

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Государственное образовательное учреждение высшего профессионального образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

О.В. Стукач

**ПРОГРАММНЫЙ КОМПЛЕКС STATISTICA
В РЕШЕНИИ ЗАДАЧ
УПРАВЛЕНИЯ КАЧЕСТВОМ**

*Рекомендовано в качестве учебного пособия
Редакционно-издательским советом
Томского политехнического университета*

Издательство
Томского политехнического университета
2011

УДК 658.562:004(075.8)

ББК 30.607–7я73

С88

Стукач О.В.

С88 Программный комплекс Statistica в решении задач управления качеством: учебное пособие / О.В. Стукач; Томский политехнический университет. – Томск: Изд-во Томского политехнического университета, 2011. – 163 с.

В пособии даётся систематизированное изложение методологии решения проблемы повышения качества с использованием методов теории вероятностей и математической статистики. Подробно рассмотрена работа с универсальным пакетом «Statistica» по системному подходу к обработке данных: анализу закономерностей в данных, всестороннему и последовательному исследованию статистической информации, формированию статистических выводов. Материал позволяет по-новому взглянуть на методы статистического анализа процессов и использовать их как комплекс системных мероприятий по повышению качества управления, а также как лабораторный курс по изучению использования статистических методов в промышленном управлении, для изучения классификации и поиска максимально точной и прагматичной информации о структуре данных.

Рекомендуется студентам направления подготовки 221700 «Стандартизация и метрология» (квалификация «бакалавр» и «магистр») для изучения курса «Программные статистические комплексы».

УДК 658.562:004(075.8)

ББК 30.607–7я73

Рецензенты

Доктор технических наук, профессор ТУСУРа

А.А. Шелупанов

Кандидат технических наук, доцент ТУСУРа

В.И. Карнышев

© ГОУ ВПО НИ ТПУ, 2011

© Стукач О.В., 2011

© Обложка. Издательство Томского
политехнического университета, 2011

ПРЕДИСЛОВИЕ

Оптимальные стратегии управления качеством сегодня становятся основным фактором создания длительного конкурентного преимущества и роста инвестиционной привлекательности. В этом смысле это ресурс, причём более важный, чем деньги или товарно-материальные ценности. Российские компании в конкурентной борьбе зачастую проигрывают западным не только из-за технологической отсталости и неэффективного управления. Важная проблема – плохое качество продукции и услуг.

Повышение качества становится одним из направлений совершенствования бизнеса. Наиболее очевидным способом является его статистическое моделирование. В результате наблюдается более глубокое проникновение в изучаемые процессы, в самую природу явлений. Анализ статистических данных позволяет быстро выявить и оценить все необходимые характеристики процесса. Затем можно провести крупномасштабное компьютерное моделирование как нормально протекающего процесса, так и выходящего из-под контроля с целью сравнения разнообразных улучшающих вмешательств и выбора оптимального из них. Можно также отметить применение статистического моделирования в оценке эффективности инвестиций, прогнозировании сбыта и так далее. Поэтому путь математического моделирования процессов и последовательного установления логических причинно-следственных связей для обеспечения возможности наблюдения, контроля и управления ими – это эффективное средство при решении различных проблем.

Одним из важнейших элементов системы менеджмента качества (СМК) на всех этапах жизненного цикла продукции в соответствии с требованиями стандартов серии ИСО 9000 является применение статистических методов. Использование статистических методов способствует пониманию изменчивости показателей качества продукции и, следовательно, может помочь предприятию повысить результативность и эффективность принимаемых решений.

Статистическое мышление необходимо для каждого участника процесса, а для этого необходимо знать статистические методы, которые доступны для всех за счёт своей простоты, достигнутой в семи инструментах контроля качества. Каждый служащий компании или организации, используя статистические методы для анализа и контроля процессов, тем самым способствует повышению качества, эффективности производства и снижению затрат.

Но для такой сферы, как управление качеством остаются огромные малоиспользуемые резервы вследствие неглубокого понимания теории вероятностей и математической статистики. Как справедливо указано в работе [1], «основным препятствием для широкого применения на микрокомпьютерах методов моделирования является недостаток знаний и воображения у руководителей, принимающих решения. Кроме того, не хватает специалистов по управлению, хорошо разбирающихся в моделировании и умело избегающих подводных камней».

Аналитики добились выдающихся успехов по применению компьютерных методов многомерного анализа, кластерного анализа, исследования временных рядов и других методов. Но эти методы слабо используются в управлении качеством. Ещё остается громадная область применения универсальных статистических пакетов для решения задач, математически простых, но имеющих большое практическое значение.

Современная конкурентная среда продуцирует огромный информационный поток, лишая правильного восприятия действительности. Без современных технологий интеллектуального анализа данных долговременное управление реальными процессами и принятие правильных решений невозможно. Статистическое управление позволяет грамотно собрать данные, описать их структуру, понять и увидеть закономерности в массе вероятностных явлений. Статистические методы – это удивительно мощный инструмент управления. Даже простейшие методы визуального анализа позволяют прояснить сложную ситуацию, тщательно скрытую за нагромождением информации, исследовать её и принять доказательное решение.

Учебное пособие написано по материалам лабораторного курса «Программные статистические комплексы». Все практические расчёты и примеры рассматриваются в процедурах пакета Statistica. Поэтому в первой главе приводится минимально необходимое для профессиональной работы описание структуры и возможностей пакета. Подробно с описанием пакета можно ознакомиться по специальной литературе [2–4]. Во второй главе изложена методика разведочного анализа данных. Предварительный анализ включает построение графиков и вычисление простейших статистик. Разведочный анализ также даёт возможность определить множество методов, которые пригодны для последующего анализа. Третья глава содержит материал по проверке статистических гипотез, в основном, для целей контроля и управления качеством. Четвёртая глава посвящена корреляционному и регрессионному анализу, то есть установлению факта взаимной зависимости слу-

чайных величин и вида этой зависимости. В современных условиях эти зависимости, как технические, так и экономические, сильно нелинейны. Подходящие модели этих зависимостей даёт регрессионный анализ, обсуждаемый в главе 5. Подробно рассмотрены методы регрессионного анализа: множественная регрессия, процедуры пошагового выбора наиболее значимых факторов, вопросы проверки значимости и адекватности моделей. В шестой главе проводится анализ процессов методами построения различных карт контроля качества. Здесь рассматривается идеология «шести сигм» применительно к процессам. Важное место в анализе процессов занимают данные, к которым зачастую неприменимы впрямую различные статистические модели. Кластерный анализ, методы которого исследуются в седьмой главе, служит примером того, как извлечь информацию из данных любого объёма в любой ситуации. Методы кластерного анализа работают всегда, поэтому их можно применять как для разведочного анализа, так и при исследованиях разнородных множеств. Восьмая глава посвящена выявлению факторов, непосредственно влияющих на изучаемые процессы, то есть факторному анализу. Необходимость в нём встречается часто и требует профессиональной подготовки. Методы иллюстрируются многочисленными практическими примерами.

Научная статистика предоставляет в распоряжение методы, полезные для проведения углубленной исследовательской работы. Но использование этих методов не освобождает от необходимости думать. Основная цель этого учебного пособия состоит в том, чтобы научить читателя статистически мыслить, а не просто заучить понятия математической статистики.

Все методы, развитые в книге, сопровождаются примерами. При изложении материала автор старался высветить идею того или иного метода, не прибегая к математической строгости, так как многие методы интуитивно понятны и нашли применение до того, как были обоснованы математиками. Многие доказательства, столько характерные для фундаментальных математических работ либо опущены, либо заменены подтверждающими примерами. В пособии нет строгого обоснования методов, а продемонстрированы их успешное применение для конкретных практических задач. Поступая аналогичным образом, читатели могут использовать рассмотренные методы для решения своих задач с обнадеживающим результатом. Автор стремился не к разнообразию примеров, а к разнообразию методов, которые могут быть применены к однотипным задачам.

В конце книги приведен список цитированной литературы по статистическим методам. Конечно, список работ не является исчерпывающим, но в нём нашли отражение все основные труды по промышленной статистике с качеством «выше среднего».

Литературы по промышленной статистике издано немало. Ещё больше книг создано по компьютерной обработке данных с помощью статистических пакетов. Но они в основном являются описанием программного интерфейса, а изложение статистических задач и методов их решения сделано весьма неполно. Редкое исключение из этого правила – работа [5], под влиянием которой была написана настоящая книга.

Полезные замечания и идеи, которые улучшили работу, были высказаны участниками тренингов по статистическому управлению процессами на различных предприятиях страны. Автор благодарит коллег В.Н. Борикова и Н.Н. Казанцеву [6] за многие мысли, в явном и неявном виде присутствующие в материале книги.

Известно, что статистические расчёты требуют большого количества наблюдений. Но формат книги не может вместить полного объёма данных, и автору приходилось сокращать их так, чтобы принципиально не изменить решения и выводов, которые были сделаны исходя из реального объёма данных. Возможно, это не совсем удалось. В этой связи хочется снова процитировать старую, но добрую книгу [1], лучше всего характеризующую данную проблему: «Немного найдется областей, для которых высказывание «малое знание опасно» более истинно, чем для теории вероятностей и математической статистики: понятия иногда довольно трудноуловимы, а неправильные ответы обычно не являются «очевидно неправильными», как в других областях математики... Эта теория является настоящим минным полем для тех, кто невнимательно изучил и некритически применяет основные принципы. Да и вообще всем нам свойственно ошибаться».

ГЛАВА 1. РАЗВЕДОЧНЫЙ ВИЗУАЛЬНЫЙ АНАЛИЗ ДАННЫХ И СТРУКТУРА ПРОГРАММЫ STATISTICA

1.1. Сбор и анализ данных

В условиях предприятия можно собрать огромное множество данных. Когда существует намерение применить какой-то практический способ выполнения работы, естественно оценить, пригоден он или нет. Обычно решение принимается на основе прошлых результатов и опыта, или же за основу берутся традиционные методы. В случае заводской работы, когда данные собираются на протяжении реального производственного процесса, процедурные методы вводятся исходя из полученной информации. Производственная процедура будет наиболее эффективной, если сделана её надлежащая оценка. Для этого очень важны данные с рабочих мест:

- данные и их последующая оценка, формирующие базис для действий и решений. Поскольку заводские операции различаются в зависимости от конкретной производственной процедуры, данные целесообразно классифицировать по их назначению;

- данные, позволяющие понять действительную ситуацию. Эти данные собирают, чтобы проверить разброс в размерах деталей, происходящих из-за наладки станка, или проконтролировать процент дефектных единиц, содержащихся в поступающей партии. По мере роста количества данных их нужно, для облегчения понимания и дальнейшего объяснения, статистически организовать. Тогда можно будет провести оценивание и сравнить состояние поступившей партии или производственного процесса с установленными стандартными или заданными величинами и т. д.;

- данные для анализа. Эти данные используются, например, для выявления связи между дефектом и причиной. Такие данные собирают путем изучения прошлых результатов и проведения новых испытаний, при этом для получения корректной информации используют различные статистические методы;

- данные для управления производственным процессом. Данные этого вида, полученные после проверки качества продукции, могут использоваться с целью установления, нормально ли отлажен производст-

венный процесс или нет. Для такого оценивания применяют контрольные карты, на основе которых принимают соответствующие меры;

- данные для регулировок оборудования;
- данные для приёмки и забракования. Формы этих данных используются для приемки или забракования деталей и изделий после контроля. Существует два метода контроля: сплошной и выборочный, на основе полученной информации решают, что делать с деталями или продукцией.

Данные служат основой действий. После оценивания фактического состояния, выявленного посредством данных, можно принимать надлежащие меры. Первый критический шаг – определить, представляют ли данные типичную ситуацию. Иными словами, достоверно ли собраны данные, чтобы с их помощью выявить факты, и позволяют ли собранные, проанализированные и прошедшие сравнения данные выявить факты. Первая часть формулировки относится к задачам выборочных методов, вторая – к статистической обработке данных. Необходимо все-сторонне рассматривать цели сбора данных, подходящие методы взятия выборок и сортировки данных. Не следует неоправданно много набирать данные какого-то конкретного типа только потому, что их легко собирать. То же справедливо и в отношении неполных данных, которые бывают удобны для сбора, но недостаточно результативны и удовлетворительны.

Нужно, чтобы данные представляли факты, а применяемые статистические методы приводили к достоверной оценке собранных данных. Основа решения может быть найдена только после сравнения с ситуацией в целом, как она представлена на гистограмме или контрольной карте.

Даже понимая потребность в наличии данных, на многих рабочих местах зачастую трудно получить их в численных значениях.

Цель сбора данных – не в нахождении их численных значений, а в создании базы для принятия решений. Сами данные могут быть выражены в любой форме. В общем случае данные можно разделить на данные измерений (длина, вес, время и т. п.) и данные подсчётов (число дефектных единиц в партии, число конкретных дефектов, процент дефектных единиц и т. д.). Кроме того, существуют данные по относительной выгоде, данные числовых последовательностей и данные функций распределения, которые более сложны, но полезны для тех, кто имеет дело с экспериментом, чтобы на их основе делать выводы.

После того как данные собраны, их анализируют. Нужная информация при этом извлекается с помощью использования статистических методов. Следовательно, данные необходимо организовывать таким образом, чтобы облегчить дальнейший анализ.

Прежде всего нужно чётко записывать природу данных. Между их сбором и проведением анализа может пройти значительное время. Больше того, листы данных могут пригодиться в других случаях и для другого применения. Необходимо записывать не только цель измерений, но и дату, и используемый прибор, и фамилию проводившего измерения и метод и т. д. Кроме того, записывать данные нужно в такой форме, чтобы их легко было использовать в дальнейшем.

1.2. Общие сведения о пакете Statistica

Универсальная интегрированная система, предназначенная для статистического анализа, визуализации данных и разработки пользовательских приложений Statistica – это современный пакет, в котором реализованы все новейшие компьютерные и математические методы статистического анализа данных. Программа имеет несколько тысяч зарегистрированных пользователей во всем мире, является наиболее динамично развивающимся статистическим пакетом и мировым лидером на рынке статистического программного обеспечения [2–5].

Для того чтобы собранные данные грамотно обработать и извлечь из них максимум информации, требуются немалые усилия. Программа Statistica – это надёжный помощник и консультант. Она снабжена подсказками, какие методы анализа существуют и какие из них лучше всего подходят для тех или иных задач. Электронный учебник по статистике [7] сильно облегчает процесс освоения программы.

Система избавляет пользователя от рутинных вычислений, наглядно отображает результаты анализа, помогает оптимально спланировать будущие эксперименты и создаёт высококачественные отчёты, оставляя специалисту удовольствие интерпретации результатов и формулировки выводов. Система содержит полный набор классических и современных методов анализа данных, что позволяет гибко организовать работу. Помимо общих статистических и графических средств, в системе имеются специализированные модули, например, для проведения социологических исследований, решения промышленных и других задач, при решении которых возникает проблема анализа статистических данных.

Система обладает следующими общепризнанными достоинствами:

- содержит полный набор классических и продвинутых методов анализа данных;
- легка в освоении подготовленным пользователем;
- полностью совместима с приложениями операционной системы Windows;
- является средством построения приложений в конкретных областях;
- данные системы Statistica легко конвертировать в различные базы данных и электронные таблицы;
- в комплект поставки входят специально подобранные примеры, позволяющие систематически осваивать методы анализа;
- поддерживает большинство Интернет-форматов;
- поддерживает высококачественную графику, позволяющую эффективно визуализировать данные и проводить графический анализ;
- содержит язык программирования, который позволяет расширять систему и запускать её из других Windows-приложений.

На рис. 1.1 представлено главное меню, которое появляется при запуске пакета Statistica.

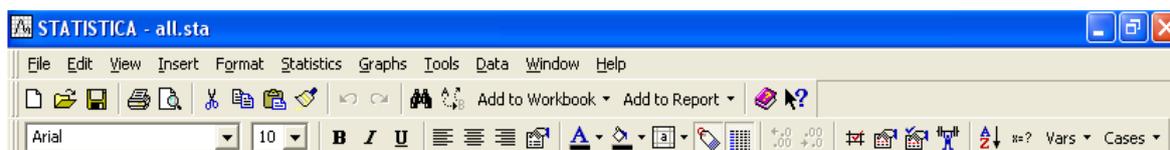


Рис. 1.1. Главное меню пакета Statistica

Панель состоит из следующих опций:

- файл (file);
- редактирование (edit);
- просмотр (view);
- вставка (insert);
- формат (format);
- статистика (statistics);
- графики (graphs);
- инструменты (tools);
- данные (data);
- окно (window);
- справка (help).

Подробно о назначении всех инструментальных кнопок можно узнать, если нажать кнопку . При этом раскрывается соответствующее окно с системой поиска и механизмом гиперссылок.

При работе в программе Statistica наиболее часто используются опции Статистика (Statistics) и Графики (Graphs), диалоговые окна которых показаны на рис. 1.2–1.3.



Рис. 1.2. Диалоговое окно опции Statistics

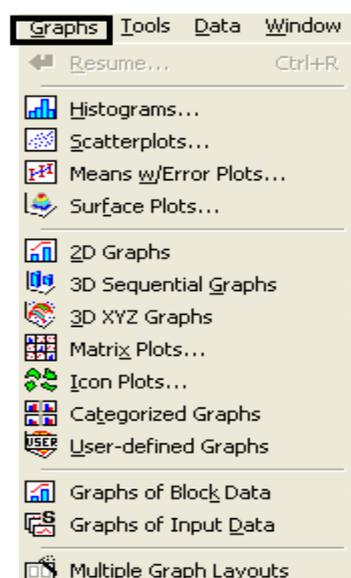


Рис. 1.3. Диалоговое окно опции Graphs

Statistica позволяет:

- построить различные графики: гистограммы (Graphs/Histograms), графики рассеяния (Graphs/Scatterplots), круговые диаграммы (Graphs/ 2D Graphs/ Pie Charts), построить 3D (3D XYZ Graphs) и другие графики;
- вычислить вероятность, среднее значение и т. д., построить графики различных распределений с помощью вероятностного калькулятора (Statistics/Probability Calculator);
- построить диаграмму Парето (Statistics/ Industrial Statistics&Six Sigma/ Quality Control Charts/ Pareto chart analysis);
- построить диаграмму причин и результатов (Statistics/ Industrial Statistics&Six Sigma/ Process Analysis/ Cause-effect diagrams);
- построить контрольные карты (Statistics/ Industrial Statistics&Six Sigma/ Quality Control Charts);

- провести кластерный анализ (Statistics/ Multivariable Exploratory Techniques/ Cluster Analysis);
- провести нелинейное оценивание – регрессионный анализ (Statistics/ Advanced Linear/ Nonlinear Models/Nonlinear Estimation);
- провести корреляционный анализ (Statistics/Basic Statistics/ Correlation Matrices);
- рассчитать статистические характеристики переменных (Statistics/ Basic Statistics/ Descriptive Statistics);
- провести анализ временных рядов (Statistics/ Advanced Linear/ Nonlinear Model / Time Series Analysis/ Forecasting);
- организовать анализ с помощью других статистических методов, используемых в промышленности для обработки данных.

1.3. Запуск программы Statistica

Рабочее окно пакета Statistica приведено на рис. 1.4. Чистый лист представляет собой таблицу из строк и столбцов (трафарет). Перемещаться по листу из ячейки в ячейку можно с помощью стрелок и клавишей Enter или щелчком левой кнопки мыши в нужной ячейке. Ячейка, в которой стоит курсор, обведена чёрной контурной линией и называется активной.

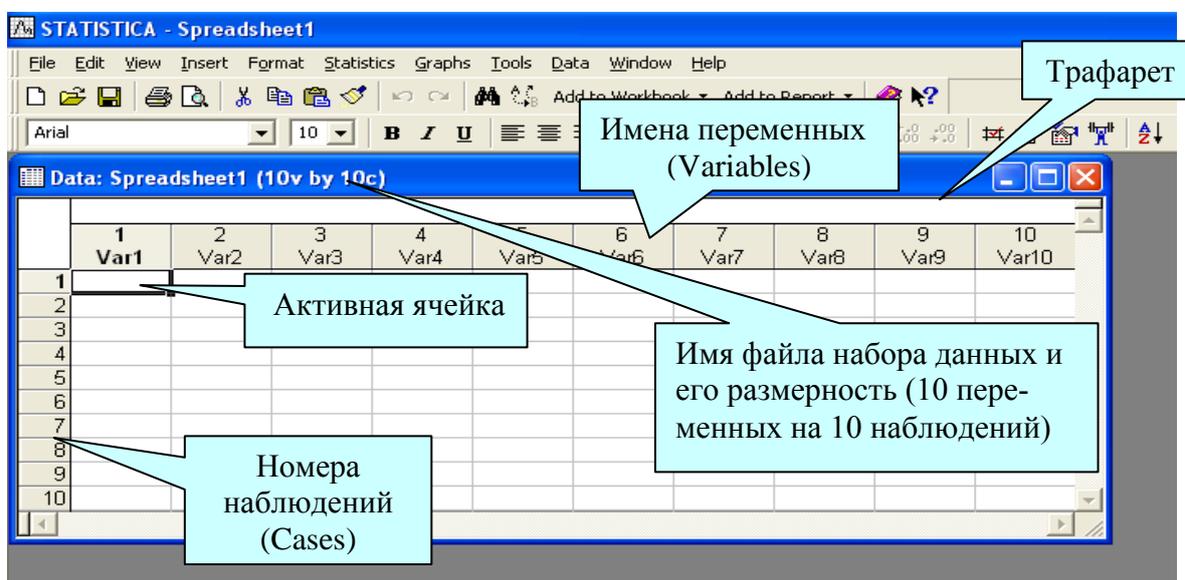


Рис. 1.4. Рабочее окно пакета Statistica

1.4. Структура ввода и редактирования данных

Набор данных в пакете Statistica – это прямоугольная таблица, столбцам которой соответствуют обрабатываемые переменные (Variables), а строкам отвечают наблюдения (Cases) значений переменных. В отличие от электронной таблицы Excel, где строки и столбцы могут быть интерпретированы пользователем по собственному желанию, в программе Statistica всё подчинено обработке случайных переменных.

Для создания нового набора данных нужно, прежде всего, завести файл с трафаретом таблицы нужных размеров. Для этого необходимо использовать модуль *File/ New*. В раскрывшемся диалоговом окне, приведенном на рис. 1.5, необходимо выбрать нужное количество столбцов (Variables) и строк (Cases). При нажатии опции Insert в основном меню или кнопки Vars на панели инструментов становятся доступными команды редактирования переменных (столбцов): Add (добавить новые переменные), Delete (удалить переменные), Move (переместить) и др. При нажатии кнопки Cases становятся доступными аналогичные команды редактирования строк.

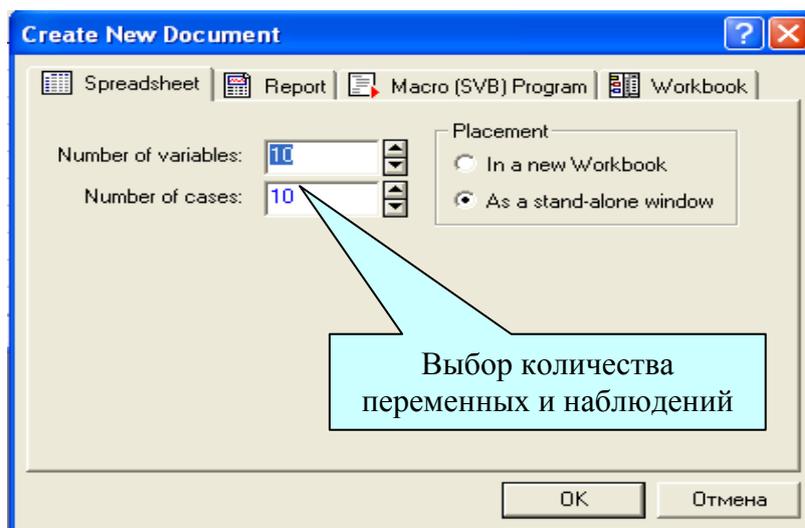


Рис. 1.5. Создание нового документа

Наблюдениям и переменным в трафарете можно дать содержательные названия. В любом случае наблюдения нумеруются. Имена переменных лучше всегда делать содержательными, а не абстрактными Var1, Var2 и т. д. (как на рис. 1.4). Для этого необходимо дважды щёлк-

нуть левой кнопкой мыши по переменной в трафарете, в результате чего появится диалоговое окно (рис. 1.6).

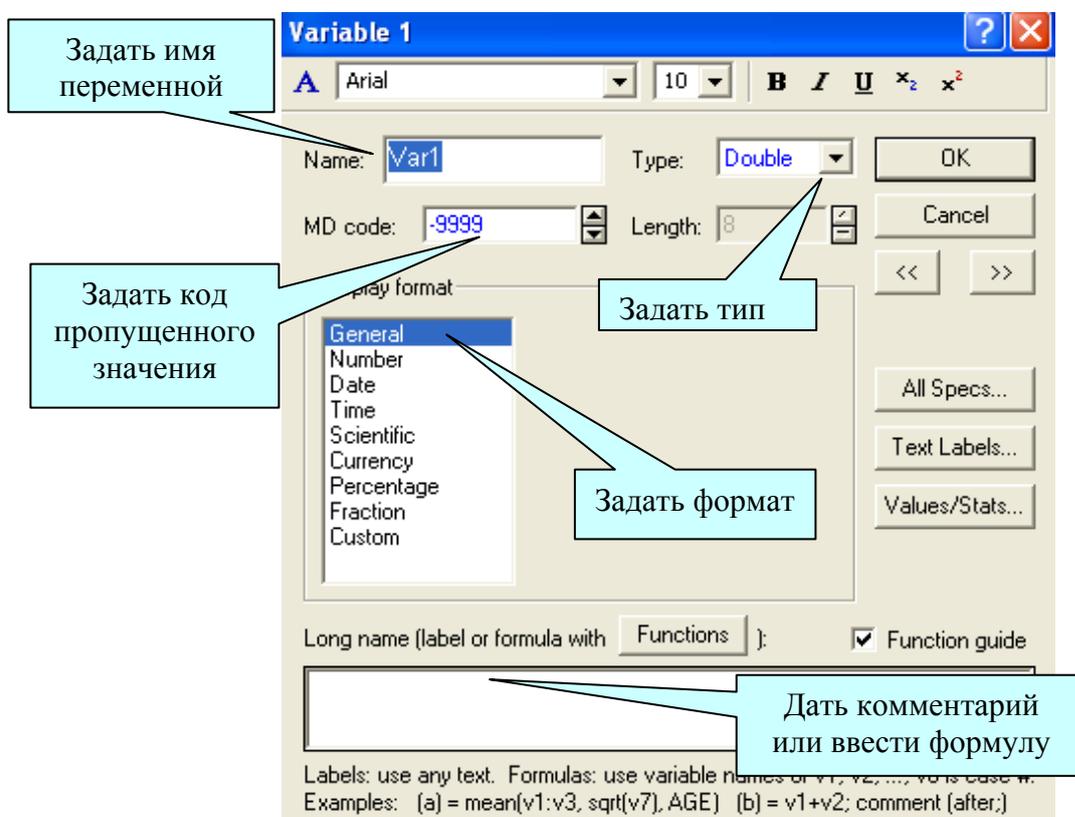


Рис. 1.6. Изменение параметров переменных

Кроме имени (Name) для каждой переменной надо указать так называемый код пропущенного значения (MD Code). По умолчанию этот код равен «-9999», и он отмечает в памяти для процедур обработки пакета, что на самом деле в определённой клетке трафарета реального значения нет. Изображается пропущенное значение на экране в наборе данных пробелом. Из обязательных атрибутов переменной надо указать тип и формат её значений. Тип (Type) определяет, будет ли переменная числовой, текстовой, датой, временем и прочее, а формат (Format) описывает размеры значений переменной. При этом формат каждой переменной нужно определить особенно тщательно. По умолчанию он числовой с фиксированной точкой, где под все значащие цифры, знак числа и десятичную точку отведено 8 символов, 3 из которых предназначены для дробной части. Значениям переменной можно также дать развернутый содержательный комментарий (Long Name).

В этом же поле можно задать формулу, по которой будет рассчитываться выбранная переменная; например, можно написать $=v1+v2$, и тогда выбранная переменная может быть пересчитана по указанной формуле: найдена сумма первой и второй переменной. В формулах переменные можно обозначать буквой v с указанием номера (например, $v1$ означает первый столбец) или написать действительные названия переменных. Чтобы пересчёт состоялся, нажмите кнопку «ОК» и согласитесь с предложением «Recalculate the variable now». Другой способ – нажать кнопку *Vars* и выбрать команду *Recalculate*. После точки с запятой в поле формулы можно написать комментарий.

Приведём пример **многомерной** таблицы с данными (табл. 1.1).

Таблица 1.1

Данные по ремонту оборудования

Дата	Установка	Оборудование	Дефект	Цена потерь, руб.	Результат
02.05.2007	ТВА160	ЧПТВА	Остановка	5500	Не устранён
03.05.2007	ДС158	1015	Погрешность	4600	Откалиброван
06.05.2007	ТВА160	ЧПТВА	Остановка	3250	Не устранён
09.05.2007	ДС158	ПК	Сбой	5180	Устранён
10.05.2007	SPESO	Фильтр	Поврежд. цепи	6380	Отремонтирован
21.05.2007	ДС158	Горелка	Бурс	1500	Отремонтирован
25.05.2007	МАП	ЧПМАП	Остановка	7560	Отремонтирован
14.06.2007	ТВА160	Термо	Износ	2000	Замена
17.06.2007	ДС1581	510	Поврежд. цепи	1100	Устранён
19.06.2007	МАП	Фильтр	Пурф	1700	Отремонтирован
22.06.2007	ТВА160	ЧПТВА	Остановка	5940	Не устранён
23.06.2007	МАП	Фильтр	Пурф	2460	Отремонтирован
23.06.2007	ТВА160	ЧПТВА	Остановка	1750	Не устранён
10.07.2007	ДС158	Пневмо	Остановка	4300	Отремонтирован
15.07.2007	SPESO	Горелка	Не разжигается	4300	Отремонтирован
19.07.2007	МАП	Термо	Износ	5690	Замена
29.07.2007	ДС158	Горелка	Бурс	2100	Отремонтирован
05.08.2007	ТВА160	ЧПТВА	Остановка	4000	Не устранён
08.08.2007	ТВА160	Горелка	Помеха	2500	Устранён
11.08.2007	МАП	Фильтр	Поврежд. цепи	7760	Отремонтирован

С созданными файлами можно выполнять следующие операции:

- открытие файла данных: в меню *File* необходимо выбрать *Open* и открыть интересующий файл;
- сохранение файла: в меню *File* необходимо выбрать *Save as...* дать имя файлу и указать место, где сохранить файл;
- импорт файла данных Excel (*.xls), dBase (*.dbf), ASCII (например, *.txt): в меню *File* необходимо выбрать *Import Data* (импорт данных);
- печать файла (в меню *File* необходимо выбрать *Print...*).
- Файлы данных в программе Statistica имеют расширение *sta*.

1.5. Визуальный анализ данных

Визуальные методы анализа данных чрезвычайно важны для предварительного исследования. Многие скрытые явления становятся отчетливыми, если для них найти подходящее графическое представление. Кроме того, многие сложные задачи решаются чрезвычайно простыми методами описательной статистики.

График – это чертёж, показывающий соотношение статистических величин при помощи разнообразных геометрических и изобразительных средств. В пакете Statistica графический анализ проводится через опцию *Graphs* (рис. 1.7).

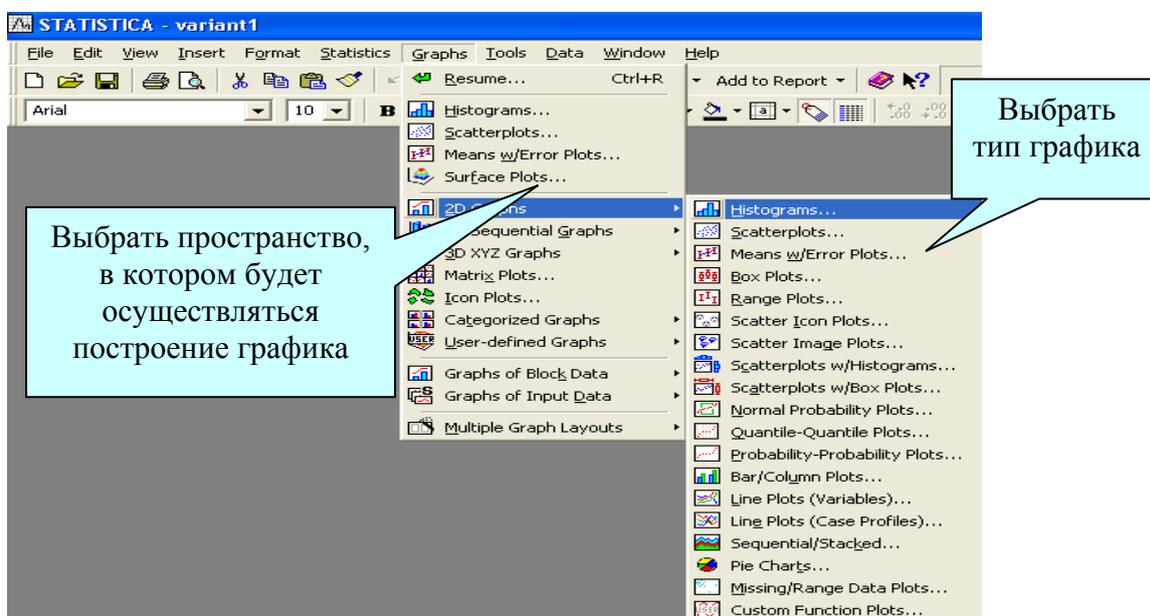


Рис. 1.7. Графический анализ в пакете Statistica

Опция Graphs позволяет построить различные виды графиков. Рассмотрим наиболее распространённые из них.

1.6. Диаграмма рассеяния

Диаграммой рассеяния называется представление элементов выборки как точек на плоскости. Диаграмма строится по команде *Graphs/Scatterplots*. В появившемся окне (рис. 1.8) необходимо нажать кнопку *Variables:* и указать переменные – аргумент и функцию. Во вкладке *Advanced* можно указать тип подгоночной функции (*Fit*) или отключить её (*Off*). Опция *Graph type:* позволяет построить множество графиков на разных (*Regular*) или одной (*Multiple*) сетке.

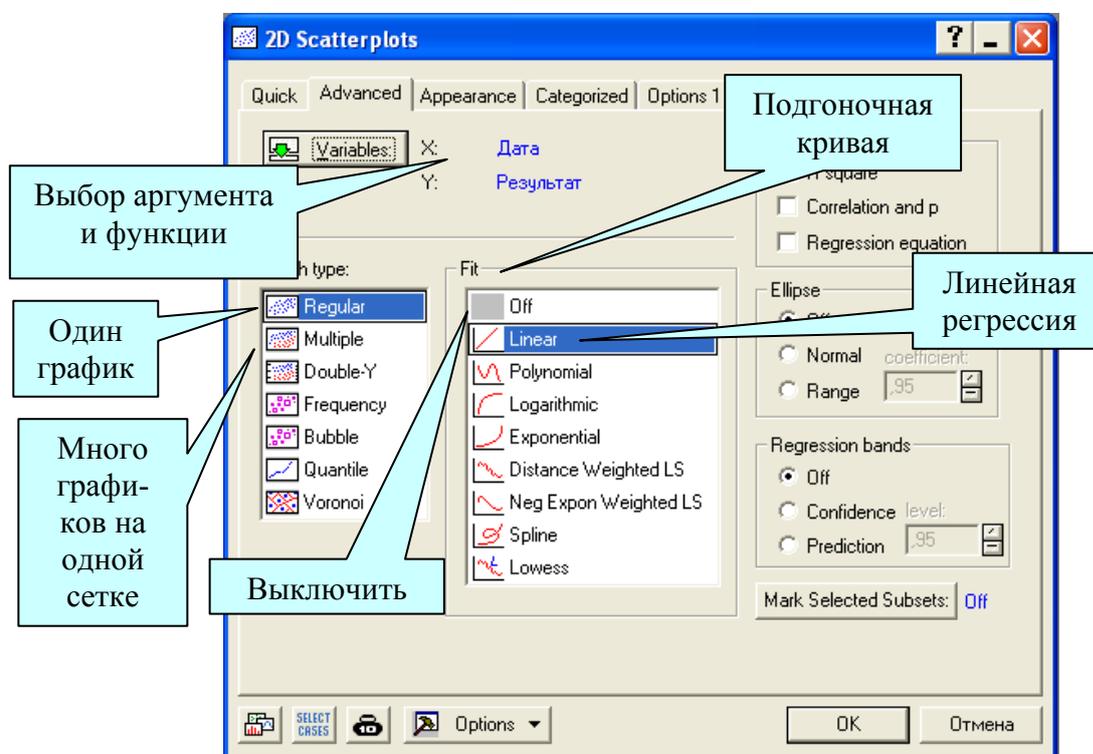


Рис. 1.8. Окно настройки графика

Построим диаграмму рассеяния для данных из табл. 1.1 (рис. 1.9).

Прямая на диаграмме рассеяния – это график простой линейной регрессии $y = -703,1678 + 0,0207x$.

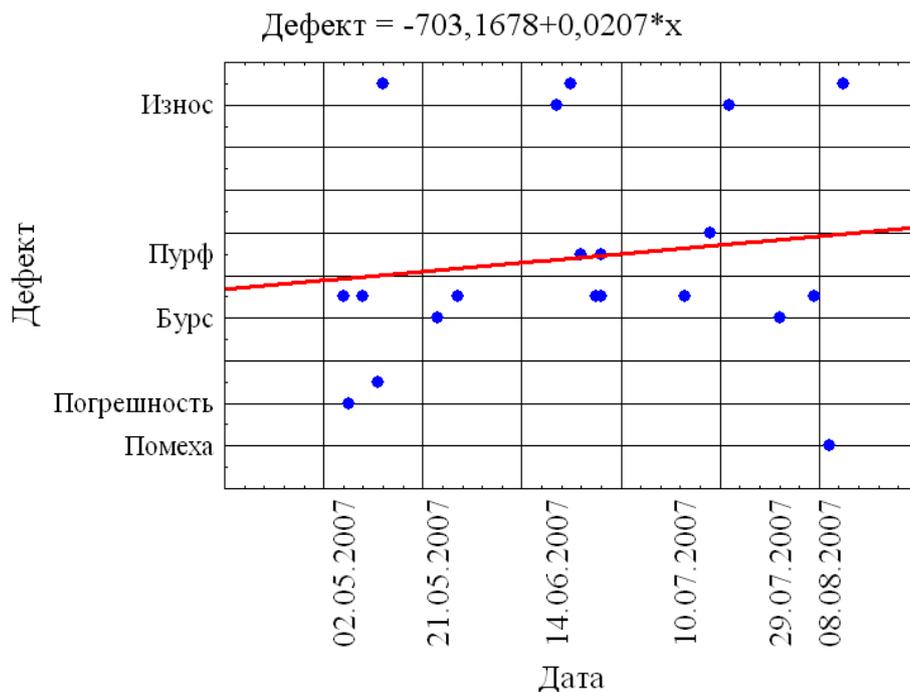


Рис. 1.9. Пример диаграммы рассеяния

1.7. Трёхмерный визуальный анализ данных

Трёхмерный визуальный анализ позволяет наглядно анализировать данные в трёхмерном пространстве, например, строить трёхмерное изображение последовательностей исходных данных (наблюдений) для одной или нескольких выбранных переменных (рис. 1.10). Трёхмерные представления значений каждой переменной не перекрываются, как на двухмерном графике, а строятся как значения какой-то поверхности в пространстве. С помощью трёхмерного визуального анализа можно обнаружить сложные нелинейные взаимосвязи между переменными.

Зная исходную функцию, можно построить любой график, как двумерный, так и трёхмерный. В отличие от большинства других типов графиков, для пользовательского графика не требуется выбирать переменные. Вместо этого для построения графика программа запросит ввод формулы. Построим, например, график в трёхмерном пространстве (рис. 1.10). Для построения графика необходимо ввести функцию в поле *Function* (рис. 1.11). Получившийся график представлен на рис. 1.12.

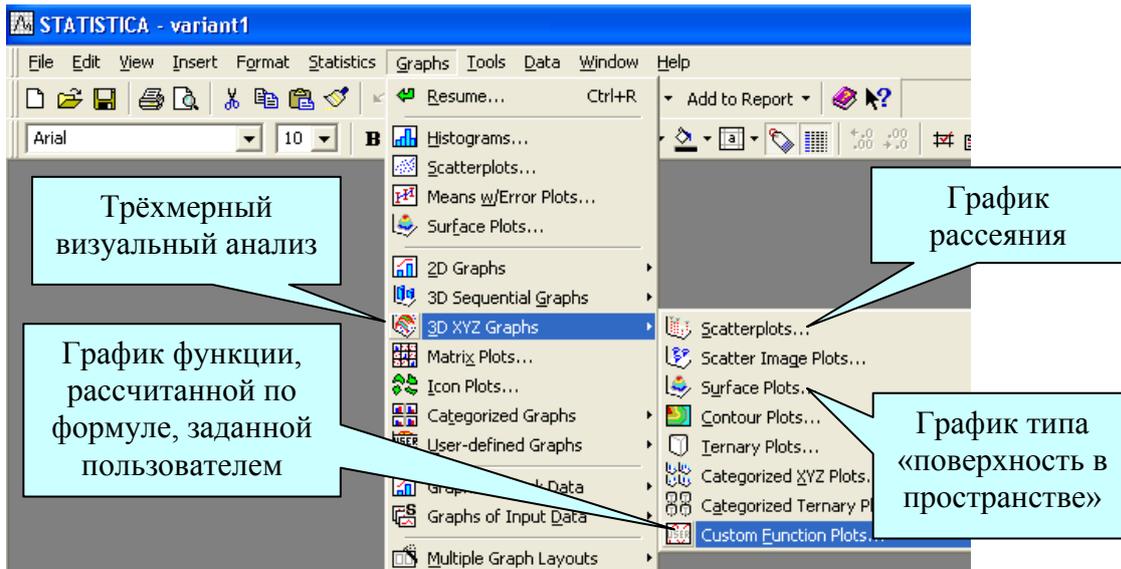


Рис. 1.10. Построение пользовательских графиков

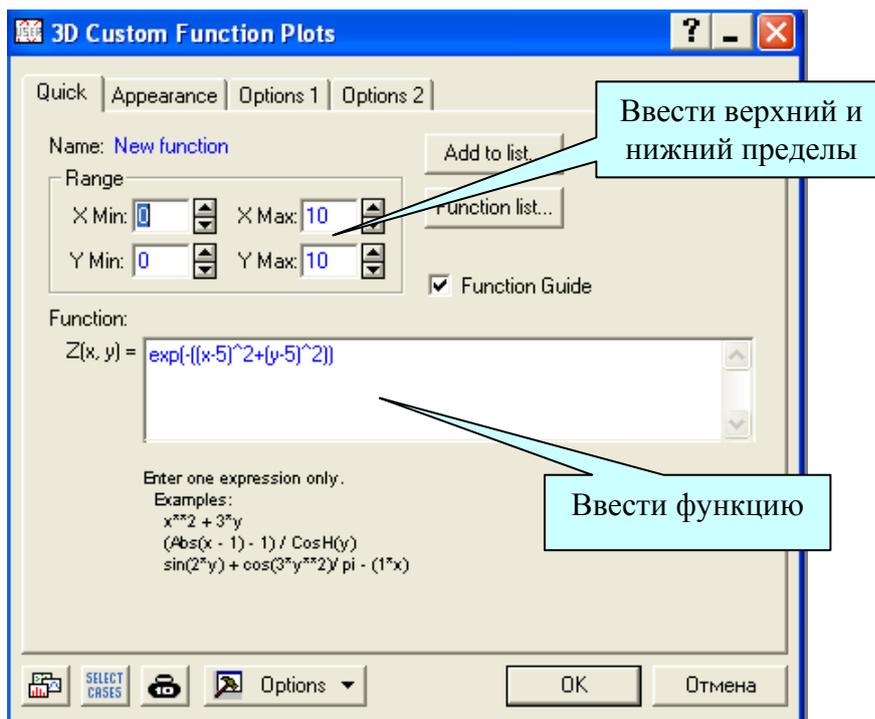


Рис. 1.11. Окно для построения графика

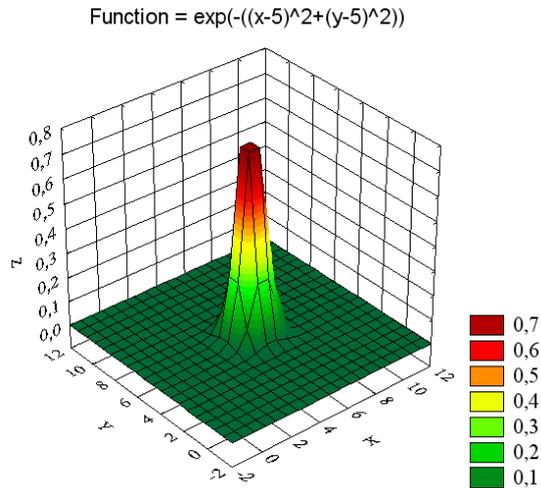


Рис. 1.12. Трёхмерный график

1.8. Круговые диаграммы

Круговые диаграммы весьма показательны, когда количество данных не велико (рис. 1.13). Данные представлены в процентах.

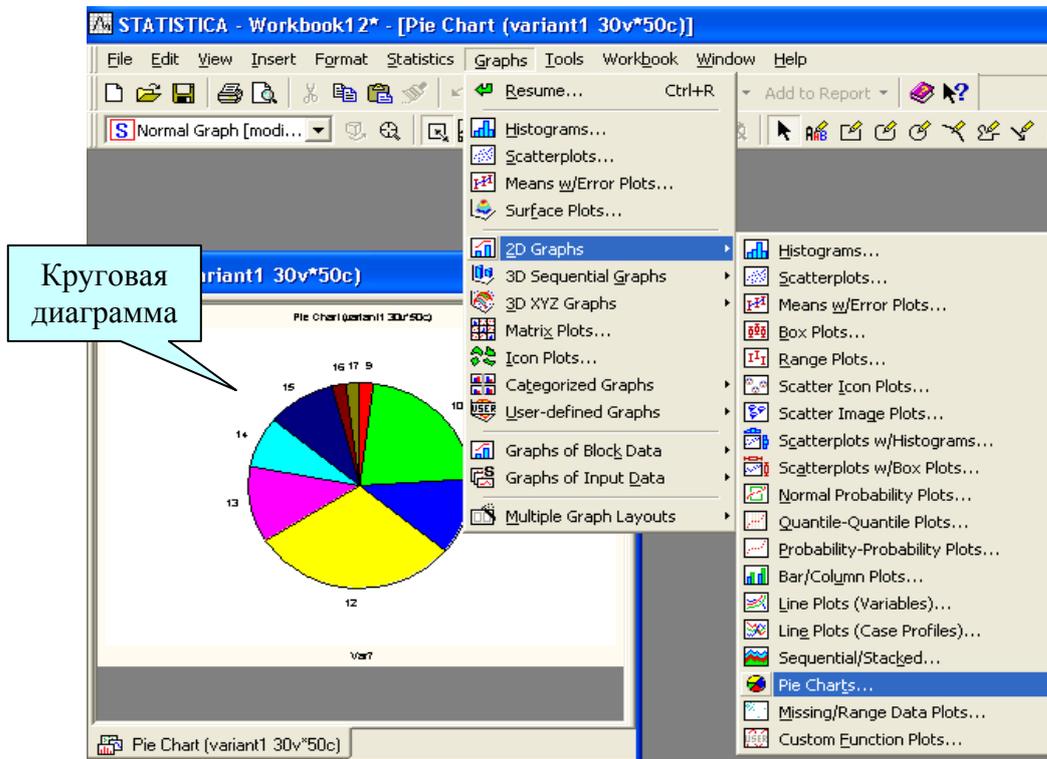


Рис. 1.13. Построение круговой диаграммы

Построим круговые диаграммы для примера (табл. 1.1), указав в окне выбора переменных сразу две переменные: «Установка» и «Дефект». Диаграммы показаны на рис. 1.14–1.15. По построенным круговым диаграммам легко определить установку, вызвавшую большинство ремонтных остановок за эксплуатационный период май–август 2007 года. Это ТВА160 – 35 % от всех остановок заводов.

1.9. Построение гистограмм

Любой производственный процесс характеризуется определённым распределением заданного показателя качества продукции [6]. Это распределение обусловлено его физической природой, условиями реализации и множеством случайных и неслучайных факторов, действующих на процесс. Знание реального процесса позволяет принимать необходимое управленческое решение с целью выявления и последующего устранения причин, ухудшающих характеристики процесса и с целью его дальнейшего совершенствования.

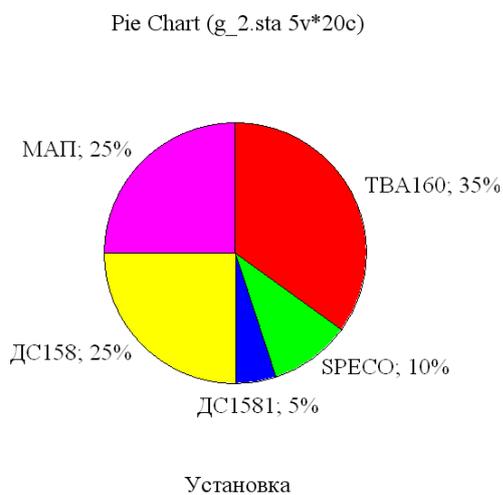


Рис. 1.14. Круговая диаграмма для переменной «Установка»



Рис. 1.15. Круговая диаграмма для переменной «Дефект»

Представление о реальном процессе можно получить путём построения гистограмм. Часто первый шаг визуального анализа данных состоит в построении гистограмм для всех переменных.

Гистограмма строится для исследуемого показателя на основе выборки. Она состоит из прямоугольников, горизонтальные стороны которых равны частичным интервалам, а вертикальные – числу измеренных объектов, показатель которых попал в тот или иной интервал. Частичные интервалы – это малые отрезки, на которые разбивается область возможных значений показателя. Гистограммы позволяют увидеть, как распределены значения переменных по интервалам группировки, то есть как часто переменные принимают значения из различных интервалов. Гистограмма наглядно показывает, какие диапазоны значений исследуемой переменной являются наиболее частыми, насколько сильно они различаются между собой, как сконцентрировано большинство наблюдений вокруг среднего, является ли распределение симметричным или нет, имеет ли оно одну моду или несколько мод, то есть является ли мультимодальным.

Для построения гистограммы в программе Statistica можно воспользоваться командой *Graphs / 2D Graphs / Histograms* (рис. 1.16).

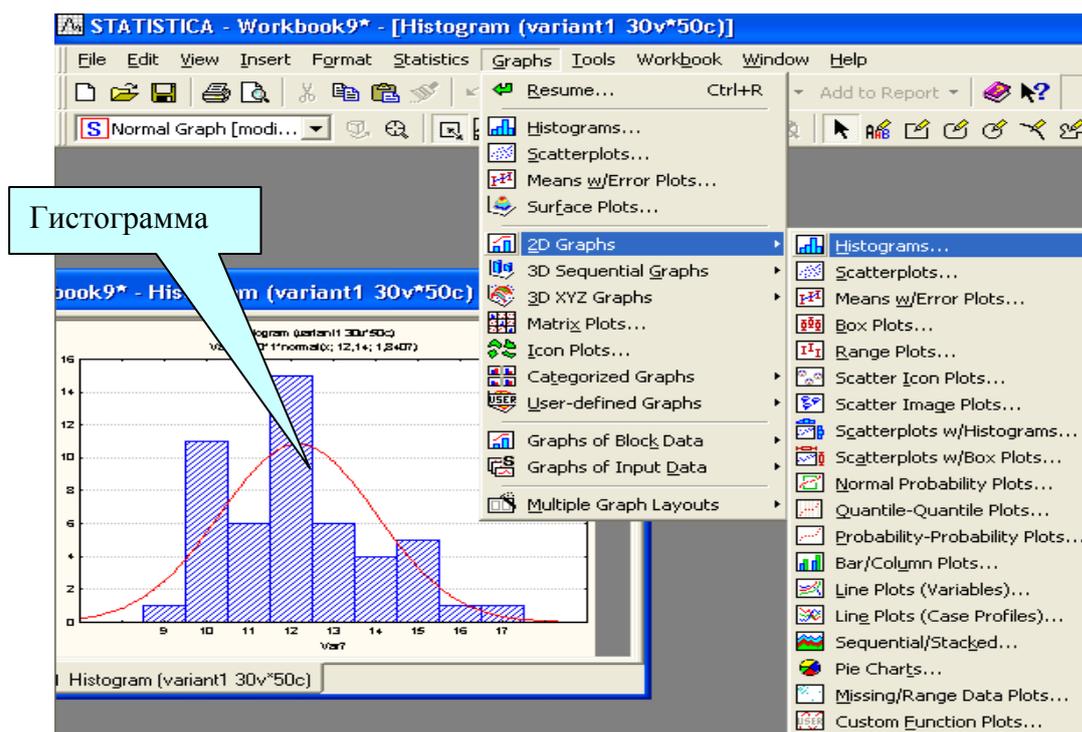


Рис. 1.16. Построение гистограммы

Раскроется диалоговое окно (рис. 1.17).

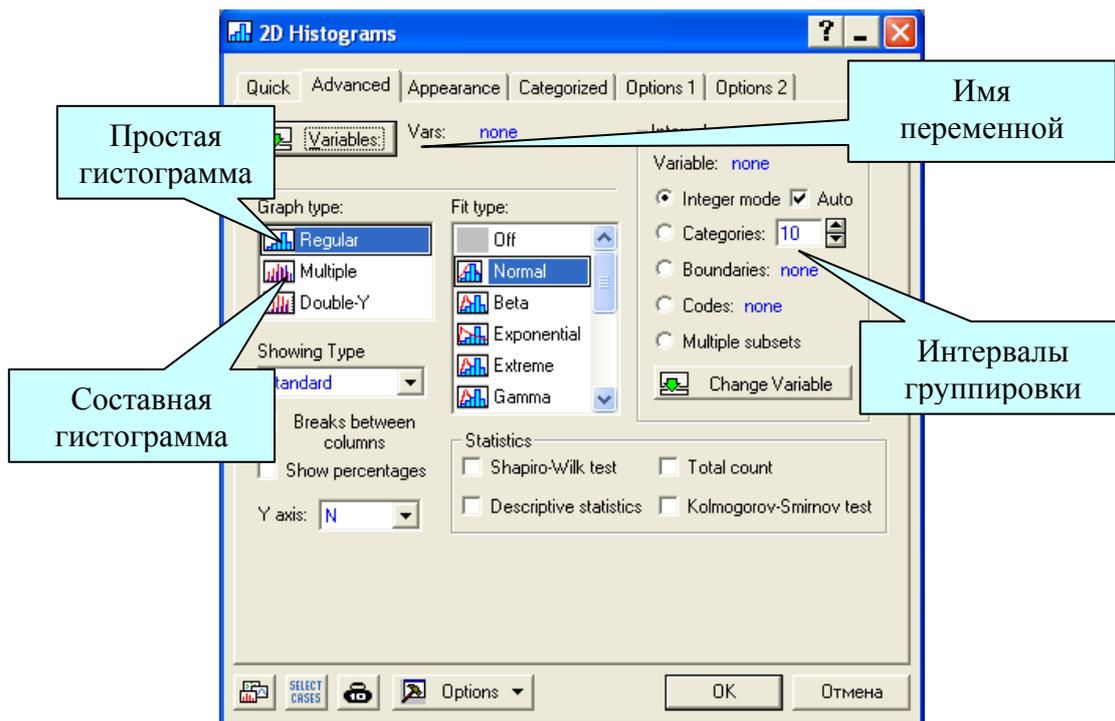


Рис. 1.17. Окно для построения гистограмм

Прежде всего, в этом окне следует определиться с именем переменной, для которой будет строиться гистограмма (*Variables:*). Результат выбора можно отследить в поле *Vars:*. Затем в поле *Graph Type:* следует выбрать графический тип гистограммы. По умолчанию установлен обычный (*Regular*).

Далее в поле *Categories* (классы) нужно определиться с типом классов группировки. Если данные *дискретны* (прерывисты, например, как у пуассоновской выборки), то кнопкой лучше задать режим *Integer Mode*. Если данные *непрерывны* (например, как у нормальной или равномерной выборки), то кнопкой следует выбрать режим *Categories:*. При отмеченном чекбоксе «» у статуса *Auto* это означает, что данные будут сгруппированы по классам одинаковой длины в пределах выборочного размаха. Число классов здесь нужно указать в соответствующем окне с цифрами. Как число, так и границы классов у гистограммы можно задать произвольными. Для этого следует использовать опцию *Boundaries*.

Наконец в поле *Fit Type:* по линейке прокрутки можно выбрать вид ожидаемой подгоночной кривой к графику гистограммы. К примеру, для нормальной выборки логично выбрать *Normal*; а для пуассоновской выборки логично указать *Poisson*; но для равномерной выборки следует

задать *Off*, поскольку форма плотности распределения здесь очевидна и отображать её не следует. Гистограмма строится по клавише *OK* в правом нижнем углу окна (рис. 1.17).

Построим гистограмму для переменной «Дата» из табл. 1.1.

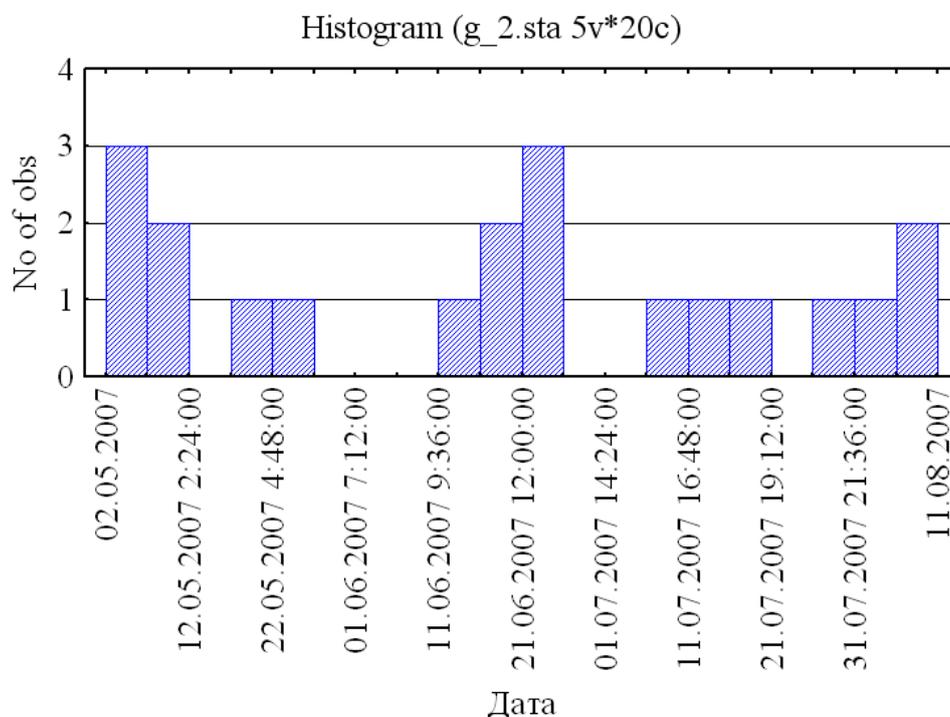


Рис. 1.18. Гистограмма для переменной «Дата»

Из рис. 1.18 видим, что наибольшее число поломок оборудования было в первых числах мая и в последнюю неделю июня. Хорошо погуляли!

С помощью специальных статистических критериев, например, с помощью критерия хи-квадрат, можно удостовериться, насколько правилен этот вывод. В данном примере различие между случаями отказов оборудования небольшое, но и число наблюдений мало. Если бы подобное различие имело место для 100 дней, то, очевидно, мы отнесли бы его на счёт случайной ошибки и не приняли бы во внимание.

В разведочном анализе данных гистограмма – обязательный шаг. Гистограмма представляет интерес по следующим причинам:

- по форме распределения можно охарактеризовать природу исследуемой переменной (например, бимодальность распределения может означать, что выборка неоднородна и состоит из наблюдений, принадлежащих двум различным множествам;

- многие статистики критериев основаны на определенных предположениях о виде распределения. Гистограммы помогают визуально проверить выполнение этих предположений.

Построение гистограммы является быстрым и наглядным методом получения информации о виде и характере процесса на основе относительно небольшого объёма выборки. Это обуславливает его широкое применение при анализе, настройке и наладке процесса и позволяет в дальнейшем в случае необходимости применять по отношению к процессу целенаправленные предупреждающие или корректирующие управляющие воздействия.

1.10. Задания для самостоятельной работы

Задание 1. Визуализация значений заданных переменных с использованием статистических графиков.

Для файла с исходными данными <http://ieeetpu.ru/statlab/var1.sta> построить гистограммы (команда *Graphs/Histograms*) для двух переменных на одном графике в зависимости от номера вашего варианта N: VarN, VarN+1. Использовать опцию *Multiple* (несколько графиков на одной сетке) во вкладке *Quick*.

Для тех же переменных построить столбчатую диаграмму (*Graphs/2D Graphs/Bar Columns Plots*).

Для переменной VarN построить круговую диаграмму (*Pie Chart - Counts*). Обратить внимание, как строится график *Pie Chart* при изменении переменной *Categories*.

Для переменных VarN, VarN+1, VarN+2 построить 3D график (*Graphs/3D XYZ Graphs/Surface Plots*). Обратить внимание, что трёхмерный график можно разворачивать на любой угол в подменю «свойства графика / все свойства». Настроить графики, подписав переменные и оси.

Задание 2. Решить с помощью вероятностного калькулятора следующую задачу.

Известно, что в некоторой стране рост мужчин приближенно имеет нормальное распределение со средним 176 см и стандартным отклонением 7,63 см. Какова вероятность того, что рост случайно встреченного вами мужчины будет не менее 186 см?

Задание 3. Решить с помощью вероятностного калькулятора следующую задачу.

Вы попали в страну, где рост мужчин приближенно имеет нормальное распределение со средним 173 см и стандартным отклонением 8,65 см. Какова вероятность того, что рост случайно встреченного мужчины будет не менее 195 см?

Задание 4. Исходные данные (файл http://ieeee.tpu.ru/statlab/var_6_2.sta) представляют собой базу дефектных ведомостей по ремонту техники. Провести статистические испытания (на ваше усмотрение) и интерпретировать полученные результаты. Необходимо выявить:

1. Наиболее проблемное оборудование.
2. Машинистов, на долю которых приходится наибольшее число дефектов.
3. Периоды времени, когда появление дефектов наиболее вероятно.
4. Является ли появление дефектов абсолютно случайным или существует какая-то особая причина, требующая выявления и устранения?

ГЛАВА 2. ПЕРВИЧНАЯ ОБРАБОТКА ДАННЫХ И ВЫЧИСЛЕНИЕ ЭЛЕМЕНТАРНЫХ СТАТИСТИК

2.1. Вероятность и достоверность

Знание теории вероятностей, математической статистики и умение применить их на практике является одним из самых важных и полезных элементов образования. В огромном количестве случаев такие знания окажут неоценимую помощь в информационной работе, оградят от многих ошибок [8].

Основой статистических методов является теория вероятностей. Большинство из нас специально не изучало теории вероятностей и математической статистики и либо слабо разбирается, либо вовсе не знакомо с ней. Ближе всего мы подходим к статистике, изучая приближённые вычисления. К сожалению, на этом знакомство со случайностью в школе обычно и заканчивается. Учителя явно боятся знакомить детей с вопросами, на которые нельзя дать точный ответ.

«Сомнительно, чтобы где-нибудь, помимо банка, где клерки пересчитывают грязными руками чужие медяки, точность, на которую способна арифметика, имела какую-либо ценность»
В. Плэтт [8]

В современной жизни вряд ли найдется область, где нельзя было бы с пользой применить, пусть в самой простой форме, научной статистики. Кем бы вы ни были, если в процессе работы вам приходится интерпретировать фактический материал, вы можете обойтись без статистики, но её незнание отрицательно скажется на результатах вашей работы.

Многие ошибочно считают, что уяснение теории вероятностей им не под силу. Для того чтобы рассеять это заблуждение, следует напомнить, что существуют разные степени знакомства с теорией вероятностей, каждой из которых достаточно, чтобы с пользой применять эту теорию в работе.

Можно научиться «мыслить категориями теории вероятностей», уяснив смысл примерно двух десятков терминов, например, таких, как вероятность, кривая нормального распределения, среднее значение, медиана, мода, среднее квадратичное отклонение, средняя квадратичная

ошибка, случайная ошибка, дисперсия, корреляция, статистическая значимость. Особенно важно понять характер различия между средними величинами, а также такие термины, как квартиль, ошибки выборочного метода, доверительные интервалы и т. д. Таким путём можно получить представление о теории вероятностей и здраво судить о соответствующих понятиях. Хотя производить необходимые вычисления так научиться нельзя. Это более серьёзная степень владения предметом теории вероятностей. Третья, высшая степень – изучить или вновь освоить методы математического анализа, логику и математическую статистику так, чтобы стать специалистом в этой области и получить возможность справиться со многими трудностями, связанными с применением теории вероятностей в работе. «Мышление категориями теории вероятностей» и восприятие мира через призму статистики помогает вырабатывать правильное представление о явлениях, которые мы изучаем, и является ценным методом решения многих задач.

2.2. Генеральная совокупность и выборка

Множество всех обследуемых объектов называется генеральной совокупностью. Если это множество содержит небольшое число элементов, то возможно полное обследование всех его элементов. Однако в большинстве случаев в силу того, что генеральная совокупность имеет очень много элементов либо её элементы труднодоступны, либо по другим причинам обследуется некоторая часть генеральной совокупности – выборка. В этом случае основные характеристики генеральной совокупности оцениваются (то есть определяются приближённо) по выборке. Соответствующие статистики называются «выборочное среднее», «выборочная дисперсия» и т. д. Очевидно, что не всякая выборка правильно отражает свойства генеральной совокупности. Например, нельзя судить о среднем душевом доходе населения по выборке, составленной из доходов служащих финансовых компаний. Выборка должна давать правильное, неискажённое представление о генеральной совокупности, или, как говорят, должна быть репрезентативной. Для такой выборки представление о параметрах технологических процессов будет отражать реальное положение, если пропорции между вероятностями появления показателя качества продукции в выборке соответствуют пропорциям в генеральной совокупности.

Если свойства генеральной совокупности заранее неизвестны, то за неимением лучшего следует использовать простой случайный выбор.

Это означает, что все элементы генеральной совокупности должны иметь равные шансы попасть в выборку.

2.3. Простейшие описательные статистики

Так как значения переменных не постоянны, нужно описывать их изменчивость. Для этого придуманы описательные или дескриптивные статистики: минимум, максимум, среднее, дисперсия, стандартное отклонение, медиана, квартили, мода и так далее. Идея этих статистик очень проста: вместо того чтобы рассматривать все значения переменной, а их может быть очень много, вначале стоит посмотреть описательные статистики. Они дают общее представление о значениях, которые принимает переменная.

Расчёт описательных статистик производится при помощи модуля *Statistics/ Basic Statistics/ Tables*. В этом модуле объединены наиболее часто использующиеся на начальном этапе обработки данных процедуры. В стартовой панели модуля приводится перечень статистических процедур этого модуля (рис. 2.1).

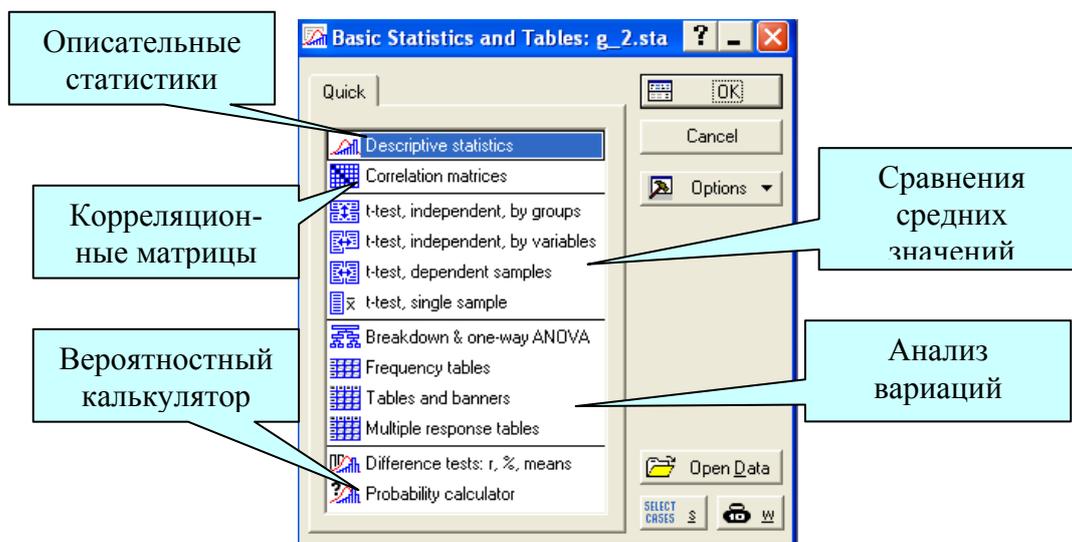


Рис. 2.1. Стартовое окно модуля с перечнем статистических процедур

При вызове модуля *Descriptive statistics* (Описательные статистики) появляется диалоговое окно (рис. 2.2), в котором при помощи кнопки *Variables* следует выбрать переменные для анализа. Для построения гистограмм и таблиц частот используются кнопки *Frequency tables* и

Histograms соответственно. Чтобы выбрать статистики, подлежащие вычислению, нужно воспользоваться вкладкой *Advanced* этого диалогового окна.

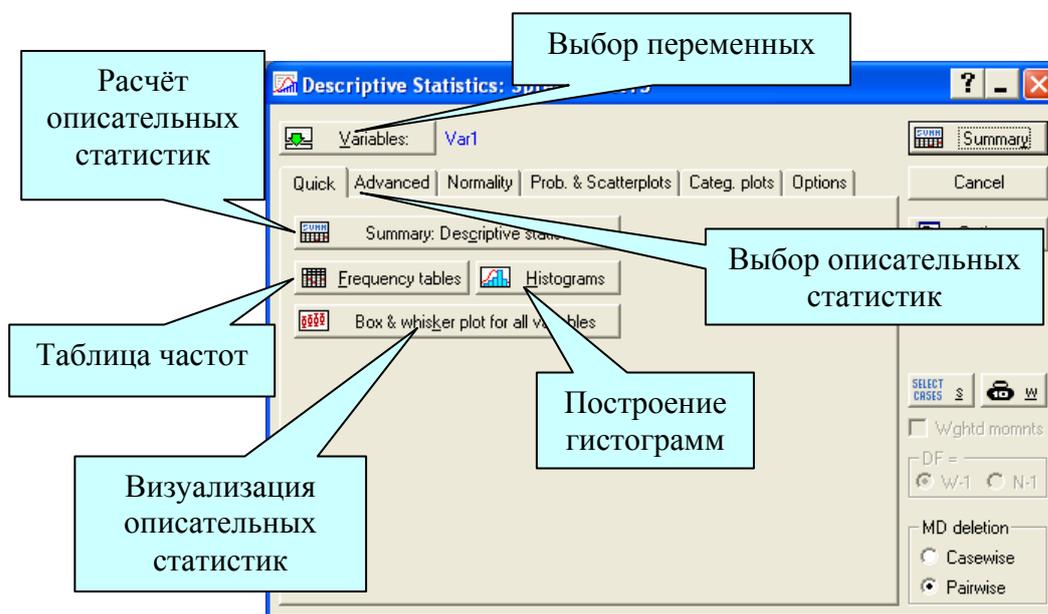


Рис. 2.2. Диалоговое окно модуля «Descriptive statistics»

Статистики, подлежащие вычислению, следует отметить. Возможен расчёт следующих описательных статистик [9].

Valid N – объем выборки.

Mean – среднее арифметическое. Это наиболее часто используемое среднее, поскольку в расчет здесь принимаются все без исключения значения. Часто называемое просто средним, среднее арифметическое определяется как сумма наблюдений, делённая на их количество.

Среднее значение случайной величины представляет собой наиболее типичное, наиболее вероятное её значение, своеобразный «центр», вокруг которого разбросаны все значения признака. Точно так же, как люди могут иметь различные мнения по поводу местонахождения центра города, есть и различные способы оценки среднего значения набора данных [1]. Примерами различных типов средних значений служат среднее арифметическое, полусумма крайних значений, медиана, мода, геометрическое среднее и гармоническое среднее.

Median – медиана. Медианой является такое значение случайной величины, которое разделяет все наблюдения выборки на две равные по численности части.

Медиана – это величина, находящаяся посередине набора данных, когда в нём все наблюдения упорядочены по возрастанию; если число наблюдений чётно, то имеются два «срединных» значения, и медиана равна их полусумме. Мода представляет собой наиболее часто встречающееся значение, и поэтому в некоторых наборах данных могут быть две или более моды, имеющие одну и ту же частоту.

Квантиль – такая величина, что заданная часть наблюдений меньше этой величины или равна ей. Для медианы эта часть равна одной второй **Sum** – сумма.

Standard Deviation – стандартное отклонение.

Стандартное отклонение (или среднее квадратическое отклонение) является мерой изменчивости (вариации) признака. Оно показывает, на какую величину в среднем отклоняются наблюдения от среднего значения признака.

Variance – дисперсия.

Дисперсия является мерой изменчивости, вариации признака и представляет собой средний квадрат отклонений наблюдений от среднего значения признака. В отличие от других показателей вариации дисперсия может быть разложена на составные части, что позволяет тем самым оценить влияние различных факторов на вариацию признака. Дисперсия – один из важнейших показателей, характеризующих явление или процесс, один из основных критериев возможности создания достаточно точных моделей.

Standard error of mean – стандартная ошибка среднего.

Стандартная ошибка среднего – это величина, на которую отличается среднее значение выборки от среднего значения генеральной совокупности при условии, что распределение близко к нормальному.

95 % confidence limits of mean – 95%-й доверительный интервал для статистического анализа.

Это интервал, в который с вероятностью 0,95 попадает среднее значение признака генеральной совокупности. Интервал выбирается при помощи вкладки *Categ. plots* (рис. 2.2). Наиболее часто используется вероятность 0,95 (95 %). Вероятности 0,95 соответствует уровень значимости 0,05 (5 %), установленный по умолчанию.

Minimum, maximum – минимальное и максимальное значения.

Очень часто решающее значение имеют крайние величины, как самые высокие (максимальная нагрузка), так и самые низкие (самое слабое звено цепи)

Lower, upper quartiles – нижняя и верхняя квартили.

Квартилями называются такие величины Q_1 и Q_3 , что одна четвертая часть наблюдений меньше или равна Q_1 и три четверти наблюдений меньше или равны Q_3 . Ясно, что мы можем подобным образом определить и величину Q_2 , которая в этом случае является медианой. Часто величину Q_1 называют нижней квартилью, а величину Q_3 – верхней. Разность между ними называется интерквартильной шириной.

Quartile range – интерквартильная широта.

Range – размах.

Расстояние между наибольшим (maximum) и наименьшим (minimum) значениями признака. Размах – одна из ключевых характеристик при оценке качества процессов.

Skewness – асимметрия.

Асимметрия характеризует степень смещения вариационного ряда относительно среднего значения по величине и направлению. В симметричной кривой коэффициент асимметрии равен нулю. Если правая ветвь кривой, начиная от вершины) больше левой (правосторонняя асимметрия), то коэффициент асимметрии больше нуля. Если левая ветвь кривой больше правой (левосторонняя асимметрия), то коэффициент асимметрии меньше нуля.

Standard error of Skewness – стандартная ошибка асимметрии.

Kurtosis – эксцесс.

Эксцесс характеризует степень концентрации случаев вокруг среднего значения и является своеобразной мерой крутизны кривой. В кривой нормального распределения эксцесс равен нулю. Если эксцесс больше нуля, то кривая распределения характеризуется островершинностью. При отрицательном эксцессе кривая более пологая по сравнению с нормальным распределением.

Standard error of Kurtosis – стандартная ошибка эксцесса.

К сожалению, пакет Statistica не рассчитывает такие часто применяемые статистики, как коэффициент вариации и относительная ошибка среднего значения. Но их определение не составляет большого труда [9].

Коэффициент вариации есть отношение стандартного отклонения к среднему значению, умноженное на 100%:

$$\text{Коэффициент Вариации} = \frac{\text{StandardDeviation}}{\text{Mean}} \cdot 100\%$$

Коэффициент вариации, как дисперсия и стандартное отклонение, является показателем изменчивости признака. Коэффициент вариации не зависит от единиц измерения, поэтому удобен для сравнительной оценки различных статистических совокупностей. При величине коэффициента вариации до 10 % изменчивость оценивается как слабая, 11–25 % – средняя, более 25 % – сильная.

Относительная ошибка среднего значения (%) – отношение стандартной ошибки среднего к среднему значению, умноженное на 100 % (для вероятности 0,68):

$$\text{Относительная Ошибка Среднего Значения} = \frac{\text{StandardErrorOfMean}}{\text{Mean}} \cdot 100\%$$

Это процент расхождения между генеральной и выборочной средней, показывает, на сколько процентов можно ошибиться, если утверждать, что генеральная средняя равна выборочной средней. Если относительная ошибка не превышает 5 %, то точность исследований (точность опыта) оценивается как хорошая, до 10 % – удовлетворительная.

Точность 3–5 % является вполне достаточной для большинства задач.

2.4. Примеры вычисления описательных статистик

Приведём пример вычисления описательных статистик для любимой около-экономическими кругами задачи – «расчёта среднего дохода» на предприятии, в регионе, в стране. Для эксперимента возьмём следующую выборку (рис. 2.3). Мы будем условно считать, что вышли на улицу и спросили первых 12 попавшихся нам человек о размере их дохода (в условных денежных единицах) предполагая, что 12 наблюдений – это репрезентативная выборка и её вполне достаточно для формулировки выводов. Здесь же дана сравнительная характеристика среднего значения, медианы и моды в виде схемы из книги [8].

Рассчитаем среднее арифметическое, являющееся в данном случае оценкой математического ожидания, медиану, моду и квартили. Результаты расчётов в системе Statistica приведены на рис. 2.4.

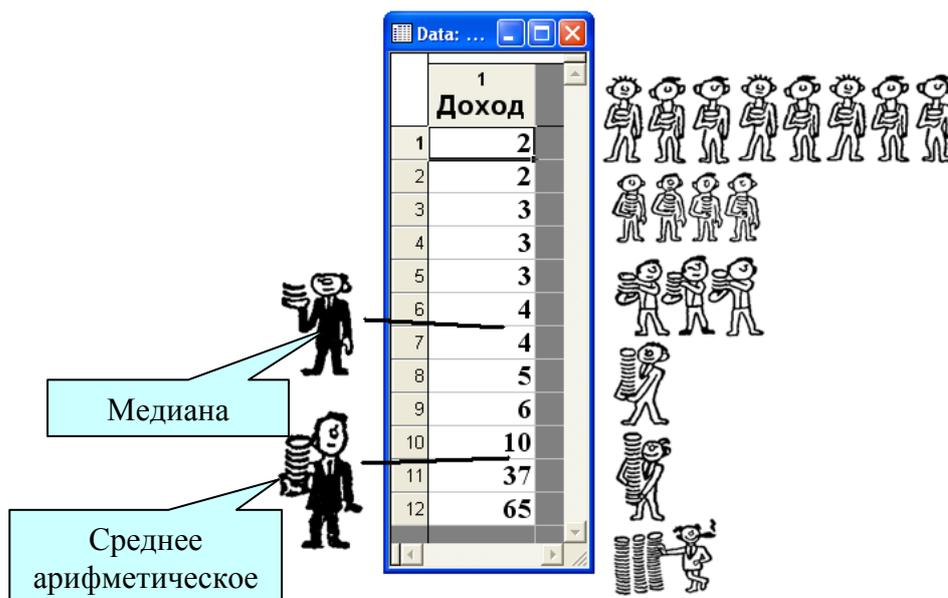


Рис. 2.3. Данные для эксперимента по расчёту описательных статистик

Variable	Mean	Geometric Mean	Median	Mode	Frequency of Mode	Lower Quartile	Upper Quartile	Range	Quartile Range	Std.Dev.
Доход	12,00	5,73	4,00	3,00	3	3,00	8,00	63,00	5,00	19,29

Рис. 2.4. Описательные статистики для примера из рис. 2.3

Рассуждения, основанные на средних значениях, зачастую поверхностны [8]. «Средний доход» ничего не говорит о реальном положении дел. Он составил по расчёту 12 условных денежных единиц, в то время как нищее большинство – четверть всей выборки имеет доход ниже 3,0 (нижняя квартиль), а три четверти – ниже 8,0. Такая характеристика, как медиана гораздо лучше описывает нашу выборку, свидетельствуя, что половина населения в выборке имеет доход ниже 4,0. Это опять таки меньше «среднего дохода». Но если уж вычислять среднее, то хотя бы среднее геометрическое, а не арифметическое. Существуют две причи-

ны использования «среднего дохода» в газетно-телевизионной похвальбе чиновников друг перед другом:

1. Это делается случайно: население статистически неграмотно и понимает только «среднее арифметическое».

2. Это делается специально: неграмотными проще управлять.

Автор склоняется ко второму варианту, но не навязывает его читателю. Читатель знает, что такое «медиана» и обладает статистическим мышлением.

2.5. Визуализация описательных статистик

Для визуализации описательных статистик можно построить «графики коробок» («ящики с усами»). С помощью этого графика можно быстро оценить данные на предмет структуры распределения, наличия неправдоподобных измерений, однородности наблюдений и так далее.

Это легко можно сделать при помощи кнопки *Box & Whisker plot for all variables* окна *Descriptive statistics* (рис. 2.2). Предварительно необходимо обратиться к вкладке *Options* и установить одно из четырёх положений:

Median/Quart./Range – Медиана / Квартили / Размах;

Mean/SE/SD – Среднее / Ошибка среднего / Стандартное отклонение;

Mean/SD/1.96SD – Среднее / Стандартное отклонение / Интервал $1,96 \cdot$ стандартного отклонения;

Mean/SE/1.96*SE – Среднее / Ошибка среднего / Интервал $1,96 \cdot$ ошибки среднего.

Визуализация описательных статистик рассматриваемого примера при помощи графика «ящик с усами» для среднего и медианы представлена на рис. 2.5 и 2.6 соответственно.

Как видно, в этом случае построение графика для среднего значения не очень показательно. Правильнее сказать, из него ничего не видно. По рис. 2.6 можно определить, что выборка крайне неоднородна. Наличие длинного верхнего уса говорит о том, что людей с высоким доходом, вытягивающих статистику «среднего дохода» вверх, очень немного. Основная масса нищеты – это 75 % населения (три квартиля снизу до 8 денежных единиц). Половина выборки имеет доход от 3 до 8 денежных единиц, что опять таки ниже «среднего дохода».

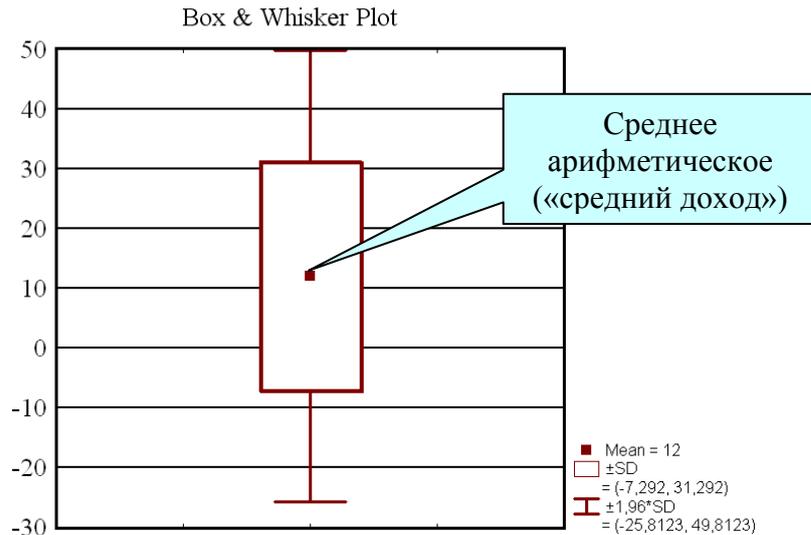


Рис. 2.5. Описательные статистики для среднего значения

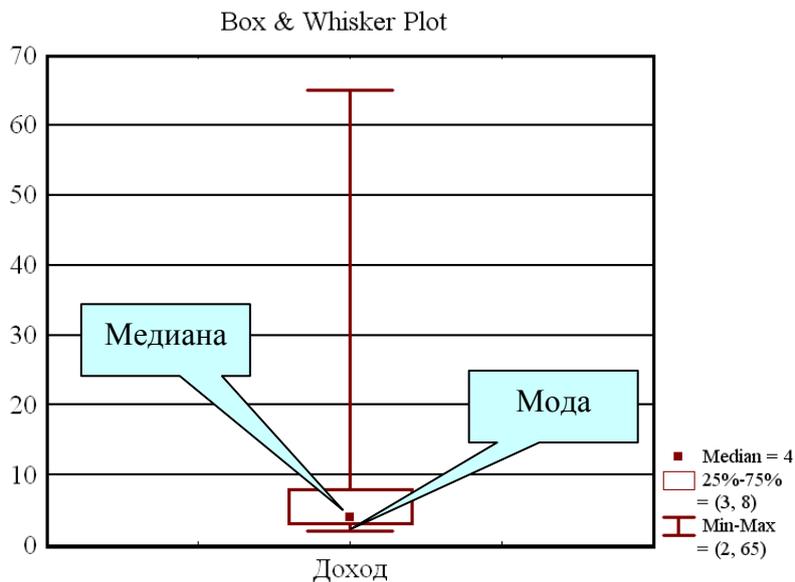


Рис. 2.6. Описательные статистики для медианы

Тот, кто знаком с теорией вероятностей, понимает, что медиана или мода лучше выражают тенденцию повторяемости большого количества величин, чем среднее арифметическое значение.

В модуле описательных статистик можно представить распределение переменных на гистограммах. Для этого предназначена кнопка *Histograms*. Для нашего примера с доходом (рис. 2.3) гистограмма изображена на рис. 2.7. Гистограмма подтверждает выводы, сделанные из

графиков коробок о неправомерности оценки выборки только по одному среднему арифметическому значению.

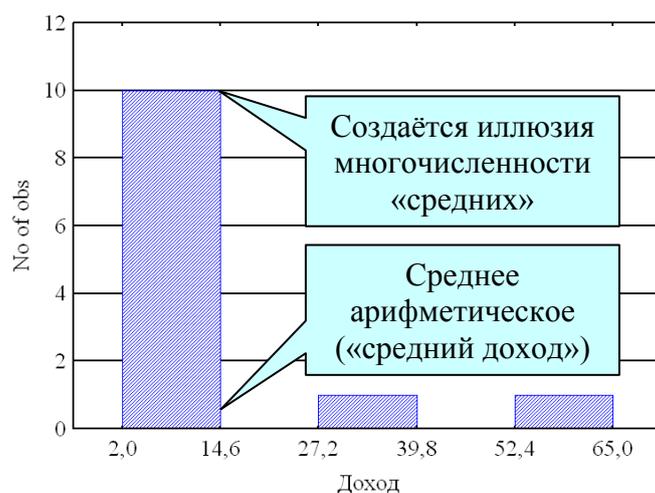


Рис. 2.7. Гистограмма для данных из рис. 2.3

На гистограмму при необходимости можно наложить плотность нормального распределения, проверить близость распределения к нормальному виду при помощи критериев Колмогорова–Смирнова, Лиллиефорса; вычислить статистику Шапиро–Уилкса. Для этого в группе опций *Distribution* необходимо установить флажок напротив соответствующих статистик. Значения статистик показываются прямо на гистограммах.

Чем меньше величина статистики Колмогорова–Смирнова, тем ближе распределение случайной величины к нормальному.

2.6. Правило трёх частей

Автор не придумал ничего лучше, чем просто процитировать аналогичный раздел из книги [8], сократив его.

«Впервые я услышал о правиле трёх частей от одного работника, который думал, что он сделал великое открытие. В обязанности этого работника входило помогать фермерам района лучше вести хозяйство – использовать для посева лучшие сорта пшеницы, применять больше удобрений и т. п. Он проводил беседы, распространял специальную литературу и другими способами пропагандировал передовые методы ведения сельского хозяйства. Он рассказал мне, что, несмотря на все ста-

рания, ему никогда не удавалось убедить больше одной трети наиболее передовых фермеров принять его рекомендации. Последняя часть наиболее отсталых фермеров не желала вводить никаких улучшений. Рано или поздно в силу определённых экономических факторов последние разорялись и лишались своих ферм. В другой раз такие же соображения высказал мне один опытный профессор университета. В первые годы после второй мировой войны в колледж принимали слишком много студентов. При таком большом количестве студентов было трудно, а зачастую невозможно хорошо организовать преподавание. Профессор сказал мне: «В таких условиях вы не можете уделить достаточное внимание каждому студенту. Приходится расходовать своё время с наибольшей пользой. В самом начале работы выявите наиболее способных студентов, составляющих первую четверть группы. Этими студентами в дальнейшем можно не заниматься. Они в состоянии все усвоить сами и наверняка сдадут экзамены. Затем *как можно скорее выявите самых слабых студентов, составляющих последнюю четверть группы. На них не следует тратить время. Они не принесут славы ни вам, ни университету. Вероятно, им не удастся получить диплом инженера.*

Чтобы расходовать своё время с максимальным эффектом, тратьте его почти целиком на студентов со средними способностями, составляющими половину группы. Они нуждаются в вашей помощи и обладают достаточными способностями, чтобы извлечь из неё пользу».

Разделение любой изучаемой группы людей или организаций на три части практически полезно. С помощью этого метода мы выделяем в рамках любой группы людей, которые занимают положение в середине группы и со временем воспримут передовые методы, и, наконец, людей, плетущихся в хвосте. ***Рано или поздно в силу экономической или интеллектуальной конкуренции последняя часть группы исчезает.***

В течение столетия первый и второй законы термодинамики направляли развитие мысли в области естественных наук. Быть может, описанное нами правило трёх частей заслуживает того, чтобы его назвать первым законом человеческой динамики. Возможно, этот закон окажет аналогичное влияние на развитие общественных наук».

2.7. Нормальное распределение

Нормальное распределение, иногда называемое гауссовским, играет важную роль в статистике по многим причинам. Распределение

большого числа статистик является нормальным или может быть получено из нормального с помощью некоторых преобразований.

Нормальное распределение даёт хорошую модель для реальных явлений, в которых:

- имеется сильная тенденция данных группироваться вокруг центра;
- положительные и отрицательные отклонения от центра равновероятны;
- частота отклонений быстро падает, когда отклонения от центра становятся большими.

В основе нормального закона лежит центральная предельная теорема: «Если случайная величина X представляет сумму большого числа взаимно независимых случайных величин, влияние каждой из которых на всю сумму мало в сравнении с суммарным влиянием всех остальных, то X имеет распределение, близкое к нормальному».

Именно в силу этого все другие распределения стремятся к нормальному распределению с увеличением числа воздействующих на случайную величину факторов.

Точная форма нормального распределения $f(x)$ (характерная «колоколообразная кривая») определяется только двумя параметрами: средним μ и стандартным отклонением σ (рис. 3.11):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (3.1)$$

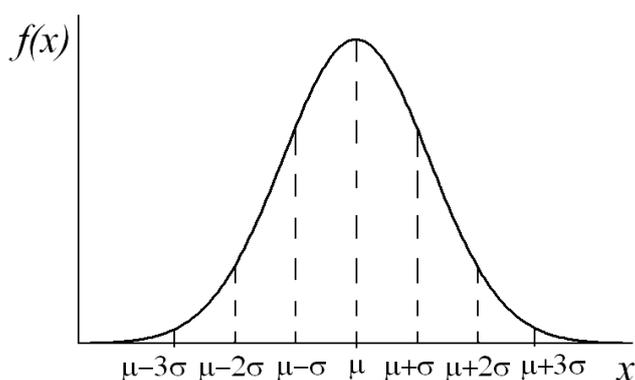


Рис. 2.8. Плотность вероятности нормального распределения

Множество величин на практике имеют нормальное распределение, например, распределение приращений индексов развитых стран, курсы акций, распределение погрешностей измерения, отклонение большинст-

ва параметров продукции от номинальных величин при её изготовлении и т. д.

Характерное свойство нормального распределения состоит в том, что 68,27 % из всех его наблюдений лежат в диапазоне одного стандартного отклонения от среднего $[\mu - \sigma, \mu + \sigma]$, диапазон два стандартных отклонения $[\mu - 2\sigma, \mu + 2\sigma]$ включает 95,45 % значений, диапазон три стандартных отклонения $[\mu - 3\sigma, \mu + 3\sigma]$ включает 99,73 % значений (рис. 3.11). Таким образом, за пределами $\pm 3\sigma$ относительно μ вероятность появления случайной величины не превышает значения 0,27 %. Это знаменитое правило «три сигма», чрезвычайно популярное на практике.

Обычно отдельные величины группируются вокруг определённого среднего значения и по мере удаления от него дисперсия всё более и более увеличивается. Величины, наиболее удалённые от среднего значения, могут существенным образом отличаться от основной массы величин данной группы. В каждом конкретном случае нужно чётко знать, что представляет интерес: основная масса величин или крайние для данной группы величины.

Степень отклонения крайних величин от среднего зависит обычно от трёх факторов:

- от состава выборки;
- от размера изучаемой выборки;
- от характера выборки.

О нормальности распределения можно судить по графику, который называется «нормальный вероятностный график». Его легко построить при помощи опции *Normal probability plots* окна «*Descriptive statistics*». Чем ближе распределение к нормальному виду, тем лучше значения ложатся на прямую линию (рис. 2.9).

Этот метод оценки является фактически глазомерным. В сомнительных случаях проверку на нормальность можно продолжить с использованием специальных статистических критериев (Колмогорова-Смирнова, хи-квадрат). Однако детальная проверка гипотезы о нормальности выборки требует довольно значительных объемов выборки (не менее 100 наблюдений).

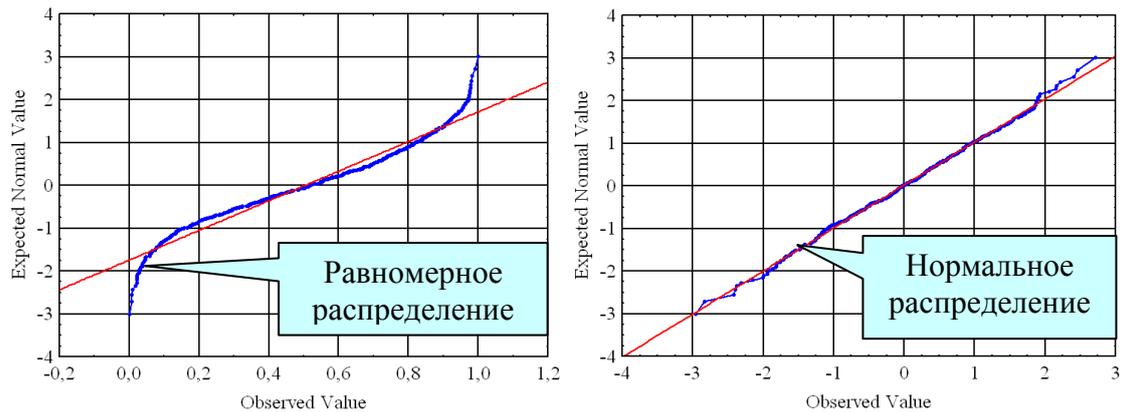


Рис. 2.9. Нормальный вероятностный график

2.8. Технологическое рассеяние и допуск на контролируемый показатель качества

Причиной появления того или иного значения случайной величины является то, что она формируется под воздействием большого числа влияющих факторов [6]. Промышленное производство связано с тем, что на его контролируемые показатели влияет множество факторов, неизбежным следствием чего является распределение показателей по нормальному закону. Так, при производстве деталей для машин, приборов и оборудования материалы могут иметь некоторый разброс в свойствах, например, иметь разные физические характеристики от партии к партии. Станки и оборудование каждый раз настраиваются с некоторыми вариациями, в процессе работы изнашиваются, а резцы тупятся. Процесс измерения параметра сопровождается погрешностями измерения, присущими как средствам и методам измерения, так и операторам. Внешние условия, в которых протекает процесс, могут испытывать колебания, например, изменяется температура, влажность, давление. В приёмах выполнения различных операций проявляются индивидуальные предпочтения операторов и т. д.

Все эти факторы оказывают влияние на контролируемые показатели качества продукции, которые будут распределены в соответствии с нормальным законом со средним μ и стандартным отклонением σ (рис. 2.8).

Одной из основных характеристик технологического процесса наряду с μ и σ является *поле рассеяния*, или *полное технологическое рассеяние*. Это область значений контролируемого показателя, в которой

он появляется с вероятностью, близкой к единице. Для нормального закона такой областью считают интервал $[\mu - 3\sigma, \mu + 3\sigma]$, в котором вероятность появления контролируемого показателя равна 0,9973. То есть поле рассеяния – это интервал, равный 6σ .

Любой контролируемый показатель продукции (признак качества) задаётся номинальным (расчётным или требуемым) значением показателя $x_{\text{НОМ}}$ и полем допуска $\Delta_{\text{ИЗД}}$ на этот показатель и определяется как

$$\Delta_{\text{ИЗД}} = x_{\text{В}} - x_{\text{Н}},$$

где $x_{\text{В}}$ и $x_{\text{Н}}$ – верхнее и нижнее допустимое значение показателя.

При этом требованием к процессу является совпадение центра поля рассеяния μ с номинальным значением $x_{\text{НОМ}}$ (рис. 2.10) для симметричного допуска или с центром поля допуска, если центр не симметричен относительно $x_{\text{НОМ}}$.

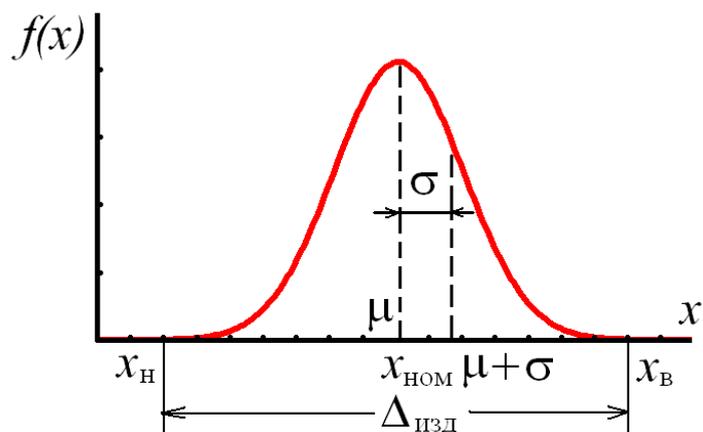


Рис. 2.10. Технологическое рассеяние и поле допуска на контролируемый показатель

Если $\mu = x_{\text{НОМ}}$, то максимальное число единиц продукции будет иметь значение контролируемого показателя, близкое к расчётному.

2.9. Настройка, наладка и качество технологических процессов

Полное технологическое рассеяние процесса может располагаться по отношению к допуску на показатель качества по-разному. Если центр поля рассеяния μ совпадает с $x_{\text{НОМ}}$, то говорят, что процесс *настроен* (рис. 2.11, а). Но такой процесс не гарантирует отсутствие бра-

ка, т. к. поле рассеяния, равное 6σ , превышает величину $\Delta_{\text{изд}}$. Процесс называется *налаженным*, когда поле рассеяния $6\sigma \leq \Delta_{\text{изд}}$.

Налаженный процесс тоже не гарантирует отсутствие брака, что показано на рис. 2.11, б. Чтобы процесс не давал брака, необходимо, чтобы он был настроен и налажен (рис. 2.11, в).

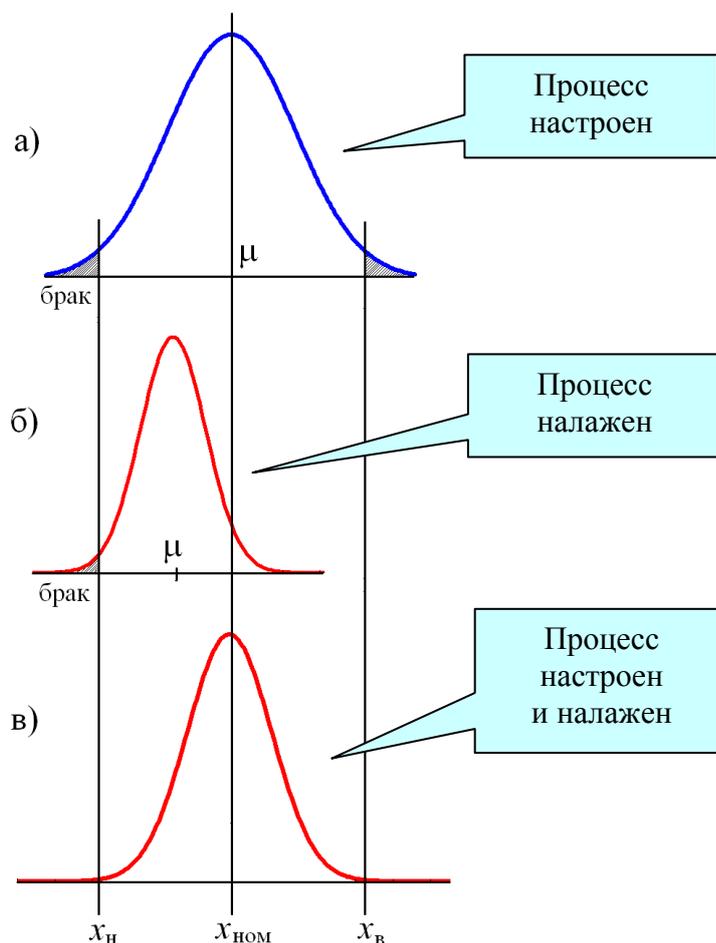


Рис. 2.11. Расположение технологического рассеяния относительно поля допуска

Количественной характеристикой качества процесса служит *коэффициент способности (точности)*, который определяется формулой

$$C_p = \frac{\Delta_{\text{изд}}}{6\sigma}.$$

При $C_p \geq 1,33$ процесс считается точным (способным). В этом случае допуск на признак качества $\Delta_{\text{изд}} \geq 8\sigma$ и полное технологическое рассеяние целиком лежит внутри поля допуска. Процесс в этом случае обла-

дает запасом точности, равным σ с каждой стороны поля допуска. Если в этом случае средство контроля выбрано так, что его погрешность не превышает 0,1–0,2 от допуска на параметр, то брак практически исключён.

При $1 \leq C_p < 1,33$ процесс удовлетворительный (условно способный). И хотя объективно в этом случае брак отсутствует, погрешности измерения при контроле продукции в некоторых случаях могут приводить к тому, что продукция, у которой значения показателя качества находятся вблизи границ допуска, может либо неправильно браковаться (годные в браке или ложный брак), либо неправильно приниматься (брак в годных или ложные годные).

При $C_p < 1$ процесс неудовлетворительный (неспособный), так как здесь полное технологическое рассеяние больше поля допуска, и брак в этом случае неизбежен.

2.10. Оценка качества технологических процессов в системе

Statistica

Оценка способности и качества производственного процесса легко проводится в системе Statistica. Для этого необходимо сформировать данные о процессе в виде измеряемых величин или их средних значений и запустить одну из процедур в меню *Statistics/ Industrial Statistics & Six Sigma/ Process Analysis/*. Далее необходимо ввести номинальные характеристики процесса – среднее значение, верхнюю и нижнюю границу. В результате будут рассчитаны характеристики процесса в соответствии с разд. 2.9 и коэффициент способности. Можно построить наглядную гистограмму с нанесёнными на неё характеристиками качества процесса. Интерфейс модуля оценки качества интуитивно понятен, и читатель без труда разберётся с ним самостоятельно.

2.11. Задания для самостоятельной работы

Задание 1. Исследование средних величин.

Конвертировать файл с данными в систему Statistica в соответствии с табл. 2.1.

Исследование средних величин

Номер варианта	Имя файла с данными	Используемые переменные
1	http://ieeep.tpu.ru/statlab/mmvb.txt	1 и 2 столбцы (ежедневное изменение индекса Московской межбанковской валютной биржи)
2	http://ieeep.tpu.ru/statlab/mmvb.txt	1 и 3 столбцы (ежедневное изменение стоимости чистых активов ПИФ ММВБ)
3	http://ieeep.tpu.ru/statlab/deposit.txt	1 и 2 столбцы (ежедневное изменение стоимости пая ПИФ депозитный)
4	http://ieeep.tpu.ru/statlab/deposit.txt	1 и 3 столбцы (ежедневное изменение стоимости чистых активов пая ПИФ депозитный)

Построить гистограмму для выбранной второй или третьей переменной. Сравнить построение гистограммы для разных значений интервалов группировки, которые можно изменить в окне построения гистограммы.

Задание 2. Составить новую таблицу, разделив выбранную переменную по месяцам. При этом каждая переменная новой таблицы будет соответствовать одному месяцу. С помощью модуля *Statistics/ Basic Statistics/ Tables* исследовать изменение среднего арифметического значения переменной и медианы. Усреднения проводить каждый месяц. Построить графики «ящик с усами» для всех средних значений и медиан (кнопка *Box & Whisker plot for all variables* окна *Descriptive statistics*). На графике соединить средние значения прямыми линиями. Сделать выводы.

ГЛАВА 3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

3.1. Статистические модели

Во многих случаях требуется решить, справедливо ли некоторое суждение. Если мы считаем, что исходные данные для таких суждений в той или иной мере носят случайный характер, то и ответы можно дать лишь с определённой степенью уверенности, и имеется некоторая вероятность ошибиться. Поэтому при ответе на подобные вопросы хотелось бы не только уметь принимать наиболее обоснованные решения, но и оценивать вероятность ошибочности принятого решения [10].

Весь статистический анализ основан на идее случайного выбора. Мы понимаем, что имеющиеся данные появились как результат случайного выбора из некоторой генеральной совокупности, нередко – воображаемой. Поскольку мы приняли вероятностную точку зрения на происхождение наших данных (т. е. считаем, что они получены путем случайного выбора), то все дальнейшие суждения, основанные на этих данных, будут иметь вероятностный характер. Всякое утверждение будет верным лишь с некоторой вероятностью. И с некоторой вероятностью оно может оказаться неверным.

Какую вероятность следует считать малой? Ответ зависит от того, какой опасностью грозит нам ошибка. При проверке статистических гипотез, например, полагают малыми вероятности, начиная с 0,05–0,01.

Рассмотрение вероятностных задач в строгой математической постановке приводит к понятию статистической гипотезы. В этой главе рассматриваются вопросы о способах проверки статистических гипотез в программе Statistica.

3.2. Статистические гипотезы

Термин «гипотеза» означает предположение, которое не только вызывает сомнения, но и которое мы собираемся в данный момент проверить.

Нулевая гипотеза H_0 – это гипотеза об отсутствии различий. Это то, что мы хотим опровергнуть, если стоит задача доказать значимость различий

Она содержит число 0: $x_1 - x_2 = 0$, где x_1 и x_2 – сопоставляемые значения признаков.

Альтернативная гипотеза H_1 – это гипотеза о значимости различий.

Это то, что мы хотим доказать, поэтому иногда её называют экспериментальной гипотезой

Бывают задачи, когда мы хотим доказать незначимость различий, то есть подтвердить нулевую гипотезу. Однако чаще требуется доказать значимость различий, ибо они более информативны в поиске нового.

Проверка гипотез осуществляется с помощью критериев статистической оценки различий.

3.3. Статистические критерии

Если гипотезу можно проверить непосредственно, не возникает никаких проблем. Но если прямого способа проверки нет, приходится прибегать к проверкам косвенным. Это значит, что приходится довольствоваться проверкой некоторых следствий, которые логически вытекают из гипотезы. Если некоторое явление логически неизбежно следует гипотезы, но в природе не наблюдается, то это значит, что гипотеза неверна. С другой стороны, если происходит то, что при гипотезе происходить не должно, это тоже означает ложность гипотезы. Заметим, что подтверждение следствия ещё не означает справедливости гипотезы, поскольку правильное заключение может вытекать и из неверной предпосылки. Поэтому косвенным образом доказать гипотезу нельзя, хотя опровергнуть – можно. Отсюда успех адвокатов.

Статистический критерий – это правило, по которому принимается решение о принятии истинной и отклонении ложной гипотезы с высокой вероятностью

Критерии делятся на параметрические и непараметрические.

Параметрические критерии – это критерии, включающие в формулу расчёта параметры распределения, то есть средние и дисперсии (t-критерий Стьюдента, критерий F и др.).

Непараметрические критерии – это критерии, не включающие в формулу расчёта параметров распределения и основанные на оперировании частотами или рангами (Q-критерий Розенбаума, критерий Уилкоксона и др.).

При нормальном распределении признака параметрические критерии обладают большей мощностью, чем непараметрические критерии. Они способны отвергать нулевую гипотезу, если она неверна. Поэтому во всех случаях, когда сравниваемые выборки взяты из нормально распределённых совокупностей, следует отдавать предпочтение параметрическим критериям.

Проверка данных на соответствие нормальному закону распределения очень важна для данных промышленной статисти-

В случае очень больших отличий распределений признака от нормального вида следует применять непараметрические критерии, которые в этой ситуации оказываются часто более мощными. В ситуациях, когда варьирующие признаки выражаются не в численной форме, применение непараметрических критериев оказывается единственно возможным.

3.4. Проверка гипотез с помощью критериев

Схема проверки гипотез с помощью статистических критериев состоит из следующих трёх шагов.

1. Вычисляется эмпирическое (или фактическое, реальное) значение критерия $F_{\text{эмп}}$. Вычисляется число степеней свободы и уровень значимости.

2. По таблицам критических значений для выбранного критерия находится так называемая критическая точка (или критическое значение) $F_{\text{кр}}$.

3. По соотношению эмпирического и критического значений критерия судят о том, подтверждается или опровергается нулевая гипотеза. Например, если $F_{\text{эмп}} > F_{\text{кр}}$, гипотеза H_0 отвергается.

В системе Statistica это делается автоматически.

В большинстве случаев для того, чтобы различия признавались значимыми, необходимо, чтобы эмпирическое значение критерия превышало критическое, хотя есть критерии (например, Манна–Уитни или

критерий знаков), в которых нужно придерживаться противоположного правила.

Число степеней свободы равно числу классов вариационного ряда минус число условий, при которых он был сформирован. К числу таких условий относятся объём выборки, средние и дисперсии.

Уровень значимости – это вероятность отклонения нулевой гипотезы, в то время как она верна

Обычно при проверке статистических гипотез принимают три уровня значимости: 5%-й (вероятность ошибочной оценки $\alpha = 0,05$), 1%-й ($\alpha = 0,01$) и 0,1%-й ($\alpha = 0,001$). В промышленной статистике часто считают достаточным 5%-й уровень значимости. При этом нулевую гипотезу не отвергают, если в результате исследования окажется, что вероятность ошибочности оценки относительно правильности принятой гипотезы превышает 5 %, т. е. $\alpha > 0,05$. Если же $\alpha < 0,05$, то принятую гипотезу следует отвергнуть на взятом уровне значимости. Ошибка при этом возможна не более чем в 5 % случаев, т. е. она маловероятна. При более ответственных исследованиях уровень значимости может быть уменьшен до 1 % или даже до 0,1 %.

В пакете Statistica значение задаваемого уровня значимости не используется. Как правило, в выходных данных содержатся выборочные значения статистики критерия и вероятность того, что случайная величина превышает это выборочное значение при условии, что верна гипотеза H_0 . Эта вероятность называется *p*-значением (*p*-level).

3.5. Ошибки при принятии гипотез

Ошибка, состоящая в том, что правильная гипотеза отклонена, в то время как она верна, называется ошибкой I рода
Ошибка, состоящая в том, что правильная гипотеза принята, в то время как она неверна, называется ошибкой II рода

При приёмочном контроле ошибка первого рода приводит к браковке партии с допустимой долей брака (риск производителя). При контроле производства – к вмешательству в налаженный процесс производства (ложная тревога). Ошибка второго рода приводит к принятию

партии с недопустимой долей брака (риск потребителя). При контроле производства – приводит к вмешательству в процесс производства, вышедший за допустимые границы.

3.6. Проверка гипотез о виде распределения

При проверке гипотез о параметрах генеральной совокупности контролируемого показателя предполагается, что закон распределения известен. Однако на практике это не всегда имеет место. И тогда необходимо определить, какому закону распределения подчиняется исследуемая случайная величина.

В конкретных задачах, как правило, всегда имеется некоторое основание предполагать, что закон распределения имеет определенный вид F (например, нормальный, Рэлея, Пуассона и т. д.). Это предположение может быть сделано, например, на основе построения гистограммы или на основе физического смысла исследуемого показателя.

В этом случае необходимо проверить гипотезу H_0 : генеральная совокупность распределена по закону F . Конкурирующей гипотезой будет гипотеза H_1 : генеральная совокупность не распределена по закону F .

Для решения этой задачи используют статистические критерии, называемые *критериями согласия*.

Теория вероятностей позволяет пользоваться несколькими критериями согласия: критерий Пирсона (критерий χ^2), критерий Колмогорова, Смирнова и др.

Здесь ограничимся только проверкой гипотез с помощью критерия Пирсона. Его достоинство по сравнению с другими критериями состоит в том, что он может быть применен к самым различным законам распределения, тогда как другие критерии применимы только к вполне определенным законам.

Пусть имеется выборка наблюдений случайной величины. Проверяется гипотеза H_0 , утверждающая, что случайная величина имеет функцию распределения $F(x)$. Проверка гипотезы H_0 при помощи критерия χ^2 в системе Statistica осуществляется по следующей схеме.

1. Формируются исходные данные, состоящие из n наблюдений одной переменной Var1. В качестве примера возьмём результаты измерения диаметров заклёпок – 200 наблюдений [11]:

13,39 13,33 13,56 13,38 13,43 13,37 13,53 13,40 13,25 13,37
13,28 13,34 13,50 13,38 13,38 13,45 13,47 13,62 13,45 13,39

13,53 13,58 13,32 13,27 13,42 13,40 13,57 13,46 13,33 13,40
 13,57 13,36 13,43 13,38 13,26 13,52 13,35 13,29 13,48 13,43
 13,40 13,39 13,50 13,52 13,39 13,39 13,46 13,29 13,55 13,31
 13,29 13,33 13,38 13,61 13,55 13,40 13,20 13,31 13,46 13,13
 13,43 13,51 13,50 13,38 13,44 13,62 13,42 13,54 13,31 13,58
 13,41 13,49 13,42 13,45 13,34 13,47 13,48 13,59 13,20 14,56
 13,55 13,44 13,50 13,40 13,48 13,29 13,31 13,42 13,32 13,48
 13,43 13,26 13,58 13,38 13,48 13,45 13,29 13,32 13,24 13,38
 13,34 13,14 13,31 13,51 13,59 13,32 13,52 13,57 13,62 13,29
 13,23 13,37 13,64 13,30 13,40 13,58 13,24 13,32 13,52 13,50
 13,43 13,58 13,63 13,48 13,34 13,37 13,18 13,50 13,45 13,60
 13,38 13,33 13,57 13,28 13,32 13,40 13,40 13,33 13,20 13,44
 13,34 13,54 13,40 13,47 13,28 13,41 13,39 13,48 13,42 13,46
 13,28 13,46 13,37 13,53 13,43 13,30 13,45 13,40 13,45 13,40
 13,33 13,39 13,56 13,46 13,26 13,35 13,42 13,36 13,44 13,41
 13,43 13,51 13,51 13,24 13,34 13,28 13,37 13,54 13,43 13,35
 13,52 13,23 13,48 13,48 13,54 13,41 13,51 13,44 13,36 13,36
 13,53 13,44 13,69 13,66 13,32 13,26 13,51 13,38 13,46 13,34

2. По команде *Statistics/ Distribution Fitting* в стартовом окне (рис. 3.1) выбираем вид случайной величины – непрерывная (*Continuous Distributions*, установлена по умолчанию) или дискретная (*Discrete Distributions*), вид распределения (по умолчанию предлагается нормальное), *OK*. Кнопкой *Variables* выбираем переменную.

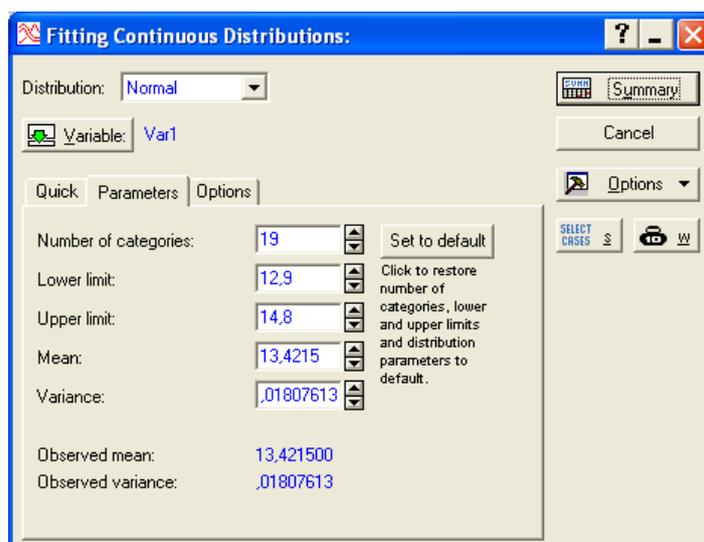


Рис. 3.1. Значения оценок параметров при проверке гипотезы

3. Во вкладке *Parameters* того же окна появятся оценки параметров. Число интервалов группировки (*Number of categories*) можно при необходимости изменить.

4. Нажмите кнопку *Summary*. На экран выводится таблица для расчёта статистики критерия – распределение случайной величины по интервалам. В таблице частот нужны столбцы *Observed Frequency* (наблюдаемые частоты) и *Expected Frequency* (ожидаемые частоты). Сравним графически наблюдаемые и ожидаемые частоты: запишем соответствующие столбцы в таблицу данных и построим график рассеяния (команды *Graphs/ Scatterplots/ Variables/ ОК*). Наблюдаем существенное различие между переменными, так как точки плохо укладываются на прямую (рис. 3.2). Если бы переменные были одинаковы, все наблюдения лежали бы на прямой с уравнением $Var2=Var1$.

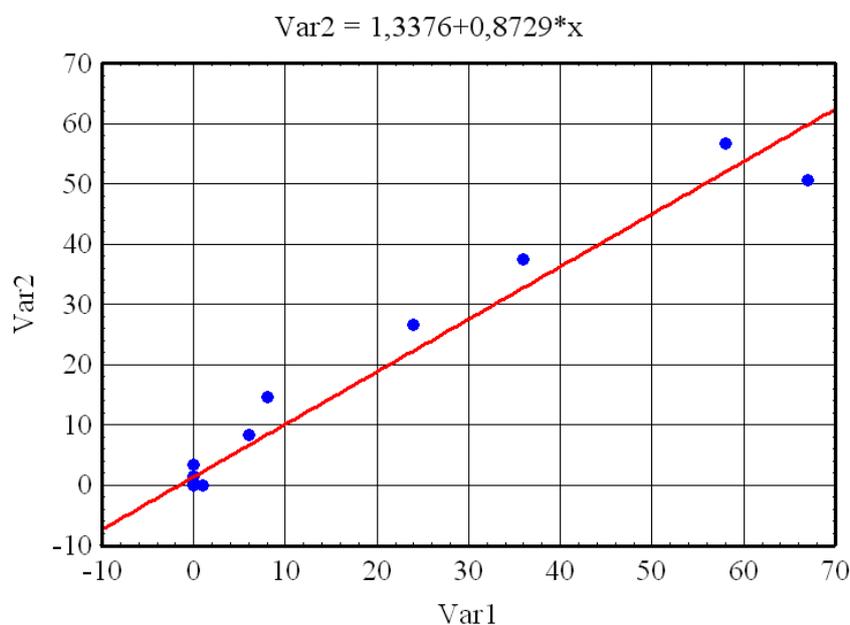


Рис. 3.2. График значений *Observed Frequency* от *Expected Frequency*

Вверху таблицы выводится значение статистики критерия χ^2 (*Chi-Square*), число степеней свободы (*df*) и вычисленный уровень значимости *p-level*. Для нашего примера получено:

Variable: Var1, Distribution: Normal
Chi-Square = 11,99951, df = 3 (adjusted), p = 0,0073.

Значение вероятности $p=P(\chi^2_3 > 11,999) = 0,007$ означает, что если гипотеза верна, вероятность получить 12 или больше равна 0,007. Это слишком мало, чтобы поверить в нормальность распределения. Гипотезу о нормальности отклоняем.

Если посмотреть гистограмму наблюдений (рис 3.3), видно, что в выборке имеется одно anomальное значение 14,56 (188-е по счёту), которое могло появиться в результате какой-либо ошибки (при записи наблюдений, при перепечатке или попалась деталь с другого станка и т. д.). Удалим его и снова проверим гипотезу.

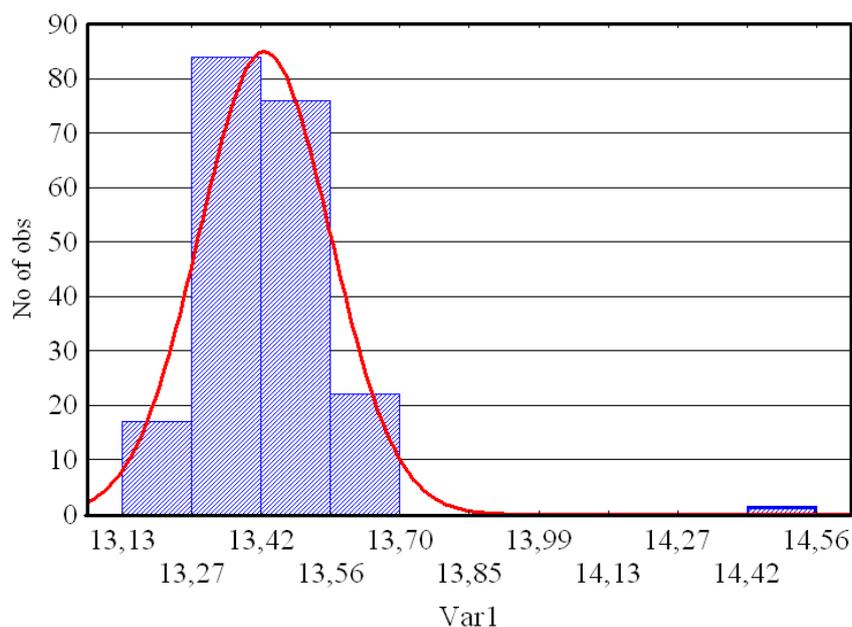


Рис. 3.3. Гистограмма исходной переменной Var1

Удаление одного наблюдения, если оно типично, не может изменить характеристики совокупности из 200 элементов. Если же изменение происходит, это наблюдение типичным не является и должно быть удалено. Если повторить проверку гипотезы для «цензурированной» выборки, можно убедиться в том, что наблюдения не противоречат гипотезе о нормальности.

3.7. Проверка гипотез об однородности выборок

Пусть имеются выборки, извлечённые из различных совокупностей. Требуется проверить гипотезу о том, что исходные совокупности распределены одинаково. В системе Statistica эта гипотеза проверяется в

модуле *Statistics/ Advanced Linear/Nonlinear models/ Log-Linear Analysis of Frequency Tables*.

Пусть, к примеру, имеются данные о наличии примесей (P1–P4) в углеродистой стали, выплавляемой двумя заводами Z1, Z2 (рис. 3.4) [11].

	1 P1	2 P2	3 P3	4 P4
Z1	82	535	1173	1714
Z2	63	429	995	1307

Рис. 3.4. Пример проверки однородности выборок

Проверим гипотезу о том, что распределения содержания нежелательной примеси одинаковы на этих заводах.

1. В строке *Input file:* выбираем *Frequencies w/out coding variables* (частоты без кодирующих переменных). Кнопкой *Variables* вводим все переменные (*Select all*). Кнопкой *Specify Table* (спецификация таблицы) в ячейках *No. of levels:* вводим 4 и 2 (рис. 3.5).

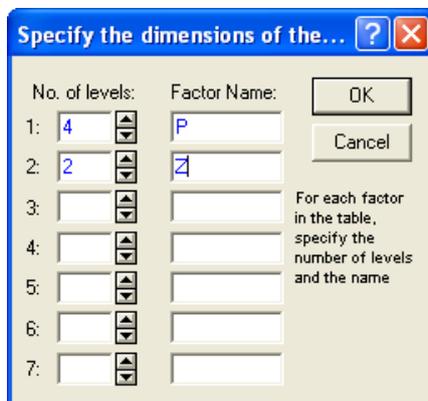


Рис. 3.5. Определение спецификации таблицы

2. Дважды нажимаем *OK* и во вкладке *Advanced* получившегося окна выполним *Test all marginal & partial association models*.

3. В таблице *Results of Fitting all K-Factor Interactions* в последней строке получаем столбца значение статистики критерия χ^2 (*Chi-Square*), равное 3,59, число степеней свободы (*Degrs. of Freedom*) $df=3$ и уровень

значимости 0,30887. Эта величина не больше критической. Следовательно, гипотезу об одинаковом распределении содержания примеси в металле на двух заводах можно принять.

3.8. Задания для самостоятельной работы

Задание 1. Теорема Хинчина утверждает, что среднее арифметическое

$$\bar{X} = \left| \frac{X_1 + X_2 + \dots + X_n}{n} \right|$$

независимых случайных величин $X_j, j=1, 2, \dots, n$, имеющих одно и то же распределение и конечное математическое ожидание m , сходится по вероятности при $n \rightarrow \infty$ к m . Таким образом, при заданном ε и достаточно большом n событие

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

можно считать практически достоверным.

Постепенно увеличивая n , с помощью пакета Statistica показать, выполняется теорема или нет для заданного закона распределения случайных чисел $X \in [0; 1]$. Закон распределения случайных чисел выбрать из таблицы по номеру своего варианта (табл. 3.1):

Таблица 3.1

Варианты задания и распределения

Номер варианта	Распределение	Функция в пакете Statistica
1	равномерное	Rnd
2	нормальное	Rndnormal
3	Пуассона	Poisson

Указание. Равномерно распределённые случайные числа генерируются программно. Для заполнения переменной случайными числами из интервала $[0;1]$ необходимо щёлкнуть правой кнопкой мыши на имени переменной и выбрать команды *Fill/Standardize Block / Fill Random Values*. Среднее значение вычисляется тем же способом командой *Statistics of Block Data / Block Columns / Means*.

Заполнить переменную вычисленными значениями можно, если дважды щёлкнуть левой кнопкой мыши по имени переменной и в поя-

вившемся окне «*Long name (label or formula)*» записать формулу. Например, формула =rndnormal(2) позволяет получить нормально распределённые случайные числа в интервале [0;2].

Задание 2. Сформировать набор данных для последующего анализа в программе Statistica, состоящий из 1 переменной и 150 наблюдений. Переменную заполнить числами из табл. 3.2. При этом N-му варианту соответствуют элементы выборки, расположенные в 15-ти следующих строчках таблицы, начиная с N-й (объем выборки при этом $n = 150$).

Таблица 3.2

Варианты задания и элементы выборки

N										
1	48	30	43	44	30	34	32	43	40	46
2	25	21	34	49	39	37	45	49	31	49
3	43	46	34	35	42	30	41	34	42	22
4	38	40	26	47	34	42	38	20	38	36
5	30	13	41	40	40	15	35	11	38	45
6	37	12	38	36	14	39	32	54	43	39
7	23	30	32	36	32	34	49	18	49	50
8	37	20	44	28	44	35	45	34	33	41
9	43	45	50	14	33	39	41	39	46	31
10	40	52	44	39	35	54	33	42	42	36
11	44	51	45	19	34	44	40	37	43	32
12	33	42	40	35	37	13	48	48	50	32
13	40	48	45	23	36	36	42	40	37	30
14	44	50	46	39	31	48	44	42	36	51
15	44	50	54	37	33	34	42	43	43	47
16	33	48	18	42	15	32	34	14	39	45
17	48	26	31	34	38	36	46	49	40	48
18	42	47	35	34	41	33	41	35	43	42
19	39	37	47	27	33	22	37	19	19	37
20	43	41	30	39	38	36	36	34	42	46
21	39	44	37	35	43	38	33	47	45	38
22	37	48	38	52	40	45	44	42	38	40
23	44	46	37	34	41	37	41	39	30	38

24	32	41	48	36	51	36	33	39	45	40
25	34	41	38	34	33	27	51	45	27	38
26	42	37	46	41	47	36	30	45	41	40
27	37	37	39	42	48	41	36	39	33	47
28	43	49	27	31	41	46	40	36	36	42
29	41	46	33	37	47	35	31	29	30	36
30	48	38	37	34	40	34	36	50	48	39
31	30	38	43	41	44	45	38	37	46	50
32	41	48	41	43	47	37	42	34	32	44
33	37	48	46	41	41	37	37	48	49	46
34	38	44	50	37	47	27	48	37	46	38
35	48	47	38	52	34	36	34	41	41	32
36	31	43	34	46	37	40	41	39	32	42
37	47	33	51	41	40	45	37	36	27	36
38	37	42	46	35	34	38	45	36	20	40
39	34	48	30	51	33	41	44	42	39	39
40	45	45	41	40	36	27	50	44	41	48
41	36	36	32	32	36	49	27	45	30	35
42	40	38	45	40	40	50	42	37	50	39
43	43	38	30	59	42	41	33	42	38	44
44	44	41	47	52	51	38	50	39	50	48
45	49	43	52	50	30	30	26	50	27	49
46	27	49	46	39	47	26	49	52	29	44
47	51	53	48	49	53	45	27	43	48	44

По данной выборке объёма $n = 150$ построить статистический ряд:

1	2	..	e
n_1	2	..	e

где $x_1 < x_2 < \dots < x_e$ элементы выборки, записанные в порядке возрастания, n_i – частоты появления одинаковых значений случайной величины y .

Задание 3. На основе статистического ряда построить сгруппированную выборку. Для этого задается определенный отрезок $[a, b]$, внутри которого расположены все элементы исследуемой выборки, число интервалов k , на которое делится этот отрезок. Находятся длины интер-

валов $h = \frac{b-a}{k}$, концы интервалов $x_i = a + (i-1)h$, середины интервалов $z_i = \frac{1}{2}(x_i + x_{i+1})$ и соответствующие эмпирические частоты m_i (m_i – число элементов выборки, попавших в i -й интервал), $i = 1, 2, \dots, k$. Результаты вычислений заносятся в табл. 3.3.

Таблица 3.3

Группировка выборки

Номер интервала	Границы интервала	Середины интервалов	Эмпирические частоты
i	x_i, x_{i+1}	z_i	m_i
1			
2			
...			
...			
k			

Принять уровень значимости $\alpha = 0,05$, отрезок $[24,5; 54,5]$, число интервалов $k = 10$.

Задание 4. Построить график эмпирической функции распределения

$$F^*(x) = \begin{cases} 0 & x \leq z_1 \\ \frac{1}{n}(m_1 + \dots + m_i), & \text{при } z_i < x \leq z_{i+1}, (i = 1, 2, \dots, k-1). \\ 1 & x > z_k \end{cases}$$

и гистограмму.

Известно, что гистограмма строится из прямоугольников с основаниями $[x_i, x_{i+1}]$ и высотами $\frac{m_i}{nh}$. Проверить, выполняются ли эти условия при построении гистограммы в программе Statistica.

Задание 5. Найти выборочное среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^k m_i z_i$, исправленную выборочную дисперсию $S^2 = \frac{1}{n-1} \sum_{i=1}^k m_i \left(z_i - \bar{x} \right)^2$; исправленное выборочное среднеквадратическое отклонение $S = \sqrt{S^2}$.

Проверить гипотезу о нормальном распределении случайной величины X с математическим ожиданием $a = \bar{x}$ и среднеквадратическим отклонением $\sigma = S$ с помощью критерия χ^2 Пирсона.

Для этого вычислить теоретические частоты попадания случайной величины X в i -й интервал np_i , где

$$p_i = p\{x_i \leq X < x_{i+1}\} = \Phi\left(\frac{x_{i+1} - \bar{x}}{S}\right) - \Phi\left(\frac{x_i - \bar{x}}{S}\right).$$

Значения функции Лапласа $\Phi(x) = \frac{2}{\sqrt{2\pi}} \int_0^x e^{-\frac{u^2}{2}} du$ находятся по таблице или с помощью вероятностного калькулятора.

Если при некотором i эмпирическая или теоретическая частота меньше 5, тогда этот интервал объединяют с соседним, при этом теоретические и эмпирические частоты суммируются. После объединения получают r интервалов ($r \leq k$).

Составляется статистика χ^2 Пирсона

$$\chi^2_{\text{набл.}} = \sum_{i=1}^r \frac{(m_i - np_i)^2}{np_i}.$$

Затем по закону уровня значимости α и числу степеней свободы $\nu = r - 3$ находится критическая точка $\chi^2_{\alpha, \nu}$ по таблице квантилей распределения χ^2 . Если $\chi^2_{\text{набл.}} > \chi^2_{\alpha, \nu}$, то гипотеза отвергается. Если $\chi^2_{\text{набл.}} \leq \chi^2_{\alpha, \nu}$, гипотеза принимается.

Задание 6. Построить график плотности вероятности

$f(x) = \frac{1}{\sqrt{2\pi}S} e^{-\frac{(x-\bar{x})^2}{2S^2}}$ случайной величины X . Какой закон распределения X вы наблюдаете?

Задание 7. Проанализировать набор данных из задания 2, состоящий из 1 переменной и 150 наблюдений в программе Statistica. Целью задания является проверка гипотезы о нормальном распределении случайной величины по критерию χ^2 автоматически. Для этого выполнить следующие действия: *Statistics/ Distribution Fitting* (подбор распределений)/ *Continuous Distributions* (непрерывные распределения)/ *OK/ Normal* (нормальное распределение)/ *Variable/ Summary*. На экран выводится таблица для расчёта статистики критерия.

Во вкладке *Parameters* того же окна появятся оценки параметров. Число интервалов группировки (*Number of categories*) можно при необходимости изменить.

Для вычерчивания измеряемого и ожидаемого распределения нажимаем соответствующую кнопку (*Plot of observed and expected distribution*). Появится гистограмма, вверху которой написано рассчитанное значение χ^2 (*Chi-Square test*), число степеней свободы (*df*) и уровень значимости (*p*). Именно *p*-уровень представляет собой вероятность ошибки, связанной с распространением наблюдаемого результата на всю выборку.

Сравнить графически наблюдаемые (*Observed Frequency*) и ожидаемые частоты (*Expected Frequency*): записать соответствующие столбцы в отдельную таблицу и построить график рассеяния (команды *Graphs/ Scatterplots/ Variables/ OK*). Переменные существенно различаются, если точки плохо укладываются на прямую линию. (Если бы переменные были одинаковы, все наблюдения лежали бы на прямой с уравнением $\text{Var2}=\text{Var1}$).

Является ли исследуемая переменная нормально распределённой? Сравнить полученный результат с заданием 6 и сделать выводы.

Задание 8. Таблицу из восьми переменных ($\text{Var1} \dots \text{Var8}$) и 500 наблюдений заполнить случайными числами из интервала $[0;1]$. Найти $\text{Var9}=\text{Var1}+\text{Var2}+\dots+\text{Var8}$. Для этого необходимо дважды щёлкнуть левой кнопкой мыши по Var9 и в появившемся окне «Long name (label or formula)» записать формулу

$$=v1+v2+v3+v4+v5+v6+v7+v8$$

или

$$=\text{sum}(v1:v8).$$

Построить гистограммы для Var1 и Var9 отдельно. Какие распределения вы получили? Объяснить полученный результат.

Задание 9. Для этих же данных проверить гипотезу о нормальном распределении случайных величин Var1 Var9 по критерию χ^2 автоматически, как это делалось в задании 7. Сравнить с результатами Задания 8 и объяснить полученный результат.

ГЛАВА 4. РЕГРЕССИЯ, КОРРЕЛЯЦИЯ И СОВПАДЕНИЕ

4.1. Зависимость

Основная задача регрессионного и корреляционного анализа состоит в выявлении связи между случайными переменными. Например, на свободном рынке обычно наблюдается большая степень корреляции между размером урожая и рыночными ценами на соответствующую продукцию сельского хозяйства. Часто корреляция привлекает наше внимание к причинно-следственным связям, существующим между изучаемыми двумя рядами величин. В области естественных и общественных наук установление существенной корреляции часто заставляет нас искать возможные связи между явлениями, которые в противном случае могли остаться незамеченными.

В экономике в большинстве случаев между переменными величинами существуют зависимости, когда каждому значению одной переменной соответствует не какое-то определённое, а множество возможных значений другой переменной. Иначе говоря, каждому значению одной переменной соответствует определённое условное распределение другой переменной. Такая зависимость получила название *статистической*.

Возникновение понятия статистической связи обуславливается тем, что зависимая переменная подвержена влиянию неконтролируемых или неучтённых факторов, а также тем, что измерение значений переменных неизбежно сопровождается некоторыми случайными ошибками.

Статистическая зависимость между двумя переменными, при которой каждому значению одной переменной соответствует определённое условное математическое ожидание (среднее значение) другой, называется *корреляционной*.

Функциональная зависимость представляет собой частный случай корреляционной. При функциональной зависимости с изменением значений некоторой переменной x однозначно изменяется определённое значение переменной y , при корреляционной – определённое среднее значение (математическое ожидание) y , а при статистической – определённое распределение переменной y . Каждая корреляционная зависимость является статистической, но не каждая статистическая зависимость является корреляционной.

Статистические связи между переменными можно изучать методами корреляционного и регрессионного анализа. Основной задачей корреляционного анализа является выявление связи между случайными переменными и оценка её степени. Основной задачей регрессионного анализа является установление формы и изучение зависимости между переменными.

4.2. Корреляция

Корреляция определяет степень, с которой значения двух переменных «пропорциональны» друг другу. *Пропорциональность* означает просто *линейную зависимость*. Корреляция высокая, если на графике зависимость «можно представить» прямой линией (с положительным или отрицательным углом наклона). Таким образом, это простейшая регрессионная модель, описывающая зависимость одной переменной от одного фактора.

В производственных условиях обычно информации, полученной из диаграмм рассеяния при условии их корректного построения, бывает достаточно для того, чтобы оценить степень зависимости y от x . Но в ряде случаев требуется дать количественную оценку степени связи между величинами x и y . Такой оценкой является коэффициент корреляции.

Коэффициент корреляции – это показатель, оценивающий тесноту линейной связи между признаками

Отметим основные характеристики этого показателя.

- Он может принимать значения от -1 до $+1$. Знак «+» означает, что связь прямая (когда значения одной переменной возрастают, значения другой переменной также возрастают), «-» означает, что связь обратная.
- Чем ближе коэффициент к $|1|$, тем теснее линейная связь. При величине коэффициента корреляции менее $0,3$ связь оценивается как слабая, от $0,31$ до $0,5$ – умеренная, от $0,51$ до $0,7$ – значительная, от $0,71$ до $0,9$ – тесная, $0,91$ и выше – очень тесная.
- Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится.

- При $r = \pm 1$ корреляционная связь представляет линейную функциональную зависимость. При этом все наблюдаемые значения располагаются на общей прямой. Её ещё называют линией регрессии.

- При $r = 0$ линейная корреляционная связь отсутствует. При этом групповые средние переменных совпадают с их общими средними, а линии регрессии параллельны осям координат.

Равенство $r = 0$ говорит лишь об отсутствии линейной корреляционной зависимости (некоррелированности переменных), но не вообще об отсутствии корреляционной, а тем более, статистической зависимости.

Основываясь на коэффициентах корреляции, вы не можете *строго* доказать причинной зависимости между переменными, однако можете определить *ложные* корреляции, т. е. корреляции, которые обусловлены влияниями «других», остающихся вне вашего поля зрения переменных. Основная проблема ложной корреляции состоит в том, что вы не знаете, кто является её носителем. Тем не менее, если вы знаете, где искать, то можно воспользоваться *частные корреляции*, чтобы контролировать (*частично исключённое*) влияние определённых переменных.

«Как только я собираюсь в баню с друзьями, так в стране очередное ЧП случается»
министр МЧС С.К. Шойгу

Корреляция, совпадение или необычное явление сами по себе ничего не доказывают, но они могут привлечь внимание к отдельным вопросам и привести к дополнительному исследованию. Хотя корреляция прямо не указывает на причинную связь, она может служить ключом к разгадке причин. При благоприятных условиях на её основе можно сформулировать гипотезы, проверяемые экспериментально, когда возможен контроль других влияний, помимо тех немногочисленных, которые подлежат исследованию [10].

Очень важно установить логическую связь между двумя рядами явлений или двумя совпадающими во времени явлениями, либо же дать им разумное объяснение.

Иногда вывод об отсутствии корреляции важнее наличия сильной корреляции. Нулевая корреляция двух переменных может свидетельствовать о том, что никакого влияния одной переменной на другую не существует, при условии, что мы доверяем результатам измерений.

4.3. Корреляционный анализ в программе Statistica

Корреляционный анализ в программе Statistica проводят с помощью модуля *Statistics/ Basic Statistics/ Correlation Matrices*. В стартовом окне (рис. 4.1) для расчёта квадратной матрицы используется кнопка *One variable list*. С помощью кнопки *Two lists (rect. matrix)* можно ограничиться выводом только необходимых переменных, если не требуются все возможные парные корреляции. Из списка выбирают переменные, между которыми будут рассчитаны парные коэффициенты корреляции. После нажатия на кнопку *Summary* или *Correlations* на экране появится корреляционная матрица (рис. 4.2).

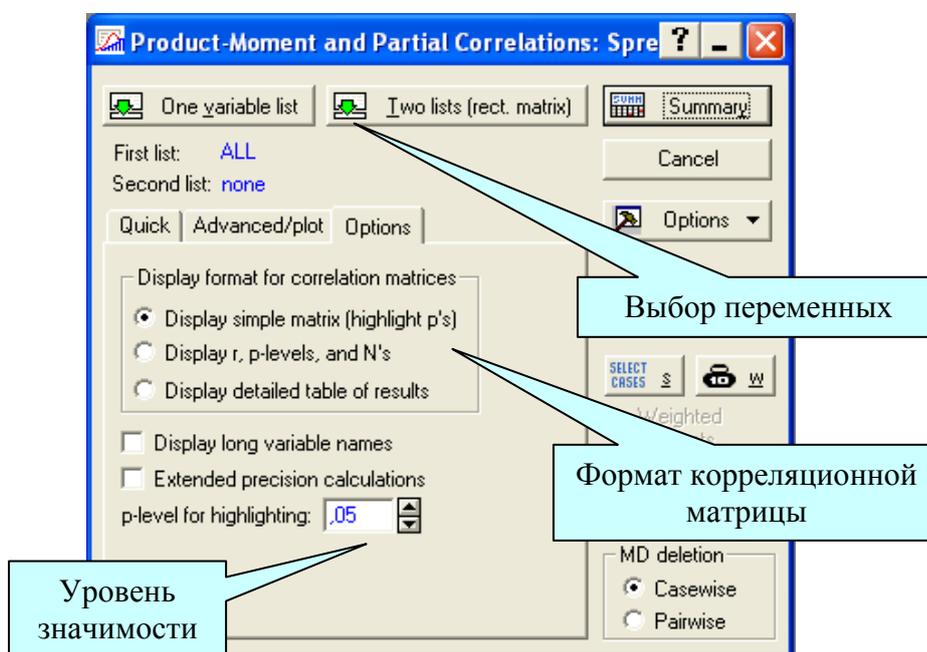


Рис. 4.1. Стартовое окно корреляционного анализа

Процедура *Correlation matrices* сразу же дает возможность проверить достоверность рассчитанных коэффициентов корреляции. Значение коэффициента корреляции может быть высоким, но не достоверным, случайным. Чтобы увидеть вероятность нулевой гипотезы (p), гласящей о том, что коэффициент корреляции равен нулю, нужно в опции *Display format for correlation matrices* (рис. 4.3) установить переключатель на вторую строку *Display r, p-levels, and N's*. Но даже если этого не делать и оставить переключатель в первом положении *Display simple matrix (highlight p's)*, статистически значимые на уровне 0,05 ко-

коэффициенты корреляции будут выделены в корреляционной матрице на экране красным цветом, а при распечатке помечены звёздочкой. Третье положение переключателя опции *Display Detailed table of results* позволяет просмотреть результаты корреляционного анализа в деталях. Флажок опции *MD deletion* устанавливается для исключения из обработки всей строки файла данных, в которой есть хотя бы одно пропущенное значение.

Correlations (Spreadsheet1)						
Marked correlations are significant at p < ,05000						
N=10 (Casewise deletion of missing data)						
Variable	Var1	Var2	Var3	Var4		
Var1	1,00	0,68	-0,38	0,27		
Var2	0,68	1,00	-0,06	0,11		
Var3	-0,38	-0,06	1,00	-0,56		
Var4	0,27	0,11	-0,56	1,00		

Рис. 4.2. Корреляционная матрица

Для построения диаграмм рассеяния необходимо во вкладке *Quick* стартового модуля *Statistics/ Basic Statistics/ Correlation Matrices* нажать кнопку *Scatterplot matrix for selected variables*. В результате этих действий появится графическое изображение зависимостей, пример которого приведён на рис. 4.5. Остаётся только сделать выводы.

Проведённая прямая в каждой диаграмме рассеяния называется *прямой регрессии* или прямой, построенной *методом наименьших квадратов*. Последний термин связан с тем, что сумма *квадратов* расстояний (вычисленных по оси ординат) от наблюдаемых точек до прямой является минимальной. Заметим, что использование *квадратов* расстояний приводит к тому, что оценки параметров прямой сильно реагируют на выбросы.

По главной диагонали матрицы строятся гистограммы. Понятно, что любая переменная стопроцентно коррелирует сама с собой, и строить линию регрессии не имеет смысла.

Во многих исследованиях первый шаг анализа состоит в вычислении корреляционной матрицы всех переменных и проверке значимых (ожидаемых и неожиданных) корреляций. После того как это сделано, следует понять общую природу обнаруженной статистической значимости: понять, почему одни коэффициенты корреляции значимы, а другие нет.

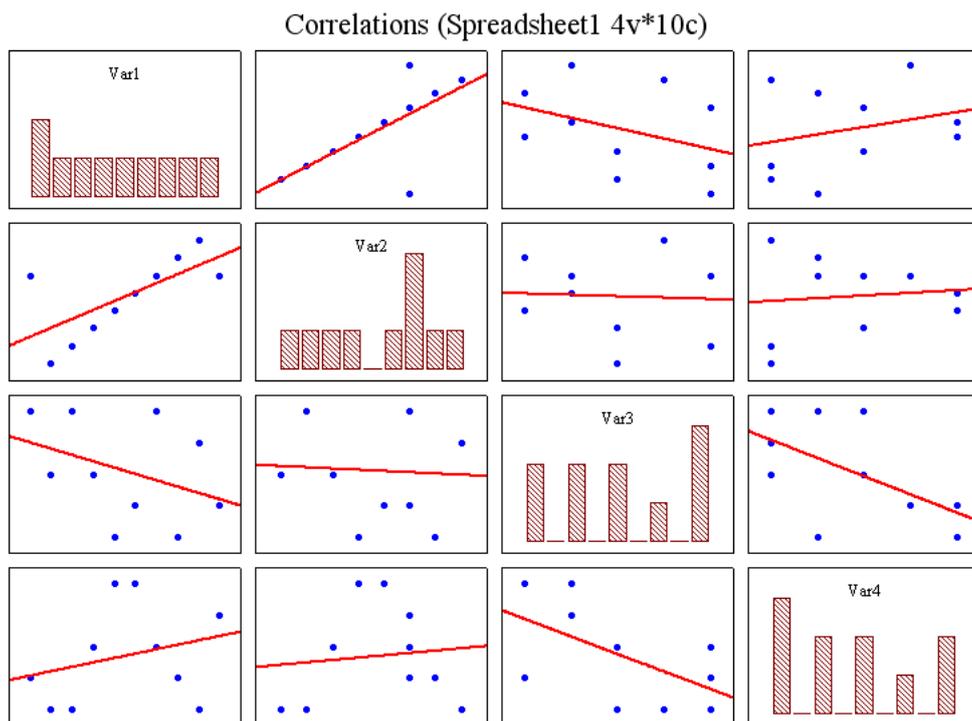


Рис. 4.3. Диаграммы рассеяния

Но следует иметь в виду, что если используется несколько критериев, значимые результаты могут появляться «удивительно часто», и это будет происходить чисто случайным образом. Например, коэффициент, значимый на уровне 0,05, будет встречаться чисто случайно один раз в каждом из 20 подвергнутых исследованию коэффициентов. Нет способа автоматически выделить «истинную» корреляцию. Поэтому следует подходить с осторожностью ко всем не предсказанным или заранее не запланированным результатам и попытаться соотнести их с другими (надёжными) результатами. В конечном счёте, самый убедительный способ проверки состоит в проведении повторного экспериментального исследования. Такое положение является общим для всех методов анализа, использующих множественные сравнения и статистическую значимость.

Рассмотрим пример решения практической задачи о производительности землеройной техники. Из-за сезонного характера работ неизбежны простои. Но поскольку простой техники обходится дорого, руководство предприятия интересовало пути сокращения простоев, в частности, в летние месяцы. В табл. 4.1 приведены данные о работе и простое всего парка в машино-часах.

Таблица 4.1

Данные по производительности землеройной техники

Месяц	Простой	Работа
ноябрь	1130,01	4137,63
декабрь	734,42	3704,00
январь	265,40	1328,40
февраль	586,60	1961,60
март	666,70	1939,70
апрель	1232,00	3116,00
май	3888,35	8509,35
июнь	5465,39	12588,89
июль	7412,33	14875,50
август	7168,66	15388,08
сентябрь	7416,68	15450,67
октябрь	5001,41	11944,82

Сначала имеет смысл отобразить данные на графике. Чтобы построить два графика на одной сетке, необходимо выбрать модуль *Graphs/ Scatterplots...* После чего появится диалоговое окно, в котором необходимо выбрать вкладку *Advanced*. Далее следует выбрать необходимые переменные и тип графика (рис. 4.4). После нажатия на кнопку *OK* график будет выведен.

После построения графика его можно отредактировать. В частности, изменить тип шрифта и кегль, изменить фон, толщину линий графиков и сетки. Наиболее часто приходится менять масштаб по осям. Для этого надо дважды щёлкнуть левой кнопкой мыши по размерной сетке, в появившемся окне *Axis Layout* выбрать *Scaling/ Mode: Manual* и указать нужные значения *Minimum* и *Maximum*. Готовый результат представлен на рис. 4.5.

Графики рассеяния (рис. 4.5) и корреляционный анализ (табл. 4.2) показали, что сезонность не является фактором, влияющим на простой. Налицо почти линейная зависимость между работой и простоями (рис. 4.8), то есть чем больше техника находится в работе, тем дольше она будет простаивать.

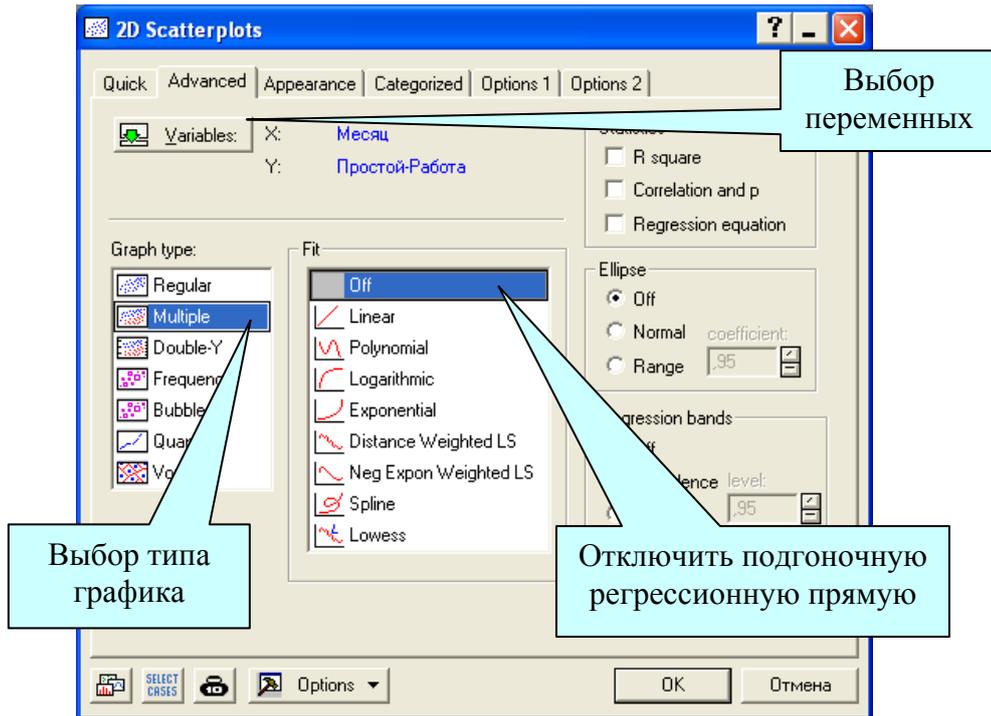


Рис. 4.4. Диалоговое окно построения графиков

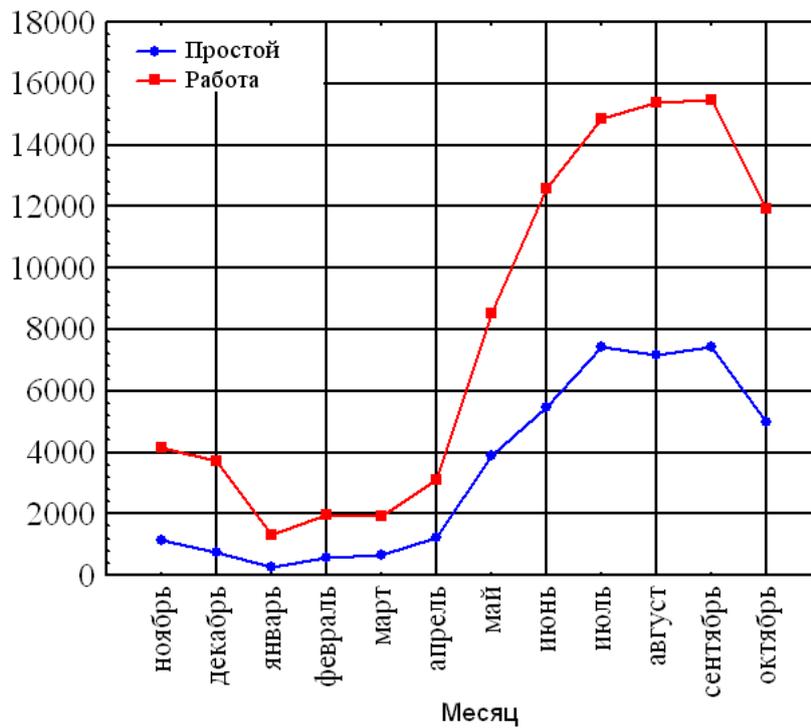


Рис. 4.5. Диаграммы рассеяния

Таблица 4.2

Корреляционная матрица для производительности землеройной техники

	Месяц	Простой	Работа
Месяц	1,00	0,86*	0,85*
Простой	0,86*	1,00	0,99*
Работа	0,85*	0,99*	1,00

* – коэффициенты, значимые по уровню 0,05

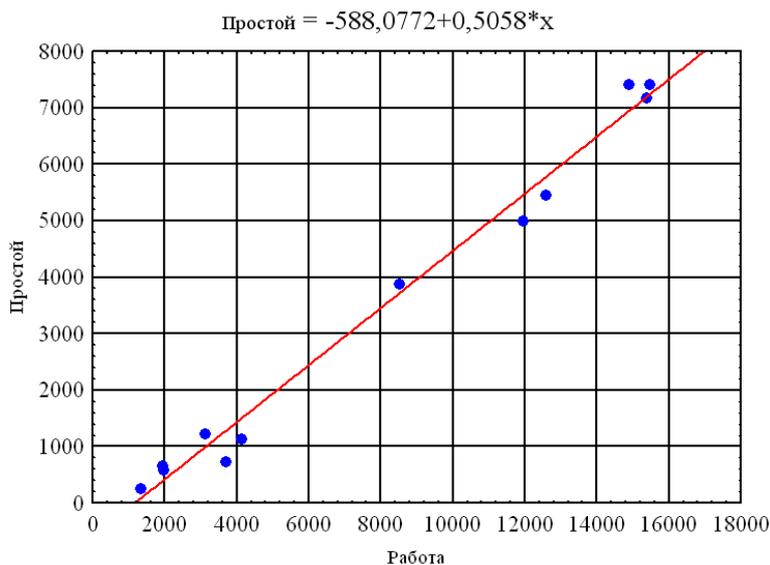


Рис. 4.6. Зависимость времени простоя от времени работы

Понятно, что для решения задачи сокращения простоев техники нужно выявить влияющие факторы и искать статистическую зависимость от них. Этот пример показывает, что степень связи между любыми двумя переменными, независимо от того, как эта связь выражена, зависит от характера измерения переменных.

4.4. Ранговая корреляция

На практике часто изучают связи между порядковыми переменными, измеренными в так называемой порядковой шкале. В этой шкале можно установить лишь порядок, в котором объекты выстраиваются по степени проявления признака (например, качество жилищных условий, тестовые баллы, экзаменационные оценки). Если, скажем, по некоторой дисциплине два студента имеют оценки «отлично» и «удовле-

творительно», то можно лишь утверждать, что уровень подготовки по этой дисциплине первого студента лучше, чем второго, но нельзя сказать, на сколько.

Оказалось, что в таких случаях проблема оценки тесноты связи разрешима, если упорядочить, или ранжировать объекты анализа по степени выраженности измеряемых признаков. При этом каждому объекту присваивается определённый номер, называемый рангом. Например, объекту с наименьшим проявлением (значением) признака присваивается ранг 1, следующему за ним – 2 и т. д. Объекты можно располагать и в порядке убывания проявления признака.

Ранжируя попарно связанные значения признаков, можно видеть, как они распределяются относительно друг друга. Если возрастающим значениям одного признака соответствуют возрастающие значения другого, то между ними существует положительная связь. Если же при возрастании значений одного признака значения другого последовательно уменьшаются, это указывает на наличие отрицательной связи между ними. При отсутствии корреляции ранжированным значениям одного признака будут соответствовать самые различные значения другого.

Определив ранги значений переменных, по их разностям можно судить о степени зависимости одного признака от изменений другого.

Коэффициент ранговой корреляции Спирмена находится по формуле:

$$r = 1 - \frac{6 \sum_{i=1}^n (r_i - s_i)^2}{n^3 - n},$$

где r_i и s_i – ранги i -го объекта по переменным x и y , n – число пар наблюдений (объём выборки). Если ранги всех объектов равны ($r_i = s_i$, $i = 1, 2, \dots, n$), то $r = 1$, то есть наблюдается полная прямая связь.

Рассмотрим вычисление ранговой корреляции между рейтингом подразделения и премиальным фондом (табл. 4.3).

Таблица 4.3

Рейтинг подразделения и премиальный фонд

Рейтинг (Var 1)	13	16	29	35	36	41	75	89
Фонд (Var 2)	120	110	110	140	150	130	150	130

Выберите модуль *Statistics/ Nonparametrics*, в появившемся стартовом окне выберите пункт *Correlations (Spearman Kendall tau, gamma)* / кнопка *OK*. В открывшемся диалоговом окне выберите исследуемые признаки кнопкой *Variables (List 1 – Var 1, List 2 – Var 2)*. После нажатия на кнопку *Spearman rank R* получим окно с результатами корреляционного анализа (рис. 4.7).

		Spearman Rank Order Correlations			
		MD pairwise deleted			
		Marked correlations are significant at p <,05000			
Pair of Variables		Valid N	Spearman R	t(N-2)	p-level
Var1	& Var2	8	0,618284	1,926931	0,102279

Число наблюдений

Коэффициент корреляции Спирмена

Уровень значимости

Рис. 4.7. Результаты корреляционного анализа

Коэффициент корреляции Спирмена равен 0,618 с уровнем p-level 0,10. Это означает, что связь рейтинга, выражающего результативность работы, и премиального фонда статистически незначима по уровню 0,05.

Коэффициент ранговой корреляции τ Кендалла вычисляется по формуле

$$\tau = 1 - \frac{4k}{n(n-1)},$$

где k – число инверсий (нарушений порядка) в ряду рангов второй переменной при условии, что ранги первой переменной упорядочены.

В пакете *Statistica* коэффициент ранговой корреляции Кендалла вычисляется в процедуре *Statistics/ Nonparametrics*, в появившемся стартовом окне выберите пункт *Correlations (Spearman Kendall tau, gamma)* / кнопка *OK*. В открывшемся диалоговом окне выберите исследуемые признаки кнопкой *Variables (List 1 – Var 1, List 2 – Var 2)*. После нажатия на кнопку *Kendall Tau* во вкладке *Advanced* получим окно с результатами корреляционного анализа (рис. 4.8).

		Kendall Tau Correlations				
		MD pairwise deleted				
		Marked correlations are significant at $p < ,05000$				
Pair of Variables		Valid N	Kendall Tau	Z	p-level	p-exact 1-tailed
Var1	& Var2	8	0,415761	1,440238	0,149800	,138

Число наблюдений

Коэффициент корреляции Кендалла

Z-статистика

Рис. 4.8. Результаты корреляционного анализа

В качестве примера использованы данные из табл. 4.3. Так выборочное значение $\tau = 0,416$, то на уровне значимости $\alpha = 0,05$, что меньше p , ранговая корреляция незначима. Так как квантиль распределения $N(0, 1)$ $u_{0,95} = 1,645$, что больше выборочного значения Z , коэффициент ранговой корреляции τ незначимо отличается от нуля. Это означает, что связь рейтинга, выражающего результативность работы, и премиального фонда статистически незначима по уровню 0,05. Квантиль распределения $N(0, 1)$ $u_{0,95}$ вычисляется с помощью вероятностного калькулятора.

В заключение раздела отметим, что рассмотренные примеры отличаются малым числом наблюдений. Для надёжного результата общее число наблюдений не должно быть меньше 50. Несоблюдение этого требования не гарантирует достаточно точных выводов, которые делают на основании выборочных показателей.

4.5. Основы регрессионного анализа

Регрессионный анализ является одним из наиболее распространённых методов обработки экспериментальных данных при изучении зависимостей в физике, биологии, экономике, технике и других областях. Он заключается в определении аналитического выражения, в котором изменение одной величины (называемой зависимой или результативным признаком) y обусловлено влиянием одной или нескольких независимых величин (факторов) x_1, x_2, \dots, x_n , а множество всех прочих

факторов, также оказывающих влияние на зависимую величину, принимается за постоянные и средние значения.

Регрессия может быть однофакторной (парной) и многофакторной (множественной). Если в качестве объясняющих факторов использовать только три фактора x_1, x_2, x_3 , то регрессионная модель может быть записана в виде:

$$y = f(x_1, x_2, x_3) + \varepsilon,$$

где $f(x_1, x_2, x_3)$ – **неслучайная** составляющая отклика y , зависящая от x_1, x_2, x_3 , а ε – остаток или случайная составляющая, обусловленная влиянием на отклик y множества неучтённых и непредсказуемых факторов, а также ошибок измерений зависимой переменной.

Для простой (парной) регрессии в условиях, когда достаточно полно установлены причинно-следственные связи, можно использовать графическое изображение. При множественности причинных связей невозможно чётко разграничить одни причинные явления от других. В этом случае наиболее приемлемым способом определения зависимости (уравнения регрессии) является метод перебора различных уравнений, реализуемый с помощью компьютера.

Существуют различные регрессионные модели, определяемые выбором функции $f(x_1, x_2, x_3)$:

- простая линейная регрессия $y = b_0 + b_1x + \varepsilon$;
- множественная регрессия $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_{m-1}x_{m-1} + \varepsilon$;
- полиномиальная регрессия $y = b_0 + b_1x + b_2x^2 + \dots + b_{m-1}x^{m-1} + \varepsilon$;
- регрессионная модель общего вида:

$$y = b_0 + b_1f_1(x_1, x_2, \dots, x_n) + \dots + b_{m-1}f_{m-1}(x_1, x_2, \dots, x_m) + \varepsilon, \quad (4.1)$$

где $f_i(x_1, x_2, \dots, x_n), i = \overline{1, m-1}$ – заданные функции факторов.

Параметры b_0, b_1, \dots, b_{m-1} называются коэффициентами регрессии. В приведённые регрессионные модели коэффициенты b_0, b_1, \dots, b_{m-1} входят линейно. Такие модели называют линейными, а математические методы анализа этих моделей – линейным регрессионным анализом.

В некоторых случаях нелинейные модели с помощью специальных линеаризирующих преобразований могут быть преобразованы в линейные. Например, функция $y = b_0x^{b_1}$ с помощью логарифмирования $\ln y = \ln b_0 + b_1 \ln x$ и замены переменных $\tilde{y} = \ln y, \hat{b}_0 = \ln b_0, x = \ln x$ преобразуется в линейную по параметрам $y = \hat{b}_0 + b_1x$.

После выбора вида регрессионной модели, используя результаты наблюдений зависимой переменной и факторов, нужно вычислить оценки (приближённые значения) параметров регрессии, а затем проверить значимость и адекватность модели результатам наблюдений.

Порядок проведения регрессионного анализа следующий:

- выбор модели регрессии, что включает в себе предположение о зависимости функций регрессии от факторов;
- оценка параметров регрессии в выбранной модели методом наименьших квадратов;
- проверка статистических гипотез о регрессии.

4.6. Пример проведения регрессионного анализа данных

Воспользуемся данными из табл. 4.1 для построения приближённой зависимости времени простоя техники от времени работы и месяца. На существование этой зависимости, причём линейной, указывает корреляционный анализ. Имея зависимость, выраженную в виде формулы, можно прогнозировать время простоя на следующий период и оценить недополученную прибыль в результате простоев, что так любят делать экономисты.

Следствием грамотной математической модели всегда является управленческое решение

Линейный регрессионный анализ выполняется в модуле *Statistics/Multiple Regression*. В стартовом диалоговом окне этого модуля при помощи кнопки *Variables* указываются зависимая (*dependent*) и независимые (*independent*) переменные. В поле *Input file* указывается тип файла с данными:

Raw Data – данные в виде строчной таблицы (по умолчанию);

Correlation Matrix – данные в виде корреляционной матрицы.

В стартовом окне можно задать и дополнительные опции и параметры анализа. Например, можно выбрать определённое подмножество наблюдений для анализа или приписать вес переменным. Также можно задать и опции, которые относятся непосредственно к статистической процедуре: задать правило обработки пропущенных данных, выбрать метод анализа по умолчанию и др.

Для вывода результатов и их анализа нажмите на кнопку *OK*. Система произведет вычисления, и на экране появится окно результатов (рис. 4.9). Оно имеет простую структуру: верхняя часть окна – информационная, нижняя содержит функциональные кнопки, позволяющие всесторонне просмотреть результаты анализа.

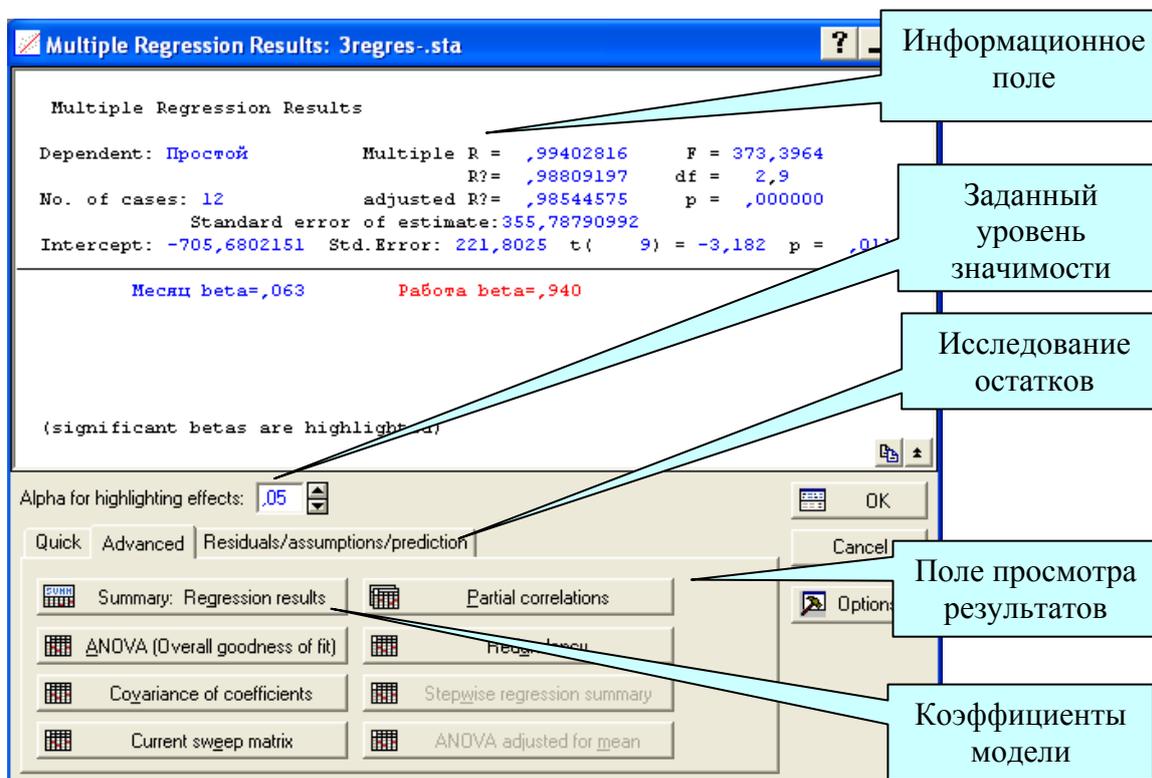


Рис. 4.9. Окно результатов регрессионного анализа

Рассмотрим вначале информационную часть окна. В ней содержится краткая информация о проведённом анализе.

Dependent – имя зависимой переменной. В нашем случае это «Простой».

No. of cases – число наблюдений, по которым построена регрессия. В примере число равно 12.

Multiple R – коэффициент множественной корреляции. Эта статистика полезна в множественной регрессии, когда вы хотите описать зависимости между переменными. Она может принимать значения от 0 до 1 и характеризует тесноту линейной связи между зависимой и всеми независимыми переменными.

R^2 – квадрат коэффициента множественной корреляции (R^2), называемый коэффициентом детерминации:

$$R^2 = \frac{SSR}{SST}, \quad (4.2)$$

где SSR – сумма квадратов, объясненная уравнением регрессии (Sum of Squares about Regression), SST – полная сумма квадратов (Total Sum of Squares).

Коэффициент детерминации является одной из основных статистик в данном окне, он показывает долю общего разброса (относительно выборочного среднего зависимой переменной), которая объясняется построенной регрессией. Чем ближе коэффициент детерминации к единице, тем качественнее найдена модель (объясняет поведение большего числа точек).

Коэффициент детерминации, определяемый выражением (4.2), обладает одним существенным недостатком. При равенстве числа независимых переменных q числу наблюдений n величина R^2 равна 1. По мере добавления переменных в уравнение значение R^2 неизбежно возрастает. Это ведет к неоправданному предпочтению моделей с большим числом независимых переменных. Отсюда следует, что необходима поправка к R^2 , которая бы учитывала число переменных и наблюдений. В результате получаем скорректированный коэффициент детерминации (adjusted R^2) \bar{R}^2 :

$$\bar{R}^2 = 1 - \frac{n-1}{n-q-1}(1-R^2).$$

Включение новой переменной в регрессионное уравнение увеличивает R^2 не всегда, а только в том случае, когда частный F -критерий при проверке гипотезы о значимости включаемой переменной больше или равен 1. В противном случае включение новой переменной уменьшает значение коэффициентов детерминации. Таким образом, скорректированный R^2 можно с большим успехом (по сравнению с R^2) применять для выбора наилучшего подмножества независимых переменных в регрессионном уравнении.

F -критерий используется для оценки адекватности регрессионной модели, определяет отношение дисперсии оценки модели к дисперсии остатка и равен

$$F = \frac{SSR/q}{SSE/(n-q-1)},$$

где SSE – сумма квадратов остатков.

Всякая сумма квадратов связана с числом степеней свободы. Это разность между числом различных опытов и числом констант, найденных по этим опытам независимо друг от друга. Например, для *SSE* число степеней свободы равно числу опытов n минус $(q + 1)$ коэффициентов регрессии.

Standard 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.

F – значения F-критерия для проверки гипотезы $b_1 = 0$.

df – число степеней свободы F-критерия.

p – уровень значимости.

t – *t*-критерий для проверки гипотезы о равенстве нулю свободного члена уравнения. Если p больше заданного уровня значимости Alpha (рис. 4.11), то гипотеза $b_0 = 0$ принимается.

Beta – коэффициенты b уравнения.

В информационной части прежде всего нужно смотреть на значение коэффициента детерминации. В нашем примере он равен 0,988... Это значит, что построенная регрессия объясняет 98,8 % разброса значений переменной «Простой» относительно среднего. Это хороший результат.

Далее смотрим на значение *F*-критерия и уровень его значимости p . *F*-критерий используется для проверки гипотезы, утверждающей, что между зависимой переменной «Простой» и независимой переменной «Работа» нет линейной зависимости, т. е. $b_1 = 0$, против альтернативы « b_1 не равен нулю». В данном примере большое значение *F*-критерия 373,3964 и даваемый в окне уровень значимости $p = 0,0112$ показывают, что построенная регрессия значима.

При помощи кнопок диалогового окна *Multiple Regressions Results* (рис. 4.9) результаты регрессионного анализа можно просмотреть более детально. Щелкните далее на кнопку *Summary:Regression results* (краткие результаты регрессии). Вы увидите таблицу с результатами анализа, приведённую на рис. 4.10.

Во втором столбце таблицы (*Beta*) выводятся стандартизованные коэффициенты регрессии, в третьем (*Std.Err. of Beta*) – их стандартные отклонения. В случае множественной регрессии стандартизованные ко-

эффиценты регрессии используются для сравнения влияния на зависимую переменную факторов, имеющих различную размерность.

Regression Summary for Dependent Variable: Простой (3regres-.sta)						
R= ,99402816 R?= ,98809197 Adjusted R?= ,98544575						
F(2,9)=373,40 p<,00000 Std.Error of estimate: 355,79						
N=12	Beta	Std.Err. of Beta	B	Std.Err. of B	t(9)	p-level
Intercept			-705,680	221,8025	-3,18157	0,011157
Месяц	0,062537	0,069694	51,152	57,0059	0,89731	0,392925
Работа	0,940149	0,069694	0,479	0,0355	13,48967	0,000000

Кoeffициенты регрессии

СКО коэффицентов

Уровень значимости

Рис. 4.10. Краткие результаты регрессии

В четвёртом столбце таблицы имеются оценки неизвестных параметров модели:

$$b_0 = -705,680;$$

$$b_1 = 51,152;$$

$$b_2 = 0,479;$$

в пятом столбце (St.Err. of B) – их стандартные отклонения.

Итак, искомая модель зависимости времени простоя техники от времени работы и месяца имеет вид:

$$\text{Простой} = -705,680 + 51,152 * \text{Месяц} + 0,479 * \text{Работа} + \varepsilon. \quad (4.3)$$

Из модели очевидна необходимость снижения сезонности работ.

В шестом и седьмом столбцах таблицы (рис. 4.12) выводятся t -статистики и соответствующие уровни значимости для проверки гипотезы о равенстве нулю коэффицентов регрессии. Для нашего примера гипотеза для b_0 и b_2 отклоняется.

4.7. Оценка адекватности модели по остаткам

Для оценки адекватности модели необходимо исследовать **остатки**. *Остатки* – это разность между исходными (наблюдаемыми) значениями зависимой переменной и предсказанными (модельными, *Predic-*

ted values) значениями по формуле (4.3). Остатки должны быть нормально распределены, иметь нулевое среднее значение и постоянную дисперсию, независимо от величин зависимых и независимых переменных. Модель должна быть адекватна на всех отрезках интервала изменения зависимой переменной. Вначале для оценки адекватности модели лучше всего использовать визуальные методы и затем, если потребуется, перейти к статистическим критериям.

Для исследования остатков в окне результатов регрессионного анализа (рис. 4.9) необходимо выбрать вкладку *Residuals/assumptions/prediction* и нажать кнопку *Perform residual analysis*. Для оценки адекватности модели построим нормальный вероятностный график остатков. В отобразившемся окне, перейдя к вкладке *Quick*, необходимо нажать кнопку *Normal plot of residuals* (рис. 4.11). Полученный график остатков приведён на рис. 4.12.

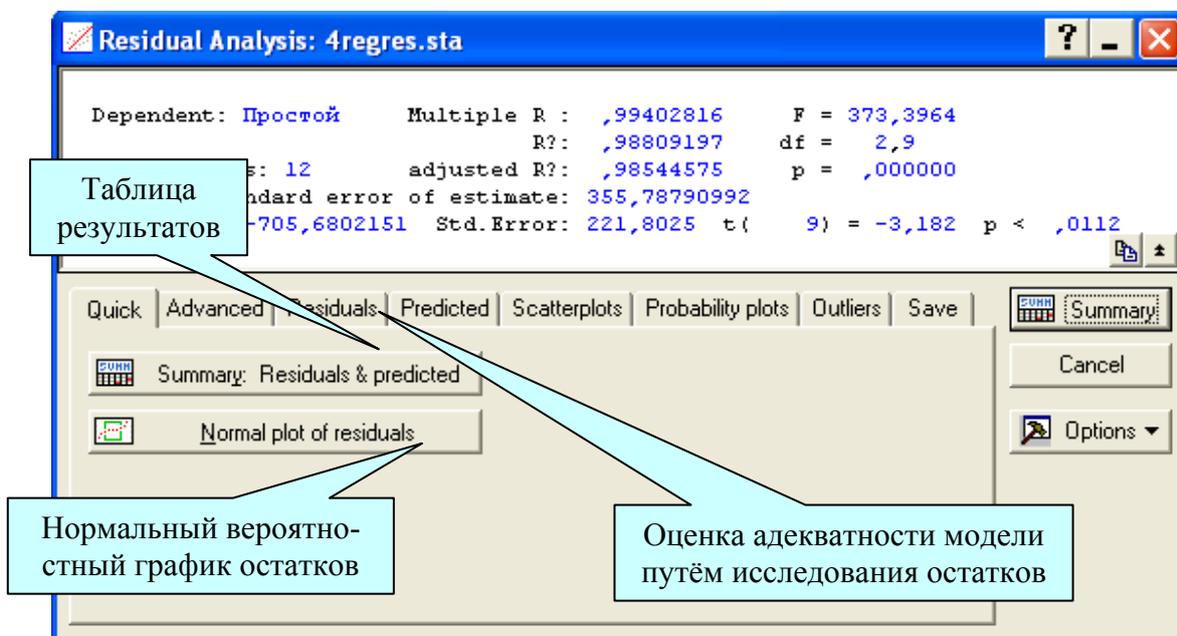


Рис. 4.11. Окно исследования остатков

Из графика видно, что остатки достаточно хорошо ложатся на прямую, которая соответствует нормальному закону. Поэтому предположение о нормальном распределении ошибок выполнено.

Для выявления нестабильности дисперсии ошибки уравнения можно построить график зависимости регрессионных остатков от предсказанного значения зависимой переменной. Во вкладке *Scatterplots* на-

жмите кнопку *Predicted vs. residuals* (рис. 4.13). В результате будет строен график, приведённый на рис. 4.14.

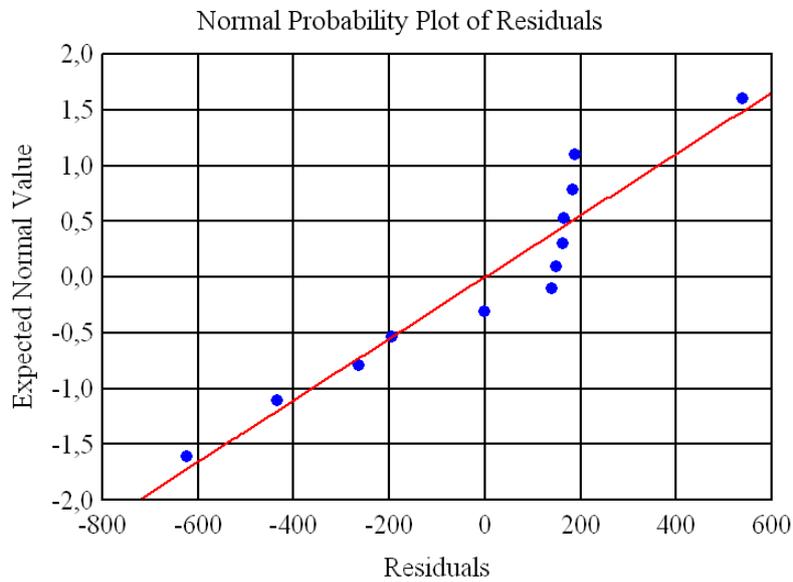


Рис. 4.12. Нормальный вероятностный график остатков

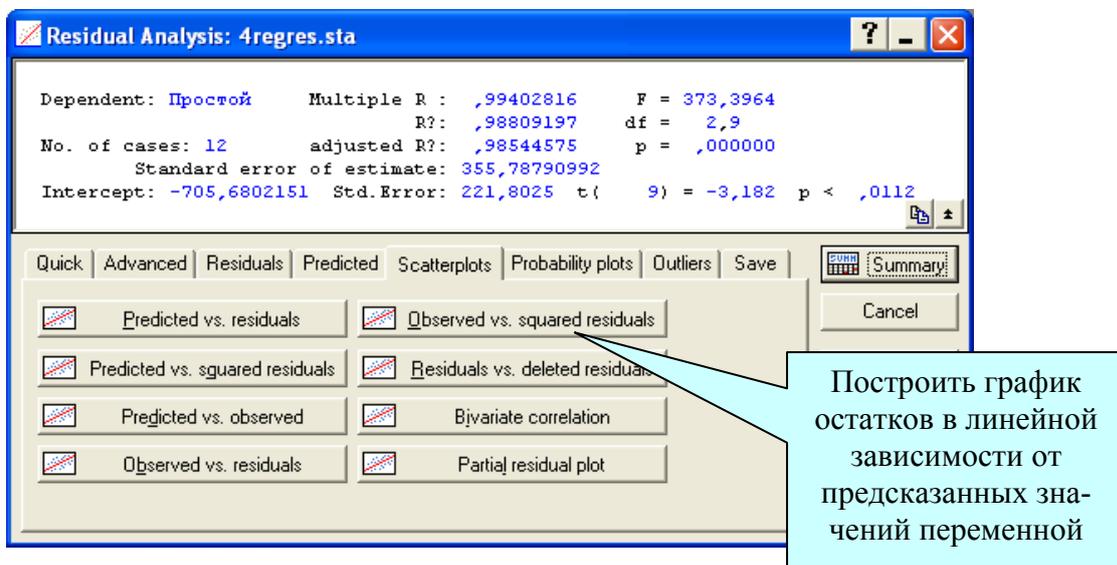


Рис. 4.13. Окно анализа остатков

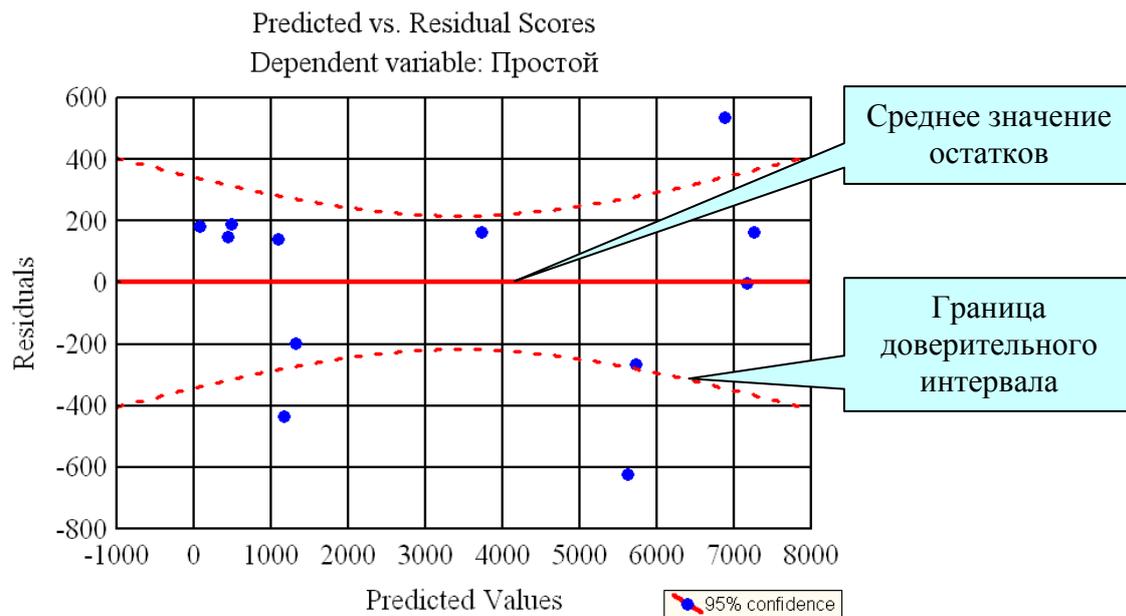


Рис. 4.14. График остатков в линейной зависимости от предсказанных значений

Из этого графика видно, что остатки хаотично разбросаны относительно прямой, в их поведении нет закономерностей. Нет оснований говорить, что остатки связаны между собой, нет также резко выделяющихся остатков. Отсюда можно заключить, что модель достаточно адекватно описывает данные.

Рассмотрим наиболее характерные формы этих графиков.

1. Выделяющиеся точки графика: некоторые из остатков могут по абсолютной величине сильно превосходить все остальные остатки. Если максимальное значение остатка больше некоторого заранее выбранного числа, то наблюдение, имеющее такой остаток, является аномальным (выброс). Величина остатка аномального наблюдения зависит от объема выборки n .

Задача исключения аномальных данных не простая. С одной стороны, одно единственное такое наблюдение может обесценить все результаты регрессионного анализа. Тогда эти наблюдения нужно удалить. С другой стороны, автоматическое удаление резко выделяющихся наблюдений без установления причин их возникновения оправдано только в хорошо обкатанных регрессионных моделях, в которых основной интерес представляют только большинство данных.

2. Если остатки попадают в горизонтальную полосу с центром на оси абсцисс, как обозначено штриховыми линиями на рис. 4.14, модель можно рассматривать как адекватную.

3. Если эта полоса непрерывно расширяется, когда y возрастает, это указывает на отсутствие постоянства дисперсии. В этом случае нужно стабилизировать дисперсию путем преобразований или перейти к взвешенному метод наименьших квадратов [12].

4. Если эта полоса имеет вид линейного тренда (возрастает или убывает), то анализ ошибочен. Отрицательные остатки соответствуют малым значениям предсказанных значений y , положительные остатки – большим значениям. Этот результат может получиться и при ошибочном пропуске свободного члена b_0 . Другой способ исправить ситуацию – включить в регрессионную модель фактор, зависящий от номера наблюдения (или времени).

5. Если график имеет криволинейный вид, то модель неадекватна. В регрессионной модели не учтены факторы, оказывающие существенное влияние на зависимую переменную y . Необходимо вводить дополнительные члены (например, квадратичные и взаимодействия) или провести преобразование наблюдений, а затем повторить все вычисления и анализ остатков.

Очень удобным визуальным способом оценки адекватности регрессионной модели является анализ графика опытных и полученных по регрессионному уравнению значений зависимой переменной. Он строится при помощи кнопки *Predicted vs. observed* окна анализа остатков (рис. 4.15).

Из рис. 4.15 хорошо видно, что линейный вид модели хорошо описывает взаимосвязь переменной «Простой» от месяца и времени работы. Эта связь носит линейный характер.

Важно просмотреть графики зависимости остатков от каждой из независимых переменных. Эти графики полезны для обнаружения нелинейной зависимости от переменных. Их легко просмотреть при помощи кнопки *Residuals vs. independent var.* вкладки *Residuals* (рис. 4.13). Остатки должны быть нормально распределены, т. е. на графике они должны представлять приблизительно горизонтальную полосу одинаковой ширины на всем ее протяжении. Коэффициент корреляции (r) между регрессионными остатками и переменными должен равняться нулю. Присутствие нелинейного тренда в регрессионных остатках вызывает сомнение в адекватности модели и говорит о необходимости пересмотра модели – преобразования или ввода новых переменных, перехода к

нелинейной модели. Может, например, оказаться, что в исходную модель нужно включить в слагаемое x^2 или перейти от x к $\log x$. Линейная модель регрессии предполагает, что переменные не взаимодействуют друг с другом, и изменение одного из них не оказывает никакого влияния на значения других. Чтобы проверить справедливость этого предположения, нужно построить график остатков от произведения x_1x_2 . Если график имеет вид линейного тренда, то в модель нужно ввести x_1x_2 .

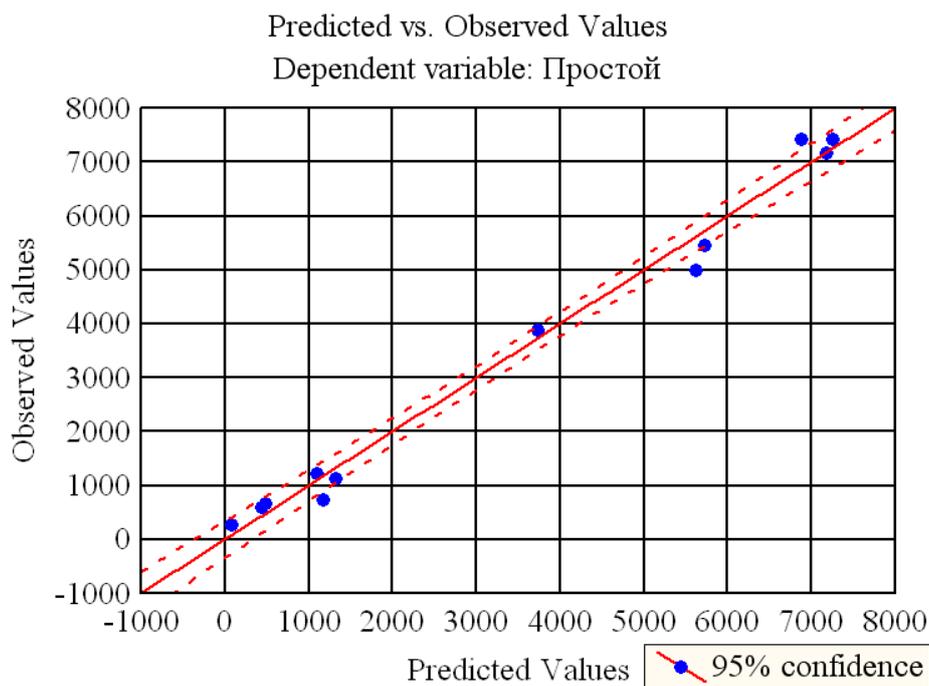


Рис. 4.15. График зависимости наблюдаемых значений зависимой переменной от полученных по регрессионному уравнению

Следует заметить, что мы имеем очень небольшое число данных – всего 12. Поэтому мы используем графические методы оценки адекватности модели. В сложных задачах графические и статистические методы оценки адекватности должны естественно дополнять друг друга.

Кнопка *Redundancy* предназначена для поиска выбросов. Выбросы – это остатки, которые значительно превосходят по абсолютной величине остальные. Выбросы дают данные, которые являются не типичными по отношению к остальным данным и требуют выяснения причин их возникновения. Выбросы должны исключаться из обработки, если они вызваны ошибками измерения. Для выделения выбросов, имеющих в регрессионных остатках, предложены следующие метрики:

1. Расстояние Р.Д. Кука (*Cook's Distance*) показывает расстояние между коэффициентами уравнения регрессии после исключения из обработки каждой точки данных. Большое значение показателя Кука указывает на сильно влияющее наблюдение.

2. Расстояние Махаланобиса (*Mahalanobis Distance*) показывает, насколько каждое наблюдение отклоняется от центра статистической совокупности.

Просмотр величин остатков и специальных критериев для их оценки осуществляется при помощи кнопки Summary окна исследования остатков (рис. 4.13).

4.8. Корреляционный и дисперсионный анализ модели

Частная корреляция – это корреляция между двумя переменными, когда одна или больше из оставшихся переменных удерживаются на постоянном уровне. Частные коэффициенты корреляции, как и парные, могут принимать значения от -1 до $+1$. Кнопка *Partial correlations* окна результатов регрессионного анализа (рис. 4.11) позволяет просмотреть частные коэффициенты корреляции (*Partial Cor.*) между переменными (рис. 4.18).

Variables currently in the Equation; DV: Простой (4regres.sta)							
Variable	Beta in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(9)	p-level
Месяц	0,062537	0,286559	0,032639	0,272400	0,727600	0,89731	0,392925
Работа	0,940149	0,976152	0,490682	0,272400	0,727600	13,48967	0,000000

Рис. 4.16. Результаты расчёта частных коэффициентов корреляции

В идеальной регрессионной модели независимые переменные вообще не коррелируют друг с другом. В самом деле, если две независимые переменные сильно коррелированы с откликом и друг с другом, то достаточно включить в уравнение только одну из них. Обычно включают ту переменную, значения которой легче и дешевле измерять.

Сильная взаимная коррелированность независимых переменных в нашем уравнении затрудняет анализ влияния отдельных факторов на зависимую переменную. Сильная коррелированность переменных в моделях, разрабатываемых для промышленных приложений, является частым явлением. Это приводит к увеличению ошибок уравнения,

уменьшению точности оценивания. Общая эффективность использования регрессионной модели снижается. Поэтому выбор независимых переменных, включаемых в регрессионную модель, необходимо проводить очень тщательно.

Кнопка *ANOVA (Overall goodness of fit)* окна результатов регрессионного анализа (рис. 4.9) позволяет ознакомиться с результатами дисперсионного анализа уравнения регрессии (рис. 4.17). В строках таблицы дисперсионного анализа уравнения регрессии записаны источники вариации: *Regress.* – обусловленная регрессией, *Residual* – остаточная, *Total* – общая. Значения столбцов таблицы: *Sums of Squares* – сумма квадратов, *df* – число степеней свободы, *Mean Squares* – среднее квадратическое значение, *F* – значение *F*-критерия, *p-level* – вероятность нулевой гипотезы для *F*-критерия.

Analysis of Variance; DV: Простой (4regres.sta)					
Effect	Sums of Squares	df	Mean Squares	F	p-level
Regress.	94532789	2	47266395	373,3964	0,000000
Residual	1139265	9	126585		
Total	95672054				

Рис. 4.17. Результаты дисперсионного анализа уравнения регрессии

Из рис. 4.17 видим, что *F*-критерий полученного уравнения регрессии значим на 0,05-уровне. Вероятность нулевой гипотезы (*p-level*) значительно меньше 0,05, что говорит об общей значимости уравнения регрессии.

Кнопка *Predict dependent variable* позволяет рассчитать по полученному регрессионному уравнению значение зависимой переменной по значениям независимых переменных, которые необходимо ввести в появляющемся диалоговом окне.

Кнопка *Descriptive statistics* позволяет просмотреть описательные статистики и корреляционную матрицу с парными коэффициентами корреляции переменных, участвующих в регрессионной модели.

4.9. Фиксированная нелинейная регрессия

Как отмечалось в п. 4.5, в некоторых случаях нелинейные модели с помощью специальных линеаризирующих преобразований могут быть

преобразованы в линейные. Рассмотрим порядок нахождения коэффициентов уравнений нелинейной регрессии, которые через преобразования переменных могут быть приведены к линейной модели.

В качестве примера рассмотрим экономические показатели некоторого предприятия за три квартала текущего года (табл. 4.3). Предположим, что необходимо определить, как влияют на полученную прибыль (y) доходы (x_1), фонд оплаты труда рабочих (x_2) и накладные расходы (x_3). Полученная формула, например, позволит составить прогноз на следующий месяц и оценить значимость каждого фактора.

Таблица 4.3

Экономические показатели некоторого предприятия

Месяц	Прибыль	Доходы	ФОТ рабочих	Накладные расходы
	y	x_1	x_2	x_3
Январь	2839,6	1675,9	567,9	757,6
Февраль	3354,9	2050,2	696,0	900,9
Март	4302,6	2382,1	795,6	954,7
Апрель	6690,5	2798,3	880,6	896,1
Май	5414,0	2735,3	1012,4	974,6
Июнь	4805,3	2552,7	843,1	1017,4
Июль	5680,1	2987,6	1092,5	1072,3
Август	5315,2	3171,8	1046,9	1054,0
Сентябрь	4724,5	2902,7	1099,3	1070,6

В качестве метода разведочного анализа выберем построение трёхмерных графиков. Из рис. 4.18 и 4.19 видно, что зависимость переменной y от x_1 , x_2 и x_3 явно нелинейная. Кроме того, форма графиков подозрительно напоминает гиперболу. Следовательно, можно попытаться найти модель в виде:

$$y = b_1/x_1 + b_2/x_2 + b_3/x_3 + \varepsilon \quad (4.4)$$

Характерно, что экономисты предприятия планировали прибыль исходя из линейной модели, неадекватность которой была очевидна для них же самих.

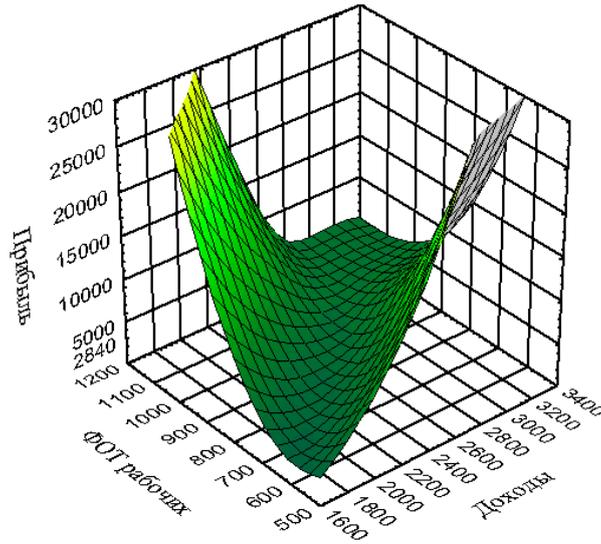


Рис. 4.18. График зависимости переменной y от x_1 и x_2

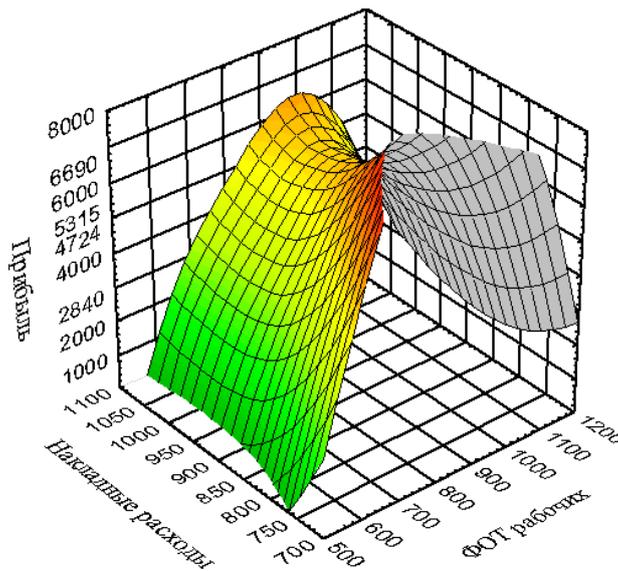


Рис. 4.19. График зависимости переменной y от x_2 и x_3

После запуска модуля фиксированной регрессии *Statistics/ Advanced Linear/Nonlinear Models/Fixed Nonlinear Regression* и выбора переменных после нажатия на кнопку *OK* в диалоговом окне *Non-linear Components Regression* (рис. 4.20) можно выбрать типы преобразования переменных в виде широко распространённых математических функ-

ций. Для нашего примера это $1/x$. Если потребуются какие либо иные преобразования переменных, то эти преобразования нужно делать в таблице с исходными данными, а затем включить полученные фиктивные переменные в качестве зависимых в регрессионную модель.

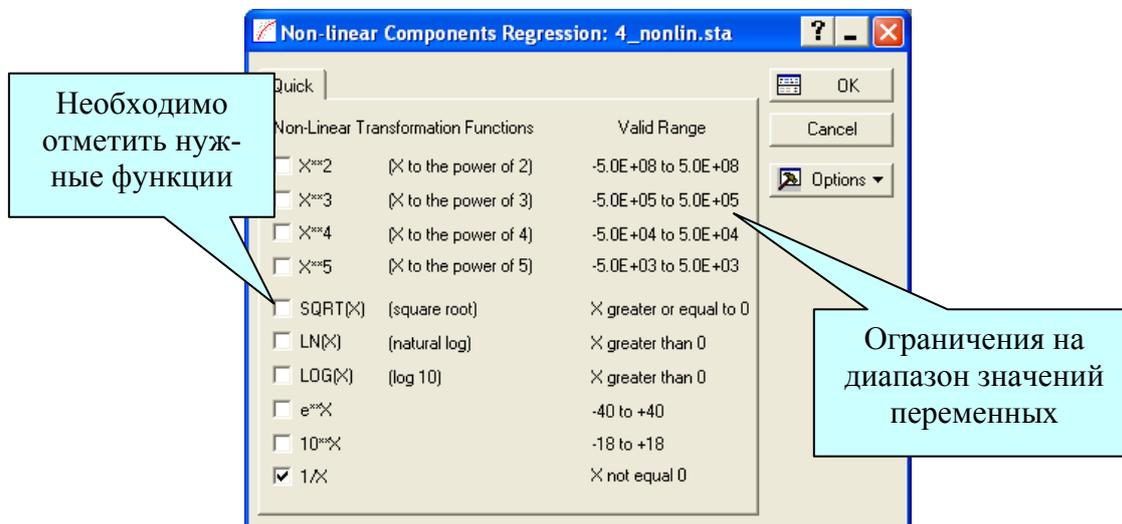


Рис. 4.20. Окно выбора типов преобразования переменных

После того, как тип преобразования переменных определён, необходимо уточнение зависимой и независимых переменных фиксированной нелинейной регрессионной модели. Оно производится на следующем шаге при помощи кнопки *Variables* диалогового окна «Уточнение модели» – *Model Definition* (рис. 4.21). В этом окне установим значение поля **Intercept** на «Set to zero», что позволит получить регрессионную модель без свободного члена уравнения (4.4), то есть $b_0 = 0$. Читатель может самостоятельно убедиться в том, что включение свободного члена не приводит к существенному улучшению модели.

Зависимой переменной (Dependent variables) в нашем случае будет «Прибыль» y ; независимыми (Independent variables) – обратные величины 3, 4 и 5 переменной по списку, то есть $1/x_1$, $1/x_2$ и $1/x_3$ (рис. 4.22).

Результаты появляются при нажатии кнопки ОК (рис. 4.23). Уравнение (4.4) с найденными коэффициентами имеет вид:

$$y = -21671821,4/x_1 - 2444052,3/x_2 + 15608932,1/x_3 + \varepsilon \quad (4.5)$$

Все коэффициенты уравнения значимы по уровню 0,05. Уравнение объясняет 99,24 % ($R^2 = 0,9924$) вариации зависимой переменной. По анализу остатков можно убедиться в адекватности полученной модели.

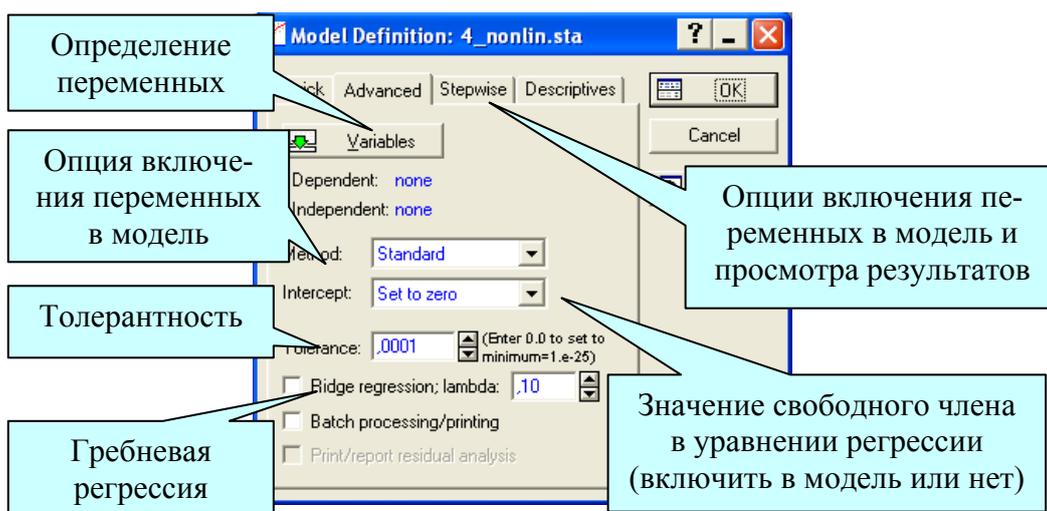


Рис. 4.21. Диалоговое окно Model Definition

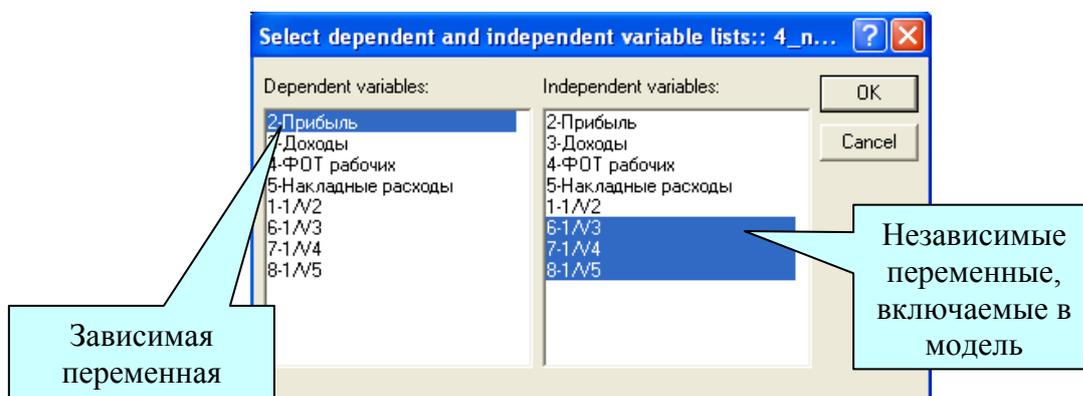


Рис. 4.22. Диалоговое окно выбора переменных

Regression Summary for Dependent Variable: Прибыль (4_.sta)						
R= ,99619175 R ² = ,99239801 Adjusted R ² = ,98859702						
F(3,6)=261,09 p<,00000 Std.Error of estimate: 525,35						
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(6)	p-level
N=9						
1/V3	-1,80743	0,775576	-21671821	9299465	-2,33044	0,058606
1/V4	-0,59693	0,713163	-2444052	2919948	-0,83702	0,434647
1/V5	3,34061	0,300167	15608932	1402524	11,12917	0,000031

Рис. 4.23. Результаты регрессионного анализа

Ошибка уравнения составляет 525,35. Если сравнить абсолютную величину ошибки со средним значением зависимой переменной:

$$525,35/4791,8527*100 \% = 10,96 \%,$$

то заметим, что она довольно велика. Следовательно, модель нуждается в совершенствовании.

4.10. Пошаговая регрессия

Поиск наилучшей регрессионной модели – это громоздкий процесс. При помощи опции *Method* (рис. 4.21) пользователь может отказаться от стандартного регрессионного анализа (Standard) и воспользоваться методами пошагового включения переменных в регрессионную модель (*Forward stepwise*) или пошагового исключения переменных (*Backward step wise*) из регрессионной модели. Эти методы можно использовать в сложных системах с большим числом переменных. Опция *Displaying results* вкладки *Stepwise* позволяет просматривать итоговые результаты регрессионного анализа (*Summary only*) или после каждого шага включения или исключения переменных (*At each step*).

Воспользуемся методом пошагового включения переменных для нахождения наилучшего регрессионного уравнения для данных из табл. 4.4. В качестве независимых переменных, которые потенциально могут быть включены в модель для y примем переменные x_1, x_2, x_3 , их обратные значения $1/x_1, 1/x_2, 1/x_3$ и логарифмы $\ln(x_1), \ln(x_2), \ln(x_3)$. В результате окно уточнения переменных модели примет вид (рис. 4.24).

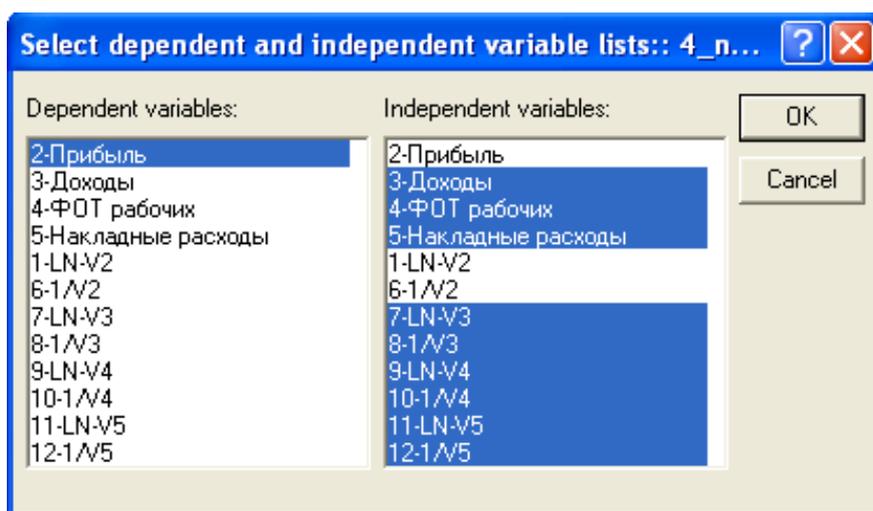


Рис. 4.24. Диалоговое окно выбора переменных

Для пошаговых методов можно установить величину *Tolerance* (допуск) и величины частного *F*-критерия для включения в модель (*F to enter*) и исключения из неё (*F to remove*) (рис. 4.21). Величина допуска является границей для включения в модель переменных, допуск на которые меньше установленного. Если величина допуска мала, то переменная несёт малую дополнительную информацию, она незначима и включение её в модель не целесообразно. Характерно, что новая независимая переменная, включаемая в модель, может сильно повлиять на зависимую переменную. Наоборот, если она включается в модель после других переменных, она может уже мало влиять на зависимую переменную, например, из-за сильной коррелированности с переменными, уже включёнными в модель. По умолчанию в пакете Statistica переменная включается в модель, если частный *F*-критерий больше или равен 1. Численное значение *F*-критерия для включения никогда не выбирается меньшим, чем численное значение *F*-критерия для исключения.

Воспользуемся установками по умолчанию и запустим анализ. В результате процедуры пошагового включения переменных в регрессионную модель получено следующее уравнение (рис. 4.25):

$$y = 85001,1 - 22705835,2/x_1 - 10350,98 * \ln(x_3) + \varepsilon \quad (4.6)$$

Regression Summary for Dependent Variable: Прибыль						
R= ,96521693 R²= ,93164372 Adjusted R²= ,90885829						
F(2,6)=40,888 p<,00032 Std.Error of estimate: 356,81						
N=9	Beta	Std.Err. of Beta	B	Std.Err. of B	t(6)	p-level
Intercept			85001	17684	4,80674	0,002980
1/V3	-1,72306	0,230057	-22705835	3031601	-7,48972	0,000293
LN-V5	-0,98550	0,230057	-10351	2416	-4,28372	0,005184

Рис. 4.25. Результаты регрессионного анализа, полученные методом *Forward stepwise*

Из рис. 4.25 видно, что все коэффициенты уравнения значимы по уровню 0,05 ($p\text{-level} < 0,05$). Это уравнение объясняет 93,16 % ($R^2 = 0,9316$) вариации зависимой переменной. Средняя ошибка составляет 356,81, что почти в 1,5 раза меньше, чем в модели (4.5).

Итак, согласно модели (4.5), прибыль тем больше, чем больше фонд оплаты труда. В уточнённой модели (4.6), прибыль зависит от дохода и накладных расходов, но не зависит от фонда оплаты труда. Ком-

пьютер исключил как незначимую эту переменную, поступив куда умнее руководителя предприятия, вечно экономящего на зарплате.

В заключение раздела отметим, что некоторые закономерности в данных можно найти чисто математическим путём, между тем как непосредственное наблюдение не позволяет установить даже их присутствия.

4.11. Наилучшие регрессионные модели

Технические навыки при работе в системе Statistica – это ремесло, которому может научиться каждый. Поиск наилучшей регрессионной модели – это искусство, у которого нет рецептов. С одной стороны, для получения надёжных прогнозов значений отклика y в модель нужно включать как можно больше независимых переменных. С другой стороны, с увеличением их числа возрастает дисперсия прогноза и увеличиваются затраты, связанные с получением информации о дополнительных переменных, поэтому желательно включать в уравнение как можно меньше переменных. Тем не менее, существуют некоторые общие требования к регрессионным моделям:

Регрессионная модель должна объяснять не менее 80 % вариации зависимой переменной, т. е. $R^2 > 0,8$.

Чем меньше сумма квадратов остатков, чем меньше стандартная ошибка оценки и чем больше R^2 , тем лучше уравнение регрессии.

Коэффициенты уравнения регрессии и его свободный член должны быть значимы по уровню 0,05.

Стандартная ошибка оценки зависимой переменной по уравнению должна составлять не более 5 % среднего значения зависимой переменной.

Остатки от регрессии должны быть без заметной автокорреляции ($r < 0,3$), нормально распределены и без систематической составляющей.

Отметим, что понятие «наилучшая регрессионная модель» является субъективным, так как нет никакой единой статистической процедуры для выбора соответствующего подмножества независимых переменных.

4.12. Гребневая регрессия

В основе рассмотренного ранее регрессионного анализа лежит метод наименьших квадратов. Его недостатком является относительно небольшая устойчивость к изменениям входных данных. В настоящее

время широко стали применяться альтернативные регрессионные модели, одной из которых является *гребневая регрессия*, которая отличается устойчивостью для случаев сильной коррелированности зависимых переменных друг с другом. В отличие от метода наименьших квадратов, дающего несмещённые оценки коэффициентов уравнения, в методе гребневой регрессии оценки смещённые, но при этом они имеют меньшую дисперсию. Поэтому такие оценки могут давать более точные и приемлемые для практического использования модели.

Для расчёта коэффициентов уравнения гребневой регрессии следует отметить чекбокс в опции *Ridge regression* диалогового окна *Model Definition* (рис. 4.21).

При практическом использовании метода гребневой регрессии одним из основных вопросов является выбор параметра λ (*lambda*). Существуют численные методы расчёта этого параметра, но чаще используют простой опытный подход: начинают расчёт при $\lambda = 0$, увеличивают параметр с малым шагом, например, 0,001 и следят за ошибкой регрессии и коэффициентами уравнения. Ошибка не должна увеличиваться, а коэффициенты должны стабилизироваться и при дальнейшем увеличении параметра мало изменяться. Значение принятого параметра λ является мерой смещения оценок от истинного значения, поэтому стараются не придавать λ слишком больших значений. Обычно λ выбирают меньше 0,5. При $\lambda = 0$ уравнение имеет коэффициенты классического метода наименьших квадратов.

4.13. Задания для самостоятельной работы

Задание 1. С помощью модуля *Statistics/Basic Statistics/Correlation Matrices* рассчитать коэффициенты корреляции для переменных, состоящих из строк VarN, VarN+1 и VarN+2 (см. Табл. 4.4), где N – номер варианта. Выбирать опции *One variable list u Summary: Correlation matrix* (таблица с коэффициентами корреляции); *Scatterplot matrix for selected variables* (графическое отображение зависимостей). Построить корреляционную матрицу. Сделать выводы о взаимной зависимости переменных.

Транспонирование строк и столбцов при вставке данных из таблицы можно провести с помощью приложения Excel. Скопируйте данные в Excel через буфер обмена (Ctrl-C, Ctrl-V, Ctrl-C), создайте новую таб-

лицу (Ctrl-N), выполните операцию «Правка/ Специальная вставка/ (отметить чекбокс «транспонирование»).

Таблица 4.4

Варианты задания и переменные

N										
1	48	30	43	44	30	34	32	43	40	46
2	25	21	34	49	39	37	45	49	31	49
3	43	46	34	35	42	30	41	34	42	22
4	38	40	26	47	34	42	38	20	38	36
5	30	13	41	40	40	15	35	11	38	45
6	37	12	38	36	14	39	32	54	43	39
7	23	30	32	36	32	34	49	18	49	50
8	37	20	44	28	44	35	45	34	33	41
9	43	45	50	14	33	39	41	39	46	31
10	40	52	44	39	35	54	33	42	42	36
11	44	51	45	19	34	44	40	37	43	32
12	33	42	40	35	37	13	48	48	50	32
13	40	48	45	23	36	36	42	40	37	30
14	44	50	46	39	31	48	44	42	36	51
15	44	50	54	37	33	34	42	43	43	47
16	33	48	18	42	15	32	34	14	39	45
17	48	26	31	34	38	36	46	49	40	48
18	42	47	35	34	41	33	41	35	43	42
19	39	37	47	27	33	22	37	19	19	37
20	43	41	30	39	38	36	36	34	42	46
21	39	44	37	35	43	38	33	47	45	38
22	37	48	38	52	40	45	44	42	38	40
23	44	46	37	34	41	37	41	39	30	38
24	32	41	48	36	51	36	33	39	45	40
25	34	41	38	34	33	27	51	45	27	38
26	42	37	46	41	47	36	30	45	41	40
27	37	37	39	42	48	41	36	39	33	47
28	43	49	27	31	41	46	40	36	36	42
29	41	46	33	37	47	35	31	29	30	36
30	48	38	37	34	40	34	36	50	48	39
31	30	38	43	41	44	45	38	37	46	50

N										
32	41	48	41	43	47	37	42	34	32	44
33	37	48	46	41	41	37	37	48	49	46
34	38	44	50	37	47	27	48	37	46	38
35	48	47	38	52	34	36	34	41	41	32
36	31	43	34	46	37	40	41	39	32	42
37	47	33	51	41	40	45	37	36	27	36
38	37	42	46	35	34	38	45	36	20	40
39	34	48	30	51	33	41	44	42	39	39
40	45	45	41	40	36	27	50	44	41	48
41	36	36	32	32	36	49	27	45	30	35
42	40	38	45	40	40	50	42	37	50	39
43	43	38	30	59	42	41	33	42	38	44
44	44	41	47	52	51	38	50	39	50	48
45	49	43	52	50	30	30	26	50	27	49
46	27	49	46	39	47	26	49	52	29	44
47	51	53	48	49	53	45	27	43	48	44

Задание 2. Изменить те же данные следующим образом. Первую переменную (пусть это Var1) оставить неизменной, а Var2 сделать равной $2*Var1$; Var3 сделать равной $2*Var1+Var1^2$. Для этого необходимо дважды щелкнуть левой кнопкой мыши по имени переменной и в появившемся окне «*Long name (label or formula)*» записать формулу

$$=2*v1$$

для Var2 или

$$=2*v1+v1*v1$$

для Var3 соответственно.

Рассчитать коэффициенты корреляции, построить корреляционную матрицу. Сделать выводы о взаимной зависимости переменных.

Задание 3. Исходные данные (файл http://ieeetpu.ru/statlab/var_6_1.sta) представляют собой журналы рейтинговых оценок, полученных студентами кафедры КИСМ (контрольные точки, оценка за экзамен, дисциплина). Провести статистические испытания (на ваше усмотрение) и интерпретировать полученные результаты. Необходимо дать ответ на следующие вопросы.

1. Найти закон распределения экзаменационных оценок. Обосновать свой выбор. Распределены ли экзаменационные оценки по нормальному закону? Как это выявить? Согласуется ли полученный результат с общепринятым о случайных процессах?

2. Распределены ли рейтинговые оценки по нормальному закону? Доказать с помощью программы Statistica и сравнить с предыдущими результатами.

3. Зависят ли результаты экзаменов от рейтинга в семестре? Какими методами можно доказать или опровергнуть закономерность? Воспользоваться этими методами и подробно описать проделанную работу.

Задание 4. Исходные данные –

http://ieeep.tpu.ru/statlab/ege_2006.rar

http://ieeep.tpu.ru/statlab/ege_2007.rar

http://ieeep.tpu.ru/statlab/ege_2008.rar

http://ieeep.tpu.ru/statlab/ege_2009.rar

представляют собой статистику основных результатов единого государственного экзамена. Провести статистические испытания (на ваше усмотрение) и интерпретировать полученные результаты. Цель исследования – ответить на вопросы о том, 1) насколько средняя успеваемость по физике и математике в Томской области лучше или хуже, чем в целом по России; 2) как и насколько существенно она изменяется с 2006 по 2009 годы.

Задание 5. Исходные данные (файл <http://ieeep.tpu.ru/statlab/rector.pdf>) представляют собой отчёт ректора ЮФУ за 2009 год.

Провести статистические испытания (на ваше усмотрение) и интерпретировать полученные результаты. Ответить на следующие вопросы.

1. Как зависит уровень зарплаты в подразделениях университета от фонда оплаты труда.

2. Как изменится доходность бюджета при увеличении зарплаты на 10 %.

3. Какие улучшающие вмешательства возможны для увеличения доходной части бюджета университета.

ГЛАВА 5. НЕЛИНЕЙНЫЕ МОДЕЛИ ПРОЦЕССОВ

5.1. Нелинейная регрессия

Если модель сильно нелинейна по параметрам, предполагаемый вид нелинейной функции (4.1) может быть задан пользователем. В этом случае для оценки коэффициентов регрессии b необходимо воспользоваться модулем «Nonlinear Estimation» – «Нелинейное оценивание».

Проведём нелинейный регрессионный анализ данных на модельном примере поиска зависимости $y=f(x)$ (табл. 5.1).

Таблица 5.1

Данные для регрессионного анализа

Номер	Аргумент x	Функция y
1	0,5	0,0964
2	1,0	-0,6562
3	1,5	-1,185
4	2,0	-1,4474
5	2,5	-1,4406
6	3,0	-1,2017
7	3,5	-0,8018
8	4,0	-0,3314
9	4,5	0,1159
10	5,0	0,4611
11	5,5	0,6542
12	6,0	0,6814
13	6,5	0,5667
14	7,0	0,3642
15	7,5	0,145
16	8,0	-0,0185
17	8,5	-0,07
18	9,0	0,0187
19	9,5	0,2405
20	10,0	0,5535

Предварительно построим график данных (модуль *Graphs / Scatter-plots*). На панели настройки установим *Graph Type: Regular; Fit: Spline*.

График представлен на рис. 5.1. Из графика видно, что зависимость между аргументом и функцией сильно нелинейная.

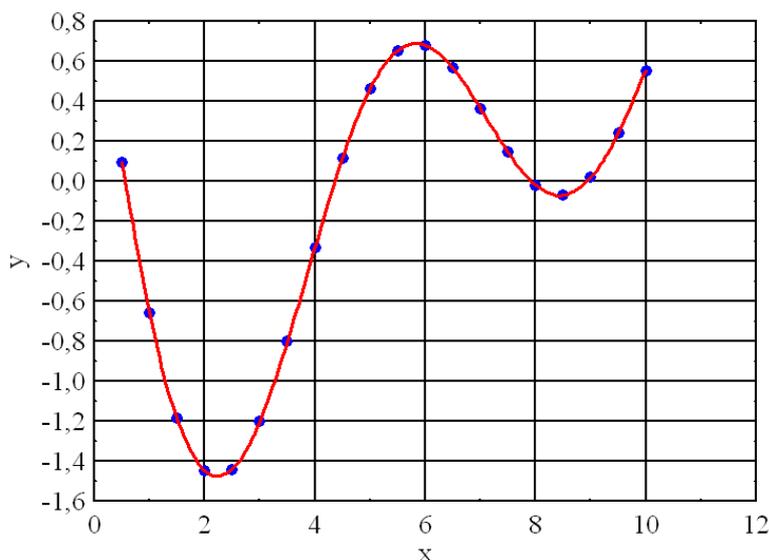


Рис. 5.1. График функции (табл. 5.1)

Вызовем модуль *Statistics / Advanced Linear/Nonlinear models / Non-linear estimation / User-specified regression, custom loss function* (определяемая пользователем регрессия и функция потерь). Кнопкой «*Function to be estimated*» введём функцию (рис. 5.2):

$$v2 = \exp(b1 * v1) + \sin(b2 * v1) + \sin(b3 * v1).$$

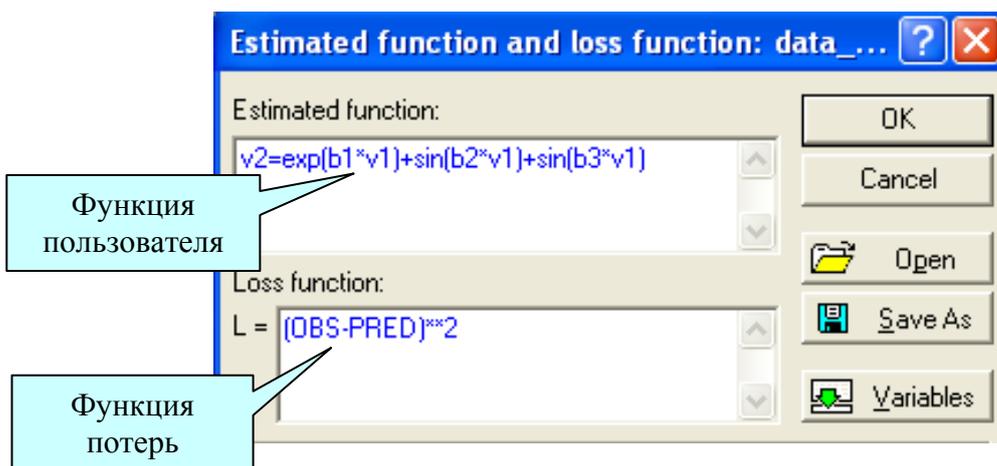


Рис. 5.2. Окно ввода функции пользователя

Функция потерь (Loss function) по умолчанию определена как сумма квадратов разностей наблюдаемых и предсказанных значений

$((OBS-PRED)**2)$, следовательно, оценки параметров вычисляются методом наименьших квадратов. Аналогичный модуль *User-specified regression, least squares* не требует ввода функции потерь.

В появившемся далее окне *Model Estimation* (метод оценивания) во вкладке «*Advanced*» можно задать вычислительный метод, например, *Quasi-Newton*, максимальное число итераций (*Maximum number of iterations*), критерий сходимости (*Convergence criterion*), кнопкой «*Start values:*» можно задать начальные значения искомых коэффициентов, кнопкой «*Initial step sizes:*» – шаг изменения параметров. Чтобы получить среднеквадратические ошибки оценок параметров, нужно включить опцию *Asymptotic standard errors*.

Выбор начальных значений параметров является очень важным, так как неудачное начальное приближение может привести к медленной сходимости и даже к расходимости процесса вычислений.

Если теперь запустить программу оценивания параметров, нажав *OK*, то может появиться следующее сообщение: «*Error in function: change start values/precision/step size*» (вычисления не могут выполняться: следует изменить начальные значения параметров/ точность вычислений/ величину шага).

Результаты вычислений (см. табл. 5.2) содержат: значение функции потерь (*Loss function*) по шагам итераций (*Final value*), оценки параметров, коэффициент множественной корреляции R , долю дисперсии исходных данных, объясняемую моделью (коэффициент детерминации R^2).

Таблица 5.2

Результаты процедуры оценивания

Model is: $v2=\exp(b1*v1)+\sin(b2*v1)+\sin(b3*v1)$	
Dependent variable: y	Independent variables: 1
Loss function: $(OBS-PRED)**2$	
Final value: ,000000026	
Proportion of variance accounted for: ,999999997 R = ,999999999	

Оценки параметров, их среднеквадратические значения, t -статистики для проверки гипотезы о равенстве нулю коэффициентов регрессии и соответствующие уровни значимости выводятся при нажатии кнопки «*Summary: Parameters & standard errors*» (рис. 5.3). Из рис. 5.3 видно, что регрессия высокосignификантна. Это можно доказать, подставив

найденные коэффициенты в формулу и рассчитав значения функции для соответствующих значений аргумента (табл. 5.1).

Model: $v_2 = \exp(b_1 * v_1) + \sin(b_2 * v_1) + \sin(b_3 * v_1)$			
Dep. var: y Loss: (OBS-PRED)**2			
Final loss: ,000000026 R=1,0000 Variance explained: 100,00%			
N=20	b1	b2	b3
Estimate	-0,50	-0,90	-0,50
Std.Err.	0,00	0,00	0,00
t(17)	-18859,65	-302830,87	-131206,14
p-level	0,00	0,00	0,00

Рис. 5.3. Результаты оценивания коэффициентов

Нажав кнопку «*Fitted 2D function & observed values*» на панели результатов, получим график функции и исходных данных. Панель результатов содержит также различные опции для анализа остатков, позволяющие проверить гипотезу об адекватности модели результатам наблюдений.

5.2. Полиномиальная регрессия

У рассмотренной в п. 5.1 методики есть существенный недостаток: невозможно доказать правильность выбора модельной функции даже в случае высокого коэффициента детерминации. Можно попытаться аппроксимировать данные полиномом. В случае слабой нелинейности полинома второго-третьего порядка вполне бывает достаточно для практических целей. В пакете Statistica реализованы полиномиальные модели вида:

$$y(x) = b_0 + b_1x + b_2x^2 + \varepsilon$$

$$y(x) = b_0 + b_1x + b_2x^2 + b_3x^3 + \varepsilon.$$

5.3. Регрессионное моделирование в экономике

Экономические процессы принципиально отличаются от технологических большей степенью неопределённости. Предпочтения людей меняются фантастически непредсказуемо, и это заставляет искать изощрённые математические модели, более-менее удовлетворяющие практическим целям планирования и организации производства. Не смотря

на то, что современная экономическая наука достигла небывалых высот в использовании новейших достижений математического анализа, на предприятиях (по крайней мере тех, которые консультировал автор) наблюдается полное отсутствие экономического моделирования и планирования, так как в сегодняшней конкурентной среде трудно воспринимать всерьёз примитивные линейные модели, которые можно было там лицезреть.

Рассмотрим статью [13], в которой на примере Амурского государственного университета проведено исследование спроса населения на образовательные услуги с помощью регрессионного анализа, т. е. подбирается некоторая функция, которая в среднем отражает зависимость количества поданных заявлений от группы факторов. Статья была выбрана из бесчисленного множества подобных ей. Она ярче всего демонстрирует результат применения регрессионного анализа методом «не приходя в сознание».

Спрос оценивался по количеству поданных заявлений (зависимая переменная y). В качестве независимых переменных, влияющих на количество поданных заявлений, были выбраны x_1 – средняя пенсия, x_2 – рождаемость; x_3 – количество выпускников школ, x_4 – средняя заработная плата. Исходные данные из статьи [13] приведены в табл. 5.2.

Таблица 5.3

Данные для регрессионного анализа (спрос на образовательные услуги)

Год	y	x_1	x_2	x_3	x_4
1994	1191	100	10200	7400	300
1995	2003	229	10100	7400	611
1996	3256	348	9800	7500	902
1997	4842	395	9400	7900	1103
1998	6260	413	9750	8500	1201
1999	6800	536	9200	9500	1302
2000	7188	817	9400	10000	1917
2001	7350	1119	10100	9900	2756

С использованием методов множественной регрессии была получена следующая модель:

$$y = 0,30x_1 + 0,20x_2 + 0,70x_3 + 0,75x_4 + \varepsilon \quad (5.1)$$

Коэффициент детерминации составил $R^2=0,72$, минимальное значение среднеквадратической ошибки равно $S=35,41$. По тем же данным с помощью пакета Statistica мы получили другую модель:

$$y = -24,4724x_1 - 1,252x_2 + 1,8532x_3 + 10,6119x_4 + \varepsilon \quad (5.2)$$

с коэффициентом детерминации $R^2=0,997$, минимальным значением среднеквадратической ошибки $S=0,16$. Данная модель выгодно отличается от (5.1). Это убедительно демонстрирует график рассеяния (рис. 5.4).

Формальное применение статистических методов без скрупулёзного анализа их пригодности для обработки конкретного типа данных, как правило, приводит к совершенно невероятным результатам

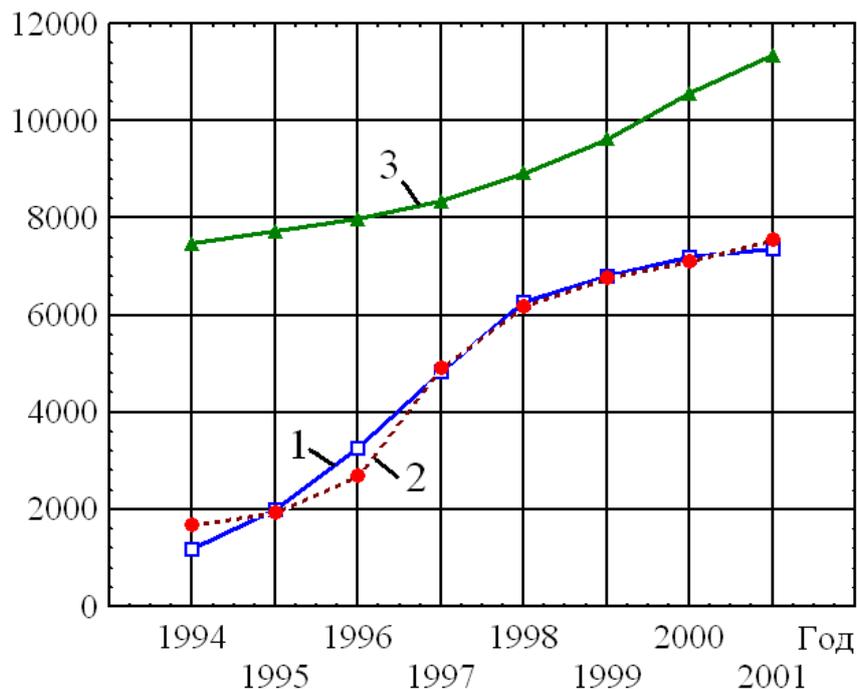


Рис. 5.4. График исходных данных (1), регрессионная модель, найденная с помощью программы Statistica (2) и в работе (3)

По модели (5.1) в статье [13] сделан вывод о том, что наибольшее влияние на результат у оказывает средняя заработная плата (x_4). На втором месте – количество выпускников школ (x_3). Отмечено, что с ростом

рождаемости (x_2) соответственно увеличивается и количество выпускников школ, являющихся потенциальными потребителями образовательных услуг образовательных учреждений, т. е. и количество поданных заявлений возрастает. Указанный вывод отрицает необходимость фактора x_2 в модели. Что касается заработной платы и пенсии, то в [13] отмечено, что чем выше средняя заработная плата, тем большее количество родителей сможет дать детям образование, т. е. с ростом зарплаты и пенсии количество поданных заявлений также возрастет. Этот вывод малоубедителен, так как никакие родители не пожалеют никаких денег на образование детей. В модели (5.2) коэффициенты для факторов x_1 и x_2 отрицательные, что совершенно не удивительно для такого набора исходных данных. Расчёты подтверждают незначимость фактора x_1 в модели.

5.4. Задания для самостоятельной работы

Задание 1. Прочитать файл с исходными данными <http://ieeee.tpu.ru/statlab/variant6.sta>. Считать первый столбец независимой переменной (Argument), а остальные (Var-N) – зависимой переменной. Номер зависимой переменной соответствует вашему варианту по списку группы. Построить график зависимости второй переменной Var-N от первой: *Graphs / 2D Scatterplot*; во вкладке *Advanced* выбрать *Off*, в опции «свойства графика» соединить точки линиями. Выбирая в настройках *Advanced* аппроксимирующую функцию, определить, какая функция является наилучшей аппроксимацией для предложенных данных. Объяснить, почему.

Задание 2. В модуле "Nonlinear Estimation" – «Нелинейное оценивание» собраны процедуры, позволяющие оценить нелинейные зависимости между данными. Вы можете выбрать различные модели зависимостей, задать собственную функцию, выбрать метод оценивания неизвестных параметров.

Провести нелинейный регрессионный анализ данных задания 1. Для этого воспользоваться модулем *Statistics / Advanced Linear/ Nonlinear models / Nonlinear estimation / user-specified regression, least squares*. Кнопкой *Function to be estimated* ввести модельную функцию (см. табл. 5.4). Кнопкой *Variables* ввести зависимую переменную.

Таблица 5.4

Расчётные формулы и интервал для выбора начального приближения всех коэффициентов b_1 , b_2 , b_3 для вариантов заданий

Вариант	Формула, записанная в терминах программы Statistica	нижняя граница интервала	верхняя граница интервала
1	$v2=\exp(b1*v1)+\sin(b2*v1)+\sin(b3*v1)$	-0,9	-0,4
2	$v3=\exp(b1*v1)+\sin(b2*v1)+\sin(b3*v1)$	-1,3	-0,4
3	$v4=\exp(b1*v1)+\sin(b2*v1)+\sin(b3*v1)$	-2,1	-0,4
4	$v5=\exp(b1*v1)+\cos(b2*v1)+\sin(b3*v1)$	-1,8	-0,2
5	$v6=\exp(b1*v1)+\sin(b2*v1)+\cos(b3*v1)$	-1,4	-0,4
6	$v7=\exp(b1*v1)+\cos(b2*v1)+\cos(b3*v1)$	-1,5	-0,5
7	$v8=\exp(b1*v1)+\cos(b2*v1)+\cos(b3*v1)$	-1,9	-0,1
8	$v9=\exp(b1*v1)+\sin(b2*v1)+\cos(b3*v1)$	-2,1	-0,4
9	$v10=\sin(b1*v1)+\sin(b2*v1)+\exp(b3*v1)$	-1,2	-0,3
10	$v11=\sin(b1*v1)+\sin(b2*v1)+\exp(b3*v1)$	-1,3	0,1
11	$v12=\sin(b1*v1)+\sin(b2*v1)+\exp(b3*v1)$	-2,1	-0,3
12	$v13=\cos(b1*v1)+\sin(b2*v1)+\exp(b3*v1)$	-1,8	0,1
13	$v14=\sin(b1*v1)+\cos(b2*v1)+\exp(b3*v1)$	-1,1	1,3
14	$v15=\cos(b1*v1)+\cos(b2*v1)+\exp(b3*v1)$	-1,3	-0,3
15	$v16=\cos(b1*v1)+\cos(b2*v1)+\exp(b3*v1)$	-1,5	-0,8
16	$v17=\sin(b1*v1)+\cos(b2*v1)+\exp(b3*v1)$	-1,5	-0,5
17	$v18=\exp(b1*v1)+\sin(b2*v1)+\sin(b3*v1)$	-0,9	-0,1
18	$v19=\exp(b1*v1)+\sin(b2*v1)+\sin(b3*v1)$	-0,9	0,1
19	$v20=\exp(b1*v1)+\sin(b2*v1)+\sin(b3*v1)$	-2,1	-0,8
20	$v21=\exp(b1*v1)+\cos(b2*v1)+\sin(b3*v1)$	-1,7	0,1
21	$v22=\exp(b1*v1)+\sin(b2*v1)+\cos(b3*v1)$	-1,1	1,6
22	$v23=\exp(b1*v1)+\cos(b2*v1)+\cos(b3*v1)$	-1,3	0,1
23	$v24=\exp(b1*v1)+\cos(b2*v1)+\cos(b3*v1)$	-1,1	-0,5
24	$v25=\exp(b1*v1)+\sin(b2*v1)+\cos(b3*v1)$	-1,5	-0,1
25	$v26=\exp(b1*v1)+\sin(b2*v1)+\sin(b3*v1)$	-1,1	-0,1
26	$v27=\exp(b1*v1)+\sin(b2*v1)+\sin(b3*v1)$	-0,9	0,2
27	$v28=\exp(b1*v1)+\sin(b2*v1)+\sin(b3*v1)$	-2,1	-0,3
28	$v29=\exp(b1*v1)+\cos(b2*v1)+\sin(b3*v1)$	-1,8	0,1
29	$v30=\exp(b1*v1)+\sin(b2*v1)+\cos(b3*v1)$	-1,1	1,6
30	$v31=\exp(b1*v1)+\cos(b2*v1)+\cos(b3*v1)$	-1,4	0,1

В появившемся окне во вкладке *Advanced* задать критерий сходимости 10^{-3} (по умолчанию 10^{-6}), кнопкой *Start values* ввести произвольные начальные значения искомым коэффициентов (b_1 , b_2 , b_3).

Примечание: Указанные в табл. 5.4 интервалы приведены для удобства разумного выбора начального приближения для коэффициентов, чтобы не очень долго промучиться со сходимостью решения к истинному. Эти значения даны только для справки.

Показать, что полученные коэффициенты нелинейной регрессии далеки от реальных при неудачном начальном приближении. Включить типичные графики в отчёт. Как зависит успех расчётов от хорошего начального приближения? Как это влияет на скорость сходимости? Сделать соответствующие выводы. (*Примечание:* скорость сходимости можно оценить по истории итераций, которую можно просмотреть, нажав кнопку *Iterations history*).

Найти коэффициенты нелинейной регрессии. Провести анализ остатков и показать, что найдено наилучшее решение.

Задание 3. Ввести в формулу модельной функции точное значение для двух коэффициентов, найденных в предыдущем задании и вновь провести регрессионный анализ поиска только одного коэффициента, введя для него «хорошее» начальное приближение. Насколько уменьшилась ошибка регрессии в этом случае? Увеличилась ли скорость сходимости?

Задание 4. Исходные данные см. в табл. 5.5 по номеру варианта

Таблица 5.5

Файлы с данными для вариантов заданий

Номер варианта	Файл с данными для работы
1	http://ieee.tpu.ru/statlab/sale1.sta
2	http://ieee.tpu.ru/statlab/sale2.sta
3	http://ieee.tpu.ru/statlab/income1.sta
4	http://ieee.tpu.ru/statlab/income2.sta
5	http://ieee.tpu.ru/statlab/income3.sta
6	http://ieee.tpu.ru/statlab/income4.sta
7	http://ieee.tpu.ru/statlab/income5.sta

Построить разумную с вашей точки зрения нелинейную регрессионную модель по предложенным данным. Оценить адекватность модели по остаткам.

Задание 5. Исходные данные (файл http://ieeetpu.ru/statlab/var_6_3.sta) представляют собой объём продаж щебней разных фракций (переменные) во времени (наблюдения). Провести статистические испытания (на ваше усмотрение) и интерпретировать полученные результаты. Необходимо выявить лидеров продаж, построить регрессионные модели и показать их пригодность или непригодность для построения прогноза продаж на следующий месяц.

Задание 4. Исходные данные (файл http://ieeetpu.ru/statlab/var_6_4.sta) представляют собой статистику времени работы и простоя разного оборудования на промышленной площадке (переменные) по месяцам (наблюдения). Целью улучшающего вмешательства является сокращение времени простоя оборудования. Провести статистические испытания (на ваше усмотрение) и интерпретировать полученные результаты. Необходимо выявить наиболее и наименее проблемное оборудование, построить регрессионные модели и показать их пригодность или непригодность для построения прогноза времени работы и простоя на следующий месяц. Является ли простой абсолютно случайным, или существует какая-то особая причина, требующая выявления и устранения?

Задание 5. Исходные данные (файл http://ieeetpu.ru/statlab/var_6_5.sta) представляют собой статистику отгрузки щебней разных фракций (var4) во времени (var2, var3). Известна вместимость каждой машины и её гаражный номер (var5, var6, var7, var8, var10), грузополучатель (var9) и номер заявки (var1). Целью улучшающего вмешательства является увеличение объёмов продаж и повышение эффективности работы грузовиков. Провести статистические испытания (на ваше усмотрение) и интерпретировать полученные результаты. Необходимо выявить клиентов, которые приносят компании наибольшую прибыль, наиболее проблемные моменты в работе, эффективность использования грузовиков. Являются ли продажи абсолютно случайным явлением или существует возможность их прогнозирования на ближайшее время?

Задание 6. Исходные данные (файл http://ieeetpu.ru/statlab/var_6_6.sta) представляют собой статистику отгрузки щебней разных фракций (var6) во времени (var2, var3). Известна вместимость каждой машины и её гаражный номер (var7, var8, var9, var10, var12), грузополучатель (var11) и номер заявки (var1). Целью улучшающего вмешательства является сокращение времени простоя грузовиков на площадке

($\text{var5}=\text{var4}-\text{var3}$) и повышение эффективности работы грузовиков. Провести статистические испытания (на ваше усмотрение) и интерпретировать полученные результаты. Необходимо оценить степень связи времени простоя с влияющими на него факторами. Необходимо выявить клиентов, которые приносят компании наибольшую прибыль, наиболее проблемные моменты в работе, эффективность использования грузовиков. Является ли простой грузовиков абсолютно случайным явлением или существует возможность влияния на него простыми мерами?

Задание 7. Исходные данные (файл http://ieeee.tpu.ru/statlab/var_6_7.sta) представляют собой статистику отгрузки щебней разных фракций (var6) во времени (var2 , var3). Известна вместимость каждой машины и её гаражный номер (var10 , var12), грузополучатель (var11) и номер заявки (var1). Целью улучшающего вмешательства является сокращение времени простоя грузовиков на площадке ($\text{var5}=\text{var4}-\text{var3}$) и повышение эффективности работы грузовиков. Провести статистические испытания (на ваше усмотрение) и интерпретировать полученные результаты. Необходимо оценить степень связи времени простоя с влияющими на него факторами. Необходимо выявить клиентов, которые приносят компании наибольшую прибыль, наиболее проблемные моменты в работе, эффективность использования грузовиков. Является ли простой грузовиков абсолютно случайным явлением или существует возможность влияния на него простыми мерами?

ГЛАВА 6. КОНТРОЛЬ КАЧЕСТВА

6.1. Статистические методы контроля качества

Качество относится к числу важнейших критериев функционирования предприятия в условиях относительно насыщенного рынка и преобладающей неценовой конкуренции. Эффективно управлять процессами производства – значит активно использовать экономические и организационные рычаги воздействия на разработку, производство и эксплуатацию изделий. Качество продукции обеспечивается в первую очередь самим изготовителем на всех этапах жизненного цикла, начиная с проектирования и разработки, а также непрерывно в процессе производства. Для того чтобы выпускать продукцию высокого технического уровня и качества, необходимо эффективно управлять процессами формирования этих комплексных и обобщающих характеристик изделий.

При осуществлении контроля качества производится обязательный сбор данных, а затем их обработка. Но данные, касающиеся даже одного и того же параметра изделия, не могут быть многократно получены при идентичных условиях, так как в ходе процесса меняются отдельные процессы и обстоятельства. Поэтому при операциях, относящихся к контролю качества, приходится иметь дело с большим числом данных, характеризующих те или иные параметры изделия, условия процесса и т. д. Эти данные при повторных измерениях всегда оказываются несколько отличающимися от полученных в другое время и при других условиях, то есть всегда наблюдается разброс данных. Анализируя разброс данных, можно найти решение возникшей в процессе производства проблемы.

6.2. Цели управления качеством с помощью статистических методов

При повторяющихся рабочих процессах, прежде всего при серийном и массовом производстве, определённые факторы снижения качества становятся типичными. Использование математико-статистических методов даёт возможность исследовать протекание технологического процесса.

В таком случае говорят, что процесс изготовления является статистически управляемым. Статистические методы позволяют обнаружить: где, когда, кем и при каких обстоятельствах вызваны те или иные помехи в производственном процессе. Это повышает чувство ответственности всех участников производственного процесса, способствует тесному сотрудничеству и рождает новое отношение к понятию «качество».

Сами по себе эти методы часто не указывают непосредственно на причину брака, но показывают, где её искать, или, как говорят, «освещают тёмные углы» производственных процессов [6]

6.3. Диаграмма причин и результатов

Когда решается задача анализа возможных причин, ответственных за тот или иной дефект или проблему, целесообразно эти причины определённым образом упорядочить, провести их классификацию, выявить максимально возможное их количество без риска упустить какую-нибудь из них. При этом очень важно обеспечить наглядность, т. е. ситуацию, при которой все причины и их отношение к результату постоянно находились бы в поле зрения.

Объектами исследования с помощью причинно-следственных диаграмм могут быть: появление дефектности изделий, увеличение расходов на устранение брака, падение спроса на продукцию на рынке, управление персоналом и т. д.

Диаграмму причин и результатов впервые внедрил в производственную практику профессор Токийского университета Каору Исикава (1953 г.).

Диаграмма причин и результатов — это диаграмма, которая показывает отношение между показателями качества и воздействующими на него факторами

Безусловно, это один из наиболее элегантных и широко используемых методов среди так называемых семи простых инструментов контроля качества. Иначе диаграмму Исикавы называют причинно-следственной диаграммой или «рыбий скелет» (рис. 6.1).

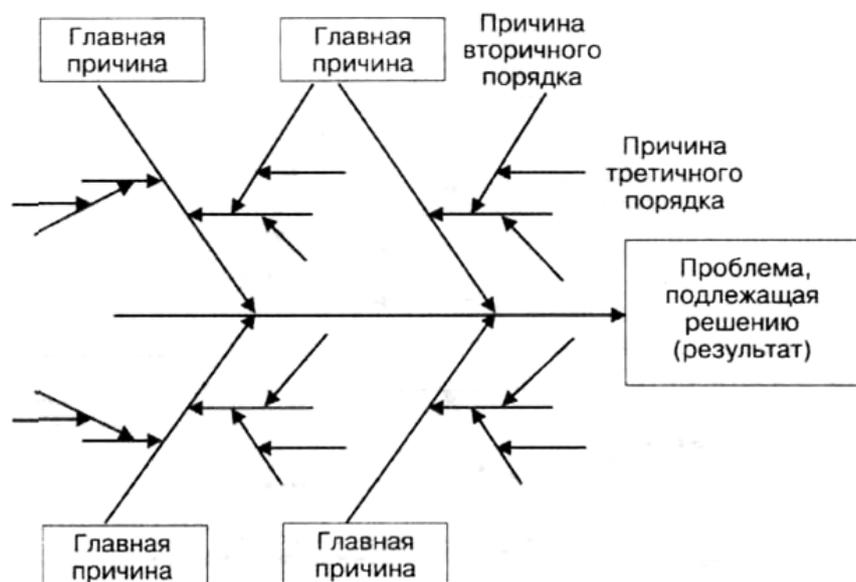


Рис. 6.1. Диаграмма причин и результатов

Для построения причинно-следственной диаграммы данные заносятся в таблицу рабочего окна, как показано на рис. 6.2. Затем в основном рабочем окне системы в выпадающем меню выберите команду *Statistics/ Industrial Statistica & Six Sigma/ Process Analysis* (рис. 6.3).

Варка супа					
	1	2	3	4	5
	Продукты	Технология варки	Условия варки	Повар	Оборудов. кухни
1	Вода	Закрытая крышка кастрюли	Время	Опыт работы	Исправная плита
2	Соль	Сначала варить бульон	Температура плиты	Квалификация	Большая кастрюля
3	Мясо	Снимать пену при варке бульона		Знание рецепта	Острый нож
4	Картофель	Нарезать продукты ножом			Доска для нарезки
5	Морковь	Не допускать разваривания			
6	Лук				
7	Лапша				
8	Приправы				

Рис. 6.2. Пример данных для причинно-следственной диаграммы

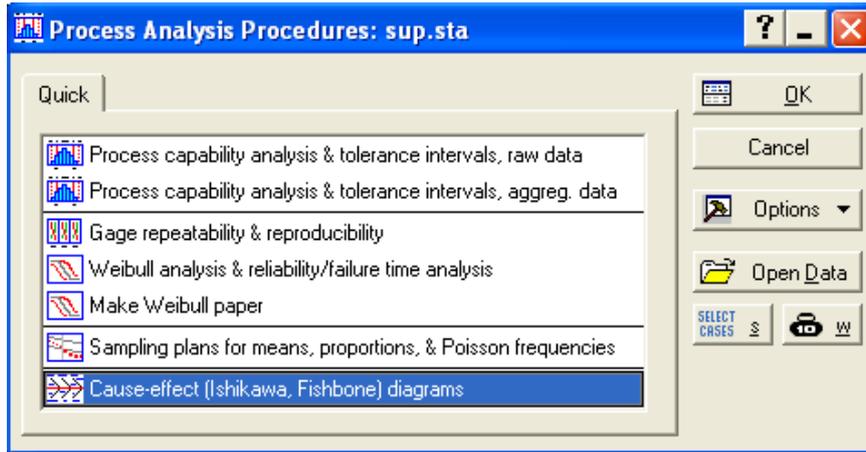


Рис. 6.3. Диалоговое окно выбора диаграммы Исикавы

В появившемся окне выберите команду *Cause-effect (Ishikawa, Fishbone) diagrams* и нажмите *OK*. Появится окно, показанное на рис. 6.4, в котором с помощью кнопки *Variables* необходимо отметить, какие факторы будут находиться сверху «хребта рыбы», а какие внизу. С помощью вкладок *Arrows* и *Font sizes* можно выбрать размер шрифтов для надписей, толщину и угол наклона линий «костей». Пример диаграммы показан на рис. 6.5. Все линии и надписи на диаграмме можно изменить и передвинуть. Дорабатывать диаграмму можно с помощью панели рисования, что и сделано на рис. 6.5.

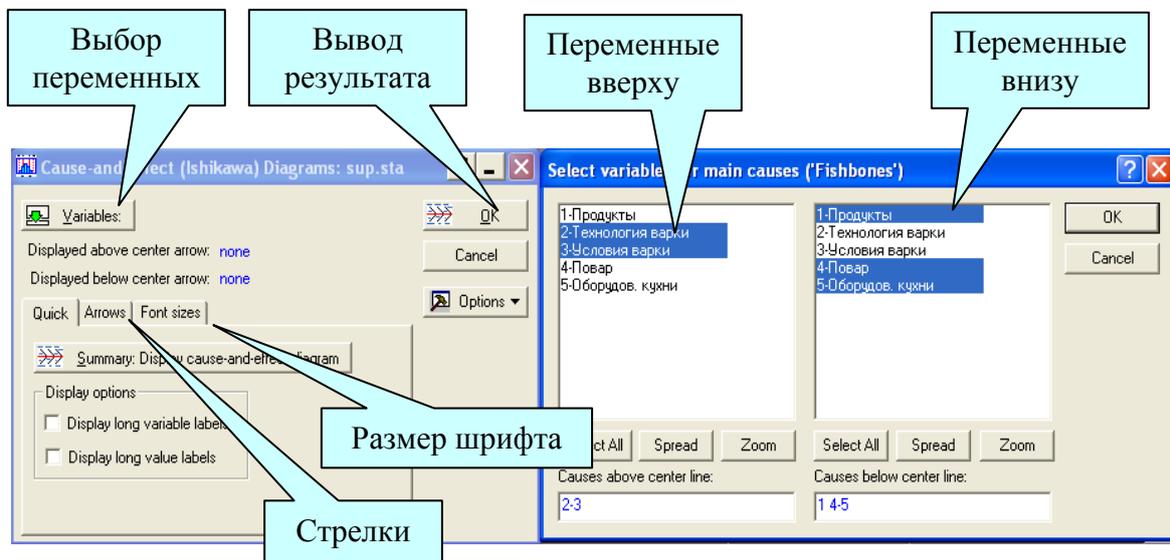


Рис.6.4. Окна выбора переменных для причинно-следственной диаграммы

Cause-And-Effect Diagram

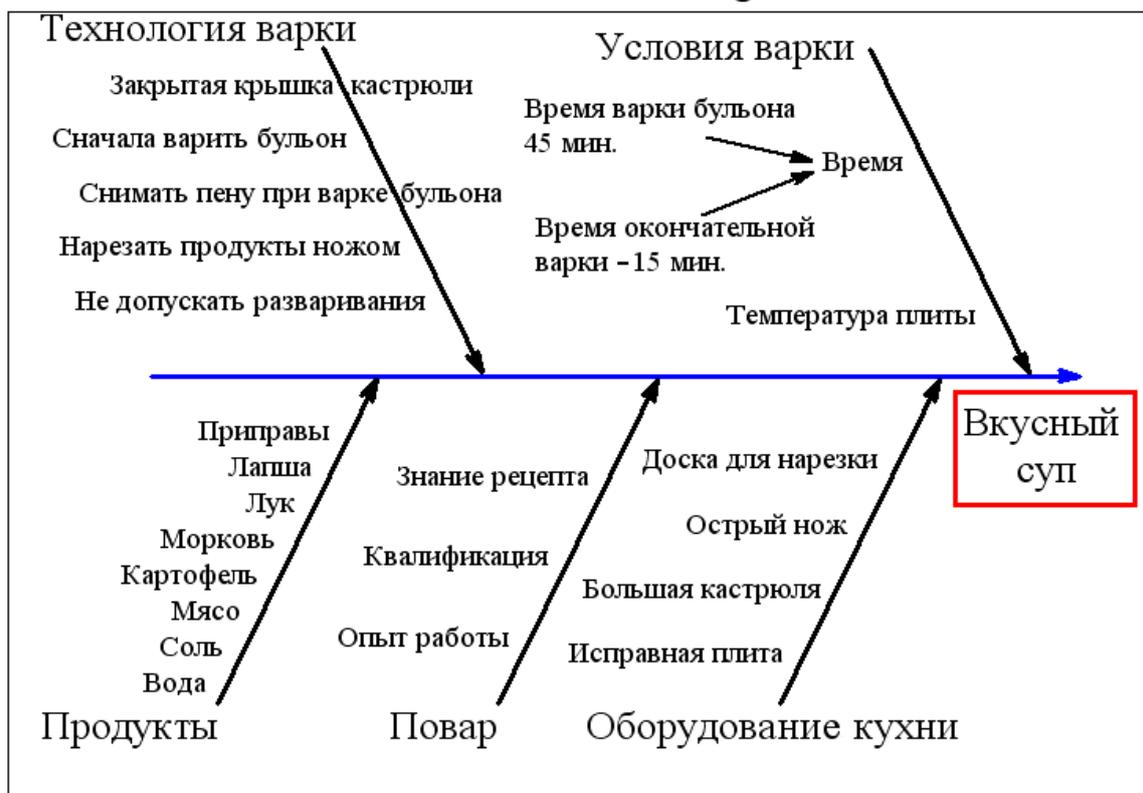


Рис. 6.5. Причинно-следственная диаграмма

Нанесите на диаграмму всю необходимую информацию: её название; наименование изделия, процесса или группы процессов; имена участников процесса; дату и т. д. Это можно сделать с помощью панели рисования, доступной в программном окне.

Построенную диаграмму Исикавы необходимо постоянно совершенствовать. Это позволяет получить действительно ценный документ, который поможет в решении и других проблем, которые могут возникнуть в дальнейшем не только в связи с рассматриваемым показателем качества, но и при возникновении других дефектов или несоответствий (рис. 6.6).

Дальнейшая работа будет состоять в том, чтобы на основе наблюдений за реальным процессом, в результате которого потеря качества, установить действительную связь между исследуемым показателем качества и выбранными факторами (причинами), которые оказывают на него наибольшее негативное воздействие.



Рис 6.6. Причинно-следственная диаграмма для некачественного фотокопирования

6.4. Закон 80/20

Смысл закона, восходящего к работам социолога Вильфредо Парето, состоит в том, что за 80 % результата отвечает 20 % причин.

Поскольку подавляющую долю эффекта определяет лишь небольшая доля элементов, дающих максимальный вклад, их влияние оказывается непропорционально велико, поэтому этот закон также называют принципом дисбаланса.

Под «результатом» процесса может пониматься, например, суммарный объем продаж многономенклатурного товара, благосостояние населения страны, объем товара на складе, количество жителей городов и т. п. Важным является то, чтобы число составляющих (количество ас-

сортиментных позиций, население страны, количество городов и т. д.), было бы велико.

Популярность закона Парето определяется с одной стороны его чрезвычайной простотой и наглядностью, а с другой стороны – возможностью применения в анализе очень широкого круга процессов. Например:

- 80 % пыли подметается с 20 % пола, по которому чаще всего ходят;
- 80 % стирки уходит на 20 % одежды, которую чаще всего носят;
- 80 % покупок делают 20 % покупателей;
- 80 % телефонных звонков делают 20 % абонентов;
- 80 % продукции выпускают 20 % предприятий;
- 80 % работы делают 20 % людей;
- 80 % людей считают, что они входят в эти 20 %;
- 80 % пользования файлами осуществляется в пределах 20 % файлов;
- 80 % времени, отдаваемого чтению, тратится на 20 % газетных страниц;
- 80 % прибыли дают только 20 % клиентов;
- 80 % потерь на производстве дают только 20 % видов дефектов, а оставшиеся 80 % видов дефектов обуславливают остальные 20 % потерь.

Конечно, соотношение 80/20 не является абсолютным и универсальным, хотя, как правило, отклонения от этого соотношения не очень велики.

6.5. Анализ Парето

В большинстве случаев подавляющее число дефектов и связанных с ними материальных потерь возникает из-за относительно небольшого числа причин. Таким образом, выяснив причины появления основных дефектов, можно устранить почти все потери, сосредоточив усилия на ликвидации именно этих причин.

Диаграмма Парето – это инструмент, позволяющий распределить усилия для разрешения возникающих проблем и выявить основные причины, которые нужно проанализировать в первую очередь

С помощью анализа Парето можно выявить, какой из видов дефектов приносит наибольшие потери во времени или в материалах, какие

дефекты встречаются наиболее часто. Можно анализировать экономические проблемы предприятия, социальные процессы в больших коллективах, психологические проблемы в группах и много других проблем, возникающих в производственной, экономической, социальной и других сферах деятельности [6]. Диаграммы Парето применять целесообразно только в том случае, когда анализируется большое число видов дефектов или причин их появления и когда выявление группы существенных причин затруднено.

Диаграмма Парето по результатам деятельности предназначена для выявления главной проблемы. Она отражает нежелательные результаты деятельности: дефекты, поломки, отказы, ремонты, возвраты продукции, объём потерь, затраты, нехватку запасов, ошибки в составлении счетов, срыв сроков поставок и прочее.

Диаграмма Парето по причинам отражает причины проблем, возникающих в ходе производства. Она используется для выявления главной из них: исполнитель работы, оборудование, сырьё, метод работы, измерения.

Построение диаграммы Парето начинают с классификации возникающих проблем по отдельным факторам (например, проблемы, относящиеся к браку, к работе оборудования или исполнителей и т. д.). Затем производят сбор и анализ по каждому фактору, чтобы выяснить, какие из этих факторов являются преобладающими при решении проблем.

В качестве примера рассмотрим данные по ремонту оборудования (табл. 1.1) и построим диаграмму Парето для дефектов и вызванных ими потерь (4 и 5 столбцы таблицы). Выберем модуль *Statistics/ Industrial Statistic & Six Sigma/ Quality Control Charts/ Pareto chart analysis* (рис. 6.7). В появившемся диалоговом окне, приведённом на рис. 6.8, необходимо выбрать формат для ввода данных и нажать *OK*. Если диаграмма строится только по причинам, используются настройки по умолчанию – *Codes (requires tabulation of data codes)*. Если диаграмма строится по причинам и стоимости, выбираем опцию *Codes and counts (one variable with defect type, one variable with counts)*, как показано на рис. 6.8.

В появившемся диалоговом окне осталось выбрать переменные так, как показано на рис. 6.9 и нажать *OK*. В результате будет построена диаграмма Парето, приведённая на рис. 6.10.

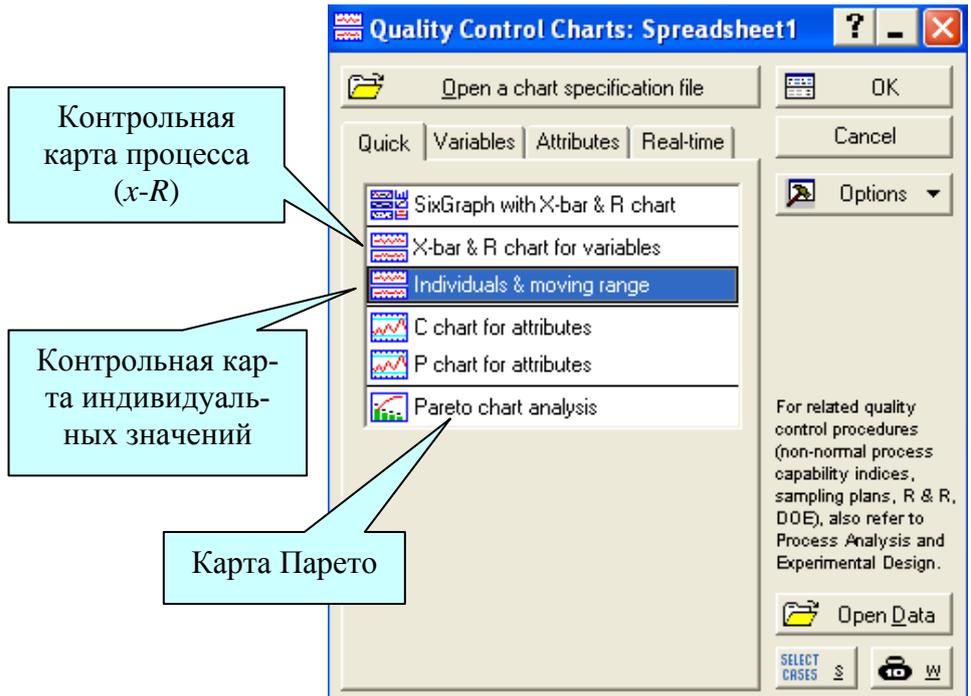


Рис 6.7. Окно выбора типа контрольной карты

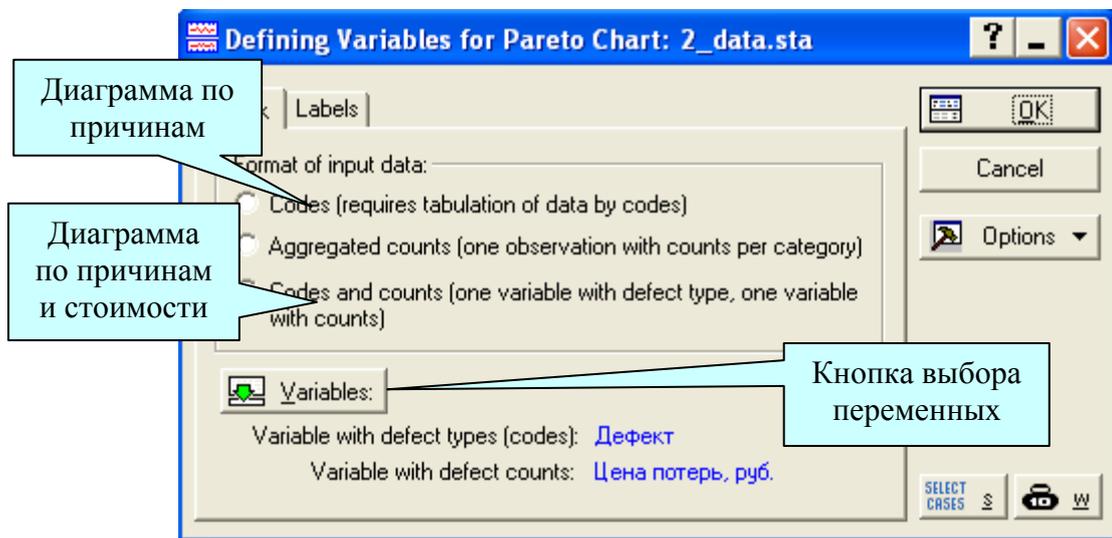


Рис. 6.8. Выбор формата данных для диаграммы Парето

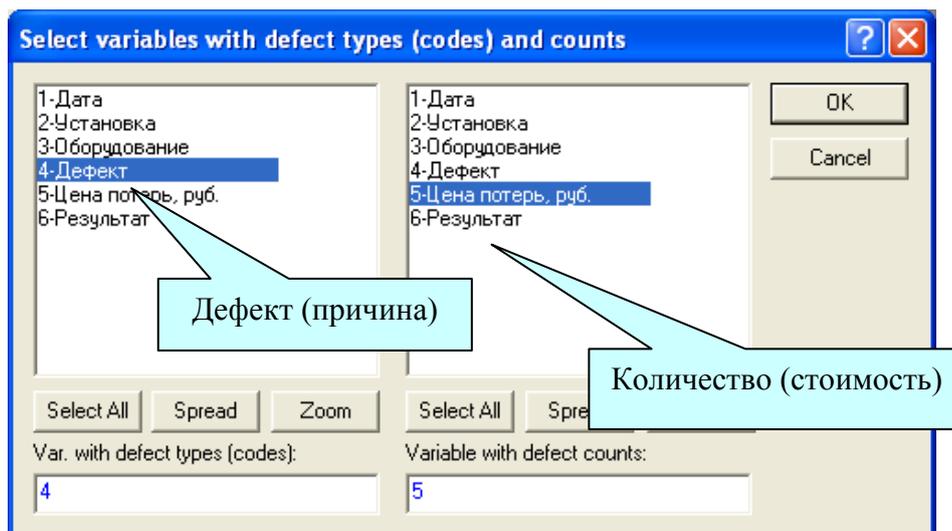


Рис 6.9. Окно выбора переменных для диаграммы Парето

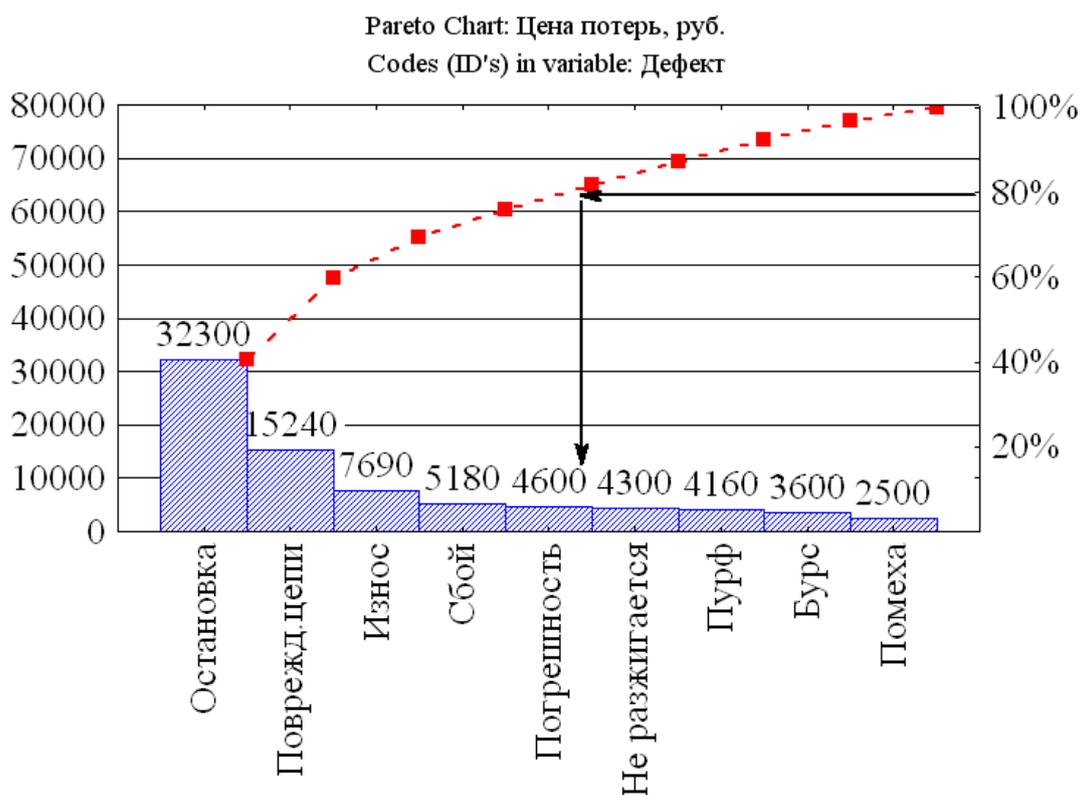


Рис. 6.10. Диаграмма Парето

Диаграмму можно отредактировать с помощью панели рисования и с помощью настроек панели «свойства графиков». На рис. 6.10 таким образом отмечено 80 % дефектов – остановка, повреждение цепи, износ,

сбой, погрешность. Остальные дефекты дают только 20 % потерь. Диаграмму можно вывести в виде таблицы, если вернуться к окну построения диаграммы и нажать кнопку *Display chart summary*. Накопленный процент отображается в последней строке таблицы.

С помощью диаграммы по результатам выявляются существенные дефекты. Затем из них выбирается дефект, который встречается наиболее часто, после чего *выдвигаются предположения о том, какие причины могут быть ответственны за этот дефект*. Здесь можно использовать в качестве метода анализа диаграмму Исикавы. Далее на основе дополнительных наблюдений строится диаграмма Парето по причинам и из них выявляются существенные, которые и устраняются в первую очередь. Подобным образом последовательно устраняются все существенные дефекты, выявленные с помощью диаграммы по результатам.

После устранения существенных дефектов снова строится диаграмма Парето по результатам и выявляются существенные дефекты среди оставшихся. Эти дефекты снова анализируются с помощью диаграммы по причинам и затем устраняются.

6.6. Карты контроля качества

Изготовление продукции всегда связано с непостоянством условий производства [6]. Это приводит к изменениям качества изготавливаемых изделий. При хорошо спланированном и правильно осуществляемом процессе эти изменения незначительны. В таком случае говорят, что процесс является *статистически подконтрольным*. Как правило, производственные процессы протекают в статистически регулируемом состоянии, однако случаются ситуации, когда под воздействием случайных причин процесс выходит из состояния статистического контроля. В таких случаях необходимо как можно быстрее обнаружить причину этих вариаций, что без применения специальных методов сделать порой весьма трудно.

Для решения этой задачи используется механизм, разработанный в 1924 году американским инженером Вальтером Шухартом, базирующийся на использовании контрольных карт, часто называемых картами Шухарта. Карты контроля качества, или контрольные карты служат для постоянного контроля за тем, чтобы производственный процесс оставался статистически подконтрольным. Основная цель применения контрольных карт – быстрое обнаружение характера изменений в производственных процессах *по результатам наблюдения за параметрами*

продукции с целью поиска их причин и корректировки процесса ещё до того, как начнёт появляться бракованная продукция.

Все описанные ранее статистические методы дают возможность зафиксировать состояние процесса в определённый момент времени. В отличие от них метод контрольных карт позволяет отслеживать состояние процесса во времени и более того – воздействовать на процесс до того, как он выйдет из-под контроля.

Контрольные карты – это линейные графики для оценки управляемости процесса по результатам сравнения отдельных измерений с заданными контрольными границами (рис. 6.11).

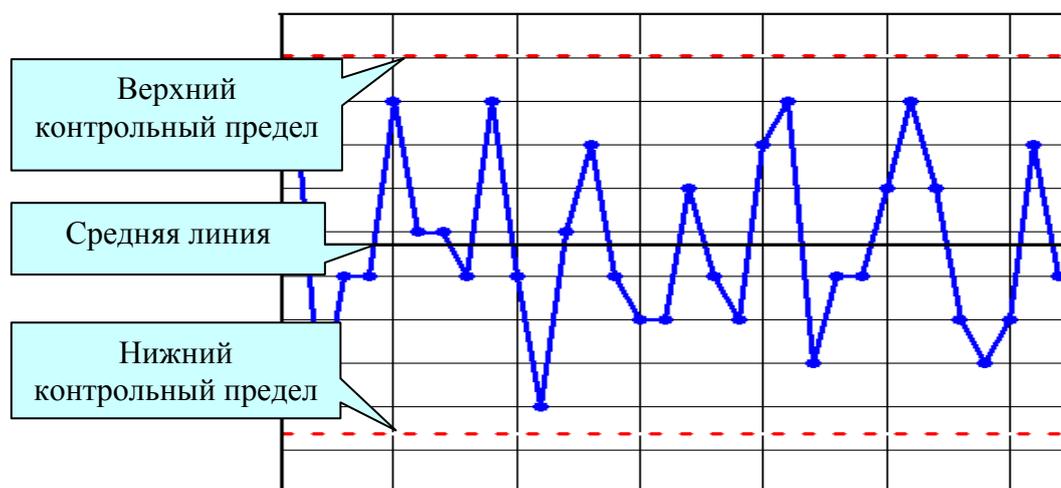


Рис 6.11. Пример контрольной карты

Всякая контрольная карта состоит обычно из трёх линий. Центральная (средняя) линия представляет собой требуемое среднее значение характеристики контролируемого параметра качества. Две другие линии, одна из которых находится над центральной – верхний контрольный предел (UCL – Upper Control Level), а другая под ней – нижний контрольный предел (LCL – Lower Control Level), представляют собой максимально допустимые пределы изменения значений контролируемой характеристики (показателя качества), чтобы считать процесс удовлетворяющим предъявляемым к нему требованиям.

Контрольные карты применяются как для анализа количественных данных, когда результаты измерений показателя качества непрерывны и выражаются в числовой форме, так и в случае, когда информация об объектах дискретна и ограничена выводом типа «годен»–«не годен». В первом случае применяются контрольные карты по количественному

признаку, во втором – по альтернативному. Подробнее о видах карт можно прочитать в работе [6].

6.7. Контрольная карта индивидуальных значений

Эта карта применяется, когда наблюдение производится над сравнительно небольшим числом объектов, и все они подвергаются контролю. Чаще всего это бывает при наладке и настройке процесса, когда преследуется цель его предварительного исследования. Карта удобна тогда, когда процесс протекает в реальном времени и есть возможность оперативного вмешательства в него в случае выхода параметра качества за допустимые пределы.

Порядок построения карты следующий.

1. Данные измерения исследуемой величины x_i регистрируются *последовательно с равным шагом*. Предположим, например, что необходимо контролировать концентрацию некоторого вещества в химическом процессе. Вы наблюдаете процесс в реальном времени в течении 32 часов и снимаете с датчиков нужную характеристику каждый час (табл. 6.1, первый столбец).

Таблица 6.1

Наблюдаемые значения концентрации вещества

Наблюдаемое значение (x_i)	Номер наблюдения в выборке
102	1
95	2
98	3
98	4
102	1
99	2
99	3
98	4
102	1
98	2
95	3
99	4
101	1
98	2

97	3
97	4
100	1
98	2
97	3
101	4
102	1
96	2
98	3
98	4
100	1
102	2
100	3
97	4
96	1
97	2
101	3
98	4

2. Запустить модуль *Statistics/ Industrial Statistics & Six Sigma/ Quality Control Charts*. На стартовой панели (рис. 6.7) выбрать *Individuals & moving range (отдельные наблюдения и скользящий размах)* и нажать кнопку *OK*.

3. В появившемся диалоговом окне выбрать переменную с измерениями – *Measurements (observations)* (рис. 6.12) и нажать *OK*. В результате будет построена контрольная карта, приведённая на рис. 6.13. Имеется возможность группировки данных, если наблюдений слишком много. При этом у каждой выборки будет вычислено среднее значение, которое наносится на карту. Для группировки необходимо указать переменную *Part identifiers (code numbers)*, где должны быть номера выборок. Если объём каждой выборки постоянный, это можно указать прямо в окне на рис. 6.12, отметив чекбокс *Constant number of samples per part* и введя нужный объём выборки.

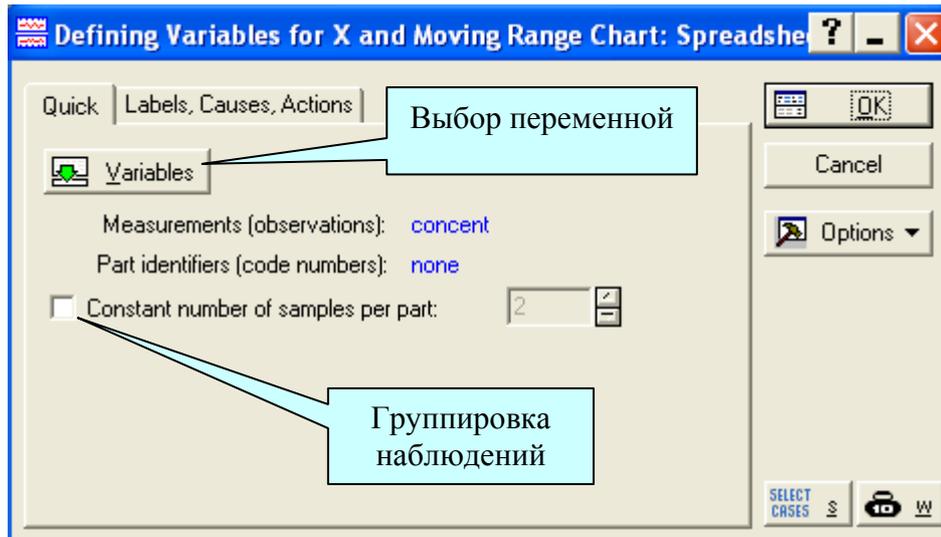


Рис 6.12. Окно выбора переменной для контрольной карты

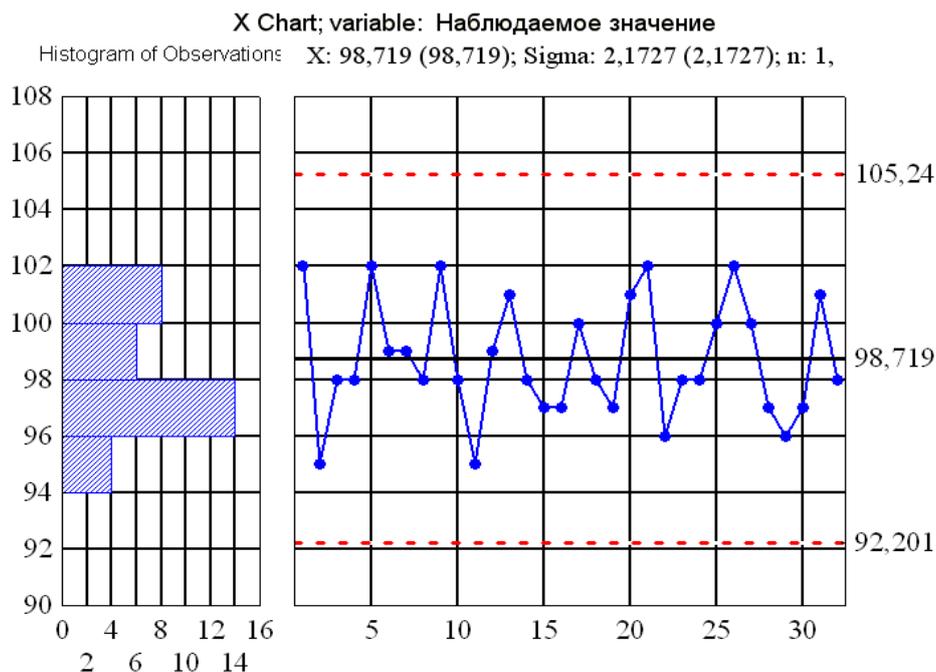


Рис 6.13. Контрольная карта индивидуальных значений

В зависимости от расположения точек относительно границ регулирования и средней линии можно сделать заключение о состоянии процесса. В данном случае все точки лежат внутри границ регулирования и их разброс относительно средней линии можно считать относительно равномерным. Исследуемый процесс находится в состоянии ста-

тистического регулирования. Если на контрольную карту нанести поле допуска и сравнить его расположение с границами регулирования, то можно сделать заключение относительно настройки и наладки процесса.

Недостатком \bar{x} -карты является то, что она не даёт наглядного представления как о динамике изменения наладки процесса, так и о динамике уровня его настройки, т. е. не позволяет судить об изменении во времени величины поля рассеяния и положения его центра. Поэтому применение этой карты ограничивается, как правило, предварительной оценкой процесса.

6.8. Контрольная карта средних значений и размахов

Эта карта применяется при массовом производстве. Достоинство её состоит, во-первых, в том, что она позволяет отслеживать во времени как настройку процесса, так и его наладку, а во-вторых, выводы относительно характеристик делаются на основе малых выборок из большого числа рассматриваемых единиц продукции, что существенно удешевляет контроль текущих характеристик процесса [6].

Порядок построения карты в системе Statistica следующий.

1. Все единицы продукции последовательно во времени подразделяются на группы, из которых в последствии будут браться малые мгновенные выборки. Группы могут быть представлены как выработка за час, смену или за другой промежуток времени.

2. Из каждой группы берётся случайная выборка объёмом не более 10-ти единиц продукции. Чаще всего объём выборки составляет 4–5 единиц. Таких последовательных во времени выборок следует взять не менее 20–25 штук.

В примере (табл. 6.1) результаты измерений записаны в первом столбце, номера групп – во втором.

3. Запустить модуль *Statistics/ Industrial Statistics & Six Sigma/ Quality Control Charts*. На стартовой панели (рис. 6.7) выбрать *X-bar & R chart for variables* и нажать кнопку *OK*.

4. В появившемся диалоговом окне выбрать переменную с измерениями – *Measurements* и переменную – номера выборок *Sample Idents (opt.)* и нажать *OK*. В результате будет построена контрольная карта, приведённая на рис. 6.14. Здесь также если объём каждой выборки постоянный, это можно указать прямо в окне выбора переменных, отметив

чекбокс *Constant number of samples per part* и введя нужный объём выборки.

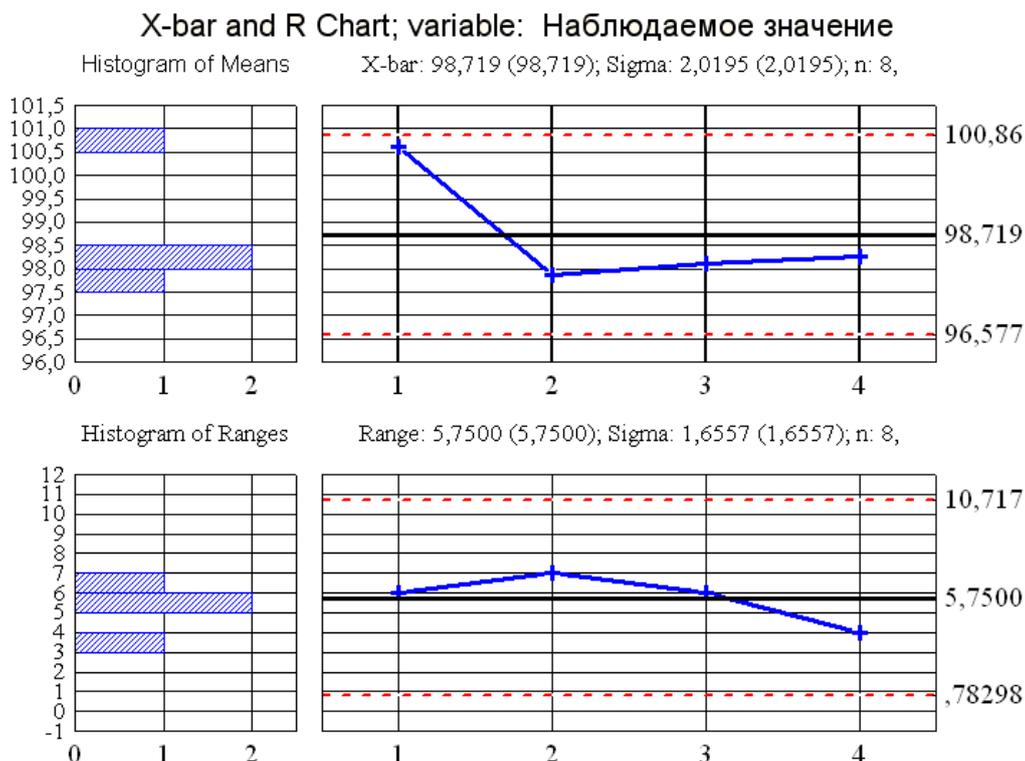


Рис 6.14. Контрольная карта средних значений и скользящих размахов

На x-карте скользящих средних все точки попадают внутрь контрольных границ. На контрольной карте скользящих размахов все точки также находятся внутри контрольных границ. Размахи служат оценкой изменчивости характеристик, поэтому можно сказать, что концентрация вещества подчиняется требованиям статистического контроля по уровню средних и изменчивости.

6.9. Чтение контрольных карт

Достоинство контрольных карт в управлении процессом состоит в том, что они дают точное представление о состоянии объекта управления (процесса) с помощью анализа карты. Это позволяет быстрее принимать необходимые корректирующие меры, если процессу угрожает выход из-под контроля и возможность появления брака.

Выход процесса из-под контроля оценивается по следующим критериям.

Точки выходят за контрольные пределы (UCL, LCL) или лежат на них

Процесс вышел из состояния статистического контроля, т. е. стал нестабильным, и характеристики его изменились (рис. 6.15). Если при этом выхода за границы допуска нет, то вмешательство в процесс не требуется.

Если на процесс действуют только обычные причины, то такой процесс называется стабильным. Настройка и разброс стабильного процесса со временем не меняются.

Для стабильного процесса вероятность выхода за контрольные границы среднего и размаха в группе очень мала (меньше 0,01). И если точка всё-таки вышла за контрольные границы, то, скорее всего, это является следствием воздействия особой причины.

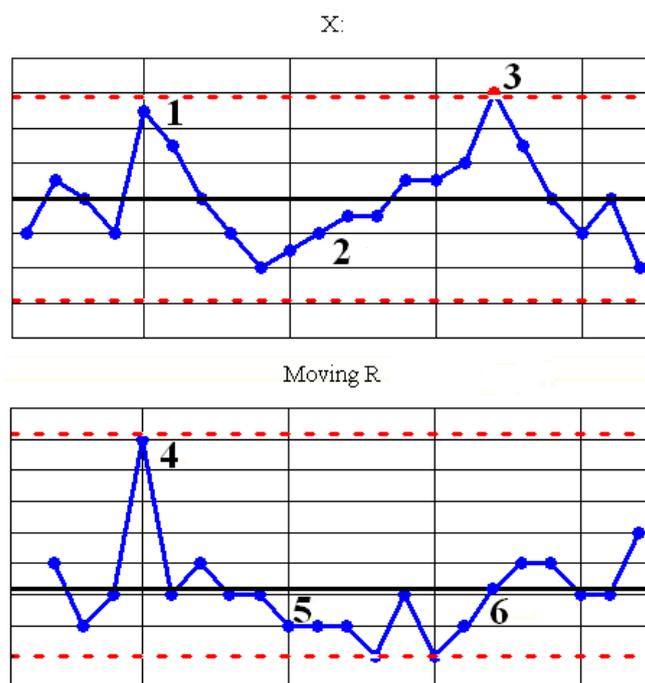


Рис. 6.15. Выход точки за контрольный предел, дрейф и серия точек

При построении x - R карты могут возникнуть следующие ситуации [14].

1. За границами регулирования находятся точки на R -карте и соответствующие им точки на x -карте (рис. 6.15, точки 1, 4). Это означает,

что за счёт обычных (внутренних) причин увеличилось технологическое рассеяние, т. е. возросла величина σ . В этом случае следует заняться поиском и устранением причин разладки процесса.

2. За границами регулирования находятся точки на x -карте, но при этом соответствующие им точки на R -карте лежат в границах регулирования (рис. 6.15, точки 3, 7). Поскольку по R -карте выхода за границы регулирования нет, полное технологическое рассеяние остаётся прежним, т. е. наладка процесса не изменяется. Значит, есть все основания предполагать, что выход за границы регулирования по x -карте произошёл потому, что распределение по x сместилось в сторону больших или меньших значений контролируемого признака. Это, как правило, является результатом воздействия на процесс какой-то особой внешней причины, изменяющей его настройку. Дальнейшие действия должны быть связаны с поиском и устранением этой причины.

3. За границами регулирования наблюдаются точки на R -карте, а также соответствующие им и не соответствующие точки на x -карте. Это говорит о наличии как обычных, так и особых причин, ухудшающих процесс.

Часто встречается ситуация, когда влияние первой обнаруженной особой причины настолько велико, что из-за неё не видно влияния других причин. В этом случае соответствующую точку исключают из набора данных и строят карту заново. Влияние других причин становится видимым. Таким образом, последовательно, шаг за шагом обнаруживая особое поведение точек на контрольной карте и устанавливая их причины, делают процесс прозрачным, доступным нашему пониманию [15].

Наблюдается серия точек

Серия – это такое состояние процесса, при котором последовательные точки лежат по одну сторону от средней линии (рис. 6.16, точки 5). Число таких точек называется длиной серии. Процесс нестабилен, если:

- серия состоит из 7 точек и более;
- 10 точек из 11 лежат по одну сторону от средней линии;
- не менее 12 точек из 14 лежат по одну сторону от средней линии;
- не менее 16 точек из 20 лежат по одну сторону от средней линии.

Причиной серии является внешнее воздействие на процесс, которое сдвигает центр рассеяния в ту или иную сторону от средней линии, изменяя настройку процесса.

Наблюдается дрейф

Дрейф – это не менее 7 поднимающихся или ниспадающих точек (рис. 6.15, точки 2, 5). Причинами появления дрейфа могут быть, например, такие факторы, как постепенный рост (падение) температуры окружающей среды, износ технологического оборудования, появление в средствах измерения прогрессирующих погрешностей, изменение физических и химических параметров процесса и другие неслучайные причины.

Две и более близлежащих точки приближаются к границам регулирования (лежат за пределами 2-сигмовых границ)

Точки считаются приблизившимися к границам регулирования, если они находятся за пределами плюс-минус 2σ относительно средней линии, т. е. на расстоянии большем, чем $2/3$ расстояния от средней линии до границы регулирования в так называемой *зоне внимания* (рис. 6.16).

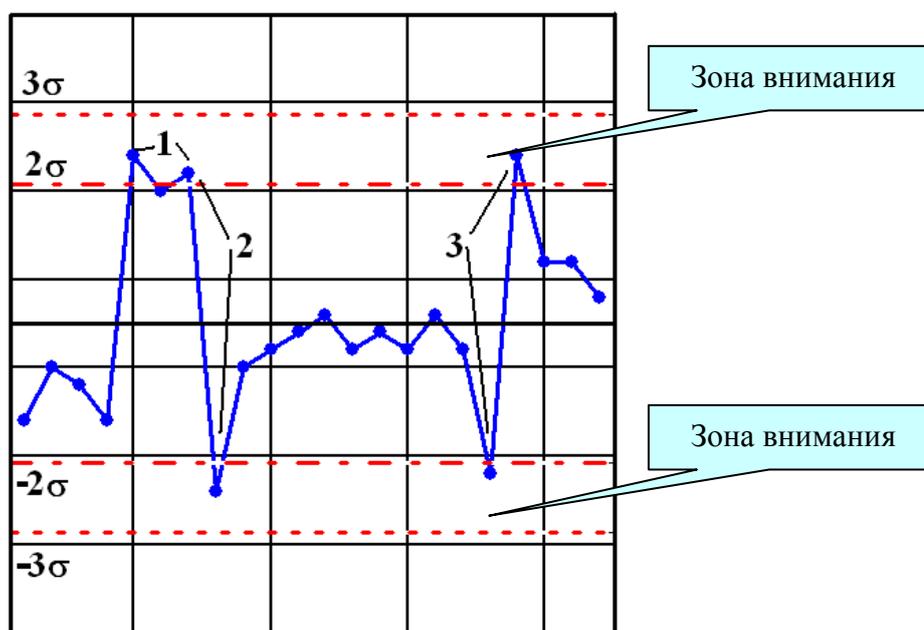


Рис. 6.16. Приближение к границам регулирования

Если выход точки за границы регулирования незначителен и в дальнейшем больше не повторяется, то вполне возможно, что этот факт не говорит о дестабилизации процесса. Но если выход за границы ка-

кой-либо одной точки составляет заметную величину, то вмешательство в процесс с целью его совершенствования в любом случае необходимо.

Наблюдается периодичность

Наличие подъёмов и спадов с примерно одинаковыми интервалами (рис. 6.17) также говорит о нестабильности процесса, причиной которой может быть воздействие на процесс внешнего периодически изменяющегося фактора.

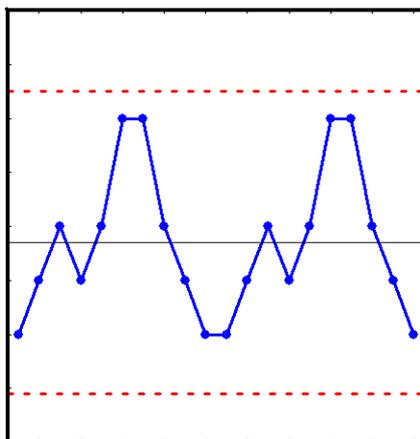


Рис. 6.17. Периодичность

Точки приближаются к средней линии

Точки считаются приблизившимися к средней линии, если они лежат внутри полуторасигмовой зоны, то есть внутри линий, делящих пополам расстояние от средней линии до границ регулирования. В этом случае следует изменить способ разбиения на выборки или группы, поскольку может оказаться, что смешаны данные из разных распределений.

Отмеченные в каждом рассмотренном случае выходы процессов из состояния статистического регулирования несут в себе потенциальную угрозу получения брака в недалеком будущем, ибо однажды возникшие нестабильности в процессе всегда имеют тенденцию со временем нарастать.

Таким образом, контрольные карты и их грамотный анализ позволяют прогнозировать характер протекания производственных процессов в будущем и вовремя их останавливать для корректировки с целью предупреждения возможного появления бракованной продукции.

6.10. Контрольные карты накопленных сумм

Карты Шухарта нечувствительны к малым возмущениям процессов. При достаточно долговременном контроле информация о начальном этапе процесса теряется. В отличие от рассмотренных, контрольные карты накопленных сумм – это карты с памятью. Они могут быть более чувствительными к возмущениям, т. е. уже в самом начале сдвига уровня настройки процесса или изменения технологического рассеяния они указывают на необходимость вмешательства в процесс.

Таким образом, контрольные карты накопленных сумм следует применять в тех случаях, когда даже незначительные смещения уровня настройки процесса недопустимы и подлежат скорейшему устранению.

Для построения контрольной карты накопленных сумм на стартовой панели (рис. 6.7) необходимо выбрать вкладку *Variables*, а в ней – *CuSum chart for individuals* и нажать *OK*. Дальнейшие действия аналогичны рассмотренному алгоритму для построения карт индивидуальных и средних значений.

6.11. Задания для самостоятельной работы

Задание 1. В системе Statistica построить диаграмму причин и результатов с помощью модуля *Statistics / Industrial Statistics & Six Sigma/ Process Analysis/ Cause–Effect [Ishikawa, Fishbone] diagrams*.

Задание 2. С помощью модуля *Statistics/ Industrial Statistics & Six Sigma/ Quality Control Charts/ Pareto chart analysis* построить диаграмму Парето. Причины и их число выбрать самостоятельно, можно любые, не относящиеся к какому-либо процессу. Диаграммы построить:

- только для переменной 1 «причина», без учета значимости причины для общего вклада в качество анализируемого процесса. Для этого на вкладке *Quick* выбрать *Codes (requires tabulation of data by codes)* – установлен по умолчанию, а в окне выбора переменных указать одну переменную с перечнем причин;

- для переменных 1 «причина» и 2 «число», с учётом значимости причины для общего вклада в качество анализируемого процесса. Для этого на вкладке *Quick* выбрать *Codes and counts (one variable with defect type, one variable with counts)*.

В чём состоит разница между этими картами Парето?

Задание 3. С целью выяснения причин брака составлен контрольный листок в предположении, что причинами могут быть рабочий, станок или смена (табл. 6.2). Определить, кто виноват, если это возможно.

Таблица 6.2

Распределение дефектов по рабочим, станкам и сменам

Рабочий	Станок	1 смена	2 смена	3 смена	Число дефектов на станках	Сумма дефектов рабочего
А	А 1			•	1	24
	А 2	• •	•		3	
	А 3	•	••••• • ••	••••• •• •••••	20	
Б	Б 1	• •	••••• • ••	••••• •	15	45
	Б 2	•	•	•••	5	
	Б 3	• •	••••• • •••	••••• ••• ••••• ••	25	
В	В 1	•	•	• •	4	52
	В 2	• •	• •	• •• •	8	
	В 3	•••	••••• •• ••••• ••••• •••	••••• • ••••• ••• •••	40	

Задание 4. Прочитать файл с данными <http://ieeep.tpu.ru/statlab/variant3.sta>. С помощью модуля *Statistics / Industrial Statistics & Six Sigma/ Quality Control Charts/ Individuals & moving range* построить контрольную X-R карту индивидуальных значений для одной переменной своего

варианта (VarN) и провести анализ качества производственного процесса. Проверить машинное построение карты расчётом средней линии.

Задание 5. Прочитать файл с данными <http://ieeep.tpu.ru/statlab/variant3.sta>. С помощью модуля *Statistics / Industrial Statistics & Six Sigma/ Quality Control Charts/ X bar & R chart for variables* построить контрольную X-R карту производственного процесса для одной переменной своего варианта с группировкой по времени (VarN, Time) и провести анализ качества процесса. Сравнить построение карты с результатами предыдущего задания. Чем отличается построение карты индивидуальных значений от карты процесса?

Задание 6. С помощью модуля *Statistics / Industrial Statistics & Six Sigma/ Quality Control Charts/ X-bar & R Chart for variables* построить контрольную X-R карту процесса для одной переменной своего варианта (VarN) и временных интервалов (Time, Day) и провести анализ качества производственного процесса.

Задание 7. Построить контрольную C-карту индивидуальных значений для одной переменной своего варианта (VarN) и провести анализ качества процесса.

Задание 8. Исследовались 8 прядильных машин на предмет числа обрывов нити пряжи. За десять обследований продолжительностью по 15 минут было обнаружено следующее число обрывов нити (см. файл <http://ieeep.tpu.ru/statlab/indstat1.sta>). Определить среднее число обрывов нити и стандартное отклонение для каждой машины и выборки. Построить гистограммы и сделать вывод о качестве работы прядильных машин.

Задание 9. Для исследования коррозии серии образцов изделий, изготовленные в различных условиях, были подвергнуты климатическим воздействиям. Результаты измерений 10 серий приведены в файле <http://ieeep.tpu.ru/statlab/indstat2.sta>. Определить среднее значение и стандартное отклонение для каждой серии. Построить гистограммы и сделать вывод о качестве изделий.

Задание 10. На заводе осуществляется отливка стальных деталей из в объеме 5 плавков за смену по 4 тонны каждая. В плавках содержание

кремния контролируется экспресс-анализом. Содержание кремния не должно превышать 1 %. Результаты экспресс-анализа представлены в файле <http://ieeep.tpu.ru/statlab/indstat3.sta>. Определить среднее значение и стандартное отклонение содержания кремния для смен и партии. Построить гистограммы и сделать вывод о качестве процесса.

Задание 11. В лаборатории измерялось разрывное усилие образцов проволоки одной марки на 21 машине. Данные представлены в файле <http://ieeep.tpu.ru/statlab/indstat4.sta>. Определить среднее значение и стандартное отклонение разрывного усилия для каждой машины и всех машин. Построить гистограммы и сделать вывод о качестве работы машин.

Задание 12. На производстве осуществляется контроль диаметра подшипников, получаемых от поставщика для конвейера по сборке двигателей. Для контроля качества случайным образом отбирается 50 изделий, которые группируются по 5 штук в группу. Данные контроля представлены в файле <http://ieeep.tpu.ru/statlab/indstat5.sta>. Определить среднее значение и стандартное отклонение диаметра подшипника для каждой группы и всех групп. Построить гистограммы и сделать вывод о качестве поставляемой партии подшипников.

Задание 13. На производстве осуществляется контроль спиралей ламп для прожекторов. Контролируемым показателем качества является падение напряжения. Данные испытаний представлены в файле <http://ieeep.tpu.ru/statlab/indstat6.sta>. Определить среднее значение и стандартное отклонение контролируемого параметра для каждой партии и всех испытываемых образцов. Построить гистограммы и сделать вывод о качестве выпускаемых партий ламп.

Задание 14. Исходные данные (файл http://ieeep.tpu.ru/statlab/var_6_8.sta) представляют собой промеры толщины асфальта и уклона дороги при её асфальтировании. Смысл переменных: var1 – пикет, var2 – уклон дороги, var3 – толщина укатанного асфальта слева от оси, var4 – по оси, var5 – справа по оси. СНиП допускает толщину асфальта 6 ± 1 см и уклон от 10 до 20 единиц. Провести статистические испытания (на ваше усмотрение) и интерпретировать полученные результаты. Оценить качество асфальтирования дороги.

Задание 15. Исходные данные (файл <http://ieee.tpu.ru/statlab/bad.xls>) представляют собой статистику лабораторных анализов биологически активных добавок в процессе их изготовления. Смысл ячеек таблицы:

столбец А – номер партии;

строка 2 – содержание марганца (С), магния (D), кальция (E) и цинка (F) в продукте «БАД» (G); погрешности их измерения (I–M);

ячейки C1-G1: содержание в продукте «БАД» марганца, магния, кальция и цинка по ТУ;

столбец В – номер часа, в течение которого продукт готовится и по окончании которого отбираются пробы.

Провести статистические испытания (на ваше усмотрение) и интерпретировать полученные результаты. Оценить качество процессов изготовления БАД.

ГЛАВА 7. КЛАСТЕРНЫЙ АНАЛИЗ

7.1. Общие сведения о кластерном анализе

Кластерный анализ включает в себя набор различных алгоритмов классификации. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры. Например, биологи ставят цель разбить животных на различные виды, чтобы содержательно описать различия между ними. Задача кластерного анализа состоит в разбиении исходной совокупности объектов на группы схожих, близких между собой объектов. Эти группы называют кластерами. Другими словами, *кластерный анализ – это один из способов классификации объектов по их признакам* [5]. Желательно, чтобы результаты классификации имели содержательную интерпретацию.

Результаты, полученные методами кластерного анализа, применяют в самых различных областях. В маркетинге – это сегментация конкурентов и потребителей. В психиатрии для успешной терапии является решающей правильная диагностика симптомов, таких как паранойя, шизофрения и т. д. В менеджменте важна классификация поставщиков, выявление схожих производственных ситуаций, при которых возникает брак. В социологии – разбиение респондентов на однородные группы [16, 17]. В инвестировании важно сгруппировать ценные бумаги по сходству в тенденции доходности, чтобы составить оптимальный инвестиционный портфель. В общем, всякий раз, когда необходимо классифицировать большое количество информации такого рода и представлять её в виде, пригодном для дальнейшей обработки, кластерный анализ оказывается весьма полезным и эффективным.

Достоинство кластерного анализа состоит в том, что он работает даже тогда, когда данных мало и не выполняются требования нормальности распределений случайных величин и другие требования классических методов статистического анализа

Кластерный анализ позволяет провести объективную классификацию любых объектов, которые охарактеризованы рядом признаков. Из этого можно извлечь ряд преимуществ.

1. Полученные кластеры можно интерпретировать, то есть описывать, какие же собственно группы существуют.

2. Отдельные кластеры можно выбраковывать. Это полезно в тех случаях, когда при наборе данных допущены определённые ошибки, в результате которых значения показателей у отдельных объектов резко отклоняются. При применении кластерного анализа такие объекты попадают в отдельный кластер.

3. Для дальнейшего анализа могут быть выбраны только те кластеры, которые обладают интересующими характеристиками.

7.2. Методы кластеризации

В пакете Statistica реализуются следующие методы кластеризации.

1. Иерархические алгоритмы – древовидная кластеризация. В основе иерархических алгоритмов лежит идея последовательной кластеризации. На начальном шаге каждый объект рассматривается как отдельный кластер. На следующем шаге некоторые из ближайших друг к другу кластеров будут объединяться в отдельный кластер.

2. Метод К-средних. Этот метод используется наиболее часто. Он относится к группе так называемых эталонных методов кластерного анализа. Число кластеров K задаётся пользователем.

3. Двухходовое объединение. При использовании этого метода кластеризация проводится одновременно как по переменным (столбцам), так и по результатам наблюдений (строкам). Результатами процедуры являются описательные статистики по переменным и наблюдениям, а также двумерная цветная диаграмма, на которой цветом отмечаются значения данных. По распределению цвета можно составить представление об однородных группах.

7.3. Нормирование переменных для кластеризации

Разбиение исходной совокупности объектов на кластеры связано с вычислением расстояний между объектами и выбора объектов, расстояние между которыми наименьшее из всех возможных.

Наиболее часто используется привычное всем нам евклидово (геометрическое) расстояние. Эта метрика отвечает интуитивным представлениям о близости объектов в пространстве (как будто расстояния между объектами измерены рулеткой). Но для данной метрики на расстояние между объектами могут сильно влиять изменения масшта-

бов (единиц измерения). Например, если один из признаков измерен в миллиметрах, а затем его значение переведены в сантиметры, евклидово расстояние между объектами сильно изменится. Это приведет к тому, что результаты кластерного анализа могут значительно отличаться от предыдущих.

Если переменные измерены в разных единицах измерения, то требуется их предварительная нормировка, то есть преобразование исходных данных, которое переводит их в безразмерные величины.

В пакете Statistica нормировка любой переменной x выполняется по формуле:

$$x_{\text{норм}} = \frac{x - \mu}{\sigma}.$$

Для этого нужно щёлкнуть правой кнопкой мыши по имени переменной и в открывшемся меню выбрать последовательность команд: *Fill/ Standardize Block/ Standardize Columns*.

Значения нормированной переменной станут равными нулю, а дисперсии – единице.

7.4. Метод К-средних в программе Statistica

Метод К-средних (K-means) разбивает множество объектов на заданное число K различных кластеров, расположенных на возможно больших расстояниях друг от друга. Обычно, когда результаты кластерного анализа методом К-средних получены, можно рассчитать средние для каждого кластера по каждому измерению, чтобы оценить, насколько кластеры различаются друг от друга. В идеале вы должны получить сильно различающиеся средние для большинства измерений, используемых в анализе. Значения F -статистики, полученные для каждого измерения, являются другим индикатором того, насколько хорошо соответствующее измерение дискриминирует кластеры.

В качестве примера рассмотрим результаты опроса 17-ти сотрудников предприятия по удовлетворённости показателями качества служебной карьеры. В табл. 7.1 даны ответы на вопросы анкеты по десятибалльной шкале (1 – минимальный балл, 10 – максимальный). Имена переменных соответствуют ответам на следующие вопросы: СЛЦ – сочетание личных целей и целей организации; ОСО – ощущение справедливости в оплате труда; ТБД – территориальная близость к дому; ОЭБ – ощущение экономического благосостояния; КР – карьерный рост; ЖСР – желание сменить работу; ОСБ – ощущение социального благо-

получия. Данный пример представляет собой малую часть исследования, с полным вариантом которого можно ознакомиться в работе [16].

Таблица 7.1

Результаты опроса сотрудников

Сотрудник	Показатели качества служебной карьеры						
	СЛЦ	ОСО	ТБД	ОЭБ	КР	ЖСР	ОСБ
1	4	3	10	4	3	5	5
2	6	3	9	4	7	2	1
3	5	4	5	4	6	4	1
4	5	3	6	3	3	7	4
5	7	10	10	8	8	1	3
6	8	4	7	5	8	2	3
7	3	2	5	2	4	8	2
8	3	2	3	3	5	10	3
9	7	3	3	3	3	7	2
10	2	2	3	1	2	10	1
11	9	3	5	4	7	2	2
12	5	5	2	3	2	5	3
13	1	5	5	2	6	6	1
14	7	5	2	7	7	1	4
15	2	2	4	2	2	1	5
16	3	3	1	3	5	1	2
17	7	3	3	4	8	5	3

Используя эти данные, необходимо разделить сотрудников на группы и для каждой из них выделить наиболее эффективные рычаги управления. При этом различия между группами должны быть очевидными, а внутри группы респонденты должны быть максимально похожи.

На сегодняшний день большинство социологических опросов дает лишь процентное соотношение голосов: считается основное количество положительно ответивших, либо процент неудовлетворённых, но системно этот вопрос не рассматривают. Чаще всего опрос не показывает тенденции изменения ситуации. В некоторых случаях необходимо считать не количество человек, которые «за» или «против», а расстояние, или меру сходства, то есть определять группы людей, которые относятся к какому-либо вопросу примерно одинаково.

Для выявления на основе данных опроса некоторых реально существующих взаимосвязей признаков и порождения на этой основе их типологии можно использовать процедуры кластерного анализа. Наличие

каких-либо априорных гипотез социолога при работе процедур кластерного анализа не является необходимым условием.

В программе Statistica кластерный анализ выполняется следующим образом.

1. Создать файл данных из табл. 7.1.

2. Выбрать модуль *Statistics/ Multivariable Exploratory Techniques/ Cluster Analysis*. Нажать *OK*, в результате чего появится диалоговое окно, приведённое на рис. 7.1. В появившемся окне выбрать метод кластеризации *K-means clustering* и нажать *OK*.

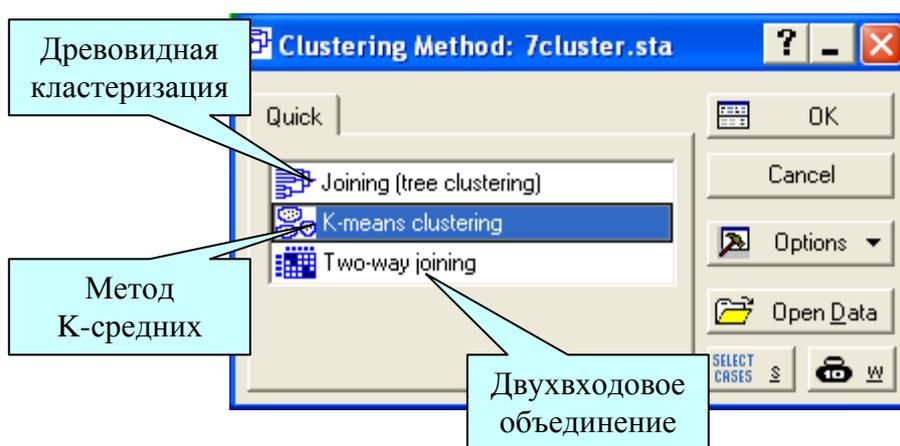


Рис. 7.1. Диалоговое окно выбора метода кластеризации

3. В появившемся диалоговом окне (рис. 7.2) необходимо установить следующие настройки.

Выбрать переменные кнопкой *Variables*.

Выбрать объекты кластеризации: это могут быть переменные – столбцы (*Variables (columns)*), либо наблюдения – строки (*Cases (Rows)*). Сначала проведём кластеризацию по строкам *Cases(rows)*.

Выбрать число кластеров. Этот выбор делает пользователь, исходя из собственных предположений о числе групп схожих объектов. При выборе количества кластеров руководствуйтесь следующим: количество кластеров, по возможности, не должно быть слишком большим. Расстояние, на котором объединялись объекты данного кластера, должно быть, по возможности, гораздо меньше расстояния, на котором к этому кластеру присоединяется ещё что-либо.

Нас интересует, например, как соотносятся ответы на вопросы анкеты у рядовых сотрудников и руководства предприятия. Поэтому выбираем $K=2$. Для дальнейшей сегментации можно увеличивать число кластеров.

Далее необходимо выбрать начальное разбиение объектов по кластерам (Initial cluster centers). Пакет Statistica предлагает 1) выбрать наблюдения с максимальным расстоянием между центрами кластеров; 2) рассортировать расстояния и выбрать наблюдения с постоянными интервалами (установка по умолчанию); 3) взять первые наблюдения за центры и присоединять остальные объекты к ним. Для наших целей подходит вариант 1).

Многие алгоритмы кластеризации часто «навязывают» данным не присущую им структуру и дезориентируют исследователя [10]. Поэтому крайне необходимо применять несколько алгоритмов кластерного анализа и делать выводы на основании общей оценки результатов работы алгоритмов

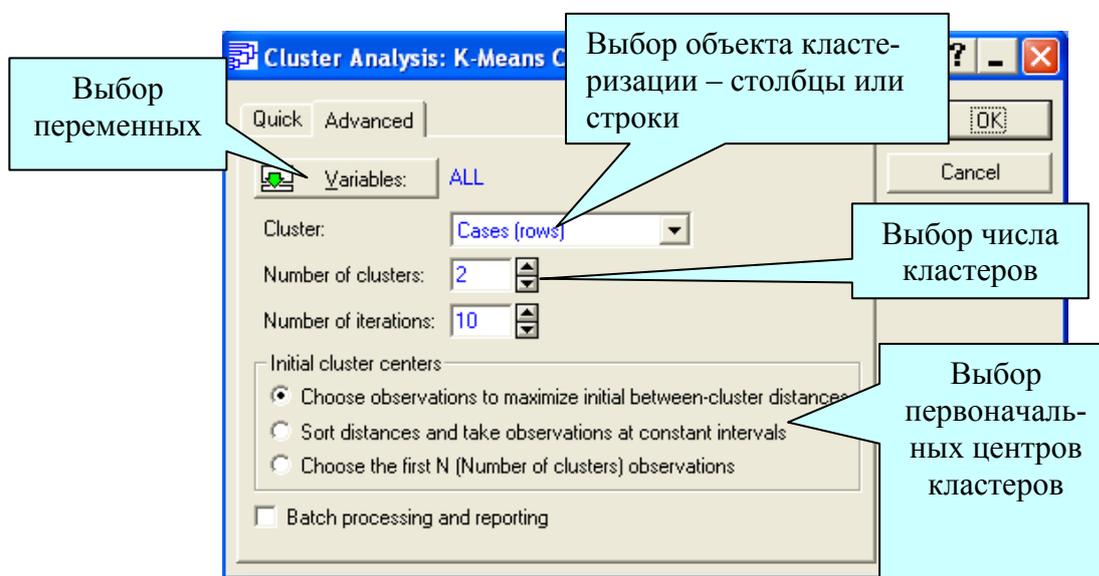


Рис. 7.2. Диалоговое окно выбора числа кластеров и типа кластеризации

4. Результаты анализа можно посмотреть в появившемся диалоговом окне, приведённом на рис. 7.3. Если выбрать вкладку *Graph of means*, будет построен график координат центров кластеров, который приведён на рис. 7.4. Каждая ломаная линия на этом графике соответствует одному из кластеров. Каждое деление горизонтальной оси графика соответствует одной из переменных, включенных в анализ. Вертикальная ось соответствует средним значениям переменных для объектов, входящих в каждый из кластеров.

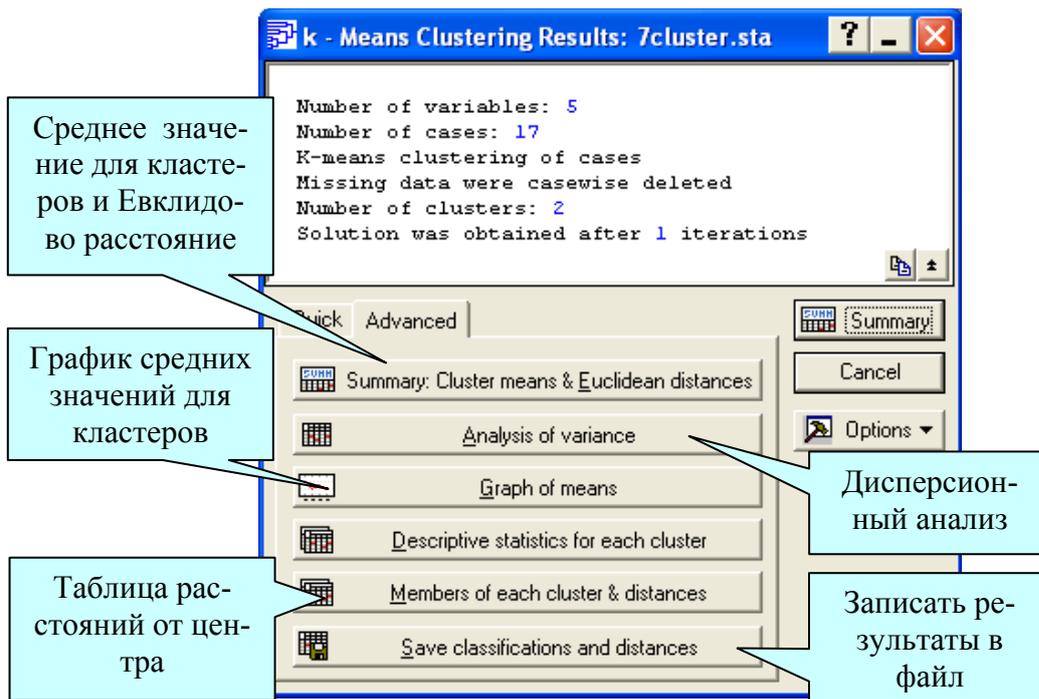


Рис. 7.3. Построение графика координат центров кластеров

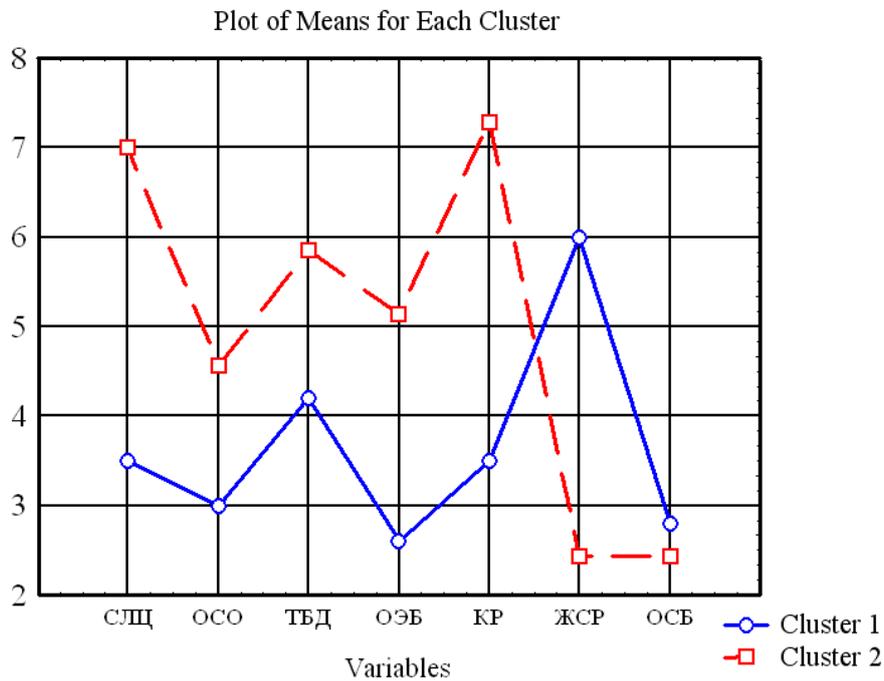


Рис. 7.4. График координат центров кластеров

Глядя на рис. 7.4, можно отметить, что просматриваются существенные отличия в отношении двух групп людей к служебной карьере почти по все вопросам. Лишь в одном вопросе наблюдается полное единодушие – в ощущении социального благополучия (ОСБ), вернее, отсутствии такового (2,5 балла из 10). Предположим, что кластер 1 в основном состоит из рабочих, а кластер 2 – из руководства. Руководители больше удовлетворены карьерным ростом (КР), сочетанием личных целей и целей организации (СЛЦ). У них выше уровень ощущения экономического благосостояния (ОЭБ) и ощущения справедливости в оплате труда (ОСО). Территориальная близость к дому (ТБД) волнует их меньше, чем рабочих, вероятно, из-за меньших проблем с транспортом. Также у руководителей меньше желания сменить работу (ЖСР).

Не смотря на то, что работники делятся на две категории, они подобным образом отвечают на большинство вопросов. Другими словами, если что-то не устраивает общую группу работников, то же самое не устраивает и высшее руководство, и наоборот. Согласование графиков позволяет сделать выводы о том, что благосостояние одной группы отражается на благосостоянии другой.

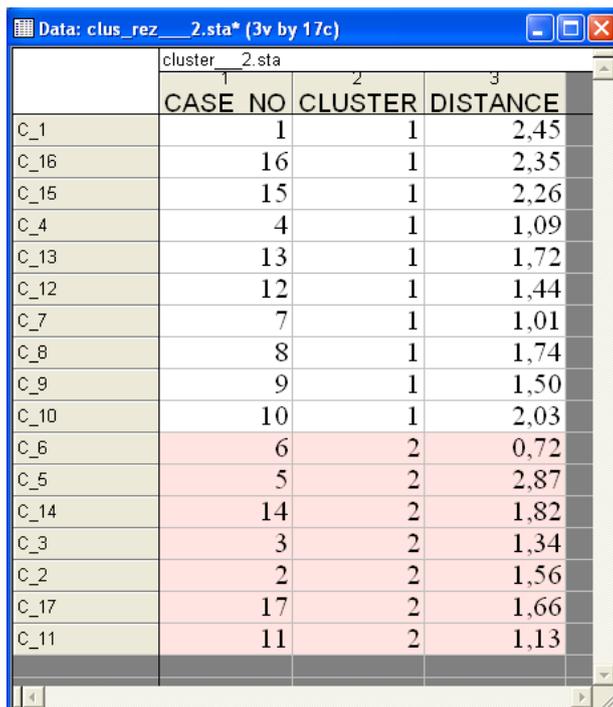
Сотрудники кластера 1 не довольны территориальной близостью к дому. Данной группой является основная часть работников, которые в основном приходят на предприятие с разных сторон города. Следовательно, можно предложить главному руководству направить часть прибыли на строительство жилья для сотрудников предприятия.

Просматриваются существенные отличия в отношении двух групп людей к служебной карьере. Те сотрудники, которых устраивает карьерный рост, у которых высоко совпадение личных целей и целей организации, не имеют желание сменить работу и ощущают удовлетворённость результатами труда. И наоборот, сотрудников, желающих сменить работу и неудовлетворённых результатами труда, не устраивают изложенные показатели. Высшему руководству следует обратить особое внимание на сложившуюся ситуацию.

Результаты дисперсионного анализа по каждому признаку выводятся по нажатию кнопки *Analysis of variance* (рис. 7.3). Выводятся суммы квадратов отклонения объектов от центров кластеров (*SS Within*) и суммы квадратов отклонений между центрами кластеров (*SS Between*), значения *F*-статистики и уровни значимости *p*. Для нашего примера уровни значимости для двух переменных довольно велики, что объясняется малым числом наблюдений. В полном варианте исследования, с которым можно ознакомиться в работе [17], гипотезы о равенстве сред-

них для центров кластеров отклоняются на уровнях значимости меньше 0,01.

Кнопка *Save classifications and distances* (рис. 7.3) выводит номера объектов, входящих в каждый кластер и расстояния объектов до центра каждого кластера. Рекомендуется рассортировать столбец с номером кластера (CLUSTER) по возрастанию (кнопкой A-Z на панели инструментов). Для рассматриваемого примера отсортированная таблица результатов будет иметь вид, показанный на рис. 7.5.



	cluster_1	cluster_2	cluster_3
	CASE_NO	CLUSTER	DISTANCE
C_1	1	1	2,45
C_16	16	1	2,35
C_15	15	1	2,26
C_4	4	1	1,09
C_13	13	1	1,72
C_12	12	1	1,44
C_7	7	1	1,01
C_8	8	1	1,74
C_9	9	1	1,50
C_10	10	1	2,03
C_6	6	2	0,72
C_5	5	2	2,87
C_14	14	2	1,82
C_3	3	2	1,34
C_2	2	2	1,56
C_17	17	2	1,66
C_11	11	2	1,13

Рис. 7.5. Состав каждого кластера и расстояния объектов от центра

В таблице показаны номера наблюдений (CASE_NO), составляющие кластеры с номерами CLUSTER и расстояния от центра каждого кластера (DISTANCE). Информация о принадлежности объектов к кластерам может быть записана в файл и использоваться в дальнейшем анализе. В данном примере сравнение полученных результатов с анкетами показало, что кластер 1 состоит, в основном, из рядовых работников, а кластер 2 – из менеджеров.

Таким образом, можно заметить, что при обработке результатов анкетирования кластерный анализ оказался мощным методом, позволяющим сделать выводы, к которым невозможно прийти, построив гисто-

грамму средних или посчитав процентное соотношение различных показателями качества трудовой жизни.

7.5. Древоподобная кластеризация

Древоподобная кластеризация – это пример иерархического алгоритма, принцип работы которого состоит в последовательном объединении в кластер сначала самых близких, а затем и всё более отдалённых друг от друга элементов. Большинство из этих алгоритмов исходит из матрицы сходства (расстояний), и каждый отдельный элемент рассматривается вначале как отдельный кластер.

После загрузки модуля кластерного анализа (рис. 7.1) и выбора *Joining (tree clustering)*, в окне ввода параметров кластеризации можно изменить следующие параметры:

- Исходные данные (*Input*). Они могут быть в виде матрицы исследуемых данных (*Raw data*) и в виде матрицы расстояний (*Distance matrix*).

- Кластеризацию (*Cluster*) наблюдений (*Cases (raw)*) или переменных (*Variable (columns)*), описывающих состояние объекта.

- Меры расстояния (*Distance measure*). Здесь возможен выбор следующих мер: евклидово расстояние (*Euclidean distances*), квадрат евклидова расстояния (*Squared Euclidean distances*), расстояние городских кварталов (манхэттенское расстояние, *City-block (Manhattan) distance*), расстояние Чебышёва (*Chebychev distance metric*), степенное расстояние (*Power...*), процент несогласия (*Percent disagreement*).

- Метод кластеризации (*Amalgamation (linkage) rule*). Здесь возможны следующие варианты: одиночная связь (метод ближайшего соседа) (*Single Linkage*), полная связь (метод наиболее удалённых соседей) (*Complete Linkage*), невзвешенное попарное среднее (*Unweighted pair-group average*), взвешенное попарное среднее (*Weighted pair-group average*), невзвешенный центроидный метод (*Unweighted pair-group centroid*), взвешенный центроидный метод (медиана) (*Weighted pair-group centroid (median)*), метод Уорда (*Ward's method*).

В результате кластеризации строится горизонтальная или вертикальная дендрограмма – график, на котором определены расстояния между объектами и кластерами при их последовательном объединении. Древоподобная структура графика позволяет определить кластеры в зависимости от выбранного порога – заданного расстояния между кластерами. Кроме того, выводится матрица расстояний между исходными

объектами (*Distance matrix*); средние и среднеквадратичные отклонения для каждого исходного объекта (*Distriptive statistics*).

Для рассмотренного примера (табл. 7.1) проведём кластерный анализ переменных с установками по умолчанию. Результирующая дендрограмма изображена на рис. 7.6. На вертикальной оси дендрограммы откладываются расстояния между объектами и между объектами и кластерами. Так, расстояние между переменными ОЭБ и ОСО равно пяти. Эти переменные на первом шаге объединяются в один кластер. Горизонтальные отрезки дендрограммы проводятся на уровнях, соответствующих пороговым значениям расстояний, выбираемым для данного шага кластеризации.

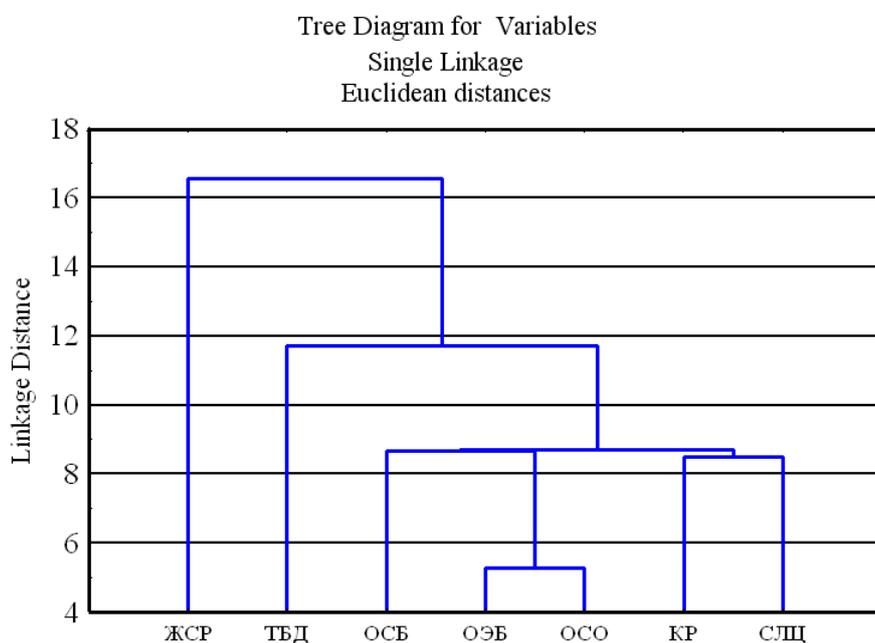


Рис. 7.6. Дендрограмма

Из графика видно, что вопрос «желание сменить работу» (ЖСР) образует отдельный кластер. Вообще, желание свалить куда угодно посещает всех в равной степени. Далее отдельный кластер составляет вопрос о территориальной близости к дому (ТБД). По степени важности он стоит на втором месте, что подтверждает вывод о необходимости строительства жилья, сделанный по результатам исследования методом К-средних. Ощущение экономического благосостояния (ОЭБ) и справедливости в оплате труда (ОСО) объединены – это блок экономических вопросов. Карьерный рост (КР) и сочетание личных целей и целей орга-

низации (СЛЦ) также объединены. Другие методы кластеризации, а также выбор других видов расстояний не приводит к существенному изменению дендрограммы.

7.6. Задания для самостоятельной работы

Задание 1. Исходные данные – файл http://ieee.tpu.ru/statlab/fact_ank.sta представляют собой результаты анкетного опроса работников предприятия по вопросам качества трудовой жизни. Для анализа выбрано 15 вопросов (переменные) и 41 наблюдение – ответы на вопросы по 10-бальной шкале (1 – не удовлетворяет, 4 – скорее не удовлетворяет, 7 – скорее удовлетворяет, 10 – удовлетворяет). Ответило 15 руководителей (первые 15 наблюдений) и 26 рабочих.

Провести кластерный анализ с целью:

- выявить наиболее проблемные вопросы, волнующие всех работников;
- выявить вопросы, на которые руководители и рабочие отвечают диаметрально противоположно.

По результатам анализа дать рекомендации для принятия управленческих решений: что необходимо сделать в первую очередь для улучшения качества трудовой жизни предприятия?

Задание 2. Исходные данные – файл <http://ieee.tpu.ru/statlab/wood-en.sta> представляют собой статистику отдела сбыта предприятия по производству стройматериалов. Для каждого вида продукции указан объём продаж по месяцам и выручка.

Провести кластерный анализ с целью:

сегментировать продукцию по месяцам года;

повышения выручки путём рационализации выпуска нужной продукции по месяцам.

Предположить существование трёх кластеров, на которые сегментируются выручка и объём продукции. По результатам анализа дать рекомендации для принятия управленческих решений: что необходимо сделать для улучшения качества производственного процесса.

ГЛАВА 8. ФАКТОРНЫЙ АНАЛИЗ

8.1. Идея факторного анализа

В основе различных методов факторного анализа лежит следующая гипотеза: наблюдаемые или измеряемые параметры являются лишь косвенными характеристиками изучаемого объекта, в действительности существуют внутренние (скрытые, не наблюдаемые непосредственно) параметры и свойства, число которых мало и которые определяют значения наблюдаемых параметров. Эти внутренние параметры принято называть *факторами*.

Цель факторного анализа – сконцентрировать исходную информацию, выражая большое число рассматриваемых признаков через меньшее число более ёмких внутренних характеристик явления, которые не поддаются непосредственному измерению

Установлено, что выделение и последующее наблюдение за уровнем общих факторов даёт возможность обнаруживать предотказные состояния объекта на очень ранних стадиях развития дефекта [18]. Факторный анализ позволяет отслеживать стабильность корреляционных связей между отдельными параметрами. Именно корреляционные связи между параметрами, а также между параметрами и общими факторами содержат основную диагностическую информацию о процессах. Применение инструментария пакета Statistica при выполнении факторного анализа исключает необходимость использования дополнительных вычислительных средств и делает анализ наглядным и понятным для пользователя.

Результаты факторного анализа будут успешными, если удастся дать интерпретацию выявленных факторов, исходя из смысла показателей, характеризующих эти факторы.

8.2. Сущность факторного анализа

Приведём несколько основных положений факторного анализа [18]. Пусть для матрицы $X[1:p, 1:n]$ измеренных параметров объекта

существует ковариационная (корреляционная) матрица $C[1:p, 1:p]$, где p – число параметров, n – число наблюдений. Путем линейного преобразования

$$X = QY + U \quad (8.1)$$

можно уменьшить размерность исходного факторного пространства $X[1:p]$ до уровня $Y[1:p']$, при этом $p' \ll p$. Матрица $Y[1:p'; 1:n]$ содержит ненаблюдаемые факторы, которые по существу являются гиперпараметрами, характеризующими наиболее общие свойства анализируемого объекта. Общие факторы чаще всего выбирают статистически независимыми, что облегчает их физическую интерпретацию. Вектор наблюдаемых признаков $X[1:p]$ имеет смысл следствия изменения этих гиперпараметров.

Матрица $U[1:p', 1:n]$ состоит из остаточных факторов, которые включают в основном ошибки измерения признаков $x(i)$. Прямоугольная матрица $Q[1:p, 1:p']$ содержит факторные нагрузки, определяющие линейную связь между признаками и гиперпараметрами.

Факторные нагрузки – это значения коэффициентов корреляции каждого из исходных признаков с каждым из выявленных факторов. Чем теснее связь данного признака с рассматриваемым фактором, тем выше значение факторной нагрузки. Положительный знак факторной нагрузки указывает на прямую (а отрицательный знак – на обратную) связь данного признака с фактором. Таким образом, данные о факторных нагрузках позволяют сформулировать выводы о наборе исходных признаков, отражающих тот или иной фактор, и об относительном весе отдельного признака в структуре каждого фактора.

Модель факторного анализа (8.1) похожа на модель многомерного регрессионного (4.1) анализа. Принципиальное отличие модели (8.1) в том, что вектор $Y[1:p']$ – это ненаблюдаемые факторы, а в регрессионном анализе – это регистрируемые параметры. В правой части уравнения (8.1) неизвестными являются матрица факторных нагрузок $Q[1:p, 1:p']$ и матрица значений общих факторов $Y[1:p']$.

Пакет статистического анализа Statistica позволяет в диалоговом режиме вычислить матрицу факторных нагрузок, а также значения нескольких заранее заданных главных факторов.

8.3. Факторный анализ в системе Statistica

Рассмотрим последовательность выполнения факторного анализа на примере обработки результатов анкетного опроса работников пред-

приятия (глава 7, табл. 7.1). Требуется выявить основные факторы, которые определяют качество трудовой жизни.

На первом этапе необходимо отобрать переменные для проведения факторного анализа. Используя корреляционный анализ, исследователь пытается выявить взаимосвязь исследуемых признаков, что, в свою очередь, даёт ему возможность выделить полный и безызбыточный набор признаков путём объединения сильно коррелирующих признаков.

Если проводить факторный анализ по всем переменным, то результаты могут получиться не совсем объективными, так как некоторые переменные определяются другими данными, и не могут регулироваться сотрудниками рассматриваемой организации.

Для того чтобы понять, какие показатели следует исключить, построим по имеющимся данным матрицу коэффициентов корреляции в Statistica: *Statistics/ Basic Statistics/ Correlation Matrices/ Ok*. В стартовом окне этой процедуры *Product-Moment and Partial Correlations* (рис. 4.3) для расчёта квадратной матрицы используется кнопка *One variable list*. Выбираем все переменные (*select all*), *Ok*, *Summary*. Получаем корреляционную матрицу (рис. 8.1).

Correlations							
Marked correlations are significant at p < ,05000							
N=17 (Casewise deletion of missing data)							
Variable	СПЦ	ОСО	ТБД	ОЭБ	КР	ЖСР	ОСБ
СПЦ	1,00	0,32	0,21	0,69	0,58	-0,44	0,07
ОСО	0,32	1,00	0,37	0,75	0,45	-0,43	0,02
ТБД	0,21	0,37	1,00	0,39	0,27	-0,21	0,13
ОЭБ	0,69	0,75	0,39	1,00	0,68	-0,61	0,26
КР	0,58	0,45	0,27	0,68	1,00	-0,48	-0,22
ЖСР	-0,44	-0,43	-0,21	-0,61	-0,48	1,00	-0,22
ОСБ	0,07	0,02	0,13	0,26	-0,22	-0,22	1,00

Рис. 8.1. Корреляционная матрица

Если коэффициент корреляции изменяется в пределах от 0,7 до 1, то это означает сильную корреляцию показателей. В этом случае можно исключить одну переменную с сильной корреляцией. И наоборот, если коэффициент корреляции мал, можно исключить переменную из-за того, что она ничего не добавит к общей сумме. В нашем случае сильной

корреляции между какими-либо переменными не наблюдается, и факторный анализ будем проводить для полного набора переменных.

Для запуска факторного анализа необходимо вызвать модуль *Statistics/ Multivariate Exploratory Techniques* (многомерные исследовательские методы)/ *Factor Analysis* (факторный анализ). На экране появится окно модуля *Factor Analysis* (рис. 8.2).

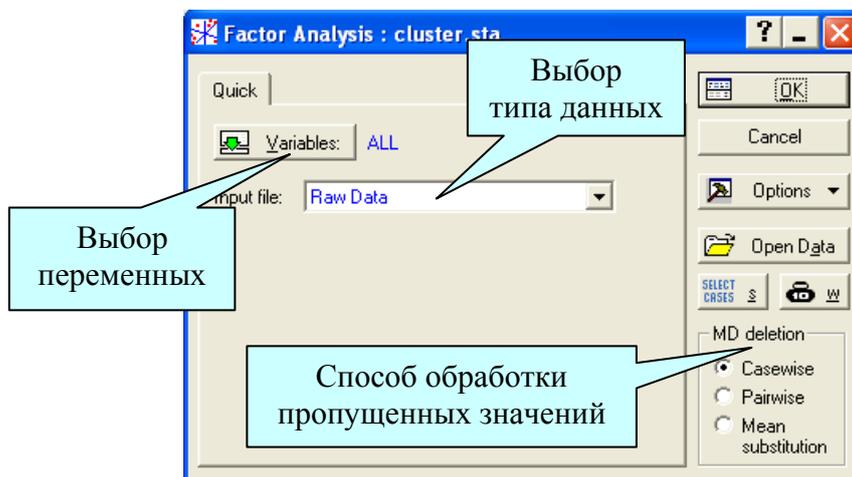


Рис. 8.2. Стартовое окно модуля «Факторный анализ»

Для анализа выбираем все переменные электронной таблицы; *Variables* (переменные): *select all, Ok*. В строке *Input file* (тип файла входных данных) указывается *Raw Data* (исходные данные). В модуле возможны два типа исходных данных – *Raw Data* (исходные данные) и *Correlation Matrix* – корреляционная матрица.

В разделе *MD deletion* задаётся способ обработки пропущенных значений:

- *Casewise* – способ исключения пропущенных значений (по умолчанию);
- *Pairwise* – парный способ исключения пропущенных значений;
- *Mean substitution* – подстановка среднего вместо пропущенных значений.

Способ *Casewise* состоит в том, что в электронной таблице, содержащей данные, игнорируются все строки, в которых имеется хотя бы одно пропущенное значение. Это относится ко всем переменным. В способе *Pairwise* игнорируются пропущенные значения не для всех переменных, а лишь для выбранной пары.

Выберем способ обработки пропущенных значений *Casewise*.

Statistica обработает пропущенные значения тем способом, который указан, вычислит корреляционную матрицу и предложит на выбор несколько методов факторного анализа.

После нажатия кнопки *Ok* появляется окно *Define Method of Factor Extraction* (определить метод выделения факторов), рис. 8.3.

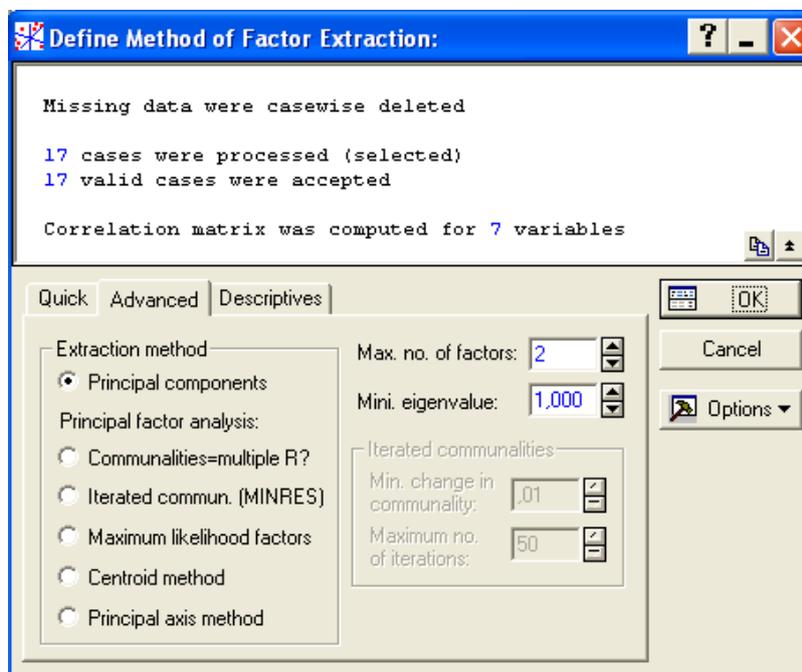


Рис. 8.3. Окно метода выделения факторов

Верхняя часть окна является информационной. Здесь сообщается, что пропущенные значения обработаны методом *Casewise*. Обработано 17 наблюдений и 17 наблюдений принято для дальнейших вычислений. Корреляционная матрица вычислена для 7 переменных. Нижняя часть окна содержит 3 вкладки: *Quick*, *Advanced*, *Descriptives*.

Во вкладке *Descriptives* (описательные статистики) имеются две кнопки:

 Review correlations, means, standard deviations

– просмотреть корреляции, средние и стандартные отклонения;

 Compute multiple regression analyses

– построить множественную регрессию.

Нажав на первую кнопку, можно посмотреть средние и стандартные отклонения, корреляции, ковариации, построить различные графики и гистограммы.

Во вкладке *Advanced*, в левой части, выберем метод (*Extraction method*) факторного анализа: *Principal components* (метод главных компонент). В правой части выбираем максимальное число факторов (2). Задаётся либо максимальное число факторов (*Max no of factors*), либо минимальное собственное значение: 1 (*eigenvalue*).

Нажимаем *Ok*, и Statistica быстро произведёт вычисления. На экране появляется окно *Factor Analysis Results* (результаты факторного анализа). Как говорилось ранее, результаты факторного анализа выражаются набором факторных нагрузок. Поэтому далее будем работать с вкладкой *Loadings*, рис 8.4.

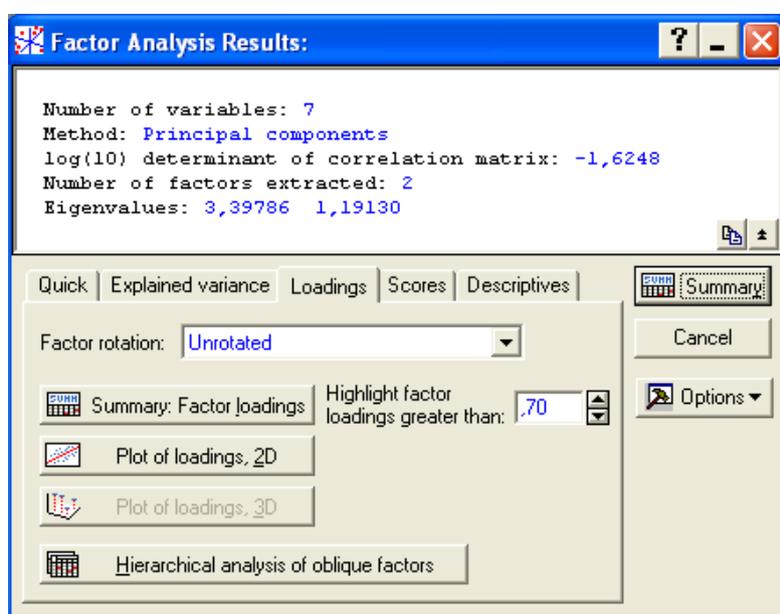


Рис. 8.4. Окно результатов факторного анализа

Верхняя часть окна – информационная:
Number of variables (число анализируемых переменных): 7;
Method (метод выделения факторов): *Principal components* (главных компонент);
Log (10) determinant of correlation matrix (десятичный логарифм детерминанта корреляционной матрицы): $-1,6248$;
Number of factors extracted (число выделенных факторов): 2;
Eigenvalues (собственные значения): 3,39786 и 1,19130.

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

Factor rotation – вращение факторов, в данном выпадающем окне можно выбрать различные повороты осей. С помощью поворота системы координат можно получить множество решений, из которого необходимо выбрать интерпретируемое решение.

Из всех предлагаемых методов, мы сначала посмотрим результат анализа без вращения системы координат – *Unrotated*. Если полученный результат окажется интерпретируемым и будет нас устраивать, то на этом можно остановиться. Если нет, можно вращать оси и посмотреть другие решения.

Щёлкаем по кнопке  и смотрим факторные нагрузки численно (рис. 8.5).

Factor Loadings (Unrotated) Extraction: Principal components (Marked loadings are > ,700000)		
Variable	Factor 1	Factor 2
СПЦ	-0,738919	-0,111165
ОСО	-0,748467	-0,025743
ТБД	-0,492660	0,185867
ОЭБ	-0,946896	0,100490
КР	-0,776403	-0,456091
ЖСР	0,724918	-0,168255
ОСБ	-0,155000	0,947260
Expl.Var	3,397863	1,191296
Prp.Totl	0,485409	0,170185

Рис. 8.5. Факторные нагрузки

Напомним, что *факторные нагрузки* – это значения коэффициентов корреляции каждой из переменных с каждым из выявленных факторов.

Значение факторной нагрузки, большее 0,7 показывает, что данный признак или переменная тесно связан с рассматриваемым фактором. Чем теснее связь данного признака с рассматриваемым фактором, тем выше значение факторной нагрузки. Положительный знак факторной нагрузки указывает на прямую (а отрицательный знак – на обратную) связь данного признака с фактором.

Итак, из таблицы факторных нагрузок было выявлено два фактора. Первый определяет ОСБ – ощущение социального благополучия. Остальные переменные обусловлены вторым фактором. Характерно, что все вопросы качества трудовой жизни предприятия составляют один фактор, а ощущение социального благополучия не связано с ним: это второй (политический) фактор.

В строке *Expl. Var* (рис. 8.5) приведена дисперсия, приходящаяся на тот или иной фактор. В строке *Prp. Totl* приведена доля дисперсии, приходящаяся на первый и второй фактор. Следовательно, на первый фактор приходится 48,5 % всей дисперсии, а на второй фактор – 17,0 % всей дисперсии, всё остальное приходится на другие неучтённые факторы. В итоге, два выявленных фактора объясняют 65,5 % всей дисперсии.

Посмотрим результат факторного анализа графически, нажав на кнопку *Plot of loadings, 2D*.

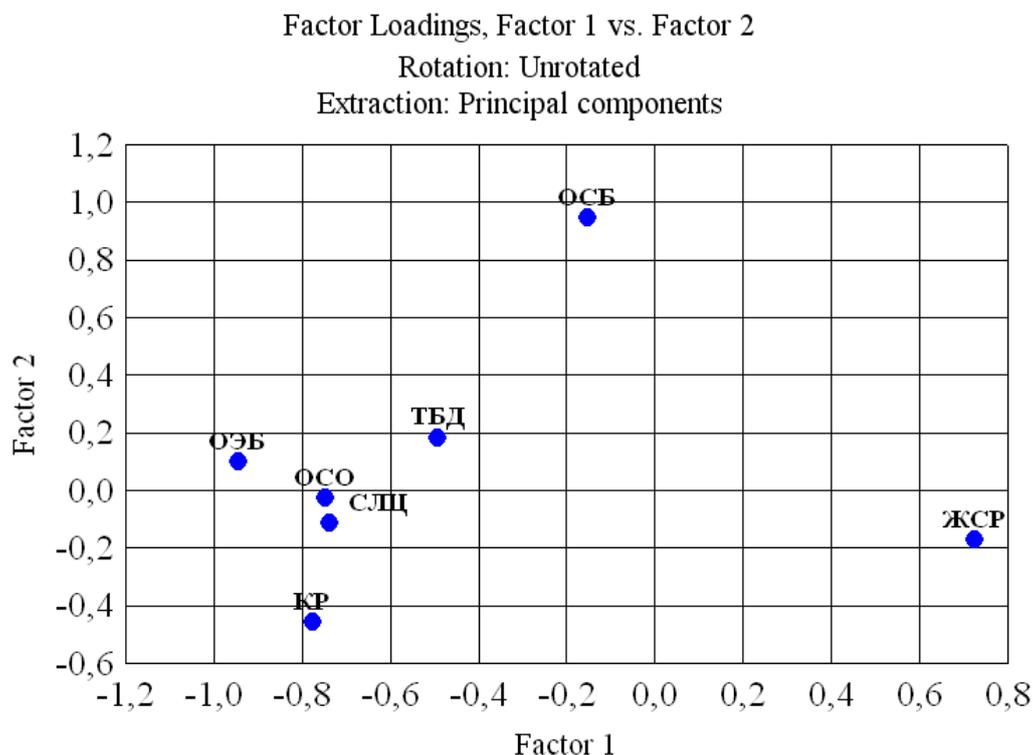


Рис. 8.6. График факторных нагрузок

Здесь мы также видим две группы факторов – ОСБ и остальное множество переменных, из которых выделяется ЖСР – желание сме-

нить работу. Имеет смысл исследовать это желание более основательно на основе сбора дополнительных данных.

8.4. Выбор и уточнение количества факторов

Как только получена информация о том, сколько дисперсии выделил каждый фактор, можно возвратиться к вопросу о том, сколько факторов следует оставить. По своей природе это решение произвольно. Но имеются некоторые общеупотребительные рекомендации, и на практике следование им даёт наилучшие результаты.

Количество общих факторов (гиперпараметров) определяется путём вычисления собственных чисел (рис. 8.7) матрицы $X[1:p, 1:n]$ в модуле факторного анализа. Для этого во вкладке *Explained variance* (рис. 8.4) необходимо нажать кнопку *Scree plot*.

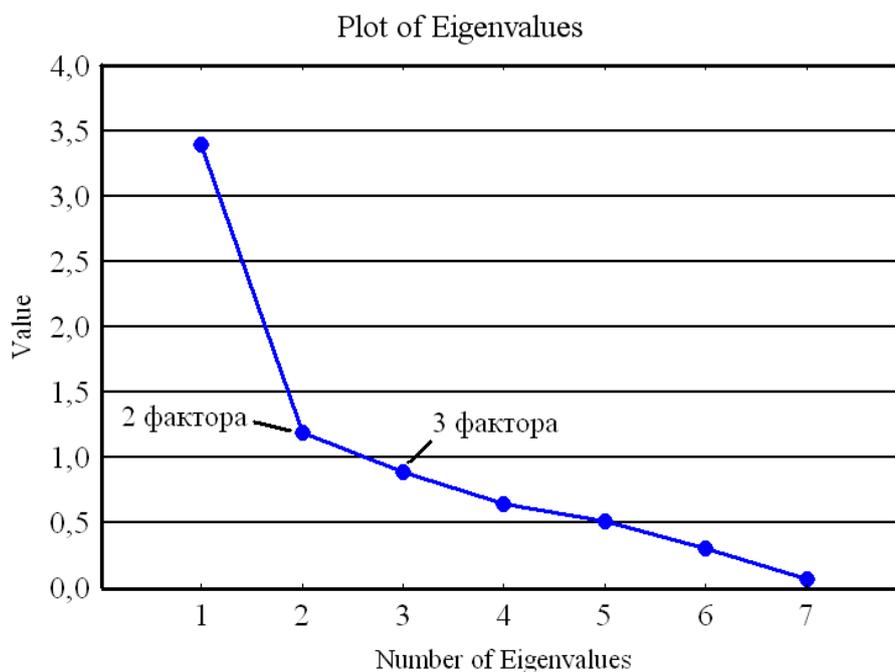


Рис. 8.7. График собственных чисел

Максимальное число общих факторов может быть равно количеству собственных чисел матрицы параметров. Но с увеличением числа факторов существенно возрастают трудности их физической интерпретации.

Сначала можно отобрать только факторы, с собственными значениями, большими 1. По существу, это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий используется наиболее широко. В приведённом выше примере на основе этого критерия следует сохранить только 2 фактора (две главные компоненты).

Можно найти такое место на графике (рис. 8.7), где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только «факториальная осыпь». В соответствии с этим критерием можно оставить в примере 2 или 3 фактора.

Из рис. 8.7 видно, что третий фактор незначительно увеличивает долю общей дисперсии.

8.5. Рекомендации по применению факторного анализа

Факторный анализ параметров позволяет выявить на ранней стадии нарушение рабочего процесса (возникновение дефекта) в различных объектах, которое часто невозможно заметить путём непосредственного наблюдения за параметрами. Это объясняется тем, что нарушение корреляционных связей между параметрами возникает значительно раньше, чем изменение одного параметра. Такое искажение корреляционных связей позволяет своевременно обнаружить факторный анализ параметров. Для этого достаточно иметь массивы зарегистрированных параметров.

Можно дать общие рекомендации по использованию факторного анализа вне зависимости от предметной области.

- На каждый фактор должно приходиться не менее двух измеренных параметров.
- Число измерений параметров должно быть больше числа переменных.
- Количество факторов должно обосновываться, исходя из физической интерпретации процесса.
- Всегда следует добиваться того, чтобы количество факторов было намного меньше числа переменных.

Пространство исходных признаков должно быть представлено в однородных шкалах измерения, т. к. это позволяет при вычислении использовать корреляционные матрицы. В противном случае возникает проблема «весов» различных параметров, что приводит к необходимо-

сти применения при вычислении ковариационных матриц. Отсюда может появиться дополнительная проблема повторяемости результатов факторного анализа при изменении количества признаков. Следует отметить, что указанная проблема просто решается в пакете Statistica путем перехода к стандартизированной форме представления параметров. При этом все параметры становятся равнозначными по степени их связи с процессами в объекте исследования.

8.6. Задания для самостоятельной работы

Задание 1. Исходные данные (файл <http://ieee.tpu.ru/statlab/medic.sta>) представляют собой удельные показатели ресурсов здравоохранения и социальных ресурсов для районов Новосибирской области и г. Новосибирска. Цель работы – исследовать взаимосвязь между заболеваемостью, инвалидностью и ресурсами здравоохранения, социальными ресурсами. Исследовать влияние на заболеваемость и инвалидность ресурсов здравоохранения. Найти наиболее значимые ресурсы здравоохранения и социальные ресурсы. Сформировать группы районов, обладающих близкими параметрами по различным показателям ресурсов здравоохранения и социальных ресурсов и по сумме всех показателей.

Задание 2. Исходные данные – файл <http://ieee.tpu.ru/statlab/sde.sta> представляют собой статистику затрат строительной компании. Найти факторы, существенно влияющие на накладные расходы. Сформировать группы переменных с близким объёмом расходов.

ЗАКЛЮЧЕНИЕ

Использование статистических методов управления процессами неизбежно выводит любую компанию на новый уровень конкурентоспособности. Статистические методы позволяют разработать стратегию развития компании на основе прогнозирования динамики основных показателей и соотношений между ними. Кроме того, большое значение для успешной работы компании имеют статистические методы контроля и анализа качества продукции. Несмотря на простоту методов управления процессами, они представляют собой мощный механизм повышения качества и могут использоваться для решения весьма обширного круга задач, когда приходится принимать решения в условиях действия многочисленных влияющих на процесс факторов.

Не смотря на разнообразие сфер применения статистики, имеются общие методы статистической работы, которыми нужно руководствоваться всегда и везде. Данная книга даёт представление об этих основных статистических методах и о том, как надо работать в пакете Statistica. Дополнительную информацию о методах промышленной статистики можно прочитать в работах [19, 20].

Учебное пособие теоретически обобщает и развивает методы промышленной статистики. Автор надеется, что книга вызвала интерес к рассматриваемым в ней вопросам. Он сочтёт себя вполне удовлетворённым, если хотя бы некоторые читатели, охотно расстающиеся с автором и его книгой, получившие целостное представление о проблеме статистического управления процессами, захотят ознакомиться с работами, указанными в списке литературы, и также приложить свои знания, опыт и математическое образование к рассматриваемым проблемам.

ЛИТЕРАТУРА

1. Теннант-Смит Дж. Бейсик для статистиков: Пер. с англ. / Дж. Теннант-Смит. – М.: Мир, 1988. – 208 с. – ISBN 5-03-000725-3.
2. Боровиков В.П. Программа Statistica для студентов и инженеров / В.П. Боровиков. – М.: Компьютер пресс, 2000. – 301 с.
3. Боровиков В.П. Statistica: искусство анализа данных на компьютере / В.П. Боровиков. – СПб.: Питер, 2001. – 656 с.
4. Боровиков В.П. Популярное введение в программу STATISTICA / В.П. Боровиков. – М.: Компьютер пресс, 1998. – 267 с.
5. Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL: учебное пособие / Э.А. Вуколов. – М.: ФОРУМ: ИНФРА-М, 2004. – 464 с. – (Профессиональное образование). – ISBN 5-8199-0141-X (ФОРУМ). – ISBN 5-16-002003-9 (ИНФРА-М).
6. Казанцева Н.Н. Статистический контроль и статистические методы управления качеством: учебное пособие / Н.Н. Казанцева. – Томск: Изд-во ТПУ, 2004. – 116 с.
7. Электронный учебник по статистике [Электронный ресурс]. – М.: StatSoft, Inc. – 2001. – Режим доступа: <http://www.statsoft.ru/home/textbook/default.htm>.
8. Плэтт В. Информационная работа стратегической разведки / В. Плэтт. – М.: Изд-во иностранной литературы, 1958.
9. Использование пакета Statistica 5.0 для статистической обработки опытных данных: методические указания / Сост.: С.В. Кабанов. – Саратов: Изд-во Саратов. гос. агр. ун-та, 2000. – 90 с.
10. Берестнева О.Г. Компьютерный анализ данных: учебное пособие / О.Г. Берестнева, Е.А. Муратова, А.М. Уразаев. – Томск: Изд-во ТПУ, 2003. – 204 с. : ил. – Библиогр.: С. 200–201. – ISBN 5-98298-004-8.
11. Горицкий Ю.А. Практикум по статистике с пакетом STATISTICA. Учебное пособие по курсу «Математическая статистика» / Ю.А. Горицкий. – М.: Изд-во МЭИ, 2000. – 44 с. – ISBN 5-7046-0573-7.
12. Каримов Р.Н. Обработка экспериментальной информации. Ч. 1. Разведочный анализ. Анализ качественных данных / Р.Н. Каримов. – Саратов: Саратовский госуд. техн. ун-т, 2002. – 112 с.
13. Каширина И.Б. Экономико-математическая модель прогнозирования спроса на образовательные услуги / И.Б. Каширина, В.Г. Мыслик // Моделирование систем. – 2002. – № 2 (4). – С. 46–53.

14. Eickelmann N. Statistical Process Control: What You Don't Measure Can Hurt You! / N. Eickelmann, A. Anant // IEEE Software. – 2003. – № 3. – P. 49–51.
15. Кочетков Е.П. Диалог консультанта с руководителем подразделения / Е.П. Кочетков. – Нижний Новгород: «Вектор-ТиС», 2003. – 112 с. – ISBN 5-93126-031-5.
16. Волкова Н.А. Кластерный анализ результатов социологического опроса работников предприятия / Н.А. Волкова, О.В. Стукач // Вестник Ульяновского государственного технического университета. – 2005. – № 2. – С. 68–72.
17. Мартюшева П.В. Кластерный анализ как инструмент менеджмента качества для обработки социологических опросов на промышленном предприятии / П.В. Мартюшева, О.В. Стукач // Доклады томского государственного университета систем управления и радиоэлектроники. – 2007. – Вып. 1 (15). – С. 71–76. – ISSN 1818-0442.
18. Рыбалко В.В. Параметрическое диагностирование энергетических объектов на основе факторного анализа в среде Statistica / В.В. Рыбалко // Exponenta Pro. – 2004. – № 2 (6). – С. 78–83.
19. Статистические методы повышения качества / Под ред. Х. Кумэ. – Пер. с англ. – М.: Финансы и статистика, 1990. – 304 с.
20. Ефимов В.В. Статистические методы в управлении качеством продукции: учебное пособие / В.В. Ефимов, Т.В. Барт. – М.: КНОРУС, 2006. – 240 с. – ISBN 5-85971-262-6.

СОДЕРЖАНИЕ

Предисловие	3
Глава 1. Разведочный визуальный анализ данных и структура программы Statistica	7
1.1. Сбор и анализ данных	7
1.2. Общие сведения о пакете Statistica	9
1.3. Запуск программы Statistica	12
1.4. Структура ввода и редактирования данных	13
1.5. Графический анализ данных	16
1.6. Диаграмма рассеяния	17
1.7. Трёхмерный визуальный анализ данных	18
1.8. Круговые диаграммы.....	20
1.9. Построение гистограмм	21
1.10. Задания для самостоятельной работы	25
Глава 2. Первичная обработка данных и вычисление элементарных статистик	27
2.1. Вероятность и достоверность.....	27
2.2. Генеральная совокупность и выборка	28
2.3. Простейшие описательные статистики	29
2.4. Примеры вычисления описательных статистик	33
2.5. Визуализация описательных статистик	35
2.6. Правило трёх частей	37
2.7. Нормальное распределение	38
2.8. Технологическое рассеяние и допуск на контролируемый показатель качества	41
2.9. Настройка, наладка и качество технологических процессов.....	42
2.10. Задания для самостоятельной работы	44
Глава 3. Проверка статистических гипотез	46
3.1. Статистические модели.....	46
3.2. Статистические гипотезы	46
3.3. Статистические критерии	47
3.4. Проверка гипотез с помощью критериев	48
3.5. Ошибки при принятии гипотез	49
3.6. Проверка гипотез о виде распределения.....	50
3.7. Проверка гипотез об однородности выборок	53
3.8. Задания для самостоятельной работы	55

Глава 4. Регрессия, корреляция и совпадение	61
4.1. Зависимость	61
4.2. Корреляция	62
4.3. Корреляционный анализ в программе Statistica	64
4.4. Ранговая корреляция	69
4.5. Основы регрессионного анализа	72
4.6. Пример проведения регрессионного анализа данных	74
4.7. Оценка адекватности модели по остаткам	78
4.8. Корреляционный и дисперсионный анализ модели	84
4.9. Фиксированная нелинейная регрессия	85
4.10. Пошаговая регрессия	90
4.11. Наилучшие регрессионные модели	92
4.12. Гребневая регрессия	92
4.13. Задания для самостоятельной работы	93
Глава 5. Нелинейные модели процессов	97
5.1. Нелинейная регрессия	97
5.2. Полиномиальная регрессия	100
5.3. Регрессионное моделирование в экономике	100
5.4. Задания для самостоятельной работы	103
Глава 6. Контроль качества	108
6.1. Статистические методы контроля качества	108
6.2. Цели управления качеством с помощью статистических методов	108
6.3. Диаграмма причин и результатов	109
6.4. Закон 80/20	113
6.5. Анализ Парето	114
6.6. Карты контроля качества	118
6.7. Контрольная карта индивидуальных значений	120
6.8. Контрольная карта средних значений и размахов	123
6.9. Чтение контрольных карт	124
6.10. Контрольные карты накопленных сумм	129
6.11. Задания для самостоятельной работы	129
Глава 7. Кластерный анализ	134
7.1. Общие сведения о кластерном анализе	134
7.2. Методы кластеризации	135
7.3. Нормирование переменных для кластеризации	135
7.4. Метод К-средних в программе Statistica	136
7.5. Древоподобная кластеризация	143
7.6. Задания для самостоятельной работы	145

Глава 8. Факторный анализ	146
8.1. Идея факторного анализа.....	146
8.2. Сущность факторного анализа.....	146
8.3. Факторный анализ в системе Statistica	147
8.4. Выбор и уточнение количества факторов	154
8.5. Рекомендации по применению факторного анализа	155
8.6. Задание для самостоятельной работы	156
Заключение	157
Литература.....	158

Учебное издание

СТУКАЧ Олег Владимирович

**ПРОГРАММНЫЙ КОМПЛЕКС STATISTICA
В РЕШЕНИИ ЗАДАЧ
УПРАВЛЕНИЯ КАЧЕСТВОМ**

Учебное пособие

Издано в авторской редакции

Научный редактор
*кандидат технических наук,
доцент В.Н. Бориков*

Дизайн обложки *А.И. Сидоренко*

**Отпечатано в Издательстве ТПУ в полном соответствии
с качеством предоставленного оригинал-макета**

Подписано к печати 30.09.2011. Формат 60x84/16. Бумага «Снегурочка».
Печать XEROX. Усл. печ. л. 9,48. Уч.-изд. л. 8,57.
Заказ 1381-11. Тираж 100 экз.



Национальный исследовательский Томский политехнический университет
Система менеджмента качества
Издательства Томского политехнического университета сертифицирована
NATIONAL QUALITY ASSURANCE по стандарту BS EN ISO 9001:2008



ИЗДАТЕЛЬСТВО  **ТПУ**. 634050, г. Томск, пр. Ленина, 30
Тел./факс: 8(3822)56-35-35, www.tpu.ru