Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

Ю.М. Жученко

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ИНФОРМАЦИИ С ПРИМЕНЕНИЕМ ПЕРСОНАЛЬНЫХ КОМПЬЮТЕРОВ ПРАКТИЧЕСКОЕ ПОСОБИЕ

Гомель 2007

Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

Ю.М. Жученко

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ИНФОРМАЦИИ С ПРИМЕНЕНИЕМ ПЕРСОНАЛЬНЫХ КОМПЬЮТЕРОВ

ПРАКТИЧЕСКОЕ ПОСОБИЕ

для студентов IV курса

специальность 1-31 01 01 02 “Биология (научно-педагогическая деятельность)”

Гомель 2007

УДК 519.22 : 517.97 (075.8)
ББК 22.172 + 22.161. 8 я73
Ж 94

Рецензенты:
А.С. Неверов, профессор, доктор технических наук**;**
заведующий кафедрой химии учреждения образования
«Белорусский государственный университет транспорта»

А.М. Дворник, профессор, доктор биологических наук;
кафедра физиологии человека и животных учреждения образования
«Гомельский государственный университет
имени Франциска Скорины»

Жученко Ю.М.

Ж 94 Статистическая обработка информации с применением персональных компьютеров [Текст] : [практическое пособие для студентов IV курса специальности 1-31 01 01 02 “Биология (научно-педагогическая деятельность)”] / Ю. М. Жученко; М-во образования РБ, Гомельский государственный университет им. Ф. Скорины. – Гомель : ГГУ им Ф. Скорины, 2007.–105 с.

ISBN

Целью подготовки издания практического пособия является оказание помощи студентам в усвоении основ курса вариационной статистики и обработки результатов экспериментов с применением табличного редактора EXCEL и пакета статистического анализа STATISTCA 6.0.

Курс лекций адресован студентам специальности 1-31 01 01 02 “Биология (научно-педагогическая деятельность)”

УДК 519.22 : 517.97 (075.8)
ББК 22.172 + 22.161. 8 я73

ISBN

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 6
СИСТЕМА ТАБЛИЧНЫЙ РЕДАКТОР EXCEL 8
ТЕМА 1 ОПИСАТЕЛЬНАЯ СТАТИСТИКА 8
1.1 Гистограмма 10
ТЕМА 2 КОРРЕЛЯЦИЯ 11
ТЕМА 3 РЕГРЕССИОННЫЙ АНАЛИЗ 12
ТЕМА 4 ДИСПЕРСИОННЫЙ АНАЛИЗ 15
4.1 Однофакторный дисперсионный анализ 15
4.2 Двухфакторный дисперсионный анализ с повторениями 16
СИСТЕМА STATISTICA 6 18
ТЕМА 5 ПЕРВИЧНЫЙ АНАЛИЗ СТАТИСТИЧЕСКИХ ДАННЫХ В СИСТЕМЕ STATISTICA — МОДУЛЬ BASIC STATISTICS/TABLES (ОСНОВНЫЕ СТАТИСТИКИ/ТАБЛИЦЫ) 18
5.1 Вероятностный калькулятор 18
5.2 Нормальное распределение 21
5.3 Правила 2 и 3 сигма 24
5.4 Распределение хи-квадрат 26
5.5 t-распределение Стьюдента 27
5.6 F-распределение 29
5.7 Логарифмически-нормальное распределение 30
ТЕМА 6 БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ И ИГРОВЫЕ ЗАДАЧИ 32
ТЕМА 7 ОПИСАТЕЛЬНАЯ СТАТИСТИКА. ПЕРВИЧНЫЙ АНАЛИЗ СТАТИСТИЧЕСКИХ ДАННЫХ В STATISTICA6 39
ТЕМА 8 КОРРЕЛЯЦИОННЫЙ АНАЛИЗ В STATISTICA6 43
ТЕМА 9 РЕГРЕССИОННЫЙ АНАЛИЗ В СИСТЕМЕ STATISTICA — МОДУЛЬ MULTIPLE REGRESSION (МНОЖЕСТВЕННАЯ РЕГРЕССИЯ) 50

9.1 Описание модели 50
9.2 Постановка задачи 51
ТЕМА 10 ДИСПЕРСИОННЫЙ АНАЛИЗ 60
10.1 Однофакторный дисперсионный анализ 63
10.2 Двухфакторный анализ 69
ТЕМА 11 КЛАССИФИКАЦИЯ ДАННЫХ В СИСТЕМЕ STATISTICA. МОДУЛЬ DISCRIMINANT ANALYSIS (ДИСКРИМИНАНТНЫЙ АНАЛИЗ) 75
11.1 Постановка задачи, методы решения, ограничения 76
11.2 Предположения и ограничения 78
11.3 Классификация цветов ириса 79
ТЕМА 12 КЛАСТЕРНЫЙ АНАЛИЗ (НА ПРИМЕРЕ АВТОМОБИЛЕЙ РАЗНЫХ МАРОК) 88
12.1 Запуск модуля Кластерный анализ 93
12.2 Открытие файла данных 94
12.3 Выбор метода 96
12.4 Выбор переменных, установка начальных значений, запуск вычислительной процедуры метода k-средних 97
12.5 Просмотр результатов кластеризации 98
ЛИТЕРАТУРА 104

ВВЕДЕНИЕ

Современная биология не может развиваться без применения основ математической статистики. Математика требуется, прежде всего, при описании биологических множеств, популяций, штаммов, сортов, пород, линий, посевов, стад, подопытных групп. Математические методы необходимы для исчерпывающего извлечения информации о типичных объектах, их разнообразии, структуре этого разнообразия, о системах биологических взаимоотношений и взаимодействий, о разных биоценозах, о влияниях разных факторов на биологические объекты, развивающиеся в различных условиях.

Некоторые биологические вопросы не могут быть решены без применения специальных математических методов. К таким вопросам относятся сравнение выборочных групп по изучаемым показателям и определение достоверности результатов такого сравнения с заданной вероятностью безошибочных прогнозов, определение достаточной численности подопытных объектов, измерение силы влияния различных факторов на биологические процессы и явления и т.д. Современный этап развития науки характеризуется широким применением средств вычислительной техники. Применительно к современным персональным компьютерам в арсенале математической обработки информации пользователем существует огромное количество программных продуктов. Среди них важное место занимают табличный процессор Microsoft Excel и программа STATISTICA 6 для статистического анализа данных в среде Windows.

Представленное практическое пособие является знакомством с указанными программами, а материал расположен таким образом, что вы можете повторить все описанные действия вслед за нами на своем компьютере. Упражнения и задачи для самостоятельной работы позволят углубить понимание разбираемых примеров. Ранее каждый шаг

исследований, начиная от представления данных, перевода их в нужный формат, проверки, группировки, сортировки, графической интерпретации, подготовки программ обработки до просмотра результатов, был трудной задачей. Теперь достаточно двух-трех щелчков мыши, чтобы огромные объемы данных чрезвычайно быстро преобразовались, обработались и появились на экране в виде графиков, диаграмм и таблиц.

На простых и доступных примерах, взятых из различных сфер жизни, показаны возможности этих систем по статистической обработке данных (описательная статистика, корреляция и регрессия, дискриминантный анализ и др.).

В основу пособия положены материалы, опубликованные в работах (1-6) и адаптированы к современным версиям табличного процессора Microsoft Excel и пакета статистического анализа STATISTCA 6.0.

Практическое пособие может быть адресовано не только студентам специальности 1-31 01 01 02 “Биология (научно-педагогическая деятельность)”, но и самому широкому кругу читателей, работающих на персональных компьютерах.

СИСТЕМА ТАБЛИЧНЫЙ РЕДАКТОР EXCEL

ТЕМА 1 ОПИСАТЕЛЬНАЯ СТАТИСТИКА

1.1 Первичный анализ статистических данных в EXCEL

1.2 Гистограмма

1.3 Первичный анализ статистических данных в EXCEL

Формат представления исходных данных для выполнения первичного статистического анализа выполняется записью информации в одну строку или один столбец (в примере исходные данные и карманы в табличном редакторе EXCEL представлены столбцами).

Таблица 1.1. Исходные данные

x_i	7,1	7,7	3,6	8,3	8,8	10,4	8,9	9	8,9	14	9,7	9,4	8,5	15,9	12,6
	9,1	6,2	10,7	13,8	13,6	15,2	3,4	9,3	13,3	6,7	7,9	4,9	4,5	8	17,1

Для корректной статистической обработки необходимо определить величину класса (кармана) по формуле:

$k=(X_{max}-X_{min})/(1+3,322*\lg N)$, где N–число наблюдений. В данном случае $k=(17,1-3,4)/(1+3,322*\lg 30)=2,32$. (Табл.1.2)

Табл.1.2

Карманы	3.4	5.7	8.0	10.4	12.7	15.0	17.3

Шаг 1. В системе **Excel** в меню **Сервис** открыть модуль **Анализ данных** (рис. 1.1).

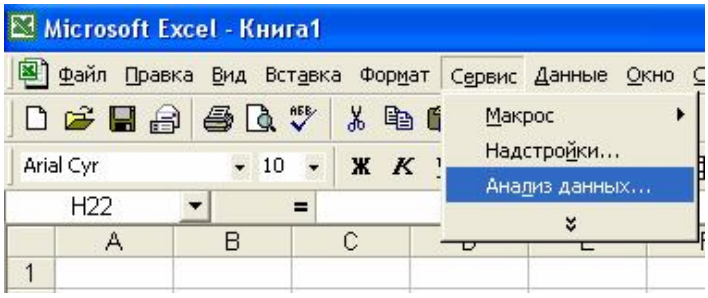


Рис. 1.1.

Шаг 2. В модуле **Анализ данных** выбрать **Описательная статистика,** после чего щелкнуть мышкой **ok** (рис. 1.2). В появившемся окне

выполнить операции и установки, как показано на рис 1.3. Щелкнуть мышкой **ok**. Результат обработки появится в указанном поле (выходной интервал C1). В таблице 1.3 показаны результаты статистической обработки.

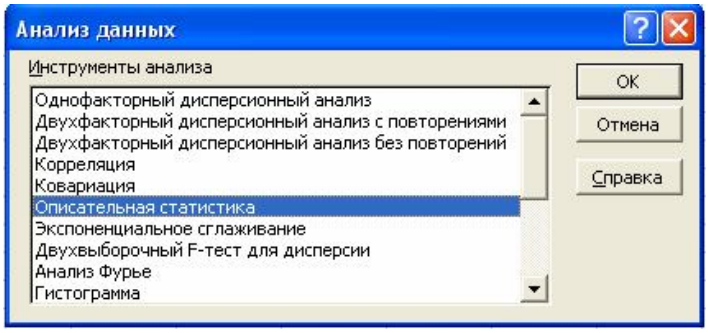


Рис. 1.2. Окно Анализа данных

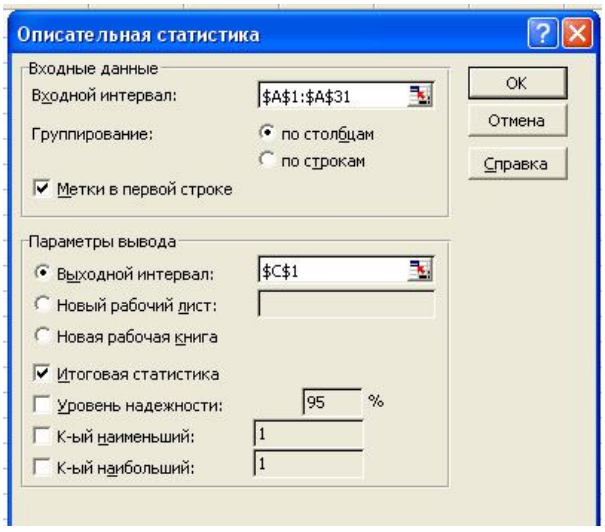


Рис. 1.3. Стартовая панель

Таблица 1.2. Описательная статистика (Результат обработки)

Среднее	9,55
Стандартная ошибка	0,650247521
Медиана	8,95
Мода	8,9
Стандартное отклонение	3,561552354
Дисперсия выборки	12,68465517
Эксцесс	-0,37736441
Асимметричность	0,338895597
Интервал	13,7
Минимум	3,4
Максимум	17,1
Сумма	286,5
Счет	30

1.1 Гистограмма

Вернуться в модуль **Анализ данных** выбрать опцию **Гистограмма,** после чего щелкнуть мышкой **ok**. В появившемся окне выполнить операции и установки, как показано на рис. 1.4. Щелкнуть мышкой **ok**. Результат обработки появится в указанном поле (выходной интервал G1, рис. 1.5).

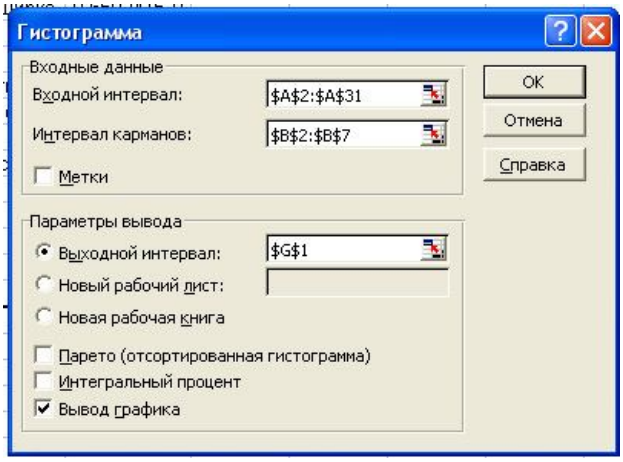


Рис. 1.4. Стартовая панель

O	P
Карман	*Частота*
3,4	1
6,2	4
9	11
11,8	6
14,6	5
17,4	3
Еще	0

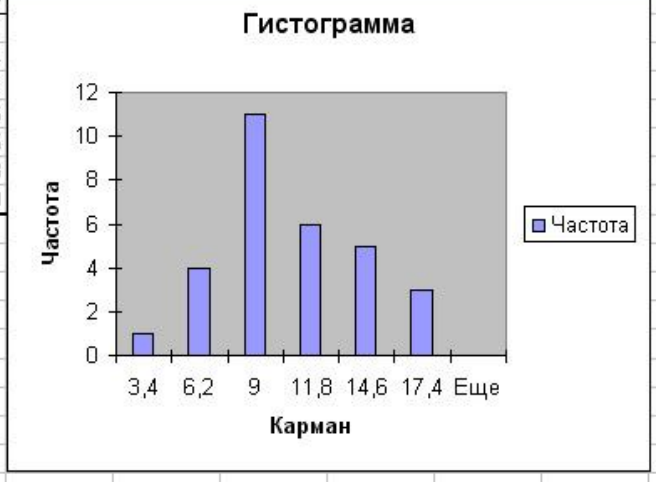


Рис. 1.5. Гистограмма

ТЕМА 2 КОРРЕЛЯЦИЯ

Исходные данные представлены в табл. 2.1 (в табличном редакторе EXCEL данные представлены двумя столбцами).

Открыть модуль **Анализ данных** выбрать опцию **Корреляция,** после чего щелкнуть мышкой **ok**. В появившемся окне выполнить операции и установки, как показано на рис. 2.1. Щелкнуть мышкой **ok**. Результат обработки появится в указанном поле (выходной интервал, E1, табл. 2.2).

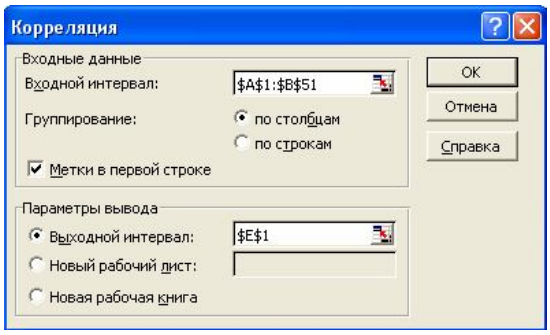


Рис. 2.1. Стартовая панель

Таблица 2.1. Исходные данные

X	Y	X	Y	X	Y
3,4	14,3	8,4	19,8	10,7	21,3
3,6	14,9	8,5	19,9	11,6	21,3
4,5	17,3	8,8	19,9	12	21,8
4,8	17,3	8,9	20,1	12,3	22
4,9	17,4	8,9	20,1	12,6	22,1
5,2	17,5	8,9	20,1	12,7	22,4
5,4	17,6	8,9	20,1	13,3	22,7
5,7	17,6	9	20,2	13,6	23,5
6,2	17,6	9	20,3	13,8	24,2
6,7	17,8	9,1	20,3	14	24,4
7,1	18	9,3	20,5	15	25,2
7,5	18	9,4	20,6	15,2	25,2
7,7	18,1	9,7	20,9	15,8	25,3
7,8	18,1	9,7	21	15,9	25,7
7,9	18,6	9,9	21,1	16,6	26,8
8	19,7	10,1	21,1	17,1	27,5
8,3	19,7	10,4	21,2		

Табл. 2.2. Результат обработки

	X	Y
X	1	0,983328
Y	0,983328	1

ТЕМА 3 РЕГРЕССИОННЫЙ АНАЛИЗ

Для выполнения регрессионного анализа использовались исходные данные табл. 2.1.

Открыть модуль **Анализ данных** выбрать опцию **Регрессия,** после чего щелкнуть мышкой **ok**. В появившемся окне выполнить операции и установки, как показано на рис. 3.1. Щелкнуть мышкой **ok**. Результат обработки появится в указанном поле (выходной интервал, E1, табл. 3.2–3.4).

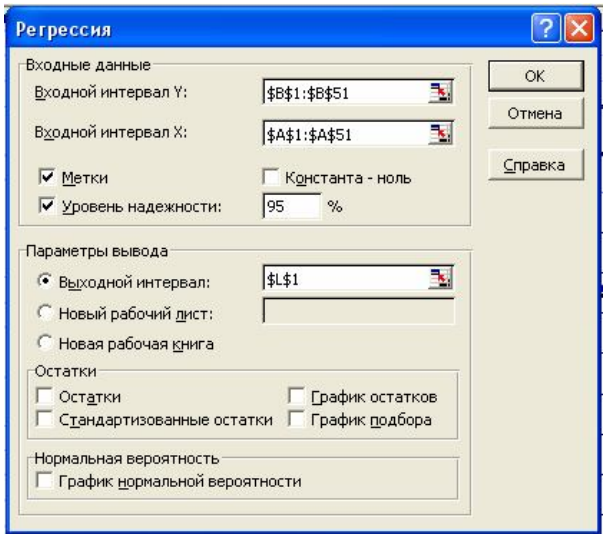


Рис. 3.1. Стартовая панель

Табл. 3.2. Результат обработки

Множественный R	0,9833276
R-квадрат	0,9669331
Нормированный R-квадрат	0,9662443
Стандартная ошибка	0,5320122
Наблюдения	50

Табл. 3.3. Дисперсионный анализ

	df	SS	MS	F	Значимость F
Регрессия	1	397,27	397,27	1403,61	3,4E-37
Остаток	48	13,586	0,28		
Итого	49	410,86			

Табл. 3.4. Регрессионный анализ

	Коэффи-циенты	Стандар-тная ошибка	t-статис-тика	P-значе-ние	Нижние 95%	Верхние 95%	Нижние 95,0%	Верхние 95,0%
Y-пересечение	12,67	0,224	56,628	1,3E-45	12,218	13,118	12,219	13,1189
X	0,815	0,0218	37,465	3,4E-37	0,772	0,8596	0,7729	0,8596

Таким образом, для данного массива данных получена очень надежная регрессия с высоким коэффициентом корреляции:

Y=12,668+0,8158*X, r=0,9833276

Примечание. Операцию регрессии и корреляции можно выполнить в системе Excel, используя модуль **Мастер диаграмм**.

В системе **Excel** открыть модуль **Мастер диаграмм** (рис. 3.2). Выбрать **Тип диаграммы–Точечная**. Щелкнуть по кнопке **Далее**. Выбрать диапазон данных (Рис. 3.3), оформить график и нажать **Готово.**

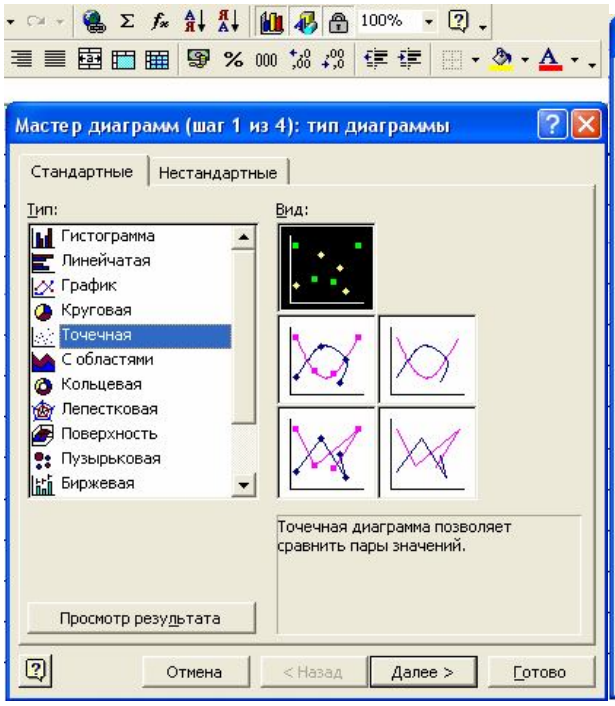


Рис. 3.2. Стартовая панель

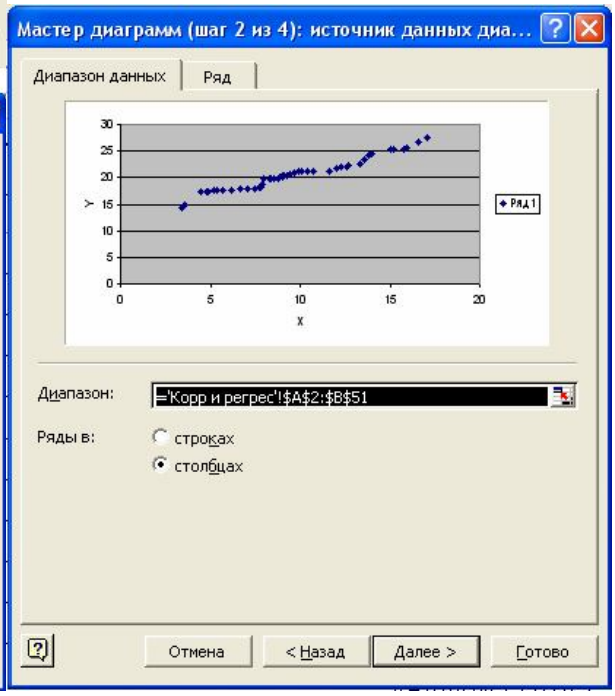


Рис. 3.3.

В системе **Excel** щелкнуть правой кнопкой по точкам и выбрать опцию **Добавить линию тренда**. Выбрать **Тип – линейная** (рис. 3.4). В опции **Параметры** выбрать установки, как показано на рис. 3.5.

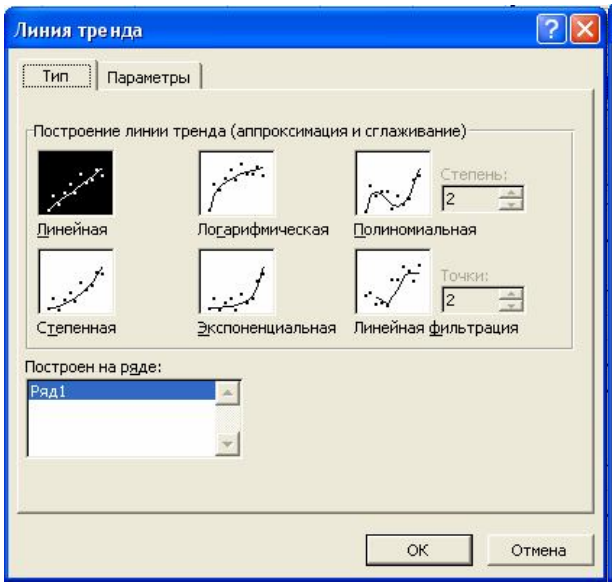


Рис. 3.4.

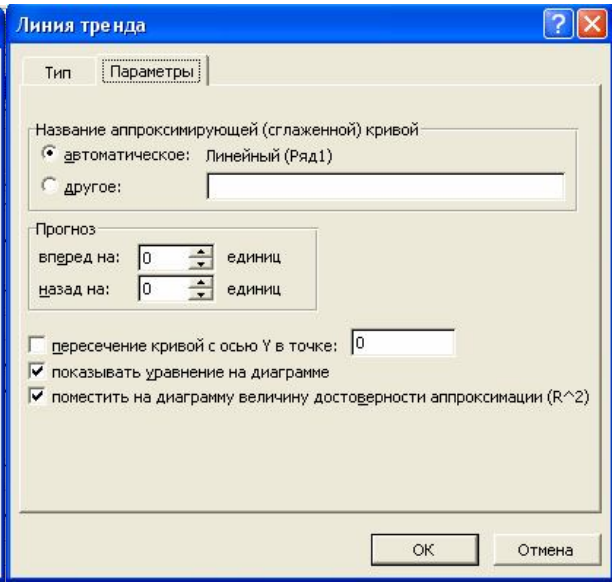


Рис. 3.5.

Отредактированная диаграмма представлена на рис. 3.6.

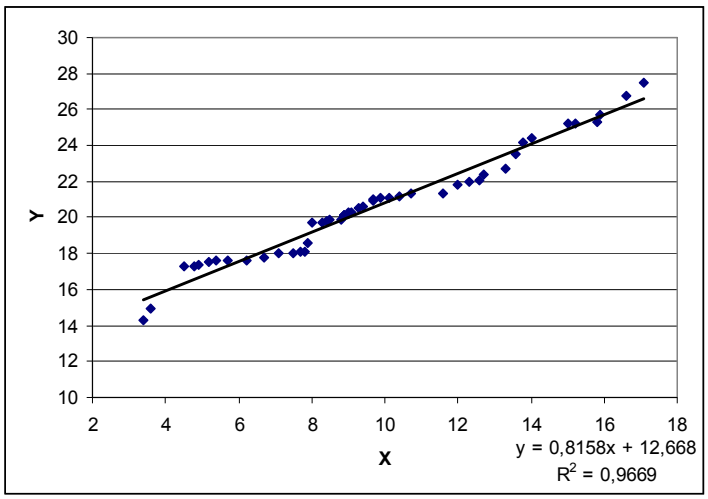


Рис. 3.6. Отредактированная диаграмма

Уравнение регрессии и R^2 находится в правом нижнем углу диаграммы. Как видно, что они такие же, как и при выполнении регрессионного анализа.

ТЕМА 4 ДИСПЕРСИОННЫЙ АНАЛИЗ

4.1 Однофакторный дисперсионный анализ

4.2 Двухфакторный дисперсионный анализ с повторениями

4.1 Однофакторный дисперсионный анализ

Для выполнения регрессионного анализа использовались исходные данные табл. 4.1 (в табличном редакторе EXCEL данные представлены двумя столбцами).

Таблица 4.1. Исходные данные

a	2	3	1	4	3	6	3	5	6	4
b	6	9	9	7	6	6	3	6	5	6

Открыть модуль **Анализ данных** выбрать опцию **Однофакторный дисперсионный анализ,** после чего щелкнуть мышкой **ok**. В появившемся окне выполнить операции и установки, как показано на рис. 4.1. Щелкнуть мышкой **ok**. Результат обработки появится в указанном поле (выходной интервал, M1, табл. 4.2–4.3).

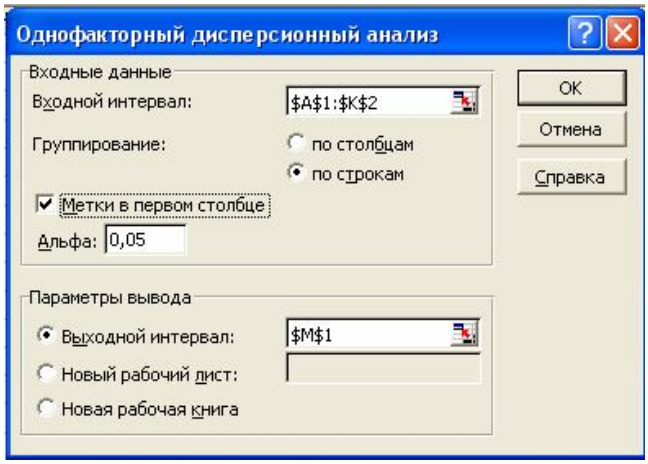


Рис. 4.1. Стартовая панель

Таблица 4.2. Статистические параметры

Группы	Счет	Сумма	Среднее	Дисперсия
a	10	37	3,7	2,677778
b	10	63	6,3	3,122222

Таблица 4.3. Результаты дисперсионного анализа

Источник вариации	SS	df	MS	F	P-Значение	F критическое
Между группами	33,8	1	33,8	11,65517	0,003094	4,413863
Внутри групп	52,2	18	2,9			
Итого	86	19				

В рассмотренном примере *F*-критерий (критерий Фишера) показывает, что различие между средними статистически значимо (значимо на уровне 0.003). Поскольку различие между средними значениями *значимо*, нулевая гипотеза *отвергается* и принимается альтернативная гипотеза о существовании различия между средними.

4.2 Двухфакторный дисперсионный анализ с повторениями

Для выполнения регрессионного анализа использовались исходные данные табл. 5.1.

Таблица 5.1. Исходные данные

	a	b
o	58	49
o	84	55
o	39	48
p	72	74
p	72	74
p	64	85

Открыть модуль **Анализ данных** выбрать опцию **Двухфакторный дисперсионный анализ с повторениями,** после чего щелкнуть мышкой **ok**. В появившемся окне выполнить операции и установки, как показано на рис. 5.1. Щелкнуть мышкой **ok**. Результат обработки появится в указанном поле (выходной интервал, O1, табл. 5.2–5.3).

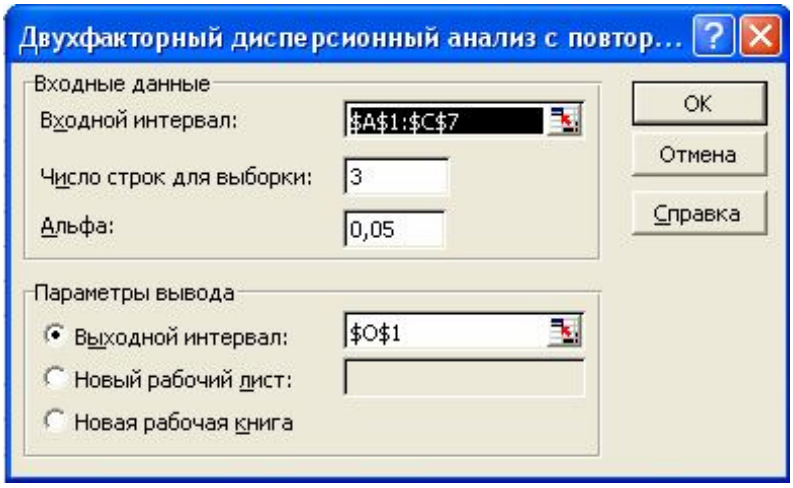


Рис. 5.1. Стартовая панель

Таблица 5.2. Статистические параметры

	a	b	Итого
o			
Счет	3	3	6
Сумма	181	152	333
Среднее	60,333333	50,666667	55,5
Дисперсия	510,33333	14,333333	237,9
p			
Счет	3	3	6
Сумма	208	233	441
Среднее	69,333333	77,666667	73,5
Дисперсия	21,333333	40,333333	45,5
Итого			
Счет	6	6	
Сумма	389	385	
Среднее	64,833333	64,166667	
Дисперсия	236,96667	240,56667	

Таблица 5.3. Дисперсионный анализ

Источник вариации	SS	df	MS	F	P-Значение	F критическое
Выборка (2 фактор)	972	1	972	6,6310404	0,0328668	5,317645
Столбцы (1 фактор)	1,3333333	1	1,3333333	0,0090961	0,926364	5,317645
Взаимодействие	243	1	243	1,6577601	0,2339021	5,317645
Внутри	1172,6667	8	146,58333			
Итого	2389	11				

В рассмотренном примере *F*-критерий показывает, что различие между средними статистически значимо за счет влияния второго фактора (значимо на уровне 0.033). Сила влияния этого фактора составляет около 40%.

СИСТЕМА STATISTICA 6

ТЕМА 5 ПЕРВИЧНЫЙ АНАЛИЗ СТАТИСТИЧЕСКИХ ДАННЫХ В СИСТЕМЕ STATISTICA — МОДУЛЬ BASIC STATISTICS/TABLES (ОСНОВНЫЕ СТАТИСТИКИ/ТАБЛИЦЫ)

6.1 Вероятностный калькулятор

6.2 Нормальное распределение

6.3 Правила 2 и 3 сигма

6.4 Распределение хи-квадрат

6.5 t-распределение Стьюдента

6.6 F-распределение

6.7 Логарифмически-нормальное распределение

Запустить модуль: Basic Statistics/Tables (Основные статистики/таблицы).

Прежде всего, необходимо познакомиться с вероятностным калькулятором, имеющемся в этом модуле и делающем ненужными многие таблицы вероятностных распределений, которые опубликованы в книгах по статистике.

Вероятностный калькулятор находится в модуле Basic Statistics/Tables (Основные статистики/таблицы). С его помощью можно решать многие статистические задачи.

Проделайте следующие действия. Вы на примерах познакомитесь с возможностями калькулятора и с большинством употребляемых в статистике распределений вероятности.

Эти распределения используются в таблицах вывода системы STATISTICA.

5.1 Вероятностный калькулятор

Запустите модуль Basic Statistics/Tables (Основные статистики/таблицы) из переключателя модулей. Высветите в стартовой панели модуля Basic Statistics/Tables (Основные статистики/таблицы) строку Probability

calculator (Вероятностный калькулятор) рис. 6.1, 6.2.

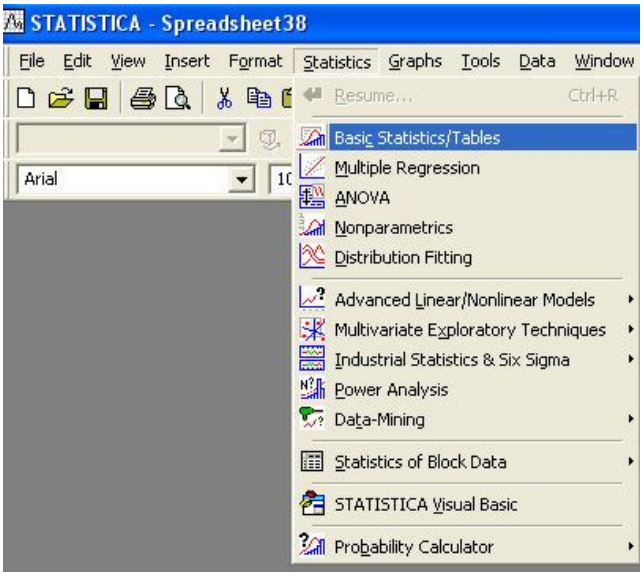


Рис. 6.1

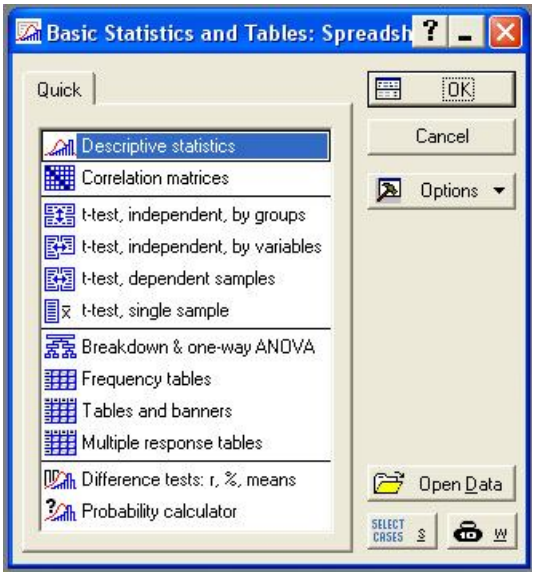


Рис. 6.2.

Нажмите кнопку ОК. Перед вами откроется окно Probability Distribution Calculator (Калькулятор вероятностных распределений) (рис. 6.3):

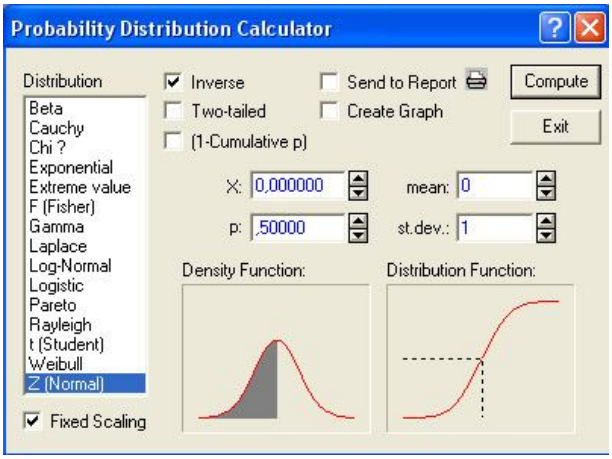


Рис. 6.3. Калькулятор вероятностных распределений

Окно имеет следующую структуру–в левой части список распределений **Distribution (Распределение).** Многие стандартные распределения в этом окне можно выбрать, высвечивая их названия в списке слева: *Бета, Коши, хи-квадрат, нормальное, логнормальное, распределение Стьюдента* и т. д. Выберем, например, в списке строчку **Z(Normal) – Нормальное распределение.** Автоматически справа появляются поля, где можно задать параметры нормального распределения: **среднее – mean и стандартное**

отклонение – st. dev. (см. рисунок 6.3). Система по умолчанию запишет в них стандартные значения: **среднее=0, стандартное отклонение=1.** Эти значения можно изменить, поместив курсор мыши в эти поля, щелкнуть левой кнопкой и ввести с клавиатуры нужные величины.

Одновременно с выбором распределения в левом списке справа в калькуляторе появляются графики нормальной плотности и функции распределения: **Density Function (Функция плотности), Distribution Function (Функция распределения).**

В поле p задается уровень вероятности. Поместите курсор мыши в это поле и щелкните левой кнопкой. Наберите далее любое значение в интервале от 0 до 1. После нажатия на кнопку **Compute (Вычислить)** (в правом верхнем углу калькулятора) в строке Z появится соответствующий квантиль.

То же можно сделать и в обратную сторону — по заданному значению Z вычислить уровень вероятности p. Задав какое-либо значение, щелкните по кнопке **Compute (Вычислить)** в правом верхнем углу. В строке p появится уровень для данного значения Z.

Опции в верхней части окна имеют следующее назначение: Inverse (Обратная функция распределения), Two-tailed (Двухсторонний), 1-Cumulative p, Print (Печать), Create graph (Создать график).

Если пометить опцию **Create graph (Создать график) и** нажать далее кнопку **Compute (Вычислить), то** на экране появится график плотности и функции распределения (задайте в строке p какое-либо значение, например, 0) (рис. 6.4).

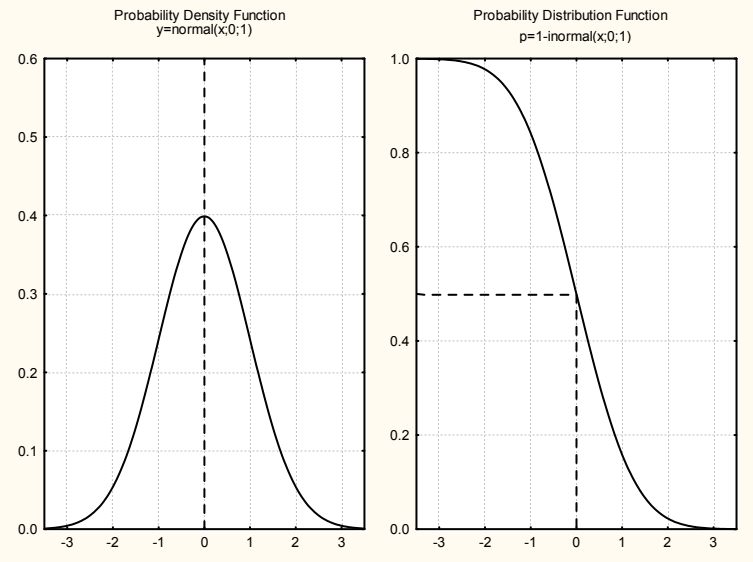


Рис. 6.4. Плотность и функция распределения стандартной нормальной величины

Таким образом, вероятностный калькулятор заменяет многие таблицы. Теперь, вместо того чтобы использовать таблицы распределений, вы можете использовать данный калькулятор.

5.2 Нормальное распределение

Наиболее часто встречающееся в статистике и в теории вероятностей распределение — это нормальное распределение.

Известно, что случайные ошибки в экономических рядах, рядах, возникающих в природе, имеют приблизительно нормальное распределение. Рост взрослых людей так же можно приближенно описать нормальным распределением.

Нормальное распределение имеет два параметра:

mean — среднее;

standard deviation — стандартное отклонение.

Эти параметры задаются в окне вероятностного калькулятора.

Иногда стандартное отклонение называют среднеквадратическим отклонением.

Перечислим некоторые признаки нормального распределения.

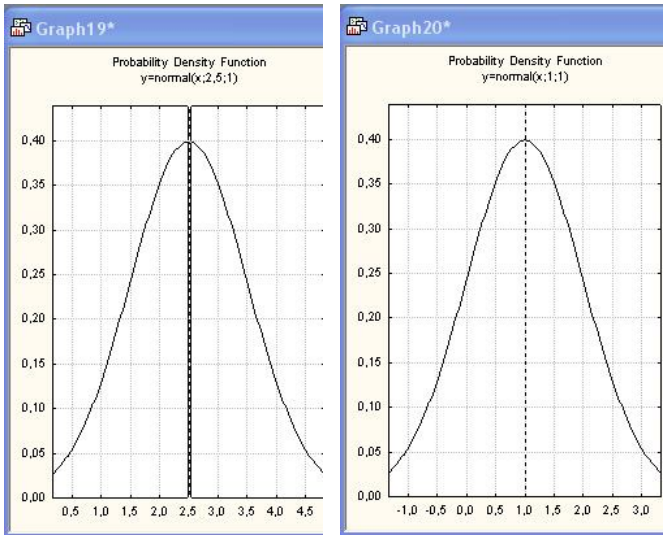
Плотность нормального распределения симметрична относительно

среднего. Среднее значение определяет меру расположения плотности. Среднее значение нормального распределения совпадает с медианой и модой.

Зададим различные значения среднего, оставив пока без изменения стандартное отклонение. Будем считать, что оно равно 1.

Откройте вероятностный калькулятор в поле **mean (среднее),** задав вначале 1. В поле p зададим значение 0.5 (в данном примере это чисто техническая установка).

Выберите опцию **Create graph (Создать график)** и нажмите далее кнопку **Compute (Вычислить),** на экране появится график плотности (рис. 6.5 а):



а) б)

Рис. 6.5. Плотность нормального распределения со средним а)–1; б)–2.5

Повторите те же действия, задав в поле **mean (среднее)** значение 2.5. Вы увидите следующий график (рис. 6.5 б):

Посмотрите внимательно на эти графики. Вы видите, что график плотности нормального распределения сдвигается по оси ординат при изменении среднего. Можно сказать и более точно: при возрастании среднего графики сдвигаются вправо.

Пик плотности нормального распределения находится в точке с

ординатой, равной среднему значению, а плотность симметрична относительно этого значения.

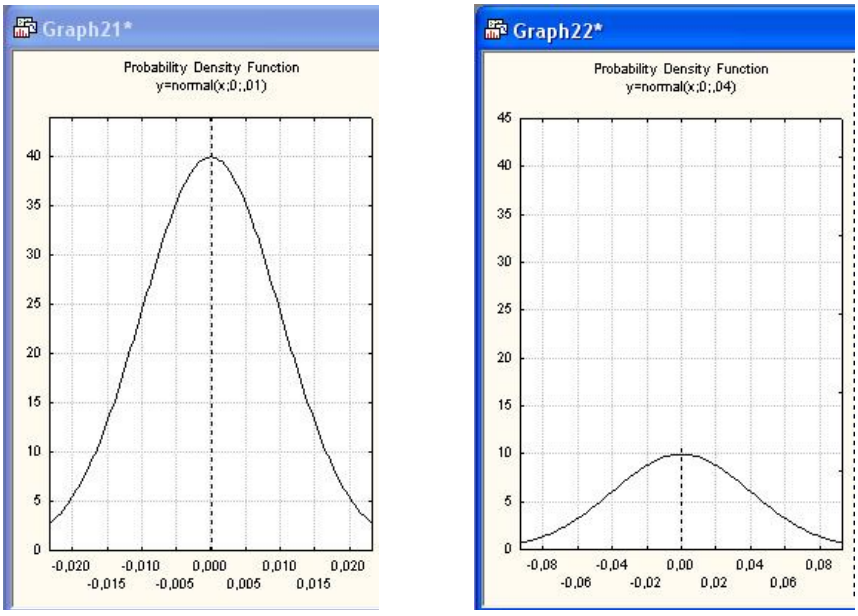
Это значение задается в поле: **mean (среднее).**

Посмотрим, как меняется плотность распределения при изменении другого параметра – стандартного отклонения.

Зададим различные значения стандартного отклонения, считая, что среднее фиксировано и равно 0.

Напомним факт, известный из курса элементарной теории вероятностей, что дисперсия дает меру рассеяния плотности вероятности.

Корень квадратный из дисперсии равен стандартному отклонению. Дисперсию часто обозначают сигмой в квадрате, а стандартное отклонение — просто сигмой. Дисперсия и стандартное отклонение положительны. Дисперсия может сколь угодно приближаться к 0, но может принимать и сколь угодно большое значение, при этом, очевидно, изменяется и распределение вероятности. Покажем на графиках, как изменяется плотность нормального распределения при уменьшении и увеличении дисперсии (рис. 6.6 а, 6.6 б).



а) б)

Рис. 6.6. Плотность нормального распределения со средним 0 и дисперсией а)–0.01 и б)–0.04

Итак, при увеличении дисперсии плотность нормального распределения расплывается или рассеивается относительно среднего значения, при уменьшении дисперсии она, наоборот, сжимается, концентрируясь возле одной точки – точки максимального значения.

Рассмотрим пример использования нормального распределения.

Пример: Известно, что в некоторой стране рост взрослых мужчин приближенно имеет нормальное распределение со средним 176.6 см и стандартным отклонением 7.63 см.

Какова вероятность того, что рост наугад выбранного мужчины не больше 185 см и не меньше 175 см?

Решение.

Шаг 1. Откройте вероятностный калькулятор. Выберите в списке распределений Z(Normal) **(Нормальное распределение).**

Шаг 2. Задайте:

в поле mean — среднее 175.6,

в поле st.dev. — стандартное отклонение 7.63.

Шаг 3. В поле Z задайте 185. Нажмите кнопку **Compute (Вычислить).**

В поле p появилось значение 0.891022. Запомните это значение как **p1.**

Шаг 4. В поле Z задайте 175. Нажмите кнопку **Compute (Вычислить).**

В поле p появилось значение 0.468661. Запомните это значение как **p2.**

Шаг 5. Вычтите **p2** из **p1.** Вы получите 0.422361.

Итак, с вероятностью 0.422361 встреченный вами мужчина имеет рост не ниже 175 и не выше 185 сантиметров.

5.3 Правила 2 и 3 сигма

Правила 2 и 3 сигма полезно знать. Они часто используются на практике. Смысл этих правил состоит в том, что если от точки среднего, или от точки максимума плотности нормального распределения отложить вправо и влево соответственно два и три стандартных отклонения (2 и 3 сигма), то площадь под графиком нормальной плотности, подсчитанная по

этому промежутку, будет соответственно равна 95.45% и 99.73% всей площади под графиком.

Другими словами это можно выразить следующим образом: 95.45% и 99.73% всех независимых наблюдений из нормального распределения лежит в пределах 2-х и 3-х стандартных отклонений от среднего значения.

Это правило также легко проверить с помощью вероятностного калькулятора. Выберите нормальное распределение в списке распределений, задайте, например, стандартные параметры: среднее 0, стандартное отклонение 1, пометьте опцию **Two-tailed (Двухсторонний),** в строке X задайте 2 (два стандартных отклонения), нажмите **Compute,** в строке p появится значение 0.954500, см. рис. 6.7.

В поле **Density Function (Функция плотности)** вероятностного калькулятора показана заштрихованная площадь под графиком плотности, в поле p показано значение 0.9545. Переходя к процентам, имеем 95.45%. Заштрихованная площадь составляет 95.45% всей площади под графиком.

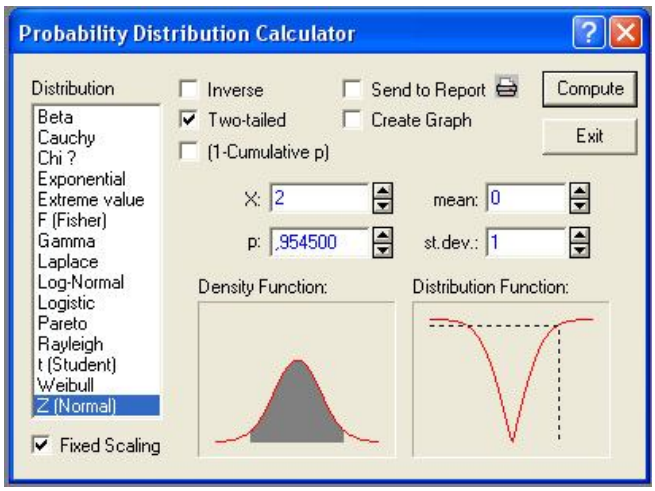


Рис. 6.7. Иллюстрация к правилу 2 сигм

Сделайте тоже самое для 3 сигм. Выберите нормальное распределение, задайте стандартные параметры: среднее 0, стандартное отклонение 1, пометьте опцию Two-tailed в строке X задайте 3 (три стандартных отклонения), **нажмите Compute,** в строке p появится значение 0.997300 (рис. 6.8).

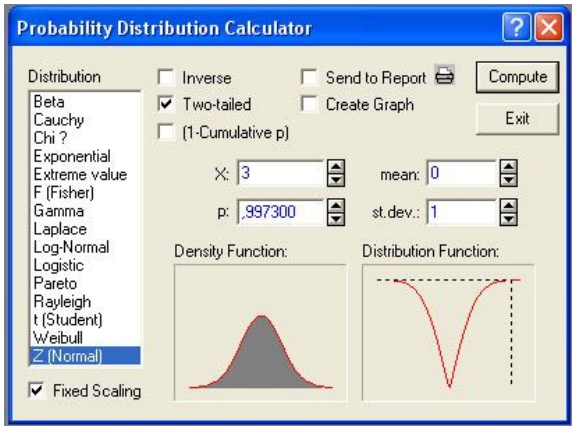


Рис. 6.8. Иллюстрация к правилу 3 сигм

Данные правила действуют при любых значениях среднего и стандартного отклонения нормального закона.

5.4 Распределение хи-квадрат

Случайная величина, имеющая распределение хи-квадрат, определяется как сумма квадратов k независимых стандартных нормальных величин. Нормальные случайные величины – это величины, имеющие нормальное распределение. Число k в определении хи-квадрата называется числом степеней свободы. В частном случае, когда k=l случайная величина хи-квадрат равна квадрату стандартной нормальной величины. Итак, это распределение имеет только один параметр – число степеней свободы, являющийся целым положительным числом.

В списке распределений вероятностного калькулятора выберите **Chi I — хи-квадрат-распределение** (рис. 6.9):

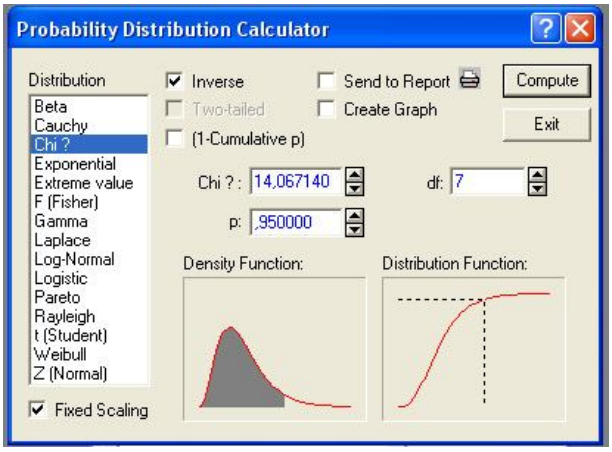


Рис. 6.9. Задание хи-квадрат-распределения в вероятностном калькуляторе

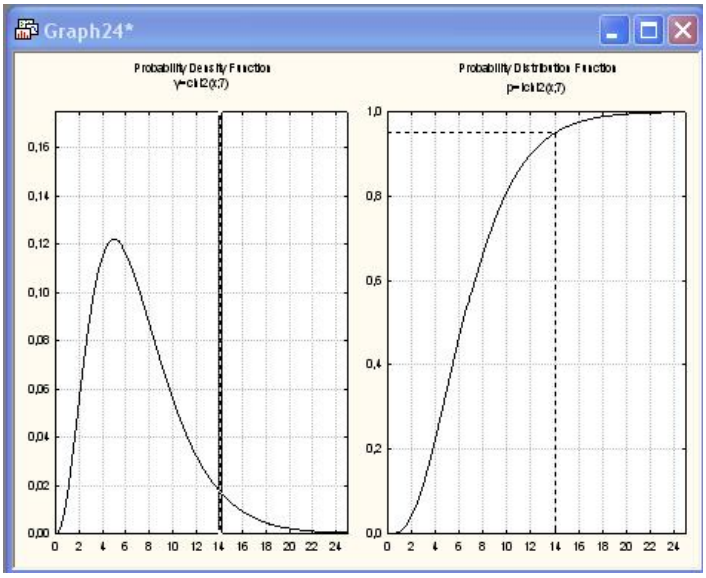
В строке *df* задайте 7 — число степеней свободы.

В поле p задайте 0.95. Нажмите кнопку **Compute (Вычислить),** в строке Chi I вы увидите *0.95 — квантиль хи-квадрат-распределения* с 7 степенями свободы.

Выберите далее опцию **Создать график** и вновь щелкните на кнопке **Compute (Вычислить)** либо просто нажмите **ENTER** на клавиатуре, вы увидите график плотности и функции распределения хи-квадрат с 7 степенями свободы (рис. 6.10).

Обратите внимание на то, что это распределение несимметрично и сосредоточено только на положительной полуоси.

Распределение хи-квадрат играет важную роль при исследовании оценки дисперсии нормальной выборки, а также при проверке зависимостей в таблицах сопряженности и в критериях согласия.



I

Рис. 6.10. График плотности и функции распределения случайной величины хи-квадрат с *7* степенями свободы

5.5 t-распределение Стьюдента

t-распределение важно в тех случаях, когда рассматриваются оценки среднего и неизвестна дисперсия выборки. В этом случае используют выборочную дисперсию и t-распределение.

t-распределение возникает в таблицах вывода регрессионного анализа.

Это одно из важнейших распределений, наряду с нормальным и распределением хи-квадрат.

t-распределение k-степенями свободы сосредоточено на всей действительной оси, симметрично относительно 0. Среднее t-распределения равно 0, дисперсия равна k/(k-2).

В списке распределений вероятностного калькулятора выберите t (Student) (t-распределение Стьюдента) (рис. 6.11).

В строке *df* задайте 5 – число степеней свободы. Пометьте опцию **Create Graph (Создать график).**

В поле p задайте 0.5. Нажмите кнопку **Compute (Вычислить),** на экране вы увидите следующий график (рис. 6.12).

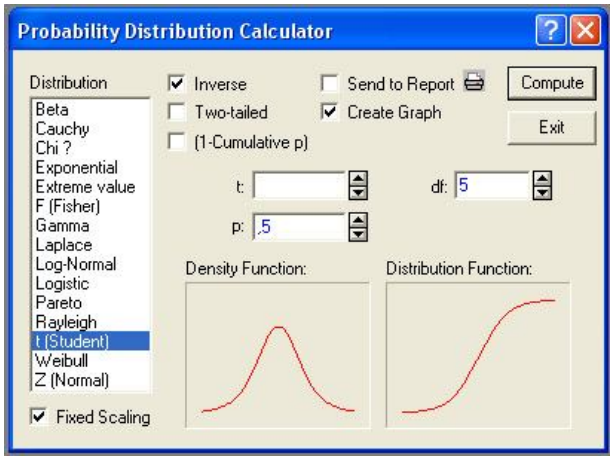


Рис. 6.11. Задание распределения Стьюдента в вероятностном калькуляторе

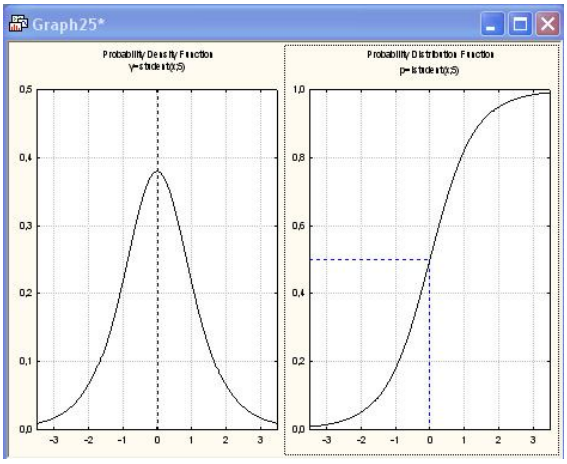


Рис. 6.12. Плотность и функция t-распределения Стьюдента с 5 степенями свободы

При больших степенях свободы (больших 30) t-распределение практически совпадает со стандартным нормальным распределением.

Плотность t-распределение деформируется при возрастании числа степеней свободы следующим образом: пик увеличивается, хвосты более круто идут к 0, кажется, как будто плотность сжимается с боков.

В такой деформации плотности легко убедиться с помощью вероятностного калькулятора. Задайте в поле df (степень свободы)

значение 50. Нажав кнопку Compute (Вычислить), на экране вы увидите следующий график (рис. 6.13).

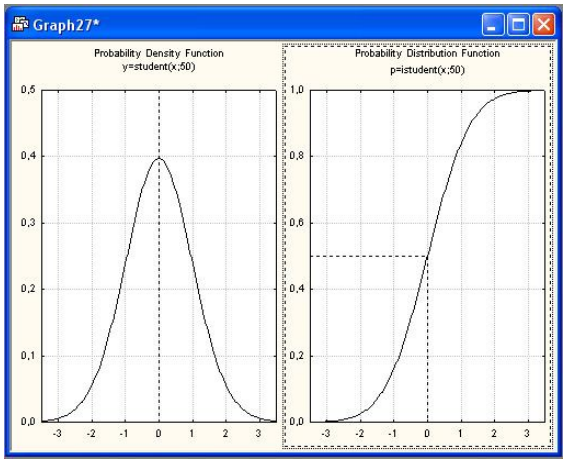


Рис. 6.13. Плотность и функция распределения Стьюдента с 50 степенями свободы

Сравнив график плотности распределения Стьюдента с большим числом степеней свободы, например 50, и график плотности стандартного нормального распределения, вы убедитесь, что они очень похожи.

5.6 F-распределение

F-распределение возникает в регрессионном, дисперсионном и дискриминантном анализе, а также в других видах многомерного анализа данных.

Далее оно будет неоднократно встречаться в таблицах вывода системы STATISTICA. Именно поэтому с ним следует ознакомиться подробнее.

Случайная величина, имеющая F-распределение с парой степеней свободы m, n, определяется как отношение двух независимых случайных величин, имеющих распределение хи-квадрат со степенями свободы m и n с умножением на нормировочный сомножитель n/m.

F-распределение сосредоточено на положительной полуоси. Это распределение в отличие от нормального несимметрично. Покажем, как построить график F-распределения и вычислить его процентные точки.

В списке распределений вероятностного калькулятора выберите F (F-распределение) (рис. 6.14).

Задайте в поле df1 (степень свободы 1) значение 10, в поле df2 (степень свободы 2) — значение 11. Пометьте опцию Create Graph (Создать График).

В поле p задайте 0.5. Нажав кнопку Compute (Вычислить), на экране вы увидите следующий график (рис. 6.15):

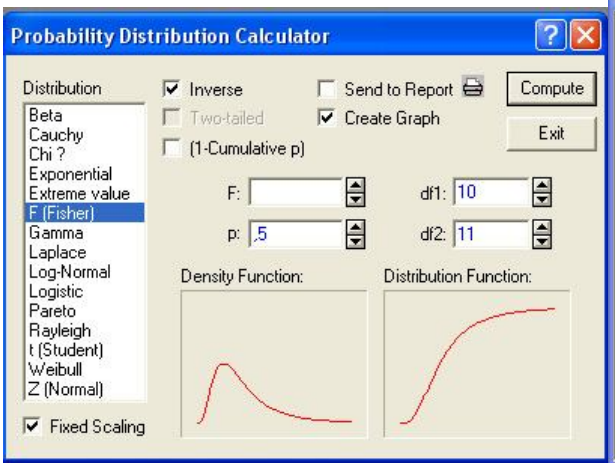


Рис. 6.14. Задание F-распределения в вероятностном калькуляторе

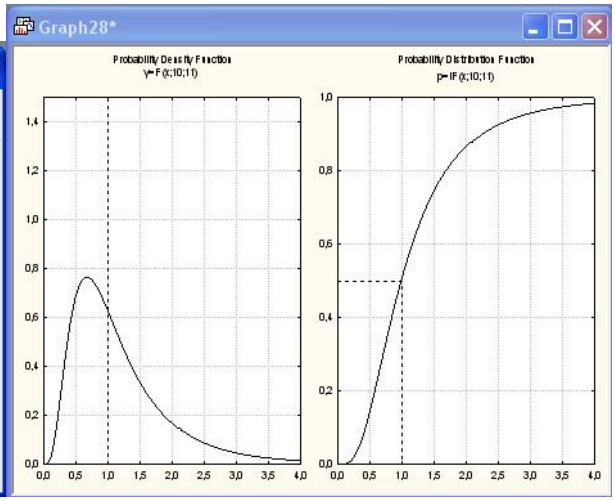


Рис. 6.15. Плотность и функция F-распределения со степенями свободы 10, 11

5.7 Логарифмически-нормальное распределение

Говорят, что случайная величина X имеет логарифмически-нормальное распределение, если величина Ln (X) является нормальной. Словами это можно выразить так: логарифм логарифмически-нормальной величины является нормальной величиной. Так как нормальное распределение описывается двумя параметрами, то и логарифмически-нормальное распределение также имеет два параметра.

Плотность распределения имеет одно максимальное значение и несимметрично. График плотности логарифмически-нормального распределения показан на рис. 6.16.

В списке распределений вероятностного калькулятора выберите **Log-**

Normal – Логарифмически-нормальное распределение (рис. 6.17).

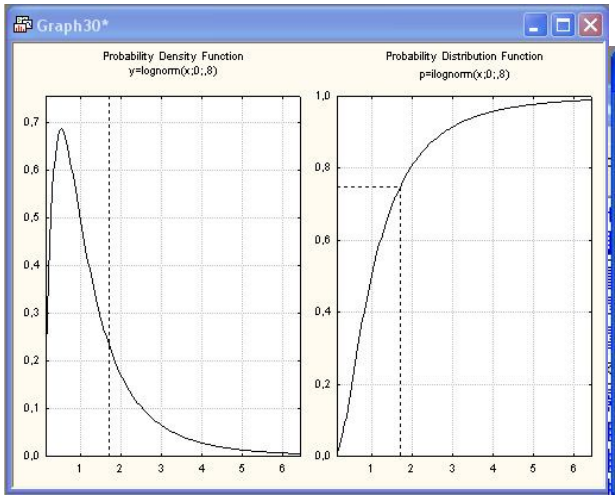


Рис. 6.16. Плотность логарифмически-нормального распределения

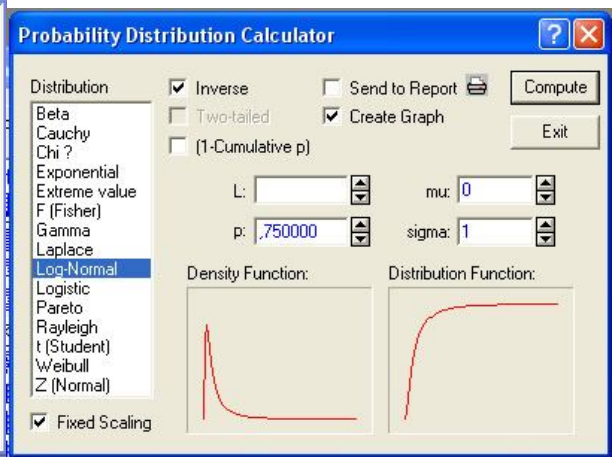


Рис. 6.17. Задание логарифмически-нормального распределения в вероятностном калькуляторе

Также легко, например, вычислить процентные точки и построить графики других необходимых для расчетов распределений: бета-распределение, распределение Коши и т.д. Для этого необходимо задать в вероятностном калькуляторе вид распределения и параметры.

ТЕМА 6 БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ И ИГРОВЫЕ ЗАДАЧИ

Биномиальное распределение является важнейшим дискретным распределением, то есть распределением, которое сосредоточено всего лишь в нескольких точках. Этим точкам биномиальное распределение приписывает ненулевые вероятности. Таким образом, биномиальное распределение отличается от всех разобранных выше распределений (нормального, хи-квадрат и др.), которые приписывают *нулевые* вероятности отдельно выбранным точкам и называются непрерывными.

Лучше всего понять биномиальное распределение исходя из следующего эксперимента. Представьте, вы бросаете монету. Пусть вероятность выпадения герба есть p, а вероятность выпадения решки есть *q=1–p* (мы рассматриваем самый общий случай, когда монета несимметрична, имеет, например, смещенный центр тяжести). Выпадения герба считаем успехом, а выпадение решки – неудачей. Тогда число выпавших гербов (или решек) имеет биномиальное распределение.

Параметрами биномиального распределения являются вероятность успеха *p (q=1–p)* и число испытаний *n*. Точная формула для вероятности *m*-успехов в *n*-испытаниях такая:

$$p(m;n) = B(m;n)\cdot p^{n}\cdot(1-p)^{n-m} \quad \text{m=0,1 ... n,}$$

где $B(m;n) = \dfrac{n!}{(n-m)!\cdot m!}$ – биномиальный коэффициент.

На практике часто нужно вычислять отдельные биномиальные вероятности и суммировать их далее по определенному множеству целых чисел. Это достаточно трудоемкая процедура (представьте, что *n* и *m* – большие числа). В литературе существуют обширные таблицы.

В нынешней реализации вероятностного калькулятора в STATISTICA нет биномиального распределения, однако биномиальное распределение реализовано в языке STATISTICA BASIC, и им легко воспользоваться.

Создайте пустую электронную таблицу testsm.sta вида (рис. 7.1).

Дважды щелкните на имени переменной var1 и откройте диалоговое окно спецификации переменной var1.

В нижней части окна в поле **Long name** запишите формулу, как показано на рисунке ниже.

Нажмите кнопку OK в правом верхнем углу окна (рис. 7.2).

Согласно этой формуле программа вычислит вероятности успеха и занесет их в таблицу в значения первой переменной. Теперь таблица примет вид, представленный на рис. 7.3.

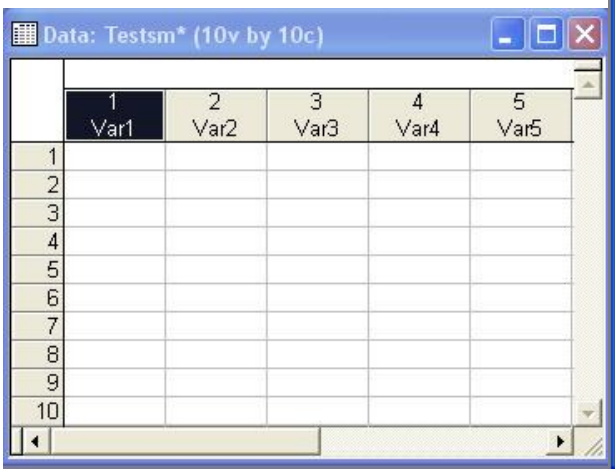


Рис. 7.1. Пустая электронная таблица testsm.sta

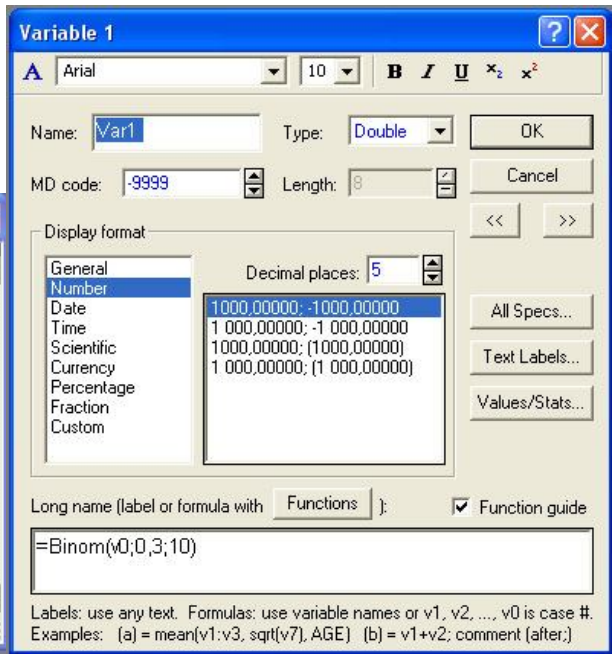


Рис. 7.2. Задание формулы вычисления биномиальных вероятностей

В данной таблице вероятность успеха – выпадения герба – равна 0.3. Из таблицы вы видите, что вероятность выпадения *ровно* одного герба в 10 бросаниях – 0.12106, вероятность выпадения *ровно* двух гербов в 10 бросаниях – 0.2334 и т.д.

Вероятность успеха легко изменить, сделав ее равной, например, 0.5. Это означает, что бросается симметричная монета и вероятность успеха

равна вероятности неудачи (вероятность выпадения герба равна вероятности выпадения решки).

Дважды щелкните на имени переменной ВЕРОЯТ и откройте окно спецификации переменной var1.

В нижней части окна в поле **Long name** измените формулу, вместо 0.3 запишите 0.5.

Нажмите кнопку ОК в правом верхнем углу. Программа вычислит новые биномиальные вероятности и занесет их в электронную таблицу (рис. 7.4).

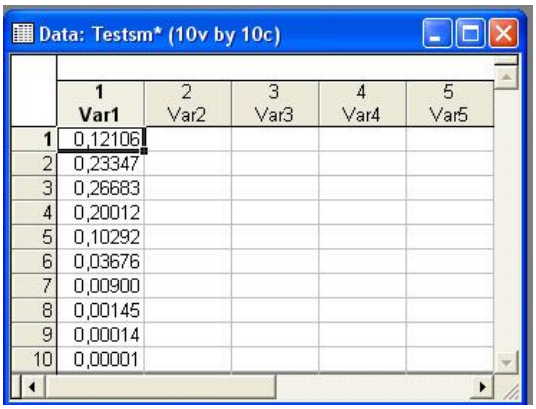


Рис. 7.3. Электронная таблица с биномиальными вероятностями (вероятность успеха 0.3, число испытаний 10)

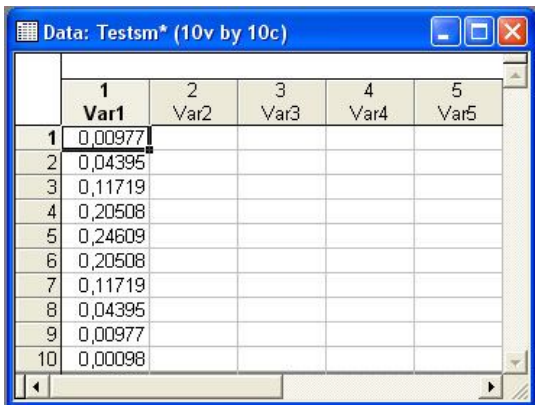


Рис. 7.4. Электронная таблица с биномиальными вероятностями (вероятность успеха 0.5, число испытаний 10)

Заметьте, что максимальная вероятность в этой таблице приходится на значение 5, что и понятно из соображений симметрии.

Если вы забудете функцию, которая вычисляет биномиальные вероятности в системе, то воспользуйтесь средством **Function Browser.**

Нажав кнопку **Functions** в окне спецификации переменной, вы откроете диалоговое окно **Function Browser,** в котором легко выбрать нужную функцию биномиального распределения.

Задача шевалье де Мере

Классическим и вместе с тем забавным является пример шевалье де Мере, когда ставший известным в веках, благодаря своей любознательности азартный игрок спросил себя: стоит ему ставить на выпадение двух шестерок одновременно при бросании двух костей 24 раза или нет? Его собственные вычисления показали, что стоит, так как вероятность данного события при 24 бросаниях костей больше 1/2. Как же он удивился, когда с течением времени обнаружил, что постоянно оказывается в проигрыше! Оскорбленный де Мере во всем обвинил статистику. И только знаменитый Паскаль указал ему на заблуждение: оказывается, вероятность данного события 0,49, следовательно, в длинной серии игр, состоящих в 24 подбрасываниях двух костей, выигрыш происходит лишь в 49%, а не в более чем 50% игр, как рассчитывал де Мере. Шевалье обычно играл всю ночь, и для него было важно, чтобы в более чем половине игр он был в выигрыше.

Сейчас мы «мгновенно» решим эту задачу с помощью самых простых средств STATISTICA.

Создайте рабочий файл play.sta. Дважды щелкните на имени переменной и откройте окно спецификации переменной var1.

В нижней части окна в поле **Long name** запишите формулу, как показано на рис. 7.5. Число испытаний в задаче шевалье 24. Вероятность успеха равна 1/36, потому что с такой вероятностью при бросании двух костей выпадают шестерки.

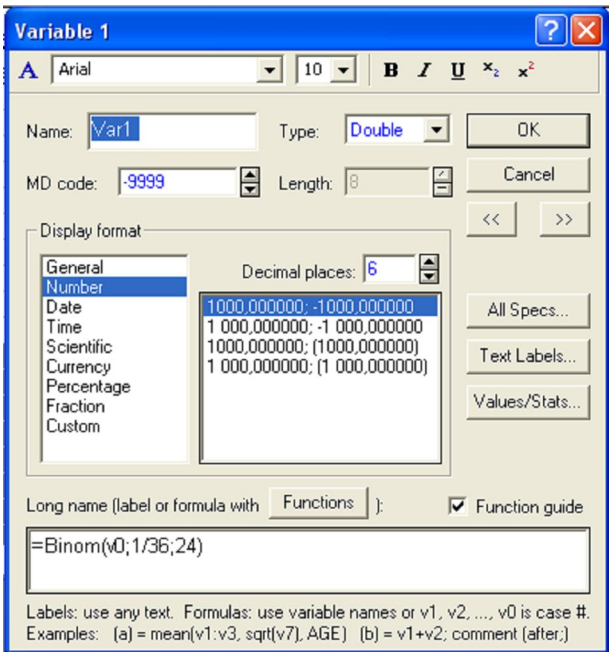


Рис. 7.5. Задание формулы биномиального распределения в задаче шевалье

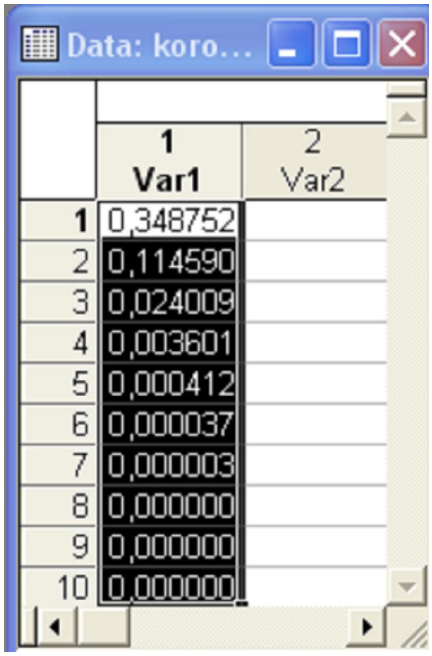


Рис. 7.6. Вероятности выпадения шестерок при 24 бросаниях двух костей

Нажмите кнопку ОК в правом верхнем углу. Программа вычислит биномиальные вероятности.

В первом столбце этой таблицы даны последовательно вероятности выпадения двух шестерок один раз, два раза, три раза и т.д.

Шевалье де Мере спросил, стоит ли ему ставить на выпадение двух шестерок одновременно при бросании двух костей 24 раза или нет?

Нам нужно вычислить вероятность выпадения по крайней мере одной пары шестерок. Следовательно, все эти вероятности нужно сложить. Сделав это, вы получите ответ в классической задаче. Вероятность выпадения по крайней мере одной пары шестерок при 24 бросаниях пары костей равна 0.49140.

Таким образом, в длинной серии игр, состоящих из 24 бросаний пары костей, игрок, ставящий на выпадение двух шестерок одновременно, в среднем устойчиво проигрывает.

Но вот вопрос: как изменить условия игры, чтобы находиться в

выигрыше?

Будем ли мы выигрывать, если игра состоит из 25 бросаний, то есть мы ставим на выпадение пары шестерок в 25 бросаниях.

Измененная задача шевалье де Мере

Итак, предположим, что шевалье де Мере изменил условия игры и стал ставить на выпадение пары шестерок в 25 бросаниях. Оказывается, увеличение числа бросков всего на 1 делает игру уже выигрышной.

В этом мы сейчас убедимся. Будем по-прежнему работать с файлом play.sta.

Повторите все действия предыдущей задачи с переменной var2. Дважды щелкните на имени переменной и откройте окно спецификации переменной var2 (рис. 7.7).

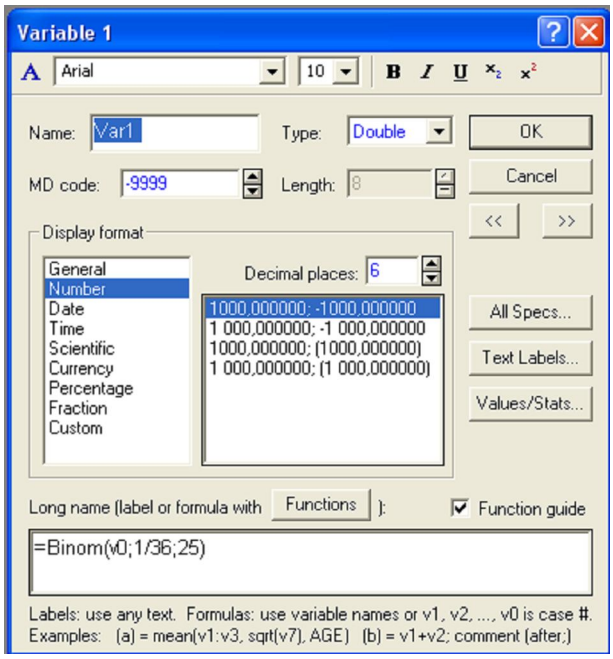


Рис. 7.7. Задание формулы биномиального распределения в измененной задаче шевалье

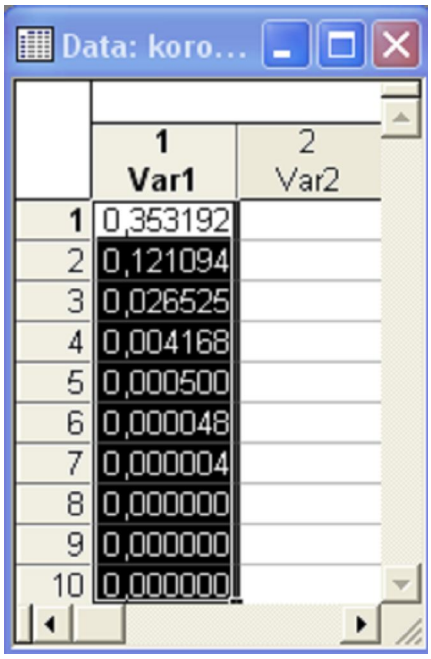


Рис. 7.8. Вероятности выпадения шестерок при 25 бросаниях двух костей

Заметьте, единственное отличие этой формулы от формулы в исходной задаче шевалье в том, что мы поставили число испытаний 25 вместо 24.

Нажмите кнопку OK в правом верхнем углу. Программа вычислит

новые биномиальные вероятности и занесет их в значения переменной var2.

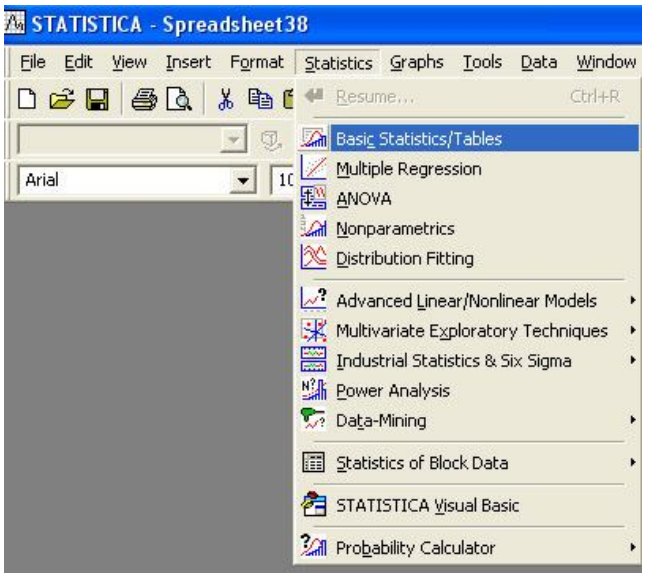
Теперь файл play.sta будет выглядеть следующим образом (рис. 7.8).

Складывая значения в столбце, легко найти, что вероятность выпадения по крайней мере одной пары шестерок в 25 подбрасываниях пары костей больше 0.5.

Если бы шевалье де Мере играл в такую игру, он находился бы в среднем в выигрыше, так как в более чем 50% игр, состоящих из 25 подбрасываний пары костей, по крайней мере один раз выпадали бы шестерки.

ТЕМА 7 ОПИСАТЕЛЬНАЯ СТАТИСТИКА. ПЕРВИЧНЫЙ АНАЛИЗ СТАТИСТИЧЕСКИХ ДАННЫХ В STATISTICA6

Запустите модуль Basic Statistics/Tables (Основные статистики/таблицы) из Переключателя модулей. Высветите в стартовой панели модуля Basic Statistics/Tables (Основные статистики/таблицы) строку Descriptive statistics (Описательная статистика) рис. 8.1



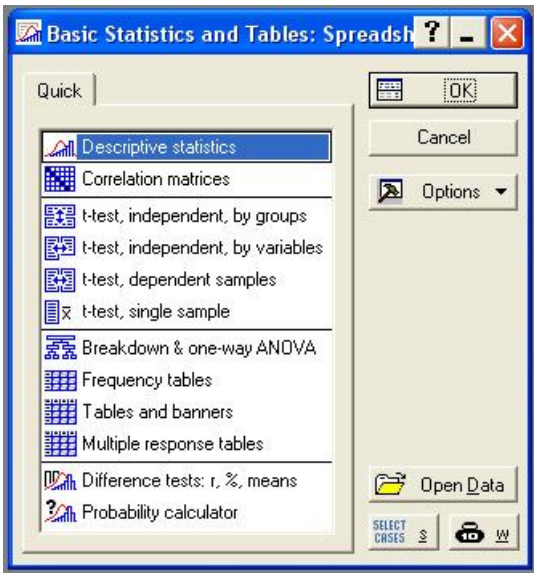


Рис. 8.1 Рис. 8.2. Описательная статистика

Шаг 1. Нажмите кнопку ОК. Перед вами откроется окно Descriptive statistics (Описательная статистика) (рис. 8.3, 8.4).

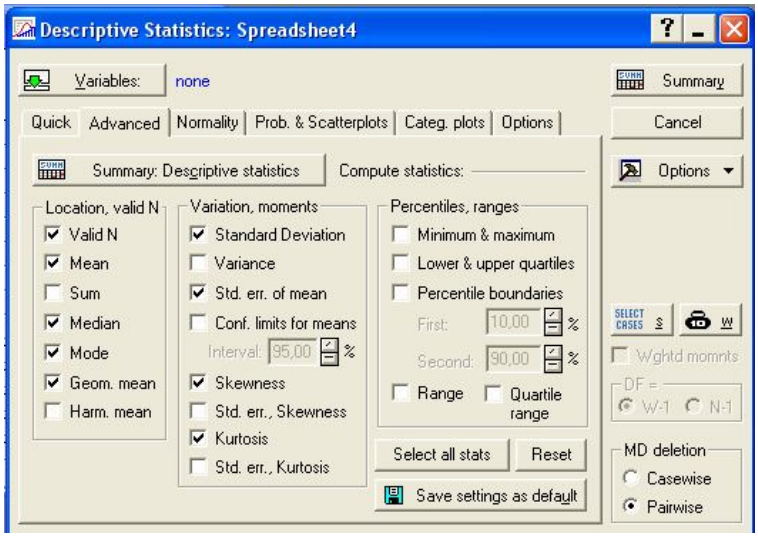
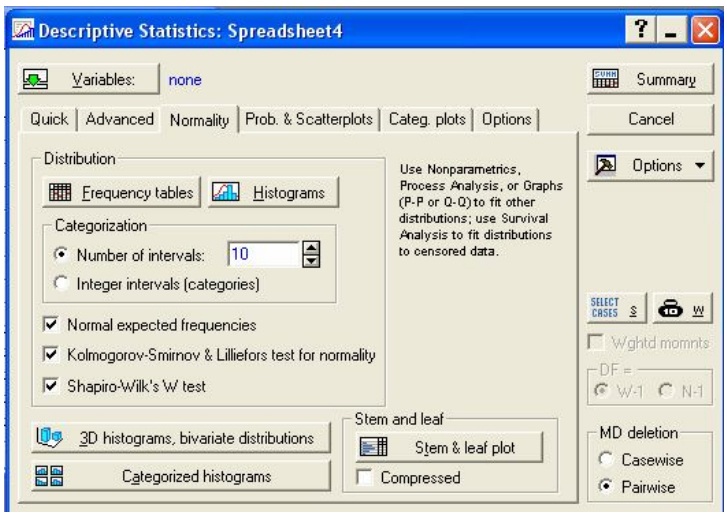


Рис. 8.3.

Выполните установки, как показано на рис. 8.3 и 8.4.

Шаг 2. Загрузите в систему STATISTICA исходные данные (рис. 8.5)



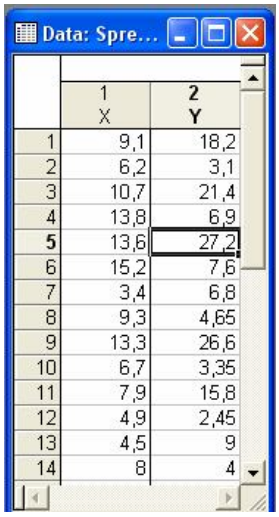
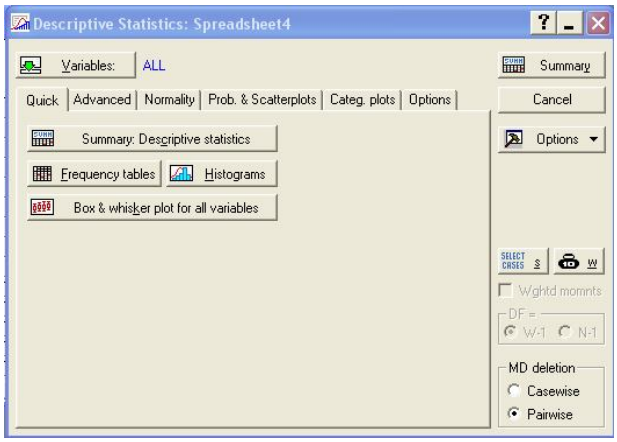


Рис. 8.4. Рис. 8.5.

Шаг 3. Возвратитесь в окно Descriptive statistics (Описательная статистика). В появившемся окне выберите переменные для анализа (рис.8.6.). Выбор переменных осуществляется с помощью кнопки **Переменные (Variables)**. Нажав кнопку **Variables** в окне рис. 8.6 выберите переменные (в данном случае X и Y) (рис. 8.7).



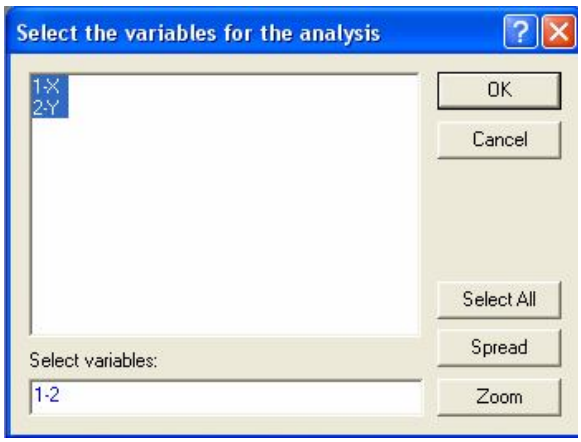


Рис. 8.6. Диалоговое окно результатов Рис. 8.7. Окно выбора переменных

Шаг 4. В диалоговом окне результатов последовательно нажмите кнопки:

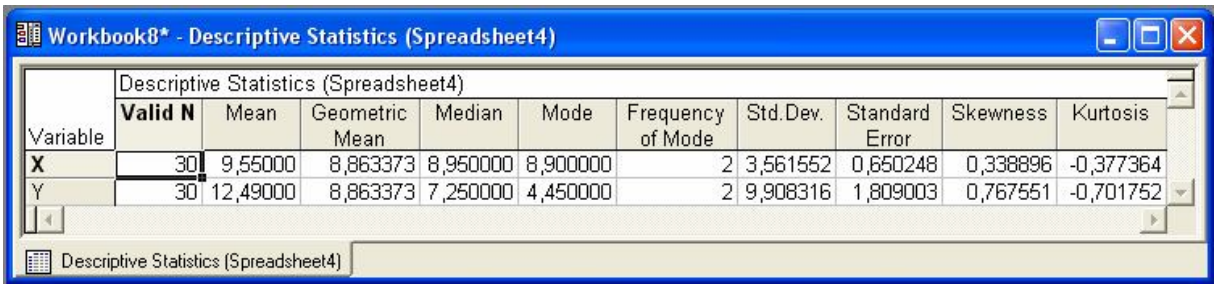
описательная статистика – Summary Descriptive statistics;

таблица частот – Frequency Tables;

гистограммы – Histograms.

Результаты статистической обработки представлены на рис. 8.8 – 8.12.

Из представленной обработки следует, что массив данных X описывается функцией нормального распределения, а Y – нет, что достаточно хорошо видно из рис. 8.11 и 8.12. Более подробную информацию можно извлечь из анализа результатов обработки, представленных в таблицах (рис 9.8–9.10).



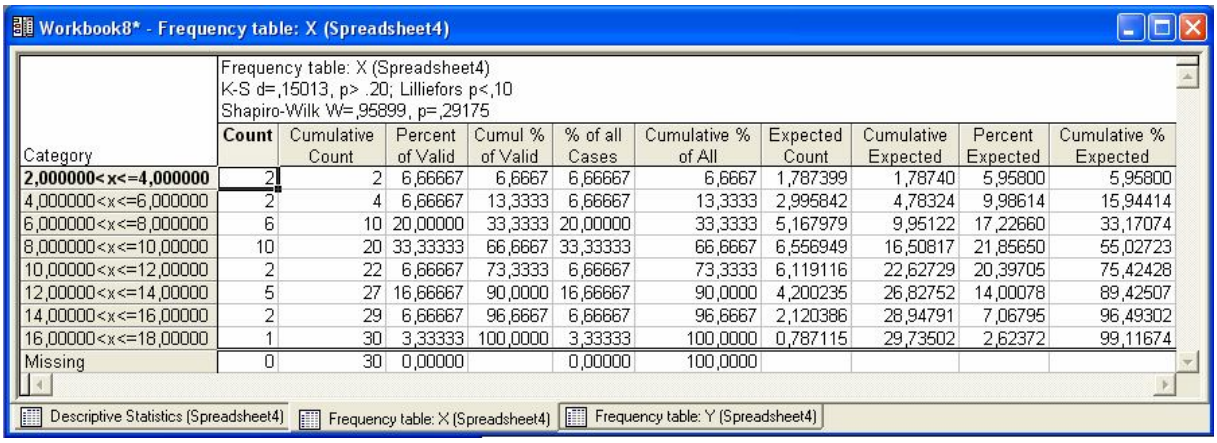
Workbook8* - Descriptive Statistics (Spreadsheet4)

Descriptive Statistics (Spreadsheet4)

Variable	Valid N	Mean	Geometric Mean	Median	Mode	Frequency of Mode	Std.Dev.	Standard Error	Skewness	Kurtosis
X	30	9,55000	8,863373	8,950000	8,900000	2	3,561552	0,650248	0,338896	-0,377364
Y	30	12,49000	8,863373	7,250000	4,450000	2	9,908316	1,809003	0,767551	-0,701752

Descriptive Statistics (Spreadsheet4)

Рис. 8.8.



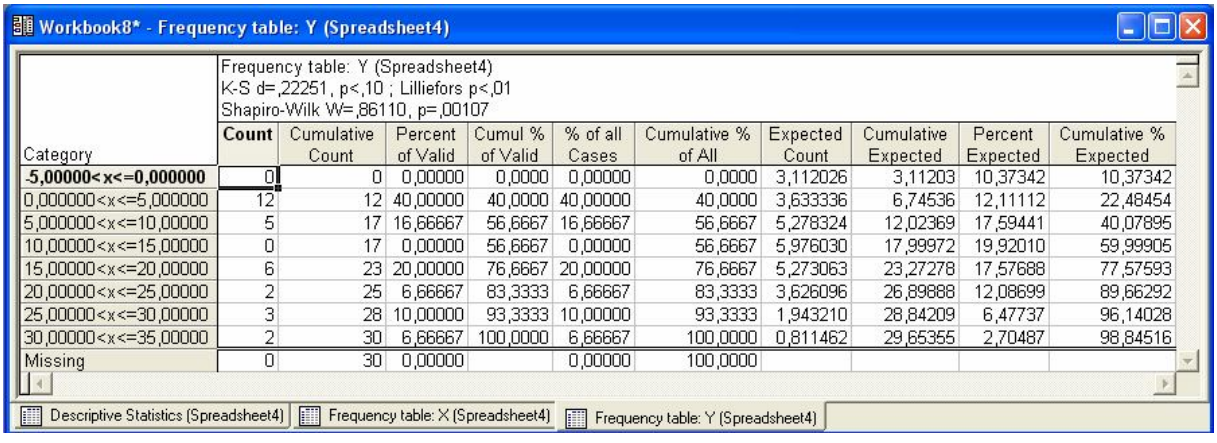
Workbook8* - Frequency table: X (Spreadsheet4)

Frequency table: X (Spreadsheet4)
K-S d=,15013, p> .20; Lilliefors p<,10
Shapiro-Wilk W=,95899, p=,29175

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All	Expected Count	Cumulative Expected	Percent Expected	Cumulative % Expected
2,000000<x<=4,000000	2	2	6,66667	6,6667	6,66667	6,6667	1,787399	1,78740	5,95800	5,95800
4,000000<x<=6,000000	2	4	6,66667	13,3333	6,66667	13,3333	2,995842	4,78324	9,98614	15,94414
6,000000<x<=8,000000	6	10	20,00000	33,3333	20,00000	33,3333	5,167979	9,95122	17,22660	33,17074
8,000000<x<=10,00000	10	20	33,33333	66,6667	33,33333	66,6667	6,556949	16,50817	21,85650	55,02723
10,00000<x<=12,00000	2	22	6,66667	73,3333	6,66667	73,3333	6,119116	22,62729	20,39705	75,42428
12,00000<x<=14,00000	5	27	16,66667	90,0000	16,66667	90,0000	4,200235	26,82752	14,00078	89,42507
14,00000<x<=16,00000	2	29	6,66667	96,6667	6,66667	96,6667	2,120386	28,94791	7,06795	96,49302
16,00000<x<=18,00000	1	30	3,33333	100,0000	3,33333	100,0000	0,787115	29,73502	2,62372	99,11674
Missing	0	30	0,00000		0,00000	100,0000				

Descriptive Statistics (Spreadsheet4) | Frequency table: X (Spreadsheet4) | Frequency table: Y (Spreadsheet4)

Рис. 8.9.



Workbook8* - Frequency table: Y (Spreadsheet4)

Frequency table: Y (Spreadsheet4)
K-S d=,22251, p<,10 ; Lilliefors p<,01
Shapiro-Wilk W=,86110, p=,00107

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All	Expected Count	Cumulative Expected	Percent Expected	Cumulative % Expected
-5,00000<x<=0,000000	0	0	0,00000	0,0000	0,00000	0,0000	3,112026	3,11203	10,37342	10,37342
0,000000<x<=5,000000	12	12	40,00000	40,0000	40,00000	40,0000	3,633336	6,74536	12,11112	22,48454
5,000000<x<=10,00000	5	17	16,66667	56,6667	16,66667	56,6667	5,278324	12,02369	17,59441	40,07895
10,00000<x<=15,00000	0	17	0,00000	56,6667	0,00000	56,6667	5,976030	17,99972	19,92010	59,99905
15,00000<x<=20,00000	6	23	20,00000	76,6667	20,00000	76,6667	5,273063	23,27278	17,57688	77,57593
20,00000<x<=25,00000	2	25	6,66667	83,3333	6,66667	83,3333	3,626096	26,89888	12,08699	89,66292
25,00000<x<=30,00000	3	28	10,00000	93,3333	10,00000	93,3333	1,943210	28,84209	6,47737	96,14028
30,00000<x<=35,00000	2	30	6,66667	100,0000	6,66667	100,0000	0,811462	29,65355	2,70487	98,84516
Missing	0	30	0,00000		0,00000	100,0000				

Descriptive Statistics (Spreadsheet4) | Frequency table: X (Spreadsheet4) | Frequency table: Y (Spreadsheet4)

Рис. 8.10.

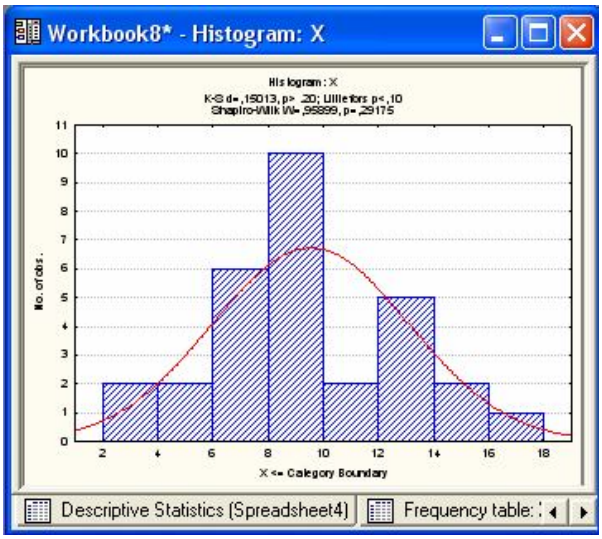


Рис. 8.11.

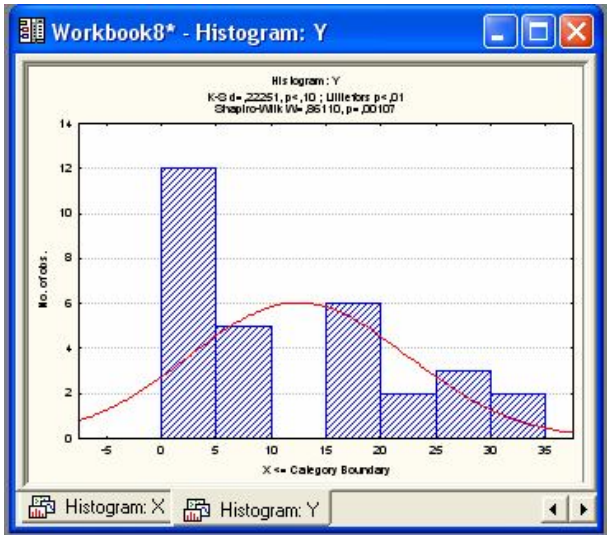


Рис. 8.12.

ТЕМА 8 КОРРЕЛЯЦИОННЫЙ АНАЛИЗ В STATISTICA6

Определение корреляции. Корреляция представляет собой меру зависимости переменных. Наиболее известна корреляция Пирсона. При вычислении корреляции Пирсона предполагается, что переменные измерены, как минимум, в интервальной шкале. Некоторые другие коэффициенты корреляции могут быть вычислены для менее информативных шкал. Коэффициенты корреляции изменяются в пределах от –1.00 до +1.00. Обратите внимание на крайние значения коэффициента корреляции. Значение -1.00 означает, что переменные имеют строгую отрицательную корреляцию. Значение +1.00 означает, что переменные имеют строгую положительную корреляцию. Отметим, что значение 0.00 означает отсутствие корреляции. Наиболее часто используемый коэффициент корреляции *Пирсона r* называется также *линейной* корреляцией, т.к. измеряет степень линейных связей между переменными.

Простая линейная корреляция (Пирсона–r). Корреляция Пирсона (далее называемая просто *корреляцией*) предполагает, что две рассматриваемые переменные измерены, по крайней мере, в интервальной шкале. Она определяет степень, с которой значения двух переменных "пропорциональны" друг другу. Важно, что значение коэффициента корреляции не зависит от масштаба измерения. Например, корреляция между ростом и весом будет одной и той же, независимо от того, проводились измерения в *дюймах* и *фунтах* или в *сантиметрах* и *килограммах*. *Пропорциональность* означает просто *линейную зависимость*. Корреляция высокая, если на графике зависимость "можно представить" прямой линией (с положительным или отрицательным углом наклона).

Проведенная прямая называется *прямой регрессии* или прямой, построенной *методом наименьших квадратов*. Последний термин связан с тем, что сумма *квадратов* расстояний (вычисленных по оси Y) от

наблюдаемых точек до прямой является минимальной. Заметим, что использование *квадратов* расстояний приводит к тому, что оценки параметров прямой сильно реагируют на выбросы.

Как интерпретировать значения корреляций. Коэффициент корреляции Пирсона (r) представляет собой меру линейной зависимости двух переменных. Если возвести его в квадрат, то полученное значение коэффициента детерминации r^2) представляет долю вариации, общую для двух переменных (иными словами, "степень" зависимости или связанности двух переменных). Чтобы оценить зависимость между переменными, нужно знать как "величину" корреляции, так и ее значимость.

Значимость корреляций. Уровень значимости, вычисленный для каждой корреляции, представляет собой главный источник информации о надежности корреляции. Как объяснялось выше, значимость определенного коэффициента корреляции зависит от объема выборок. Критерий значимости основывается на предположении, что распределение остатков (т.е. отклонений наблюдений от регрессионной прямой) для зависимой переменной *y* является нормальным (с постоянной дисперсией для всех значений независимой переменной *x*).

Выбросы. По определению, выбросы являются нетипичными, резко выделяющимися наблюдениями. Так как при построении прямой регрессии используется сумма *квадратов* расстояний наблюдаемых точек до прямой, то выбросы могут существенно повлиять на наклон прямой и, следовательно, на значение коэффициента корреляции. Поэтому единичный выброс (значение которого возводится в квадрат) способен существенно изменить наклон прямой и, следовательно, значение корреляции.

Заметим, что если размер выборки относительно мал, то добавление или исключение некоторых данных (которые, возможно, не являются "выбросами", как в предыдущем примере) способно оказать существенное

влияние на прямую регрессии (и коэффициент корреляции). Это показано в следующем примере, где мы назвали исключенные точки "выбросами"; хотя, возможно, они являются не выбросами, а экстремальными значениями.

Обычно считается, что выбросы представляют собой случайную ошибку, которую следует контролировать. К сожалению, не существует общепринятого метода автоматического удаления выбросов (тем не менее, см. следующий раздел). Чтобы не быть введенными в заблуждение полученными значениями, необходимо проверить на диаграмме рассеяния каждый важный случай значимой корреляции. Очевидно, выбросы могут не только искусственно увеличить значение коэффициента корреляции, но также реально уменьшить существующую корреляцию.

Количественный подход к выбросам. Некоторые исследователи применяют численные методы удаления выбросов. Например, исключаются значения, которые выходят за границы ±2 стандартных отклонений (и даже ±1.5 стандартных отклонений) вокруг выборочного среднего. В ряде случаев такая "чистка" данных абсолютно необходима. К сожалению, в общем случае, определение выбросов субъективно, и решение должно приниматься индивидуально в каждом эксперименте (с учетом особенностей эксперимента или "сложившейся практики" в данной области). Следует заметить, что в некоторых случаях относительная частота выбросов к численности групп может быть исследована и разумно проинтерпретирована с точки зрения самой организации эксперимента.

ПРИМЕР

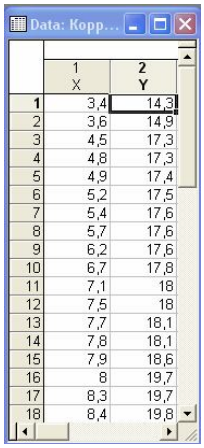


Рис.9.1. Исходные данные

X–независимая переменная

Y–зависимая переменная

Проведем анализ в модуле Основная статистика (Basic statistics/Tables).

Рассмотрим и установим связь между X и Y.

Шаг 1. Из Переключателя модулей STATISTICA откройте модуль **Основная статистика – Basic statistics/Tables**. Высветите название модуля и далее щелкните мышью по названию модуля: **Basic statistics/Tables** (рис. 9.2).

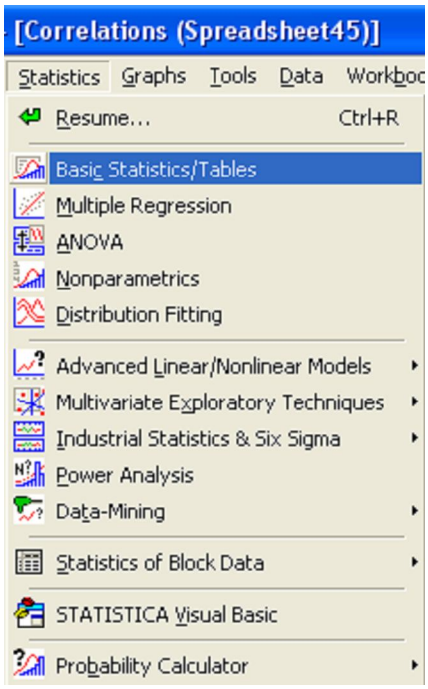


Рис. 9.2. Стартовая панель модуля Basic statistics/Tables

Шаг 2. На экране появится (рис. 9.3). Щелкните мышью по названию **Корреляционная матрица – Correlation matrics.**

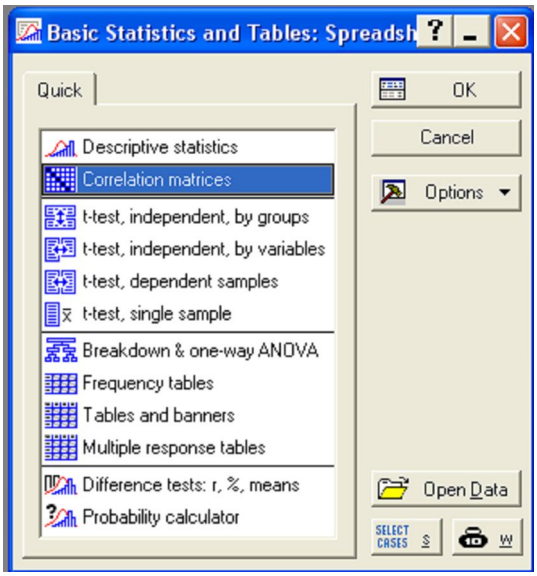


Рис. 9.3.

Шаг 3. Выберите переменные для анализа. Выбор переменных осуществляется с помощью кнопки **Two list,** находящейся в центре верхней части панели (рис. 9.4).

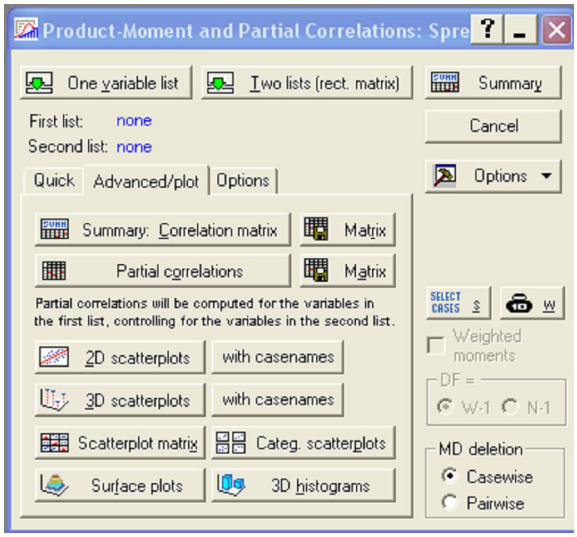


Рис. 9.4. Стартовая панель модуля **Correlation matrics**.

После того как кнопка будет нажата, диалоговое окно **Выбрать списки зависимых и независимых переменных — Select one or two variable list** — появится на вашем экране (рис. 9.5).

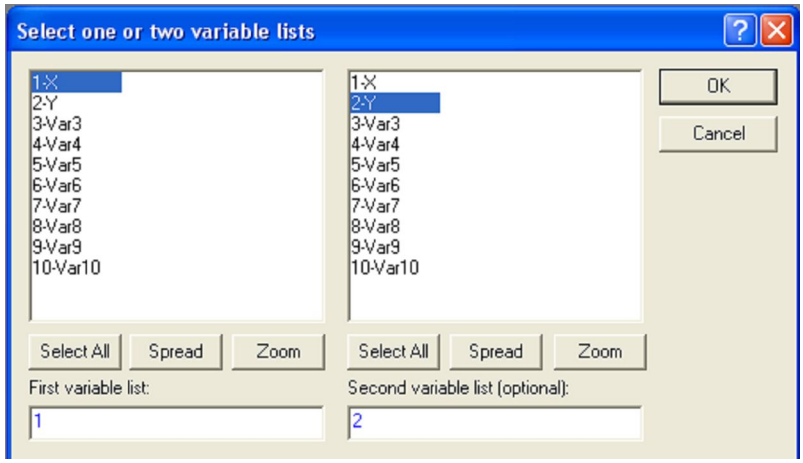


Рис. 9.5. Окно выбора переменных для анализа

Шаг 4. Высветив имя переменной в правой части окна, выберите переменную в левой части окна.

После нажатия кнопки ok в режиме **Options** выполните установки, показанные на рис. 9.6, подсветив **Displey detaled table of results.**

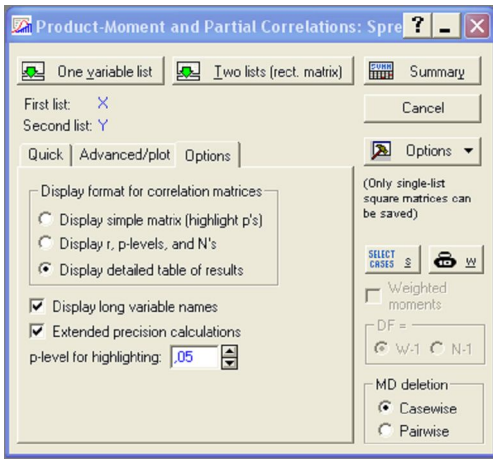


Рис. 9.6

Шаг 5. После нажатия кнопки **Summary** программа произведет расчеты корреляции межу X и Y, и через секунду на экране появится следующее окно результатов (рис. 9.7):

среднее;

стандартное отклонение;

значение коэффициента корреляции r;

значение коэффициента детерминации r2;

t–критерий;

p – уровень значимости;

число коррелируемых пар;

свободный член – 12,66835.

коэффициент при независимой переменной – 0.815797.

В нашем примере r = 0.98... Это очень хорошее значение (подсвечено красным цветом), показывающее, что построенная регрессия объясняет более 90% разброса значений переменной X относительно среднего.

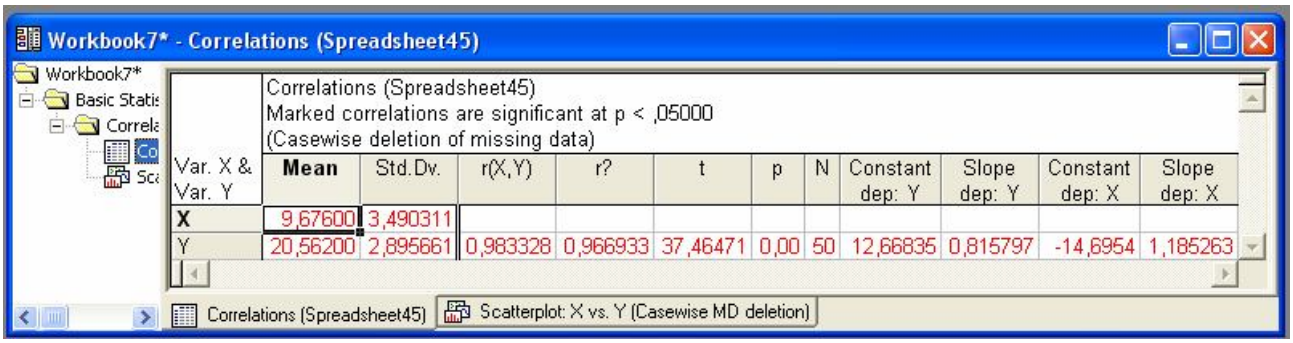


Рис. 9.7. Результат расчета корреляции

Из таблицы мы видим, что оцененная модель имеет вид:

Y = 0.815797*X +12,66835

На графике данные с подогнанной прямой имеют вид (рис. 9.8):

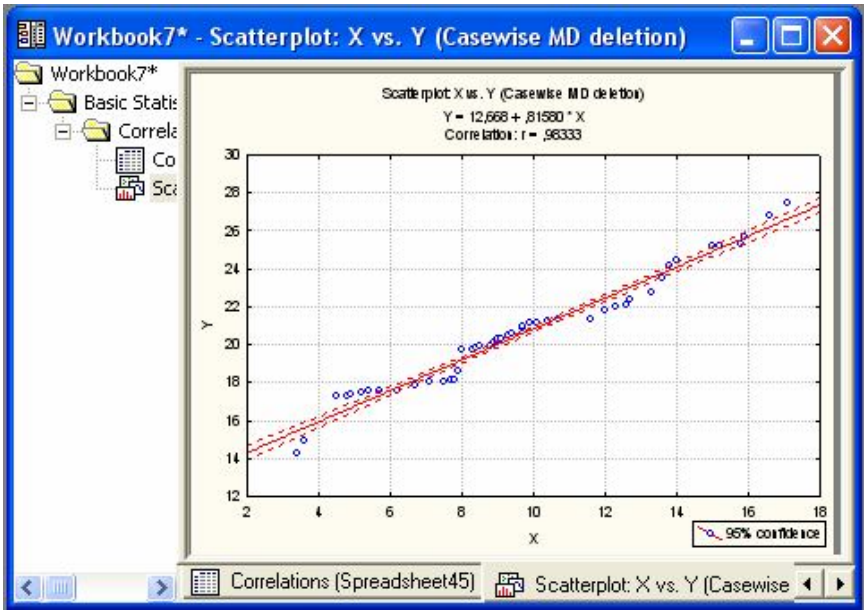


Рис. 9.8. Линейная регрессия для данных: X и Y

ТЕМА 9 РЕГРЕССИОННЫЙ АНАЛИЗ В СИСТЕМЕ STATISTICA — МОДУЛЬ MULTIPLE REGRESSION (МНОЖЕСТВЕННАЯ РЕГРЕССИЯ)

9.1 Описание модели

9.2 Постановка задачи

Типичной практической задачей является задача определения зависимостей в системе данных.

Предположим, вы наблюдаете значения пары переменных X и Y и хотите найти зависимость между ними.

Переменная X носит название **независимой переменной, или** предиктора, переменная Y называется зависимой **переменной,** или откликом.

Данная терминология связана с тем, что мы хотим определить именно зависимость Y от X или предсказать, какими будут значения Y при данных значениях X.

Значение переменной X в i-м опыте будем обозначать через X(i), соответствующее значение величины Y обозначим через Y(i), 0<i<=n.

Итак, вы наблюдаете значения независимой X(i) и соответствующие им значения зависимой Y(i), 0<i<=n, и хотите оценить зависимость Y от X. В статистике подобные задачи решаются в рамках регрессионной модели. Мы будем рассматривать самую простую регрессионную модель — линейную. Однако и в рамках этой модели могут быть решены самые разнообразные практические задачи.

Регрессионный анализ в системе STATISTICA проводится в модуле **Множественная регрессия (Multiple regression).**

9.1 Описание модели

Дадим точное описание линейной регрессионной модели, в рамках которой мы будем исследовать зависимость Y от X.

Мы постулируем, что наблюдаемые величины связаны между собой

регрессионной зависимостью вида:

Y(i) = Bl* X(i) + BO +e(i),

0<i<=n,

где **B1, B0** — неизвестные константы, **e(i)** — ненаблюдаемые случайные величины (наблюдаются только **X(i), Y(i), 0<i<=n) со средним 0** (как говорят, являются несмещенными) и неизвестной дисперсией, не меняющейся от опыта к опыту.

Иногда случайные величины **e(i),** 0<i<=n называют ошибками наблюдения. Относительно **e(i)** предполагается, что они не коррелированы в разных опытах. Кроме того, часто предполагается, что ошибки имеют нормальное распределение. В этом случае некоррелированность влечет независимость.

Можно рассматривать и более общие линейные модели, например, с несколькими независимыми переменными:

Y(i) = Bl*Xl(i) + B2*X2(i) + ... + BK *XK(i) + BO +e(i),

0<i<=n,

где **B0, B1,** B2, ... BK — неизвестные коэффициенты.

Эта модель также может исследоваться в модуле Множественная регрессия (Multiple regression).

9.2 Постановка задачи

Общая задача состоит в том, чтобы по наблюдениям (X(1),Y(2)), ... (X(n),Y(n)):

оценить параметры модели **B1, B0** наилучшим образом;

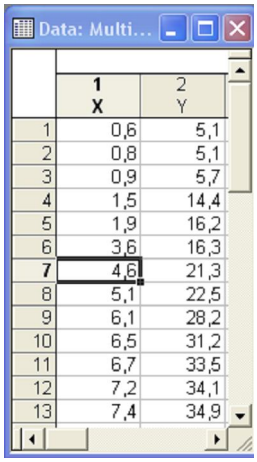
построить доверительные интервалы для **B1, B0;**

проверить гипотезу о значимости регрессии;

оценить степень адекватности модели и т.д.

Ниже на примерt показано, как решается данная задача в системе STATISTICA.

ПРИМЕР



	1 X	2 Y
1	0,6	5,1
2	0,8	5,1
3	0,9	5,7
4	1,5	14,4
5	1,9	16,2
6	3,6	16,3
7	4,6	21,3
8	5,1	22,5
9	6,1	28,2
10	6,5	31,2
11	6,7	33,5
12	7,2	34,1
13	7,4	34,9

Рис. 10.1. Исходные данные

X–независимая переменная

Y–зависимая переменная

Проведем анализ в модуле **Множественная регрессия.**

Рассмотрим и установим связь между X и Y.

Шаг 1. Из Переключателя модулей STATISTICA откройте модуль **Множественная регрессия — Multiple** regression. Высветите название модуля и далее щелкните мышью по названию модуля: **Multiple regression.**

Шаг 2. На экране появится стартовая панель модуля (рис. 10.2).

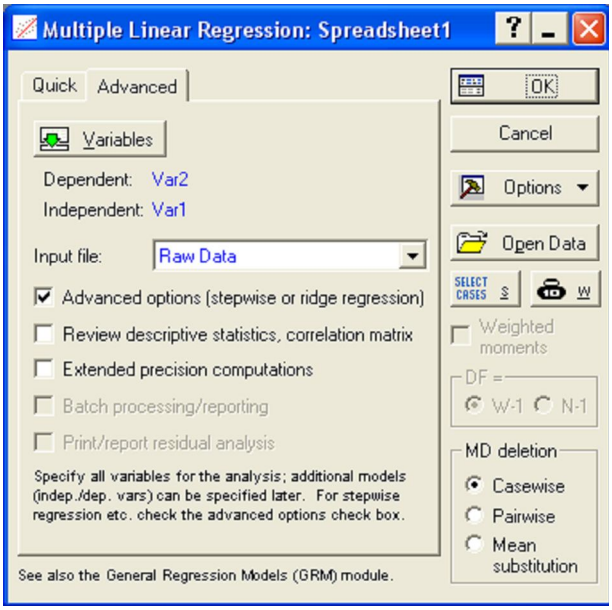


Рис. 10.2. Стартовая панель модуля Множественная регрессия

Выполните установки, как показано на рис. 10.2. Далее выберите

переменные для анализа. Выбор переменных осуществляется с помощью кнопки **Переменные (Variables),** находящейся в левом верхнем углу панели.

После того как кнопка будет нажата, диалоговое окно Выбрать списки зависимых и независимых переменных — Select dependent and independent variable list — появится на вашем экране (рис. 10.3).

Высветив имя переменной в левой части окна, выберите зависимую переменную. Высветив имя переменной в правой части окна, выберите независимую переменную.

В данном примере независимой переменной является X, зависимой — Y. Высветив имена этих переменных, как показано на рисунке, нажмите кнопку ОК в правом верхнем углу окна **Select dependent and independent variable list.** Вы вновь окажетесь в стартовой панели модуля. Переменные для анализа выбраны.

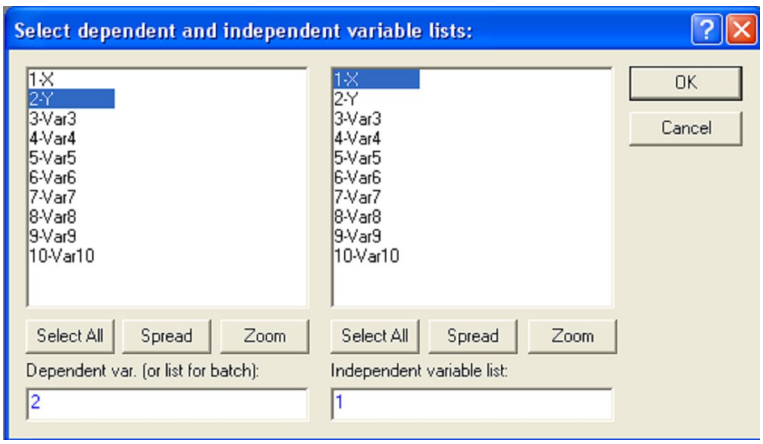


Рис. 10.3. Окно выбора переменных для анализа

Нажмите кнопку ОК в правом углу стартовой панели.

Шаг 3. На экране перед вами появится диалоговое окно **Построение модели — Model Definition** (рис. 10.4).

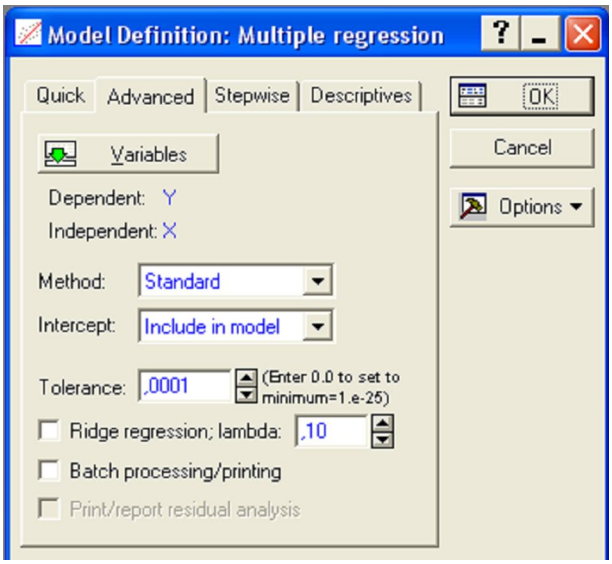


Рис. 10.4. Окно построения модели в модуле Множественная регрессия

В данном окне выберите стандартный метод оценивания, в опции **Method (Метод): Стандартный (Standard).** Далее нажмите кнопку ОК.

Программа произведет оценивание параметров модели стандартным методом, и через секунду на экране появится следующее диалоговое окно результатов.

Шаг 4. В диалоговом окне **Результаты Множественной регрессии — Multiple Regression Results** просмотрите результаты оценивания. Результаты можно просмотреть в численном и графическом виде.

Окно результатов анализа имеет следующую структуру: верх окна — информационный. Он состоит из двух частей: в первой части содержится основная информация о результатах оценивания, во второй высвечиваются значимые регрессионные коэффициенты. Внизу окна **Результаты множественной регрессии** находятся функциональные кнопки, позволяющие всесторонне просмотреть результаты анализа (рис. 10.5).

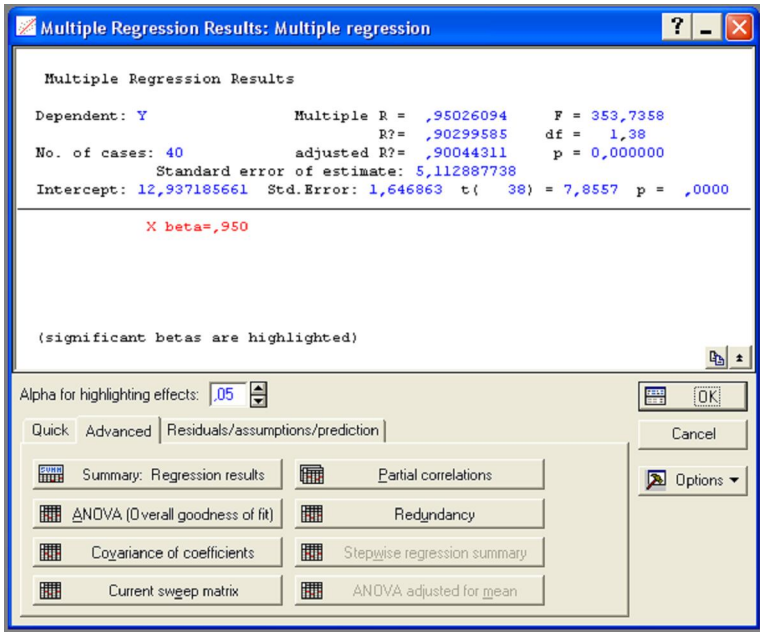


Рис. 10.5. Окно оценивания параметров в примере

Рассмотрим вначале информационную часть окна. В ней содержатся краткие сведения о результатах анализа. А именно:

Dep. Var. (Имя зависимой переменной). В данном случае – Y.

No. of Cases (Число случаев, по которым построена регрессия). В примере это число равно 40.

Multiple R (Коэффициент множественной корреляции).

R^2 (Квадрат коэффициента множественной корреляции), обычно называемый коэффициентом детерминации. Коэффициент детерминации является чрезвычайно важной характеристикой, поэтому мы подробно обсудим его. Коэффициент детерминации является одной из основных статистик в данном окне, он показывает долю общего разброса (относительно выборочного среднего зависимой переменной), которая объясняется построенной регрессией.

Adjusted R^2 (Скорректированный коэффициент детерминации), определяемый как:

Adjusted $R^2 = 1–(1– R^2)* n/(n-p)$,

где n — число наблюдений в модели, p — число параметров

модели (число независимых переменных плюс 1, так как в модель включен свободный член).

Std. Error of estimate (Стандартная ошибка оценки). Эта статистика является мерой рассеяния наблюдаемых значений относительно регрессионной прямой.

Intercept (Оценка свободного члена регрессии). Значение коэффициента B0 в уравнении регрессии.

Std. Error (Стандартная ошибка оценки свободного члена).Стандартная ошибка коэффициента B0 в уравнении регрессии.

t(df) and p-value (Значение t-критерия и уровень p). t-критерий используется для проверки гипотезы о равенстве 0 свободного члена регрессии.

F — значения F-критерия.

df — число степеней свободы F-критерия.

p — уровень значимости.

В информационной части посмотрим, прежде всего, на значения коэффициента детерминации. Значения коэффициента детерминации лежат в пределах от 0 до 1. В нашем примере RI = 0.90... Это очень хорошее значение, показывающее, что построенная регрессия объясняет более 90% разброса значений переменной X относительно среднего.

Далее посмотрите на значение F-критерия и уровень значимости p. F-критерий используется для проверки гипотезы о значимости регрессии. В данном случае для проверки гипотезы, утверждающей, что между зависимой переменной X и независимой переменной Y нет линейной зависимости, то есть B1 = 0, против альтернативы **B1** не равен 0. В данном примере мы имеем большое значение F-критерия — 353.7358 и даваемый в окне уровень значимости p = 0.0000, показывающие, что построенная регрессия высоко значима.

Рассмотрим вторую часть информационного окна. В этой части система сама говорит нам о значимых регрессионных коэффициентах, высвечивая строку: **X beta = 0.950** и на пояснение **значимые beta высвечены — significant beta's are highlighted.** Отметим, что в данном случае **beta** есть стандартизованный коэффициент **B1,** то есть коэффициент при независимой переменной **X.**

Перейдем в функциональную часть окна результатов.

Прежде всего, нажмите кнопку **Итоговый результат регрессии — Regression summary.** На экране появится электронная таблица вывода — **spredsheet,** в которой представлены итоговые результаты оценивания регрессионной модели (Рис. 10.6).

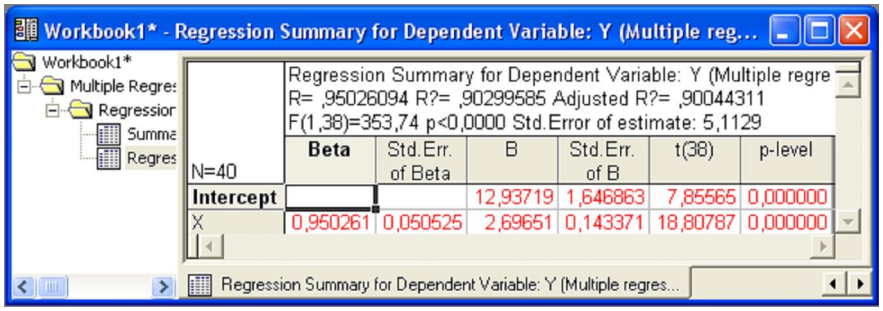


Рис. 10.6. Итоговая таблица регрессии

Это стандартная таблица вывода регрессионного анализа. В первом столбце таблицы даны значения коэффициентов **beta — стандартизованные коэффициенты регрессионного уравнения,** во втором — стандартные ошибки **beta,** в третьем –точечные оценки параметров модели:

Свободный член **B0 = 12,93719**.

Коэффициент B1 (при независимой переменной X) = 2,69651.

Далее, стандартные ошибки для BO, B1, значения статистик t-критерия и т.д.

Из таблицы мы видим, что оцененная модель имеет вид:

Y = 2,69651*X +12,93719

На графике данные с подогнанной прямой имеют вид (рис. 10.7):

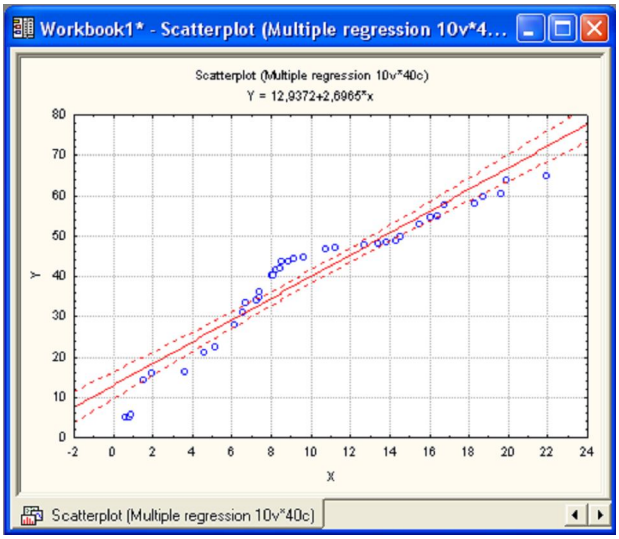


Рис. 10.7. Линейная регрессия для данных: X и Y

Шаг 5. Оценка адекватности модели.

Важным элементом анализа является оценка адекватности модели.

После того как доказана адекватность модели, полученные результаты можно уверенно использовать для дальнейших действий.

Анализ адекватности основывается на анализе остатков.

Остатки представляют собой **разности** между наблюдаемыми значениями и модельными, то есть значениями, подсчитанными по модели с оцененными параметрами.

В STATISTICA в модуле **Множественная регрессия** имеется специальное диалоговое окно, в котором проводится всесторонний анализ остатков.

Нажмите кнопку Анализ остатков — Residual Analysis.

Следующее диалоговое окно **Анализ остатков — Residual Analysis** появится на экране (рис. 10.8).

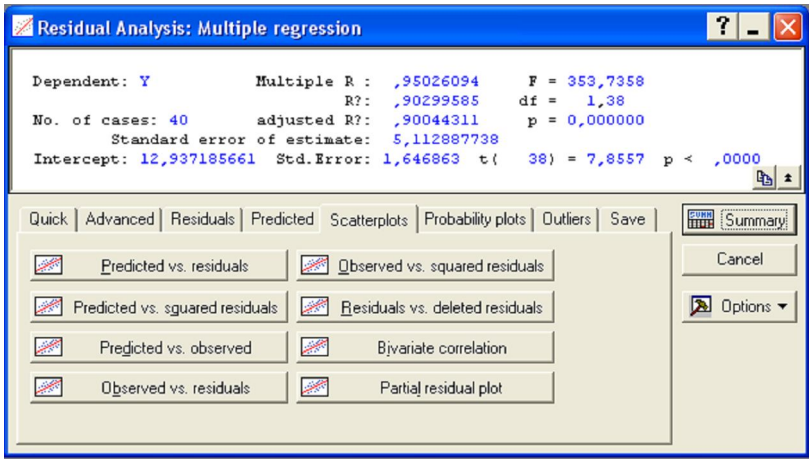


Рис. 10.8. Диалоговое окно Анализ остатков в модуле

Нажмите в этом окне, например, кнопку **Obs&residuals.** На экране появится график (рис. 10.9), который говорит о достаточной адекватности модели.

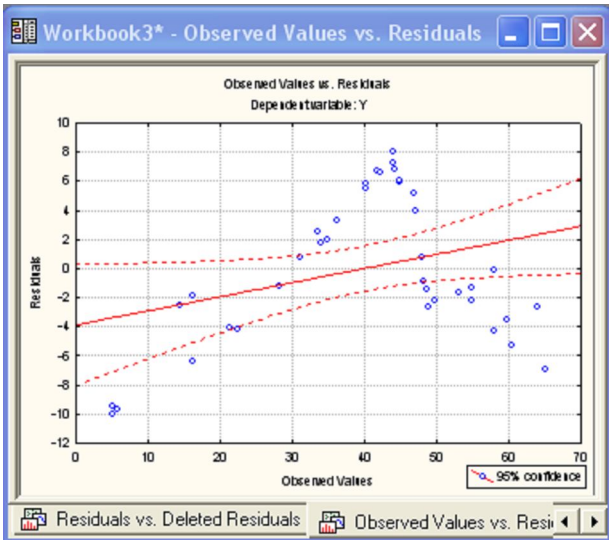


Рис. 10.9. График наблюдаемые переменные-остатки

Часто, если остатки не являются нормальными, а также для стабилизации дисперсии применяют преобразования зависимых и независимых переменных, например логарифмическое преобразование зависимых переменных или извлечение квадратного корня.

ТЕМА 10 ДИСПЕРСИОННЫЙ АНАЛИЗ

10.1 Однофакторный дисперсионный анализ

10.2 Двухфакторный анализ

Процедура сравнения средних называется дисперсионным анализом. В действительности, это связано с тем, что при исследовании статистической значимости различия между средними двух (или нескольких) групп, на самом деле сравниваются (т.е. анализируются) выборочные дисперсии. Фундаментальная концепция дисперсионного анализа предложена Фишером в 1920 году.

Для выборки объема n выборочная дисперсия вычисляется как сумма квадратов отклонений от выборочного среднего, деленная на *n-1* (объем выборки минус единица). Таким образом, при фиксированном объеме выборки *n* дисперсия есть функция суммы квадратов (отклонений), обозначаемая, для краткости, *SS* (от английского Sum of Squares - Сумма квадратов). В основе дисперсионного анализа лежит разделение дисперсии на части или компоненты. Рассмотрим следующий набор данных:

	Группа 1	Группа 2
Наблюдение 1 Наблюдение 2 Наблюдение 3	2 3 1	6 7 5
Среднее Сумма квадратов (СК)	2 2	6 2
Общее среднее Общая сумма квадратов	4 28	

Средние двух групп существенно различны (*2* и *6* соответственно). Сумма квадратов отклонений *внутри* каждой группы равна *2*. Складывая их, получаем 4. Если теперь повторить эти вычисления *без учета* групповой принадлежности, то есть, если вычислить *SS* исходя из общего среднего этих двух выборок, то получим величину *28*. Иными словами, дисперсия (сумма квадратов), основанная на внутригрупповой изменчивости, приводит к гораздо меньшим значениям, чем при вычислении на основе общей изменчивости (относительно общего

среднего). Причина этого, очевидно, заключается в существенной разнице между средними значениями, и это различие между средними и объясняет существующее различие между суммами квадратов. В самом деле, если использовать для анализа этих данных модуль *Дисперсионный анализ*, то будет получена следующая таблица, называемая таблицей дисперсионного анализа:

	ГЛАВНЫЙ ЭФФЕКТ				
	SS	ст. св.	MS	F	p
Эффект	24.0	1	24.0	24.0	.008
Ошибка	4.0	4	1.0		

Как видно из таблицы, общая сумма квадратов *SS* = 28 разбита на компоненты: сумму квадратов, обусловленную *внутригрупповой* изменчивостью (*2+2=4*; см. вторую строку таблицы) и сумму квадратов, обусловленную различием средних значений между группами (28–(2+2)=24; см первую строку таблицы). Заметим, что *MS* в этой таблице есть средний квадрат, равный *SS,* деленная на число степеней свободы (ст. св).

***SS* ошибок и *SS* эффекта.** Внутригрупповая изменчивость (*SS*) обычно называется остаточной компонентой или дисперсией *ошибки.* Это означает, что обычно при проведении эксперимента она не может быть предсказана или объяснена. С другой стороны, *SS эффекта* (или компоненту дисперсии между группами) можно объяснить различием между средними значениями в группах. Иными словами, принадлежность к некоторой группе *объясняет* межгрупповую изменчивость, т.к. нам известно, что эти группы обладают разными средними значениями.

Проверка значимости. Объясняются причины, по которым многие критерии используют отношение объясненной и необъясненной дисперсии. Примером такого использования является сам дисперсионный анализ. Проверка значимости в дисперсионном анализе основана на сравнении компоненты дисперсии, обусловленной межгрупповым

разбросом (называемой *средним квадратом эффекта* или $MS_{эффект}$) и компоненты дисперсии, обусловленной внутригрупповым разбросом (называемой *средним квадратом ошибки* или $MS_{ошибка}$. Если верна нулевая гипотеза (равенство средних в двух популяциях), то можно ожидать сравнительно небольшое различие выборочных средних из-за чисто случайной изменчивости. Поэтому, при нулевой гипотезе, внутригрупповая дисперсия будет практически совпадать с общей дисперсией, подсчитанной без учета групповой принадлежности. Полученные внутригрупповые дисперсии можно сравнить с помощью F-критерия, проверяющего, действительно ли отношение дисперсий значимо больше 1. В рассмотренном выше примере *F*-критерий показывает, что различие между средними статистически значимо (значимо на уровне 0.008).

Подводя итоги, можно сказать, что целью дисперсионного анализа является проверка статистической значимости различия между средними (для групп или переменных). Эта проверка проводится с помощью разбиения суммы квадратов на компоненты, т.е. с помощью разбиения общей дисперсии (вариации) на части, одна из которых обусловлена случайной ошибкой (то есть внутригрупповой изменчивостью), а вторая связана с различием средних значений. Последняя компонента дисперсии затем используется для анализа статистической значимости различия между средними значениями. Если это различие *значимо*, нулевая гипотеза *отвергается* и принимается альтернативная гипотеза о существовании различия между средними.

Взаимодействия высших порядков. В то время как объяснить попарные взаимодействия еще сравнительно легко, взаимодействия высших порядков объяснить значительно сложнее.

В общем случае взаимодействие между факторами описывается в виде изменения одного эффекта под воздействием другого. Двухфакторное

взаимодействие можно описать как изменение главного эффекта фактора, характеризующего сложность задачи, под воздействием второго фактора. Для взаимодействия трех факторов можно сказать, что взаимодействие двух факторов изменяется под воздействием третьего. Если изучается взаимодействие четырех факторов, можно сказать, что взаимодействие трех факторов, изменяется под воздействием четвертого фактора, т.е. существуют различные типы взаимодействий на разных уровнях четвертого фактора. Оказалось, что во многих областях взаимодействие пяти или даже большего количества факторов не является чем-то необычным.

10.1 Однофакторный дисперсионный анализ

Ввести исходные данные из табл 4.1, как показано на рис. 10.1.

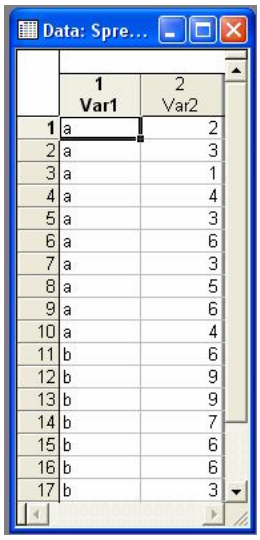


Рис. 10.1. Исходные данные

Var2–независимая переменная;

Var1–факторы.

Проведем анализ в модуле **ANOVA.**

Шаг 1. Из Переключателя модулей STATISTICA откройте модуль **ANOVA**. Высветите название модуля и далее щелкните мышью по названию модуля: **ANOVA** (Рис. 10.2).

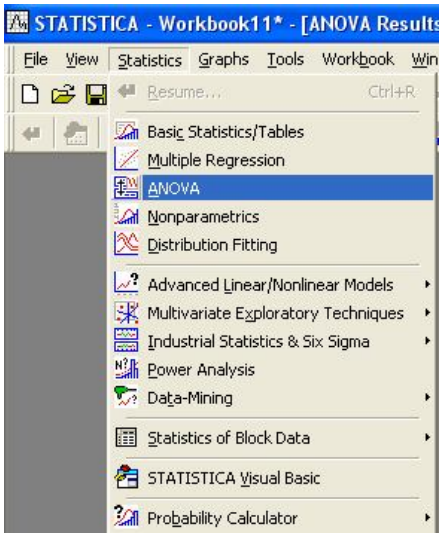


Рис. 10.2. Основное меню

Шаг 2. На экране появится стартовая панель модуля (рис. 10.3). Выполнить установки, как показано на рис. 10.3.

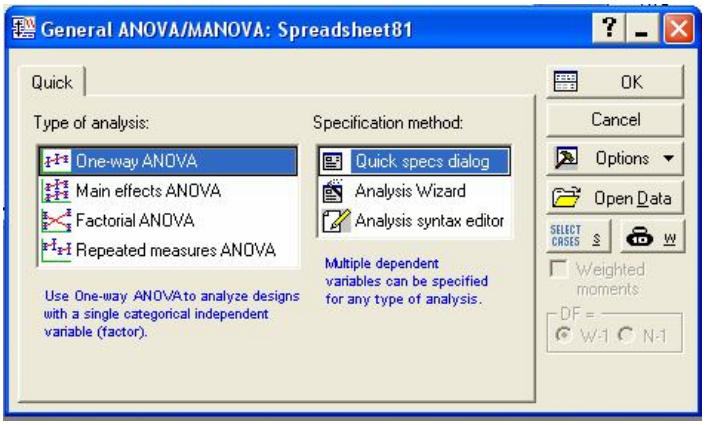


Рис. 10.3. Стартовая панель модуля

Шаг 3. После нажатия кнопки ok в появившемся окне выберите переменные для анализа (рис. 10.4.).

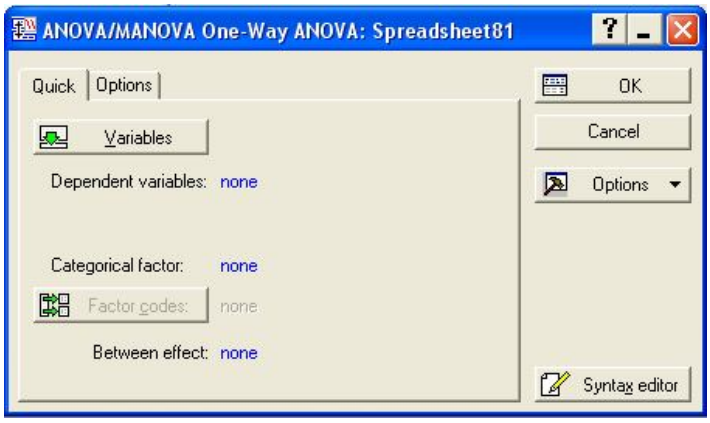


Рис. 10.4.

Выбор переменных осуществляется с помощью кнопки **Переменные (Variables),** находящейся в левом верхнем углу панели. После того как кнопка будет нажата, диалоговое окно **Выбрать списки зависимых переменных и факторов — Select dependent variables and categorical predictor (factor)** — появится на вашем экране (рис. 10.5).

Шаг 4. В левой части окна имя переменной выберите зависимую переменную, а в правой – фактор. После нажатия кнопки ok в появившемся окне выберите **All**, нажав кнопку **Factor codes** (рис. 10.6.).

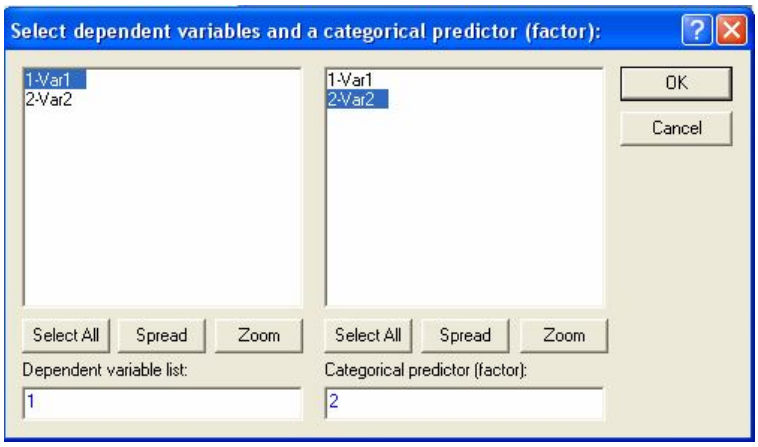


Рис. 10.5. Окно выбора переменных для анализа

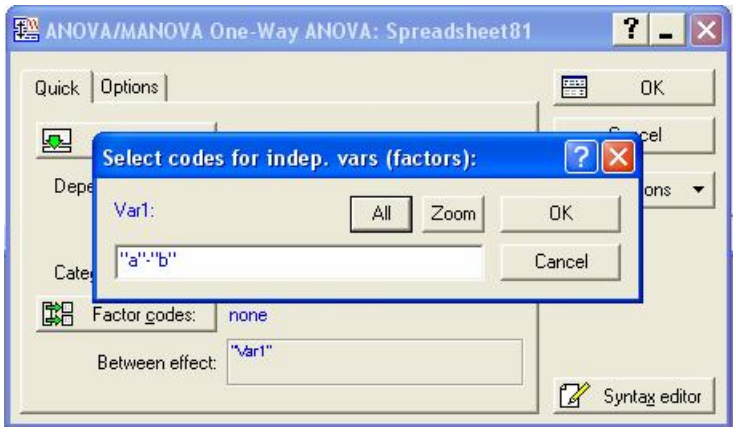


Рис. 10.6.

Нажмите кнопку ОК в правом углу стартовой панели.

Шаг 5. На экране перед вами появится диалоговое окно **Результаты — Anova Results** (рис. 10.7). В данном окне выберите **Univariate Results (Результат дисперсионного анализа).** Далее нажмите кнопку ОК.

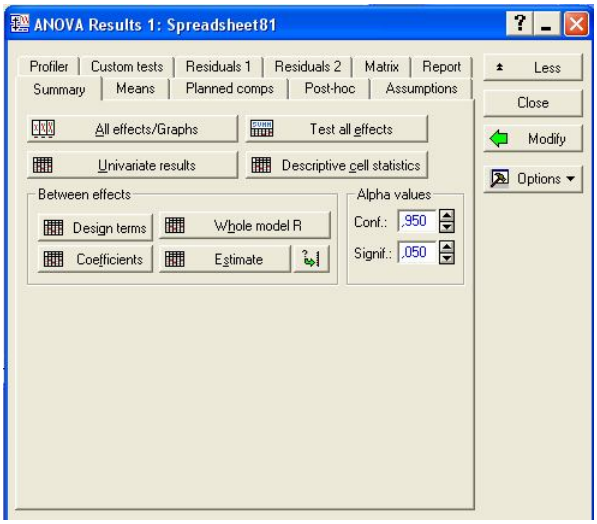


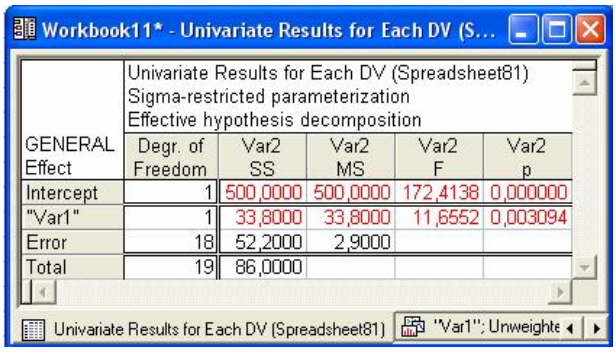
Рис. 10.7. Диалоговое окно результатов

В окне результатов рис. 10.8 представлены результаты дисперсионного анализа:

между группами – Var1;

внутри групп – Error.

В рассмотренном примере *F*-критерий показывает, что различие между средними статистически значимо (значимо на уровне 0.003). Поскольку различие между средними значениями *значимо*, нулевая гипотеза *отвергается* и принимается альтернативная гипотеза о существовании различия между средними (Результат в строке: Между группами – Var1 подсвечивается красным цветом).



GENERAL Effect	Degr. of Freedom	Var2 SS	Var2 MS	Var2 F	Var2 p
Intercept	1	500,0000	500,0000	172,4138	0,000000
"Var1"	1	33,8000	33,8000	11,6552	0,003094
Error	18	52,2000	2,9000		
Total	19	86,0000			

Рис. 10.8. Результаты дисперсионного анализа

Визуализация результатов дисперсионного анализа представлена на рис. 10.9 и 10.11.

Шаг 6. В диалоговом окне результатов (Рис. 10.7) нажать кнопку **All**

effects / Grafs, затем в появившемся окне ok. Результат представлен на рис. 10.9.

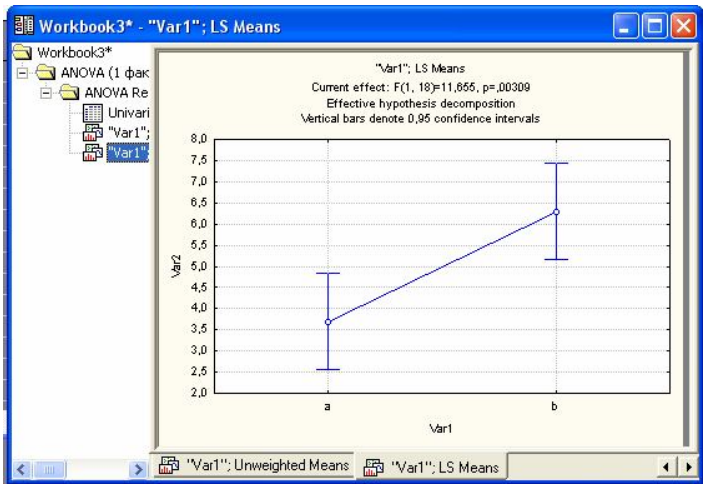


Рис. 10.9.

Шаг 6. Из основного меню (рис. 10.10) выберете модуль **Графика – Graphs,** выберете опцию **Box plots** и щелкните мышкой. Затем выполните установки, как это показано на рис. 10.11.

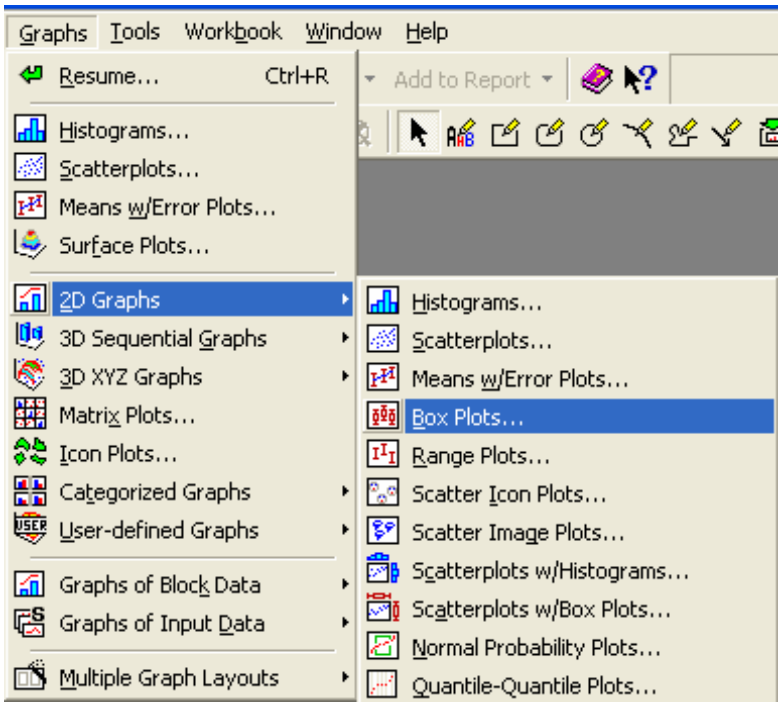


Рис. 10.10.

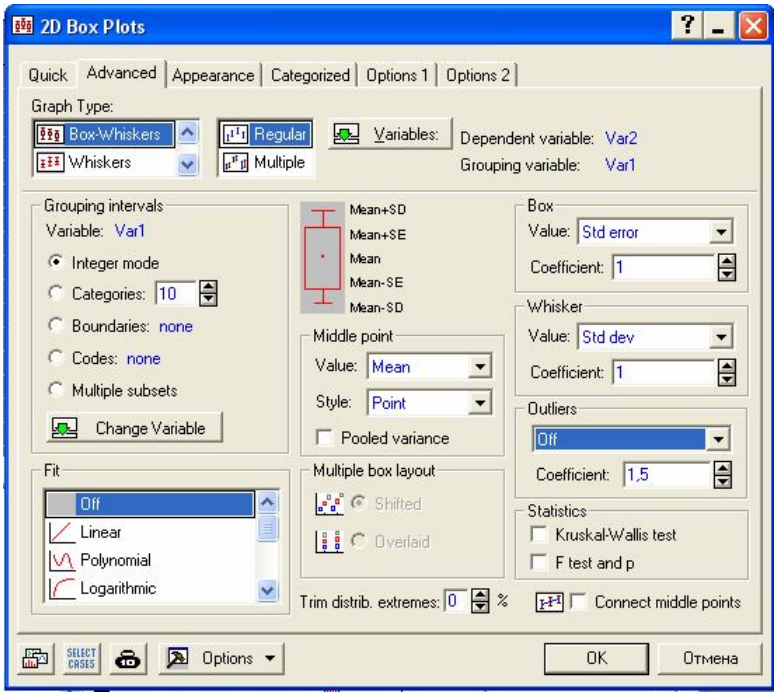


Рис. 10.11.

Нажмите кнопку ok и вы получите визуализацию статистических параметров при воздействии факторов a и b (рис. 10.12), для каждого из которых показаны:

среднее;

стандартное отклонение;

стандартная ошибка.

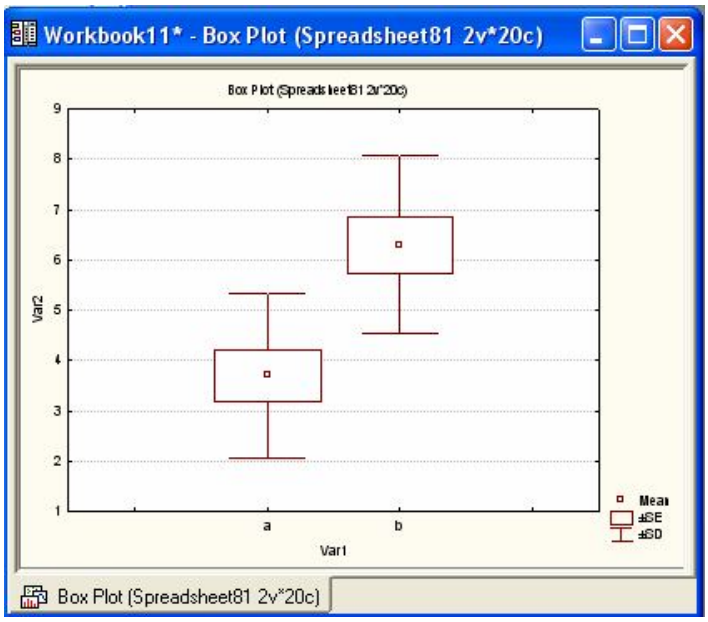


Рис. 10.12.

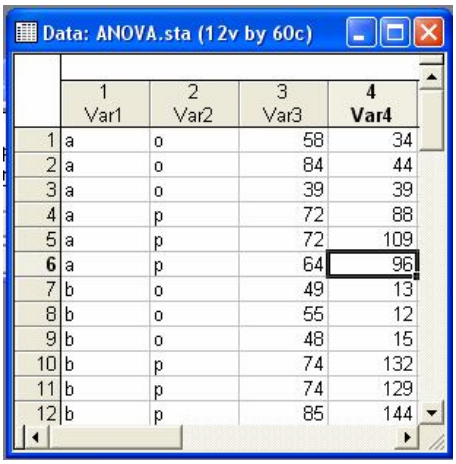
10.2 Двухфакторный анализ

Ввести исходные данные из табл 5.1, как показано на рис. 10.13.

Var3–независимая переменная;

Var1, Var2 –Факторы.

Проведем анализ в модуле **ANOVA.**



	1 Var1	2 Var2	3 Var3	4 Var4
1	a	o	58	34
2	a	o	84	44
3	a	o	39	39
4	a	p	72	88
5	a	p	72	109
6	a	p	64	96
7	b	o	49	13
8	b	o	55	12
9	b	o	48	15
10	b	p	74	132
11	b	p	74	129
12	b	p	85	144

Рис. 10.13. Исходные данные

Шаг 1. Из Переключателя модулей STATISTICA откройте модуль **ANOVA**. Высветите название модуля и далее щелкните мышью по названию модуля: **ANOVA** (Рис. 10.14).

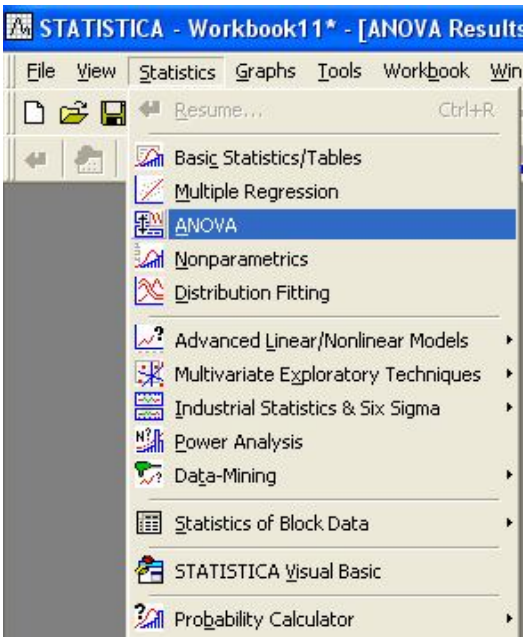


Рис. 10.14.

Шаг 2. На экране появится стартовая панель модуля (рис. 10.15).

Выполнить установки, как показано на рис. 10.15.

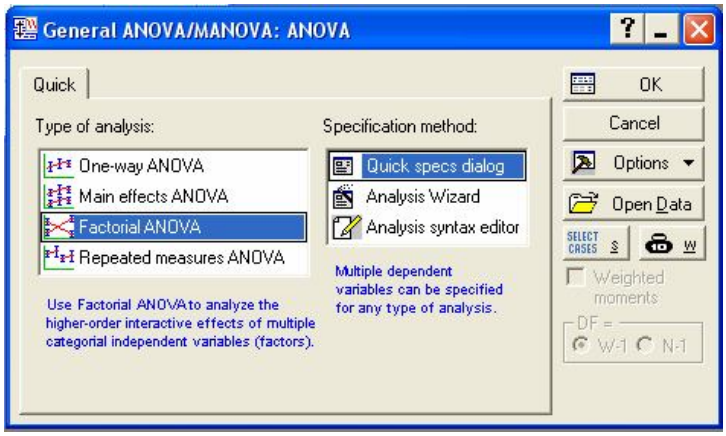


Рис. 10.15. Стартовая панель модуля

Шаг 3. После нажатия кнопки ok в появившемся окне выберите переменные для анализа (рис. 10.16.). Выбор переменных осуществляется с помощью кнопки **Переменные (Variables),** находящейся в левом верхнем углу панели. После того как кнопка будет нажата, диалоговое окно **Выбрать списки зависимых переменных и факторов — Select dependent variables and categorical predictor (factor)** — появится на вашем экране (рис. 10.17).

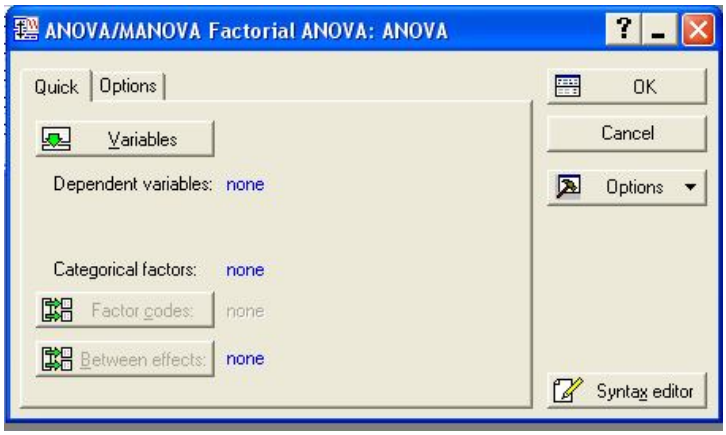


Рис. 10.16.

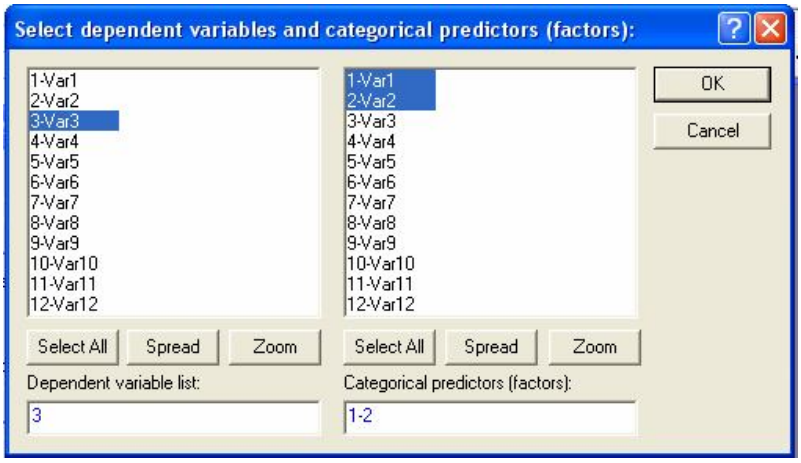


Рис. 10.17.

Шаг 4. В левой части окна имя переменной выберите зависимую переменную, а в правой – факторы. После нажатия кнопки ok в появившемся окне выберите **All**, нажав кнопку **Factor codes** (рис. 10.18.).

Нажмите кнопку ОК в правом углу стартовой панели.

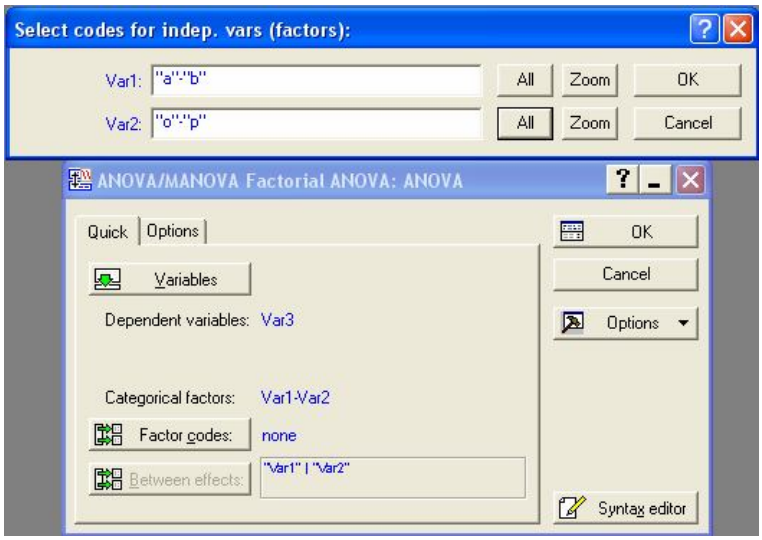


Рис. 10.18.

Шаг 5. На экране перед вами появится диалоговое окно **Результаты — Anova Results** (рис. 10.19). В данном окне выберите **Univariate Results (Результат дисперсионного анализа).** Далее нажмите кнопку ОК.

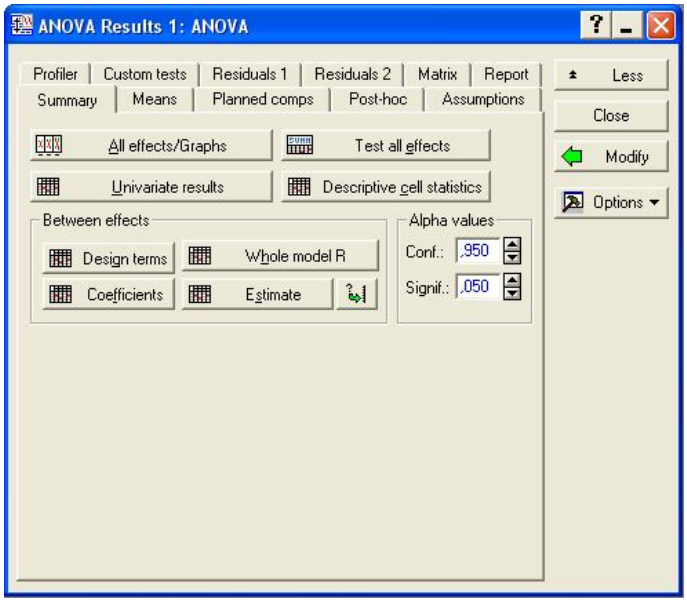


Рис. 10.19. Диалоговое окно результатов

В окне результатов рис. 10.20 представлены результаты дисперсионного анализа:

между группами, фактор 1 – Var1;

между группами, фактор 2 – Var2;

взаимодействие – Var1*Var2;

внутри групп – Error.

В рассмотренном примере *F*-критерий показывает, что различие между средними статистически значимо за счет влияния второго фактора (значимо на уровне 0.033). Сила влияния этого фактора составляет около 40%. (Результат в строке: Между группами фактор 2 – Var2 подсвечивается красным цветом).

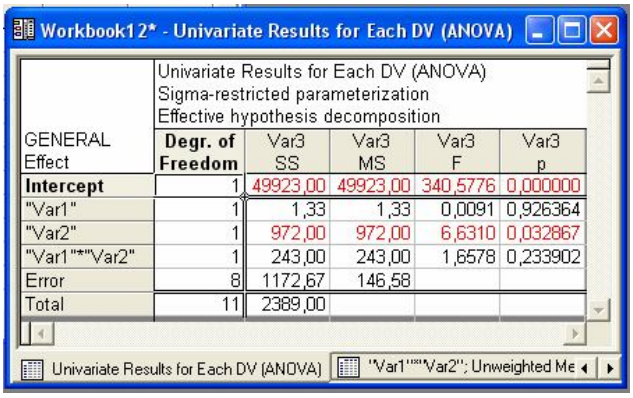


Рис. 10.20. Результаты дисперсионного анализа

Визуализация результатов дисперсионного анализа представлена на рис. 10.21, 10.23 и 10.24.

Шаг 6. В диалоговом окне результатов (Рис. 10.19) нажать кнопку **All effects / Grafs**, затем в появившемся окне ok. Результат представлен на рис. 10.21.

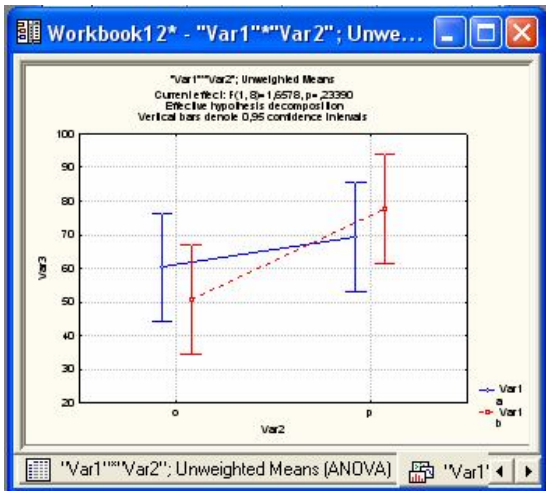


Рис. 10.21.

Шаг 6. Из основного меню (рис. 10.22) выберете модуль **Графика – Graphs,** выберете опцию **Box plots** и щелкните мышкой. Затем выполните установки, как это показано на рис. 10.23.

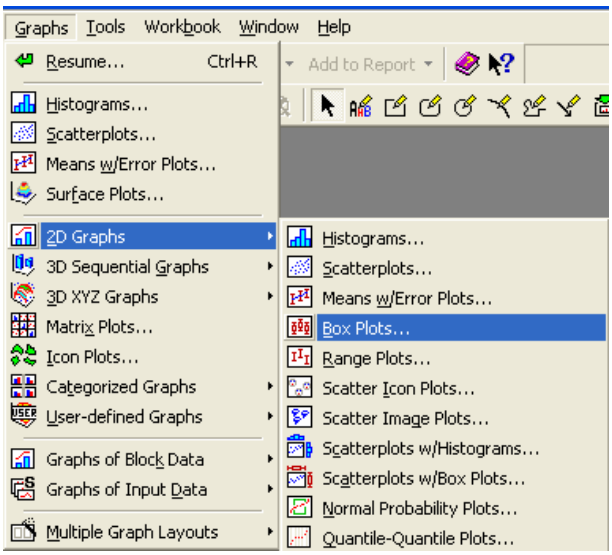


Рис. 10.22.

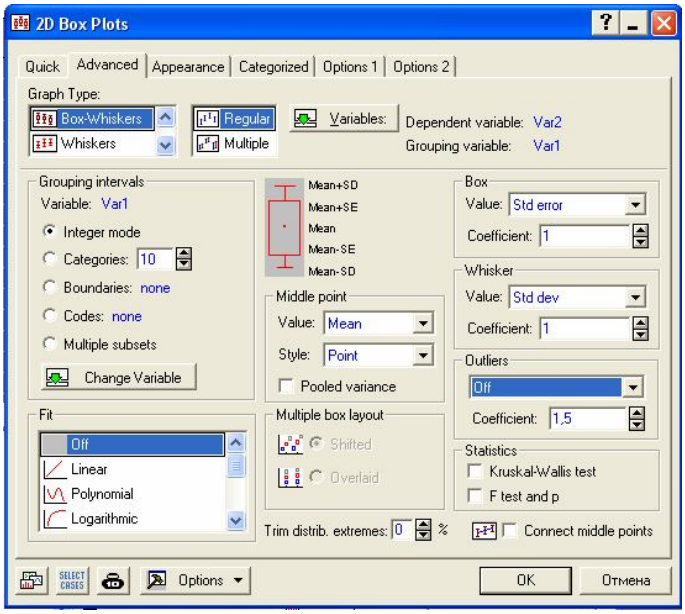


Рис. 10.23.

Нажмите кнопку ok и вы получите визуализацию статистических параметров при воздействии факторов (рис. 10.24, 10.25) для каждого из которых показаны:

среднее;

стандартное отклонение;

стандартная ошибка.

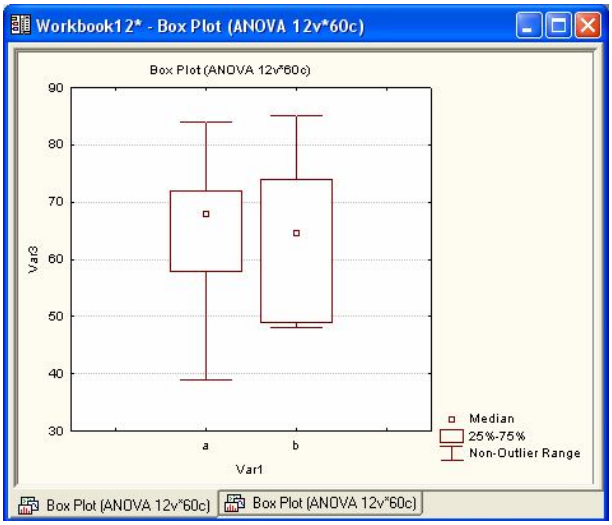


Рис. 10.24.

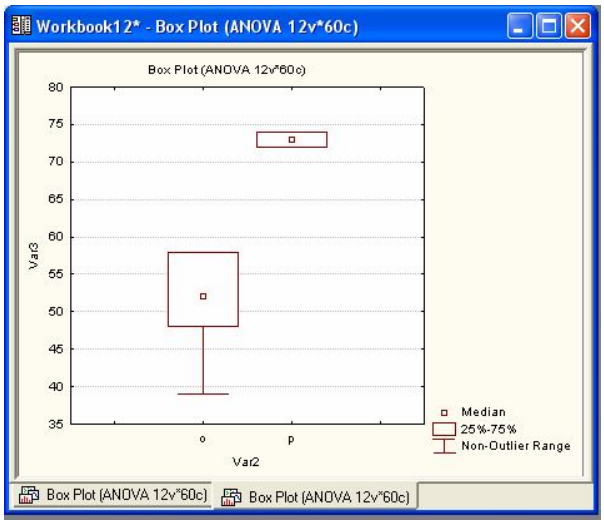


Рис. 10.25.

ТЕМА 11 КЛАССИФИКАЦИЯ ДАННЫХ В СИСТЕМЕ STATISTICA. МОДУЛЬ DISCRIMINANT ANALYSIS (ДИСКРИМИНАНТНЫЙ АНАЛИЗ)

11.1 Постановка задачи, методы решения, ограничения

11.2 Предположения и ограничения

11.3 Классификация цветов ириса

Дискриминантный анализ является одним из методов многомерного статистического анализа. Цель дискриминантного анализа состоит в том, чтобы на основе измерения различных характеристик (признаков, параметров) объекта классифицировать его, то есть отнести к одной из нескольких групп (классов) некоторым оптимальным способом. Под оптимальным способом понимается либо минимум математического ожидания потерь, либо минимум вероятности ложной классификации. Этот вид анализа является многомерным, так как измеряется несколько параметров объекта, по крайней мере, больше одного, например, температура, влажность в технологическом процессе, давление, состав крови, температура больного и т.д.

Типичные области применения дискриминантного анализа — биология, медицина, управление производством, экономика, геология, контроль качества.

В медицине объектом исследования является пациент, когда по результатам измерений различных параметров, проведения диагностических тестов врач определяет, например, необходимо ли хирургическое вмешательство при лечении.

В управлении производством принимается решение по отнесению поступающего сырья или продукции к одному из нескольких типов.

В экономике важно решение по отнесению клиента к определенному классу при выдаче кредита.

Широкий круг задач, возникающих на практике и связанных с

классификацией, можно решить методами дискриминантного анализа.

В модуле **Discriminant analysis (Дискриминантный анализ)** системы STATISTICA имеется широкий набор средств, обеспечивающих проведение дискриминантного анализа данных, визуализации и интерпретации результатов.

11.1 Постановка задачи, методы решения, ограничения

Предположим, имеется n объектов с m характеристиками. В результате измерений каждый объект характеризуется вектором $x_1 \ldots x_m$, $m > 1$. Задача состоит в том, чтобы по результатам измерений отнести объект к одной из нескольких групп (классов) $G_1, \ldots G_k$, $k \geq 2$. Иными словами, нужно построить решающее правило, позволяющее по результатам измерений параметров объекта указать группу, к которой он принадлежит. Число групп заранее известно, также известно, что объект заведомо принадлежит к определенной группе.

Пусть X – пространство значений вектора измерений. Решающее правило называется *нерандомизированным*, если пространство X разбито на k непересекающихся областей; при попадании измерения параметров объекта в k-ю область объект относится к k-й группе.

Решающее правило называется *рандомизированным*, если для каждого вектора наблюдений x задана вероятность pi(x), с которой объект принадлежит i-й группе, $p_i(x) >= 0$, $p_1(x) + \ldots + p_k(x) = 1$, i=1,...k.

Очевидно, при использовании решающего правила возникают потери, вызванные тем, что объект неправильно классифицирован — отнесен к классу i, когда в действительности он принадлежит классу j (i не равно j).

Если можно измерить убыток r(ij) при неправильной классификации объекта, то вводят средние потери, к которым приводит применение данного правила, и пытаются найти правило, минимизирующее эти средние потери.

Если значение потерь трудно оценить численно, то при построении

оптимального правила используют критерий минимальной вероятности ложной классификации.

В дискриминантном анализе можно задать априорные вероятности принадлежности объекта к определенному классу. На практике эти вероятности оцениваются из массива экспериментальных данных.

Так как массив экспериментальных данных накапливается, то эти оценки постепенно уточняются. При этом можно учесть различные факторы, влияющие на принадлежность объекта к определенному классу, например, если поступает мука в хлебное производство, то можно учесть сезонные факторы: вероятность того, что мука будет лучшего качества осенью выше той же вероятности весной.

В случае двух групп объектов дискриминантный анализ эквивалентен множественной регрессии (зависимой переменной является номер группы).

Независимые переменные с наибольшими стандартизированными коэффициентами регрессии дают наибольший вклад в предсказание принадлежности объекта к группе.

В модуле **Discriminant analysis (Дискриминантный анализ)** реализовано два общих метода дискриминантного анализа: стандартный и пошаговый (включения и исключения). Данные методы дискриминантного анализа аналогичны методам множественной регрессии. В случае двух групп методом наименьших квадратов строится регрессионная прямая (зависимая переменная — номер группы, все остальные переменные – независимые). Если групп несколько, то можно представить себе, что вначале строится дискриминация между группами 1 и 2, затем между 2 и 3, и так далее.

В пошаговом методе модель строится последовательно по шагам. Для метода включения STATISTICA на каждом шаге оценивает вклад в функцию дискриминации не включенных в модель переменных.

Переменная, дающая наибольший вклад, включается в модель, далее система переходит к следующему шагу. Если применяется так называемый пошаговый метод исключения, то вначале в модель включаются все переменные, затем производится их последовательное исключение.

Близкими к методам дискриминантного анализа являются методы дисперсионного анализа, кластерного и факторного анализов, а также, как уже говорилось, методы множественной регрессии. Отличие кластерного анализа от дискриминантного в том, что в нем заранее не фиксировано число групп (кластеров).

11.2 Предположения и ограничения

Дискриминантный анализ «работает» при выполнении ряда предположений.

Предположение о том, что наблюдаемые величины – измеряемые характеристики объекта – имеют нормальное распределение. Это предположение следует проверять. В модуле имеются специальные опции, позволяющие быстро построить гистограммы и графики на вероятностной бумаге. Специальные тесты на нормальность имеются в модуле ANOVA/ MANOVA. Следует заметить, что умеренные отклонения от этого предположения не являются фатальными.

Предположение об однородности дисперсий и ковариаций наблюдаемых переменных в разных классах (отличие между классами имеется только в средних). Умеренные отклонения от этого предположения также допустимы.

Широкий набор статистик и опций для тестирования различных предположений дискриминантного анализа, в частности так называемый М-критерий Бокса, содержится также в модуле **ANOVA/MANOVA – Дисперсионный анализ.**

Методы, реализованные в модуле, являются линейными. Функции классификации и дискриминантные функции являются линейными

комбинациями наблюдаемых величин.

Сделаем важное замечание о проверке предположений анализа. Дискриминантный анализ может быть проведен и когда основные предположения не выполняются (предположение о нормальности и равенстве ковариационных матриц). Задача состоит в интерпретации результатов. В конечном счете, наиболее важным критерием правильности построенного классификатора является практика. И если окажется, что в результате построен классификатор, «работающий» на практике, то это будет достижением.

В связи с этим мы рекомендуем проводить проверку с разумной степенью точности, сосредоточив основные усилия на построении модели.

11.3 Классификация цветов ириса

Знакомство с возможностями проведения дискриминантного анализа в системе STATISTICA лучше всего начать с разбора апробированного примера. Таким примером является классический пример Фишера – анализа цветков ириса.

Задача состоит в том, чтобы по результатам измерения длины и ширины чашелистиков и лепестков цветков ириса отнести ирис к одному из трех типов: SETOSA, VERSICOL, VIRGINIC.

Данные для этого примера имеются в файле *Irisdat.sta.* В файле содержатся результаты измерений 150 цветков ириса, по 50 каждого типа.

Шаг 1. Из Переключателя модулей STATISTICA откройте стартовую панель модуля **Discriminate function analysis (Дискриминантный функциональный анализ)** (рис. 11.1)**.**

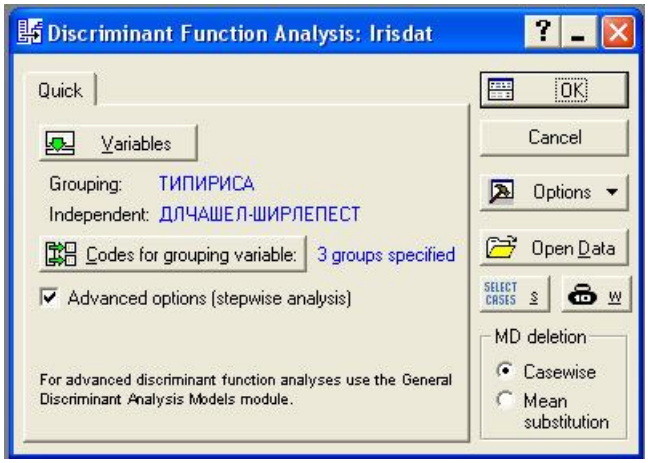


Рис. 11.1. Стартовая панель модуля Дискриминантный анализ

Шаг 2. Нажмите кнопку Open Data (Открыть данные) и откройте файл данных *Irisdat.sta* из каталога *Examples*. Следующий файл данных появится на экране (11.2).

Data: Irisdat* (5v by 150c)

Fisher (1936) iris data: length & width of sepals and petals, 3 types

	1 ДЛЧАШЕЛ	2 ШИРЧАШЕЛ	3 ДЛЛЕПЕСТ	4 ШИРЛЕПЕСТ	5 ТИПИРИСА
1	5	3,3	1,4	0,2	SETOSA
2	6,4	2,8	5,6	2,2	VIRGINIC
3	6,5	2,8	4,6	1,5	VERSICOL
4	6,7	3,1	5,6	2,4	VIRGINIC
5	6,3	2,8	5,1	1,5	VIRGINIC
6	4,6	3,4	1,4	0,3	SETOSA
7	6,9	3,1	5,1	2,3	VIRGINIC
8	6,2	2,2	4,5	1,5	VERSICOL
9	5,9	3,2	4,8	1,8	VERSICOL
10	4,6	3,6	1	0,2	SETOSA
11	6,1	3	4,6	1,4	VERSICOL
12	6	2,7	5,1	1,6	VERSICOL

Рис. 11.2. Файл данных Iris.sta

Шаг 3. Нажмите **кнопку Variables (Переменные)** и выберите переменные для анализа.

В качестве Группирующей переменной – Grouping variable выберите переменную IRISTYPE — ТИПИРИСА.

В качестве Независимых переменных – Independent variables выберите переменные ДЛИНА ЧАШЕЛИСТИКА, ШИРИНА ЧАШЕЛИСТИКА, ДЛИНА ПЕСТИКА, ШИРИНА ПЕСТИКА.

Сделайте установки, как показано на рисунке 11.1.

Шаг 4. Нажмите кнопку ОК и откройте диалоговое окно **Model Definition (Определение модели,** рис. 11.3**).**

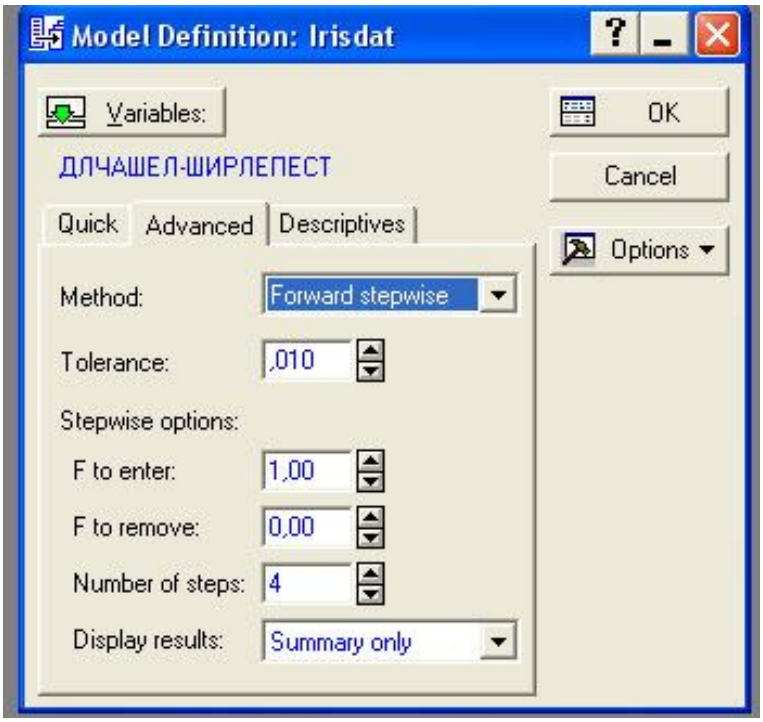


Рис. 11.3. Окно определения модели дискриминантного анализа

Сделайте установки, как показано на рисунке 11.3. Нажмите кнопку ОК и запустите вычислительную процедуру, реализующую пошаговый метод включения.

Шаг 5. Всесторонне просмотрите итоги в диалоговом окне **Результаты дискриминантного анализа** (рис. 11.4).

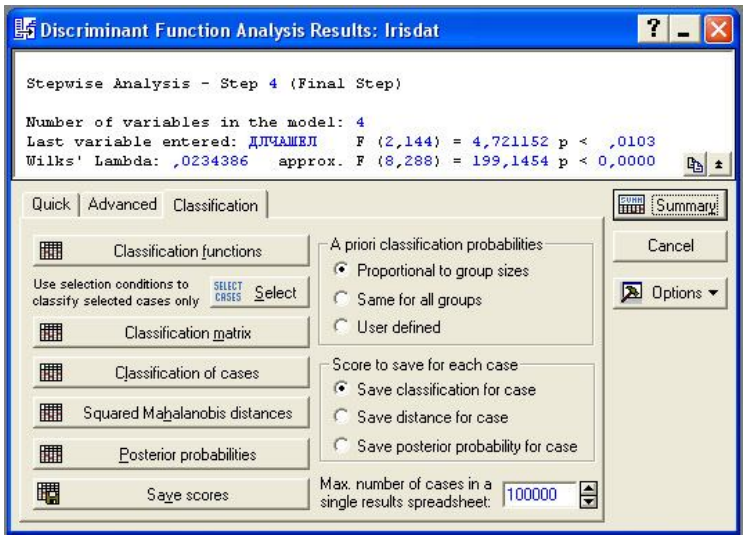


Рис. 11.4. Окно результатов дискриминантного анализа данных из файла

Iris.sta

Информационная часть окна сообщает, что использован:

Stepwise analysis — Пошаговый анализ, Step 4 (Final step) – Шаг 4 (Заключительный шаг);

Number of variables in the model – Число переменных в модели: 4;

Last variable entered – Последняя включенная переменная: ДЛЧАШЕЛИ, соответствующее значение статистики F-критерия F(2, 144) = 4.72, уровень значимости p < 0.0103;

Wilks lambda – Значение лямбды Уилкса: 0.0234;

approx. F (4,292) = 199.1454 — Приближенное значение F-статистики, связанной с лямбдой Уилкса;

p – уровень значимости F-критерия для значения 199.1454

Значения статистики лямбда Уилкса лежат в интервале [0,1]

Значения статистики Уилкса, лежащие около 0, свидетельствуют о хорошей дискриминации. Значения статистики Уилкса, лежащие около 1, свидетельствуют о плохой дискриминации. Иными словами, это можно выразить следующим образом: если значения лямбда Уилкса близки к 0, то мощность дискриминации (мощность = 1–вероятность ошибки) близка к 1, если лямбда Уилкса близка к 1, то мощность близка к 0.

Шаг 6. Нажмите кнопку Variables in the model (Переменные, включенные в модель).

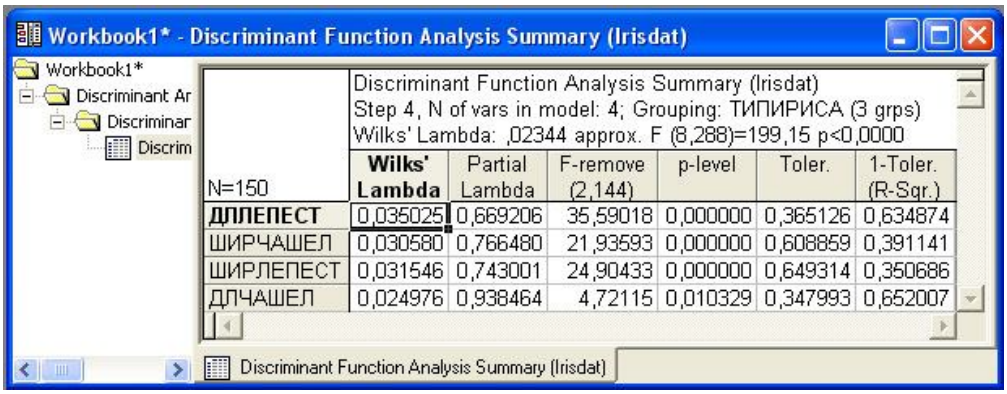


Рис. 11.5. Итоговая таблица анализа данных из файла iris.sta

Шаг 7. Просмотрите разделение групп на графике. Для этого

инициируйте кнопку Canonical analysis & graphs (Канонический анализ и графики). В появившемся диалоговом окне Canonical Analysis (Канонический анализ) нажмите кнопку Scatterplot of canonical scores (Диаграмма рассеяния канонических значений). На экране появится следующий график (рис 11.6):

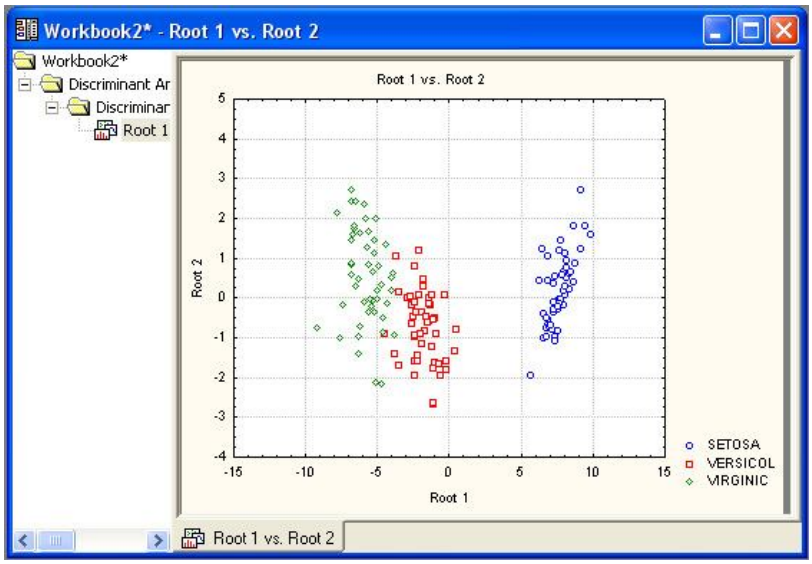


Рис. 11.6. Разделение трех типов ириса

Шаг 8. Просмотрите функции классификации. В диалоговом окне Результаты дискриминантного анализа нажмите кнопку Classification functions (Функции классификации, рис. 11.7).

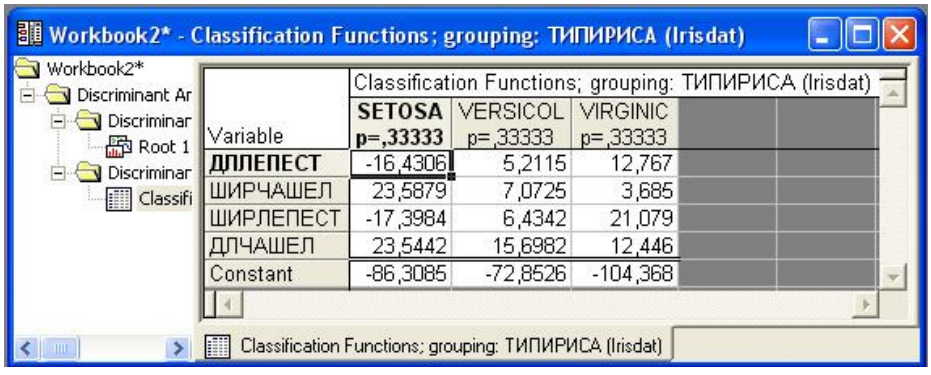


Рис. 11.7. Функции классификации, построенные пошаговым методом вперед (Forward stepwise)

С помощью этих функций можно вычислить классификационные значения (метки) для вновь наблюдаемых цветков по формулам:

SETOSA = -16.43*ДЛЛЕПЕСТ + 23.69*ШИРЧАШЕЛ -

17.4*ШИРЛЕПЕС + 23.54*ДЛЧАШЕЛИ - 86.31

VERSICOL = 5.21*ДЛЛЕПЕСТ + 7.07*ШИРЧАШЕЛ - 6.43*ШИРЛЕПЕС + 15.70*ДЛЧАШЕЛИ - 72.85

VIRGINIC = 12.76*ДЛЛЕПЕСТ + 3.69*ШИРЧАШЕЛ - 21.08*ШИРЛЕПЕС + 12.5*ДЛЧАШЕЛИ - 104.37

Пусть вы имеете новый цветок со значениями: ДЛЛЕПЕСТ, ШИРЧАШЕЛ, ШИРЛЕПЕС, ДЛЧАШЕЛИ

К какому типу ириса его отнести? Формально следует подставить эти значения в приведенные выше формулы и вычислить классификационные значения **SETOSA, VERSICOL, VIRGINIC.**

Новый цветок относится к тому классу, для которого классификационное значение максимально.

Конечно, построенные классификационные функции могут быть определены в электронных таблицах как формулы, и для каждого добавленного случая по ним могут быть вычислены классификационные метки. Таким образом, каждый новый объект автоматически относится к определенному классу.

Шаг 9. Расстояния Махаланобиса.

Нажмите кнопку **Squared Mahalanobis distance (Квадрат расстояния Махаланобиса) и** вы увидите таблицу с квадратами расстояния Махаланобиса от точек (случаев) до центров групп (рис. 11.8):

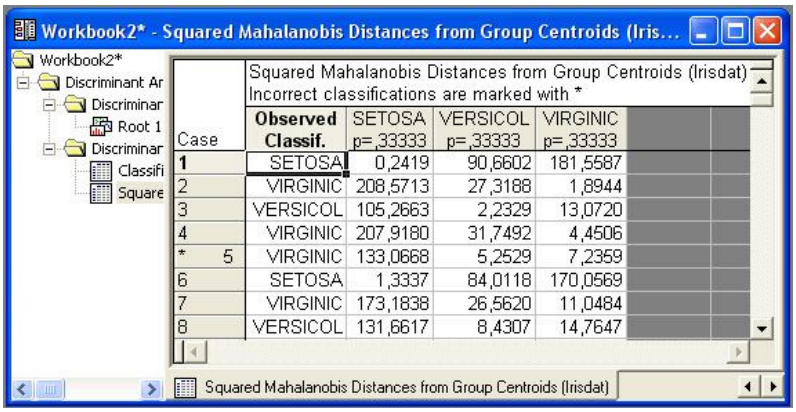


Рис. 11.8. Расстояния Махаланобиса для данных из файла iris.sta

Случай относится к группе, до которой расстояние Махаланобиса минимально.

Шаг 10. Апостериорные вероятности.

Рассмотрите группу опций внизу диалогового окна Результаты дискриминантного анализа: A priori classifications probabilities (Априорные вероятности классификации). До анализа вы задаете для каждого случая (в данном примере цветка) вероятность, с какой он принадлежит к определенному классу. После того как анализ выполнен, можно пересчитать эти вероятности и получить апостериорные вероятности классификации. Нажав кнопку Posterior probabilities (Апостериорные вероятности), вы увидите таблицу с апостериорными вероятностями принадлежности объекта к определенному классу (рис.11.9).

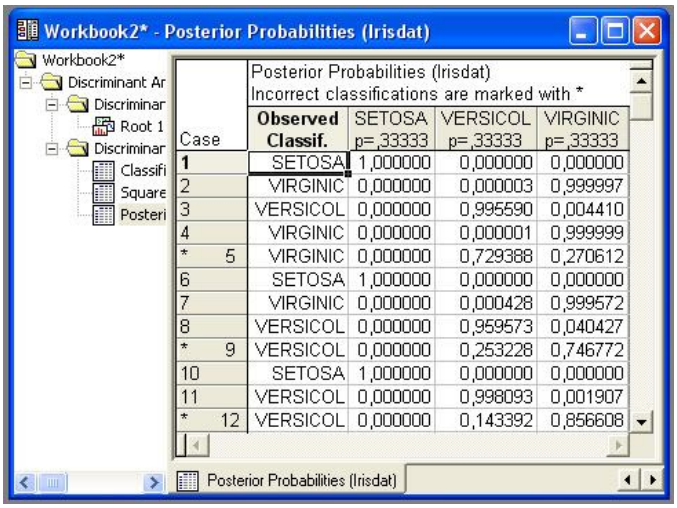


Рис. 11.9. Таблица апостериорных вероятностей

Интерпретация данной таблицы очень проста. В первом столбце указан тип ириса для каждого случая. Во втором, третьем, четвертом столбце даны апостериорные вероятности отнесения каждого цветка к определенному типу.

Цветок относится к группе с максимальной апостериорной вероятностью.

Знаком * отмечаются неправильно классифицированные при использовании данного правила случаи (5, 9, 12).

Шаг 11. Классификация новых случаев.

Не закрывая диалога **Результаты дискриминантного анализа,** добавьте в таблицу исходных данных новый случай, например:

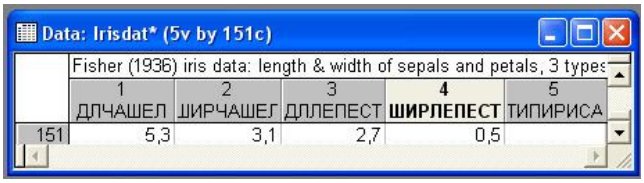


Рис. 11.10. Новое наблюдение в данных Iris.sta

Для того чтобы понять, к какому классу относится этот объект, нажмите **кнопку Posterior probabilities (Апостериорные вероятности),** вы увидите ту же таблицу с постериорными вероятностями, к которой будет добавлена строка (рис. 11.11):

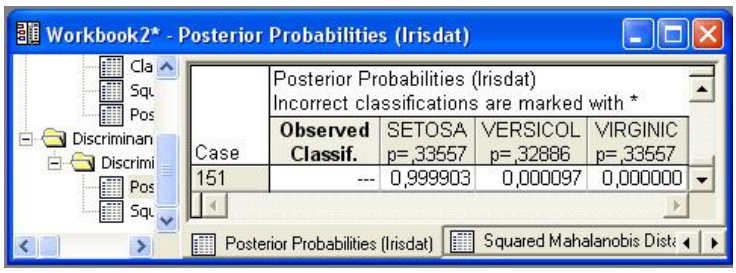


Рис. 11.11. Классификация нового наблюдения

Итак, новое наблюдение с вероятностью 0.999 можно отнести к типу SETOSA.

Нажмите кнопку **Squared Mahalanobis distance (Квадрат расстояния Махаланобиса),** и вы увидите таблицу с квадратами расстояния Махаланобиса. В последней строке таблицы мы видим расстояния нового случая до групповых центров:

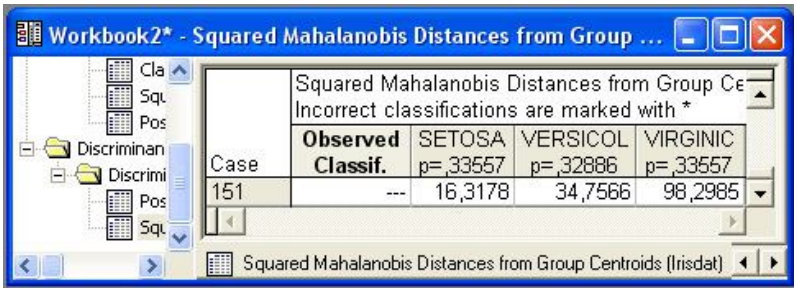


Рис. 11.12. Расстояние Махаланобиса от нового наблюдения до центров групп

Опять расстояние от нового наблюдения до центра групп минимально для группы SETOSA. Следовательно, с высокой степенью вероятности новый цветок — это ирис типа SETOSA.

ТЕМА 12 КЛАСТЕРНЫЙ АНАЛИЗ (НА ПРИМЕРЕ АВТОМОБИЛЕЙ РАЗНЫХ МАРОК)

12.1 Запуск модуля Кластерный анализ

12.2 Открытие файла данных

12.3 Выбор метода

12.4 Выбор переменных, установка начальных значений, запуск вычислительной процедуры метода k-средних

12.5 Просмотр результатов кластеризации

Кластерный анализ объединяет различные процедуры, используемые для проведения классификации. В результате применения этих процедур исходная совокупность объектов разделяется на кластеры или группы (классы) схожих между собой объектов. Под кластером обычно понимают группу объектов, обладающую свойством плотности (плотность объектов внутри кластера выше, чем вне его), дисперсией, отделимостью от других кластеров, формой (например, кластер может иметь очертания гиперсферы или эллипсоида), размером. Конечно, данное определение не является строгим (строгого определения не существует вообще). Если вы взглянете на географическую карту и увидите на ней горы или созвездия на звездном небе, то поймете, что такое кластеры.

Наиболее часто методы кластерного анализа используются в социологии, маркетинговых исследованиях, экономике, биологии, медицине, археологии.

Сложность задач кластерного анализа состоит в том, что реальные объекты являются многомерными, то есть описываются не одним, а несколькими параметрами (представьте, что объекты — это персональные компьютеры), и объединение объектов в группы проводится в пространстве многих измерений, что весьма нетривиально. Кроме того, данные могут носить нечисловой характер.

В целом методы кластеризации делятся на *агломеративные* (от слова

агломерат — скопление) и *итеративные дивизивные* (от слова division —-деление, разделение).

В агломеративных, или объединительных методах происходит последовательное объединение наиболее близких объектов в один кластер. Процесс такого последовательного объединения можно показать на графике в виде дендрограммы, или дерева объединения. Это удобное представление позволяет наглядно представить кластеризацию агломеративными алгоритмами.

Исходными данными для анализа могут быть собственно объекты и их параметры, например, в нашей, рассматриваемой далее задаче с автомобилями разных марок в файле STATISTICA в случаях (строках таблицы) записаны марки машин, в переменных (столбцах): price — цена, acceler — время в секундах, необходимое для того, чтобы разогнаться с места до скорости 60 миль в час, и другие параметры. Данные для анализа могут быть также представлены матрицей расстояний между объектами, в которой на пересечении строки с номером i и столбца с номером j записано расстояние между i-м и j-м объектом.

Если расстояния не даны сразу, то агломеративные алгоритмы начинаются с вычисления расстояний между объектами.

Переход от объектов к *расстояниям* между объектами – важный момент.

Расстояние между объектами – одна из мер сходства. Интуитивно понятно, что, чем меньше расстояние между объектами, тем они более схожи. Но как выбрать естественную метрику, то есть как *естественно* для данной задачи измерить расстояние между объектами?

Часто используют обычную *евклидову* метрику, например, если объект описывается двумя параметрами, то он может быть изображен точкой на плоскости, а расстояние между объектами – это расстояние между точками, вычисленное по теореме Пифагора. Вы просто возводите в

квадрат расстояния по каждой координате, суммируете их и из полученной суммы извлекаете квадратный корень. Если вы не будете возводить в квадрат покоординатные расстояния, а просто возьмете их абсолютные значения и просуммируете, то получите так называемое *манхэттенское расстояние*, или «расстояние городских кварталов». Такое расстояние связано с перемещением человека по улицам города, а не с движением по ровной местности.

Представьте, что вы находитесь в городе. Здесь существуют определенные правила перемещения и, соответственно, правила вычисления пройденного расстояния. Перемещаться можно только по улицам (нельзя, например, пересечь квартал или дом по диагонали). Аналогия в декартовой плоскости приводит к перемещениям только по линиям, параллельным осям координат, и, соответственно, к манхэттенскому расстоянию.

В программе STATISTICA доступны следующие меры сходства объектов: евклидова метрика, квадрат евклидовой метрики, манхэттенское расстояние, или «расстояние городских кварталов», метрика Чебышева, метрика Минковского, пирсоновский коэффициент корреляции (точнее, 1–пирсоновский коэффициент корреляции), коэффициент совстречаемости (точнее, 1–коэффициент совстречаемости).

В программе STATISTICA все эти испытания проводятся в естественной среде в диалоговом режиме.

В STATISTICA реализованы следующие методы кластеризации — агломеративные **методы: joining (tree clustering), two-way joining,** а также метод **k-средних — k-means clustering.**

Обычно перед началом классификации данные стандартизуются (вычитается среднее и производится деление на корень квадратный из дисперсии). Полученные в результате стандартизации переменные имеют нулевое среднее и единичную дисперсию. Эту операцию успешно можно

проводить в программе EXCEL. Рассматриваемые нами далее данные уже стандартизованы.

В STATISTICA можно выбрать следующие правила иерархического объединения кластеров:

Single linkage — метод одиночной связи,

Complete linkage — метод полной связи,

Unweighted pair group average — невзвешенный метод «средней связи»,

Weighted pair group average — взвешенный метод «средней связи»,

Weighted centroid pair group (median) — взвешенный центроидный метод,

Ward method — метод Уорда.

Данные алгоритмы различаются правилами объединения объектов в кластеры.

В методе одиночной связи на первом шаге объединяются два объекта, имеющие между *собой максимальную* меру сходства. На следующем шаге к ним присоединяется объект с максимальной мерой сходства с *одним* из объектов кластера. Таким образом, процесс продолжается далее. Итак, для включения объекта в кластер требуется максимальное сходство лишь с одним членом кластера. Отсюда и название метода одиночной связи, нужна только одна связь, чтобы присоединить объект к кластеру: связь нового элемента с кластером определяется только по одному из элементов кластера. Недостатком этого метода является образование слишком больших «продолговатых» кластеров.

Метод полных связей позволяет устранить указанный недостаток. Здесь мера сходства между объектом – кандидатом на включение в кластер и всеми членами кластера не может быть меньше некоторого порогового значения. В методе средней связи мера сходства между кандидатом и

членами кластера усредняется, например, берется просто среднее арифметическое мер сходства.

Идея еще одного агломеративного метода – метода Уорда состоит в том, чтобы проводить объединение, дающее минимальное приращение внутригрупповой суммы квадратов отклонений. Замечено, что метод Уорда приводит к образованию кластеров примерно равных размеров и имеющих форму гиперсфер.

Рассмотрим еще итеративный метод группировки **k-средних — k-means clustering.** Данный метод работает непосредственно с *объектами,* а не с матрицей сходства.

В методе **k-средних** объект относится к тому классу, расстояние до которого минимально. Расстояние понимается как евклидово расстояние, то есть объекты рассматриваются как точки евклидова пространства.

Как определить евклидово расстояние, мы уже знаем. Но как определить расстояние от *объекта* до *совокупности* объектов? Оказывается, это можно сделать следующим способом: каждый класс объектов имеет центр тяжести (рассмотрите, как и ранее, простейший случай — представьте, что объект имеет только два параметра, тогда его можно изобразить точкой на плоскости, а группа объектов – это просто группа точек).

Расстояние между объектом и классом есть расстояние между объектом и центром класса. Но как вычислить центр класса?. Например, взять средние по каждому параметру. Тогда расстояние между объектом и группой объектов вполне определено и алгоритм может работать.

Представьте, что число объектов в группе равно 2. Соедините эти точки отрезком прямой и найдите его середину. Это и будет центр тяжести группы, состоящей из двух точек. Расстояние от этого центра до исходной точки будет искомым расстоянием.

Принципиально метод **k-средних** «работает» следующим образом:

вначале задается некоторое разбиение данных на кластеры (число кластеров определяется пользователем); вычисляются центры тяжести кластеров;

происходит перемещение точек: каждая точка помещается в ближайший к ней кластер;

вычисляются центры тяжести новых кластеров;

шаги 2, 3 повторяются, пока не будет найдена стабильная конфигурация (то есть кластеры перестанут изменяться)или число итераций не превысит заданное пользователем. Итоговая конфигурация и является искомой.

Таким образом, мы надеемся, вы получили первое представление о методах классификации. Возможно, в будущем, поработав со STATISTICA, вам удастся придумать свой метод кластеризации или модифицировать существующие. Кластерный анализ открыт для новых идей.

Перейдем к рассмотрению примера.

В нашем примере рассматриваются автомобили разных марок, которые различаются ценой, расходом горючего и некоторыми техническими характеристиками, например, разгоном — временем, необходимым для того, чтобы достичь скорости 60 миль в час.

12.1 Запуск модуля Кластерный анализ

В рабочем окне **STATISTICA** выберете название модуля – **Cluster Analysis (Кластерный анализ),** высветите его имя и щелкните на его имени.

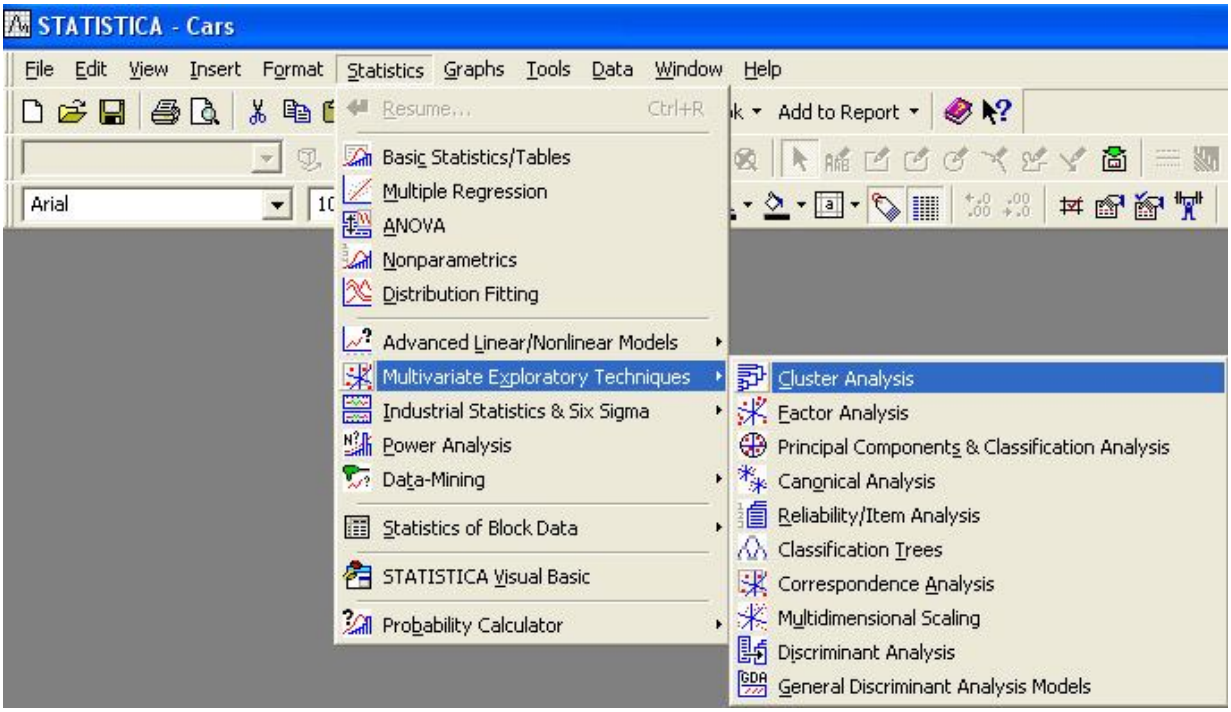


Рис. 12.1. Запуск модуля Кластерный анализ

На экране появится стартовая панель модуля **Claster Analysis (Кластерный анализ)** (рис. 12.2).

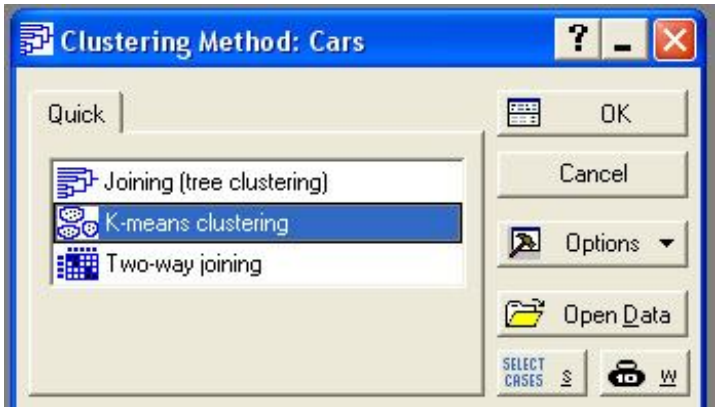


Рис. 12.2. Стартовая панель модуля Кластерный анализ

12.2 Открытие файла данных

Стандартным образом, нажав кнопку **Open Data,** откройте окно выбора файла (рис. 12.3).

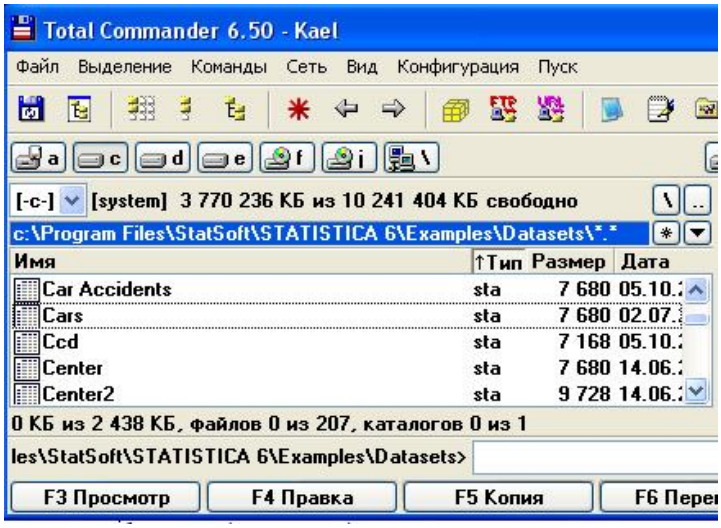


Рис. 12.3. Выбор файла с данными об автомобилях

Выберите в этом окне файл Cars.sta, как показано на рисунке, и два раза нажмите левую кнопку мышки**.** Файл выбран, и вы вернетесь обратно, в стартовую панель модуля.

В рабочем окне, сзади стартовой панели, вы видите открытый файл с данными (рис. 12.4).

Data: Cars (5v by 22c)

	Performance, fuel economy, and approximate price for various automobiles				
	1 PRICE	2 ACCELERATION	3 BRAKING	4 HANDLING	5 MILEAGE
Acura	-0,521072363	0,477252671	-0,00657103855	0,381619066	2,07875356
Audi	0,865652474	0,208033216	0,31869537	-0,0913735792	-0,677061608
BMW	0,495859184	-0,801539742	0,192202878	-0,0913735792	-0,153805564
Buick	-0,613520685	1,68874022	0,933087475	-0,20962174	-0,153805564
Corvette	1,23544576	-1,8111127	-0,494470651	0,972859872	-0,677061608
Chrysler	-0,613520685	0,0734234878	0,427117506	-0,20962174	-0,153805564
Dodge	-0,705969008	-0,195795968	0,481328574	0,145122743	-0,153805564
Eagle	-0,613520685	1,21760617	-4,19889364	-0,20962174	-0,677061608
Ford	-0,705969008	-1,54189324	0,987298543	0,145122743	-1,7235737
Honda	-0,42862404	0,409947807	-0,00657103855	0,0268745821	0,369450479
Isuzu	-0,79841733	0,409947807	-0,0607821066	-4,23005922	1,0671252
Mazda	0,126065894	0,679167263	-0,133063531	0,499867227	-1,7235737
Mercedes	1,05054912	0,00611862399	0,119921454	-0,0913735792	-0,153805564
Mitsub.	-0,613520685	-1,00345433	0,0837807415	0,381619066	0,718287842
Nissan	-0,42862404	0,0734234878	-0,00657103855	0,263370905	0,997357732
Olds	-0,613520685	-0,734234878	0,40904715	0,381619066	2,11363729
Pontiac	-0,613520685	0,679167263	0,535539642	0,145122743	0,195031798
Porsche	3,4542055	-2,21494188	-0,295696735	0,618115388	-1,02589897
Saab	0,588307506	0,679167263	0,246413946	0,263370905	0,0206131169
Toyota	-0,0588307506	1,21760617	0,22834359	0,73636355	-0,851480289
VW	-0,705969008	-0,128491104	0,101851098	0,381619066	0,195031798
Volvo	0,218514217	0,611862399	0,13799181	-0,20962174	0,369450479

Рис. 12.4. Файл cars.sta с данными автомобилей разных марок

Из информации в верхней части окна вы видите, что в **файле** Cars.sta записаны цена автомобиля, технические характеристики, количество миль,

пройденных на одном галлоне бензина.

Всего в файле содержатся данные о 22 машинах разных марок**.** Марки машин – это случаи.

Переменные в этом файле:

PRICE – цена;

ACCELE - HANDLI –технические характеристики;

MILAGE – расход горючего (количество миль, пройденных на одном галлоне бензина).

Все характеристики машин уже стандартизованы: из значений переменной price вычтена средняя цена и разность поделена на корень квадратный из дисперсии.

Наша задача – разбить автомобили на несколько групп, в которых автомобили мало отличаются друг от друга (существенно меньше, чем в целом в совокупности).

Задача эта сложна, так как мы сравниваем машины не по какому-то одному параметру, а по нескольким параметрам одновременно.

Вы видите, что по одним характеристикам одни машины близки друг к другу, по другим – другие. В конечном итоге разбиение на группы – тоже не самоцель. Конечно, число параметров можно увеличить. Очевидно, разбив машины на группы, мы можем лучше в целом представить их совокупность, с тем, чтобы затем более обоснованно принимать решение, например при покупке или обмене одной машины на другую.

Если бы машины сравнивались по одному параметру, например по расходу горючего, мы выбрали бы машину с меньшим расходом топлива на одну милю. Все машины были бы упорядочены в одну линию, и задача не представляла бы проблем.

Однако параметров несколько, и ситуация существенно усложняется.

12.3 Выбор метода

Посмотрите на стартовую панель. В главной ее части находится список

методов кластерного анализа, реализованных в STATISTICA.

В списке методов высветите k-means **(k-средних)** (см. рис. 12.2) и нажмите кнопку ОК в правом верхнем углу панели.

Диалоговое окно метода **k-means** появится на экране (рис. 12.5):

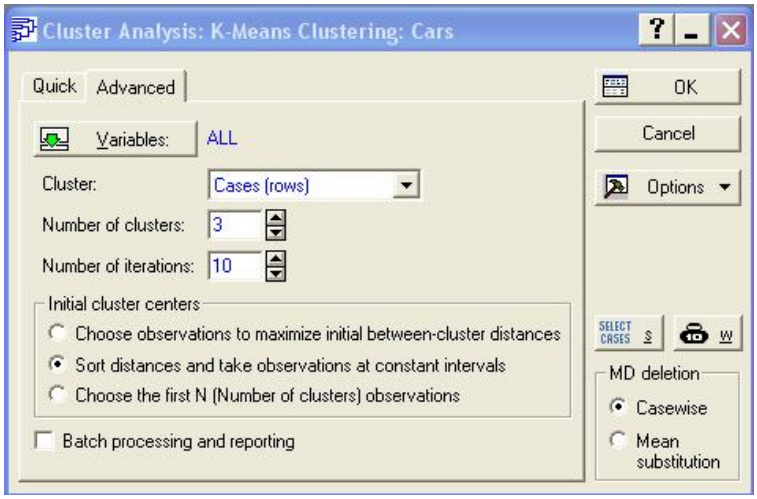


Рис. 12.5. Диалоговое окно метода k-means

12.4 Выбор переменных, установка начальных значений, запуск вычислительной процедуры метода k-средних

Начните работать в данном окне. Прежде всего, выберите переменные для анализа.

Нажмите кнопку **Variables (Переменные)** в левом верхней углу текущего окна и откройте диалоговое окно: **Select variable for the analysis** (рис. 12.6).

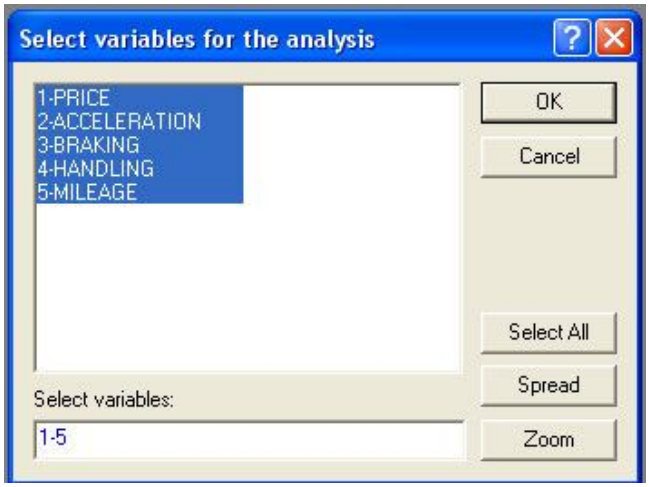


Рис. 12.6. Выбор переменных для кластерного анализа

Так как мы будем разбивать машины на группы и учитывать все параметры, то нажмите вначале кнопку **Select All (Выбрать все),** а затем нажмите кнопку **ОК.**

Посмотрите далее на поле **Cluster (Кластер),** находящееся ниже **кнопки Variables (Переменные).** Нажав на стрелку в этом поле, выберите пункт меню **Cases (Случаи).** Альтернативный выбор был **бы Variables (Переменные).** Так следует поступить, если нужно кластеризировать переменные.

В данном примере мы кластеризируем машины, которые являются случаями в исходном файле данных, поэтому мы и выбрали пункт Cases.

В поле **Number of clusters (Число кластеров)** нужно определить число групп, на которые мы хотим разбить автомобили. Запишите в это поле число 3.

Таким образом, мы будем разбивать машины на 3 кластера.

В строке **Number of iterations (Число итераций)** задается максимальное число итераций, используемых при построении классов. Задайте, например, число 11.

В строке **Missing data** задается способ обработки пропущенных значений в данных (например, для какой-то машины отсутствует значение некоторого параметра). В данном примере пропусков в данных нет и обработки пропущенных значений не происходит.

Группа опций **Initial cluster centers** позволяет задать начальные центры кластеров.

Сделайте установки, как показано на рисунке 12.5.

После того как все установки сделаны, нажмите кнопку ОК в верхнем правом углу окна **k-means Clustering** и запустите вычислительную процедуру.

12.5 Просмотр результатов кластеризации

Спустя несколько секунд после нажатия кнопки **ОК** в k-means

Clustering окно результатов появится на экране (рис. 12.7):

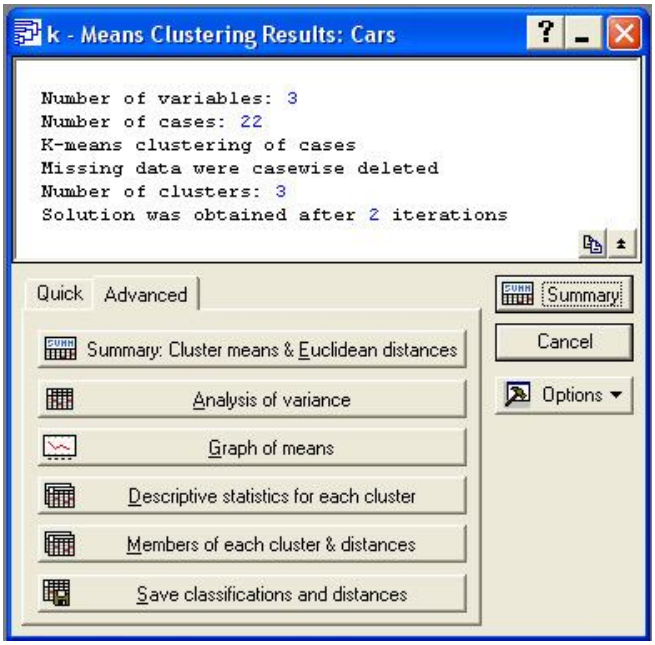


Рис. 12.7. Окно результатов кластеризации машин по методу средних

В верхней части окна записана информация: число переменных число случаев, метод кластеризации, число кластеров, а также сообщение о том, после скольких итераций найдено решение:

Solution was obtained after 3 iterations — Решение найдено после 3 итераций.

Кнопки в нижней части окна позволяют провести анализ результатов кластеризации.

Кнопка Analysis of variation (Дисперсионный анализ) позволяет просмотреть таблицу дисперсионного анализа.

Кнопка Cluster Means&Euclidean Distances позволяет вывести таблицы, в первой из которых указаны средние для каждого кластера (усреднение производится внутри кластера), во второй указаны евклидовы расстояния и квадраты евклидовых расстояний между кластерами.

Кнопка **Graph of means** позволяет посмотреть средние значения для каждого кластера на линейном графике.

Кнопка **Descriptive Statistics for each clusters** открывает электронную таблицу с описательными статистиками для каждого кластера (среднее,

дисперсия и т.д.)

Кнопка **Save classifications and distances** позволяет сохранить результаты классификации в файле STATISTICA для дальнейшего исследования.

Нам, конечно, интересно посмотреть, как распределились машины по кластерам. Нажмите для этого кнопку **Member of each cluster&distances.**

На экране появятся 3 электронные таблицы с названиями машин, отнесенных к определенным кластерам (рис. 12.8 — 12.10):

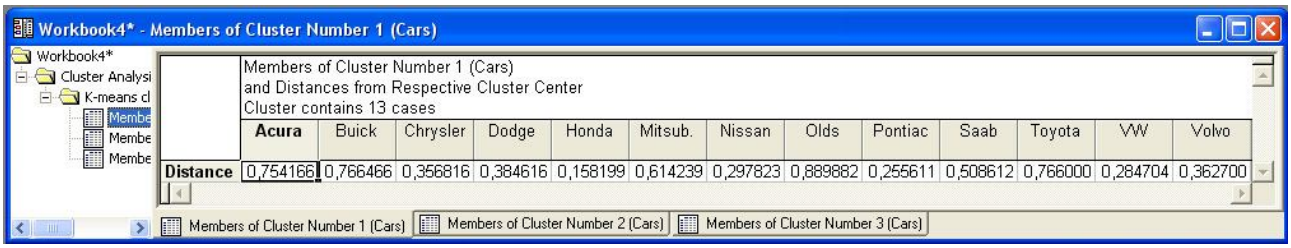


Рис. 12.8. Первый кластер

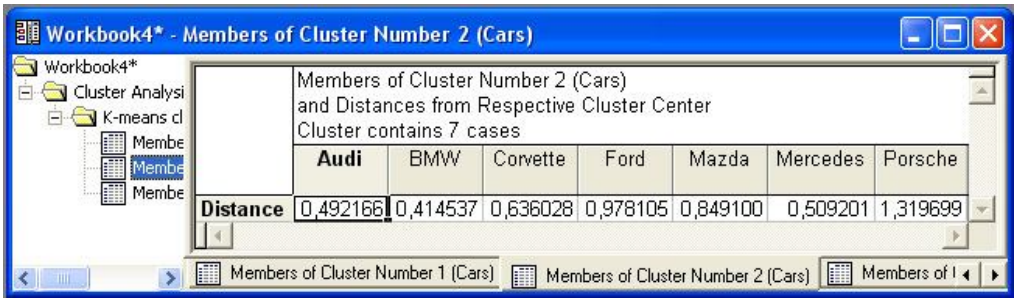


Рис. 12.9. Второй кластер

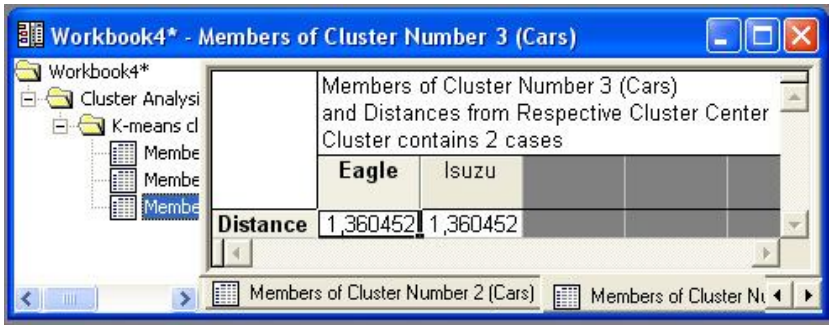


Рис. 12.10. Третий кластер

В строках таблиц указано расстояние от каждой машины до центра кластера.

Нажмите на кнопку **Cluster means&Euclidean distances.** На экране

появится таблица, в которой даны евклидовы расстояния между средними кластеров (по каждому из параметров внутри кластера вычисляется среднее, получается 3 точки в пятимерном пространстве, и между ними находится расстояние) (рис. 12.11).

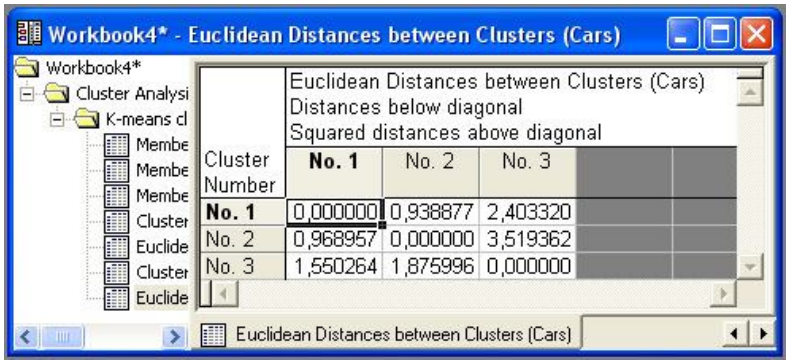


Рис. 12.11. Расстояния между кластерами

Из таблицы вы видите, что расстояние между первым и вторым кластером 0.969, а например, между вторым и третьим –1.876.

Над диагональю в таблице даны квадраты расстояний между кластерами.

С помощью кнопки **Graph of means (График средних)** строятся следующие графики средних значений характеристик машин для каждого кластера (рис. 12.12):

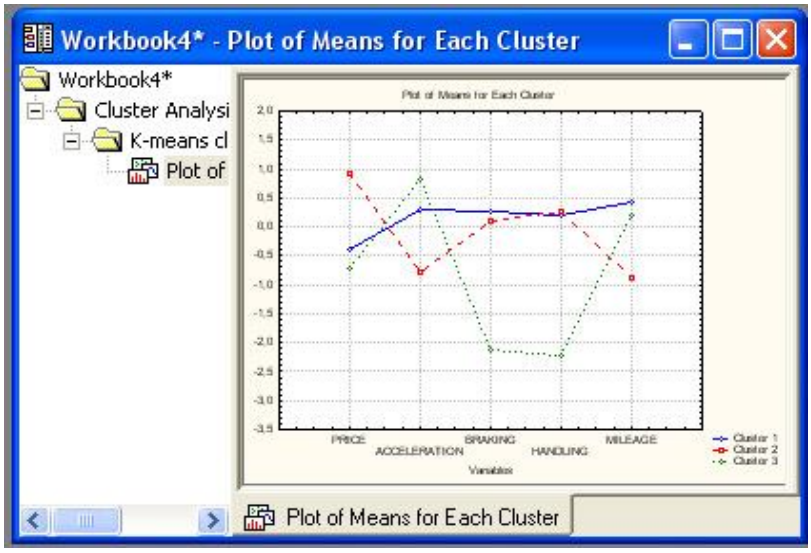


Рис. 12.12. График средних для каждого кластера

Закроем окно результатов и вернемся в начальное окно метода k-

средних.

Изменим переменные для анализа.

Нажмите кнопку **Variables (Переменные)** в левом верхнем углу текущего окна и откройте диалоговое окно: **Select variables for the analysis.** Сделайте в нем установки, как показано на рис. 12.13 (мы выберем теперь только 3 параметра, характеризующих машины):

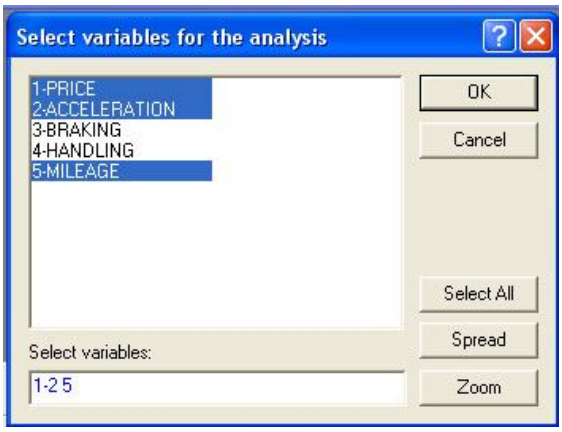


Рис. 12.13. Выбор части переменных для кластерного анализа методом k-средних

Повторите действия, описанные ранее. Нажмите кнопку **Graph of means (График средних),** постройте следующие графики средних значений характеристик машин для каждого кластера (рис. 12.14):

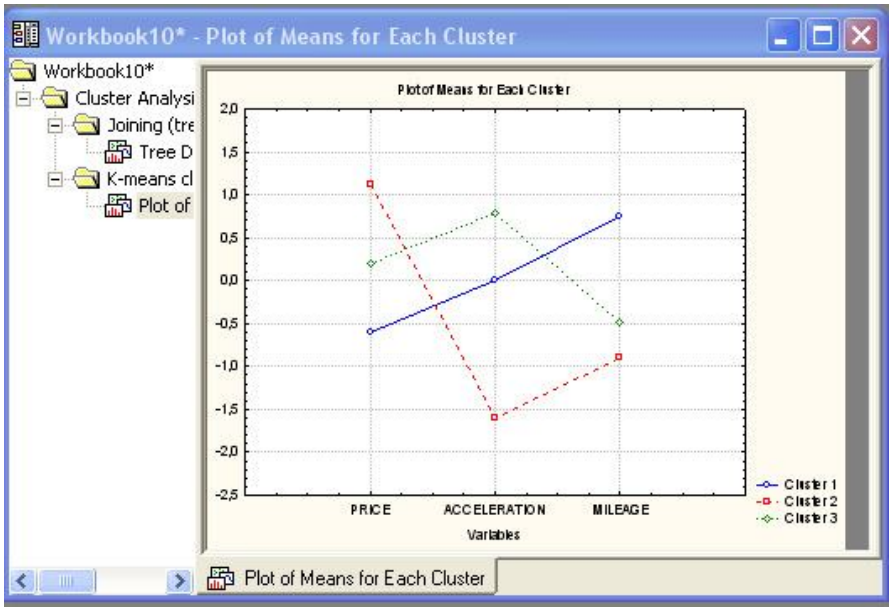


Рис. 12.14. График средних для новых кластеров

Заметьте, что состав групп изменился. Теперь машины более отчетливо группируются. Мы «пожертвовали» размерностью, сократили число параметров и получили более отчетливо выраженные группы (сравните с рис. 12.12).

Поэкспериментируйте с этими данными. Возможно, вам удастся найти оптимальную кластеризацию.

После того как вы поработаете с этим примером, обязательно попробуйте расклассифицировать другие свои собственные данные.

В системе реализованы также и другие методы кластеризации, в частности так называемый **two-way joining,** в котором кластеризируются случаи и переменные одновременно.

Если вы воспользуетесь Joining **(tree clustering),** то сможете увидеть дендрограмму, или дерево объединения (рис. 12.15), о котором мы говорили вначале.

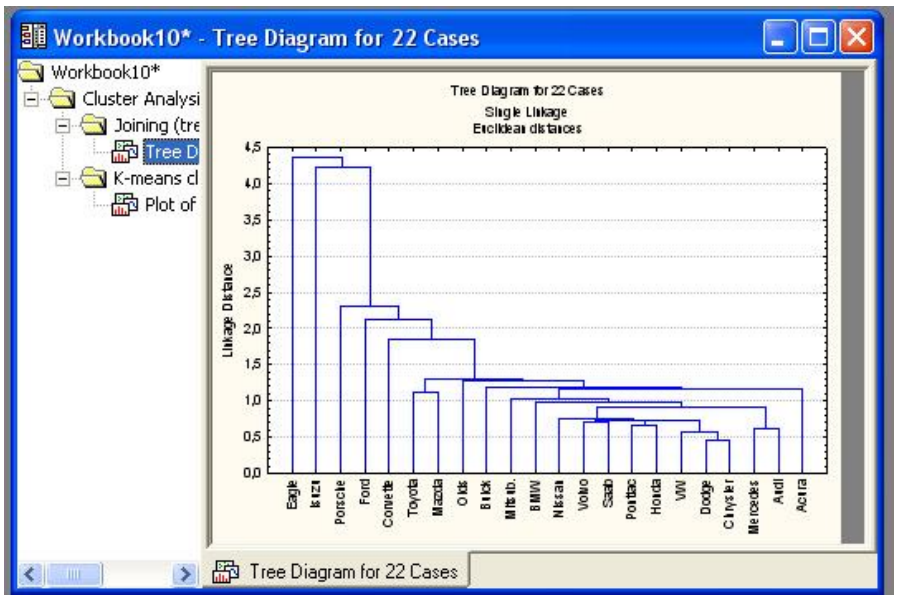


Рис. 12.15. Дерево объединения машин разных марок в кластер методом одиночной связи

ЛИТЕРАТУРА

1. Биометрия [Текст]: [учебник / Лакин Г.Ф.]. – М.: Высшая школа, 1990. – 244 с.
2. Урбах В.Ю. Статистический анализ в биологических и медицинских исследованиях [Текст]: [учебник / Урбах В.Ю.]. – М.: Медицина, 1975. –345 с.
3. Плохинский Н.А. Биометрия [Текст]: [учебник / Плохинский Н.А.]. – М.: МГУ, 1970. – 368 с.
4. Свалов Н.Н. Вариационная статистика [Текст]: [учебник / Свалов Н.Н.]. – М.: Лесная промышленность, 1977. – 177 с.
5. Боровиков В.П. Популярное введение в программу STATISTICA: [Текст]: учеб. пособие для студ. высш. учеб. заведений/ В.П. Боровикова. – М.: КомпьютерПресс, 1998. 269 с.
6. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистические методы в медико-биологических исследованиях с использованием Excel: [Текст]: учеб. пособие для студ. высш. учеб. заведений/ С.Н. Лапач и др. – К.: МОРИОН, 2000. 105 с.

Учебное издание

Жученко Юрий Михайлович

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ИНФОРМАЦИИ С ПРИМЕНЕНИЕМ ПЕРСОНАЛЬНЫХ КОМПЬЮТЕРОВ ПРАКТИЧЕСКОЕ ПОСОБИЕ

для студентов IV курса

специальность 1-31 01 01 02 "Биология (научно-педагогическая деятельность)"

Редактор В.И. Шкредова

Корректор В.В. Калугина

Лицензия № 02330/0133208 от 30.04.04.
Подписано в печать _____. Формат 60х84 1/16.
Бумага писчая №1. Гарнитура «Таймс». Усл. п. л. ____.
Уч.- изд. л. ____. Тираж 100 экз. Заказ № _____.

Отпечатано с оригинал-макета на ризографе
Учреждения образования
«Гомельский государственный университет
имени Франциска Скорины»
лицензия №02330/0056611 от 16.02.04.
246019, г. Гомель, ул. Советская, 104