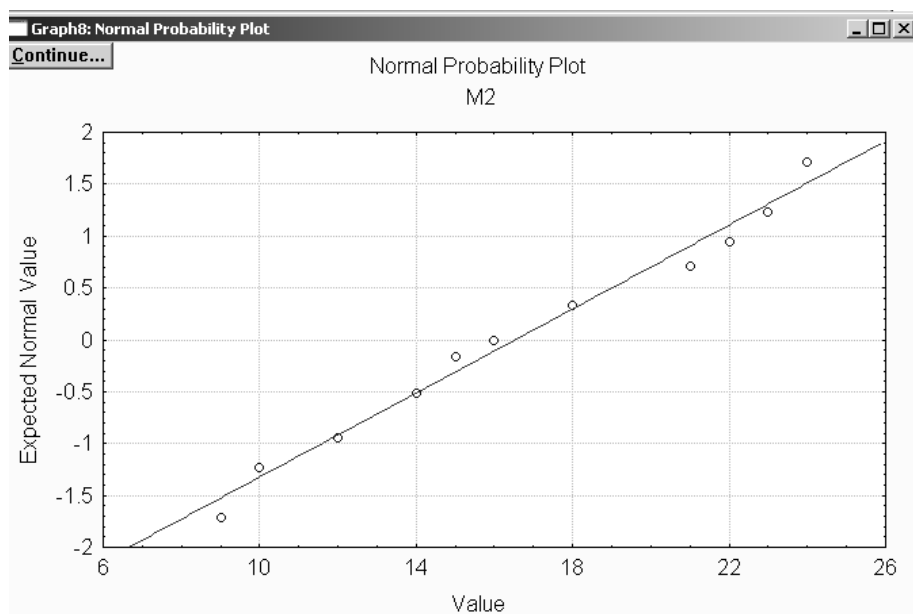


Рис. 2.19

## Контрольные вопросы

1. Какие оценки параметров называются точечными? Перечислите основные свойства точечных оценок.
2. Каковы точечные оценки математического ожидания и дисперсии?
3. В чем состоит метод максимального правдоподобия?
4. Доказать несмещенность и состоятельность выборочной средней как оценки математического ожидания.
5. Как определяется несмещенная дисперсия?
6. Перечислите основные распределения, используемые в статистических расчетах. Как определяются квантили этих распределений? От чего они зависят?
7. Используя таблицы, найдите квантили  
 $U_{0.1}; \chi^2_{0.95}(20); t_{0.05}(10); F_{0.1}(10,20)$
8. Как строится доверительный интервал для математического ожидания? Дисперсии?
9. Какая гипотеза называется нулевой? Альтернативной? В чем состоят ошибки первого и второго рода?

10. В какой последовательности проводится проверка параметрической гипотезы?
11. Почему граница критической двухсторонней области определяется квантилями  $Z_{\alpha/2}$  и  $Z_{1-\alpha/2}$ ?
12. Как проверяется гипотеза о равенстве двух дисперсий, если математические ожидания известны? Неизвестны?
13. Какие критерии используются для проверки гипотез о виде распределения?
14. В чем состоит критерий согласия хи-квадрат?

## ДИСПЕРСИОННЫЙ АНАЛИЗ

### 3.1.

#### Однофакторный дисперсионный анализ

Во многих практических ситуациях представляет интерес влияние того или иного фактора на рассматриваемый признак.

Пусть, например, оценка качества поверхности детали проводится с помощью  $l$  приборов и необходимо исследовать влияние фактора «прибор» на результат измерений. Если приборов два, то проверка нулевой гипотезы о равенстве их средних показаний проводится обычными методами проверки статистических гипотез. Если же  $l > 2$ , то используются методы дисперсионного анализа.

Проверяется нулевая гипотеза  $H_0: m_1 = m_2 = \dots = m_l$  об отсутствии влияния на результативный признак  $X$  фактора  $A$ , имеющего  $l$  уровней  $A_k$ ,  $k = 1, \dots, l$ . Основная идея дисперсионного анализа состоит в том, чтобы сопоставить дисперсию за счет воздействия фактора  $A$  с дисперсией, обусловленной случайными причинами. Если различие между ними не существенно, то влияние фактора  $A$  на признак  $X$  незначительно. Если же различие между факторной и остаточной дисперсиями значимо, то это говорит о влиянии фактора  $A$  на рассматриваемый признак  $X$ .

Предполагается, что случайная величина  $X$  имеет нормальное распределение с математическим ожиданием  $m_k$ , зависящим от уровня фактора  $A_k$ , и постоянной дисперсией  $\sigma^2$ . В качестве исходных данных используются выборочные значения величины  $X$ , полученные для каждого уровня фактора  $A$ ; число элементов выборки на каждом уровне равно  $n$ , тогда общее число наблюдений  $nl$ ,  $x_{ik}$  - результат

номер опыта	Уровни фактора А				
	A1	A2	... A <sub>k</sub> ...	A <sub>l</sub>	
1	X <sub>11</sub>	X <sub>12</sub>	... X <sub>1k</sub> ...	X <sub>1l</sub>	
2	X <sub>21</sub>	X <sub>22</sub>	... X <sub>2k</sub> ...	X <sub>2l</sub>	
...	...	...	-----	...	
i	X <sub>i1</sub>	X <sub>i2</sub>	... X <sub>ik</sub> ...	X <sub>il</sub>	
...	...	...	-----	...	
n	X <sub>n1</sub>	X <sub>n2</sub>	... X <sub>nk</sub> ...	X <sub>nl</sub>	

$i$ -го наблюдения ( $i=1, \dots, n$ ) за  $k$ -тым уровнем фактора  $A$  ( $k=1, \dots, l$ ).

Выборочная средняя, соответствующая  $k$ -му уровню фактора  $A$ , (групповая средняя) вычисляется по формуле:

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik}; \quad (3.1)$$

общая выборочная средняя есть

$$\bar{x} = \frac{1}{nl} \sum_{k=1}^l \sum_{i=1}^n x_{ik} = \frac{1}{l} \sum_{k=1}^l \bar{x}_k. \quad (3.2)$$

Для вычисления дисперсии найдем суммы квадратов.

Общая сумма квадратов – это сумма квадратов отклонений наблюдаемых значений  $x_{ik}$  от общей выборочной средней:

$$Q = \sum_{k=1}^l \sum_{i=1}^n (x_{ik} - \bar{x})^2 = \sum_{k=1}^l \sum_{i=1}^n x_{ik}^2 - nl\bar{x}^2. \quad (3.3)$$

Факторная сумма квадратов (обусловленная влиянием фактора  $A$ ) - это сумма квадратов отклонений групповых средних от общей средней:

$$Q_A = n \sum_{k=1}^l (\bar{x}_k - \bar{x})^2 = n \sum_{k=1}^l \bar{x}_k^2 - nl\bar{x}^2. \quad (3.4)$$

Остаточная сумма квадратов характеризует рассеяние внутри группы:

$$Q_e = \sum_{k=1}^l \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2. \quad (3.5)$$

На практике эта сумма определяется из основного тождества дисперсионного анализа, в соответствии с которым

$$Q = Q_A + Q_e. \quad (3.6)$$

Разделив суммы квадратов на соответствующее число степеней свободы, найдем соответствующие дисперсии (иногда их называют средними суммами квадратов):

$$\begin{aligned} S^2 &= \frac{Q}{nl-1}, \\ S_A^2 &= \frac{Q_A}{l-1}, \\ S_e^2 &= \frac{Q}{l(n-1)}. \end{aligned} \quad (3.7)$$

Если нулевая гипотеза о равенстве средних справедлива, то эти дисперсии являются несмещенными оценками дисперсий генеральной совокупности. Значительное превышение дисперсии  $S_A^2$  над дисперсией  $S_e^2$  можно объяснить различием средних в группах. Поэтому для проверки нулевой гипотезы используется отношение этих средних, которое имеет распределение Фишера

$$F = \frac{S_A}{S_e} = \frac{\frac{1}{l-1} Q_A}{\frac{1}{l(n-1)} Q_e} \quad (3.8)$$

с числом степеней свободы  $(l-1)$  и  $l(n-1)$ . Гипотеза  $H_0: m_1 = m_2 = \dots = m_l$  не противоречит результатам наблюдений при заданном уровне значимости  $\alpha$ , если

$$F > F_{1-\alpha}(l-1, l(n-1)),$$

в этом случае считается, что фактор  $A$  не оказывает существенного влияния на признак  $X$ .

Результаты расчета обычно сводятся в таблицу.



Источник дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия	Выборочное значение статистики Фишера
Фактор $A$	$Q_A$	$l - 1$	$S_A^2$	$F$
Остаток	$Q_e$	$l(n - 1)$	$S_e^2$	
Общая	$Q$	$ln - 1$	$S^2$	

### 3.2.

## Многофакторный дисперсионный анализ

В двухфакторном дисперсионном анализе проверяется влияние на результативный признак  $X$  двух факторов  $A$  и  $B$  и их взаимодействия. Фактор  $A$  имеет  $l$  уровней  $A_j$ ,  $j = 1, \dots, l$ ; фактор  $B$  –  $r$  уровней  $B_k$ ,  $k = 1, \dots, r$ . При каждом сочетании уровней  $A_j B_k$  делается  $n$  наблюдений. Общее число наблюдений  $nlr$ .

Проверяются три нулевые гипотезы: об отсутствии влияния на результативный признак  $X$  фактора  $A$ , фактора  $B$  и их взаимодействия  $AB$ .

Пусть  $X_{ijk}$  – результат  $i$ -го наблюдения ( $i = 1, \dots, n$ ) при  $j$ -ом уровне фактора  $A$  и  $k$ -ом уровне фактора  $B$ . Тогда средняя, соответствующая сочетанию уровней  $A$  и  $B$ :

$$\bar{x}_{jk} = \frac{1}{n} \sum_{i=1}^n x_{ijk};$$

средняя, соответствующая уровню  $A_j$ :

$$\bar{x}_{0j} = \frac{1}{nr} \sum_{i=1}^n \sum_{k=1}^r x_{ijk} = \frac{1}{r} \sum_{k=1}^r \bar{x}_{jk};$$

средняя, соответствующая уровню  $B_k$ :

$$x_{k0} = \frac{1}{nl} \sum_{i=1}^n \sum_{j=1}^l x_{ijk} = \frac{1}{l} \sum_{j=1}^l \bar{x}_{jk},$$

общая средняя

$$\bar{x} = \frac{1}{nlr} \sum_{i=1}^n \sum_{j=1}^l \sum_{k=1}^r x_{ijk} = \frac{1}{l} \sum_j \bar{x}_{0j} = \frac{1}{r} \sum_{k=1}^r x_{k0}.$$

По аналогии с однофакторным анализом справедливо тождество

$$Q = Q_A + Q_B + Q_{AB} + Q_e,$$

где общая сумма квадратов:

$$Q = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x})^2,$$

сумма квадратов, учитывающая влияние фактора  $A$ :

$$Q_A = nr \sum_j (\bar{x}_{0j} - \bar{x})^2;$$

сумма квадратов, учитывающая влияние фактора  $B$ :

$$Q_B = nl \sum_k (\bar{x}_{0k} - \bar{x})^2;$$

сумма квадратов, учитывающая взаимодействие факторов  $A$  и  $B$ :

$$Q_{AB} = n \sum_j \sum_k (\bar{x}_{jk} - x_{0j} - x_{k0} + \bar{x})^2;$$

остаточная сумма квадратов:

$$Q_e = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_{jk})^2;$$

соответствующие дисперсии:

$$S^2 = \frac{Q}{nlr - 1},$$

$$S_A^2 = \frac{Q_A}{l - 1},$$

$$S_B^2 = \frac{Q_B}{r - 1},$$

$$S_{AB}^2 = \frac{Q_{AB}}{(l - 1)(r - 1)},$$

$$S_e^2 = \frac{Q_e}{lr(n - 1)}.$$

Проверка нулевых гипотез осуществляется с использованием статистик Фишера:

$$F_A = \frac{S_A^2}{S_e^2},$$

$$F_B = \frac{S_B^2}{S_e^2},$$

$$F_{AB} = \frac{S_{AB}^2}{S_e^2},$$

которые сравниваются с соответствующими квантилями. Например, гипотеза  $H_0$  об отсутствии влияния взаимодействия факторов  $A$  и  $B$  на результативный признак  $X$  принимается, если

$$F_{AB} < F_{1-\alpha}[(l-1)(r-1), lr(n-1)].$$

Результаты оформляются в виде таблицы.

Источник дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия	Выборочное значение статистики Фишера
Фактор $A$	$Q_A$	$l - 1$	$S_A^2$	$F_A$
Фактор $B$	$Q_B$	$r - 1$	$S_B^2$	$F_B$
Взаимодействие $AB$	$Q_{AB}$	$(l - 1)(r - 1)$	$S_{AB}^2$	$F_{AB}$
Остаток	$Q_e$	$rl(n - 1)$	$S_e^2$	
Общая	$Q$	$lrn - 1$	$S^2$	

Алгоритм трехфакторного дисперсионного анализа аналогичен двухфакторному. Оценивается влияние факторов  $A$ ,  $B$ ,  $C$ , их попарного взаимодействия  $AB$ ,  $BC$ ,  $AC$  и общего взаимодействия  $ABC$  на результативный признак  $X$ . Фактор  $A$  имеет  $l$  уровней, фактор  $B$  –  $r$  уровней, фактор  $C$  –  $q$

уровней. При каждом сочетании уровней проводятся по  $n$  измерений, то есть общее число измерений  $nlrq$ .

Таблица трехфакторного анализа имеет вид:

Источник дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия	Выборочное значение статистики Фишера
Фактор $A$	$Q_A$	$l - 1$	$S_A^2$	$F_A$
Фактор $B$	$Q_B$	$r - 1$	$S_B^2$	$F_B$
Фактор $C$	$Q_C$	$q - 1$	$S_C^2$	$F_C$
Взаимодействие $AB$	$Q_{AB}$	$(l - 1)(r - 1)$	$S_{AB}^2$	$F_{AB}$
Взаимодействие $BC$	$Q_{BC}$	$(r - 1)(q - 1)$	$S_{BC}^2$	$F_{BC}$
Взаимодействие $AC$	$Q_{AC}$	$(l - 1)(q - 1)$	$S_{AC}^2$	$F_{AC}$
Взаимодействие $ABC$	$Q_{ABC}$	$(l - 1)(r - 1) * (q - 1)$	$S_{ABC}^2$	$F_{ABC}$
Остаток	$Q_e$	$lrq(n - 1)$	$S_e^2$	
Общая	$Q$	$lrqn - 1$	$S^2$	

Для проверки нулевой гипотезы, например, об отсутствии влияния общего взаимодействия  $ABC$  значение статистики Фишера

$$F_{ABC} = \frac{S_{ABC}^2}{S_e^2} = \frac{\frac{1}{(l - 1)(r - 1)(q - 1)} Q_{ABC}}{\frac{1}{lrq(n - 1)} Q_e}$$

сравнивается с квантилью

$$F_{1-\alpha}[(l - 1)(r - 1)(q - 1), lrq(n - 1)]$$

### 3.3.

#### Примеры расчета

Пример 3.1. Оценить влияние технологии чистовой обработки (три вида технологий) на точность изготовления детали. Проводятся по 4 замера (при каждом виде технологии) отклонения размера детали от номинала в мкм. Принять  $\alpha = 0,05$ .

Номер замера	Вид технологии		
	1	2	3
1	1	2	3
2	2	1	2
3	2	3	2
4	1	2	3

1. Используем алгоритм однофакторного дисперсионного анализа; имеем  $n = 4$ ,  $l = 3$ .

Групповые средние

$$\bar{x}_1 = \frac{1}{n} \sum x_{i1} = \frac{1}{4}(1+2+2+1) = \frac{3}{2};$$

$$\bar{x}_2 = \frac{1}{4}(2+1+3+2) = 2;$$

$$\bar{x}_3 = \frac{1}{4}(3+2+2+3) = \frac{5}{2}.$$

2. Общая средняя

$$\bar{x} = \frac{1}{l} \sum \bar{x}_k = \frac{1}{3}(\frac{3}{2} + 2 + \frac{5}{2}) = 2.$$

3. Общая сумма квадратов

$$Q = \sum \sum x_{ik}^2 - nl\bar{x}^2 = 1^2 + 2^2 + 2^2 + 1^2 + 2^2 + 1^2 + 3^2 + 2^2 + 3^2 + 2^2 + 2^2 + 3^2 - 4 \cdot 3 \cdot 2^2 = 6.$$

4. Факторная сумма квадратов

$$Q_A = n \sum \bar{x}_k^2 - nl\bar{x}^2 = 4(\frac{9}{4} + 4 + \frac{25}{4}) - 4 \cdot 3 \cdot 2^2 = 2.$$

5. Остаточная сумма квадратов

$$Q_{\epsilon} = Q - Q_A = 6 - 4 = 2$$

6. Заполним таблицу однофакторного дисперсионного анализа:

Источник дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия	Выборочное значение статистики Фишера
Фактор $A$	2	2	1	9/4
Остаток	4	9	4/9	
Общая	6	11	6/11	

7. Находим по таблице квантиль распределения Фишера

$$F_{1-\alpha}(l-1, l(n-1)) = F_{0,95}(2,9) = 4,26$$

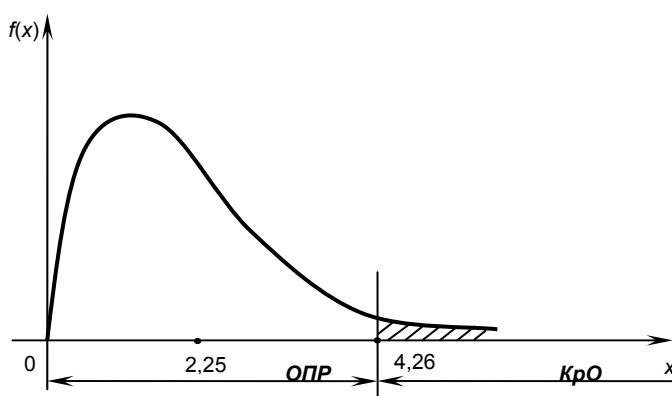


Рис. 3.1

Так как выборочное значение статистики Фишера  $F = 9/4 = 2,25$  оказалось меньше критического, 4,26 (см. рис. 3.1), то нулевая гипотеза принимается, то есть в данном случае влияние технологии изготовления на точность детали несущественно.

Пример 3.2. Требуется оценить влияние давления (фактор  $A$ , 4 уровня), температуры при прессовании (фактор  $B$ , 4 уровня) и времени выдержки в пресс-форме (фактор  $C$ , 3 уровня) на предел прочности болтов из стекловолокнита, если при каждом сочетании уровней испытывалось по 5 образцов. В результате предварительной обработки опытных данных

найденны значения сумм квадратов:  $Q_A = 22400$ ,  $Q_B = 3200$ ,  $Q_C = 2700$ ,  $Q_{AB} = 3800$ ,  $Q_{AC} = 4600$ ,  $Q_{BC} = 1900$ ,  $Q_{ABC} = 10300$ ,  $Q = 108500$ . Принять уровень значимости  $\alpha = 0,05$ .

Учитывая, что из условия задачи  $l = r = 4$ ,  $q = 3$ ,  $n = 5$ , заполняем таблицу трехфакторного анализа. При этом

$$Q_e = Q - Q_A - Q_B - Q_C - Q_{AB} - Q_{AC} - Q_{BC} - Q_{ABC} = 59600$$

Источник дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия	Выборочное значение статистики Фишера	Критическое значение
Фактор $A$	22400	3	7467	24,09	2,60
Фактор $B$	3200	3	1067	3,44	2,60
Фактор $C$	2700	2	1350	4,35	3,00
$AB$	3800	9	422	1,36	1,88
$BC$	1900	6	317	1,02	2,10
$AC$	4600	6	767	2,47	2,10
$ABC$	10300	18	572	1,85	1,61
Остаток	59600	192	310		
Общая	108500	239	454		

Для удобства сравнения в таблицу добавлена колонка с критическими значениями статистики Фишера, определяемыми по таблицам.

Сравнивая последние две колонки таблицы видим, что все три рассматриваемых фактора, а также взаимодействие между температурой и временем выдержки ( $AC$ ) и общее взаимодействие – оказывают влияние на предел прочности болтов.

### 3.4. Дисперсионный анализ в Excel

Требуется оценить влияние квалификации наладчиков (фактор *A*) на рассеяние диаметров шариков. Замеры отклонения диаметра от номинала для каждого из пяти наладчиков проводились по 6 раз:

№	A1	A2	A3	A4	A5
1	1,2	0,6	0,9	1,7	1
2	1,1	1,1	0,6	1,4	1,4
3	1	0,8	0,8	1,3	1,1
4	1,3	0,7	1	1,6	0,9
5	1,1	0,7	1	1,2	1,2
6	0,8	0,9	1,1	1,3	1,5

Проверяется нулевая гипотеза о равенстве математических ожиданий отклонения для всех пяти наладчиков, то есть предполагается, что квалификация наладчика не влияет на точность изготовления шариков.

Для проведения анализа воспользуйтесь инструментом *Однофакторный дисперсионный анализ* пакета *Анализ данных*. В качестве исходных данных введите таблицу замеров.

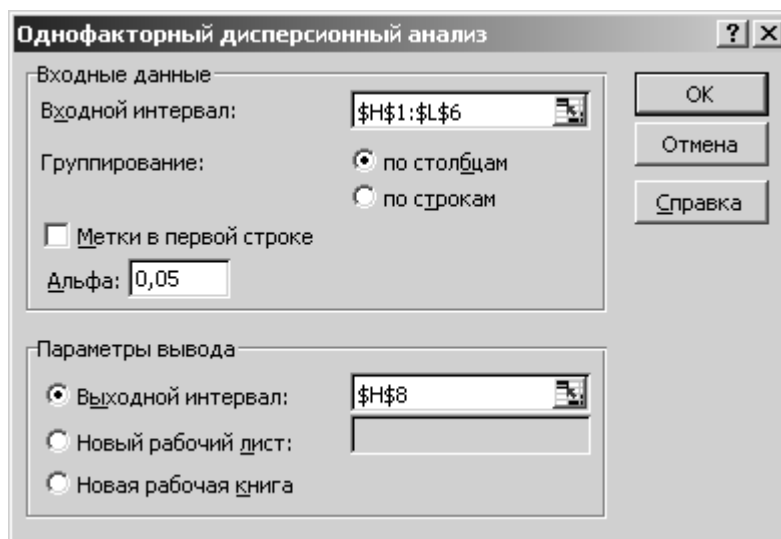


Рис. 3.2

Выводятся две таблицы. В первой таблице приводятся статистические характеристики для каждого наладчика, во второй (ANOVA) – результаты



анализа, в частности, значение статистики Фишера ( $F$ ) и граница критической области ( $F$  критическое).

Однофакторный дисперсионный анализ						
ИТОГИ						
<i>Группы</i>	<i>Счет</i>	<i>Сумма</i>	<i>Среднее</i>	<i>Дисперсия</i>		
Столбец 1	6	6,5	1,0833	0,0297		
Столбец 2	6	4,8	0,8	0,032		
Столбец 3	6	5,4	0,9	0,032		
Столбец 4	6	8,5	1,4167	0,0377		
Столбец 5	6	7,1	1,1833	0,0537		
Дисперсионный анализ						
<i>Источник вариации</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p-знач.</i>	<i>F<sub>кр</sub></i>
Между группами	1,4087	4	0,3522	9,518	8E-05	2,758
Внутри групп	0,925	25	0,037			
Итого	2,3337	29				

Рис. 3.3

Если выборочное значение статистики оказалось меньше критического, нулевая гипотеза принимается. В данном примере выборочное значение статистики – 9,52 – оказалось больше критического 2,76, то есть значение статистики Фишера попало в критическую область: нулевая гипотеза о незначимости квалификации наладчиков отвергается.

В пакете анализа имеются и инструменты для проведения двухфакторного дисперсионного анализа (с повторениями и без повторений).

### 3.5.

## Дисперсионный анализ в Statistica

Предположим, что изучается, как влияет тип магазина на товарооборот. В магазинах трех типов фиксируется товарооборот за каждый из 8 месяцев работы (в млн руб).

Маг. 1	Маг. 2	Маг. 3
19	20	16
23	20	15
26	32	18
18	37	26
20	40	19
20	24	17
18	22	19
35	18	18

Для проведения дисперсионного анализа создайте файл из 24 строк и 2 столбцов, введите в первый столбец номер магазина (группирующая переменная), а во второй – данные о товарообороте ( $3 \times 8 = 24$  значения).

Откройте модуль с основными статистиками (*Basic Statistic / Tables*), загрузите метод *Группировка и однофакторный дисперсионный анализ (Breakdown & one-way ANOVA)*, выберите в поле *Анализ – Подробный анализ выбранных таблиц (Detailed analysis of individual tables)*, введите переменные – группирующую (*Grouping*) – столбец 1, и зависимую (*Depended*) – столбец 2. После двух щелчков *ОК* появится таблица результатов (*Descriptive statistics and correlations by groups – Results*) (рис. 3.4)

Щелкните по кнопке *Дисперсионный анализ (Analysis of variance)*. В появившемся окне рассчитаны основные статистики, включая *F*-статистику Фишера и *p*-значение, которое значительно превышает обычно принимаемое значение 0,05; таким образом, принимается гипотеза о равенстве товарооборотов в разных магазинах: фактор «магазин» не оказывает влияния на результативный признак «товарооборот».

Для проведения многофакторного дисперсионного анализа и более подробного однофакторного может быть использован специальный модуль ANOVA/MANOVA (см. рис. 1.11)

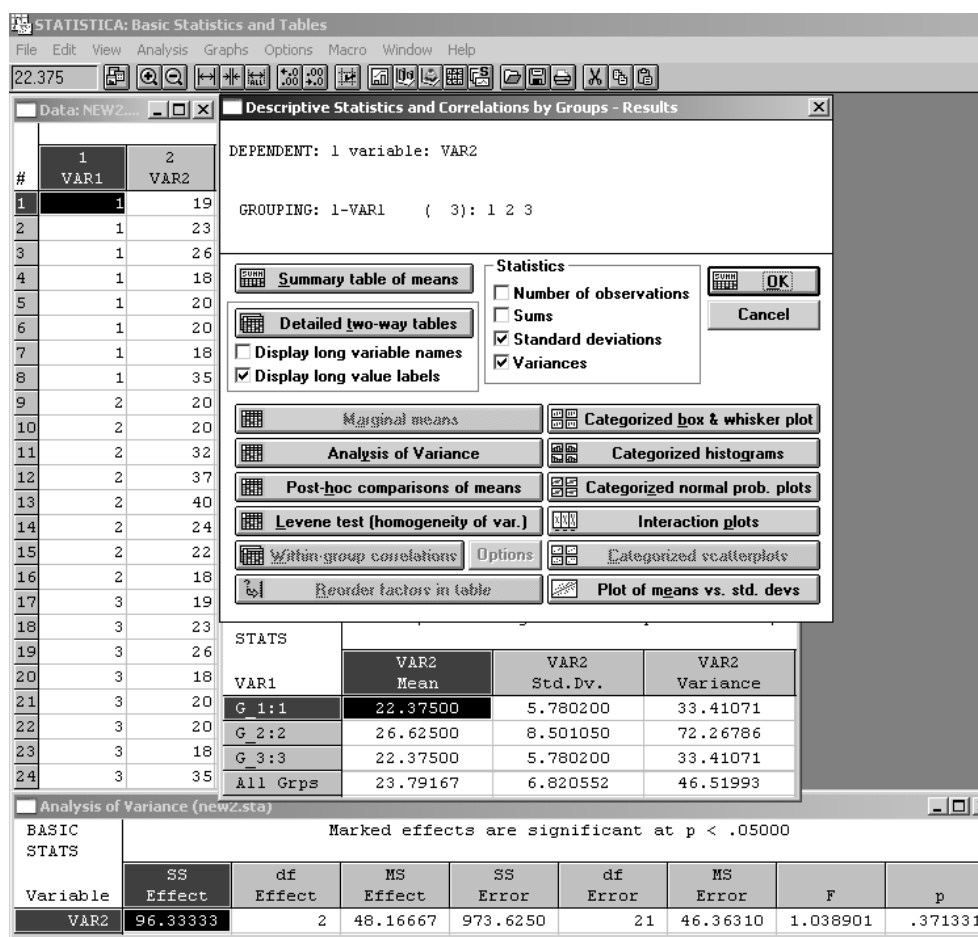


Рис. 3.4

## Контрольные вопросы

1. Доказать основное тождество однофакторного дисперсионного анализа.
2. Почему для проверки нулевых гипотез в дисперсионном анализе используется отношение дисперсий?
3. С помощью графика функции распределения Фишера пояснить, в каких случаях принимается, а в каких отвергается нулевая гипотеза.
4. Какие предположения о случайной величине  $X$  используются в дисперсионном анализе?
5. Какие гипотезы проверяются в двухфакторном дисперсионном анализе?
6. Как вычислить остаточную сумму квадратов в трехфакторном дисперсионном анализе?

7. Как вычисляется статистика Фишера при проверке гипотезы о влиянии фактора  $A$ ? Взаимодействия факторов  $AB$ ? Общего взаимодействия трех факторов  $ABC$ ?
8. От чего зависит критическое значение статистики Фишера?

## КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

### 4.1.

#### Коэффициент корреляции

Любая случайная величина  $X$  есть функция элементарного события  $\omega$ , входящего в пространство элементарных событий  $\Omega$ . Если каждому элементарному событию  $\omega$  ставится в соответствие  $k$  случайных величин  $X_i$  ( $i = 1, \dots, k$ ), то говорят о  $k$ -мерной случайной величине. Например, состояние любого технического объекта характеризуется набором нескольких случайных величин; если в результате эксперимента определяются координаты точки плоскости – имеем двумерную случайную величину (или двумерный вектор); если в процессе изготовления детали измеряется три размера – трехмерный случайный вектор и т. д.

Значение одной величины может не зависеть от того, какие значения приняли другие величины – в этом случае они называются независимыми. Если значение одной величины однозначно определяет значение другой, то такие величины связаны функциональной зависимостью. *Корреляционный анализ* устанавливает степень тесноты взаимосвязи между случайными величинами. Эта связь может быть более или менее тесной. Парная корреляция изучает взаимосвязи между двумя случайными величинами, множественная – между большим числом величин.

По аналогии с одномерной случайной величиной введем для двумерного вектора понятие центрального момента. Центральным моментом порядка  $(k + s)$  двумерного дискретного случайного вектора  $(X, Y)$  называется число

$$\mu_{k,s} = M[(X - m_x)^k \cdot (Y - m_y)^s] = \sum \sum (x_i - m_x)^k \cdot (y_j - m_y)^s \cdot P_{ij}, \quad (4.1)$$

где  $m_x$  и  $m_y$  – математические ожидания,  $P_{ij} = P\{X = x_i, Y = y_j\}$ .

Центральный момент порядка  $(1 + 1)$  называется *ковариацией*:

$$K_{XY} = \mu_{11} = M[(X - m_x)(Y - m_y)] = m_{xy} - m_x m_y, \quad (4.2)$$

а отношение ковариации к произведению среднеквадратичных отклонений

$$\rho = \frac{K_{XY}}{\sigma_x \sigma_y} \quad (4.3)$$

– *коэффициентом корреляции*. Коэффициент корреляции, по модулю не превышающий единицы  $|\rho| \leq 1$ , определяет степень линейной зависимости между случайными величинами  $X$  и  $Y$ . При  $\rho > 0$  корреляция называется положительной (в этом случае с увеличением  $X$  растет и  $Y$ ), при  $\rho < 0$  – отрицательной. Если  $\rho = 0$ , случайные величины  $X$  и  $Y$  называются *некоррелированными*; это не означает, что эти величины не связаны между собой, но линейной связи между ними нет. Если же  $|\rho| = 1$ , значит, величины  $X$  и  $Y$  связаны функциональной зависимостью типа  $Y = aX + b$ .

На практике считается, что при  $|\rho| < 0,2$  линейная связь между  $X$  и  $Y$  практически отсутствует; при  $|\rho| = 0,2 - 0,5$  – связь слабая; при  $|\rho| = 0,5 - 0,75$  – средняя; при  $|\rho| = 0,75 - 0,95$  – сильная. При  $|\rho| > 0,95$  практически имеет место функциональная связь.

Пусть  $(x_i, y_i)$ ,  $i = 1, \dots, n$  – выборка объема  $n$  из наблюдений случайного двумерного вектора  $(X, Y)$ . Изображая элементы выборки  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  точками плоскости в декартовой системе координат, получим *диаграмму рассеивания* (облако точек, корреляционное поле).

Для выборочного вектора с учетом того, что

$$\begin{aligned} p_{ij} &= \frac{1}{n}; \\ m_X^* &= \bar{x}; \\ m_Y^* &= \bar{y}, \end{aligned}$$

имеем:

$$K_{XY}^* = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} (\sum x_i y_i - n \bar{x} \bar{y})$$

$$\sigma_x^* = \sqrt{D_x^*} = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} = \sqrt{\frac{1}{n} (\sum x_i^2 - n \bar{x}^2)} ;$$

$$\sigma_y^* = \sqrt{\frac{1}{n} (\sum y_i^2 - n \bar{y}^2)} ,$$

тогда выборочный коэффициент корреляции запишется в виде

$$r = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)(\sum y_i^2 - n \bar{y}^2)}} . \quad (4.4)$$

## 4.2.

### Проверка значимости корреляции

Пусть  $r$  – выборочный коэффициент корреляции, вычисленный по выборке объема  $n$  из генеральной совокупности, имеющей нормальное распределение. Требуется при заданном уровне значимости  $\alpha$  проверить нулевую гипотезу  $H_0$ :  $\rho = 0$  о равенстве нулю коэффициента корреляции генеральной совокупности.

Если нулевая гипотеза будет отвергнута, то говорят о *значимости* коэффициента корреляции, а значит о том, что случайные величины  $X$  и  $Y$  коррелированы. Если нулевая гипотеза принимается, то коэффициент корреляции незначим, и случайные величины  $X$  и  $Y$  некоррелированы.

Для проверки гипотезы  $H_0$  используется статистика

$$t = r \sqrt{(n-2) / \sqrt{1-r^2}} , \quad (4.5)$$

имеющая распределение Стьюдента с числом степеней свободы  $(n-2)$ .

Пусть, например, альтернативная гипотеза  $H_1$ :  $\rho < 0$  тогда граница критической области определяется квантилью  $t_\alpha(n-2)$ ; если же  $H_1$ :  $\rho \neq 0$

определяются границы двухсторонней критической области  $t_{\alpha/2}(n-2)$  и  $t_{1-\alpha/2}(n-2)$ .

### 4.3.

## Множественная корреляция

Изучается степень тесноты линейной связи между  $k$  случайными величинами  $X_1, X_2, \dots, X_n$ . Выборка представляется в виде матрицы  $X$ , состоящей из результатов  $n$  наблюдений за каждым из  $k$  элементов случайного вектора:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1l} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2l} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{il} & \dots & x_{ik} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nl} & \dots & x_{nk} \end{pmatrix}$$

– размерность этой матрицы  $n \times k$ :  $n$  строк,  $k$  столбцов. В первом столбце представлены  $n$  значений случайной величины  $X_1$  во втором –  $n$  значений  $X_2$  и т. д. По этим данным можно построить *ковариационную матрицу*:

$$K = \begin{pmatrix} K_{11} & K_{12} & \dots & K_{1k} \\ \dots & \dots & \dots & \dots \\ K_{k1} & K_{k2} & \dots & K_{kk} \end{pmatrix}, \quad (4.6)$$

где

$$K_{ij} = K_{ji} = M[(x_i - \bar{x}_i)(x_j - \bar{x}_j)] \quad i = 1, \dots, k; j = 1, \dots, k,$$

т.е. матрица симметрична; элементы главной диагонали

$$K_{ii} = M[(x_i - \bar{x}_i)^2] = D_{x_i}$$

– дисперсии соответствующей случайной величины  $X_{i\cdot}$ . Также строится корреляционная матрица



$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1m} & \dots & r_{1k} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{21} & r_{22} & r_{23} & \dots & r_{2m} & \dots & r_{2k} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & r_{k3} & \dots & r_{km} & \dots & 1 \end{pmatrix} \quad (4.7)$$

– симметричная с единичными диагональными элементами. Недиagonальные элементы этой матрицы – это выборочные коэффициенты парной корреляции, определяемые как

$$r_{lm} = \frac{\sum x_{il}x_{im} - n\bar{x}_l\bar{x}_m}{\sqrt{[\sum x_{il}^2 - n\bar{x}_l^2][\sum x_{im}^2 - n\bar{x}_m^2]}};$$

здесь  $l = 1, \dots, k; m = 1, \dots, k; i = 1, \dots, n$ .

$$\bar{x}_l = \frac{1}{n} \sum x_{il}; \bar{x}_m = \frac{1}{n} \sum x_{im};$$

$x_{il}$  – результат  $i$ -го наблюдения за случайной величиной  $X_l$ .

Коэффициенты парной корреляции при множественной корреляции могут привести к неправильным выводам при изучении тесноты связи между двумя случайными величинами  $X_l$  и  $X_m$ , так как на связь между этими двумя величинами могут оказывать влияние и другие компоненты  $k$ -мерного случайного вектора.

Для исключения влияния других случайных величин определяют *частный коэффициент корреляции*, показывающий меру взаимосвязи между двумя величинами при исключении влияния других. Частный коэффициент корреляции выражается через элементы корреляционной матрицы  $R$ . Например, частный коэффициент корреляции между случайными величинами  $X_1$  и  $X_2$  равен

$$r_{12 \cdot 3, 4, \dots, k} = -\frac{R_{12}}{\sqrt{R_{11}R_{22}}}, \quad (4.8)$$

где  $R_{lm}$  – алгебраическое дополнение элемента  $r_{lm}$  корреляционной матрицы  $R$ . Напомним, что алгебраическим дополнением элемента  $r_{lm}$  называется

определитель, получаемый из определителя матрицы  $R$  вычеркиванием  $l$ -ой строки и  $m$ -ого столбца, умноженный на  $(-1)^{l+m}$ .

*Множественный коэффициент корреляции* характеризует тесноту связи между одной переменной (результативной) и остальными, входящими в  $k$ -мерный вектор. Если, например, результативной является случайная величина  $X_1$ , то множественный коэффициент корреляции есть

$$r_1 = r_{1/23\dots k} = \sqrt{1 - \frac{|R|}{R_{11}}}, \quad (4.9)$$

где  $|R|$  – определитель корреляционной матрицы. Квадрат множественного коэффициента корреляции называется *коэффициентом детерминации*. Если  $r_1^2 = 1$ , то величина  $X_1$  является линейной комбинацией случайных величин  $X_2, X_3, \dots, X_n$ . Если же  $r_1^2 = 0$ , то величина  $X_1$  не коррелирована ни с одной из случайных величин  $X_2, X_3, \dots, X_n$ . Чем лучше  $X_1$  приближается линейными комбинациями  $X_2, X_3, \dots, X_n$ , тем ближе коэффициент детерминации к единице.

Значимость парных коэффициентов корреляции определяется с использованием статистики Стьюдента. По аналогии проверяется значимость частных коэффициентов корреляции; для этого используется статистика

$$t = \frac{r\sqrt{n-3}}{\sqrt{1-r^2}}; \quad (4.10)$$

(здесь  $r$  – соответствующий частный коэффициент корреляции), имеющая распределение Стьюдента с числом степеней свободы  $(n-3)$ .

Для проверки значимости коэффициента детерминации используется критерий Фишера. Выборочное значение статистики

$$F = \frac{\frac{1}{k-1}r_1^2}{\frac{1}{n-k}(1-r_1^2)} \quad (4.11)$$

сравнивается с критическим значением, зависящим от уровня значимости, вида альтернативной гипотезы и чисел степеней свободы  $(k-1)$  и  $(n-k)$ .

#### 4.4.

#### Примеры расчета

Пример 4.1. При производственных испытаниях определяется толщина сердцевины сверла  $X$  в мм и стойкость – время работы сверла до затупления  $Y$  в мин. Провести корреляционный анализ связи между этими показателями.

$X$	0,75	0,79	0,81	0,82	0,84	0,85
$Y$	14	23	42	39	46	40
$X$	0,86	0,89	0,90	0,94	0,95	0,98
$Y$	42	45	49	51	85	78

1. Объем выборки  $n = 12$ . Выборочные средние

$$\bar{x} = (0,75 + 0,79 + \dots + 0,98) / 12 = 0,865;$$

$$\bar{y} = (14 + 23 + \dots + 78) / 12 = 46,167.$$

2. Выборочный коэффициент корреляции

$$r = (0,75 \cdot 14 + \dots + 0,98 \cdot 78 - 12 \cdot 0,865 \cdot 46,167) / [(0,75^2 + \dots + 0,98^2 - 12 \cdot 0,865^2)(14^2 + \dots + 78^2 - 12 \cdot 46,167^2)]^{1/2} = 0,90.$$

3. Проверим значимость корреляции: выборочное значение статистики Стьюдента

$$t = 0,90 \cdot [(12 - 2) / (1 - 0,90^2)]^{1/2} = 6,61;$$

критическое значение при правостороннем критерии на уровне значимости  $\alpha = 0,05$

$$t_{1-\alpha}(n-2) = t_{0,95}(10) = 1,812;$$

выборочное значение статистики попало в критическую область, нулевая гипотеза отвергается; следовательно, между толщиной сердцевины сверла и стойкостью имеет место сильная корреляция.

На рис. 4.1 показана соответствующая диаграмма рассеяния.

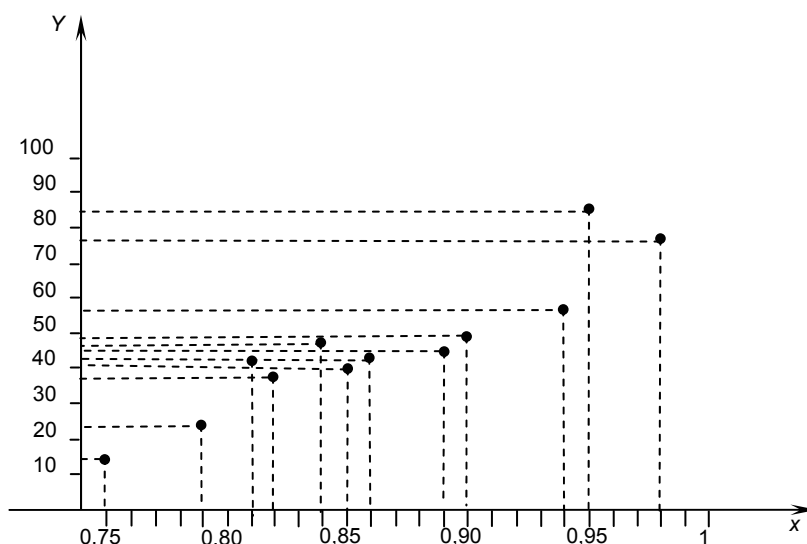


Рис. 4.1

Пример 4.2. При полевых испытаниях подземных стальных трубопроводов получены значения деформации трубопровода  $X$  (мм) в зависимости от жесткости  $Y$  (кгс/см) основания траншеи, в которую укладывается трубопровод:

X	1,08	0,94	0,96	0,73	0,64	0,68	0,63	0,60	0,67	0,52
Y	5,7	7,2	10,1	11,2	13,4	13,7	13,9	14,2	16,0	18,2

Определить коэффициент корреляции и проверить его значимость на уровне значимости  $\alpha = 0,05$  при альтернативной гипотезе  $H_1: \rho < 0$

1. Объем выборки  $n = 10$ .

Выборочная средняя по  $x$ :

$$\bar{x} = \frac{1}{n} \sum x_i = \frac{1}{10} (1,08 + \dots + 0,52) = 0,745 \text{ мм},$$

выборочная средняя по  $y$ :

$$\bar{y} = \frac{1}{n} \sum y_i = \frac{1}{10} (5,7 + \dots + 18,2) = 12,36 \text{ кгс/см}^2$$

2. Выборочный коэффициент корреляции найдем по формуле:

$$r = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)(\sum y_i^2 - n \bar{y}^2)}} = \frac{1,08 \cdot 5,7 + \dots + 0,52 \cdot 18,2 - 10 \cdot 0,745 \cdot 12,36}{\sqrt{(1,08^2 + \dots + 0,52^2 - 10 \cdot 0,745^2)(5,7^2 + \dots + 18,2^2 - 10 \cdot 12,36^2)}} = -0,934$$

3. Найдем выборочное значение  $t$ -статистики для проверки значимости коэффициента корреляции:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0,934\sqrt{8}}{\sqrt{1-0,934^2}} = -7,43.$$

4. Альтернативная гипотеза  $H_1: \rho < 0$  поэтому границей критической области является квантиль Стьюдента  $t_{\alpha}(n-2)$ : по таблице находим

$$t_{0,05}(8) = -t_{0,95}(8) = -1,860$$

5. Видим, что выборочное значение статистики  $t = 7,43$  попало в критическую область, поэтому гипотеза  $H_0: \rho = 0$  о незначимости коэффициента корреляции отклоняется, коэффициент корреляции значим, а т. к.  $r = -0,934$ , то между деформацией трубопровода и жесткостью основания существует сильная корреляция.

Пример 4.3. Исследовалось влияние на ползучесть бетона ( $X_1$ ), расхода цемента на 1 м<sup>3</sup> бетона ( $X_2$ ) и влажности среды ( $X_3$ ).

X1	X2	X3
27	340	80
64	300	75
123	250	68
147	180	63
189	140	59
214	110	52
327	70	48
412	60	40

Построить корреляционную матрицу и определить выборочные частные коэффициенты корреляции. Проверить значимость частных коэффициентов корреляции. Вычислить коэффициент детерминации и проверить его значимость. Принять  $\alpha = 0,1$ .

1. Объем выборки  $n = 8$ .

Выборочные средние

$$\bar{x}_1 = \frac{1}{8}(27 + \dots + 412) = 187,88,$$

$$\bar{x}_2 = \frac{1}{8}(340 + \dots + 60) = 181,25,$$

$$\bar{x}_3 = \frac{1}{8}(80 + \dots + 40) = 60,62.$$

2. Коэффициенты парной корреляции найдем по формуле

$$\begin{aligned} r_{12} = r_{21} &= \frac{\sum x_{i1}x_{i2} - n\bar{x}_1\bar{x}_2}{\sqrt{(\sum x_{i1}^2 - n\bar{x}_1^2)(\sum x_{i2}^2 - n\bar{x}_2^2)}} = \\ &= \frac{27 \cdot 340 + \dots + 412 \cdot 60 - 8 \cdot 187,77 \cdot 181,25}{\sqrt{(27^2 + \dots + 412^2 - 8 \cdot 187,88)(340^2 + \dots + 60^2 - 8 \cdot 181,25^2)}} = -0,934. \end{aligned}$$

Строим корреляционную матрицу:

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix} = \begin{pmatrix} 1 & -0,934 & -0,977 \\ -0,934 & 1 & 0,979 \\ -0,977 & 0,979 & 1 \end{pmatrix}.$$

3. Найдем частные коэффициенты корреляции. Для этого вначале вычислим алгебраические дополнения элементов матрицы  $R$ .

$$R_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 0,0416$$

$$R_{22} = \begin{vmatrix} 1 & r_{13} \\ r_{31} & 1 \end{vmatrix} = 1 - r_{13}^2 = 0,0455$$

$$R_{33} = \begin{vmatrix} 1 & r_{21} \\ r_{21} & 1 \end{vmatrix} = 1 - r_{12}^2 = 0,1276$$

$$R_{12} = -\begin{vmatrix} r_{12} & r_{23} \\ r_{31} & 1 \end{vmatrix} = -r_{21} + r_{23}r_{31} = -0,0225$$

$$R_{23} = -\begin{vmatrix} 1 & r_{12} \\ r_{31} & r_{32} \end{vmatrix} = -r_{32} + r_{12}r_{31} = -0,0665$$

$$R_{13} = \begin{vmatrix} r_{21} & 1 \\ r_{31} & r_{32} \end{vmatrix} = r_{21}r_{32} - r_{31} = 0,0626$$

Выборочные частные коэффициенты определяются по формуле

$$r_{12/3} = -\frac{R_{12}}{\sqrt{R_{11}R_{22}}} = \frac{0,0225}{\sqrt{0,0416 \cdot 0,0456}} = 0,517$$

$$r_{31/2} = -\frac{R_{13}}{\sqrt{R_{11}R_{33}}} = \frac{-0,0625}{\sqrt{0,0416 \cdot 0,1276}} = -0,859$$

$$r_{23/1} = -\frac{R_{23}}{\sqrt{R_{22}R_{33}}} = \frac{0,0665}{\sqrt{0,0455 \cdot 0,1276}} = 0,873$$

4. Для проверки значимости частных коэффициентов корреляции найдем выборочные значения статистики Стьюдента:

$$t_{12} = r_{12/3} \sqrt{n-3} / \sqrt{1-r_{12/3}^2} = 0,517 \sqrt{5} / \sqrt{1-0,517^2} = 1,35$$

$$t_{31} = -3,752; t_{23} = 4,000$$

В качестве альтернативной примем гипотезу  $H_1: \rho \neq 0$ . Границы критической области  $t_{\alpha/2}(n-3)$  и  $t_{1-\alpha/2}(n-3)$  найдем по таблице

$$t_{0,05}(5) = -t_{0,95}(5) = -2,015$$

Видим, что коэффициент  $r_{12/3}$  оказался незначимым (значение статистики  $t_{12}$  – в области принятия решения) коэффициенты  $r_{13/2}$  и  $r_{23/1}$  – значимы.

5. Находим коэффициент детерминации, рассматривая переменную  $X_1$  (ползучесть бетона) как результирующую:

$$r_1^2 = 1 - |R|/R_{11},$$

где  $|R|$  – определитель корреляционной матрицы:

$$|R| = 1 \cdot R_{11} + r_{12}R_{12} + r_{13}R_{13} = 1 \cdot 0,0146 + 0,934 \cdot 0,0225 - 0,977 \cdot 0,0626 = 0,00145 \quad \text{тогда}$$

$$r_1^2 = 1 - 0,00145 / 0,0416 = 0,965$$

6. Для проверки значимости множественного коэффициента детерминации найдем выборочное значение статистики Фишера

$$F = \frac{\frac{1}{k-1} r_1^2}{\frac{1}{n-k} (1-r_1^2)},$$

где  $n = 8$ ,  $k = 3$ . Тогда

$$F = \frac{0,965^2 / 2}{(1 - 0,965^2) / 5} = 33,83$$

В качестве альтернативной прием гипотезу  $H_1: \rho > 0$ . Тогда граница критической области определяется квантилью  $F_{1-\alpha}(k-1, n-k)$ , которую найдем по таблице  $F_{0,9}(2, 5) = 3.78$ .

Видим, что выборочное значение статистики Фишера попало в критическую область, поэтому гипотеза  $H_0$  о незначимости отвергается, коэффициент детерминации значим, что указывает на существование корреляционной связи между ползучестью бетона с одной стороны и расходом цемента и влажностью с другой.

## 4.5. Корреляционный анализ в Excel

Для построения диаграммы рассеяния используется *Мастер диаграмм* / Тип диаграммы: *Точечная*. Построим диаграмму рассеяния для данных из примера 4.1: результат показан на рис. 4.2.

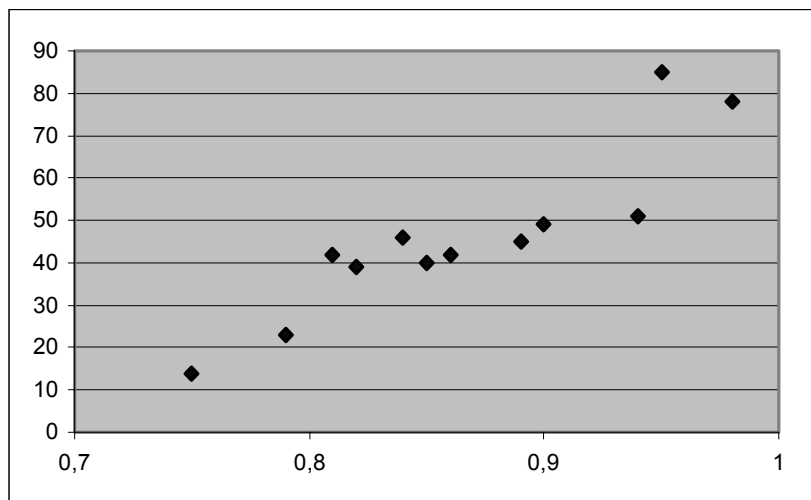


Рис. 4.2



Для расчета коэффициента корреляции и проверки его значимости могут быть использованы встроенные функции КОРРЕЛ (коэффициент корреляции) и СТЬЮДРАСПОБР (для вычисления квантилей распределения Стьюдента). Обратите внимание на ввод уровня значимости  $\alpha$  в последней функции: функция предназначена для использования в двустороннем критерии, у нас по условию задачи – правосторонний (т. е. односторонний) критерий, поэтому введено удвоенное значение уровня значимости. Исходные данные введены в ячейках B1:M2, функция СЧЕТ – в ячейке N21. Результаты приведены на рис. 4.3.

$n =$	=СЧЁТ(B1:M1)	12
$r =$	=КОРРЕЛ(B1:M1;B2:M2)	0,90
$t =$	=N22*КОРЕНЬ((N21-2)/(1-N22^2))	6,61
$\alpha =$	0,05	0,05
$tkp =$	=СТЮДРАСПОБР(2*N25;N21-2)	1,81

Рис. 4.3

Для расчета выборочного коэффициента корреляции также можно воспользоваться инструментом анализа данных *Корреляция* (рис. 4.4).

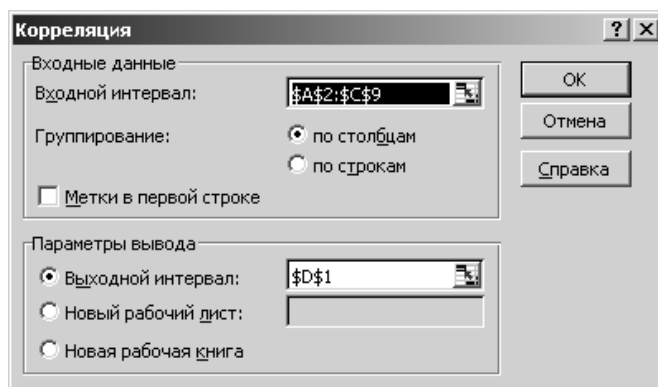


Рис. 4.4

Это особенно удобно, если требуется найти парные коэффициенты корреляции для нескольких переменных.

	A	B	C	D	E	F	G	H
1	X1	X2	X3		Столбец 1	Столбец 2	Столбец 3	
2	27	340	80	Столбец 1	1			
3	64	300	75	Столбец 2	-0,933105997	1		
4	123	250	68	Столбец 3	-0,976416733	0,979783825	1	
5	147	180	63					
6	189	140	59					
7	214	110	52					
8	327	70	48					
9	412	60	40					
10								

Рис. 4.5

На рис. 4.5 приведены выборочные данные и результат расчета для данных из примера 4.3.

## 4.6. Корреляционный анализ в Statistica

Для анализа степени тесноты линейной связи между переменными может быть построена корреляционная матрица. Выберите в стартовой панели команду *Корреляционные матрицы*, в окне *Корреляция Пирсона*; задайте один из двух возможных типов корреляционных матриц, квадратную или прямоугольную. Введите все три переменные M1, M2 и NEWVAR из таблицы исходных данных для анализа.

После щелчка *OK* получите корреляционную матрицу. Красным цветом в ней выделены корреляции, значимые на уровне значимости 0,05: такой оказалась корреляция между переменными M2 и NEWVAR.

Variable	M1	M2	NEWVAR
M1	1.00	-.46	.33
M2	-.46	1.00	.68
NEWVAR	.33	.68	1.00

Рис. 4.6

Щелчком по кнопке *Матрица* можно построить матричный график с гистограммами по каждой переменной, диаграммами рассеяния между каждой

парой переменных и соответствующими линиями регрессии (рис. 4.7), удобный для визуальной оценки переменных и связей между ними.

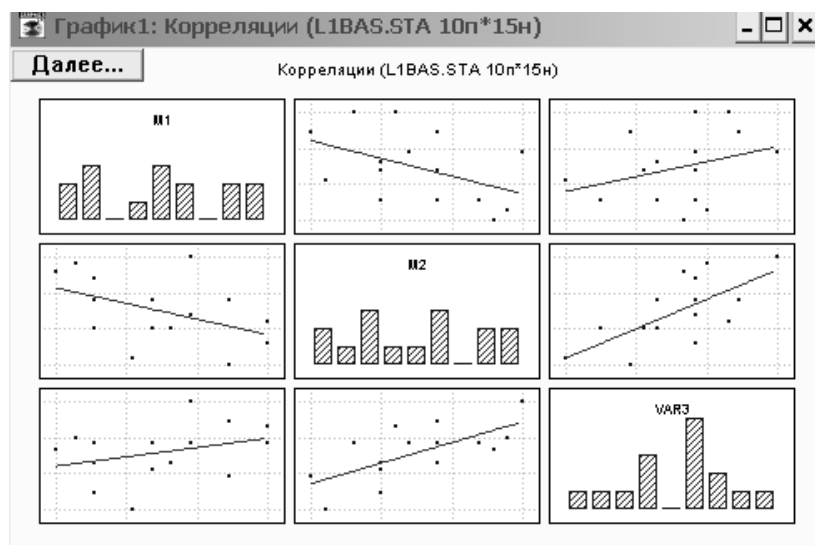


Рис. 4.7

## Контрольные вопросы

1. Какая зависимость называется стохастической?
2. Что означает некоррелированность случайных величин  $X$  и  $Y$ ?
3. В каком случае коэффициент корреляции равен по модулю единице?
4. Выведите формулу для определения ковариации двумерной выборки.
5. Как проверить значимость коэффициента парной корреляции?
6. Как строится ковариационная матрица?
7. Как вычисляются коэффициенты корреляционной матрицы?
8. Что означает равенство коэффициента детерминации нулю? единице?
9. Для чего определяется частный коэффициент корреляции?
10. Как проверить значимость коэффициента детерминации?

## РЕГРЕССИОННЫЙ АНАЛИЗ

### 5.1.

#### Парная линейная регрессия

Регрессионный анализ – раздел прикладной статистики, изучающий связь между зависимой переменной  $Y$  и одной или несколькими независимыми переменными. Вначале рассмотрим парный анализ, когда независимая переменная одна. Пусть эта переменная  $X$  принимает некоторые фиксированные значения  $x_1, x_2, \dots, x_n$ . Соответствующие значения зависимой переменной  $Y$  имеют разброс вследствие погрешности измерений и различных неучтенных факторов и оказались равными  $y_1, y_2, \dots, y_n$ .

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

$$Y = \beta_0 + \beta_1 x + \varepsilon, \quad (5.1)$$

где  $\beta_0$  и  $\beta_1$  – параметры линейной регрессии,  $\varepsilon$  – случайная ошибка наблюдения; предполагается, что математическое ожидание  $M(\varepsilon) = 0$ , а дисперсия  $D(\varepsilon) = \sigma^2$  постоянна.

Задача регрессионного анализа сводится к оценке параметров регрессии  $\beta_0$  и  $\beta_1$ , проверке гипотезы о значимости модели и оценке её адекватности – достаточно ли хорошо согласуется модель с результатами наблюдений?

Для оценки параметров регрессии используется *метод наименьших квадратов*: в качестве оценок принимаются такие значения  $\beta_0$  и  $\beta_1$ , которые минимизируют сумму квадратов отклонений наблюдаемых значений  $y_i$  от расчетных точек  $\tilde{y}_i$ . Для парной линейной модели эти оценки определяются по формулам:

$$\begin{aligned}\tilde{\beta}_1 &= Q_{xy} / Q_x, \\ \tilde{\beta}_0 &= \bar{y} - \tilde{\beta}_1 \bar{x},\end{aligned}\tag{5.2}$$

где

$$Q_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \bar{x} \bar{y},\tag{5.3}$$

$$Q_x = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n \bar{x}^2.\tag{5.4}$$

Расчетное значение  $y_i = \beta_0 + \beta_1 x_i$ . Разности между наблюдаемыми и расчетными значениями  $y_i - \tilde{y}_i$ , называются *остатками*, а соответствующая сумма квадратов – остаточной суммой квадратов:

$$Q_e = \sum (y_i - \tilde{y}_i)^2.\tag{5.5}$$

Воспользуемся алгоритмом однофакторного дисперсионного анализа, где

$$Q_y = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n \bar{y}^2\tag{5.6}$$

– общая сумма квадратов, а сумма квадратов, обусловленная регрессией, есть

$$Q_R = \sum (\tilde{y}_i - \bar{y})^2 = \beta_1^2 Q_x\tag{5.7}$$

Тогда остаточную сумму квадратов можно вычислить из тождества

$$Q_y = Q_R + Q_e.\tag{5.8}$$

Линейная регрессионная модель называется *незначимой*, если параметр  $\beta_1 = 0$ . Для проверки гипотезы  $H_0: \beta = 0$  используется статистика Фишера

$$F = \frac{Q_R}{\frac{1}{n-2} Q_e},\tag{5.9}$$

которая при заданном уровне значимости  $\alpha$  сравнивается с квантилью  $F_{1-\alpha}(1, n - 2)$  с числом степеней свободы 1 и  $(n - 2)$ ; если оказывается

$$F > F_{1-\alpha}(1, n - 2),$$

то гипотеза  $H_0$  отклоняется и говорят, что регрессионная модель статистически значима.

Кроме значимости проверяется и *адекватность* модели. Иногда адекватность проверяется по диаграмме рассеивания с нанесенной расчетной прямой. Если же адекватность неочевидна, то проводят специальную проверку. В этом случае необходимо иметь несколько результатов наблюдений  $y_{ij}$  при одних и тех же значениях  $x_i$ ,  $i = 1, \dots, m$ ;  $j = 1, \dots, n_i$ , где  $m$  – количество различных значений  $x_i$ . Очевидно,  $\sum_{i=1}^m n_i = n$ . Если модель адекватна результатам наблюдений, то средние из  $n_i$  наблюдений  $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$  должны быть близки к вычисленным значениям  $\tilde{y}_i$ , то есть сумму квадратов

$$Q_n = \sum_{i=1}^m (\bar{y}_i - \tilde{y}_i)^2$$

можно рассматривать как меру неадекватности модели, остаточную сумму квадратов можно представить в виде суммы

$$Q_e = Q_n + Q_p,$$

где  $Q_p$  – сумма квадратов чистой ошибки

$$Q_p = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

Для проверки адекватности модели используется статистика Фишера

$$F = \frac{\frac{1}{m-2} Q_n}{\frac{1}{n-m} Q_p}.$$

Если выборочное значение этой статистики оказывается меньше критического значения  $F_{1-\alpha}(m-2, n-m)$ , то гипотеза об адекватности линейной модели принимается. Если же это условие не выполняется, то используют одну из нелинейных моделей.

Проверка адекватности модели не всегда возможна. Если нет дополнительных измерений  $Y$ , ограничиваются сравнением статистики  $F$  с  $F_{1-\alpha}$ . Если

$$F > 4 F_{1-\alpha}(m-2, n-m),$$

то модель считается пригодной для использования при прогнозе значений  $Y$  по известным значениям  $X$ .

## 5.2. Парная нелинейная регрессия

В общем случае нелинейная регрессионная модель (нелинейная по фактору  $X$ , но линейная по параметрам  $\beta_j$ ) имеет вид

$$Y = \beta_0 + \beta_1 \varphi_1(x) + \beta_2 \varphi_2(x) + \dots + \beta_{k-1} \varphi_{k-1}(x) + \varepsilon, \quad (5.10)$$

где  $\beta_j$  – неизвестные параметры, а  $\varphi_j(x)$  – известные базисные функции. Они могут быть степенными  $\varphi_j(x) = x^j$ , тригонометрическими  $\varphi_j(x) = \sin(\lambda_j x)$  и т. д.

Используя метод наименьших квадратов, для оценки параметров можно получить нормальную систему:

$$\begin{aligned} n\beta_0 + \beta_1 \sum \varphi_1(x_i) + \beta_2 \sum \varphi_2(x_i) + \dots + \beta_{k-1} \sum \varphi_{k-1}(x_i) &= \sum y_i, \\ \beta_0 \sum \varphi_1(x_i) + \beta_1 \sum \varphi_1^2(x_i) + \beta_2 \sum \varphi_1(x_i) \varphi_2(x_i) + \dots + \beta_{k-1} \sum \varphi_1(x_i) \varphi_{k-1}(x_i) &= \sum \varphi_1(x_i) y_i \\ \beta_0 \sum \varphi_{k-1}(x_i) + \beta_1 \sum \varphi_1(x_i) \varphi_{k-1}(x_i) + \dots + \beta_{k-1} \sum \varphi_{k-1}^2(x_i) &= \sum \varphi_{k-1}(x_i) y_i \end{aligned}$$

в частности, если рассматривается параболическая модель

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon, \quad (5.11)$$

имеем  $k = 3$ ,  $\varphi_1(x) = x$ ,  $\varphi_2(x) = x^2$  и нормальная система примет вид:

$$\begin{aligned} n\beta_0 + \beta_1 \sum x_i + \beta_2 \sum x_i^2 &= \sum y_i, \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 + \beta_2 \sum x_i^3 &= \sum x_i y_i, \\ \beta_0 \sum x_i^2 + \beta_1 \sum x_i^3 + \beta_2 \sum x_i^4 &= \sum x_i^2 y_i, \end{aligned} \quad (5.12)$$

для гиперболической модели

$$\begin{aligned} Y &= \beta_0 + \beta_1/x + \varepsilon \\ n\beta_0 + \beta_1 \sum \frac{1}{x_i} &= \sum y_i \\ \beta_0 \sum \frac{1}{x_i} + \beta_1 \sum \frac{1}{x_i^2} &= \sum \frac{x_i}{y_i}. \end{aligned} \quad (5.13)$$

Мы рассмотрели регрессионные модели, нелинейные по фактору  $X$ , но линейные по параметрам  $\beta_j$ . Во многих практических задачах зависимость между  $X$  и  $Y$  нелинейна и по параметрам. В этом случае по возможности пытаются свести нелинейную по параметрам модель к модели, линейной по параметрам.

Пусть, например, зависимость между переменными  $z$  и  $x$  имеет вид

$$z = \frac{1}{\beta_0 + \beta_1 x}.$$

Представим ее в виде

$$\frac{1}{z} = \beta_0 + \beta_1 x$$

и введем новую переменную  $y = 1/z$ , тогда получим модель

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

линейную по параметрам. Если

$$z = e^{\beta_0 + \beta_1 x},$$

то, прологарифмировав:

$$\ln z = \beta_0 + \beta_1 x,$$

и введя обозначение

$$y = \ln z,$$

также получим линейную модель.

По аналогии с парной линейной регрессией, проводится проверка значимости и адекватности модели.

Очевидно, для одного набора опытных данных  $(x_i, y_i)$  можно использовать различные модели, которые окажутся и значимыми, и адекватными. Для характеристики качества той или иной модели используется коэффициент корреляции, показывающий степень тесноты линейной связи между опытными значениями  $y_i$ , и их предсказаниями  $\tilde{y}_i$  по модели.



По формуле для выборочного коэффициента корреляции имеем:

$$r = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} = \frac{\sum y_i \tilde{y}_i - n\bar{y}\bar{\tilde{y}}}{\sqrt{(\sum y_i^2 - n\bar{y}^2)(\sum \tilde{y}_i^2 - n\bar{\tilde{y}}^2)}},$$

где

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \varphi_1(x_i) + \dots + \tilde{\beta}_{k-1} \varphi_{k-1}(x_i); \bar{\tilde{y}} = \frac{1}{n} \sum y_i \tilde{y}_i = \frac{1}{n} \sum \tilde{y}_i.$$

Чем ближе коэффициент корреляции к единице (по модулю), при условии его значимости, тем более качественной считается модель из набора моделей одинаковой размерности. Вообще, как для парной, так и ниже рассматриваемой множественной регрессии, для оценки качества модели используются помимо коэффициента корреляции и другие критерии. Различают при этом внутренние, смешанные и внешние меры качества.

### 5.3.

## Множественная регрессия

Если случайная величина  $Y$  зависит от нескольких независимых переменных  $x_1, x_2, \dots, x_n$ , то исследование зависимости между  $Y$  и  $x_j$  ( $j = 1, \dots, k - 1$ ) составляет предмет *множественного* регрессионного анализа.

Регрессионную модель представим в виде

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1} + \varepsilon \quad (5.14)$$

или в матричной форме

$$Y = X\beta + \varepsilon, \quad (5.15)$$

где

$$Y = (y_1 \ y_2 \ \dots \ y_n)^T$$

– вектор наблюдений, содержащий  $n$  значений  $Y_i$  (случайные величины), индекс " $T$ " означает транспонирование матрицы;

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,k-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2,k-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n,k-1} \end{pmatrix}$$

– регрессионная матрица размера  $n \times k$ , содержащая элементы  $x_{ij}$  – результаты  $i$ -го наблюдения за входными функциями  $x_j$ ;  $k$  – количество параметров;  $x_{ij}$  – неслучайные величины (в общем случае – базисные функции входных параметров);

$$\beta = (\beta_0 \ \beta_1 \ \dots \ \beta_{k-1})^T$$

– вектор неизвестных параметров регрессии, подлежащих оцениванию (неслучайные величины);

$$\varepsilon = (\varepsilon_1 \ \varepsilon_2 \ \dots \ \varepsilon_n)^T$$

– вектор ошибок, содержащий неизвестные погрешности наблюдений  $\varepsilon_i$  (случайные величины, распределенные по нормальному закону, некоррелированные и статистически независимые, с нулевым математическим ожиданием и постоянной дисперсией).

Обычно значения выходной случайной величины  $Y$  называют откликом, а входные величины  $x_j$  – регрессорами. Очевидно, если в модели парной нелинейной регрессии  $\varphi_j(x)$  обозначить через новые переменные  $x_j$  ( $j = 1, \dots, k - 1$ ), то модель

$$Y = \beta_0 + \beta_1 \varphi_1(x) + \beta_2 \varphi_2(x) + \dots + \beta_{k-1} \varphi_{k-1}(x) + \varepsilon$$

может также рассматриваться с позиций множественного регрессионного анализа.

Оценки параметров модели по методу наименьших квадратов определяются по формуле

$$\tilde{\beta} = (X^T X)^{-1} \times X^T Y \quad (5.16)$$

где  $(X^T X)^{-1}$  – матрица, обратная матрице  $X^T X$ .

При решении задачи поиска оптимальной регрессии описанная процедура является предварительной; точное решение проблемы предполагает (помимо использования внутренних, смешанных и внешних мер) проверку соблюдения условий применения регрессионного анализа и вычислительную адаптацию к их нарушениям.

Для проверки значимости рассматриваемой модели в качестве нулевой используется гипотеза  $H_0: \beta_1 = \beta_2 = \dots = \beta_{k-1}$  о том, что все регрессоры  $x_j$  ( $j = 1, \dots, k - 1$ ) не оказывают существенного влияния на отклик. Статистика Фишера

$$F = \frac{\frac{1}{k-1} Q_R}{\frac{1}{n-k} Q_e} \quad (5.17)$$

сравнивается с квантилью  $F_{1-\alpha}(k-1, n-k)$ . Здесь

$$\begin{aligned} Q_R &= \tilde{\beta}^T X^T Y - n\bar{y}^2 \\ Q_e &= Q_y - Q_R \\ Q_y &= \sum y_i^2 - n\bar{y}^2 \end{aligned} \quad (5.18)$$

Если гипотеза  $H_0$  отклоняется, то проверяется значимость каждого

регрессора:  $H_0^{(j)}: \beta_j = 0$ , то есть предположение о том, что регрессор  $X_j$  статистически незначим. Используется статистика Стьюдента

$$t_j = \frac{|\tilde{\beta}_j|}{S_j}, \quad (5.19)$$

где  $s_j$  - среднеквадратическое отклонение параметра  $\beta_j$ , которое можно найти по формуле:

$$S_j = \sqrt{\frac{Q_e}{n-k} c_{jj}}, \quad (5.20)$$

где  $c_{jj}$  -диагональные элементы матрицы  $(X^T X)^{-1}$ . Найденное значение сравнивается с квантилью  $t_{1-\alpha/2}(n-k)$ . Если какой-либо из параметров оказался незначимым, соответствующий регрессор  $x_j$  из модели исключается.

Для новой модели заново выполняют все расчеты и сопоставляют ее точность с исходной моделью. Такой подход к поиску оптимальной регрессионной модели называется структурной идентификацией.

Кроме того, можно оценить степень важности каждого регрессора путем анализа частных коэффициентов корреляции и таким образом проранжировать регрессоры по степени их важности для модели.

В частном случае двух регрессоров

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

если не предполагается использование ЭВМ, удобнее провести расчет в такой последовательности. Найдем суммы

$$\begin{aligned} Q_{x1} &= \sum (x_{i1} - \bar{x}_1)^2 = \sum x_{i1}^2 - n\bar{x}_1^2 \\ Q_{x2} &= \sum (x_{i2} - \bar{x}_2)^2 = \sum x_{i2}^2 - n\bar{x}_2^2 \\ Q_y &= \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 \\ Q_{x1x2} &= \sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = \sum x_{i1}x_{i2} - n\bar{x}_1\bar{x}_2 \\ Q_{x1y} &= \sum (x_{i1} - \bar{x}_1)(y_i - \bar{y}) = \sum x_{i1}y_i - n\bar{x}_1\bar{y} \\ Q_{x2y} &= \sum (x_{i2} - \bar{x}_2)(y_i - \bar{y}) = \sum x_{i2}y_i - n\bar{x}_2\bar{y} \end{aligned}$$

Введем обозначения:

$$X = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 \\ \dots & \dots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 \end{pmatrix}, Y = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \dots \\ y_n - \bar{y} \end{pmatrix},$$

$$X^T X = \begin{pmatrix} Q_{x1} & Q_{x1x2} \\ Q_{x1x2} & Q_{x2} \end{pmatrix};$$

тогда

$$(X^T X)^{-1} = \begin{pmatrix} Q_{x2} & -Q_{x1x2} \\ -Q_{x1x2} & Q_{x1} \end{pmatrix} \frac{1}{|X^T X|},$$

где  $|X^T X|$  - определитель матрицы  $X^T X$ ;

$$X^T Y = \begin{pmatrix} Q_{x1y} \\ Q_{x2y} \end{pmatrix}.$$

Оценки параметров регрессии запишутся в виде

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = (X^T X)^{-1} X^T Y = \begin{pmatrix} Q_{x2} & -Q_{x1x2} \\ -Q_{x1x2} & Q_{x1} \end{pmatrix} \begin{pmatrix} Q_{x1y} \\ Q_{x2y} \end{pmatrix} \frac{1}{|X^T X|}.$$

Для проверки гипотез используются те же статистики, что и ранее, причем

$$Q_Q = \tilde{\beta}^T X^T Y = (\tilde{\beta}_1 \tilde{\beta}_2) \begin{pmatrix} Q_{x1y} \\ Q_{x2y} \end{pmatrix},$$

а коэффициенты

$$c_{11} = Q_{x2} / |X^T X|$$

$$c_{22} = Q_{x1} / |X^T X|.$$

## 5.4.

### Примеры расчета

Пример 5.1. Исследуется зависимость между пределом прочности прессованной детали  $y$  (МПа) и температурой при прессовании  $x$  (град.). Предполагается наличие линейной зависимости между этими показателями. Экспериментально получены следующие данные:

$x$	120	125	130	135	140	145	150	155	160	165
$y$	110	107	105	98	100	95	95	92	86	83

Объем выборки  $n = 10$ . Выборочные средние

$$\bar{x} = (120 + 125 + \dots + 165) / 10 = 142,5 ;$$

$$\bar{y} = (110 + 107 + \dots + 83) / 10 = 97,1.$$

Найдем оценки параметров линейной регрессии:

$$Q_{xy} = 120 \cdot 110 + 125 \cdot 107 + \dots + 165 \cdot 83 - 10 \cdot 142,5 \cdot 97,1 = -1172,5;$$

$$Q_x = 120^2 + 125^2 + \dots + 165^2 - 10 \cdot 142,5^2 = 2062,5,$$

тогда

$$\tilde{\beta}_1 = -1172,5 / 2062,5 = -0,57;$$

$$\tilde{\beta}_0 = 97,1 - (-0,57) \cdot 142,5 = 178,11.$$

Уравнение линейной регрессии

$$y = 178,11 - 0,57x.$$

Диаграмма рассеяния и расчетная прямая показаны на рис.5.7.

Проверим значимость регрессии:

$$Q_R = 0,57^2 \cdot 2062,5 = 666,55,$$

$$Q_y = 110^2 + 107^2 + \dots + 83^2 - 10 \cdot 97,1^2 = 692,9,$$

$$Q_e = 692,9 - 666,55 = 26,35,$$

тогда

$$F = 666,55 \cdot (10 - 2) / 26,35 = 202,36.$$

Критическое значение статистики Фишера:

$$F_{0,95}(1, 8) = 5,32.$$

Гипотеза о незначимости отклоняется, регрессионная модель значима.

Пример 5.2. Проведены испытания стального образца на растяжение: при заданных нагрузках  $x$  (кН) определяется удлинение  $Y$  (мм); для некоторых значений нагрузки удлинения измерялись трижды. Предполагается, что сталь имеет линейное упрочнение, т. е. связь между удлинением и нагрузкой прямо пропорциональная.

Определить параметры линейной модели, проверить ее значимость и адекватность, принимая уровень значимости  $\alpha=0,05$ .

X	20	21	21	21	22	23	23	23	24	25	25	25
Y	0,2	1,2	1,0	1,1	1,9	2,8	2,8	2,7	3,9	5,0	4,8	4,7

1. Имеем: объем выборки  $n = 12$ ;

$$\bar{x} = \frac{1}{n} \sum x_i = \frac{1}{12} (20 + 21 \cdot 3 + \dots + 25 \cdot 3) = 22,75$$

$$\bar{y} = \frac{1}{n} \sum y_i = \frac{1}{12} (0,2 + 1,2 + \dots + 4,7) = 2,675$$

$$Q_{xy} = \sum x_i y_i - n \bar{x} \bar{y} = 20 \cdot 0,2 + \dots + 25 \cdot 4,7 - 12 \cdot 22,75 \cdot 2,675 = 31,825$$

$$Q_x = \sum x_i^2 - n \bar{x}^2 = 20^2 + \dots + 3 \cdot 25^2 - 12 \cdot 22,75^2 = 34,25$$

2. Находим оценки параметров линейной регрессии:

$$\tilde{\beta}_1 = Q_{xy} / Q_x = 31,825 / 34,25 = 0,93$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x} = 2,675 - 0,93 \cdot 22,75 = -18,46$$

тогда уравнение линейной регрессии

$$y = -18,46 + 0,93x$$

3. Проверим значимость модели:

$$Q_R = \tilde{\beta}_1^2 Q_x = 0,93^2 \cdot 34,25 = 29,559$$

$$Q_y = \sum y_i^2 - n \bar{y}^2 = 0,2^2 + \dots + 4,7^2 - 12 \cdot 2,675^2 = 29,743$$

$$Q_e = Q_y - Q_R = 29,743 - 29,559 = 0,184$$

Статистика Фишера определяется как

$$F = \frac{Q_R}{\frac{1}{n-2} Q_e} = \frac{29,559}{\frac{1}{10} 0,184} = 1606,5$$

Критические значения находим по таблице

$$F_{1-\alpha}(1, n - 2) = F_{0,95}(1, 10) = 4,96;$$

откуда следует, что регрессия значима (гипотеза о незначимости отклоняется:  $1606,5 \gg 4,96$ ).

4. Проверим адекватность модели. Количество различных значений  $X_j$   $m = 6$ ; имеем:

$$n_2 = 3; \bar{y}_2 = \frac{1}{3}(1,2 + 1,1 + 1,0) = 1,1$$

$$n_4 = 3; \bar{y}_4 = \frac{1}{3}(2,8 + 2,8 + 2,7) = 2,767$$

$$n_6 = 3; \bar{y}_6 = \frac{1}{3}(5,0 + 4,8 + 4,7) = 4,833$$

$$n_1 = n_3 = n_5 = 1; \text{ для } n_i = 1 \ y_{ij} - \bar{y}_i = 0, \text{ т.к. } \bar{y}_i = y_{ij},$$

тогда сумма квадратов чистой ошибки

$$Q_p = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = (1,2 - 1,1)^2 + (1,1 - 1,1)^2 + (1,0 - 1,1)^2 + (2,8 - 2,767)^2 + (2,8 - 2,767)^2 + \\ + (2,7 - 2,767)^2 + (5,0 - 4,833)^2 + (4,8 - 4,833)^2 + (4,7 - 4,833)^2 = 0,073$$

Сумма квадратов, обусловленная неадекватностью,

$$Q_n = Q_e - Q_p = 0,184 - 0,073 = 0,111.$$

Выборочное значение статистики Фишера

$$F = \frac{\frac{1}{m-2} Q_n}{\frac{1}{n-m} Q_p} = \frac{\frac{1}{4} 0,111}{\frac{1}{6} 0,073} = 1,97$$

Критическое значение

$$F_{1-\alpha}(m - 2, n - m) = F_{0,95}(4, 6) = 4,53.$$

Так как выборочное значение статистики оказалось меньше критического, гипотеза об адекватности модели результатам наблюдений принимается.

Пример 5.3. Определяется давление в системе  $Y$  (МПа) в зависимости от времени выдержки  $x$  (мин.).



X	0	1	2	3	4
Y	0,40	0,20	0,10	0,06	0,04

Возможна аппроксимация опытных данных параболической зависимостью  $Y = \beta_0 + \beta_1 x + \beta_2 x^2$  или прямой  $Y = \beta_0 + \beta_1 x$ .

(Строго говоря количество измерений  $(x_i, y_i)$  должно превышать количество оцениваемых параметров в 5÷15 раз. В последующих примерах для простоты расчетов используется малое число измерений).

1. Найдем параметры параболической регрессии, используя соответствующую систему, в которой при  $n = 5$

$$\sum x_i = 0 + 1 + \dots + 4 = 10;$$

$$\sum x_i^2 = 0^2 + 1^2 + \dots + 4^2 = 30 \quad \sum x_i^3 = 100 \quad \sum x_i^4 = 354$$

$$\sum y_i = 0,40 + 0,20 + \dots + 0,04 = 0,80$$

$$\sum x_i y_i = 0 \cdot 0,40 + 1 \cdot 0,20 + \dots + 4 \cdot 0,04 = 0,74$$

$$\sum x_i^2 y_i = 0^2 \cdot 0,40 + 1^2 \cdot 0,20 + \dots + 4^2 \cdot 0,04 = 1,78$$

Нормальная система примет вид:

$$5\tilde{\beta}_0 + 10\tilde{\beta}_1 + 30\tilde{\beta}_2 = 0,80$$

$$10\tilde{\beta}_0 + 30\tilde{\beta}_1 + 100\tilde{\beta}_2 = 0,74$$

$$30\tilde{\beta}_0 + 100\tilde{\beta}_1 + 354\tilde{\beta}_2 = 1,78$$

Первое уравнение умножаем на  $(-2)$  и складываем со вторым, затем его же умножаем на  $(-6)$  и складываем с третьим; получим

$$10\tilde{\beta}_1 + 40\tilde{\beta}_2 = -0,86$$

$$40\tilde{\beta}_1 - 174\tilde{\beta}_2 = -3,02$$

Умножая первое уравнение на  $(-4)$  и складывая со вторым, найдем

$$14\tilde{\beta}_2 = 0,42; \tilde{\beta}_2 = 0,030$$

$$\tilde{\beta}_1 = -0,086 - 4\tilde{\beta}_2; \tilde{\beta}_1 = -0,206$$

$$\tilde{\beta}_0 = 0,074 - 3\tilde{\beta}_1 - 103\tilde{\beta}_2; \tilde{\beta}_0 = 0,392,$$

то есть искомое уравнение

$$y = 0,392 - 0,206x + 0,030x^2$$

2. Для оценки качества полученной модели найдем коэффициент корреляции. По найденному уравнению вычислим  $y_i$ :

X	0	1	2	3	4
Y	0,392	0,216	0,100	0,044	0,048

Определим величины, входящие в формулу для расчета коэффициента корреляции:

$$\sum y_i \tilde{y}_i = 0,40 \cdot 0,392 + \dots + 0,04 \cdot 0,048 = 0,21456$$

$$\bar{y} = \frac{1}{5} \cdot 0,80 = 0,16 \quad \bar{\tilde{y}} = \frac{1}{5} (0,392 + \dots + 0,048) = 0,16$$

$$\sum y_i^2 = 0,40^2 + \dots + 0,04^2 = 0,2152$$

$$\sum \tilde{y}_i^2 = 0,392^2 + \dots + 0,048^2 = 0,21456$$

тогда коэффициент корреляции

$$r = \frac{\sum y_i \tilde{y}_i - n \bar{y} \bar{\tilde{y}}}{\sqrt{(\sum y_i^2 - n \bar{y}^2)(\sum \tilde{y}_i^2 - n \bar{\tilde{y}}^2)}} = \frac{0,21456 - 5 \cdot 0,16 \cdot 0,16}{\sqrt{(0,2152 - 5 \cdot 0,16^2)(0,21456 - 5 \cdot 0,16^2)}} = 0,996$$

3. Найдем параметры прямой. Имеем:

$$\tilde{\beta}_1 = Q_{xy} / Q_x = (0,74 - 5 \frac{10}{5} 0,16) / 30 - 5 \frac{10}{5} \cdot \frac{10}{5} = -0,086$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x} = 0,16 + 0,086 \frac{10}{5} = 0,332$$

то есть искомое уравнение

$$y = 0,332 - 0,086x.$$

Найдем соответствующий коэффициент корреляции. Вычисляем  $y_i$ . По аналогии с предыдущим находим:

X	0	1	2	3	4
Y	0,332	0,246	0,160	0,074	-0,012

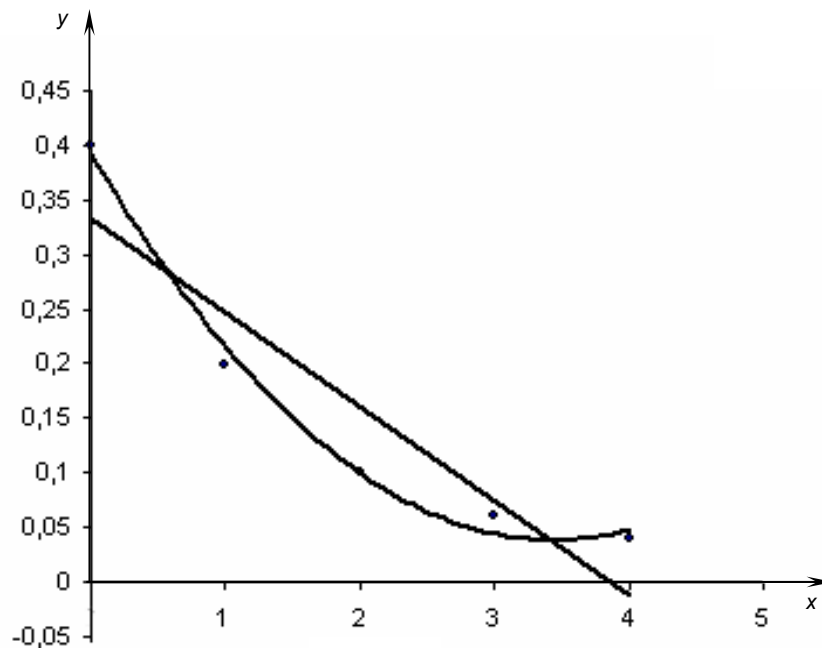


Рис. 5.1.

$$\sum y_i \tilde{y}_i = 0,20196$$

$$\tilde{\bar{y}} = 0,16$$

$$\sum \tilde{y}_{2i} = 0,20196$$

тогда коэффициент корреляции

$$r = \frac{0,20196 - 0,5 \cdot 0,16 \cdot 0,16}{\sqrt{(0,2152 - 5 \cdot 0,16^2)(0,20196 - 0,5 \cdot 0,16^2)}} = 0,921$$

5. Сравнивая значения коэффициентов корреляции, видим, что параболическая модель существенно лучше отображает результаты наблюдений. Это же видно и из графиков (рис. 5.1).

Пример 5.3. При исследовании обрабатываемости цветных сплавов у в зависимости от относительного удлинения  $X_1(\%)$  и предела прочности  $X_2$  (МПа) получены следующие опытные данные:

$X_1$	3	4	5	6	6	10
$X_2$	260	60	150	150	240	400
$Y$	0,98	1,27	1,15	1,14	1,02	0,81

Провести регрессионный анализ модели  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$  на уровне значимости  $\alpha = 0,05$ .

1. Найдем вначале необходимые для расчета суммы. Объем выборки:

$$\begin{aligned}
 n &= 6; \sum x_{i1} = 3 + 4 + \dots + 10 = 34; \bar{x}_1 = 5,667; \\
 \sum x_{i2} &= 260 + 60 + \dots + 400 = 1260; \bar{x}_2 = 210 \\
 \sum y_i &= 0,98 + 1,27 + \dots + 0,81 = 6,37; \bar{y} = 1,062; \\
 \sum x_{i1}^2 &= 3^2 + 4^2 + \dots + 10^2 = 222 \\
 \sum x_{i2}^2 &= 260^2 + 60^2 + \dots + 400^2 = 333800 \\
 \sum y_i^2 &= 0,98^2 + \dots + 0,81^2 = 6,892 \\
 \sum x_{i1}x_{i2} &= 3 \cdot 260 + 4 \cdot 60 + \dots + 10 \cdot 400 = 8110 \\
 \sum x_{i1}y_i &= 3 \cdot 0,98 + \dots + 10 \cdot 0,81 = 3483 \\
 \sum x_{i2}y_i &= 260 \cdot 0,98 + \dots + 400 \cdot 0,81 = 1243,3 \\
 Q_{x1} &= \sum x_{i1}^2 - n\bar{x}_1^2 = 222 - 6 \cdot 5,667^2 = 29,311 \\
 Q_{x2} &= \sum x_{i2}^2 - n\bar{x}_2^2 = 333800 - 6 \cdot 210^2 = 69200 \\
 Q_y &= \sum y_i^2 - n\bar{y}^2 = 6,892 - 6 \cdot 1,062^2 = 0,125 \\
 Q_{x1x2} &= \sum x_{i1}x_{i2} - n\bar{x}_1\bar{x}_2 = 8110 - 6 \cdot 5,667 \cdot 210 = 969,58 \\
 Q_{x1y} &= \sum x_{i1}y_i - n\bar{x}_1\bar{y} = 3483 - 6 \cdot 5,667 \cdot 1,062 = -1,280 \\
 Q_{x2y} &= \sum x_{i2}y_i - n\bar{x}_2\bar{y} = 1243,3 - 6 \cdot 210 \cdot 1,062 = -94,82
 \end{aligned}$$

2. Определяем параметры регрессионной модели.

$$\begin{aligned}
 |X^T X| &= Q_{x1} Q_{x2} - Q_{x1x2}^2 = 29,311 \cdot 69200 - 969,58^2 = 1088236 \\
 \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} &= \begin{pmatrix} Q_{x2} & -Q_{x1x2} \\ -Q_{x1x2} & Q_{x1} \end{pmatrix} \begin{pmatrix} Q_{x1y} \\ Q_{x2y} \end{pmatrix} \frac{1}{|X^T X|} = \begin{pmatrix} 69200 & -969,58 \\ -969,58 & 29,311 \end{pmatrix} \begin{pmatrix} -1,280 \\ -94,82 \end{pmatrix} \frac{1}{1088236} = \begin{pmatrix} 0,0031 \\ -0,0014 \end{pmatrix}
 \end{aligned}$$

Искомая модель имеет вид:

$$Y = 1,138 + 0,0031x_1 - 0,0014x_2.$$

3. Проверим гипотезу  $H_0 : \beta_1 = \beta_2 = 0$ . Сумма квадратов, обусловленная регрессией

$$Q_R = (\tilde{\beta}_1 \tilde{\beta}_2) \begin{pmatrix} Q_{x1y} \\ Q_{x2y} \end{pmatrix} = (0,0031 - 0,0014) \begin{pmatrix} -1,280 \\ -94,82 \end{pmatrix} = 0,094$$

остаточная сумма квадратов

$$Q_e = Q_y - Q_R = 0,125 - 0,094 = 0,031$$

Выборочное значение статистики Фишера равно

$$F = \frac{\frac{1}{2} Q_R}{\frac{1}{3} Q_e} = \frac{\frac{1}{2} 0,094}{\frac{1}{3} 0,031} = 4,56$$

Критическое значение оказывается меньше выборочного

$$F_{1-\alpha}(k-1, n-k) = F_{0,95}(2, 3) = 9,55 > 4,56,$$

поэтому гипотеза  $H_0$ , принимается, это означает, что регрессоры  $X_1$  и  $X_2$  (относительное удлинение и предел прочности) не влияют существенно на отклик (обрабатываемость).

Если бы это влияние оказалось существенным, то далее потребовалась бы проверка значимости каждого регрессора.

## 5.5. Регрессионный анализ в Excel

### Парная регрессия

Для проведения регрессионного анализа в электронных таблицах имеется несколько различных средств. Во-первых, это встроенные статистические функции:

ОТРЕЗОК (для расчета коэффициента  $\beta_0$  в парной линейной регрессии, определяющего отрезок, отсекаемый линией регрессии по оси  $y$ ),

НАКЛОН (для расчета коэффициента  $\beta_1$  в парной линейной регрессии, определяющего наклон линии регрессии),

ЛИНЕЙН (для расчета множественной линейной регрессии),  
ТЕНДЕНЦИЯ (для прогноза по множественной линейной регрессии),  
ПРЕДСКАЗ (для прогноза по парной линейной регрессии),  
ЛГРФПРИБЛ (для расчета экспоненциальной регрессии)

$$y = \beta_0 \beta_1^{x_1} \beta_2^{x_2} \dots \beta_k^{x_k},$$

часто используемой в экономико-статистических расчетах, в частности, при анализе динамики различных явлений),

РОСТ (для прогноза по экспоненциальной регрессии) и другие.

Во-вторых, для построения парных регрессий можно использовать инструмент *Линия тренда*, позволяющий построить линейную и несколько видов нелинейной регрессии: рассчитать уравнение, коэффициент детерминации, построить графики, дать прогноз.

Наконец, для проведения регрессионного анализа удобен (особенно для множественной линейной регрессии) инструмент *Регрессия* из пакета *Анализ данных*.

Вначале рассмотрим технологию применения этого инструмента при проведении парного линейного регрессионного анализа.

Построим зависимость предела прочности прессованной детали от температуры при прессовании (см. пример 5.1).

Введите значения  $x$  и  $y$  в два столбца электронной таблицы и откройте окно *Регрессия* (рис. 5.2). При заполнении полей этого окна имеется возможность установить (при необходимости) константу  $\beta_0$ , равную нулю, изменить уровень значимости (по умолчанию уровень надежности 0,95 соответствует уровню значимости 0,05). При необходимости рассчитываются остатки или стандартизированные остатки. Могут быть выведены графики остатков, нормальной вероятности и график подбора: диаграмма рассеяния с нанесенной на нее расчетной линией регрессии.

Поставьте флажки для вывода остатков (при этом одновременно будут найдены и прогнозируемые значения отклика) и построения графика подбора.

На рис. 5.3 показаны результаты расчета. В таблице Регрессионная статистика приведены, в частности, коэффициент детерминации  $R$ -квадрат и стандартная ошибка, в таблице Дисперсионный анализ рассчитана статистика Фишера и приведено  $p$ -значение, определяющее значимость модели: регрессионная модель значима, если вероятность ошибки  $p$  меньше заданного уровня значимости (напомним, что по умолчанию оно равно 0,05). В таблице с коэффициентами модели приведены оценки  $\beta_0$  ( $Y$ -пересечение) и

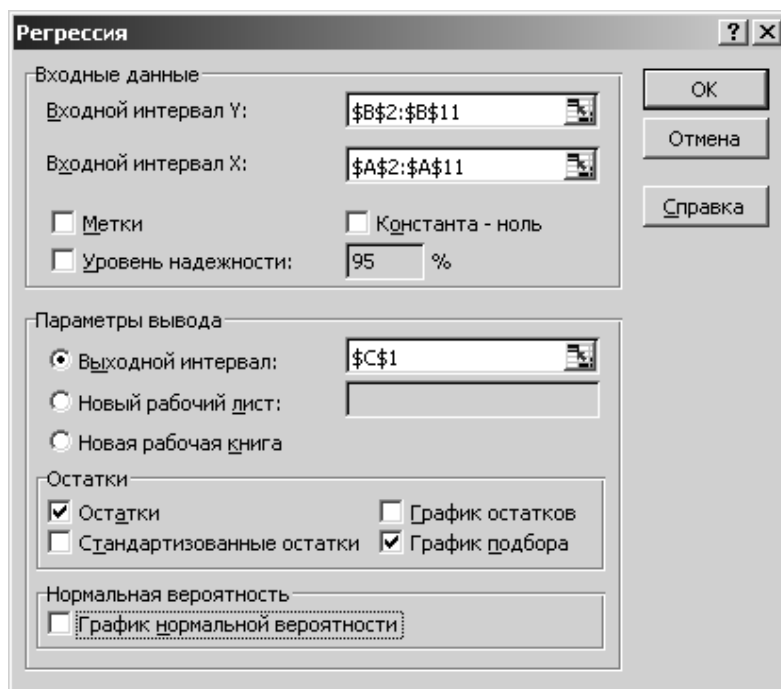


Рис. 5.2

$\beta_1$  (Переменная  $X_1$ ), их стандартные ошибки, значения статистик Стьюдента, их  $p$ -значения, доверительные интервалы. В таблице Вывод остатка, кроме остатков, приведены прогнозируемые (предсказанные) значения  $y$ .

Из этих таблиц следует, что искомая модель имеет вид

$$y = 178,108 - 0,568x_1;$$

$X$	$Y$	ВЫВОД ИТОГОВ	
120	110	Регрессионная статистика	
125	107	Множественный $R$	0,981
130	105	$R$ -квадрат	0,962
135	98	Нормир. $R$ -квадрат	0,957

140	100	Стандарт. ошибка		1,815			
145	95	Наблюдения		10			
150	95	Дисперсионный анализ					
155	92		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	Значимость
160	86	Регрессия	1	666,55	666,5	202,4	5,8E-07
165	83	Остаток	8	26,352	3,294		
		Итого	9	692,9			
		Коэфф.	<i>Ст. ошибка</i>	<i>t</i>	<i>P</i>	<i>Нижн.</i>	<i>Верхние 95</i>
Y-перес		178,109	5,7236	31,118	1E-09	164,9	191,308
Пер X 1		-0,568	0,04	-14,23	6E-07	-0,66	-0,4763
		ВЫВОД ОСТАТКА					
		<i>Наблюд.</i>	<i>Предсказанное Y</i>	<i>Остатки</i>			
		1	109,89	0,1091			
		2	107,05	-0,048			
		3	104,21	0,7939			
		4	101,36	-3,364			
		5	98,52	1,4788			
		6	95,68	-0,679			
		7	92,84	2,1636			
		8	89,99	2,0061			
		9	87,15	-1,152			
		10	84,31	-1,309			

Рис. 5.3

она значима, поскольку значимость  $p = 5,8 \cdot 10^{-7} \ll 0,05$ ; коэффициент детерминации  $R^2 = 0,962$ .

Рассмотрим теперь решение этой же задачи с использованием инструмента Линия тренда. По исходным данным, используя мастер диаграмм, постройте точечную диаграмму (рис. 5.4) и вызовите контекстное меню, щелкнув правой кнопкой мыши по одной из точек диаграммы.



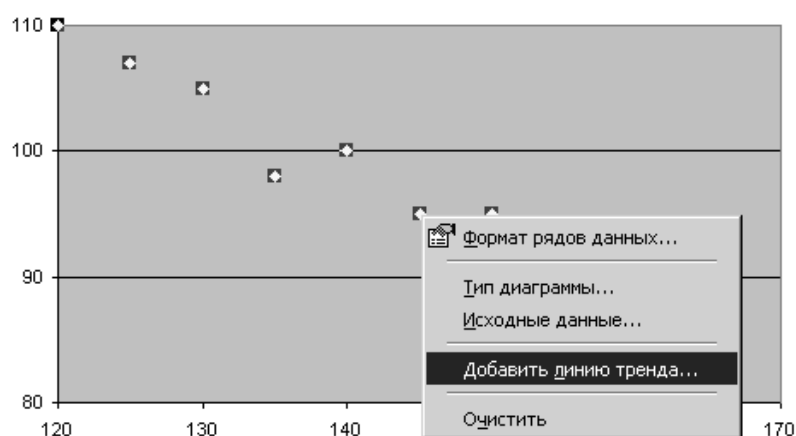


Рис. 5.4

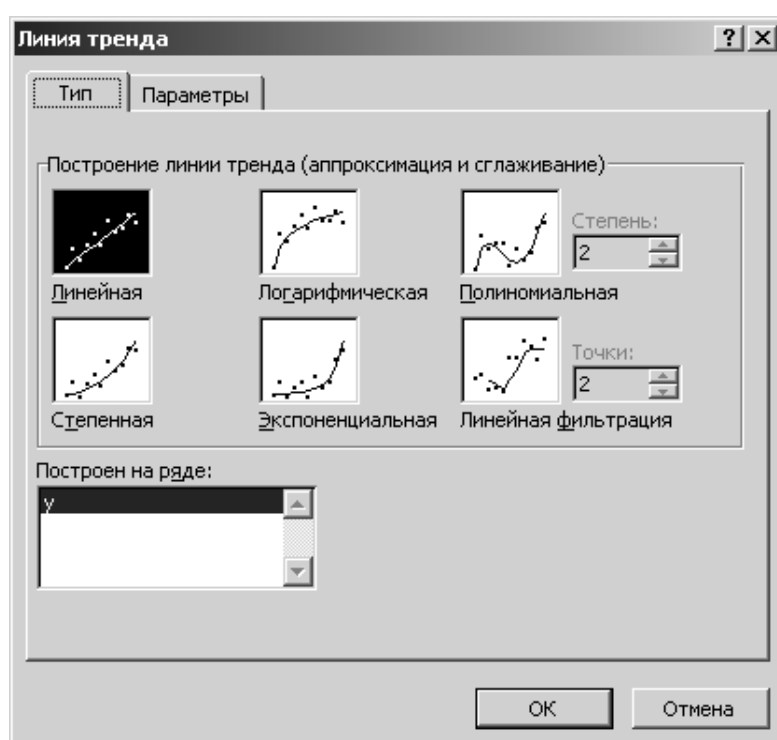


Рис. 5.5

Далее выбирается тип линии тренда (рис. 5.5) и устанавливаются параметры (рис. 5.6). При необходимости здесь же можно ввести наименование линии, сделать прогноз, установить на нулевое значение параметр  $\beta_0$ . На рис. 5.7 показан построенный график с уравнением модели и коэффициентом детерминации.

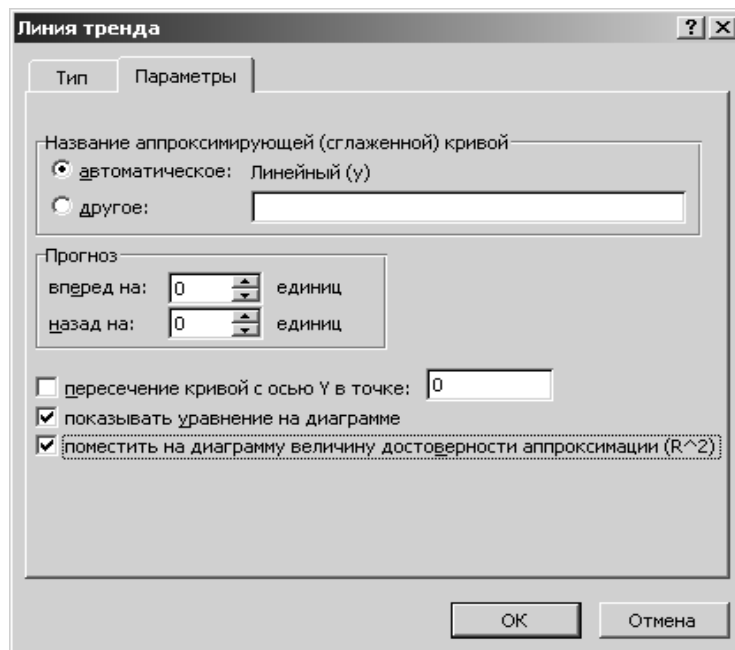


Рис. 5.6

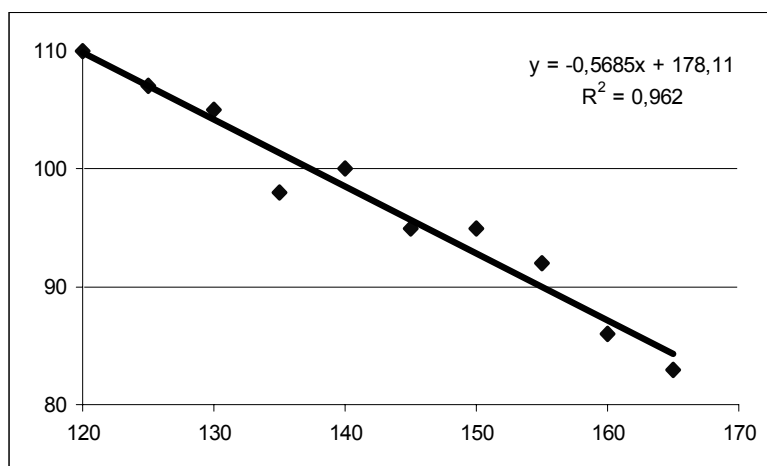


Рис. 5.7

Используя этот же метод, найдите самостоятельно зависимость давления в системе от времени выдержки (см. пример 5.3). Опробуйте не только те модели, которые были рассмотрены при выполнении примера, но и другие: экспоненциальную, полиномы различных степеней. Обратите внимание на то обстоятельство, что не любая из имеющихся моделей может быть выбрана. Почему? Выберите по возможности оптимальную модель: с достаточно высоким коэффициентом детерминации, но не слишком громоздкую (очевидно, что чем выше степень полинома, тем ближе кривая линия к опытным точкам).

В каком случае коэффициент детерминации точно равен единице? Поясните этот результат.

### Множественная регрессия

Изучалось влияние на влажность вафельного листа  $y$  времени выдержки листа в печи  $x_1$ , температуры печи  $x_2$  и влажности теста  $x_3$ . Проведено 20 наблюдений:

№	$y$	$x_1$	$x_2$	$x_3$	№	$y$	$x_1$	$x_2$	$x_3$
1	3,1	2,5	180	63	11	2,9	3	180	63
2	3,4	2,5	180	64	12	3,0	3	180	64
3	3,5	2,5	180	65	13	3,1	3	180	65
4	3,2	2,5	180	63	14	2,8	3	180	63
5	3,3	2,5	180	64	15	2,9	3	180	64
6	3,4	2,5	200	65	16	2,9	3	200	65
7	3,2	2,5	200	63	17	2,7	3	200	63
8	3,3	2,5	200	64	18	2,8	3	200	64
9	3,4	2,5	200	65	19	2,9	3	200	65
10	3,2	2,5	200	63	20	2,8	3	200	63

Требуется построить модель множественной линейной регрессии, предполагая наличие линейной связи между влажностью вафельного листа и тремя указанными факторами.

Введите исходные данные в столбцы. Воспользуйтесь инструментом *Регрессия* из пакета *Анализ данных*. При вводе входного интервала  $X$  выделите мышью все три столбца с независимыми переменными. Результаты расчета частично показаны на рис. 5.8. Полученная модель имеет вид:

$$y = -1,0506 - 0,84x_1 - 0,0041 x_2 + 0,1132x_3.$$

ВЫВОД ИТОГОВ					
Регрессионная статистика					
<i>R</i>	0,9716				
<i>R</i> -квад	0,9441				
Норм. <i>R</i>	0,9336				
Ст.ошиб.	0,0631				
Наблюд.	20				
Дисперсионный анализ					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	Значимость <i>F</i>
Регресс.	3	1,0744	0,35813	90,07	3,104E-10
Остаток	16	0,0636	0,00398		
Итого	19	1,138			

	Коэффициенты	<i>Ст. ошибка</i>	<i>t-стат.</i>	<i>P-Значение</i>
<i>Y</i> -пересечение	-1,0506	1,1045	-0,9512	0,35564
Переменная <i>X</i> 1	-0,84	0,0564	-14,894	8,5E-11
Переменная <i>X</i> 2	-0,0041	0,0014	-2,9095	0,01024
Переменная <i>X</i> 3	0,1132	0,0171	6,62251	5,9E-06

Рис. 5.8

Модель значима (см. проверку значимости по *F*-критерию), все факторы также значимы: это следует из того, что все *p*-значения для переменных меньше, чем 0,05.

Если бы некоторые из факторов (регрессоров) оказались незначимы, можно было бы попытаться построить новую модель, удалив их из нее.

Более корректно в этой ситуации воспользоваться пошаговой регрессией. Из-за отсутствия в Excel средств пошаговой регрессии, следует обратиться к системе Statistica или к пакету СПОР.

Решите еще одну задачу, приведенную в справке Excel.

Застройщик оценивает группу зданий в деловом районе. Его интересуют общая площадь здания  $x_1$ , количество офисов  $x_2$ , количество входов  $x_3$ , время эксплуатации здания  $x_4$ .

Наугад выбираются 11 зданий из 1500. Исходные данные приведены в таблице (0,5 входа означает вход только для доставки корреспонденции),  $y$  – цена здания в тыс. у. е.

№	$x_1$	$x_2$	$x_3$	$x_4$	$y$
1	2310	2	2	20	142
2	2333	2	2	12	144
3	2356	3	1,5	33	151
4	2379	3	2	43	150
5	2402	2	3	53	139
6	2425	4	2	23	169
7	2448	2	1,5	99	126
8	2471	2	2	34	142
9	2494	3	3	23	163
10	2517	4	4	55	169
11	2540	2	3	22	149

Предполагается наличие линейной связи между ценой и факторами.

Найдите коэффициенты модели. Проверьте значимость модели и факторов.

Модель имеет вид

$$y = 56,587 + 0,02556x_1 + 12,618x_2 + 2,709x_3 - 0,2318x_4.$$

Знак «минус» перед  $x_4$  означает, что с увеличением времени эксплуатации стоимость офиса снижается. Модель значима (по статистике Фишера), коэффициент детерминации достаточно высокий, все факторы значимы (по статистике Стьюдента).

Застройщик выбрал здание площадью 2500 кв. метров, с тремя офисами, двумя входами, время эксплуатации – 25 лет. Определим его оценочную стоимость по полученной модели:

$$y = 56,587 + 0,02556 \cdot 2500 + 12,618 \cdot 3 + 2,709 \cdot 2 - 0,2318 \cdot 25 = 158.$$

Таким образом, прогнозируемая стоимость здания составит 158 тыс. у. е.

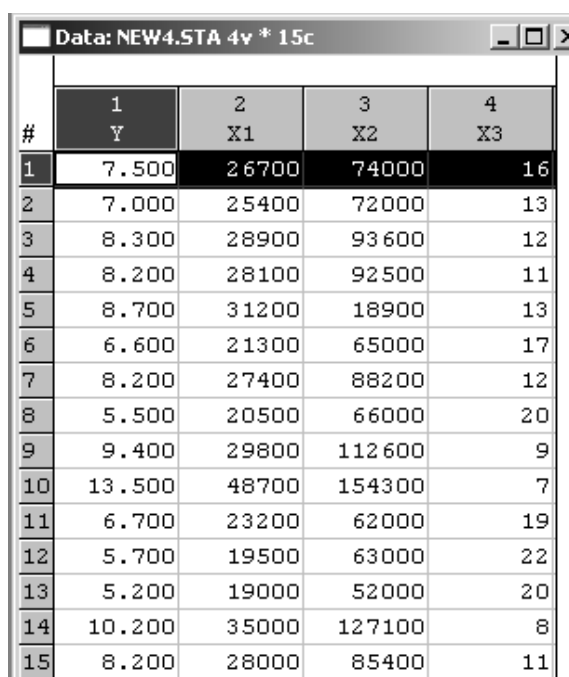
## 5.6. Регрессионный анализ в Statistica

Анализируется производительность труда на предприятиях  $Y$  в зависимости от численности работников  $X1$ , среднегодовой зарплаты  $X2$  и непроизводственных расходов  $X3$ . Собраны данные по 15 предприятиям.

Подготовьте файл исходных данных для построения линейной регрессионной модели, создав таблицу из 15 строк и 4 столбцов (рис. 5.9).

### Построение линейной модели

В переключателе модулей выберите модуль множественной линейной регрессии (*Multiple Regression*) и загрузите созданный файл.



#	1 Y	2 X1	3 X2	4 X3
1	7.500	26700	74000	16
2	7.000	25400	72000	13
3	8.300	28900	93600	12
4	8.200	28100	92500	11
5	8.700	31200	18900	13
6	6.600	21300	65000	17
7	8.200	27400	88200	12
8	5.500	20500	66000	20
9	9.400	29800	112600	9
10	13.500	48700	154300	7
11	6.700	23200	62000	19
12	5.700	19500	63000	22
13	5.200	19000	52000	20
14	10.200	35000	127100	8
15	8.200	28000	85400	11

Рис. 5.9

В окне *Multiple Regression* введите переменные *Variable* – зависимую *Dependent Y* и независимые *Independent X1, X2, X3*, установите *Input file* – *raw data* (исходные данные; возможна альтернатива – ввод корреляционной матрицы) и *Mode* – *standard* (обычная линейная модель; здесь возможно использование и фиксированной нелинейной модели – будет рассмотрено

далее; нелинейные модели произвольного вида могут быть построены с использованием модуля *Nonlinear Estimation*). После щелчка *OK* появляется окно *Model Definition* (рис. 5.10), с помощью которого можно при необходимости установить нулевое значение постоянной  $\beta_0$  (*Intercept – set to zero*), перейти к пошаговой регрессии (*Forward stepwise / Backward stepwise*) или построению гребневых оценок (*Ridge regression*).

Выберите *Method – standard* (стандартный метод); *Intercept – include in model* (свободный член  $\beta_0$  включить в модель). Практически сразу появляется окно с результатами расчета *Multiple Regression Results* (рис. 5.11).

**Model Definition**

**Variables**

Independent: X1-X3  
Dependent: Y

**Method:** Standard

**Intercept:** Include in model

**Tolerance:** .000100 (Enter 0.0 to set to minimum=1.e-25)

☐ **Ridge regression; lambda:** .100

**Stepwise Multiple Regression:**

**F to enter:** 1.00000 **F to remove:** 0.00000

**Number of steps:** 3

**Displaying results:** Summary only

☐ **Batch processing/printing**

☐ **Print residual analysis**

**Review Correlation matrix/means/SD**

**OK** **Cancel**

Рис. 5.10

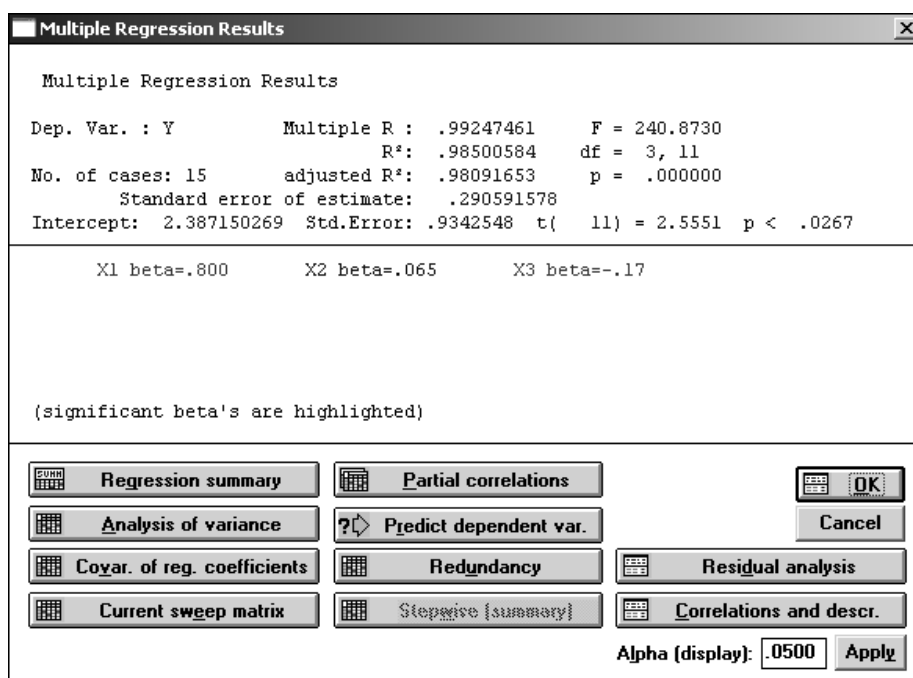


Рис. 5.11

В верхней (информационной) части окна приведена краткая сводка результатов. Наиболее важными являются  $F$ -статистика (используется для проверки значимости модели, модель считается значимой, если это значение превышает критическое; в пакете Statistica, как уже отмечалось ранее, для проверки значимости приводится  $p$ -значение – вероятность того, что модель незначима (как правило, модель считается значимой при  $p < 0.05$ ); коэффициент детерминации  $R^2$  (квадрат коэффициента корреляции между опытными и прогнозируемыми значениями; чем он ближе к единице, тем лучше модель соответствует опытными данным); значения параметров стандартизованной модели  $b_j$  – значения при значимых факторах (на заданном уровне значимости, обычно 0,05) выделены красным цветом.

Более подробные результаты можно получить с помощью кнопок. Щелкните по кнопке *Regression Summary*; здесь приведены коэффициенты ВЕТА и их стандартные ошибки для стандартизованных переменных, значения  $B$  ( $\beta_j$ ), их стандартные ошибки, значения  $t$ -статистик и  $p$ -значения, показывающие значимость каждого из факторов, используемых в модели (рис. 5.12).



Regression Summary for Dependent Variable: Y (new4.sta)						
R= .99247461 RI= .98500584 Adjusted RI= .98091653 F(3,11)=240.87 p<.00000 Std.Error of estimate: .29059						
N=15	BETA	St. Err. of BETA	B	St. Err. of B	t(11)	p-level
Intercpt			2.387150	.934255	2.55514	.026749
X1	.800153	.071822	.000226	.000020	11.14077	.000000
X2	.065413	.055954	.000004	.000004	1.16905	.267095
X3	-.165485	.069934	-.073739	.031162	-2.36630	.037396

Рис. 5.12

Analysis of Variance (new4.sta)					
Continue...	Sums of Squares	df	Mean Squares	F	p-level
Regress.	61.02045	3	20.34015	240.8730	.000000
Residual	.92888	11	.08444		
Total	61.94933				

Рис. 5.13

В окне *Analysys of variance* можно просмотреть результаты дисперсионного анализа регрессионной модели (рис. 5.13); щелкнув по кнопке *Correlations and descr.stats* можно просмотреть коэффициенты корреляции между переменными, их средние значения и стандартные отклонения; с использованием кнопки *Residual analysis* проводится подробный анализ остатков (расхождений между опытными и прогнозируемыми значениями). Опробуйте эти действия.

Полученная модель имеет вид  $Y = 2,387 + 0,000226X_1 + 0,00004X_2 - 0,07374X_3$ ; модель значима, два из трех факторов значимы (незначим фактор  $X_2$ , для него  $p = 0,267$ , что гораздо больше 0,05, при котором фактор считается значимым), коэффициент детерминации 0,985 (очень хорошее значение, практически равное единице).

### Построение нелинейных моделей

Часто используются модели мультипликативного типа (степенные) вида  $Y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} \dots x_k^{\beta_k}$ . Это одна из разновидностей фиксированных нелинейных моделей. Для линеаризации модели используется логарифмирование. Его можно провести, подготовив в таблице исходных данных столбцы с

логарифмами соответствующих величин. Однако при небольшом числе переменных можно использовать более удобные средства работы с моделями такого типа, встроенными в пакет. В окне *Multiple Regression* выберите *Mode – Fixed non-linear* (рис. 5.14). После щелчка *OK* в окне *Non-linear components Regression* станут доступны различные нелинейные преобразования переменных (рис. 5.15). Выберите *LN(X)* – в результате будут автоматически подсчитаны данные с натуральными логарифмами от всех выбранных переменных. Далее в окне *Model Definition* укажите – зависимая переменная *LN<sub>Y</sub>*, независимые *LN<sub>X1</sub> ... LN<sub>X3</sub>* (рис. 5.16).

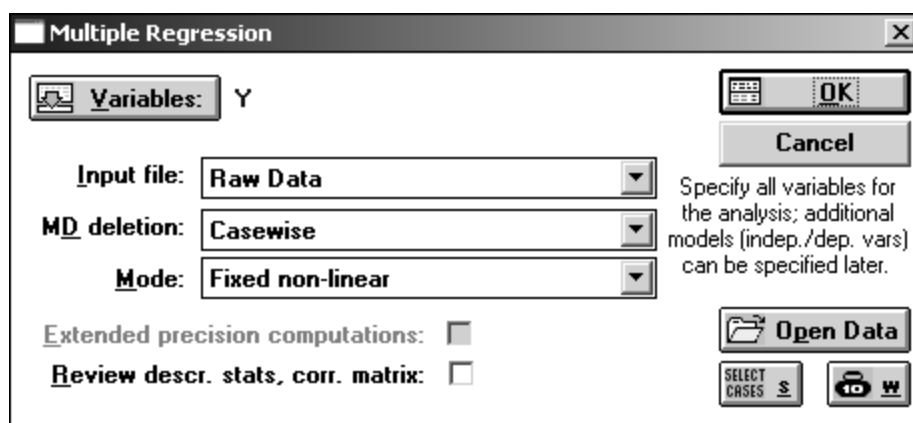


Рис. 5.14

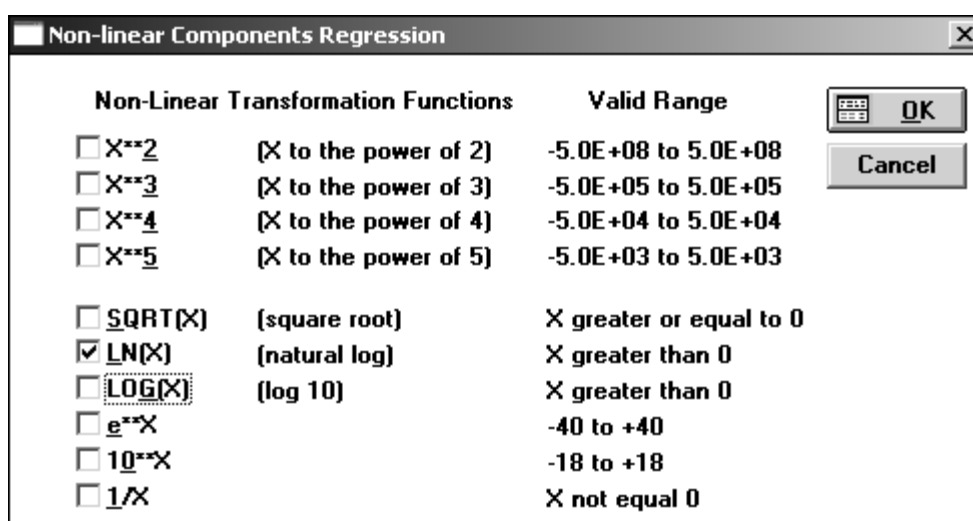


Рис. 5.15

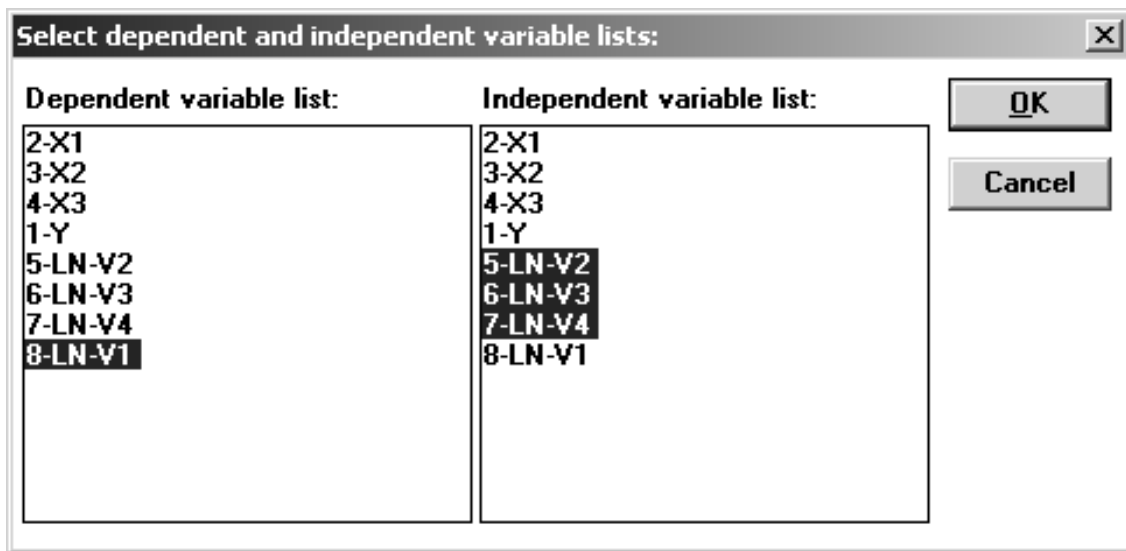


Рис. 5.16

Постройте модель. Проверьте ее значимость и значимость каждого из факторов. Лучше или хуже (по каким показателям) эта модель по сравнению с линейной?

Data: NEW4.STA 7v * 15c							
#	1 Y	2 X1	3 X2	4 X3	5 X1X2	6 X2X3	7 X3X1
1	7.500	26700	74000	16	19758E5	1184000	427200
2	7.000	25400	72000	13	18288E5	936000	330200
3	8.300	28900	93600	12	270504E4	1123200	346800
4	8.200	28100	92500	11	259925E4	1017500	309100
5	8.700	31200	18900	13	58968E4	245700	405600
6	6.600	21300	65000	17	13845E5	1105000	362100
7	8.200	27400	88200	12	241668E4	1058400	328800
8	5.500	20500	66000	20	1353E6	1320000	410000
9	9.400	29800	112600	9	335548E4	1013400	268200
10	13.500	48700	154300	7	751441E4	1080100	340900
11	6.700	23200	62000	19	14384E5	1178000	440800
12	5.700	19500	63000	22	12285E5	1386000	429000
13	5.200	19000	52000	20	988E6	1040000	380000
14	10.200	35000	127100	8	44485E5	1016800	280000
15	8.200	28000	85400	11	23912E5	939400	308000

Рис. 5.17

Опробуем теперь построение неполной квадратичной модели  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$ , учитывающей не только влияние

самих факторов, но и их парных взаимодействий. Для этого в таблицу исходных данных добавьте столбцы, значения элементов которых рассчитываются как попарные произведения факторов (например, в окне спецификации столбца  $X1X2$  введите формулу  $=X1*X2$  и щелкните по кнопке  $x=?$  на панели инструментов – пересчитать данные) (рис. 5.17).

Повторите регрессионный анализ, введя в качестве независимых переменных 6 регрессоров – три фактора и три взаимодействия. Дайте заключение о качестве модели.

### Пошаговая регрессия

Иногда, как в рассмотренном примере, некоторые регрессоры (или факторы) могут оказаться незначимыми. В такой ситуации используют пошаговый отбор регрессоров. При пошаговой регрессии с исключением в исходную модель вначале включаются все

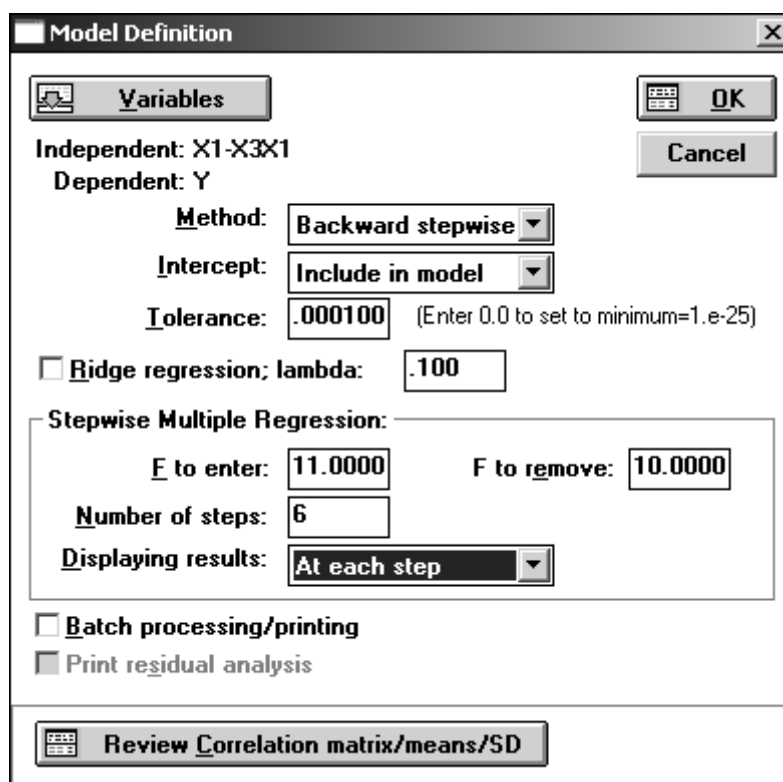


Рис. 5.18

регрессоры, и производится последовательное исключение тех из них, которые несут существенное влияние на отклик  $Y$ . При пошаговой регрессии с включением последовательно включаются в модель члены в порядке убывания их влияния

на отклик. Укажите в качестве независимых переменных все 6 регрессоров последней модели. Выберите вместо метода *Standard – Backward stepwise* (пошаговая регрессия с исключением) с отображением результатов на каждом шаге – *Displaying results – At each step*.

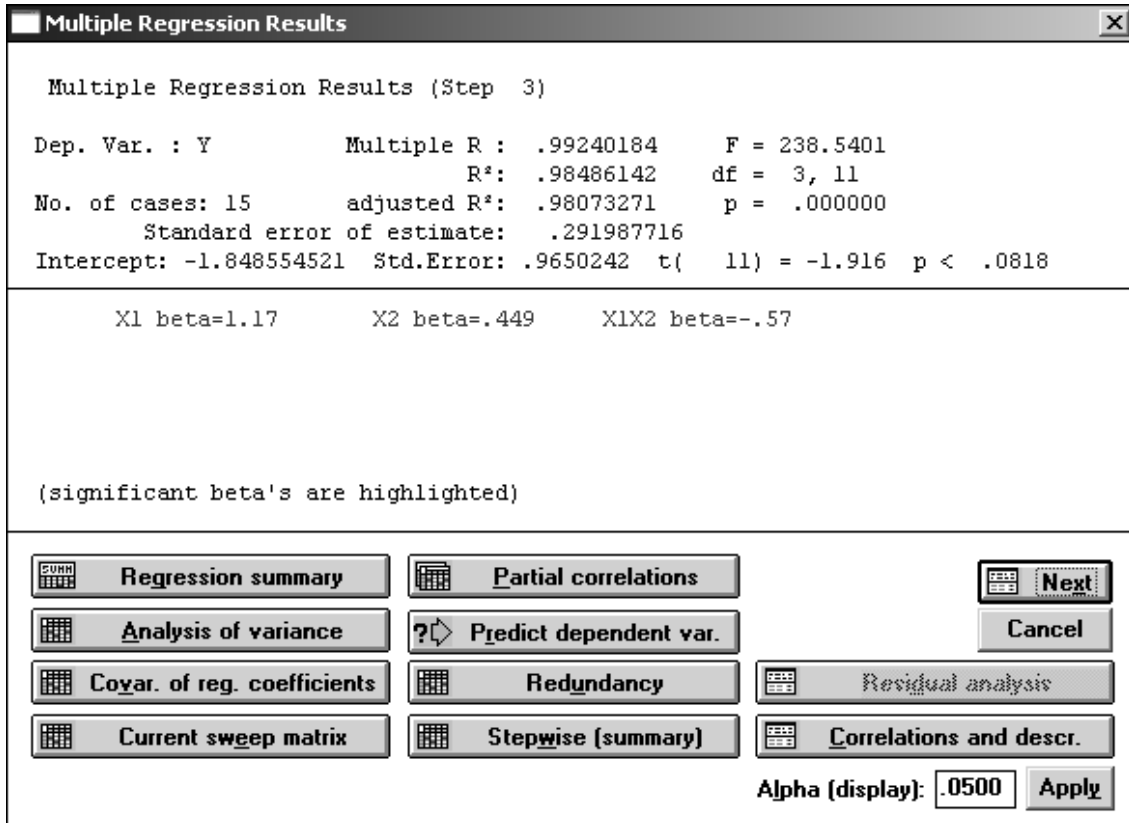


Рис. 5.19

Щелкая по кнопке *Next*, доведите процесс до значимости всех входящих в модель регрессоров. Окончательный результат просмотрите, щелкнув по кнопке *Regression Summary* (рис. 5.20). Сравните коэффициент детерминации этой модели с линейной. Проведите по аналогии пошаговую регрессию с включением *Forward stepwise*. Сравните полученные результаты.

Regression Summary for Dependent Variable: Y (new4.sta)						
Continue...						
R= .99240184 RI= .98486142 Adjusted RI= .98073271 F(3,11)=238.54 p<.00000 Std.Error of estimate: .29199						
N=15	BETA	St. Err. of BETA	B	St. Err. of B	t(11)	p-level
Intercpt			-1.84855	.965024	-1.91555	.081767
X1	1.165859	.121349	.00033	.000034	9.60750	.000001
X2	.449287	.158770	.00003	.000010	2.82979	.016378
X1X2	-.566871	.243021	-.00000	.000000	-2.33260	.039687

Рис. 5.20

При отсутствии пакета Statistica можно воспользоваться пакетом «Система поиска оптимальных регрессий» (СПОР) [6], включающем помимо рассмотренных методов множественной и пошаговой регрессии и более точную процедуру включения с исключением, а также ряд других процедур структурно-параметрической идентификации. СПОР имеет свою достаточно удобную библиотеку описаний (инструкций для пользователя), приведенных в [6].

## Контрольные вопросы

1. В чем заключается проверка значимости парной регрессионной модели?
2. Привести примеры адекватных и неадекватных моделей с иллюстрацией на графиках.
3. Используя нормальную систему, вывести уравнения для оценки параметров регрессии  $y = \beta_0 + \beta_1 x^3$ .
4. Преобразовать нелинейную по параметрам модель  $y = \beta x^{\beta_1}$  в линейную модель.
5. Сформулировать основные предположения регрессионного анализа.
6. Вывести формулы для определения параметров множественной регрессии (в матричном виде).
7. Как проверяется значимость регрессоров в множественном анализе?
8. Как вычислить среднеквадратичное отклонение  $s_j$  параметра множественной регрессии?

### Задание 1.

## ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Задание №1 включает одну комплексную задачу.

#### Условие задачи:

Для заданной выборки определить числовые характеристики (выборочное среднее, дисперсию смещенную и несмещенную, стандартное отклонение, коэффициенты асимметрии и эксцесса), построить графики выборочной функции распределения и гистограмму частот, приняв число интервалов равным 8; в предположении нормальности распределения данных построить 95% доверительный интервал для математического ожидания генеральной совокупности.

Варианты 1-10. По результатам механических испытаний партии стальных образцов получены значения предела прочности (в МПа):

854, 903, 872, 892\*, 933\*\*, 881, 919, 903, 868, 932, 904, 865, 897, 868\*, 905, 943\*\*, 901, 868\*, 947\*\*, 908, 895, 853, 893, 878, 862, 857, 928, 919\*, 925, 901, 911, 883\*, 947\*\*, 945, 881, 884, 939, 891, 885, 902, 938, 864\*, 904, 895, 872, 896\*, 878, 913, 875, 894, 878, 935, 878, 918, 891, 873\*.

(К значениям, отмеченным \*, прибавить  $3N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $V$ , где  $V$  – порядковый номер группы в потоке<sup>1</sup>).

Варианты 11-20. Износ режущего инструмента через определенное время обработки детали на станке составил (в мкм):

54\*\*, 103\*, 72, 92, 83, 81, 79, 53\*\*, 68, 82, 94, 65, 97, 110\*, 78, 82, 63, 101\*, 68, 87, 98, 95, 53\*\*, 93, 78, 62, 57, 88, 99, 105\*, 66, 73, 67, 101\*, 91, 83, 57, 55\*\*, 81,

---

<sup>1</sup> Значения  $V$  и  $N$  дает преподаватель.

83, 89, 91, 85, 102, 88, 108\*, 93, 58, 67, 104\*, 78, 85, 78, 85, 78, 108, 86, 91, 93, 88, 75, 68, 94, 115\*, 84, 101.

(От значений, отмеченных \*, отнять  $2N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $V$ , где  $V$  – порядковый номер группы в потоке<sup>2</sup>).

Варианты 21-30. По результатам контроля партии штампованных деталей получены следующие значения длины (в мм):

204, 196, 202, 203, 210\*, 201, 199, 203, 198, 202, 195, 205, 208\*, 194, 195, 202, 203, 207\*, 200, 199, 201, 198, 197, 198, 195, 203, 209\*, 203, 202, 197, 198, 199, 215\*, 201, 201, 203, 197, 145, 201, 204, 199, 209\*, 205, 201, 204, 199, 201, 212\*, 202, 198, 197, 204, 205, 202, 196, 197, 214\*, 206.

(От значений, отмеченных \*, отнять  $N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $V$ , где  $V$  – порядковый номер группы в потоке<sup>2</sup>).

---

<sup>2</sup> Значения  $V$  и  $N$  дает преподаватель.



## Задание 2.

# МЕТОДЫ АНАЛИЗА ДАННЫХ

Задание №2 включает три задачи.

**Условие задачи №1 (дисперсионный анализ):**

Варианты 1-10. Менеджер по продажам в сети супермаркетов хочет знать, влияет ли расположение рекламных щитов на объем продаж товара. Для каждого из трех видов щитов отобрано случайным образом по 6 магазинов, расположенных в соответствующем районе. Объемы продаж за месяц (млн руб.) приведены в таблице.

Расположение щитов		
1	2	3
8,1*	3,9	4,8
7,5	4,4	6,0*
5,8	2,8**	4,4**
6,6	3,4	5,8
5,9	4,8*	6,2*
4,8**	3,6	4,9

(От значений, отмеченных \*, отнять  $0,1N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $V$ , где  $V$  – порядковый номер группы в потоке<sup>3</sup>).

Варианты 11-20. Компания, производящая спортивные товары, желает сравнить расстояние, которое пролетают мячи, изготовленные по 4 разным технологиям. По каждой технологии произведено по 10 мячей. Мячи переданы для испытания в спортивный клуб, где испытаны в течение короткого

---

<sup>3</sup> Значения  $V$  и  $N$  дает преподаватель.

промежутка времени при одинаковых погодных условиях. Результаты испытаний в м.:

Технология			
1	2	3	4
206**	203**	217	213
226*	223	230	231
208	206	221	221
224*	223	227	222
206	205	218	229
229*	234*	231*	235*
204	204	224	213
228*	219	225	228
209	210	211**	214
221	233	229	225

(От значений, отмеченных \*, отнять  $N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $V$ , где  $V$  – порядковый номер группы в потоке <sup>4</sup>).

Варианты 21-30. Проверить влияние на износостойкость детали материала (три вида), из которого она изготовлена. Получены данные по износостойкости пяти деталей для каждого материала: время работы детали до износа, тыс. час.

Материал 1	1,25	1,32**	1,28	1,26	1,29
Материал 2	1,12*	1,15*	1,26	1,19	1,21
Материал 3	1,32	1,33**	1,34**	1,29	1,30

<sup>4</sup> Значения  $V$  и  $N$  дает преподаватель.

(От значений, отмеченных \*, отнять  $0,01N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $0,02V$ , где  $V$  – порядковый номер группы в потоке<sup>5</sup>).

### Условие задачи №2 (парная регрессия):

Для заданной выборки возможно применение линейной или параболической парной регрессионной модели. Построить обе модели и определить, какая из них лучше аппроксимирует опытные данные. В качестве критерия качества модели использовать коэффициент детерминации. На диаграмме рассеивания показать линии, соответствующие построенным моделям.

Варианты 1-10. Установить связь между максимальным напряжением изгиба в зубчатом колесе  $x$  (МПа) и числом циклов  $y$  (тыс. циклов) до разрушения:

$x$	900	850	800	750	700	650	600	550	500	450	400
$y$	62**	64**	70	81	94	111	120	212	347*	542*	1230*

(От значений, отмеченных \*, отнять  $10N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $V$ , где  $V$  – порядковый номер группы в потоке<sup>5</sup>).

Варианты 11-20. За каждым из 9 менеджеров по сбыту закреплена определенная территория. В таблице приведены численность населения на этой территории  $x$  в тыс. чел. и объемы продаж, обеспеченные соответствующим менеджером,  $y$  в млн руб.

$x$	4,96	8,26	9,09	12,25*	4,73	13,68*	3,58	2,77**	4,64
$y$	2,69**	3,54	3,32	3,54	2,25	5,15	2,02	1,71	3,26

(От значений, отмеченных \*, отнять  $0,1N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $0,04V$ , где  $V$  – порядковый номер группы в потоке<sup>6</sup>).

<sup>5</sup> Значения  $V$  и  $N$  дает преподаватель.

Варианты 21-30. В таблице приведены данные о величине списка почтовой рассылки  $x$  в тыс. фамилий и объеме продаж  $y$  в тыс. у.е. по группе каталогов.

$x$	168	21	94	39	249	43	589	41
$y$	5200	2400	3600	2000	7300	2500	15700	2500

(От значений, отмеченных \*, отнять  $0,1N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $0,04V$ , где  $V$  – порядковый номер группы в потоке<sup>6</sup>).

**Условие задачи №3 (множественная регрессия):**

Для заданной выборки провести множественный регрессионный анализ модели  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ .

Варианты 1-10. Дана таблица экспериментальных данных зависимости производительности  $y$  выпуска колец подшипников (тыс. шт.) от содержания механических примесей  $x_1$  (мг/л) соды  $x_2$  (г/л) и нитрата натрия  $x_3$  (г/л) в смазочно-охлаждающей жидкости, используемой в процессе шлифования колец.

$x_1$	309	220	90**	100	156	110
$x_2$	1.8*	4.0	5.6	5.1*	7.5	6.9
$x_3$	1.8	4.0*	5.6*	5.1	6.6	7.6
$y$	61	54	65	53	56	54
$x_1$	140	200	135	46**	40**	32**
$x_2$	6.5	6.4	6.7	6.9	8.5	7.5
$x_3$	8.0	9.2	8.3*	1.5*	1.9*	2.0
$y$	57	70	82	57	51	68

---

<sup>6</sup> Значения  $V$  и  $N$  дает преподаватель.

(От значений, отмеченных \*, отнять  $0,1N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $10V$ , где  $V$  – порядковый номер группы в потоке<sup>7</sup>).

Варианты 11-20. Анализируется зависимость урожайности зерновых культур  $y$  от количества используемых тракторов  $x_1$ , комбайнов  $x_2$  и расхода удобрений  $x_3$ . Приведены данные по 18 хозяйствам.

$x_1$	16	4	25	48*	21	21	7	4	5
$x_2$	25	27	29	39	26	30	28	26	24
$x_3$	3,1**	5,8	3,2	4,2	3,9	3,3	4,2	2,3	2,0**
$y$	98	85	91	99	97	87	125*	77	70
$x_1$	33	17	24	93*	17	6	3	15*	1
$x_2$	32	31	32	41	27	29	25	29	20
$x_3$	12,2	7,2	2,6	4,1	8,5	1,2	0,9**	2,1	4,2
$y$	140*	98	109	119*	98	71	73	84	84

(От значений, отмеченных \*, отнять  $0,1N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $V$ , где  $V$  – порядковый номер группы в потоке<sup>7</sup>).

Варианты 21-30. Предполагается, что зарплата работников предприятия  $y$  определяется их стажем работы  $x_1$ , продолжительностью обучения  $x_2$  и возрастом  $x_3$ . Собранные данные по 21 работнику представлены в таблице.

(От значений, отмеченных \*, отнять  $0,1N$ , где  $N$  – порядковый номер студента в группе (вариант); отмеченным \*\* – прибавить  $V$ , где  $V$  – порядковый номер группы в потоке<sup>7</sup>).

---

<sup>7</sup> Значения  $V$  и  $N$  дает преподаватель.

$x_1$	21	6	18	15	16	9	5
$x_2$	10	9	9	11	11	11	12
$x_3$	48	26	35	51	44	32	23
$y$	18900	15100	22100	21500	22400	17100	17600
$x_1$	10	8	17	6	10	10	3
$x_2$	12	12	13	13	13	14	14
$x_3$	33	28	39	26	46	38	26
$y$	15200	19300	26500	12700	24400	21000	17100
$x_1$	8	6	4	7	4	3	3
$x_2$	14	15	16	16	17	18	18
$x_3$	32	47	28	34	26	32	34
$y$	23500	23100	21000	28100	23800	25900	25800

# Образец оформления задания

---

Федеральное агентство по образованию  
Ульяновский государственный технический университет  
Кафедра «Прикладная математика и информатика»

ИНДИВИДУАЛЬНЫЙ ТИПОВОЙ РАСЧЕТ №1  
по дисциплине  
«Теория вероятностей и математическая статистика»

Выполнил:  
студент группы ПМд-21  
И. И. Иванов  
(вариант  $N = 11$ ,  $V = 2$ )

Проверил:  
П. П. Петров

Ульяновск  
2009

Условие задачи:

Для заданной выборки определить числовые характеристики (выборочное среднее, дисперсию смещенную и несмещенную, стандартное отклонение, коэффициенты асимметрии и эксцесса), построить графики выборочной функции распределения и гистограмму частот, приняв число интервалов, равным 8; в предположении нормальности распределения данных построить 95% доверительный интервал для математического ожидания генеральной совокупности.

Износ режущего инструмента через определенное время обработки детали на станке составил (в мкм):

56, 81, 72, 92, 83, 81, 79, 55, 68, 82, 94, 65, 97, 88, 78, 82, 63, 79, 68, 87, 98, 95, 55, 93, 78, 62, 57, 88, 99, 83, 66, 73, 67, 79, 91, 83, 57, 57, 81, 83, 89, 91, 85, 102, 88, 86, 93, 58, 67, 82, 78, 85, 78, 85, 78, 108, 86, 91, 93, 88, 75, 68, 94, 93, 84, 101.

Решение:

А. Расчет с использованием калькулятора.

Объем выборки  $n = 66$ .

Выборочное среднее

$$\bar{x} = \frac{1}{n} \sum x_i = \frac{1}{66} (56 + 81 + \dots + 84 + 101) = \dots;$$

выборочная дисперсия

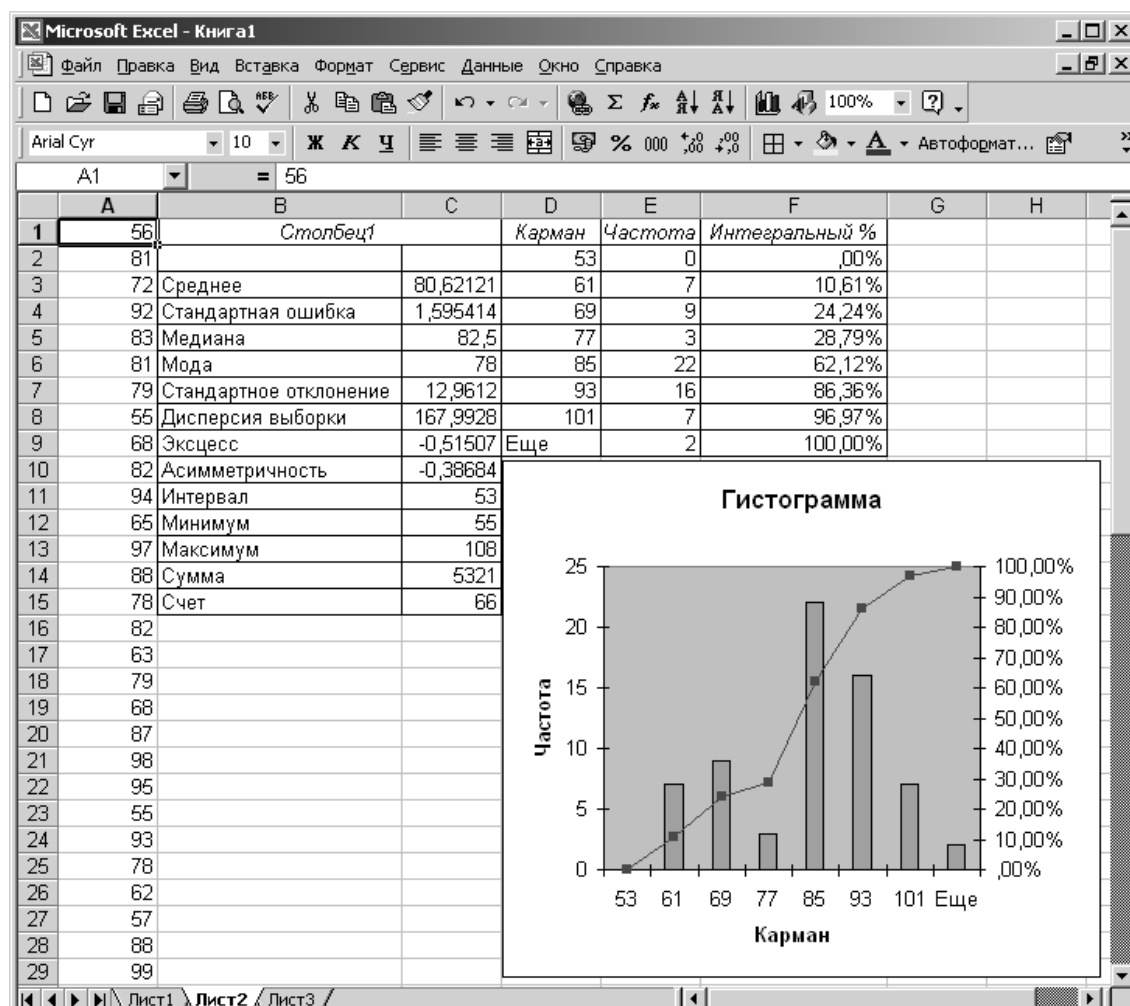
$$D_x^* = \frac{1}{n} \sum x_i^2 - \bar{x}^2 = \frac{1}{66} (56^2 + 81^2 + \dots + 84^2 + 101^2) = \dots;$$

...

(все этапы расчета выполняются с использованием микрокалькулятора).



Б. Расчет с использованием электронных таблиц Excel (или системы Statistica).



(Копируются экранные формы со всеми расчетами и графиками).

## Список использованной литературы

---

1. Айвазян, С. А. Прикладная статистика и основы эконометрики / С. А. Айвазян, В. С. Мхитарян. – М. : ЮНИТИ, 1998. – 1022 с.
2. Болотин, В. В. Применение методов теории вероятностей и теории надежности в расчетах сооружений / В. В. Болотин. – М. : Стройиздат, 1971. – 255 с.
3. Боровиков, В. Statistica: Искусство анализа данных на компьютере / В. Боровиков. – СПб : Питер, 2001. – 656 с.
4. Боровиков, В. П. Прогнозирование в системе Statistica в среде Windows / В. П. Боровиков, Г. И. Ивченко. – М. : Финансы и статистика, 1999. – 384 с.
5. Валеев, С. Г. Регрессионное моделирование при обработке наблюдений / С. Г. Валеев. – М. : Наука, 1991. – 272 с. (2-е изд. : Валеев, С. Г. Регрессионное моделирование при обработке данных / С. Г. Валеев. – Казань : ФЭН, 2001. – 296 с.).
6. Валеев, С. Г. Система поиска оптимальных регрессий / С. Г. Валеев, Г. Р. Кадырова. – Казань : ФЭН, 2003. – 160 с.
7. Валеев, С. Г. Прикладная статистика. Методические указания к типовым расчетам / С. Г. Валеев, В. Н. Клячкин. – Ульяновск : УлПИ, 1992. – 56 с.
8. Вуколов, Э. А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов Statistica и Excel / Э. А. Вуколов. – М. : ИНФРА-М, 2004. – 464 с.
9. Дубров, А. М. Математико-статистический анализ на программируемых микрокалькуляторах / А. М. Дубров, В. С. Мхитарян, Л. И. Трошин. – М. : Финансы и статистика, 1991. – 176 с.
10. Калихман, И. Л. Вероятность и статистика / И. Л. Калихман, Е. М. Четыркин. – М. : Финансы и статистика, 1982. – 320 с.
11. Кацев, П. Г. Статистические методы исследования режущего инструмента / П. Г. Кацев. – М. : Машиностроение, 1974. – 231 с.
12. Клячкин, В. Н. Статистические методы в управлении качеством: компьютерные технологии / В. Н. Клячкин. – М. : Финансы и статистика, 2007. – 304 с.
13. Левин, Д. Статистика для менеджеров с использованием Excel / Д. Левин, Д. Стефан, Т. Кребиль. – М. : Вильямс, 2004. – 1312 с.
14. Макарова, Н. В. Статистика в Excel / Н. В. Макарова, В. Я. Трофимец. – М. : Финансы и статистика, 2002. – 368 с.
15. Макаров А. А. Анализ данных на компьютере / Ю. Н. Тюрин, А. А. Макаров. – М. : ИНФРА-М, Финансы и статистика, 1995. – 384 с.
16. Сборник задач по математике для вузов. Часть 3. Теория вероятностей и математическая статистика / Под ред. А. В. Ефимова. – М. : Наука, 1990. – 428 с.

17. Сигел, Э. Практическая бизнес-статистика / Э. Сигел,. – М. : Вильямс, 2004. – 1056 с.
18. Солонин, И. С. Математическая статистика в технологии Машиностроения / И. С. Солонин. – М. : Машиностроение, 1972. – 216 с.
19. Степнов, М. Н. Статистические методы обработки результатов механических испытаний / М. Н. Степнов. М. : Машиностроение, 1985. – 232 с.
20. Шеффе, Г. Дисперсионный анализ / Г. Шеффе. – М. : Наука, 1980. – 512 с.

<b>Глава 1. Описательная статистика</b>	3
1.1. Способы представления выборки	3
1.2. Числовые характеристики выборки	5
1.3. Пример расчета	7
1.4. Описательная статистика в Excel	10
1.5. Описательная статистика в Statistica	15
Контрольные вопросы	20
<b>Глава 2. Оценка параметров и проверка гипотез</b>	22
2.1. Точечные оценки параметров	22
2.2. Интервальные оценки	24
2.3. Проверка параметрических гипотез	28
2.4. Критерии согласия	31
2.5. Примеры расчета	33
2.6. Проверка гипотез в Excel	35
2.7. Оценка параметров и проверка гипотез в Statistica	42
Контрольные вопросы	44
<b>Глава 3. Дисперсионный анализ</b>	46
3.1. Однофакторный дисперсионный анализ	46
3.2. Многофакторный дисперсионный анализ	49
3.3. Примеры расчета	53
3.4. Дисперсионный анализ в Excel	56
3.5. Дисперсионный анализ в Statistica	58
Контрольные вопросы	59
<b>Глава 4. Корреляционный анализ</b>	61
4.1. Коэффициент корреляции	61
4.2. Проверка значимости корреляции	63
4.3. Множественная корреляция	64
4.4. Примеры расчета	67
4.5. Корреляционный анализ в Excel	72
4.6. Корреляционный анализ в Statistica	74
Контрольные вопросы	75
<b>Глава 5. Регрессионный анализ</b>	76
5.1. Парная линейная регрессия	76
5.2. Парная нелинейная регрессия	79
5.3. Множественная регрессия	82
5.4. Примеры расчета	86
5.5. Регрессионный анализ в Excel	93
5.6. Регрессионный анализ в Statistica	102
Контрольные вопросы	110
<b>Варианты индивидуальных заданий</b>	111
Задание №1. Описательная статистика	111
Задание №2. Методы анализа данных	113
Образец оформления задания	119
Список использованной литературы	122