

Министерство образования Республики Беларусь  
Учреждение образования  
«Белорусский государственный университет  
информатики и радиоэлектроники»

Кафедра экономической информатики

Т.М.Музычина, С.А.Поттосина

**ИСПОЛЬЗОВАНИЕ ПАКЕТА ПРИКЛАДНЫХ ПРОГРАММ  
«СТАТИСТИКА»**

**ЛАБОРАТОРНЫЙ ПРАКТИКУМ**  
по курсу «Эконометрика»  
для студентов экономических специальностей БГУИР  
всех форм обучения

Минск 2004

УДК 330.115(075.8)  
ББК 65.051 я 73  
М 89

Рецензент:

зав.кафедрой «Информационные технологии автоматизированных систем» БГУИР, доктор техн. наук, проф. В.С.Муха

**Т. М. Музыкаина**

М 89      Использование пакета прикладных программ «Статистика»: Лаб. практикум по курсу «Эконометрика» для студ. эконом. спец. БГУИР всех форм обуч./Т.М. Музыкаина, С.А. Поттосина. –Мн.: БГУИР, 2004. — 71 с.: ил.  
ISBN 985-444-611-5.

В лабораторном практикуме представлены четыре лабораторные работы, проводимые с использованием пакета *STATISTICA* в среде *Windows*. Даны краткие теоретические сведения по анализу данных, построению регрессионных зависимостей, прогнозированию временных рядов, анализу качества производственных процессов, а также варианты заданий. Примеры выполнения заданий способствуют наглядному представлению и более глубокому усвоению технологии работы с пакетом *STATISTICA* для специалистов в области экономической информатики.

УДК 330.115(075.8)  
ББК 65.051 я 73

ISBN 985-444-611-5

© Музыкаина Т.М., Поттосина С.А., 2004  
© БГУИР, 2004

## СОДЕРЖАНИЕ

Лабораторная работа №1. Анализ, обработка и представление данных в рамках пакета STATISTICA.....	4
1.1. Основные статистические модули пакета STATISTICA.....	4
1.2. Промышленная статистика.....	8
1.3. Особенности работы с пакетом STATISTICA.....	8
1.4. Основные элементы диалога в модуле «Основные статистики».....	11
1.5. Варианты заданий.....	20
Лабораторная работа №2. Регрессионный анализ эконометрической информации.....	22
2.1. ОБЩИЕ СВЕДЕНИЯ О МОДУЛЕ.....	22
2.2. Примеры реализации диалога в модуле «Множественная регрессия».....	23
2.3. Варианты заданий.....	31
Лабораторная работа № 3. Анализ временных рядов и прогнозирование.....	37
3.1. ЭЛЕМЕНТЫ ДИАЛОГА В МОДУЛЕ TIME SERIES ANALYSIS/FORECASTING.....	37
(Анализ временных рядов/Прогнозирование).....	37
3.2. Диалог ARIMA.....	39
3.3. Экспоненциальное сглаживание и прогнозирование.....	48
3.4. Варианты заданий.....	50
Лабораторная работа № 4. Статистический контроль производственного процесса.....	52
4.1. ПОСТАНОВКА ЗАДАЧИ.....	52
4.2. Диалог модуля «Контроль качества».....	52
4.3. Диалог модуля «Анализ процесса».....	55
4.4. Варианты заданий.....	58
Литература.....	60
Приложения.....	60

# Лабораторная работа №1

## АНАЛИЗ, ОБРАБОТКА И ПРЕДСТАВЛЕНИЕ ДАННЫХ В РАМКАХ ПАКЕТА *STATISTICA*

*Цель работы* - получить навыки работы с пакетом *STATISTICA*.

### 1.1. Основные статистические модули пакета *STATISTICA*

Статистические процедуры системы *STATISTICA* сгруппированы в нескольких специализированных статистических модулях. В каждом модуле можно выполнить определенный способ обработки данных, не обращаясь к процедурам из других модулей.

Ниже приводится краткое описание статистических процедур, реализованных в конкретных статистических модулях.

#### **Модуль «Основные статистики и таблицы»**

Если вы хотите провести предварительную обработку данных, осуществить разведочный анализ данных, определить зависимости между ними, разбить их различными способами на группы, просмотреть эти группы визуально и определить взаимосвязи между данными, запустите модуль *Basic Statistics/Tables* (*Основные статистики/Таблицы*). Обычно с этого модуля начинается работа в системе. Этот статистический модуль включает в себя приведенные ниже группы статистических процедур.

**Описательные статистики, группировки, разведочный анализ.** *STATISTICA* предлагает широкий выбор методов разведочного статистического анализа. Система может вычислить практически все описательные статистики, включая медиану, моду, квартили, определенные пользователем процентиля, средние и стандартные отклонения, доверительные интервалы для среднего, коэффициенты асимметрии, эксцесса (с их стандартными ошибками), гармоническое и геометрическое среднее, а также многие другие описательные статистики. Возможен выбор критериев для тестирования нормальности распределения (критерий Колмогорова–Смирнова, Лилиефорса, Шапиро–Уилкса). Широкий выбор графиков помогает проведению разведочного анализа данных.

**Корреляции.** Этот подраздел включает большое количество средств, позволяющих исследовать зависимости между переменными. Возможно вычисление практически всех общих мер зависимости, включая коэффициент корреляции Пирсона, коэффициент ранговой корреляции Спирмена, коэффициент сопряженности признаков и многие другие. Корреляционные матрицы могут быть вычислены и для данных с пропусками, используя специальные методы обработки

пропущенных значений.

***t*-критерии (и другие критерии для групповых различий).** В этом подразделе представлены *t*-критерии для зависимых и независимых выборок, а также статистика Хоттелинга. Данные критерии могут быть вычислены и в модуле *ANOVA/MANOVA*.

**Таблицы частот и таблицы кросс-табуляций.** В данном подразделе содержится обширный набор процедур, обеспечивающих табулирование непрерывных, категориальных, дихотомических переменных, а также переменных, полученных в результате многовариативных опросов. Вычисляются как кумулятивные, так и относительные частоты. Доступны тесты для кросс-табулированных частот. Вычисляются статистики Пирсона, максимального правдоподобия, Йетс-коррекция, хи-квадрат, статистики Фишера и многие другие. Каскады сложных графиков для многократно классифицированных данных могут быть просмотрены интерактивно.

### **Модуль «Множественная регрессия»**

Основное назначение данного модуля — построение зависимостей между многомерными переменными, подбор простой линейной модели и оценка ее адекватности.

Модуль «Множественная регрессия» включает в себя исчерпывающий набор средств множественной линейной и фиксированной нелинейной (в частности, полиномиальной, экспоненциальной, логарифмической и др.) регрессии, включая пошаговые, иерархические и другие методы, а также ридж-регрессию.

### **Модуль «Дисперсионный анализ» (ANOVA/MANOVA–модуль)**

Этот модуль позволяет дать оценку степени воздействия известных факторов на измеряемые данные. *ANOVA/MANOVA*–модуль представляет собой набор процедур общего одномерного и многомерного дисперсионного и ковариационного анализа.

В модуле доступны решения многих задач в наиболее прямых "функциональных" терминах. Даже пользователи, имеющие малый опыт работы с *ANOVA*, могут анализировать достаточно сложные проекты с помощью системы *STATISTICA*.

### **Модуль «Дискриминантный анализ»**

Этот модуль предназначен для решения задач, в которых по результатам измерений необходимо отнести объект к одному из нескольких классов, например, в медицине при обследовании больных, или в геологии при оценке перспективности месторождений, или в банковской деятельности .

### **Модуль «Непараметрическая статистика и подгонка распределений»**

Данный модуль предназначен для проверки различных гипотез о характере распределения имеющихся данных. Модуль содержит обширный набор непараметрических критериев согласия, в частности критерий Колмогорова–

Смирнова, ранговые критерии Манна–Уитни, Вальда–Вольфовица, Вилкоксона и многие другие.

### **Модуль «Факторный анализ»**

Этот модуль помогает выделить основные общие факторы качественного характера, влияющие на наблюдаемые характеристики сложного объекта и связи между ними, например основные социально-экономические факторы или факторы, влияющие на результаты голосования. Модуль содержит широкий набор методов и опций, снабжающих пользователя исчерпывающими средствами факторного анализа.

### **Модуль «Многомерное шкалирование»**

Данный модуль помогает представить данные о близости объектов какой-либо простой пространственной моделью, в которой объекты интерпретируются, например, как города на обычной карте, а различия между ними есть просто расстояния (в частности, данные о странах, политических партиях, автомобилях и т. д.), и всесторонне диагностировать модель.

Модуль содержит инструментарий для выполнения многомерного шкалирования. Матрицы подобия, различия и корреляции могут быть вычислены для большого числа переменных (до 90 переменных) с размерностью до 9 компонент. Начальная конфигурация может быть вычислена автоматически с помощью анализа главных компонент либо задана пользователем. Доступны всевозможные метрики.

### **Модуль «Кластерный анализ»**

Предназначение данного модуля — произвести сложную иерархическую классификацию данных или выделить в них кластеры.

### **Модуль «Лог-линейный анализ»**

Этот модуль позволяет проанализировать сложные многоуровневые таблицы, как это бывает, например, в социологии при анализе массовых опросов. Модуль содержит полный инструментарий лог-линейных процедур для многоуровневых таблиц частот. Могут быть проанализированы оба типа таблиц — полные и неполные (со структурными нулями). Таблицы частот могут быть построены из необработанных данных или непосредственно введены в систему. В процессе обработки пользователь может просматривать как исходную таблицу, так и маргинальные таблицы. Также имеется процедура автоматического выбора наилучшей модели.

### **Модуль «Нелинейное оценивание»**

Данный модуль предназначен для определения нелинейных зависимостей в данных, подгонки к ним функциональных кривых. Модуль предоставляет возможность осуществить подгонку к наблюдаемым данным кривой, по существу, любого типа. Одна из уникальных возможностей модуля (в отличие от

традиционных программ нелинейного оценивания) в том, что в нем не накладываются ограничения на размер файла обрабатываемых данных.

Оценки могут быть построены с помощью оценок метода наименьших квадратов, метода максимального правдоподобия или заданной пользователем функции потерь.

### **Модуль «Каноническая корреляция»**

Модуль включает в себя широкий набор процедур для выполнения канонического корреляционного анализа, исследования связи между двумя множествами переменных. Модуль может обрабатывать векторные данные или корреляционные матрицы и вычислять все стандартные канонические корреляционные статистики, включая собственные векторы, собственные значения, коэффициенты избыточности, канонические веса, нагрузки, критерии значимости для каждого корня и т.д.

### **Модуль «Анализ длительности жизни»**

Если данные представляют собой таблицы длительности жизни, как, например, в медицине или страховании, и вы хотите проанализировать их, то следует воспользоваться этим модулем.

### **Модуль «Анализ временных рядов и прогнозирование»**

Анализ временных рядов активно применяется в деловых, научных и инженерных приложениях. Модуль предлагает широкий набор методов анализа. Имеются большие возможности преобразования исходного временного ряда, интересный набор графиков и статистических процедур. Для восстановления пропущенных наблюдений могут быть использованы интерполяционные методы.

Модуль состоит из нескольких общих процедур, интегрированных вокруг динамического графического представления временных рядов и их сглаживающих/моделирующих преобразований. Пользователь может смоделировать одновременно несколько рядов и выполнить интерактивный "что-если" анализ, наблюдая ряд на графике. Подобно всем графикам в системе *STATISTICA*, графики рядов могут быть представлены на экране в увеличенном режиме, сжатыми, развернутыми или сдвинутыми относительно друг друга.

Методы преобразования рядов включают следующие процедуры: исключение среднего, тренда, взвешенное скользящее среднее, медианное сглаживание, фильтрацию, взятие разностей с любым сдвигом и многое другое.

### **Модуль «Моделирование структурными уравнениями (SEPATH)»**

Данный модуль помогает построить и тестировать различные модели, объясняющие структуру связей между наблюдаемыми переменными. Моделирование структурными уравнениями — мощное средство многомерного статистического анализа, развитое в последние годы и имеющее целью соединить статистические методы с методами теории систем.

## 1.2. Промышленная статистика

Пакет *STATISTICA* обеспечивает обширный выбор специализированных статистических методов для промышленных приложений (большинство из которых фактически может быть эффективно использовано и для непромышленных приложений, например в общественных науках, биологии, маркетинге). Эти методы разбиты на три модуля: “Карты контроля качества”, “Анализ процессов” и “Планирование эксперимента”.

### Модуль «Карты контроля качества»

В модуле предлагается широкий выбор методов анализа качества, включая карты контроля качества. Система предлагает гибкую реализацию Паретто-карт,  $\bar{X}$ -столбчатых карт,  $R$ ,  $S$  (рассеяния),  $C$ ,  $NP$  (биномиальных подсчетов),  $P$  (биномиальных пропорций),  $U$ ,  $CUSUM$  (кумулятивных сумм),  $MA$  (скользящего среднего),  $EWMA$  (скользящего среднего с экспоненциальными весами) – карт, многомерных карт (карты Хоттелинга). Эти карты могут использовать параметры, вычисленные на основе данных либо определенные пользователем.

### Модуль «Анализ процессов»

Модуль включает инструментарий для построения выборочных планов, анализа производительности процессов, калибровочного анализа повторяющихся партий продукции.

### Модуль «Планирование эксперимента»

Модуль содержит возможности строить и анализировать все общие планы экспериментов в промышленности и прикладных областях.

## 1.3. Особенности работы с пакетом STATISTICA

### 1.3.1. Создание отчета

В некоторых случаях полезно создавать файл регистрации всех результатов анализа данных, т.е. содержимого всех таблиц *Scrollsheet* и/или графиков, которые появляются на экране. Такой файл называется отчетом (рис.1.1).

Отчет, в который автоматически выводятся все результаты анализа, называется автоотчетом. Кроме создания автоотчета можно "вручную" направлять в отчет отдельные *Scrollsheet* или графики. Для того чтобы создать автоотчет, необходимо выполнить следующие шаги:

1. Выбрать опцию *Automatically Print All Scrollsheet* (Автоматически печатать все *Scrollsheet*) в диалоговом окне *Page/Output Setup* (Параметры



страницы/Вывода ), доступном из спускающегося меню **File** (Файл), или с помощью двойного нажатия мыши на поле **Output** (Вывод) в строке состояния в низу окна **STATISTICA**. В диалоговом окне «Параметры страницы/Вывода» можно задать:

- канал вывода (принтер, текстовый файл, окно с отчетом),
- заголовок отчета,
- опции автоматического создания отчета.

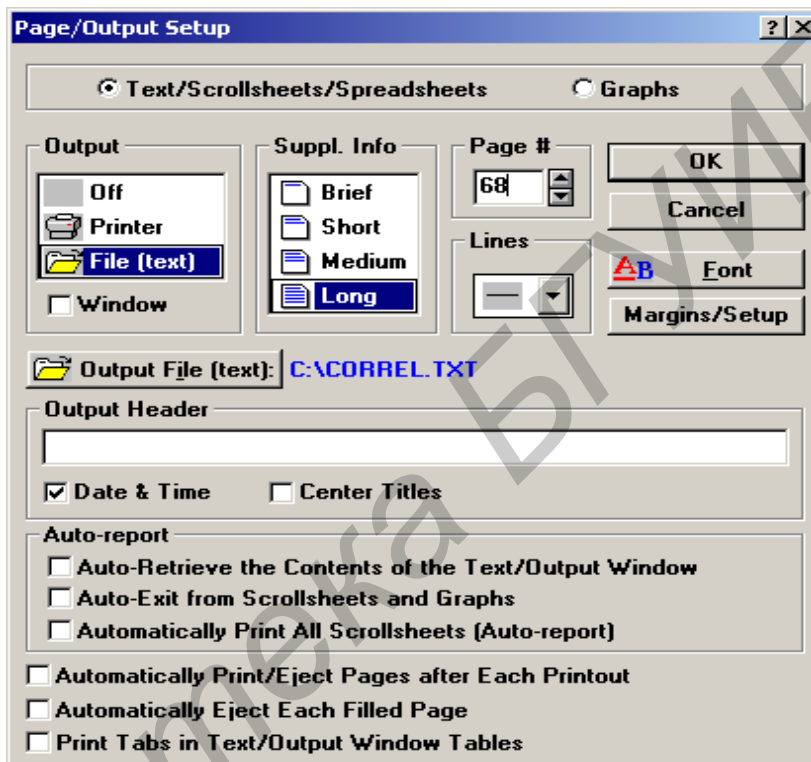


Рис. 1.1. Диалоговое окно, в котором задаются основные параметры канала вывода

2. После этого все таблицы, которые будут появляться на экране, автоматически выведутся в файл с отчетом, который открыт в отдельном окне на рабочем пространстве системы **STATISTICA**.

### 1.3.2. Использование команды **Paste Link** (Специальная вставка) для задания динамического обмена данными

Эта команда доступна из меню **Edit** (Правка) и предназначена для установления динамического обмена данными **DDE** (*Dynamic Data Exchange*), подготовленными в каком-либо другом **Windows** – приложении (источник или сервер), и электронной таблицей в **STATISTICA** (клиент).

Для того чтобы установить динамический обмен данными, необходимо выполнить следующие действия:

- 1) в исходном документе (например, таблице **MS Excel**) выделить данные,

которые следует связать;

2) выбрать команду «Копировать» в меню «Правка»;

3) в электронной таблице *STATISTICA*, с которой следует связать данные, установить указатель туда, куда их следует вставить, и нажать левую клавишу мыши;

4) выбрать команду *Paste Link* (Специальная вставка) в меню *Edit* (Правка);

5) выбрать нужный формат, а затем нажать кнопку "Связать".

Динамическая связь установлена. При таком способе задания связи вся информация о ней хранится в так называемых длинных именах переменных и может быть просмотрена и отредактирована в диалоговом окне *Current Specs* (Текущие спецификации переменной). Можно также редактировать эти связи, используя команду *Link* (Связи).

Команда *Link* (Связи) из меню *Edit* (Правка) открывает диалоговое окно *Links Manager* (Менеджера связей). В нем можно редактировать динамические связи *DDE* (Dynamic Data Exchange) между сервером (*Windows*– приложение, в котором подготовлены данные) и клиентом (документом системы *STATISTICA*).

### 1.3.3. Графические методы анализа данных

Ниже представлены графические методы анализа данных, которыми располагает пакет *STATISTICA*.



**2М столбчатые диаграммы.** На столбчатой диаграмме последовательность значений представлена в виде столбцов (одному наблюдению соответствует один столбец). Если выбрано несколько переменных, то для каждой из них будет построен отдельный график.



**Столбчатая диаграмма по левой оси Y.** На этой диаграмме для каждой точки данных (т.е. для каждой пары координат XY) рисуется один горизонтальный столбец, соединенный с левой осью Y. Вертикальное положение столбца определяется координатой Y точки данных, а его длина - соответствующим значением переменной X.



**2М диаграммы размаха.** На диаграмме размаха стандартная ошибка, стандартное отклонение, минимум-максимум или константа представлены в виде отрезка (с горизонтальными черточками на обоих концах).



**2М гистограммы.** Эти гистограммы являются графическим представлением распределения частот выбранных переменных, на которых для каждого интервала (класса) рисуется столбец, высота которого пропорциональна частоте класса.



**2М линейные графики.** На линейных графиках отдельные точки

данных соединены линией.



**2М круговая диаграмма.** На этой диаграмме пропорции отдельных значений переменной  $X$  представлены в виде круговых секторов.



**Стандартный нормальный вероятностный график.** Данный график строится следующим образом. Сначала все значения упорядочиваются по рангу. По этим рангам рассчитываются значения  $z$  (стандартизованные значения нормального распределения) в предположении, что данные имеют нормальное распределение. Эти значения  $z$  откладываются по оси  $Y$  графика. Если наблюдаемые значения, откладываемые по оси, распределены нормально, то все значения на графике должны попасть на прямую линию. Если наблюдаемые значения не являются нормально распределенными, их расположение отклоняется от прямой. На этом графике можно легко обнаружить выбросы. Если наблюдается очевидное несовпадение, а данные располагаются относительно прямой определенным образом (например, в виде буквы  $S$ ), то к ним, вероятно, нужно применить какое-либо преобразование.



**"Лица Чернова".** Этот тип пиктографиков составляет отдельную группу. Здесь каждое наблюдение представляет собой схематичное изображение лица, определенным чертам которого соответствуют относительные значения выбранных переменных.



**Категоризованные графики.** Эти графики представляют собой наборы двухмерных, трехмерных, тернарных или  $n$ -мерных графиков (таких как гистограммы, диаграммы рассеяния, линейные графики, поверхности, тернарные диаграммы рассеяния и пр.). Эти "входящие" графики располагаются последовательно в одном графическом окне, позволяя сравнивать структуру данных для каждой из указанных подгрупп.

#### 1.4. Основные элементы диалога в модуле «Основные статистики»

В модуле *Basic Statistics/Tables* (Основные статистики/Таблицы) объединены процедуры, наиболее часто используемые на начальном этапе обработки данных, когда выясняется структура, определяются зависимости между данными, производится их группировка.

Статистики, используемые в данном модуле, в основном очень просты - это среднее, коэффициенты корреляции, дисперсия. Другие статистики, математически более сложные и на первый взгляд непривычные, в действительности также имеют простой смысл и доступны. Перечислим основные задачи, которые можно решить в данном модуле.

1. Оценка вероятностных законов распределения и выборочных характеристик случайной величины  $X$  по наблюдениям за ее значениями в независимых опытах, другими словами, по выборке  $(X_1, \dots, X_n)$ .

2. Оценка параметров законов распределения случайных величин.
3. Проверка гипотез о виде законов распределения и о параметрах этих распределений.

### 1.4.1. Основные характеристики случайных величин и их выборочные аналоги

Одной из задач математической статистики является оценка различных вероятностных характеристик случайной величины по ее наблюдениям в независимых опытах или, что то же самое, по выборке  $X_1, \dots, X_n$ . Соответствующие характеристики называются выборочными, так как они построены по выборке. Рассмотрим некоторые теоретические характеристики случайных величин и их выборочные аналоги.

К числу важнейших характеристик случайной величины  $X$  относятся математическое ожидание или среднее значение - обозначение  $E_X$  и дисперсия - обозначение  $D_X$ . Среднее значение представляет собой характеристику положения, а дисперсия – меру отклонения от этого положения. Их выборочными аналогами являются выборочные среднее и дисперсия. Выборочной характеристикой среднего  $E_X$  является величина  $m_n = (X_1 + \dots + X_n)/n$ , называемая выборочным средним. Выборочная дисперсия  $s_n^2$  определяется как  $s_n^2 = ((X_1 - m_n)^2 + \dots + (X_n - m_n)^2)/(n - 1)$ . Корень квадратный из выборочной дисперсии  $s_n$  есть выборочное стандартное отклонение. К характеристикам положения относятся также коэффициент асимметрии  $ass = E(X - E_X)^3 / \sqrt{D_X}^3$ , где  $\sqrt{\phantom{x}}$  – функция извлечения квадратного корня, и коэффициент эксцесса  $eks = E(X - E_X)^4 / (D_X)^2$ . Для получения оценок коэффициентов асимметрии и эксцесса следует заменить фигурирующие в определениях асимметрии и эксцесса средние и дисперсии их выборочными значениями.

Иногда выборочная асимметрия и эксцесс используются для проверки гипотезы о том, что выборка извлечена из нормального распределения. Известно, что для нормального распределения  $ass = 0$  (нормальное распределение симметрично), а  $eks = 3$ . Если выборочные значения асимметрии далеки от 0, а выборочные значения эксцесса далеки от 3, то это служит основанием усомниться в гипотезе о нормальности выборки. Конечно, существуют более тонкие методы тестирования гипотезы нормальности.

Рассмотрим еще одну характеристику случайной величины – квантиль. Обозначим функцию распределения наблюдаемой случайной величины  $X$  через  $F(x)$ . Предположим, что  $p$  есть фиксированное число в интервале  $(0,1)$ . Тогда  $p$ -квантиль (обозначение  $x_p$ ) распределения  $F$  определяется как  $x_p = \sup\{x: F(x) \leq p\}$ . Если  $F$  строго монотонна, то  $x_p$  определяется как корень уравнения:  $P - F(x_p)$ .

Таким образом, квантиль  $x_p$  как функция  $p$  есть обратная функция  $F$ . Квантили с определенными значениями  $p$  имеют специальные названия:

- 0,5-квантиль называется медианой,
- 0,75-квантиль называется верхней квартилью,

- 0,25-квантиль называется нижней квартилью.

Выборочная квантиль получается, если в определении квантили вместо функции  $F$  взять выборочную или эмпирическую функцию распределения  $F_n$ , т.е. оценку функции распределения  $F$ , построенную по выборке  $X_1, \dots, X_n$ .

Эмпирическая функция распределения или выборочная функция распределения определяется как  $F_n(x) = \sum \text{Ind}(X_i < x)/n$ , где  $\text{Ind}(\ )$  есть индикаторная функция, определяемая следующим образом: для любых двух чисел  $a, b$   $\text{Ind}(a < b)$  равен 1, если  $a < b$ , и равен 0 в противном случае. При безграничном увеличении объема выборки  $n$  эмпирическая функция распределения  $F_n$  сходится к функции распределения  $F$ . Таким образом, эмпирическая функция распределения является состоятельной оценкой функции распределения.

В пакете **STATISTICA** легко вычислить разнообразные выборочные характеристики, выбирая пункты меню **Statistics** ( *Статистики* ) в диалоге **Описательные статистики** в модуле «Основные статистики/Таблицы».

### 1.4.2. Гистограммы

Оценкой функции распределения является эмпирическая функция распределения, а оценкой плотности функции распределения – гистограмма. Пусть  $X_1, \dots, X_n$  - исходная выборка. Предположим, что все значения выборки принадлежат отрезку  $[a, b]$ . Разобьем отрезок  $[a, b]$  на  $N$  интервалов:

$$[a, d_1), [d_1, d_2) \dots [d_N, b].$$

Обозначим через  $n_j$  число наблюдений (элементов выборки), попавших в  $j$ -й интервал. Пусть  $d_0 = a$ . Определим функцию  $f(x)$ ,  $a \leq x < b$  равенством:

$$f(x) = f_j \text{ при } d_{j-1} \leq x, j = 0, 1, \dots$$

Определенная таким образом функция  $f$  называется гистограммой. Очевидно, что гистограмму можно определить и через эмпирическую функцию. Действительно, при выполнении следующих условий:

- 1)  $n, N$  безгранично увеличиваются,
- 2)  $n \cdot \max(d_j - d_{j-1})$  безгранично возрастает с ростом  $n, N$

гистограмма приближается к плотности распределения. Условие 2 показывает, что разбиение множества значений выборки на классы при построении гистограммы не должно быть слишком мелким. Тогда гистограмма будет состоятельной оценкой плотности.

Иногда опускают деление на  $n$  и откладывают по оси ординат при изображении гистограмм на графиках просто частоты попадания наблюдений в соответствующий интервал.

Ниже приведены примеры различных гистограмм с наложенными на них плотностями нормального распределения, когда использовано различное число интервалов группировки. Данные взяты из файла **series\_g** в **STATISTICA**. В этом файле представлено число всех пассажиров, перевезенных авиалиниями США за месяц. Наблюдения велись в течение 12 лет. Из графиков на рис. 1.2,

1.3 и 1.4 отчетливо видно, как меняется поведение гистограммы в зависимости от длины интервала группировки.

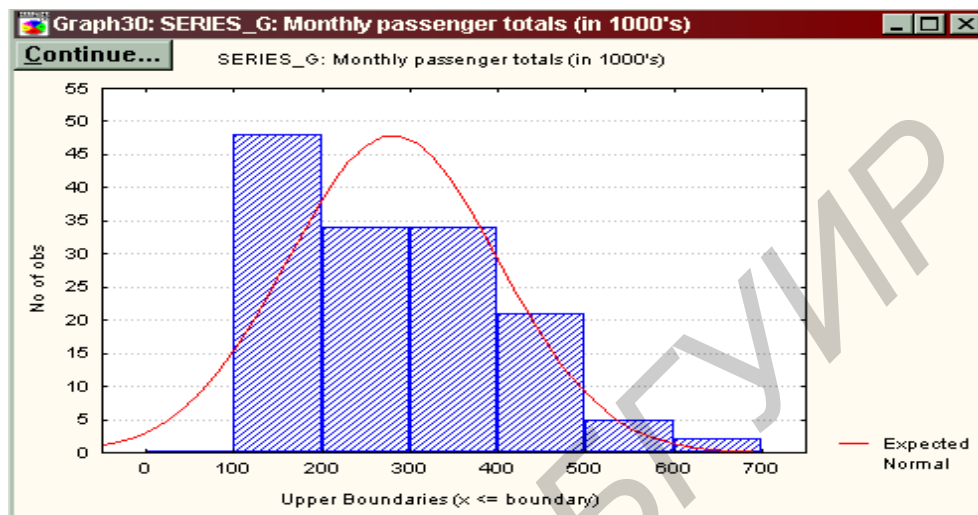


Рис. 1.2. Гистограмма с наложенной плотностью нормального распределения.  
Длина интервала группировки 100

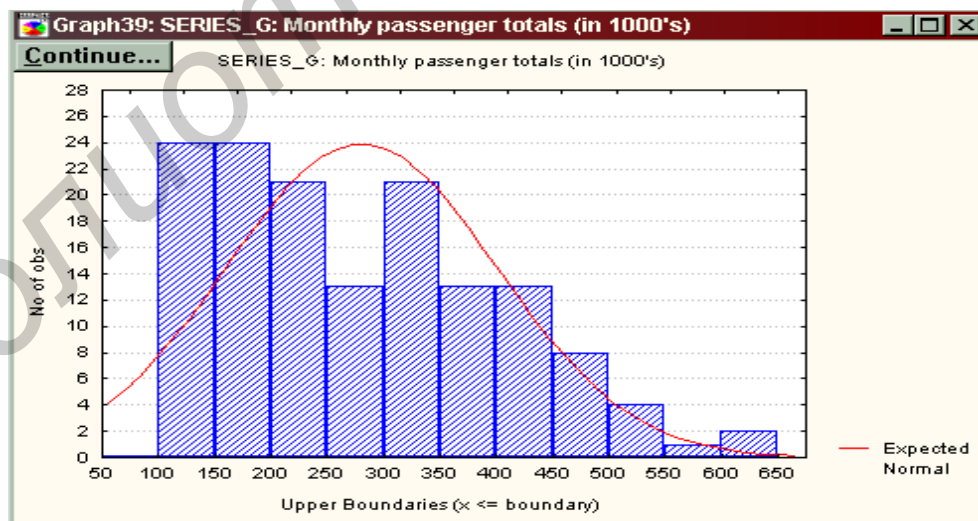


Рис. 1.3. Гистограмма с наложенной плотностью нормального распределения.  
Длина интервала группировки 50

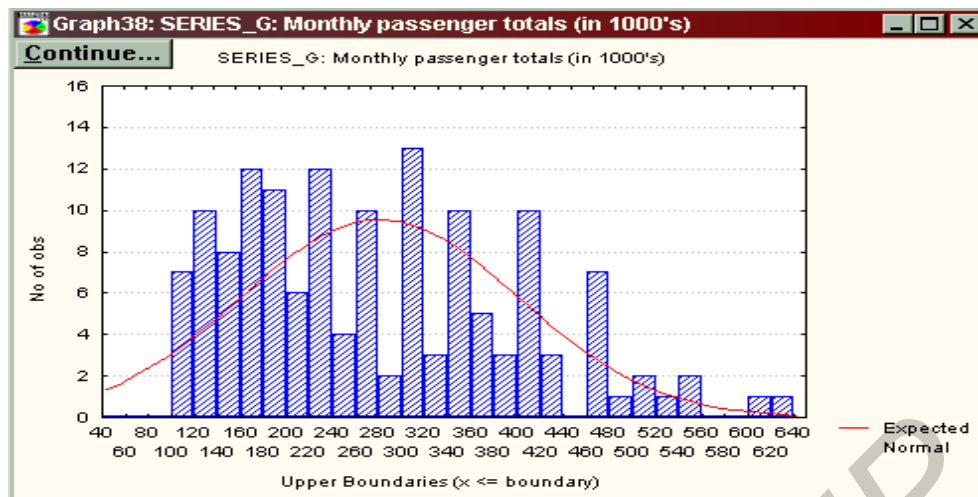


Рис. 1.4. Гистограмма с наложенной плотностью нормального распределения.  
Длина интервала группировки 20

### 1.4.3. Ковариация и корреляция

Перейдем к рассмотрению числовых характеристик двух случайных величин  $X, Y$ - ковариации и коэффициента корреляции. Ковариацией двух случайных величин  $X, Y$  называется величина

$$\text{Cov}(X, Y) = E((X - E_X) \cdot (Y - E_Y)),$$

где  $E$  - символ математического ожидания.

Корреляция – стохастическая линейная зависимость между случайными величинами. Коэффициент корреляции двух случайных величин  $X, Y$  характеризует степень линейной зависимости между двумя случайными величинами и определяется как

$$\text{Cor}(X, Y) = \text{Cov}(X, Y) / \sqrt{D_X \cdot D_Y}).$$

Значения коэффициента корреляции лежат на отрезке  $(-1, 1)$ . Если  $X, Y$  связаны линейной зависимостью:  $Y = a \cdot X + b$ , где  $a, b$  - некоторые константы, то  $|\text{Cor}(X, Y)| = 1$ . Более точно,  $\text{Cor}(X, Y) = 1$ , если  $a > 0$ .  $\text{Cor}(X, Y) = -1$ , если  $a < 0$ .

Если при увеличении значений одной величины увеличиваются значения другой, то говорят - величины положительно коррелированы. Напротив, если при увеличении значений одной величины значения другой уменьшаются, тогда говорят, что случайные величины коррелированы отрицательно. Нулевая корреляция означает, что линейной зависимости нет.

Если  $X, Y$  – нормальные случайные величины, то из равенства нулю коэффициента корреляции следует независимость этих величин.

Выборочные ковариация и коэффициент корреляции получаются, если в определяющих их формулах математические ожидания и дисперсии заменить на их выборочные аналоги.

Выборочные коэффициенты корреляции между различными случайными величинами легко построить в **STATISTICA** в диалоге **Correlation matrices** (Корреляционные матрицы).

#### 1.4.4. Некоторые типы случайных величин и функций распределения

В модуле *Основные статистики/Таблицы* предусмотрена возможность работы с различными законами распределений. Для этого нужно воспользоваться калькулятором вероятностных распределений.

**Нормальное (гауссовское) распределение.** Случайная величина  $X$  называется нормальной или гауссовской, если ее плотность распределения имеет вид:

$$n_x = (1/\sqrt{2 \cdot \pi} \cdot \sigma) \cdot \exp\{- (x - m)^2 / (2 \cdot \sigma^2)\}.$$

Параметр  $m$  в данной формуле есть среднее, а  $\sigma^2$  - дисперсия нормального распределения. Обычно нормальный закон обозначается как  $N(m, \sigma^2)$ .

**Распределение хи-квадрат.** Пусть  $X_1, \dots, X_k$  - независимые нормальные случайные величины со средним 0 и дисперсией 1. Определим случайную величину  $Y_k$  как сумму квадратов этих величин:  $Y = X_1^2 + \dots + X_k^2$ .

Случайная величина  $Y_k$  имеет распределение хи-квадрат с  $k$  степенями свободы.

**Распределение Стьюдента ( $t$ -распределение)** Пусть случайная величина  $X$  имеет стандартное нормальное распределение, а случайная величина  $Y$  имеет распределение хи-квадрат с  $k$  степенями свободы. Определим случайную величину  $t$ , имеющую  $t$ -распределение с  $k$  степенями свободы:

$$t(k) = \sqrt{k} \cdot X/Y.$$

**Распределение Фишера ( $F$ -распределение)** Пусть  $X_1, X_2$  - независимые случайные величины, каждая из которых имеет распределение хи-квадрат с числом степеней свободы  $n_1, n_2$ . Определим случайную величину  $F(n_1, n_2) = (n_2/n_1) \cdot (X_1/X_2)$ .

Определенная таким образом случайная величина  $F(n_1, n_2)$  имеет  $F$ -распределение с числом степеней свободы  $(n_1, n_2)$ .

#### 1.4.5. Задачи проверки гипотез

Важной задачей обработки статистических данных является проверка гипотез относительно наблюдаемых случайных величин. Рассмотрим основные понятия, связанные с проверкой статистических гипотез. Лучше всего начать с самого простого случая - проверки двух альтернативных гипотез. Пусть наблюдаются значения случайной величины  $X$ , имеющей функцию распределения, которую обозначим через  $F_0$ . Функция распределения  $F_0$  неизвестна, и относительно нее выдвигается основная гипотеза:  $H_0: F_0 = F$ .

Альтернативной гипотезой является гипотеза:  $H_1: F_0 = G$ .

Предполагается, что функции распределения  $F$  и  $G$  не равны между собой, иными словами, имеется, по крайней мере, одна точка, где их значения различны.

Итак, по выборке  $X_1, \dots, X_n$  нужно проверить гипотезу  $H_0$  против альтернативы  $H_1$ . Очевидно, что решение принять одну из этих гипотез, может



привести к ошибкам. Возможны ошибки двух родов: принять  $H_1$ , когда верна  $H_0$ , - ошибка 1-го рода; принять  $H_0$ , когда верна  $H_1$ , - ошибка 2-го рода. Вероятности этих ошибок обозначим через  $alpha$  и  $beta$  соответственно. Заметим, что минимизировать вероятности этих ошибок в общем случае нельзя. На практике фиксируют уровень вероятности ошибки первого рода  $alpha$  и стараются минимизировать ошибку второго рода -  $beta$ .

Таким образом, при проверке двух альтернативных гипотез необходимо действовать следующим образом: задать некоторый уровень значимости  $alpha^*$  и строить критерий - правило выбора  $H_0$  или  $H_1$ , у которого вероятность ошибки первого рода не превосходит  $alpha^*$ , а вероятность ошибки второго рода минимальна. Часто критерий задается критериальной статистикой  $z(X_1, \dots, X_n)$ , отображающей наблюдаемые значения выборки на действительную прямую, и критической областью  $W$  действительной прямой.

При попадании наблюдаемых значений статистики  $z$  в область  $W$  гипотеза  $H_0$  отвергается (принимается гипотеза  $H_1$ ). При попадании значений статистики  $z$  в дополнительную область к области  $W$  гипотеза  $H_0$  принимается (гипотеза  $H_1$  отвергается).

Величина  $alpha^*$  называется уровнем значимости критерия. Он отражает степень вашего доверия к гипотезе  $H_0$ . Чем меньше  $alpha^*$ , тем нужны более веские основания, чтобы отвергнуть  $H_0$ . На практике обычно берут  $alpha^* = 0,05, 0,01$  или  $0,001$ . Величина  $1 - beta$  называется мощностью критерия.

В случае двух простых гипотез - функции распределения  $F$  и  $G$  определены точно, критерий максимальной мощности существует и называется критерием Неймана - Пирсона. В общем случае - при тестировании сложных гипотез - таких критериев нет. Отсюда возникает разнообразие критериев при тестировании гипотез, в частности  $t$ -критерий и критерий Колмогорова-Смирнова.

### **$t$ -критерий**

Пусть имеются две выборки: выборка  $X_1, \dots, X_{n_1}$  объема  $n_1$  и выборка  $Y_1, \dots, Y_{n_2}$  объема  $n_2$ . Распределение выборок является нормальным с равными дисперсиями и математическими ожиданиями  $m_1, m_2$  соответственно.

Вам требуется проверить гипотезу  $H_0: m_1 = m_2$ , согласно которой средние двух выборок совпадают, против альтернативы:  $H_1: m_1$  не равно  $m_2$ . Это типичная задача проверки гипотез. Для тестирования данных гипотез применяется  $t$ -критерий. Статистика критерия имеет вид:

$$t(n_1, n_2) = \frac{m(n_1) - m(n_2)}{s(n_1, n_2) \cdot \sqrt{(1/n_1 + 1/n_2)}},$$

где  $m(n_1) = \sum_{i=1}^{n_1} \frac{X(i)}{n_1}, \quad m(n_2) = \sum_{i=1}^{n_2} \frac{Y(i)}{n_2},$

$$s^2 = \left( \sum_{i=1}^{n_1} (X(i) - m_1)^2 + \sum_{i=1}^{n_2} (Y(i) - m_2)^2 \right) / (n_1 + n_2 - 1)$$

Критическая область для  $t$ -критерия с уровнем значимости  $alpha^*$  определяется как

$W = \{t: |t| > t(alpha^*)\}$ , где  $t(alpha^*)$  - корень уравнения:

$1 - F(t(alpha^*)) + F(-t(alpha^*)) = alpha^*$ , где  $F$  - функция распределения Стьюдента (прил. 2) с  $(n_1, n_2)$  степенями свободы.

В системе **STATISTICA** решение данной задачи выполняется мгновенно в диалоге **t-test - t-критерий**. Достаточно задать переменные для анализа щелчком по кнопке **Variables - Переменные** и щелкнуть далее по кнопке **T-test**.

### Критерий согласия Колмогорова – Смирнова

Пусть наблюдаются значения случайной величины  $X$  и необходимо проверить гипотезу  $H_0$ : случайная величина имеет непрерывную функцию распределения  $F$  против альтернативы: функция распределения случайной величины  $X$  не равна  $F$ .

Такие альтернативы называются общими, так как в них не фиксирован вид функции распределения, а просто утверждается, что она не равна  $F$ . Критерии для тестирования такой гипотезы называют критериями согласия, так как они проверяют, в какой степени наблюдаемые значения величины  $X$  согласуются с функцией распределения  $F$ .

По выборке  $X_1, \dots, X_n$  строится эмпирическая функция  $F_n$ , ранее определенная. Для проверки гипотезы  $H_0$  используют критерий Колмогорова – Смирнова, для чего находят значение статистики  $D_n = \sup |F_n(x) - F(x)|$ . Данная статистика показывает, насколько эмпирическая функция отклоняется от функции распределения, заданной по гипотезе  $H_0$ . Величина  $D_n$  измеряет степень согласия  $F$  с  $F_n$ . Далее задается уровень значимости  $alpha^*$  и строится критическая область. Пусть  $K$  – функция распределения статистики Колмогорова – Смирнова. Определим критическую точку  $D(alpha^*)$  из уравнения

$$1 - K(D(alpha^*) \cdot \sqrt{n}) = alpha^*$$

Будем отвергать  $H_0$ , если вычисленное на основании наблюдаемых данных значение  $D_n$  превышает уровень  $D(alpha^*)$ . Этот критерий имеет заданный уровень значимости  $alpha^*$ . Отметим, что функция  $K$ , а значит,  $alpha^*$  не зависит от функции распределения  $F$ . Этот факт имеет очень важное методическое значение.

Функция  $K$  табулирована, и всегда можно решить данное уравнение, по крайней мере, приблизительно. В **STATISTICA** все эти вычисления делаются автоматически.

Часто приходится проверять гипотезу о том, что наблюдаемые величины распределены нормально. Здесь также можно применять критерий Колмогорова – Смирнова. Достаточно выбрать опцию **Критерий Колмогорова - Смирнова** и

система произведет все нужные вычисления.

Если нужно проверить гипотезу о согласии распределения данных с распределениями, отличными от нормального, следует использовать модуль «Непараметрическая статистика». Здесь в диалоговом окне **Distributing Fitting (Подгонка распределения)** можно выбрать различные функции распределения: нормальное, равномерное, экспоненциальное, гамма, логарифмически нормальное, хи-квадрат, биномиальное, пуассоновское, геометрическое, Бернулли и др.

#### 1.4.6. Сравнение средних двух групп в модуле Basic Statistics/Tables

*Постановка задачи* : пусть имеются данные опроса мужчин и женщин об их предпочтениях напитка (*пепси* или *кока*). Нам необходимо проверить гипотезу о равенстве предпочтений в выборе напитка у разных полов.

Для решения поставленной задачи применим *t*-критерий, который, как было указано ранее, позволяет сравнивать средние двух групп.

Загрузите модуль **Basic Statistics/Tables (Основные статистики/Таблицы)**. Откройте файл **adstudy.sta** с исходными данными, который находится в папке **Examples** пакета **STATISTICA**. В этом файле содержится 25 переменных, имеющих 50 значений (случаев), так как в опросе принимало участие 50 человек. В первом столбце **CASENAME** записаны имена опрошенных, во втором столбце **GENDER** – пол, а в третьем столбце **ADVERT** – что предпочитают – *пепси* или *кока*, далее идут переменные (от **MEASUR1** до **MEASUR23**), носящие технический характер.

Загрузите стартовое окно модуля **Basic Statistics/Tables (Основные статистики/Таблицы)**. Щелкните в появившемся окне модуля на строке: **T-test for independent samples (t-критерий для независимых выборок)**. Задайте в строке **Input File – One record per case (use a grouping variable) (Одна запись на случай (используя группирующую переменную))**.

Далее выберите группирующую переменную и независимую переменную. Группирующей переменной является **GENDER (Пол)**, зависимой переменной – **ADVERT (Предпочтение)**.

Щелчком по кнопке **Variables** раскройте окно выбора переменных. В левом списке выберите группирующую переменную (допускается выбор одной переменной), в правом – зависимые переменные (можно выбрать несколько переменных). Щелкните далее **OK** или просто нажмите **ENTER** на клавиатуре. В появившемся окне вновь наберите в строке **Code for Group 1 (Код для группы 1) Male**, а в строке **Code for Group 2 (Код для группы 2) – Female**.

Щелкните **OK**, на экране появится следующая таблица:

Grouping: GENDER: Gender of the subjects (May 15, 1996).							
Continue...		Group 1: MALE Group 2: FEMALE					
Variable	t-value	df	p	Valid N MALE	Valid N FEMALE	Std.Dev. MALE	S
ADVERT	1,205214	48	,234029	28	22	,507875	

Рис 1.5. Таблица результатов тестирования

В таблице представлен уровень значимости  $p$ . Уровень значимости - это число, показывающее, насколько значима для вас высказанная гипотеза. Если у вас имеется большая уверенность в правильности гипотезы, вы задаете вероятность ошибки как можно меньше, уменьшая тем самым вероятность неверного решения - отвергнуть гипотезу, когда она верна. Чем тяжелее последствия ошибки первого рода, тем меньшим следует выбирать уровень значимости. Нужно, однако, помнить, что при уменьшении уровня значимости увеличивается вероятность ошибки второго рода. Обычно задают следующие небольшие уровни значимости: 0,1, 0,05 и т. д. Но и в больших уровнях есть свой резон, в конечном счете, это ваше право - вы будете платить за неверное решение, если отвергнете правильную гипотезу. В данном случае уровень значимости  $p$  равен 0,234029. На основании этого мы должны признать, что собранные данные не дают серьезных оснований говорить о том, что пол человека влияет на выбор типа напитка.

## 1.5. ВАРИАНТЫ ЗАДАНИЙ

1. Ознакомиться с основными модулями системы *STATISTICA*.
2. Подготовить отчет для отображения результатов работы в виде *RTF*-файла.
3. Создать электронную таблицу с исходными данными. Преобразовать исходные данные - добавить переменную *NEWVAR* в соответствии с вариантами задания, указанными ниже.
4. Создать файл *Excel*, содержащий статистические данные. Установить связь с данными из этого файла.
5. Ознакомиться с принципами работы модуля *Basic statistic/Tables* (*Основные статистики/Таблицы*).
6. Исследовать основные характеристики исходных данных: выборочное среднее и дисперсия, асимметрия и эксцесс, ковариация, коэффициент корреляции.
7. Построить гистограммы.
8. Проверить гипотезу о равенстве средних двух выборок.
9. Построить:
  - графики категоризации всех типов, при этом разбить данные по категориям: квартал, тип, сорт и т. д.;
  - графики пользовательские *CUSTOM 2D* и *CUSTOM 3D*.

**Вариант****NEWVAR**

1	$(VAR1+VAR2)/3$
2	$lg(VAR1)*cos(VAR2)$
3	$sin(VAR1)/tg(VAR2)$
4	$2cos^3(VAR1+VAR2)$
5	$2tg^5(VAR1+VAR2)$
6	$exp(VAR1+3VAR2)/(VAR1-VAR2)$
7	$exp(VAR1)+VAR2$
8	$(VAR1+3VAR2)^2/(VAR1-VAR2)^2$
9	$2VAR1+2.3VAR2$
10	$VAR1*VAR2+lg(VAR1+VAR2)$
11	$VAR2+sin(VAR2)/VAR1$
12	$VAR2+cos(VAR1)$

Библиотека БГУИР

## Лабораторная работа №2

# РЕГРЕССИОННЫЙ АНАЛИЗ ЭКОНОМЕТРИЧЕСКОЙ ИНФОРМАЦИИ

*Цель работы* - научиться строить регрессионные модели, проводить анализ качества модели.

### 2.1. Общие сведения о модуле

Методы регрессионного анализа, позволяющие моделировать статистические зависимости между двумя или несколькими переменными, представлены в **STATISTICA** модулем *Multiple Regression* (*Множественная регрессия*). В нем реализованы различные методы множественной линейной, пошаговой и фиксированной нелинейной регрессии (в частности, полиномиальной, экспоненциальной, логарифмической). **STATISTICA** позволяет вычислить все необходимые статистики и оценить адекватность построенной модели. Анализ остатков и выбросов может быть проведен при помощи широкого набора графиков, включая разнообразные точечные графики, графики частичных корреляций и многие другие. Система прогноза позволяет пользователю выполнять анализ "что - если".

Методы общей нелинейной регрессии реализованы в модуле *Nonlinear Estimation* (*Нелинейное оценивание*). Он позволяет строить произвольную регрессионную модель, задаваемую некоторой алгебраической формулой, которая может быть нелинейной как по переменным, так и по параметрам. Для расчета модели могут использоваться различные итерационные алгоритмы минимизации. Программа осуществляет полный контроль за всеми аспектами вычислительных процедур (начальное значение, размер шага, критерий сходимости и т.д.). Большинство обычных нелинейных регрессионных моделей задано в модуле и может быть просто выбрано из меню.

При работе с моделями множественной регрессии необходимо провести предварительный анализ целесообразности включения выбранных переменных в регрессионную модель.

Установка флажка в поле *Review descr. stats, corr. Matrix* (*Обзор описательных статистик, корреляционная матрица*) позволит провести предварительный анализ исходных переменных и построить корреляционную матрицу, анализ которой дает возможность сделать важные выводы о структуре связей между выбранными переменными.

Не рекомендуется включать в модель переменные, слабо связанные с результативным признаком, а также переменные, тесно связанные друг с другом. Когда между объясняющими переменными существует ощутимая линейная зависимость, говорят о существовании мультиколлинеарности. В этом случае решение становится неустойчивым, незначительное изменение состава выборки

(значений признаков) или состава объясняемых переменных может вызвать кардинальное изменение модели, что делает ее использование малоприменимым в практических целях. Наиболее распространенные в таких случаях приемы: исключение одной из двух сильно связанных переменных, использование гребневой регрессии, переход от первоначальных переменных к их главным компонентам.

Если сбросить флажок в поле *Perform default analysis* (Метод анализа по умолчанию), то появляется доступ к диалоговому окну *Model Definition*, открывающему возможность дополнительного выбора методов анализа, среди которых имеются методы пошаговой (*Stepwise*) и гребневой (*Ridge*) регрессии.

Методы пошаговой регрессии позволяют из множества независимых переменных отобрать только те, которые наиболее значимы для адекватного описания многопараметрической регрессии. В модуле реализованы две процедуры отбора переменных, каждая из которых может давать различный конечный набор переменных: последовательное включение (*Forward stepwise*) и последовательное исключение (*Backward stepwise*).

Гребневая регрессия используется для получения более устойчивых оценок параметров регрессионной модели в условиях мультиколлинеарности переменных.

## 2.2. Примеры реализации диалога в модуле «Множественная регрессия»

Рассмотрим примеры решения задачи построения и анализа качества модели множественной регрессии.

В табл. 2.1 даны курсы покупки и продажи акций «Иркутскэнерго» и «Красноярскэнерго». Ниже представлены расшифровки использованных в таблице обозначений:

- *Дата* — дата сделки;
- *Иркут1* — цена покупки акции «Иркутскэнерго»;
- *Иркут2* — цена продажи акции «Иркутскэнерго»;
- *Краснояр1* — цена покупки акции «Красноярскэнерго»;
- *Краснояр2* — цена покупки акции «Красноярскэнерго»;

**Пример 1.** Провести анализ зависимости цены покупки акций «Иркутскэнерго» от цены продаж в модуле *Multiple Regression* (Множественная регрессия) по следующей схеме:

- 1) открыть модуль;
- 2) выбрать список зависимых и независимых переменных;
- 3) провести оценивание параметров модели;
- 4) проверить качество полученных оценок параметров;
- 5) провести анализ адекватности модели.

Таблица 2.1

№ п.п.	Дата	Иркут1	Иркут2	Краснояр1	Краснояр2
1	22/10/96	0.0966	0.1103	0.1345	0.1758
2	23/10/96	0.0980	0.1110	0.1362	0.1792
3	24/10/96	0.0986	0.1117	0.1362	0.1808
4	25/10/96	0.0992	0.1121	0.1362	0.1808
5	28/10/96	0.0988	0.1115	0.1362	0.1808
6	29/10/96	0.0992	0.1114	0.1372	0.1858
7	30/10/96	0.0979	0.1121	0.1372	0.1858
8	31/10/96	0.0981	0.1149	0.1378	0.1867
9	01/11/96	0.0972	0.1118	0.1409	0.1800
10	04/11/96	0.0974	0.1118	0.1399	0.1800
11	05/11/96	0.0981	0.1119	0.1350	0.1800
12	06/11/96	0.0990	0.1127	0.1322	0.1800
13	10/11/96	0.0986	0.1127	0.1280	0.1825
14	11/11/96	0.0987	0.1129	0.1273	0.1860
15	12/11/96	0.0996	0.1137	0.1236	0.1717
16	13/11/96	0.1000	0.1145	0.1239	0.1808
17	14/11/96	0.1003	0.1156	0.1258	0.1875
18	15/11/96	0.1016	0.1163	0.1288	0.1756
19	18/11/96	0.1018	0.1161	0.1347	0.1983
20	19/11/96	0.1039	0.1176	0.1386	0.2206
21	20/11/96	0.1068	0.1236	0.1532	0.2228

**Шаг 1.** В пакете *STATISTICA* откройте модуль *Multiple regression* (*Множественная регрессия*). Высветите название модуля и далее нажмите кнопку *Switch to* (*Переключиться в*) либо просто дважды щелкните мышью по названию модуля: *Multiple regression*.

**Шаг 2.** На экране появится стартовая панель модуля. Нажмите кнопку *Open Data* (*Открыть данные*) и откройте файл данных, в котором находятся исходные данные. Далее выберите переменные для анализа. Выбор переменных осуществляется с помощью кнопки *Variables* (*Переменные*), находящейся в левом верхнем углу панели.

После того как кнопка будет нажата, диалоговое окно *Select dependent and independent variable list* (*Выбрать списки зависимых и независимых переменных*) появится на вашем экране. Высветив имя переменной в левой части окна, выберите зависимую переменную. Высветив имя переменной в правой части окна, выберите независимую переменную. То же можно сделать, просто набрав номера переменных в строках: *Dependent variable list* (*Список зависимых переменных*) и *Independent variable list* (*Список независимых переменных*).

В данном примере независимой переменной является переменная *Иркут1*, зависимой— *Иркут2*. Высветив имена этих переменных, нажмите кнопку *OK* в правом верхнем углу окна *Select dependent and independent variable list* (*Выбрать списки зависимых и независимых переменных*). Вы вновь окажетесь в стартовой панели модуля. Переменные для анализа выбраны. Никаких



дополнительных установок в стартовой панели в данном случае производить не нужно. Нажмите кнопку **OK** в правом углу стартовой панели.

**Шаг 3.** На экране перед вами появится диалоговое окно *Model Definition* (Построение модели) (рис. 2.1).

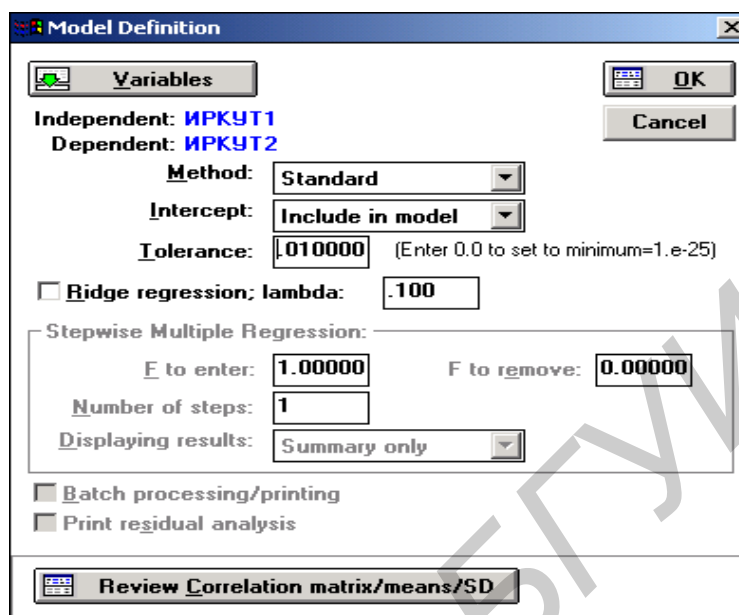


Рис. 2.1. Окно построения модели в модуле *Множественная регрессия*

В данном окне выберите стандартный метод оценивания в опции *Method* (Method): **Standard** (Стандартный). Далее нажмите кнопку **OK**. Программа произведет оценивание параметров модели стандартным методом, и через секунду на экране появится диалоговое окно результатов, показанное на рис. 2.2.

**Шаг 4.** В диалоговом окне *Multiple Regression Results* (Результаты множественной регрессии) просмотрите результаты оценивания, которые представлены в численном и графическом виде.

Окно результатов анализа имеет следующую структуру: верх окна — информационный. Он состоит из двух частей: в первой части содержится основная информация о результатах оценивания, во второй высвечиваются значимые регрессионные коэффициенты. В низу окна *Результаты множественной регрессии* находятся функциональные кнопки, позволяющие всесторонне просмотреть результаты анализа (рис. 2.2).

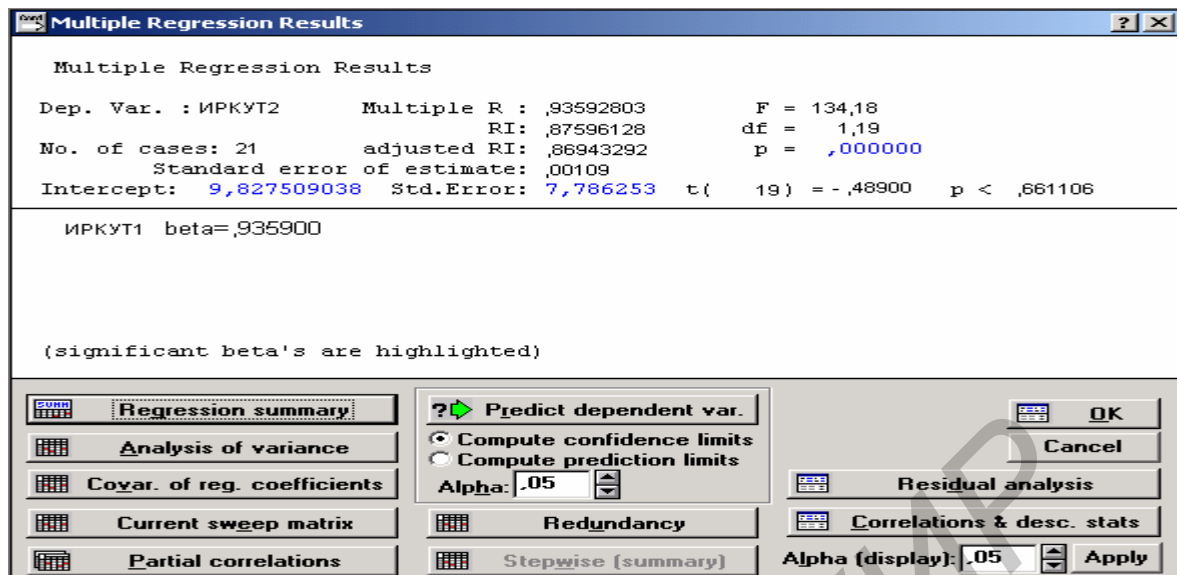


Рис. 2.2. Окно *Результаты множественной регрессии*

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

- **Dep. Var.** (*Имя зависимой переменной*) - в данном случае *Иркут2*;
- **No. of Cases** (*Число наблюдений, по которым построена регрессия*) - в примере это число равно 21;
- **Multiple R** (*Коэффициент множественной корреляции*);
- **R-square** ( $R^2$  - *коэффициент детерминации, квадрат коэффициента множественной корреляции*) используется для статистической оценки тесноты связи между результативным и объясняющими показателями. Он выражает долю объясненной изучаемыми факторами дисперсии результативного признака и служит важной характеристикой качества построенной модели. Этот коэффициент может принимать значения от 0 до 1. Несмещенной оценкой  $R^2$  служит скорректированный на потерю степеней свободы коэффициент множественной детерминации (*Adjusted R<sup>2</sup>*);
- **Adjusted R-square** (*Скорректированный коэффициент детерминации*), определяемый как

$$\text{Adjusted } R\text{-square} = 1 - (1 - R^2) \cdot (n / (n - p)),$$

где  $n$  — число наблюдений в модели,  $p$  — число параметров модели (число независимых переменных плюс 1, так как в модель включен свободный член);

- **Std. Error of estimate** (*Стандартная ошибка оценки*) - эта статистика является мерой рассеяния наблюдаемых значений относительно регрессионной прямой;
- **Intercept** (*Оценка свободного члена регрессии*) - значение коэффициента  $B_0$  в уравнении регрессии;
- **Std. Error** (*Стандартная ошибка оценки свободного члена*) - стандартная ошибка коэффициента  $B_0$  в уравнении регрессии;
- **t(df) and p-value** (*Значение t-критерия и уровень p*) - t-критерий используется для проверки гипотезы о равенстве 0 свободного члена регрессии

(см. прил. 2);

•  $F$ - критерий Фишера, определяющий значимость полученной модели. Оценивает вероятность случайного отклонения от нуля коэффициента детерминации при отсутствии связи между элементами совокупности. Желательно, чтобы полученный минимальный уровень значимости  $F$ -критерия ( $p$ -level) был меньше 0,05 (см. прил. 2);

•  $df$  — число степеней свободы  $F$ -критерия;

•  $p$  — уровень значимости.

В информационной части посмотрим прежде всего на значения коэффициента детерминации, которые лежат в пределах от 0 до 1. В нашем примере  $R^2 = 0,876$ . Это значение показывает, что построенная регрессия объясняет более 87,6% разброса значений переменной *Иркут2* относительно среднего.

Далее посмотрите на значение  $F$ -критерия и уровень значимости  $p$ .  $F$ -критерий используется для проверки гипотезы о значимости регрессии. В данном случае для проверки гипотезы, утверждающей, что между зависимой переменной *Иркут2* и независимой переменной *Иркут1* нет линейной зависимости, т. е.  $B_1 = 0$ , против альтернативы:  $B_1$  не равен 0. В данном примере мы имеем большое значение  $F$ -критерия — 134,18, а представленный в окне уровень значимости  $p = 0,00$ , показывает, что построенная регрессия высоко значима.

Рассмотрим вторую часть информационного окна. В ней представлена информация о значимых и незначимых оценках регрессионных коэффициентов. При этом высвечивается строка: *Иркут1 beta = 0,936* и приводится пояснение **Significant beta's are highlighted** (Значимые beta высвечены). Отметим, что в данном случае  $beta$  есть стандартизованный коэффициент  $B_1$ , т. е. коэффициент при независимой переменной *Иркут1*.

Перейдем в функциональную часть окна результатов.

Прежде всего нажмите кнопку **Regression summary** (Итоговый результат регрессии). На экране появится **Spreadsheet** (Электронная таблица вывода), в которой представлены итоговые результаты оценивания регрессионной модели (рис. 2.3).

Continue...	R= .93592803 RI= .87596128 Adjusted RI= .86943292 F(1,19)=134.18 p<.00000 Std.Error of estimate: .00109					
N=21	BETA	St. Err. of BETA	B	St. Err. of B	t (19)	p-level
Intercept			-.005000	.010200	-.48900	.661106
ИРКУТ1	.935900	.080800	1.192500	.102900	11.58400	.000000

Рис. 2.3. Итоговая таблица регрессии

В первом столбце таблицы даны значения коэффициентов  $beta$  (стандартизованные коэффициенты регрессионного уравнения), во втором — стандартные ошибки этих коэффициентов, в третьем — точечные оценки параметров модели:

• свободный член  $B_0 = -0,0050$ ;

- коэффициент  $B_1$  (при независимой переменной  $Иркут1$ ) = 1,1925. Далее представлены стандартные ошибки для  $B_0$ ,  $B_1$ , значения статистик t-критерия и т.д. По итоговой таблице регрессии можно построить модель, которая имеет вид

$$Иркут2 = 1,1925 \cdot Иркут1 - 0,0050.$$

**Шаг 5.** Оценка адекватности модели - важный элемент анализа. После того как доказана адекватность модели, полученные результаты можно уверенно использовать для дальнейших действий.

Анализ адекватности основывается на анализе остатков. Остатки представляют собой разности между наблюдаемыми значениями и модельными, т.е. значениями, подсчитанными по модели с оцененными параметрами.

Графики зависимости регрессионных остатков от экспериментальных значений исходных переменных позволяют проверить предположения об однородности и независимости ошибок, являющихся предпосылками применения метода наименьших квадратов, и локализовать выбросы. Если указанные допущения выполняются, графики будут представлять собой симметричное, случайное и равномерное распределения точек. Графики эмпирической функции распределения остатков на нормальной вероятностной бумаге (*Probability plots*) и гистограммы позволяют проверить справедливость предположения о нормальном распределении остатков. Кроме этого, имеется возможность вычислить статистику Дарбина–Уотсона (*Darbin-Watson Stat*) для проверки наличия автокорреляции в остатках, вывести на экран (*Display residuals&pred*) и сохранить в файле (*Save residuals*) информацию о наблюдаемых и подобранных по модели значениях результативного показателя и остатках.

Рекомендуется также построить график линейной зависимости предсказанных (подобранных по модели) значений зависимой переменной от наблюдаемых (*Predicted & Observed*), что позволяет наглядно изобразить результаты регрессионного анализа.

В модуле *Множественная регрессия* имеется специальное диалоговое окно, в котором проводится всесторонний анализ остатков. Нажмите кнопку *Residual Analysis* (*Анализ остатков*) - на экране появится диалоговое окно *Residual Analysis* (*Анализ остатков*). Нажмите в этом окне, например, кнопку *Obs&residuals*, на экране появится график (рис. 2.4), который говорит о достаточной адекватности модели.

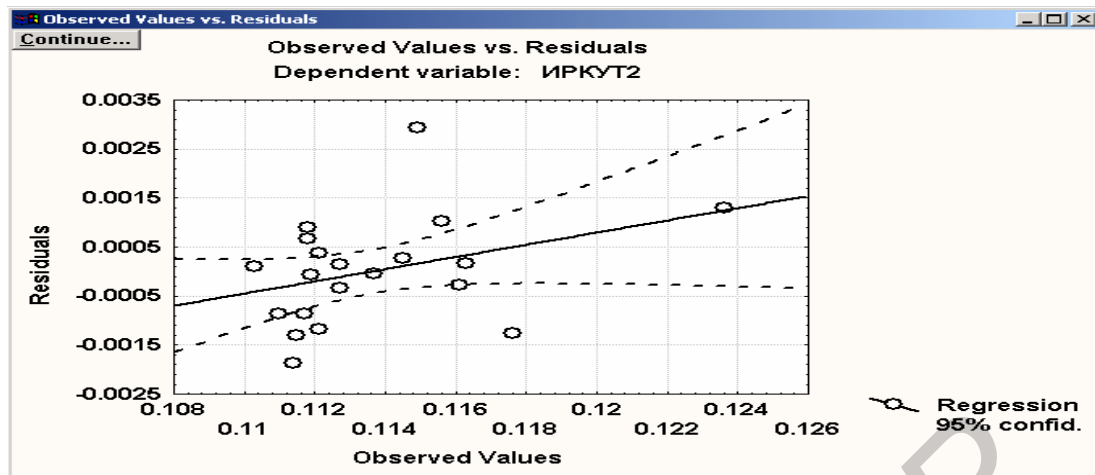


Рис. 2.4. График наблюдаемых переменных-остатков

Часто, если остатки не распределены по нормальному закону, а также для стабилизации дисперсии данных применяют преобразования зависимых и независимых переменных, например логарифмическое преобразование зависимых переменных или извлечение квадратного корня.

**Пример 2.** Построить зависимость между ценой продаж акций «Красноярскэнерго» и «Иркутскэнерго».

Основные наши действия те же, что и в предыдущем примере. В данном примере независимой переменной является *Иркут2* — цена продажи «Иркутскэнерго», зависимой — *Краснояр2* — цена продажи «Красноярскэнерго».

Сделайте соответствующие установки в окне **Select dependent and independent variable list** и нажмите кнопку **OK**. Система произведет вычисления, и на экране появится следующее окно результатов (рис. 2.5).

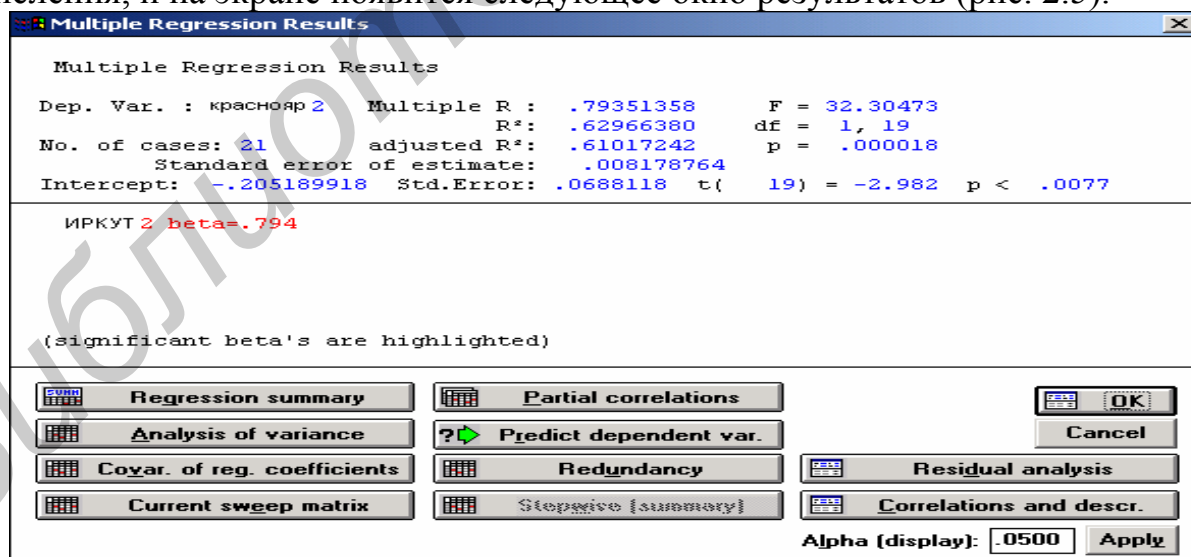


Рис. 2.5. Окно результатов в примере с акциями «Красноярскэнерго», «Иркутскэнерго»

Нажмите кнопку **Regression summary** (*Итоговый результат регрессии*), на экране появится электронная таблица с численными результатами оценивания регрессионной модели. Затем нажмите кнопку **Residual Analysis** (*Анализ*

остатков), в появившемся диалоговом окне посмотрите, как связаны остатки с наблюдаемыми значениями (рис. 2.6).

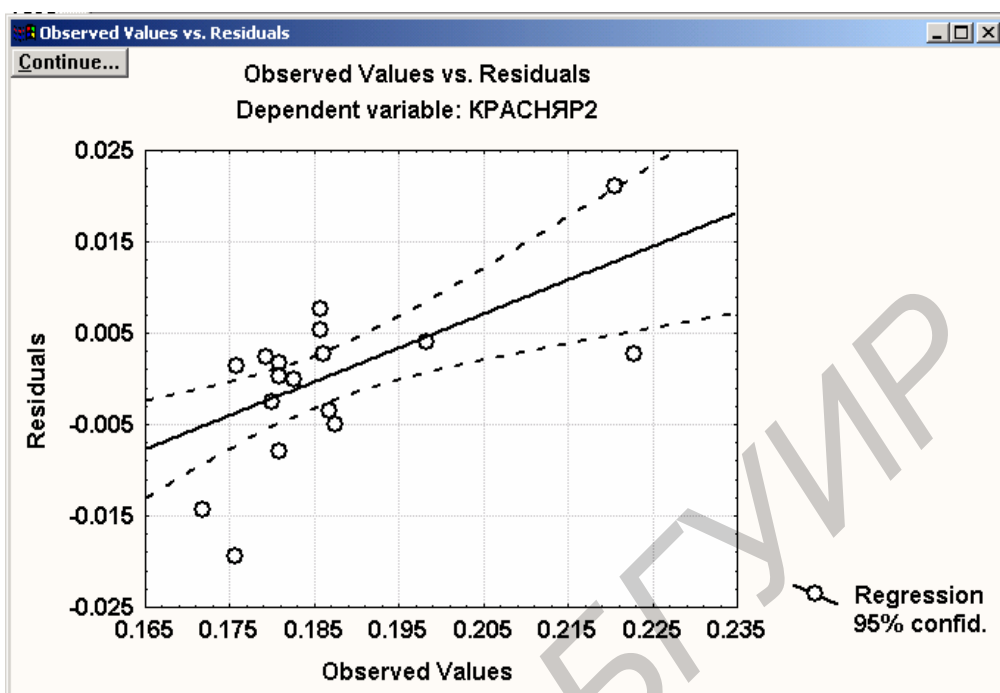


Рис. 2.6. График наблюдаемых переменных-остатков.

В окне *Анализ остатков* нажмите кнопку *Pred&observed(F)* и посмотрите, как наблюдаемые значения связаны с предсказанными с помощью построенной модели (рис. 2.7).

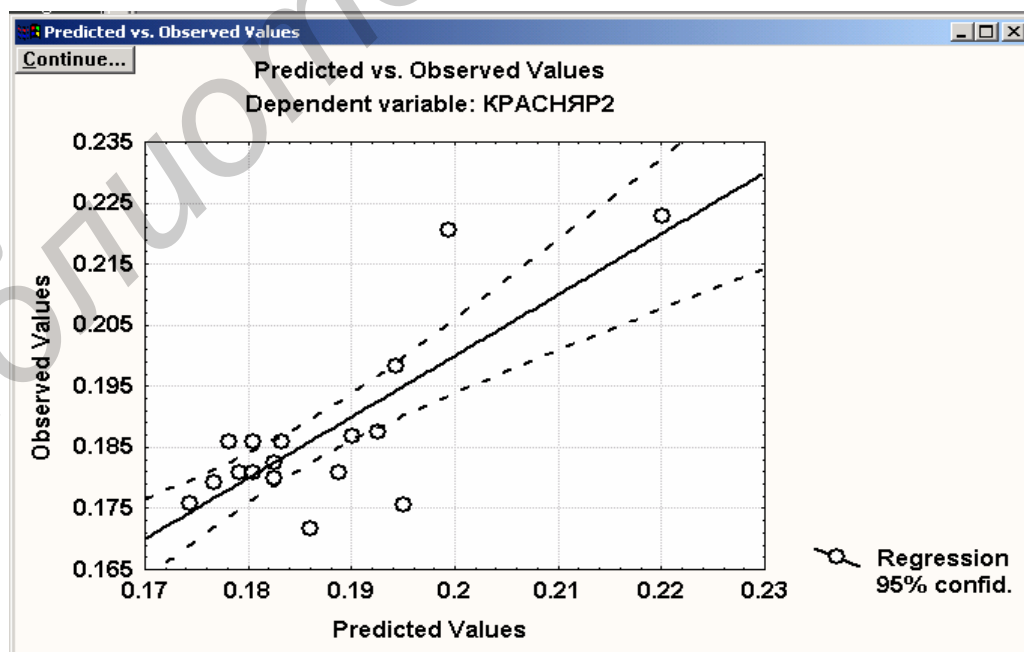


Рис. 2.7. График наблюдаемых и предсказанных значений

Из графиков 2.6, 2.7 видно, что модель достаточно адекватно описывает данные. Следовательно, с ее помощью можно делать достаточно точные

выводы о зависимости стоимости акций «Красноярскэнерго» и «Иркутск-энерго» за рассматриваемый промежуток времени .

## 2.3. ВАРИАНТЫ ЗАДАНИЙ

Ниже приведены данные к задачам *A, B, C*.

**Задача А.** В вариантах заданий 1–12 по выборочным данным необходимо исследовать зависимость между показателями  $X$ ,  $Y$  и построить множественную регрессионную модель, для чего следует:

- установить наличие связи между исследуемыми показателями графическим методом (построить корреляционное поле);
- построить регрессионную модель с использованием не менее двух уравнений регрессии и отобрать лучшее по критерию минимума остаточной дисперсии;
- оценить адекватность модели.

Для измерения интенсивности связи между показателями:

- вычислить коэффициент корреляции, коэффициент детерминации, индекс корреляции;
- вычислить ошибки коэффициента корреляции и параметров модели с заданной доверительной вероятностью;
- оценить значимость коэффициента регрессии модели по критерию Стьюдента.

Предусмотреть создание отчета.

Зависимость между показателями  $X_1$ ,  $X_2$ ,  $X_3$  реализованной продукции и балансовой прибылью  $Y$  предприятий одной из отраслей промышленности характеризуется данными, представленными ниже согласно варианту.

1

$X_1$	2	3	4	3	2	6	5	7	8	12	9
$X_2$	1.2	1.8	2.0	2.5	3.0	3.2	3.5	4.9	5.0	6.2	7.3
$X_3$	1.7	2.2	8.6	1.3	3.4	3.9	4.7	5.8	3.6	6.4	7.2
$Y$	20	75	41	82	106	129	145	180	210	250	262

2

$X_1$	5.5	10.5	12.6	15.3	16.0	17.2	18.9	19.4	20.1	21.6	22.0
$X_2$	9	10	12	13	15	17	19	21	25	27	29
$X_3$	1.2	1.8	2.0	2.5	3.0	3.2	3.5	4.9	5.0	6.2	7.3
$Y$	20	25	34	30	36	37	40	46	58	69	80

3

$X_1$	1.2	2.8	3.4	4.6	5.2	6.4	7.8	8.3	9.1	9.9	10.5
$X_2$	1.2	1.8	2.0	2.5	3.0	3.2	3.5	4.9	5.0	6.2	7.3
$X_3$	2	3	4	3	2	6	5	7	8	12	9
$Y$	20	50	57	63	22	75	60	81	87	102	95

4

$X_1$	10.2	15.3	18.4	20.5	24.7	25.6	27.3	28.3	29.6	30.1	31.0
$X_2$	1.4	2.6	3.2	4.8	5.6	6.3	7.7	8.1	9.5	10.2	11.3
$X_3$	3	2	5	4	6	7	9	11	5	4	9
$Y$	20	15	57	47	69	74	89	107	48	52	78

Зависимость между показателями  $X_1$ ,  $X_2$ ,  $X_3$  основных фондов и объемом валовой продукции  $Y$  предприятия одной из отраслей промышленности характеризуется следующими данными:

5

$X_1$	1.1	2.3	3.5	4.1	5.7	6.6	7.3	8.5	9.8	10.1	12.0
$X_2$	1.2	2.8	3.4	4.6	5.2	6.4	7.8	8.3	9.1	9.9	10.5
$X_3$	1.4	2.6	3.2	4.8	5.6	6.3	7.7	8.1	9.5	10.2	11.3
$Y$	20	25	31	32	40	56	52	60	61	70	75

6

$X_1$	1.5	2.6	3.5	4.8	5.9	6.3	7.2	8.9	9.5	11.1	15.0
$X_2$	10.2	15.3	18.4	20.5	24.7	25.6	27.3	28.3	29.6	30.1	31.0
$X_3$	1.1	2.3	3.5	4.1	5.7	6.6	7.3	8.5	9.8	10.1	12.0
$Y$	21	26	30	31	39	54	51	63	65	72	78

7

$X_1$	1.4	2.6	3.2	4.8	5.6	6.3	7.7	8.1	9.5	10.2	11.3
$X_2$	1.1	2.3	3.5	4.1	5.7	6.6	7.3	8.5	9.8	10.1	12.0
$X_3$	1.5	2.6	3.5	4.8	5.9	6.3	7.2	8.9	9.5	11.1	15.0
$Y$	30	35	41	43	50	61	68	73	79	81	93

8

$X_1$	25	30	32	37	41	53	59	63	71	69	80
$X_2$	11.4	16.8	17.2	21.5	25.8	27.9	28.4	30.1	31.6	34.8	37.2
$X_3$	1.2	2.8	3.4	4.6	5.2	6.4	7.8	8.3	9.1	9.9	10.5
$Y$	25	30	32	37	41	53	59	63	71	69	80



Зависимость между показателями  $X_1, X_2, X_3$  располагаемого дохода и объемом частного потребления  $Y$  в определенном периоде одной из стран характеризуется данными, представленными ниже:

9

$X_1$	7.0	7.9	8.2	8.9	9.4	9.9	10.7	11.2	12.1	15.7	16.0
$X_2$	5.5	10.5	12.6	15.3	16.0	17.2	18.9	19.4	20.1	21.6	22.0
$X_3$	10.2	15.3	18.4	20.5	24.7	25.6	27.3	28.3	29.6	30.1	31.0
$Y$	11.4	16.8	17.2	21.5	25.8	27.9	28.4	30.1	31.6	34.8	37.2

10

$X_1$	5.5	10.5	12.6	15.3	16.0	17.2	18.9	19.4	20.1	21.6	22.0
$X_2$	1.5	2.6	3.5	4.8	5.9	6.3	7.2	8.9	9.5	11.1	15.0
$X_3$	9	10	12	13	15	17	19	21	25	27	29
$Y$	13	15	14	17	16	19	20	22	28	30	32

11

$X_1$	9	10	12	13	15	17	19	21	25	27	29
$X_2$	1.2	2.8	3.4	4.6	5.2	6.4	7.8	8.3	9.1	9.9	10.5
$X_3$	5.5	10.5	12.6	15.3	16.0	17.2	18.9	19.4	20.1	21.6	22.0
$Y$	7.1	7.9	8.3	10.6	13.6	15.2	17.8	16.3	17.9	18.9	20.6

12

$X_1$	1.4	2.6	3.2	4.8	5.6	6.3	7.7	8.1	9.5	10.2	11.3
$X_2$	1.1	2.3	3.5	4.1	5.7	6.6	7.3	8.5	9.8	10.1	12.0
$X_3$	7.0	7.9	8.2	8.9	9.4	9.9	10.7	11.2	12.1	15.7	16.0
$Y$	13	15	16	17	18	21	22	20	25	24	26

**Задача В.** Исследование зависимости между расходами на жилье ( $y$ ), располагаемым личным доходом ( $x$ ) и индексом реальных цен на жилье ( $p$ ) показало, что имеет место следующая логарифмическая регрессионная модель:

$$\log y_t = \alpha + \beta_1 \log x_t + \beta_2 \log p_t + u_t.$$

Индекс  $t$  показывает, что текущие расходы на жилье связаны с текущим доходом. Данные по  $y_t, x_t, p_t$  за рассматриваемый период времени приведены в табл. 2.2.

Таблица 2.2

Расходы на жилье ( $y$ ), располагаемый личный доход ( $x$ ) и индекс реальных цен или относительной цены на жилье ( $p$ ) в США (1959–1983 гг.)

Год	$y_t$	$x_t$	$p_t$
1959	60.9	479.7	104.5
1960	64.0	489.7	104.5
1961	67.9	503.8	105.1
1962	70.7	524.9	105.0
1963	74.0	542.3	104.8
1964	77.4	580.8	104.5
1965	81.6	616.3	104.0
1966	85.3	646.9	102.6
1967	98.1	673.5	102.2
1968	93.5	701.3	100.9
1969	98.4	722.5	100.0
1970	102.0	751.6	99.6
1971	106.4	779.2	100.0
1972	112.5	810.3	100.0
1973	118.2	865.3	99.1
1974	124.2	858.4	95.1
1975	128.3	875.8	93.3
1976	134.9	906.8	93.7
1977	141.3	942.9	94.5
1978	148.5	988.8	94.7
1979	154.8	1015.5	93.8
1980	159.8	1021.6	93.0
1981	164.8	1049.3	94.2
1982	167.5	1058.3	96.7
1983	171.3	1095.4	99.2

Поскольку расходы на жилье подвержены инерции и медленно согласуются с изменением доходов и цен, можно предложить иные регрессионные модели, а именно:

$$\log y_t = \alpha + \beta_1 \log x_{t-1} + \beta_2 \log p_{t-1} + u_t,$$

$$\log y_t = \alpha + \beta_1 \log x_t + \beta_2 \log p_{t-2} + u_t.$$

В этих моделях используются лаговые переменные  $x_{t-1}$ ,  $p_{t-1}$ ,  $x_t$ ,  $p_{t-2}$ , для получения значений которых нужно просто сдвинуть данные для  $x_t$ ,  $p_t$  в табл. 2.2 на один (два) уровня ниже.

В вариантах 1-12 необходимо построить:

1) логарифмическую регрессию между расходами на жилье и доходами с

запаздыванием на один год и сравнить результаты с такой же регрессией без запаздывания на тот же период;

2) логарифмическую регрессию между расходами на жилье и доходами с запаздыванием на два года и сравнить результаты с такой же регрессией без запаздывания на тот же период;

3) логарифмическую регрессию между расходами на жилье и относительной ценой с запаздыванием на один год и сравнить результаты с такой же регрессией без запаздывания на тот же период;

4) линейную регрессию между расходами на жилье и доходами с запаздыванием на один год и сравнить результаты с такой же регрессией без запаздывания на тот же период;

5) линейную регрессию между расходами на жилье и доходами с запаздыванием на два года и сравнить результаты с такой же регрессией без запаздывания на тот же период;

6) линейную регрессию между расходами на жилье и относительной ценой с запаздыванием на один год и сравнить результаты с такой же регрессией без запаздывания на тот же период;

7) линейную регрессию между расходами на жилье и относительной ценой с запаздыванием на два года и сравнить результаты с такой же регрессией без запаздывания на тот же период;

8) линейное уравнение регрессии объема расходов на жилье от располагаемых доходов, индекса цен на жилье (относительной ценой) и объемов этих расходов с лагом на один период;

9) логарифмическое уравнение регрессии объема расходов на жилье от располагаемых доходов, индекса цен на жилье (относительной ценой) и объемов этих расходов с лагом на один период;

10) логарифмическое уравнение регрессии объема расходов на жилье от располагаемых доходов, индекса цен на жилье (относительной ценой) и объемов этих расходов с лагом на два периода;

11) логарифмическое уравнение регрессии объема расходов на жилье от располагаемых доходов с лагом на один период, индекса цен на жилье (относительной ценой) и объемов этих расходов с лагом на один период;

12) линейное уравнение регрессии объема расходов на жилье от располагаемых доходов с лагом на два периода, индекса цен на жилье (относительной ценой) и объемов этих расходов с лагом на один период.

**Задача С. 1.** В табл. 2.3 приведены данные об эксплуатации самолетов: возраст самолета и стоимость. Построив линейную регрессионную модель, установите, каким образом стоимость эксплуатации самолета связана со временем, в течение которого он эксплуатируется. Проанализируйте результаты. Оцените адекватность модели, проведя анализ остатков.

В данных об эксплуатации самолетов исключите повторные наблюдения. Проанализируйте эти данные. Ответьте на следующие вопросы:

- изменилось ли уравнение регрессии;
- как изменился коэффициент детерминации;

- является ли регрессия значимой;
- адекватна ли подобранная модель исходным данным.

2. В табл. 2.4 приведены данные зависимости цены дома от его полезной площади. Стоимость дома измеряется в долларах, площадь – в квадратных футах (1 фут равен 30,48 см). Постройте линейную регрессионную модель вида  $Цена = B_0 + B_1 \cdot Площадь$ . Исследуйте зависимость влияния полезной площади дома на цену и проанализируйте результаты. Проведите анализ остатков и оцените адекватность модели.

3. В табл. 2.5 приведены данные о цене за бутылку марочного портвейна при оптовой продаже в зависимости от года, когда вино было заложено. Визуализируйте данные. Исследуйте регрессионную модель вида  $Цена = B_0 + B_1 \cdot (1972 - год)$ . Проверьте значимость регрессии, оцените адекватность модели исходным данным.

Таблица 2.3

Данные об эксплуатации самолетов		
№ п.п.	Возраст (г.)	Стоимость (у.е.)
1	4.5	619.0
2	4.5	1049.0
3	4.5	1033.0
4	4.0	495.0
5	4.0	723.0
6	4.0	681.0
7	5.0	890.0
8	5.0	1522.0
9	5.5	987.0
10	5.0	1194.0
11	0.5	163.0
12	0.5	182.0
13	6.0	764.0
14	6.0	1373.0
15	1.0	978.0
16	1.0	466.0
17	1.0	549.0

Таблица 2.4

Данные о цене и площади дома		
№ п.п.	Цена (дол, США)	Площадь (кв.ф.)
1	14650.00	1317.000
2	12850.00	1080.000
3	17900.00	1688.000
4	21977.00	1738.000
5	13900.00	1040.000
6	13100.00	1404.000
7	19750.00	1558.000
8	11600.00	976.000
9	12750.00	1077.000
10	14550.00	1315.000
11	17230.00	1620.000
12	21900.00	1720.000
13	18980.00	1438.000
14	18230.00	1412.000
15	17940.00	1650.000
16	16230.00	1590.000
17	15470.00	1499.000

Таблица 2.5

Данные о цене портвейна	
Год	Цена (у.е.)
1890	50.00
1900	35.00
1920	25.00
1931	11.98
1934	15.00
1935	13.00
1940	6.98
1941	10.00
1944	5.99
1948	8.98
1950	6.98
1952	4.99
1955	5.98
1960	4.98

# Лабораторная работа № 3

## АНАЛИЗ ВРЕМЕННЫХ РЯДОВ И ПРОГНОЗИРОВАНИЕ

*Цель работы* - научиться обрабатывать временные ряды в рамках пакета *Statistica*, делать прогноз значения наблюдаемого показателя на определенный период.

### 3.1. Элементы диалога в модуле *Time Series Analysis/Forecasting* (Анализ временных рядов/Прогнозирование)

Временными рядами называют ряды измерений, зависящие от дискретного или непрерывного аргумента (времени), подвергающиеся нерегулярным случайным воздействиям или флуктуациям и допускающие только статическое описание.

Общее назначение модуля – построить простую модель, описывающую временной ряд, сгладить ее, спрогнозировать будущие значения временного ряда на основе наблюдаемых значений до данного момента, построить регрессионные зависимости одного ряда от другого, провести спектральный или Фурье– анализ ряда и т.д.

Выберите модуль *Time Series Analysis/Forecasting* (Анализ временных рядов/Прогнозирование) в переключателе модулей, высветив его и щелкнув **OK**, либо дважды щелкнув по его пиктограмме. Перед вами появится стартовая панель этого модуля, имеющая следующую структуру.

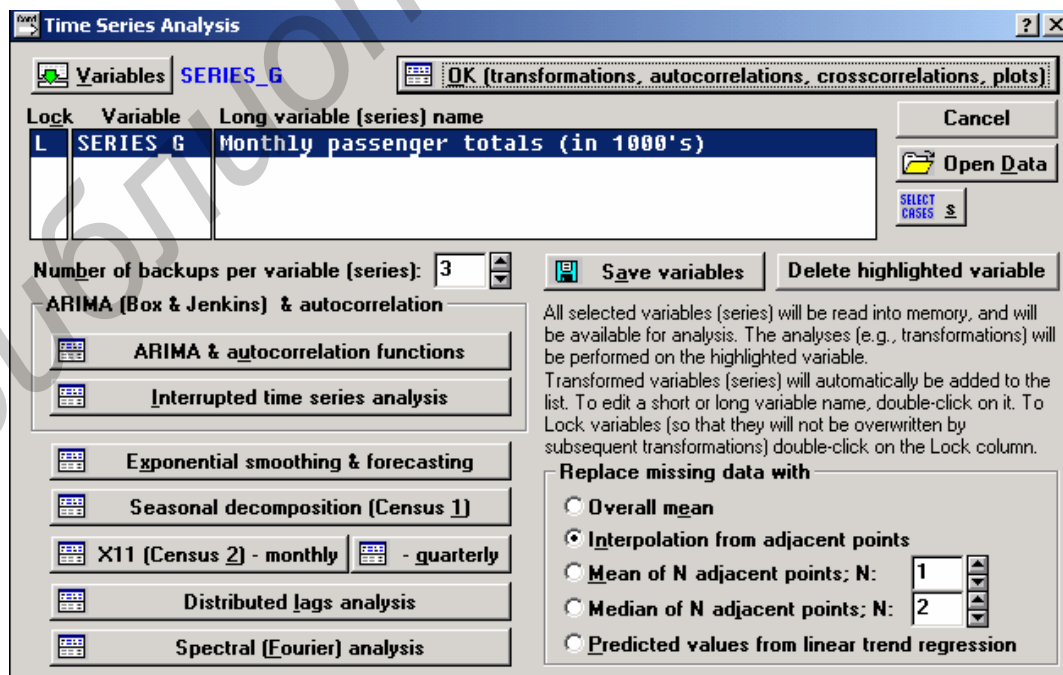


Рис. 3.1. Стартовая панель модуля *Time Series Analysis/Forecasting* (Анализ временных рядов/Прогнозирование)

В нижней функциональной части панели находятся функциональные кнопки, открывающие различные методы обработки временных рядов.

Кнопка **Open data** (*Открыть данные*) (в верхнем правом углу панели) открывает файл данных для обработки.

С помощью кнопки **Variables** (*Переменные*) (в левом верхнем углу) открывается диалоговое окно выбора переменных из открытого файла данных. Имена и расширенные имена выбранных переменных появляются в верхней, информационной части окна.

Нажав кнопку **Delete highlighted variable** (*Удалить высвеченные переменные*) из диалога удаляются переменные, высвеченные в информационной части панели. Расположенная рядом кнопка **Save variables** (*Сохранить переменные*) позволяет сохранить высвеченные переменные в файле данных системы. Потребность в данной кнопке возникает, например, когда производятся преобразования временного ряда и какие-то преобразования нужно сохранить. Последовательность действий здесь следующая: вы высвечиваете имя соответствующей переменной в информационном окне и нажимаете соответствующую кнопку.

Верхняя кнопка **OK** (*Transformations, correlations, crosscorrelations, plots*) — *Да* (*Преобразования, корреляции, кросскорреляции, графики*) позволяет преобразовать ряд, вычислить корреляции и автокорреляции, построить графики. Опции в правом нижнем углу задают способы обработки пропущенных значений ряда.

**STATISTICA** предлагает следующие методы для анализа временных рядов (см. функциональные кнопки, открывающие эти возможности, в левой части стартовой панели, представленные на рис. 3.1).

**ARIMA** - метод авторегрессии и проинтегрированного скользящего среднего;

**Interrupted time series analysis** - анализ прерванного временного ряда (модели интервенции для ARIMA);

**Exponential smoothing & forecasting** - экспоненциальное сглаживание и прогнозирование;

**Seasonal decomposition (1, 2)** - сезонная декомпозиция 1 и 2 (в том числе месячная и квартальная);

**Distributed lags analysis** - анализ распределенных лагов (регрессионная модель для двух временных рядов);

**Spectral (Fourier) analysis** — спектральный (Фурье) анализ.

Термин **ARIMA** часто не переводится, иногда вместо него будем просто говорить «авторегрессия». Это важный класс параметрических моделей, который широко применяется в статистике. В системе **STATISTICA ARIMA** реализованы методологии Бокса и Дженкинса.

Как уже говорилось, в **STATISTICA** имеется возможность проводить анализ прерванной авторегрессии, рассматривать так называемые модели **ARIMA** с интервенцией — **Interrupted time series analysis** (*Анализ прерванных временных рядов*). Необходимость в такого рода анализе возникает, когда с некоторого момента резко изменяются условия реализации наблюдаемого ряда

(изменились правила торгов на бирже, ускорилось биение сердца больного после приема лекарства).

Внешнее воздействие на ряд, вызываемое разными причинами, может быть как короткотечным, импульсивным, так и длительным, устойчивым. В момент воздействия временной ряд резко меняется, но далее вновь описывается авторегрессионной моделью.

В **STATISTICA** предлагаются различные методы внешнего воздействия.

Методы **Exponential smoothing & forecasting** — экспоненциальное сглаживание и прогнозирование — позволяют сгладить наблюдаемый ряд, выделить из него шум и спрогнозировать будущие значения. Различные виды трендов и сезонность могут быть учтены в модели. Доступны модели с аддитивным и мультипликативным шумами.

Методы **Seasonal decomposition (1,2)** — сезонная декомпозиция 1 и 2 — позволяют анализировать аддитивные модели временных рядов, когда временной ряд  $X_t$  представляется в виде

$$X_t = F(t) + S(t) + E(t), t = 0, 1, 2, \dots,$$

где  $F(t)$  — тренд (детерминированная функция);  $S(t)$  — сезонная составляющая;  $E(t)$  — случайная составляющая, а также мультипликативные модели временных рядов, в которых случайная составляющая входит множителем:

$$X_t = F(t) \cdot S(t) \cdot E(t).$$

В модели может быть добавлена циклическая составляющая, отличающаяся от сезонной большим периодом.

Методы **Distributed lags analysis** (Анализ распределенных лагов) позволяют построить регрессию одного ряда на другой. Это важно, например, в том случае, когда вы хотите предсказать значения зависимого ряда на основе измерений со сдвигом независимого ряда, как это часто бывает в экономике, где одни измерения опережают другие. Иногда такой анализ называют анализом распределенных запаздываний, так как один ряд запаздывает относительно другого. Распределяя запаздывания независимого ряда и приписывая им разные веса, возможно как можно точнее приблизить значения зависимого временного ряда.

### 3.2. Диалог **ARIMA**

*Постановка задачи* : предположим, мы имеем данные о количестве общих авиаперевозок в заданном месяце. Необходимо подобрать к ряду авторегрессионную модель, оценить ее параметры и на основе наблюдаемых значений спрогнозировать перевозки пассажиров на несколько месяцев вперед.

Для работы загрузите опцию **ARIMA & autocorrelation functions** (**ARIMA и автокорреляционные функции**) в стартовой панели модуля **Анализа временных рядов**. Затем откройте эти данные из файла *Series\_g.sta*, который поставляется

вместе с пакетом *STATISTICA* и находится в папке *Examples*. Все месяцы пронумерованы от 1 до 144.

Подгонка авторегрессионной модели проводится в несколько этапов, которые будут последовательно описаны ниже.

### Этап 1 – приведение ряда к стационарному

Прежде чем подогнать к временному ряду авторегрессионную модель, его следует сделать стационарным. Вначале необходимо просмотреть данные графически. Для этого в правой части стартовой панели *ARIMA* расположена группа функциональных кнопок *Review and plot variables* (*Просмотреть переменные и построить график*). Воспользуйтесь этими кнопками. Иницируйте вначале верхнюю кнопку *Plot* (*График*).

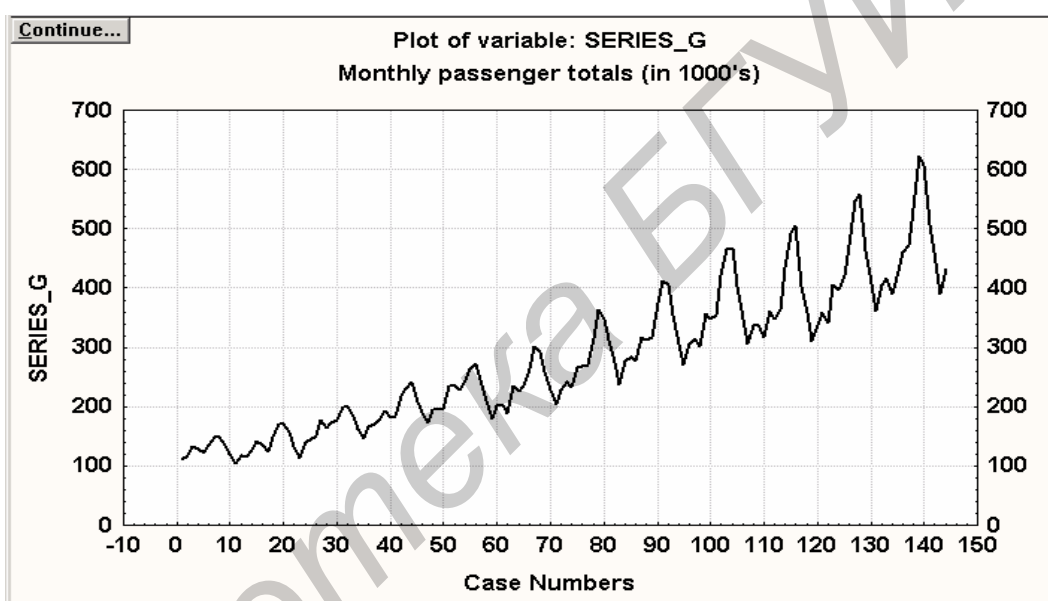


Рис. 3.2. Графическое представление исходных данных временного ряда

Просмотрев ряд на графике (рис. 3.2), вернитесь в стартовое окно *ARIMA*, щелкнув по кнопке *Continue* (*Продолжить*) в левом верхнем углу графика. Наблюдаемый ряд не является стационарным, как это хорошо видно на графике. Мы будем последовательно преобразовывать ряд, делая его раз за разом все более похожим на стационарный.

Существуют различные способы преобразования, позволяющие сделать ряд более стационарным. Модель *ARIMA* является наиболее простой из возможных моделей. Она представляет значения ряда, наблюдаемые в данный момент, в виде конечной линейной комбинации предыдущих значений самого ряда (отсюда сам термин — авторегрессия) и линейной комбинации значений временного ряда с независимыми значениями (отсюда термин — скользящее среднее). Формальные определения моделей следующие:



**Модель авторегрессии (АР) порядка p:**

$$\begin{aligned}x_n &= f_1 x_{n-1} + \dots + f_p x_{n-p} + a_n, \quad n \geq q, \\x_0 &= a_0, \\x_1 &= f_1 x_0 + a_1, \\&\dots \dots \dots \dots \dots \dots \dots \\x_p &= f_1 x_{p-1} + \dots + f_p a_0 + a_p,\end{aligned}$$

где  $f_1, \dots, f_p$  – параметры модели.

**Модель скользящего среднего (СС) порядка q** определяется как:

$$\begin{aligned}x_n &= a_n - t_1 a_{n-1} - \dots - t_q a_{n-q}, \quad n \geq q, \\x_0 &= a_0, \\x_1 &= a_1 - t_1 a_1, \\&\dots \dots \dots \dots \dots \dots \dots \\x_q &= a_q - t_1 a_{q-1} - \dots - t_q a_0,\end{aligned}$$

где  $t_1, \dots, t_q$  – константы (параметры модели),  $s^2$  – дисперсия белого шума, которая также является параметром модели.

**Модель авторегрессии и скользящего среднего (АРСС) :**

$$x_n = f_1 x_{n-1} + \dots + f_p x_{n-p} + a_n - t_1 a_{n-1} - \dots - t_q a_{n-q}.$$

Несмотря на простоту, это очень мощная модель временных рядов, позволяющая часто с достаточной степенью точности прогнозировать новые значения ряда.

Различные способы преобразования ряда доступны с помощью кнопки **Other transformations & plots** (*Другие преобразования и графики*) на панели инструментов **ARIMA**. Иницируйте эту кнопку. Перед вами откроется окно **Transformation of Variables** (*Преобразования переменных*). 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 показывать вам график преобразованных данных после каждого преобразования ряда.

Обратите внимание также на следующие кнопки на стартовой панели **ARIMA**:

**Autocorrelations** — Автокорреляции;

**Partial autocorrelations** — Частные автокорреляции.

Инициировав эти кнопки, вы можете просмотреть автокорреляции и частные автокорреляции преобразованного ряда.

Автокорреляции и частные автокорреляции — это те характеристики, которые позволяют определить, насколько близок ряд к стационарному, а также сориентироваться в преобразованиях, которые следует сделать, чтобы привести ряд к стационарному. После каждого шага преобразований целесообразно инициировать эти кнопки, чтобы определить направление своих дальнейших

действий. Существует простая методика, доступная практически каждому, позволяющая по автокорреляциям и частным автокорреляциям судить о близости ряда к стационарному.

После щелчка по кнопке *Continue* (*Продолжить*) в углу графика вы вновь вернетесь в окно *Transformation of variables* (*Преобразование переменных*). На экране появится окно *Time Series Transformations* (*Преобразования временного ряда*). В появившемся окне можно выбрать различные способы преобразования временного ряда. Выберите какую-нибудь опцию, затем нажмите кнопку **OK** (*Transform*) — *Да* (*Преобразовать*) в правом верхнем углу. Программа произведет выбранное вами преобразование. На каждом шаге можно выполнить *только одно преобразование значений высвеченной переменной*.

Все преобразования переменных программа запишет в информационной части диалогового окна. Чтобы выполнить несколько преобразований, следует повторить эти действия несколько раз. По умолчанию, в системе каждый раз будет высвечиваться переменная, преобразованная на последнем шаге.

В окне *Преобразования временного ряда* имеется несколько групп преобразований, объединенных по функциональному признаку. Эти преобразования можно применять последовательно, одно за другим. Конечная цель всех преобразований — привести ряд к удобному для анализа виду.

В левом верхнем углу даны преобразования, которые можно получить применением к ряду некоторых стандартных функций, — новый временной ряд получается преобразованием  $f: x_{\text{новый}} = f(x_{\text{старый}})$ .

Возможны следующие преобразования ряда:

- *Add a constant* — прибавить константу к значениям ряда;
- *Power* — возвести в степень;
- *Inverse power* — возвести в обратную степень;
- *Natural log* — взять натуральный логарифм;
- *Exponent* — выполнить экспоненциальное преобразование.

Функциональная кнопка *Transformations for spectrum analysis* (*Преобразования для спектрального анализа*) выполнит преобразования ряда, необходимые для выполнения спектрального анализа.

В правой части окна имеется группа опций *Two-series transformations* (*Преобразования двух временных рядов*). Пусть вы анализируете пару временных рядов  $X$  и  $Y$ . Выбрав опцию *Difference* (*Разность*), вы вычислите новое значение ряда  $X$  в точках  $t$  на временной оси по формуле

$$X_t = X_t - Y(t - lag),$$

где значения сдвига (запаздывания или лага) задаются в поле *lag* (сдвиг).

Выбрав опцию *Residualizing* (*Остаточный*), вы вычислите новые значения ряда по формуле

$$X_t = X_t - (a + b \cdot Y(t - lag)),$$

где параметры  $a$  и  $b$  могут быть либо заданы, либо оценены из данных.

В последнем случае следует выбрать опцию *Estimate a and b from data* (Оценить  $a$  и  $b$  из данных).

В группе опций, расположенных ниже, *Shift relative starting point of series* (Сдвинуть относительную начальную точку ряда) предлагаются возможности сдвинуть ряд вперед и назад.

Опции *Filtering and other techniques* (Фильтрация и другие методы) дают возможность для фильтрации ряда и осуществления следующих преобразований ряда:

*Differencing* (Вычисление последовательных разностей) - новые значения ряда  $x_t$ , вычисляются из старых по формуле  $x_t = x_t - x(t - lag)$ ;

*Integrate* (Суммировать) - новые значения ряда  $x_t$  вычисляются из старых по формуле  $x_t = x_t + x(t - lag)$ .

В окне *Преобразования временного ряда* имеются также следующие опции: выделение среднего, стандартизация (нормировка), выделение тренда, сглаживание – скользящее среднее, медиана, простое экспоненциальное сглаживание.

После того, как разобраны опции окна *Преобразования временного ряда*, вернемся вновь к ряду *SERIES\_G*. Наша ближайшая цель — сделать ряд стационарным. Тогда к нему можно подобрать авторегрессионную модель. Будем последовательно применять несколько преобразований к временному ряду месячных авиаперевозок. Для уменьшения амплитуд колебаний временных рядов часто используют логарифмическое преобразование. Применим преобразование *Natural log* (Натуральный логарифм).

После этого можно просмотреть график преобразованных данных, нажав кнопку *Plot* (График) на стартовой панели *ARIMA*. На графике видно, что после преобразования ряда его колебания существенно уменьшились.

Далее возьмем первую разность ряда, опцию которой можно найти в правом нижнем углу окна *Time Series Transformations* (Преобразования временного ряда). Щелкните по кнопке *OK (Transform)* — Да (Преобразовать). В появившемся окне *Time Series Transformations* (Преобразования временного ряда) выберите преобразование *Differencing* ( $x = x - x(lag)$ ), положив  $lag = 1$ . Смысл этого преобразования в том, что из текущего значения ряда вычитается предыдущее (со сдвигом 1) и результат представляется в качестве значения нового ряда.

Преобразовывая ряд с помощью разности первого порядка, мы избавляемся от линейного тренда в ряде.

Так как ряд имеет сезонную составляющую, следует взять сезонную разность. Для этого выберите снова преобразование *Differencing* ( $x = x - x(lag)$ ) и в качестве лага (сдвига) задайте 12,  $lag = 12$  – это так называемый сезонный сдвиг. Данное преобразование исключает сезонную составляющую с периодом 12. Щелкните по кнопке *OK* (Да).

Щелкните далее по кнопке *Plot* (График), после сделанных преобразований — логарифмирования, разность с лагом 1, разность с лагом 12, вы увидите на рисунке ряд, который очень похож на стационарный (рис. 3.3).

Закончите преобразования ряда.

Нажав в окне *Transformation of variables* (Преобразование переменных) кнопку *Exit* (Выход), вы вновь окажетесь в окне *ARIMA*.

Вернемся вновь в стартовую панель *Single Series ARIMA* (Единичная *ARIMA*) — это можно сделать, например, щелкнув по кнопке *Continue* (Продолжить) на графике, далее закрыть появившееся окно *Преобразование переменных* кнопкой *Exit* (Выход). После нажатия этих кнопок система автоматически вернется в нужное окно.

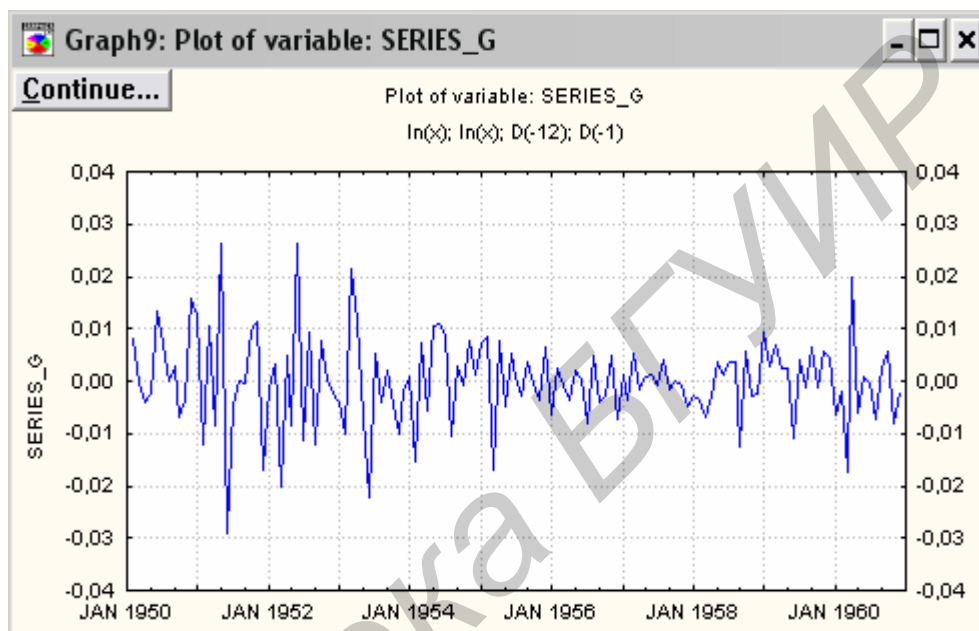


Рис. 3.3. График месячных перевозок пассажиров после применения трех преобразований: *Ln*, *Differencing* с лагом 1, *Differencing* с лагом 12

## Этап 2 – идентификация или подгонка *ARIMA* к временному ряду

В стартовой панели *Single Series ARIMA* (Единичная *ARIMA*) произведем идентификацию модели, т.е. определим, какие параметры и какое их количество должно присутствовать в модели.

В *ARIMA* имеется всего 4 типа параметров ( $p$ ,  $P$ ,  $q$ ,  $Q$ ), которые надо определить. В левой верхней части панели *Single Series ARIMA* (Единичная *ARIMA*) есть специальная группа опций, позволяющих идентифицировать модель. Эта группа называется *Arima model parametr*s (Параметры *ARIMA*). В группе имеются следующие опции:

$p$  – *Autoregressive* (параметр авторегрессии),

$P$  – *Seasonal* (сезонный параметр авторегрессии),

$q$  – *Motiving average* (параметр скользящего среднего),

$Q$  – *Seasonal* (сезонный параметр скользящего среднего).

В строке возле каждого параметра задается число параметров данного типа. По крайней мере один из этих параметров должен присутствовать в модели. Рекомендуется не брать большое количество параметров, а вначале поработать с малым количеством параметров.

В примере с перевозками пассажиров в диалоговом окне *Single Series*

**ARIMA** сделаем следующие установки:

$$p = P = 0$$
$$q = Q = 1.$$

В окне **Transform variable (series) prior to analysis** (*Преобразование переменной (ряда) до анализа*) пометьте нужные опции, указывая преобразования, ранее производимые с исходным рядом: **Natural Log** (*Натуральное логарифмирование*) и **Difference** (*Разность*) с лагом 1 и лагом 12. В "технических" строках **No. of passes** задайте 1. Далее выберите метод оценивания неизвестных параметров  $q$  и  $Q$ .

### Этап 3 – оценивание неизвестных параметров **ARIMA**

Метод оценивания выбирается в левой нижней части окна с заголовком **Estimation of Maximum Likelihood** (*Оценивание методом максимального правдоподобия*). Система предлагает две вычислительные процедуры, реализующие метод максимального правдоподобия:

**Approximate** (*Приближенный*);

**Exact** (*Точный*).

Выберите, например, метод **Exact** (*Точный*). В окне **Variables** (*Переменные*) в верхней части диалогового окна не забудьте высветить первую строку — исходный временной ряд. Отметим, что вы могли бы сразу задать преобразования данных в этой части окна, а не производить преобразование данных в диалоговом режиме, как было сделано до этого.

Щелчком кнопки **OK** (**Begin parameter estimation**) — *Да (Начать оценку параметров)* запустите процедуру оценивания. В окне появится строка **Parameter estimation process converged** (*Процесс оценки параметров сошелся*). Если оценки параметров вас не удовлетворяют, например слишком велика функция потерь, значения которой даны в первой колонке, щелкните по кнопке **Cancel** (*Отменить*) в правом нижнем углу.

Вернитесь в стартовую панель **ARIMA**, измените метод оценивания, его начальные установки либо вернитесь к фазе идентификации модели и снова идентифицируйте модель, т.е. задайте новые параметры  $p$ ,  $P$ ,  $q$ ,  $Q$ . Если же вы находите оценки приемлемыми, щелкните по кнопке **OK** и всесторонне просмотрите результаты. Окно **Single Series ARIMA Results** (*Результаты ARIMA*) откроется перед вами:

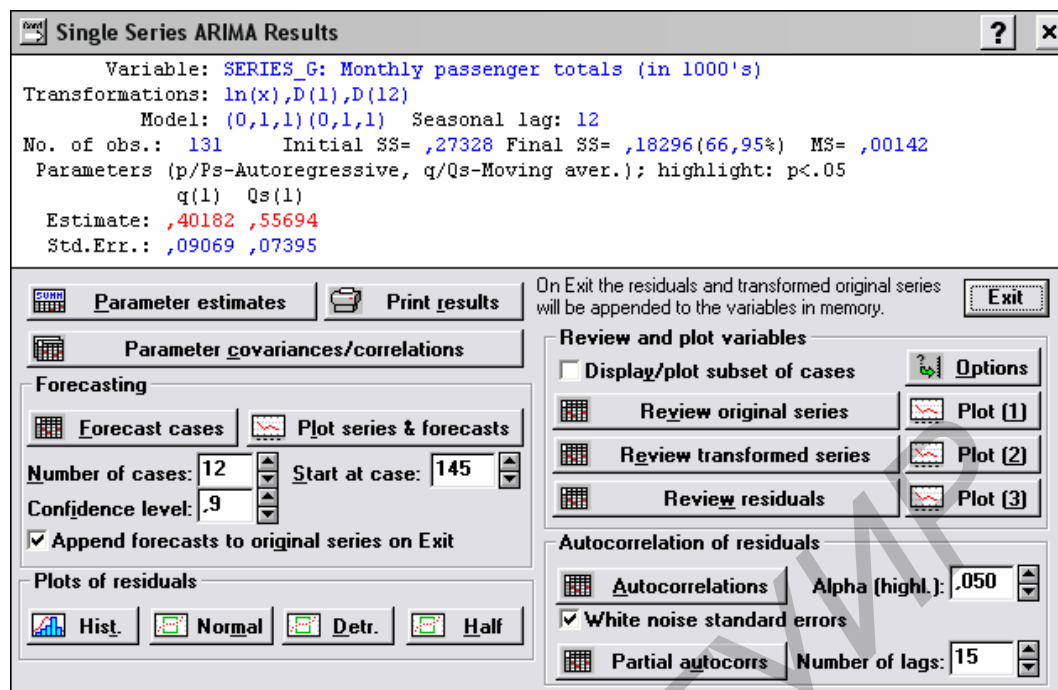


Рис. 3.4. Окно *Результаты ARIMA*

В данном окне вы можете просмотреть результаты оценивания и оценить качество подгонки. В верхней части окна дана информация о модели и оценках ее параметров. Просмотрите численные оценки, щелкнув по кнопке *Parameter estimates* (*Оценки параметров*). Перед вами появится следующая таблица:

Paramet.	Param.	Asympt. Std. Err.	Asympt. t(129)	p	Lower 95% Conf	Upper 95% Conf
q(1)	,401823	,090686	4,430945	,000020	,222399	,581247
Qs(1)	,556937	,073949	7,531376	,000000	,410627	,703246

Рис. 3.5. Таблица оценок параметров *ARIMA*

В первой колонке этой таблицы — оценки параметров, во второй — асимптотическая стандартная ошибка, в третьей — значения t-критерия, в четвертой — уровни значимости, в пятой и шестой — соответственно верхние и нижние границы 95%-ных доверительных интервалов для соответствующих неизвестных параметров модели. Например, из первой строчки вы видите, что интервал (0,222399, 0,581247) с вероятностью 0,95 покрывает истинное значение параметра  $q(1)$ . Число 0,401823, приведенное в первой колонке, есть **точечная оценка** неизвестного параметра  $q(1)$ .

#### Этап 4 – оценка качества модели или степени ее адекватности данным

В окне *Результаты ARIMA* посмотрим графики остатков ряда. Последние представляют собой разности наблюдаемого временного ряда и

значений, вычисленных на модели (модельных) значений. Анализ остатков позволяет сделать заключения о качестве подогнанной модели. **STATISTICA** позволяет всесторонне проанализировать остатки между наблюдаемыми и предсказанными значениями.

Обратите внимание на группу кнопок в левом нижнем углу под заголовком **Plot of residuals** (*Графики остатков*). Щелкните, например, по кнопке **Normal plot** (*Графики на нормальной вероятностной бумаге*) – это график остатков, построенный на нормальной вероятностной бумаге. Графики остатков чрезвычайно важны при определении степени адекватности подогнанной модели данным.

Вся общая методология для подгонки регрессионных моделей сохраняется и при подгонке моделей **ARIMA**. Полезно просмотреть гистограмму остатков и увидеть, насколько их распределение согласуется с нормальным распределением. Щелкнем по кнопке **Histogram** (*Гистограмма*) – это гистограмма значений с наложенной нормальной плотностью. Из графика видно, что распределение остатков похоже на нормальное.

"Есть ли зависимость между остатками?" — вот тот вопрос, который естественно задать далее. Для ответа на него воспользуйтесь группой кнопок, находящихся в правом нижнем углу окна просмотра результатов. Эта группа кнопок называется **Autocorrelations of residuals** — *Автокорреляции остатков*. Щелкните, например, по кнопке **Autocorrelations** — *Автокорреляции*, вы увидите, что остатки практически некоррелированы. То же самое можно увидеть, если инициировать кнопку **Partial autocorrelations** (*Частные автокорреляции*).

Итак, мы видим, что модель достаточно адекватно описывает наблюдаемый временной ряд.

Оценка качества модели или степени ее адекватности данным — важный этап подгонки модели **ARIMA**. Без оценки адекватности вы не можете доверять прогнозу, построенному с помощью модели. Если модель не адекватна данным, прогноз с помощью нее не имеет смысла.

Теперь, после выполнения этих этапов, можно приступить к построению прогноза.

В левой части окна **Результаты ARIMA** имеется набор опций **Forecasting** (*Прогнозирование*). Эти опции важны, и мы опишем их детально. В строке **Number of cases** (*Число случаев*) задайте число случаев, на которое вы хотите спрогнозировать ряд, например, если на два года вперед, то задайте число 24. Можно задать **Confidence level** (*Уровень доверия*) или коэффициент доверия, измеряющих надежность прогнозируемых значений ряда. Обычно в качестве уровня доверия берут 0,9 либо 0,95. Однако бывают ситуации, когда целесообразно брать существенно меньшие значения уровня доверия, например 0,7.

В строке **Start at case** (*Начать со случая*) указывается номер случая, с которого следует начать прогноз. Выберите опцию **Append forecasts to original series on Exit** (*Добавьте прогнозируемые величины к концу наблюдаемого ряда*). Кнопка **Forecast cases** (*Прогнозируемые случаи*) позволяет просмотреть прогноз

в электронной таблице на несколько случаев (месяцев) вперед.

Щелкните по кнопке *Plot series&forecasts* (*График ряда и прогнозы*), вы увидите то, к чему мы стремились, — график ряда с прогнозируемыми значениями на 24 месяца вперед (рис.3.6).

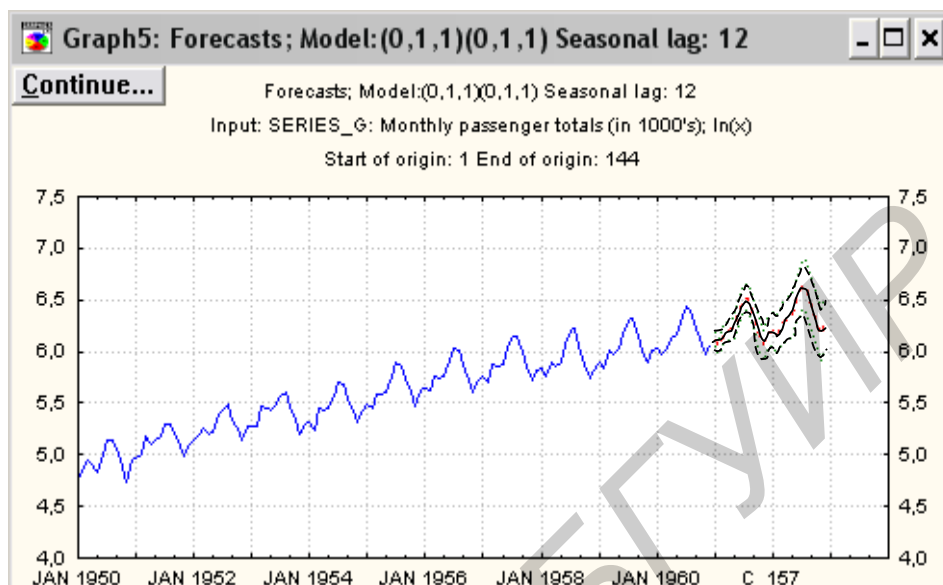


Рис. 3.6. График авиаперевозок с построенным прогнозом

Используя этот прогноз, вы можете строить стратегию своих действий на будущее, например добиваться лучшей загрузки самолетного парка, рассчитывать оптимальную цену билетов при предварительных продажах и т.д.

### 3.3. Экспоненциальное сглаживание и прогнозирование

Экспоненциальное сглаживание является существенно более простым методом, чем *ARIMA*, и в ряде случаев позволяет строить приемлемые прогнозы наблюдаемых временных рядов. Суть метода состоит в том, что исходный ряд  $x(t)$  сглаживается и образуется новый временной ряд  $y(t)$ , поведение которого можно прогнозировать. Простое экспоненциальное сглаживание описывается формулой

$$y(t) = \alpha \cdot x(t) + (1 - \alpha) \cdot x(t - 1),$$

где  $\alpha$  – некоторый фиксированный параметр,  $0 < \alpha < 1$ .

В стартовой панели модуля *Анализ временных рядов/Прогнозирование* инициализируйте кнопку *Exponential Smoothing&Forecasting* (*Экспоненциальное сглаживание и прогнозирование*). На экране появится окно *Seasonal and Non-Seasonal Exponential Smoothing* (*Сезонное и несезонное экспоненциальное сглаживание*) (рис. 3.7).



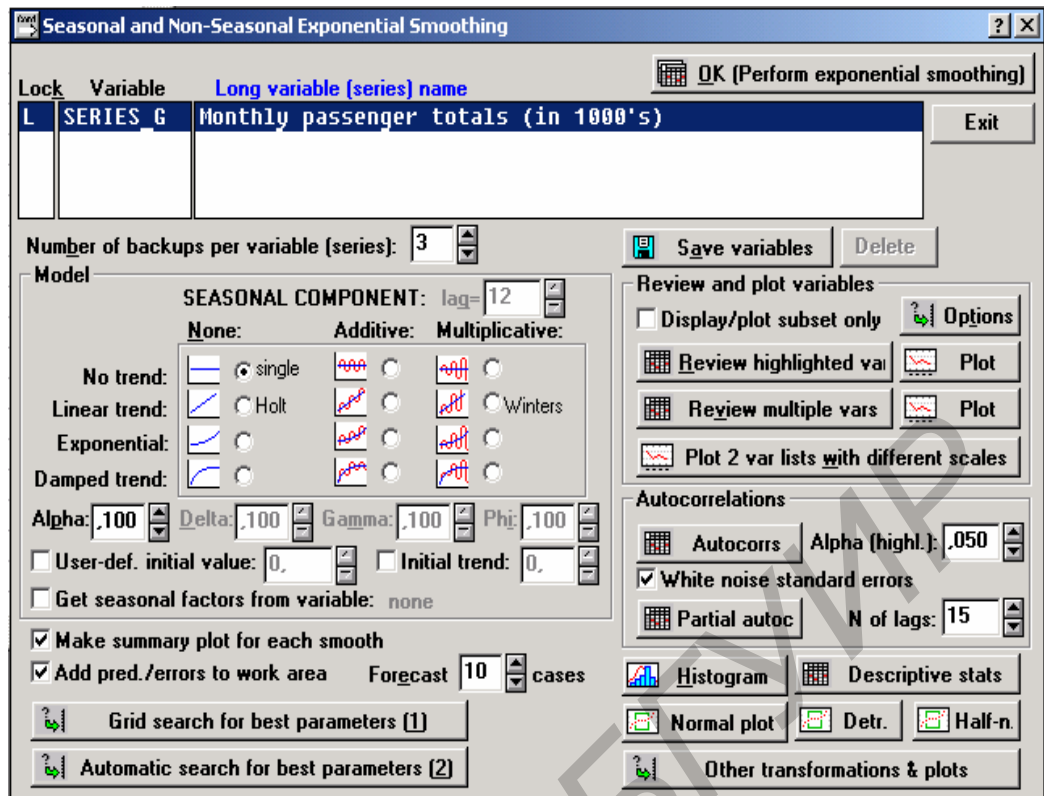


Рис. 3.7. Окно модуля *Сезонное и несезонное экспоненциальное сглаживание*

Вы знаете, что ряд имеет аддитивную сезонную составляющую с лагом 12, поэтому задайте в строке *Seasonal component* (*Сезонная компонента*) лаг 12. Далее в столбце *Additive* (*Аддитивный*) выберите линейный тренд. Как только вы задали качественные характеристики, описывающие поведение ряда, система сама предложит на выбор 3 параметра – *Alpha*, *Delta*, *Gamma*, которые нужно определить. Две кнопки внизу панели служат для этого: *Grid Search for best parameters* (*Поиск на сетке лучших параметров*) и *Automatic search for best parameters* (*Автоматический поиск лучших параметров*).

Щёлкните по кнопке *Поиск лучших параметров*, в появившемся окне задаются параметры сетки, на которой происходит поиск, и начальные значения неизвестных параметров (рис. 3.8).

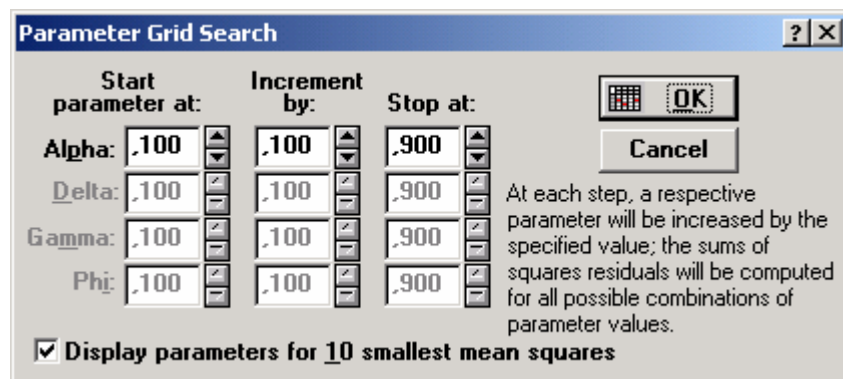


Рис. 3.8. Окно задания параметров сетки

Система переберёт все значения на сетке и определит лучшие значения параметров. Установите все значения  $Alpha$ ,  $Delta$ ,  $Gamma$ , равные 0,1, и параметра *Stop at* (*Остановиться на*), равного 0,9. Щёлкните по кнопке **OK**, на экране появится таблица, в которой в самой верхней строке даны лучшие значения для рассматриваемого примера:  $alpha = 0,2$   $delta = 0,9$ ,  $gamma = 0,1$ .

Щёлкнув по кнопке *Continue* (*Продолжить*), вернитесь в окно *Сезонное и несезонное экспоненциальное сглаживание*. Щёлкните по кнопке **OK** (*Perform exponential smoothing*) – *Да* (*Выполнить экспоненциальное сглаживание*) и на экране последовательно появятся таблица и график. Визуально прогноз похож на тот, что получался при использовании модели *ARIMA* (рис. 3.9). Если сравнить их численно, то мы увидим, что в прогнозах есть отличия. Если взять верхние и нижние границы 90% доверительных интервалов для модели *ARIMA*, то прогноз по методу экспоненциального сглаживания попадает в этот интервал.

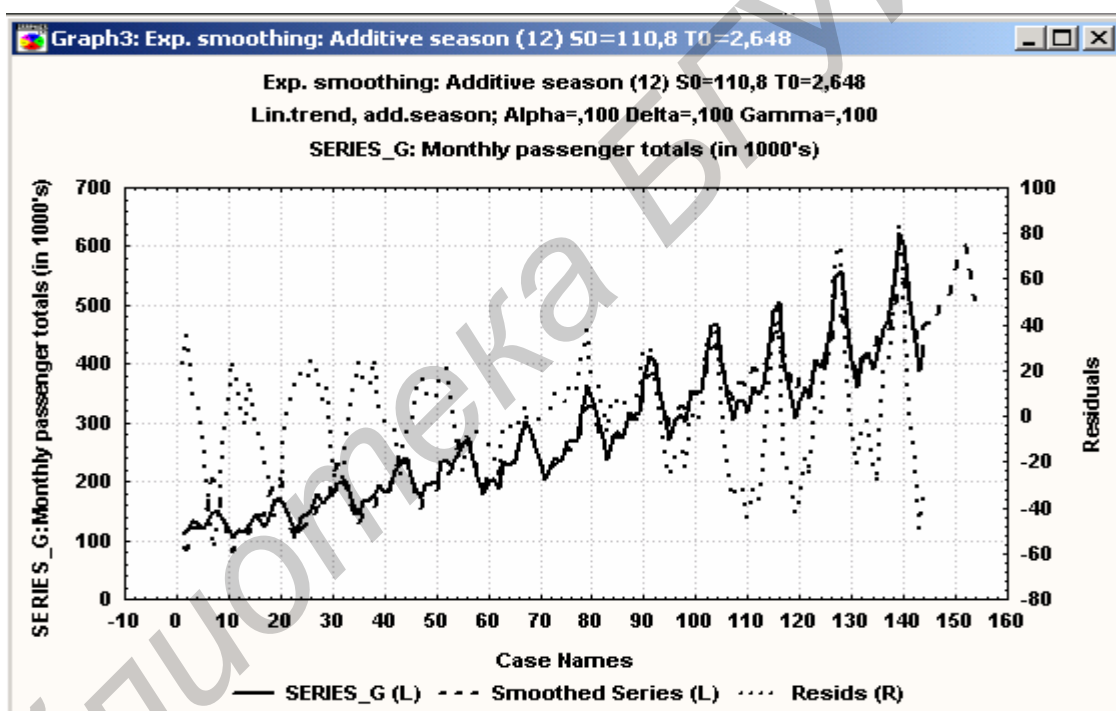


Рис. 3.9. График прогноза при экспоненциальном сглаживании

Модель можно проверить на автокорреляцию остатков, нажав на кнопку *Autocorrelations* (*Автокорреляции*), после чего появится график автокорреляционной функции.

### 3.4. ВАРИАНТЫ ЗАДАНИЙ

На основании имеющихся данных курса акций компании *IBM* за 160 дней (прил. 1) сделать прогноз на  $i$  дней вперед, используя:

- модель *ARIMA* (авторегрессии - скользящего среднего);
  - экспоненциальное сглаживание;
- и сравнить полученные результаты.

Предусмотреть автоматическое создание отчета по результатам проводимого анализа.

1. Прогноз на 100 дней. Исходные данные находятся в варианте 1.
2. Прогноз на 50 дней. Исходные данные находятся в варианте 2.
3. Прогноз на 20 дней. Исходные данные находятся в варианте 3.
4. Прогноз на 30 дней. Исходные данные находятся в варианте 4.
5. Прогноз на 60 дней. Исходные данные находятся в варианте 5.
6. Прогноз на 120 дней. Исходные данные находятся в варианте 6.
7. Прогноз на 40 дней. Исходные данные находятся в варианте 7.
8. Прогноз на 70 дней. Исходные данные находятся в варианте 8.
9. Прогноз на 15 дней. Исходные данные находятся в варианте 9.
10. Прогноз на 20 дней. Исходные данные находятся в варианте 10.
11. Прогноз на 65 дней. Исходные данные находятся в варианте 11.
12. Прогноз на 35 дней. Исходные данные находятся в варианте 12.

Данные вариантов заданий следует представить в файле с исходными данными (файл с расширением *sta*) в виде столбца (*variable*) со 160-ю случаями (*cases*).

# Лабораторная работа № 4

## СТАТИСТИЧЕСКИЙ КОНТРОЛЬ ПРОИЗВОДСТВЕННОГО ПРОЦЕССА

*Цель работы* – провести анализ качества выполнения производственного процесса.

### 4.1. Постановка задачи

Необходимо осуществить контроль качества выполнения производственного процесса предприятия, производящего агрегаты. Контролируется точность выполнения отверстий в лопастях агрегатов. Для реализации задачи взять наблюдение за ходом выполнения процесса под статистический контроль. По результатам такого контроля сделать выводы о пригодности производственного процесса.

Для решения поставленной задачи следует использовать  $X$ - и  $R$ -контрольные карты.  $X$ -карта, или карта средних, применяется для отображения количественных характеристик процесса на основе выборочных средних. На  $X$ -картах также строятся контрольные границы, позволяющие определить, когда средние выходят за допустимый уровень.  $R$ -карта, или карта размахов, предназначена для контроля изменчивости процесса. Мерой изменчивости является выборочный размах или разность между максимальным и минимальным значениями выборки.

### 4.2. Диалог модуля «Контроль качества»

Для начала работы необходимо запустить модуль *Quality Control* (Контроль качества). Предположим, что исходные данные представляют собой результаты 100 измерений диаметра отверстий, которые находятся в файле *Vane.sta* (рис. 4.1). Переменная названа *Vane* (лопасть).

Перед началом рассмотрим описательные статистики и визуализируем данные. Для этого в файле данных выделим переменную *Vane* и щелкнем по ней правой кнопкой мыши. В появившемся списке опций выберем *Быстрые описательные статистики* | *Описательные для VANE*. Выберем в меню *Graphs* | *Stats 2D Graphs* | *Scatterplots* (Графики | Стат 2М графики | Диаграммы рассеяния). В качестве переменной  $X$  выберем *Vane*, в качестве  $Y$  – *num* и нажмем кнопку *OK*. Появится окно с диаграммой рассеяния, показанное на рис. 4.2.

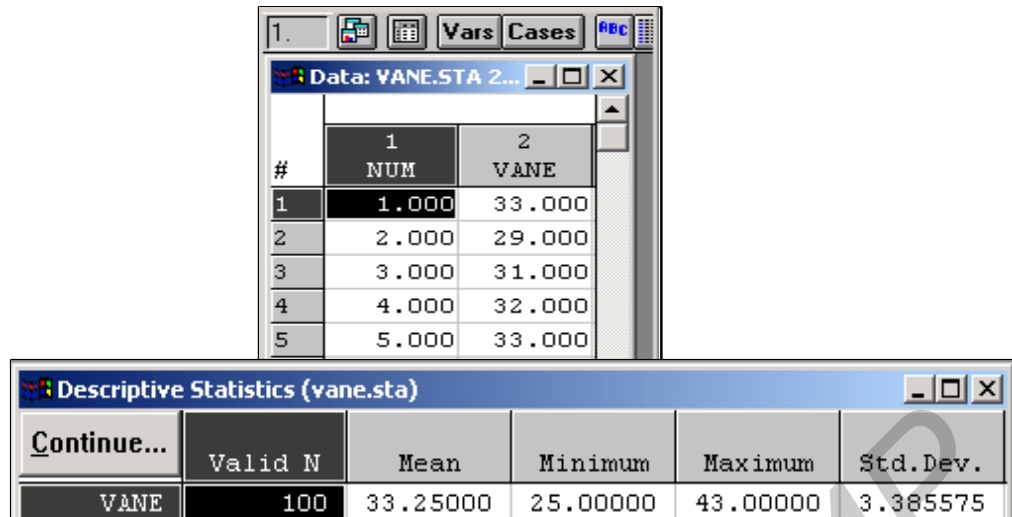


Рис. 4.1. Файл данных и описательные статистики для переменной *Vane*

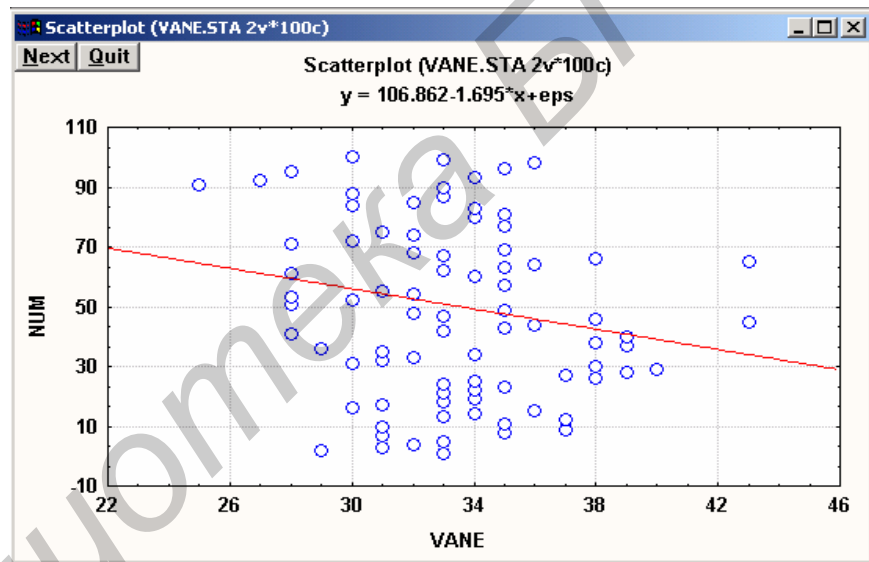


Рис. 4.2. Диаграмма рассеяния для переменной *Vane*

Запустим модуль *Quality Control Charts* (Карты контроля качества). На стартовой панели выберем *Charts for variables; Means (X-bar) & variability (R, S, SI)* (Карты для переменных; Средние (X-карта) и переменные (R, S, SI)) и нажмем кнопку **OK**. В появившемся диалоговом окне выберем *R-Charts*. В окне *Defining Variables for Range Chart* (Определение переменных для карты размаха) в качестве переменной с измерениями выберем *vane*.

Нажмем кнопку **OK**, появятся графики, содержащие X- и R- карты (рис. 4.3 и 4.4).

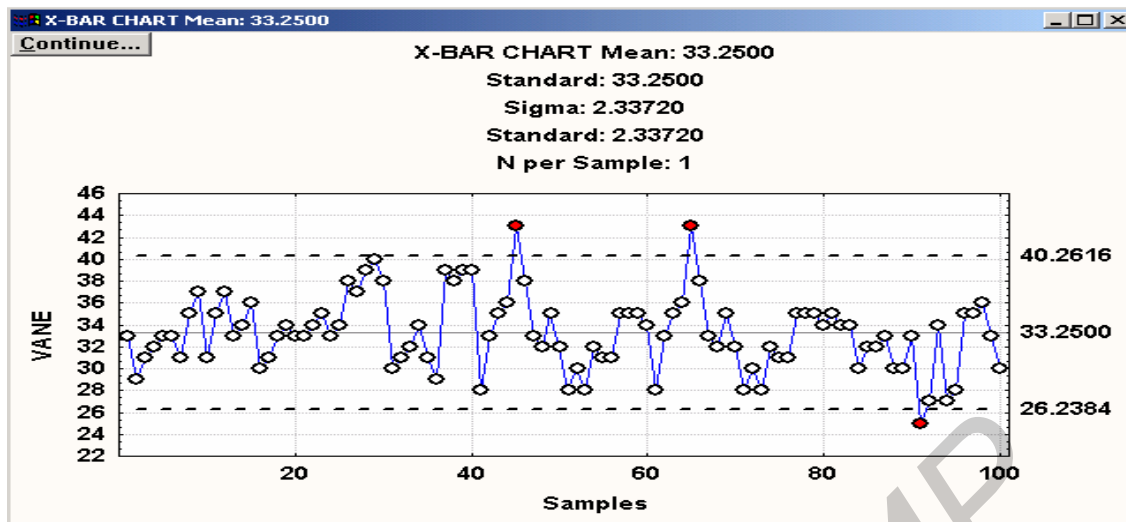


Рис. 4.3. X- карты для переменной *Vane*

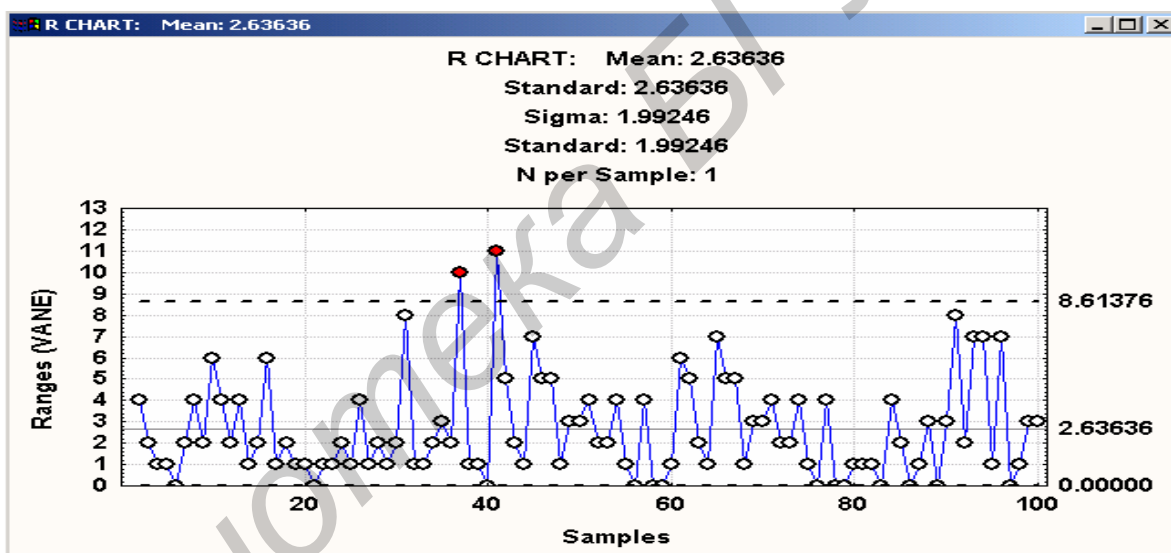


Рис. 4.4. R- карты для переменной *Vane*

На первом графике приведены значения верхнего контрольного предела, центральной линии и нижнего контрольного предела для каждой карты. Пунктиром выделены подгруппы измерений, которые выходят за контрольные пределы: три значения для X-карты и два – для R-карты. Можно точно определить эти значения, для чего нажмем на кнопку *Descriptive statistics* (Описательные статистики). Появятся две таблицы *Vane: X-Bar – X- карта* и *Vane: R-Chart – R- карта*, в которых содержатся данные по подгруппам измерений. В них красным цветом должны быть выделены подгруппы, не попавшие в интервал между контрольными пределами.

Затем необходимо обнаружить и устранить причины потери качества, определяемые условиями конкретного производства. Однако этот вопрос в данной работе рассматриваться не будет. После этого следует скорректировать контрольные карты, для чего необходимо удалить экземпляры

неудовлетворительного качества. После перерасчета системой всех параметров контрольных карт можно заново вывести контрольные карты, скорректированные соответствующим образом. После осуществления процедуры исключения все измерения будут находиться внутри контрольных пределов .

Очевидно, что даже отлаженный производственный процесс нуждается в периодической проверке. Покажем, как это делается.

### 4.3. Диалог модуля «Анализ процесса»

По результатам предыдущей обработки мы убедились, что изучаемый процесс теперь удовлетворяет требованиям качества. Возникает вопрос, насколько он соответствует заданным спецификациям. Для проведения такого анализа используется модуль *Process Analysis* (Анализ процесса), в котором имеется специальная процедура *Process (machine) capability analysis* (Анализ пригодности процесса). Предположим, что в нашем случае номинальный диаметр отверстия на лопастях составляет 30. Допуски лежат в пределах от 20 до 40.

Произведем проверку исследуемого процесса на удовлетворение этим требованиям. Для этого случайным образом отберем 100 изготовленных лопаток и сделаем на них замеры диаметров отверстия. Добавим серию новых замеров в файл данных, назвав новую переменную *capvane*.

Запустим модуль *Process Analysis* (Анализ производственного процесса) и откроем файл данных с добавленной переменной *capvane*. На стартовой панели выберем *Process (machine) capability analysis & tolerance intervals* (Анализ пригодности процесса и доверительные интервалы) и нажмем **OK**. В появившемся диалоговом окне выберем *capvane* в качестве переменной для анализа.

Нажмем кнопку **OK**, появится окно *Process capability analysis* (Анализ пригодности процесса). В поле *Type (Type)* выберем *Lower, Nominal, Upper* (Нижний предел, Номинал, Верхний предел). Нажмем кнопку *Specifications* (Спецификации) и зададим требуемые значения: 30 – для номинала, 20 – для нижнего предела и 40 – для верхнего предела. Тогда диалоговое окно *Анализ пригодности процесса* должно выглядеть, как показано на рис. 4.5.

Нажмем кнопку *Process capability indices* (Индексы пригодности процесса). Появится таблица с различными индексами, характеризующими соответствие распределения произведенных замеров требованиям спецификации (рис. 4.6).

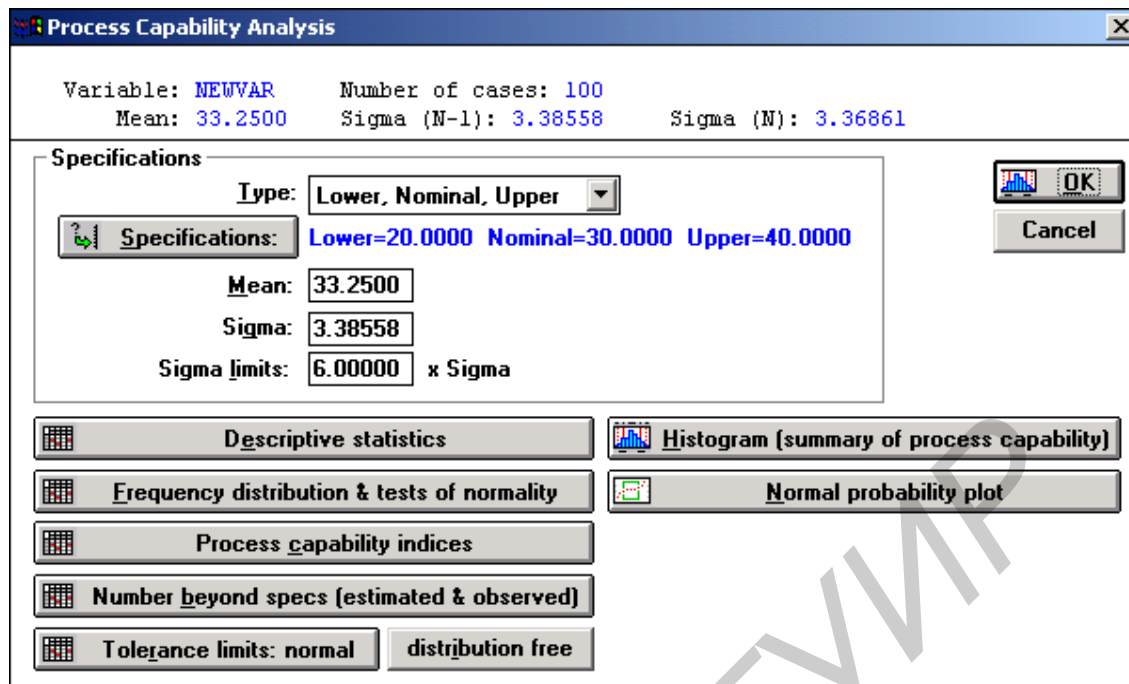


Рис. 4.5. Диалоговое окно *Анализ пригодности процесса*

Variable: CAPVANE (vane.sta)	
PROCESS	-3.000 *Sigma=23.0933
ANALYSIS	+3.000 *Sigma=43.4067
Capability Index	Value
Lower Specification Limit	20.00000
Nominal Specification	30.00000
Upper Specification Limit	40.00000
CP (potential capability)	1.23610
CR (capability ratio)	1.01567
CPK (demonstrated excellence)	.81600
ZL (lower specification z)	-3.91366
ZU (upper specification z)	1.99375
K (non-centering correction)	.34000
CPM (potential capability II)	.70856

Рис. 4.6. Индексы пригодности процесса

Индексы выведенной на экран таблицы следующие:  
*Lower Specification Limit* (Нижняя граница допуска);  
*Nominal Specification* (Номинал);  
*Upper Specification Limit* (Верхняя граница допуска);  
*CP (potential capability)* – CP (Потенциальная пригодность);  
*CR (capability ratio)* – CR (Отношение пригодности);  
*CPK (demonstrated excellence)* – CPK (Подтвержденное качество);  
*ZL (lower specification z)* – ZL (Нижний индекс пригодности);  
*ZU (upper specification z)* – ZU (Верхний индекс пригодности);



***K (non-centering correction)*** – *K (Нецентрирующая поправка)*;

***CPM (potential capability II)*** – *CPM (Потенциальная пригодность)*.

Индекс *CP* представляет собой разность между границами допуска, деленную на 6 стандартных отклонений. В данном случае *CP* равен 1,236 – это хороший результат.

*CPK* – односторонний индекс. Он равен расстоянию от среднего до ближайшей границы допуска, деленному на 3 стандартных отклонения. Здесь *CPK* составляет 0,816. Довольно большое отличие приведенных двух индексов друг от друга говорит о том, что распределение рассматриваемых замеров плохо центрировано относительно верхней и нижней границ допуска.

Индекс *K* представляет собой разницу между номиналом и средним значением распределения, деленную на половину расстояния между границами допуска. Так как этот индекс в нашем случае равен 0,34, можно сказать, что среднее значение распределения лежит на 34%-м пути от центра к верхней границе допуска.

Вернитесь в окно ***Process Capability Analysis*** (*Анализ пригодности процесса*). Для настроек гистограммы укажите в соответствующих опциях ***Lower limit*** (*Нижняя граница*) = 18, ***Upper limit*** (*Верхняя граница*) = 44, ***No. of categories*** (*Число категорий*) = 14.

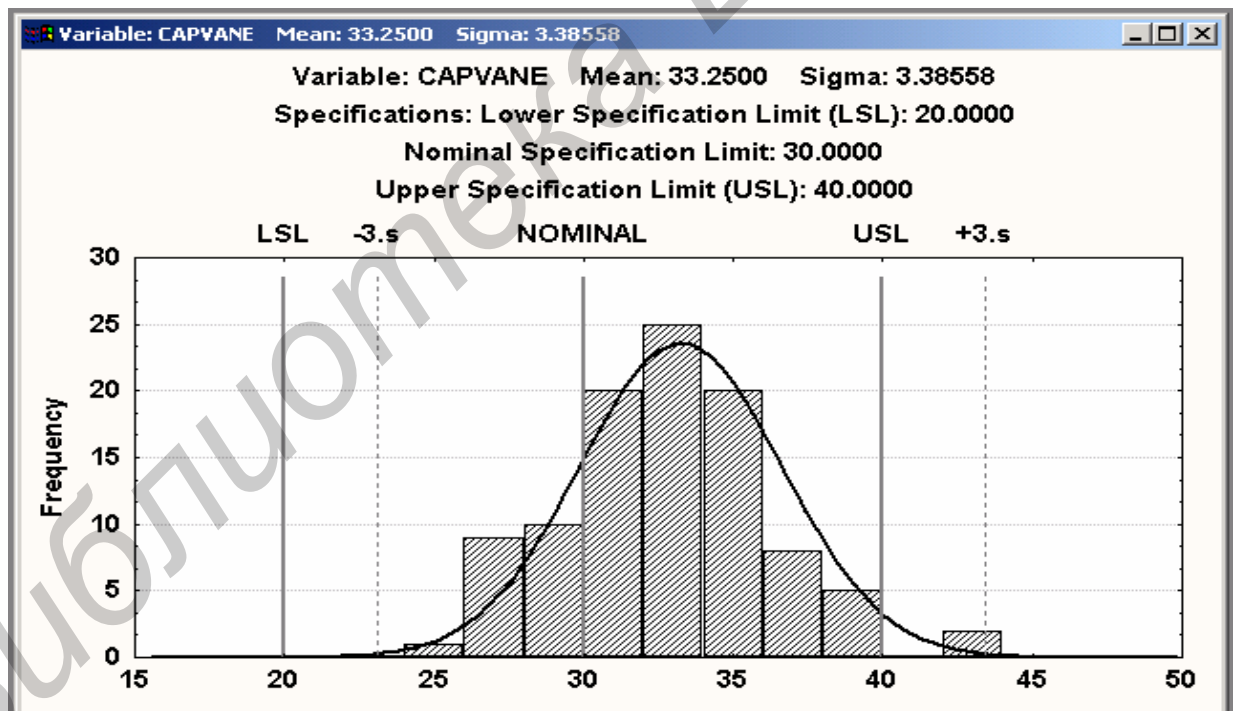


Рис. 4.7. Графическое отображение результатов анализа

На рис. 4.7 хорошо видно, что распределение результатов измерений приблизительно попадает в интервал между границами допуска. Вместе с тем оно смещено относительно центра, т. е. относительно номинала, поэтому можно сделать вывод о невыполнении инженерных требований. Руководству следует принять соответствующие меры.

#### 4.4. ВАРИАНТЫ ЗАДАНИЙ

1. По выборочным данным произвести статистический контроль процесса производства изделия *A*. Работу производить в модуле *Quality Control* (Контроль качества). В табл. 4.1 в столбце *Rad* обозначен радиус измеренного отверстия производимого изделия.

Таблица 4.1

Значения контрольных измерений радиуса отверстий изделия *A*

№ п.п.	<i>Rad</i>	№ п.п.	<i>Rad</i>	№ п.п.	<i>Rad</i>	№ п.п.	<i>Rad</i>	№ п.п.	<i>Rad</i>
1	43	21	43	41	48	61	37	81	45
2	39	22	44	42	43	62	42	82	44
3	41	23	45	43	42	63	44	83	44
4	42	24	43	44	45	64	45	84	40
5	43	25	44	45	42	65	47	85	42
6	43	26	40	46	48	66	43	86	42
7	41	27	41	47	43	67	43	87	43
8	45	28	43	48	42	68	45	88	40
9	47	29	44	49	45	69	47	89	40
10	41	30	43	50	42	70	46	90	43
11	45	31	43	51	41	71	45	91	45
12	47	32	44	52	45	72	47	92	45
13	43	33	45	53	45	73	42	93	46
14	44	34	43	54	45	74	45	94	43
15	46	35	44	55	44	75	49	95	40
16	40	36	48	56	37	76	37	96	45
17	41	37	43	57	45	77	40	97	45
18	43	38	42	58	45	78	37	98	46
19	44	39	45	59	45	79	41	99	43
20	43	40	42	60	44	80	40	100	40

2. Установить, насколько процесс производства изделия *B* соответствует заданным спецификациям. Для проведения анализа использовать модуль *Process Analysis* (Анализ процесса). В табл.4.2 указаны результаты измерений *D* (диаметра) отверстия производимого изделия.

## Результаты измерений

№ п.п.	D	№ п.п.	D	№ п.п.	D	№ п.п.	D	№ п.п.	D
1	52	21	48	41	45	61	48	81	48
2	51	22	50	42	50	62	45	82	50
3	49	23	41	43	47	63	52	83	44
4	46	24	51	44	48	64	49	84	47
5	47	25	53	45	48	65	49	85	46
6	43	26	46	46	49	66	49	86	54
7	47	27	43	47	51	67	50	87	48
8	47	28	44	48	46	68	46	88	48
9	49	29	46	49	51	69	49	89	51
10	48	30	48	50	48	70	49	90	45
11	50	31	47	51	46	71	46	91	48
12	48	32	45	52	49	72	48	92	51
13	50	33	48	53	52	73	48	93	49
14	46	34	53	54	48	74	50	94	46
15	46	35	53	55	50	75	50	95	49
16	51	36	50	56	47	76	42	96	50
17	49	37	47	57	53	77	51	97	48
18	45	38	49	58	52	78	47	98	49
19	47	39	54	59	56	79	51	99	44
20	50	40	48	60	47	80	51	100	47

## ЛИТЕРАТУРА

1. Боровиков В. П. Программа STATISTICA для студентов и инженеров. 2-е изд.– М.: КомпьютерПресс, 2001.–301 с.

2. Методические указания, программа и контрольные задания по курсу "Экономико-математические модели и методы" для студентов экономических специальностей заочной формы обучения/ Сост. С. А. Поттосина.– Мн.: БГУИР, 1999.

3. Боровиков В. П., Боровиков И. П. STATISTICA – статистический анализ и обработка данных в среде Windows. Изд. 2-е, стереотип. – М.: Информ.-издат. дом "Филинь", 1998.– 608 с.

4. Алексахин С. В., Балдин А. В., Николаев А. Б., Строганов В. Ю. Прикладной статистический анализ: Учеб. пособие для вузов.– М.: ПРИОР, 2001. – 224 с.

5. <http://www.statsoft.ru/home/portal/applications/industry/vanes.htm>.

6. <http://www.statsoft.ru/home/portal/applications/industry/gage.htm>.

Библиотека БГУИР

# ПРИЛОЖЕНИЯ

## Приложение 1

### Курс акций компании IBM

Вариант 1

460	474	545	592
457	465	549	596
452	466	545	596
459	467	545	595
462	471	547	598
459	471	543	598
463	467	540	595
479	473	539	592
493	481	532	592
490	488	517	588
492	490	527	582
498	489	540	576
499	489	542	578
497	485	538	589
496	491	541	585
490	492	541	580
489	494	547	579
478	499	553	584
487	498	559	581
491	500	557	581
487	497	557	577
482	494	560	577
479	495	571	578
478	500	571	580
479	504	569	586
477	513	575	583
479	511	580	581
475	514	584	576
479	510	585	571
476	509	590	575
476	515	599	575
478	519	603	573
479	523	599	577
477	519	596	582
476	523	585	584
475	531	587	579
475	547	585	572
473	551	581	577
474	547	583	571
474	541	592	560

Вариант 2

545	592	549	547
549	596	556	548
545	596	557	549
545	595	563	553
547	598	564	553
543	598	567	552
540	595	561	551
539	592	559	550
532	592	553	553
517	588	553	554
527	582	553	551
540	576	547	551
542	578	550	545
538	589	544	547
541	585	541	547
541	580	532	537
547	579	525	539
553	584	542	538
559	581	555	533
557	581	558	525
557	577	551	513
560	577	551	510
571	578	552	521
571	580	553	521
569	586	557	521
575	583	557	523
580	581	548	516
584	576	547	511
585	571	545	518
590	575	545	517
599	575	539	520
603	573	539	519
599	577	535	519
596	582	537	519
585	584	535	518
587	579	536	513
585	572	537	499
581	577	543	485
583	571	548	454
592	560	546	462

Вариант 3

549	547	473	339
556	548	482	350
557	549	486	351
563	553	475	350
564	553	459	345
567	552	451	350
561	551	453	359
559	550	446	375
553	553	455	379
553	554	452	376
553	551	457	382
547	551	449	370
550	545	450	365
544	547	435	367
541	547	415	372
532	537	398	373
525	539	399	363
542	538	361	371
555	533	383	369
558	525	393	376
551	513	385	387
551	510	360	387
552	521	364	376
553	521	365	385
557	521	370	385
557	523	374	380
548	516	359	373
547	511	335	382
545	518	323	377
545	517	306	376
539	520	333	379
539	519	330	386
535	519	336	387
537	519	328	386
535	518	316	389
536	513	320	394
537	499	332	393
543	485	320	409
548	454	333	411
546	462	344	409

Вариант 4

472	486	557	604
469	477	561	608
464	478	557	608
471	479	557	607
474	483	559	610
471	483	555	610
475	479	552	607
491	485	551	604
505	493	544	604
502	500	529	600
504	502	539	594
510	501	552	588
511	501	554	590
509	497	550	601
508	503	553	597
502	504	553	592
501	506	559	591
490	511	565	596
499	510	571	593
503	512	569	593
499	509	569	589
494	506	572	589
491	507	583	590
490	512	583	592
491	516	581	598
489	525	587	595
491	523	592	593
487	526	596	588
491	522	597	583
488	521	602	587
488	527	611	587
490	531	615	585
491	535	611	589
489	531	608	594
488	535	597	596
487	543	599	591
487	559	597	584
485	563	593	589
486	559	595	583
486	553	604	572

Вариант 5

604	561	559	485
608	568	560	494
608	569	561	498
607	575	565	487
610	576	565	471
610	579	564	463
607	573	563	465
604	571	562	458
604	565	565	467
600	565	566	464
594	565	563	469
588	559	563	461
590	562	557	462
601	556	559	447
597	553	559	427
592	544	549	410
591	537	551	411
596	554	550	373
593	567	545	395
593	570	537	405
589	563	525	397
589	563	522	372
590	564	533	376
592	565	533	377
598	569	533	382
595	569	535	386
593	560	528	371
588	559	523	347
583	557	530	335
587	557	529	318
587	551	532	345
585	551	531	342
589	547	531	348
594	549	531	340
596	547	530	328
591	548	525	332
584	549	511	344
589	555	497	332
583	560	466	345
572	558	474	356

Вариант 6

559	485	351	420
560	494	362	405
561	498	363	403
565	487	362	400
565	471	357	408
564	463	362	399
563	465	371	395
562	458	387	400
565	467	391	394
566	464	388	396
563	469	394	394
563	461	382	395
557	462	377	395
559	447	379	400
559	427	384	407
549	410	385	404
551	411	375	398
550	373	383	395
545	395	381	389
537	405	388	376
525	397	399	381
522	372	399	367
533	376	388	362
533	377	397	365
533	382	397	352
535	386	392	362
528	371	385	361
523	347	394	370
530	335	389	372
529	318	388	372
532	345	391	378
531	342	398	371
531	348	399	368
531	340	398	367
530	328	401	379
525	332	406	369
511	344	405	373
497	332	421	367
466	345	423	360
474	356	421	355

Вариант 7

507	521	592	639
504	512	596	643
499	513	592	643
506	514	592	642
509	518	594	645
506	518	590	645
510	514	587	642
526	520	586	639
540	528	579	639
537	535	564	635
539	537	574	629
545	536	587	623
546	536	589	625
544	532	585	636
543	538	588	632
537	539	588	627
536	541	594	626
525	546	600	631
534	545	606	628
538	547	604	628
534	544	604	624
529	541	607	624
526	542	618	625
525	547	618	627
526	551	616	633
524	560	622	630
526	558	627	628
522	561	631	623
526	557	632	618
523	556	637	622
523	562	646	622
525	566	650	620
526	570	646	624
524	566	643	629
523	570	632	631
522	578	634	626
522	594	632	619
520	598	628	624
521	594	630	618
521	588	639	607

Вариант 8

596	594	520	386
603	595	529	397
604	596	533	398
610	600	522	397
611	600	506	392
614	599	498	397
608	598	500	406
606	597	493	422
600	600	502	426
600	601	499	423
600	598	504	429
594	598	496	417
597	592	497	412
591	594	482	414
588	594	462	419
579	584	445	420
572	586	446	410
589	585	408	418
602	580	430	416
605	572	440	423
598	560	432	434
598	557	407	434
599	568	411	423
600	568	412	432
604	568	417	432
604	570	421	427
595	563	406	420
594	558	382	429
592	565	370	424
592	564	353	423
586	567	380	426
586	566	377	433
582	566	383	434
584	566	375	433
582	565	363	436
583	560	367	441
584	546	379	440
590	532	367	456
595	501	380	458
593	509	391	456

Вариант 9

592	639	596	594
596	643	603	595
592	643	604	596
592	642	610	600
594	645	611	600
590	645	614	599
587	642	608	598
586	639	606	597
579	639	600	600
564	635	600	601
574	629	600	598
587	623	594	598
589	625	597	592
585	636	591	594
588	632	588	594
588	627	579	584
594	626	572	586
600	631	589	585
606	628	602	580
604	628	605	572
604	624	598	560
607	624	598	557
618	625	599	568
618	627	600	568
616	633	604	568
622	630	604	570
627	628	595	563
631	623	594	558
632	618	592	565
637	622	592	564
646	622	586	567
650	620	586	566
646	624	582	566
643	629	584	566
632	631	582	565
634	626	583	560
632	619	584	546
628	624	590	532
630	618	595	501
639	607	593	509

Вариант 10

528	542	613	660
525	533	617	664
520	534	613	664
527	535	613	663
530	539	615	666
527	539	611	666
531	535	608	663
547	541	607	660
561	549	600	660
558	556	585	656
560	558	595	650
566	557	608	644
567	557	610	646
565	553	606	657
564	559	609	653
558	560	609	648
557	562	615	647
546	567	621	652
555	566	627	649
559	568	625	649
555	565	625	645
550	562	628	645
547	563	639	646
546	568	639	648
547	572	637	654
545	581	643	651
547	579	648	649
543	582	652	644
547	578	653	639
544	577	658	643
544	583	667	643
546	587	671	641
547	591	667	645
545	587	664	650
544	591	653	652
543	599	655	647
543	615	653	640
541	619	649	645
542	615	651	639
542	609	660	628

Вариант 11

615	541	407	476
616	550	418	461
617	554	419	459
621	543	418	456
621	527	413	464
620	519	418	455
619	521	427	451
618	514	443	456
621	523	447	450
622	520	444	452
619	525	450	450
619	517	438	451
613	518	433	451
615	503	435	456
615	483	440	463
605	466	441	460
607	467	431	454
606	429	439	451
601	451	437	445
593	461	444	432
581	453	455	437
578	428	455	423
589	432	444	418
589	433	453	421
589	438	453	408
591	442	448	418
584	427	441	417
579	403	450	426
586	391	445	428
585	374	444	428
588	401	447	434
587	398	454	427
587	404	455	424
587	396	454	423
586	384	457	435
581	388	462	425
567	400	461	429
553	388	477	423
522	401	479	416
530	412	477	411

Вариант 12

613	660	617	615
617	664	624	616
613	664	625	617
613	663	631	621
615	666	632	621
611	666	635	620
608	663	629	619
607	660	627	618
600	660	621	621
585	656	621	622
595	650	621	619
608	644	615	619
610	646	618	613
606	657	612	615
609	653	609	615
609	648	600	605
615	647	593	607
621	652	610	606
627	649	623	601
625	649	626	593
625	645	619	581
628	645	619	578
639	646	620	589
639	648	621	589
637	654	625	589
643	651	625	591
648	649	616	584
652	644	615	579
653	639	613	586
658	643	613	585
667	643	607	588
671	641	607	587
667	645	603	587
664	650	605	587
653	652	603	586
655	647	604	581
653	640	605	567
649	645	611	553
651	639	616	522
660	628	614	530



Стандартное нормальное (*u*) распределение

Значения, приведенные в табл. П. 2.1, представляют собой величины площадей под стандартной нормальной (гауссовой) кривой от 0 до соответствующего *u*-значения, как показано на рис. П. 2.1. Например, величина этой площади между значениями 0 и 2.36 показана в ячейке, находящейся на пересечении строки 2.3 и графы 0.06, и составляет 0.4909. Значение площади между 0 и отрицательным значением находится на пересечении строки и графы, которые в сумме соответствуют абсолютному значению заданной величины. Например, площадь под кривой от -1.3 до 0 равна площади под кривой между 1.3 и 0, поэтому ее значение находится на пересечении строки 1.3 и графы 0.00 (и составляет 0.4032). График функции плотности вероятности нормального распределения показан на рис. П. 2.2. График функции распределения показан на рис. П. 2.3.

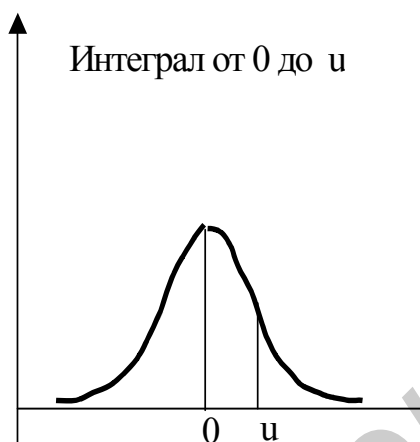


Рис. П. 2.1. Нормальное распределение

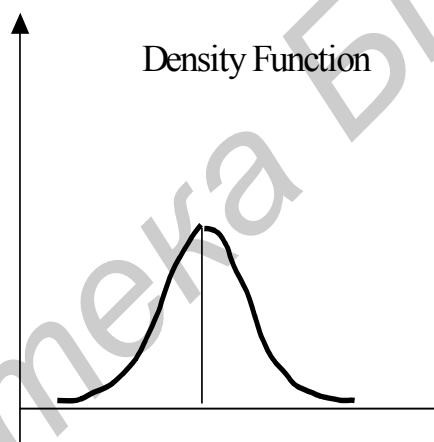


Рис. П. 2.2. Функция плотности вероятности

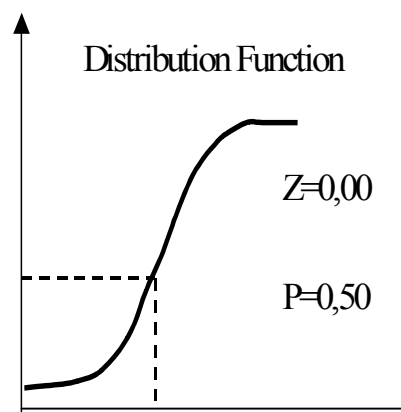


Рис. П. 2.3. Функция распределения

Таблица П. 2.1

Нормальное (*u*) распределение

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>	<i>11</i>
<b>0.0</b>	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
<b>0.1</b>	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
<b>0.2</b>	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
<b>0.3</b>	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
<b>0.4</b>	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
<b>0.5</b>	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
<b>0.6</b>	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549

1	2	3	4	5	6	7	8	9	10	11
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

### Распределение Стьюдента

Форма распределения Стьюдента (рис. П. 2.4) зависит от числа степеней свободы. В верхней части табл. П. 2.2 приведена вероятность получения значения, большего, чем указаны в соответствующей ячейке. Критическое значение  $t$ -распределения с 6-ю степенями свободы, соответствующее вероятности 0.05, находится на пересечении графы 0.05 и строки 6:  $t(0.05,6) = 1,943180$ . Графики функции плотности вероятности и функции распределения показаны на рис. П. 2.5 и рис. П. 2.6 соответственно.

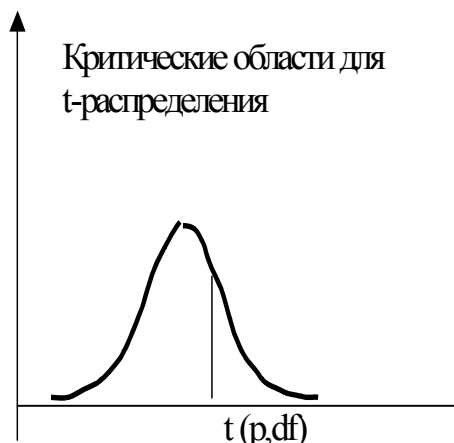


Рис. П. 2.4. Распределение Стьюдента

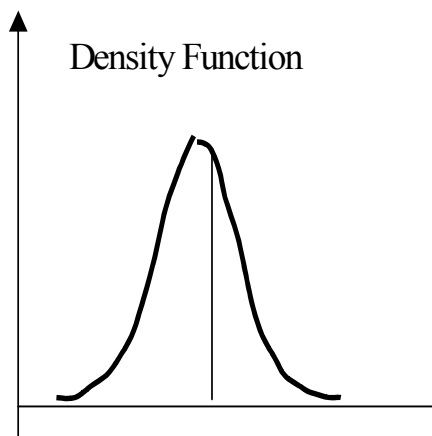


Рис. П. 2.5. Функция плотности вероятности

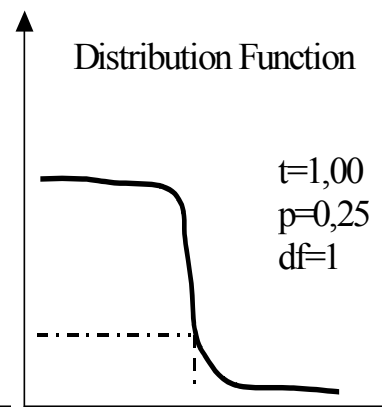


Рис. П. 2.6. Функция распределения

Таблица П. 2.2

Распределение Стьюдента

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	2	3	4	5	6	7	8	9
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066

1	2	3	4	5	6	7	8	9
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

### F-распределение

F-распределение является асимметричным и обычно используется в дисперсионном анализе. Такую плотность распределения имеют величины, являющиеся отношением двух величин, имеющих *хи-квадрат* – распределение, при этом соответствующее F-распределение определяется двумя значениями числа степеней свободы. На рис. П. 2.9 показано распределение  $F(10,10)$ . Первый индекс всегда соответствует числу степеней свободы для числителя, второй – для знаменателя. Этот порядок является существенным, поскольку  $F(10,12)$  не равно  $F(12,10)$ . В приведенной ниже табл. П. 2.3 в графах показано число степеней свободы числителя, а в строках – число степеней свободы для знаменателя. Критическое значение F-распределения для вероятности  $\alpha = 0,05$  и степеней свободы 10 и 12 находится на пересечении графы со значением 10 (числитель) и строки со значением 12 (знаменатель). В табл. П. 2.3, для  $F(0.05, 10, 12) = 2,7534$ . На рис. П. 2.8 и рис. П. 2.9 показаны функция плотности вероятности и функция распределения соответственно.

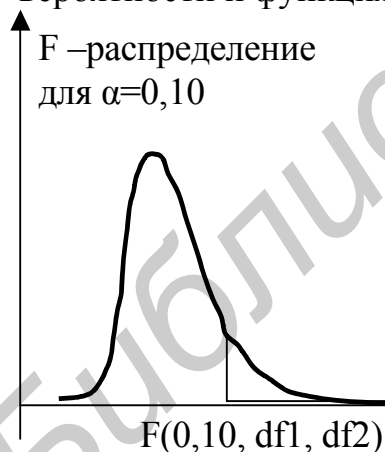


Рис. П. 2.7. F-распределение

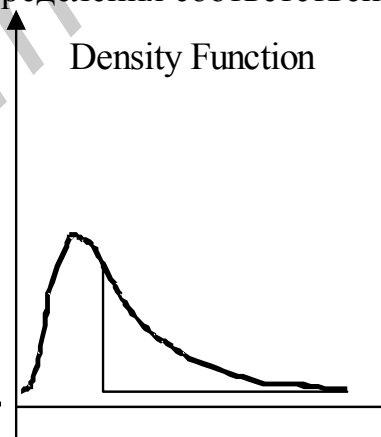


Рис. П. 2.8. Функция плотности вероятности

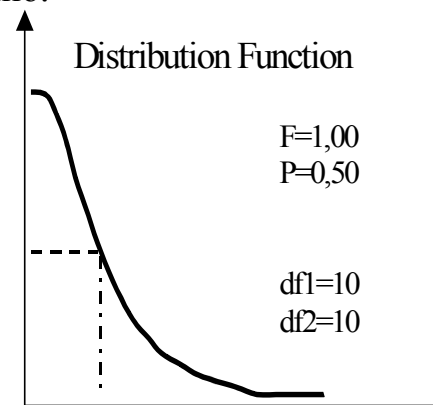


Рис. П. 2.9. Функция распределения

Таблица П. 2.3

 $F$ - распределение для  $\alpha = 0,05$ 

df2/ df1	1	2	3	4	5	6	7	8	9	10	12	15
<i>l</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>	<i>11</i>	<i>12</i>	<i>13</i>
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	243.9060	245.9499
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4125	19.4291
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446	8.7029
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117	5.8578
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777	4.6188
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9999	3.9381
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5107
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839	3.2184
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0061
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	2.7186
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037	2.5331
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.5342	2.4630
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4034
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247	2.3522
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3807	2.3077
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421	2.2686
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.3080	2.2341
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660	2.3210	2.2504	2.1757
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2258	2.1508
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	2.2747	2.2036	2.1282
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834	2.1077
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0889

Окончание табл. П.2.3

<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>	<i>11</i>	<i>12</i>	<i>13</i>
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.1479	2.0716
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.1323	2.0558
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900	2.1179	2.0411
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229	2.1768	2.1045	2.0275
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.0148
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	2.0035	1.9245
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401	1.9926	1.9174	1.8364
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588	1.9105	1.8337	1.7505
inf	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.7522	1.6664

Учебное издание

Музычина Татьяна Михайловна,  
Поттосина Светлана Анатольевна

**ИСПОЛЬЗОВАНИЕ ПАКЕТА ПРИКЛАДНЫХ ПРОГРАММ  
«СТАТИСТИКА»**

**ЛАБОРАТОРНЫЙ ПРАКТИКУМ**  
по курсу «Эконометрика»  
для студентов экономических специальностей БГУИР  
всех форм обучения

Редактор Н.А. Бебель

Корректор Е.Н. Батурчик

---

Подписано в печать 05.05.2004.

Бумага офсетная.

Уч.-изд.л. 4,0.

Печать ризографическая.

Тираж 150 экз.

Формат 60x84 1/16.

Усл.печ.л. , .

Заказ 567.

---

Издание и полиграфическое исполнение: Учреждение образования

“Белорусский государственный университет информатики и радиоэлектроники”

Лицензия на осуществление издательской деятельности № 02330/0056964 от 01.04.2004.

Лицензия на осуществление полиграфической деятельности № 02330/0133108 от 30.04.2004.

220013, Минск, П. Бровки, 6.