А. Г. Сыса, Е. П. Живицкая

СТАТИСТИЧЕСКИЙ АНАЛИЗ В БИОЛОГИИ И МЕДИЦИНЕ

Учреждение образования
«Международный государственный экологический
институт имени А. Д. Сахарова»
Белорусского государственного университета

А. Г. Сыса, Е. П. Живицкая

СТАТИСТИЧЕСКИЙ АНАЛИЗ В БИОЛОГИИ И МЕДИЦИНЕ

Учебно-методическое пособие

Минск
«ИВЦ Минфина»
2018

УДК 531:57:61(075.8)
ББК 51.1(02)я7
С97

Р е ц е н з е н т ы:
кандидат технических наук, доцент, доцент кафедры «Экология» Белорусского национального технического университета *С. А. Лаптенок*;
кандидат биологических наук, научный сотрудник РУП «Научно-практический центр гигиены» *Е. К. Власенко*

Сыса, А. Г.

С97 Статистический анализ в биологии и медицине / А. Г. Сыса, Е. П. Живицкая. – Минск : ИВЦ Минфина, 2018. – 140 с.

ISBN 978-985-7205-08-0.

В пособии помещены учебно-методические материалы, соответствующие учебной программе проведения практических и лабораторных занятий по дисциплине «Статистический анализ в биологии и медицине» со студентами II ступени высшего образования.

Главы практикума включают пошаговые инструкции по реализации различных видов математико-статистического анализа в Statistica. Особое внимание уделяется получаемым результатам и их интерпретации.

Предназначается обучающимся специальности 1-33 81 01 Прикладная иммунология II ступени высшего образования МГЭИ им. А. Д. Сахарова БГУ.

УДК 531:57:61(075.8)
ББК 51.1(02)я7

ISBN 978-985-7205-08-0

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Статистика в биологии и медицине является одним из инструментов анализа экспериментальных данных и клинических наблюдений, а также языком, с помощью которого сообщаются полученные математические результаты. Однако это не единственная задача статистики в медицине. Математический аппарат широко применяется в диагностических целях, при решении классификационных задач и поиске новых закономерностей, а также для постановки новых научных гипотез. Поскольку процессы обработки информации, которые являются трудоемкими и занимают много времени, в настоящее время компьютеризированы. Следовательно, решение задач данного типа связано с применением различных компьютерных программ, содержащих математико-статистические процедуры обработки данных. Эти программы, часто именуемые статистическими пакетами, содержат всевозможные процедуры и методы, направленные на работу с большими массивами данных, так что основная трудность на данный момент состоит не в реализации самой процедуры, а в грамотном ее выборе в соответствии с ситуацией.

В качестве базового программного продукта, иллюстрирующего возможности статистического анализа, выбрана программа для обработки статистической информации Statistica. В настоящее время это одно из наиболее удобных и широко распространенных средств статистического анализа, обладающих достаточно удобным интерфейсом и включающих в себя практически полный перечень инструментов и методов, необходимых для решения задач по обработке биомедицинских данных.

Учебно-методическое пособие предназначено, главным образом, для специалистов медицинского и биологического профиля, занимающихся научными исследованиями, магистрантов и аспирантов. Пособие может быть использовано и для студентов старших курсов, изучающих статистический анализ в рамках биомедицинских специальностей.

Материалы лабораторных занятий, включенных в данное издание, являются авторскими.

Глава 1. Описательная статистика и подгонка распределений

1.1. Описательная статистика и подгонка распределений

Распределение. Генеральная и выборочная совокупности. Параметры распределения (среднее, дисперсия, стандартное отклонение, стандартная ошибка среднего, медиана, мода, процентили, размах и коэффициент вариации). Оценка параметров распределения по выборке. Нормальное распределение. Проверка соответствия распределения нормальному закону.

Для быстрого ознакомления с полученными данными и выявления в них явных закономерностей на начальном этапе статистического анализа бывает полезно построить гистограмму или полигон распределения, которые представляют собой графики, отражающие связь между значениями изучаемого признака и частотой встречаемости этих значений.

Необходимо построить полигон распределения для представленных ниже данных о количестве клеток (в тысячах) в 25 образцах: 5 4 5 5 4 5 4 3 5 4 7 5 6 1 6 4 4 4 2 3 5 5 4 2 3.

Для того чтобы программа смогла построить полигон распределения, из имеющихся данных нужно сформировать вариационный ряд, то есть двойной ряд чисел, в котором содержатся значения анализируемого признака и частоты их встречаемости. Переименуйте столбец Var 2 и Var 3 в «Число клеток» и «Количество образцов» соответственно. Теперь сформируем вариационный ряд. Для этого в разделе основного меню **Statistics** (Статистические процедуры) выберем модуль **Basic Statistics/Tables** (Основные статистические показатели/Таблицы), а в нем – опцию **Frequency tables** (Таблицы частот). В появившемся диалоговом окне необходимо указать, какую именно переменную мы собираемся анализировать. Для этого служит кнопка **Variables** (Переменные). При нажатии на нее появится еще одно окно (**Select the variables for the analysis**), основная часть которого занята списком имеющихся в таблице переменных. Дважды кликните по пункту «Клетки» (рис. 1.1), а затем нажмите кнопку **Summary: Frequency tables** (Результат: Таблицы частот).

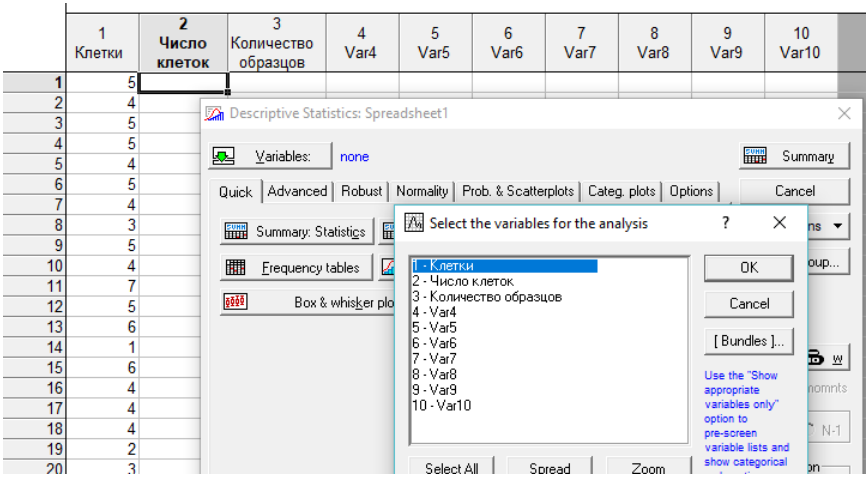


Рис. 1.1. Выбор переменной для расчета таблицы частот

В итоге программа сформирует таблицу, представляющую собой «расширенный» вариант вариационного ряда (рис. 1.2).

Frequency table: Клетки (Spreadsheet1)
K-S d=,20129, p> .20; Lilliefors p<,05

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
0,000000<x<=1,000000	1	1	4,00000	4,0000	4,00000	4,0000
1,000000<x<=2,000000	2	3	8,00000	12,0000	8,00000	12,0000
2,000000<x<=3,000000	3	6	12,00000	24,0000	12,00000	24,0000
3,000000<x<=4,000000	8	14	32,00000	56,0000	32,00000	56,0000
4,000000<x<=5,000000	8	22	32,00000	88,0000	32,00000	88,0000
5,000000<x<=6,000000	2	24	8,00000	96,0000	8,00000	96,0000
6,000000<x<=7,000000	1	25	4,00000	100,0000	4,00000	100,0000
Missing	0	25	0,00000		0,00000	100,0000

Рис. 1.2. Таблица частот для данных

В этой таблице имеются следующие столбцы:

- Category (Категория): содержит ранжированные значения анализируемой переменной, отмеченные в выборке. В случае с нашим примером видим, что переменная изменялась от 1 до 7.
- Count (Счет): здесь приведены частоты, с которыми встречались отмеченные значения переменной.
- Cumulative count: накопленные частоты.
- Percent: процент, который составляет каждая из частот от общего числа наблюдений.
- Cumulative percent: накопленные процентные доли частот.

• Последняя строка итоговой таблицы называется Missing (Отсутствующие) – она имеет отношение к неотмеченным в выборке значениям переменной. Таковых в нашем примере нет (встречались все возможные значения числа – от 1 до7), в связи с чем на пересечении столбца Count и строки Missing видим 0.

Внесите «вручную» необходимые значения из полученной таблицы частот в нашу исходную таблицу с данными. Теперь есть все необходимое, чтобы построить полигон распределения. В разделе главного меню **Graphs** (Графики) выберите **2D Graphs** (Двухмерные графики) > **Line plots (Variables)** (Линейные графики (по переменным)). В появившемся диалоговом окне (рис. 1.3) выберите закладку **Advanced** (Расширенные настройки). На ней в поле **Graph type** (Тип графика) выберите **XY Trace**, а в выпадающем меню **Display points** (Отображать точки) – **On** (Включить). Наконец, откройте закладку **Options 1**, разыщите на ней выпадающее меню **Case labels** (Подписи наблюдений) и снимите флажок. Вы можете не выполнять последнюю операцию, если хотите, чтобы значения количества клеток на графике отображались.

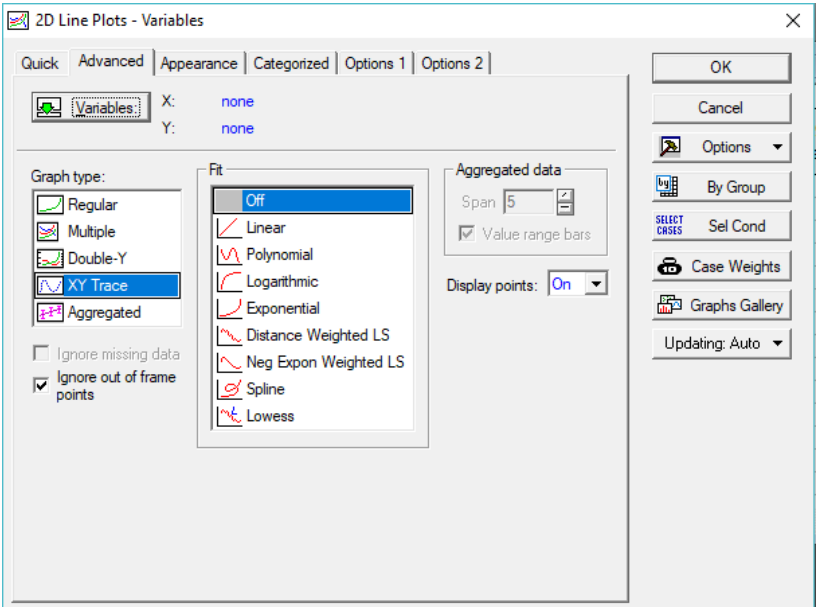


Рис. 1.3. Диалоговое окно модуля 2D Line plots на закладке Advanced

Теперь необходимо «задать» программе, какой из столбцов таблицы с данными соответствуют числу клеток (ось X), а какой – частоте встречаемости (ось Y). Для этого вернитесь на закладку **Advanced**

и нажмите **Variables** (Переменные). В результате появится окно с двумя списками переменных. В левом списке выделите пункт «Число клеток», а в правом – «Количество проб». Далее нажимаем кнопку ОК, затем еще раз ОК, и получаем график (рис. 1.4). Заметьте: полученный график является составной частью рабочей книги Workbook, как это ранее было с итоговой таблицей анализа **Frequency Tables**. Если кликнуть один раз по полученному графику правой кнопкой мыши и из контекстного меню выбрать опцию **Copy Graph**, можно скопировать график в буфер обмена и затем вставить в документ практически любого другого Windows-приложения, например, MS Word или Excel. Кроме того, график можно сохранить как самостоятельный файл с расширением *.stg*. Для этого необходимо выделить иконку графика в каталоге рабочей книги и, удерживая нажатой левую клавишу мыши, перетащить ее за пределы рабочей книги (рис. 1.4). В результате рисунок окажется в отдельном окне. Теперь, кликнув по нему правой кнопкой мыши, можно применить команду **Save Graph** (Сохранить график).

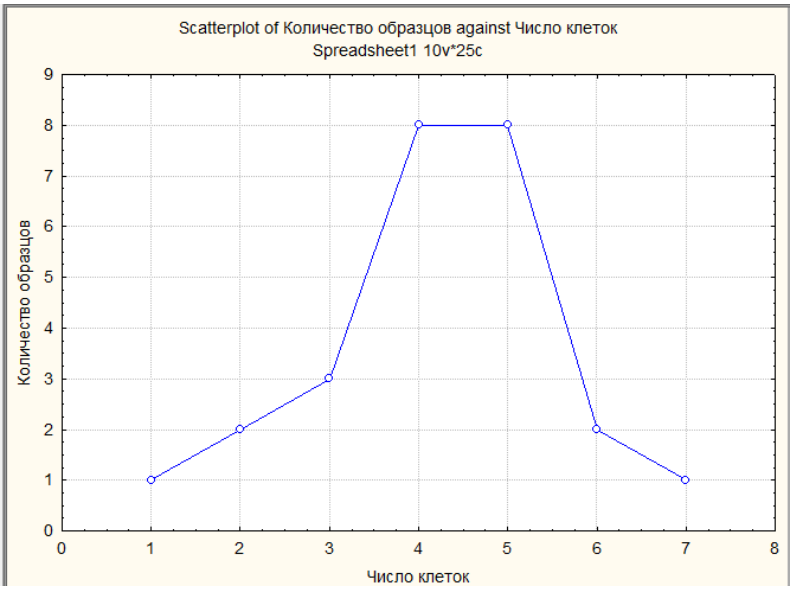


Рис. 1.4. Полигон распределения для данных о числе клеток в пробах

Построение гистограмм

Ниже представлены морфометрических характеристики эндометриоидного овариального рака (площадь клетки в мкм^2): 65 68 71 74 76 76 76 76 76 79 79 79 79 79 79 79 79 79 79 82 82 85 85 85 85 85 88 88 97 103.

Размах значений площади клеток достаточно велик (103 – 65 = 38 мкм). Кроме того, в этом ряду некоторые значения отсутствуют (например, не встречены клетки с длиной 66, 78, 83 мкм). Для графического изображения частотного распределения в данном случае лучше подходит гистограмма, а не полигон распределения.

Создайте новый файл данных с одной переменной (назовите ее, например, «Площадь клеток») и 30 наблюдениями. Введите приведенные выше значения длины клеток в столбец этой таблицы данных.

Для построения гистограммы достаточно выполнить следующие действия:

• В основном меню программы выбрать **Graphs** > **2D Graphs** > **Histograms** (Гистограммы).

• В появившемся окне (рис. 1.5) выбрать закладку **Advanced**. Нажав на кнопку **Variables**, выбрать для анализа переменную «Площадь клеток». В поле **Fit type** (Тип подгонки) выбрать Off, а в выпадающем меню **Y axis** (Ось Y) – %. Остальные настройки можно оставить без изменений.

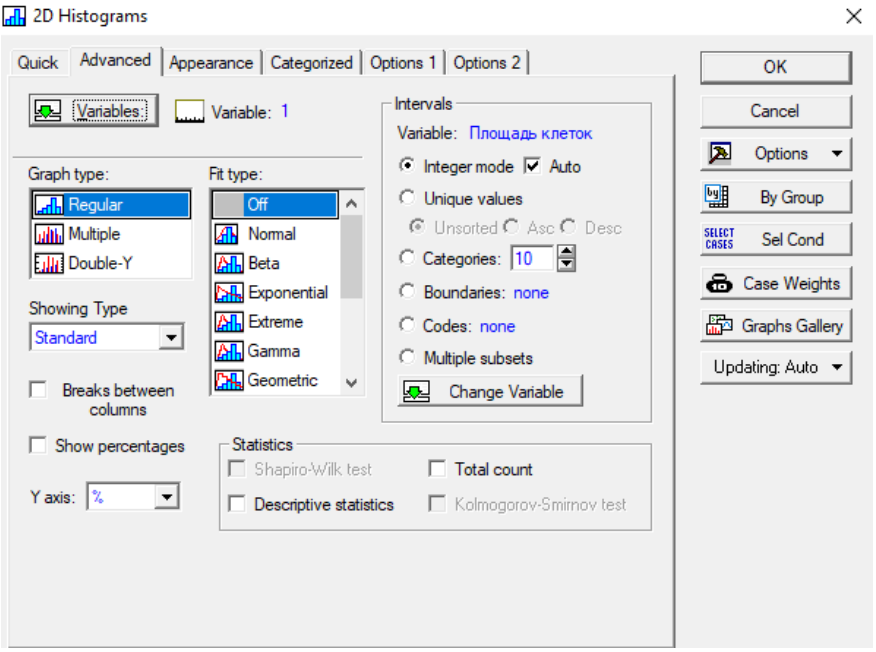


Рис. 1.5. Диалоговое окно модуля 2D Histograms на закладке Advanced

• Нажмите на кнопку OK. В результате у вас должен получиться график, подобный приведенному на рис. 1.6.

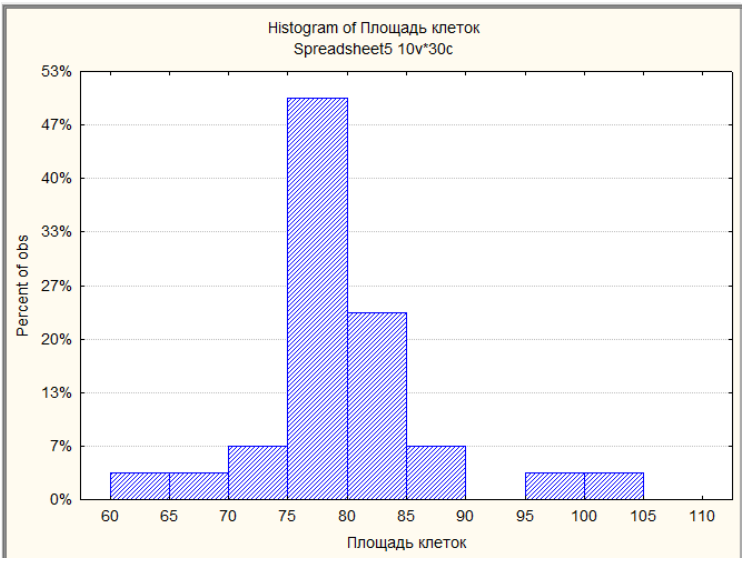


Рис. 1.6. Частотное распределение значений площади клеток

Расчет параметров описательной статистики

Расчет параметров описательной статистики в программе Statistica выполняется при помощи модуля **Descriptive statistics** (Описательная статистика) (рис. 1.7).

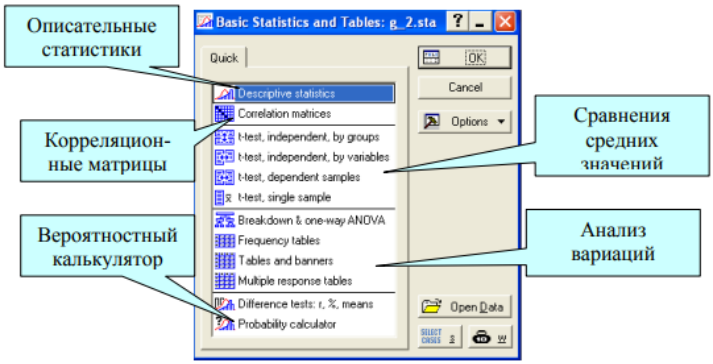


Рис. 1.7. Стартовое окно модуля с перечнем статистических процедур

Для его запуска:

- Войдите в раздел **Statistics** основного меню и выберите в нем пункт **Basic statistics/Tables**. В появившемся окне дважды кликните по пункту **Descriptive statistics**.

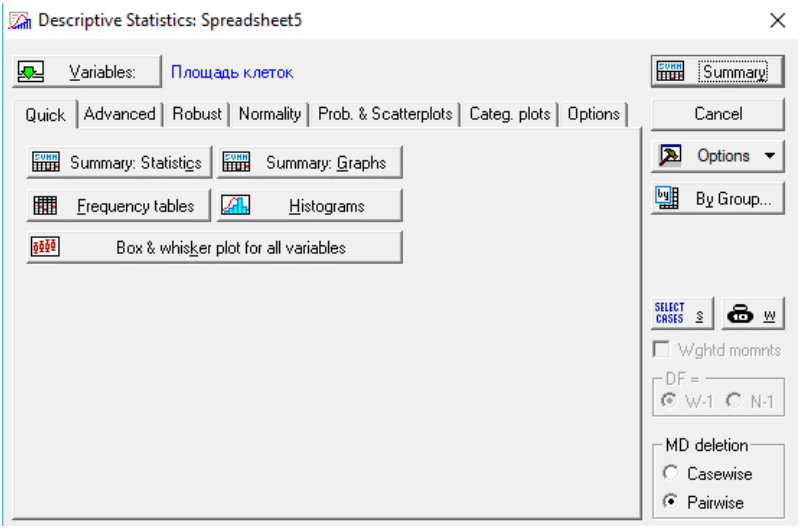


Рис. 1.8. Модуль Descriptive Statistics на закладке Quick

В диалоговом окне модуля **Descriptive statistics** (рис. 1.8) присутствует ряд элементов, встречающиеся в большинстве модулей программы, например:

- кнопка **Variables**, с помощью которой выбираются анализируемые переменные;
- кнопка **Summary** (Результат) – выводит результаты анализа;
- кнопка **Options** (Опции) – позволяет настроить внешний вид программы и окон вывода результатов анализа;
- стандартная для Windows кнопка Cancel (Отмена). Кроме того, это окно имеет несколько закладок.

По умолчанию перед пользователем первой предстает закладка **Quick** (Быстро). Находясь на ней, можно выполнить следующие операции:

- Рассчитать показатели описательной статистики – **кнопка Summary: Descriptive statistics**. Перечень рассчитываемых показателей определяется настройками, заданными на другой закладке окна – **Advanced**.
- Получить таблицу частот встречаемости каждого значения анализируемой переменной – кнопка **Frequency tables**;
- Построить частотное распределение значений анализируемой переменной в виде гистограммы – кнопка **Histograms**. Автоматически вместе с гистограммой программа нарисует теоретически ожидаемую нормальную кривую, глядя на которую можно заключить, подчиняются ли анализируемые данные нормальному закону распределения.

• Построить для выбранной переменной (или для нескольких переменных одновременно) так называемую диаграмму размаха – кнопка **Box & whisker plot for all variables**.

Для расчета подробного перечня показателей описательной статистики следует воспользоваться другой закладкой модуля – **Advanced**. Основную часть этой закладки занимает список статистических показателей:

• Valid N – объем выборки;
• Mean – арифметическая средняя;
• Sum – сумма значений анализируемой переменной;
• Median – медиана;
• Mode – мода;
• Geom. mean – геометрическая средняя;
• Harm. mean – гармоническая средняя;
• Standard Deviation – стандартное отклонение;
• Variance – дисперсия;
• Std. err. of mean – стандартная ошибка средней;
• Conf. limits for means: Interval % – доверительные пределы для средних: ширина доверительного интервала;
• Skewness – коэффициент асимметрии;
• Std. err., Skewness – стандартная ошибка коэффициента асимметрии;
• Kurtosis – коэффициент эксцесса;
• Std. err., Kurtosis – стандартная ошибка коэффициента эксцесса;
• Minimum & maximum – минимальное и максимальное значения;
• Lower & upper quartiles – нижний и верхний квартили;
• Percentile boundaries: First & Second: первый и второй процентили;
• Range – размах;
• Quartile range – интерквартильный размах.

На закладке **Advanced** имеются также следующие кнопки:

• Select all stats – позволяет выбрать для расчета сразу все имеющиеся статистические показателей;
• Reset – сброс «галочек» у всех показателей;
• Save settings as default – используя эту кнопку, можно сохранить определенный набор показателей, которые программа будет предлагать для расчета по умолчанию при каждом запуске модуля **Descriptive Statistics**.

Поставьте соответствующие «галочки» и рассчитайте Valid N (объем выборки), Mean (арифметическая средняя), Median (медиана), Mode (мода), Standard Deviation (стандартное отклонение), Std. err. of mean (стандартная ошибка средней), Lower & upper quartiles (нижний и верхний квартили); Range (размах); Quartile range – интерквартильный размах.

Следующей за **Advanced** идет закладка **Normality** (Нормальность) (рис. 1.9). Это важная составляющая модуля описательной статистики, которой предстоит пользоваться очень часто. Здесь можно определить, насколько статистически значимо частотное распределение анализируемых данных отличается от нормального распределения. Наиболее важными элементами этой закладки являются:

- кнопки **Frequency tables** и **Histograms**;
- Поле **Categorization** (Категоризация): воспользовавшись опцией Number of intervals, можно задать количество «столбиков» на гистограмме. Эта опция используется в случаях, когда анализируемый биологический признак является непрерывным. Если же он дискретен, то есть выражается только целыми числами, следует выбрать опцию Integral intervals (Categories).
- Опция **Normal expected frequencies** (Ожидаемые нормальные частоты): при ее выборе и последующем нажатии на кнопку Frequency tables программа выдаст таблицу, которая помимо фактических частот численных значений переменной, будет содержать также теоретически ожидаемые нормальные частоты.
- Тесты, применяемые для проверки соответствия анализируемых данных закону нормального распределения – Kolmogorov-Smirnov & Lilliefors test for normality и ShapiroWilk's W test.

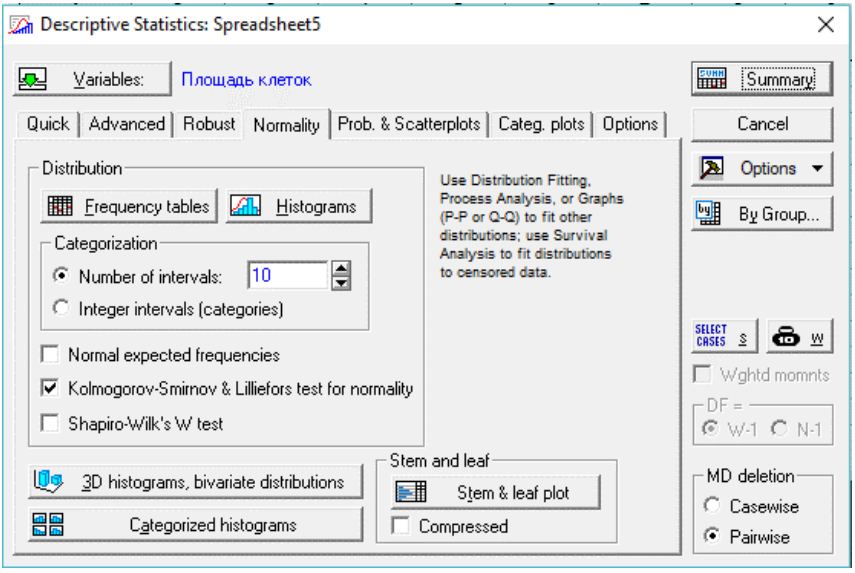


Рис. 1.9. Модуль Descriptive Statistics на закладке Normality

В ряде случаев полезной может оказаться и закладка **Prob. & Scatterplots** (Вероятностные графики и диаграммы рассеяния), следующая за закладкой Normality. В частности, с ее помощью можно построить двух- и трехмерные графики зависимости между анализируемыми переменными, а также проверить данные на нормальность с использованием графика нормальных вероятностей (Normal probability plot).

Диаграммы диапазонов

Результаты практически любого статистического анализа воспринимаются гораздо легче, когда они представлены в графической форме. Рассмотрим несколько типов графиков, которые наиболее часто используются в биологических исследованиях.

Диаграммы диапазонов (wisker plots) удобны для описания временной динамики или пространственного градиента исследуемых величин. Точки на таких графиках чаще всего соответствуют средней арифметической или медиане анализируемого признака. Отличительной особенностью является наличие у точек так называемых «усов» (от англ. «whiskers») – вертикально или горизонтально отходящих линий, длина которых соответствует величине выбранного исследователем показателя разброса данных (минимум и максимум, стандартное отклонение, дисперсия, квартили) или точности оценки генеральных параметров (стандартная ошибка, доверительный интервал).

Рассмотрим следующий пример. В течение 5 месяцев – с мая по сентябрь – ребенку выполняли измерение веса три раза в месяц. Полученные данные приведены на рис. 1.10 (создайте аналогичный файл данных и сохраните его – он потребуется нам также при рассмотрении следующего раздела).

	1 Месяц	2 Вес
1	май	12,9
2	май	13,0
3	май	13,0
4	июнь	13,5
5	июнь	13,6
6	июнь	13,8
7	июль	14,0
8	июль	14,0
9	июль	13,8
10	август	14,2
11	август	14,5
12	август	14,5
13	сентябрь	15,0
14	сентябрь	15,2
15	сентябрь	15,3

Рис. 1.10. Пример оформления данных для построения диаграммы диапазонов

Изобразим графически динамику среднемесячного веса. Обратите внимание на то, каким образом данные располагаются в таблице (рис. 1.10). Чтобы программа «поняла», какие из наблюдений относятся к конкретному месяцу, был введен дополнительный столбец «Месяц». В нем перечислены названия месяцев, а в соседнем столбце – «Вес» – приведены сами значения исследуемой переменной. Такой способ оформления данных характерен для многих видов статистического анализа, реализованных в Statistica, и будет неоднократно встречаться нам в дальнейшем. Столбец «Сезон» в терминах программы называется «группирующей переменной» (**Grouping variable**), а столбец, в котором непосредственно находятся значения исследуемого признака – зависимой переменной (**Dependent variable**). Последнее название является очень подходящим, поскольку указывает на то, что, например, температура воды зависит от сезона года. Для построения диаграммы диапазонов необходимо в разделе **Graphs** основного меню выбрать **2D Graphs**, а затем – **Means w/Error plots** (Графики средних с ошибками). Вешний вид появляющегося в результате этого окна представлен на рис. 1.11.

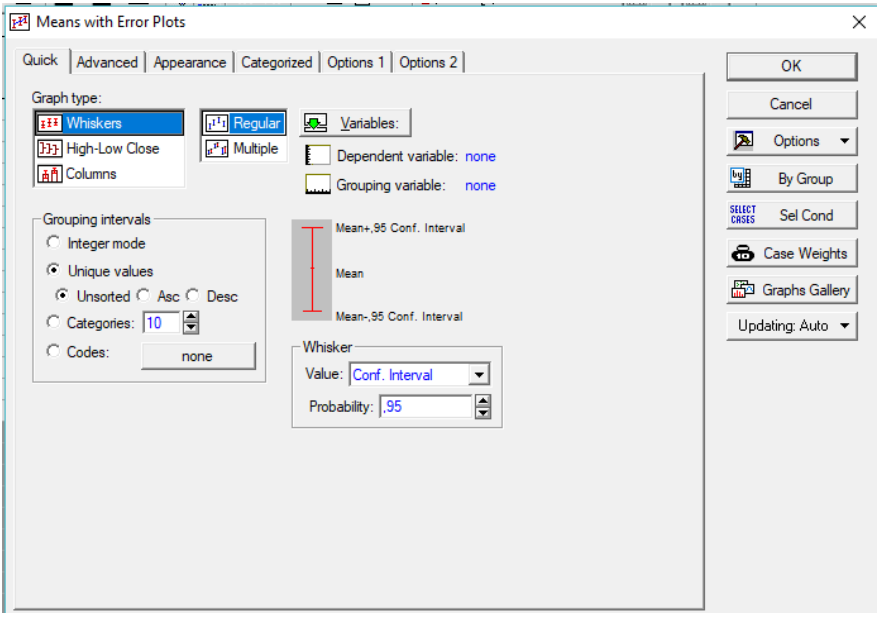


Рис. 1.11. Окно настройки параметров диаграммы диапазонов

Как обычно, начинать следует с указания переменных, которые будут участвовать в анализе – для этого необходимо воспользоваться кнопкой **Variables**. Появится диалоговое окно (**Select variables for means with**

error plots) с двумя списками имеющихся в таблице переменных. В левом списке необходимо выбрать зависимую переменную (в нашем случае это «Вес»), а в правом – группирующую переменную («Месяц»). После этого – нажать кнопку ОК. Далее в поле **Grouping intervals** (Группирующие интервалы) нужно указать программе, на какие интервалы ей следует разбить ось X. В нашем примере вдоль оси X должны располагаться названия месяцев. Чтобы сообщить это программе, нажимаем кнопку **Codes** (Коды), а на появляющейся панели – кнопки **All** (Все) и ОК (Пояснение: в качестве кодов в нашем примере выступают названия месяцев. Поскольку мы хотим, чтобы на графике были отображены данные для всех месяцев, в течение которых выполнялись измерения температуры, необходимо нажать кнопку **All**). Осталось указать, чему на графике будут соответствовать «усы», отходящие от точек. Для этого служит поле **Whisker**. Предположим, мы хотим, чтобы длина «усов» была бы равна доверительному интервалу. В выпадающем меню **Value** (Значение) выбираем **Conf.Interval** (Доверительный интервал), а в поле **Probability** оставим по умолчанию 0,95 (рис. 1.11). Теперь все основные настройки завершены. После нажатия на кнопку ОК можно будет увидеть график, подобный приведенному на рис. 2.12.

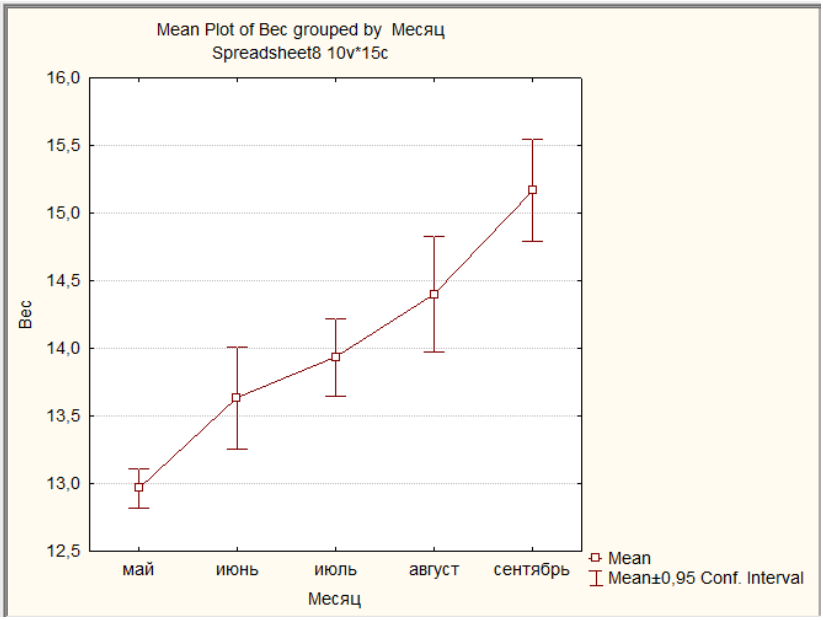


Рис. 1.12. Диаграмма диапазонов, построенная по данным веса ребенка

Диаграммы размахов

Диаграммы размаха, или «ящики с усами» (от англ. «box-and-whisker plots»), получили свое название за характерный вид: точку, соответствующую средней арифметической или медиане, окружает вертикально расположенный прямоугольник («ящик»), длина которого соответствует одному из показателей разброса или точности оценки генерального параметра. Дополнительно от этого прямоугольника отходят «усы», также соответствующие по длине одному из показателей разброса или точности. Таким образом, графики этого типа позволяют дать очень полную статистическую характеристику для каждой анализируемой выборки. Диаграммы размаха можно использовать для визуальной экспресс-оценки разницы между двумя или более группами (например, между датами отбора проб, экспериментальными группами, участками пространства и т. п.).

Для построения диаграммы диапазонов необходимо в разделе **Graphs** основного меню выбрать **2D Graphs**, а затем – **Box plots**. На рис. 1.13 представлен внешний вид данного модуля, открытый на закладке **Advanced**.

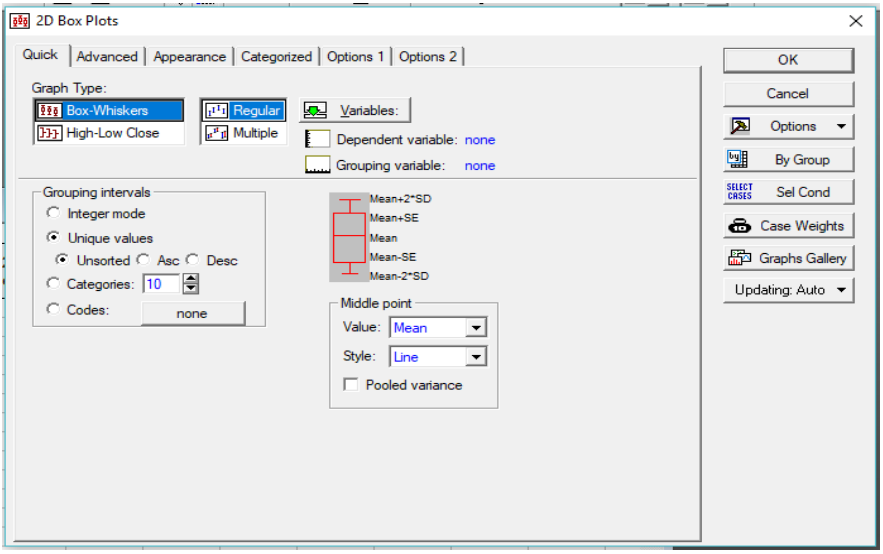


Рис. 1.13. Окно настройки параметров диаграммы размахов

Предположим, что мы хотим визуально сравнить, различается ли среднемесячный вес ребенка в июне и сентябре. Для построения графика необходимо установить следующие настройки (см. рис. 1.13):

• На закладке **Advanced** нажать на кнопку **Variables** и указать, какая из переменных является зависимой (**Dependent**) («Вес»), а какая – группирующей (**Grouping**) («Месяц»).

• В поле **Grouping intervals** выбрать опцию **Codes**, а затем нажать кнопку **Specify codes** (Определить коды), чтобы указать программе, какие именно месяцы будут участвовать в анализе. В появившемся окне ввести через пробел слова «Июнь» и «Сентябрь».

• В выпадающем меню **Value** поля **Middle point** (Средняя точка) выбрать **Mean** (Арифметическая средняя). Так мы сообщим программе, что на графике в качестве точек ей следует изображать средние значения веса.

• В выпадающем меню **Value** поля **Box** выбрать, статистический показатель, который будет изображен в виде «ящика» (например, **Std error** – Стандартная ошибка). **Coefficient** выставить на 1.

• В выпадающем меню **Value** поля **Whisker** выбрать статистический показатель, который будет изображен в виде «усов» (например, **Std dev** – Стандартное отклонение).

• В поле **Outliers** (Выбросы) выбрать **Off** (Отключить). (Пояснение: в результате этого действия программа не будет изображать на графике точки-выбросы, то есть значения признака, которые слишком велики или слишком малы по сравнению с остальными значениями в выборке.) Остальные настройки можно оставить без изменений. После нажатия на кнопку ОК появится график, подобный приведенному на рис. 1.14.

На полученном графике хорошо видно, что среднемесячный вес в сентябре был значительно выше, чем в июне.

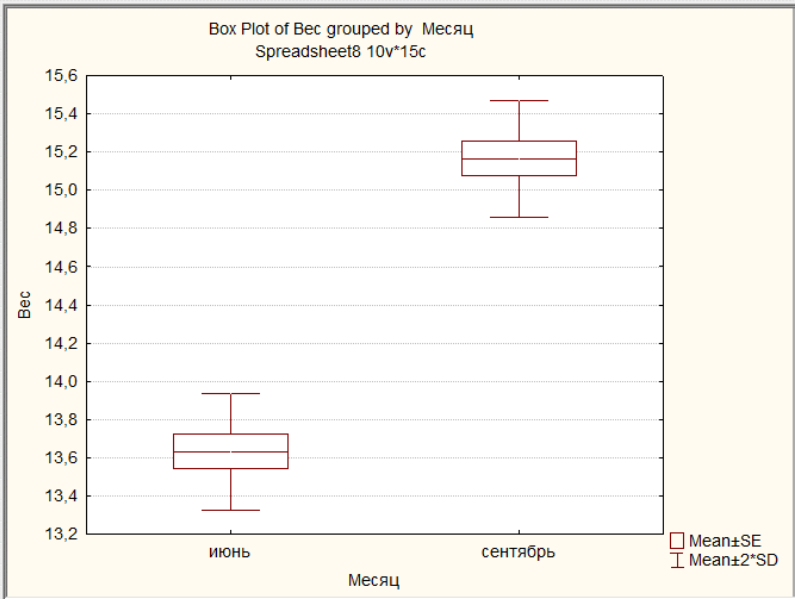


Рис. 1.14. Диаграмма размахов, построенная по данным о весе за июнь и сентябрь

Круговые диаграммы

Круговые диаграммы (pie charts) удобны при анализе качественных признаков. Например, такой график хорошо подойдет для описания соотношения стадий болезни в изучаемой выборке. Допустим, при обследовании 15 человек получены результаты, приведенные на рис. 1.15.

	1 Стадия	V
1	первая	
2	первая	
3	вторая	
4	третья	
5	первая	
6	третья	
7	вторая	
8	четвертая	
9	вторая	
10	первая	
11	третья	
12	четвертая	
13	вторая	
14	первая	
15	первая	

Рис. 1.15. Пример оформления данных для построения круговой диаграммы

Для построения круговой диаграммы, секторы которой были бы пропорциональны долям каждого из вариантов стадии, необходимо выполнить следующее:

- В разделе **Graphs** главного меню выбрать **2D Graphs** > **Pie chart.**
- В появившемся окне перейти на закладку **Advanced** (рис. 1.16).

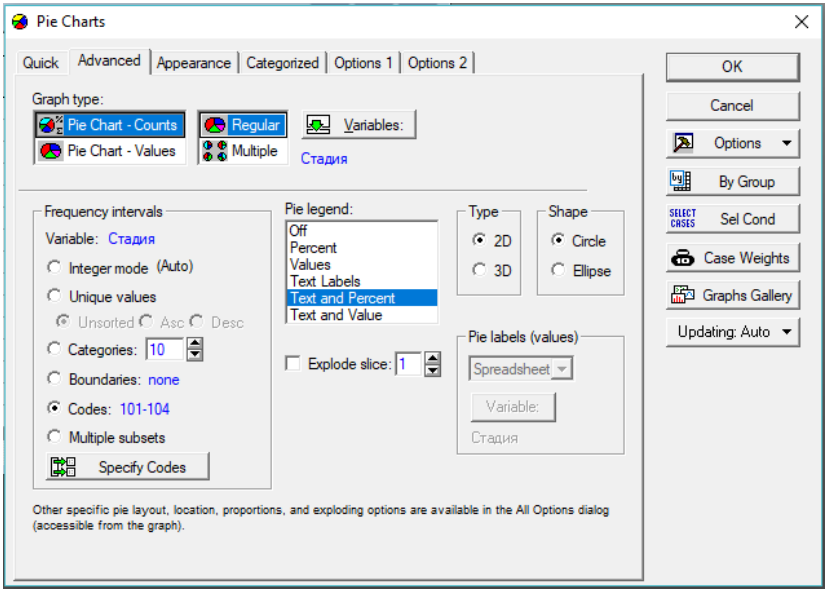


Рис. 1.16. Окно настройки параметров круговой диаграммы

- В поле **Graph type** (Тип графика) выбрать **Pie chart** – **Counts** (Круговая диаграмма – Счет). Данная опция позволяет построить график на основе исходных данных – программа сама подсчитает, сколько в анализируемой совокупности было пациентов первой, второй, третьей и четвертой стадии заболевания. Если бы мы ввели в таблицу предварительно рассчитанные доли каждой из стадий, то в поле **Graph type** следовало бы выбрать **Pie chart – Values** (Круговая диаграмма – Значения). Однако при этом пришлось бы оформить таблицу с данными несколько по-иному.
- В поле **Frequency intervals** (Интервалы частот) выбрать опцию **Codes** и нажать кнопку **Specify codes**. На появившейся панели нажать кнопку All, а затем OK.
- В поле **Pie legend** (Легенда диаграммы) выбрать вариант того, как будут подписаны сегменты круговой диаграммы. Например, при выделении Text and Percent будут отображены названия вариантов стадий и частота (%), с которой встречается каждый вариант в выборке. Остальные настройки можно оставить неизменными. Нажатие на кнопку OK приведет к построению графика, подобного приведенному на рис. 1.17.

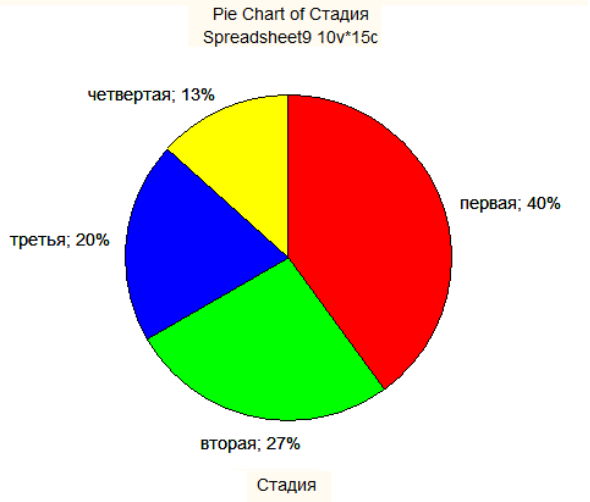


Рис. 1.17. Круговая диаграмма, построенная по данным об окраске цветков

1.2. Оценка выборочных характеристик с использованием специальных функций. Подбор закона и параметров распределения

Подбор закона и параметров распределения. Проверка на нормальность распределения.

Как известно, существующие методы статистического анализа можно подразделить на две группы – параметрические и непараметрические. Важным условием, определяющим возможность применения параметрических методов, является подчинение анализируемых данных закону нормального (Гауссова) распределения, которое имеет характерный колоколообразный вид. В то же время непараметрические методы выполнения этого условия не требуют. Установлено, что в подавляющем большинстве случаев (около 75 %) распределения биологических признаков существенно отличаются от нормального. Тем не менее, очень многие исследователи-биологи совершают ошибку, применяя параметрические методы анализа для ненормально распределенных признаков. Часто это приводит к выводам, не соответствующим действительности. Во избежание указанной ошибки, любой анализ биологических признаков должен сопровождаться проверкой нормальности их распределения. Для этого существует достаточно широкий набор методов. Мы рассмотрим три подхода, реализованные в программе Statistica.

Подгонка распределения

На рис. 1.18 представлены результаты измерения роста у 20 студентов мужского пола. Необходимо установить, распределены ли эти данные по нормальному закону.

	1 Рост, см	V
1	164	
2	168	
3	175	
4	175	
5	182	
6	175	
7	186	
8	180	
9	176	
10	175	
11	187	
12	169	
13	172	
14	170	
15	183	
16	178	
17	187	
18	173	
19	180	
20	179	

Рис. 1.18. Данные о росте 20 студентов (см)

В программе Statistica имеется специальный модуль – **Distribution fitting** (Подгонка распределения), позволяющий проверить соответствие анализируемых данных целому ряду математических распределений. Его можно запустить из раздела главного меню Statistics, или нажав кнопку на дополнительной панели инструментов. Внешний вид окна этого модуля приведен на рис. 1.19.

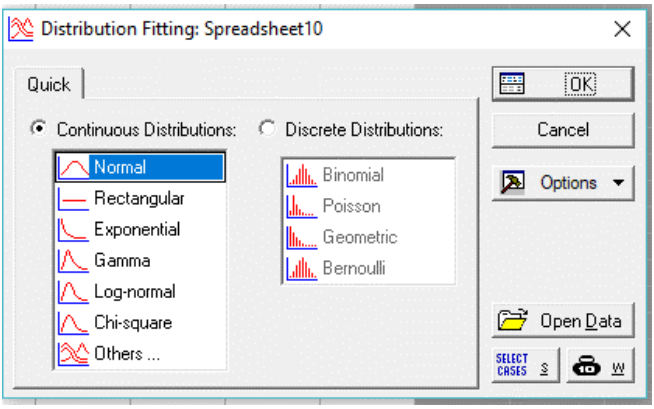


Рис. 1.19. Диалоговое окно модуля Distribution fitting

Поскольку нам необходимо проверить, подчиняются ли данные о росте студентов нормальному распределению, в списке непрерывных распределений (**Continuous distributions**) выбираем **Normal**, после чего нажимаем кнопку OK (рис. 1.19). Далее появится еще одно диалоговое

окно (рисунок не приводится), где необходимо указать, какую именно переменную мы хотим проанализировать, и как. Переменная для анализа задается путем нажатия кнопки **Variables**. На вкладке **Parameters** установите число интервалов 6 в окне **Number of categories**. На вкладке **Options** снимите «галочку» в окне **Chi-Square test**. Остальные настройки можно оставить без изменений. Нажав кнопку **Plot of observed and expected distributions** (График наблюдаемого и ожидаемого распределений), получим гистограмму распределения данных о росте студентов и колоколообразную красную кривую, соответствующую ожидаемому нормальному распределению (у этого ожидаемого распределения те же средняя арифметическая и стандартное отклонение, что и в анализируемой совокупности данных) (рис. 1.20).

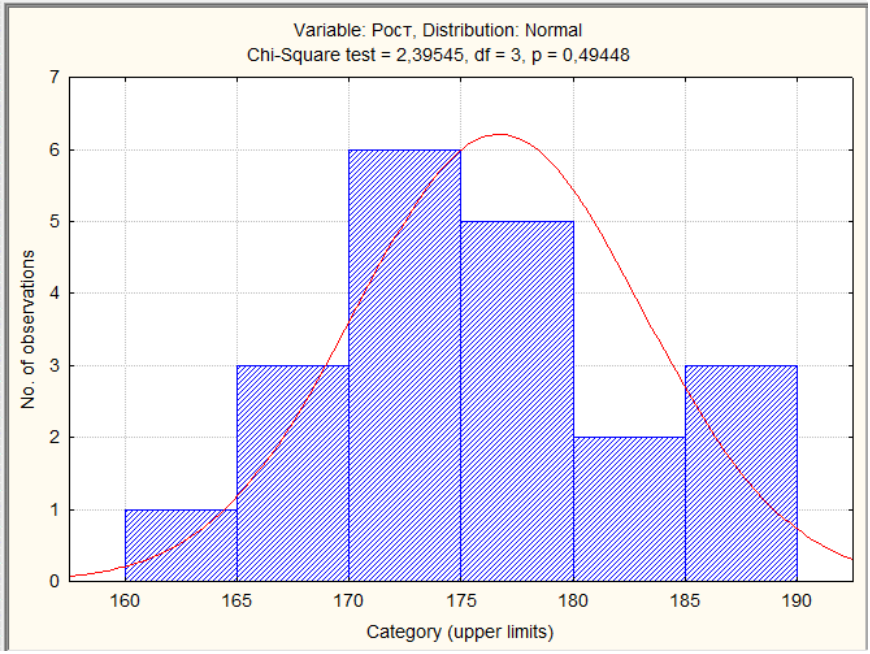


Рис. 1.20. Результат анализа, выполненного с помощью модуля Distribution fitting

Глядя на полученный рисунок, можно сказать, что в целом распределение значений роста студентов соответствует нормальному (столбики гистограммы образуют колоколообразную фигуру). Это заключение, основанное на визуальном анализе распределения, имеет и более строгое подтверждение в виде результатов теста χ^2 (Chi-square test, см. в верхней части графика на рис. 1.20). В данном случае этот тест проверяет нулевую гипотезу о том, что наблюдаемое распределение анализируемого

признака не отличается от теоретически ожидаемого нормального распределения. Поскольку вероятность ошибиться, отклонив эту гипотезу оказалась намного больше 0,05 (P = 0,49448), мы принимаем, что гипотеза действительно верна. Иными словами, распределение значений роста студентов статистически не отличается от нормального распределения.

Тесты Колмогорова–Смирнова и Шапиро–Уилка

Следует отметить, что мощность теста хи-квадрат (χ^2) при проверке нормальности распределения анализируемых данных относительно невысока (другими словами, его применение достаточно часто приводит к ошибочному выводу о нормальности распределения). Поэтому лучше воспользоваться другими тестами. Их можно найти в уже рассмотренном выше модуле **Descriptive Statistics** (Описательная статистика). После запуска этого модуля необходимо открыть закладку **Normality** и в поле **Distribution** (Распределение) разыскать опции **Kolmogorov-Smirnov and Lilliefors test for normality** (Тест Колмогорова–Смирнова и Лиллифорса на нормальность) и **Shapiro-Wilk's W test** (W-тест Шапиро–Уилка). Равно как и критерий χ^2, эти тесты проверяют нулевую гипотезу об отсутствии различий между наблюдаемым распределением признака и теоретическим ожидаемым нормальным распределением. Наиболее предпочтительным, особенно при небольших выборках (n = 3÷50) является использование W-критерия Шапиро-Уилка, поскольку он обладает наибольшей мощностью в сравнении со всеми перечисленными критериями (чаще выявляет различия между распределениями в тех случаях, когда они действительно есть). Для выбора того или иного теста, достаточно поставить флажок рядом с его названием. После выбора анализируемой переменной (кнопка **Variables**) и нажатия кнопки **Histograms** программа создаст гистограмму распределения значений признака и ожидаемую нормальную кривую (рис. 1.21).

Результаты выбранных тестов на нормальность автоматически располагаются в заголовке этого графика. При p > 0,05 можно заключить, что анализируемое распределение не отличается от нормального. В примере с данными о росте студентов для теста Шапир—Уилка получаем p = 0,7979 (рис. 1.21), что подтверждает сделанный ранее вывод о нормальности распределения этих данных.

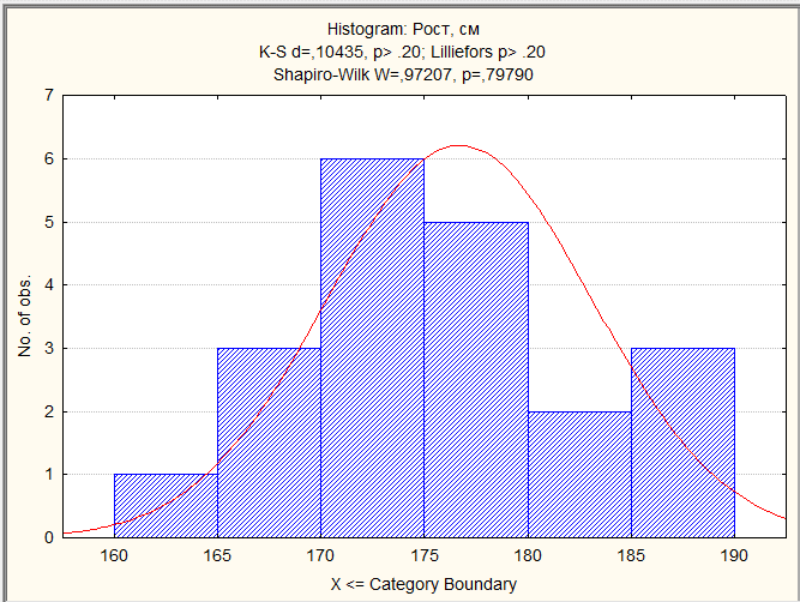


Рис. 1.21. Результат проверки нормальности распределения данных, выполненной при помощи модуля Descriptive Statistics

<u>График нормальных вероятностей</u>

В модуле **Descriptive Statistics** реализован еще один способ проверки данных на нормальность распределения. Он заключается в использовании графика нормальных вероятностей, или «вероятностной бумаги». Такой график изображает зависимость ожидаемых нормальных частот значений признака от их реальных частот. Очевидно, что если между наблюдаемым и ожидаемым распределениями нет никакой разницы, точки на этом графике выстроятся строго вдоль прямой. Иначе они образуют фигуру отличную от прямой. На этом принципе и основано применение «вероятностной бумаги». Для построения графика такого типа необходимо в модуле **Descriptive Statistics** перейти на закладку **Prob. & Scatterplots** (Вероятностные графики и диаграммы рассеяния) и нажать на кнопку **Normal probability plot** (График нормальных вероятностей). В результате появится график, подобный приведенному на рис. 1.22.

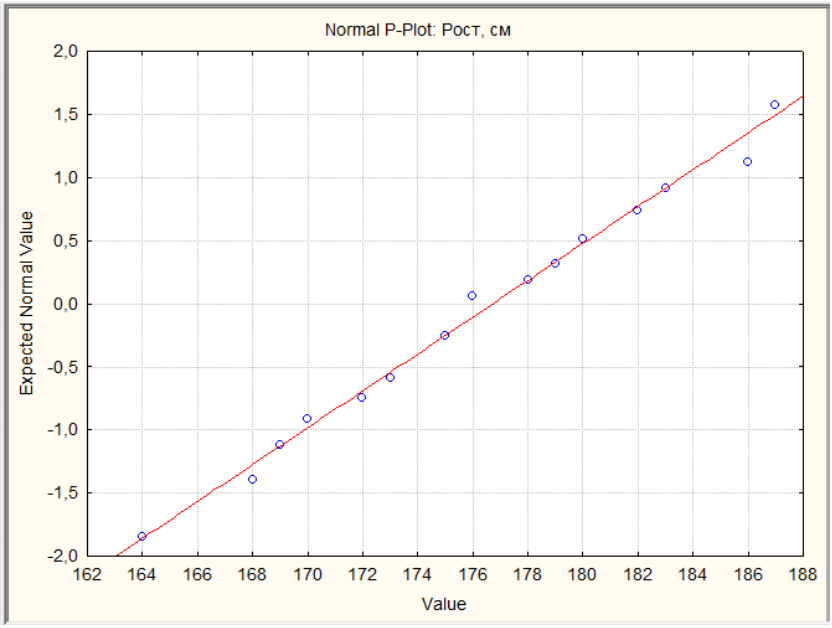


Рис. 1.22. Проверка нормальности распределения данных о росте студентов с использованием графика нормальных вероятностей

Точки на этом рисунке плотно выстраиваются вдоль теоретически ожидаемой прямой, что еще раз подтверждает нормальность распределения данных о росте студентов.

Глава 2. Классические методы и критерии статистики

2.1. Анализ качественных признаков

Оценка статистической значимости различий. Формулировка нулевой и альтернативной гипотез. Уровень значимости критерия.

В предыдущей главе мы производили анализ количественных признаков. Примером таких признаков служат артериальное давление, количество дней госпитализации, время послеродовой активности и т. д. Единицей их измерения могут быть миллиметры ртутного столба, часы или дни. Над значениями количественных признаков можно производить арифметические действия. Можно, например, сказать, что артериальное давление снизилось на какое-то количество единиц. Кроме того, их можно упорядочить: расположить в порядке возрастания или убывания.

Однако очень многие признаки невозможно измерить числом. Например, можно быть либо мужчиной, либо женщиной, либо, больным либо здоровым. Это качественные признаки. Они не связаны между собой никакими арифметическими соотношениями, упорядочить их также нельзя. Единственный способ описания качественных признаков состоит в том, чтобы подсчитать число объектов, имеющих одно и то же значение. Кроме того, можно подсчитать, какая доля от общего числа объектов приходится на то или иное значение.

Сравнение частот при наличии таблиц сопряженности 2×2 в двух несвязанных выборках с помощью критерия хи-квадрат

Условия и ограничения применения критерия хи-квадрат Пирсона

1. Сопоставляемые показатели должны быть измерены в номинальной шкале (например, пол пациента – мужской или женский) или в порядковой (например, степень артериальной гипертензии, принимающая значения от 0 до 3).

2. Данный метод позволяет проводить анализ не только четырехпольных таблиц, когда и фактор, и исход являются бинарными переменными, то есть имеют только два возможных значения (например, мужской или женский пол, наличие или отсутствие определенного заболевания в анамнезе...). Критерий хи-квадрат Пирсона может применяться и в случае анализа многопольных таблиц, когда фактор и (или) исход принимают три и более значений.

3. Сопоставляемые группы должны быть независимыми, то есть критерий хи-квадрат не должен применяться при сравнении наблюдений «до-после». В этих случаях проводится тест Мак-Немара (при сравнении

двух связанных совокупностей) или рассчитывается Q-критерий Кохрена (в случае сравнения трех и более групп).

4. При анализе четырехпольных таблиц ожидаемые значения в каждой из ячеек должны быть не менее 10. В том случае, если хотя бы в одной ячейке ожидаемое явление принимает значение от 5 до 9, критерий хи-квадрат должен рассчитываться с поправкой Йейтса. Если хотя бы в одной ячейке ожидаемое явление меньше 5, то для анализа должен использоваться точный критерий Фишера.

5. В случае анализа многопольных таблиц ожидаемое число наблюдений не должно принимать значения менее 5 более чем в 20 % ячеек.

Рассмотрим учебный пример. Гемодиализ позволяет сохранить жизнь людям, страдающим хронической почечной недостаточностью. При гемодиализе кровь больного пропускают через искусственную почку – аппарат, удаляющий из крови продукты обмена веществ. Искусственная почка подсоединяется к артерии и вене больного: кровь из артерии поступает в аппарат и оттуда, уже очищенная – в вену. Так как гемодиализ проводится регулярно, больному устанавливают артериовенозный шунт. В артерию и вену на предплечье вводят тефлоновые трубки; их концы выводят наружу и соединяют друг с другом. При очередной процедуре гемодиализа трубки разъединяют между собой и присоединяют к аппарату. После диализа трубки вновь соединяют, и кровь течет по шунту из артерии в вену. Завихрения тока крови в местах соединения трубок и сосудов приводят к тому, что шунт часто тромбируется. Тромбы приходится регулярно удалять, а в тяжелых случаях даже менять шунт. Руководствуясь тем, что аспирин препятствует образованию тромбов, Г. Хартер и соавт. решили проверить, нельзя ли снизить риск тромбоза назначением небольших доз аспирина (160 мг/сут). Было проведено контролируемое испытание. Все больные, согласившиеся на участие в испытании и не имевшие противопоказании к аспирину, были случайным образом разделены на две группы: 1-я получала плацебо, 2-я – аспирин. Ни врач, дававший больному препарат, ни больной не знали, был это аспирин или плацебо. Такой способ проведения испытания (он называется двойным слепым) исключает «подсуживание» со стороны врача или больного, и, хотя технически сложен, дает наиболее надежные результаты. Исследование проводилось до тех пор, пока общее число больных с тромбозом шунта не достигло 25. Группы практически не различались по возрасту, полу и продолжительности лечения гемодиализом.

В 1-ой группе тромбоз шунта произошел у 18 из 25 больных, во 2-ой – у 6 из 19 (табл. 2.1). Можно ли говорить о статистически значимом различии доли больных с тромбозом, а тем самым об эффективности аспирина? Таблица результатов исследования представлена в следующем виде:

Таблица 2.1

Влияние аспирина на тромбоз: таблица сопряженности

Показатели	Плацебо	Аспирин
Тромбоз есть	18	6
Тромбоза нет	7	13

Нулевая гипотеза: аспирин не влияет на возникновение тромбоза шунта.

Уровень значимости принимается 0,05.

Запустите программу Statistica, создайте новый документ. В меню выберите **Анализ — Непараметрическая статистика/ Statistics-Nonparametric>— Таблицы 2×2/ 2×2 Tables(X/V/Phi, McNemar, Fisher exact) >OK.**

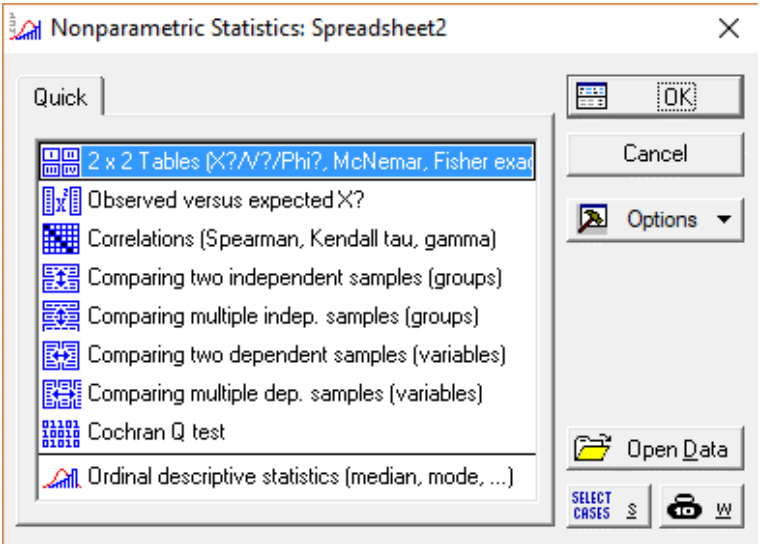


Рис. 2.1. Вид окна выбора метода анализа

В появившемся окне введите значения из полученной таблицы сопряженности. При этом левый столбец соответствует левому столбику таблицы (Плацебо), а правый соответствует правому (Аспирин). Аналогичная ситуация и со строками (рис. 2.2).

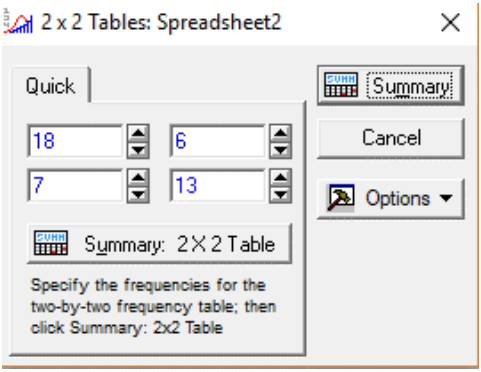


Рис. 2.2. Вид окна ввода данных

Нажмите **Summary**. Появится таблица с результатами статистической обработки (рис. 2.3).

	2 x 2 Table (Spreadsheet2)		
	Column 1	Column 2	Row Totals
Frequencies, row 1	18	6	24
Percent of total	40,909%	13,636%	54,545%
Frequencies, row 2	7	13	20
Percent of total	15,909%	29,545%	45,455%
Column totals	25	19	44
Percent of total	56,818%	43,182%	
Chi-square (df=1)	7,11	p= ,0077	
V-square (df=1)	6,95	p= ,0084	
Yates corrected Chi-square	5,58	p= ,0182	
Phi-square	,16168		
Fisher exact p, one-tailed		p= ,0087	
two-tailed		p= ,0138	
McNemar Chi-square (A/D)	,52	p= ,4725	
Chi-square (B/C)	0,00	p=1,0000	

Рис. 2.3. Вид таблицы с результатами статистической обработки

Так, из таблицы следует, что у больных, принимавших аспирин, тромбозы наблюдались в 13,6 % случаев против 40,9 % больных, принимавших плацебо. Однако необходимо оценить статистическую значимость полученного различия с помощью правильно подобранного критерия.

Так как в данном случае анализ проводился таблицы сопряженности 2×2, то необходимо учитывать поправку Йейтса. Исходя из полученных значений **критерия хи-квадрат** (5,58) и **вероятности p** (0,0182), следует заключить, что видимые различия в клетках таблицы сопряженности значимы. Поэтому нулевая гипотеза **отвергается**. Аспирин действительно положительно влияет на снижение вероятности возникновения тромбоза шунта.

Сравнение качественных признаков (выраженных в частотах) в 2-х независимых группах с помощью точного метода Фишера

Условия применимости критерия Фишера:

1. Сравниваемые переменные должны быть измерены в номинальной шкале и иметь только два значения, например, артериальное давление в норме или повышено, исход благоприятный или неблагоприятный, послеоперационные осложнения есть или нет.

2. Точный критерий Фишера предназначен для сравнения двух независимых групп, разделенных по факторному признаку. Соответственно, фактор также должен иметь только два возможных значения.

3. Критерий подходит для сравнения очень малых выборок: точный критерий Фишера может применяться для анализа четырехполных таблиц в случае значений ожидаемого явления менее 5, что является ограничением для применения критерия хи-квадрат Пирсона, даже с учетом поправки Йейтса.

4. Точный критерий Фишера бывает односторонним и двусторонним. При одностороннем варианте точно известно, куда отклонится один из показателей. Например, во время исследования сравнивают, сколько пациентов выздоровело по сравнению с группой контроля. Предполагают, что терапия не может ухудшить состояние пациентов, а только либо вылечить, либо нет.

Двусторонний тест оценивает различия частот по двум направлениям. То есть оценивается вероятность как большей, так и меньшей частоты явления в экспериментальной группе по сравнению с контрольной группой.

Рассмотрим учебный пример. Т. Бишоп изучил эффективность высокочастотной стимуляции нерва в качестве обезболивающего средства при удалении зуба (табл. 2.2). Все больные подключились к прибору, но в одних случаях он работал, в других был выключен. Ни стоматолог, ни больной не знали, включен ли прибор. Позволяют ли следующие данные считать высокочастотную стимуляцию нерва действенным анальгезирующим средством?

Таблица 2.2

Эффективность высокочастотной стимуляции нерва в качестве обезболивающего средства при удалении зуба (таблица сопряженности)

	Прибор включен	Прибор выключен	Всего
Боли нет	7	2	9
Боль есть	1	8	9
Всего	8	10	18

Нулевая гипотеза: пациенты испытывают боль одинаково часто как при включенном, так и при выключенном приборе.

Рассчитаем ожидаемые числа. При включенном приборе находились 8 пациентов, при выключенном 10. Боль возникла у 9 из 18, то есть в 50 % случаев, не возникла в 9 из 18 (также в 50 %). Если мы приняли за нулевую гипотезу предположение о том, что прибор не влияет на частоту возникновения боли, то рассчитав, сколько составляет 50 % от 10 и 8, мы узнаем ожидаемые числа. Ожидаемые числа для пациентов с болью при включенном приборе и выключенном приборе составляют 5 и 4. Таким же образом рассчитываем ожидаемы числа для пациентов без боли. Они также составят 50 %, но от чисел 9 и 9. Это будут 4,5 и 4,5. Следовательно, необходимо использовать точный критерий Фишера, который представлен двусторонним и односторонним. Двусторонний вариант любого критерия выбирают, если нужно проверить, существуют ли статистически значимые различия. Односторонний вариант любого критерия выбирают, если нужно показать, что изучаемый фактор повышает или понижает результат. Проверим справедливость нулевой гипотезы с помощью двустороннего критерия.

Повторим действия, аналогичные, описанным выше. В меню выберите **Анализ — Непараметрическая статистика / Statistics-Nonparametric > Таблицы 2×2/ 2×2 Tables(X/V/Phi, McNemar, Fisher exact) >OK.**

Введите в строки значения частот соответственно таблице. Получится следующее окно (рис. 2.4):

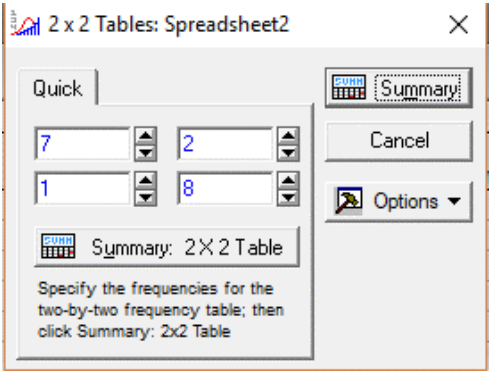


Рис. 2.4. Вид окна ввода данных

Нажмите **Summary**. Появится таблица с результатами статистической обработки.

	2 x 2 Table (Spreadsheet2)		
	Column 1	Column 2	Row Totals
Frequencies, row 1	7	2	9
Percent of total	38,889%	11,111%	50,000%
Frequencies, row 2	1	8	9
Percent of total	5,556%	44,444%	50,000%
Column totals	8	10	18
Percent of total	44,444%	55,556%	
Chi-square (df=1)	8,10	p= ,0044	
V-square (df=1)	7,65	p= ,0057	
Yates corrected Chi-square	5,63	p= ,0177	
Phi-square	,45000		
Fisher exact p, one-tailed		p= ,0076	
two-tailed		p= ,0152	
McNemar Chi-square (A/D)	0,00	p=1,0000	
Chi-square (B/C)	0,00	p=1,0000	

Рис. 2.5. Вид таблицы с результатами статистической обработки

Из табл. 2.5 видно, что при включенном приборе 38,9 % пациентов не испытывали боли, а при отключенном приборе боль отсутствовала в 11,1 % случаев. Вероятность нулевой гипотезы по двустороннему критерию Фишера (two-tailed) равна p = 0,0152, что ниже 0,05. Следовательно, мы отвергаем нулевую гипотезу.

Таким образом, применение электроанальгезии статистически значимо (на 27,8 %) снижает частоту ощущения боли (p = 0,0152 по двустороннему критерию Фишера).

Сравнение качественных признаков (выраженных в частотах) в двух связанных выборках с помощью критерия МакНемара

Тест хи-квадрат по методу МакНемара применяют исключительно для дихотомических переменных. При этом для двух зависимых переменных выясняют, происходят ли какие-либо изменения в структуре распределения их значений. В большинстве наблюдений сравнение проводят с учетом временного фактора по схеме «до-после».

Рассмотрим учебный пример. При ринитах широко используют альфа-адреномиметики в качестве вазоконстрикторов, позволяющих уменьшить отек слизистой оболочки носа и снизить секрецию слизи, обеспечивая тем самым восстановление носового дыхания. Известно много препаратов данной фармакологической группы. Детской больницей было закуплено два из них: спироназ и ринолизин.

Под наблюдение взяли 170 детей. Каждому ребенку в первый день закапывали только спироназ, во второй – только ринолизин. Через 10 мин оценивали состояние носового дыхания. Получили следующие результаты (табл. 2.3):

Таблица 2.3

Эффективность действия вазоконстрикторов

Препарат	Эффект	
	Нос заложен	Нос дышит
Спироназ	33	53
Ринолизин	48	36

Необходимо выяснить, какой из препаратов эффективнее.

Нулевая гипотеза: оба препарата дают одинаковую частоту случаев восстановления носового дыхания.

В меню выберите **Анализ — Непараметрическая статистика/ Statistics-Nonparametric >— Таблицы 2×2/ 2×2 Tables(X/V/Phi, McNemar, Fisher exact) >OK**.

Введите в строки значения частот соответственно таблице. Обратите внимание, что это не таблица сопряженности, поскольку каждый объект исследуется дважды. Получится следующее окно:

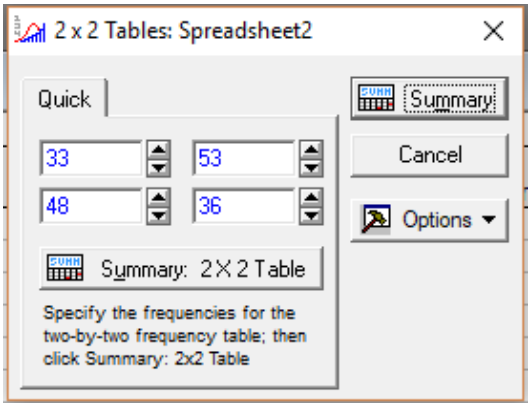


Рис. 2.6. Вид окна ввода данных

Нажмите **Summary**. Появится таблица с результатами статистической обработки.

	2 x 2 Table (Spreadsheet2)		
	Column 1	Column 2	Row Totals
Frequencies, row 1	33	53	86
Percent of total	19,412%	31,176%	50,588%
Frequencies, row 2	48	36	84
Percent of total	28,235%	21,176%	49,412%
Column totals	81	89	170
Percent of total	47,647%	52,353%	
Chi-square (df=1)	6,00	p= ,0143	
V-square (df=1)	5,97	p= ,0146	
Yates corrected Chi-square	5,27	p= ,0217	
Phi-square	,03531		
Fisher exact p, one-tailed		p= ,0107	
two-tailed		p= ,0210	
McNemar Chi-square (A/D)	,06	p= ,8097	
Chi-square (B/C)	,16	p= ,6906	

Рис. 2.7. Вид таблицы с результатами статистической обработки

Из табл. 2.7 видно, что спироназ помог примерно 31,2 % пациентам, ринолизин – 28,2 %.

Критерий МакНемара, подчиняющийся распределению χ^2, представлен в двух вариантах – Chi-square (A/D) и Chi-square (B/C). В нашем случае вероятность нулевой гипотезы определяем по Chi-square (B/C). Данный критерий равен 0,16, а соответствующая вероятность равна 0,69, что выше критического уровня значимости 0,05. Исходя из этого, принимаем нулевую гипотезу.

Таким образом, статистически значимых различий между спироназом и ринолизином по способности восстанавливать носовое дыхание нет ($\chi^2 = 0,16$; $p = 0,69$).

Следует отметить, что если бы мы ошибочно использовали для оценки критерии Хи-квадрат или Фишера (двусторонний), то получили бы противоположный вывод. Поскольку p=0,014 и p=0,021 соответственно, то ошибочно признали бы верной альтернативную гипотезу о различном влиянии препаратов на носовое дыхание.

Тест Кохрена для повторных испытаний

Q критерий Кохрена – непараметрический критерий для проверки значимости различия двух и более воздействий на группы; воздействие (отклик) является дихотомической переменной (принимает два значения – 0/1; да/нет). Если переменные в другой шкале, их нужно дихотомировать.

Q критерий Кохрена является развитием критерия хи-квадрат Макнемара.

Допустим, что в предыдущей задаче мы добавили еще один препарат. Каждому ребенку в первый день закапывали только спироназ, во второй – только ринолизин, в третий – оксиметазолин. Через 10 мин оценивали состояние носового дыхания. Если эффект наблюдался, то кодировали цифрой 1, если нет – цифрой 0. Получили следующие результаты (рис. 2.8):

	1 Спироназ	2 Ринолизин	3 Оксиметазолин
1	0	0	0
2	1	1	0
3	0	0	0
4	0	0	0
5	1	1	0
6	1	1	0
7	1	1	0
8	0	1	0
9	1	0	0
10	0	0	0
11	1	1	1
12	1	1	1
13	1	1	0
14	1	1	0
15	1	1	0
16	1	1	1
17	1	1	0
18	1	1	0

Рис. 2.8. Вид таблицы данных

В меню выберите **Анализ — Непараметрическая статистика/ Statistics-Nonparametric >— Q критерий Кохрена / Cochran Q test >OK**. Нажав кнопку **Переменные/ Variables,** выберите все переменные. Нажмите **Summary.**

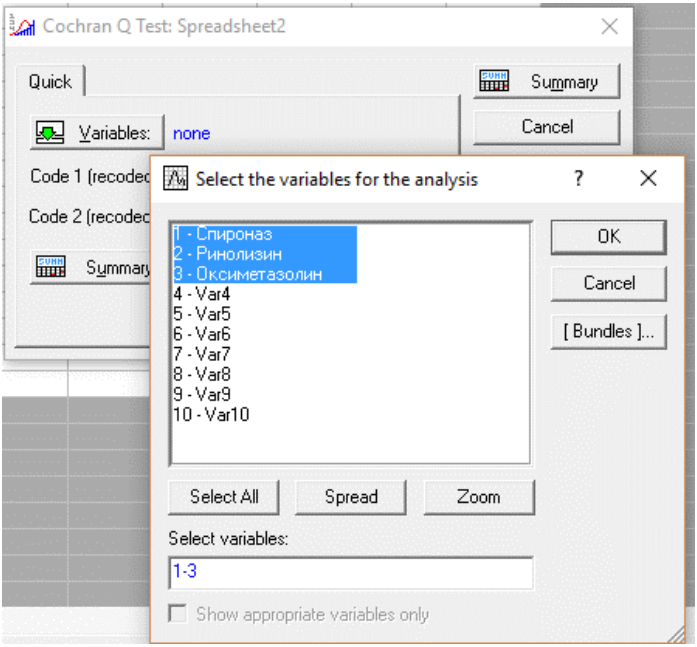


Рис. 2.9. Вид окна выбора переменных

Появится таблица результатов анализа Кохрена (рис. 2.10).

Cochran Q Test (Spreadsheet10)
Number of valid cases:18
Q = 18,18182, df = 2, p < ,000113

Variable	Sum	Percent 0's	Percent 1's
Var1	13,00000	27,77778	72,22222
Var2	13,00000	27,77778	72,22222
Var3	3,00000	83,33333	16,66667

Рис. 2.10. Вид таблицы с результатами статистической обработки

В заголовке окна значение коэффициента Q = 18,18 при уровне значимости меньше 0,000113, что свидетельствует о том, что группы отличаются друг от друга. Таким образом, можем говорить о наличии различий действия препаратов на заложенность носа.

Сравнение двух качественных признаков в двух несвязанных выборках, выраженных в процентах (сравнение относительных частот внутри одной группы и в двух группах)

Необходимо выяснить, отличаются ли две группы больных по наличию какого-либо признака, наличие которого в обеих группах кодируется единицей, а отсутствие – нулем.

Рассмотрим следующий пример. Галотан и морфин по-разному влияют на артериальное давление. Однако для клинициста важнее знать, есть ли различие в операционной летальности? В исследовании из 122 больных, оперированных под галотановой анестезией, умерли16, то есть 13,1 %. При использовании морфина умерли 20 из 134, то есть 14,9 %. Летальность при использовании галотана оказалась примерно на 2 % ниже, чем при использовании морфина. Можно ли считать, что морфин опаснее галотана?

Нулевая гипотеза: летальность при анестезии галотаном и морфином одинакова.

Поскольку выборки независимы, объем их больше 100, данные представлены в виде процентов, то необходимо использовать z-критерий для долей.

В строке операционного меню (рис. 2.11) выбираем **Анализ > Statistic > Basic Statistic/Tables > Difference tests:r, % > OK**

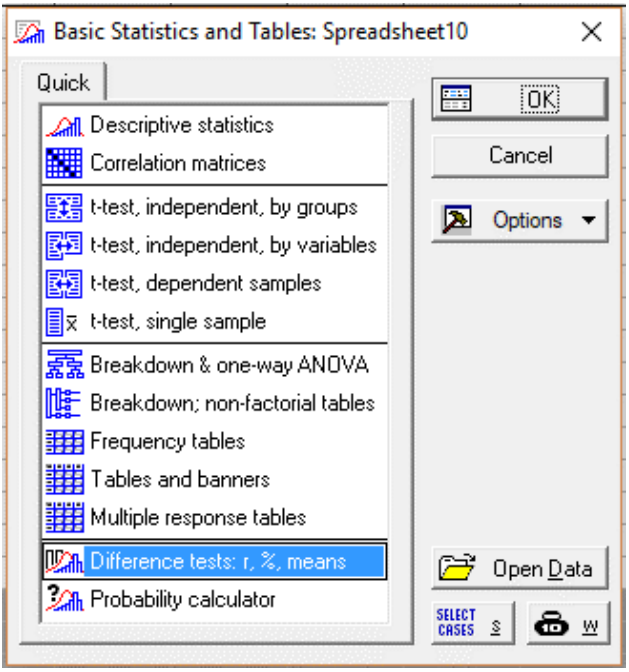


Рис. 2.11. Вид окна выбора метода анализа данных

В новом диалоговом окне в поле **Сравнение двух пропорций (долей)/ Difference between two proportions** в строку **Pr.1** введите процент летальности в группе с галотаном в виде десятичной дроби. Для этого разделите 13,1 % на 100 % и введите число 0,131. Аналогично в строку **Pr.2** – число 0,149. Затем введите объемы выборок $N_1 = 122$, $N_2 = 134$. Тип критерия **Двусторонний /Two sided > Compute** (рис. 2.12).

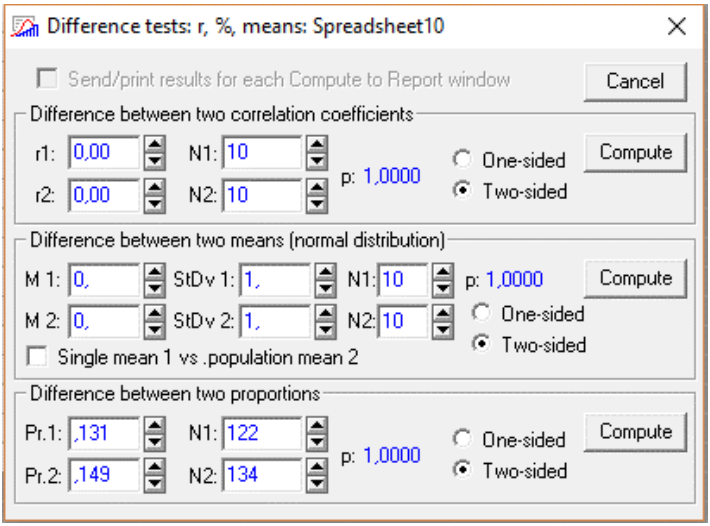


Рис. 2.12. Вид окна ввода данных

Появится окно анализа (рис. 2.13). из которого следует, что вероятность нулевой гипотезы p = 0,6792.

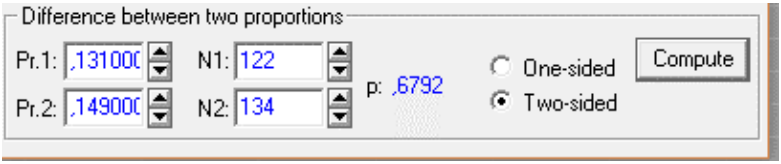


Рис. 2.13. Вид таблицы с результатами статистической обработки

Поскольку 0,6792 больше критического уровня значимости 0,05, то принимаем нулевую гипотезу. Таким образом, статистически значимые различия в летальности при анестезии галотаном и морфином отсутствуют (p = 0,6792).

Замечание: в том случае, когда изучаемый признак в одной из групп встречался 0 % или 100 %, для расчета критерия значимости нужно вводить не о или 1, а числа, близкие к ним, например, 0,0000000000001 и 0,999999999. Такая замена не влияет на результат вычислений.

2.2. Гипотеза о равенстве средних двух генеральных совокупностей. Введение в корреляционный анализ

Гипотеза об однородности дисперсий. Оценка корреляции двух случайных величин.

Одной из обычных задач в биологических исследованиях является сравнение арифметических средних двух групп (например, экспериментальной и контрольной). Классическим методом, позволяющим ее решать, является t-тест Стьюдента, или просто «t-тест». Нулевая гипотеза, проверяемая в ходе данного теста, заключается в том, что обе группы происходят из одной генеральной совокупности; другими словами, что наблюдаемые различия между средними значениями сравниваемых выборок случайны и не вызваны действием изучаемого фактора. Тест Стьюдента относится к группе параметрических методов анализа. Его корректное применение требует выполнения трех условий:

- обе выборки должны быть независимыми: свойства одной из них никак не должны быть связаны со свойствами другой (известно, например, что женщины в среднем ниже мужчин, однако это не является результатом того, что мужчины оказывают какое-то особое влияние на рост женщин – дело здесь в генетических особенностях пола);
- обе выборки должны подчиняться нормальному закону распределения;
- между дисперсиями выборок не должно быть статистически значимой разницы (однородность дисперсий).

К сожалению, многие исследователи-биологи игнорируют перечисленные условия при выполнении теста Стьюдента, что часто приводит к ошибочным результатам. Наиболее «опасным» является несоблюдение требования о нормальности распределения значений признака в сравниваемых группах.

Рассмотрим, как тест Стьюдента можно выполнить при помощи программы Statistica. На рис. 2.14 приведены данные о содержании гемоглобина в крови (г/л) больных сахарным диабетом и здоровых людей. Выборки сопоставимы по полу, возрасту и т. п. Применим *t*-тест для сравнения средних значений этих двух независимых выборок. (*Примечание.* В учебных целях допустим, что обе выборки распределены нормально, и что их дисперсии различаются незначительно).

	1 Группа	2 Гемоглобин
1	Больные	120
2	Больные	113
3	Больные	125
4	Больные	118
5	Больные	116
6	Больные	114
7	Больные	119
8	Здоровые	116
9	Здоровые	117
10	Здоровые	121
11	Здоровые	114
12	Здоровые	116
13	Здоровые	118
14	Здоровые	123
15	Здоровые	120

Рис. 2.14. Пример оформления данных для выполнения *t*-теста для независимых выборок

Обратите внимание на то, как оформлены данные на рис. 2.1. В табл. имеются две переменные. Одна из них – группирующая (**Grouping variable**) («Группа») – содержит коды, указывающие принадлежность данных о содержании гемоглобина к конкретной группе. Другая – зависимая переменная (**Dependent variable**) («Гемоглобин») – содержит собственно данные. Возможен и другой вариант оформления – данные для каждой группы («Больные» и «Здоровые») можно было бы просто внести в отдельные столбцы, не используя группирующую переменную.

Для выполнения t-теста для независимых выборок необходимо выполнить следующие действия:

- Запустить соответствующий модуль (рис. 2.15) из меню **Statistics > Basic statistics/Tables > t-test, independent, by groups** (если в таблице с данными есть группирующая переменная) или t-test, independent, by variables (если данные внесены в самостоятельные столбцы). (*Примечание.* мы рассмотрим вариант теста, при котором группирующая переменная присутствует; рис. 2.14).
- В открывшемся окне нажать кнопку **Variables** и указать программе, какая из переменных является группирующей, а какая – зависимой (рис. 2.16).
- Нажать на кнопку **Summary: T-tests** (рис. 2.15).

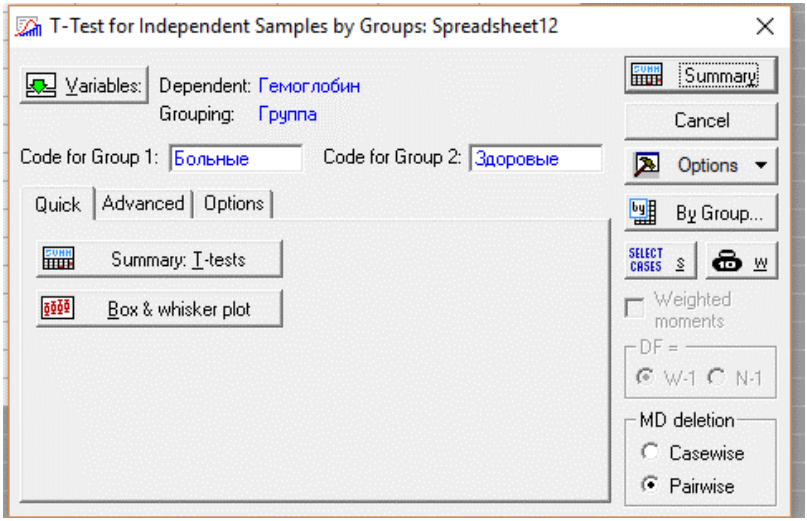


Рис. 2.15. Вид окна модуля t-теста для независимых выборок

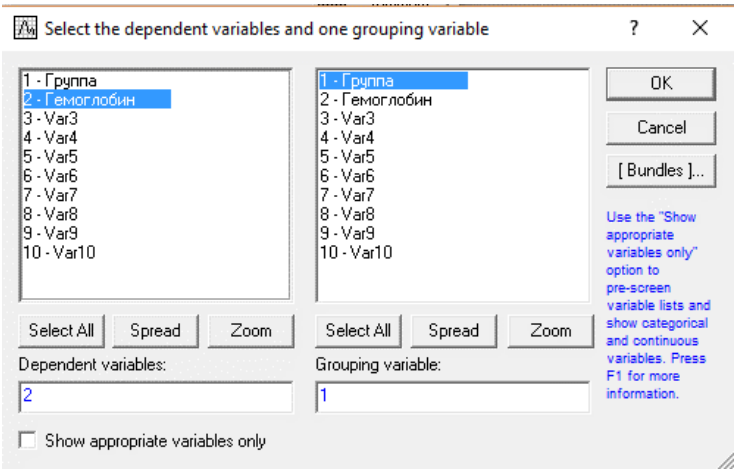


Рис. 2.16. Вид окна выбора переменных для включения в t-тест

В итоге программа создаст рабочую книгу, содержащую таблицу с результатами t-теста. Эта таблица имеет несколько столбцов (рис. 2.17):

- Mean (Больные): среднее значение содержания гемоглобина в группе «Больные»;
- Mean (Здоровые): среднее значение содержания гемоглобина в группе «Здоровые»;
- *t*-value: значение рассчитанного программой *t*-критерия Стьюдента;

• - df: число степеней свободы;

• P: вероятность ошибочно отвергнуть нулевую гипотезу об отсутствии различий между средними (см. выше). Фактически это самый главный интересующий нас результат анализа. В рассматриваемом примере P > 0,05, на основании чего можно сделать вывод об отсутствии статистически значимых различий между средними значениями содержания гемоглобина из разных групп наблюдения.

• Valid N (Северная): объем выборки «Больные»;

• Valid N (Южная): объем выборки «Здоровые»;

• Std. dev. (Больные): стандартное отклонение выборки «Больные»;

• Std. dev. (Здоровые): стандартное отклонение выборки «Здоровые»;

• F-ratio, Variances: значение F-критерия Фишера, с помощью которого проверяется гипотеза о равенстве дисперсий в сравниваемых выборках (см. выше условия применения теста Стьюдента);

• P, Variances: вероятность ошибки для F-теста Фишера. Поскольку в нашем случае P > 0,05, можно заключить, что дисперсии сравниваемых выборок не различаются (т. е. условие однородности дисперсий выполняется).

T-tests; Grouping: Группа (Spreadsheet12)
Group 1: Больные
Group 2: Здоровые

Variable	Mean Больные	Mean Здоровые	t-value	df	p	Valid N Больные	Valid N Здоровые	Std.Dev. Больные	Std.Dev. Здоровые	F-ratio Variances	p Variances
Гемоглобин	117,8571	118,1250	-0,146732	13	0,885595	7	8	4,059087	2,997022	1,834327	0,445585

Рис.2.17. Результаты выполнения t-теста для независимых выборок

Случай зависимых выборок

С зависимыми выборками исследователь имеет дело каждый раз, когда измерения значений изучаемого признака выполняются на одних и тех же объектах. Рассмотрим следующий пример. Для выяснения эффективности нового лекарства у пациентов измеряли САД до применения препарата и после. Необходимо выяснить, различается ли уровень АД в зависимости от применения лекарства (рис. 2.18).

5 До	6 После
225	192
241	202
226	206
220	196
236	196
232	214
224	198
230	194
290	179

Рис. 2.18. Пример оформления данных для выполнения *t*-теста для зависимых выборок.

Поскольку давление измерялось у одних и тех же людей, то выборки, полученные в результате двух описанных экспериментов, являются зависимыми. Чтобы сравнить средний уровень САД воспользуемся *t*-тестом для зависимых выборок. (*Примечание.* В учебных целях допустим, что условие о нормальности распределения данных выполняется). Для выполнения этого варианта *t*-теста необходимо:

• Запустить соответствующий модуль из меню **Statistics > Basic statistics/Tables > t-test, dependent samples**.

• В открывшемся окне нажать на кнопку **Variables** и указать программе первую (**First variable**) и вторую (**Second variable**) переменные, участвующие в анализе.

Нажать на кнопку **Summary: T-tests**.

В результате появится таблица с результатами, очень похожая на ту, что мы уже видели при выполнении t-теста для независимых выборок. Она содержит следующие столбцы (рис. 2.19):

• Mean – средние значения САД для каждой из сравниваемых групп;

• Std. dv. – стандартные отклонения для каждой группы;

• N – число наблюдений;

• Diff. – средняя разница;

• Std. dv. diff. – стандартное отклоение для средней разницы;

• t – значение t-критерия;

• df – число степеней свободы;

• P – вероятность ошибочно отвергнуть нулевую гипотезу о том, что средние величины САД в сравниваемых группах не различаются. Поскольку в нашем случае P < 0,05, можно смело заключить, что средние значения САД при использовании нового лекарства значительно различаются (*Примечание.* При наличии различий, результаты анализа в Statistica обычно (но не во всех модулях!) выделяются красным цветом).

T-test for Dependent Samples (Spreadsheet12)
Marked differences are significant at p < ,05000

Variable	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
До	236,0000	21,25441						
После	197,4444	9,70967	9	38,55556	28,33774	4,081717	8	0,003526

Рис. 2.19. Результаты выполнения t-теста для зависимых выборок

<u>Сравнение выборочной средней с константой</u>

В ряде случаев возникает необходимость сравнить выборочную среднюю не с другой выборочной средней, а с определенной константой. Допустим, что, согласно государственному стандарту, ПДК неко-

торого загрязнителя составляет 100 единиц. При замерах содержания этого вещества в 10 пробах городской почвы получены следующие значения: 108, 99, 112, 100, 101, 98, 95, 105, 90, 102. Необходимо установить, превышает ли среднее содержание загрязнителя в исследованных образцах почвы предельно допустимую концентрацию?

Для ответа на поставленный вопрос необходимо воспользоваться анализом, который в программе Statistica называется t-test for single means (*t*-тест для средних, рассчитанных по одной выборке). Его можно найти в меню **Statistics > Basic Statistics/Tables > t-test, single sample**. В результате появится окно, внешний вид которого представлен на рис. 2.20.

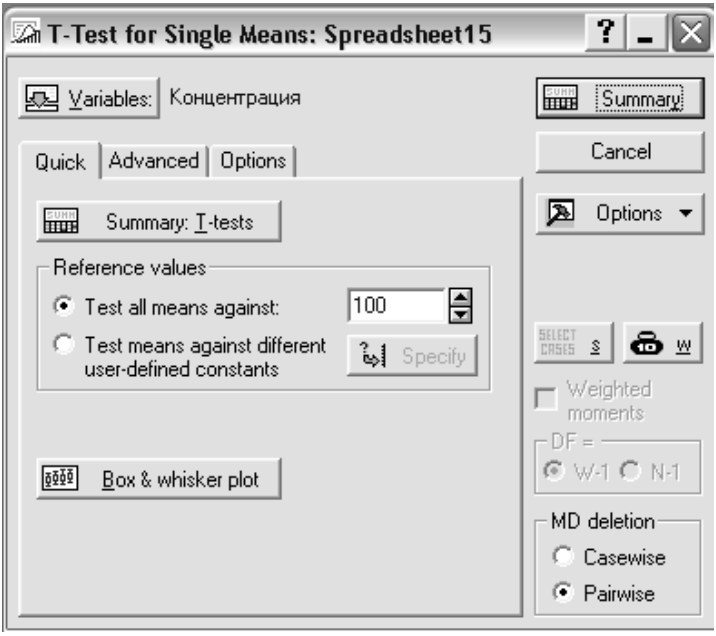


Рис. 2.20. Окно настройки t-теста для средних, рассчитанных по одной выборке

Нажав на кнопку **Variables**, необходимо выбрать столбец, который содержит анализируемые данные. Таких переменных можно ввести несколько (например, если бы аналогичные отборы проб почвы производились в разные месяцы, то в анализ можно было бы включить и эти данные). В таком случае программа сравнит все выборки с одним «контрольным» значением. Последнее задается в поле **Reference values** (Контрольные значения) (рис. 2.20). В нашем примере контролем является ПДК = 100 – его и следует внести напротив опции **Test all means against...** (Сравнить все средние с...). При необходимости, включенные в

анализ переменные можно сравнить с несколькими контрольными значениями (это достигается путем активации опции **Test means against user-defined constants** (Сравнить средние с константами, заданными пользователем)).

После нажатия на кнопку **Summary** появится рабочая книга, содержащая таблицу с результатами анализа (рис. 2.21). В этой таблице имеются следующие столбцы:

- Mean – среднее значение, рассчитанное на основе выборочных данных (в нашем случае – средняя концентрация загрязнителя в 10 пробах почвы);
- Std dv. – стандартное отклонение;
- N – объем выборки;
- Std. err. – стандартная ошибка;
- Reference constant – контрольное значение;
- df – число степеней свободы;
- P – вероятность ошибочно отвергнуть нулевую гипотезу о том, что выборочная средняя не отличается от контрольной величины. В нашем случае P > 0,05. Таким образом, несмотря на некоторое превышение средней концентрации загрязнителя в некоторых пробах, в целом это превышение это является незначительным.

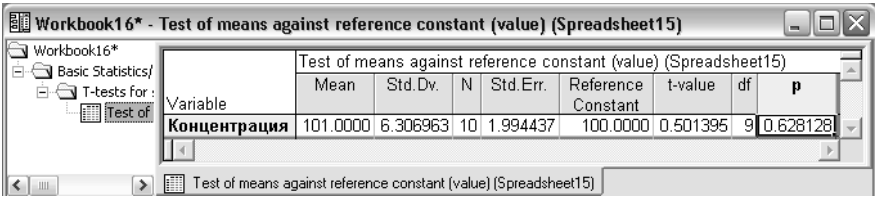


Рис. 2.21. Результаты сравнения выборочной средней с контрольным значением

2.3. Непараметрические аналоги статистических критериев

Использование рангового критерия Уилкоксона-Манна-Уитни. Критерий хи-квадрат. Точный тест Фишера; критерии Мак-Немара и Кохрана-Мантеля-Хензеля. Оценка статистической мощности при сравнении долей.

Если значения признака в двух сравниваемых группах распределены ненормально, применение параметрического *t*-теста для их сравнения будет часто приводить к искаженным результатам. В таких случаях следует воспользоваться соответствующим непараметрическим аналогом теста Стьюдента. Для сравнения двух независимых ненормально распре-

деленных выборок используется U-тест Манна–Уитни (Mann–Whitney U-test). В программе Statistica этот тест выполняется следующим образом:

• В меню **Statistics** выбрать **Nonparametrics**, а затем **Comparing two independent samples** (Сравнение двух независимых выборок).

В появившемся окне (рис. 2.22) нажать на кнопку **Variables** и выбрать зависимую и независимую переменные (для примера воспользуемся теми же данными по содержанию гемоглобина в крови людей из разных групп (рис. 2.14).

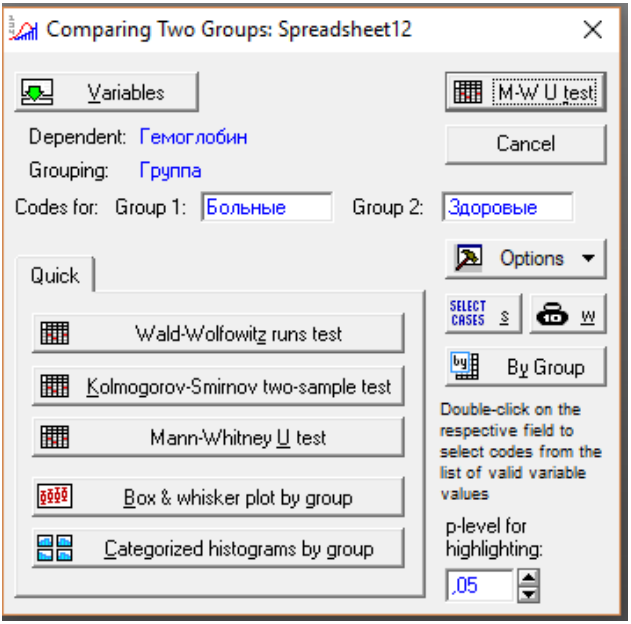


Рис. 2.22. Окно Comparing two groups модуля Nonparametrics (при сравнении независимых выборок)

Нажать на кнопку **Mann–Whitney U-test** или **M-W U test**. Внешний вид появляющегося после этого окна представлен на рисунке 2.23. Самое главное, на что следует обратить внимание в итоговой таблице теста – это величина вероятности ошибки P. При большом числе наблюдений в выборках (20 и более) значение P необходимо искать в 5-м столбце таблицы (вслед за «Z»), иначе – в 7-м (вслед за «Z-adjusted»). При $P < 0,05$ делается вывод о наличии статистически значимой разницы между сравниваемыми выборками (*Примечание.* В отличие от t-теста, тест Манна-Уитни сравнивает не средние значения выборок, а суммы рангов по каждой из них. Ранг – положение определенного значения изучаемого признака в упорядоченном по убыванию или возрастанию ряду).

variable	Mann-Whitney U Test (Spreadsheet12) By variable Группа Marked tests are significant at p <,05000 Rank Sum Больные	Rank Sum Здоровые	U	Z	p-level	Z adjusted	p-level	Valid N Больные	Valid N Здоровые	2*1sided exact p
Гемоглобин	53,50000	66,50000	25,50000	-0,289319	0,772338	-0,291144	0,770941	7	8	0,778866

Рис. 2.23. Результаты выполнения теста Манна–Уитни

Случай зависимых выборок

Если две зависимые выборки распределены ненормально, то для их сравнения следует применить тест Уилкоксона (Wilcoxon matched pair test), который можно найти там же, где и тест Манна–Уитни (**Statistics > Nonparametrics > Comparing dependent samples**).

Далее необходимо:

В появившемся окне (рис. 2.24) нажать кнопку **Variables** и задать переменные для анализа (для примера используем данные, представленные на рис. 2.18.

- Нажать кнопку **Wilcoxon matched pair test**.

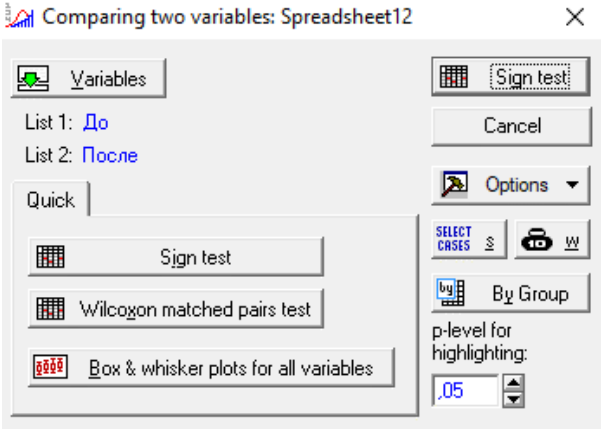


Рис. 2.24. Окно Comparing two groups модуля Nonparametrics (при сравнении зависимых выборок)

- В итоговой таблице (рис. 2.25) найти величину P. При P < 0,05 можно сделать вывод о наличии статистически значимой разницы между сравниваемыми выборками.

Pair of Variables	Wilcoxon Matched Pairs Test (Spreadsheet12) Marked tests are significant at p <,05000 Valid N	T	Z	p-level
До & После	9	0,00	2,665570	0,007686

Рис. 2.25. Результаты выполнения теста Уилкоксона

Корреляционный анализ

Корреляционный анализ позволяет сделать заключение не только о том, какова связь между двумя признаками по направлению (прямая или обратная), но и, что очень важно, выразить ее количественно при помощи коэффициента корреляции – величины, изменяющейся от –1 до +1. Чем ближе коэффициент к 1 (по модулю), тем сильнее связь между признаками. Знак коэффициент указывает на направлении зависимости.

Коэффициент корреляции Пирсона

Использование коэффициента корреляции Пирсона для оценки степени связи между двумя признаками предполагает выполнение следующих двух обязательных условий:

- значения обоих анализируемых признаков распределены нормально;
- связь между признаками является линейной.

Способы проверки данных на нормальность распределения мы рассмотрели ранее. О том, как установить линейность зависимости между признаками, изложено ниже. Предположим, необходимо выяснить наличие связи между концентрацией гемоглобина и количеством эритроцитов в образцах. Для этого были выполнены соответствующие измерения в 12 образцах. Полученные данные приведены на рис. 2.26.

	1 Гемоглобин, г/л	2 Эритроциты, млн.клеток
1	10,4	7,4
2	10,8	7,6
3	11,1	7,9
4	10,2	7,2
5	10,3	7,4
6	10,2	7,1
7	10,7	7,4
8	10,5	7,2
9	10,8	7,8
10	11,2	7,7
11	10,6	7,8
12	11,4	8,3

Рис. 2.26. Пример оформления данных
для расчета коэффициента корреляции Пирсона

Для расчета коэффициента корреляции Пирсона необходимо выполнить следующее:

- Запустить модуль анализа из меню **Statistics > Basic Statistics/Tables > Correlation Matrices** (Корреляционные матрицы).

• В появившемся окне выбрать переменные, которые должны участвовать в анализе. Для этого нужно нажать либо кнопку **One variable list** (Один список переменных) либо **Two lists (rect. matrix)** (Два списка (прямоугольная матрица)). В первом случае анализируемые переменные последовательно выбираются из одного списка, а во втором – из двух (рис. 2.27).

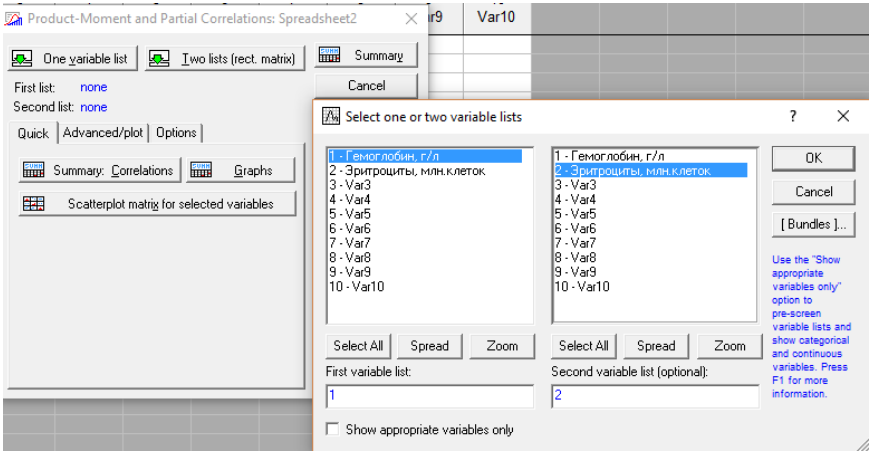


Рис. 2.27. Вид окна выбор переменных
для расчета коэффициента корреляции Пирсона

• Проверить условия применимости коэффициента Пирсона (см. выше). Для визуальной оценки выполнения этих условий можно нажать кнопку **Scatterplot matrix for selected variables** (Диаграмма рассеяния для выбранных переменных). В результате программа построит точечный график, по осям которого будут отложены значения соответствующих переменных (рис.2.28). Диагональная линия на этом графике служит для оценки линейности связи между анализируемыми признаками. Если точки-наблюдения укладываются вдоль этой линии на близком расстоянии, можно говорить о существовании прямолинейной зависимости. Вместе с диаграммой рассеяния программа строит также распределения значений анализируемых признаков в виде гистограмм, по форме которых можно проверить условие о нормальности распределения. *(Примечание.* В рассматриваемом примере условие нормальности распределения явно не выполняется, однако в учебных целях мы продолжим расчет коэффициента корреляции Пирсона).

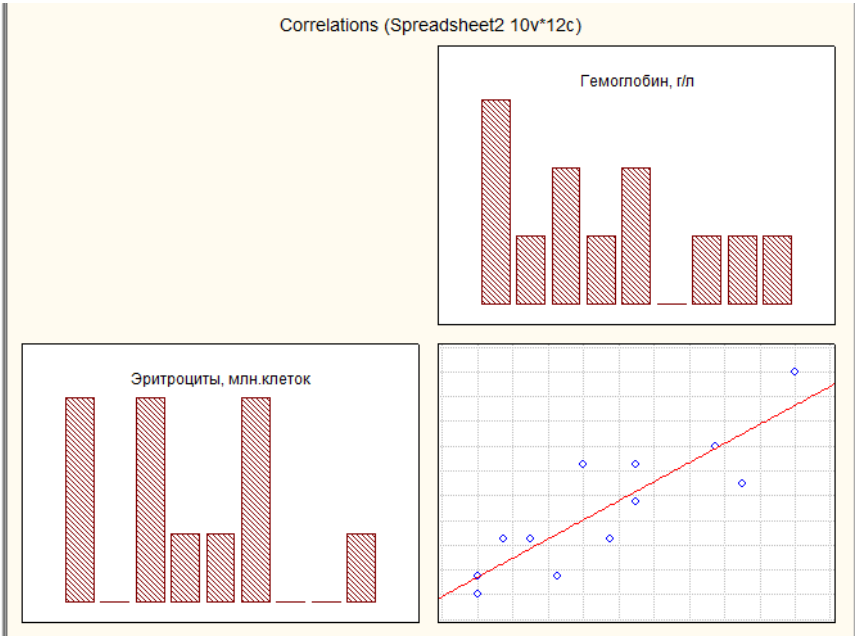


Рис. 2.28. Визуальная оценка условий применимости коэффициента корреляции Пирсона

• Нажать на кнопку **Summary: Correlation matrix** (Результат: Корреляционная матрица). В результате появится таблица, содержащая рассчитанный программой коэффициент корреляции (рис. 2.29).

Variable	Correlations (Spreadsheet2) Marked correlations are significant at p < ,05000 N=12 (Casewise deletion of missing data) **Эритроциты, млн.клеток**
Гемоглобин, г/л	0,87

Рис. 2.29. Результат расчета коэффициента Пирсона

В нашем случае коэффициент оказался очень высоким (r = 0,87), что указывает на существование тесной связи между концентрацией гемоглобина и количеством эритроцитов. Одновременно с расчетом коэффициента программа оценивает и его статистическую значимость, то есть проверяет нулевую гипотезу о том, что в действительности связь между признаками отсутствует. Статистически значимые коэффициенты корреляции Пирсона в программе Statistica выделяются красным цветом (P < 0,05).

Сравнение двух коэффициентов корреляции Пирсона

В ряде случаев возникает необходимость сравнить два коэффициента корреляции, рассчитанных на основе разных выборок. Рассмотрим следующий пример. В ходе обследования 98 больных людей выяснилось, что коэффициент корреляции между концентрацией гемоглобина и количеством эритроцитов у них составляет 0,78. У здоровых людей (обследовано 95 лиц) эта же зависимость выражалась несколько более высоким коэффициентом – 0,84. Случайна ли разница между этими двумя коэффициентами?

Нулевая гипотеза, которую нам предстоит проверить, заключается в том, что разница действительно случайна. Для ее проверки воспользуемся модулем программы, который называется **Difference tests** (Тесты на различие). Он находится в меню **Statistics > Basic Statistics/Tables > Difference tests: r, %, means**. Внешний вид модуля приведен на рис. 2.29.

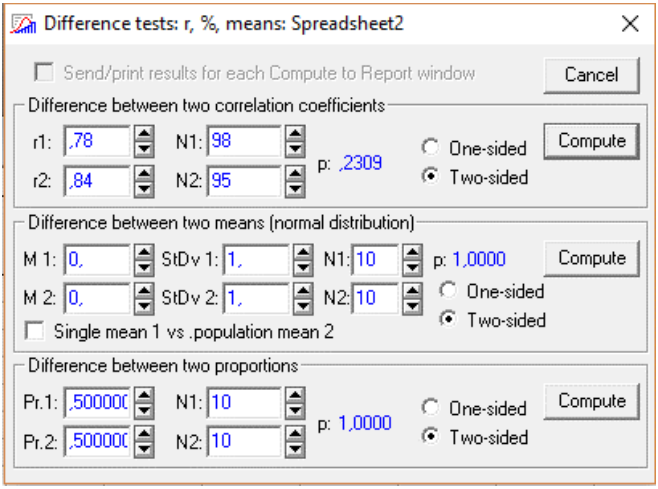


Рис. 2.29. Окно модуля Difference tests

Рассмотрим рис. 2.29. В верхней части изображенного на нем диалогового окна имеется поле Difference between two correlation coefficients (Различие между двумя коэффициентами корреляции) с четырьмя ячейками, в которые необходимо ввести соответствующие данные для выполнения анализа. В ячейки **r1** и **r2** вводятся значения сравниваемых коэффициентов корреляции, в **N1** и **N2** – объемы выборок, на основе которых эти коэффициенты были рассчитаны. Рядом с ячейками нужно указать, какой вариант теста мы собираемся выполнять – одно- (One-sided) или двухсторонний (Two-sided). Поскольку мы просто хотим выяснить наличие разницы между коэффициентами, оставим Two-sided. Если бы стояла необходимость проверить гипотезу о том, что один из

коэффициентов значительно больше другого, то нужно было бы выбрать опцию One-sided. Наконец, следует нажать кнопку Compute (Рассчитать). В том же поле программа покажет, чему равна вероятность справедливости нулевой гипотезы. В нашем примере P = 0,2309, что больше обычно принимаемого уровня «0,05». Следовательно, наблюдаемая разница между коэффициентами корреляции является случайной.

Коэффициент корреляции Спирмена

При расчете коэффициента корреляции Пирсона между концентрацией гемоглобина и количеством эритроцитов (рис. 2.14) мы столкнулись с тем, что значения этих признаков не были распределены нормально (рис. 2.16). В подобных ситуациях применение коэффициента Пирсона может приводить к выводам, не соответствующим действительности. Вместо него следует воспользоваться одним из непараметрических коэффициентов корреляции. Из последних наиболее обычен ранговый коэффициент корреляции Спирмена. Рассчитаем его для тех же данных о концентрации гемоглобина и количестве эритроцитов (рис. 2.14). Для этого необходимо выполнить следующее:

- Запустить модуль **Nonparametric correlations** (Непараметрические корреляции) из меню **Statistics > Nonparametrics > Correlations (Spearman, Kendall tau, gamma)** (Корреляции (Спирмена, тау Кендалла, гамма)).
- В появившемся окне (рис. 2.30) нажать на кнопку **Variables** и выбрать столбцы, содержащие необходимые данные.

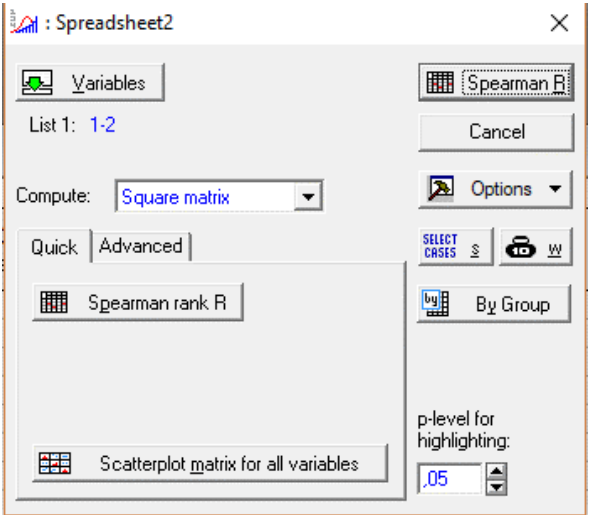


Рис. 2.30. Модуль непараметрического корреляционного анализа

• Нажать кнопку **Spearman R** или **Spearman rank R**. Появится таблица с результатами анализа (рис. 2.31), которая содержит столбцы Valid N (число наблюдений), Spearman R (коэффициент корреляции Спирмена), t(N-2) (значение критерия Стьюдента для числа степеней свободы n-2), и P (вероятность ошибки для нулевой гипотезы об отсутствии связи между признаками).

Spearman Rank Order Correlations (Spreadsheet2)
MD pairwise deleted
Marked correlations are significant at p <,05000

Variable	**Гемоглобин, г/л**	Эритроциты, млн.клеток
Гемоглобин, г/л	1,000000	0,851085
Эритроциты, млн.клеток	0,851085	1,000000

Рис. 2.31. Результат расчета коэффициента корреляции Спирмена

В нашем примере коэффициент корреляции Спирмена оказался несколько ниже рассчитанного ранее коэффициента Пирсона (0,85 против 0,87). При этом он является в высокой степени статистически значимым ($p << 0,05$).

Коэффициент ассоциации (связанности)

Коэффициенты корреляции Пирсона и Спирмена применяются для оценки связи между количественными признаками. Однако многие биологические исследования сопряжены с изучением качественных величин (окраска, пол, наличие или отсутствие определенного состояния и т. п.). Для таких признаков также можно определить степень «связанности». Один из показателей, который позволяет это сделать – коэффициент ассоциации φ, который изменяется от 0 до 1. Чем ближе φ к 1, тем сильнее связь.

Рассмотрим пример: совокупность из 111 мышей разделили на две группы: по 57 и 54 мыши. Первой группе мышей сделали инъекцию патогенных бактерий при последующем введении сыворотки с антителами. Животные из второй группы служили контролем – им сделали только инъекции бактерий. После некоторого времени инкубации оказалось, что всего погибло 38 мышей, а выжило 73. Из погибших 13 принадлежали первой группе, а 25 – ко второй (контрольной). *Вопрос:* «Есть ли связь между введением сыворотки и выживаемостью мышей?». Данные описанного эксперимента удобно представить в виде таблицы сопряженности размером 2×2 (= четырехпольная таблица, или таблица с двумя входами) (табл. 2.4).

Таблица 2.4

Результат эксперимента

	Погибло	Выжило
Бактерии + сыворотка	13	44
Только бактерии	25	29

Для расчета коэффициента ассоциации необходимо:

• Запустить модуль анализа четырехпольных таблиц из меню **Statistics > Nonparametrics > 2×2 Tables**.

• В ячейки появившейся панели (рис. 2.31) ввести данные о численности мышей в каждой из экспериментальных групп (в соответствии с приведенной выше табл. 2.4).

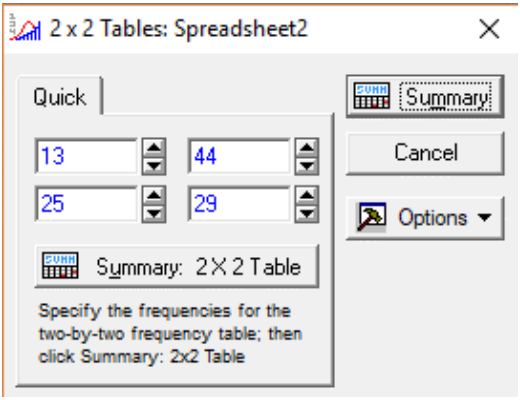


Рис. 2.31. Модуль анализа таблиц сопряженности 2×2

• Нажать кнопку **Summary**. В результате этого появится таблица с большим набором статистических показателей (рис. 2.32).

	2 x 2 Table (Spreadsheet2)		
	Column 1	Column 2	Row Totals
Frequencies, row 1	13	44	57
Percent of total	11,712%	39,640%	51,351%
Frequencies, row 2	25	29	54
Percent of total	22,523%	26,126%	48,649%
Column totals	38	73	111
Percent of total	34,234%	65,766%	
Chi-square (df=1)	6,80	p= ,0091	
V-square (df=1)	6,73	p= ,0095	
Yates corrected Chi-square	5,79	p= ,0161	
Phi-square	,06122		
Fisher exact p, one-tailed		p= ,0078	
two-tailed		p= ,0102	
McNemar Chi-square (A/D)	5,36	p= ,0206	
Chi-square (B/C)	4,70	p= ,0302	

Рис. 2.32. Результат анализа таблицы сопряженности 2×2

Нам необходима строка Phi-square, в которой находится значение коэффициента ассоциации, возведенное в квадрат для придания ему положительного значения. В нашем случае фи-квадрат равен 0,061. После извлечения корня получаем $\varphi = 0{,}247$. Таким образом, полученное значение фи указывает на достаточно слабую связь между введением сыворотки и выживаемостью зараженных мышей. Заметьте, что модуль 2×2 Tables можно использовать также для сравнения частот бинарных качественных признаков в двух группах. Так, в таблице (рис. 6.9) помимо коэффициента ассоциации приведены значения: критерия χ^2 без поправки Йетса, а также точного критерия Фишера; критерия Мак-Нимара для зависимых групп. Анализируя, например, результат теста χ^2, можно заключить, что несмотря на установленный нами слабый эффект от введения сыворотки с антителами, выживаемость мышей в контрольной и экспериментальной группах все же статистически значимо различается ($P = 0{,}0091$; см. строку Chi-square (df = 1) на рис. 2.32).

Глава 3. Линейные модели в дисперсионном анализе

3.1. Линейные модели в дисперсионном анализе

Условия применения дисперсионного анализа. Внутригрупповая и межгрупповая дисперсии. Процедура вычисления критерия F. Критерий Стьюдента как вариант дисперсионного анализа для сравнения двух выборок. Дисперсионный анализ повторных измерений в медицинских экспериментах.

Дисперсионный анализ (ANalysis Of VAriance, ANOVA) используется для выявления влияния на изучаемый показатель некоторых факторов, обычно не поддающихся количественному измерению, а принадлежащих к номинальной или порядковой шкалам. Указанные факторы часто называют (как и в регрессионном анализе) независимыми переменными, а исследуемый с точки зрения влияния на него выбранных факторов показатель – зависимой переменной (объясняемой переменной, результирующим признаком).

Суть дисперсионного анализа заключается в разложении вариации зависимой переменной на части, соответствующие раздельному и совместному влиянию на нее независимых переменных с тем, чтобы посредством статистических методов установить приемлемость ряда гипотез о значимости такого влияния.

В зависимости от числа факторов модели дисперсионного анализа подразделяются на однофакторные и многофакторные (двухфакторные и т. д.). Конкретные значения каждого из факторов принято называть их уровнями. Множество уровней – свое для каждого из факторов, оно фиксировано (то есть не зависит от номера эксперимента, доставляющего выборочные значения интересующих показателей) и конечно. Вообще говоря, допускаются и факторы с количественными характеристиками, но область изменения каждого из них – это тоже индивидуальное, фиксированное и конечное, но уже числовое множество. Описанная модель «устройства» факторов обычно именуется ***детерминированной*** (ниже обозначается M_1, fixed-effects model). В более сложной ***случайной или стохастической*** модели (M_2, random effects model) уровни каждого фактора в данном наблюдении получают случайную выборку из генеральной совокупности всех его уровней. В ***смешанной*** модели (M_3, mixed-effects model) уровни одних факторов заранее фиксированы, тогда как уровни других – случайные выборки.

Параметрический однофакторный дисперсионный анализ

Для проверки эффективности двух новых препаратов был выполнен следующий эксперимент. Было в каждую из 27 лунок планшета было добавлено 10000 клеток. Затем в 9 из них добавили используемый в практике препарат, в следующие 9 – препарат с модификацией 1, а в оставшиеся 9 – препарат с модификацией 2. Через 3 дня было посчитано количество клеток в каждой лунке. Полученные данные приведены на рис. 3.1. *Вопрос:* «Различается ли средний прирост клеточной культуры в зависимости от типа препарата?».

	1 Препарат	2 Количество
1	Модификация 0	19200
2	Модификация 0	20200
3	Модификация 0	20600
4	Модификация 0	19600
5	Модификация 0	19600
6	Модификация 0	21400
7	Модификация 0	19800
8	Модификация 0	19400
9	Модификация 0	17900
10	Модификация 1	22500
11	Модификация 1	24100
12	Модификация 1	22600
13	Модификация 1	22000
14	Модификация 1	23600
15	Модификация 1	23200
16	Модификация 1	22400
17	Модификация 1	23000
18	Модификация 1	20900
19	Модификация 2	22520
20	Модификация 2	24000
21	Модификация 2	22610
22	Модификация 2	22050
23	Модификация 2	23680
24	Модификация 2	23210
25	Модификация 2	22490
26	Модификация 2	22960
27	Модификация 2	21000

Рис. 3.1. Пример оформления данных для выполнения однофакторного дисперсионного анализа

Воспользуемся однофакторным дисперсионным анализом (поскольку проверяется влияние лишь одного фактора – типа препарата). Для его выполнения необходимо:

• Запустить модуль One-way ANOVA (рис. 3.2) из меню **Statistics > ANOVA**.

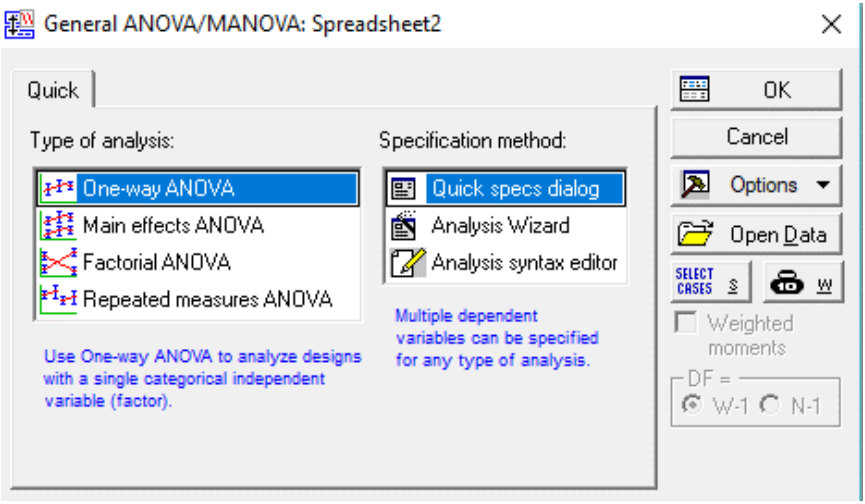


Рис. 3.2. Окно модуля однофакторного дисперсионного анализа

• Нажать на кнопку **Variables** и выбрать зависимую («Количество») и группирующую переменные («Препарат»). Нажать на кнопки: **Factor codes > All** (так мы укажем программе, что в анализе должны участвовать все задействованные в эксперименте группы) > OK. В итоге появится окно с 8 закладками (рис. 3.3). Автоматически программа откроет его на закладке **Quick** (Быстро).

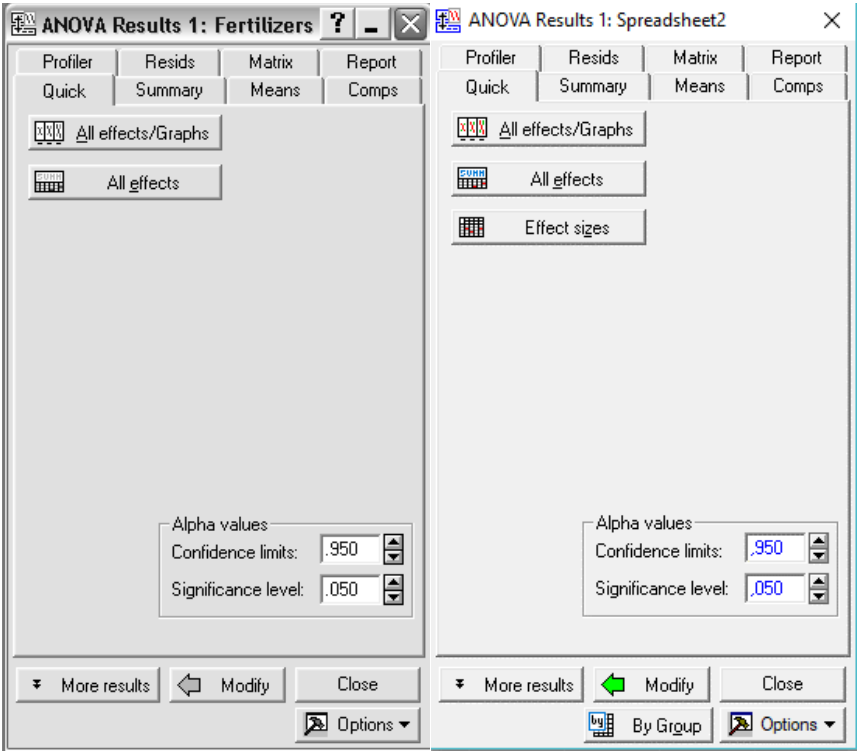


Рис. 3.3. Окно выбора результатов дисперсионного анализа

Результаты анализа можно получить уже на данном этапе, если нажать на кнопку **All effects** (Все эффекты). Однако рассматриваемый вариант дисперсионного анализа является параметрическим, то есть предполагает выполнение следующих обязательных условий в отношении данных: 1) в каждой из сравниваемых групп значения анализируемого признака распределяются нормально; 2) групповые дисперсии однородны (между ними нет статистически значимой разницы). Кроме того, все сравниваемые выборки должны быть независимыми. Поэтому перед получением результатов анализа следует проверить, выполняются ли указанные условия и поступаем ли мы корректно, используя данный вариант дисперсионного анализа.

Для проверки условий ANOVA необходимо выполнить следующее:

• Нажать на кнопку **More results** (Дополнительные результаты), расположенную в нижней части окна **ANOVA Results** > Открыть закладку **Assumptions** (Допущения). В результате этого появится окно, представленное на рис. 3.4.

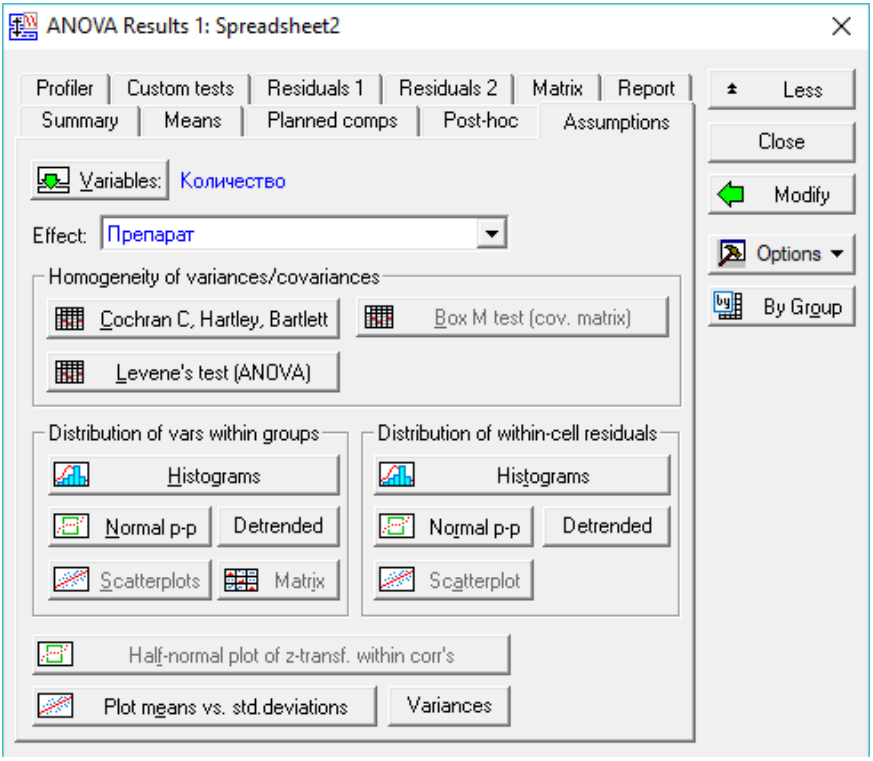


Рис. 3.4. Окно дополнительных результатов дисперсионного анализа на закладке Assumptions

• Для проверки однородности групповых дисперсий в поле **Homogeneity of variances/covariances** нажать на кнопку **Levene's test** (тест Левена). Если результат этого теста указывает на отсутствие различий между дисперсиями ($P > 0{,}05$), то применение параметрического варианта дисперсионного анализа является обоснованным. В нашем примере различий действительно нет ($P = 0{,}993$) (проверьте самостоятельно).

• Для проверки нормальности распределения анализируемых данных необходимо воспользоваться одной из опций, доступных в поле **Distribution of variables within groups** (Распределение переменных внутри групп). *Примечание.* Если число наблюдений в сравниваемых группах невелико, лучше использовать график нормальных вероятностей (кнопка **Normal p-p**). Если же их много, то можно оценить характер распределений, построив гистограммы (кнопка **Histograms**). При нажатии на одну их этих кнопок программа предложит список групп, участвующих в анализе. Пример графика, построенного на «вероятностной бумаге» для

«используемого препарата», приведен на рис. 3.5. Видно, что точки-наблюдения тесно укладываются вдоль теоретически ожидаемой прямой. Аналогичная ситуация характерна и для остальных двух групп из рассматриваемого примера. Таким образом, анализируемые данные по урожайности удовлетворяют обоим условиям ANOVA.

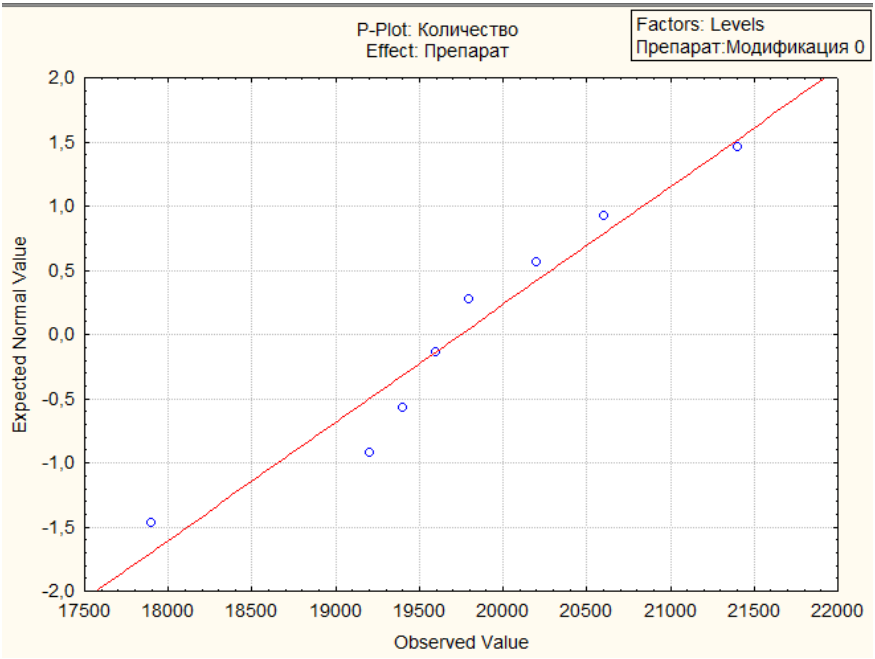
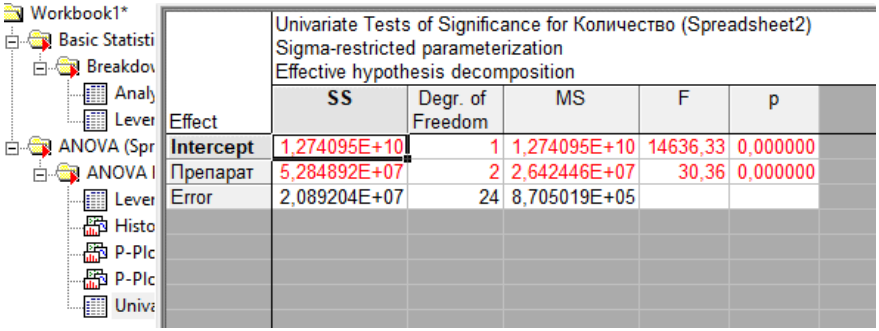


Рис. 3.5. Окно дополнительных результатов дисперсионного анализа на закладке Assumptions

• Наконец, необходимо на закладке **Summary** (рис. 3.4) нажать кнопку **Test all effects** (Проверить все эффекты). В появившейся таблице результатов (рис. 3.6) необходимо разыскать ячейку с величиной ошибки P для нулевой гипотезы об отсутствии связи между количеством клеток в лунке и типом препарата (строка «Препарат»). Поскольку в нашем примере $P < 0,05$, можно заключить, что среднее количество клеток статистически значимо различается в зависимости от использованного препарата.



Univariate Tests of Significance for Количество (Spreadsheet2)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	1,274095E+10	1	1,274095E+10	14636,33	0,000000
Препарат	5,284892E+07	2	2,642446E+07	30,36	0,000000
Error	2,089204E+07	24	8,705019E+05		

Рис. 3.6. Результаты однофакторного дисперсионного анализа

Кроме этого, результаты анализа можно получить, выбрав **All effects/Graphs**. Появится таблица со значениями анализа (рис. 3.7).

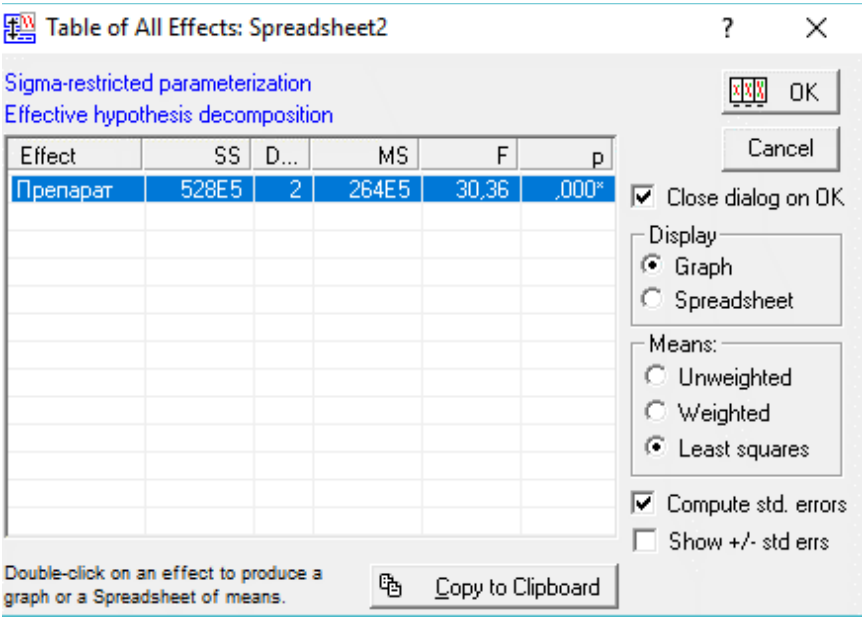


Рис. 3.7. Вид таблицы с результатами дисперсионного анализа

Из таблицы следует, что значение критерия Фишера равно 9,354, а уровень значимости равен 0,000*. Таким образом, отвергается нулевая гипотеза и принимается альтернативная, о влиянии типа препарата на рост клеток.

Если необходимо визуально сравнить различие, то в открытом окне нажмите ОК. Появится графическое представление выборочных данных с основными параметрами (рис. 3.8).

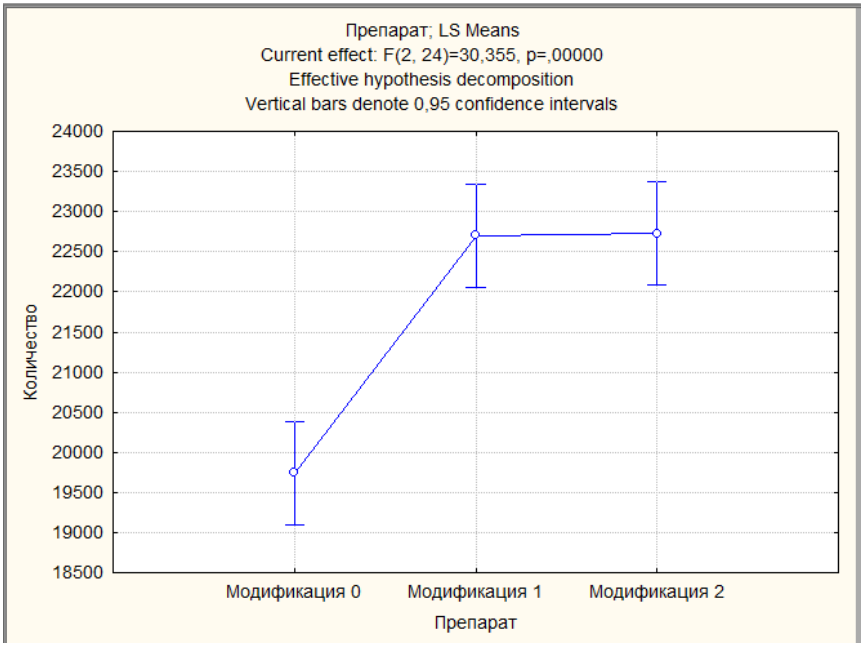


Рис. 3.8. Графическое представление выборочных данных дисперсионного анализа

Апостериорный анализ

Важно помнить, что дисперсионный анализ позволяет проверить лишь гипотезу об отсутствии различий между сравниваемыми группами в целом. Однако с его помощью невозможно узнать, какие именно группы различаются между собой. Для выяснения этого необходимо воспользоваться методами множественных сравнений, являющихся частью апостериорного анализа (Post-hoc analysis). Механизм их работы заключается в проведении попарных сравнений средних значений всех групп, включенных в дисперсионный анализ.

Для выполнения множественных сравнений необходимо открыть закладку **Post hoc** (рис. 3.9) в окне дополнительных результатов дисперсионного анализа (**More results**).

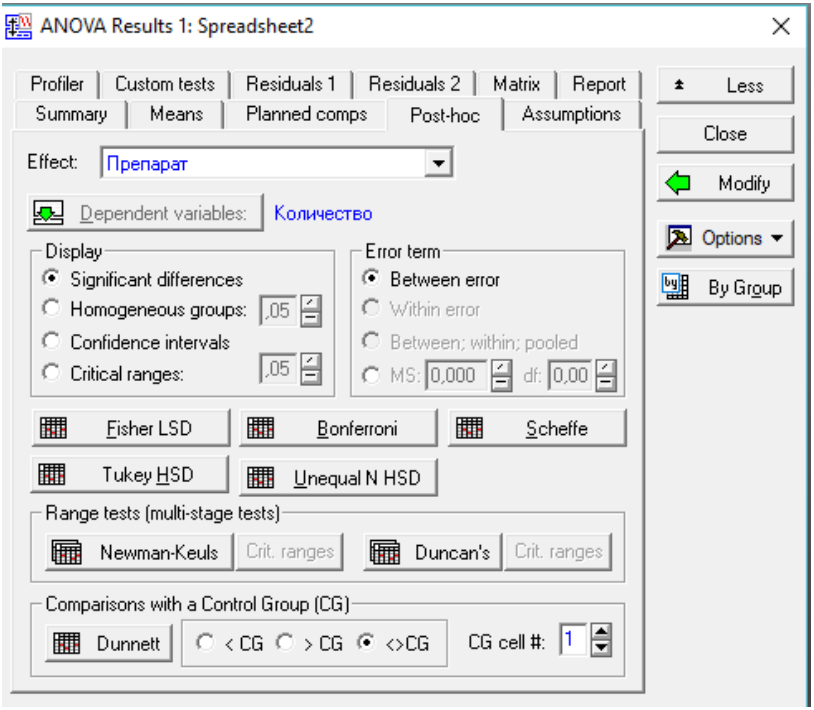


Рис. 3.9. Окно дополнительных результатов дисперсионного анализа на закладке Post-hoc

Программа Statistica предлагает ряд тестов для множественных сравнений, несколько различающихся по мощности:

Fisher LSD,
Bonferroni,
Scheffe,
Tukey HSD,
Newman-Keuls,
Duncan's,
Dunnet.

Наиболее часто используемыми являются тесты Тьюки (Tukey HSD) и Ньюмена-Кейлса (Newman-Keuls). Нажатие на кнопку соответствующего теста приводит к появлению рабочей книги с матрицей значений P. Из рис. 3.10, например, видно, что статистически значимая разница в количестве клеток существует между парами препаратов «Модификация 0 – Модификация 1» и «Модификация 0 – модификация 2» ($P < 0,05$), тогда как оба новых препарата по эффективности не различаются ($P > 0,05$) (выполнен тест Тьюки для выборок с одинаковыми объемами).

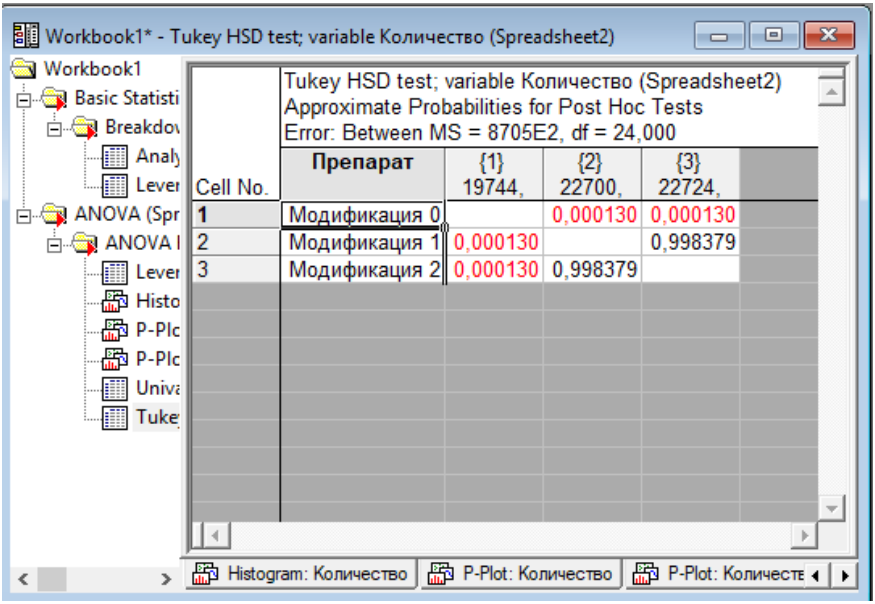


Рис. 3.10. Результат выполнения теста Тьюки

3.2. Протокол разведочного анализа данных. Структура модельных объектов дисперсионного анализа

Модели двух- и многофакторного дисперсионного анализа.

В примере с тремя типами препаратов мы не обращали внимания на то, какой применялся тип питательной смеси. Лунки выбирались случайным образом, а это значит, что просто в силу случая они могли оказаться разными по свойствам среды, что, в свою очередь, также могло сказаться на росте клеток. Теперь мы учтем это обстоятельство и выполним двухфакторный дисперсионный анализ (фактор 1 – тип препарата, фактор 2 – тип питательной смеси (1 – DMEM, 2 – RPMI 1640)). Поскольку в анализ включен дополнительный фактор, в таблицу с данными необходимо добавить еще одну группирующую переменную, которая будет содержать коды типов питательной смеси (рисунок 3.11). Для выполнения двухфакторного дисперсионного анализа в программе Statistica необходимо выполнить следующее:

• Запустить модуль **Factorial ANOVA** (Факторный дисперсионный анализ) из меню **Statistics > ANOVA**.

• В появившемся окне (рисунок не приводится) нажать кнопку **Variables** и выбрать зависимую («Количество») и группирующие пере-

менные («Препарат» и «Тип»). Нажать на кнопки: **Factor codes > All > ОК > ОК**. В итоге появится уже знакомое вам по рис. 3.3 окно с 8 закладками. (*Примечание.* При изложении дальнейшего материала предполагается, что анализируемые данные успешно прошли проверку на нормальность распределения и однородность групповых дисперсий – см. выше).

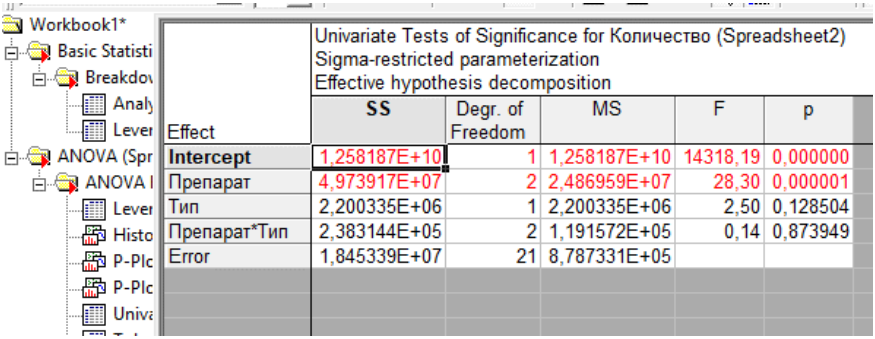
	1 Препарат	2 Тип	3 Количество
1	Модификация 0	1	19200
2	Модификация 0	2	20200
3	Модификация 0	1	20600
4	Модификация 0	2	19600
5	Модификация 0	1	19600
6	Модификация 0	1	21400
7	Модификация 0	2	19800
8	Модификация 0	2	19400
9	Модификация 0	2	17900
10	Модификация 1	1	22500
11	Модификация 1	2	24100
12	Модификация 1	2	22600
13	Модификация 1	1	22000
14	Модификация 1	1	23600
15	Модификация 1	1	23200
16	Модификация 1	2	22400
17	Модификация 1	1	23000
18	Модификация 1	2	20900
19	Модификация 2	2	22520
20	Модификация 2	1	24000
21	Модификация 2	1	22610
22	Модификация 2	1	22050
23	Модификация 2	1	23680
24	Модификация 2	2	23210
25	Модификация 2	1	22490
26	Модификация 2	2	22960
27	Модификация 2	2	21000

Рис. 3.11. Пример оформления данных для выполнения двухфакторного дисперсионного анализа
Обозначения: «Модификация 0», «Модификация 1», и «Модификация 2» – типы препаратов; «1» и «2» – питательная смесь DMEM и RPMI 1640

Нажать на кнопку **All effects**. Это приведет к появлению таблицы с результатами дисперсионного анализа (рис. 3.12). Самое главное для нас в этой таблице – это вторая и третья строки. В конце этих строк приведены вероятности ошибок для нулевых гипотез об отсутствии влияния типа препарата и типа смеси на количество клеток. Видно, что лишь в случае с типом препарата $P < 0,05$. Это говорит о значительном влиянии данного фактора на прирост клеток. Доказать подобный же эффект для типа смеси нам в данном эксперименте не удалось ($P > 0,05$). Строка

«Препарат * Тип» касается взаимного влияния исследуемых факторов на прирост. Как видим, взаимодействие между этими факторами также отсутствует (P > 0,05).

Аналогичным образом в программе Statistica можно выполнить дисперсионный анализ и с большим количеством факторов.



Univariate Tests of Significance for Количество (Spreadsheet2)
Sigma-restricted parameterization
Effective hypothesis decomposition

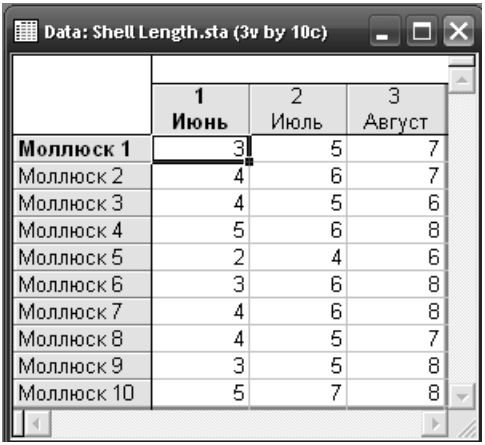
Effect	SS	Degr. of Freedom	MS	F	p
Intercept	1,258187E+10	1	1,258187E+10	14318,19	0,000000
Препарат	4,973917E+07	2	2,486959E+07	28,30	0,000001
Тип	2,200335E+06	1	2,200335E+06	2,50	0,128504
Препарат*Тип	2,383144E+05	2	1,191572E+05	0,14	0,873949
Error	1,845339E+07	21	8,787331E+05		

Рис. 3.12. Результаты двухфакторного дисперсионного анализа

Дисперсионный анализ Фридмана

Рассмотренные выше варианты дисперсионного анализа помимо обязательных условий о нормальности и однородности групповых дисперсий предполагали также, что сравниваемые группы являются независимыми. В случае с зависимыми выборками необходимо воспользоваться дисперсионным анализом Фридмана (Friedman ANOVA). Следует отметить, что, являясь непараметрическим методом, анализ Фридмана не требует нормальности распределения данных и однородности дисперсий.

На рис. 3.13 представлены данные, полученные в ходе наблюдений за изменениями длины раковины моллюска Sphaerium sp. в летние месяцы. Десять особей были посажены в специальный садок, который затем установили в озере в типичном для моллюска биотопе. Каждый месяц измеряли длину раковины у всех 10 моллюсков. Необходимо выяснить, произошли ли существенные изменения средней длины раковины к концу опыта.



	1 Июнь	2 Июль	3 Август
Моллюск 1	3	5	7
Моллюск 2	4	6	7
Моллюск 3	4	5	6
Моллюск 4	5	6	8
Моллюск 5	2	4	6
Моллюск 6	3	6	8
Моллюск 7	4	6	8
Моллюск 8	4	5	7
Моллюск 9	3	5	8
Моллюск 10	5	7	8

Рис. 3.13. Пример оформления данных для выполнения дисперсионного анализа Фридмана

Поскольку измерения раковины выполнялись на одних и тех же особях Sphaerium, три полученные выборки являются зависимыми. Сравним их с помощью дисперсионного анализа Фридмана. Для этого необходимо:

• Запустить модуль анализа (рис. 3.14) из меню **Statistics > Nonparametrics > Comparing multiple dependent samples** (Сравнение нескольких зависимых выборок).

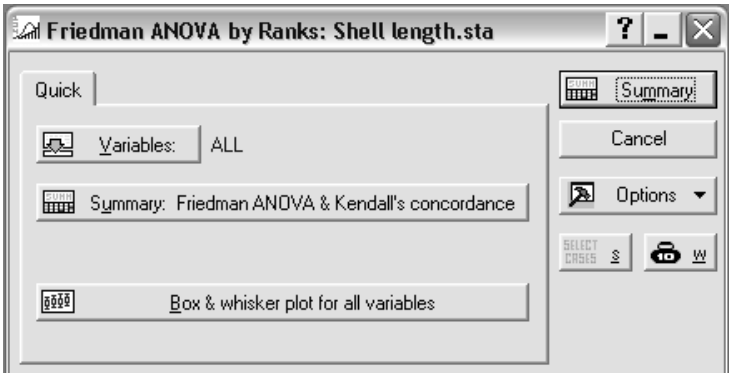


Рис. 3.14. Модуль дисперсионного анализа Фридмана

• Нажать кнопку **Variables** и выбрать переменные, которые должны участвовать в анализе.

• Нажать кнопку **Summary: Friedman ANOVA and Kendall's concordance** (Результат: ANOVA по Фридману и критерий согласованности Кендалла).

• В появившейся таблице с результатами необходимо отыскать величину ошибки P для нулевой гипотезы о том, что в течение трех месяцев наблюдений длина раковины у моллюсков существенно не изменилась. Эта величина находится в заголовке таблицы (рис. 3.15). При $P < 0,05$ (как в нашем случае) можно сделать вывод о наличии статистически значимых различий между группами.

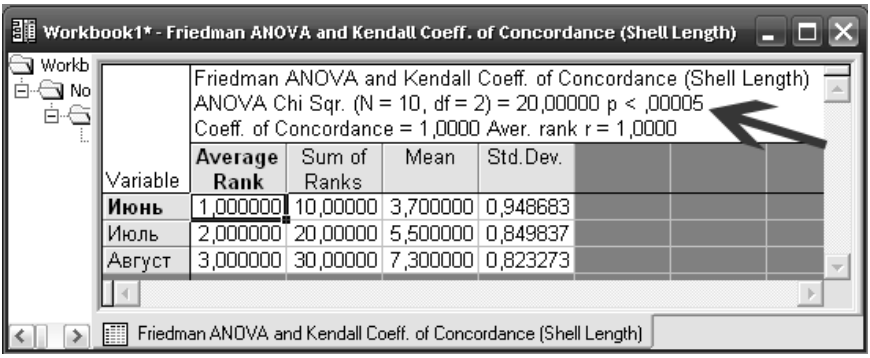


Рис. 3.15. Результаты дисперсионного анализа Фридмана (вероятность ошибки P указана стрелкой)

Дисперсионный анализ Крускала–Уоллиса

В медико-биологических исследованиях данные распределены по нормальному закону достаточно редко. И даже если выборочные единицы отбираются из нормально распределенных генеральных совокупностей, объем выборок часто оказывается слишком малым для того, чтобы вообще сделать какие-либо выводы относительно вида распределения. Это делает параметрический дисперсионный анализ неприменимым. Выходом становится использование непараметрического дисперсионного анализа Крускала–Уоллиса (или H-теста) (Kruskal–Wallis ANOVA), хотя он и обладает несколько меньшей мощностью в сравнении с параметрическим вариантом.

Для изучения иммуномодулирующего действия диметилсульфоксида (ДМСО) на функциональную активность полиморфноядерных гранулоцитов периферической крови человека выделенные нейтрофилы подвергались инкубации с ДМСО различной концентрации (3 %, 5 %, 8 %). Прединкубированные с различными концентрациями ДМСО нейтрофилы вносили в лунки полистирольного планшета (21 лунка). После проведения определенных манипуляций спектрофотометрически оценивали

оптическую плотность в каждой лунке. На рис. 3.16 представлены данные п о оптической плотности при разных концентрациях ДМСО. Как видно, число наблюдений для каждой из концентраций невелико (n = 7). При этом даже если объединить все имеющиеся данные в одну совокупность, окажется, что их распределение далеко от нормального (проверьте самостоятельно).

	1 Концентрация	2 Оптическая плотность
1	3% ДМСО	0
2	3% ДМСО	0
3	3% ДМСО	0,2
4	3% ДМСО	0,72
5	3% ДМСО	0
6	3% ДМСО	0,16
7	3% ДМСО	1,2
8	5% ДМСО	1,12
9	5% ДМСО	0
10	5% ДМСО	0
11	5% ДМСО	1,4
12	5% ДМСО	1,68
13	5% ДМСО	0,56
14	5% ДМСО	0,9
15	8% ДМСО	3,46
16	8% ДМСО	1,28
17	8% ДМСО	1,16
18	8% ДМСО	0
19	8% ДМСО	0
20	8% ДМСО	3
21	8% ДМСО	2,86

Рис. 3.16. Пример оформления данных для выполнения дисперсионного анализа Крускала–Уоллиса

Чтобы выяснить, различается ли оптическая плотность при разных концентрациях ДМСО, применим дисперсионный анализ Крускала–Уоллиса. Для этого в программе Statistica необходимо выполнить следующее:

- Запустить модуль анализа (рис. 3.17) из меню **Statistics > Nonparametrics > Comparing multiple independent samples** (Сравнение нескольких независимых выборок).

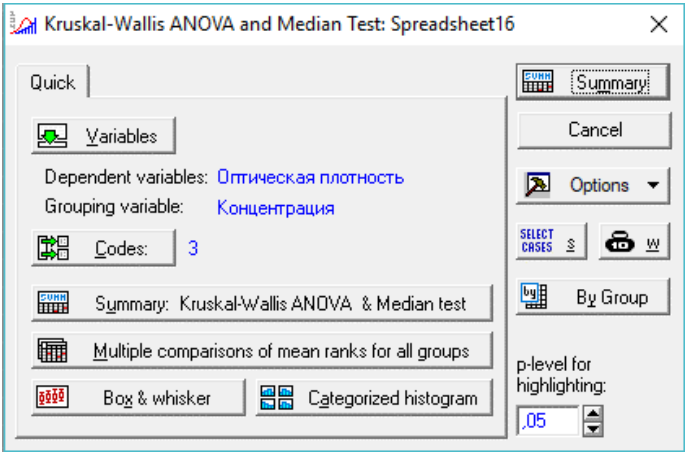


Рис. 3.17. Модуль дисперсионного анализа Крускала–Уоллиса

• Нажать кнопку **Variables** и выбрать зависимую («Оптическая плотность») и группирующую («Концентрация») переменные. Нажать на кнопки: **Factor codes > All > OK > OK**.

• Нажать на кнопку **Summary: Kruskal-Wallis ANOVA and Median test** (*Результат.* ANOVA по Крускалу–Уоллису и медианный тест).

В появившейся таблице с результатами (рис. 3.18) необходимо отыскать величину ошибки P для нулевой гипотезы о том, что оптическая плотность нейтрофилов при исследованных концентрациях не различается. Если P > 0,05 (как в нашем примере), то следует вывод об отсутствии различий между сравниваемыми группами. Вместе с результатом Kruskal–Wallis ANOVA на отдельном листе в рабочей книге программа выдает также результаты *медианного теста*. Этот тест проверяет ту же нулевую гипотезу, что и H-тест Крускала–Уоллиса, однако является менее мощным.

Kruskal-Wallis ANOVA by Ranks; Оптическая плотность (Spreadsheet16)
Independent (grouping) variable: Концентрация
Kruskal-Wallis test: H (2, N= 21) =3,565653 p =,1682

Depend.: Оптическая плотность	Code	Valid N	Sum of Ranks
3% ДМСО	102	7	55,00000
5% ДМСО	103	7	78,00000
8% ДМСО	104	7	98,00000

Рис. 3.18. Результаты дисперсионного анализа Крускала–Уоллиса

Сделайте вывод о различии между сравниваемыми группами.

Глава 4. Регрессионные модели зависимостей между количественными переменными

4.1. Линейные регрессионные модели зависимостей между количественными переменными

Уравнение линейной регрессии. Оценка параметров уравнения регрессии по выборочным данным медицинских исследований. Остаточная дисперсия. Стандартные ошибки коэффициентов уравнения линейной регрессии. Проверка статистической значимости линейной зависимости.

Хотя коэффициент корреляции и позволяет количественно охарактеризовать степень связи, с его помощью невозможно предсказать, чему в среднем будет равно значение одного признака при заданном значении другого признака. Решить эту задачу позволяет регрессионный анализ.

Оценка коэффициентов линейной регрессии

Достаточно часто связь между двумя биологическими признаками имеет линейный характер, что, как известно, можно выразить в виде уравнения:

$$y = a + bx,$$

где y и x – анализируемые признаки;

a – свободный член уравнения; при b = 0 получаем y = a, то есть a – это точка, в которой линия регрессии пересекается с осью OY (эту точку называют также «y-пересечением», или «Intercept»);

b – коэффициент регрессии, отражающий угол наклона линии регрессии. Чем больше b отличается от 0, тем сильнее связь между анализируемыми признаками.

Даже если связь между биологическими признаками носит нелинейный характер (например, экспоненциальный), практически всегда можно выделить участки, хорошо аппроксимируемые линейной регрессией. Поэтому именно с нее мы и начнем рассмотрение регрессионного анализа. Приведенное выше уравнение можно использовать для описания связи между двумя признаками лишь при выполнении следующих обязательных условий:

- Зависимость между признаками носит линейный характер;
- Оба признака распределены нормально.

На рис. 4.1 приведены данные о систолическом давлении крови у людей разного возраста.

	1 Возраст, лет	2 Давление, мм.рт.ст.
1	30	108
2	30	110
3	40	125
4	40	120
5	40	118
6	50	132
7	50	137
8	50	134
9	60	148
10	60	151
11	60	146
12	60	147
13	70	162
14	70	156
15	70	164
16	70	159

Рис. 4.1. Пример оформления данных для выполнения регрессионного анализа

Рассчитаем коэффициенты линейного регрессионного уравнения, описывающего связь между этими двумя биологическими параметрами. Расчет коэффициентов регрессионных уравнений можно выполнить в нескольких модулях программы Statistica. Мы воспользуемся модулем **Multiple Regression Analysis** (Анализ множественной регрессии). Для выполнения регрессионного анализа необходимо:

• Запустить соответствующий модуль из меню **Statistics** или нажать на кнопку на дополнительной панели инструментов.

• В появившемся окне нажать на кнопку **Variables** и указать, какая из анализируемых переменных является зависимой (**Dependent variable**), а какая – независимой (**Independent variable**) (в нашем примере систолическое давление зависит от возраста). Нажать кнопку ОК. В итоге появится окно (рис. 4.2), которое уже на данном этапе анализа содержит некоторые важные его результаты:

а) Dependent: имя зависимой переменной;

б) No. of cases: число наблюдений;

в) Intercept: значение свободного члена регрессионного уравнения;

г) Std. error: стандартная ошибка свободного члена регрессионного уравнения;

д) Multiple R: коэффициент множественной корреляции;

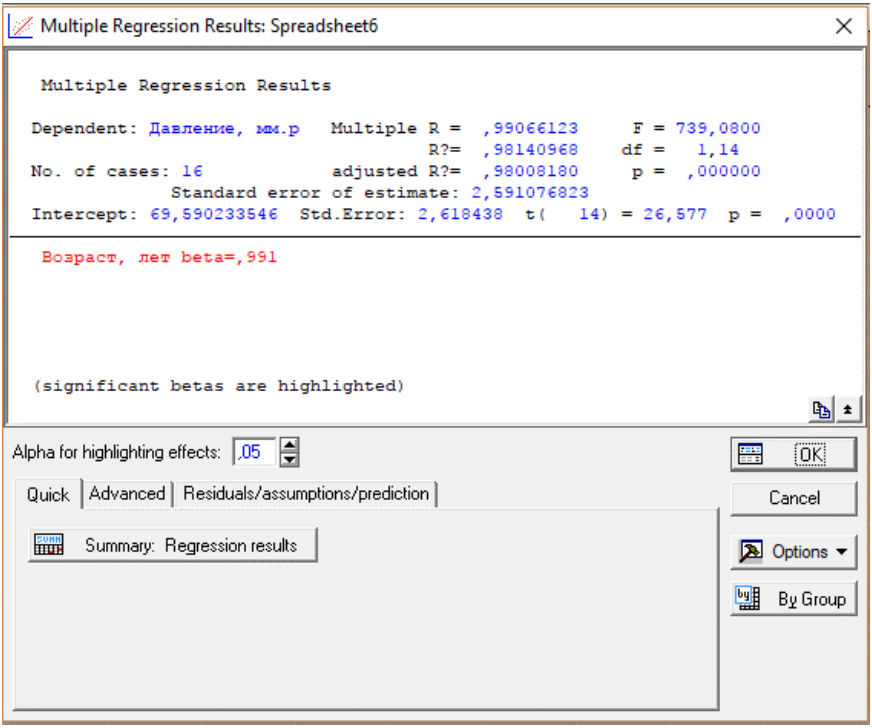


Рис. 4.2. Окно предварительных результатов регрессионного анализа

е) R^2: коэффициент детерминации. Это очень важный показатель в регрессионном анализе. Он изменяется от 0 до 1 и отражает «качество» рассчитанной регрессии, показывая долю (%) общего разброса выборочных точек, которая «объясняется» построенной регрессией (например, при $R^2 = 0{,}85$, следует вывод о том, что 85 % дисперсии зависимой переменной y объясняется вариацией независимой переменной x);

ж) Adjusted R^2: скорректированный на число степеней свободы коэффициент детерминации (Adjusted R-square = 1 - (1 - R-square)×[n/(n - p)], где n – число наблюдений, p – число независимых переменных плюс 1);

з) Standard error of estimate: параметр, отражающий степень разброса выборочных значений относительно линии регрессии;

и) F, df и p: F-критерий, число степеней свободы, принятое при его расчете, и вероятность ошибки для нулевой гипотезы F-теста. F-тест в регрессионном анализе применяется для оценки статистической значимости модели. При $p < 0{,}05$ можно заключить, что рассчитанная регрессия удовлетворительно описывает связь между исследуемыми признаками;

к) t(df) и p: критерий Стьюдента t используется для проверки нулевой гипотезы о равенстве 0 свободного члена регрессионного уравнения. P – вероятность ошибки для этой нулевой гипотезы;

л) beta: стандартизованный коэффициент регрессии – это коэффициент регрессии, который мы получили бы в случае предварительной стандартизации обеих переменных (при таком преобразовании, когда их средние значения стали бы равны 0, а стандартные отклонения -1). Расчет beta позволяет оценить, в какой степени значения зависимой переменной определяются значениями независимой переменной. Beta может оказаться особенно полезным показателем при включении в анализ нескольких независимых переменных, выражающихся в разных единицах измерения – в таком случае коэффициент отражал бы удельный вклад каждой из этих переменных в вариацию зависимой переменной. При наличии одной независимой переменной коэффициент beta идентичен Multiple R.

• Нажать кнопку **Summary: Regression results** (Результаты регрессионного анализа). Появится таблица со следующими результатами анализа (рис. 4.3):

а) Beta: стандартизованный коэффициент регрессии;

б) Std. err. of beta: стандартная ошибка стандартизованного коэффициента регрессии;

в) B: один из самых важных столбцов в этой таблице, поскольку именно он содержит искомые значения свободного члена регрессионного уравнения (в строке Intercept) и коэффициента регрессии (нижняя строка таблицы);

г) Std. err. of B: стандартные ошибки коэффициентов уравнения;

д) t(df): значения t-критерия Стьюдента, который используется для проверки гипотезы о равенстве обоих коэффициентов уравнения 0;

е) p-level: вероятность ошибки для нулевой гипотезы о равенстве коэффициентов уравнения нулю.

Regression Summary for Dependent Variable: Давление, мм.рт.ст. (Spreadsheet6)
R= ,99066123 R?= ,98140968 Adjusted R?= ,98008180
F(1,14)=739,08 p<,00000 Std.Error of estimate: 2,5911

N=16	Beta	Std.Err. of Beta	B	Std.Err. of B	t(14)	p-level
Intercept			69,59023	2,618438	26,57700	0,000000
Возраст, лет	0,990661	0,036440	1,29830	0,047756	27,18603	0,000000

Рис. 4.3. Результаты регрессионного анализа

Из рис. 4.3 видно, что оба коэффициента регрессии статистически значимо отличаются от 0 ($p << 0{,}001$) и что в целом построенная регрессионная модель отлично описывает связь между возрастом людей и си-

столическим давлением крови (R 2 = 98 %). Само же рассчитанное уравнение мы можем записать следующим образом:

$$H = 1{,}298 \times A + 69{,}590,$$

где H – давление, A – возраст человека.

Важной частью регрессионного анализа является анализ остатков (остатки представляют собой разности между наблюдаемыми значениями зависимой переменной и теми ее значениями, которые предсказываются регрессионной моделью). Он запускается путем нажатия кнопки **Perform residual analysis** (Выполнить анализ остатков) на закладке **Residuals / Assumptions / Predictions** (Остатки / Условия / Предсказания) (рис. 4.2, 4.4).

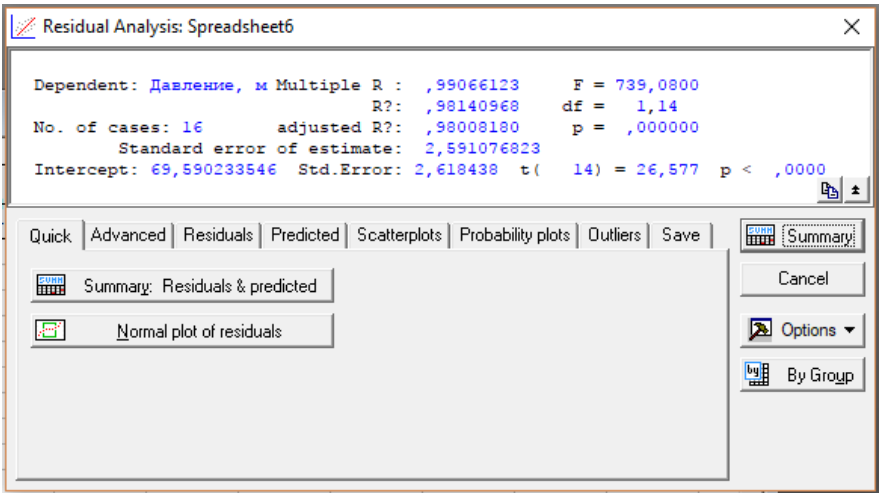


Рис. 4.4. Подмодуль анализа остатков модуля множественного регрессионного анализа

Первое, что нужно проверить в отношении остатков – это нормальность их распределения. Для этого на закладке **Quick** подмодуля анализа остатков (рис. 4.4) можно нажать кнопку Normal plot of residuals, чтобы построить график нормальных вероятностей. Если точки на этом графике достаточно тесно укладываются вдоль теоретически ожидаемой прямой, то можно заключить, что остатки распределяются нормально (рис. 4.5). Иначе линейная регрессионная модель для анализируемых переменных будет неприменима (в ряде случаев, однако, помогает трансформация данных, см. ниже).

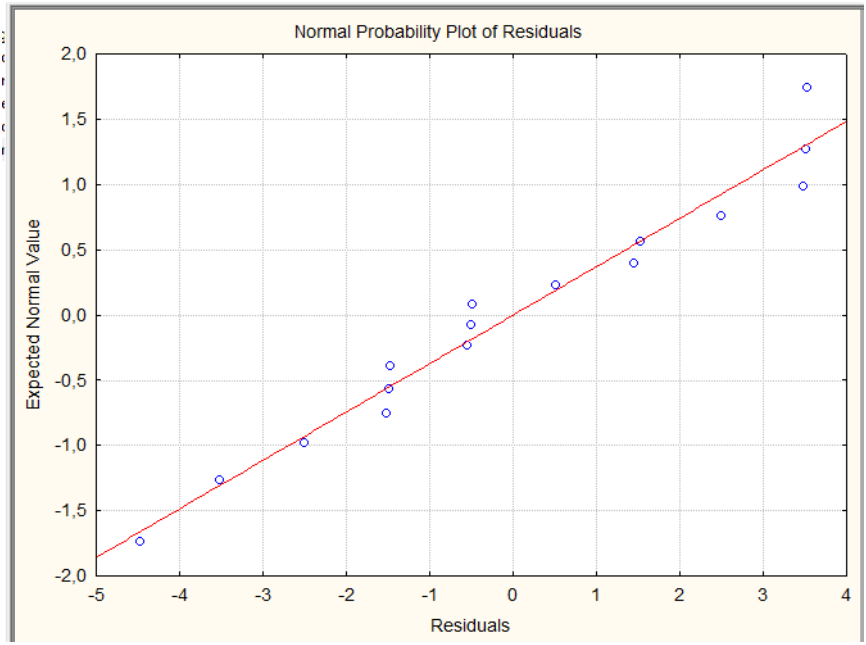


Рис. 4.5. Результат проверки нормальности распределения остатков с помощью графика нормальных вероятностей

Второе условие в отношении остатков состоит в том, что их дисперсия должна оставаться неизменной во всем диапазоне значений анализируемых переменных. Для проверки этого условия на закладке **Scatterplots** (Диаграммы рассеяния); (рис. 4.4) можно нажать кнопку **Predicted vs. residuals**, чтобы построить график зависимости значений остатков от предсказываемых моделью значений зависимой переменной. Если проверяемое условие выполняется, то точки на этом графике будут располагаться хаотично, не проявляя никакой закономерности (рис. 4.6).

Если же в расположении точек имеется тенденция (разброс увеличивается слева направо, точки тесно укладываются вдоль прямой и т. п.), линейный регрессионный анализ также неприменим (однако и в этом случае может помочь трансформация исходных данных, см. ниже).

В рассмотренном примере оба условия в отношении остатков выполняются, что еще раз подтверждает адекватность рассчитанной регрессионной модели для описания связи между артериальным давлением и возрастом людей.

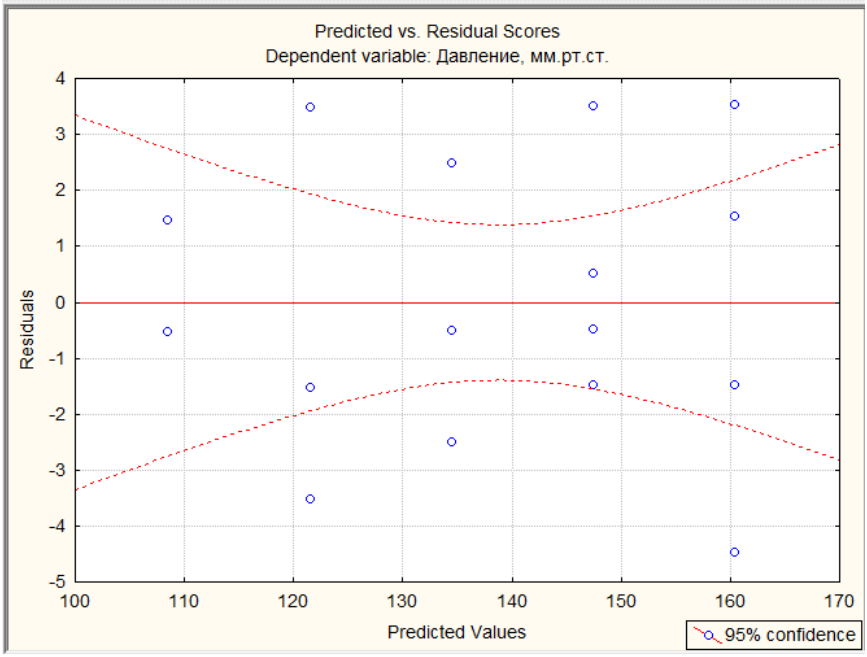


Рис. 4.6. Результат проверки однородности дисперсии остатков

<u>Трансформация нелинейно связанных признаков</u>

Многие биологические признаки проявляют нелинейный характер связи. Например, размеры тела и интенсивность метаболических процессов имеют степенную или экспоненциальную зависимость. Как же быть, если встает задача по расчету уравнения зависимости между подобными переменными? Иногда в таких случаях помогает определенная трансформация исходных данных, которая позволяет перевести их в другую шкалу измерения и тем самым «выровнять» нелинейную зависимость между признаками. На рис. 4.7 приведены данные объема тимуса ($см^3$) и количества вырабатываемых $СД8^+$. Необходимо рассчитать уравнение, описывающее связь между этими двумя признаками.

	1 Объем, см^3	2 CD8+, млн.
1	0,556	0,018
2	0,600	0,021
3	0,694	0,035
4	0,779	0,055
5	0,849	0,13
6	0,954	0,28
7	1,099	0,48
8	1,102	0,54
9	1,204	0,72
10	1,205	0,731

Рис. 4.7. Пример двух нелинейно связанных биологических признаков

Характер связи между двумя переменными можно проверить еще до запуска модуля регрессионного анализа. Для этого достаточно построить диаграмму рассеяния (**Graphs > Scatterplots**), подобную приведенной на рис. 4.8.

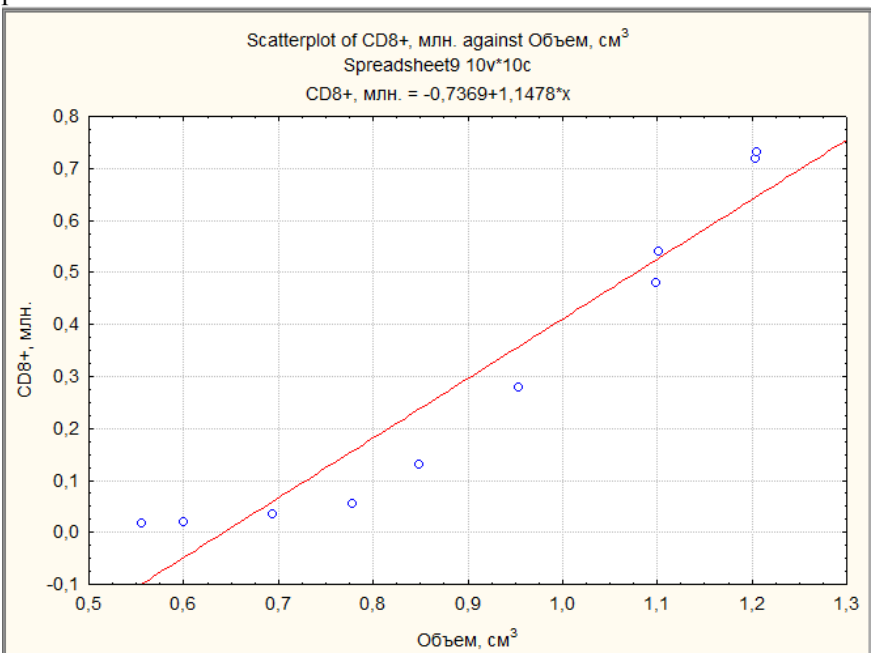


Рис. 4.8. Визуальная оценка характера связи между двумя признаками при помощи диаграммы рассеяния

Из этого рисунка видно, что связь между размером тимуса и выработкой СД8$^+$ далека от линейной и больше напоминает степенную зави-

симость вида y = ax+b. Линейный регрессионный анализ здесь неприменим. Однако степенные и экспоненциальные зависимости обычно легко можно привести к линейным путем логарифмирования значений одного или (чаще) обоих анализируемых признаков. Такую трансформацию в программе Statistica выполнить очень просто.

При подготовке данных к анализу в окне настройки свойств переменных последним можно присваивать т.н. длинные имена (Long name) в виде формул. Прологарифмируем столбец 1, содержащий значения объема тимуса (рис. 4.7). Для этого воспользуемся столбцом 3, который следует за данными по интенсивности дыхания. Кликнув два раза по его заголовку, мы попадем в окно настроек свойств переменной. В поле Long name необходимо ввести формулу =log10(v1), где v1 – это столбец с данными об объеме тимуса. В поле Name введем короткое имя переменной, например, «Log S» (рис. 4.9).

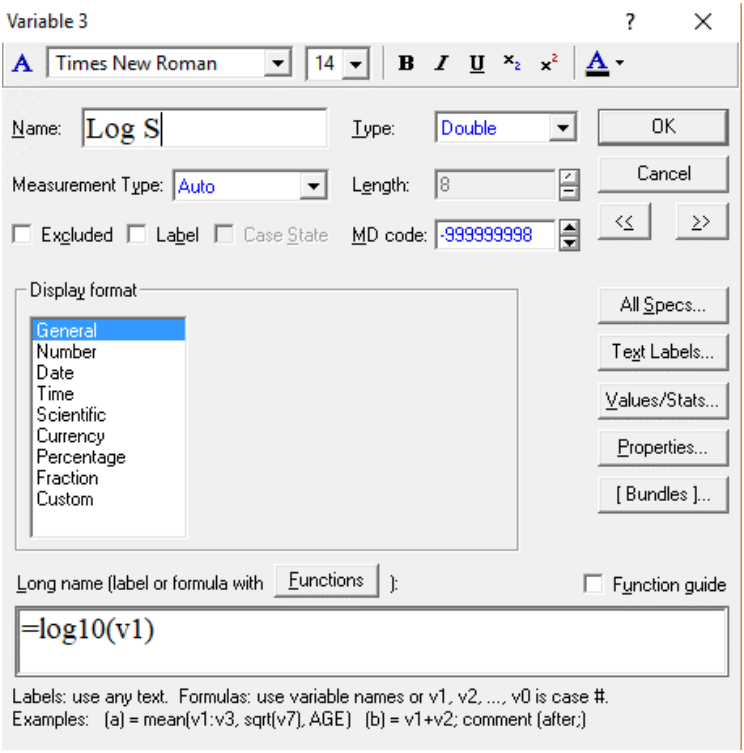


Рис. 4.9. Введение формулы =log10(v1) в поле длинного имени переменной

После нажатия на кнопку OK появится небольшая панель с надписью «Expression OK. Recalculate the variable now?» (Формула введена

правильно. Пересчитать значения переменной?). Нажимаем Yes. В результате в третьем столбце таблицы с данными появятся пересчитанные значения первого столбца. Аналогичную операцию логарифмирования (рис. 4.10) следует выполнить и для столбца с данными выработки СД8^{+}. Для этого можно воспользоваться 4-м столбцом таблицы и в качестве его длинного имени ввести формулу =log10(v2).

	1 Объем, см^3	2 CD8+, млн.	3 Log S	4 Log CD8+
1	0,556	0,018	-0,2549252	-1,7447275
2	0,600	0,021	-0,2218487	-1,6777807
3	0,694	0,035	-0,1586405	-1,455932
4	0,779	0,055	-0,1084625	-1,2596373
5	0,849	0,13	-0,0710923	-0,8860566
6	0,954	0,28	-0,0204516	-0,552842
7	1,099	0,48	0,04099769	-0,3187588
8	1,102	0,54	0,04218159	-0,2676062
9	1,204	0,72	0,08062649	-0,1426675
10	1,205	0,731	0,08098705	-0,1360826

Рис. 4.10. Результат логарифмирования первого и второго столбцов в таблице с данными

Если теперь мы построим диаграмму рассеяния для прологарифмированных значений анализируемых признаков, то увидим, что точки гораздо теснее укладываются вдоль прямой линии и мы можем применить обычный регрессионный анализ для нахождения зависимости между признаками (рис. 4.11). Регрессионное уравнение для прологарифмированных переменных запишется в виде $\log y = b \times \log x + \log a$.

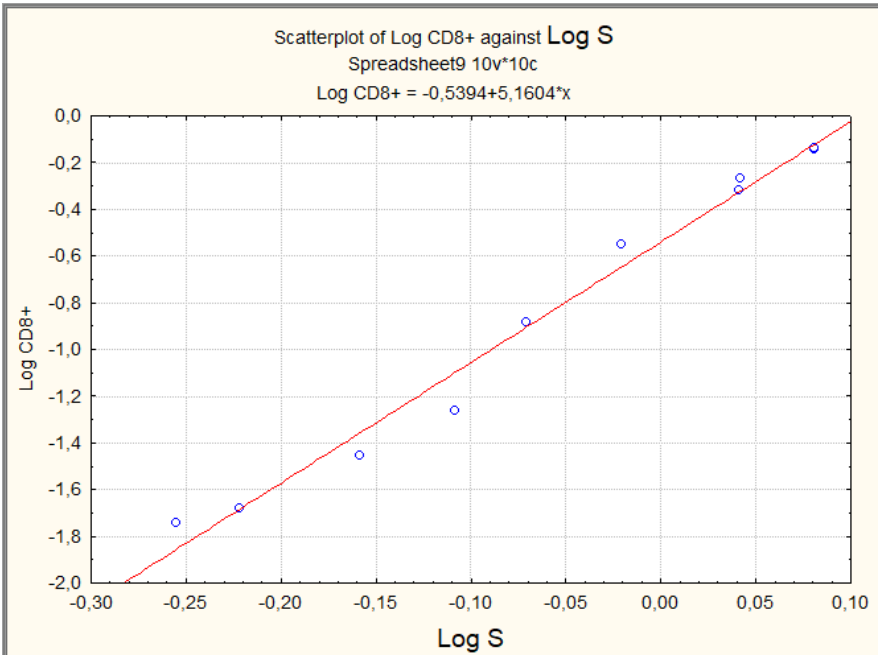


Рис. 4.11. Визуальная оценка характера связи между двумя признаками после их логарифмирования

Необходимо отметить, что процедура трансформации исходных данных часто применяется не только для «выравнивания» нелинейных связей между признаками. Часто логарифмирование или иные распространенные способы трансформации позволяют привести асимметрично распределенные данные к нормальному распределению, а также добиться однородности дисперсии в группах, подлежащих анализу с использованием параметрических методов.

4.2. Нелинейные модели регрессии

Полиномиальные и нелинейные модели регрессии

Трансформация данных не всегда позволяет привести зависимость между двумя признаками к линейной форме. Кроме того, научный интерес может представлять уравнение именно нелинейной связи. В программе Statistica можно рассчитать уравнение зависимости практически любого функционального вида. Для этого служит специальный модуль **Nonlinear estimations** (Нелинейные оценивания) (**Statistics > Advanced Linear/Nonlinear models > Nonlinear estimations**, либо кнопка на допол-

нительной панели инструментов). Применим этот анализ к приведенным выше данным о связи между объемом тимуса ($см^3$) и количеством вырабатываемых $СД8^+$ (рис. 4.7). Проверим, имеется ли между полученными данными степенная зависимость вида $y = ax^b$. В списке опций модуля **Nonlinear estimations** выберем **User specified regression, least squares** (Заданная пользователем регрессия, метод наименьших квадратов) (рис. 4.12).

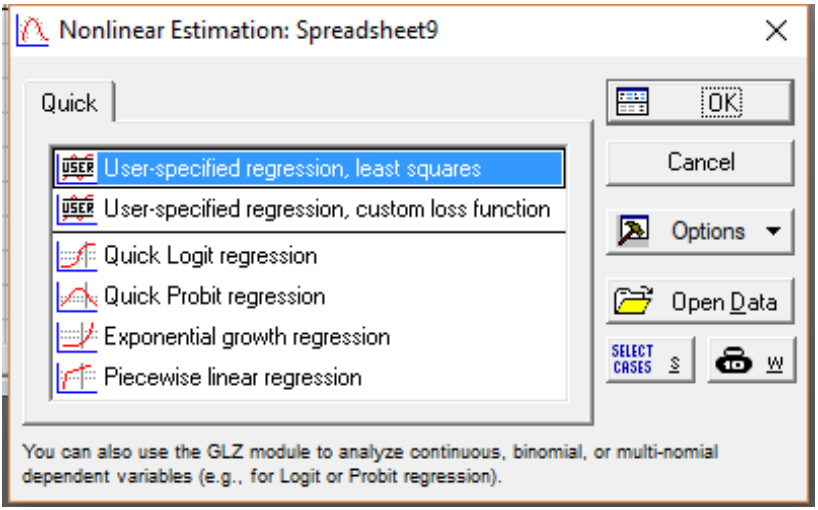


Рис. 4.12. Окно модуля Nonlinear estimations

В появившемся в результате этого окне необходимо нажать кнопку **Function to be estimated** (Оцениваемая функция) и затем ввести формулу v2 = a*v1^b (рис. 4.13).

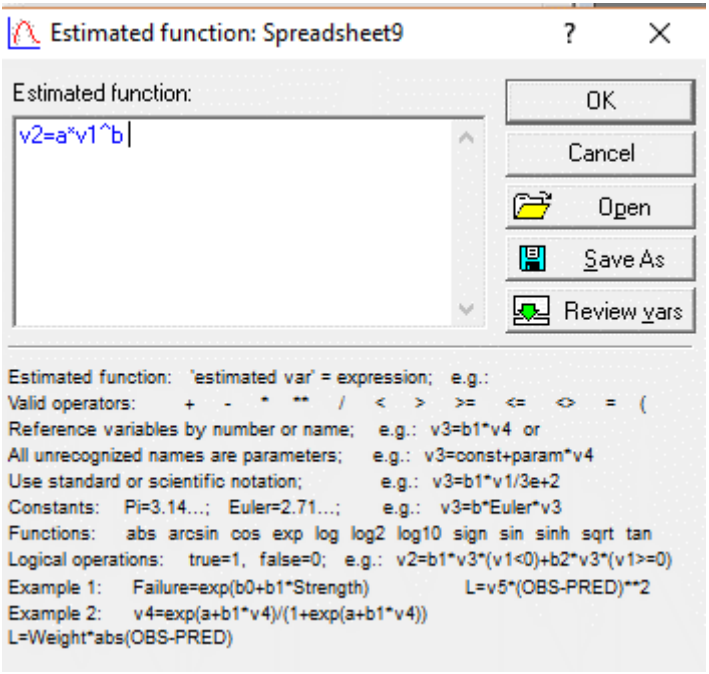


Рис. 4.13. Окно выбора результатов нелинейного регрессионного анализа

После последовательного нажатия на кнопки **ОК > ОК** откроется диалоговое окно, в котором можно выбрать алгоритм расчета коэффициентов уравнения, получить параметры описательной статистики для каждой из переменных (закладка **Review**), и т. п. Мы оставим все настройки без изменений и нажмем кнопку **ОК**. Откроется новое окно, в котором можно выполнить анализ остатков (закладка **Residuals**), проследить последовательность того, как программа рассчитывала коэффициенты уравнения (кнопка **Iteration history**), получить параметры описательной статистики и т. д.

Само же это уравнение можно записать в виде:

$$R = 0{,}3128 \times L4{,}6276.$$

Глава 5. Многомерные данные: снижение размерности

5.1. Факторный анализ

Проблема анализа многомерных данных. Снижение размерности исходного признакового пространства и отбор наиболее информативных показателей. Сущность методов факторного анализа и их классификация. Общий вид линейной модели факторного анализа и основные задачи факторного анализа. Проблема общности в факторном анализе, способы вычисления оценок общностей. Общий алгоритм факторного анализа. Геометрическое представление наблюдаемых объектов в пространстве элементарных признаков и латентных факторов.

Факторный анализ является одним из разделов многомерного статистического анализа. Он основан на многомерном нормальном распределении, то есть каждый из используемых признаков изучаемого объекта должен иметь нормальный закон распределения. Факторный анализ исследует внутреннюю структуру ковариационной и корреляционной матриц системы признаков изучаемого объекта.

Пусть в изучаемой выборке отобрано N проб. В каждой из них измерены значения K признаков и получены значения случайных многомерных нормально распределенных величин:

$$X_t = X_{1t}, X_{2t}, \ldots, X_{kt}),$$

где $t = 1, 2, \ldots, N$.

Ясно, что эти значения случайных многомерных величин обусловлены какими-то объективными причинами, которые будем называть факторами. Предполагается, что число этих факторов всегда меньше, чем число K измеряемых параметров (признаков) изучаемого объекта. Эти факторы являются скрытыми, их нельзя непосредственно измерить и поэтому они представляются гипотетическими. Однако имеются методы их выявления, которые и составляют сущность факторного анализа.

В факторном анализе решаются следующие задачи:

Определить количество действующих факторов и указать их относительную интенсивность

Выявить признаковую структуру факторов, то есть показать, какими признаками объекта обусловлено действие того или иного фактора и в какой относительной мере

Выявить факторную структуру изучаемых признаков объекта, то есть показать долю влияния каждого из факторов на значение того или иного признака этого объекта

Воссоздать в факторном координатном пространстве облик изучаемого объекта, используя вычисляемые значения факторов для каждого наблюдения исходной выборочной совокупности.

Пусть в каждой пробе пациента мы измерили четыре характеристики, которые обусловлены действием двух факторов F_1 и F_2. Фактор F_1 действует на все четыре характеристики объекта, а фактор F_2 действует лишь на два признака X_1 и X_2.

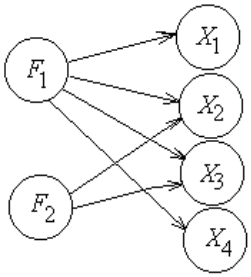


Рис. 5.1. Модель факторного анализа

Значит значения признаков X_1 и X_4 определяются только фактором F_1, а признаки X_2 и X_3 определяются совокупным действием фактором F_1 и F_2. Но мы всего этого пока не знаем и перед нами стоит задача оценить интенсивность влияния факторов F_1 и F_2 на признаки X_i и выделить в значениях X_i те части, которые обусловлены действием каждого из факторов F_1 и F_2 в отдельности.

Для решения этой задачи предполагают, что X_i линейно зависят от F_m (m = 1,2). Для нашего случая имеем

$$X_i = a_{i1} \cdot F_1 + a_{i2} \cdot F_2,$$

где I = 1, 2, 3, 4.

a_{i1},·a_{i2} – коэффициенты, называемые факторными нагрузками.

Существует **<u>две модели факторного анализа</u>**:

1) ***<u>метод главных компонент (МГК)</u>***, в котором наблюдаемые значения каждого из признаков X_1 представляются в виде линейных комбинаций факторных нагрузок a_{ij} и факторов F_j, где j = 1, 2, …, m, причем $m<<k$.

$$X_i = \sum_{j=1}^{m} a_{ij} F_j,$$

где m – число факторов.

2) ***модель собственного факторного анализа (ФА)***, когда наблюдаемые значения определяются не только факторами, но и действием локальных случайных причин

$$X_i = \sum_{j=1}^{m} a_{ij} F_j + e_i.$$

Факторный анализ позволяет решить две важные проблемы исследователя: описать объект измерения всесторонне и компактно. С помощью факторного анализа возможно выявление скрытых переменных факторов, отвечающих за наличие линейных статистических корреляций между наблюдаемыми переменными.

Две основных цели факторного анализа:

1) определение взаимосвязей между переменными (классификация переменных), то есть «объективная R-классификация»;

2) сокращение числа переменных необходимых для описания данных.

При анализе в один фактор объединяются сильно коррелирующие между собой переменные и, как следствие, происходит перераспределение дисперсии между компонентами, получается максимально простая и наглядная структура факторов. После объединения коррелированность компонент внутри каждого фактора между собой будет выше, чем их коррелированность с компонентами из других факторов. Эта процедура также позволяет выделить латентные переменные, что бывает особенно важно при анализе социальных представлений и ценностей. Например, анализируя оценки, полученные по нескольким шкалам, исследователь замечает, что они сходны между собой и имеют высокий коэффициент корреляции. Он может предположить, что существует некоторая латентная переменная, с помощью которой можно объяснить наблюдаемое сходство полученных оценок. Такую латентную переменную называют фактором.

Практическое выполнение факторного анализа начинается с проверки его условий. В обязательные условия факторного анализа входят:

- все признаки должны быть количественными;
- число наблюдений должно быть не менее чем в два раза больше числа переменных;
- выборка должна быть однородна;
- исходные переменные должны быть распределены симметрично;
- факторный анализ осуществляется по коррелирующим переменным.

Основные понятия факторного анализа: **фактор** — скрытая переменная и **нагрузка** — корреляция между исходной переменной и фактором.

Сущностью факторного анализа является ***процедура вращения факторов***, то есть перераспределения дисперсии по определенному методу. Цель ортогональных вращений — определение простой структуры факторных нагрузок, целью большинства косоугольных вращений является определение простой структуры вторичных факторов, то есть косоугольное вращение следует использовать в частных случаях. Поэтому ортогональное вращение предпочтительнее.

Вращение бывает:

- ортогональным,
- косоугольным.

При первом виде вращения каждый последующий фактор определяется так, чтобы максимизировать изменчивость, оставшуюся от предыдущих, поэтому факторы оказываются независимыми, некоррелированными друг от друга (к этому типу относится *метод главных компонент*). Второй вид – это преобразование, при котором факторы коррелируют друг с другом. Преимущество косоугольного вращения состоит в следующем: когда в результате его выполнения получаются ортогональные факторы, можно быть уверенным, что эта ортогональность действительно им свойственна, а не привнесена искусственно. Существует около 13 методов вращения в обоих видах, в статистической программе Statistica доступны пять: три ортогональных, один косоугольный и один комбинированный, однако из всех наиболее употребителен ортогональный метод «варимакс». Метод «варимакс» максимизирует разброс квадратов нагрузок для каждого фактора, что приводит к увеличению больших и уменьшению малых значений факторных нагрузок. В результате простая структура получается для каждого фактора в отдельности.

Главной проблемой факторного анализа является ***выделение и интерпретация главных факторов***. При отборе компонент исследователь обычно сталкивается с существенными трудностями, так как не существует однозначного критерия выделения факторов, и потому здесь неизбежен субъективизм интерпретаций результатов. Существует несколько часто употребляемых критериев определения числа факторов. Некоторые из них являются альтернативными по отношению к другим, а часть этих критериев можно использовать вместе, чтобы один дополнял другой.

Критерий Кайзера или критерий собственных чисел предложен Кайзером, и является, вероятно, наиболее широко используемым. Отбираются только факторы с собственными значениями равными или большими 1. Это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается.

Критерий каменистой осыпи или критерий отсеивания является графическим методом, впервые предложенным психологом Кэттелом. Собственные значения возможно изобразить в виде простого графика. Кэттел предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только «факториальная осыпь». «Осыпь» является геологическим термином, обозначающим обломки горных пород, скапливающиеся в нижней части скалистого склона. Однако этот критерий отличается высокой субъективностью и, в отличие от предыдущего критерия, статистически необоснован. Недостатки обоих критериев заключаются в том, что первый иногда сохраняет слишком много факторов, в то время как второй может сохранить слишком мало факторов. Однако оба критерия вполне хороши при нормальных условиях, когда имеется относительно небольшое число факторов и много переменных. На практике возникает важный *вопрос:* «Когда полученное решение может быть содержательно интерпретировано?». В этой связи предлагается использовать еще несколько критериев.

Критерий значимости особенно эффективен, когда модель генеральной совокупности известна и отсутствуют второстепенные факторы. Но критерий непригоден для поиска изменений в модели и реализуем только в факторном анализе по методу наименьших квадратов или максимального правдоподобия.

Критерий доли воспроизводимой дисперсии. Факторы ранжируются по доле детерминируемой дисперсии, когда процент дисперсии оказывается несущественным, выделение следует остановить. Желательно, чтобы выделенные факторы объясняли более 80 % разброса. Недостатки критерия: во-первых, субъективность выделения, во-вторых, специфика данных может быть такова, что все главные факторы не смогут совокупно объяснить желательного процента разброса. Поэтому главные факторы должны вместе объяснять не меньше 50,1 % дисперсии.

Критерий интерпретируемости и инвариантности сочетает статистическую точность с субъективными интересами. Согласно ему, главные факторы можно выделять до тех пор, пока будет возможна их ясная интерпретация. Она, в свою очередь, зависит от величины факторных нагрузок, то есть, если в факторе есть хотя бы одна сильная нагрузка, то он может быть интерпретирован. Возможен и обратный вариант – если сильные нагрузки имеются, однако интерпретация затруднительна, то от этой компоненты предпочтительно отказаться.

Практика показывает, что если вращение не произвело существенных изменений в структуре факторного пространства, то это свидетельствует о его устойчивости и стабильности данных. Возможны еще два варианта:

• сильное перераспределение дисперсии – результат выявления латентного фактора;

• очень незначительное изменение (десятые, сотые или тысячные доли нагрузки) или его отсутствие вообще, если при этом сильные корреляции могут иметь только один фактор, то это однофакторное распределение.

Последнее возможно, например, когда на предмет наличия определенного свойства проверяются несколько социальных групп, однако искомое свойство есть только у одной из них.

5.2. Метод главных компонент

Метод главных компонент. Формальная постановка задачи. Аппроксимация данных линейными многообразиями. Метод главных компонент через критерий максимизации разброса в данных. Поиск ортогональных проекций с наибольшим рассеянием. Поиск ортогональных проекций с наибольшим среднеквадратичным расстоянием между точками. Матрица преобразования к главным компонентам. Оценка числа главных компонент по правилу сломанной трости. Механическая аналогия и метод главных компонент для взвешенных данных.

Имеется система переменных $Y_1...Y_3$, $X_4...X_{17}$. Значения переменных или признаков для каждого из наблюдений известны. Имеющаяся информация представлена в виде матрицы, каждая строка которой состоит из значений одного показателя для каждого из объектов исследования. Предполагается, что каждый элемент этой матрицы является результатом воздействия некоторого числа гипотетических общих факторов и одного характерного фактора, то есть представляют собой линейную комбинацию ненаблюдаемых, гипотетических, непосредственно не измеряемых факторов. Следовательно, любой метод факторного анализа имеет одну главную задачу: представить результирующий фактор в виде линейной комбинации некоторого числа общих факторов и одного характерного фактора.

Рассмотрим процедуру решения практической задачи методом факторного анализа в системе Statistica.

Откройте модуль **Факторный анализ** с помощью переключателя модулей **Statistica**.

Дважды щелкните на имени модуля либо, выделив его, щелкните на кнопке **Switch to**. На экране появиться стартовая панель модуля **Factor Analysis** (**Факторный анализ**) (рис. 5.2).

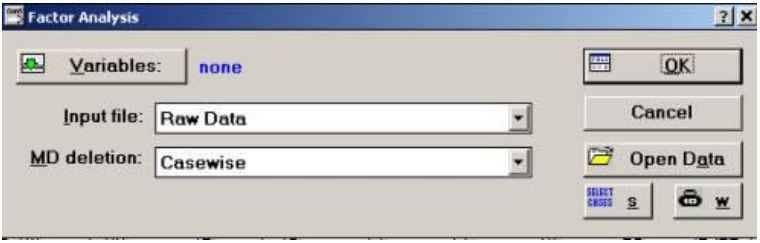


Рис. 5.2. Вид стартовой панели модуля Factor Analysis (факторный анализ)

В строке **Input File** (**Файл входных данных**) укажите тип исходного файла, с которым вы будете работать.

В модуле возможны следующие типы исходных данных:

Correlation Matrix (**Корреляционная матрица**);

Raw Data (**Исходные данные**).

Выберите, например, **Raw Data**. Это обычный файл данных, где по строкам записаны значения переменных.

В строке **Missing Data** (**Пропущенные данные**) задайте способ обработки пропущенных значений.

Casewise – способ исключения пропущенных случаев. Заключается в том, что в электронной таблице, содержащей данные, игнорируются все строки (случаи), в которых имеется хотя бы одно пропущенное значение. Это относится ко всем переменным. В таблице остаются только случаи, в которых нет ни одного пропуска.

Pairwise – парный способ исключения пропущенных значений (игнорируются пропущенные случаи не для всех переменных, а лишь для выбранной пары). Случаи, в которых нет пропусков, используются в обработке, например, при поэлементном вычислении корреляционной матрицы, когда последовательно рассматриваются все пары переменных. Очевидно, в способе **Pairwise** остается больше наблюдений для обработки, чем в способе **Casewise**. Тонкость, однако, состоит в том, что в **Pairwise** оценки различных коэффициентов корреляции строятся по разному числу наблюдений.

Mean Substitution – подстановка среднего вместо пропущенных значений.

Выберите **Casewise** – способ исключения пропущенных случаев.

Откройте файл, в котором содержатся данные для анализа. Щелкните мышью на кнопке **Open Data** – Открыть данные. Перед вами появиться окно **Open Data File** – Открыть файл данных.

Выберите файл. Выделите его имя и нажмите **OK** (либо просто дважды щелкните мышью по имени файла). Выбрав файл, вы автомати-

чески окажетесь в стартовой панели модуля **Factor analysis (Выбрать переменные для факторного анализа**).

В данном окне (рис. 5.3) выберите переменные либо, высвечивая их мышью в представленном списке, удерживая ее левую кнопку и двигаясь вверх либо вниз, либо, например, набрав номер переменных в нижней строке. (Максимальное число переменных –300).

Кнопка **Sellect All** (**Выбрать все**) позволяет выбрать все переменные сразу. Это самый быстрый способ.

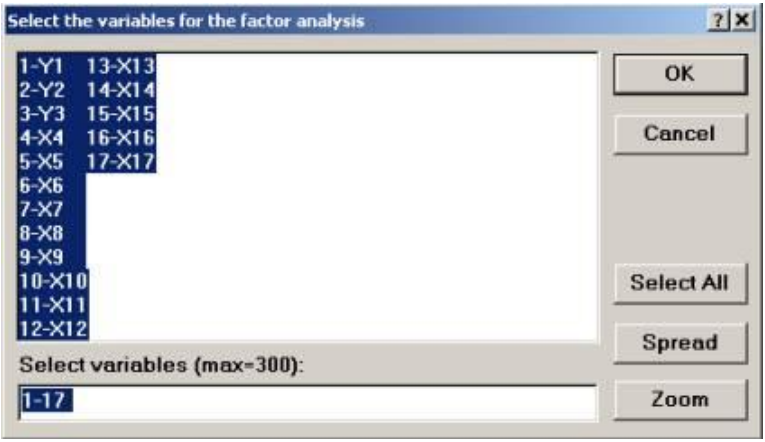


Рис. 5.3. Вид окна выбора переменных

Щелкните по кнопке **Spread (Распахнуть**), просмотрите расширенное описание переменных. Щелкните по кнопке **Sellect All** (**Выбрать все**) и выберите все переменные.

Щелкнув в стартовом окне модуля на кнопку **OK**, вы начнете анализ выбранных переменных.

Statistica обработает пропущенные значения тем способом, какой вы ей указали, вычислит корреляционную матрицу и предложит на выбор несколько методов факторного анализа.

Вычисление корреляционной матрицы (если она не задается сразу) – первый этап факторного анализа!

Щелкните на кнопку **OK**, перед вами появиться окно **Define Method of Factor Extraction** (**Определить метод выделения факторов**). Данное окно имеет следующую структуру (рис. 5.4).

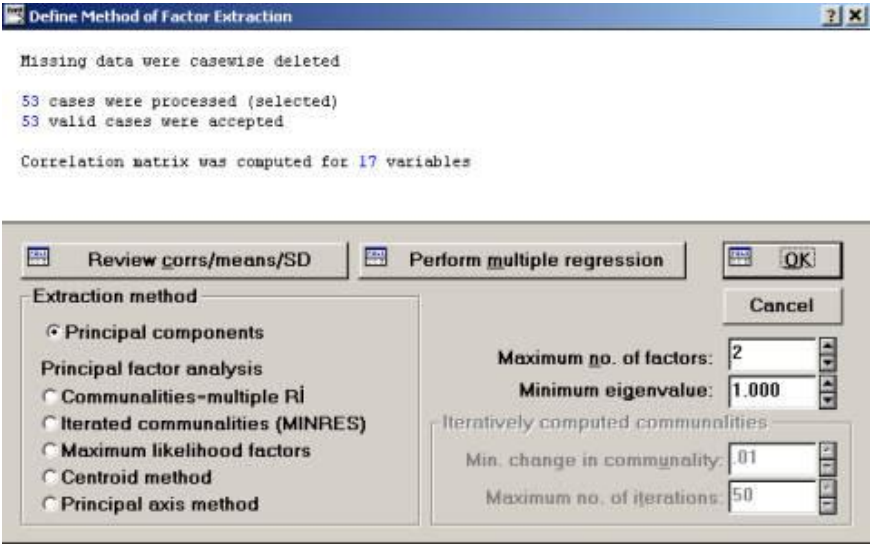


Рис. 5.4. Вид окна выбора метода выделения факторов

Первая (верхняя) часть окна является информационной: здесь сообщается, что пропущенные значения обработаны методом **Casewise.** Обработано 53 случая, из них 53 приняты для дальнейших вычислений. Корреляционная матрица вычислена для 17 первых.

Вторая (нижняя) часть диалогового окна окно **Define Method of Factor Extraction (Определить метод выделения факторов)** содержит 4 функциональных кнопки и опции выбора метода – левая часть окна, а также поля, в которых проводятся установки для итеративного вычисления общностей.

Обратите внимание на поля в правой части окна:

Maximum no.of factors – Максимальное число факторов;

Minimum eigenvalue – Минимальное собственное значение.

Эти поля определяют число факторов, которые будут выделены системой. Собственные значения, меньше указанного в поле, игнорируются.

Рассмотрим функциональные клавиши окна. Из 4 функциональных кнопок 2 – стандартные: кнопка **OK** нажимается для запуска вычислительной процедуры, когда выбран метод и проведены установки в полях. Кнопка **Cancel (Отменить)** прекращает работу в окне и возвращает в стартовое окно модуля.

Инициировав кнопку **Review corrs/means/SD (Просмотреть корреляции/средние/стандартные отклонения)**, вы откроете окно **Просмотреть описательные статистики**.

С помощью кнопки **Perform multiple regression (Выполнить множественную регрессию)**, не выходя из модуля, вы можете выполнить множественную регрессию.

Группа опций, объединенных под заголовком **Extraction method** – метод выделения, позволяет выбрать метод обработки.

Как отмечалось в математическом анализе, в зависимости от критерия оптимальности возможен анализ либо методом **Principal components** – методом главных компонент, либо одним из методов, объединенных в группу **Principal factor analysis – анализ главных (общих) факторов**.

Система предлагает следующие методы в группе **Principal factor analysis– анализ главных факторов**:

Communalities = multiple R2** (общности равны квадрату коэффициента множественной корреляции);

Iterated communalities (MINRES) – метод итеративных общностей (минимальных остатков);

Centroid method – центроидный метод;

Principal axis method – метод главных осей.

В окне **Define Method of Factor Extraction (Определить метод выделения факторов**) щелкните на кнопке **Review corrs/means/SD (Просмотреть корреляции /средние/стандартные отклонения**).

Перед вами появилось окно просмотра описательных статистик для анализируемых данных (рис. 5.5), где можно посмотреть средние, стандартные отклонения, корреляции, ковариации, построить различные графики.

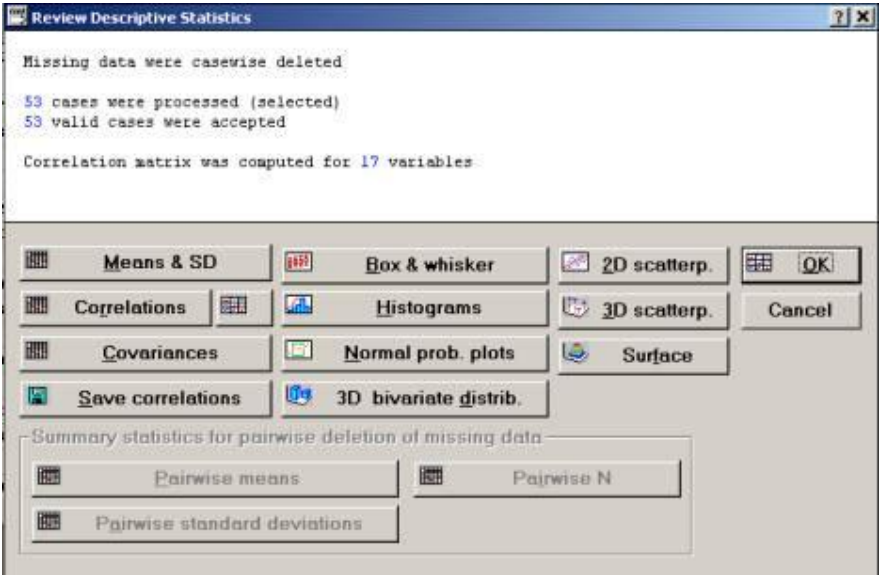


Рис. 5.5. Вид окна просмотра описательных статистик

Щелкните на кнопке **Correlations** (**Корреляции**). Вы увидите на экране корреляционную матрицу выбранных ранее переменных (рис. 5.6).

Correlations (data.sta)

Continue... Casewise deletion of MD N=53

Variable	Y1	Y2	Y3	X4	X5	X6	X7	X8	X9	X10	X11	X12
Y1	1.00	.55	.13	.10	.28	.03	-.17	-.06	-.09	.04	.26	.44
Y2	.55	1.00	.04	-.37	.13	.06	-.12	.13	-.09	-.08	.24	.83
Y3	.13	.04	1.00	.08	-.07	-.19	.22	.41	.04	.08	.07	.02
X4	.10	-.37	.08	1.00	-.14	-.05	-.04	-.07	.11	-.19	-.07	-.36
X5	.28	.13	-.07	-.14	1.00	-.10	-.10	-.05	-.12	.07	.04	.03
X6	.03	.06	-.19	-.05	-.10	1.00	-.20	-.24	.02	-.05	.23	-.00
X7	-.17	-.12	.22	-.04	-.10	-.20	1.00	.17	-.23	-.06	.05	.01
X8	-.06	.13	.41	-.07	-.05	-.24	.17	1.00	-.10	-.16	-.01	.14
X9	-.09	-.09	.04	.11	-.12	.02	-.23	-.10	1.00	.31	-.24	-.10
X10	.04	-.08	.08	-.19	.07	-.05	-.06	-.16	.31	1.00	-.05	-.14
X11	.26	.24	.07	-.07	.04	.23	.05	-.01	-.24	-.05	1.00	.31
X12	.44	.83	.02	-.36	.03	-.00	.01	.14	-.10	-.14	.31	1.00
X13	-.10	-.42	.06	.35	.02	.08	-.06	-.09	-.07	.05	-.05	-.37
X14	.20	.29	-.11	-.01	-.09	-.17	.07	.20	-.00	-.27	.14	.62
X15	-.09	-.30	-.20	.22	.01	.28	.16	-.13	-.02	-.15	-.08	-.29
X16	-.25	-.28	-.15	-.14	.11	-.10	-.06	.08	.07	.30	-.28	-.31
X17	.02	-.06	-.33	-.01	-.06	.22	-.16	-.25	-.05	-.08	-.04	-.05

Рис. 5.6. Вид корреляционной матрицы выбранных переменных

Сохраните корреляционную матрицу, инициировав кнопку **Save correlations** (**Сохранить корреляции**). Далее в модуле **Факторный анализ** при дальнейших исследованиях можно сразу работать с корреляционной матрицей.

Щелкните на кнопку **Continue** (**Продолжить**) в верхнем левом углу таблицы и вернитесь в окно **Define Method of factor Extraction** (**Определить метод выделения факторов**).

Выберите опцию **Principal components** (**Главные компоненты**) и щелкните на кнопке **OK**.

Система быстро произведет вычисления, и на экране появится окно **Factor Analysis Results** (**Результаты факторного анализа**) (рис. 5.7).

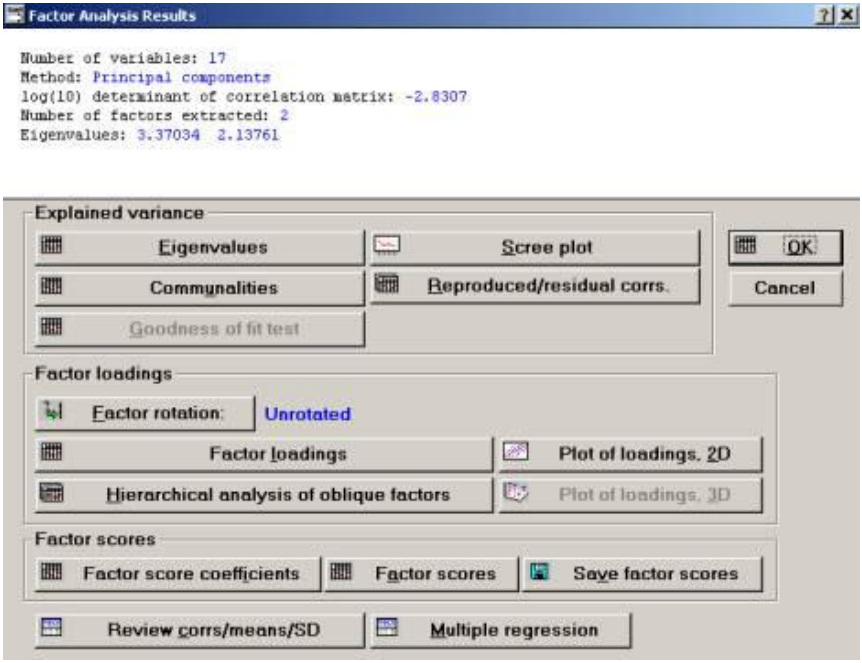


Рис. 5.7. Вид окна результатов факторного анализа

В верхней части окна **Результаты факторного анализа** дается информационное сообщение:

Number of variables (число анализируемых переменных) – 17;

Method (метод анализа – главные компоненты);

log(10) determination of correlation matrix (десятичный логарифм детерминанта корреляционной матрицы) – 2,8307;

Number of Factor extraction (число выделенных факторов) – 2;

Eigenvalues (собственные значения) – 3,37034; 2,13761.

В нижней части окна находятся функциональные кнопки, позволяющие всесторонне просмотреть результаты анализа численно и графически.

Щелкните на кнопку **Plot of Loadings 2D** (**Двумерный график нагрузок**) и посмотрите результаты факторного анализа на графике (рис. 5.8).

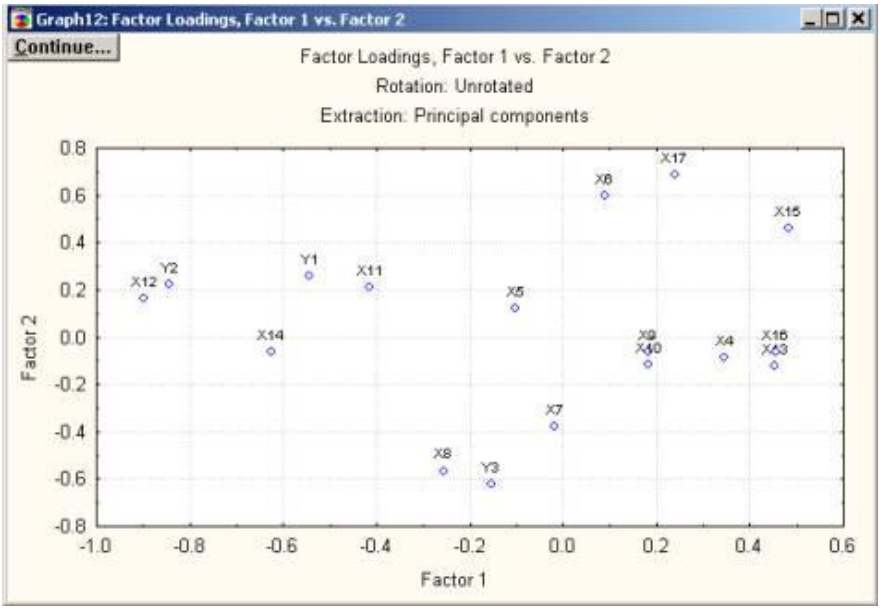
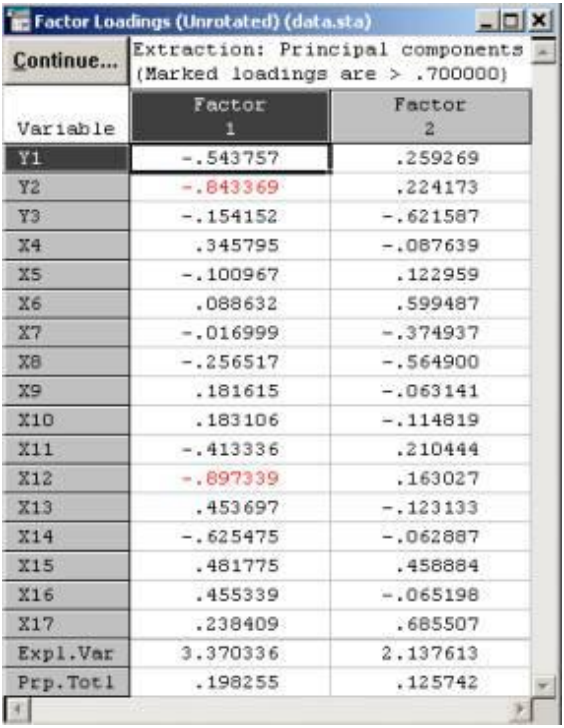


Рис. 5.8. Вид графика нагрузок

Щелкните на кнопке **Factor Loadings** (**Факторные нагрузки**), и посмотрите нагрузки численно (рис. 5.9).



Variable	Factor 1	Factor 2
Y1	-.543757	.259269
Y2	-.843369	.224173
Y3	-.154152	-.621587
X4	.345795	-.087639
X5	-.100967	.122959
X6	.088632	.599487
X7	-.016999	-.374937
X8	-.256517	-.564900
X9	.181615	-.063141
X10	.183106	-.114819
X11	-.413336	.210444
X12	-.897339	.163027
X13	.453697	-.123133
X14	-.625475	-.062887
X15	.481775	.458884
X16	.455339	-.065198
X17	.238409	.685507
Expl.Var	3.370336	2.137613
Prp.Totl	.198255	.125742

Рис. 5.9. Вид таблицы факторных нагрузок

Рассматривая решение на графике, обратите внимание на нагрузки параметров выделенные красным цветом. Их трудно проинтерпретировать, возникает вопрос, какой смысл придать второму фактору. В этом случае целесообразно прибегнуть к повороту осей, надеясь получить решение, которое можно интерпретировать в предметной области.

Щелкните на кнопку **Factor rotation** (**Вращение факторов**).

Цель вращения – получение простой структуры, при которой большинство наблюдений находится вблизи осей координат. При случайной конфигурации наблюдений невозможно получить простую структуру.

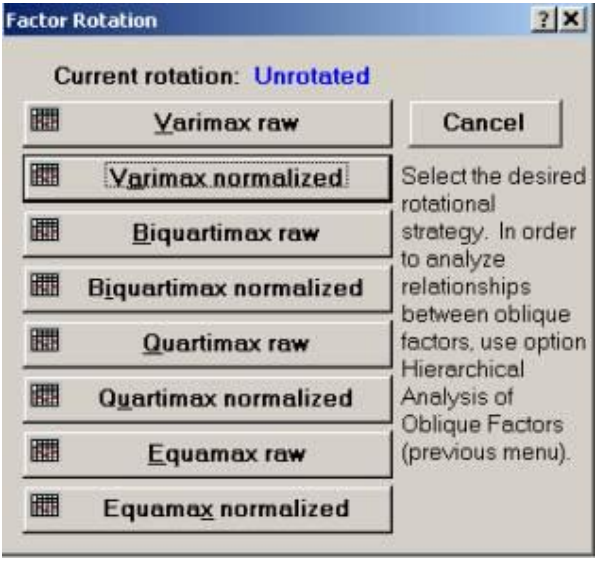


Рис. 5.10. Вид таблицы факторных нагрузок

В данном окне (рис. 5.10) вы можете выбрать различные повороты оси. Окно предлагает несколько возможностей оценить и найти нужный поворот следующими методами:

- **Varimax** – Варимакс;
- **Biquartimax** – Биквартимакс;
- **Quartimax** – Квартимакс;
- **Equamax** – Эквимакс.

Дополнительный термин в названии методов – **normalized** (нормализованные) – указывает на то, что факторные нагрузки в процедуре нормализуются: делятся на корень квадратный из соответствующей дисперсии. Термин **raw** (исходные) показывает, что вращаемые нагрузки не нормализованы.

Инициируйте кнопку **Varimax normalized** (**Варимакс нормализованный**).

Система произведет вращение факторов методом нормализованного варимакса, и окно **Factor Analysis Results** (**Результаты факторного анализа**) снова появится на мониторе. Вновь инициируйте в этом окне кнопку **Plot of Loadings 2D** (**Двумерный график нагрузок**). Вы опять увидите график нагрузок (рис. 5.11).

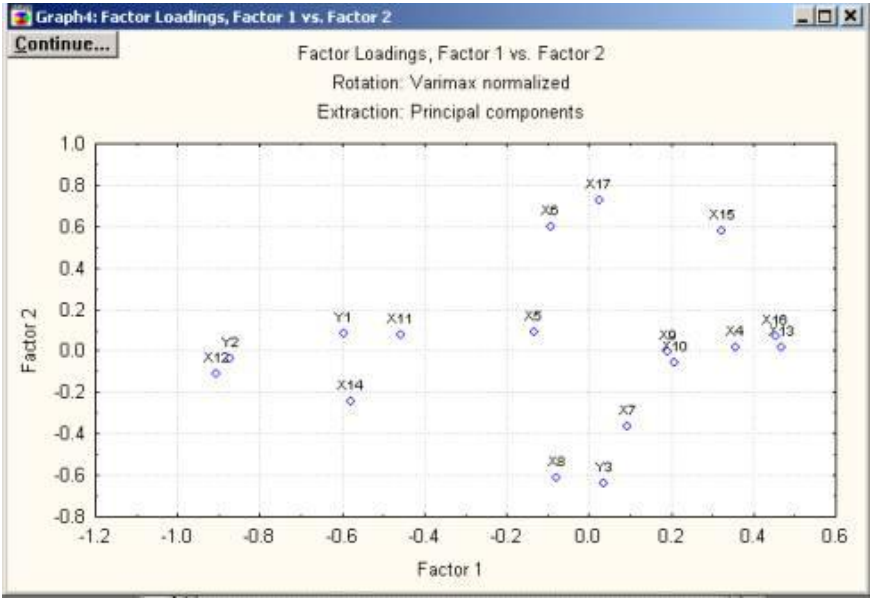
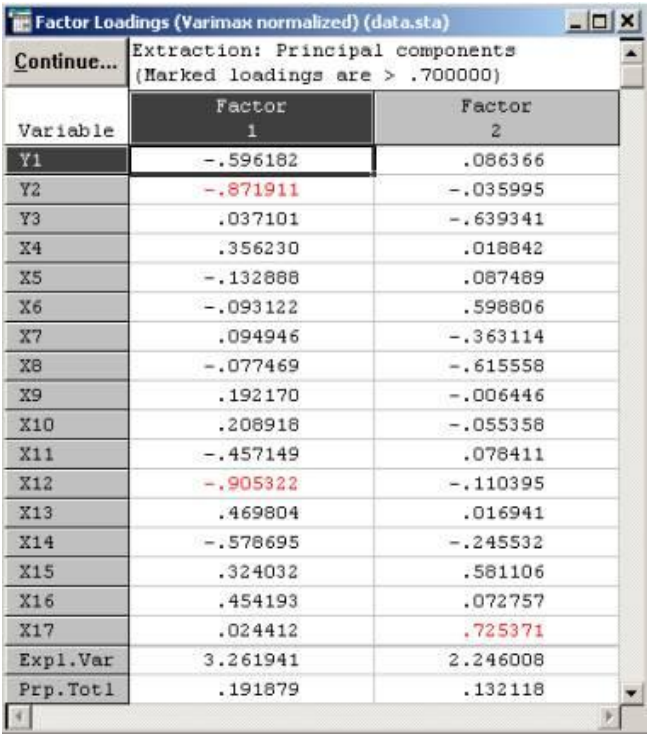


Рис. 5.11. Вид графика нагрузок после вращения факторов методом нормализованного варимакса

Конечно, этот график немного отличается от предыдущего. Посмотрим еще нагрузки численно, инициировав кнопку **Факторные нагрузки**. Щелкните на кнопке **Factor loadings**, вы откроете окно, изображенное на рис. 5.12.



Variable	Factor 1	Factor 2
Y1	-.596182	.086366
Y2	-.871911	-.035995
Y3	.037101	-.639341
X4	.356230	.018842
X5	-.132888	.087489
X6	-.093122	.598806
X7	.094946	-.363114
X8	-.077469	-.615558
X9	.192170	-.006446
X10	.208918	-.055358
X11	-.457149	.078411
X12	-.905322	-.110395
X13	.469804	.016941
X14	-.578695	-.245532
X15	.324032	.581106
X16	.454193	.072757
X17	.024412	.725371
Expl.Var	3.261941	2.246008
Prp.Totl	.191879	.132118

Рис. 5.12. Вид таблицы факторных нагрузок

Теперь найденное решение уже можно интерпретировать. Факторы чаще интерпретируют по нагрузкам. Первый фактор теснее всего связан с *Y*2 и *X*12. Второй фактор (только с одним показателем) *X*17.

Возникает *вопрос*: «Сколькими же факторами следует ограничиваться на практике?» Для этого в программном пакете Statistica существует критерий **Scree plot** (**Критерий каменистой осыпи**). В окне **Factor Analysis Results** нажмите кнопку **Scree plot** получите следующий график (рис. 5.13).

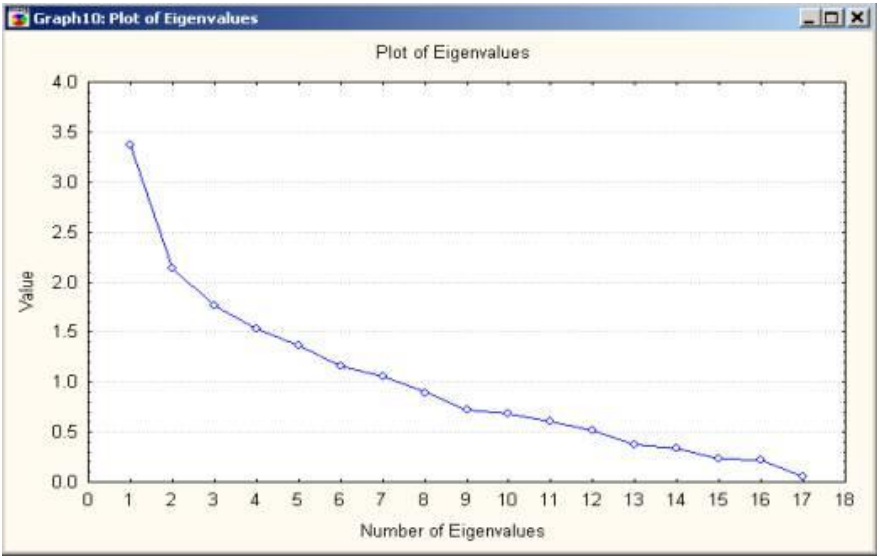


Рис. 5.12. Вид графика собственных значений

В точках с координатами 1 и 2 осыпание замедляется наиболее существенно, следовательно, можно ограничиваться одним или двумя факторами.

5.3. Методы классификации без обучения

Методы кластерного анализа. Общая постановка задачи классификации без обучения. Расстояния между объектами и меры близости объектов друг к другу. Расстояние между классами объектов. Проблема использования метрики при измерении расстояний. Выбор функции расстояния в непрерывных пространствах: Манхэттенская, Евклидова и Чебышевская метрики. Измерение расстояний в полярной системе координат. Классификация объектов при известном и неизвестном числе классов с учетом выбора метрики.

Кластеризация – это разделение множества входных векторов на группы (кластеры) по степени «схожести» друг на друга.

Кластеризация в Data Mining приобретает ценность тогда, когда она выступает одним из этапов анализа данных, построения законченного аналитического решения. Аналитику часто легче выделить группы схожих объектов, изучить их особенности и построить для каждой группы отдельную модель, чем создавать одну общую модель для всех данных. Таким приемом постоянно пользуются в медико-биологических исследо-

ваниях, выделяя группы пациентов, нуклеотидных повторностей, иммунологических параметров и разрабатывая для каждой из них отдельную стратегию анализа.

Для того, чтобы сравнивать два объекта, необходимо иметь ***критерий***, на основании которого будет происходить сравнение. Как правило, таким критерием является расстояние между объектами.

Есть множество мер расстояния, рассмотрим несколько из них.

Евклидово расстояние — наиболее распространенное расстояние. Оно является геометрическим расстоянием в многомерном пространстве:

$$\rho(x, x') = \sqrt{\sum_{i}^{n} (x_i - x'_i)^2}.$$

Квадрат Евклидова расстояния – применяется для придания большего веса более отдаленным друг от друга объектам. Это расстояние вычисляется по формуле:

$$\rho(x, x') = \sum_{i}^{n} (x_i - x'_i)^2.$$

Расстояние городских кварталов (**Манхэттенское расстояние**) – расстояние средних разностей по координатам:

$$\rho(x, x') = \sum_{i}^{n} \left| x_i - x'_i \right|.$$

В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако отметим, что для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).

Расстояние Чебышева. Это расстояние может оказаться полезным, когда желают определить два объекта как «различные», если они различаются по какой-либо одной координате (каким-либо одним измерением):

$$\rho(x, x') = \max(\left| x_i - x'_i \right|).$$

Степенное расстояние. Иногда желают прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Это может быть достигнуто с использованием степенного расстояния:

$$\rho(x, x') = \sqrt[r]{\sum_{i}^{n} (x_i - x'_i)^p},$$

где r и p – параметры, определяемые пользователем. Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра – r и p – равны двум, то это расстояние совпадает с расстоянием Евклида.

Выбор расстояния (критерия схожести) лежит полностью на исследователе. При выборе различных мер результаты кластеризации могут существенно отличаться.

Алгоритм k-means (k-средних)

Наиболее простой, но в то же время достаточно неточный метод кластеризации в классической реализации. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k. Действие алгоритма таково, что он стремится минимизировать среднеквадратичное отклонение на точках каждого кластера. Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров.

Проблемы алгоритма k-means:

- необходимо заранее знать количество кластеров. Нами был предложен метод определения количества кластеров, который основывался на нахождении кластеров, распределенных по некоему закону (в моем случае все сводилось к нормальному закону). После этого выполнялся классический алгоритм k-means, который давал более точные результаты.
- алгоритм очень чувствителен к выбору начальных центров кластеров. Классический вариант подразумевает случайный выбор кластеров, что очень часто являлось источником погрешности. Как вариант решения, необходимо проводить исследования объекта для более точного определения центров начальных кластеров. В нашем случае на начальном этапе предлагается принимать в качестве центов самые отдаленные точки кластеров.
- не справляется с задачей, когда объект принадлежит к разным кластерам в равной степени или не принадлежит ни одному из них.

Рассмотрим процедуру решения практической задачи методом кластерного анализа в системе Statistica.

Задачей кластерного анализа является организация наблюдаемых данных в наглядные структуры. Для решения данной задачи в кластер-

ном анализе используются следующие методы: **Joining (tree clustering)** (*иерархические агломеративные методы или древовидная кластеризация*), **K-means clustering** (*метод К средних*), **Two-way joining** (*двувходовое объединение*).

Рассмотрим процесс формирования выборок в системе Statistica:

1. Из переключателя модулей **Statistica** откройте модуль **Cluster Analysis** (*Кластерный Анализ*). Высветите название модуля и далее нажмите кнопку **Switch to** (*Переключиться в*) либо просто дважды щелкните мышью по названию модуля **Cluster Analysis.**

2. На экране появится стартовая панель модуля (рис. 5.13) **Clustering Method** (*методы кластерного анализа*): **Joining (tree clustering)** (*иерархические агломеративные методы или древовидная кластеризация*), **K-means clustering** (*метод К средних*), **Two-way joining** (*двувходовое объединение*). Разберем каждый из этих методов.

Рис. 5.13. Стартовая панель модуля Clustering Method
(методы кластерного анализа)

Joining (tree clustering) (иерархические агломеративные методы).

1. Откроем файл **(Open Data)** date_1.sta. После выбора **Joining (tree clustering)** и нажатия кн. ОК появляется окно **Cluster Analysis: Joing (Tree Clustering)** (окно ввода режимов работы для иерархических агломеративных методов) (рис. 5.14), в котором кн. **Variables** позволяет выбрать переменные участвующие в классификации. Нажмем на кн. **Variables** и выберем все переменные **Select All**. После соответствующего выбора и нажмем кнопку **OK.**

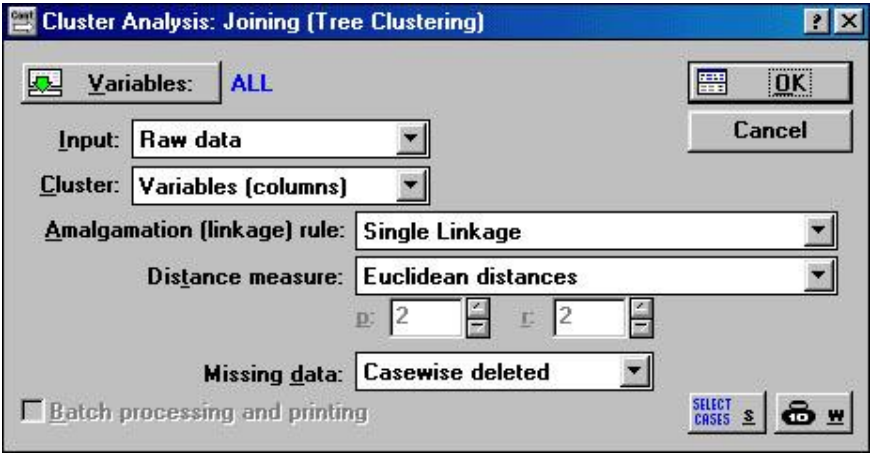


Рис.5.14. Cluster Analysis: Joing (Tree Clustering) (окно ввода режимов работы для иерархических агломеративных методов)

Также можно задать **Input** (*тип входной информации*) и **Cluster** (*режим классификации* (*по признакам или объектам*)). Можно указать **Amalgamation (linkage) rule** (*правило объединения*) и **Distance measure** (*метрика расстояний*). **Codes for grouping variable** (*коды для групп переменной*) будут указывать количество анализируемых групп объектов. **Missing data** (*пропущенные переменные*) позволяет выбрать либо построчное удаление переменных из списка, либо заменить их на средние значения. **Open Data** – позволяет открыть файл с данными. Причем можно указать условия выбора наблюдений из базы данных кнопку **Select Cases**. Можно задавать веса переменным, выбрав их из списка -кнопку **W**. Проставьте значения, как показано на рис. 5.14.

3. После задания всех необходимых параметров и нажатия кн. **ОК** будут произведены вычисления, а на экране появится окно, содержащее результаты кластерного анализа **«Joining Results» (**рис. 5.15).

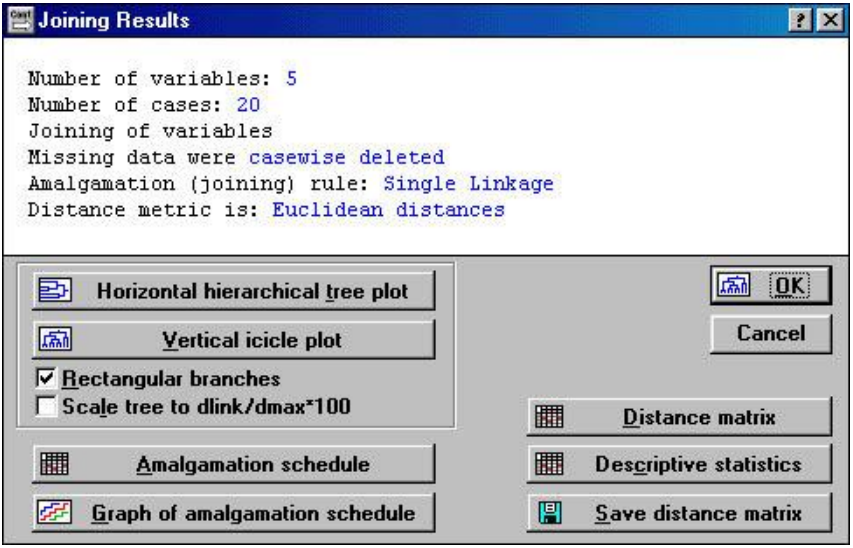


Рис. 5.15. Окно, содержащее результаты кластерного анализа «Joining Results»

Вывод результатов и их анализ

Информационная часть диалогового окна **Joining Results Discriminant Function Analisis Results** (*результаты анализа кластерных функций*) сообщает, что

- Number of variables – число переменных;
- Number of cases – число наблюдений;
- Missing data were casewise deleted – осуществлена классификация наблюдений или переменных (зависит от уровня параметра в строке Cluster в предыдущем окне настроцки.)
- Amalgation (joing) rule – правило объединения кластеров (название иерархического агломеративного метода, заданного в строке Amalgation rules, а в предыдущем окне настрйки);
- Distanse.metric is – Метрика расстояния (зависит от установки в строке Distance measure впредыдущем окне настройки.

Пользователь может вызвать на экран горизонтальную и вертикальную диаграмму (Horizontal hierachical plot или Vertical icicle plot). Наиболее традиционное – вертикальное представление (рис. 5.16).

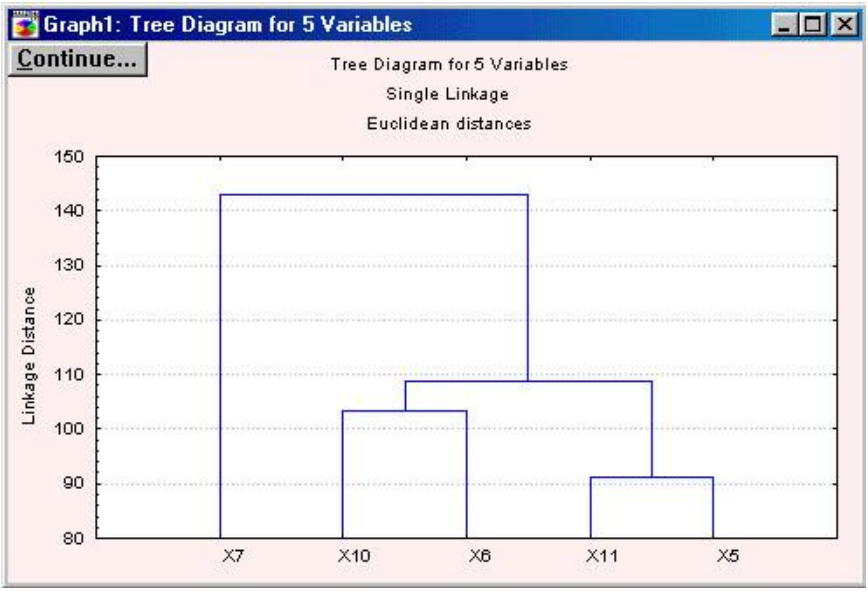


Рис. 5.16. Вид вертикальной дендрограммы

Теперь представим себе, что постепенно (очень малыми шагами) вы «ослабляете» ваш критерий тем, какие объекты являются уникальными, а какие нет. Другими словами, вы понижаете порог, относящийся к решению об объединении двух или более объектов в один кластер. В результате, вы *связываете* вместе все большее и большее число объектов и агрегируете (*объединяете*) все больше и больше кластеров, состоящих из значительно различающихся элементов. Окончательно, на последнем шаге все объекты объединяются вместе. Когда данные имеют ясную «структуру» в терминах кластеров объектов, сходных между собой, тогда эта структура, скорее всего, должна быть отражена в иерархическом дереве различными ветвями. В результате успешного анализа методом объединения появляется возможность обнаружить кластеры (ветви) и интерпретировать их.

Чтобы вернуться в окно, содержащее другие результаты кластерного анализа, необходимо щелкнуть по Continue.

Щелчком мыши можно раскрыть строку Amalgamation schedule, содержащую протокол объединения кластеров (рис. 5.17).

Amalgamation Schedule (date_22.sta)

Continue...

Single Linkage
Euclidean distances

linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5
91,07861	X5	X11			
103,1825	X6	X10			
108,6525	X5	X11	X6	X10	
142,8405	X5	X11	X6	X10	X7

Рис. 5.17. Вид таблицы эвклидовых расстояний между наблюдениями

В заголовке указан иерархический агломеративный метод и метрика расстояния. Таблица может занимать несколько окон.

Следующей в окне результатов идет кнопка **Graph of amalgamation schedule.** После щелчка раскрывается окно, содержащее ступенчатое, графическое изображение изменений расстояний при объединении кластеров (рис. 5.18).

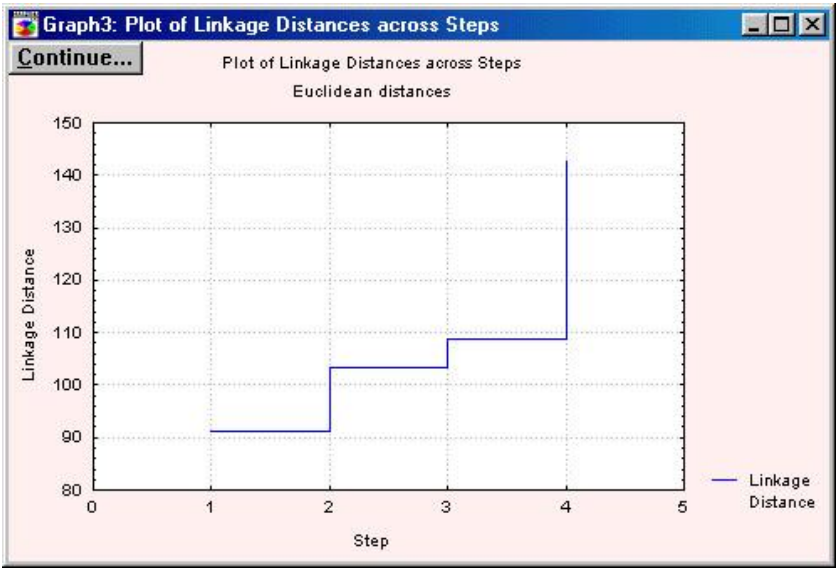


Рис. 5.18. Ступенчатый график изменений расстояний при объединении кластеров

Вернувшись в основное окно результатов и классификации. Для просмотра же матрицы расстояний необходимо осуществить щелчок на строке Distance matrix (рис. 5.19).

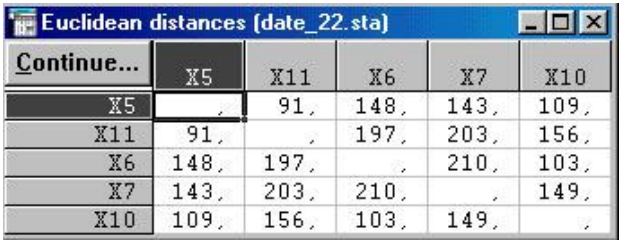


Рис. 5.19. Вид таблицы матрицы расстояний

В основном окне результатов классификации имеется строка **Save distance matrix as:** (*Сохранить матрицу расстояний как:*), позволяющая задать имя файла, в котором будет сохранена матрица расстояний, которая в дальнейшем будет подвергнута обработке.

Строка **Discriptive statistics** содержит такие важнейшие описательные статистики, как среднее **(means)** и среднеквадратическое отклонение **(standart deviations)** для каждого наблюдения. При проведении классификации *n* объектов по *k* признакам, для пользователя представляют большой интерес значения этих показателей для каждого признака. Для того чтобы эти характеристики рассчитывались именно по признакам необходимо вернуться в основное окно настройки параметров и задать в строке **Cluster** значение **«variables (columns)»**.

K-means clustering (метод К средних)

Суть этого метода состоит в следующем: исследователь заранее определяет количество классов (k) на которые необходимо разбить имеющиеся наблюдения, и первые *k* наблюдений становятся центрами этих классов. Для каждого следующего наблюдения рассчитываются расстояния до центров кластеров и данное наблюдение относится к тому кластеру, расстояние до которого было минимальным. После чего для этого кластера (в котором увеличилось количество наблюдений) рассчитывается новый центр тяжести (как среднее по каждому показателю) по всем включенным в кластер наблюдениям.

Предположим, вы уже имеете гипотезы относительно числа кластеров (по наблюдениям или по переменным). Вы можете указать системе образовать ровно три кластера так, чтобы они были настолько различны, насколько возможно. Это именно тот тип задач, которые решает алгоритм метода К-средних. В общем случае метод К-средних строит ровно К-различных кластеров, расположенных на возможно больших расстояниях друг от друга.

1. Из стартовой панели модуля (рис. 5.13) **Clustering Method** (*методы кластерного анализа*) выберем **K-means clustering** (*метод К–средних*). Откроем файл **(Open Data) date_2.sta.**

2. После нажатия кн. **ОК** появляется окно **Cluster Analysis: K-means clustering** (*метод К-средних*) (рис. 5.20), в котором кн. **Variables** позволяет выбрать переменные участвующие в классификации. Нажмем на кн. **Variables** и выберем все переменные **Select All**.

В строке **Cluster** указывается как ведется классификация: при запуске установлен режим **Variables (colums)** – классифицируются переменные на основании их наблюдений. Однако в подавляющем большинстве случаев используется режим **Cases (rows)** – классифицируются наблюдения. Для того чтобы включить режим **Cases (rows),** надо нажать на кнопку в конце строки, после чего в открывшемся окошке подвести курсор на надпись **Cases (rows)** и нажать левую кнопку.

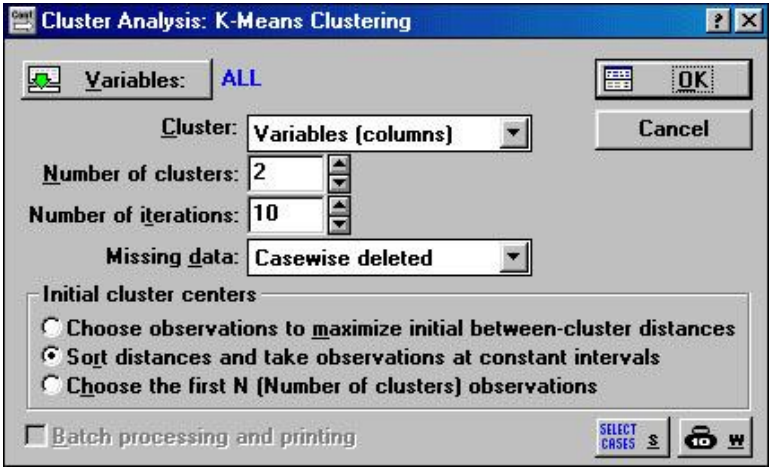


Рис. 5.20. Cluster Analysis: K-means clustering
(окно ввода режимов работы для метода К-средних)

В строке **Number of iterations** указывается количество итераций в расчетах кластеров. Как правило, установленных по умолчанию 10 итераций вполне достаточно. В строке **Missing data** устанавливается режим работы с теми наблюдениями (или переменными, если установлен режим **Variables (columns)** в строке **Cluster**), в которых пропущены данные. Если установить режим **Subsituted by means** (*заменять на среднее*), то вместо пропущенного числа будет использовано среднее по этой переменной (или наблюдению). Переключение в режим **Subsitituted by means** выполняется аналогично переключениям в строке **Cluster.** После соответствующего выбора нажмем кн. **OK.** Будут произведены вычисления и появится новое окно**: «K-means Clustering Results»** (рис. 5.21).

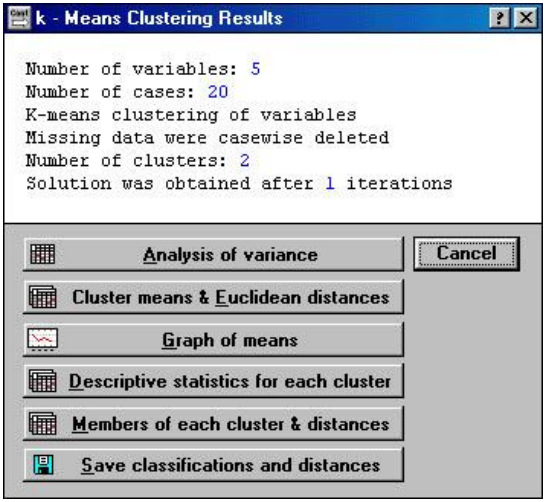


Рис. 5.21. Окно, содержащее результаты кластерного анализа

Вывод результатов и их анализ

В верхней части окна (в том же порядке, как они идут на экране):

- Количество переменных.
- Количество наблюдений.
- Классификация наблюдений (или переменных, зависит от установки в предыдущем окне в строке **Cluster**) методом K-средних.
- Наблюдения с пропущенными данными удаляются (или: изменяются средними значениями. Зависит от установки в предыдущем окне в строке Missing data).
- Количество кластеров.
- Решение достигнуто после : итераций.

В нижней части окнарасположены кнопки для вывода различной информации по кластерам.

1. **Analysis of Variance** (*анализ дисперсии*). После нажатия появляется таблица (рис. 5.22), в которой приведена межгрупповая и внутригрупповая дисперсии, где строки – переменные (наблюдения), столбцы – показатели для каждой переменной: дисперсия между кластерами, число степеней свободы для межклассовой дисперсии, дисперсия внутри кластеров, число степеней свободы для внутриклассовой дисперсии, F-критерий, для проверки гипотезы о неравенстве дисперсий. Проверка данной гипотезы похожа на проверку гипотезы в дисперсионном анализе, когда делается предположение о том, что уровни фактора не влияют на результат.

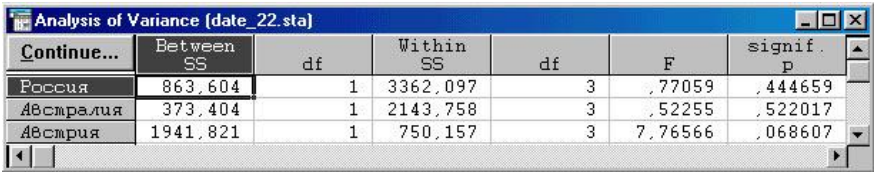


Рис. 5.22. Вид таблицы с результатами дисперсионного анализа

2. **Cluster Means & Euclidean *Distances*** *(средние значения в кластерах и евклидово расстояние*). Выводятся две таблицы. В первой (рис. 5.23) указаны средние величины класса по всем переменным (наблюдениям). По вертикали указаны номера классов, а по горизонтали переменные (наблюдения).

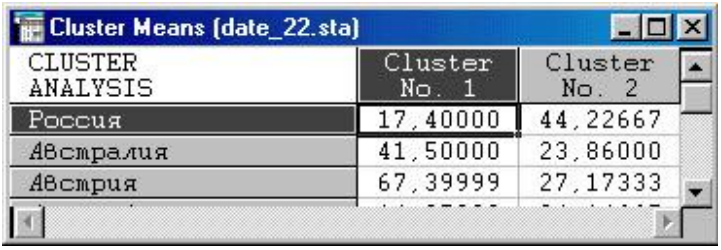


Рис. 5.23. Вид таблицы средних величин класса
по всем переменным (наблюдениям)

Во второй таблице (рис. 5.24) приведены расстояния между классами. И по вертикали и по горизонтали указаны номера кластеров. Таким образом, при пересечении строк и столбцов указаны расстояния между соответствующими классами. Причем выше диагонали (на которой стоят нули) указаны квадраты, а ниже просто Евклидово расстояние.

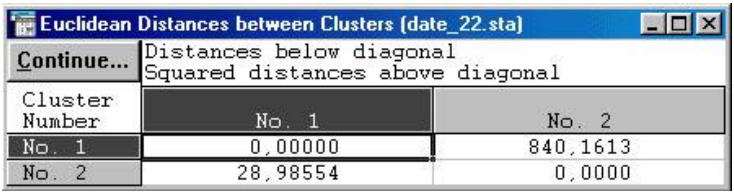


Рис. 5.24. Вид таблицы расстояния между классами

3. **Graph of means** представляет собой графическое изображение (рис. 5.25) информации, содержащейся в таблице, выводимой при нажатии кнопку **Analysis of Variance** (*анализ дисперсии*). На графике показаны средние значения переменных для каждого кластера.

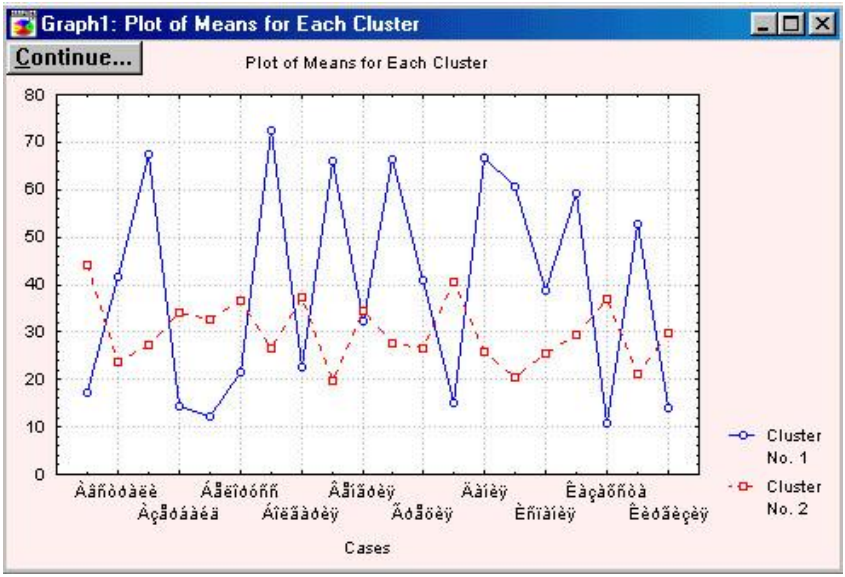


Рис. 5.25. Графическое изображение средние значения переменных для каждого кластера

По горизонтали отложены участвующие в классификации переменные, а по вертикали – средние значения переменных в разрезе получаемых кластеров.

4. **Descriptive Statistics for each cluster** (*описательная статистика для каждого кластера*). После нажатия этой кнопки выводятся окна, количество которых равно количеству кластеров. В каждом таком окне в строках указаны переменные (наблюдения), а по горизонтали их характеристики, расчитанные для данного класса: среднее, несмещенное среднеквадратическое отклонение, несмещенная дисперсия.

5. **Members for each cluster & distances.** Выводится столько окон, сколько задано классов. В каждом окне указывается общее число элементов, отнесенных к этому кластеру, в верхней строке указан номер наблюдения (переменной), отнесенной к данному классу и Евклидово расстояние от центра класса до этого наблюдения (переменной). Центр класса – средние величины по всем переменным (наблюдениям) для этого класса.

6. **Save classifications and distances.** Позволяет сохранить в формате программы статистику таблицы, в которой содержатся значения всех переменных, их порядковые номера, номера кластеров, к которым они отнесены, и Евклидовы расстояния от центра кластера до наблюдения. Записанная таблица может быть вызвана любым блоком или подвергнута дальнейшей обработке.

Обычно, когда результаты кластерного анализа методом K-средних получены, можно рассчитать средние для каждого кластера по каждому измерению, чтобы оценить, насколько кластеры различаются друг от друга. В идеале вы должны получить сильно различающиеся средние для большинства, если не для всех измерений, используемых в анализе (в нашем случае (рис. 5.25). Значения переменных пересекаются, но все же мы можем наблюдать достаточно четкие различия кластеров. Для более отчетливой группировки следует сократить число параметров. Значения *F*-статистики, полученные для каждого измерения, являются другим индикатором того, насколько хорошо соответствующее измерение дискриминирует кластеры. Так как у нас решение найдено после одной итерации (меньше чем мы задали), то можно сделать вывод о том, что итоговая конфигурация является искомой.

Глава 6. Многомерные данные: дискриминантный анализ

6.1. Деревья решений

Термины, применяемые в деревьях решений. Методология деревьев решений. Задачи и области применения деревьев решений. Преимущества и недостатки деревьев решений. Методы деревьев решений. Процедура Деревья классификации. Настройка вывода модели. Сохранение результатов. Результаты процедуры – деревья классификации. Проверка модели.

Метод деревьев решений (*decision trees*) является одним из наиболее популярных методов решения задач классификации и прогнозирования. Иногда этот метод *Data Mining* также называют деревьями решающих *правил*, деревьями классификации и регрессии.

Как видно из последнего названия, при помощи данного метода решаются задачи классификации и прогнозирования.

Если зависимая, то есть целевая *переменная* принимает дискретные значения, то при помощи метода дерева решений решается задача классификации.

Если же зависимая *переменная* принимает непрерывные значения, то *дерево* решений устанавливает зависимость этой переменной от независимых переменных: решает задачу численного прогнозирования.

Впервые деревья решений были предложены Ховилендом и Хантом (Hoveland, Hunt) в конце 50-х годов прошлого века. Самая ранняя и известная работа Ханта и др. «Эксперименты в индукции» («Experiments in *Induction*»), в которой излагается суть деревьев решений, была опубликована в 1966 г.

В наиболее простом виде *дерево* решений – это способ представления *правил* в иерархической, последовательной структуре. Основа такой структуры – ответы «Да» или «Нет» на ряд вопросов.

На рис. 6.1 приведен пример дерева решений, задача которого ответить на вопрос: «Играть ли в гольф?». Чтобы решить задачу, то есть принять *решение, играть* ли в гольф, следует отнести текущую ситуацию к одному из известных классов (в данном случае – «играть» или «не играть»). Для этого требуется ответить на ряд вопросов, которые находятся в узлах этого дерева, начиная с его корня.

Первый узел нашего дерева «Солнечно?» является *узлом проверки* – условием. При положительном ответе на вопрос осуществляется переход к левой части дерева, называемой левой ***ветвью,*** при отрицательном – к правой части дерева. Таким образом, ***внутренний узел*** дерева является ***узлом проверки*** определенного условия. Далее идет следующий вопрос

и т. д., пока не будет достигнут ***конечный узел*** дерева, являющийся *узлом решения*. Для нашего дерева существует два типа *конечного узла*: «играть» и «не играть» в гольф.

В результате прохождения от корня дерева (иногда называемого корневой вершиной) до его вершины решается задача классификации: выбирается один из классов – «играть» и «не играть» в гольф.

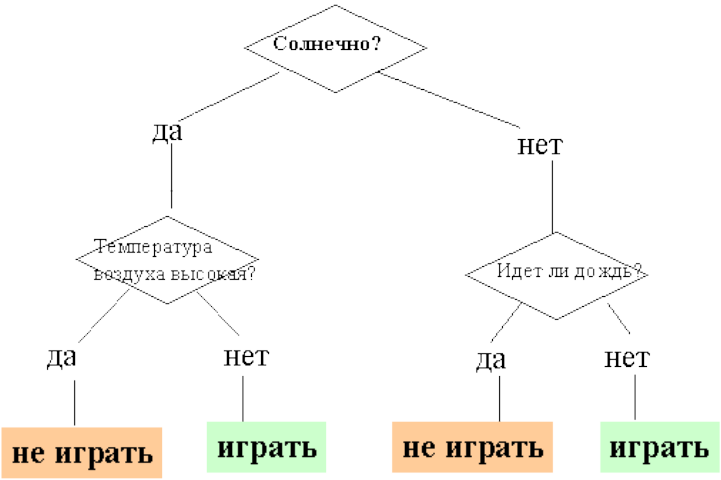


Рис. 6.1. Дерево решений «Играть ли в гольф?»

Целью построения дерева решения в нашем случае является *определение* значения категориальной зависимой переменной.

Итак, для нашей задачи основными элементами дерева решений являются:

Корень дерева: «Солнечно?»

Внутренний узел дерева или *узел проверки*: «Температура воздуха высокая?», «Идет ли дождь?»

Лист, *конечный узел* дерева, *узел решения* или *вершина*: «Играть», «Не играть»

Ветвь дерева (случаи ответа): «Да», «Нет».

В рассмотренном примере решается задача *бинарной классификации*, то есть создается дихотомическая классификационная модель. Пример демонстрирует работу так называемых бинарных деревьев.

В узлах бинарных деревьев *ветвление* может вестись только в двух направлениях: существует возможность только двух ответов на поставленный вопрос («да» и «нет»).

Бинарные деревья являются самым простым, частным случаем деревьев решений. В остальных случаях ответов и, соответственно, *ветвей* дерева, выходящих из его *внутреннего узла*, может быть больше двух.

Каждая *ветвь* дерева, идущая от *внутреннего узла*, отмечена *предикатом расщепления*. Последний может относиться лишь к одному *атрибуту расщепления* данного узла. Характерная особенность *предикатов расщепления*: каждая *запись* использует уникальный *путь* от корня дерева только к одному узлу-решению. Объединенная *информация* об *атрибутах расщепления* и *предикатах расщепления* в узле называется ***критерием расщепления*** (splitting *criterion*).

Таким образом, для данной задачи (как и для любой другой) может быть построено множество деревьев решений различного качества, с различной прогнозирующей точностью.

Качество построенного дерева решения весьма зависит от правильного выбора *критерия расщепления*. Над разработкой и усовершенствованием критериев работают многие исследователи.

Метод деревьев решений часто называют «наивным» подходом. Но благодаря целому ряду преимуществ, данный метод является одним из наиболее популярных для решения задач классификации.

Преимущества деревьев решений

Интуитивность деревьев решений. Классификационная модель, представленная в виде дерева решений, является интуитивной и упрощает понимание решаемой задачи. Результат работы алгоритмов конструирования деревьев решений, в отличие, например, от нейронных сетей, представляющих собой «черные ящики», легко интерпретируется пользователем. Это свойство деревьев решений не только важно при отнесении к определенному классу нового объекта, но и полезно при интерпретации модели классификации в целом. *Дерево* решений позволяет понять и объяснить, почему конкретный *объект* относится к тому или иному классу.

Деревья решений дают возможность извлекать *правила* из *базы данных* на **естественном языке**. Пример *правила*: Если Возраст > 35 и Доход > 200, то выдать *кредит*.

Деревья решений позволяют создавать классификационные модели в тех областях, где аналитику достаточно сложно формализовать знания.

Алгоритм конструирования дерева решений **не требует от пользователя выбора входных атрибутов** (независимых переменных). На *вход алгоритма* можно подавать все существующие атрибуты, *алгоритм* сам выберет наиболее значимые среди них, и только они будут использованы для построения дерева. В сравнении, например, с нейронными сетями, это значительно облегчает пользователю работу, поскольку в нейронных сетях выбор количества входных атрибутов существенно влияет на время обучения.

Точность моделей, созданных при помощи деревьев решений, сопоставима с другими методами построения классификационных моделей (*статистические методы*, нейронные сети).

Разработан ряд **масштабируемых алгоритмов**, которые могут быть использованы для построения деревьев решения на сверхбольших базах данных; *масштабируемость* здесь означает, что с ростом числа примеров или записей *базы данных* время, затрачиваемое на обучение, то есть построение деревьев решений, растет линейно. Примеры таких алгоритмов: SLIQ, SPRINT.

Быстрый процесс обучения. На построение классификационных моделей при помощи алгоритмов конструирования деревьев решений требуется значительно меньше времени, чем, например, на обучение нейронных сетей.

Большинство алгоритмов конструирования деревьев решений имеют возможность специальной обработки **пропущенных значений**.

Многие классические *статистические методы*, при помощи которых решаются задачи классификации, могут работать только с числовыми данными, в то время как деревья решений работают и с числовыми, и с **категориальными** типами данных.

Многие *статистические методы* являются параметрическими, и *пользователь* должен заранее владеть определенной информацией, например, знать вид модели, иметь гипотезу о виде зависимости между переменными, предполагать, какой вид распределения имеют данные. Деревья решений, в отличие от таких методов, строят непараметрические модели. Таким образом, деревья решений способны решать такие задачи *Data Mining*, в которых отсутствует априорная *информация* о виде зависимости между исследуемыми данными.

Процесс конструирования дерева решений. Напомним, что рассматриваемая нами задача классификации относится к стратегии *обучения с учителем*, иногда называемого индуктивным обучением. В этих случаях все объекты тренировочного набора данных заранее отнесены к одному из предопределенных классов.

Алгоритмы конструирования деревьев решений состоят из этапов «построение» или «*создание*» дерева (*tree* building) и «*сокращение*» дерева (*tree pruning*). В ходе *создания* дерева решаются вопросы выбора *критерия расщепления* и остановки обучения (если это предусмотрено алгоритмом). В ходе этапа *сокращения* дерева решается вопрос отсечения некоторых его *ветвей*.

Пример использования дерева решений в ППП Statistica

Предположим, что у вас имеются данные о координатах – Долготе (Longitude) и Широте (Latitude) – для 37 циклонов (табл. 6.1), достигающих силы урагана, по двум классификациям циклонов – Baro и Trop.

Приведенный ниже модельный набор данных использовался для целей иллюстрации в работе Elsner, Lehmiller, Kimberlain (1996), авторы которой исследовали различия между бароклинными и тропическими циклонами в Северной Атлантике.

Таблица 6.1

Данные о координатах для 37 циклонов

N/N	LONGITUD	LATITUDE	CLASS	№/№	LONGITUD	LATITUDE	CLASS
1	2	3	4	5	6	7	8
1	59.00	17.00	BARO	21	66.00	14.00	TROP
2	59.50	21.00	BARO	22	66.00	17.00	TROP
3	60.00	12.00	BARO	23	66.50	17.00	TROP
4	60.50	16.00	BARO	24	66.50	18.00	TROP
5	61.00	13.00	BARO	25	66.50	21.00	TROP
6	61.00	15.00	BARO	26	67.00	14.00	TROP
7	61.50	17.00	BARO	27	67.50	18.00	TROP
8	61.50	19.00	BARO	28	68.00	14.00	BARO
9	62.00	14.00	BARO	29	68.50	18.00	BARO
10	63.00	15.00	TROP	30	69.00	13.00	BARO
11	63.50	19.00	TROP	31	69.00	15.00	BARO
12	64.00	12.00	TROP	32	69.50	17.00	BARO
13	64.50	16.00	TROP	33	69.50	19.00	BARO
14	65.00	12.00	TROP	34	70.00	12.00	BARO
15	65.00	15.00	TROP	35	70.50	16.00	BARO
16	65.00	17.00	TROP	36	71.00	17.00	BARO
17	65.50	16.00	TROP	37	71.50	21.00	BARO
18	65.50	19.00	TROP				
19	65.50	21.00	TROP				
20	66.00	13.00	TROP				

Откроем модуль **Classification Tree**.

Введем или откроем файл с исходными данными.

Выполним команду *Analysis/Resume Analysis*, на экране появится окно, изображенное на рис. 6.2.

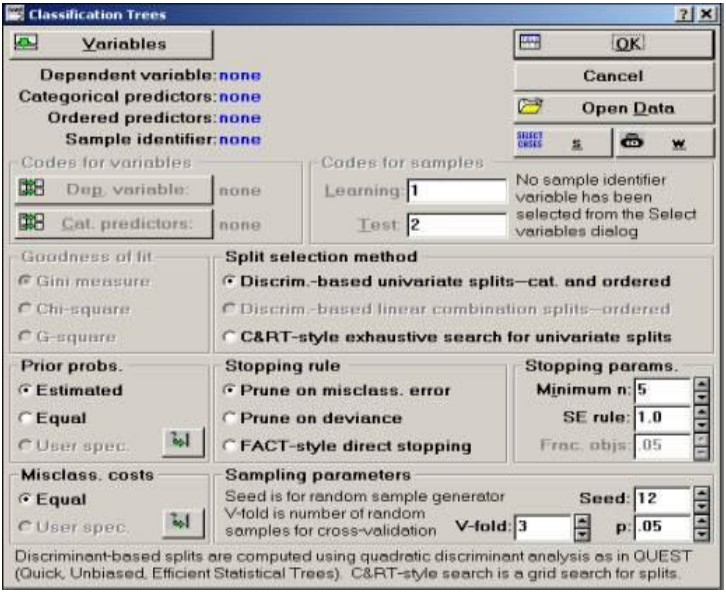


Рис. 6.2. Результат выполнения команды *Analysis/Resume Analysis*

В разделе **Split selection method** выберем метод **CART (C&RT-style);**

Нажмите кнопку **Variables** и выделите переменные, как показано на рисунке, и нажмите **OK**. Появится окно (рис. 6.3).

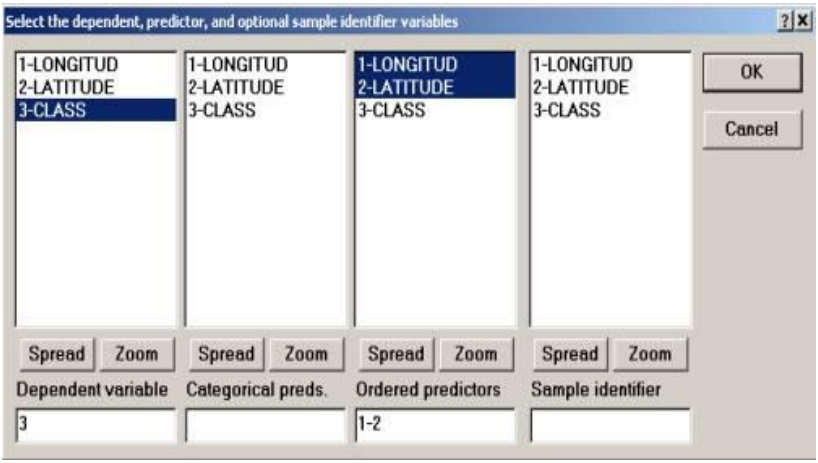


Рис. 6.3. Окно выбора переменных

В разделе **Codes for variables** нажмите кнопку **Dep. Variable**, а в появившемся окне – соответственно кнопку **All,** а затем **OK** (рис. 6.4).

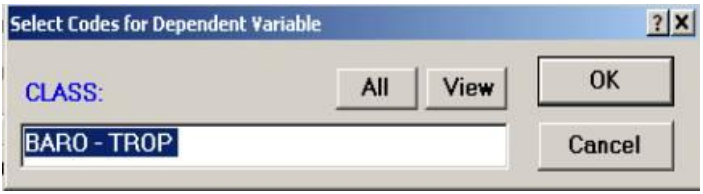


Рис. 6.4. Окно выбора взаимосвязи переменных

В итоге должны получить следующее окно (рис. 6.5).

Рис. 6.5. Результат выполнения команды *Analysis/Resume Analysis* после задания переменных

Нажав **OK**, получим окно (рис. 6.6).

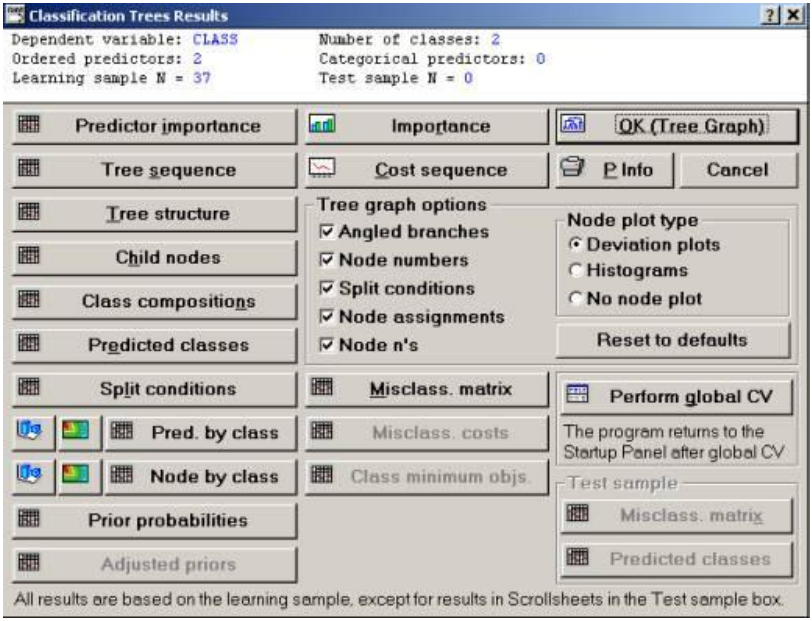


Рис. 6.6. Результат выполнения алгоритма деревьев решений

В открытом диалоговом окне поочередно будем нажимать кнопки для вывода результатов и анализа данных: (**OK (Tree Graph), Tree structure, Importance**) и получать соответствующие картинки.

«Дерево классификации» для переменной Class, использующее опцию **Полный перебор деревьев с одномерным ветвлением по методу CART**, сумело правильно классифицировать все 37 циклонов. «Граф дерева» для этого «дерева классификации» показан на рис. 6.7.

В заголовке «графа» приведена общая информация, согласно которой полученное «дерево классификации» имеет 2 ветвления и 3 терминальные вершины. Терминальные вершины (или, как их иногда называют, листья) – это узлы «дерева», начиная с которых никакие решения больше не принимаются. На рис. 6.7 терминальные вершины показаны красными пунктирными линиями, а остальные – так называемые решающие вершины или вершины ветвления – сплошными черными линиями. Началом «дерева» считается самая верхняя решающая вершина, которую иногда также называют корнем «дерева». На рисунке она расположена в левом верхнем углу и помечена цифрой 1. Первоначально все 37 циклонов приписываются к этой корневой вершине и предварительно классифицируются как *Baro* – на это указывает надпись *Baro* в правом верхнем углу вершины. Класс *Baro* был выбран для начальной классификации потому, что число циклонов *Baro* немного больше, чем циклонов *Trop*

(см. *гистограмму*, изображенную внутри корневой вершины). В левом верхнем углу «графа» имеется надпись – легенда, указывающая, какие столбики гистограммы вершины соответствуют циклонам *Baro* и *Trop*.

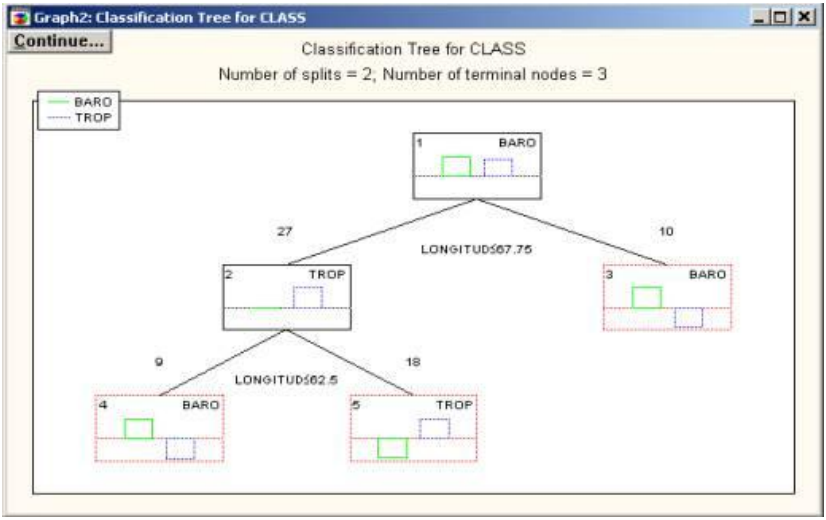


Рис. 6.7. Граф дерева классификации

Корневая вершина разветвляется на две новых вершины. Под корневой вершиной имеется текст, описывающий схему данного ветвления. Из него следует, что циклоны, имеющие значение Долготы, меньшее или равное 67.75, отнесены к вершине номер 2 и предположительно классифицированы как *Trop*, а циклоны с Долготой, большей 67.75, приписаны к вершине 3 и классифицированы как *Baro*. Числа 27 и 10 над вершинами 2 и 3 соответственно обозначают число наблюдений, попавших в эти две дочерние вершины из родительской корневой вершины. Затем точно так же разветвляется вершина 2. В результате 9 циклонов со значениями Долготы, меньшими или равными 62.5, приписываются к вершине 4 и классифицируются как *Baro*, а остальные 18 циклонов с Долготой, большей 62.5, – к вершине 5 и классифицируются как *Trop*.

На «графе дерева» вся эта информация представлена в простом, удобном для восприятия виде. Если теперь мы посмотрим на гистограммы терминальных вершин «дерева», расположенных в нижней строке, то увидим, что «дерево классификации» сумело абсолютно правильно расклассифицировать циклоны. Каждая из терминальных вершин «чистая», то есть не содержит неправильно классифицированных наблюдений. Вся информация, содержащаяся в «графе дерева», продублирована в таблице результатов *Структура дерева*, которая приведена на рис. 6.8.

Tree Structure (barotrop.sta)

Continue...

Child nodes, observed class n's, predicted class, and split condition for each node

Node	Left branch	Right branch	n in cls BARO	n in cls TROP	Predict. class	Split constant	Split variable
1	2	3	19	18	BARO	-67.7500	LONGITUD
2	4	5	9	18	TROP	-62.5000	LONGITUD
3			10	0	BARO		
4			9	0	BARO		
5			0	18	TROP		

Рис. 6.8. Вид таблицы результатов *Структура дерева*

Следует обратить внимание на то, что в таблице результатов вершины с 3-й по 5-ю помечены как терминальные, так как в них не происходит ветвления. Обратите также внимание на знаки *постоянных ветвления*, например, –67.75 для вершины 1.

Если делаются одномерные ветвления, то каждой предикторной переменной можно приписать ранг по шкале от 0 до 100 в зависимости от степени ее влияния на отклик зависимой переменной. В нашем примере очевидно, что Долгота (*Longitude*) имеет большую важность, а Широта (Latitude) – относительно небольшую (рис. 6.9).

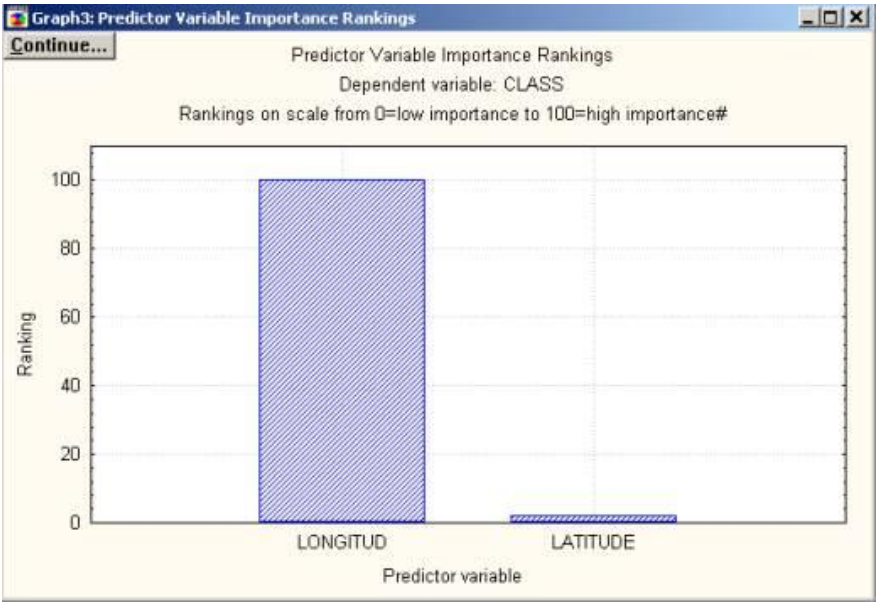


Рис. 6.9. Диаграмма рангов каждой предикторной переменной

6.2. Метод опорных векторов

Свойство метода опорных векторов. Линейный метод опорных векторов. Нелинейный классификатор. Свойства. Параметр выбора. Обобщенный метод опорных векторов в случае существование множества классов (мультиклассовый).

Машина опорных векторов (SVM, Support Vector Machine) является одним из наиболее популярных и универсальных алгоритмов машинного обучения. Данный алгоритм может применяться как для решения задач классификации, так и для восстановления регрессии.

Классификация, в частности, имеет довольно широкое применение: от распознавания образов или создания спам-фильтров до вычисления распределения горячих алюминиевых частиц в ракетных выхлопах.

Задача классификации состоит в определении к какому классу, как минимум двух изначально известных относится данный объект. Обычно таким объектом является вектор в *n*-мерном вещественном пространстве $\mathbb{R}^n$. Координаты вектора описывают отдельные атрибуты объекта. Например, цвет ***c***, заданный в модели RGB, является вектором в трехмерном пространстве: ***c***=(*red, green, blue*).

Если классов всего два («болен / здоров», «подействует препарат / не подействует препарат», «красное / черное»), то задача называется бинарной классификацией. Если классов несколько – многоклассовая (мультиклассовая) классификация. Также могут иметься образцы каждого класса – объекты, про которые заранее известно, к какому классу они принадлежат. Такие задачи называют *обучением с учителем*, а известные данные называются *обучающей выборкой*. (*Примечание.* Если классы изначально не заданы, то перед нами *задача кластеризации.*)

Метод опорных векторов, рассматриваемый в данном разделе, относится к обучению с учителем.

Таким образом, математическая формулировка задачи классификации такова:

пусть **X** – пространство объектов (например, $\mathbb{R}^n$),

Y – наши классы (например, ***Y*** = {–1,1}).

Дана обучающая выборка: (x_1, y_1), …, (x_m, y_m). Требуется построить функцию $F : X \to Y$ (классификатор), сопоставляющий класс ***y*** произвольному объекту ***x.***

Метод опорных векторов изначально относится к бинарным классификаторам, хотя существуют способы заставить его работать и для задач мультиклассификации.

Идею метода удобно проиллюстрировать на следующем простом примере: даны точки на плоскости, разбитые на два класса (рис. 6.10). Проведем линию, разделяющую эти два класса (красная линия на

рис. 6.10). Далее все новые точки (не из обучающей выборки) автоматически классифицируются следующим образом:

точка выше прямой попадает в класс **A**,

точка ниже прямой – в класс **B**.

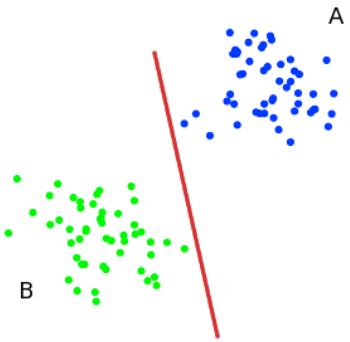


Рис. 6.10. Пример разбития точек плоскости на два класса

Такую прямую назовем разделяющей прямой. Однако, в пространствах высоких размерностей прямая уже не будет разделять наши классы, так как понятие «ниже прямой» или «выше прямой» теряет всякий смысл. Поэтому вместо прямых необходимо рассматривать гиперплоскости – пространства, размерность которых на единицу меньше, чем размерность исходного пространства. В нашем примере гиперплоскость – это обычная двумерная плоскость.

Следует иметь в виду, что существует несколько прямых, разделяющих два класса (рис. 6.11).

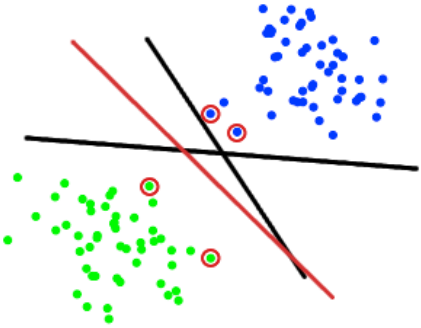


Рис. 6.11. Семейство прямых, разбивающих точки плоскости на два класса

С точки зрения точности классификации, лучше всего выбрать прямую, расстояние от которой до каждого класса максимально. Другими словами, выберем ту прямую, которая разделяет классы наилучшим образом (красная прямая на рис. 6.11). Такая прямая, а в общем случае – гиперплоскость, называется оптимальной разделяющей гиперплоскостью.

Вектора (точки в двумерном пространстве), лежащие ближе всех к разделяющей гиперплоскости, называются опорными векторами (support vectors). На рис. 6.11 они помечены красным.

На практике случаи, когда данные можно разделить гиперплоскостью, или, как еще говорят, *линейно*, довольно редки. Пример линейной неразделимости можно видеть на рис. 6.12.

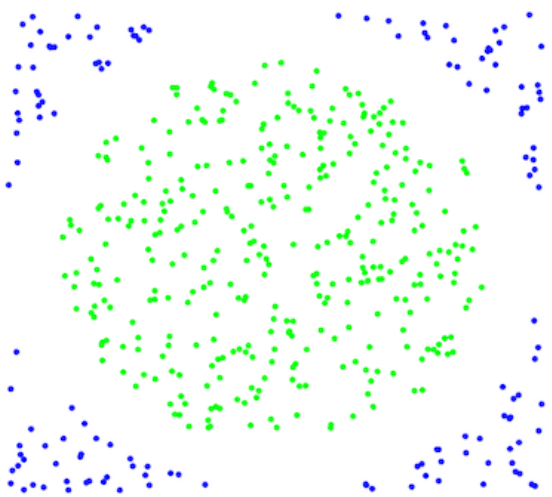


Рис. 6.12. Пример линейной неразделимости точек на плоскости

В этом случае поступают так: все элементы обучающей выборки вкладываются в пространство **X** более высокой размерности с помощью специального отображения $\varphi : \mathbb{R}^n \to X$. При этом отображение φ выбирается так, чтобы в новом пространстве **X** выборка была *линейно* разделима.

Классифицирующая функция F принимает вид $F(\mathbf{x}) = \operatorname{sign}\left(\langle \mathbf{w}, \varphi(\mathbf{x}) \rangle + b\right)$. Выражение $\mathbf{k}(\mathbf{x}, \mathbf{x}') = \langle\, \varphi(\mathbf{x}), \varphi(\mathbf{x}')\, \rangle$ называется *ядром* классификатора. С математической точки зрения, ядром может служить любая положительно определенная симметричная функция двух переменных. Положительная определенность необходимо для того, чтобы соответствующая функция Лагранжа в задаче оптимизации была ограничена снизу, то есть задача оптимизации была бы корректно определена. Точность классификатора зависит от выбора ядра.

Чаще всего на практике встречаются следующие ядра:

Полиномиальное: $\mathbf{k}(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + \text{const})^d$

Радиальная базисная функция: $\mathbf{k}(\mathbf{x}, \mathbf{x}') = e^{-\gamma ||\mathbf{x} - \mathbf{x}'||^2}, \gamma > 0.$

Гауссова радиальная базисная функция: $\mathbf{k}(\mathbf{x}, \mathbf{x}') = e^{-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}}.$

Сигмоид: $\mathbf{k}(\mathbf{x}, \mathbf{x}') = \tanh(\kappa \langle \mathbf{x}, \mathbf{x}' \rangle + c), \ \kappa > 0, c < 0.$

Пример использования линейного классификатора в ППП Statistica

Для простоты, разберем двумерный случай с линейным разделением точек. Имеются данные о **типах** ириса.

По **длине стебля** и **ширине цветка** необходимо **определить тип** ириса.

Структура данных представлена на рис. 6.13.

	1 SLENGTH	2 PWIDTH	3 FLOWER
37	5,5	0,2	Setosa
38	4,9	0,1	Setosa
39	4,4	0,2	Setosa
40	5,1	0,2	Setosa
41	5,0	0,3	Setosa
42	4,5	0,3	Setosa
43	4,4	0,2	Setosa
44	5,0	0,6	Setosa
45	5,1	0,4	Setosa
46	4,8	0,3	Setosa
47	5,1	0,2	Setosa
48	4,6	0,2	Setosa
49	5,3	0,2	Setosa
50	5,0	0,2	Setosa
51	7,0	1,4	Versicol
52	6,4	1,5	Versicol
53	6,9	1,5	Versicol
54	5,5	1,3	Versicol
55	6,5	1,5	Versicol
56	5,7	1,3	Versicol
57	6,3	1,6	Versicol
58	4,9	1,0	Versicol
59	6,6	1,3	Versicol
60	5,2	1,4	Versicol
61	5,0	1,0	Versicol
62	5,9	1,5	Versicol

Рис. 6.13. Вид таблицы данных в Statistica

Переменные:

- SLENGTH – длина стебля,
- PWIDTH – ширина цветка,
- FLOWER – тип ириса.

В результате анализа было найдено 18 векторов (по 9 для каждого класса).

```
Dataset IrisSNN:
  Dependent: FLOWER
  Independents: SLENGTH, PWIDTH
  Sample size = 75 (Train), 25 (Test), 100 (Overall)

Support Vector machine results:
  SVM type: Classification type 1 (capacity=1,000)
  Kernel type: Linear
  Number of support vectors = 18 (18 bounded)
  Cross-validation accuracy (%) = 100,000
  Support vectors per class: 9 (Setosa), 9 (Versicol),

  Class. accuracy (%) = 100,000(Train), 100,000(Test),
```

Рис. 6.14. Результат выполнения алгоритма

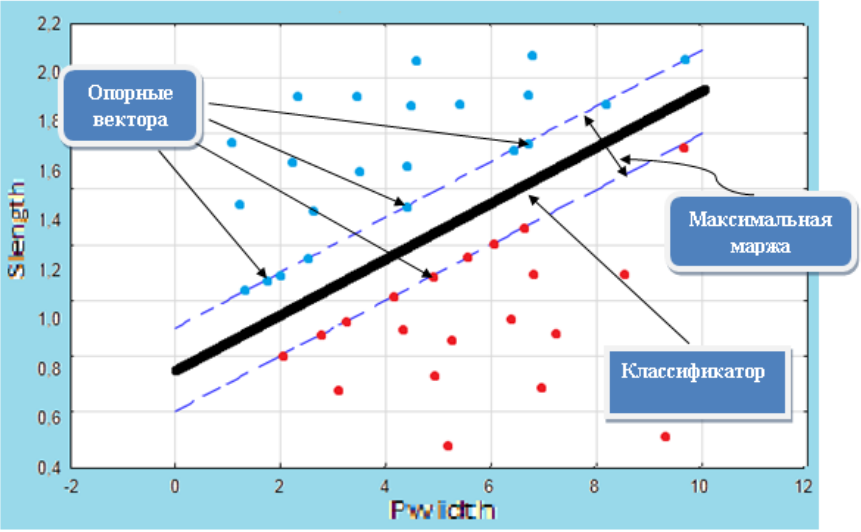


Рис. 6.15. Графическое представление работы алгоритма метода опорных векторов в Statistica

Максимальная маржа составила 1,722.

	SVM model specifications (decision constants) SVM: Classification type 1 (C=1,000000) Kernel: Linear
Decision Constant	**Value**
1	-1,72201

Рис. 6.16. Вид таблицы с выводом значения максимально маржи

Пример обобщенного метода опорных векторов в случае существования множества классов

Данные о трех типах цветков представлены в виде таблицы в файле *Irisdat.sta* из папки примеров *STATISTICA*.

	Fisher (1936) iris data: length & width of sepals and petals, 3 types of Iris				
	1 SEPALLEN	2 SEPALWID	3 PETALLEN	4 PETALWID	5 IRISTYPE
1	5,0	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	VERSICO
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	VERSICO
9	5,9	3,2	4,8	1,8	VERSICO
10	4,6	3,6	1,0	0,2	SETOSA
11	6,1	3,0	4,6	1,4	VERSICO
12	6,0	2,7	5,1	1,6	VERSICO
13	6,5	3,0	5,2	2,0	VIRGINIC
14	5,6	2,5	3,9	1,1	VERSICO
15	6,5	3,0	5,5	1,8	VIRGINIC
16	5,8	2,7	5,1	1,9	VIRGINIC
17	6,8	3,2	5,9	2,3	VIRGINIC
18	5,1	3,3	1,7	0,5	SETOSA
19	5,7	2,8	4,5	1,3	VERSICO

Рис. 6.17. Вид таблицы данных в Statistica

В качестве непрерывных переменных выступают:

- SEPALLEN – длина чашелистика,
- SEPALWID – ширина чашелистика,
- PETALWID – ширина лепестка,
- PETALLEN – длина лепестка.

В качестве категориальной переменной *Iristype* (тип ириса).
Выберем переменные для анализа как показано на рис. 6.18.

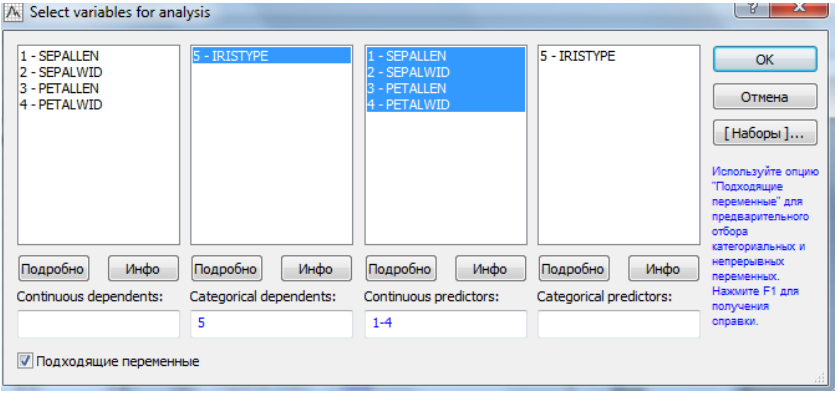


Рис. 6.18. Окно выбора переменных для анализа

В качестве зависимой переменной выступает категориальная переменная – тип ириса. В качестве независимых – все остальные.

Включим поиск по сетке констант как в пункте *Регрессия*.

В результате анализа было найдено 40 векторов.

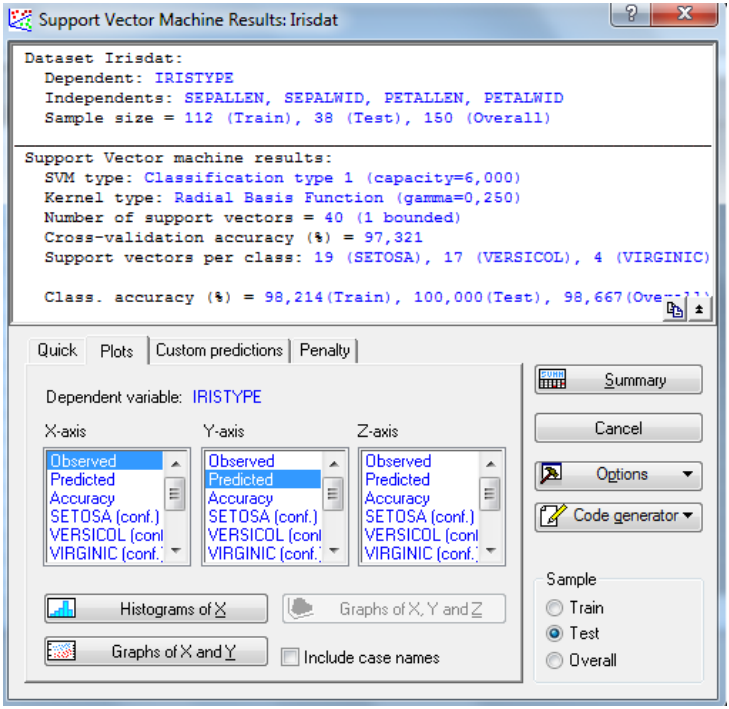


Рис. 6.19. Результат выполнения алгоритма

Case Name	IRISTYPE Dependent	IRISTYPE Predicted	IRISTYPE Accuracy	SETOSA Confidence	VERSICOL Confidence	VIRGINIC Confidence
1	SETOSA	SETOSA	Correct	0,666667	0,333333	0,000000
2	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
4	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
9	VERSICOL	VERSICOL	Correct	0,000000	0,666667	0,333333
16	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
21	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
22	VERSICOL	VERSICOL	Correct	0,000000	0,666667	0,333333
27	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
30	VERSICOL	VERSICOL	Correct	0,000000	0,666667	0,333333
34	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
35	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
41	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
42	SETOSA	SETOSA	Correct	0,666667	0,333333	0,000000
45	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
53	SETOSA	SETOSA	Correct	0,666667	0,333333	0,000000
54	SETOSA	SETOSA	Correct	0,666667	0,333333	0,000000
77	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
78	SETOSA	SETOSA	Correct	0,666667	0,333333	0,000000
82	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
86	VERSICOL	VERSICOL	Correct	0,000000	0,666667	0,333333
89	VIRGINIC	VIRGINIC	Correct	0,000000	0,333333	0,666667
98	VERSICOL	VERSICOL	Correct	0,000000	0,666667	0,333333
100	SETOSA	SETOSA	Correct	0,666667	0,333333	0,000000
101	SETOSA	SETOSA	Correct	0,666667	0,333333	0,000000
112	SETOSA	SETOSA	Correct	0,666667	0,333333	0,000000
114	VERSICOL	VERSICOL	Correct	0,000000	0,666667	0,333333
115	SETOSA	SETOSA	Correct	0,666667	0,333333	0,000000

Рис. 6.20. Вид таблицы с результатами выполнения алгоритма

Построим диаграммы рассеяния контрольной выборки: по оси X – независимая переменная, по оси Y –SETOSA, VERSICOL, VIRGIN.

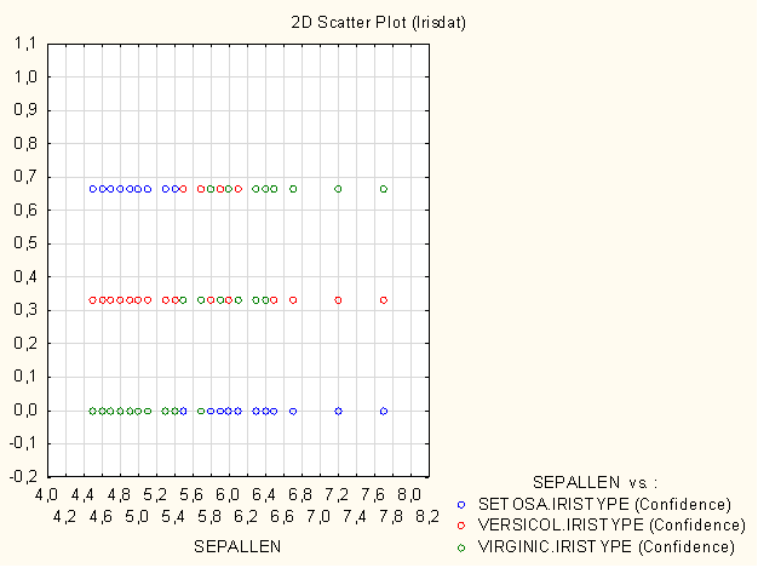


Рис. 6.21. Диаграмма рассеяния контрольной выборки в осях «длина чашелистика / тип ириса»

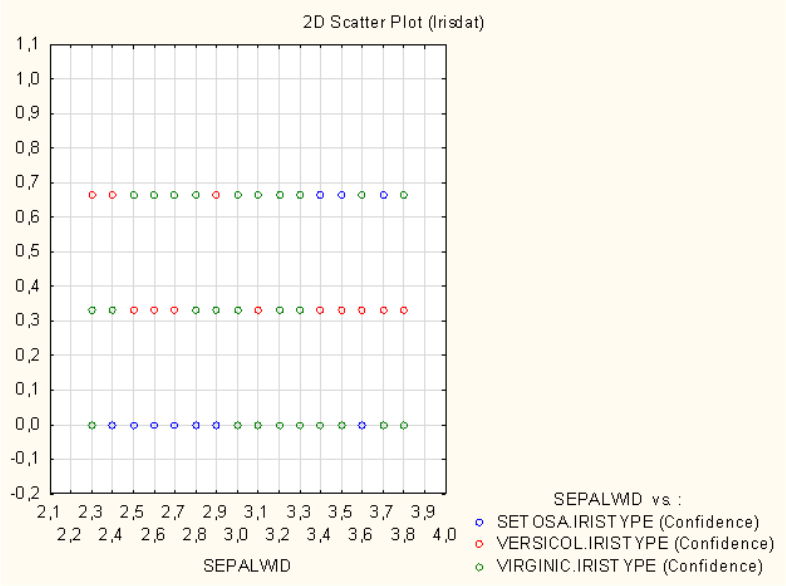


Рис. 6.22. Диаграмма рассеяния контрольной выборки в осях «ширина чашелистика / тип ириса»

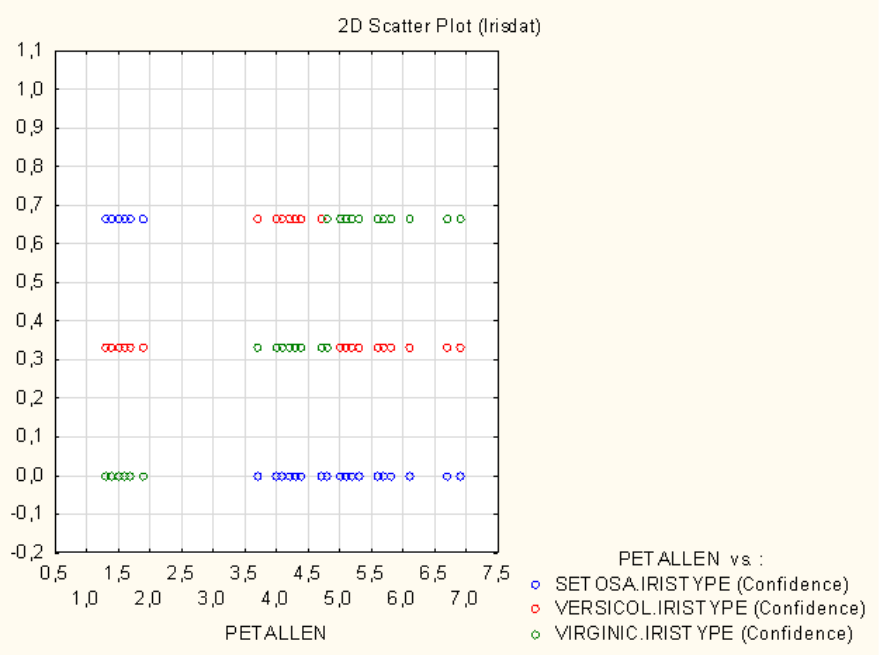


Рис. 6.23. Диаграмма рассеяния контрольной выборки в осях «длина лепестка / тип ириса»

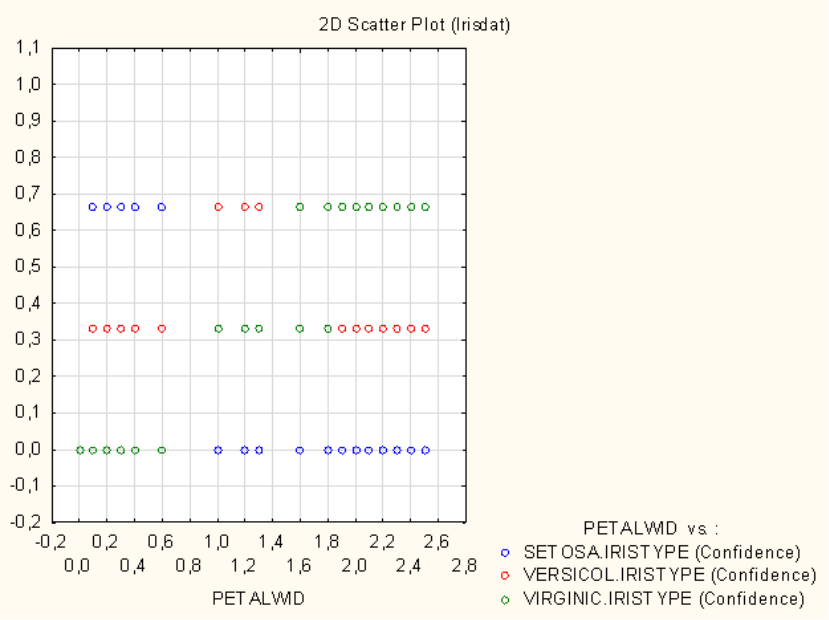


Рис. 6.24. Диаграмма рассеяния контрольной выборки в осях «ширина лепестка / тип ириса»

На контрольной выборке метод опорных векторов проявил себя великолепно – 100 % точность!

ЛИТЕРАТУРА

Основная:

1. *Боровиков В. П.* Популярное введение в современный анализ данных в системе STATISTICA: учеб. пособие для вузов. Москва: РиС, 2015. – 288 с.

2. *Горяинова Е. Р.* Прикладные методы анализа статистических данных: учебное пособие / Е. Р. Горяинова, А. Р. Панков, Е. Н. Платонов. – Москва: ИД ГУ ВШЭ, 2012. – 310 с.

3. *Лесковец Ю.* Анализ больших наборов данных / Ю. Лесковец, А. Раджараман. – Москва: ДМК, 2016. – 498 с.

4. *Крянев А. В.* Метрический анализ и обработка данных / А. В. Крянев, Г. В. Лукин, Д. К. Удумян. – Москва: Физматлит, 2012. – 308 с.

5. *Кулаичев А. П.* Методы и средства комплексного анализа данных: учебное пособие. – Москва: Форум, НИЦ ИНФРА-М, 2013. – 512 с.

6. Медицинская статистика понятным языком / Ашис Банержи. – Москва: Практ. медицина, 2014.

Дополнительная:

7. *Халафян А. А.* Учебник STATISTICA 6. Статистический анализ данных – 3-е изд. – Москва: Бином, 2009. – 512 с.

8. *Новикова Н. М.* Статистические методы в биологии и медицине: учеб.-метод. пособие. – Минск: МГЭУ им. А. Д. Сахарова, 2009. – 88 с.

9. *Кабаков Р.* R в действии. Анализ и визуализация данных в программе R. – Москва: ДМК, 2016. – 588 с.

10. Биометрика – журнал для медиков и биологов, сторонников доказательной биомедицины [Электронный ресурс]. URL: http://www.biometrica.tomsk.ru.

11. *Орлов А. И.* Прикладная статистика: учебник. – Москва: Изд-во «Экзамен», 2004. – 656 с.

12. *Чиркин А. А.* Клинический анализ лабораторных данных. – Витебск: Мед. лит-ра, 2008. – 384 с.

13. *Петри А.* Наглядная статистика в медицине / А. Петри, К. Сэбин. – Москва: Издательский дом ГЭОТАР-МЕД, 2003 – 143 с.

Полнотекстовые базы данных

Современные профессиональные базы данных, информационные, справочные и поисковые системы:

Aquatic Conservation, Biodiversity and Conservation, Ecological Research, Ecosystems, Ecotoxicology, Environmental and Ecological Statistics,

Environmental International, Environmental Health, Environmental Management, Environmental Manager, Environmental Monitoring and Assessment, Environmental Pollution, Environmental Science and Technology, Environmetrics, European Environment, European Journal of Forest Research, Evolutionary Ecology, Journal of Environmental Monitoring, Journal of Chemical Ecology, Journal of Health and Place, Journal of Plant Research, Land Degradation and Rehabilitation, Landscape and Ecological Engineering, Landscape and Urban Planing, Naturwissenschaften, Population Ecology, Urban Ecosystems.

Интернет-ресурсы

Электронный учебник по пакету Statistica: http://www.statsoft.ru/home/textbook/esc.html

Учебное издание

Сыса Алексей Григорьевич,
Живицкая Елена Петровна

СТАТИСТИЧЕСКИЙ АНАЛИЗ В БИОЛОГИИ И МЕДИЦИНЕ

Учебно-методическое пособие

Редактор *Л. М. Кореневская*
Компьютерная верстка *Д. В. Головач*
Технический редактор *А. В. Красуцкая*

Подписано в печать 05.02.2018. Формат 60×90 1/16.
Бумага офсетная. Печать офсетная.
Усл. печ. л. 8,75. Уч.-изд. л. 4,10.
Тираж 100 экз. Заказ № 82.

Республиканское унитарное предприятие «Информационно-вычислительный центр Министерства финансов Республики Беларусь».
Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий
№ 1/161 от 27.01.2014, № 2/41 от 29.01.2014.
Ул. Кальварийская, 17, 220004, г. Минск.