А.П. КУЛАИЧЕВ

МЕТОДЫ И СРЕДСТВА КОМПЛЕКСНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ

УЧЕБНОЕ ПОСОБИЕ

переработанное и дополненное

Допущено
Учебно-методическим объединением
по классическому университетскому образованию
в качестве учебного пособия для вузов
по дисциплинам «Математическая статистика» и «Информатика»

Москва

2017

УДК 519.2(075.8)
ББК 22.172я73
К90

Кулаичев А.П.

К90 Методы и средства комплексного статистического анализа данных : учеб. пособие / А.П. Кулаичев. — . — М. — 484 с. — (Высшее образование: Бакалавриат).

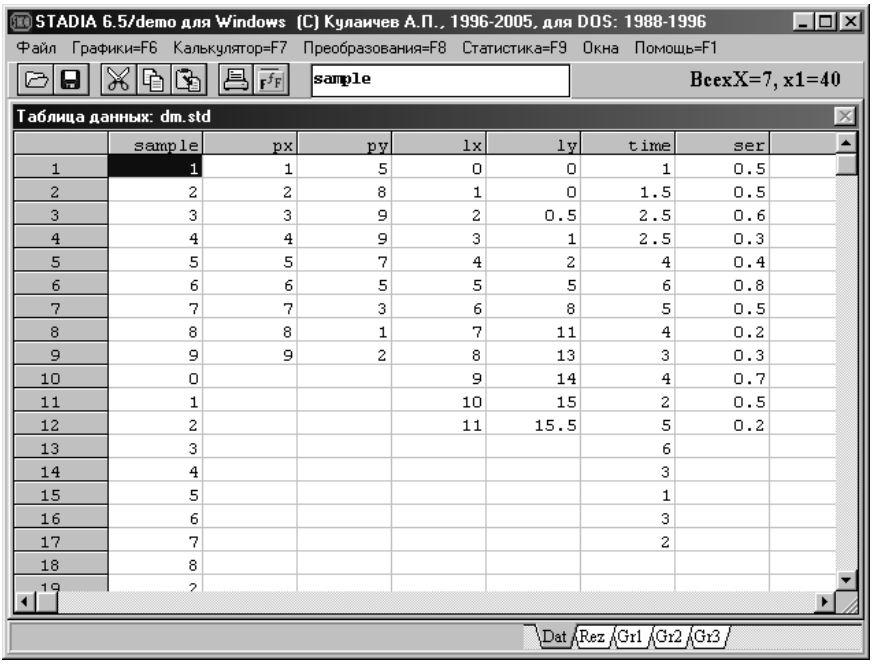
Книга представляет собой обобщающее учебное пособие по математической статистике. В конспективной и доступной форме, с использованием наглядных предметных примеров из различных областей приложения рассмотрены все основные статистические разделы, понятия, методы и средства анализа данных на компьютере.

Обучающим инструментом для практического освоения излагаемых методов является универсальный российский статистический пакет STADIA, ставший в данной области своеобразным стандартом де-факто.

Для студентов, аспирантов и преподавателей вузов, а также для специалистов разного профиля, связанных с анализом информации в различных областях науки, техники, производства, медицины, управления, планирования, экономики, бизнеса и др.

УДК 519.2(075.8)
ББК 22.172я73

О Г Л А В Л Е Н И Е



	sample	px	py	lx	ly	time	ser
1	1	1	5	0	0	1	0.5
2	2	2	8	1	0	1.5	0.5
3	3	3	9	2	0.5	2.5	0.6
4	4	4	9	3	1	2.5	0.3
5	5	5	7	4	2	4	0.4
6	6	6	5	5	5	6	0.8
7	7	7	3	6	8	5	0.5
8	8	8	1	7	11	4	0.2
9	9	9	2	8	13	3	0.3
10	0			9	14	4	0.7
11	1			10	15	2	0.5
12	2			11	15.5	5	0.2
13	3					6	
14	4					3	
15	5					1	
16	6					3	
17	7					2	
18	8						
19	9						

Рис. 2.1. Экран статистической системы с электронной таблицей

Запуск системы и конец сеанса. Чтобы начать сеанс работы, нужно из директории STADIA запустить одноименное приложение. На экране монитора появляется типичный для приложений Windows экран системы, изображенный на рис. 2.1.

Чтобы закончить сеанс работы, следует нажать кнопку закрытия головного окна, или клавишу [F10], или выполнить подпункт «*Выход*» из пункта «*Файлы*» в верхней командной строке.

Помощь и подсказки. Чтобы вспомнить назначение клавиш и полей экранного меню, переведите на них указатель мыши и прочтите подсказку в нижней строке экрана.

Для получения более подробной экранной помощи по текущему контексту следует вызвать экранный справочник нажатием клавиши [F1] или из пункта «*Помощь*» верхней командной строки.

Составные компоненты

Основные архитектурные компоненты статистической системы вызываются из верхней командной строки или нажатием *быстрых* клавиш [F1] — [F10] (что более быстро и надежно).

1. Электронная таблица является центральным компонентом и предназначена для ввода, хранения, просмотра и редактирования исходных

данных. В этой таблице данные представляются в виде матрицы или вектора, где столбцы соответствуют переменным или выборкам, а строки — значениям переменных, измерениям или объектам. Элементы таблицы могут содержать как числовые, так и символьные (*номинальные*) значения.

Возможности работы с электронной таблицей становятся доступны при активизации страницы [*Dat*] . Доступный объем электронной таблицы определен конкретной модификацией статпакета (см. выше).

2. Файловая подсистема обеспечивает ввод различного типа информации из дисковых файлов и запись данных на диск, вызывается по нажатию клавиш [F3] , [F4] или из выкидного меню пункта «*Файлы*» верхней командной строки, а также по нажатию левых кнопок в инструментальной строке экрана. По сравнению с типичным для Windows «*Open–Close*»–диалогом здесь произведен ряд существенных усовершенствований (см. разд. 3.2).

3. Блок преобразований предназначен для выполнения различных преобразований над исходными данными (алгебраические, логические, матричные, комбинаторные и другие преобразования) и вызов его производится по нажатию клавиши [F8] или же по выполнению пункта «*Преобразования*» из верхней командной строки.

4. Калькулятор обеспечивает оперативное выполнение различных вспомогательных вычислений по вводимым выражениям и вызывается по нажатию клавиши [F7] или же по выполнению пункта «*Вычисления*» из верхней командной строки.

5. Графопостроитель обеспечивает построение различных графиков исходных данных по нажатию клавиши [F6] или же по выполнению пункта «*График*» из верхней командной строки.

6. Графический редактор позволяет редактировать графики данных и графики результатов анализа, сохранять их в дисковых файлах или выводить на печать. Возможности редактирования графиков в виде инструментальной линейки кнопок становятся доступными при активизации любой из страниц графиков ⌊*Gr*$_i$⌋, i = 1–15.

7. Блок статистики содержит набор процедур, реализующих вычисления по наиболее употребительным статистическим методам и вызов его меню производится по нажатию клавиши [F9] или же по выполнению пункта «*Статистика*» из верхней командной строки.

8. Текстовый редактор результатов позволяет просматривать и редактировать выдачу числовых результатов анализа, сохранять ее в дисковом файле, выдавать на печать, переносить во внешние пакеты. Текстовый редактор становится доступным при активизации страницы результатов анализа ⌊*Rez*⌋.

9. Экранный справочник состоит из серии разделов с сетью перекрестных электронных ссылок и содержит описание всех операций и математических методов, а его вызов производится по нажатию клавиши

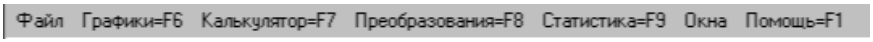
2.2. Порядок диалога

Экранное пространство (см. рис. 2.1) включает следующие типичные для Windows-приложений постоянные строки или панели.

1. Строка заголовка.

В этой строке располагаются также кнопки закрытия, свертывания и расширения головного окна.

2. Командная строка. Вторая *командная строка* содержит команды и выкидные списки команд на выполнение основных операций, часть из которых дублируются инструментальными кнопками и «горячими» клавишами быстрого вызова:



3. Строка кнопок. Третья *строка инструментальных кнопок* с пиктограммами дублирует наиболее часто выполняемые операции:

В этой строке располагается окно редактирования текущей ячейки электронной таблицы, указывается число определенных переменных и размер текущей переменной, а также присутствуют постоянные кнопки ряда операций, одинаково работающие для любой активной страницы, а именно:

- чтение и запись содержимого активной страницы (см. разд. 3.2);

- операции с буфером обмена: вырезание, копирование и вставка (см. разд. 2.5);

- выдача на печать содержимого активной страницы;

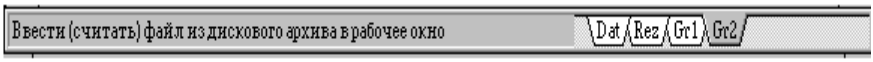
- изменение шрифта активной страницы (см. ниже «Настройки»).

При активизации страницы электронной таблицы в этой строке появляется также и поле редактирования содержимого текущей ячейки таблицы, а также указание числа переменных в таблице *BcexX*= и числа значений текущей переменной *Xi*=.

4. Рабочее окно.

Основная часть экрана составляет рабочее пространство пользователя, в котором он и проводит большую часть времени: вводит данные, обдумывает результаты анализа и просматривает графики и т. п. Это пространство организовано в форме страничной картотеки (см. ниже).

5. Строка подсказки и закладок. Последняя строка экспонирует оперативные подсказки к пунктам и полям ввода различных меню:

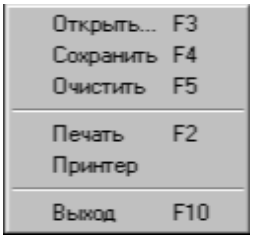


Подсказка появляется, если пару секунд остановить мышь на некотором экранном элементе.

Правая часть этой строки отведена для ярлыков–закладок страниц картотеки (см. ниже).

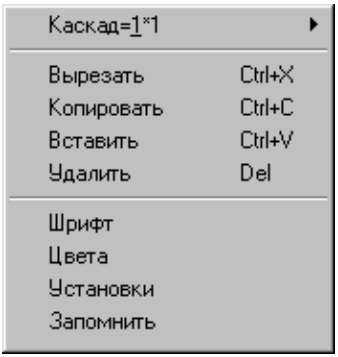
Командная строка

Система выкидных меню верхней командной строки включает три разворачивающихся пункта: файлы, окна и помощь. Другие пункты верхней командной строки невыкидные и их исполнение сразу приводит к вызову соответствующих макрокомпонентов и реализующих их диалогов.



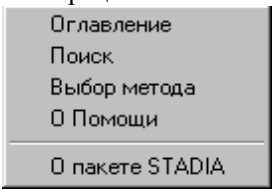
Файлы. Выкидное меню пункта «*Файлы*» включает операции чтения, записи, очистки, печати текущей страницы, настройки принтера (по стандартному Windows-диалогу [29]) и выхода из STADIA. Работа этих пунктов дублируется быстрыми функциональными клавишами F2 — F5 , F9

Окна. Выкидное меню пункта «*Окна*» включает следующие операции:



- «*Каскад*» или набор вариантов совместной визуализации страниц (рис. 2.2 и пояснения к нему);
- операции с буфером обмена (см. разд. 2.5): вырезать, копировать, вставить, удалить (действующие одинаково для любой активной страницы);
- выполнение различных настроек (см. ниже) и их запоминание для следующих сеансов работы;
- список–переключатель доступных экранных страниц.

Помощь. Выкидное меню пункта «*Помощь*» включает следующие операции:



- переход в оглавление экранного справочника;
- поиск задаваемого контекста в справочнике;
- «*Выбор метода*» — позволяет подобрать для имеющихся данных подходящий метод анализа;
- «*О помощи*» — знакомит со справочником;
- «*О STADIA*» — выдает краткие сведения о пакете и условиях поставки.

Универсальные операции. Операции чтения, записи, очистки, печати из пункта «Файлы», операции с буфером обмена и настройки шрифта из пункта «Окна» действуют одинаково в различных контекстах: электронная таблица, результаты анализа, графики.

Страничная картотека

Страничная картотека является эффективной альтернативой оконному захламлению рабочего экранного пространства. Благодаря такому решению, каждый компонент использует максимум свободного пространства, а пользователь в значительной степени избавляется от неадекватных *швейцарских* обязанностей (переместить, свернуть–развернуть, закрыть–открыть окно).

Страницами картотеки являются: *электронная таблица, редактор результатов анализа* и от одного до 15 *графиков*. Переключение страниц производится нажатием его ярлыка–закладки. Эти закладки располагаются в правой части нижней строки экрана:

и имеют следующие обозначения:

- ⌊*Dat*⌋ — страница электронной таблицы;
- ⌊*Rez*⌋ — страница редактора результатов анализа;
- ⌊*Gr*$_i$⌋ — страницы графиков i = 1–15.

Страницы электронной таблицы и редактора результатов анализа являются постоянными и имеют привычные правую–вертикальную и нижнюю–горизонтальную линейки перемещения. Однако для быстрого движения по странице удобно также пользоваться быстрыми клавишами [PageUp], [PageDown], [Home], [End].

Графические страницы. Графические страницы возникают при выводе графика и могут быть *удалены* нажатием на левую верхнюю кнопку закрытия страницы или на клавишу [F5].

После вывода на 15-ю графическую страницу следующий вывод производится на первую графическую страницу, замещая там имеющийся график. Чтобы не терять в результате этого нужную информацию, следует своевременно удалять ненужные графические страницы.

Перемещение. Графические страницы можно *перемещать*. Для этого подведите указатель мыши к ярлыку некоторой страницы, нажмите ***правую*** кнопку мыши и, не отпуская ее, ведите указатель (он изменит свою привычную форму на стрелку с пачкой листков) до ярлыка страницы назначения. Здесь отпустите кнопку мыши.

Каскад окон. Имеется возможность одновременной визуализации нескольких страниц (рис. 2.2), чему служит пункт «*Каскад*» из раздела «*Окна*» верхней командной строки.

Здесь предоставлен выбор из следующих альтернатив:

- одна страница на экране;
- две страницы на экране;
- три страницы на экране;
- четыре страницы — по две в ряд;

- шесть страниц — по три в ряд.

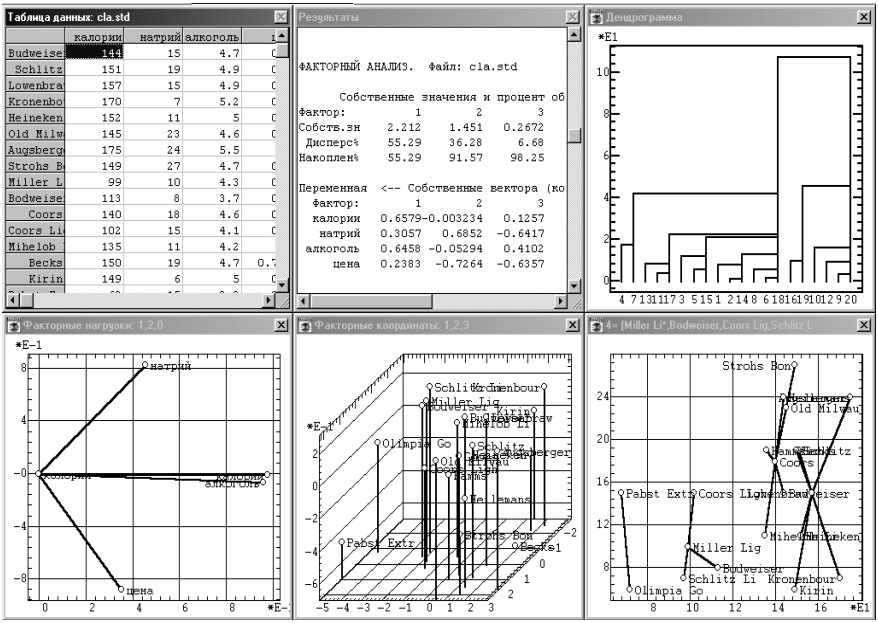


Рис. 2.2. Каскад из шести страниц в рабочем окне

Особенно полезна такая операция для одновременного представления на экране нескольких графиков для их визуального сравнения.

Диалоговые панели

Диалог базируется на двух типах экранных панелей: *меню выбора* и *бланки ввода*.

Меню выбора включают серию пунктов–кнопок, нажатие на одну из которых приводит к выполнению соответствующей альтернативы. При этом каждая такая экранная кнопка традиционно сопровождается отдельной *быстрой* клавишей вызова, в качестве которых используются порядковые цифровые и алфавитные клавиши. Такая сквозная «нумерация» позволяет пользователю быстро вырабатывать автоматизмы на выполнение рутинных цепочек операций.

Бланки ввода содержат поля, в которые следует вводить значения или выражения. В отличие от большинства Windows–приложений, введенные в такие поля значения сохраняются от вызова к вызову и от сеанса к сеансу, что устраняет нудную необходимость повторного ввода одних и тех же значений.

Некоторые диалоговые панели сочетают свойства меню и бланка: в них кнопки сопровождаются полями ввода значений для уточнения выполнения выбранной процедуры.

Отмена (сброс) меню/бланка дублируется быстрой клавишей [Esc], а исполнение — клавишей [Enter]. Эти клавиши аналогичны кнопкам «*Утвердить*», «*Отменить*» (или левой верхней кнопке закрытия окна), присутствующим в некоторых экранных меню.

Основные цвета меню могут быть изменены (см. ниже в «Настройки»).

Ограничение. По техническим причинам в бланках ввода не действует клавиша [Del], убирающая следующий символ, вместо нее следует пользоваться клавишей [BackSpace], убирающей предыдущей символ.

Бланки выбора переменных. Очень часто из электронной таблицы нужно выбрать только часть переменных для анализа, выполнения различных преобразований, построения графиков и т. п.

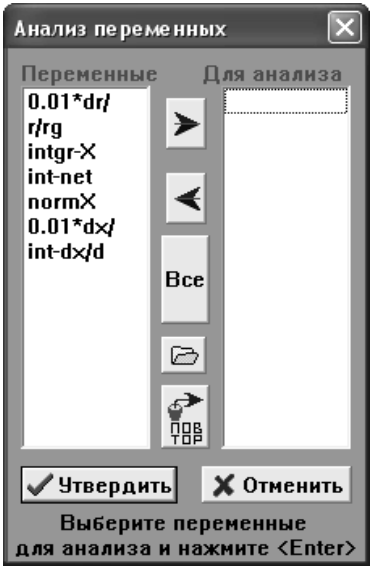


Рис. 2.3. Бланк выбора переменных

Такой выбор осуществляется в специальных бланках (рис. 2.3). И хотя вид таких банков может немного меняться в зависимости от контекста, но общая композиция и принципы работы с ними одинаковы:

- слева имеется поле списка переменных из электронной таблицы;
- справа расположены поля списков выбранных переменных.

Между полями находятся кнопки переноса выделенных переменных из одного поля в другое и кнопка выбора всех переменных «*Все*». Отдельные переменные можно также переносить по двойному щелчку мыши на имени переменной в левом списке (что значительно быстрее).

При повторных вызовах бланка, можно не выбирать снова те же переменные, а повторить их предыдущий выбор нажатием на экранную кнопку «*Повтор*».

Здесь же имеется кнопка с пиктограммой чтения, которая позволяет выявлять различия и вычислять корреляции (разд. 6.3, 6.4, 7.1-7.5) между выбранными парами переменных. Для этого необходимо предварительно подготовить и записать в текущий архив текстовый файл (тип .TXT), в котором построчно перечислить номера пар анализируемых переменных, разделенных пробелом.

После завершения формирования списка анализируемых переменных следует нажать кнопку «*Утвердить*» (дублируется клавишей [Enter]). Кнопка «*Отменить*» (дублируется клавишей [Esc]) отменяет бланк выбора переменных.

Настройки

Общесистемные настройки касаются принтера, шрифтов, размера головного окна, ряда вычислительных параметров, визуализации электронной таблицы и цветов меню.

Принтер. Установка принтера производится выполнением подпункта «*Принтер*» из выпадающего меню пункта «*Файлы*» верхней командной строки экрана. При этом появляется стандартный Windows–бланк установок с выпадающим списком принтеров [29], из которого надо выбрать подключенный к вашему компьютеру принтер и нажать [Enter]. Отказаться от переустановки принтера можно нажатием [Esc].

Шрифты. Установка шрифта производится независимо для трех страничных компонентов: *электронная таблица, редактор результатов и графический редактор*. Для переустановки шрифта необходимо перейти к экранной странице соответствующего компонента и выполнить подпункт «*Шрифт*» из выпадающего меню пункта «*Окна*» (можно также нажать седьмую слева кнопку из третьей экранной строки). При этом появляется стандартный Windows–бланк со списком доступных шрифтов, размеров и стилей и цвета букв. Установите в этом бланке нужные значения и нажмите [Enter]. Отказаться от переустановки шрифта можно нажатием [Esc].

Исходно для этих компонентов установлены наиболее подходящие шрифты. Однако в конфигурации Windows на вашем компьютере может не быть конкретного шрифта или же он может не содержать символов кириллицы. С другой стороны, даже если такой шрифт и имеется, то ваш принтер может содержать его вариант без русских букв, тогда на печать текст будет выдаваться греческими буквами. Все это — не наши проблемы, а проблемы Microsoft, но решать их придется вам посредством подбора и переустановки шрифта для соответствующей экранной страницы.

Вычислительные параметры. Установка вычислительных параметров касается:

а) числа значащих цифр в экранной выдаче результатов анализа (исходное число равно 4);

б) критического уровня значимости нулевой гипотезы α (см. разд. 5.4) для выдачи словесной интерпретации результатов анализа и доверительных интервалов (исходное значение равно 0.05).

Изменение этих значений производится из подпункта «*Установки*» выпадающего меню пункта «*Окна*», вслед за ним появляется экранный бланк (рис. 2.4) с полями ввода значений этих двух параметров. Введите новые значения и нажмите кнопку «*Утвердить*» или клавишу [Enter] или же откажитесь от ввода, нажав клавишу [Esc].

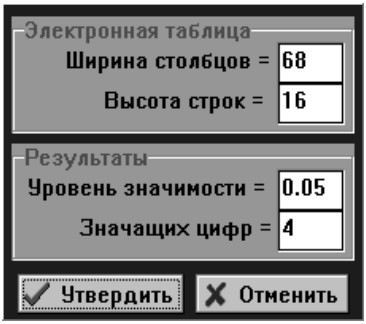


Рис. 2.4. Бланк установки вычислительных параметров и ячеек электронной таблицы

Электронная таблица. Установки, относящиеся к электронной таблице (см. рис. 2.1), касаются экранных размеров ее ячеек (измеряются в пикселях) и производятся в том же выкидном меню, что и установки вычислительных параметров. Переустановка этих размеров бывает необходима в случае перехода к более крупному шрифту.

Цвета меню. Изменение цвета экранных меню производится из подпункта «*Цвета*» выпадающего меню пункта «*Окна*», вслед за чем появляется специальный бланк установки цветов. В качестве основных компонентов меню выступают следующие:

- фоновая панель исходно синего цвета;
- фронтальные панели исходно зеленого цвета;
- заголовочные надписи исходно красного цвета;
- надписи–пояснения исходно синего цвета.

Для переустановки цвета любого компонента следует щелкнуть мышью по сопоставленной ему кнопке, после чего появляется стандартный Windows-бланк цветовой палитры [29], в котором надо выбрать нужный цвет и нажать кнопку «*Утвердить*» или клавишу [Enter]. Внизу бланка находятся две дополнительные кнопки установки стандартной цветной и черно–белой палитры. Для сброса бланка нажмите клавишу [Esc] или же левую верхнюю кнопку закрытия меню.

Сохранить настройки. Все произведенные настройки полезно сохранить для следующих сеансов, выполнив пункт «*Запомнить*» выкидного меню «*Окна*» верхней командной строки.

2.3. Использование формул

Систематическое использование вычислений, задаваемых пользователем по вводимой формуле (везде, где это возможно и осмысленно), существенно расширяет аналитические и комбинаторные возможности пакета. Для этого имеются специальные типовые *бланки формульного ввода*, а также бланки составления и редактирования формул. И хотя эти бланки в деталях могут различаться в зависимости от контекста, но основные принципы работы с ними едины.

Рассмотрим типовой бланк формульного ввода, приведенный на рис. 2.6, вверху которого указано название в зависимости от контекста.

В таком бланке имеется восемь независимых полей для ввода восьми различных формул. Поскольку поля ввода запоминающие, то их содер-

жимое сохраняется от вызова к вызову и от сеанса к сеансу и повторный ввод уже имеющейся формулы не требуется.

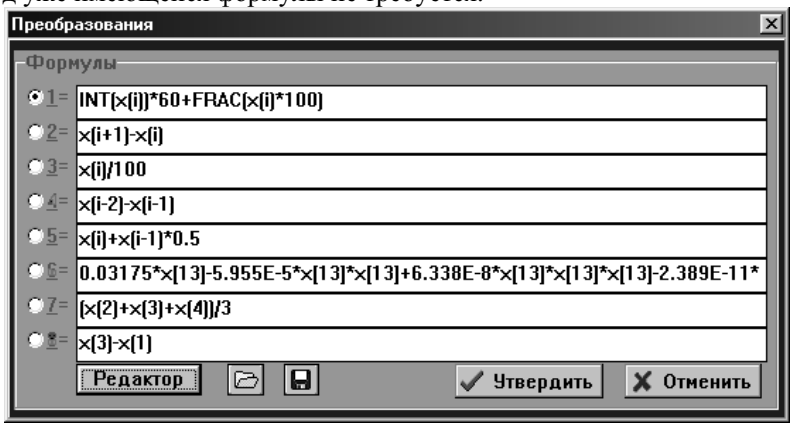


Рис. 2.6. Бланк формульного ввода

В бланке имеются две кнопки с пиктограммами чтения и записи. Эти кнопки позволяют сохранять содержимое бланка в архиве в специальных файлах и считывать в бланк формулы из архива. Чтение и запись файлов осуществляется по стандартному диалогу файловой системы (см. разд. 3.2). Таким образом, можно использовать не только память восьми формул экранного бланка, но и произвольное количество таких наборов, имеющихся в дисковом архиве. Исполнение выбранной формулы происходит по следующим действиям:

- нажатие клавиши — порядкового номера формулы;
- двойной щелчок мышью по формуле;
- зажигание фонарика слева от формулы и нажатие кнопки «*Утвердить*» или клавиши Enter .

Ошибки в формуле типа несоответствия скобок, неправильного обозначения функций и переменных и т. п. выявляются только на этапе вычислений с выдачей соответствующих диагностик (см. разд. 2.7).

Редактор формул. Ввести новую формулу можно непосредственно в рассмотренный бланк формул, однако часто удобно пользоваться специальным *формульным редактором*, вызываемым по нажатию кнопки «*Редактор*», после чего появляется бланк, изображенный на рис. 2.7 (вид этого бланка может незначительно варьироваться в зависимости от контекста). В этом бланке имеются следующие компоненты:

- верхнее поле вводимой формулы;
- левый список переменных из электронной таблицы или формальных переменных, допустимых в данном контексте (см. разд. 3.4, 10.6);

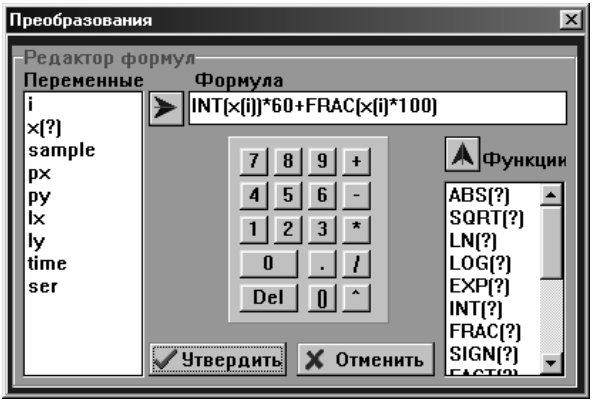


Рис. 2.7. Бланк редактора формул

- правый список допустимых алгебраических функций: абсолютное значение (ABS), корень квадратный (SQRT), натуральный и десятичный логарифмы (LN, LOG), экспонента (EXP), целая и дробная части аргумента (INT, FRAC), знак аргумента (SIGN={−1, 0, +1}), факториал от аргумента (FACT), случайное число по равномерному закону распределения в интервале от 0 до значения аргумента (RAND), прямые и обратные тригонометрические функции (SIN, COS, TAN — аргумент в радианах, ASIN, ACOS, ATAN — результат в радианах);
- кнопки переноса элементов из списков в поле формулы;
- центральная цифровая и операционная клавиатура, включает арифметические операции сложения, вычитания, умножения, деления, возведения в степень (+, –, *, /, ^), модуль (*a%b* — «взять *a* по модулю *b*»);
- кнопки утверждения составленной формулы или ее отмены с закрытием редактора формул.

Перенос элемента из списка в поле формулы производится по двойному щелчку мышью или же после его выделения в списке и нажатия на кнопку переноса. Для изменения порядка выполнения операций в формуле служат круглые скобки. Утвержденная формула будет занесена в текущее поле бланка формул (у которого горит фонарик).

2.4. Экранная помощь и совет

Электронный справочник. Контекстная справка вызывается (рис. 2.8) по нажатию клавиши F1 и дает пояснения по работе с активным меню, окном или выполняемым методом анализа, преобразованием и т. п.

Войдя в справочник, можно далее путешествовать по электронным ссылкам или же перейти в оглавление и вызвать нужную рубрику, а также произвести поиск нужного термина по индексному указателю.

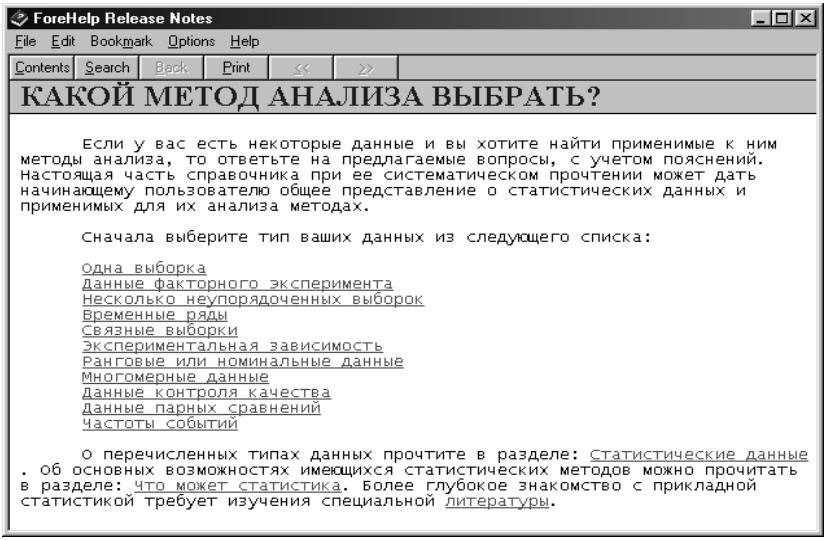


Рис. 2.8. Электронный справочник в режиме экспертной системы

Экспертная система. Особым компонентом справочника является *экспертная система* (рис. 2.8), которая позволит пользователю подобрать метод анализа для имеющихся данных. Для входа в экспертную систему нужно выполнить команду *«Помощь»* из верхней командной строки и выполнить пункт «*Выбор метода*» или перейти в оглавление справочника и вызвать пункт «*Выбор метода*».

Далее в последовательных экранах необходимо уточнить тип данных, предназначенных для анализа. После ряда таких уточнений будет предложено несколько вариантов статистического анализа ваших данных. Сделайте последний выбор, и на экране появится описание требуемого метода.

2.5. Буфер обмена

Буфер межпрограммного обмена или *Clipboard* является удобным изобретением Windows. Этот буфер может сохранять как текстовую, так и графическую информацию, забранную из любых страниц и бланков или внешних приложений. Имеющуюся в буфере информацию можно вставить в любое активное окно или меню практически любого приложения.

Следует, однако, помнить, что при очередном заборе в буфер новой информации предыдущее содержимое буфера теряется.

[Ctrl]+[C] или [Ctrl]+[Ins] = *копировать* выделенную информацию из текущего меню или окна в буфер обмена.

[Ctrl]+[X] или [Shift]+[Del] = удалить выделенную информацию из текущего меню или окна и перенести ее в буфер обмена.

[Ctrl]+[V] или [Shift]+[Ins] = *вставить* информацию из буфера обмена в активное окно, электронную таблицу или меню.

Способы выделения информации зависят от типа окна или меню. Как правило, выделение производится движением мыши с нажатой левой кнопкой.

Вышеприведенные сочетания клавиш межбуферного обмена дублируются постоянными инструментальными кнопками в третьей строке экрана, а также соответствующими пунктами выкидного меню пункта «*Окна*» из верхней командной строки.

Внимание! Иногда клавишные операции с буфером обмена перестают работать. Это значит, что вы случайно переключились в русский регистр и вместо [Ctrl]+[x] генерируется [Ctrl]+[ч].

2.6. Диагностика ошибок

Большинство ошибок пользователя (неправильный ответ, нарушение требований статистического метода, неверные данные и т. п.) обнаруживается на начальном этапе выполнения операции и сопровождается соответствующими диагностиками на появляющейся информационной панели.

Однако в отдельных случаях некоторая ошибка, которую не удалось сразу обнаружить, может повлечь прерывание выполнения очередной операции с выдачей системного предупреждения. Тогда необходимо проверить, правильно ли подготовлены исходные данные для выбранного метода и корректны ли действия, предшествующие ошибке. Ниже приведен список предупреждений об ошибках в алфавитном порядке.

!ДАННЫЕ НЕ ДЛЯ ДВУХ РЕГРЕССИЙ! — данные в электронной таблице не соответствуют методу сравнения двух регрессий.

!ДАННЫЕ НЕ ДЛЯ ТАБУЛЯЦИИ! — введенные в электронную таблицу данные не подходят для категориального анализа.

!ДИСПЕРСИЯ=0 ДЛЯ НЕКОТОРОЙ ПЕРЕМЕННОЙ! — некоторая переменная обладает нулевой дисперсией.

!ДЛИНА ПЕРЕМЕННОЙ БОЛЬШЕ ДОПУСТИМОЙ! — число значений некоторой переменной больше максимально допустимого объема (определяется поставляемой модификацией STADIA).

Заменено недопустимых аргументов = n — при выполнении преобразования часть недопустимых аргументов заменено нулями.

!ИЗМЕРЕНИЙ < 3! — для факторного анализа у выбранных переменных слишком мало измерений.

!ИЗМЕРЕНИЙ МЕНЬШЕ ЧЕМ ПЕРЕМЕННЫХ! — матрица данных содержит измерений меньше, чем переменных, что не допускается выбранным методом.

!МАЛО ЗНАЧЕНИЙ ПЕРЕМЕННЫХ! — у выбранных переменных слишком мало измерений.

!МЕТОД НЕПРИМЕНИМ ПРИ ТАКИХ ПАРАМЕТРАХ! — исходные данные или установленные параметры не удовлетворяют требованиям используемого метода.

!МНОГО ПЕРЕМЕННЫХ! — в дискриминантном анализе выбрано слишком много переменных (определяется поставляемой модификацией STADIA).

!МНОГО ПЕРЕМЕННЫХ ИЛИ ИЗМЕРЕНИЙ! — в кластерном анализе выбрано слишком много переменных или измерений.

Нарушен файловый состав STADIA — на жестком диске недостает некоторых системных файлов — необходимо восстановить конфигурацию с дискеты-копии.

!НЕВЕРНАЯ НУМЕРАЦИЯ КЛАССОВ ИЛИ КЛАССОВ<2! — номера классов в данных для дискриминантного анализа не являются целыми числами в интервале от 0 до *N* без пропусков.

!НЕДОСТАТОЧНО ДАННЫХ ДЛЯ ГРАФИКА! — для выбранного типа графика недостаточно указанных переменных.

!НЕКВАДРАТНАЯ МАТРИЦА! — используемый метод требует квадратной матрицы данных.

!НЕРАВНАЯ ДЛИНА ПЕРЕМЕННЫХ — МЕТОД НЕПРИМЕНИМ! — выбранный метод анализа требует указания переменных, имеющих равное количество значений.

!НЕСООТВЕТСТВИЕ МОДЕЛИ! — в регрессионной модели число коэффициентов меньше двух или больше числа измерений, или переменных меньше двух.

!НЕТ МЕСТА НА ДИСКЕ! — диск не содержит достаточно свободного места для записи файла данных.

!НЕ УКАЗАН ФАЙЛ! — не указан файл чтения или записи.

!ОШИБКА ЧИСЛА ГРАДАЦИЙ ФАКТОРА 1! — число уровней первого фактора, указанное в диалоге двухфакторного анализа с повторными измерениями не кратно числу переменных в матрице данных.

!2 > ПЕРЕМЕННЫХ > n! — для исполняемого метода выбрано слишком мало или много переменных.

!ПЕРЕМЕННЫХ < 3! — в дискриминантном анализе указано малое число переменных.

!ПЕРЕМЕННЫХ > n! для исполняемого метода выбрано слишком много переменных (определяется поставляемой модификацией STADIA).

Переполнена память: *n* — попытка ввести в электронную таблицу данных, больше ее размера *n* (определяется поставляемой модификацией STADIA).

!ПРЕВЫШЕН ЛИМИТ ИТЕРАЦИЙ! — в регрессионной процедуре удовлетворительное решение не найдено в допустимом числе итераций.

!СИНГУЛЯРНАЯ МАТРИЦА, РЕШЕНИЕ НЕВОЗМОЖНО! — данные, использованные для выбранного метода, имеют особенное распределе-

ние, не позволяющее однозначно рассчитать модель (например, все экспериментальные точки расположены на m–1-мерной плоскости).

!ЧИСЛО БИНОВ ГИСТОГРАММЫ < 3! — указано малое число интервалов для расчета гистограммы.

Узнайте у поставщика пароль по коду: <code> — произошли изменения в конфигурации компьютера, следует связаться с поставщиком STADIA и узнать у него пароль по выдаваемому коду.

Нижеследующие диагностики относятся к вводимым формулам различного типа. Большинство диагностик сопровождается указанием контекста обнаружения ошибки или порядковым номером символа в формуле.

!ПЕРЕМЕННЫХ В ФОРМУЛЕ > 64!

!ТЕКСТОВАЯ КОНСТАНТА НЕДОПУСТИМА!

!НЕТ КОНЦА ТЕКСТА!

!ОШИБКА КОНСТАНТЫ!

Нет такого разделителя:

Нет такой функции:

Нет такой переменной:

Нет закрывающей скобки

Ошибка операции:

Глава 3

РАБОТА С ДАННЫМИ

«Ибо много званых,
а мало избранных»
[от Матф., 22, 14]

Настоящая глава содержит все сведения, необходимые для оперирования с информацией, предназначенной для последующего анализа: ввод данных с клавиатуры или из дискового архива, импорт/экспорт данных в различных форматах и предварительные преобразования данных.

3.1. Электронная таблица

Электронная таблица (рис. 3.1) представляет собой *матрицу* для ввода, сохранения, редактирования и преобразования данных, в которой *столбцы* отвечают переменным, а *строки* — измерениям или объектам. Элементы таблицы могут содержать как числовые, так и символьные (номинальные) значения, однако последние используются, в основном, не для вычислений, а в информационных целях. Верхние серые ячейки таблицы предназначены для наименований переменных, а левые серые ячейки — для наименований измерений или объектов.

Таблица данных: s40.std

	5км	15км	10км	15-1кр	15-2кр	15-3кр	x7
VO	20.15	58.04	37.37	18.55	19.20	19.49	
DT	21.15	61.08	39.46	19.50	20.28	20.50	
AS	21.31	61.22	?	20.40	20.25	20.17	
VZ	21.32	60	40.5	19.50	20.10	20.08	
AB	21.37	60.25	39.35	19.34	20.26	20.25	
AK	21.40	64.50	42.31	21.47	21.34	21.35	
AC	22.00	63.27	42.03	20.44	21.16	21.37	
JB	23.55	70.42	47.05	23.07	23.43	23.52	
JU	23.56	68.48	45.56	22.30	22.58	23.20	
VG	24.09	72.33	?	23.32	24.23	24.37	
BU	24.42	73.43	46.54	23.50	24.35	25.18	
SA	24.55	68.15	44.52	23.22	24.08	23.45	
IC	25.15	74.15	48.10	24.22	25.00	24.53	
EJ	26.05	77.27	51.05	24.43	25.22	26.17	
SP	26.40	76.13	49.46	25.13	25.27	25.33	
16							
17							

Рис. 3.1. Электронная таблица с матрицей данных, включающей имена переменных и имена объектов

Ввод с клавиатуры. Чтобы ввести или изменить конкретное значение в электронной таблице просто сделайте его позицию *активной* — темного цвета (щелкнув по ней указателем мыши или используя клавиши движения курсора) и наберите новое значение (нажатие после этого [Enter] влечет переход к следующей позиции).

Изображение чисел. Для представления значений в таблице отведено восемь позиций, поэтому, если значение превышает этот размер, то воспользуйтесь расширенным полем редактирования в третьей экранной строке, содержащей инструментальные кнопки.

Числа обычно записываются в *научной нотации*, когда очень большие и очень малые значения представлены с десятичным множителем, показатель которого следует за символом «*E*», например, 2.8E–3 соответствует 0.0028.

Общие операции. В рамках электронной таблицы работают общие операции чтения, записи и очистки содержимого (см. разд. 3.2), печати (см. разд. 2.2) и буфера обмена (см. разд. 2.5).

Размер ячеек. Исходные экранные размеры ячеек электронной таблицы могут быть переустановлены (см. «Настройки» разд. 2.2).

Символьные переменные. В качестве значений переменных могут быть использованы не только числовые, но и *символьные значения*. Числовые значения могут быть преобразованы в символьные и обратно посредством операции *кодирования* (см. разд. 3.4).

Наименования переменных. Для изменения *наименования переменной* выделите ее столбец щелчком мыши по верхней серой ячейке и произведите изменение имени в расширенном поле редактирования в третьей строке инструментальных кнопок.

Наименования объектов. Для изменения *наименования объекта* выделите его строку щелчком мыши по левой серой ячейке и произведите изменение имени в расширенном поле редактирования в третьей строке инструментальных кнопок.

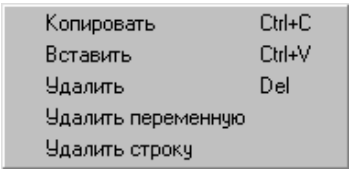
Переходы. *Переход* к следующим позициям таблицы осуществляется клавишами движения курсора, а смена страниц — клавишами [PageUp], [PageDown]. Для *быстрого перемещения* по таблице используйте линейки прокрутки внизу и справа экрана, управляемые мышью.

Удаления. *Удаление* числа в текущей позиции производится клавишей [Del]. Для удаления всей переменной выделите весь ее столбец и нажмите клавишу [Del].

Операции с фрагментами. В таблице можно также выделять и отдельные *фрагменты данных*. Для этого подведите мышь к верхнему левому фрагменту данных, нажмите левую клавишу и, не отпуская ее, ведите мышь к правому нижнему краю фрагмента. Далее выделенный фрагмент можно удалить (клавишей [Del]), скопировать или забрать в буфер межпрограммного обмена или вставить из него (см. разд. 2.5).

Перемещение переменных. Можно также *переместить* отдельные переменные в таблице. Для этого подведите указатель мыши к имени переменной, нажмите ***правую*** кнопку мыши и, не отпуская ее, ведите указатель (он изменит свою привычную форму на стрелку с листком) до имени переменной назначения. Здесь отпустите кнопку мыши, и переменная будет вставлена в указанное место. При необходимости можно затем удалить исходную переменную.

Более сложные операции с операции с данными в электронной таблице производятся средствами блока преобразований (см. разд. 3.4).



Контекстное меню. Контекстное меню вызывается нажатием правой кнопки мыши и дублирует ряд операций с буфером обмена и удаления строк и переменных.

Представление данных. Форма представления исходных данных (одна или две переменные, матрица или псевдоматрица данных) должна отвечать требованиям конкретного статистического метода. С этими требованиями можно познакомиться в описании статистических методов.

Ограничения. Объем электронной таблицы определяется поставленной модификацией пакета и может доходить до 64 000, а число переменных (столбцов) — до 500.

3.2. Чтение, запись и удаление файлов

Ввод данных в электронную таблицу и запись содержащихся в ней данных в дисковый архив возможны, когда страница ⌊*Dat*⌋ с электронной таблицей является активной. Операции чтения, записи и очистки действуют аналогичным образом для страницы результатов (разд. 5.5) и для страниц графиков (разд. 4.1).

F3 – *чтение.* Операция чтения данных выполняется из выкидного меню пункта «*Файлы*», а также по нажатию на быструю функциональную клавишу F3 или же на кнопку экранной строки инструментов.

Слияние файлов. Если в рабочем пространстве уже имеются данные, то перед чтением файлов выдается запрос: «*Очистить область от данных (Да/Нет)*?» В случае отрицательного ответа вводимые данные будут добавлены к имеющимся, что позволяет сливать несколько файлов данных в электронной таблице.

F4 – *запись.* Операция записи содержимого активной страницы выполняется из выкидного меню пункта «*Файлы*», а также по нажатию на

быструю функциональную клавишу [F4] или же на вторую слева кнопку из третьей экранной строки.

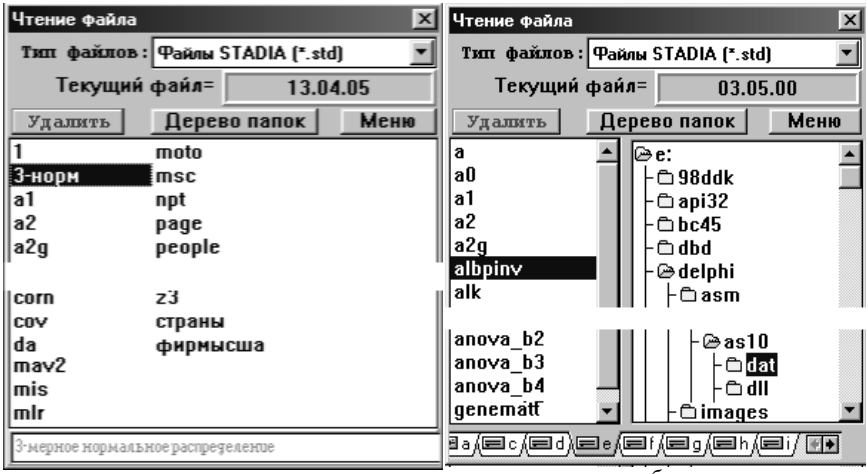
[F5] — *очистить*. Чтобы очистить содержимое активной страницы нажмите клавишу [F5] или же выполните соответствующий подпункт из пункта «*Файлы*» верхней командной строки.

Форматы данных. Числовые данные на диске могут храниться в числовом, текстовом и *DBF*– форматах. *Числовой формат* более компактен и более быстр при чтении и записи. Записанные в этом формате файлы имеют тип *.STD*.

DBF–формат обеспечивает совместимость с системами управления базами данных *DBase–II, III, IV* и с большим числом других пакетов, допускающих экспорт/импорт файлов в этом формате.

Файлы, записанные в *текстовом формате* (*ASCII*–файлы), могут быть обработаны любым внешним текстовым редактором. В текстовом файле данных первая строка может содержать (или же не содержать) имена переменных, последующие строки содержат значения переменных. В качестве разделителей для чисел и имен переменных в строке могут использоваться пробелы, знаки табуляции или запятые. Фрагменты строк, начинающиеся с символа «точка с запятой», считаются комментариями и при вводе игнорируются.

Бланк чтения/записи. *Чтение/запись* файлов данных производится средствами стандартного экранного бланка (рис. 3.2–3.4).



а *б*

Рис. 3.2. Бланк чтения файлов: *а* — обычный вид; *б* — с окном папок

Для лучшей ориентировки относительно предметного смысла файлов в нижней строке бланка выводится *комментарий* к текущему файлу (рис. 3.2, *а*). Комментарий может быть изменен перед записью файла (см. ниже).

Для *перемещения* по большим архивам, оглавления которых не умещаются на экране, служат линейки перемещения и клавиши [PageDown], [PageUp]. Кроме того, для быстрого перехода к нужному файлу можно нажать клавишу первого символа из его имени — и оглавление сдвинется к имени первого из таких файлов.

Смена архива. Если нужные файлы расположены в другом архиве (папке), то нажмите кнопку *«Дерево папок»* и в правой части бланка появится дерево папок текущего диска, а внизу — список логических дисков (рис. 3.2, *б*). Выберите нужный диск и папку в нем и в бланке будет выдано оглавление установленного архива (чтобы развернуть папку, может потребоваться щелчок мышью по ней). Для отмены окна папок еще раз нажмите кнопку *«Дерево папок»*.

Чтение файла. При *чтении* файла щелкните дважды по имени файла в оглавлении текущего архива или же выберите файл из оглавления и нажмите [Enter]. Если же вы передумали, то нажмите клавишу [Esc] для отмены бланка или используйте кнопку закрытия окна.

Объединение файлов. Если в электронной таблице уже имеются данные, то при чтении файлов выдается запрос: «*Очистить электронную таблицу (Да/Нет)*?» В случае отрицательного ответа вводимые данные будут добавлены к имеющимся, что позволяет объединять несколько файлов данных для совместного анализа и перекомпоновки.

Поиск в комментариях. Для лучшей ориентации в больших архивах в бланке чтения предусмотрено выкидное меню (показано в правом левом углу рис. 3.3), вызываемое по нажатию кнопки *«Меню»*.

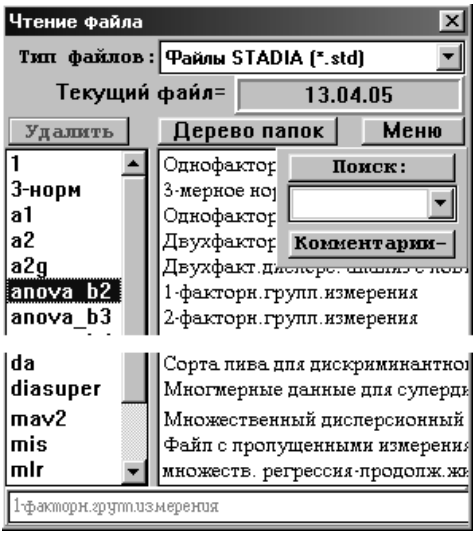


Рис. 3.3. Бланк чтения с выкидным меню и окном комментариев файлов

В этом меню есть кнопка поиска файлов и кнопка комментариев. При нажатии кнопки комментариев в правом полуокне бланка появляются комментарии к списку файлов (рис. 3.3). Для отмены окна комментариев повторно нажмите кнопку *«Комментарии»*.

Кнопка *«Поиск»* позволяет выводить в список только те файлы, в комментариях которых встречается заданный контекст. Контекст или образец поиска задается в нижерасположенном поле ввода. Например, при задании в поле ввода контекста «*мер*», в список будут отобраны только файлы, содержащие в

своих комментариях *«измерения», «размер», «меры»* и т. п.

Поле ввода снабжено кнопкой выкидного списка ранее введенных контекстов, из которого можно просто выбрать нужный контекст. Для того чтобы изменить имеющийся контекст в списке, надо сначала его выбрать из списка, а затем отредактировать в поле ввода.

При исполнении поиска с пустым контекстом выдается весь список файлов.

Импорт данных. Для *импорта* данных, записанных в другом формате (например, в текстовом формате или *DBF-III* форматах), выберите соответствующий пункт из *выпадающего списка* типа файлов, после чего оглавление обновится с выдачей списка *TXT*– или *DBF*–файлов. После этого выполните вышеописанный выбор по оглавлению.

Данные для чтения в *текстовом формате* должны удовлетворять следующим условиям:

- каждая текстовая строка содержит последовательность значений переменных для одной строки матрицы данных;
- в качестве разделителей между числами может использоваться любое число пробелов, а также запятые или символы табуляции;
- первая текстовая строка может включать список имен переменных; если же первое значение первой строки является числом, то вся она интерпретируется как первая строка данных;
- при необходимости импортировать имена объектов необходимо их перечислить в первом столбце текстового файла, а первым в строке имен переменных поставить слово «объекты».

Примечание. Есть некоторая тонкость при импорте текстовых файлов, столбцы которых имеют разную длину. В этом случае в качестве разделителей элементов в строках предпочтительно иметь знаки табуляции или запятые, а недостающие до квадратной матрицы элементы должны быть представлены пустыми местами. В противном случае после ввода некоторые элементы могут быть смещены в левые столбцы.

Запись файла. Для *записи* файла необходимо в бланке записи (рис. 3.4) ввести имя файла в поле ввода и нажать кнопку *«Записать файл»* или клавишу [Enter]. Для записи данных в уже имеющийся файл достаточно дважды щелкнуть по имени файла в оглавлении. При одинарном щелчке по имени файла оно переносится в поле ввода имени файла записи.

Если же вы передумали, то нажмите клавишу [Esc] для отмены бланка записи или используйте кнопку закрытия окна.

Рис. 3.4. Бланк записи файлов

Комментарий. Перед записью файла очень полезно сформировать или обновить его *комментарий* в поле ввода нижней строки бланка. Для этого переведите туда курсор и введите комментарий или отредактируйте имеющийся. Наличие подобных комментарием очень облегчит ориентировку в необозримых архивах данных, которые рано или поздно возникнут. О использовании комментариев см. выше в чтении файлов.

Экспорт данных. Для *экспорта* данных в другом формате (например, в текстовом *ASCII*– или *DBF*– форматах), выберите соответствующий пункт из выпадающего меню типа файлов, после чего оглавление обновится с выдачей списка *TXT*–, *TAB*– или *DBF*–файлов. После этого выполните вышеописанные операции записи.

Данные для записи в *текстовом формате* будут разделяться пробелами или же знаками табуляции в зависимости от выбранного типа (*TXT* или *TAB*). Первая строка импортируемого файла содержит имена переменных. Если левый (объектный, серого цвета) столбец электронной таблицы содержит оригинальные имена объектов, то первым столбцом текстового файла будет список имен объектов. При этом первым в строке имен переменных будет слово «Объекты».

Удаление файла. Чтобы удалить (стереть) файл из текущего архива, выделите его наименование в оглавлении и нажмите на бланке кнопку *«Удалить»*, подтвердив затем необходимость удаления этого файла.

Перенос данных. Перенос данных из одного пакета в другой обычно производится через буфер обмена. При переносе данных из STADIA в другие пакеты обычно проблем не возникает. Обратный же перенос иногда наталкивается на проблемы, не поддающиеся рациональному объяснению. Иногда проблемы возникают и при экспорте-импорте данных в текстовом формате (ASCII-файлов)

В таких случаях в качестве промежуточного звена удобно использовать редактор MS Word. Более того, там имеются удобные средства работы с таблицами, позволяющие переводить многоколоночный текст в таблицу и обратно, удалять и вставлять в таблице строки и столбцы, перемещать их содержимое и т. п. Если колонки разделены многими пробелами,

то их можно сократить до одного пробела повторным выполнением операции замены.

Для переноса данных в STADIA полезно предварительно использовать их экспорт из пакета-источника в формате DBASE-III. В этом случае часто удается переносить и наименования переменных.

Пример. При переносе данных в некоторых случаях может потребоваться определенная степень изобретательности. Пусть, например, вы имеете таблицу 14.1.2 в Worde. И вам нужно перенести первый, второй и последний числовой столбцы в электронную таблицу, чтобы выполнить кластерный анализ и получить дендрограмму рис. 14.1.3. Напрямую через буфер обмена это сделать не удастся, поскольку значения в каждой колонке таблицы содержатся в одной ее ячейке, а не в 15-ти отдельных ячейках. В данном случае изобрести можно следующее: 1) создать новый документ Word; 2) выделить колонку в документе 1 и перенести ее в документ 2; 3) выделить полученный текст и преобразовать ее в таблицу; 4) перенести полученную таблицу (из одного столбца и 15 строк) в переменную электронную таблицы; 5) так поступить с каждой колонкой исходной таблицы.

3.3. Калькулятор

[F7] — *калькулятор.*

Калькулятор (рис. 3.5) обеспечивает выполнение различных вспомогательных вычислений по вводимым выражениям и вызывается по нажатию клавиши [F7] или по выполнению раздела «*Калькулятор*» из верхней экранной командной строки.

В *экранном бланке* калькулятора имеются следующие элементы:

- верхнее поле *результата* вычисления;
- поле *формулы* вычислений с выкидным списком ранее введенных формул;
- левый список допустимых *параметров* с кнопкой их переноса в поле формулы, а именно:
  - *i, j* — текущая переменная и текущее измерение в электронной таблице (выделенная щелчком мыши);
  - *x(?,?)* — содержимое ячейки электронной таблицы, определенной двумя параметрами, разделенными запятой: номером столбца и номером строки;
  - *a, b, c, d* — обозначения ячеек таблицы сопряженности размера 2×2 для вычислений по формуле (см. разд. 7.6);

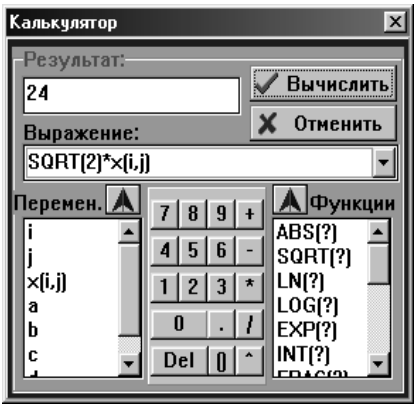


Рис. 3.5. Калькулятор

- правый список допустимых алгебраических и тригонометрических функций (см. разд. 2.3) с кнопкой их переноса в поле формулы;
- панель, воспроизводящая *цифровую клавиатуру*, для набора чисел и арифметических операций (см. разд. 2.3);
- кнопки выполнения вычислений (дублируется клавишей [Enter]) и отмены калькулятора (дублируется клавишей [Esc]).

Набор формулы (после активизации ее поля перемещением указателя мыши со щелчком по левой кнопке) можно производить и полностью с клавиатуры, что значительно быстрее.

Для перенесения в поле ввода уже готовых формул из различных меню и окон целесообразно пользоваться буфером обмена (см. разд. 2.5).

В выкидном списке формул вычислений можно заготовить до 10 часто используемых формул. Для того чтобы ввести новую формулу в список, надо щелкнуть мышью по соответствующей позиции в списке, после чего в поле ввода формулы ввести новую формулу. После этого она будет сохранена в этой позиции списка формул.

Примеры. Для вычисления $\sqrt{\sin^2(0{,}5)+\cos^2(0{,}5)}$ наберите:

*SQRT(SIN(*0.5)^2+*COS(*0.5)^2)* .

Для суммирования текущей ячейки электронной таблицы с тем же измерением следующего столбца наберите: *x(i,j)+x(i*+1,*j)* .

3.4. Преобразования

Блок преобразования данных содержит обширный набор алгебраических, тригонометрических, матричных и других операций, необходимых для преобразования исходных данных в электронной таблице к нужному виду.

[F8] — *блок преобразований.*

Для вызова меню выбора преобразования (рис. 3.6) нужно нажать клавишу [F8] или выполнить пункт «*Преобразования*» в верхней экранной строке команд.

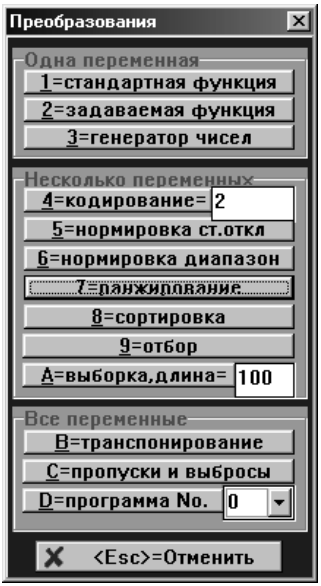


Рис. 3.6. Меню выбора преобразования

Группы преобразований. Операции преобразования разбиты на три группы в зависимости от того, изменяются ли при этом значения одной переменной, нескольких переменных или же всех переменных (матричные операции).

1. Операции над одной переменной. Эти операции производятся над выделенной переменной электронной таблицы. Результат преобразования записывается в ту же самую переменную. Для выделения переменной щелкните по ее имени мышью.

В данную группу входят три типа операций:

- стандартные алгебраические функции;
- функции, задаваемые по вводимой формуле;
- генератор чисел.

Выполнение стандартных и задаваемых функций может проводиться также над выделенным фрагментом электронной таблицы, при этом в такой фрагмент могут входить несколько переменных или все переменные.

2. Операции над несколькими переменными. Эти операции состоят в изменении значений выбранных переменных в электронной таблице:

- *кодирование* значений (замена кодом) по заданному условию;
- *нормировка* значений;
- *ранжирование* значений;
- *сортировка* (переупорядочение) значений по возрастанию/убыванию;
- *отбор* значений по заданному условию;
- *выборка* случайного подмножества значений.

3. Матричные операции. Эти операции проводятся над всем содержимым электронной таблицы:

- транспонирование матрицы данных;
- анализ и замена пропущенных значений.

Стандартные функции. Выполнение пункта «*Стандартная функция*» приводит к появлению меню выбора *стандартных операций* (рис. 3.7), имеющего также поля ввода значений двух параметров *a* и *b*, которые требуются для некоторых операций.

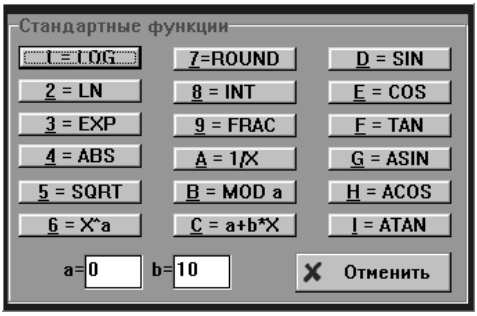


Рис. 3.7. Меню стандартных функций

В число *стандартных функций* входят следующие:

- *логарифмы* десятичный (*LOG*) и натуральный (*LN*), определены только для положительных чисел;
- *экспонента* (*EXP*);
- абсолютное значение (*ABS*);
- квадратный корень (*SQRT*), определен только для положительных чисел;
- *модуль* (*MOD*) по основанию *a*;
- округление (*ROUND*);
- выделение целой (*INT*) или дробной (*FRAC*) части;
- деление *1/X* (не определено для *X=0*);
- возведение *X* в степень a (*X^a*, при этом дробная степень отрицательных чисел не определена);
- *тригонометрические функции* от аргумента, значения которого выражены в радианах: синус (*SIN*), косинус (*COS*), тангенс (*TAN*);
- *обратные тригонометрические функции*, значения которых выражены в радианах: арксинус (*ASIN*), арккосинус (*ACOS*), арктангенс (*ATAN*);
- *линейное преобразование* вида *Y=a+b*X* от одной переменной; при *a* = 0 значения *X* умножаются на константу; при *b* = 0 все значения равны константе *a*.

Примечание: неопределенные значения функций заменяются нулями с выдачей числа замененных значений.

Задаваемая функция. Выполнение пункта «*Задаваемая функция*» приводит к появлению типового *бланка формул* (см. рис. 2.6), в который надо ввести новую *формулу преобразований* или же выбрать одну из имеющихся в бланке формул и нажать [Enter] или экранную кнопку «*Утвердить*».

Для каждого значения текущей переменной или в выделенном фрагменте электронной таблицы вычисляется новое значение на основании введенной формулы. Если в этой формуле фигурируют другие переменные из электронной таблицы, то они, естественно, должны иметь одинаковую с текущей переменной размерность.

В качестве переменных в формулах могут использоваться имена переменных из электронной таблицы, а также формальные обозначения *x(?)*, где позиция *"?"* может быть: а) целым числом 1, 2, 3,..., указывающим порядковый номер переменной в электронной таблице; б) или же арифметическим выражением с индексом *i*, обозначающим номер текущей пере-

менной. Последний вариант во многих случаях более универсален. Действительно, если в формуле используются явные имена переменных, то она оперирует только с так же обозначенными переменными. В случае же обозначений *x(?)* формула применима ко всем данным, имеющим одинаковое количество одинаково расположенных переменных, независимо от их конкретных наименований. Формулы же с индексом *i* можно применять и для преобразования различных фрагментов электронной таблицы.

Внимание: В пустую переменную нельзя занести вычисленные по формуле значения, поскольку количество вычисляемых значений определяется размерностью текущей переменной. В таких случаях надо предварительно сдублировать в пустую переменную значения какой-либо переменной из электронной таблицы.

Примеры: 1) формула *(x(i)+x(i+1))/2* вычисляет полусумму значений текущей и следующей переменной и заносит результат в текущую переменную; 2) чтобы значения текущей переменной, означающей время, записанное в виде <*минуты*>.<*секунды*>, перевести в секунды следует использовать формулу *INT(x(i))*60+FRAC(x(i))*100*.

Генератор чисел. Выполнение пункта «*Генератор чисел*» приводит к появлению меню выбора типа генератора (рис. 3.8).

В этом меню предоставляется выбор из следующих возможностей:

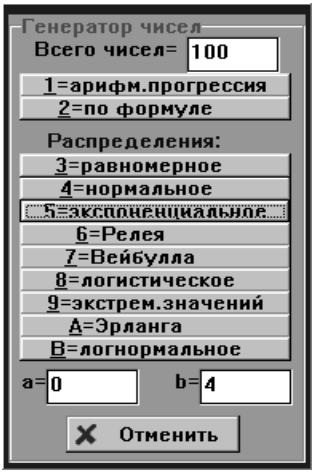


Рис. 3.8. Меню генератора чисел

- генерация чисел по закону арифметической прогрессии (возрастающей или убывающей) $a+b*i$, $i = 0, n-1$;
- генерация чисел по *задаваемой формуле* от аргумента *X*, где *X* изменяет свои значения по арифметической прогрессии $a+b*i$, $i = 0, n-1$;
- генераторы *случайных чисел,* распределенных по следующим законам:
  а) по *равномерному закону* в диапазоне от *a* до *b*;
  б) по *нормальному закону* со средним значением *a* и стандартным отклонением *b*;
  в) по другим законам распределения (см. разд. 12.3).

В верхнем поле ввода бланка необходимо предварительно указать количество *n* генерируемых чисел, а в двух нижних полях ввести значения параметров *a, b*, если они нужны для выбираемого генератора.

Отмена меню генератора происходит по нажатию клавиши Esc .

Выбор генератора по формуле приводит к появлению типового *бланка формул* (см. рис. 2.6), в который надо ввести новую формулу генератора или же выбрать одну из имеющихся в бланке формул и нажать Enter

или экранную кнопку «*Утвердить*». В формуле генератора допустимо использовать формальную переменную *x*, принимающую ряд последовательных значений от значения *a*, возрастающих с шагом *b*.

Например, чтобы получить переменную в виде суммы трех синусоид с уменьшающимся рядом периодов и амплитуд: 1, 1/2, 1/4, нужно сгенерировать три переменные по формулам: *SIN(x/100), SIN(x/50)/2, SIN(x/25)/4*, а затем сложить их значения по формуле: *x(i)+x(i+1)+x(i+2)*.

Кодирование. Операция *кодирования* означает замену значений выбранных переменных некоторым *кодом*, которым может быть как число, так и символьное обозначение (слово, текст). Такая замена производится только для тех значений переменных, которые удовлетворяют вводимому *логическому условию* (см. ниже).

Перед выполнением этой операции в меню преобразований необходимо в расположенное справа (от кнопки кодирования) поле ввода ввести числовой или символьный *код*, а после нажатия на кнопку «*Кодирование*» — указать подлежащие выборке переменные. Это производится в появляющемся типовом бланке выбора переменных (см. рис. 2.3).

После этого появляется типовой *бланк формул* (см. рис. 2.6), в который надо ввести логическое условие или же выбрать одно из имеющихся в бланке условий и нажать Enter или экранную кнопку «*Утвердить*».

В условиях, в дополнение к средствам раздела 2.3, можно использовать также отношения: равно, не равно, больше, меньше, больше или равно, меньше или равно (=, <>, >, <, >=, <=) и логические операции *И, ИЛИ, НЕТ* (&, |, ~). В отношениях *равно* и *не равно* в качестве операндов можно употреблять также и *символьные константы (литералы)*, заключенные в апострофы или кавычки. В этом случае будет проверяться полное текстуальное совпадение (несовпадение) операндов, например, логическое условие *x(1)='TRUE'* будет выполняться для всех значений первой переменной, которые являются литеральной константой (номинальным значением) *TRUE*.

В качестве переменных в формулах условий можно использовать не только обозначения, аналогичные вышерассмотренным в пункте «*Задаваемая функция*», но и формальные обозначения «*x*» для любого значения выбранных для кодирования переменных, а также обозначение «*j*» для номера строки. Примеры: а) логическое условие *(x>=0&x<=1)|(x>=10&x<=11)*, выполняется для всех значений выбранных для кодирования переменных, находящихся в диапазоне (0, 1) или в диапазоне (10, 11); б) логическое условие *j%3=0* выполняется для каждой третьей строки (% — операция «взять модуль»); в) логическое условие *(j<11)|(j>19)&(j<21)* выполняется для первого и третьего десятка строк.

Примечание. Замена значений кодом производится оперативно в электронной таблице, поэтому, если в список кодируемых переменных включены и переменные, фигурирующие в логических условиях, то на

некотором шаге (после их кодирования) это может привести к искажению результатов дальнейшего кодирования.

Нормировка значений выбранных переменных в электронной таблице может производиться двумя способами:

- *по диапазону значений*, когда из каждого значения переменной вычитается *минимальное значение* и результат делится на *диапазон значений* (т. е. на разность между максимальным и минимальным значениями) данной переменной, при этом все значения становятся положительными и не превышающими единицу;
- *по стандартному отклонению* (так называемая *нормализация* или *стандартизация*), когда из каждого значения переменной вычитается *среднее значение,* и это делится на *стандартное отклонение* данной переменной (полученные значения центрированы нулем).

После выбора операции нормировки в меню преобразований необходимо указать подлежащие нормировке переменные. Это производится в появляющемся типовом *бланке выбора переменных* (см. рис. 2.3).

Ранжирование. Операция *ранжирования* осуществляет замену числовых значений выбранных переменных на их *ранги*, т. е. на целые числа, являющиеся порядковыми номерами этих значений в ряду, упорядоченному по возрастанию значений. Совпадающие значения заменяются средними рангами, которые могут принимать дробные значения. В системе STADIA практически все ранговые статистические методы сами производят предварительное ранжирование переменных. Однако преобразование ранжирования может быть полезным в наглядных и вычислительных целях.

После выбора этой операции в меню преобразований следует указать подлежащие ранжированию переменные, что производится в появляющемся типовом *бланке выбора переменных* (см. рис. 2.3).

Сортировка. Данная операция позволяет *переупорядочить* измерения избранных переменных из электронной таблицы (*сортируемые* переменные) по возрастанию значений *сортирующих* переменных.

При выполнении этой операции появляется специальный экранный бланк (рис. 3.9), который включает следующие элементы:

- три вертикальных списка:
  - список переменных электронной таблицы;
  - список *сортирующих* переменных;
  - список *сортируемых* переменных;
- кнопки переноса переменных из одного списка в другой и обратно;
- кнопку выбора всех переменных в качестве сортируемых;
- фонарики режима сортировки: по возрастанию значений сортирующих переменных или по убыванию этих значений;
- кнопки утверждения и отмены выбора.

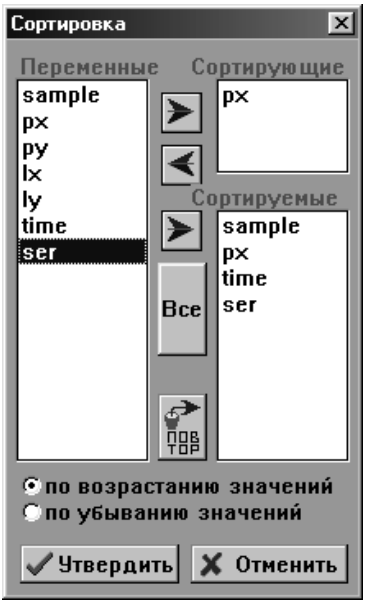


Рис. 3.9. Бланк сортировки значений переменных

Необходимо выбрать от одной до трех сортирующих переменных и не менее одной сортируемой переменной.

При наличии несколько сортирующих переменных сортировка имеет многоступенчатый характер:

- сначала она производится по значениям первой сортирующей переменной;
- затем в случае одинаковых значений первой сортируемой переменной для этих значений производится сортировка по значениям второй сортирующей переменной и т. д.

Для утверждения и завершения выбора (и начала сортировки) нажмите клавишу [Enter] или кнопку «*Утвердить*» на бланке.

Для отмены выбора и сортировки нажмите клавишу [Esc] или кнопку «*Отменить*» на бланке.

Отбор. Эта операция позволяет оставлять в электронной таблице только те измерения указанных переменных, которые удовлетворяют вводимому в бланк формул *логическому условию*, построенному аналогично операции кодирования.

Выборка. При выполнении статистического анализа часто возникает следующая задача: имеется достаточно объемная выборка, т. е. измеренные значения некоторой переменной. Однако исследователь не уверен, что эти значения получены достаточно *случайным образом* в ходе корректно организованных измерений. В таких случаях для повышения статистической *репрезентативности* полезно для анализа из такой выборки отобрать случайным образом некоторое подмножество значений (получить *подвыборку*).

Перед выполнением этой операции в меню преобразований необходимо в расположенное справа поле ввести число отбираемых случайным образом измерений (размер подвыборки), а после нажатия на кнопку «*Выборка*» — указать подлежащие выборке переменные, что производится в появляющемся типовом *бланке выбора переменных* (см. рис. 2.3).

Транспонирование. В результате *транспонирования* матрицы данных в электронной таблице строки становятся столбцами, а столбцы — строками.

Пример

Задача. В антропометрическом исследовании ста человек фиксировались их пол, рост, вес, возраст и другие показатели (файл PEOPLE). Однако по признаку пола данные были собраны вперемешку. При статистическом же анализе встала задача оценки различий между мужчинами и женщинами (см. пример разд. 6.4). Поэтому нужно рассортировать в электронной таблице данные по признаку пола.

Преобразования: Добиться нужного результата можно операцией сортировки, но она применима к количественным переменным, а здесь признак пола представлен номинальными значениями: *«м», «ж»*. Поэтому сначала нужно дважды выполнить операцию кодировки над переменной «пол» с кодами 1 и 2 и условиями поиска: *x='м'* и *x='ж'*. Затем можно выполнить операцию сортировки над всеми переменными по возрастанию значений сортирующей переменной «пол». В результате сначала (по строкам) будут идти измерения, относящиеся к мужчинам, а затем к женщинам. В заключение следует выделить фрагмент, относящийся к женщинам, удалить его в буфер обмена, а затем вставить содержимое буфера в свободные переменные электронной таблицы.

3.5. Пропуски и выбросы

Пропуски. Даже в прекрасно организованных и проведенных экспериментах (измерениях, сборах данных) некоторые наблюдения могут быть зарегистрированы неверно или не зарегистрированы совсем. Например, экспериментальное животное может погибнуть, пациент — не придти на назначенный прием, очередной препарат — оказаться испорченным, а регистрирующий прибор — отказать.

Ввод подобных *пропущенных* значений в электронную таблицу производится посредством набора любого *нечислового* значения, например, буквы или «?».

При работе статистических процедур все пропущенные значения автоматически игнорируются, т. е. не учитываются в вычислениях. При этом для методов, использующих всю матрицу данных или парные данные, игнорируются все наблюдения (строки), в которых встречается хотя бы один пропуск. Количество пропущенных и игнорированных (логически удаленных) значений указывается в заголовке выдачи результатов статистического анализа. Однако такое игнорирование может привести к потерям ценной для анализа информации, поэтому часто возникает необходимость каким-то образом заменить пропущенные значения, сохранив валидные парные или многомерные измерения.

Выбросы. Кроме этого, в анализируемых данных нередко присутствуют значения с большими отклонениями от среднего — так называемые *выбросы.* Они могут быть следствием некорректной регистрации данных

и существенно искажать результаты статистического анализа, а нередко могут приводить и к совершенно неверным выводам (см. в примерах к этому разделу). Поэтому такие данные необходимо обнаружить перед применением статистического метода, чувствительного к выбросам, и полностью их удалить или же некоторым образом заменить на значения, не нарушающие статистические закономерности.

Поиск пропущенных значений и выбросов осуществляется по операции «Пропуски и выбросы» *Блока преобразований* (см. рис. 3.6). При этом появляется экранное меню (рис. 3.10).

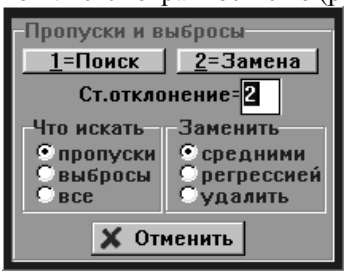


Рис. 3.10. Меню поиска и замены пропусков и выбросов

При выполнении операции поиска в экранную страницу результатов ⌊*Rez*⌋ выдается таблица *пропущенных значений* и *выбросов*. По этой таблице можно визуально оценить характер распределения пропущенных элементов в матрице данных. Над столбцами таблицы указаны номера переменных, а в строках — номера измерений. Пропущенные значения отмечены –?– , а выбросы — их величиной в стандартных отклонениях от среднего значения (см. *Пример 1*).

В меню (рис. 3.10) можно задать поиск только выбросов, только пропусков или того и другого. Там же можно задать нижнюю границу для поиска выбросов значением стандартного отклонения от среднего, превышение которого будет интерпретироваться как выброс.

Замена пропущенных значений и выбросов осуществляется нажатием кнопки «Замена» в меню рис. 3.10 и может быть произведена одним из двух методов (устанавливается фонариком режима в меню): заменой каждого пропущенного значения *средним значением*, вычисленным для соответствующей переменной, или заменой *регрессионными значениями*.

Второй метод применим к парным переменным или к многомерным данным и обеспечивает более корректную и дифференцированную замену, используя информацию о переменных, значения которых изменяются синхронно с анализируемой переменной. В соответствии с этим методом для каждой анализируемой переменной, содержащей пропущенные значения, выбирается парная переменная по условию максимума коэффициента корреляции. Затем по парной переменной вычисляется линейная регрессия с анализируемой переменной. Все пропущенные значения анализируемой переменной заменяются регрессионными значениями, вычисленными для соответствующих значений парной переменной в качестве аргумента полученной регрессионной модели. Однако если парная переменная в соответствующем измерении также содержит пропущенное значение, то замена производится по методу средних.

Установкой соответствующих фонариков в меню можно задать замену только выбросов, только пропусков или того и другого, а также произвести не их замену, а удаление. В последнем случае, если все переменные имеют одинаковую длину, то они могут являться парными переменными или многомерными данными, в таком случае при положительном ответе на вопрос будут удаляться целиком строки с пропусками/выбросами.

При выполнении замены выдача результатов поиска в экранную страницу результатов ⌊*Rez*⌋ не производится.

Следует также обратить внимание на то, что при наличии очень больших по величине выбросов средние и регрессионные значения могут быть сильно смещены, поэтому однократная замена может быть не стопроцентно эффективной. В таком случае следует повторно проанализировать данные на наличие выбросов и в случае необходимости повторно выполнить их замену.

Пример 1

Задача. В эксперименте были измерены значения четырех показателей у восьми животных, при этом некоторые значения не были зафиксированы (табл. 3.1, файл MIS):

Таблица 3.1. Измерения четырех показателей у 8 животных с пропусками

Объект	x1	x2	x3	x4
1	1.1	44.3	–2	0.3
2	?	48.1	?	0.97
3	2.8	51	0.1	0.78
4	8.399	?	44	–0.56
5	?	35	?	2
6	5.5	33	2	?
7	5	32.3	1.7	–3.4
8	8.899	–1.1	0	1.1

Произведем анализ пропущенных значений и выбросов:

Результат:

```
                Пропущн=6,   Выбросов=3
    1     2      3      4
2 | -?-          -?-
4 |        -?-    2
5 | -?-          -?-
6 |                     -?-
7 |                    -2.1
8 |       -2.1
```

Выводы: Выявлено наличие выбросов: в 8–м значении переменной x2, в 4–м значении переменной x3 и в 7–м значении переменной x4 с амплитудой –2,1, 2 и –2,1 стандартных отклонений.

Произведем замену пропущенных значений по методу средних и по методу регрессии, совместив их для наглядности (результаты выделены жирным шрифтом).

Результат:

	Метод средних				Метод регрессии			
	x1	x2	x3	x4	x1	x2	x3	x4
1	1.1	44.3	-2	0.3	1.1	44.3	-2	0.3
2	**5.28**	48.1	**7.63**	0.97	**2.46**	48.1	**7.633**	0.97
3	2.8	51	0.1	0.78	2.8	51	0.1	0.78
4	8.39	**34.65**	44	-0.56	8.399	**8.504**	44	-0.56
5	**5.28**	35	**7.63**	2	**4.238**	35	**7.633**	2
6	5.5	33	2	**0.17**	5.5	33	2	**0.289**
7	5	32.3	1.7	-3.4	5	32.3	1.7	-3.4
8	8.89	-1.1	0	1.1	8.899	-1.1	0	1.1

Пример 2

Задача. На кафедре ВНД МГУ исследовался условный рефлекс на болевой стимул у кошки с измерением его латентного периода до и после введения блокатора *D1*-рецепторов SCH 23390 (табл. 3.2, файл LATPER)[1]:

Таблица 3.2. Латентность реакции кошки на стимул до и после введения блокатора боли

До	575 650 450 475 550 500 500 500 500 650 525 550 550 500 625 2525 600 475 500 450 550 1850 500 475 575
После	550 625 650 750 575 700 600 775 775 750 625 1050 600 900 700 525 625 1125 700 725 825

Попробуем выявить различия между средними значениями латентных периодов по параметрическому критерию Стьюдента (см. разд. 6.4) и непараметрическому критерию Вилкоксона (см. разд. 7.2).

Результаты:

```
КРИТЕРИЙ ФИШЕРА И СТЬЮДЕНТА.  Файл:latper Переменные: до, после
Статистика Стьюдента=0.574, Значимость=0.5757, степ.своб = 44
   Гипотеза 0: <Нет различий между выброчными средними>

КРИТЕРИИ СДВИГА (ПОЛОЖЕНИЯ). Файл:latper Переменные: до, после
Вилкоксон=406, Z=4.016, Значимость=2.987E-5, степ.своб = 25,21
   Гипотеза 1: <Есть различия между медианами выборок>
Ван дер Варден=-11.3, Z=-3.58, Значимость=0.00017, степ.своб=25,21
   Гипотеза 1: <Есть различия между медианами выборок>
```

Выводы: Анализ дает нам парадоксальные результаты. Критерий Стьюдента не обнаруживает различий с высокой степенью достоверности (P=0.5757), а два непараметрических критерия выявляют различия с высоким уровнем значимости, близким к нулю. В чем здесь дело?

Если же мы повнимательнее присмотримся к исходным данным, то обнаружим в них несколько значений, заметно отличающихся от большинства других по величине. Для более точного исследования этих отклонений выполним процедуру поиска пропусков и выбросов:

[1] Данные предоставлены В.И. Майоровым.

Р е з у л ь т а т ы:

```
Пропущн=0,   Выбросов=4
     i         до     после
    12|                  2.1
    16|       3.9
    18|                  2.6
    22|       2.5
```

В ы в о д ы: В исходных данных присутствуют 4 выброса, величиной более двух стандартных отклонений, что привело к сильному сдвигу средних значений двух выборок, поэтому чувствительный к выбросам критерий Стьюдента не обнаружил их различий. Следовательно, необходимо удалить эти экстремальные значения из исходных данных и повторить анализ.

Р е з у л ь т а т ы:

```
КРИТЕРИЙ ФИШЕРА И СТЬЮДЕНТА.  Файл: latper
Переменные: до, после
Статистика Стьюдента=5.872, Значимость=1.523E-5, степ.своб = 40
   Гипотеза 1: <Есть различия между выборочными средними>
```

В ы в о д ы: После удаления выбросов критерий Стьюдента не подтверждает нулевую гипотезу об отсутствии различий средних, так как ранее это показали нечувствительные к выбросам непараметрические критерии.

Глава 4

ГРАФИЧЕСКИЕ СРЕДСТВА

«Когда субъект не справляется с притоком новых данных естественным путем концептивного аналогизирования,
он становится жертвой перцептивного аналогизирования.
Этот процесс известен также под названием
"метафорическая деформация".
Теперь вам ясно?»
[Отис Бландерс Клент]

Графическое представление данных и результатов является важнейшим средством визуального анализа данных, поскольку человеческий глаз во взаимодействии с головным мозгом является точнейшим и мощнейшим аналитическим инструментом, способным увидеть и оценить то, что не доступно никакому математическому алгоритму. В этом плане пакет STADIA имеет исчерпывающие современные возможности, просто и доступно организованные.

Данная глава содержит общий обзор средств графического вывода, доступных из различных процедур и организованных по единым принципам с возможностью их гибкого комбинирования и взаимодействия. Однако эти средства не претендуют на уровень презентационной графики, для чего предназначены специальные графические редакторы типа *Corel Draw, PhotoShop, HarwardGraphics* и другие.

4.1. Графический диалог

Использование средств графического представления возможно для:

а) исходных *данных*;

б) *результатов* анализа.

Построение *графиков данных* производится по нажатию клавиши [F6] (или исполнением пункта *«Графики»* из верхней командной строки), а построение *графика результатов* анализа — при выполнении конкретного статистического метода. В обоих случаях график выдается в отдельную *экранную страницу* $\lfloor Gr_i \rfloor$, , i = 1–15.

Графические страницы. Допускается создание до 15 *графических страниц*. После вывода на 15-ю страницу следующий вывод производится на первую графическую страницу, замещая имеющийся там график. Чтобы не терять в результате этого нужную информацию, следует своевременно *удалять* ненужные графические страницы. Этого можно сделать нажатием на кнопку закрытия окна, но проще нажать *быструю* клавишу [F5], когда страница активна.

Графические страницы можно *перемещать* (например, для сопоставления графиков). Для этого подведите указатель мыши к ярлыку некоторой страницы, нажмите ***правую*** кнопку мыши и, не отпуская ее, ведите указатель (он изменит свою привычную форму на стрелку с пачкой листков) до ярлыка страницы назначения. Здесь отпустите кнопку мыши.

Для визуального сопоставления графиков полезна их одновременная визуализация посредством команды *«Каскад»* из пункта *«Окна»* верхней командной строки (см. рис. 2.2).

Дополнительные инструменты. При активизации экранной страницы с графиком в третьей инструментальной строке экрана появляется ряд дополнительных *кнопок* общего графического назначения, посредством которых можно модифицировать уже построенный график:

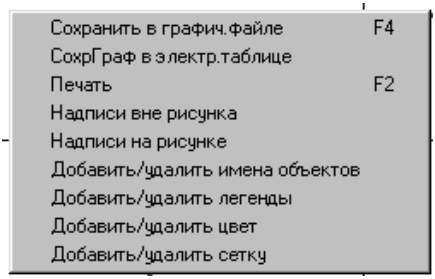
а именно (слева направо):

- кнопка «*СохрГраф*» сохранения данных с графика в электронной таблице;
- кнопка изменения толщины линий на графике (действует не на все типы графиков);
- кнопка переключения: цветное/черно–белое изображение (полезно использовать перед выводом графика на принтер или для вставки рисунка в статью);
- кнопка добавления подрисуночных надписей и легенд / наименований объектов;
- кнопка добавления/снятия координатной сетки.

Наряду с этим, отдельно для категорий научной графики и деловой графики контекстно появляются еще и дополнительные инструментальные кнопки (см. ниже).

Кнопка «СохрГраф». Кнопка *сохранения графика* является чрезвычайно удобным средством для перенесения графических данных в *электронную таблицу* для дополнительного анализа и визуализации, поскольку в этом случае для сохраненных данных становятся доступны все другие возможности пакета. При этом в электронную таблицу (в первые ее свободные переменные) заносятся координаты *X, Y* всех точек с графика. Особенно это полезно для сохранения результатов уже проведенного анализа или графиков данных, полученных в результате дополнительных вычислений (сплайны, сглаживание и т. п.).

Кнопка надписей. Кнопка выполнения *подрисуночных надписей* вызывает экранный бланк, в который можно ввести необходимые надписи по осям и под рисунком, а также заказать вывод легенд и наименований объектов. *Легенды* представляют собой список *маркеров* кривых с их наименованиями, располагаемый справа от графика (см. ниже).



Контекстное меню. При нажатии на правую кнопку мыши в активном графике появляется контекстное меню, команды которого дублируют ряд операций: сохранение графика в графическом формате, перенос с графика координат в электронную таблицу, вывод графика на печать, создание надписей, добавление имен объектов, легенд, координатной сетки, изменение цвета.

Надписи на рисунке

Бланк ввода надписей (рис. 4.1) для уже построенного графика вызывается нажатием кнопки «*Надписи*». Этот бланк содержит следующие элементы:

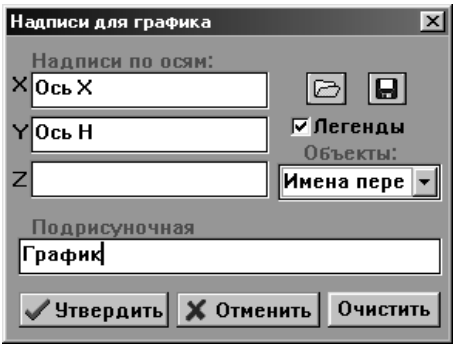


Рис. 4.1. Бланк ввода подрисуночных надписей

- четыре горизонтальных поля ввода надписей по осям *X, Y, Z* и подрисуночной надписи;
- фонарик разрешения вывода *легенд* на рисунок;
- выкидной список для выбора переменной, содержащей наименования объектов;
- кнопка чтения надписей из дискового архива;
- кнопка записи текущих надписей в дисковый архив;
- кнопка очистки полей бланка;
- кнопка *«Утвердить»* для вывода произведенных установок на рисунок (дублируется клавишей [Enter]);
- кнопка отмены бланка без утверждения установок (дублируется клавишей [Esc]).

Введенные в бланк надписи сохраняются в полях ввода данного бланка от вызова к вызову и от сеанса к сеансу.

Архив надписей. Чтобы не вводить каждый раз типовые надписи, имеется возможность хранить надписи в *дисковом архиве*. Для этого предназначены две кнопки с *пиктограммами* чтения и записи. Первая операция позволяет *считывать* в экранный бланк надписи из архивных файлов, а вторая — *записывать* текущие надписи в архивный файл указанного наименования. Порядок выполнения этих операций идентичен операциям чтения и записи файлов данных (см. разд. 3.2), за исключением того, что в оглавление архива выдается только список файлов надписей (тип STS).

Легенды. *Легенды* представляют собой пояснения к различным компонентам графика, располагаемые справа от него. Для *научной графики* легенды маркируют различные кривые с указанием имен соответствующих им *Y*–переменных. Для *диаграммы рассеяния* легенды маркируют отдельные точки (объекты). Для *деловой графики* легенды (имена объектов или переменных) зависят от расположения осей. При этом для круговой диаграммы легенды присутствуют в виде наименований секторов.

Имена объектов. Часто на графике необходимо отобразить обозначения или номера объектов из электронной таблицы (например, на диаграмме рассеяния). Такие имена могут храниться в левом объектном столбце или в любой переменной в электронной таблице. Тогда эту переменную полезно указать в поле «*Объекты*» бланка. При выборе таких обозначений следует стремиться к их компактности, иначе обозначения будут на графике накладываться друг на друга. При отсутствии указания переменной с именами объектов в ряде случаев, по умолчанию, используются порядковые номера объектов.

Средства редактирования

В рамках активной *графической страницы* можно выполнять ряд операций *редактирования* изображения, связанных с изменением его размеров, перемещением или удалением отдельных компонентов, набором дополнительных надписей и др.

Изменить размеры. Чтобы *изменить размер* графика, необходимо:

1) переместить указатель мыши к одной из ограничивающих график координатных осей (для *круговой диаграммы* условные оси находятся за ее пределами), при этом форма указателя изменится на двунаправленную стрелку;
2) нажать левую кнопку мыши и вести границу в ее новое место, после чего отпустить кнопку мыши.

Чтобы переместить сам график в другое место (достаточно редко выполняемая операция), нужно последовательно переместить его границы.

Операции с фрагментами рисунка. Чтобы *выделить* некоторый *фрагмент* (участок) на рабочей *графической странице* для последующего перемещения или удаления, необходимо подвести указатель мыши в предполагаемую левую верхнюю точку фрагмента, нажать левую кнопку мыши и, не отпуская кнопку, тянуть появившийся пунктирный четырехугольник к предполагаемой правой нижней точке фрагмента.

Работа с буфером обмена. Выделенный фрагмент можно удалить или скопировать в *буфер обмена*, а затем его можно вставить в любом месте графика. Такими средствами через буфер обмена можно перемещать на графике различные его элементы (см. также разд. 2.5).

Чтобы *отменить* выделение фрагмента пунктирным четырехугольником, следует нажать левую кнопку мыши в любом другом месте страницы.

Надписи с клавиатуры. Чтобы сделать дополнительную надпись на самом рисунке, нужно произвести выделение в требуемом месте (т. е. движением мыши с нажатой левой клавишей построить пунктирный прямоугольник нужного размера), нажать на клавишу “пробел” и в появившемся бланке ввести текст надписи установленным шрифтом. Это следует делать в завершение оформления графика (перед выводом на печать или сохранением в графическом файле), поскольку надписи на рисунке не сохраняются при его перестроениях.

Общие операции

В рамках активной *графической страницы* действует ряд общих операций, доступных из верхней экранной строки команд, дублируемых горячими функциональными клавишами, а также постоянными инструментальными кнопками в третьей экранной строке:

- *чтение* графического файла из дискового архива в активную графическую страницу;
- *запись* графической страницы в виде графического файла в дисковый архив;
- операции с *буфером обмена*: вырезка, копирование, вставка, удаление выделенного участка изображения;
- *печать* активной графической страницы на принтере;
- переустановка текущего графического *шрифта*.

Чтение и запись графических файлов производится в формате *BMP* (диалог чтения/записи аналогичен разд. 3.2).

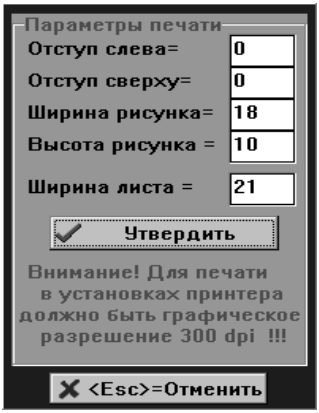


Рис. 4.2. Бланк ввода размеров рисунка на печати

Печать. При выполнении операции печати графика появляется экранный бланк (рис. 4.2), в котором надо ввести в сантиметрах размеры и положение рисунка (отступы от левого и верхнего края листа, ширину и высоту, но не более формата A4), а также указать ширину листа бумаги в сантиметрах.

Расположение листа печати (книжное или ландшафтное) изменяется в установках принтера (команда *«Принтер»* пункта *«Файлы»* верхней командной строки).

Внимание: по техническим причинам для правильного вывода на печать графиков для принтера должно быть установлено графическое разрешение 300 dpi.

Изменение шрифта. При выдаче на рисунок надписей могут возникнуть различные шрифтовые проблемы, которые следует решать посредством переустановки шрифта для графических страниц (см. разд. 2.2).

Графики данных

Чтобы построить *график данных,* нужно нажать клавишу F6 (или же выполнить пункт «*График*» в верхней командной сроке), что приводит к вызову *меню выбора типа графика* данных (рис. 4.3).

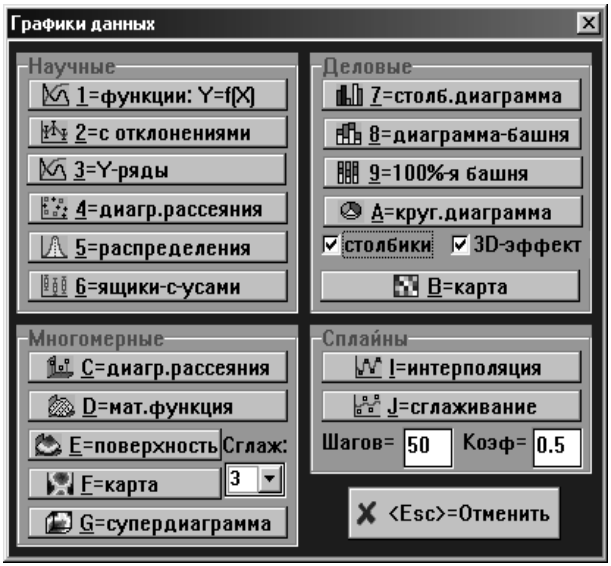


Рис. 4.3. Меню выбора графика данных

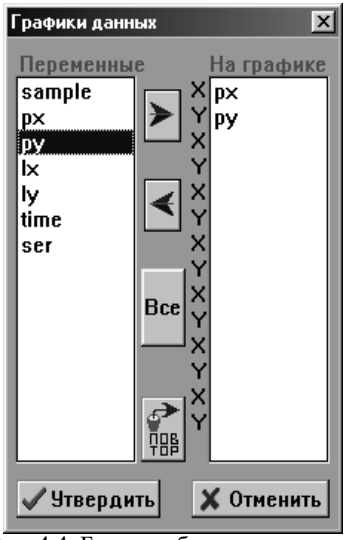


Рис. 4.4. Бланк выбора переменных для графика

В этом меню следует выбрать нужный тип графика посредством нажатия на соответствующую экранную кнопку или же на сопоставленную ей клавишу быстрого вызова. Доступные в STADIA графические формы разбиты на четыре группы: научная графика, деловая графика, многомерная графика и сплайны (являются специальным типом научной графики).

Масштабирование графика и разметка осей производятся автоматически. В конце координатных осей могут быть указаны масштабные десятичные множители вида: **Ei*, где *i* — показатель десятичной степени.

Далее в появившемся *бланке выбора переменных* (рис. 4.4) следует отобрать переменные из электронной таблицы, отображаемые на графике (выбор таких переменных зависит от типа графика, см. также по-

яснения к рис. 2.3).

4.2. Научная графика и сплайны

Научная графика включает наиболее употребительные в научных и инженерных исследованиях формы двухмерных графиков:

1) *функциональный график*, изображающий одну или несколько экспериментальных или теоретических зависимостей вида *Y=f(X)*, представленных множеством пар значений переменных *X, Y*;
2) *график с отклонениями*, аналогичен функциональному, но в нем каждое значение *Y* интерпретируется как некоторое *среднее значение*, и для него еще указывается третья переменная *dY*, представляющая *интервал ошибки* или *стандартное отклонение* значений;
3) *Y-ряды* аналогичны функциональным графикам, но не в зависимости от некой другой переменной, а в порядке своих значений, в этой же форме удобно представлять и *временные ряды*;
4) *диаграмма рассеяния (скаттерграмма)*, где данные интерпретируются не как зависимость *Y=f(X)*, а как множество пар значений *X,Y*, каждая из которых представляет координаты некоторой точки в двухмерном пространстве;
5) *график распределения* выборочных значений представляет изображение одной или нескольких выборок (переменных) в порядке возрастания их значений;
6) *ящики с усами* являются модификацией получившего распространение способа компактного совместного изображения многих выборок с указанием их средних значений и стандартных отклонений;
7) сплайны можно отнести к специальному разделу научной функциональной графики, где промежутки между экспериментальными точками сглаживаются/интерполируются посредством специальных кубических парабол — сплайнов:
   а) сплайн–*интерполяция* обеспечивает прохождение сплайнов непосредственно через заданные точки;
   б) сплайн–сглаживание обеспечивает прохождение сплайнов на некотором удалении от заданных точек с меньшими колебаниями.

При активизации экранной страницы с научным графиком в третьей экранной строке появляется две дополнительные инструментальные *кнопки*:

1) кнопка изменения *формы* текущего графика имеет четыре состояния, циклически изменяемые при каждом нажатии:
   - график в виде линий;
   - график в виде линий с маркерами точек;
   - график в виде маркированных точек (*диаграмма рассеяния*);
   - график в виде столбиков (*столбиковая диаграмма*);

2) кнопка *номера* зависимости *Y=f(X)* устанавливает одну из нескольких экспонируемых на графике зависимостей в качестве текущей (форма которой изменяется первой кнопкой).

Этими кнопками можно изменить форму представления любой зависимости на функциональном и некоторых других типах графиков.

Функциональный график. В экранном *бланке выбора переменных* необходимо указать одну или несколько пар переменных *X, Y*, представляющих экспериментальные зависимости *Y=f(X)* (рис. 4.5).

В бланке выбора можно указать и одну переменную *Y*, тогда по оси абсцисс будут расположены порядковые числа значений *Y*.

Y-ряды. В *бланке выбора переменных* необходимо указать одну или несколько переменных *Y*. Вид графика аналогичен функциональному, однако ось *X* представляет порядковые номера значений *Y*. Форма графиков может быть также изменена вышеупомянутыми функциональными кнопками.

График с отклонениями. В *бланке выбора переменных* необходимо указать одну или несколько триад переменных: *X, Y* и отклонения *dY* (рис. 4.6).

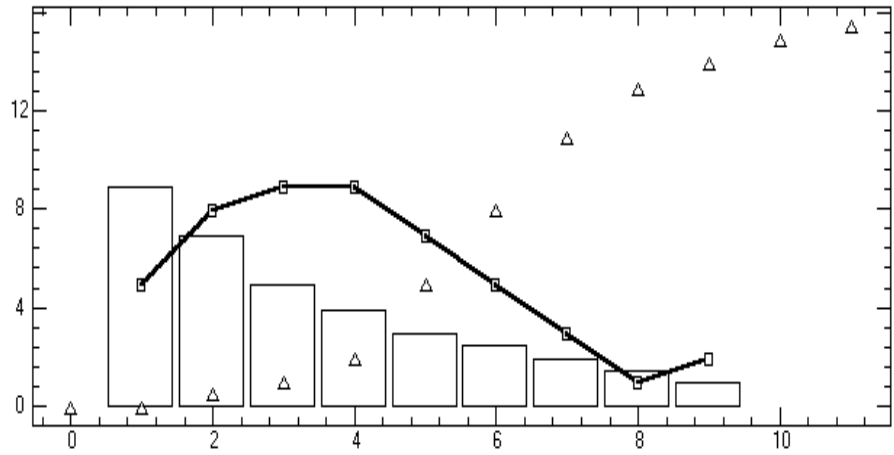


Рис. 4.5. Три функциональных графика в трех формах изображении: линиями, точками (диаграмма рассеяния) и столбиками (столбиковая диаграмма)

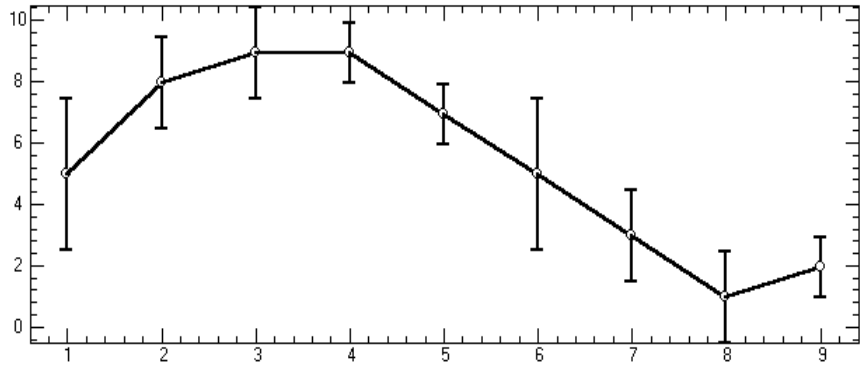


Рис. 4.6. Функциональный график с отклонениями

Диаграмма рассеяния. В *бланке переменных* необходимо указать одну или несколько пар переменных, соответствующих координатам *X* и *Y* экспонируемых на графике точек.

После вывода графика для каждой точки можно добавить наименования объектов. Для этого следует вызвать *бланк подрисуночных надписей* (см. рис. 4.1) нажатием на соответствующую инструментальную графическую кнопку и в этом бланке установить признак *«Легенды»*. Можно также вместо номеров проставить *символьные обозначения* точек, установив в бланке надписей *переменную–маркер*.

График распределения (Кетле). В *бланке переменных* необходимо указать одну или несколько переменных, представляющих экспонированные на графике выборки.

Значения переменных будут упорядочены по возрастанию и изображены как функция их порядкового номера (рис. 4.7).

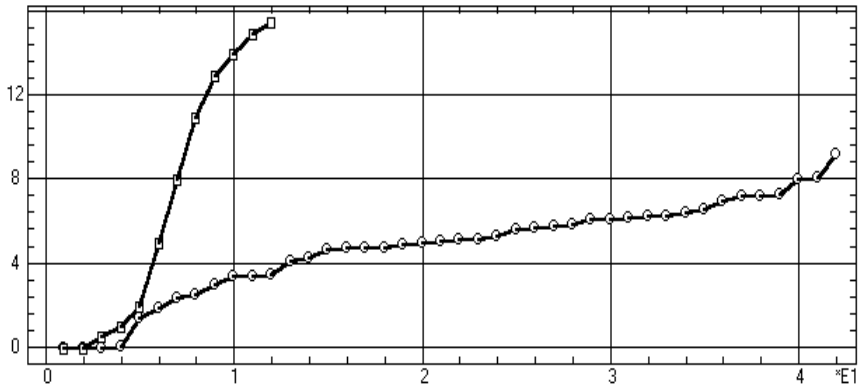


Рис. 4.7. График распределения значений двух выборок (график Кетле)

Ящик с усами. В *бланке выбора переменных* необходимо указать не менее двух переменных, представляющих экспонированные на графике выборки.

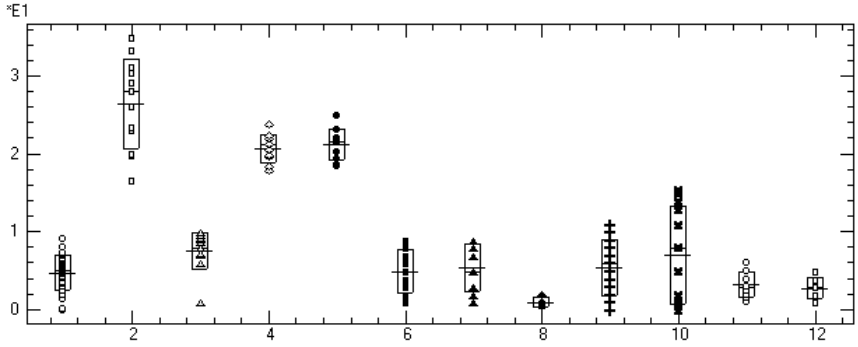


Рис. 4.8. «Ящики с усами» для 12 выборок

Значения каждой выбранной переменной представляются в виде серии точек в отдельной вертикальной колонке (рис. 4.8). Для каждой такой колонки рисуется прямоугольник, соответствующий *стандартному отклонению* для этой переменной (в положительную и отрицательную стороны), внутри которого горизонтальными линиями указываются положения *среднего значения* (более длинная линия) и *медианы* переменной.

Сплайны. *Кубические сплайны* вида: $s = a_0+a_1*x+a_2*x^2+a_3*x^3$ используются для достижения плавного перехода между заданными точками экспериментальной функции с непрерывностью первой и второй *производных* во всей области (рис. 4.9). Имеется два вида таких графиков:

- график *сплайн–интерполяции* зависимости *Y* от *X*: сплайн проходит точно через заданные точки;
- график *сплайн–сглаживания* зависимости *Y* от *X*: сплайн слаживает экспериментальную зависимость, проходя между ее точками.

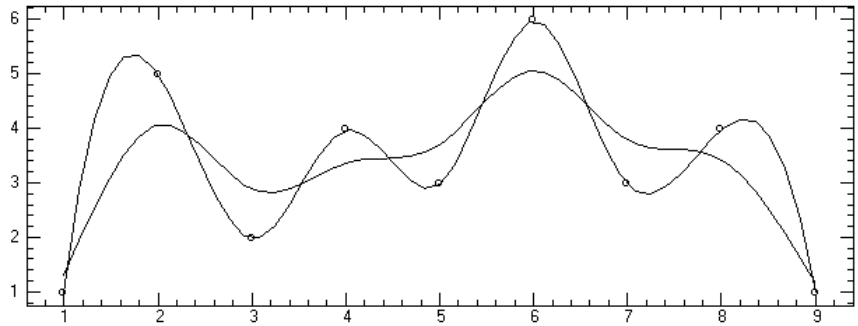


Рис. 4.9. Сплайны интерполяции и сглаживания (круги — экспериментальные точки)

Перед нажатием на кнопку сплайн–графика в нижерасположенном поле ввода (см. рис. 4.3) необходимо указать число точек, в которых будут вычислены значения сплайна, в интервале между каждыми двумя экспериментальными точками.

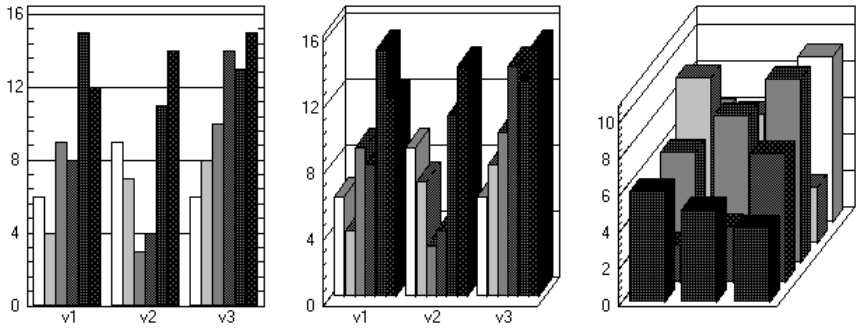
Для графика сплайн–сглаживания в соседнем поле ввода дополнительно необходимо указать значение *коэффициента сглаживания*. Этот коэффициент указывает среднее расстояние, на которое сглаживающий сплайн будет отстоять от заданных точек. Чем ближе этот коэффициент к нулю, тем ближе к экспериментальным точкам будет проходить сплайн. Значение коэффициента, равное 1, соответствует одному *среднеквадратичному отклонению* для экспериментальных значений *Y*. Сплайн–сглаживание, в частности, оказывается очень эффективным для сглаживания сильно зашумленных временных рядов и экспериментальных зависимостей (см. в примере к разд. 9.4).

После выбора сплайн–графика в последующем *бланке выбора переменных* необходимо выбрать две переменные из электронной таблицы в качестве *X* и *Y* (можно указать и одну переменную *Y*, тогда по оси абсцисс будут порядковые номера значений).

4.3. Деловая графика

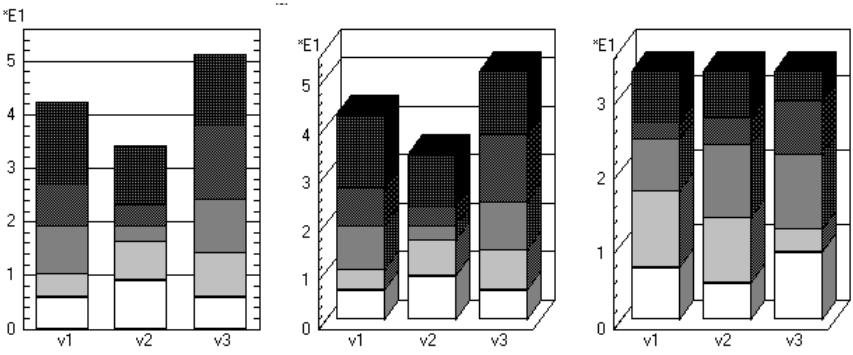
Раздел *деловой графики* объединяет наиболее употребительные формы изображения данных *гуманитарного* и *экономического* характера, представленных в виде матрицы со значениями нескольких переменных (столбцы), измеренных у ряда объектов (строки). При этом преимущественно представляет интерес визуальное сравнение различных объектов по значению одной или нескольких переменных.

Из меню графиков данных можно выбрать следующие типы диаграмм:



а *б* *в*

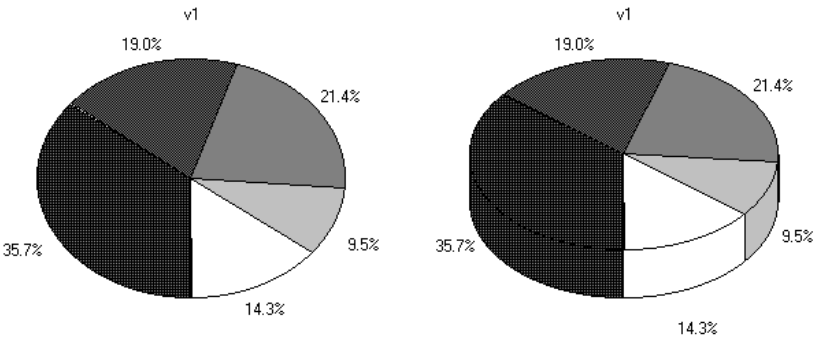
Рис. 4.10. Столбиковые диаграммы: *а* — без трехмерного эффекта; *б* — с трехмерным эффектом; *в* — трехмерная столбиковая диаграмма



а *б* *в*

Рис. 4.11. Башенные диаграммы:
а — без трехмерного эффекта; *б* — с трехмерным эффектом; *в* — 100%–ная

- *столбиковая диаграмма* (рис. 4.10) обеспечивает последовательно линейное расположение значений объектов для каждой последующей переменной;
- *диаграмма–башня* (рис. 4.11), в отличие от столбиковой, изображает значения объектов для каждой последующей переменной;
- *100%*–ная *башня* (рис. 4.11) представляет вариант диаграммы–башни, у которой каждая вертикальная колонка (объект) нормирована на 100%;
- *круговая диаграмма* (рис. 4.12) изображает соотношение значений некоторой переменной для ряда объектов в виде секторов круга;
- *карта* (рис. 4.13) изображает соотношение значений переменных (по горизонтали) для ряда объектов (по вертикали) в форме матрицы, где значения (клетки матрицы) кодируются в цветной или черно-белой шкале.



а *б*

Рис. 4.12. Круговая диаграмма: Ъ
а — без трехмерного эффекта; *б* — с трехмерным эффектом

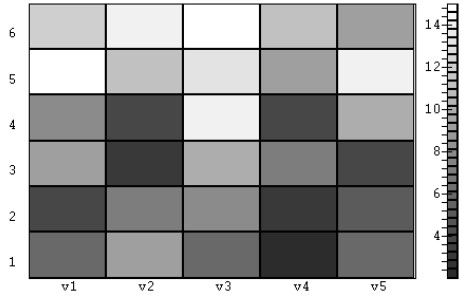


Рис.4.13. Карта

Кроме этого, можно уточнить исходную форму представления каждой из диаграмм посредством двух специальных переключателей, находящихся внизу панели деловой графики (см. рис 4.2):

- переключатель «*столбики*» определяет, будет ли выбранная диаграмма изображаться столбиками или же линиями (лентами);
- переключатель «3D–эффект» определяет, будут ли координатная рамка и сами диаграммы иметь трехмерный эффект;

После выбора типа графика в появляющемся экранном *бланке переменных* необходимо указать переменные, экспонируемые на графике.

Графические признаки и инструменты.

В разделе деловой графики доступно свыше 30 различных форм диаграмм, которые классифицируются по четырем признакам: накопленность, форма, трехмерность и оси.

Эти признаки управляются дополнительными инструментальными кнопками с пиктограммами, контекстно появляющимися в третьей экранной строке при активизации страницы с деловым графиком. Для круговой диаграммы появляется еще пятая дополнительная инструментальная кнопка — номер объекта.

Таким образом, независимо от исходной формы графика, в экранной странице можно его изменить, пользуясь четырьмя–пятью дополнительными инструментами, которые рассмотрим подробнее.

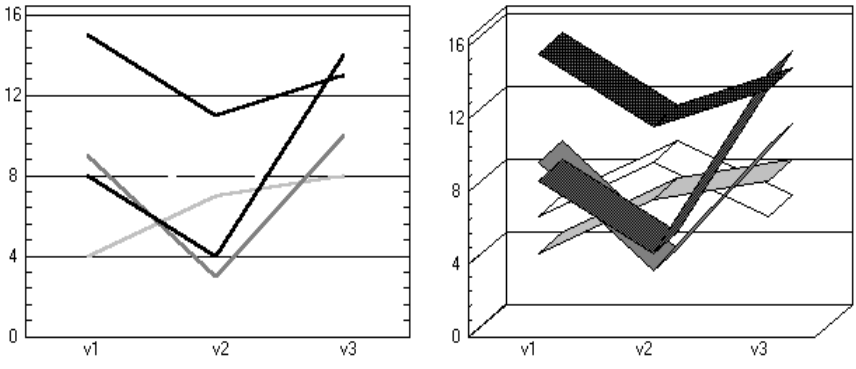
Накопленность. Этот признак–инструмент имеет три градации, циклически изменяемых при каждом нажатии:

- *столбиковая диаграмма* (рис. 4.10);
- *диаграмма–башня* (рис. 4.11);
- *100%*–ная *башня* позволяет сопоставить процентные представимости каждой переменной у разных объектов.

Форма представления. Этот признак–инструмент также имеет три градации, циклически изменяемые при каждом нажатии:

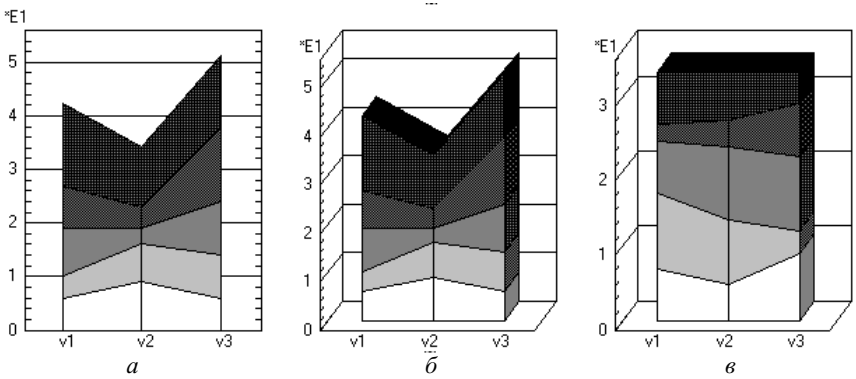
- *столбиковая* форма (рис. 4.10, 4.11);
- *линейная* форма (рис. 4.14, 4.15) изображает изменение значений каждой переменной у последовательных объектов в виде обычного графика, состоящего из линий (или ленточек при трехмерности);

- форма круговой диаграммы (рис. 4.12).



а *б*

Рис. 4.14. Ленточные диаграммы:
а — без трехмерного эффекта; *б* — с трехмерным эффектом



а *б* *в*

Рис. 4.15. Ленточные накопленные диаграммы:
а – без трехмерного эффекта; *б* – с трехмерным эффектом; *в* – 100%–ная

Трехмерность. Этот признак–инструмент со значениями *Да/Нет* определяет наличие у координатной рамки и самих диаграмм трехмерного эффекта.

Для столбиковой диаграммы со столбиковой или линейной формой представления этот признак имеет еще одну дополнительную градацию — изображение диаграммы в виде реально трехмерного графика (см. рис. 4.10).

Оси. Этот признак–инструмент имеет два значения:

- на диаграмме по горизонтали располагаются объекты;
- на диаграмме по горизонтальной оси располагаются переменные.

Исходное значение признака «*Оси*» обеспечивает расположение столбиковых и башенных диаграмм в порядке переменных и изображение круговой диаграммы со значениями первой выбранной переменной для всех объектов.

В этом случае под горизонтальной осью располагаются наименования переменных, а справа от графика в качестве *легенд* — наименования объектов (если легенды и наименования объектов заданы в бланке надписей). Для круговой диаграммы у секторов указаны наименования объектов с их процентными соотношениями, а над диаграммой — наименование текущей переменной (процентное содержание секторов менее 1% не указывается).

При смене осей позиции наименований переменных и объектов взаимно меняются. В случае круговой диаграммы при повторном нажатии на кнопку «Оси» диаграмма изображает или значения объектов для избранной переменной или же значения переменных для избранного объекта.

Объект. Этот признак–инструмент с цифрой–номером принадлежит только круговой диаграмме и в зависимости от значения признака «*Оси*» определяет:

- порядковый номер объекта, значения переменных которого представлены на диаграмме;
- порядковый номер переменной, значения которой в порядке объектов представлены на диаграмме.

При нулевом значении этого признака выводится суммарная круговая диаграмма, где каждый сектор получен суммированием значений по соответствующему измерению.

Комбинируя значения рассмотренных признаков, можно уже после построения графика подобрать для него наиболее приемлемую форму из более 30 доступных вариантов.

4.4. Трехмерная графика

Трехмерная графика объединяет формы представления многомерных данных, главным образом, из научных и технических приложений:

- *диаграмма рассеяния* множества триад значений *X, Y, Z*, представляющих координаты точек в трехмерном пространстве;
- *поверхность* представляет изображение поверхности в трехмерном пространстве, заданной алгебраической формулой;
- *поверхность сглаживания* представляет изображение поверхности в трехмерном пространстве, сглаживающую множество точек, заданных координатами *X, Y, Z*;
- *картирование* аналогично сглаживанию, но результат представляется в виде двумерной карты, на которой высоты сглаживающей поверхности представлены в цветной или черно–белой тональной шкале;

- *супердиаграмма* предназначена для визуализации экстрамногомерных данных (более трех измерений).

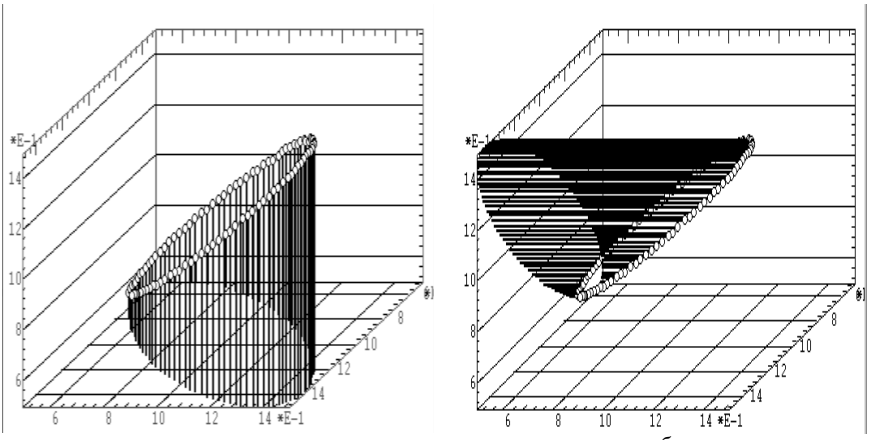
На последние три типа графиков действуют инструментальные кнопки изменения цвета и надписей, но не действуют кнопки толщины линий и сетки, а установка легенд не имеет эффекта. Перерисовка этих графиков в отдельных случаях может занимать несколько секунд.

Диаграмма рассеяния (скаттерграмма). Трехмерная диаграмма рассеяния (рис. 4.16) изображает множество триад значений *X, Y, Z* в виде точек в трехмерном пространстве.

Для ее построения в экранном бланке выбора перемененных (см. рис. 4.4) необходимо указать одну или несколько троек переменных, соответствующих координатам *X, Y, Z*.

После этого в третьей инструментальной строке появляются еще две дополнительные *кнопки*:

Вторая слева кнопка «*Проекция*», сменяющая при нажатии свое обозначение на «*X*», «*Y*», «*Z*», «*», «+», «*0*», изменяет проекцию следов точек и их связи, соответственно: проекции на одну из трех координатных плоскостей (рис. 4.16, *а, б*); связывая точки с геометрическим центром; связывая точки с началом координат; связывая точки между собой с ближайшим соседом.



а *б*

Рис. 4.16. Трехмерная диаграмма рассеяния с проекцией следов точек:
а — на плоскость X–Y; *б* — на плоскость Y–Z

Вращение графика. Первая из этих кнопок «*Вращение*» включает/выключает режим вращения графика. В режиме вращения график приобретает вид, изображенный на рис. 4.17. Здесь исходная вращаемая система координат обозначена зеленым цветом, а проекции точек указаны в новой неподвижной на рисунке системе координат, изображенной черным цветом.

В режиме вращения на графике появляются шесть управляющих кнопок, они управляют вращением графика вправо или влево относительно трех координатных осей.

Седьмая кнопка «*СохрВращ*» сохраняет результат вращения при возврате к диаграмме рассеяния.

Поэтому, чтобы сохранить результат вращения в электронной таблице, следует: 1) нажать кнопку «*СохрВращ*»; 2) снять режим вращения; 3) нажать кнопку «*СохрГраф*». Такой прием можно применить и перед построением поверхности сглаживания и для многих других целей.

Нижерасположенное поле ввода позволяет менять угловой шаг при нажатии на шесть кнопок вращения.

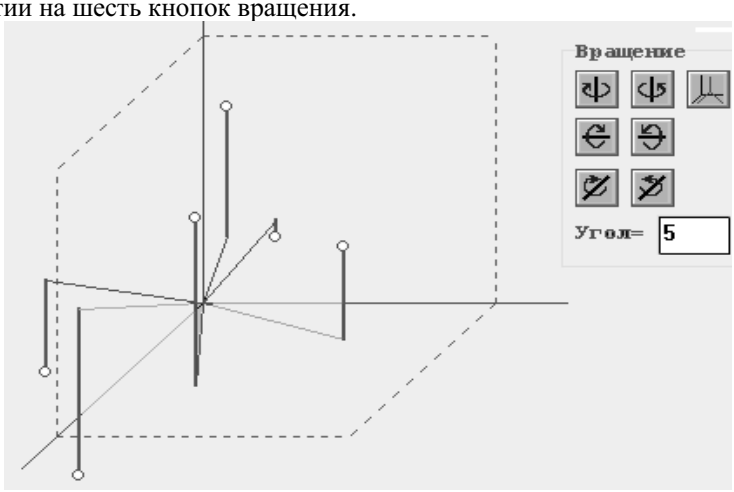


Рис. 4.17. Трехмерная диаграмма рассеяния в режиме вращения

Поверхность сглаживания. Данный тип графика представляет собой результат *интерполирующего сглаживания* множества точек в трехмерном пространстве, заданных координатами *X, Y, Z* переменные для которых необходимо выбрать в типовом бланке (см. рис. 4.4). Сглаживание выполняется по *полевому* методу (см. ниже: супердиаграмма). Сглаживающая поверхность располагается в *аксонометрической* проекции.

На рис. 4.18 приведена поверхность, сглаживающая следующее распределение точек:

```
X:  1,   1,   2,   2,  2,   3,  3,  4;
Y:  1,   3,   1,   2,  4,   3,  4,  2;
Z: 20, 100, 100, 200, 20, 150, 50, 20.
```

Для обзора поверхности с другой стороны нужно произвести вращение исходного множества точек, используя средства диаграммы рассеяния.

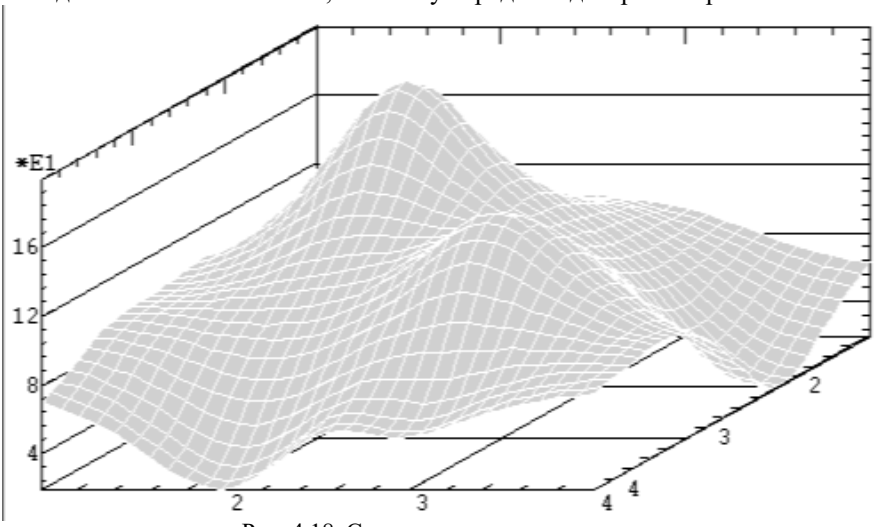


Рис. 4.18. Сглаживающая поверхность

Сглаживающая карта. *Карта* (рис. 4.19) представляет собой двухмерное изображение *сглаживающей поверхности*, на которой высоты кодируются в цветовой или черно–белой тональной шкале, приведенной справа от карты.

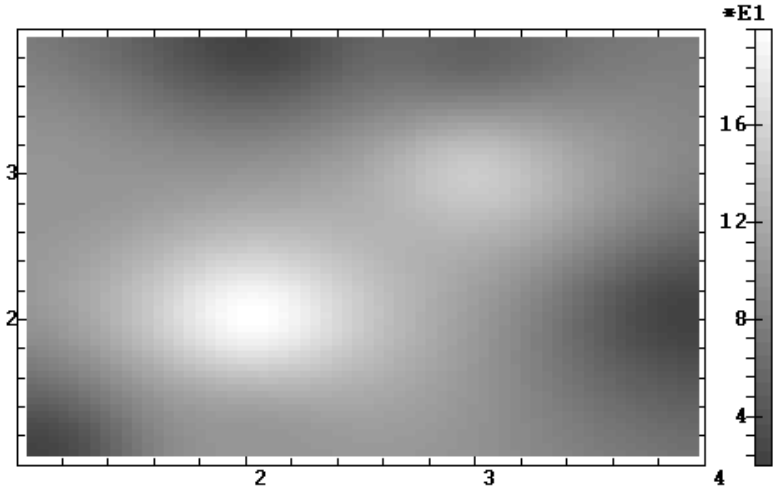


Рис. 4.19. Карта сглаживания

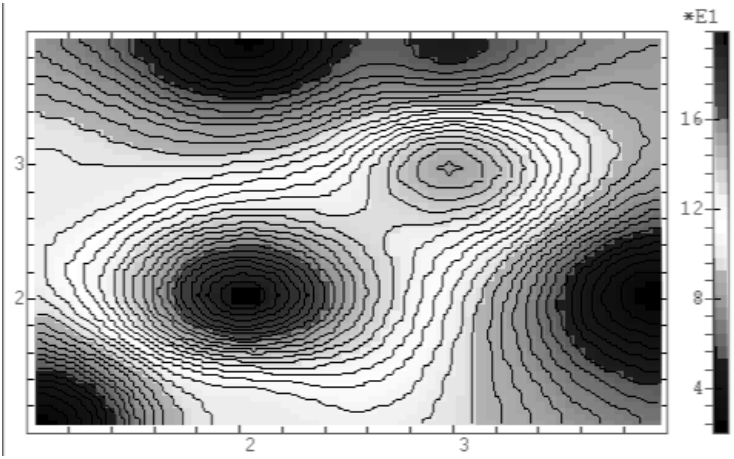


Рис. 4.20. Карта сглаживания с геодезическими уровнями

При нажатии на инструментальную кнопку «*Сетка»* на карте появляются геодезические линии (рис. 4.20).

При нажатии на инструментальную кнопку «*СохрГраф»* координаты *X, Y, Z* карты с заданным шагом по осям *X, Y* сохраняются в электронной таблице.

Трехмерная поверхность. График *трехмерной поверхности* строится по вводимой формуле, представляющей зависимость *Z=f(X,Y)*. На рис. 4.21 приведена поверхность, построенная для интервала значений *X*=0–5 и интервала значений *Y*=0–7 по формуле:

6**x***x***y***y*–36**x***x***x***y*–24**x***y***y*+143**x***y*+15**x*/(*y*+2)–2**y*/(*x*+5)+5.

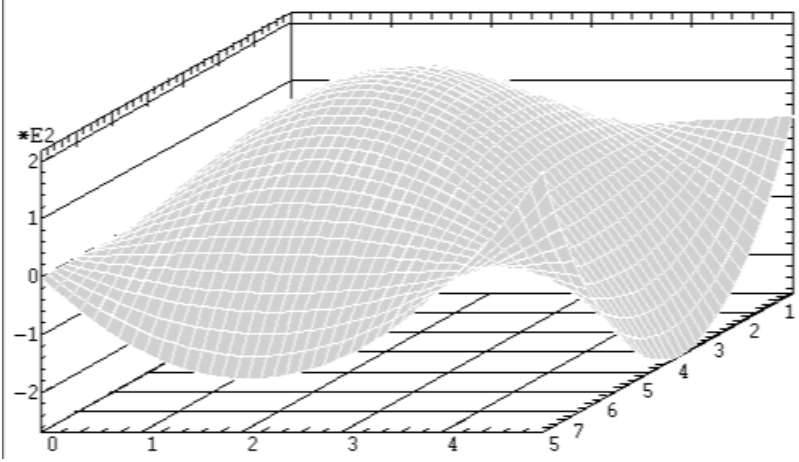


Рис. 4.21. Трехмерная поверхность

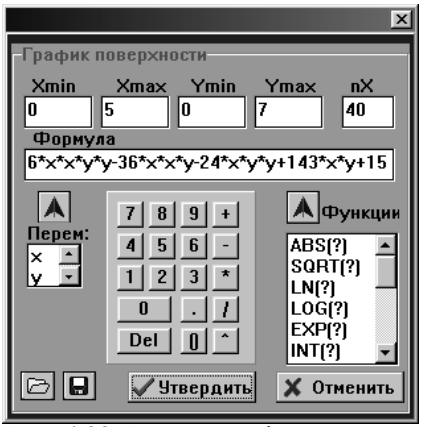


Рис. 4.22. Банк ввода формулы поверхности

Экранный бланк для ввода формулы и параметров трехмерной поверхности (рис. 4.22) содержит следующие элементы:

- два поля для левой и правой границы по координате *X*;
- два поля для левой и правой границы по координате *Y*;
- поле числа *градаций* в установленных диапазонах значений *X* и *Y*;
- поле вводимой *формулы* поверхности;
- поле осей координат *X, Y* с кнопкой их переноса в поле формулы;
- список допустимых алгебраических и тригонометрических функций с кнопкой их переноса в поле формулы (см. разд. 2.3);
- панель цифровой *клавиатуры*, для набора чисел и арифметических операций (см. разд. 2.3);
- кнопки обращения к *архиву формул* (чтение/запись);
- кнопка утверждения введенной формулы вычислений (она дублируется клавишей [Enter]);
- кнопка «*Отменить*» для отмены формулы и выхода из редактора (она дублируется клавишей [Esc]).

Содержимое полей бланка сохраняется от вызова к вызову и от сеанса к сеансу, поэтому не нужно каждый раз заново вводить одни и те же формулы.

Кроме того, имеется удобная возможность хранить различные формулы в *дисковом архиве* (см. разд. 3.2). Для этого предназначены две кнопки с *пиктограммами* чтения и записи. Первая операция позволяет считывать в экранный бланк формулу из архивного файла, а вторая — записывать формулу из бланка в архивный файл.

Ввод формулы можно осуществлять, пользуясь исполнительными инструментами бланка, однако это можно производить и полностью с клавиатуры (после активизации ее поля посредством перемещения указателя мыши с нажатием левой кнопки), что значительно быстрее.

Для перенесения в поле ввода уже готовых формул из различных меню и окон эффективно пользоваться *буфером обмена* (см. разд. 2.5).

Супердиаграмма. Визуализация многомерных данных настоятельно необходима не только во многих естественных науках (физика, химия, биология, геология, океанология, космология, психология, социология и других), но и в большинстве их практических и технических приложений: концентрация химического элемента в пробах, взятых на различных глу-

бинах и в различных точках земной коры (океана, атмосферы), плотность частиц в некотором объеме межгалактического пространства, изменение характеристик квантовых взаимодействий в пространстве множества определяющих их физических параметров, топология различных математических пространств и т. п.

Поэтому концепция супердиаграммы была сформулирована еще в 1993 г. при проектировании комплексной электрофизиологической лаборатории CONAN [48] для визуализации многомерных данных ЭЭГ–анализа. В пакете STADIA эта концепция получила свое дальнейшее развитие.

Определение. *Супердиаграммой* (СД) называется динамичный четырехмерный график (*гиперкуб*, рис. 4.23), представляющий выбранный срез многомерного пространства. Архитектурно гиперкуб включает подвижную плоскость в пространстве трех переменных (трехмерный куб), на которой в виде цветной карты изображено сглаженное распределение значений (амплитуд) четвертой переменной — избранного *показателя*.

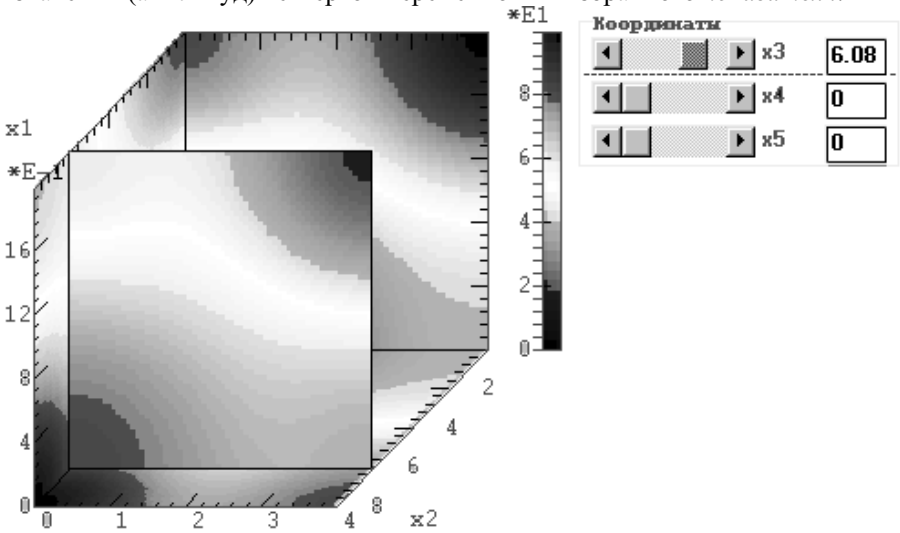


Рис. 4.23. Супердиаграмма

Терминология. Пусть имеется множество точек в *n*–мерном пространстве (такое множество численно представляется в виде обычной матрицы данных: столбцы — переменные или координатные оси, строки — измерения, объекты или значения координат). Представленные в СД измерения такого *n*–мерного пространства (переменные) подразделяются на три группы:

1) выделенная *W–переменная* (переменная–*показатель*) предназначена для картирования ее значений (амплитуд) на подвижной плоскости гиперкуба;

2) три другие переменные (*x1, x2, x3*) выбираются в качестве осей визуализируемого гиперкуба;
3) остальные *i–е переменные* (в количестве *n–4*) определяют конкретное 4–мерное сечение, отображаемое гиперкубом в *n*–мерном пространстве.

Справа от гиперкуба располагаются подвижные ползунки для изменения значений *x3* и *i*–х переменных с указанием обозначения соответствующей переменной с полем ее текущего значения.

Диалог. В *бланке выбора переменных* нужно указать *x1, x2, x3, W* и *i–е* переменные из матрицы данных. Подвижное сечение управляется клавишами горизонтального перемещения или верхним из ползунков. Переход к новому 4–мерному сечению производится простым перемещением ползунков.

Изначально в качестве *x1, x2, x3* координат принимаются первые три из выбранных переменных, а в качестве *W* — последняя из выбранных переменных. Смена *x1, x2, x3* координат производится двойным щелчком по полю значения соответствующей *x3* переменной. Замена *x3* переменной на *i*–ю переменную производится двойным щелчком по полю значения соответствующей *i*–й переменной.

Вычисленная на подвижной плоскости карта может быть запомнена в электронной таблице по нажатию кнопки «*СохрГраф*». Далее сохраненные данные могут быть использованы для построения поверхности или ее вращения.

Серия диаграмм. При нажатии на кнопку *«Сетка»* в третьей инструментальной строке происходит переключение супердиаграммы в режим

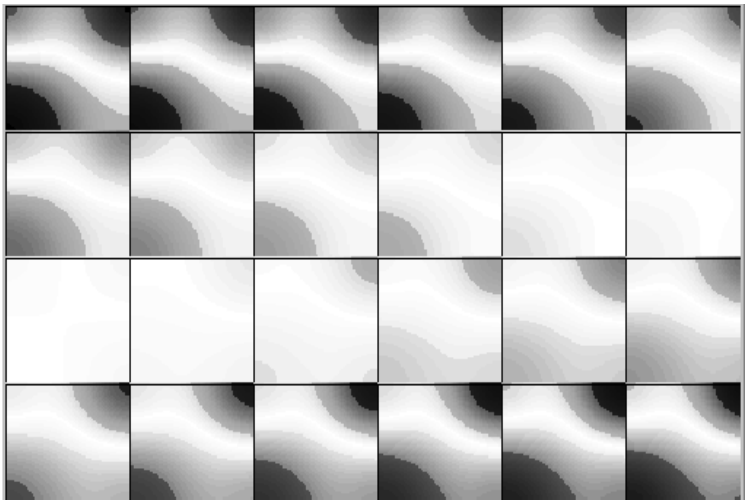


Рис. 4.24. Серийное картирование супердиаграммы

серийного картирования гиперкуба (по срезам *x3*, рис. 4.24) и обратно.

Алгоритм. Для расчета карты предложена простая и самоочевидная *полевая* модель: точки рассматриваются как источники зарядов, величина которых определяется значением *W*–переменной. Отсюда суммарная *напряженность поля* в конкретной точке рассчитывается по обычному в физике закону обратных квадратов (как в законах Ньютона и Кулона). Такой *полевой* метод дает результаты, близкие к интерполяции кубическими сплайнами, но требует существенно меньше вычислений. С другой стороны, он намного более прост и нагляден, чем многоступенчатые «эвристические» алгоритмы двухмерного сглаживания, использованные в известных и мощных западных пакетах *Serfer/Grapher* (научная графика), *IMSL* (библиотека статистических процедур на фортране) и других.

Глава 5

СТАТИСТИЧЕСКИЕ СРЕДСТВА

«Чтобы правильно задать вопрос,
нужно знать большую часть ответа»
[Р.Шекли. Ответчик]

Блок статистического анализа содержит набор процедур, реализующих широко употребительные и устоявшиеся методы анализа данных и представления результатов.

В этой главе мы сделаем следующий шаг популярного введения в статистический анализ, начатый в разд. 1.2. Более строгое изложение назначения и применимости статистических методов содержится в главах 6—13 с избранными примерами. И наконец, для углубленного изучения математической статистики следует обратиться к рекомендованной литературе.

Естественно, что мы не могли для каждого метода привести примеры на все мыслимые области его применения. Однако любой конкретный пример следует рассматривать как универсальный и обобщенный — просто замените название исходных данных на данные из вашей области, а постановку задачи — на вашу постановку задачи, и пример превратится в ваш пример.

Везде далее под *экспериментом* мы будем понимать процесс сбора информации об объекте исследования (или явлении), связанный с измерением или регистрацией значений, характеризующих объект переменных.

5.1. Статистический диалог

Чтобы провести *статистический анализ,* необходимо выполнить ряд последовательных шагов.

1. Ввод данных. Прежде всего, нужно ввести данные в *электронную таблицу*, принимая во внимание, что обрабатываемые данные должны соответствовать выбранному методу анализа.

2. Выбор метода. После этого следует вызвать меню статистических методов[1] (рис. 5.1) нажатием клавиши F9 или выполнением пункта «*Статистика*» в верхней командной строке. В этом меню нажмите кнопку нужного метода или же на клавиатуре — сопоставленную ей клавишу быстрого вызова.

---

[1] В отличие от западных статпакетов меню методов позволяет обозреть одномоментно все предоставляемые возможности в соответствии с классическими подразделами математической статистики, а не будучи скрытым в многоуровневых выпадающих списках под неадекватными наименованиями.

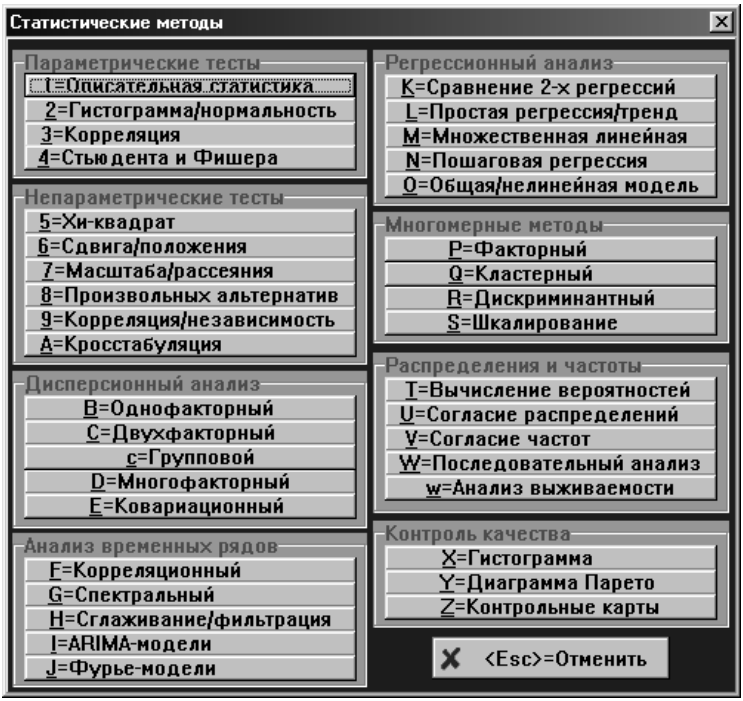


Рис. 5.1 Меню статистических методов

3. Выбор переменных. Для большинства методов далее появляется *бланк выбора переменных* (см. рис. 2.3), в котором можно отобрать подлежащие анализу переменные из электронной таблицы, принимая во внимание, что:

а) статистические тесты, оперирующие с одной выборкой, выполняются над каждой из выбранных переменных;

б) тесты для двух выборок могут быть выполнены для всех пар из избранного множества переменных;

в) многомерные методы требуют не менее двух переменных.

Ряд статистических методов (*кросстабуляция, дисперсионный анализ, шкалирование*) требует уже определенным образом подготовленных данных в электронной таблице без выбора переменных.

4. Диалог. Далее протекает диалог, характерный для выполняемого метода с выдачей числовых результатов анализа и их интерпретации в экранную страницу ⌊*Rez*⌋ текстового редактора (см. разд. 5.5), а графиков результатов — в графические страницы (разд. 4.1). В пакете STADIA реализован *прямоточный статистический диалог* в соответствии с естественной последовательностью исполнения методов, в котором вопросы и

меню появляются по ходу исполнения процедуры, упрощая тем самым на каждом шаге проблемы выбора[1].

5. Графики результатов. Большинство статистических процедур позволяют представлять результаты анализа в виде графиков с выдачей их в очередную экранную страницу ⌊*Gr*$_i$⌋, i = 1–15. В этом случае имеются следующие варианты:

- некоторые обязательные графики выдаются в графическую страницу безусловно;
- перед выдачей второстепенных графиках появляется запрос типа «*Выдать график (Да/Нет)*» и при положительном ответе выдается соответствующий график;

ряд графиков имеет информационный характер и их полезно просмотреть для уточнения хода дальнейшего анализа, в этом случае после выдачи графика в верхней части экрана появляется *ждущее меню* (рис. 5.2), чтобы оставить или удалить график.

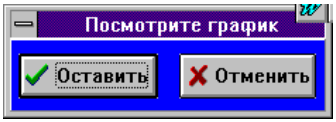


Рис. 5.2. Ждущее меню просмотра графика

По окончании выполнения статистической процедуры можно дополнительно поработать со страницами числовых результатов и графиков средствами текстового редактора (см. разд. 5.5) и графопостроителя (см. разд. 4.1).

Перенос результатов в электронную таблицу. Средства переноса результатов анализа в электронную таблицу открывают для них доступ ко всем возможностями пакета, в частности, с целью повторного анализа, преобразования или сохранения результатов анализа.

Для переноса числовых результатов необходимо их забрать в *буфер обмена* (см. разд. 2.5) из страницы результатов, перейти в электронную таблицу и в нужном месте вставить содержимое буфера обмена.

Полученные в графической форме результаты могут быть перенесены в электронную таблицу посредством специальной инструментальной клавиши «*СохрГраф*» (см. разд. 4.1). Координаты точек каждого графика переносятся в первые свободные столбцы электронной таблицы.

Для ускорения выполнения длинных последовательных рутинных операций в ходе статистического диалога очень полезен механизм макрокоманд (см. разд. 2.6).

---

[1] Это коренным образом отличается от реализованной в западных статпакетах иерархической системы вызываемых самим пользователем меню, каждая с умопомрачительным количеством настраиваемых режимов.

5.4. Текстовый редактор результатов

Текстовый редактор результатов анализа позволяет просматривать и редактировать выдачу числовых результатов анализа, сохранять ее в дисковом файле или выдавать на печать. Тем самым этот редактор выполняет сугубо вспомогательные функции и нисколько не претендует на статус издательской системы, для чего следует пользоваться специальными пакетами типа *MS Word, Page Maker, Venture* и другими.

Текстовый редактор становится доступным при активизации страницы результатов анализа ⌊*Rez*⌋ (рис. 5.7). Он поддерживает большинство типичных для всех подобных редакторов операций: передвижение указателя текущей позиции (в виде вертикальной палочки), ввод текста с клавиатуры, выделение фрагментов текста, удаление символов и фрагментов, забор фрагментов в буфер обмена и вставление фрагментов из буфера.

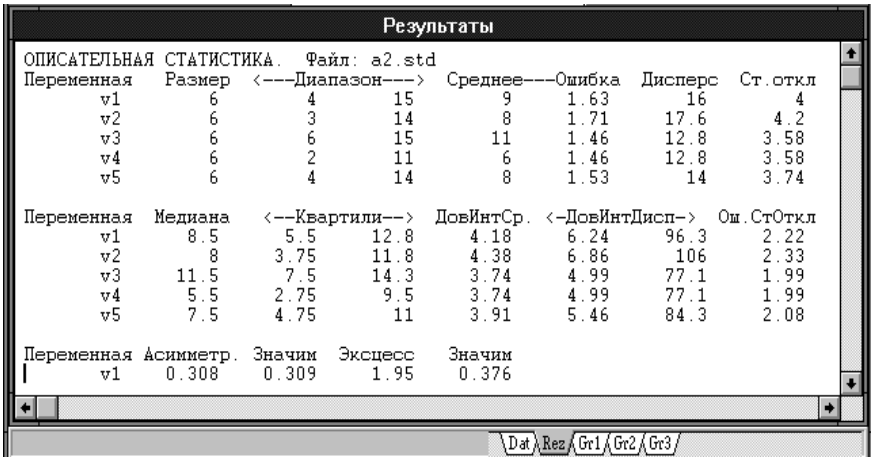


Рис. 5.7. Страница редактора результатов

Ограничение. По техническим причинам в редакторе не действует клавиша [Del], убирающая следующий символ, вместо нее следует пользоваться клавишей [BackSpace], убирающей предыдущей символ.

Представление чисел. Результаты в окно выдаются в так называемой *научной нотации*, когда очень большие и очень малые числа представляются в виде мантиссы и показателя десятичной степени. Например, число 4.8E–5 означает 4,8, умноженное на 10 в минус пятой степени, т. е. 0.000048, а 4.8E5 означает 480 000.

В электронную таблицу. Любой результат из страницы редактора может быть перенесен в электронную таблицу для повторного анализа и построения графиков.

Общие операции. Редактор поддерживает также общие операции: чтение и запись в отношении текстовых файлов, изменение шрифта и выдачи результатов на печать.

Операция *записи* позволяет сохранять выдачу результатов в дисковом файле, а операция *очистки* очищает страницу результатов.

Операция *чтения* полезна для компоновки общего отчета о полученных в разное время результатах. По такой операции содержимое страницы результатов полностью замещается. Поэтому перед выполнением чтения полезно забрать в буфер имеющуюся выдачу (разд. 2.5).

Операция *печати* выдает имеющиеся в редакторе результаты на принтер (изменение шрифта печати рассмотрено разд. 2.2).

Проблема шрифта. Исходно для выдачи результатов анализа в редактор установлен моноширинный системный экранный шрифт, все символы которого, включая пробелы, имеют одинаковую ширину. Это удобно для хорошего позиционирования заголовков и таблиц относительно нижеследующих колонок числовых результатов, поэтому вся выдача результатов рассчитана именно на использование подобного моноширинного шрифта.

Однако на вашем компьютере при просмотре результатов и выдаче их на печать могут возникнуть различные шрифтовые проблемы, которые следует решать посредством переустановки шрифта для страницы результатов (см. разд. 2.3).

5.5. Обозначения, учебная версия и примеры

В формулах описания статистических методов используются следующие общие обозначения:

x_i, y_j — элементы выборок X, Y с числом значений i=1,...,n, j=1,...,m;

x_{ij} — матрица с i=1,...,m переменными и j=1,...,n измерениями;

M (M_x, M_y) — выборочное среднее (для выборок X и Y);

S^2 (S_x^2, S_y^2) — выборочная дисперсия (для выборок X и Y);

S (S_x, S_y) — выборочное стандартное отклонение (для X и Y);

N — общее число значений;

$|x|$ — абсолютное значение x;

ln(x) — натуральный логарифм x;

exp(x) — экспонента x;

α — критический уровень значимости нулевой гипотезы;

P — уровень значимости статистического критерия, вычисленный для конкретных выборочных значений;

T — статистика, подчиняющаяся распределению Стьюдента;

F — статистика, подчиняющаяся распределению Фишера;

Z — статистика, подчиняющаяся нормальному распределению.

В примерах, иллюстрирующих применение статистических методов, приводится постановка задачи, исходные данные, экранная выдача результатов и выводимые на экран или на печать графики. Учебная версия STADIA с файлами примеров свободно доступна по адресу: http://statsoft.msu.ru/stadia.zip.

Следует прочитать инструкцию и запустить инсталлятор. По завершении инсталляции на жестком диске образуется папка \STADIA с поддиректорией \DAT, в которой содержатся учебные файлы данных, использованные в качестве примеров по статистическим методам (табл. 5.1). По этим примерам можно самостоятельно освоить работу со всеми процедурами.

Учебная и свободно доступная версия STADIA позволяет обрабатывать учебные файлы данных и данные, введенные с клавиатуры в текущем сеансе. Ограничения: невозможно сохранить введенные данные на диске, переносить данные через буфер обмена, обрабатывать матрицы объемом более 400 чисел.

Таблица 5.1. Учебные файлы данных

Имя файла	Содержимое	Разделы
3-норм	3-мерное нормальное распределение	11.1
a1	Урожайность четырех сортов пшеницы	8.2.1
a2	Урожайность пяти сортов картофеля	8.3
a2g	Выдыхаемый азот при четырех диетах питания	8.3
anova_b2	Двигательная активность крыс в ответ на инъекции мидозалама	8.4
anova_b3	Частоты выполнения профилактики при контактах с загрязненной средой	8.4
anova_b4	Частоты нажатия на педальку крысами в ответ на болевой раздражитель	8.4
anova_b5	Ошибки вождения трех марок автомобилей на трех типах дорог	8.4
anova_bs	Число приступов головной боли при релаксации и без нее	8.4
birds	Метрические показатели птиц	11.3
chmstr*	Показатели успеваемости студентов	10.3
cla	Показатели сортов немецкого пива	11.2, 11.3
cor1	Соотношения возраста, длины стопы и времени решения задач у 22 детей	6.3
corn	Урожайность пшеницы в СССР	9.6,10.3
cov	Влияние тренировки на способность подойти к живой змее	8.6
diasuper	Многомерные данные для супердиаграммы	4.4
doll	Динамика курса доллара и депозитов в 1994 г.	14.4

d-riga	Игровые показатели Динамо-Рига в сезоне 1980 г.	14.2
frea	Время выточки детали при трех способах обработки	8.2.3
jon	Время выполнения операций в зависимости от мотивации	8.2.2
latper	Латентные периоды условного рефлекса у кошки	3.5
life2	Данные о выживаемости пациентов	12.6
lr2	Сила сжатия кисти в зависимости от размера предплечья	10.2
mav,mav1	Урожайность картофеля в зависимости от предпосевной обработки и агротехники	8.5
mis	Данные с пропущенными значениями	3.5
mlr*	Показатели жизни в 75 странах мира	10.4–10.6
moto	Разрушающее испытание мотоцикла	9.4
msc	Сравнение политических лидеров эпохи перестройки	11.4
npt	Данные об уровнях радиоактивности, куриных эмбрионах, анализе химических препаратов, привесе свиней, усвояемости консервов	8.2–8.5
oil	Динамика мировых цен на нефть Брент в 2001–2004 гг.	9.6
page	Прочность хлопка в зависимости от количества удобрения	8.2.3
people*	Антропометрические данные 100 людей: рост, вес, возраст, цвет глаз и волос	6.1–6.4, 7.1–7.6
qcontr	Примеры контроля качества продукции	13.1–13.3
s40	Протоколы лыжных соревнований	14.1
seqan	Число бракованных изделий разного типа	12.5
sheffe	Показатели прочности шести сплавов	8.2.1
spec	Динамика авиаперевозок и численности насекомых	9.1–9.6
tab	Число случаев тромбоза при приеме двух препаратов	7.6
tales	Оценки героев детских сказок	11.1
testa	Результаты 10 тестов на профпригодность	11.1,11.2
town	Расстояния между городами СССР	11.4
tst	Продуктивность пшеницы и картофеля	7.3, 7.4
z3*	Велоэргометрические испытания	14.3
РФ	Курс ЦБ и индекс промышленного производства в 1996–1999 гг.	14.4.1
РФ96–99*	Ежемесячные измерения 64 экономических показателей РФ в 1996–1999 гг.	14.4.3
Фирмысша*	Экономические показатели 266 фирм США	14.4.2

Звездочкой помечены данные, размер которых превосходит возможности учебной версии, поэтому в учебных файлах они представлены сокращенно.

Глава 6

ПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ

«То, что уже есть, не требует доказательств.
Все доказательства суть попытки чем–то стать»
[Зе Краггаш. О неумолимости правдоподобного]

Рассматриваемые в данной главе методы базируются на предположении о том, что аализируемые выборки подчиняются нормальному закону распределения, достаточно объемны (включают не менее 10–15 значений) и содержат количественные измерения. Поэтому перед их применением необходимо убедиться в допустимости этой гипотезы. Такая проверка может быть выполнена по критериям *Смирнова* (Колмогорова–Смирнова), *омега–квадрат* и *хи–квадрат* (разд. 6.2), а также по коэффициентам *асимметрии* и *эксцесса* (разд. 6.1).

Предлагаемые процедуры позволяют вычислять основные статистические параметры, характеризующие выборку (разд. 6.1), строить график распределения ее значений и проверять гипотезу о нормальности (разд. 6.2), оценивать степень различия двух выборок по средним значениям, по дисперсии (разд. 6.4), или различия в синхронности изменения значений парных выборок (*корреляции*, разд. 6.3).

6.1. Описательная статистика

Действия и результаты. Для анализа нужно выбрать из электронной таблицы одну или несколько переменных (бланк рис. 2.3 и пояснения к нему).

Сначала вычисляются основные выборочные характеристики: размер выборки, диапазон значений, выборочное среднее (*M*), ошибка вычисления среднего (*eM*), выборочные дисперсия и стандартное отклонение (S^2, *S*).

Далее, по подтверждению, может быть выдана дополнительная статистика (см. *Пример*):

1) медиана и квартили, размах 100(1–α)% доверительного интервала (полуинтервал) среднего (*dM*), границы доверительного интервала дисперсии (S_1, S_2), ошибка стандартного отклонения (*eS*);
2) коэффициенты асимметрии (*Sw*) и эксцесса (*Ku*) с уровнями значимости *P* нулевой гипотезы об отсутствии различий выборочного распределения от нормального распределения по каждому из коэффициентов (проверка производится по нормально распределенным статистикам Z_S, Z_K). Если $P>0.05$, нулевая гипотеза может быть принята. Такой способ проверки нормальности распределения хорошо работает для малых выборок, если же объем выборки превосходит 15–20 значений, то предпочтительнее пользоваться специальными критериями (см. разд. 6.2).

Затем, по подтверждению, вычисленные средние значения и стандартные отклонения могут быть сохранены в электронной таблице для дальнейшего анализа.

Ограничение. Размер выборки должен быть больше 4 и меньше *l*, где l = 16000, 5000, 1000, 100 при объеме матрицы данных в 64 000, 20 000, 4000 и 400 чисел.

Пример

З а д а ч а. В антропометрическом исследовании было случайным образом отобрано 50 мужчин и 50 женщин в возрасте от 15 до 70 лет, у которых были зарегистрированы следующие показатели: рост, вес, возраст, цвет глаз и цвет волос (файл PEOPLE). Характер распределения показателей роста, веса и возраста в выборке иллюстрирует рис. 6.1, а соотношения *вес–рост* и *вес–возраст* показаны на диаграммах рассеяния (рис. 6.2).

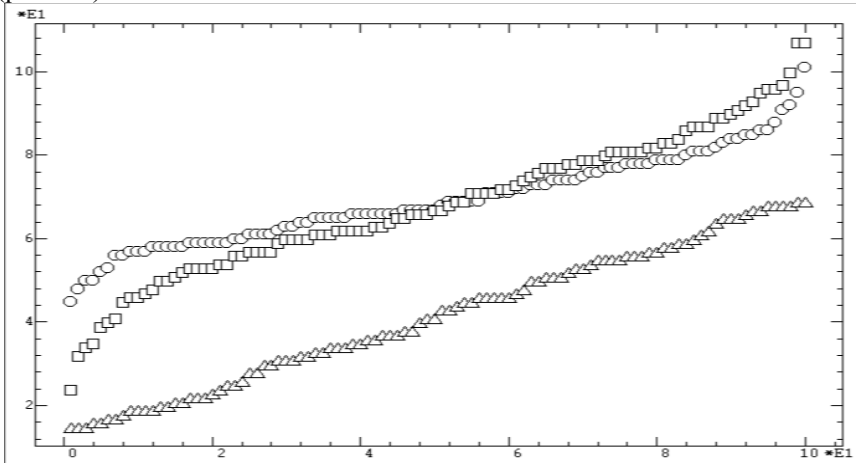
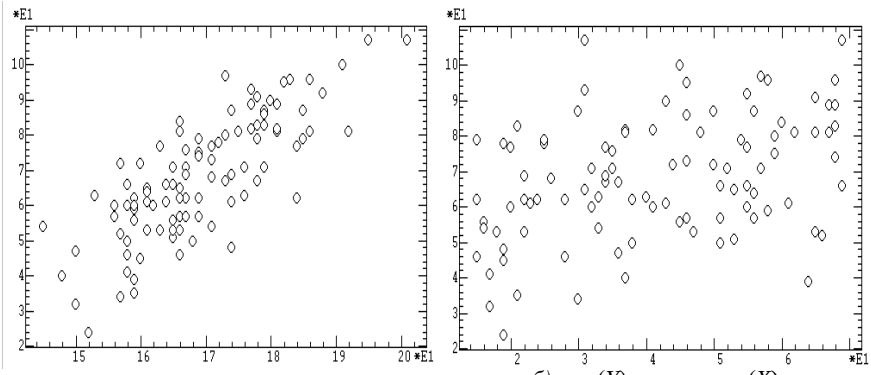


Рис. 6.1. График Кетле для распределения значений антропометрических показателей: квадраты — *рост–100*; круги — *вес*; треугольники — *возраст*; по горизонтальной оси — номер измерения в упорядоченном по возрастанию значений ряду



а) вес (*Y*) — рост (*X*) *б)* вес (*Y*) — возраст (*X*)

Рис. 6.2. Диаграммы рассеяния антропометрических показателей; разметка осей произведена в десятках единиц измерения:

На предварительном этапе исследования необходимо вычислить оценки описательной статистики для этих показателей.

Р е з у л ь т а т ы:

```
ОПИСАТЕЛЬНАЯ СТАТИСТИКА.                Файл: people.std
Перемен. Размер <-Диапазон-> Среднее-Ошибка Дисперс Ст.откл Сумма
рост      100     145  201     169   1.09     119    10.9   1.69E4
вес       100      24  107     68.3  1.71     293    17.1   6.83E3
возраст   100      15   69     41.7  1.65     272    16.5   4.17E3
Переменная  Медиана <-Квартили-> ДовИнтСр <-ДовИнтДисп-> Ош.СтОткл
рост          168      161  178     2.13     89.6   161      2.9
вес            67       57   81     3.35      221   398      4.56
возраст        42       28  55.8    3.23      205   369      4.39
Переменная Асимметр. Значим  Эксцесс   Значим
рост        0.323    0.087    2.85     0.42
вес        -0.054    0.41     2.67     0.273
возраст     0.0072   0.488    1.79     0.0057
```

В ы в о д ы: В данной выборке наблюдаются следующие средние показатели для мужчин и женщин: вес = 68.3 кг, рост = 169 см, возраст = 41,7 года, при стандартных отклонениях в 10.9 кг, 17.1 см, 16.5 лет. По оценкам асимметрии и эксцесса нормальному распределению соответствует показатели роста и веса (оба уровня значимости существенно больше 0.05), но по оценке асимметрии проходит и показатель возраста. Можно также отметить заметную асимметрию роста, что видимо определено систематически меньшим ростом у женщин (т. е. по этому показателю в выборке присутствуют две различающиеся популяции). С другой стороны, вес не имеет заметной асимметрии, что можно объяснить большим средним весом у женщин, несмотря на их меньший рост.

Продолжение анализа данного примера см. в следующем разделе.

6.2. Гистограмма и проверка распределения на нормальность

Назначение. *Гистограмма* является общеупотребительной формой представления выборочного распределения. Для ее вычисления диапазон изменения выборочных значений разбивают на некоторое число равных интервалов (*бинов*) и подсчитывают число значений, попадающих в каждый бин. При графическом представлении гистограммы на каждом интервале строится прямоугольник (столбик), высота которого пропорциональна числу выборочных значений в бине (рис. 6.4).

Действия. Для анализа нужно выбрать из электронной таблицы одну переменную, представляющую анализируемую выборку (рис. 6.3).

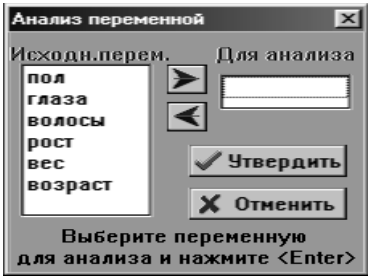


Рис. 6.3. Бланк выбора переменной для расчета гистограммы

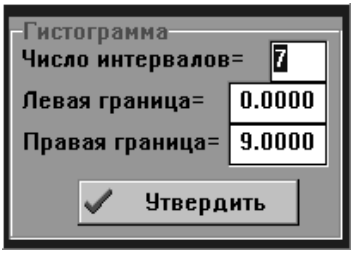


Рис. 6.4. Бланк установки параметров гистограммы

Затем можно уточнить число интервалов и область определения гистограммы (рис. 6.4). В качестве штатного числа интервалов подсказывается значение, вычисленное по эвристической формуле: $int(1.5+3.3\cdot\log_{10}(N))$, а область определения принята равной диапазону выборочных значений. Можно увеличить число интервалов, чтобы гистограмма была более подробной. Установка другого размера области определения полезна, например, когда нужно визуально сравнить несколько гистограмм, с их построением в одинаковом интервале по оси *X*.

Результаты. Для каждого интервала гистограммы выводятся следующие значения (см. *Пример 1*): левая граница интервала в исходных единицах и в единицах стандартного отклонения; число выборочных значений, попавших в интервал; накопленное число выборочных значений до текущего интервала включительно (последние два параметра приводятся в натуральном и в процентном выражении).

Затем проводится проверка нулевой гипотезы об отсутствии различий между выборочным и нормальным распределениями и выдача трех различных статистик:

- *Колмогорова D* с уровнем значимости *P*;
- *омега–квадрат* ω^2 с уровнем значимости *P*;
- *хи–квадрат* χ^2 с уровнем значимости *P*.

При $P>0.05$ нулевая гипотеза может быть принята.

Графическая выдача содержит изображение гистограммы с наложенной кривой нормального распределения (рис. 6.7).

По подтверждению вычисленные частоты гистограммы заносятся в первую свободную переменную матрицы данных, что полезно для последующего анализа, использующего распределения в качестве исходных данных (см. разд. 7.1).

Ограничения:

1. Размер выборки должен быть больше 4 и меньше l, где l = 16000, 5000, 1000, 100 при объеме матрицы данных в 65000, 20000, 4000 и 400 чисел.
2. Используемые аппроксимации критериев для вычисления уровня значимости получены в предположении «n стремится к бесконечности» и достаточно точны для достаточно больших выборок (n>10–15) и в области значений P=0.15–0.01.

Пример 1

Задача. В антропометрическом исследовании ста человек фиксировались, в частности, их вес и возраст (файл PEOPLE). В примере к разд. 6.1 было показано, что по оценкам асимметрии и эксцесса нормальному распределению соответствуют показатели роста и веса. Возраст же не соответствует нормальному распределению по показателю эксцесса Для более наглядного представления о характере распределений этих показателей полезно построить гистограммы и провести более тщательную проверку соответствия их нормальному распределению. Сделаем это для показателей веса и возраста.

Результаты:

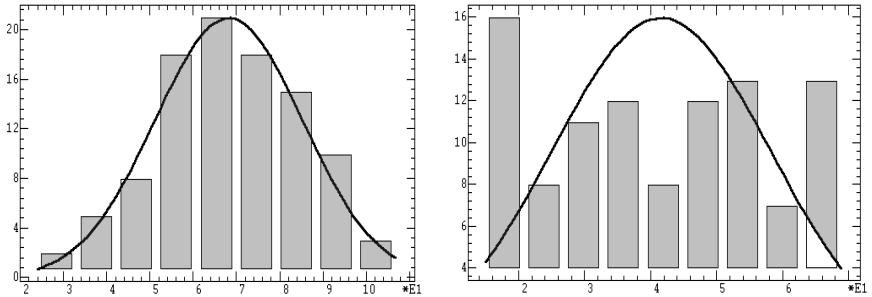
ГИСТОГРАММА И ТЕСТ НОРМАЛЬНОСТИ. Файл: people.std
Переменная: вес

X-лев.	X-станд	Частота	%	Накопл.	%
24	-2.59	2	2	2	2
33.2	-2.05	5	5	7	7
42.4	-1.51	8	8	15	15
51.7	-0.97	18	18	33	33
60.9	-0.432	21	21	54	54
70.1	0.107	18	18	72	72
79.3	0.646	15	15	87	87
88.6	1.18	10	10	97	97
97.8	1.72	3	3	100	100
107	2.26				

Колмогоров=0.0547, Значимость=0.977, степ.своб = 100
Гипотеза 0: <Распределение не отличается от нормального>
Омега-квадрат=0.0379, Значимость=1.04, степ.своб = 100
Гипотеза 0: <Распределение не отличается от нормального>

```
      Хи-квадрат=1.94, Значимость=0.926, степ.своб = 6
         Гипотеза 0: <Распределение не отличается от нормального>
      ГИСТОГРАММА И ТЕСТ НОРМАЛЬНОСТИ.  Файл: people.std
                                          Переменная: возраст
   X-лев. X-станд  Частота        %   Накопл.        %
      15    -1.62       16       16       16       16
      21    -1.25        8        8       24       24
      27    -0.89       11       11       35       35
      33   -0.525       12       12       47       47
      39   -0.161        8        8       55       55
      45    0.203       12       12       67       67
      51    0.567       13       13       80       80
      57    0.931        7        7       87       87
      63     1.29       13       13      100      100
      69     1.66
Колмогоров=0.0809, Значимость=0.128, степ.своб = 100
   Гипотеза 0: <Распределение не отличается от нормального>
Омега-квадрат=0.172, Значимость=0.0115, степ.своб = 100
   Гипотеза 1: <Распределение отличается от нормального>
Хи-квадрат=41.7, Значимость=6.72E-6, степ.своб = 6
         Гипотеза 1: <Распределение отличается от нормального>
```

В ы в о д ы: Согласно результирующим уровням значимости всех трех критериев (P>>0.05) можно принять гипотезу о нормальном распределении веса людей (рис. 6.7, *а*).



а *б*

Рис. 6.7. Гистограммы выборочных распределений
с графиками плотности вероятности нормального распределения:
а — вес в выборке из 100 человек; б — возраст в той же выборке

Что же касается возраста, то по двум критериям (чувствительным к различиям «на концах») его распределение не соответствует нормальному закону (на глаз это хорошо видно по рис. 6.7, *б*). Это соответствует содержательным соображениям о том, что в случайной выборке возраст людей должен быть распределен достаточно равномерно. Это предположение будет проверено при рассмотрении примера 3 разд. 7.1.

Продолжение анализа данного примера см. в следующем разделе.

Пример 2

Задача. Очень полезной практикой для выработки внутреннего «ощущения случайности» случайных процессов, а также для понимания сравнительной чувствительности вышерассмотренных критериев является следующее простое упражнение. Сгенерируйте несколько случайных выборок согласно нормальному закону распределения достаточно большого размера (операция *«Генератор чисел»* из *Блока преобразований*, разд. 3.4), проверьте их распределение на нормальность и визуально сравните их гистограммы.

Ниже приведены результаты проверки по критерию Колмогорова четырех подобных выборок, включающих по 100 элементов.

Результаты:

```
Колмогоров=0.073, Значимость=0.259, степ.своб = 100
Колмогоров=0.049, Значимость=0.058, степ.своб = 100
Колмогоров=0.048, Значимость=0.242, степ.своб = 100
Колмогоров=0.057, Значимость=0.076, степ.своб = 100
```

Выводы: Как видно из полученных результатов, все выборки подчиняются нормальному закону распределения, но уровни значимости нулевой гипотезы колеблются в довольно широких пределах, иногда приближаясь вплотную к критической границе.

В качестве самостоятельного упражнения можно сравнить результаты по другим критериям, а также их зависимость от размеров выборок.

6.3. Линейная корреляция

Назначение. *Параметрический коэффициент корреляции Пирсона r* является индикатором ***линейной связи*** между парными переменными (о парных переменных см. разд. 5.2), подчиняющимися нормальному закону распределения.

Действия и результаты. Для анализа нужно выбрать из электронной таблицы две или несколько переменных (см. рис. 2.3).

Вычисляется коэффициент корреляции *r Пирсона* со статистикой Стьюдента и уровнем значимости *P* нулевой гипотезы «r=0». Если P>0.05, коэффициент корреляции может быть признан незначимым.

В случае нескольких выбранных переменных выдается диагональная матрица коэффициентов корреляции с указанием критического значения r_0 и числа значимых коэффициентов корреляции (r>r_0). Эта матрица (по подтверждению) может быть сохранена в электронной таблице для последующего использования (например, в разделах многомерной статистики).

В том случае, если для анализа выбрано три переменные, то вычисляется также диагональная матрица частых коэффициентов корреляции $r_{xy\text{-}z}$ между парами переменных с выдачей критического значения r_0 и числа значимых коэффициентов.

Если анализ производится между заданными парами переменных (см. рис. 2.3, кнопка с пиктограммой чтения), то в текстовый файл Л-КОР.txt записываются вычисленные коэффициенты с уровнями значимости.

Ограничение. Объемы выборок должны быть равны и содержать не менее четырех значений и не более l, где l = 16000, 5000, 1000, 100 при объеме матрицы данных в 64 000, 20 000, 4000 и 400 чисел.

Пример 1

Задача. Целью исследования было выявление зависимости между продуктивностью пшеницы и картофеля на соседних полях. Для этого были использованы данные по урожайности [ц/га] за 12 последовательных лет (табл. 6.3.1, переменные LC1, LC2 в файле TST)

Таблица 6.3.1. Урожайность пшеницы и картофеля за 12 лет [ц/га]

Пшен	20,1	23,6	26,3	19,9	16,7	23,2	31,4	33,5	28,2	35,3	29,3	30,5
Карт	7,2	7,1	7,4	6,1	6	7,4	9,4	9,2	8,8	10,4	8	9,7

Для выявления связи этих двух показателей данные были подвергнуты корреляционному анализу.

Результаты:

```
ПАРАМЕТРИЧЕСКАЯ КОРРЕЛЯЦИЯ. Файл: tst.std Переменные: lc1, lc2
Коэфф.корреляции=0.8516 T=5.138, Значимость=0.0006, степ.своб = 10
       Гипотеза 0: <Коэффициент корреляции отличен от нуля>
```

Выводы: Проверка нулевой гипотезы (уровень значимости равен 0.0006, что существенно меньше 0.05) выявляет значимую корреляцию между урожайностью пшеницы и картофеля.

Пример 2

Задача. Целью исследования было выявление связей между антропометрическими и умственными показателями детей. В экспериментах у 22 детей (табл. 6.3.2, файл COR1) измерялись: время решения математических задач, возраст и размер стопы.

Таблица 6.3.2. Результаты исследования детей

Ученик	Время решения задачи, мин	Возраст, лет	Размер стопы, см
1	4,6	10,1	17,2
2	2,8	10,6	18
3	5,4	10,3	17
4	4	10,1	16,7
5	2,6	10,9	17,8
6	4,4	10,2	17,4
7	4,6	9,9	16,5
8	4,4	10,1	16,5
9	5,2	9,3	15,1
10	2,8	10,8	17,8
11	5,2	9,4	15,5
12	3,6	10,2	16,9
13	3,8	9,8	16,9
14	6	9,1	15
15	4,2	10,1	16,7
16	4,4	10,2	17
17	4	10,2	16,5
18	4,4	10,1	17,2
19	4,2	10	16,1
20	3,6	10,3	16,7
21	3,4	10,5	17,2
22	2	11,1	20

Для выявления связи этих признаков данные были подвергнуты корреляционному анализу.

Результаты:

```
ПАРАМЕТРИЧЕССКАЯ КОРРЕЛЯЦИЯ.  Файл: cor1.std
            Корреляционная матрица
             время  возраст
  возраст   -0.869
    стопа   -0.822    0.906
 Критическое значение=0.419
 Число значимых коэффициентов=3 (100%)

            Частные корреляции
             время  возраст
  возраст   -0.515
    стопа   -0.163    0.682
 Критическое значение=0.429
 Число значимых коэффициентов=2 (66%)
```

Выводы: Расчет коэффициентов парной корреляции выявил удивительный факт — наличие высокой обратной связи между умственными способностями и размером стопы ребенка (корреляция –0.8538 по абсолютной величине значительно превышает критическое значение 0.4186). Однако расчет частных корреляций показал, что эти признаки не связаны друг с другом (частная корреляция –0.1637 по абсолютной величине много ниже критического уровня 0.429), поскольку оба они сильно зависят от возраста.

Пример 3

Задача. Продолжим анализ антропометрических данных из примеров к разд. 6.1, 6.2 (файл PEOPLE) с целью исследования корреляций между показателями веса, роста и возраста людей. Уже визуальный анализ диаграмм рассеяния (см. рис. 6.2) выявляет хорошую коррелированность веса и роста и меньшую, но все же заметную корреляцию между весом и возрастом.

Результаты (сокращенно):

```
ПАРАМЕТРИЧЕССКАЯ КОРРЕЛЯЦИЯ.  Файл: cor1.std
            Корреляционная матрица
              рост      вес
      вес    0.784
  возраст    0.106    0.385
 Критическое значение=0.194
 Число значимых коэффициентов=2 (66%)
```

Выводы: Значимые корреляции обнаружены между весом и ростом, а также между весом и возрастом, хотя последняя связь оказалась в 2 раза слабее первой (как мы и предполагали выше). Связь между возрастом и ростом мала и незначима. Она, видимо, является следствием продолжения роста человека от 15 до 25 лет, что компенсируется небольшим уменьшением роста в преклонном возрасте. Анализ частных коэффициентов корреляции здесь не проводится, поскольку нет причин считать наличие сильного влияния связей с третьим показателем на взаимную связь двух других.

Продолжение анализа данного примера см. в следующем разделе.

6.4. Критерии Стьюдента и Фишера

Назначение. Критерий Стьюдента для двух выборок, подчиняющихся нормальному закону распределения, проверяет нулевую гипотезу об отсутствии различий (о равенстве) их выборочных средних, а критерий Фишера — гипотезу о равенстве дисперсий двух выборок. Для парных переменных существует отдельная формулировка критерия Стьюдента, более чувствительная к их различиям.

Действия и результаты. Для анализа нужно выбрать из электронной таблицы две или несколько переменных (см. рис. 2.3).

Выдача включает значения следующих статистик:

- статистика *Фишера F* (она равна отношению дисперсий выборок);
- статистика *Стьюдента T* (в зависимости от результатов сравнения дисперсий применяются различные формулы вычисления *T*–статистики);
- разность средних и половина ее доверительного интервала *dM*;
- в случае равенства размеров выборок выдается также статистика Стьюдента, применимая для *парных переменных* (о парных переменных см. разд. 5.2, в западных статпакетах используются наименования *одновыборочный и двухвыборочный критерии* Стьюдента).

Для каждой статистики вычисляется уровень значимости *P* соответствующей нулевой гипотезы отсутствия различий. Если $P>0.05$, нулевая гипотеза может быть принята.

В случае нескольких выбранных переменных или задания конкретных пар переменных в текстовом файле (см. рис. 2.3, кнопка с пиктограммой чтения) указанные вычисления производятся для всех пар переменных и дополнительно выдается критическое значение Стьюдента с учетом поправки Бонферрони. Значимости нулевых гипотез по Стьюденту для по-

следующего использования записываются в текстовые файлы НепарС.TXT, ПарС.TXT (непарные и парные выборки).

Пример 1

З а д а ч а. Целью исследования было сравнение влияния на урожайность пшеницы двух агротехнических методов, применяемых на двух соседних полях в течение 10 лет (табл. 6.4.1, переменные *f1*, *f2* в файле TST).

Таблица 6.4.1. Урожайность пшеницы за 10 лет [ц/га] на соседних полях при двух агротехнических методах

метод1	20	17,9	20,6	22	21,4	23,8	21,4	19,8	18,4	22,5
метод2	22,1	18,5	19,4	22,1	21,7	24,9	21,6	20,3	18,7	23,1

Р е з у л ь т а т ы:

```
КРИТЕРИИ ФИШЕРА И СТЬЮДЕНТА.  Файл: tst.std  Переменные: f1, f2
Статистика Фишера=0.81, Значимость=0.379, степ.своб = 9,9
   Гипотеза 0: <Нет различий между выборочными дисперсиями>
Статистика Стьюдента=0.534, Значимость=0.606, степ.своб = 18
Разность средних=0.46, доверит.интервал=0.522
   Гипотеза 0: <Нет различий между выброчными средними>
Стьюдент для парных данных=1.76, Значимость=0.11, степ.своб = 9
   Гипотеза 0: <Нет различий между выброчными средними>
```

В ы в о д ы: Как можно видеть из полученных результатов анализа, ни критерий Стьюдента, ни критерий Фишера не выявляют заметных различий между рассматриваемыми методами сбора урожая (полученные уровни значимости 0.379 и 0.6056 существенно больше 0.05).

Отметим также, что в данном случае выборки можно рассматривать как парные переменные, поскольку поля являются соседними и не отличаются ни по почве, ни по климату, ни по другим условиям, а измерения проводились синхронно по годам. Однако даже более чувствительный к различиям парный критерий Стьюдента не выявляет таковых.

Количественную оценку диапазона значений для генеральной разности средних относительно выборочной разности 0.46 дает доверительный интервал 0.522, который включает и нулевое значение с перекрытием в 0.522–0.46=0.062, т. е. с «запасом» на $0.062/0.46 \cdot 100 = 13{,}5\%$.

П р и м е р 2

З а д а ч а. Продолжим исследование антропометрических данных из примеров к разд. 6.1–6.3 (файл PEOPLE) с целью исследования различий между показателями веса, роста и возраста мужчин и женщин. Перед этим необходимо произвести ряд преобразований, чтобы разделить данные по мужчинам и женщинам (см. пример к разд. 3.4)

Р е з у л ь т а т ы (сокращенно):

```
КРИТЕРИЙ ФИШЕРА И СТЬЮДЕНТА.  Файл: people.std
                              Переменные: рост-м, рост-ж
Статистика Стьюдента=3.97, Значимость=0.000327, степ.своб = 98
   Гипотеза 1: <Есть различия между выборочными средними>
Разность средних=8.03, доверит.интервал=0.000661
С поправкой Бонферрони: значимость=0.000259, степ.своб=294,
      крит.значимость=0.00333
   Гипотеза 1: <Есть различия между выборочными средними>

                              Переменные: вес-м, вес-ж
Статистика Стьюдента=4.33, Значимость=0.000141, степ.своб = 98
   Гипотеза 1: <Есть различия между выборочными средними>
Разность средних=13.7, доверит.интервал=0.000444
```

```
С поправкой Бонферрони: значимость=0.000107, степ.своб=294,
      крит.значимость=0.00333
   Гипотеза 1: <Есть различия между выборочными средними>

                              Переменные: возраст-м, возраст-ж
Статистика Стьюдента=0.141, Значимость=0.883, степ.своб = 98
   Гипотеза 0: <Нет различий между выборочными средними>
Разность средних=0.468, доверит.интервал=2.93
```

В ы в о д ы: Результаты показывают наличие различий в весе и росте по признаку пола (уровни значимости намного меньше критического уровня, да и доверительные интервалы составляют незначительный процент от разностей средних), но не в возрасте (уровень значимости нулевой гипотезы очень высок, а доверительные интервалы в шесть раз превышают разность средних). В данном случае поправку Бонферрони учитывать не следует, поскольку здесь имеет место не сравнение различий нескольких групп по одному показателю, а различие двух групп по трем разным показателям.

Продолжение анализа данного примера см. разд. 7.6.

Глава 7

НЕПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ

«Ищите же прежде Царствия Божия
и правды Его»
[от Матф: 6,33]

Назначение. Большинство статистических процедур опираются на допущение о *нормальном распределении* исходных данных. Для ненормально распределенных данных, ранговых выборок и выборок малого объема более эффективно применять так называемые *непараметрические методы*, не базирующиеся на каком-либо предположении о законе распределения данных, а использующие, как правило, только предположения о случайном характере исходных данных и о непрерывности генеральной совокупности, из которой они извлечены.

7.1. Критерий хи–квадрат

Назначение. Критерий хи–квадрат оценивает различие двух выборок по форме распределения их значений, представленных в виде гистограмм. В отличие от других непараметрических методов он применим к достаточно представительным выборкам, включающим не менее 15–20 элементов и представленных в форме гистограммы (см. разд. 6.2). Число интервалов в гистограмме должно быть не менее 4, а каждый интервал должен включать не менее 3–4 выборочных значений (в противном случае рекомендуется соединить соседние интервалы гистограммы).

Разновидности критерия. Данный раздел включает два варианта критерия хи–квадрат:

1. *Критерий однородности* двух независимых выборок проверяет гипотезу отсутствия различий между двумя выборочными распределениями.
2. *Критерий согласия* выборочного распределения и предполагаемого теоретического распределения.

Исходные данные. Исходные данные представляют собой гистограммы двух распределений (двух эмпирических или эмпирического и теоретического).

Предварительное преобразование исходной выборки в форму гистограммы может быть произведено в процедуре «Гистограмма и нормальность» (см. разд. 6.2).

Действия и результаты. Для анализа нужно выбрать из электронной таблицы две переменные (см. рис. 2.3), представляющие собой гистограммы сравниваемых распределений.

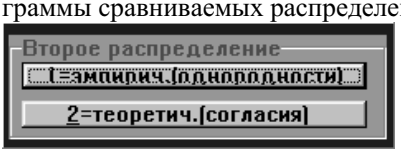


Рис. 7.1. Меню выбора критерия хи-квадрат

Далее нужно уточнить вариант критерия хи–квадрат (рис. 7.1).

После этого вычисляется значение статистики хи–квадрат и уровень значимости P соответствующей нулевой гипотезы. При $P>0.05$ нулевая гипотеза может быть принята.

Интервалы гистограммы, включающие менее трех выборочных значений, в процессе вычислений автоматически сливаются с соседними.

Ограничения: Число значений обоих переменных (интервалов у гистограмм) должно быть одинаковым и не менее четырех.

Пример 1

Задача. Необходимо сравнить частоты событий, измеренные в эксперименте, с теоретическими частотами (табл. 7.1.1, переменные Chi1 и Chi3 в файле NPT).

Таблица 7.1.1. Экспериментальные и теоретические частоты событий

Эксперимент	7	11	13	19	16	7	7
Теория	19	16	7	11	13	7	7

Результаты:

```
КРИТЕРИЙ ХИ-КВАДРАТ.  Файл: npt.txt
                    Переменные: chi1, chi3
Хи-квадрат=10.71, Значимость=0.0978, степ.своб = 6
   Гипотеза 0: <Нет различий между двумя распределениями>
```

Выводы: Вычисленные уровни значимости критерия хи–квадрат согласия позволяют принять гипотезу о соответствии экспериментальных данных предполагаемому теоретическому распределению (P=0.0978 больше 0.05).

Пример 2

Задача. На предприятии имеются данные о числе работников с заработной платой в заданных пределах для двух возрастных категорий (табл. 7.1.2, переменные CH1, CH2 из файла NPT). Необходимо проверить гипотезу об отсутствии различий в оплате труда между двумя возрастными категориями работников.

Таблица 7.1.2. Экспериментальные и теоретические частоты событий

Зарплата	100-120	120-140	140-160	160-180	180-200	200-220
Категория1	71	430	1072	1609	1178	158
Категория2	54	324	894	1202	903	112

Поскольку данные представляют собой гистограммы, то для сравнения используем критерий однородности этих двух экспериментальных распределений, как принадлежащих одной генеральной совокупности.

Результаты:

```
КРИТЕРИЙ ХИ-КВАДРАТ.  Файл: npt.txt   Переменные: chi1, chit
Хи-квадрат=3.218, Значимость=0.781, степ.своб = 6
   Гипотеза 0: <Нет различий между двумя распределениями>
```

В ы в о д ы: Полученный уровень значимости хи–квадрат критерия однородности (P=0.781 существенно больше 0.05) позволяет принять нулевую гипотезу об отсутствии различий в оплате труда.

П р и м е р 3

З а д а ч а. Продолжим анализ антропометрических измерений из раздела 6.1. В разд. 6.2 было показано, что показатель возраста в анализируемой выборке имеет распределение, отличающееся от нормального, и было высказано предположение о равномерном его распределении в представленном диапазоне 15–70 лет. Проверим это предположение. Для этого необходимо при построении гистограммы подтвердить необходимость сохранения ее в матрице данных. Кроме того, принимая во внимание, что объем нашей выборки составляет 100 человек, установим при построении гистограммы число бинов равным 10. Тогда проще будет сформировать вторую переменную, необходимую для данного метода, содержащую гистограмму теоретического распределения с 10 бинами, по 10 элементов в каждом.

Р е з у л ь т а т ы:

```
КРИТЕРИЙ ХИ-КВАДРАТ.  Файл: people.std
Хи-квадрат=8.6, Значимость=0.475, степ.своб = 9
   Гипотеза 0: <Нет различий между двумя распределениями>
```

В ы в о д ы: Полученный уровень значимости хи–квадрат критерия согласия позволяет принять нулевую гипотезу об отсутствии различий между распределением выборочных значений показателя возраста и равномерным распределением.

7.2. Критерии различия сдвига (положения)

Назначение. Критерии различия сдвига направлены на проверку следующих гипотез:

а) отсутствие различий во взаимном положении (*медианах*) двух независимых совокупностей, например наблюдений одних объектов без «обработки» и других объектов после обработки с анализом систематического *сдвига* значений второй выборки как результата обработки;

б) сдвиг выборок друг относительно друга равен значению d;

в) медиана одной анализируемой выборки равна значению d.

В случае б) необходимо предварительно все значения второй выборки Y уменьшить на величину d: $y_i=y_i–d$, что можно сделать посредством операции линейного преобразования $ax+b$ в *Блоке преобразований* с параметрами $a=–d$, b=1 (см. разд. 3.4).

В случае в) необходимо подготовить вспомогательную парную выборку, все элементы которой равны d.

Действия и результаты. Для анализа нужно выбрать из электронной таблицы две или несколько переменных (см. рис. 2.3).

В результатах вычисляются:

- значение статистики *W* Вилкоксона (*Wilcoxon*) — сумма рангов R_{Xi} элементов одной из выборок в объединенной ранжированной выборке;
- значение статистики *V* Ван дер Вардена (*van der Varden*), основанную на использовании метода «произвольных меток».

Когда объемы двух выборок совпадают, дополнительно вычисляются следующие две статистики, отвечающие более мощным критериям, применимым в случае парных данных (о парных данных см. разд. 5.2, в западных статпакетах используются наименования *одновыборочный и двухвыборочный критерии*):

- значение статистики W_1 Вилкоксона — сумма рангов абсолютных значений разностей парных элементов двух выборок, вычисленная для положительных разностей;
- значение статистики знаков *S*, определенное как число положительных разностей парных элементов двух выборок.

Для каждой статистики вычисляется нормальная аппроксимация (*Z*–статистика) и уровень значимости *P* нулевой гипотезы об отсутствии различий в сдвиге двух выборок по отношению друг к другу. Если $P>0.05$, нулевая гипотеза может быть принята.

В случае нескольких выбранных переменных или задания конкретных пар переменных в текстовом файле (см. рис. 2.3, кнопка с пиктограммой чтения) указанные вычисления производятся для всех пар переменных. Значимости нулевых гипотез по Вилкоксону для последующего использования записываются в текстовые файлы Непар.TXT, Пар.TXT (непарные и парные выборки).

Пример 1

Задача. В эксперименте измерены уровни радиоактивности для двух групп препаратов (табл. 7.2.1, переменные WT1, WT2 из файла NPT).

Таблица 7.2.1. Уровни радиоактивности двух препаратов [имп/с]

Препарат 1	340 343 322 349 332 320 313 304 329
Препарат 2	318 321 318 301 312

Необходимо оценить достоверность различий между этими препаратами в средних значениях.

Результаты:

```
КРИТЕРИИ СДВИГА (ПОЛОЖЕНИЯ). Файл: npt.txt Переменные: wt1, wt2
Вилкоксон=82, Z=-1.935, Значимость=0.0264, степ.своб = 9,5
   Гипотеза 1: <Есть различия между медианами выборок>
Ван дер Варден=2.944, Z=1.902, Значимость=0.0285, степ.своб = 9,5
   Гипотеза 1: <Есть различия между медианами выборок>
```

Выводы: Применение критериев различия сдвига Вилкоксона и Ван дер Вардена, как можно видеть (полученные уровни значимости 0.0266 и 0.0285 меньше 0.05), позволяют принять гипотезу о различии между препаратами.

Пример 2

Задача. В эксперименте оценивалась светочувствительность куриных эмбрионов в темноте и на свету (по числу клевков по скорлупе). Измерения проводились попеременно у 25 эмбрионов (табл. 7.2.2, переменные SIG1, SIG2 в файле NPT).

Ожидается, что реакции на световой стимул будет соответствовать положительный сдвиг. Необходимо проверить это предположение.

Таблица 7.2.2. Светочувствительность 25 куриных эмбрионов в темноте и на свету [числу клевков в минуту по скорлупе]

Свет	6	14	26	7	8	23	11	9	19	26	18	18	18	14
Темнота	5	21	73	25	3	77	59	13	36	46	9	25	59	38

Свет	55	15	30	21	27	8	24	21	18	23	31
Темнота	70	36	55	46	25	30	39	46	71	31	33

Результаты:

```
КРИТЕРИИ СДВИГА (ПОЛОЖЕНИЯ). Файл: npt.txt Переменные: sig1, sig2
Вилкоксон=455, Z=3.54, Значимость=0.000199, степ.своб = 25,25
   Гипотеза 1: <Есть различия между медианами выборок>
```

```
Ван-дер-Варден=-10.7, Z=-3.19, Значимость=0.0008, степ.своб=25,25
   Гипотеза 1: <Есть различия между медианами выборок>
Для парных данных:
Вилкоксон=17.5, Z=-3.9, Значимость=4.8E-5, степ.своб = 2,25
   Гипотеза 1: <Есть различия между медианами выборок>
Знаков=4, Z=-3.2, Значимость=0.0007, степ.своб = 2,25
   Гипотеза 1: <Есть различия между медианами выборок>
```

Выводы: Анализируемые данные представляют парные переменные, поскольку измерения проводились у одних и тех же эмбрионов. Результаты применения критериев сдвига для парных выборок выявляют различие между двумя выборками (вычисленные уровни значимости 4.8E–5 и 0.0007 существенно меньше значения 0.05), что говорит о наличии реакции на свет у эмбрионов.

Пример 3

Задача. Применим данный метод для проверки гипотезы о том, что медиана совокупности, представленной выборкой SIG1 *(из Примера 2*) имеет значение 20. Для этого введем в переменную MSIG 25 константу со значением 20.

Результаты:

```
КРИТЕРИИ СДВИГА (ПОЛОЖЕНИЯ). Файл: npt.txt Переменные: sig1, msig
Вилкоксон=600, Z=0.7778, Значимость=0.2183, степ.своб = 25,25
   Гипотеза 0: <Нет различий между медианами выборок>
Ван дерВарден=-1.981, Z=-0.5916, Значимость=0.277, степ.своб=25,25
   Гипотеза 0: <Нет различий между медианами выборок>
Для парных данных:
Вилкоксон=137.5, Z=-0.6727, Значимость=0.2505, степ.своб = 2,25
   Гипотеза 0: <Нет различий между медианами выборок>
Знаков=11, Z=-0.4, Значимость=0.3445, степ.своб = 2,25
   Гипотеза 0: <Нет различий между медианами выборок>
```

Выводы: Результаты анализа позволяют принять нулевую гипотезу о равенстве медианы значению 20.

7.3. Критерии различия масштаба (рассеяния)

Назначение. Представленные здесь методы основаны на предположении о равенстве *медиан* сравниваемых выборок и направлены:

- на проверку гипотезы об отсутствии различий в *масштабах* (в разбросе или рассеянии значений) двух выборок из независимых совокупностей, например, наблюдений одних объектов без «обработки» и других объектов после обработки с анализом изменения рассеяния значений второй выборки как результата обработки;
- на проверку гипотезы о том, что отношение масштабов выборок равно заданной величине g.

Предварительные преобразования. В последнем случае необходимо предварительно изменить значения второй выборки *Y*: $y_i=(y_i-m_0)/g$, где m_0 — общая медиана двух выборок. Для этого следует в *Блоке преобразований* дважды выполнить операцию линейного преобразования: первый раз — с параметрами $a=-m_0$, $b=1$ и второй раз — с параметрами $a=0$, $b=1/g$ (см. разд. 3.4).

Если медианы генеральных совокупностей, из которых извлечены выборки, не равны по величине, но их значения известны: m_1, m_0, то настоящий метод можно применить, предварительно модифицировав одну из выборок, например выборку *Y* по формуле $y_i=y_i-m_2+m_1$. Это можно сделать посредством операции линейного преобразования в *Блоке преобразований* с параметрами $a=m_1-m_0$, $b=1$.

Если же медианы не равны и неизвестны, то следует подтвердить гипотезу об отсутствии различий сдвига (см. разд.. 7.2) или же использовать метод для обнаружения произвольных альтернатив (см. разд. 7.4).

Действия и результаты. Для анализа нужно выбрать из электронной таблицы две или несколько переменных (см. рис. 2.3).

Вычисляются значения статистик *W* Ансари–Бредли (*Ansari–Bradly*) и *K* Клотца (*Klotz*), которые являются концептуальными аналогами статистик Вилкоксона и Ван дер Вардена (см. разд. 7.2).

Для каждой исходной статистики вычисляется нормальная аппроксимация (*Z*–статистика) и уровень значимости *P* нулевой гипотезы о отсутствии различий в разбросе значений двух выборок. Если $P>0.05$, нулевая гипотеза может быть принята.

В случае нескольких выбранных переменных подобные вычисления производятся для всех пар переменных.

Связи. Процедура учитывает наличие *связей* между выборочными значениями: одинаковые ранги заменяются средними рангами.

Ограничения: размер выборки должен быть не больше *l*, где *l* = 16000, 5000, 1000, 100 при объеме матрицы данных в 64000, 20000, 4000 и 400 чисел.

Пример

Задача. Был произведен полумикроанализ на железо 20 препаратов железистой сыворотки (применяется от малокровия) с использованием традиционного и нового методов (табл. 7.3.1, переменные AB1, AB2 в файле NPT).

Внедрение нового метода возможно в случае, когда он не приводит к существенному ухудшению точности измерений, выраженному в возрастании разброса результатов измерений. Необходимо проверить эту гипотезу.

Таблица 7.3.1. Результаты полумикроанализа на железо 20 препаратов железистой сыворотки

Метод1	111	107	100	99	102	106	109	108	104	99
Метод2	107	108	106	98	105	103	110	105	104	100
Метод1	101	96	97	102	107	113	116	113	110	98
Метод2	96	108	103	104	114	114	113	108	106	99

Результаты:

```
КРИТЕРИИ СДВИГА (ПОЛОЖЕНИЯ). Файл: npt.std Переменные: ab1, ab2
Для парных данных:
Вилкоксон=86, Z=-0.364, Значимость=0.358, степ.своб = 2,19
   Гипотеза 0: <Нет различий между медианами выборок>
Знаков=8, Z=-0.459, Значимость=0.323, степ.своб = 2,19
   Гипотеза 0: <Нет различий между медианами выборок>

КРИТЕРИИ МАСШТАБА (РАССЕЯНИЯ). Файл: npt.std Переменные: ab1, ab2
Ансари-Бредли=185.5, Z=-1.327, Значимость=0.0922, степ.своб=20,20
   Гипотеза 0: <Нет различий между выборками в масштабах>
Клотц=19.2, Z=0.6874, Значимость=0.2459, степ.своб = 20,20
   Гипотеза 0: <Нет различий между выборками в масштабах>
```

Выводы: Применение критериев сдвига (результаты здесь не приводятся) позволяет принять нулевую гипотезу отсутствия различий на уровне значимости 0.02. Поэтому применение критериев масштаба Ансари–Бредли и Клотца к этим данным допустимо. Их результаты показывают отсутствие достоверных различий в разбросе значений сравниваемых выборок (полученные уровни значимости 0.0922 и 0.246 больше 0.05), т. е. точность измерений не уменьшается.

7.4. Критерии интегральных различий

Назначение. Критерии этого класса предназначены для обнаружения всех возможных отклонений от гипотезы об идентичности двух совокупностей.

Действия и результаты. Для анализа нужно выбрать из электронной таблицы две или несколько переменных (см. рис. 2.3).

Вычисляются значение статистики *D* Смирнова (часто ошибочно называемой статистикой Колмогорова–Смирнова) и уровень значимости *P* нулевой гипотезы об отсутствии интегральных различий между выборками. Если *P*>0.05, нулевая гипотеза может быть принята.

В случае нескольких выбранных переменных подобные вычисления производятся для всех пар переменных

Ограничения: размер выборки должен быть не больше 16000, 5000, 1000, 100 при объеме матрицы данных в 64000, 20000, 4000 и 400 чисел.

Пример

Задача. Был измерен привес свиней при двух различных рационах питания (табл. 7.4.1, переменные KST1, KST1 в файле NPT):

Таблица 7.4.1. Привес свиней при двух рационах питания

Рацион1	11.5, 26, 29.1, 19.7, 2.3, 22.6, 30.9, 10.8, 23.2, 38.8, 21.5
Рацион2	18.4, 15.5, 25.2, 16.9, 24, 13.3, 17.9, 13.2

Требуется оценить достоверность различий этих двух рационов.

Результаты:

```
КРИТЕРИЙ КОЛМОГОРОВА-СМИРНОВА. Файл:npt.txt Переменные: kst1, kst2
Смирнов=0.4773, Значимость=0.2425, степ.своб = 11,8
   Гипотеза 0: <Нет интегральных различий между выборками>
```

В ы в о д ы: Применение критерия Смирнова позволяет принять гипотезу об отсутствии различий между этими двумя рационами (вычисленный уровень значимости 0.2425 существенно больше 0.05).

7.5. Ранговая корреляция

Назначение. Свободные от распределения методы оценки *корреляции* (понятие корреляционной связи переменных рассмотрено в разд. 6.3) предназначены для проверки гипотезы о некоррелируемости (независимость, отсутствие соответствия или ассоциативности) двух *парных переменных*, извлеченных из непрерывной двумерной совокупности. Эти методы используют коэффициент ранговой корреляции Спирмена и коэффициент конкордации Кенделла, также они применимы и к ранговым данным. Связь номинальных переменных оценивается по методу кросстабуляции (разд. 7.6).

Действия и результаты. Для анализа нужно выбрать из электронной таблицы две или несколько переменных (см. рис. 2.3). В результатах вычисляются:

- коэффициент *конкордации t* Кенделла (*Kendall*), вычисляемый как число всех пар значений одной выборки, для которых соответствующие пары значений другой выборки имеют одинаковую тенденцию (возрастание или уменьшение значений), минус число пар с противоположной тенденцией, деленное на общее число пар;
- коэффициент ранговой корреляции *r* Спирмена (*Spearman*), который является непараметрическим эквивалентом коэффициента корреляции Пирсона применительно к предварительно ранжированным выборкам (при условии отсутствия совпадающих рангов); он обладает большей чувствительностью в случаях: а) асимметричных выборок; б) выборок, связанных монотонной, но нелинейной зависимостью.

Для каждого коэффициента вычисляется нормальная аппроксимация (*Z*–статистика) и уровень значимости *P* гипотезы о равенстве нулю коэффициента корреляции. Если $P>0.05$, нулевая гипотеза может быть принята.

В случае нескольких выбранных переменных подобные вычисления производятся для всех пар переменных с дополнительной выдачей диагональной матрицы коэффициентов корреляции. Эта матрица по подтверждению может быть сохранена в электронной таблице для последующего использования (например, в разделах многомерной статистики).

Множественные сравнения. Нередко производится вычисление ранговых корреляций между некоторым набором выборок, в результате чего делаются общие выводы. В этом случае необходима коррекция критического уровня значимости (см. в разд. 5.4).

Ограничения: размер выборки должен быть не больше l, где l = 16000, 5000, 1000, 100 при объеме матрицы данных в 64000, 20000, 4000 и 400 чисел.

где: m, l — число последовательностей совпадающих рангов в выборках X,Y; m_{xj}, m_{yj} — длины этих последовательностей.

Пример

З а д а ч а. Была измерена степень усвояемости (L–Хантера) для 9 партий консервированного тунца (переменная NCOR1 в файле NPT) и параллельно были получены результаты опроса по 6–балльной шкале, усредненные по 80 потребителям (переменная NCOR2).

Таблица 7.5.1. Оценки усвояемости для 10 партий консервированного тунца

L-Хантера	44.4	45.9	41.9	53.3	44.7	44.1	50.7	45.2	60.1
Опрос	2.6	3.1	2.5	5	3.6	4	5.2	2.8	3.8

Исходно предполагается, что мера Хантера положительно связана с баллами опроса. Необходимо подтвердить или опровергнуть это предположение.

Р е з у л ь т а т ы:

```
НЕПАРАМЕТРИЧЕСКАЯ КОРРЕЛЯЦИЯ.  Файл: npt.txt
                          Переменные: ncor1, ncor2
Кенделл=0.4444, Z=1.668, Значимость=0.0476, степ.своб = 9
    Гипотеза 1: <Есть корреляция между выборками>
Спирмен=0.6, Z=1.691, Значимость=0.0453, степ.своб = 9
    Гипотеза 1: <Есть корреляция между выборками>
```

В ы в о д ы: Исходное предположение подтверждается результатами анализа, согласно которым нулевая гипотеза об отсутствии связи между двумя признаками может быть отвергнута на уровнях значимости, равной 0.0476, 0.0453.

Примечание: Точные таблицы распределения Кенделла дают для этого примера уровень значимости 0.006 вместо асимптотического 0.0476.

7.6. Анализ таблиц сопряженности

Назначение. Данный метод предназначен для анализа двумерных *таблиц сопряженности* или *кросстабуляции* двух ранговых или номинальных переменных (*признаков*) с проверкой гипотезы о независимости переменных.

Примером двух номинальных переменных могут служить: пол с градациями мужской и женский, цвет волос с градациями светлый, темный, рыжий для некоторой исследуемой группы людей.

Анализ таблиц сопряженности размера 2×2 может быть осуществлен также методом сравнения частот событий (разд. 12.4), когда попарно проверяются различия частот встречаемости событий в клетках таблицы сопряженности (в отношениях к общему числу событий).

Исходные данные могут быть представлены двумя способами:

1) в виде готовой таблицы кросстабуляции размером $m \cdot n$, в которой столбцы отвечают различным значениям первой переменной (i=1–n),

строки — различным значениям второй переменной (j=1–m), а каждый элемент таблицы указывает количество объектов из исследуемой популяции, имеющих соответствующее сочетание значений переменных;

2) в виде значений двух парных (целочисленных или номинальных) переменных для N объектов, в этом случае процедура предварительно создает таблицу кросстабуляции; значения переменных в матрице данных должны быть представлены в виде целых чисел (номинальных кодов) или же символьных номинальных обозначений, объединение этих двух форм представления данных недопустимо.

Процедура распознает второй тип данных по наличию только двух переменных с числом значений, большим 5. Поэтому, во избежание коллизии, таблицы кросстабуляции 2×5 и больше следует представлять большей размерностью по горизонтали.

Результаты представляются в виде шести таблиц:

1) наблюденные частоты признаков x_{ij} (таблица кросстабуляции);
2) процентные частоты признаков для рядов;
3) процентные частоты признаков для столбцов;
4) процентные частоты признаков для всей таблицы;
5) ожидаемые частоты E_{ij} признаков в предположении их независимости;
6) остаточные частоты x_{ij}–E_{ij}.

Эти таблицы традиционны для данного метода, но им редко уделяется пристальное внимание. Перед таблицами приводятся номинальные обозначения для столбцов и строк.

Далее вычисляется ряд статистик, используемых для оценки различных аспектов связи между двумя номинальными переменными (*признаками*, см. формулы).

```
        40.9     15.9|     56.8%
        13.6     29.5|     43.2%
        -------------|
        54.5%     45.5%
          Ожидаемые частоты признаков:
        13.6     11.4
        10.4     8.64
          Остаточные частоты признаков (набл-ожид):
        4.36    -4.36
       -4.36     4.36
Хи-квадрат =7.19, Значимость=0.00734, степ.своб = 1
   Гипотеза 1: <Есть связь между признаками>
V-коэфф.Краммера =0.404
Ламбда Гудмана и Крускала: симметр, ряд, столб =0.333, 0.316, 0.35
   Тау-b Кенделла =0.402   Тау-c Кенделла =0.397
   Гамма Гудмана и Кенделла =0.696  d(x,y)Соммера=0.404, 0.4
```

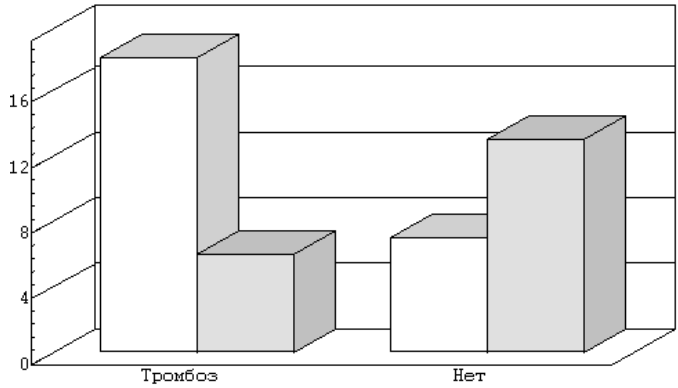


Рис. 7.2. Столбиковая диаграмма кросстабуляции: тромбоз – аспирин

В ы в о д ы: Вычисленный уровень значимости критерия хи–квадрат позволяет принять гипотезу о зависимости тромбоза от приема или неприема аспирина (P=0.007 меньше 0.05). Значения последующих коэффициентов раскрывают различные аспекты выявленной взаимосвязи этих признаков, согласно их вышерассмотренным свойствам. Диаграмма (рис. 7.2) показывает значительную симметричную связь между признаками: большим случаям тромбоза при контроле соответствуют меньше случаев при приеме аспирина. Эту симметрию отражают и большие значения ряда коэффициентов (лямбда, тау-c, гамма и $d(x,y)$).

Данный пример может быть также проанализирован методом попарного сравнения частот событий (разд. 12.4) в клетках таблицы, для чего надо предварительно разделить число событий на число испытуемых каждой категории.

П р и м е р 2

З а д а ч а. Продолжим анализ антропометрических данных из примеров к разделам 6.1–6.4 (файл PEOPLE) с целью исследования связи ме-

жду цветом глаз и цветом волос (предварительно из матрицы данных следует удалить все переменные, кроме двух анализируемых). Здесь мы имеем в качестве исходных данных не готовую таблицу кросстабуляции, а две нативные переменные (глаза, волосы), причем их значения выражены не в числовой, а в номинальной шкале (синий, зеленый, коричневый, черный, светлый, каштановый).

Р е з у л ь т а т ы (сокращенно):

```
   КРОССТАБУЛЯЦИЯ.  Файл: people.std
   Столбцы: светлый русый черный коричн
   Строки: синий серый коричн
           Наблюденные частоты признаков
          26       10        3        2|       41
          14       18       11        4|       47
           2        6        3        1|       12
   -----------------------------------|
          42       34       17        7|      100
Хи-квадрат =14.5, Значимость=0.0241, степ.своб = 6
   Гипотеза 1: <Есть связь между признаками>
Коэфф. Фи =0.381   Коэфф.сопряж. Пирсона =0.356
V-коэфф. Граммера =0.27
Лямбда Гудмана и Крускала: симметр,ряд,столб =0.18, 0,0566, 0.293
Tay-b Кенделла =0.312   Tay-c Кенделла =0.297
Гамма Гудмана и Кенделла =0.471  d(x,y)Соммера=0.332, 0.294
```

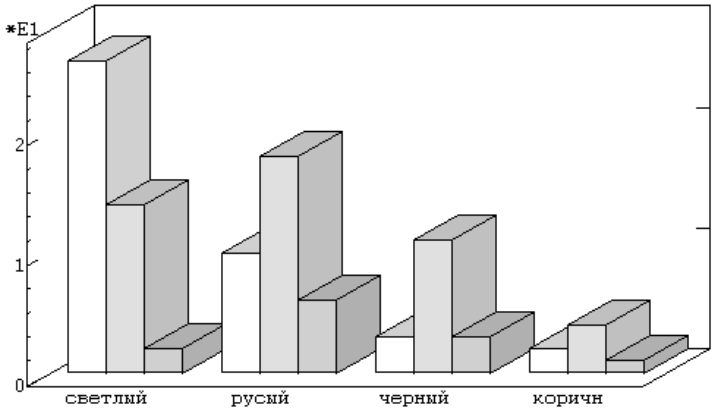


Рис. 7.3. Столбиковая диаграмма кросстабуляции: цвет глаз – цвет волос

В ы в о д ы: Результаты анализа свидетельствуют, что имеет место несомненная связь между цветом глаз и цветом волос (P=0.024 меньше 0.05). Диаграмма (рис. 7.3) показывает значительную асимметричность связи между признаками: большие-меньшие встречаемости цвета глаз воспроизводятся для большинства цвета волос. Это отражают и сравнительно меньшие значения коэффициентов симметрии *лямбда*.

8.2. Однофакторный дисперсионный анализ

8.2.1. Параметрические методы

Назначение. С помощью данного метода в зависимости от типа модели по исследуемому фактору (с фиксированными или же со случайны-

ми эффектами) на основе параметрического критерия Фишера проверяется одна из двух нулевых гипотез:

- средние значения для групп откликов, измеренных при различных значениях фактора, не имеют существенных различий между собой (*модель 1*);
- дисперсия средних значений для групп откликов, измеренных при различных значениях фактора, не отлична от нуля (*модель 2*).

В случае наличия факторного эффекта нередко представляет интерес более детальный анализ на наличие различий между конкретными уровнями фактора или группами уровней. Эту задачу решает метод парных сравнений Шеффе (*Scheffe*), применимость которого предполагает условие гомогенности (однородности) дисперсий (нулевую гипотезу отсутствия различия дисперсий можно проверить средствами разд. 7.3).

Исходные данные представляются в виде псевдоматрицы (т. е. столбцы не обязаны быть одинаковой длины), в которой переменные отвечают различным уровням исследуемого фактора и каждая переменная содержит отклики, измеренные при соответствующем значении фактора.

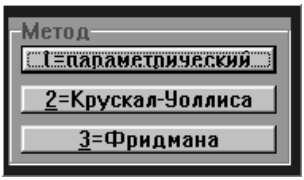


Рис. 8.5. Меню выбора метода однофакторного анализа

Действия и результаты. Сначала нужно выбрать из электронной таблицы переменные, соответствующие различным уровням фактора (см. бланк рис. 2.3), а затем выбрать параметрический метод анализа (меню — рис. 8.5).

Выдача результатов включает стандартную дисперсионную таблицу со столбцами: сумма квадратов, число степеней свободы, средняя сумма квадратов, сила влияния фактора (по Снедекору), а строки содержат межгрупповые, внутригрупповые и общие значения (см. формулы и примеры).

Далее вычисляется статистика Фишера F с уровнем значимости P. Если $P>0.05$, нулевая гипотеза об отсутствии влияния фактора может быть принята.

Затем выдаются значения параметров однофакторной модели m_x, a_i (факторные эффекты, см. разд. 8.1) с доверительными интервалами $d(m_x)$, $d(a_i)$. В случае наличия факторного эффекта выдается таблица парных сравнений Шеффе:

Переменные	Разность	Интервал	Значим	Гипотеза H1
1-2	0.395	0.4982	0.0183	Да
1-3	0.325	0.5252	0.4326	
1-4	0.09167	0.4794	0.9898	

...

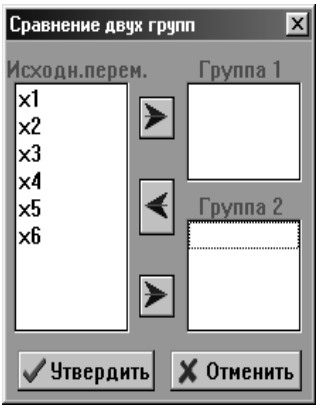


Рис. 8.6. Бланк выбора групп откликов для сравнений Шеффе

В этой таблице для всех пар уровней исследуемого фактора приведены следующие параметры (по столбцам): разность средних значений, размах доверительного интервала разности, уровень значимости нулевой гипотезы об отсутствии различий между средними значениями и рекомендации по принятию альтернативной гипотезы.

Далее можно продолжить анализ Шеффе уже по групповому сравнению факторного эффекта для двух выбранных групп откликов. Для этого в последующем бланке (рис. 8.6) необходимо сформировать две группы переменных из электронной таблицы. Такой бланк будет повторяться до его отмены.

=

с $N–k$ степенями свободы для уровня значимости α с оценкой по Т-критерию Стьюдента.

Доверительный интервал сравнения Шеффе между двумя группами откликов с числом измерений n_i, $n_j = \sqrt{\frac{k\,F_{k,N-k}\,SE}{(N-k)}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$.

Пример 1

З а д а ч а. В исследовании оценивалась урожайность четырех различных сортов пшеницы, выращиваемых на нескольких (различных по числу) участках примерно одного почвенного типа (табл. 8.2.1, файл A1). Необходимо выяснить, отличаются ли эти сорта по урожайности.

Таблица 8.2.1. Урожайность четырех сортов пшеницы [ц/га]

	Сорт 1	Сорт 2	Сорт 3	Сорт 4
Участки	17	15.8	17.4	15.7
	17.2	17	16.6	16.8
	16.1	16.4	16.2	15.1
	17		15.6	15.2
	16.8		15.5	
			17.2	

Р е з у л ь т а т ы:

```
   1-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: a1.std параметрический

Источник    Сум.квадр  Ст.своб Ср.квадр  Сила влияния
  Факт.1        2.824        3    0.9412     -0.1505
  Остат.        6.436       14    0.4597
  Общая          9.26       17    0.5447

F(фактор1)=2.047, Значимость=0.1528, степ.своб = 3,14
   Гипотеза 0: <Нет влияния фактора на отклик>
```

В ы в о д ы: Дисперсионный анализ показывает отсутствие заметного влияния на урожайность фактора сорта пшеницы при P=0.152.

Пример 2

З а д а ч а. Исследовались показатели прочности шести сплавов (в испытаниях нескольких образцов), из которых четвертый сплав является стандартным (табл. 8.2.2, файл SHEFFE). Следует верифицировать наличие общих различий между всеми сплавами, после чего оценить попарные различия.

Таблица 8.2.2. Показатели прочности шести легированных сплавов

	Сплав1	Сплав2	Сплав3	Стандарт	Сплав4	Сплав5
Об-	15,1	15,3	15,1	15,3	15,2	15,2
раз-	15	15,5	15,6	14,9	14,8	15,1
цы	15,4	15,7	15,4	15	15,4	15,3
	15	15,8	15,7	14,9	15	15,1
		15,3		15,2	15	
				14,9		

Р е з у л ь т а т ы:

```
1-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: sheffe.std
          параметрический

Источник   Сум.квадр  Ст.своб Ср.квадр Сила влияния
  Факт.1      1.013        5    0.2026     0.02548
  Остат.     0.9143       22   0.04156
  Общая       1.927       27   0.07138

F(фактор1)=4.874, Значимость=0.004, степ.своб = 5,22
   Гипотеза 1: <Есть влияние фактора на отклик>

             Парные сравнения Шеффе
Переменные  Разность  Интервал    Значим  Гипотеза H1
    1-2       0.395    0.4981    0.1836
    1-3       0.325     0.525    0.4326
    1-4     0.09167    0.4793    0.9898
    1-5       0.045    0.4981    0.999
    1-6        0.05     0.525    0.9989
    2-3        0.07    0.4981    0.9966
    2-4      0.4867    0.4496    0.0282       Да
    2-5        0.44    0.4696    0.0763
    2-6       0.345    0.4981    0.3104
    3-4      0.4167    0.4793    0.1172
    3-5        0.37    0.4981    0.241
    3-6       0.275     0.525    0.6117
    4-5     0.04667    0.4496    0.9986
    4-6      0.1417    0.4793    0.9432
    5-6       0.095    0.4981    0.9899
x2,x3-x1,x5,x6:0.3583  0.3548    0.0413       Да
```

В ы в о д ы: Однофакторный дисперсионный анализ показывает, что сплавы в целом различны по фактору прочности на уровне значимости 0.004. Последующий анализ парных сравнений Шеффе выявляет значимые отличия сплава 2 от стандарта. Заметные отличия от стандарта можно наблюдать и у сплава 3. Тем самым эти два сплава, видимо, имеют структуру, отличную от группы сплавов 1, 5, 6. Дополнительное сравнение двух групп сплавов выявляет значимые различия между ними.

8.2.2. Непараметрические методы Крускала-Уоллиса и Джонкхриера

Назначение. Непараметрические (*ранговые*) методы однофакторного анализа для нескольких независимых выборок, полученных при различных уровнях исследуемого фактора, оценивают факторный эффект с помощью двух различных подходов.

результат

Исходные данные представляются в виде псевдоматрицы x_{ij}, которая в каждом своем столбце содержит измерения, произведенные при соответствующем уровне исследуемого фактора. Число измерений в каждом столбце может быть различным.

В случае использования критерия Джонкхиера выборки должны быть расположены по предполагаемому возрастанию факторного эффекта.

Если исходные данные не приведены в ранговую форму, то процедура выполняет их общее (сквозное) ранжирование.

Действия и результаты. Обращение к анализу Крускала–Уоллиса производится из пункта однофакторного дисперсионного анализа с уточнением метода в последующем меню (рис. 8.5).

Вычисляется *H*–статистика Крускала–Уоллиса с уровнем значимости *P*, а в случае утвердительного ответа на вопрос об упорядоченности факторных эффектов по столбцам матрицы данных (по возрастанию) — также и *J*–статистика Джонкхиера.

Если $P>0.05$, соответствующая нулевая гипотеза может быть принята. Такой вывод корректен, когда число объектов и значений фактора достаточно велико, в противном случае желательно сравнить значение статистики с критическим в таблице соответствующего распределения из статистического справочника.

Связи. Процедура учитывает наличие *связей* в виде одинаковых рангов, производя коррекцию *H*–статистики.

Ограничение. Число измерений не должно превышать 16000, 5000, 1000, 100 при объеме матрицы данных в 64000, 20000, 4000 и 400 чисел.

Пример

Задача. В эксперименте измерялось время выполнения монотонных производственных операций в зависимости от мотивационного фактора. Было сформировано три группы рабочих: группа 1 — контрольная; рабочие группы 2 получали только общие сведения о требуемой производительности, а рабочие группы 3 получили полную информацию, включая карты пооперационной разбивки работы (табл. 8.2.3, файл JON). Требуется оценить, влияет ли полученная информация о производительности операций на саму производительность труда.

Таблица 8.2.3. Время выполнения монотонных операций [с] в зависимости от полноты предварительно получаемой информации

	Контроль	Общая информация	Полная информация
Рабочие (разные)	40	38	48
	35	40	40
	38	47	45
	43	44	43
	44	40	46
	41	42	44

Результаты:

```
1-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: jon.std

Краскал-Уоллис=4.361, Значимость=0.1129, степ.своб = 2
   Гипотеза 0: <Нет влияния фактора на отклик>
Джонкхиер=79, Значимость=0.0216, степ.своб = 3,18
   Гипотеза 1: <Есть влияние фактора на отклик>
```

Выводы: Хотя критерий Крускала–Уоллиса не выявляет влияния основного фактора, но более конкретизированный анализ с использованием критерия Джонкхиера выявляет значимое влиянием уровня чистой мотивации на производительность труда.

Примечание: Точные таблицы распределения Крускала–Уоллиса дают для этого примера уровень значимости 0.139 вместо асимптотического 0.1129

8.2.3. Непараметрические методы Фридмана и Пейджа

Назначение. Рассмотренные здесь методы часто относят к двухфакторной модели дисперсионного анализа. На самом же деле они реализуют непараметрические аналоги однофакторного анализа групповых измере-

ний (см. разд. 8.4), когда действию различных уровней фактора последовательно подвергается каждый из заданного набора объектов.

Исходные данные представляются в виде матрицы, в которой столбцы ($j = 1,...,m$) соответствуют различным значениям основного фактора, а строки ($i = 1,...,n$) — различным значениям второго фактора, а ранжирование должно быть отдельным для каждой строки.

В случае использования критерия Пейджа обработки (столбцы) должны быть упорядочены по предполагаемому возрастанию факторного эффекта.

Если исходные данные не приведены в ранговую форму, то процедура выполняет их ранжирование отдельно по измерениям.

Действия и результаты. Обращение к данной процедуре производится из пункта однофакторного дисперсионного анализа с уточнением метода в меню (рис. 8.5).

Вычисляется статистика S Фридмана с уровнем значимости P, а в случае утвердительного ответа на вопрос об упорядоченности факторных эффектов по столбцам матрицы данных (по возрастанию) — также и статистика L–Пейджа.

Если $P > 0.05$, соответствующая нулевая гипотеза может быть принята. Такой вывод корректен, когда число значений обоих факторов достаточно велико, в противном случае желательно сравнить вычисленную статистику с критическим значением в таблице распределения Фридмана или Пейджа из статистического справочника.

Пример 1

Задача. В эксперименте измерялось среднее время выточки детали на станке. Работу выполняли 22 рабочих попеременно с использованием трех различных методов обработки (табл. 8.2.4, в файле FREA представлены три переменные *Метод1, Метод2, Метод3* с 22 измерениями каждая). Требуется оценить, различаются ли эти три метода в плане влияние на производительность работы.

Таблица 8.2.4. Время выточки детали [мин] при использовании трех методов работы

Рабочий	1	2	3	4	5	6	7	8	9	10	11
Метод1	5,4	5,85	5,2	5,55	5,9	5,45	5,4	5,45	5,25	5,85	5,25
Метод2	5,5	5,7	5,6	5,5	5,85	5,55	5,4	5,5	5,15	5,8	5,2
Метод3	5,55	5,75	5,5	5,4	5,7	5,6	5,35	5,35	5	5,7	5,1

Рабочий	12	13	14	15	16	17	18	19	20	21	22
Метод1	5,65	5,6	5,05	5,5	5,45	5,55	5,45	5,5	5,65	5,7	6,3
Метод2	5,55	5,35	5	5,5	5,55	5,55	5,5	5,45	5,6	5,65	6,3
Метод3	5,45	5,45	5,95	5,4	5,5	5,35	5,55	5,25	5,4	5,55	6,26

Результаты:

```
2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: frea.std

Фридман=11.14, Значимость=0.0038, степ.своб = 2
   Гипотеза 1: <Есть влияние фактора на отклик>
```

Выводы: Результат анализа позволяет принять гипотезу о различии трех методов обработки по влиянию на производительность труда на уровне значимости 0.0038.

Пример 2

Задача. В эксперименте измерялась прочность хлопка в зависимости от количества калийного удобрения, внесенного в почву при его выращивании. Исследование проводилось на трех различных полях, на каждом из которых выделено пять делянок для испытаний пяти возрастающих доз внесения калийного удобрения (табл. 8.2.5, файл PAGE).

Таблица 8.2.5. Прочность хлопка в зависимости от дозы внесения в почву калийного удобрения

	Доза 1	Доза 2	Доза 3	Доза 4	Доза 5
Поле 1	7,46	7,17	7,76	8,14	7,62
Поле 2	7,68	7,57	7,73	8,15	8
Поле 3	7,21	7,80	7,74	7,87	7,93

Р е з у л ь т а т ы:

```
2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: page.std

Фридман=8.8, Значимость=0.0662, степ.своб = 4
   Гипотеза 0: <Нет влияния фактора на отклик>
Пейдж=158, Значимость=0.0039, степ.своб = 5,3
   Гипотеза 1: <Есть влияние фактора на отклик>
```

В ы в о д ы: Хотя критерий Фридмана не выявляет влияния основного фактора, но более конкретизированный анализ с использованием критерия Пейджа выявляет значимое увеличение прочности хлопка в зависимости от количества вносимого в почву калийного удобрения на уровне значимости 0.039.

Примечание: Точные таблицы распределения Фридмана дают для этого примера уровень значимости 0.0725 вместо асимптотического 0.0662.

8.3. Двухфакторный дисперсионный анализ

Назначение. Посредством данного метода в зависимости от типа модели по каждому фактору (с фиксированными или же со случайными эффектами) с помощью параметрического критерия Фишера проверяется одна из двух нулевых гипотез:

- средние значения для групп откликов, измеренных при различных значениях фактора, не имеют существенных различий между собой (*модель 1*);
- дисперсия средних значений для групп откликов, измеренных при различных значениях фактора, не отлична от нуля (*модель 2*).

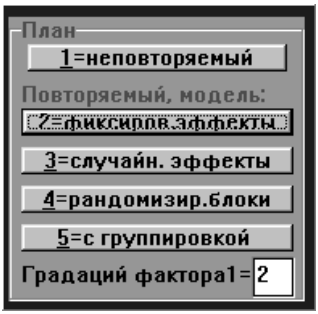


Рис. 8.7. Меню методов двухфакторного дисперсионного анализа

Разновидности метода. Имеется две разновидности метода в зависимости от того, производились ли *повторные измерения* при каждом сочетании значений двух исследуемых факторов или нет: уточнение метода производится из меню выбора (рис. 8.7).

1. Нет повторных измерений. При эксперименте без повторных измерений исходные данные должны представлять собой матрицу размером $n \cdot m$, в которой столбцы отвечают различным уровням первого фактора $j = 1,...,m$, а строки — различным уровням второго фактора $i = 1,...,n$, а каждая ячейка содержит один отклик, измеренный при соответствующем сочетании уровней исследуемых факторов.

Выдача результатов включает дисперсионную таблицу со столбцами: сумма квадратов, число степеней свободы, средняя сумма квадратов, сила влияния фактора (по Снедекору), а строки содержат значения для первого и второго факторов, а также остаточные и общие параметры (см. формулы и примеры).

Далее для каждого фактора вычисляется статистика Фишера F с уровнем значимости P. Если $P>0.05$, нулевая гипотеза об отсутствии влияния соответствующего фактора может быть принята.

2. Есть повторные измерения. При эксперименте с повторными измерениями исходные данные должны представлять собой псевдоматрицу (не обязательно одинаковой длины столбцов), в которой переменные (столбцы $i = 1,...,n{\cdot}m$) располагаются в порядке изменения уровней сначала первого фактора A, а затем второго фактора B, а именно: A_1B_1, A_2B_1, ..., A_1B_2, A_2B_2, ...

Поскольку такое представление данных может отвечать различным сочетаниям числа градаций факторов, то в поле ввода меню выбора метода (рис. 8.7) необходимо указать число уровней первого фактора, после чего нажать кнопку используемой модели:

0 = с фиксированными эффектами;
1 = со случайными эффектами;
2 = с рандомизированными блоками;
3 = с группировкой.

Выдача результатов включает дисперсионную таблицу со столбцами: сумма квадратов, число степеней свободы, средняя сумма квадратов, сила влияния фактора (по Снедекору), а строки содержат значения для первого и для второго факторов, для эффекта межфакторного взаимодействия, а также остаточные и общие параметры (см. формулы и примеры).

Далее для каждого факторного эффекта вычисляется статистика Фишера F с уровнем значимости P. Если $P>0.05$, нулевая гипотеза об отсутствии соответствующего факторного эффекта может быть принята. Если эффект взаимодействия не обнаружен, то проводится дополнительный анализ по факторам A и B, но без учета их взаимодействия. Такой дополнительный анализ, как правило, дает более низкий уровень значимости нулевых гипотез. Полученными результатами рекомендуется пользоваться, если уровень значимости гипотезы отсутствия взаимодействия факторов достаточно высок ($P>0.1$).

Затем выдаются значения параметров однофакторной модели m_x, a_i, b_i (факторные эффекты, см. разд. 8.1) с доверительными интервалами $d(m_x)$, $d(a_i)$, $d(b_i)$.

Пример 1

З а д а ч а. В эксперименте фиксировалась урожайность пяти сортов картофеля, выращенных на пяти участках одинакового размера и почвенного состава, при этом каждый из этих участков обрабатывался одним из шести сортов удобрений (табл. 8.3.1, файл A2).

Таблица 8.3.1. Урожайность пяти сортов картофеля [ц/га] в зависимости от вида вносимого в почву удобрения

	Карт.1	Карт.2	Карт.3	Карт.4	Карт.5
Удобр.1	6	9	6	2	6
Удобр.2	4	7	8	3	5
Удобр.3	9	3	10	7	4
Удобр.4	8	4	14	4	10
Удобр.5	15	11	13	9	14
Удобр.6	12	14	15	11	9

Необходимо выяснить, различна ли в среднем урожайность разных сортов картофеля независимо от применяемого удобрения и различна ли эффективность используемых удобрений независимо от сорта.

Р е з у л ь т а т ы:

```
2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: a2.std
         Факторный план: неповторяемый

Источник   Сум.квадр  Ст.своб Ср.квадр  Сила влияния
  Факт.1      79.2        4     19.8      -0.3788
  Факт.2       248        5     49.6       0.6308
  Остат.       118       20      5.9
  Общая      445.2       29    15.35

F(фактор1)=3.356, Значимость=0.0291, степ.своб = 4,20
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор2)=8.407, Значимость=0.0003, степ.своб = 5,20
   Гипотеза 1: <Есть влияние фактора на отклик>
```

В ы в о д ы: Дисперсионный анализ выявляет существенное влияние сорта картофеля и вида удобрения на урожайность (уровни значимости 0.0291 и 0.0003 меньше 0.05).

П р и м е р 2

З а д а ч а. В эксперименте измерялось количество выдыхаемого азота для четырех режимов питания и для двух возрастных категорий пациентов. В исследовании участвовало по три разных пациента в каждой из восьми групп «возраст–диета» (табл. 8.3.2, в файле A2G имеется восемь столбцов с тремя измерениями: для диет 1–4 возраста 1 и диет 1–4 возраста 2). Необходимо определить влияние диет, возрастных групп и их взаимодействия на количество выдыхаемого азота.

Таблица 8.3.2. Объем выдыхаемого азота (в литрах) при четырех диетах питания и для двух возрастных групп пациентов

	Диета 1	Диета2	Диета3	Диета4
Возраст 1	4,097 4,859 3,54	4,368 5,668 3,752	4,169 5,709 4,416	4,928 5,608 4,94
Возраст 2	2,87 4,648 3,848	3,579 5,393 4,374	4,403 4,496 4,688	4,905 5,208 4,806

С иллюстративными целями ниже производятся два расчета для модели с фиксированными эффектами и для модели со случайными эффектами.

Результаты при фиксированных эффектах:

```
2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: a2g.std
Факторный план: повторяемый с фиксированными эффектами

Источник   Сум.квадр  Ст.своб Ср.квадр Сила влияния
  Факт.1       3.63        3     1.21   -0.216
  Факт.2      0.335        1    0.335   -0.332
  Межфак     0.0453        3   0.0151   -0.333
  Остат.       7.83       16    0.489
  Общая        11.8       23    0.515
```

```
F(фактор1)=2.47, Значимость=0.0984, степ.своб = 3,16
   Гипотеза 0: <Нет влияния фактора на отклик>
F(фактор2)=0.685, Значимость=0.51, степ.своб = 16
   Гипотеза 0: <Нет влияния фактора на отклик>
F(межфакт)=0.0308, Значимость=0.992, степ.своб = 3,16
   Гипотеза 0: <Нет влияния фактора на отклик>
  После объединения межфакторной и остаточной СК:
F(фактор1)=2.92, Значимость=0.06, степ.своб = 3,19
   Гипотеза 0: <Нет влияния фактора на отклик>
F(фактор2)=0.808, Значимость=0.566, степ.своб = 19
   Гипотеза 0: <Нет влияния фактора на отклик>

          Параметры модели:
Среднее = 20.9, доверит.инт.=1.8
Эффект1-1 = -16.2, доверит.инт.=14.3
Эффект1-1 = -16.5, доверит.инт.=14.3
Эффект2-1 = -16.9, доверит.инт.=4.02
Эффект2-2 = -16.4, доверит.инт.=4.02
Эффект2-3 = -16.2, доверит.инт.=4.02
Эффект2-4 = -15.8, доверит.инт.=4.02
```

Результаты при случайных эффектах:

```
2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: a2g.std
Факторный план: повторяемый, со случайными эффектами

Источник   Сум.квадр  Ст.своб Ср.квадр Сила влияния
  Факт.1        3.63        3     1.21   -0.216
  Факт.2       0.335        1    0.335   -0.332
  Межфак      0.0453        3   0.0151   -0.333
  Остат.        7.83       16    0.489
  Общая         11.8       23    0.515
F(фактор1)=80.1, Значимость=0.0023, степ.своб = 3,3
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор2)=22.2, Значимость=0.0002, степ.своб = 3
   Гипотеза 1: <Есть влияние фактора на отклик>
F(межфакт)=0.0308, Значимость=0.992, степ.своб = 3,16
   Гипотеза 0: <Нет влияния фактора на отклик>
  После объединения межфакторной и остаточной СК:
F(фактор1)=2.92, Значимость=0.06, степ.своб = 3,19
   Гипотеза 0: <Нет влияния фактора на отклик>
F(фактор2)=0.808, Значимость=0.566, степ.своб = 19
   Гипотеза 0: <Нет влияния фактора на отклик>
```

В ы в о д ы: Для модели с фиксированными эффектами не обнаружено значимых эффектов факторов или их взаимодействия. Для модели со случайными эффектами оба фактора оказались значимыми. Поскольку же их взаимодействие незначимо, то исследователю может показаться целесообразным произвести объединение, так как число степеней свободы остаточной дисперсии (знаменателя F-критерия) слишком мало для оценки эффектов. Однако в результате объединения оба фактора становятся незначимыми. Эти изменения в оценке значимости показывают, насколько различными могут оказаться результаты в зависимости от отношения исследователя к типу модели и объединению дисперсий.

8.4. Дисперсионный анализ групповых измерений

Назначение. В ряде областей исследования (в психологии, медицине, биологии, агрономии и некоторых других) достаточно часто встречаются ситуации, когда имеется фиксированная *группа объектов* исследования (subjects), которая последовательно подвергается действию различных уровней одного, двух или более факторов, в результате чего у каждого объекта измеряется по одному отклику для каждого уровня фактора. Такая схема эксперимента, очевидно, существенно отличается от классической схемы факторного исследования, когда измерения отклика при различных сочетаниях значений факторов производятся у различных объектов. Важным здесь является то, что отклики измеряются у одних и тех же объектов при различных уровнях факторов и несут в себе следы их субъективной вариабельности. Тем самым оказывается возможным вычислить эту субъективную вариабельность и удалить ее из остаточной (случайной) дисперсии, тем самым повысив достоверность выводов о влиянии главных факторов.

собственных делянок совершенно независимо от выбора делянок для высевания других сортов.

Разновидности метода. Выбор конкретного метода производится по меню рис. 8.8.

Поясним методы дисперсионного анализа групповых измерений[1] на примере условного эксперимента, в котором проверяется действие на группу пациентов некоторого постоянно принимаемого медицинского препарата с течением времени (фактор 1), оцениваемого измерением некоторого физиологического показателя (отклик или зависимая переменная) 1 раз в неделю. Если к этой схеме мы не делаем никаких дополнительных добавлений, то она соответствует *однофакторному групповому* эксперименту.

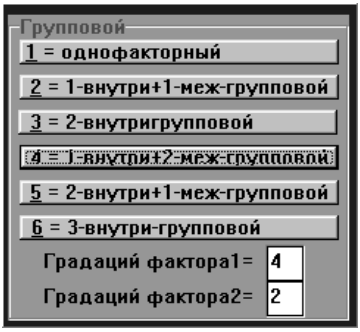


Рис. 8.8. Меню выбора метода группового дисперсионного анализа

Пусть теперь мы разделили пациентов на две (или более) группы: принимающих традиционный препарат и принимающих новый препарат. В этом случае мы имеем схему *одного внутригруппового* (within–subjects) фактора (неделя) и *одного межгруппового* (between–subjects) фактора (препарат). Если же одна и та же группа пациентов принимает сначала традиционный препарат, а затем новый, то получается схема с *двумя внутригрупповыми* факторами.

Пусть теперь мы разделили пациентов по полу (мужчины и женщины). Тогда мы имеем схему *трехфакторного* исследования, в которой два фактора (пол и препарат) являются *межгрупповыми* и один фактор (неделя) является *внутригрупповым*.

Предположим теперь, что все пациенты принимают сначала первый, а затем второй препарат. Тогда у нас получается схема *трехфакторного* исследования с *двумя внутригрупповыми* факторами (препарат и время) и *одним межгрупповым* фактором (пол).

Пусть теперь в последней схеме исследования мы не различаем пациентов по полу, но зато каждого пациента при приеме каждого препарата в течение некоторого фиксированного числа недель подвергаем сначала одному, а затем другому методу физиотерапии. В таком случае мы имеем схему эксперимента с *тремя внутригрупповыми* факторами.

---

[1] Реализованные в данном разделе методы следуют монографии: David C. Howell. Statistical Methods for Psychology – fourth edition. Duxbury Press, Belmont, 1997.

Примечание: Реальная практика дисперсионного анализа обычно ограничивается трехфакторными схемами в связи как с практическими ограничениями возможности постановки экспериментов большей факторности, так и с ограничениями осмысления их результатов и аргументации в обсуждениях с коллегами и оппонентами.

Исходные данные представляются в виде матрицы или псевдоматрицы, вид которой зависит от типа факторного эксперимента.

Однофакторный эксперимент. При этой схеме исходные данные должны представлять собой матрицу размером $m \cdot n$, в которой столбцы отвечают различным уровням первого фактора $j = 1,...,m$, по строкам расположены объекты $i = 1,...,n$, а каждая ячейка содержит один отклик, измеренный у данного объекта при соответствующем уровне фактора.

Схема с двумя факторами. При этой схеме исходные данные должны представлять собой матрицу размером $M \cdot n$, в которой по строкам расположены объекты $i = 1,...,n$, а по столбцам — сначала измерения для первого уровня первого фактора (внутригруппового для схемы со смешанными факторами или первого межгруппового для схемы с двумя межгрупповыми факторами) для всех последовательных уровней второго фактора, затем второго уровня первого фактора для всех последовательных уровнях второго фактора и т. д. Тем самым $M=ms$, где m — число уровней первого фактора, а s — число уровней второго фактора. При этих схемах в меню выбора метода анализа (рис. 8.8) необходимо ввести число уровней первого фактора.

Схема с тремя факторами. При этой схеме исходные данные должны представлять собой матрицу размером Mn, в которой по столбцам расположены объекты $i = 1,...,n$, а по строкам — измерения сначала для 1-го уровня первого фактора и 1-го уровня второго фактора при всех последовательных значениях уровня третьего фактора, затем измерения для 1-го уровня первого фактора и 2-го уровня второго фактора при всех последовательных значениях уровня третьего фактора и т. д., а в конце (по такой же схеме) — для последнего уровня первого фактора и последнего уровня второго фактора при всех последовательных значениях уровня третьего фактора. Тем самым $M=mst$, где m — число уровней первого фактора, s — число уровней второго фактора, t — число уровней третьего фактора. При этих схемах в меню выбора метода анализа (рис. 8.8) необходимо ввести число уровней первого и второго факторов.

Результаты анализа групповых измерений включают дисперсионную таблицу со столбцами: сумма квадратов, число степеней свободы, средняя сумма квадратов, а строки содержат значения для первого, второго и т. д. факторов, а также остаточные и общие параметры (см. формулы и примеры).

Далее для каждого фактора вычисляется статистика Фишера F и уровень значимости P. Если $P>0.05$, нулевая гипотеза об отсутствии влияния соответствующего фактора может быть принята.

Затем для визуальной проверки составной симметрии выдается матрица ковариаций и поправочные коэффициенты ε и ε'. Если эти коэффициенты не слишком велики по своим значениям: $\varepsilon<0.75$ или $\varepsilon>0.75$ и $\varepsilon'<0.98$, то проводится перепроверка нулевых гипотез при скорректированных степенях свободы.

Пример 1

Задача. Рассмотрим однофакторный эксперимент, в котором девять пациентов используют релаксационную методику для ослабления мигрирующей головной боли (фиксируется число приступов в неделю). Исследование проводилось в течение пяти недель, причем релаксационная техника использовалась только в последние две недели. Тем самым стоит задача выявить влияние фактора релаксации на фоне первых двух контрольных недель. Полученные данные приведены в табл. 8.4.1 (файл ANOVA_BS, где переменные (столбцы) соответствуют неделям, а строки — пациентам).

Таблица 8.4.1. Частота головных болей при релаксационной терапии

	Недели				
Пациенты	1	2	3	4	5
1	21	22	8	6	6
2	20	19	10	4	4
3	17	15	5	4	5
4	25	30	13	12	17
5	30	27	13	8	6
6	19	27	8	7	4
7	26	16	5	2	5
8	17	18	8	1	5
9	26	24	14	8	9

В меню рис. 8.8 выбираем «1-факторный» групповой анализ.

Результаты:

```
ГРУППОВОЙ 1-ФАКТОРНЫЙ АНАЛИЗ.  Файл: anova_bs.std
Источник  Сум.квадр  Ст.своб Ср.квадр
Групповая    486.7        8
```

```
      Факт.1          2449          4     612.3
      Остаточн.      230.4         32       7.2
      Общая           3166         44
      F(фактор1)=85.04, Значимость=0, степ.своб = 4,32
         Гипотеза 1: <Есть влияние фактора на отклик>
      Матрица ковариаций
            21      11.75       9.25      7.833      7.333
         11.75       28.5      13.75      16.38      13.38
          9.25      13.75       11.5      8.583      8.208
         7.833      16.38      8.583      11.69      10.82
         7.333      13.38      8.208      10.82      16.94
Поправочные коэффициенты: e=0.6845  e'=1
      С учетом поправочниго коэффициента=0.6845
      F(фактор1)=85.04, Значимость=0, степ.своб = 2.738,21.9
   Гипотеза 1: <Есть влияние фактора на отклик>
```

В ы в о д ы: Результаты проведенного анализа выявляют влияние фактора релаксационной тренировки (уровень значимости близок к нулю). Отметим, что если бы мы использовали здесь обычную технику однофакторного анализа (применимую и в случае пяти различных групп пациентов в каждой из пяти недель и поэтому приводящую для нашего случая к завышению остаточной дисперсии), то получили бы значение критерия Фишера, равное 34,15 с 40 степенями свободы, которое хотя и тоже значимо, но в 2 с лишним раза меньше вышеприведенного.

П р и м е р 2

З а д а ч а. Рассмотрим двухфакторный эксперимент с одним межгрупповым фактором, в котором исследуется двигательная активность крыс в ответ на инъекции мидозалама (King, 1986). В первое время после инъекции обычно наблюдается снижение моторики, однако, как и в случае морфина, здесь быстро развивается привыкание. Кинг попытался выяснить, можно ли объяснить степень этого привыкания только внешними условиями. В предэкспериментальной фазе двум тестовых группам (из восьми животных каждая) вводились инъекции препарата в течение нескольких дней, а контрольной группе из восьми животных вводился нейтральный физиологический раствор. В заключительный (тестовый) день всем трем группам были сделаны инъекции мидозалама, причем одна из тестовых групп находилась в своем обычном помещении (среде), а другая тестовая группа была перенесена в новую окружающую среду. Если априорная гипотеза Кинга верна, то реакции контрольной группы должны быть близки к реакциям группы, сменившей среду обитания, поскольку в обоих этих группах исключался фактор условного привыкания, определенный неизменностью среды обитания. Поскольку процесс метаболизма лекарства составляет около одного часа, измерения двигательной активности проводились в шести последовательных 5-минутных интервалах, и собранные данные приведены в табл. 8.4.2 (файл ANOVA_B2, где переменные (столбцы) соответствуют временным интервалам для трех последовательных групп, а строки — животным).

Очевидно, что межгрупповой эффект суммируется из различия между группами и различия между объектами в каждой группе, в то время как внутригрупповой (временной) эффект имеет три компонента: главный эффект повторно-временных измерений и его два взаимодействия с групповыми различиями и с различиями объектов.

Таблица 8.4.2. Двигательная активность крыс в ответ на инъекцию мидозалама для шести временных интервалов и трех экспериментальных групп

	Временные интервалы					
Группы	1	2	3	4	5	6
1. Контроль	150	44	71	59	132	74
	335	270	156	160	118	230
	149	52	91	115	43	154
	159	31	127	212	71	224
	159	0	35	75	71	34
	292	125	184	246	225	170
	297	187	66	96	209	74
	170	37	42	66	114	81
2. Та же среда	346	175	177	192	239	140
	426	329	236	76	102	232
	359	238	183	123	183	30
	272	60	82	85	101	98
	200	271	263	216	241	227
	366	291	263	144	220	180
	371	364	270	308	219	267
	497	402	294	216	284	255
3. Новая среда	282	186	225	134	189	169
	317	31	85	120	131	205
	362	104	144	114	115	127
	338	132	91	77	108	169
	263	94	141	142	120	195
	138	38	16	95	39	55
	329	62	62	6	93	67
	292	139	104	184	193	122

В меню (рис. 8.8) выбираем: 1-меж- и 1-внутригрупповой анализ, устанавливая в этом меню шесть градаций первого фактора.

Р е з у л ь т а т ы:

```
ГРУППОВОЙ 1+1-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.
Файл: anova_b2.std

Источник  Сум.квадр  Ст.своб Ср.квадр
Групповая  6.705E5       23
Факт.1     3.997E5        5  7.995E4
Межфакт.   8.082E4       10     8082
Остат.1    2.812E5      105     2678
Факт.2     2.858E5        2  1.429E5
Остат.2    3.847E5       21  1.832E4
Общая      1.432E6      143

F(фактор1)=29.85, Значимость=0, степ.своб = 5,70
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор2)=7.801, Значимость=0.0055, степ.своб = 2,14
   Гипотеза 1: <Есть влияние фактора на отклик>
F(межфатор.)=3.018, Значимость=0.0034, степ.своб = 10,70
   Гипотеза 1: <Есть влияние фактора на отклик>
```

```
Матрица ковариаций
     6388       4696       2240      681.6       2018       1924
     4696       7864       4181       2462       2892       3532
     2240       4181       3912       2697       2162       3298
    681.6       2462       2697       4601       2249       3085
     2018       2892       2162       2249       3717      989.3
     1924       3532       3298       3085      989.3       5228

Поправочные коэффициенты: e=0.6569  e'=0.8674

  С учетом поправочниго коэффициента=0.6569
F(фактор1)=29.85, Значимость=0, степ.своб = 3.285,45.99
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор2)=7.801, Значимость=0.0055, степ.своб = 2,14
   Гипотеза 1: <Есть влияние фактора на отклик>
F(межфатор.)=3.018, Значимость=0.0121, степ.своб=6.569,45.99
Гипотеза 1: <Есть влияние фактора на отклик>
```

В ы в о д ы: Результаты проведенного анализа выявляют существенное влияние как внутригруппового (временного) фактора, так и межгрупповые различия от внешних условий, а также и их взаимодействие (уровни значимости 0, 0.0055 и 0.0034 меньше 0.05). На графике факторных эффектов (рис. 8.9) видно, что активность животных существенно падает в первом 5-минутном интервале, причем ее уровень в неизменной среде превышает активность других двух групп, но падение активности продолжается в следующих интервалах, тогда как в других группах она немного повышается.

Последующий анализ может идти по пути исследования *простых эффектов* посредством сведения двухфакторной схемы к различным вариантам однофакторной, например, для исследования различий между группами в первом и последнем временном интервале (классический однофакторный анализ с повторными измерениями), или для исследования влияния фактора времени в каждой из трех групп по отдельности (однофакторный групповой анализ).

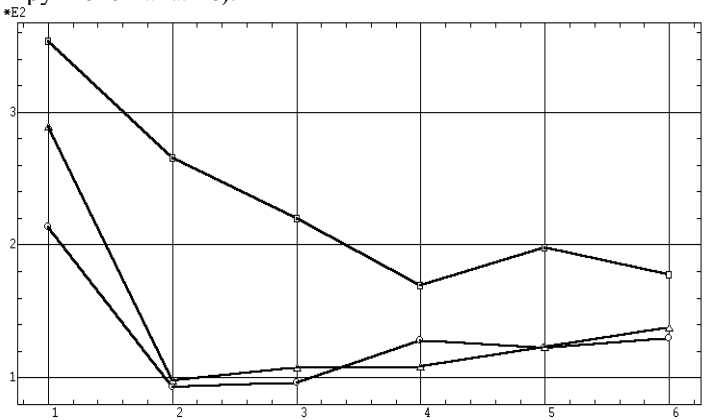


Рис. 8.9. График факторных эффектов. По горизонтальной оси — уровни первого фактора, порядок графиков (сверху вниз): та же среда, контроль, новая среда

З а д а ч а. Модифицируем теперь условно схему данного эксперимента, считая что во всех группах были одни и те же крысы (т. е. как бы имея план с двумя внутригрупповыми факторами).

Р е з у л ь т а т ы:

```
ГРУППОВОЙ 2-ФАКТОРНЫЙ АНАЛИЗ.  Файл: anova_b2.std
Источник  Сум.квадр  Ст.своб Ср.квадр
Групповая  6.378E4         7
Факт.1     3.997E5         5   7.995E4
Остат.1    1.018E5        35    1.72E4
Факт.2     2.858E5         2   1.429E5
Остат.2    3.209E5        14   2.292E4
Межфакт12  8.082E4        10      8082
Остат.12   6.021E5        70       8602
Общая      1.432E6       143
F(фактор1)=27.49, Значимость=0, степ.своб = 5,35
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор2)=6.234, Значимость=0.0115, степ.своб = 2,14
   Гипотеза 0: <Есть влияние фактора на отклик>
F(межфатор.12)=0.9395, Значимость=0.5036, степ.своб = 10,70
   Гипотеза 0: <Нет влияния фактора на отклик>
Матрица ковариаций
     6388      4696      2240     681.6      2018      1924
     4696      7864      4181      2462      2892      3532
     2240      4181      3912      2697      2162      3298
    681.6      2462      2697      4601      2249      3085
     2018      2892      2162      2249      3717     989.3
     1924      3532      3298      3085     989.3      5228
Поправочные коэффициенты: e=0.6569  e"=0.8674
  С учетом поправочниго коэффициента=0.6569
F(фактор1)=27.49, Значимость=0, степ.своб = 3.285,45.99
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор2)=6.234, Значимость=0.0115, степ.своб = 2,14
   Гипотеза 1: <Есть влияние фактора на отклик>
F(межфатор.)=0.9395, Значимость=0.5171,степ.своб=6.569,45.99
   Гипотеза 0: <Нет влияния фактора на отклик>
```

В ы в о д ы: Отличие полученных результатов от вышерассмотренной схемы с одним внутригрупповым и одним межгрупповым факторами состоит в резком усилении незначимости межфакторного взаимодействия и в двукратном увеличении уровня значимости нулевой гипотезы для второго фактора.

П р и м е р 3

З а д а ч а. Рассмотрим трехфакторный эксперимент (схема с одним внутригрупповым фактором) по использованию индивидуальных профилактических приемов для предотвращения угрозы заболевания при контактах с потенциально загрязненной средой (табл. 8.4.3, файл ANOVA_B3, где строки отвечают испытуемым, а столбцы — последовательным уровням факторов в последовательности: время, методика, пол). Использовались два вида обучения, которые проходили две группы испытуемых: лекционный курс и индивидуальная тренировка выполнения необходимой профилактики. Каждая группа включала 10 мужчин и 10 жен-

щин, у которых по данным опроса выяснялась частота выполнения профилактики до и после обучения, а также спустя 6 месяцев и 1 год.

Таблица 8.4.3. Профилактика от заболевания

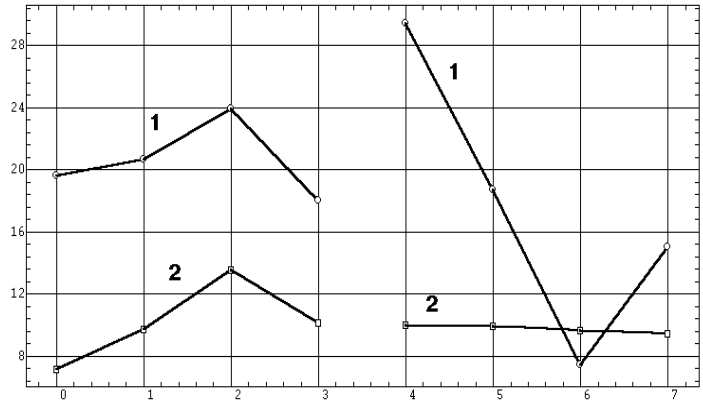
	Тренировка				Демонстрация			
	До	После	6 мес	1 год	До	После	6 мес	1 год
Мужчины	7	22	13	14	0	0	0	0
	25	10	17	24	69	56	14	36
	50	36	49	23	5	0	0	5
	16	38	34	24	4	24	0	0
	33	25	24	25	35	8	0	0
	10	7	23	26	7	0	9	37
	13	33	27	24	51	53	8	26
	22	20	21	11	25	0	0	15
	4	0	12	0	59	45	11	16
	17	16	20	10	40	2	33	16
Женщины	0	6	22	26	15	28	26	15
	0	16	12	15	0	0	0	0
	0	8	0	0	6	0	23	0
	15	14	22	8	0	0	0	0
	27	18	24	37	25	28	0	16
	0	0	0	0	36	22	14	48
	4	27	21	3	19	22	29	2
	26	9	9	12	0	0	5	14
	0	0	14	1	0	0	0	0
	0	0	12	0	0	0	0	0

Таким образом, данная схема отвечает трехфакторному плану с одним внутригрупповым фактором (время) и двумя межгрупповыми факторами (методика обучения и пол). В рис. 8.8 выбираем: «1-внутригрупповой» и «2-межгрупповой» анализ, устанавливая в этом меню четыре градации первого фактора и две градации второго фактора

Р е з у л ь т а т ы:

```
ГРУППОВОЙ 1+2-ФАКТОРНЫЙ АНАЛИЗ.  Файл: anova_b3.std
Источник  Сум.квадр  Ст.своб Ср.квадр
Групповая  2.149E4       39
Факт.1       274.1        3     91.36
Межфакт12     1378        3     459.3
Межфакт13    779.9        3       260
Межфак123    476.4        3     158.8
Остат.1    1.101E4      108     101.9
Факт.2       107.3        1     107.3
Межфакт23    63.76        1     63.76
Факт.3        3358        1      3358
Остат.2    1.796E4       36     498.9
Общая       3.54E4      159
F(фактор1)=0.8965, Значимость=0.5521, степ.своб = 3,108
   Гипотеза 0: <Нет влияния фактора на отклик>
F(фактор2)=0.215, Значимость=0.8254, степ.своб = 36
   Гипотеза 0: <Нет влияния фактора на отклик>
F(фактор3)=6.731, Значимость=0, степ.своб = 36
   Гипотеза 1: <Есть влияние фактора на отклик>
F(межфатор.12)=4.507, Значимость=0.0054, степ.своб = 3,108
   Гипотеза 1: <Есть влияние фактора на отклик>
F(межфатор.13)=2.551, Значимость=0,0583, степ.своб = 3,108
   Гипотеза 0: <Нет влияния фактора на отклик>
```

```
F(межфатор.23)=0.1278, Значимость=0.8944, степ.своб = 36
   Гипотеза 0: <Нет влияния фактора на отклик>
F(межфатор.123)=1.558, Значимость=0.2023, степ.своб = 3,12
   Гипотеза 0: <Нет влияния фактора на отклик>
  Матрица ковариаций
    335.7       206     74.55     131.6
      206       251     75.25       100
    74.55     75.25       146     64.17
    131.6       100     64.17     168.1
Поправочные коэффициенты: e=0.8649  e"=1
```



а *б*

Рис. 8.10. График факторных эффектов (по горизонтальной оси — уровни первого фактора): для тренированной группы (*а*); для *лекционной* группы (*б*); 1 — мужчины; 2 — женщины

В ы в о д ы: Как показывают полученные результаты, значимым является только фактор пола (уровень значимости близок к нулю), а также взаимодействие факторов *время–методика* (уровень значимости 0.0054). Изучение графика факторных эффектов (рис. 8.10) показывает, что женщины пользуются профилактическими приемами реже, чем мужчины, однако это может свидетельствовать и о том, что первые просто реже контактируют с потенциально загрязненной средой. Этот же график уточняет тенденцию обнаруженного межфакторного взаимодействия: обе тренируемые группы усиливают профилактические действия с течением времени, тогда как “лекционные” группы не изменяют или даже снижают свой первоначальный уровень. Поскольку поправочные коэффициенты ε и ε' велики по своим значениям, повторная проверка гипотез со скорректированными степенями свободы не проводится.

Дальнейший анализ может идти по пути детального исследования *простых эффектов* путем сведения трехфакторной схемы к различным комбинациям двухфакторной: пол и время в тренируемых (“лекционных”) группах, пол и методика при различных временных срезах, методика и время для каждого пола и т. п.

Пример 4

Задача. Рассмотрим трехфакторный эксперимент (схема с двумя внутригрупповыми факторами) по выработке у крыс условного рефлекса на болевой раздражитель (табл. 8.4.4, файл ANOVA_B4, где строки отвечают крысам, а столбцы — последовательным уровням факторов в последовательности: фаза, цикл, группа). Нормально крысы научены нажимать на педальки в своем вольере для получения пищи, чем они и занимаются значительную часть времени. Далее одной группе из 8 крыс (A) в фазе I периодически подают звуковой сигнал (предупредительный стимул), после чего на педальки поступает слабомощный электрический заряд для болевого воздействия, что влечет снижения активности крыс по нажатию на педальку в результате научения.

Таблица 8.4.4. Выработка условного рефлекса у крыс на предупредительный стимул болевого раздражителя

	Циклы							
	1		2		3		4	
	Фаза		Фаза		Фаза		Фаза	
Группа	I	II	I	II	I	II	I	II
A	1	28	22	48	22	50	14	48
	21	21	16	40	15	39	11	56
	15	17	13	35	22	45	1	43
	30	34	55	54	37	57	57	68
	11	23	12	33	10	50	8	53
	16	11	18	34	11	40	5	40
	7	26	29	40	25	50	14	56
	0	22	23	45	18	38	15	50
B	1	6	16	8	9	14	11	33
	37	59	28	36	34	32	26	37
	18	43	38	50	39	15	29	18
	1	2	9	8	6	5	5	15
	44	25	28	42	47	46	33	35
	15	14	22	32	16	23	32	26
	0	3	7	17	6	9	10	15
	26	15	31	32	28	22	16	15
C	33	43	40	52	39	52	38	48
	4	35	9	42	4	46	23	51
	32	39	38	47	24	44	16	40
	17	34	21	41	27	50	13	40
	44	52	37	48	33	53	33	43
	12	16	9	39	9	59	13	45
	18	42	3	62	45	49	60	57
	13	29	14	44	9	50	15	48

В фазе II эту группу крыс переводят в другой вольер (изменение окружающей среды), где звуковой сигнал не подкрепляется последующим электрическим шоком. И так проводятся четыре последовательных цикла из этих двух фаз. В этой группе в результате выработанного условного рефлекса частота нажатий на педальку в фазе II также должна быть ниже по сравнению с отсутствием звукового стимула, однако она должна повышаться от цикла к циклу в связи с влиянием фактора “безопасной сре-

ды" в фазе II. У второй группы крыс (B) аналогичный рефлекс вырабатывается в фазе I не на звуковой, а на световой стимул, поэтому в новом вольере фазы II они в ответ на звуковой стимул должны менее реагировать на звук, особенно в первых двух циклах. Третья группа крыс (C) отличалась от первой только тем, что обе фазы выполнялись в том же самом вольере, т. е. не присутствовал фактор "безопасной среды". В эксперименте фиксировалось число нажатий на педальки после подачи предупредительного стимула.

Таким образом, данная схема отвечает трехфакторному плану с двумя внутригрупповыми факторами (фаза и цикл) и одним межгрупповым фактором (три различные условно-рефлекторные методики). В меню рис. 8.8 выбираем: «2-внутригрупповой» и «1-межгрупповой анализ», устанавливая две градации первого фактора и четыре градации второго фактора.

Р е з у л ь т а т ы:

```
ГРУППОВОЙ 2+1-ФАКТОРНЫЙ АНАЛИЗ.  Файл: anova_b4.std

Источник  Сум.квадр  Ст.своб Ср.квадр
Групповая  2.034E4       23
Факт.1      1.17E4        1   1.17E4
Межфакт13     4054        2     2027
Остат.1       1893       21    90.12
Межфакт12    741.5        3    247.2
Межфак123     1274        6    212.3
Остат.1-3     3859       63    61.26
Факт.2        2727        3      909
Межфакт23     1047        6    174.5
Остат.2       4761       63    75.58
Факт.3        4617        2     2308
Остат.3    1.572E4       21    748.7
Общая       5.24E4      191

F(фактор1)=129.9, Значимость=0, степ.своб = 21
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор2)=12.03, Значимость=0, степ.своб = 3,63
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор3)=3.083, Значимость=0.0656, степ.своб = 2,21
   Гипотеза 0: <Нет влияния фактора на отклик>
F(межфатор.12)=4.035, Значимость=0.0109, степ.своб = 3,63
   Гипотеза 0: <Нет влияния фактора на отклик>
F(межфатор.13)=22.49, Значимость=0, степ.своб = 2,21
   Гипотеза 1: <Есть влияние фактора на отклик>
F(межфатор.23)=2.309, Значимость=0.044, степ.своб = 6,63
   Гипотеза 0: <Нет влияния фактора на отклик>
F(межфатор.123)=3.466, Значимость=0.0053, степ.своб = 6,63
   Гипотеза 1: <Есть влияние фактора на отклик>
```

В ы в о д ы: Как показывают полученные результаты, значимыми являются первые два фактора (их уровни значимости близки к нулю), т. е. имеется сильное влияние как фазы (подкрепляемый и неподкрепляемый предупредительный стимул), так и цикла (т. е. временного дообучения тому, что фаза II безопасна). С другой стороны, неожиданно оказывается малозначимым третий фактор группы (уровень значимости 0.0656 выше уровня 0.05), т. е. фактор дополнительной смены внешних условий между

фазами или смены модальности предупредительного стимула. Вместе с тем оказываются значимыми все три парных межфакторных взаимодействия (0.0109, 0, 0.044), так же как и трехфакторное взаимодействие (0.0053).

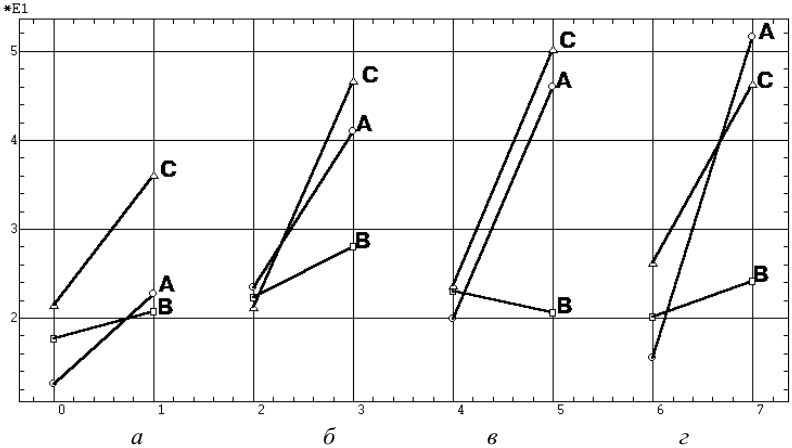


Рис. 8.11. График факторных эффектов с первого по четвертый циклы (*а, б, в, г*); по горизонтальной оси — уровни первого фактора; на графиках по вертикали указаны обозначения групп A, B, C

Кроме того, рисунок факторных эффектов (рис. 8.11) показывает, что первая группа демонстрирует значительно меньшее обучение “безопасному предупреждающему стимулу” в фазе II, т. е. смена внешних условий оказывает дополнительное пролонгирующее шоковое ожидание. Поэтому, если исключить из анализа одну из групп B или C, то и эффект третьего фактора окажется значимым. Аналогичным образом, исключая ту или иную фазу или выбирая отдельные циклы (т. е. сводя трехфакторную схему к двухфакторной), можно изучать более тонкие взаимодействия между исследуемыми факторами и группами.

Пример 5

З а д а ч а. Рассмотрим схему трехфакторно-внутригруппового эксперимента (табл. 8.4.5, файл ANOVA_B5, где строки отвечают водителям, а столбцы — последовательным уровням факторов в последовательности: модель, время суток, дорога): имеются три автомобиля различных моделей: дешевая, средняя, дорогая (М, С, Д), три типа дорожного покрытия (грунтовая, щебенка и шоссе) и два времени суток (день и ночь). Три водителя управляют каждым из трех автомобилей во всех комбинациях условий и у них фиксируется число ошибок вождения.

Таблица 8.4.5. Число ошибок вождения у трех водителей в зависимости от модели автомобиля, времени суток и дорожного покрытия

	Ночь	День

Дорога:	М	С	Д	М	С	Д
Грунтовая	10	8	6	5	4	3
	9	8	5	4	3	3
	8	7	4	4	1	2
Щебенка	9	7	5	4	3	3
	10	6	4	4	2	2
	7	4	3	3	3	2
Шоссе	7	6	3	2	2	1
	4	5	2	2	3	2
	3	4	2	1	0	1

В меню рис. 8.8 выбираем: «3-внутригрупповой» анализ, устанавливая в этом меню три градации первого фактора и две градации второго фактора.

Р е з у л ь т а т ы:

```
ГРУППОВОЙ 3-ФАКТОРНЫЙ АНАЛИЗ.  Файл: anova_b5.std

Источник  Сум.квадр  Ст.своб Ср.квадр
Групповая    24.11        2
Факт.1       51.44        2    25.72
Остат.1      1.111        4   0.2222
Факт.2       140.2        1    140.2
Остат.2      2.333        2    1.167
Факт.3       56.78        2    28.39
Остат.3     0.1111        4  0.02778
Межфакт12    16.78        2    8.389
Остат.12    0.8889        4   0.2222
Межфакт13    8.778        4    2.194
Остат.13     4.667        8   0.5833
Межфакт23    5.444        2    2.722
Остат.23     5.222        4    1.306
Межфак123    2.778        4   0.6944
Остат.123    2.889        8   0.3611
Общая        323.5       53

F(фактор1)=92.6, Значимость=0, степ.своб = 2,4
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор2)=120.1, Значимость=0.0002, степ.своб = 2
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор3)=1022, Значимость=0.0002, степ.своб = 2,4
   Гипотеза 1: <Есть влияние фактора на отклик>
F(межфатор.12)=37.75, Значимость=0.0041, степ.своб = 2,4
   Гипотеза 1: <Есть влияние фактора на отклик>
F(межфатор.13)=3.762, Значимость=0,0524, степ.своб = 4,8
   Гипотеза 0: <Нет влияния фактора на отклик>
F(межфатор.23)=2.085, Значимость=0.2396, степ.своб = 2,4
   Гипотеза 0: <Нет влияния фактора на отклик>
F(межфатор.123)=1.923, Значимость=0.1997, степ.своб = 4,8
   Гипотеза 0: <Нет влияния фактора на отклик>
```

В ы в о д ы: Как показывают полученные результаты, на число ошибок вождения оказывают влияние все три исследуемых фактора (их уровни значимости близки к нулю), а также наблюдается значимое взаимодействие факторов модели автомобиля и времени суток (уровень значимости 0.0041 существенно ниже уровня 0.05).

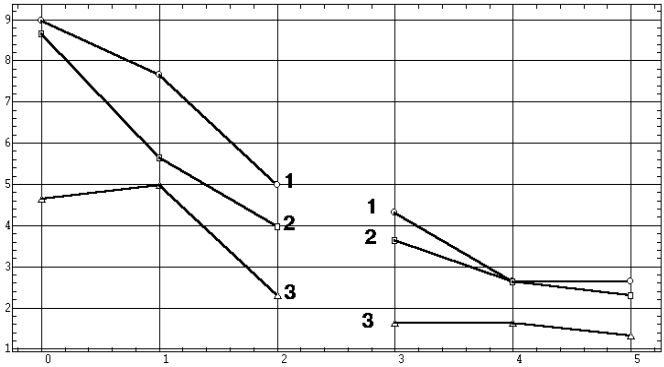


Рис. 8.12. График факторных эффектов (по горизонтальной оси — уровни первого фактора) для ночи и для дня (*а, б)*; 1 —; 2 — средние; 3 — дорогие

Как показывает график факторных эффектов (рис. 8.12), число ошибок вождения существенно падает при улучшении дорожного покрытия, при переходе к более дорогому автомобилю и в дневное время суток по сравнению с ночным.

8.5. Многофакторный дисперсионный анализ

Назначение. В этом разделе представлена универсальная процедура дисперсионного анализа, базирующаяся на методе множественной линейной регрессии. Такой подход удобен для анализа данных, отнесение которых к той или иной модели затруднительно. Кроме того, он позволяет обрабатывать данные и многофакторных (m>2) экспериментов. Процедура не выделяет эффектов межфакторного взаимодействия. Однако она позволяет выявлять факторные эффекты даже в том случае, когда произведены измерения не при всех сочетаниях значений факторов, т. е. в случае неполного факторного планирования.

дом двухфакторного

Исходные данные представляют собой матрицу размером (m+1)n, n — число измерений, в которой в качестве первых m переменных содержатся значения градаций m факторов, а m+1–я переменная содержит значения отклика, измеренного при указанных градациях фактора. Каждый фактор должен иметь не менее двух градаций, значения которых нумеруются целыми числами, начиная с 1. Для каждого фактора должны быть произведены измерения по крайней мере при двух его уровнях, при этом допускаются повторные измерения при каждом сочетании значений факторов. Общее число измерений должно быть больше числа факторов.

Результаты. На экран выдается стандартная таблица дисперсионного анализа и результаты проверки каждой гипотезы (см. в разд. 8.2, 8.3).

Пример

З а д а ч а. Исследовалось влияние трех агротехнических технологий выращивания картофеля и двух методов предпосевной обработки семян на урожайность. Для каждого сочетания значений этих двух факторов было выделено разное число участков. Полученные данные об урожайности картофеля представлены в табл. 8.5.1.

Таблица 8.5.1. Урожайность картофеля [ц/га] в зависимости от предпосевной обработки и агротехнических методов выращивания

	Фактор F_1 – агротехнический метод		
Фактор F_2 – предпосевная обработка	1	2	3
1	17.5 16	13.2	12.8 10.4 9.9
2	10.1 8.6 11.3	5.4 3.7	10.3

Такого типа данные в электронной таблице для анализа представляются в виде матрицы, содержащей три переменные: значения первого и второго факторов и значение измеренного отклика (файл MAV):

F_1	F_2	Y
1	1	17.5
1	1	16.2
2	1	13.2
3	1	12.8
3	1	10.4
3	1	9.9
1	2	10.1
1	2	8.6
1	2	11.3
2	2	5.4
2	2	3.7
3	2	10.3

Р е з у л ь т а т ы:

```
МНОГОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: mav.std

Источник Сум.квадр Ст.своб Ср.квадр Сила влияния
Фактор 1    92.15     2      46.07      0.1648
F(фактор 1)=6.073, Значимость=0.0247, степ.своб = 2,8
   Гипотеза 1: <Есть влияние фактора на отклик>
Фактор 2    117.3     1      117.3      0.663
F(фактор 2)=17.66, Значимость=0, степ.своб = 8
   Гипотеза 1: <Есть влияние фактора на отклик>
  Остат.      36.59          8       4.574
```

В ы в о д ы: Результаты анализа показывают значимое влияние на урожайность как предпосевной обработки семян, так и агротехнической технологии.

З а д а ч а 2: Эти же данные представляется возможным обработать методом двухфакторного анализа с повторными измерениями (см. разд. 8.3, в соответствии с требованиями метода данные представлены в файле MAV1) и сравнить результаты.

Р е з у л ь т а т ы:

```
2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл: mav1.std
          Факторный план: повторяемый, с фиксированными эффектами

Источник Сум.квадр Ст.своб Ср.квадр     F  Значимость Сила влияния
  Факт.1     39.8       2     19.9     10.8    0.0109     0.429
  Факт.2     84.1       1     84.1     45.7   1.67E-5     0.483
  Межфак     33.1       2     16.6        9    0.0161      0.4
  Остат.       11       6     1.84
  Общая       168      11     15.3
```

```
F(фактор1)=10.8, Значимость=0.0109, степ.своб = 2,6
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор2)=45.7, Значимость=1.67E-5, степ.своб = 6
   Гипотеза 1: <Есть влияние фактора на отклик>
F(межфакт)=9, Значимость=0.0161, степ.своб = 2,6
   Гипотеза 1: <Есть влияние фактора на отклик>
```

Выводы: При сравнении этих результатов с предыдущими можно заметить некоторые расхождения в суммах квадратов и степенях свободы, поскольку многофакторный метод не учитывает межфакторных взаимодействий. Тем не менее результирующие значимости нулевых гипотез находятся в очень хорошем согласии.

8.6. Ковариационный анализ

Назначение. Основной задачей ковариационного анализа (так же как и обычного однофакторного дисперсионного анализа) является проверка влияния качественного или количественного фактора на отклик (см. разд. 8.2). Однако здесь при каждом измерении вместе с значением отклика регистрируются значения одной или нескольких сопутствующих переменных (количественных со–факторов), которые также могут оказывать влияние на отклик, но это влияние желательно исключить при проверке основного факторного эффекта, т. е. требуется рафинировать основной эффект от влияния сопутствующих переменных.

Исходные данные представляют собой матрицу из m переменных по n измерений: первые m–2 переменных являются сопутствующими переменными; m–1–я переменная является откликом в смысле дисперсионного анализа. Каждое значение m–й переменной представляет номер уровня фактора (целое число), при котором было произведено данное измерение. Матрица должна быть упорядочена по возрастанию значений m–й переменной. Чтобы привести матрицу к упорядоченному виду можно воспользоваться преобразованием сортировки (см. разд. 3.4).

Результаты. Сначала производится проверка гипотез о равенстве средних значений сопутствующих переменных и отклика (гомогенность) для групп, соответствующих различным уровням фактора. Нулевая гипотеза гомогенности сопутствующих переменных говорит о сбалансированности проведенного эксперимента, когда их значения для различных групп примерно одинаково распределены. Принятие нулевой гипотезы гомогенности переменной отклика может быть следствием отсутствия факторного эффекта.

Далее производится множественный линейный регрессионный анализ первых m–1 переменных в соответствии с описанием разд. 9.4.

В завершение процедуры производится однофакторный дисперсионный анализ m–ой переменной, значения которой скорректированы вычитанием вычисленных регрессионных значений (выдача результатов аналогична разд. 8.2).

Пример

З а д а ч а. В эксперименте изучалось влияние тренировки на способность человека близко подойти к устрашающему объекту (живой змее), прежде чем он почувствует дискомфорт или беспокойство. Набрали 4 группы по 10 добровольцев (всего 40 испытуемых), с которыми провели и различные по используемому манекену змеи предварительные тренировки. Эти четыре группы соответствовали четырем уровням *фактора тренировки F*, влияние которого следовало выяснить в результате анализа.

Далее провели реальные испытания, фиксируя минимальное расстояние приближения к живой змее (переменная *Y*). Однако это расстояние может зависеть от множества других сопутствующих факторов (смелость, возраст, острота зрения и пр.), поэтому при неудачном разбиении испытуемых на группы статистический результат может быть сильно искажен. В данном случае решили учитывать один сопутствующий фактор — природную смелость, оцениваемую аналогичной пробой до начала тренировки (переменная *X*). Данные исследования представлены в табл. 8.6.1, а для анализа эти данные представлены (файл COV) значениями трех парных переменных: *X* — все значения до тренировки; *Y* — все значения после тренировки; *F* — соответствующие значения фактора.

Таблица 8.6.1. Расстояние приближения к живой змее (в дюймах)

F-тип манекена	1	2	3	4	1	2	3	4
Испытуемые	X-до тренировки				Y-после тренировки			
1	25	17	32	10	25	11	24	8
2	13	9	30	29	25	9	18	17
3	10	19	12	7	12	16	2	8
4	25	25	30	17	30	17	24	12
5	10	6	10	8	37	1	2	7
6	17	23	8	30	25	12	0	26
7	9	7	5	5	31	4	0	8
8	18	5	11	29	26	3	1	29
9	27	30	5	5	28	26	1	29
10	17	19	25	13	29	20	10	9

Р е з у л ь т а т ы:

```
КОВАРИАЦИОННЫЙ АНАЛИЗ.  Файл: cov.std

Гомогенность X1: F=0.07494, Значимость=0.9723, степ.своб = 3,36
   Гипотеза 0: <Нет влияния фактора на отклик>
Гомогенность X2: F=8.85, Значимость=0.0003, степ.своб = 3,36
   Гипотеза 1: <Есть влияние фактора на отклик>

  Коэфф.         a0          a1
Значение  4.262E-12    0.6427
Ст.ошиб.      0.9321     0.1047
 Значим.      0.9955          0

Источник  Сум.квадр. Степ.св Средн.квадр.
Регресс.      1310        1         1310
Остаточн      1321       38        34.75
     Вся      2630       39

Множеств R      R^2   R^2прив  Ст.ошиб.     F    Значим
    0.7056   0.4979   0.4847    5.895    37.68       0
   Гипотеза 1: <Есть влияние фактора на отклик>

Источник   Сум.квадр  Ст.своб  Ср.квадр  Сила влияния
  Факт.1       1940        3       646.6       0.3666
  Остат.       1321       35       37.73
  Общая        3260       38        85.8
F(фактор1)=17.63, Значимость=0, степ.своб = 3,36
   Гипотеза 1: <Есть влияние фактора на отклик>
```

В ы в о д ы: Предварительные результаты показывают, что значения сопутствующей переменной достаточно однородно распределены по уровням фактора, а значения переменной–отклика существенно различаются в этих же группах. Регрессионная модель также достаточно хорошо воспроизводит зависимость отклика от сопутствующей переменной. Заключительный дисперсионный анализ позволяет отклонить гипотезу об отсутствии влияния фактора на уровне значимости, близком к нулю, и принять гипотезу о присутствии фактора тренировки.

Глава 9

АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

«Вслед за дождем пошел снег,
а за ним — два дурака»
[Р.Шекли. Обмен разумов]

Временной ряд — это совокупность последовательных измерений некоторой переменной (*процесса*), произведенных через одинаковые интервалы значений некоторого параметра (чаще всего — времени или пространственной координаты).

9.1. Анализ и прогнозирование тренда

Назначение. Анализ тренда предназначен для исследования закона изменения или дрейфа локального среднего значения временного ряда с построением математической модели тренда и с прогнозированием на этой основе будущего поведения временного ряда. Анализ тренда производится методами простой или общей регрессии (см. разд. 10.3, 10.6). При их исполнении необходимо выбрать из электронной таблицы одну переменную, представляющую анализируемый временной ряд или две переменные, вторая из которых представляет значения временного параметра.

Результаты. При анализа тренда можно получить следующие результаты:

- опробовать несколько математических моделей тренда и выбрать ту, которая с большей точностью описывает динамику изменения временного ряда;
- построить прогноз будущего поведения временного ряда на основании выбранной модели тренда с $100(1-\alpha)\%$-ным доверительным интервалом;
- удалить тренд из временного ряда с целью обеспечения его стационарности, необходимой для корреляционного и спектрального анализа (для этого после расчета регрессионной модели следует выполнить анализ остатков и сохранить остатки в матрице данных).

Пример прогнозирования тренда для временного ряда изменения среднестатистической урожайности зерновых приведен разд. 10.3.

Фурье-модели (разд. 9.6) позволяют учитывать тренд при прогнозировании временного ряда.

9.2. Корреляционный анализ

Назначение. Корреляционный анализ является средством выявления доминирующих корреляций и их *лагов* (задержек) и периодичностей в одном процессе *X* (*автокорреляция*) или между двумя процессами *X, Y* (*кросскорреляция*). Высокие корреляции могут служить индикатором причинно–следственных связей или взаимодействий внутри одного процесса или между двумя процессами, а величина лага указывает временную задержку в передаче взаимодействия.

;

Действия и результаты. После запуска процедуры следует выбрать (бланк рис. 9.1, см. также пояснения к рис. 2.3) одну или две переменные для автокорреляционного или кросскорреляционного анализа.

В этом же бланке необходимо задать значения следующих трех параметров:

- *размерность* временного шага анализируемого ряда для привязки результатов к реальной временной шкале;
- длина m сдвигаемого фрагмента первого ряда, выраженная в числе включаемых в него измерений (если $m < 4$, или $m > n$, то принимается $m = n/2$, где n — длина ряда);

- *сдвиг* этого фрагмента i_0, т. е. его положение относительно начала ряда;
- признак вычисления номинальной корреляции c_{xy}^N работает в режиме интервальной корреляции.

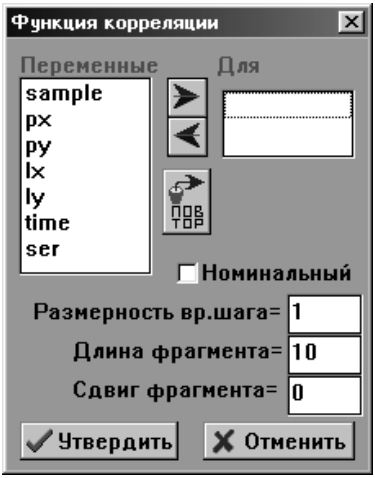


Рис. 9.1. Экранный бланк выбора переменных и установки параметров корреляционного анализа

Если i_0=0 и m=0, то вычисляется классическая корреляционная функция c_{xy}, в противном случае вычисляется интервальная или номинальная корреляционная функция.

Если для анализа выбрана одна переменная, то вычисляются значения *автокорреляционной функции*, которая позволяет определить, в какой степени динамика изменения заданного фрагмента воспроизводится в сдвинутых во времени его же отрезках. Если выбраны две переменные, то вычисляются значения *кросскорреляционной функции*, которая позволяет определить, в какой степени динамика изменения заданного фрагмента первого ряда воспроизводится в сдвинутых во времени фрагментах второго ряда. Если временные ряды имеют разную длину, то выдается предупредительная диагностика и в качестве n принимается длина более короткого ряда.

Выдача результатов включает также критическое значение для нулевой гипотезы «меньшие значения корреляций не отличны от нуля» на уровне значимости α аналогично коэффициенту корреляции Пирсона (разд. 6.3).

В завершение процедуры выдается график авто– или кросскорреляционной функции.

Полученные результаты можно перенести с графика в электронную таблицу для последующего анализа и использования нажатием инструментальной кнопки «*СохрГраф*».

Ограничение. Размер временного ряда должен быть не меньше 8 и не больше l, где l = 16000, 5000, 1000, 100 при объеме матрицы данных в 64000, 20000, 4000 и 400 чисел.

Примеры

Задача. Некая коммерческая фирма уделяла важное внимание работе на рынке фьючерсных контрактов. Поэтому для выработки обос-

нованной стратегии желательно понимание динамики фьючерсного курса и его несомненной связи с текущим курсом доллара.

Для этого используются (табл. 9.2.1) результаты долларовых торгов на ММВБ в период с 3.10.94 по 15.12.94 (переменная dol из файла SPEC) и результаты фьючерсных долларовых торгов с поставкой 15.12.94 (торги производятся каждый день, исключая субботу и воскресенье, переменная fut).

Таблица 9.2.1. Результаты долларовых и фьючерсных торгов на ММВБ в период с 3.10.94 по 15.12.94

доллар	2643, 2668, 2808, 2833, 3081, 3926, 2994, 2988, 2996, 2996
фьючерс	3104, 3228, 3386, 3524, 3684, 3524, 3524, 3585, 3582, 3590
доллар	3005, 3015, 3030, 3036, 3046, 3055, 3075, 3085, 3093, 3099
фьючерс	3626, 3550, 3485, 3484, 3487, 3476, 3478, 3485, 3482, 3458
доллар	3102, 3102, 3102, 3118, 3131, 3143, 3157, 3175, 3187, 3198
фьючерс	3368, 3303, 3281, 3315, 3358, 3403, 3379, 3379, 3370, 3370
доллар	3201, 3228, 3232, 3234, 3249, 3275, 3275, 3292, 3306, 3338
фьючерс	3380, 3376, 3377, 3355, 3350, 3362, 3355, 3350, 3359, 3380
доллар	3368, 3383
фьючерс	3380, 3380

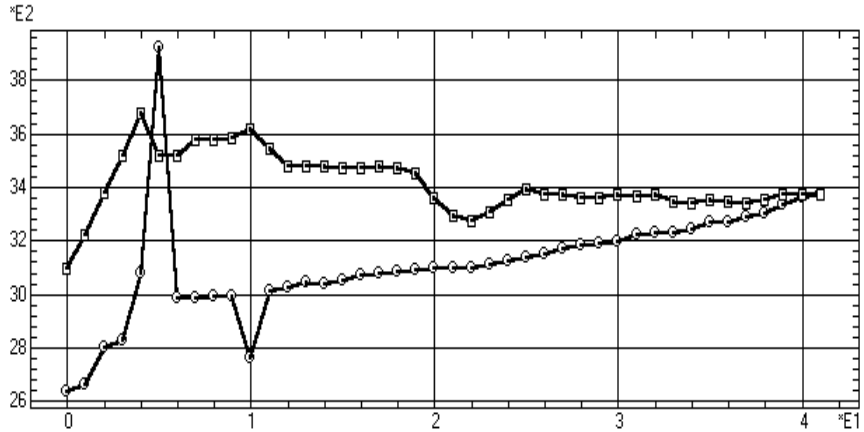


Рис. 9.2. Котировка доллара на ММВБ (кресты) и фьючерсных поставок (квадраты) с 3.10 по 15.12 1994 г.

Визуальный анализ. В курсе доллара (рис. 9.2) выделяется пик знаменитого «*черного вторника*» (11.10.94) и линейная тенденция с небольшими колебаниями, отражающая методику «циркуля и линейки», распространенную тогда в высших планирующих органах. Во фьючерсном курсе имеется вполне понятная тенденция приближения к реальному курсу (при приближении дня поставки 15.12.94) с нерегулярными и достаточно высокоамплитудными колебаниями. Интересно, что курс фьючерсных контрактов рос при начальном повышении курса доллара, но заметно не отреагировал собственно на *черный вторник*. Тем самым, уже

простое изучение графиков временных рядов дает нам достаточно много предварительной информации.

Постановка задачи. Очевидно, что упомянутые процессы развиваются в некоторой информационной среде, обладающей собственной инерционностью и упругостью, что определено существующими организационной и инфраструктурой, менталитетом участников и заинтересованных лиц и кучей других факторов, не поддающихся непосредственному учету. Эти динамические свойства налагают определенные ограничения на передачу взаимодействия от одного процесса к другому, на периодичность колебаний и крутизну фронтов роста и спада.

Варианты анализа. Прямое прогнозирование курса доллара мало что может дать, сверх моделирования общей возрастающей тенденции. Предсказание резких изменений типа *черного вторника* (см. разд. 14.4) возможно только в непосредственной близости от такого рода событий и при наличии сильно развитой интуиции.

Если же попробовать прогнозировать фьючерсные котировки, то даже применение таких изощренных методов как модели Бокса–Дженкинса, вплоть до 12–го авторегрессионного порядка, дает нам лишь общее представление о динамике средней тенденции на очень короткое будущее, а различные конкретные генерации прогнозов существенно отличаются друг от друга. Возможно, что более надежные результаты дало бы моделирование и прогнозирование изменений курса (первая производная) или скорости этих изменений (вторая производная, характеризующая инерционность процессов).

Сначала попробуем построить интервальную автокорреляционную функцию временного ряда *fut*, выбрав в качестве сдвигаемого фрагмента 10 первых измерений, содержащих резкий выброс. Поэтому автокорреляционную функцию такого ряда можно рассматривать как переходную характеристику реакции системы на начальное возмущение.

Р е з у л ь т а т ы:

```
КОРРЕЛЯЦИОННЫЙ АНАЛИЗ.  Файл: spec.std
 Переменные: fut,fut  Критич.значение=0.621
```

Обсуждение. Как следует из рис. 9.3 автокорреляционная функция на первых 18 лагах достаточно монотонно уменьшается в область отрицательных значений, т. е. динамика временного ряда становится все более обратной по сравнению с начальным участком.

Однако на 18—20 лаге наступает резкий, но короткий переход к практически 100%–ной коррелированности с начальным участком. Такая задержка и характер отложенной реакции может служить важным показателем свойств исследуемой среды.

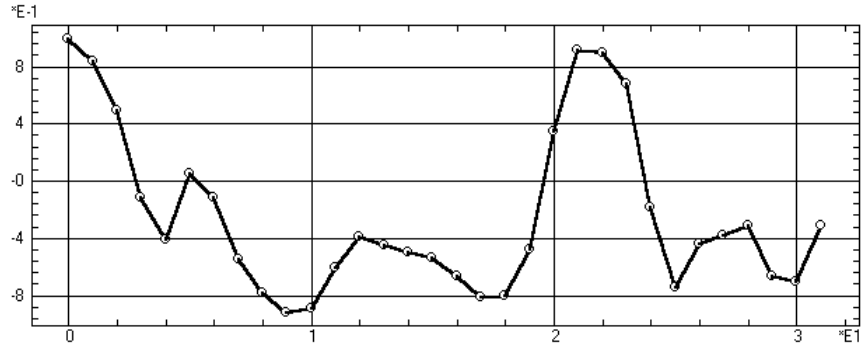


Рис. 9.3. Автокорреляционная функция фьючерсных котировок доллара. По горизонтальной оси — величина лага

Продолжение анализа. Построим теперь интервальную кросскорреляционную функцию, отражающую влияние процесса *fut* на процесс *dol*.

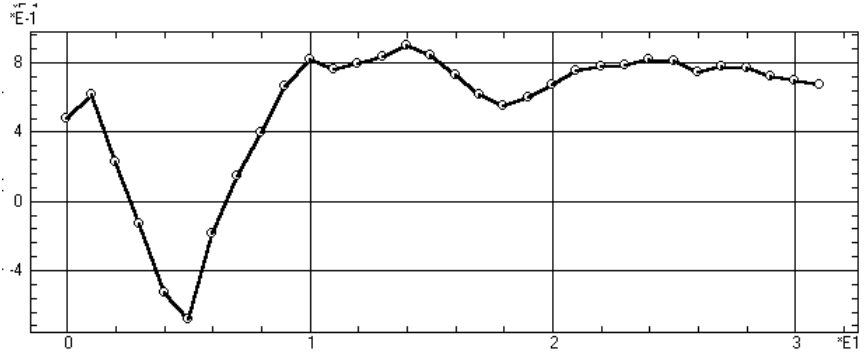


Рис. 9.4. Кросскорреляционная функция фьючерсных котировок доллара на курс доллара

Обсуждение. Полученная кросскорреляционная функция (рис. 9.4) достаточно монотонна, кроме начального отрицательного выброса, что подтверждает вполне очевидное заключение: фьючерсные котировки доллара мало влияют на сам курс доллара.

Продолжение анализа. Построим теперь интервальную кросскорреляционную функцию, отражающую влияние процесса *dol* на процесс *fut*. Для сравнения вычислим также и классическую корреляционную функцию, а для большей наглядности совместим их числовые выдачи и графики.

Результаты:

```
КОРРЕЛЯЦИОННЫЙ АНАЛИЗ.  Файл: spec.std  Переменные: dol, fut
  Сдвиг  Интерв.корр. Класс.корр. Крит.значение
Критич.значение=0.621
        0      0.483    -0.0036         0.304
```

1	0.29	-0.185	0.308
2	0.296	-0.237	0.312
3	0.0362	-0.305	0.316
4	0.00264	-0.289	0.321
5	0.489	-0.159	0.325
6	-0.0451	-0.226	0.33
7	-0.547	-0.3	0.335
8	-0.572	-0.269	0.34
9	-0.527	-0.216	0.345
10	-0.487	-0.144	0.351
11	-0.212	-0.0355	0.357
12	0.000573	0.0329	0.363
13	-0.00285	0.0442	0.369
14	-0.079	0.0304	0.376
15	-0.39	-0.0492	0.383

. . . . . .

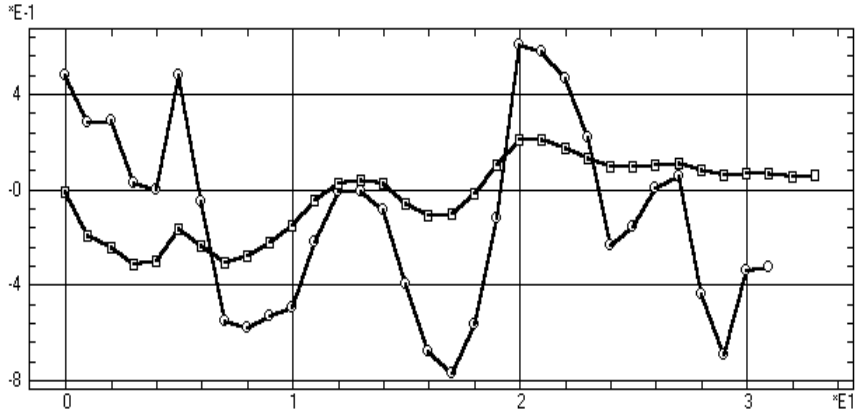


Рис. 9.5. Кросскорреляционная функция курса доллара на его фьючерсные котировки (квадраты — классическая корреляция, круги — интервальная корреляция)

Обсуждение. Интервальная корреляционная функция (рис. 9.5[1]) достоверно воспроизводит отложенную на 18–20 лагов резкую положительную реакцию фьючерсных котировок на начальный скачок курса доллара, отмеченную еще в автокорреляционной функции. Что же касается классической корреляционной функции, то по ее графику нельзя сделать никаких существенных выводов, поскольку все ее значения не отличны от нуля, а колебания незначительны.

Полученные статистические результаты дают нам несравненно более значимые материалы для выработки финансовой стратегии, чем моделирование и прогнозирование, предупреждая о том, что следует обращать

[1] В качестве технической детали отметим, что совместный график корреляций получен следующим образом: вычислены две корреляционные функции с их индивидуальными графиками, данные с каждого графика перенесены в электронную таблицу нажатием инструментальной кнопки «*СохрГраф*» (в результате в электронной таблице появились две пары новых переменных), после чего построен функциональный график этих переменных.

особое внимание на динамику процессов спустя 18–20 дней после очередного возмущения. Эти же материалы дают пищу специалистам для размышлений о физических свойствах среды, передающей взаимодействие процессов с такими большими задержками.

Что дальше? Здесь мы рассмотрели экстремальный случай реакции системы на резкое начальное возмущение. Дальнейший анализ может развиваться в направлении исследования взаимодействия процессов в обычных условиях, не содержащих резких колебаний. Оставляем эту задачу в качестве учебной читателям.

9.3. Спектральный анализ

Назначение. Одним из общепринятых способов анализа структуры стационарных временных рядов является использование *дискретного преобразования Фурье* для оценки спектральной плотности или спектра ряда.

Этот метод может применяться:

- для получения описательных статистик одного временного ряда или же статистик зависимостей между двумя временными рядами;
- для выявления периодических и квазипериодических свойств временных рядов;
- для проверки адекватности моделей, построенных другими методами;
- для сжатого представления данных;
- для интерполяции динамики временных рядов.

Действия. Сначала из электронной таблицы необходимо выбрать (бланк рис. 9.6, см. также рис. 2.3) одну или две переменные для спектрального или кроссспектрального анализа. В этом же бланке необходимо задать значения следующих параметров:

- размерность временного шага анализируемого ряда для привязки результатов к реальной временной и частотной шкалам;
- длина *k* анализируемого отрезка временного ряда, выраженная в числе включаемых в него измерений, при длине=0 анализируется весь временной ряд;
- сдвиг очередного отрезка Δt относительно предыдущего, при $\Delta t=0$ сдвиг равен длине отрезка;
- тип временного окна сглаживания для подавления в спектре эффекта *вытекание мощности*;
- тип усреднения частотных характеристик, вычисленных на последовательных отрезках временного ряда.

Рис. 9.6. Экранный бланк выбора переменных и параметров спектрального анализа

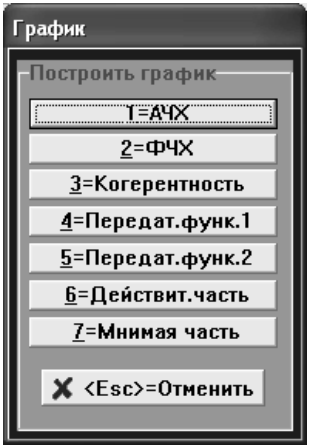


Рис. 9.7. Экранное меню выбора спектрального графика

Результаты. Вычисляются значения *амплитудно–частотной характеристики* $A(i)$ и значения *фазо–частотной характеристики* $\varphi(i)$. В случае кроссспектра вычисляются также значения *передаточных функций* $H_1(i)$, $H_2(i)$, *когерентности* $\gamma(i)$, действительной и мнимой частей АЧХ. Эти результаты могут быть представлены в виде графиков выбором из последующего меню (рис. 9.7), по горизонтальной оси — частота. Запросы графиков повторяются до нажатия кнопки «*Отменить*».

Полученные результаты с графика можно перенести в электронную таблицу для последующего анализа нажатием инструментальной кнопки «*СохрГраф*».

Спектральный анализ может быть проведен повторно с преобразованием исходного временного ряда. Так, если в процессе выявлены сильные сезонные колебания и требуется провести более детальное исследование несезонных закономерностей, то перед повторным анализом следует подавить сезонные изменения трансформацией временного ряда посредством фильтрации, сезонного центрирования, нормирования или дифференцирования (см. разд. 3.4, 9.4).

Спектральный анализ проводится с использованием тригонометрических вычислений, поэтому для него не действуют ограничения быстрого преобразования Фурье по модулю два.

Ограничение. Размер временного ряда не может превосходить l, где l = 16000, 5000, 1000, 100 при объеме матрицы данных в 64000, 20000, 4000 и 400 чисел.

Пример 1

Задача. Необходимо проанализировать временную динамику авиаперевозок. Рассмотрим ежемесячные данные о расстояниях, пройденных самолетами Великобритании за шесть последовательных лет (переменная FLIGHT файл SPEC, в тысячах миль).

Сначала проведем визуальный анализ этого временного ряда (рис 9.8). Как можно заметить, эти данные характеризуются: а) линейным трендом с общей нестационарной тенденцией к возрастанию значений; б) сезонными колебаниями с периодом 12 мес. Для того чтобы нам исследовать различные типы периодичности в рассматриваемом процессе методом спектрального анализа, необходимо предварительно провести нормирование данных в *Блоке преобразований* (разд. 3.4) и обеспечить стационарность процесса удалением линейного тренда посредством выполнения процеду-

ры простой регрессии с записью остатков в матрицу данных. Эти остатки и используются в дальнейшем анализе.

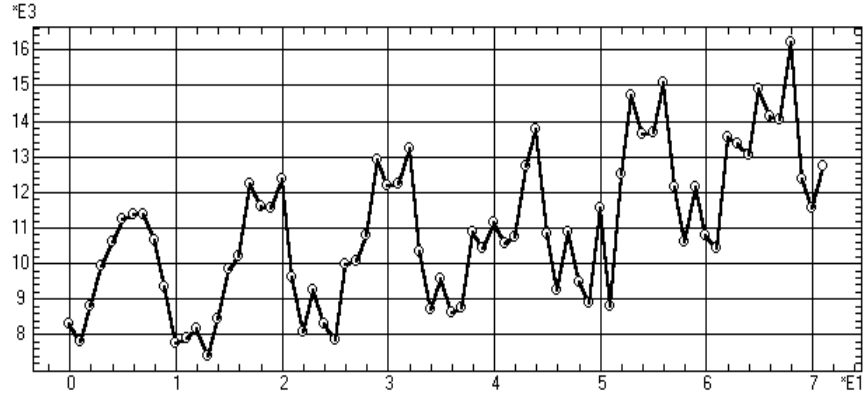


Рис. 9.8. График процесса *Flight* — ежемесячные данные о расстояниях, проходимых самолетами Великобритании

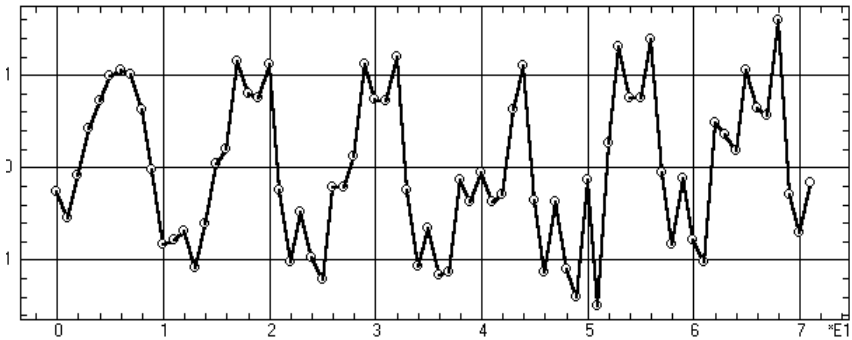


Рис. 9.9. График процесса *Flight* после удаления линейного тренда и нормализации

В результате этих операций мы получаем исходные данные для спектрального анализа, изображенные на графике рис. 9.9.

Проведем теперь спектральный анализ этого процесса.

Р е з у л ь т а т ы:

```
СПЕКТРАЛЬНЫЙ АНАЛИЗ.  Файл: spec1.std  Переменные: flight,flight
```

Обсуждение: В полученном спектре (рис. 9.10) виден доминирующий пик на частоте с периодом 1 год (что можно было бы наблюдать и на графике корреляционной функции). Однако вместе с этим наблюдается и второй, меньший пик для периода 0,3333 (4 месяца). Он достаточно удален от основного пика, поэтому маловероятно его появление вследствие эффекта вытекания мощности.

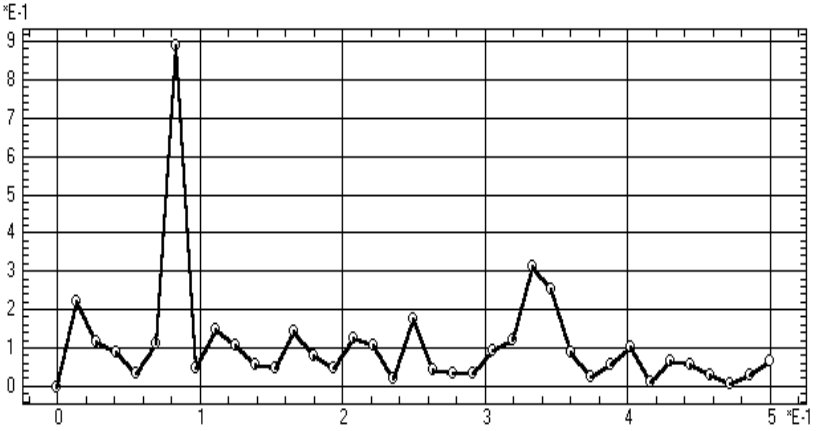


Рис. 9.10. Амплитудно–частотная характеристика процесса *Flight*

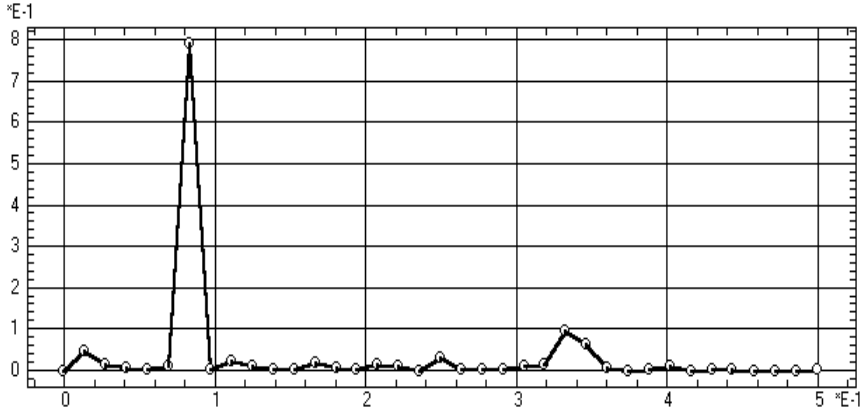


Рис. 9.11. Периодограмма процесса *Flight*

Для сравнения приведем периодограмму (рис 9.11), которая заменяет АЧХ в большинстве западных статистических пакетах (*StatGraphics, SPSS, Statistica*). Как можно легко заметить, этот график не дает никакой информации о любых спектральных составляющих, кроме подавляющего основного пика.

Продолжение анализа. Если мы хотим провести более тщательное исследование процесса, не связанное с сезонностью, то необходимо повторить анализ с предварительным удалением частоты 1/12 сезонным фильтром $1–B^{12}$ и частоты 4/12 одним из фильтров $(1+B+B^2)/3$ или $1–\sqrt{3}\,B+B^2$ с использованием средств разд. 9.3.

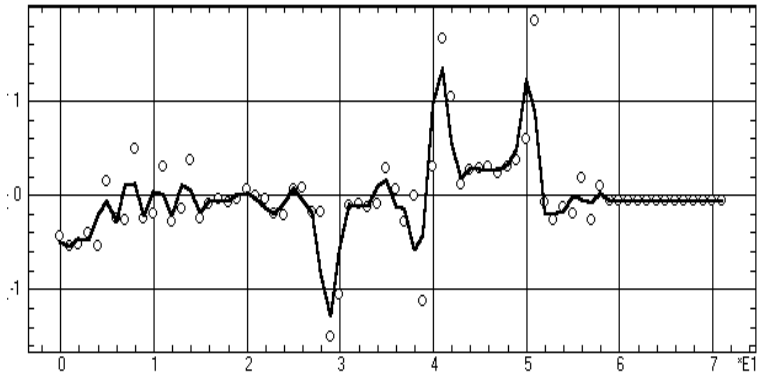


Рис. 9.12. График процесса *Flight* после фильтрации

В результате этих операций график исходных данных приобретает вид, приведенный на рис 9.12. А после выполнения спектрального анализа мы получим график спектра, изображенный на рис. 9.13.

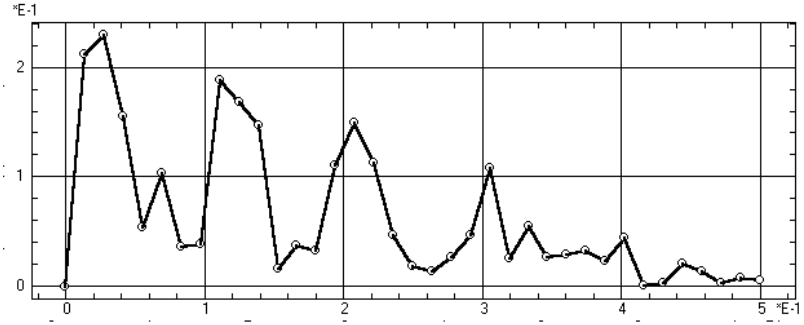


Рис. 9.13. Амплитудный спектр отфильтрованного процесса *Flight*

Обсуждение: Как можно заметить по графику спектра, в исследуемом процессе после проведенных фильтраций осталась, в основном, только случайная компонента и ее спектр сосредоточен преимущественно на низких частотах. Дальнейший интерес может представлять сравнение автокорреляционных функций, соответствующих вышерассмотренным спектрам.

Кроме того, для различных месяцев среднее значение и дисперсия могут существенно различаться. Поэтому можно проверить, что получится, если осуществить предварительную сезонную стандартизацию по месяцам. Продолжение анализа примера см. в следующем разделе.

П р и м е р 2

З а д а ч а. Необходимо исследовать временные взаимосвязи между числом отложенных яиц и числом взрослых насекомых, измеренных с пе-

риодичностью 2 дня (переменные *Eggs* и *Insects* в файле SPEC). Объединенный график этих двух временных рядов приведен на рис. 9.14.

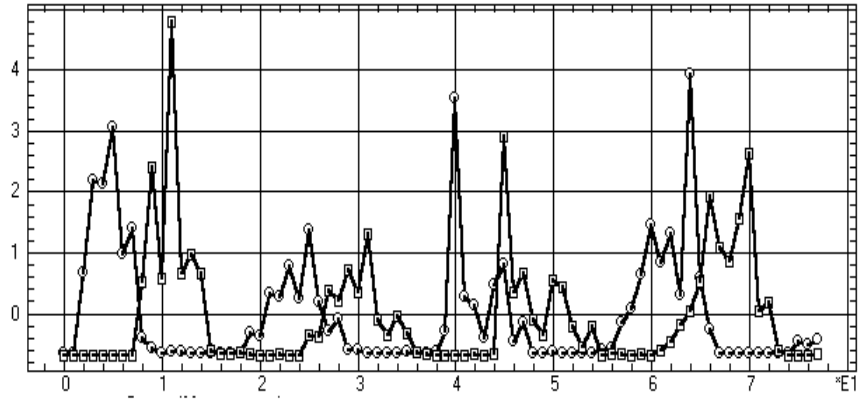


Рис. 9.14. График временных рядов *Eggs* (круги) и *Insects* (квадраты)

Как можно заметить, временные ряды не имеют заметного тренда и поэтому перед спектральным анализом можно провести только стандартизацию каждого ряда.

Р е з у л ь т а т ы:

```
СПЕКТРАЛЬНЫЙ АНАЛИЗ.  Файл: spec1.std Переменные: eggs, insects
```

Обсуждение: Как видно из рис. 9.15, мощность кросс–спектра сосредоточена преимущественно на низких частотах (при дополнительном анализе с усреднением легко убедиться, что на этих частотах и значения когерентности являются устойчиво высокими, что позволяет принять с доверием оценки фазы и передаточной функции).

Выделяется пик с периодом 39 дней, что можно интерпретировать как среднее время развития и жизни насекомого.

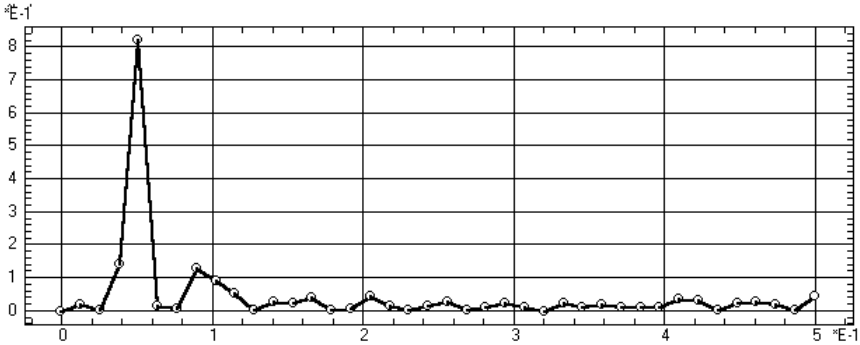


Рис. 9.15. Кросс-спектр временных рядов *Eggs* и *Insects*

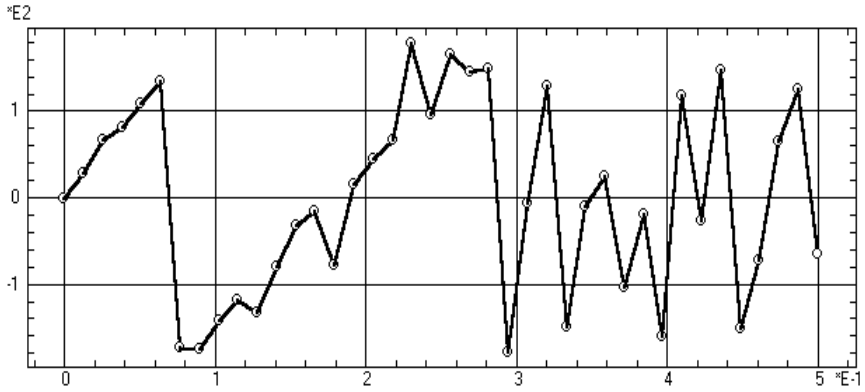


Рис. 9.16. График фазы кросс-спектра для рядов *Eggs* и *Insects*

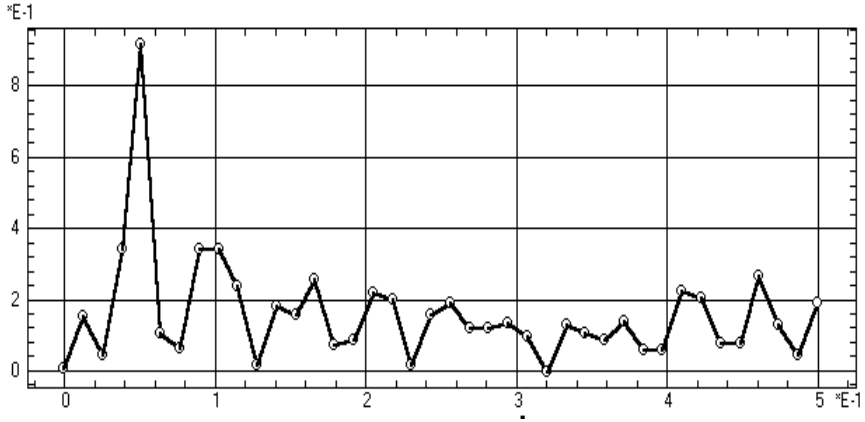


Рис. 9.17. График передаточной функции для рядов *Eggs* и *Insects*

Фаза (рис. 9.16) имеет явный линейный тренд на низких частотах, что иллюстрирует систематическое смещение рядов друг относительно друга (от момента снесения яйца до превращения во взрослое насекомое). В то же время передаточная функция является относительно постоянной в этой области. Поскольку рассматриваемые процессы связаны обратной связью, интерпретация передаточной функции должна выполняться с осторожностью.

Эти результаты позволяют выдвинуть гипотезу о том, что на число взрослых насекомых в каждом следующем поколении влияет только число яиц в предыдущем поколении, а не их распределение во времени.

Для более отчетливого выделения информации о взаимных связях двух процессов можно провести повторный анализ с различными степенями сглаживания исходных процессов и результатов. Полезной при анализе может быть и кросс–корреляционная функция, поскольку наклон фазы в известной степени соответствует лагу (в данном случае равное –6) мак-

симума корреляционной функции. Поэтому можно вычислить спектр еще раз, выполняя выравнивание с параметром, близким к –6.

9.4. Сглаживание и фильтрация

Назначение. Методы *сглаживания и фильтрации* предназначены для преобразования временного ряда y_t с удалением из него высокочастотных, низкочастотных или *сезонных* колебаний. В настоящий раздел включены наиболее распространенные в практике преобразования. Во многом аналогичные цели могут быть достигнуты с использованием *фурье–моделей* (см. разд. 9.6).

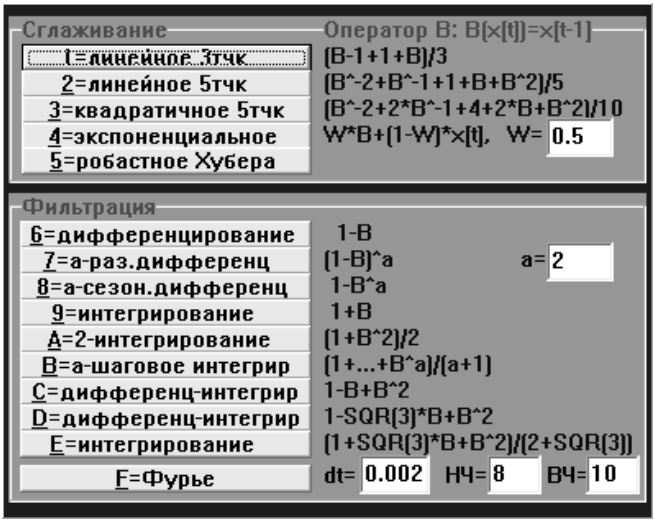


Рис. 9.18. Меню выбора метода сглаживания/фильтрации

Действия. Сначала из электронной таблицы необходимо выбрать (см. бланк рис. 6.4) переменную, представляющую анализируемый временной ряд, после чего выбрать метод сглаживания или фильтрации (меню рис. 9.18).

Рис. 9.19. Меню выбора метода робастного сглаживания

В случае *робастного сглаживания* необходимо выбрать (меню рис. 9.19) тип окна, установив прежде значения двух параметров:

- ширина окна в виде числа попадающих в него измерений;
- значение масштабной константы Хубера.

Метод робастного сглаживания применим также и к экспериментальным зависимостям (см. разд. 11.3), поскольку он допус-

кает, что шаг по *X* может быть не только постоянным, как в случае временного ряда, но и плавающим.

При методе Фурье производится вычисление спектра, из которого вырезается частотная полоса НЧ-ВЧ, после чего повторным преобразованием Фурье получается отфильтрованный сигнал. Для привязке к шкале частот в поле dt необходимо задать размерность временного шага анализируемого временного ряда. При dt=1 необходимо установить НЧ=*i*/*n*, ВЧ=*j*/*n*, где *n* – длина временного ряда, *i, j* – порядковые номера гармоник амплитудного спектра (см. разд. 9.3) для выделяемой полосы частот.

Результаты. Строится график сглаживающей кривой с наложенной точной диаграммой значений временного ряда.

Сглаженные значения можно перенести с графика в электронную таблицу для последующего анализа нажатием инструментальной кнопки «*СохрГраф*».

Сплайны: Для сглаживания сильно зашумленных рядов чрезвычайно эффективным является также использование сглаживающих сплайнов (см. разд. 4.2, а также ниже в примере 1) с малым числом шагов и малым значением коэффициента сглаживания.

Ограничения: Размер временного ряда не может превосходить *l*, где *l* = 16000, 5000, 1000, 100 при объеме матрицы данных в 64000, 20000, 4000 и 400 чисел. В методах фильтрации, требующих дополнительно введения параметра *a*, размер временного ряда должен быть *n*>*a*.

Пример 1

Задача. В разрушающем техническом эксперименте регистрировалось ускорение головы манекена водителя мотоцикла после столкновения его с неподвижным жестким препятствием (файл MOTO). Необходимо исследовать эту зависимость.

Обсуждение: Обратимся сначала к визуальному анализу (рис. 9.20). Прежде всего бросается в глаза большая зашумленность экспериментальных данных, вызванная погрешностями измерений ускорения. Требуется некоторым образом аппроксимировать эту зависимость, чтобы высветить ее скрытую шумом природу и получить более точные оценки закона изменения ускорения.

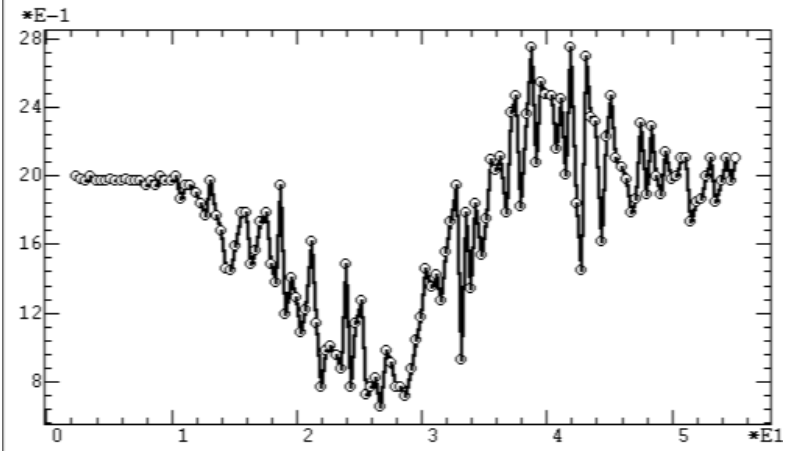


Рис. 9.20. Ускорение [g] головы водителя мотоцикла в зависимости от времени после столкновения с препятствием [ms]

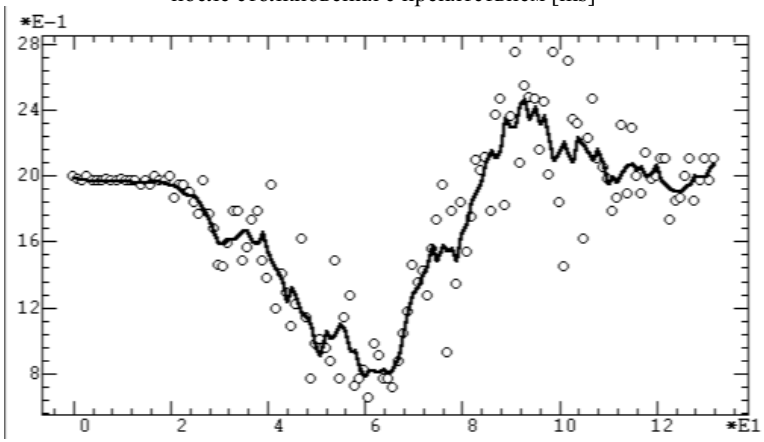


Рис. 9.21. Сглаживание скользящим средним по пяти точкам

Для столь зашумленных рядов более качественное сглаживание дают не методы скользящего среднего, а методы робастного сглаживания, менее чувствительного в резким колебаниям процесса.

Действительно, как легко видеть по сравнению графиков результатов (рис. 9.21, 9.22) метод робастного сглаживания Хубера (W=0.05, шири-

на=5, константа=10, треугольное окно) дает намного более чем сглаживание скользящим средним по пяти точкам. Еще более впечатляющий эффект дает применение сглаживающего сплайна с числом шагов 2 и коэффициентом 0.2 (см. разд. 4.2).

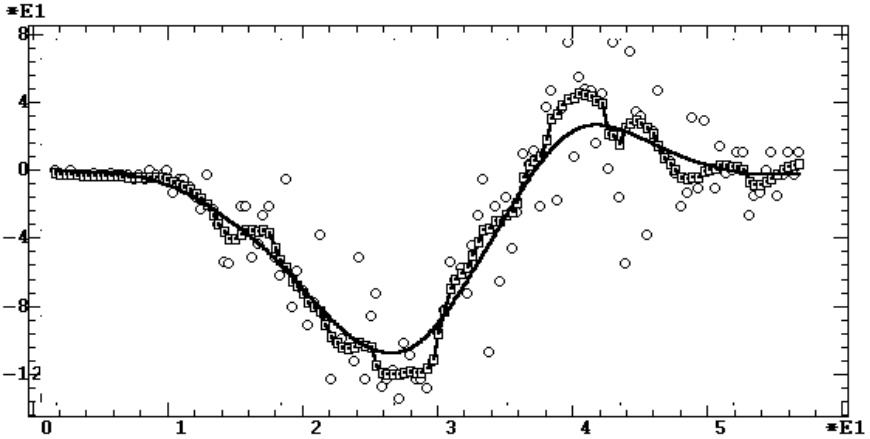


Рис. 9.22. Робастное сглаживание (квадраты) и сглаживание сплайном (сплошная линия)

Пример 2

Задача. На примере временной динамики авиаперевозок (рис. 9.8) визуально оценим результаты различных методов сглаживания и фильтрации.

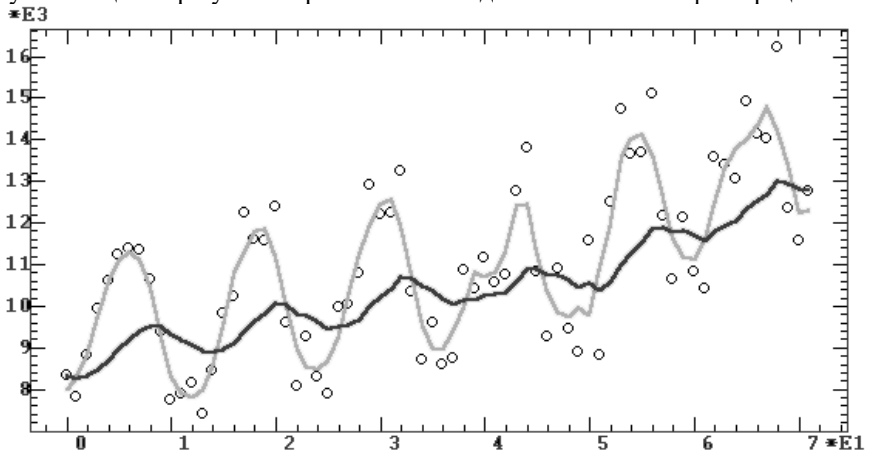


Рис. 9.23. Сглаживание ряда *Flight* (круги) скользящим средним по трем точкам (серое) и экспоненциальное сглаживание с коэффициентом w=0,1, способствующим выделению тренда (черное)

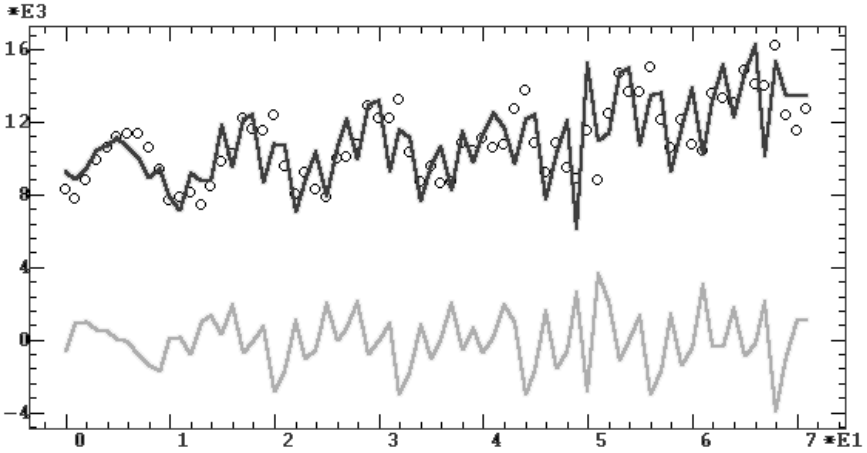


Рис. 9.26. Фильтрация ряда *Flight* (круги) дифференцированием (серое) и дифференцированием–интегрированием (черное)

Продолжение анализа данного примера см. в следующем разделе.

9.5. Авторегрессионные модели

Назначение. Построение моделей авторегрессии и проинтегрированного скользящего среднего (ARIMA) считается полезным для описания и прогнозирования поведения как стационарных временных рядов, так и нестационарных процессов, проявляющих однородные колебания вокруг изменяющегося среднего значения.

С другой стороны, построение подобных моделей крайне сложно для понимания и реализации[1]. В качестве альтернативы нами предложены более простые и наглядные *фурье–модели* (разд. 9.6), которые часто позволяют получать и более точные прогнозы, при этом учитывающие наличие в процессе тренда и скачков.

Действия. Сначала из электронной таблицы нужно выбрать (см. бланк рис. 6.4) переменную, представляющую анализируемый временной ряд.

Затем по подтверждению можно получить график частной автокорреляционной функции и график ошибки прогноза в зависимости от числа членов авторегрессионной модели, который позволяет предварительно оценить необходимое число параметров модели AR по положению минимума ошибки прогноза.

---

[1] Объем программного кода ARIMA–процедуры сопоставим с совокупным кодом таких сложнейших процедур многомерной статистики, как кластерный, дискриминантный, факторный анализ и многомерное шкалирование вместе взятые.

Установка параметров модели. Начальные значения основных параметров модели устанавливаются в экранном бланке рис. (9.28):

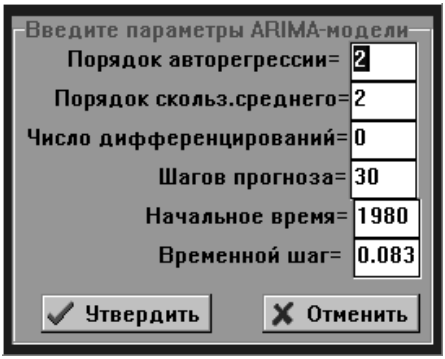


Рис. 9.28. Бланк ввода параметров ARIMA–модели

p_0 — число авторегрессионных параметров;

q_0 — число параметров модели скользящего среднего;

d_0 — количество дифференцирований временного ряда;

f — число шагов прогнозирования будущих значений временного ряда;

t_0 — начальное значение временного параметра;

dt — шаг временного параметра.

После заполнения всех позиций нажмите Enter или экранную кнопку «*Утвердить*».

Далее начинается процесс вычислений, который включает предварительное и окончательное оценивание параметров модели.

Предварительное оценивание состоит в подборе числа параметров модели и в приближенной оценке их значений. На каждом шаге предварительного оценивания выдается информационная надпись, содержащая параметры текущей модели: p_i, q_i, d_i, а также значение статистики *хи–квадрат,* уровень значимости нулевой гипотезы об адекватности модели временному ряду и подсказка о принятии или непринятии нулевой гипотезы.

На каждом последующем шаге увеличивается на единицу число авторегрессионных членов p_i, число параметров модели скользящего среднего q_i или число дифференцирований временного ряда d_i. При этом для предшествующих параметров устанавливаются их исходные значения (p_0, q_0 или d_0), что обеспечивает перебор всех комбинаций числа параметров в диапазонах $p_i = <p_0, p_0+2>$; $q_i = <q_0, q_0+2>$; $d_i=<d_0, d_0+2>$.

Процесс предварительного оценивания прекращается по принятию нулевой гипотезы или по исчерпанию допустимого числа параметров.

Итерационный процесс вычисления окончательных оценок параметров модели продолжается до достижения требуемой точности оценки или по окончании 25 итерационных шагов. На каждом шаге выдается информационная надпись, содержащая номер шага и текущее условие окончания итераций.

Результаты. По окончании итерационного процесса выдаются оцененные значения параметров авторегрессионной модели и модели скользящего среднего.

Затем выводится таблица, в которой для каждого шага прогнозирования указываются номер шага, среднее значение прогноза, стандартная

ошибка прогноза, доверительный интервал прогноза для установленного критического уровня значимости α, а также три варианта прогноза.

После этого по подтверждению могут быть выданы следующие графики:

- график временного ряда со средним значением и интервалом стандартной ошибки прогноза;
- графики временного ряда с тремя вариантами прогноза.

Далее производится анализ остатков с выдачей статистики *Дурбина–Ватсона* и уровня значимости нулевой гипотезы отсутствия коррелированности остатков. Здесь же по подтверждению может быть выдан график остатков.

Полученные результаты можно перенести в электронную таблицу для последующего анализа и построения комплексных графиков с числовой выдачи результатов посредством *буфера обмена* (см. разд. 2.4) или напрямую с графиков данных нажатием инструментальной кнопки «*СохрГраф*».

Ограничение. Размер временного ряда не может превосходить l, где l = 10560, 3300, 660, 66 при объеме матрицы данных в 64000, 20000, 4000 и 400 чисел.

Пример

Задача. Построим ARIMA-моделѝ для временного ряда авиаперевозок (см. рис. 9.8) после его нормирования и удаления линейного тренда (см. рис. 9.9).

Обсуждение: Рассмотрим первый результат анализа — график ошибки прогноза на рис. 9.29. Как можно заметить, минимум ошибки прогноза имеет место при 3–4 параметрах в модели авторегрессии. Поэтому в качестве начальных условий можно указать четыре параметра авторегрессионной модели и 0 параметров в модели скользящего среднего, поскольку в анализируемом ряде шумовая составляющая очевидно невелика по сравнению с закономерной периодической составляющей и, следовательно, в первом приближении можно ограничиться простейшей моделью белого шума.

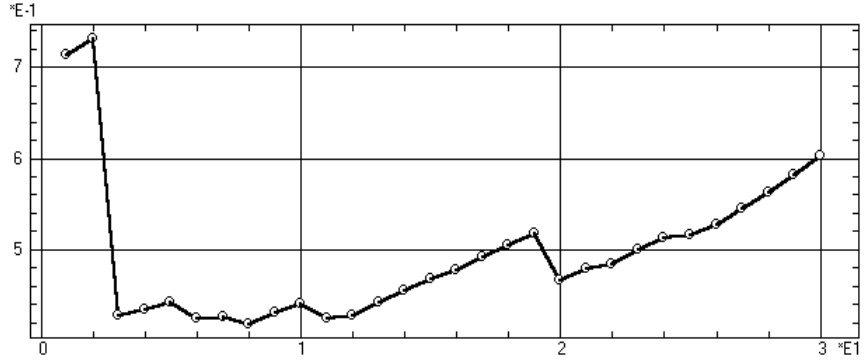


Рис. 9.29. График ошибки прогноза AR–модели от числа параметров

Результаты:

```
АВТОРЕГРЕССИЯ БОКСА-ДЖЕНКИНСА.  Файл: spec1.std Переменная flight
      Модель: AR=4, MA=0, дифф=0, хи2=5.12, ст.своб=7,значим=0.645
```

```
          Гипотеза 0: <Модель адекватна временному ряду>
                    Коэффициенты авторегрессионной модели
     a0        a1        a2        a3        a4        a5        a6
  -1.6E-5    0.662   -0.196     0.527     -0.74
             Коэффициенты модели скользящего среднего
Шум    b1        b2        b3        b4        b5        b6
     0.21
Время   Ср.прогн   Ст.ошиб   Довер.инт   Прогн1   Прогн2   Прогн3
1986     -1.305    0.5239       1.026    -1.887    1.033   -1.919
1986    -0.9899    0.6284       1.231    -1.239   0.8491  -0.9013
1986    0.03444    0.6411       1.255  - 0.4842   0.256   -0.528
1986    -0.3534    0.7046       1.38     -1.765  -0.09418 -0.6128
1986      0.203    0.7055       1.382  - 0.1943   0.5309   2.414
1986      0.954    0.7557       1.48       0.9131  0.5516   2.497
1986     0.3801    0.7639       1.496      0.3331  0.3516   1.029
1987     0.4332    0.8058       1.578      0.9709  0.4742   2.571
1987      0.565    0.8479       1.66    - 0.07651 0.164    0.9153
1987    -0.2161    0.8481       1.661   - 0.9155  0.6421  -0.8492
1987    -0.3066    0.8481       1.661   - 0.4995  0.6371   0.215
1987    -0.1833    0.8499       1.664   - 0.8315  0.7391  -0.9581
   . . . . . .
      Корреляция остатков Дурбина-Ватсона=1.11, значимость=0.344
         Гипотеза 0: <Модель адекватна временному ряду>
```

Обсуждение: Приведенная выдача результатов показывает, что произведенный выбор параметров позволил получить достаточно адекватное описание временного ряда 4-х параметрической моделью, с помощью которой произведен прогноз изменения интенсивности авиаперевозок на 30 месяцев вперед. Далее выдаются графики прогноза и остатков.

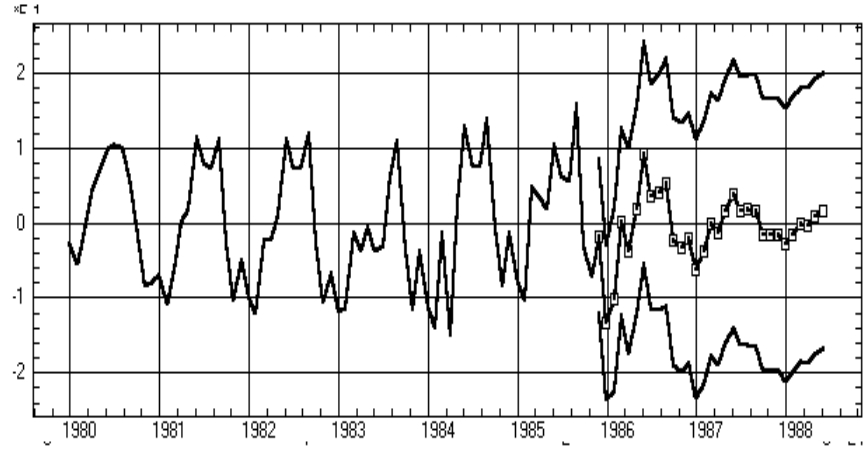


Рис. 9.30. График временного ряда со средним прогнозом и 95%–ным доверительным интервалом ARIMA–прогноза

Как можно заметить из рис. 9.30, амплитуда сезонного изменения среднего прогноза существенно уменьшается с течением времени (в соотнесении с более стабильной сезонной вариабельностью исходного ряда), а зона доверительного интервала (т. е. ошибки прогноза) достаточно широка, что говорит о недостаточной точности и устойчивости модели.

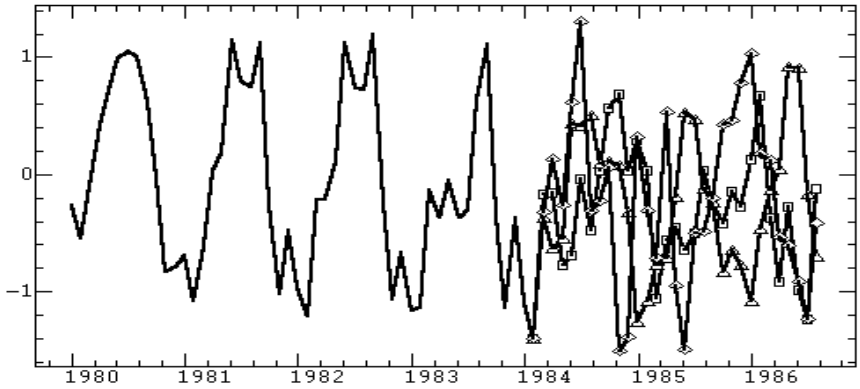


Рис. 9.31. Временной ряд с тремя генерациями ARIMA–прогноза

Значительное влияние шумовой составляющей выявляется и на рис 9.31, где три генерации прогнозов существенно различаются между собой вплоть до прямо противоположных предсказаний в некоторых точках. Величина остатков (рис. 9.32) также заметно возрастает от начала к концу ряда. Все выявленные недостатки в большой степени относятся собственно к рассматриваемому методу.

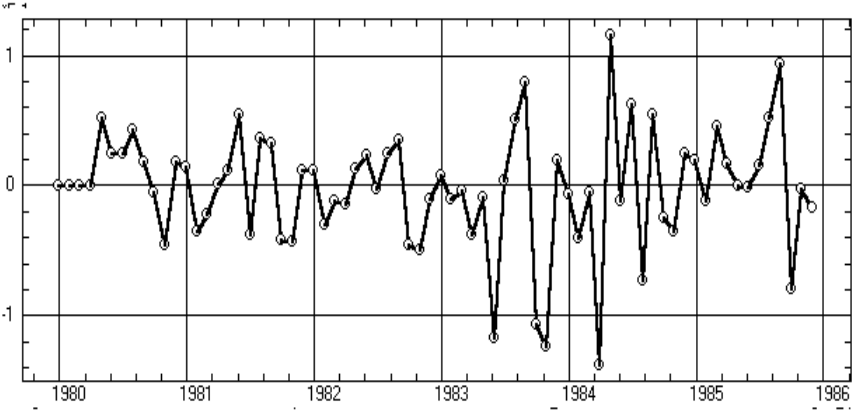


Рис. 9.32. График распределения ARIMA–остатков

В качестве варианта дальнейшего анализа можно предложить предварительную трансформацию временного ряда методом логарифмирования, что позволяет сместить интерес исследования с абсолютного изменения числа пассажиров на его относительное изменение.

Для достижения стационарности можно также попробовать применение дифференциального фильтра для удаления линейной тенденции и сезонного 12–месячного фильтра для устранения сезонных изменений.

Далее можно построить модели для различных отрезков временного ряда с целью выяснения однородности данных в смысле применимости одной и той же модели к первой и второй их половине.

Продолжение анализа данного примера см. в следующем разделе.

9.6. Фурье–модели

Назначение. Данный метод разработан нами в 1995 г. как средство моделирования *нестационарных* временных рядов, имеющих выраженные гармонические составляющие. *Фурье–модели*, в отличие от большинства других, являются многоцелевым инструментом и могут применяться как для прогнозирования, так и для интерполяции, фильтрации и сглаживания временных рядов.

Действия и результаты. Если в электронной таблице содержится несколько переменных, то нужно выбрать представляющую анализируемый временной ряд (см. бланк рис. 6.4). Последующие действия и результаты включают следующие шаги.

1. Предварительная коррекция процесса. Выделению гармонических составляющих и выбору параметров спектральной модели часто мешает наличие в процессе *тренда* и высокоамплитудных, непериодических *скачков* (в литературе используется также термин «интервенции»). Удаление этих компонентов производится установками меню (рис. 9.33). Произведенные коррекции будут автоматически восстановлены при прогнозировании временного ряда.

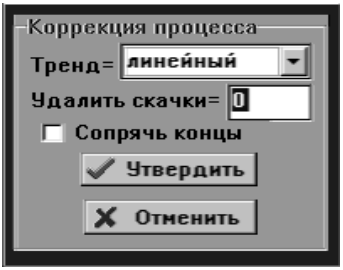


Рис. 9.33. Бланк коррекции процесса

Для нивелирования тренда предлагаются две модели: линейная и параболическая. Применимость той или иной модели может быть предварительно оценена методом простой регрессии (разд. 10.2) с доступным там же прогнозированием. Следует подчеркнуть, что линейная модель часто является более предпочтительной, несмотря на отклонения процесса от линейности, поскольку эти отклонения будут автоматически учтены спектральной моделью.

Скачки характеризуются экстраординарным значением производной (скорости изменения процесса). Для их удаления в поле ввода бланка рис. 9.32 следует задать уровень «нормальных» значений производной. Оценить этот уровень можно, вычислив предварительно производную и посмотрев на ее график (методом дифференцирования из разд. 9.4). Для более точной оценки полезно сохранить данные с графика в электронной таблице (кнопка *«СохрГраф»*) и там изучить значения производной в цифровом виде. При нулевом значении уровня производной скачки не удаляются.

Для процессов монотонного характера полезно дополнительно установить режим «сопряжения концов», который состоит в повторном удалении «псевдотренда» так, чтобы начальные и конечные амплитуды скорректированного процесса совпадали, иначе при прогнозировании может наблюдаться скачок в начале прогноза (вследствие математических свойств Фурье-преобразования). Для процессов с высокочастотными колебаниями *сопряжение концов* не обязательно, а иногда и нежелательно.

График скорректированного временного ряда выдается для визуальной оценки и изучения. При неудовлетворительном результате коррекцию можно повторить.

2. Оценочный график АЧХ. Оценку гармонических составляющих процесса для принятия решения о разделении детерминированных и шумовых составляющих легко произвести по графику *амплитудно–частотной характеристики* или амплитудного спектра (см. разд. 9.3). Шкала амплитуд на графике для удобства представлена в процентах от максимальной амплитуды.

3. Параметры модели. После этого уже можно осмысленно определить необходимые параметры фурье-модели (бланк рис. 9.34):

- число шагов прогноза развития процесса (при нулевом значении этого параметра прогноз не строится);

Рис. 9.34. Бланк параметров фурье–модели

- число делений временного ряда для *экспоненциального усреднения* спектра; значение 1 соответствует усреднению спектра со спектром отрезка в 1/2 длины временного ряда от его конца, значение 2 — с отрезками 1/2 и 1/4 длины временного ряда и т. д. (при значении 0 усреднение не производится, на длине отрезка менее 8 отсчетов усреднение автоматически завершается, поэтому строго контролировать верхний предел этого значения не требуется);
- верхняя и нижняя границы диапазона частот удаляемых спектральных составляющих (*фильтрация*), при нулевых значениях фильтрация не производится;
- амплитудный порог, ниже которого удаляются все спектральные составляющие (при нулевом значении удаление не производится);
- признак необходимости *адаптации* усеченной модели к временному ряду (при отсутствии признака адаптационный механизм не запускается, для полной спектральной модели адаптация эффекта не имеет);
- начальное время процесса и его временной шаг необходимы только для временной шкалы результатов прогнозирования, во всех других результатах время без ущерба для общности представляется последовательностью натуральных чисел.

Нажатие на кнопку «*Утвердить*» продолжает дальнейший анализ в соответствии с установленными параметрами, а нажатие на кнопку «*Отменить*» приводит к завершению работы процедуры.

Нулевые значения параметров соответствуют *чистой* фурье-модели, точно воспроизводящей временной ряд.

Продолжение анализа зависит от наличия или отсутствия заказа прогнозирования.

4. Результаты без прогноза. В случае отсутствия *прогнозирования* в качестве числовых результатов выдается последовательность значений временной модели анализируемого ряда и *среднеквадратичного шума* (характеризует отличие модели от исходного временного ряда), а также совместный график скорректированного временного ряда и его фурье-модели, т. е. без восстановления тренда и скачков). Для визуализации модели исходного временного ряда следует фиктивно заказать один шаг прогноза.

Далее по подтверждению можно возвратиться к новым установкам параметров модели. Такой цикл позволяет последовательно удалять из модели отдельные спектральные составляющие.

5. Анализ остатков. В случае наличия прогноза по подтверждению может быть произведен *анализ остатков*, т. е. разностей между моделью и временным рядом. В качестве числовых результатов выдается таблица значений следующих показателей (аналогично методу простой регрессии, см. разд. 10.1):

```
X  Yэксп  Yмодл  остаток  Ст.остат  Ст.ошиб  Довер.инт
```

Затем выдается график *регрессионных остатков* для последовательных точек временного ряда.

6. Прогноз. Выдается таблица параметров фурье-модели, включающая следующие значения:

```
Амплитуда    Период    Фаза
```

Затем выдается таблица *прогноза*, включающая значения следующих показателей (аналогично методу простой регрессии, см. разд. 10.1):

```
Хпрогн    Yпрогн    Ст.ошиб    Довер.инт
```

В завершение выдается график временного ряда, совмещенный с графиком временной модели с зоной прогноза. В зоне прогноза указаны верхняя и нижняя границы *доверительного интервала* прогнозируемых значений.

Полученные результаты можно перенести в электронную таблицу для последующего анализа и построения комплексных графиков с числовой выдачи результатов посредством буфера обмена (*Clipboard*) или напрямую с графиков нажатием инструментальной кнопки «*СохрГраф*».

Ограничение. Размер временного ряда не может меньше 8 или превосходить *l*–*d*, где *d* — число шагов прогноза: *l* = 9000, 2800, 570, 57 при объеме матрицы данных в 64 000, 20 000, 4000 и 400 чисел.

Пример 1

Задача. Данный пример предназначен для сравнения фурье-моделей с другими методами анализа временных рядов. Поэтому в качестве исходных данных используем хорошо знакомый в данной главе временной ряд изменения авиаперевозок (см. рис. 9.9). Внимательное изучение исходного временного ряда показывает, что в нем с течением времени нарастают амплитуды высокочастотных нерегулярных составляющих, существенно искажающие главный годичный ритм (нестационарность).

Шаг 1. Поскольку временной ряд уже нормирован и центрирован предварительная коррекция модели (тренд, скачки) не нужна. Амплитудный спектр процесса приведен на рис. 9.10. Для начала произведем только фильтрацию высокочастотных составляющих спектра в диапазоне от 20 до 36. Для этого в бланке параметров фурье-модели (рис. 9.34) следует

заполнить позиции верхнего и нижнего фильтров, остальные позиции оставим нулевыми.

Р е з у л ь т а т ы:

```
ФУРЬЕ-МОДЕЛИ Файл: spec1.std              Переменная flight
Ст.откл.шума=0.343
```

Обсуждение: В результате фильтрации спектр процесса приобрел вид рис. 9.35. Как можно видеть по графику результата фильтрации (рис. 9.36), удаление высокочастотных составляющих из спектра действительно привело к желаемому сглаживанию быстрых колебаний временного ряда. Полученный результат близок к результату фильтрации ряда скользящим средним (см. рис. 9.24). Удаленные компоненты соответствуют шуму (характеризует различие между моделью и временным рядом) со стандартным отклонением 0,343.

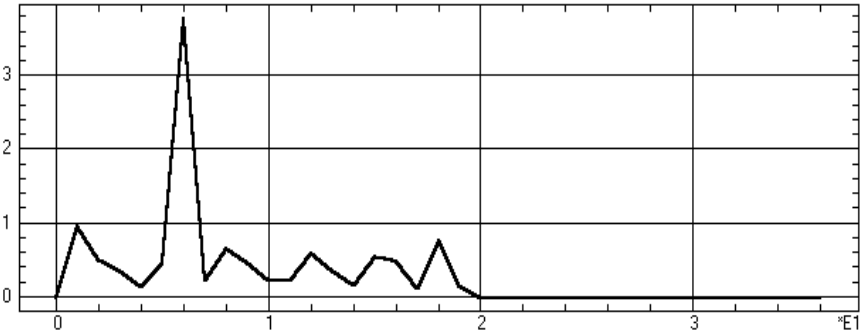


Рис. 9.35. График спектральной модели после удаления высоких частот

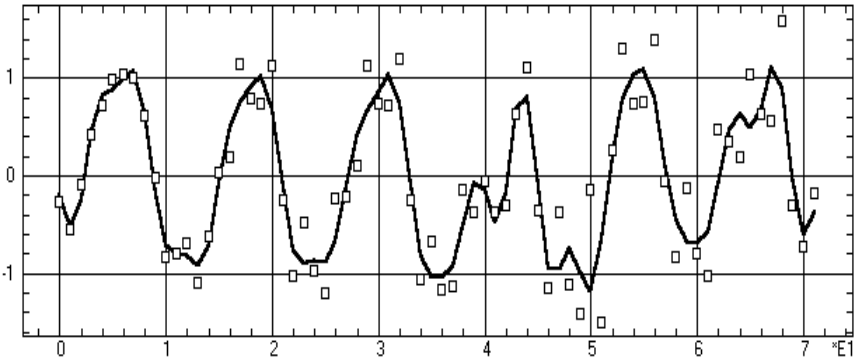


Рис. 9.36. График фильтрации временного ряда авиаперевозок

Ш а г 2. Удалим теперь из спектра основную спектральную гармонику (сезонный компонент) в диапазоне частот 4–8.

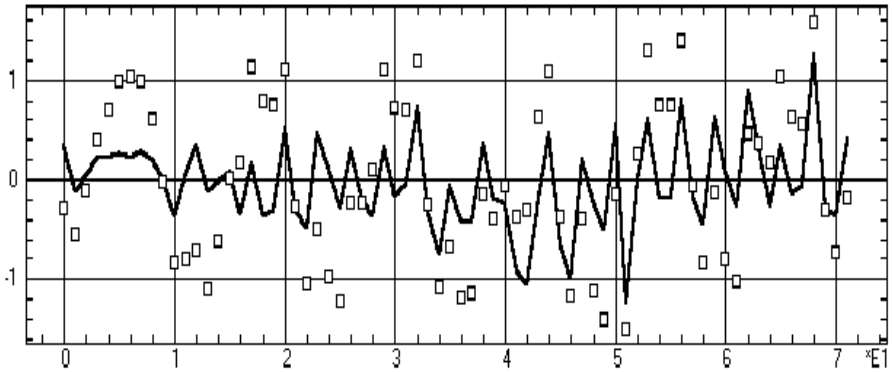


Рис. 9.37. График спектральной модели после сезонной фильтрации

Обсуждение: Сравним полученный результат (рис 9.37) с результатом сезонной фильтрации из разд. 9.4 (см. рис. 9.27). Как можно заметить, фурье–фильтрация обеспечивает удаление сезонной составляющей без резких выбросов, характерных для рис. 9.27, кроме того результат охватывает всю длину временного ряда, без присущего сезонному дифференцированию «обрезания» конца ряда.

Ш а г 3. Используем теперь нефильтрованную («чистую») фурье–модель для прогнозирования авиаперевозок на 30 шагов, т. е. на два с половиной года. Для этого в бланке рис. 9.34 заполним только поле прогноза.

Обсуждение: Как видно из полученного результата (рис 9.38), *чистая* фурье–модель полностью повторяет временной ряд и ее прогноз является повторением начального участка процесса.

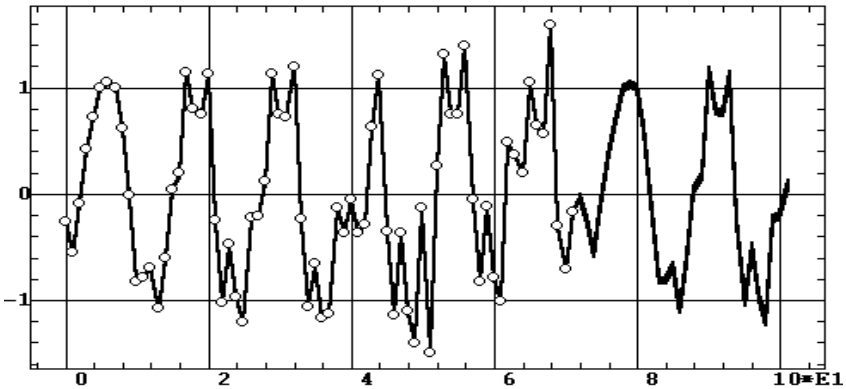


Рис. 9.38. Временной ряд авиаперевозок с прогнозом на 30 шагов

Ш а г 4. Изменим модель, удалив из спектра низкоамплитудные составляющие, не превышающие 15% от максимума.

Р е з у л ь т а т ы (сокращенно):

Файл: spec1.std Переменная flight

Параметры Фурье-модели

Амплитуда	Период	Фаза
0.00571	0	180
0.217	73	-9.34
0.894	12.2	161
0.145	9.13	-38.8
0.145	6.08	-105
0.172	4.06	-15
0.313	3.04	71.3
0.258	2.92	-152

Ст.отклонение остатков (различие модели и процесса)=0.0817

X	Yэксп	Yмодл	остаток	Ст.остат	Ст.ошиб	Довер.инт
0	-0.256	-0.457	0.202	0.706	0.296	0.582
1	-0.537	-0.212	-0.325	-1.12	0.295	0.581
2	-0.081	-0.166	0.0851	0.301	0.295	0.58
3	0.432	0.279	0.153	0.538	0.295	0.58
4	0.737	0.658	0.0789	0.279	0.294	0.579
5	1.01	0.659	0.347	1.21	0.294	0.579
6	1.06	0.963	0.097	0.342	0.294	0.578
7	1.02	1.25	-0.234	-0.807	0.294	0.578
8	0.635	0.786	-0.151	-0.519	0.293	0.577
9	-0.00816	-0.0947	0.0865	0.306	0.293	0.577
10	-0.822	-0.511	-0.311	-1.08	0.293	0.576

. . .

Хпрогн	Упрогн	Ст.ошиб	Довер.инт
72	-0.457	0.296	0.582
73	-0.212	0.296	0.583
74	-0.166	0.297	0.584
75	0.279	0.297	0.584
76	0.658	0.297	0.585
77	0.659	0.298	0.586
78	0.963	0.298	0.587
79	1.25	0.298	0.587
80	0.786	0.299	0.588

. . .

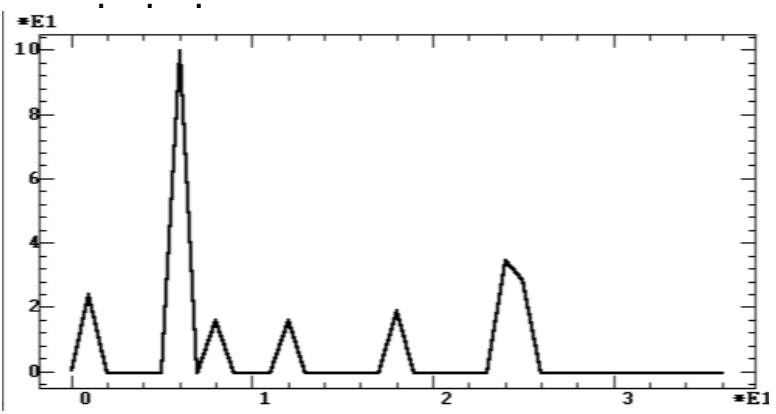


Рис. 9.39. График модели после удаления низкоамплитудных составляющих

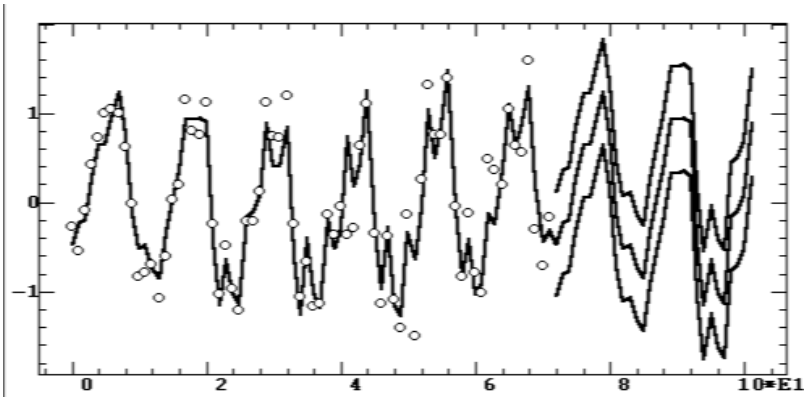


Рис. 9.40. Временной ряд авиаперевозок с фурье–моделью и с прогнозом на 30 шагов

Обсуждение: Здесь с иллюстративными целями приведена также и числовая выдача результатов с параметрами модели, анализом остатков и прогнозом. Амплитудный спектр модели приведен на рис. 9.39.

Как видно из результата моделирования и прогноза (рис 9.40), модель уже не повторяет временной ряд. По сравнению с рис. 9.38 появились высокочастотные колебания на начальном участке процесса, но снизилась их амплитуда на конечном участке. Однако прогноз остается повторением начального участка смоделированного ряда.

Ш а г 5. Добавим к данной модели еще и адаптацию.

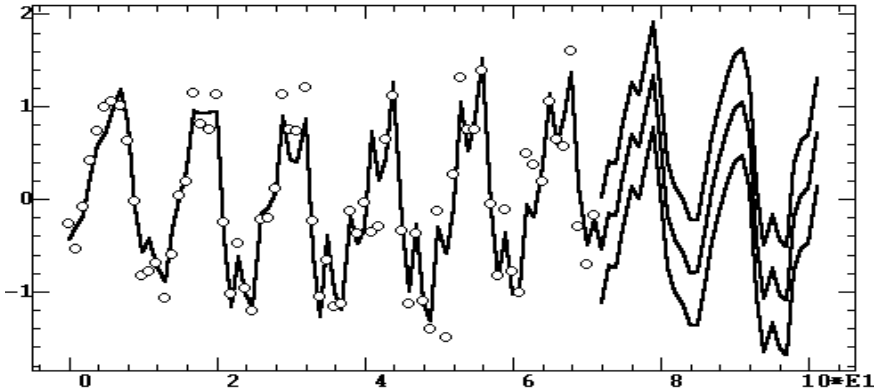


Рис. 9.41. Фурье–модель с адаптацией

Обсуждение: Как видно из рис 9.41 за счет адаптации само моделирование процесса по сравнению с рис. 9.40 более близко к чистой фурье-модели (см. рис. 9.38) и к исходному временному ряду, к тому же прогноз уже не повторяет начало временного ряда.

Ш а г 6. Добавим теперь в модель еще и два усреднения.

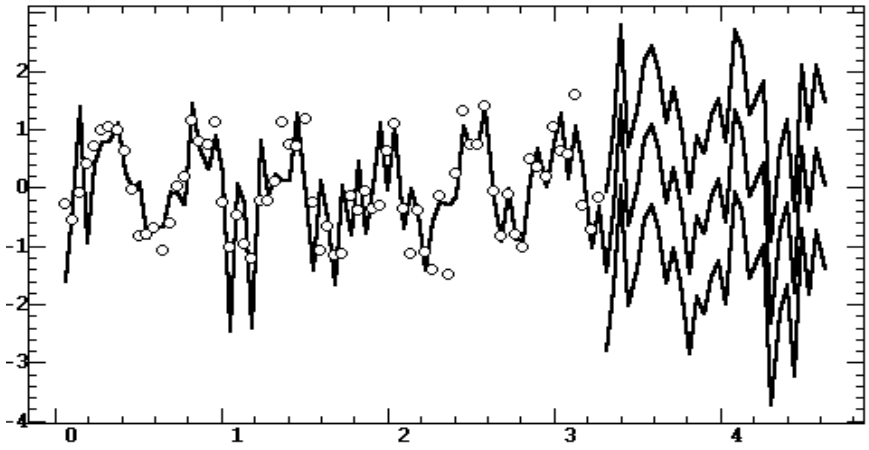


Рис. 9.42. Фурье–модель с адаптацией и двумя усреднениями

Обсуждение: Как видно из рис 9.42 за счет усреднения по сравнению с рис. 9.41 в зоне прогнозирования более представлены высокочастотные колебания, характерные для последнего участка временного ряда. Поэтому в отношении прогноза эту модель следует признать более адекватной, хотя при моделировании начального участка авиаперевозок там появились не характерные для него высокочастотные колебания.

Сравнение методов

З а д а н и е: Проведем сравнение предложенного фурье–метода с авторегрессионным подходом (разд. 9.5). Для этого исключим из рассматриваемого временного ряда его последнюю треть и сравним точность прогнозирования на этом участке, достигаемую с использованием каждого метода (рис. 9.38). Для фурье–модели используем уже опробованную схему двукратного усреднения с адаптацией и очисткой спектра.

Р е з у л ь т а т ы:

X	Y	Yфурье	Yarima
50	-0.1187	-0.861	-0.151
51	-1.482	0.281	-0.0169
52	0.2741	0.383	-0.0822
53	1.321	-0.0206	0.824
54	0.7636	0.453	0.769
55	0.7658	1.33	0.199
56	1.405	1.16	0.373
57	-4.042E-2	0.176	0.047
58	-0.8149	-0.457	-0.62
59	-0.1097	-0.869	-0.505
60	-0.7781	-1.63	-0.464

X	Y	Yфурье	Yarima
61	-1.003	-1.89	-0.661
62	0.4955	-0.778	-0.27
63	0.3761	0.543	0.0779
64	0.2025	0.623	0.0485
65	1.06	0.141	0.272
66	0.6503	0.375	0.403
67	0.5779	0.817	0.139
68	1.603	0.484	0.0277
69	-0.2873	-0.0613	-0.0376
70	-0.7006	-0.0925	-0.311
71	-0.1595	-0.361	-0.385

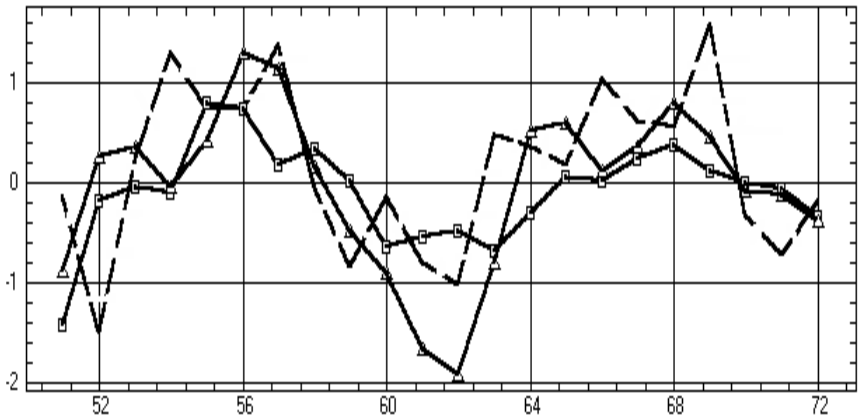


Рис. 9.43. Сравнение фурье–прогноза (треугольники) и ARIMA–прогноза (квадраты) с временным рядом авиаперевозок (пунктирная линия)

Обсуждение: В числовых результатах выше приведены истинные значения прогнозируемого временного ряда *Y* и прогностические значения фурье– и ARIMA–моделей; *Yфурье, Yarima.* Для количественного сравнения обоих методов вычислим среднее стандартное отклонение *S* для разностей между прогностическими и истинными значениями (*Y–Yфурье, Y–Yarima*) и получим: *S*=0,578 для фурье–модели и *S*=0,641 для ARIMA–модели.

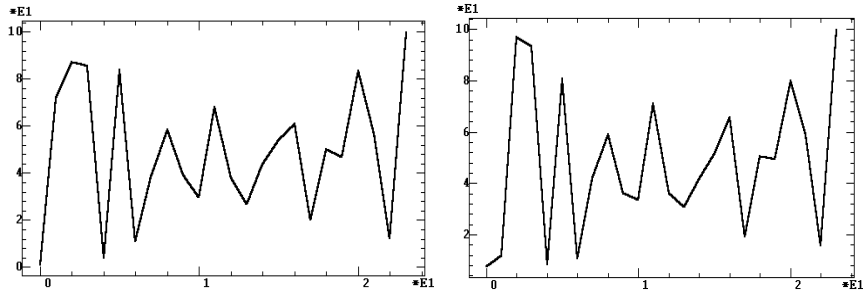
Таким образом фурье–модель дает прогноз в среднем на 9,8%=(0,641–0,578)/0,641*100 точнее, чем широко популярная, но алгоритмически неизмеримо более сложная и трудная для понимания ARIMA–модель.

Заключение: На основании выводов по рассмотренным примерам можно сделать следующий вывод: предложенные фурье–модели являются многоцелевым инструментом исследования и в каждой из своих областей применения дают результаты не хуже, а зачастую – и лучше, чем популярные и давно известные аналоги.

Пример 2

Задача. Попробуем смоделировать и спрогнозировать натуральный (непреобразованный) временной ряд изменения урожайности зерновых в СССР из примера 1 к разд. 10.3. Поскольку, как там показано, для этого ряда более подходит параболическая модель тренда по сравнению с линейной, то для сравнения используем обе эти модели. Кроме того в этом процессе не наблюдается явных скачков, что следует из самой природы процесса (примеры с удалением скачков рассмотрены разд. 14.4).

Результаты (сокращенно и без числовой выдачи):

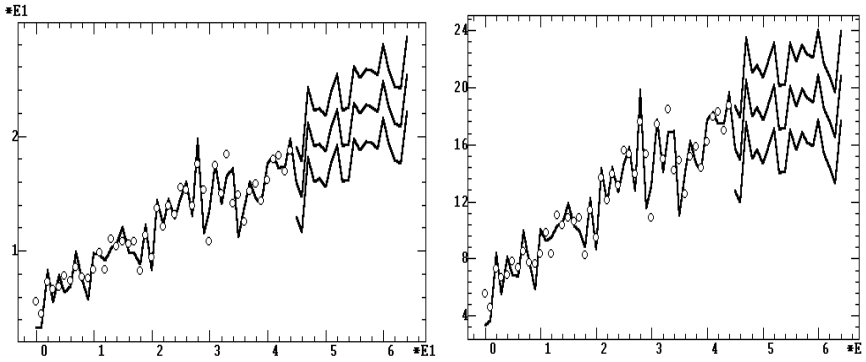


а *б*

Рис. 9.44. Спектры изменения урожайности зерновых после удаления тренда:
а — линейный тренд; *б* — полиномиальный тренд

Обсуждение: Как видно из графиков полных спектральных моделей (рис. 9.44) малозначимые компоненты можно отсечь на уровне 30% от амплитуды максимальной гармоники. В связи с небольшой длиной временного ряда выполним только два усреднения с адаптацией, дополнительной фильтрации делать не будем, выполним прогноз изменения урожайности зерновых на 20 лет вперед, т. е. до 2009 г.

Р е з у л ь т а т ы (сокращенно и без числовой выдачи):



а *б*

Рис. 9.45. Модель и прогноз урожайности зерновых в СССР:
а — линейный тренд; *б* — полиномиальный тренд

Обсуждение: Как видно из графиков моделей и прогнозов, параболическая модель (рис. 9.45) предсказывает в среднем практически полную стагнацию урожайности на 20 лет. Линейная же модель (рис. 9.45, *а*) предсказывает в среднем рост приблизительно линейного характера, на котором, однако, внимательным взглядом можно заметить продолжение замедления роста, характерное для последней трети процесса. Тем самым линейную модель тренда в данном случае следует признать более адекватной средним тенденциям временного ряда. Почему это происходит, несмотря на лучшее соответствие временного ряда параболической модели, выяв-

ленное в примере 1 разд. 10.3. Как можно заметить из сравнения спектров рис. 9.44 в линейной модели намного более выражена самая низкочастотная гармоника. И именно она компенсирует сравнительно меньшую адекватность модели линейного тренда, добавляя к нему длинно периодическую синусоиду. Это было уже отмечено в основной части данного раздела.

Еще следует отметить, что обе модели проявляют существенные высокочастотные колебания урожайности, характерные для последней половины временного ряда, определенную двумя заказанными в модели усреднениями.

Пример 3

З а д а ч а. Рассмотрим ежемесячную динамику цен на нефть Брент с 2001 по 2004 гг.[1] (рис. 9.46, файл OIL). Этот процесс визуально также имеет явную нелинейность, но в отличие от предыдущего примера не затухающую, а возрастающую. Поэтому здесь также попробуем сравнить две модели линейную и параболическую. Кроме того, в этом процессе не наблюдается явных скачков (примеры с удалением скачков рассмотрены разд. 14.4).

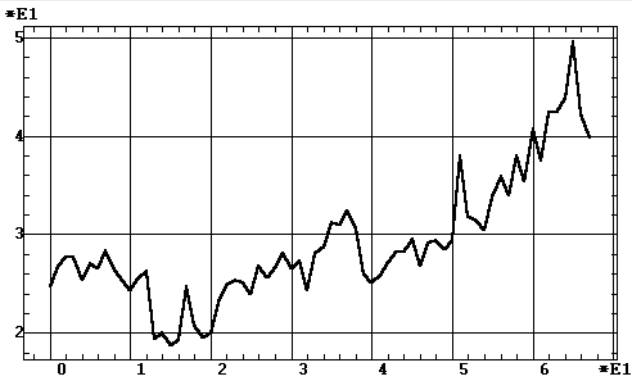
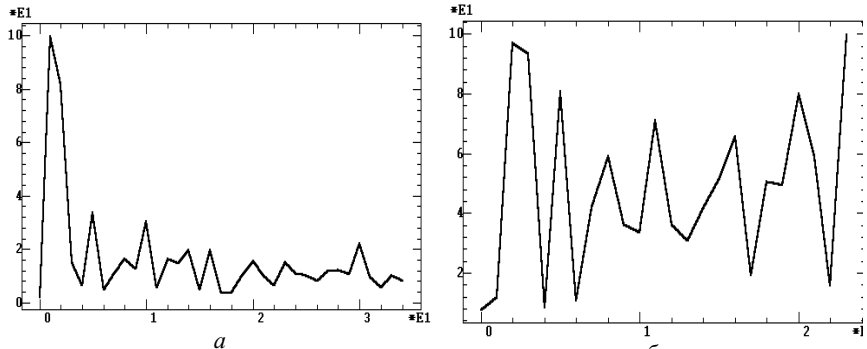


Рис. 9.46. Ежемесячная динамика цен на нефть Брент с 2001 по 2004 гг.

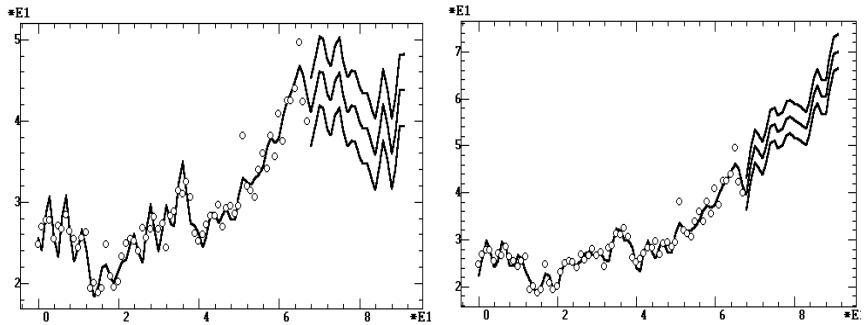
Обсуждение: Как видно из графиков полных спектральных моделей (рис. 9.47) малозначимые компоненты можно попробовать отсечь на уровне 10%. При линейном тренде превалирующую амплитуду имеет низкочастотная составляющая вследствие высокоамплитудных и длинно периодических колебания остатков относительно тренда. Полиномиальная модель намного более адекватна процессу и в ней в равной степени представлены низко- и высокочастотные составляющие. Остальные параметры спектральной модели оставим аналогичными примеру 2 и выполним прогноз на два последующих года.

[1] Данные предоставлены научным сотрудником ВНИИ внешнеэкономических связей В.А. Ярных



а *б*

Рис. 9.47. Спектры цен на нефть Брент после удаления тренда:
а — линейный тренд; *б* — полиномиальный тренд



а *б*

Рис. 9.48. Модели и прогнозы цен на нефть Брент:
а — линейный тренд; *б* — полиномиальный тренд

Обсуждение: Как видно из графиков моделей и прогнозов, линейная модель (рис. 9.48, *а*) за счет сильной низкочастотной составляющей в спектре (рис. 9.47, *а*) предсказывает в снижение цен на нефть в последующие два года. В противоположность этому параболическая модель (рис. 9.48, *б*) предсказывает продолжение тенденции примерно линейного роста цен на нефть. Однако из содержательных соображений в начале 2005 г. (т. е. не имея апостеорной информации) можно было бы сказать, что оба этих прогноза недостаточно реалистичны, и истина должна лежать где-то посредине. Возможно, в качестве модели тренда данного процесса более подошла бы парабола третьей или более высокой степени.

Глава 10

РЕГРЕССИОННЫЙ АНАЛИЗ

«Есть в Искаженном Мире закономерность,
но есть и отсутствие закономерности.
Ни что в нем не обязательно, ни что не необходимо»
[Зе Краггаш. О неумолимости правдоподобного]

Назначение. Во многих практических задачах, исследующих различного рода зависимости, необходимо на основании экспериментальных данных выразить зависимую переменную в виде некоторой математической функции от независимых переменных, т. е. построить *регрессионную модель*. Методы регрессионного анализа позволяют:

1) производить расчет различного вида регрессионных моделей с определением значений *параметров* модели (коэффициентов при независимых переменных);
2) проверить гипотезу адекватности модели имеющимся наблюдениям;
3) использовать модель для предсказания или *прогнозирования* значений зависимой переменной при новых или ненаблюденных значениях независимых переменных.

10.1. Общие регрессионные результаты

Стандартная последовательность результатов регрессионного анализа включает следующие компоненты (см. в примерах раздела 10.3):

1. Уравнение регрессии или модель, записанную в общем виде.

2. Таблица значений коэффициентов модели со стандартными ошибками вычисления каждого коэффициента.

3. Таблица дисперсионного анализа (см. ниже формулы) со столбцами: `Источник Сумма квадратов Степ.свободы Средн.сумма квадр` с тремя строками для параметров: регрессионные, остаточные и общие.

4. Таблица проверки нулевой гипотезы со следующими статистическими характеристиками:

- множественный коэффициент корреляции *R* между зависимой переменной и независимыми переменными;
- R^2 (квадрат *R* или *коэффициент детерминации*);
- приведенная или несмещенная оценка R^2;
- стандартная ошибка вычислений;
- значение статистики Фишера *F* и уровень значимости *P* нулевой гипотезы о равенстве нулю коэффициента множественной корреляции.

Нулевая гипотеза. Принятие последней нулевой гипотезы означает отсутствие соответствия между исходными данными и математической моделью, иными словами — модель неадекватно описывает экспериментальные данные. Если $P > 0{,}05$, нулевая гипотеза может быть принята, а модель отвергается как неадекватная экспериментальным данным.

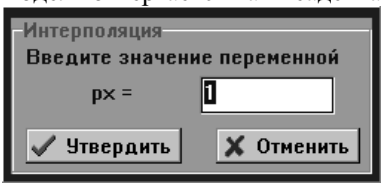


Рис. 10.2. Бланк ввода X для интерполяции Y

5. *Интерполяция.* Далее по полученному уравнению регрессии можно рассчитать *прогностические* (будущие) или же *интерполяционные* (промежуточные) значения зависимой переменной *Y* для заданных значений независимой(мых) переменной(ных) *X*. Задание значений *X* производится в бланке рис. 10.2.

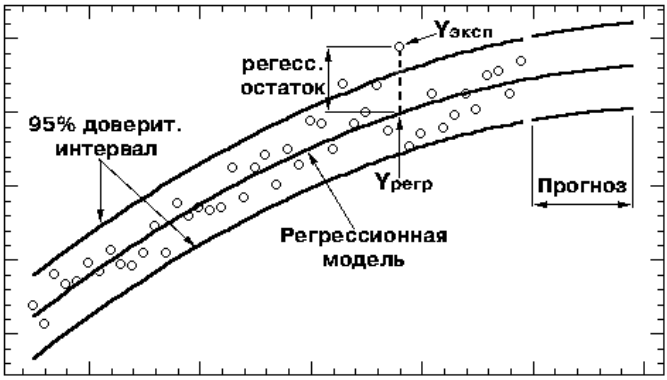


Рис. 10.3. Компоненты регрессионных графиков

6. *Регрессионный график.* Затем строится регрессионный график (рис. 10.3), на котором представлена диаграмма рассеяния экспериментальных точек с регрессионной кривой и зоной ее 95% доверительного интервала.

7. Дальнейший анализ может включать (рис. 10.4) следующие возможности:

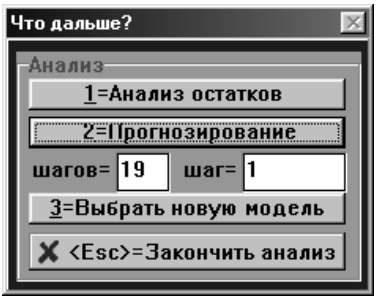


Рис. 10.4. Меню выбора дальнейшего анализа

- анализ регрессионных остатков;
- прогнозирование регрессионных значений в будущем;
- переход к выбору и анализу новой регрессионной модели;
- завершение анализа.

Анализ остатков. В ходе анализа *регрессионных остатков* выдается таблица (см. примеры разд. 10.3), где для каждого экспериментального значения параметра $X_{эксп}$ приведены следующие значения (рис. 10.3): экспериментальное значение $Y_{эксп}$, регрессионное значение $Y_{регр}$ (вычисленное для $X_{эксп}$ по уравнению регрессии), регрессионный остаток $Y_{эксп}–Y_{регр}$, остаток в единицах стандартного отклонения, стандартная ошибка $dY_{регр}$ регрессионного значения $Y_{регр}$, а также 95%–ный доверительный интервал $iY_{регр}$ регрессионного значения $Y_{регр}$.

Далее по подтверждению могут быть выданы два рисунка:

1) зависимость регрессионных остатков ($Y_{эксп}–Y_{регр}$ по оси Y) от экспериментальных значений $Y_{эксп}$ (по оси X);
2) зависимость регрессионных остатков (ось Y) от экспериментальных значений $X_{эксп}$.

Эти графики позволяют проверить предположения о линейности, однородности (гомогенности) и независимости ошибок и локализовать выбросы. Если указанные допущения выполняются, то на графике не будет наблюдаться заметной зависимости между остаточными и регрессионными или параметрическими значениями, т. е. оба графика будут представлять собой симметричное, случайное и равномерное распределение точек. Появление заметной закономерности в распределении остатков является индикатором определенной неадекватности модели экспериментальным данным.

Сохранение остатков. Остатки можно перенести с графика в электронную таблицу для последующего анализа остатков нажатием инструментальной кнопки «*СохрГраф*».

Дальнейший анализ остатков может состоять в построении гистограммы их распределения и проверки этого распределения на нормальность

средствами процедуры разд. 6.2; проверки коррелированности остатков с зависимой или независимой переменной в разд. 6.3 и т. д.). Чаще всего эта возможность используется для удаления тренда у временного ряда (см. разд. 9.1).

Прогноз. При выборе *прогноза* в правое поле меню дальнейшего анализа (рис. 10.3) нужно ввести число точек прогноза и величину шага прогноза и нажать кнопку прогноза. За этим выдается таблица, которая для каждого прогнозируемого X содержит следующие параметры: значение независимой переменной X_i, регрессионный отклик $Y_{регр}$, стандартную ошибку прогноза индивидуального значения отклика $dY_{регр}$ и 95%–ный доверительный интервал прогноза $iY_{регр}$. Строится график экспериментальных точек, регрессионной кривой, продолженной в зону прогноза с границами 95%–ного доверительного интервала прогноза (рис.10.3).

Частные отличия результатов многопараметрического регрессионного анализа рассмотрены в разд. 10.4.

10.2. Сравнение двух линий регрессии

Назначение. Для двух экспериментальных зависимостей производится сравнение их линий регрессии с проверкой нулевой гипотезы об отсутствии различий между ними.

Исходные данные представляются в виде псевдоматрицы, содержащей две пары переменных X_1, Y_1 и X_2, Y_2 с числом значений n и m.

Результаты включают значения T–статистики (Стьюдента) с уровнем значимости P для двух гипотез: о равенстве угловых коэффициентов регрессий и об отсутствии сдвига между ними. Если $P>0.05$, соответствующая нулевая гипотеза может быть принята.

Пример

Задача. При ревматоидном артрите из-за болезненности суставов происходит частичная атрофия мышц, что сказывается на физической силе. Чтобы проверить статистическую значимость этих изменений, в группе больных и в группе здоровых людей измерялись зависимости силы сжатия кисти от поперечного сечения предплечья (файл LR2). Результаты исследования изображены на рис. 10.5.

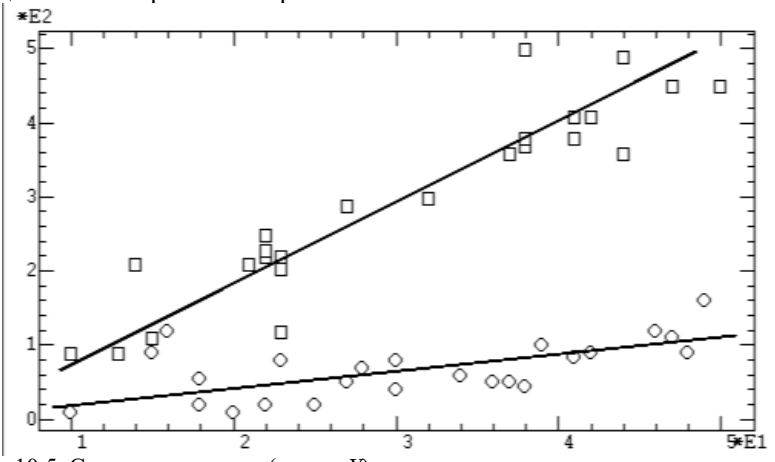


Рис. 10.5. Сила сжатия кисти (кг, ось *Y*) в зависимости от поперечного сечения предплечья (см^2, ось *X*) для здоровых людей (квадраты) и больных артритом (круги)

Результаты:

```
СРАВНЕНИЕ ДВУХ РЕГРЕССИЙ.  Файл: lr2.std
T(паралельность)=7.83, Значимость=9.27E-7, степ.своб = 44
   Гипотеза 1: <Есть различия между коэффициентами наклона>
T(равенство средних)=19.8, Значимость=8.32E-10, степ.своб = 44
    Гипотеза 1: <Есть различия в положении регрессионных прямых>
```

Выводы: Согласно результатам анализа выявлены различия как углов наклона, так и в сдвиге регрессионных зависимостей (оба уровня значимости близки к нулю).

10.3. Простая регрессия

Назначение. Процедура простой регрессии предоставляет возможность строить наиболее употребительные регрессионные модели для экспериментальных зависимостей от одной переменной, а также для временных рядов.

Если в предлагаемом списке моделей нет подходящей, то следует обратиться к разд. 10.6, где по формуле можно задать любую алгебраическую модель.

В ходе анализа можно получить следующие результаты:

- выбрать из нескольких математических моделей ту, которая с большей точностью описывает экспериментальную зависимость;
- вычислять интерполяционные значения;
- построить прогноз на будущее на основе выбранной модели с 100(1−α)%-ным доверительным интервалом;
- провести анализ регрессионных остатков.

Исходные данные представляют собой две парные *X* и *Y* переменные из электронной таблицы, или одну переменную *Y*, представляющую временной ряд для анализа его тренда (см. разд. 9.1).

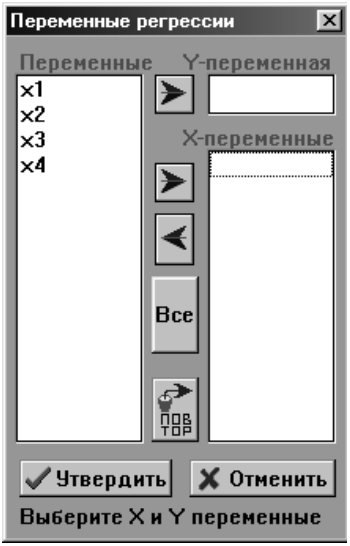


Рис. 10.6. Бланк выбора переменных для регрессионного анализа

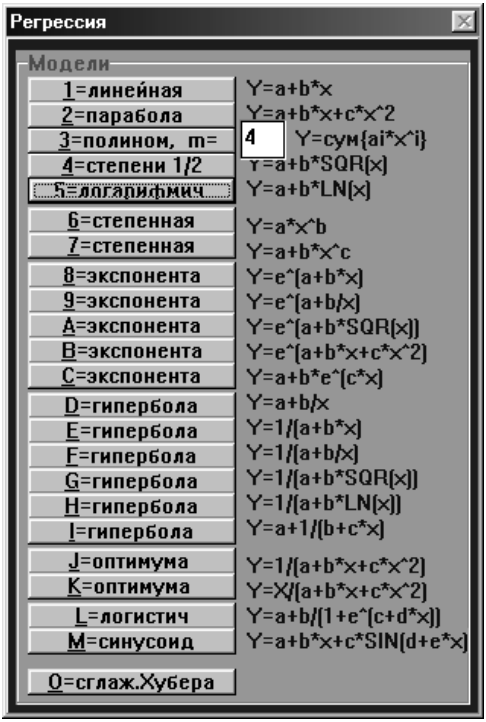


Рис. 10.7. Меню выбора регрессионной модели

Действия и результаты.

1. Сначала нужно выбрать из электронной таблицы для анализа две парные *X* и *Y* переменные или же одну переменную *Y*, представляющую временной ряд (см. рис. 10.6, рис. 2.3). Бланк выбора переменных отличается наличием двух полей: выбранной *Y*–переменной и *X*–переменных (в

Пример 1

Задача. Рассмотрим данные Госкомстата СССР по средней урожайности зерновых культур (центнеры с гектара) с 1945 по 1989 годы (файл CORN). Попробуем проанализировать закономерности изменения этого кардинально важного для всей экономики и благосостояния страны показателя, чтобы иметь возможность строить достоверные прогнозы на будущее.

Для этого нам, прежде всего, полезно визуально изучить график данного временного ряда (рис. 10.20).

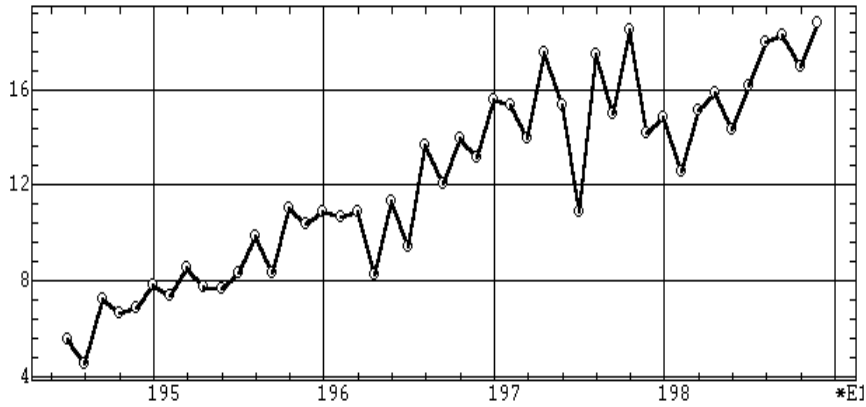


Рис. 10.20. Изменение урожайности зерновых в СССР с 1945 по 1989 гг.

Как легко заметить, в динамике урожайности преобладает линейно возрастающая тенденция с ежегодными нерегулярными колебаниями, поэтому естественным представляется описание этих данных линейной регрессионной мо-

делью. Можно также отметить: два периода увеличения нестабильности в 1962–1966 гг. и в 1973–1981 гг., два периода стагнации в 1958–1962 гг. (эпоха позднего Хрущева) и в 1970–1982 гг. (эпоха позднего Брежнева), а также ускорение возрастания урожайности в 1963–1973 гг. (эпоха раннего Брежнева) и в 1983–1989 гг. (послебрежневская эпоха и эпоха раннего Горбачева). Вот сколько много полезных и ассоциативных исторических выводов можно сделать из простого просмотра исходных данных.

Перейдем собственно к регрессионному анализу, применив линейную модель и вычислив по ней прогностическое значение на 2000 год.

Р е з у л ь т а т ы:

```
ПРОСТАЯ РЕГРЕССИЯ.  Файл: corn.std  Переменные: data, zerno

Модель: линейная  Y = a0+a1*x
  Коэфф.        a0        a1
Значение      -529     0.275
Ст.ошиб.      36.3    0.0185
 Значим.         0         0

Источник  Сум.квадр. Степ.св Средн.квадр.
Регресс.       574        1       574
Остаточн       111       43      2.59
     Вся       686       44

Множеств R      R^2  R^2прив  Ст.ошиб.        F   Значим
  0.91521   0.8376  0.83383    1.6094      222        0
Гипотеза1:<Регрессионная модель адекватна экспериментальн.данным>

   data=2000, zerno=21.274

    Хэксп     Yэксп     Yрегр  остаток Ст.остат  Ст.ошиб Довер.инт
   1.95E3       5.6      6.14   -0.543   -0.341     1.68      3.34
   1.95E3       4.6      6.42    -1.82    -1.14     1.67      3.33
   1.95E3       7.3      6.69    0.607    0.381     1.67      3.32
   1.95E3       6.7      6.97   -0.268   -0.169     1.66      3.31
   1.95E3       6.9      7.24   -0.344   -0.216     1.66      3.31
   1.95E3       7.9      7.52    0.381     0.24     1.66       3.3
                    .  .  .  .  .
```

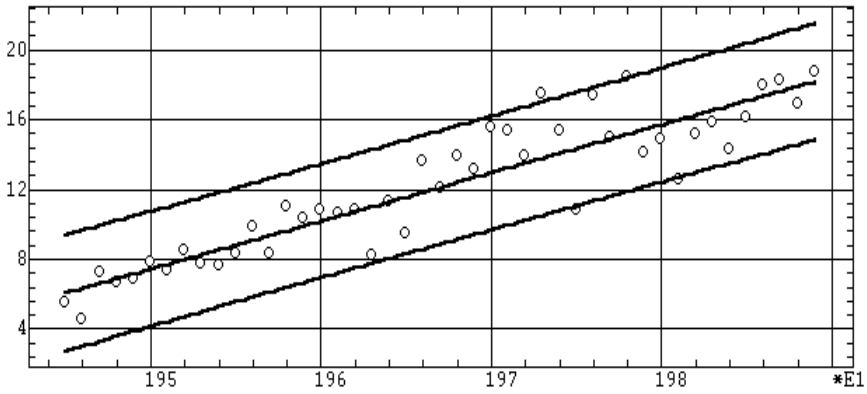
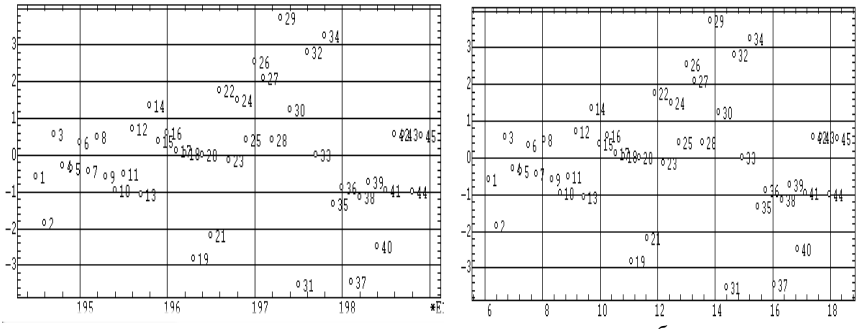


Рис. 10.21. Линейная регрессионная модель с зоной доверительного интервала

Обсуждение результатов. Как следует из числовых результатов, линейная модель адекватна экспериментальным данным (значимость нулевой гипотезы близка к нулю). На регрессионном графике (рис. 10.21) лишь незначительное число экспериментальных точек выходит за доверительный интервал, а распределение остатков (рис. 10.22) достаточно однородно.



а *б*

Рис. 10.22. Регрессионные остатки (по оси *Y):*
а — по годам; *б* — относительно регрессионных значений

Если сохранить остатки в электронной таблице, то при дальнейшем анализе корреляций и распределений легко выяснить, что их распределение не коррелировано как по *X*, так и по *Y* и нормально распределено, и это дополнительно подтверждает адекватность модели. Мы получили по построенной модели для 2000 г. прогноз урожайности более 21 центнера с га, однако можно попытаться построить и детальный погодовой прогноз (рис. 10.23).

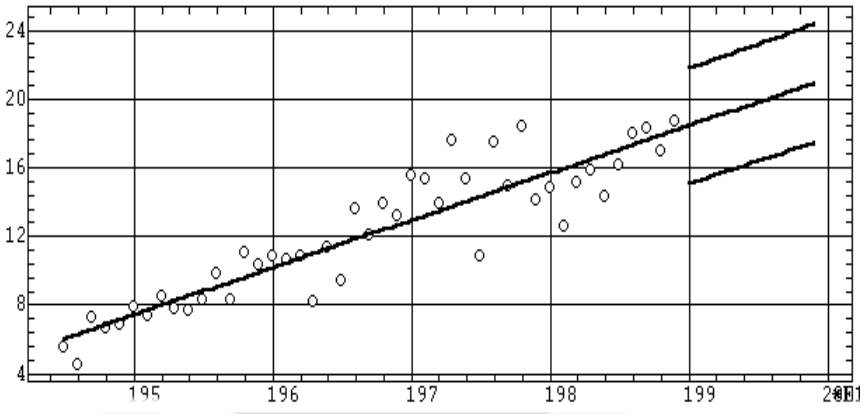


Рис. 10.23. Прогнозирование урожайности пшеницы с 1990 по 2000 гг. с зоной доверительного интервала прогноза индивидуальных значений

Выдача прогноза:

Хпрогн	Yпрогн	Ст.ошиб	Довер.инт
1.99E3	18.5	1.68	3.35
1.99E3	18.8	1.69	3.36
1.99E3	19.1	1.69	3.37
1.99E3	19.3	1.7	3.38
1.99E3	19.6	1.7	3.39
2E3	19.9	1.71	3.4
2E3	20.2	1.71	3.41
2E3	20.4	1.72	3.42
2E3	20.7	1.73	3.43

Обсуждение: Если сравнить среднюю урожайность зерновых в 1990—1996 гг. с нашим прогнозом, то выявленная линейная тенденция продолжает сохраняться, а ежегодные колебания вполне укладываются в доверительный интервал прогноза. Полученные результаты могут навести на крайне интересные геополитические выводы:

1) средняя тенденция увеличения урожайности дает некоторую надежду на оттягивание глобального голода и вымирания человечества;
2) отчаянные попытки «догнать и перегнать Америку» и внедрение кукурузы в 60-х годах, а также криминальный развал народного хозяйства в 90-х годах не смогли, к счастью, резко переломить общую урожайную тенденцию;
3) экспоненциальное увеличение научных знаний и технологий также не позволяет изменить стабильность коэффициента наклона, видимо, компенсируясь соразмерно прогрессирующей деградацией земель.

Продолжение анализа. Но не будем торопиться, а еще раз повнимательнее присмотримся к динамике роста урожайности (см. рис. 10.20). И зададимся вопросом: так ли она уж строго линейна, не видится ли в ней признаков замедления роста, особенно на последнем отрезке?

Для ответа на этот вопрос можно сделать следующее: разбить интервал наблюдения на четыре последовательных отрезка, подсчитать на каждом среднее значение и вычислить разность средних.

Для последовательных 11-летних периодов средствами описательной статистики мы получим следующие средние значения: 7,173, 10,47, 14,61, 15,93, а их разности 3,3, 2,14, 1,32 показывают очевидное замедление роста. Поэтому продолжим анализ с подбором более адекватной модели. После перебора возможных вариантов наиболее приемлемой следует признать параболическую модель.

Р е з у л ь т а т ы:

```
ПРОСТАЯ РЕГРЕССИЯ.  Файл: corn.std  Переменные: data, zerno
Модель: парабола  Y = a0+a1*x+a2*x^2
  Коэфф.       a0       a1       a2
Значение   -10445     10.36   -0.0025
Ст.ошиб.   6035.5     6.137    0.0015
 Значим.    0.087     0.095    0.104
Источник  Сум.квадр. Степ.св Средн.квадр.
Регресс.   581.19        2   290.59
Остаточн   104.65       42   2.4917
```

```
      Вся    685.84        44
Множеств R      R^2  R^2прив  Ст.ошиб.       F   Значим
  0.92055  0.84741  0.84014   1.5785    116.6        0
Гипотеза 1: <Регрессионная модель адекватна эксперимент.данным>
      data=2000, zerno=18.915
```

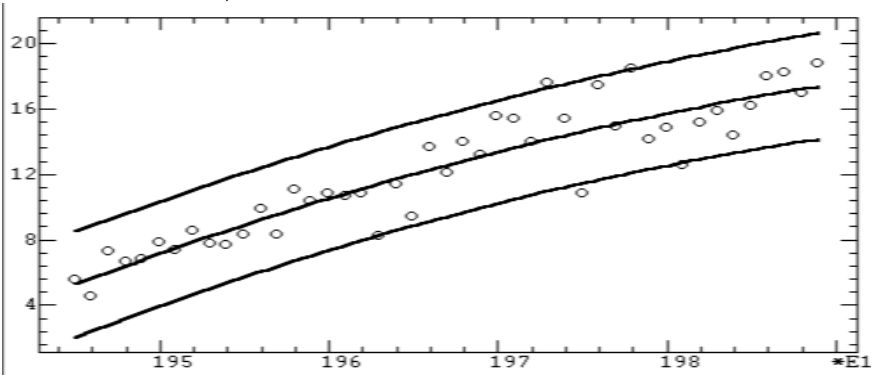


Рис. 10.24. Параболическая регрессионная модель с зоной доверительного интервала

Обсуждение результатов. Как легко заметить по итоговой статистике (более высокая множественная корреляция, меньше стандартная ошибка), параболическая модель является более адекватной для данного временного ряда. И эта модель в прогнозе дает заметно меньшую среднюю урожайность зерновых к 2000 году (19 вместо 21 ц/га).

Пример 2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1		Общая	Практикум						Коллоквиумы						Синтезы						Контрольные					Экзамен
2		сумма	1	2	3	4	5	Сум	1	2	3	4	5	Сум	1	2	3	4	5	Сум	1	2	3	4	Сум	
3	Авраменко	228.3	6	4	7	7	6	30	15	15	17	17	17	81	4	4	4	4	4	20	12.0	8.5	19.5	9.5	49.5	48.3
4	Батырев	130.1	6	3	7	5	4	25	14	14	13	13	12	66	4	4	4	4	4	20	6.5	9.0	2.5	0.5	18.5	0.6
5	Година	236.1	6	4	7	7	6	30	20	20	20	20	20	100	4	4	4	4	4	20	18.5	24.5	19.5	22.0	84.5	1.6
6	Ефимов Н.	18.8	6	4				10						0						0		2.5	2.0	3.5	8.0	0.8
7	Иоутси	195.0	6	4	7	7	6	30	15	16	18	20	20	89	4	4	4	4	4	20	8.0	8.0	20.5	18.0	54.5	1.5
8	Коваленко	300.3	6	4	7	7	6	30	19	19	19	20	20	97	4	4	4	4	4	20	20.0	22.0	22.5	15.0	79.5	73.8
9	Кружалова	300.0	6	4	7	7	6	30	18	19	19	20	20	96	4	4	4	4	4	20	23.0	19.0	25.0	12.5	79.5	74.5
10	Ломако	232.8	6	4	7	7	6	30	19	20	19	20	20	98	4	4	4	4	4	20	24.5	17.0	22.0	18.0	81.5	3.3
11	Меньшени	63.0	6	4				10	18					18	4					4	11.0	12.5	7.0		30.5	0.5
12	Морозов Д.	241.0	6	4	7	7	6	30	20	20	19	20	19	98	4	4	4	4	4	20	25.0	23.5	22.5	20.5	91.5	1.5
13	Новикова	293.2	6	4	7	7	6	30	19	19	19	19	20	96	4	4	4	4	4	20	21.0	14.5	23.0	10.0	68.5	78.7
14	Носкова	295.2	6	4	7	7	6	30	18	19	19	19	19	94	4	4	4	4	4	20	16.0	16.0	21.5	19.0	72.5	79.2
15	Попов М.В.	291.0	6	4	7	7	6	30	19	18	18	18	19	92	4	4	4	4	4	20	23.0	11.5	20.5	15.5	70.5	78.5
16	Простакова	237.5	6	4	7	7	6	30	19	20	20	20	18	97	4	4	4	4	4	20	22.0	24.0	25.0	15.5	86.5	4.0
17	Пурецкий	310.8	6	4	7	7	6	30	19	19	19	20	19	96	4	4	4	4	4	20	24.0	22.5	18.0	14.5	79.0	85.8
18	Рагузин	314.0	6	4	7	7	6	30	19	18	19	20	19	95	4	4	4	4	4	20	20.5	22.0	23.0	12.0	77.5	91.5
19	Ракша	222.5	6	4	7	7	6	30	17	16	15	16	15	79	4	4	4	4	4	20	17.0	15.0	13.0	9.0	54.0	39.5
20	Рассохина	286.5	6	4	7	7	6	30	18	19	19	20	19	95	4	4	4	4	4	20	18.5	18.0	22.5	17.0	76.0	65.5
21	Самохин	302.7	6	4	7	7	6	30	18	18	18	18	17	89	4	4	4	4	4	20	19.5	17.0	22.5	14.0	73.0	90.7
22	Смогоржев	185.5	6	4	7	7	6	30	16	17	18	17	18	86	4	4	4	4	4	20	8.0	12.5	14.5	13.0	48.0	1.5
23	Стрелков	224.3	6	4	7	7	6	30	18	16	17	16	18	85	4	4	4	4	4	20	7.0	5.5	15.0	9.0	36.5	52.8
24	Фролов	239.3	6	4	7	7	6	30	17	16	18	19	18	88	4	4	4	4	4	20	14.0	16.0	16.0	5.5	51.5	50.3
25	Чистяков	179.0	6	4	7	7	6	30	16	14	16	17	17	80	4	4	4	4	4	20	5.0	8.0	11.0	5.5	29.5	19.5
26	Шишин	235.6	6	4	7	7	6	30	19	20	19	20	20	98	4	4	4	4	4	20	23.5	22.5	20.0	20.3	86.3	1.3

Рис. 10.25. Таблица учета семестровой успеваемости студентов одной группы химфака МГУ

З а д а ч а. На химическом факультете МГУ учебная часть ведет детальный анализ успеваемости студентов в течение каждого семестра по следующим четырем разделам[1]: практикумы, коллоквиумы, синтезы, контрольные работы (рис. 10.25). В каждом разделе в течение семестра проводятся по 4–5 занятий и по каждому занятию для каждого студента выставляется бальная оценка. Затем оценки по каждому разделу суммируются, вычисляется общая сумма баллов, проводится экзамен, по которому выставляется 110-балльная оценка.

Определяющими для успеваемости студента по семестру являются: сумма баллов за контрольные работы (максимум 100 баллов) и общая сумма баллов (максимум 255 баллов).

Преподавательский опыт говорит, что существует достаточно строгая связь между эффективностью работы студентов в семестре и их оценками на экзамене, который с учетом этой связи является достаточной формальностью. Поэтому учебная часть решила разработать метод априорного выставления экзаменационных оценок студентам по результатам их семестровой работы. Если же студент не согласен с априорной оценкой, то он может попытаться ее улучшить, пойдя на экзамен. Это позволило бы: 1) разгрузить преподавателей от излишней экзаменационной работы; 2) избавить студентов от дополнительного стресса, связанного с подготовкой и сдачей экзаменов.

Визуальный анализ. Сначала надо выяснить, действительно ли существует предполагаемая зависимость. Лучше всего это будет видно на диаграмме рассеяния. Возьмем для примера зависимость общего балла и экзаменационного балла по 246 студентам (переменные *Экз* и *Сумма* из файла CHMSTR).

В ы в о д ы. С некоторым упрощением подходя к данной задаче будем рассуждать следующим образом. Как следует из диаграммы (рис. 10.26), в области ненулевых экзаменационных оценок (по оси ординат) наблюдается близкая к линейной зависимость между анализируемыми показателями. Однако в исходных данных явно присутствует инородная и довольно большая группа студентов, имеющих нулевые экзаменационные оценки.

Выяснить истинную причину такого положения можно только при более глубоком изучении реальной ситуации. То есть здесь необходим переход от формального анализа к предметному анализу. В результате выясняется, что в эту инородную группу входят четыре категории студентов: а) не сдавшие экзамен по причине плохих знаний (пропустившие много семестровых занятий (от 0 до 100 баллов по оси абсцисс); б) не сдавшие экзамен по причине экзаменационного волнения (от 100 до 200 баллов по

[1] Данные предоставлены профессором МГУ В.П. Зломановым

оси абсцисс); в) не пришедшие на экзамен; г) получившие «автомат» за отличную работу в семестре (от 200 до 250 баллов по оси абсцисс).

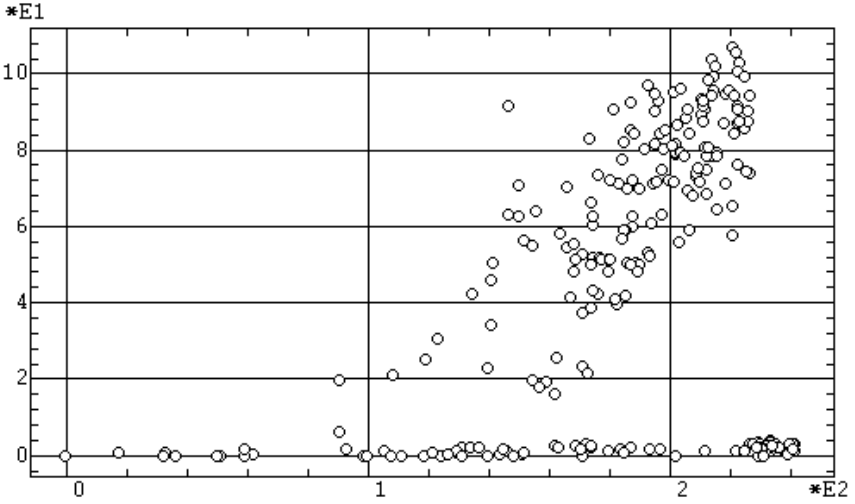


Рис. 10.26. Диаграмма рассеяния *экзаменационный балл = f(сумма баллов за семестр*)

Преобразование данных. Присутствие таких инородных данных может сильно исказить результаты анализа, поэтому необходимо их удалить. Проще всего это сделать, выполнив операцию *сортировки* из *Блока преобразований* (разд. 3.4) для переменных *Экз* и *Сумма* по возрастанию значений переменной *Экз*, после чего удалить из матрицы данных все строки, содержащие малые значения переменной *Экз*. В результате останутся данные по 156 студентам (переменные *Экз1* и *Сумма1* из файла CHMSTR).

А н а л и з. Рассчитаем для рафинированных таким образом данных регрессионную модель, по которой вычислим экзаменационные оценки для значений 150 и 200 суммы экзаменационных баллов:

```
ПРОСТАЯ РЕГРЕССИЯ.  Файл: chmstr.std Переменные: сумма1, экз1

Модель: линейная  Y = a0+a1*x
  Коэфф.        a0        a1
Значение    -49.83    0.6207
Ст.ошиб.     8.434   0.04386
 Значим. 4.511E-6 3.98E-10
Источник  Сум.квадр. Степ.св Средн.квадр.
Регресс.  4.438E4         1  4.438E4
Остаточн  3.412E4       154    221.6
     Вся   7.85E4       155
Множеств R     R^2  R^2прив  Ст.ошиб.       F   Значим
  0.75189  0.56534  0.56252    14.885   200.3 2.45E-16
  Гипотеза  1:  <Регрессионная  модель  адекватна  эксперимен-
таль.данным>
     сумма1=150, Y=43.27
     сумма1=200, Y=74.31
```

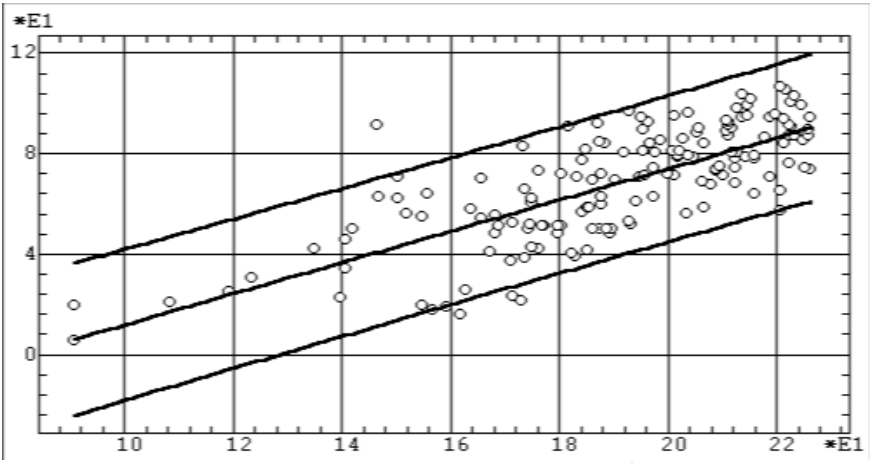


Рис. 10.27. Линейная регрессия *экзаменационный балл = f(сумма баллов за семестр)* с доверительным интервалом

Выводы:

1. Линейная регрессионная модель адекватна экспериментальным данным на пренебрежительно малом уровне значимости нулевой гипотезы.
2. Модель позволяет рассчитать априорные экзаменационные оценки для студентов по сумме их семестровых данных. Так при сумме баллов 200 студенту может быть предложена экзаменационная оценка в 74 балла, что в пересчете на 5-бальную систему от максимума в 110 баллов составляет 74/110·5=3,364.

Анализ остатков. Однако такое предложение (вывод 2) будет статистически неполноценным, поскольку кроме регрессионной прямой надо учитывать еще и статистический разброс оценок. Поэтому надо продолжить регрессионный анализ: 1) выполнить анализ регрессионных остатков с построением их графика (рис. 10.28); 2) сохранить остатки с графика в электронной таблице нажатием на инструментальную кнопку *«СохрГраф»*; 3) вычислить описательную статистику для остатков; 4) при вычислении априорной экзаменационной оценки корректировать ее с учетом вычисленного стандартного отклонения регрессионных остатков.

Результаты:

```
ОПИСАТЕЛЬНАЯ СТАТИСТИКА.  Файл: chmstr.std
Перемен. Размер ←Диапазон-> Среднее-Ошибка Дисперс Ст.откл  Сумма
      x8    156  -36.7 50.2  -1.3E-8 1.18   220.1   14.84   -2E-6
```

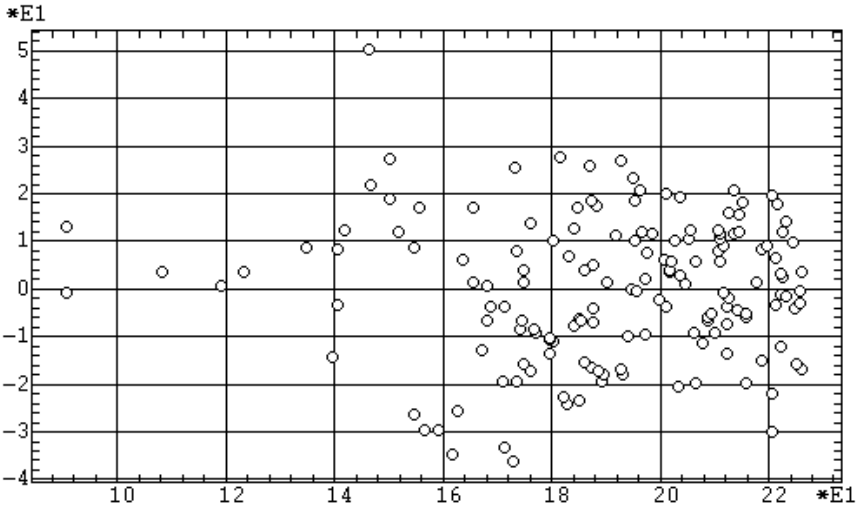


Рис. 10.29. Диаграмма рассеяния регрессионных остатков по значениям независимой переменной *Сум*

В ы в о д ы: Стандартное отклонение регрессионных остатков в пересчете на 5-балльную систему составляет 14,84/110·5=0,6745. Поэтому при суммарном балле, равном 200, экзаменационная оценка может варьироваться от 3,364–0,6745 до 3,364+0,6745 (с учетом того, что диапазон одного стандартного отклонения включает 68% популяции). При выборе из этого диапазона следует учитывать действие случайных факторов, ведущих к снижению оценки на экзамене: влияние стресса, неудачного вопроса в билете, усталость преподавателя и т. п. Поэтому представляется разумным предлагать студентам оценки из верхней части указанного диапазона.

Обратная задача. Кроме вышерассмотренной для преподавателя актуальна и обратная задача: иметь таблицу, в которой для проходных экзаменационных баллов были бы даны соответствующие им диапазоны суммарных баллов. Для этого нужно по вышерассмотренной схеме произвести обратный регрессионный анализ для зависимости суммы баллов от экзаменационной оценки. При этом вычислим суммы баллов для оценок 5 (110 баллов), 4 (88 баллов), 3 (66 баллов), 2 (44 балла) и оценки 74 балла для сравнения с ранее полученными результатами.

Р е з у л ь т а т ы:

```
ПРОСТАЯ РЕГРЕССИЯ.  Файл: chmstr.std Переменные: экз1, сумма1

Модель: линейная  Y = a0+a1*x
  Коэфф.        a0         a1
Значение     128.1    0.9108
Ст.ошиб.     4.629   0.06436
```

```
 Значим. 1.01E-12 3.98E-10

Источник  Сум.квадр. Степ.св Средн.квадр.
Регресс.  6.513E4        1  6.513E4
Остаточн  5.007E4      154    325.1
     Вся  1.152E5      155

Множеств R     R^2  R^2прив  Ст.ошиб.       F   Значим
  0.75189  0.56534  0.56252   18.032    200.3 2.45E-16
Гипотеза 1: <Регрессионная модель адекватна эксперимент.данным>

     экз1=110, сумма1=228.3
     экз1=88, сумма1=208.3
     экз1=66, сумма1=188.2
     экз1=44, сумма1=168.1
     экз1=74, сумма1=195.5
```

Выводы:

1. Линейная регрессионная модель адекватна имеющимся данным.
2. Определены средние значения суммы баллов для основных экзаменационных оценок: 228, 208, 188, 168, 195.
3. Полученное значение 195,5 суммы баллов для экзаменационной оценки 74 немного не совпадает со значением 200, использованным выше в прямой регрессии (расхождение составляет менее 2,5%). Это является следствием несовпадения прямой и обратной регрессии, обсуждаемого во введении к данной главе.
4. Если продолжить анализ регрессионных остатков по вышерассмотренной схеме, то можно определить стандартное отклонение регрессии, равное 17,97.

Направления дальнейшего анализа. Имеющиеся данные предоставляют богатейший материал для разнообразных дальнейших исследований. Сформулируем только одну из интересных в этом плане задач.

Различные группы студентов ведут разные преподаватели. У каждого из них имеются свои правила преподавания и выставления оценок, что сказывается как на знаниях студентов, так и на полученных ими баллах. Сделав допущение, что группы студентов примерно равны по средним способностям, можно для каждого преподавателя вычислить средний бал выставляемых ими оценок, а также выявить их попарные статистические различия и группировки. Зная это, можно оценить качество преподавания предмета различными преподавателями, сравнивая баллы, выставленные их студентам другими преподавателями. Все это может быть плодотворным материалом для организационных и методических выводов по коррекции учебного процесса с целью повышения уровня преподавания, а также ликвидации систематического завышения или занижения оценок.

10.4. Множественная линейная регрессия

Исходные данные представляются в виде матрицы, содержащей одну или несколько независимых переменных *X* и зависимую переменную *Y*.

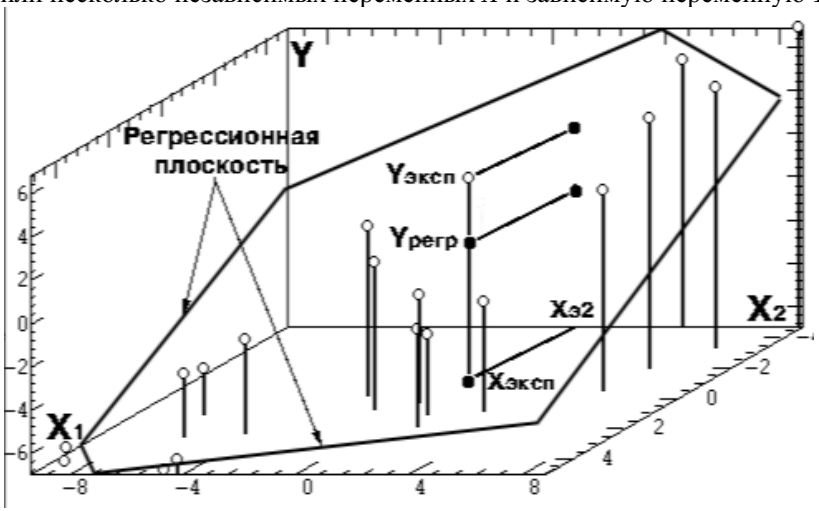


Рис. 10.30. Компоненты 2-параметрической линейной регрессии

Действия и результаты. Сначала из электронной таблицы надо выбрать независимую и зависимые переменные для анализа (см. бланк рис. 10.5). Последовательность результатов отличается от однопараметрического случая (см. разд. 10.1) следующим:

1. Для интерполяции (прогнозирования) новых значений *Y* по модели необходимо ввести не одно значение *X*, а последовательно координаты всех независимых переменных (см. бланк рис. 10.2).
2. Интервальное прогнозирование (подобно простой регрессии) не производится, его заменяет интерполяция п.1.
3. Многомерная гиперплоскость не допускает прямого наглядного графического представления (аналогично графику регрессии для однопараметрического случая), его заменяют графики проекций регрессионных значений на избранные плоскости исходных переменных п.4. *Y*эксп, *Y*регр, *Y*-эксп–*Y*регр, *X*эксп.
4. Продолжение анализа включает следующие возможности (меню рис. 10.31):

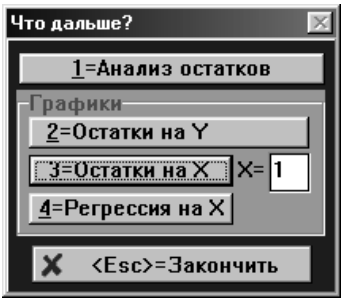


Рис. 10.31. Меню дальнейшего анализа множественной регрессии

- числовой анализ регрессионных остатков (аналогично разд. 10.1);
- график регрессионных остатков $Y_{эксп}–Y_{регр}$ соответственно экспериментальным значениям $Y_{эксп}$ зависимой переменной Y (аналогично разд. 10.1);
- график регрессионных остатков от значений $X_{эi}$ независимой переменной X_i, порядковый номер которой i указан в поле ввода;
- график регрессионных значений $Y_{регр}$ для экспериментальных значений $X_{эi}$ независимой переменной X_i (порядковый номер i которой указан в поле ввода рис. 10.5), т. е. график проекции регрессионных значений на плоскость $Y–X_i$.

Ограничение. Метод неприменим, если $m>30$ или $n<m+1$ и $n(m+1)>l$, где l — объем матрицы данных 64 000, 20 000, 4000 и 400 чисел.

Пример

Задача. Мировая статистика накопила богатейший материал по множеству экономических и социальных показателей развития различных стран. Эти данные, в частности, могут быть полезны для исследования разного рода зависимостей и прогнозирования. Многие из таких зависимостей носят линейный характер, или же могут быть приближены к таковому посредством простых алгебраических преобразований исходных показателей.

Однако линейное моделирование осмыслено не для всего множества показателей, а только для тех из них, которые взаимосвязаны. Поэтому на первом этапе такого исследования необходимо вычислить кросс-корреляционную матрицу и отобрать для дальнейшего анализа группы показателей, взаимно связанных большими корреляциями.

На таком предварительном этапе для данного примера из данных по 108 государствам мы отобрали следующие взаимосвязанные показатели[1] (файл MLR): средняя продолжительность жизни, детская смертность, потребление калорий, уровень рождаемости, уровень смертности, количество детей, взаимно связанными высокими корреляциями:

```
ПАРАМЕТРИЧЕССКАЯ КОРРЕЛЯЦИЯ.  Файл: mlr.std
          Корреляционная матрица
           ДЛ.ЖИЗНИ ДЕТСМЕРТ  КАЛОРИИ  УР.РОЖД УР.СМЕРТ
 ДЕТСМЕРТ   -0.946
  КАЛОРИИ    0.765   -0.777
  УР.РОЖД   -0.829    0.867   -0.762
 УР.СМЕРТ   -0.756    0.646   -0.355    0.432
 КОЛДЕТЕЙ   -0.809    0.845   -0.696    0.974     0.48
```

[1] Данные из архива SPSS.

```
Критическое значение=0.224
Число значимых коэффициентов=15 (100%))
```

Выявление изменений в средней продолжительности жизни требует достаточно длительного времени наблюдения. Изменения же в других показателях могут быть отслежены значительно быстрее. Поэтому актуальной является задача предсказания изменения продолжительности жизни при изменении других, связанных с ней показателей. На первом этапе такого исследования предположим, что продолжительность жизни линейно зависит от других социальных показателей, и для решения поставленной задачи рассчитаем линейную многопараметрическую регрессионную модель. После этого осуществим предсказание длительности жизни для избранного набора значений других показателей.

Р е з у л ь т а т ы:

```
МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ.  Файл: mlr.std

Y=ДЛ.ЖИЗНИ          ДЕТСМЕРТ  КАЛОРИИ  УР.РОЖД УР.СМЕРТ КОЛДЕТЕЙ
  Коэфф.       a0        a1        a2        a3        a4        a5
Значение     74.9   -0.123  0.00247   -0.332   -0.787      1.3
Ст.ошиб.     3.19   0.0194 0.000846    0.121   0.0914     0.68
 Значим. 4.68E-11  3.79E-6  0.00496  0.00787  1.77E-7   0,0564

Источник  Сум.квадр. Степ.св Средн.квадр.
Регресс.    7.19E3        5   1.44E3
Остаточн       375       69     5.43
     Вся    7.56E3       74

Множеств R     R^2  R^2прив  Ст.ошиб.       F   Значим
  0.97491  0.95046  0.94687    2.3305     265 4.74E-12
Гипотеза 1: <Регрессионная модель адекватна эксперименталь.данным>

ДЕТСМЕРТ=20, КАЛОРИИ=2E3, УР.РОЖД=10, УР.СМЕРТ=5, КОЛДЕТЕЙ=3
ДЛ.ЖИЗНИ=74
```

Обсуждение: Как можно заметить, построенная линейная модель адекватна экспериментальным данным (уровень значимости гипотезы о равенстве нулю коэффициента множественной корреляции близок к нулю). Однако, рассматривая значимости регрессионных коэффициентов при различных показателях, можно заметить, что количеством детей в дальнейшем можно пренебречь, поскольку значимость гипотезы о равенстве его коэффициента *a5* нулю (*P=0,0564*) выше критического уровня. Поэтому повторим анализ с исключением этого показателя.

Р е з у л ь т а т ы:

```
МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ.  Файл: mlr.std

Y=ДЛ.ЖИЗНИ          ДЕТСМЕРТ  КАЛОРИИ  УР.РОЖД УР.СМЕРТ
  Коэфф.       a0        a1        a2        a3        a4
Значение     72.3   -0.126  0.00292   -0.118   -0.734
Ст.ошиб.     2.94   0.0197 0.000828   0.0481   0.0889
 Значим. 3.33E-11  3.33E-6  0.00109   0.0161  2.59E-7

Источник  Сум.квадр. Степ.св Средн.квадр.
Регресс.    7.17E3        4   1.79E3
Остаточн       395       70     5.64
     Вся    7.56E3       74
```

```
Множеств R      R^2  R^2прив  Ст.ошиб.       F   Значим
  0.97356  0.94782  0.94484    2.3745      318 4.43E-12
Гипотеза 1: <Регрессионная модель адекватна эксперименталь.данным>
ДЕТСМЕРТ=20, КАЛОРИИ=2Е3, УР.РОЖД=10, УР.СМЕРТ=5, ДЛ.ЖИЗНИ=70.7
```

Обсуждение: Уточненная модель осталась практически на том же уровне адекватности (при некотором увеличении значения *F*-статистики), но можно заметить, что при тех же значениях параметров она предсказывает несколько меньшую продолжительность жизни (70,7 лет вместо 74). Теперь можно произвести анализ регрессионных остатков и построение результирующих графиков.

Хэксп	Yэксп	Yрегр	остаток	Ст.остат	Ст.ошиб	Довер.инт
7.3	74	73.1	0.902	0.391	2.36	4.64
6.7	73	72.1	0.862	0.373	2.36	4.64
7.2	74	70.9	3.05	1.32	2.36	4.64
25.6	68	69.2	-1.16	-0.503	2.35	4.62
106	53	52.6	0.424	0.184	2.38	4.68
75	59	57.8	1.22	0.53	2.35	4.62
39.3	60	64.6	-4.6	-1.99	2.34	4.61
66	57	62.9	-5.88	-2.55	2.34	4.62
118	47	45.3	1.71	0.742	2.39	4.71
105	46	44	1.96	0.848	2.38	4.68

. . .

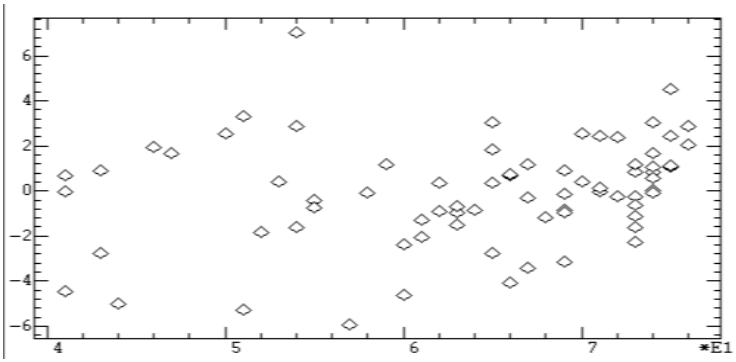
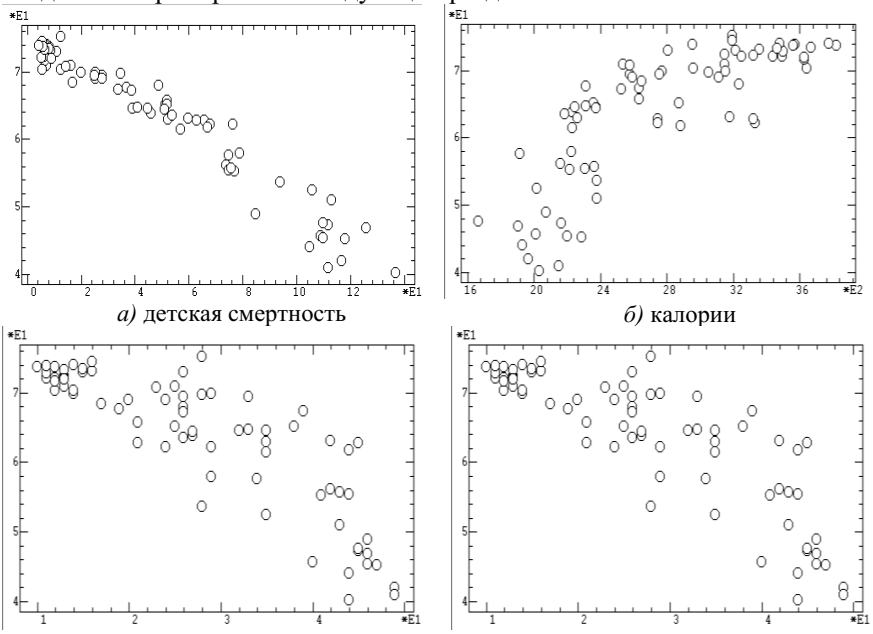


Рис. 10.32. Регрессионные остатки (по оси *Y*) относительно регрессионных значений

Обсуждение: Визуальное изучение распределения регрессионных остатков (рис. 10.32) не выявляет какой-либо зависимости между ними. Заметна только более плотная концентрация в области больших длительностей жизни. Это говорит о том, что страны с меньшей длительностью жизни более сильно различаются между собой.

Изучение регрессионных графиков показывает следующее: 1) продолжительность жизни практически линейно связана с детской смертностью (рис. 10.33, *а*), что вполне соответствует смысловой взаимосвязи этих показателей; 2) хорошая линейная связь прослеживается и с уровнем смертности (рис. 10.33, *г*); 3) связи с уровнем рождаемости и особенно с потреблением калорий носят нелинейный характер, поэтому для большей адекватности в отношение прогноза возможно будет оправдано введение в модель

нелинейных членов по этим показателям (см. разд. 10.6). Продолжение анализа данного примера см. в следующем разделе.



а) детская смертность *б)* калории

в) уровень рождаемости *г)* уровень смертности

Рис. 10.33. Распределения регрессионных значений (по оси *Y*) относительно исходных показателей

10.5. Пошаговая регрессия

Назначение. Пошаговая регрессия в рамках линейной многопараметрической модели (см. разд. 10.4) позволяет из множества исходных переменных производить отбор тех независимых переменных, которые наиболее значимы для адекватного представления исходных данных. Тем самым этот метод позволяет, во–первых, построить более простую, сокращенную модель, а, во–вторых, в последующем сборе данных не регистрировать значения несущественных переменных.

Исходные данные представляются в виде, аналогичном методу множественной регрессии (см. разд. 10.4).

Действия и результаты. Сначала необходимо выбрать из электронной таблицы независимую и зависимые переменные (см. бланк рис. 10.6), а также установить значения параметров пошаговой регрессии (рис. 10.34):

Рис. 10.34. Бланк установок пошаговой регрессии

1) признак критерия отбора: по *P*– или *F*–уровню;
2) *P*– или *F*–уровень критерия отбора;
3) уровень толерантности *T* (только для метода включения).

После этого нужно выбрать метод отбора пере-

менных (нажатие соответствующей кнопки).

В начале анализа для каждой переменной вычисляется среднее и стандартное отклонение, а также матрица корреляций/ковариаций между переменными.

На каждом шаге включения указывается наименование введенной переменной и выполняется стандартная выдача множественной линейной регрессии (см. разд. 10.4), которая дополняется изменением квадрата коэффициента множественной корреляции (R^2) и значениями *F* и *P* для нулевой гипотезы «*изменение* $R^2 = 0$».

Далее для всех включенных в модель переменных выдаются значения:

- регрессионного коэффициента *B*;
- стандартной ошибки вычисления коэффициента *B*;
- стандартизированного регрессионного коэффициента бета (получаемого посредством умножения *B* на отношение стандартных отклонений S_x/S_y);
- значений *F* и *P* для нулевой гипотезы о равенстве коэффициента *B* нулю.

Затем для всех переменных, не включенных в модель, выдаются значения *B*, ошибки *B*, *бета*, *F* и *P* для нулевой гипотезы изменения *R*, частичного коэффициента корреляции и толерантности *T*.

На каждом шаге удаления указывается наименование удаленной переменной и выполняется выдача, аналогичная методу включения за исключением статистики по переменным вне регрессионной модели.

Заключительная выдача результатов и диалог имеют стандартный вид (см. разд. 10.1) для случая многопараметрической модели.

Ограничения:

1. Метод неприменим, если $m > 43$ или $n < m+1$.
2. Размер выборок не может превосходить $l/(m+3)$, где *l* — объем матрицы данных в 64000, 20000, 4000 или 400 чисел.

Примеры

Задача. Мы продолжим анализ данных мировой социальной статистики из примера к разд. 10.4 с целью определить методом шаговой регрессии необходимый и достаточный набор переменных, объясняющих показатель средней продолжительности жизни.

Результаты:

```
ПОШАГОВАЯ РЕГРЕССИЯ.  Файл: mlr.std

*** Метод включения. Шаг No.1, введена переменная:ДЕТСМЕРТ
  Коэфф.       a0         a1
Значение      75.5     -0.247
Ст.ошиб.     0.603    0.00992
 Значим. 9.65E-14    2.6E-11

Источник  Сум.квадр. Степ.св Средн.квадр.
Регресс.    6.77E3        1     6.77E3
Остаточн       797       73       10.9
     Вся    7.56E3       74

Множеств R      R^2  R^2прив  Ст.ошиб.       F   Значим
  0.94583   0.8946  0.89316    3.3047      620 3.24E-14
Гипотеза 1: <Регрессионная модель адекватна эксперименталь.данным>
Измен.R^2         F   Значим
    0.895       620  2.6E-11

--------------  Переменные в уравнении ---------------
 Переменн. Коэфф.B  Ст.ош.B     Бета         F   Значим
 ДЕТСМЕРТ    -0.247  0.00992   -0.946       620  2.6E-11
----------------- Переменные не в уравнении ---------------------
 Переменн. Коэфф.B  Ст.ош.B     Бета       F Значим  Частн.R Толер.
  КАЛОРИИ  0.00136  0.00107   0.0763    1.61  0.206    0.148   0.396
  УР.РОЖД  -0.0308   0.0621   -0.038   0.246 0.627    0,0584  0.249
 УР.СМЕРТ   -0.568   0.0924   -0.249    37.8  4.9E-6   0.587   0.583
 КОЛДЕТЕЙ   -0.178    0.371   -0.034    0.23  0.639    0,0564  0.287

    *** Метод включения. Шаг No.2, введена переменная:УР.СМЕРТ
  Коэфф.       a0         a1         a2
Значение      79.1     -0.205     -0.568
Ст.ошиб.     0.768     0.0106     0.0924
 Значим. 1.45E-13   1.5E-10   4.95E-6

                              ...

    *** Метод включения. Шаг No.3, введена переменная:КАЛОРИИ
  Коэфф.       a0         a1         a2         a3
Значение      68.9     -0.159     -0.675   0.00332
Ст.ошиб.      2.68     0.0151     0.0885  0.000839
 Значим. 2.39E-11    2.4E-8    5.6E-7  0.000384

                              ...

    *** Метод включения. Шаг No.4, введена переменная:УР.РОЖД
  Коэфф.       a0         a1         a2         a3         a4
Значение      72.3     -0.126     -0.734   0.00292     -0.118
```

```
Ст.ошиб.      2.94   0.0197   0.0889 0.000828   0.0481
 Значим. 3.33E-11  3.33E-6  2.59E-7  0.00109   0.0161

Источник  Сум.квадр. Степ.св Средн.квадр.
Регресс.    7.17E3        4   1.79E3
Остаточн      395       70     5.64
     Вся    7.56E3       74

Множеств R     R^2  R^2прив  Ст.ошиб.       F   Значим
  0.97356  0.94782  0.94484    2.3745     318 4.43E-12
Гипотеза 1: <Регрессионная модель адекватна эксперименталь.данным>
Измен.R^2         F   Значим
  0.00446      5.99   0.0161

--------------  Переменные в уравнении ---------------
 Переменн. Коэфф.B  Ст.ош.B     Бета         F   Значим
 ДЕТСМЕРТ    -0.126   0.0197   -0.484        41  3.33E-6
 УР.СМЕРТ    -0.734   0.0889   -0.323      68.3  2.59E-7
  КАЛОРИИ  0.00292 0.000828    0.164      12.4  0.00109
  УР.РОЖД    -0.118   0.0481   -0.145      5.99   0.0161
------------------ Переменные не в уравнении --------------------
 Переменн. Коэфф.B Ст.ош.B  Бета     F   Значим  Частн.R    Толер.
 КОЛДЕТЕЙ      1.3    0.68  0.25  3.67   0,0564    0.225    0.042
```

Обсуждение: На последнем шаге процедуры формируется модель, включающая четыре переменные за исключением количества детей. Это подтверждает вывод, сделанный в примере к разд. 10.4, что этот показатель незначим для предсказания.

Отметим, что при пошаговой регрессии сравнительную значимость объясняющих переменных представляется возможным количественно оценить толерантностью, определяющей последовательность включения в модель.

Продолжение анализа данного примера см. в следующем разделе.

10.6. Общая регрессия

Назначение. Данная процедура позволяет строить произвольную регрессионную модель, задаваемую некоторой алгебраической формулой, которая может быть нелинейной как по переменным, так и по параметрам. Для расчета модели используется итерационный алгоритм минимизации *наискорейшего спуска*.

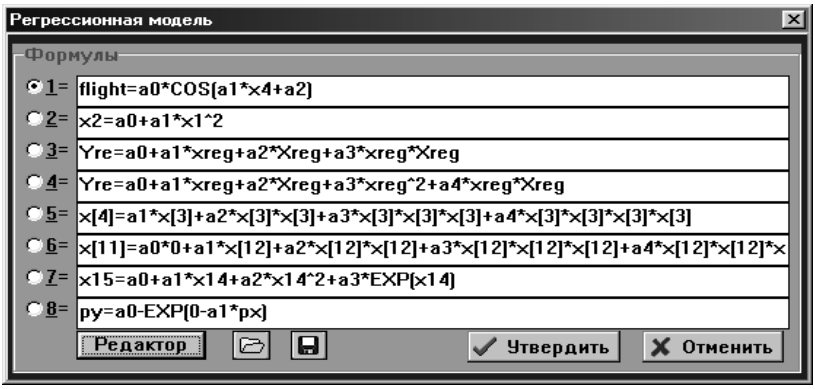


Рис. 10.35. Бланк ввода регрессионных моделей

Исходные данные представляются в виде матрицы, содержащей одну или несколько независимых переменных *X* и зависимую переменную *Y*.

Действия и результаты. Сначала нужно выбрать или ввести заново требуемую формулу модели в одну из восьми позиций (бланк — рис. 10.35, см. также разд. 2.3).

Редактор моделей Ввод новых формул можно производить непосредственно в этот бланк, однако часто удобно пользоваться формульным редактором, вызываемым по нажатию кнопки «*Редактор*», после чего появляется бланк редактора регрессионных моделей (рис. 10.36).

Работа этого бланка аналогична типовому формульному редактору (см. разд. 2.3) за следующими добавлениями:

- кроме поля формулы имеется поле зависимой переменной со своей кнопкой переноса из списка переменных электронной таблицы;
- в правой части бланка размещен список обозначений коэффициентов регрессионной модели: *a*0, *a*1, *a*2,...;
- при окончательном формировании модели независимая переменная приписывается справа к самой формуле с разделением знаком *равенство*.

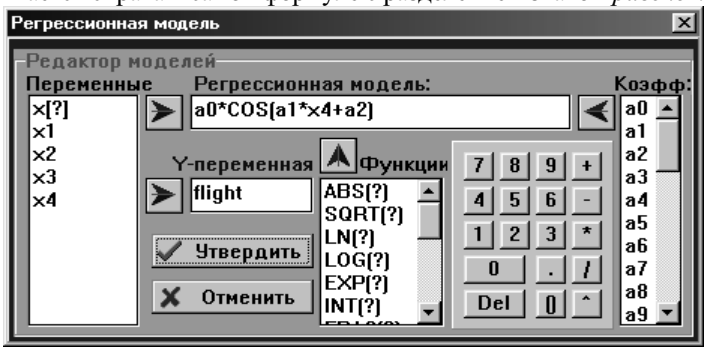


Рис. 10.36. Бланк редактора регрессионной модели

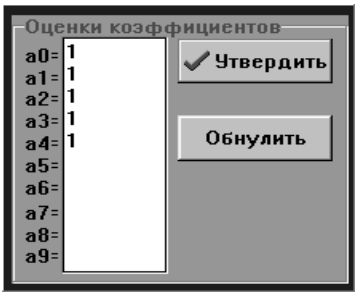


Рис. 10.37. Бланк ввода начальных оценок коэффициентов нелинейной модели

Начальные оценки. Поскольку сходимость алгоритма и правильность решения в ряде случаев может существенно зависеть от начальных значений регрессионных коэффициентов, то желательно ввести их начальные оценки (бланк — рис. 10.37),

Дело в том, что сложные нелинейные модели на фоне пологих минимумов могут иметь и более глубокие минимумы в виде узких дыр и щелей. Алгоритм минимизации при неудачном выборе начальных условий может не заметить этих дыр и скатиться в один из локальных минимумов или же привести к переполнению разрядной сетки промежуточных вычислений.

Если в этом бланке нажать на кнопку «*Обнулить*», то начальные оценки параметров будут установлены равными единице.

Выдача результатов процедуры общей регрессии стандартная (см. разд. 10.1) с учетом поправки для множественной регрессии (см. разд. 10.4). В случае однопараметрической регрессии последовательность графической выдачи аналогична разд. 10.3.

сходимости процесса неадекватности модели

2. Метод неприменим при $m>30$ или $m>n$, где n — число измерений, m — число регрессионных коэффициентов, или $l/(m+3)>n$, где l — объем матрицы данных в 64000, 20000, 4000 или 400 чисел.

Пример

Задача. Мы продолжим анализ данных мировой социальной статистики из примера к разд. 10.4. Там мы обнаружили нелинейный характер зависимости средней продолжительности жизни от количества потребляемых калорий и уровня рождаемости (рис 10.26). Попробуем учесть эти нелинейности введением в модель двух квадратичных членов. Формула нелинейной модели, вводимая в бланк формул, будет выглядеть следующим образом:

```
a0+a1*ДЕТСМЕРТ+a2*КАЛОРИИ+a3*КАЛОРИИ^2+a4*УР.РОЖД+a
5*УР.РОЖД^2+a6*УР.СМЕРТ
```

Как и в примере раздела 10.4 вычислим прогноз при тех же значениях параметров модели.

Результаты:

```
ОБЩАЯ (+НЕЛИНЕЙНАЯ) РЕГРЕССИЯ.  Файл: mlr.std

  Модель:
ДЛ.ЖИЗНИ=a0+a1*ДЕТСМЕРТ+a2*КАЛОРИИ+a3*КАЛОРИИ^2+a4*УР.РОЖД+a5*УР.Р
ОЖД^2+a6*УР.СМЕРТ
  Коэфф.      a0       a1       a2       a3       a4       a5       a6
Значение      88     -0.1  0.0035 -3.7E-7    -1.04   0.0151    -1.13
Ст.ошиб.    9.42   0.0194 0.00623  1.1E-6    0.236  0.00377    0.128
 Значим. 8.2E-8 2.95E-5   0.583   0.737  0.00014 0.000352  1.46E-7

Источник  Сум.квадр. Степ.св Средн.квадр.
Регресс.    7.25E3        6   1.21E3
Остаточн       318       68     4.67
     Вся    7.56E3       74
Множеств R      R^2  R^2прив  Ст.ошиб.       F   Значим
  0.97876  0.95797  0.95426    2.1621     258 3.99E-12
Гипотеза 1: <Регрессионная модель адекватна эксперименталь.данным>
ДЕТСМЕРТ=20, КАЛОРИИ=2E3,  УР.РОЖД=10,  УР.СМЕРТ=5, ДЛ.ЖИЗНИ =76.9
```

Выводы: Модель адекватна исходным данным, однако коэффициенты *a*2, *a*3 при показателе потребления калорий не отличны от нуля, поэтому скорректируем модель, удалив наиболее недостоверный квадратичный член, и повторим анализ.

Результаты:

```
ОБЩАЯ (+НЕЛИНЕЙНАЯ) РЕГРЕССИЯ.  Файл: mlr.std

  Модель:
ДЛ.ЖИЗНИ=a0+a1*ДЕТСМЕРТ+a2*КАЛОРИИ+a3*УР.РОЖД+a4*УР.РОЖД^2+a5*УР.С
МЕРТ
  Коэфф.       a0       a1       a2       a3       a4       a5
Значение     90.6   -0.101  0.00142    -1.02   0.0148    -1.13
Ст.ошиб.     5.25   0.0189 0.000836    0.228  0.00365    0.127
 Значим. 3.99E-10  2.03E-5   0.0896 0.000116 0.000309  1.25E-7

Источник  Сум.квадр. Степ.св Средн.квадр.
Регресс.    7.25E3        5   1.45E3
Остаточн       318       69     4.61
     Вся    7.56E3       74
Множеств R      R^2  R^2прив  Ст.ошиб.       F   Значим
  0.97873   0.9579  0.95485    2.1482     314 3.29E-12
Гипотеза 1: <Регрессионная модель адекватна эксперименталь.данным>
```

```
ДЕТСМЕРТ=20,  КАЛОРИИ=2E3,  УР.РОЖД=10, УР.СМЕРТ=5, ДЛ.ЖИЗНИ=77
```

В ы в о д ы: Как можно заметить, значимость гипотезы о равенстве нулю коэффициента при потреблении калорий при исключении квадратичного члена резко снизилась до уровня, близкого к критическому (0,0896 против 0,583). Можно сделать вывод, что нелинейность по этому показателю определяется зависимостью ее от уровня смертности: чем более калорийно питание, тем меньше смертность. Почему возможна такая связанная нелинейная зависимость по двум показателям?

Представим себе мысленно более простой и наглядный пример: двумерную регрессионную плоскость в трехмерном пространстве y–$x1$–$x2$, причем эта плоскость линейна по координате $x1$, но изогнута по координате $x2$. Далее представим, что на этой плоскости рассыпаны наши экспериментальные точки, однако эти точки рассыпаны не хаотично, а группируются вдоль биссектрисы угла $x1$–$x2$. Тогда проекции точек как на координатную плоскость y–$x1$, так и на координатную плоскость y–$x2$ будут изогнуты, несмотря на то, что сама регрессионная плоскость изогнута только по одной координате.

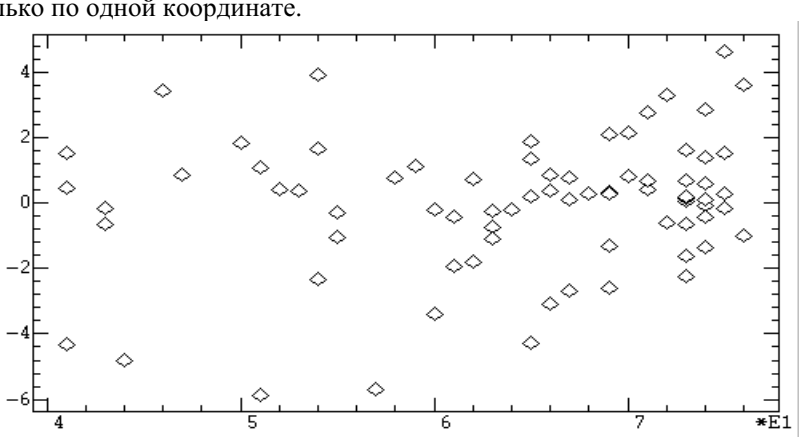


Рис. 10.38. Регрессионные остатки относительно регрессионных значений по оси *X*

Проведя далее анализ регрессионных остатков и построив график их распределения (рис. 10.38) можно заметить уменьшение диапазона разброса остатков (по оси ординат) по сравнению с линейной моделью (см. рис. 10.32). Это является дополнительным свидетельством большей адекватности нелинейной модели.

Учитывая близость к критической значимости коэффициента $a2$ можно оставить его в модели и использовать эту окончательную модель для прогнозирования средней продолжительности жизни. При введенных параметрах прогноз дает нам длительность жизни в 77 лет.

Глава 11

МНОГОМЕРНЫЕ МЕТОДЫ

«Знаешь ли ты, что можешь распространить себя везде, в любом направлении, которое я указал?»
[Жуан Матус]

Назначение. Многомерные методы предоставляют вычислительные и графические средства для исследования различных форм ассоциации (сходства, близости, группировки) данных, представленных в виде множества *переменных*, значения которых относятся к некоторому множеству *объектов*, *наблюдений* или *измерений*. При этом все переменные рассматриваются как равноправные, без разделения на зависимые и независимые (в отличие от методов дисперсионного и регрессионного анализа, некоторые из которых также работают с многомерными данными).

11.1. Факторный анализ

Введение. Факторный анализ[1] является наиболее значимым и часто используемым из многомерных методов, а методы кластеризации и дискриминантного анализа дают более наглядные результаты, если они применяются к данным, уже прошедшим процедуру факторизации. Однако, в большинстве учебников этот раздел излагается непредметно, несистематично, ненаглядно или же чрезмерно формально. Ниже мы следуем дидактике, доказавшей свою действенность в нашей многолетней преподавательской практике.

Во введении в факторной анализ надо обратить внимание аудитории на следующие принципиальные моменты:

- опосредованность и относительность исходных переменных, не отражающих истинные действующие факторы;
- эти скрытые факторы могут проявляться в нескольких переменных (коррелированность);
- сила влияния фактора проявляется в диапазоне изменения переменных (дисперсии;
- первая задача анализа — найти главные факторы;
- простой и понятный предметный иллюстрирующий пример.

Исходные данные представляются в виде матрицы, содержащей данные одного из следующих двух типов:

- значения m переменных для n объектов (матрица размером $m \cdot n$);
- квадратная матрица корреляций между m переменными (матрица размером $m \cdot m$).

В случае исходных данных типа «переменные–объекты» процедура предварительно позиционирует значения переменных на нуль вычитани-

[1] Естественно–научными мы называем исследования первичных природных закономерностей посредством измерительных приборов, что их принципиально отличает от исследований мнений людей на заданную тему (гумманитарно–опросные исследования).

ем средних значений переменных, а при методе корреляций (см. ниже «Данные и корреляции») значения дополнительно нормируются на стандартные отклонения переменных — это следует помнить для понимания последующих результатов.

Вторая форма полезна тогда, когда используется уже готовая матрица корреляций, например вычисленная во внешнем пакете. Однако в этом случае невозможен постфакторный анализ распределения и взаимозависимостей объектов.

Действия и результаты следуют естественной линейной последовательности выполнения факторного анализа[1].

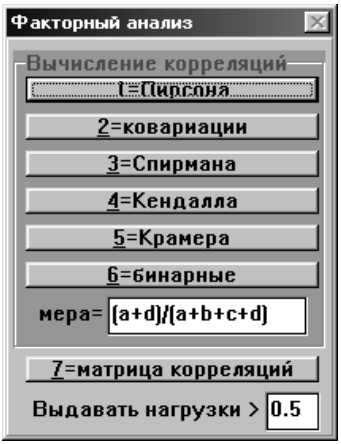


Рис. 11.2. Меню выбора формы исходных данных и корреляций

1. Данные и корреляции. Сначала следует уточнить тип исходных данных (рис. 11.2) и вид корреляций:

Если исходные данные представлены в форме корреляционной матрицы, то она сразу используется для факторного анализа (кнопка 7 меню рис. 11.2). Если же данные представляют собой значения *m* переменных для *n* объектов, то процедура предварительно вычисляет для анализа корреляционную матрицу, и в этом случае следует уточнить метод вычисления корреляций:

1) параметрические коэффициенты корреляции Пирсона (см. разд. 6.3) являются наиболее употребительной формой корреляции в факторном анализе в случае числовых и нормально распределенных исходных данных;
2) ковариации представляют собой взаимные вариации между переменными, ковариация переменной с самой собой является ее дисперсией; математически ковариация представляет собой числитель из формулы коэффициента корреляции, деленный на размер выборки; использование ковариационной матрицы сравнительно менее употребительно и позволяет в вычислениях учитывать не только степень взаимосвязанности (коррелированности) переменных, но и абсолютную величину ковариаций;
3), 4) непараметрические коэффициенты корреляции Спирмана и конкордации Кенделла (см. разд. 7.5) применимы в случае ненормально распределенных числовых данных и ранговых переменных;
5) коэффициенты связности Крамера (см. разд. 7.6) применимы для номинальных и ранговых переменных; значения переменных в матрице данных должны быть представлены в виде целых чисел (номинальных кодов или рангов) или же символьных номинальных обозначений, объединение этих двух форм представления данных недопустимо; для каждой пары переменных процедура предварительно рассчитывает таблицу кросстабуляции; ***примечание:*** число градаций значений переменных не должно превышать 20;

[1] За многочисленные полезные предложения по представлению результатов факторного анализа мы особо признательны А.Н. Гусеву.

6) коэффициенты связности бинарных переменных, вычисляемые по вводимой формуле (поле ввода *Мера*, рис. 11.2); значения переменных в матрице данных должны быть представлены кодами 1, 2 или же двумя символьными константами, объединение этих двух форм представления данных недопустимо; для каждой пары переменных процедура предварительно рассчитывает таблицу кросстабуляции размера 2×2 (см. разд. 7.6):

a	*b*
c	*d*

где обозначения *a, b, c, d* должны быть использованы в формуле вычисления коэффициента связности переменных, значения коэффициента должны находиться в диапазоне 0–1. Выбор формулы зависит от того, какую роль (позитивный, негативный или нейтральный смысл) исследователь отводит частотам встречаемости пар значений признаков *a, b, c, d*, например: *(a+d)/(a+b+c+d); a/(a+b+c+d); 2a/(2a+b+c); a/(a+2(b+c)); (ad–bc)/(a+b+c+d)*^2 и др.

В этом же меню можно установить абсолютное значение уровня нечувствительности для числовой выдачи значений *факторных нагрузок* (см. ниже), что удобно для лучшей ориентации в объемных выдачах (с исключением из них малых по абсолютной величине нагрузок) при большом числе переменных.

2. Корреляции. В ряде случаев здесь бывает полезным вывести таблицу «*переменная–среднее–ст.отклонение*» и диагональную (симметричную) матрицу коэффициентов корреляции между парами переменных (здесь и ниже результаты преимущественно взяты из примера данных, рассмотренных во введении к данному разделу):

```
Переменная  Среднее  Ст.отклон.
       x1  -0.5883    4.078
       x2  -0.7884      4.2
       x3  -0.6719    4.144

Корреляционная матрица
              x1       x2       x3
       x2    0.907
       x3    0.968    0.871
  Критическое значение=0.1941
 Число значимых коэффициентов=3 (100%)
```

С учетом того, что корреляционная матрица симметрична и на диагонали ее стоят единицы, приводится только ее поддиагональная часть. Выводится также критическое значение (для коэффициентов, значения которых меньше критического, принимается нулевая гипотеза «коэффициент не отличен от нуля») и выдается число и процент значимых коэффициентов.

3. Главные компоненты. Изложение результатов факторного анализа следует начинать с метода главных компонент, как наиболее интуитивно познаваемого и допускающего наглядные геометрические иллюстрации (рис. 11.1).

4. График «каменистая осыпь» представляет собой график распределения собственных значений факторов (по оси *Y*) в порядке убывания их величины (рис. 11.4). Этот график используется для отбора числа значимых факторов.

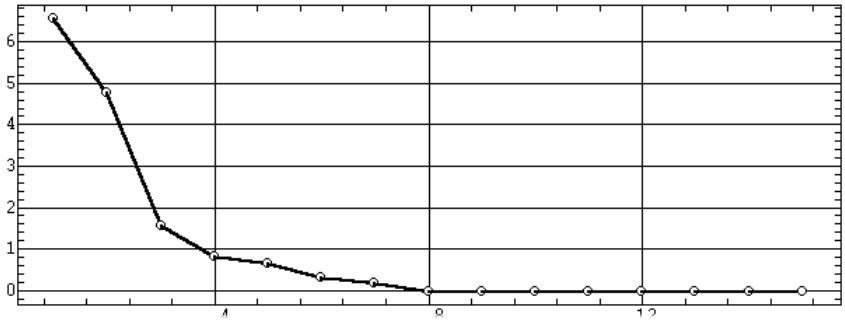


Рис. 11.4. Пример графика собственных значений главных компонент

5. Выбор числа значимых факторов. Полученные результаты позволяют произвести выбор числа значимых или, как их обычно называют, *общих факторов* для последующего анализа или вращения.

Для этого можно использовать три критерия:

1) критерий Кайзера состоит в отбрасывании компонент, собственные значения которых меньше единицы, что далеко не всегда оправдано[1];
2) критерий Кеттелла использует на графике собственных значений точку перегиба к его выполаживанию (на рис. 11.4 это соответствует 3–4 факторам);
3) отбрасывание компонент, суммарно отражающих менее 5–20% общей дисперсии.

Неучитываемые в дальнейшем малозначимые компоненты называются *специфическими (*или *характерными) факторами*.

6. Собственные вектора. Во многих учебниках после рассмотрения собственных значений сразу производится скачек к сравнительно сложным для восприятия факторным нагрузкам. Однако дидактически правильным является первоначальное рассмотрение *собственных векторов* (вектора главных компонент), которые допускают прямое геометри-

[1] Мотивируется это следующим образом: поскольку дисперсия нормализованных переменных равна единице, то компоненты с меньшими собственными значениями (а они отражают факторную дисперсию) менее значимы, нежели сами переменные. Однако часто встречается ситуация (например, при малом числе переменных или при высокой корреляции только между двумя из них), когда уже второй компонент имеет собственное значение, меньшее единицы, а тогда по рассматриваемому критерию факторный анализ становится бессмысленным.

7. Факторные нагрузки. Математически *факторная нагрузка* равна векторному коэффициенту a_{ij} перехода от переменной j к фактору i, умноженному на корень квадратный из собственного значения фактора:

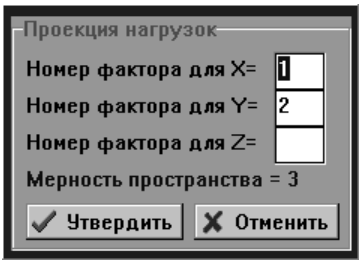
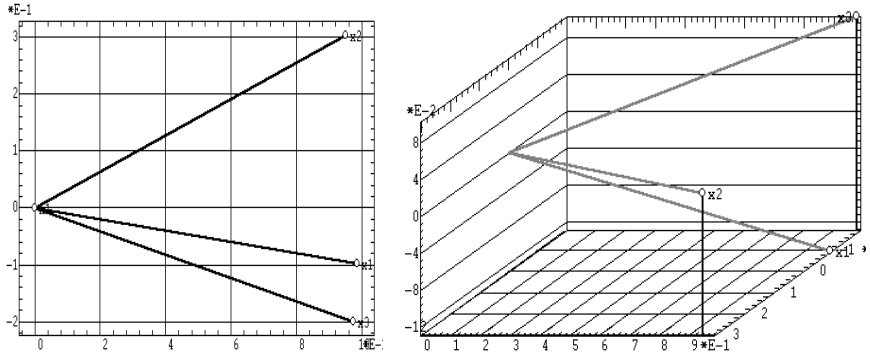


Рис. 11.6. Бланк выбора факторов для проекции нагрузок

Графики нагрузок. Для получения наглядного представления о соотношении величин факторных нагрузок полезно построить их векторные графики. Для этого в бланке (рис. 11.6) можно указать номера двух факторов и тогда выводимый векторный график будет плоским (рис. 11.7, *а*) или же указать три фактора и тогда будет выдан трехмерный векторный график (рис. 11.7, *б*). Запросы новых проекций нагрузок повторяются до отме-

ны бланка.



а *б*

Рис. 11.7. Проекции факторных нагрузок (для рассматриваемого примера): *а* — на плоскость факторов 1–2; *б* — в пространстве факторов 1–3

На этих графиках: чем ближе вектор к факторной оси, тем более переменная выражается через данный фактор и тем менее — через перпендикулярный фактор. Чем длиннее вектор, тем сравнительно более весомо представлена переменная в факторах.

10. Проекции объектов в факторных координатах. По результатам факторного анализа часто бывает необходимо получить факторные координаты объектов (неудачный синоним — *факторные значения*[1]) и построить их проекции на основные факторные плоскости.

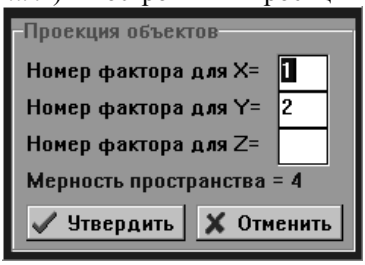
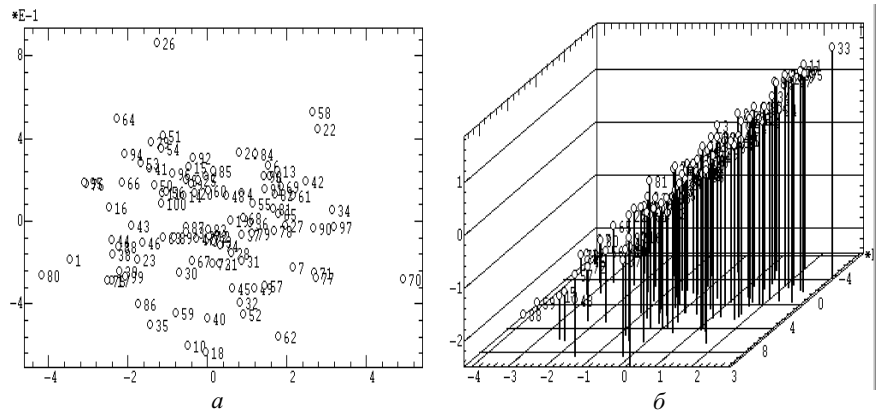


Рис. 11.7. Бланк выбора факторов для проекции объектов

Если в следующем экранном бланке (рис. 11.7) указать два фактора, то будет построена двумерная диаграмма рассеяния (рис. 11.8, *а*), если же указать три фактора, то будет выдана трехмерная диаграмма (рис. 11.8, *б*). Запросы новых проекций объектов повторяются до отмены бланка.

[1] Термин *«факторные значения»*, являющийся дословной калькой с английского, следует признать крайне неудачным. Действительно, что по-русски значит *«факторные значения»*? Это – значения факторов, но никак не факторные координаты объектов!



а *б*

Рис. 11.8. Проекции объектов (для рассматриваемого примера):
а — на факторную плоскость 1–2; *б* — в пространство факторов 1–3

Координаты проекций объектов с графика можно сохранить в электронной таблице для дальнейшего анализа (например, для кластеризации, сравнения различий, регрессионного анализа и т. п.) нажатием на инструментальную кнопку *«СохрГраф»*.

Примечание. Проекции объектов выдаются в натуральном виде, поэтому размерность по факторной оси, имеющей большее собственное значение, будет больше, чем по менее значимому фактору, а дисперсия измерений по конкретному фактору равна его собственному значению[1].

11. Факторное прогнозирование[1]. Во многих исследованиях встает задача прогнозирования значений некоторой исходной переменной при заданных значениях факторов. В контексте факторного анализа это можно назвать также обратной задачей, задачей восстановления или интерполяции значений исходной переменной.

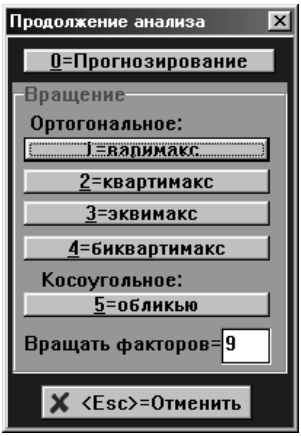


Рис. 11.9. Меню продолжения анализа

Поясним это на примерах. Пусть в проекции измерений на факторную плоскость прослеживается некая явная функциональная зависимость между ними. Тогда законной будет постановка задачи прогнозирования этой зависимости в одном из двух направлений ее продолжения и проецирования этого прогноза на различные исходные переменные. Другим примером (даже в отсутствие упомянутой функциональной зависимости) является интерес исследователя к поведению исходных переменных в области промежуточных факторных значений, не зафиксированных в эксперименте. Это соответствует задаче интерполяции. Особый интерес представляет случай многомерных данных типа связных временных рядов, когда рассматриваемая постановка соответствует классической задаче прогнозирования временных рядов (см. в разд. 14.4.3).

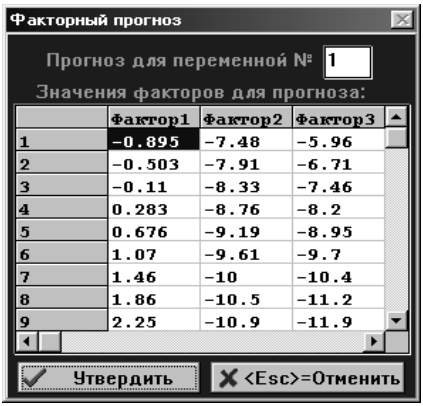


Рис. 11.10. Бланк ввода факторных значений для прогнозирования

После выбора факторного прогнозирования (рис. 11.9) нужно в факторную таблицу (рис. 11.10) ввести значения главных факторов, по которым будет вычислен прогноз, а также указать порядковый номер прогнозируемой переменной из электронной таблицы. Факторная таблица позволяет вводить до 100 наборов значений (строки) для 2–7 главных факторов (столбцы). Ввод значений в факторную таблицу может осуществляться вручную или из буфера обмена комбинацией клавиш [Ctrl] + [V] или [Shift] + [Ins]. В результате вычисляются значения указанной исходной переменной для каждого из введенных наборов значений факторов.

Факторное прогнозирование не рекомендуется применять после вращения в связи с рядом искажений факторной структуры. Методика факторного прогнозирования детально проиллюстрирована в разд. 14.4.3.

12. Вращение факторов. После выделения факторов многие источники рекомендуют произвести *вращение* избранных факторных векторов в определенном этими факторами подпространстве. Целью вращения провозглашается получение более просто интерпретируемой системы факторов (*простая структура*), при которой каждая переменная имеет большие нагрузки на малое число факторов и малые нагрузки на остальные факторы.

Следует категорически заявить, что формулировка подобной цели в методическом плане принадлежит к области лженауки. Это называется подгонкой исходных данных. Представьте себе, что Уотсон и Крик начали бы вращать рентгеноскопические данные с целью облегчить себе жизнь и более быстро получить максимально простую модель ДНК. Тем не менее в связи с распространенностью этой ереси[1] приведем некоторые сведения о вращениях.

[1] Многих сбивает также с толку кочующее по учебникам утверждение, что методы факторного анализа (см. ниже) дают решение с точностью до вращения факторов, из чего делается вывод, что вращение принципиально не меняет найденного решения. Однако факторное решение обеспечивает хорошее приближение общностей переменных (= сумма нагрузок) к исходным корреляциям, а общности переменных (в отличие от нагрузок) при вращении действительно не меняются. Но такое «обоснование» аналогично следующей наглядной ситуации. Пусть гражданин А платит за жилье миллион, а гражданин Б – тысячу. И мы говорим: давай-

Для вращения необходимо указать число существенных факторов, которые будут вращаться, и выбрать метод вращения в меню рис. 11.9.

Выдача результатов. По окончании вращения выдаются следующие таблицы и графики:

1. Таблица *общностей* и *специфичностей* каждой переменной:

```
Вращение: варимакс, число факторов=2
Переменная  Общность  Специфичность
       x1   0.9844    0.01641
       x2   0.9968    0.000704
       x3   0.9909    0.01077
```

Общность переменной представляет собой сумму квадратов ее нагрузок на общие факторы, а специфичность — сумму квадратов нагрузок на специфические факторы.

В случае облимин–вращения (косоугольное вращение) выдается также матрица взаимных корреляций факторов, иллюстрирующая степень их взаимной неортогональности:

те-ка повращаем ситуацию до наоборот, ведь общая-то квартплата от этого не изменится!

```
Корреляции факторов
  Фактор:        1         2         3
        2  0.08756
        3  -0.3361  -0.1582
Критическое значение=0.5671 Число значимых коэффициентов=0 (0%)
```

2. Таблица собственных значений факторов и процентов объяснимой дисперсии (аналогична вышерассмотренной):

```
Собств.значения и %объясняемой дисперсии факторов после вращения
Фактор:          1         2         3
Собств.зн     1.673     1.299
 Дисперс%     55.77      43.3
  Накопл%     55.77     99.07
```

3. Таблица коэффициентов вращения факторных осей относительно главных компонент:

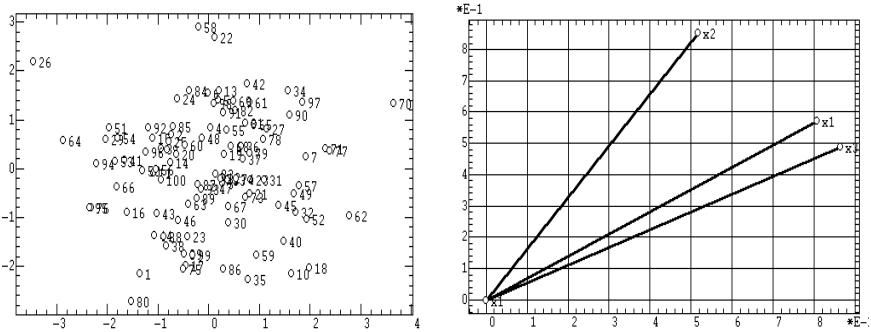
```
Коэффициенты вращения факторных осей
Нов.факторы:     1         2         3
        1   0.7547    0.6561
        2  -0.6561    0.7547
```

4. Таблица факторных нагрузок после вращения (аналогична вышерассмотренной):

```
Переменная <-------- Факторные нагрузки после вращения ---------->
     Фактор:         1         2
          x1     0.809    0.5744
          x2    0.5184    0.8532
          x3     0.866    0.4908
```

Примечания: а) поскольку вращаются только координатные оси в подпространстве выбранных факторов, то факторные нагрузки специфических факторов не изменяются; б) после вращения собственные значения уже не являются суммой квадратов нагрузок.

5. Графики проекций объектов на факторные оси и векторные графики факторных нагрузок (рис. 11.6, 11.8). Запросы новых плоскостей проекций (рис. 11.5, 11.7) продолжаются до их отмены.
6. Далее диалог возвращается к запросу нового вращения (рис. 11.9), который можно отметить, закончив тем самым факторный анализ.



а *б*

Рис. 11.11. Проекции результатов вращения на факторную плоскость 1–2:
а — объекты; *б* — факторные нагрузки

$i=1$ $i=1$ $i=k+1$

Примеры

Здесь мы рассмотрим три примера применения факторного анализа в психологии, социологии, политологии и тестировании профессиональной пригодности, к чему примыкают многочисленные задачи опросного характера. Примеры из естественнонаучных областей рассмотрены в разд. 14.2–14.4.

Пример 1

Задача. Данный пример выполнен в психодиагностической методике «*Сказка*» для определения уровня развития детей[1], в которой ребенку предлагается оценить по трехбалльной шкале «*Да*»–«*Не знаю*»–«*Нет*» (выражаются баллами 1, 0, –1) черты характера и личностные особенности (свойства) известных сказочных героев: Айболит, Буратино, Кот в сапогах, Снежная королева, Карабас-Барабас, Карлсон, Мальвина и Пьеро.

Важное примечание. Здесь мы имеем очень важный дидактический пример перехода от ранговой шкалы измерений (1, 0, –1) к метрической, более приемлемой для факторного анализа. Такой переход осуществляется посредством операции усреднения ответов многих респондентов.

В данной методике анализ факторной структуры в пространстве свойств героев позволяет выявить структуру сознания ребенка и судить об уровне сформированности и усвоенности общественных стереотипов и норм поведения. Для этого индивидуальные результаты ответов ребенка сравниваются с факторными свойствами нормативно-усредненной матри-

[1] Петренко В.Ф. Психосемантика сознания. М.: МГУ, 1988.

цы, полученной из ответов преподавателей начальной школы (табл. 11.1, файл TALES, в котором переменные соответствуют строкам табл. 11.1).

Таблица 11.1. Характеристики сказочных персонажей

Характер	Айболит	Кот	Карлсон	СнКор	Кар-Бар	Бурат	Мальвина	Пьер
ВернДруг	1.00	0.87	0.47	-0.98	-0.80	0.80	0.10	0.93
Смелый	0.93	0.87	0.20	0.30	-0.13	0.87	-0.33	0.13
Красивый	-0.27	0.53	-0.20	1.00	-0.80	-0.33	0.93	0.47
Добрый	1.00	0.73	0.60	-1.00	-1.00	0.80	0.21	0.87
Хитрый	-0.87	0.9	0.50	0.60	0.50	0.40	-0.93	-0.93
Жадный	-1.00	-0.67	0.53	0.98	1.00	-0.80	-0.87	-1.00
Плакса	-1.00	-0.95	-0.93	-0.67	0.00	-0.95	0.30	0.90
Умный	1.00	0.87	0.58	0.10	-0.27	0.43	0.44	0.37
Ябеда	-1.00	-1.00	-0.67	-1.00	0.78	-0.91	0.87	-0.83
Воспитан	0.93	0.93	-0.80	0.00	-0.90	-0.70	1.00	0.87
Хвастун	-1.00	0.27	1.00	0.33	0.60	0.77	0.00	-1.00
Умелый	1.00	0.93	0.20	0.03	-0.47	0.10	0.40	0.23
Драчун	-0.98	0.53	0.20	0.70	1.00	0.73	-1.00	-0.98
Шалун	-0.97	0.27	1.00	-0.87	0.00	1.00	-1.00	-0.80
Веселый	0.63	0.87	1.00	-1.00	-0.93	1.00	0.01	-1.00

Ниже производится анализ этой матрицы усредненных ответов преподавателей.

Р е з у л ь т а т ы:

```
ФАКТОРНЫЙ АНАЛИЗ.  Файл: tale1.txt
Переменная   Среднее  Ст.отклон.
Верн.друг     0.299     0.791
   Смелый     0.322      0.51
 Красивый     0.166     0.655
   Добрый     0.276     0.821
   Хитрый    0.0213     0.785
   Жадный    -0.229       0.9
   Плакса    -0.412     0.723
    Умный      0.44     0.404
    Ябеда     -0.47     0.807
 Воспитан     0.166     0.862
  Хвастун     0.121     0.758
   Умелый     0.303     0.481
   Драчун     0.025     0.867
    Шалун    -0.171      0.86
  Веселый    0.0725     0.924

Корреляционная матрица (сокращенно):
         Верн.друг Смелый Красивый Добрый Хитрый  Жадный  Плакса
   Смелый    0.482
 Красивый   -0.103  -0.218
   Добрый     0.99   0.448  -0.061
   Хитрый   -0.351   0.324  -0.18   -0.386
   Жадный   -0.843   -0.29  -0.174  -0.838   0.622
   Плакса   -0.128  -0.822   0.255  -0.14   -0.573  -0.146
    Умный    0.817   0.633   0.133   0.834  -0.237  -0.683  -0.452
    Ябеда   -0.473  -0.726  -0.109  -0.461  -0.202   0.207   0.506
 Воспитан    0.426   0.048   0.624   0.417  -0.605  -0.69    0.281
  Хвастун   -0.405   0.004  -0.236  -0.379   0.799   0.603  -0.427
```

Умелый	0.699	0.579	0.287	0.699	-0.282	-0.681	-0.353
Драчун	-0.499	0.225	-0.316	-0.543	0.94	0.682	-0.449
Шалун	0.2	0.331	-0.525	0.196	0.67	0.188	-0.498
Веселый	0.64	0.659	-0.248	0.671	0.192	-0.389	-0.71

. . .

Собственные значения и процент объясняемой дисперсии факторов

Фактор:	1	2	3	4	5	6
Собств.зн	6.57	4.79	1.58	0.853	0.663	0.335
Дисперс%	43.8	32	10.6	5.69	4.42	2.23
Накоплен%	43.8	75.8	86.3	92	96.4	98.7

Обсуждение: Как видно из результатов и рис. 11.12 в ходе вычисления главных компонент можно выделить компактную систему из трех–четырех основных факторов, отражающих 86–92% дисперсии объектов–героев (вместо исходного 15–мерного пространства свойств).

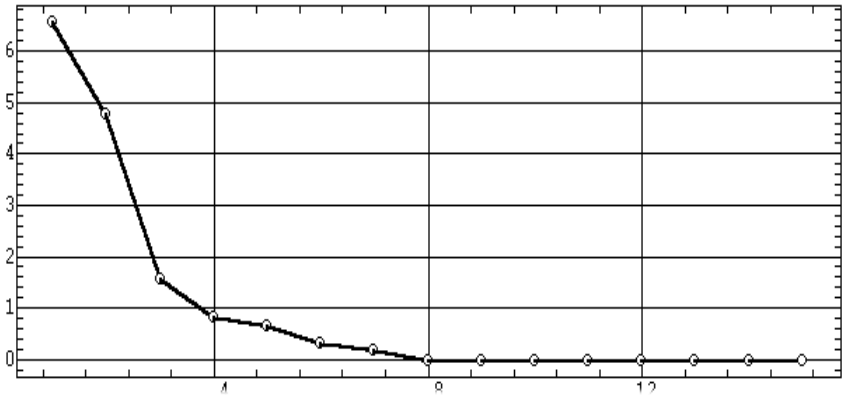
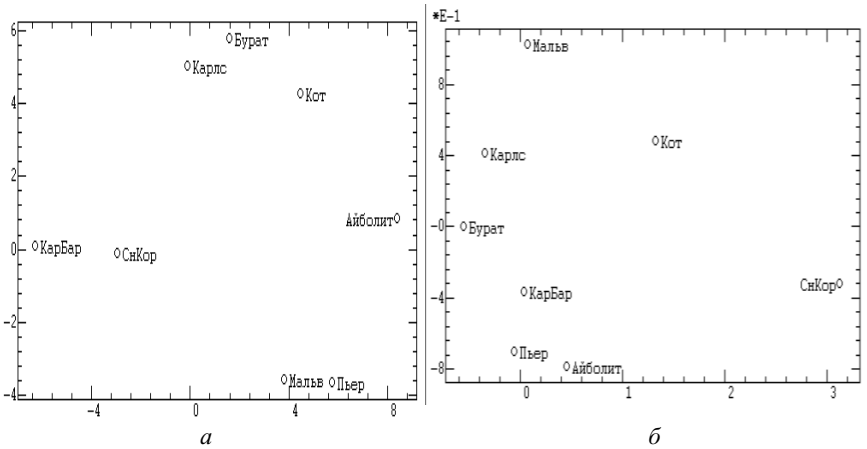


Рис. 11.12. График собственных значений факторов (по оси *Y*) относительно номеров факторов (по оси *X*)

Продолжение результатов:

Переменная <Собственные вектора (коэфф. поворота факторных осей)>

Верн.друг	0.334	0.141	-0.271	-0.0457	0.257	-0.0888
Смелый	0.163	0.351	0.231	-0.309	-0.139	-0.312
Красивый	0.0953	-0.198	0.496	0.525	0.4	0.129
Добрый	0.337	0.137	-0.276	0.0229	0.241	0.0947
Хитрый	-0.229	0.31	0.247	0.115	0.181	-0.299
Жадный	-0.348	0.0346	0.197	-0.0774	-0.137	0.464
Плакса	-0.0217	-0.399	-0.256	0.02	0.392	-0.219
Умный	0.347	0.173	0.0808	0.137	-0.149	0.185
Ябеда	-0.168	-0.231	-0.338	0.435	-0.51	-0.356
Воспитан	0.305	-0.209	0.255	0.166	-0.0395	-0.34
Хвастун	-0.27	0.261	-0.0646	0.455	0,0559	0.0881
Умелый	0.349	0.0905	0.225	0.141	-0.239	-0.0828
Драчун	-0.293	0.257	0.154	-0.0481	0.12	-0.458
Шалун	-0.108	0.379	-0.307	0.159	0.301	-0.00969
Веселый	0.175	0.359	-0.163	0.341	-0.208	0.123



а *б*

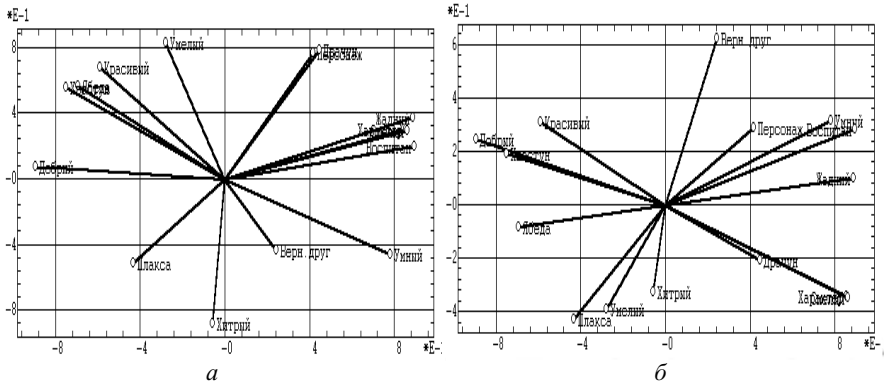
Рис. 11.13. Объекты (герои) в проекции на плоскость главных компонент: *а* — факторная плоскость 1–2; *б* — факторная плоскость 3–4

Обсуждение: Как видно из числовой выдачи и рис. 11.13 исследуемые объекты–герои в проекции на плоскости первых трех факторов образуют хорошо различимые группировки, что дает надежду на хорошую дальнейшую интерпретацию результатов.

Продолжение результатов:

Переменная	<------ Факторные нагрузки до вращения ---------->						
Верн.друг	0.857	0.309	-0.341	-0.0422	0.209	-0,0514	-0.0757
Смелый	0.417	0.769	0.291	-0.285	-0.113	-0.18	0.152
Красивый	0.244	-0.433	0.625	0.485	0.325	0.0745	0.127
Добрый	0.864	0.3	-0.348	0.0212	0.196	0,0548	-0.023
Хитрый	-0.588	0.679	0.311	0.106	0.147	-0.173	-0.183
Жадный	-0.892	0.0757	0.248	-0.0714	-0.112	0.268	-0.217
Плакса	-0,0555	-0.875	-0.322	0.0185	0.319	-0.127	-0.0973
Умный	0.89	0.378	0.102	0.126	-0.121	0.107	-0.11
Ябеда	-0.432	-0.506	-0.425	0.402	-0.415	-0.206	-0.0155
Воспитан	0.783	-0.456	0.32	0.153	0.0322	-0.197	-0.11
Хвастун	-0.692	0.572	-0.0813	0.42	0.0455	0,051	0.0739
Умелый	0.895	0.198	0.283	0.13	-0.194	-0.048	-0.15
Драчун	-0.751	0.563	0.194	-0.0444	0.0978	-0.265	-0.003
Шалун	-0.277	0.829	-0.387	0.147	0.245	-0.00561	-0.0616
Веселый	0.449	0.787	-0.205	0.315	-0.169	0.0712	0.0614

Обсуждение: Как видно из числовой выдачи и рис. 11.14 исходные свойства проецируются в сравнимой степени на несколько факторов, что может затруднить их предметную интерпретацию. Поэтому выберем три основных фактора и проведем их варимакс–вращение в пространстве 15 исходных переменных (свойств) с целью получения более простой структуры проекции.



а *б*

Рис. 11.14. Факторные нагрузки исходных свойств героев:
а — факторная плоскость 1–2; *б* — факторная плоскость 3–4

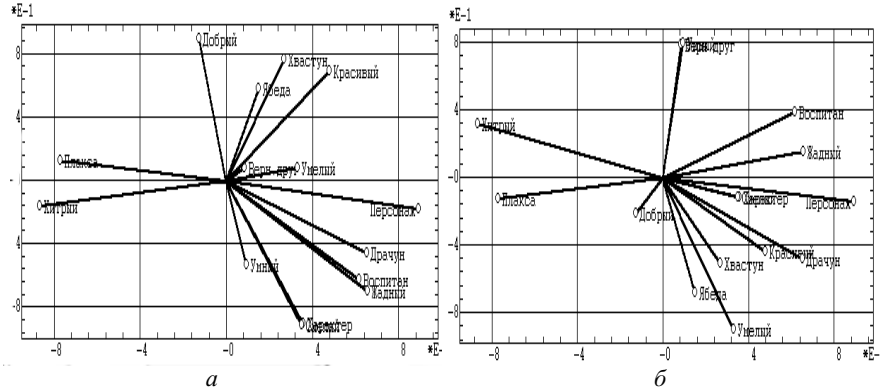
Продолжение результатов:

Переменная	Общность	Специфичность
Верн.друг	0.946	0,0546
Смелый	0.85	0.15
Красивый	0.637	0.363
Добрый	0.957	0.043
Хитрый	0.904	0.0968
Жадный	0.863	0.137
Плакса	0.872	0.128
Умный	0.945	0,055
Ябеда	0.623	0.377
Воспитан	0.924	0.0759
Хвастун	0.813	0.187
Умелый	0.92	0.0799
Драчун	0.918	0.0821
Шалун	0.914	0.0863
Веселый	0.863	0.138

Переменная	Факторные нагрузки после вращения		
Верн.друг	0.359	-0.897	-0.114
Смелый	0.894	-0.178	-0.14
Красивый	0.0865	0.0875	0.789
Добрый	0.351	-0.907	-0.111
Хитрый	0.48	0.694	-0.438
Жадный	-0.125	0.897	-0.207
Плакса	-0.864	-0.154	0.32
Умный	0.659	-0.698	0.154
Ябеда	-0.768	0.134	-0.123
Воспитан	0.0967	-0.532	0.795
Хвастун	0.152	0.58	-0.674
Умелый	0.62	-0.618	0.393
Драчун	0.271	0.769	-0.503
Шалун	0.333	0.0791	-0.892
Веселый	0.655	-0.455	-0.476

Обсуждение: Как видно из сравнения рис. 11.14 и 11.15, в результате вращения получена более простая структура, в которой исходные переменные (свойства) преимущественно проецируются на один из трех

главных факторов, что облегчает задачу интерпретации факторов в терминах исходных переменных.



а *б*

Рис. 11.15. Факторные нагрузки исходных свойств героев после вращения: *а* — факторная плоскость 1–2; *б* — факторная плоскость 3–4

Напомним, что факторная нагрузка показывает, насколько выражено в данной шкале (в переменной) содержание, которое описывает фактор. При интерпретации важно найти семантический смысл выделенных факторов, чему помогает группировка объектов и свойств по наибольшим нагрузкам основных факторов (см. табл. 11.2).

Таблица 11.2. Группировка свойств и героев по факторам

	Фактор 1	Фактор 2	Фактор 3
Отрицательн. Свойства	Плакса=-0.84 Ябеда=-0.768	Жадный=0.897 Драчун=0.769 Хитрый=0.694	Шалун=-0.892 Хвастун=-0.674
Отрицательн. Герои	КарБараб=-0.823 Мальвина=-0.52 Пьерро=-0.239	Кар-Бараб=2.48 СнежнКорол=1.88	Карлсон=-2.75 Буратино=-2.16 Кар-Бараб=-1.86
Положительн. Свойства	Смелый=0.894 Умный=0.659 Веселый=0.655 Умелый=0.62	Добрый =-0.907 Верн.друг=0.897 Умный=-0.698 Умелый=-0.618	Воспитан=0.795 Красивый=0.798
Положительн. герои	Кот в сап=2.18 Айболит=1.52	Айболит=-3.19 Пьеро= -2.35	Пьеро=3.49 Мальвина=2.97 Айболит=2.26

После исследования содержимого табл. 11.2 можно дать следующие психологические интерпретации трех факторов:

- фактор 1: «Сила личности (эго)»;
- фактор 2: «Добро — Зло»;
- фактор 3: «Социальная нормативность — акцентированность».

Следует обратить особое внимание на то, что категория *Сила личности* для современных взрослых воспитателей стоит на более высоком уровне, чем дихотомия *Добро—Зло*.

Упражнение. В качестве самостоятельного упражнения полезно заполнить исходную матрицу *герои–свойства* своими собственными ответами или ответами своего ребенка и, повторив факторный анализ, сравнить полученные результаты с нормативными. Следует, однако, предостеречь от резких педагогических выводов из такого сопоставления, поскольку приведенные нормативные результаты отражают только представления взрослого человека о добре и зле и применительно не к самому себе, а к идеальному с их точки зрения ребенку, что может быть далеко не вполне толерантно представлениям об этих категориях нормального ребенка.

Заключение. Если заменить в данной методике сказочных героев на политических лидеров, партии, государства или транснациональные корпорации, легко осознать значимость получаемых результатов и практическую ценность выводов в совершенно других и неизмеримо менее «игрушечных» областях.[1]

Пример 2

Классический дидактический пример социологического применения факторного анализа приведен в статье Д.К. Арчера, Ф.М. Шелли, П.Д. Тейлора и Э.Р. Уайта[2] для изучения факторов, влияющих на президентские выборы в США (рис. 11.16).

Сырые исходные данные представляют собой матрицу, содержащую проценты голосов, поданных за демократов на 28 выборах в 1872—1984 гг. (строки) в каждом из 50 штатов (столбцы). На основе этих данных независимо решались две следующие задачи:

1) ***выделение региональных факторов***: в этом случае строилась матрица «*штат–штат*» корреляций временных рядов изменения процента голосов (между строками в исходной матрице);
2) ***выделение временных факторов***: в этом случае строилась матрица «*выборы–выборы*» корреляций между процентами голосов всех штатов (между столбцами в исходной матрице).

Эти две корреляционные матрицы и явились исходными данными для факторного анализа. Данный пример интересен еще и тем, что в нем проводятся анализ прямой и транспонированной матриц, т. е. матрицы «переменные–измерения» и матрицы «измерения–переменные».

В результате оказалось, что большинство выявленных *региональных корреляций* можно описать всего тремя факторами (из 50 штатов), интерпретируемых как «нормальный региональный», «нормальный либераль-

---

[1] Об использовании факторного анализа в социальной и политической психологии см. статьи В.Ф.Петренко с сотрудниками в Психологическом журнале: N6, 1991 (Семантическое пространство политических партий России) и N6, 1995 (Психосемантический анализ динамики качества жизни Россиян).

[2] Scientific American, №9, 1988.

ный» и «нормальный консервативный» типы. Во *временном срезе* было выделено шесть основных факторов (из 28 выборов), интерпретируемых по двум типам «предпочтения»: 1) *социально–экономические*: «городской», «сельский», «пригородный»; 2) *региональные*: «юг», «запад», «северо-восток».

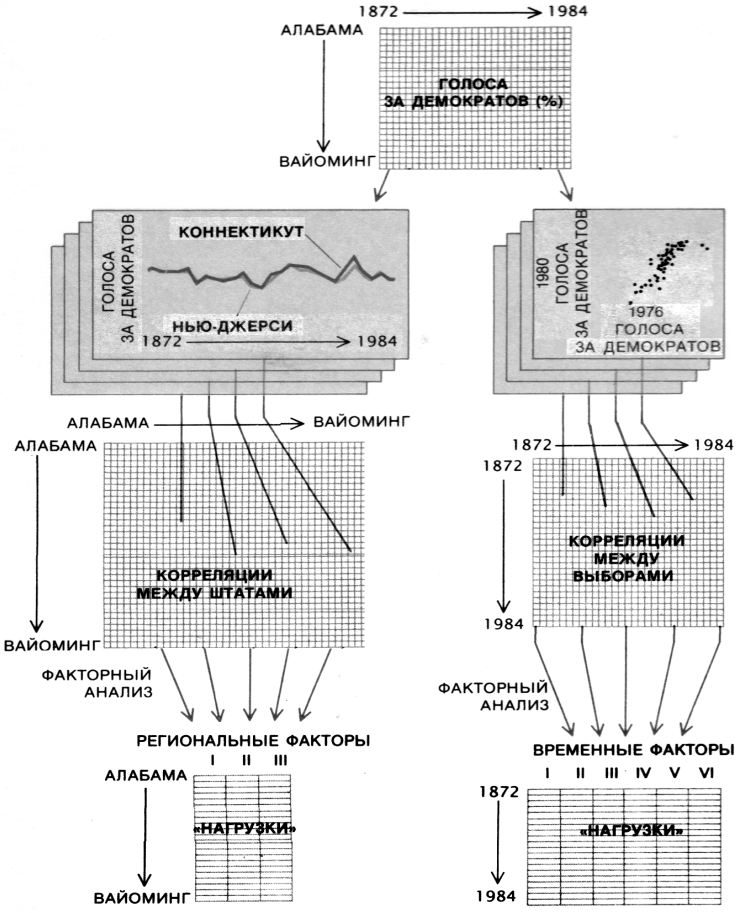


Рис. 11.16. Факторный анализ президентских выборов в США. Из исходной матрицы «число голосов, поданных за демократов» по штатам и годам выделяются: *а* — временные зависимости по каждому штату; *б* — зависимости между всеми парами выборов, по всем выборам каждая. Для этих двух типов данных вычисляются корреляционные матрицы, которые подвергаются факторному анализу. В результате выделено три основных действующих региональных и шесть временных факторов.

Каждому штату соответствует определенная *нагрузка* по каждому фактору (проекция фактора на штат) и для каждых выборов. И здесь мы имеем также прекрасный пример использования графиков изменения факторных нагрузок во времени.

Так в результате изучения временных графиков изменения нагрузок региональных факторов было выявлено существование «избирательных эпох», в каждой из которых в отдельных штатах проявляется тенденция сохранять определенный тип голосования по всей стране (рис. 11.17).

При изучении графиков изменения нагрузок временных факторов было выявлено различие региональных предпочтений перед социально-экономическими и устойчиво-периодический характер изменения обоих типов предпочтения (рис. 11.18).

Добавим к этому, что подобные графики являются плодотворным материалом для последующего построения прогностических моделей методами регрессионного и фурье-анализа.

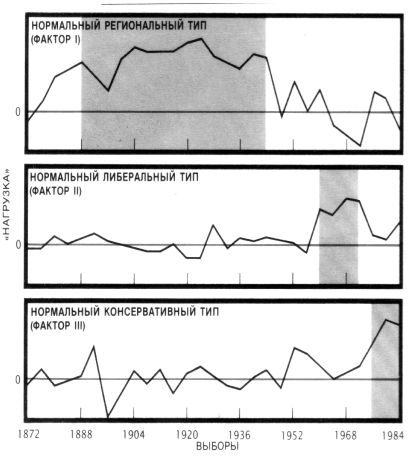


Рис. 11.17. «Избирательные эпохи», в течение которых поведение избирателей в конкретном регионе остается достаточно стабильным

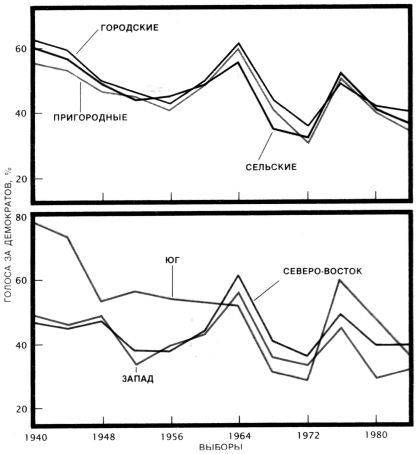


Рис. 11.18. Региональные предпочтения (внизу) играют большую роль (большее различие нагрузок), чем социально-экономические (вверху)

В результате дальнейшего осмысления результатов были сформулированы и верифицированы многие важные комплексные понятия, например: *урбанизация* и *региональное разделение*, *устойчивые* и *неустойчивые предпочтения* и др.

Чрезвычайно эффективными для визуального исследования явились также географические карты, раскрашиваемые по различным показателям факторного анализа (рис. 11.19).

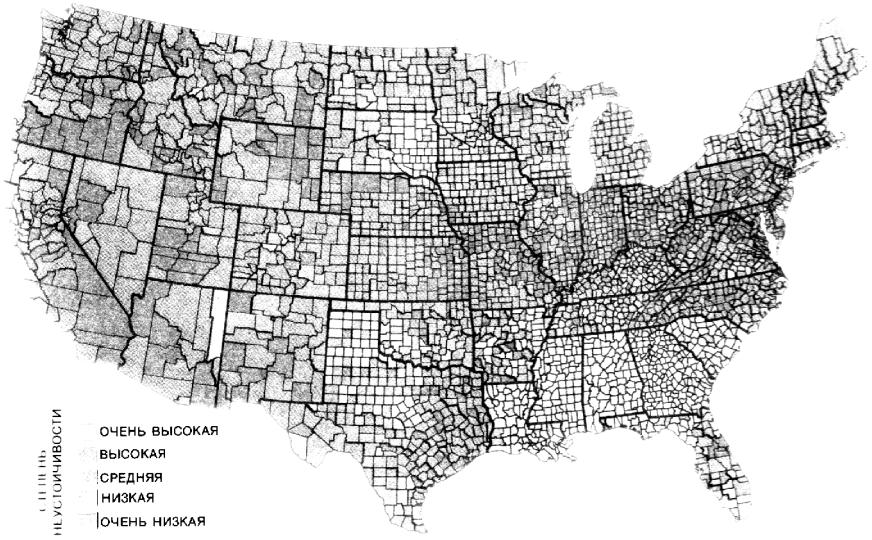


Рис. 11.19. Неустойчивость предпочтений избирателей, равно как и стабильность их симпатий, географически остается неизменной (данные по более 3000 графств). Районы с наибольшей неустойчивостью (Северо-Восток, Юг, север Среднего Запада) составляют главные поля сражения обеих партий

Таким образом, визуальное изучение наглядных факторных (географических) диаграмм и их временных графиков позволяет понять динамику и механизмы сложных социальных процессов и дать им практическую интерпретацию, важную для выработки текущей экономической и региональной политики, а также для коррекции последующих предвыборных стратегий.

Заключение. Заменим в данном примере процент голосов любым экономическим показателем и получим множество впечатляющих примеров факторных исследований из области маркетинга, менеджмента, бизнеса, экономики, геополитики и т. п.

Пример 3

З а д а ч а. При приеме на работу 18 претендентов прошли 10 специальных тестов, результаты каждого из которых оценивались по 10-балльной системе (табл. 11.3, файл TESTS[1]). Отметим, что этот тип данных в аналитическом плане эквивалентен многочисленным задачам различного типа опросов с ответами в заданной шкале предпочтений.

Интересно было бы выявить главные факторы, действующие в этой батарее из 10 тестов, а также выяснить наличие или отсутствие группировок среди претендентов.

---

[1] Данные из архива SPSS, аналитика наша.

Таблица 11.3. Результаты тестирования 18 претендентов на работу по 10-балльной системе

Тесты: _____ Испытуемые	Память на числа	Математические задачи	Находчивость в диалоге	Составление алгоритмов	Уверенность в выступлении	Командный дух	Находчивость	Сотрудничество	Признание в коллективе	Сила убеждения
1	10	10	9	10	10	10	9	10	10	9
2	10	10	4	10	5	5	4	5	4	3
3	5	4	10	5	10	4	10	5	3	10
4	10	10	9	10	10	10	9	10	10	9
5	4	3	5	4	3	10	4	10	10	5
6	10	10	4	10	5	4	3	4	5	5
7	4	4	5	5	4	10	5	10	10	6
8	4	5	3	4	5	10	4	10	10	4
9	4	5	10	4	10	5	10	4	3	10
10	10	10	4	10	5	4	4	5	4	4
11	4	5	10	5	10	4	10	4	5	10
12	10	10	9	10	10	9	9	10	10	10
13	6	5	4	3	5	10	5	10	10	5
14	4	5	10	4	10	5	10	3	4	10
15	10	10	9	10	10	9	10	9	10	10
16	6	5	3	4	4	10	4	10	10	5
17	10	10	5	10	4	5	4	3	4	5
18	4	5	10	4	10	4	10	4	4	10

Здесь мы имеем дело с ранговыми переменными, поэтому для оценки их коррелированности приемлемы коэффициенты Спирмана, Кенделла и Крамера. Для сохранения общности результатов с номинальными данными используем в факторном анализе коэффициент связности Крамера.

Р е з у л ь т а т ы:

```
ФАКТОРНЫЙ АНАЛИЗ.  Файл: tests.std

Собственные значения и процент объясняемой дисперсии факторов
Фактор:       1     2     3     4     5     6     7     8
Собств.зн  5.72  1.35  1.02  0.52  0.423  0.327  0.269  0.199
 Дисперс%  57.2  13.5  10.2   5.2   4.23   3.27   2.69   1.99
Накоплен%  57.2  70.6  80.9  86.1   90.3   93.6   96.2   98.2

Переменная <---- Факторные нагрузки до вращения ------>
  Фактор:        1         2         3         47        8
   память    0.752
 мат.зада     0.75
 прям.диа    0.885
 алгоритм    0.736
 уверенно    0.717                -0.558
 команд.д    0.715     0.538
 находчив     0.83
 сотрудни    0.724     0.538
 признани    0.697     0.517
 убеждени    0.735
```

Выводы.

1. Как видно из таблицы собственных значений, в качестве главных факторов можно с уверенностью выбрать первые три, которые покрывают 80,9% дисперсии измерений. Четвертый и последующие компоненты в 2 с лишним раза меньше по величине третьего компонента, к тому же их собственные значения существенно меньше единицы.
2. Как видно из таблицы факторных нагрузок, все тесты преимущественно проецируются на первый фактор, который можно было бы назвать фактором «общей профессиональной пригодности». На второй фактор проецируются с недоминирующими весами тесты *командный дух, сотрудничество, признание в коллективе*. Этот фактор можно было бы назвать «эффективность в коллективной работе». На третий фактор с недоменирующим весом проецируется всего лишь один и малозначимый на начальном этапе работы претендента тест *уверенность в выступлении*. В связи с этим третий фактор можно в предварительном анализе не рассматривать.

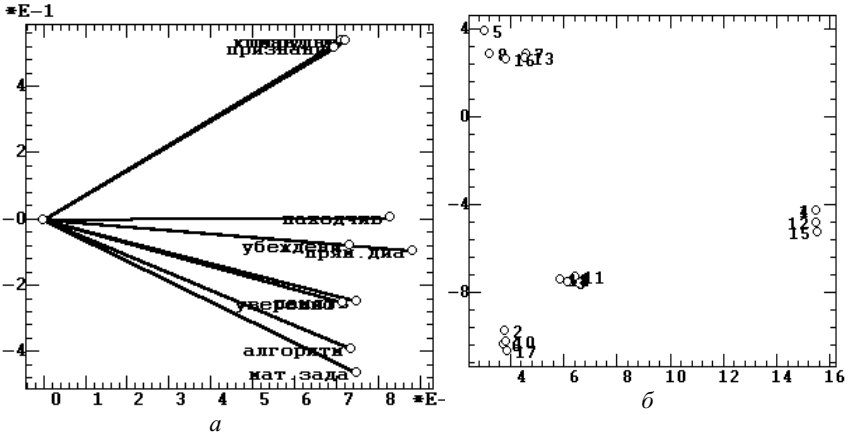


Рис. 11.20. Проекции на плоскость факторов 1–2 результатов тестирования: *а* — факторные нагрузки; *б* — претенденты

3. Как видно из графиков факторных нагрузок (рис. 11.20, *а*), две пары тестов имеют очень близкие по величине и проекции нагрузки: *командный дух, сотрудничество* и *уверенность в выступлении, память*. Поэтому эти тесты при повторном факторном анализе можно было бы исключить из рассмотрения.
4. В проекции претендентов (рис. 11.20, *б*) выделяются четыре плотные и далеко удаленные друг от друга группы претендентов: 1)={5, 7, 8, 13, 16}; 2)={2, 6, 10, 17}; 3)={3, 9, 11, 14, 18}; 4)={1, 4, 12, 15}. Группа 4 имеет заметное преимущество по первому фактору, а группа 1 — по второму фактору. Группы же 2 и 3 показали посредственную пригод-

ность по обоим факторам, поэтому по их представителям, вероятно, следует вынести отрицательное решение в отношении приема на работу.

Последующие исследования.

1. Провести факторный анализ с использованием корреляций Спирмана и Кенделла и оценить различие или сходство в результатах.
2. Исходные данные можно преобразовать к бинарному виду, например, по следующим соображениям. Установить порог оценок: удовлетворительно—неудовлетворительно (скажем, по уровню 5 баллов) и преобразовать по этому порогу данные[1]. Затем провести факторный анализ с использованием различных формул бинарных коэффициентов и оценить различие или сходство в результатах.
3. Провести кластерный анализ по агломеративным и дивизивной стратегиям и сравнить выявленные кластеры с группировками, визуально выявленными на рис. 11.20, *б*.

Эти задачи мы оставляем в качестве учебной практики читателям.

11.2. Кластерный анализ

Назначение. Метод кластерного анализа позволяет:

- строить дерево классификации *n* объектов посредством иерархического объединения их в группы или *кластеры* все более высокой общности на основе критерия минимума расстояния в пространстве *m* переменных, описывающих объекты;
- находить разбиение некоторого множества объектов на заданное число компактных кластеров.

Отметим, что кластерный анализ не содержит вычислительного механизма проверки гипотезы об адекватности получаемых классификаций. Результаты кластеризации в этом плане можно статистически верифицировать с использованием метода дискриминантного анализа (см. разд. 11.3).

Исходные данные представляются в виде матрицы размером $m \cdot n$, содержащей информацию одного из следующих трех типов:

- измерения X_{ij} значений *m* переменных для *n* объектов;
- квадратная ($m=n$) матрица расстояний между парами объектов;
- квадратная ($m=n$) матрица близостей для всех пар *n* объектов.

В матрице близостей или расстояний может быть заполнена лишь нижняя левая половина (т. е. поддиагональные элементы), а верхняя половина заполнена нулями.

Действия и результаты следуют естественной линейной последовательности выполнения кластерного анализа:

---

[1] Операцией кодирования в *Блоке преобразований* по условиям $x>5$ и $x<6$.

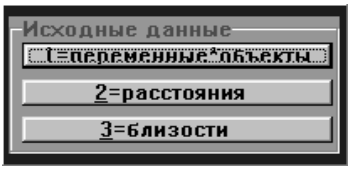


Рис. 11.21. Меню выбора типа исходных данных

1. Тип данных. Сначала необходимо указать тип исходных данных (рис. 11.21): прямоугольная матрица: переменные (столбцы) и объекты (строки), или же квадратная матрица взаимных расстояний, или матрица близостей между всеми парами объектов.

2. Метрика.

Если исходные данные представляют собой значения m переменных для n объектов, то далее необходимо выбрать (рис. 11.22) метод вычисления расстояния d_{ij} между объектами в многомерном пространстве (метрику). Пояснения и рекомендации по использованию различных метрик даны ниже.

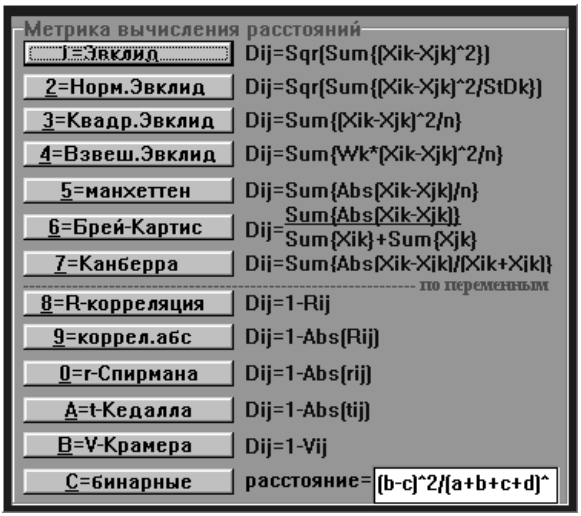


Рис. 11.22. Меню выбора метрики расстояний

3. Стратегия. После этого выбирается (рис. 11.23) стратегия объединения.

Разделяющая (дивизивная) стратегия. В случае дивизивной стратегии в том же меню необходимо указать число кластеров, на которое желательно разбить множество объектов, причем окончательное количество кластеров может получиться меньше этого числа, если затребованное разбиение для имеющихся данных невозможно.

Далее выдаются среднее внутрикластерное расстояние (по которому можно сравнивать различные варианты кластеризации текущих данных) и найденные кластеры с порядковыми номерами входящих в каждый кластер объектов, среди которых центральный по геометрическому положению объект отмечен звездочкой.

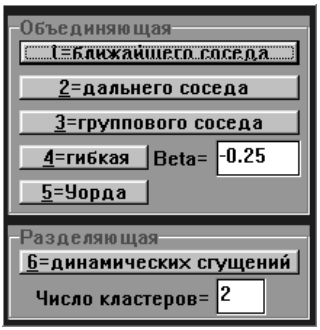


Рис. 11.23. Меню выбора стратегии классификации

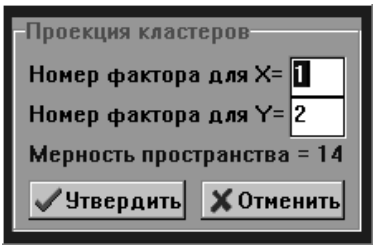


Рис. 11.24. Бланк установки переменных проекции кластеров

Затем в случае матрицы «*переменные–объекты*» запрашиваются порядковые номера двух переменных для определения плоскости проекции для выдачи графика кластеров (рис. 11.24), на котором объекты каждого кластера соединяются линиями с центральным объектом. Запросы плоскости проекции и выдача графиков повторяются до нажатия на кнопку «*Отменить*».

Полученную кластеризацию полезно затем статистически верифицировать методом дискриминантного анализа (разд. 11.3), для этого номера кластеров для объектов можно по подтверждению сохранить в электронной таблице.

Объединяющая (агломеративная) стратегия. В случае агломеративного метода задается вопрос о необходимости вывода диагональной матрицы расстояний между объектами, в которой строки будут соответствовать объектам (i = 2–m), а столбцы — объектам от 1 до i–1.

Далее производится выдача последовательности кластеров возрастающей общности с указанием номеров входящих в кластеры объектов и расстояния, на уровне которого произошло объединение каждого кластера.

В случае выбора гибкой стратегии в меню рис. 11.23 необходимо предварительно указать коэффициент «*бета*».

Дополнительные рекомендации по использованию стратегий даны ниже.

Дендрограмма. После этого строится *дендрограмма* — дерево объединения кластеров с порядковыми номерами объектов по горизонтальной оси и со шкалой расстояний по вертикальной оси.

Пропущенные значения в вычислениях не учитываются и их умеренное присутствие (до 20–30%), как правило, мало влияет на результат.

Ограничение: метод неприменим, если $n>l$ ($m>l$ для метрик, использующих коэффициент корреляции), где l = 500, 200, 89 при объеме матрицы данных в 64000, 20 000, 4000 чисел.

Пример 1

Задача. Произведено измерение четырех важных показателей для 20 популярных сортов немецкого пива (табл. 11.4, файл CLA).

Таблица 11.4. Основные показатели популярных сортов немецкого пива

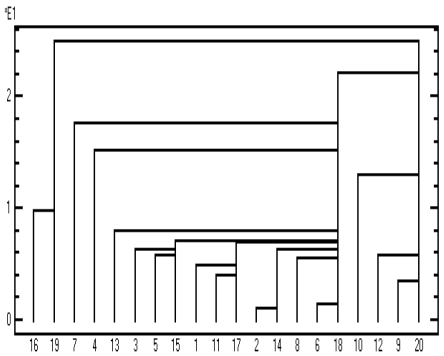
№	Сорт	Калории	Натрий	Алкоголь	Цена
1	Budweiser	144	15	4.7	0.43
2	Schlitz	151	19	4.9	0.43
3	Lowenbraw	157	15	4.9	0.48
4	Kronenbourg	170	7	5.2	0.73
5	Heineken	152	11	5	0.77
6	Old_Milwaukee	145	23	4.6	0.28
7	Augsberger	175	24	5.5	0.4
8	Strohs_Bonemain_Style	149	27	4.7	0.42
9	Miller_Lighr	99	10	4.3	0.43
10	Bodweiser_Light	113	8	3.7	0.44
11	Coors	140	18	4.6	0.44
12	Coors_Light	102	15	4.1	0.46
13	Mihelob_Light	135	11	4.2	0.5
14	Becks	150	19	4.7	0.76
15	Kirin	149	6	5	0.79
16	Pabst_Extra_Light	68	15	2.3	0.38
17	Hamms	136	19	4.4	0.43
18	Heilemans_Old_Stile	144	24	4.9	0.43
19	Olimpia_Gold_Light	72	6	2.9	0.46
20	Schlitz_Light	97	7	4.2	0.47

Необходимо произвести кластеризацию этих данных с использованием эвклидовой метрики для выявления имеющихся группировок. Это поможет рекламировать и потреблять более дешевые сорта пива, но близкие по показателям к более дорогим.

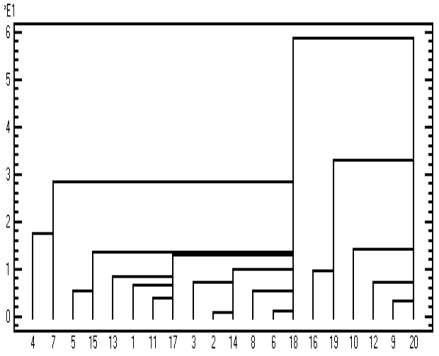
Результаты:

```
   КЛАСТЕРНЫЙ АНАЛИЗ.  Файл: cla.std        Эвклид+Ближ.сосед
            Таблица расстояний
         (1)       (2)       (3)       (4)       (5)       (6)       (7)
(  2)  8.06
(  3)    13     7.21
(  4)  27.2     22.5     15.3
(  5)  8.96     8.07     6.41     18.4
(  6)  8.06     7.22     14.4     29.7     13.9
```

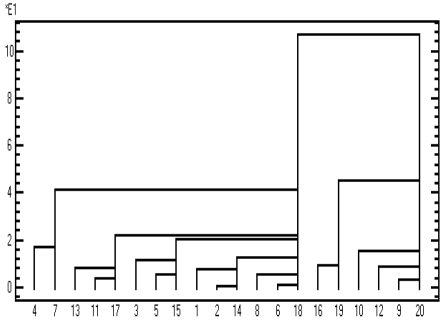
```
(  7)  32.3    24.5    20.1    17.7    26.4      30
(  8)    13    8.25    14.4      29    16.3    5.66    26.2
(  9)  45.3    52.8    58.2    71.1      53    47.8    77.3   52.8
                . . .
          К л а с т е р ы: (список объектов) -> расстояние
   (14,2) --> 1.07
   (18,6) --> 1.45
   (20,9) --> 3.61
   (17,11) --> 4.13
   (17,1,11) --> 5
   (18,8,6) --> 5.66
   (15,5) --> 5.83
   (20,12,9) --> 5.83
   (15,3,5) --> 6.41
   (18,14,2,8,6) --> 6.42
   (18,17,1,11,14,2,8,6) --> 7.07
   (18,15,3,5,17,1,11,14,2,8,6) --> 7.21
   (18,13,15,3,5,17,1,11,14,2,8,6) --> 8.07
   (19,16) --> 9.87
   (20,10,12,9) --> 13
   (18,4,13,15,3,5,17,1,11,14,2,8,6) --> 15.3
   (18,7,4,13,15,3,5,17,1,11,14,2,8,6) --> 17.7
   (20,18,7,4,13,15,3,5,17,1,11,14,2,8,6,10,12,9) --> 22.2
   (20,19,16,18,7,4,13,15,3,5,17,1,11,14,2,8,6,10,12,9) --> 25.1
```



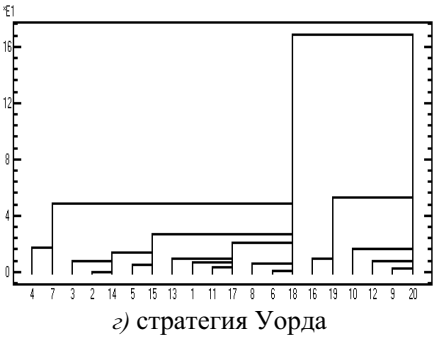
а) стратегия ближайшего соседа



б) стратегия группового соседа



в) стратегия дальнего соседа



г) стратегия Уорда

Рис. 11.25. Дендрограммы (по оси *Y* — расстояние объединения, по оси *X* — номера объектов)

Обсуждение результатов: Как видно из рис. 11.25, *а*, стратегия ближайшего соседа достаточно отчетливо выделяет три группы сортов пива: (1–3, 5, 6, 8, 11, 13–15, 17, 18); (9, 10, 12, 20); (16, 19); при этом сорта 7 и 4 несколько выпадают из этой классификации. Поэтому можно попытаться применить другую, растягивающую пространство, стратегию для более яркого разделения кластеров.

Для сравнения на рис. 11.25, *б–г* приведены дендрограммы, построенные с использованием стратегии группового и дальнего соседа и стратегии Уорда.

Как можно заметить из сравнения дендрограмм, стратегии группового соседа, дальнего соседа и Уорда по сравнению со стратегией ближайшего соседа дают все более четкое выделение кластеров и несколько отличные классификации. Так на дендрограммах рис. 11.25, *в, г* можно выделить 3–5 кластеров. В связи с этим далее применим дивизивную стратегию с эвклидовой метрикой, чтобы сравнить группировки исследуемых объектов на различное число кластеров.

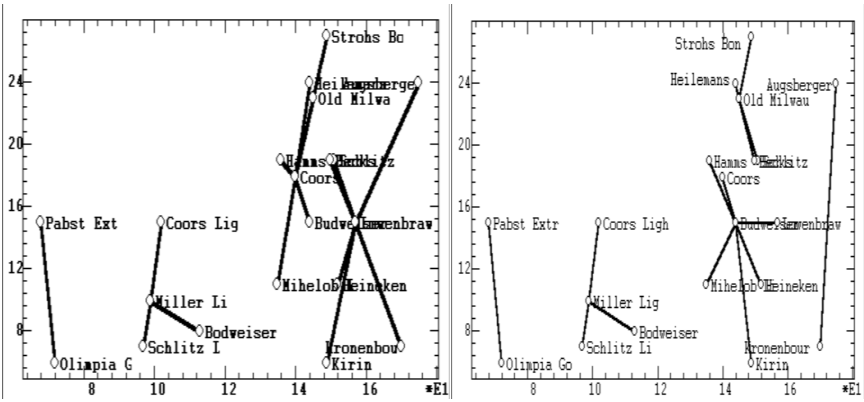
Р е з у л ь т а т ы:

```
   КЛАСТЕРНЫЙ АНАЛИЗ.  Файл: cla.std          Эвклид+Дивизивная
К л а с т е р ы:
Среднее внутрикластерное расстояние=9.81
1= (Miller Li*,Bodweiser,Coors Lig,Pabst Ext,Olimpia G,Schlitz L)
2= (Schlitz,Lowenbraw*,Kronenbou,Heineken,Augsberge,Becks,Kirin)
3= (Budweiser,Old Milwa,Strohs Bo,Coors*,Mihelob
    L,Hamms,Heilemans)

К л а с т е р ы:
Среднее внутрикластерное расстояние=7.37
1= (Pabst Ext*,Olimpia G)
2= (Schlitz,Lowenbraw*,Kronenbou,Heineken,Augsberge,Becks,Kirin)
3= (Budweiser,Old Milwa,Strohs Bo,Coors*,Mihelob
    L,Hamms,Heilemans)
4= (Miller Li*,Bodweiser,Coors Lig,Schlitz L)

К л а с т е р ы:
Среднее внутрикластерное расстояние=6.4
1= (Pabst Ext*,Olimpia G)
2= (Schlitz,Old Milwa*,Strohs Bo,Becks,Heilemans)
3= (Budweiser*,Lowenbraw,Heineken,Coors,Mihelob L,Kirin,Hamms)
4= (Miller Li*,Bodweiser,Coors Lig,Schlitz L)
5= (Kronenbou*,Augsberge)
```

Обсуждение: Как показывают числовые результаты, переход от трех к четырем кластерам влечет существенное повышение компактности группировки, о чем свидетельствует значительное уменьшение среднего внутрикластерного расстояния (с 9,81 до 7,37). Переход же от четырех к пяти кластерам сопровождается более чем в 2 раза меньшим сокращением внутрикластерного расстояния (с 7,37 до 6,4). Тем самым группировку на три кластера следует признать неудачной из-за большого внутрикластерного расстояния.



а *б*

Рис. 11.26. Дивизивная кластеризация сортов пива в проекции на плоскость *калории* (ось *X*)—*натрий* (ось *У*): *а* — на 4 кластера; *б* — на 5 кластеров

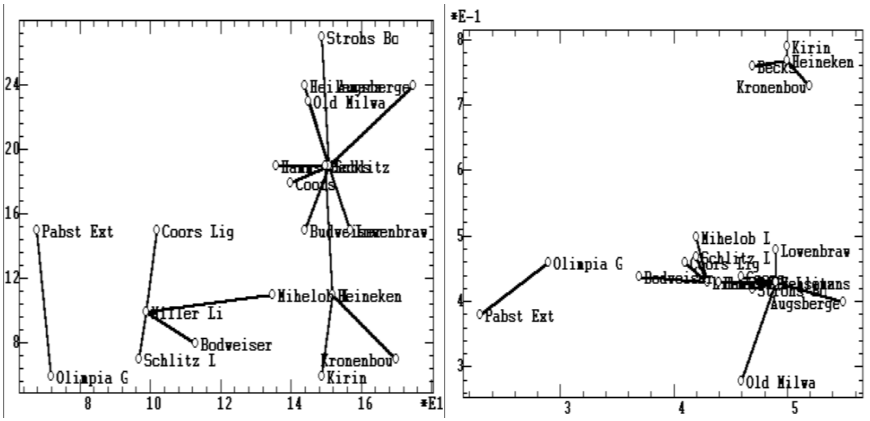
Чтобы выбрать лучшую группировку из оставшихся двух, близких по внутрикластерному расстоянию, обратимся к визуальному анализу соответствующих им диаграмм кластеризации (рис. 11.26). Из сравнения диаграмм видно, что разбиение на пять кластеров (рис. 11.26, *б*) менее логично, поскольку приводит к выделению кластера из двух, очень далеко отстоящих друг от друга сортов Kronenbou и Augsberge.

Из визуального исследования диаграмм можно также вывести еще одно интересное наблюдение: кластеризация сортов пива идет преимущественно по оси калорий. В чем здесь дело? Обратившись к нашим исходным данным (табл. 11.4), можно заметить существенную разницу в значениях переменных: значение переменной *калории* в 10 раз превышают значения переменной *натрий* и еще больше превышают значения других переменных. Поэтому и удаленность объектов по шкале калорий подавляюще превалирует при кластеризации по сравнению с удаленностью по другим шкалам. Чтобы сделать переменные равноправными, следует вместо эвклидовой метрики использовать нормированную эвклидову метрику.

Р е з у л ь т а т ы:

```
КЛАСТЕРНЫЙ АНАЛИЗ.  Файл: cla.std  Норм.Эвклид.+Дивизивная
                        К л а с т е р ы:
Среднее внутрикластерное расстояние=0.813
1= (Pabst Ext*,Olimpia G)
2= (Kronenbou,Heineken*,Becks,Kirin)
3= (Budweiser,Schlitz*,Lowenbraw,Old Milwa,Augsberge,Strohs
    Bo,Coors,Hamms,Heilemans)
4= (Miller Li*,Bodweiser,Coors Lig,Mihelob L,Schlitz L)
```

Обсуждение: Как видно из диаграмм кластеризации (рис. 11.27) теперь группировка сортов пива идет не только по шкале калорий, но и по трем другим шкалам, тем самым более адекватно учитывая вклады всех переменных.



а *б*

Рис. 11.27. Результаты дивизивной кластеризация в нормированной эвклидовой метрике: *а* — плоскость *калории–натрий; б* — плоскость *алкоголь–цена*

Сравнивая две диаграммы рис. 11.26, *а* и 11.27, *а*, следует отметить малое изменение двух кластеров в области небольших калорийностей и заметную перегруппировку двух кластеров в области высоких калорийностей.

Далее можно задаться вопросом, а все ли переменные одинаково значимы для кластеризации, нет ли среди них второстепенных, которые можно было бы не учитывать. Например, могут быть пары переменных, связанных четкой функциональной зависимостью, тогда из такой пары можно в анализе учитывать только одну переменную, поскольку она однозначно определяет значения другой переменной. Для поиска таких связанных переменных можно было бы воспользоваться корреляционным анализом, выбрать пары с высокими коэффициентами корреляции, а затем рассмотреть их диаграммы рассеяния для визуальной оценки степени функциональной зависимости. Однако можно поступить проще — провести еще одну кластеризацию, но уже не объектов, а переменных. Для этого надо просто использовать метрику коэффициентов корреляции. При этом выберем стратегию ближайшего соседа, сжимающую пространство, чтобы на дендрограмме визуально выделялись только очень близкие друг к другу группировки.

Р е з у л ь т а т ы:

```
КЛАСТЕРНЫЙ АНАЛИЗ.  Файл: cla.std Корреляция+Ближ.сосед
                        К л а с т е р ы:
                  (список объектов) -> расстояние
(3,1) --> 0.079
(3,2,1) --> 0.588
(4,3,2,1) --> 0.668
1=калории, 2=натрий, 3=алкоголь, 4=цена
```

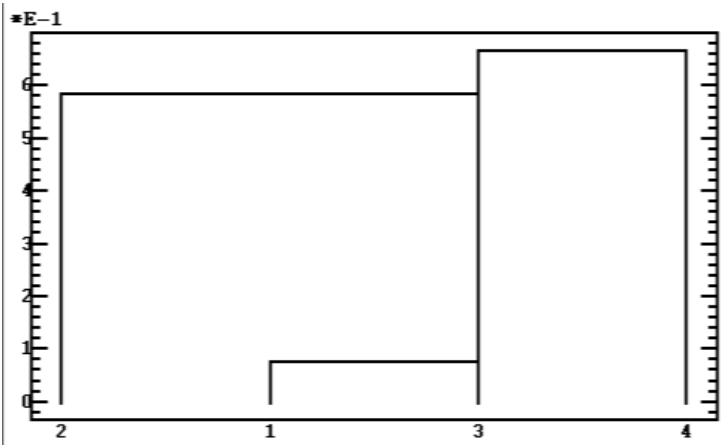


Рис. 11.28. Дендрограмма классификации переменных: по оси *Y* — расстояние, по оси *X* — номер переменной

Обсуждение: Как видно из числовой выдачи и дендрограммы (рис. 11.28), две переменные 1 и 3 (содержание калорий и алкоголя) очень близки друг к другу (коррелированы): они объединены на расстоянии 0,079, тогда как следующие два объединения происходят на расстояниях в 7 с лишним раз больших. Поэтому одну из этих переменных можно было бы вполне исключить из рассмотрения, повторить кластерный анализ и сравнить результаты.

Интересно было бы также провести факторный анализ, выяснить, какие факторы являются главными, вычислить координаты объектов в пространстве главных факторов и уже в этом пространстве провести новую кластеризацию и сравнить полученные результаты.

Предлагаем эти задачи в качестве учебной практики читателям. Дополнительные примеры разнопланового применения кластерного анализа рассмотрены в гл. 14. Анализ данного же примера будет продолжен в следующем разделе.

Пример 2

Задача. Возвратимся к данным тестирования профессиональной пригодности примера 3 к разд. 11.1 и к результатам их факторного анализа. В частности, там были обнаружены практически совпадающие по главным факторам две пары переменных и четыре плотные группы тестируемых. Интересно было бы методом кластерного анализа проверить эти результаты. Используем метрику коэффициента связности Крамера и стратегию дальнего соседа (для более четкого разделения кластеров).

Результаты:

```
КЛАСТЕРНЫЙ АНАЛИЗ.  Файл: tests.std
                      Крамера+Дальн.сосед
```

```
                    К л а с т е р ы:
                (список переменных) -> расстояние
(5,3) --> 0.216  увереннос,прям.диал
(8,6) --> 0.225  сотруднич,команд.ду
(4,1) --> 0.247  алгоритмы,память
(4,2,1) --> 0.29 алгоритмы,мат.задач,память
(10,7) --> 0.322 убеждение,находчиво
(9,8,6) --> 0.358 признание,сотруднич,команд.ду
(10,5,3,7) --> 0.379 убеждение,увереннос,прям.диал,находчиво
(10,4,2,1,5,3,7) --> 0.664
убеждение,алгоритмы,мат.задач,память,увереннос,прям.диал,находчиво
(10,9,8,6,4,2,1,5,3,7) --> 0.718
```

В ы в о д ы: Как видно из числовой выдачи, кластерный анализ подтвердил близость только одной пары тестов: *командный дух* и сотрудничество, которые были объединены на втором шаге кластеризации с расстоянием 0,225. Что касается теста *уверенность в выступлении*, то он объединен не с тестом на *память*, а с *находчивостью в прямом диалоге*, тест же *память* оказался близок к *составлению алгоритмов*.

Такие изменения становятся понятными, если мы возвратимся к графикам факторных нагрузок (рис. 11.20, *а*), на которых видно, что пара тестов *уверенность в выступлении* — *память*, непосредственно соседствует именно с *находчивостью в прямом диалоге* и *составлением алгоритмов*. Смещение связей объясняется тем, что кластерный анализ использует всю информацию о тестах, в то время как проекция нагрузок на плоскость первых двух факторов (рис. 11.20, *а*), как это следует из процента накопленной дисперсии в таблице факторных нагрузок, отражает только 70,6% исходной информации, поэтому в проекциях на другие факторные плоскости эти тесты могут сочетаться в других комбинациях.

Продолжение анализа. Теперь можно перейти к классификации претендентов с использованием той же метрики Крамера. Такое применение метрики допустимо, поскольку переменные однородны (все являются тестами) и число градаций их значений одинаково. Перед повторением кластерного анализа надо предварительно транспонировать матрицу данных, чтобы претенденты располагались по столбцам, а тесты — по строкам (поскольку, как было отмечено выше, метрика Крамера по умолчанию классифицирует переменные, а не объекты).

В ы в о д ы: На полученной дендрограмме (рис. 11.29) можно выделить три-четыре кластера. Выпишем эти кластеры и сравним с визуально выделенными группами в факторном анализе (рис. 11.20, *б*)

```
КлАн: {5,8,13,16} {6,15,17} {3,9,11,14,18} {1,2,4,7,10,12}
ФаАн: {5,7,8,13,16} {2,6,10,17} {3,9,11,14,18} {1,4,12,15}
```

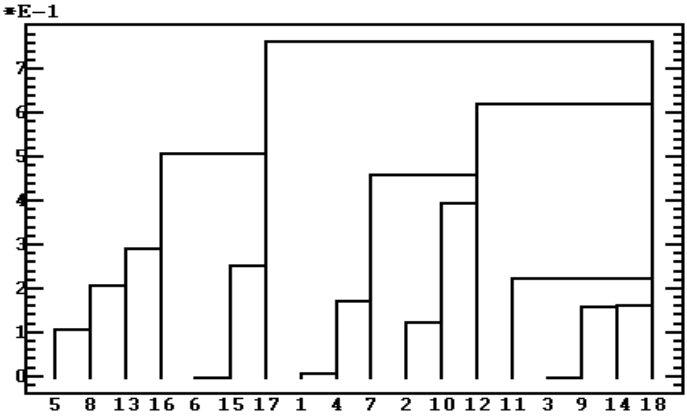


Рис. 11.29. Дендрограмма классификации претендентов

Как можно заметить, четыре группы претендентов в значительной степени сохранили свой состав. Перемещение отдельных претендентов (между первой, второй и четвертой группами, они отмечены жирным шрифтом) можно также объяснить полным использованием в кластерном анализе исходной информации.

Здесь особо следует подчеркнуть, что упомянутое полное использование информации может иметь и свои негативные стороны, поскольку вся (нерафинированная) информация включает и случайные (шумовые) компоненты, которые могут искажать или смещать главные закономерности.

Последующие исследования.

1. Провести кластерный анализ с использованием дивизивной стратегии с разбиением на четыре кластера и сравнить результаты.
2. Провести кластерный анализ с использованием корреляций Спирмана и Кенделла и оценить различие или сходство в результатах.
3. Исходные данные преобразовать к бинарному виду (аналогично примеру 3 к разд. 11.1). Провести кластерный анализ с использованием различных формул бинарных коэффициентов и оценить различие или сходство в результатах.

Оставляем эти задачи в качестве учебных читателям.

11.3. Дискриминантный анализ

Назначение. Дискриминантный анализ позволяет:

- проверить гипотезу о непротиворечивости предполагаемой классификации заданного множества n объектов на k классов в m–мерном пространстве переменных X_j; $j = 1–m$;
- классифицировать новые объекты.

В ходе вычислений ищется набор дискриминирующих функций d_l, обеспечивающих классификацию объектов на заданное число k классов:

$$d_l = b_{l0} + b_{l1}X_1 + ... + b_{lm}X_m\,, \quad l = 1, ..., k\,.$$

При отнесении объекта к классу его координаты подставляются в дискриминирующие функции всех классов. Класс объекта определяется максимумом из полученных значений дискриминирующих функций.

Исходные данные представляются в виде матрицы размером $(m+1)n$, в которой первые m столбцов содержат значения m переменных для n объектов, а $m+1$–я переменная в качестве своих значений содержит для каждого объекта номер его класса (натуральные числа от 1 до k, где k — число классов). Объекты (строки) в матрице могут быть неупорядочены относительно номеров классов.

Если кроме вычисления дискриминантной функции нужно с ее помощью классифицировать ряд новых объектов, то такие объекты также исходно включаются в матрицу данных с номером класса 0.

Результаты. Выдача результатов включает:

- суммарное межкластерное расстояние Махаланобиса D^2 (*Mahalanobis*) между классами с уровнем значимости P для нулевой гипотезы «$D^2=0$» (см. ниже в примечаниях) по хи–квадрат критерию от вычисленного значения D^2 с $m(k–1)$ степенями свободы;
- коэффициенты дискриминирующей функции, обеспечивающей отнесение объектов к данному классу, отдельно для каждого класса;
- таблицу, где для каждого объекта l указываются:
  - номер его класса r;
  - расстояние Махаланобиса Dl^2 (от объекта до центра класса);
  - уровень значимости P_l нулевой гипотезы «$Dl^2=0$» (*объект может быть отнесен к данному классу*) по критерию хи–квадрат от вычисленного значения Dl^2 с m степенями свободы;
  - апостеорная вероятность p_l отнесения объекта к его классу.

Если $P>0.05$, соответствующая нулевая гипотеза отнесения объекта к соответствующему классу может быть принята. Для вновь классифицируемых объектов их номера классов отмечаются символом «звездочка». Для объектов, априорный класс которых изменен, ставится восклицательный знак.

По подтверждению можно выдать не полную, а сокращенную таблицу результатов, в которой указываются только вновь классифицированные объекты, объекты, изменившие свой класс, и объекты с уровнем значимости нулевой гипотезы «$Dl^2=0$» ниже критического. Это удобно для повышения обозримости результатов в случае большого числа объектов.

Пример 1

Задача. Проведем дискриминантную верификацию результатов дивизивной классификации сортов немецкого пива из примера к разд. 11.2. Для приведения исходных данных к требуемому в данном случае виду при выполнении кластерного анализа подтвердим необходимость сохранения номеров кластеров в электронной таблице. Начнем с четырех кластеров, выделенных по эвклидовой метрике (рис. 11.26, *а*).

Результаты:

```
ДИСКРИМИНАНТНЫЙ АНАЛИЗ.  Файл: cla.std

    Расстояние Махаланобиса=84.9, значимость=1.28E-5
   Гипотеза 1: <Межкластерное расстояние отлично от нуля>
    Класс  <-Коэффициенты дискриминантной функции: b[0],b[1],...->
        1      -100      1.06     0.934      28.7      99.9
        2      -383       2.5      1.21      49.3       169
        3      -299      2.17       1.2      45.3       144
        4      -200      1.42     0.848      47.3       117

    Объект   Класс       D^2     Значим  Апостеор.вероят.
Budweiser        3       1.2      0.878        0.991
  Schlitz        3!     2.31      0.678        0.821
Lowenbraw        2     2.79      0.593        0.853
Kronenbou        2     6.23      0.182            1
 Heineken        2     1.82      0.768            1
Old Milwa        3     1.48       0.83            1
Augsberge        2     7.12       0.13            1
Strohs Bo        3     3.53      0.474        0.962
Miller Li        4     1.89      0.756            1
Bodweiser        4     7.53       0.11            1
    Coors        3    0.143      0.998        0.998
Coors Lig        4     1.42       0.84            1
Mihelob L        3     3.91      0.418            1
    Becks        2      6.6      0.158        0.997
    Kirin        2      3.2      0.525        0.999
Pabst Ext        1     2.34      0.673            1
    Hamms        3    0.772      0.942            1
Heilemans        3     2.61      0.626        0.979
Olimpia G        1     2.34      0.673            1
Schlitz L        4      1.7      0.791            1
```

Обсуждение результатов: Как показала проверка гипотезы о равенстве нулю расстояния Махаланобиса «D^2=0» (ее уровень значимости очень низок — 0,0000128), кластеры достаточно компактны и хорошо разделены. Для большинства сортов пива гипотеза о нулевых расстояниях до центров соответствующих кластеров «Dl^2=0» принимается с очень высокими уровнями значимости (от 0,417 до 0,998). Можно отметить сравнительно меньшие значимости для Kronenbou, Augsberge, Becks, что можно объяснить их принадлежностью ко второму достаточно разбросанному классу (рис. 11.26, *а*, правый кластер в области высоких калорийностей).

Обращает также на себя внимание изменение класса пива Schlitz с второго на третий. Судя по диаграмме рис. 11.26, *а*, этот сорт действительно находится на краю кластера 2 в непосредственной близости от

граничных объектов кластера 3. Здесь проявилось определенное различие в алгоритмах кластерного и дискриминантного анализа. Дивизивные группировки кластерного анализа базируются на выделении центрального объекта в каждом кластере и поиске близлежащих к нему объектов. При этом положение такого объекта в общем не совпадает с геометрическим центром кластера. Дискриминантный же анализ оценивает группировки по максимуму дискриминирующей функции, положение которого в ряде случаев более соответствует геометрическому центру кластера.

В качестве учебной задачи оставляем читателям дискриминантный анализ результатов кластеризации в нормированной эвклидовой метрике (рис. 11.27) и сравнения результатов с вышерассмотренными. Отметим только что в этом случае смен кластеров наблюдаться не будет.

Пример 2

Таблица 11.5. Метрические данные 45 птиц

№	Крыло	Голова	Клюв	Лапы	Вес	Пол
1	25.8	9.5	4.2	14.6	380	0
2	25.1	9.6	3.8	14.5	355	0
3	24.8	9.4	3.8	15	355	1
4	25.9	9.3	3.9	15.5	375	0
5	25.8	9.6	4.5	15.6	360	2
6	24.8	9.4	4.4	14.7	355	1
7	25.6	9.6	4.1	14.5	356	0
8	25.3	9.4	4.4	15.4	360	2
9	25.5	9.9	4	15.1	330	2
10	24.2	8.6	3.9	13.6	350	1
11	25	9.3	3.9	15.1	338	2
12	24.7	9.3	3.8	14.5	335	0
13	24.3	9.7	4	13.9	310	2
14	25.1	9.3	3.9	15.2	332	2
15	24.7	8.9	3.8	15.1	340	2
16	25.5	9.4	4.1	15.2	333	0
17	25.7	9.7	3.9	13.8	323	2
18	25.9	9.7	4	15.6	325	2
19	25.1	9.4	4.6	14.9	325	0
20	24.7	8.6	3.9	13.9	345	1
21	25.3	8.8	3.9	14.3	340	1
22	26.1	9.3	3.8	15	327	1
23	26	9.4	3.9	14.3	316	0
24	24.8	8.4	4	14.4	336	1
25	25.6	8.8	4.2	15	330	2
26	24.6	8.4	4	14.8	330	2
27	25.1	8.6	3.8	14.4	329	0
28	25.1	8.3	3.9	14.5	338	1
29	24.9	8.9	3.8	14.6	314	0
30	25.3	8.4	3.8	13.9	335	1
31	24.7	9.3	4	15.4	285	0
32	24.7	8.2	3.7	14.3	320	1
33	25.4	8.6	4	13.8	310	1
34	25.7	8.6	3.9	13.8	315	2
35	23.8	8.4	3.7	14.1	300	1
36	24.8	8.8	3.9	14.3	290	0
37	25.3	8.4	4.1	14.3	305	1
38	24.9	8.6	3.8	14	292	1
39	25.9	8.3	3.8	13.8	317	0
40	25	8.2	3.8	13.3	305	1

41	26	8.5	3.7	13.7	300	2
42	25	8.4	3.8	13.3	290	1
43	25.7	8.4	3.9	14.8	300	1
44	24.8	8	3.6	13.2	294	0
45	25.3	8.2	3.6	13.3	292	1

З а д а ч а. В исследовании одного вида птиц фиксировались их метрические данные и половая принадлежность: 1 — самец, 2 — самка (табл. 11.5, файл BIRDS[1]). Последний признак у ряда особей по разным причинам не удалось определить (значение 0 в столбце «Пол» табл. 11.5). Поэтому встала задача найти классифицирующую функцию, по которой можно было бы приписать пол неопределенным особям.

Р е з у л ь т а т ы (сокращенно):

```
ДИСКРИМИНАНТНЫЙ АНАЛИЗ.  Файл: birds.std

    Расстояние Махаланобиса=2.46, значимость=0.783
Гипотеза 0: <Межкластерное расстояние не отлично от нуля>
    Класс  <- Коэффициенты дискриминантной функции:a[0],a[1],...->
        1  -1.37E3     90.2     10.6     17.1     6.48    0.732
        2  -1.42E3     90.8     13.4     18.2     7.56    0.702

    Объект   Класс      D^2   Значим Вероят.отнесения
        1        2*     12.9    0.024    0.699
        2        2*     7.64    0.177    0.718
        3        2!     5.98    0.308     0.68
        4        2*     12.7   0.0263    0.752
        6        2!     6.34    0.274    0.754
        7        2*      5.1    0.403     0.82
       12        2*     3.29    0.656    0.618
       13        2      11.2   0.0467    0.848
       15        1!     4.61    0.465    0.537
       16        2*    0.672    0.984    0.912
       19        2*     14.8   0.0111    0.932
       22        2!      4.6    0.466     0.88
       23        2*     3.36    0.645    0.871
       26        1!     3.24    0.662    0.801
       27        1*    0.858    0.973    0.769
       29        2*     2.21     0.82    0.549
       31        2*     18.6  0.00228     0.96
       34        1!     2.39    0.793     0.73
       36        2*     6.31    0.277    0.592
       39        1*     4.33    0.503     0.87
       41        1!     4.94    0.423    0.732
       44        1*     3.31    0.653    0.971
```

Обсуждение результатов: Как показала проверка гипотезы о равенстве нулю расстояния Махаланобиса (ее уровень значимости очень высок), кластеры или некомпактны, или же плохо разделены. Для окончательного вывода следует обратиться к анализу значимостей нулевых гипотез для объектов.

Из изучения сокращенной таблицы результатов можно сделать следующие выводы:

[1] Данные с сокращениями из архива SPSS.

1) семь из 31 ранее классифицированных объектов (№ 3, 6, 15, 22, 26, 34, 41) были отнесены к противоположенному классу с высокими уровнями значимости, а для объекта №13 значимость отнесения к классу 2 оказалась ниже критической;
2) для четырех из 14 вновь классифицированных объектов (№ 1, 4, 19, 31) значимость отнесения к соответствующему классу оказалась также ниже критической.

В процентном отношении отмеченные отклонения не слишком велики и не могут являться достаточным основанием для признания всей классификации неудачной. Однако они могут служить основанием для более глубокого изучения ситуации. Во-первых, следует посмотреть, нет ли погрешностей в самой методике сбора исходных данных, т. е. метрологических или субъективных погрешностей (это за пределами наших возможностей, поскольку сборщики данных недоступны). Во-вторых, полезно визуально изучить пространственное распределение измерений, чтобы сделать вывод о хорошем или плохом их разделении по половому признаку. Для этого следует провести факторный анализ с выдачей диаграммы рассеяния объектов на плоскость первых двух главных компонент. Оставляем эту задачу в качестве учебной читателям.

11.4. Шкалирование

Во многих областях исследования (в психологии, социологии, биологии, лингвистике и других) часто затруднительно или невозможно произвести непосредственное измерение переменных, характеризующих изучаемую популяцию объектов, а можно лишь экспертным или каким–то другим образом оценить взаимную близость или же различия между парами объектов. В то же время для детального анализа популяции объектов (в том числе и рассмотренными выше многомерными методами) желательно оперировать именно с числовыми переменными, характеризующими каждый объект индивидуально.

Назначение. Задачей данного метода является построение метрического пространства небольшой размерности, в которое может быть погружен многомерный граф, узлы которого составляют объекты, а длина ребер пропорциональна расстоянию между объектами.

Исходные данные представляются в виде квадратной матрицы взаимных расстояний d_{ij} между n объектами (в этой матрице может быть заполнена только левая нижняя половина, а остальная часть может быть заполнена нулями).

Действия и результаты. Анализ протекает в следующей последовательности.

1. Начальное приближение. Вычисляется начальное приближение метрического пространства расположения объектов методом главных

компонент (аналогично разд. 11.1) и на экран выводится таблица, где для каждого пространственного измерения приводятся:

- собственное значение, пропорциональное части общей дисперсии экспериментальных данных, приходящейся на данное изменение;
- процент полной дисперсии, приходящейся на каждое измерение;
- процент накопленной дисперсии.

Малозначительные измерения, собственные значения которых составляют менее 2% от накопленной дисперсии, в выдаче опускаются.

2. Методы шкалирования. Далее нужно выбрать метод шкалирования из следующих возможностей (рис. 11.30):

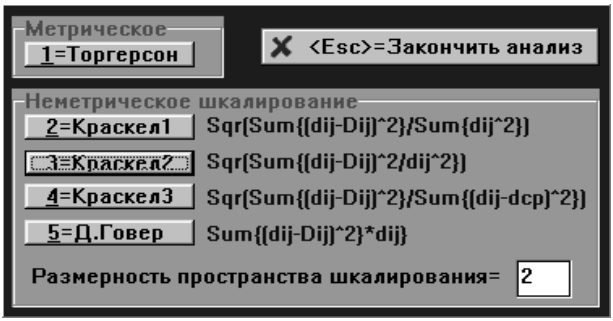


Рис. 11.30. Меню выбора метода шкалирования

- метрическое шкалирование по методу Торгерсона (*Torgerson*);
- неметрическое шкалирование по методам Шепарда–Крускала (*Shepard–Kruskal*) и Говера (*Gover*), которые различаются формулами оценки стресса (стресс — показатель невязки или несовпадения между исходными и вычисленными различиями между объектами, выступающий в качестве функционала алгоритма минимизации);
- закончить анализ.

Метрики. В случае неметрического шкалирования необходимо указать размерность пространства шкалирования (не превышающего количества рассчитанных на первом этапе компонент) и выбрать метрику или метод вычисления расстояний между объектами (рис. 11.31).

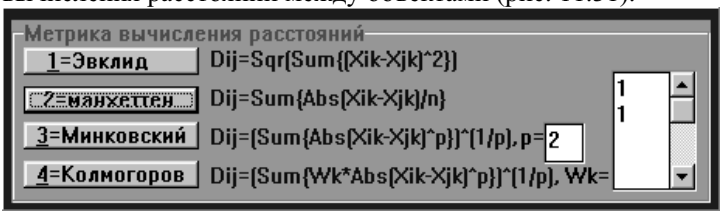


Рис. 11.31. Меню выбора метрики шкалирования

В случае метрик Минковского и Колмогорова в поле ввода необходимо указать показатель степени метрики $0<p<5$, а в случае метрики Колмогорова необходимо ввести значения весов для координатных осей $w_k>0$.

3. Результаты. Основными результатами анализа являются:

- таблица координат объектов;
- рисунок проекции объектов на плоскость двух указанных в бланке (см. рис. 11.7) координатных осей, при этом запросы очередной плоскости проекции повторяются до отмены бланка;
- значения стресса, вычисленные по различным формулам.
- коэффициент корреляции между исходными расстояниями (близостями) объектов и расстояниями между ними в построенном пространстве шкалирования (в случае метрического шкалирования выдается последовательность коэффициентов корреляции для 2-мерного, 3-мерного и т. д. пространств).

Вычисленные координаты объектов с рисунка можно сохранить в электронной таблице для последующего анализа (например, для кластеризации, сравнения различий, регрессионного анализа и т. п.) нажатием на инструментальную графическую кнопку «*СохрГраф*».

После этого анализ может быть повторен с п.2.

Ограничение. Метод неприменим, если $m>l$, где l = 250, 140, 62 при объеме матрицы данных в 64000, 20000, 4000 чисел.

Пример 1

З а д а ч а. В сентябре 1990 г. был проведен опрос по оценке в трехбалльной шкале взаимных различий между восемью ведущими в то время политическими деятелями (табл. 11.6, файл MSC).

Таблица 11.6. Различия между популярными политическими лидерами 1990 г.

Popov	Gidaspov	Gorbi	Bush	Eltsin	Lansberg	Ligachev	Sobchak
0	-	-	-	-	-	-	-
2	0	-	-	-	-	-	-
2	1	0	-	-	-	-	-
1	2	1	0	-	-	-	-
1	2	2	2	0	-	-	-
1	3	2	1	2	0	-	-
3	1	2	3	2	3	0	-
1	3	1	1	1	1	3	0

Требуется определить основные социологические и политические факторы, действующие на мнение респондентов, а также исследовать взаимное положение политических лидеров в пространстве этих факторов.

Произведем непараметрическое шкалирование этих данных в эвклидовой метрике с минимизацией по стрессу 1 Краскела в 2-мерном пространстве и с выдачей графика проекции на плоскость первых двух компонент.

Р е з у л ь т а т ы:

```
МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ.  Файл: msc.std

              Собственные значения координатных осей шкалирования
Координата:       1         2         3         4         5         6
Собств.зн      9.92      3.22      2.14      1.39
 Дисперс%      59.5      19.3      12.8      8.32
Накоплен%      59.5      78.8      91.7       100

        Метод: Краскел1+Эвклид
           Координаты объектов в результате шкалирования
Координата:     1         2         3         4         5         6
    Popov     0.764    -0.651
 Gidaspov     -1.44     0.463
    Gorbi    -0.338     0.807
     Bush     0.667     0.732
   Eltsin   -0.0465     -1.08
 Lansberg      1.49       0.2
 Ligachev     -1.97    -0.382
  Sobchak     0.873   -0.0874
Стресс 1-4 = 0.132; 1.16, 0.338, 3.54
```

Обсуждение: Как следует из числовой выдачи, исходные данные уверенно погружаются в 4-мерное пространство, в котором первые два фактора покрывают 79% дисперсии. Проекция объектов на плоскость первых двух факторов (рис. 11.32) с учетом политической ориентации лидеров позволяет достаточно уверенно интерпретировать первую главную ось (абсцисс) как «*демократичность*» или «*популярность*», а вторую — как «*признанность в качестве официального политического деятеля*».

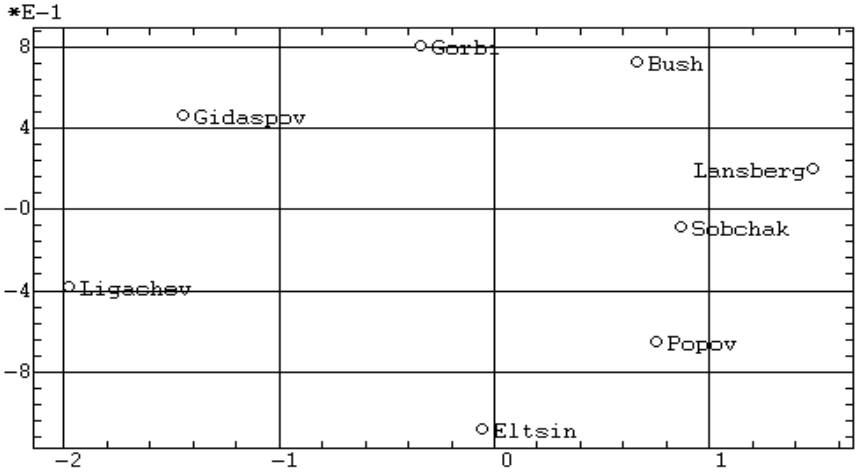


Рис. 11.32. Расположение политических лидеров 1991 г. в плоскости первых двух факторов

Для сравнения приведем результаты шкалирования методом Торгерсона (рис. 11.33). Заметное отличие заключается только в сближении Попова и Собчака. Отметим, что для данного примера формула оценки

стресса Говера дает результаты, близкие к Торгерсону, а все три формулы стресса Краскела дают практически идентичные результаты.

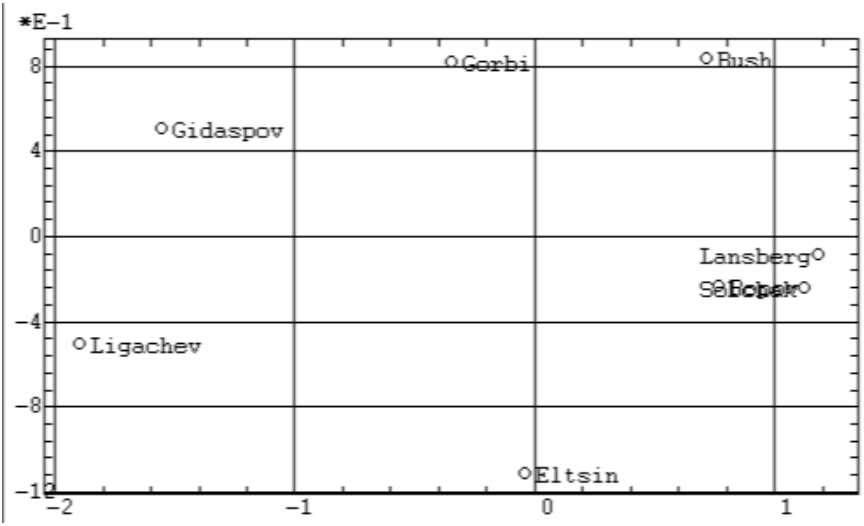


Рис. 11.33. Результаты метрического шкалирования методом Торгенсона

Пример 2

Проверим эффективность процедуры многомерного шкалирования на следующем наглядном примере. Возьмем карту СССР и на ней выберем координаты крупных городов. Далее измеряем взаимные расстояния между всеми парами городов и округляем их со случайными погрешностями, дабы сделать граф расстояний неплоским (табл. 11.7, файл TOWN).

Таблица 11.7. Приблизительные расстояния между городами СССР

Москва	Киев	СПБ	Архан	Саратов	Пермь	Краснод	Минск	Казань	Астрах
0	-	-	-	-	-	-	-	-	-
47	0	-	-	-	-	-	-	-	-
39	65	0	-	-	-	-	-	-	-
62	105	47	0	-	-	-	-	-	-
45	68	83	92	0	-	-	-	-	-
72	116	93	69	60	0	-	-	-	-
75	54	109	135	56	116	0	-	-	-
42	27	43	87	78	113	80	0	-	-
45	85	74	67	32	32	87	86	0	-
81	86	119	130	39	78	45	105	66	0

Обработаем полученную матрицу взаимных расстояний процедурой многомерного шкалирования и сравним полученный результат с расположением городов на карте СССР.

В ы в о д ы: Результаты шкалирования (рис. 11.34) вполне соответствуют реальному распределение городов на карте СССР.

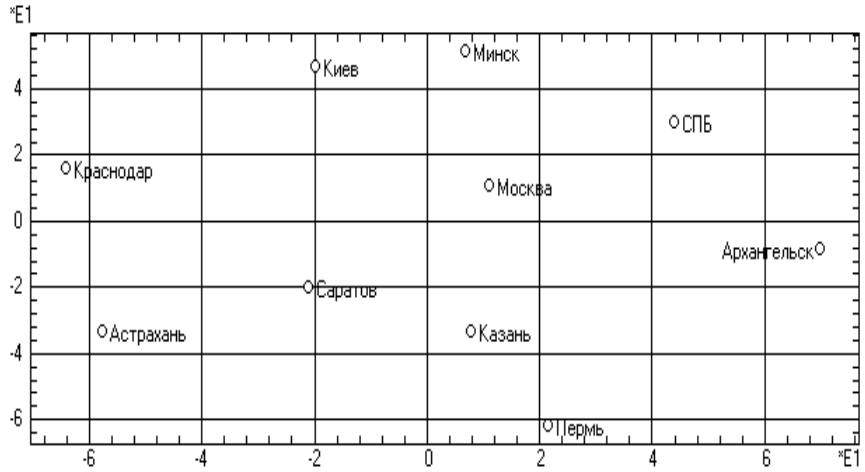


Рис.11.34. Расположение городов СССР (запад вверху)

Глава 12

ВЕРОЯТНОСТИ И ЧАСТОТЫ

«В–Себе–и–для–Себя–Сущий–Дух»
[Гегель. Феноменология абсолютного духа]

В данной главе рассматриваются специальные разделы математической статистики, связанные с вычислением вероятности событий для различных законов распределения (разд. 12.2), критерии согласия эмпирического и теоретического распределений (разд. 12.3), согласия частот событий (*долей*, разд. 12.4), последовательный анализ отклонений наблюдаемой частоты от заданного уровня (разд. 12.5) и анализ кривых выживаемости (разд. 12.6).

12.1. Случайные величины и распределения

В теории вероятностей различают два основных класса случайных величин:

а) дискретные, множество значений которых представляет собой конечную или счетную последовательность;

б) непрерывные, значения которых принадлежат к некоторому диапазону и могут отличаться друг от друга на сколь угодно малую величину.

12.2. Вычисления вероятностей

Назначение. В данном разделе рассмотрены вычисления значений функции вероятности наиболее употребительных в практике дискретных и непрерывных распределений.

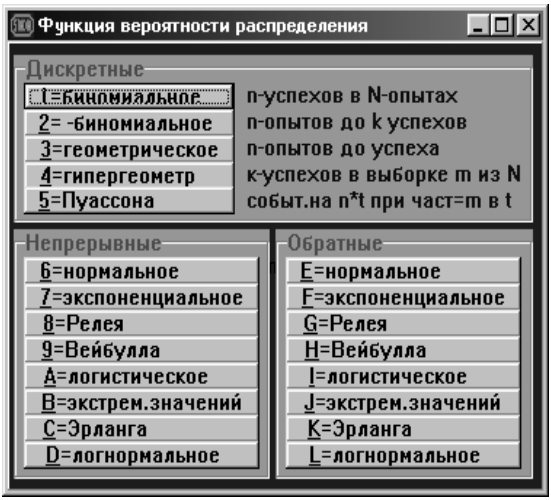


Рис. 12.1. Меню выбора распределения

Действия и результаты.

1. Сначала следует выбрать тип дискретного или непрерывного распределения (меню — рис. 12.1), после чего ввести параметры распределения. Для каждого непрерывного распределения в меню имеются два варианта: прямое и обратное распределение.

2. Прямое распределение. В случае прямого распределения необходимо указать значения одного или двух параметров распределения (рис. 12.2).

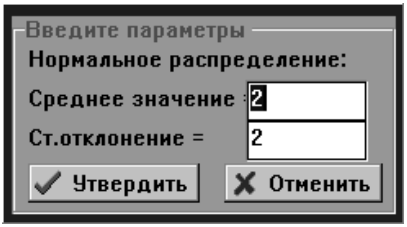


Рис. 12.2. Бланк установки параметров прямого распределения

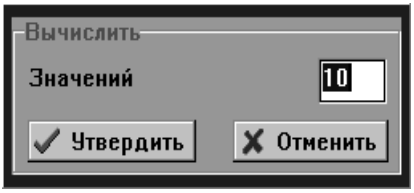


Рис. 12.3. Бланк установки числа вычисляемых значений дискретного распределения

Дискретное распределение. В случае дискретного распределения нужно указать число вычисляемых значений распределения (рис. 12.3). Будут вычислены все указанные значения функции вероятности и плотности вероятности.

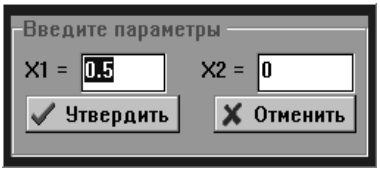


Рис. 12.4. Бланк установки двух вычисляемых значений непрерывного распределения

Непрерывное распределение. В случае непрерывного распределения нужно указать два значения аргумента *X1* и *X2* (рис. 12.4), для которых вычисляются значения функции распределения.

Будут вычислены значения вероятности, соответствующие введенным параметрам и вероятность попадания *X* в диапазон *X1–X2* (в случае ввода одного значения *X1* полагается *X2*=0). Затем запросы пар значений *X* будут продолжены до отмены бланка ввода.

3. Обратное распределение. В случае обратного распределения необходимо ввести значение вероятности *P*, после чего будет вычислено соответствующее значение параметра *X*.

В результатах указываются название распределения и выборочные значения среднего, стандартного отклонения и дисперсии.

Числовые результаты дополняются графиками функции вероятности распределения и плотности вероятности (рис. 12.5, 12.6).

Ограничения:

$n\log(1-P) > -39$ — биномиальное распределение;

$k\log(1-P) > -39$ — отрицательное биномиальное распределение;

$n<N-S$ или $n<S$ — гипергеометрическое распределение.

Пример 1

Задача. Вычислить 13 значений функции вероятности биномиального распределения с вероятностью успеха 0,7 и числом испытаний 10 и построить график распределения.

Результаты:

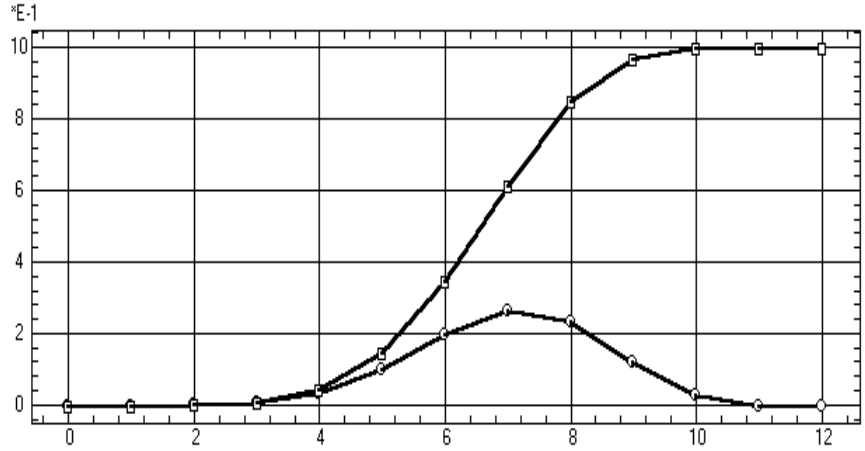


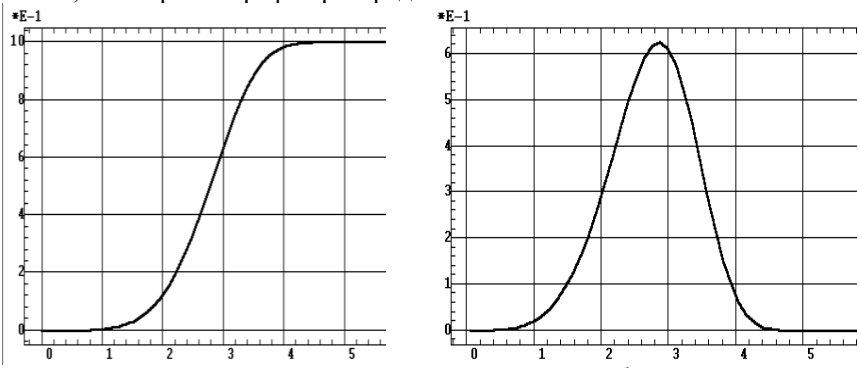
Рис. 12.5. График вероятности и плотности вероятности числа успехов X (X = 0, 1, 2,...) в 10 испытаниях при вероятности успеха в одном испытании, равной 0,7 (биномиальное распределение)

```
ВЫЧИСЛЕНИЕ ВЕРОЯТНОСТЕЙ. Распределение биномиальное: 0.7, 10, 2
          Среднее=7, Дисперсия=2.1, Ст.отклонение=4.41

Функция распределения вероятностей
         r      P(=r)     P(X<=r)
         0    5.9E-6     5.9E-6
         1  0.000138   0.000144
         2   0.00145    0.00159
         3     0.009     0.0106
         4    0.0368     0.0473
         5     0.103       0.15
         6       0.2       0.35
         7     0.267      0.617
         8     0.233      0.851
         9     0.121      0.972
        10    0.0282          1
```

Пример 2

Задача. Вычислить несколько значений функции вероятности распределения Вейбулла с параметрами: масштаба, равным 5, и формы, равным 3, и построить график распределения.



а *б*

Рис. 12.6. Графики распределения Вейбулла с параметрами: формы, равным 3, и масштаба, равным 5:
а — функции вероятности; *б* — функции плотности вероятности

Результаты:

```
ВЫЧИСЛЕНИЕ ВЕРОЯТНОСТЕЙ.  Распределение Вейбулла:2,3
Среднее=1.79, Дисперсия=2.81, Ст.отклонение=7.92
```

x1	F(x1)	x2	F(x2)	F(x1)-F(x2)
1	2	0.118	0.632	0.515
2	3	0.632	0.966	0.334
3	4	0.966	1	0.0339
4	5	1	1	0.000335
5	6	1	1	1.64E-7
6	7	1	1	1.82E-12
7	8	1	1	0

12.3. Согласие распределений

Назначение. Метод предназначен для проверки гипотезы об отсутствии различий между эмпирическим и теоретическим распределениями. В реализации использован ряд современных результатов по аппроксимации распределений статистик Колмогорова и омега–квадрат для различного типа распределений.

Действия и результаты. Сначала нужно выбрать для анализа переменную из электронной таблицы (см. рис. 7.2), представляющую экспериментальное распределение.

Далее следует выбрать тип непрерывного теоретического распределения (рис. 12.7).

Результаты включают:

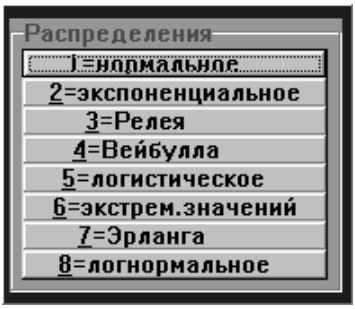


Рис. 12.7. Меню выбора теоретического распределения

ления (рис. 12.8).

- значение *статистики Колмогорова D* с уровнем значимости *P* нулевой гипотезы об отсутствии различий между эмпирическим и теоретическим распределениями;
- значение *статистики омега–квадрат W* с уровнем значимости *P* нулевой гипотезы.

Далее, по подтверждению, может быть построен график функции вероятности теоретического распределения с диаграммой рассеяния эмпирического распреде-

Пример 1

Задача. Выполнить проверку соответствия теоретического распределения и эмпирического распределения выборки объемом 30 чисел, полученной генерацией по распределению Вейбулла с параметрами: масштаба, равным 2, и параметром формы, равным 3, и выдать совместный график теоретического и эмпирического распределений.

Результаты:

```
СОГЛАСИЕ РАСПРЕДЕЛЕНИЙ.  Распределение Вейбулла: 1.98, 3.67
      Колмогоров=0.0666, Значимость=1, степ.своб = 30
      Гипотеза 0: <Распределение не отличается от теоретического>
      Омега-квадрат=0.016, Значимость=1, степ.своб = 30
      Гипотеза 0: <Распределение не отличается от теоретического>
```

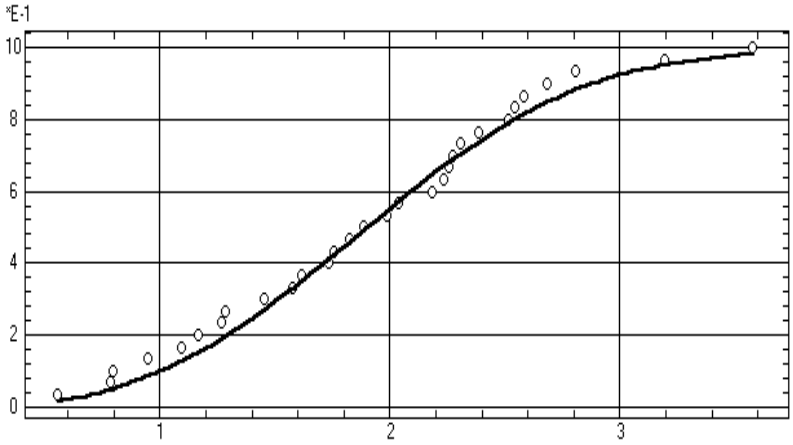


Рис. 12.8. Графики эмпирического и теоретического (Вейбулла) распределений

12.4. Согласие частот событий (долей)

Назначение. Данный метод проверяет одну из двух гипотез:

- о равенстве наблюдаемых частот появления двух событий (сравнение двух долей);
- о равенстве частоты теоретической вероятности события;
- а также позволяет вычислить минимальное число наблюдений, необходимое для определения частоты некоторого события с заданной точностью.

Действия и результаты. Сначала надо ввести исходные данные и выбрать варианта анализа (рис. 12.9).

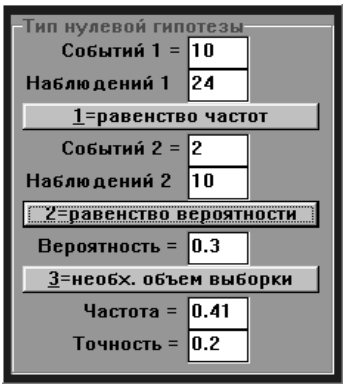


Рис. 12.9. Меню выбора метода согласия частот

В случае сравнения двух частот нужно ввести две пары чисел *m1*, *n1* и *m2*, *n2* (число наблюденных событий и число наблюдений). После этого выдается значение *Z*–статистики и уровень значимости *P* нулевой гипотезы отсутствия различий между двумя частотами. Вычисляется также величина разности частот и половина ее доверительного интервала *df*.

Использование доверительного интервала разности частот в ряде случаев предпочтительнее, в частности, оно позволяет количественно сопоставлять разные результаты (подробнее см. разд. 6.1).

В случае сравнения наблюдаемой частоты с теоретической вероятностью данного события следует ввести три параметра: число наблюденных событий 1, число наблюдений 1 и теоретическую вероятность события. После этого выдается значение *Z*–статистики и уровень значимости *P* нулевой гипотезы отсутствия различий между наблюденной частотой и теоретической вероятностью.

Для вычисления минимально необходимого числа наблюдений в меню следует ввести экспериментальную вероятность события и желаемую точность ее оценки.

Ограничение: *Z*–аппроксимация для критерия равенства частот применима при достаточно большом числе наблюдений: *m1*, *n1–m1*, *m2*, *n2–m2* > 5.

Пример 1

Задача. В одной выборке объема 1000 оказалось 20 бракованных изделий, в другой выборке объема 900 выявлено 30 бракованных изделий. Необходимо проверить, совпадают ли частоты появления бракованных изделий. Большое число наблюдений позволяет использовать *Z*–статистику.

Результаты:

```
СОГЛАСИЕ ЧАСТОТ.
Z =-1.956, значимость=5.04E-2
Гипотеза 0: <Частоты событий совпадают>
Разность частот=0.0133, доверит.интервал=0.0121
```

Выводы: Результаты анализа не позволяют принять ни одну из гипотез о соотношении частот появления брака (уровень значимости близок к критическому значению 0,05), поэтому требуется накопление дополнительных данных. Это подтверждает и противоположный вывод по сравнению разности частот с доверительным интервалом, который, несмотря на близость к разности частот, не включает нулевое значение.

Пример 2

Задача. Продолжим анализ соотношений встречаемости тромбоза при приеме аспирина и контроле из примера 1 разд. 7.6: при приеме аспирина наблюдалось 6 случаев тромбоза у 19 больных, а в контрольной группе — 18 случаев у 25 больных. Большое число наблюдений позволяет использовать *Z*–статистику.

Результаты:

```
СОГЛАСИЕ ЧАСТОТ.
Z=2.8,  Значимость=0.00508
   Гипотеза 1: <Частоты событий не совпадают>
Разность частот=0.377, доверит.интервал=0.248
```

Выводы: Результаты анализа позволяют с достаточной уверенностью (P=0,0029 значительно ниже критического значения 0,05, а доверительный интервал относительно разности частот не включает нулевое значение) принять гипотезу о различии этих двух частот, что подтверждает результаты примера 1 из разд. 7.6.

12.5. Последовательный анализ

Назначение. Метод последовательного анализа позволяет в ходе последовательных наблюдений изменения показателя x посредством проверки нулевой гипотезы $x=x_0$ относительно альтернативы $x<x_0+d$ решать,

продолжить ли далее процесс наблюдений, или же их объем уже достаточен для принятия той или иной гипотезы.

Этот метод особенно полезен в тех случаях, когда каждое новое наблюдение связано с существенными организационными или материальными затратами.

Исходные данные представляют собой две переменные:

1) последовательность моментов измерений t_i;

2) последовательность измерений исследуемого показателя x_i.

Действия и результаты. Сначала из электронной таблицы следует выбрать две переменные t и x для анализа (см. типовое меню рис. 2.3).

Далее нужно ввести три параметра нулевой и альтернативной гипотез (рис. 12.10): x_0, оценку стандартного отклонения исследуемого показателя S_x и d.

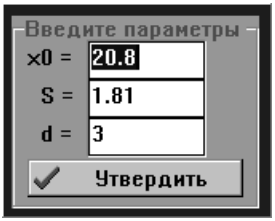


Рис. 12.10. Бланк установки параметров последовательного анализа

Строится график накопленной суммы $x_i–x_0$ и ограниченная двумя линиями область L_+, L_- принятия нулевой гипотезы по выбранному критическому уровню значимости α.

Новые испытания рекомендуется продолжать до тех пор, пока график не стабилизируется внутри области принятия нулевой гипотезы или же выйдет из этой области вверх (альтернатива $x>x_0+d$) или вниз (альтернатива $x<x_0–d$).

Пример

З а д а ч а. Было установлено, что затраты на предпосевную обработку семян окупятся, если прирост урожайности составит не менее d=3 ц/га при средней урожайности необработанных семян x_0=20,8 ц/га. Многолетние наблюдения показали, что в норме S_x=1,81 ц/га. В табл. 12.1 приведены урожайности обработанных семян за 10 последовательных лет (файл SEQ).

Таблица 12.1. Урожайность зерновых при предпосевной обработке семян

1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
26.7	21	24.1	27.1	25.1	23	26.2	19.4	21.8	23.4

Необходимо проверить, достаточно ли произведенных оценок урожайности (испытаний) для подтверждения гипотезы об окупаемости затрат.

В ы в о д ы: Как показывает график результатов последовательного анализа, оценки урожайности постепенно и уверенно выходят за допусковые границы, поэтому можно принять альтернативную гипотезу и закончить испытания уже на 5–6 году.

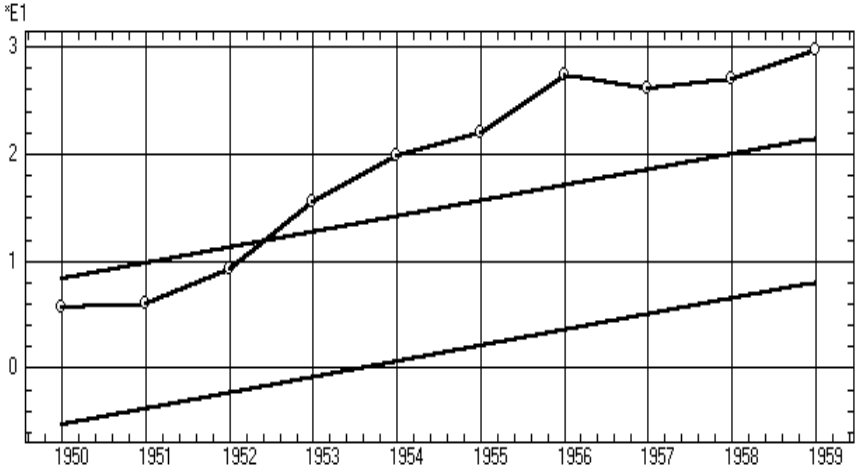


Рис. 12.11. График последовательного анализа эффективности предпосевной обработки семян с зоной принятия нулевой гипотезы

12.6. Анализ выживаемости

Назначение. *Кривая выживаемости* (рис. 12.12) задает вероятность *пережить* любой из моментов времени после некоторого стартового события, например после начала лечения. Здесь термин «пережить» означает недостижение некоторого интересующего исследователя *исхода*. Кривые такого рода позволяют описать продолжительность самых разнообразных процессов. В них в качестве *исхода* может выступать не только смерть, но и любые другие события, в том числе и не только нежелательные события. Например, можно изучать лечение какого-либо нелетального заболевания (исход — ремиссия), эффективность контрацепции (исход — беременность), долговечность протеза (исход — поломка) и т. п.

Важной характеристикой кривой выживаемости является ее *медиана* или интервал времени, который переживают 50% исследуемой популяции. Аналогично медиане могут быть вычислены и другие *процентили* выживаемости.

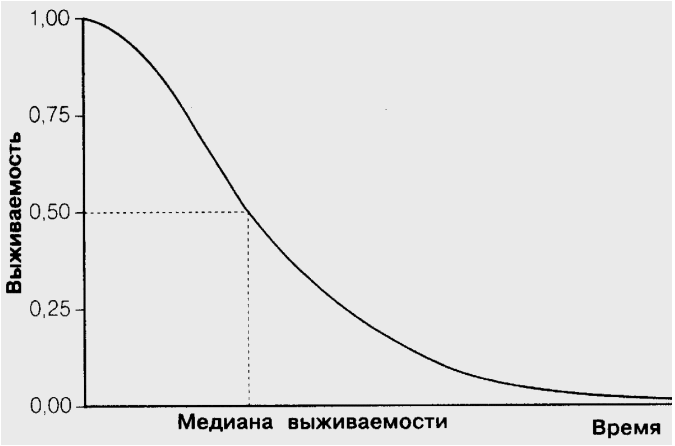


Рис. 12.12. Кривая выживаемости показывает процент выживших из наблюдаемой популяции в каждый момент времени от стартового события

График на рис. 12.12 представляет идеальную кривую выживаемости для случая бесконечной или очень большой популяции, когда временной интервал между двумя последовательными исходами близок к нулю, и когда наблюдение ведется очень длительное время, превосходящее медиану в несколько раз. В реальности же объем наблюдаемой популяции не очень велик, как и общее время наблюдения. Поэтому экспериментальная кривая выживания имеет не непрерывный характер как на рис. 12.12, а ступенчатый. На ее характер оказывают также влияние объекты, не достигшие интересующего исследователя исхода, а по каким-то причинам выбывшие из наблюдения, ставшие недоступными (сменившие место жительства, отказавшиеся от продолжения лечения и т. п.). В этом случае расчет кривой выживаемости производится моментным методом Каплана–Мейера [7].

Первой из задач анализа является построение экспериментальной кривой выживаемости, оценка ее медианы, стандартных ошибок и доверительных интервалов.

Часто также имеет место наблюдение двух популяций в двух различных *условиях* (например, при двух различных способах лечения). Тогда встает задача оценки различий между такими двумя *условиями* по различиям в соответствующих кривых выживания.

Исходные данные для построения одной кривой выживания представляются в виде двух *парных* переменных: первая переменная фиксирует моменты времени очередных исходов, а вторая переменная для каждого такого момента указывает число исходов. Если в очередной момент времени имеет место *выбывание* из-под наблюдения, то число исходов-выбываний указывается со знаком минус. Если имеют место как выбыва-

ния, так и нормальные исходы, то по ним приводятся две последовательные пары значений переменных. В случае сравнения двух кривых выживания в матрице данных приводятся значения соответствующих им двух парных переменных.

Результаты:

1) для каждой рассчитанной кривой выживаемости приводится таблица со столбцами (см. пример):
   t_i — временной рубеж очередного зафиксированного исхода/выбытия;
   p_i — вероятность пережить очередной временной рубеж, рассчитывается как отношение числа оставшихся под наблюдением n_{i+1} к числу бывших под наблюдением n_i перед временным рубежом t_i;
   v_i — вероятность выжить с начала наблюдения и после временного рубежа t_i по моментному методу рассчитывается как произведение p_j, $j = 1–i$;
   s_i — стандартная ошибка вычисления v_i;
   d_i — половина доверительного интервала для v_i получается умножением s_i на двустороннее критическое значение для стандартного нормального распределения при выбранном критическом уровне значимости α;
2) график экспериментальной кривой выживаемости $v_i=f(t_i)$ или совместный график двух кривых выживаемости;
3) в случае двух кривых выживаемости вычисляется Z-статистика и уровень значимости нулевой гипотезы об отсутствии различий в кривых выживаемости.

Рассмотренный метод сравнения двух кривых выживаемости носит название *логранговый критерий*. Аналогичную задачу решает и *критерий*

Гехана. Однако в большинстве случаев он менее предпочтителен (подробнее см. в [7]).

Пример

Задача. При остром лимфобластном лейкозе мутация предшественника лимфоцитов приводит к появлению клона лейкозных клеток, способных неограниченно делиться. Они подавляют нормальное кроветворение, вызывая иммунодефицит, анемию и тромбоцитопению, приводящие к смерти. При лечении облучением и химиотерапией для компенсации их побочных действий используют пересадку костного мозга. Для избежания отторжения костной ткани лучше использовать косный мозг близких родственником (аллотрансплантация). Однако не у всех таковые имеются или согласны на донорство, поэтому используется и пересадка собственного костного мозга (аутотрансплантация).

В табл. 12.2 приведены данные о числе умерших и выбывших из наблюдения пациентов (столбцы «выбыли») при этих двух способах пересадки костного мозга (файл LIFE2). В колонках «время» указаны месяцы исходов.

Таблица 12.2. Данные выживаемости пациентов с лейкозом при аутотрансплантации (1) и аллотрансплантации (2) костного мозга

Время1	Выбыли1	Время2	Выбыли2
1	3	1	1
2	2	2	1
3	1	3	1
4	1	4	1
5	1	6	1
6	1	7	1
7	1	12	1
10	1	20	-1
12	2	21	-1
14	1	24	1
17	1	30	-1
20	-1	60	-1
27	2	85	-2
28	1	86	-1
30	2	87	-1
36	1	90	-1
38	-1	100	-1
40	-1	119	-1
45	-1	132	-1
50	3		
63	-1		
132	-2		

Возникает задача сравнения этих двух видов трансплантации по степени выживаемости пациентов.

Обсуждение:

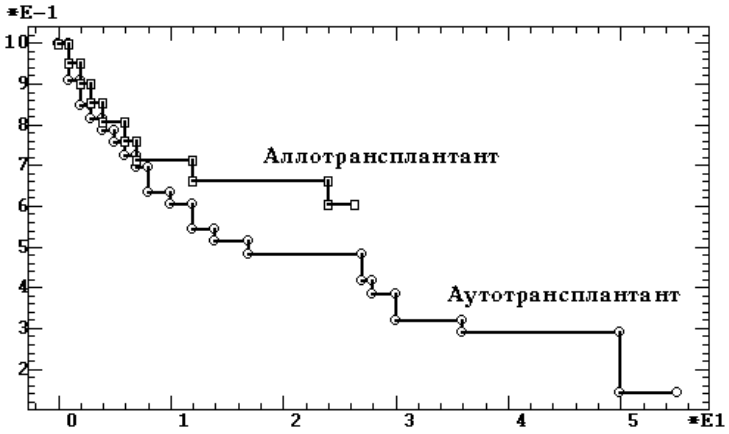


Рис. 12.13. Сравнение двух кривых выживаемости

СРАВНЕНИЕ ВЫЖИВАЕМОСТЕЙ. Файл: life2.std

Время	Пережили	Выжили	Ст.ошибка	Время	Пережили	Выжили	Ст.ошибка
1	0.909	0.909	0,05	1	0.952	0.952	0.0465
2	0.933	0.848	0.0624	2	0.95	0.905	0.0641
3	0.964	0.818	0.0671	3	0.947	0.857	0.0764
4	0.963	0.788	0.0712	4	0.944	0.81	0.0857
5	0.962	0.758	0.0746	6	0.941	0.762	0.0929
6	0.96	0.727	0.0775	7	0.938	0.714	0.0986
7	0.958	0.697	0.08	12	0.933	0.667	0.103
8	0.913	0.636	0.0837	24	0.909	0.606	0.11
10	0.952	0.606	0.0851				
12	0.9	0.545	0.0867				
14	0.944	0.515	0.087				
17	0.941	0.485	0.087				
27	0.867	0.42	0.0866				
28	0.923	0.388	0.0857				
30	0.833	0.323	0.0827				
36	0.9	0.291	0.0805				
50	0.5	0.145	0.0717				

Z=2.16, значимость=0.0153
Гипотеза 1: <Есть различия между кривыми выживаемости>

В ы в о д ы: Как показывают результаты статистического сравнения, две кривые выживаемости различаются на высоком уровне достоверности (P=0,0153). Поэтому можно обратиться к визуальному исследованию совместного графика кривых выживаемости (рис. 12.13). На графике видно, что аллотрансплантация костного мозга имеет несомненное преимущество, способствуя значительно большей выживаемости пациентов. Медиана аутотрансплантации находится в районе 17-го месяца, а к 50-му месяцу выживает только 14,5% пациентов. При аллотрансплантации медиана экспериментально не достигается (ее можно визуально прогнозировать в районе 50 месяцев), и как видно из табл. 12.2 — за счет пациентов, выбывших из-под наблюдения, по всей видимости вследствие стабилизации их состояния.