Э. А. Вуколов

ОСНОВЫ СТАТИСТИЧЕСКОГО АНАЛИЗА

Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL

2-е издание, исправленное и дополненное

Рекомендовано Советом Учебно-методического объединения вузов России по образованию в области менеджмента в качестве учебного пособия по специальности «Менеджмент организации»

ФОРУМ
МОСКВА
2008

УДК 311(075.32)
ББК 22.172я723
В88

Рецензенты:
зам. Генерального директора НИИ «Зенит» по науке,
доктор технических наук профессор *А. И. Кобзарь*;
зам. директора НИИФизпроблем,
кандидат технических наук *А. А. Васенков*

Вуколов Э. А.
В88 Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL: учебное пособие. — 2-е изд., испр. и доп. — М.: ФОРУМ, 2008. — 464 с. — (Высшее образование).

ISBN 978-5-91134-231-9

Книга является учебно-методическим пособием по теории вероятностей, статистическим методам и исследованию операций. Приведены необходимые теоретические сведения и подробно рассматривается решение задач прикладной статистики с использованием пакета STATISTICA. Излагаются основы симплекс-метода и рассматривается решение задач исследования операций средствами пакета EXCEL. Приводятся варианты заданий и методические разработки по основным разделам статистики и исследования операций.

Книга адресуется всем, кому необходимо применять статистические методы в своей деятельности, преподавателям и студентам, изучающим математическую, экономическую и прикладную статистику и методы исследования операций.

УДК 311(075.32)
ББК 22.172я723

ISBN 978-5-91134-231-9

Предисловие научного редактора

Название книги «Практикум...» не вполне отражает ее содержание и направленность. Материал, включенный в книгу, значительно шире набора практических задач и упражнений по обработке статистических данных. Книга представляет довольно редкое в учебной литературе сочетание рассказа об основных понятиях математической статистики, разбора решения задач на простых числовых примерах и показа решения тех же задач в статистическом пакете STATISTICA. Это одновременно: учебник, справочник, задачник и компьютерный практикум, охватывающий обширную тематику анализа данных и, отчасти, оптимизации. Следовало ли объединять все эти жанры в одной книге? Выиграл ли от этого читатель? На мой взгляд, несомненно, да! Типичный недостаток многих существующих учебников по математической и прикладной статистике — разрыв между теоретическим изложением и практическими расчетами реальных прикладных задач. В условиях отсутствия эффективных средств статистических расчетов на компьютерах, авторы часто вынуждены были отказываться от разбора содержательных задач в числах, так как даже простые статистические процедуры довольно трудоемки в расчетах. С другой стороны, в литературе, посвященной непосредственно статистическим пакетам и обработке данных на компьютере, изложение постановок статистических задач и методов их решения, крайне поверхностно. В этих книгах основное внимание обычно уделено разбору интерфейсов статистических пакетов. Редкими исключениями из этого правила являются книги [10, 37].

Книга, несомненно, рассчитана на широкий круг читателей: студентов, преподавателей и всех, кто анализирует данные на практике.

Для студентов книга является достаточно простым учебным пособием по курсам: математической и прикладной статистики, оптимизации, эконометрики и др. Не злоупотребляя формальным изложением и детальными доказательствами, в книге подробно разъясняются основные понятия математической статистики, оценивания, проверки гипотез, непараметрических (ранговых) методов и т. п. В приложении книги дано краткое изложение курса теории вероятностей. Книга окажется весьма полезной и при подготовке курсовых и дипломных работ, содержащих обработку практических данных.

Для преподавателя книга — готовый практикум по курсам теории вероятностей, статистики и оптимизации. По каждой теме она содержит обширные наборы типовых числовых задач для студентов и правила распределения индивидуальных задач в группах. При этом четко формулируются теоретические вопросы и понятия, которые должен знать студент по теме. Даны простые примеры на вычисления для закрепления понимания основных формул и статистических выводов. И, наконец, представлены более

сложные и содержательные задачи для компьютерного практикума. Все это делает книгу в методическом плане очень удобной для организации учебного процесса. Книга поможет преподавателю связать изложение теории обработки данных с ее реализацией в компьютерных программах. Ведь чаще всего документация пакета и его подсказка не разъясняют подробно, как именно вычисляется та или иная статистика.

Книга будет полезна всем, кто обрабатывает данные на практике, вне зависимости от области приложения. Разбор постановки задачи, помощь в выборе статистической процедуры для ее выполнения, пояснение на простом примере, как работает процедура, — именно то, что чаще всего не хватает практику, даже если он знаком с теорией.

Наконец, книга окажет неоценимую помощь всем, кто работает в пакете STATISTICA. В ней детально разбираются условия использования большинства статистических процедур пакета, порядок ввода данных и форма вывода результатов расчетов.

А. А. Макаров

Цель расчетов — не числа, а понимание.
Р. Хэмминг

Предисловие

Прикладные статистические методы широко используются в практической деятельности людей, работающих в самых различных сферах. Владение основами статистических методов необходимо специалистам, работающим в естественно — научных и инженерных областях, а также представителям гуманитарных профессий: экономистам, социологам, психологам, лингвистам.

Возможности компьютеров в обработке громадных объемов информации сделали доступными для широкого пользователя самые современные методы статистического анализа. В настоящее время разработано большое количество статистических пакетов программ, представляющих удобную современную форму программного обеспечения.

Применение статистических пакетов упрощает использование статистических методов, однако необходимо не только собрать и правильно ввести данные, выбрать тот или иной способ их обработки, но и понимать основные идеи статистических методов и, что особенно важно, предположения, при которых теоретически обоснованы эти методы. При осмысленном применении статистических методов невозможны курьезы и абсурдные выводы, которыми так богата история статистики.

Предлагаемый читателю практикум по статистическим методам написан на основе лекций, практических занятий и лабораторных работ в течение многих лет проводившихся автором в институте экономики, управления и права Московского института электронной техники (МИЭТ), а также курсов, читаемых для слушателей, получающих второе высшее образование.

Основная цель книги состоит в том, чтобы на простых примерах показать как можно использовать статистический пакет программ для выполнения статистических расчетов, провести анализ результатов и сделать выводы.

Особенностью данного практикума является то, что он содержит большой теоретический и справочный материал. Таким образом, его можно рассматривать и как учебник по прикладной статистике.

Практикум, прежде всего, может представлять интерес для широкого круга специалистов, которым необходимо применять статистические методы в своей профессиональной деятельности.

Практикум можно использовать и при изучении всех разделов традиционных учебных курсов теории вероятностей, математической и общей статистики для инженерных и экономических специальностей. Он может

быть полезен как студентам, изучающим эти дисциплины, так и преподавателям при организации и проведении практических занятий и лабораторных работ по статистическим методам и курсу исследования операций.

Несколько более подробно о структуре и содержании практикума.

Все практические расчеты и примеры рассматриваются в процедурах пакета STATISTICA. В главе 1 приводится краткое, но достаточное на наш взгляд, описание структуры пакета. Более подробно с описанием пакета можно ознакомиться по специальной литературе [15, 16, 17].

Цель второй главы — показать, как можно использовать пакет STATISTICA для вычисления вероятностей и для генерации случайных чисел с заданным распределением.

Теоретической основой статистических методов является теория вероятностей.

Основные понятия теории вероятностей: случайные события, случайные величины, распределения и числовые характеристики случайных величин постоянно используются в книге. Необходимые сведения по теории вероятностей приводятся в Приложении. К этому материалу можно обращаться по мере необходимости как справочному.

В третьей главе изложены основы статистических методов. Здесь приводятся необходимые определения, теоретические сведения, а также многочисленные примеры, поясняющие изложение.

Каждая глава содержит практическую часть, в которой подробно разбирается решение примеров в пакете STATISTICA, и приводятся дополнительные задания для самостоятельной работы.

Суть статистических методов находится «на кончике пера»: так же как нельзя научиться считать без практики, не зная правил арифметики и алгебры, а используя только калькулятор, точно также нельзя понять статистические методы используя только компьютер и не занимаясь вычислениями «вручную». Поэтому в практических работах, по возможности, использовался следующий подход. Сначала рассматривается простая задача на данную тему и решается «вручную», без компьютера. Затем эта же задача решается с применением пакета STATISTICA на компьютере. Результаты анализируются и сравниваются, и, после этого, предлагается ряд дополнительных заданий для самостоятельной работы.

Пакет STATISTICA содержит большое число статистических процедур. Однако разобраться с особенностями процедуры, предположениями, при которых она может быть использована, и, наконец, с вопросами выбора той или иной процедуры при решении конкретной задачи можно только обратившись к специальной литературе.

Для студента или специалиста, желающего использовать статистические методы в своей области знаний, такую литературу либо трудно найти, либо весьма сложно читать, так как она написана на достаточно высоком математическом уровне. Описание статистических методов в руководствах, имеющихся в документации пакетов обычно не полно и мало понятно.

Для таких читателей в практикуме рассмотрены отдельные разделы статистического анализа, выходящие за рамки учебных программ, но имеющих большое значение в практике научных и инженерных исследований. Это, прежде всего, непараметрические методы математической статистики

(глава 4) и однофакторный дисперсионный анализ (глава 5). Подробно рассмотрены методы регрессионного анализа: множественная регрессия, процедуры пошагового выбора наиболее значимых факторов, вопросы проверки значимости и адекватности моделей, корреляционный анализ (глава 6). Простейшие методы анализа временных рядов: скользящие средние, вычисление сезонных индексов и прогнозирование на основе экспоненциального сглаживания изложены в главе 7. В главе 8 рассмотрены основы кластерного анализа.

В последней главе рассматривается решение задач, относящихся к тематике курсов «Методы оптимизации», «Дискретная математика», «Исследование операций». Это сделано в связи с тем, что исследование операций, как и экономическая статистика входят в общеобразовательный стандарт дисциплин для таких специальностей как менеджмент, маркетинг, муниципальное управление и др. Мы надеемся, что рассмотрение методов исследования операций в одной книге со статистическими методами представит определенные удобства для читателей, обучающихся этим специальностям.

Цель этой главы состоит в том, чтобы на простых примерах показать, как задачи подобного вида могут быть решены на компьютере средствами пакета Excel.

Чтобы упростить работу с пакетом в приложении приводится англо-русский словарь терминов пакета STATISTICA и статистических терминов.

Автор выражает глубокую признательность и благодарность научному редактору, кандидату физико-математических наук, ведущему научному сотруднику НИИ Механики МГУ, доценту кафедры математики Высшей школы экономики А. А. Макарову.

Серьезная профессиональная работа А. А. Макарова и общение с ним во многом способствовали улучшению содержания и структуры книги.

Большую помощь в техническом оформлении книги на компьютере оказали студенты института экономики и управления МИЭТ. Всем им и особенно К. Панкратову, К. Миляеву, Е. Ушаковой, В. Сухушиной, А. Чамовой, Е. Вербицкой, А. Курьянову, Н. Бартко, И. Семиной, О. Панасенко и многим другим автор выражает глубокую благодарность.

Во втором издании исправлены описки и технические неточности. В первую главу добавлен п. 1.4: **Некоторые особенности версии 6.1**.

Глава 1
СТРУКТУРА ПАКЕТА STATISTICA

Универсальный статистический пакет STATISTICA разработан и производится фирмой Statsoft, Inc. (США). Последние версии пакета полностью совместимы со средой Windows. В России пакет распространяется фирмой Statsoft Russia. Подробную информацию о пакете можно получить на сайте **http://www.statsoft.ru**. Работа с пакетом подробно описана в ряде книг на русском языке [15, 16, 17], кроме того пакет содержит понятную и хорошо структурированную документацию в системе Help. В связи с этим мы ограничимся очень кратким описанием структуры пакета.

1.1. Модули пакета STATISTICA

Пакет STATISTICA имеет модульную структуру.

Модули открываются простым щелчком мыши из **Переключателя модулей** (STATISTICA Module Switcher) (рис. 1.1), который выводится на экран при запуске пакета. Инструментальная кнопка **Переключателя модулей** расположена в третьей строке рабочего окна STATISTICA (см. рис. 1.2).

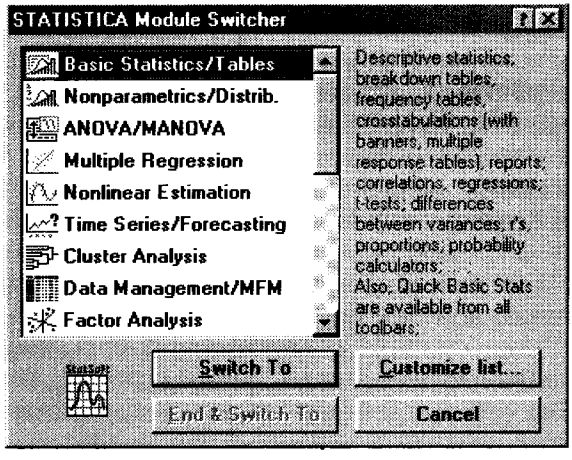


Рис. 1.1. Переключатель модулей STATISTICA

Каждый модуль может работать независимо от других модулей системы. В данной книге используются следующие модули:

- **Data Management** (Управление данными);
- **Basic Statistics/Tables** (Основные статистики/Таблицы);

- **Nonparametrics/Distributions** (Непараметрические методы и подгонка распределений);
- **Multiple Regression** (Множественная регрессия);
- **ANOVA/MANOVA** (Одно-/Многофакторный дисперсионный анализ);
- **Time Series/Forecasting** (Временные ряды/Прогнозирование);
- **Cluster Analysis** (Кластерный анализ).

Переключение модулей

Чтобы переключиться из одного модуля в другой модуль нужно нажать кнопку **Переключателя модулей** , затем высветить название модуля в **Переключателе модулей** и нажать кнопку **Switch to (Переключиться в)**, либо просто дважды щелкнуть на название модуля. После этого на экране появится рабочее окно системы со стартовой панелью модуля.

Рабочее окно STATISTICA

На рис. 1.2 показано окно STATISTICA после запуска модуля Basic Statistics/Tables (основные статистики и таблицы), стартовая панель модуля удалена.

Название модуля высвечивается в верхней строке окна.

Вторая сверху строка окна (строка меню) содержит ряд пунктов меню общих для Windows приложений: File — Файл, Edit — Правка, View — Вид, Window — Окно, Help — Помощь. Меню Analysis — Анализ выводит на эк-

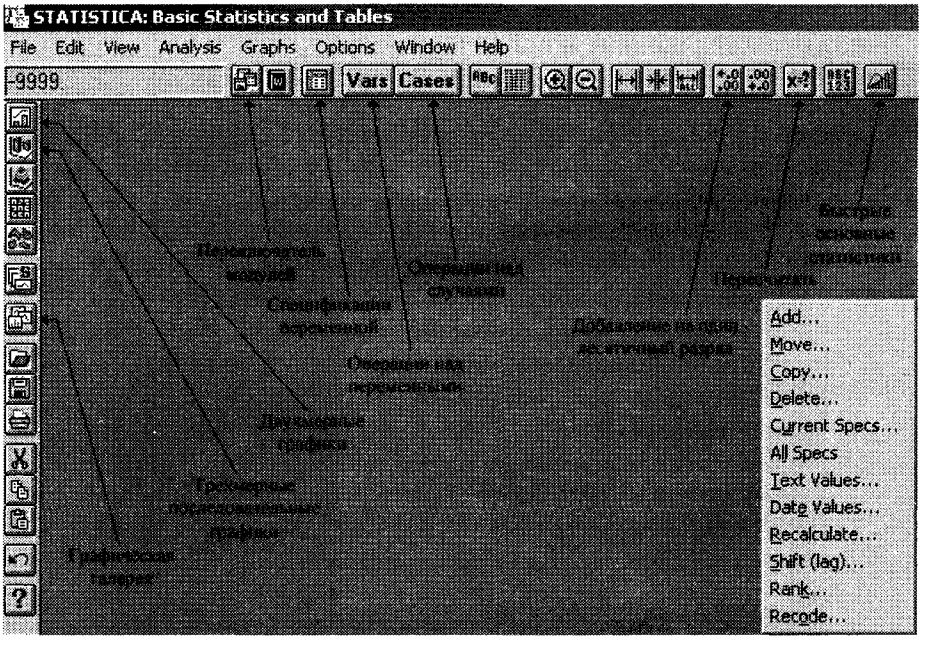


Рис 1.2. Окно STATISTICA (в правом нижнем углу команды редактирования переменных — кнопка Vars)

ран стартовую панель (Startup Panel) любого модуля. Меню Graphs — Графики предназначено для построения различных графиков (такое же назначение имеют верхние инструментальные кнопки расположенные в вертикальном ряду с левой стороны экрана). В меню Options — Опции задаются значения параметров конфигурации пакета.

Третья сверху строка окна и вертикальный ряд с левой стороны окна содержат инструментальные кнопки для быстрого доступа к командам меню. Назначение наиболее часто используемых кнопок указано на рис. 1.2. Часть кнопок снабжена общепринятыми пиктограммами.

Работа в модуле

Работа в каждом модуле имеет общие черты. Нужно выполнить следующие действия:

- ввести данные или открыть файл данных;
- выбрать переменные для анализа;
- выбрать метод анализа из меню в стартовой панели модуля;
- выбрать конкретную вычислительную процедуру и задать ее параметры;
- произвести запуск вычислительной процедуры;
- выбрать следующий шаг анализа.

Стартовая панель модуля (Startup Panel)

Стартовая панель любого модуля всегда вызывается с помощью меню Analysis (см. рис. 1.2). Стартовая панель модуля дает возможности:

1) открыть файл данных (кнопка Open data);
2) выбрать переменные для анализа (кнопка Variables);
3) выбрать метод анализа данных.

Стартовая панель модуля Basic Statistics/Tables приведена на рис. 1.2а.

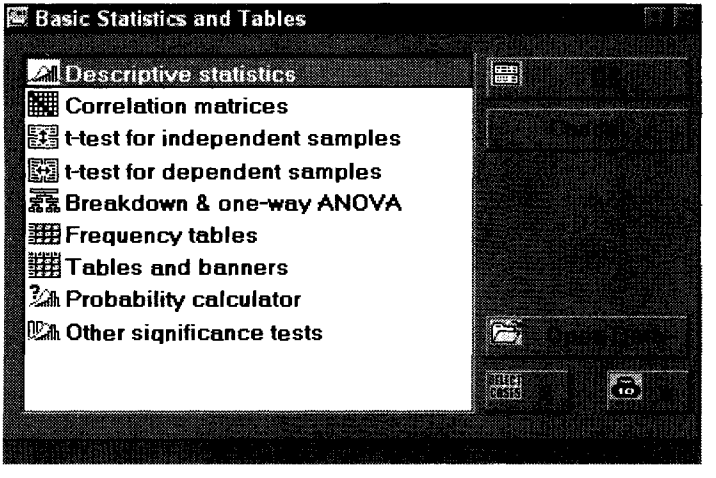


Рис. 1.2а. Стартовая панель модуля Basic Statistics/Tables

1.2. Структура, ввод и редактирование данных

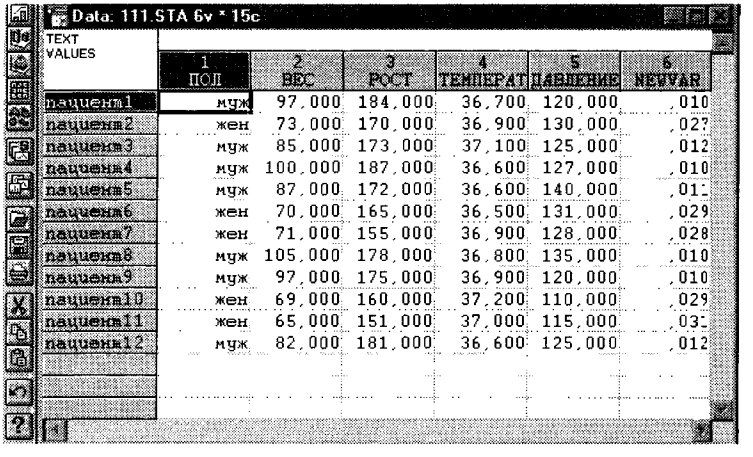
Файлы данных в пакете STATISTICA организованы аналогично файлам в электронных таблицах (например Excel). Такой файл можно рассматривать как таблицу (Spreadsheet), в которой столбцы являются переменными (Variables), а в строках записываются значения переменных — наблюдения (Cases).

На рис. 1.3 приведен файл содержащий данные медицинского обследования двенадцати пациентов. Для каждого пациента определялись: пол, вес, рост, температура и давление. Таким образом, файл содержит значение пяти переменных, каждая из которых представлена двенадцатью наблюдениями.

Имя файла (111.sta) и размер таблицы, содержащей файл, шесть переменных и пятнадцать наблюдений, 6v * 15c, указаны в заголовке таблицы. Три последние строки таблицы не содержат данных. В вычислительных процедурах пакета строки таблицы не содержащие данных рассматриваются как пропущенные значения (missing). Пропущенные значения учитываются в некоторых процедурах пакета, в частности при выполнении частотной табуляции (Frequency tables) в модуле Basic Stat./Tables. (см. главу 3, п. 3.4).

В файле на рис. 1.3 переменная **ПОЛ** представлена в нечисловой (текстовой) форме (Text Values).

В пакете STATISTICA при вводе текстовой переменной каждому значению ставится в соответствие числовая метка (число). Таким образом, значения текстовых переменных имеют двойную запись. Например, при вводе переменной **ПОЛ** значению муж соответствует 1, а значению жен — 2. Чтобы перевести текстовые переменные в числовую форму нужно нажать кнопку  на панели инструментов. Задать или изменить числовые метки для выделенных переменных можно, нажав инструментальную кнопку Vars (Modify Variables) и выбрав в выпадающем меню (см. рис. 1.2, меню в правом нижнем углу) кнопку Text Values... (либо нажать инструментальную кнопку ).


Data: 111.STA 6v * 15c

TEXT VALUES

	1 ПОЛ	2 ВЕС	3 РОСТ	4 ТЕМПЕРАТ	5 ДАВЛЕНИЕ	6 NEWVAR
пациент1	муж	97,000	184,000	36,700	120,000	,010
пациент2	жен	73,000	170,000	36,900	130,000	,027
пациент3	муж	85,000	173,000	37,100	125,000	,012
пациент4	муж	100,000	187,000	36,600	127,000	,010
пациент5	муж	87,000	172,000	36,600	140,000	,011
пациент6	жен	70,000	165,000	36,500	131,000	,029
пациент7	жен	71,000	155,000	36,900	128,000	,028
пациент8	муж	105,000	178,000	36,800	135,000	,010
пациент9	муж	97,000	175,000	36,900	120,000	,010
пациент10	жен	69,000	160,000	37,200	110,000	,029
пациент11	жен	65,000	151,000	37,000	115,000	,031
пациент12	муж	82,000	181,000	36,600	125,000	,012

Рис. 1.3. Данные в файле 111.sta

Рассмотрите несколько файлов из директории EXAMPLES. Откройте файл cardata. В этом файле переменными (Vars) являются характеристики 155 автомобилей: мощность двигателя, вес, год выпуска и т. д., а наблюдениями (Cases) — рассматриваемые автомобили.

В файле adstudy переменными являются характеристики конкретных людей: пол (Gender), предпочтения (Advert) и другие, а наблюдениями — фамилии людей.

Файлы данных в STATISTICA имеют расширение **sta**. Такие файлы могут иметь практически неограниченное число строк (если нужно — миллионы), количество столбцов ограничено числом **4092**. Таблицы, имеющие более чем 4092 столбца (вплоть до **32 тысяч** столбцов) могут быть оформлены как ***мегафайлы*** (***megafiles***). Мегафайлы имеют расширение **mfm**.

1.2.1. Ввод данных

Создание файла данных. В меню **File** нужно выбрать **New Data** (Новые данные). Далее в диалоговом окне нужно задать название файла.

После того как таблица для нового файла появится на экране (по умолчанию размер таблицы 10v*10c) нужно, используя инструментальные кнопки Vars и Cases установить необходимое число переменных и число строк.

Далее, щелкнув два раза на имени выбранной переменной, в окне Variable Specs... (рис. 1.4) задать имя переменной, формат представления чисел и т. д. Для этих же целей более удобно использовать опцию All Specs (спецификации всех переменных) в выпадающем меню, появляющемся при нажатии инструментальной кнопки Vars (см. рис. 1.2, меню в правом нижнем углу).

Заполнение таблицы данных в программе STATISTICA осуществляется с помощью клавиатуры и мыши (включая выделение блоков, копирование,

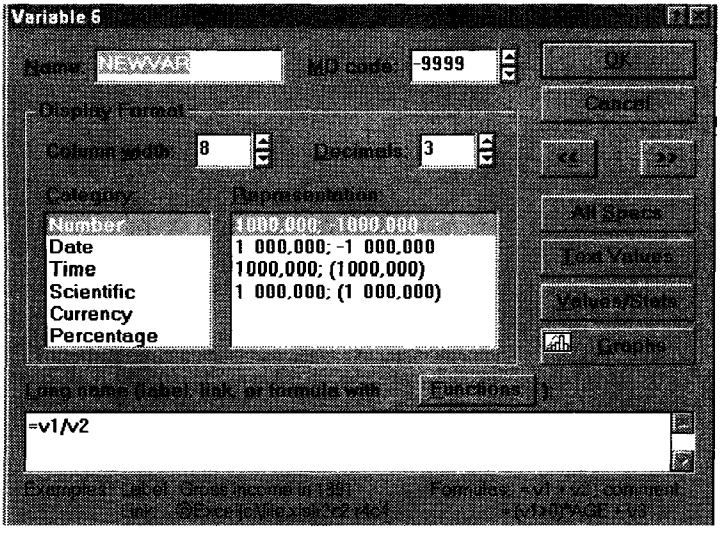


Рис. 1.4. Окно спецификации переменной

вставку и др.) в полной аналогии с электронными таблицами, такими как Excel. Десятичные знаки отделяются запятой, например: 25,33.

Открытие файла данных. В меню **File** (Файл) необходимо выбрать **Open Data** (Открыть данные) и открыть интересующий файл.

Сохранение файла. В меню **File** необходимо выбрать Save as... (сохранить как) и указать место, где сохранить файл.

Импорт файла данных. В меню **File** необходимо выбрать **Import Data** (Импорт данных). Можно импортировать данные самых разных типов: **Excel (*.xls)**, **dBase (*.dbf)**, **ASCII** (например, ***.txt**), **Megafile Manager (*.mfm)** и др.

Экспорт файла данных. В меню **File** необходимо выбрать **Export Data** (Экспорт данных).

1.2.2. Редактирование данных

При нажатии кнопки **Vars** на панели инструментов становятся доступными команды редактирования переменных-столбцов (см. рис. 1.2 меню в правом нижнем углу): **Add** (добавить новые переменные), **Delete** (удалить переменные), **Move** (переместить) и др. При нажатии кнопки **Cases** становятся доступными аналогичные команды редактирования строк.

Если дважды щелкнуть на имени переменной (например, **Var1**), то появится окно спецификации переменной **(Variable specs)** (рис. 1.4), посредством которого можно редактировать выбранную переменную: задать имя переменной (Name), формат для ее представления в числовом виде (Column width — число знаков и Decimals — число десятичных знаков), числовые метки для текстовых переменных (Text Value), построить графики (Graphs).

Обратим внимание на поле **Long name (label, link, or formula)**. В этом поле можно задать формулу, по которой будет рассчитываться выбранная переменная; например, можно написать **=v1/v2**; и тогда выбранная переменная может быть пересчитана по указанной формуле. В формулах по умолчанию переменные можно обозначать буквой **v** с указанием номера переменной (например, **v2** означает второй столбец); альтернативный способ заключается в написании действительных названий переменных, например: **=вес/рост**. Чтобы пересчет действительно состоялся, нажмите **OK** и согласитесь с предложением **«Recalculate the variable now»** (другой способ: нажать кнопку **Vars** и выбрать команду **Recalculate**). Если после формулы поставить точку с запятой, то далее в том же поле можно написать любой комментарий.

Источники данных. Для того чтобы получить файлы данных, которые могли бы быть использованы на этапе обучения работе в пакете STATISTICA, можно пойти различными путями:

- использовать файлы—примеры, которые поставляются вместе с системой и находятся в каталоге **stat\examples**;
- создать файлы данных и наполнить их данными, анализ которых представляет интерес;
- импортировать в программу STATISTICA данные, хранящиеся в форматах других программ, например данные из пакета Excel;

- используя встроенные в пакет генераторы случайных чисел и язык STATISTICA BASIC, создать таблицы данных, имитирующие реальные статистические выборки. Удобства такого способа заключаются в быстроте заполнения таблиц и возможности управления статистическими характеристиками генерируемых данных (см. главу 2, п. 2.3).

1.3. Вычисление основных статистик и построение графиков

Вычислить основные статистики для переменных можно используя инструментальную кнопку **Quick Basic Stats**.

На рис. 1.5 показаны результаты вычисления основных статистик для переменных Вес, Рост, Температура, Давление из файла представленного на рис. 1.3.

	Valid N	Mean	Confid. -95,000%	Confid. 95,000	Sum
ВЕС	12	83,4167	74,6025	92,2308	1001,000
РОСТ	12	170,9167	163,7259	178,1074	2051,000
ТЕМПЕРАТ	12	36,8167	36,6763	36,9570	441,800
ДАВЛЕНИЕ	12	125,5000	120,1807	130,8193	1506,000

Рис. 1.5. Основные статистики для переменных Вес—Давление

Valid N — число наблюдений, **mean** — оценки математических ожиданий, **Confid** — 95 % — границы 95 % доверительных интервалов для математических ожиданий и др. (более подробно см. главу 3, п. 3.4).

В пакете STATISTICA существует функция **Graphs Gallery (Графическая Галерея)**, на рабочем столе (рис. 1.2) она имеет вид и расположена в середине левого вертикального ряда инструментальных кнопок. Щелкнув по этой кнопке, попадаем в окно (рис. 1.6). В левой части окна выбирается пространство, в котором будет осуществляться построение графика, в правой выбирается тип графика.

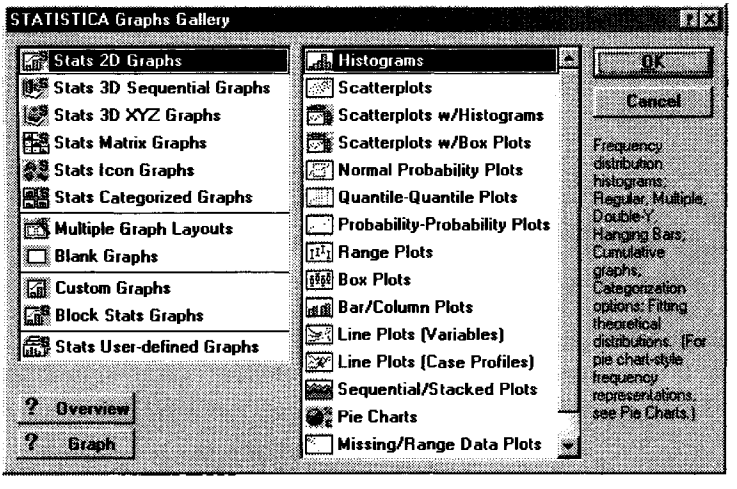


Рис. 1.6. Graphs Gallery (Графическая Галерея) на рабочем столе

В этом диалоговом окне представлены различные виды графиков. Рассмотрим наиболее простые графики:

- столбцовую диаграмму (Bar/Column Plots);
- диаграмму рассеяния (Scatterplots);
- построение пользовательских графиков (Custom Function Plots).

Построение столбцовой диаграммы для переменной Вес. Graphs Gallery → Stats 2D Graphs → Bar/Column Plots — в появившемся окне (рис. 1.7) устанавливаем: имя переменной **Вес, Graph Type: Regular, Control Limits: off, OK.**

Полученный график приведен на рис. 1.8.

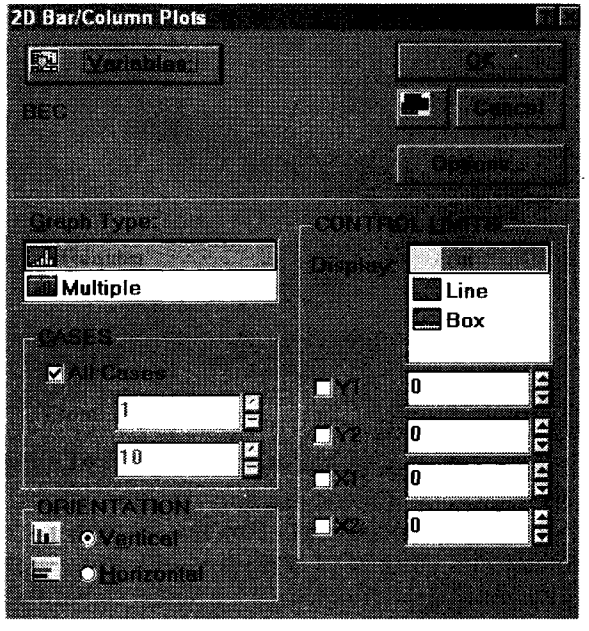


Рис. 1.7. Окно для построения столбцовой диаграммы

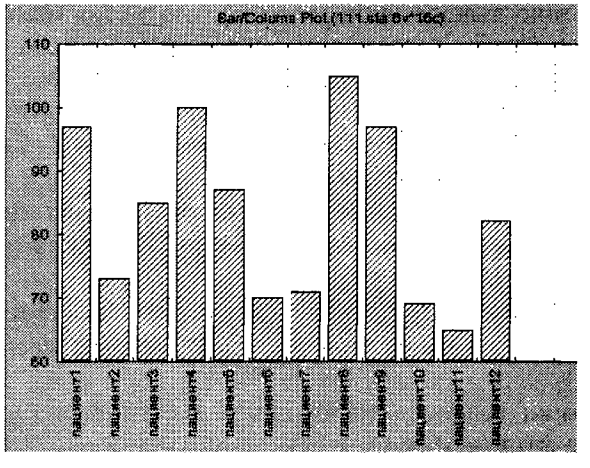


Рис. 1.8. Столбцовая диаграмма для переменной Вес

Диаграмма рассеяния для переменных Вес и Рост. Graphs → Stats 2D Graphs... → Scatterplots... — вводим значения по осям X и Y (нажав на кнопку **Variables** и выбрав переменные Вес и Рост) — OK. Диаграмма рассеяния показана на рис. 1.9.

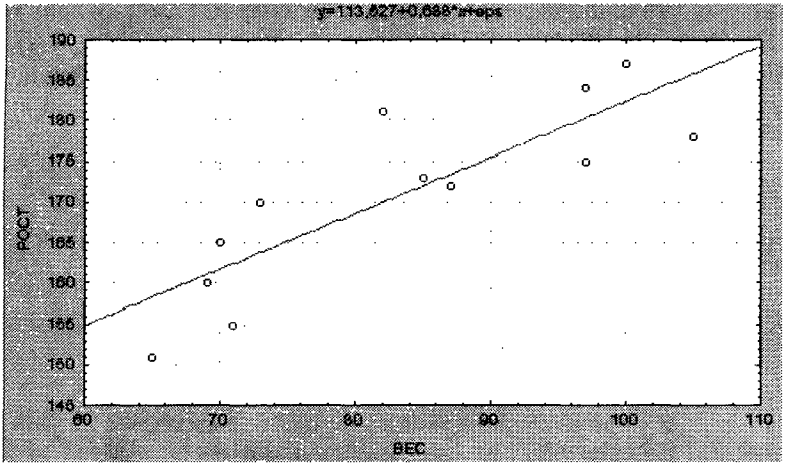


Рис. 1.9. Диаграмма рассеяния для переменных Вес и Рост

Прямая на диаграмме рассеяния — график простой линейной регрессии $y = 113{,}527 + 0{,}688 \cdot x$, где y — Рост, а x — Вес (см. главу 6, п. 6.1).

В пакете STATISTICA можно достаточно просто построить различные статистические графики: полигоны частот, график накопленных относительных частот (огиву), гистограммы. Причем графики строятся не только для переменных, записанных в таблицу исходных данных, но и для результатов вычислений, представленных в виде таблицы. Такие таблицы появляются при одновременной обработке нескольких переменных: вычислений описательных статистик и в ряде других операций, а также при частотной табуляции. Каждый столбец таблицы результатов: mean, min, max, std. dev... при выводе описательных статистик или count, percent, cum. percent в таблице частот есть новая переменная, которая может быть представлена на графике. Чтобы построить график, нужно сделать щелчок правой кнопкой мыши на заголовке соответствующего столбца (имени переменной) и в появившемся меню выбрать **Custom Graphs → 2D Graphs → line plot**.

Заслуживает внимания **построение пользовательских графиков (Custom Function Plots)**, так как, зная исходную функцию, можно построить любой график (как двумерный, так и трехмерный).

Построим, например, график в трехмерном пространстве. **Graphs → Stat 3D XYZ Graphs → Custom Function Plots**. Открывшееся окно показано на рис. 1.10.

Для построения графика, необходимо ввести функцию в поле **Enter Function (введите функцию)** (рис. 1.11).

Получившийся график представлен на рис. 1.12.

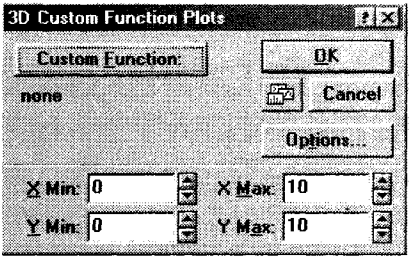


Рис. 1.10. Окно для построения графика

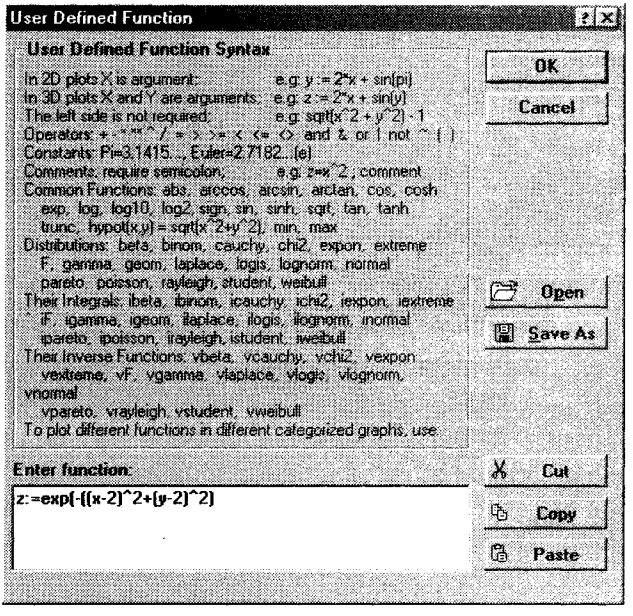


Рис. 1.11. Окно для ввода функции

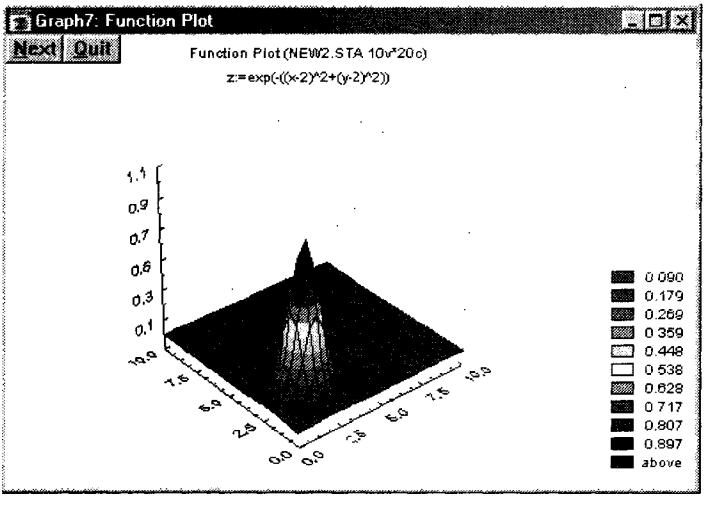


Рис. 1.12. Трехмерный график

1.4. Некоторые особенности версии 6.1

В п. 1.1 кратко описано окно пакета в версиях 5.5 и 6.0 на английском языке. В настоящее время фирма Statsoft Russia распространяет версию 6.1 на русском языке.

Основные отличия окна **STATISTICA** в версии 6.1 состоят в отсутствии **Переключателя модулей**. Выбор необходимого модуля для статистической обработки данных осуществляется в меню **Анализ** (в строке основных меню окна **STATISTICA**) (рис. 1.13).

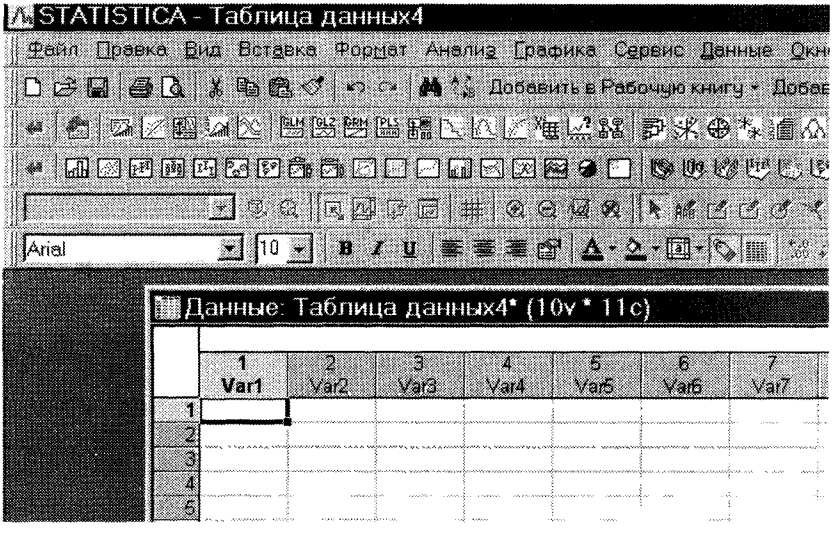


Рис. 1.13

Модуль **Временные ряды и Прогнозирование**, а также модуль **Нелинейное оценивание**, используемые в данной книге, содержатся в подменю **Углубленные методы анализа**. Модуль **Кластерный анализ** находится в подменю **Многомерный разведочный анализ**.

Для ввода новых данных нужно создать таблицу данных, используя меню **Файл—Создать**.

В новой версии окно **STATISTICA** не содержит кнопок **Vars** (переменные) и **Cases** (наблюдения). Вместо этого в строке основных меню имеются кнопки **Данные** и **Вставка** которые открывают меню для всех операций редактирование данных и таблицы данных.Кроме этого имеется множество других возможностей для редактирования.

Результаты статистического анализа удобно сохранять в специальных файлах — **Рабочих книгах**.

Глава 2
ВЫЧИСЛЕНИЕ ВЕРОЯТНОСТЕЙ И МОДЕЛИРОВАНИЕ РАСПРЕДЕЛЕНИЙ СЛУЧАЙНЫХ ВЕЛИЧИН В ПАКЕТЕ STATISTICA

Теоретической основой статистических методов является теория вероятностей. Основные понятия теории вероятностей: случайные события и их вероятности, случайные величины, распределения и числовые характеристики случайных величин постоянно используются в книге. Необходимые сведения по теории вероятностей приводятся в Приложении. К этому материалу можно обращаться по мере необходимости как справочному. Читателю основательно подзабывшему основы теории вероятностей, советуем внимательно прочитать эту часть, особенно разделы П.2 и П.3, в которых описываются случайные величины и их распределения.

Цель этой главы — показать, как можно использовать пакет STATISTICA для вычисления вероятностей и квантилей распределения случайных величин, построения различных вероятностных графиков и для генерации случайных чисел с заданным распределением (моделирования распределения).

Здесь и всюду в дальнейшем будут использоваться следующие обозначения.

X, Y, Z, ... — обозначения случайных величин — прописные латинские буквы.

x, y, z, a, b, ... — строчные буквы — значения, которые могут принимать случайные величины.

$P[X = a]$ — вероятность события, состоящего в том, что дискретная случайная величина X принимает значение a.

$P[a < X < b]$ — вероятность события, состоящего в том, что случайная величина X принимает значение из интервала (a, b).

$F(x)$ — функция распределения, $f(x)$ — плотность распределения непрерывной случайной величины X (см. Приложение П.2.3, П.3.1).

Далее мы рассмотрим вероятностные функции пакета STATISTICA для основных дискретных и непрерывных распределений.

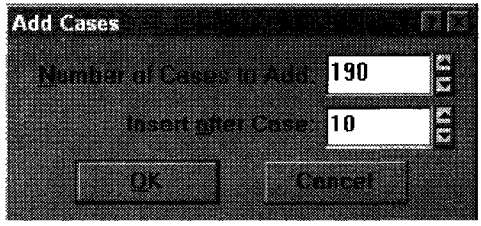
Предварительно переключитесь в модуль Basic Statistics and Tables (основные статистики и таблицы).

Откройте новый файл для выполнения расчетов и записи результатов моделирования: **File → New Data → OK**, введите имя файла, например: **mod → OK**.

В открывшейся таблице для ввода исходных данных имеется 10 переменных (var1 ÷ var10) и 10 строк. Увеличим число строк до 200. Для этого

нажмите кнопку Cases — наблюдения (см. рис. 1.2) и выполните следующие действия: **Cases → Add** (добавить число строк до 200).

В появившееся окно **Add Cases**, введите число добавляемых строк (Number of cases to Add) — 190, **OK**.



В заголовке файла появится надпись: Data: mod. sta 10v * 200c.

2.1. Вычисление вероятностей для дискретных случайных величин

Случайная величина X, имеющая биномиальное распределение $B(n, p)$, принимает значения: 0, 1, 2, ..., n. Распределение случайной величины X определяется следующей формулой

$$p_k = P[X = k] = C_n^k p^k q^{n-k}, \quad k = 0, 1, ..., n,$$

где $q = 1 - p$, C_n^k — число сочетаний из n по k, $C_n^k = \dfrac{n!}{(n-k)!k!}$ (см. П.2.6).

В пакете STATISTICA для вычисления вероятностей биномиального распределения используются две функции:

1. **Binom (*x*; *p*; *n*)** — вероятность, того что случайная величина X, имеющая биномиальное распределение с параметрами n и p, примет значение x, т. е. $P[X = x]$.

2. **IBinom (*x*; *p*; *n*)** — суммарная накопленная вероятность, $P[X \le x]$ (аналог функции распределения в точке x), $\sum_{k=0}^{k=x} P[X = k]$.

Чтобы выполнить вычисления с помощью этих функций и ввести результаты вычислений в какую-либо переменную (например, Var1) нужно открыть окно спецификации Var1 (**Variable Specs...**), и в поле **long name** записать функцию **=Binom(*x*; *p*; *n*)** или **=IBinom(*x*; *p*; *n*)** и задать вместо x, p и n необходимые значения.

Пример 2.1. Пусть случайная величина X имеет биномиальное распределение с параметрами $n = 5$ и $p = 0{,}3$. Вычислим вероятности событий: $P[X = 2]$ и $P[X \le 4]$.

Решение. Установите курсор на Var1 и нажмите правую кнопку мыши. В открывшемся меню выберите **Variable Specs...** (рис. 2.1).

В поле **long name** введите функцию **=Binom(2;0,3;5)**, **OK.** На экране появляется сообщение: Expression OK. Recalculate the Variable now? (Формула

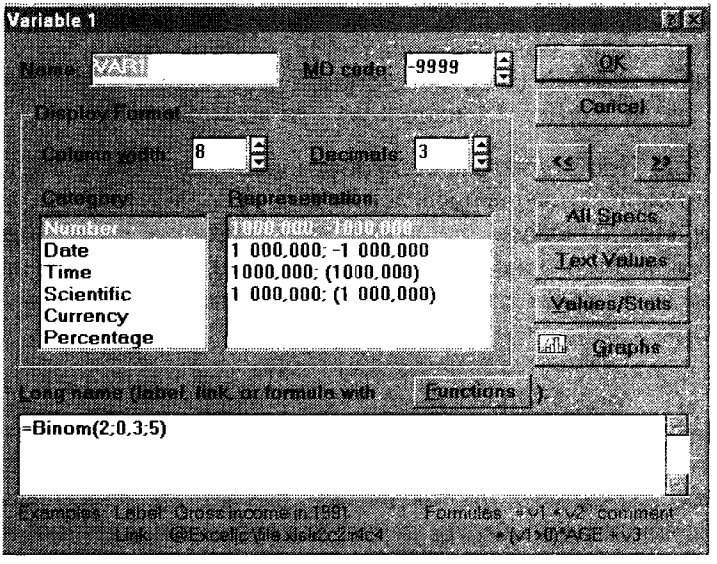


Рис. 2.1. Окно **Variable Specs...** для **Var1**

записана верно. Пересчитать переменную сейчас?), да. Поля переменной Var1 заполняются числами 0,309 = $P[X = 2]$. (Проверьте!). Аналогично, введя в поле **long name** функцию **=IBinom(4;0,3;5)**, получим $P[X \le 4] = 0{,}998$.

В поле **long name** можно вставить соответствующую формулу, выбрав ее из списка после нажатия кнопки **FUNCTIONS** (функции) в окне **Variable Specs...** . (рис. 2.1) и нажав кнопку **Insert**. Список функций показан на рис. 2.2.

Если в окне списка функций (рис. 2.2) нажать кнопку **SYNTAX**, то появится текст, поясняющий правила для записи формул и назначение функ-

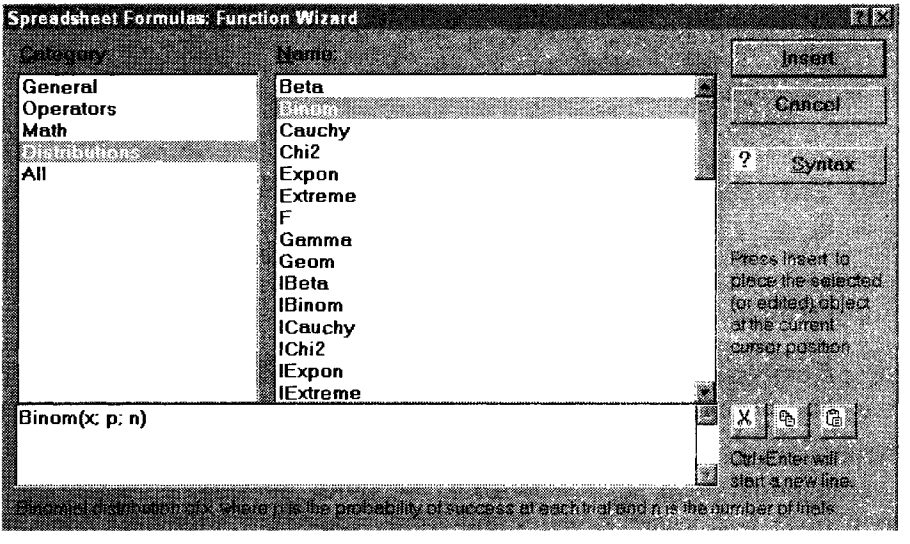


Рис. 2.2. Список функций

ций. Функции, используемые в вероятностных расчетах, поясняются в дополнительном тексте, который появляется, если щелкнуть мышью по **Distribution functions and their integrals** (функции распределения и их интегралы).

Решите снова пример 2.1 используя эти возможности.

Пример 2.2. Построим график вероятностей для случайной величины X имеющей биномиальное распределение $B(n, p)$.

Решение. Предположим, что число экспериментов $n = 15$, а вероятность появления «успеха» в одном эксперименте $p = 0{,}3$. Вычислим вероятности p_k появления k «успехов» в пятнадцати экспериментах по формуле биномиального распределения

$$p_k = C_n^k p^k (1-p)^{n-k},\ k = 0,\ 1,\ 2,\ ...,\ 15 \text{ при } n = 15 \text{ и } p = 0{,}3.$$

Значения k определим с помощью оператора V0. Оператор V0 вводит в переменную номера строк: 1, 2, 3, ... соответственно, оператор V0-1 вводит номера строк: 0, 1, 2... .

Чтобы вычислить и занести вероятности p_k в переменную (например) Var4, нужно в поле long name переменной Var4 ввести формулу

=Binom(V0-1;0,3;15).

После пересчета по этой формуле в столбце Var4 получим значения вероятностей: $p_0 = 0{,}004748$; $p_1 = 0{,}030520$, ...; $p_{15} = 0{,}000001$. (Чтобы вывести значения с такой точностью надо увеличить число десятичных знаков используя соответствующую кнопку на панели инструментов (см. рис. 1.2)). Для построения графика войдем в меню **Graphs → Stats 2D Graphs → Bar/Column Plots...** .

Установим значение переменной Variables: Var4, OK. График показан на рис. 2.3.

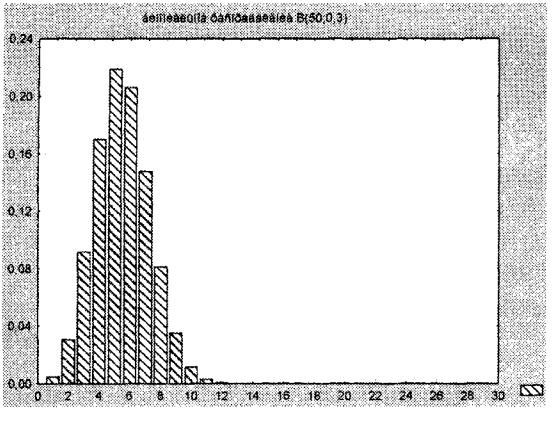


Рис. 2.3. Значения вероятностей биномиального распределения B(15;0,3)

Чтобы увидеть как изменяются свойства биномиального распределения при увеличении числа экспериментов, увеличим значение n с 15 до 50. Для этого в поле **long name** переменной Var4 установим: **=Binom(V0-1;0,3;50)**. После пересчета построим график (рис. 2.4).

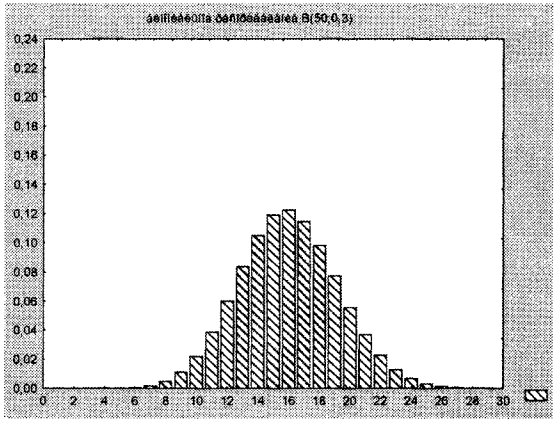


Рис. 2.4. Значения вероятностей биномиального распределения B(50;0,3)

Распределение на рис. 2.4 по своему виду близко к нормальному распределению с математическим ожиданием $m = 50 \cdot 0{,}3 = 15$ (вспомните теорему Муавра—Лапласа, Приложение, П.4).

Аналогично вычисляются вероятности и рассматриваются свойства распределения Пуассона (используем функцию **Poisson(*k*; λ)**), **Poisson(*k*; λ)** $= \dfrac{\lambda^k}{k!} e^{-\lambda} = P[X = k]$, $k = 0, 1, 2, \ldots$ и геометрического распределения (используем функцию **Geom(x;p)**).

Geom(k;p) $= P[X = k] = (1 - p)^k \cdot p = q^k p$, $k = 0, 1, 2, \ldots$ (см. Приложение П2.6).

С помощью функций **IPoisson(x;λ)** и **IGeom(x;p)** вычисляются соответствующие накопленные вероятности: $P[X \le x]$.

Постройте график для вероятностей биномиального распределения B(50;0,01). Убедитесь, что полученное распределение близко к распределению Пуассона с параметром $\lambda = 50 \cdot 0{,}01 = 0{,}5$.

2.2. Вычисление вероятностей и квантилей для непрерывных случайных величин

Для непрерывных распределений, например, нормального (см. Приложение, П.3.5) вычисляются три функции:

1. ***Normal***$(x; \mu; \sigma)$ — плотность распределения;

$\text{normal}(x; \mu; \sigma) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$ — это значение функции плотности нормального распределения $N(\mu, \sigma^2)$ в точке x, где μ — математическое ожидание, σ^2 — дисперсия, а σ — среднеквадратическое (стандартное) отклонение.

2. ***Inormal***$(x; \mu; \sigma) = \dfrac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{\frac{(t-\mu)^2}{2\sigma^2}} dt$ — значение функции распределения нормального распределения $N(\mu; \sigma^2)$ в точке x.

3. ***Vnormal*** (x; μ; σ) — значение функции обратной к функции ***Inormal***(x; μ; σ) в точке x. Обратная функция к функции распределения используется для вычисления квантилей и моделирования распределений случайных величин (см. ниже, п. 2.3).

Квантиль распределения случайной величины X порядка p, $[0 < p < 1]$ будет обозначаться x_p. Напомним, что для непрерывной случайной величины X, x_p есть решение уравнения: $F(x_p) = p$ или $P[X < x_p] = p$ (см. Приложение, П.3.3).

Таким образом, **Vnormal(p;0;1)** $= x_p$, где x_p — квантиль порядка p стандартного нормального распределения $N(0, 1)$ с математическим ожиданием $\mu = 0$ и среднеквадратическим отклонением $\sigma = 1$. Например, **Vnormal(0,95;0;1)** = 1,645 (проверьте!). Таблица квантилей распределения $N(0, 1)$ приведена в Приложении 3.

Пример 2.3. Использование функций плотности $f(x)$ и функции распределения $F(x)$ для построения графиков.

Построим соответствующие графики для $X \sim N(m = 7, \sigma^2 = 4)$ — случайной величины, имеющей нормальное распределение с математическим ожиданием $m = 7$ и дисперсией $\sigma^2 = 4$. Для построения графиков $f(x)$ и $F(x)$ нужно задать сетку значений x с постоянным шагом. Для этого в файле данных для **Var1** в поле **long name** введите формулу =(V0-7)/10. Напомним, что оператор V0 вводит номера строк: 1, 2, ..., 100. После пересчета по этой формуле в переменной **Var1** получим значения x: –0,6; –0,5; –0,3; ...

Для переменных **Var2** и **Var3** в поле **long name** соответственно введем формулы: **=Normal(v1;7;2)** и **=INormal(v1;7;2)**. После пересчета в переменной V2 получим значение плотности распределения, а в V3 — значение функции распределения.

Чтобы построить графики войдем в меню **Graphs** → **Stats 2D Graphs** → **Scatterplots...** Тип графика **Double-Y** (рис. 2.5). Установим значения переменных:

- **X — Var1**;
- **Left Y — Var2**;
- **Right Y — Var3**;
- **Fit → Off**.

После выполнения получим графики (рис. 2.6).

Аналогичным образом задаются функции плотности, функции распределения и функции обратные к функциям распределения для других непрерывных случайных величин. Приведем список обозначений функций плотности для распределений, которые будут использоваться в дальнейшем изложении.

Expon(*x*; λ) — плотность экспоненциального распределения с параметром λ (см. Приложение, П.3.4);

Chi2(*x*; *k*) — плотность распределения χ^2 с k степенями свободы;

F(*x*; k_1; k_2) — плотность распределения Фишера с k_1 и k_2 степенями свободы;

Student(*x*; *k*) — плотность распределения Стьюдента с k степенями свободы.

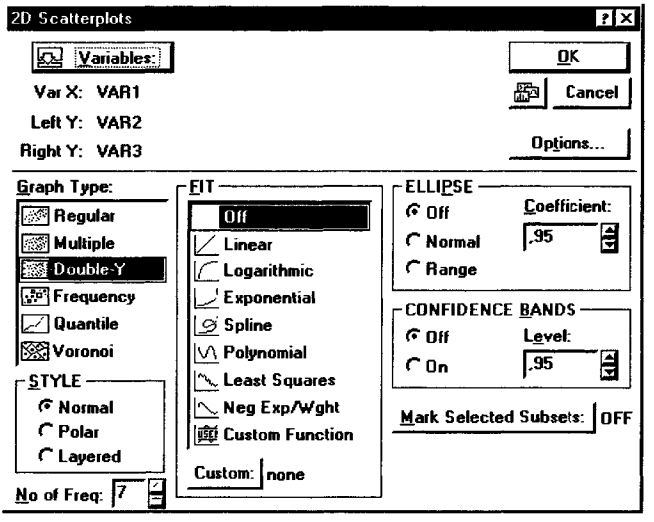


Рис. 2.5. Установка вида графика

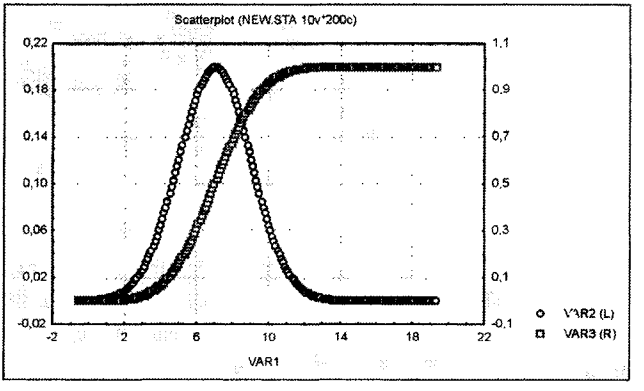


Рис. 2.6. Графики плотности и функции распределения $N(7; 4)$

Необходимые сведения по трем последним распределениям, играющим, наряду с нормальным распределением, важнейшую роль в статистических расчетах, приведены в главе 3 (см. п. 3.2.2).

Работа с Probability Distr. Calculator (вероятностный калькулятор)

Для непрерывных распределений все необходимые расчеты можно выполнить, используя вероятностный калькулятор. Войдите в модуль **Basic Statistics and Tables** → **Analysis** → **Startup Panel** → **Probability Calculator** (рис. 2.7).

Чтобы понять работу с вероятностным калькулятором, откройте таблицу функции распределения стандартного нормального закона $N(0; 1)$:

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

Эта таблица приведена в Приложении 3. В поле **Distribution** (слева) выберите нормальное распределение: **Z(normal)**, математическое ожидание — ***mean*** = 0, стандартное отклонение — **st.dev.** = 1. Если теперь ввести значения ***x*** (например $x = 0{,}1$) и щелкнуть левой кнопкой на **Compute**, то в поле p появится число равное вероятности события $Z < x$, где Z — случайная величина, имеющая нормальное распределение $N(0, 1)$, $Z \sim N(0, 1)$, т. е. $p = P[Z < x] = \Phi(x)$.

Так, при $x = 0{,}1$, получим $p = P[Z < 0{,}1] = \Phi(0, 1) = 0{,}539828$ (проверьте по таблице значения функции $\Phi(x)$!).

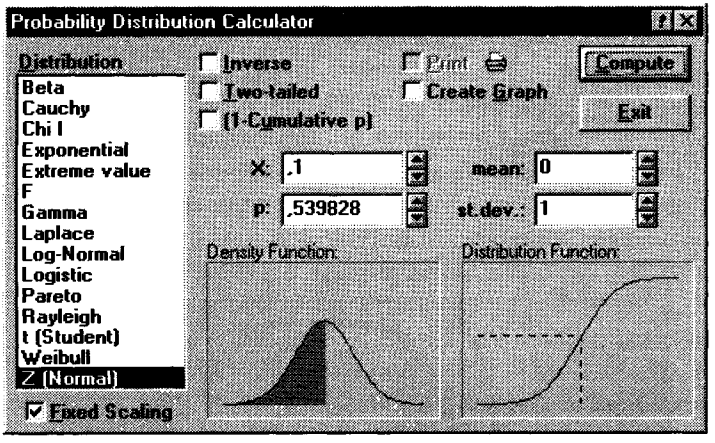


Рис. 2.7. Окно вероятностного калькулятора

На графике плотности распределения (**Density Function**) вычисленная вероятность p равна заштрихованной части площади под графиком плотности распределения (вся площадь под графиком плотности = 1). На графике функции распределения (**Distribution Function**), p — это ордината $\Phi(x)$, соответствующая абсциссе x. Точные значения на графиках можно получить щелкнув левой кнопкой на **Create graph**.

Напомним определение квантили порядка p для непрерывного распределения случайной величины X с плотностью $f(x)$ и функцией распределения $F(x)$: это число x_p удовлетворяющее условию

$$P[X < x_p] = F(x_p) = p \quad \text{или} \quad \int_{-\infty}^{x_p} f(x)dx = p.$$

Таким образом, используя вероятностный калькулятор можно: 1) по значению квантили $x_p = x$ найти порядок квантили p; 2) по значению порядка квантили p найти соответствующую квантиль $x_p = x$.

Задания

1. Составьте таблицу квантилей порядка 0,01; 0,05; 0,1; 0,9; 0,95; 0,99 для распределений $N(0, 1)$, Стьюдента (t-распределения) ($k = 10$), распределения хи-квадрат ($k = 19$), распределения Фишера (F) ($k_1 = 10$, $k_2 = 15$).

2. Для случайной величины $X \sim N(1, \sigma^2 = 4)$ вычислить вероятности следующих событий: $P[X < 2]$, $P[X \geq 3]$, $P[0 < X < 3]$, $P[|X| < 1]$, $P[|X| \geq 2]$, $P[|X - 1| < 1]$, $P[|X - 2| > 1]$.

Расчеты проведите используя функцию $\Phi(x)$ (см. Приложение, П.3.5) и сравните результаты со значениями, вычисленными с помощью вероятностного калькулятора. Покажите на графиках плотности распределения те фигуры, площади которых соответствуют указанным вероятностям.

3. Объясните, что вычисляет вероятностный калькулятор, если используются опции **two-tailed** и **1-Cumulative** *p*. Приведите примеры расчетов с использованием этих опций.

Построение графиков плотностей и функций распределений с использованием опции Graphs

График любой функции из списка функций, используемых в вероятностных расчетах, строится так: **Graphs** → **Stat 2D Graphs** → **Custom Function Plots...**

Custom Function: в поле введем, например, **normal(*x*; 7; 2)** — функцию плотности нормального распределения с $m = 7$ и $\sigma = 2$; далее надо ввести диапазон изменения x: $x_{min} = 0$, $x_{max} = 14$ и нажать **OK**. График плотности приведен на рис. 2.8.

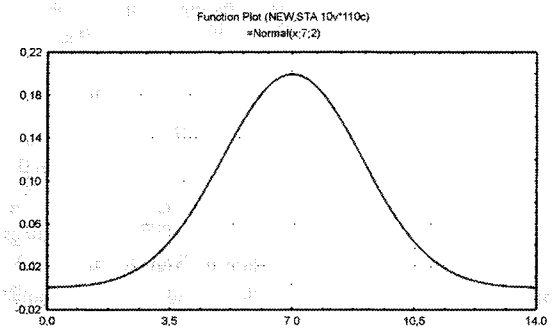


Рис. 2.8. График плотности нормального распределения с параметрами $m = 7$, $\sigma = 2$

2.3. Моделирование распределений случайных величин

В задачах статистического анализа сложных систем, например при разработке систем автоматического проектирования (САПР), исследовании систем массового обслуживания, широко используется метод моделирования выборки из генеральной совокупности с заданным законом распределения.

Пусть случайная величина X имеет функцию распределения $F(x)$. Как известно из теории вероятностей, случайная величина $Y = F(X)$ имеет равномерное распределение $R(0, 1)$ (см. Приложение, П.3.4). Отсюда следует, что случайная величина X может быть получена из равномерно распределенной случайной величины Y по формуле $X = F^{-1}(Y)$, где F^{-1} — функция, обратная к F (заведомо существующая для случайных величин непрерывного типа).

Метод моделирования выборки из генеральной совокупности с законом распределения $F(x)$ реализуется следующим алгоритмом:

$$x_j = F^{-1}(y_j), \quad j = 1, 2, ..., n,$$

где $y_1, y_2, ..., y_n$ — выборка из генеральной совокупности с равномерным распределением $R(0, 1)$, являющаяся последовательностью случайных чисел.

Алгоритм получения выборки из генеральной совокупности с законом распределения $F(x)$ поясняется на рис. 2.9.

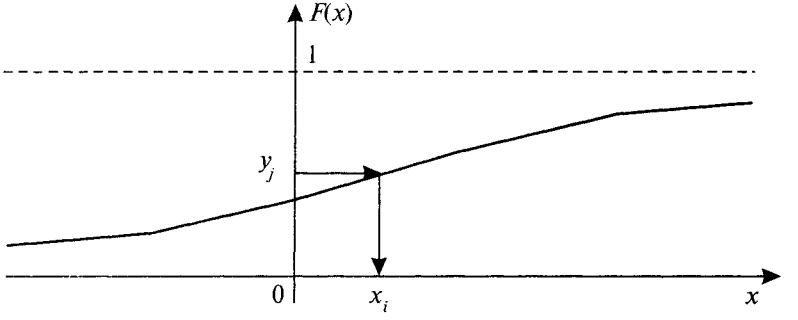


Рис. 2.9. Получение выборки из генеральной совокупности с заданным законом распределения $F(x)$

Таким образом, для моделирования выборки из непрерывного распределения с функцией распределения $F(x)$, нужно сначала получить выборку из генеральной совокупности, имеющей равномерное распределение $R(0, 1)$, а затем использовать функцию, обратную к функции распределения $F(x)$, соответствующей случайной величины.

В пакете **STATISTICA** распределение $R(0, 1)$ моделируется с помощью функции: **=rnd(1)**.

Для примера смоделируем выборку из равномерного распределения $R(0, 1)$ и запишем результат в переменной **Var4**. Курсор на поле **Var4** → щелчок правой кнопкой мыши → **Variable Specs...** → в поле **long name** введем формулу: **=rnd(1)** → **OK**.

После выполнения пересчета 200 значений переменной **Var4** будут заполнены числами, представляющими случайную выборку наблюдений из генеральной совокупности, имеющей равномерное распределение $R(0, 1)$. Чтобы получить выборку из нормального распределения $N(m = 7; \sigma^2 = 4)$ и записать ее в переменную **Var5**, нужно в поле **long name** переменной **Var5** записать формулу: **=Vnormal(V4;7;2)**.

Можно в качестве аргумента функции **Vnormal** сразу записать **rnd(1)**, тогда соответствующая формула будет: **=Vnormal(rnd(1);7;2)**.

Аналогично моделируются выборки для любого непрерывного распределения.

Задание

Методом моделирования получите выборки объема 200 для следующих распределений: хи-квадрат ($k = 10$), Стьюдента $T(k)$, $k = 7$ и Фишера $F(k_1, k_2)$, $k_1 = 5$, $k_2 = 8$.

Постройте гистограммы выборок. Для этого поставьте курсор на имя переменной, щелкните по правой кнопке мыши, в выпадающем меню выберите **Quick Stats Graphs ⇒ Histogram of**.

Сравните полученные гистограммы с графиками плотностей для моделируемых распределений.

2.4. Практические работы по теории вероятностей

2.4.1. Работа 1. Законы больших чисел. Центральная предельная теорема и ее следствия

Рассмотрим несколько статистических экспериментов, иллюстрирующих законы больших чисел и центральную предельную теорему (формулировки теорем приведены в Приложении, П.4).

Напомним основные следствия из центральной предельной теоремы и теоремы Муавра—Лапласа:

1. Пусть $x_1, x_2, \ldots, x_n$ — выборка объема n из некоторой генеральной совокупности с математическим ожиданием m и дисперсией σ^2.

Выборочное среднее

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

при больших n имеет асимптотически (при $n \to \infty$) нормальное распределение $N\left(m, \frac{\sigma^2}{n}\right)$.

2. Пусть h — относительная частота появления события A в n независимых случайных экспериментах, тогда h имеет асимптотически (при $n \to \infty$) нормальное распределение $N\left(p, \frac{pq}{n}\right)$, где p — вероятность появления события A в одном эксперименте, $q = 1 - p$.

3. Пуассоновская случайная величина с параметром λ имеет асимптотически (при $\lambda \to \infty$) нормальное распределение $N(\lambda, \lambda)$.

4. Случайная величина $\chi^2(k)$, имеющая распределение хи-квадрат с k степенями свободы, имеет асимптотически (при $k \to \infty$) нормальное распределение $N(k, 2k)$.

5. Случайная величина $T(k)$, имеющая распределение Стьюдента с k степенями свободы, имеет асимптотически (при $k \to \infty$) стандартное нормальное распределение $N(0, 1)$.

1. Теорема Бернулли

Теорема Бернулли утверждает, что в последовательности n независимых экспериментов при $n \to \infty$ частота h появления события A сходится по

вероятности к вероятности события A, $P(A) = p$. Это означает, что для любого сколь угодно малого $\varepsilon > 0$

$$P[|h - p| < \varepsilon] \to 1 \quad \text{при} \quad n \to \infty.$$

Отсюда следует, что при достаточно большом числе экспериментов n вероятность события, состоящего в том, что наблюдаемая частота h появления события A и его вероятность p будут отличаться не более чем на заданное значение ε, будет очень близка к 1, так что событие $|h - p| < \varepsilon$ можно считать *практически достоверным*. Найдем число экспериментов n, при котором событие $|h - p| < \varepsilon$ будет выполняться с вероятностью $1 - \alpha$ (α — мало). Воспользуемся следствием 2 из центральной предельной теоремы: при больших n частота h имеет приближенно нормальное распределение с математическим ожиданием p, ($M[h] = p$) и дисперсией $\frac{pq}{n}$, $\left(D[h] = \frac{pq}{n}\right)$:

$$h \underset{n\to\infty}{\sim} N\left(p, \frac{pq}{n}\right),$$

следовательно,

$$P[|h - p| < \varepsilon] \approx 2\Phi\left(\frac{\varepsilon}{\sqrt{pq/n}}\right) - 1, \quad \text{(см. Приложение, П.3.5).}$$

Используя очевидное неравенство $pq \le \frac{1}{4}$, получим

$$2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) - 1 \ge 2\Phi(2\varepsilon\sqrt{n}) - 1.$$

Для определения n имеем уравнение:

$$2\Phi(2\varepsilon\sqrt{n}) - 1 = 1 - \alpha,$$

следовательно,

$$\Phi(2\varepsilon\sqrt{n}) = 1 - \frac{\alpha}{2} \quad \text{и} \quad 2\varepsilon\sqrt{n} = u_{1-\frac{\alpha}{2}},$$

где $u_{1-\frac{\alpha}{2}}$ — квантиль порядка $1 - \frac{\alpha}{2}$ стандартного нормального распределения $N(0; 1)$:

$$\Phi(u_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}.$$

Окончательно получаем

$$n \ge \frac{(u_{1-\frac{\alpha}{2}})^2}{4\varepsilon^2}.$$

В частности, при $\varepsilon = 0{,}1$ и $1 - \alpha = 0{,}99$, $u_{0{,}975} = 2{,}576$ (проверьте, используя статистический калькулятор!), получим, что необходимое число экспериментов n составит не менее чем 166:

$$n \ge 166.$$

При $\varepsilon = 0{,}1$ и $1 - \alpha = 0{,}9973$ получим

$$n \geq 232.$$

Если взять $\varepsilon = 0{,}05$ и $1 - \alpha = 0{,}99$, необходимое число экспериментов составит

$$n \geq 664.$$

Выполнение расчетов в пакете STATISTICA

Рассмотрим эксперимент с подбрасыванием симметричной монеты $p = \frac{1}{2}$, $q = 1 - p = \frac{1}{2}$. Результаты этого эксперимента моделируем, генерируя случайную величину X_k:

$$X_k = \begin{cases} 1, & \text{если в } k\text{-м подбрасывании выпал «герб»} \\ 0, & \text{если в } k\text{-м подбрасывании выпала «решка»} \end{cases}, \quad k = 1, 2, \ldots, n.$$

Частота h появления герба при n подбрасываниях монеты, равна

$$h = \frac{\sum_{k=1}^{n} X_k}{n}.$$

Чтобы выполнить расчеты в пакете STATISTICA, переключитесь в модуль **Basic Stat**. Создайте новый файл, для этого выберите **New Data** из меню **File**. Эта команда доступна также по комбинации клавиш CTRL + N. В появившемся диалоговом окне **New Data: Specify File Name** введите имя файла и нажмите **OK**. STATISTICA автоматически откроет пустую электронную таблицу. В заголовке окна электронной таблицы отображается имя файла и его размер. Размер таблицы по умолчанию принят **10v*10c** (10 переменных и 10 пронумерованных строк). Преобразуем ее в таблицу с 664 строками. Для этого воспользуемся кнопкой Cases на панели инструментов и командой **Add** (рис. 2.10).

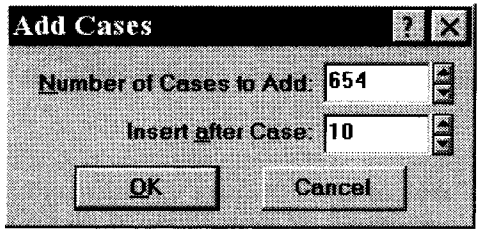


Рис. 2.10

В появившемся диалоговом окне в разделе **Number of Cases to Add** задайте количество добавляемых случаев (654) и нажмите **OK**.

Генерируем значения случайных величин X_k, $k = 1, 2, \ldots, 664$ в переменной **VAR1**. Дважды щелкните на **VAR1** левой кнопкой мыши. На экране появится окно спецификации данной переменной (рис. 2.1). В поле **long name** введите функцию: **=trunc(rnd(1) + 0,5)** и нажмите **OK**.

Оператор **trunc** вычисляет целую часть значений случайной величины, имеющей равномерное распределение $R(0,5; 1,5)$. Далее вычисляем частоту h, определяя среднее значение (mean) для **VAR1**. Для этого войдите в меню **Analysis** и выберите команду **Quick Basic Stats**. Результаты расчетов приведены на рис. 2.11.

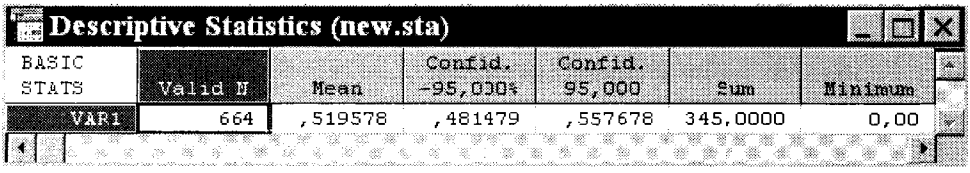


Рис. 2.11. Результаты расчетов

Сравнивая полученное значение с 0,5, убедитесь, что

$$|h - 0{,}5| \le 0{,}05.$$

Задание

Проведите эксперимент 5 раз, запишите значения h и $|h - 0{,}5|$.

2. Теорема Хинчина

Теорема Хинчина утверждает, что среднее арифметическое

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

для независимых случайных величин X_i, $i = 1, 2, \ldots, n$, имеющих одно и тоже распределение и конечное математическое ожидание m, сходится по вероятности при $n \to \infty$ к m. Таким образом, при заданном ε и достаточно большом n событие

$$\left| \frac{X_1 + X_2 + \ldots + X_n}{n} - m \right| < \varepsilon$$

можно считать практически достоверным.

Найдем число слагаемых n, при котором событие $|\overline{X} - m| < \varepsilon$ будет выполняться с вероятностью $1 - \alpha$, (α — мало). Воспользуемся следствием 1 из центральной предельной теоремы: при $n \to \infty$ среднее арифметическое независимых одинаково распределенных случайных величин X_i, $M[X_i] = m$, $D[X_i] = \sigma^2$, $i = 1, 2, \ldots, n$, имеет приближенно нормальное распределение с математическим ожиданием m и дисперсией σ^2/n:

$$\overline{X} \sim N(m, \sigma^2/n).$$

Отсюда следует, что

$$P[|\overline{X} - m| < \varepsilon] \approx 2\Phi\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right) - 1.$$

Для определения n имеем уравнение

$$2\Phi\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right) - 1 = 1 - \alpha,$$

или

$$\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) = 1 - \frac{\alpha}{2}.$$

Отсюда получаем

$$n \geq \frac{(u_{1-\frac{\alpha}{2}})^2 \sigma^2}{\varepsilon^2}.$$

В частности, если X_i — независимые равномерно распределенные случайные величины, $X_i \sim R(0, 1)$, $i_i = 1, 2, ..., n$, $m = 0{,}5$; $\sigma^2 = \frac{1}{12}$, при $n = 8$ событие

$$\left|\frac{X_1 + ... + X_8}{8} - 0{,}5\right| < 0{,}2$$

выполняется с вероятностью 0,968. Проверьте!

Выполнение в пакете STATISTICA

Создайте новый файл данных и присвойте ему имя. Появившуюся таблицу **10v * 10c** преобразуйте в таблицу с 8-ю переменными. Для этого на панели инструментов нажмите кнопку **Vars** и в выпадающем меню выберите **Delete** и удалите переменные **Var9** и **Var10.**

Генерируем значения 8-ми случайных величин $X_i \sim R(0, 1)$, $i = 1, 2, ..., 8$. Для этого:

1) выделите всю таблицу, щелкнув мышью по полю на пересечении строк и столбцов;

2) выберите в меню **Edit** команду **Fill/Standardize Block**. В выпадающем меню выберите команду **Fill Random Values** и нажмите **OK**.

Таким образом, мы получили по 10 реализаций (по числу строк) 8-ми независимых случайных величин $X_i \sim R(0, 1)$.

Вычислим среднее арифметическое (***means***) по каждой реализации (по каждой строке). Для этого щелкните правой кнопкой мыши на **VAR1**, в появившемся выпадающем меню выберите команду **Block Stats/Rows**, а затем **Means** (средние значения) и нажмите **OK**.

В 9-м столбце будут вычислены 10 средних арифметических. Выпишите значения $\overline{X}_i$ и значения $\left|\overline{X}_i - 0{,}5\right|$, $i = 1,2, ..., 10$.

Подтверждается ли утверждение теоремы Хинчина?

Задание

1. Проверьте утверждение теоремы Хинчина для независимых случайных величин X_i, имеющих экспоненциальное распределение с параметром $\lambda = 2$, $i = 1, 2, ..., 8$; $\varepsilon = 0{,}2$; $1 - \alpha = 0{,}95$.

Указание: для моделирования используйте следующую возможность записи функций в поле **long name** для каждой переменной. Поставьте курсор

на имя любой переменной, щелкните правой кнопкой мыши, выберите **Variable Specs**. В появившемся окне нажмите кнопку **All Specs** (все спецификации).

В таблице спецификаций поля **long name** можно заполнять последовательно, используя процедуру копирования. В меню выберите **Edit** и команду **Copy**, затем установите курсор в новое поле и выполните команду **Paste**.

2. Проверьте, что если X_i имеет распределение Коши **(Cauchy)**, то утверждение теоремы Хинчина не выполняется. Почему?

3. Центральная предельная теорема

Центральная предельная теорема утверждает, что сумма n независимых случайных величин при определенных условиях, асимптотически (при $n \to \infty$), имеет нормальное распределение. В случае, когда все слагаемые независимы и имеют одно и тоже распределение, например $X_i \sim R(0, 1)$, $i = 1, 2, ..., n$, эти условия выполнены.

Выполнение теоремы демонстрируем следующим образом. Генерируем результаты наблюдений восьми случайных величин $X_i \sim R(0, 1)$, $i = 1, 2, ..., 8$. Образуем суммы:

$$S_2 = X_1 + X_2, \quad S_3 = X_1 + X_2 + X_3, \ ...,$$

$$S_6 = X_1 + X_2 + ... + X_6, \quad S_8 = S_6 + X_7 + X_8.$$

Построим гистограммы для X_1, S_2, S_3, ..., S_8. Графики гистограмм по мере увеличения числа слагаемых приближаются к симметричной колоколообразной кривой — графику плотности нормального распределения.

Выполнение в пакете STATISTICA

Создайте новый файл данных **12v * 100c**. Дважды щелкните на **VAR1**, в появившемся окне нажмите на кнопку **Variable Specs**, а затем **All Specs** (все спецификации).

Задайте последовательно имена переменным: X_1, X_2, ..., X_7, X_8, S_2, S_4, S_6, S_8. В переменные X_1, X_2, ..., X_7, X_8 сгенерируем выборки из $R(0; 1)$, записав в поля **long name** выражение **=rnd(1)**.

В поля **long name** переменных S_2, S_4, S_6, S_8 запишите соответствующие функции, чтобы вычислить суммы:

$$X_1 + X_2, \quad S_2 + X_3 + X_4, \quad S_4 + X_5 + X_6, \quad S_6 + X_7 + X_8.$$

Результаты заполнения окна **All Specs** см. на рис. 2.12.

Далее установите курсор на имя любой переменной и щелкните по правой кнопке мыши, в выпадающем меню выберите **Recalculate** (пересчитать). В появившемся окне выберите кнопку **All Variables**, все переменные и нажмите **OK**.

Затем постройте гистограммы для S_6 и S_8. Для этого поставьте курсор на имя переменной, щелкните по правой кнопке мыши, в выпадающем меню выберите **Quick Stats Graphs** ⇒ **Histogram of...** ⇒ **Normal Fit**.

	Name	MD Code	Format	
1	X1	-9999	8.3	=rnd(1)
2	X2	-9999	8.3	=rnd(1)
3	X3	-9999	8.3	=rnd(1)
4	X4	-9999	8.3	=rnd(1)
5	X5	-9999	8.3	=rnd(1)
6	X6	-9999	8.3	=rnd(1)
7	X7	-9999	8.3	=rnd(1)
8	X8	-9999	8.3	=rnd(1)
9	S2	-9999	8.3	=x1+x2
10	S4	-9999	8.3	=s2+x3+x4
11	S6	-9999	8.3	=s4+x5+x6
12	S8	-9999	8.3	=s6+x7+x8

Рис. 2.12. Панель **All Specs**

Гистограммы S_6 и S_8 показывают, что распределение суммы независимых равномерно распределенных случайных вличин уже при числе слагаемых $n = 6$ и $n = 8$ начинает приближаться к нормальному распределению: график плотности нормального распределения показан на рис. 2.13 сплошной кривой.

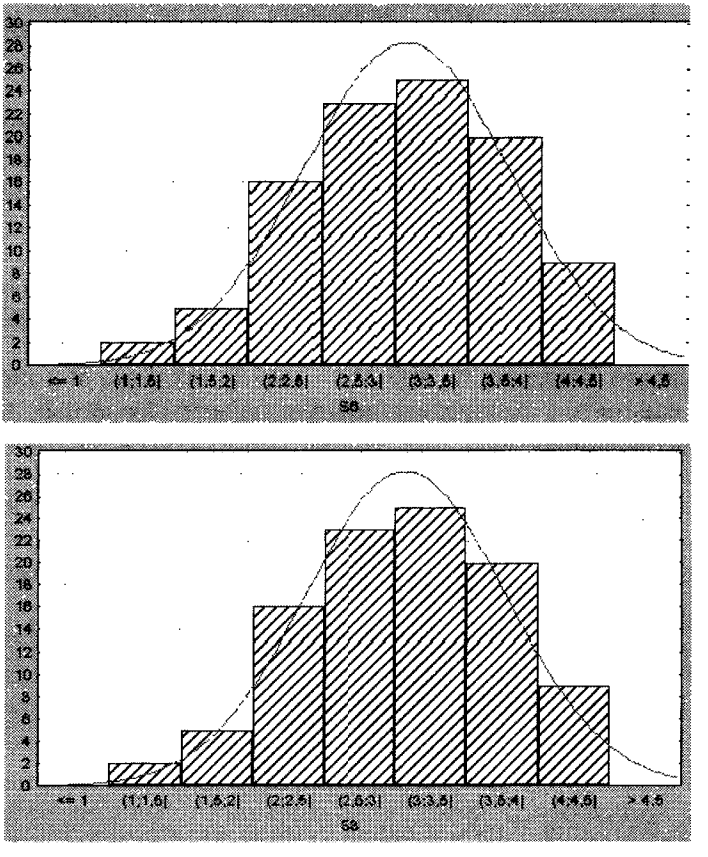


Рис. 2.13. Гистограммы сумм S_6 и S_8

Задание

Повторите вычисления для слагаемых, имеющих *экспоненциальное распределение*, $\lambda = 2$. Постройте гистограммы для сумм слагаемых.

2.4.2. Работа 2. Характеристики основных вероятностных распределений. Моделирование распределений случайных величин

1. *Основные понятия.* Дискретные и непрерывные случайные величины. Ряд распределения. Функция и плотность распределения и их свойства. Числовые характеристики случайных величин: математическое ожидание, дисперсия, среднее квадратическое отклонение, мода, медиана, начальные и центральные моменты, асимметрия, эксцесс, квантиль порядка p. Необходимые сведения приводятся в Приложении «Основы теории вероятностей», см. П2, П3 и в главе 3, см. п. 3.2.

2. *Задание.*

I. Изучить основные свойства, характеристики и зависимость от параметров следующих распределений: биномиального $B(n, p)$, пуассоновского $Pu(a)$, геометрического $Ge(p)$, равномерного $R(a, b)$, показательного $Ex(a)$, нормального $N(m, D(X))$, а также распределений хи-квадрат, Стьюдента $T(k)$ и Фишера $F(k_1, k_2)$. (Определения и основные свойства этих распределений приводятся в главе 3, п. 3.2.2).

Для каждого из распределений студент должен знать:

1) определение математической или физической модели, приводящие к данному распределению, а также область, где оно встречается и используется;

2) функцию и плотность распределения, параметры, математическое ожидание, дисперсию, особенности формы распределения и асимптотические свойства;

3) вычисление квантилей x_p заданного порядка и вероятностей;

4) связь данного распределения с другими распределениями.

II. Для своего варианта V, где V — номер студента в списке группы, определить параметры распределений для девяти случайных величин: X_1, X_2, ..., X_9. Вычисленные значения параметров нужно записать в таблицу.

1) X_1 имеет биномиальное распределение $B(n, p)$.

Параметры n и p определить по следующим формулам:

$$n = \begin{cases} 10, & \text{если } 1 \le V \le 10, \\ 15, & \text{если } 10 < V \le 20, \\ 20, & \text{если } 21 \le V, \end{cases}$$

$$p = \begin{cases} (V \bmod 10)/10, & \text{если } V \ne 10 \text{ и } V \ne 20, \\ 0{,}1, & \text{если } V = 10, \\ 0{,}5, & \text{если } V = 20. \end{cases}$$

Указание. Функция V mod a равна остатку от деления числа V на a.

Например: 21 mod 10 = 1; 21 mod 3 = 0; 21 mod 6 = 3.

2) X_2 имеет распределение Пуассона $Pu(\lambda)$; параметр λ определяется по формуле

$$\lambda = (V \bmod 3) + 5.$$

3) X_3 имеет геометрическое распределение $Ge(p)$; параметр p взять равным параметру p для биномиального распределения $B(n, p)$.

4) X_4 имеет равномерное распределение $R(a, b)$; параметры a и b определить по формулам:

$$a = (V \bmod 10) - 9,$$

$$b = a + 10.$$

5) X_5 имеет экспоненциальное распределение $Ex(\lambda)$; параметр λ взять равным значению p для биномиального распределения $B(n, p)$.

6) X_6 имеет нормальное распределение $N(m, D[X])$; значения параметров определить по формулам:

$$m = (V \bmod 10) - 5,$$

$$D[X] = (V \bmod 3) + 1.$$

7) X_7 имеет распределение хи-квадрат, $\chi^2(k)$. Параметр k определяется по формуле

$$k = (V \bmod 10) + 7.$$

8) X_8 имеет распределение Стьюдента, $T(k)$. Параметр k определяется по формуле

$$k = (V \bmod 10) + 2.$$

9) X_9 имеет распределение Фишера, $F(k_1, k_2)$. Параметры k_1 и k_2 определяются по формулам:

$$k_1 = (V \bmod 10) + 7,$$

$$k_2 = (V \bmod 5) + 8.$$

III. Выполнить следующие расчеты:

1. Для каждого из распределений $X_1 \ldots X_6$ определить *точные значения* математического ожидания $M[X]$, дисперсии $D[X]$, $P[2 \le X \le 4]$. Для непрерывных распределений X_4, X_5, X_6 вычислить значения квантилей порядков 0,25; 0,5; 0,75.

Параметры распределений и результаты представить в виде таблицы.

2. Вычислить и показать на графиках плотности распределений соответствующую вероятность $P[|X - m| \ge k\sigma]$ для X_4, X_5, X_6 при $k = 1, 2, 3$.

3. Сформулировать правило «3-х сигм» для X_6.

4. Исследовать асимптотические (при $n \to \infty$, $k \to \infty$, $k_1 \to \infty$, $k_2 \to \infty$) свойства распределений: биномиального при фиксированной вероятности успеха p, Стьюдента, хи-квадрат и Фишера. Привести эскизы графиков плотностей для исходных и асимптотических распределений.

5. Методом моделирования получить выборки объема 100 для каждого из непрерывных распределений $X_4...X_9$. Параметры распределений взять из таблицы, составленной для п. II. Используя полученные выборки, для каждого из распределений $X_4...X_9$ найти оценки $M[X]$, $D[X]$, квантилей порядков 0,25; 0,5; 0,75, а также оценки моды, медианы, эксцесса и асимметрии. Результаты занести в таблицу для п. II. Сравнить точные значения с оценками по выборке.

Указания.

1. Цель работы 2 состоит в том, чтобы, наряду с выполнением обычных расчетов, использовать возможности пакета STATISTICA.

2. При обработке результатов моделирования распределений (получении оценок) можно воспользоваться командой **Quick Basic.Stats**. (меню **Analysis**), либо опцией **Descriptive statistics** (описательные статистики). Эта опция открывается из стартовой панели модуля **Basic Statistics and Tables**.

Подробно работа в этой опции описана в главе 3 (см. п. 3.4, работа 1).

3. Оценки параметров, вычисляемые по выборке, это приближенные значения этих параметров. Проверьте это по своей таблице. Повторите процедуру моделирования и сравните полученные оценки с предыдущими результатами. Объясните полученные результаты и читайте главу 3 (особенно п. 3.2) — в этом суть статистических методов.

Глава 3
ОСНОВЫ СТАТИСТИЧЕСКИХ МЕТОДОВ

3.1. Основные понятия и методы статистического описания

3.1.1. Типы статистических данных

Статистические данные представляют собой наблюдаемые или измеряемые значения одного или нескольких признаков обследуемой совокупности объектов. Различают *количественные и качественные признаки.* Значения количественных признаков могут быть непрерывными (вес, суточная производительность, годовой объем производства) или дискретными (количество детей, год выпуска продукции, количество холодильников, проданных в течение дня). Примерами *качественных признаков* являются, например, пол, семейное положение, цвет кожи, качество товара. В свою очередь качественные признаки в зависимости от вида данных делятся на *номинальные* (классификационные) и *ординальные* (порядковые). Говорят также, что соответствующие качественные признаки измеряются в номинальной или порядковой шкале. Разница между этими шкалами состоит в следующем.

Признак, измеряемый *в номинальной шкале*, принимает одно значение из конечного числа заведомо установленных градаций. Примерами признаков, измеряемых в номинальной шкале, являются пол (мужской, женский), цвет, марка автомобиля, тип строительного материала, классификация животных и т. п. Статистические данные, измеряемые в номинальных шкалах, представляются в виде таблиц, в которых приводятся частоты появления той или иной градации признака. Часто номинальные данные появляются при обработке социологических опросов. Например, может представлять интерес вопрос о предпочтительности той или иной группы избирателей к каждому из кандидатов на пост президента России. В этом случае результаты опроса могут быть представлены в виде прямоугольной таблицы, содержащей m строк (по числу кандидатов) и n столбцов (по числу групп избирателей), в каждой клетке которой записывается число избирателей данной группы, отдавших предпочтение соответствующему кандидату. Таблицы таких данных называются *таблицами сопряженности* размера $m \times n$.

Значения качественных признаков, измеряемых *в ординальной шкале*, могут быть упорядочены. Примерами таких признаков являются тестовые баллы и школьные оценки, качество условий жизни (плохое, удовлетворительное, хорошее, очень хорошее), сила ветра, оцениваемая по шкале

Бофорта (штиль, слабый ветер, умеренный ветер, свежий ветер, шторм и т. д.). Для признаков, измеряемых в ординальных шкалах, операции сложения и вычитания не имеют смысла. Так, нельзя сказать, что студент, получивший на экзамене «пять» по статистике знает предмет на одну единицу лучше, чем студент, получивший по этому предмету «четыре», поскольку для знаний не существует единицы измерения. Однако можно сказать, что первый студент знает статистику лучше, чем второй.

Для представления значений ординальных признаков в числовой форме используется следующий способ. Все значения признака записываются в порядке возрастания в виде ряда. Каждому значению поставим в соответствие натуральное число, равное его номеру в ряду. Это число называется *рангом*. Например, качество условий жизни (плохое, удовлетворительное, хорошее, очень хорошее) будет представлено рангами 1, 2, 3, 4. Для ординальных признаков, представленных в виде рангов, разработаны специальные статистические методы, позволяющие измерять степень близости признаков (например, ранговая корреляция), проверять гипотезы о виде распределения, проводить дисперсионный анализ (более подробно см. главу 4).

Для данных, представленных в номинальной шкале, также не определены операции сложения и вычитания. Эти данные (в отличие от ординальных признаков) не могут быть упорядочены и, следовательно, оцифрованы с помощью рангов. Применяя специальные статистические методы для номинальных признаков, можно проверить гипотезы о независимости признаков и о принадлежности двух или нескольких выборок к одному виду. Для оцифровки номинальных признаков используются числовые метки. Выбор меток в зависимости от цели статистического анализа может проводиться по различным критериям [9].

3.1.2. Генеральная совокупность и выборка

Множество всех обследуемых объектов называется *генеральной совокупностью*. Если это множество содержит небольшое число элементов, то возможно полное обследование всех его элементов. Однако в большинстве случаев в силу того, что генеральная совокупность имеет очень много элементов либо ее элементы труднодоступны, либо по другим причинам обследуется некоторая часть генеральной совокупности — *выборка*. В этом случае основные характеристики генеральной совокупности (их называют статистиками: среднее, дисперсия и т. д.) оцениваются (т. е. определяются приближенно) по выборке. Соответствующие статистики называются «выборочное среднее», «выборочная дисперсия» и т. д. Очевидно, что не всякая выборка правильно отражает свойства генеральной совокупности. Например, нельзя судить о среднем душевом доходе населения по выборке, составленной из доходов служащих финансовых компаний. Выборка должна давать правильное, неискаженное представление о генеральной совокупности, или, как говорят, быть *репрезентативной*. Если свойства генеральной совокупности заранее неизвестны, то, за неимением лучшего, следует использовать *простой случайный выбор*. Это означает, что все элементы генеральной совокупности должны иметь равные шансы попасть в выборку.

Например, при выяснении мнения всех студентов университета по какому-либо вопросу, выборка, составленная из студентов первого курса, не будет репрезентативной. Процедуру случайного выбора можно организовать, например, так. Запишем фамилии всех студентов на отдельные карточки, которые затем тщательно перетасуем, и из всего множества карточек отберем нужное количество. Ответы выбранных таким способом студентов составят репрезентативную выборку. Если требуется, чтобы в выборке были представлены элементы различных групп, составляющих генеральную совокупность, используется процедура *типического отбора*. Так, если студенты первого курса составляют 15 % всех студентов университета, то и в выборке они должны составлять 15 %. В некоторых случаях необходимо учитывать не только курс, но и специализацию студентов, если это может повлиять на результаты опроса.

Как правило, статистические данные в силу ошибок измерений, влияния внешней среды, присущей индивидуумам случайной изменчивости, имеют разброс. Рассмотрим, например, результаты выборочного контроля партии расфасованной продукции. С большой вероятностью можно сказать, что в выборке не найдется ни одной пары пакетов, имеющих один и тот же вес. Основная задача статистики состоит в получении осмысленных заключений именно из такого типа данных, т. е. данных, подверженных случайной изменчивости.

Математическая модель статистических данных содержит детерминированную и случайную составляющие. В простейшей модели случайная компонента — это случайная величина. Например, математическая модель данных, представляющих вес расфасованной продукции, есть сумма двух величин: номинального веса m пакета (детерминированная компонента) и отклонения истинного веса пакета от номинального. Это отклонение (имеющее случайный характер) можно рассматривать как сумму очень большого числа случайных факторов всегда имеющих место в производственных условиях (износ оборудования, влажность и температура продукта и др.). Следовательно, в силу центральной предельной теоремы (см. Приложение, П.4) отклонение от номинального веса — это случайная величина, имеющая нормальное распределение $N(0, \sigma^2)$ с нулевым математическим ожиданием и некоторой дисперсией σ^2.

В связи с этим генеральная совокупность определяется как множество значений случайной величины X, представляющей модель данных. В рассмотренном примере это бесконечная генеральная совокупность, имеющая нормальное распределение $N(m, \sigma^2)$.

Выборка определяется следующим образом. Пусть случайная величина X в случайном эксперименте E наблюдается n раз в предположении, что условия проведения эксперимента, а следовательно, и распределение наблюдаемой случайной величины X не изменяются от эксперимента к эксперименту. Этот новый составной эксперимент связан с n-мерной случайной величиной — случайным вектором $(X_1, X_2, ..., X_n)$, где X_j — случайная величина, соответствующая j-му эксперименту. Очевидно, X_j, $j = 1, 2, ..., n$ — независимые в совокупности случайные величины, каждая из которых имеет тот же закон распределения, что и случайная величина X.

Закон распределения случайной величины X называется *распределением генеральной совокупности*, а случайный вектор $(X_1, X_2, ..., X_n)$ — *выборочным вектором*. Числа $x_1, ..., x_n$, получаемые на практике при n-кратном повторении эксперимента E в неизменных условиях, представляют собой реализацию выборочного вектора $(X_1, X_2, ..., X_n)$ и называются выборкой $(x_1, ..., x_n)$ объема n.

Выборку $(x_1, ..., x_n)$, при необходимости, можно рассматривать как точку выборочного пространства, т. е. множества, на котором задано распределение вектора $(X_1, X_2, ..., X_n)$.

Аналогично определяется выборка, когда случайный эксперимент E связан с несколькими случайными величинами. Например, выборка объема n из двумерной генеральной совокупности есть последовательность $(x_1, y_1), ..., (x_n, y_n)$ пар значений случайных величин X и Y, принимаемых ими при n независимых повторениях случайного эксперимента E.

Таким образом, в математической статистике выборка объема n понимается двояко: как последовательность n чисел, являющихся результатами наблюдения случайной величины X: $x_1, ..., x_n$, и как выборочный вектор, т. е. совокупность n независимых случайных величин $X_1, X_2, ..., X_n$, каждая из которых имеет одно и то же распределение, совпадающее с распределением наблюдаемой случайной величины X. При этом выборка как совокупность чисел $(x_1, ..., x_n)$ есть некоторая возможная реализация выборочного вектора $(X_1, X_2, ..., X_n)$. В каком именно смысле используется понятие выборки в том или ином месте, должно быть ясно, во-первых, из обозначений: числа обозначаются строчными буквами, а случайные величины — прописными, а во-вторых, из контекста. Понятие выборочного вектора является основополагающим в математической теории статистики: теории статистического оценивания, при построении доверительных интервалов, проверке статистических гипотез.

3.1.3. Представление данных в виде таблиц и графиков

Если элементы исходной выборки упорядочить по величине (т. е. представить их в виде *вариационного ряда*) и отметить вертикальными черточками (штрихами) их повторяемость, то получится *статистический ряд* выборки. Количество штрихов для данного элемента выборки называют его *частотой*.

Пример 3.1. Записать в виде вариационного и статистического рядов выборку

$$5, 3, 7, 10, 5, 5, 2, 10, 7, 2, 7, 7, 4, 2, 4.$$

Решение. Объем выборки $n = 15$.

Упорядочив элементы выборки по величине, получим вариационный ряд:

$$2, 2, 2, 3, 4, 4, 5, 5, 5, 7, 7, 7, 7, 10, 10.$$

Статистический ряд запишем в виде табл. 3.1.

Таблица 3.1. **Статистический ряд для примера 3.1**

Элементы, x_i	2	3	4	5	7	10
Штрихи	///	/	//	///	////	//
Частота, n_i	3	1	2	3	4	2

Для контроля находим: $\sum_i n_i = 15$.

Разность между максимальным и минимальным элементами выборки называется размахом R.

В примере 3.1 размах выборки R равен

$$R = x_{\max} - x_{\min} = 10 - 2 = 8.$$

При большом объеме выборки ее элементы объединяют в группы (разряды), представляя результаты опытов в виде *группированного статистического ряда.* Для этого интервал, содержащий все элементы выборки, разбивается на k непересекающихся интервалов. Вычисления значительно упрощаются, если интервалы имеют одинаковую длину $b \approx R/k$. После того, как интервалы для группировки выбраны, определяют частоты — количество n_i элементов выборки, попавших в i-й интервал (элемент, совпадающий с правой границей интервала, относится к последующему интервалу). Получающийся статистический ряд в верхней строке содержит середины интервалов группировки, а в нижней — частоты n_i ($i = 1, 2, ..., k$).

Наряду с частотами одновременно подсчитываются также *накопленные частоты* $\sum_{j=1}^{i} n_j$, *относительные частоты* n_i/n и *накопленные относительные частоты* $\sum_{j=1}^{i} n_j/n$, $i = 1, 2, ..., k$. Полученные результаты сводятся в таблицу, называемую *таблицей частот группированной выборки.*

Группировка выборки вносит погрешность в дальнейшие вычисления, которая растет с уменьшением числа интервалов. В зависимости от объема выборки число интервалов выбирают от 6 до 15.

Пример 3.2. Представить выборку из 55 наблюдений в виде таблицы частот, используя 7 интервалов группировки.

Выборка:

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

Решение. Размах выборки $R = 23,8 - 10,1 = 13,7$. Длина интервала группировки $b = 13,7/7 \approx 2$. В качестве первого интервала удобно взять интервал 10—12.

Результаты группировки сведены в табл. 3.2.

Таблица 3.2. **Таблица частот для примера 3.2**

Номер интервала, i	Границы интервала	Середина интервала, x_i	Частота, n_i	Накопленная частота, $\sum_{j=1}^{i} n_j$	Относительная частота, n_i/n	Накопленная относительная частота, $\sum_{j=1}^{i} n_j/n$
1	10—12	11	2	2	0,0364	0,0364
2	12—14	13	4	6	0,0727	0,1091
3	14—16	15	8	14	0,1455	0,2546
4	16—18	17	12	26	0,2182	0,4728
5	18—20	19	16	42	0,2909	0,7637
6	20—22	21	10	52	0,1818	0,9455
7	22—24	23	3	55	0,0545	1,0000

Для наглядного представления выборки используют *гистограмму* и *полигон* частот.

Гистограммой частот группированной выборки называется функция, постоянная на интервалах группировки и принимающая на каждом из них значения n_i/b, $i = 1,2, \ldots, k$ соответственно. Площадь ступенчатой фигуры над графиком гистограммы равна объему выборки n. Так как значения гистограммы пропорциональны самим частотам, обычно по оси ординат откладывают значения частот n_i, а не значения n_i/b.

Аналогично определяется гистограмма относительных частот. Площадь соответствующей ступенчатой фигуры для нее равна единице. При увеличении объема выборки и уменьшении интервала группировки гистограмма относительных частот является статистическим аналогом плотности распределения $f(x)$ генеральной совокупности.

Полигоном частот называется ломаная с вершинами в точках $(x_i, n_i/b)$, $i = 1, 2, \ldots, k$, а полигоном относительных частот — ломаная с вершинами в точках $(x_i, n_i/nb)$, $i = 1,2, \ldots, k$. Таким образом, полигон относительных частот получается из полигона частот сжатием по оси Oy в n раз. Как и в случае гистограммы, при построении полигонов по оси ординат откладывают значения частот или относительных частот.

Огивой (*многоугольником накопленных частот*) называется ломаная, вершины которой имеют абсциссы, совпадающие с правыми границами интервалов группировки, и ординаты, совпадающие со значениями накопленных частот для соответствующих интервалов. Если в качестве ординат вершин ломаной принимаются значения накопленных частот в процентах $((\sum n_i \cdot 100\,\%)/n)$, то полученный график называют процентной огивой (рис. 3.1).

Например, используя огиву, получим, что 50 % всех наблюдений не превышает значения 18,2, а значения, не превышающие 16, составляют 25,46 %.

Пусть $x_1, x_2, \ldots, x_n$ — выборка объема n из генеральной совокупности с функцией распределения $F(x)$ и пусть $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ — вариационный ряд выборки. *Выборочным* (*эмпирическим*) *распределением*, соответствующим $x_1, x_2, \ldots, x_n$, называется распределение дискретной случайной величины, при-

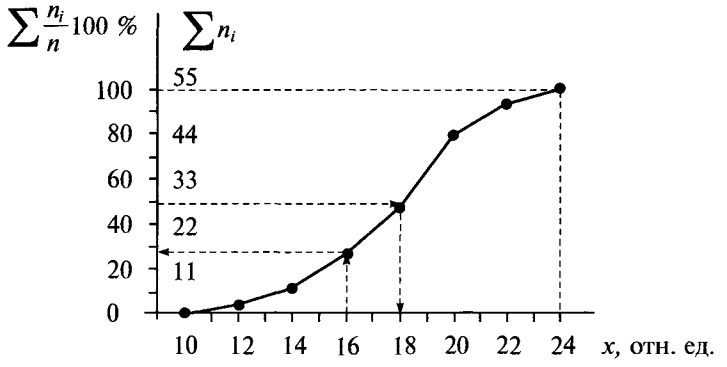


Рис. 3.1. Огива для данных примера 3.2

нимающей значения x_1, x_2, ..., x_n (среди которых могут быть и совпадающие) с вероятностями $1/n$. При этом вероятности для совпадающих значений складываются.

Эмпирическая функция распределения $F_n^*(x)$ — это ступенчатая функция, имеющая скачки, кратные числу $1/n$ в точках $x^{(1)}$, $x^{(2)}$, ..., $x^{(n)}$:

$$F_n^*(x) = \begin{cases} 0, & x \le x^{(1)} \\ k/n, & x^{(k)} < x \le x^{(k+1)},\ k = 1, 2, \ldots, n-1 \\ 1, & x > x^{(n)}. \end{cases}$$

Эмпирическая функция распределения $F_n^*(x)$ является статистическим аналогом функции распределения $F(x)$ генеральной совокупности.

Пример 3.2 (продолжение). Построить гистограмму, полигон частот и огиву.

Решение. По результатам группировки (см. табл. 3.2) строим гистограмму частот (рис. 3.2). Соединяя отрезками ломаной середины верхних оснований прямоугольников, из которых состоит полученная гистограмма, получаем соответствующий полигон частот (рис. 3.3). Огива приведена на рис. 3.1. На вертикальной оси нанесены две шкалы: накопленные частоты

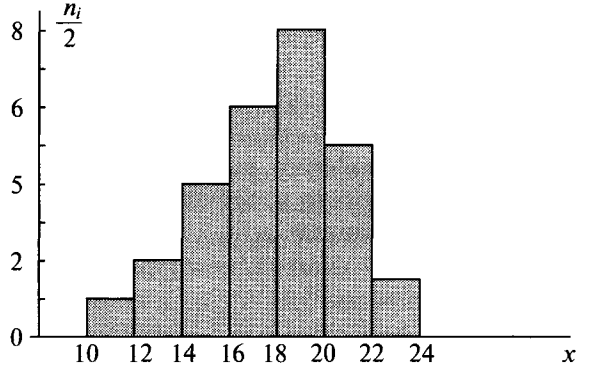


Рис. 3.2. Гистограмма к примеру 3.2

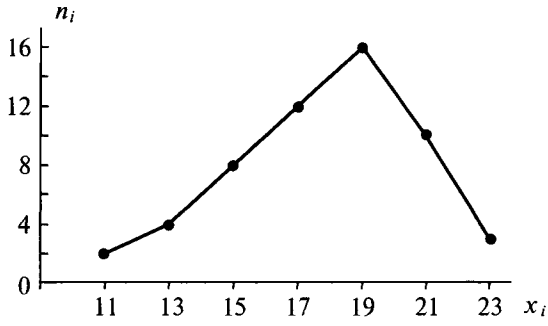


Рис. 3.3. Полигон частот к примеру 3.2

и их значения в процентах. Таким образом, по рис. 3.1 можно определить значение процентной огивы.

3.1.4. Оценка характеристик генеральной совокупности по выборке

Представление данных в виде гистограмм и полигонов дает информацию о распределении генеральной совокупности. Соответствующие гипотезы проверяются по статистическим критериям (см. п. 3.3). Однако часто требуется охарактеризовать генеральную совокупность некоторыми количественными показателями, которые определяют положение центра распределения, рассеяние (разброс) и асимметрию, что дает возможность сравнить одну совокупность данных с другой. По выборке можно определить приближенные значения (оценки) этих числовых характеристик, которые называются *выборочными характеристиками.*

Пусть $x_1, x_2, \ldots, x_n$ — выборка объема n из генеральной совокупности.

Выборочный начальный момент r-го порядка определяется как соответствующий момент выборочного распределения и вычисляется по формуле

$$\alpha_r^* = \frac{1}{n}\sum_{i=1}^{n} x_i^r, \; r = 1, 2, \ldots .$$

Выборочный начальный момент первого порядка обычно называют средним арифметическим и обозначают $\bar{x}$:

$$\bar{x} = \alpha_1^* = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Среднее арифметическое выборки является в определенном смысле «хорошей» оценкой математического ожидания m генеральной совокупности для симметричных распределений.

Выборочный центральный момент r-го порядка определяется по формуле

$$\mu_r^* = (1/n)\sum_{i=1}^{n} (x_i - \bar{x})^r .$$

Если выборка $x_1, x_2, \ldots, x_n$ представлена в виде статистического ряда, причем частота появления элемента x_i, $i = 1, 2, \ldots, k$, в выборке равна n_i, $\sum_{i=1}^{k} n_i = n$, то выборочные начальные и центральные моменты r-го порядка вычисляются по формулам:

$$\alpha_r^* = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^r, \tag{1}$$

$$\mu_r^* = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^r, \quad r = 1, 2, \ldots \tag{2}$$

Для выборки большого объема представленной в виде *таблицы частот группированной* выборки (см. пример 3.2, табл. 3.2), выборочные начальные и центральные моменты рассчитываются по формулам (1) и (2), в которых в качестве x_i, $i = 1, 2, \ldots, k$ принимается среднее значение i-го интервала группировки, а n_i — частота попадания в i-й интервал.

Начальные выборочные моменты α_r^* связаны с центральными выборочными моментами μ_r^*, следующими соотношениями:

$$\mu_2^* = \alpha_2^* - \bar{x}^2,$$

$$\mu_3^* = \alpha_3^* - 3\alpha_2^*\bar{x} + 2\bar{x}^3,$$

$$\mu_4^* = \alpha_4^* - 4\alpha_3^*\bar{x} + 6\alpha_2^*\bar{x}^2 - 3\bar{x}^4 \text{ и т. д.}$$

Для вычисления характеристик генеральной совокупности по выборке также используют *выборочные квантили* различных порядков.

Выборочная квантиль x_p порядка p, $1 < p < 0$, определяется как элемент вариационного ряда выборки $x^{(1)}$, $x^{(2)}$, ..., $x^{(n)}$ с номером $[np] + 1$, где $[a]$ — целая часть a.

В статистической практике используют ряд квантилей, имеющих специальные названия:

персентили: P_1, P_2, ..., P_{99} — квантили порядков 0,01; 0,02; ...; 0,99;

децили: D_1, D_2, ..., D_9 — квантили порядков 0,10; 0,20; ...; 0,90;

квартили: Q_1, Q_2, Q_3 — квантили порядков 0,25; 0,50; 0,75.

Для сгруппированных выборок оценки выборочных квантилей можно определить по *огиве* (см. пример 3.2 и рис. 3.1).

Оценка характеристик положения

Наиболее распространенными оценками характеристик положения являются *среднее арифметическое выборки* (*выборочное среднее*), *выборочная медиана* и *выборочная мода*. В дальнейшем будем опускать термин «выборочная», имея, однако, в виду, что любая оценка, вычисляемая по выборке является всего лишь приближенным значением соответствующей характеристики генеральной совокупности.

Для выборки объема n: $x_1, x_2, \ldots, x_n$ *среднее арифметическое* равно

$$\bar{x} = \frac{1}{n}\sum_i x_i.$$

Если выборка представлена в виде статистического ряда, где n_i есть частота элемента x_i, среднее арифметическое вычисляется по формуле

$$\bar{x} = \frac{1}{n}\sum_i n_i x_i.$$

По этой же формуле вычисляется среднее арифметическое и для сгруппированной выборки, причем в качестве x_i берется середина интервала.

Модой d называется элемент выборки, имеющий наибольшую частоту.

Пусть $x^{(1)}$, $x^{(2)}$, ..., $x^{(n)}$ — вариационный ряд выборки.

Медианой называется число h, которое делит вариационный ряд на две части, содержащие равное количество элементов. Если объем выборки n — нечетное число ($n = 2k + 1$), то $h = x^{(k+1)}$, т. е. h является средним элементом вариационного ряда. Если же $n = 2k$, то $h = \frac{1}{2}(x^{(k)} + x^{(k+1)})$.

Пример 3.3. Определить среднее, моду и медиану для выборки

5, 6, 8, 2, 3, 1, 1, 4.

Решение. Представим данные в виде вариационного ряда

1, 1, 2, 3, 4, 5, 6, 8.

Среднее $\bar{x} = \frac{1}{8}(1 + 1 + 2 + 3 + 4 + 5 + 6 + 8) = 3{,}75.$

Все элементы входят в выборку по одному разу, кроме 1, следовательно, мода $d = 1$. Так как $n = 8$, то медиана

$$h = \frac{1}{2}(3 + 4) = 3{,}5.$$

Мода предоставляет важную информацию для производителей товаров, конструкторов, работников торговли. Например, производитель часов должен знать, по какой цене распродается основная часть его продукции, в магазине бытовой техники должны знать наиболее популярные марки товара и т. д.

Соотношение между средним, модой и медианой в зависимости от формы функции плотности распределения показывают рис. 3.4—3.6.

Среднее геометрическое. Пусть x_1, x_2, ..., x_n — положительные числа. Среднее геометрическое вычисляется по формуле

$$x_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}.$$

Среднее геометрическое применяется для характеристики данных, заданных через равные промежутки времени, и определяет среднюю долю относительных изменений. Такие данные могут характеризовать явления роста, например, доходы по вкладам, эксплуатационные расходы, прирост населения и т. д.

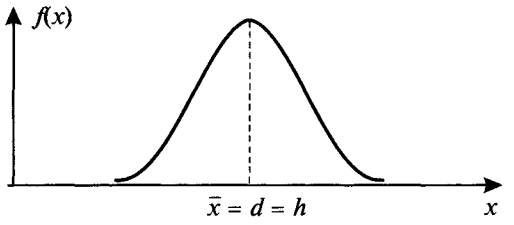


Рис. 3.4. Симметричное унимодальное распределение

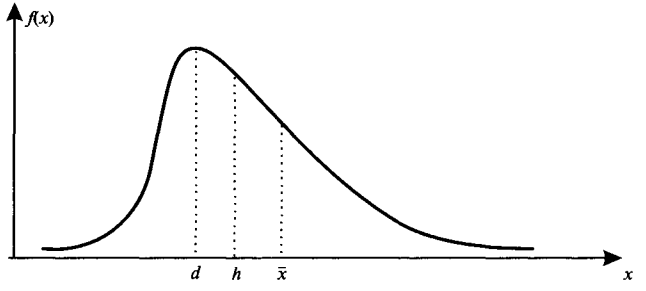


Рис. 3.5. Скошенное справа распределение (асимметрия > 0)

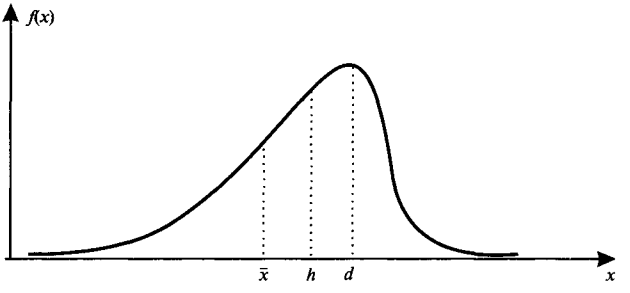


Рис. 3.6. Скошенное слева распределение (асимметрия < 0)

Между средним геометрическим и средним арифметическим имеет место следующее соотношение

$$\sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \le \frac{x_1 + x_2 + \ldots + x_n}{n},$$

причем знак равенства имеет место при $x_1 = x_2 = \ldots = x_n$.

Логарифм среднего геометрического вычисляется по формуле

$$\lg x_g = \frac{\sum_{i=1}^{n} \lg x_i}{n}.$$

Среднее гармоническое x_H вычисляется по формуле

$$\frac{1}{x_H} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}.$$

Среднее гармоническое используется для характеристики данных, размерность которых выражена отношением различных физических величин: км/ч; л/на 100 км и т. д.

Между средним арифметическим, средним гармоническим и средним геометрическим существует следующее соотношение

$$\bar{x} \geq x_g \geq x_H .$$

Оценка характеристик рассеяния

Наиболее распространенными мерами рассеяния являются *размах*, *средний межквартильный размах*, *персентильный размах*, *дисперсия* и *среднее квадратическое отклонение*.

Размах определяется как разность между максимальным и минимальным значениями выборки

$$R = x_{\max} - x_{\min} = x^{(n)} - x^{(1)}.$$

Пусть $x^{(1)}$, $x^{(2)}$, ..., $x^{(n)}$ — вариационный ряд выборки.

Вариационный ряд делится тремя квартилями (Q_1, Q_2, Q_3) на 4 равные части, т. е. Q_1 — это значение, ниже которого лежит 25 % наблюдений, Q_2 — 50 % наблюдений (Q_2 равен медиане), Q_3 — 75 % наблюдений.

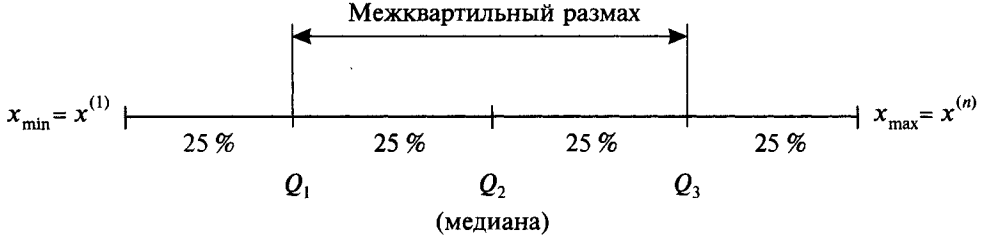


Рис. 3.7

Средний межквартильный размах равен половине разности верхнего и нижнего квартилей

$$\frac{Q_3 - Q_1}{2}.$$

Персентильный размах равен разности 90- и 10-го персентилей:

$$P_{90} - P_{10} = x_{0,9} - x_{0,1}.$$

Дисперсия для выборки негруппированных данных определяется по формуле

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - n \cdot \bar{x}^2}{n-1},$$

для сгруппированных данных — по формуле

$$S^2 = \frac{\sum n_i (x_i - \bar{x})^2}{n-1} = \frac{\sum n_i x_i^2 - n\bar{x}^2}{n-1}.$$

Оценку дисперсии можно вычислить, используя второй выборочный центральный момент μ_2^*:

$$S^2 = \frac{n}{n-1} \cdot \mu_2^*.$$

Среднее квадратическое отклонение S определяется как арифметический квадратный корень из дисперсии $S = \sqrt{S^2}$.

В качестве меры относительного разброса данных используют *коэффициент вариации*

$$V = \frac{S}{\bar{x}} \quad \text{или} \quad C_V = \frac{S}{\bar{x}} \cdot 100\ \%.$$

Выбор той или иной характеристики рассеяния выборки неоднозначен. Наиболее просто вычисляется размах, так же просто вычисляются средний межквартильный размах и персентильный размах, причем, как и все оценки, основанные на квантилях, они оказываются более устойчивыми к поведению распределений на «хвостах» распределений и более стабильны для распределений, отклоняющихся от нормального. Ошибки в данных приводят к значительным колебаниям среднего арифметического и дисперсии, в то время как оценки, основанные на квантилях, в меньшей степени подвержены такого рода колебаниям.

Оценка характеристик формы распределения

На рис. 3.4—3.6 приведены три вида унимодальных распределений с различной асимметрией. Асимметрия обычно измеряется *коэффициентом асимметрии* S_{k1} (skewness):

$$S_{k1} = \frac{\mu_3^*}{\sigma^3},$$

где μ_3^* — выборочный центральный момент третьего порядка.

Значительно проще вычисляется другой показатель асимметрии S_{k2} на основе квартилей распределения

$$S_{k2} = \frac{x_{0,75} + x_{0,25} - 2 \cdot x_{0,5}}{x_{0,75} - x_{0,25}} = \frac{Q_3 + Q_1 - 2 \cdot Q_2}{Q_3 - Q_1}.$$

«Острота» пика распределения определяется коэффициентом эксцесса (kurtosis).

Коэффициент эксцесса вычисляется по формуле

$$K = \frac{\mu_4^*}{\sigma^4} - 3,$$

где μ_4^* — выборочный центральный момент четвертого порядка.

Для нормального распределения коэффициенты асимметрии и эксцесса равны нулю.

3.2. Принципы статистического оценивания. Классификация оценок

Часто объектом статистического анализа являются результаты наблюдения случайной величины X, вид распределения которой известен. Например, если имеются данные о росте новобранцев в возрасте 20 лет, то опыт обработки такого рода данных дает основание утверждать, что наблюдаемая случайная величина X имеет нормальное распределение $N(m, \sigma^2)$. Задача состоит в нахождении приближенных значений-оценок параметров распределения m и σ^2 по выборке.

Аналогично, ставится задача об оценке параметров других распределений.

В общем случае, пусть $x_1, \ldots, x_n$ — выборка наблюдений случайной величины X с известным законом распределения $F(x, \theta)$, где параметр θ неизвестен. Оценка $\tilde{\theta}_n$ параметра θ по выборке объема n вычисляется как значение некоторой функции элементов выборки (такие функции называются **статистиками**)

$$\tilde{\theta}_n = \tilde{\theta}_n(x_1, \ldots, x_n).$$

Как правило, существуют несколько статистик, которые можно использовать для оценки одного и того же параметра. Например, в случае если распределение генеральной совокупности симметрично относительно математического ожидания m, для оценки параметра m можно взять выборочное среднее

$$\tilde{m}_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

или выборочную медиану h

$$\tilde{m}_2 = h.$$

Оценка параметра θ при помощи любой статистики дает приближенное значение параметра. Более того, значения статистик для разных выборок объема n из одной генеральной совокупности будут различны (обратитесь к таблице, содержащей результаты работы 2 в главе 2: моделирование распределений случайных величин). В качестве оценки следует взять такую статистику, значения которой для различных выборок из данной генеральной совокупности были бы «в среднем» близки к истинному значению параметра. Желательно также, чтобы с увеличением объема выборки точность оценки возрастала.

Для выбора статистики, которая определяет значение оценки $\tilde{\theta}_n$, имеющее наименьший разброс около истинного значения параметра θ, полученную выборку наблюдений $x_1, \ldots, x_n$ рассматривают, как одну из возможных реализаций выборочного вектора $(X_1, \ldots, X_n)$ (см. п. 3.1.2), а оценку — как функцию его элементов

$$\tilde{\theta}_n = \tilde{\theta}_n(X_1, \ldots, X_n).$$

Распределение случайной величины $\tilde{\theta}_n$ называют **выборочным распределением оценки** (или просто распределением оценки).

Лучшая оценка параметра может быть получена путем сравнения выборочных распределений нескольких подходящих статистик. На рис. 3.8 приведены эскизы графиков плотностей распределений для трех статистик $\tilde{\theta}_1$, $\tilde{\theta}_2$ и $\tilde{\theta}_3$, используемых для оценки одного параметра θ. Очевидно, статистика $\tilde{\theta}_1$ предпочтительнее статистик $\tilde{\theta}_2$ и $\tilde{\theta}_3$, так как плотность $f_{\tilde{\theta}_1}$ теснее сосредоточена около истинного значения θ, чем плотности $f_{\tilde{\theta}_2}$ и $f_{\tilde{\theta}_3}$, и, следовательно, значения $\tilde{\theta}_1$ в «среднем» будут ближе к θ, чем значения $\tilde{\theta}_2$ и $\tilde{\theta}_3$. Однако часто выборочное распределение той или иной статистики получить невозможно. В этом случае для выбора наилучшей оценки используют несколько свойств, которые характеризуют качество оценок параметров и которыми должна обладать «хорошая» оценка. Это следующие свойства.

1. **Состоятельность.** Оценка $\tilde{\theta}_n = \tilde{\theta}_n(X_1, \ldots, X_n)$ называется состоятельной оценкой параметра θ, если, при $n \to \infty$, $\tilde{\theta}_n$ сходится по вероятности к θ.

Состоятельность оценки $\tilde{\theta}_n$ во многих случаях может быть установлена с помощью следующей теоремы.

Теорема 1 (достаточное условие состоятельности). Если $M[\tilde{\theta}_n] \to \theta$ и $D[\tilde{\theta}_n] \to 0$ при $n \to \infty$, то $\tilde{\theta}_n$ — состоятельная оценка параметра θ.

2. **Несмещенность.** Оценка $\tilde{\theta}_n$ называется несмещенной оценкой параметра θ, если ее математическое ожидание равно оцениваемому параметру, т. е. $M[\tilde{\theta}_n] = \theta$.

Разность $M[\tilde{\theta}_n] - \theta$ называется **смещением**, или **систематической ошибкой**. На рис. 3.8 $f_{\tilde{\theta}_3}$ — плотность смещенной оценки $\tilde{\theta}_3$ параметра θ, смеще-

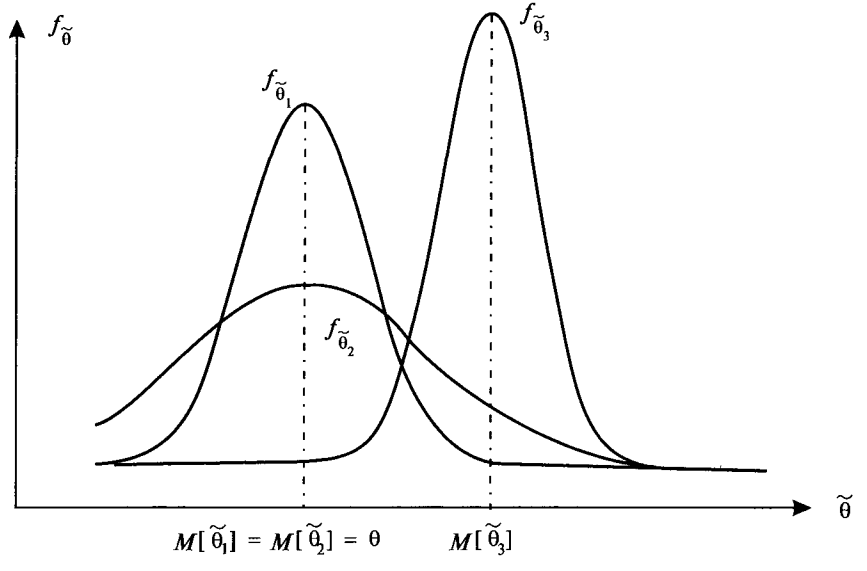


Рис. 3.8. Графики плотностей распределения статистик, используемых для оценки одного параметра θ

ние равно $M[\tilde{\theta}_3] - \theta$. Для несмещенных оценок систематическая ошибка оценивания равна нулю.

3. Пусть для оценки параметра θ можно использовать две несмещенные оценки $\tilde{\theta}_1$ и $\tilde{\theta}_2$. Если $D[\tilde{\theta}_1] < D[\tilde{\theta}_2]$, то говорят, что оценка $\tilde{\theta}_1$ **более эффективна**, чем оценка $\tilde{\theta}_2$.

На рис. 3.8 показаны плотности двух несмещенных оценок $\tilde{\theta}_1$ и $\tilde{\theta}_2$: $M[\tilde{\theta}_1] = M[\tilde{\theta}_2] = \theta$. Для этих оценок $D[\tilde{\theta}_1] < D[\tilde{\theta}_2]$ и, следовательно, оценка $\tilde{\theta}_1$ более эффективна, чем $\tilde{\theta}_2$.

В математической статистике разработаны различные методы нахождения оценок параметров распределений: метод моментов, метод максимального правдоподобия, метод наименьших квадратов (см. главу 6) и другие. Подробное изложение соответствующей теории можно найти в литературе, например [22, 18, 14].

3.2.1. Несмещенные и состоятельные оценки математического ожидания и дисперсии генеральной совокупности

Пусть $x_1, x_2, \ldots, x_n$ выборка из генеральной совокупности с конечным математическим ожиданием m и дисперсией σ^2. В качестве оценки математического ожидания возьмем выборочное среднее

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Чтобы проверить несмещенность и состоятельность выборочного среднего $\bar{x}$ как оценки m, рассмотрим статистику $\bar{x}$ как функцию выборочного вектора $(X_1, \ldots, X_n)$, т. е. $\bar{X} = \frac{1}{n}\sum X_i$. По определению имеем: $M[X_i] = m$ и $D[X_i] = \sigma^2$, $i = 1, 2, \ldots, n$, причем X_i — независимые в совокупности случайные величины. Следовательно,

$$M[\bar{X}] = M\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} M[X_i] = \frac{1}{n} nm = m,$$

$$D[\bar{X}] = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} D[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Отсюда, по определению, получаем, что $\bar{X}$ — несмещенная оценка m. Так как $D[\bar{X}] \to 0$ при $n \to \infty$, то, в силу **теоремы 1**, $\bar{X}$ является состоятельной оценкой математического ожидания m генеральной совокупности.

Рассмотрим оценку дисперсии.

Покажем, что статистика

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

будет несмещенной и состоятельной оценкой дисперсии генеральной совокупности.

Предварительно преобразуем сумму квадратов следующим образом:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}[(x_i - m) - (\bar{x} - m)]^2 =$$

$$= \sum_{i=1}^{n}(x_i - m)^2 - 2(\bar{x} - m)\sum_{i=1}^{n}(x_i - m) + \sum_{i=1}^{n}(\bar{x} - m)^2 =$$

$$= \sum_{i=1}^{n}(x_i - m)^2 - 2n(\bar{x} - m)^2 + n(\bar{x} - m)^2 = \sum_{i=1}^{n}(x_i - m)^2 - n(\bar{x} - m)^2.$$

Для проверки несмещенности, представим статистику s^2 как функцию выборочного вектора $(X_1, ..., X_n)$ и вычислим математическое ожидание S^2, используя результат предыдущего преобразования

$$M[S^2] = \frac{1}{n-1}M\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n-1}M\left[\sum_{i=1}^{n}(X_i - m)^2 - n(\bar{X} - m)^2\right] =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}M[(X_i - m)^2] - \frac{n}{n-1}M[(\bar{X} - m)^2] = \frac{n\cdot\sigma^2}{n-1} - \frac{n}{n-1}\cdot\frac{\sigma^2}{n} = \sigma^2.$$

Здесь мы воспользовались тем, что $M[(\bar{X} - m)^2] = D[\bar{X}] = \frac{\sigma^2}{n}$.

Оценка дисперсии, определяемая как выборочный центральный момент второго порядка

$$S_0^2 = \mu_2^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

будет смещенной оценкой, со смещением $-\frac{\sigma^2}{n}$, это следует из того, что обе оценки связаны простым соотношением $S_0^2 = \frac{n-1}{n}S^2$, следовательно, $M[S_0] = \frac{n-1}{n}M[S^2] = \frac{n-1}{n}\sigma^2$. Смещение оценки S_0^2 равно

$$M[S_0^2] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

Можно показать [13], что дисперсия оценки S^2 равна

$$D[S^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right),$$

т. е., если $\mu_4 < \infty$, то $D[S^2] \to 0$ при $n \to \infty$. В силу **теоремы 1** S^2 будет состоятельной оценкой дисперсии генеральной совокупности.

Если среднее генеральной совокупности n известно, то несмещенной оценкой дисперсии является статистика

$$S_0^2 = \frac{1}{n}\sum(X_i - m)^2.$$

Действительно,

$$M[S_0^2] = M\left[\frac{1}{n}\sum(X_i - m)^2\right] = \frac{1}{n}\sum M[(X_i - m)^2] = \frac{1}{n} n\sigma^2 = \sigma^2.$$

Свойства несмещенности и состоятельности не связаны друг с другом, т. е. ни одно из этих свойств не определяет другое. Например, оценка S_0^2 является смещенной оценкой дисперсии, однако эта оценка, как и оценка S^2, является состоятельной. Это следует из того, что обе оценки связаны простым соотношением

$$S_0^2 = \frac{n-1}{n} S^2.$$

Часто для оценки данного параметра может быть найдено несколько состоятельных и даже несмещенных оценок. Например, в случае выборки из генеральной совокупности имеющей нормальное распределение, наряду с выборочным средним, выборочная медиана является несмещенной и состоятельной оценкой среднего генеральной совокупности. Дисперсии этих оценок равны соответственно $\frac{\sigma^2}{n}$ и $\frac{\pi\sigma^2}{2n}$. Естественно, что при сравнении этих оценок следует отдать предпочтение выборочному среднему, как оценке, имеющей меньшую дисперсию. Такая оценка будет в среднем меньше отклоняться от истинного значения параметра, чем оценка с большей дисперсией, и будет более эффективной.

Однако требования несмещенности и малой дисперсии могут быть несовместимы.

Чтобы найти разумный компромисс между величинами смещения и дисперсии, можно минимизировать средний квадрат ошибки, т. е. в качестве оценки параметра θ выбрать $\tilde{\theta}$ из условия минимума математического ожидания случайной величины $(\tilde{\theta} - \theta)^2$.

3.2.2. Распределения основных статистик в случае нормально распределенной генеральной совокупности: распределения хи-квадрат, Стьюдента и Фишера

Основные результаты сложившейся к настоящему времени классической теории математической статистики получены для методов, использующих выборки из генеральной совокупности, имеющей нормальное распределение. Для понимания методов математической статистики достаточно знания определений и свойств распределения χ^2, Стьюдента и Фишера, которые приведены ниже.

Распределения основных статистик, вычисляемых по выборке из нормально распределенной генеральной совокупности, связаны с распределениями $\chi^2(k)$, Стьюдента $T(k)$ и Фишера $F(k_1, k_2)$. Приведем определения и некоторые свойства этих распределений. Всюду в дальнейшем квантили порядка p будут обозначаться следующим образом:

u_p — квантиль стандартного нормального распределения $N(0, 1)$;

$\chi_p^2(k)$ — квантиль распределения χ^2 с k степенями свободы;

$t_p(k)$ — квантиль распределения Стьюдента с k степенями свободы;

$F_p(k_1, k_2)$ — квантиль распределения Фишера с k_1 и k_2 степенями свободы.

Таблицы выше перечисленных квантилей приведены в [1].

В пакете **STATISTICA** квантили распределений определяются при помощи вероятностного калькулятора (см. главу 2).

Распределением χ^2 с k степенями свободы называется распределение случайной величины $\chi^2(k)$, равной сумме квадратов k независимых нормально распределенных по стандартному нормальному закону $N(0, 1)$ случайных величин U_i, $i = 1, 2, ..., k$, т. е. распределение случайной величины

$$\chi^2(k) = U_1^2 + \ldots + U_k^2.$$

Распределение χ^2 с k степенями свободы там, где это не вызывает недоразумений, будет обозначаться также $\chi^2(k)$.

Плотность распределения $f_{\chi^2}(x)$ определяется формулой

$$f_{\chi^2}(x) = \begin{cases} 0, & x \le 0, \\ \dfrac{1}{2^{k/2}\Gamma\left(\dfrac{k}{2}\right)} \cdot x^{\frac{k-2}{2}} \cdot e^{-\frac{x}{2}}, & x > 0, \end{cases}$$

где $\Gamma(\alpha)$ — гамма функция или интеграл Эйлера второго рода

$$\Gamma(\alpha + 1) = \int_0^\infty x^\alpha e^{-x} dx.$$

График функции $f_{\chi^2}(x)$ приведен на рис. 3.9.

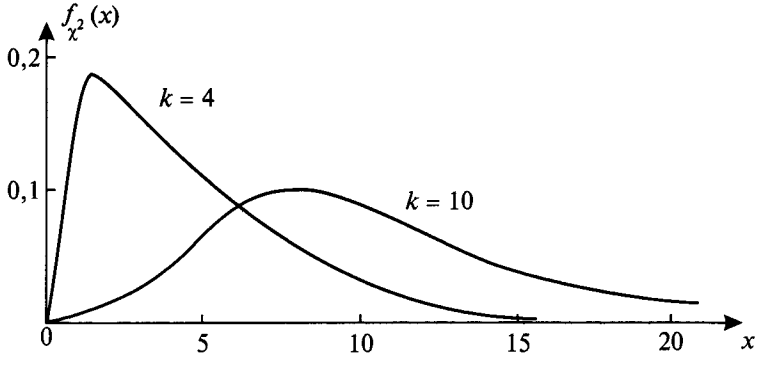


Рис. 3.9. Плотности распределения $\chi^2(k)$ при числе степеней свободы $k = 4$ и 10

Среднее и дисперсия распределения $\chi^2(k)$ равны соответственно

$$M[\chi^2(k)] = k, \quad D[\chi^2(k)] = 2k.$$

Если $\chi^2(k_1)$ и $\chi^2(k_2)$ — независимые случайные величины, имеющие распределение χ^2 с k_1 и k_2 степенями свободы соответственно, то сумма этих случайных величин имеет распределение χ^2 с $k_1 + k_2$ степенями свободы

$$\chi^2(k_1) + \chi^2(k_2) = \chi^2(k_1 + k_2).$$

Распределение $\chi^2(k)$ при больших значениях k ($k > 30$) с достаточной для практических расчетов точностью аппроксимируется нормальным распределением.

Распределением Стьюдента с k степенями свободы называется распределение случайной величины $T(k)$, равной отношению двух независимых случайных величин U и $\sqrt{\chi^2(k)/k}$, т. е.

$$T(k) = \frac{U}{\sqrt{\chi^2(k)/k}},$$

где U имеет стандартное нормальное распределение $N(0, 1)$. Распределение Стьюдента с k степенями свободы будет также обозначаться $T(k)$.

Распределение Стьюдента с k степенями свободы имеет плотность $f_T(x)$ (рис. 3.10):

$$f_T(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}}\left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \quad -\infty < x < +\infty,$$

среднее, $M[T(k)] = 0$ и дисперсию, $D[T(k)] = \frac{k}{k-2}$, $k > 2$. Плотность распределения Стьюдента симметрична относительно оси ординат, следовательно, для квантилей распределения Стьюдента $t_p(k)$ имеет место соотношение $t_p(k) = -t_{1-p}(k)$.

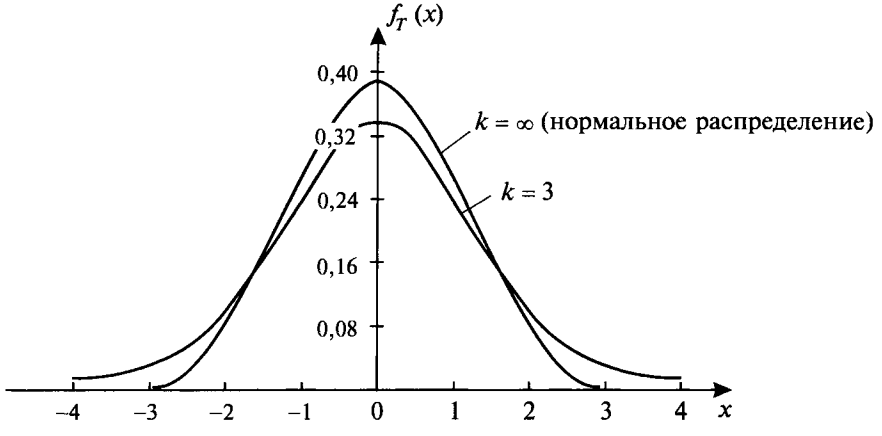


Рис. 3.10. Плотности распределения Стьюдента и нормального распределения

При больших k ($k > 30$) для квантилей $t_p(k)$ распределения Стьюдента выполнено приближенное равенство $t_p(k) \approx u_p$.

Распределением Фишера с k_1 и k_2 степенями свободы называется распределение случайной величины $F(k_1, k_2)$, равной отношению двух независимых случайных величин $\chi^2(k_1)/k_1$ и $\chi^2(k_2)/k_2$, т. е.

$$F(k_1, k_2) = \frac{\chi^2(k_1)/k_1}{\chi^2(k_2)/k_2}.$$

Распределение Фишера с k_1 и k_2 степенями свободы будет также обозначаться $F(k_1, k_2)$. Распределение Фишера с k_1 и k_2 степенями свободы имеет плотность $f_F(x)$ (рис. 3.11):

$$f_F(x) = \begin{cases} 0, & x \le 0, \\ \dfrac{\Gamma\left(\dfrac{k_1 + k_2}{2}\right)}{\Gamma\left(\dfrac{k_1}{2}\right)\Gamma\left(\dfrac{k_2}{2}\right)} \left(\dfrac{k_1}{k_2}\right)^{\frac{k_1}{2}} \dfrac{x^{\frac{k_1}{2}-1}}{\left(1 + \dfrac{k_1}{k_2} x\right)^{\frac{k_1+k_2}{2}}}, & x > 0, \end{cases}$$

среднее, $M[F] = \dfrac{k_2}{k_2 - 2}$, $k_2 > 2$; $D[F] = \dfrac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$, $k_2 > 4$.

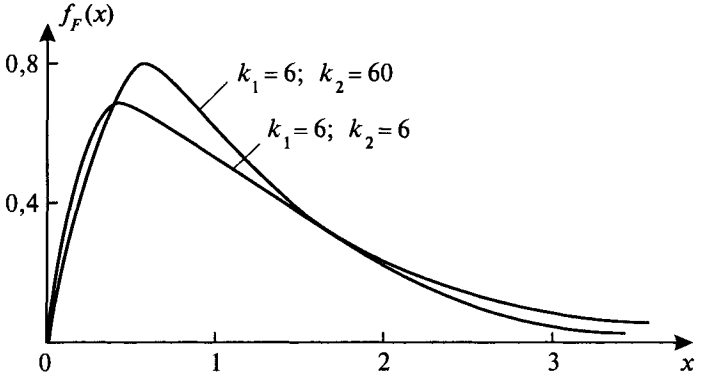


Рис. 3.11. Плотности распределения Фишера

Квантили распределения Фишера порядка p и $1 - p$ связаны следующей формулой

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

Между случайными величинами, имеющими нормальное распределение, распределение χ^2, Стьюдента и Фишера, имеют место соотношения:

$$T^2(k) = F(1, k),$$

$$F(k, \infty) = \frac{\chi^2(k)}{k},$$

$$\chi^2(1) = U^2,$$

$$t^2_{1-\frac{\alpha}{2}}(k) = F_{1-\alpha}(1, k).$$

3.2.3. Распределение выборочной дисперсии и некоторых нормированных статистик

Распределения основных статистик, вычисляемых на основе выборки объема n из нормально распределенной $N(m, \sigma^2)$ генеральной совокупности, можно получить из следующих фактов.

1. Выборочное среднее $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ имеет нормальное распределение

$$N\left(m, \frac{\sigma^2}{n}\right).$$

2. Выборочная дисперсия в случае, когда математическое ожидание m генеральной совокупности известно, $S_0^2 = \frac{1}{n}\sum (X_i - m)^2$ связана со случайной величиной, имеющей распределение χ^2 с n степенями свободы, соотношением

$$S_0^2 = \frac{\sigma^2}{n}\chi^2(n).$$

Этот результат следует из определения случайной величины $\chi^2(n)$ как суммы квадратов n независимых случайных величин U_i, $i = 1, 2, ..., n$, каждая из которых имеет стандартное нормальное распределение $N(0, 1)$.

3. Выборочная дисперсия $S^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$ связана со случайной величиной, имеющей распределение χ^2 с $(n-1)$ степенями свободы, соотношением

$$S^2 = \frac{\sigma^2}{n-1}\chi^2(n-1).$$

Этот результат является следствием следующей теоремы.

Теорема 2. Пусть $x_1, x_2, ..., x_n$ — выборка из нормально распределенной генеральной совокупности $N(m, \sigma^2)$, а $\bar{x} = \frac{1}{n}\sum x_i$ и $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$ — соответственно статистики, являющиеся оценками среднего и дисперсии генеральной совокупности. Тогда статистики $\bar{X}$ и S^2, рассматриваемые как функции выборочного вектора, являются независимыми случайными величинами, причем статистика $\frac{n-1}{\sigma^2}S^2$ имеет распределение $\chi^2(n-1)$.

4. Статистика $\frac{\bar{X} - m}{S/\sqrt{n}}$ является случайной величиной, имеющей распределение Стьюдента с $(n-1)$ степенью свободы, т. е.

$$\frac{\bar{X} - m}{S/\sqrt{n}} = T(n-1).$$

Этот результат легко получить, воспользовавшись следствием **Теоремы 2** и выполняя преобразования

$$\frac{\bar{X}-m}{S/\sqrt{n}}=\frac{\bar{X}-m}{\sqrt{S^2/n}}=\frac{\bar{X}-m}{\sqrt{\dfrac{\sigma^2\cdot\chi^2(n-1)}{n(n-1)}}}=\frac{\dfrac{\bar{X}-m}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{\chi^2(n-1)}{n-1}}}.$$

Так как в числителе стоит статистика $\dfrac{\bar{X}-m}{\sigma/\sqrt{n}}$, имеющая стандартное нормальное распределение $N(0;1)$, а в знаменателе независимая от нее случайная величина $\sqrt{\dfrac{\chi^2(n-1)}{n-1}}$, то, по определению распределения Стьюдента, отношение этих случайных величин имеет распределение Стьюдента с $(n-1)$ степенями свободы.

5. Пусть S_1^2 и S_2^2 — выборочные дисперсии, вычисленные по независимым выборкам объема n_1 и n_2 из двух нормально распределенных генеральных совокупностей соответственно с дисперсиями σ_1^2 и σ_2^2, тогда отношение $\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ имеет распределение Фишера с (n_1-1) и (n_2-1) степенями свободы, т. е.

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}=F(n_1-1;n_2-1).$$

Здесь мы воспользовались определением распределения Фишера.

3.2.4. Интервальные оценки. Доверительный интервал и доверительная вероятность

При статистической обработке результатов наблюдений часто необходимо найти не только оценку $\tilde{\theta}$ неизвестного параметра θ, но и охарактеризовать точность этой оценки. С этой целью вводится понятие доверительного интервала.

Доверительным интервалом для параметра θ называется интервал (θ_1, θ_2), накрывающий истинное значение θ с заданной вероятностью $p = 1-\alpha$, т. е.

$$P[\theta_1 < \theta < \theta_2] = 1-\alpha.$$

Число $1-\alpha$ называется **доверительной вероятностью**, а значение α — **уровнем значимости**. Статистики $\theta_1 = \theta_1(x_1, \ldots, x_n)$ и $\theta_2 = \theta_2(x_1, \ldots, x_n)$, определяемые по выборке $x_1, \ldots, x_n$ из генеральной совокупности с неизвестным параметром θ, называются соответственно **нижней** и **верхней границами доверительного интервала**.

Доверительному интервалу можно дать следующую статистическую интерпретацию: в большой серии независимых экспериментов, в каждом из которых получена выборка объема n, в среднем $(1-\alpha)\cdot 100\,\%$ из общего

числа построенных доверительных интервалов содержат истинное значение параметра θ.

Длина доверительного интервала, характеризующая точность интервального оценивания, зависит от объема выборки n и доверительной вероятности $1-\alpha$: при увеличении объема выборки длина доверительного интервала уменьшается, а с приближением доверительной вероятности к единице — увеличивается. Выбор доверительной вероятности определяется конкретными условиями. Обычно используются значения $1-\alpha$, равные 0,90; 0,95; 0,99.

При решении некоторых задач применяются односторонние интервалы, границы которых определяются из условий

$$P[\theta < \theta_2] = 1-\alpha \text{ или } P[\theta_1 < \theta] = 1-\alpha.$$

Эти интервалы называются соответственно **левосторонними** и **правосторонними доверительными интервалами**.

Чтобы найти доверительный интервал для параметра θ, необходимо знать закон распределения статистики $\tilde{\theta} = \tilde{\theta}(x_1, \ldots, x_n)$, значение которой является оценкой параметра θ. При этом для получения доверительного интервала наименьшей длины при данном объеме выборки n и заданной доверительной вероятности $1-\alpha$ в качестве оценки $\tilde{\theta}$ параметра θ следует брать наиболее эффективную оценку.

Один из методов построения доверительных интервалов состоит в следующем. Предположим, что существует статистика $Y = Y(\tilde{\theta}, \theta)$ такая, что:

а) закон распределения Y известен и не зависит от θ;

б) функция $Y(\tilde{\theta}, \theta)$ непрерывна и строго монотонна по θ.

Пусть, далее, $1-\alpha$ — заданная доверительная вероятность, а $y_{\alpha/2}$ и $y_{1-\frac{\alpha}{2}}$ — квантили распределения статистики Y порядков $\alpha/2$ и $1-\frac{\alpha}{2}$ соответственно. Тогда с вероятностью $1-\alpha$ выполняется неравенство

$$y_{\alpha/2} < Y(\tilde{\theta}, \theta) < y_{1-\frac{\alpha}{2}}.$$

Решая неравенство относительно θ, найдем границы θ_1 и θ_2 доверительного интервала для θ.

Рассмотрим доверительные интервалы для параметров нормально распределенной генеральной совокупности.

Доверительный интервал для среднего. Пусть $x_1, \ldots, x_n$ — выборка n наблюдений случайной величины X, которая имеет нормальное распределение $N(m, \sigma^2)$. В качестве оценки m возьмем выборочное среднее $\bar{x}$. Предположим, что дисперсия генеральной совокупности σ^2 известна. Рассмотрим статистику $Y = \dfrac{\bar{X} - m}{\sigma}$, имеющую стандартное нормальное распределение $N(0, 1)$ независимо от m и других параметров. Кроме того, Y как функция m, непрерывна и монотонна. Найдем квантили $u_{\frac{\alpha}{2}}$ и $u_{1-\frac{\alpha}{2}}$ стандартного нормального распределения $N(0, 1)$. Следовательно, вероятность события,

состоящего в том, что $u_{\frac{\alpha}{2}} < Y < u_{1-\frac{\alpha}{2}}$ будет равна заданной доверительной вероятности $1 - \alpha$, т. е.

$$P\left[u_{\frac{\alpha}{2}} < Y < u_{1-\frac{\alpha}{2}}\right] = 1 - \alpha.$$

Решая неравенство $u_{\frac{\alpha}{2}} < \dfrac{\overline{X} - m}{\sigma/\sqrt{n}} < u_{1-\frac{\alpha}{2}}$ относительно m, получим, что с вероятностью $1 - \alpha$ выполняется следующее условие

$$\overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} < m < \overline{X} - \frac{\sigma}{\sqrt{n}} u_{\frac{\alpha}{2}}.$$

Используя соотношение для квантилей стандартного нормального распределения: $u_{\frac{\alpha}{2}} = -u_{1-\frac{\alpha}{2}}$, найденный доверительный интервал для m можно записать так:

$$\overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} < m < \overline{X} + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}.$$

Если дисперсия генеральной совокупности σ^2 неизвестна, то по выборке наблюдений определяют оценку дисперсии

$$\tilde{\sigma}^2 = S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2.$$

Доверительный интервал для среднего в этом случае находят, используя статистику

$$Y = \frac{\overline{X} - m}{S/\sqrt{n}},$$

имеющую распределение Стьюдента с $n - 1$ степенью свободы. По заданной доверительной вероятности $1 - \alpha$ находим квантили $t_{\frac{\alpha}{2}}(n-1)$ и $t_{1-\frac{\alpha}{2}}(n-1)$ распределения Стьюдента с $(n - 1)$ степенью свободы.

Решая неравенство

$$t_{\frac{\alpha}{2}}(n-1) < \frac{\overline{X} - m}{S/\sqrt{n}} < t_{1-\frac{\alpha}{2}}(n-1)$$

относительно m и используя соотношение для квантилей распределения Стьюдента: $t_{\frac{\alpha}{2}}(n-1) = -t_{1-\frac{\alpha}{2}}(n-1)$, получим, что доверительный интервал для среднего при неизвестной дисперсии равен

$$\overline{X} - \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1) < m < \overline{X} + \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1).$$

Доверительный интервал для дисперсии. В зависимости от того, известно или неизвестно среднее генеральной совокупности, для нахождения доверительного интервала применяют различные формулы. Рассмотрим случай, когда среднее генеральной совокупности неизвестно. В этом случае в качестве оценки дисперсии используют выборочную дисперсию S^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Статистика $Y = \dfrac{(n-1)S^2}{\sigma^2}$ имеет распределение $\chi^2(n-1)$. Определим квантили $\chi^2_{\frac{\alpha}{2}}(n-1)$ и $\chi^2_{1-\frac{\alpha}{2}}(n-1)$, удовлетворяющие условию

$$P\left[\chi^2_{\frac{\alpha}{2}}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{1-\frac{\alpha}{2}}(n-1)\right] = 1-\alpha.$$

Решая неравенство

$$\chi^2_{\frac{\alpha}{2}}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{1-\frac{\alpha}{2}}(n-1)$$

относительно σ^2, получим доверительный интервал для дисперсии

$$\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}.$$

Доверительные интервалы для разности средних. При обработке результатов наблюдений часто необходимо оценить разность средних двух совокупностей. Такая задача возникает при сравнении средних двух совокупностей: если доверительный интервал для разности средних накрывает нуль, то следует считать, что различие средних двух совокупностей незначимо, и вызвано случайной изменчивостью величин и ошибками измерений. При этом вероятность того, что такое утверждение не верно, не превышает уровня значимости α. Пусть сравниваются средние двух генеральных совокупностей, имеющие нормальное распределение соответственно $N(m_1, \sigma_1^2)$ и $N(m_2, \sigma_2^2)$. По выборкам объема n_1 и n_2 из этих совокупностей найдем оценки средних: $\bar{X}_1$ и $\bar{X}_2$. Предположим, что дисперсии обеих совокупностей σ_1^2 и σ_2^2 известны, тогда дисперсия разности $\bar{X}_1 - \bar{X}_2$ равна $\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$, а статистика

$$Y = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

имеет стандартное нормальное распределение $N(0, 1)$. Рассуждая так же, как и при выводе доверительного интервала для среднего при известной

дисперсии, получим, что доверительный интервал для разности средних имеет вид

$$(\bar{X}_1 - \bar{X}_2) - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < m_1 - m_2 < (\bar{X}_1 - \bar{X}_2) + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Если дисперсии генеральных совокупностей неизвестны, но можно считать, что они равны, т. е. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, то доверительный интервал для разности средних находим так. Определим оценку дисперсии σ^2, используя результаты обеих выборок по формуле

$$\tilde{\sigma}^2 = S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

где S_1^2 — несмещенная оценка дисперсии, определяемая по выборке объема n_1, а S_2^2 — несмещенная оценка дисперсии определяемая по выборке объема n_2. Нетрудно показать, что S^2 — несмещенная оценка дисперсии σ^2. Действительно, так как $M[S_1^2] = M[S_2^2] = \sigma^2$, то $M[S^2] = \sigma^2$. Распределение этой оценки связано с распределением $\chi^2(n_1 + n_2 - 2)$ следующим образом

$$S^2 = \frac{(n_1 - 1)\dfrac{\sigma^2\chi^2(n_1 - 1)}{(n_1 - 1)} + (n_2 - 1)\dfrac{\sigma^2\chi^2(n_2 - 1)}{(n_2 - 1)}}{n_1 + n_2 - 2} = \frac{\sigma^2\chi^2(n_1 + n_2 - 2)}{n_1 + n_2 - 2}.$$

Используя статистику

$$Y = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}},$$

имеющую распределение Стьюдента с $(n_1 + n_2 - 2)$ степенями свободы, получим, что доверительный интервал для разности средних имеет вид

$$(\bar{X}_1 - \bar{X}_2) - t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < m_1 - m_2 < (\bar{X}_1 - \bar{X}_2) +$$

$$+ t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Пример 3.4. Можно ли считать, что средние двух нормально распределенных совокупностей равны, если выборочные средние и дисперсии, вычисленные по двум выборкам объема $n_1 = 16$ и $n_2 = 9$, равны соответственно $\bar{x}_1 = 12{,}57$; $s_1^2 = 0{,}91$; $\bar{x}_2 = 11{,}87$; $s_2^2 = 0{,}95$.

Предполагается, что дисперсии обеих совокупностей равны. Уровень значимости α принять равным 0,05.

Решение. Найдем доверительный интервал для разности средних для доверительной вероятности $1-\alpha = 0{,}95$. Предварительно вычислим оценку дисперсии σ^2 используя оценки для дисперсий по обеим выборкам

$$s^2 = \frac{15 \cdot 0{,}91 + 8 \cdot 0{,}95}{16 + 9 - 2} \approx 0{,}924.$$

Найдем квантиль $t_{0,975}(16 + 9 - 2) = 2{,}069$. По формуле получим доверительный интервал для разности средних

$$(12{,}57 - 11{,}87) - 2{,}069 \cdot 0{,}961\sqrt{\frac{1}{16} + \frac{1}{9}} < m_1 - m_2 < (12{,}57 - 11{,}87) +$$

$$+ 2{,}069 \cdot 0{,}961\sqrt{\frac{1}{16} + \frac{1}{9}}.$$

или

$$-0{,}123 < m_1 - m_2 < 1{,}533.$$

Так как доверительный интервал для разности средних покрывает нуль, то на 5 % уровне значимости следует считать, что средние генеральных совокупностей равны.

3.2.5. Оценка доли элементов совокупности, обладающих некоторым признаком

Такая задача часто встречается при обработке результатов переписей или социологических исследований. Например, нас может интересовать число безработных в данном городе или процент населения живущих ниже уровня бедности. Пусть генеральная совокупность состоит из N элементов, из которых D элементов обладают некоторым свойством или признаком. Отношение $\frac{D}{N} = p$ определяет долю элементов, обладающих данным признаком в генеральной совокупности. Несмещенной и состоятельной оценкой p по выборке объема n будет относительная частота $h = \frac{x}{n}$, где x — число элементов, обладающих данным свойством в выборке объема n. Можно показать, что дисперсия h: $D[h] = \frac{pq}{n-1} \cdot \frac{N-n}{N}$, где $q = 1 - p$, а $\frac{N-n}{N}$ — поправка на конечность генеральной совокупности.

Если доля отбора $\frac{n}{N}$ не превышает 5 %, то этой поправкой можно пренебречь.

Можно найти распределение случайной величины X — числа элементов, обладающих данным свойством среди элементов выборки объема n. Это гипергеометрическое распределение.

Вероятность того, что в выборке объема n будет ровно x элементов, обладающих данным свойством, вычисляется по формуле

$$P[X = x] = \frac{C_D^x C_{N-D}^{n-x}}{C_N^n}, \text{ где } N \geq n > 0,\ N - D > 0,\ x = 0,\ 1,\ \ldots,\ n.$$

В случае, если N и $N - D$ велики по сравнению с n, можно считать, что значение $p = \frac{D}{N}$ при последовательном отборе элементов выборки не изменяется и случайная величина X будет иметь биномиальное распределение

$$P[X = x] = C_n^x p^x (1 - p)^{n-x}, \quad x = 0,\ 1,\ \ldots,\ n.$$

В этом случае задача оценки доли сводится к оцениванию параметра p для биномиального распределения.

В силу следствия из центральной предельной теоремы (см. Приложение, П.4), статистика h имеет асимптотически (при $n \to \infty$) нормальное распределение

$$h \sim N\left(p, \frac{pq}{n}\right).$$

Этот результат используется для определения приблизительных границ доверительного интервала для p при большом числе испытаний n.

Пусть $1 - \alpha$ — заданная доверительная вероятность. Рассмотрим статистику $U = \frac{(h - p)}{\sqrt{\frac{pq}{n}}}$, которая имеет, при $n \to \infty$, приблизительно стандартное нормальное распределение $N(0, 1)$ независимо от значения p. При больших значениях n тогда имеем

$$P\left[\left|\frac{h - p}{\sqrt{pq/n}}\right| < u_{1-\alpha/2}\right] \approx 1 - \alpha,$$

где $u_{1-\alpha/2}$ — квантиль стандартного нормального распределения $N(0, 1)$. Отсюда получаем, что с вероятностью $1 - \alpha$ выполняется равенство

$$h - u_{1-\alpha/2}\sqrt{pq/n} < p < h + u_{1-\alpha/2}\sqrt{pq/n}.$$

Заменяя значения p и q в левой и правой частях неравенства их оценками $\tilde{p} = h$ и $\tilde{q} = 1 - h$, получаем, что приближенно границы доверительного интервала для вероятности «успеха» p имеют вид

$$h - u_{1-\alpha/2}\sqrt{h(1 - h)/n} < p < h + u_{1-\alpha/2}\sqrt{h(1 - h)/n}.$$

Пример 3.5. При проверке 100 деталей из большой партии обнаружено 10 бракованных деталей.

а. Найти 95%-й приближенный доверительный интервал для доли бракованных деталей во всей партии.

б. Какой минимальный объем выборки следует взять для того, чтобы с вероятностью 0,95 можно было утверждать, что доля бракованных деталей по всей партии отличается от частоты появления бракованных деталей в выборке не более чем на 1 %?

Решение.

а. Оценка доли бракованных деталей в партии по выборке равна

$$\tilde{p} = h = 10/100 = 0{,}1.$$

Используя вероятностный калькулятор, либо по таблице в Приложении 3 находим квантиль стандартного нормального распределения $u_{1-\frac{\alpha}{2}} = u_{0{,}975} = 1{,}96$. По приведенной выше формуле 95%-й доверительный интервал для p будет

$$0{,}041 < p < 0{,}159.$$

б. Представим доверительный интервал в виде неравенства

$$|h - p| < u_{1-\frac{\alpha}{2}}\sqrt{h(1-h)/n},$$

которое выполняется с вероятностью $\approx 1 - \alpha = 0{,}95$. Так как по условию задачи $|h - p| \le 0{,}01$, то для определения n получим неравенство

$$u_{0{,}975}\sqrt{h(1-h)/n} \le 0{,}01.$$

Отсюда следует, что

$$1{,}96 \cdot \sqrt{0{,}1(1-0{,}1)/n} \le 0{,}01 \text{ и } n \ge (0{,}3 \cdot 196)^2 = 3457{,}44.$$

Следовательно, минимальный объем выборки $n \approx 3458$.

3.3. Проверка статистических гипотез

3.3.1. Основные понятия

Во многих случаях результаты наблюдений используются для проверки предположений (гипотез) относительно тех или иных свойств распределения генеральной совокупности. В частности, такого рода задачи возникают при сравнении различных технологических процессов или методов обработки по определенным измеряемым признакам, например по точности, производительности и т. д.

Пусть X — наблюдаемая дискретная или непрерывная случайная величина. **Статистической гипотезой** H называется предположение относительно параметров или вида распределения случайной величины X. Статистическая гипотеза H называется **простой**, если она однозначно определяет распределение случайной величины X; в противном случае, гипотеза H называется **сложной**. Например, простой гипотезой является предположение о том, что случайная величина X распределена по нормальному закону $N(0, 1)$; если же высказывается предположение, что случайная величина имеет нормальное распределение $N(m, 1)$, где $a \le m \le b$, то это сложная гипотеза. Другим примером сложной гипотезы является предположение о том, что непрерывная случайная величина X с вероятностью 1/3 принимает значение из интервала (1, 5); в этом случае распределение случайной величины X может быть любым из класса непрерывных распределений.

Часто распределение случайной величины X известно и по выборке наблюдений необходимо проверить предположения о значении параметров этого распределения. Такие гипотезы называются **параметрическими**. В этой главе рассматривается проверка параметрических гипотез.

Проверяемая гипотеза называется **нулевой** гипотезой и обозначается H_0. Наряду с гипотезой H_0 рассматривают одну из **альтернативных (конкурирующих)** гипотез H_1. Например, если проверяется гипотеза H_0 о параметре θ H_0: $\theta = \theta_0$, где θ_0 — известное значение, в качестве альтернативной гипотезы можно рассмотреть одну из следующих гипотез

$$H_1^{(1)}: \theta > \theta_0;\ H_1^{(2)}: \theta < \theta_0;\ H_1^{(3)}: \theta \neq \theta_0;\ H_1^{(4)}: \theta = \theta_1,$$

где θ_1 — известное значение. Выбор альтернативной гипотезы определяется конкретной формулировкой задачи.

Правило, по которому принимается решение принять или отклонить гипотезу H_0, называется *критерием*. Так как решение принимается на основе выборки наблюдений случайной величины X, необходимо выбрать подходящую статистику Z, называемую в этом случае *статистикой критерия*. При проверке простой параметрической гипотезы H_0: $\theta = \theta_0$ в качестве статистики критерия обычно выбирают ту же статистику, что и для оценки параметра θ.

Проверка статистической гипотезы основывается на принципе, в соответствии с которым маловероятные события считаются невозможными, а события, имеющие большую вероятность, — достоверными. Этот принцип можно реализовать следующим образом. Перед анализом выборки назначается некоторая малая вероятность α, называемая **уровнем значимости**. Пусть V — множество значений статистики Z, а V_k — такое подмножество, что, при условии истинности гипотезы H_0, вероятность попадания статистики критерия в V_k равна α:

$$P[Z \in V_k / H_0] = \alpha.$$

Пусть $z_в$ — выборочное значение статистики Z, вычисленное по выборке наблюдений. Критерий формулируется следующим образом: отклонить гипотезу H_0, если $z_в \in V_k$; принять гипотезу H_0, если $z_в \in V \setminus V_k$. Критерий, основанный на использовании заранее заданного уровня значимости α, называют **критерием значимости**. Множество V_k всех значений статистики критерия Z, при которых принимается решение отклонить гипотезу H_0, называется **критической областью**; область $V \setminus V_k$ называется областью **принятия гипотезы** H_0.

Уровень значимости α определяет «размер» критической области V_k. Положение критической области на множестве значений статистики Z зависит от формулировки альтернативной гипотезы H_1. Например, если проверяется гипотеза $H_0: \theta = \theta_0$, а альтернативная гипотеза H_1 формулируется как $H_1: \theta > \theta_0$ ($\theta < \theta_0$), то критическая область размещается на правом (левом) «хвосте» плотности распределения статистики Z, т. е. имеет вид неравенства: $Z > z_{1-\alpha}$($Z < z_\alpha$), где $z_{1-\alpha}$ и z_α — квантили распределения статистики Z соответственно порядка $1 - \alpha$ и α вычисленные при условии, что верна гипотеза H_0. В этом случае критерий называется **односторонним**, соответственно право- и

левосторонним. Если альтернативная гипотеза формулируется как $H_1 : \theta \neq \theta_0$, критическая область размещается на обоих «хвостах» плотности распределения Z (рис. 3.12), т. е. определяется совокупностью неравенств:

$$Z < z_{\frac{\alpha}{2}} \text{ и } Z > z_{1-\frac{\alpha}{2}}.$$

В этом случае критерий называется **двусторонним**.

Таким образом, проверка статистической гипотезы при помощи критерия значимости может быть разбита на следующие этапы:

1. Сформулировать проверяемую H_0 и альтернативную H_1 гипотезы.
2. Назначить уровень значимости α.
3. Выбрать статистику Z критерия для проверки гипотезы H_0.
4. Определить выборочное распределение статистики Z критерия при условии, что верна гипотеза H_0.
5. Определить критическую область V_k в зависимости от формулировки альтернативной гипотезы одним из неравенств: $Z > z_{1-\alpha}$, $Z < z_{\alpha}$, или совокупностью неравенств $Z < z_{\frac{\alpha}{2}}$ и $Z > z_{1-\frac{\alpha}{2}}$.
6. Получить выборку наблюдений и вычислить выборочное значение статистики $z_в$ критерия.
7. Принять статистическое решение:
   - если $z_в \in V_k$, отклонить гипотезу H_0 как не согласующуюся с результатами наблюдений;
   - если, $z_в \in V \setminus V_k$ принять гипотезу H_0, т. е. считать, что гипотеза H_0 не противоречит результатам наблюдений.

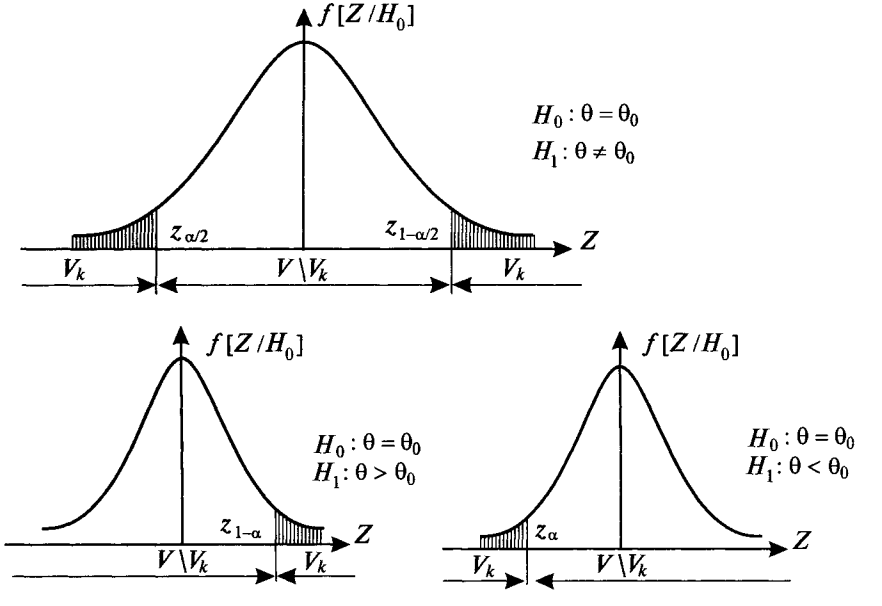


Рис. 3.12. Размещение критической области при различных альтернативных гипотезах

Замечания.

1. Обычно на этапах 4—7 используют статистику, квантили которой табулированы: статистику с нормальным распределением $N(0, 1)$, статистику Стьюдента, χ^2 или Фишера. Однако интерпретацию решения и вычисление вероятностей ошибок, допускаемых при проверке гипотез, удобно проводить для статистики, являющейся непосредственно оценкой параметра θ, т. е. статистики $\tilde{\theta}$.

2. В табл. 3.3 и 3.4 приводятся критерии значимости для проверки гипотез о дисперсиях и средних нормально распределенных генеральных совокупностей.

Таблица 3.3. **Критерии значимости для проверки гипотез о дисперсиях нормально распределенных генеральных совокупностей**

Проверяемая гипотеза H_0	Предположение относительно m	Статистика Z критерия	Распределение Z: $f(z/H_0)$	Область принятия гипотезы H_0 для двустороннего критерия	Альтернативная гипотеза и область принятия гипотезы H_0 для правостороннего критерия
$\sigma^2 = \sigma_0^2$	m известно	$\dfrac{nS_0^2}{\sigma_0^2}$	$\chi^2(n)$	$\chi^2_{\alpha/2}(n) < \dfrac{ns_0^2}{\sigma_0^2} < \chi^2_{1-\frac{\alpha}{2}}(n)$	H_1: $\sigma^2 > \sigma_0^2$ $\dfrac{ns_0^2}{\sigma_0^2} < \chi^2_{1-\alpha}(n)$
	m неизвестно, $\tilde{m} = \bar{x}$	$\dfrac{(n-1)S^2}{\sigma_0^2}$	$\chi^2(n-1)$	$\chi^2_{\frac{\alpha}{2}}(n-1) < \dfrac{(n-1)s^2}{\sigma_0^2} < \chi^2_{1-\frac{\alpha}{2}}(n-1)$	H_1: $\sigma^2 > \sigma_0^2$ $\dfrac{(n-1)s^2}{\sigma_0^2} < \chi^2_{1-\alpha}(n-1)$
$\sigma_1^2 = \sigma_2^2$	m_1 и m_2 известны	$\dfrac{S_{01}^2}{S_{02}^2}$; $s_{01}^2 > s_{02}^2$	$F(n_1, n_2)$	$\dfrac{s_{01}^2}{s_{02}^2} < F_{1-\frac{\alpha}{2}}(n_1, n_2)$	H_1: $\sigma_1^2 > \sigma_2^2$ $\dfrac{s_{01}^2}{s_{02}^2} < F_{1-\alpha}(n_1, n_2)$
	m_1 и m_2 неизвестны, $\tilde{m}_1 = \bar{x}_1$, $\tilde{m}_2 = \bar{x}_2$	$\dfrac{S_1^2}{S_2^2}$; $s_1^2 > s_2^2$	$F(n_1 - 1, n_2 - 1)$	$\dfrac{s_1^2}{s_2^2} < F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$	H_1: $\sigma_1^2 > \sigma_2^2$ $\dfrac{s_1^2}{s_2^2} < F_{1-\alpha}(n_1 - 1, n_2 - 1)$

Критерий Бартлетта для сравнения дисперсий нескольких совокупностей

H_0	Предположение относительно m_i	Статистика Z критерия	Распределение Z $f(z/H_0)$	Область принятия гипотезы H_0
$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_l^2$	m_i неизвестны, $i = 1, 2, \ldots, l$	$\dfrac{1}{c}\left[\sum_{i=1}^{l}(n_i - 1)\ln S^2 - \sum_{i=1}^{l}(n_i - 1)\ln S_i^2\right]$, где $c = 1 + \dfrac{1}{3(l-1)}\left[\sum_{i=1}^{l}\dfrac{1}{n_i - 1} - \dfrac{1}{\sum_{i=1}^{l}(n_i - 1)}\right]$, $S^2 = \dfrac{\sum(n_i - 1)S_i^2}{\sum(n_i - 1)}$	$\chi^2_{1-\alpha}(l-1)$	$z_в < \chi^2_{1-\alpha}(l-1)$

Таблица 3.4. **Критерии значимости для проверки гипотез о средних нормально распределенных генеральных совокупностей**

Проверяемая гипотеза, H_0	Предположение относительно, σ^2	Статистика Z критерия	Распределение Z: $f(z/H_0)$	Область принятия гипотезы H_0 для двустороннего критерия	Альтернативная гипотеза и область принятия гипотезы H_0 для правостороннего критерия
$m = m_0$	σ^2 известна	$\dfrac{\bar{X} - m_0}{\sigma/\sqrt{n}}$	$N(0, 1)$	$\dfrac{\lvert\bar{x} - m_0\rvert}{\sigma/\sqrt{n}} < u_{1-\frac{\alpha}{2}}$	H_1: $m > m_0$; $\dfrac{\bar{x} - m_0}{\sigma/\sqrt{n}} < u_{1-\alpha}$
	σ^2 неизвестна	$\dfrac{\bar{X} - m_0}{S/\sqrt{n}}$	$T(n-1)$	$\dfrac{\lvert\bar{x} - m_0\rvert}{s/\sqrt{n}} < t_{1-\frac{\alpha}{2}}(n-1)$	H_1: $m > m_0$; $\dfrac{\bar{x} - m_0}{s/\sqrt{n}} < t_{1-\alpha}(n-1)$
$m_1 = m_2$	σ_1^2 и σ_2^2 известны	$\dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$N(0, 1)$	$\dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} < u_{1-\frac{\alpha}{2}}$	H_1: $m_1 > m_2$; $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} < u_{1-\alpha}$
	σ_1^2 и σ_2^2 неизвестны, причем гипотеза H_0: $\sigma_1^2 = \sigma_2^2$ принимается, $\tilde{\sigma}_1^2 = s_1^2$, $\tilde{\sigma}_2^2 = s_2^2$	$\dfrac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$, где $S^2 = \left(\dfrac{(n_1-1)S_1^2}{n_1+n_2-2} + \dfrac{(n_2-1)S_2^2}{n_1+n_2-2}\right)$	$T(n_1+n_2-2)$	$\dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} < t_{1-\frac{\alpha}{2}}(n_1+n_2-2)$	H_1: $m_1 > m_2$; $\dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} < t_{1-\alpha}(n_1+n_2-2)$
	σ_1^2 и σ_2^2 неизвестны, причем гипотеза H_0: $\sigma_1^2 = \sigma_2^2$ отклоняется, $\tilde{\sigma}_1^2 = s_1^2$, $\tilde{\sigma}_2^2 = s_2^2$	$\dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$	$T(k)$, где $k \approx \dfrac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1+1} + \dfrac{\left(s_2^2/n_2\right)^2}{n_2+1}} - 2$	$\dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} < t_{1-\frac{\alpha}{2}}(k)$	H_1: $m_1 > m_2$; $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} < t_{1-\alpha}(k)$

3. В статистических пакетах обычно не используется значения задаваемого уровня значимости α. Как правило в выходных данных содержатся выборочные значения $z_в$ статистики критерия Z и вероятность того, что случайная величина Z (при условии, что верна гипотеза H_0) превышает выборочное значение $z_в$, т. е. значение

$$p = P[Z > |z_в|\big|_{H_0}] .$$

Эта вероятность называется p-значением (иногда она обозначается как p-level).

При двусторонней проверке p-значение равно

$$p = 2P[Z > |z_в||_{H_0}] .$$

Если $p > \alpha$, где α заданный уровень значимости, гипотеза H_0 принимается на уровне значимости p. Если $p < \alpha$ — гипотеза H_0 отклоняется, так как $z_в$ попадает в критическую область, причем вероятность ошибки первого рода равна p (см. ниже).

Пример 3.6. По паспортным данным автомобильного двигателя расход топлива на 100 км пробега составляет 10 л. В результате изменения конструкции двигателя ожидается, что расход топлива уменьшится. Для проверки проводятся испытания 25 случайно отобранных автомобилей с модернизированным двигателем, причем выборочное среднее расходов топлива на 100 км пробега по результатам испытаний составило $\bar{x} = 9{,}3$ л. Предположим, что выборка расходов топлива получена из нормально распределенной генеральной совокупности со средним m и дисперсией $\sigma^2 = 4$ л^2. Используя критерий значимости, проверить гипотезу, утверждающую, что изменение конструкции двигателя не повлияло на расход топлива.

Решение. Проверяется гипотеза о среднем m нормально распределенной генеральной совокупности.

Проверку гипотезы проведем по этапам:

1. Проверяемая гипотеза $H_0 : m = 10$, альтернативная гипотеза $H_1 : m < 10$.

2. Выберем уровень значимости $\alpha = 0{,}05$.

3. В качестве статистики критерия используем оценку математического ожидания — выборочное среднее $\bar{X}$.

4. Так как выборка получена из нормально распределенной генеральной совокупности, выборочное среднее $\bar{X}$ также имеет нормальное распределение с дисперсией $\sigma^2/n = 4/25$. При условии, что верна гипотеза H_0, математическое ожидание этого распределения равно 10. Нормированная статистика $U = \dfrac{\bar{X} - 10}{\sqrt{4/25}}$ имеет стандартное нормальное распределение $N(0, 1)$.

5. Альтернативная гипотеза $H_1 : m < 10$ предполагает уменьшение расхода топлива, следовательно, нужно использовать односторонний критерий. Критическая область определяется неравенством $U < u_\alpha$. Квантиль стандартного нормального распределения $N(0, 1)$ порядка 0,05 равна

$$u_{0,05} = -u_{0,95} = -1{,}645.$$

6. Выборочное значение нормированной статистики критерия равно

$$u_в = \frac{9{,}3 - 10}{\sqrt{4/25}} = -1{,}75.$$

7. Статистическое решение: так как выборочное значение статистики критерия принадлежит критической области, гипотеза H_0 отклоняется: на уровне значимости $\alpha = 0{,}05$ следует считать, что изменение конструкции двигателя привело к уменьшению расхода топлива.

Граница критической области для исходной статистики $\bar{x}_k$ может быть найдена из соотношения

$$\frac{\bar{x}_k - 10}{\sqrt{4/25}} = -1{,}645,$$

откуда получаем $\bar{x}_k = 9{,}342$, т. е. критическая область для статистики $\bar{X}$ определяется неравенством: $\bar{X} < 9{,}342$.

3.3.2. Ошибки первого и второго рода. Мощность критерия

Если для проверки статистической гипотезы H_0 применяется критерий значимости, то, в соответствии с принципом проверки гипотез, гипотеза H_0 отклоняется при попадании статистики критерия в критическую область. Если тем не менее гипотеза H_0 верна, то принимаемое решение неверно. Ошибка, совершаемая при отклонении правильной гипотезы H_0, называется **ошибкой первого рода**. Очевидно, вероятность ошибки первого рода равна вероятности попадания статистики критерия в критическую область V_k при условии, что верна гипотеза H_0, т. е. равна уровню значимости α:

$$P[Z \in V_k / H_0] = \alpha. \qquad (1)$$

Ошибка второго рода происходит в том случае, если гипотеза H_0 принимается, но в действительности верна альтернативная гипотеза H_1. Вероятность ошибки второго рода β можно вычислить (при простой альтернативной гипотезе H_1) по формуле

$$\beta = P[Z \in V \setminus V_k / H_1]. \qquad (2)$$

Пример 3.7. В условиях примера 3.6 предположим, что наряду с гипотезой $H_0 : m = 10$ л рассматривается простая альтернативная гипотеза $H_1 : m = 9$ л. В качестве статистики критерия снова возьмем выборочное среднее $\bar{X}$, а критическую область зададим неравенством $\bar{X} < 9{,}44$. Найти вероятности ошибок первого и второго рода для критерия с такой критической областью.

Решение. Найдем вероятность ошибки первого рода. Статистика критерия $\bar{X}$, при условии, что верна гипотеза H_0, имеет нормальное распределение $N(10; 4/25)$. По формуле (1), находим

$$\alpha = P[\bar{X} < 9{,}44 / H_0 : m = 10] = \Phi\left[\frac{9{,}44 - 10}{\sqrt{4/25}}\right] = \Phi(-1{,}4) \approx 0{,}08.$$

Это означает, что принятый критерий классифицирует приблизительно 8 % автомобилей, имеющих расход 10 л бензина на 100 км пробега, как автомобили, имеющие меньший расход топлива.

При условии, что верна гипотеза $H_1 : m_1 = 9$, статистика критерия $\bar{X}$ имеет нормальное распределение $N(9; 4/25)$. Вероятность ошибки второго рода по формуле (2) равна:

$$\beta = P[\bar{X} \geq 9{,}44 / H_1 : m = 9] = 1 - \Phi\left[\frac{9{,}44 - 9}{\sqrt{4/25}}\right] = 1 - \Phi(1{,}1) \approx 0{,}136.$$

Следовательно, в соответствии с принятым критерием, 13,6 % автомобилей, имеющих расход топлива 9 л на 100 км пробега, классифицируются как автомобили, имеющие расход 10 л.

Пусть проверяется гипотеза $H_0 : \theta = \theta_0$, а V_k — критическая область размера α. **Функцией мощности критерия** $\mu(V_k, \theta)$ называется вероятность отклонения гипотезы H_0 в функции параметра θ, т. е.

$$\mu(V_k, \theta) = P[Z \in V_k / \theta].$$

Вероятность отклонения гипотезы H_0 при конкретном значении параметра θ называется **мощностью критерия**. Очевидно, что $\mu(V_k, \theta_0) = \alpha$. Если альтернативная гипотеза H_1 простая, причем $H_1 : \theta = \theta_1$, то мощность критерия равна $1 - \beta$, т. е.

$$\mu(V_k, \theta_1) = 1 - \beta.$$

Обычно строят график функции мощности, вычисляя мощность критерия при нескольких значениях параметра θ.

3.3.3. Определение объема выборки при заданных вероятностях ошибок первого и второго рода

При заданной вероятности ошибки первого рода α вероятность ошибки второго рода может быть уменьшена за счет увеличения объема выборки. Если при этом вероятность ошибки второго рода не превышает заданное значение β, минимальный объем выборки n можно найти из решения системы

$$\begin{cases} P[Z \in V_k / H_0] = \alpha; \\ P[Z \in V \setminus V_k / H_1] \le \beta. \end{cases}$$

Аналитическое решение системы возможно только в простейших случаях.

Пример 3.8. Какой минимальный объем выборки n следует взять в условиях примера 3.6, чтобы при проверке гипотезы $H_0 : m = 10$ л против альтернативной гипотезы $H_1 : m = 9$ л ошибка первого рода была бы равна $\alpha = 0{,}01$, а ошибка второго рода не превышала 0,1? Какова критическая область в этом случае?

Решение. Так как альтернативная гипотеза H_1 предполагает меньшее значение параметра m, критическая область определяется неравенством $V_k : \bar{X} < \bar{x}_k$. По условию задачи имеем

$$\begin{cases} P[\bar{X} < \bar{x}_k / H_0 : m = 10] = \Phi\left[\dfrac{\bar{x}_k - 10}{\sqrt{4/n}}\right] = 0{,}01; \\ P[\bar{X} \ge \bar{x}_k / H_1 : m = 9] = 1 - \Phi\left[\dfrac{\bar{x}_k - 9}{\sqrt{4/n}}\right] \le 0{,}1. \end{cases}$$

Эту систему можно записать так:

$$\begin{cases} \dfrac{\bar{x}_k - 10}{2}\sqrt{n} = u_{0,01} = -2{,}326; \\ \dfrac{\bar{x}_k - 9}{2}\sqrt{n} \geq u_{0,9} = 1{,}282. \end{cases}$$

Исключая $\bar{x}_k$, получим, что $n \geq 53$. Подставляя наименьшее значение n в первое уравнение системы, найдем границу критической области

$$\bar{x}_k = 10 - \frac{2 \cdot 2{,}326}{\sqrt{53}} \approx 9{,}361.$$

Следовательно, критическая область V_k определяется неравенством $\bar{X} < 9{,}361$.

Пример 3.9. При измерении производительности двух агрегатов получены следующие результаты (в кг вещества за час работы):

Агрегат	Номер измерения				
	1	2	3	4	5
Агрегат А	14,1	10,1	14,7	13,7	14,0
Агрегат Б	14,0	14,5	13,7	12,7	14,1

Можно ли считать, что производительности агрегатов А и Б одинаковы, в предположении, что обе выборки получены из нормально распределенных генеральных совокупностей? Принять $\alpha = 0{,}1$.

Решение. Проверяется гипотеза $H_0 : m_1 = m_2$ при альтернативной гипотезе $H_1 : m_1 \neq m_2$. Вычислим оценки средних и дисперсий

$$\bar{x}_1 = 13{,}32,\ \bar{x}_2 = 13{,}80;\ s_1^2 = 3{,}37,\ s_2^2 = 0{,}46.$$

Предварительно проверим гипотезу о равенстве дисперсий $H_0 : \sigma_1^2 = \sigma_2^2$ (табл. 3.3):

$$\frac{s_1^2}{s_2^2} \approx \frac{3{,}37}{0{,}46} \approx 7{,}33.$$

Так как $F_{1-\alpha/2}(n_1 - 1, n_2 - 1) = F_{0,95}(4, 4) = 6{,}39$, то гипотеза о равенстве дисперсий отклоняется. Для проверки гипотезы о равенстве средних используем статистику Уэлча из нижней строки табл. 3.4. Вычислим выборочное значение статистики Уэлча:

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{|13{,}32 - 13{,}80|}{\sqrt{3{,}37/5 + 0{,}46/5}} \approx 0{,}55.$$

Число степеней свободы $k \approx \dfrac{(3{,}37/5 + 0{,}46/5)^2}{\dfrac{(3{,}37/5)^2}{6} + \dfrac{(0{,}46/5)^2}{6}} - 2 \approx 6$. Так как квантиль распределения Стьюдента $t_{0,95}(6) = 1{,}943$, то гипотеза о равенстве средних принимается.

Пример 3.10. В таблице приведены результаты измерений производительности 6 агрегатов и оценки дисперсий s_i^2, $i = 1, 2, ..., 6$, этих измерений. Используя эти данные, проверить гипотезу о равенстве дисперсий σ_i^2. Принять $\alpha = 0{,}1$.

№ измерения	Агрегаты					
	1	2	3	4	5	6
1	14,0	14,1	14,0	14,5	12,5	14,0
2	14,5	10,1	12,3	14,2	12,3	14,0
3	14,0	14,7	12,8	15,0	11,5	13,5
4	12,7	13,7	11,0	14,7	12,9	14,7
5	14,1	14,0	13,1	13,5	12,8	13,6
s_i^2	0,46	3,37	1,22	0,33	0,31	0,22

Решение. Для проверки гипотезы $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_6^2$ воспользуемся критерием Бартлетта (табл. 3.3). Предварительно вычислим

$$S^2 = \frac{4 \cdot (0{,}46 + 3{,}37 + 1{,}22 + 0{,}33 + 0{,}31 + 0{,}22)}{4 \cdot 6} \approx 0{,}98,$$

$$c \approx 1 + \frac{1}{3 \cdot (6-1)} \cdot \left[6 \cdot \frac{1}{4} - \frac{1}{6 \cdot 4} \right] \approx 1{,}097.$$

Выборочное значение $z_в$ статистики критерия равно

$$z_в \approx \frac{1}{1{,}097} \cdot [6 \cdot 4 \cdot \ln 0{,}98 - 4 \cdot (\ln 0{,}46 + \ln 3{,}37 + \ln 1{,}22 + \ln 0{,}33 +$$

$$+ \ln 0{,}31 + \ln 0{,}22)] \approx 11{,}07.$$

Так как $\chi^2_{0{,}90}(5) = 9{,}24$, а $z_в \approx 11{,}07$, гипотеза о равенстве дисперсий отклоняется.

Проверка статистических гипотез с использованием критериев значимости может быть проведена на основе соответствующих доверительных интервалов. При этом одностороннему критерию значимости соответствует односторонний доверительный интервал, а двустороннему критерию значимости — двусторонний доверительный интервал. Гипотеза $H_0 : \theta = \theta_0$ принимается, если значение θ_0 накрывается соответствующим доверительным интервалом; в противном случае — гипотеза H_0 отклоняется.

Если проверяется гипотеза $H_0 : \theta_1 = \theta_2$, то рассматривается доверительный интервал для разности $\theta_1 - \theta_2$. Гипотеза H_0 принимается, если доверительный интервал для разности параметров $\theta_1 - \theta_2$ накрывает нулевое значение. Исключение составляет проверка гипотезы о равенстве дисперсий $H_0 : \sigma_1^2 = \sigma_2^2$, так как доверительный интервал строится для отношения дисперсий, гипотеза H_0 в этом случае принимается, если доверительный интервал накрывает значение, равное единице.

3.3.4. Проверка гипотез о виде распределения по критерию χ^2

Пусть x_1, x_2, ..., x_n — выборка наблюдений случайной величины X. Проверяется гипотеза H_0, утверждающая, что X имеет функцию распределения $F(x)$.

Проверка гипотезы H_0 при помощи критерия χ^2 осуществляется по следующей схеме. По выборке наблюдений находят оценки неизвестных параметров предполагаемого закона распределения случайной величины X. Далее, область возможных значений случайной величины X разбивается на r множеств Δ_1, Δ_2, Δ_r, например, r интервалов в случае, когда X — непрерывная случайная величина, или r групп, состоящих из отдельных значений, для дискретной случайной величины X.

Пусть n_k — число элементов выборки, принадлежащих множеству Δ_k, $k = 1, 2, ..., r$. Очевидно, что $\sum_{k=1}^{r} n_k = n$. Используя предполагаемый закон распределения случайной величины X, находят вероятности p_k того, что значение X принадлежит множеству Δ_k, т. е. $p_k = P[X \in \Delta_k]$, $k = 1, 2, ..., r$. Очевидно, что $\sum_{k=1}^{r} p_k = 1$. Полученные результаты можно представить в виде следующей таблицы:

	Число наблюдений				Всего
	Δ_1	Δ_2	...	Δ_r	
Наблюдаемое	n_1	n_2	...	n_r	n
Ожидаемое	np_1	np_2	...	np_r	n

Выборочное значение статистики критерия вычисляется по формуле

$$\chi^2_{\text{в}} = \sum_{k=1}^{r} \frac{(n_k - np_k)^2}{np_k}.$$

Гипотеза H_0 согласуется с результатами наблюдений на уровне значимости α, если

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

где $\chi^2_{1-\alpha}(r - l - 1)$ — квантиль порядка $1 - \alpha$ распределения χ^2 с $(r - l - 1)$ степенями свободы, а l — число неизвестных параметров распределения, оцениваемых по выборке; если же $\chi^2_{\text{в}} \geq \chi^2_{1-\alpha}(r - l - 1)$, то гипотеза H_0 отклоняется.

Замечание. Критерий χ^2 использует тот факт, что случайные величины $\frac{n_k - np_k}{\sqrt{np_k}}$, $k = 1, 2, ..., r$, имеют распределения, близкие к нормальному $N(0, 1)$. Чтобы это утверждение было достаточно точным, необходимо, чтобы для всех интервалов выполнялось условие $np_k \geq 5$. Если для некоторых интервалов это условие не выполняется, то их следует объединить с соседними.

Пример 3.11. Проверка гипотезы о распределении по закону Пуассона. В первых двух столбцах табл. 3.5 приведены данные об отказах аппаратуры за 10 000 ч работы. Общее число обследованных экземпляров аппаратуры $n = 757$, при этом наблюдался $0 \cdot 427 + 1 \cdot 235 + 2 \cdot 72 + 3 \cdot 21 + 4 \cdot 1 + 5 \cdot 1 = 451$ отказ. Проверить гипотезу о том, что число отказов X имеет распределение Пуассона:

$$p_k = P[X = k] = \frac{\lambda^k}{k!} e^{-\lambda},\ k = 1, \ldots, \text{ при } \alpha = 0{,}01.$$

Таблица 3.5

Число отказов, k	Количество случаев, в которых наблюдалось k отказов, n_k	$p_k = \frac{0{,}6^k}{k!} e^{-0{,}6}$	Ожидаемое число случаев с k отказами, np_k
0	427	0,54881	416
1	235	0,32929	249
2	72	0,09879	75
3	21	0,01976	15
4	1	0,00296	2
5	1	0,00036	0
≥6	0	0,00004	0
Сумма	757	—	—

Решение. Оценка параметра λ равна среднему числу отказов: $\tilde{\lambda} = 451/757 \approx 0{,}6$. По таблице распределения Пуассона с $\lambda = 0{,}6$ находим вероятности p_k и ожидаемое число случаев с k отказами (третий и четвертый столбцы табл. 3.5).

Для $k = 4$, 5 и 6 значения $np_k < 5$, поэтому объединяем эти строки со строкой для $k = 3$. В результате получим значения, приведенные в табл. 3.6.

Таблица 3.6

k	n_k	np_k	$\frac{(n_k - np_k)^2}{np_k}$
0	427	416	0,291
1	235	249	0,787
2	72	75	0,120
≥3	23	17	2,118
—	—	—	$\chi^2_в = 3{,}316$

Так как по выборке оценивался один параметр λ, то $l = 1$, число степеней свободы равно $4 - 1 - 1 = 2$. По таблице квантилей распределения χ^2 (см. например [1] или воспользуйтесь вероятностным калькулятором) находим $\chi^2_{0,99}(2) = 9{,}21$, следовательно, гипотеза о распределении числа отказов по закону Пуассона принимается.

Пример 3.12. Проверить гипотезу о нормальном распределении по выборке 55 наблюдений:

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

Принять $\alpha = 0{,}1$.

Решение. Объем выборки $n = 55$. Для проверки гипотезы о нормальном распределении нужно найти оценки математического ожидания и дисперсии. Имеем

$$\tilde{m} = \bar{x} = \frac{1}{n}\sum_{i=1}^{55} x_i \approx 17{,}87,$$

$$\tilde{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{55}(x_i - \bar{x})^2 \approx 8{,}62.$$

Воспользуемся результатами группировки выборки в примере 3.2 (см. табл. 3.2), расширив первый и последний интервалы. Результаты группировки приведены во втором и третьем столбцах табл. 3.7.

Таблица 3.7

Номер интервала, k	Границы интервала, Δ_k	Наблюдаемая частота, n_k	Вероятность попадания в интервал, Δ_k, p_k	Ожидаемая частота, np_k	np_k	$n_k - np_k$	$\frac{(n_k - np_k)^2}{np_k}$
1	$-\infty$—12	2	0,0228	1,254	5,274	0,725	0,010
2	12—14	4	0,0731	4,020			
3	14—16	8	0,1686	9,273	9,273	–1,273	0,175
4	16—18	12	0,2576	14,168	14,168	–2,168	0,332
5	18—20	16	0,2484	13,662	13,662	–2,338	0,400
6	20—22	10	0,1519	8,354	12,633	0,366	0,011
7	22—$+\infty$	3	0,0778	4,279			
	Сумма	55	1,0001	55	55	—	0,928

В четвертом столбце табл. 3.5 приведены вероятности p_k, вычисляемые по формуле

$$p_k = P[X \in \Delta_k] = \Phi\left(\frac{b_k - \bar{x}}{s}\right) - \Phi\left(\frac{a_k - \bar{x}}{s}\right), \quad k = 1, 2, \ldots, 7,$$

где a_k и b_k — соответственно нижняя и верхняя границы интервала Δ_k, а значения функции $\Phi(x)$ берутся из таблицы (см. Приложение 3 или воспользуйтесь вероятностным калькулятором). В пятом столбце приводятся ожидаемые частоты np_k, а в шестом — значение np_k после объединения первых двух и последних двух интервалов.

Так как после объединения осталось $r = 5$ интервалов, а по выборке определены оценки двух параметров — математического ожидания и дисперсии, т. е. $l = 2$, то число степеней свободы равно $5 - 2 - 1 = 2$. По таблице квантилей распределения хи-квадрат (см. например [1] или воспользуйтесь вероятностным калькулятором) находим $\chi^2_{0,90}(2) = 4{,}61$. Выборочное значение статистики критерия равно $\chi^2_в = 0{,}928$, следовательно, гипотеза о нормальном распределении принимается.

Рассмотрим решение примера 3.12 в пакете STATISTICA. Запустите модуль Nonparametrics/Distrib. и введите выборку из примера 3.12 в переменную Var1.

Для проверки гипотезы о нормальном распределении по критерию χ^2 в пакете STATISTICA выполните следующие действия.

В меню Analysis выберите команду Startup Panel. В выпадающем меню выберите раздел Distribution Fitting (подбор распределений). Далее, в Continuous Distributions (непрерывные распределения) выберите Normal (нормальное распределение). Нажмите OK. В меню Fitting Continuous Distributions нажмите на кнопку Variable и выделите переменную VAR1. После выделения переменной на экране появятся оценки математического ожидания и дисперсии. По умолчанию число интервалов группировки равно 18. Это число можно изменить. Нажмите OK. На экран выводится таблица для расчета статистики критерия, значение статистики критерия (оно равно 1,83935), число степеней свободы (df) и вычисленный уровень значимости $p = 0{,}7652667$ (см. рис. 3.12а).

Так как вычисленный уровень значимости $p = 0{,}7652667$ превышает значение заданного уровня значимости $\alpha = 0{,}1$, то гипотеза о нормальном распределении принимается.

Variable VAR1 ; distribution: Normal (112.sta)

Kolmogorov-Smirnov d = ,0504031, p = n.s.
Chi-Square: 1,839395, df = 4, p = ,7652667 (df adjusted)

Upper Boundary	observed freq-cy	cumulativ observed	percent observed	cumul. % observed	exp fr
<= 9	0	0	0,00000	0,0000	
10,	0	0	0,00000	0,0000	
11,	1	1	1,81818	1,8182	
12,	1	2	1,81818	3,6364	
13,	1	3	1,81818	5,4545	1
14,	3	6	5,45455	10,9091	2
15,	4	10	7,27273	18,1818	3
16,	4	14	7,27273	25,4545	5
17,	5	19	9,09091	34,5455	6
18,	7	26	12,72727	47,2727	7
19,	7	33	12,72727	60,0000	7
20,	9	42	16,36364	76,3636	6
21,	6	48	10,90909	87,2727	4
22,	4	52	7,27273	94,5455	3
23,	1	53	1,81818	96,3636	2
24,	2	55	3,63536	100,0000	1
25,	0	55	0,00000	100,0000	

Рис. 3.12а. Таблица для расчета статистики хи-квадрат

3.4. Работы по статистическим методам

3.4.1. Работа 1. Оценивание характеристик генеральной совокупности по выборке. Методы группировки. Построение таблицы частот и гистограмм

1. Основные понятия

Генеральная совокупность, выборка, статистический и вариационные ряды, группировка, таблица частот группированной выборки, распределение выборки, эмпирическая функция распределения, гистограмма частот.

Выборочные характеристики генеральной совокупности: среднее, мода, медиана, дисперсия, асимметрия, эксцесс, начальные и центральные моменты, выборочные квантили и квартили, размах выборки.

Критерии для выбора наилучшей оценки параметров распределения: состоятельность, несмещенность, эффективность.

2. Варианты для вычислений в работах 1—3

1.	1	2	2	4	3	3	1	1	4	2	1	3	2	1	1	2	2	5	6	7
2.	0	6	2	3	5	8	3	2	1	9	4	4	9	1	3	2	6	1	2	4
3.	2	4	1	3	1	2	2	2	5	1	5	1	4	2	0	4	3	3	1	0
4.	2	4	5	3	1	7	5	4	5	7	6	6	5	2	2	1	4	7	2	1
5.	4	4	8	5	9	3	9	3	3	3	7	5	3	6	7	8	4	6	5	9
6.	8	4	9	3	7	3	4	9	6	9	5	8	4	7	7	4	3	5	4	9
7.	3	7	9	5	7	9	9	5	2	6	7	4	3	9	8	8	8	6	5	6
8.	14	10	9	6	9	6	6	8	10	9	11	7	8	9	7	12	8	7	8	13
9.	8	3	3	10	4	10	8	12	5	11	5	5	6	6	4	11	7	3	4	11
10.	7	8	12	11	8	8	12	10	8	12	5	11	9	10	11	10	12	8	7	9
11.	13	14	20	6	11	6	10	11	10	12	3	11	17	8	12	19	11	13	8	11
12.	12	11	11	16	14	16	10	9	11	9	11	13	9	15	11	10	15	14	14	11
13.	9	9	14	11	10	12	7	19	17	14	10	8	8	11	12	10	11	14	8	13
14.	11	10	12	17	17	19	11	15	5	11	11	14	6	7	14	11	13	13	8	5
15.	21	14	7	13	17	18	15	17	8	14	20	11	10	11	13	20	18	19	17	11
16.	19	13	12	21	14	19	16	14	8	11	15	13	17	17	15	18	15	20	16	13
17.	15	21	18	15	18	23	23	16	13	19	14	15	21	22	22	14	16	22	17	14
18.	18	17	15	23	12	28	17	17	18	10	25	20	17	18	23	23	21	27	25	22
19.	31	15	17	15	17	18	17	18	23	20	23	20	25	20	25	18	20	18	19	19
20.	18	21	19	17	19	17	19	21	24	19	23	21	18	20	22	24	19	20	22	18
21.	25	19	24	22	20	16	16	19	22	22	23	21	20	23	21	19	17	17	18	21
22.	19	13	26	16	27	19	23	32	18	18	14	17	21	22	23	24	25	30	31	26
23.	25	26	21	24	22	23	20	20	22	26	24	21	22	23	25	25	20	21	22	23
24.	12	20	23	27	17	23	23	21	25	20	22	25	21	17	15	13	14	24	18	19
25.	30	25	25	26	16	30	30	23	27	25	25	17	18	19	18	27	22	23	23	20

3. Задание

По выборке из своего варианта (объем выборки = 20) выполнить следующие расчеты и задания (использовать методы из п. 3.1.3, 3.1.4):

1. Построить статистический и вариационный ряды.
2. Вычислить оценки математического ожидания, моды и медианы, несмещенную и смещенную оценки дисперсии, размах выборки.
3. Построить таблицу частот и накопленных частот для сгруппированной выборки (число интервалов равно 4).
4. Построить гистограмму частот и относительных частот.
5. Ввести данные в пакет STATISTICA, выполнить все расчеты п. 1—4, сравнить результаты и записать в отчет.

4. Выполнение задания в пакете STATISTICA

Дана выборка объема 20:

11 10 12 17 17 19 11 15 5 11 11 14 6 7 14 11 13 13 8 5.

Запустите программу STATISTICA. Создайте новый файл для ввода данных, выполняя следующие операции **File-New Data...** Вместо имени файла **new.sta** введите новое имя, например, свою фамилию, чтобы в дальнейшем сохранить данные и показать их преподавателю. После ввода имени нажмите **OK**. Файл данных создан. Расширение имени файла **.sta** присваивается автоматически. В поле Var1 введите выборку приведенную выше. Если не достаточно строк (как правило, программа по умолчанию предлагает 10) воспользуйтесь кнопкой 

Cases (Выбор), затем **Add (добавить)** и в строке **Number of Cases to Add** (количество добавляемых строк) запишите число десять (для данного примера).

Для построения вариационного ряда, щелкните по кнопке переключателя модулей и вызовите модуль, который называется **Data Management (Управление данными)**, далее в подпункте **Analysis (Анализ)** главной строки выберите функцию **Sort (Сортировка)** (рис. 3.13) (убедитесь, что поле, в котором вы работаете, не изменилось — **Var1**). Сортировку данных можно проводить как в порядке возрастания (**Ascen**), так и в порядке убывания (**Desc**). Обратите внимание на следующие особенности процедуры **Sort**:

1. При выполнении процедуры сортировки, например, по переменной Var1, перемещаются *все строки* электронной таблицы.
2. Если в столбце значений переменной, по которой проводится сортировка, есть незаполненные (пропущенные) значения, то они считаются равными нулю и, следовательно, также участвуют в процедуре, а соответствующие им строки электронной таблицы перемещаются на соответствующие места.

После построения вариационного ряда в **Переключателе модулей** выберите **Basic Statistics/Tables (Основные статистики/таблицы)** и нажмите кнопку **Switch to (Переключиться в)**. Стартовая панель модуля вызывается кнопкой **Analysis** главной строки. В стартовой панели модуля выберите подменю **Descriptive Statistics** (рис. 3.14). Кнопка **More Statistics** отвечает за подсчет тех величин, которые вы хотите найти.

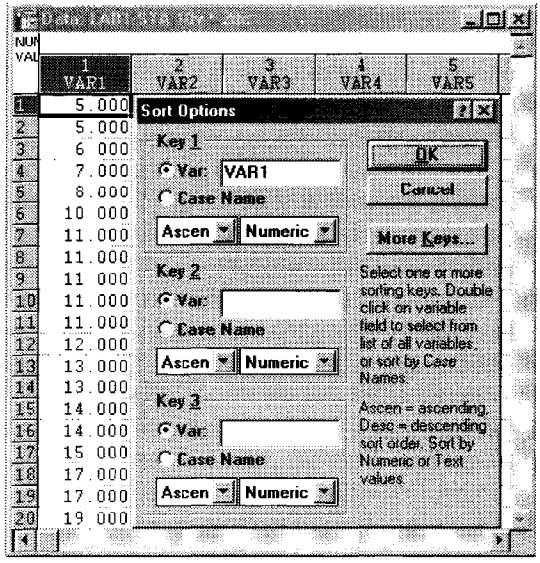


Рис. 3.13. Вид меню функции сортировки

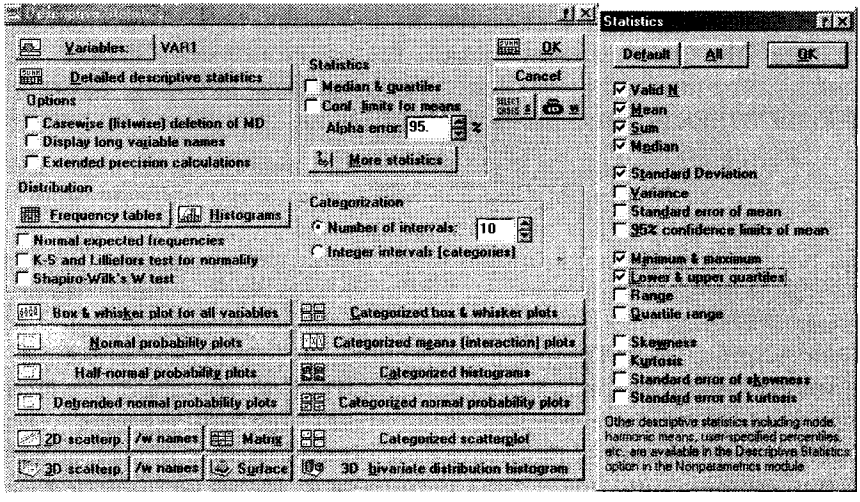


Рис. 3.14. Вид меню функции Descriptive Statistic (слева) и More Statistics (справа)

Программа предлагает вычисление следующих статистик:

Valid N — число элементов выборки (n);

Mean — среднее значение (выборочное среднее $\bar{x}$ — оценка математического ожидания);

Sum — сумма;

Median — оценка медианы;

Standard Deviation — стандартное отклонение (среднее квадратическое отклонение s);

Variance — несмещенная оценка дисперсии s^2;

Standard error of mean — стандартная ошибка среднего $= s/\sqrt{n}$;

95 % confidence limits of mean — 95%-е доверительные интервалы для среднего (математического ожидания генеральной совокупности);

Minimum & maximum — максимальное и минимальное значение выборки;

Lower & upper quartiles — верхний и нижний квартили;

Range — размах (разность между максимумом и минимумом);

Quartiles range — разность между верхним и нижним квартилем;

Skewness — выборочный коэффициент асимметрии;

Kurtosis — выборочный коэффициент эксцесса;

Standard error of skewness — стандартная ошибка коэффициента асимметрии;

Standard error of kurtosis — стандартная ошибка коэффициента эксцесса.

Отметьте необходимые характеристики, введите имя анализируемой переменной **Var1** в левом верхнем углу и нажмите OK.

Полученные результаты представлены в виде таблицы (рис. 3.14а).

Для проведения группировки выборки, в стартовой панели модуля Basic Stat./Tables выберем процедуру **Frequency tables** (рис. 3.15).

В этом диалоговом окне можно либо задать размер интервала — **Step size** (причем даже не обязательно задавать начальное значение, компьютер может сделать это автоматически при установленной «галочке» напротив слов **at minimum**), либо просто задать количество интервалов командой **No. of exact intervals** (установив количество интервалов равное четырем, для

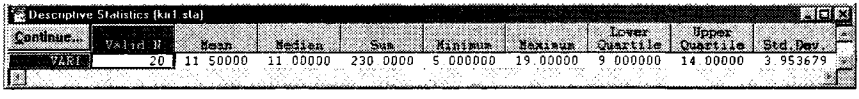


Рис. 3.14а. Таблица результатов

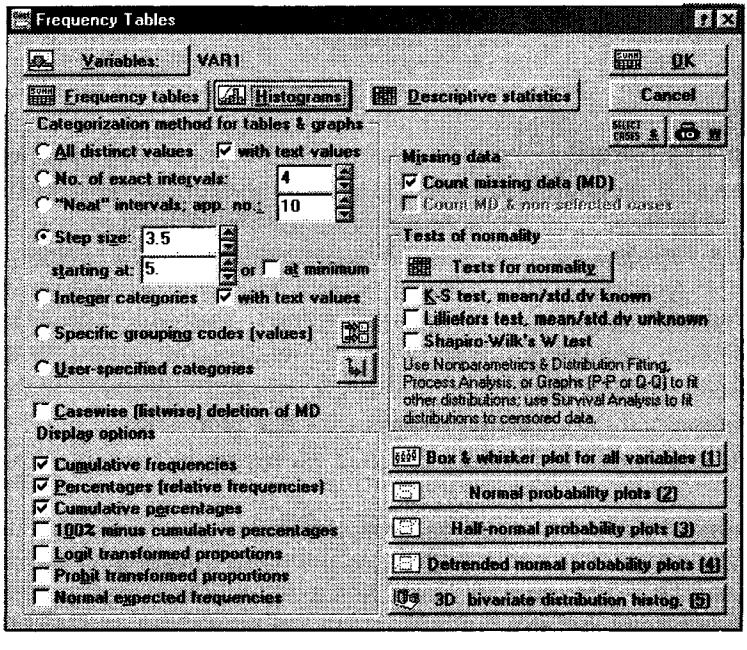


Рис 3.15. Диалоговое окно Frequency tables

данного примера). Результаты группировки данных (при установке размера интервала Step size: 3,5) приведены на рис. 3.16.

Для построения гистограммы воспользуемся этим же окном (рис. 3.15), в котором присутствует кнопка Histograms, отвечающая за построение.

Итоговый вид гистограммы представлен на рис. 3.17.

VAR1 (kir1.sta)

Continue...	Count	Cumul Count	Percent	Cumul Percent
5.00000<=x<8.50000	5	5	25.00000	25.0000
8.50000<=x<12.0000	7	12	35.00000	60.0000
12.0000<=x<15.5000	5	17	25.00000	85.0000
15.5000<=x<19.0000	3	20	15.00000	100.0000
Missing	0	20	0.00000	100.0000

Рис. 3.16. Результат группировки выборки

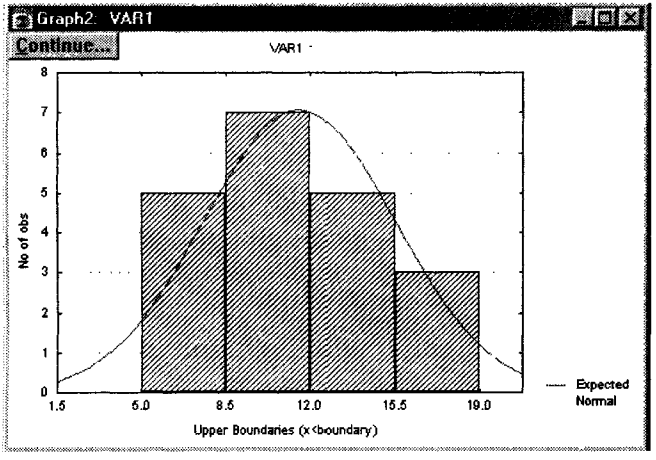


Рис. 3.17. Итоговый вид гистограммы

Кривая **Exp Normal** — график плотности нормального распределения с математическим ожиданием $\bar{x} = 11{,}5$; и средним квадратическим отклонением $s = 3{,}953679$.

Построим график накопленных относительных частот (огиву) для переменной Var1. Для этого нужно выполнить следующие операции: **Analysis** → **Frequency tables**, ввести имя переменной Var1 и нажать кнопку **Frequency tables** на панели процедуры (рис. 3.15), в появившейся таблице частот выбрать столбец **Cumul. percent** и щелкнуть правой кнопкой мыши по его имени, в меню выбрать **Custom Graphs...** → **line plot, OK**.

График накопленных относительных частот представлен на рис. 3.18.

Замечание. Если некоторые ячейки столбца-переменной не заполнены, то они учитываются как отсутствующие значения (**missing**). Например, если в файле 1.sta 22v * 30c, содержащем 30 строк, в качестве переменной **Var1** ввести подряд с первой строчки 20 чисел, то наблюдения 21—30 являются пропущенными значениями. При вычислении таблицы частот, пропущен-

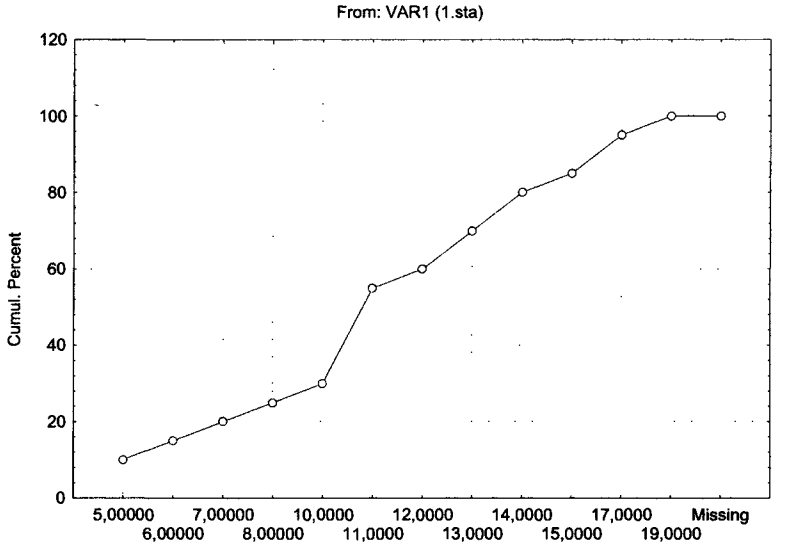


Рис. 3.18. График накопленных относительных частот

ные значения учитываются при подсчете относительных частот (**Percent**), накопленных относительных частот (**Cumul. Percent**): эти значения будут вычисляться делением не на 20 (по числу введенных чисел), а на 30, т. е. на число строк в переменной **Var1**.

3.4.2. Работа 2. Доверительные интервалы. Проверка гипотез о параметрах и виде распределения

1. Основные понятия

Выборка как система независимых случайных величин — выборочный вектор.

Распределение выборки и основных выборочных характеристик: выборочного среднего и выборочной дисперсии в случае нормально распределенной генеральной совокупности. Определения и свойства распределения χ^2, распределений Стьюдента и Фишера. Доверительные интервалы для среднего и дисперсии. Статистические гипотезы о параметрах и их проверка по критериям значимости. Статистика критерия. Уровень значимости. Критическая область. Ошибки 1 и 2 рода. Проверка гипотез о виде распределения. Критерий χ^2.

2. Задание

По выборке из своего варианта, используя результаты расчетов в Работе 1, выполнить следующие расчеты и задания:

1. Вычислить доверительные интервалы для среднего и дисперсии нормально распределенной генеральной совокупности при доверительной вероятности равной $1 - \alpha = 0{,}95$, $1 - \alpha = 0{,}90$ (см. п. 3.2.4).

2. На уровне значимости $\alpha = 0{,}01$ проверить гипотезы (п. 3.3.1):

а) $H_0 : m = m_0$, где $m_0 = \bar{x} + 0{,}5s$, где $\bar{x}$ — оценка среднего, а s — оценка среднего квадратического отклонения;

б) $H_0 : \sigma^2 = a_0$, $a_0 = 2s$.

3. На уровне значимости $\alpha = 0{,}05$ проверить гипотезу о нормальном распределении генеральной совокупности по критерию χ^2 (п. 3.3.4).

4. Ввести данные в пакет STATISTICA, выполнить расчеты п. 1—3, сравнить результаты и записать в отчет.

3. Выполнение задания в пакете STATISTICA

Доверительные интервалы для среднего нормально распределенной генеральной совокупности для заданной доверительной вероятности вычисляются в модуле **Basic Statistics/Tables** в подменю **Descriptive Statistics** (см. Работу 1). Например, 95 % доверительный интервал для среднего вычисляется по выборке объема 20 из Работы 1 и имеет границы: (9,649621; 13,35038). Используя этот результат можно на 5 % уровне значимости проверить гипотезу $H_0 : m = m_0$, при альтернативной гипотезе $H_1 : m \neq m_0$, где $m_0 = x + 0{,}5s = 11{,}5 + 0{,}5 \cdot 3{,}953679 = 13{,}47884$. Так как значение $m_0 = 13{,}47...$ не входит в 95%-й доверительный интервал, то гипотеза $H_0 : m = 13{,}47...$ отклоняется на уровне значимости 5 %.

Доверительные интервалы для дисперсии в пакете STATISTICA не вычисляются.

Рассмотрим проверку гипотез о виде распределения по критерию χ^2. Для примера проверим гипотезу о том, что выборка в Работе 1 (переменная **VAR1**) получена из генеральной совокупности, имеющей нормальное распределение. В качестве параметров распределения математического ожидания m и дисперсии σ^2 примем оценки этих параметров mean $= \bar{x} = 11{,}5$ и variance $= s^2 = 15{,}63$.

Для вычисления статистики хи-квадрат запустите модуль **Nonparametrics/Distrib**.

В меню **Analysis** выберите команду **Startup Panel**. В выпадающем меню выберите раздел **Distribution Fitting** (подгонка распределений). Далее, в **Continuous Distributions** (непрерывные распределения) выберите **Normal** (нормальное распределение). В меню **Fitting Continuous Distributions** нажмите на кнопку **Variable** и выделите переменную **VAR1**. Нажмите **OK**. В строке **Distribution** выберите пункт **Normal**.

Число интервалов группировки определяется программой автоматически. Пользователь может изменить число интервалов группировки или принять значение, предлагаемое программой.

Для сравнения со сделанными ранее расчетами установите число интервалов = 4, нижний предел = 5, верхний предел = 19. Нажав OK, получим таблицу для вычисления статистики χ^2. В данном примере число наблюдений (observed frequency) в последнем четвертом интервале = 3.

Так как это значение меньше чем 5, то при подсчете статистики χ^2 последний интервал объединяется с предпоследним, третьим интервалом и

число интервалов становится равным трем, число степеней свободы для статистики χ^2 равно 0 (3 – 2 – 1 = 0). Таким образом, использовать пакет STATISTICA для проверки гипотезы о нормальном распределении для выборки столь малого объема по критерию χ^2 в данном примере нельзя.

Интересно, что гипотеза о том, что выборка получена из генеральной совокупности, имеющей распределение χ^2 с числом степеней свободы равным 11,5 (среднему выборки) принимается на уровне значимости $p = 0{,}607$. Чтобы получить этот результат нажмите **Continue...** и в появившемся меню в строке **Distribution** установите **Chi-Square** и нажмите кнопку **Graph**.

Результаты процедуры содержат: результаты группировки, значение статистики χ^2 с одной степенью свободы (d.f. = 1, так как 3 – 1 – 1 = 1) и вычисленный уровень значимости $p = P[\chi^2(1) > 0{,}256] = 0{,}607$. Так как вычисленный уровень значимости p больше заданного уровня значимости $\alpha = 0{,}05$, то гипотеза принимается.

Заметим, что для проверки гипотезы о виде распределения по критерию χ^2 необходимо иметь выборку значительно большего объема, чем 20.

Дополнительные задания

• Смоделируйте несколько выборок объема 200 из нормального, экспоненциального и равномерного распределений и проверьте соответствующие гипотезы по критерию χ^2 используя меню **Fitting Continuous Distributions**.

• Решите следующие задачи (1—9).

В каждой задаче:

а) определите оценки среднего, дисперсии, медианы, нижнего и верхнего квартилей, коэффициентов асимметрии и эксцесса;

б) постройте 90 % доверительные интервалы для среднего и дисперсии;

в) постройте гистограммы используя 5 и 8 интервалов;

г) определите подходящее распределение и проверьте гипотезу о виде распределения по критерию χ^2.

1. Ниже приведен вес (в килограммах) 100 пациентов, желающих пройти курс лечения, чтобы снизить вес.

103	90	95	106	101	79	98	91	79	87
120	93	88	111	82	84	86	81	86	98
79	83	91	108	105	117	107	97	94	101
106	93	82	121	107	84	87	99	88	111
86	82	79	83	106	106	82	91	85	114
70	79	89	78	112	90	103	82	79	84
98	86	96	90	96	103	83	89	96	99
100	97	87	77	117	87	88	110	104	82
82	61	110	82	95	92	110	108	103	117
94	99	104	102	103	85	95	89	77	93

Используя данные, постройте огиву.
Определите процент пациентов имеющих вес более чем 100 кг.

2. Длины 25 танкеров проходящих через канал (в метрах) таковы:

66	65	96	80	71
93	66	96	75	61
69	61	51	84	58
73	77	89	69	92
57	56	55	78	96

Постройте огиву, которая поможет ответить на следующий вопрос. Пошлина собирается со всех танкеров, длина которых превышает 60 м. Какая доля танкеров пройдет через канал не уплачивая пошлины?

3. В среднем рыболовное судно за один рейс вылавливает 5 тыс. кг рыбы. Данные улова в 20 последних рейсах судна следующие:

6500	6700	3400	3600	2000
7000	5600	4500	8000	5000
4600	8100	6500	9000	4200
4800	7000	7500	6000	5400

Постройте огиву, которая поможет ответить на следующие вопросы.
Какова доля среднестатистического улова?
Какой улов представляет собой среднее значение в данной выборке?
Каков улов в 80 % рейсов?

4. В течение 50 дней фиксировалось время для набора титульного листа газеты. Данные (до десятой доли минуты) представлены ниже:

20,8	28,0	21,9	20,0	20,7	20,9	25,0	22,0	28,0	20,1
25,3	20,7	25,0	21,2	23,8	23,3	20,9	29,0	23,5	19,5
23,7	20,3	23,6	19,0	25,1	25,0	19,5	24,1	24,2	21,8
21,3	21,5	23,1	19,9	24,2	24,1	19,8	23,9	28,0	23,9
19,7	24,2	23,8	20,7	23,8	24,3	21,1	20,9	21,6	27

Расположите данные в виде вариационного ряда.
Постройте частотное распределение и распределение накопленных частот, используя интервал в 0,8 мин.
Постройте полигон частот.
Используя данные, постройте огиву.
Определите процент случаев, в которых страница набирается не более чем за 24 мин.

5. Данные, отражающие еженедельный рост ржи (в сантиметрах), следующие:

0,4	1,9	1,5	0,9	0,3	1,6	0,4	1,5	1,2	0,8
0,9	0,7	0,9	0,7	0,9	1,5	0,5	1,5	1,7	1,8

Представьте данные в виде вариационного ряда.

Используя интервалы длиной 0,25, постройте гистограмму относительных частот.

Постройте огиву и определите долю ржи, которая вырастает более чем на 1 см в неделю.

Какова величина среднего еженедельного роста ржи?

6. Были собраны данные о продолжительности ожидания прибытия автомобиля реанимации к пациенам:

Время ожидания, мин									
12	16	21	20	24	3	11	17	29	18
26	4	7	14	25	1	27	15	16	5

Представьте данные в виде вариационного ряда.

Используя 6 классов, постройте гистограмму частот

Как долго ждут автомобиля реанимации 75 % пациентов?

Определите среднее время ожидания автомобиля.

7. Менеджер компании фиксирует время (в мин), которое идет на переналадку и текущий ремонт оборудования в шахте в течение рабочей смены. Результаты 35 последних наблюдений приведены ниже:

60	72	126	110	91	115	112
80	66	101	75	93	129	105
113	121	93	87	119	111	97
102	116	114	107	113	119	100
110	99	139	108	128	84	99

Представьте данные в виде вариационного ряда.

Если среднее время простоя оборудования составляет 108 мин, то во скольких случаях оборудование простаивало более 108 мин, а во скольких — менее?

Постройте график относительных накопленных частот с 10-минутными интервалами.

8. Производительность труда бригады шахтеров (в тоннах угля за смену) следующая:

356	331	299	391	364	317	386
360	281	360	402	411	390	362
311	357	300	375	427	370	383
322	380	353	371	400	379	380
369	393	377	389	430	340	368

Постройте график относительных накопленных частот с шестью равными интервалами.

Во скольких случаях производительность была ниже 380 т за смену, а во скольких — выше?

9. Менеджер по техническому обеспечению в крупной авиакомпании решил проверить партию болтов, полученную от нового поставщика. 25 болтов из этой партии были отправлены на экспертизу для определения предельного усилия на излом. Результаты экспертизы приведены в тыс. кг:

67,0	62,3	56,8	64,0	66,1
54,4	60,5	64,5	62,9	57,0
64,4	59,3	58,9	64,0	61,2
56,7	58,5	64,4	53,8	60,3
68,5	57,0	57,3	63,9	62,7

Представьте данные в виде вариационного ряда.

Какая часть болтов выдержит усилие более, чем 54 432 кг, а какая часть — более, чем 68 040 кг?

По стандарту болт должен выдерживать усилие не менее чем 63 504 кг. Какая доля выборки окажется непригодной для использования в корпусе самолета?

3.4.3. Работа 3. Доверительные интервалы для разности средних и отношения дисперсий

Проверка гипотез о равенстве средних и дисперсий двух нормально распределенных генеральных совокупностей.

1. Основные понятия

Доверительные интервалы для разности средних и отношения дисперсий.

2. Задание

По выборкам из своего и следующего по номеру вариантов выполнить следующие расчеты и задания (п. 3.3.1):

1. Вычислить доверительный интервал для отношения дисперсий при доверительной вероятности равной 0,95.

2. Проверить гипотезу о равенстве дисперсий двух генеральных совокупностей.

3. Если в п. 2 гипотеза о равенстве дисперсий принимается, то вычислить доверительный интервал для разности средних и проверить гипотезу о равенстве средних на уровне значимости $\alpha = 0{,}05$.

4. Если в п. 2 гипотеза о равенстве дисперсий отклоняется, то проверить гипотезу о равенстве средних по критерию Уэлча (нижняя строка табл. 3.4).

5. Ввести данные в пакет STATISTICA, выполнить все расчеты п. 2—4, сравнить результаты и записать в отчет.

3. Выполнение задания в пакете STATISTICA

Проверим гипотезу о равенстве средних двух генеральных совокупностей по двум выборкам объема 20. Одна из выборок записана в переменной Var1 и анализировалась в Работах 1 и 2. Другая выборка: 21, 14, 7, 13, 17, 18, 15, 17, 8, 14, 20, 11, 10, 11, 13, 20, 18, 19, 17, 11 записывается в переменную **Var2**. Предполагается, что обе выборки получены из генеральных совокупностей, имеющих нормальное распределение, следовательно, для проверки гипотезы о равенстве средних используется статистика Стьюдента. В пакете STATISTICA для проверки гипотезы о равенстве средних последовательно используем: **Basic statistics** → **T-test for independent samples** (*T*-критерий для независимых выборок) → **Each variable contains the data for one group** (каждая выборка представлена как переменная). В соответствующие поля вводим **Var1** и **Var2** , результаты вычислений выводятся в таблице : средние обеих выборок $\bar{x}_1 = 11{,}5$; $\bar{x}_2 = 14{,}7$; объемы выборок $n_1 = 20$; $n_2 = 20$, средние квадратические отклонения выборок $s_1 = 3{,}954$; $s_2 = 4{,}144$; значение *t*-статистики

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}} \approx -2{,}499,$$

где $s = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$.

Число степеней свободы *t*-статистики: $df = 20 + 20 - 2 = 38$.

Приводится вычисленный уровень значимости

$$p = P[\,|T(n_1 + n_2 - 2)| \rangle\, |\,t\,|] = P[\,|T(38)| \rangle\ \ 2{,}499] = 0{,}0169,$$

где $T(38)$ — случайная величина, имеющая распределение Стьюдента с 38 степенями свободы.

Этот результат показывает, что на уровне значимости $\alpha = 0{,}05$ гипотеза о равенстве средних отклоняется (гипотеза принимается на уровне значимости $\alpha = 0{,}0169$).

Приведенная выше *t*-статистика может быть использована только в случае, если дисперсии обеих генеральных совокупностей равны. Гипотезу

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, можно проверить используя статистику s_2^2 / s_1^2 (в числитель ставится бóльшая оценка дисперсии): гипотеза H_0 принимается, если

$$\frac{s_2^2}{s_1^2} < F_{1-\frac{\alpha}{2}}(n_2 - 1, n_1 - 1),$$

где $F_{1-\frac{\alpha}{2}}(n_2 - 1, n_1 - 1)$ — квантиль распределения Фишера порядка $1 - \alpha/2$ с $n_2 - 1$ и $n_1 - 1$ степенями свободы.

Для рассматриваемого примера значение статистики $s_2^2 / s_1^2 = 1{,}098$, а квантиль $F_{0{,}975}(19,19) = 2{,}526$ (это значение можно вычислить в Probability Calculator). Так как $s_2^2 / s_1^2 < F_{0{,}975}(19,19)$, гипотеза о равенстве дисперсий принимается на уровне значимости $\alpha = 0{,}05$ и применение t-статистики правомерно.

Проверку гипотезы о средних двух нормально распределенных генеральных совокупностей можно провести, используя и другую опцию пакета STATISTICA: **Basic Stat./Tables → Other significance tests → Difference between two means**. В этой же опции можно проверить гипотезы о сравнении коэффициентов корреляции и долей признака двух генеральных совокупностей.

4. Используя пакет STATISTICA, решите следующие задачи

1. При исследовании влияния 2-х типов покрытия на удельную проводимость телевизионных трубок получены следующие результаты (в условных единицах):

№ трубки	1	2	3	4	5	6
1-й тип	6	5	12	9	10	—
2-й тип	14	11	0	5	6	8

Можно ли считать, что тип покрытия влияет на удельную проводимость трубок? Принять $\alpha = 0{,}10$.

2. Чтобы определить, какое влияние оказывает температура окружающей среды на систематическую ошибку угломерного инструмента, проведены измерения горизонтального угла объекта δ утром ($t = 10$ °C) и днем ($t = 26$ °C). Результаты измерений δ (в угловых секундах) следующие:

Утром	38,2	36,4	37,7	36,1	37,9	37,8	—	—
Днем	39,5	38,7	37,8	38,6	39,2	39,1	38,9	39,2

Можно ли считать, что температура окружающей среды влияет на систематическую ошибку угломерного инструмента? Принять $\alpha = 0{,}05$.

3. На двух станках А и В производят одну и ту же продукцию, контролируемую по внутреннему диаметру изделия. Из продукции станка А была взята выборка из 16 изделий, а из продукции станка В — выборка из 25 изделий. Выборочные оценки средних и дисперсий контролируемых разме-

ров $\bar{x}_A = 37{,}5$ мм при $s_A^2 = 1{,}21$ мм^2 и $\bar{x}_в = 36{,}8$ мм при $s_в^2 = 1{,}44$ мм^2. Используя 2-х сторонний критерий, проверить гипотезу о равенстве математических ожиданий контролируемых размерах в продукции обоих станков, если: а) $\alpha = 0{,}05$; б) $\alpha = 0{,}10$.

3.4.4. Работа 4. Группировка данных по классифицирующему признаку

Пример 3.11. Ниже приведены данные по 30 предприятиям:

A	B	C
1	27	21
2	28	35
3	41	38
4	44	46
5	55	51
6	33	30
7	37	38
8	49	50
9	56	61
10	37	34
11	38	35
12	49	39
13	26	19
14	29	36
15	20	24
16	47	40
17	36	35
18	56	60
19	57	48
20	45	43
21	39	45
22	46	48
23	60	46
24	55	57
25	53	34
26	42	42
27	41	47
28	35	30
29	33	41
30	46	27

Здесь A — номер предприятия; B — среднегодовая стоимость основных производственных фондов, млн. руб.; C — стоимость произведенной продукции, млн. руб.

Чтобы рассмотреть зависимость между среднегодовой стоимостью основных производственных фондов и стоимостью продукции, произведите группировку предприятий по среднегодовой стоимости производственных фондов, образовав 4 группы с равными интервалами.

По каждой группе и совокупности всех предприятий определите:

1) число предприятий;

2) среднегодовую стоимость основных производственных фондов — всего и в среднем на одно предприятие;

3) стоимость произведенной продукции — всего и в среднем на одно предприятие;

4) фондоемкость продукции — определяется путем деления столбца B на столбец C;

5) фондоотдачу (эффективность использования основных фондов) — определяется путем деления столбца C на столбец B;

Результаты группировки представьте в сводной таблице.

Решение. Создадим новый документ (**File** → **Open Other** → **New Data**, или просто **Ctrl** + **N**). Установим число строк в электронной таблице (**Cases**) равное 30.

В первый и третий столбцы введем производственные фонды (B), а во второй — производство продукции (C). В четвертый столбец введем фондоемкость продукции — определяется путем деления столбца B на столбец C; в пятый — фондоотдачу (эффективность использования основных фондов) — определяется путем деления столбца C на столбец B.

Назовем первый столбец ФОНДЫ, второй — ПРОИЗВ, третий — ФОНДЫ1, четвертый — ФЕ, пятый — ФО. Для этого двойным щелчком мыши вызываем диалоговое окно Variable1 и в поле name (имя) пишем название первой колонки — ФОНДЫ (рис. 3.19).

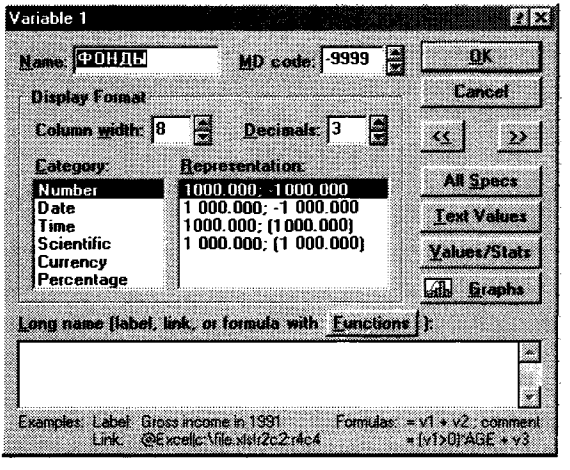


Рис. 3.19. Диалоговое окно для VAR1

Далее введем названия следующих переменных (используйте кнопку >>). В поле Long name переменных ФЕ и ФО введите функции соответственно =V1/V2 и =V2/V1 и выполните команду Recalculate (перерасчет) в меню Vars либо нажав инструментальную кнопку  .

Определим минимальное и максимальное значение среднегодовой стоимости производственных фондов по всем предприятиям.

Это уже делалось в работе 3.1 (используем опцию **Descriptive Statistics** рис. 3.13). Получим: max = 60, min = 20.

Найдем R, размах выборки, $R = 60 - 20 = 40$. Длина интервала группировки $l = \frac{R}{k} = \frac{40}{4} = 10$, где k — число интервалов, $k = 4$. В данном примере определим следующие четыре интервала группировки: [20,30); [30,40); [40,50); [50,60].

Чтобы произвести группировку переменной ФОНДЫ1 нужно щелчком правой кнопки мыши по столбцу ФОНДЫ1 вызывать меню и выбирать подпункт **Modify Variable(s) (Изменение переменных) — Recode (перекодировка)** (рис. 3.20).

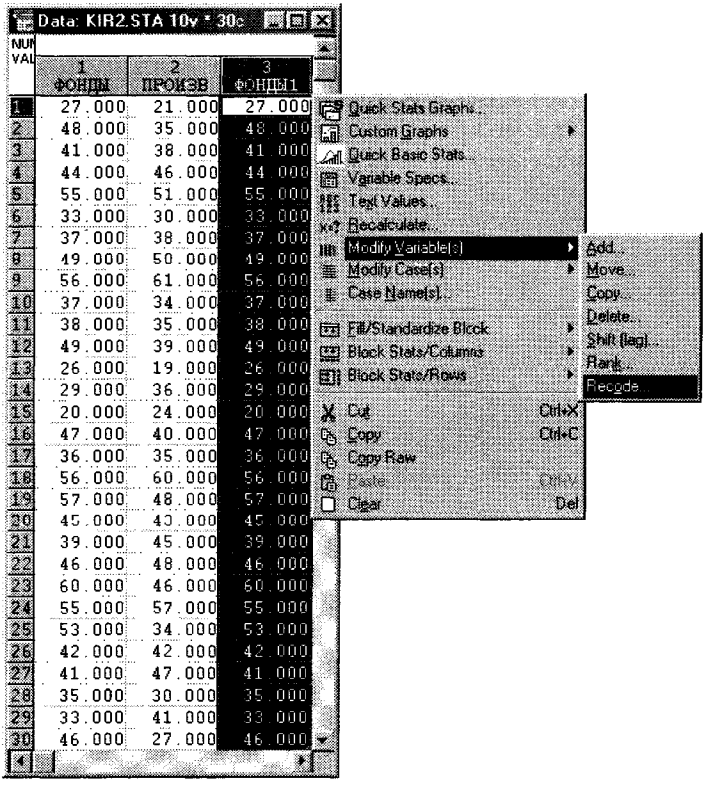


Рис. 3.20. Команды для преобразования переменной ФОНДЫ1

В появившемся окне (рис. 3.21) задаем границы интервалов следующими командами:

1) V3 >= 20 and V3 < 30

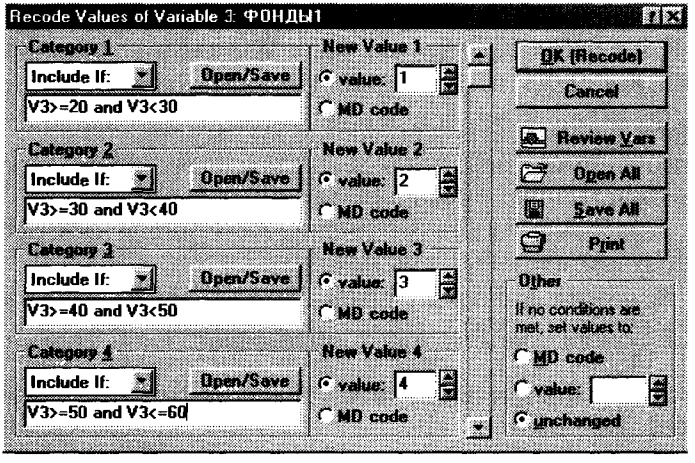


Рис. 3.21. Запись границ интервалов

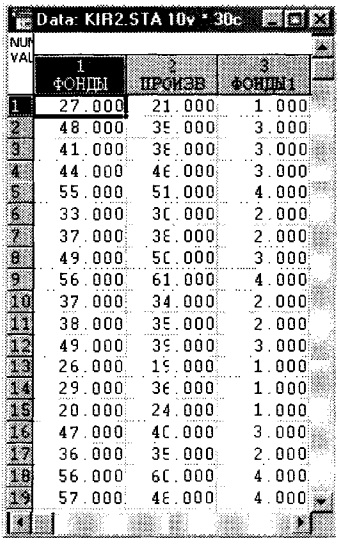
2) V3 > = 30 and V3 < 40

3) V3 > = 40 and V3 < 50

4) V3 > = 50 and V3 < = 60, где V3 — третий столбец (ФОНДЫ1), OK.

В третьем столбце вместо введенных нами данных появятся номера групп 1, 2, 3, 4, в которые попали предприятия (рис. 3.22).

Теперь предприятия можно упорядочить по третьему столбцу. Вызовите модуль **Data Management (Управление данными)**, в подпункте **Analysis (Анализ)** главной строки выберите функцию **Sort (Сортировка)**, убедитесь, что поле, в котором вы работаете, не изменилось — ФОНДЫ1. После выполнения сортировки по трем столбцам в порядке возрастания (Ascen) предприятия будут упорядочены по четырем группам.


Data: KIR2.STA 10v * 30c

	1 ФОНДЫ	2 ПРОИЗВ	3 ФОНДЫ1
1	27.000	21.000	1.000
2	48.000	35.000	3.000
3	41.000	36.000	3.000
4	44.000	46.000	3.000
5	55.000	51.000	4.000
6	33.000	30.000	2.000
7	37.000	36.000	2.000
8	49.000	50.000	3.000
9	56.000	61.000	4.000
10	37.000	34.000	2.000
11	38.000	35.000	2.000
12	49.000	39.000	3.000
13	26.000	19.000	1.000
14	29.000	36.000	1.000
15	20.000	24.000	1.000
16	47.000	40.000	3.000
17	36.000	35.000	2.000
18	56.000	60.000	4.000
19	57.000	46.000	4.000

Рис. 3.22. Разбиение предприятий по четырем группам

Для определения необходимых данных по группам предприятий вернитесь в модуль **Basic Stat/Tables** и в меню стартовой панели модуля выберите опцию **Breakdown and one-way ANOVA** (разбиение и однофакторный дисперсионный анализ).

В качестве группирующей переменной (**grouping**) выберете **Var3** (ФОНДЫ1), а в качестве зависимых переменных (**dependent**): **Var1**, **Var2**, **Var4** и **Var5**, OK.

В окне результатов (рис. 3.23) укажите необходимые статистики для групп предприятий: **Number of obser.** (число наблюдений в группе), **Sums** (суммы показателей по группам).

Далее, в левой части окна результатов (рис. 3.23) нажмите верхнюю кнопку: **Summary table of means** (таблица средних).

В полученной таблице (рис. 3.24) содержатся все необходимые результаты по группам предприятий для переменных ФОНДЫ, ПРОИЗВ, ФЕ и ФО.

Рассмотрите таблицу результатов (рис. 3.24). Столбцы таблицы: **means**, **N**, **sums**, ... — суть новые переменные. Это позволяет провести их дальнейший анализ. Определим, например, как изменяется средняя фондоемкость продукции по группам предприятий. Построим график переменной ФЕ means (предварительно нужно удалить последнюю строку таблицы — **All Grps**). Для этого щелчком правой кнопки мыши на имени переменной ФЕ means вызовите меню и выполните следующие действия: **Custom Graphs** → **2D Graphs-OK**. Полученный график приведен на рис. 3.25.

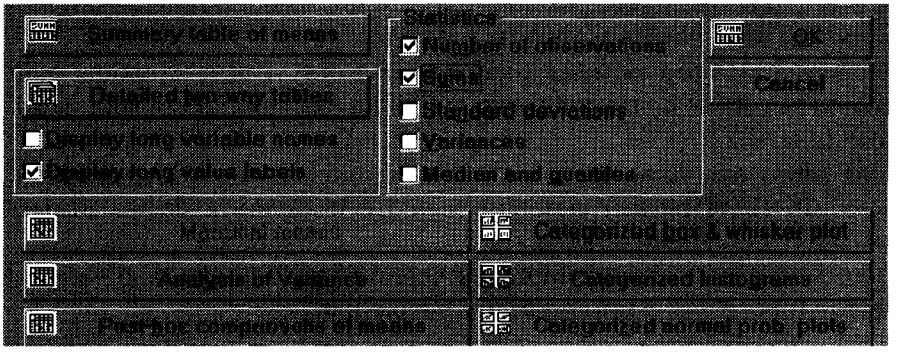


Рис. 3.23. Окно результатов классификации и дисперсионного анализа

Summary Table of Means (41.sta)
N=30 (No missing data in dep. var. list)

ФОНДЫ1	ФОНДЫ Means	ФОНДЫ N	ФОНДЫ Sum	ПРОИЗВОД Means	П
G_1:1	26,00000	5	130,000	27,00000	
G_2:2	36,00000	8	288,000	36,00000	
G_3:3	45,00000	10	450,000	42,00000	
G_4:4	56,00000	7	392,000	51,00000	
All Grps	42,00000	30	1260,000	40,00000	

Рис. 3.24. Таблица результатов группировки

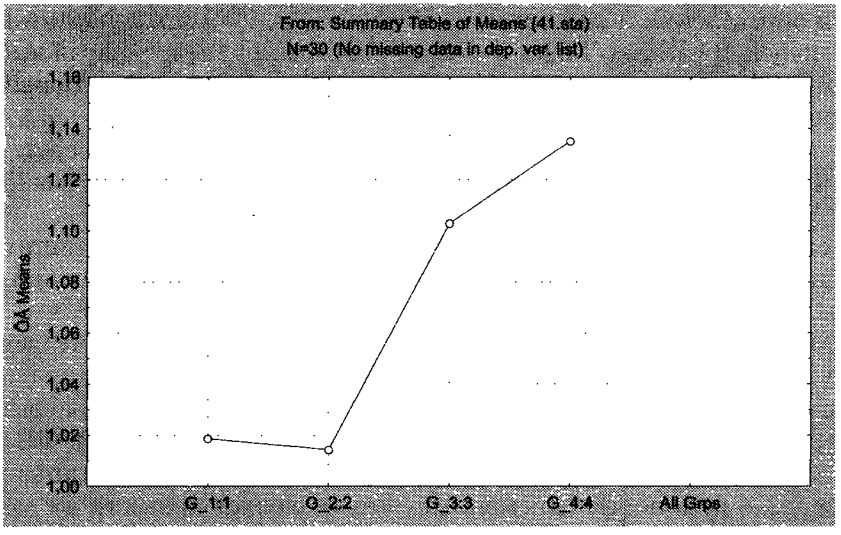


Рис. 3.25. График переменной ФЕ means

Варианты заданий к работе № 4

Задание 1.

Рассмотрите результаты по группам предприятий для примера 3.11 (рис. 3.24). Постройте график переменной ФО means. Наблюдается ли зависимость между номером группы и фондоотдачей?

Используя буфер обмена (процедуры Copy—Paste) столбцы таблицы можно переписать в электронную таблицу, содержащую файл исходных данных (рис. 3.22).

Для этого щелчком правой кнопки мыши на имени переменной (например, ФОНДЫ means) вызовите меню и выполните необходимые действия.

Можно выделить и скопировать через буфер обмена всю таблицу результатов (рис. 3.24). Выполните эту операцию.

Задание 2.

Для выявления зависимости между объемом розничного товарооборота и издержками обращения *сгруппируйте* предприятия торговли по объему розничного товарооборота, образовав, пять групп с равными интервалами.

По каждой группе предприятий торговли и их совокупности *определите*:

1) число предприятий;

2) объем розничного товарооборота — всего и в среднем на одно предприятие;

3) сумму издержек обращения — всего и в среднем на одно предприятие;

4) относительный уровень издержек обращения (удельный вес издержек обращения в объеме розничного товарооборота).

Результаты группировки *представьте* в таблице. Наблюдается ли зависимость между номером группы и удельным весом издержек обращения?

№ п/п	Розничный товарооборот, млн руб.	Издержки обращения, млн руб.
1	510	30
2	560	33
3	800	46
4	465	31
5	225	16
6	390	25
7	640	39
8	405	26
9	200	15
10	425	34
11	570	37
12	472	28
13	250	19
14	665	38
15	650	36
16	620	35
17	380	24
18	550	38
19	750	44
20	660	36
21	450	27
22	563	34
23	400	26
24	553	38
25	772	45

Задание 3.

Для изучения зависимости между объемом работ и накладными расходами *произведите* группировку по объему работ, образовав четыре группы предприятий с равными интервалами.

По каждой группе и совокупности предприятий *определите*:

1) число предприятий;

2) объем работ — всего и в среднем на одно предприятие;

3) объем накладных расходов — всего и в среднем на одно предприятие;

4) долю накладных расходов в объеме произведенных работ.

Результаты *представьте* в таблице. Существует ли зависимость между номером группы и долей накладных расходов?

№ п/п	Объем работ, млн руб.	Накладные расходы, млн руб.
1	9,0	2,7
2	10,3	3,0
3	7,0	2,5
4	5,2	2,2
5	6,4	2,5
6	9,5	2,7
7	14,0	4,0
8	13,0	4,0
9	5,0	2,0
10	7,4	2,6
11	9,3	2,6
12	8,0	3,2
13	10,2	2,3
14	10,0	3,0
15	12,0	2,6
16	15,0	3,0
17	16,0	5,0
18	17,0	4,3
19	21,0	5,0
20	19,0	4,8
21	12,5	4,0
22	8,0	2,0
23	11,0	3,0
24	6,5	2,0
25	13,0	5,0

Задание 4.

Для изучения зависимости между размером нераспределенной прибыли и инвестициями в основные фонды *произведите* группировку по размеру нераспределенной прибыли, образовав четыре группы предприятий с равными интервалами.

По каждой группе предприятий и совокупности в целом *определите*:

1) число предприятий;

2) размер нераспределенной прибыли — всего и в среднем на одно предприятие;

3) размер инвестиций в основные фонды — всего и в среднем на одно предприятие;

4) долю инвестиций в объеме нераспределенной прибыли.

Результаты группировки *представьте* в сводной таблице. Существует ли зависимость между номером группы и долей инвестиций?

№ п/п	Нераспределенная прибыль, млн руб.	Инвестиции в основные фонды, млн руб.
1	2,3	0,03
2	3,4	0,30
3	4,3	0,40
4	5,0	0,60
5	6,0	1,00
6	2,0	0,16
7	3,6	0,20
8	4,2	0,30
9	5,8	1,00
10	4,7	0,60
11	2,7	0,11
12	3,8	0,40
13	4,5	0,70
14	4,8	0,70
15	4,4	0,50
16	5,5	0,80
17	5,6	0,70
18	4,1	0,30
19	3,6	0,30
20	5,7	0,90

Задание 5.

Рассмотрите данные о стоимости однокомнатных квартир (Приложение 1.2).

Выполните следующие расчеты ($\alpha = 0{,}05$):

1. Для стоимости и общей площади квартир найдите оценки следующих параметров: средних, моды, медианы, максимальных и минимальных значений, нижних и верхних квартилей.

2. Определите оценки среднего, доверительные интервалы и постройте гистограммы стоимости квартир отдельно:

- для квартир расположенных на первом и последнем этажах;
- для квартир в кирпичных домах;
- для квартир расположенных не на первых и последних этажах и не в кирпичных домах.

Можно ли считать, что стоимости квартир в двух последних категориях существенно различаются?

3. Сравните средние цены на квартиры в Коньково и на Ленинском проспекте. Различаются ли эти цены?

4. Используя процедуру **Break-down and one-way ANOVA** из модуля **Basic Statistics and Tables**, определите число продаваемых квартир в каждом районе и их средние стоимости и минимальные и максимальные значения. (В качестве группирующей (Grouping) переменной возьмите переменную Region).

Постройте гистограммы (используйте опцию **Categorized histograms**).

4. Используя процедуру **Tables and Banners** из модуля **Basic Statistics**, постройте 2×2 таблицу сопряженности для переменных CAT и FLOOR. Определите сколько квартир из приведенного списка (69 квартир) расположены не на первых и не на последних этажах не кирпичных домов.

Глава 4
НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Основные методы математической статистики: оценка параметров распределения, проверка статистических гипотез, дисперсионный анализ применяются в предположении, что распределение генеральной совокупности известно. В частности t-критерий для сравнения средних двух генеральных совокупностей и однофакторный дисперсионный анализ для сравнения средних нескольких совокупностей пригодны только в случае нормального распределения последних. Однако часто встречаются данные для которых эти предположения не выполняются. Например, результаты социологических опросов обычно имеют форму ответов вида «да» или «нет» и представляются в виде таблиц, содержащих частоты положительных и отрицательных ответов. Традиционные методы математической статистики не могут быть использованы для обработки таких данных. В этих случаях используются непараметрические методы, т. е. методы независящие от распределения генеральной совокупности.

Непараметрические методы применяются для качественных данных, представленных в номинальной шкале и для данных, измеряемых в порядковой шкале (т. е. представленных в виде рангов), а также для количественных данных в том случае, когда распределение генеральной совокупности неизвестно.

В пакете STATISTICA непараметрические процедуры выполняются в модуле **Nonpametrics/Distrib**. Стартовая панель модуля приведена на рис. 4.1.

Мы последовательно опишем соответствующие методы и приведем примеры выполнения процедур.

В модуле **Nonpametrics/Distrib** содержится большое количество процедур.

При решении конкретной задачи необходимо выбрать тот или иной метод. Помощь в таком выборе может оказать следующая классификация непараметрических методов, используемых для проверки гипотезы о том, что анализируемые данные — это выборки из однородных генеральных совокупностей. Заметим, что понятие однородности генеральных совокупностей понимается достаточно широко: это могут быть генеральные совокупности, имеющие одну и ту же функцию распределения, либо совокупности, у которых совпадают характеристики положения (средние, медианы) и/или характеристики разброса (дисперсии).

Первым критерием для выбора метода является, очевидно, вид шкалы, в которой представлены исходные данные.

Вторым критерием является вид выборок (независимые или связанные) и их количество.

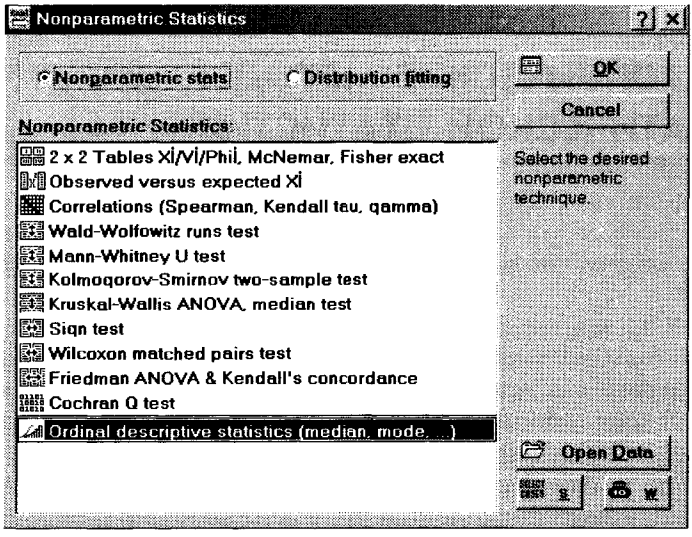


Рис. 4.1. Стартовая панель модуля Nonpametrics/Distrib

Поясним понятие связанной выборки. Если над каждым из n объектов или индивидуумов проводятся два наблюдения: одно до, а другое после некоторого воздействия (приема лекарства, обучения, рекламной компании, обработки тем или иным способом и т. д.), то результаты наблюдений представляют две связанные (зависимые) выборки объема n. В случае если каждый из n объектов подвергается k воздействиям, то результаты наблюдений представляют k связанных выборок объема n. Например, множество оценок, проставленных k судьями каждому из n спортсменов — это k связанных выборок объема n, измеренных в порядковой шкале. Таким образом, рассматриваемые ниже методы можно классифицировать следующим образом.

1. Исходные данные: две независимые выборки объемов n_1 и n_2.

Проверяемая гипотеза H_0: выборки принадлежат однородным генеральным совокупностям.

Методы:

1) критерий серий Вальда—Вольфовица;

2) критерий Манна—Уитни;

3) двухвыборочный критерий Колмогорова—Смирнова.

2. Исходные данные: пары наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$ двух признаков X и Y, измеренных в порядковых или количественных шкалах.

Проверяемая гипотеза H_0: признаки X и Y некоррелированны.

Меры статистической зависимости: ранговый коэффициент корреляции Спирмена, коэффициент корреляции τ Кендалла.

3. Исходные данные: k независимых выборок объемов $n_1, n_2, \ldots, n_k$.

Проверяемая гипотеза H_0: выборки принадлежат однородным генеральным совокупностям.

Методы:

1) однофакторный дисперсионный анализ Краскела—Уоллиса;

2) медианный критерий.

4. Исходные данные: две связанные выборки объемов *n*.

Проверяемая гипотеза H_0 : выборки принадлежат однородным генеральным совокупностям.

Методы:

1) критерий знаков;

2) критерий Вилкоксона.

5. Исходные данные: *k* связанных выборок объемов *n*.

Проверяемая гипотеза H_0 : выборки принадлежат однородным генеральным совокупностям.

Методы:

1) двухфакторный анализ Фридмана;

2) меры связи — коэффициент конкордации Кендалла.

6. Связанные выборки, измеряемые в номинальной шкале.

1) Исходные данные: две связанные выборки объемов *n* переменных *X* и *Y*, каждая из которых принимает два значения (0,1; +, –; и т. д.).

Проверяемая гипотеза H_0 : эффект воздействия отсутствует.

Метод: критерий Макнимара.

2) Исходные данные: *k* связанных выборок объемов *n* переменных X_1, X_2, ..., X_k, каждая из которых принимает два значения.

Проверяемая гипотеза H_0 : эффект воздействия отсутствует.

Метод: критерий Кокрена.

7. Выборки, измеряемые в номинальной шкале.

1) Исходные данные: выборки двух случайных переменных *X* и *Y*, каждая из которых принимает два значения.

Проверяемая гипотеза H_0 : *X* и *Y* независимы.

Метод: анализ таблицы сопряженности 2 × 2 (точный критерий Фишера, критерий χ^2).

2) Исходные данные: выборки двух переменных *X* и *Y*, представленных в номинальных шкалах. *X* принимает *k* значений, *Y* — *r* значений.

Проверяемая гипотеза H_0 : *X* и *Y* — независимы.

Метод: анализ таблицы сопряженности $k \times r$ (критерий χ^2). Анализ таких таблиц проводится в модуле **Basic Stat and Tables**, опция **Tables and banners**.

Далее методы и процедуры непараметрической статистики приводятся по порядку их расположения в стартовой панели модуля (рис. 4.1).

4.1. Таблицы сопряженности 2 × 2, статистики χ^2, φ, критерий Макнимара, точный критерий Фишера (2 × 2 Tables Xi/Vi/Phi, McNemar, Fisher exact)

В таблице сопряженности 2 × 2 записываются частоты для двух случайных переменных *X* и *Y*, каждая из которых принимает два значения: 0 и 1, «да» и «нет» и т. д.

Пример 4.1. Чтобы определить отношение телезрителей разного пола к телевизионной передаче опросили 60 человек: 35 мужчин и 25 женщин. Оказалось, что 25 мужчин одобряют, а 10 — не одобряют эту передачу. В то же время 16 женщин высказывают свое отрицательное отношение к передаче, а 9 — положительное. Выяснить зависит ли отношение к передаче от пола телезрителей.

Решение. Данные можно записать в виде таблицы сопряженности 2×2:

Пол	Отношение к передаче	
	за	против
Мужчины	25	10
Женщины	9	16

Формально задача состоит в определении независимости двух рассматриваемых признаков X (пол) и Y (отношение к передаче) или в проверке нулевой гипотезы H_0 : отношение к передаче не зависит от пола, при альтернативной гипотезе H_1 : отношение к передаче зависит от пола.

Для проверки гипотезы H_0 : X и Y независимы применяется *критерий* χ^2.

Чтобы пояснить необходимые расчеты запишем таблицу сопряженности 2×2 в следующем виде:

Пол	Отношение к передаче		
	за	против	сумма по строкам
Мужчины	$n_{11} = a$	$n_{12} = b$	$n_{1\cdot} = a + b$
Женщины	$n_{21} = c$	$n_{22} = d$	$n_{2\cdot} = c + d$
Сумма по столбцам	$n_{\cdot 1} = a + c$	$n_{\cdot 2} = b + d$	$n = a + b + c + d$

В рассматриваемом примере эта таблица имеет вид:

Пол	Отношение к передаче		
	за	против	сумма по строкам
Мужчины	25	10	35
Женщины	9	16	25
Сумма по столбцам	34	26	60

Статистика критерия χ^2 использует разности между наблюдаемыми частотами a, b, c, d и ожидаемыми частотами a_0, b_0, c_0, d_0, вычисляемыми при условии, что гипотеза H_0 о независимости признаков верна:

$$a_0 = \frac{(a+b)(a+c)}{n} = \frac{35 \cdot 34}{60} \approx 19{,}83;$$

$$b_0 = \frac{(a+b)(b+d)}{n} = \frac{35 \cdot 26}{60} \approx 15{,}17;$$

$$c_0 = \frac{(c+d)(a+c)}{n} = \frac{25 \cdot 34}{60} \approx 14{,}17;$$

$$d_0 = \frac{(c+d)(b+d)}{n} = \frac{25 \cdot 26}{60} \approx 10{,}83.$$

Выборочное значение статистики $\chi^2_в$ вычисляется по формуле:

$$\chi^2_в = \frac{(a-a_0)^2}{a_0} + \frac{(b-b_0)^2}{b_0} + \frac{(c-c_0)^2}{c_0} + \frac{(d-d_0)^2}{d_0} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

При $n \to \infty$ статистика $\chi^2_в$ имеет распределение хи-квадрат с одной степенью свободы. Если ожидаемые частоты ≤ 5, то выборочное значение статистики $\chi^2_в$ вычисляют с поправкой Йетса на непрерывность:

$$\chi^2_в = \frac{(|a-a_0|-0{,}5)^2}{a_0} + \frac{(|b-b_0|-0{,}5)^2}{b_0} + \frac{(|c-c_0|-0{,}5)^2}{c_0} +$$

$$+ \frac{(|d-d_0|-0{,}5)^2}{d_0} = \frac{n\left(ad-bc-\frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

Гипотеза H_0 принимается на уровне значимости α, если $\chi^2_в < \chi^2_{1-\alpha}(1)$, где $\chi^2_{1-\alpha}(1)$ — квантиль распределения хи-квадрат с одной степенью свободы порядка $1-\alpha$.

Для данного примера выборочное значение $\chi^2_в = 7{,}45$, а с поправкой Йетса — $\chi^2_в = 6{,}08$. Так как $\chi^2_{0{,}95}(1) = 3{,}84$ (проверьте, используя статистический калькулятор!) и $\chi^2_в > 3{,}84$, то гипотеза H_0 отклоняется: на уровне значимости $\alpha = 0{,}05$ следует считать, что отношение к передаче зависит от пола.

Эти же результаты получим, введя данные в соответствующую процедуру пакета STATISTICA. Таблица результатов приведена на рис. 4.2.

P-значения для статистики χ^2 и статистики χ^2, скорректированной по Йетсу, соответственно равны: 0,0063 и 0,0137. Таким образом, на уровне значимости $\alpha = 0{,}05$ гипотеза H_0 отклоняется. В таблице результатов приводится мера связи между переменными X и Y — *коэффициент фи-квадрат* (средний коэффициент сопряженности):

$$\varphi^2 = \frac{\chi^2_в}{n} = 0{,}124.$$

2 x 2 Table (nonparametr.sta)

Continue...	Column 1	Column 2	Row Totals
Frequencies, row 1	25	10	35
Percent of total	41,667%	16,667%	58,333%
Frequencies, row 2	9	16	25
Percent of total	15,000%	26,667%	41,667%
Column totals	34	26	60
Percent of total	56,667%	43,333%	
Chi-square (df=1)	7,45	p= ,0063	
V-square (df=1)	7,33	p= ,0068	
Yates corrected Chi-square	6,08	p= ,0137	
Phi-square	,12424		
Fisher exact p, one-tailed		p= ,0066	
two-tailed		p= ,0087	
McNemar Chi-square (A/D)	1,56	p= ,2115	
Chi-square (B/C)	0,00	p=1,0000	

Рис. 4.2. Результаты процедуры 2 × 2 Tables...

Значение φ^2 изменяется от 0 (между переменными нет зависимости) до 1 (между переменными имеется абсолютная зависимость, т. е. все частоты расположены на одной из диагоналей таблицы 2 × 2).

Если суммарный объем n выборки небольшой ($n \le 30$), то для проверки гипотезы H_0 применяется *критерий Фишера*, (см. [14], с. 345). Односторонние (*one-tailed*) и двусторонние (*two-tailed*) уровни значимости (p) для критерия Фишера (*Fisher exact p*) вычисляются и приводятся в таблице результатов выполнения процедуры для таблицы сопряженности 2 × 2 (рис. 4.2): $p = 0{,}0066$ и $p = 0{,}0087$.

Критерий значимости изменений Макнимара применяется, если исходные данные — две связанные выборки. Над одним и тем же объектом или индивидуумом проводятся два наблюдения: одно до, другое после некоторого воздействия (приема лекарства, обучения, рекламной компании и т. д.). Отрицательный результат или ответ обозначим знаком «–», а положительный — знаком «+».

Пример 4.2. Двести покупателей магазина бытовой техники дали ответы на вопрос: «Хотите ли вы купить кухонный комбайн новой марки?» до и после того как им был показан рекламный ролик. Частоты ответов приведены в таблице 2 × 2:

До	После	
	–	+
+	$a = 10$	$b = 71$
–	$c = 74$	$d = 45$

Показывают ли эти результаты, что просмотр рекламного ролика оказал эффективное воздействие на покупателей?

Решение. Очевидно, что эффективность воздействия определяется числом покупателей изменивших свое мнение с «+» на «–» и с «–» на «+», т. е.

частотами в клетках a и d. Задача состоит в проверке нулевой гипотезы H_0: в генеральной совокупности доля p покупателей, изменивших ответ с «+» на «–» равна доле q покупателей, изменивших ответ с «–» на «+» или проверке гипотезы $H_0: p = q = \frac{1}{2}$ при альтернативной гипотезе $H_1: p \neq q \neq \frac{1}{2}$.

Объем выборки n из биномиального распределения равен сумме частот в клетках a и d, $n = a + d = 55$. При $n \geq 50$ для проверки гипотезы H_0 используется статистика χ^2. Выборочное значение статистики $\chi^2_в$ вычисляется по формуле

$$\chi^2_в = \frac{(|a - d| - 1)^2}{a + d} \approx 21{,}02.$$

Так как $\chi^2_в$ больше квантили распределения $\chi^2_{0,95}(1) = 3{,}84$, гипотеза H_0 отклоняется на уровне значимости $\alpha = 0{,}05$. Таким образом, результаты свидетельствуют о том, что рекламный ролик оказал эффективное воздействие на покупателей. При $n < 50$ для определения границ критической области нужно использовать накопленные значения вероятностей биномиального распределения с $p = \frac{1}{2}$.

В пакете STATISTICA накопленные вероятности биномиального распределения (при $p = \frac{1}{2}$) вычисляются при помощи функции **IBinom (x;0,5;n)**. Нажмите кнопку **Functions** в окне спецификации переменных. В появившемся диалоговом окне **Function Wizard** выберите нужную функцию биномиального распределения (рис. 4.3): в окне **Category** выберите **Distributions**, в окне **Name** выберите **IBinom**. Нажмите кнопку **Insert**. Функция биномиального распределения появится в окне спецификации переменной.

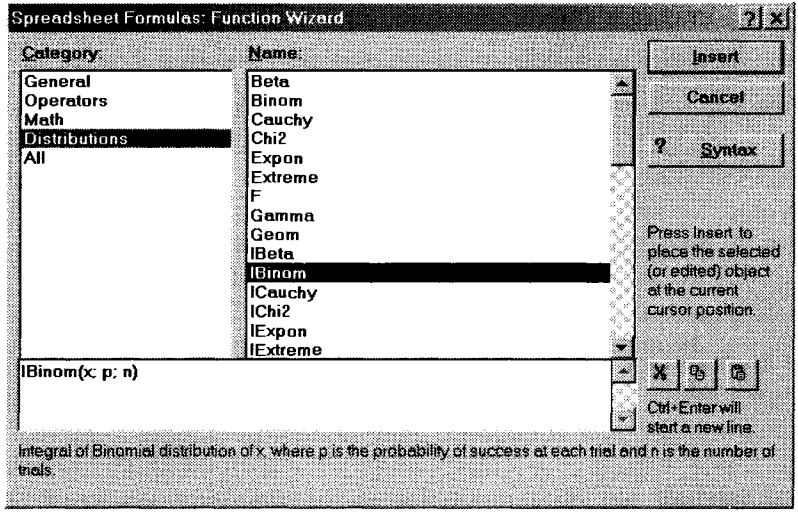


Рис. 4.3. Выбор биномиального распределения в конструкторе функций

Для рассматриваемого примера, при $n = 55$, $\alpha = 0{,}05$ и для двусторонней гипотезы H_1 критические значения будут $x = 19$, (**IBinom(19;0,5;55)** $\approx 0{,}015$) и $x = 55 - 19 = 36$, (**IBinom(36;0,5;55)** $\approx 0{,}9850$, $1 - 0{,}985 = 0{,}015$), так как соответствующие вероятности должны приблизительно быть равными $\alpha/2 = 0{,}025$. Наблюдаемые частоты попадают в критическую область: $a = 10 < 19$, $d = 45 > 36$ и, следовательно, на уровне значимости $\alpha = 0{,}05$ гипотеза H_0 отклоняется.

4.1.1. Задачи

Решите следующие задачи. Определите меры связи χ^2 и φ^2, прокомментируйте результаты.

Задача 1. Исследуются два производственных процесса изготовления поршневых колец. Используя критерий χ^2, проверьте гипотезу о равенстве процента брака в обоих процессах по следующим данным при $\alpha = 0{,}01$:

Кольца	Процесс	
	1	2
Годные	195	149
Бракованные	5	2

Задача 2. В течение месяца завод поставил предприятию 200 корпусов, из которых 3 оказались дефектными. В следующий месяц было поставлено 850 корпусов, из которых 7 оказалось дефектными. Изменилась ли доля дефектных корпусов в поставках завода? Принять $\alpha = 0{,}01$.

Задача 3. 1000 человек классифицировали по признаку дальтонизма. По приведенным ниже данным проверить, есть ли зависимость между наличием дальтонизма и полом человека, при $\alpha = 0{,}05$.

	Мужчины	Женщины
Дальтоники	38	6
Не дальтоники	442	514

Задача 4. Во время эпидемии гриппа изучалась эффективность прививок против этого заболевания. Получены следующие результаты:

После прививки		Без прививки	
заболели	не заболели	заболели	не заболели
4	192	34	111

Указывают ли эти результаты на эффективность прививок? Принять $\alpha = 0{,}01$.

4.2. Статистика χ^2 для сравнения наблюдаемых и ожидаемых частот (Observed versus expected Xi)

Процедура **Observed versus expected Xi** (см. рис. 4.1) использует статистику χ^2 для проверки согласия наблюдаемых и ожидаемых частот. Ожидаемые частоты могут вычисляться на основе теоретической модели некоторого предполагаемого закона распределения случайной величины (вспомните критерий χ^2 для проверки гипотезы о виде распределения, см. главу 3, п. 3.3.4). Выборочное значение статистики $\chi^2_в$ вычисляется по формуле

$$\chi^2_в = \sum_{k=1}^{r} \frac{(n_k - n'_k)^2}{n'_k},$$

где n_k — наблюдаемые частоты; n'_k — ожидаемые частоты, $k = 1, 2, ..., r$.

Если теоретическая модель верна, то при $r \to \infty$ выборочная статистика $\chi^2_в$ имеет распределение χ^2 с числом степеней свободы $(r - l - 1)$, где l — число параметров распределения, оцениваемых по выборке, либо число связей, налагаемых моделью. Гипотеза о согласии наблюдаемых и ожидаемых частот принимается, если

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

где $\chi^2_{1-\alpha}(r - l - 1)$ — квантиль распределения $\chi^2(r - l - 1)$ порядка $1 - \alpha$; α — заданный уровень значимости; в противном случае гипотеза о согласии частот отклоняется.

Пример 4.3. Метод получения случайных чисел был применен 250 раз, при этом получены следующие результаты:

Число	0	1	2	3	4	5	6	7	8	9
Частота появления, n_k	27	18	23	31	21	23	28	25	22	32

Можно ли считать, что числа 0—9 появляются с одной и той же вероятностью? Принять $\alpha = 0{,}10$.

Решение. Если предположение верно, то частоты появления цифр должны быть равны и составлять $n'_k = 25$, $k = 1, 2, ..., 10$. Вычисленное значение статистики $\chi^2_в$ равно

$$\chi^2_в = 7{,}2.$$

Так как квантиль $\chi^2_{0,9}(9) = 14{,}7$, что превышает $\chi^2_в$, то гипотеза о согласии наблюдаемых и ожидаемых частот принимается на уровне значимости $\alpha = 0{,}10$: следует считать, что метод действительно дает числа, появляющиеся с одной и той же частотой.

Такой же результат получим, введя данные в модуль **Nonpametrics/Distrib** пакета STATISTICA (рис. 4.3*а*).

Observed vs. Expected Frequencies ...

Continue..

Chi-Square = 7,200000 df = 9 p < ,616308

Case	observed VAR1	expected VAR2	O - E	(O-E)**2 /E
C: 1	27,0000	25,0000	2,00000	,160000
C: 2	18,0000	25,0000	-7,00000	1,960000
C: 3	23,0000	25,0000	-2,00000	,160000
C: 4	31,0000	25,0000	6,00000	1,440000
C: 5	21,0000	25,0000	-4,00000	,640000
C: 6	23,0000	25,0000	-2,00000	,160000
C: 7	28,0000	25,0000	3,00000	,360000
C: 8	25,0000	25,0000	0,00000	0,000000
C: 9	22,0000	25,0000	-3,00000	,360000
C: 10	32,0000	25,0000	7,00000	1,960000
Sum	250,0000	250,0000	0,00000	7,200000

Рис. 4.3*а*. Решение примера 4.3

Вычисленный уровень значимости $p = P[\chi^2(9) > 7{,}2] = 0{,}616307$, что больше, чем заданный уровень значимости $\alpha = 0{,}10$, следовательно, гипотеза о согласии наблюдаемых и ожидаемых частот принимается.

4.2.1. Задачи

Задача 1. Числа выпадений герба при 20 подбрасываниях двух монет распределились следующим образом:

Количество гербов	0	1	2
Число подбрасываний	4	8	8

Согласуются ли эти результаты с предположениями о симметричности монет и независимости результатов подбрасываний? Принять $\alpha = 0{,}05$.

Задача 2. Ниже приводятся данные о фактических объемах сбыта продукции (в условных единицах) в пяти районах:

Район	1	2	3	4	5
Фактический объем сбыта	110	130	70	90	100

Согласуются ли эти результаты с предположением о том, что сбыт продукции в этих районах должен быть одинаковым? Принять $\alpha = 0{,}01$.

Задача 3. На экзамене студент отвечает только на один вопрос по одной из трех частей курса. Анализ вопросов, заданных 60 студентам, показал, что 23 студента получили вопросы из первой, 15 — из второй и 22 — из третьей части курса. Можно ли считать, что студент с равной вероятностью получит вопрос по любой из трех частей курса? Принять $\alpha = 0{,}10$.

Задача 4. В цехе с 10 станками ежедневно регистрировалось число вышедших из строя станков. Всего было проведено 200 наблюдений, результаты которых приведены ниже:

Число выбывших станков	0	1	2	3	4	5	6	7	8	9	10
Число зарегистрированных случаев	41	62	45	22	16	8	4	2	0	0	0

Проверить гипотезу H_0 о том, что число выбывших из строя станков имеет распределение Пуассона. Принять $\alpha = 0{,}05$.

Задача 5. Во время второй мировой войны на Лондон упало 537 самолетов-снарядов. Вся территория Лондона была разделена на 576 участков площадью по 0,25 км^2. Ниже приведены числа участков n_k, на которые упало k снарядов:

k	0	1	2	3	4	5 и больше
n_k	229	211	93	35	7	1

Согласуются ли эти данные с гипотезой о том, что число снарядов, упавших на каждый из участков, имеет распределение Пуассона? Принять $\alpha = 0{,}05$.

Задача 6. Ниже приводятся данные о числе деталей, поступающих на конвейер в течение 600 двухминутных интервалов:

Число деталей	0	1	2	3	4	5	6
Число интервалов	400	167	29	3	0	0	1

Используя критерий χ^2, проверить гипотезу H_0 о пуассоновском распределении числа деталей при $\alpha = 0{,}05$.

В следующих заданиях при $\alpha = 0{,}10$ необходимо проверить гипотезу H_0 о том, что выборки получены из нормально распределенной генеральной совокупности.

Задача 7. Рост 1004 девушек в возрасте 16 лет (см):

Границы интервала	134—137	137—140	140—143	143—146	146—149	149—152	152—155
Частота	1	4	16	53	121	197	229

Границы интервала	155—158	158—161	161—164	164—167	167—170	170—173
Частота	186	121	53	17	5	1

Задача 8. 200 отклонений размера вала от номинального значения (мкм):

Середина интервала	–0,14	–0,12	–0,10	–0,08	–0,06	–0,04	–0,02
Частота	3	8	11	20	27	36	29

Середина интервала	0,00	0,02	0,04	0,06	0,08	0,10	0,12
Частота	18	17	17	8	4	1	1

Задача 9. Величина контрольного размера 68 деталей, изготовленных на одном станке (мм):

Границы интервала	2,9—3,9	3,9—4,9	4,9—5,9	5,9—6,9	6,9—7,9
Частота	5	15	23	19	6

4.3. Коэффициенты ранговой корреляции Спирмена и τ Кендалла (Correlations Spearman, Kendall tau)

В этой опции вычисляются непараметрические меры взаимозависимости между двумя случайными переменными, измеренными в порядковой шкале. Коэффициенты ранговой корреляции Спирмана и τ Кендалла можно применить и к данным, измеренным в количественных шкалах, наряду с коэффициентом корреляции Пирсона (см. ниже, глава 6, п. 6.1). Коэффициенты ранговой корреляции (как и большинство непараметрических оценок) менее чувствительны к выбросам и погрешностям в результатах наблюдений и, в этом смысле, являются более устойчивыми и надежными мерами взаимозависимости по сравнению с коэффициентом корреляции Пирсона.

Коэффициент ранговой корреляции Спирмена

Пусть (x_i, y_i), $i = 1, 2, ..., n$ — выборка наблюдений двух переменных X и Y, измеренных в порядковой или количественной шкалах. Предположим, что среди элементов выборки x_i и y_i, $i = 1, 2, ..., n$ нет совпадающих элементов (случай с совпадающими элементами рассматривается ниже). Упорядочим элементы x_i по возрастанию (т. е. запишем вариационный ряд $x^{(1)}, x^{(2)}, ..., x^{(n)}$) и каждому x_i поставим в соответствие ранг x_i^I — номер элемента x_i в вариационном ряду. Очевидно, наименьший элемент выборки $x^{(1)}$ будет иметь ранг 1, а наибольший элемент $x^{(n)}$ — ранг n. Аналогичным образом определим ранги y_i^I элементов y_i, $i = 1, 2, ..., n$. Каждой паре

(x_i, y_i) соответствует пара рангов (x_i', y_i'). Коэффициент ранговой корреляции Спирмена вычисляется по формуле

$$r_s = 1 - \frac{6\sum(x_i' - y_i')^2}{n(n^2 - 1)}. \qquad (1)$$

Полученное значение r_s называется *выборочным коэффициентом ранговой корреляции Спирмена* ρ_s.

Коэффициент ρ_s по модулю не превосходит единицу: $|\rho_s| \le 1$. Большие (по модулю) значения выборочного коэффициента r_s показывают, что между случайными величинами X и Y есть зависимость (в этом случае говорят, что коэффициент ранговой корреляции ρ_s — значим).

Зная выборочное значение коэффициента ранговой корреляции r_s, можно проверить гипотезу о незначимости ρ_s $H_0 : \rho_s = 0$, наблюдаемые случайные величины X и Y некоррелированны. Для этого используются таблицы критических значений $\rho_s : \rho(\alpha, n)$, где α — заданный уровень значимости, n — объем выборки (см. [10], табл. П3.10, с. 524). Если $|r_s| < \rho(\alpha, n)$, то гипотеза $H_0 : \rho_s = 0$ принимается на уровне значимости 2α при альтернативной гипотезе $H_1 : \rho_s \ne 0$.

Если $r_s > 0$ и $\rho_s < \rho(\alpha, n)$, то гипотеза $H_0 : \rho_s = 0$ принимается на уровне значимости α при альтернативной гипотезе $H_1 : \rho_s > 0$.

Пример 4.4. Вычислить коэффициент ранговой корреляции для следующей выборки

x	68,8	63,3	75,5	67,2	71,3	72,8	76,5	63,5	69,9	71,4
y	167	113,3	159,9	153,6	150,8	181,2	173,1	115,4	125,6	166,2

Проверить значимость ранговой корреляции при $\alpha = 0{,}10$.

Решение. Определим ранги элементов исходной выборки. Предварительно перепишем исходную выборку, упорядочив ее элементы по верхней строке (т. е. по значению x_i), в результате получим

x	63,3	63,5	67,2	68,8	69,9	71,3	71,4	72,8	75,5	76,5
y	113,3	115,4	153,6	167	125,6	150,8	166,2	181,2	159,9	173,1

Определим ранги для значения y_i. Вариационный ряд для y_i имеет вид:

i	1	2	3	4	5	6	7	8	9	10
$y^{(i)}$	113,3	115,4	125,6	150,8	153,6	159,9	166,2	167	173,1	181,2

Таким образом, упорядоченной по элементам x_i выборке соответствует следующая последовательность пар рангов и их разностей:

x_i'	1	2	3	4	5	6	7	8	9	10
y_i'	1	2	5	8	3	4	7	10	6	9
$x_i' - y_i'$	0	0	–2	–4	2	2	0	–2	3	1

Выборочное значение коэффициента ранговой корреляции Спирмена r по формуле (1), равна

$$r_s \approx 0{,}745.$$

Для данного примера $n = 10$, $\alpha = 0{,}10$. Чтобы проверить гипотезу H_0: $\rho_s = 0$ при альтернативной гипотезе H_1: $\rho_s \neq 0$, по таблице ([10], табл. П3.10, с. 524) найдем $\rho(0{,}05;\ 10) = 0{,}564$.

Так как $r_s > 0{,}564$, то ранговая корреляция значима.

Коэффициент ранговой корреляции τ Кендалла

Вычисляется по формуле

$$\tau = 1 - \frac{4k}{n(n-1)},$$

где k — число инверсий в ряду рангов второй переменной (y_i') (при условии, что ранги первой переменной (x_i') упорядочены).

В примере 4.4 последовательности рангов следующие:

x_i'	1	2	3	4	5	6	7	8	9	10
y_i'	1	2	5	8	3	4	7	10	6	9

Найдем число инверсий (нарушений порядка) в последовательности y_i', $i = 1, 2, \ldots, 10$.

Числа 1 и 2 инверсий не образуют; число 5 образует две инверсии, так как стоит перед числами 3 и 4; 8 — образует четыре инверсии с числами 3, 4, 7 и 6; 7 — образует одну инверсию; 10 — две. Таким образом число инверсий $k = 9$

$$\tau = 1 - \frac{4 \cdot 9}{10(10-1)} = 0{,}6.$$

Для проверки значимости коэффициента ранговой корреляции Кендалла можно воспользоваться таблицей критических значений $\tau(\alpha, n)$ (см. [10], табл. П3.9, с. 522). Для двусторонней альтернативы H_1: $\tau \neq 0$, критическое значение $\tau(0{,}05;\ 10) = 0{,}422$. Так выборочное значение $\tau = 0{,}6$, то на уровне значимости $\alpha = 0{,}10$, гипотезу H_0: $\tau = 0$ следует отклонить — ранговая корреляция значима.

Для проверки значимости τ при больших объемах выборки используется статистика

$$Z = \sqrt{\frac{9n(n-1)}{2(2n+5)}}\,\tau.$$

При больших значениях n статистика Z имеет (приближенно) стандартное нормальное распределение $N(0, 1)$.

Для данного примера выборочное значение Z равно

$$z_{\text{в}} = \sqrt{\frac{9 \cdot 10 \cdot 9}{2(2 \cdot 10 + 5)}} \cdot 0{,}6 \approx 2{,}4149.$$

Так как квантиль распределения $N(0, 1)$: $u_{0,95} = 1{,}645$, что меньше $z_{\text{в}}$, то коэффициент ранговой корреляции τ значимо отличается от нуля.

В пакете STATISTICA коэффициенты ранговой корреляции вычисляются в процедуре **Correlations** (**Spearmen, Kendall tau, gamma**) (см. стартовую панель модуля рис. 4.1).

Уровень значимости p для коэффициента ранговой корреляции Кендалла τ вычисляется с помощью статистики $Z = \sqrt{\frac{9n(n-1)}{2(2n+5)}}\,\tau$:

$$p = P[|Z| > z_{\text{в}}].$$

В примере 4.4 выборочное значение статистики Z, $z_{\text{в}} \approx 2{,}4149$, и, следовательно, вычисленный уровень значимости p равен

$$p = P[|Z| > 2{,}4149] \approx 0{,}01574.$$

Так как это значение превышает заданный уровень значимости $\alpha = 0{,}01$, то коэффициент ранговой корреляции τ значимо отличен от нуля. Как и коэффициент ранговой корреляции Спирмена ρ_s, коэффициент τ по модулю не превосходит единицу: $|\tau| \le 1$. Значения ± 1 эти коэффициенты ранговой корреляции ρ_s и τ принимают в случае, когда последовательности рангов x_i', y_i', $i = 1, 2, \ldots, n$ совпадают, либо расположены во взаимно обратном порядке.

Если два или более элементов вариационного ряда совпадают, то этим элементам присваивается один и тот же ранг, равный среднему арифметическому их номеров. Например, вариационному ряду: 0, 1, 2, 2, 2, 4, 8 будет соответствовать следующая последовательность рангов: 1, 2, 4, 4, 4, 6, 7, так как третьему, четвертому и пятому элементам вариационного ряда (они совпадают и равны 2) присваивается ранг $\frac{3+4+5}{3} = 4$.

В случае совпадающих рангов для расчета ранговых коэффициентов корреляции ρ_s и τ используют скорректированные формулы. Выборочное значение коэффициента ранговой корреляции Спирмена r_s вычисляется по следующей формуле

$$r_s = \frac{\frac{1}{6}(n^3 - n) - \sum (x_i' - y_i')^2 - T_x - T_y}{\sqrt{\left[\frac{1}{6}(n^3 - n) - 2T_x\right]\left[\frac{1}{6}(n^3 - n) - 2T_y\right]}},$$

где $T_x = \frac{1}{12}\sum_{t=1}^{m_x}[(n_t)^3 - n_t]$, $T_y = \frac{1}{12}\sum_{l=1}^{m_y}[(n_l)^3 - n_l]$.

Здесь m_x — число групп совпадающих рангов в последовательности рангов x_i', n_t — число совпадающих рангов в группе с номером t, $t = 1, 2, \ldots, m_x$.

Аналогично, m_y — число групп совпадающих рангов в последовательности y_i^I, n_l — число совпадающих рангов в группе с номером l, $l = 1, 2, ..., m_y$.

Скорректированная формула для вычисления коэффициента ранговой корреляции Кендалла имеет вид

$$\tau^I = \frac{\tau - \dfrac{2(U_x + U_y)}{n(n-1)}}{\sqrt{\left(1 - \dfrac{2U_x}{n(n-1)}\right)\left(1 - \dfrac{2U_y}{n(n-1)}\right)}},$$

где τ — значение коэффициента ранговой корреляции Кендалла, вычисленное без поправки; $U_x = \frac{1}{2}\sum_{t=1}^{m_x} n_t(n_t - 1)$, $U_y = \frac{1}{2}\sum_{l=1}^{m_y} n_l(n_l - 1)$, где значения n_t и n_l были определены выше.

Пример 4.5. Объемы продаж в двух магазинах бытовой техники в течение 10 дней составили (в тыс. руб.)

x	19	15	17	18	17	18	21	21	15	13
y	19	17	17	17	17	19	20	19	15	14

Определить коэффициенты ранговой корреляции.

Решение. Определим ранги исходной выборки. Предварительно упорядочим элементы выборки по элементам первой строки (x):

i	1	2	3	4	5	6	7	8	9	10
x	13	15	15	17	17	18	18	19	21	21
y	14	15	17	17	17	17	19	19	19	20

Вторая строка (y) также записана в порядке возрастания. Поэтому можно сразу записать последовательность пар рангов, присваивая повторяющимся элементам равные ранги по правилу среднего арифметического:

x^I	1	2,5	2,5	4,5	4,5	6,5	6,5	8	9,5	9,5
y^I	1	2	4,5	4,5	4,5	4,5	8	8	8	10
$x^I - y^I$	0	0,5	–2	0	0	2	–1,5	0	1,5	–0,5

$$\sum (x^I - y^I)^2 = 0{,}25 + 4 + 4 + 2{,}25 + 2{,}25 + 0{,}25 = 13, \quad n = 10,$$

$$T_x = \frac{1}{12}[(2^3 - 2) + (2^3 - 2) + (2^3 - 2) + (2^3 - 2)] = \frac{24}{12} = 2,$$

$$T_y = \frac{1}{12}[(4^3 - 4) + (3^3 - 3)] = \frac{84}{12} = 7,$$

$$r_s = \frac{1/6\,(10^3 - 10) - 13 - 2 - 7}{\sqrt{[1/6\,(10^3 - 10) - 2 \cdot 2][1/6\,(10^3 - 10) - 2 \cdot 7]}} \approx 0{,}917.$$

Далее вычислим коэффициент τ. Так как в упорядоченной по x^l последовательности пар, во второй строке (y^l) инверсий нет, то $k = 0$, и коэффициент ранговой корреляции Кенделла равен

$$\tau = 1 - \frac{4k}{n(n-1)} = 1.$$

Чтобы определить скорректированное значение τ^l, предварительно вычислим

$$U_x = \frac{1}{2}[2 \cdot 1 + 2 \cdot 1 + 2 \cdot 1 + 2 \cdot 1] = 4; \; U_y = \frac{1}{2}[4 \cdot 3 + 3 \cdot 2] = 9.$$

Таким образом τ^l равно

$$\tau^l = \frac{1 - \frac{2(4+9)}{10 \cdot 9}}{\sqrt{\left(1 - \frac{2 \cdot 4}{10 \cdot 9}\right)\left(1 - \frac{2 \cdot 9}{10 \cdot 9}\right)}} \approx 0{,}833.$$

Оба коэффициента ранговой корреляции ρ_s и τ — значимы на уровне значимости $\alpha = 0{,}10$, так как их выборочные значения превышают критические значения 0,564 и 0,422 (см. пример 4.4).

4.3.1. Задачи

В следующих задачах вычислите коэффициенты ранговой корреляции Спирмена и τ Кендалла. Проверьте значимость полученных результатов, сравните коэффициенты ранговой корреляции и прокомментируйте их.

Задача 1. Спортсмены, ранги которых при построении по росту были 1, 2, ..., 10, заняли на состязаниях следующие места:

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

Как велика ранговая корреляция между ростом и быстротой бега?

Задача 2. Цветные диски, имеющие порядок оттенков 1, 2, ..., 15, были расположены испытуемым в следующем порядке:

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

Охарактеризовать способность испытуемого различать оттенки цветов с помощью коэффициентов ранговой корреляции между действительными и наблюдаемыми результатами.

Задача 3. Найти коэффициент ранговой корреляции между урожайностью пшеницы и картофеля на соседних полях по следующим данным:

Годы	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
Пшеница, (ц)	20,1	23,6	26,3	19,9	16,7	23,2	31,4	33,5	28,2	35,3	29,3	30,5
Картофель, (ц)	7,2	7,1	7,4	6,1	6,0	7,3	9,4	9,2	8,8	10,4	8,0	9,7

Задача 4. Для контрольной партии интегральных схем по нескольким параметрам определено значение критерия годности K. Найти коэффициенты ранговой корреляции между значениями K и удельного сопротивления p-кармана R_p, а также между значениями R_p и напряжением отсечки V_0 по следующим данным:

K	0,226	0,187	0,678	0,141	0,197	0,339	0,421	0,141	0,127	0,819
R_p, $\frac{\text{Ом} \cdot \text{мм}^2}{\text{м}}$	905	1004	1119	1200	1340	1261	1140	1190	1060	1130
V_0, в	1,2	1,9	1,7	1,5	4,5	2,2	2,3	2,4	1,8	1,4

Проверить значимость полученных коэффициентов при $\alpha = 0{,}10$.

Задача 5. Измерения длины головы (x) и длины грудного плавника (y) у 16 окуней дали результаты (в мм):

x	66	61	67	73	51	59	48	47	58	44	41	54	52	47	51	45
y	38	31	36	43	29	33	28	25	36	26	21	30	20	27	28	26

а. Найти коэффициенты ранговой корреляции. Проверить значимость полученного результата при $\alpha = 0{,}05$.

б. Найти коэффициент корреляции Пирсона и проверить его значимость при $\alpha = 0{,}05$ в предположении, что выборка наблюдений получена из нормально распределенной двумерной совокупности.

Задача 6. Связь между массой тела (x) и количеством гемоглобина в крови (y) у павианов-гамадрилов характеризуется следующими данными:

Масса тела, кг	18	17,7	19	18	19	22	21	21	20	30
Гемоглобин (по Сали)	70	74	72	80	77	80	80	89	76	86

а. Найти коэффициенты ранговой корреляции и проверить их значимость при $\alpha = 0{,}05$.

б. Найти коэффициент корреляции Пирсона.

4.4. Критерий серий Вальда—Вольфовица (Wald—Wolfowitz runs test)

Критерий серий применяется для проверки гипотезы H_0, утверждающей, что две группы данных представляют случайные независимые выборки с объемами n_1 и n_2 из одной генеральной совокупности, т. е. не отличаются друг от друга по наблюдаемому признаку.

Результаты наблюдений записываются в виде вариационного ряда объединенной выборки, а принадлежность данных к той или иной группе определяется с помощью кодирующей переменной, принимающей два значения (0 и 1, + и –, 1 и 2 и так далее). Полученную таким образом последовательность назовем *последовательностью кодов.*

Серией в последовательности кодов называется всякая подпоследовательность, состоящая из одинаковых кодов и ограниченная противоположными кодами, либо находящаяся в начале или конце исходной последовательности. Например, в последовательности кодов: 0 1 0 0 0 1 1 1 1 1 0 0 имеется пять серий: (0), (1), (0 0 0), (1 1 1 1 1), (0 0).

Статистикой критерия является число серий N в последовательности кодов. Если гипотеза H_0 верна, то обе выборки должны быть хорошо перемешаны в общем вариационном ряду и число серий N должно быть велико. Если же выборки получены из генеральных совокупностей с разными распределениями, различающимися средними значениями или разбросом, то число серий N, по-видимому, будет мало.

Критическая область определяется неравенствами: $N \le N_1$ и $N \ge N_2$, где значения N_1 и N_2 определяются по объему выборок n_1 и n_2 и уровню значимости α (см., например, [1] табл. П11, $\alpha = 0{,}05$).

При больших объемах выборок ($n_1 > 20$ и/или $n_2 > 20$) для проверки гипотезы H_0 можно использовать статистику Z:

$$Z = \frac{\left| N - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1 \right) \right| - \frac{1}{2}}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}}.$$

Если гипотеза H_0 верна, то Z имеет (приближенно) стандартное нормальное распределение $N(0, 1)$. Гипотеза H_0 принимается на уровне значимости α, если выборочное значение статистики Z, $z_в$ удовлетворяют неравенству: $|z_в| \le u_{1-\frac{\alpha}{2}}$, где $u_{1-\frac{\alpha}{2}}$ — квантиль нормального распределения $N(0, 1)$ порядка $1 - \frac{\alpha}{2}$; если $|z_в| > u_{1-\frac{\alpha}{2}}$, то гипотеза H_0 отклоняется.

Пример 4.6. При изучении иностранного языка в двух группах студентов использовались две различные методики. После изучения части курса студенты обеих групп написали диктант. Количество ошибок в диктанте таково:

1 группа: 31, 26, 33, 11, 13, 5, 18, 1, 2, 16, 17, 23, 20, 21, 9;

2 группа: 12, 7, 4, 8, 3, 6, 10, 25, 22, 24, 15, 19, 14, 36, 34, 32, 27, 29, 30, 35, 28.

Можно ли считать, что применение разных методик не приводит к существенному различию в результатах диктанта? Принять $\alpha = 0{,}01$.

Решение. Проверяемая гипотеза H_0: обе выборки получены из одной генеральной совокупности. Альтернативная гипотеза H_1: выборки получены из разных генеральных совокупностей, т. е. разные методики приводят к различным успехам в изучении языка.

Для проверки гипотезы используем критерий серий. Присвоим элементам первой группы код 1, а элементам второй группы код 0. Объединим выборки, запишем вариационный ряд и составим последовательность кодов:

1 1 0 0 1 0 0 0 1 0 1 0 1 0 0 1 1 1 0 1 1 0 1 0 0 1 0 0 0 0 1 0 1 0 0 0

Число серий в последовательности кодов $N = 22$.

Первая группа состоит из $n_1 = 15$ элементов, а вторая — из $n_2 = 21$ элемента. Для проверки гипотезы H_0 используем статистику Z. Выборочное значение Z вычисляется по приведенной выше формуле и равно

$$z_в \approx 1{,}044.$$

Так как это значение меньше квантили распределения $N(0, 1)$ $u_{0,995} = 2{,}576$, то гипотеза H_0 не отклоняется, т. е. различные методики обучения не повлияли на результаты диктанта.

Чтобы решить задачу в пакете STATISTICA надо записать данные в две переменные. В одну переменную **(dependent variables)** надо последовательно занести обе выборки, а в другую **(grouping variables)** — коды, определяющие принадлежность элементов к той или иной выборке (рис. 4.4).

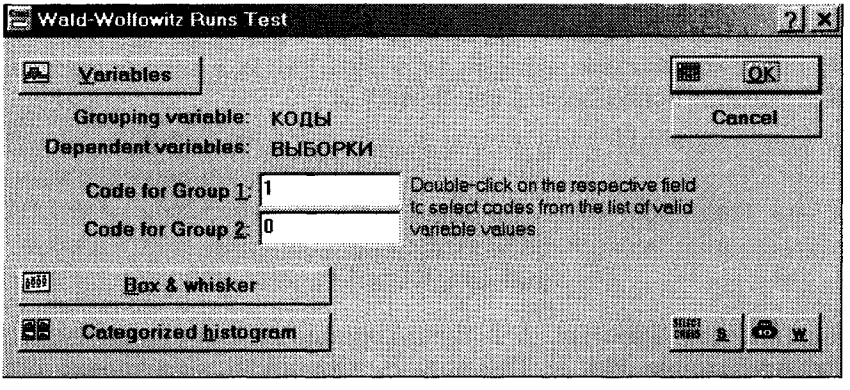


Рис. 4.4. Критерий серий: ввод данных

В результате выполнения процедуры получим:

No. *of runs* (число серий) = 22,

No. *of ties* (число совпадающих значений) = 0,

Z adjstd (скорректированное на непрерывность выборочное значение Z) = 1,044466

p-level

$$(p = P[|Z| > z_в]) = 0{,}296278.$$

В пакете STATISTICA решение выглядит следующим образом (рис. 4.5). Таким образом, гипотеза H_0 не отклоняется.

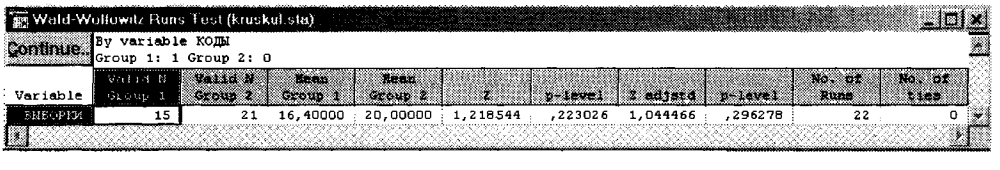


Рис. 4.5. Решение примера 4.6

4.5. Критерий Манна—Уитни (Mann—Whitney U test)

Критерий применяется для сравнения двух независимых выборок объемов n_1 и n_2 и проверяет гипотезу H_0, утверждающую, что выборки получены из однородных генеральных совокупностей и, в частности, имеют равные средние и медианы, т. е. применяется в тех же условиях, что и критерий серий.

Статистика W критерия определяется следующим образом. Расположим $n_1 + n_2$ значений объединенной выборки в порядке возрастания, т. е. в виде вариационного ряда. Каждому элементу ряда поставим в соответствие номер в ряду — *ранг*.

Если несколько элементов ряда совпадают по величине, то каждому из них присваивается ранг, равный среднему арифметическому их номеров. Последний элемент в ранжированной объединенной выборке должен иметь ранг $n_1 + n_2$. Этот факт можно использовать при проверке правильности ранжирования.

Пусть R_1 — сумма рангов первой выборки, R_2 — сумма рангов второй выборки. Вычислим значения w_1 и w_2, которые определяются формулами:

$$w_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1,$$

$$w_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2.$$

Правильность вычислений проверяется по формуле

$$w_1 + w_2 = n_1 n_2.$$

Выборочное значение $w_в$ статистики критерия W есть наименьшее из чисел w_1 и w_2.

В таблицах (см. [11], табл. H_1—H_4, с. 157—160) приводятся нижние и верхние критические значения для выборочных значений $w_в$ при $n_1, n_2 \le 20$ для различных уровней значимости α. Если выборочное значение статистики критерия не выходит за пределы, определенные критическими значениями, то гипотеза H_0 не противоречит результатам наблюдений.

Если объем каждой из выборок больше 8, то проверку гипотезы H_0 можно проводить, используя статистику

$$Z = \frac{W - \frac{1}{2} n_1 n_2}{\sqrt{\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)}},$$

имеющую (при условии, что верна гипотеза H_0) приблизительно стандартное нормальное распределение $N(0, 1)$. В этом случае гипотеза H_0 отклоняется на уровне значимости α, если выборочное значение $z_в$ статистики Z удовлетворяет неравенству

$$z_в < u_\alpha \ (z_в > u_{1-\alpha})$$

при левосторонней (правосторонней) альтернативной гипотезе H_1 и если

$$|z_в| > u_{1-\frac{\alpha}{2}}$$

при двусторонней альтернативной гипотезе H_1.

Пример 4.7. Измерялось напряжение пробоя у диодов, отобранных случайным образом из двух партий. Результаты измерения (в вольтах) следующие:

1-я партия	50	41	48	60	46	60	51	42	62	54	42	46
2-я партия	38	40	47	51	63	50	63	57	59	51	—	—

Имеются ли основания утверждать, что напряжение пробоя у диодов обеих партий равно? Решить пример, используя критерий Манна—Уитни. Принять $\alpha = 0{,}10$.

Решение. Составим вариационный ряд, отмечая принадлежность элемента к первой партии черточкой сверху. В результате получим следующую ранжированную последовательность:

Элемент	38	40	$\overline{41}$	$\overline{42}$	$\overline{42}$	$\overline{46}$	$\overline{46}$	47	$\overline{48}$	$\overline{50}$	50
Ранг	1	2	3	4,5	4,5	6,5	6,5	8	9	10,5	10,5

Элемент	$\overline{51}$	51	51	$\overline{54}$	57	59	$\overline{60}$	$\overline{60}$	$\overline{62}$	63	63
Ранг	13	13	13	15	16	17	18,5	18,5	20	21,5	21,5

Найдем суммы рангов для каждой партии

$$R_1 = 129{,}5, \quad R_2 = 123{,}5.$$

Так как $n_1 = 12$, $n_2 = 10$, то

$$w_1 = 12 \cdot 10 + \frac{12 \cdot (12+1)}{2} - 129{,}5 = 68{,}5,$$

$$w_2 = 12 \cdot 10 + \frac{10 \cdot (10+1)}{2} - 123{,}5 = 51{,}5.$$

Выборочное значение $w_в$ статистики критерия таково

$$w_в = 51{,}5.$$

Так как $n_1 > 8$, $n_2 > 8$, то для проверки гипотезы H_0 используем статистику Z. Выборочное значение этой статистики определяется по формуле

$$z_в = \frac{51{,}5 - \frac{1}{2} \cdot 12 \cdot 10}{\sqrt{\frac{1}{12} \cdot 12 \cdot 10(12+10+1)}} \approx -0{,}56.$$

Проверяемое предположение соответствует двусторонней альтернативной гипотезе, следовательно, значение $|z_в|$ сравнивается с квантилью стандартного нормального распределения $u_{1-\frac{\alpha}{2}}$, которая определяется по таблице

$$u_{1-\frac{\alpha}{2}} = u_{0,95} = 1{,}645.$$

Так как $|-0{,}56| < 1{,}645$, то гипотеза H_0 о равенстве напряжения пробоя у диодов обеих партий не отклоняется.

В пакете STATISTICA решение выглядит следующим образом (рис. 4.6).

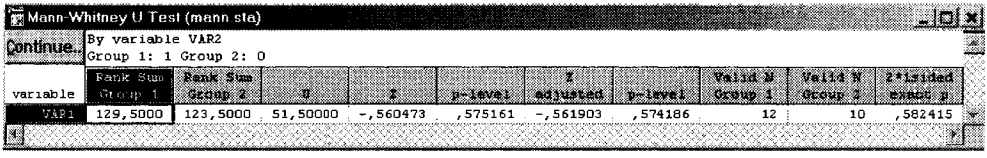


Рис. 4.6. Решение примера 4.6

Таким образом, утверждение о том, что напряжение пробоя у диодов обеих партий равно, следует принять.

4.5.1. Задачи

Решите следующие задачи, используя критерии серий Вальда—Вольфовица и Манна—Уитни. Сравните и прокомментируйте результаты.

Задача 1. По выборкам из двух партий микросхем после операции легирования поликремния измерялось удельное сопротивление. Результаты замеров следующие:

1-я партия	52,2	33	76	32,5	49,5	32,5	191,5	112,5
2-я партия	119	17,5	43,5	43,5	90,5	40	50	108

1-я партия	52,9	114,8	33,7	69,1	112,5	48,5	16,5
2-я партия	62,4	16,5	97,5	96	46	—	—

Можно ли утверждать, что обе партии получены из одной генеральной совокупности? Принять $\alpha = 0{,}10$.

Задача 2. В условиях предыдущей задачи после операции разгонки бора измерена глубина слоя диффузии и получены следующие результаты (мкм):

1-я партия	9,8	9,8	8,6	8,6	9,2	9,2	9,8
2-я партия	8,6	9,2	10,4	9	9,8	9,2	9,6

1-я партия	9	10	9,4	9	11,2	10,8	9,2	9,4
2-я партия	10	9,8	9,0	9,8	8,7	8,6	—	—

Можно ли считать, что глубина слоя диффузии в микросхемах из обеих партий различна? Принять $\alpha = 0{,}10$.

Задача 3. Для того чтобы повысить объем продаж фирма торгующая сыром через сеть специальных магазинов решила провести специальную рекламную акцию. Приведенные ниже данные отражают объем продаж по дням, во время которых рекламная акция проводилась (верхняя строка таблицы), и по дням, в которые она не проводилась (нижняя строка таблицы).

Объемы продаж (в сотнях)
18 21 23 15 19 26 17 18 22 20 18 21 27
22 17 15 23 25 20 26 24 16 17 23 21

Определите повлияла ли рекламная акция на повышение объема продаж, $\alpha = 0{,}05$.

Задача 4. Профессор Ньютон решил определить, быстрее или медленнее его наиболее способные студенты сдают письменные тесты: быстрее потому, что они быстрее вспоминают усвоенные навыки или медленнее потому, что на запись всего, что они знают уходит больше времени. В частности, при решении задач по физике он записал полученные студентами отметки в порядке сдачи их работ:

Порядок сдачи работ	Отметки									
1 — 10	94	70	85	89	92	98	63	88	74	85
11 — 20	69	90	57	86	79	72	80	93	66	74
21 — 30	50	55	47	59	68	63	89	51	90	88

а. Студентов, набравших 90 и более баллов профессор считает наиболее способными студентами. Может ли он при уровне значимости 5 % считать, что сдача работ этими студентами носила случайный характер?

б. Можно ли считать, что студенты, набравшие 60 или более баллов, которые считаются прошедшими тест, сдали свои работы в случайной последовательности в отличие от тех кто не прошел тест? Уровень значимости также 5 %.

Задача 5. В биохимическом исследовании, проведенном методом меченных атомов, по результатам изучения 8 препаратов контрольной серии получены следующие показания счетчика импульсов (в импульсах в минуту):

Опыт	340	343	322	349	332	320	313	304
Контроль	318	321	318	301	312	—	—	—

Можно ли считать, что полученные значения опытной и контрольной серий различны? Принять $\alpha = 0{,}10$.

Задача 6. Длина тела личинок щелкуна, обитающих в посевах озимой ржи и проса (выраженная в мм), варьируется следующим образом:

В посевах ржи	7	10	14	15	12	16	12
В посевах проса	11	12	16	13	18	15	—

На основании этих проб создается впечатление о более крупных размерах личинок щелкунов, обитающих на просе. Проверить это предположение. Принять $\alpha = 0{,}01$.

Задача 7. Изучалось влияние кобальта на увеличение массы кроликов. Опыт проводился на двух группах животных — опытной и контрольной. Возраст кроликов колебался в пределах от 1,5 до 2 месяцев. Исходная масса тела особей находилась в пределах от 500 до 600 г. Опыт длился 8 недель. Обе группы содержались на одном и том же кормовом рационе, но в отличие от контрольных, опытные кролики каждый день получали в виде водного раствора по 0,06 г хлористого кобальта на 1 кг массы тела. За время опыта у животных наблюдались следующие прибавки в массе (за 1 неделю):

Контрольные	560	580	600	420	530	490	580	470
Опытные	692	700	621	640	561	680	630	—

Можно ли считать, что добавки хлористого кобальта действительно дают прибавку массы тела? Принять $\alpha = 0{,}10$.

Задача 8. Двум группам испытуемых предлагалось провести опознание трех очертаний цифры 5. Результаты эксперимента (время опознания в секундах) следующие:

1-я группа	25	28	27	29	26	24	28	23	30	25	26	25
2-я группа	18	19	31	32	17	15	41	35	38	13	14	—

Можно ли считать, что результаты для первой и второй групп различны? Принять $\alpha = 0{,}05$.

Задача 9. Для контроля настройки двух станков-автоматов, производящих детали по одному чертежу, определили отклонения от номинальных размеров у нескольких деталей, изготовленных на обоих станках. В результате получили следующие данные (в мкм):

Станок А	44	–14	32	8	–50	20	–35	15	10	–8	–20	5
Станок В	52	–49	61	–35	–48	18	–45	35	23	21	–59	–19

Проверить гипотезу H_0 о том, что отклонения от номинальных размеров на обоих станках в среднем не отклоняется на уровне значимости $\alpha = 0{,}10$.

Используя критерий серий Вальда—Вольфовица решите следующие задачи.

Задача 10. Для 13 деталей получены следующие отклонения контрольного размера от номинального значения (в мкм):

8; 10; 5; –5; –9; 7; 6; –11; –4; –4; 15; 21; –3.

Можно ли считать, что полученная выборка представляет результаты случайных и независимых наблюдений? Принять $\alpha = 0{,}05$.

Задача 11. При подбрасывании монеты 45 раз последовательность результатов ($Г$ — выпадение герба, $Р$ — выпадение решки) имела следующий вид:

ГГГГГГРРРРГГГРРРГРГРРРРРР
ГГГГРРРРРГГГРГГГГРРР.

Является ли такая последовательность случайной выборкой? Принять $\alpha = 0{,}05$.

Задача 12. Глубина слоя диффузии, определенная по выборке из партии микросхем, имеет следующие значения (в мкм):

9,8; 9,8; 8,6; 9,2; 9,8; 9,0; 10,0; 9,4; 9,0; 11,2; 10,8; 9,2; 9,4.

Проверить гипотезу H_0 о том, что полученные результаты распределены случайным образом. Принять $\alpha = 0{,}05$.

Задача 13. Национальный банк отметил пол первых 40 клиентов, посетивших банк во вторник, в следующей последовательности: M — мужчины, F — женщины:

M, F, M, M, M, M, F, F, M, M, M, F, M, M, M, M, M, F, F, M,
F, M, M, M, F, M, M, M, M, M, M, F, M, M, M, M, M, F, F, M.

Уровень значимости $\alpha = 0{,}05$. Проверьте, случаен ли характер последовательности?

4.6. Двухвыборочный тест Колмогорова—Смирнова (Kolmogorov—Smirnov two-sample test)

Применяется для проверки гипотезы о том, что две независимые выборки $x_1, x_2, \ldots, x_{n_1}$ и $y_1, y_2, \ldots, y_{n_2}$ получены из одной генеральной совокупности, т. е. функции распределения $F_1(x)$ и $F_2(y)$ двух генеральных совокупностей равны (в этом случае говорят, что генеральные совокупности *однородны*):

$$H_0: F_1(x) \equiv F_2(y)\big|_{y=x},$$

при альтернативной гипотезе $H_1: F_1(x) \neq F_2(y)\big|_{y=x}$.

Статистикой критерия является максимальная разница между эмпирическими функциями распределения, построенными по выборкам

$$D = \max\left|F_1^*(x) - F_2^*(y)\right|.$$

Критические значения для статистики D приводятся в таблицах ([20]).

При больших значениях n_1 и n_2 (> 40) используются следующие критические значения:

при $\alpha = 0{,}05$ $D_{\text{крит}} = 1{,}36 \cdot \sqrt{\dfrac{n_1 n_2}{n_1 + n_2}}$;

при $\alpha = 0{,}10$, $D_{\text{крит}} = 1{,}22 \cdot \sqrt{\dfrac{n_1 n_2}{n_1 + n_2}}$.

Пример 4.8. По данным примера 4.7 проверить гипотезу о том, что обе выборки получены из одной генеральной совокупности.

Решение. Переключаемся в модуль **Nonpametrics/Distrib** ⇒ **Kolmogorov—Smirnov two-sample test**.

Ввод данных осуществляется так же как и в опцию **Mann-Whitney U test**: обе выборки вводятся в одну переменную (**dependent variables**), а коды вводятся в группирующую переменную (**grouping variables**). После выполнения процедуры получим значение $D = 0{,}216667$ (*max neg. differuc.* = 0,216667) и *p-level* > 0,10. Следовательно, на уровне значимости $\alpha = 0{,}10$ гипотеза H_0 не отклоняется: обе выборки получены из одной генеральной совокупности.

4.7. Однофакторный дисперсионный анализ Краскела—Уоллиса и медианный критерий (Kruskal—Wallis ANOVA and median test)

Критерий Краскела—Уоллиса служит для проверки гипотезы H_0 : k выборок объемов n_1, n_2, ..., n_k получены из одной генеральной совокупности, т. е. является обобщением U-критерия *Манна—Уитни* на случай, когда число выборок $k > 2$.

Статистика критерия H определяется следующим образом. Все выборки записываются в одну последовательность. Эта последовательность записывается в порядке возрастания, т. е. в виде вариационного ряда. Для каждого элемента выборки определяется ранг (так же как в U-критерии). Пусть R_i — сумма рангов i-й выборки, $i = 1, 2, ..., k$. Для контроля можно использовать тождество

$$\sum_{i=1}^{k} R_i \equiv \frac{n(n+1)}{n},$$

где n — число элементов объединенной выборки: $n = \sum_{i=1}^{k} n_i$.

Статистика критерия H определяется так:

$$H = \frac{12}{n(n+1)}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right) - 3(n+1).$$

Если гипотеза H_0 верна, то при $n_i \geq 5$ и $k \geq 4$ статистика H имеет приблизительно распределение χ^2 с $(k-1)$ степенями свободы. Гипотеза H_0 отклоняется на уровне значимости α, если выборочное значение $H_в$ статистики H удовлетворяет условию

$$H_в > \chi^2_{1-\alpha}(k-1),$$

где $\chi^2_{1-\alpha}(k-1)$ — квантиль распределения χ^2 порядка $(1-\alpha)$ с $(k-1)$ — степенью свободы.

Для $n_i \leq 8$, $k=3$; $n_i \leq 4$, $k=4$; $n_i \leq 3$, $k=5$; $n_i \leq 3$, $k=6$, $i=1, 2, ..., k$ имеются точные таблицы критических значений (см. [10], табл. П3.7, с. 515).

Пример 4.9. Ниже приводятся данные о содержании иммуноглобулина *IgA* в сыворотке крови (в мг %) у больных четырех возрастных групп:

Возрастная группа	Содержание *IgA* (мг %)										
1	83	85	82	82	84	—	—	—	—	—	—
2	84	85	85	86	86	87	—	—	—	—	—
3	86	87	87	87	88	88	88	88	88	89	90
4	89	90	90	91	91	—	—	—	—	—	—

Проверить гипотезу о том, что содержание иммуноглобулина у всех возрастных групп совпадает. Принять $\alpha = 0{,}01$.

Решение. Для проверки гипотезы H_0 воспользуемся критерием Краскела—Уоллиса. Суммы рангов по выборкам и объемы выборок равны:

$R_1 = 17{,}5$, $n_1 = 5$;

$R_2 = 52$, $n_2 = 6$;

$R_3 = 186$, $n_3 = 11$;

$R_4 = 122{,}5$, $n_4 = 5$.

$$n = \sum_{i=1}^{4} n_i = 27.$$

Выборочное значение статистики критерия H:

$$H_в = 21{,}99548.$$

Так как квантиль распределения χ^2:

$$\chi^2_{0{,}95}(4-1) = 7{,}81,$$

что меньше $H_в$, то гипотеза H_0 отклоняется на уровне значимости $\alpha = 0{,}05$: данные свидетельствуют о различном содержании иммуноглобулина в крови больных разных возрастных групп.

Этот же результат получим, введя данные в процедуру **Kruskal—Wallis ANOVA and median test**. Четыре выборки вводятся подряд в одну переменную (**dependent var**), а коды выборок (1, 2, 3, 4) вводятся в группирующую переменную (**grouping var**).

После выполнения процедуры получим значение

$$H = 21{,}99548$$

и значение

$$p = P[\chi^2(3) > H_в] = 0{,}0001,$$

что меньше, чем заданный уровень значимости $\alpha = 0{,}10$, следовательно, гипотеза H_0 отклоняется.

В пакете STATISTICA решение выглядит следующим образом (рис. 4.7).

Kruskal-Wallis ANOVA by Ranks (new.sta)

NONPAR STATS: Independent (grouping) variable: VAR2
Kruskal-Wallis test: H (3, N= 27) = 21,99548 p =,0001

Depend.: VAR1	Code	Valid N	Sum of Ranks
Group 1	1	5	17,5000
Group 2	2	6	52,0000
Group 3	3	11	186,0000
Group 4	4	5	122,5000

Рис. 4.7. Решение примера 4.9

Медианный критерий используется для проверки нулевой гипотезы о том, что все k выборок получены из генеральных совокупностей, имеющих равные медианы. Процедура применения критерия состоит в следующем. Все выборки объединяются в одну выборку объема $n = n_1 + n_2 + \ldots + n_k$. Эта выборка записывается в виде вариационного ряда и определяется общая медиана: если ряд содержит нечетное число элементов, то медиана равна среднему члену вариационного ряда; если ряд содержит четное число элементов, то медиана равна среднему арифметическому двух средних элементов. Далее для каждой выборки определяется число элементов, лежащих ниже или совпадающих с медианой, и число элементов, лежащих выше медианы. Результаты (частоты) заносятся в таблицу сопряженности $2 \times k$.

Для проверки гипотезы H_0: все k генеральных совокупностей имеют равные медианы, можно использовать статистику χ^2:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} = \sum \frac{f_0^2}{f_e} - n,$$

где f_0 — наблюдаемые частоты, f_e — ожидаемые частоты при условии, что гипотеза H_0 верна.

Если гипотеза H_0 верна, статистика χ^2 имеет распределение хи-квадрат с $(2-1)(k-1) = k-1$ числом степеней свободы. Гипотеза H_0 отклоняется если

$$\chi^2_в > \chi^2_{1-\alpha}(k-1),$$

где $\chi^2_в$ — выборочное значение статистики χ^2, а $\chi^2_{1-\alpha}(k-1)$ — квантиль распределения $\chi^2(k-1)$ порядка $(1-\alpha)$.

Пример 4.10. Успеваемость студентов четырех групп оценивается по 100-балльной шкале. Оценки студентов приведены ниже. Можно ли считать, что медианы оценок студентов по группам действительно различны? Принять $\alpha = 0{,}05$.

Группа			
1	2	3	4
77	44	26	7
31	78	70	28
59	38	55	19
48	20	61	39
40	25	73	55
59	29	61	36
57	51	63	19
22	74	79	11
13	54	50	9
16	56	45	80
22	47	33	9
25	40	45	10

Решение. Для решения примера используем медианный критерий.

Результаты применения медианного критерия следующие: общая медиана = 42, таблица сопряженности 2 × 4 для наблюдаемых частот f_0 имеет вид:

	Группа 1	Группа 2	Группа 3	Группа 4	Всего
≤ медианы	$a = 7$	$b = 5$	$c = 2$	$d = 10$	$a + b + c + d = 24$
> медианы	$e = 5$	$f = 7$	$g = 10$	$k = 2$	$e + f + g + k = 24$
	$a + e = 12$	$b + f = 12$	$c + g = 12$	$d + k = 12$	$n = 48$

Ожидаемые частоты f_e определяются следующим образом:

$$\text{клетка } a = \frac{(a + b + c + d)(a + e)}{n} = 6;$$

$$\text{клетка } b = \frac{(a + b + c + d)(b + f)}{n} = 6;$$

$$\text{клетка } c = \frac{(a + b + c + d)(c + g)}{n} = 6;$$

$$\text{клетка } d = \frac{(a + b + c + d)(d + k)}{n} = 6;$$

$$\text{клетка } e = \frac{(e+f+g+k)(a+e)}{n} = 6;$$

$$\text{клетка } f = \frac{(e+f+g+k)(b+f)}{n} = 6;$$

$$\text{клетка } g = \frac{(e+f+g+k)(c+g)}{n} = 6;$$

$$\text{клетка } k = \frac{(e+f+g+k)(d+k)}{n} = 6.$$

Выборочное значение статистики χ^2:

$$\chi^2_{\text{в}} = 11{,}33.$$

Квантиль $\chi^2_{1-\alpha}(k-1) = \chi^2_{0,95}(4-1) = 7{,}81$. Так как выборочное значение $\chi^2_{\text{в}} = 11{,}33$ превышает квантиль $\chi^2_{1-\alpha}(k-1) = 7{,}81$, то гипотезу H_0: медианы генеральных совокупностей равны следует отклонить. При решении задачи в пакете STATISTICA используется вычисленный уровень значимости p:

$$p = P[\chi^2(3) > 11{,}33] = 0{,}0101.$$

Так как это значение меньше, чем заданный уровень значимости $\alpha = 0{,}05$, то гипотеза H_0 отклоняется. Следует считать, что медианы оценок студентов по группам различны.

4.7.1. Задачи

1. Решите следующие задачи, используя однофакторный анализ Краскела—Уоллиса.

2. К этим же задачам примените медианный критерий: сформулируйте и проверьте соответствующие гипотезы. Сравните результаты с результатами в п. 1.

Задача 1. Три группы водителей обучались по различным методикам. После окончания срока обучения был произведен тестовый контроль над случайно отобранными водителями из каждой группы. Получены следующие результаты:

Номер группы, k	Число ошибок, допущенных водителями, x_{ik}	Сумма по каждой группе, x_k	Число контролируемых водителей, n_k
1	1 3 2 1 0 2 1	10	7
2	2 3 2 1 4 - -	12	5
3	4 5 3 - - - -	12	3

На уровне значимости $\alpha = 0{,}05$ проверить гипотезу об отсутствии влияния различных методик обучения на результаты тестового контроля водителей.

Задача 2. В таблице приведены розничные цены (в условных единицах) на три модели обуви. Определите наличие разницы в розничных ценах на эти модели. Уровень значимости $\alpha = 0{,}1$.

Модель *A*	89	90	92	81	76	88	85	95	97	86	100
Модель *B*	78	93	81	87	89	71	90	96	82	85	
Модель *C*	80	88	86	85	79	80	84	85	90	92	

Задача 3. Компания, специализирующаяся по доставке подарков почтой, располагает следующими данными по количеству полученных заказов по трем видам оплаты.

Заказ по кредитной карточке	78	64	75	45	82	69	60
Чеком	110	70	53	51	61	68	
Наличными	90	68	70	54	74	65	59

Можно ли считать, что число заказов не зависит от вида оплаты? $\alpha = 0{,}05$.

Задача 4. Фирма имеет три магазина. Фирма составляет ежедневный отчет по количеству покупателей посетивших каждый магазин и сделавших покупки. Ниже приводится выборка данных. Можно ли утверждать, что на уровне значимости $\alpha = 0{,}05$ в каждом из трех магазинов побывало одинаковое количество покупателей, сделавших покупки?

Магазин 1	99	64	101	85	79	88	97	95	90	100
Магазин 2	83	102	125	61	91	96	94	89	93	75
Магазин 3	89	98	56	105	87	90	87	101	76	89

Задача 5. Утомленная изучением статистики студентка Катя посетила несколько магазинов, чтобы определить, действительно ли цены на простоквашу значительно различаются в зависимости от сорта. Ее наблюдения приводятся ниже. Может ли Катя сделать вывод, что цены на простоквашу действительно зависят от сорта? $\alpha = 0{,}05$.

Цена (в условных единицах)			
Сорт *A*	Сорт *B*	Сорт *C*	Сорт *D*
61	52	47	67
55	58	52	63
57	54	49	68
60	55	49	69
58	57		65
62			

Задача 6. Для производителей новых препаратов по лечению нервных расстройств важно знать действие их препаратов на двигательные функции организма, в частности, на координацию движений. Проверено действие четырех препаратов. Испытуемым предлагались тесты на ловкость, и подсчитывалось количество сделанных ими ошибок. Результаты тестов приводятся ниже:

	Количество ошибок в движениях						
Препарат 1	245	258	239	241	235	242	
Препарат 2	277	276	269	274	270	275	
Препарат 3	215	232	225	247	226	230	222
Препарат 4	241	253	237	246	340	300	240

Различаются ли все четыре препарата по степени воздействия на координацию движений при $\alpha = 0{,}05$?

Задача 7. Компания хочет установить расценки на рекламные объявления на стендах в зависимости от их месторасположения. Определим «популярность» стендов как количество людей, которые рассматривают щит в течение пятиминутного интервала. В таблице указано количество людей, останавливающихся у щитов в течение нескольких пятиминутных интервалов:

Рекламный щит 1	30	45	26	44	18	38	42	29	
Рекламный щит 2	29	38	36	21	36	18	17	30	32
Рекламный щит 3	32	44	40	43	24	28	18		

Одинаковы ли «популярности» стендов? Принять $\alpha = 0{,}05$.

Задача 8. Инвестор хочет знать, существуют ли значительные различия в доходах от акций, облигаций и инвестиционных фондов. Он взял случайные выборки каждого способа вложения капиталов и получил следующие данные:

	Доходы (в %)						
Акции	2,0	6,0	2,0	2,1	6,2	2,9	3,0
Облигации	4,0	3,1	2,2	5,3	5,9	5,5	
Инвестиционные фонды	3,5	3,1	2,9	6,0	4,5	3,2	

а) определите основную H_0 и альтернативную H_1 гипотезы;

б) проверьте гипотезу H_0 при $\alpha = 0{,}05$;

в) сформулируйте окончательный вывод.

Задача 9. За период 1986—1988 в автомобильных катастрофах погибло более 75 тысяч человек. Прибегнув к этой мрачной статистике, страховой институт дорожной безопасности рассчитал уровень смертности для 103 самых широко используемых моделей. Автомобили были разбиты по категориям, как-то: микроавтобусы, 4-х дверные автомобили, 2-х дверные или спортивные автомобили, а также автомобили особого класса. Далее автомобили в каждой категории были разбиты по размеру: маленькие, средние и большие. Посмотрите на уровни смертности (количество смертей на 10 000 зарегистрированных автомобилей) для 4-х дверных автомобилей:

Большие	1,2	1,3	1,4	1,5	1,5	1,5	1,6	1,8		
Средние	1,1	1,2	1,2	1,2	1,3	1,3	1,3	1,3	1,4	1,4
	1,5	1,6	1,6	1,6	1,7	1,7	1,8	1,9	2,0	2,3
	2,3	2,4	2,5	2,6	2,9					
Малые	1,1	1,5	1,6	1,7	1,8	2,0	2,0	2,0	2,3	2,5
	2,6	2,8	3,2	4,1						

Проверьте гипотезу о равенстве уровней смертности для трех категорий при уровне значимости 5 %.

4.8. Критерий знаков (Sign test)

Критерий знаков применяется для проверки гипотезы H_0 об однородности генеральных совокупностей по попарно связанным выборкам. Такая задача возникает, например, при сравнении двух измерительных приборов. При этом используют n объектов и над каждым из них производят по одному измерению с помощью обоих приборов.

Обозначим x_i и y_i, $i = 1, 2, ..., n$, результаты измерения i-го объекта, полученные соответственно при помощи первого и второго приборов. Если сравниваемые выборки получены из однородных совокупностей, то значения x_i и y_i взаимозаменяемы, и, следовательно, вероятности появления положительных и отрицательных разностей $x_i - y_i$ равны. Вероятности появления нулевых разностей равны нулю в силу предполагаемой непрерывности распределения измеряемого признака. Таким образом, вероятности появления положительных и отрицательных разностей равны $\frac{1}{2}$, т. е.

$$P[x_i - y_i > 0] = P[x_i - y_i < 0] = \frac{1}{2},$$

$i = 1, 2, ..., l$, где l — число ненулевых разностей, $l \le n$.

Нулевые разности могут появиться из-за случайных погрешностей или ошибок округления, и соответствующие им пары наблюдений, исключаются из рассмотрения.

Статистикой критерия знаков является число знаков «+» или «–» в последовательности знаков разностей парных выборок (x_i, y_i), $i = 1, 2, ..., l$.

В дальнейшем, для определенности, берется число знаков «+». При условии, что знаки разностей $x_i - y_i$ *независимы*, число знаков «+» имеет *биномиальное распределение* с параметрами $p = \frac{1}{2}$ и l, т. е. $B(l, \frac{1}{2})$. Задача сводится к проверке гипотезы $H_0 : p = \frac{1}{2}$ при одной из альтернативных гипотез $H_1^{(1)}: p > \frac{1}{2}$, $H_1^{(2)}: p < \frac{1}{2}$, $H_1^{(3)}: p \neq \frac{1}{2}$.

Пусть r — наблюденное число знаков «+», а α — заданный уровень значимости. Гипотеза H_0 отклоняется, если при альтернативной гипотезе $H_1^{(1)}: p > \frac{1}{2}$ выполняется неравенство

$$\sum_{i=r}^{l} C_l^i \left(\frac{1}{2}\right)^l \leq \alpha, \tag{1}$$

где C_l^i — число сочетаний, $C_l^i = \frac{l!}{i!(l-i)!}$ или, при альтернативной гипотезе $H_1^{(2)}: p < \frac{1}{2}$ выполняется неравенство

$$\sum_{i=0}^{r} C_l^i \left(\frac{1}{2}\right)^l \leq \alpha, \tag{2}$$

или, наконец, при альтернативной гипотезе $H_1^{(3)}: p \neq \frac{1}{2}$ выполняется одно из неравенств

$$\begin{aligned} &\sum_{i=0}^{r} C_l^i \left(\frac{1}{2}\right)^l \leq \frac{\alpha}{2} \text{ или} \\ &\sum_{i=r}^{l} C_l^i \left(\frac{1}{2}\right)^l \leq \frac{\alpha}{2}. \end{aligned} \tag{3}$$

Если при соответствующих альтернативных гипотезах неравенства (1)—(3) не выполняются, то гипотеза H_0 не противоречит результатам наблюдений и принимается на уровне значимости α.

Вероятности в левых частях неравенств (1)—(3) вычисляются либо непосредственно, либо с помощью нормальной аппроксимации биномиального распределения (теорема Муавра—Лапласа, Приложение, П.4).

Обозначим число знаков «+» как значение случайной величины X. Если гипотеза $H_0 : p = \frac{1}{2}$ верна, то X имеет приближенно нормальное распределение с математическим ожиданием $lp = \frac{l}{2}$ и дисперсией $lpq = \frac{l}{4}$, где

$q = 1 - p = 1/2$ и вероятность события $[X \le r]$ вычисляется по приближенной формуле

$$P[X \le r] = \sum_{i=0}^{r} C_l^i \left(\frac{1}{2}\right)^l \approx \Phi\left[\frac{r - \frac{l}{2}}{\sqrt{\frac{l}{4}}}\right],$$

где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ — функция распределения стандартного нормального закона $N(1, 0)$.

При $l \le 50$ более точный результат получается, если использовать «поправку на непрерывность». Окончательная формула в этом случае имеет вид

$$P[X \le r] \approx \Phi\left[\frac{r - \frac{l}{2} + 0{,}5}{\sqrt{\frac{l}{4}}}\right].$$

При решении задач в пакете STATISTICA выводятся следующие результаты: число ненулевых разностей l, процент разностей со знаком «+»: $\frac{r}{l} \cdot 100\ \%$, модуль нормированной статистики критерия $Z = \frac{\left|r - \frac{l}{2}\right| + 0{,}5}{\sqrt{\frac{l}{4}}}$ и вычисленный уровень значимости (для двусторонней проверки)

$$p\text{-}level = P[|Z| > z_в],$$

где $z_в$ — выборочное значение статистики Z.

Пример 4.11. Предполагается, что один из двух приборов, определяющих скорость автомобиля, имеет систематическую ошибку. Для проверки этого предположения определили скорость десяти автомобилей, причем скорость каждого фиксировалась одновременно двумя приборами.

v_1, км/ч	70	85	63	54	65	80	75	95	52	55
v_2, км/ч	72	86	62	55	63	80	78	90	53	57

Позволяют ли эти результаты утверждать, что второй прибор дает завышенные значения скорости?

Принять $\alpha = 0{,}10$.

Решение. В предположении, что скорости движения автомобилей не зависят друг от друга, задачу можно решить, применяя критерий знаков.

Последовательность знаков разностей $v_1 - v_2$: –, –, + , –, + , 0, –, + , –, –. Число ненулевых разностей $l = 9$, а число положительных разностей $r = 3$. Выборочное значение статистики критерия $z_в$ равно:

$$z_в = \frac{\left|3 - \frac{9}{2}\right| + 0{,}5}{\sqrt{\frac{9}{4}}} = 0{,}666667.$$

Вычисленный уровень значимости $p = 0{,}504985$ достаточно высок и больше заданного уровня значимости $\alpha = 0{,}10$, следовательно, гипотеза H_0 не отклоняется. Следует считать, что различие в показаниях приборов вызвано случайными ошибками.

В пакете STATISTICA решение выглядит следующим образом (рис. 4.8).

Sign Test (stest.sta)

Continue...	No. of Non-ties	Percent v < V	Z	p-level
VAR1 & VAR2	9	66,66666	,666667	,504985

Рис. 4.8. Решение примера 4.11

4.9. Критерий Вилкоксона для связанных пар наблюдений (Wilcoxon watched pairs test)

Критерий Вилкоксона, так же как и критерий знаков, используется для проверки гипотезы H_0 об однородности двух генеральных совокупностей по попарно связанным выборкам.

Отличие состоит в том, что используется информация об относительных размерах разностей элементов двух выборок. Проверка по критерию осуществляется следующим образом. Абсолютные значения ненулевых разностей упорядочиваются в порядке возрастания и определяются их ранги. Равным разностям присваивается средний ранг. Далее вычисляются суммы рангов для отрицательных разностей R_n и положительных разностей R_p. Для проверки расчетов используется тождество

$$R_n + R_p = \frac{n(n+1)}{2}.$$

Статистикой критерия T является число равное наименьшему значению суммы рангов R_n и R_p. Критические значения T определяются по таблице ([14], с. 289, табл. 67). При $n > 25$ используют статистику Z

$$Z = \frac{\left|T - \frac{n(n+1)}{4}\right|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}.$$

При условии, что гипотеза H_0 верна, Z имеет (приближенно) стандартное нормальное распределение $N(0, 1)$. Гипотеза H_0 отклоняется на уровне значимости α (при двухсторонней альтернативе), если

$$z_в > u_{1-\frac{\alpha}{2}},$$

где $z_в$ — выборочное значение статистики Z, а $u_{1-\frac{\alpha}{2}}$ — квантиль стандартного нормального распределения $N(0, 1)$ порядка $1-\frac{\alpha}{2}$.

Пример 4.12. Проверим гипотезу H_0 об однородности генеральных совокупностей по данным в примере 4.11, используя критерий Вилкоксона.

Решение. Значения разностей $v_1 - v_1$ следующие: –2, –1, 1, –1, 2, 0, –3, 5, –1, –2.

Упорядочим абсолютные значения ненулевых разностей, отметим их знак и определим ранги (нулевая разность не учитывается):

Номер	1	2	3	4	5	6	7	8	9
Модули разностей	1	1	1	1	2	2	2	3	5
Знак	+	–	–	–	+	–	–	–	+
Ранги	2,5	2,5	2,5	2,5	6	6	6	8	9

Сумма рангов отрицательных разностей $R_n = 27{,}5$, для положительных разностей сумма рангов $R_p = 17{,}5$, таким образом, статистика критерия $T = 17{,}5$. По таблице ([14], табл. 67) критическое значение $T_{кр}$ для $n = 9$ на 5 % уровне значимости для двустороннего критерия равно 5, так как $T > 5$, то гипотеза H_0 принимается.

В пакете STATISTICA выводятся следующие результаты процедуры **Wilcoxon watched pairs test** (рис. 4.9):

значение T-статистики = 17,5;

значение Z-статистики $\approx 0{,}592$;

p-level = $P[|Z| > 0{,}592] \approx 0{,}553$.

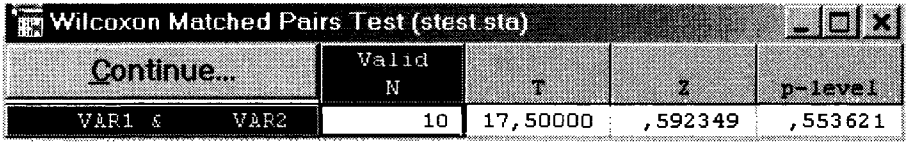


Рис. 4.9. Решение примера 4.12

4.9.1. Задачи

Решите задачи, используя критерий знаков и критерий Вилкоксона. Сравните и прокомментируйте результаты.

Задача 1. Сравнивалось действие двух экстрактов вируса табачной мозаики. Для этого каждая из двух половин листа натиралась соответствующим препаратом. Число пораженных мест приводятся ниже:

Экстракт A	20	39	43	13	28	26	17	49	36
Экстракт B	31	22	45	6	21	13	17	46	31

Можно ли считать, что действие этих экстрактов различно?

Принять $\alpha = 0{,}01$.

Задача 2. Ниже приводится время (в секундах) решения контрольных задач одиннадцатью учащимися до и после специальных упражнений по устному счету. Можно ли считать, что эти упражнения улучшили способности учащихся в решении задач?

Принять $\alpha = 0{,}01$.

До упражнения	87	61	98	90	93	74	83	72	81	75	83
После упражнения	50	45	79	90	88	65	52	79	84	61	52

Задача 3. Для десяти человек была предложена специальная диета. После двухнедельного питания по этой диете масса их тела изменилась следующим образом:

Масса до диеты (кг)	68	80	92	81	70	79	78	66	57	76
Масса после диеты (кг)	60	84	87	79	74	71	72	67	57	70

1. Можно ли рекомендовать эту диету для людей, желающих похудеть?
2. Оказывает ли эта диета какое-либо существенное действие на массу тела?

Принять $\alpha = 0{,}10$.

Задача 4. Проверить предположение о том, что предлагаемый лечебный препарат не меняет состав крови (в частности, число лейкоцитов), если препарат испытывался на десяти особях, а последующий анализ крови дал следующие результаты:

0,97; 1,05; 1,09; 0,88; 1,01; 1,14; 1,03; 1,07; 0,94; 1,02

(числа выражают отношение числа лейкоцитов в опыте к числу лейкоцитов в норме). Принять $\alpha = 0{,}01$.

Задача 5. Изучалось влияние черного и апрельского пара на урожай ржи. Опыт длился шесть лет. Учитывалась масса 1000 зерен в граммах. Результаты опыта следующие:

Год посева	1	2	3	4	5	6
По черному пару	31,1	24,0	24,6	28,6	29,1	30,1
По апрельскому пару	31,6	24,2	24,8	19,1	29,9	31,0

Можно ли считать, что урожай ржи по апрельскому пару значимо выше, чем по черному? Проверить это предположение, если $\alpha = 0{,}05$.

Задача 6. Изменение урожайности при применении одного из видов предпосевной обработки семян характеризуется следующими данными (в центрах с гектара):

Год	1972	1973	1974	1975	1976	1977	1978	1979	1980
Необработанные семена	20,0	17,9	20,6	22,0	21,4	23,8	21,4	19,8	18,4
Обработанные семена	22,1	18,5	19,4	22,1	21,7	24,9	21,6	20,3	18,3

Можно ли считать, что предпосевная обработка увеличивает урожайность? Принять $\alpha = 0{,}05$.

4.10. Двухфакторный анализ Фридмана и коэффициент конкордации Кендалла (Friedman ANOVA and Kendall's concordance)

Рассмотрим следующую задачу. Кинопленка четырех видов была представлена трем экспертам для определения лучшей из них. Каждому эксперту предложили упорядочить пленки по степени предпочтения. Баллы (ранги), проставленные экспертами, приведены в таблице. Наибольший балл соответствует пленке самого лучшего качества.

Эксперты, k	Вид пленки, n			
	1	2	3	4
1	2	1	3	4
2	2	1	4	3
3	2	1	4	3
Σ	6	3	11	10

В данной задаче на результат оценки качества пленки оказывают влияние два фактора: вид пленки (способ изготовления, обработки) и индивидуальные особенности экспертов при оценке пленок одного и того же вида. В примере это приводит к трем связанным выборкам (строкам) объема 4.

Требуется определить, различаются ли виды пленок и согласованы ли оценки экспертов. Если оценки экспертов не согласованы, т. е. являются *независимыми*, то им, очевидно, нельзя доверять, так как их оценки носят случайный характер, на который не оказывают влияния представленные пленки.

Эту задачу можно обобщить.

Пусть таблица результатов оценки или наблюдений n объектов состоит из k строк и n столбцов. В строках записываются k ранжированных переменных, причем длины ранжировок (объемы выборки) равны n. Строки таблицы можно рассматривать как k связанных выборок объема n. Связность выборок следует из того, что выборки суть — повторные наблюдения на одних и тех же n объектах. Если объекты не различаются между собой, суммы рангов по столбцам также не будут различаться. Нулевая гипотеза H_0: между столбцами нет различия — проверяется с помощью статистики Фридмана F.

Выборочное значение статистики F, $F_в$ вычисляется по формуле

$$F_в = \frac{12}{kn(n+1)} \sum_{j=1}^{n} \left[\sum_{i=1}^{k} R_{ij} - \frac{1}{2} k(n+1) \right]^2 = \frac{12}{kn(n+1)} \sum_{j=1}^{n} \left(\sum_{i=1}^{k} R_{ij} \right)^2 - 3k(n+1),$$

где R_{ij} — ранг j-го объекта, присваемый i-м экспертом.

Если гипотеза H_0 верна, то при $k \to \infty$ статистика F имеет распределение хи-квадрат с $(n-1)$ степенями свободы.

Гипотеза H_0 отклоняется на уровне значимости α, если

$$F_в > \chi^2_{1-\alpha}(n-1),$$

где $\chi^2_{1-\alpha}(n-1)$ — квантиль распределения $\chi^2(n-1)$ порядка $1-\alpha$.

Мерой согласия различных ранжировок n объектов является коэффициент конкордации (согласия) Кендалла W:

$$W = \frac{F}{k(n-1)}.$$

Коэффициент конкордации W лежит в пределах: $0 \le W \le 1$. $W = 1$ тогда и только тогда, когда все k ранжировок совпадают. Статистическая значимость W проверяется на основе того, что статистика $k(n-1)W$ при $k \to \infty$ имеет (приближенно) распределение хи-квадрат с $(n-1)$ степенями свободы. Если $n \le 7$, то для проверки статистической значимости W используют таблицы критических значений (см. [21], табл. П16б).

В случае, когда в ранжировках (в строках таблицы) имеются совпадающие ранги, вычисляется скорректированная статистика F':

$$F' = \frac{\sum_{j=1}^{n} \left[\sum_{i=1}^{k} R_{ij} - \frac{1}{2} k(n+1) \right]^2}{\frac{1}{12} kn(n+1) - \frac{1}{n-1} \sum_{i=1}^{k} T_i},$$

где $T_i = \frac{1}{12} \sum_{t=1}^{m} [(n_t)^3 - n_t]$, $i = 1, 2, \ldots, k$. Здесь m — число групп повторяющихся рангов в i-ой ранжировке, n_t — число совпадающих рангов в группе с номером t, $t = 1, 2, \ldots, m$.

Пример 4.13. Вычислим статистику Фридмана F и коэффициент конкордации W по данным задачи об экспертах кинопленки.

Решение. В этом примере число ранжировок $k = 3$, объем выборки $n = 4$. Суммы рангов по столбцам: 6, 3, 11, 10. Значение выборочной статистики критерия $F_в$ равно

$$F_в = \frac{12}{3 \cdot 4 \cdot (4+1)}(6^2 + 3^2 + 11^2 + 10^2) - 3 \cdot 3 \cdot (4+1) = 8{,}2;$$

при $\alpha = 0{,}05$, $\chi^2_{0,95}(3) = 7{,}81$.

Следовательно, на уровне значимости $\alpha = 0{,}05$ гипотеза H_0 отклоняется: следует считать, что виды пленок, по мнению экспертов, различны.

Коэффициент конкордации W равен

$$W = \frac{8{,}2}{3 \cdot (4-1)} = \frac{8{,}2}{9} = 0{,}91.$$

Большое значение W свидетельствует о согласованности оценок экспертов.

При решении задачи в модуле **Nonparametrics/Distrib.** пакета STATISTICA данные вводятся как n переменных (**vars**), каждая из которых имеет k значений (**cases**). Данные можно вводить в виде любых чисел, ранжировка производится при вычислениях.

Таблица результатов содержит по каждой переменной (столбцу): средний ранг, сумму рангов, среднее и стандартное отклонение, выборочное значение статистики Фридмана $F_в$, уровень значимости

$$p = P[\chi^2(n-1) > F_в],$$

коэффициент конкордации W и среднее значение рангового коэффициента корреляции Спирмена $\overline{r_s}$ вычисленное как среднее по всем возможным парам ранжировок (число пар равно $\frac{k(k-1)}{2} = C_k^2$):

$$\overline{r_s} = \frac{kW - 1}{k - 1}.$$

В пакете **STATISTICA** решение примера 4.13 имеет следующий вид (рис. 4.10).

Friedman ANOVA and Kendall Coefficient of Concordance

Continue...

ANOVA Chi Sqr. (N = 3, df = 3) = 8,200000 p < ,04207
Coeff. of Concordance = ,91111 Aver. rank r = ,86667

Variable	Average Rank	Sum of Ranks	Mean	Std.Dev.
VAR1	2,000000	6,00000	2,000000	--
VAR2	1,000000	3,00000	1,000000	--
VAR3	3,666667	11,00000	3,666667	,577350
VAR4	3,333333	10,00000	3,333333	,577350

Рис. 4.10. Решение примера 4.13

4.11. *Q*-критерий Кокрена (Cochran *Q*-test)

Критерий применяется в следующей ситуации. Предположим, что n объектов подвергается k различным воздействиям или условиям. Результаты записываются в виде «да—нет», «1»—«0», «+»—«–» и так далее.

Для примера рассмотрим следующую задачу.

Во время презентации четырех новых сортов мороженого пятнадцати покупателям было предложено попробовать все сорта мороженого и высказать свое отношение к каждому сорту в следующем виде: 0 — нравится, 1 — не нравится. Ответы покупателей записаны в следующей таблице:

Покупатели, n	Вид мороженого, k				$\sum = v_i$
	1	2	3	4	
1	0	0	0	0	0
2	1	1	0	0	2
3	1	1	0	0	2
4	0	0	0	0	0
5	1	0	1	1	3
6	1	1	0	0	2
7	1	1	0	1	3
8	1	0	0	1	2
9	0	1	1	0	2
10	1	0	0	0	1
11	1	1	0	1	3
12	0	1	1	1	3
13	1	0	0	0	1
14	1	0	0	1	2
15	1	1	1	1	4
$\sum = u_i$	11	8	4	7	30

Требуется проверить гипотезу H_0: все сорта мороженого нравятся покупателям в равной степени. Альтернативная гипотеза H_1 утверждает, что сорта мороженого нравятся покупателям в разной степени. Для проверки гипотезы H_0 используется критерий Кокрена. Статистика критерия Q вычисляется следующим образом:

1. Найдем число единиц в каждом столбце $u_1, u_2, ..., u_k$ и число единиц в каждой строке $v_1, v_2, ..., v_n$.

2. Вычислим следующие суммы и суммы квадратов:

$$\sum_{i=1}^{k} u_i, \quad \sum_{j=1}^{n} v_j, \quad \sum_{i=1}^{k} u_i^2, \quad \sum_{j=1}^{n} v_j^2.$$

3. Вычислим статистику критерия Q по формуле

$$Q = \frac{(k-1)\left[k\left(\sum_{i=1}^{k} u_i^2\right) - \left(\sum_{i=1}^{k} u_i\right)^2\right]}{k\left(\sum_{j=1}^{n} v_j\right) - \sum_{j=1}^{n} v_j^2}.$$

При условии, что гипотеза H_0 верна, распределение статистики Q сходится при $n \to \infty$ к распределению хи-квадрат с $(k-1)$ степенями свободы. Таким образом, гипотеза H_0 отклоняется на уровне значимости α, если

$$Q_в > \chi^2_{1-\alpha}(k-1),$$

где $\chi^2_{1-\alpha}(k-1)$ — квантиль распределения $\chi^2(k-1)$ порядка $(1-\alpha)$, а $Q_в$ — выборочное значение статистики Q.

Пример 4.14. Проверить гипотезу H_0 по данным в приведенной выше таблице.

Решение. Вычислим суммы по столбцам и строкам таблицы: результаты приведены в исходной таблице.

Далее вычислим:

$$\sum_{i=1}^{4} u_i = 11 + 8 + 4 + 7 = 30,$$

$$\sum_{j=1}^{15} v_j = 30,$$

$$\sum_{i=1}^{4} u_i^2 = 11^2 + 8^2 + 4^2 + 7^2 = 250,$$

$$\sum_{j=1}^{15} v_j^2 = 78.$$

Выборочное значение статистики Q равно

$$Q_в = \frac{(4-1)[4 \cdot 250 - 30^2]}{4 \cdot 30 - 78} \approx 7{,}143.$$

Так как $\chi^2_{0,095}(3) = 7{,}81$, что больше $Q_в$, то на уровне значимости $\alpha = 0{,}05$ гипотеза H_0 не отклоняется: следует считать, что все сорта мороженого нравятся покупателям в равной степени.

Cochran's Q Test (qtest.sta)

Continue

Number of valid cases: 15
Q = 7,142857, df = 3, p < ,067502

Variable	Sum	Percent 0's	Percent 1's
VAR1	11,00000	26,66666	73,33334
VAR2	8,00000	46,66667	53,33333
VAR3	4,00000	73,33334	26,66667
VAR4	7,00000	53,33333	46,66667

Рис. 4.11. Решение примера 4.14

При решении задачи в опции **Cochran Q-test** данные вводятся как k переменных (**vars**), каждая из которых имеет n значений (**cases**), записанных как 0 и 1, либо двумя другими кодами, значение которых надо установить перед выполнением процедуры. Таблица результатов, показанная на рисунке 4.11, содержит: суммы для каждой переменной, проценты нулей и единиц, выборочное значение статистики Q, число степеней свободы статистики хи-квадрат, $df = k - 1$ и уровень значимости

$$p = P[\chi^2(k-1) > Q_в].$$

Глава 5
ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

5.1. Основные понятия

Пусть результаты наблюдений составляют l независимых выборок (групп), полученных из l нормально распределенных генеральных совокупностей, которые имеют, вообще говоря, различные средние m_1, m_2, ...m_l и равные дисперсии D. Проверяется гипотеза о равенстве средних H_0: $m_1 = m_2 = ... = m_l$. На практике такая задача возникает при исследовании влияния, которое оказывает изменение некоторого фактора на измеряемую величину. Например, если измерения проводятся на l различных приборах, то можно исследовать влияние фактора «прибор» на результаты измерения. Суть однофакторного дисперсионного анализа состоит в следующем.

Пусть x_{ik} обозначает i-й элемент k-й выборки, $i = 1, 2, ..., n_k$; $k = 1, 2, ..., l$; $\bar{x}_k$ — выборочное среднее k-й выборки

$$\bar{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik},$$

$\bar{x}$ — общее выборочное среднее

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{l}\sum_{i=1}^{n_k} x_{ik},$$

n — общее число наблюдений, $n = \sum_{k=1}^{l} n_k$.

Сумма квадратов отклонений наблюдений x_{ik} от общего среднего $\bar{x}$ может быть представлена так:

$$\sum_{k=1}^{l}\sum_{i=1}^{n_k}(x_{ik} - \bar{x})^2 = \sum_{k=1}^{l} n_k(\bar{x}_k - \bar{x})^2 + \sum_{k=1}^{l}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)^2.$$

Это основное тождество дисперсионного анализа, запишем его в виде

$$Q = Q_1 + Q_2.$$

Q — сумма квадратов отклонений наблюдений от общего среднего, Q_1 — сумма квадратов отклонений выборочных средних групп от общего среднего (между группами), Q_2 — сумма квадратов отклонений наблюдений от выборочных средних групп (внутри групп).

Основное тождество легко проверяется, если возвести в квадрат обе части очевидного равенства: $(x_{ik} - \bar{x}) = [(\bar{x}_k - \bar{x}) + (x_{ik} - \bar{x}_k)]$, затем просуммировать обе части по i ($i = 1, 2, ..., n_k$) и k ($k = 1, 2, ..., l$) и учесть, что

$$\sum_{k=1}^{l}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)(\bar{x}_k - \bar{x}) = 0$$

в силу определения средних $\bar{x}_k$ и $\bar{x}$.

Если верна гипотеза H_0 о равенстве средних, то можно показать [22], что статистики Q_1/D *и* Q_2/D независимы и имеют распределение χ^2 соответственно с $l - 1$ и $n - l$ степенями свободы. Следовательно, статистики

$$S_1^2 = \frac{Q_1}{l-1} \text{ и } S_2^2 = \frac{Q_2}{n-l}$$

являются несмещенными оценками неизвестной дисперсии ошибок наблюдений D. Оценка S_1^2 характеризует рассеяние групповых средних, а оценка S_2^2 — рассеяние внутри групп, которое обусловлено случайными вариациями результатов наблюдений. Значительное превышение величины S_1^2 над значением величины S_2^2 можно объяснить различием средних в группах. Отношение этих оценок при условии, что верна гипотеза H_0, имеет распределение Фишера с $l - 1$ и $n - l$ степенями свободы

$$\frac{S_1^2}{S_2^2} = \frac{Q_1/(l-1)}{Q_2/(n-l)} = F(l-1, n-l).$$

Эта статистика используется для проверки гипотезы H_0 о равенстве средних. Гипотеза не противоречит результатам наблюдений, если выборочное значение $F_в$ статистики F меньше квантили распределения Фишера $F_{1-\alpha}(l-1, n-l)$. Если $F_в$ больше $F_{1-\alpha}(l-1, n-l)$, то гипотеза H_0 отклоняется и следует считать, что среди средних $m_1, m_2, ..., m_l$ имеется хотя бы два неравных друг другу.

Пример 5.1. Удобрения для комнатных растений фасуются в пакеты весом по 0,5 кг. Из партии пакетов, расфасованных в течение суток, случайным образом отобрали 30 пакетов. Они были распределены по трем различным условиям хранения. После хранения в течение одной недели определялось содержание влаги в продукте, хранящемся в каждом пакете.

Данные о содержании влаги приводятся ниже.

Условия хранения	Содержание влаги, %
1	10,1 7,3 5,6 6,2 8,4 8,1 8,0 7,6 5,3 7,2
2	11,7 12,2 11,8 7,8 8,9 9,9 12,4 11,0 10,3 13,8 10,5 9,8 9,1
3	10,2 12,0 8,8 8,7 10,5 11,0 9,1

На уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что условия хранения продукта не оказывают влияния на содержание влаги.

Предполагается, что выборки получены из независимых нормально распределенных совокупностей с одной и той же дисперсией.

Решение. Задача состоит в проверке гипотезы $H_0 : m_1 = m_2 = m_3$, где m_k — математическое ожидание случайной величины — содержание влаги в продукте с k-м условием хранения, $k = 1, 2, 3$. В нашем случае число уровней фактора «условия хранения продукта», $l = 3$, общий объем всей выборки: $n = 10 + 13 + 7 = 30$.

Вычисления удобно проводить в такой последовательности.

Вычислим суммы элементов выборок для каждого уровня фактора, по группам $x_{.k} = \sum_{i=1}^{n_k} x_{ik}$; $x_{.1} = 73{,}8$; $x_{.2} = 139{,}2$; $x_{.3} = 70{,}3$.

Сумма всех элементов выборки равна

$$x_{..} = \sum_{k=1}^{l} x_{.k} = 73{,}8 + 139{,}2 + 70{,}3 = 283{,}3,$$

а сумма их квадратов будет

$$\sum_{k=1}^{l}\sum_{i=1}^{n_k} x_{ik}^2 \approx 2801{,}61.$$

Далее получаем:

$$Q = \sum_{k=1}^{l}\sum_{i=1}^{n_k}(x_{ik} - \bar{x})^2 = \sum_{k=1}^{l}\sum_{i=1}^{n_k}(x_{ik})^2 - \frac{1}{n}(x_{..})^2 = 2801{,}61 - \frac{1}{30}(283{,}3)^2 \approx 126{,}31,$$

$$Q_1 = \sum_{k=1}^{l} n_k(\bar{x}_k - \bar{x})^2 = \sum_{k=1}^{l}\frac{1}{n_k}x_{.k}^2 - \frac{1}{n}(x_{..})^2 = \frac{1}{10}73{,}8^2 + \frac{1}{13}139{,}2^2 +$$

$$+\frac{1}{7}70{,}3^2 - \frac{1}{30}(283{,}3)^2 \approx 65{,}869,$$

$$Q_2 = Q - Q_1 = 126{,}31 - 65{,}869 = 60{,}441.$$

Вычисляем выборочное значение статистики F:

$$F_{\text{в}} = \frac{Q_1/(l-1)}{Q_2/(n-l)} = \frac{65{,}869/(3-1)}{60{,}441/(30-3)} \approx 14{,}712.$$

Используя вероятностный калькулятор или таблицы квантилей распределения Фишера, находим $F_{0{,}95}(2{,}27) = 3{,}35$. Так как $F_{\text{в}} = 14{,}712 > 3{,}35$, то на уровне значимости $\alpha = 0{,}05$ гипотеза о равенстве средних отклоняется: условия хранения продукта оказывают значимое влияние на содержание влаги.

Линейные контрасты. Если гипотеза H_0 о равенстве средних отклоняется, то требуется определить, какие именно группы имеют значимое различие средних.

Для этих целей используется *метод линейных контрастов.* Линейный контраст Lk определяется как линейная комбинация

$$Lk = \sum_{k=1}^{l} c_k m_k,$$

где c_k, $k = 1, 2, ..., l$ — константы, однозначно определяемые из формулировки проверяемых альтернативных гипотез, причем

$$\sum_{k=1}^{l} c_k = 0.$$

Оценка линейного контраста Lk равна

$$\widetilde{Lk} = \sum_{k=1}^{l} c_k \bar{x}_k,$$

а оценка дисперсии линейного контраста Lk вычисляется по формуле

$$S^2_{Lk} = \widetilde{D}[Lk] = \widetilde{\sigma}^2 \sum_{k=1}^{l} \frac{c_k^2}{n_k} = \frac{Q_2}{n-l} \sum_{k=1}^{l} \frac{c_k^2}{n_k}.$$

Границы доверительного интервала для Lk имеют вид

$$\widetilde{Lk} \pm s_{Lk} \sqrt{(l-1) \cdot F_{1-\alpha}(l-1, n-l)}.$$

Пример 5.2. В условиях примера 5.1 при двусторонних альтернативных гипотезах проверить гипотезы $H_0^{(1)}: m_1 = m_2$; $H_0^{(2)}: m_1 = m_3$; $H_0^{(3)}: m_2 = m_3$; $H_0^{(4)}: \frac{1}{2}(m_1 + m_3) = m_2$.

Решение. В соответствии с проверяемыми гипотезами $H_0^{(i)}$, $i = 1, 2, 3, 4$, определяются линейные контрасты:

$Lk_1 = m_1 - m_2$;	$c_1 = 1$,	$c_2 = -1$,	$c_3 = 0$;
$Lk_2 = m_1 - m_3$;	$c_1 = 1$,	$c_2 = 0$,	$c_3 = -1$;
$Lk_3 = m_2 - m_3$;	$c_1 = 0$,	$c_2 = 1$,	$c_3 = -1$;
$Lk_4 = 1/2(m_1 + m_2) - m_3$;	$c_1 = 1/2$,	$c_2 = 1/2$,	$c_3 = -1$.

Найдем границы доверительных интервалов для линейных контрастов Lk_i, $i = 1, 2, 3, 4$.

Предварительно вычислим оценки линейных контрастов и их дисперсий. Выборочные средние по группам равны: $\bar{x}_1 = 7{,}38$, $\bar{x}_2 \approx 10{,}71$, $\bar{x}_3 \approx 10{,}04$.

Оценка дисперсии ошибок наблюдений:

$$\widetilde{\sigma}^2 = \frac{Q_2}{n-l} = \frac{60{,}441}{30-3} \approx 2{,}239.$$

Вычислим оценки контрастов и их дисперсий:

$$\widetilde{Lk}_1 = 7{,}38 - 10{,}71 = -3{,}33,\ s^2_{Lk_1} = 2{,}239 \cdot \left(\frac{1}{10} + \frac{1}{13}\right) \approx 0{,}396;$$

$$\widetilde{Lk}_2 = 7{,}38 - 10{,}04 = -2{,}66,\ s^2_{Lk_2} = 2{,}239 \cdot \left(\frac{1}{10} + \frac{1}{7}\right) \approx 0{,}544;$$

$$\widetilde{Lk}_3 = 10{,}71 - 10{,}04 = 0{,}67,\ s^2_{Lk_3} = 2{,}239 \cdot \left(\frac{1}{13} + \frac{1}{7}\right) \approx 0{,}492;$$

$$\widetilde{Lk}_4 = \frac{1}{2}(7{,}38 + 10{,}71) - 10{,}04 = -0{,}995,$$

$$s^2_{Lk_4} = 2{,}239 \cdot \left(\frac{(1/2)^2}{10} + \frac{(1/2)^2}{13} + \frac{1}{7} \right) \approx 0{,}419.$$

По таблице (см. [1], либо воспользуйтесь вероятностным калькулятором) находим квантиль распределения Фишера $F_{1-\alpha}(l-1, n-l) = F_{0{,}95}(2,27) = 3{,}35$. Чтобы определить доверительные интервалы для линейных контрастов, предварительно вычислим

$$\sqrt{(l-1)F_{1-\alpha}(l-1, n-l)} = \sqrt{(3-1)\cdot 3{,}35} \approx 2{,}59.$$

Таким образом, доверительные границы для контрастов Lk_i, $i = 1, 2, 3, 4$, равны соответственно $-3{,}33 \pm 1{,}63$; $-2{,}66 \pm 1{,}91$; $0{,}67 \pm 1{,}82$; $-0{,}995 \pm 1{,}68$.

Так как нулевое значение накрывается доверительными интервалами для Lk_3 и Lk_4, то гипотезы $H_0^{(3)}$ и $H_0^{(4)}$ принимаются, гипотезы $H_0^{(1)}$ и $H_0^{(2)}$ отклоняются. Таким образом, значимо различны средние первой и второй групп, а также средние первой и третьей групп.

5.2. Решение примера в пакете STATISTICA

Решим пример 5.1, используя модуль **Basic Statistics and Tables**, опция **Breakdown and one-way ANOVA** (однофакторный дисперсионный анализ). Можно использовать также модуль **ANOVA/MANOVA**. Создадим таблицу с двумя столбцами P и G и 30 строками; в P занесем данные по влажности продукта, в G — обозначения групп, определяющих условия хранения: группы 1, 2, 3. В стартовой панели модуля **Basic Statistics** выберем процедуру **Breakdown and one-way ANOVA** (классификация и однофакторный дисперсионный анализ), рис. 5.1.

В окне для задания параметров процедуры (рис. 5.2) в опции **Analysis** выберем: **Detailed Analysis of Individual tables** и введем переменные для ана-

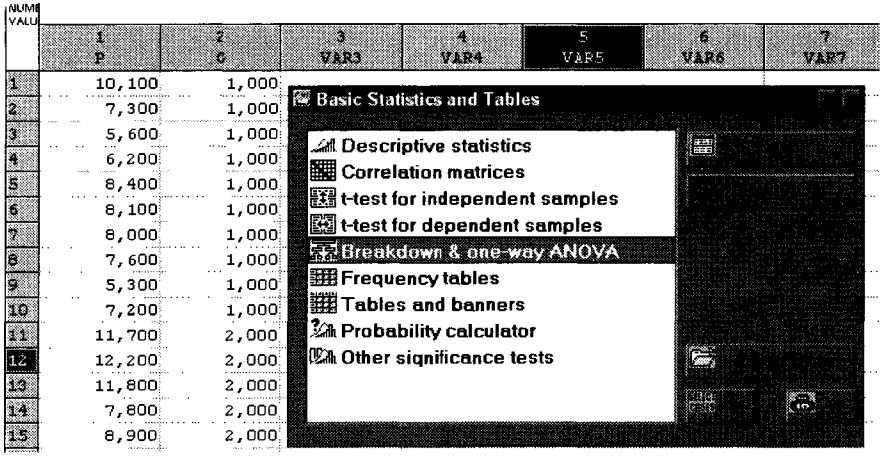


Рис. 5.1. Исходные данные для примера и меню для выбора процедуры

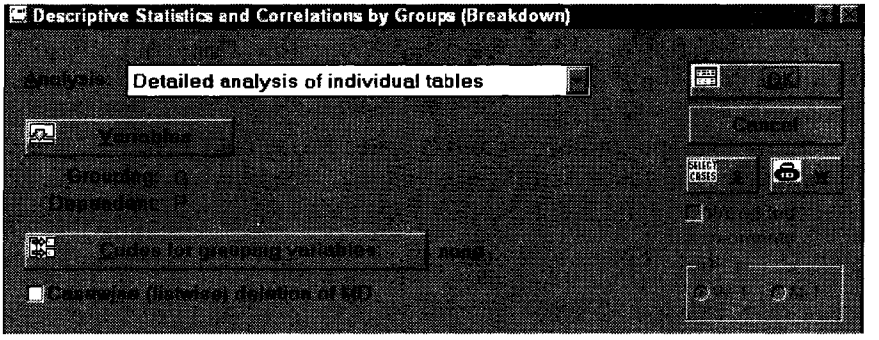


Рис. 5.2. Выбор данных для дисперсионного анализа

лиза (кнопка **Variables**). Как **Grouping variables** (группирующие переменные) следует указать столбец *G*, а как **Dependent variables** (зависимые переменные — отклики) столбец *P*, OK.

В окне результатов (рис. 5.3) отметим следующие статистики: **Number of observations** (количество наблюдений), **Standard deviations** (стандартные отклонения) и **Variances** (дисперсии).

Чтобы вычислить эти статистики для каждой из групп нажмите кнопку **Summary table of means** (таблица средних) в левой части окна. Таблица результатов приводится на рис. 5.4.

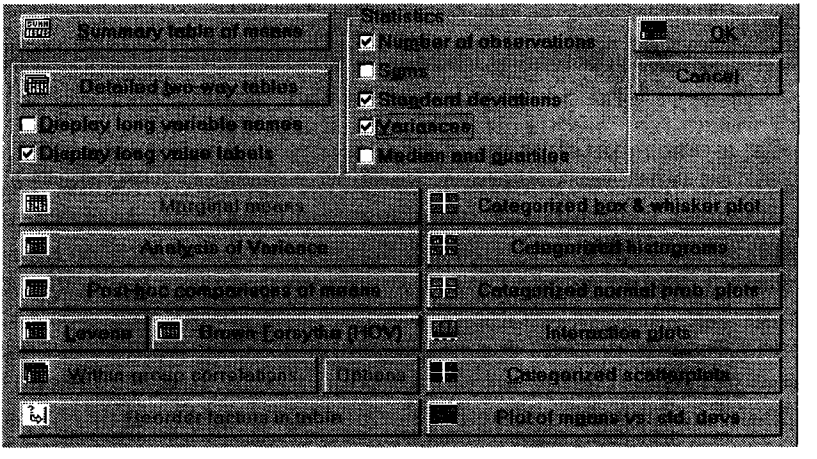


Рис. 5.3. Окно для выбора результатов расчета

Summary Table of Means (113.sta)

N=30 (No missing data in dep. var. list)

G	P Means	P N	P Std.Dev.	P Variance
G 1:1	7,38000	10	1,426573	2,035111
G 2:2	10,70769	13	1,656030	2,742436
G 3:3	10,04286	7	1,239432	1,536190
All Grps	9,44333	30	2,087018	4,355644

Рис. 5.4. Таблица средних

Возвратимся в окно **Descriptive Stats and ... Results** (описательные статистики и результаты) (рис. 5.3) (используя кнопку **Continue**) и выполним дисперсионный анализ, нажав кнопку **Analysis of Variance** (дисперсионный анализ).

В таблице дисперсионного анализа (рис. 5.5) приводятся:

сумма квадратов отклонений выборочных средних групп $\bar{x}_k$ от общего среднего $\bar{x}$ (между группами), Q_1 (**SS Effect** = 65,871); число степеней свободы для Q_1, $l-1$ (**df Effect** = 2); отношение $\frac{Q_1}{l-1}$ — среднее значение суммы квадратов (**MS Effect** = 32,936).

Далее приводятся: сумма квадратов отклонений результатов наблюдений x_{ik} от выборочных средних групп $\bar{x}_k$ (внутри групп), Q_2 (**SS Error** = 60,442); число степеней свободы для Q_2, $n-l$ (**df Error** = 27); отношение $\frac{Q_2}{n-l}$ — среднее значение суммы квадратов (**MS Error** = 2,238); выборочное значение F-статистики, $F_в = 14{,}712$ и вычисленный уровень значимости $p = P[F(2,27) > F_в] = 0{,}000048$.

Analysis of Variance (113.sta)

Marked effects are significant at p < ,05000

Variable	SS Effect	df Effect	MS Effect	SS Error	df Error
P	65,87129	2	32,93565	60,44237	27

Рис. 5.5. Таблица дисперсионного анализа

Так как вычисленный уровень значимости p меньше заданного уровня значимости $\alpha = 0{,}05$, то гипотеза о равенстве средних отклоняется. Таким образом, условия хранения продукта значимо влияют на содержание влаги в продукте.

Определим, какие виды условий хранения продукта приводят к значимому различию в содержании влаги. Для ответа на этот вопрос возвращаемся в окно **Descriptive Stats and ... Results** (рис. 5.3) и выполняем **Post - hoc comparisons of means** (сравнение средних) по методу Шеффе (**Sheffe test**).

В таблице попарного сравнения средних (рис. 5.6) указаны уровни значимости для проверки гипотез о равенстве средних для всех пар уровней фактора G. Гипотеза о том, что математическое ожидание второй группы

Scheffe Test; Variable: P (113.sta)

Marked differences are significant at p < ,05000

G	{1} M=7,3800	{2} M=10,708	{3} M=10,043
G 1:1 {1}		,000068	,004890
G 2:2 {2}	,000068		,642827
G 3:3 {3}	,004890	,642827	

Рис. 5.6. Таблица попарного сравнения средних

равно математическому ожиданию третьей группы принимается на уровне значимости $p \approx 0{,}643$, гипотеза о равенстве математических ожиданий первой и второй групп отклоняется, $p \approx 0{,}000068$. Также отклоняется гипотеза о равенстве математических ожиданий первой и третьей групп, $p \approx 0{,}00489$.

5.3. Проверка предположений дисперсионного анализа

Напомним, что при применении дисперсионного анализа предполагается, что исходные данные — независимые выборки наблюдений, полученные из нормально распределенных генеральных совокупностей имеющих одну и ту же дисперсию. При выполнении анализа в пакете STATISTICA выполнение этих предположений можно проверить. Один из способов проверки нормальности состоит в том, что исходные данные (по группам) наносятся на специальный график — вероятностную бумагу (см. например [14, 19]). Чтобы выполнить эту процедуру нужно в окне результатов (рис. 5.3) нажать кнопку **Categorized normal prob. plots** (категоризованные нормальные вероятностные графики), затем выбрать **Normal prob. plot**, **OK**. Для исходных данных примера 5.1 получим следующие графики (рис. 5.7).

Точки, соответствующие нормально распределенным данным, укладываются на прямые. Как показывают графики, исходные данные достаточно плотно группируются относительно прямых.

Для проверки выполнения предположения о равенстве дисперсий по группам используются так называемые критерии однородности дисперсий (**Test of Homog. of Variances**). Один из таких критериев (критерий Бартлетта) был рассмотрен в главе 3, пример 3.3.

В пакете STATISTICA для этих целей можно использовать критерий Левена (**Levene test....**) либо критерий Брауна—Форсайта (**Brown—Forsythe Test...**).

Чтобы выполнить эти процедуры нужно нажать соответствующие кнопки в окне результатов дисперсионного анализа (рис. 5.3). Для исход-

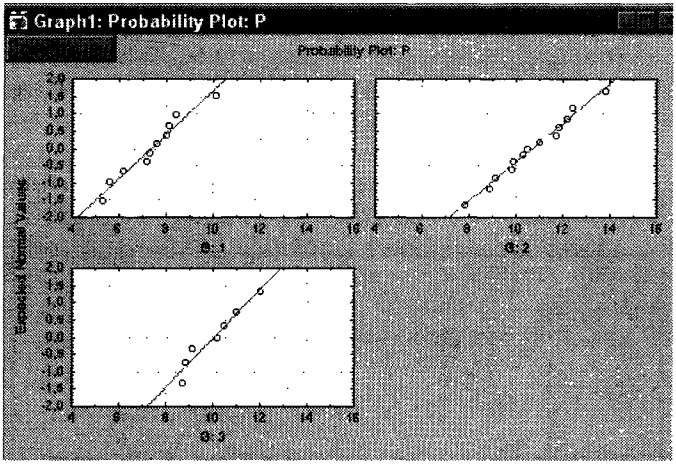


Рис. 5.7. Исходные данные примера 5.1 на нормальной вероятностной бумаге

ных данных примера 5.1 гипотеза о равенстве дисперсий по группам принимается по этим двум критериям на уровнях значимости, соответственно, $p \approx 0{,}640$ и $p \approx 0{,}669$.

Здесь необходимо сделать следующее замечание.

Задача рассматриваемого примера 5.1 состоит в том, чтобы продемонстрировать технику вычислений. Применение дисперсионного анализа (как и большинства других статистических методов) для получения обоснованных и практически важных выводов требует значительно больше исходных данных. Только в этом случае мы можем говорить о проверке нормальности и других предположений. Вопрос о том насколько оправдано применение статистических методов, основанных на «нормальной теории», к данным, распределение которых отличается от нормального, является достаточно сложным (см., например, [13, 23]). В связи с этим, в тех случаях, когда выполнение предположений дисперсионного анализа проблематично, следует использовать также и соответствующие непараметрические процедуры (см. главу 4) и сравнить результаты.

5.4. Задания для самостоятельного решения

В задачах 1 и 2 требуется проверить гипотезу о равенстве средних в трех группах.

Если гипотеза принимается, то найти несмещенные оценки средних в группах и дисперсии ошибок наблюдений.

В случае, если гипотеза отклоняется, провести попарное сравнение средних, используя метод линейных контрастов.

Задача 1. Проверьте гипотезу о равенстве средних по следующим трем выборкам:

1	2	3
6	14	12
5	11	4
12	5	7
9	6	—
10	—	—

$\alpha = 0{,}05$.

Задача 2. Проверьте гипотезу о равенстве средних по данным о товарообороте трех магазинов в течение шести месяцев (в млн руб.).

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

$\alpha = 0{,}10$.

Задача 3. На химическом заводе разработаны два варианта технологического процесса. Чтобы оценить, как изменится дневная производительность при переходе на работу по новым вариантам технологического процесса, завод в течение 10 дней работает по каждому из вариантов.

Дневная производительность завода приводится в таблице:

День работы	Существующая схема	Вариант 1	Вариант 2
1	46	74	52
2	48	82	63
3	73	64	64
4	52	72	48
5	72	84	70
6	44	68	78
7	66	76	68
8	46	88	70
9	60	70	54
10	48	60	75

Можно ли считать, что производительность завода изменилась при переходе на новые варианты технологического процесса? Принять $\alpha = 0{,}05$.

Задача 4. В трех магазинах, продающих товары одного вида, данные товарооборота за 8 месяцев работы (в тыс. руб.) составили следующую сводку

Магазин	Месяц							
	1	2	3	4	5	6	7	8
1	19	23	26	18	20	20	18	35
2	20	20	32	27	40	24	22	18
3	16	15	18	26	19	17	19	18

Требуется проверить гипотезу H_0 о равенстве среднего товарооборота в магазинах. Если гипотеза принимается, то найти несмещенные оценки среднего и дисперсии. Предполагается, что выборки получены из независимых нормально распределенных совокупностей с одной и той же дисперсией. Проверьте, выполняются ли эти предположения.

Принять $\alpha = 0{,}10$.

Задача 5. В приложении 1.2 приведены данные о стоимости однокомнатных квартир. Проверьте гипотезу о равенстве средней стоимости квартир в различных районах. Если гипотеза отклоняется, проведите попарное сравнение средних, используя метод линейных контрастов. Проверьте, выполняются ли предположения дисперсионного анализа для исходных данных.

$\alpha = 0{,}05$.

Задача 6. Проведите однофакторный дисперсионный анализ по данным примера 3.11 и задания 1 (п. 3.4, глава 3).

Фактором является группа предприятия, а переменными: среднегодовая стоимость ОПФ, производство продукции, фондоемкость и фондоотдача.

Уровень значимости $\alpha = 0{,}05$.

Сделайте выводы по результатам дисперсионного анализа.

Задача 7. Проведите однофакторный дисперсионный анализ по данным к заданиям 2, 3, 4 (п. 3.4). Фактором является группа предприятия.

В качестве переменных возьмите:

Задание 2. Розничный товарооборот; издержки обращения, удельный вес издержек.

Задание 3. Объем работ; накладные расходы, долю накладных расходов.

Задание 4. Нераспределенная прибыль; инвестиции, доля инвестиций.

Уровень значимости $\alpha = 0{,}05$.

Сделайте выводы по результатам дисперсионного анализа.

Глава 6
РЕГРЕССИОННЫЙ АНАЛИЗ

Во многих случаях исследуются объекты, характеризующиеся несколькими признаками. Например, у каждого человека можно измерить рост, вес, частоту пульса и ряд других физиологических показателей; работу торгового предприятия можно оценивать по объему товарооборота и величине прибыли. Совокупность данных такого типа представляет выборку из многомерной генеральной совокупности. Для таких данных интерес представляет не только определение характеристик распределения каждого признака, но и то, насколько тесно эти признаки связаны между собой, можно ли по значению одного признака сделать какие-либо выводы о предполагаемом значении другого признака и т. д.

Регрессионный анализ — это один из наиболее известных статистических методов, применяемых для решения задач такого рода. Основная цель регрессионного анализа состоит в определении связи между некоторой характеристикой Y наблюдаемого явления или объекта и величинами $x_1, x_2, ..., x_m$, которые обусловливают, объясняют изменения Y. Переменная Y называется *зависимой переменной* (откликом), объясняющие переменные $x_1, x_2, ..., x_m$ называются *предикторами*, *регрессорами* или *факторами*.

Например, нас может интересовать как зависит стоимость однокомнатной квартиры (y) от площади комнаты (x_1), размера кухни (x_2), удаленности дома от метро (x_3) и других тому подобных факторов. Для ответа на этот вопрос необходимо собрать данные по однокомнатным квартирам предлагаемым к продаже.

Пример исходных данных такого рода для 69 квартир приведен в Приложении 1.2.

В данном случае нужно выяснить как эти факторы связаны с ценой квартиры, какой фактор является наиболее важным при прогнозе стоимости, имеется ли в исходных данных квартиры, обладающие какими-либо специфическими свойствами (выбросы).

Если в рассматриваемом примере в качестве объясняющих факторов использовать только три определенных выше фактора x_1, x_2, x_3, то регрессионная модель может быть записана в виде

$$Y = f(x_1, x_2, x_3) + \varepsilon,$$

где $f(x_1, x_2, x_3)$ — детерминированная составляющая отклика Y, зависящая от x_1, x_2, x_3, а ε — случайная составляющая.

Случайная составляющая ε обусловлена влиянием на стоимость квартиры множества неучтенных факторов (среди которых могут быть и такие

непредсказуемые факторы, как человеческая реакция или мода), а также ошибок наблюдений или измерений зависимой переменной.

Замечание. Часто (например, в пакете STATISTICA) объясняющие переменные x_1, x_2, ..., x_m называют независимыми переменными. Такое название во многих случаях не соответствует реальной ситуации: «независимые» переменные могут быть зависимы и влиять одна на другую. Во многих случаях термин «независимые переменные» используется в другом контексте: это переменные, значения которых в процессе определения отклика, могут устанавливаться произвольно, независимо.

Существуют различные регрессионные модели, определяемые выбором функции $f(x_1, x_2, \ldots x_m)$:

1) простая линейная регрессия

$$Y = \beta_0 + \beta_1 x + \varepsilon;$$

2) множественная регрессия

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1} + \varepsilon;$$

3) полиномиальная регрессия

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_{k-1} x^{k-1} + \varepsilon;$$

4) регрессионная модель общего вида:

$$Y = \beta_0 + \beta_1 \varphi_1(x_1, x_2, \ldots, x_m) + \ldots + \beta_{k-1} \varphi_{k-1}(x_1, x_2, \ldots, x_m) + \varepsilon,$$

где $\varphi_i(x_1, x_2, \ldots, x_m)$, $i = 1, 2, \ldots, k-1$, — заданные функции факторов.

Коэффициенты β_0, β_1, ..., β_{k-1} называются ***параметрами регрессии.***

В приведенные регрессионные модели параметры β_0, β_1, ..., β_{k-1} входят линейно. Такие модели называют ***линейными (по параметрам) моделями***, а математические методы анализа этих моделей — *линейным регрессионным анализом.*

Модель $y = \beta_0 e^{\beta_1 x_1} + \beta_1 e^{\beta_2 x_2}$ нелинейна по параметрам. В некоторых случаях нелинейные модели с помощью специальных линеаризирующих преобразований могут быть преобразованы в линейные. Рассмотрим несколько примеров.

1. Функция $y = \beta_0 x^{\beta_1}$ при $x > 0$ с помощью логарифмирования и замены переменных преобразуется так: $\ln y = \ln \beta_0 + \beta_1 \ln x$. Произведя замену переменных $y' = \ln y$; $\beta_0' = \ln \beta_0$; $x' = \ln x$, получим линейную по параметрам функцию

$$y' = \beta_0' + \beta_1 x'.$$

2. Функция $y = \dfrac{ax}{b + x}$ преобразуется так:

$$b + x = a\frac{x}{y} \quad \text{или} \quad \frac{x}{y} = \frac{b}{a} + \frac{1}{a}x.$$

После замены переменных $y' = \frac{x}{y}$, $\beta_0 = \frac{b}{a}$, $\beta_1 = \frac{1}{a}$ получим

$$y' = \beta_0 + \beta_1 x.$$

3. Логистическая функция $y = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$ при помощи преобразования $y' = \ln\left(\frac{y}{1-y}\right)$ принимает вид:

$$y' = \beta_0 + \beta_1 x.$$

После выбора вида регрессионной модели, используя результаты наблюдений зависимой переменной и факторов, нужно вычислить оценки (приближенные значения) параметров регрессии, а затем проверить значимость и адекватность модели результатам наблюдений.

6.1. Простая линейная регрессия

Так называется простейшая регрессионная модель описывающая зависимость переменной Y от одного фактора x.

6.1.1. Коэффициент корреляции и простая линейная регрессия, оценка параметров регрессии методом наименьших квадратов

Пусть (x_i, y_i), $i = 1, 2, 3, ..., n$ — выборка наблюдений из двумерной генеральной совокупности. Предварительное представление о зависимости между случайными величинами X и Y можно получить, отображая элементы выборки как точки на плоскости. Такое представление выборки называется ***диаграммой рассеяния***.

При построении диаграммы рассеяния рекомендуется масштабы по осям x и y выбирать так, чтобы значения обоих признаков укладывались на отрезках приблизительно равной длины.

Возможны различные варианты расположения «облака» точек, по которым можно судить о виде и степени взаимосвязи между признаками X и Y (рис. 6.1, *а—г*).

Количественной характеристикой степени линейной зависимости между случайными величинами X и Y является ***коэффициент корреляции*** ρ (см. Приложение, П.2.7).

Оценка коэффициента корреляции ρ, по выборке наблюдений (x_i, y_i), $i = 1, 2, ..., n$, вычисляется по формуле

$$r = \frac{Q_{xy}}{\sqrt{Q_x \cdot Q_y}},$$

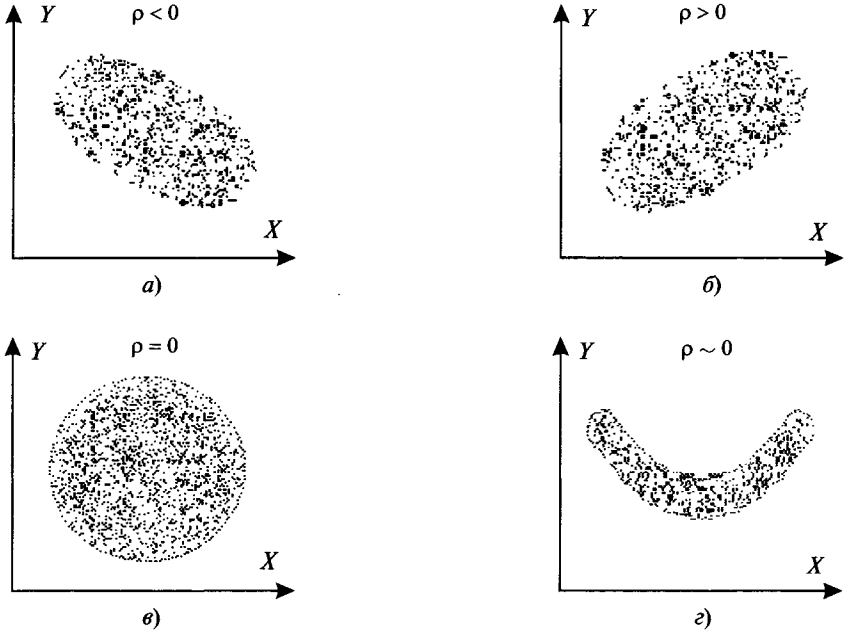


Рис. 6.1. Варианты расположения «облака» точек

где

$$Q_x = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n};$$

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

$$Q_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n};$$

$$\bar{x} = \frac{1}{n}\sum x_i,\ \bar{y} = \frac{1}{n}\sum y_i.$$

Для коэффициента корреляции справедливы следующие утверждения:

1) $-1 \le \rho \le 1$;

2) если $|\rho| = 1$, то между X и Y имеет место функциональная линейная зависимость, все точки (x_i, y_i) будут лежать на прямой (рис. 6.1, *д*);

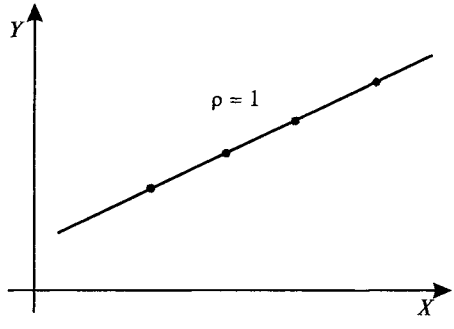


Рис. 6.1*д*. Функциональная линейная зависимость между X и Y при $\rho = 1$

3) если $\rho = 0$, то говорят, что X и Y некоррелированны, т. е. между ними нет линейной зависимости (см. рис. 6.1, *в—г*);

4) если X и Y имеют двумерное нормальное распределение, то из равенства $\rho = 0$ следует, что они статистически независимы.

В случае, когда между случайными величинами X и Y существует достаточно тесная линейная статистическая зависимость $|r| > 0$, ее можно аппроксимировать уравнением линейной регрессии Y на x:

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

где β_0 и β_1 — параметры линейной регрессии; x — независимая переменная (фактор, предиктор); y — зависимая переменная (отклик).

При этом предполагается, что независимая переменная x измеряется точно, а Y является случайной величиной. Таким образом, исследуют, как в «среднем» изменяются значения зависимой переменной Y при изменении независимой переменной x.

В тех случаях, когда признаки X и Y равнозначны (например, рост и вес), аналогично регрессии Y на x рассматривают линейную регрессию X на y: $x = \beta_0' + \beta_1' y$.

Если случайный вектор (X, Y) имеет двумерное нормальное распределение (см. Приложение, П.3.6), то линейная регрессия Y на x равна условному математическому ожиданию и имеет вид

$$M[Y/X = x] = m_y + \rho\frac{\sigma_y}{\sigma_x}(x - m_x),$$

а регрессия X на y равна

$$M[X/Y = y] = m_x + \rho\frac{\sigma_x}{\sigma_y}(y - m_y),$$

где m_x, m_y и σ_x, σ_y — соответственно математические ожидания и средние квадратические отклонения X и Y, ρ — коэффициент корреляции.

Для оценки параметров линейной регрессии Y на x по результатам наблюдений (x_i, y_i), $i = 1, 2, 3, \ldots, n$, используется метод наименьших квадратов: в качестве оценок параметров берут значения $\tilde{\beta}_0$ и $\tilde{\beta}_1$, минимизирующие $Q(\beta_0, \beta_1)$ сумму квадратов отклонений значений зависимой переменной y_i от значений $\tilde{y}_i$, вычисляемых по уравнению регрессии: $\tilde{y}_i = \beta_0 + \beta_1 x_i$, $i = 1, 2, \ldots, n$,

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Из необходимых условий минимума функции $Q(\beta_0, \beta_1)$:

$$\frac{\partial Q}{\partial \beta_0} = 0; \quad \frac{\partial Q}{\partial \beta_1} = 0$$

получаем формулы для вычисления оценок параметров β_1 и β_0 регрессии Y на x:

$$\tilde{\beta}_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} = \frac{Q_{xy}}{Q_x}; \quad (1)$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}. \quad (2)$$

Аналогично, для регрессии X на y оценки параметров β_1' и β_0' вычисляются по формулам

$$\tilde{\beta}_1' = \frac{Q_{xy}}{Q_y};$$

$$\tilde{\beta}_0' = \bar{x} - \tilde{\beta}_1' \bar{y}.$$

Уравнения

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$$

и

$$x = \tilde{\beta}_0' + \tilde{\beta}_1' y = \bar{x} + r\frac{s_x}{s_y}(y - \bar{y}),$$

где s_x и s_y — оценки средних квадратических отклонений σ_x и σ_y:

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{Q_x}{n}}, \quad s_y = \sqrt{s_y^2} = \sqrt{\frac{Q_y}{n}},$$

r — оценка коэффициента корреляции ρ, называются *выборочными уравнениями линейной регрессии.*

Прямые регрессии пересекаются в точке с координатами $\bar{x}$ и $\bar{y}$ и образуют «ножницы». При $|r| = 1$ обе прямые совпадают, при $|r| = 0$ — они перпендикулярны друг другу.

Между коэффициентом корреляции и параметрами регрессии имеются следующие соотношения:

$$\sqrt{\tilde{\beta}_1 \cdot \tilde{\beta}_1'} = |r|; \quad \tilde{\beta}_1 = r\frac{s_y}{s_x}; \quad \tilde{\beta}_1' = r\frac{s_x}{s_y}.$$

6.1.2. Предположения, при которых проводится регрессионный анализ. Статистический анализ простой линейной регрессии

Для выборки наблюдений (x_i, y_i), $i = 1, 2, ..., n$ простая линейная регрессия определяет n уравнений:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

где $x_1, x_2, ..., x_n$ — значения фактора x, а $y_1, y_2, ..., y_n$ соответствующие им значения, зависимой переменной Y, полученные как результаты независимых экспериментов или наблюдений; ε_i — ошибки наблюдений зависимой переменной, имеющие случайный характер.

В регрессионном анализе предполагается, что случайные величины ε_i и ε_j, $i \neq j$, $i, j = 1, 2, ..., n$ некоррелированны, имеют нулевое математическое ожидание: $M[\varepsilon_i] = 0$ и равные дисперсии: $D[\varepsilon_i] = \sigma^2$, $i = 1, 2, ..., n$.

При статистическом анализе регрессионной модели предполагается также, что случайные ошибки наблюдений имеют нормальное распределение: $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, ..., n$. В этом случае ε_i будут независимыми случайными величинами.

Задача линейного регрессионного анализа состоит в том, чтобы по результатам наблюдений (x_i, y_i), $i = 1, 2, ..., n$:

а) получить наилучшие точечные и интервальные оценки неизвестных параметров β_0, β_1 и σ^2;

б) проверить статистические гипотезы о параметрах модели;

в) проверить, достаточно ли хорошо модель согласуется с результатами наблюдений (адекватность модели результатам наблюдений).

Оценки параметров линейной регрессии (1) и (2), получаемые по методу наименьших квадратов, (МНК-оценки), при любом законе распределения ошибок наблюдений ε_i, $i = 1, 2, ..., n$ имеют следующие свойства:

1) МНК-оценки являются ***линейными функциями результатов наблюдений*** y_i, $i = 1, 2, ..., n$ и ***несмещенными оценками параметров***, т. е. $M[\tilde{\beta}_j] = \beta_j$, $j = 0, 1$.

2) МНК-оценки имеют ***минимальные дисперсии*** в классе несмещенных оценок, являющихся линейными функциями результатов наблюдений (теорема Гаусса—Маркова).

Если ошибки наблюдений ε_i некоррелированны и имеют нормальное распределение, т. е. $\varepsilon_i \sim N(0, \sigma^2)$, то в дополнение к свойствам 1 и 2 выполняется свойство

3) МНК-оценки ***совпадают с оценками, вычисляемыми по методу максимального правдоподобия.***

Функция

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x$$

определяет *выборочную (эмпирическую) регрессию Y на x.* Эмпирическая регрессия является оценкой предполагаемой (теоретической) линейной регрессии по результатам наблюдений. Разности между наблюдаемыми значениями переменной Y при $x = x_i$, $i = 1, 2, ..., n$ и расчетными значениями $\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i$ называются *остатками* и обозначаются e_i:

$$e_i = y_i - \tilde{y}_i, \; i = 1, 2, ..., n.$$

Сумма квадратов остатков e_i, $Q_e = \sum e_i^2 = \sum (y_i - \tilde{y}_i)^2$, называется ***остаточной суммой квадратов.***

Качество аппроксимации результатов наблюдений (x_i, y_i), $i = 1, 2, ..., n$, выборочной регрессией определяется величиной ***остаточной дисперсии*** S^2, вычисляемой по формуле

$$S^2 = \frac{\sum e_i^2}{n-2} = \frac{1}{n-2} \sum [y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_i)]^2 = \frac{Q_e}{n-2}.$$

Остаточная дисперсия S^2 используется для оценки обычно неизвестной дисперсии ошибок наблюдений σ^2.

Если модель согласуется с результатами наблюдений (адекватна результатам наблюдений), то остаточная дисперсия является ***несмещенной оценкой дисперсии ошибок наблюдений*** σ^2, т. е. $M[S^2] = \sigma^2$.

Всюду в дальнейшем будем предполагать, что ошибки наблюдений ε_i, $i = 1, 2, ..., n$ имеют нормальное распределение: $\varepsilon_i \sim N(0, \sigma^2)$ и независимы.

Это предположение эквивалентно тому, что результаты наблюдений y_i, $i = 1, 2, ..., n$ являются реализациями независимых нормально распределенных случайных величин Y_i:

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, 2, ..., n.$$

Оценки $\tilde{\beta}_0$ и $\tilde{\beta}_1$ параметров линейной регрессии, вычисленные по формулам (1) и (2), как линейные функции Y_i, $i = 1, 2, ..., n$ также будут случайными величинами, имеющими нормальное распределение:

$$\tilde{\beta}_0 \sim N(\beta_0, D[\tilde{\beta}_0]),$$

$$\tilde{\beta}_1 \sim N(\beta_1, D[\tilde{\beta}_1]),$$

где оценки дисперсии $\tilde{\beta}_0$ и $\tilde{\beta}_1$ соответственно равны:

$$D[\tilde{\beta}_0] = \frac{\sigma^2(\sum x_i^2)}{nQ_x}, \quad D[\tilde{\beta}_1] = \frac{\sigma^2}{Q_x}.$$

Если линейная регрессия адекватна результатам наблюдений, то можно показать [13, 22] , что статистика Q_e/σ^2 имеет распределение χ^2 с $(n-2)$ степенями свободы, т. е.

$$\frac{Q_e}{\sigma^2} = \chi^2(n-2).$$

Распределение статистики $\frac{Q_e}{\sigma^2}$ не зависит от распределений оценок $\tilde{\beta}_0$ и $\tilde{\beta}_1$. Несмещенная оценка дисперсии ошибок наблюдений S^2 связана с распределением $\chi^2(n-2)$ следующим соотношением

$$\frac{S^2}{\sigma^2} = \frac{\chi^2(n-2)}{n-2}. \tag{3}$$

Используя приведенные выше результаты можно вычислить доверительные интервалы для параметров линейной регрессии.

Рассмотрим нормированную статистику

$$\frac{\tilde{\beta}_i - \beta_i}{\sqrt{\frac{S^2}{\sigma^2} D[\tilde{\beta}_i]}}, \quad i = 0,1. \tag{4}$$

Преобразуем статистику (4), используя соотношение (3):

$$\frac{\widetilde{\beta}_i - \beta_i}{\sqrt{\frac{S^2}{\sigma^2} D[\widetilde{\beta}_i]}} = \frac{\frac{\widetilde{\beta}_i - \beta_i}{\sqrt{D[\widetilde{\beta}_i]}}}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\frac{\widetilde{\beta}_i - \beta_i}{\sqrt{D[\widetilde{\beta}_i]}}}{\sqrt{\frac{\chi^2(n-2)}{n-2}}} = T(n-2).$$

Таким образом, статистика (4) может быть представлена, как отношение двух независимых случайных величин: в числителе — случайная величина, имеющая стандартное нормальное распределение $N(0, 1)$, а в знаменателе — $\sqrt{\frac{\chi^2(n-2)}{n-2}}$ и, следовательно, имеет распределение Стьюдента с $(n-2)$ степенями свободы (см. п. 3.2.2 главы 3). Используя этот результат, получаем, что интервалы для параметров линейной регрессии вычисляются по формулам:

$$\widetilde{\beta}_0 \pm t_{1-\frac{\alpha}{2}}(n-2) \cdot S\sqrt{\frac{\sum x_i^2}{nQ_x}} \text{ или } \widetilde{\beta}_0 \pm t_{1-\frac{\alpha}{2}}(n-2)\sqrt{D[\widetilde{\beta}_0]};$$

$$\widetilde{\beta}_1 \pm t_{1-\frac{\alpha}{2}}(n-2) \cdot S\sqrt{\frac{1}{Q_x}} \text{ или } \widetilde{\beta}_1 \pm t_{1-\frac{\alpha}{2}}(n-2)\sqrt{D[\widetilde{\beta}_1]},$$

где $t_{1-\frac{\alpha}{2}}(n-2)$ — квантиль распределения Стьюдента с $(n-2)$ степенями свободы порядка $1-\frac{\alpha}{2}$; S — оценка среднего квадратического ошибок наблюдений: $S = \sqrt{\frac{Q_e}{n-2}}$, $D[\widetilde{\beta}_i]$ — дисперсия оценки параметра $\widetilde{\beta}_i$, $i = 0,1$.

Доверительный интервал для дисперсии ошибок наблюдений σ^2 имеет вид

$$\frac{(n-2)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-2)} < \sigma^2 < \frac{(n-2)S^2}{\chi^2_{\frac{\alpha}{2}}(n-2)},$$

где $\chi^2_p(n-2)$ — квантили распределения χ^2 с $(n-2)$ степенями свободы порядка p, а S^2 — оценка дисперсии ошибок наблюдений.

В практических вычислениях остаточную сумму квадратов Q_e получают из тождества

$$\sum(y_i - \bar{y})^2 = \sum(\widetilde{y}_i - \bar{y})^2 + \sum(y_i - \widetilde{y}_i)^2,$$

которое записывается в виде

$$Q_y = Q_R + Q_e,$$

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

$$Q_R = \sum(\widetilde{y}_i - \bar{y})^2 = \widetilde{\beta}_1 \cdot Q_{xy} = \widetilde{\beta}_1^2 Q_x = \frac{Q_{xy}^2}{Q_x}.$$

Вывод тождества $Q_y = Q_R + Q_e$ приведен ниже в п. 6.3.3.

Величина Q_R называется ***суммой квадратов, обусловленной регрессией.***

Линейная регрессионная модель называется ***незначимой***, если параметр $\beta_1 = 0$. В этом случае, так как между зависимой переменной y и фактором x линейной зависимости нет, либо она выражена очень слабо (см. рис. 6.1, *в*), рассматривать линейную регрессию нет смысла. Для проверки гипотезы $H_0: \beta_1 = 0$ используют либо доверительный интервал для параметра β_1 либо статистику Фишера F:

$$F = \frac{Q_R(n-2)}{Q_e} = \frac{\tilde{\beta}_1^2 Q_x}{S^2}.$$

Если гипотеза $H_0: \beta_1 = 0$ верна, то статистика F имеет распределение Фишера с 1 и $(n-2)$ степенями свободы.

Гипотеза $H_0: \beta_1 = 0$ принимается на уровне значимости α, если выборочное значение статистики Фишера $F_в$ будет меньше квантили распределения Фишера $F_{1-\alpha}(1, n-2)$, т. е. $F_в < F_{1-\alpha}(1, n-2)$. В противном случае, гипотеза H_0 отклоняется.

В случае, когда гипотеза $H_0: \beta_1 = 0$ *отклоняется*, говорят, что регрессионная модель ***статистически значима.*** Из этого не следует, конечно, что модель хорошо согласуется с результатами наблюдений, т. е. адекватна им.

Полезной характеристикой линейной регрессии является ***коэффициент детерминации*** R^2, вычисляемый по формуле

$$R^2 = \frac{Q_R}{Q_y} = 1 - \frac{Q_e}{Q_y}.$$

Коэффициент детерминации R^2 равен той доле разброса результатов наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$ относительно горизонтальной прямой $y = \bar{y}$, которая объясняется регрессионной моделью. Величина $R = +\sqrt{R^2}$ является оценкой коэффициента корреляции между результатами наблюдений y_i и вычисляемыми значениями $\tilde{y}_i$, предсказываемыми регрессией, т. е.

$$R = \tilde{\rho}_{Y\tilde{Y}} = r_{y\tilde{y}}.$$

В случае простой линейной регрессии Y на x (одной независимой переменной x) между коэффициентом R и выборочным коэффициентом корреляции r_{xy} имеется следующее соотношение

$$r_{xy} = (\text{знак}\,\tilde{\beta}_1) R.$$

Напомним, что коэффициент корреляции ρ_{xy} определяет степень линейной зависимости между случайными величинами X и Y.

6.1.3. Проверка выполнения предположений регрессионного анализа по остаткам. Доверительные интервалы для прогноза

Линейная регрессионная модель называется ***адекватной***, если предсказанные по ней значения переменной Y согласуются с результатами наблюдений. Грубая оценка адекватности модели может быть проведена непо-

средственно по графику остатков, т. е. разностей между наблюдаемыми значениями y_i и вычисленными по уравнению регрессии Y на x значениями $\tilde{y}_i$, $i = 1, 2, ..., n$. Если модель адекватна, то остатки e_i являются реализациями случайных ошибок наблюдений ε_i, $i = 1, 2, ..., n$, которые, в силу предположений, должны быть независимыми нормально распределенными случайными величинами с нулевыми средними и равными дисперсиями σ^2. Всякое отклонение от предложений относительно ошибок наблюдений ε_i должно отразиться на поведении остатков e_i, $i = 1, 2, ..., n$. Различные графики остатков дают возможность определить те или иные отклонения [4, 18].

График стандартизированных остатков $d_i = e_i / S$, где S — оценка стандартного отклонения ошибок наблюдений в функции предсказанного значения зависимой переменной $\tilde{y}_i$, $i = 1, 2, ..., n$, позволяет обнаружить следующие дефекты регрессионной модели и исходных данных.

1) Наличие выбросов, т. е. таких остатков, которые по абсолютному значению значительно превосходят все остальные остатки d_i. Например такие остатки d_i, для которых $| d_i | > 3$.

2) Нарушение условия постоянства дисперсии ошибок для всех наблюдений. Если все остатки укладываются в симметричную относительно нулевой линии полосу, то дисперсии ошибок наблюдений можно считать постоянными.

3) Криволинейный характер графика остатков показывает, что в регрессионной модели не учтены факторы, оказывающие существенное влияние на зависимую переменную Y.

График стандартизированных остатков d_i в функции номера наблюдения i, $i = 1, 2, ..., n$, (что совпадает в некоторых задачах с графиком по времени) может показывать наличие корреляции между последовательными значениями d_i или указывать на непостоянство дисперсии ошибок наблюдений (если остатки не укладываются в симметричную относительно нулевой линии полосу).

В случае если остатки на этом графике лежит в пределах полосы постоянной ширины, но имеют линейный или криволинейный тренд, то в регрессионную модель необходимо включить фактор зависящий от номера наблюдения (или времени).

Если ошибки наблюдений ε_i, $i = 1, 2, ..., n$ имеют нормальное распределение $N(0, \sigma^2)$, то и остатки e_i также должны иметь нормальное распределение. Гипотезу о нормальном распределении остатков при достаточно большом объеме выборки n можно проверить с помощью критерия χ^2 (см. главу 3, п. 3.3) или критерия Колмогорова—Смирнова. В статистических пакетах проверка выполнения этого условия обычно выполняется на специальном графике — вероятностной бумаге [14, 19]. Нормально распределенные остатки укладываются на прямую. Точки значительно удаленные от прямой можно рассматривать как выбросы. Если появление выбросов объясняется грубыми ошибками в исходных данных, то они должны быть исключены из дальнейшего анализа. В противном случае выбросы могут указывать на неадекватность модели.

Нарушение предположения о некоррелированности ошибок наблюдений ε_i, $i = 1, 2, ..., n$ приводит к тому, что в последовательности остатков e_i обнаруживается сериальная корреляция, т. е. корреляция между остатками e_i, $i = 1, 2, ..., n$, отстоящими друг от друга на k шагов. Наличие сериальной корреляции в последовательности остатков проверяется с помощью критерия Дарбина—Уотсона.

Статистика критерия d вычисляется по формуле

$$d = \sum_{i=2}^{n} \frac{(e_i - e_{i-1})^2}{Q_e}.$$

Критерий Дарбина—Уотсона позволяет проверить гипотезу H_0: все сериальные корреляции равны 0, $\rho_k = 0$, $k = 1, 2, ...$ при альтернативной гипотезе H_1: $\rho_k = \rho^k$, $\rho \neq 0$, $|\rho| < 1$.

Процедура проверки состоит в следующем. В зависимости от числа наблюдений n, числа оцениваемых параметров k модели и уровня значимости α по таблице (см. Приложение 2) находят два числа d_1 и d_2. В зависимости от формулировки альтернативной гипотезы H_1 решение принимается по одному из следующих правил:

1) H_1: $\rho > 0$:

H_0 принимается, если $d > d_2$,

H_0 отклоняется, если $d < d_1$,

при $d_1 \le d \le d_2$ решение не принимается;

2) H_1: $\rho < 0$:

H_0 принимается, если $4 - d > d_2$,

H_0 отклоняется, если $4 - d < d_1$,

при $d_1 \le 4 - d \le d_2$ решение не принимается;

3) H_1: $\rho \neq 0$:

H_0 принимается на уровне значимости 2α, если $d > d_2$ или $4 - d > d_2$,

H_0 отклоняется на уровне значимости 2α, если $d < d_1$ или $4 - d < d_1$.

Если гипотеза H_0 отклоняется, то либо ошибки наблюдений в исходных данных коррелированны (в этом случае для оценки параметров нужно применять другие методы, например взвешенный метод наименьших квадратов [33]), либо в модели не учтен один или несколько существенных факторов, влияющих на зависимую переменную, либо неправильно выбрана форма связи между переменными.

Более тщательная проверка адекватности регрессионной модели может быть проведена, если для зависимой переменной Y проведены повторные наблюдения. В этом случае для проверки адекватности модели используется следующая процедура дисперсионного анализа.

Пусть при i-м значении фактора x, $x = x_i$, проведено n_i повторных наблюдений зависимой переменной Y, $i = 1, 2, ..., m$. Объем всей выборки $n = \sum_{i=1}^{m} n_i$. Обозначим y_{ij}, $j = 1, 2, ..., n_i$, результаты повторных наблюдений Y при i-м значении фактора x, $x = x_i$. Если модель адекватна данным, то

средние $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, $i = 1, 2, \ldots, m$ должны быть близки к значениям $\tilde{y}_i$, предсказанным регрессионной моделью

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i.$$

Мерой неадекватности модели будет сумма квадратов отклонений $(\bar{y}_i - \tilde{y}_i)$:

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

Чем меньше Q_n, тем лучше результаты наблюдений согласуются с моделью. Возведя обе части тождества $y_{ij} - \tilde{y}_i = (\bar{y}_i - \tilde{y}_i) + (y_{ij} - \bar{y}_i)$ в квадрат и просуммировав их по i и по j, получим, что остаточная сумма квадратов Q_e может быть разбита на две суммы Q_n и Q_p:

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

или

$$Q_e = Q_n + Q_p.$$

Второе слагаемое Q_p называется *суммой квадратов чистой ошибки.*

Если модель адекватна результатам наблюдений, то статистики $\frac{Q_n}{\sigma^2}$ и $\frac{Q_p}{\sigma^2}$ независимы и имеют распределение χ^2 соответственно с $(m - k)$ и $(n - m)$ степенями свободы, где k — число оцениваемых параметров. Для простой линейной регрессии число оцениваемых параметров, $k = 2$.

В этом случае статистика

$$F = \frac{Q_n/(m-2)}{Q_p/(n-m)}$$

имеет распределение Фишера с $(m - 2)$ и $(n - m)$ степенями свободы.

Выборочное значение статистики F, $F_в$ сравнивается с квантилью распределения Фишера $F_{1-\alpha}(m-2, n-m)$. Если $F_в < F_{1-\alpha}(m-2, n-m)$, то гипотеза об адекватности модели принимается на уровне значимости α. В противном случае модель не адекватна результатам наблюдений.

Если регрессионная модель значима и адекватна результатам наблюдений, то она может быть использована для определения прогноза $\tilde{y}(x_0)$ при заданном значении независимой переменной $x = x_0$.

Доверительный интервал для *среднего* значения Y при $x = x_0$ определяется по формуле

$$\tilde{y}(x_0) \pm t_{1-\alpha/2}(n-2) \cdot S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Q_x}},$$

а доверительный интервал для *индивидуального* значения Y при $x = x_0$ вычисляется по формуле:

$$\tilde{y}(x_0) \pm t_{1-\alpha/2}(n-2) \cdot S\sqrt{1 + 1/n + \frac{(x_0 - \bar{x})^2}{Q_x}},$$

где $\tilde{y}(x_0) = \tilde{\beta}_0 + \tilde{\beta}_1 x_0$.

Задачу регрессионного анализа удобно записывать в матричном виде. Введем следующие обозначения:

регрессионная матрица $(n \times 2)$ $A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, вектор $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$,

вектор параметров модели $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$,

вектор ошибок наблюдений $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$.

Тогда простая линейная регрессия определяется матричным уравнением

$$Y = A\beta + \varepsilon.$$

Метод наименьших квадратов дает оценку β, вычисляемую по формуле

$$\tilde{\beta} = (A^T A)^{-1} A^T Y, \tag{5}$$

где A^T — матрица, транспонированная к матрице A; $A^T A = B$ — информационная матрица; $B^{-1} = (A^T A)^{-1}$ — матрица, обратная к матрице $B = (A^T A)$. Вывод этой формулы приводится ниже (см. п. 6.3).

Сумма квадратов, обусловленная регрессией, определяется по формуле

$$Q_R = \beta^T A^T Y - n(\bar{y})^2.$$

Остаточная сумма квадратов: $Q_e = Q_y - Q_R$.

Ковариационная матрица K для оценок параметров регрессии вычисляется по формуле

$$K = S^2 (A^T A)^{-1} = S^2 B^{-1}.$$

Дисперсии оценок параметров — диагональные элементы матрицы K:

$$D[\tilde{\beta}_0] = S^2(b_{11}),$$

$$D[\tilde{\beta}_1] = S^2(b_{22}),$$

где b_{ii}, $i = 1, 2$ — диагональные элементы матрицы B^{-1}.

6.2. Практические задания

6.2.1. Работа 1. Простая линейная регрессия

1. Основные понятия

Диаграмма рассеяния. Зависимые переменные и факторы. Предположения регрессионного анализа. Простая линейная регрессия Y на x и X на y. Параметры регрессии и их оценка методом наименьших квадратов (МНК). Остатки. Остаточная сумма квадратов. Оценка дисперсии ошибок наблюдений. Коэффициент детерминации. Оценка коэффициента корреляции.

2. Варианты 1—25 заданий для работ 1 и 2

Каждый вариант содержит пять значений фактора x и зависимой переменной y.

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7
–2,2	–4,0	–0,5	3,3	1,2	2,8	2,8	4,0	1,4	–0,7	–1,0	–1,2	2,7	1,0
–0,1	0,2	0,9	0,5	–3,0	–1,1	0,6	4,1	–2,3	3,9	2,5	2,4	0,2	2,8
3,1	5,4	1,5	0,0	–0,4	3,0	3,0	2,9	0,2	1,6	3,3	5,3	–1,2	2,9
–0,2	0,7	0,6	1,0	2,3	4,3	–1,6	5,9	4,8	–2,7	2,2	3,9	–0,5	3,2
1,0	3,5	–0,2	1,7	–0,3	1,3	–0,7	5,5	1,2	–0,6	2,0	0,5	–0,7	2,5
x_8	y_8	x_9	y_9	x_{10}	y_{10}	x_{11}	y_{11}	x_{12}	y_{12}	x_{13}	y_{13}	x_{14}	y_{14}
3,3	3,8	–0,2	4,5	3,5	2,0	4,2	1,5	1,8	7,1	6,2	2,5	5,3	3,5
1,1	3,7	0,8	6,2	1,5	–0,6	5,3	1,8	4,3	7,8	3,1	2,0	2,8	0,7
–1,4	–1,1	–1,2	3,2	3,5	4,7	1,4	2,5	0,0	1,8	1,7	0,6	3,0	2,1
2,7	3,5	–0,5	2,9	1,8	–0,7	0,3	2,9	1,3	2,3	8,0	3,6	3,5	1,3
0,8	2,5	1,0	5,3	–0,3	0,4	2,0	2,8	3,2	4,7	6,1	1,5	3,0	2,5
x_{15}	y_{15}	x_{16}	y_{16}	x_{17}	y_{17}	x_{18}	y_{18}	x_{19}	y_{19}	x_{20}	y_{20}	x_{21}	y_{21}
5,6	1,0	3,7	2,8	9,2	0,5	1,7	2,3	4,8	6,8	4,4	4,7	7,1	5,4
6,1	2,5	3,7	2,5	1,2	7,8	4,9	5,0	3,0	7,8	3,8	6,4	4,5	6,5
6,1	2,2	3,9	1,3	4,3	5,3	0,6	2,5	5,8	6,1	2,4	4,3	8,3	4,6
5,6	5,5	3,9	1,3	6,5	4,5	8,0	7,1	4,5	6,8	3,4	4,3	6,7	5,4
2,7	6,0	1,9	3,4	4,5	4,2	1,0	3,1	6,5	4,5	2,3	2,7	10,4	4,0
x_{22}	y_{22}	x_{23}	y_{23}	x_{24}	y_{24}	x_{25}	y_{25}						
2,7	5,0	6,6	1,7	6,7	2,9	5,5	4,0						
8,5	–0,5	6,9	2,5	5,5	3,2	8,1	5,6						
5,0	3,7	3,4	5,7	8,2	6,1	8,5	5,7						
6,8	0,7	7,7	1,1	4,8	2,7	5,9	3,6						
4,6	3,2	8,3	1,9	3,1	1,4	7,8	4,0						

3. Задание

По выборке из своего варианта выполнить следующие расчеты и задания:

1. Построить диаграмму рассеяния выборки (построение сделать точно на бумаге в клеточку или миллиметровке).

2. Вычислить оценки параметров линейной регрессии Y на x: $y = \beta_0 + \beta_1 x$ и X на y: $x = \beta'_0 + \beta'_1 y$, используя суммы квадратов Q_y, Q_x, Q_{xy} (формулы (1), (2)).

3. Нанести графики прямых регрессий Y на x и X на y на диаграмму рассеяния.

4. Для линейной регрессии Y на x вычислить остатки e_i, $i = 1, 2, ..., n$; остаточную сумму квадратов $Q_e = \sum e_i^2$; оценку дисперсии ошибок наблюдений S^2, коэффициент детерминации R^2 и оценку коэффициента корреляции r.

5. Для линейной регрессии $y = \beta_0 + \beta_1 x$ выписать матрицу A, транспонированную матрицу A^T, информационную матрицу $B = A^T A$ и обратную матрицу к матрице $B = (A^T A)$. Найти оценки β_0 и β_1, используя формулу для расчета оценок в матричном виде (формула (5)). Сравнить результаты вычислений с оценками, полученными в п. 2.

6. Ввести данные в пакет STATISTICA, выполнить п. 1—4, сравнить результаты расчетов и полученные графики, записать в отчет результаты.

Чтобы показать технику вычислений, рассмотрим пример расчета простой линейной регрессии с небольшим объемом данных. Читатель, знакомый с техникой вычислений, может пропустить выкладки и перейти сразу к решению в пакете STATISTICA.

Пример 6.1. Пример простой линейной регрессии Y на x. Исходные данные: результаты наблюдений зависимой переменной (y) и фактора (x) следующие:

y	x
4,0	5,5
5,6	8,1
5,7	8,5
3,6	5,9
4,0	7,8

Решение.

1. По данным примера вычислим суммы квадратов Q_y, Q_x и сумму произведений Q_{xy}; $n = 5$. Предварительно найдем средние значения:

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{5}(5{,}5 + 8{,}1 + 8{,}5 + 5{,}9 + 7{,}8) = 7{,}16;$$

$$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{5}(4 + 5{,}6 + 5{,}7 + 3{,}6 + 4) = 4{,}58;$$

$$Q_x = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n(\bar{x})^2 = (5{,}5^2 + 8{,}1^2 + 8{,}5^2 + 5{,}9^2 + 7{,}8^2) -$$
$$- 5 \cdot (7{,}16)^2 = 263{,}76 - 5 \cdot 51{,}266 = 7{,}432;$$

$$Q_y = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2 = (4^2 + 5{,}6^2 + 5{,}7^2 + 3{,}6^2 + 4^2) -$$
$$- 5 \cdot (4{,}58)^2 = 108{,}81 - 5 \cdot 20{,}976 = 3{,}928;$$

$$Q_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \cdot \bar{x} \cdot \bar{y} = (5{,}5 \cdot 4 + 8{,}1 \cdot 5{,}6 + 8{,}5 \cdot 5{,}7 +$$
$$+ 5{,}9 \cdot 3{,}6 + 7{,}8 \cdot 4) - 5 \cdot 7{,}16 \cdot 4{,}58 = 168{,}25 - 5 \cdot 7{,}16 \cdot 4{,}58 = 4{,}289.$$

Оценки параметров линейной регрессии $y = \beta_0 + \beta_1 x$, по формулам (1) и (2) равны:

$$\tilde{\beta}_1 = \frac{Q_{xy}}{Q_x} = \frac{4{,}289}{7{,}432} \approx 0{,}577;$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \cdot \bar{x} = 4{,}58 - 0{,}577 \cdot 7{,}16 \approx 0{,}451.$$

Таким образом, уравнение линейной регрессии Y на x имеет вид

$$y = 0{,}451 + 0{,}577x.$$

Аналогично, оценки параметров линейной регрессии X на y:

$$\tilde{\beta}_1' = \frac{Q_{xy}}{Q_y} \approx 1{,}091;\ \tilde{\beta}_0' = \bar{x} - \tilde{\beta}_1' \bar{y} = 7{,}16 - 1{,}091 \cdot 4{,}58 \approx 2{,}163.$$

Уравнение линейной регрессии X на y имеет вид

$$x = 2{,}163 + 1{,}091y.$$

2. Диаграмма рассеяния исходных данных и прямая регрессии Y на x показана на рис. 6.2.

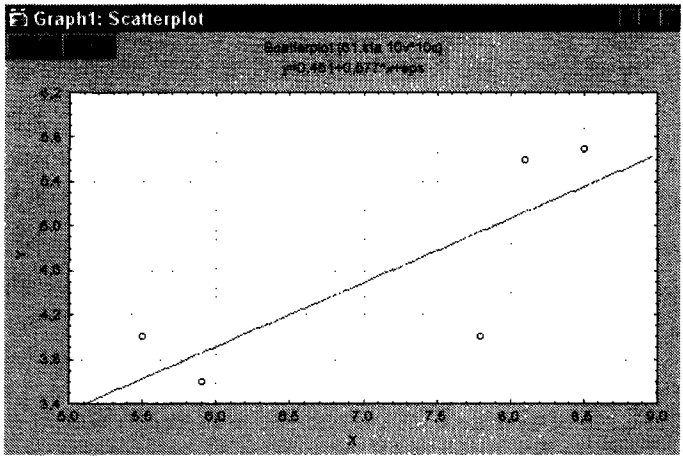


Рис. 6.2. Диаграмма рассеяния и прямая регрессии Y на x

3. Для линейной регрессии Y на x вычислим остатки:

$$e_i = y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_i),\ i = 1, 2, \ldots, 5;$$

$$e_1 = 4 - (0{,}451 + 0{,}577 \cdot 5{,}5) = 0{,}377;$$

$$e_2 = 5{,}6 - (0{,}451 + 0{,}577 \cdot 8{,}1) = 0{,}478;$$

.....

$$e_5 = 4 - (0{,}451 + 0{,}577 \cdot 7{,}8) = -0{,}949.$$

Остаточная сумма квадратов Q_e:

$$Q_e = (0{,}377)^2 + (0{,}478)^2 + (0{,}35)^2 + (-0{,}25)^2 + (-0{,}949)^2 \approx 1{,}457.$$

Оценка дисперсии ошибок наблюдений

$$S^2 = \frac{Q_e}{n-k} = \frac{1{,}457}{5-2} \approx 0{,}486,$$

где k — число оцениваемых параметров; для простой линейной регрессии $k = 2$.

Коэффициент детерминации R^2:

$$R^2 = 1 - \frac{Q_e}{Q_y} = 1 - \frac{1{,}457}{3{,}928} \approx 0{,}629.$$

Оценка коэффициента корреляции r:

$$r = \frac{Q_{xy}}{\sqrt{Q_x \cdot Q_y}} = \frac{4{,}286}{\sqrt{7{,}438 \cdot 3{,}928}} \approx 0{,}793.$$

4. Вычислим оценки параметров линейной регрессии Y на x в матричном виде, используя формулу (5):

$$\tilde{\beta} = (A^T A)^{-1} A^T Y,$$

где $\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix}$; A — регрессионная матрица: $A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \ldots & \ldots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} 1 & 5{,}5 \\ 1 & 8{,}1 \\ 1 & 8{,}5 \\ 1 & 5{,}9 \\ 1 & 7{,}8 \end{pmatrix}$;

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix} = \begin{pmatrix} 4 \\ 5{,}6 \\ 5{,}7 \\ 3{,}6 \\ 4 \end{pmatrix}.$$

Последовательно вычисляем:

$$A^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 5{,}5 & 8{,}1 & 8{,}5 & 5{,}9 & 7{,}8 \end{pmatrix},$$

$$B = A^T A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 5{,}5 & 8{,}1 & 8{,}5 & 5{,}9 & 7{,}8 \end{pmatrix} \cdot \begin{pmatrix} 1 & 5{,}5 \\ 1 & 8{,}1 \\ 1 & 8{,}5 \\ 1 & 5{,}9 \\ 1 & 7{,}8 \end{pmatrix} = \begin{pmatrix} 5 & 35{,}8 \\ 35{,}8 & 263{,}76 \end{pmatrix}.$$

Определитель матрицы B:

$$|B| = \det(A^T A) = 37{,}16.$$

Обратная матрица к матрице B:

$$B^{-1} = \frac{1}{|B|} \cdot B^* = \frac{1}{37{,}16} \cdot \begin{pmatrix} 263{,}76 & -35{,}8 \\ -35{,}8 & 5 \end{pmatrix} = \begin{pmatrix} 7{,}098 & -0{,}963 \\ -0{,}963 & 0{,}135 \end{pmatrix},$$

где B^* — присоединенная матрица к матрице B, составленная из алгебраических дополнений к элементам матрицы B.

Далее вычисляем произведения матриц

$$B^{-1} \cdot A^T = \begin{pmatrix} 7{,}098 & -0{,}963 \\ -0{,}963 & 0{,}135 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 5{,}5 & 8{,}1 & 8{,}5 & 5{,}9 & 7{,}8 \end{pmatrix} =$$

$$= \begin{pmatrix} 1{,}7992 & -0{,}7056 & -1{,}0910 & 1{,}4139 & -0{,}4166 \\ -0{,}2234 & 0{,}1265 & 0{,}1803 & -0{,}1695 & 0{,}0861 \end{pmatrix}.$$

Окончательно

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix} = B^{-1} A^T Y =$$

$$= \begin{pmatrix} 1{,}7992 & -0{,}7056 & -1{,}0910 & 1{,}4139 & -0{,}4166 \\ -0{,}2234 & 0{,}1265 & 0{,}1803 & -0{,}1695 & 0{,}0861 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 5{,}6 \\ 5{,}7 \\ 3{,}6 \\ 4 \end{pmatrix} = \begin{pmatrix} 0{,}4509 \\ 0{,}5767 \end{pmatrix}.$$

Сравнивая полученные значения с результатами в п. 2 видно, что расхождение имеется только в третьем десятичном знаке.

5. Выполнение задания в пакете STATISTICA.

Решение. 1. Откройте новый файл данных. В таблице удалите ненужные столбцы (**Var-Delete**) и строки наблюдений (**Cases-Delete**). Дайте имена переменным: Y — зависимая переменная (**Dependent**), X — фактор (независимая переменная — **Independent**). В ячейки таблицы введите данные.

2. Построим график исходных данных. Для этого можно воспользоваться меню **Graphs — графики** и выбрать необходимый тип графика. В нашем примере мы воспользуемся двумерными диаграммами рассеяния (**Stats 2D Graphs → Scatterplots**).

В диалоговом окне при помощи кнопки **Variables — Переменные** выберите необходимые переменные, которые вы хотите отобразить графически и необходимый тип графика. Диаграмма рассеяния с прямой регрессии Y на x показана на рис. 6.2.

3. В **Переключателе модулей (STATISTICA Module Switcher)** выберите модуль Множественная регрессия (Multiple Regression). После запуска модуля на экране откроется стартовая панель модуля (рис. 6.3). Далее выберите переменные для анализа (воспользуйтесь кнопкой **Variables**). В качестве зависимой переменной (***Dependent***) выберите Y, в качестве независимой (***Independent***) — X. После определения зависимых и независимых переменных на стартовой панели нажмите **OK**. Появится окно с результатами вычислений (рис. 6.4.).

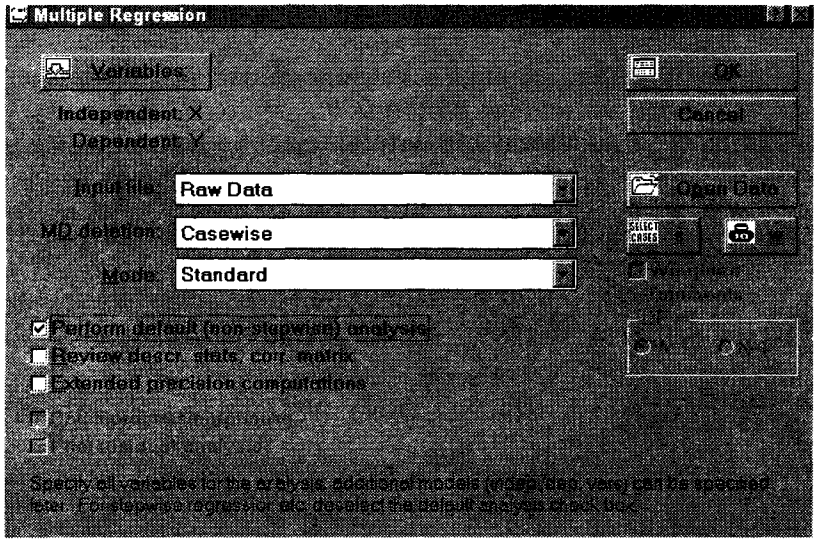


Рис. 6.3. Стартовая панель модуля Multiple Regression

В диалоговом окне **Результаты Множественной регрессии — Multiple Regression Results** просмотрите результаты оценивания. Результаты можно просмотреть в численном и графическом виде. Окно результатов анализа имеет следующую структуру: верх окна — информационный. Он состоит из двух частей: в первой части содержится основная информация о результатах оценивания, во второй высвечиваются значимый *стандартизованный* регрессионный коэффициент X-beta=,793; стандартизованный коэффициент регрессии вычисляется по формуле

$$\tilde{\beta}_{01} = \tilde{\beta}_1 \cdot (s_x / s_y),$$

где s_x и s_y — оценки среднеквадратических отклонений для переменных x и y.

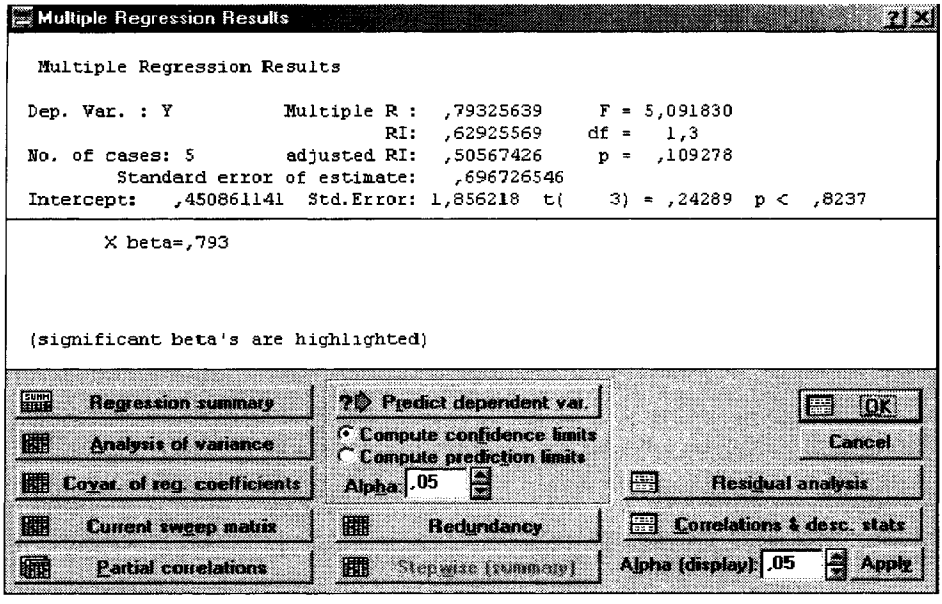


Рис. 6.4. Окно результатов множественной регрессии

Внизу окна **Результаты множественной регрессии** находятся функциональные кнопки, позволяющие всесторонне рассмотреть результаты анализа.

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

— ***Dep.Var.*** — имя зависимой переменной (Y);

— ***No. of Cases*** — число наблюдений (объем выборки n), по которым построена регрессия ($n = 5$);

— ***Multiple R*** — коэффициент множественной корреляции (описывает степень линейной зависимости между Y и факторами); в случае простой линейной регрессии равен модулю коэффициента корреляции;

— ***R — square — RI*** — квадрат коэффициента множественной корреляции (коэффициент детерминации). Если регрессионная модель значима, то коэффициент детерминации равен той доле дисперсии ошибок наблюдений, которая объясняется регрессионной моделью.

Коэффициент детерминации, вычисляется по формуле

$$R^2 = 1 - \frac{Q_e}{Q_y};$$

— ***Adjusted R-square: adjusted RI*** — скорректированный коэффициент детерминации

$$R_1^2 = 1 - \frac{Q_e/(n-k)}{Q_y/(n-1)},$$

где n — число наблюдений, а k — число оцениваемых параметров регрессионной модели; для простой линейной регрессии $k = 2$, так как определяются оценки двух параметров β_0 и β_1;

— ***Std. Error of estimate*** — среднее квадратическое отклонение ошибок наблюдений

$$S = \sqrt{S^2} = \sqrt{\frac{Q_e}{(n-k)}};$$

— ***Intercept*** — оценка свободного члена регрессии ($\tilde{\beta}_0$);

— ***Std. Error*** — стандартная ошибка оценки свободного члена $\sqrt{D[\tilde{\beta}_0]}$;

— ***t(n – k) and p-value*** — выборочное значение t-статистики и вычисленного уровня значимости p.

t-статистика используется для проверки гипотезы $H_0 : \beta_0 = 0$:

$$t = \frac{\tilde{\beta}_0}{\sqrt{D[\tilde{\beta}_0]}}.$$

Уровень значимости $p = P[T(n-k) > |t_в|]$, где $T(n-k)$ — случайная величина, имеющая распределение Стьюдента с $(n-k)$ степенями свободы, $t_в$ — выборочное значение t-статистики.

Если $p > \alpha$, где α — заданный уровень значимости, то гипотеза $H_0 : \beta_0 = 0$ принимается.

В данном случае $p = 0{,}823749$, следовательно гипотеза $H_0 : \beta_0 = 0$ принимается.

— ***F*** — выборочное значение F-статистики, $F_в$.

F-статистика используется для проверки гипотезы $H_0 : \beta_1 = 0$.

Если гипотеза $H_0 : \beta_1 = 0$ верна, то статистика F имеет распределение Фишера с $(k-1)$ и $(n-k)$ степенями свободы.

Гипотеза H_0 принимается на уровне значимости α, если выборочное значение статистики F, $F_в$, меньше $F_{1-\alpha}(k-1, n-k)$ — квантили распределения Фишера порядка $1-\alpha$. Если гипотеза $H_0 : \beta_1 = 0$ принимается, то *регрессионная модель незначима.*

— ***df*** — число степеней свободы F-статистики: $(k-1; n-k)$.

— ***p*** — вычисленный уровень значимости.

Вычисленный уровень значимости p: $p = P[F(k-1; n-k) > F_в]$, где $F_в$ — выборочное значение F-статистики.

Если $p < \alpha$, то гипотеза $H_0 : \beta_1 = 0$ отклоняется; если $p > \alpha$, то гипотеза $H_0 : \beta_1 = 0$ принимается.

В данном примере $p = 0{,}109278$, следовательно гипотеза $H_0 : \beta_1 = 0$ принимается на уровне значимости $\alpha = 0{,}05$. *Регрессионная модель незначима.*

Функциональные кнопкн. При нажатии кнопки **Regression Summary — Результаты регрессии** на экране появится следующая таблица с результатами анализа (рис. 6.5.):

Во втором столбце таблицы (**BETA**) выводится стандартизованный коэффициент регрессии β_{01}:

$$\beta_{01} = \tilde{\beta}_1 \cdot (s_x / s_y),$$

где s_x и s_y — оценки среднеквадратических отклонений для переменных x и y.

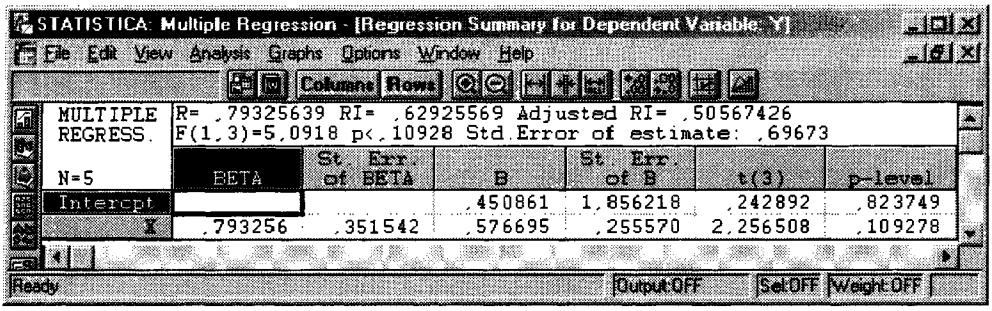


Рис. 6.5. Результаты регрессии

Стандартизированные коэффициенты регрессии — безразмерные величины.

В случае множественной регрессии стандартизованные коэффициенты регрессии используются для сравнения влияния на зависимую переменную факторов, имеющих различную размерность.

В четвертом столбце (*B*) приведены МНК-оценки коэффициентов регрессии: $\widetilde{\beta}_0$ и $\widetilde{\beta}_1$.

В пятом столбце (***St.Err. of B***) — их стандартные отклонения $S_{\widetilde{\beta}_i} = \sqrt{D[\widetilde{\beta}_i]}$; $i = 0, 1$.

В шестом столбце — t-статистики для проверки гипотезы $H_0 : \beta_i = 0$:

$$t_i = \frac{\widetilde{\beta}_i}{\sqrt{D[\widetilde{\beta}_i]}}; \; i = 0, 1.$$

В седьмом столбце — соответствующие уровни значимости

$$p = P[T(n-k) > |t_i|].$$

В данном случае гипотеза $H_0 : \beta_1 = 0$ принимается на уровне значимости $\alpha = 0{,}05$.

Вычисленный уровень значимости $p > \alpha$. Это означает, что регрессионная модель незначима. Гипотеза $H_0 : \beta_0 = 0$ также принимается при $\alpha = 0{,}05$.

Чтобы просмотреть и проанализировать остатки, войдите в меню **Residual Analysis** (анализ остатков), нажав соответствующую кнопку в нижней правой части панели результатов вычислений (рис. 6.4).

Это меню представлено на рис. 6.6.

Чтобы просмотреть остатки и их график, нажмите в левой нижней части этого меню кнопку **Plots of residuals(A)** (графики остатков (A)). Выбрав опцию **Raw residuals** (значения остатков), получим график остатков, наблюдаемые значения (**observed value**) зависимой переменной Y, предсказанные значения Y (**predicted**), остатки (**residuals**) и стандартизированные остатки (**Standard Residual**) вычисляемые по формуле $\frac{e_i}{S}$, $i = 1, 2, ..., n$, где S — оценка среднеквадратического отклонения ошибок наблюдений, $S \approx 0{,}7$ (рис. 6.7).

Остаточная сумма квадратов Q_e (**Residual**) сумма квадратов, обусловленная регрессией Q_R (**Regress**) и сумма квадратов отклонений зависимой пе-

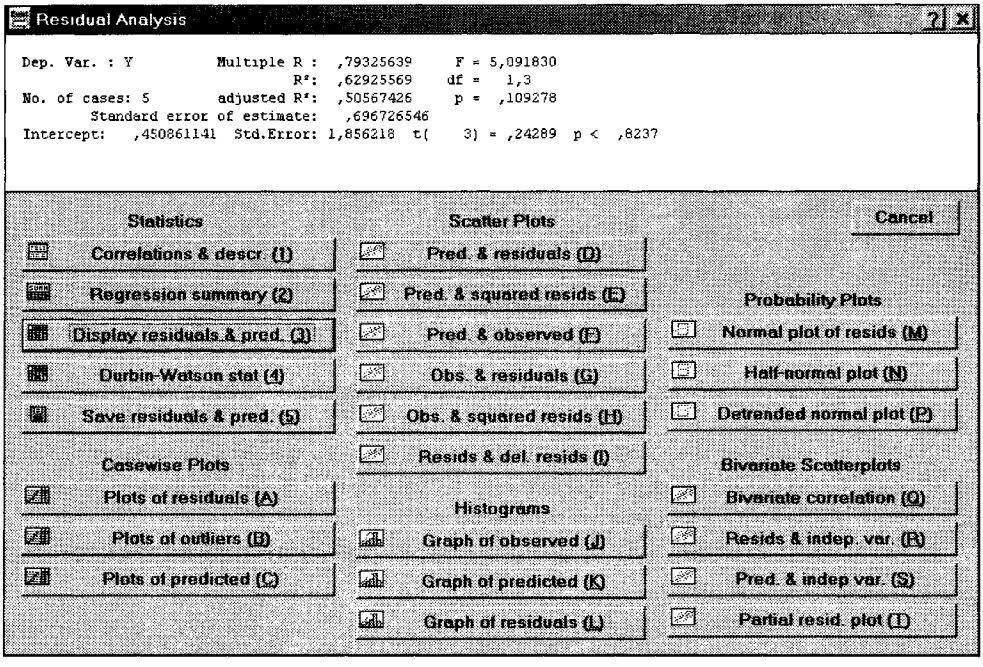


Рис. 6.6. Меню для анализа остатков

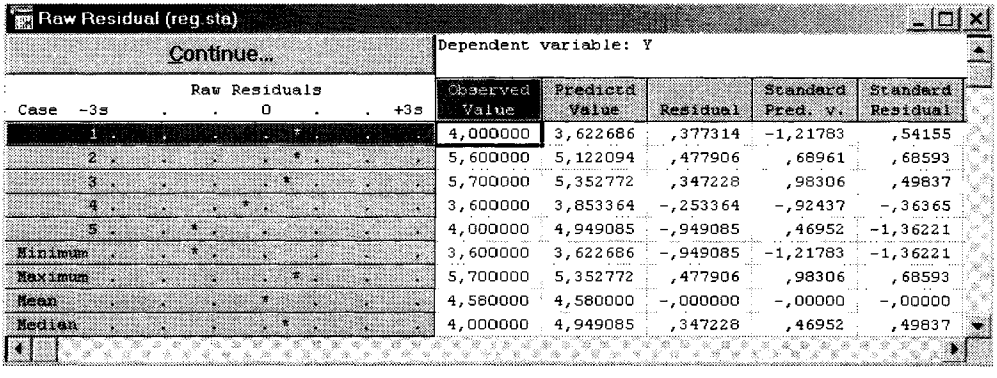


Рис. 6.7. График остатков (слева) и их значения (столбец Residual) (справа)

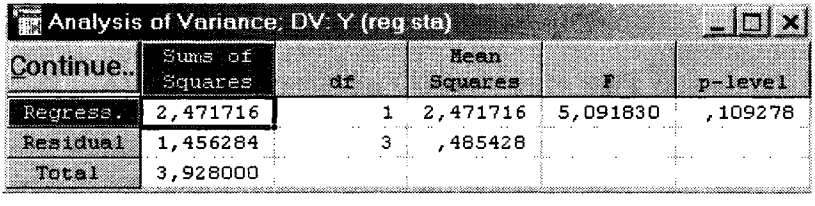


Рис. 6.8. Дисперсионный анализ

ременной Y от среднего Q_y (**Total**) вычисляются при нажатии кнопки **Analysis of Variance** (дисперсионный анализ) на панели результатов вычислений (рис. 6.4). Результаты дисперсионного анализа приведены на рис. 6.8.

В этой же таблице приведены соответствующие значения числа степеней свободы (*df*), средние квадраты, *F*-статистика для проверки гипотезы о незначимости регрессионной модели и вычисленный уровень значимости *p*.

В данном примере гипотеза о незначимости регрессионной модели по *F*-критерию также принимается, т. к. $p \approx 0{,}11$, что больше обычно задаваемого уровня значимости $\alpha = 0{,}05$.

6.2.2. Работа 2. Проверка значимости и адекватности простой линейной регрессии. Прогнозирование

1. Основные понятия

Ковариационная матрица оценок параметров регрессионной модели. Доверительные интервалы для параметров регрессионной модели. Проверка гипотез о равенстве параметров нулю. Разложение суммы квадратов отклонений результатов наблюдений от среднего (Q_y) на сумму квадратов обусловленную регрессией (Q_R) и остаточную на сумму квадратов (Q_e). Смысл тождества: $Q_y = Q_R + Q_e$ и проверка гипотезы о незначимости регрессионной модели. Проверка адекватности по графику остатков. Критерий Дарбина—Уотсона. Проверка гипотезы о нормальном распределении остатков. Проверка адекватности по повторным наблюдениям. Доверительные интервалы для среднего предсказанного значения и для индивидуального предсказанного значения.

2. Задание

По выборке из своего варианта, используя результаты расчетов полученные в работе 1, выполнить следующие расчеты и задания:

1. Вычислить ковариационную матрицу оценок параметров регрессионной модели.
2. Вычислить доверительные интервалы для параметров регрессии и для дисперсии ошибок наблюдений при доверительной вероятности 0,95.
3. Вычислить сумму квадратов, обусловленную регрессией по одной из формул

$$Q_R = \tilde{\beta}_1 Q_{xy} = \tilde{\beta}_1^2 Q_x = \frac{Q_{xy}^2}{Q_x}.$$

4. Проверить тождество: $Q_y = Q_R + Q_e$.
5. Проверить гипотезу о незначимости модели $H_0 : \beta_1 = 0$ по *F*-критерию Фишера и используя доверительный интервал для β_1.
6. Построить график остатков.
7. Вычислить статистику Дарбина—Уотсона.
8. Вычислить доверительные интервалы для среднего предсказанного значения и индивидуального предсказанного значения $\tilde{Y}(x_0)$. В качестве x_0 взять два значения

$$x_{01} = \frac{x_{\min} + x_{\max}}{2} \quad \text{и} \quad x_{02} = x_{\max} + 2,$$

где $x_{\min}$ и $x_{\max}$ минимальное и максимальное значение x в заданной выборке.

Границы доверительных интервалов для предсказанных значений нанести на график, содержащий прямую регрессии Y на x и диаграмму рассеяния. Доверительную вероятность взять равной 0,90.

9. Ввести данные в пакет STATISTICA, выполнить п. 1—8. Сравнить результаты расчетов и записать их в отчет.

Пример 6.1 (продолжение). Продолжим решение примера 6.1 по пунктам задания в работе 2.

1. Ковариационная матрица оценок параметров регрессионной модели K вычисляется по формуле

$$K = S^2 (A^T A)^{-1} = S^2 B^{-1} =$$

$$= 0{,}486 \begin{pmatrix} 7{,}098 & -0{,}963 \\ -0{,}963 & 0{,}135 \end{pmatrix} \approx \begin{pmatrix} 3{,}449 & -0{,}468 \\ -0{,}468 & 0{,}065 \end{pmatrix}.$$

Таким образом имеем:

$$D[\tilde{\beta}_0] = 3{,}449, \quad D[\tilde{\beta}_1] = 0{,}065, \quad \operatorname{cov}(\tilde{\beta}_0, \tilde{\beta}_1) = -0{,}468.$$

В пакете STATISTICA выводятся значения стандартных отклонений (***St. Error of B***):

$$\sqrt{D(\tilde{\beta}_0)} = \sqrt{3{,}449} \approx 1{,}856 \quad \text{и} \quad \sqrt{D(\tilde{\beta}_1)} = \sqrt{0{,}065} \approx 0{,}255$$

(см. рис. 6.5. Результаты регрессии).

2. Доверительные интервалы для параметров линейной регрессии вычисляются по следующим формулам:

$$\text{для } \beta_0\text{: } \tilde{\beta}_0 \pm t_{1-\frac{\alpha}{2}}(n-k) \cdot \sqrt{D[\tilde{\beta}_0]};$$

$$\text{для } \beta_1\text{: } \tilde{\beta}_1 \pm t_{1-\frac{\alpha}{2}}(n-k) \cdot \sqrt{D[\tilde{\beta}_1]},$$

где $t_{1-\frac{\alpha}{2}}(n-k)$ — квантиль распределения Стьюдента с $(n-k)$ степенями свободы порядка $1-\dfrac{\alpha}{2}$.

При доверительной вероятности $1-\alpha = 0{,}95$, $t_{0{,}975}(5-2) = t_{0{,}975}(3) = 3{,}182$ (используйте статистический калькулятор!)

Окончательно имеем следующие значения доверительных интервалов:

$$\text{для } \beta_0\text{: } 0{,}451 \pm 3{,}182 \cdot \sqrt{3{,}449} = 0{,}451 \pm 5{,}909,$$

$$\text{для } \beta_1\text{: } 0{,}577 \pm 3{,}182 \cdot \sqrt{0{,}065} = 0{,}577 \pm 0{,}811.$$

Таким образом оба коэффициента регрессии β_0 и β_1 *незначимы* на уровне значимости $\alpha = 0{,}05$, т. к. 95%-е доверительные интервалы для β_0 и β_1 включают нуль.

В пакете STATISTICA (см. рис. 6.5) вычисляются значения t-статистик для проверки гипотезы $H_0: \beta_0 = 0$

$$t = \frac{\widetilde{\beta}_0}{\sqrt{D[\widetilde{\beta}_0]}} = \frac{0{,}451}{1{,}856} \approx 0{,}243$$

и для проверки гипотезы $H_0: \beta_1 = 0$

$$t = \frac{\widetilde{\beta}_1}{\sqrt{D[\widetilde{\beta}_1]}} = \frac{0{,}577}{0{,}255} \approx 2{,}256.$$

Обе гипотезы принимаются на уровне значимости соответственно:

$$p = 0{,}824 \quad \text{и} \quad p = 0{,}109.$$

Доверительный интервал для дисперсии ошибок наблюдений определяется по формуле

$$\frac{(n-k)\cdot S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-k)} < \sigma^2 < \frac{(n-k)\cdot S^2}{\chi^2_{\frac{\alpha}{2}}(n-k)},$$

где $\chi^2_{1-\frac{\alpha}{2}}(n-k)$ и $\chi^2_{\frac{\alpha}{2}}(n-k)$ — квантили распределения χ^2 с $(n-k)$ степенями свободы. При доверительной вероятности $1-\alpha = 0{,}95$ имеем (используйте статистический калькулятор!) при $n = 5$ и $k = 2$:

$$\chi^2_{0{,}975}(3) = 7{,}81, \quad \chi^2_{0{,}025}(3) = 0{,}216.$$

Таким образом доверительный интервал для дисперсии ошибок наблюдений имеет вид

$$\frac{(5-2)\cdot 0{,}486}{7{,}81} < \sigma^2 < \frac{(5-2)\cdot 0{,}486}{0{,}216}$$

или окончательно

$$0{,}187 < \sigma^2 < 6{,}75.$$

3. Сумма квадратов, обусловленная регрессией Q_R:

$$Q_R = \widetilde{\beta}_1 Q_{xy} = 0{,}577 \cdot 4{,}286 \approx 2{,}472$$

(сравните результаты расчета с результатами дисперсионного анализа, рис. 6.8).

4. Проверяем тождество $Q_y = Q_R + Q_e$: $3{,}928 \approx 2{,}472 + 1{,}457 = 3{,}929$.

5. Проверим гипотезу $H_0: \beta_1 = 0$ о незначимости регрессионной модели по критерию Фишера.

Выборочное значение статистики Фишера F равно

$$F_в = \frac{Q_R/(k-1)}{Q_e/(n-k)} = \frac{2{,}472/1}{1{,}457/3} = 5{,}091.$$

Так как $F_в$ меньше квантили распределения Фишера $F_{1-\alpha}(k-1, n-k) = F_{0,95}(1,3) = 10,13$, то гипотеза $H_0: \beta_1 = 0$ не отклоняется: регрессионная модель *незначима* (сравните этот результат со значениями F-статистики и p-уровня на рис. 6.8).

Тот же результат получим используя 95%-й доверительный интервал для β_1: (–0,235; 1,387).

Так как 95%-й доверительный интервал для β_1 накрывает 0, гипотеза $H_0: \beta_1 = 0$ принимается на уровне значимости $\alpha = 0,05$.

6. График остатков. В данном примере число остатков очень мало ($n = 5$) поэтому сделать какие-либо выводы о выполнении предположений регрессионного анализа по остаткам нельзя. Более того, так как регрессионная модель незначима, то проверка этих предложений лишена смысла.

7. Вычислим статистику Дарбина—Уотсона

$$d = \sum_{i=2}^{n} \frac{(e_i - e_{i-1})^2}{Q_e} =$$

$$= \frac{(0,48 - 0,378)^2 + (0,35 - 0,48)^2 + (-0,25 - 0,35)^2 + (-0,947 + 0,25)^2}{1,457} \approx$$

$$\approx \frac{0,873}{1,457} \approx 0,5989.$$

Для $n = 5$ критических значений статистики Дарбина—Уотсона в таблице (Приложение 2) нет. Поэтому проверить гипотезу о некоррелированности остатков при столь малом числе наблюдений нельзя.

8. Вычислим доверительные интервалы для предсказанных значений. Здесь надо иметь в виду, что если регрессионная модель незначима и неадекватна результатам наблюдений, как это имеет место в данном примере, то эту модель использовать для прогноза нельзя. Мы приведем соответствующие расчеты, чтобы продемонстрировать только технику вычислений.

Найдем предсказанное значение Y в точках:

$$x_{01} = \frac{x_{\min} + x_{\max}}{2} = \frac{5,5 + 8,5}{2} = 7,$$

$$x_{02} = x_{\max} + 2 = 8,5 + 2 = 10,5,$$

$$\tilde{y}(x_{01}) = \tilde{\beta}_0 + \tilde{\beta}_1 \cdot x_{01} = 0,454 + 0,576 \cdot 7 = 4,486,$$

$$\tilde{y}(x_{02}) = 0,454 + 0,576 \cdot 10,5 = 6,502.$$

Границы доверительного интервала для *среднего предсказанного значения* (**confidence limit**) вычисляются по формуле

$$\tilde{y}(x_0) \pm t_{1-\frac{\alpha}{2}}(n-2) \cdot S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Q_x}}$$

или по более общей формуле (см. ниже, п. 6.3.2):

$$\tilde{y}(x_0) \pm t_{1-\frac{\alpha}{2}}(n-k) \cdot S\sqrt{a^T(x_0) \cdot B^{-1} \cdot a(x_0)},$$

где $a^T(x)$ — вектор-строка регрессионной матрицы A; в случае простой линейной регрессии: $a^T(x) = (1, x)$.

В данном примере, при доверительной вероятности $1 - \alpha = 0{,}90$ имеем при $x_{01} = 7$, $(t_{0,95}(3) = 2{,}353)$:

$$4{,}486 \pm 2{,}353 \cdot \sqrt{0{,}486} \cdot \sqrt{\frac{1}{5} + \frac{(7 - 7{,}16)^2}{7{,}458}} = 4{,}486 \pm 0{,}740.$$

По более общей формуле

$$a^T(x_0) \cdot B^{-1} \cdot a(x_0) = (1;7) \cdot \begin{pmatrix} 7{,}098 & -0{,}963 \\ -0{,}963 & 0{,}135 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 7 \end{pmatrix} = 0{,}231.$$

Таким образом, доверительный интервал для среднего предсказанного значения равен

$$4{,}486 \pm 2{,}353 \cdot \sqrt{0{,}486} \cdot \sqrt{0{,}231} = 4{,}486 \pm 0{,}789.$$

Чтобы вычислить *доверительный интервал для индивидуального предсказанного значения* (**prediction limit**) оценка дисперсии $D[\tilde{y}(x_0)]$ должна включать еще один источник вариации — разброс относительно линии регрессии, определяемый дисперсией S^2. Таким образом, доверительный интервал для индивидуального значения вычисляется по формуле

$$\tilde{y}(x_0) \pm t_{1-\frac{\alpha}{2}}(n-2) \cdot S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Q_x}}$$

или, в общем случае (см. ниже, п. 6.3.2):

$$\tilde{y}(x_0) \pm t_{1-\frac{\alpha}{2}}(n-k) \cdot S\sqrt{1 + a^T(x_0) \cdot B^{-1} \cdot a(x_0)}\ .$$

В рассматриваемом примере для индивидуального предсказанного значения Y при $x_{01} = 7$, получим следующие значения границ доверительного интервала

$$4{,}486 \pm 2{,}353 \cdot \sqrt{0{,}486} \cdot \sqrt{1 + \frac{1}{5} + \frac{(7 - 7{,}16)^2}{7{,}458}} = 4{,}486 \pm 1{,}780$$

или по общей формуле

$$4{,}486 \pm 2{,}353 \cdot \sqrt{0{,}486} \cdot \sqrt{1 + 0{,}231} = 4{,}486 \pm 1{,}820.$$

Аналогично вычисляются значения границ доверительных интервалов для среднего и индивидуального предсказанного значения Y при $x = 10{,}5$. Соответственно, имеем:

$$6{,}502 \pm 2{,}353 \cdot \sqrt{0{,}486} \cdot \sqrt{\frac{1}{5} + \frac{(10{,}5 - 7{,}16)^2}{7{,}458}} = 6{,}502 \pm 2{,}136;$$

$$6{,}502 \pm 2{,}353 \cdot \sqrt{0{,}486} \cdot \sqrt{1 + \frac{1}{5} + \frac{(10{,}5 - 7{,}16)^2}{7{,}458}} = 6{,}502 \pm 2{,}693.$$

9. Выполнение задания в пакете STATISTICA.

Основные моменты статистического анализа результатов расчетов для простой линейной регрессии в пакете STATISTICA мы уже прокомментировали. Более подробно проверка адекватности регрессионной модели по анализу остатков будет рассмотрена ниже в примере 6.2.

Рассмотрим вычисление предсказанных значений и доверительных интервалов для них.

Вычисления выполняются при нажатии кнопки **Predict dependent variable** (предсказанное значение зависимой переменной) в окне **Multiple Regression Results** (рис. 6.4). Предварительно надо задать уровень значимости α и вид вычисляемого доверительного интервала: **Confidence limits** — доверительный интервал для среднего предсказанного значения; или **Prediction limits** — доверительный интервал для индивидуального предсказанного значения.

Нажав кнопку и задав значение независимой переменной, например, $x_{01} = 7$, в таблице результатов (рис. 6.9) получим предсказанное значение: $\tilde{y}(x_0) \approx 4{,}488$ и 90%-е доверительные интервалы для среднего предсказанного значения: (3,748; 5,227).

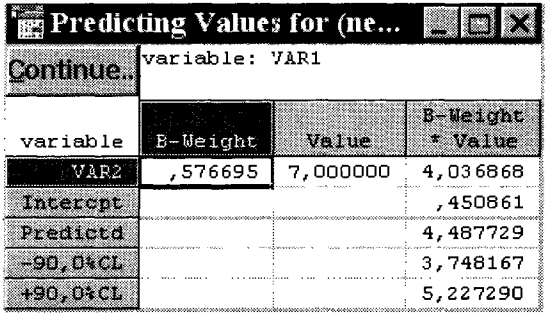


Рис. 6.9. Вычисление предсказанного значения

6.2.3. Задания для самостоятельной работы.

Ниже приводятся девять задач регрессионного анализа с одной независимой переменной.

Для каждой задачи, используя пакет STATISTICA, найдите уравнение регрессии и проведите статистический анализ регрессионной модели.

1. Постройте диаграмму рассеяния, используя следующую последовательность действий: **Graphs** → **Stat 2D Graphs** → **Scatterplots...** → введите переменные X и Y, задайте:

1) **Graph Type: Regular**.

2) **Fit** (модель): последовательно устанавливайте следующие модели:

Linear ($y = \beta_0 + \beta_1 x + \varepsilon$);

Logarithmic ($y = \beta_0 + \beta_1 \lg x + \varepsilon$);

Exponential ($y = \beta_0 \exp(\beta_1 x)$);

Polynomial ($y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots$).

Просматривая каждый график, выберите две наиболее подходящие с вашей точки зрения модели.

Указания.

1) Модель $Y = \beta_0 \exp(\beta_1 x)$ нелинейна по параметрам, но легко преобразуется в линейную следующим преобразованием и заменой переменных:

$\ln y = \ln \beta_0 + \beta_1 x$;

$y' = \ln y$, $\beta'_0 = \ln \beta_0$;

$y' = \beta'_0 + \beta_1 x$.

2) Для анализа регрессионных моделей, содержащих нелинейные функции независимой переменной x (x^2, x^3, $\log(x)$, $\sqrt{x}$, $1/x$, 10^x) воспользуйтесь опцией Fixed non-linear. Эта опция открывается из стартовой панели модуля Multiple Regression (рис. 6.3) в окне Mode (по умолчанию в этом окне установлена опция Standard).

2. По результатам регрессионного анализа выбранных вами обеих моделей в модуле **Multiple Regression**, определите модель, имеющую наибольший коэффициент детерминации R^2.

3. Проведите статистический анализ для выбранной регрессионной модели:

- Определите доверительные интервалы и проверьте гипотезы о значимости параметров регрессии.
- Найдите остаточную сумму квадратов Q_e оценку дисперсии ошибок наблюдений S^2, коэффициент детерминации R^2, оценку коэффициента корреляции.
- Проверьте значимость регрессионной модели по F-критерию.
- Определите остатки. Постройте график остатков. Проверьте выполнение предложений регрессионного анализа:
  — дисперсия остатков постоянна;
  — остатки некоррелированны;
  — остатки имеют нормальное распределение $N(0, \sigma^2)$.
  Сделайте вывод об адекватности регрессионной модели результатам наблюдений.
- Используя выбранную модель регрессии, определите предсказанное значение зависимой переменной Y при следующем значении независимой переменной:

$$x_0 = \max(x) + S_x,$$

где S_x — оценка среднего квадратического отклонения переменной x. Определите доверительные интервалы для среднего и индивидуального предсказанного значения. Для всех расчетов принять $\alpha = 0{,}05$.

Задачи

1. Президента компании интересует зависимость между приростом годового дохода и качеством работы коммерческих агентов в будущем году. Он выбрал 12 агентов и определил размеры дохода, приносимого компании каждым из них (в процентах от окладов), а также количество продаж, проведенных каждым агентом в течение года:

Размер дохода x, %	7,8	6,9	6,7	6,0	6,9	5,2	6,3	8,4	7,2	10,1	10,8	7,7
Количество продаж, y	64	73	42	49	71	46	32	88	53	84	85	93

Определите регрессионную модель по этим данным.

2. Используя приведенные ниже данные, установите, есть ли значимая зависимость между объемом инвестиций и ценой за акцию?

Объем инвестиций x, млн руб.	108	4,4	3,5	3,6	39	68,4	7,5	5,5	375	12	51
Цена за акцию y, руб.	12	4	5	6	13	19	8,5	5	15	6	12

3. Автосервисное предприятие имеет следующие данные по стоимости ежегодного технического обслуживания автомобилей определенной марки в зависимости от времени эксплуатации.

Стоимость тех. обслуживания, y (тыс. руб.)	5,3	5,2	6,0	5,7	6,6	6,8	8,1	6,9	10,3	4,0	2,5
Время эксплуатации, x (лет)	5	4	5	6	7	8	10	8	11	3	2

Определите регрессионную модель для этих данных.

4. Ниже приведены данные характеризующие стоимость продукции в зависимости от количества исходного сырья:

Стоимость продукции y, (тыс. руб.)	10	7	5	6	7	6	4	3	8	9
Количество исходного сырья x, тонны	0,25	0,20	0,16	0,17	0,19	0,18	0,15	0,14	0,21	0,21

Определите регрессионную модель для этих данных.

5. Для 25 предприятий розничной торговли получены следующие данные:

Розничный товарооборот x, млн руб.	Издержки обращения y, млн руб.
510	30
560	33
800	46
465	31
225	16
390	25
640	39
405	26
200	15
425	34
570	37
472	28
250	19
665	38
650	36
620	35
380	24
550	38
750	44
660	36
450	27
563	34
400	26
553	38
772	45

Определите регрессионную модель для этих данных.

6. Определите регрессионную модель, описывающую зависимость нераспределенной прибыли (млн руб.) от инвестиций в основные фонды (млн руб.) по результатам обследования 20 предприятий:

Нераспределенная прибыль y, млн руб.	Инвестиции в основные фонды x, млн руб.
2,3	0,3
3,4	0,30
4,3	0,40
5,0	0,60
6,0	1,00
2,0	0,16
3,6	0,20
4,2	0,30
5,8	1,00
4,7	0,60
2,7	0,11
3,8	0,40
4,5	0,70
4,8	0,70
4,4	0,50
5,5	0,80
5,6	0,70
4,1	0,30
3,6	0,30
5,7	0,90

7. По результатам отчетов о работе 30 малых предприятий получены такие показатели:

Среднегодовая стоимость основных производственных фондов x, млн руб.	Производство продукции y, млн руб.
27	21
28	35
41	38
44	46

Продолжение задачи 7

Среднегодовая стоимость основных производственных фондов x, млн руб.	Производство продукции y, млн руб.
55	51
33	30
37	38
49	50
56	61
37	34
38	35
49	39
26	19
29	36
20	24
47	40
36	35
56	60
57	48
45	43
39	45
46	48
60	46
55	57
53	34
42	42
41	47
35	30
33	41
46	27

Определите регрессионную модель, описывающую зависимость между этими показателями.

8. Анализ работы 25 предприятий сферы услуг дал следующие результаты:

Объем работ x, млн руб.	Накладные расходы y, млн руб.
9,0	2,7
10,3	3,0
7,0	2,5
5,2	2,2
6,4	2,5
9,5	2,7
14,0	4,0
13,0	4,0
5,0	2,0
7,4	2,6
9,3	2,6
8,0	3,2
10,2	2,3
10,0	3,0
12,0	2,6
15,0	3,0
16,0	5,0
17,0	4,3
21	5,0
19,0	4,8
12,5	4,0
8,0	2,0
11,0	3,0
6,5	2,0
13,0	5,0

Найдите регрессионную модель для этих данных.

9. Найдите регрессионные модели $Y=f(X)+\varepsilon$ для данных приведенных ниже:

X — независимая переменная;

Y_1—Y_6 — зависимые переменные.

Номер наблюдения	X	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1	1,2	66,6	63,0	14,4	99,2	1,8	10,0
2	2,7	62,8	74,1	21,5	89,9	6,4	6,3
3	4,3	63,8	71,0	24,7	85,5	7,8	1,3
4	4,4	58,5	77,6	27,0	87,4	6,0	2,2
5	4,5	64,3	81,7	28,4	84,2	7,1	1,2
6	4,6	60,6	79,6	30,5	87,2	8,7	2,2
7	4,8	58,5	64,6	20,3	87,8	7,7	,6
8	5,9	51,5	97,4	25,7	86,5	7,5	,2
9	6,0	58,4	89,0	29,1	80,4	8,8	1,2
10	6,1	52,1	77,1	30,4	82,2	9,7	,1
11	7,3	52,5	96,5	30,7	80,2	8,2	–,1
12	7,7	44,2	92,0	29,1	85,3	10,1	,1
13	7,8	51,7	82,8	31,9	77,4	9,7	–,2
14	8,0	52,2	79,8	31,4	80,1	8,2	–,0
15	9,5	42,0	114,7	35,2	77,8	11,6	,1
16	9,7	40,4	102,7	35,0	78,0	9,9	,5
17	11,8	38,0	117,2	38,3	76,6	10,4	,1
18	12,6	29,5	128,2	39,9	79,5	12,7	,1
19	12,8	29,3	141,0	37,5	82,4	12,4	,2
20	13,0	27,9	139,5	35,5	77,4	11,2	–,9
21	13,1	31,6	147,9	39,4	76,2	12,2	–,1
22	13,3	32,4	143,9	40,0	77,4	12,3	–,6
23	13,7	19,1	169,5	39,9	76,0	13,1	–1,0
24	14,0	28,4	152,8	38,2	76,6	12,6	–,3
25	14,4	24,0	162,3	40,3	68,8	12,6	,5
26	14,8	20,6	161,9	39,1	76,2	13,6	,4
27	15,3	13,9	167,5	38,8	76,5	14,2	–,6
28	16,1	12,5	188,6	40,1	75,4	13,1	1,3
29	17,0	5,6	212,9	37,9	73,8	12,4	,5
30	17,4	1,4	222,3	43,7	77,1	12,7	,1

6.3. Множественная регрессия

Если переменная Y зависит от нескольких факторов x_1, x_2, ..., x_{k-1}, то регрессионная модель определяется уравнением множественной регрессии

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1} + \varepsilon.$$

Это уравнение содержит k параметров: β_0, β_1, ..., β_{k-1}; ε — вектор случайных ошибок наблюдений.

Исходные данные для регрессионного анализа представляют собой результаты наблюдений зависимой переменной Y и факторов x_1, x_2, ..., x_{k-1} и записываются в виде таблицы:

y	x_1	x_2	...	x_{k-1}
y_1	x_{11}	x_{21}	...	$x_{(k-1)1}$
y_2	x_{12}	x_{22}	...	$x_{(k-1)2}$
...	...	...	...	...
y_n	x_{1n}	x_{2n}	...	$x_{(k-1)n}$

Регрессионный анализ простой линейной регрессии (п. 6.1) обобщается на случай множественной регрессии. Для нахождения оценок параметров β_i, $i = 0, 1, \ldots, k-1$ по результатам наблюдений используется метод наименьших квадратов (МНК): в качестве оценок параметров принимаются значения $\tilde{\beta}_0$, $\tilde{\beta}_1$, ..., $\tilde{\beta}_{k-1}$, минимизирующие сумму квадратов отклонений зависимой переменной y_i, $i = 1, 2, \ldots, n$ от значений $\tilde{y}_i$, вычисляемых по уравнению множественной регрессии:

$$\tilde{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{k-1} x_{(k-1)i}. \qquad (1)$$

Таким образом, оценки, вычисляемые по методу наименьших квадратов (МНК-оценки), определяются из условия минимума функции k переменных $Q(\beta_0, \beta_1, \ldots, \beta_{k-1})$:

$$Q(\beta_0, \beta_1, \ldots, \beta_{k-1}) = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 =$$

$$= \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_{k-1} x_{(k-1)i})]^2. \qquad (2)$$

Из необходимых условий минимума функции (2):

$$\frac{\partial Q}{\partial \beta_i} = 0, \; i = 0, 1, \ldots, k-1$$

следует, что МНК-оценки $\tilde{\beta}_0$, $\tilde{\beta}_1$, ..., $\tilde{\beta}_{k-1}$ являются решениями системы k линейных алгебраических уравнений:

$$\begin{cases} \beta_0 n + \beta_1 \sum x_{1i} + \ldots + \beta_{k-1} \sum x_{(k-1)i} = \sum y_i \\ \beta_0 \sum x_{1i} + \beta_1 \sum x_{1i}^2 + \ldots + \beta_{k-1} \sum x_{1i} x_{(k-1)i} = \sum y_i x_{1i} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \beta_0 \sum x_{(k-1)i} + \beta_1 \sum x_{1i} x_{(k-1)i} + \ldots + \beta_{k-1} \sum x_{(k-1)i}^2 = \sum y_i x_{(k-1)i} \end{cases} \quad (3)$$

Система уравнений (3) называется *нормальной* системой.

Первое уравнение системы (3) дает следующее соотношение между МНК-оценками:

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}_1 - \tilde{\beta}_2 \bar{x}_2 - \ldots - \tilde{\beta}_{k-1} \bar{x}_{k-1},$$

где $\bar{y}$, $\bar{x}_1$, ..., $\bar{x}_{k-1}$ — средние значения результатов наблюдений зависимой переменной y и факторов x_1, x_2, ..., x_{k-1}. Это важный результат (посмотрите на аналогичное соотношение для простой линейной регрессии).

Анализ множественной регрессии на основе системы нормальных уравнений (3) является трудоемкой и неэффективной процедурой. Пробиться сквозь дебри индексов и знаков суммирования нам поможет аппарат матричной алгебры. Эта изящная математическая теория дает возможность выпукло представить основные идеи регрессионного анализа и получить практически важные выводы, не только для множественной регрессии, но и для полиномиальной регрессии и регрессии общего вида (см. введение к этой главе). Вся разница состоит только в том как определена регрессионная матрица A.

Введем матричные обозначения:

регрессионная $(n \times k)$ матрица A: $\begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{(k-1)1} \\ 1 & x_{12} & x_{22} & \cdots & x_{(k-1)2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{(k-1)n} \end{pmatrix}$;

$Y = (y_1, y_2, \ldots, y_n)^T$ — вектор значений зависимой переменной;

$\beta = (\beta_0, \beta_1, \ldots, \beta_{k-1})^T$ — вектор значений параметров модели;

$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ — вектор ошибок наблюдений зависимой переменной.

Тогда множественная линейная регрессия определяется матричным уравнением

$$Y = A \cdot \beta + \varepsilon.$$

6.3.1. Оценка параметров регрессионной модели по результатам наблюдений

Для оценки параметров модели по результатам наблюдений метод наименьших квадратов (МНК) применяется следующим образом.

Введем вектор невязки (вектор ошибок наблюдений):

$$\varepsilon = Y - A\beta.$$

Предположим, что ранг матрицы A равен k и, следовательно, столбцы матрицы A — линейно-независимые векторы $A_1, A_2, \ldots, A_k$.

Вектор $A\beta$ является линейной комбинацией столбцов матрицы A и может быть записан в виде

$$A\beta = A_1\beta_0 + A_2\beta_1 + \ldots + A_k\beta_{k-1}.$$

Это означает, что вектор $A\beta$ лежит в линейном подпространстве L_k размерности k, причем $L_k \subset L_n$, где L_n — линейное пространство векторов размерности n. Так как вектор наблюдений Y принадлежит L_n, то в качестве решения системы

$$Y = A\beta$$

можно взять вектор $\tilde{\beta}$, минимизирующий модуль вектора невязки

$$\|\varepsilon\| = \|Y - A\beta\|.$$

В геометрической интерпретации (см. рис. 6.10) вектор $A\tilde{\beta}$ будет ортогональной проекцией вектора Y на подпространство L_k.

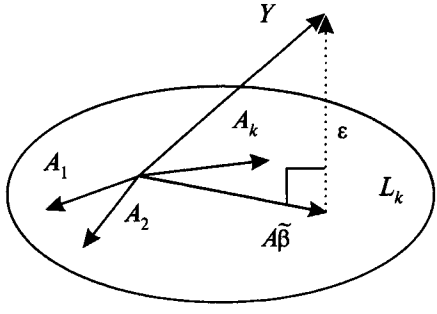


Рис. 6.10. Геометрическая интерпретация метода наименьших квадратов

Из условия ортогональности вектора невязки и подпространства L_k следует, что скалярное произведение вектора ε и любого вектора $A\alpha \subset L_k$, $\alpha \neq 0$, равно нулю:

$$(A\alpha)^T(Y - A\beta) = 0$$

или

$$\alpha^T(A^TY - A^TA\beta) = 0.$$

Решив уравнение $A^TY - A^TA\beta = 0$, получим, что решением несовместной системы $Y = A\beta$, минимизирующим модуль вектора невязки ε, будет вектор $\tilde{\beta}$:

$$\tilde{\beta} = (A^TA)^{-1}A^TY.$$

Полученная формула определяет оценки параметров линейной регрессионной модели по результатам наблюдений.

6.3.2. Статистический анализ МНК-оценок. Оценка качества аппроксимации данных с помощью линейной регрессионной модели

МНК-оценки параметров модели, вычисляемые по формуле

$$\tilde{\beta} = (A^T A)^{-1} A^T Y$$

представляют собой линейные функции случайного вектора Y и, следовательно, будут случайными величинами. Этот результат является следствием того, что оценки параметров модели вычисляются по выборке наблюдений переменных Y, x_1, x_2, ... и изменяются при вариациях наблюдений.

Задача статистического анализа состоит в исследовании свойств оценок, нахождении доверительных интервалов для параметров линейной модели и проверке гипотез о параметрах.

Статистический анализ проводится при следующих предположениях:

- вектор ошибок (невязки) имеет нулевое математическое ожидание: $M[\varepsilon] = 0$ — это означает, что отсутствуют систематические ошибки наблюдений;
- дисперсии ошибок наблюдений постоянны: $D[\varepsilon_i] = \sigma^2$, $i = 1, 2, ..., n$, ошибки наблюдений ε_i некоррелированны: $\mathrm{cov}[\varepsilon_i, \varepsilon_j] = 0$, $i \neq j$, $i, j = 1, 2, ..., n$ и, следовательно, ковариационная матрица ошибок наблюдений $\mathrm{cov}(\varepsilon)$ имеет вид: $\mathrm{cov}(\varepsilon) = \sigma^2 I$, где I — единичная $(n \times n)$ матрица;
- вектор ошибок ε имеет n-мерное нормальное распределение $N(0, \sigma^2 I)$. При этом из предположения о некоррелированности ошибок наблюдений следует их независимость.

Вектор наблюдений зависимой переменной Y в линейной регрессионной модели можно записать в виде

$$Y = A\beta + \varepsilon.$$

Отсюда следует, что Y имеет n-мерное нормальное распределение с математическим ожиданием

$$M[Y] = M[A\beta + \varepsilon] = A\beta = m_Y.$$

Ковариационная матрица вектора Y, $\mathrm{cov}(Y)$, вычисляется по формуле

$$\mathrm{cov}(Y) = M[(Y - m_Y)(Y - m_Y)^T] = M[\varepsilon\varepsilon^T] = \mathrm{cov}(\varepsilon) = \sigma^2 I.$$

Пусть C — постоянная $(k \times n)$ матрица. Легко проверяются следующие свойства:

$$M[CY] = CM[Y],$$

$$\mathrm{cov}(CY) = C\,\mathrm{cov}(Y)\,C^T.$$

Обозначим квадратную симметрическую матрицу $A^T A$ размера $(k \times k)$ как B: $B = A^T A$. Матрица B называется информационной матрицей. Запишем вектор МНК-оценок $\tilde{\beta}$ в виде

$$\tilde{\beta} = \mathrm{B}^{-1} A^T Y = B^{-1} A^T (A\beta + \varepsilon) = \beta + B^{-1} A^T \varepsilon.$$

Найдем математическое ожидание вектора $\tilde{\beta}$:

$$M[\tilde{\beta}] = \beta + B^{-1}A^T M[\varepsilon] = \beta.$$

Это означает, что МНК-оценки $\tilde{\beta}$ — несмещенные оценки вектора параметров регрессии β.

Вычислим ковариационную матрицу $\mathrm{cov}(\tilde{\beta})$, используя вышеприведенную формулу для ковариации произведения CY:

$$\mathrm{cov}(\tilde{\beta}) = \mathrm{cov}(B^{-1}A^T Y) = B^{-1}A^T \mathrm{cov}(Y) AB^{-1} = \sigma^2 B^{-1}.$$

Таким образом, вектор $\tilde{\beta}$, как линейная функция вектора Y, имеет k-мерное нормальное распределение с математическим ожиданием и ковариационной матрицей соответственно $M[\tilde{\beta}] = \beta$, $\mathrm{cov}(\tilde{\beta}) = \sigma^2 B^{-1}$.

Подставим МНК-оценки $\tilde{\beta}$ в уравнение невязки

$$Y - A\tilde{\beta} = e.$$

Вектор e (вектор остатков) определяет разницу между результатами наблюдений и значениями, предсказываемыми моделью. Квадрат модуля вектора остатков, равный $e^T e$, называется *остаточной суммой квадратов* Q_e, которая характеризует качество аппроксимации данных регрессионной моделью.

Линейная регрессионная модель называется *адекватной результатам наблюдений*, если предсказанные по ней значения зависимой переменной Y согласуются с результатами наблюдений. Если модель адекватна, то остатки e_i являются реализациями случайных ошибок наблюдений ε_i, $i = 1, 2, \ldots, n$, и, следовательно, в силу вышеприведенных предположений независимы и имеют нормальное распределение $N(0, \sigma^2)$.

Проверка выполнения этих предположений различными статистическими методами лежит в основе оценки адекватности модели по остаткам. Если для каждого или для некоторых значений независимой переменной имеются несколько значений зависимой переменной Y (повторные наблюдения), то проверка адекватности проводится на основе дисперсионного анализа (см. ниже, п. 6.3.4).

В дальнейшем предполагается, что регрессионная модель адекватна результатам наблюдений, тогда можно показать [22], что отношение $\dfrac{Q_e}{\sigma^2}$ имеет распределение χ^2 с $(n - k)$ степенями свободы, где σ^2 — дисперсия ошибок наблюдений; n — число наблюдений; k — число параметров модели. Отсюда следует, что статистика

$$S^2 = \frac{Q_e}{n - k}$$

будет несмещенной оценкой дисперсии ошибок наблюдений σ^2, которая, как правило, неизвестна. Доверительный интервал для σ^2 имеет вид

$$\frac{Q_e}{\chi^2_{1-\frac{\alpha}{2}}(n-k)} < \sigma^2 < \frac{Q_e}{\chi^2_{\frac{\alpha}{2}}(n-k)},$$

где $\chi^2_{1-\frac{\alpha}{2}}(n-k)$, $\chi^2_{\frac{\alpha}{2}}(n-k)$ — квантили распределения $\chi^2(n-k)$ соответственно порядков $1-\alpha/2$ и $\alpha/2$, а α — заданный уровень значимости.

Рассмотрим координаты вектора МНК-оценок. Имеем

$$M[\tilde{\beta}_i]=\beta_i,\ D[\tilde{\beta}_i]=\sigma^2(B^{-1})_{ii},\quad i=0,\ 1,\ 2,\ \ldots,\ k-1,$$

где $(B^{-1})_{ii}$ — диагональный элемент матрицы B^{-1} находящийся на пересечении i-й строки и i-го столбца. Оценка параметра β_i, $\tilde{\beta}_i$, имеет нормальное распределение с математическим ожиданием $M[\tilde{\beta}_i]=\beta_i$:

$$\tilde{\beta}_i \sim N(\beta_i, D[\tilde{\beta}_i]).$$

Так как (для модели адекватной результатам наблюдений) имеет место соотношение

$$\frac{Q_e}{\sigma^2}=\chi^2(n-k),$$

то для оценки неизвестной дисперсии ошибок наблюдений σ^2 воспользуемся статистикой $S^2=\dfrac{Q_e}{n-k}$. Из двух последних соотношений следует, что статистика S^2/σ^2 связана с распределением $\chi^2(n-k)$ следующей формулой

$$\frac{S^2}{\sigma^2}=\frac{\chi^2(n-k)}{n-k}.$$

Рассмотрим нормированную статистику

$$\frac{\dfrac{\tilde{\beta}_i-\beta_i}{\sqrt{D[\tilde{\beta}_i]}}}{\sqrt{\dfrac{S^2}{\sigma^2}}}.$$

Ее можно преобразовать так: $\dfrac{\dfrac{\tilde{\beta}_i-\beta_i}{\sqrt{D[\tilde{\beta}_i]}}}{\sqrt{\dfrac{S^2}{\sigma^2}}}=\dfrac{\dfrac{\tilde{\beta}_i-\beta_i}{\sigma\sqrt{(B^{-1})_{ii}}}}{\sqrt{\dfrac{\chi^2(n-k)}{n-k}}}=\dfrac{\tilde{\beta}_i-\beta_i}{S\sqrt{(B^{-1})_{ii}}}=T(n-k).$

Таким образом, так как в числителе — случайная величина, имеющая стандартное нормальное распределение $N(0, 1)$, а в знаменателе — независимая от числителя случайная величина $\sqrt{\dfrac{\chi^2(n-k)}{n-k}}$, то эта статистика имеет распределение Стьюдента с $(n-k)$ степенями свободы. С учетом этого результата получим, что доверительные интервалы для параметров β_i регрессионной модели имеют вид

$$\tilde{\beta}_i - t_{1-\frac{\alpha}{2}}(n-k)\cdot S\sqrt{(B^{-1})_{ii}} < \beta_i < \tilde{\beta}_i + t_{1-\frac{\alpha}{2}}(n-k)\cdot S\sqrt{(B^{-1})_{ii}},\ i=1,\ 2,\ \ldots,\ k-1,$$

где $t_{1-\frac{\alpha}{2}}(n-k)$ — квантиль распределения Стьюдента порядка $1-\frac{\alpha}{2}$ с $(n-k)$ степенями свободы; α — заданный уровень значимости.

Рассмотрим доверительный интервал для предсказанного среднего значения.

Обозначим строку регрессионной матрицы A через $a^T(x)$. Для простой линейной регрессии $Y=\beta_0+\beta_1 x$ это вектор-строка $a^T(x)=(1,x)$; для множественной регрессии $Y=\beta_0+\beta_1 x+\ldots+\beta_{k-1}x_{k-1}$ это вектор $a^T(x)=(1,\ x_1,\ x_2,\ \ldots,\ x_{k-1})$; для полиномиальной регрессии $Y=\beta_0+\beta_1 x+\ldots+\beta_{k-1}x^{k-1}$ это вектор $a^T(x)=(1,\ x,\ x^2,\ \ldots,\ x^{k-1})$.

Значение зависимой переменной Y, предсказанное регрессионной моделью в точке x_0, можно вычислить так: $\hat{Y}(x_0)=\tilde{\beta}^T\cdot a(x_0)$, где $\tilde{\beta}$ — вектор-столбец МНК-оценок.

Так как математическое ожидание $M[\hat{Y}]=\beta^T a(x_0)$, то дисперсию для $\hat{Y}$ можно вычислить, используя следующие преобразования

$$D[\hat{Y}]=M[(\hat{Y}-M[\hat{Y}])^2]=M[(\hat{Y}-M[\hat{Y}])^T(\hat{Y}-M[\hat{Y}])]=$$

$$=M[a^T(x_0)\cdot(\tilde{\beta}^T-\beta^T)^T\cdot(\tilde{\beta}^T-\beta^T)a(x_0)]=$$

$$=a^T(x_0)\cdot M[(\tilde{\beta}-\beta)\cdot(\tilde{\beta}-\beta)^T]\cdot a(x_0)=$$

$$=a^T(x_0)\mathrm{cov}(\tilde{\beta})\cdot a(x_0)=\sigma^2 a^T(x_0)B^{-1}a(x_0).$$

Оценка дисперсии $\hat{Y}$ по результатам наблюдений равна $S^2_{\hat{Y}}$:

$$S^2_{\hat{Y}}=S^2 a^T(x_0)B^{-1}a(x_0),$$

где $S^2=\frac{Q_e}{n-k}$ — несмещенная оценка дисперсии ошибок наблюдений.

Предсказанное среднее значение $\hat{Y}$ имеет нормальное распределение, поэтому статистика

$$\frac{\hat{Y}-M[\hat{Y}]}{S_{\hat{Y}}}=T(n-k)$$

имеет распределение Стьюдента с $(n-k)$ степенями свободы. Границы доверительного интервала для предсказанного среднего значения при доверительной вероятности $(1-\alpha)$ имеют вид

$$\hat{Y}(x_0)\pm t_{1-\frac{\alpha}{2}}(n-k)S_{\hat{Y}}=\hat{Y}(x_0)\pm t_{1-\frac{\alpha}{2}}(n-k)S\sqrt{a^T(x_0)B^{-1}a(x_0)},$$

где $t_{1-\frac{\alpha}{2}}(n-k)$ — квантиль распределения Стьюдента с $(n-k)$ степенями свободы.

Границы доверительного интервала для предсказанного индивидуального значения вычисляются по формуле

$$\hat{Y}(x_0) \pm t_{1-\frac{\alpha}{2}}(n-k)S\sqrt{1 + a^T(x_0) \cdot B^{-1}a(x_0)}.$$

Границы доверительного интервала зависят от значения независимых переменных в точке x_0.

6.3.3. Дисперсионный анализ и проверка гипотез о параметрах линейной регрессии

Преобразуем остаточную сумму квадратов Q_e:

$$Q_e = e^T e = (Y - A\tilde{\beta})^T (Y - A\tilde{\beta}) = Y^T Y - 2\tilde{\beta}^T A^T Y + \tilde{\beta}^T A^T A\tilde{\beta}. \qquad (4)$$

МНК-оценки $\tilde{\beta}$ были получены из условия минимума Q_e и, следовательно, в силу необходимых условий минимума Q_e, выполняется условие

$$\frac{\partial Q_e}{\partial \beta}\Big|_{\beta=\tilde{\beta}} = 0.$$

Вычислив производную Q_e по вектору β: $\frac{\partial Q_e}{\partial \beta}\Big|_{\beta=\tilde{\beta}}$ и приравняв ее к нулю, получим

$$\frac{\partial Q_e}{\partial \beta}\Big|_{\beta=\tilde{\beta}} = -2A^T Y + 2A^T A\tilde{\beta} = 0.$$

Применив это соотношение в формуле (4), найдем, что остаточную сумму квадратов Q_e можно записать в следующем виде

$$Q_e = e^T e = Y^T Y - \tilde{\beta}^T A^T Y.$$

Обозначим сумму квадратов отклонений y_i от $\bar{y}$ как Q_y:

$$Q_y = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = Y^T Y - n\bar{y}^2,$$

где $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Тогда

$$Q_e = Y^T Y - \tilde{\beta}^T A^T A = Q_y - (\tilde{\beta}^T A^T Y - n\bar{y}^2) = Q_y - Q_R,$$

где

$$Q_R = \tilde{\beta}^T A^T Y - n\bar{y}^2 = \tilde{\beta}^T A^T Y - \frac{1}{n}(\sum y_i)^2.$$

Q_R называется ***суммой квадратов, обусловленной регрессией***. Таким образом, мы получили основное тождество дисперсионного анализа для линейной регрессии

$$Q_y = Q_R + Q_e.$$

Можно показать [22], что если модель адекватна данным, то статистики, входящие в основное тождество, независимы и связаны с распределением χ^2 следующим образом

$$\frac{Q_y}{\sigma^2} = \chi^2(n-1); \quad \frac{Q_R}{\sigma^2} = \chi^2(k-1); \quad \frac{Q_e}{\sigma^2} = \chi^2(n-k).$$

Этот результат используется для проверки гипотезы $H_0: \beta_1 = \beta_2 = ... = \beta_{k-1} = 0$, утверждающей, что факторы x_1, x_2, ..., x_{k-1} не улучшают предсказание Y по сравнению с $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Статистикой критерия для проверки гипотезы H_0 является отношение

$$F = \frac{Q_R/(k-1)}{Q_e/(n-k)}.$$

Если гипотеза H_0 верна, то F имеет распределение Фишера с $k-1$ и $n-k$ степенями свободы. Гипотеза H_0 принимается на уровне значимости α, если вычисленное значение статистики F меньше квантили распределения Фишера порядка $1-\alpha$: $F_{1-\alpha}(k-1, n-k)$. В этом случае говорят, что ***регрессионная модель незначима***. Если гипотеза H_0 отклоняется, то ***регрессионная модель*** называется ***значимой***.

Полезной характеристикой значимой регрессионной модели является ***коэффициент детерминации*** R^2, вычисляемый по формуле

$$R^2 = \frac{Q_R}{Q_y} = 1 - \frac{Q_e}{Q_y}.$$

Коэффициент детерминации равен той доле разброса результатов наблюдений y_i, $i = 1, 2, ..., n$ относительно горизонтальной прямой $y = \bar{y}$, которая объясняется регрессионной моделью. Арифметический корень из R^2 равен коэффициенту корреляции между наблюдаемыми y_i и предсказываемыми моделью значениями зависимой переменной $\tilde{y}_i$:

$$R_{y\tilde{y}} = +\sqrt{R^2} = R.$$

R называют также ***множественным коэффициентом корреляции***, так как он является мерой линейной зависимости между Y и факторами $x_1, x_2, ..., x_{k-1}$.

Приведенный коэффициент детерминации $\overline{R}^2$ вычисляется по формуле

$$\overline{R}^2 = 1 - \frac{Q_e/(n-k)}{Q_y/(n-1)}.$$

Относительно параметров регрессии можно также проверить гипотезу $H_0: \beta_i = 0$, $i = 1, 2, ..., k-1$, утверждающую, что переменная x_i не улучшает предсказание Y по сравнению с предсказанием, получаемым с помощью регрессии Y по остальным $(k-2)$ факторам. Гипотеза $H_0: \beta_i = 0$, как и гипотеза $H_0: \beta_i = \beta_i^{(0)}$, где $\beta_i^{(0)}$ — заданная константа, проверяется с помощью доверительного интервала для β_i: если доверительный интервал для β_i на-

крывает число $\beta_i^{(0)}$, то гипотеза $H_0: \beta_i = \beta_i^{(0)}$ принимается на уровне значимости α.

Рассмотрим проверку еще одной гипотезы: о равенстве нулю некоторого подмножества из $(k-1)$ параметров регрессии. Предположим, что это подмножество состоит из первых m параметров. Тогда гипотеза $H_0: \beta_1 = \beta_2 = ... = \beta_m = 0$ утверждает, что введение в регрессионную модель m факторов $x_1, x_2, ..., x_m$ не улучшает предсказание Y по сравнению с предсказанием, полученным с помощью регрессии Y по факторам $x_{m+1}, x_{m+2}, ..., x_{k-1}$. Для проверки H_0 сначала вычисляются параметры регрессии Y по факторам $x_{m+1}, x_{m+2}, ..., x_{k-1}$ и находится остаточная сумма квадратов Q_e' для «урезанной» модели, а затем определяются параметры регрессии Y по факторам $x_1, x_2, ..., x_{k-1}$ и вычисляется остаточная сумма квадратов Q_e для полной модели. Статистика критерия равна отношению

$$F = \frac{(Q_e' - Q_e)/m}{Q_e/(n-k)}.$$

Если гипотеза H_0 верна, то F имеет распределение Фишера с m и $n-k$ степенями свободы. H_0 принимается на уровне значимости α, если $F < F_{1-\alpha}(m, n-k)$, где $F_{1-\alpha}(m, n-k)$ — квантиль распределения Фишера порядка $1-\alpha$.

6.3.4. Проверка адекватности модели

Для проверки адекватности полученной модели результатам наблюдений надо найти остатки, т. е. разности между наблюдаемыми и предсказанными моделью значениями переменной Y. Вектор остатков равен

$$e = Y - A\tilde{\beta}.$$

Далее вычисляются статистики, на основе которых можно проверить выполнение основных предположений регрессионного анализа и адекватность полученной модели (см. выше п. 6.1.3).

Для адекватной модели, кроме некоррелированности остатков, их нормального распределения, должно выполняться условие гомоскедаксичности, т. е. постоянства дисперсии ошибок наблюдений для всех наблюдений. Оценка выполнимости этого условия проводится по графику остатков в зависимости от номера наблюдений: если все остатки укладываются в симметричную относительно нулевой линии полосу, то, можно считать, что дисперсия ошибок наблюдений постоянна.

Более тщательная проверка адекватности регрессионной модели может быть проведена, если для зависимой переменной Y проведены повторные наблюдения. В этом случае для проверки адекватности модели используется следующая процедура дисперсионного анализа.

Пусть при i-м наборе независимых переменных проведено n_i повторных наблюдений переменной Y, $i = 1, 2, ..., m$. Объем всей выборки $n = \sum_{i=1}^{m} n_i$. Обозначим y_{ij}, $j = 1, 2, ..., n_i$ результаты повторных наблюдений Y

при i-м наборе независимых переменных. Если модель адекватна данным, то средние $\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n_i} y_{ij}$, $j = 1, 2, ..., m$ должны быть близки к значениям $\tilde{y}_i$, предсказанным регрессионной моделью

$$\tilde{Y} = A\tilde{\beta}.$$

Мерой неадекватности модели будет сумма квадратов

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

Чем меньше будет Q_n, тем лучше результаты наблюдений согласуются с моделью. Возведя обе части тождества $y_{ij} - \tilde{y}_i = (\bar{y}_i - \tilde{y}_i) + (y_{ij} - \bar{y}_i)$ в квадрат и просуммировав их по i и по j, получим, что остаточная сумма квадратов может быть разбита на две суммы

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

или

$$Q_e = Q_n + Q_p,$$

где второе слагаемое Q_p называется суммой квадратов чистой ошибки.

Если модель адекватна результатам наблюдений, то статистики $\frac{Q_n}{\sigma^2}$ и $\frac{Q_p}{\sigma^2}$ независимы и имеют распределение χ^2 соответственно с $(m - k)$ и $(n - m)$ степенями свободы, где k — число оцениваемых параметров в уравнении множественной регрессии.

В этом случае статистика

$$F = \frac{Q_n/(m-k)}{Q_p/(n-m)}$$

имеет распределение Фишера с $(m - k)$ и $(n - m)$ степенями свободы.

Вычисленное значение статистики сравнивается с квантилью распределения Фишера $F_{1-\alpha}(m-k, n-m)$. Если $F < F_{1-\alpha}(m-k, n-m)$, то гипотеза об адекватности модели принимается на уровне значимости α.

6.3.5. Вычислительные проблемы регрессионного анализа: мультиколлинеарность и плохая обусловленность информационной матрицы

Вычисление МНК-оценок параметров линейной регрессионной модели по формуле

$$\tilde{\beta} = (A^T A)^{-1} A^T Y$$

предполагает, что регрессионная матрица A имеет ранг k, где k — число параметров модели, а информационная матрица

$$B = A^T A$$

является невырожденной, т. е. определитель матрицы B не равен нулю: $|B| \neq 0$.

Если $|B| = 0$, то ранг матрицы A будет меньше k. Это условие является следствием того, что между столбцами матрицы A существует линейная зависимость, т. е. хотя бы один из них является линейной комбинацией других столбцов. Если столбцы матрицы A рассматривать как векторы в n-мерном линейном пространстве, то некоторые из них будут коллинеарны. Это явление называется **строгой мультиколлинеарностью**. Случай, когда линейная зависимость между столбцами матрицы выполняется лишь приблизительно, т. е. когда $|B| \approx 0$, называется **мультиколлинеарностью**. При этом одно или несколько собственных чисел матрицы $B = A^T A$ и определитель матрицы будут близки к нулю.

Основные следствия мультиколлинеарности таковы:

1) падает точность оценивания: ошибки оценок некоторых параметров становятся очень большими, резко возрастают дисперсии оценок;

2) коэффициенты регрессионной модели сильно коррелированны между собой, что затрудняет их интерпретацию;

3) некоторые переменные становятся незначимыми и должны исключаться из модели, хотя истинная причина состоит не в том, что эти переменные не влияют на зависимую переменную, а в том, что выборочные данные не позволяют отобразить это влияние;

4) оценки параметров становятся неустойчивыми: добавление нескольких наблюдений приводит к большим изменениям в оценках параметров.

Мультиколлинеарность имеет место при больших значениях элементов обратной матрицы $B^{-1} = (A^T A)^{-1}$ и, следовательно, при больших значениях дисперсий оценок параметров регрессии. В этом случае и предсказание, и остаточная сумма квадратов будут неточными. Мультиколлинеарность можно легко определить, используя стандартизацию исходных данных.

Рассмотрим модель множественной регрессии

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \varepsilon. \qquad (*)$$

Преобразуем исходные данные: — зависимую переменную Y и факторы переменные x_1, x_2, ..., x_{k-1}, используя процедуру стандартизации по следующим формулам:

$$y'_i = \frac{y_i - \bar{y}}{s_y}, \quad x'_{ji} = \frac{x_{ji} - \bar{x}_j}{s_{x_j}},$$

где $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$, $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ji}$, $s_{x_j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2}$,

$j = 1, 2, \ldots, k-1$, $i = 1, 2, \ldots, n$.

Регрессионная модель в этом случае имеет вид

$$y'_i = \beta'_1 x'_{1i} + \beta'_2 x'_{2i} + \ldots + \beta'_{k-1} x'_{k-1i} + \varepsilon_i. \qquad (**)$$

В этой модели свободный член (аналогичный β_0) отсутствует.

Записав регрессионную матрицу A' и вектор Y' в виде

$$A' = \begin{pmatrix} x'_{11} & x'_{21} & \dots & x'_{k-11} \\ x'_{12} & x'_{22} & \dots & x'_{k-12} \\ \dots & \dots & \dots & \dots \\ x'_{1n} & x'_{2n} & \dots & x'_{k-1n} \end{pmatrix}, \quad Y' = \begin{pmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_n \end{pmatrix},$$

представим регрессионную модель так:

$$Y' = A'\beta' + \varepsilon.$$

Элементы информационной матрицы $(A')^T A' = B'$ равны

$$B'_{fg} = \sum_{i=1}^{n} x'_{fi} x'_{gi} = \frac{\sum_{i=1}^{n} (x_{fi} - \bar{x}_f)(x_{gi} - \bar{x}_g)}{\sqrt{\sum_{i=1}^{n} (x_{fi} - \bar{x}_f)^2 (x_{gi} - \bar{x}_g)^2}},$$

где $f, g = 1, 2, \dots, k-1$.

Таким образом, элементами матрицы B' являются коэффициенты корреляции между факторами $x_1, x_2, \dots, x_{k-1}$:

$$B' = R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1(k-1)} \\ r_{21} & 1 & \dots & r_{2(k-1)} \\ \dots & \dots & \ddots & \dots \\ r_{(k-1)1} & r_{(k-1)2} & \dots & 1 \end{pmatrix}.$$

Между параметрами моделей (*) и (**) существуют простые соотношения

$$\beta_j = \beta'_j \cdot \frac{s_y}{s_{x_j}}, \quad j = 1,2, \dots, k-1; \quad \beta_0 = \bar{y} - \sum_{j=1}^{k-1} \left(\beta'_j \cdot \frac{s_y}{s_{x_j}} \right) \bar{x}_j = \bar{y} - \sum_{j=1}^{k-1} (\beta_j \bar{x}_j).$$

Если какой-либо внедиагональный элемент матрицы R, $|r_{fg}| \geq 0{,}9$, то между соответствующими переменными x_f и x_g существует сильная линейная зависимость и возможны эффекты мультиколлинеарности.

Наличие мультиколлинеарности определяется и по определителю корреляционной матрицы: в этом случае $|R| \ll 1$, при строгой мультиколлинеарности $|R| = 0$.

Более точен следующий метод [33]. Строгая мултиколлинеарность означает, что некоторый фактор x_i есть линейная комбинация всех других факторов или подмножества этих факторов. Возьмем в качестве зависимой переменной x_i, а в качестве факторов — все остальные факторы x_i, $i = 1, 2, \dots, i-1, i+1, \dots, k-1$. Найдем оценки параметров множественной регрессии и коэффицент детерминации R_i^2. Переменную x_i относят к мультиколлинеарным, если коэффициент детерминации R_i^2, определяющий ли-

нейную зависимость фактора x_i от всех других факторов, больше коэффициента детерминации R^2 между зависимой переменной Y и всеми факторами. Для каждой переменной x_i вычисляют F-статистику

$$F_i = \frac{R_i^2/(k-2)}{(1-R_i^2)/(n-k+1)}, \quad i = 1, 2, ..., k$$

и проверяют значимость каждой статистики F_i, сравнивая ее с квантилью распределения Фишера $F_{1-\alpha}(k-2, n-k+1)$. Значения F_i показывают, какие из переменных могут быть в большей степени подвержены мультиколлинеарности.

В линейной алгебре для определения мультиколлинеарности используют числа обусловленности. Число обусловленности определяется как отношение максимального собственного числа λ_{max} матрицы $B = (A^TA)$ к ее минимальному собственному числу λ_{min}.

Можно считать, что при $\frac{\lambda_{max}}{\lambda_{min}} \geq 10^5 — 10^6$ имеет место сильная мультиколлинеарность.

Плохая обусловленность матрицы B часто проявляется при использовании полиномиальных моделей, особенно если измерения фактора x выполняются через равные интервалы, а при высоких степенях x (выше 4) проявляется почти всегда.

Для устранения или уменьшения мультиколлинеарности используются следующие меры:

1. Привлечение дополнительной информации. В некоторых случаях (например при исследовании технологических процессов) устранить мультиколлиниарность можно добавляя к исходным данным результаты новых наблюдений с использованием методов планирования эксперимента [5].

2. Преобразование множества факторов в несколько ортогональных множеств. При этом применяются методы многомерного статистического анализа: факторный анализ и метод главных компонент, а также специальные процедуры регрессионного анализа: регрессия на главные компоненты, гребневая регрессия. Для полиномиальной регрессии вместо системы функций 1, x, x^2, ... следует использовать функции 1, $(x-\bar{x})$, $(x-\bar{x})^2$, ... либо (что более эффективно) ортогональную систему функций, например полиномы Чебышева [4, 18].

3. Стандартизация и центрирование исходных данных.

4. Последовательное включение факторов в регрессионную модель. При этом на каждом шаге необходимо определить улучшит ли новый фактор модель, появляются или нет признаки мультиколлинеарности.

6.3.6. Пример множественной регрессии

Рассмотрим пример множественной регрессии с тремя факторами и одной зависимой переменной.

Пример 6.2. Руководство авиакомпании по результатам анализа деятельности 15 своих представительств получило следующие данные за март месяц:

Y	x_1	x_2	x_3
79,3	2,5	10,0	3,0
200,1	5,5	8,0	6,0
163,2	6,0	12,0	9,0
200,1	7,9	7,0	16,0
146,0	5,2	8,0	15,0
177,7	7,6	12,0	9,0
30,9	2,0	12,0	8,0
291,9	9,0	5,0	10,0
160,0	4,0	8,0	4,0
339,4	9,6	5,0	16,0
159,6	5,5	11,0	7,0
88,3	3,0	12,0	8,0
237,5	6,0	6,0	10,0
107,2	5,0	10,0	4,0
155,0	3,5	10,0	4,0

где Y (зависимая переменная) — общий доход от проданных билетов, млн руб.; x_1 — средства на развитие компаний в регионе, млн руб.; x_2 — число конкурирующих компаний; x_3 — процент пассажиров, летавших бесплатно.

Найти уравнение множественной регрессии. Проверить значимость и адекватность регрессионной модели. Существенно ли влияет на доход число пассажиров, летавших бесплатно? Какой доход (в среднем) может ожидать компания, вложившая в развитие 2,5 млн руб., если число конкурирующих компаний в регионе равно десяти, а число пассажиров, летавших бесплатно по разным причинам, составляет 3 %. Принять уровень значимости $\alpha = 0{,}05$.

Решение в пакете STATISTICA. Проведите те же операции в модуле **Multiple Regression**, что и в работе **6.1**: введите данные — **Analysis** → **Startup Panel**: ***Variables*: *dependent var-Y*, *independent var-X1*, *X2*, *X3*, OK** → **Regression Summary**. Результаты регрессионного анализа приведены на рис. 6.11.

Уравнение множественной регрессии имеет вид: $Y = 170{,}76 + 25{,}42x_1 - 13x_2 - 2{,}7x_3$.

Из данной таблицы видно, что гипотеза $H_0: \beta_3 = 0$ принимается на уровне значимости $p = 0{,}267$, так как $p > \alpha = 0{,}05$. Остальные коэффициенты регрессионной модели значимы.

Regression Summary for Dependent Variable: VAR1

Continue..

R= ,95117377 RI= ,90473155 Adjusted RI= ,87874924
F(3,11)=34,82: p<,00001 Std.Error of estimate: 27,798

N=15	BETA	St. Err. of BETA	B	St. Err. of B	t(11)	p-level
Intercpt			170,7600	52,08625	3,27841	,007355
VAR2	,731272	,139913	25,4233	4,86420	5,22661	,000283
VAR3	-,415135	,119008	-13,0035	3,72776	-3,48829	,005074
VAR4	-,145960	,124879	-2,7059	2,31510	-1,16881	,267186

Рис. 6.11. Результаты регрессионного анализа

Проверим гипотезу о незначимости регрессионной модели. Для этого используем опцию **Analysis of Variance** (дисперсионный анализ).

Результаты дисперсионного анализа приведены в таблице (рис. 6.11*а*).

Из таблицы видно, что статистика критерия Фишера, вычисляемая по формуле

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

равна $F(3,11) = 34{,}821$, так как $p = 0{,}000007$, что меньше, чем $\alpha = 0{,}05$, то гипотеза о незначимости модели отклоняется.

Так как коэффициент β_3 незначим, пересчитаем уравнение множественной регрессии используя два фактора x_1 и x_2. Результаты регрессионного анализа приводятся на рис. 6.12.

Уравнение множественной регрессии имеет вид: $Y = 159{,}86 + 22{,}39x_1 - 12{,}53x_2$.

Коэффициенты регрессионной модели β_0, β_1, β_2 значимы (соответствующие уровни значимости равны соответственно: 0,009; 0,00017; 0,0059).

Analysis of Variance; DV: VAR1 (mhozh_reg.s...

Continue..

	Sums of Squares	df	Mean Squares	F	p-level
Regress.	80720,39	3	26906,80	34,82106	,000007
Residual	8499,88	11	772,72		
Total	89220,26				

Рис. 6.11*а*. Таблица дисперсионного анализа

Continue..

R= ,94493383 RI= ,89289994 Adjusted RI= ,87504993
F(2,12)=50,022 p<,00000 Std.Error of estimate: 28,219

N=15	BETA	St. Err. of BETA	B	St. Err. of B	t(12)
Intercpt			159,8629	52,02090	3,07305
X1	,643971	,120099	22,3882	4,17535	5,36199
X2	-,400069	,120099	-12,5316	3,76193	-3,33115

Рис. 6.12. Результаты регрессионного анализа (факторы x_1 и x_2)

Регрессионная модель значима: $F = 50{,}022$, уровень значимости $p = 0{,}000002$.

Чтобы проверить выполнение предположений регрессионного анализа и адекватность модели рассмотрим остатки. Для этого используем опцию Residual Analysis (анализ остатков).

Начнем с проверки гипотезы о том, что все сериальные корреляции в последовательности остатков равны нулю (гипотеза H_0). Для проверки этой гипотезы используется критерий Дарбина—Уотсона (см. выше п. 6.1.3).

Чтобы проверить гипотезу H_0, в окне **Multiple Regression Results** (рис. 6.4) выберите опцию **Residual Analysis** (рис. 6.6), а затем — **Durbin-Watson stat**. Результат приводится на рис. 6.12*а*.

Durbin-Watson d (62.sta)

Summary: and serial correlation of residuals

	Durbin-Watson d	Serial Corr.
Estimate	1,896936	-,058644

Рис. 6.12*а*. Статистика Дарбина—Уотсона

В данном случае статистика Дарбина—Уотсона $d = 1{,}8969$, что больше табличного значения $d_2 = 1{,}75$ (см. Приложение 2), следовательно, гипотеза H_0 : все сериальные корреляции равны нулю принимается на уровне значимости $2\alpha = 0{,}1$.

Построим график остатков. Для этого в окне **Residual Analysis** (рис. 6.6) нужно выбрать опцию **Plots of Residuals (A)**. Результаты приводятся на рис. 6.13.

Все остатки укладываются в симметричную относительно нулевой линии полосу шириной $\pm 2S$. Это означает, что, по-видимому, дисперсии ошибок наблюдений постоянны.

Window Dependent variable: Y

Raw Residuals -3s . . 0 . . +3s

	Residual	Standard Pred. y.	Stande Residu
1	-11,2178	-1,04146	-,3
2	17,3545	,18116	,6
3	19,3866	-,33495	,6
4	-48,9087	1,05958	-1,7
5	-30,0291	,09212	-1,0
6	-1,9345	,13992	-,0
7	-23,3606	-1,52211	-,8
8	-6,7988	1,71830	-,2
9	10,8368	-,26403	,3
10	27,2682	1,89637	,9
11	14,4492	-,31722	,5
12	11,6512	-1,22532	,4
13	18,4973	,66180	,6
14	-39,2883	-,29949	-1,3
15	42,0940	-,74467	1,4
4	-48,9087	-1,52211	-1,7

Рис. 6.13. График остатков

Теперь проверим гипотезу о нормальности распределения остатков. Для этого в том же окне (**Residual Analysis**) необходимо выбрать опцию **Normal Probability Plot of Residuals**. Результаты выполнения процедуры представлены на специальном графике (рис. 6.14).

Из графика (рис. 6.14) видно, что точки расположены близко к прямой, значит, можно предположить, что остатки распределены по нормальному закону. Гипотезу о нормальном распределении остатков можно также проверить по критерию χ^2 или критерию Колмогорова—Смирнова.

Таким образом, можно считать, что предположения регрессионного анализа выполняются. Распределение остатков на рис. 6.13 (случайное, без каких-либо закономерностей) показывает, что регрессионная модель адекватна результатам наблюдений и может быть использована для прогнозирования. Для выполнения прогноза в окне **Multiple Regression Results** (рис. 6.4) нужно выбрать опцию **Predict Dependent Var**, в появившемся окне нужно ввести значения факторов x_1, x_2 и задать уровень значимости $\alpha = 0{,}05$.

В появившемся окне (рис. 6.15) приведены результаты прогноза: при $x_1 = 2{,}5$, $x_2 = 10$: в первом столбце приведены оценки параметров регрессии $\tilde{\beta}_i$, $i = 1, 2$; во втором — значения факторов x_i.

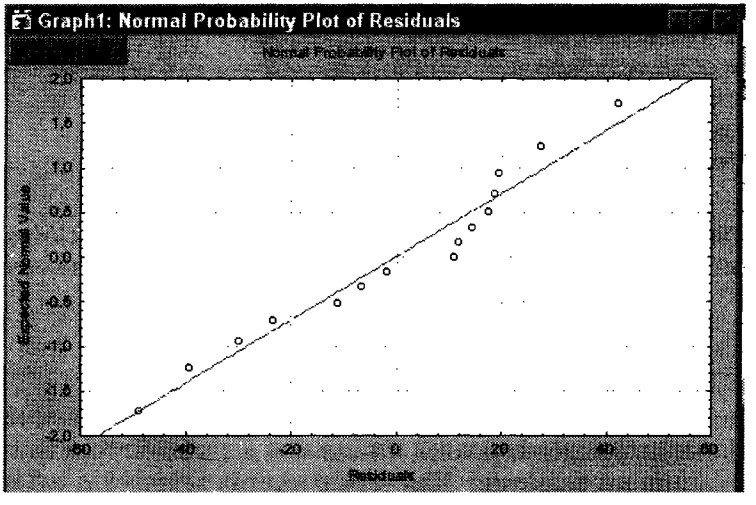


Рис. 6.14. Остатки на графике **Normal Probability Plot**

Predicting Values for (62.sta)

	variable: Y		
variable	B-Weight	Value	B-Weight * Value
X1	22,3882	2,50000	55,970
X2	-12,5316	10,00000	-125,316
Intercpt			159,863
Predictd			90,518
-95,0%CL			62,374
+95,0%CL			118,662

Рис. 6.15. Результаты прогноза

Предсказываемое значение Y выведено в строке **Predicted**, ниже вычислены 95 % доверительные интервалы для среднего предсказанного значения $Y = 90{,}518$.

6.3.7. Задания для самостоятельного решения

Задание 1

В приложении 1.1 (табл. П1 и П2) приведены варианты 10 заданий по регрессионному анализу. Для каждой задачи выполните следующие задания.

1. Найдите уравнение множественной регрессии. Выполните дисперсионный анализ. Проверьте значимость регрессионной модели. Найдите оценку дисперсии ошибок наблюдений, коэффициенты детерминации и множественной корреляции. Определите доверительные интервалы для параметров регрессии, проверьте гипотезу о значимости параметров и гипотезу $H_0: \beta_1 = \beta_2 = 0$, где β_1 и β_2 — коэффициенты регрессии для первого и второго факторов.

2. Определите остатки. Постройте график остатков. Проверьте выполнение предположений регрессионного анализа:

- дисперсия остатков постоянна;
- остатки некоррелированны;
- остатки имеют нормальное распределение $N(0,\sigma^2)$.

Сделайте вывод об адекватности регрессионной модели результатам наблюдений.

3. Используя модель множественной регрессии, определите предсказанное значение зависимой переменной Y при следующих значениях факторов

$$x_{0i} = \max x_i + 2S_i,$$

где S_i — оценка среднего квадратического отклонения переменной x_i, $i = 1, 2, \ldots, p$. Определите доверительные интервалы для среднего и индивидуального предсказанного значения. Для всех расчетов принять $\alpha = 0{,}05$.

Задание 2

Решите следующие задачи и проведите полный статистический анализ результатов.

Задачи.

1. Используя приведенные ниже данные, найдите уравнение множественной регрессии и ответьте на следующие вопросы:

1) каковы оценки коэффициентов регрессии и стандартные ошибки этих оценок?

2) каков коэффициент детерминации?

3) каково ожидаемое или прогнозируемое значение для Y при $x_1 = 5{,}8$, $x_2 = 4{,}2$, $x_3 = 5{,}1$?

Y	x_1	x_2	x_3
64,7	3,5	5,3	8,5
80,9	7,4	1,6	2,6
24,6	2,5	6,3	4,5
43,9	3,7	9,4	8,8
77,7	5,5	1,4	3,6
20,6	8,3	9,2	2,5
66,9	6,7	2,5	2,7
34,3	1,2	2,2	1,3

2. Используя приведенные ниже данные, найдите уравнение множественной регрессии и ответьте на следующие вопросы:

1) каковы оценки коэффициентов регрессии и стандартные ошибки этих оценок?

2) каков коэффициент детерминации?

3) каков 95%-й доверительный интервал для предсказанного среднего значения Y при x_1, x_2, x_3 и x_4, равных 52,4; 41,6; 35,8; 3 соответственно?

x_1	x_2	x_3	x_4	Y
21,4	62,9	21,9	–2	22,8
51,7	40,7	42,9	5	93,7
41,8	81,8	69,8	2	64,9
11,8	41,0	90,9	–4	19,2
71,6	22,6	12,9	8	55,8
91,9	61,5	30,9	1	23,1

3. Владелец бухгалтерской фирмы считает, что целесообразно прогнозировать заранее количество налоговых деклараций, приходящихся на период с 1 марта по 15 апреля, так как в этом случае он сможет лучше спланировать работу на этот период. Он предполагает, что при таком прогнозе могут быть использованы следующие факторы. Данные об этих факторах и количестве налоговых деклараций приведены ниже:

Экономический индекс, x_1	Население в радиусе 1 км от фирмы, x_2, тыс. чел.	Средний доход в районе, x_3, тыс. руб.	Количество деклараций на период с 1 марта по 15 апреля, Y, тыс.
99	10,188	21,465	2,306
106	8,566	22,228	1,266
100	10,557	27,665	1,422
129	10,219	25,200	1,721
179	9,662	26,300	2,544

1. Определите уравнение множественной регрессии для этих данных.
2. Какой процент дисперсии данных описывается этим уравнением?

4. Мы пытаемся предсказать годовой спрос на продукцию, используя следующие факторы:

— цена за одну единицу продукции, руб.;
— доход потребителя, руб.;
— замена (цена на заменитель этого товара), руб.

Данные были собраны за 14 лет.

Год	Спрос	Цена	Доход	Замена
1	40	9	400	10
2	45	8	500	14
3	55	8	700	13
4	60	7	800	11
5	70	6	900	15
6	65	6	1000	16
7	65	8	1100	17
8	75	5	1200	22
9	75	5	1300	19
10	80	5	1400	20
11	100	3	1500	23
12	90	4	1600	18
13	95	3	1700	24
14	85	4	1800	21

1. Можно ли точно определить знак («+» или «–») регрессионных коэффициентов для факторов. Дайте краткое объяснение. (Заметьте это не статистический вопрос. Вы должны просто подумать о смысле регрессионного коэффициента.)
2. Найдите уравнение множественной регрессии.
3. Найдите коэффициент детерминации для данного примера и объясните его смысл.
4. Найдите оценку дисперсии ошибок наблюдений для данного примера и объясните ее смысл.
5. Используя уравнение, оцените, какой спрос можно ожидать, если цена — 6 руб., доход — 1200 руб. и цена на заменитель — 17 руб.

5. Леня Голубков собирается продать дом. Чтобы решить, какую цену запросить, он собрал данные о 12 недавних продажах. Он принимал во внимание цену, площадь дома, количество этажей, количество ванных и возраст дома.

Цена, тыс. долл.	Площадь, сотни кв. футов	Этажность	Количество ванных	Возраст дома, года
49,65	8,9	1	1,0	2
67,95	9,5	1	1,0	6
81,15	12,6	2	1,5	11
81,60	12,9	2	1,5	8
91,50	19,0	2	1,0	22
95,25	17,6	1	1,0	17
100,35	20,0	2	1,5	12
104,25	20,6	2	1,5	11
112,65	20,5	1	2,0	9
149,70	25,1	2	2,0	8
160,65	22,7	2	2,0	18
232,50	40,8	3	4,0	12

Примечание. 1,5 ванны означает, что в доме имеется одна ванная и одна комната с душем.

1. Найдите уравнение множественной регрессии.
2. Каков коэффициент детерминации для этого уравнения и что он определяет?
3. Если дом имеет площадь 1800 кв. футов, один этаж, полторы ванны и возраст 6 лет, по какой цене Леня сможет продать дом?

6. Сталелитейная корпорация рассматривала факторы, влияющие на объем ежегодно продаваемой стали. Руководство предполагает, что важнейшими являются следующие факторы: годовой национальный уровень инфляции, средняя цена за тонну импортируемой стали, сбивающая цены корпорации, количество автомобилей, которое планируют выпустить производители машин в данном году. Были собраны следующие данные за последние 7 лет:

Год	Продано, млн т	Темп инфляции, %	Средняя цена за тонну импортируемой стали, тыс. долл.	Количество машин, млн штук
1	4,2	3,1	3,10	6,2
2	3,1	3,9	5,00	5,1
3	4,0	7,5	2,20	5,7
4	4,7	10,7	4,50	7,1
5	4,3	15,5	4,35	6,5
6	3,7	13,0	2,60	6,1
7	3,5	11,0	3,05	5,9

1. Найдите уравнение множественной регрессии для этих данных.

2. Какой процент дисперсии данных описывается этим уравнением?

3. Сколько тонн стали сможет продать корпорация, если в данном году темп инфляции будет 7,1, автомобилестроители планируют выпустить 6 млн машин, а средняя цена импортной стали — 3,5 долл. за т?

Задание 3

В приложении 1.2 приведены данные о стоимости однокомнатных квартир. Проведите регрессионный анализ этих данных взяв в качестве зависимой переменной стоимость квартир, а в качестве факторов данные о площадях квартиры, кухни, жилой комнаты, а также удаленность квартиры от центра и метро. Определите уравнения множественной регрессии используя следующие множества исходных данных:

а) все данные;

б) данные о квартирах не расположенных на первых и последних этажах;

в) данные о квартирах расположенных в некирпичных домах и не на первых и не на последних этажах;

г) данные о квартирах расположенных в Коньково.

Проверьте значимость и адекватность регрессионных моделей на уровне значимости $\alpha = 0{,}05$.

Определите среднюю стоимость однокомнатной квартиры в одном из районов Юго-западной части Москвы, если общая площадь равна 40 м^2, площадь кухни равна 9,5 м^2, а площадь комнаты составляет 21 м^2, при условии, что удаленность от центра равна 9 км, а до ближайшей станции метро надо пройти 7 мин. Найдите 95%-е доверительные интервалы для среднего и индивидуального предсказанного значения. Какова будет средняя стоимость данной квартиры, если она располагается не в кирпичном доме и не на первом или не на последнем этажах?

Сколько будет стоить данная квартира в Коньково?

6.4. Пошаговая регрессия

В случае множественной регрессии с большим числом факторов необходимо классифицировать эти переменные по степени их важности для предсказания зависимой переменной *Y*. Далее, следует исключить из анализа факторы наименее существенные для предсказания *Y*, а также переменные сильно коррелированные с другими, уже включенными в анализ, факторами. Эту задачу можно определить как выбор «наилучшей» регрессии, т. е. определение минимального набора факторов, достаточно точно предсказывающих *Y*.

Одним из методов отбора наиболее существенных факторов является пошаговая регрессия (**Stepwise Multiple Regression**). В пакете STATISTICA реализованы две процедуры пошаговой регрессии **Backward stepwise** и **Forward stepwise**.

Процедура **Backward stepwise** (называемая также метод исключения) состоит в следующем.

1. На нулевом шаге проводится регрессионный анализ для всех факторов. Каждый фактор x_m, $m = 1, 2, ..., k - 1$ проверяется на значимость, что равносильно проверке гипотезы о незначимости соответствующего коэффициента регрессии, т. е. проверке гипотезы $H_0: \beta_m = 0$. В процедуре пошаговой регрессии для проверки гипотезы H_0 используется статистика F_m: $F_m = (t_m)^2$, где t_m — статистика Стьюдента

$$t_m = \frac{\tilde{\beta}_m}{\sqrt{D[\tilde{\beta}_m]}}.$$

Если гипотеза H_0 верна, то t_m имеет распределение Стьюдента с $(n - k)$ степенями свободы, где n — число наблюдений, k — число оцениваемых параметров, а F_m имеет распределение Фишера $F(1, n - k)$ соответственно с одной и $(n - k)$ степенями свободы.

2. Наименьшая величина F_m, $m = 1, 2, ..., k - 1$, например F_e, сравнивается с заданным значением F_0 — величиной F-удаления (F-to remove).

Если $F_e < F_0$, то фактор x_l исключаем из анализа и рассчитываем новое уравнение регрессии по $(k - 2)$ факторам. Далее переходим к следующему шагу.

Если $F_e > F_0$, то регрессионное уравнение остается без изменений.

Пример 6.3. Рассмотрим работу процедуры **Backward stepwise** на примере следующих данных [32]:

x_1	x_2	x_3	x_4	Y
7	26	6	60	78,5
1	29	15	52	74,3
11	56	8	20	104,3
11	31	8	47	87,6
7	52	6	33	95,9
11	55	9	22	109,2
3	71	17	6	102,7
1	31	22	44	72,5
2	54	18	22	93,1
21	47	4	26	115,9
1	40	23	34	83,8
11	66	9	12	113,3
10	68	8	12	109,4

Решение в пакете STATISTICA. Вводим данные и переключаемся в модуль **Multiple Regression**. Вводим переменные: **Dependent var: Y**, **Independent var: X1, X2, X3, X4**. Чтобы выполнить пошаговую регрессию, на стартовой панели модуля (рис. 6.3) нужно снять ограничение: **Perform default (non stepwise) analysis** и нажать **OK**. В появившейся панели **Model Definition** (рис. 6.16) нужно выполнить следующие установки: **Metod: Backward stepwise; Display result: At each step** (в этом случае будут выводиться результаты на каждом шаге). Значения *F*-включения (***F*-to enter**) и *F*-удаления (***F*-to remove**) по умолчанию равны 11 и 10. Их можно изменять в зависимости от опыта и конкретной задачи (обычно их устанавливают соответственно равными 4 и 3,9). Остальные установки можно оставить без изменения.

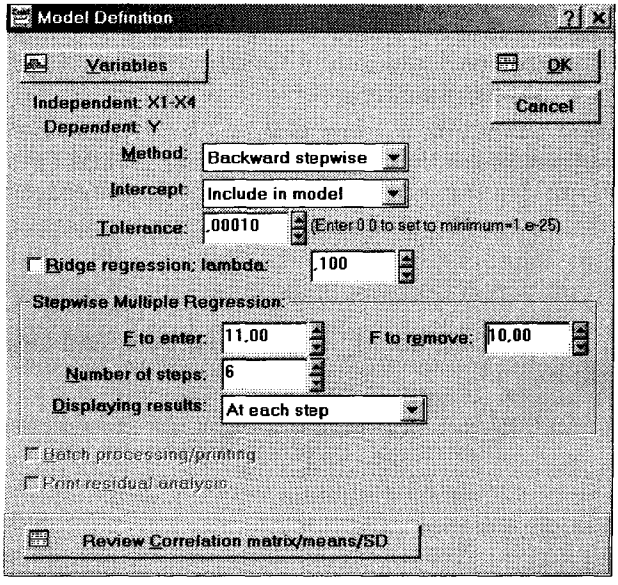


Рис. 6.16. Панель Model Definition

На нулевом шаге в уравнение регрессии вводятся все четыре фактора. Нажав кнопку **Regression Summary**, получим следующие результаты (рис. 6.17):

Regression Summary for Dependent Variable: Y (15.sta)

Continue...

R= ,99114864 RI= ,98237562 Adjusted RI= ,97356343
F(4,8)=111,48 p<,00000 Std.Error of estimate: 2,4460

N=13	BETA	St. Err. of BETA	B	St. Err. of B	t(8)	p-level
Intercpt			62,40537	70,07096	,890602	,399134
X1	,606512	,291220	1,55110	,74477	2,082660	,070822
X2	,527706	,748670	,51017	,72379	,704858	,500901
X3	,043390	,321330	,10191	,75471	,135031	,895923
X4	-,160287	,788917	-,14406	,70905	-,203174	,844072

Рис. 6.17. Результаты на нулевом шаге пошаговой регрессии

Наименьшее значение (по модудю) t-статистики, $t_3(8) \approx 0{,}135$ имеет переменная x_3, соответствующий коэффициент регрессии β_3 — незначим ($H_0: \beta_3 = 0$ принимается, так как p-level = 0,895). Значение F-статистики, $F_3 = 0{,}0182$ меньше, чем ***F*-to remove**, следовательно, переменная x_3 исключается из дальнейшего анализа. На втором шаге исключается переменная x_4 ($t_4(8) = -1{,}365$, $F_4 = 1{,}863$, $p = 0{,}205$). На третьем (последнем) шаге обе оставшиеся переменные x_1 и x_2 не исключаются, так как соответствующие статистики $t_1(10) = 12{,}104$; $F_1 = 146{,}5$ и $t_2(10) = 14{,}442$, $F_2 = 208{,}6$ больше значения ***F*-to remove**. Оба коэффициента регрессии β_1 и β_2 — значимы. Окончательное уравнение регрессии:

$$Y = 52{,}577 + 1{,}468x_1 + 0{,}662x_2.$$

Это уравнение значимо ($F = 229{,}5$). Нажав кнопку **Stepwise (summary)**, получим результаты выполнения процедуры по шагам, в частности, изменение коэффициента детерминации R^2 (рис. 6.18).

Summary of Stepwise Regression; DV: Y (15.sta)

MULTIPLE REGRESS.	Step +in/-out	Multiple R	Multiple R-square	R-square change	F - to enter/rem	p-level	Variable included
X3	-1	,991128	,982335	-,000040	,018233	,895266	3
X4	-2	,989282	,978678	-,003657	1,863262	,202170	2

Рис. 6.18. Результаты пошаговой регрессии

Процедура **Forward Stepwise** организована в обратном направлении по сравнению с предыдущей процедурой удаления переменных.

1. На первом шаге в уравнение регрессии включается фактор, имеющий наибольший коэффициент корреляции с Y, например x_l. Определяется уравнение простой линейной регрессии: $Y = \beta_0 + \beta_1 x_l$ и проверяется значимость x_l, т. е. проверяется гипотеза $H_0: \beta_1 = 0$.

2. Если x_l значима, то на втором шаге вычисляется F-статистика для включения в модель каждой из $k - 2$ оставшихся переменных. Переменная x_j, $j \neq l$, имеющая наибольшее значение F_j, включается в модель. Определяется уравнение регрессии с двумя факторами x_l и x_j. Далее проверяется значимость полученного уравнения регрессии и вычисляются частные F-статистики: F_l и F_j. Наименьшая из этих величин сравнивается со значением F-удаления (***F*-to remove**). В зависимости от результата сравнения переменная x_l или x_j либо удаляется из уравнения, либо сохраняется. Такая проверка всех выбранных переменных проводится на каждом шаге. Поэтому может оказаться, что переменная, включенная в уравнение на предыдущем шаге, может быть удалена из анализа на следующих шагах.

3. Предположим, что на предыдущих шагах в регрессионную модель включено c из $k - 1$ факторов. На следующем шаге для каждого из оставшихся $(k - c - 1)$ факторов вычисляются значения F-включения: $F_i = (t_i)^2$, $i = 1, 2, \ldots, (k - c - 1)$. При условии, что верна гипотеза $H_0: \beta_i = 0$, статистика F_i имеет распределение Фишера $F(1, n - c - 2)$ соответственно с одной и $(n - c - 2)$ степенями свободы.

Если наибольшее значение F_i меньше заданного значения F-включения (***F*-to enter**), то процесс заканчивается.

Процесс пошаговой регрессии заканчивается, также если в регрессионную модель включены все переменные, либо если превышено заданное число шагов.

Пример 6.3 (продолжение). Рассмотрим работу процедуры **Forward Stepwise**. Используем те же данные. На панели **Model Definition** (рис. 6.16) введем **Metod: Forward stepwise**. Нажав внизу панели кнопку **Review Correlation matrix**... и затем, выбрав **Correlations**, рассмотрим корреляционную матрицу переменных x_1, x_2, x_3, x_4, Y (рис. 6.19):

Correlations (15.sta)

Continue..	X1	X2	X3	X4	Y
X1	1,000000	,228579	-,824134	-,245445	,730717
X2	,228579	1,000000	-,139242	-,972955	,816253
X3	-,824134	-,139242	1,000000	,029537	-,534671
X4	-,245445	-,972955	,029537	1,000000	-,821305
Y	,730717	,816253	-,534671	-,821305	1,000000

Рис. 6.19. Корреляционная матрица

Наибольший коэффициент корреляции с Y имеет переменная x_4 ($r_{yx_4} = -0{,}8213$), x_4 включается в регрессионную модель на первом шаге. Переменная x_4 — значима (см. **Regression Summary** на первом шаге: $t_4(11) \approx -4{,}775$, $F_4 = (-4{,}775)^2 = 22{,}79$, $p = 0{,}0005$), уравнение простой линейной регрессии имеет вид

$$Y = 117{,}568 - 0{,}738x_4.$$

Также значимо полученное уравнение регрессии: $F(1{,}11) = 22{,}799$, $p < 0{,}00058$.

На втором шаге в уравнение регрессии включается переменная x_1 (см. **Summary of Stepwise...** на втором шаге: $F_4 = 22{,}79$, $p = 0{,}001009$; $F_1 = 108{,}22$, $p = 0{,}000003$), причем, обе переменные значимы. Уравнение регрессии также значимо:

$$F = 176{,}627,\ p \approx 0.$$

На третьем шаге в уравнение регрессии включается переменная x_2.

В полученном уравнении переменные x_4 и x_2 незначимы:

1) для x_4: $t_4(9) = -1{,}365$, $F_4 = (-1{,}365)^2 = 1{,}863$, $p = 0{,}205$.

2) для x_2: $t_2(9) = 2{,}242$, $F_2 = (2{,}242)^2 = 5{,}026$, $p = 0{,}052$.

Следовательно, если в качестве ***F*-to remove** установить значение 3,9, то переменная x_4 будет удалена из уравнения.

Заметим, что квантили порядка $(1-\alpha)$ распределения Фишера для уровней значимости $\alpha = 0{,}10$ и $\alpha = 0{,}05$ равны:

$$F_{0,9}(1,9) = 3{,}36 \text{ и } F_{0,95}(1,9) = 5{,}12,$$

т. е. и при уровне значимости $\alpha = 0{,}10$, и при $\alpha = 0{,}05$ переменную x_4 следует удалить из уравнения регрессии. Окончательный результат совпадает с результатом процедуры **Backward stepwise**:

$$Y = 52{,}577 + 1{,}468x_1 + 0{,}662x_2.$$

Нажав кнопку **Stepwise (summary)** получим результаты выполнения процедуры по шагам (рис. 6.20).

Summary of Stepwise Regression; DV: Y (15 stn)

Continue..	Step +in/-out	Multiple R	Multiple R-square	R-square change	F - to entr/rem	p-level	Variable included
X4	1	,821305	,674542	,674542	22,7985	,000752	1
X1	2	,986139	,972471	,297929	108,2239	,000001	2
X2	3	,991128	,982335	,009864	5,0259	,048851	3
X4	-4	,989282	,978678	-,003657	1,8633	,202170	2

Рис. 6.20. Результаты выполнения процедуры пошаговой регрессии

6.4.1. Задания для самостоятельной работы

Задание 1

В приложении 1.1 (табл. П1 и П2) приведены варианты 10 заданий по регрессионному анализу. Для каждой задачи выполните следующие задания.

1. Используя пошаговую регрессию, определите минимальное число факторов достаточно точно предсказывающих зависимую переменную Y. Используйте обе процедуры **Backward** и **Forward Stepwise**. Подберите подходящие значения F-включения и F-удаления для каждой процедуры. Сравните и проанализируйте результаты обеих процедур.

2. Используя наиболее существенные факторы, найдите уравнение множественной регрессии. Выполните дисперсионный анализ. Проверьте значимость регрессионной модели. Найдите оценку дисперсии ошибок наблюдений, коэффициенты детерминации и множественной корреляции. Определите доверительные интервалы для параметров регрессии, проверьте гипотезу о значимости параметров и гипотезу $H_0: \beta_1 = \beta_2 = 0$, где β_1 и β_2 — коэффициенты регрессии для первого и второго из отобранных факторов.

3. Определите остатки. Постройте график остатков. Проверьте выполнение предположения регрессионного анализа:

- дисперсия остатков постоянна;
- остатки некоррелированны;
- остатки имеют нормальное распределение $N(0,\sigma^2)$.

Сделайте вывод об адекватности регрессионной модели результатам наблюдений.

4. Используя модель множественной регрессии, определите предсказанное значение зависимой переменной Y при следующих значениях выбранных p факторов

$$x_{0i} = \max x_i + 2S_i,$$

где S_i — оценка среднего квадратического отклонения переменной x_i, $i = 1, 2, ..., p$. Определите доверительные интервалы для среднего и индивидуального предсказанного значения. Для всех расчетов принять $\alpha = 0{,}05$.

Задание 2

В приложении 1.2 приведены данные о стоимости однокомнатных квартир. Используя процедуры пошаговой регрессии определите минимальное число факторов достаточно точно определяющих стоимость квартиры. Выясните изменяются ли переменные, определяемые пошаговой регрессией, если используются не все данные, а определенные подмножества данных. Рассмотрите следующие подмножества данных:

а) все данные;

б) данные о квартирах, не расположенных на первых и последних этажах;

в) данные о квартирах, расположенных в некирпичных домах и не на первых и не на последних этажах;

г) данные о квартирах, расположенных в Коньково.

6.5. Корреляционный анализ

В корреляционном анализе исследуется взаимозависимость между k случайными величинами $X_1, X_2, ..., X_k$. Предполагается, что выборка получена из генеральной совокупности, имеющей ***k-мерное нормальное распределение.***

Для случайных величин X_1 и X_2 основной характеристикой взаимозависимости является *парный коэффициент корреляции*

$$\rho_{12} = \frac{\operatorname{cov}(X_1, X_2)}{\sigma_1 \sigma_2},$$

где $\operatorname{cov}(X_1, X_2) = M[(X_1 - m_1)(X_2 - m_2)]$ — ковариация; m_1, m_2 — математические ожидания X_1 и X_2; σ_1 и σ_2 — их средние квадратические отклонения.

Коэффициент корреляции ρ_{12} определяет **степень линейной зависимости** между X_1 и X_2.

В случае k случайных величин $X_1, X_2, ..., X_k$ парные коэффициенты корреляции ρ_{ij}, $i, j = 1, 2, ..., k$, $i \neq j$, образуют симметрическую корреляционную $(k \times k)$ матрицу:

$$R = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{12} & 1 & \dots & \rho_{2k} \\ \dots & \dots & \dots & \dots \\ \rho_{1k} & \rho_{2k} & \dots & 1 \end{pmatrix}.$$

Однако связь между любой парой случайных величин может быть обусловлена влиянием других случайных величин из данной совокупности. Например, при рассмотрении случайных величин X_1, X_2, X_3 возможны такие взаимосвязи (рис. 6.21):

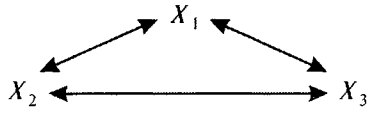


Рис. 6.21. Возможные взаимосвязи между X_1, X_2 и X_3

Чтобы определить взаимосвязи между двумя переменными, независимо от третьей, вычисляются ***частные коэффициенты корреляции***: $\rho_{12\cdot3}$, $\rho_{13\cdot2}$, $\rho_{23\cdot1}$, где $\rho_{12\cdot3}$ — частный коэффициент корреляции между X_1 и X_2, $\rho_{13\cdot2}$, $\rho_{23\cdot1}$ — определяются аналогично.

Оценку частного коэффициента корреляции, например $\tilde{\rho}_{12\cdot3} = r_{12\cdot3}$, можно получить, исключая линейные связи между X_1 и X_3, а также между X_2 и X_3.

Пусть $\begin{pmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{22} \\ \dots & \dots & \dots \\ x_{1n} & x_{2n} & x_{3n} \end{pmatrix}$ — выборка наблюдений случайных величин X_1, X_2, X_3 объема n.

Найдем линейную регрессию X_1 на X_3 и X_2 на X_3:

$$x_{1i} = a_{1\cdot3} + b_{13}x_{3i} + u_i,$$

$$x_{2i} = a_{2\cdot3} + b_{23}x_{3i} + v_i, \quad i = 1, 2, \dots, n.$$

где оценки параметров соответствующих регрессионных моделей: $a_{1\cdot3}$ и $a_{2\cdot3}$ — оценки свободных членов; b_{13} и b_{23} — оценки регрессионных коэффициентов, а u_i и v_i — остатки, т. е. величины, необъясненные регрессией X_1 на X_3 и X_2 на X_3.

Оценка частного коэффициента корреляции между X_1 и X_2, определяется как оценка парной корреляции между остатками u_i и v_i:

$$r_{12\cdot3} = \frac{\sum u_i v_i}{\sqrt{\sum u_i^2}\sqrt{\sum v_i^2}},$$

средние остатков u_i и v_i равны нулю, так как параметры регрессии оцениваются методом наименьших квадратов.

Преобразуя эту формулу, можно показать [33], что

$$r_{12\cdot3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}.$$

Квадрат частного коэффициента корреляции $r_{12\cdot3}^2$ называется ***частным коэффициентом детерминации***. Он определяет долю дисперсии X_1, которая объясняется регрессией X_1 на X_2 при фиксированном значении X_3.

Аналогично определяются остальные частные коэффициенты корреляции и детерминации.

В общем случае, пусть l и h — какие-либо две переменные из набора $X_1, X_2, ..., X_k$, c — подмножество из оставшихся $k-2$ переменных. Частный коэффициент корреляции $\rho_{lh \cdot c}$ определяет степень линейной зависимости между l и h, при условии, что влияние переменных из c исключено.

Частные коэффициенты первого порядка определяются по формуле

$$\rho_{lh \cdot d} = \frac{\rho_{lh} - \rho_{ld}\rho_{hd}}{\sqrt{1-\rho_{ld}^2}\sqrt{1-\rho_{hd}^2}}.$$

Последовательно применяя эту рекуррентную формулу, получим

$$\rho_{lh \cdot cd} = \frac{\rho_{lh \cdot c} - \rho_{ld \cdot c}\rho_{hd \cdot c}}{\sqrt{1-\rho_{ld \cdot c}^2}\sqrt{1-\rho_{hd \cdot c}^2}},$$

где c — любое подмножество оставшихся переменных.

Для множественной линейной регрессии представляет интерес следующий случай. Пусть переменная X_1 взята в качестве зависимой переменной: $X_1 = Y$, X_m — любая переменная из набора $X_2, X_3, ..., X_k$, а c — подмножество всех оставшихся $(k-2)$ переменных, тогда $r_{yx_m \cdot c}$ — мера линейной зависимости Y от X_m после вычитания эффекта, обусловленного зависимостью этих переменных с переменными из подмножества c. Проверка гипотезы о том, что при фиксированных значениях переменных из c вклад переменной X_m в предсказание Y незначим, т. е. гипотезы $H_0: \rho_{yx_m \cdot c} = 0$, эквивалентна проверке гипотезы $H_0: \beta_m = 0$. Проверка последней гипотезы выполняется с помощью статистики

$$t_m = \frac{\tilde{\beta}_m}{\sqrt{D[\tilde{\beta}_m]}}.$$

Если гипотеза $H_0: \beta_m = 0$ верна, то статистика t_m имеет распределение Стьюдента с $n-k$ степенями свободы.

В пакете STATISTICA результаты множественной регрессии содержат значения t_m-статистик. Частный коэффициент корреляции можно вычислить по формуле

$$r_{yx_m \cdot c} = \frac{t_m}{\sqrt{t_m^2 + n - k}}, \qquad m = 2, 3, ..., k.$$

Частный коэффициент корреляции $r_{yx_m \cdot c}$, где c — некоторое подмножество из $p < k-2$ переменных, определяет «качество» X_m для предсказания Y после исключения влияния на Y переменных из подмножества c. Сравнивая $r_{yx_m \cdot c}$ для всех X_m, не входящих в c, можно упорядочить переменные по их важности для предсказания Y относительно c.

Пример 6.4. Рассмотрим взаимосвязи следующих показателей эффективности производства по данным 25 однотипных машиностроительных предприятий [8]:

X — выработка валовой продукции на одного работающего (производительность труда), млн руб.;

Y — выпуск валовой продукции на один рубль среднегодовой стоимости производственных фондов (фондоотдача), руб.;

Z — материалоемкость — стоимость материалов в валовой продукции, %:

X	Y	Z
6,0	2,0	25
4,9	0,8	30
7,0	2,7	20
6,7	3,0	21
5,8	1,0	28
6,1	2,1	26
5,0	0,9	30
6,9	2,6	22
6,8	3,0	20
5,9	1,1	29
5,0	0,8	27
5,6	2,2	25
6,0	2,4	24
5,7	2,2	25
5,1	1,3	30
5,2	1,5	24
7,3	2,7	20
6,1	2,4	27
6,2	2,2	28
5,9	2,0	26
6,0	2,0	26
4,8	0,9	31
7,3	3,2	19
7,2	3,3	20
7,0	3,0	20

Решение в пакете STATISTICA. Анализ проводится в модуле **Multiple Regression**. В качестве зависимой переменной возьмем X, а Y и Z определим как факторы. Прежде всего рассмотрим корреляционную матрицу. В нижней части стартовой панели модуля (рис. 6.3) надо отметить опцию **Review descr. stats. Correlation matrix..., OK**. Корреляционная матрица и диаграммы рассеивания для всех пар переменных выводится при нажатии кнопок **Correlations** и **Graph** (рис. 6.22).

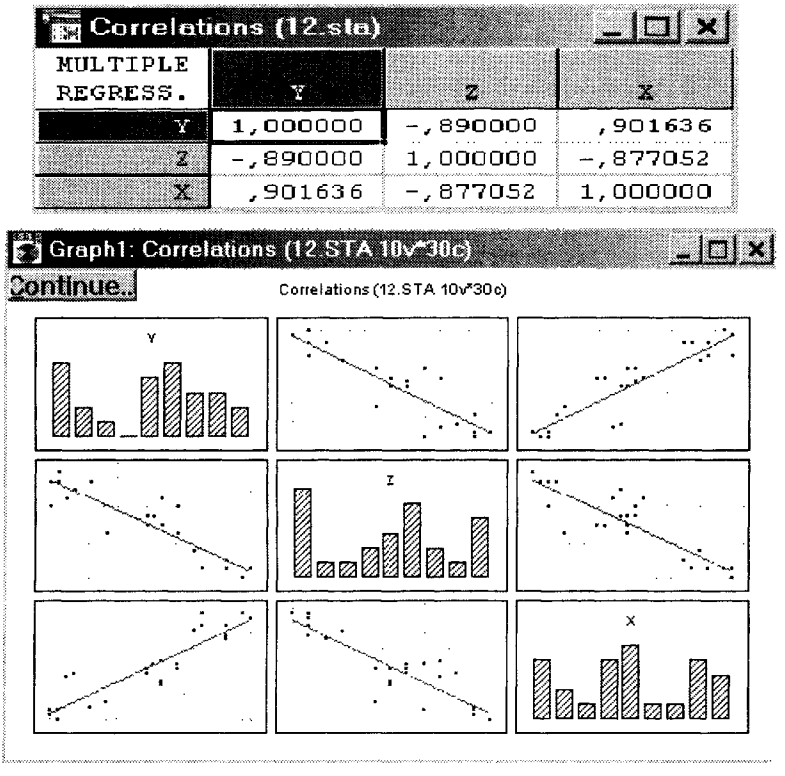


Рис. 6.22. Корреляционная матрица, гистограммы и диаграммы рассеяния для исходных данных

Коэффициенты корреляции и диаграммы рассеяния свидетельствуют о сильной линейной зависимости между всеми парами переменных.

Чтобы определить оценки частных корреляций $r_{xy \cdot z}$ и $r_{xz \cdot y}$ нужно выполнить регрессионный анализ, нажав **OK**, и в левом нижнем углу окна результатов регрессии (рис. 6.4) нажать кнопку **Partial correlations (частные корреляции)**.

В результате получим следующие частные корреляции, t-статистики и их уровни значимости:

$$r_{xy \cdot z} \approx 0{,}553; \quad t_y \approx 3{,}11; \quad p \approx 0{,}005;$$

$$r_{xz \cdot y} \approx -0{,}378; \quad t_z \approx -1{,}92; \quad p \approx 0{,}068.$$

В пакете STATISTICA результат выглядит следующим образом (рис. 6.23).

Variables currently in the Equation: DV: X (12.sta)

Continue..	Beta in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(22)	p-level
Y	,582299	,552681	,265505	,207900	,792100	3,11054	,005098
Z	-,358806	-,378273	-,163602	,207900	,792100	-1,91668	,068359

Рис. 6.23. Частные корреляции для примера 6.4

Так как $r_{xy}^2 > r_{xy\cdot z}^2$, то взаимозависимость между X и Y частично обусловлена влиянием Z на X и Y. Аналогично объясняется взаимозависимость между X и Z: она практически полностью определена влиянием Y на X и Z, так как частный коэффициент корреляции $r_{xz\cdot y}$ незначим ($p = 0{,}068$).

Чтобы вычислить оценку частного коэффициента корреляции $r_{yz\cdot x}$, нужно снова войти в стартовую панель (рис. 6.3) и выбрать в качестве зависимой переменной Y, а в качестве факторов взять X и Z, затем нажать **OK**. Нажав кнопку **Partial correlations** на панели результатов регрессии, получим

$$r_{yz\cdot x} \approx -0{,}477; \quad t_z = -2{,}549; \quad p \approx 0{,}018.$$

Таким образом взаимозависимость между Y и Z в значительной степени обусловлена влиянием X на Y и Z.

Для рассматриваемых переменных X, Y, Z не существует односторонней причинно-следственной связи, поэтому в качестве регрессионной модели следует взять модель, соответствующую наибольшему частному коэффициенту корреляции: $r_{xy\cdot z} = 0{,}553$. В этом случае мы получаем два уравнения регрессии: X на Y и Z; Y на X и Z. Анализируем оба уравнения и сравниваем коэффициенты детерминации: $R_{x\cdot yz}^2 \approx 0{,}840$; $R_{y\cdot xz}^2 \approx 0{,}856$.

Выбирая уравнение, определяющее наибольший коэффициент детерминации, окончательно получаем регрессионную модель Y на X и Z в виде

$$y = 1{,}1034 + 0{,}534x - 0{,}092z.$$

6.5.1. Задания для самостоятельной работы

Рассмотрите взаимосвязи факторов в следующих заданиях и определите подходящую регрессионную модель.

Задание 1

На предприятии существует 16 научно-производственных отделов, занятых выпуском различной продукции, работ, услуг. Параметры каждого отдела определяются четырьмя признаками:

x_1 — стоимость активной части основных производственных фондов, тыс. руб.;

x_2 — среднемесячный объем работ отдела, тыс. руб.;

x_3 — удельный вес работ/услуг отдела по внутрифирменной кооперации, %;

x_4 — среднемесячная прибыль отдела, тыс. руб.

Исходные данные по отделам приведены ниже.

№ отдела	Значения признаков			
	x_1	x_2	x_3	x_4
1	699	190	53	11
2	532	211	19	42
3	650	152	46	14
4	768	216	67	17
5	67	106	0	32
6	322	397	26	52
7	736	180	49	18
8	501	239	11	60
9	293	391	16	66
10	300	396	29	87
11	73	160	0	22
12	862	199	51	22
13	112	136	0	29
14	289	388	31	74
15	512	195	6	58
16	490	201	9	65

Задание 2

Ниже приведены значения основных факторов сельскохозяйственного производства для 20 районов:

x_1 — число тракторов на 100 га;

x_2 — число зерноуборочных комбайнов на 100 га;

x_3 — число орудий поверхностной обработки почвы на 100 га;

x_4 — количество удобрений, расходуемых на гектар (т/га);

x_5 — количество химических средств защиты растений, расходуемых на гектар (ц/га).

Номер наблюдения	x_1	x_2	x_3	x_4	x_5
1	1,59	0,26	2,05	0,32	0,14
2	0,34	0,28	0,46	0,59	0,66
3	2,53	0,31	2,46	0,30	0,31
4	4,63	0,40	6,44	0,43	0,59
5	2,16	0,26	2,16	0,39	0,16
6	2,16	0,30	2,69	0,32	0,17

Продолжение задания 2

Номер наблюдения	x_1	x_2	x_3	x_4	x_5
7	0,68	0,29	0,73	0,42	0,23
8	0,35	0,26	0,42	0,21	0,08
9	0,52	0,24	0,49	0,20	0,08
10	3,42	0,31	3,02	1,37	0,73
11	1,78	0,30	3,19	0,73	0,17
12	2,40	0,32	3,30	0,25	0,14
13	9,36	0,40	11,51	0,39	0,38
14	1,72	0,28	2,26	0,82	0,17
15	0,59	0,29	0,60	0,13	0,35
16	0,28	0,26	0,30	0,09	0,15
17	1,64	0,29	1,44	0,20	0,08
18	0,09	0,22	0,05	0,43	0,20
19	0,08	0,25	0,03	0,73	0,20
20	1,36	0,26	0,17	0,99	0,42

Задание 3

Проведите корреляционный анализ данных из Приложения 1.1 табл. П2. Варианты заданий приведены в табл. П1.

Задание 4

В Приложении 1.2 приведены данные о стоимости однокомнатных квартир. Вычислите корреляционную матрицу для переменных price, *distc*, *distm*, *totsq*, *kitsq*, *livsq*. Дайте экономическую интерпретацию парных коэффициентов корреляции. Найдите частные коэффициенты корреляции между переменными *price* и *totsq*, *price* и *kitsq* при условии, что влияние остальных переменных исключено. Объясните полученные результаты.

6.6. Нелинейная регрессия

Если модель нелинейна по параметрам, например, имеет вид

$$y = \Theta_1 + \Theta_2 \cos(x_1 + \Theta_3 x_2),$$

то для оценки параметров Θ_1, Θ_2, Θ_3 также можно использовать метод наименьших квадратов, однако, в этом случае приходится решать систему нелинейных уравнений. Для этого используются различные приближенные вычислительные методы.

Важным моментом является выбор начальных значений параметров, так как их неудачный выбор может привести к медленной сходимости и даже к расходимости процесса вычислений.

Чтобы ознакомиться с особенностями используемых вычислительных методов и методов выбора начальных значений оцениваемых параметров нужно обратиться к специальной литературе [18, 24, 31].

В пакете STATISTICA оценки параметров нелинейной регрессии вычисляются в модуле **Nonlinear Estimation** (нелинейное оценивание).

Результаты вычислений содержат: число итераций, значение функции потерь по шагам итераций (***Loss***), оценки параметров, долю дисперсии исходных данных, объясняемую моделью (коэффициент детерминации R^2) и значение $R = \sqrt{R^2}$ (коэффициент множественной корреляции).

Кроме этих данных, можно получить среднеквадратические ошибки оценок параметров $\widetilde{\Theta}_i$ (***Std. Err.***), t-статистики и уровни их значимости (***p-level***). Эти результаты позволяют проверить гипотезу $H_0: \Theta_i = 0$ и найти доверительные интервалы для Θ_i.

Чтобы вывести эти результаты нужно на панели выбора метода оценивания параметров (**Model Estimation**) (рис. 6.27) включить опцию **Asymptotic Standard errors** (асимптотические стандартные ошибки) и после окончания процесса вычисления оценок параметров на панели результатов нажать кнопку **Parameters and Standard errors** (параметры и стандартные ошибки).

Панель результатов содержит также различные опции для анализа остатков, позволяющие проверить гипотезу об адекватности модели результатам наблюдений.

Пример 6.5. По следующим данным найдите наилучшие оценки параметров в уравнении

$$y = a + be^{-cx}.$$

Определить 95%-е доверительные интервалы для параметров.

Данные:

y	–4,8	–3	–4,2	20	53,4	51,6
x	95,9	48,2	19,5	5,4	1,4	0,4

Решение в пакете STATISTICA. Введем данные в файл и присвоим им соответствующие имена Y и X. Предварительно рассмотрим график данных: **Graphs** → **Stats 2D Graphs** → **Statterplots**. На панели настройки (**2D Statterplots**) установим **Graph Type**: **Regular**; **FIT**: ***off***; **Options** → **Case Labels**: **Case Name**, **OK**, **OK**. График данных представлен на рис. 6.24.

Переключимся в модуль **Nonlinear Estimation** (нелинейное оценивание) и в открывшемся меню выбираем **User-specified regression** (определяемая пользователем регрессия) (рис. 6.25).

В окно **Function to be estimated...** (оцениваемая функция) вводим функцию $y = a + b * \exp(-c * x)$ (рис. 6.26).

Функция потерь (***Loss function***) по умолчанию определена как сумма квадратов разностей наблюдаемых и предсказанных значений (***(OBS-PRED)**2***), следовательно, оценки параметров вычисляются методом наименьших квадратов, **OK**.

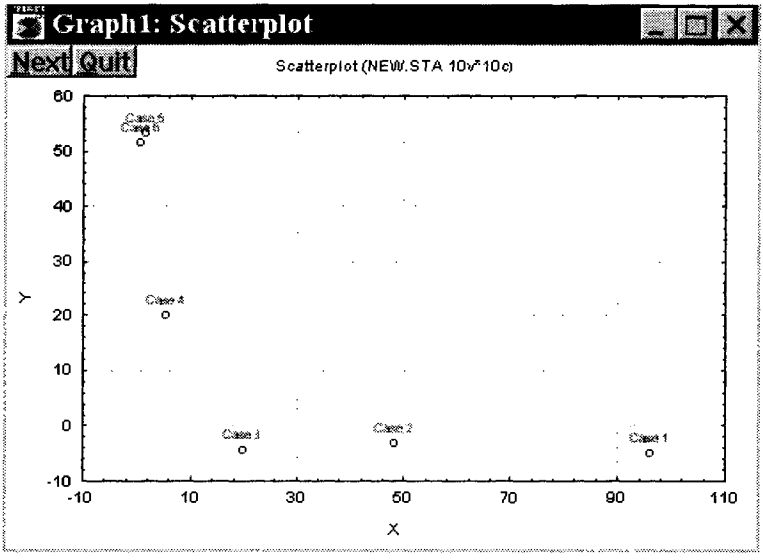


Рис. 6.24. Исходные данные для примера 6.5

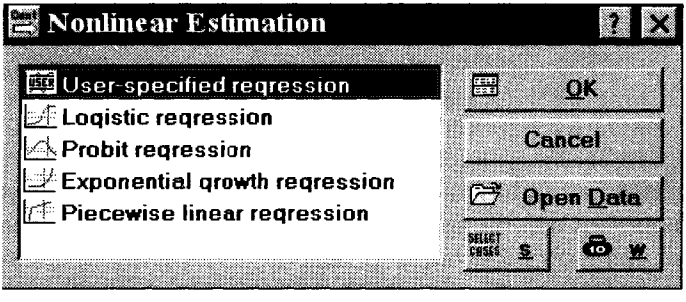


Рис. 6.25. Стартовая панель модуля Nonlinear Estimation

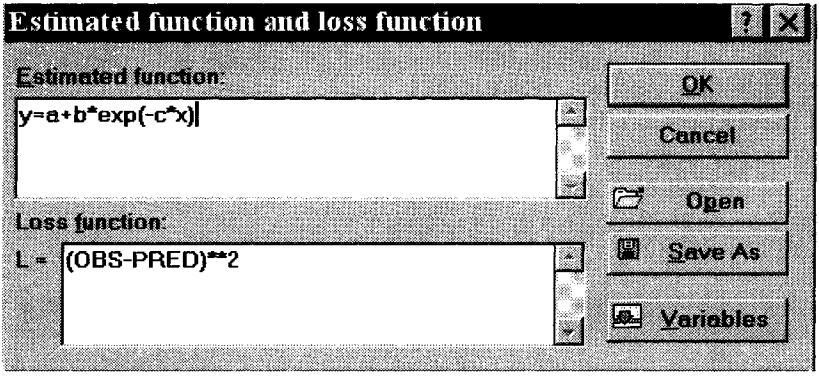


Рис. 6.26. Ввод функции

В открывшейся панели **Model Estimation** (метод оценивания) (рис. 6.27) выбирается вычислительный метод, например, **Quasi-Newton**, можно задать начальные значения параметров, число итераций, шаг изменения парамет-

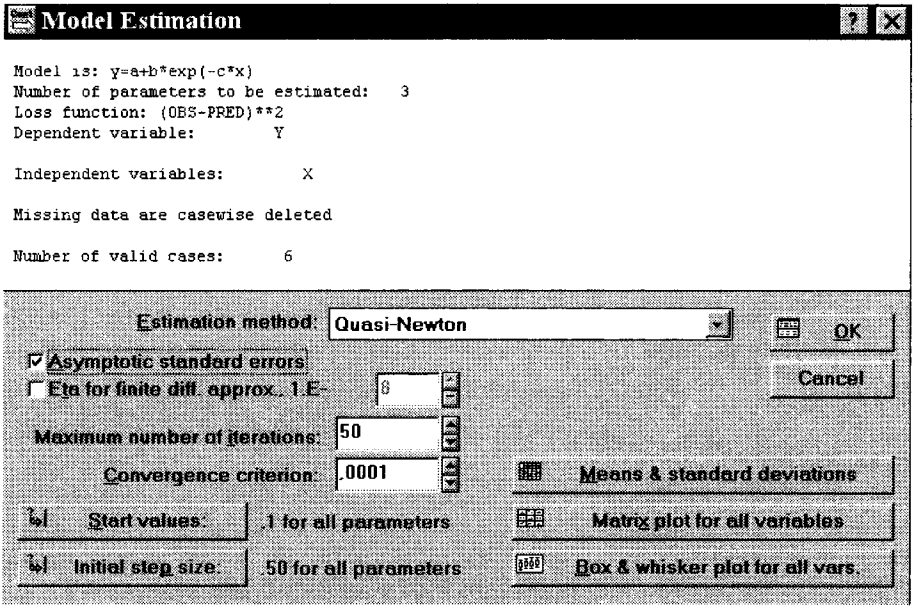


Рис. 6.27. Панель для выбора метода оценивания

ров. Чтобы получить среднеквадратические ошибки оценок параметров нужно включить опцию **Asymptotic Standard errors**.

Если теперь запустить программу оценивания параметров, нажав **OK**, то может появиться следующее сообщение: **Error in function: change start values/precision/step size** (вычисления не могут выполняться: следует изменить начальные значения параметров/точность вычислений/величину шага).

В данном примере не были определены начальные значения параметров (по умолчанию для всех параметров они задаются равными 0,1). Их можно определить достаточно просто. Используя таблицу исходных данных имеем: при $x = 0{,}4$; $y(0{,}4) = 51{,}6$, т. е. $a + b \approx 51{,}6$. Далее, для используемой функции при $x \to \infty$, $y \to a$. Так как при $x = 95{,}9$ $y = -4{,}8$ то начальное значение $a \approx -4{,}8$, начальное значение $b \approx 51{,}6 - (-4{,}8) \approx 55$. Начальное значение параметра c примем равным 0,1.

Для вычисления начальных значений можно также составить систему трех (по числу параметров) уравнений, подставляя в функцию исходные данные, и попытаться, хотя бы приближенно, найти решение. После ввода начальных значений, запустив процедуру снова, получим следующие результаты (рис. 6.28).

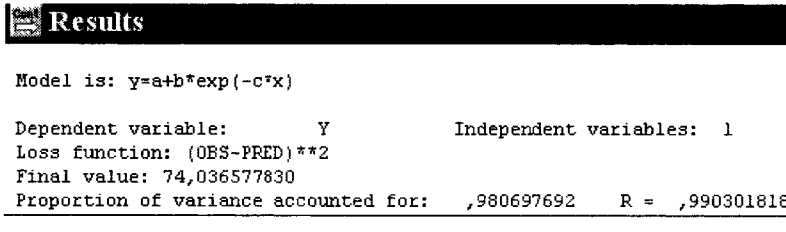


Рис. 6.28. Результаты процедуры оценивания

Окончательное значение функции потерь (***Final value***) ≈ 74,04.

Коэффициент детерминации $R^2 \approx 0{,}981$.

Нажав кнопку **Fitted 2D function...** на панели результатов, получим график функции и исходных данных (рис. 6.29).

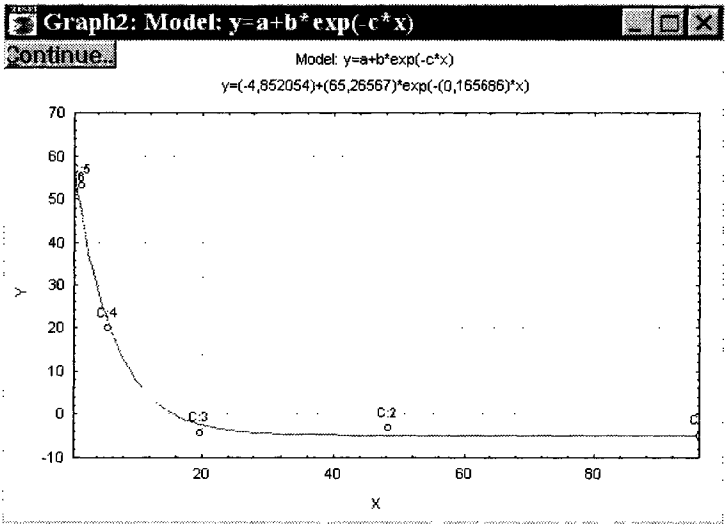


Рис. 6.29. График функции

Оценки параметров, их среднеквадратические значения, *t*-статистики для проверки гипотезы $H_0: \Theta_i = 0$ и соответствующие уровни значимости выводятся при нажатии кнопки **Parameters and Standard errors** (рис. 6.30).

Model: y=a+b*exp(-c*x) (new.sta)

Continue..

Dep. var: Y Loss: (OBS-PRED)**2
Final loss: 74,036577830 R=,99030 Variance explained: 98,070%

N=6	A	B	C
Estimate	-4,85205	65,26568	,165686
Std.Err.	3,09913	5,51709	,039442
t(3)	-1,56562	11,82973	4,200753
p-level	,21541	,00130	,024620

Рис. 6.30. Результаты оценивания параметров

Доверительные интервалы для параметров при доверительной вероятности $1 - \alpha$ определяются по формуле

$$\widetilde{\Theta}_i \pm t_{1-\frac{\alpha}{2}}(n-k) \cdot \sqrt{D[\widetilde{\Theta}_i]},\ i = 1, 2, \ldots, k,$$

где n — объем выборки, k — число оцениваемых параметров, $t_{1-\frac{\alpha}{2}}(n-k)$ — квантиль распределения Стьюдента порядка $1 - \dfrac{\alpha}{2}$ с $(n-k)$ степенями свободы.

Для данного примера при доверительной вероятности $1-\alpha=0{,}95$, $t_{0{,}975}(3)=3{,}182$, следовательно, доверительные интервалы будут такие:

a: $-4{,}852 \pm 3{,}182 \cdot 3{,}099 = -4{,}852 \pm 9{,}861$;

b: $65{,}265 \pm 3{,}182 \cdot 5{,}518 = 65{,}265 \pm 17{,}558$;

c: $0{,}165 \pm 3{,}182 \cdot 0{,}039 = 0{,}165 \pm 0{,}124$.

Гипотеза H_0: $a=0$ принимается, так как $p=0{,}215$; гипотезы H_0: $b=0$ и H_0: $c=0$ отклоняются, так как соответственно $p=0{,}001$ и $p=0{,}025$.

График функции и значение коэффициента детерминации $R^2=0{,}98$ показывают, что модель, по-видимому, адекватна результатам наблюдений.

6.6.1. Задания для самостоятельной работы

1. Найдите оценки параметров a и b функции Кобба—Дугласа:

$$Q = Q_0\left(\frac{L}{L_0}\right)^a\left(\frac{K}{K_0}\right)^b,$$

где Q — объем производства; L — трудовые ресурсы; K — капитал; а Q_0, L_0 и K_0 — фиксированные значения этих переменных: $Q_0=51$, $L_0=19$, $K_0=16$.

Значения Q при определенных трудовых ресурсах L и стоимости капитала K заданы в таблице

L \ K	6	11	16	21	26
9	6	12	19	27	36
14	11	12	35	50	66
19	17	32	51	72	96
24	22	42	67	95	126
29	27	52	83	118	156

2. Прибыль P, получаемая фирмой, определяется формулой $P=kx-c$, где x — объем производства (т), k — цена одной тонны продукции (руб./т), а c — издержки производства (руб.).

Предположим, что $k=40$ руб./т, а переменные x и c имеют следующее значение:

x	0	1	2	3	4	5	6	7	8	9	10	11
c	50	100	128	148	162	180	200	222	260	305	360	425

Найдите оценки параметров модели $P(x)$:

$$P(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

Используя полученную модель найдите объем производства обеспечивающий максимальную прибыль и точку самоокупаемости, т. е. такое значение x при котором прибыль равна нулю.

3. В табл. 6.1 приведены результаты наблюдений зависимой переменной z_i, $i = 1, 2, ..., 7$ и независимых переменных x и y.

Найдите оценки параметров для следующих моделей

$$z_i = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2, \quad i = 1, 2, ..., 7.$$

Таблица 6.1

№	x	y	z_1	z_2	z_3	z_4	z_5	z_6	z_7
1	3,000	–2,000	11,800	10,800	6,300	11,800	11,800	11,800	11,800
2	3,000	1,000	43,400	60,400	40,900	43,400	43,400	43,400	43,400
3	3,000	–3,000	1,800	–5,200	–4,700	1,800	1,800	1,800	1,800
4	2,000	3,000	68,800	90,800	67,800	68,800	68,800	68,800	68,800
5	2,000	–1,000	23,400	29,400	18,400	23,400	23,400	23,400	23,400
6	2,000	0,000	39,600	49,600	35,600	39,600	39,600	39,600	39,600
7	1,000	2,000	33,600	46,600	31,100	33,600	33,600	33,600	33,600
8	1,000	1,000	28,200	39,200	24,700	28,200	28,200	28,200	28,200
9	1,000	–3,000	18,600	21,600	11,100	18,600	18,600	18,600	18,600
10	0,000	2,000	28,200	36,200	25,200	28,200	28,200	28,200	28,200
11	0,000	1,000	7,800	15,800	3,800	7,800	7,800	7,800	7,800
12	0,000	–3,000	21,200	29,200	13,200	21,200	21,200	21,200	21,200
13	–1,000	2,000	19,600	22,600	16,100	19,600	19,600	19,600	19,600
14	–1,000	–3,000	24,600	37,600	16,100	24,600	24,600	24,600	24,600
15	–1,000	2,000	40,600	43,600	37,100	40,600	40,600	40,600	40,600
16	–2,000	2,000	22,800	20,800	18,800	22,800	22,800	22,800	22,800
17	–2,000	1,000	25,400	27,400	20,400	25,400	25,400	25,400	25,400
18	–2,000	0,000	22,600	28,600	16,600	22,600	22,600	22,600	22,600
19	–3,000	1,000	16,400	15,400	10,900	16,400	16,400	16,400	16,400
20	–3,000	3,000	2,800	–10,200	–0,700	2,800	2,800	2,800	2,800
21	–3,000	–2,000	52,800	69,800	44,300	52,800	52,800	52,800	52,800

4. Приведенные ниже данные представляют значения трех зависимых переменных y_1, y_2, y_3 и независимой переменной x. Найдите оценки параметров a и b для следующих моделей

$$y_i = a(1 + e^{bx}), \quad i = 1, 2, 3.$$

Проверьте, хорошо ли подходит к этим данным полиномиальная модель?

x	y_1	y_2	y_3
1,0	9,0	4,4	6,7
2,0	9,0	8,1	12,5
3,0	16,0	9,4	14,3
4,0	20,0	10,2	15,8
5,0	21,0	12,3	17,1
6,0	22,0	11,5	18,0
7,0	23,0	13,2	20,0
8,0	30,0	15,2	36,3
9,0	41,0	18,5	45,1
10,0	50,0	31,0	60,0

5. Найдите оценки параметров a, b и c для модели

$$y = a(10)^{\frac{bx}{c+x}}$$

по следующим данным:

y	4,1	8,5	16,5	32,2	65,1	99	151,2	225	340,9	424	523	675	780
x	0	10	20	30	40	50	60	70	85	90	95	100	105

Глава 7
АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

7.1. Основные характеристики и компоненты временного ряда

Временным рядом называется последовательность наблюдений, упорядоченная по времени: y_1, y_2, ..., y_n , где y_t — числа, представляющие наблюдения некоторой переменной в n равностоящих моментов времени $t = 1, 2, ..., n$. Примерами данных, которые необходимо изучать во времени являются: цены на товар, деловая активность, национальный валовой продукт. Особенностью, выделяющей анализ временных рядов, является зависимость данных, причем характер этой зависимости может определяться положением наблюдений в последовательности.

Основные задачи анализа:

1) прогнозирование, на основе знания прошлого;

2) сжатое описание характерных особенностей ряда;

3) управление процессом, порождающим ряд.

В теории временных рядов разработаны различные методы исследования и анализа: корреляционный и спектральный анализ, методы сглаживания и фильтрации, модели авторегрессии и скользящего среднего [25, 26, 34, 35, 36].

В анализе временных рядов, как и в большинстве статистических методов, предполагается, что исходные данные содержат детерминированную и случайную составляющие. В общем случае детерминированная составляющая может быть представлена в виде комбинации следующих компонент:

а) *тренда* определяющего главную тенденцию временного ряда;

б) более или менее регулярных колебаний относительного тренда — *циклов*;

в) периодических колебаний; такие колебания называются *сезонной составляющей.*

Временной ряд может быть представлен различными математическими моделями.

Пусть u_t — тренд, W_t, S_t, ε_t — соответственно циклическая, сезонная и случайная остаточная составляющие.

Аддитивная модель записывается в виде

$$y_t = u_t + W_t + S_t + \varepsilon_t.$$

Мультипликативная модель имеет вид

$$y_t = u_t W_t S_t \varepsilon_t$$

и при переходе к логарифмам сводится к аддитивной модели.

Если предположить, что сезонная составляющая S_t пропорциональна сумме тренда и циклической составляющей $(u_t + W_t)$, $S_t = (u_t + W_t)C_t$, то временной ряд будет представлен в виде *смешанной модели*

$$y_t = (u_t + W_t)(1 + C_t) + \varepsilon_t.$$

Выбор модели зависит от конкретной совокупности явлений, определяющих данный временной ряд и их взаимосвязей.

Представление временного ряда в виде той или иной композиции его компонент естественно приводит к идее последовательного выделения этих компонент и прогнозирования на основе полученной модели.

Пример 7.1. Поясним выбор модели на примере временного ряда, представляющего объем ежемесячных розничных продаж в США в период с 1953 по 1964 гг. (рис. 7.1).

Из графика видно, что данные имеют линейный тренд и сезонную составляющую с периодом 12 месяцев. Наибольшее число продаж ежегодно приходится на декабрь, что связано с рождественскими праздниками. Наименьшее число продаж ежегодно приходится на февраль. При этом амплитуда сезонных колебаний относительно тренда возрастает год от года. Для таких данных подходит мультипликативная модель

$$y_t = u_t W_t S_t \varepsilon_t.$$

При постоянной амплитуде сезонных колебаний возможно, что более подходящей моделью была бы аддитивная модель

$$y_t = u_t + W_t + S_t + \varepsilon_t.$$

Рассмотрим результаты выделения отдельных компонент для мультипликативной модели временного ряда в процедуре **Seasonal decomposition** (сезонная декомпозиция) модуля **Times series** (Временные ряды) пакета STATISTICA. Эта процедура и выполнение ее в пакете STATISTICA будут подробно рассмотрены в пунктах 7.3 и 7.6.

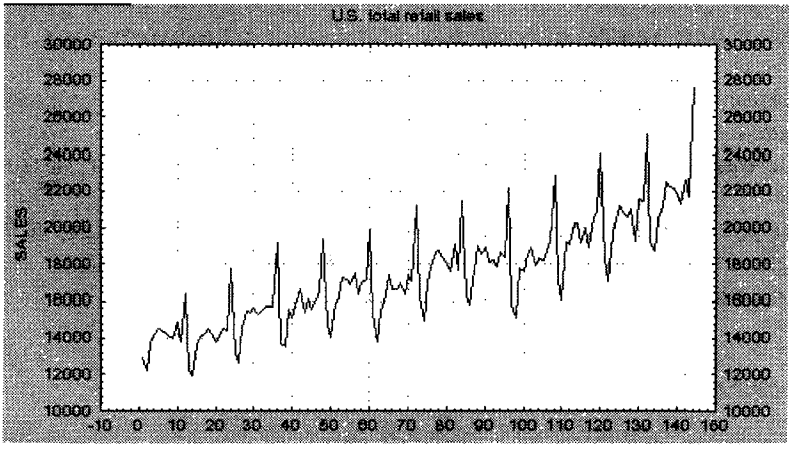


Рис. 7.1. Ежемесячный объем продаж в США с 1953 по 1964 гг.

Тренд и циклическая компонента, выделенные методом простого скользящего среднего по 12 точкам, показаны на рис. 7.2.

Сезонная компонента (в процентах) показана на рис. 7.3.

На рис. 7.4 приведены данные, из которых исключена сезонная компонента, а на рис. 7.5 показана остаточная случайная составляющая, т. е. данные, из которых исключен тренд и сезонная составляющая (остаточная составляющая показана в очень большом масштабе).

В данном примере временной ряд имеет простую структуру.

На рис. 7.6 представлены ежедневные цены (в долларах за баррель) для фьючерсов на нефть марки «Brend» с 15.07 по 22.09.1988 г.

В этом примере данные имеют квадратичный тренд, циклическую и случайную остаточную составляющие.

Параметры тренда: $u_t = \beta_0 + \beta_1 t + \beta_2 t^2$, $t = 1, 2, \ldots, 55$ можно определить методами регрессионного анализа [см. главу 6].

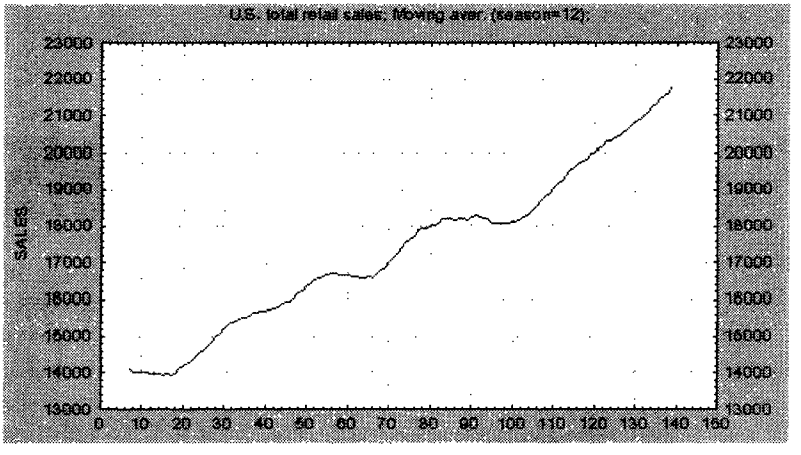


Рис. 7.2. Результаты простого скользящего среднего по 12 точкам

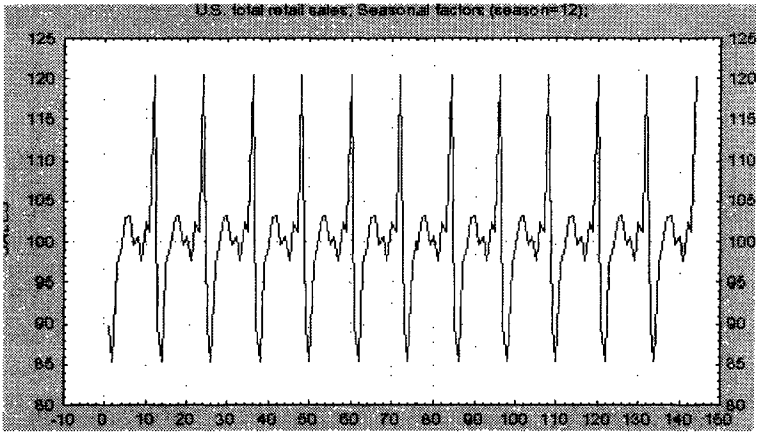


Рис. 7.3. Сезонная компонента

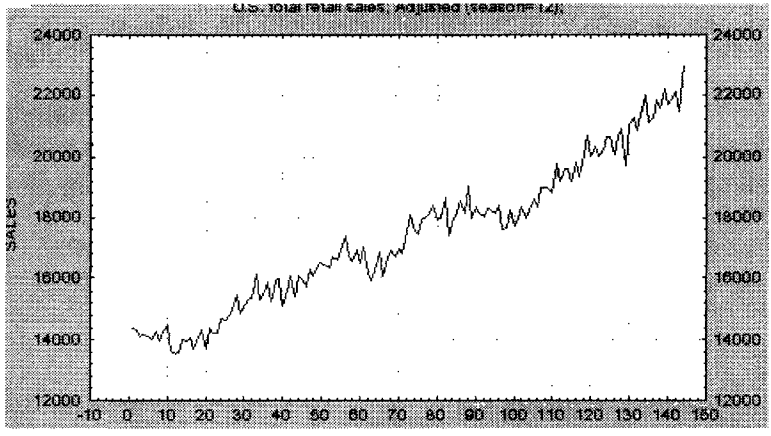


Рис. 7.4. Данные скорректированные на сезонную составляющую

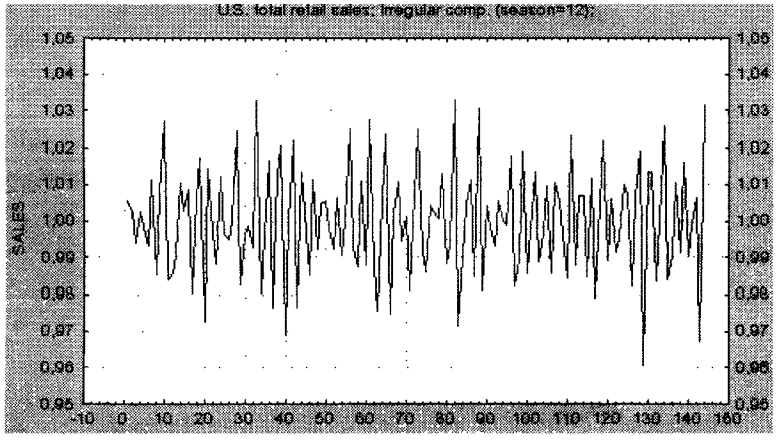


Рис. 7.5. Случайная остаточная составляющая

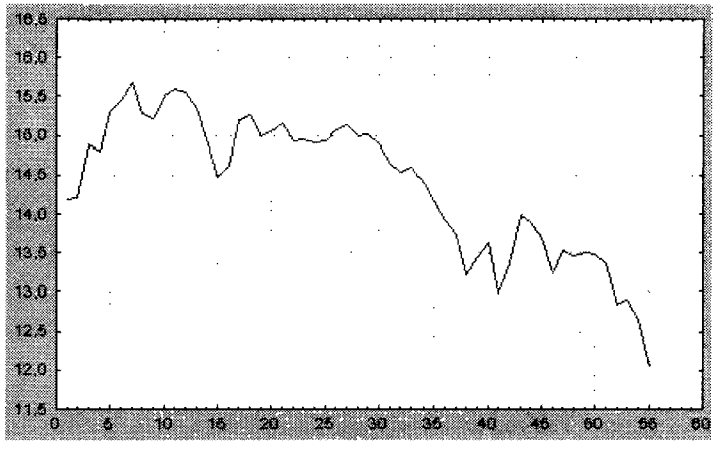


Рис. 7.6. Ежедневные цены на нефть

7.1.1. Числовые характеристики временного ряда и их оценка по результатам наблюдений

С математической точки зрения временной ряд является реализацией *случайного процесса с дискретным целочисленным параметром t.*

Случайным процессом с дискретным целочисленным параметром t называется семейство случайных величин $\{Y(t)\}$, где параметр t принимает значения из множества целых чисел.

Случайный процесс является обобщением понятия случайного вектора на случай бесконечного числа компонент. Случайный процесс называют также случайной функцией и обозначают как $Y(t)$.

Математическим ожиданием случайного процесса $Y(t)$ является неслучайная функция $m(t) = M[Y(t)]$; при каждом значении t $m(t)$ равно математическому ожиданию соответствующей случайной величины. Аналогично определяется дисперсия случайного процесса $D(t) = D[Y(t)]$. Важнейшей характеристикой временного ряда является автокорреляционная функция:

$$\rho_k = \frac{M[Y(t) - m(t))(Y(t + k) - m(t + k))]}{\sigma(t)\sigma(t + k)},$$

где $\sigma(t) = \sqrt{D[Y(t)]}$, $k = 0, 1, ..., n - 2$; $t = 1, 2, ..., n$.

Величина k, определяющая временной интервал между случайными величинами $Y(t)$ и $Y(t + k)$, для которых вычисляется коэффициент корреляции ρ_k, называется *лагом.*

Совокупность значений ρ_k, а также их график как функции от k называют *коррелограммой.*

Обширный класс случайных процессов представляют стационарные процессы.

Временной ряд y_t, где $t = 1, 2, ..., n$, называется *стационарным*, если в порождающем его случайном процессе $Y(t)$ последовательные группы случайных величин Y_{t+1}, Y_{t+2}, ..., Y_{t+k} имеют одну и ту же многомерную функцию распределения при любых значениях t и k. Таким образом механизм, генерирующий временной ряд, хотя и имеет вероятностный характер, остается неизменным во времени и, следовательно, не зависит от начала отсчета. Для стационарных рядов математическое ожидание и дисперсия — константы:

$$M[Y(t)] = m;$$

$$D[Y(t)] = \sigma^2,$$

а автокорреляционная функция ρ_k не зависит от t, причем

$$\rho_k = \rho_{-k} \quad \text{и} \quad \rho_0 = 1.$$

Остаточная случайная компонента временного ряда обычно представляет стационарный случайный процесс. График такого процесса приведен на рис. 7.5.

Если для случайного процесса $Y(t)$ математическое ожидание постоянно, а значения автокорреляционной функции ρ_k зависит только от k (и не зависят от t), то случайный процесс называется *стационарным в широком смысле.*

Для оценки числовых характеристик случайного процесса $Y(t)$ (математического ожидания, дисперсии, автокорреляционной функции) необходи-

мо иметь множество независимых реализаций случайного процесса. Однако при анализе временных рядов в экономике, социологии и в других практически важных областях, мы имеем дело лишь с одной реализацией временного ряда.

При определенных условиях оценки числовых характеристик временного ряда можно вычислить по одной реализации.

Пусть $y_1, y_2, ..., y_n$ — значения наблюдаемого временного ряда. Вычислим «среднее по реализации»

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

Можно показать [25, 35, 36], что среднее по реализации $\bar{Y}$ временного ряда является несмещенной и состоятельной оценкой математического ожидания m если выполнены следующие условия:

1) временной ряд является стационарным в широком смысле;

2) $\lim\limits_{n\to\infty} \frac{1}{n}\sum\limits_{k=1}^{n} \rho_k = 0.$

Таким образом поведение автокорреляционной функции определяет возможность оценки числовых характеристик стационарного временного ряда по одной реализации.

Оценка автокорреляционной функции по значениям временного ряда $y_1, y_2, ..., y_n$ вычисляется следующим образом. Рассмотрим множество пар: $(y_1, y_2), (y_2, y_3), ..., (y_{n-1}, y_n)$. Это реализации двумерной случайной величины $(Y(t), Y(t+1))$ по которым можно вычислить оценку коэффициента корреляции ρ_1 (см. главу 6, п. 6.1.1). Аналогично определяется оценка коэффициента корреляции ρ_2 по $(n-2)$ парам $(y_1, y_3), (y_2, y_4), ..., (y_{n-2}, y_n)$. Оценки коэффициента корреляции по значениям временного ряда называют ***сериальными корреляциями*** и обозначают r_k, $k = 1, 2, ...$. Сериальные корреляции вычисляют по формуле

$$r_k = \frac{\frac{1}{n-k}\sum\limits_{t=1}^{n-k}(y_t - \frac{1}{n-k}\sum\limits_{t=1}^{n-k} y_t)(y_{t+k} - \frac{1}{n-k}\sum\limits_{t=1}^{n-k} y_{t+k})}{\left\{\frac{1}{n-k}\sum\left[y_t - \frac{1}{n-k}\sum\limits_{t=1}^{n-k} y_t\right]^2 \cdot \frac{1}{n-k}\sum\limits_{t=1}^{n-k}\left[y_{t+k} - \frac{1}{n-k}\sum\limits_{t=1}^{n-k} y_{t+k}\right]^2\right\}^{\frac{1}{2}}} . \quad (1)$$

Если в (1) заменить оценки среднего значения ряда по $n-k$ наблюдениям, оценкой среднего всего ряда $y_1, ..., y_n$, а в знаменатель подставить оценку дисперсии всего ряда, то формула значительно упрощается

$$r_k \approx \frac{\frac{1}{n-k}\sum\limits_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\frac{1}{n}\sum(y_t - \bar{y})^2},$$

где

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t . \quad (2)$$

График значений r_k как функции от лага k называется ***выборочной коррелограммой***. Коррелограмма показывает насколько сильна линейная зависимость между членами ряда разделенными $(k-1)$ наблюдениями ($k = 1, 2, 3, \ldots$).

Выборочная коррелограмма позволяет определить наличие сезонной компоненты во временном ряде. На рис. 7.7 представлены коррелограмма временного ряда для примера 7.1. Предварительно из исходного ряда был удален тренд. Так как для анализа ряда была использована мультипликативная модель, то для удаления тренда нужно значения исходного ряда разделить на значения скользящих средних (рис. 7.2), представляющих тренд.

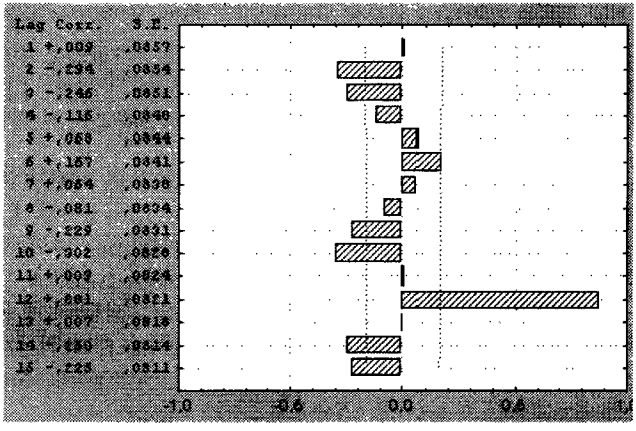


Рис. 7.7. Коррелограмма ряда без тренда для примера 7.1

Значение $r_{12} \approx 0{,}9$ значительно превышает все остальные сериальные корреляции, что указывает на наличие в ряде сезонной составляющей с периодом равным 12 месяцев.

На коррелограмме указаны 95%-е доверительные интервалы для сериальных корреляций (они определяются двумя вертикальными штриховыми линиями, симметричными относительно нулевой линии).

Границы доверительных интервалов вычисляются по формуле: $-\frac{1}{n} \pm \frac{2}{\sqrt{n}}$, где n число элементов временного ряда (объем выборки) [10]. В примере 7.1 $n = 144$ и границы 95%-го доверительного интервала будут (–0,174; 0,160). Таким образом сериальные корреляции r_2, r_3, r_9, r_{10}, r_{12}, r_{14}, r_{15} значимо не равны нулю.

7.2. Определение тренда и сглаживание временного ряда

Пусть временной ряд представим в виде суммы тренда и остаточной составляющей, т. е.

$$y_t = u_t + \varepsilon_t.$$

В ряде случаев тренд является известной функцией времени. Если эта функция зависит линейно от параметров, то для определения тренда ис-

пользуются методы регрессионного анализа (см. главу 6). Если тренд нельзя представить простой функцией времени на всем рассматриваемом интервале, то применяют методы сглаживания на основе скользящего среднего.

Сглаживание временного ряда означает представление тренда в данной точке посредством среднего значения ряда, вычисленного в окрестности данной точки. Используемое для сглаживания количество точек в окрестности называют *базой*. Если значения ряда из окрестности данной точки входят с одним и тем же весом, то операция сглаживания называется *простым скользящим средним*.

Пример 7.2. Сглаживание временного ряда на основе простого скользящего среднего по трем точкам (база равна 3). Рассмотрим следующие данные:

$$y_t = 6{,}00;\ 8{,}82;\ 8{,}94;\ 8{,}05;\ 9{,}75;\ 11{,}51;\ 13{,}69;\ 12{,}04;\ 14{,}76;\ 16{,}18;\ 17{,}11;\ 14{,}99;\ 15{,}01;\ 16{,}00;\ 15{,}26;\ 11{,}75,$$

где $t = 1, 2, \ldots, 16$ — номер наблюдений, проведенных через равные интервалы времени.

Решение. Сглаженные значения вычисляют последовательно: как средние арифметические первых трех значений: $(6 + 8{,}82 + 8{,}94)/3 = 7{,}92$, следующей тройки значений: $(8{,}82 + 8{,}94 + 8{,}05)/3$ и т. д. Таким образом при вычислениях каждое значение исходного ряда входит с весом 1/3. Результаты сглаживания приведены в табл. 7.1 и на рис. 7.8.

Таблица 7.1

t	y_t	Сглаженный ряд
1	6,00	—
2	8,82	7,92
3	8,94	8,60
4	8,05	8,91
5	9,75	9,77
6	11,51	11,65
7	13,69	12,41
8	12,04	13,50
9	14,76	14,33
10	16,18	16,02
11	17,11	16,09
12	14,99	15,70
13	15,01	15,33
14	16,00	15,42
15	15,26	14,34
16	11,75	—

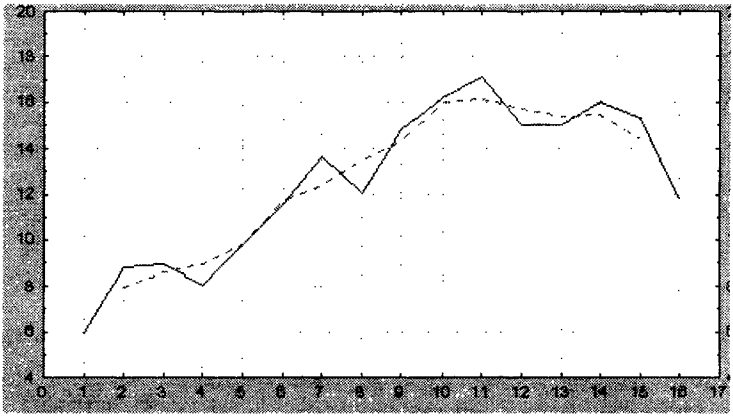


Рис. 7.8. — — исходный ряд; --- — сглаженный ряд, полученный при помощи простого скользящего среднего по трем точкам

Пример 7.3. Сглаживание ряда на основе простого скользящего среднего по четырем точкам; центрированное скользящее среднее.

К данным из примера 7.2 применим процедуру простого скользящего среднего по четырем точкам.

Решение. Вычисляем последовательно среднее арифметическое для первых четырех значений: (6 + 8,82 + 8,94 + 8,05)/4 = 7,95, затем для следующей четверки значений: (8,82 + 8,94 + 8,05 + 9,75)/4 = 8,89, (8,94 + 8,05 + + 9,75 + 11,51)/4 = 9,56 и так далее. Вычисленные значения представляют тренд соответственно между вторым и третьим наблюдениями, между третьим и четвертым наблюдениями, между четвертым и первым наблюдениями и так далее. Такое неудобство возникает всегда, если для сглаживания используется база из четного числа точек. В этом случае полученные значения центрируют, применяя к ним процедуру простого скользящего среднего по двум точкам. Последовательно вычисляем: (7,95 + 8,89)/2 = 8,42, (8,89 + 9,56)/2 = 9,23 и т. д. Полученные значения представляют значения тренда для третьего наблюдения, для четвертого наблюдения и так далее.

Результаты сглаживания приведены в табл. 7.2 и на рис. 7.9. Сравнивая сглаженные ряды для примеров 2 и 3, можно заметить, что тренд, полученный центрированным скользящим средним по четырем точкам лучше отражает главную тенденцию ряда.

Таблица 7.2

t	y_t	Сглаженный ряд, простое скользящее среднее по 4 точкам	Сглаженный ряд, центрированное скользящее среднее
1	6,00	—	—
2	8,82	7,9525	—
3	8,94	8,8900	8,42
4	8,05	9,5625	9,23
5	9,75	10,7500	10,16

Продолжение табл. 7.2

t	y_t	Сглаженный ряд, простое скользящее среднее по 4 точкам	Сглаженный ряд, центрированное скользящее среднее
6	11,51	11,7475	11,25
7	13,69	13,0000	12,37
8	12,04	14,1675	13,58
9	14,76	15,0225	14,59
10	16,18	15,7600	15,39
11	17,11	15,8225	15,79
12	14,99	15,7725	15,80
13	15,01	15,3150	15,54
14	16,00	14,5050	14,91
15	15,26		—
16	11,75	—	—

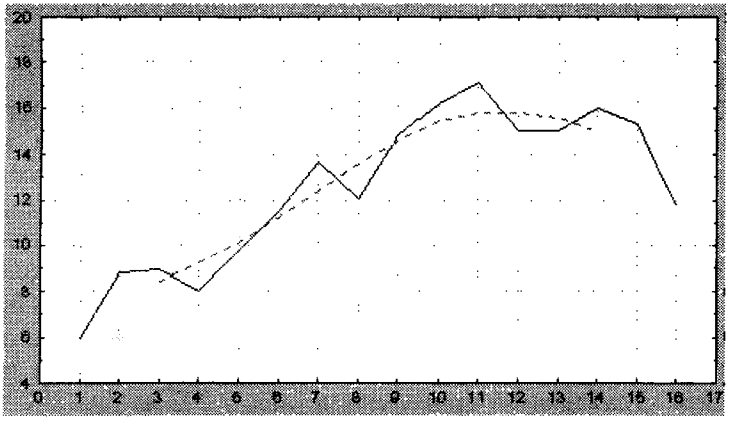


Рис. 7.9. — — исходный ряд; --- — сглаженный ряд, полученный при помощи центрированного скользящего среднего по четырем точкам

Метод скользящих средних используется и как промежуточная процедура выделения тренда в сезонной декомпозиции временного ряда (см. ниже, п. 7.3). В этом случае базу для процедуры скользящего среднего нужно выбирать равной или кратной периоду сезонных колебаний.

7.2.1. Процедура скользящего среднего с весами

Этот метод применяется если тренд временного ряда имеет вид нелинейной функции.

Любую гладкую функцию можно достаточно точно аппроксимировать полиномом в окрестности заданной точки. Рассмотрим следующую про-

цедуру, эквивалентную вычислению скользящего среднего с весами для базы содержащей $2m+1$ членов временного ряда. Подбираем полином порядка k к первой группе, содержащей $2m+1$ членов ряда и используем этот полином для определения значения тренда в $(m+1)$-й средней точке группы. Далее подбираем полином того же порядка к группе, содержащей 2, 3, ..., $2m+2$ члены ряда и определяем тренд в $(m+2)$-й точке и так далее. Покажем, что эта процедура эквивалентна вычислению скользящего среднего с весами, определенными в зависимости от числа точек в базе и порядка полинома.

Пример 7.4. Предположим, что полином третьего порядка подбирается к группам из семи точек. Найти весовые коэффициенты для процедуры скользящего среднего.

Решение. Чтобы упростить дальнейшие вычисления в качестве базы возьмем значения временного ряда y_t при значениях $t=-3, -2, -1, 0, 1, 2, 3$. Искомый полином, представляющий тренд u_t временного ряда, имеет вид

$$u_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3.$$

Определим параметры β_i, $i=0, 1, 2, 3$, используя метод наименьших квадратов, т. е. из условия минимума суммы квадратов

$$Q = \sum_{t=-3}^{3} (y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \beta_3 t^3)^2.$$

Из необходимых условий минимума функции Q, получим следующую систему уравнений:

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = 2\sum (y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \beta_3 t^3)(-1) = 0; \\ \dfrac{\partial Q}{\partial \beta_1} = 2\sum (y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \beta_3 t^3)(-t) = 0; \\ \dfrac{\partial Q}{\partial \beta_2} = 2\sum (y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \beta_3 t^3)(-t^2) = 0; \\ \dfrac{\partial Q}{\partial \beta_3} = 2\sum (y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \beta_3 t^3)(-t^3) = 0. \end{cases}$$

Преобразовав эти уравнения и используя то, что при $t=-3, -2, -1, 0, 1, 2, 3$:

$$\sum t = \sum t^3 = \sum t^5 = 0;$$

$$\sum t^2 = 2(9 + 4 + 1) = 28;$$

$$\sum t^4 = 2(81 + 16 + 1) = 196;$$

$$\sum t^6 = 2(729 + 64 + 1) = 1588.$$

Получим следующую систему уравнений:

$$\begin{cases} \sum y_t = 7\beta_0 + 28\beta_2; \\ \sum t y_t = 28\beta_2 + 196\beta_3; \\ \sum t^2 y_t = 28\beta_0 + 196\beta_2; \\ \sum t^3 y_t = 196\beta_1 + 1588\beta_3. \end{cases} \quad (3)$$

Так как необходимо вычислить значение тренда в средней точке базы, т. е. при $t = 0$: $u_0 = \beta_0$, то, решая совместно первое и третье уравнение системы (3), получим

$$\beta_0 = \frac{1}{21}\left[7\left(\sum_{t=-3}^{3} y_t\right) - \sum_{t=-3}^{3} t^2 y_t\right] =$$

$$= \frac{1}{21}[7(y_{-3} + y_{-2} + y_{-1} + y_0 + y_1 + y_2 + y_3) - (9y_{-3} + 4y_{-2} + y_{-1} + 0 + y_1 + 4y_2 + 9y_3)] =$$

$$= \frac{1}{21}[-2y_{-3} + 3y_{-2} + 6y_{-1} + 7y_0 + 6y_1 + 3y_2 - 2y_3].$$

Таким образом значение тренда в какой-либо точке вычисляется как сумма ряда в семи точках, взятых с весами

$$\frac{1}{21}[-2,\ 3,\ 6,\ 7,\ 6,\ 3,\ -2],$$

причем значение тренда вычисляется в центральной, четвертой, точке базы.

В силу симметрии весов их можно записать короче

$$\frac{1}{21}[-2,\ 3,\ 6,\ 7,\ \ldots].$$

Для примера рассмотрим ряд, описываемый точно кубической зависимостью: $y_t = (t-1)^3$.

При $t = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ получим следующие значения ряда: 0, 1, 8, 27, 64, 125, 216, 343, 512, 729. Вычислим значение тренда при $t = 4$, используя приведенные выше весовые коэффициенты. Имеем

$$u_4 = \frac{1}{21}[-2y_1 + 3y_2 + 6y_3 + 7y_4 + 6y_5 + 3y_6 - 2y_7] =$$

$$= \frac{1}{21}[-2 \cdot 0 + 3 \cdot 1 + 6 \cdot 8 + 7 \cdot 27 + 6 \cdot 64 + 3 \cdot 125 - 2 \cdot 216] = 27.$$

Полученное значение точно совпадает со значением временного ряда y_t при $t = 4$, так как при описании ряда использовался полином третьей степени.

Аналогично вычисляются веса при подборе к $2m + 1$ точкам полинома порядка k. Веса находят из условия минимума суммы квадратов

$$Q = \sum_{t=-m}^{m} (y_t - \beta_0 - \beta_1 t - \ldots - \beta_k t^k)^2.$$

Приведем веса для подсчета скользящего среднего при подборе полиномов различной степени:

Число точек в базе	Веса при использовании полиномов второй и третьей степени
5	$\frac{1}{35}[-3, 12, 17, ...]$
7	$\frac{1}{21}[-2, 3, 6, 7, ...]$
9	$\frac{1}{231}[-21, 14, 39, 54, 59, ...]$
Число точек в базе	Веса при использовании полиномов четвертой и пятой степени
7	$\frac{1}{231}[5, -30, 75, 131, ...]$
9	$\frac{1}{429}[15, -55, 30, 135, 179, ...]$

Вычисление тренда в конце ряда

При определении тренда методом скользящих средних по $2m + 1$ точке не вычисляются значения в первых и последних m точках ряда. Покажем, как можно получить формулы для вычисления последних (что наиболее важно) значений. Продолжим рассмотрение примера 7.4 с подбором полинома 3-го порядка. Решая систему (3) определим значения β_1, β_2, β_3:

$$\beta_1 = \frac{1}{1512}[397\sum ty_t - 49\sum t^3 y_t],$$

$$\beta_2 = \frac{1}{84}[-4\sum y_t + \sum t^2 y_t],$$

$$\beta_3 = \frac{1}{216}[-7\sum ty_t + \sum t^3 y_t].$$

Значения тренда в конце ряда при ($t = 1, 2, 3$) будут:

$$u_1 = \beta_0 + \beta_1 + \beta_2 + \beta_3 = \frac{1}{42}[y_{-3} - 4y_{-2} + 2y_{-1} + 12y_0 + 19y_1 + 16y_2 - 4y_3];$$

$$u_2 = \beta_0 + 2\beta_1 + 4\beta_2 + 8\beta_3 = \frac{1}{42}[4y_{-3} - 7y_{-2} - 4y_{-1} + 6y_0 + 16y_1 + 19y_2 + 8y_3];$$

$$u_3 = \beta_0 + 3\beta_1 + 9\beta_2 + 27\beta_3 = \frac{1}{42}[-2y_{-3} + 4y_{-2} + y_{-1} - 4y_0 - 4y_1 + 8y_2 + 39y_3].$$

Подробные таблицы, содержащие веса для расчета скользящих средних приведены в [25].

Влияние процедуры выделения тренда на случайную остаточную величину

Предположим, что имеется случайный ряд

$$y_t = \varepsilon_t,$$

где ε_t — некоррелированные случайные величины, $M[\varepsilon_t] = 0$; $D[\varepsilon_t] = \sigma^2$; $\operatorname{cov}(\varepsilon_t, \varepsilon_{t+k}) = 0$, $t = 1, 2, ..., n$; $k = 1, 2, ..., n$.

Пусть $a_1, a_2, ..., a_{2m+1}$ — веса, используемые для вычисления скользящего среднего с базой, содержащей $2m + 1$ точки. Значение тренда в центральной $(m + 1)$-ой точке будет

$$u_{t+m+1} = \sum_{k=1}^{2m+1} a_k \varepsilon_{t+k}.$$

Вычислим математическое ожидание и дисперсию сглаженного ряда

$$M[u_t] = 0;\ D[u_t] = \sigma^2 \sum_{k=1}^{2m+1} a_k^2.$$

Так как $|a_k| < 1$, то дисперсия сглаженного ряда будет меньше дисперсии исходного ряда σ^2. Найдем ковариации и сериальные корреляции ρ_k сглаженного ряда

$$\operatorname{cov}(u_{t+m+1}, u_{t+m+1+k}) = M[(a_1\varepsilon_{t+1} + a_2\varepsilon_{t+2} + ...)(a_1\varepsilon_{t+1+k} + a_2\varepsilon_{t+2+k} + ...)] =$$
$$= \sigma^2 \sum_{j=1}^{2m+1-k} a_j a_{j+k}.$$

Таким образом

$$\rho_k = \frac{\operatorname{cov}(u_t, u_{t+k})}{\sigma^2} = \frac{\sum_{j=1}^{2m+1-k} a_j a_{j+k}}{\sum_{j=1}^{2m+1} a_j^2}$$

и, следовательно, сглаженный ряд имеет ненулевые сериальные корреляции. В силу этого в сглаженном ряду появится циклическая составляющая.

Появление систематических колебаний как результат процедуры сглаживания временного ряда с помощью скользящих средних называется эффектом Слуцкого—Юла.

7.2.2. Понижение порядка полиномиального тренда при помощи процедуры последовательного взятия разностей

Пусть y_t, $t = 1, 2, ..., n$ — исходный временной ряд. Составим ряд, состоящий из первых разностей: $\Delta y_1, ..., \Delta y_{n-1}$, где $\Delta y_t = y_{t+1} - y_t$.

Если ряд y_t имеет линейный тренд

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t,$$

то ряд Δy_t, $t = 1, 2, ..., n - 1$, составленный из первых разностей, не будет содержать тренда. Действительно

$$\Delta y_t = y_{t+1} - y_t = \beta_0 + \beta_1(t + 1) + \varepsilon_{t+1} - (\beta_0 + \beta_1 t + \varepsilon_t) = \beta_1 + \varepsilon_{t+1} - \varepsilon_t.$$

При этом, если случайная компонента ε_t ряда y_t состояла из последовательности взаимно независимых случайных величин, то случайная компонента $(\varepsilon_{t+1} - \varepsilon_t)$ ряда из первых разностей y_t будет состоять из последовательностей взаимно коррелированных случайных величин.

Аналогично, можно показать, что если исходный ряд y_t имеет квадратичный тренд, то ряд составленный из вторых разностей $\Delta^2 y_1$, $\Delta^2 y_2$, ..., $\Delta^2 y_{n-2}$, где $\Delta^2 y_t = \Delta y_{t+1} - \Delta y_t = y_{t+2} - 2y_{t+1} + y_t$ не будет содержать тренда.

Таким образом процедуру последовательного взятия разностей можно использовать для удаления и понижения порядка полиномиального тренда в исходном временном ряду.

Еще раз обратим внимание на то, что при использовании этой процедуры случайная составляющая будет состоять из последовательности взаимно коррелированных случайных величин, и это может существенно усложнить ее дальнейший анализ.

Если ряд содержит тренд, представляемый полиномом степени k и случайную составляющую ε_t, то ряд из разностей k-го порядка, $\Delta^k y_t$ должен содержать только случайную составляющую. Предположим, что ε_t — некоррелированные случайные величины с нулевым математическим ожиданием и постоянной дисперсией

$$M[\varepsilon_t] = 0; \quad D[\varepsilon_t] = \sigma^2, \quad t = 1, 2, \ldots$$

Рассмотрим как влияет вычисление последовательных разностей на математическое ожидание и дисперсию случайной компоненты. Последовательно вычисляя разности, получим

$$\Delta^r \varepsilon_t = \varepsilon_{t+r} - C_r^1 \varepsilon_{t+r-1} + C_r^2 \varepsilon_{t+r-2} - \ldots + (-1)^r \varepsilon_t,$$

где C_r^m — биномиальные коэффициенты, $C_r^m = \dfrac{r!}{m!(r-m)!}$.

Вычисляя математическое ожидание и дисперсию для разностей k-го порядка, имеем:

$$M[\Delta^r \varepsilon_t] = 0;$$

$$D[\Delta^r \varepsilon_t] = \sigma^2[1 + (C_r^1)^2 + (C_r^2)^2 + \ldots + 1] = \sigma^2 C_{2r}^r,$$

так как сумма в скобках равна коэффициенту при x^2 в произведении двух биномов: $(x+1)^r (x+1)^r$.

Если в результате вычисления последовательных разностей k-го порядка тренд исключен, то оценка дисперсии случайной составляющей вычисляется по формуле

$$V_k = \tilde{\sigma}^2 = \frac{S_k^2}{C_{2k}^2},$$

где S_k^2 — оценка дисперсии ряда $\Delta^k y_t$. Порядок тренда можно оценить, вычисляя последовательно разности исходного ряда: Δy_t, $\Delta^2 y_t$, ... и определяя оценки дисперсии V_1, V_2, ... Порядок тренда не превышает числа k, начиная с которого оценки дисперсии V_k, V_{k+1}, ... становятся приблизительно

постоянными. Этот метод не всегда приводит к решению. Это может быть связано с тем, что процедура последовательных разностей искажает другие компоненты ряда, значения V_1, V_2, ... коррелированны, недостаточно членов в исходном ряду и т. д.

Обычно для оценки тренда используют полиномы не выше третьего порядка. Для оценки трендов имеющих большую скорость роста можно использовать другие функции (например экспоненциальную функцию).

7.3. Определение сезонной составляющей ряда (сезонных индексов) и сезонная декомпозиция временного ряда

Предположим, что исходный ряд представлен мультипликативной моделью

$$y_t = u_t W_t S_t \varepsilon_t.$$

Если в ряду записаны квартальные данные, то сезонные эффекты определяются сезонными индексами, вычисляемыми для каждого квартала: S_1, S_2, S_3, S_4 в процентах. Сумма квартальных сезонных индексов $S_1 + ... + S_4 = 400$ %.

Если ряд представляет месячные данные, то вычисляются месячные сезонные индексы: S_1, S_2, ..., S_{12}; сумма месячных сезонных индексов: $S_1 + S_2 + ... + S_{12} = 1200$ %.

В случае, когда сезонная составляющая имеет период, равный семи (например, для данных по дням недели), вычисляются семь сезонных индексов S_1, S_2, ..., S_7; суммы индексов $S_1 + S_2 + ... + S_7 = 700$ %.

Рассмотрим вычисление квартальных сезонных индексов. Процедура состоит из нескольких шагов.

Шаг 1. Выделяются тренд и циклическая составляющая при помощи процедуры центрированного скользящего среднего по четырем точкам, так как ряд представляет квартальные данные.

Шаг 2. Вычисляются сезонная и остаточная составляющие в процентах делением исходных данных на значение тренда и циклической составляющей, полученных на шаге 1:

$$\frac{u_t w_t S_t \varepsilon_t}{u_t w_t} \cdot 100 = S_t \varepsilon_t \cdot 100.$$

Шаг 3. Вычисляются средние результатов шага 2 для каждого квартала: S_1', S_2', S_3', S_4'. При этом минимальные и максимальные значения в совокупностях квартальных данных отбрасываются.

Шаг 4. Определяется корректирующий коэффициент:

$$K = 400/(S_1' + S_2' + S_3' + S_4').$$

Шаг 5. Определяются скорректированные сезонные индексы: $S_i = kS_i'$, $i = 1, 2, 3, 4$, сумма которых в точности равна 400 %.

Пример 7.5. Определение квартальных сезонных индексов методом отношения к скользящему среднему.

Найти сезонные индексы для квартальных данных объема продаж мороженого (в тоннах) в течение пяти лет.

Решение. В табл. 7.3 приведены исходные данные y_t (столбец 3), центрированные скользящие средние по четырем точкам, определяющие тренд u_t (столбец 4). В этом примере можно считать, что циклическая составляющая отсутствует. В столбце 5 вычислены отношения исходных данных y_t к тренду u_t в процентах, определяющие сезонную и остаточную составляющие.

Таблица 7.3

Годы	Кварталы	Объем продаж, y_t	Шаг 1. 4-квартальные центрирован. скользящие средние, u_t	Шаг 2. Отношение исх. данных к скользящим средним в %, $(y_t/u_t)100$	S_t, сезонные индексы	$(y_t/S_t)100$	ε_t, остаточная составляющая
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	1	8,8	—	—	63,54	—	—
	2	13,5	—	—	100,96	—	—
	3	18,9	14,17	133,33	135,08	13,99	0,99
	4	15,0	14,61	102,65	100,42	14,94	1,02
2	1	9,8	15,32	63,95	63,54	15,42	1,01
	2	16,0	15,96	100,23	100,96	15,85	0,99
	3	22,1	16,34	135,27	135,08	16,36	1,00
	4	16,9	16,70	101,20	100,42	16,83	1,01
3	1	10,9	17,21	63,33	63,54	17,15	1,00
	2	17,8	17,70	100,56	100,96	17,63	1,00
	3	24,4	18,07	134,99	135,08	18,06	1,00
	4	18,5	18,55	99,73	100,42	18,42	0,99
4	1	12,3	19,27	63,81	63,54	19,36	1,00
	2	20,2	19,91	101,44	100,96	20,01	1,00
	3	27,8	20,27	137,11	135,08	20,58	1,01
	4	20,2	20,79	97,17	100,42	20,12	0,97
5	1	13,5	21,81	61,89	63,54	21,25	0,97
	2	23,1	22,69	101,82	100,96	22,88	1,01
	3	33,1	—	—	135,08	—	—
	4	21,9	—	—	100,42	—	—

Шаг 3. Запишем данные столбца 5 в виде табл. 7.4 и вычислим среднее по кварталам, отбрасывая наибольшие и наименьшие значения.

Таблица 7.4

	Кварталы			
Год	1	2	3	4
1	—	—	~~133,33~~	~~102,65~~
2	~~63,95~~	~~100,23~~	135,27	101,20
3	63,33	100,56	134,99	99,73
4	63,81	101,44	~~137,11~~	~~97,17~~
5	~~61,89~~	~~101,82~~	—	—
Сумма	127,14	202	270,26	200,93
Среднее, S_i'	63,57	101,00	135,13	100,46

В последней строке табл. 7.4 вычислены некоррелированные сезонные индексы S_i', $i = 1, 2, 3, 4$.

Шаг 4. Вычисляем корректирующий коэффициент:

$$K = 400/(63{,}57 + 101 + 135{,}13 + 100{,}43) \approx 0{,}9996.$$

Шаг 5. Вычисляем скорректированные сезонные индексы:

$$S_1 = 0{,}9996 \cdot 63{,}57 = 63{,}54,$$
$$S_2 = 0{,}9996 \cdot 101 = 100{,}96,$$
$$S_3 = 0{,}9996 \cdot 135{,}13 = 135{,}08,$$
$$S_4 = 0{,}9996 \cdot 100{,}46 = 100{,}42.$$

Сезонные индексы выписаны в столбце 6 (см. табл. 7.3). В столбце 7 приведены исходные данные без сезонной составляющей: $(y_t/S_t)100$, а в столбце 8 — остаточная компонента ε_t, полученная делением столбца 7 на столбец 4.

Замечания. 1. Приведенный выше метод определения сезонной составляющей не имеет достаточного математического обоснования, однако широко используется в практике обработки временных рядов. Имеются несколько модификаций этого метода. В частности: тренд можно определять по уравнению регрессии; при определении сезонных индексов вычисляются не средние арифметические, а медианы; в некоторых случаях рекомендуется применять сглаживание к данным без сезонной составляющей и так далее.

2. Если временной ряд представлен аддитивной моделью, то на первом и третьем шагах выполняются те же операции, что и для мультипликативной модели. На втором, четвертом и пятом шагах метод используется так.

Шаг 2. После выделения тренда и циклической составляющей, вычитают эти компоненты из исходных данных: $y_t - (u_t + W_t) = S_t + \varepsilon_t$.

Шаг 4. Сумма сезонных индексов для аддитивной модели должна быть равна нулю, поэтому корректирующий коэффициент вычисляется по формуле

$$c = (S_1' + S_2' + S_3' + S_4')/4.$$

Шаг 5. Вычисляют скорректированные сезонные индексы

$$S_i = S_i' - c,\ i = 1, 2, 3, 4.$$

3. Если временной ряд представлен месячными данными, то вычисляется двенадцать сезонных индексов по той же схеме.

4. Качество той или иной модели временного ряда оценивается по остаточной компоненте, в частности по значению ее дисперсии, некоррелированности остатков (по критерию Дарбина—Уотсона, см. п. 6.1), а также по значению ошибки прогноза.

7.3.1. Прогонозирование ряда по тренду и сезонной составляющей

Для ряда, представленного мультипликативной моделью

$$y_t = u_t S_t \varepsilon_t, \quad t = 1, 2, ..., n,$$

прогнозируемое значение для $t = n + k$, $\tilde{y}_{n+k}$, вычисляется по формуле

$$\tilde{y}_{n+k} = \tilde{u}_{n+k} \cdot \frac{S_i}{100},$$

где S_i — значение сезонного индекса, соответствующего моменту времени $t = n + k$, а $\tilde{u}_{n+k}$ — прогнозируемое значение тренда.

В случае аддитивной модели прогнозируемое значение на момент $t = n + k$ вычисляется по формуле

$$\tilde{y}_{n+k} = \tilde{u}_{n+k} + S_i.$$

Если для оценки тренда использовался метод скользящих средних (как в примере 7.5), то для вычисления прогнозируемого значения тренда $\tilde{u}_{n+k}$ нужно воспользоваться подходящей функцией времени, например, полиномиальной регрессией $u_t = \beta_0 + \beta_1 t + \beta_2 t^2 + ...$. Параметры β_0, β_1, β_2, ... определяются методом наименьших квадратов (см. главу 6), в качестве исходных данных можно взять сглаженные значения ряда, вычисленные методом скользящих средних.

Для оценки ошибки прогноза используется следующий метод. Временной ряд $y_1, y_2, ..., y_n$ — разбивается на две части:

$y_1, y_2, ..., y_m$ — значение ряда в период предыстории;

$y_{m+1}, y_{m+2}, ..., y_{m+d}$ — значения ряда в период прогноза, где d — прогнозируемый период, $m + d = n$.

По значениям ряда в период предыстории $y_1, y_2, ..., y_m$ оценивается тренд, сезонные индексы и вычисляются значения прогноза $\tilde{y}_{m+1}, \tilde{y}_{m+2}, ..., \tilde{y}_{m+d}$. Далее определяются ошибки прогноза $e_{m+1} = y_{m+1} - \tilde{y}_{m+1}$, $e_{m+2} = y_{m+2} - \tilde{y}_{m+2}$, ...

Точность прогноза определяется по средней абсолютной относительной ошибке прогноза (MAPE) в процентах и зависит от выбора значений m и d:

$$\text{MAPE} = e(m,d) = \frac{1}{d}\sum_{k=1}^{d}\left|\frac{y_{m+k} - \tilde{y}_{m+k}}{y_{m+k}}\right| \cdot 100\ \%,$$

где y_{m+k} — фактическое значение временного ряда; $\tilde{y}_{m+k}$ — прогнозируемое значение.

Пример 7.6. Прогнозирование и вычисление ошибки прогнозирования для данных примера 7.2.

Решение. Рассчитаем параметры квадратичного тренда по m первым значениям ряда. Результаты расчетов приведены в табл. 7.5, где даны три варианта выбора периодов предыстории и прогноза и соответствующие уравнения трендов, и в табл. 7.6 с исходными данными, результатами прогноза и ошибками прогноза по каждому варианту расчета.

Таблица 7.5

№ варианта	Период предыстории, m	Период прогноза, $d = 16 - m$	Уравнение тренда, вычисленное по данным периода предыстории
1	10	6	$6{,}1447 + 0{,}6394 \cdot t + 0{,}0341 \cdot t^2$
2	12	4	$5{,}1973 + 1{,}1726 \cdot t - 0{,}0184 \cdot t^2$
3	14	2	$4{,}5784 + 1{,}4940 \cdot t - 0{,}0474 \cdot t^2$

Таблица 7.6

Исходные данные		Прогнозируемые значения, $\tilde{y}_t$		
t	y_t	Вариант 1	Вариант 2	Вариант 3
1	6,00	6,8182	6,3514	6,0250
2	8,82	7,5599	7,4687	7,3768
3	8,94	8,3698	8,5491	8,6338
4	8,05	9,2478	9,5925	9,7960
5	9,75	10,1941	10,5991	10,8633
6	11,51	11,2085	11,5688	11,8359
7	13,69	12,2911	12,5016	12,7136
8	12,04	13,4419	13,3975	13,4964
9	14,76	14,6608	14,2565	14,1845
10	16,18	15,9480	15,0786	14,7777
11	17,11	17,3033	15,8639	15,2761
12	14,99	18,7268	16,6122	15,6797
13	15,01	20,2185	17,3236	15,9885
14	16,00	21,7784	17,9982	16,2025
15	15,26	23,4065	18,6358	16,3216
16	11,75	25,1027	19,2366	16,3459
Ошибки прогноза, %		43,95	28,42	23,33

Как видно из табл. 7.6, оценка ошибки прогноза, определяемая таким методом, зависит от выборов предыстории и прогноза.

7.4. Прогнозирование на основе экспоненциального сглаживания

Простая процедура прогноза строится при помощи усреднения всех прошлых значений временного ряда, которые используются с весами, убывающими по геометрическому или экспоненциальному закону. Эта процедура называется *экспоненциальным сглаживанием.*

Простое экспоненциальное сглаживание предполагает, что временной ряд y_t, $t = 1, 2, ..., n$ представлен следующей моделью

$$y_t = a_0 + \varepsilon(t),$$

где $\varepsilon(t)$ — случайная ошибка, а значение a_0 постоянно на данном временном интервале, но может медленно изменяться со временем. Для определения a_0 применяется процедура скользящего среднего, в которой используются все предыдущие наблюдения с весами, изменяющимися по геометрическому или экспоненциальному закону. Последнему наблюдению y_n приписывается наибольший вес, предпоследнему y_{n-1} — вес меньший, чем последнему и т. д.

Таким образом, для временного ряда: $y_1, y_2, y_3, ..., y_n$ прогноз следующего значения $\tilde{y}_{n+1}$ определяется по формуле

$$\tilde{y}_{n+1} = (1-\beta)(y_n + \beta y_{n-1} + \beta^2 y_{n-2} + \ldots), \tag{4}$$

где $0 < \beta < 1$. Сумма весов равна 1:

$$(1-\beta)(1+\beta+\beta^2+\ldots) = (1-\beta)\frac{1}{1-\beta} = 1.$$

Формулу (4) можно преобразовать к более удобному виду

$$\tilde{y}_{n+1} = (1-\beta)y_n + \beta\{(1-\beta)[y_{n-1} + \beta y_{n-2} + \beta^2 y_{n-3} + \ldots]\} =$$
$$= (1-\beta)y_n + \beta\tilde{y}_n, \tag{5}$$

так как выражение в фигурных скобках дает прогноз предыдущего значения $\tilde{y}_n$.

Обычно формулу для вычисления прогноза на основе простого экспоненциального сглаживания записывают в виде

$$\tilde{y}_{n+1} = \alpha y_n + (1-\alpha)\tilde{y}_n,$$

где $\alpha = 1 - \beta$ называется параметром экспоненциального сглаживания по Брауну, по имени ученого впервые предложившего данную процедуру.

Этот метод обобщается на случай, когда ряд представлен более сложными моделями.

Линейная модель временного ряда имеет вид

$$y_t = a_0 + a_1 t + \varepsilon(t).$$

Приведем окончательные формулы для расчета значений $a_0(n)$ и $a_1(n)$, по предыдущим значениям $a_0(n-1)$ и $a_1(n-1)$ [25]:

$$a_0(n) = a_0(n-1) + a_1(n-1) + (1-\beta^2)e_t,$$

$$a_1(n) = a_1(n-1) + (1-\beta)^2 e_t,$$

где e_t — ошибка прогнозирования на один шаг вперед: $e_t = y_n - a_0(n-1) - a_1(n-1)$.

Прогноз на один шаг вычисляется по формуле

$$\tilde{y}_{n+1} = a_0(n) + a_1(n),$$

а прогноз на k шагов вычисляется по формуле

$$\tilde{y}_{n+k} = a_0(n) + a_1(n) \cdot k.$$

В качестве начальных значений a_0 и a_1 принимают оценки соответствующих коэффициентов линейного тренда.

При выборе параметра $\alpha = 1 - \beta$, нужно иметь в виду, что вес наблюдения y_{n-k}, отстоящего на k периодов, будет равен: $\alpha(1-\alpha)^k = (1-\beta)\beta^k$.

Если начальные условия достоверны, то α берут небольшим ($\alpha \approx 0$). При выборе α близким к единице при расчете прогноза учитываются в большей степени последние наблюдения. Небольшие изменения α мало сказываются на результатах прогноза. В некоторых случаях хорошие результаты дает определение α по формуле

$$\alpha = 2/(m+1),$$

где m — число наблюдений, входящих в интервал сглаживания ряда. В этом случае суммарный вес последних m наблюдений

$$\alpha \sum_{k=0}^{m-1} (1-\alpha)^k \approx 0{,}865 \quad (\text{при } m \geq 100),$$

т. е. ~87 % веса имеют последние наблюдения.

В более сложных моделях экспоненциального сглаживания оцениваются различные тренды и сезонная составляющая (см. п. 7.6). Такие модели определяются несколькими параметрами: α, δ, γ, φ. Для оценки параметров используются специальные методы оптимизирующие критерии качества прогноза.

При обработке временных рядов с помощью пакетов статистических программ параметры процедур можно подбирать, используя графики результатов, а также интегральные характеристики качества процедур. В пакете STATISTICA вычисляются следующие характеристики качества прогноза:

Mean error (средняя ошибка), $\mathrm{ME} = \frac{1}{n}\sum_{k=1}^{n}(y_k - \tilde{y}_k)$;

Sums of squares (сумма квадратов ошибок), $\mathrm{SSE} = \sum_{k=1}^{n}(y_k - \tilde{y}_k)^2$;

Mean square (средняя квадратическая ошибка), $\mathrm{MSE} = \frac{1}{n}\sum_{k=1}^{n}(y_k - \tilde{y}_k)^2$;

Mean absolute error (средняя абсолютная ошибка), $\mathrm{MAE} = \frac{1}{n}\sum_{k=1}^{n}|y_k - \tilde{y}_k|$;

Mean absolute percentage error (средняя абсолютная относительная ошибка), $\text{MAPE} = \frac{1}{n}\sum_{k=1}^{n}\left|\frac{y_k - \tilde{y}_k}{y_k}\right| \cdot 100\ \%$;

Mean percentage error (средняя относительная ошибка),

$$\text{MPE} = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{y_k - \tilde{y}_k}{y_k}\right) \cdot 100\ \%,$$

где y_k и $\tilde{y}_k$, $k = 1, 2, \ldots, n$ — соответственно исходные данные и результаты процедуры прогноза.

Недостатком средней ошибки (ME) является то, что положительные и отрицательные ошибки компенсируют друг друга, поэтому ME не является хорошим показателем качества прогноза. Средняя абсолютная ошибка (MAE) по сравнению со средней квадратической ошибкой (MSE) более устойчива по отношению к выбросам. Относительные ошибки позволяют учесть тот факт, что при прогнозе, например месячных данных, они могут достаточно сильно изменяться от месяца к месяцу (в зависимости от сезона). При расчете средней относительной ошибки (MPE) отрицательные и положительные относительные ошибки будут компенсировать друг друга. Поэтому для оценки качества прогноза (для всего ряда) лучше использовать среднюю абсолютную относительную ошибку (MAPE).

7.5. Стационарные временные ряды. Процессы авторегрессии первого и второго порядков

Стационарные временные ряды представляют данные в которых тренд, сезонная и циклическая составляющие либо отсутствуют, либо исключены. Для стационарных рядов математическое ожидание и дисперсия — константы:

$$M[Y(t)] = m;$$

$$D[Y(t)] = \sigma^2,$$

а автокорреляционная функция k-го порядка не зависит от t, причем

$$\rho_k = \rho_{-k}\ .$$

Можно показать, что для стационарного ряда значения сериальных корреляций ρ_k связаны определенными соотношениями.

Рассмотрим автоковариационную и автокорреляционную матрицы стационарного процесса $Y(t)$:

$$K = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{n-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{n-1} & \gamma_{n-2} & \gamma_{n-3} & \cdots & \gamma_0 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix},$$

где $\gamma_k = \text{cov}(Y(t),\ Y(t+k))$, $\gamma_0 = D[Y(t)] = \sigma^2$.

Пусть L_t — линейная функция случайных величин $Y(t), Y(t-1), \ldots, Y(t-n+1)$:

$$L_t = a_1 Y(t) + a_2 Y(t-1) + \ldots + a_n Y(t-n+1), \quad a_i \geq 0,\ i = 1, 2, \ldots, n.$$

Так как для стационарного процесса ковариация

$$\operatorname{cov}[Y(i),\ Y(j)] = \gamma_{|i-j|},$$

то дисперсия L_t равна

$$D[L_t] = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \gamma_{|i-j|} \geq 0.$$

Это означает, что автоковариационная и автокорреляционная матрицы стационарного процесса положительно полуопределенные и главные миноры этих матриц неотрицательны.

В, частности, при $n = 2$, имеем

$$\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix} \geq 0,$$

и, следовательно, $1 - \rho_1^2 \geq 0$, $|\rho_1| \leq 1$.

Для $n = 3$ должны выполняться следующие условия:

$$\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix} \geq 0, \quad \begin{vmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{vmatrix} \geq 0,$$

$$\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix} \geq 0, \text{ т. е. } |\rho_1| \leq 1, \quad |\rho_2| \leq 1,$$

$$-1 \leq \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \leq 1.$$

Например, если $\rho_1 = 0{,}9$, то $0{,}72 \leq \rho_2 \leq 1$.

Одной из основных моделей стационарных временных рядов является авторегрессионный процесс порядка k:

$$y_{t+k} + \alpha_1 y_{t+k-1} + \alpha_2 y_{t+k-2} + \ldots + \alpha_k y_t = \varepsilon_{t+k},$$

или

$$y_t = -\alpha_1 y_{t-1} - \alpha_2 y_{t-2} - \ldots - \alpha_k y_{t-k} + \varepsilon_t, \tag{6}$$

где ε_t — случайная составляющая, имеющая нулевое математическое ожидание и независящая от $y_{t-1}, y_{t-2}, \ldots$

Простейшими процессами этого вида являются процессы авторегрессии первого и второго порядков.

Процесс авторегрессии первого порядка определяется выражением

$$y_t = -\alpha_1 y_{t-1} + \varepsilon_t$$

или

$$y_t = \rho y_{t-1} + \varepsilon_t. \tag{7}$$

Умножая (7) на y_{t-1} и вычисляя математическое ожидание произведения $y_t \cdot y_{t-1}$, получим

$$\operatorname{cov}(y_t, y_{t-1}) = M[y_t \cdot y_{t-1}] =$$

$$= M[y_{t-1}(\rho y_{t-1} + \varepsilon_t)] = \rho D[y_t], \quad \text{т. е. } \rho_1 = \rho.$$

Аналогично, умножая (7) на y_{t-k} и вычисляя математическое ожидание, имеем

$$\operatorname{cov}(y_t, y_{t-k}) = \rho \operatorname{cov}(y_{t-1}, y_{t-k}).$$

Разделив это выражение на $D[y_t]$ получим авторреляцию k-го порядка

$$\rho_k = \rho \cdot \rho_{k-1} = \rho^2 \cdot \rho_{k-2} = \ldots = \rho^k.$$

Таким образом авторреляции процесса выражаются через первую автокорреляцию $\rho_1 = \rho : \rho_k = \rho^k$. На рис. 7.10 приведен график огибающей автокорреляционной функции процесса авторегрессии первого порядка (при $\rho > 0$).

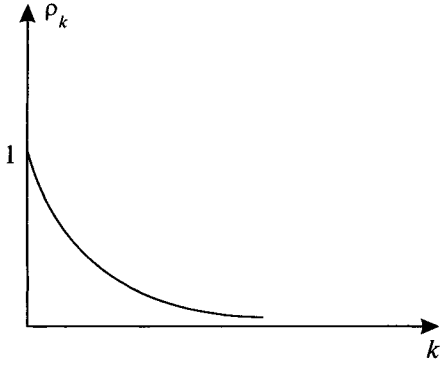


Рис. 7.10

График этого процесса представляет колебания более или менее регулярного типа.

Авторегрессионный процесс второго порядка определяется формулой

$$y_t = -\alpha_1 y_{t-1} - \alpha_2 y_{t-2} + \varepsilon_t, \tag{8}$$

где ε_t — случайная компонента, независимая от y_{t-1} и y_{t-2}, как и в авторегрессии первого порядка процессе. Умножая (8) последовательно на y_{t-1} и y_{t-2} и взяв математические ожидания, получим систему уравнений

$$\begin{cases} \rho_1 + \alpha_1 + \alpha_2 \rho_1 = 0, \\ \rho_2 + \alpha_1 \rho_1 + \alpha_2 = 0. \end{cases}$$

Решая систему уравнений, имеем:

$$\rho_1 = -\frac{\alpha_1}{1+\alpha_2};$$

$$\rho_2 = -\alpha_2 + \frac{\alpha_1^2}{1+\alpha_2}.$$

Умножая (8) на y_{t-k} и взяв математическое ожидание находим уравнение для вычисления автокорреляций более высокого порядка

$$\rho_k + \alpha_1\rho_{k-1} + \alpha_2\rho_{k-2} = 0.$$

Авторегрессионные процессы первого и второго порядка используются для описания стационарных временных рядов, в частности, для описания остаточной компоненты [10, 17, 34].

7.6. Анализ временных рядов в пакете STATISTICA

7.6.1. Работа 1. Определение тренда методом скользящих средних. Анализ сезонной составляющей

1. Основные понятия

Автокорреляционная функция. Сериальные корреляции. Аддитивная и мультипликативная модели временного ряда. Метод скользящих средних. Сезонные индексы. Случайная составляющая временного ряда.

2. Варианты заданий для работ 1 и 2

1	2	3	4	5	6	7	8
1,65	23,46	0,54	30,42	20,89	12,60	3,54	15,48
2,59	14,86	2,16	30,56	20,11	18,92	7,81	9,29
6,18	20,14	5,39	29,90	16,41	17,08	12,83	8,26
6,26	21,59	3,48	21,67	18,95	15,51	6,73	5,45
6,44	18,98	4,54	26,31	21,43	8,97	6,29	10,49
7,16	21,77	7,99	28,13	16,54	14,52	15,88	14,47
10,56	20,27	7,95	24,06	11,55	12,77	12,27	9,46
10,93	16,86	7,01	20,55	14,39	12,96	7,84	8,79
9,53	16,23	9,89	24,35	20,66	5,55	10,71	12,96
10,64	18,55	12,35	18,12	15,31	11,09	14,60	15,37
17,43	14,87	12,91	18,69	9,34	9,23	17,48	11,82
14,72	11,98	14,42	14,88	11,39	5,03	12,97	11,34

Продолжение вариантов заданий для работ 1 и 2

1	2	3	4	5	6	7	8
15,50	14,41	14,13	11,66	11,34	2,15	11,34	20,84
15,01	13,42	18,67	19,83	10,07	8,95	23,82	16,58
17,83	10,44	16,95	14,10	5,95	8,04	19,97	12,47
18,43	8,26	15,84	10,16	4,59	5,68	11,51	7,05
17,69	8,86	19,23	10,08	8,74	0,14	18,07	15,08
19,80	9,53	22,05	5,82	9,96	5,85	22,11	16,97
22,64	6,88	22,59	8,46	3,03	4,21	23,12	13,51
22,86	4,10	21,15	5,50	3,17	2,56	15,52	13,45
21,56	7,61	23,98	3,60	4,45	0,08	20,03	16,55
22,16	4,92	26,45	8,44	4,06	3,87	24,36	18,47
25,82	1,79	29,80	3,04	0,16	1,10	27,02	21,73
26,50	0,10	27,41	0,00	1,52	0,85	21,31	14,04

9	10	11	12	13	14	15	16
12,19	23,75	18,47	76,88	8,48	24,78	3,07	10,22
8,41	28,00	14,87	69,88	10,43	22,55	6,26	10,06
14,68	33,01	21,51	74,55	18,97	30,85	7,46	13,34
8,64	16,78	9,07	59,75	6,37	23,88	6,48	11,92
32,94	18,16	16,02	72,21	9,86	27,78	1,64	8,81
22,61	20,05	11,12	66,85	1,29	12,71	5,41	8,10
45,92	3,18	23,45	69,91	13,23	25,25	6,18	12,51
23,63	16,11	6,45	68,05	8,50	25,70	16,93	11,16
18,59	21,66	14,21	72,59	11,68	34,44	2,71	8,77
36,22	20,16	8,18	42,83	10,17	23,18	6,94	4,87
50,10	24,71	14,50	67,04	14,18	29,81	8,35	10,57
46,22	15,63	3,86	56,63	2,79	22,26	11,59	10,37
23,63	16,27	10,14	61,10	26,63	22,97	5,98	6,88
47,30	18,99	9,99	44,88	15,69	16,37	10,77	9,13
40,03	21,12	14,47	52,90	20,32	22,82	14,71	10,31
56,53	8,34	0,65	46,03	17,28	14,19	14,66	7,13
38,41	14,96	8,97	46,72	22,87	16,40	11,77	3,52
51,47	17,17	2,47	46,48	23,80	7,23	27,10	0,14

Продолжение вариантов заданий для работ 1 и 2

9	10	11	12	13	14	15	16
6,29	20,24	12,58	31,63	28,81	13,05	9,69	6,35
35,41	8,31	3,12	21,72	28,59	4,63	22,31	5,30
67,79	12,36	6,81	21,40	35,68	3,19	19,73	1,46
74,21	14,59	0,43	11,40	35,72	4,55	25,88	1,09
79,12	21,72	4,65	10,06	39,44	0,94	29,00	2,40
45,10	28,69	5,91	0,42	40,04	11,07	32,18	1,92

17	18	19	20	21	22	23	24
11,54	0,54	6,86	11,43	10,41	4,89	15,45	5,93
0,80	4,33	3,91	7,60	7,70	3,10	11,94	3,88
12,76	3,73	6,66	12,15	10,39	5,19	11,93	5,08
11,18	5,18	6,38	10,39	10,73	1,02	18,66	5,98
8,90	2,50	8,35	11,44	12,31	6,25	12,69	7,77
8,49	3,72	6,16	10,94	9,58	5,06	10,01	6,67
11,38	4,78	7,68	13,54	11,53	5,96	8,81	6,55
10,93	5,72	7,12	11,87	11,55	6,27	10,86	6,27
9,40	3,69	8,61	13,35	13,98	6,56	11,49	8,23
9,30	4,80	5,87	11,72	10,07	6,43	10,78	6,61
12,43	6,35	7,76	13,58	11,44	6,45	10,38	7,40
11,03	6,89	7,07	10,56	11,00	6,26	13,07	7,48
10,88	6,38	8,37	11,04	11,16	7,00	10,81	8,08
11,33	5,93	8,69	8,96	9,49	4,51	12,73	7,00
13,86	9,17	6,83	11,38	10,41	5,93	12,11	6,16
14,98	9,31	6,17	9,26	9,15	6,53	15,74	5,73
12,66	4,07	6,98	9,38	8,48	6,98	17,71	7,23
12,98	9,47	3,84	8,04	5,41	8,96	15,31	3,86
18,09	12,28	4,75	10,98	6,44	5,78	11,15	5,63
17,49	13,32	4,05	7,95	6,15	5,87	18,12	5,66
14,97	9,87	4,88	7,67	3,17	6,21	20,81	5,71
14,42	12,73	0,51	4,69	0,47	3,33	19,90	2,62
21,29	16,73	2,60	7,16	1,80	5,21	19,15	3,89
20,66	17,05	0,90	4,17	1,26	4,63	22,43	3,44

3. Задание

Используя выборку данных из своего варианта (24 значения), выполнить следующие задания:

1. Построить график ряда.

2. Вычислить сглаженные ряды, используя простые скользящие средние по:

а) трем точкам;

б) четырем точкам (после сглаживания провести центрирование);

в) пяти точкам.

Сглаженные ряды нанести на три отдельных графика вместе с исходными данными.

3. Рассчитать четыре сезонных индекса для исходного ряда по аддитивной модели ряда. Построить на одном графике:

а) исходные данные y_t;

б) центрированные скользящие средние (оценка тренда) u_t;

в) сезонные индексы = S_1, S_2, S_3, S_4;

г) данные без сезонной составляющей $V_t = y_t - S_i$;

д) остатки $e_t = V_t - u_t = y_t - S_i - u_t$.

4. Повторить расчеты из пункта 3 для мультипликативной модели ряда и построить графики а)—д).

5. Найти дисперсии остатков для обеих моделей ряда. Сравнить результаты и выбрать подходящую модель.

6. Ввести данные в пакет STATISTICA. Выполнить все расчеты в п. 1)—5), сравнить результаты и записать их в отчет.

4. Выполнение задания в пакете STATISTICA

Для того, чтобы войти в модуль Анализ временных рядов и прогнозирование необходимо в **Переключателе модулей (Statistica Module Switcher)** выбрать модуль **Time series/Forecasting**. Общее назначение модуля — построить простую модель, описывающую ряд, сгладить его, спрогнозировать будущие значения временного ряда на основе наблюдаемых до данного момента, построить регрессионные зависимости одного ряда от другого, провести спектральный или Фурье-анализ ряда и т. д.

В этом модуле реализованы следующие методы анализа временных рядов:

- ARIMA-АРПСС: модель авторегрессии и проинтегрированного скользящего среднего;
- Interrupted time series analysis — анализ прерванного временного ряда (модели интервенции для АРПСС);
- Exponential smoothing & forecasting — экспоненциальное сглаживание и прогнозирование;
- Seasonal decomposition (Census 1) — сезонная декомпозиция;
- X11 (Census 2) — monthly — quarterly — X11 метод — месячно — квартально — специальный метод сезонной декомпозиции;
- distributed lags analysis — анализ распределенных лагов;
- Spectral (Fourier) analysis — спектральный (Фурье) анализ.

Мы рассмотрим простейшие методы анализа временных рядов.

Более подробно методы анализа временных рядов см. в [17, 25, 26, 34, 35, 36].

1. Выделение тренда методом скользящих средних

В качестве примера возьмем следующие данные: 8,8; 13,5; 18,9; 15,0; 9,8; 16,0; 22,1; 16,9; 10,9; 17,8; 24,4; 18,5; 12,3; 20,2; 27,8; 20,2; 13,5; 23,1; 33,1; 21,9; 13,7; 24; 33,5; 22,1.

Введите данные в пакет STATISTICA.

Для начала анализа необходимо вызвать стартовую панель модуля, для этого войдите в меню **Analysis — Анализ** и выберите команду **Startup Panel**. Далее выберите переменную для анализа (воспользуйтесь кнопкой Variables). Кнопка **Variables — Переменные**, расположенная в левом верхнем углу, открывает диалоговое окно выбора переменных из файла данных. Нажмите ее и откройте диалоговое окно **Select the variables for the time series analysis — Выбрать переменные для анализа временных рядов**. В окне можно выбрать переменные для анализа (максимальное число переменных 20). Выбор можно осуществить либо высвечивая имена в верхней части окна, либо задавая номера переменных в нижней строке.

Высветите имена переменных — в данном случае одной переменной, и вновь нажмите кнопку OK. После определения переменной на экране появится следующее окно (рис. 7.11).

После того как файл открыт и выбраны переменные для анализа, в информационной части панели в поле **Variable — Переменная** появятся имена переменных, расширенные имена автоматически отображаются в графе **Long variable (series) name — Длинное имя переменной (рядов)**.

Слева от имени анализируемых переменных стоит значок **L** в графе **Lock**, означающий, что переменные закрыты на ключ и не могут быть удалены без прерывания анализа.

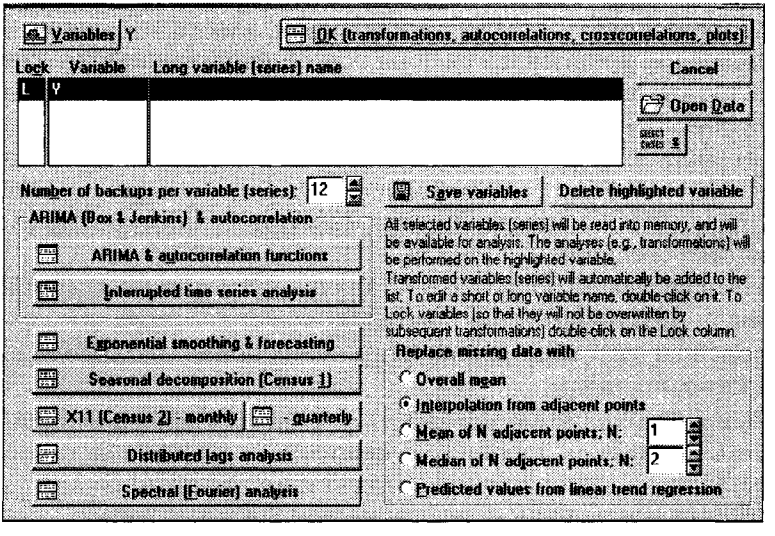


Рис. 7.11. Стартовая панель модуля

Весь дальнейший диалог происходит именно с этими переменными, которые можно преобразовывать, анализировать, но нельзя удалить из текущего анализа.

В процессе работы ряды многократно преобразовываются, однако, не все преобразования необходимы; чтобы не хранить лишнюю информацию, их следует удалить из диалога. Для этого служит кнопка **Delete highlighted variable — Удалить высвеченные переменные**.

Напротив, некоторые переменные нужно сохранить для дальнейшего анализа, например, для того чтобы применить альтернативный способ обработки.

Кнопка **Save variables — Сохранить переменные** сохраняет высвеченные переменные в файле данных STATISTICA. Сохраненную таким образом переменную можно проанализировать впоследствии в любом модуле STATISTICA.

В верхней части стартовой панели расположена опция **Number of backups per variables (series) — Число резервов для переменных (рядов)**, которая определяет число преобразований ряда в текущем диалоге. Если число преобразований превысит указанное в опции число, то система сделает запрос: сохранять очередное преобразование?

Построим график исходных данных. Для этого в окне Стартовой панели необходимо нажать на кнопку OK, в результате чего на экране появится следующее окно (рис. 7.12).

Обратите внимание на опцию **Plot variables (series) after each transformation — Построить график переменных (ряда) после каждого преобразования в данном окне**. После того как вы установите ее, система будет автоматически выдавать график преобразованных данных после каждого преобразования ряда. Опция **Display/plot subset of cases only** позволяет просмотреть численно или построить график только для определенного подмножества данных.

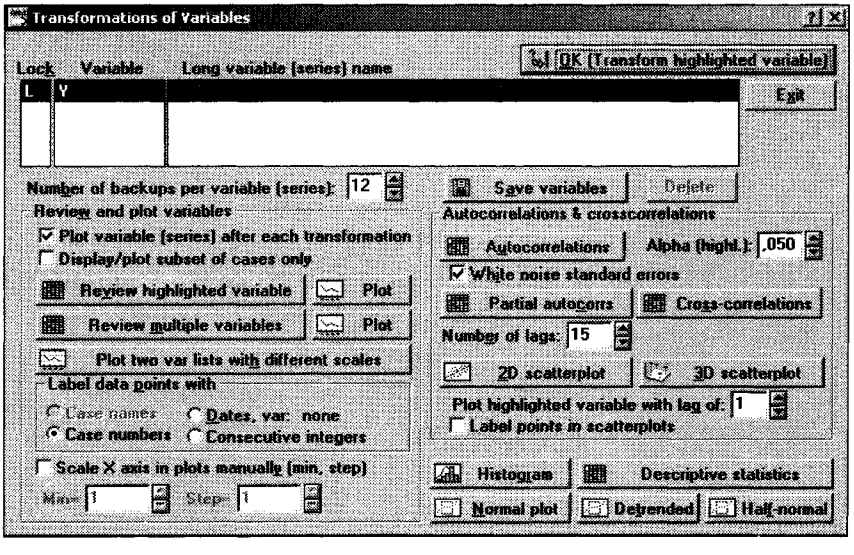


Рис. 7.12. Окно для выбора преобразования переменных

Следует обратить внимание на кнопки:

- **Review highlighted variable**, эта функция выводит на экран значения выделенной переменной;
- **Plot** (верхняя), выводит график выделенной переменной;
- **Review multiple variables**, эта функция выводит численные результаты для нескольких выделенных переменных представляющих результаты преобразования временного ряда (включая исходные данные);
- **Plot** (нижняя), выводит графики выделенных переменных.

Построим график исходных данных (рис. 7.13).

Чтобы определить тренд методом скользящих средних необходимо в этом же окне еще раз нажать на кнопку OK. В результате появится следующее окно (рис. 7.14).

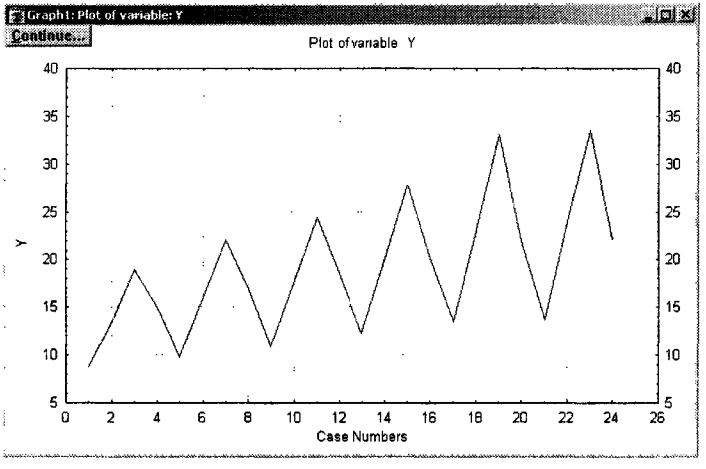


Рис. 7.13. График исходных данных

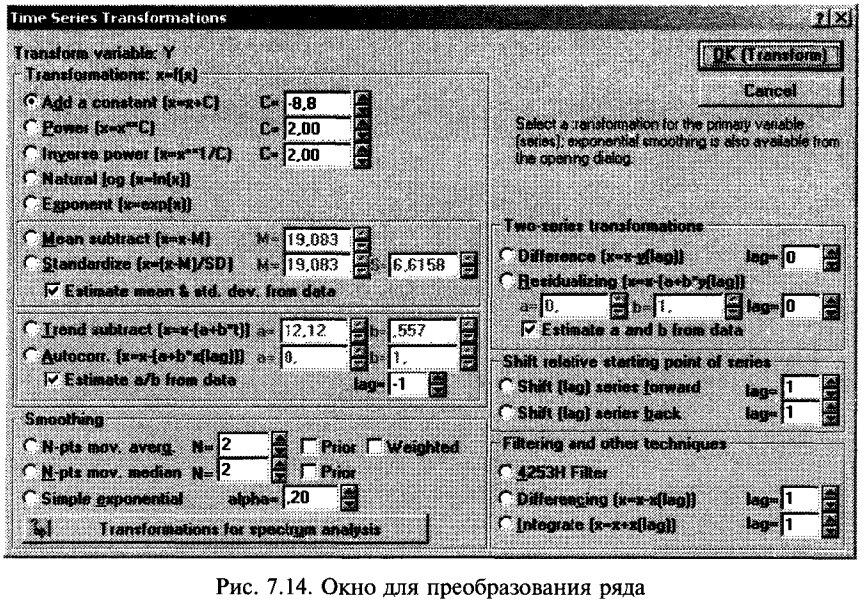


Рис. 7.14. Окно для преобразования ряда

В левой нижней части окна, в поле **Smoothing — Сглаживание** необходимо выделить опцию **N-pts mov. averg. — сглаживание по методу скользящих средних**. Данная функция позволяет произвести сглаживание по двум, трем и более точкам ($N = 2, 3, ...$). Установите $N = 4$ (сглаживание по четырем точкам) и нажмите на кнопку OK. На экране появится график сглаженного ряда.

Для того, чтобы просмотреть одновременно исходные данные и результаты процедуры простого скользящего среднего по четырем точкам нужно нажать кнопку **Continue...** и воспользоваться функцией **Review multiple variables** в окне преобразования переменных (рис. 7.12). Функция Plot позволит просмотреть скользящие средние на графике одновременно с исходными данными (рис. 7.15).

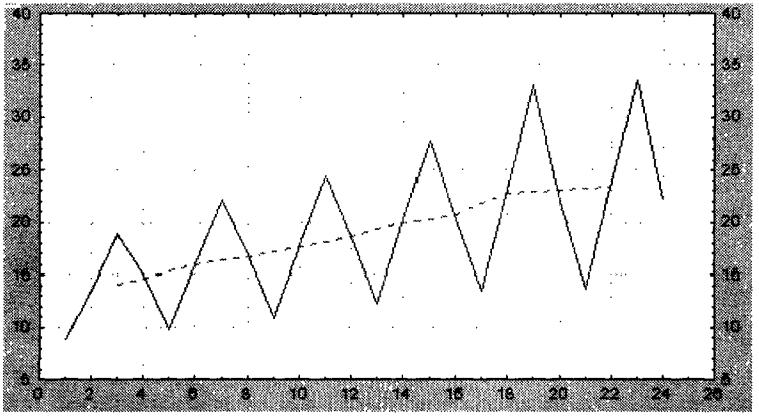


Рис. 7.15. Исходные данные и результаты сглаживания по четырем точкам

Если необходимо просмотреть результаты сглаживания численно, нужно нажать кнопку **Review multiple variables** (рис. 7.12), выбрать переменные и нажать OK.

Чтобы вернуться в стартовую панель модуля нажмите **Exit (Выход)**.

2. Преобразования временных рядов. Вычисление коррелограммы

В левой части окна Time series Transformations — преобразования временного ряда (рис. 7.14) имеется ряд опций позволяющих выполнить различные преобразования исходного временного ряда: добавление константы, возведение в целую и дробную степень, логарифмирование, стандартизация данных и др. [17, 10]. Преобразования можно выполнять последовательно. Одна из целей преобразований состоит в том, чтобы сделать ряд стационарным. В случае линейного тренда это можно сделать с помощью опции Trend subtract (удаление тренда). Параметры a и b линейного тренда $a + bt$ могут задаваться или оцениваться по исходным данным. В правом нижнем углу размещена опция Differencing — взятие разностей определенного порядка (см. п. 7.2.2).

Чтобы вычислить коррелограмму надо переключиться в окно Transformations of Variables — преобразования переменных (рис. 7.12) и выбрать

опцию Autocorrelations (автокорреляции) в правой верхней части окна. Число вычисляемых сериальных корреляций задается в окне Number of lags. На рис. 7.16 приведена коррелограмма для исходного ряда. Коррелограмма показывает, что *ряд имеет сезонную составляющую с периодом равным четырем.*

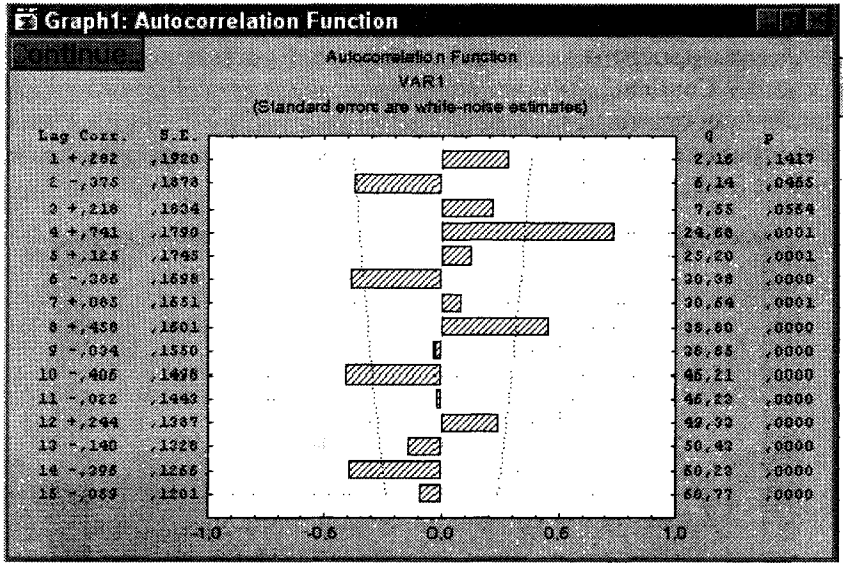


Рис. 7.16. Коррелограмма

3. Сезонная декомпозиция

Seasonal decomposition — Сезонная декомпозиция позволяет выделить в ряде сезонную компоненту, обозначаемую ***S***, тренд-циклическую компоненту — **TC** и нерегулярную (случайную) составляющую — **I**. Модель может быть мультипликативной или аддитивной.

Нажмите кнопку **Seasonal decomposition (Census 1) — Сезонная декомпозиция** на стартовой панели модуля (рис. 7.11) и откройте диалог **Сезонная декомпозиция** (рис. 7.17).

В центральной части панели находятся опции, позволяющие задать модель ряда. Эти опции объединены в группу **Seasonal model — Сезонная модель**:

- **Additive** — Аддитивная;
- **Multiplicative** — Мультипликативная.

В опции **Seasonal lag — Сезонный лаг** задается число сезонных индексов.

Следующая группа опций позволяет определить следующие составляющие:

- **Moving averages** — Скользящие средние;
- **Ratios/Differences** — Отношения/Разности (если модель мультипликативная, берется отношение, если аддитивная — разность исходного ряда и тренда);

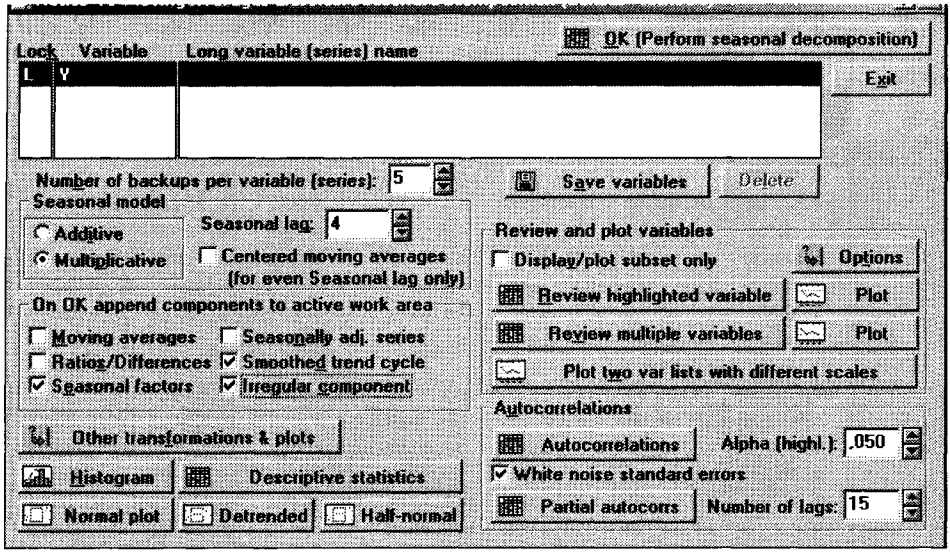


Рис. 7.17. Панель для установки параметров сезонной декомпозиции

- **Seasonal factors** — Сезонные индексы;
- **Seasonal adj. series** — Ряд без сезонной составляющей;
- **Smoothed trend cycle** — Сглаженная тренд — циклическая компонента;
- **Irregular components** — Нерегулярная (случайная) составляющая.

Запустите процедуру сезонной декомпозиции, нажав кнопку **OK (Perform seasonal decomposition) — Да (Выполнить сезонную декомпозицию).** Результаты расчетов для мультипликативной модели выводится в виде таблицы (рис. 7.18).

Для того чтобы на один график вывести графики нескольких компонент надо отметить их на левой части панели (рис. 7.17), нажать OK и затем выбрать их с помощью кнопки **Review Multiple Variables** и нажать **Plot**.

Для расчета сезонных индексов и сезонной декомпозиции мультипликативной модели в пакете STATISTICA используется метод, отличающийся от метода, описанного в п. 7.3. Таблица, приведенная на рис. 7.18, получилась в результате следующей последовательности преобразований временного ряда (Y). Во втором столбце (Moving Averages) приведены простые скользящие средние по четырем точкам временного ряда (без центрирования), причем в таблице значения скользящих средних смещены вниз на $\frac{1}{2}$ строки: среднее арифметическое четырех первых точек $\frac{8{,}8 + 13{,}5 + 18{,}9 + 15}{4} = 14{,}05$ определяется как значение скользящего среднего в третьей точке и т. д. В третьем и четвертом столбцах (Rations и Seasonal Factors) вычисляются соответственно отношения элементов исходного ряда к скользящему среднему (в процентах) и скорректированные сезонные индексы.

В пятом столбце (Adjusted Series) вычисляется ряд скорректированный на сезонные индексы, т. е. ряд без сезонной составляющей (вычисляется

Seasonal Decomposition: Multipl. season (4) (laba72.sta)

Continue... Y

Case	Y	Moving Averages	Ratios	Seasonal Factors	Adjusted Series	Smoothed Trend-c.	Irreg. Compon.
1	8,80000			63,4753	13,86366	13,45559	1,030327
2	13,50000			100,7384	13,40105	13,70591	,977757
3	18,90000	14,05000	134,5196	136,4323	13,85302	14,20655	,975115
4	15,00000	14,30000	104,8951	99,3540	15,09753	14,79562	1,020405
5	9,80000	14,92500	65,6616	63,4753	15,43908	15,36992	1,004499
6	16,00000	15,72500	101,7488	100,7384	15,88273	15,89231	,999397
7	22,10000	16,20000	136,4198	136,4323	16,19851	16,33243	,991800
8	16,90000	16,47500	102,5797	99,3540	17,00988	16,81367	1,011670
9	10,90000	16,92500	64,4018	63,4753	17,17203	17,21753	,997358
10	17,80000	17,50000	101,7143	100,7384	17,66953	17,63905	1,001728
11	24,40000	17,90000	136,3128	136,4323	17,88433	18,08692	,988799
12	18,50000	18,25000	101,3699	99,3540	18,62029	18,67847	,996885
13	12,30000	18,85000	65,2520	63,4753	19,37762	19,30423	1,003802
14	20,20000	19,70000	102,5381	100,7384	20,05194	19,84617	1,010369
15	27,80000	20,12500	138,1366	136,4323	20,37640	20,28239	1,004635
16	20,20000	20,42500	98,8984	99,3540	20,33134	20,80730	,977125
17	13,50000	21,15000	63,8298	63,4753	21,26811	21,66288	,981777
18	23,10000	22,47500	102,7809	100,7384	22,93069	22,46936	1,020531
19	33,10000	22,90000	144,5415	136,4323	24,26111	22,84231	1,062113
20	21,90000	22,95000	95,4248	99,3540	22,04239	22,73006	,969746
21	13,70000	23,17500	59,1154	63,4753	21,58320	22,81089	,946180
22	24,00000	23,27500	103,1149	100,7384	23,82409	23,11482	1,030685
23	33,50000	23,32500	143,6227	136,4323	24,55430	23,54069	1,043058
24	22,10000			99,3540	22,24369	23,75363	,936433

Рис. 7.18. Результаты сезонной декомпозиции

делением элементов исходного ряда (Y) на сезонные индексы и умножением результата на 100).

В шестом столбце вычисляется сглаженная тренд-циклическая составляющая (Smoothed Trend-c.), т. е. приводятся результаты сглаживания ряда скорректированного на сезонные индексы. Сглаживание выполняется с помощью процедур скользящего среднего по пяти точкам с весами 1, 2, 3, 2, 1. Например, значение в третьей точке вычисляется по формуле

$$\frac{1 \cdot 13{,}86 + 2 \cdot 13{,}40 + 3 \cdot 13{,}85 + 2 \cdot 15{,}10 + 1 \cdot 15{,}44}{9} \approx 14{,}02.$$

Значения сглаженного ряда в первых двух точках и в последних двух точках вычисляются по специальным формулам.

В последнем, седьмом столбце, приводится остаточная (случайная) компонента ряда (Irreg. Compon.). Остаточная компонента вычисляется делением значений скорректированного ряда (пятый столбец) на значение сглаженного ряда (шестой столбец).

Таким образом результаты расчетов сезонных индексов и компонент временного ряда выполненных по методу приведенному в п. 7.3 и в пакете STATISTICA будут незначительно отличаться.

7.6.2. Работа 2. Прогнозирование по тренду и сезонной составляющей. Прогнозирование временного ряда методом экспоненциального сглаживания

1. Основные понятия

Вычисление ошибок прогноза. Прогнозирование методом экспоненциального сглаживания.

2. Задание

Используя выборку данных из своего варианта (24 точки), выполнить следующие задания:

1. По выбранной модели найти значения прогноза для 25, 26, 27 и 28 точек (см. п. 7.3.1). Для прогноза использовать оценки сезонных индексов, полученных в работе 1. Для оценки тренда можно, в зависимости от данных, вычислить оценки параметров линейной ($u_t = \beta_0 + \beta_1 t$) или квадратичной функции тренда ($u_t = \beta_0 + \beta_1 t + \beta_2 t^2$) по данным скользящих средних по четырем точкам.

2. Вычислить среднеквадратическую (MSE) и среднюю абсолютную относительную (MAPE) ошибки прогноза.

Экспоненциальное сглаживание в пакете STATISTICA.

Нажмите кнопку **Exponential Smoothing & Forecasting — Экспоненциальное сглаживание и прогнозирование** на стартовой панели модуля **Time Series Analysis** (рис. 7.11).

На экране появится стартовая панель диалога **Seasonal and Non-Seasonal Exponential Smoothing — Сезонное и несезонное экспоненциальное сглаживание** (рис. 7.19):

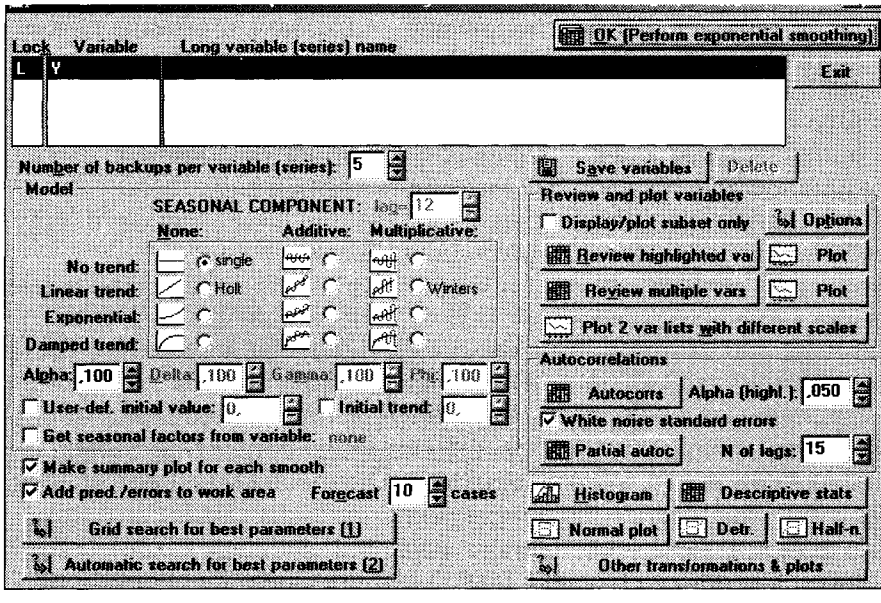


Рис. 7.19. Панель установки параметров **экспоненциального сглаживания**

Группа опций **Model — Модель,** позволяет, исходя из особенностей ряда, определить модель экспоненциального сглаживания. Для определения модели нужно задать *сезонную компоненту, тренд и параметры сглаживания.* Это можно сделать в следующих опциях:

- **Seasonal component** — Сезонная компонента;
- **None** — Нет;
- **Additive** — Аддитивная;
- **Multiplicative** — Мультипликативная;
- **No trend** — Нет тренда;
- **Linear trend** — Линейный тренд;
- **Exponential** — Экспоненциальный тренд
- **Damped trend** — Демпфированный (затухающий) тренд.

В полях **Alpha, Delta, Gamma, Phi** задаются параметры экспоненциального сглаживания.

Параметр **Alpha** необходим для всех моделей экспоненциального сглаживания. Остальные параметры нужны для специальных моделей. Параметр **Delta** — сезонный сглаживающий параметр, необходим лишь в сезонных моделях. Параметры **Gamma и Phi** являются параметрами *сглаживания тренда. Рассмотрим оставшиеся опции*:

- **User-def. initial value** — Определяемое пользователем начальное значение. Можно задать начальное значение сглаженного ряда;
- **Initial trend** — Начальный тренд. Можно задать начальное значение тренда. Если опция не используется, то начальное значение тренда оценивается;
- **Get seasonal factors from variables** — Оценить сезонные индексы по данным.

Следующие две опции относятся к представлению результатов на графиках:

- **Make summary plot for each smooth** — Сделать итоговый график для каждого сглаживания;
- **Add pred/errors to work area** — Добавить сглаженный ряд/остатки в рабочую область.

В опции **Forecast cases — Прогноз наблюдений** указывается, на сколько значений вперед будет прогнозироваться исходный ряд.

Простое экспоненциальное сглаживание без учета тренда и сезонной составляющей выполняется при установке **Single**.

В качестве примера рассмотрим процедуру экспоненциального сглаживания для данных из работы 1: 8,8; 15,5; 18,9; 15; 9,8; ...

Рассматриваемый ряд содержит линейный тренд и сезонную составляющую, поэтому для прогноза нужно использовать модель экспоненциального сглаживания по Винтеру (**Winters**). Для того чтобы определить параметры α, δ, γ можно использовать либо поиск по сетке (**Grid Search...**) либо автоматический поиск (**Autom. Search...**). Оптимальные значения параметров $\alpha = 0{,}662$, $\delta = 1{,}0$ и $\gamma = 0{,}0$ определялись автоматическим поиском.

Средняя квадратическая ошибка прогноза составляет 0,844 (рис. 7.21). Результаты прогноза на графике можно считать очень хорошими (рис. 7.20).

Более подробно о процедуре экспоненциального сглаживания см. [25, 17].

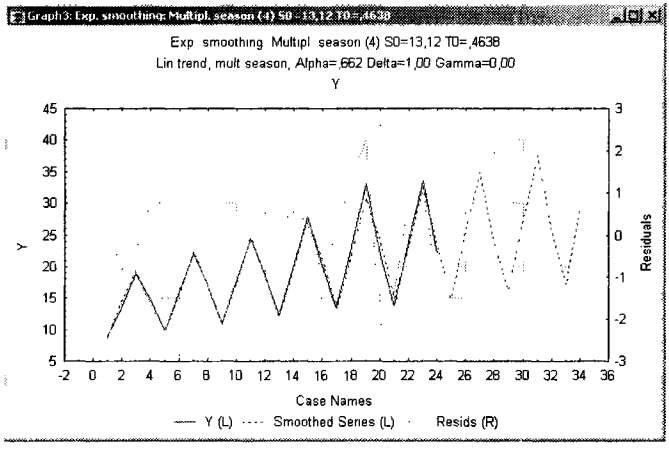


Рис. 7.20. Результаты экспоненциального сглаживания с использованием модели Винтера

Exp. smoothing: Multipl. season (4) S0=13,12 T0=,4638

TIME	Lin.trend, mult.season; Alpha=,662 Delta=1,00 Gamma=0,00
SERIES	Y
Summary of error	Error
Mean error	,0143141836212
Mean absolute error	,6754858278553
Sums of squares	20,2717894705036
Mean square	,8446578946043
Mean percentage error	-,2917919808520
Mean abs. perc. error	3,3473227425123

Рис. 7.21. Ошибки прогноза

7.7. Задачи для самостоятельного решения

1. Количество персональных компьютеров в Университете за последние 6 лет составило соответственно:

Год	1	2	3	4	5	6
Количество PC	50	110	350	1,020	1,950	3,710

(а) Постройте график по этим данным.

(б) Найдите уравнение линейного и квадратичного трендов.

(в) Оцените количество PC, которое будет использоваться в Университете на седьмой год, используя уравнения линейного и квадратичного трендов.

2. Следующая таблица описывает изменение почтовых тарифов в течение одиннадцати лет.

Год	1	2	3	4	5	6	7	8	9	10	11
Тариф (коп.)	5	5	8	8	10	13	15	18	20	22	25

(а) Постройте график по этим данным.
(б) Найдите уравнение линейного и квадратичного трендов.

3. Компания, специализирующаяся на производстве очистительных устройств, зафиксировала следующий объем продаж за последние 9 лет.

Год	1	2	3	4	5	6	7	8	9
Количество (×1 000 000 руб.)	13	15	19	21	27	35	47	49	57

(а) Постройте график по этим данным.
(б) Найдите уравнение линейного и квадратичного трендов.
(в) Какая из оценок тренда будет наиболее точной?

4. Хозяин фирмы по производству лодок составил следующую таблицу, содержащую квартальные доходы за последние 5 лет (×100 000 руб.):

Год	Весна	Лето	Осень	Зима
1	102	120	90	78
2	110	126	95	83
3	111	128	97	86
4	115	135	103	91
5	122	144	110	98

(а) Рассчитать 4-квартальные центрированные скользящие средние.
(б) Вычислить отношение (в процентах) исходных данных к скользящему среднему.
(в) Найти сезонные индексы для каждого квартала.

5. Президент фармацевтической компании, выписала следующие процентные ставки за каждый квартал в течение 4 лет:

Год	Весна	Лето	Осень	Зима
1	5.6	6.8	6.3	5.2
2	5.7	6.7	6.4	5.4
3	5.3	6.6	6.1	5.1
4	5.4	6.9	6.2	5.3

(а) Вычислить 4-квартальные центрированные скользящие средние и отобразить их на графике по текущим датам.
(б) Найти сезонные индексы для каждого квартала. Какой вывод вы можете сделать по полученным результатам?

6. Федеральный Заповедник вывел процентное отношение текущей суммы к 4-квартальному центрированному скользящему среднему значению, исходя из поквартальных счетов за 4-х годичный период:

Год	Весна	Лето	Осень	Зима
1	87	106	86	125
2	85	110	83	127
3	84	105	87	128
4	88	104	88	124

Вычислите сезонные индексы за каждый квартал. Сравните их с индексами, вычисленными в задании 5.

7. Производитель автомобильных ключей вывел следующие процентные отношения данных к 4-квартальному центрированному скользящему среднему значению в расходах фирмы в каждом квартале за последние 6 лет:

Год	Весна	Лето	Осень	Зима
1	108	128	94	70
2	112	132	88	68
3	109	134	84	73
4	110	131	90	69
5	108	135	89	68
6	106	129	93	72

Вычислите сезонные индексы за каждый квартал.

8. Заведующий учебным отделом в университете составил следующую таблицу посещаемости занятий студентами за последние 5 лет:

Год	Осень	Зима	Весна	Лето
1	220	203	193	84
2	235	208	206	76
3	236	206	209	73
4	241	215	206	92
5	239	221	213	115

Определите сезонные индексы.

9. Горный курорт посещало следующее количество гостей в течение каждого сезона за последние 5 лет. Вычислите сезонный индекс для каждого квартала. Если летом на курорте работало 50 человек, сколько человек нужно нанять зимой?

Год	Весна	Лето	Осень	Зима
1	200	300	125	325
2	175	250	150	375
3	225	300	200	450
4	200	350	225	375
5	175	300	200	350

10. Строительная компания собрала цифры о количестве домов, строительство которых было начато в каждом квартале за последние 5 лет.

Год	Весна	Лето	Осень	Зима
1	8	10	7	5
2	9	10	7	6
3	10	11	7	6
4	10	12	8	7
5	11	13	9	8

(а) Вычислите сезонный индекс для каждого квартала.

(б) Если бы объем капитала компании непосредственно зависел от количества начатых домов, на сколько процентов он должен был бы уменьшаться с лета до зимы?

11. Комиссия определила расход энергии, исходя из следующего ежеквартального расхода натурального газа, в млн $м^3$:

Год	Зима	Весна	Лето	Осень
1	293	246	231	282
2	301	252	227	291
3	304	259	239	296
4	306	265	240	300

(а) Определите сезонные индексы и исключите из данных сезонную составляющую.

(б) Методом наименьших квадратов вычислите параметры линейного тренда.

(в) Постройте графики текущих данных, данных без сезонной составляющей и без тренда.

12. Положение на рынке местного производителя пива представляют следующие цифры:

Продажи за квартал

Год	I	II	III	IV
1	19	24	38	25
2	21	28	44	23
3	23	31	41	23
4	24	35	48	21

(а) Вычислите сезонные индексы для этих данных (используйте центрированные средние значения за 4 квартала).

(б) Исключите сезонную составляющую из данных.

13. Используя данные из задачи 12:

(а) Методом наименьших квадратов найдите параметры прямой, которая наилучшим способом характеризует основную тенденцию временного ряда в данных о продаже пива.

(б) Определите циклическую компоненту в этом временном ряду, исключив тренд из исходных данных.

Глава 8
КЛАСТЕРНЫЙ АНАЛИЗ

8.1. Основные понятия

Пусть X_1, X_2, ..., X_n — исходная совокупность объектов, каждый из которых задан набором p признаков. Например, объектами могут быть пациенты клиники, а признаками — физические данные (вес, давление и т. д.) и результаты амбулаторного обследования каждого пациента (содержание сахара в крови, уровень гемоглобина и т. д.)

Задача кластерного анализа состоит в разбиении исходной совокупности объектов на группы схожих, близких между собой объектов. Эти группы называют кластерами или таксонами.

Другими словами, кластерный анализ это один из способов классификации объектов по их признакам. Желательно, чтобы результаты классификации имели содержательную интерпретацию.

Результаты, полученные методами кластерного анализа применяются в самых разнообразных областях. Например, в области медицины кластеризация заболеваний и симптомов заболеваний приводит к классификациям используемым для выбора методов лечения. Кластерный анализ широко применяется в маркетинговых исследованиях. В общем, всякий раз, когда необходимо классифицировать большое количество информации такого рода и представить ее в виде пригодном для дальнейшей обработки кластерный анализ оказывается весьма полезным и эффективным.

Фактически, кластерный анализ является «набором» различных *алгоритмов* «распределения объектов по кластерам».

В настоящее время известно огромное число алгоритмов кластеризации. Их разнообразие объясняется не только разными вычислительными методами, но и различными концепциями, лежащими в основе кластеризации [2, 27].

Одна из концепций состоит в построении разбиения исходного множества объектов доставляющего оптимальное значение определенной целевой функции. Большая группа методов кластеризации использует в качестве целевой функции внутригрупповую сумму квадратов: разбиение каждого множества должно быть таково, чтобы оно минимизировало внутригрупповые суммы квадратов. Эти методы используют евклидову метрику и называются методами минимальной дисперсии.

Большинство алгоритмов кластеризации основано на использовании эвристических методов. Дать рекомендации для выбора того или иного ме-

тода кластеризации можно только в общих чертах, а основной критерий выбора — практическая полезность результата.

Пусть X_1, X_2, ..., X_n — объекты, каждый из которых задан набором p признаков. Распределения объектов по кластерам на однородные в некотором смысле группы должно удовлетворять критерию оптимальности, который выражается в терминах расстояния $\rho(X_i,X_j)$ между любой парой объектов рассматриваемой совокупности.

В качестве расстояния (метрики) может быть взята любая неотрицательная действительная функция $\rho(X_i, X_j)$, определенная на множестве X_1, X_2, ..., X_n и удовлетворяющая следующим условиям:

а) $\rho(X_i, X_j) = 0$ тогда и только тогда, когда $X_i = X_j$;

б) $\rho(X_i, X_j) = \rho(X_j, X_i)$;

в) $\rho(X_i, X_j) \le \rho(X_i, X_k) + \rho(X_k, X_j)$.

Выбор расстояния между объектами неоднозначен и в этом состоит основная сложность.

Наиболее популярной метрикой является евклидова. Эта метрика отвечает интуитивным представлениям близости. При этом на расстояние между объектами могут сильно влиять изменения масштабов (единиц измерения) по осям. Например, если один из признаков измерен в метрах, а затем его значение переведены в сантиметры (т. е. умножены на 100), то евклидово расстояние между объектами сильно изменится и это приведет к тому, что результаты кластерного анализа могут значительно отличаться от предыдущих.

Если признаки измерены в разных единицах измерения, то требуется их предварительная нормировка — такое преобразование исходных данных, которое переводит их в безразмерные величины.

Наиболее известные способы нормировки следующие:

$$z_1 = \frac{x - \bar{x}}{\sigma}, \quad z_2 = \frac{x}{\bar{x}}, \quad z_3 = \frac{x}{x'}, \quad z_4 = \frac{x}{x_{\max}}, \quad z_5 = \frac{x - \bar{x}}{x_{\max} - x_{\min}},$$

где z_i, $i = 1, 2, ..., 5$ — нормированное значение;

x — исходное значение, $\bar{x}$ и σ — соответственно среднее и среднее квадратическое отклонение x, x' — эталонное (нормативное) значение, $x_{\max}$ и $x_{\min}$ — наибольшее и наименьшее значение x.

В пакете STATISTICA нормировка любой переменной выполняется по формуле $\frac{x - \bar{x}}{\sigma}$. Для этого нужно щелкнуть правой кнопкой мыши на имени переменной и в открывшемся меню выбрать: **Fill/Standardize Block** → **Standardize Columns**.

Нормировка, особенно по формуле $\frac{x - \bar{x}}{\sigma}$, сильно искажает геометрию исходного пространства, что может изменить результаты кластеризации.

Выбор метрики для каждой задачи должен производиться с учетом целей кластеризации, свойств признаков анализируемых объектов, вероятностной структуры данных и т. п. (см., например, [27]).

Наиболее употребительные метрики следующие (в скобках указано английское обозначение некоторых метрик, используемых в пакете STATISTICA в опции ***Distance measure***).

1. Евклидова метрика (***Euclidean distance***):

$$\rho_E(X_i, X_j) = \left[\sum_{k=1}^{p}(X_{ki} - X_{kj})^2\right]^{\frac{1}{2}},$$

где X_{ki} — значение k-го признака i-го объекта.

2. «Взвешенная» евклидова метрика:

$$\rho_{вE}(X_i, X_j) = \left[\sum_{k=1}^{p} W_k (X_{ki} - X_{kj})^2\right]^{\frac{1}{2}},$$

где W_k — «вес» k-го признака. Применяется в тех случаях, когда каждому признаку можно приписать «вес», пропорциональный степени важности данного признака в задаче классификации. Цель «взвешивания» признака состоит в том, чтобы обеспечить максимальную дискриминирующую способность признака для разделения на кластеры.

3. l_m-нормы:

$$\rho_m(X_i, X_j) = \left[\sum_{k=1}^{p}|X_{ki} - X_{kj}|^m\right]^{\frac{1}{m}}.$$

В частности, при $m = 1$ получаем меру l_1 — манхэттоновское расстояние или расстояние городских кварталов (***City-block*** (***Manhatten***) ***distance***).

Это расстояние является просто средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для евклидова расстояния. Однако отметим, что для этой метрики влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).

4. Супремум — норма (расстояние Чебышева (***Chebychev distance metric***))

$$\rho_\infty(X_i, X_j) = \sup_{k=1,2,\ldots,p}\{|X_{ki} - X_{kj}|\}.$$

Это расстояние может оказаться полезным, если желают определить два объекта как «различные», если они различаются по какому-либо одному признаку.

5. Степенное расстояние (***Power***):

$$\rho_p(X_i, X_j) = \left\{\sum_{k=1}^{p}(X_{ki} - X_{kj})^m\right\}^{\frac{1}{r}}.$$

Параметры m и r задаются пользователем. При $m = r = 2$ ρ_p совпадает с обычной евклидовой метрикой ρ_E.

6. Хеммингово расстояние

$$\rho_H(X_i, X_j) = \frac{\text{число случаев: } X_{ki} \neq X_{kj}}{p}.$$

ρ_H используется для признаков измеряемых в номинальной шкале и принимающих два значения. В пакете STATISTICA используется связанная с ρ_H метрика: процент несогласия (***Percent disagreement***).

7. Метрика Махаланобиса, определяемая формулой

$$\rho_0(X_i, X_j) = [(X_i - X_j)^T \Lambda^T \sum_X^{-1} \Lambda (X_i - X_j)]^{\frac{1}{2}},$$

где $\sum_x$ — ковариационная матрица генеральной совокупности, из которой извлекаются объекты X_i и X_j; Λ — симметричная неотрицательно-определенная матрица весовых коэффициентов, выбираемая обычно диагональной.

8. Коэффициент корреляции Пирсона (Pearson *r*):

$$\rho_k = \frac{\sum_{k=1}^{p}(X_{ki} - \overline{X}_i)(X_{kj} - \overline{X}_j)}{\sqrt{\sum_{k=1}^{p}(X_{ki} - \overline{X}_i)^2 \cdot \sum_{k=1}^{p}(X_{kj} - \overline{X}_j)^2}},$$

где $\overline{X}_i = \frac{1}{p}\sum_{k=1}^{p} X_{ki}$; $\overline{X}_j = \frac{1}{p}\sum_{k=1}^{p} X_{kj}$.

Метрика ρ_k не удовлетворяет условиям метрики (проверьте!), но активно используется в статистических исследованиях как мера взаимозависимости двух объектов.

Процедуры классификации на основе методов кластерного анализа используют расстояния между множествами объектов. Эти расстояния можно ввести различными способами. Пусть S_i — i-й класс (группа, кластер), n_i — число элементов в i-м классе, $\overline{X}(i)$ — «центр тяжести» i-го класса. Компоненты вектора $\overline{X}(i)$ вычисляются по формуле

$$\overline{X}_j(i) = \frac{1}{n_i}\sum_{k=1}^{n_i} X_{jk},\ j = 1, 2, \ldots, p.$$

Наиболее употребительные меры расстояния между классами следующие.

1. Расстояние, измеряемое по принципу «ближайшего соседа»

$$\rho_{\min}(S_l, S_m) = \min_{X_i \in S_l;\ Y_j \in S_m} \rho(X_i, Y_j).$$

В этом методе расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. В результате кластеры представляют длинные «цепочки».

2. Расстояние, измеряемое по принципу «дальнего соседа»:

$$\rho_{\max}(S_l, S_m) = \max_{X_i \in S_l;\ Y_j \in S_m} \rho(X_i, Y_j).$$

Расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т. е. «наиболее удаленными соседями»). Метод обычно работает очень хорошо, если кластеры не имеют удлиненную форму.

3. Расстояние, измеряемое по «центрам тяжести» классов

$$\rho(S_l, S_m) = \rho(\bar{X}(l), \bar{Y}(m)).$$

4. Расстояние, измеряемое по принципу «средней связи»

$$\rho_{\text{ср}}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{Y_j \in S_m} \rho(X_i, Y_j).$$

5. Обобщенное расстояние (по А. Н. Колмогорову):

$$\rho_r(S_l, S_m) = \left[\frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{Y_j \in S_m} \rho^r(X_i, Y_j) \right]^{\frac{1}{r}}.$$

Обобщенное расстояние ρ_r при $r = 1$, $\rho_1(S_l, S_m) = \rho_{\text{ср}}(S_l, S_m)$ называется средним расстоянием между классами S_l и S_m, соответствующим данной метрике ρ.

Можно показать, что при $r \to \infty$, $\rho_r = \rho_{\max}(S_l, S_m)$, а при $r \to -\infty$, $\rho_r = \rho_{\min}(S_l, S_m)$.

В процедурах кластеризации, использующих последовательное объединение элементов и классов, применяется следующая формула для пересчета расстояния между классом S_l и классом $S_{m,q} = S_m \cup S_q$, являющимся объединением двух классов S_m и S_q:

$$\rho_r(S_l, S_{m,q}) = \left\{ \frac{n_m [\rho_r(S_l, S_m)]^r + n_q [\rho_r(S_l, S_q)]^r}{n_m + n_q} \right\}^{\frac{1}{r}},$$

где n_m и n_q — число элементов соответственно в классах S_m и S_q.

С этой же целью используют также следующую формулу

$$\rho(S_l, S_{m,q}) = \alpha \rho_{lm} + \beta \rho_{lq} + \gamma \rho_{mq} + \delta |\rho_{lm} - \rho_{lq}|, \qquad (1)$$

где α, β, γ, и δ — числовые коэффициенты, значения которых определяет выбор той или иной меры расстояния между классами. Например, при $\alpha = \beta = -\delta = {}^1/_2$, $\gamma = 0$ — расстояние определяется по принципу ближайшего соседа; при $\alpha = \beta = \delta = {}^1/_2$, $\gamma = 0$ — расстояние определяется по принципу дальнего соседа; при $\alpha = n_m/(n_m + n_q)$, $\beta = n_q/(n_m + n_q)$, $\gamma = \delta = 0$, получим расстояние между классами, определяемое как среднее из расстояний между всеми парами элементов, из которых один берется из одного класса, а второй из другого класса.

6. Статистическое расстояние между классами

$$\rho_S(S_l, S_m) = \frac{n_l n_m}{n_l + n_m} (\bar{X} - \bar{Y})^T (\bar{X} - \bar{Y}).$$

Использование меры ρ_S обосновывается следующим образом. Рассмотрим класс, содержащий n элементов: X_1, X_2, ..., X_n, причем размерность каждого элемента равна p, а «центр тяжести» класса равен $\bar{X}$.

В качестве меры рассеяния элементов используют матрицу рассеяния

$$S_X = \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T.$$

След $\mathrm{tr}S_x$ матрицы S_x (т. е. сумму диагональных элементов) называют статистическим рассеянием множества элементов X_1, X_2, ..., X_n или внутригрупповой суммой квадратов

$$\mathrm{tr}S_X = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{p} (X_{ki} - \bar{X}_k)^2 = \sum_{i=1}^{n} (X_i - \bar{X})^T (X_i - \bar{X}).$$

Можно показать, что

$$\mathrm{tr}S_X = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^T (X_i - X_j) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_E^2 (X_i, X_j),$$

где суммирование проводится для значений $i < j$.

Таким образом, использование $\mathrm{tr}S_x$ в качестве меры рассеяния класса связано с евклидовой метрикой. Определитель матрицы рассеяния $|S_x|$ называют обобщенной дисперсией множества элементов X_1, X_2, ..., X_n.

Рассмотрим два класса S_l и S_m и их объединение $S_l \cup S_m = S(l, m)$.

Тогда матрица рассеяния для класса $S(l, m)$ равна

$$S_X(l,m) = \sum_{i=1}^{n_l} (X_i - M)(X_i - M)^T + \sum_{j=1}^{n_m} (Y_j - M)(Y_j - M)^T,$$

где

$X_i \in S_l$, $i = 1, 2, ..., n_l$;

$Y_j \in S_m$, $j = 1, 2, ..., n_m$;

$$M = \frac{1}{n_l + n_m} \left(\sum_{i=1}^{n_l} X_i + \sum_{j=1}^{n_m} Y_j \right) = \frac{1}{n_l + n_m} (n_l \bar{X} + n_m \bar{Y}).$$

Матрицу рассеяния S_X для класса $S(l, m)$ можно преобразовать так

$$S_X(l,m) = S_X(l) + S_X(m) + \frac{n_l n_m}{n_l + n_m} (\bar{X} - \bar{Y})(\bar{X} - \bar{Y})^T.$$

Матрица $\frac{n_l n_m}{n_l + n_m} (\bar{X} - \bar{Y})(\bar{X} - \bar{Y})^T$ называется матрицей межгруппового рассеяния, а след этой матрицы есть статистическое расстояние между классами S_l и S_m:

$$\mathrm{tr} \frac{n_l n_m}{n_l + n_m} (\bar{X} - \bar{Y})(\bar{X} - \bar{Y})^T = \frac{n_l n_m}{n_l + n_m} (\bar{X} - \bar{Y})^T (\bar{X} - \bar{Y}) = \rho_S(S_l, S_m).$$

Меры расстояния между классами могут быть также построены на основе вероятностных распределений каждого из классов.

8.2. Методы кластерного анализа в пакете STATISTICA

В модуле ***Cluster Analysis*** пакета STATISTICA реализуются следующие методы кластеризации:

- соединения (древовидная кластеризация), ***Joining*** (***tree clustering***);
- метод *K*-средних (***K-means clustering***);
- двухвходовое объединение (***Two-way joining***).

8.2.1. Иерархические алгоритмы

Первая опция (***Joining***) представляет группу так называемых иерархических алгоритмов кластеризации. В основе этих алгоритмов лежит идея последовательной кластеризации. Пусть исходное множество содержит n объектов X_1, X_2, X_3, ..., X_n.

В качестве расстояния между объектами X_i и X_j выбирается некоторая метрика ρ. Выбор метрики необходимо сделать в опции *distance measure* панели *Joining*.

На начальном шаге каждый объект рассматривается как отдельный кластер. На следующем шаге некоторые из ближайших друг к другу кластеров будут объединяться в один новый кластер. В зависимости от выбора меры, по которой определяется расстояние между кластерами, реализуются следующие методы объединения объектов в кластеры (выбор осуществляется в зависимости от меры расстояния между кластерами в опции: ***Amalgamation*** (***linkage***) ***rule***).

1. **Метод одиночной связи** (***Single Linkage***). Кластеры объединяются исходя из расстояния, измеряемого по методу «ближайшего соседа». Группы, между которыми расстояния самые маленькие, объединяются. Каждое объединение уменьшает число групп на единицу. Расстояние между группами определяется как расстояние между ближайшими членами групп. Метод приводит к «цепным» кластерам.

2. **Метод полной связи** (***Complete Linkage***). Расстояние между группами определяется как расстояние измеряемое по принципу «дальнего соседа». Расстояние между объединяемыми кластерами равно диаметру наименьшей сферы, содержащей оба кластера. Метод создает компактные кластеры в виде гиперсфер, которые плохо объединяются с другими кластерами. Если кластеры имеют удлиненную форму, то метод не работает.

3. **Метод невзвешенного попарного среднего** (***Unweighted pair-group average***). Расстояние между кластерами определяется по принципу «средней связи».

4. **Метод взвешенного попарного среднего** (***Weighted pair-group average***). Расстояние между кластерами определяется по принципу «средней связи», но с учетом в качестве весов числа объектов, содержащихся в кластерах.

5. **Невзвешенный центроидный метод** (***Unweighted pair-group centroid***). Расстояния между кластерами определяется как расстояние между их «центрами тяжести»

$$\rho(S_l, S_m) = \rho(\overline{X}, \overline{Y}).$$

6. **Взвешенный центроидный метод (*Weighted pair-group centroid*).** Расстояние между классами определяется как расстояние между их «центрами тяжести», но с учетом весов, определяемых по количеству объектов в каждом кластере (т. е. с учетом размеров кластеров).

7. **Метод Уорда (Ward's metod).** В этом методе в качестве целевой функции используется сумма квадратов расстояний между каждым элементом и «центром тяжести» класса, содержащего этот элемент. Кластеризация представляет последовательную процедуру, на каждом шаге которой объединяются два таких класса, при объединении которых происходит минимизация статистического расстояния между классами ρ_s, вычисляемого по формуле

$$\rho_s = \frac{n_l n_m}{n_l + n_m}(\bar{X} - \bar{Y})^T(\bar{X} - \bar{Y}).$$

Рассмотрим работу иерархического алгоритма кластеризации на простом примере.

Пример 8.1. Провести кластеризацию четырех объектов методами одиночной связи (*Single Linkage*) и полной связи (*Complete Linkage*). Каждый объект определяется двумя признаками:

Признак	Объект			
	1	2	3	4
x_i	0	–1	1	4
y_i	–2	0	2	0

Решение. Расстояние между объектами X_i и X_j определим как квадрат евклидовой метрики

$$\rho_{ij} = (x_i - x_j)^2 + (y_i - y_j)^2, \quad i, j = 1, 2, 3, 4; \quad j \neq i.$$

Таким образом, расстояние между первым и вторым объектами равно:

$$\rho_{12} = (0 + 1)^2 + (-2 - 0)^2 = 5,$$

а между первым и третьим объектами

$$\rho_{13} = (0 - 1)^2 + (-2 - 2)^2 = 17 \text{ и т. д.}$$

На **первом шаге** матрица расстояний между объектами D_1 имеет вид

$$D_1 = \begin{pmatrix} 0 & 5 & 17 & 20 \\ 5 & 0 & 8 & 25 \\ 17 & 8 & 0 & 13 \\ 20 & 25 & 13 & 0 \end{pmatrix}.$$

Наиболее близки первый и второй объекты: $\rho_{12} = 5$, следовательно, эти объекты объединяются в один кластер.

На **втором шаге** имеем следующие кластеры:

Кластеры	1	2	3
Объекты	(1, 2)	3	4

Определяем расстояние между кластерами по методу «ближайшего соседа»:

$$\rho_{12}^{(2)} = \rho_{(1,2),3} = \min(\rho_{13}; \rho_{23}) = \min(17; 8) = 8;$$

$$\rho_{13}^{(2)} = \rho_{(1,2),4} = \min(\rho_{14}; \rho_{24}) = \min(20; 25) = 20;$$

$$\rho_{23}^{(2)} = \rho_{3,4} = 13.$$

Для вычисления расстояния между кластерами можно воспользоваться формулой (1):

$$\rho(S_l, S_{(m,q)}) = \alpha\rho_{lm} + \beta\rho_{lq} + \gamma\rho_{mq} + \delta\,|\,\rho_{lm} - \rho_{lq}\,|.$$

Расстояние по принципу «ближайшего соседа» определяется при $\alpha = \beta = \frac{1}{2}$, $\gamma = 0$, $\delta = -\frac{1}{2}$. Таким образом, например, расстояние между первым и вторым кластерами $\rho_{12}^{(2)}$ по формуле (1) равно

$$\rho_{12}^{(2)} = \rho_{(1,2)3} = \frac{1}{2}\rho_{13} + \frac{1}{2}\rho_{23} - \frac{1}{2}|\rho_{12} - \rho_{23}| = \frac{1}{2}17 + \frac{1}{2}8 - \frac{1}{2}(17 - 8) = 8.$$

Матрица расстояний D_2 между тремя кластерами на втором шаге имеет вид

$$D_2 = \begin{pmatrix} 0 & 8 & 20 \\ 8 & 0 & 13 \\ 20 & 13 & 0 \end{pmatrix}.$$

Как следует из матрицы расстояний D_2 наиболее близки первый и второй кластеры: $\rho_{12}^{(2)} = 8$, следовательно, эти кластеры объединяются в один кластер.

Таким образом, на **третьем шаге** имеем следующие кластеры:

Номера кластеров	1	2
Состав кластеров (в скобках указаны номера кластеров на втором шаге)	(1, 2)	3
Состав кластеров (в скобках указаны номера исходных объектов)	(1, 2, 3)	4

Определяем расстояние между кластерами

$$\rho_{12}^{(3)} = \rho_{(1,2),3}^{(2)} = \min(\rho_{13}^{(2)}; \rho_{23}^{(2)}) = \min(20; 13) = 13.$$

Матрица расстояний D_3 между двумя кластерами на третьем шаге

$$D_3 = \begin{pmatrix} 0 & 13 \\ 13 & 0 \end{pmatrix}.$$

На последнем, **четвертом шаге**, оба кластера объединяются.

Теперь рассмотрим работу алгоритма, в случае, когда расстояние между кластерами определяется по принципу «дальней связи» – *Complete Linkage.*

На **первом шаге** объединяются наиболее близкие первый и второй объекты: $\rho_{12} = 5$.

На **втором шаге** имеем следующие кластеры:

Кластер	1	2	3
Объекты	(1, 2)	3	4

Далее определяем расстояние между объектами по принципу «дальней связи»:

$$\rho_{12}^{(2)} = \rho_{(1,2),3} = \max(\rho_{13}; \rho_{23}) = \max(17; 8) = 17;$$

$$\rho_{13}^{(2)} = \rho_{(1,2),4} = \max(\rho_{14}; \rho_{24}) = \max(20; 25) = 25;$$

$$\rho_{23}^{(2)} = \rho_{3,4} = 13.$$

Матрица расстояний D_2 между тремя кластерами на втором шаге имеет вид

$$D_2 = \begin{pmatrix} 0 & 17 & 25 \\ 17 & 0 & 13 \\ 25 & 13 & 0 \end{pmatrix}.$$

Таким образом наиболее близки второй и третий кластеры: $\rho_{32}^{(2)} = 13$. Эти кластеры объединяются и на **третьем шаге** имеем следующие кластеры:

Номера кластеров	1	2
Состав кластеров (в скобках указаны номера кластеров на 2-м шаге)	1	(2, 3)
Состав кластеров (в скобках указаны номера исходных объектов)	(1, 2)	(3, 4)

Определяем расстояние между кластерами

$$\rho_{12}^{(2)} = \rho_{1,(2,3)} = \max(\rho_{12}^{(2)}; \rho_{23}^{(2)}) = \max(17; 25) = 25.$$

Матрица расстояний

$$D_3 = \begin{pmatrix} 0 & 25 \\ 25 & 0 \end{pmatrix}.$$

На последнем, **четвертом шаге**, оба кластера объединяются.

8.2.2. Выполнение иерархических процедур в пакете STATISTICA

Для реализации любого метода кластеризации из группы иерархических процедур ***Joining*** (***tree clustering***) необходимо сделать следующие установки:

1) выбрать переменные для анализа (***Variables***);

2) определить вид входных данных (***Input***): можно вводить таблицу с координтами объектов (***Raw data***) либо сразу матрицу расстояний между объектами (***Distance matrix***);

3) определить объекты кластеризации: это могут быть переменные (столбцы) (***Variables*** (***columns***)) либо наблюдения (строки) — ***Cases*** (***rows***). В последнем случае каждая строка таблицы исходных данных есть объект;

4) выбрать метрику, определяющую расстояние между кластерами — ***Amalgamation*** (***linkage***) ***rule***;

5) выбрать метрику, определяющую расстояние между объектами — ***Distance measure***.

Результаты кластеризации имеют следующий вид:

1) строится горизонтальная или вертикальная дендрограмма — график, на котором определены расстояния между объектами и кластерами при их последовательном объединении. Древовидная структура графика позволяет определить кластеры в зависимости от выбранного порога — заданного расстояния между кластерами;

2) выводится матрица расстояний между исходными объектами (*Distance matrix*);

3) выводятся средние и среднеквадратичные отклонения для каждого исходного объекта (***Distiptive statistics***).

Рассмотрим решение **примера 8.1** в пакете STATISTICA.

Нажмите кнопку ***Module Switcher*** на панели инструментов, в появившемся окне выберете модуль ***Cluster Analysis***, а затем ***Joining*** (***tree clustering***). В новом окне выполните следующие настройки:

а) нажмите на кнопку ***Variables*** и введите имена двух переменных x и y, в которых записаны исходные данные примера 8.1;

б) в разделе ***Input*** введите ***Raw data*** (*исходные данные*);

в) в разделе ***Cluster*** выберите ***Cases*** (***rows***). При этой установке объекты кластеризации — двумерные наблюдения с координатами x_i и y_i, $i = 1, 2, 3, 4$;

г) в разделе ***Amalgamation*** (***linkage***) ***rule*** выберите ***Single Linkage*** (*метод одиночной связи*);

д) в разделе ***Distance measure*** выберите ***Squared Euclidean distances*** (*квадрат евклидовой метрики*) и нажмите OK.

В появившемся окне нажмите на кнопку ***Vertical icicle plot***. На экране появится дендрограмма (рис. 8.1), показывающая объединение объектов, расстояние между которыми является наименьшим, в кластеры.

На вертикальной оси дендрограммы откладываются расстояния между объектами и между объектами и кластерами. Так, расстояние между объектами C_1 и C_2 равно 5 (см. матрицу расстояний D_1 в примере 8.1). Эти объекты объединяются в один кластер на первом шаге.

Расстояние между этим кластером и объектом C_3 равно 8 (см. матрицу расстояний D_2). Объект C_3 объединяется с кластером (C_1, C_2) на втором шаге. Наконец, расстояние между объектом C_4 и кластером (C_1, C_2, C_3) равно 13 (см. матрицу расстояний D_3).

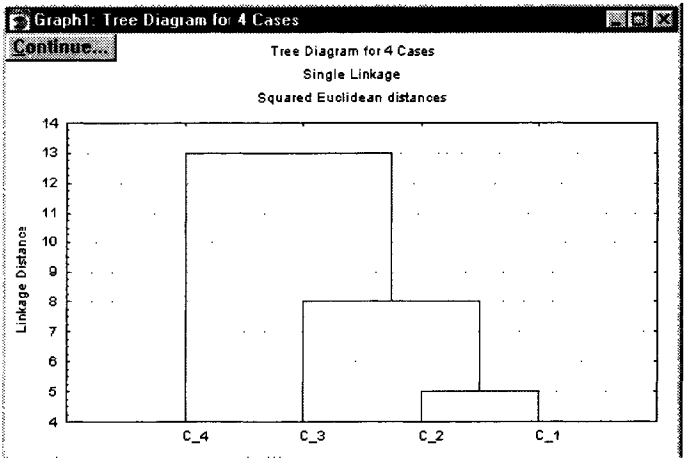


Рис. 8.1. Дендрограмма при методе одиночной связи

Таким образом, горизонтальные отрезки дендрограммы проводятся на уровнях, соответствующих пороговым значениям расстояний, выбираемым для данного шага кластеризации.

Кластеризация методом одиночной связи (ближайшего соседа) приводит к образованию одного кластера (пороговое расстояние равно 13).

Далее последовательно нажмите ***Continue...*** и ***Cancel*** и в окне установок процедуры в разделе ***Amalgamation...*** выберите ***Complete Linkage***. После выполнения процедуры появится следующая дендрограмма (рис. 8.2).

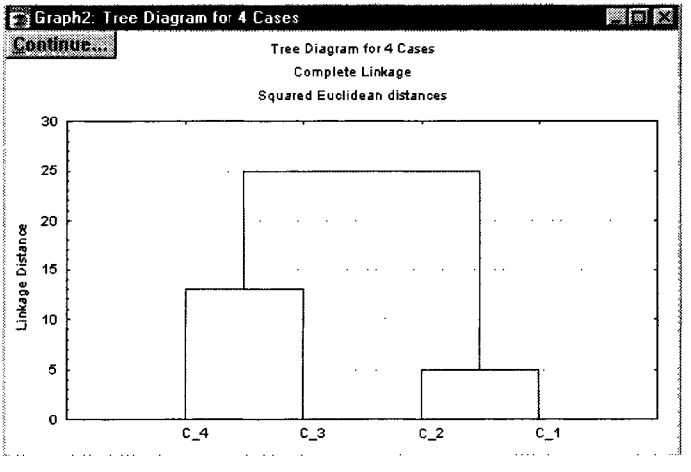


Рис. 8.2. Дендрограмма при методе полной связи

Вторая дендрогамма показывает, что кластеризация методом полной связи (дальнего соседа) при таком же пороговом расстоянии равным 13, приводит к образованию двух кластеров.

Сравните полученные дендрограммы с результатами решения примера 8.1.

8.2.3. Метод *K*-средних

Метод *K*-средних относится к группе так называемых эталонных методов кластерного анализа. Число кластеров *K* задается пользователем. Процедура состоит в следующем. На первом шаге определяют *K* кластеров — эталонов (это могут быть, например, первые *K* объектов). Далее каждый объект присоединяется к ближайшему эталону. В качестве критерия используется минимальное расстояние внутри кластера относительно среднего. Как только объект включается в кластер, среднее пересчитывается. После пересчета эталона объекты снова распределяются по ближайшим кластерам и т. д. Процедура заканчивается при стабилизации процесса, т. е. при стабилизации центров тяжести.

Пример 8.2. Провести классификацию $n = 10$ объектов, каждый из которых характеризуется тремя признаками: x, y и z. Таблица данных приведена на рис. 8.3.

Data: NEW.STA 10v * 10c

	X	Y	Z
1	,252	2,100	2,600
2	,322	2,700	2,600
3	,339	2,000	2,000
4	,406	1,900	1,000
5	,492	2,200	2,000
6	,507	2,300	2,000
7	,695	2,300	2,400
8	,751	2,100	2,100
9	,903	2,400	2,100
10	,971	2,400	2,500

Рис. 8.3. Данные для примера 8.2

Решение в пакете STATISTICA.

1. Визуализация данных (в трехмерном случае).

В меню ***Graphs*** выберите ***Stats 3D XYZ Graphs***. В выпадающем меню выберите команду ***Scatterplots***, в появившемся окне нажмите на кнопку ***Variables*** и задайте ***X***, ***Y***, ***Z*** (рис. 8.4). Затем нажмите на кнопку ***Options*** и в разделе ***Display*** включите ***Case Name*** (имена наблюдений), нажмите OK.

На экране появится диаграмма рассеяния для исходных данных (рис. 8.5). По диаграмме видно, что объекты образуют три кластера.

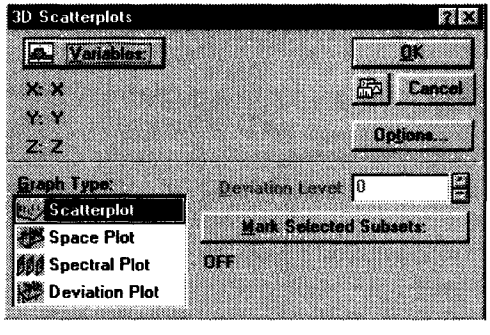


Рис. 8.4. Ввод данных для построения диаграммы рассеяния

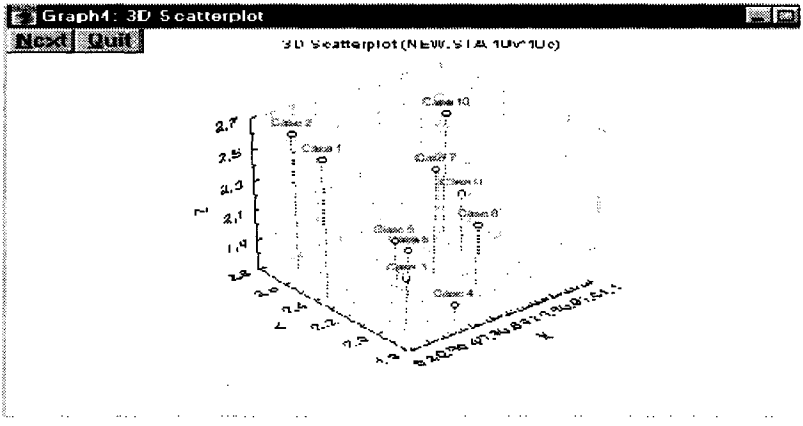


Рис. 8.5. Диаграмма рассеяния для данных примера 8.2

2. Проведем кластерный анализ с помощью метода *K*-средних (***K-means clustering***). На панели инструментов нажмите кнопку ***Module Switcher***, в появившемся окне выберите ***Cluster Analysis*** и в стартовой панели модуля выберете ***K-means clustering***. В новом окне выполните следующие настройки:

а) нажмите на кнопку ***Variables*** и введите переменные ***X***, ***Y***, ***Z***;

б) в разделе ***Cluster*** выберите ***Cases*** (***rows***);

в) в разделе ***Number of clusters*** задайте число кластеров, равное трем;

г) задайте число итераций;

д) выберете один из трех методов для начального определения центров кластеров (эталонов): либо выбираются первые *K* объектов, либо выбираются объекты наиболее отстоящие друг от друга, либо отстоящие друг от друга на одинаковом расстоянии.

После выбора установок нажмите OK.

3. Результаты кластеризации.

а. ***Analysis of variance*** — результаты дисперсионного анализа по каждому признаку ***X***, ***Y***, ***Z*** (рис. 8.6): выводятся суммы квадратов отклонения объектов от центров кластеров (***SS Within***) и суммы квадратов отклонений между центрами кластеров (***SS Between***), значения *F*-статистики и уровни значимости *p*.

Analysis of Variance (new.sta)

Continue...	Between SS	df	Within SS	df	F	signif p
X	,421231	2	,131911	7	11,17657	,006623
Y	,377333	2	,191667	7	6,89044	,022183
Z	,614333	2	,134667	7	15,96658	,002464

Рис. 8.6. Результаты дисперсионного анализа

В данном примере уровни значимости равны: 0,0066; 0,0222; 0,0024, т. е. по *X*, *Y* и *Z* гипотезы о равенстве средних для центров кластеров отклоняются на уровне значимости $\alpha > 0{,}0222$.

б. Выводятся координаты центров и матрицы расстояний между центрами (рис. 8.7).

Euclidean Distances between Clusters (new.sta)

Continue...

Distances below diagonal
Squared distances above diagonal

Cluster Number	No. 1	No. 2	No. 3
No. 1	0,000000	,117488	,212961
No. 2	,342765	0,000000	,136288
No. 3	,461477	,369171	0,000000

Рис. 8.7. Матрица расстояний между центрами кластеров

в. График распределения центров кластеров (рис. 8.8):

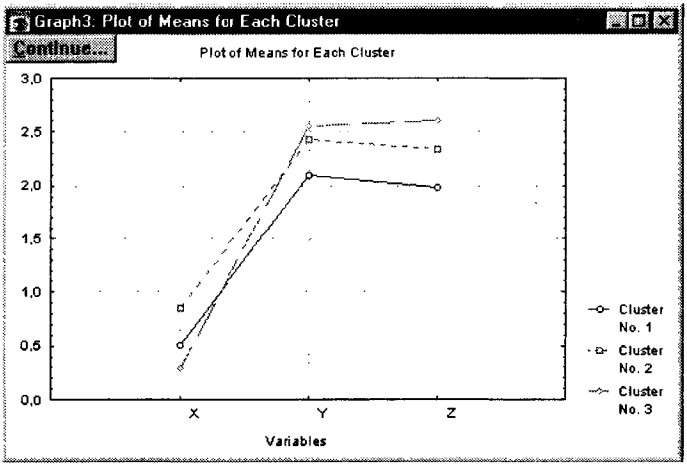


Рис. 8.8. График координат центров кластеров

г. Статистики для каждого кластера по координатам *X*, *Y*, *Z*: средние центров, стандартные отклонения и дисперсии.

д. Номера объектов, входящих в каждый кластер и расстояния объектов до центра каждого кластера.

В данном примере объекты распределились следующим образом:

кластер 1: {3, 4, 5, 6, 8};

кластер 2: {7, 9, 10};

кластер 3: {1, 2}.

8.2.4. Двухвходовое объединение

Кластеризация проводится одновременно как по переменным (столбцам), так и по результатам наблюдений (строкам). Для примера рассмотрим десять экономических показателей (производительность труда, индекс снижения себестоимости, рентабельность и т. д.) для каждого из 50 предприятий (Приложение 1.1). Нас могут интересовать однородные по этим показателям группы предприятий (кластеризация производится по предприятиям (строкам)), а также однородные группы экономических показателей (кластеризация проводится по переменным — столбцам). Процедура двухвходового объединения используется в тех случаях, когда можно ожи-

дать, что одновременная кластеризация по переменным (столбцам) и наблюдениям (строкам) дает возможность получить осмысленные кластеры. Результаты процедуры: описательные статистики по переменным и наблюдениям, а также двумерная цветная диаграмма (*Two-way joining graph*), на которой цветом отмечаются значения данных. По распределению цвета можно составить представление об однородных группах.

8.3. Задачи для самостоятельного решения

1. Используя иерархические алгоритмы проведите кластеризацию данных из примера 8.2. Сравните результаты с результатами в примере 8.2.

2. На предприятии существует 16 научно-производственных отделов, занятых выпуском различной продукции, работ, услуг. Поскольку виды деятельности, количество работающих, рентабельность отделов, существенно различаются между собой, б*ыло решено сгруппировать отделы в несколько однородных групп, а затем для каждой группы разработать свою систему премирования.*

После тщательного анализа выбрали четыре признака, с помощью которых описывались важные (для указанной цели) параметры каждого отдела:

X_1 — стоимость активной части основных производственных фондов, тыс. руб.;

X_2 — среднемесячный объем работ отдела, тыс. руб.;

X_3 — удельный вес работ/услуг отдела по внутрифирменной кооперации, %;

X_4 — среднемесячная прибыль отдела, тыс. руб.

Исходные данные по отделам приведены ниже.

№ отдела	Значения признаков			
	X_1	X_2	X_3	X_4
1	699	190	53	11
2	532	211	19	42
3	650	152	46	14
4	768	216	67	17
5	67	106	0	32
6	322	397	26	52
7	736	180	49	18
8	501	239	11	60
9	293	391	16	66
10	300	396	29	87

Продолжение исходных данных

№ отдела	Значения признаков			
	X_1	X_2	X_3	X_4
11	73	160	0	22
12	862	199	51	22
13	112	136	0	29
14	289	388	31	74
15	512	195	6	58
16	490	201	9	65

Проведите кластеризацию отделов используя иерархические алгоритмы (Joining):

а) используя исходные данные;

б) используя стандартизованные данные, т. е. данные, преобразованные по формуле

$$\frac{X_{ij} - \bar{X}_j}{S_j},$$

где

X_{ij} — i-е значение j-го признака, $i = 1, 2, ..., 16$; $j = 1, 2, 3, 4$;

$\bar{X}_j = \frac{1}{16}\sum_{i=1}^{16} x_{ij}$ — оценка среднего для j-го признака;

$S_j = \sqrt{\frac{1}{15}\sum_{i=1}^{16}(x_{ij} - \bar{x}_j)^2}$ — оценка среднего квадратического отклонения для j-го признака.

Процедуру стандартизации данных можно выполнить непосредственно в таблице, используя следующую последовательность действий: курсор на имени переменной → нажать правую кнопку мыши → в выпадающем меню выбрать **File/Standardize Block** → **Standardize Columns** → OK.

Сравните результаты кластеризации. По результатам кластеризации определите число кластеров и их состав. Найдите статистические характеристики каждого кластера.

Проведите кластеризацию используя метод K-средних (число кластеров задайте равным 4). Сравните результаты (составы кластеров).

3. Ниже приведены значения основных факторов сельскохозяйственного производства для 20 районов:

x_1 — число тракторов на 100 га;

x_2 — число зерноуборочных комбайнов на 100 га;

x_3 — число орудий поверхностной обработки почвы на 100 га;

x_4 — количество удобрений, расходуемых на гектар (т/га);

x_5 — количество химических средств защиты растений, расходуемых на гектар (ц/га).

Районы	Факторы				
	x_1	x_2	x_3	x_4	x_5
1	1,59	0,26	2,05	0,32	0,14
2	0,34	0,28	0,46	0,59	0,66
3	2,53	0,31	2,46	0,30	0,31
4	4,63	0,40	6,44	0,43	0,59
5	2,16	0,26	2,16	0,39	0,16
6	2,16	0,30	2,69	0,32	0,17
7	0,68	0,29	0,73	0,42	0,23
8	0,35	0,26	0,42	0,21	0,08
9	0,52	0,24	0,49	0,20	0,08
10	3,42	0,31	3,02	1,37	0,73
11	1,78	0,30	3,19	0,73	0,17
12	2,40	0,32	3,30	0,25	0,14
13	9,36	0,40	11,51	0,39	0,38
14	1,72	0,28	2,26	0,82	0,17
15	0,59	0,29	0,60	0,13	0,35
16	0,28	0,26	0,30	0,09	0,15
17	1,64	0,29	1,44	0,20	0,08
18	0,09	0,22	0,05	0,43	0,20
19	0,08	0,25	0,03	0,73	0,20
20	1,36	0,26	0,17	0,99	0,42

1) проведите кластеризацию районов используя несколько иерархических алгоритмов (Joining):

а) используя исходные данные;

б) используя стандартизованные данные;

2) проведите кластеризацию используя метод *K*-средних (число кластеров задайте равным 3). Сравните составы кластеров и их характеристики.

4. Проведите кластерный анализ данных из Приложения 1.1 табл. П2. Варианты заданий приведены в табл. П1.

5. Проведите кластерный анализ данных об однокомнатных квартирах (Приложение 1.2).

Используйте процедуры Joining (Complete Linkage) и метод *K*-средних (рассмотрите 6 кластеров).

Попробуйте интерпретировать полученные кластеры, используя категории качества и стоимости квартир.

Глава 9
РЕШЕНИЕ ЗАДАЧ ИССЛЕДОВАНИЯ ОПЕРАЦИЙ В EXCEL

Исследование операций — это раздел современной прикладной математики, ориентированный на построение, разработку и применение математических моделей принятия оптимальных решений. Область приложений для методов исследования операций весьма обширна: это инженерно-технические, технико-экономические, социально-экономические задачи, а также задачи управления в различных сферах.

Методы исследования операций часто рассматривают как инструмент для принятия оптимального решения. Нахождение оптимального решения во многих практически важных задачах предполагает анализ и выбор в определенном множестве допустимых решений элемента, удовлетворяющего тому или иному критерию оптимальности. Например, это может быть выбор такого допустимого плана производства продукции, при котором предприятие получит максимальную прибыль. Здесь «допустимый план» означает план, который может быть реально выполнен с учетом всех возможностей предприятия, т. е. с учетом ограничений на материальные, энергетические, людские и тому подобные ресурсы. В данном случае критерием оптимальности является достижение максимума целевой функции — дохода. Таким образом, рассматриваемая задача — это задача нахождения максимума целевой функции на множестве допустимых решений. Последнее определяется рядом ограничений на переменные задачи, которые задаются в виде неравенств и равенств. Задачи такого вида обычно называют *задачами распределения ресурсов*.

Аналогично определяется *транспортная задача*: необходимо найти допустимый план перевозок минимизирующий транспортные расходы.

Математические методы решения задач такого типа разработаны достаточно подробно. Это, прежде всего, методы линейного и нелинейного программирования [28, 31].

Цель настоящей главы состоит в том, чтобы на простых примерах показать, как задачи подобного вида могут быть решены на компьютере средствами пакета EXCEL.

В EXCEL имеется процедура «Поиск решения» (Solver), реализующая численные методы решения задач нелинейного программирования: метод Ньютона и метод сопряженных градиентов, а также симплекс-метод для решения задач линейного программирования. Использование этой процедуры для пользователя, который знаком с основами этих методов и умеющего правильно поставить задачу, не представляет труда.

Необходимо выполнить следующие действия.

1. Ввести данные и необходимые формулы в рабочий лист Excel.
2. Вызвать процедуру «Поиск решения» из меню *Сервис*. Окно процедуры приведено на рис. 9.10.
3. Указать адрес ячейки, содержащий формулу для вычисления целевой функции и задать вид оптимизации (максимум, минимум, заданное значение).
4. Указать адреса рабочих (изменяемых) ячеек, которые будут содержать значения переменных задачи (см. ниже, пример 9.1).
5. Задать ограничения, которым должны удовлетворять переменные.
6. Используя опцию *Параметры* задать численный метод для решения задачи и параметры этой процедуры.
7. Нажать кнопку *Выполнить*.

Замечание. Поиск решения — это надстройка Excel и она должна быть предварительно загружена. Чтобы загрузить эту процедуру нужно выполнить команду *Сервис* → *Надстройки*, установить флажок у надстройки *Поиск решения* и нажать OK.

В данной главе на простых примерах будут рассмотрены несколько задач линейного программирования. Достаточно подробно, на примере решения задачи распределения ресурсов, рассматриваются графический метод решения и симплекс-метод в алгебраическом и табличном вариантах. Это поможет заинтересованному читателю самостоятельно решать задачи из приведенных вариантов заданий, грамотно поставить и решить задачи с использованием процедуры *Поиск решения* (см. также [29]).

9.1. Методы решения задач линейного программирования (ЛП)

Пример 9.1. Задача распределения ресурсов. Предприятие изготавливает и продает краску двух видов: для внутренних и внешних работ. Для производства краски используется два исходных продукта A и B. Расходы продуктов A и B на 1 т соответствующих красок и запасы этих продуктов на складе приведены в таблице:

Исходный продукт	Расход продуктов (в тоннах на 1 т краски)		Запас продукта на складе (т)
	краска для внутренних работ	краска для внешних работ	
A	1	2	3
B	3	1	3

Цена за 1 т краски для внутренних работ составляет 2000 руб., краска для наружных работ продается по 1000 руб. за 1 т. Требуется определить какое количество краски каждого вида следует производить предприятию, чтобы получить **максимальный доход**.

Составление математической модели задачи.

1) *Переменные задачи.*

Обозначим: x_1 — количество производимой краски для внутренних работ; x_2 — соответствующее количество краски для наружных работ.

2) *Ограничения*, которым должны удовлетворять переменные задачи:

$x_1, x_2 \geq 0$;

по расходу продукта А: $x_1 + 2x_2 \leq 3$;

по расходу продукта В: $3x_1 + x_2 \leq 3$.

В левых частях последних двух неравенств определены расходы продуктов А и В, а в правых частях неравенств записаны запасы этих продуктов.

3) *Целевая функция задачи.*

Обозначим Z — доход от продажи краски (в тысячах рублей), тогда целевая функция задачи записывается так:

$$Z = 2x_1 + x_2.$$

Таким образом, задача состоит в том, чтобы найти $\max Z = 2x_1 + x_2$, при ограничениях:

$$x_1 + 2x_2 \leq 3, \quad \text{(A)}$$

$$3x_1 + x_2 \leq 3, \quad \text{(B)}$$

$$x_1, x_2 \geq 0.$$

Так как переменные задачи x_1 и x_2 входят в целевую функцию и ограничения задачи *линейно*, соответствующая задача оптимизации называется *задачей линейного программирования* (ЛП).

9.1.1. Графическое решение задачи ЛП

В рассматриваемом примере содержатся только две переменные x_1 и x_2, поэтому задачу можно решить графически.

1. На плоскости x_1Ox_2 строим область допустимых значений переменных, определяемую ограничениями задачи:

$$x_1 + 2x_2 \leq 3, \quad \text{(A)}$$

$$3x_1 + x_2 \leq 3, \quad \text{(B)}$$

$$x_1, x_2 \geq 0.$$

Последнее ограничение определяет первый квадрант плоскости. Чтобы построить множество точек удовлетворяющих неравенству (А) нанесем на плоскость график прямой, определяющий границу этого множества

$$x_1 + 2x_2 = 3. \quad \text{(A)}$$

Приведем это уравнение к виду: $\frac{x_1}{a} + \frac{x_2}{b} = 1$. Это уравнение прямой «в отрезках» и для построения этой прямой используются две точки $(a, 0)$ и $(0, b)$ (рис. 9.1).

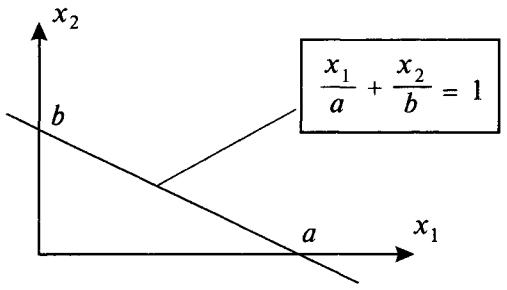


Рис. 9.1

Приведя уравнение (А) к виду прямой в отрезках, получим

$$\frac{x_1}{3} + \frac{x_2}{3/2} = 1.$$

Аналогично, для ограничения (В), уравнение прямой в отрезках будет

$$\frac{x_1}{1} + \frac{x_2}{3} = 1.$$

Построим обе прямые на плоскости. Множества точек, удовлетворяющие неравенствам (А) и (В) будут полуплоскости, лежащие под соответствующими прямыми, а множество допустимых значений переменных будет пересечением (общей частью) этих полуплоскостей, лежащее в первом квадранте: четырехугольник *ABCD* (рис. 9.2).

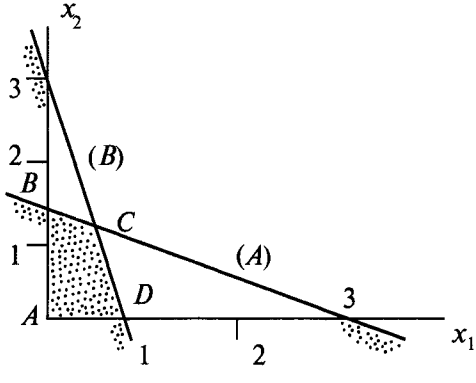


Рис. 9.2

2. На множестве допустимых решений (*ABCD*) найдем точку, в которой целевая функция $Z = 2x_1 + x_2$ имеет максимальное значение. Для этого рассмотрим *линии уровня* целевой функции. Линией уровня называется множество точек, на которых функция принимает постоянное значение

$$Z = 2x_1 + x_2 = k,$$

где k — задаваемая постоянная.

При $k = 1$ уравнение линии уровня будет

$$2x_1 + x_2 = 1,$$

или (в отрезках):

$$\frac{x_1}{1/2} + \frac{x_2}{1} = 1.$$

При $k = 2$, аналогично

$$2x_1 + x_2 = 2,$$

или

$$\frac{x_1}{1} + \frac{x_2}{2} = 1.$$

Нанеся линии уровня на область допустимых решений (рис. 9.3), получим, что при увеличении значения Z соответствующая линия уровня перемещается параллельно предыдущей вправо и вверх. Таким образом, точкой из многоугольника $ABCD$, в которой целевая функция Z имеет максимальное значение будет вершина C. Эта точка и определяет решение задачи.

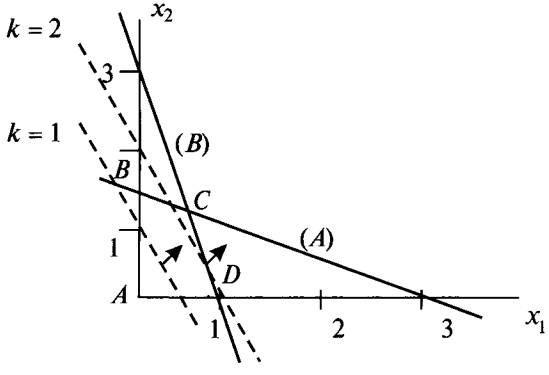


Рис. 9.3

3. Вычисление координат оптимальной точки (C).

Точка C лежит на пересечении прямых (A) и (B), поэтому, чтобы определить ее координаты надо решить систему уравнений:

$$\begin{cases} x_1 + 2x_2 = 3, & \text{(A)} \\ 3x_1 + x_2 = 3. & \text{(B)} \end{cases}$$

Решение:

$$x_1^* = 0{,}6;\ x_2^* = 1{,}2;$$

максимальное значение Z:

$$Z^* = 2 \cdot 0{,}6 + 1{,}2 = 2{,}4.$$

9.1.2. Алгебраическое решение задачи ЛП симплекс-методом

Если число переменных в задаче больше трех, то графический метод применить нельзя.

Анализ графического решения (рис. 9.3) дает возможность сформулировать следующие свойства задачи ЛП, которые справедливы для любого числа переменных и ограничений.

1. *Область допустимых значений переменных U — выпуклое множество.* В примере 9.1 — это четырехугольник *ABCD.* Если число ограничений в задаче с двумя переменными будет больше четырех, то область допустимых значений будет выпуклым многоугольником.

Это очень важное свойство всех задач ЛП, оно является ключевым для их решения симплекс-методом. Остановимся на нем более подробно. Напомним определение выпуклого множества. Множество точек U называется *выпуклым*, если для любых точек N_1 и N_2, принадлежащих множеству U, все точки отрезка N_1N_2 принадлежат U. Примеры выпуклого и невыпуклого множеств приведены на рис. 9.3*а*:

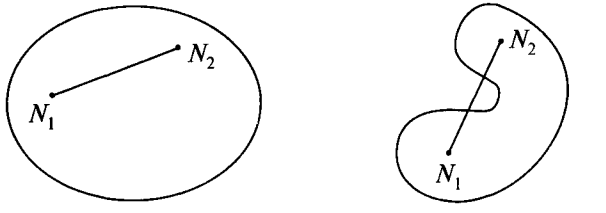


Рис. 9.3*а*. Выпуклое и невыпуклое множества

Множества точек, удовлетворяющие линейным неравенствам:

$$a_1x_1 + a_2x_2 \le b \quad \text{или} \quad a_1x_1 + a_2x_2 \ge b$$

есть полуплоскости (и выпуклые множества).

Аналогично, решения линейных неравенств:

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n \le b \quad \text{или} \quad a_1x_1 + a_2x_2 + \ldots + a_nx_n \ge b$$

образуют полупространства (и также будут выпуклыми множествами).

Решение системы линейных неравенств, задающих ограничения задачи ЛП, образует пересечение (или общую часть) этих полупространств. А пересечение выпуклых множеств есть выпуклое множество (доказательство этого утверждения сделайте самостоятельно).

2. *Если решение задачи* ЛП *существует и единственно, то оно достигается в вершине (угловой точке) множества U.* Это свойство также является ключевым для процедуры решения симплекс-методом. В примере 9.1 максимум целевой функции достигается в вершине *C*. Конечно, если линией уровня целевой функции Z будет прямая параллельная, например, отрезку

CD, то решением задачи будут все точки отрезка CD. В этом случае задача ЛП имеет бесчисленное множество решений. На практике такая возможность встречается крайне редко (из-за ошибок округления присутствующих всегда при расчетах на компьютере). Поэтому в дальнейшем мы будем рассматривать только случай, когда решение задачи ЛП достигается в угловой точке выпуклого множества U и, в этом случае, оно будет единственно, так как U — выпуклое множество.

3. *Может ли задача* ЛП *не иметь решения?* Да, такие случаи встречаются часто, и тому есть две причины.

1) множество допустимых решений U — пустое множество. Так бывает в тех случаях, когда ограничения задачи противоречат друг другу. Задача поставлена некорректно, ограничения задачи следует пересмотреть;

2) множество U — неограничено. Если при нахождении максимума целевой функции Z множество U неограничено «сверху», то Z может принимать сколь угодно большие значения и задача решения не имеет. Аналогичная ситуация будет при нахождении минимума Z, если U неограничено «снизу». В этом случае при постановке задачи не учтены реальные ограничения, их необходимо добавить.

Идея алгебраического решения задачи симплекс-методом основана на свойствах 1 и 2, сформулированных выше.

Предположим, что мы нашли координаты какой-либо угловой точки A_0 (вершины) множества допустимых решений U. Возьмем точку A_0 в качестве начального решения задачи и вычислим значение целевой функции Z в точке A_0, $Z(A_0)$. Далее рассмотрим угловые точки A_1, ..., A_k множества U, которые являются соседними (смежными) с точкой A_0 и вычислим $Z(A_1)$, $Z(A_2)$, ..., $Z(A_k)$. Если $Z(A_0) \geq Z(A_i)$, $i = 1, 2, ..., k$, то задача ЛП решена (множество допустимых решений U выпукло!) и решение есть точка A_0. Если $Z(A_0) < Z(A_i)$, $i = 1, 2, ..., k$, то в качестве новой (начальной) угловой точки берем точку A_l, в которой целевая функция имеет наибольшее значение: $Z(A_l) = \max\limits_{1 \leq j \leq k} Z(A_j)$ и повторяем процедуру снова (рассматриваем угловые точки смежные с A_l и, если во всех этих точках целевая функция принимает значение больше, чем $Z(A_l)$, переходим в точку с наибольшим значением Z). Так как число угловых точек выпуклого множества U конечно, то конечная последовательность таких переходов позволяет найти угловую точку являющуюся решением задачи.

Таким образом, для того чтобы решить задачу симплекс-методом, нужно, прежде всего, ответить на вопросы:

1) как определяется угловая точка выпуклого множества U;

2) как выполнить переход из заданной угловой точки в соседнюю, смежную угловую точку;

3) как определяется начальная угловая точка;

Решение этих вопросов составляет содержание определенных теорем в теории ЛП [28]. Здесь мы приведем необходимые определения и результаты этой теории без доказательства и продемонстрируем технику вычислений симплекс-методом на простом примере.

Решение задачи ЛП симплекс-методом

Рассмотрим задачу в примере 9.1: найти $\max Z = 2x_1 + x_2$ при ограничениях:

$$x_1 + 2x_2 \le 3, \quad \text{(A)}$$

$$3x_1 + x_2 \le 3, \quad \text{(B)}$$

$$x_{1,2} \ge 0.$$

Приведем задачу к *стандартному* виду. Для этого запишем ограничения (A) и (B) в виде равенств, добавляя в левые части неравенств неотрицательные числа x_3, x_4, которые называются *балансовыми* переменными:

$$x_1 + 2x_2 + x_3 = 3, \quad \text{(A)}$$

$$3x_1 + x_2 + x_4 = 3, \quad \text{(B)}$$

$$x_{1,2,3,4} \ge 0.$$

Здесь x_3 и x_4 — новые неизвестные. Теперь множество допустимых решений U — это множество *неотрицательных* решений системы двух уравнений:

$$\begin{cases} x_1 + 2x_2 + x_3 = 3, \\ 3x_1 + x_2 + x_4 = 3. \end{cases} \quad (1)$$

Эта система имеет бесчисленное множество решений, которые можно найти из *общего решения*, выразив его через переменные x_3 и x_4:

$$\begin{cases} x_3 = 3 - x_1 - 2x_2, \\ x_4 = 3 - 3x_1 - x_2. \end{cases} \quad (2)$$

Переменные x_1 и x_2 в правой части (2) называются *свободными*, а переменные x_3 и x_4 в левой части (2) называются *базисными* переменными. Например, при значениях свободных переменных $x_1 = 0{,}5$; $x_2 = 1$ из общего решения (2) получим, что $x_3 = 0{,}5$; $x_4 = 0{,}5$. Точка с координатами (0,5; 1; 0,5; 0,5) является решением (1) и, следовательно, принадлежит множеству U. Если систему (1) решить относительно двух любых других переменных, например x_1 и x_3, то общее решение будет записано в другом виде (используя второе уравнение, выражаем x_1 через x_2 и x_4, затем исключаем x_1 из первого уравнения):

$$\begin{cases} x_1 = 1 - 1/3\,x_2 - 1/3\,x_4, \\ x_3 = 2 - 5/3\,x_2 - 1/3\,x_4. \end{cases} \quad (3)$$

В общем решении (3) свободными переменными являются x_2 и x_4, а базисными переменными x_1 и x_3.

В общем случае число базисных переменных равно числу линейно независимых уравнений — ограничений при записи задачи в стандартном виде. Число базисных и свободных переменных для данной задачи остается неизменным, однако, при переходе от одной записи общего решения к

другой, свободные переменные могут становиться базисными и наоборот. Так, при переходе от общего решения (2) к общему решению (3), свободная переменная x_1 стала базисной, а базисная переменная x_4 стала свободной.

Имеет место следующая *теорема*:

угловые точки множества допустимых решений U — это такие решения системы уравнений—ограничений, для которых:

1) свободные переменные равны нулю;

2) базисные переменные — неотрицательны.

Верно и обратное утверждение: если для некоторого решения системы уравнений—ограничений выполнены условия 1) и 2), то это решение определяет угловую точку множества допустимых решений U.

Рассмотрим задачу в примере 9.1. Из общего решения в форме (2), приравняв свободные переменные x_1 и x_2 к нулю, получаем угловую точку A: ($x_1 = 0$; $x_2 = 0$; $x_3 = 3$; $x_4 = 3$); из общего решения в форме (3), приравняв свободные переменные x_2 и x_4 к нулю, получаем угловую точку D: ($x_1 = 1$; $x_2 = 0$; $x_3 = 2$; $x_4 = 0$).

Обратим внимание на то, что угловые точки A и D — смежные (см. рис. 9.3). Переход от A к D можно осуществить, выполнив преобразование общего решения в форме (2) к общему решению в форме (3). При этом одна свободная переменная (x_1) становится базисной, а одна базисная переменная (x_4) становится свободной. Можно показать, что такое преобразование всегда приводит к переходу в какую-либо угловую точку, смежную с начальной (исходной) угловой точкой.

Для полного определения симплекс-метода нам осталось определить выбор начальной угловой точки.

Если в исходной задаче все ограничения представляют неравенства вида «≤», то начальная угловая точка определяется сразу при приведении задачи к стандартному виду. Обратимся к примеру 9.1. Из системы (1) можно сразу получить общее решение (2), взяв в качестве базисных переменных балансовые переменные x_3 и x_4. Свободные переменные x_1, x_2 в угловой точке должны быть равны нулю. Таким образом, определена начальная угловая точка A: ($x_1 = 0$; $x_2 = 0$; $x_3 = 3$; $x_4 = 3$).

Если в исходной задаче некоторые ограничения имеют вид равенств или неравенств вида «≥», то соответствующие балансовые переменные будут либо отсутствовать, либо входить в уравнение стандартной системы с отрицательным знаком и не могут быть базисными. В этом случае для определения начальной точки используются специальные методы (например, метод искусственного базиса [28]). Мы не будем рассматривать этот метод, так как при решении задачи ЛП в пакете Excel начальное решение определять не нужно, это выполняет сама программа.

Рассмотрим решение примера 9.1 симплекс-методом. Предварительные этапы уже выполнены: исходная задача преобразована к стандартному виду (1), определено общее решение (2):

$$\begin{cases} x_3 = 3 - x_1 - 2x_2, \\ x_4 = 3 - 3x_1 - x_2 \end{cases} \qquad (2)$$

и определена начальная угловая точка A: ($x_1 = 0$; $x_2 = 0$; $x_3 = 3$; $x_4 = 3$). Среди угловых точек, являющихся решениями системы (1), нужно найти угловую точку, в которой целевая функция:

$$Z = 2x_1 + x_2 \tag{4}$$

будет иметь максимальное значение. В начальной угловой точке целевая функция равна нулю: $Z(A) = 0$.

Симплекс-метод выполняется последовательно, по шагам. На каждом шаге выполняется переход из предыдущей угловой точки в такую смежную угловую точку, в которой Z принимает наибольшее значение.

Шаг 1

a. Среди свободных переменных x_1 и x_2 выберем одну, которая в смежной угловой точке станет базисной. Очевидно, следует выбрать переменную x_1, так как в записи целевой функции (4) при x_1 стоит бóльший коэффициент — 2, чем при x_2, т. е. приращение Z при переходе в смежную угловую точку с базисной переменной x_1 будет больше.

b. Среди базисных переменных x_3 и x_4 выберем одну, которая станет свободной. Для этого запишем (2) с учетом того, что в смежной угловой точке x_2 остается свободной переменной и, следовательно, $x_2 = 0$:

$$\begin{cases} x_3 = 3 - x_1, \\ x_4 = 3 - 3x_1. \end{cases} \tag{5}$$

Из (5) следует, что при увеличении x_1 среди базисных переменных x_3 и x_4 первой обратится в нуль x_4, поэтому эту базисную переменную следует сделать свободной в новой угловой точке. Выбор базисной переменной более удобно выполнять, используя следующее правило: найдем отношения для свободных членов и коэффициентов при x_1 в правых частях (5). Для первой строки (x_3) это отношение 3/1 = 3, для второй (x_4) — 3/3 = 1. Минимальное отношение (3/3 = 1) соответствует x_4, следовательно, эта базисная переменная и будет свободной для новой угловой точки.

c. Выполняем преобразование (2) так, чтобы свободными переменными стали x_2 и x_4, а базисными x_1 и x_3. Из второго уравнения (2) получаем

$$x_1 = 1 - 1/3x_2 - 1/3x_4. \tag{6}$$

Подставив этот результат в первое уравнение (2), получим

$$x_3 = 2 - 5/3x_2 + 1/3x_4. \tag{7}$$

Таким образом, новая угловая точка имеет координаты: $x_1 = 1$; $x_2 = 0$; $x_3 = 2$; $x_4 = 0$. Это точка D (см. рис. 9.3).

d. Запишем Z как функцию свободных переменных x_2 и x_4. Для этого в (4) исключим x_1, используя (6):

$$Z = 2(1 - 1/3x_2 - 1/3x_4) + x_2 = 2 + 1/3x_2 - 2/3x_4, \tag{8}$$

$$Z(D) = 2.$$

Шаг 2

a. Из (8) следует, что Z можно увеличить, если в смежной угловой точке свободную переменную x_2 сделать базисной. Переменная x_4 остается свободной.

b. В (6) и (7) при $x_4 = 0$ найдем отношения: $1/(1/3) = 3$ и $2/(5/3) = 6/5$. Минимальное отношение (6/5) получается для (7). Таким образом, базисная переменная x_3 в новой угловой точке будет свободной.

c. Выполним преобразование (6) и (7) так, что свободными переменными будут x_3 и x_4, а базисными x_1и x_2.

Из (7) находим

$$x_2 = 6/5 - 3/5x_3 + 1/5x_4. \tag{9}$$

Подставляя (9) в (6), получим

$$x_1 = 3/5 + 1/5x_3 - 2/5x_4. \tag{10}$$

Таким образом, новая угловая точка имеет координаты $x_1 = 3/5$; $x_2 = 6/5$; $x_3 = 0$; $x_4 = 0$. Это точка C.

d. Запишем Z как функцию свободных переменных x_3 и x_4: исключим из (8) переменную x_2:

$$\begin{gathered} Z = 2 + 1/3(6/5 - 3/5x_3 + 1/5x_4) - 2/3x_4 = 12/5 - 1/5x_3 - 3/5x_4, \\ Z(C) = 12/5. \end{gathered} \tag{11}$$

Шаг 3

a. Из (11) следует, что за счет перевода свободных переменных x_3 и x_4 в базисные увеличить Z нельзя. Таким образом, решение задачи дает угловая точка C. Решение задачи ЛП симплекс-методом найдено.

Более удобно решение задачи ЛП симплекс-методом проводить с помощью симплекс-таблиц.

По существу последовательное преобразование симплекс-таблиц и есть выполнение шагов симплекс-метода, т. е. последовательный переход из одной угловой точки множества U в одну из смежных угловых точек. Это можно увидеть сравнивая выполнение последовательных шагов симплекс-метода, приведенных выше, с правилами симплекс-преобразования и последовательностью симплекс-таблиц.

9.1.3. Решение задачи ЛП в симплекс-таблицах

1. Приведение задачи к стандартному виду

Вводя вспомогательные (балансовые) переменные x_3 и x_4 в левые части неравенств (A) и (B), запишем ограничения в виде уравнений:

$$x_1 + 2x_2 + x_3 = 3, \tag{A}$$

$$3x_1 + x_2 + x_4 = 3. \tag{B}$$

Целевая функция $Z = 2x_1 + x_2$ при решении задачи в симплекс-таблицах записывается так:

$$Z - 2x_1 - x_2 = 0. \tag{C}$$

2. Составление первой симплекс-таблицы

Симплекс-таблица составляется из коэффициентов при x_1, x_2, x_3, x_4 и чисел, стоящих в правых частях уравнений—ограничений задачи: в первой строке записываются элементы уравнения (А), во второй — (В). В последней строке симплекс-таблицы записываются коэффициенты и правая часть целевой функции (С). Таким образом, симплекс-таблица содержит две строки коэффициентов (по числу ограничений задачи) и строку коэффициентов целевой функции. Число столбцов в симплекс-таблице равно числу переменных задачи плюс один столбец правых частей (b):

	x_1	x_2	x_3	x_4	b
(А)	1	2	1	0	3
(В)	3	1	0	1	3
(С)	–2	–1	0	0	0

Переменные, для которых столбцы коэффициентов состоят из одной единицы и нулей, называются *базисными* (в приведенном примере x_3 и x_4 — базисные переменные). Число базисных переменных равно числу ограничений задачи и не меняется при симплекс-преобразовании. Остальные переменные называются *свободными* (x_1 и x_2).

Симплекс-таблица определяет частное решение системы уравнений—ограничений:

$$\begin{cases} x_1 + 2x_2 + x_3 = 3, & \text{(А)} \\ 3x_1 + x_2 + x_4 = 3, & \text{(В)} \end{cases}$$

при котором свободные переменные равны нулю ($x_1 = 0$, $x_2 = 0$), а базисные переменные равны правым частям соответствующих строк ($x_3 = 3$, $x_4 = 3$), т. е. угловую точку A (рис. 9.3).

Значение целевой функции Z всегда равно числу, стоящему в правом нижнем углу таблицы ($Z = 2 \cdot 0 + 1 \cdot 0 = 0$). Первая симплекс-таблица соответствует начальному решению задачи ЛП ($x_1 = 0$, $x_2 = 0$, $x_3 = 3$, $x_4 = 3$, $Z = 0$).

Симплекс-метод состоит в последовательном перемещении по вершинам многоугольника допустимых решений. Каждой вершине соответствует своя симплекс-таблица, которая получается из предыдущей при помощи симплекс-преобразования. Симплекс-преобразованию предшествует выбор разрешающей строки и разрешающего столбца.

В качестве *разрешающего столбца* берут столбец, у которого коэффициент в строке целевой функции является отрицательным и наибольшим по модулю. Если в данной симплекс-таблице строка целевой функции не содержат отрицательных коэффициентов, то решение задачи ЛП закончено и симплекс-таблица определяет решение задачи, при котором целевая функция Z принимает максимальное значение.

Разрешающая строка определяется по отношениям коэффициентов столбца b к соответствующим коэффициентам *разрешающего столбца. Раз-*

решающей будет строка, для которой это отношение *минимально*. При этом для нулевых и отрицательных коэффициентов разрешающего столбца отношения не вычисляются (и соответствующие строки не могут быть разрешающими).

Для первой симплекс-таблицы разрешающим столбцом является первый столбец (свободная переменная x_1 будет преобразована в базисную). Среди отношений коэффициентов столбца b к коэффициентам разрешающего столбца: 3/1 и 3/3 минимальным будет отношение 3/3: разрешающей строкой будет вторая строка (базисная переменная x_4 будет преобразована в свободную).

x_1	x_2	x_3	x_4	b
1	2	1	0	3
→ [3]	1	0	1	3
–2	–1	0	0	0
↑				

На пересечении разрешающего столбца и разрешающей строки находится *разрешающий элемент* (он выделен знаком □).

Задача симплекс преобразования состоит в том, чтобы на месте разрешающего элемента получить единицу, а все остальные элементы разрешающего столбца сделать нулевыми.

При этом допускается выполнение только двух операций со строками симплекс-таблицы:

а) *разрешающую* строку можно делить (умножать) на любое число;

б) из любой строки можно вычитать элементы *разрешающей* строки или к любой строке можно прибавлять элементы *разрешающей* строки.

Шаг 1. Выполним преобразование первой симплекс-таблицы.

1) Делим элементы разрешающей строки на 3:

x_1	x_2	x_3	x_4	b
1	2	1	0	3
1	1/3	0	1/3	1
–2	–1	0	0	0

2) Из элементов первой строки вычитаем элементы второй (*разрешающей*) строки (при этом первый элемент разрешающего столбца будет равен нулю):

x_1	x_2	x_3	x_4	b
0	5/3	1	–1/3	2
1	1/3	0	1/3	1
–2	–1	0	0	0

3) К элементам третьей строки прибавляем элементы второй (*разрешающей*) строки, предварительно умножив их на два (при этом третий элемент разрешающего столбца будет равен нулю):

x_1	x_2	x_3	x_4	b	
0	5/3	1	–1/3	2	
1	1/3	0	1/3	1	
0	–1/3	0	2/3	2	← $Z(D)$

Преобразование закончено. Полученной симплекс-таблице соответствует следующее решение:

базисные переменные: $x_1 = 1$, $x_3 = 2$ (их значения определяются с помощью стрелок ↳, указанных в симплекс-таблицах;

свободные переменные: $x_2 = 0$, $x_4 = 0$.

Точка с координатами $x_1 = 1$, $x_2 = 0$ — это вершина D (см. рис. 9.3). Значение целевой функции $Z(D) = 2$ (в правом нижнем углу таблицы).

Шаг 2. Так как в строке коэффициентов целевой функции есть отрицательный коэффициент (–1/3 во втором столбце), то преобразование продолжается. Второй столбец является разрешающим (свободная переменная x_2 переводится в базисную), минимальным среди отношений: $\frac{2}{5/3} = \frac{6}{5}$ и $\frac{1}{1/3} = 3$ является первое число, следовательно, разрешающей строкой является первая строка (базисная переменная x_3 переводится в свободную).

	x_1	x_2	x_3	x_4	b
→	0	**5/3**	1	–1/3	2
	1	1/3	0	1/3	1
	0	–1/3	0	2/3	2

Выполнив симплекс-преобразование, получим:

x_1	x_2	x_3	x_4	b	
0	1	3/5	–1/5	6/5	
1	0	0	1/3	3/5	
0	0	1/5	3/5	12/5	← $Z(C)$

Шаг 3. Так как в строке коэффициентов целевой функции нет отрицательных, решение задачи закончено.

Оптимальное решение таково:

базисные переменные: $x_1^* = 3/5 = 0{,}6$; $x_2^* = 6/5 = 1{,}2$;

свободные переменные: $x_3^* = 0$; $x_4^* = 0$.

Точка с координатами $x_1^* = 0{,}6$ и $x_2^* = 1{,}2$ это вершина C (см. рис. 9.3).

Максимальное значение дохода (целевой функции):

$$Z^*(C) = 12/5 = 2{,}4.$$

9.1.4. Решение задачи распределения ресурсов в EXCEL

1. Ввод данных примера 1 в таблицу EXCEL показан на рис. 9.4.

	Перем енные			
Имя	Краска 1	Краска 2	Доход	
	2	1		
	Ограни чения			
Ресурс			Расход	Запасы
A	1	2		3
B	3	1		3

Рис. 9.4

На рис. 9.4 «Краска 1» обозначает краску для внутренних работ, «Краска 2» — краску для наружных работ.

Для переменных задачи x_1 и x_2 отведены ячейки B3 и C3. Эти ячейки называются *рабочими, или изменяемыми* ячейками. *В изменяемые ячейки ничего не заносится* и в результате решения задачи в этих ячейках будут записаны оптимальные значения переменных.

В ячейку D4 вводится формула для вычисления целевой функции задачи (дохода) $Z = 2x_1 + x_2$. Чтобы сделать это надо выполнить следующие действия:

1) курсор в D4;

2) курсор на кнопку f_x (мастер функций);

3) в появившемся окне выбрать в левом столбце — «Математические», а в правом столбце — «СУММПРОИЗВ» (рис. 9.5);

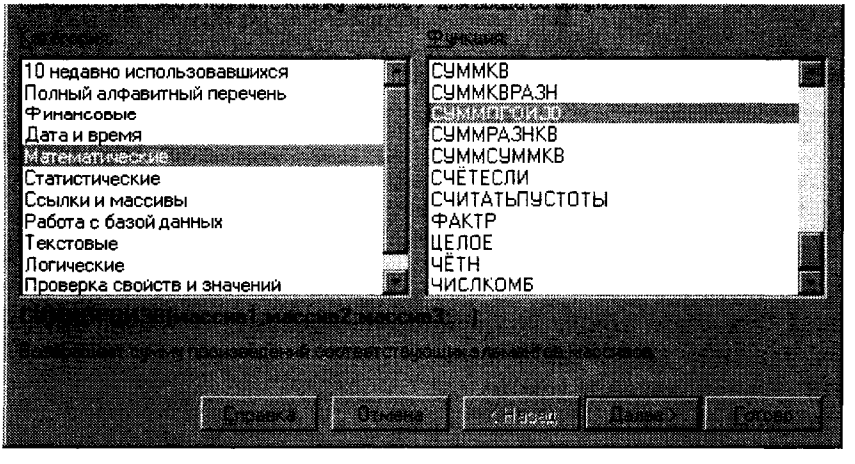


Рис. 9.5

4) в окне мастера функций нажать *Далее>*, в появившемся окне (рис. 9.6) в поле «массив 1» ввести (протаскивая курсор мыши по ячейкам)

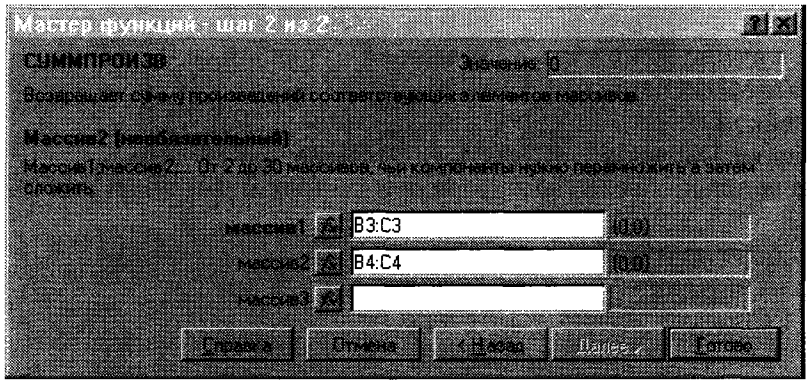


Рис. 9.6

адреса изменяемых ячеек B3:C3. В поле «массив 2» вводятся адреса ячеек содержащих цены на краски B4:C4, после нажать *Готово.*

В ячейку D7 вводится формула для вычисления израсходованного количества продукта A: $x_1 + 2x_2$, а в ячейку D8 вводится формула для израсходованного количества продукта B: $3x_1 + x_2$. Обе формулы вводятся аналогично целевой функции (рис. 9.7 и 9.8).

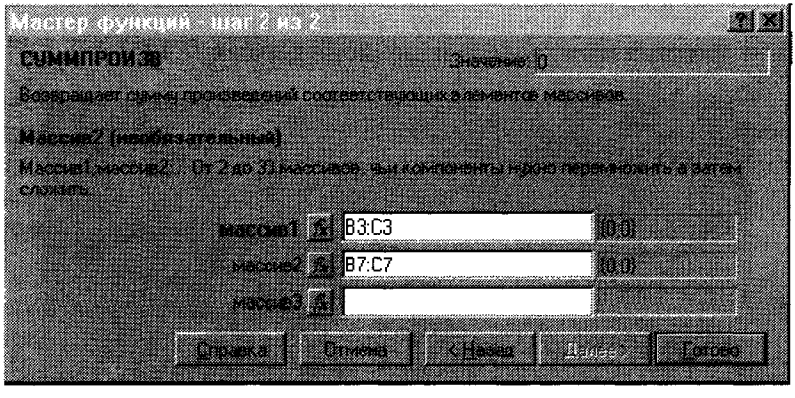


Рис. 9.7

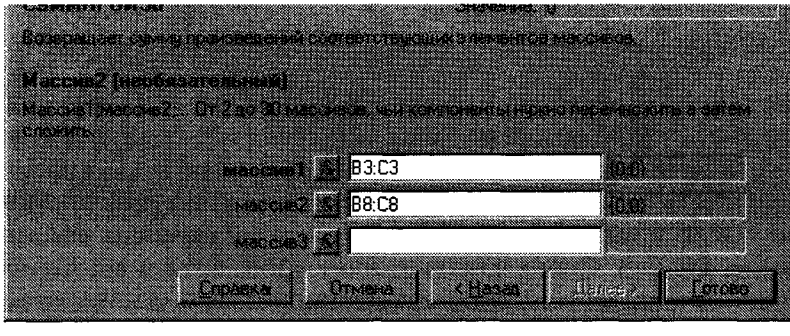


Рис. 9.8

Проверить результаты ввода можно следующим образом: при установке курсора в ячейку D4 в строке ввода должно появиться:

«=СУММПРОИЗВ(B3:C3;B4:C4)».

В ячейку D7: «=СУММПРОИЗВ(B3:C3;B7:C7)».

В ячейки D8: «=СУММПРОИЗВ(B3:C3;B8:C8)».

Окончательно после ввода формул и данных экран имеет вид (рис. 9.9).

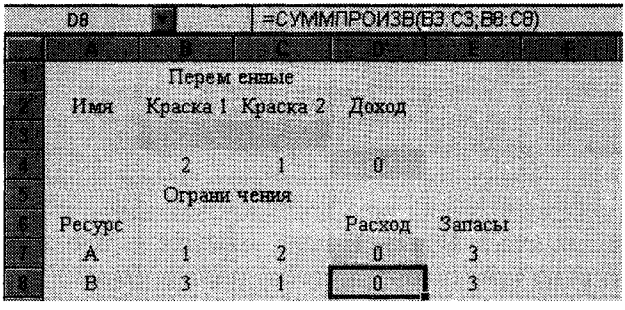


Рис. 9.9

2. Работа в окне «Поиск решения»

В меню «Сервис» выбираем процедуру «Поиск решения».

В появившемся окне (рис. 9.10) нужно установить адрес ячейки D4, содержащей формулу для вычисления целевой функции, значение целевой функции — *максимальное*, адреса изменяемых ячеек: B3:C3.

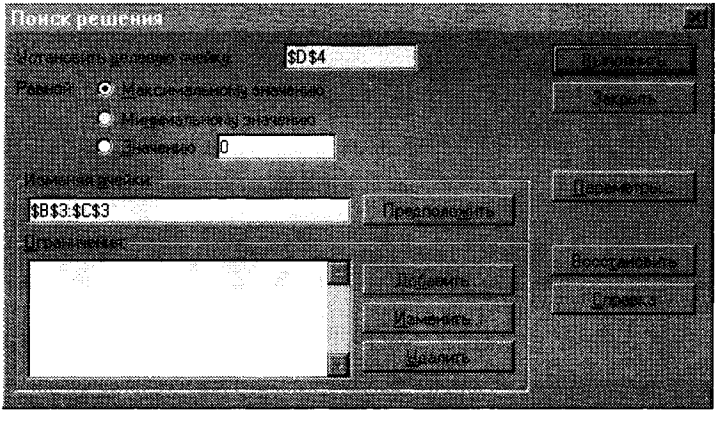


Рис. 9.10

Чтобы ввести ограничения задачи, нужно нажать кнопку «Добавить». В появившемся диалоговом окне (рис. 9.11) слева ввести адрес D7 (израсходованное количество продукта A), затем выбрать знак <= и в правой части ввести количество продукта A на складе, равное 3 (или адрес ячейки E7).

После ввода нажать кнопку «Добавить» и аналогично ввести второе ограничение: D8 <= 3. Снова нажать кнопку «Добавить» и ввести ограничение: B3:C3 >= 0 (соответствующее ограничению x_1, x_2 >= 0). После ввода последнего ограничения нажать OK. После ввода всех ограничений окно «Поиска решений имеет» будет иметь следующий вид (рис. 9.12).

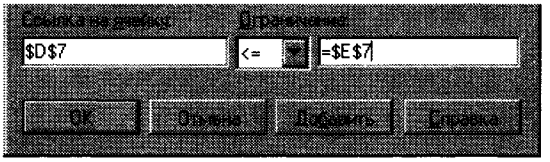


Рис. 9.11

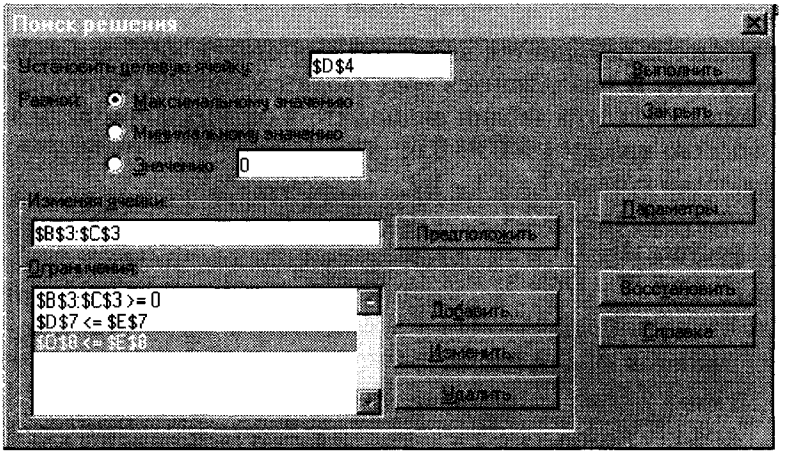


Рис. 9.12

3. Настройка параметров решения задачи

В окне «Поиск решения» нажать «Параметры» в появившемся окне (рис. 9.13) установить флажок в пункте «Линейная модель». В этом случае при решении задачи будет использоваться симплекс-метод. Остальные значения можно оставить без изменения. После нажать кнопку OK.

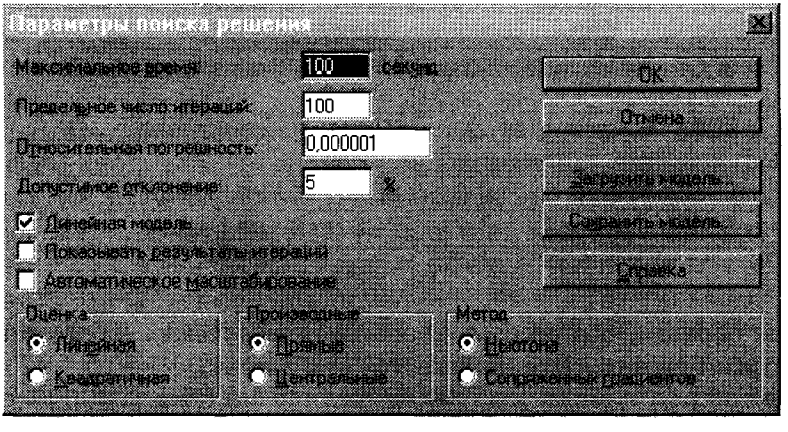


Рис. 9.13

Для решения задачи в окне «Поиск решения» нажать кнопку *«Выполнить»*. Если решение найдено, то появляется диалоговое окно для выбора вида отчета (рис. 9.14).

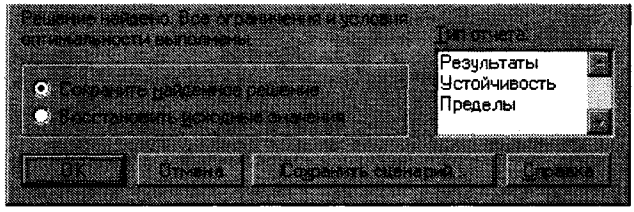


Рис. 9.14

Для просмотра результатов выбираем тип отчета: «Результаты» и нажимаем кнопку OK. В появившихся трех таблицах (рис. 9.15) приводятся результаты поиска. Из этих таблиц видно, что в оптимальном решении:

производство краски 1 = B3 = 0,6;

производство краски 2 = C3 = 1,2;

при этом доход = D4 = 2,4;

расход ресурса A = D7 = 3;

расход ресурса B = D8 = 3.

Целевая ячейка (Макс)

Ячейка	Имя	Исходно	Результат
$D$4	Доход	2,4	2,4

Изменяемые ячейки

Ячейка	Имя	Исходно	Результат
$B$3	Краска 1	0,6	0,6
$C$3	Краска 2	1,2	1,2

Ограничения

Ячейка	Имя	Значение	Формула	Состояние	Разница
$D$7	A Расход	3	$D$7<=$E$7	связанное	0
$D$8	B Расход	3	$D$8<=$E$8	связанное	0
$B$3	Краска 1	0,6	$B$3>=0	не связан.	0,6
$C$3	Краска 2	1,2	$C$3>=0	не связан.	1,2

Рис. 9.15

Таким образом, оба ресурса являются *дефицитными* (соответствующие этим ресурсам ограничения называются *связанными*).

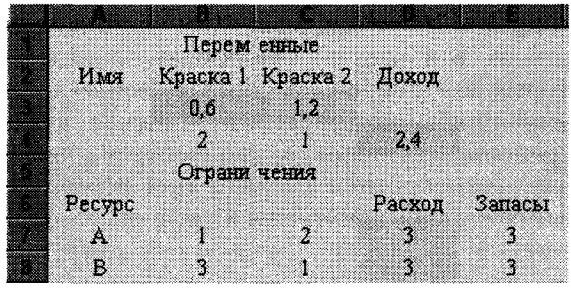
«Отчет по результатам» состоит из трех таблиц (рис. 9.15):

в таблице 1 приводятся сведения о целевой функции;

в таблице 2 приводятся значения переменных задачи;

в таблице 3 показаны результаты поиска для ограничений задачи.

Первоначальная таблица EXCEL заполняется результатами, полученными при решении (рис. 9.16).



	A	B	C	D	E
1		Перем енные			
2	Имя	Краска 1	Краска 2	Доход	
3		0,6	1,2		
4		2	1	2,4	
5		Ограни чения			
6	Ресурс			Расход	Запасы
7	A	1	2	3	3
8	B	3	1	3	3

Рис. 9.16

Целочисленное линейное программирование

Большой класс задач ЛП предполагает, что все или некоторые переменные задачи должны принимать целые (дискретные) значения. Для решения таких задач используются специальные методы: метод ветвей и границ, метод Гомори (см. [28]).

При решении задач такого вида с помощью процедуры «Поиск решения» никаких дополнительных сложностей не возникает: при вводе ограничений задачи нужно указать дополнительное ограничение — все или часть переменных *целые*.

Например, если пример 9.1 решать как целочисленную задачу ЛП и дополнительно к ограничениям (см. рис. 9.12) задать ограничение: x_1 и x_2 — *целые*, то решением задачи будет точка D, $x_1 = 1$, $x_2 = 0$, $Z(D) = 2$.

9.2. Транспортная задача

Транспортные модели (задачи) представляют специальный класс задач линейного программирования. Такие модели используются для разработки наиболее экономичных планов перевозки продукции одного вида из нескольких пунктов (например, складов) в пункты назначения (например, магазины).

Транспортные модели также применяются при составлении расписаний, управлении запасами, управлении движением капиталов, назначении персонала (см. п. 9.3) и во многих других задачах подобного вида.

В свою очередь транспортные модели и их модификации представляют собой частный случай так называемых сетевых моделей, в рамках которых можно сформулировать и решить большое число практически важных задач, в частности, задачу нахождения наикратчайшего пути между двумя пунктами по существующей сети дорог (см. п. 9.4).

Транспортную задачу и ее модификации можно решать симплекс-методом. Однако специфика ограничений в этих задачах позволила разработать более эффективные вычислительные процедуры для их решения, в частности, метод потенциалов для транспортной задачи [28]. По существу метод потенциалов воспроизводит все шаги симплекс-метода и здесь рассматриваться не будет.

Пример 9.2. Фирма обслуживающая туристов прибывающих на отдых, должна разместить их в 4 отелях: «Морской», «Солнечный», «Слава» и «Уютный», в которых забронировано соответственно 5, 15, 15 и 10 мест. Пятнадцать туристов прибывают по железной дороге, двадцать пять прилетают очередным рейсом в аэропорт, а пять человек прибудут на теплоходе на морской вокзал. Транспортные расходы при перевозке одного туриста из пунктов прибытия в отели приведены в табл. 9.1.

Таблица 9.1

Исходный пункт, i		Пункт назначения (отели), j			
		Морской	Солнечный	Слава	Уютный
		1	2	3	4
Железнодорожный вокзал	1	10	0	20	11
Аэропорт	2	12	7	9	20
Морской вокзал	3	0	14	16	18

В условиях жесткой конкуренции фирма должна минимизировать свои расходы, значительную часть которых составляют именно транспортные расходы. Требуется определить такой план перевозки туристов из пунктов прибытия в отели, при котором суммарные транспортные расходы будут минимальны и все туристы будут размещены в отелях.

I. Математическая модель задачи

1. *Переменные задачи.* Обозначим количество туристов, которые будут перевозиться из пункта i в отель j как X_{ij} ($i = 1, 2, 3$; $j = 1, 2, 3, 4$). Это переменные задачи, значения которых должны быть определены в процессе решения. Например, X_{23} — это число туристов, которое должно быть перевезено из аэропорта (пункт 2) в отель «Слава» (пункт 3). В задаче содержится $3 \cdot 4 = 12$ переменных.

2. *Ограничения на переменные задачи.* Очевидно, что все переменные задачи неотрицательные и целые числа, т. е.

$$X_{ij} \geq 0, \tag{1}$$

$$X_{ij} \text{ — целые числа,} \tag{2}$$

где $i = 1, 2, 3$; $j = 1, 2, 3, 4$.

Кроме этого, должны удовлетворяться следующие условия. Число туристов, вывозимых с железнодорожного вокзала (пункт 1) равно 15, поэтому

$$X_{11} + X_{12} + X_{13} + X_{14} = \sum_{j=1}^{4} X_{1j} = 15. \tag{3}$$

Аналогично, для аэропорта (пункт 2):

$$X_{21} + X_{22} + X_{23} + X_{24} = \sum_{j=1}^{4} X_{2j} = 25, \tag{4}$$

и для морского вокзала (пункт 3):

$$X_{31} + X_{32} + X_{33} + X_{34} = \sum_{j=1}^{4} X_{3j} = 5. \quad (5)$$

По условию задачи в отеле «Морской» (пункт 1) забронировано 5 мест, поэтому

$$X_{11} + X_{21} + X_{31} = \sum_{i=1}^{3} X_{i1} = 5. \quad (6)$$

Аналогично, для отеля «Солнечный» (пункт 2):

$$X_{12} + X_{22} + X_{32} = \sum_{i=1}^{3} X_{i2} = 15. \quad (7)$$

Для отеля «Слава» (пункт 3):

$$X_{13} + X_{23} + X_{33} = \sum_{i=1}^{3} X_{i3} = 15. \quad (8)$$

Для отеля «Уютный» (пункт 4):

$$X_{14} + X_{24} + X_{34} = \sum_{i=1}^{3} X_{i4} = 10. \quad (9)$$

Обычно транспортная задача записывается в виде таблицы, где в ячейках помещаются переменные задачи (X_{ij}), а в правом верхнем углу ячейки стоят стоимости перевозки из пункта i в пункт j (C_{ij}). В крайнем правом столбце и нижней строке таблицы записываются числа определяющие ограничения задачи (в данном примере — это число туристов в исходных пунктах и число мест в пунктах назначения — отелях).

Для примера 2 таблица имеет вид (табл. 9.2):

Таблица 9.2

Исходный пункт, i	Пункт назначения (отели), j				Число туристов в исходном пункте
	1	2	3	4	
1	10 X_{11}	0 X_{12}	20 X_{13}	11 X_{14}	15
2	12 X_{21}	7 X_{22}	9 X_{23}	20 X_{24}	25
3	0 X_{31}	14 X_{32}	16 X_{33}	18 X_{34}	5
Число мест в отеле	5	15	15	10	$\sum = 45$ / $\sum = 45$

Транспортная задача, для которой суммы чисел в последнем столбце и нижней строке равны, называется *сбалансированной*: 15 + 25 + 5 = 45,

5 + 15 + 15 + 10 = 45. Если транспортная задача не сбалансирована, то в таблицу добавляется еще одна строка или столбец. Причем стоимости перевозки в добавленных ячейках принимаются равными нулю.

Рассмотрим пример несбалансированной транспортной задачи. Предположим, что в аэропорт прибыло не пять, а десять туристов. Сумма чисел в последнем столбце будет равна: 15 + 25 + 10 = 50. Чтобы сбалансировать задачу вводим пятый столбец (фиктивный отель) с пятью местами. Таблица для сбалансированной транспортной задачи в этом случае будет иметь вид (табл. 9.3):

Таблица 9.3

Исходный пункт, i	Пункт назначения (отели), j					Число туристов в исходном пункте
	1	2	3	4	5	
1	10 X_{11}	0 X_{12}	20 X_{13}	11 X_{14}	0 X_{15}	15
2	12 X_{21}	7 X_{22}	9 X_{23}	20 X_{24}	0 X_{25}	25
3	0 X_{31}	14 X_{32}	16 X_{33}	18 X_{34}	0 X_{35}	10
Число мест в отеле	5	15	15	10	5	$\sum = 50$ / $\sum = 50$

3. *Целевая функция.* Вернемся к исходной задаче (табл. 9.2). Транспортные расходы на перевозку туристов в отели вычисляются по формуле

$$Z = \sum_{i=1}^{3}\sum_{j=1}^{4} C_{ij}X_{ij} = 10 \cdot X_{11} + 0 \cdot X_{12} + 20 \cdot X_{13} + \ldots + 18 \cdot X_{34}. \qquad (10)$$

Окончательно транспортная задача имеет вид (табл. 9.2). Нужно найти такие значения переменных X_{ij} ($i = 1, 2, 3$; $j = 1, 2, 3, 4$) при которых целевая функция, определяемая формулой (10), будет иметь минимальное значение и будут выполнены ограничения (1)—(9):

$X_{ij} \geq 0$, где X_{ij} — целые числа ($i = 1, 2, 3$; $j = 1, 2, 3, 4$).

$$\sum_{j=1}^{4} X_{1j} = 15; \qquad \sum_{i=1}^{3} X_{i1} = 5;$$

$$\sum_{j=1}^{4} X_{2j} = 25; \qquad \sum_{i=1}^{3} X_{i2} = 15;$$

$$\sum_{j=1}^{4} X_{3j} = 5; \qquad \sum_{i=1}^{3} X_{i3} = 15;$$

$$\sum_{i=1}^{3} X_{i4} = 10.$$

II. Решение транспортной задачи в процедуре EXCEL «Поиск решения»

При решении транспортной задачи в EXCEL задача должна быть предварительно *сбалансирована.*

1. *Ввод данных.* Вводим данные табл. 9.1, 9.2 в ячейки EXCEL (рис. 9.17).

В ячейках B3:E5 введены стоимости перевозок (табл. 9.1).

В ячейках F3:F5 находится число прибывающих туристов, а в ячейках B6:E6 находится число мест в отелях. Ячейки B8:E10 — рабочие (изменяемые) ячейки, в которых будут вычисляться значения переменных задачи X_{ij}.

В ячейках F8:F10 нужно записать формулы для вычисления левых частей ограничений (3)—(5):

в F8 должна быть сумма ячеек B8:E8;

в F9 должна быть сумма ячеек B9:E9;

в F10 должна быть сумма ячеек B10:E10.

Формулы для вычисления левых частей ограничений (6)—(9) введем в ячейки B11:E11:

в B11 должна быть сумма ячеек B8:B10;

в C11 должна быть сумма ячеек C8:C10;

в D11 должна быть сумма ячеек D8:D10;

в E11 должна быть сумма ячеек E8:E10;

Целевую функцию поместим в ячейку G3:

G3:СУММПРОИЗВ (B3:E5; B8:E10).

Таблица исходных данных имеет вид (рис. 9.17):

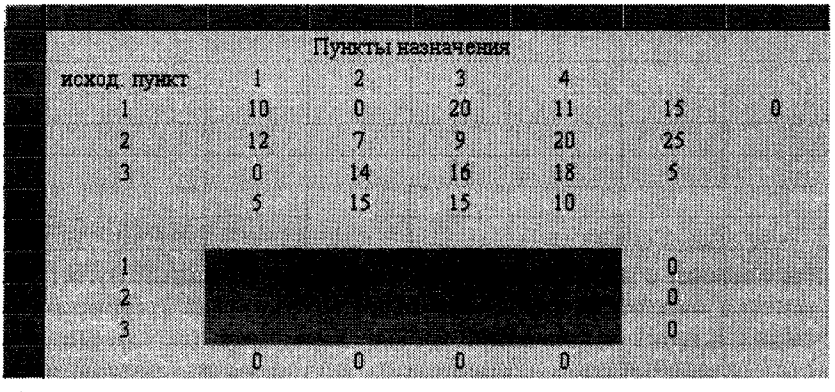


Рис. 9.17

2. *Заполнение окна процедуры «Поиск решения».*

Целевая функция: G3.

Значение целевой функции: min.

Изменяемые ячейки: B8:E10.

Ограничения задачи:

F8:F10 = F3:F5 (формулы (3)—(5));

B11:E11 = B6:E6 (формулы (6)—(9));

B8:E10 ≥ 0 (1) и B8:E10 — целые числа (2).

В окне «Параметры» установить «Линейная модель», что соответствует решению задачи симплекс-методом. Результаты заполнения окна показаны на рис. 9.18.

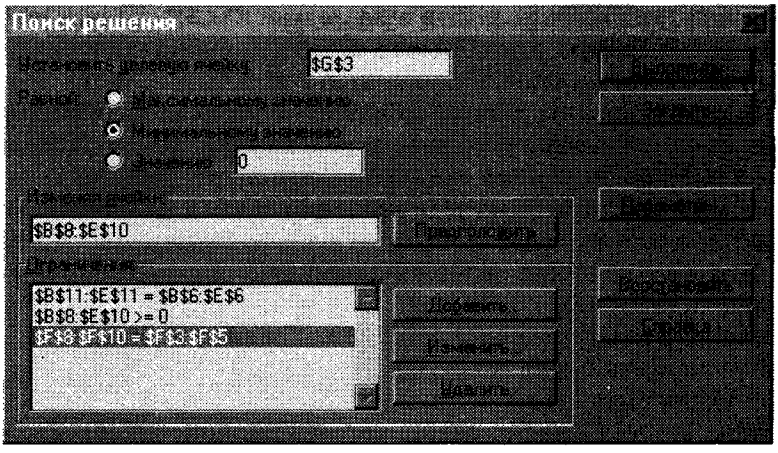


Рис. 9.18

3. Выполнив процедуру «Поиск решения», получим следующие результаты (рис. 9.19):

		Пункты назначения					
2	исход. пункт	1	2	3	4		
3	1	10	0	20	11	15	315
4	2	12	7	9	20	25	
5	3	0	14	16	18	5	
6		5	15	15	10		
7							
8	1	0	5	0	10	15	
9	2	0	10	15	0	25	
10	3	5	0	0	0	5	
11		5	15	15	10		

Рис. 9.19

Таким образом с железнодорожного вокзала (исходный пункт 1) следует 10 туристов отвезти в отель «Уютный» (пункт 4) и 5 туристов в отель «Солнечный» (пункт назначения 2); из аэропорта (исходный пункт 2) 10 туристов отвезти в отель «Солнечный» (пункт назначения 2) и 15 туристов в отель «Слава» (пункт назначения 3); туристов прибывающих на морской вокзал (исходный пункт 3) нужно отправить в отель «Морской» (пункт назначения 1). Все эти результаты видны в конечной таблице (рис. 9.19). При этом суммарная стоимость транспортных расходов составит 315 руб. (ячейка G3).

9.3. Задача о назначениях

В этой задаче необходимо выбрать претендентов на имеющиеся вакансии таким образом, чтобы сумма баллов полученных отобранными претендентами при тестировании была бы максимальной. Таким образом, нужно отобрать лучших среди претендентов.

Задачу о назначениях можно рассматривать как транспортную задачу, в которой претенденты представляют «исходные пункты», а вакансии — «пункты назначения». Стоимость «перевозки» из пункта i в пункт j равна C_{ij} — баллу, набранному i-м претендентом при тестировании на j-ю вакансию.

Заметим, что формулировки задачи о назначениях могут быть чрезвычайно разнообразны, но, кратко, это всегда выбор лучшего исполнителя для работы.

Пример 9.3. В конкурсе на занятие пяти вакансий (V_1, V_2, V_3, V_4, V_5) участвуют семь претендентов (P_1, P_2, P_3, P_4, P_5, P_6, P_7). Результаты тестирования каждого претендента, на соответствующие вакансии, даны в виде матрицы — C (тестирование производилось по десятибалльной системе).

Определить, какого претендента и на какую вакансию следует принять, причем так, чтобы сумма баллов отобранных претендентов оказалась максимальной.

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	7	5	7	6	7
P_2	6	4	8	4	9
P_3	8	6	4	3	8
P_4	7	7	8	5	7
P_5	5	9	7	9	5
P_6	6	8	6	4	7
P_7	7	7	8	6	4

I. Математическая модель задачи

1. *Переменные задачи.*

Введем переменные x_{ij}, принимающие два значения:

$x_{ij} = 0$, если i-й претендент (P_i) не принимается на j-ю вакансию (V_j).

$x_{ij} = 1$, если i-й претендент (P_i) принимается на вакансию (V_j).

$i = 1, 2, ..., 7$; $j = 1, 2, ..., 5$.

2. *Ограничения на переменные задачи.*

Очевидно, что все переменные задачи неотрицательные и двоичные числа: $x_{ij} \geq 0$ и x_{ij} — могут принимать значения 0 или 1.

Кроме того, так как каждый претендент может занять только одну вакансию и все вакансии должны быть заняты, должны удовлетворяться следующие ограничения:

$$\sum_{i=1}^{7} x_{ij} = 1, \quad j = 1, 2, ..., 7,$$

$$\sum_{j=1}^{5} x_{ij} = 1, \quad i = 1,2, ..., 5.$$

Другими словами, в матрице (x_{ij}) суммы элементов по каждой строке и суммы элементов по каждому столбцу должны быть равны единицам. Это условие означает, что выбор претендентов должен быть таким, чтобы в матрице (x_{ij}), представляющей решение задачи, было бы по одной единице в каждой строке и по одной единице в каждом столбце, остальные элементы матрицы должны равняться нулю.

3. *Целевая функция в задаче о назначениях.*

Необходимо выбрать претендентов так, чтобы суммарное число очков, набранное ими было бы максимальным. Суммарное число набранных очков вычисляется по формуле:

$$Z = \sum_{i=1}^{7}\sum_{j=1}^{5} x_{ij}c_{ij};$$

$$Z = c_{11}x_{11} + c_{12}x_{12} + \ldots + c_{75}x_{75} = 7x_{11} + 5x_{12} + \ldots + 4x_{75}.$$

Математическая модель задачи записывается так:

$$\text{найти} \quad \max Z = \sum_{i=1}^{7}\sum_{j=1}^{5} x_{ij}c_{ij}$$

при ограничениях:

$$x_{ij} \geq 0 \text{ и } x_{ij} \text{ — двоичные числа, } i = 1, 2, \ldots, 7; \quad j = 1, 2, \ldots, 5;$$

$$\sum_{i=1}^{7} x_{ij} = 1, \quad j = 1, 2, \ldots, 7;$$

$$\sum_{j=1}^{5} x_{ij} = 1, \quad i = 1, 2, \ldots, 5.$$

Таким образом, задача о назначениях есть частный случай транспортной задачи.

II. Решение задачи о назначениях при помощи преобразования матрицы (*C*)

Рассмотрим решение задачи о назначениях, в которой нужно найти минимум функции *Z*. Предварительно задачу о назначениях нужно сбалансировать. В рассматриваемом примере эта процедура выполняется добавлением двух столбцов (две фиктивные вакансии) с нулевыми результатами тестирования:

$$C = \begin{pmatrix} 7 & 5 & 7 & 6 & 7 & 0 & 0 \\ 6 & 4 & 8 & 4 & 9 & 0 & 0 \\ 8 & 6 & 4 & 3 & 8 & 0 & 0 \\ 7 & 7 & 8 & 5 & 7 & 0 & 0 \\ 5 & 9 & 7 & 9 & 5 & 0 & 0 \\ 7 & 8 & 6 & 4 & 7 & 0 & 0 \\ 6 & 7 & 8 & 6 & 4 & 0 & 0 \end{pmatrix}.$$

Задача нахождения максимального значения функции Z эквивалентна задаче нахождения минимума для функции $-Z = \sum_{i=1}^{7}\sum_{j=1}^{5} x_{ij}(-c_{ij})$, матрица $(-C)$ имеет вид:

$$-C = \begin{pmatrix} -7 & -5 & -7 & -6 & -7 & 0 & 0 \\ -6 & -4 & -8 & -4 & -9 & 0 & 0 \\ -8 & -6 & -4 & -3 & -8 & 0 & 0 \\ -7 & -7 & -8 & -5 & -7 & 0 & 0 \\ -5 & -9 & -7 & -9 & -5 & 0 & 0 \\ -7 & -8 & -6 & -4 & -7 & 0 & 0 \\ -6 & -7 & -8 & -6 & -4 & 0 & 0 \end{pmatrix}.$$

Можно показать (функция Z линейна!), что при вычитании из всех элементов столбца или строки матрицы (C) одного и того же числа, решения x_{ij}, при которых функция $Z = \sum_{i=1}^{7}\sum_{j=1}^{5} x_{ij}c_{ij}$, имеет минимум (или максимум) не меняется. Поэтому матрицу (C) преобразуем по следующему правилу. В каждой строке (C) и в каждом столбце образуют нули, вычитая минимальные элементы из соответствующих строк или столбцов. Если среди нулевых элементов матрицы (C) можно получить *допустимое* решение задачи, то оно является оптимальным. Напомним, что *допустимым решением* является такой выбор из нулей, при котором выбирается по одному нулю в каждой строке и по одному нулю в каждом столбце.

В рассматриваемом примере в каждой строке матрицы (C) нули есть (они появились в результате добавления фиктивных вакансий). Чтобы образовать нули в первых пяти столбцах матрицы $(-C)$, определяем минимальные элементы в этих столбцах: –8, –9, –8, –9, –9 и вычитаем эти элементы из соответствующих столбцов матрицы. В результате получим следующую матрицу (рис. 9.20):

$$C = \begin{pmatrix} 1 & 4 & 1 & 3 & 2 & 0 & 0 \\ 2 & 5 & 0 & 5 & 0 & 0 & 0 \\ 0 & 3 & 4 & 6 & 1 & 0 & 0 \\ 1 & 2 & 0 & 4 & 2 & 0 & 0 \\ 3 & 0 & 1 & 0 & 4 & 0 & 0 \\ 1 & 1 & 2 & 5 & 2 & 0 & 0 \\ 2 & 2 & 0 & 3 & 5 & 0 & 0 \end{pmatrix}.$$

Рис. 9.20

Так как из нулевых элементов нельзя получить допустимое решение (в первой и шестой строках, а также в четвертой и седьмой строках нули стоят на одном и том же месте), то алгоритм продолжается следующим образом:

а) минимальным количеством горизонтальных и вертикальных прямых вычеркиваем все нули;

б) среди невычеркнутых элементов находим минимальный элемент;

в) вычитаем минимальный элемент из всех невычеркнутых элементов;

г) к элементам, стоящим на пересечении вертикальных и горизонтальных прямых, прибавляем минимальный элемент.

Среди множества получаемых нулевых элементов определяем допустимое решение. Если допустимое решение найти нельзя, повторяем шаги а, б, в, г снова.

Процедура вычеркивания элементов и ее результат показаны на рис. 9.21. Минимальный среди невычеркнутых элементов равен единице. На рис. 9.22 показан результат после вычитания единицы из невычеркнутых элементов и прибавления единицы к элементам, стоящим на пересечении прямых. Допустимое решение соответствует отмеченным элементам.

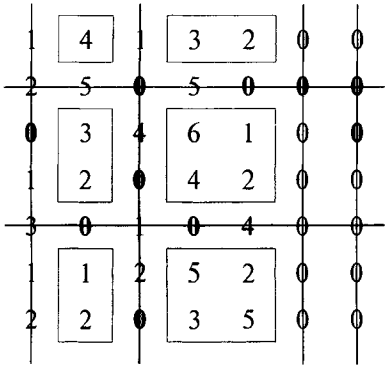


Рис. 9.21

1	3	1	2	1	**0**	0
3	5	1	5	**0**	1	1
0	2	4	5	0	0	0
1	1	**0**	3	1	0	0
4	0	2	**0**	4	1	1
1	**0**	2	4	1	0	0
2	1	0	2	4	0	**0**

Рис. 9.22

Перенеся полученное решение на исходную матрицу (C) (рис. 9.23), получим, что претенденты P_1 и P_7 попадают на фиктивные вакансии и не принимаются на работу. P_2 принимается на пятую вакансию, P_3 — на первую, P_4 — на третью, P_5 — на четвертую, P_6 — на вторую. Сумма баллов, полученная при данном решении равна: $9 + 8 + 8 + 9 + 8 = 42$.

7	5	7	6	7	**0**	0
6	4	8	4	**9**	0	0
8	6	4	3	8	0	0
7	7	**8**	5	7	0	0
5	9	7	**9**	5	0	0
6	**8**	6	4	7	0	0
7	7	8	6	4	0	**0**

Рис. 9.23

Решение задачи в процедуре EXCEL «Поиск решения»

1. *Ввод данных.* Вводим данные задачи в EXCEL, при этом нужно ввести 2 столбца (6- и 7-й) с нулевыми значениями для сбалансирования задачи. Результаты заполнения таблицы EXCEL можно увидеть на рис. 9.24.

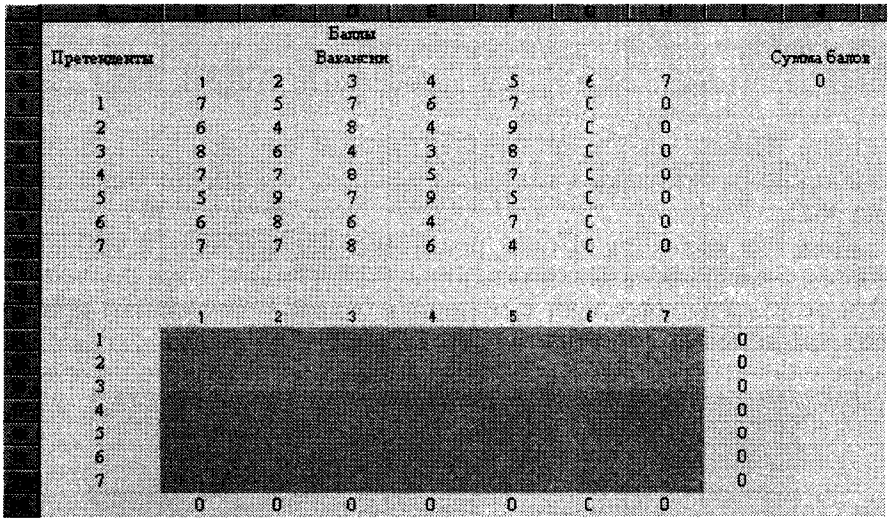


Рис. 9.24

В ячейках B4:F10 введены результаты тестирования претендентов, а в ячейках G4:H10 введены нули, что соответствует фиктивным вакансиям.

Ячейки B14:H20 являются изменяемыми ячейками для нашей процедуры.

В ячейках B21:H21 находятся суммы значений соответствующих столбцов изменяемых ячеек. Так, в ячейке B21 находится сумма ячеек B14:B20. Аналогично в ячейках:

в C21 находится сумма ячеек C14:C20;

в D21 находится сумма ячеек D14:D20;

в E21 находится сумма ячеек E14:E20;

в F21 находится сумма ячеек F14:F20.

в G21 находится сумма ячеек G14:G20;

в H21 находится сумма ячеек H14:H20.

В ячейках I14:I20 находятся суммы значений соответствующих строк изменяемых ячеек. Так, в ячейке I14 находится сумма ячеек B14:H14. Аналогично в ячейках:

в I15 находится сумма ячеек B15:H15;

в I16 находится сумма ячеек B16:H16;

в I17 находится сумма ячеек B17:H17;

в I18 находится сумма ячеек B18:H18;

в I19 находится сумма ячеек B19:H19;

в I20 находится сумма ячеек B20:H20.

Целевая функция заносится в ячейку J3 и вычисляется по формуле «СУММПРОИЗВ(B4:H10;B14:H20)».

2. *Заполнение окна процедуры «Поиск решения».*

Целевая функция: J3.

Значение целевой функции: max.

Изменяемые ячейки: B14:H20.

Ограничения задачи:

B21:H21 = 1 и I14:I20 = 1 (все свободные рабочие места должны быть заняты и все претенденты размещены по вакансиям);

B14:H20 ≥ 0 (изменяемые ячейки должны иметь положительные значения);

B14:H20 — двоичные числа.

В окне «Параметры» установить «Линейная модель», что соответствует решению задачи симплекс-методом. Результаты заполнения окна показаны на рис. 9.25.

3. *Выполнив процедуру* «Поиск решения», мы получили в первоначальной таблице следующие результаты (рис. 9.26).

Эти результаты совпадают с решением задачи, полученным преобразованием матрицы (*C*).

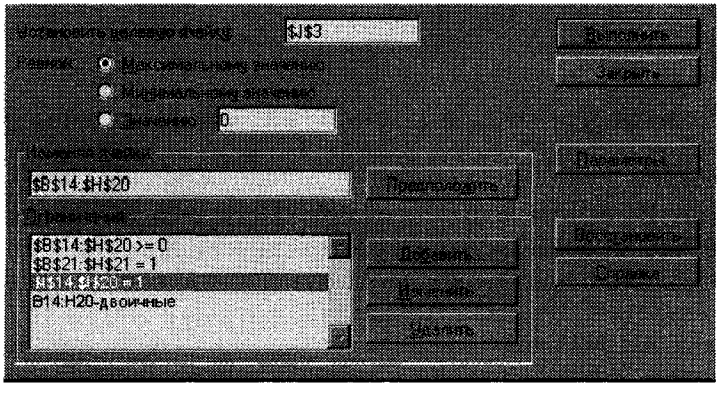


Рис. 9.25

	A	B	C	D	E	F	G	H	I	J
1				Баллы						
2	Претенденты			Вакансии						Сумма баллов
3		1	2	3	4	5	6	7		42
4	1	7	5	7	6	7	0	0		
5	2	6	4	8	4	9	0	0		
6	3	8	6	4	3	8	0	0		
7	4	7	7	8	5	7	0	0		
8	5	5	9	7	9	5	0	0		
9	6	6	8	6	4	7	0	0		
10	7	7	7	8	6	4	0	0		
11										
12										
13		1	2	3	4	5	6	7		
14	1	0	0	0	0	0	1	0	1	
15	2	0	0	0	0	1	0	0	1	
16	3	1	0	0	0	0	0	0	1	
17	4	0	0	1	0	0	0	0	1	
18	5	0	0	0	1	0	0	0	1	
19	6	0	1	0	0	0	0	0	1	
20	7	0	0	0	0	0	0	1	1	
21		1	1	1	1	1	1	1		

Рис. 9.26

9.4. Сетевые модели. Определение наикратчайшего пути между вершинами

Сеть (граф) состоит из множества вершин (узлов) и множества дуг (ребер), соединяющих вершины. Например, сеть на рис. 9.27 состоит из семи вершин и множества ориентированных дуг соединяющих вершины. Длины дуг могут в зависимости от задачи определять различные характеристики. Это может быть расстояние между вершинами, стоимость или время проезда из одной вершины в другую, пропускная способность дуги и т. д.

С помощью сетевых моделей можно поставить и решить большое число практически важных задач исследования операций.

Приведем несколько примеров:

- определение наикратчайшего пути между двумя пунктами на существующей сети дорог;
- проектирование сети кабельного телевидения, имеющей наименьшую стоимость;
- проектирование газопроводов имеющих максимальную пропускную способность;
- транспортная задача и ее модификации.

Все эти задачи можно сформулировать и решить как задачи линейного программирования. Тем не менее для решения сетевых моделей разработаны более эффективные вычислительные процедуры, учитывающие их специфику.

Пример 9.4. Определить наикратчайший путь между вершиной 1 и вершиной 7 на сети, представленной на рис. 9.27.

Для решения этой задачи в процедуре EXCEL «Поиск решения», представим ее как транспортную задачу с промежуточными пунктами. Будем считать, что транспортные расходы при перевозке одной единицы груза равны (в условных единицах) расстояниям между вершинами. Одна единица груза отправляется из вершины 1 (исходный пункт) и должна прибыть в вершину 7 (пункт назначения). Вершины 2, 3, 4, 5, 6 рассматриваются как промежуточные пункты, которые являются одновременно и исходными пунктами и пунктами назначения.

Требуется определить такую последовательность вершин, по которым должна перемещаться единица груза отправленная из вершины 1, при ко-

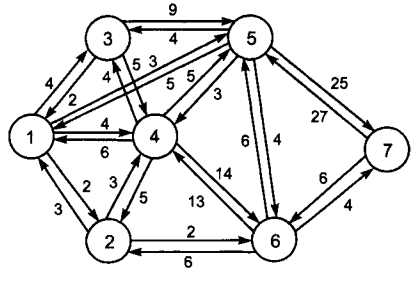


Рис. 9.27

торой стоимость транспортных расходов будет минимальна и груз попадет в вершину 7.

Так как транспортные расходы при перемещении груза из одной вершины в другую равны расстоянию между вершинами, то последовательность вершин при которой транспортные расходы будут минимальными определяет наикратчайший путь из вершины 1 в вершину 7. Матрица транспортных расходов, соответствующая данному графу представлена на рис. 9.28.

Исходные пункты	Пункты назначения						Количество груза, отправленного из пункта
	2	3	4	5	6	7	
1	2	4	4	3	*M*	*M*	1
2	0	*M*	3	*M*	2	*M*	0
3	*M*	0	5	9	*M*	*M*	0
4	5	4	0	5	14	*M*	0
5	*M*	4	3	0	4	25	0
6	6	*M*	13	6	0	4	0
Количество груза, прибывшего в пункт	0	0	0	0	0	1	

Рис. 9.28

Буквой *M* обозначается случай когда между соответствующими вершинами нет пути. В качестве *M* берут число, значительно большее длины самого большего пути. В данной задаче наибольший путь между 5- и 7-й вершинами, поэтому можно взять, например, $M = 50$. Для промежуточных пунктов 2, 3, 4, 5, 6 должны быть предусмотрены буферные емкости *B*. Буферная емкость должна быть не меньше, чем количество груза, которое перемещается в сети описываемой графом. В данной задаче $B = 1$. После введения буферных емкостей в первый столбец и нижнюю строку таблицы и замены $M = 50$, получим транспортную задачу, представляющую *задачу о назначениях* (рис. 9.29). Решим эту задачу методом преобразования матрицы транспортных расходов (см. пример 9.3).

Исходные пункты	Пункты назначения						Количество груза, отправленного из пункта
	2	3	4	5	6	7	
1	2	4	4	3	50	50	1
2	0	50	3	50	2	50	1
3	50	0	5	9	50	50	1
4	5	4	0	5	14	50	1
5	50	4	3	0	4	25	1
6	6	50	13	6	0	4	1
Количество груза, прибывшего в пункт	1	1	1	1	1	1	

Рис. 9.29

Последовательные преобразования матрицы транспортных расходов показаны на рис. 9.30, 9.31, 9.32.

2	4	4	3	50	50
0	50	3	50	2	50
50	0	5	9	50	50
5	4	0	5	14	50
50	4	3	0	4	25
6	50	13	6	0	4

Рис. 9.30

0	2	2	1	48	48
0	50	3	50	2	50
50	0	5	9	50	50
5	4	0	5	14	50
50	4	3	0	4	25
6	50	13	6	0	4

Рис. 9.31

0	2	2	1	48	44
0	50	3	50	2	46
~~50~~	~~0~~	~~5~~	~~9~~	~~50~~	~~46~~
~~5~~	~~4~~	~~0~~	~~5~~	~~14~~	~~46~~
~~50~~	~~4~~	~~3~~	~~0~~	~~4~~	~~21~~
~~6~~	~~50~~	~~13~~	~~6~~	~~0~~	~~0~~

Рис. 9.32

На рис. 9.31 показаны результаты вычитания минимального элемента первой строки (он равен 2) из первой строки, на рис. 9.32 приведены результаты вычитания минимального элемента из шестого столбца (он равен 4) и результат вычеркивания строк и столбцов с нулями. На рис. 9.33 показаны результаты вычитания из невычеркнутых элементов минимального элемента (он равен единице) и результат вычеркивания строк и столбцов второй раз.

0	1	1	0	47	43
0	49	2	49	1	45
~~51~~	~~0~~	~~5~~	~~9~~	~~50~~	~~46~~
~~6~~	~~4~~	~~0~~	~~5~~	~~14~~	~~46~~
51	4	3	0	4	21
~~7~~	~~50~~	~~13~~	~~6~~	~~0~~	~~0~~

Рис. 9.33

На рис. 9.34 приведены окончательные результаты преобразования и результаты допустимого выбора из множества нулей.

[0]	0	0	0	46	42
0	48	1	48	[0]	44
52	[0]	5	10	50	46
7	4	[0]	6	14	46
51	3	2	[0]	3	20
8	50	13	7	0	[0]

Рис. 9.34

Перенеся эти результаты на исходную таблицу (рис. 9.28), получим новую таблицу (рис. 9.35).

Исходные пункты	Пункты назначения						Количество груза, отправленного из пункта
	2	3	4	5	6	7	
1	2	4	4	3	*M*	*M*	1
2	0	*M*	3	*M*	2	*M*	0
3	*M*	0	5	9	*M*	*M*	0
4	5	4	0	5	14	*M*	0
5	*M*	4	3	0	4	25	0
6	6	*M*	13	6	0	4	0
Количество груза, прибывшего в пункт	0	0	0	0	0	1	

Рис. 9.35

Наикратчайший путь из вершины 1 в вершину 7 определяется следующей траекторией

$$1 \to 2 \to 6 \to 7.$$

Длина наикратчайшего пути равна: $2 + 2 + 4 = 8$.

Решение задачи в процедуре EXCEL «Поиск решения»

1. *Ввод данных.* Переносим данные задачи в EXCEL. Результаты заполнения таблицы EXCEL можно увидеть на рис. 9.36.

В ячейках B4:G9 введены длины путей из исходных пунктов в пункты назначения.

Ячейки B12:G17 являются изменяемыми ячейками для нашей процедуры.

В ячейках B18:G18 находятся суммы значений соответствующих столбцов изменяемых ячеек. Так в ячейке B18 находится сумма ячеек B12:B17. Аналогично в ячейках:

в C18 находится сумма ячеек C12:C17;

в D18 находится сумма ячеек D12:D17;

в E18 находится сумма ячеек E12:E17;

в F18 находится сумма ячеек F12:F17;

в G18 находится сумма ячеек G12:G17.

В ячейках H12:H17 находятся суммы значений соответствующих строк изменяемых ячеек. Так в ячейке H12 находится сумма ячеек B12:G12. Аналогично в ячейках:

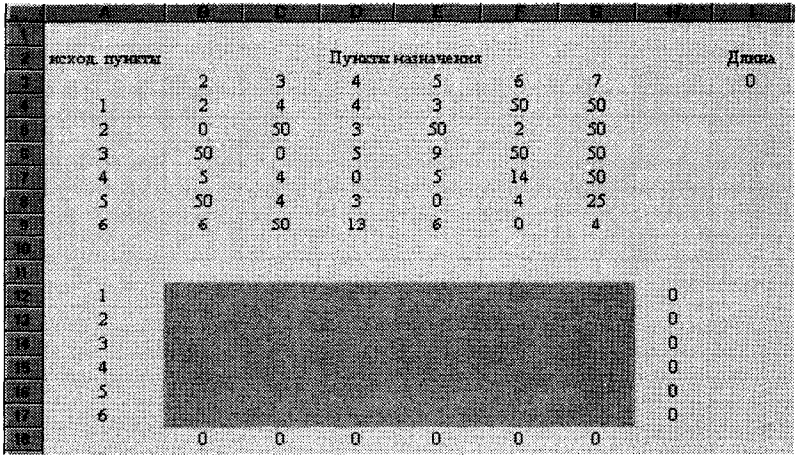
в H13 находится сумма ячеек B13:G13;

в H14 находится сумма ячеек B14:G14;

в H15 находится сумма ячеек B15:G15;

в H16 находится сумма ячеек B16:G16;

в H17 находится сумма ячеек B17:G17.



	A	B	C	D	E	F	H	I
1								
2	исход. пункты			Пункты назначения				Длина
3		2	3	4	5	6	7	0
4	1	2	4	4	3	50	50	
5	2	0	50	3	50	2	50	
6	3	50	0	5	9	50	50	
7	4	5	4	0	5	14	50	
8	5	50	4	3	0	4	25	
9	6	6	50	13	6	0	4	
10								
11								
12	1							0
13	2							0
14	3							0
15	4							0
16	5							0
17	6							0
18		0	0	0	0	0	0	

Рис. 9.36

Целевая функция заносится в ячейку I3 и вычисляется по формуле «СУММПРОИЗВ (B4:G9;B12:G17)».

2. *Заполнение окна процедуры «Поиск решения».*

Целевая функция: I3.

Значение целевой функции: min.

Изменяемые ячейки: B12:G17.

Ограничения задачи:

B18:G18 = 1 и H12:H17 = 1;

B12:G17 ≥ 0 (ячейки должны иметь положительные значения);

B12:G17 — двоичные числа.

В окне «Параметры» установить «Линейная модель», что соответствует решению задачи симплекс-методом. Результаты заполнения окна показаны на рис. 9.37.

3. *Выполнив процедуру «Поиск решения»*, в первоначальной таблице (рис. 9.36) получим следующие результаты (рис. 9.38).

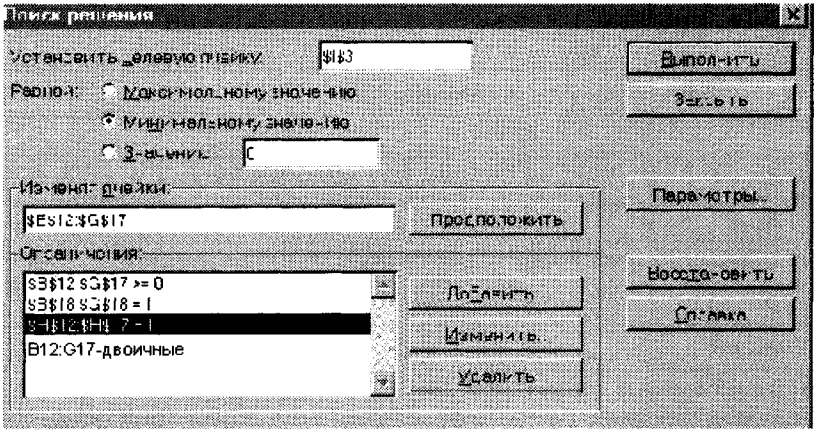
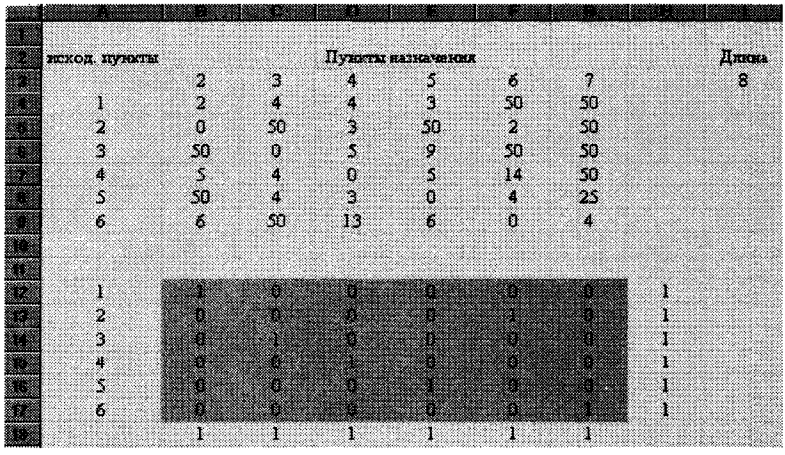


Рис. 9.37



	A	B	C	D	E	F	G	H
1								
2	исход. пункты			Пункты назначения				Длина
3		2	3	4	5	6	7	8
4	1	2	4	4	3	50	50	
5	2	0	50	3	50	2	50	
6	3	50	0	5	9	50	50	
7	4	5	4	0	5	14	50	
8	5	50	4	3	0	4	25	
9	6	6	50	13	6	0	4	
10								
11								
12	1	1	0	0	0	0	0	1
13	2	0	0	0	0	1	0	1
14	3	0	1	0	0	0	0	1
15	4	0	0	1	0	0	0	1
16	5	0	0	0	1	0	0	1
17	6	0	0	0	0	0	1	1
18		1	1	1	1	1	1	

Рис. 9.38

Эти результаты совпадают с решением данной задачи при помощи преобразования матрицы транспортных расходов: $1 \rightarrow 2 \rightarrow 6 \rightarrow 7$, длина = 8.

9.5. Варианты заданий по курсу «Исследование операций»

1. Варианты для задачи распределения ресурсов

Задание 1. Составьте математическую модель и решите задачу графическим методом.

Задание 2. Найдите решение задачи, используя симплекс-метод.

Задание 3. Решите задачу в пакете EXCEL и сравните результаты.

Вариант № 1

Предприятие, располагающее ресурсами сырья четырех видов *A*, *B*, *C* и *D*, может производить продукцию двух видов P_1, P_2. В таблице указаны затраты ресурсов на изготовление 1 т продукции, объем ресурсов и прибыль, получаемая от продажи 1 т соответствующей продукции.

Вид сырья	Вид продукции		Объем ресурсов, т
	P_1	P_2	
A	4	1	7
B	1	2	10
C	3	1	6
D	6	1	10
Прибыль, руб.	7	2	

Определить ассортимент выпускаемой продукции, при котором полученная прибыль будет максимальной.

Вариант № 2

Для изготовления двух видов изделий *A* и *B* завод использует в качестве сырья алюминий и медь. На изготовление изделий заняты токарные и фрезерные станки. Исходные данные задачи приведены в таблице:

Вид ресурсов	Объем ресурсов	Нормы расходов на 1 изделие	
		A	*B*
Алюминий, кг	4	0	1
Медь, кг	7	4	1
Токарные станки, станко-час	5	2	1
Фрезерные станки, станко-час	10	6	1
Прибыль на 1 изделие, тыс. руб.		4	3

Определить ассортимент выпускаемой продукции, при котором полученная прибыль будет максимальной.

Вариант № 3

Фирма производит два вида продуктов K_1 и K_2. Для изготовления продуктов применяются машины *A*, *B*, *C* и *D*. Время необходимое для изготовления продуктов K_1 и K_2 на равных машинах, допустимое время использования машин, а также прибыль от продажи продуктов приведены в таблице:

Машины	Допустимое время (в часах)	Необходимое время (в часах)	
		K_1	K_2
A	10	5	1
B	9	4	2
C	5	1	2
D	7	1	3
Прибыль от продажи продуктов, тыс. руб.		5	2

Какое количество каждого продукта необходимо произвести, чтобы прибыль была максимальной?

Вариант № 4

Для изготовления двух видов изделий *A* и *B* завод использует в качестве сырья алюминий и медь. На изготовление изделий заняты токарные и фрезерные станки. Исходные данные задачи приведены в таблице:

Вид ресурсов	Объем ресурсов	Нормы расходов на 1 изделие	
		A	*B*
Алюминий, кг	2	1	0
Медь, кг	6	1	1
Токарные станки, станко-час	7	2	1
Фрезерные станки, станко-час	10	4	1
Прибыль на 1 изделие, тыс. руб.		3	2

Определить ассортимент выпускаемой продукции, при котором полученная прибыль будет максимальной.

Вариант № 5

Фирма производит два вида продуктов K_1 и K_2. Для изготовления продуктов применяются машины *A*, *B*, *C* и *D*. Время необходимое для изготовления продуктов K_1 и K_2 на равных машинах, допустимое время использования машин, а также прибыль от продажи продуктов приведены в таблице:

Машины	Допустимое время (в часах)	Необходимое время (в часах)	
		K_1	K_2
A	4	0	1
B	7	4	1
C	5	2	1
D	10	6	1
Прибыль от продажи продуктов, тыс. руб.		10	4

Какое количество каждого продукта необходимо произвести, чтобы прибыль была максимальной?

Вариант № 6

Предприятие, располагающее ресурсами сырья четырех видов *A*, *B*, *C* и *D*, может производить продукцию двух видов P_1, P_2. В таблице указаны затраты ресурсов на изготовление 1 т продукции, объем ресурсов и прибыль, получаемая от продажи 1 т соответствующей продукции.

Вид сырья	Вид продукции		Объем ресурсов, т
	P_1	P_2	
A	0	1	5
B	1	0	4
C	2	1	9
D	2	1	6
Прибыль, руб.	2	5	

Определить ассортимент выпускаемой продукции, при котором полученная прибыль будет максимальной.

Вариант № 7

Для изготовления двух видов изделий A и B завод использует в качестве сырья алюминий и медь. На изготовление изделий заняты токарные и фрезерные станки. Исходные данные задачи приведены в таблице:

Вид ресурсов	Объем ресурсов	Нормы расходов на 1 изделие	
		A	B
Алюминий, кг	12	3	2
Медь, кг	20	1	4
Токарные станки, станко-час	7	2	1
Фрезерные станки, станко-час	3	1	0
Прибыль на 1 изделие, тыс. руб.		4	1

Определить ассортимент выпускаемой продукции, при котором полученная прибыль будет максимальной.

Вариант № 8

Фирма производит два вида продуктов K_1 и K_2. Для изготовления продуктов применяются машины A, B, C и D. Время необходимое для изготовления продуктов K_1 и K_2 на равных машинах, допустимое время использования машин, а также прибыль от продажи продуктов приведены в таблице:

Машины	Допустимое время (в часах)	Необходимое время (в часах)	
		K_1	K_2
A	4	0	1
B	5	2	1
C	7	4	1
D	3	2	0
Прибыль от продажи продуктов, тыс. руб.		2	3

Какое количество каждого продукта необходимо произвести, чтобы прибыль была максимальной?

Вариант № 9

Предприятие, располагающее ресурсами сырья четырех видов A, B, C и D, может производить продукцию двух видов P_1, P_2. В таблице указаны за-

траты ресурсов на изготовление 1 т продукции, объем ресурсов и прибыль, получаемая от продажи 1 т соответствующей продукции.

Вид сырья	Вид продукции		Объем ресурсов, т
	P_1	P_2	
A	0	1	4
B	2	1	5
C	4	1	7
D	2	0	3
Прибыль, руб.	6	2	

Определить ассортимент выпускаемой продукции, при котором полученная прибыль будет максимальной.

Вариант № 10

Для изготовления двух видов изделий *A* и *B* завод использует в качестве сырья алюминий и медь. На изготовление изделий заняты токарные и фрезерные станки. Исходные данные задачи приведены в таблице:

Вид ресурсов	Объем ресурсов	Нормы расходов на 1 изделие	
		A	*B*
Алюминий, кг	4	0	1
Медь, кг	5	2	1
Токарные станки, станко-час	7	4	1
Фрезерные станки, станко-час	3	2	0
Прибыль на 1 изделие, тыс. руб.		6	1

Определить ассортимент выпускаемой продукции, при котором полученная прибыль будет максимальной.

Вариант № 11

Компания производит паруса двух размеров *A* и *B* для небольших яхт. Агенты по продаже считают, что в один день на рынке может быть реализовано до 10 парусов. Для каждого паруса *A* требуется 2 м^2 материала, а для паруса *B* — 3 м^2 материала. Компания может получить 12 м^2 материала в день. Для изготовления паруса *A* требуется 14 мин машинного времени, а для изготовления паруса *B* — 50 мин. ЭВМ можно использовать 8 ч в день. Прибыль от продажи паруса типа *A* составляет 6 руб., а от продажи паруса типа *B* — 12 руб. Сколько парусов каждого типа следует выпускать в день?

Вариант № 12

Для изготовления двух видов продукции P_1 и P_2 используется четыре вида сырья: *A*, *B*, *C* и *D*. Запасы сырья ограничены: *B* — 6 т, *C* — 8 т, *D* — 5 т, *A* — 12 т. Нормы расхода сырья на изготовление 1 т продукции приведены в таблице. Составить такой план выпуска продукции, чтобы при ее реализации получить максимальную прибыль.

Вид сырья	Колич. сырья на 1 т продукции	
	P_1	P_2
A	2	3
B	1	1
C	2	1
D	0	1
Цена за тонну, руб.	2	5

Вариант № 13

Фирма производит два продукта *A* и *B*, рынок сбыта которых не ограничен. Каждый продукт должен быть обработан машинами 1, 2 и 3. Время обработки для каждого из изделий *A* и *B* приведено ниже:

Продукт	Машина		
	1	2	3
A	5	4	2
B	6	3	4

Время работы машин 1, 2, 3 соответственно 35, 32 и 40 ч в неделю. Прибыль от изделий *A* и *B* составляет соответственно 5 и 7 руб.

Фирме необходимо определить недельные нормы выпуска изделий *A* и *B* и рассчитать максимальную прибыль.

Вариант № 14

Фирма выпускает два вида продукции. В процессе производства используются три технологические операции. При изготовлении 2-го изделия технологическая операция № 2 не выполняется. Время выполнения операции (в часах) приводится в таблице.

Изделие	Технологические операции		
	1	2	3
1	1	3	1
2	2	—	4

Фонд рабочего времени ограничен:
для первой операции — 12 ч;
для второй операции — 9 ч;
для третьей операции — 6 ч;
Изучение рынка показало, что ожидаемая прибыль от продажи одного изделия видов 1 и 2 соответственно равна 4 и 7 руб.
Каков наиболее выгодный суточный объем производства каждого вида продукции?

Вариант № 15

Фирма выпускает два вида продукции. В процессе производства используются три технологические операции. При изготовлении 2-го изделия технологическая операция № 2 не выполняется. Время выполнения операции (в часах) приводится в таблице.

Изделие	Технологические операции		
	1	2	3
1	1	3	1
2	2	—	4

Фонд рабочего времени ограничен:
для первой операции — 12 ч;
для второй операции — 6 ч;
для третьей операции — 9 ч;
Изучение рынка показало, что ожидаемая прибыль от продажи одного изделия видов 1 и 2 соответственно равна 2 и 7 руб.
Каков наиболее выгодный суточный объем производства каждого вида продукции?

Вариант № 16

Фирма выпускает два вида продукции. В процессе производства используются три технологические операции. При изготовлении 2-го изделия технологическая операция № 2 не выполняется. Время выполнения операции (в часах) приводится в таблице.

Изделие	Технологические операции		
	1	2	3
1	1	3	1
2	2	—	4

Фонд рабочего времени ограничен:
для первой операции — 8 ч;
для второй операции — 5 ч;
для третьей операции — 10 ч;

Изучение рынка показало, что ожидаемая прибыль от продажи одного изделия видов 1 и 2 соответственно равна 4 и 9 руб.

Каков наиболее выгодный суточный объем производства каждого вида продукции?

Вариант № 17

Для изготовления двух видов продукции P_1 и P_2 используется четыре вида сырья: A, B, C и D. Запасы сырья ограничены: B — 9 т, C — 10 т, D — 8 т, A — 6 т. Нормы расхода сырья на изготовление 1 т продукции приведены в таблице. Составить такой план выпуска продукции, чтобы при ее реализации получить максимальную прибыль.

Вид сырья	Количество сырья на 1 т продукции	
	P_1	P_2
A	1	1
B	2	1
C	1	2
D	1	4
Цена за тонну, руб.	2	3

Вариант № 18

Фирма выпускает два вида продукции. В процессе производства используются три технологические операции. При изготовлении 2-го изделия технологическая операция № 2 не выполняется. Время выполнения операции (в часах) приводится в таблице.

Изделие	Технологические операции		
	1	2	3
1	1	3	1
2	2	—	4

Фонд рабочего времени ограничен:

для первой операции — 8 ч;

для второй операции — 9 ч;

для третьей операции — 14 ч;

Изучение рынка показало, что ожидаемая прибыль от продажи одного изделия видов 1 и 2 соответственно равна 5 и 3 руб.

Каков наиболее выгодный суточный объем производства каждого вида продукции?

Вариант № 19

Фирма производит два продукта A и B, рынок сбыта которых не ограничен. Каждый продукт должен быть обработан машинами 1, 2 и 3. Время обработки для каждого из изделий A и B приведено ниже:

Продукт	Машина		
	1	2	3
A	5	4	2
B	6	3	4

Время работы машин 1, 2, 3 соответственно 42, 38 и 38 ч в неделю. Прибыль от изделий A и B составляет соответственно 8 и 6 руб.

Фирме необходимо определить недельные нормы выпуска изделий A и B и рассчитать максимальную прибыль.

Вариант № 20

Фирма производит два продукта A и B, рынок сбыта которых не ограничен. Каждый продукт должен быть обработан машинами 1, 2 и 3. Время обработки для каждого из изделий A и B приведено ниже:

Продукт	Машина		
	1	2	3
A	5	4	2
B	6	3	4

Время работы машин 1, 2, 3 соответственно 39, 29 и 26 ч в неделю. Прибыль от изделий A и B составляет соответственно 6 и 4 руб.

Фирме необходимо определить недельные нормы выпуска изделий A и B и рассчитать максимальную прибыль.

Вариант № 21

Для изготовления двух видов продукции P_1 и P_2 используется четыре вида сырья: A, B, C и D. Запасы сырья ограничены: B — 4 т, C — 9 т, D — 5 т, A — 10 т. Нормы расхода сырья на изготовление 1 т продукции приведены в таблице. Составить такой план выпуска продукции, чтобы при ее реализации получить максимальную прибыль.

Вид сырья	Вид сырья на 1 т продукции	
	P_1	P_2
A	2	3
B	1	1
C	1	2
D	1	5
Прибыль, руб.	3	6

Вариант № 22

Фирма производит два продукта A и B, рынок сбыта которых не ограничен. Каждый продукт должен быть обработан машинами 1, 2 и 3. Время обработки для каждого из изделий A и B приведено ниже:

Продукт	Машина		
	1	2	3
A	5	4	2
B	6	3	4

Время работы машин 1, 2, 3 соответственно 25, 12 и 14 ч в неделю. Прибыль от изделий A и B составляет соответственно 8 и 5 руб.

Фирме необходимо определить недельные нормы выпуска изделий A и B и рассчитать максимальную прибыль.

Вариант № 23

Предприятие, располагающее ресурсами сырья четырех видов A, B, C и D, может производить продукцию двух видов P_1, P_2. В таблице указаны затраты ресурсов на изготовление 1 т продукции, объем ресурсов и прибыль, получаемая от изготовления 1 т соответствующей продукции.

Вид сырья	Вид продукции		Объем ресурсов, т
	P_1	P_2	
A	1	2	6
B	2	1	7
C	3	1	10
D	0	1	2
Прибыль, руб.	7	3	

Определить ассортимент выпускаемой продукции, при котором полученная прибыль будет максимальной.

Вариант № 24

Фирма производит два продукта A и B, рынок сбыта которых не ограничен. Каждый продукт должен быть обработан машинами 1, 2 и 3. Время обработки для каждого из изделий A и B приведено ниже:

Продукт	Машина		
	1	2	3
A	5	4	2
B	6	3	4

Время работы машин 1, 2, 3 соответственно 40, 20 и 18 ч в неделю. Прибыль от изделий A и B составляет соответственно 7 и 4 руб.

Вариант № 25

Компания производит паруса двух размеров A и B для небольших яхт. Агенты по продаже считают, что в один день на рынке может быть реализовано до 10 парусов. Для каждого паруса A требуется 2 м^2 материала, а для паруса B — 3 м^2 материала. Компания может получить 12 м^2 материала в день. Для изготовления паруса A требуется 14 мин машинного времени, а для изготовления паруса B — 26 мин ЭВМ можно использовать 8 ч в день. Прибыль от продажи паруса типа A составляет 6 руб., а от продажи паруса типа B — 8 руб. Сколько парусов каждого типа следует выпускать в день?

Вариант № 26

Фирма производит два вида продуктов K_1 и K_2. Для изготовления продуктов применяются машины A, B, C и D. Время необходимое для изготовления продуктов K_1 и K_2 на равных машинах, допустимое время использования машин, а также прибыль от продажи продуктов приведены в таблице:

Машины	Допустимое время (в часах)	Необходимое время (в часах)	
		K_1	K_2
A	9	1	2
B	5	1	1
C	7	2	1
D	10	3	1
Прибыль от продажи продуктов, тыс. руб.		4	3

Какое количество каждого продукта необходимо произвести, чтобы прибыль была максимальной?

Вариант № 27

Компания производит паруса двух размеров A и B для небольших яхт. Агенты по продаже считают, что в один день на рынке может быть реализовано до 10 парусов. Для каждого паруса A требуется 2 м^2 материала, а для паруса B — 3 м^2 материала. Компания может получить 12 м^2 материала в день. Для изготовления паруса A требуется 10 мин машинного времени, а для изготовления паруса B — 20 мин. ЭВМ можно использовать 8 ч в день. Прибыль от продажи паруса типа A составляет 8 руб., а от продажи паруса типа B — 3 руб. Сколько парусов каждого типа следует выпускать в день?

Вариант № 28

Компания производит паруса двух размеров A и B для небольших яхт. Агенты по продаже считают, что в один день на рынке может быть реализовано до 10 парусов. Для каждого паруса A требуется 2 м^2 материала, а для

паруса B — 3 м^2 материала. Компания может получить 12 м^2 материала в день. Для изготовления паруса A требуется 14 мин машинного времени, а для изготовления паруса B — 16 мин. ЭВМ можно использовать 8 ч в день. Прибыль от продажи паруса типа A составляет 7 руб., а от продажи паруса типа B — 9 руб. Сколько парусов каждого типа следует выпускать в день?

Вариант № 29

Компания производит паруса двух размеров A и B для небольших яхт. Агенты по продаже считают, что в один день на рынке может быть реализовано до 15 парусов. Для каждого паруса A требуется 3 м^2 материала, а для паруса B — 4 м^2 материала. Компания может получить 20 м^2 материала в день. Для изготовления паруса A требуется 14 мин машинного времени, а для изготовления паруса B — 16 мин. ЭВМ можно использовать 8 ч в день. Прибыль от продажи паруса типа A составляет 7 руб., а от продажи паруса типа B — 9 руб. Сколько парусов каждого типа следует выпускать в день?

Вариант № 30

Фирма производит два продукта A и B, рынок сбыта которых не ограничен. Каждый продукт должен быть обработан машинами 1, 2 и 3. Время обработки для каждого из изделий A и B приведено ниже:

Продукт	Машина		
	1	2	3
A	5	4	2
B	6	3	4

Время работы машин 1, 2, 3 соответственно 33, 26 и 26 ч в неделю. Прибыль от изделий A и B составляет соответственно 10 и 7 руб.

Фирме необходимо определить недельные нормы выпуска изделий A и B и рассчитать максимальную прибыль.

2. Варианты для транспортной задачи

Продукция определенного типа производится в городах A_1, A_2, A_3 и потребляется в городах B_1, B_2, B_3 и B_4.

В таблице указаны: объем производства, спрос, стоимость перевозки единицы продукции.

Составить оптимальный план перевозки продукции, при котором стоимость всех перевозок будет минимальна.

Предварительно следует проверить, сбалансирована ли данная транспортная задача. Если задача не сбалансирована, то нужно ввести фиктивных потребителей или производителей, добавляя к исходной таблице столбцы или строки.

Вариант 1

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	20	47	31	13	49
A_2	3	38	44	10	18
A_3	11	32	46	17	68
Спрос	45	30	10	45	

Вариант 2

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	47	31	13	45	34
A_2	20	47	31	13	44
A_3	4	42	41	2	68
Спрос	30	45	41	80	

Вариант 3

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	31	13	45	35	48
A_2	38	44	10	33	48
A_3	20	47	31	13	44
Спрос	40	41	45	44	

Вариант 4

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	13	45	35	7	49
A_2	47	31	13	45	47
A_3	32	46	17	27	68
Спрос	45	80	44	45	

Вариант 5

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	45	35	7	43	48
A_2	44	10	33	46	41
A_3	42	41	2	38	49
Спрос	44	12	88	44	

Вариант 6

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	35	7	43	39	45
A_2	31	13	45	35	33
A_3	47	31	13	45	19
Спрос	6	10	30	41	

Вариант 7

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	7	43	39	10	41
A_2	10	33	46	16	22
A_3	46	17	27	47	61
Спрос	38	30	19	87	

Вариант 8

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	43	39	10	40	34
A_2	13	45	35	7	18
A_3	41	2	38	44	86
Спрос	48	45	5	30	

Вариант 9

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	39	10	40	43	26
A_2	33	46	16	28	18
A_3	31	13	45	35	58
Спрос	15	50	10	22	

Вариант 10

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	10	40	43	6	16
A_2	45	35	7	43	27
A_3	17	27	47	23	68
Спрос	31	44	24	42	

Вариант 11

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	40	43	6	36	15
A_2	46	16	28	47	39
A_3	2	38	44	9	71
Спрос	50	28	36	1	

Вариант 12

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	43	6	36	45	14
A_2	35	7	43	39	48
A_3	13	45	35	7	22
Спрос	23	16	45	10	

Вариант 13

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	6	36	45	13	24
A_2	16	28	47	22	52
A_3	27	47	23	22	85
Спрос	24	18	49	20	

Вариант 14

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	36	45	13	31	34
A_2	7	43	39	10	52
A_3	38	44	9	34	81
Спрос	50	38	49	80	

Вариант 15

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	45	13	31	46	42
A_2	28	47	22	23	47
A_3	45	35	7	43	72
Спрос	30	49	44	88	

Вариант 16

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	13	31	46	19	49
A_2	43	39	10	40	88
A_3	47	23	22	47	58
Спрос	17	48	35	45	

Вариант 17

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	31	46	19	26	58
A_2	47	22	23	47	24
A_3	44	9	34	46	78
Спрос	49	36	21	49	

Вариант 18

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	46	19	26	47	54
A_2	39	10	40	43	19
A_3	35	7	43	39	44
Спрос	36	15	6	50	

Вариант 19

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	19	26	47	25	52
A_2	22	23	47	28	13
A_3	23	22	47	29	12
Спрос	10	19	10	48	

Вариант 20

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	26	47	25	20	48
A_2	10	40	43	6	28
A_3	9	34	46	15	71
Спрос	47	81	25	44	

Вариант 21

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	47	25	20	47	41
A_2	23	47	28	17	41
A_3	7	43	39	10	79
Спрос	40	46	88	37	

Вариант 22

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	25	20	47	30	32
A_2	40	43	6	36	49
A_3	22	47	29	16	46
Спрос	13	50	46	28	

Вариант 23

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	20	47	30	14	22
A_2	47	28	17	46	58
A_3	34	46	15	29	78
Спрос	43	42	50	18	

Вариант 24

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	47	30	14	45	10
A_2	43	6	36	45	61
A_3	43	39	10	40	60
Спрос	44	23	48	6	

Вариант 25

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	30	14	45	35	10
A_2	28	17	46	33	44
A_3	47	29	16	46	41
Спрос	15	43	41	6	

Вариант 26

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	14	45	35	7	22
A_2	6	36	45	13	83
A_3	46	15	29	47	56
Спрос	39	24	30	18	

Вариант 27

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	45	35	7	43	83
A_2	17	46	33	10	18
A_3	39	10	40	43	82
Спрос	47	42	15	29	

Вариант 28

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	4	5	6	10	530
A_2	8	6	3	8	405
A_3	7	10	4	11	540
Спрос	425	415	335	400	

Вариант 29

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	3	7	4	8	513
A_2	9	6	4	4	448
A_3	6	10	5	8	522
Спрос	437	417	333	396	

Вариант 30

Производители	Потребители				Объем производства
	B_1	B_2	B_3	B_4	
A_1	35	30	10	10	53
A_2	21	41	53	10	28
A_3	39	32	27	20	61
Спрос	47	38	23	24	

3. Варианты для задач о назначениях

В конкурсе на занятие пяти вакансий (V_1, V_2, V_3, V_4, V_5) участвуют семь претендентов (P_1, P_2, P_3, P_4, P_5, P_6, P_7). Результаты тестирования каждого претендента, в случае занятия им одной вакансии, даны в виде матрицы — C (тестирование производилось по десятибалльной системе).

Определить, какого претендента и на какую вакансию следует принять, причем так, чтобы сумма баллов оказалась максимальной.

Вариант № 1

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	4	6	3	6
P_2	4	5	3	6	7
P_3	8	7	8	5	8
P_4	5	5	6	8	4
P_5	8	7	5	4	9
P_6	9	5	6	6	4
P_7	5	8	8	6	5

Вариант № 2

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	4	6	7	6
P_2	7	5	6	5	7
P_3	7	7	8	9	6
P_4	6	4	3	8	6
P_5	8	4	9	6	4
P_6	4	8	6	5	7
P_7	4	8	7	8	9

Вариант № 3

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	4	7	5	7	7
P_2	6	7	3	6	8
P_3	3	4	5	8	7
P_4	7	9	8	5	6
P_5	4	6	4	8	4
P_6	5	4	9	6	5
P_7	7	5	4	7	4

Вариант № 4

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	5	7	5	8	6
P_2	7	4	6	4	8
P_3	5	8	7	8	7
P_4	4	9	4	6	5
P_5	7	8	3	5	8
P_6	9	5	7	6	7
P_7	7	6	6	5	7

Вариант № 5

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	4	6	3	6
P_2	4	5	3	6	7
P_3	5	7	8	6	5
P_4	8	7	7	8	5
P_5	5	6	8	7	9
P_6	9	5	6	8	6
P_7	8	5	6	4	7

Вариант № 6

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	7	5	7	6	7
P_2	6	4	8	4	9
P_3	8	6	4	3	8
P_4	7	7	8	5	7
P_5	5	9	7	9	5
P_6	7	8	6	4	7
P_7	6	7	8	6	4

Вариант № 7

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	7	5	7	4	7
P_2	6	4	8	6	5
P_3	7	5	6	9	5
P_4	9	5	5	8	7
P_5	5	7	7	5	8
P_6	5	6	8	6	6
P_7	8	9	6	4	7

Вариант № 8

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	4	6	3	6
P_2	4	5	3	6	7
P_3	6	7	8	5	5
P_4	8	6	5	8	8
P_5	6	8	4	6	9
P_6	9	5	6	8	6
P_7	5	7	5	7	7

Вариант № 9

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	5	6	4	6
P_2	4	7	4	7	6
P_3	8	4	7	3	7
P_4	6	6	7	7	4
P_5	5	9	4	6	4
P_6	6	4	9	4	5
P_7	4	6	5	4	9

Вариант № 10

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	8	5	4	9
P_2	7	8	7	4	8
P_3	5	4	6	8	7
P_4	4	7	8	6	6
P_5	3	6	7	8	6
P_6	7	8	4	7	8
P_7	5	7	6	8	5

Вариант № 11

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	5	6	4	6
P_2	4	7	4	7	6
P_3	8	4	7	3	7
P_4	6	6	7	7	4
P_5	5	9	4	6	4
P_6	6	4	9	4	5
P_7	4	6	5	4	9

Вариант № 12

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	4	6	3	6
P_2	4	5	3	6	7
P_3	6	7	8	5	5
P_4	8	6	5	8	8
P_5	6	8	4	6	9
P_6	9	5	6	8	6
P_7	5	7	5	7	7

Вариант № 13

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	7	5	7	4	7
P_2	6	4	8	6	5
P_3	7	5	6	9	5
P_4	9	5	5	8	7
P_5	5	7	7	5	8
P_6	5	6	8	6	6
P_7	8	9	6	4	7

Вариант № 14

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	7	5	7	6	7
P_2	6	4	8	4	9
P_3	8	6	4	3	8
P_4	7	7	8	5	7
P_5	5	9	7	9	5
P_6	7	8	6	4	7
P_7	6	7	8	6	4

Вариант № 15

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	4	6	3	6
P_2	4	5	3	6	7
P_3	5	7	8	6	5
P_4	8	7	7	8	5
P_5	5	6	8	7	9
P_6	9	5	6	8	6
P_7	8	5	6	4	7

Вариант № 16

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	5	6	8	4	6
P_2	8	9	4	7	4
P_3	4	5	7	6	8
P_4	6	8	5	7	9
P_5	5	7	6	8	5
P_6	4	8	5	8	6
P_7	7	6	5	7	8

Вариант № 17

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	5	7	5	8	6
P_2	7	4	6	4	8
P_3	5	8	7	8	7
P_4	4	9	4	6	5
P_5	7	8	3	5	8
P_6	9	5	7	6	7
P_7	7	6	6	5	7

Вариант № 18

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	4	7	5	7	7
P_2	6	7	3	6	8
P_3	3	4	5	8	7
P_4	7	9	8	5	6
P_5	4	6	4	8	4
P_6	5	4	9	6	5
P_7	7	5	4	7	4

Вариант № 19

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	4	6	7	6
P_2	7	5	6	5	7
P_3	7	7	8	9	6
P_4	6	4	3	8	6
P_5	8	4	9	6	4
P_6	4	8	6	5	7
P_7	4	8	7	8	9

Вариант № 20

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	8	5	4	9
P_2	7	8	7	4	8
P_3	5	4	6	8	7
P_4	4	7	8	6	6
P_5	3	6	7	8	6
P_6	7	8	4	7	8
P_7	5	7	6	8	5

Вариант № 21

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	5	6	4	6
P_2	4	7	4	7	6
P_3	8	4	7	3	7
P_4	6	6	7	7	4
P_5	5	9	4	6	4
P_6	6	4	9	4	5
P_7	4	6	5	4	9

Вариант № 22

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	4	6	3	6
P_2	4	5	3	6	7
P_3	6	7	8	5	5
P_4	8	6	5	8	8
P_5	6	8	4	6	9
P_6	9	5	6	8	6
P_7	5	7	5	7	7

Вариант № 23

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	7	5	7	4	7
P_2	6	4	8	6	5
P_3	7	5	6	9	5
P_4	9	5	5	8	7
P_5	5	7	7	5	8
P_6	5	6	8	6	6
P_7	8	9	6	4	7

Вариант № 24

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	7	5	7	6	7
P_2	6	4	8	4	9
P_3	8	6	4	3	8
P_4	7	7	8	5	7
P_5	5	9	7	9	5
P_6	7	8	6	4	7
P_7	6	7	8	6	4

Вариант № 25

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	4	6	3	6
P_2	4	5	3	6	7
P_3	5	7	8	6	5
P_4	8	7	7	8	5
P_5	5	6	8	7	9
P_6	9	5	6	8	6
P_7	8	5	6	4	7

Вариант № 26

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	5	5	6	9
P_2	7	8	7	4	8
P_3	5	4	6	8	7
P_4	4	7	8	6	6
P_5	3	6	7	8	7
P_6	7	9	4	5	8
P_7	5	7	6	8	5

Вариант № 27

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	5	7	5	8	6
P_2	7	4	6	4	8
P_3	5	8	7	8	7
P_4	4	9	4	6	5
P_5	7	8	3	5	8
P_6	9	5	7	6	7
P_7	7	6	6	5	7

Вариант № 28

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	4	7	5	7	7
P_2	6	7	3	6	8
P_3	3	4	5	8	7
P_4	7	9	8	5	6
P_5	4	6	4	8	4
P_6	5	4	9	6	5
P_7	7	5	4	7	4

Вариант № 29

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	4	6	7	6
P_2	7	5	6	5	7
P_3	7	7	8	9	6
P_4	6	4	3	8	6
P_5	8	4	9	6	4
P_6	4	8	6	5	7
P_7	4	8	7	8	9

Вариант № 30

$C =$

	V_1	V_2	V_3	V_4	V_5
P_1	6	8	5	4	9
P_2	7	8	7	4	8
P_3	5	4	6	8	7
P_4	4	7	8	6	6
P_5	3	6	7	8	6
P_6	7	8	4	7	8
P_7	5	7	6	8	5

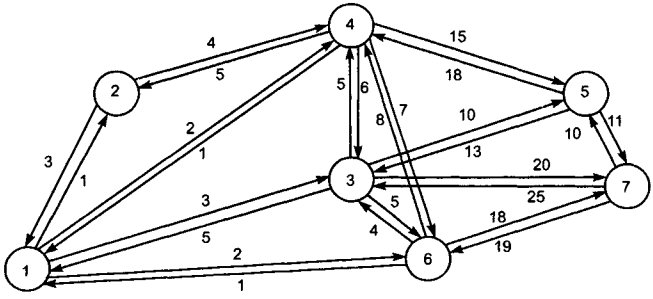
4. Варианты задач на сетях

Задания.

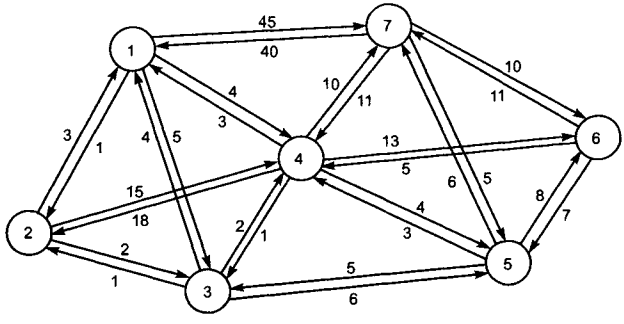
1. Представить задачу об определении наикратчайшего пути как транспортную с промежуточными пунктами. Составить матрицу задачи и решить задачу как транспортную, используя процедуру поиска решения пакета Excel.

2. Рассмотреть задачу из п. 1 как задачу о назначениях и решить эту задачу используя преобразование матрицы стоимости.

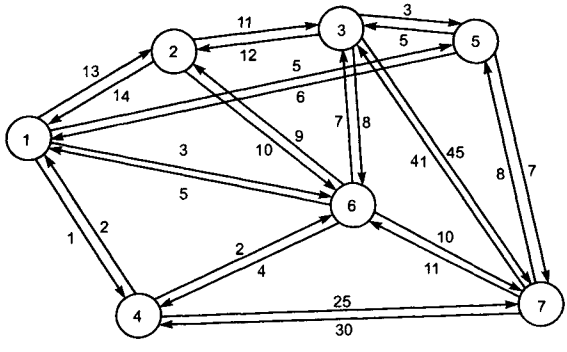
Вариант № 1



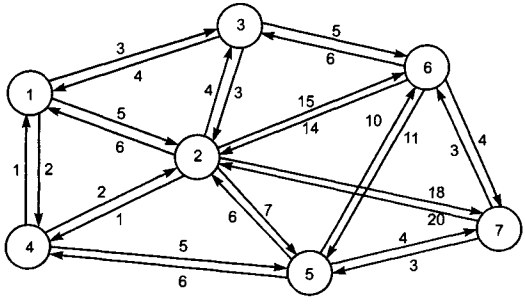
Вариант № 2



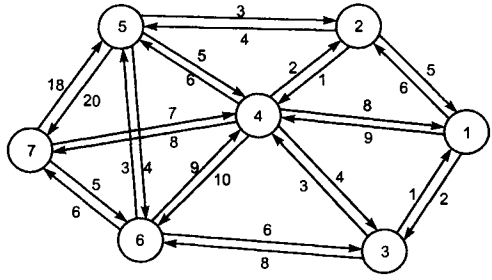
Вариант № 3



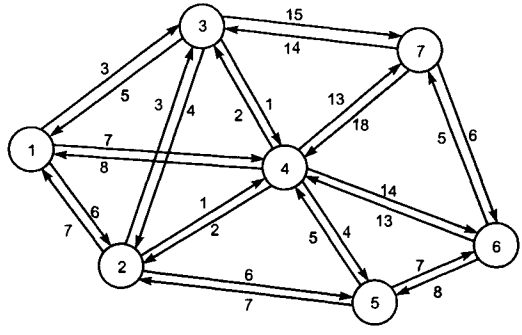
Вариант № 4



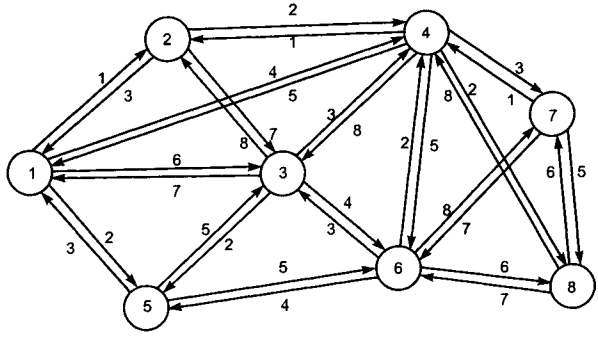
Вариант № 5



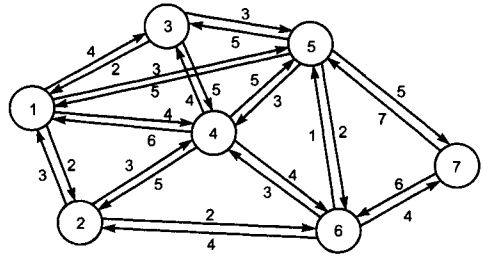
Вариант № 6



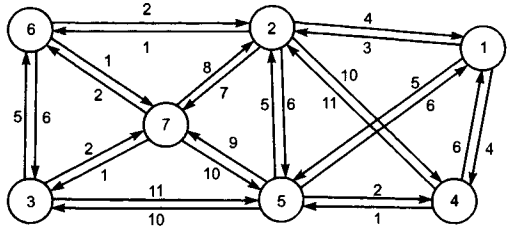
Вариант № 7



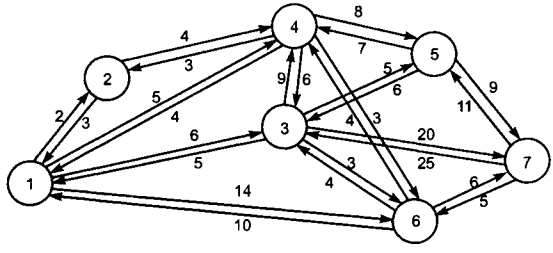
Вариант № 8



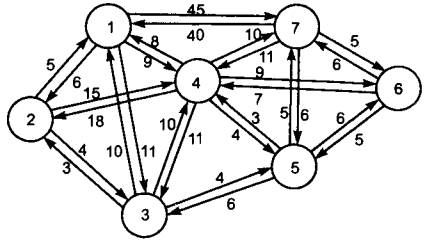
Вариант № 9



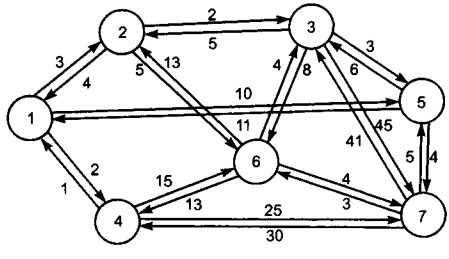
Вариант № 10



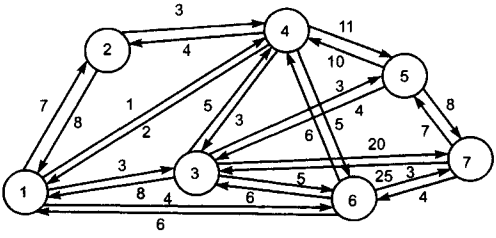
Вариант № 11



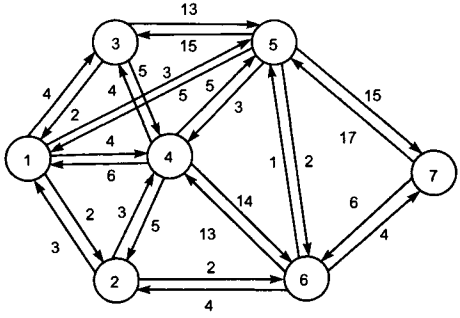
Вариант № 12



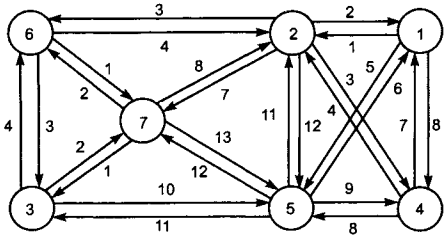
Вариант № 13



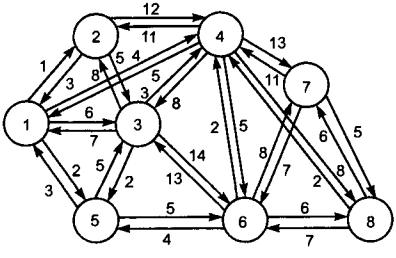
Вариант № 14



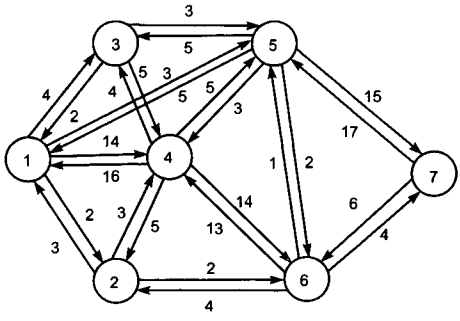
Вариант № 15



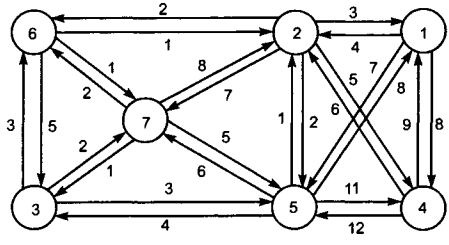
Вариант № 16



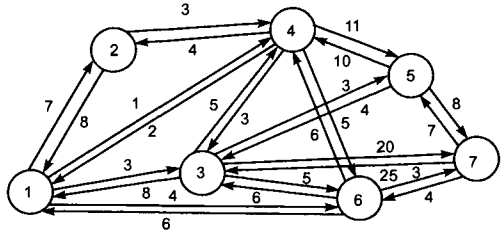
Вариант № 17



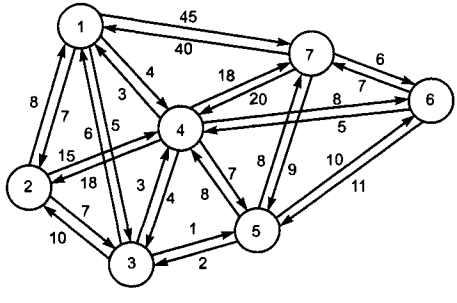
Вариант № 18



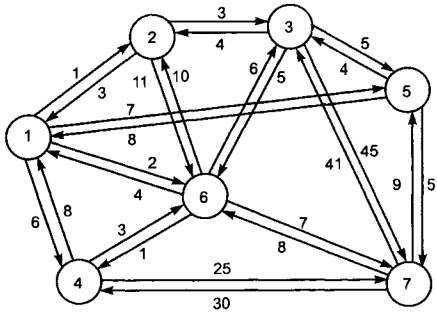
Вариант № 19



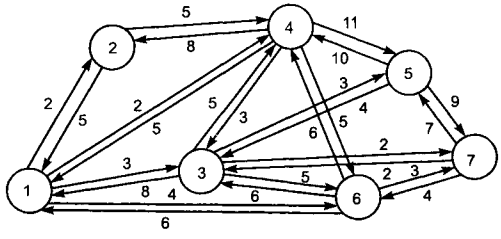
Вариант № 20



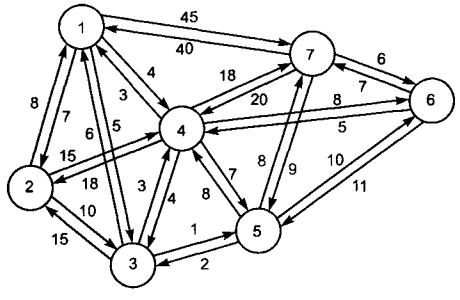
Вариант № 21



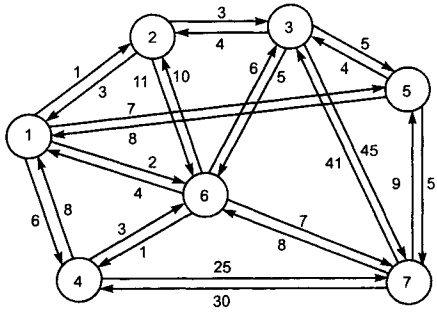
Вариант № 22



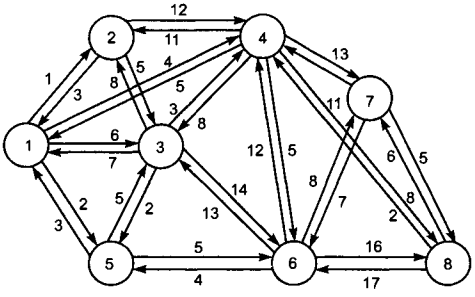
Вариант № 23



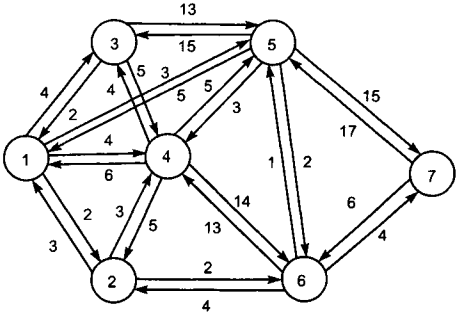
Вариант № 24



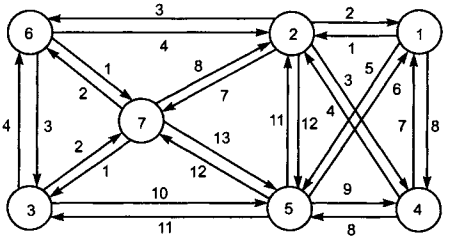
Вариант № 25



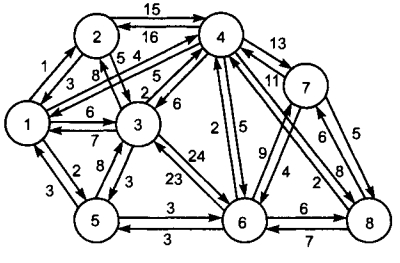
Вариант № 26



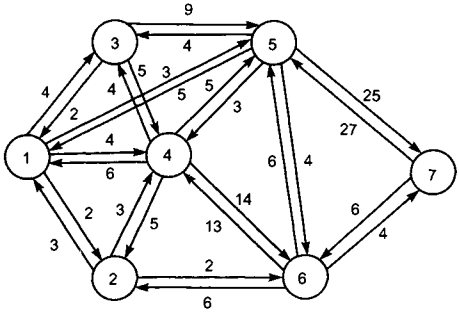
Вариант № 27



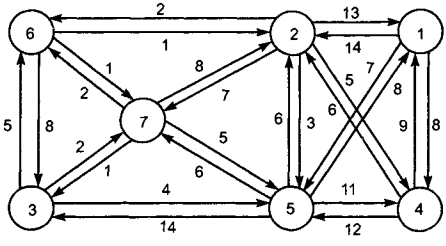
Вариант № 28



Вариант № 29



Вариант № 30



Приложение
ОСНОВЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

П.1. Случайные события

П.1.1. Статистическое определение вероятности

На практике часто встречаются эксперименты (испытания, опыты), которые можно проводить многократно, но результат которых нельзя предсказать, хотя основные условия эксперимента при его повторении остаются неизменными.

Так, при подбрасывании монеты нельзя предсказать что выпадет: герб или решка. Такие явления или опыты называются *случайными*, а их исходы — *случайными событиями*.

Объективной характеристикой того, что случайное событие A может произойти, является число $P(A)$, называемое его *вероятностью*.

Смысл и происхождение этого понятия можно объяснить следующим образом. Рассмотрим эксперимент, результатом которого являются те или иные случайные события. Предположим, что этот эксперимент можно проводить многократно в неизменных условиях и результаты предыдущих опытов не влияют на результаты последующих.

Пусть проведена серия из n экспериментов и пусть A — случайное событие, которое может произойти в результате каждого эксперимента. Определим число n_A — количество экспериментов, в которых событие A произошло, и рассмотрим отношение $h_n(A)$: $h_n(A) = n_A/n$.

Величина $h_n(A)$ называется *относительной частотой появления события* A в серии из n экспериментов. Проведем несколько серий экспериментов и для каждой серии из n экспериментов определим частоту появления события A.

Практика показывает, что частоты появления события A (при достаточно больших n) будучи, вообще говоря, различными, группируются около некоторого постоянного числа $P(A)$, т. е.

$$h_n(A) \approx P(A), \tag{1}$$

причем, чем больше n, тем выше относительная точность этого приближенного равенства. Число $P(A)$ называется *вероятностью* случайного события A, а описанное свойство (1) — *устойчивостью частот*.

Приближенное равенство (1) иногда рассматривают как *статистическое определение вероятности*. Таким образом, понятие вероятности связа-

но не с интуитивными предположениями, типа «на Марсе, вероятно, есть жизнь» или «завтра, возможно, будет хорошая погода», а с частотами появления случайных событий в таких экспериментах, для которых выполняется свойство устойчивости частот.

Пример 1. Рассмотрим распределение частот появления новорожденных девочек по данным шведской статистики за 1935 г. [3].

Месяц	1	2	3	4	5	6	7	8	9	10	11	12	За год
Всего новорожденных	7280	6957	7883	7884	7892	7609	7585	7393	7203	6903	6552	7132	88273
Девочек	3537	3407	3866	3711	3775	3665	3621	3596	3491	3391	3160	3371	42591
Частота рождения девочек	0,486	0,489	0,490	0,471	0,478	0,482	0,462	0,484	0,485	0,491	0,482	0,473	0,4825

Частоты рождения девочек очень незначительно отклоняются от средней частоты за год, равной 0,4825.

П.1.2. Пространство элементарных событий

Рассмотрим какой-либо эксперимент со случайными исходами. Результаты эксперимента — исходы будем называть случайными событиями. Условимся различать составные (или разложимые события) и элементарные (или неразложимые) события.

В результате эксперимента обязательно происходит одно и только одно элементарное событие, т. е. элементарные события взаимоисключают друг друга. Элементарные события обозначаются буквой ω.

Пример 2. Игральная кость подбрасывается один раз. Возможные случайные исходы этого эксперимента: событие A — на верхней грани выпало число очков, кратное трем, и шесть элементарных исходов ω_1, ω_2, ω_3, ω_4, ω_5, ω_6: на верхней грани выпала единица, двойка, тройка, четверка, пятерка, шестерка.

Событие A является составным, оно происходит в том случае, если на верхней грани выпадает либо тройка, либо шестерка. Таким образом, событие A можно рассматривать как совокупность (множество) элементарных исходов: $A = \{\omega_3, \omega_6\}$.

Определение 1. Множество Ω всех элементарных исходов данного эксперимента называется *множеством (пространством) элементарных событий.*

Случайные явления весьма разнообразны, и в каждом конкретном случае пространство элементарных событий Ω выбирается наиболее подходящим образом.

Пусть Ω — пространство элементарных событий некоторого эксперимента.

Для любого исхода этого эксперимента (события A) можно выделить подмножество тех элементарных исходов, наступление которых влечет за собой наступление события A. Таким образом, каждому событию A соответствует некоторое подмножество пространства элементарных событий Ω. Принадлежность элементарного события ω подмножеству A обозначают так: $\omega \in A$.

Операции над случайными событиями

Суммой событий A и B называется событие $A + B$ (или $A \cup B$), состоящее из элементарных событий, принадлежащих хотя бы одному из событий A или B (рис. 1, *а*).

Произведением событий A и B называется событие AB (или $A \cap B$), состоящее из элементарных событий, принадлежащих одновременно A и B (рис. 1, *б*).

Разностью событий A и B называется событие $A - B$ (или $A \setminus B$), состоящее из элементарных событий, принадлежащих A и не принадлежащих B (рис. 1, *в*).

Все пространство элементарных событий (событие Ω) называется *достоверным событием*, а пустое множество $\varnothing$ — *невозможным событием* (в данном эксперименте).

Событие $\overline{A} = \Omega \setminus A$ называется *противоположным событию* A (или дополнительным к событию A) (рис. 1, *г*).

События A и B называются *несовместными*, если $AB = \varnothing$ (рис. 1, *д*). Элементарные события по определению несовместны.

Если из наступления события A следует наступление B, т. е. событие B есть *следствие* события A, то это записывается так: $A \subset B$ (рис. 1, *е*).

Рис. 1, *а*—*е*, на которых демонстрируются операции над событиями, называются *диаграммами Эйлера—Венна.*

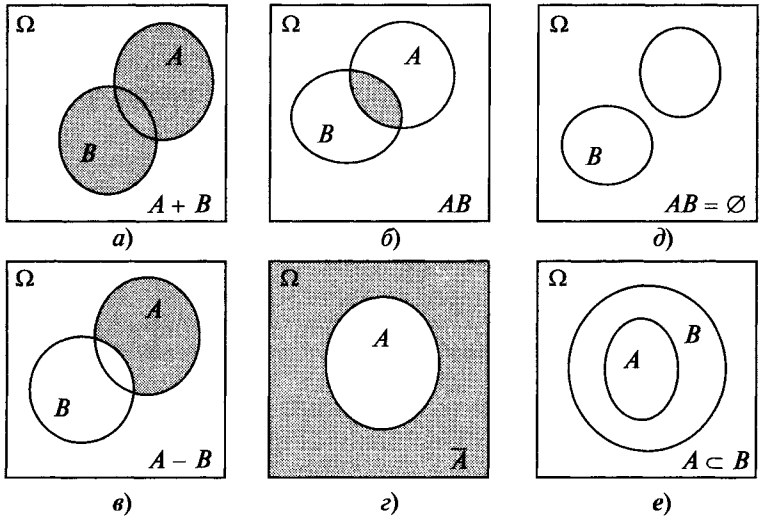


Рис. 1. Диаграммы Эйлера—Венна

Пример 3. Рассмотрим следующий случайный эксперимент: игральная кость (шесть граней) подбрасывается один раз, и наблюдается число, появляющееся на верхней грани. Множество всех элементарных (неразложимых, простейших) исходов данного эксперимента обозначим Ω. Ω состоит из шести элементов

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Пусть подмножество A соответствует событию, состоящему в том, что выпало четное число очков: $A = \{2, 4, 6\}$, а подмножество B — что выпало число очков, кратное трем: $B = \{3, 6\}$.

Тогда сумма, произведение, отрицание и разность — это следующие множества:

$$A + B = \{2, 3, 4, 6\},$$

$$AB = \{6\},$$

$$\overline{A} = \{1, 3, 5\}, \quad \overline{B} = \{1, 2, 4, 5\},$$

$$A - B = \{2, 4\}.$$

Операции над множествами удовлетворяют следующим основным законам:

1) идемпотентность: $A + A = A$, $AA = A$;

2) закон тождества: $A + \varnothing = A$, $A + \Omega = \Omega$, $A\Omega = A$, $A\varnothing = \varnothing$;

3) закон дополнения: $A + \overline{A} = \Omega$, $A\overline{A} = \varnothing$, $\overline{\overline{A}} = A$, $\overline{\Omega} = \varnothing$, $\overline{\varnothing} = \Omega$;

4) коммутативность $A + B = B + A$, $AB = BA$;

5) ассоциативность: $(A + B) + C = A + (B + C)$, $(AB)C = A(BC)$;

6) дистрибутивность: $A + BC = (A + B)(A + C)$, $A(B + C) = AB + AC$;

7) закон де Моргана: $\overline{A + B} = \overline{A} \cdot \overline{B}$, $\overline{AB} = \overline{A} + \overline{B}$.

Справедливость этих законов можно легко доказать с помощью диаграмм *Эйлера—Венна*.

П.1.3. Алгебра событий

Определение 2. Система F подмножеств Ω, удовлетворяющая условиям:

1) $\Omega \in F$;

2) из того, что $A, B \in F$, следует, что $A + B \in F$, $AB \in F$, $\overline{A}, \overline{B} \in F$

называется *алгеброй событий*.

Таким образом, алгебра F — это система подмножеств Ω, которая замкнута относительно конечного числа операций сложения, умножения и дополнения.

Замечание. Если условие 2 выполняется для счетного числа событий, то алгебра F называется σ-*алгеброй* (сигма-алгеброй).

П.1.4. Аксиоматическое определение вероятности и ее свойства

Определение 3. Пусть Ω — пространство элементарных событий данного эксперимента, а F — система подмножеств Ω, образующая алгебру событий. Числовая функция P, определенная на подмножествах Ω, которые входят в систему F, называется *вероятностью*, если выполнены следующие условия (аксиомы):

Аксиома 1. Для любого $A \in F$, $P(A) \geq 0$.

Аксиома 2. $P(\Omega) = 1$.

Аксиома 3 (аксиома конечной аддитивности). Если A и B несовместны, т. е. $AB = \varnothing$, то $P(A + B) = P(A) + P(B)$.

Аксиома 4 (аксиома непрерывности). Для любой убывающей последовательности событий из F

$$A_1 \supset A_2 \supset A_3 \supset \ldots \supset A_n \supset \ldots$$

такой, что $\prod_{n=1}^{\infty} A_n = \varnothing$, имеет место равенство

$$\lim_{n \to \infty} P(A_n) = 0.$$

Определение 4. Тройка (Ω, F, P) называется вероятностным пространством данного эксперимента.

Вероятностное пространство является математической моделью эксперимента со случайными исходами.

Обратим внимание на следующее:

1) в аксиоматическом определении вероятности определяются только для подмножеств из F, т. е. только такие подмножества рассматриваются как случайные события;

2) вероятностное пространство определяет условия, которым должна удовлетворять вероятность, но не определяет ее конкретно для того или иного эксперимента. Известно несколько моделей случайных экспериментов, в которых распределение вероятностей определяется достаточно просто, например, классическое определение вероятности для равновозможных элементарных исходов, геометрические вероятности, модели, приводящие к нормальному распределению, и другие.

Свойства вероятностей

Из аксиом, которым должны удовлетворять вероятности, следуют следующие основные свойства вероятностей.

1. Вероятность невозможного события ($\varnothing$) равна нулю, $P(\varnothing) = 0$.

Доказательство: $\varnothing + \Omega = \Omega$, следовательно, $P(\varnothing + \Omega) = P(\varnothing) + P(\Omega) = 1$. Так как $P(\Omega) = 1$, то $P(\varnothing) = 0$.

2. $P(\overline{A}) = 1 - P(A)$.

Доказательство: $A + \overline{A} = \Omega$, $P(A + \overline{A}) = P(A) + P(\overline{A}) = 1$, следовательно, $P(\overline{A}) = 1 - P(A)$.

3. Если $A \subset B$, то $P(A) \le P(B)$.

Доказательство: $B = B\Omega = B(A + \overline{A}) = BA + B\overline{A}$, но $BA = A$, (см. рис. 2), следовательно, $B = A + B\overline{A}$. $P(B) = P(A) + P(B\overline{A})$, т. е. $P(B) \ge P(A)$.

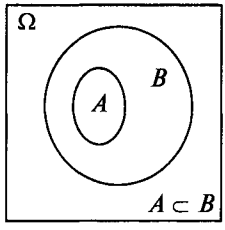


Рис. 2

4. $0 \le P(A) \le 1$.

Доказательство: $A \subset \Omega$, следовательно, $P(A) \le 1$.

5. Для любых событий A и B, $A, B \subset \Omega$, $P(A) \ge 0$, $P(B) \ge 0$ выполняется равенство: $P(A + B) = P(A) + P(B) - P(AB)$ — *теорема сложения для совместных событий.*

Доказательство (см. рис. 3): $P(A + B) = P(A - B) + P(AB) + P(B - A) =$ $= P(A) - P(AB) + P(B - A) + P(AB) = P(A) + P(B) - P(AB)$.

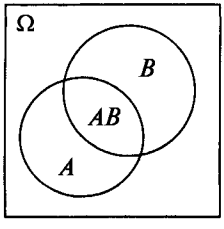


Рис. 3. Диаграмма, поясняющая вывод теоремы сложения вероятностей

П.1.5. Дискретное вероятностное пространство

Определение вероятности на дискретном пространстве элементарных событий.

Пусть Ω содержит либо конечное, либо счетное число элементарных событий: $\Omega = \{\omega_1, \omega_2, ...\}$.

F — набор всех подмножеств Ω (включая Ω и пустое множество $\varnothing$). Нетрудно показать, что F является σ-алгеброй.

Каждому элементарному исходу ω_i, $i = 1, 2, ...$ поставим в соответствие неотрицательное число $P(\omega_i) = p_i$, $i = 1, 2, ...$, такое, что $\sum_{i=1}^{\infty} p_i = 1$.

Для любого события $A \in F$ определим $P(A)$ следующим образом

$$P(A) = \sum_{\omega_i \in A} p(\omega_i). \qquad (2)$$

Можно показать, что числовая функция (2) удовлетворяет аксиомам вероятности 1—4.

Вероятностное пространство (Ω, F, P) называется в этом случае *дискретным.*

Если Ω содержит конечное число элементарных событий, например N событий, причем все элементарные исходы *равновозможны*, т. е. $p(\omega_i) = 1/N$, $i = 1, 2, ..., N$, то формулу (2) можно записать так

$$P(A) = |A| / |\Omega|, \tag{3}$$

где через $|A|$ обозначено количество элементарных исходов, составляющих множество A, а $|\Omega|$ — число всех элементарных исходов данного эксперимента, $|\Omega| = N$.

Формула (3) называется «классическим» определением вероятности.

Определение 5. Классическое определение вероятности. Если элементарные исходы равновозможны, то вероятность события A равна отношению числа исходов, благоприятствующих событию A, к числу всех элементарных исходов.

Пример 4. Из урны, содержащей 7 белых и 3 черных шара, наудачу извлекается один шар. Найти вероятность того, что будет извлечен шар белого цвета (событие A).

Решение. В данном эксперименте пространство элементарных событий Ω состоит из десяти равновозможных элементарных событий, $|\Omega| = 10$, а число элементарных событий благоприятствующих событию A равно семи, $|A| = 7$. По формуле (3)

$$P(A) = 7/10.$$

Пример 5. Из урны, содержащей 7 белых и 3 черных шара, случайным образом отбирается два шара. Какова вероятность того, что а) оба шара окажутся белыми (событие A); б) оба шара будут черными (событие B); в) шары будут разного цвета (событие C)?

Решение. Общее количество равновозможных вариантов $N = |\Omega|$ при выборке двух шаров из десяти равно числу различных пар без повторений, которое определяется как число сочетаний из 10 элементов по два, C_{10}^2

$$N = C_{10}^2 = \frac{10!}{2!\,8!} = 45.$$

а) Число благоприятствующих исходов $|A|$, которыми можно выбрать два белых шара будет равно

$$|A| = C_7^2 = \frac{7!}{2!\,5!} = 21.$$

Таким образом, искомая вероятность, по формуле (3), равна

$$P(A) = 21/45 = 7/15.$$

б) Число способов выбора двух черных шаров $|B|$ равно

$$C_3^2 = \frac{3!}{2!\,1!} = 3$$

и, следовательно, $P(B) = 3/45 = 1/15$.

в) Число способов выбора двух шаров разного цвета $|C|$ равно (по правилу произведения):

$$7 \cdot 3 = 21 \text{ и } P(C) = 21/45 = 7/15.$$

В этой задаче события A, B и C не могут произойти одновременно и одно из них обязательно происходит, поэтому: $P(A) + P(B) + P(C) = 1$. Проверьте!

П.1.6. Геометрические вероятности

В классическом определении вероятности для конечного числа равновозможных элементарных исходов вероятность события A — это доля элементарных исходов, благоприятствующих событию A.

Аналогичное этому определение вероятности используется в случае бесконечного числа «равновозможных» элементарных исходов. Это — геометрические вероятности.

Предположим, что пространство элементарных исходов Ω удовлетворяет следующим условиям:

1. Ω — квадрируемая область на плоскости (т. е. такая область, для которой можно определить площадь), $S(\Omega)$ — площадь Ω;

2. $A \subset \Omega$ — любая квадрируемая подобласть Ω, $S(A)$ — площадь A;

3. Вероятность попадания в A пропорциональна площади области A.

Тогда вероятность события A вычисляется по формуле геометрической вероятности

$$P(A) = S(A)/S(\Omega). \tag{4}$$

В случае n-мерного евклидова пространства формула (4) имеет вид

$$P(A) = \mu(A)/\mu(\Omega),$$

где $\mu(A)$ и $\mu(\Omega)$ — меры множеств A и Ω.

Пример 6. Два человека, независимо друг от друга, отправляются в один и тот же город на одну неделю в феврале месяце, причем время приезда может быть любым, от первого числа до двадцать первого февраля. Какова вероятность того, что они окажутся в городе одновременно?

Решение. Пусть x — время (в сутках) приезда в город первого, $0 \le x \le 21$, а y — время (в сутках) приезда второго человека, $0 \le y \le 21$. Введем двумерную декартову систему координат XOY. Тогда множество всех элементарных исходов Ω — квадрат со стороной равной 21 (рис. 4). Чтобы оба человека оказались в городе одновременно, разность между x и y должна быть

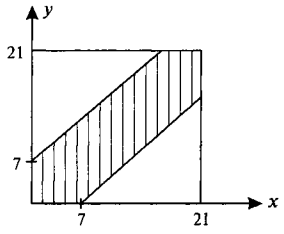


Рис. 4. Решение примера 6

не более семи суток, т. е. $|x - y| \le 7$ или $-7 \le y - x \le 7$. Этому событию соответствует та часть Ω, где удовлетворяется неравенство

$$x - 7 \le y \le x + 7.$$

Эта часть Ω представляет заштрихованный шестиугольник.

Искомая вероятность по формуле (4) вычисляется как отношение площади заштрихованной фигуры к площади квадрата

$$P(A) = [21^2 - (21 - 7)^2]/21^2 = 5/9.$$

П.1.7. Условные вероятности. Независимость событий

Пусть A и B — случайные события, которые могут произойти в одном эксперименте, $P(A) > 0$, $P(B) > 0$.

Аксиома 5. Условная вероятность $P(B/A)$ появления события B при условии, что событие A произошло определяется формулой

$$P(B/A) = P(AB)/P(A). \qquad (5)$$

Аналогично определяется условная вероятность $P(A/B)$ появления события A при условии, что событие B произошло

$$P(A/B) = P(AB)/P(B). \qquad (5')$$

Свойства условных вероятностей

1. $0 \le P(B/A) \le 1$.
2. $P(\overline{B}/A) = 1 - P(B/A)$.
3. Если событие B является следствием события A, т. е. $A \subset B$, то $P(B/A) = 1$.
4. Если событие B есть сумма событий C и D, $B = C + D$, то $P(B/A) = P((C + D)/A) = P(C/A) + P(D/A) - P(CD/A)$.

Пример 7. Из тщательно перетасованной колоды карт (36 карт) извлекается одна карта.

Найти вероятность того, что это будет туз (событие A), если известно, что извлечена карта черной масти (событие B).

Решение. В рассматриваемом случае возможно 36 равновозможных элементарных исходов. Событию A (появился туз) благоприятствует четыре исхода. Поэтому безусловная вероятность события A равна

$$P(A) = 4/36.$$

В колоде 18 карт черной масти (таким образом, событию B благоприятствуют 18 элементарных исходов), из них только две карты — тузы, поэтому условная вероятность появления события A при условии, что событие B произошло, равна

$$P(A/B) = 2/18 = 1/9.$$

По формуле (5) получаем тот же результат

$$P(A/B) = P(AB)/P(B) = (2/36)/(18/36) = 2/18 = 1/9.$$

Пример 7 показывает, что в случае, когда элементарные исходы равновозможны, формулу (5) можно легко доказать. Напомним, что в общем случае вычисление условной вероятности $P(B/A)$ по формуле (5) определяется как аксиома теории вероятностей.

Из равенств (5) и (5′) следует *теорема умножения вероятностей*

$$P(AB) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B). \qquad (6)$$

Важнейшим понятием теории вероятностей является понятие *независимости* событий.

Определение 6. Определение *независимости* событий. Если наступление события A не изменяет вероятности появления события B, т. е. $P(B/A) = P(B)$, то говорят, что событие B *не зависит* от события A.

В этом случае имеем

$$P(AB) = P(A) \cdot P(B/A) = P(A) \cdot P(B) = P(B) \cdot P(A/B),$$

и, следовательно, $P(A/B) = P(A)$. Это означает, что, если событие B не зависит от события A, то и событие A не зависит от события B. Таким образом, из независимости B от A следует независимость A от B, т. е. свойство независимости событий взаимно.

В случае независимости событий A и B теорема умножения вероятностей записывается в виде

$$P(AB) = P(A) \cdot P(B). \qquad (7)$$

Формулу (7) можно рассматривать как определение независимости событий A и B и использовать для проверки независимости событий.

Пример 8. Из колоды карт (36 карт) извлекается одна карта. Рассмотрим два события: событие A — извлечена дама, событие B — извлечена карта пиковой масти. Зависимы или независимы события A и B?

Решение. Если события A и B независимы, то справедлива формула (7). Имеем:

$$P(A) = 4/36;$$

$$P(B) = 9/36.$$

Событие AB — это извлечение пиковой дамы, следовательно

$$P(AB) = 1/36.$$

Так как равенство (7) удовлетворяется: $1/36 = 4/36 \cdot 9/36$, то события A и B независимы.

В случае трех событий A, B и C теорема умножения (6) записывается так:

$$P(ABC) = P(A) \cdot P(B/A) \cdot P(C/AB). \qquad (8)$$

Эта формула легко выводится, если обозначить событие AB как M и воспользоваться теоремой умножения (6) для событий M и C:

$$P(MC) = P(M)P(C/M) = P(A) \cdot P(B/A) \cdot P(C/AB).$$

Теорема умножения используется при вычислении вероятностей сложных событий.

Пример 9. Из урны содержащей 7 белых и 3 черных шара последовательно, без возвращения, извлекается три шара. Найти вероятность того, что извлечены три белых шара (событие A).

Решение. Введем следующие обозначения.

Событие A_1 — первый извлеченный шар белый. Событие A_2 — второй извлеченный шар белый. Событие A_3 — третий извлеченный шар белый. Тогда событие A есть произведение событий A_1, A_2 и A_3: $A = A_1A_2A_3$.

Вероятность события A можно вычислить по теореме умножения (8):

$$P(A_1A_2A_3) = P(A_1) \cdot P(A_2/A_1) \cdot P(A_3/A_1A_2).$$

Вероятности в правой части легко рассчитываются:

$$P(A_1) = 7/10;$$

$$P(A_2/A_1) = 6/9;$$

$$P(A_3/A_1A_2) = 5/8.$$

Окончательно получаем

$$P(A) = 7/10 \cdot 6/9 \cdot 5/8 = 7/24.$$

Определение 7. Три события A, B и C называются *независимыми в совокупности*, если они попарно независимы

$$P(AB) = P(A) \cdot P(B), \quad P(AC) = P(A) \cdot P(C), \quad P(BC) = P(B) \cdot P(C)$$

и выполняется равенство

$$P(ABC) = P(A)P(B)P(C).$$

Аналогично определяется независимость четырех и более событий.

Можно показать, что для независимости в совокупности недостаточно попарной независимости.

Пример 10. Подбрасывается две игральные кости. Рассмотрим следующие события: A — на первой кости выпадает нечетное число очков; B — на второй кости выпадает нечетное число очков; C — сумма очков на обеих костях нечетна.

События A, B и C попарно независимы. Действительно: $P(A) = 1/2$; $P(B) = 1/2$; $P(C) = 1/2$.

При этом событие A не зависит от B, так как $P(A) = P(A/B)$, а также не зависит от C, так как $P(A) = P(A/C)$.

Аналогично, событие B не зависит от A и от C, так как $P(B) = P(B/A) = P(B/C)$.

Следовательно, и событие C не зависит от A и от B: $P(C) = P(C/A) = P(C/B)$.

Однако, событие ABC — невозможно, т. е. $ABC = \varnothing$ и $P(ABC) = 0$. Таким образом, условие $P(ABC) = P(A) \cdot P(B) \cdot P(C)$ не выполняется. Следовательно, события A, B и C не независимы в совокупности.

П.1.8. Формула полной вероятности и формула Байеса

Пусть H_1, H_2, ..., H_k подмножества пространства элементарных событий Ω, такие что:

1) $H_i H_j = \varnothing$, $i \neq j$, $i, j = 1, 2, ..., k$;

2) $H_1 + H_2 + ... + H_k = \Omega$.

В этом случае говорят, что система подмножеств H_1, H_2, ..., H_k образует *разбиение множества* Ω. Для любого события A, являющегося подмножеством Ω, верна формула *полной вероятности*

$$P(A) = \sum_{i=1}^{k} P(H_i) \cdot P(A/H_i). \qquad (9)$$

События H_i называются гипотезами по отношению к событию A, а вероятности $P(H_i)$ трактуются как доопытные (априорные) вероятности гипотез, причем $\sum_{i=1}^{k} P(H_i) = 1$.

Пример 11. Пять неразличимых по внешнему виду урн содержат белые и черные шары в следующих количествах. В двух урнах находятся по два белых и по одному черному шару; в двух других урнах лежит по три белых и два черных шара; наконец, в одной урне лежит десять черных шаров. Наудачу выбирается одна урна и из нее извлекается один шар. Найти вероятность того, что этот шар белый.

Решение. Введем следующие гипотезы: H_1 — шар извлечен из урны 1-го состава; H_2 — шар извлечен из урны 2-го состава; H_3 — шар извлечен из урны 3-го состава.

Тогда: $P(H_1) = 2/5$; $P(H_2) = 2/5$; $P(H_3) = 1/5$.

Пусть A — событие, состоящее в том, что извлечен белый шар. Условные вероятности наступления события A при извлечении шара из урны того или иного состава равны

$$P(A/H_1) = 2/3;\ P(A/H_2) = 3/5;\ P(A/H_3) = 0.$$

По формуле (9) имеем

$$P(A) = 2/5 \cdot 2/3 + 2/5 \cdot 3/5 + 1/5 \cdot 0 = 76/150.$$

Если известно, что событие A произошло, то априорные вероятности гипотез H_i, очевидно, должны быть пересчитаны, так как появилась дополнительная информация. Апостериорные (послеопытные) вероятности гипотез H_i, при условии, что событие A произошло, вычисляются по *формуле Байеса*

$$P(H_i/A) = \frac{P(H_i)P(A/H_i)}{P(A)}, \; i = 1, 2, \ldots, k,$$

где $P(A)$ определяется по формуле полной вероятности (9).

Пример 12. Предположим, что в условиях примера 11, шар, извлеченный из наудачу выбранной урны, оказался белым — произошло событие A. Какая из двух гипотез более вероятна: шар извлечен из урны первого или второго состава?

Решение. Воспользуемся результатами решения примера 11 и формулой Байеса. Имеем:

$$P(H_1/A) = P(H_1) \cdot P(A/H_1)/P(A) = (2/5 \cdot 2/3)/(76/150) = 40/76,$$

$$P(H_2/A) = P(H_2) \cdot P(A/H_2)/P(A) = (2/5 \cdot 3/5)/(76/150) = 36/76.$$

Таким образом, более вероятно, что шар был извлечен из урны первого состава.

П.2. Дискретные случайные величины. Системы дискретных случайных величин

П.2.1. Определение дискретной случайной величины

Пусть (Ω, F, P) — дискретное вероятностное пространство.

Определение 1. Числовая функция $X = X(\omega)$ определенная на пространстве элементарных событий Ω называется *дискретной случайной величиной.*

Обозначим значения, которые принимает случайная величина X, через $x_1, x_2, \ldots, x_n, \ldots$.

Задание случайной величины как функции на Ω, означает, что каждому элементарному исходу ω соответствует определенное значение случайной величины $X(\omega)$. Одно и то же значение x_i может, вообще говоря, соответствовать нескольким элементарным событиям, определяя таким образом составное событие $X = x_i$. Обозначим вероятность этого события как $P[X = x_i] = p_i$.

Определение 2. Система равенств: $P[X = x_i] = p_i$, $i = 1, 2, \ldots, n, \ldots$ определяет распределение вероятностей дискретной случайной величины X. Очевидно, что $p_i \geq 0$, $\sum_i p_i = 1$.

Обычно распределение дискретной случайной величины записывается в виде *ряда распределения вероятностей* (табл. 1).

Таблица 1

X	x_1	x_2	...	x_n	...
$P[X = x_i]$	p_1	p_2	....	p_n	...

Пример 1. Подбрасывается две правильные игральные кости. Рассмотрим случайную величину X равную сумме очков, выпавшую на обеих костях.

В этом примере пространство элементарных событий состоит из 36 равновероятных исходов: (1, 1), (1, 2), ..., (6, 6). Случайная величина X принимает 11 значений: 2, 3, 4, ..., 12. Чтобы построить ряд распределения нужно определить вероятности составных событий $P[X = 2, 3, 4, ..., 12]$. Событию $X = 2$ соответствует один элементарный исход (1, 1), следовательно, вероятность этого события равна 1/36, $P[X = 2] = 1/36$. Событию $X = 3$ соответствует два элементарных исхода: (1, 2) и (2, 1), следовательно $P[X = 3] = 2/36$. Рассуждая, аналогично получим ряд распределения случайной величины X:

X	2	3	4	5	6	7	8	9	10	11	12
$P[X = x_i]$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Сумма вероятностей = 1, проверьте!

Можно представить ряд распределения случайной величины в виде графика называемого *многоугольником распределения*. Для примера 1 многоугольник распределения изображен на рис. 5.

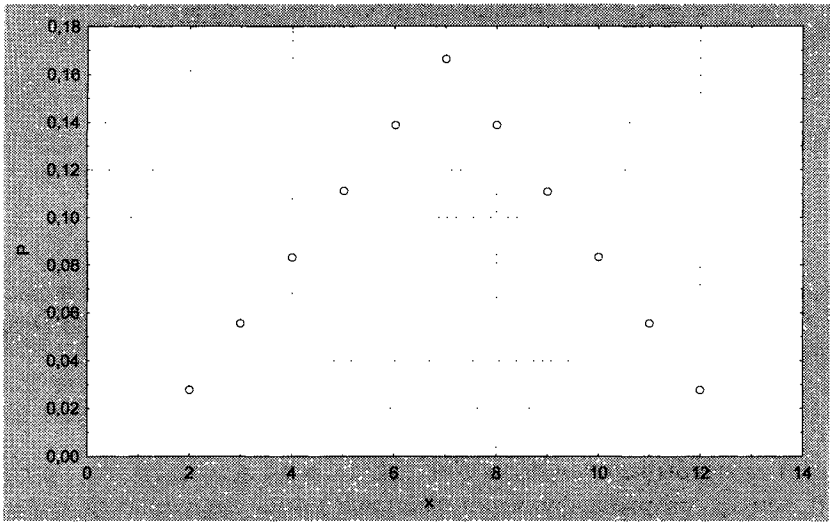


Рис. 5

П.2.2. Механическая интерпретация распределения вероятностей дискретных случайных величин

Рассмотрим дискретную случайную величину X, принимающую конечное число значений: $x_1, x_2, ..., x_n$ с вероятностями: $P[X = x_i] = p_i$, $i = 1, 2, ..., n$. Предположим, что $x_1, x_2, ..., x_n$ записаны в порядке возрастания, т. е. $x_i < x_{i+1}$. Разместим на числовой оси точки с абсциссами $x_1, x_2, ..., x_n$ и в каждую точку x_i поместим соответствующую массу, равную $p_i (\sum p_i = 1)$ (рис. 6).

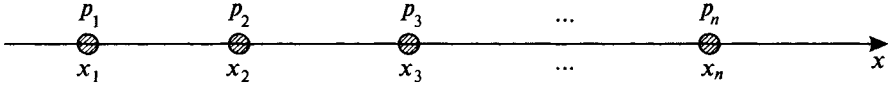


Рис. 6. Механическая интерпретация вероятностного распределения

Полученную систему материальных точек рассматривают, как механическую интерпретацию распределения случайной величины X. А именно: распределение вероятностей интерпретируется, как распределение единичной массы по n точкам с абсциссами $x_1, x_2, ..., x_n$, причем, вероятности p_i соответствует масса p_i, $i = 1, 2, ..., n$.

П.2.3. Функция распределения случайной величины

Определение 3. *Функцией распределения случайной величины* X называется функция $F(x)$, определяемая для любого действительного значения x, как вероятность события $[X < x]$, т. е.

$$F(x) = P[X < x].$$

Заметим, что в механической интерпретации функция распределения $F(x)$ есть суммарная масса материальных точек, расположенных левее точки с абсциссой x (рис. 7).

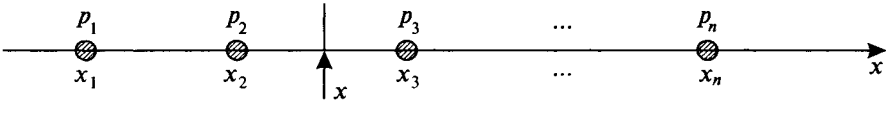


Рис. 7

Приведем (без доказательства) основные свойства функции распределения. Смысл этих свойств хорошо иллюстрируется с помощью механической интерпретации случайных величин:

1. $0 \le F(x) \le 1$.
2. $F(-\infty) = 0$; $F(+\infty) = 1$.
3. $F(x)$ — неубывающая функция x, т. е., если $x_2 > x_1$, то $F(x_2) \ge F(x_1)$.
4. В точках разрыва функция распределения непрерывна слева, т. е. $\lim\limits_{x \to a-0} F(x) = F(a)$, где $x = a$ — точка разрыва.

5. Функция распределения дискретной случайной величины X, принимающей значения $x_1, x_2, ..., x_n$ возрастает скачками в точках $x_1, x_2, ..., x_n$ на величины равные соответствующим вероятностям $P[X = x_1] = p_1$, $P[X = x_2] = p_2$, ...

Зная ряд распределения случайной величины, можно построить график функции распределения.

Пример 2. Подбрасываются две симметричные монеты. Рассмотрим случайную величину X равную числу выпавших гербов. Равновероятные элементарные исходы данного эксперимента: РР, РГ, ГР, ГГ. Ряд распределения случайной величины имеет вид:

X	0	1	2
$P[X = x_i]$	1/4	2/4	1/4

Построим график функции распределения случайной величины X. При $x < 0$ функция распределения $F(x) = 0$; в точке $x = 0$ функция распределения $F(x)$ возрастает скачком на величину равную $P[X = 0] = 1/4$. На интервале $0 < x < 1$ значение $F(x)$ постоянно, $F(x) = 1/4$. В точке $x = 1$, функция распределения $F(x)$ возрастает скачком на величину равную $P[X = 1] = 2/4$.

На интервале $1 < x < 2$ значение $F(x)$ постоянно и равно $1/4 + 2/4 = 3/4$. В точке $x = 2$ функция распределения возрастает скачком на величину равную $P[X = 2] = 1/4$. На интервале $2 < x < \infty$ значение $F(x)$ постоянно и равно $3/4 + 1/4 = 1$.

График $F(x)$ приведен на рис. 8

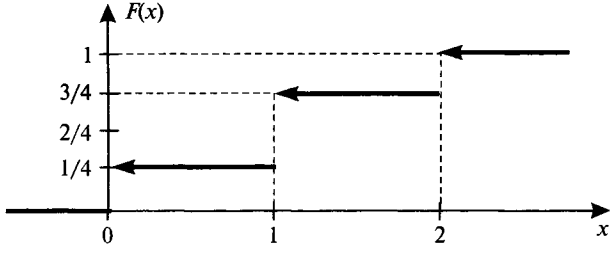


Рис. 8. Функция распределения для примера 2

П.2.4. Система двух дискретных случайных величин

Рассмотрим две случайные величины X и Y, определенные на одном дискретном вероятностном пространстве (Ω, F, P). Обозначим значения, которые принимает случайная величина X через $x_1, x_2, ..., x_n$, а значения случайной величины Y через $y_1, y_2, ..., y_m$. Распределения вероятностей X и Y обозначим соответственно $p_{x_1}, p_{x_2}, ..., p_{x_n}$ и $p_{y_1}, p_{y_2}, ..., p_{y_m}$. Вероятность события, состоящего в том, что $X = x_i$ и $Y = y_j$, обозначим как $P[X = x_i;\ Y = y_j] = p_{ij}$.

Определение 4. Система равенств $P[X = x_i;\ Y = y_j] = p_{ij},\ p_{ij} \geq 0,\ \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} = 1,$ $p_{ij} = 1,\quad i = 1, 2, ..., n,\quad j = 1, 2, ..., m$ определяет совместное распределение дискретных случайных величин X и Y или системы двух дискретных случайных (X, Y).

Распределение системы двух случайных величин (X, Y) записывают в виде таблицы распределения (табл. 2).

Таблица 2

X \ Y	y_1	y_2	...	y_m	$\sum_{j=1}^{m} p_{ij} = p_{x_i}$
x_1	p_{11}	p_{12}	...	p_{1m}	p_{x_1}
x_2	p_{21}	p_{22}	...	p_{2m}	p_{x_2}
. . .					. . .
x_n	p_{n1}	p_{n2}	...	p_{nm}	p_{x_n}
$\sum_{i=1}^{n} p_{ij} = p_{y_j}$	p_{y_1}	p_{y_2}	...	p_{y_m}	

Суммируя вероятности p_{ij} по строкам, получим распределение случайной величины X: $\sum_{j=1}^{m} p_{ij} = p_{x_i},\quad i = 1, 2, ..., n$; суммирование вероятностей p_{ij} по столбцам дает распределение случайной величины Y: $\sum_{i=1}^{n} p_{ij} = p_{y_j},\ j = 1, 2, ..., m$.

Аналогично определяется распределение системы более чем двух случайных величин.

Условная вероятность события $X = x_i$ при условии, что $Y = y_j$, $(p_{y_j} > 0)$ определяется формулой

$$P[X = x_i / Y = y_j] = \frac{P[X = x_i;\ Y = y_j]}{P[Y = y_j]} = \frac{p_{ij}}{p_{y_j}}. \qquad (1)$$

Система равенств (1) при $i = 1, 2, ..., n$ задает *условное распределение* случайной величины X при условии, что случайная величина Y принимает заданное значение $Y = y_j$.

Определение 5. Определение независимости случайных величин. Пусть таблица 2 суть таблица распределения случайных величин X и Y. Случайные величины X и Y называются *независимыми*, если события $X = x_i$ и $Y = y_j$ независимы для всех i и j таких, что $1 \leq i \leq n$, $1 \leq j \leq m$, т. е. $P[X = x_i;\ Y = y_j] = P[X = x_i] \cdot P[Y = y_j]$, или $p_{ij} = p_{x_i} \cdot p_{y_j}$.

Если X и Y независимые случайные величины, то таблица распределения (табл. 2) имеет вид таблицы умножения.

Аналогично определяется взаимная независимость более чем двух случайных величин.

Определение 6. Случайные величины X_1, X_2, ..., X_n определенные на одном дискретном вероятностном пространстве называются *взаимно независимыми*, если для любой комбинации значений x_{i_1}, x_{i_2}, ..., x_{i_n}.

$$P[X_1 = x_{i_1};\ X_2 = x_{i_2}, \ldots, X_n = x_{i_n}] = P[X_1 = x_{i_1}] \cdot P[X_2 = x_{i_2}] \ldots P[X_n = x_{i_n}].$$

Пример 3. Игральная кость подбрасывается один раз. Рассмотрим две случайные величины, связанные с этим экспериментом.

Случайная величина X принимает два значения: $x_1 = -1$, если выпало число очков кратное трем и $x_2 = 1$, если выпало число очков не кратное трем.

Ряд распределения случайной величины X имеет вид

X	-1	1
$P[X = x_i]$	2/6	4/6

Случайная величина Y равна числу выпавших очков. Ряд распределения случайной величины Y имеет вид

Y	1	2	3	4	5	6
$P[Y = y_j]$	1/6	1/6	1/6	1/6	1/6	1/6

Совместное распределение системы случайных величин X и Y задается таблицей распределения (табл. 3):

Таблица 3

X \ Y	1	2	3	4	5	6	$\sum_{j=1}^{m} p_{ij} = p_{x_i}$
-1	0	0	1/6	0	0	1/6	2/6
1	1/6	1/6	0	1/6	1/6	0	4/6
$\sum_{i=1}^{n} p_{ij} = p_{y_j}$	1/6	1/6	1/6	1/6	1/6	1/6	

Например, $P[X = -1,\ Y = 2] = p_{12} = 0$, так как при выпадении двух очков, случайная величина X не может принимать значение $x_1 = -1$. Случайные величины X и Y не являются независимыми, так как, например, $p_{12} = 0$, а $p_{x_1} \cdot p_{y_2} = \frac{2}{6} \cdot \frac{1}{6} = \frac{2}{36} \neq 0$.

П.2.5. Числовые характеристики дискретных случайных величин

Числовые характеристики случайной величины определяют основные свойства распределения: среднее значение, разброс значений случайной величины относительно среднего значения и другие. Важнейшей характеристикой случайной величины является ее *математическое ожидание* или *среднее значение.*

Определение 7. Пусть X — дискретная случайная величина, принимающая значения x_1, x_2... с вероятностями p_1, p_2, ... *Математическое ожидание* $M[X]$ случайной величины X определяется формулой

$$M[X] = \sum_i x_i p_i \qquad (2)$$

в предположении, что ряд (2) абсолютно сходится.

Если ряд (2) расходится, то говорят, что случайная величина X не имеет конечного математического ожидания.

Свойства математического ожидания

1. $M[c] = c$, где c — константа.

2. $M[cX] = cM[X]$ и $M[X + c] = M[X] + c$, где c — константа.

3. Пусть случайная величина Z является заданной функцией случайной величины X: $Z = h(X)$. Например, $Z = X^3$ или $Z = \cos X$. Случайная величина Z имеет конечное математическое ожидание $M[Z]$, вычисляемое по формуле

$$M[Z] = \sum_i h(x_i) p_i, \qquad (3)$$

при условии, что ряд (3) абсолютно сходится.

4. Пусть X_1, X_2, ..., X_n — случайные величины с конечными математическими ожиданиями. Математическое ожидание их суммы равно сумме их математических ожиданий

$$M[X_1 + X_2 + \ldots + X_n] = M[X_1] + M[X_2] + \ldots + M[X_n].$$

5. Пусть X и Y — взаимно независимые случайные величины с конечными математическими ожиданиями. Математическое ожидание произведения XY равно произведению их математических ожиданий

$$M[XY] = M[X] \cdot M[Y].$$

Это правило распространяется на любое конечное число взаимно независимых случайных величин.

Заметим, что последнее равенство для зависимых случайных величин, вообще говоря, не выполняется. Рассмотрим следующий пример.

Пример 4. Случайная величина X принимает два значения -1 и $+1$, каждое с вероятностями, равными $1/2$.

По формуле (2) имеем

$$M[X] = -1 \cdot 1/2 + 1 \cdot 1/2 = 0.$$

Пусть случайная величина $Y = X$ (это означает, что $Y(\Omega) = X(\Omega)$ для всех Ω). Тогда, очевидно, распределение Y будет то же, что и у X и $M[Y] = 0$.

Однако произведение $XY = X^2 = 1$ и $M[XY] = M[1] = 1$.

Пусть теперь случайная величина $Z = -X$ (т. е. $Z(\Omega) = -X(\Omega)$ для всех Ω).

Очевидно, распределение Z также совпадает с распределением случайной величины X. Однако $XZ = -X^2 = -1$ и $M[XZ] = M[-1] = -1$.

Пример показывает, что при одних и тех же распределениях различные зависимости между случайными величинами ($X = Y$ и $Z = -X$) приводят к различным значениям для математических ожиданий произведений XY и XZ.

6. Пусть X и Y — случайные величины с совместным распределением, задаваемым табл. 2.

Условное математическое ожидание случайной величины X при условии, что Y принимает заданное значение $Y = y_j$, вычисляется по формуле:

$$M[X/Y = y_j] = \sum_{i=1}^{n} x_i P[X = x_i/Y = y_j] = \sum_{i=1}^{n} x_i \cdot \frac{p_{ij}}{p_{y_j}}, \quad p_{y_j} > 0. \tag{4}$$

Другой важной характеристикой случайной величины X является ее *дисперсия* $D[X]$. Дисперсия характеризует разброс значений случайной величины относительно ее математического ожидания и вычисляется по формуле

$$D[X] = M[(X - M[X])^2] = \sum_{i=1} (x_i - m)^2 p_i, \tag{5}$$

где $m = M[X]$ и в предположении, что ряд (5) сходится.

Величина $\sigma = +\sqrt{D[X]}$ называется *средним квадратическим* или *стандартным отклонением* случайной величины X.

Основные свойства дисперсии:

1. $D[X] \geq 0$.

2. Если случайная величина X постоянна, т. е. $X(\Omega) = \text{const} = c$, то ее дисперсия равна нулю: $D[c] = 0$.

3. $D[cX] = c^2 D[X]$, где $c = \text{const}$.

4. $D[X + c] = D[X]$.

5. Дисперсия суммы независимых случайных величин X и Y равна сумме их дисперсий

$$D[X + Y] = D[X] + D[Y].$$

Формулу (5) можно преобразовать следующим образом

$$D[X] = M[(X - m)^2] = M[X^2] - 2mM[X] + m^2 = M[X^2] - m^2 = \sum_i x_i^2 p_i - m^2. \tag{5а}$$

Формула (5а) показывает, что для существования конечной дисперсии случайной величины X достаточно, чтобы существовало конечное математическое ожидание квадрата случайной величины $M[X^2]$.

Дисперсия является мерой разброса возможных значений случайной величины. Если дисперсия мала, то, как следует из (5), будут малы и все члены ряда $\sum_i (x_i - m)^2 p_i$, поэтому значения, для которых отклонение от среднего $x_i - m$ будут велики, должны иметь малые вероятности p_i. Таким образом, если дисперсия случайной величины X мала, то большие отклонения X от математического ожидания m маловероятны.

Этот же вывод можно получить и как следствие неравенства Чебышева.

Неравенство Чебышева: если случайная величина X имеет конечную дисперсию, то при любом $\varepsilon > 0$, $P[|X - M[X]| \geq \varepsilon] \leq \dfrac{D[X]}{\varepsilon^2}$.

Неравенство Чебышева можно записать в следующем эквивалентном виде

$$P[|X - M[X]| \geq k\sigma] \leq \frac{1}{k^2}, \qquad (6)$$

где k — константа, $k > 1$, а σ — среднее квадратическое, или стандартное отклонение $\sigma = +\sqrt{D[X]}$.

Если в (6) положить, например, $k = 10$, то $P[|X - m| \geq 10\sigma] \leq 0{,}01$.

Это означает, что в большой серии экспериментов по наблюдению случайной величины X, отклонения X от ее математического ожидания m более, чем на 10σ, будут встречаться менее, чем в 1 % случаев.

При $k = 100$ неравенство (6) показывает, что отклонения X от математического ожидания на величину бóльшую, чем 100σ, будут встречаться уже менее чем в 0,01 % наблюдений. То есть, если σ мало, то большие отклонения X от математического ожидания будут крайне редки.

Иными словами, случайная величина с очень малой дисперсией — это величина практически постоянная. В подавляющем большинстве экспериментов она почти не отклоняется от своего математического ожидания.

В механической интерпретации распределения случайной величины X как системы материальных точек, математическое ожидание суть координата центра тяжести этой системы, а дисперсия — момент инерции системы относительно центра тяжести.

Математическое ожидание и дисперсия случайной величины X являются частными случаями, так называемых *моментов.*

Определение 8. *Начальным моментом k-го порядка* α_k случайной величины X называется математическое ожидание k-й степени случайной величины X:

$$\alpha_k = M[X^k], \quad k = 1, 2, \ldots . \qquad (7)$$

Центральным моментом k-го порядка μ_k случайной величины X называется математическое ожидание k-й степени отклонения X от ее математического ожидания m:

$$\mu_k = M[(X - m)^k], \quad k = 1, 2, \ldots . \qquad (8)$$

Для дискретной случайной величины X, принимающей значения x_1, x_2, ..., x_n, ... с вероятностями соответственно p_1, p_2, ..., p_n, ... начальный и центральный моменты порядка k вычисляются по формулам:

$$\alpha_k = \sum_i x_i^k p_i,$$

$$\mu_k = \sum_i (x_i - m)^k p_i, \quad k = 1, 2, \ldots$$

при условии, что ряды в правых частях формул сходятся абсолютно.

Математическое ожидание случайной величины X — это, по определению, начальный момент первого порядка: $M[X] = \alpha_1$, а дисперсия — центральный момент второго порядка: $D[X] = \mu_2 = M[(X - m)^2]$.

Формула (5) показывает, что центральный момент второго порядка можно вычислить с помощью начальных моментов: $\mu_2 = \alpha_2 - \alpha_1^2$, так как $M[X^2] = \alpha_2$.

Известны формулы, связывающие центральные и начальные моменты более высоких порядков, например:

$\mu_3 = \alpha_3 - 3\alpha_2\alpha_1 + 2\alpha_1^3$;

$\mu_4 = \alpha_4 - 4\alpha_3\alpha_1 + + 6\alpha_2\alpha_1^2 - 3\alpha_1^4$ и т. д.

Центральные моменты третьего и четвертого порядков определяют характеристики формы распределения. *Коэффициент асимметрии (скошенности) распределения* случайной величины определяется формулой

$$a_x = \frac{\mu_3}{\sigma^3}, \tag{9}$$

где σ — среднее квадратическое, или стандартное отклонение, а *коэффициент эксцесса (островершинности)* — формулой

$$e_x = \frac{\mu_4}{\sigma^4} - 3. \tag{10}$$

Важными характеристиками распределения случайной величины X являются ее квантили.

Определение 9. Пусть p — действительное число из отрезка [0; 1]. *Квантилью* случайной величины X порядка p называется действительное число x_p, являющееся решением уравнения

$$F(x_p) = p, \tag{11}$$

где $F(x)$ — функция распределения случайной величины X.

Квантиль порядка $p = 0{,}5$ называется *медианой* h_x случайной величины X:

$$h_x = x_{0,5}.$$

Модой d_x дискретной случайной величины, принимающей значения x_1, x_2, ..., называется такое значение случайной величины, которое имеет наибольшую вероятность

$$P[X = d_x] = \max_k P[X = x_k]$$

при условии, что x_k — единственное значение, удовлетворяющее этому условию. Если такое значение не единственно, то мода случайной величины не определена.

П.2.6. Примеры дискретных распределений: биномиальное, пуассоновское и геометрическое распределения

Биномиальное распределение, $B(n, p)$. Рассмотрим случайный эксперимент, имеющий два возможных исхода: A — «успех» и $\bar{A}$ — «неудача». Предположим, что эксперимент повторяется n раз при следующих условиях:

1. Результаты последовательных экспериментов являются независимыми в совокупности случайными событиями.

2. Вероятности исходов A и $\bar{A}$ в каждом эксперименте неизменны при повторении экспериментов и равны: $P(A) = p$, $P(\bar{A}) = 1 - p = q$.

Рассмотрим случайную величину X — число «успехов» в n экспериментах. Дискретная случайная величина X принимает значения: 0, 1, 2, ..., n. Распределение случайной величины X определяется следующей формулой

$$p_k = P[X = k] = C_n^k p^k q^{n-k}, \quad k = 0, 1, ..., n, \tag{12}$$

где C_n^k — число сочетаний из n по k, $C_n^k = \dfrac{n!}{(n-k)!k!}$.

По формуле бинома Ньютона

$$\sum_{k=0}^{n} p_k = \sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p + q)^n = 1.$$

Распределение дискретной случайной величины X, определяемое формулой (12), называется *биномиальным распределением.*

Чтобы пояснить вывод формулы (12) рассмотрим сначала случай, когда эксперимент повторяется два раза ($n = 2$). Возможные исходы будут: AA, $A\bar{A}$, $\bar{A}A$, $\bar{A}\bar{A}$. Соответствующие им вероятности в силу независимости испытаний: p^2, pq, qp и q^2. В первом случае (AA) — число успехов равно двум и вероятность этого события: $P[X = 2] = p^2$, во втором и третьем случаях ($\bar{A}A$ и $A\bar{A}$) — число успехов равно единице и вероятность этого события: $P[X = 1] = 2pq$; в третьем случае $\bar{A}\bar{A}$ — число успехов равно нулю и вероятность этого события: $P[X = 0] = p^2$. Таблица распределения числа успехов X в двух экспериментах имеет вид

X	0	1	2
$P[X = k]$	q^2	$2pq$	p^2

Пусть теперь эксперимент с двумя возможными исходами A и $\bar{A}$ повторяется n раз. Вычислим вероятность того, что исход A (успех) наступит ровно k раз. Вероятность события, в котором, в первых k испытаниях будет успех (A), а в остальных $(n - k)$ испытаниях будет неуспех ($\bar{A}$) в силу независимости испытаний, будет равна

$$\underbrace{(p \cdot p \cdot \ldots \cdot p)}_{k \text{ раз}} \cdot \underbrace{(q \cdot q \cdot \ldots \cdot q)}_{(n-k) \text{ раз}} = p^k \cdot q^{n-k}.$$

Любое событие, в котором будет k успехов и $(n - k)$ неуспехов будет иметь такую же вероятность, а число таких событий определяется как коли-

чество способов выбора k мест из n, т. е. равно числу сочетаний из n по k: C_n^k. Таким образом, искомая вероятность равна $P[X = k] = C_n^k \cdot p^k \cdot q^{n-k}$.

Пример 5. В комнате горят 5 лампочек, каждая из которых в течение месяца может перегореть с вероятностью 0,3. Найти вероятность того, что в течение месяца перегорит: а) ровно три лампочки, б) хотя бы одна лампочка.

Решение. Пусть A — событие, состоящее в том, что лампочка перегорит в течение месяца. По условию задачи: $p = 0{,}3$; $q = 1 - 0{,}3 = 0{,}7$; $n = 5$. Обозначим через X число перегоревших лампочек. По формуле (12) имеем:

а) $P[X = 3] = C_5^3 \cdot (0{,}3)^3 \cdot (0{,}7)^2 \approx 0{,}132$;

б) $P[X \geq 1] = 1 - P[X = 0] = 1 - (0{,}7)^5 \approx 1 - 0{,}168 = 0{,}832$.

Вычислим математическое ожидание и дисперсию случайной величины X, имеющей биномиальное распределение. Чтобы упростить задачу, введем случайные величины X_i:

$$X_i = \begin{cases} 1, & \text{если в } i\text{-м испытании был успех — событие } A, \\ 0, & \text{если в } i\text{-м испытании был неуспех — событие } \overline{A}, \end{cases}$$

$i = 1, 2, \ldots, n$.

Распределение X_i задается таблицей

X_i	0	1
$P[X_i = k]$	$q = 1 - p$	p

Найдем математическое ожидание и дисперсию X_i, используя формулы (2) и (5)

$$M[X_i] = 0 \cdot q + 1 \cdot p = p, \tag{13}$$

$$D[X_i] = 0^2 \cdot q + 1^2 \cdot p - p^2 = p - p^2 = p(1 - p) = pq. \tag{14}$$

Очевидно, число успехов в n испытаниях X можно вычислить как сумму

$$X = X_1 + X_2 + \ldots + X_n,$$

причем $X_1, X_2, \ldots, X_n$ — независимые случайные величины, $i = 1, 2, \ldots, n$.

Так как математическое ожидание суммы случайных величин равно сумме математических ожиданий получим

$$M[X] = M[X_1] + M[X_2] + \ldots + M[X_n] = np. \tag{15}$$

Дисперсия суммы независимых случайных величин равна сумме дисперсий слагаемых, следовательно

$$D[X] = D[X_1] + D[X_2] + \ldots + D[X_n] = npq. \tag{16}$$

Распределение Пуассона, Pu(λ). Это распределение дискретной случайной величины X, принимающей значения 0, 1, 2, ... с вероятностями

$$P[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}, \tag{17}$$

где λ — параметр распределения Пуассона ($\lambda > 0$).

Нетрудно показать, что формула (17) определяет распределение вероятностей. Действительно

$$\sum_{k=0}^{\infty} P[X=k] = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

Математическое ожидание и дисперсия этого распределения равны λ: $M[X] = D[X] = \lambda$.

Пример 6. Предположим, что в большой партии изделий число бракованных изделий есть случайная величина X, имеющая распределение Пуассона с параметром $\lambda = 2$. Найти вероятность того, что в партии окажется более трех бракованных изделий.

Решение. По формуле (17) получим

$$P[X > 3] = \sum_{k=4}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda},$$

или, перейдя к противоположному событию

$$P[X > 3] = 1 - P[X \le 3] = 1 - P[X = 0] - P[X = 1] - P[X = 2] - P[X = 3] =$$

$$= 1 - e^{-2} - \frac{2}{1!} e^{-2} - \frac{2^2}{2!} e^{-2} - \frac{2^3}{3!} e^{-2} \approx 0{,}14288.$$

Приближенная формула Пуассона для вычисления вероятностей в биномиальном распределении. Пусть случайная величина X имеет биномиальное распределение. Если $n \to \infty$ и $p \to 0$ так, что $np \to \lambda$, то

$$P[X = k] = C_n^k \cdot p^k \cdot q^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots, n. \tag{18}$$

Приближение вполне приемлемо при $n > 10$, $\lambda < 0{,}1$.

Ошибка формулы (18) не превышает величины np^2.

Пример 7. Найти вероятность того, что среди 200 изготовленных деталей окажется ровно 4 бракованных, если известно, что вероятность изготовления бракованной детали составляет 0,01.

Решение. Для решения данной задачи следует использовать биномиальное распределение. Пусть X — число бракованных изделий среди изготовленных 200 деталей. Тогда искомая вероятность по формуле (12) равна

$$P[X = 4] = C_{200}^4 \cdot p^4 \cdot q^{200-4}, \text{ где } p = 001;\ q = 1 - 0{,}01 = 0{,}99.$$

В данной задаче: n — велико, p — мало, $np = \lambda = 200 \cdot 0{,}01 = 2$ и, следовательно, можно воспользоваться приближенной формулой Пуассона (18)

$$P[X = 4] \approx \frac{2^4}{4!} e^{-2} = 0{,}09022.$$

Геометрическое распределение, Ge(p). Пусть p — вероятность появления некоторого события A («успеха») в данном испытании. Испытания повто-

ряются до тех пор пока событие A не появится. Число испытаний, которое нужно провести для получения первого «успеха» есть дискретная случайная величина X, принимающая значения: 1, 2, 3, ..., k, ... Распределение случайной величины X определяется формулой

$$p_k = P[X = k] = q^{k-1}p, \tag{19}$$

где $q = 1 - p$, и называется *геометрическим*.

Сумма вероятностей p_k равна единице

$$\sum_{k=1}^{\infty} p_k = \sum_{k=1}^{\infty} q^{k-1}p = p\sum_{k=1}^{\infty} q^{k-1} = p\frac{1}{1-q} = 1.$$

Математическое ожидание и дисперсия геометрического распределения равны: $M[X] = \frac{1}{p}$, $D[X] = \frac{q}{p^2}$.

Замечание. В пакете STATISTICA (см. п. 2.1) геометрическое распределение определяется несколько иначе, а именно: это дискретная случайная величина Y, принимающая значения: 0, 1, 2, ..., и равная числу «неуспехов», предшествующих появлению первого «успеха».

Распределение случайной величины Y определяется формулой: $p_k = P[Y = k] = q^k p$.

Связь между случайными величинами Y и X очевидна: $Y = X - 1$.

Математическое ожидание и дисперсия случайной величины Y равны:

$$M[Y] = \frac{q}{p}; \quad D[Y] = \frac{q}{p^2}.$$

Пример 8. Предположим, что шанс на выигрыш в одном сеансе игры с игральным автоматом составляет 1 к 10, а денег хватает только на оплату десяти сеансов игры.

а. Какова вероятность получить выигрыш в первом, пятом и восьмом сеансах, если после получения выигрыша игра прекращается?

б. Какова вероятность хотя бы одного выигрыша, если решено провести 10 сеансов игры?

в. Какова вероятность одного выигрыша в 10 сеансах игры?

Решение.

а. Искомые вероятности определяются по формуле геометрического распределения (19) и равны:

$$p_1 = p = 0{,}1,$$

$$p_5 = pq^{5-1} = 0{,}1 \cdot (0{,}9)^4 = 0{,}0656,$$

$$p_8 = pq^{8-1} = 0{,}1 \cdot (0{,}9)^7 = 0{,}0478.$$

б. Найдем вероятность того, что в десяти сеансах не будет ни одного выигрыша (событие A):

$$P(A) = q^{10} = (0{,}9)^{10} \approx 0{,}3488.$$

Тогда получение хотя бы одного выигрыша в 10 сеансах есть событие $\bar{A}$ и его вероятность равна

$$P(\bar{A}) = 1 - P(A) \approx 1 - 0{,}3488 \approx 0{,}6512.$$

Этот же результат получим, вычислив сумму вероятностей появления 1-го успеха при k испытаниях, где $k = 1, 2, ..., 10$

$$P(\bar{A}) = \sum_{k=1}^{10} q^{k-1} \cdot p = \frac{q^{10} - 1}{q - 1} \cdot p = 1 - q^{10} = 1 - (0{,}9)^{10} = 0{,}6511.$$

в. Вероятность одного выигрыша в 10-ти сеансах игры вычисляется по формуле биномиального распределения (12) при $k = 1$, $n = 10$, $p = 0{,}1$, $q = 1 - 0{,}1 = 0{,}9$

$$P_1 = C_{10}^1 \cdot (0{,}1)^1 \cdot (0{,}9)^{10-1} = 0{,}387.$$

П.2.7. Числовые характеристики системы двух случайных величин. Ковариация и коэффициент корреляции

Рассмотрим две дискретные величины X и Y, распределение которых задается табл. 2. Числовыми характеристиками системы (X, Y) двух дискретных случайных величин являются:

- математические ожидания $M[X] = m_x$ и $M[Y] = m_y$ вычисляемые по формуле (2):

$$M[X] = m_x = \sum_{i=1}^{n} x_i p_{x_i},$$

$$M[Y] = m_y = \sum_{j=1}^{m} y_j p_{y_j},$$

- дисперсии $D[X] = \sigma_x^2$ и $D[Y] = \sigma_y^2$, где σ_x и σ_y — средние квадратические или стандартные отклонения, вычисляемые по формуле (4):

$$D[X] = \sigma_x^2 = M[(X - m_x)^2] = \sum_{i=1}^{n} (x_i - m_x)^2 p_{x_i},$$

$$D[Y] = \sigma_y^2 = M[(Y - m_y)^2] = \sum_{j=1}^{m} (y_j - m_y)^2 p_{y_j}.$$

Характеристикой взаимозависимости X и Y является их ковариация – $\mathrm{cov}(X, Y)$.

Определение 10. Пусть случайные величины X и Y имеют конечные дисперсии. *Ковариацией* X и Y называется математическое ожидание произведения центрированных случайных величин $(X - m_x)$ и $(Y - m_y)$:

$$\mathrm{cov}(X, Y) = M[(X - m_x)(Y - m_y)]. \qquad (20)$$

Используя свойства математических ожиданий, формулу (20) можно преобразовать к следующему виду

$$\mathrm{cov}(X, Y) = M[XY] - m_x M[Y] - m_x M[X] + m_x m_y = M[XY] - m_x m_y. \qquad (21)$$

Для системы дискретных случайных величин X и Y ковариация вычисляется по формуле

$$\operatorname{cov}(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i - m_x)(y_j - m_y)p_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m}x_i y_j p_{ij} - m_x m_y. \qquad (22)$$

Свойства ковариации

1. Если X и Y независимые случайные величины, то $\operatorname{cov}(X, Y) = 0$.

Действительно, по формуле (21) имеем

$$\operatorname{cov}(X, Y) = M[XY] - m_x m_y = M[X]M[Y] - m_x m_y = 0.$$

Обратное утверждение, вообще говоря, неверно.

Из того, что коэффициент ковариации равен нулю, не следует независимость случайных величин X и Y. Приведем следующий пример.

Пример 9. Пусть случайная величина X принимает значения ± 1, ± 2, причем, каждое значение принимается с одной и той же вероятностью равной 1/4. Иными словами, ряд распределения случайной величины X имеет вид

X	-2	-1	1	2
$P[X = x_i]$	1/4	1/4	1/4	1/4

Рассмотрим случайную величину $Y = X^2$. Случайная величина Y имеет ряд распределения

Y	1	4
$P[Y = y_i]$	1/2	1/2

Совместное распределение вектора (X, Y) задается таблицей распределения:

Y \ X	-2	-1	1	2
1	0	1/4	1/4	0
4	1/4	0	0	1/4

Вычислим ковариацию X и Y. Предварительно вычислим математические ожидания X и Y по формуле (2), получим

$$m_x = (-2)\cdot\frac{1}{4} + (-1)\cdot\frac{1}{4} + 1\cdot\frac{1}{4} + 2\cdot\frac{1}{4} = 0;$$

$$m_y = 1\cdot\frac{1}{2} + 4\cdot\frac{1}{2} = 2{,}5.$$

По формуле (22) вычислим ковариацию X и Y:

$$\operatorname{cov}(X,Y) = (-2)\cdot 1\cdot 0 + (-1)\cdot 1\cdot \frac{1}{4} + 1\cdot 1\cdot \frac{1}{4} + 2\cdot 1\cdot 0 + (-2)\cdot 4\cdot \frac{1}{4} +$$

$$+(-1)\cdot 4\cdot 0 + 1\cdot 4\cdot 0 + 2\cdot 4\cdot \frac{1}{4} - 0\cdot 2{,}5 = 0.$$

Таким образом, $\operatorname{cov}(X, Y) = 0$, хотя случайные величины X и Y не являются независимыми и, более того, они функционально связаны: $Y = X^2$.

2. $\operatorname{cov}(aX, bX) = ab\operatorname{cov}(X,Y)$, где a и b — константы.

3. $\operatorname{cov}(X,Y) \le \sqrt{D[X]\cdot D[Y]}$.

Это неравенство является следствием неравенства Коши—Буняковского:

$$(M[XY])^2 \le M[X^2]\cdot M[Y^2]. \tag{23}$$

Доказательство неравенства (23). Рассмотрим очевидное неравенство

$$M[(aX + Y)^2] \ge 0,$$

где a — любое действительное число, $a \ne 0$.

Преобразуем левую часть этого неравенства, используя свойства математических ожиданий

$$M[(aX + Y)^2] = M[a^2X^2 + 2aXY + Y^2] = a^2M[X^2] + 2aM[XY] + M[Y^2] \ge 0.$$

Так как полученный относительно a квадратный трехчлен принимает только неотрицательные значения, то его дискриминант будет меньше или равен нулю

$$4(M[XY])^2 - 4M[X^2]\cdot M[Y^2] \le 0.$$

Отсюда следует неравенство (23).

Для доказательства свойства 3 заменим в неравенстве (23) X на $(X - m_x)$, а Y на $(Y - m_y)$, получим:

$$(M[(X - m_x)(Y - m_y)])^2 \le M[(X - m_x)^2]\cdot M[(Y - m_y)^2]$$

или

$$(\operatorname{cov}(X,Y))^2 \le D[X]\cdot D[Y].$$

В качестве характеристики меры линейной зависимости между случайными величинами X и Y используют *коэффициент корреляции*, вычисляемый по формуле

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_x\cdot\sigma_y}. \tag{24}$$

Свойства коэффициента корреляции:

1. $|\rho(X,Y)| \le 1$, этот результат следует из свойства 3 для ковариации случайных величин X и Y.

2. Если X и Y — независимые случайные величины, то $\rho(X,Y)=0$ (см. свойство 1 для ковариации).

3. Если X и Y связаны линейной функциональной зависимостью: $Y=aX+b$, где a и b — константы, $a\neq 0$, то $|\rho(X,Y)|=1$.

Доказательство. Так как $M[Y]=aM[X]+b=am_x+b$, то имеем

$$\operatorname{cov}(X,Y)=M[(X-m_x)(Y-m_y)]=M[(X-m_x)(aX+b-am_x-b)]=$$

$$=M[(X-m_x)a(X-m_x)]=aD[X].$$

Вычислим дисперсию случайной величины $Y=aX+b$

$$D[Y]=D[aX+b]=a^2D[X].$$

Таким образом, коэффициент корреляции равен

$$\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sqrt{D[X]\cdot D[Y]}}=\frac{aD[X]}{\sqrt{D[X]\cdot a^2D[X]}}=\frac{a}{|a|}.$$

Следовательно, $\rho(X,Y)=1$, если $a>1$ и $\rho(X,Y)=-1$, если $a<0$.

Свойство 3 дает возможность трактовать коэффициент корреляции как меру (или индикатор) *линейной зависимости* между случайными величинами X и Y.

В примере 9 мы видели, что коэффициент корреляции равен нулю, то есть линейная зависимость между X и Y отсутствует, хотя случайные величины X и Y не являлись независимыми — они были связаны функционально квадратической зависимостью $Y=X^2$.

Если коэффициент корреляции между случайными величинами X и Y равен нулю ($\rho=0$), то говорят, что X и Y *некоррелированны*.

Некоррелированность случайных величин X и Y означает только, что между ними *нет линейной зависимости* и не означает статистическую независимость случайных величин X и Y.

Рассмотрим пример на вычисление числовых характеристик случайного вектора (X, Y).

Пример 10. Вычислим числовые характеристики системы случайных величин X и Y из примера 3. Распределение (X, Y) задается табл. 3.

Решение. Ряды распределений случайных величин X и Y имеют вид:

X	-1	1
$P[X=x_i]$	2/6	4/6

Y	1	2	3	4	5	6
$P[Y=y_j]$	1/6	1/6	1/6	1/6	1/6	1/6

По формуле (2) имеем:

$$M[X] = -1 \cdot \frac{2}{6} + 1 \cdot \frac{4}{6} = \frac{2}{6},$$

$$M[Y] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6}.$$

Используя формулу (5), получим:

$$D[X] = (-1)^2 \cdot \frac{2}{6} + 1^2 \cdot \frac{4}{6} - \left(\frac{2}{6}\right)^2 = \frac{8}{9}; \quad \sigma_x = \sqrt{\frac{8}{9}},$$

$$D[Y] = (1)^2 \cdot \frac{1}{6} + \ldots + 6^2 \cdot \frac{1}{6} - \left(\frac{21}{6}\right)^2 = \frac{35}{12}; \quad \sigma_y = \sqrt{\frac{35}{12}}.$$

По формуле (22), используя таблицу совместного распределения случайных величин X и Y (табл. 3), находим

$$\operatorname{cov}(X,Y) = (-1) \cdot 1 \cdot 0 + (-1) \cdot 2 \cdot 0 + (-1) \cdot 3 \cdot 1/6 + (-1) \cdot 4 \cdot 0 + (-1) \cdot 5 \cdot 0 +$$

$$+ (-1) \cdot 6 \cdot 1/6 + 1 \cdot 1 \cdot 1/6 + 1 \cdot 2 \cdot 1/6 + 1 \cdot 3 \cdot 0 + 1 \cdot 4 \cdot 1/6 + 1 \cdot 5 \cdot 1/6 +$$

$$+ 1 \cdot 6 \cdot 0 - 2/6 \cdot 21/6 = -2/3.$$

Коэффициент корреляции по формуле (24) равен

$$\rho(X, Y) = \frac{-2/3}{\sqrt{\frac{8}{9}} \cdot \sqrt{\frac{35}{12}}} \approx -0{,}414.$$

Найдем условное математическое ожидание $M[Y/X = 1]$.

Зная, что $P[X = 1] = \frac{4}{6}$, по формуле (4) имеем

$$M[Y/X = 1] = \frac{1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot 0 + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot 0}{\frac{4}{6}} = 3.$$

П.3. Непрерывные случайные величины

П.3.1. Определение непрерывной случайной величины

Непрерывные случайные величины могут принимать любые действительные значения из некоторого интервала.

Примерами таких случайных величин являются:

1) рост людей определенной возрастной группы;

2) время отказа прибора после его включения;

3) величина душевого дохода.

Формально непрерывная случайная величина X определяется следующим образом.

Определение 1. Пусть (Ω, F, P) — вероятностное пространство. Случайной величиной X называется числовая функция $X = X(\omega)$, $\omega \in \Omega$, такая, что при любом действительном x

$$\{\omega: X(\omega) < x\} \in F, \tag{1}$$

где F — система подмножеств Ω, являющаяся σ-алгеброй (см. П.1.3).

Случайное событие в (1) будем записывать как $[X < x]$. Так как операции отрицания, произведения и суммы над событиями из F, не выводят за пределы F, то события:

$$[X \geq x] = \overline{[X < x]} \in F,$$

$$[x_1 \leq X < x_2] = [X \geq x_1] \cdot [X < x_2] \in F,$$

$$[X = x] = \prod_{n=1}^{\infty} [x \leq X \leq x + \frac{1}{n}] \in F.$$

Таким образом, вероятности всех этих событий определены и могут быть вычислены с помощью функции распределения $F(x)$, где

$$F(x) = P[X < x].$$

Например: $P[X \geq x] = 1 - F(x)$; так как $[X < x_2] = [X < x_1] + [x_1 \leq X < x_2]$, то $P[X < x_2] = P[X < x_1] + P[x_1 \leq X < x_2]$, следовательно,

$$P[x_1 \leq X < x_2] = F(x_2) - F(x_1). \tag{2}$$

Свойства функции распределения $F(x)$ были сформулированы выше (см. П.2.3).

В приложениях важнейшую роль играют абсолютно непрерывные случайные величины.

Определение 2. Случайная величина X называется *абсолютно непрерывной*, если существует неотрицательная функция $f(x)$, такая, что при любом x, функция распределения $F(x)$ может быть представлена в виде

$$F(x) = \int_{-\infty}^{x} f(t)dt. \tag{3}$$

Предполагается, что $f(x)$ непрерывна всюду, за исключением конечного числа точек.

Функция $f(x)$ называется *плотностью распределения вероятностей.*

Очевидно, $P[x_1 \leq X \leq x_2] = \int_{x_1}^{x_2} f(t)dt.$

Свойства плотности распределения:

1. $f(x) \geq 0$, $-\infty < x < \infty$.

2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

3. $F'(x) = f(x)$ в точках непрерывности $f(x)$.

Свойства 1 и 2 являются характеристическими, т. е. любая функция $f(x)$ удовлетворяющая этим свойствам может быть плотностью распределения некоторой случайной величины.

Для абсолютно непрерывной случайной величины X функция распределения $F(x)$ непрерывна всюду. Это следует из свойств несобственного интеграла (3), так как $f(x)$ может иметь лишь конечное число точек разрыва. Из непрерывности $F(x)$ следует, что для непрерывной случайной величины X вероятность «попадания в точку» равна нулю

$$P[X = x] = \lim_{\Delta x \to 0} P[x \leq X \leq x + \Delta x] = \lim_{\Delta x \to 0}(F(x + \Delta x) - F(x)) = 0.$$

Вероятностный смысл плотности распределения определяется следующей формулой

$$f(x)dx = dF(x) = P[x \leq X < x + dx].$$

В связи с этим, дифференциал $f(x)dx$ называют *элементом вероятности.*

Заметим, что не всякая непрерывная функция распределения $F(x)$ может быть представлена в виде (3), т. е. не всякая непрерывная случайная величина является абсолютно непрерывной. Однако в приложениях обычно встречаются абсолютно непрерывные случайные величины, поэтому в дальнейшем будем называть абсолютно непрерывные случайные величины просто непрерывными случайными величинами.

Интересна и показательна механическая интерпретация вероятностного распределения непрерывной случайной величины X. А именно: рассмотрим распределение единичной массы на конечном или бесконечном интервале числовой оси. Функцию плотности распределения массы $f(x)$, имеющую очевидные свойства:

$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x)dt = 1,$$

можно рассматривать как плотность распределения вероятностей. Масса, сосредоточенная на полуинтервале $[x_1; x_2)$ равна вероятности попадания в этот полуинтервал и вычисляется по формуле

$$P[x_1 \leq X < x_2] = \int_{x_1}^{x_2} f(t)dt.$$

Функция распределения $F(x)$ есть, очевидно, масса, распределенная левее точки с абсциссой x

$$F(x) = P[X < x] = \int_{-\infty}^{x} f(t)dt.$$

П.3.2. Системы нескольких случайных величин

Пусть на вероятностном пространстве (Ω, F, P) заданы непрерывные случайные величины $X_1 = X_1(\omega)$, $X_2 = X_2(\omega)$, ..., $X_n = X_n(\omega)$, $\omega \in \Omega$.

Определение 3. Совместной функцией распределения $F(x_1, x_2, ..., x_n)$ случайных величин $X_1, X_2, ..., X_n$ называется вероятность события $[X_1 < x_1; X_2 < x_2; ..., X_n < x_n]$: $F(x_1, x_2, ..., x_n) = P[X_1 < x_1; X_2 < x_2; ..., X_n < x_n]$.

Неотрицательная функция n переменных $f(x_1, x_2, ..., x_n)$ называется совместной плотностью распределения случайных величин $X_1, X_2, ..., X_n$, если их совместная функция распределения может быть представлена в виде

$$F(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}...\int_{-\infty}^{x_n} f(x_1, x_2, ..., x_n)dx_1dx_2 \ ... \ dx_n.$$

Плотность распределения имеет следующие свойства:

1. $f(x_1, x_2, ..., x_n) \geq 0$;

2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty} f(x_1, x_2, ..., x_n)dx_1dx_2 \ ... \ dx_n = 1$;

3. $P[(x_1, x_2, ..., x_n) \in G] = \iint_G ...\int f(x_1, x_2, ..., x_n)dx_1dx_2 \ ... \ dx_n$.

Зная совместную плотность распределения $f(x_1, x_2, ..., x_n)$ случайных величин $X_1, X_2, ..., X_n$ можно найти плотность распределения каждой случайной величины. Для двумерного случайного вектора (X_1, X_2) с плотностью $f(x_1, x_2)$ плотность распределения случайной величины X_1, $f_1(x_1)$ равна

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2,$$

а плотность распределения случайной величины X_2, $f_2(x_2)$ равна

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_1.$$

Аналогично вычисляются одномерные плотности в случае бóльшего числа переменных.

Определение 4. Случайные величины $X_1, X_2, ..., X_n$ называются *независимыми*, если для любых действительных переменных $x_1, x_2, ..., x_n$, $F(x_1, x_2, ..., x_n) = F_1(x_1) \cdot F_2(x_2) \cdot ... \cdot F_n(x_n)$, где $F_i(x_i)$ — функция распределения случайной величины X_i, $i = 1, 2, ..., n$.

Равносильное определение независимости случайных величин $X_1, X_2, ..., X_n$ записывается так

$$f(x_1, x_2, ..., x_n) = f_1(x_1) \cdot f_2(x_2) \cdot ... \cdot f_n(x_n),$$

где $f_i(x_i)$ — плотность распределения случайной величины X_i, $i = 1, 2, ..., n$.

П.3.3. Числовые характеристики непрерывных случайных величин

Определение 5. Пусть X — непрерывная случайная величина с плотностью распределения $f(x)$. Математическим ожиданием $M[X]$ называется число $M[X] = m$, определяемое по формуле

$$M[X] = m = \int_{-\infty}^{\infty} xf(x)dx, \tag{4}$$

если интеграл (4) абсолютно сходится.

Если интеграл (4) не сходится абсолютно, то говорят, что математическое ожидание случайной величины X не существует.

Свойства математического ожидания:

1. $M[c] = c$, где c — константа.

2. $M[cX] = cM[X]$ и $M[c + X] = c + M[X]$, где c — константа.

3. Пусть X_1, X_2, ..., X_n — случайные величины с конечными математическими ожиданиями. Математическое ожидание их суммы равно сумме их математических ожиданий:

$$M[X_1 + X_2 + \ldots + X_n] = M[X_1] + M[X_2] + \ldots + M[X_n].$$

4. Пусть X и Y — взаимно независимые случайные величины с конечными математическими ожиданиями. Математическое ожидание произведения XY равно произведению их математических ожиданий:

$$M[XY] = M[X] \cdot M[Y].$$

Сравнивая формулу (4) с формулой (2) из П.2.5 для вычисления математического ожидания дискретной случайной величины X, принимающей значение x_1, x_2, ..., x_n с вероятностью p_1, p_2, ...

$$M[X] = \sum_i x_i p_i$$

видно, что при переходе от дискретных случайных величин к непрерывным случайным величинам формулы вычисления математических ожиданий преобразуются по следующему правилу: значения дискретной случайной величины x_1, x_2, ... заменяются переменной x, значения вероятностей p_1, p_2, ... заменяются элементом вероятности $f(x)dx$, а операция суммирования заменяется операцией интегрирования.

Таким образом получаем следующие формулы для вычисления числовых характеристик непрерывной случайной величины X с плотностью распределения $f(x)$:

1. Если $Z = h(X)$, где $h(x)$ — заданная функция x, то $M[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx$.

2. Дисперсия: $D[X] = \int_{-\infty}^{\infty} (x - m)^2 f(x)dx$, где $m = M[X]$.

Основные свойства дисперсии:

1) $D[X] \geq 0$;

2) если случайная величина X постоянна, т. е. $X(\Omega) = \text{const} = c$, то ее дисперсия равна нулю: $D[c] = 0$;

3) $D[cX] = c^2 D[X]$, где $c = \text{const}$;

4) $D[X + c] = D[X]$;

5) дисперсия суммы независимых случайных величин X и Y равна сумме их дисперсий

$$D[X + Y] = D[X] + D[Y].$$

3. Начальные и центральные моменты k-го порядка:

$$\alpha_k = M[X^k] = \int_{-\infty}^{\infty} x^k f(x)dx,$$

$$\mu_k = M[(X - m)^k] = \int_{-\infty}^{\infty} (x - m)^k f(x)dx, \; k = 1, 2, \ldots .$$

Предполагается, что соответствующие числовые характеристики случайной величины X существуют, если интегралы в правых частях формул абсолютно сходятся.

Важными характеристиками распределения непрерывной случайной величины X являются ее квантили.

Определение 6. Пусть p — действительное число из отрезка [0; 1]. Квантилью случайной величины X порядка p называется действительное число x_p, являющееся решением уравнения

$$F(x_p) = p, \tag{5}$$

где $F(x)$ — функция распределения случайной величины X.

Уравнение (5) имеет единственное решение, если $F(x)$ — строго монотонная функция x.

Квантиль порядка $p = 0{,}5$ называется *медианой* h_x случайной величины X:

$$h_x = x_{0,5}.$$

Модой случайной величины X называется действительное число d, являющееся точкой максимума ее плотности распределения вероятности $f(x)$, при условии, что $f(x)$ имеет единственную точку максимума; в противном случае мода случайной величины X не определена.

Пусть $f(x_1, x_2)$ — совместная плотность распределения случайных величин X_1 и X_2.

Ковариация $\text{cov}(X_1, X_2)$ вычисляется по формуле

$$\text{cov}(X_1, X_2) = M[(X_1 - m_1)(X_2 - m_2)] =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1 - m_1)(x_2 - m_2) f(x_1, x_2)dx_1 dx_2,$$

где m_1 и m_2 — математические ожидания случайных величин X_1 и X_2.

Свойства и интерпретация числовых характеристик для дискретных случайных величин, приведенные в п. П.2.5 и П.2.7, остаются теми же и для числовых характеристик непрерывных случайных величин.

П.3.4. Примеры непрерывных распределений: равномерное и экспоненциальное (показательное) распределения

Равномерное распределение на $[a, b]$, $R(a, b)$, имеет плотность, определяемую формулой

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b]; \\ 0, & x \notin [a,b]. \end{cases}$$

График плотности равномерного распределения показан на рис. 9.

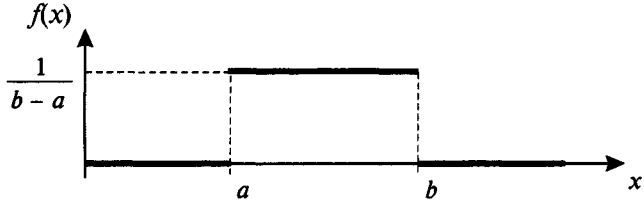


Рис. 9. График плотности равномерного распределения

Функция распределения $F(x)$ случайной величины, имеющей равномерное распределение равна

$$F(x) = \int_{-\infty}^{x} f(x)dx = \begin{cases} 0, & x \le a; \\ \dfrac{x-a}{b-a}, & a < x < b; \\ 1, & x \ge b. \end{cases}$$

График функции распределения приведен на рис. 10.

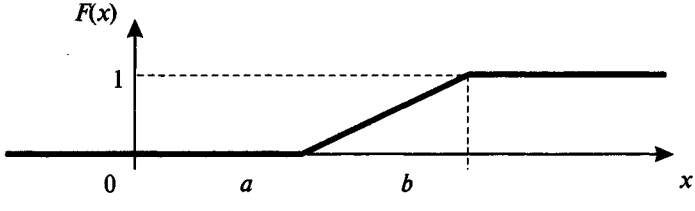


Рис. 10. График функции распределения для равномерного распределения

Найдем математическое ожидание $M[X]$ и дисперсию $D[X]$ случайной величины, имеющей равномерное распределение.

По формуле (4):

$$M[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_a^b \frac{x}{b-a}dx = \frac{1}{b-a} \cdot \frac{x^2}{2}\Big|_a^b = \frac{a+b}{2},$$

$$D[X] = M[X^2] - (M[X])^2 = \int_a^b \frac{x^2}{b-a}dx - \left(\frac{a+b}{2}\right)^2 =$$

$$= \frac{1}{b-a} \cdot \frac{x^3}{3}\Big|_a^b - \frac{(a+b)^2}{4} = \frac{(b^3-a^3)}{3 \cdot (b-a)} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

Пример 1. Цена деления шкалы измерительного прибора равна 0,2. Показание прибора округляются до ближайшего целого деления. Считая, что ошибка округления есть случайная величина, имеющая равномерное распределение, определить плотность распределения ошибки округления. Найти вероятность того, что: а) ошибка округления по абсолютной величине не превзойдет 0,05; б) будет по абсолютной величине более среднего квадратичного отклонения.

Решение. Ошибка округления X может изменяться в пределах от –0,1 до +0,1.

Плотность распределения имеет вид: $f(x) = \begin{cases} \dfrac{1}{0,2} = 5, \text{ если } x \in [-0,1;\ 0,1]; \\ 0, \text{ если } x \notin [-0,1;\ 0,1]. \end{cases}$

а) $P[|X| \le 0,05] = \int_{-0,05}^{0,05} f(x)dx = 5 \cdot x\Big|_{-0,05}^{0,05} = 0,5;$

б) $D[X] = \dfrac{(b-a)^2}{12} = \dfrac{0,2^2}{12} \approx 0,0033;$

$\sigma = \sqrt{D[X]} \approx 0,0577;$

$P[|X| > \sigma] = 1 - P[|X| \le \sigma] = 1 - \int_{-0,0577}^{0,0577} f(x)dx \approx 0,423.$

Экспоненциальное (показательное) распределение, Ex(λ), имеет плотность, определяемую формулой

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0; \\ 0, & x < 0, \end{cases}$$

где $\lambda > 0$ — параметр экспоненциального распределения.

График плотности экспоненциального распределения показан на рис. 11.

Функция распределения $F(x)$ случайной величины X, имеющей экспоненциальные распределения, равна

$$F(x) = \begin{cases} 0, \text{ при } x \le 0; \\ \int_0^x \lambda e^{-\lambda t}dt = 1 - e^{-\lambda x}, \text{ при } x > 0. \end{cases}$$

График функции распределения приведен на рис. 12.

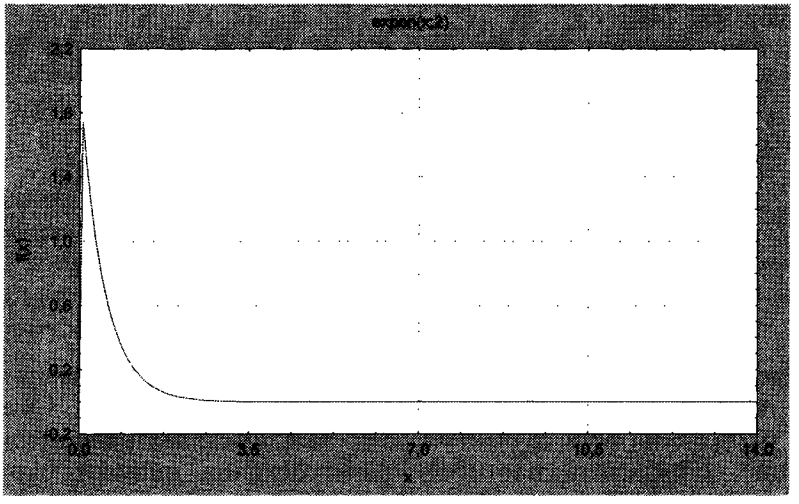


Рис. 11. График функции плотности распределения для экспоненциального закона Ex(2)

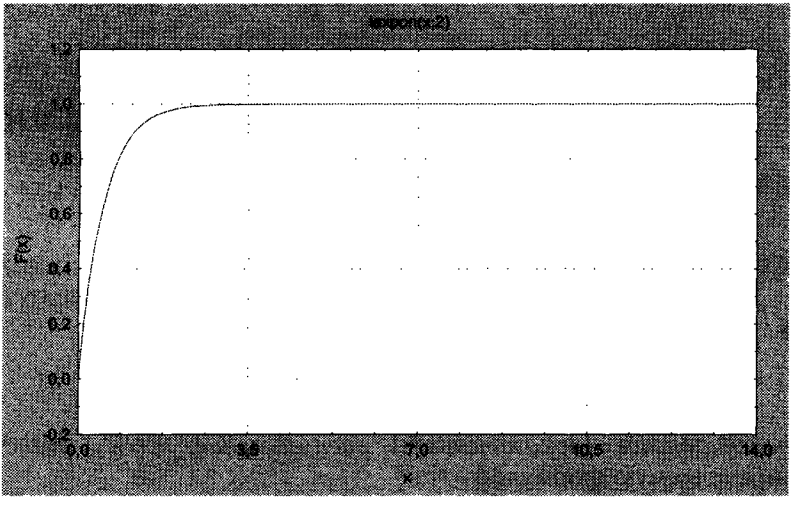


Рис. 12. График функции распределения для экспоненциального закона Ex(2)

Математическое ожидание

$$M(X) = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Дисперсия

$$D(X) = M[X^2] - (M[X])^2 = \int_0^\infty x^2\lambda e^{-\lambda x} dx - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

В приложениях теории вероятностей экспоненциальное распределение имеют случайные величины типа «времени жизни». Например, распределение времени безотказной работы прибора при постоянной интенсивности отказов. Распределение времени между моментами наступления независимых событий (таких, как телефонные звонки, приход клиентов и т. д.), появляющихся с постоянной интенсивностью, тоже имеют экспоненциальное распределение.

Пример 2. Время работы прибора — случайная величина, имеющая экспоненциальное распределение. Известно, что среднее время работы приборов данного типа равно 400 ч. Найти вероятность того, что прибор будет работать не менее 500 ч.

Решение. По условию задачи математическое ожидание случайной величины X — времени работы прибора — равно 400 ч, следовательно, $\lambda = 1/400$. Искомая вероятность

$$P[X \geq 500] = 1 - P[X < 500] = 1 - \int_0^{500} \lambda e^{-\lambda x} dx = 1 - (-e^{-\lambda x})\Big|_0^{500} =$$

$$= 1 - (1 - e^{-500/400}) = e^{-1,25} \approx 0,287.$$

П.3.5. Нормальное распределение

Нормальное распределение $N(m, \sigma^2)$ имеет плотность, определяемую формулой

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-m)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

График плотности нормального распределения приведен на рис. 13. Этот график симметричен относительно прямой $x = m$.

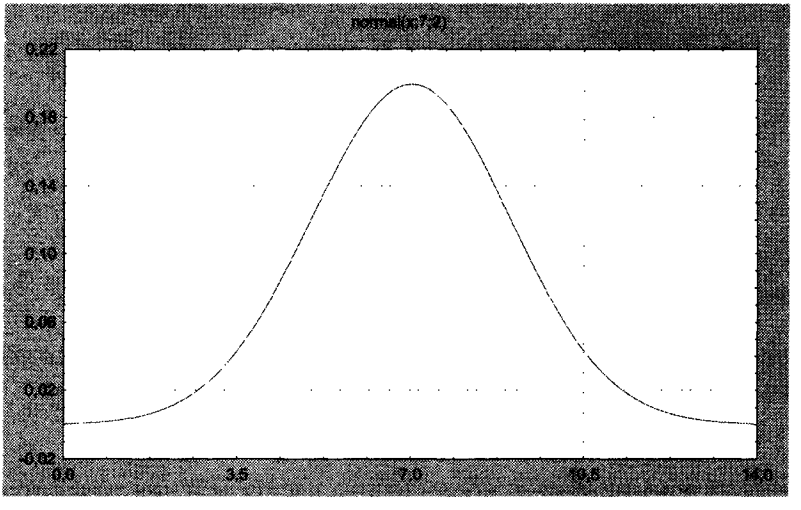


Рис. 13. График плотности нормального распределения $N(7, 4)$

Функция распределения $F(x)$ нормального распределения равна

$$F(x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-m)^2}{2\sigma^2}} dt. \qquad (5)$$

График этой функции приведен на рис. 14.

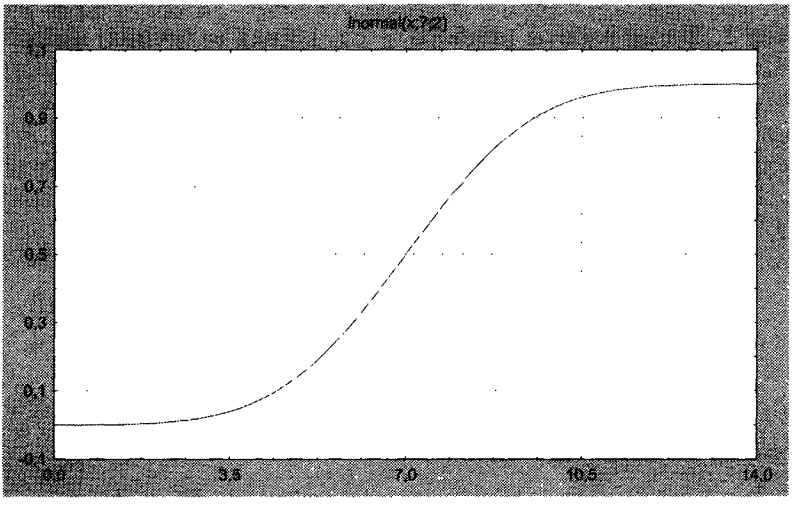


Рис. 14. График функции распределения для нормального закона $N(7, 4)$

Параметры m и σ^2 нормального распределения равны соответственно математическому ожиданию и дисперсии случайной величины X, так как:

$$M[X] = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}} dx = m,$$

$$D[X] = M[X^2] - (M[X])^2 = \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}} dx - m^2 = \sigma^2.$$

Центральные моменты нормального распределения можно вычислить из рекуррентного уравнения

$$\mu_{k+2} = (k+1)\sigma^2\mu_k, \quad k = 0, 1, 2, \ldots, \qquad (6)$$

причем $\mu_0 = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}} dx = 1.$

Из уравнения (6), в частности, следует, что для нормального распределения все центральные моменты нечетного порядка равны нулю

$$\mu_1 = \mu_3 = \mu_5 = \ldots = 0.$$

Коэффициент асимметрии a_x нормального распределения равен нулю

$$a_x = \mu_3/\sigma_3 = 0.$$

Из уравнения (6) получаем, что центральный момент четвертого порядка μ_4 равен

$$\mu_4 = 3\sigma^2\mu_2 = 3\sigma^4,$$

следовательно, коэффициент эксцесса e_x нормального распределения по формуле (10) П.2.5 равен нулю

$$e_x = \frac{\mu_4}{\sigma^4} - 3 = 0.$$

Нормальное распределение с нулевым математическим ожиданием, $m = 0$, и дисперсией, равной единице, $\sigma^2 = 1$, называется *стандартным нормальным распределением*. В дальнейшем, если случайная величина X имеет стандартное нормальное распределение, будем кратко записывать это так:

$$X \sim N(0, 1).$$

Функция плотности $\varphi(x)$ стандартного нормального закона равна

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \quad -\infty < x < \infty,$$

а функция распределения

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt. \tag{7}$$

Значения функции $\Phi(x)$ приводятся в специальных таблицах (см. Приложение 3), так как интеграл в правой части не вычисляется в элементарных функциях.

График плотности стандартного нормального распределения и распределение площадей под этим графиком приведены на рис. 15.

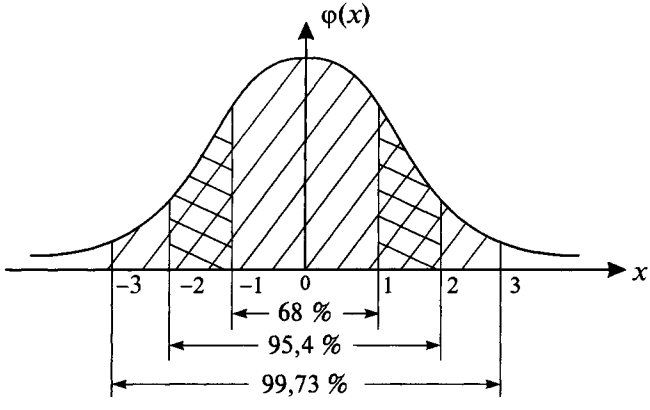


Рис. 15. График плотности стандартного нормального закона $N(0, 1)$

Так как плотность распределения стандартного нормального закона $\varphi(x)$ симметрична относительно оси ординат, то для функции распределения $\Phi(x)$ справедливо следующее свойство

$$\Phi(-x) = 1 - \Phi(x). \tag{8}$$

Значения функции $\Phi(x)$ используются при вычислении вероятности попадания нормально распределенной случайной величины X в заданный интервал.

Пусть $X \sim N(m, \sigma^2)$, тогда $P[a < X < b] = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx$.

В интеграле сделаем следующую замену переменной

$$\frac{(x-m)}{\sigma} = t; \quad \frac{dx}{\sigma} = dt.$$

Имеем, используя (7):

$$P[a < X < b] = \int_{(a-m)/\sigma}^{(b-m)/\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2}} \sigma dt = \Phi((b-m)/\sigma) - \Phi((a-m)/\sigma). \tag{9}$$

В практических задачах часто приходится вычислять вероятность попадания случайной величины $X \sim N(m, \sigma^2)$ в интервал, симметричный относительно ее математического ожидания m: $P[|X - m| < l]$.

$$P[|X - m| < l] = P[m - l < X < m + l] = \Phi((m + l - m)/\sigma) - \Phi((m - l - m)/\sigma) =$$

$$= \Phi(l/\sigma) - \Phi(-l/\sigma) = 2\Phi(l/\sigma) - 1. \tag{10}$$

Используя полученный результат, вычислим вероятность отклонения от математического ожидания нормально распределенной случайной величины на величину, равную трем среднеквадратическим отклонениям, 3σ

$$P[|X - m| < 3\sigma] = 2\Phi(3) - 1 \approx 2 \cdot 0{,}9987 - 1 \approx 0{,}9973.$$

Этот результат известен как «правило трех σ»: с вероятностью 0,9973 (практически равной единице) значение нормально распределенной случайной величины лежит в интервале $(m - 3\sigma; m + 3\sigma)$.

Пример 3. При измерении длины детали получаются случайные ошибки, подчиненные нормальному закону с нулевым математическим ожиданием и средним квадратическим отклонением, равным 10 мм. Найти вероятность того, что измерение будет проведено с ошибкой, по модулю не превосходящей 15 мм.

Решение. По условию задачи случайная величина X имеет нормальное распределение с математическим ожиданием равным нулю и средним квадратическим отклонением $\sigma = 10$ мм. Искомая вероятность вычисляется по формуле (10)

$$P[|X| < 15] = P[-15 < X < 15] = \Phi[(15 - 0)/10] - \Phi[(-15 - 0)/10] =$$

$$= \Phi(1{,}5) - \Phi(-1{,}5) = \Phi(1{,}5) - (1 - \Phi(1{,}5)) = 2\Phi(1{,}5) - 1.$$

По таблице (см. Приложение 3), получим: $\Phi(1,5) = 0,9332$, следовательно, $P[|X| < 15] = 2 \cdot 0,9332 - 1 = 0,8664$.

П.3.6. Двумерное нормальное распределение

Двумерное нормальное распределение — это распределение системы двух случайных величин (X, Y) с плотностью распределения $f(x,y)$, определяемой формулой

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2} - \right.\right.$$

$$\left.\left. - 2\rho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]\right\}. \qquad (11)$$

Плотность распределения случайной величины X, $f_1(x)$, вычисляется интегрированием $f(x,y)$ по y:

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{(x-m_1)^2}{2\sigma_1^2}\right]. \qquad (12)$$

Аналогично вычисляется плотность распределения случайной величины Y, $f_2(y)$:

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)dx = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left[-\frac{(y-m_2)^2}{2\sigma_2^2}\right]. \qquad (13)$$

Формулы (12), (13) показывают, что в случае двумерного нормального распределения с плотностью (11) компоненты X и Y имеют нормальное распределение, причем: $M[X] = m_1$, $D[X] = \sigma_1^2$, $M[Y] = m_2$, $D[Y] = \sigma_2^2$.

Ковариация X и Y равна

$$\operatorname{cov}(X, Y) = M[(X - m_1)(Y - m_2)] =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - m_1)(y - m_2) f(x,y)dxdy = \rho\sigma_1\sigma_2. \qquad (14)$$

Отсюда следует, что параметр ρ в (11) есть *коэффициент корреляции* между X и Y:

$$\rho(X,Y) = \rho.$$

Если X и Y некоррелированны, $\rho = 0$, то двумерная плотность (11) может быть представлена в виде произведения одномерных плотностей $f_1(x)$ — (12) и $f_2(y)$ — (13):

$$f(x,y) = f_1(x) \cdot f_2(y).$$

Таким образом, в случае двумерного нормального распределения *из некоррелированности компонент X и Y следует их независимость.*

Найдем плотность условного распределения случайной величины Y при условии, что $X = x$, $f(y/x)$:

$$f(y/x) = \frac{f(x,y)}{f_1(x)} =$$

$$= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_2^2}\left[(y-m_2) - \rho\frac{\sigma_2}{\sigma_1}(x-m_1)\right]^2\right\}. \qquad (15)$$

Из (15) следует, что условное распределение Y при условии, что $X = x$ — нормальное, причем условное математическое ожидание $M[Y/X = x]$ является функцией x и определяется формулой

$$M[Y/X = x] = m_2 + \rho\frac{\sigma_2}{\sigma_1}(x - m_1). \qquad (16)$$

Линейная функция x в правой части (16) называется *регрессией Y на x*. График этой линейной функции называется *прямой регрессии Y на x*.

Из (15) также следует, что условная дисперсия Y при условии, что $X = x$ определяется формулой

$$D[Y/X = x] = \sigma_2^2(1 - \rho^2).$$

Аналогично определяется условное распределение X при условии, что $Y = y$. Это распределение также нормальное, причем условное математическое ожидание равно

$$M[X/Y = y] = m_1 + \rho\frac{\sigma_1}{\sigma_2}(y - m_2).$$

Соответственно, линейная функция y в правой части называется *регрессией X на y*.

Геометрически плотность $f(x,y)$ двумерного нормального распределения (11) представляет «холмообразную» поверхность. Проекция вершины холма на плоскость xOy имеет координаты (m_1, m_2). Эта точка называется *центром рассеяния*.

Сечения поверхности $Z = f(x,y)$ плоскостями, параллельными плоскости xOy, есть кривые, определяемые уравнением

$$Q(x,y) = \frac{1}{1-\rho^2}\left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right] = \lambda^2, \qquad (17)$$

где λ — константа.

Проекциями кривых (17) на плоскость xOy будут эллипсы. Так как плотность $f(x,y)$ имеет на кривых (17) постоянное значение, то соответствующие эллипсы называют *эллипсами равных вероятностей*.

Эллипсы равных вероятностей имеют общий центр — центр рассеяния с координатами (m_1, m_2) и общие оси симметрии (они называются главными осями рассеяния ξ и η).

Квадратичная форма (17): $Q(x, y) = \lambda^2$ с помощью невырожденного линейного преобразования (поворота системы координат на некоторый угол α и параллельным переносом начала координат в центр рассеяния) может быть приведена к каноническому виду — сумме квадратов. Так как в этом случае плотность распределения может быть представлена как произведение одномерных плотностей, то это означает, что в новой системе координат случайные величины будут независимыми.

Таким образом, в случае двумерного нормального распределения, можно найти линейное преобразование переводящее систему случайных величин (X, Y) в систему независимых нормально распределенных случайных величин (ξ, η).

П4. Закон больших чисел и центральная предельная теорема

Определение 1. Последовательность случайных величин $X_1, X_2, ..., X_n, ...$ сходится по вероятности при $n \to \infty$ к случайной величине X, если для любого $\varepsilon > 0$, $\lim\limits_{n\to\infty} P[|X_n - X| \geq \varepsilon] = 0$.

Сходимость по вероятности при $n \to \infty$ записывается так: $X_n \xrightarrow{p} X$.

Закон больших чисел (теорема Хинчина)

Пусть $X_1, X_2, ..., X_n$ последовательность взаимно независимых одинаково распределенных случайных величин с конечным математическим ожиданием $m = M[X_k]$, $k = 1, 2, ..., n$. Тогда их среднее арифметическое

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

при $n \to \infty$ сходится по вероятности к m, т. е. $\overline{X} \xrightarrow{p} m$.

Следствием теоремы Хинчина является теорема Бернулли.

Теорема Бернулли

Пусть A — случайный исход некоторого эксперимента, а $P(A) = p$ — вероятность этого исхода. Предположим, что эксперимент повторяется n раз в неизменных условиях (т. е. вероятность $P(A) = p$ не изменяется при повторении экспериментов). Тогда относительная частота появления события A

$$h_n(A) = \frac{n_A}{n}$$

при $n \to \infty$ сходится по вероятности к p: $h_n(A) \xrightarrow{p} p$.

Доказательство. Пусть X_i, $i = 1, 2, ..., n$ — случайная величина, принимающая два значения: 1, если в i-м эксперименте произошло событие A, и 0, если в i-м эксперименте событие A не произошло: $P[X_i = 1] = p$, $P[X_i = 0] = 1 - p = q$.

Вычислим математическое ожидание случайной величины X_i:

$$M[X_i] = 1 \cdot p + 0 \cdot q = p,$$

и математическое ожидание их среднего арифметического

$$M\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{M[X_1] + \ldots + M[X_n]}{n} = p.$$

Случайные величины X_i, $i = 1, 2, \ldots, n$ по условию взаимно независимы, а их среднее арифметическое есть относительная частота появления события A в серии n экспериментов

$$\frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{n_A}{n} = h_n(A).$$

Следовательно, по теореме Хинчина, относительная частота появления события A в n экспериментах при $n \to \infty$ сходится по вероятности к p — вероятности появления события A в одном эксперименте.

Закон больших чисел описывает условия, при которых в случайных массовых явлениях наблюдается известный из практики факт устойчивости таких характеристик, как среднее арифметическое результатов наблюдений и частоты появления события в большой серии экспериментов. Здесь следует вспомнить статистическое определение вероятности в П.1.1. Теорема Бернулли дает математическое обоснование экспериментальным результатам, в которых наблюдается устойчивость частот при увеличении числа экспериментов.

Устойчивость среднего арифметического можно объяснить тем, что случайные отклонения от среднего, неизбежные в каждом отдельном результате, в массе однородных результатов взаимно погашаются, нивелируются, выравниваются. Вследствие этого средний результат практически перестает быть случайным и может быть предсказан достаточно точно.

В теореме Хинчина предполагалось, что случайные величины взаимно независимы и имеют одно и тоже распределение. Закон больших чисел в форме теоремы Чебышева показывает, что утверждение закона больших чисел верно и в случае, когда случайные величины X_i имеют различные распределения.

Теорема Чебышева

Пусть X_1, X_2, ..., X_n, ... последовательность взаимно независимых случайных величин с конечными дисперсиями $D[X_1]$, $D[X_2]$, ..., $D[X_n]$, ... Если существует такая константа c, что $D[X_i] \le c$, $i = 1, 2, \ldots, n, \ldots$, то среднее арифметическое случайных величин X_i: $\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$ сходится при $n \to \infty$ по вероятности к среднему арифметическому их математических ожиданий

$$\frac{M[X_1] + \ldots + M[X_n]}{n} = \frac{m_1 + \ldots + m_n}{n}.$$

Доказательство. В доказательстве этой теоремы основную роль играет неравенство Чебышева (см. П.2.5): если случайная величина имеет конечную дисперсию, то при любом $\varepsilon > 0$:

$$P[|X - M[X]| \geq \varepsilon] \leq \frac{D[x]}{\varepsilon^2}.$$

Вычислим математическое ожидание и дисперсию среднего арифметического независимых случайных величин X_i, $i = 1, 2, \ldots, n$:

$$M[\overline{X}] = M\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{M[X_1] + \ldots + M[X_n]}{n} = \frac{m_1 + \ldots + m_n}{n},$$

$$D[\overline{X}] = \frac{1}{n^2}(D[X_1] + D[X_2] + \ldots + D[X_n] \leq \frac{nc}{n^2} = \frac{c}{n}.$$

Неравенство Чебышева для случайной величины $\overline{X}$ имеет вид

$$P[|\overline{X} - \frac{m_1 + \ldots + m_n}{n}| \geq \varepsilon] \leq \frac{D[\overline{X}]}{\varepsilon^2} \leq \frac{c}{n\varepsilon^2}.$$

Так как при $n \to \infty$ для любого фиксированного ε правая часть неравенства имеет пределом нуль, то это означает, что $\overline{X} \xrightarrow{p} \frac{m_1 + \ldots + m_n}{n}$.

Приведем еще одну формулировку закона больших чисел — *теорему Маркова.*

Теорема Маркова

Если последовательность случайных величин X_1, X_2, ..., X_n, ... такова, что при $n \to \infty$, $\frac{1}{n^2}\left(D\sum_{i=1}^{n} X_i\right) \to 0$, то среднее арифметическое случайных величин X_1, X_2, ..., X_n — $\overline{X}$ сходится при $n \to \infty$ к среднему арифметическому их математических ожиданий.

Доказательство этой теоремы также следует из неравенства Чебышева. Заметим, что в теореме Маркова случайные величины могут быть зависимыми и иметь любое распределение.

Центральная предельная теорема (Ц.П.Т.)

Пусть X_1, X_2, ..., X_n, ... независимые и одинаково распределенные случайные величины с математическим ожиданием m и дисперсией σ^2. При $n \to \infty$ функция распределения случайной величины

$$\frac{\overline{X} - m}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - nm}{\sigma\sqrt{n}}$$

равномерно по x сходится к функции распределения стандартного нормального закона $\Phi(x)$, где

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

Важнейшее значение Ц.П.Т. в приложениях теории вероятностей состоит в следующем. Если наблюдаемую случайную величину можно рассматривать как сумму n независимо действующих случайных факторов, каждый из которых вносит в сумму очень малый и приблизительно равный вклад, то распределение этой случайной величины при $n \to \infty$ следует нормальному закону.

Примером применения этой теоремы на практике является нормальное распределение ошибок измерений. Действительно, всякое измерение неизбежно связано с погрешностями. Реально наблюдаемые ошибки измерений являются суммой погрешностей, вызванных многочисленными факторами (ошибки измерительного прибора, влияние атмосферных явлений, индивидуальные особенности лица проводящего измерение и др.). Причем каждый из факторов обычно лишь незначительно влияет на результат. Другими примерами являются рассеивание снарядов при стрельбе по цели или отклонение контролируемого размера изделия от номинала в массовом производстве. В этих примерах на окончательный результат также влияет очень большое число случайных факторов, каждый из которых оказывает лишь незначительный эффект.

Выше была приведена простейшая формулировка Ц.П.Т.

Более общая формулировка Ц.П.Т. предполагает, что X_1, X_2, ..., X_n, ... независимые случайные величины с различными распределениями. При этом условие конечности дисперсии $D[X_i]$, $i = 1, 2, ..., n$ заменяется условием Линдеберга, гарантирующим, что все слагаемые вносят равномерно малый вклад в общую дисперсию [3].

Из Ц.П.Т. в качестве следствия легко получить теорему Муавра—Лапласа.

Теорема Муавра—Лапласа

Пусть n_A — частота появления события A в n независимых экспериментах, тогда при $n \to \infty$

$$P\left[a < \frac{n_A - np}{\sqrt{npq}} < b\right] \approx \Phi(b) - \Phi(a), \qquad (1)$$

где $p = P(A)$ — вероятность появления события A в каждом эксперименте, $q = 1 - p$.

Доказательство. Частоту n_A появления события A в n экспериментах можно представить как сумму n независимых одинаково распределенных случайных величин $X_1, X_2, ..., X_n$:

$$n_A = X_1 + X_2 + ... + X_n = \sum_{i=1}^{n} X_i,$$

где $X_i = \begin{cases} 1, & \text{если в } i\text{-м эксперименте произошло событие } A, \\ 0, & \text{если в } i\text{-м эксперименте событие } A \text{ не произошло.} \end{cases}$

Математическое ожидание и дисперсия случайной величины X_i соответственно равны:

$$M[X_i] = 1 \cdot p + 0 \cdot q = p,$$

$$D[X_i] = M[X_i^2] - p^2 = p - p^2 = p(1 - p) = pq, \; i = 1, 2, ..., n.$$

Вычислим математическое ожидание и дисперсию частоты n_A:

$$M[n_A] = M[X_1 + X_2 + ... + X_n] = M[X_1] + M[X_2] + ... + M[X_n] = np,$$

$$D[n_A] = D[X_1 + X_2 + ... + X_n] = D[X_1] + D[X_2] + ... + D[X_n] = npq.$$

В силу Ц.П.Т., функция распределения нормированной суммы $\dfrac{\sum_{i=1}^{n} X_i - np}{\sqrt{npq}}$ при $n \to \infty$ равномерно по x сходится к функции распределения $\Phi(x)$ стандартного нормального закона, откуда и следует (1).

Теорема Муавра—Лапласа позволяет количественно оценить разброс частоты появления события A в сериях независимых повторных испытаний относительно вероятности события A, а также ответить на вопрос: сколько надо провести испытаний, чтобы отклонение наблюдаемой частоты от вероятности события A лежало в заданных пределах.

Пример 1. Пусть p — неизвестная вероятность появления события A в некотором эксперименте, который может повторяться n раз в неизменных условиях. Приблизительное значение p (оценка p) равно значению наблюдаемой относительной частоты появления события A в n экспериментах, $h_n(A)$, причем, чем больше n, тем выше относительная точность этого результата. Предположим, что нужно найти такое значение $h_n(A)$, чтобы с вероятностью большей или равной 0,95 отклонение частоты $h_n(A)$ от истинной вероятности p не превышало бы величины равной 0,05:

$$|p - h_n(A)| \le 0{,}05.$$

Сколько раз надо повторить эксперимент, чтобы удовлетворить этим условиям?

Решение. Воспользуемся теоремой Муавра—Лапласа: при $n \to \infty$ число появлений событий A в n экспериментах, n_A, имеет распределение близкое к нормальному с математическим ожиданием np и дисперсией npq:

$$n_A \sim N(np, npq).$$

По условию задачи имеем, что вероятность события $[|\frac{n_A}{n} - p| \le 0{,}05]$ должна быть не менее, чем 0,95, т. е.

$$P[|\frac{n_A}{n} - p| \le 0{,}05] \ge 0{,}95 \text{ или } P[|n_A - np| \le 0{,}05n] \ge 0{,}95.$$

Используя нормальное приближение (1) и формулу (10) П.3.4, последнее неравенство можно записать так:

$$2\Phi\left(\frac{0{,}05n}{\sqrt{npq}}\right) - 1 \ge 0{,}95.$$

Решая это неравенство, получим

$$\Phi\left(\frac{0{,}05\sqrt{n}}{\sqrt{pq}}\right) \ge 0{,}975.$$

По таблице значений функции $\Phi(x)$ (Приложение 3) определяем, что аргумент функции $\Phi(x)$ должен удовлетворять следующему неравенству

$$\frac{0{,}05\sqrt{n}}{\sqrt{pq}} \ge 1{,}96,$$

или, так как $pq \le \frac{1}{4}$, получим, что

$$\sqrt{n} \ge \frac{1{,}96 \cdot 0{,}5}{0{,}05} = 19{,}6.$$

Отсюда следует, что нужно провести не менее чем 385 экспериментов: $n \ge 385$.

Приложение 1.1

Варианты заданий по регрессионному, корреляционному и кластерному анализу

Рассматриваются следующие показатели для 50 предприятий [8]:
Y_1 — производительность труда;
Y_2 — индекс снижения себестоимости продукции;
Y_3 — рентабельность;
X_4 — трудоемкость единицы продукции;
X_5 — удельный вес рабочих;
X_6 — удельный вес покупных изделий;
X_7 — коэффициент сменности оборудования;
X_8 — премии и вознаграждения на одного работника;
X_9 — удельный вес потерь от брака;
X_{10} — фондоотдача;
X_{11} — среднегодовая численность работников;
X_{12} — среднегодовая стоимость основных производственных фондов;
X_{13} — среднегодовой фонд заработной платы работников;
X_{14} — фондовооруженность труда;
X_{15} — непроизводственные расходы.

Таблица П1. **Варианты заданий 1—10**

№ варианта	Результативный признак, Y_j	Номер факторных признаков, X_i
1	1	6, 8, 11, 12, 15
2	1	8, 11, 12, 13, 15
3	1	8, 9, 13, 14, 15
4	3	8, 9, 10, 11, 15
5	3	8, 9, 10, 12, 15
6	2	4, 5, 6, 8, 9
7	2	4, 5, 6, 7, 9
8	2	4, 5, 6, 8, 9
9	2	4, 5, 8, 9, 15
10	2	4, 5, 7, 9, 15

Таблица П2. **Таблица исходных данных**

№ предприятия	Y_1	Y_2	Y_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	9,26	204,2	13,26	0,23	0,78	0,40	1,37	1,23	0,23	1,45
2	9,38	209,6	10,16	0,24	0,75	0,26	1,49	1,04	0,39	1,30
3	12,11	222,6	13,72	0,19	0,68	0,40	1,44	1,80	0,43	1,37
4	10,81	236,7	12,85	0,17	0,70	0,50	1,42	0,43	0,18	1,65
5	9,35	62,0	10,63	0,23	0,62	0,40	1,35	0,88	0,15	1,91
6	9,87	53,1	9,12	0,43	0,76	0,19	1,39	0,57	0,34	1,68
7	8,17	172,1	25,83	0,31	0,73	0,25	1,16	1,72	0,38	1,94
8	9,12	56,5	23,39	0,26	0,71	0,44	1,27	1,70	0,09	1,89
9	5,88	52,6	14,68	0,49	0,69	0,17	1,16	0,84	0,14	1,94
10	6,30	46,6	10,05	0,36	0,73	0,39	1,25	0,60	0,21	2,06
11	6,22	53,2	13,99	0,37	0,68	0,33	1,13	0,82	0,42	1,96
12	5,49	30,1	9,68	0,43	0,74	0,25	1,10	0,84	0,05	1,02
13	6,50	146,4	10,03	0,35	0,66	0,32	1,15	0,67	0,29	1,85
14	6,61	18,1	9,13	0,38	0,72	0,02	1,23	1,04	0,48	0,88
15	4,32	13,6	5,37	0,42	0,68	0,06	1,39	0,66	0,41	0,62
16	7,37	89,8	9,86	0,30	0,77	0,15	1,38	0,86	0,62	1,09
17	7,02	62,5	12,62	0,32	0,78	0,08	1,35	0,79	0,56	1,60
18	8,25	46,3	5,02	0,25	0,78	0,20	1,42	0,34	1,76	1,53
19	8,15	103,5	21,18	0,31	0,81	0,20	1,37	1,60	1,31	1,40
20	8,72	73,3	25,17	0,26	0,79	0,30	1,41	1,46	0,45	2,22
21	6,64	76,6	19,40	0,37	0,77	0,24	1,35	1,27	0,50	1,32
22	8,10	73,01	21,0	0,29	0,78	0,10	1,48	1,58	0,77	1,48
23	5,52	32,3	6,57	0,34	0,72	0,11	1,24	0,68	1,20	0,68
24	9,37	199,6	14,19	0,23	0,79	0,47	1,40	0,86	0,21	2,30
25	13,17	598,1	15,81	0,17	0,77	0,53	1,45	1,98	0,25	1,37
26	6,67	71,2	5,23	0,29	0,80	0,34	1,40	0,33	0,15	1,51
27	5,68	90,8	7,99	0,41	0,71	0,20	1,28	0,45	0,66	1,43
28	5,22	82,1	17,50	0,41	0,79	0,24	1,33	0,74	0,74	1,82
29	10,02	76,2	17,16	0,22	0,76	0,54	1,22	0,03	0,32	2,62
30	8,16	119,5	14,54	0,29	0,78	0,40	1,28	0,99	0,89	1,75

Продолжение табл. П2

№ предприятия	Y_1	Y_2	Y_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
31	3,78	21,9	6,24	0,51	0,62	0,20	1,47	0,24	0,23	1,54
32	6,48	48,4	12,08	0,36	0,75	0,64	1,27	0,57	0,32	2,25
33	10,44	173,5	9,49	0,23	0,71	0,42	1,51	1,22	0,54	1,07
34	7,65	74,1	9,28	0,26	0,74	0,27	1,46	0,68	0,75	1,44
35	8,77	68,6	11,42	0,27	0,65	0,37	1,27	1,0	0,16	1,40
36	7,00	60,8	10,31	0,29	0,66	0,38	1,43	0,81	0,24	1,31
37	11,06	355,6	8,65	0,01	0,84	0,35	1,50	1,27	0,59	1,12
38	9,02	264,8	10,94	0,02	0,74	0,42	1,35	1,14	0,56	1,16
39	13,28	526,6	9,87	0,18	0,75	0,32	1,41	1,89	0,63	0,88
40	9,27	118,6	6,14	0,25	0,75	0,33	1,47	0,67	1,10	1,07
41	6,70	37,1	12,93	0,31	0,79	0,29	1,35	0,96	0,39	1,24
42	6,69	57,7	9,78	0,38	0,72	0,30	1,40	0,67	0,73	1,49
43	9,42	51,6	13,22	0,24	0,70	0,56	1,20	0,98	0,28	2,03
44	7,24	64,7	17,29	0,31	0,66	0,42	1,15	1,16	0,10	1,84
45	5,39	48,3	7,11	0,42	0,69	0,26	1,09	0,54	0,68	1,22
46	5,61	15,0	22,49	0,51	0,71	0,16	1,26	1,23	0,87	1,72
47	5,59	87,5	12,14	0,31	0,73	0,45	1,36	0,78	0,49	1,75
48	6,57	108,4	15,25	0,37	0,65	0,31	1,15	1,16	0,16	1,46
49	6,54	267,3	31,34	0,16	0,82	0,08	1,87	4,44	0,85	1,60
50	4,23	34,2	11,56	0,18	0,80	0,68	1,17	1,06	0,13	1,47

Продолжение табл. П2

№ предприятия	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
1	26006	167,69	47750	6,40	17,72
2	23935	186,10	50391	7,80	18,39
3	22589	220,45	43149	9,76	26,46
4	21220	169,30	41089	7,90	22,37
5	7394	39,53	14257	5,35	28,13
6	11586	40,41	22661	9,90	17,55
7	26609	102,96	52509	4,50	21,92

Продолжение табл. П2

№ предприятия	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
8	7801	37,02	14903	4,88	19,52
9	11587	45,74	25587	3,46	23,99
10	9475	40,07	16821	3,60	21,76
11	10811	45,44	19459	3,56	25,68
12	6371	41,08	12973	5,65	18,13
13	26761	136,14	50907	4,28	25,74
14	4210	42,39	6920	8,85	21,21
15	3557	37,39	5736	8,52	22,97
16	14148	101,78	26705	7,19	16,38
17	9872	47,55	20068	4,82	13,21
18	5975	32,61	11487	5,46	14,48
19	16662	103,25	32029	6,20	13,38
20	9166	38,95	18946	4,25	13,69
21	15118	81,32	28025	5,38	16,66
22	11429	67,26	20968	5,88	15,06
23	6462	59,92	11049	9,27	20,09
24	24628	107,34	45893	4,36	15,98
25	49727	512,60	99400	10,31	18,27
26	11470	53,81	20719	4,69	14,42
27	19448	80,83	36813	4,16	22,76
28	18963	59,42	33956	3,13	15,41
29	9185	36,96	17016	4,02	19,35
30	17478	91,43	34873	5,23	16,83
31	6265	17,16	11237	2,74	30,53
32	8810	27,29	17306	3,10	17,98
33	17659	184,33	39250	10,44	22,09
34	10342	58,42	19074	5,65	18,29
35	8901	59,40	18452	6,67	26,05
36	8402	49,63	17500	5,91	26,20
37	32625	391,27	7888	11,99	17,26

Окончание табл. П2

№ предприятия	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
38	31160	258,62	58947	8,30	18,83
39	46461	75,66	94697	1,63	19,70
40	13833	123,68	29626	8,94	16,87
41	6391	37,21	11688	5,82	14,63
42	11115	53,37	21955	4,80	22,17
43	6555	32,87	12243	5,01	22,62
44	11085	45,63	20193	4,12	26,44
45	9484	48,41	20122	5,10	22,26
46	3967	13,58	7612	3,49	19,13
47	15283	63,99	27404	4,19	18,28
48	20874	104,55	39648	5,01	28,23
49	19418	222,11	43799	11,44	12,39
50	3351	25,76	6235	7,67	11,64

Приложение 1.2

Стоимость однокомнатных квартир в Москве

Данные из газеты «Из рук в руки» за период с декабря 1996 г. по сентябрь 1997 г.

Была выбрана Юго-Западная часть города, в которой высок спрос на жилые площади (всего 69 наблюдений).

Переменные:

N — номер по порядку;
distc — удаленность от центра, км;
distm — удаленность от метро, мин;
totsq — общая площадь квартиры, м^2;
kitsq — площадь кухни, м^2;
livsq — площадь комнаты, м^2;
floor — этаж: 0 — первый/последний, 1 — нет;
cat — категория дома: 1 — кирпичный, 0 — нет;
price — цена квартиры, тыс. USD.

N	Region	distc	distm	totsq	kitsq	livsq	floor	cat	price
1	Фрунзенская	4	10	34,00	7,50	19,00	1	1	54
2	Ленинский проспект	5,7	7	36,00	10,00	20,00	0	0	35
3	Ленинский проспект	5,7	12	45,00	13,00	20,00	1	1	59
4	Академическая	7,6	10	35,30	10,00	20,00	1	0	35
5	Университет	8,7	6	33,00	5,50	22,00	1	0	33
6	Новые Черемушки	10,3	3	33,00	8,50	18,00	1	1	57
7	Юго-Запад	13,3	10	37,00	10,00	19,00	1	0	43
8	Коньково	14,8	2	38,00	8,50	19,10	1	0	39
9	Фрунзенская	4	15	54,00	9,20	27,20	1	1	70
10	Университет	8,7	15	35,00	6,00	20,00	0	1	43
11	Проспект Вернадского	11,4	10	31,40	5,20	21,30	1	0	33

Продолжение табл. П.1.2

N	Region	distc	distm	totsq	kitsq	livsq	floor	cat	price
12	Ленинский проспект	5,7	7	32,00	6,00	21,00	1	0	37
13	Новые Черемушки	10,3	7	38,00	8,00	19,00	0	0	33
14	Университет	8,7	10	31,60	8,80	14,00	0	0	31
15	Юго-Запад	13,3	5	32,00	8,00	17,00	1	0	37
16	Юго-Запад	13,3	10	37,00	10,00	19,00	1	0	43
17	Ленинский проспект	5,7	5	32,00	8,00	17,00	1	1	38
18	Академическая	7,6	10	37,00	8,00	19,00	1	1	51
19	Академическая	7,6	15	32,20	6,50	17,00	0	1	30
20	Коньково	14,8	3	33,00	8,00	19,00	1	0	30
21	Коньково	14,8	5	37,50	9,60	19,80	1	0	36
22	Коньково	14,8	10	33,00	7,00	19,00	1	0	33
23	Университет	8,7	15	32,00	6,00	21,50	1	0	35
24	Проспект Вернадского	11,4	5	29,70	6,00	16,10	0	0	28
25	Проспект Вернадского	11,4	15	36,00	8,60	18,00	0	0	40
26	Юго-Запад	13,3	15	36,00	10,00	19,00	0	0	33
27	Ленинский проспект	5,7	2	31,60	6,00	21,60	1	1	35
28	Ленинский проспект	5,7	5	52,00	12,00	34,00	1	1	75
29	Коньково	14,8	3	36,00	10,00	19,00	1	0	40
30	Коньково	14,8	5	33,00	8,00	18,00	1	0	30
31	Университет	8,7	5	32,00	5,50	20,10	1	0	31
32	Академическая	7,6	15	35,00	9,80	20,00	1	0	37
33	Новые Черемушки	10,3	15	38,00	10,00	19,50	1	0	40

Продолжение табл. П.1.2

N	Region	distc	distm	totsq	kitsq	livsq	floor	cat	price
34	Коньково	14,8	1	39,00	8,50	19,00	1	0	40
35	Фрунзенская	4	5	34,00	8,00	19,00	1	1	58
36	Фрунзенская	4	10	38,00	6,50	18,00	0	1	48
37	Проспект Вернадского	11,4	3	35,00	10,00	20,00	1	0	40
38	Юго-Запад	13,3	7	36,00	9,00	19,50	1	0	42
39	Новые Черемушки	10,3	7	34,00	8,00	18,00	1	1	51
40	Коньково	14,8	5	38,00	8,50	19,00	1	0	43
41	Коньково	14,8	7	33,00	6,00	19,00	1	0	30
42	Коньково	14,8	10	32,00	8,00	17,00	1	0	40
43	Коньково	14,8	10	38,00	8,50	19,10	1	0	43
44	Академическая	7,6	5	43,00	8,50	25,00	0	1	53
45	Академическая	7,6	10	30,00	6,00	18,30	1	1	28
46	Коньково	14,8	7	34,80	7,80	17,80	0	0	29
47	Коньково	14,8	15	35,00	10,00	19,60	1	0	37
48	Коньково	14,8	3	32,80	6,50	18,50	1	0	30
49	Новые Черемушки	10,3	10	39,00	9,00	19,00	1	0	45
50	Университет	8,7	15	49,00	9,00	20,50	0	1	52
51	Фрунзенская	4	3	32,00	6,20	19,00	1	1	53
52	Проспект Вернадского	11,4	10	33,00	6,50	19,00	1	0	32
53	Проспект Вернадского	11,4	15	32,30	6,00	21,90	0	0	28
54	Юго-Запад	13,3	10	30,00	7,00	19,80	1	0	34
55	Юго-Запад	13,3	10	34,00	9,00	19,00	1	0	42

Окончание таблицы П.1.2

N	Region	distc	distm	totsq	kitsq	livsq	floor	cat	price
56	Юго-Запад	13,3	7	33,00	7,00	19,00	0	0	33
57	Академическая	7,6	10	30,00	6,00	18,30	1	1	28
58	Академическая	7,6	15	32,00	6,00	18,00	1	0	30
59	Коньково	14,8	5	33,10	7,50	18,00	1	0	32
60	Коньково	14,8	2	38,00	7,50	19,00	1	0	41
61	Коньково	14,8	7	38,00	8,60	19,00	1	0	43
62	Коньково	14,8	5	37,30	6,50	19,00	1	0	31
63	Ленинский проспект	5,7	8	31,40	5,60	21,00	1	0	33
64	Ленинский проспект	5,7	7	52,00	10,00	34,00	1	1	60
65	Новые Черемушки	10,3	15	30,00	6,00	17,00	1	1	37
66	Новые Черемушки	10,3	5	36,00	11,00	20,00	1	0	41
67	Проспект Вернадского	11,4	5	28,00	6,70	14,40	1	0	35
68	Проспект Вернадского	11,4	10	31,40	5,20	21,30	1	0	33
69	Юго-Запад	13,3	5	32,00	8,00	17,00	1	0	37

Приложение 2

Таблица критических точек критерия Дарбина—Уотсона

Критические точки d_1 и d_2 для уровня 5 % ($\alpha = 0{,}05$), k — число оцениваемых параметров регрессии, n — объем выборки.

n	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	0,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	0,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79

Окончание таблицы критических точек критерия Дарбина— отсона

n	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,62	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,75	1,78

Приложение 3

Значения функции распределения $\Phi(x)$ стандартного нормального закона $N(0, 1)$:

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

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,500000	1,00	0,841345	2,00	0,977250
0,05	0,519939	1,05	0,853141	2,05	0,979818
0,10	0,539828	1,10	0,864334	2,10	0,982136
0,15	0,559618	1,15	0,874928	2,15	0,984222
0,20	0,579260	1,20	0,884893	2,20	0,986097
0,25	0,589706	1,25	0,894350	2,25	0,987776
0,30	0,617911	1,30	0,903200	2,30	0,989276
0,35	0,636831	1,35	0,911492	2,35	0,990613
0,40	0,655422	1,40	0,919243	2,40	0,991802
0,45	0,673645	1,45	0,926471	2,45	0,992857
0,50	0,691463	1,50	0,933193	2,50	0,993790
0,55	0,708840	1,55	0,939429	2,55	0,994614
0,60	0,725747	1,60	0,945201	2,60	0,995339
0,65	0,742154	1,65	0,950528	2,65	0,995975
0,70	0,758036	1,70	0,955434	2,70	0,996533
0,75	0,773373	1,75	0,959941	2,75	0,997020
0,80	0,788145	1,80	0,964070	2,80	0,997445
0,85	0,802338	1,85	0,967843	2,85	0,997814
0,90	0,815940	1,90	0,971283	2,90	0,998134
0,95	0,828944	1,95	0,974412	2,95	0,998411
				3,00	0,998650

Квантили u_p стандартного нормального распределения $N(0, 1)$

p	0,90	0,95	0,975	0,99	0,995	0,999	0,9995
u_p	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Приложение 4

Словарь терминов пакета STATISTICA и статистических терминов

Accept	Принять
Action	Действие
Add Cases	Добавить наблюдения
Add Variables	Добавить переменные
Adjust	Корректировка
Advisor	Советник
Alert Always	Сообщать всегда
ANOVA, analysis of variance	Однофакторный дисперсионный анализ
Appearance	Представление
Apply	Применить
Area Under Curve	Площадь под кривой
Assigned Cases	Связанные наблюдения
Assignment	Назначение, задание, новые данные
Assign Rank 1 to	Установить значение ранга 1 для
Associate	Связать
Attempt	Попытка
Automatic update on Exit	Автоматически обновлять при выходе
Auxiliary	Дополнительно
Background	Фон
Bands	Полосы
Banner	Заголовок

Продолжение словаря терминов

Banners	Флажки
Bar	Линейка
Baseline Errors	Исходные ошибки
Basic	Основной
Batch	Пакетный
Baund rate	Скорость передачи (бит в секунду)
Beyond	За, вне, свыше
Bias	Смещение
Biases	Склонность, предубеждение
Bivariate Distribution	Двумерное распределение
Blank	Пустой
Bookmark	Закладка
Boot	Запуск системы
Border	Рамка
Branch	Ветвь
Breakdown	Разбиение
Browse	Просмотр
Brushing	Окраска
Buttons	Кнопки
Canonical Analyses	Канонический анализ
Canonical Reduction	Каноническое преобразование
Case Name	Имена случаев (наблюдений)
Case Selection Conditions	Условия выбора случаев (наблюдений)
Cauchy distribution	Распределение Коши
Centering (of the data)	Центрирование (данных)

Продолжение словаря терминов

Character	Литера (опция, в которой выбираются гарнитура, начертание и размер (кегль) шрифта)
Chat	Электронный телефон
Check	Контроль
Chi- Squared Distribution	Распределение хи-квадрат
Clicking	Нажатие на кнопку мыши
Clipboard	Буфер промежуточного хранения
Clip Book-Viewer	Программа просматривает содержимое буфера Clipboard, сохраняет его или удаляет
Coefficient of multiple determination	Множественный коэффициент детерминации: квадрат коэффициента множественной корреляции
Collapse	Свернуть
Collapse Branch	Свернуть ветвь
Comparison	Сравнение
Compatibility	Совместимость
Complexity	Сложность
Condition Expectation	Условное матем. ожидание
Confidence interval	Доверительный интервал
Confidence Limit	Доверительный предел
Connect	Присоединять
Content	Содержание
Contiguous	Смежный
Continuity Correction	Поправка на непрерывность
Continuous Distribution	Непрерывное распределение
Control Box	Кнопка управления
Convert	Преобразовывать

Продолжение словаря терминов

Correlation Analysis	Корреляционный анализ
Correlation Coefficient	Коэффициент корреляции
Correlation Matrix	Матрица (коэффициентов) корреляции, корреляционная матрица
Covariance	Ковариация
Create Data Set	Создать набор данных
Critical Value	Критическое значение
Cumulative Probability	Интегральная (накопленная) вероятность
Cumulative Probability Distribution	Кумулятивное (накопленное) распределение вероятностей
Currency	Денежный формат (данных)
Current	Текущий
Current Spec...	Текущая спецификация
Custom Colors	Пользовательские цвета
Custom Graphs	Пользовательский график
Cut	Урезание
Data Management	Управление данными
Data Matrix	Матрица данных
Data Set Datasheet	Таблица данных
Data Set Editor	Редактор данных
Data Set Shuffle	Переметать данные
Data Values	Значение данных, данные
Decimals	Десятичные знаки
Default	По умолчанию
Define	Определять
Definition	Определение

Продолжение словаря терминов

Degrees of freedom (d.f.)	Степени свободы; число степеней свободы
Delete Cases	Удалить случаи (наблюдения)
Delimiter	Разделитель
Density Function	Функция плотности распределения вероятностей
Dependent variable	Зависимая переменная; отклик
Discrepancy	Расхождение (разность)
Discrete distribution	Дискретное распределение
Descriptive Statistics	Описательные статистики
Detail Shown	Степень подробности
Detrended Data	Данные с исключенным трендом
Deviation	Отклонение
Dial	Способ
Deletion	Вычеркивание; стирание; удаление; исключение; ликвидация; уничтожение
Destination Variables	Создаваемые переменные
Direct	Прямой
Discard	Отвергнуть
Distribution of Error	Распределение ошибок
Division	Деление
Division of Cases	Разбиение наблюдений
Double Precision Arithmetic	Вычисления с удвоенной точностью
Download	Загрузить
Draft	Чертеж
Dragging	Протягивание (мыши)
Drop-down	«Выпадающий»

Продолжение словаря терминов

Durbin—Watson test	Критерии Дарбина—Уотсона
Edit Case Names	Редактировать имена наблюдений
Eigenvalues	Собственные значения
Embedding	Встраивание (объектов)
Enlarge Set	Увеличить набор
Enough	Достаточно
Ensure	Гарантировать
Envelopes	Конверты
Error	Ошибка
Error Function	Функция ошибки
Error Mean	Среднее ошибки
Estimate	Оценка, оценивать; приблизительно подсчитывать
Estimation	Оценивание (подсчет, вычисление)
Except	Исключать
Exclude if...	Удалить, если...
Expect	Ждать; предполагаемый
Expectation	Математическое ожидание
Expected value	Математическое ожидание, среднее значение
Exponential distribution	Экспоненциальное распределение
Extreme Value	Экстремальное значение
Facile	Легкий
Feature Selection	Отбор признаков
Field	Поле
Fill Block	Заполнить блок

Продолжение словаря терминов

Fill Random Values	Заполнить значения переменных случайными величинами — числами, имеющими равномерное распределение от 0 до 1
Fit the Model	Подбор модели, подгонка модели
Flash	Мерить
Flow Control	Протокол
Fonts	Шрифт
Forward Selection Procedure	Метод включения (в регрессионном анализе)
Fractional	Дробный (ранг от 0 до 1)
Frequency	Частота
Frequency Function	Функция частот
General	Общие
Glossary	Специальный толковый словарь
Graduation	Сглаживание, нанесение кривой по точкам
Grate	Решетка
Handshake	Подтверждение
Header	Заголовок
Hidden	Скрытый
Hidden Units	Скрытые элементы
Highlight Counts	Выделить числа
Hypothesis	Гипотеза
Hypothesis Testing	Проверка гипотезы
Icon	Пиктограмма
Imaginary	Мнимая часть
Inactive	Неактивный
Include if	Включить, если

Продолжение словаря терминов

Incorrelated	Некоррелированный
Independent Samples	Независимые выборки
Independent Variable	Независимая переменная, фактор
Index	Оглавление
Input Data Matrix	Матрица исходных данных
Inputs Datasheet	Таблица входных значений
Input Variable	Входная переменная
Insertion	Выделение
Insert Object	Вставка объекта
Insufficient	Недостаточный, неподходящий
Integer	Целый
Interaction	Взаимодействие
Intercept	Свободный член (в уравнении регрессии)
Interrupted	Прерванный
Inverse of Matrix	Обращение матрицы
Involve	Включить, вовлекать
Item	Элемент данных
Iterations	Число итераций
Jittering	Разгонка (точек)
Joining	Соединение
Kurtosis	Эксцесс
Kurtosis of Frequency	Эксцесс кривой плотности распределения
Lag-1 Serial Correlation	Сериальная корреляция с единичным сдвигом
Latent Variable	Латентная (скрытая) переменная
Layout	Расположение, разметка

Продолжение словаря терминов

Least	Наименьший
Least Squares Method	Метод наименьших квадратов
Least Squares Method Equation	МНК уравнение
Least Squares Method Estimate	МНК оценка
Level of Factor	Уровень фактора
Linear Regression (Model)	Линейная регрессия (модель)
Relationship	Линейная зависимость
Trend	Тренд (временного ряда)
Links	Связи
Lock	Защитить, блокировать
Logistic	Логистический
Logistic Regression	Логистическая регрессия
Log-normal Variable	(Случайная) величина, распределенная по логарифмически нормальному закону
Loss Coefficient	Коэффициент потерь
Loss Matrix	Матрица потерь
Lower-Tailer	Односторонний критерий для нижнего «хвоста» распределения
Manuel	Руководство, инструкция
Margin	Край, граница, поле (печатной страницы)
Match Case	Учет регистра
Max/SD	Максимальное/(стандартное отклонение)
Maximum Likelihood	Максимальное правдоподобие
Mean Square	Средний квадрат
Mean Square Error	Средний квадрат ошибки

Продолжение словаря терминов

Mean Square About Regression	Средний квадрат отклонений относительно регрессии
Mean Square About Regression Due to Lack of Fit	Средний квадрат, обусловленный неадекватностью
Mean Square About Regression Due to Regression	Средний квадрат обусловленный регрессией
Mean Square About Regression Due to Residual Variation	Остаточный средний квадрат (средний квадрат, обусловленный остаточной вариацией)
Mean Square About Regression For Pure Error	Средний квадрат, характеризующий «чистую» ошибку
Mean/SD	Среднее/(стандартное отклонение)
Means	Среднее
Measure	Мера
Median	Медиана
Medium	Средняя (длительность поиска)
Merge	Объединить
Message	Сообщение, поручение
Method for Discriminating	Метод дискриминации (моделей)
Method for Discriminating of Least Squares	Метод наименьших квадратов (МНК)
Min Proportion	Минимальная доля
Mini max	Минимаксное
Missing Observations	Пропущенные наблюдения
Model Validation Technique	Метод обоснования модели
Modes	Режимы
Missing Data	Пропущенные значения
Momentum	Инерция
Mouse Pointer	Курсор мыши

Продолжение словаря терминов

Multiple Regression Calculation Correlation Coefficient	Множественный коэффициент корреляции
Multiple Regression	Множественная регрессия
Multiplicative Model	Мультипликативная модель
Multivariate	Многомерный
Move Cases	Перемещение случаев (наблюдений)
N-dimensional Multivariate Normal Distribution	*N*-мерное нормальное распределение
Negative Sereal Correlation Between Successive Residuals	Отрицательная сериальная корреляция между последовательными (соседними) остатками
Neighborhood	Окрестность
Newton—Raphson Technique	Метод Ньютона—Рафсона
Nonlinear	Нелинейный
Nonlinear Estimation	Нелинейное оценивание
Nonlinear Growth Model	Нелинейная модель роста
Nonsigular Matrix	Невырожденная матрица
Normal Deviate	Нормальное отклонение
Normal Deviate Distribution Random Variable	Нормально распределенная случайная величина
Normal Deviate Equations	Нормальные уравнения (МНК)
Normal Deviate Plot of Residuals	График остатков
Normal Distribution	Нормальное распределение
Normalization	Нормировка, стандартизация (данных)
Observations	Наблюдения
One- sided Test	Односторонний критерий
One-Way	Односторонний; однонаправленный

Продолжение словаря терминов

One-way Classification	Односторонняя классификация, классификация по одному признаку
Optimum Threshold	Оптимальный порог
Order of the Model	Порядок модели
Original Data	Исходные данные
Orthogonal Column	Ортогональные столбцы (матрицы)
Outlier	Выброс; резко выделяющееся значение
Output	Выходные данные; результат вычислений
Output Variable	Выходная переменная
Outputs Datasheet	Таблица выходных значений
Outputs Shown	Показывать при выводе
Overview	Общее представление (о каком-либо предмете); обзор
Packager	Упаковщик (объектов)
Padding	Добавление нулей (например, в ряд)
Page Layout	Просмотр пакета
Partial Correlation	Частная корреляция
Paste Special	Специальная вставка
Percentage	Проценты (представление данных в процентах); относительная (ошибка)
Percentage Point of the Distribution	Процентная точка распределения
Performance	Качество
Plot	График; кривая; диаграмма
Power	Степень
Precision	Точность
Predict	Прогнозировать, предсказывать

Продолжение словаря терминов

Predictability	Предсказуемость
Predicted (mean) Value	Предсказанное (среднее) значение
Predictive Discrepancy Sum of Squares	Сумма квадратов предсказанных расхождений
Predictive Equation (model)	Предсказывающее уравнение (модель)
Principal Components Analysis	Анализ главных компонент
Principal Component Regression	Регрессия на главных компонентах
Prior probabilities	Априорные вероятности
Probability Calculator	Вероятностный калькулятор
Probability Level	Уровень вероятности
Prompt	Подсказывать
Properties	Свойство, собственность, характеристики
Prune	Удалить
Pure Error	«Чистая ошибка» (ошибка опыта)
Pure Error Mean Square	Средний квадрат, связанный с «чистой» ошибкой
Pure Error Sum of Squares	Сумма квадратов, связанная с «чистой» ошибкой (обусловленная «чистой» ошибкой)
Raise	Увеличение
Random	Случайный
Random Arrangement of Signs	Случайное расположение знаков
Random Deviation	Случайное отклонение
Random Search	Случайный поиск
Random Variation	Случайный разброс
Range selection	Выделение диапазона ячеек
Ranks For Ties	Ранги для совпадающих значений

Продолжение словаря терминов

Rank Variables	Присвоение рангов значениям переменной
Rate	Цена, расценка
Ratio	Отношение
Raw	Исходный
Real number fields	Поля для вещественных чисел
Recalculate Variables	Пересчитать значения переменных
Receive	Получать
Recode Variables	Перекодировать переменные
Redundance	Чрезмерность, избыточность
Redial	Повторить
Refresh	Обновлять
Regression	Регрессия, зависимость
Regression Curve	Регрессионная кривая
Regression Equation	Уравнение регрессии
Regression Estimate	Регрессионная оценка
Regression Mean Squares	Средний квадрат, обусловленный регрессией
Regular	Регулярный (ранг от 0 до 1)
Reject	Отвергнуть
Rayleigh distribution	Релеевское распределение
Remove	Удалить
Repeatability	Воспроизводимость
Replace existing	Заменить существующий
Representation	Представление
Reset	Восстановить
Residual	Остаток

Продолжение словаря терминов

Residual Mean Squares	Остаточный средний квадрат
Residual Sum of Squares	Остаточная сумма квадратов
Resolution	Разрешение — количество точек на дюйм
Response	Отклик
Restore	Восстановить в прежнем размере
Resume	Возобновить, продолжить
Retrieve Defaults	Применить установки по умолчанию
Ribbon	Линейка форматирования
Ridge Regression	Гребневая регрессия, ридж-регрессия
Rounding Error	Ошибка округления
Row vector	Вектор-строка
Ruler	Координатная линейка
Run	Запустить
Run All Cases	Прогнать все наблюдения
Running	Бегущий
Runs Test	Критерий знаков
Sample	Выборка
Sample Coefficient	Выборочный коэффициент, оценка коэффициента
Sample Estimate	Выборочная оценка
Sample Size	Объем (размер) выборки
Save Defaults	Сохранить по умолчанию
Scalable	Масштабируемый
Scaled	Нормированный
Scatter Diagramm (SD)	Диаграмма рассеяния

Продолжение словаря терминов

Scientific	Научная нотация (представление чисел в научной нотации, например, 5.0314 E-02)
Scroll Bars	Линейка просмотра
S.D. (Standard Deviation) Ratio	Отношение стандартных отклонений
Selecting	Выбор
Send	Передать
Set	Множество; совокупность; семейство; ряд; последовательность
Sequential	Последовательное (приписывание рангов)
Set Cake Types	Задать типы наблюдений
Screen Catcher	Команда захвата экрана (Alt + F3)
Serial Correlation of Residuals	Сериальная корреляция остатков
Settings	Установки
Setup	Установка
Shared	Разделяемая
Shift (Lag) Variables	Сдвиг переменной
Shuffle Cases	Перемешать наблюдения
Significance Level	Уровень значимости
Significance of Regression	Значимость регрессии
Significance Test	Критерий значимости
Single Case	Одно наблюдение
Skewness of Distribution	Асимметрия распределения
Skip	Пропустить
Slope	Угловой коэффициент (наклон) (регрессии)
Smoothing Constant	Константа сглаживания

Продолжение словаря терминов

Sort Ascending	Сортировать но возрастанию
Sort Descending	Сортировать по убыванию
Source	Подача (бумаги), источник
Source Variables	Исходные переменные
Split	Разделение
Spread	Распахнуть; разброс, вариация
Square of Multiple Correlation Coefficient	Квадрат множественного коэффициента корреляции (множественный коэффициент детерминации)
Stagewise	Ступенчатый
Standard Deviation (SD)	Стандартное отклонение (среднее квадратическое отклонение)
Standardize Columns	Команда стандартизации столбцов
Startup Panel	Стартовая панель модуля
Statistically Valid	Статистически обоснованный
Stats Graphs	Статистические графики
Status Bar	Строка состояния
Stepwise	Шаговый
Stepwise Regression Procedure	Шаговый регрессионный метод
Stopping Conditions	Условия остановки
Subset	Подмножество
Sum of Squares (SS)	Сумма квадратов
Swap file	Файл подкачки
Sweep	Размах, кругозор
Switch to	Переход к другой программе
T-distribution	*T*-распределение
T-test	*t*-критерий

Продолжение словаря терминов

Target Error	Целевая ошибка
Template	Шаблон
Test	Критерии, тест, проверка
Test of Hypothesis	Проверка гипотезы
Test of Significance	Проверка значимости
Test Statistic for Ho	Статистика для проверки гипотезы Но
Text Transfer	Режим подачи текста
Text Value Labels	Метки текстовых значений
Tile	Элемент мозаичного изображения
Time Sequence	Временная последовательность
Time Series	Временной ряд
Title Bar	Линейка заголовка
Tolerance	Допустимое отклонение
Toolbar	Панель (инструментов)
Total	Всего; общий
Transformation	Преобразование
Transformation on the Observations	Преобразование наблюдений
Transpose Block	Команда для транспонирования выделенного блока (контекстное меню при нажатии правой кнопки мыши)
Transpose Data File	Команда для транспонирования файла данных (переменные становятся случаями, а случаи — переменными)
Transpose of Matrix	Транспонирование матрицы
Trial	Испытание, проба
True Model	«Истинная» модель
Truncate	Урезать

Продолжение словаря терминов

Turn-key	Под ключ
Two-State Conversion	Преобразование в два значения
Two-tailed (-side) Test	Двусторонний критерий
Two-way Table	Таблица сопряженности, таблица с двумя входами
Unadjusted	Нескорректированный, без поправок
Unexplained Variation	Необъясненная вариация
Undo	Отмена
Uniform Distribution	Равномерное распределение
Unit Number	Номер элемента
Unknown	Неизвестно
Unknown Parameters	Неизвестные параметры
Unlisted	Неизвестный
Unlock	Разблокировать
Untitled	Неопределенный, неизвестный
Update	Актуализация (выбор режима)
Updated	Модернизированный, усовершенствованный
Upper-tailed Test	Односторонний критерий для верхнего «хвоста» распределения
Valid	Действительный
Validation	Обоснованность
Validation Technique	Метод перепроверки (проверки) состоятельности
Value Label	Значение меток
Variable (dependent)	Отклик, зависимая переменная
Variable (independent)	Фактор, независимая переменная
Variance about the Regression	Дисперсия относительно регрессии

Окончание словаря терминов

Variance about Covariance Matrix	Матрица дисперсий-ковариаций
Variation	Вариация, разброс
Vector of Error	Вектор ошибок (остатков)
Vector of Observation	Вектор наблюдений
Vector of Parameters to be Estimated	Вектор оцениваемых параметров
Verbose	Подробно
Variable Definition	Определение переменной
Verify	Проверка
View	Вид
Weibull distribution	Распределение Вейбулла
Weighted Least Squares	Взвешенный метод наименьших квадратов
Win Frequencies Datasheet	Таблица частот выигрышей
Workbook	Рабочая тетрадь, рабочий журнал
Wrap	Верстка, оболочка

Литература

1. Сборник задач по математике для ВТУЗов. Теория вероятностей и математическая статистика. Под. ред. А. В. Ефимова. М.: Наука. 2-е изд., 1990; 3-е изд., 2003.

2. *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989.

3. *Гнеденко Б. В.* Курс теории вероятностей. М.: УРСС, 2001.

4. *Себер Дж.* Линейный регрессионный анализ. М.: Мир, 1980.

5. *Вучков И., Бояджиева Л., Солаков Е.* Прикладной линейный регрессионный анализ. М.: Финансы и статистика, 1987.

6. *Вентцель Е. С.* Теория вероятностей. М.: Наука, 1969.

7. *Горицкий Ю. А., Перцов Е. Е.* Практикум по статистике с пакетами Statgraphics, Statistica, SPSS. М.: МЭИ, 1997.

8. *Дубров А. М., Мхитарян В. С., Трошин Л. И.* Многомерные статистические методы. М.: Финансы и статистика, 2000.

9. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983.

10. *Тюрин Ю. Н., Макаров А. А.* Анализ данных на компьютере. М.: Инфра-М, 2003.

11. *Рунион Р.* Справочник по непараметрической статистике. М.: Финансы и статистика, 1982.

12. *Аптон Г.* Анализ таблиц сопряженности. М.: Финансы и статистика, 1982.

13. *Кендалл М., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973.

14. *Закс Л.* Статистическое оценивание. М.: Статистика, 1976.

15. *Боровиков В. П., Боровиков И. П.* Statistica. Статистический анализ и обработка данных в среде Windows. М.: Филинъ, 1997.

16. *Боровиков В.* Statistica. Искусство анализа данных на компьютере. М.: Филинъ, 1997.

17. *Боровиков В. П., Ивченко Г. И.* Прогнозирование в системе Statistica. М.: Финансы и статистика, 2000.

18. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. Книги 1 и 2. М.: Финансы и статистика, 1987.

19. Справочник по прикладной статистике. Под. ред. Э. Ллойда, У. Ледермана. Т. 1 и 2. М.: Финансы и статистика, 1989.

20. *Ликеш И., Ляга Й.* Основные таблицы математической статистики. М.: Финансы и статистика, 1985.

21. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Исследование зависимостей. М.: Финансы и статистика, 1985.

22. *Браунли К. А.* Статистическая теория и методология в науке и технике. М.: Наука, 1977.

23. Электронный учебник по статистике. Statsoft. Inc. 1999. web: http://www/statsoft.ru/

24. *Химмельблау Д.* Анализ процессов статистическими методами. М.: Мир, 1973.

25. *Кендалл М.* Временные ряды. М.: Финансы и статистика, 1981.

26. *Кендалл М., Стьюарт А.* Многомерный статистический анализ и временные ряды. М.: Наука, 1976.

27. *Мандель И. Д.* Кластерный анализ. М.: Финансы и статистика, 1988.

28. *Таха Х.* Введение в исследование операций. Т. 1 и 2. М.: Мир, 1985.

29. *Курицкий Б.* Поиск оптимальных решений средствами Excel 7.0. Спб.: BHV-Санкт-Петербург, 1997.

30. *Афифи А., Эйзен С.* Статистический анализ. М.: Мир, 1988.

31. *Химмельблау Д.* Прикладное нелинейное программирование. М.: Мир, 1975.

32. *Хальд А.* Математическая статистика с техническими приложениями. М.: ИЛ., 1956.

33. *Джонстон Дж.* Эконометрические методы. М.: Мир, 1977.

34. *Бокс Дж, Дженкинс Г.* Анализ временных рядов. Прогноз и управление. Выпуск 1 и 2. М.: Мир, 1974.

35. *Андерсон Т.* Статистический анализ временных рядов. М.: Мир, 1976.

36. *Крамер Г., Линдбеттер М.* Стационарные случайные процессы. М.: Мир, 1969.

37. *Дюк В.* Обработка данных на ПК в примерах. СПб.: Питер, 1997.

Содержание

Вуколов Эдуард Александрович

Основы статистического анализа

Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL

Учебное пособие

Редактор *А. А. Макаров*
Корректор *В. Г. Овсянникова*
Компьютерная верстка *И. В. Кондратьевой*
Оформление серии *Лары Зарецкой*

Сдано в набор 18.09.2007. Подписано в печать 17.02.2008. Формат 70×100/16
Гарнитура «Таймс». Усл. печ. л. 37,41. Уч.-изд. л. 38,2
Печать офсетная. Бумага офсетная. Тираж 2000 экз
Заказ № 1746

Издательство «ФОРУМ»
101000, Москва — Центр, Колпачный пер., д. 9а
Тел./факс: (495) 625-32-07, 625-52-43
E-mail: mail@forum-books.ru

По вопросам приобретения книг обращайтесь:

Отдел продаж «ИНФРА-М»
127282, Москва, ул. Полярная, д. 31в
Тел.: (495) 363-42-60
Факс: (495) 363-92-12
E-mail: books@infra-m.ru

Центр комплектования библиотек
119019, Москва, ул. Моховая, д. 16
(Российская государственная библиотека, кор. К)
Тел.: (495) 202-93-15

Магазин «Библиосфера» (розничная продажа)
109147, Москва, ул. Марксистская, д. 9
Тел.: (495) 670-52-18, (495) 670-52-19

Отпечатано с готовых файлов заказчика в ОАО «ИПК «Ульяновский Дом печати». 432980, г. Ульяновск, ул. Гончарова, 14