МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

ДОПОЛНИТЕЛЬНЫЕ ГЛАВЫ МАТЕМАТИКИ. СТАТИСТИЧЕСКИЙ АНАЛИЗ

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

Составители

Н.И. Кривцова, О.Е. Мойзес

Издательство
Томского политехнического университета
2015

УДК 519.23(075.8)
ББК 22.172я73
Д68

Дополнительные главы математики. Статистический анализ : учебное пособие / сост. : Н.И. Кривцова, О.Е. Мойзес ; Томский политехнический университет. – Томск : Изд-во Томского политехнического университета, 2015. – 86 с.

Д68

В пособии изложены основы статистического анализа, включающие многомерную статистическую обработку данных, корреляционный и регрессионный анализы, методы многомерной классификации. Приведены основные методы оптимизации данных. Рассмотрен статистический программный продукт для управления большим количеством экспериментальных данных.

Предназначено для студентов, обучающихся по специальности 18.04.02 «Энерго- и ресурсосберегающие процессы в химической технологии, нефтехимии и биотехнологии».

УДК 519.23(075.8)
ББК 22.172я73

Рецензенты

Кандидат технических наук
заведующая лабораторией Института химии нефти СО РАН
Н.В. Юдина

Кандидат химических наук
научный сотрудник Института химии нефти СО РАН
М.А. Копытов

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Большинство задач и исследований в химической технологии сводится к нахождению оптимальных условий. В том случае, если процесс слишком сложен и невозможно составить его детерминированную модель или информации об объекте недостаточно, применяют экспериментальный путь к решению задачи исследования при разработке нового химико-технологического процесса. Для этой цели применяют экспериментально-статистические методы, т. е. при неизвестном механизме протекающих процессов изучают зависимость отклика системы на изменение входных параметров.

Статистические методы управления качеством процесса обладают в сравнении со сплошным контролем таким важным преимуществом, как возможность обнаружения отклонения во время технологического процесса, когда можно своевременно вмешаться в процесс и скорректировать его. Таким образом, основная задача статистического анализа состоит в том, чтобы при помощи заданного набора наблюдений уловить скрытые статистические закономерности в данных, установить, как одни случайные характеристики влияют на другие характеристики, построить модель зависимости.

Уравнения математического описания в этом случае представляют собой систему эмпирических зависимостей, полученных в результате статистического обследования объекта. Эти модели называются статическими и имеют вид корреляционных и регрессионных соотношений между входными и выходными параметрами объекта.

Широкому развитию и внедрению в практику методов многомерного статистического анализа способствует развитие вычислительной техники и программного обеспечения. Пакеты прикладных программ, таких как SPSS, STATISTICA, SAS и др., обладают удобным пользовательским интерфейсом, что снимает трудности в применении статистических методов. Основные трудности при решении задач статистики заключаются в основном в сложности математического аппарата, опирающегося на линейную алгебру, теорию вероятностей и математическую статистику, а также громоздкости вычислений.

Применение программ без понимания математической сущности используемых алгоритмов способствует развитию у исследователя иллюзии простоты применения многомерных статистических методов, что может привести к неверным или необоснованным результатам. Значимые практические результаты могут быть получены только на основе профессиональных знаний в предметной области.

Пособие предназначено для изучения вопросов статистического анализа экспериментальных данных. Цель пособия – дать студентам представление об основах статистического анализа, научить обработке большого количества экспериментальных данных с использованием программного продукта STATISTICA.

Основой пособия является материал лекционных и практических занятий по дисциплине «Дополнительные главы математики» образовательной программы подготовки магистров по направлению «Энерго- и ресурсосберегающие процессы в химической технологии, нефтехимии и биотехнологии». При составлении данного пособия учитывалось, что студенты знакомы со статистическим анализом одномерных величин, планированием полного и дробного факторного экспериментов, построением матриц планирования, вычислением матриц.

В пособии уделено внимание рассмотрению интерактивной системы анализа и управления данными, современной информационной технологии статистического анализа данных – STATISTICA.

Пособие рекомендуется использовать в процессе занятий, для подготовки к контрольным работам, зачетам и экзаменам, при выполнении лабораторных работ. Оно будет полезно инженерно-техническим и научным специалистам, специализирующихся в области химической технологии.

1. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Большинство задач и исследований в химической технологии сводится к нахождению оптимальных условий. При разработке нового химико-технологического процесса задача исследования может стоять так: получить максимальный выход реакции, варьируя температуру и давление, т. е. эта задача на условный экстремум.

Существует два подхода к решению этой задачи:

1. Всестороннее исследование процесса, его физико-химических закономерностей протекания.

2. Экспериментальный путь – в том случае, если процесс слишком сложен и невозможно составить его детерминированную модель или информации об объекте недостаточно.

В то же время задачу оптимизации решать необходимо, и для этих целей применяют экспериментально-статистические методы, т. е. при неизвестном механизме протекающих процессов изучают зависимость отклика системы на изменение входных параметров. Уравнения математического описания в этом случае представляют собой систему эмпирических зависимостей, полученных в результате статистического обследования объекта. Они называются статическими и имеют вид корреляционных и регрессионных соотношений между входными и выходными параметрами объекта.

Статистические методы управления качеством процесса обладают в сравнении со сплошным контролем таким важным преимуществом, как возможность обнаружения отклонения во время технологического процесса, когда можно своевременно вмешаться в процесс и скорректировать его.

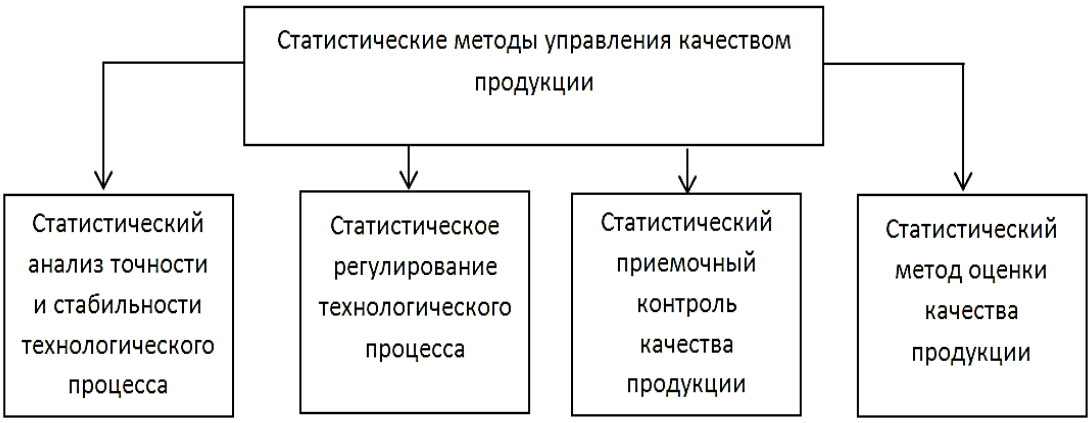


Рис. 1.1. Статистические методы управления качеством продукции

Основные области применения статистических методов управления качеством продукции представлены на рис. 1.1.

Коротко раскроем понятия, используемые на рисунке.

Статистический анализ точности и стабильности технологического процесса – это установление статистическими методами значений показателей точности и стабильности технологического процесса и определение закономерностей его протекания во времени.

Статистическое регулирование технологического процесса – это корректирование значений параметров технологического процесса по результатам выборочного контроля контролируемых параметров, осуществляемое для технологического обеспечения требуемого уровня качества продукции.

Статистический приемочный контроль качества продукции – это контроль, основанный на применении методов математической статистики для проверки соответствия качества продукции установленным требованиям и принятия продукции.

Статистический метод оценки качества продукции – это метод, при котором значения качества показателей качества продукции определяют с использованием правил математической статистики.

Общий вид статистических моделей представлен в виде:

$$y = \beta_0 + \sum_{i=1}^{N} \beta_i x_i + \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \beta_{ij} x_i x_j + \sum_{i=1}^{N} \beta_{ii} \cdot x_i^{\,2} + \ldots$$

где β_i, β_{ij} – теоретические коэффициенты, характеризующие линейные эффекты взаимодействия и квадратичные эффекты или коэффициенты уравнения регрессии.

Уравнение регрессии (полином 2) используется для построения статистических моделей объектов химической технологии. Справедлива такая модель только для объекта, на котором проводили эксперимент. Однако такие модели широко используются при решении задач оптимизации.

1.1. Элементы теории математической статистики

1.1.1. Случайная величина

При выполнении эксперимента мы имеем дело со случайными величинами.

Случайная величина – величина, значение которой принципиально нельзя предсказать, исходя из условий опыта.

Различают *непрерывные* и *случайные дискретные величины.*

Возможные значения *дискретных* случайных величин можно заранее перечислить. Дискретные переменные могут принимать только отдельные значения в некотором интервале (число атомов углерода).

Значения *непрерывной* случайной величины заранее перечислить нельзя, они непрерывно занимают некоторый промежуток. Например, переменные, связанные с непрерывным процессом, такие как t^o, давление, состав.

Для характеристики случайной величины необходимо указать, какие значения она может принимать и с какой частотой.

Пусть проведена серия опытов, которая включает n экспериментов. При этом непрерывное событие X (результат измерений) произошло m_i раз. Тогда отношение m_i/n называется **частотой появления события** X.

Частота m_i/n – тоже величина случайная и изменяется в зависимости от количества проведенных опытов. При большом числе опытов она может стабилизироваться около некоторого значения (P).

Предел, к которому стремится это отношение, при неограниченном возрастании числа экспериментов называется **вероятностью случайного события** X.

$$P\{X\} = lim\frac{m_i}{n_i}.$$

Вероятность случайного события – это отношение числа благоприятных исходов события к полному числу при большой продолжительности эксперимента.

Пример.

Соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими вероятностями, называется ***законом распределения случайной величины.***

Дискретную случайную величину также можно задать вероятностным рядом.

$$\frac{x_1 \cdot x_2 \dots \dots \dots x_n}{P(x_1) \cdot P(x_2) \dots \dots \dots . P(x_3)} - \text{ вероятностный ряд}$$

Распределение **непрерывной** случайной величины нельзя задавать при помощи вероятностей отдельных значений. Для непрерывных случайных величин принимается вероятность того, что в результате опыта значение случайной величины попадет в некоторый интервал (или в заранее намеченную совокупность чисел).

В виде функции распределения можно задать распределение как непрерывной, так и случайной дискретной величины (рис. 1.2 и 1.3). Ордината кривой, соответствующая т. x_1, есть вероятность того, что случайная величина X при испытании окажется меньше x_1.

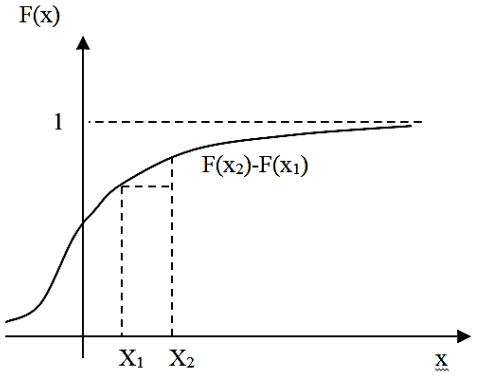


Рис. 1.2

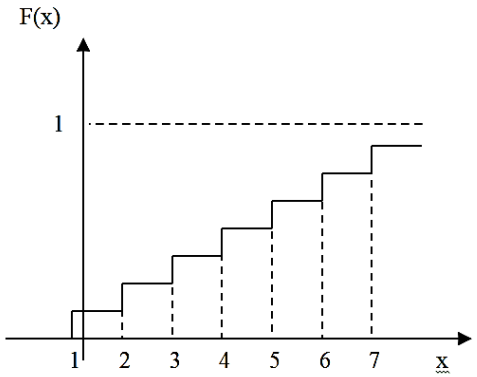


Рис. 1.3

Вероятность того, что значение случайной величины X заключено между x_1 и x_2 и равно разности функции распределения, вычисленных в этих точках.

$$P(x_1 \le X \le x_2) = F(x_2) - F(x_1).$$

Функция распределения случайной дискретной величины – всегда разрывная ступенчатая функция, скачки которой происходят в точках, соответствующих возможным значениям случайной величины, и равны вероятностям этих значений. Сумма всех скачков равна 1.

Свойства интегральной функции распределения случайной величины X:

1) $F(x) \ge 0$;

2) $F(x_2) > F(x_1)$, если $x_2 > x_1$;

3) $F(+\infty) = 1$;

4) $F(-\infty) = 0$; следовательно,

5) $0 \le F(x) \le 1$.

Для непрерывной случайной величины часто применяется производная функции распределения – это ***плотность распределения вероятности***:

$$f(x) = F'(x);$$

$$f(x) = \frac{dF(x)}{dx} = \frac{\lim_{\Delta x \to 0}(x < X \le x + \Delta x)}{\Delta x}.$$

Т. е. $F(x)$ равно отношению вероятности попадания случайной величины в интервал $(x, x + \Delta x)$ к длине этого интервала (Δx), когда Δx – бесконечно малая величина.

$$F(x) = P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f(x)dx = F(x_2) - F(x_1).$$

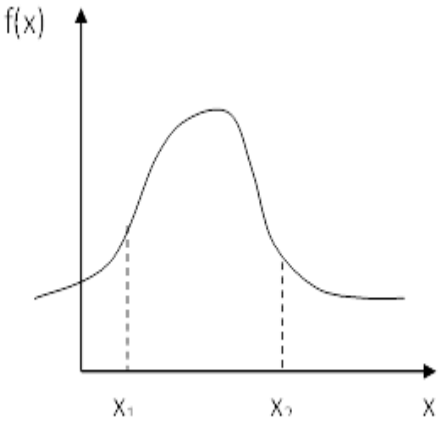


Рис. 1.4

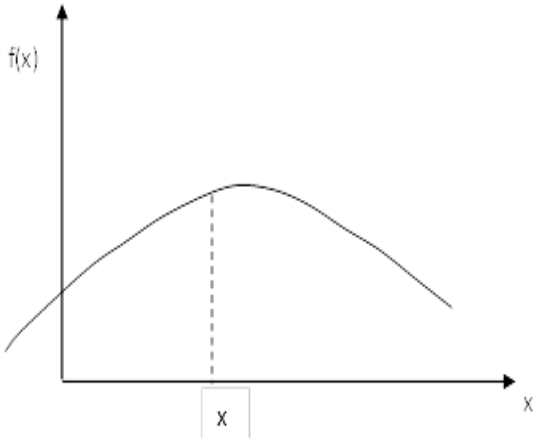


Рис. 1.5

Площадь, ограниченная осью x, прямыми $x = x_1$ и $x = x_2$ и кривой $f(x)$, равна вероятности того, что случайная величина примет значение из интервала $[x_1, x_2]$.

$$F(x) = P(-\infty \le X \le x) = \int_{-\infty}^{x} f(x)dx$$ – частный случай.

Свойства $f(x)$:

1) $f(x) \ge 0$;

2) $\int_{-\infty}^{\infty} f(x)dx = 1$ – условие нормировки.

Попадание случайной величины в интервал $-\infty < X < +\infty$ есть достоверное событие.

Также некоторые основные свойства случайных величин могут быть описаны с помощью определенных числовых характеристик.

Наиболее часто применяются на практике два параметра: математическое ожидание и дисперсия.

1.1.2. Математическое ожидание случайной величины

Математическое ожидание (среднее значение) случайной величины – это параметр, который характеризует центр рассеяния (центр распределения) случайной величины. Принято обозначать $M[x]$, m_x, m.

Для *дискретной* случайной величины:

$$m_x = M[x] = \sum_{i=1}^{n} x_i \cdot p_i.$$

Для *непрерывной* случайной величины:

$$m_x = M[x] = \int_{-\infty}^{\infty} x_i \cdot f(x)dx.$$

Т. е. математическое ожидание случайной величины приблизительно равно среднему арифметическому всех результатов, полученных при большом числе испытаний над этой величиной.

Свойства математического ожидания:

1. Математическое ожидание постоянной величины есть сама эта величина: $M[c] = C$.

2. Постоянный множитель можно выносить за знак математического ожидания: $M[cx] = c \cdot M[x]$.

3. Математическое ожидание суммы нескольких случайных величин равно сумме их математических ожиданий:

$$M[x + y + z] = M[x] + M[y] + M[z].$$

4. Математическое ожидание произведения равно произведению математического ожидания:

$$M[x \cdot y] = M[x] \cdot M[y]$$

1.1.3. Дисперсия

Дисперсия – параметр, характеризующий степень отклонения (рассеяния) случайной величины от ее среднего значения. Т. е. это параметр, характеризующий разброс значений этой величины.

Дисперсией случайной величины X называется математическое ожидание квадрата разности случайной величины и ее математического ожидания.

$$D[x] = M\left(X - M[X]\right)^2\Big),$$

$$D[x] = M\left(X - m_x\right)^2.$$

Для *дискретной* случайной величины:

$$D[x] = \sum_{i=1}^{n}\left(x_i - M[x]\right)^2 \cdot P_i.$$

Для *непрерывной* случайной величины:

$$D[x] = \int_{-\infty}^{\infty}\left(x_i - m_X\right)^2 \cdot f(x) \cdot dx.$$

Дисперсия обозначается: $D[x]$, σ_x^2, σ^2.

Величину $\sigma = \sqrt{D[x]}$ называют **среднеквадратическим отклонением**, или **стандартом.**

Свойства дисперсии:

1. Дисперсия от постоянной величины равна 0:

$$D[C] = 0.$$

2. Постоянную величину можно выносить за знак дисперсии в квадрате:

$$D[CX] = C^2D[X].$$

3. Дисперсия суммы случайной величины равна сумме дисперсий этих величин:

$$D[x_1 + x_2 + x_3 + \ \dots\ ] = D[x_1] + D[x_2] + D[x_3] + \dots + D[x_n].$$

4. $D[x] = M[x^2] - m^2_x$.

1.2. Законы распределения случайной величины

Полученное в результате эксперимента распределение необходимо аппроксимировать каким-либо теоретическим законом распределения. Такая аппроксимация позволяет математически описать и проанализировать результаты исследований.

При аппроксимации результатов эксперимента используются различные законы, например: **экспоненциальный, равномерный нормальный закон распределения (закон Гаусса).**

Это наиболее часто применяемый на практике закон распределения.

Случайная непрерывная величина X называется распределенной по нормальному закону, если ее плотность распределения имеет вид:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}; \quad \text{или} \quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{\left(x-\bar{x}\right)^2}{2\sigma^2}};$$

$$-\infty < x < \infty,$$

где M_x – математическое ожидание; σ^2 – дисперсия случайной величины; X – стандартное отклонение от математического ожидания, симметричное m; σ – среднеквадратичное отклонение (стандартное); σ и σ^2 характеризуют разброс данных, т. е. ошибку измерений.

Чем больше σ, тем больше ошибка.

График плотности распределения вероятности называется **нормальной кривой**, или **кривой Гаусса**. Это кривая колоколообразного вида, симметричная m_x. Нормальная кривая симметрична относительно $x = m_x$ и при $x \to \infty$ приближается к оси абсцисс.

При $x = m_x$ кривая имеет максимум, равный $\frac{1}{\sigma\sqrt{2\pi}}$

По мере удаления от точки m плотность распределения падает.

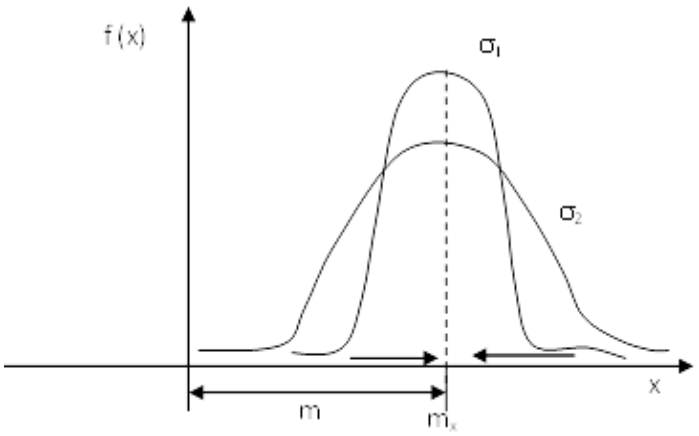


Рис. 1.6

Из графика (рис. 1.6) следует, что для нормального распределения наибольшая вероятность попадания x в окрестности математического ожидания m_x, а по обе стороны от него эта вероятность монотонно убывает.

Нормальное распределение при $m = 0$ и $\sigma = 1$ называется **стандартным.**

2. МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

Технологические объекты, как правило, характеризуются достаточно большим числом параметров, образующих многомерные векторы, и особое значение приобретают задачи изучения взаимосвязей между компонентами этих векторов, причем эти взаимосвязи необходимо выявлять на основании ограниченного числа многомерных наблюдений.

Многомерный статистический анализ – раздел математической статистики, изучающий математические методы построения оптимальных планов сбора, систематизации и обработки многомерных статистических данных, направленных на выявление характера и структуры взаимосвязей между компонентами исследуемого признака или процесса и предназначенных для получения научных и практических выводов.

Многомерный статистический анализ объединяет методы изучения статистических данных, относящихся к объектам, которые характеризуются качественными и количественными признаками.

Многомерные статистические методы среди множества возможных статистических моделей позволяет обусловлено выбрать ту модель, которая наилучшим образом описывает исходные статистические данные, характеризующие реальное поведение исследуемой совокупности объектов, а также оценить надежность и точность выводов, сделанных на основании ограниченного статистического материала.

Предположим, что рассматривается некоторая совокупность, состоящая из n нефтяных объектов, для каждого из которых известны показатели: $X1$ − содержание ароматических соединений, $X2$ − вязкость нефтепродукта, $X3$ − концентрация серы и т. п. В результате получен набор из n наблюдений над k-мерным случайным вектором $X = (X1, X2, \ldots, Xk)$.

По наблюдавшимся значениям случайного вектора $X = (X1, X2, \ldots, Xk)$ может понадобиться:

- изучить связь между его компонентами $X1, X2, \ldots, Xk$;
- определить, какие из (большого числа) рассчитанных показателей $X1, X2, \ldots, Xk$ в наибольшей степени влияют на физико-химию или на состав нефтепродукта;
- классифицировать нефтепродукты по какому-либо признаку.

Особенностью многомерного статистического анализа является то, что результаты отдельных наблюдений независимы и подчинены многомерному нормальному распределению.

Выделяют следующие преимущества многомерного статистического анализа:

1) модель приемлема для большого числа приложений;

2) только в рамках этой модели можно вычислить точное распределение выборочных характеристик.

Основная задача статистического анализа состоит в том, чтобы при помощи заданного набора наблюдений уловить скрытые статистические закономерности в данных, установить, как одни случайные характеристики влияют на другие, и построить модель зависимости.

Статистическая совокупность объектов – объекты, явления, события, которые составляют круг интересов исследователя при решении конкретной задачи анализа.

Выборка объектов – доля статистической совокупности, информация которой имеется в расположении исследователя. Выборка составляется в результате случайного отбора некоторых представителей совокупности.

Объем выборки – число представителей выборки.

Характеристика объекта – особые свойства объекта. Характеристика бывает количественная, качественная и порядковая.

Набор характеристик, или система статистических показателей, содержит различные характеристики от x_1 до x_n, которыми описываются объекты. Набор может содержать характеристики одного типа, а может включать характеристики разных типов – как количественные, так и качественные.

Пространство характеристик – множество всевозможных значений, которое может принимать набор характеристик. Набор характеристик может включать зависимые характеристики, т. е. такие характеристики, каждая из которых зависит от влияющих характеристик.

Временной ряд – набор наблюдений характеристики одного объекта в различные моменты времени.

Многомерный временной ряд представляет собой набор наблюдений нескольких характеристик одного объекта.

Модель зависимости – математическая зависимость процесса, как какая-либо одна или несколько характеристик зависят от других характеристик.

Форма записи моделей – уравнение или система уравнений, формулы, набор логических утверждений, график или дерево решений.

Данная модель используется для прогнозирования значений характеристики в зависимости от значения другой характеристики.

В результате устанавливается соответствие между множественными значениями этих характеристик.

2.1. Многомерные случайные величины и их распределение

Совокупность m функций, определенных на одном и том же множестве элементарных событий, называется m-мерной случайной величиной и записывается как

$$\zeta = \{\zeta_1 \dots \zeta_m\}.$$

Многомерные случайные величины полностью определяются ее функцией распределения вероятностей:

$$F(x_1 \dots x_m) = P(\zeta \le x_1 \dots \zeta_m \le x_m),$$

которое удовлетворяет следующим условиям:

1. $0 \le F(x_1 \dots x_m) \le 1.$
2. $F(x_1 \dots x_m)$ не убывает по каждому аргументу.
3. $\lim\limits_{\text{по всем } x_i \to \infty} F(x_1 \dots x_m).$

При этом одномерные случайные величины ζ_i и ζ_j называются независимыми, если их совместная функция распределения равна произведению одномерной функции распределения:

$$F(x_i, x_j) = F(x_i) \cdot F(x_j).$$

2.1.1. Дискретные многомерные случайные величины

Многомерная случайная величина называется дискретной, если составляющие ее случайной величины являются дискретными.

Многомерная дискретная случайная величина полностью определяется набором значений вероятностей

$$P_{i1-im} = P(\zeta_1 = x_{i1} \dots \zeta_m = x_{im}),$$

заданных для любой комбинации значений $x_{i1} \dots x_{im}$ случайной величины $\zeta_1 \dots \zeta_m$.

Функция распределения вероятности выражается через вероятности $P_{i1} \dots P_{im}$ и записывается следующим образом:

$$F(x_1 \dots x_m) = \sum P_{i1 \dots m};\ x_{i1} \le x_1 \dots x_{im} \le x_m.$$

Рассмотрим в качестве примера двумерную случайную величину (ζ, η), принимающую значения (x_i, y_i), при этом $i = 1...k$, $j = 1...l$. Вероятности всех пар значений можно представить в виде таблицы:

ζ	η					
	y_1	...	y_j	...	y_l	*Итог*
x_1	P_{11}	...	P_{1j}	...	P_{1l}	P_1
⋮	⋮	…	⋮	…	⋮	⋮
x_i	P_{i1}	...	P_{ij}	...	P_{il}	P_i
⋮	⋮	…	⋮	…	⋮	⋮
x_k	P_{k1}	...	P_{kj}	...	P_{kl}	P_k
Итог	P_1	...	P_j	...	P_l	

$$P_j = \sum_{i=1}^{k} P_{ij}\ ; \qquad P_i = \sum_{j=1}^{k} P_{ij}$$

В итоговой строке и столбце записаны суммы по столбцам и строкам. Итоговый столбец определяет одномерное распределение случайной величины ζ, а итоговая строка – одномерное распределение случайной величины η.

Если разделить все вероятности ζ-го столбца на итоговую вероятность P_j, то получим условные вероятности значений ζ, при условии что $\eta = y_i$, которые определяют условное распределение случайной величины ζ при фиксированном значении другой случайной величины $\eta = y_i$.

Аналогично определяется условное распределение η при заданном значении ζ.

Если $P_{ij} = P_i \cdot P_j$ для любых $i.j.(i \neq j)$, то случайные величины η и ζ являются независимыми.

2.1.2. Непрерывные случайные величины

Закон распределения системы непрерывной случайной величины (x, y) задается с помощью функции плотности вероятности $f(x, y)$.

Многомерная случайная величина называется непрерывной, если непрерывна ее функция распределения $F(x_1...x_m)$ и существует почти всюду функция плотности $f(x_1...x_m)$, такая, что

$$F(x_1...x_m) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_m} f(x_1...x_m)\,dx_1...dx_m\ .$$

Вероятность попадания случайной точки (x, y) в область D определяется равенством

$$P\left[(x, y) \in D\right] = \iint\limits_{D} f(x, y)\,dxdy,$$

при этом функция плотности вероятности обладает следующими свойствами:

1. $f(x, y) \geq 0.$
2. $\int\limits_{-\infty}^{+\infty} \ldots \int\limits_{-\infty}^{+\infty} f(x, y)\,dxdy = 1.$

Если все случайные точки (x, y) принадлежат конечной области D, тогда

$$\iint\limits_{D} f(x, y)\,dxdy = 1.$$

Перейти от первоначальной системы большого числа наблюдаемых взаимосвязанных факторов к системе существенно меньшего числа ненаблюдаемых факторов, определяющих вариацию первоначальных признаков, позволяют методы снижения размерности многомерного пространства. При этом потери информации практически не происходит. Методы компонентного и факторного анализа позволяют выявлять объективно существующие, но непосредственно не наблюдаемые закономерности при помощи главных компонент или факторов.

Методы многомерной классификации предназначены для разделения совокупностей объектов, характеризующихся большим числом признаков, на классы. В каждый класс должны входить объекты, в определенном смысле однородные или близкие. Такую классификацию на основании статистических данных о значениях признаков на объектах можно провести методами кластерного и дискриминантного анализа.

2.2. Математическое ожидание и дисперсия

Математическое ожидание дискретных случайных величин x и y определяется по формулам:

- для дискретных случайных величин:

$$m_x = M(x) = \sum_{i=1}^{m} \sum_{j=1}^{n} x_i \cdot P_{ij},$$

$$m_y = M(y) = \sum_{i=1}^{m} \sum_{j=1}^{n} y_i \cdot P_{ij};$$

- для непрерывных случайных величин:

$$m_x = M(x) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xf(x,y)\,dxdy;$$

$$m_y = M(y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} yf(x,y)\,dxdy.$$

Точка с координатами (m_x, m_y) называется центром рассеивания системы случайных величин (x, y).

Если случайные величины (x, y) независимы, в этом случае из законов распределения этих случайных величин можно определить математическое ожидание по формулам для одномерных величин:

$$m_x = \sum_{i=1}^{m} x_i \cdot P_i;$$

$$m_y = \sum_{j=1}^{n} y_i \cdot P_j.$$

Дисперсия дискретных случайных величин определяется по следующим формулам:

- для дискретных случайных величин:

$$D(x) = \sum_{i=1}^{m}\sum_{j=1}^{n} P_{ij}(x_i - m_x)^2,$$

$$D(y) = \sum_{i=1}^{m}\sum_{j=1}^{n} P_{ij}(y_i - m_y)^2;$$

- для непрерывных случайных величин:

$$D(x) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x - m_x)^2 f(x,y)\,dxdy,$$

$$D(y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (y - m_y)^2 f(x,y)\,dxdy.$$

Среднеквадратичное отклонение случайных величин (x, y) определяется по формулам:

$$\sigma_x = \sqrt{D(x)};$$

$$\sigma_y = \sqrt{D(y)}.$$

Важнейшую роль в теории системы случайных величин играет так называемый корреляционный момент, или ковариация:

$$Cov = M\left[\left(x - m_x\right)\left(y - m_y\right)\right];$$

$$Cov_{xy} = \sum_m \sum_n \left(x_n - m_x\right)\left(y_m - m_y\right)P_{mn};$$

$$Cov_{xy} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(x - m_x\right)\left(y - m_y\right) f\left(x, y\right)dxdy.$$

Случайные величины $\left(x, y\right)$ называются независимыми, если вероятность одной из них имеет значение, лежащее в любом промежутке области ее значений, и не зависит от того, какое значение приняла другая величина. В этом случае значение коэффициента корреляции $Cov_{xy} = 0$.

3. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

В том случает, когда отсутствует достаточный объем информации о моделируемом объекте, уравнения математического описания могут быть представлены в виде системы эмпирических зависимостей, полученных в результате статистического анализа объекта и имеющих вид регрессионных соотношений между входными и выходными параметрами объекта. В этом случае в структуре уравнений статистических моделей не отражаются физические свойства объекта моделирования. Эксперимент является основным источником информации, а обработка экспериментальных данных осуществляется методами теории вероятностей и математической статистики. При этом объект представляется в виде «черного ящика» (рис. 3.1).

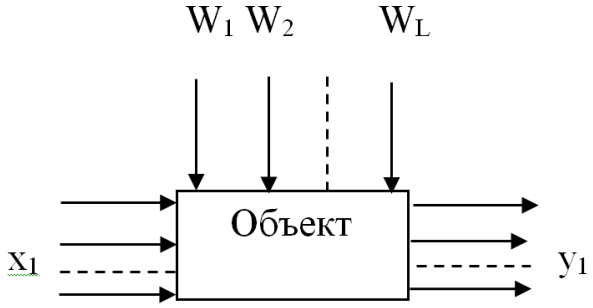


Рис. 3.1. Схематическое изображение объекта:
x_i – входные параметры; y_i – выходные параметры;
W_i – случайные воздействия, «шумы»

Математической моделью служит функция отклика, связывающая выходной параметр с входными:

$$Y = F(x_1, x_2, ..., x_n),$$

или в виде полинома

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{j=1}^{k} \beta_{uj} x_u x_j + \sum_{j=1}^{k} \beta_{jj} x_j^2 + ...$$

Поскольку в реальном процессе всегда существуют случайные воздействия, или «шумы», в этом случае изменение величины *y* носит случайный характер. Поэтому при обработке экспериментальных данных получаются так называемые выборочные коэффициенты регрессии *b*, являющиеся оценками теоретических коэффициентов β. Уравнение регрессии, полученное на основании опыта, запишется следующим образом:

$$Y = b_0 + \sum_{j=1}^{k} b_j x_j + \sum_{j=1}^{k} b_{uj} x_u x_j + \sum_{j=1}^{k} b_{jj} x_j^2 + ...$$

Вид уравнения регрессии обычно задается.

Для получения статистических моделей в виде полиномов на основе данных, собранных в пассивном эксперименте, используют методы корреляционного и регрессионного анализа.

3.1. Методы корреляционного и регрессионного анализа

Методы корреляционного и регрессионного анализа широко применяются для выявления и описания зависимостей между случайными величинами по экспериментальным данным и базируются на теории вероятности и математической статистике.

Корреляционный анализ основывается на предпосылке о том, что переменные величины y (выходной параметр) и x_i (факторы) являются случайными величинами и между ними может существовать так называемая корреляционная связь, при которой с изменением одной величины изменяется распределение другой. Для количественной оценки тесноты связи служит выборочный коэффициент корреляции.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{(n-1)\cdot S_x \cdot S_y},$$

где $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i$; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$; S_x^2, S_y^2 – выборочные дисперсии:

$$S_x^2 = \frac{\sum(x_i - \bar{x}_i)^2}{N-1};$$

$$S_y^2 = \frac{\sum(y_i - \bar{y}_i)^2}{N-1}.$$

При вычислении коэффициента корреляции удобно пользоваться следующими формулами:

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{N};$$

$$(N-1)S_x^2 = \sum x_i^2 - \frac{1}{N}(\sum x_i)^2;$$

$$S_x^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N-1};$$

$$(N-1)S_y^2 = \sum y_i^2 - \frac{1}{N}(\sum y_i)^2;$$

$$S_y^2 = \frac{\sum y_i^2 - \frac{(\sum y_i)^2}{N}}{N-1}.$$

где N – число опытов.

Выявить наличие или отсутствие корреляции между двумя величинами можно путем визуального анализа полей корреляции и оценкой величины выборочного коэффициента корреляции.

На рис. 3.2 показаны примеры корреляции между случайными величинами.

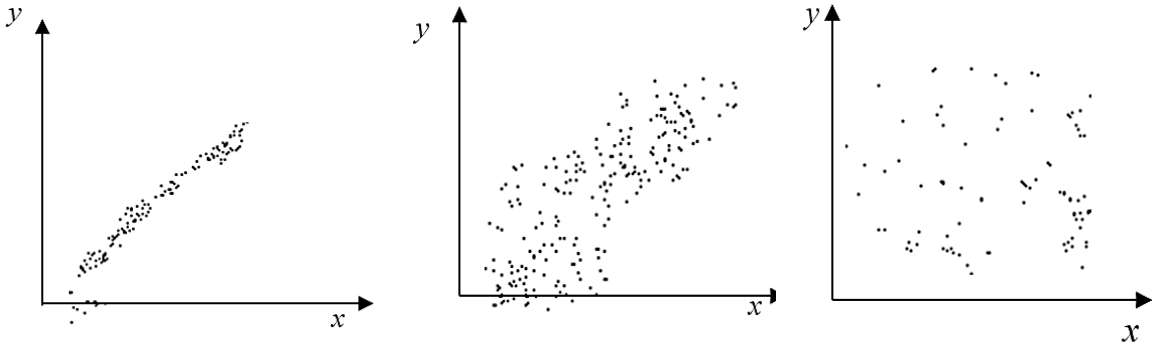


Рис. 3.2. Виды корреляции между случайными величинами

Для независимых случайных величин коэффициент корреляции равен нулю, но он может быть равен нулю для некоторых зависимых величин, которые при этом называются некоррелированными. Коэффициент корреляции характеризует не всякую зависимость, а только линейную. Если случайные величины x и y связаны точной функциональной линейной зависимостью

$$y = b_0 + b_1 x,$$
$$r_{xy} = \pm 1.$$

В том случае, когда случайные величины связаны произвольной стохастической зависимостью, коэффициент корреляции может принимать значение в пределах $-1 < r_{xy} < 1$.

Регрессионный анализ рассматривает и предполагает связь между зависимой или случайной величиной y и независимыми или неслучайными переменными $x_1, \ldots, x_i$.

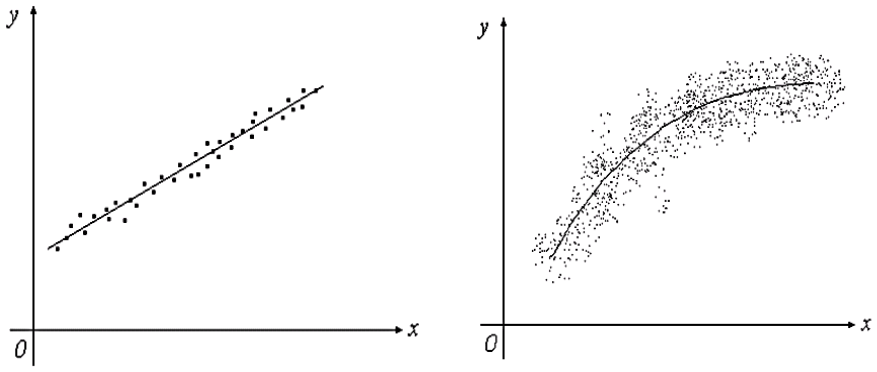
Такая связь представляется с помощью математической модели, представляющей собой уравнение или систему уравнений, которые связывают зависимую и независимую переменные.

Использование корреляционного и регрессионного анализа при обработке экспериментальных данных дает возможность построить статистическую математическую модель в виде уравнения регрессии.

Постановка задачи.

По данной выборке объема n найти уравнение приближенной регрессии и оценить допускаемую при этом ошибку, то есть нужно найти $\hat{y} = f(x)$. Эта задача решается методами корреляционного и регрессионного анализа.

По сгущениям точек (рис. 3.3) можно найти определенную зависимость, т. е. получить вид уравнения регрессии.



а *б*

Рис. 3.3. Виды регрессии: а – линейная; б – нелинейная

Если разброс точек значительный, то регрессии не будет. Следовательно, методы корреляционного и регрессионного анализа тесно связаны между собой.

Вид уравнения регрессии зависит от выбираемого метода приближения. Обычно используется метод наименьших квадратов.

$$F = \sum_{i=1}^{n} \left[y_i - f(x_i) \right]^2 = \min$$

или

$$F = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min,$$

где $y_i, \hat{y}_i$ – экспериментальные и расчетные значения выходного параметра соответственно.

Корреляционный анализ является одним из методов статистического анализа взаимосвязей нескольких признаков. Он применяется тогда, когда данные наблюдений можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону.

Основная задача корреляционного анализа состоит в оценке корреляционной матрицы генеральной совокупности по выборке и определении на ее основе оценок частных и множественных коэффициентов корреляции.

Парный, или частный, коэффициент корреляции характеризует тесноту линейной зависимости между двумя переменными на фоне действия всех остальных показателей, входящих в модель.

Множественный коэффициент корреляции характеризует тесноту линейной связи между одной переменной и остальными, входящими в модель. Изменяется в пределах от 0 до 1.

Квадрат множественного коэффициента корреляции называется множественным коэффициентом детерминации. Он характеризует долю дисперсии одной переменной, обусловленной влиянием всех остальных входящих в модель.

Исходной для анализа является матрица:

$$X = \begin{Bmatrix} x_{11} \ldots x_{1j} \ldots x_{1k} \\ x_{i1} \ldots . x_{ij} \ldots . x_{ik} \\ x_{n1} \ldots x_{nj} \ldots x_{nk} \end{Bmatrix}$$

В корреляционном анализе матрицу X рассматривают как выборку объема n из k-мерной генеральной совокупности, подчиняющейся k-мерному нормальному закону распределения.

В многомерном корреляционном анализе рассматривают две задачи:

1. Определение тесноты связи одной из переменных с совокупностью остальных ($k - 1$) переменных, включенных в анализ.

2. Определение тесноты связи между переменными при фиксировании или исключении влияния остальных переменных.

По выборке определяют оценки параметров генеральной совокупности, а именно:

- вектор средних значений $\vec{X}$;
- вектор среднеквадратичных отклонений S;
- корреляционную матрицу R порядка k – $R(k)$;

$$\vec{X} = \begin{Bmatrix} \vec{X}_1 \\ \vec{X}_2 \\ \vdots \\ \vec{X}_k \end{Bmatrix}; \qquad S = \begin{Bmatrix} S_1 \\ S_2 \\ \vdots \\ S_k \end{Bmatrix}; \qquad R = \begin{Bmatrix} 1 & r_{12} \ldots r_{1k} \\ r_{21} & 1 \ldots r_{2k} \\ \vdots & \\ r_{k1} & r_{k2} \ldots 1 \end{Bmatrix}.$$

Матрица составлена из парных выборочных коэффициентов корреляции, определенных по следующим формулам:

$$r_{jl} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_{ij} - \vec{X}_j\right)\left(X_{il} - \vec{X}_l\right)}{S_j \cdot S_l}; \quad \vec{X} = \frac{1}{n}\sum X_{ij};$$

$$S_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{ij} - \vec{X}_j\right)^2},$$

где X_{ij} – значение i-го наблюдения j-го фактора; r_{jl} – выборочный парный коэффициент корреляции, который характеризует тесноту линейной связи между показателем X_j и X_l.

Выборочный коэффициент корреляции $R_{j,1\ldots k}$ может быть вычислен по формуле

$$R_{j,1\ldots k} = \sqrt{1 - \frac{|R_k|}{R_{jj}}};$$

где $|R_k|$ – определитель матрицы R; R_{jj} – алгебраическое дополнение элемента r_{ii} той же матрицы, равное 1.

Соответственно, теснота линейных взаимосвязей одной переменной X_j с совокупностью других $(k-1)$ переменных измеряется с помощью множественного коэффициента корреляции.

Кроме того, находятся точные оценки частных и множественных коэффициентов корреляции любого порядка. Например, частный коэффициент корреляции $(k-2)$-го порядка между факторами X_1 и X_2 определяется по формуле:

$$r_{12/3,4\ldots k} = \frac{-R_{12}}{\sqrt{R_{11}R_{22}}},$$

а множественный коэффициент корреляции $(k-1)$-го порядка фактора X_1 определяется по формуле

$$r_{12/3\ldots k} = r_1 = \sqrt{1 - \frac{|R|}{R_{11}}},$$

где $|R|$ – определитель матрицы R.

С помощью множественного коэффициента корреляции R по мере приближения R к 1 делается вывод о тесноте взаимосвязи, но не о ее направлении.

Величина R^2 называется выборочным множественным коэффициентом детерминации и показывает, какую долю вариаций исследуемой переменной объясняет вариация остальных переменных. Можно показать, что множественный коэффициент корреляции значим от 0, если величина:

$$F = \frac{R^2(n-k)}{(1-R^2)(k-1)} \geq F_{\text{табл}}(q, f_1, f_2);$$

$$f_1 = k-1;$$

$$f_2 = n-k,$$

где $F_{\text{табл}}$ – табличное значение F-критерия на уровне значимости q при числе степеней свободы f_1 и f_2.

Если переменные коррелируют друг с другом, то на величине парного коэффициента корреляции частично сказывается влияние других

переменных. В связи с этим часто возникает необходимость исследовать часть корреляции между переменными при исключении влияния одной или нескольких других переменных.

Так, выборочным частным коэффициентом корреляции между переменными X_j и X_l при фиксированных значениях остальных $(k - 2)$ переменных называется выражение

$$r_{jl,1,2...k} = \frac{-r_{jl}}{\sqrt{r_{jj} r_{ll}}}.$$

Таблица 1

Сила тесноты связи

Тип связи	Значение
Сильная (тесная)	$r > 0{,}70$
Средняя	$0{,}50 < r < 0{,}69$
Умеренная	$0{,}30 < r < 0{,}49$
Слабая	$0{,}20 < r < 0{,}29$
Очень слабая	$r < 0{,}19$

Технологические объекты, как правило, характеризуются большим числом параметров, и особое значение приобретают задачи изучения взаимосвязей между данными параметрами. Статистический анализ позволяет обработать статистические данные, систематизировать их и обработать с целью выявления характера и структуры взаимосвязей для получения практических выводов.

Пример. Используя статистические методы обработки экспериментальных данных, получить вид функциональной зависимости физико-химических свойств нефти и проверить соответствие полученного уравнения регрессии эксперименту и ошибку аппроксимации.

Таблица 2

Влияние расхода сырой эмульсии на величину обводненности нефти для деэмульгатора СНПХ 4502 (dmp = 0,265 м; W = 28 %)

Концентрация химического реагента, г/т	Обводненность нефти, мас. %			
	G = 7,843e8 кг/год		*G* = 1,229e9 кг/год	
	1-я ступень	2-я ступень	1-я ступень	2-я ступень
10	3,5945	0,1897	4,0365	0,2186
20	1,0790	0,0196	1,1845	0,0217
30	0,6877	0,0081	0,7530	0,0090
40	0,5276	0,0048	0,5770	0,0053
50	0,4393	0,0034	0,4802	0,0037
60	0,3828	0,0026	0,4182	0,0028

На основании экспериментальных данных об обводненности нефти в зависимости от расхода эмульсии рассчитывают коэффициент парной корреляции (табл. 3).

Таблица 3

Значение парного коэффициента корреляции

	Обводненность нефти, мас. %			
	G = 7,843e8 кг/год		G = 1,229e9 кг/год	
	1-я ступень	2-я ступень	1-я ступень	2-я ступень
r_{xy}	−0,7828	−0,7078	−0,7800	−0,7061

Из полученных значений коэффициентов корреляции между обводненностью нефти и концентрацией химического реагента – деэмульгатора можно сделать вывод, что связь между параметрами не является линейной, однако имеет сильную тесноту связи, т. к. $R_{xy} > 0{,}70$. Это говорит о том, что при описании экспериментальных данных потребуется более сложные зависимости. Отрицательные значения коэффициентов корреляции указывают на обратную зависимость между параметрами, т. е. с увеличением концентрации деэмульгатора обводненность нефти снижается.

Далее проводим обработку экспериментальных данных в Excel с целью получения теоретической зависимости, наилучшим образом описывающей экспериментальные данные (рис. 3.4−3.7).

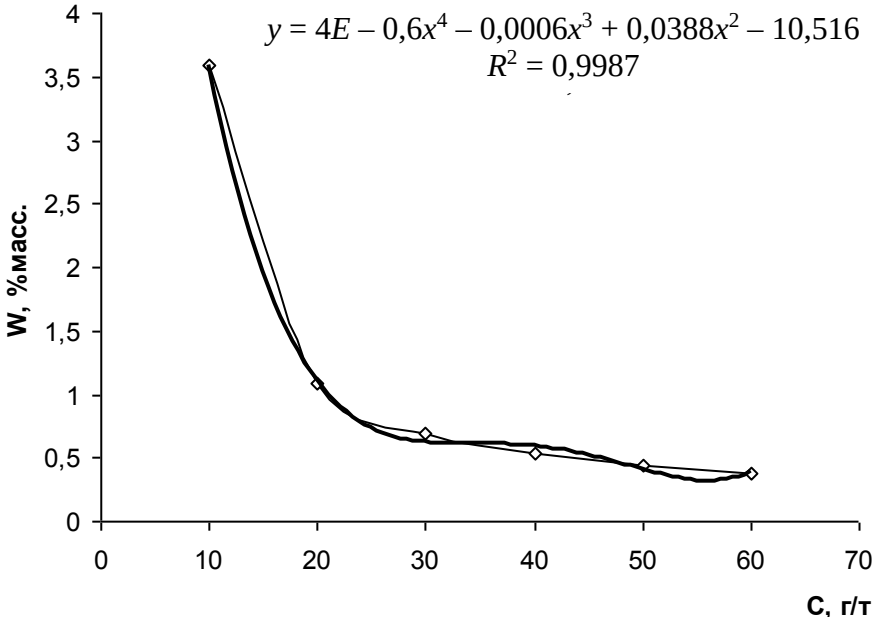


Рис. 3.4. Зависимость обводненности для расхода G = 7,843e8 кг/год на первой ступени обезвоживания

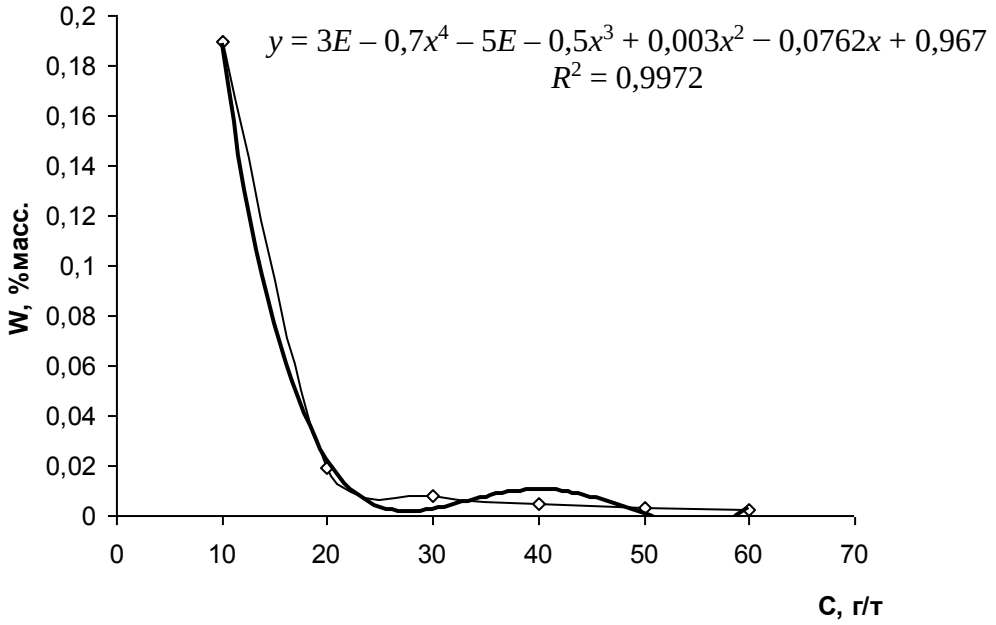


Рис. 3.5. Зависимость обводненности для расхода G = 7,843e8 кг/год на второй ступени обезвоживания

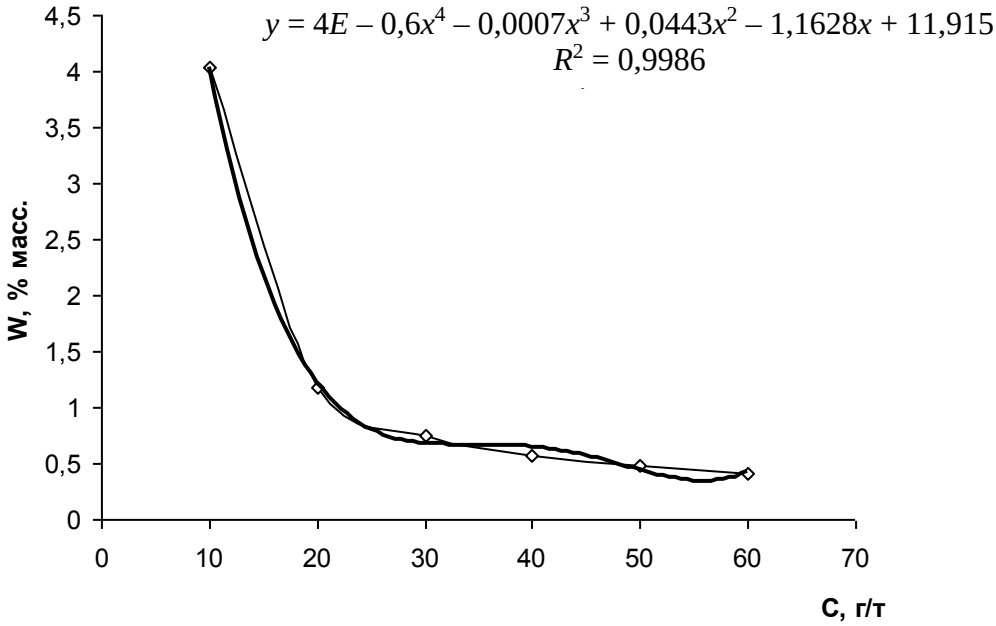


Рис. 3.6. Зависимость обводненности для расхода G = 1,229e9 кг/год на первой ступени обезвоживания

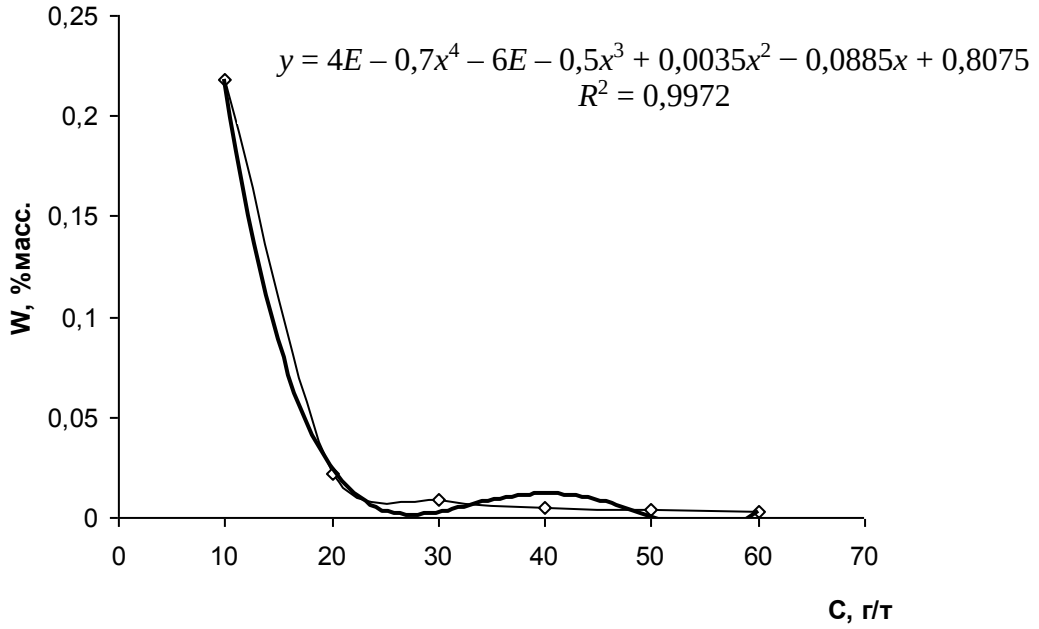


Рис. 3.7. Зависимость обводненности для расхода G = 1,229e9 кг/год на второй ступени обезвоживания

В результате обработки данных в программе Excel получили, что наилучшим образом описывает экспериментальные данные об обводненности нефти полиномиальная зависимость четвертой степени, т. к. при описании данной зависимостью доверительная вероятность максимальна. Для первых ступеней обводненности нефти доверительная вероятность при расходах G = 7,843e8 кг/год и G = 1,229e9 кг/год равна R^2 = 0,998, для вторых – R^2 = 0,997.

3.2. Статистический анализ уравнения регрессии

Регрессионный анализ состоит из трех основных этапов:

1. Оценка дисперсии воспроизводимости (оценка ошибки опыта):

а) определяется среднее по результатам опыта

$$\overline{y}_i = \frac{\sum_{u=1}^{m} y_{iu}}{m}; \qquad i = 1, ..., N;$$

где m – число параллельных опытов.

б) выборочные (построчные) дисперсии

$$S_i^2 = \frac{\sum_{u=1}^{m} \left(y_{iu} - \overline{y}_i \right)^2}{m-1};$$

в) $\sum_{i=1}^{N} S_i^{\,2}$;

г) $G = \dfrac{S^2_{\max}}{\sum S^{\,2}_{\,i}}$;

д) $S^2_{\max}$ − максимальное значение выборочной дисперсии.

Если $G < G_{\text{табл}}(q, f_1, f_2)$, при $f_1 = m - 1$; и $f_2 = N$, то дисперсия однородна;

е) рассчитывается дисперсия воспроизводимости

$$S^2_{\text{воспр}} = \frac{\sum S_i^2}{N}.$$

2. Оценка значимости коэффициентов проводится по критерию Стьюдента.

$|b_i|$ − абсолютное значение коэффициента регрессии.

S_{b_i} − среднеквадратичное отклонение i-го коэффициента.

$$S^2_{b_i} = \frac{S^2_{\text{воспр}}}{N}.$$

Если $t_{b_i} > t_{\text{табл}}(q, f)$, $f = N(m-1)$, то коэффициент значим. Если нет, то коэффициент приравнивается к 0 и из уравнения исключается.

3. Проверка модели на адекватность по критерию Фишера:

$$F = \frac{S^2_{\text{ост}}}{S^2_{\text{воспр}}};$$

$$S^2_{\text{ост}} = \frac{\sum \left(\bar{y}_i^{\,э} - \hat{y}_i \right)^2}{N - l}.$$

$$l = n + 1.$$

Если $F < F_{\text{табл}}(q, f_1, f_2)$, то линейное уравнение регрессии адекватно описывает процесс.

$$f_1 = N - 1; \; f_2 = N(m - 1).$$

Пример. В химическом процессе выход продукта реакции Y зависит от температуры x_1 и концентрации реагента x_2. Требуется найти математическое описание этого процесса.

Полученное уравнение регрессии вида

$$\hat{Y} = 5{,}95 + 0{,}82X_1 + 2{,}29X_2 + 0{,}9X_3 - 3{,}97X_4.$$

Выполнить регрессионный анализ в соответствии с приведенной выше последовательностью формул.

1. Критерий Кохрена используется для сравнения нескольких дисперсий. Дисперсия однородна (гипотеза принимается), когда выборочная дисперсия получена по одинаковым выборкам, т. е. во всех сериях экспериментов число параллельных опытов было одинаково.

Рассчитываем отношение $G = 0{,}029/0{,}035 = 0{,}824$.

Табличное значение критерия Кохрена при $q = 0{,}05$ при $f_1 = 1$; $f_2 = 8$ равно $G_m = 0{,}9798$.

$0{,}824 > 0{,}9798$, следовательно, дисперсия однородна (в противном случае необходимо было бы увеличить число параллельных опытов).

2. Рассчитываем значение дисперсии воспроизводимости:

$$S^2_{\text{воспр}} = 0{,}035/8 = 0{,}0044.$$

Критерий Стьюдента – предпосылка об отсутствии корреляции между факторами. Применяется при определении необходимого числа экспериментов для достижения заданной степени точности, для определения грубых ошибок, для сравнения между собой двух средних полученных по выборкам. Если $t > t_m$, то коэффициент значим, в противном случае – незначим и исключается из уравнения.

Для оценки значимости коэффициентов регрессии вычислим ошибку в определении коэффициентов:

$$S_b = \sqrt{S^2_{\text{воспр}}} / N = 0{,}0234;$$

$$t_0 = |b_0|/S_b = 254{,}48;$$

$$t_1 = |b1|/S_b = 34{,}96;$$

$$t_2 = |b2|/S_b = 97{,}910;$$

$$t_3 = |b3|/S_b = 38{,}36;$$

$$t_4 = |b4|/S_b = 169{,}86.$$

Табличное значение t-критерия для $q = 0{,}05$ и $f = N \cdot (m - 1)$:

$$t_m = 2{,}78.$$

Таким образом, все значения t_i больше табличного, а следовательно, все коэффициенты регрессии значимы (отличаются от 0). Следовательно, искомое уравнение имеет вид

$$\hat{Y} = 5{,}95 + 0{,}82X_1 + 2{,}29X_2 + 0{,}9X_3 - 3{,}97X_4.$$

3. Критерий Фишера применяется при проверке однородности двух дисперсий или нескольких. При этом проверяется гипотеза, можно ли считать сравниваемые дисперсии оценками одной и той же генеральной совокупности. Для проверки адекватности уравнения найдем расчетные значения функции отклика ($\hat{Y}$): $\hat{Y}_1 = 5{,}92$; $\hat{Y}_2 = 12{,}13$; $\hat{Y}_3 = 9{,}34$; $\hat{Y}_4 = 12{,}29$; $\hat{Y}_5 = -0{,}39$; $\hat{Y}_6 = 2{,}55$; $\hat{Y}_7 = -0{,}39$; $\hat{Y}_8 = 5{,}98$.

Вычисляем $S^2_{ост}$:

$$S^2_{ост} = 5{,}58.$$

Находим расчетное значение критерия Фишере по формуле

$$F = \frac{S^2_{ост}}{S^2_{воспр}},$$

где $F = 1{,}27$.

Значение критерия оптимальности, приведенное в таблице, при $q = 0{,}05$; $f_1 = 1$; $f_2 = 8$ равно $F_m = 5{,}32$.

Таким образом, $1{,}27 < 5{,}32$.

Следовательно, полученное уравнение регрессии адекватно и может быть применено для описания зависимости выхода продукта от темпера туры и концентрации реагента, уравнение является нелинейным.

Таким образом, получили уравнение регрессии следующего вида:

$$\hat{Y} = 5{,}95 + 0{,}82X_1 + 2{,}29X_2 + 0{,}9X_3 - 3{,}97X_4.$$

4. МЕТОДЫ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ. КЛАСТЕРНЫЙ АНАЛИЗ

В статистических исследованиях группировка первичных данных является базовым приемом решения задачи классификации, а поэтому и основой всей дальнейшей работы с собранной информацией.

Традиционно эта задача решается следующим образом. Из множества признаков, описывающих объект, отбирается один наиболее значимый признак, и производится группировка в соответствии со значениями данного признака. Если же требуется провести классификацию по нескольким признакам, ранжированным между собой по степени важности, то сначала производится классификация по первому признаку, затем каждый из полученных классов разбивается на подклассы по второму признаку, далее каждый из полученных классов разбивается на подклассы по третьему признаку и т. д.

В таких случаях, когда не представляется возможным упорядочить классификационные признаки, используется создание интегрального показателя (индекса), функционально зависящего от исходных признаков, с последующей классификацией по этому показателю. Данный метод является наиболее простым методом многомерной группировки.

Развитием этого подхода является вариант классификации по нескольким обобщающим показателям (главным компонентам), полученным с помощью методов факторного или компонентного анализа.

При наличии нескольких признаков (исходных или обобщенных) задача классификации может быть решена методами кластерного анализа.

Метод кластерного анализа отличается от других методов многомерной классификации отсутствием априорной информации о распределении генеральной совокупности.

Рассмотрим следующую задачу. Пусть исследуется совокупность N объектов, каждый из которых характеризуется по k-замеренным на нем признакам X. Требуется разбить эту совокупность на однородные в некотором смысле группы (классы).

При этом практически отсутствует априорная информация о характере распределения измерений X внутри классов.

Полученные в результате разбиения группы обычно называются кластерами, а методы их нахождения – **кластерным анализом**.

Кластер – группа элементов, которые характеризуются каким-либо общим свойством.

Цели кластерного анализа: систематизация множества исследуемых объектов или признаков с выделением однородных групп (кластеров). Он дает инструмент для классификации данных и выявлении в них соответствующей структуры.

Кластерный анализ может применяться в разных случаях, даже тогда, когда речь идет о простой группировке, в которой все сводится к образованию групп по признаку количественного сходства.

Кластерный анализ включает в себя набор различных алгоритмов классификации. При этом кластерный анализ не позволяет делать статистические выводы, но дает возможность изучить структуру совокупности.

Кластерный анализ позволяет рассмотреть большой, хотя имеющий ограничения, объем информации, делать его компактным и наглядным.

Кластерный анализ позволяет произвести разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того он не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет рассмотреть множество практически произвольно взятых произвольных данных.

Большинство методов кластерного анализа основаны на анализе квадратной и симметричной относительно главной диагонали матрицы коэффициентов сходства, сопряженности корреляции и т. д.

При определении корреляции между признаками сравнивается распределение двух каких-либо видов в определенной серии наблюдений и оценивается, насколько тесно совпадают эти распределения.

Наиболее часто используется иерархическая классификация, которая может быть представлена в двух основных формах: «дерево» и «вложенное множество».

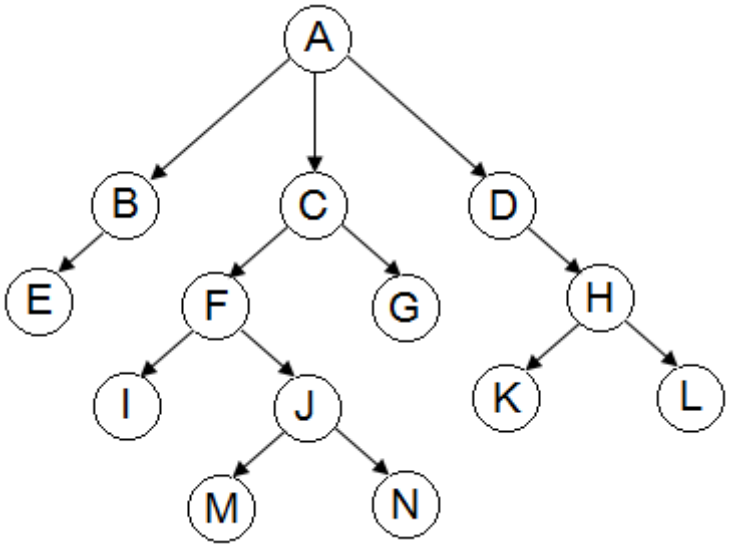


Рис. 4.1. Дерево

Дерево – специальный вид направления графа, т. е. структуры, состоящей из узлов, связанных дугами.

Дерево иерархической классификации обладает следующими свойствами:

1. Имеет только один корень или вершину.

2. Всегда имеется путь от корня до любого другого узла в дереве.

3. Каждый узел, кроме корня, имеет только одного родителя, т. е. граф не должен иметь циклов и петлей и произвольное число потоков.

4. Узлы дерева, которые не имеют потоков, называют листьями, и они соответствуют количеству классифицируемых объектов.

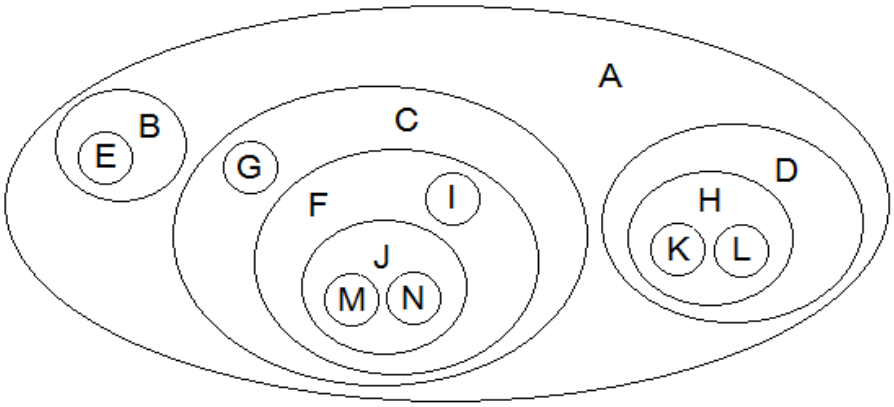


Рис. 4.2. Вложенное множество

Обычной формой представления исходных данных в задачах кластерного анализа служит прямоугольная матрица:

$$X = \begin{bmatrix} x_{11} \ldots x_{1j} \ldots x_{1k} \\ x_{i1} \ldots x_{ij} \ldots x_{ik} \\ x_{n1} \ldots x_{nj} \ldots x_{nk} \end{bmatrix}$$

Каждая строка матрицы представляет результат измерений k рассматриваемых признаков на одном из обследованных объектов.

Интерес может представлять как группировка объектов, так и группировка признаков.

Матрица X не является единственным способом представления данных в задачах кластерного анализа. Иногда исходная информация задана в виде квадратичной матрицы:

$$R = \left(r_{ij} \right), \;\; i, j = 1 \ldots n,$$

где r_{ij} – определяет степень близости i-го объекта к j-му.

В результате того, что большинство алгоритмов кластерного анализа полностью исходят из матрицы расстояний или близостей либо требуют вычисления отдельных ее элементов, и при этом данные представ-

лены в форме матрицы X, то первым этапом решения задачи поиска кластеров является выбор способа вычисления расстояний между объектами или признаками.

4.1. Расстояние между объектами (кластерами) и мера близости

Наиболее трудным и наименее формализованным в задаче классификации является определение понятия однородности объектов.

В общем случае понятие однородности объектов задается либо правилом вычисления расстояний $\rho(x_i, x_j)$ между любой парой исследуемых объектов $(x_1, x_2, ..., x_n)$, либо заданием некоторой функции $r(x_i, x_j)$, характеризующей степень близости i-го и j-го объектов.

Выбор метрики (или меры близости) является узловым моментом исследования, от которого в основном зависит окончательный вариант разбиения объектов на классы при данном алгоритме разбиения. В каждом конкретном случае этот выбор должен производиться по-своему в зависимости от целей исследования, физической и статистической природы вектора наблюдений X, априорных сведений о характере вероятностного распределения X.

Рассмотрим наиболее широко используемые в задачах кластерного анализа расстояния и меры близости.

4.1.1. Обычное евклидово расстояние

$$\rho_E(x_i, x_j) = \sqrt{\sum_{l=1}^{k} (x_{il} - x_{jl})^2};$$

$$l = 1, 2, ..., k;$$

$$i, j = 1, 2, ..., n,$$

где x_{il}, x_{jl} – величины l-й компоненты у i-го, j-го объектов соответственно.

Использование этого расстояния оправданно в следующих случаях:

1) наблюдения берутся из генеральной совокупности, имеющей многомерное нормальное распределение компоненты X, взаимно независимы и имеют одну и ту же дисперсию;

2) компоненты вектора наблюдений X одинаково важны для классификации;

3) признаковое пространство совпадает с геометрическим пространством.

4.1.2. Хеммингово расстояние

Хеммингово расстояние используется как мера различия объектов, задаваемых дихотомическими признаками. Хеммингово расстояние определяется по формуле

$$\rho_H\left(x_i, x_j\right) = \sum_{l=1}^{k} \left| x_{il} - x_{jl} \right|$$

и равно числу несовпадений значений соответствующих признаков в рассматриваемых *i*-м и *j*-м объектах.

Решение задач классификации многомерных данных предусматривает в качестве предварительного этапа исследования реализацию таких методов, которые позволяют выбрать из компонентов $(x_1, x_2, ..., x_n)$ сравнительно небольшое число наиболее информативных компонентов, т. е. уменьшить размерность наблюдаемого пространства.

В ряде процедур классификации используют понятия расстояния между группами объектов и меры близости двух групп объектов.

Пусть S_i – *i*-я группа (класс, кластер), состоящая из *n* объектов; $\overline{X}_i$ – среднее арифметическое векторных наблюдений S_i группы, т. е. «центр тяжести» *i*-й группы; $\rho(S_l, S_m)$ – расстояние между группами S_l и S_m.

Наиболее употребительными расстояниями и мерами близости между классами объектов являются:

1. Расстояние, измеряемое по принципу «ближайшего соседа»:

$$\rho_{\min}\left(S_l, S_m\right) = \min \rho\left(x_i, x_j\right);$$
$$x_i \in S_l;$$
$$x_j \in S_m.$$

2. Расстояние, измеряемое по принципу «дальнего соседа»:

$$\rho_{\max}\left(S_l, S_m\right) = \max \rho\left(x_i, x_j\right);$$
$$x_i \in S_l;$$
$$x_j \in S_m.$$

3. Расстояние, измеряемое по «центрам тяжести» групп:

$$\rho_{ц.т}\left(S_l, S_m\right) = \rho\left(\overline{X}_l, \overline{X}_m\right).$$

4. Расстояние, измеряемое по принципу «средней связи», определяется как среднее арифметическое всех попарных расстояний между представителями рассматриваемых групп:

$$\rho_{\text{ср}}\left(S_l, S_m\right) = \frac{1}{n_l n_m} \sum_{x_i \in S_l} \sum_{x_j \in S_m} \rho\left(x_i, x_j\right).$$

Существует большое количество различных способов разбиения заданной совокупности объектов на классы.

4.2. Иерархические кластер-процедуры

Иерархические, или древообразные, процедуры являются наиболее распространенными алгоритмами кластерного анализа. Они бывают двух типов: агломеративные и дивизимные. В агломератнвных процедурах начальным является разбиение, состоящее из n одноэлементных классов, а конечным − из одного класса; в дивизимных – наоборот.

Принцип работы иерархических агломератнвных процедур состоит в последовательном объединении групп элементов сначала самых близких, а затем все более отдаленных друг от друга.

Принцип работы иерархических дивизимных процедур состоит в последовательном разделении групп элементов сначала самых далеких, а затем все более близких друг от друга. Большинство этих алгоритмов исходит из матрицы расстояний.

Недостатком иерархических процедур является громоздкость их вычислительной реализации.

В качестве примера рассмотрим агломеративный иерархический алгоритм. На первом шаге алгоритма каждое наблюдение X_i рассматривается как отдельный кластер. В дальнейшем на каждом шаге работы алгоритма происходит объединение двух самых близких кластеров с учетом принятого расстояния. Работа алгоритма заканчивается, когда все наблюдения объединены в один класс.

Пример. Требуется провести классификацию $n = 6$ объектов, каждый из которых характеризуется двумя признаками:

N	1	2	3	4	5	6
X_{t1}	5	6	5	10	11	10
X_{t2}	10	12	13	9	9	7

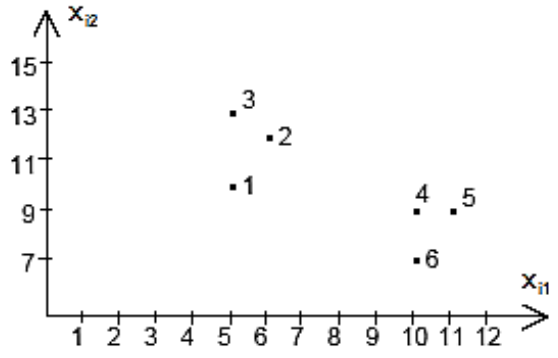


Рис. 4.3. Расположение объектов на плоскости

Воспользуемся алгомеративным иерархическим алгоритмом классификации. В качестве расстояния между объектами примем обычное евклидово расстояние. Тогда расстояние между объектами 1 и 2 равно:

$$\rho_{12} = \sqrt{(5-6)^2 + (10-12)^2} = 2,24;$$

$$\rho_{13} = \sqrt{(5-5)^2 + (10-13)^2} = 3;$$

$$\rho_{11} = 0.$$

Аналогично находим расстояния между всеми шестью объектами и строим матрицу расстояний:

$$R = \left\{\rho\left(x_i, x_j\right)\right\} = \begin{vmatrix} 0 & 2,24 & 3 & 5,1 & 6,08 & 5,83 \\ 2,24 & 0 & 1,14 & 5 & 5,83 & 6,4 \\ 3 & 1,41 & 0 & 6,4 & 7,21 & 7,81 \\ 5,1 & 5 & 6,4 & 0 & 1 & 2 \\ 6,08 & 5,83 & 7,21 & 1 & 0 & 2,24 \\ 5,83 & 6,4 & 7,81 & 2 & 2,24 & 0 \end{vmatrix}.$$

Объединение в кластеры будем производить по принципу «ближайшего соседа».

Из матрицы расстояний следует, что объекты 4 и 5 наиболее близки, а расстояние $d_{4,5}$ = 1,00, поэтому их объединяют в один кластер.

После объединения имеем пять кластеров:

Номер кластера	1	2	3	4	5
Состав кластера	(1)	(2)	(3)	(4,5)	(6)

Вновь находят расстояние между кластерами и производят дальнейшее объединение по принципу «ближайшего соседа».

Каждый раз составляем матрицу расстояний до тех пор, пока все наблюдения не будут объединены в один кластер. Результаты иерархической классификации представляют в виде дендрограммы.

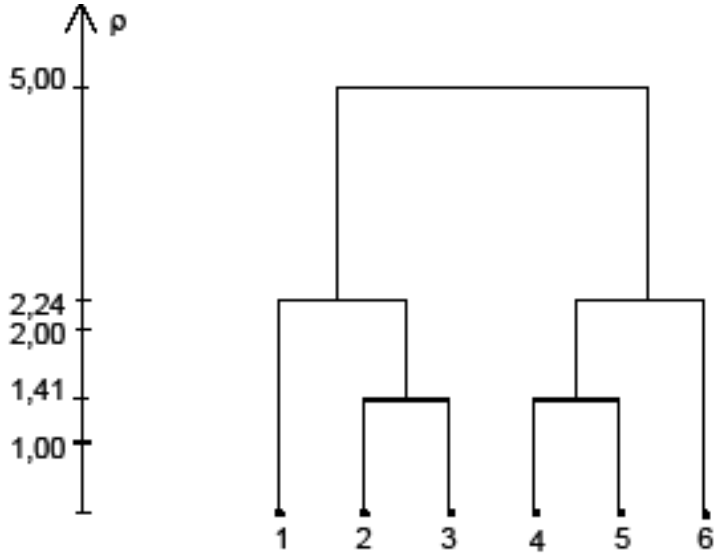


Рис. 4.4. Дендрограмма

5. СТАТИСТИЧЕСКИЕ МЕТОДЫ ОПТИМИЗАЦИИ

Оптимизация как раздел математики существует достаточно давно. Оптимизацией называют целенаправленную деятельность людей, которая заключается в получении наилучших результатов при соответствующих условиях. Другими словами, задачей оптимизации является нахождение оптимума рассматриваемой функции или оптимальных технологических условий проведения технологического процесса. Термином «оптимизация» в литературе обозначают процесс или последовательность операций, позволяющих получить уточненное решение. Оптимизация в широком смысле слова находит применение в науке, технике и в любой другой области человеческой деятельности.

В процессе проектирования обычно ставится задача определения наилучших, в некотором смысле, структуры или значений параметров объектов. Такая задача называется оптимизационной. Хотя конечной целью оптимизации является отыскание наилучшего, или оптимального, решения, обычно приходится довольствоваться улучшением известных решений, а не доведением их до совершенства. Поэтому под оптимизацией понимают скорее стремление к совершенству, которое, возможно, и не будет достигнуто.

Поиски оптимальных решений привели к созданию специальных математических методов, и уже в XVIII веке были заложены математические основы оптимизации (вариационное исчисление, численные методы и др.). Однако до второй половины XX века методы оптимизации во многих областях науки и техники применялись очень редко, поскольку практическое использование математических методов оптимизации требовало огромной вычислительной работы, которую без ЭВМ реализовать было крайне трудно, а в ряде случаев – невозможно. Особенно большие трудности возникали при решении задач оптимизации из-за большого числа параметров и их сложной взаимосвязи между собой. При наличии ЭВМ ряд задач оптимизации поддается решению.

Выбор того или иного метода оптимизации в значительной степени определяется постановкой задачи оптимизации, а также математической моделью объекта оптимизации.

Основные методы оптимизации, наиболее широко используемые в химической технологии, можно разделить на несколько групп:

1. Аналитические методы:

а) методы исследования функций классического анализа. Применяются для детерминированных процессов с критерием оптимальности в виде дифференцируемых функций;

б) метод множителей Лагранжа. Применяется для задач с ограничениями типа равенств с критерием оптимальности в виде дифференцируемых функций;

в) вариационные методы – для задач с критерием оптимальности в виде функционала;

г) принцип максимума.

2. Методы математического программирования:

а) динамическое программирование;

б) нелинейное программирование;

в) линейное программирование.

3. Градиентные методы.

4. Статистические методы.

Наиболее сложен для оптимизации случай, когда неизвестен вид целевой функции. В этом случае оптимум находится экспериментально.

Существует две области изменения выходного параметра *y*:

1. Область, удаленная от оптимума, в которой происходит значительное изменение *y*.

2. Почти стационарная область, в которой практически не происходит изменение *y*.

После того как область, удаленная от оптимума, описана линейным уравнением, используем его для оптимизации.

Нельзя рекомендовать какой-либо один метод для использования его в решении всех задач, возникающих на практике. Одни методы в этом отношении являются более общими, другие − менее общими. Большую группу методов, таких как исследования функций классического анализа, метод множителей Лагранжа, методы нелинейного программирования, на определенных этапах решения оптимальной задачи можно применять в сочетании с другими методами, например динамическим программированием или принципом максимума.

Некоторые методы специально разработаны или наилучшим образом подходят для решения оптимальных задач с математическими моделями определенного вида. Например, математический аппарат линейного программирования специально создан для решения задач с линейными ограничениями на переменные и линейными критериями оптимальности и позволяет решать большинство задач, сформулированных в такой постановке. Геометрическое программирование предназначено для решения оптимальных задач, в которых ограничения и критерии оптимальности представляются специального вида функциями – полиномами.

Для решения задач оптимизации многостадийных процессов (особенно тех, в которых состояние каждой стадии характеризуется небольшим числом переменных состояния) хорошо приспособлено динамиче-

ское программирование. Однако при наличии значительного числа этих переменных, т. е. при высокой размерности каждой стадии, применение метода динамического программирования затруднительно в результате ограничения быстродействия и объема памяти вычислительных машин.

Наилучшим путем при выборе метода оптимизации, наиболее пригодного для решения соответствующей задачи, является исследование возможностей и опыта применения различных методов оптимизации.

Методы исследования функций классического анализа представляют собой наиболее известные методы решения несложных оптимальных задач, которые хорошо известны из курса математического анализа. Обычной областью использования данных методов являются задачи с известным аналитическим выражением критерия оптимальности, что позволяет найти не очень сложное аналитическое выражение для производных. Экстремальные решения оптимальной задачи получают приравниванием к нулю производных уравнения. Для решения подобных задач аналитическим путем применяют вычислительные машины. При этом решается система конечных уравнений, чаще всего нелинейных.

При решении задач оптимальности с использованием методов исследования функций классического анализа дополнительные трудности возникают вследствие того, что система уравнений, получаемая в результате их применения, обеспечивает лишь необходимые условия оптимальности. В связи с этим все решения данной системы (а их может быть и несколько) должны быть проверены на достаточность. В результате такой проверки сначала отбрасывают решения, которые не определяют экстремальные значения критерия оптимальности, а затем среди остающихся экстремальных решений выбирают решение, удовлетворяющее условиям оптимальной задачи, т. е. наибольшему или наименьшему значению критерия оптимальности в зависимости от постановки задачи.

Методы исследования при наличии ограничений на область изменения независимых переменных можно использовать только для отыскания экстремальных значений внутри указанной области. В особенности это относится к задачам с большим числом независимых переменных (практически больше двух), в которых анализ значений критерия оптимальности на границе допустимой области изменения переменных становится весьма сложным.

Стандартная математическая задача оптимизации формулируется таким образом. Среди элементов x, образующих множества X, найти такой элемент x^*, который доставляет минимальное значение $f(x^*)$ заданной функции $f(x)$. Для того чтобы корректно поставить задачу оптимизации, необходимо задать:

1. Допустимое множество

$$X = \left\{ \vec{x} / g_i(\vec{x}) \le 0,\ i = 1, \ldots m \right\} \subset R^n .$$

2. Целевую функцию − отображение

$$f : X \to R.$$

3. Критерий поиска (max или min).

Тогда решить задачу можно тремя путями:

1) показать, что $X = \varnothing$;

2) показать, что целевая функция не ограничена снизу.

3) найти

$$\vec{x} \in X : f\left(\vec{x}\right) = \min_{\vec{x} \in X} f(\vec{x}).$$

Если минимизируемая функция не является выпуклой, то часто ограничиваются поиском локальных минимумов и максимумов: точек таких, что всюду в некоторой их окрестности для минимума и для максимума. Найти max(min) = $Z = z(x)$.

При решении оптимальной задачи методами исследования функций классического анализа приходится использовать численные методы, аналогичные методам нелинейного программирования.

5.1. Численные методы решения задач оптимизации

Нелинейное программирование занимается оптимизацией моделей задач, в которых либо ограничения, либо целевая функция, либо то и другое нелинейны.

Для выяснения трудностей решения задач данного класса, порождаемых нелинейностью, сопоставим задачи линейного и нелинейного программирования. Можно указать три характерные особенности для каждого класса (табл. 4).

Таблица 4

Сопоставление задач линейного и нелинейного программирования

Задачи линейного программирования	Задачи нелинейного программирования
1. Область Ω допустимых планов – ***выпуклое множество*** с конечным числом угловых (крайних) точек	1. Множество Ω допустимых планов может быть ***невыпуклым, несвязным***, иметь бесконечное число крайних точек
2. Экстремальное значение линейная целевая функция $z(X)$ достигает в одной из крайних точек (на ***границе*** области Ω допустимых решений)	2. Экстремум может достигаться не только на ***границе***, но и ***внутри*** области Ω допустимых решений
3. Экстремальное значение $z(X)$ целевой функции является и ***глобальным*** значением	3. Целевая функция $z(X)$ в области Ω может иметь ***несколько локальных*** экстремумов

Большинство существующих методов в нелинейном программировании можно разделить на два больших класса:

1. Прямые методы – методы непосредственного решения исходной задачи. Прямые методы порождают последовательность точек – решений, удовлетворяющих ограничениям, обеспечивающим монотонное убывание целевой функции. Недостаток: трудно получить свойство глобальной сходимости. Задачи с ограничениями в виде равенств. Метод замены переменных (МЗП).

2. Двойственные методы – методы, использующие понятие двойственности. В этом случае легко получить глобальную сходимость. Недостаток: не дают решения исходной задачи в ходе решения – оно реализуемо лишь в конце итерационного процесса.

На рис. 5.1 приводится классификация задач и методов нелинейного программирования.

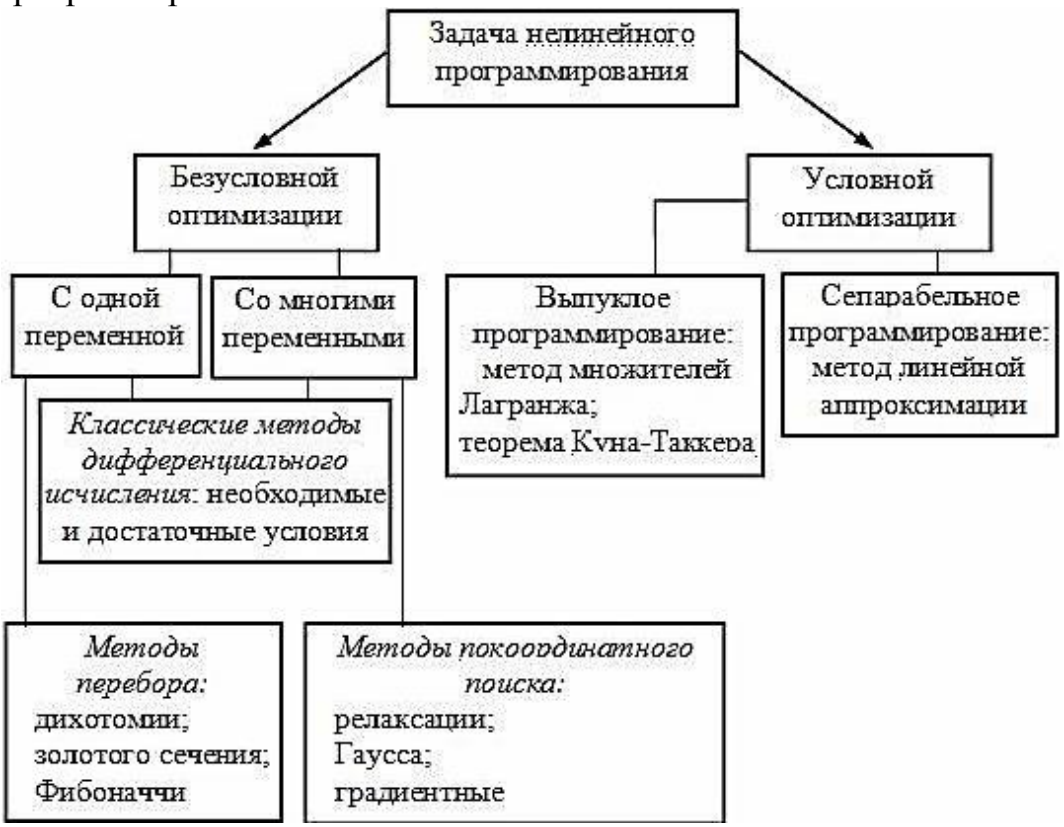


Рис. 5.1. Классификация задач и методов нелинейного программирования

Методы поиска нулей функции:

1) метод Ньютона (метод касательных);
2) метод хорд;
3) комбинированный метод;
4) метод итераций;
5) метод секущих.

Методы минимизации функций:

1) производные второго порядка;
2) одномерный поиск;
3) метод Фибоначчи;
4) метод золотого сечения;
5) метод Ньютона;
6) метод наискорейшего спуска;
7) матрица Гессе. Позволяет ответить на вопрос, является ли функция выпуклой или вогнутой;
8) метод множителей Лагранжа;
9) условия Куна–Таккера;
10) экстремум функции двух переменных.

Методы перебора применимы для отыскания экстремумов унимодальных целевых функций. Действие любого из методов поиска заключается в сужении области поиска экстремума:

а) до области заданной длины (> 0), проводя минимальное число измерений значений функции (методы дихотомии, золотого сечения);

б) до наименьших возможных размеров при заданном числе измерений n (метод Фибоначчи).

Первая формулировка целесообразна в том случае, если с каждым измерением связаны значительные затраты средств или времени, однако на поиск отпускаются неограниченные средства, которые мы все же стремимся минимизировать; вторая – когда исследователь располагает ограниченными средствами и, зная расходы, связанные с каждым измерением, стремится получить наилучший результат.

Классические методы нахождения экстремумов функций предполагают, что целевые функции непрерывные и гладкие. Для существования точки экстремума должны выполняться необходимые и достаточные условия. Необходимыми условиями существования экстремума являются требования обращения в нуль частных производных первого порядка целевой функции по каждой из переменных. Точка, найденная из необходимых условий, называется стационарной (подозрительной на оптимальную). В качестве стационарных точек могут быть точки перегиба, седловые точки и др. Поэтому необходим учет достаточных условий нахождения экстремумов функций. Он сложен для функций многих переменных как в алгебраическом, так и в вычислительном плане. Так, в случае функции двух переменных достаточным условием существования экстремума будет положительная определенность матрицы A размером 2×2 (условие Лежандра–Клебша), составленной из вторых частных производных функции. Недостатком классического метода дифференциального исчисления является и то, что он дает возможность найти

экстремум только в том случае, если он лежит внутри области определения функции. Если экстремум находится на границе области, то этот метод становится бессильным.

Методы покоординатного спуска относятся к группе приближенных методов нелинейной оптимизации и направлены на уменьшение трудностей, связанных с отысканием экстремума функции цели со сложной аналитической структурой классическими методами дифференциального исчисления. Суть этих методов заключается в продвижении от исходной точки в области определения функции к точке оптимума итеративно; в методе Гаусса – последовательно по каждой из переменных (покоординатно); в градиентных методах – одновременно по всем переменным в направлении градиента или антиградиента.

Критерием окончания итеративных процедур является равенство нулю всех частных производных целевой функции, или квадрат суммы всех частных производных целевой функции должен быть не более заданного числа e, или разность достигнутого значения целевой функции и значения в предыдущей точке должна быть не более e и др. [3].

Численное решение нелинейных (алгебраических или трансцендентных) уравнений вида $f(x) = 0$, заключается в нахождении значений x, удовлетворяющих (с заданной точностью) данному уравнению, и состоит из следующих основных этапов:

1. Отделение (изоляция, локализация) корней уравнения.

2. Уточнение с помощью некоторого вычислительного алгоритма конкретного выделенного корня с заданной точностью.

Целью первого этапа является нахождение отрезков из области определения функции, внутри которых содержится только один корень решаемого уравнения. Иногда ограничиваются рассмотрением лишь какой-нибудь части области определения, вызывающей по тем или иным соображениям интерес. Для реализации данного этапа используются графические или аналитические способы.

При аналитическом способе отделения корней полезна следующая теорема:

Непрерывная строго монотонная функция $f(x)$ имеет (и притом единственный) нуль на отрезке $[a, b]$ тогда и только тогда, когда на его концах она принимает значения разных знаков.

Достаточным признаком монотонности функции $f(x)$ на отрезке $[a, b]$ является сохранение знака производной функции.

Графический способ отделения корней целесообразно использовать в том случае, когда имеется возможность построения графика функции $y = f(x)$. Наличие графика исходной функции дает непосредственное представление о количестве и расположении нулей функции, что позво-

ляет определить промежутки, внутри которых содержится только один корень. Если построение графика функции $y = f(x)$ вызывает затруднение, часто оказывается удобным преобразовать уравнение к эквивалентному виду $f_1(x) = f_2(x)$ и построить графики функций $y = f_1(x)$ и $y = f_2(x)$. Абсциссы точек пересечения этих графиков будут соответствовать значениям корней решаемого уравнения.

Так или иначе, при завершении первого этапа должны быть определены промежутки, на каждом из которых содержится только один корень уравнения.

Для уточнения корня с требуемой точностью обычно применяется какой-либо итерационный метод, заключающийся в построении числовой последовательности $x(k)$ ($k = 0, 1, 2, \ldots$), сходящейся к искомому корню x уравнения $f(x) = 0$.

Процесс уточнения корня уравнения $f(x) = 0$ методом половинного деления (1) на отрезке $[a, b]$, при условии что функция $f(x)$ непрерывна на этом отрезке, заключается в следующем: исходный отрезок делится пополам. Если $f((a + b)/2) = 0$, то $x = (a + b)/2$ –является корнем уравнения. Если $f((a + b)/2) \neq 0$, то выбирается та из половин $[a, (a + b)/2]$ или $[(a + b)/2, b]$, на концах которой функция $f(x)$ имеет противоположные знаки. Новый суженный отрезок $[a(1), b(1)]$ снова делится пополам, и проводится то же рассмотрение и т. д. В результате на каком-то этапе либо находится точный корень уравнения, либо имеется последовательность вложенных друг в друга отрезков $[a(1), b(1)]$, $[a(2), b(2)]$, …, $[a(k), b(k)]$, для которых $f(a(k))f(b(k)) < 0$, $k = 0, 1, 2, \ldots$.

Если требуется найти корень с точностью ε, то деление отрезка пополам продолжается до тех пор, пока длина отрезка не станет меньше 2ε. Тогда середина последнего отрезка даст значение корня с требуемой точностью.

5.2. Метод Бокса–Уилсона (метод крутого восхождения по поверхности отклика)

Постановка задачи оптимизации:

Определить координату оптимальной (экстремальной) точки ($X_1^{\text{опт}}$, $X_2^{\text{опт}}$, …, $X_n^{\text{опт}}$) поверхности отклика $y = f(x_1, \ldots, x_n)$ – это градиентный метод.

Метод градиента предусматривает движение к оптимуму по кратчайшему пути, т. е. по градиенту.

Градиент – это вектор, направленный в сторону наибыстрейшего изменения функции.

Метод Бокса–Уилсона – шаговый метод движения по поверхности отклика.

$$Y^{\text{опт}} = a(X_1^{\text{опт}}, X_2^{\text{опт}}).$$

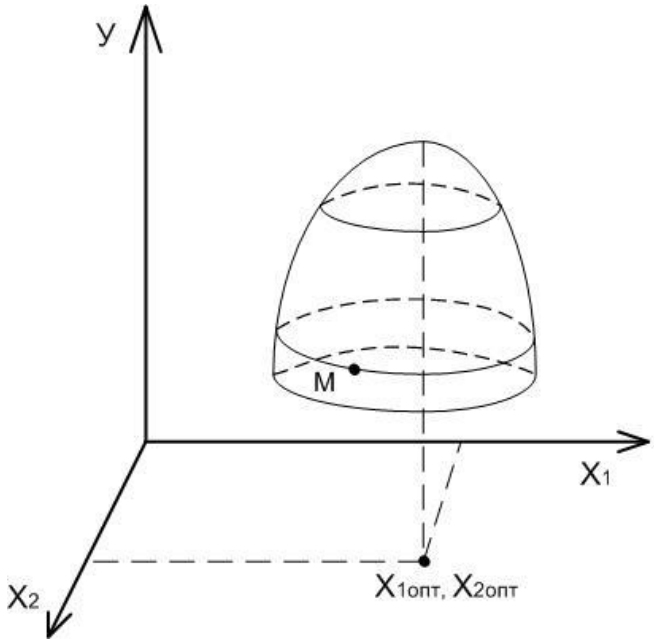


Рис. 5.2. Графическое представление метода Бокса–Уилсона

Пусть находимся в точке *M* (рис. 5.2). В этой точке ставится полный факторный эксперимент (ПФЭ) или дробный факторный эксперимент (ДФЭ) для локального описания поверхности в точке *M* линейным уравнением:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2.$$

Далее по этой поверхности отклика двигаемся по градиенту к экстремуму. До тех пор, пока наблюдается изменение *y* в лучшую сторону (либо увеличивается, либо уменьшается), т. е. минимум или максимум.

Как только *y* перестает изменяться (в данном случае расти), переносят центр планирования в точку, до которой дошли по градиенту, вновь выполняют эксперимент и строят плоскость.

Эта процедура продолжается до тех пор, пока не попадет почти в стационарную область. В этой области ставится эксперимент для описания поверхности полиномом второго порядка, затем исследуют эту поверхность для локализации экстремума.

Шаг по каждой оси дает частное производное по переменной X_i:

$$\frac{\partial f}{\partial x_1} = \frac{\partial \hat{y}}{\partial x_1} = b_1;$$

$$\frac{\partial f}{\partial x_2} = \frac{\partial \hat{y}}{\partial x_2} = b_2.$$

Чтобы двигаться к оптимуму, мы должны делать шаги, пропорциональные коэффициентам регрессии:

$$\text{grad } f = b_1\Delta x_1 + b_2\Delta x_2.$$

Расчет шагов приближения к оптимуму проводят следующим образом:

1. Вычисляют произведения коэффициентов регрессии на соответствующие интервалы варьирования различных факторов и фактор, для которого это произведение максимально, принимают за базовый.

2. Для базового фактора выбирают шаг варьирования крутого восхождения h_a, оставляя прежний интервал варьирования ΔX_i или выбирая новый, более мелкий.

3. Производят расчет шага для каждого фактора по выражению

$$\frac{b_i \cdot \Delta x_i}{a} \cdot h_a = h_i.$$

Коэффициенты берутся со своими знаками.

Движение начинают от основного уровня. При первом опыте факторы будут получать значения:

$$X_i = X_i^0 + h_i.$$

5.3. Метод деления отрезка пополам (метод дихотомии)

Рассмотрим поиск максимума в данном случае на отрезке AB (рис. 5.3):

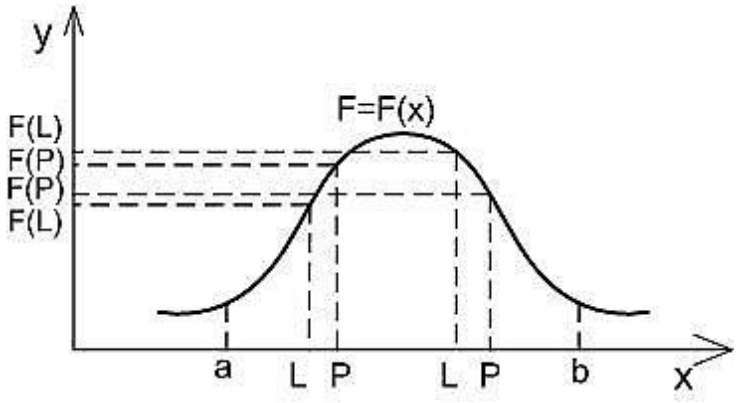


Рис. 5.3 Графическое представление метода деления отрезка пополам

Алгоритм:

1. Делим отрезок AB пополам точкой L и рассчитываем в точке L значение функции $F(L)$.

2. Выбираем малое приращение $\Delta = \frac{b-a}{10}$ или $\Delta = \frac{b-a}{100}$ ($\Delta \le \varepsilon$) и откладываем его влево или вправо от точки L и рассчитываем в точке P значение функции.

3. Сравниваем значения функций $F(L)$ и $F(P)$: если $F(L) > F(P)$, то максимум может находиться в левой части отрезка.

4. Отбрасываем ту часть отрезка, где максимума нет (в данном случае правую), и переносим точку B в точку P или L.

5. Получили отрезок AB. Алгоритм повторяется.

Новый отрезок AB делится пополам. Если $F(L) < F(P)$, то отбрасываем левый отрезок.

Вновь переносим точку A в L и повторяем процедуру деления отрезка пополам.

Процедура продолжается до тех пор, пока не выполнится заданное условие

$$\delta = |b - a| \le \varepsilon .$$

5.4. Метод золотого сечения

Поиск оптимума основан на делении отрезка на две части, при этом отношение длины всего отрезка к большей его части равно отношению большей части к меньшей:

$$\frac{l}{(l-m)} = \frac{(l-m)}{m};$$

$$l \cdot m = (l-m)^2;$$

$$l \cdot m = l^2 - 2lm + m^2;$$

$$l^2 - 3lm + m^2 = 0;$$

$$m = \frac{3l}{2} - \sqrt{\left(\frac{3l}{2}\right)^2 l^2} = \frac{3l}{2} - \sqrt{\frac{9l^2}{4} - l^2} = \frac{3l}{2} - \frac{l}{2}\sqrt{5} = \frac{l}{2}\left(3 - \sqrt{5}\right) = 0{,}382l,$$

где $m = 0{,}382l = (1 - 0{,}618)l$ – точка золотого сечения.

Рассмотрим поиск оптимума максимума методом золотого сечения.

Делим отрезок AB слева и справа в отношении золотого сечения. Получаем точки L и P (рис. 5.4). Рассчитываем значения функций в точках L и P и сравниваем их.

Если $F(L) > F(P)$, то максимум может находиться либо на участке AL, либо LP на участке PB максимума быть не может, если функция унимодальна.

Отбрасываем PB и переносим точку B в точку P. Получаем новый отрезок AB.

На новом отрезке уже есть точка L – точка золотого сечения.

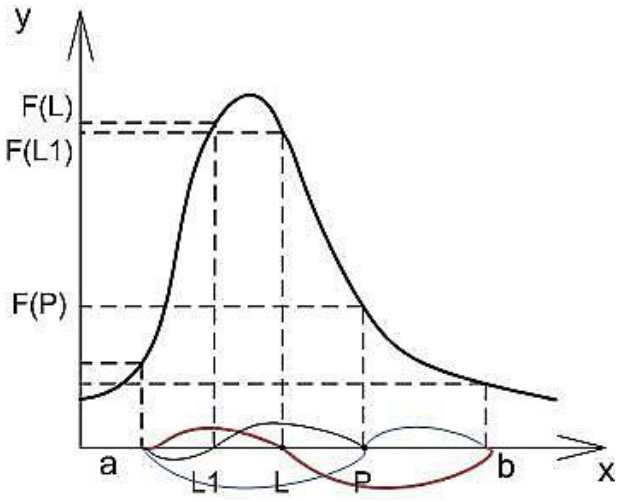


Рис. 5.4 Графическое представление метода золотого сечения

Процедура поиска продолжается до тех пор, пока не выполнится условие

$$|b - a| \leq \varepsilon.$$

5.5. Метод сканирования

Метод сканирования заключается в последовательном просмотре значений функций в ряде точек и нахождения среди них такой точки, в которой значение критерия оптимальности имеет экстремальное значение, т. е. максимум или минимум. Применяется данный метод к непрерывным функциям.

Сканированием можно исследовать как функцию одной, так и нескольких переменных.

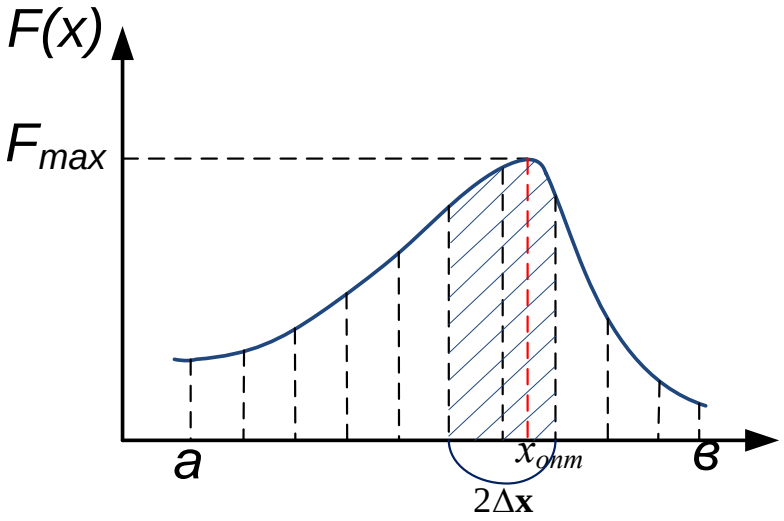


Рис. 5.5. Графическое представление метода сканирования

Рассмотрим одномерное сканирование:

Возьмем *AB*, на котором требуется отыскать экстремум целевой функции. Этот отрезок *AB* называют интервалом неопределенности функции.

В данном методе точку экстремума не обязательно определять абсолютно точно, достаточно сильно сузить интервал.

Таким образом, в одномерном случае задача поиска экстремума сводится к сужению интервала неопределенности.

Выберем целое число *K* значений целевой функции $\Delta x = \frac{b-a}{k-1}$ и в точках рассчитаем значения функции.

Выбираем максимальное значение функции и сужаем интервал неопределенности до $2\Delta x$ симметричных относительно максимальной точки.

Переносим концы отрезка и получаем новый отрезок *AB*.

Новый интервал неопределенности вновь разбивается на $\Delta x = \frac{b-a}{k-1}$.

Интегрирование (поиск оптимума) продолжается до тех пор, пока мы не сузили интервал неопределенности до той точки, которую мы задали сами.

Из всех представленных методов оптимизации удобнее всего применять метод дихотомии, так как при этом методе требуется произвести небольшое число итераций. Метод Бокса−Уилсона, наоборот, требует большого числа итераций и поэтому неудобен для практического применения.

6. ИНТЕРАКТИВНАЯ СИСТЕМА АНАЛИЗА И УПРАВЛЕНИЯ ДАННЫМИ – STATISTICA

STATISTICA − это инструмент разработки пользовательских приложений в различных областях промышленности. Все аналитические инструменты, имеющиеся в системе, доступны пользователю и могут быть выбраны с помощью альтернативного пользовательского интерфейса. Гибкая и мощная технология доступа к данным позволяет эффективно работать как с таблицами данных на локальном диске, так и с удаленными хранилищами данных.

Система обладает следующими общепризнанными достоинствами:

- содержит полный набор классических методов анализа данных: от основных методов статистики до продвинутых методов, что позволяет гибко организовать анализ;
- является средством построения приложений в конкретных областях;
- в комплект поставки входят специально подобранные примеры, позволяющие систематически осваивать методы анализа;
- отвечает всем стандартам Windows, что позволяет сделать анализ высокоинтерактивным;
- система может быть интегрирована в Интернет;
- поддерживает веб-форматы: HTML, JPEG, PNG;
- легка в освоении, и, как показывает опыт, пользователи из всех областей применения быстро осваивают систему;
- данные системы STATISTICA легко конвертировать в различные базы данных и электронные таблицы;
- поддерживает высококачественную графику, позволяющую эффектно визуализировать данные и проводить графический анализ;
- является открытой системой: содержит языки программирования, которые позволяют расширять систему, запускать ее из других Windows-приложений, например из Excel.

STATISTICA позволяет проводить исчерпывающий, всесторонний анализ данных, представлять результаты анализа в виде таблиц и графиков, автоматически создавать отчеты о проделанной работе. С помощью удобной системы подсказок можно обучаться не только работе с самим пакетом, но и современным методам статистического анализа.

6.1. Общие принципы работы с программой

Данные в системе STATISTICA организованы в виде электронных таблиц, как в привычной для пользователей программе Excel. Файл содержит наблюдения и переменные. Наблюдения можно рассматривать как эквивалент записей в базах данных (или строк электронной таблицы), а переменные − как эквивалент полей (столбцов электронной таблицы). Каждое наблюдение состоит из набора значений переменной.

В пакете STATISTICA все операции, включая копирование, перетаскивание и автоматическое заполнение ячеек, производятся так же, как в популярных электронных таблицах. При нажатии правой кнопки мыши появляется всплывающее меню, где точно так же предлагается перечень операций, которые можно выполнить над выделенным объектом.

Система STATISTICA предоставляет всесторонние возможности по импорту и экспорту данных, в том числе и из таблиц Excel.

В системе имеется также менеджер мегафайлов (доступный из модуля *Управление данными*), который позволяет работать с очень большими файлами, содержащими до 32 000 переменных.

6.1.1. Интерфейс программы

STATISTICA состоит из набора модулей (рис. 6.1), в каждом из которых собраны тематически связные группы процедур. При переключении модулей можно либо оставлять открытым только одно окно приложения STATISTICA, либо все вызванные ранее модули, поскольку каждый из них может выполняться в отдельном окне (как самостоятельное приложение Windows).

Быстро переключаться из одного модуля в другой можно: а) щелкая мышью на значках модулей на рабочем столе; б) активизируя соответствующее окно приложения (если оно уже было открыто) или в) выбирая модули в диалоговом окне *Переключатель модулей*, причем эту операцию можно настроить так, чтобы было удобно обращаться к модулям, которые используются чаще всего.

Интерфейс системы может быть настроен на конкретный пользовательский проект: можно задать отображение стольких диалоговых окон, таблиц результатов, графиков, сколько в данном случае необходимо.

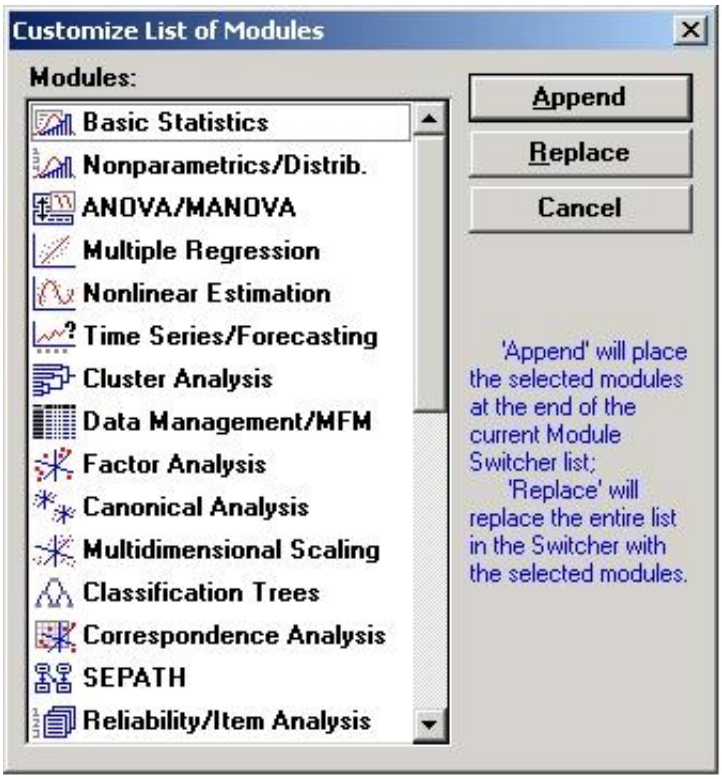


Рис. 6.1. Окно настройки модулей

Система включает следующие модули методов анализа:

Статистики и таблицы.

Исчерпывающий набор описательных статистик, таблицы сопряженности, таблицы флагов и заголовков, кросс-табуляция многомерных откликов и многомерных дихотомий, вычисление корреляционных матриц, обработка пропущенных данных, *t*-критерии для зависимых и независимых выборок, критерии однородности дисперсии, однофакторный дисперсионный анализ.

Непараметрическая статистика.

Непараметрические критерии, ранговые корреляции, подгонка распределений.

Множественная регрессия.

Пошаговая регрессия с включением и исключением предикторов, нелинейная регрессия, ридж-регрессия, построение прогнозов, всесторонний анализ остатков, вычисление прогнозов и доверительных интервалов для прогнозируемых значений (можно анализировать очень большие модели – до 500 переменных).

Нелинейное оценивание.

Подгонка любой задаваемой пользователем функции, задаваемая пользователем функция потерь, разрывная регрессия.

Временные ряды и прогнозирование.

Широкий выбор моделей анализа временных рядов, включая модели АРПСС − авторегрессии и проинтегрированного скользящего среднего, модели с интервенцией, анализ распределенных лагов, спектральный анализ чрезвычайно длинных временных рядов, преобразования рядов, включая быстрое преобразование Фурье, и многие другие процедуры углубленного анализа.

Кластерный анализ.

Широкий набор процедур кластерного анализа, включая иерархическое объединение, двухвходовое объединение, метод к-средних; алгоритмы оптимизированы для анализа очень больших проектов, например, методом к-средних можно анализировать 400 000 наблюдений с 10 переменными.

Факторный анализ.

Процедуры факторного анализа и анализа главных компонент, ортогональные и косоугольные факторы, иерархический анализ косоугольных факторов и др.

Канонический анализ.

Вычисление канонических переменных и канонических корней.

Многомерное шкалирование.

Анализ расстояний, матриц сходств и различия, диаграмма Шепарда и др.

Деревья классификации.

Современные методы построения деревьев классификации с категориальными и порядковыми предикторами и различными функциями потерь.

Анализ соответствий.

Современные методы анализа таблиц сопряженности.

Структурное моделирование.

Построение структурных моделей, продвинутый факторный анализ.

Надежность и позиционный анализ.

Методы построения вопросников, оценка надежности позиций и др.

Дискриминантный анализ.

Процедуры всестороннего дискриминантного анализа, разнообразные статистики и графическое представление результатов.

Логлинейный анализ.

Всесторонний анализ многовходовых таблиц сопряженности, автоматическое построение лучшей модели.

Анализ выживаемости.

Анализ таблиц жизни, оценки Каплана–Мейера, регрессионные модели: Кокса, логнормальная, экспоненциальная, зависящие от времени ковариаты, разнообразные статистики и критерии.

Дисперсионный анализ.

Полный набор методов одномерного и многомерного дисперсионного анализа, фиксированные и переменные ковариаты, апостериорные критерии, контрасты, проверка предположений дисперсионного анализа, планы с повторными измерениями, иерархически вложенные планы, планы с пропущенными ячейками и многое другое.

Компоненты дисперсии.

Смешанные модели дисперсионного анализа, оценка компонент дисперсии.

6.1.2. Графические возможности программы

STATISTICA обладает огромными возможностями для построения графиков непосредственно из таблиц исходных данных и таблиц результатов, причем графика и анализ данных тесно интегрированы. Например, если после вычисления корреляционной матрицы у пользователя возникает потребность в графическом представлении корреляционной зависимости, то достаточно поместить курсор на соответствующий коэффициент корреляции, нажать правую кнопку мыши и в появившемся меню выбрать пункт *Быстрые статистические графики*, а затем одну из диаграмм рассеяния. На экране появится требуемый график. В разных модулях системы имеются свои специальные графики, учитывающие особенности получаемых в них результатов.

Один из способов построения графиков в системе STATISTICA – используя окно *Галерея графиков* (рис. 6.2).

Отсюда быстро и легко вызываются все статистические и пользовательские графики, пустые графические окна и статистические графики пользователя. Для этого нужно выделить название нужного типа графика и дважды щелкнуть на нем (или нажать кнопку *ОК*).

Пользовательские и статистические графики.

Помимо специализированных графиков, которые вызываются непосредственно из итогового диалогового окна любой программы статистической обработки, существуют еще два основных типа графиков, доступных из меню или панели инструментов любой таблицы: пользовательские графики и статистические графики.

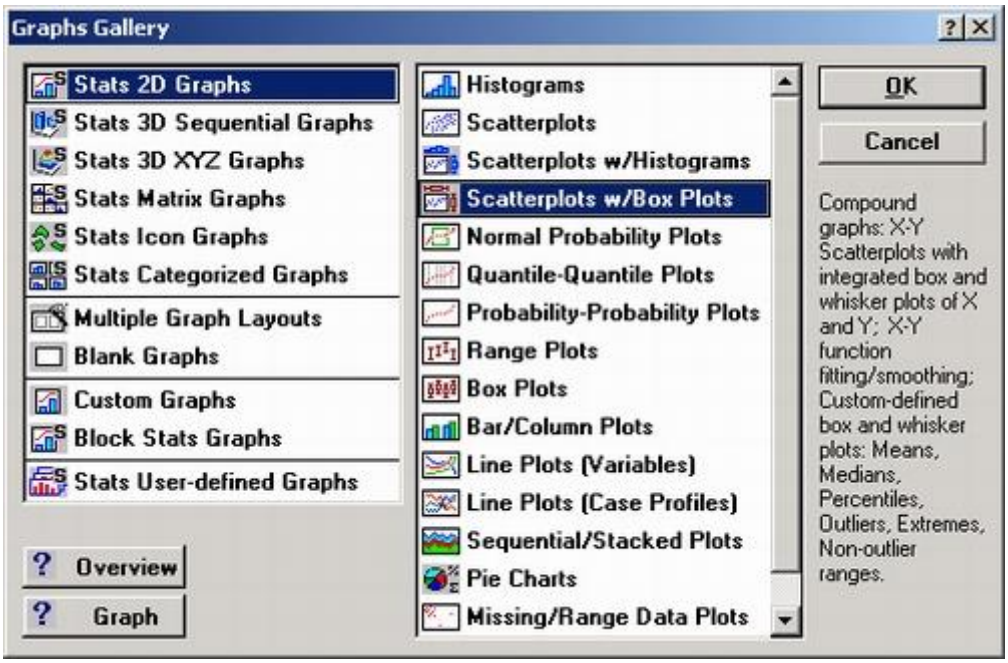


Рис. 6.2. Окно «Галерея графиков»

Главное различие между двумя основными типами графиков заключается в источнике данных для отображения. Более подробно эти различия описаны в следующих разделах: пользовательские графики, статистические графики, быстрые статистические графики, блоковые статистические графики.

Пользовательский график дает возможность отобразить любую заданную пользователем комбинацию значений из таблиц исходных данных или таблиц результатов (а также из любой комбинации их строк и/или столбцов). В меню предлагается пять типов таких графиков: 2М пользовательские графики, 3М пользовательские последовательные графики, 3М пользовательские диаграммы рассеяния и поверхности, пользовательские матричные графики и пользовательские пиктографики. При выборе одного из них открывается соответствующее диалоговое окно, где для отображения на графике можно задать диапазон данных текущей таблицы. Начальный выбор данных для построения графика, предлагаемый в этом диалоговом окне, определяется положением курсора в текущей таблице. В каждом диалоговом окне пользовательского графика при задании параметров предусмотрена возможность выбора определенного вида графика.

В отличие от пользовательских графиков, которые представляют собой средство наглядного отображения числовых данных любых таблиц, **статистические графики** предлагают сотни заранее определенных типов графических представлений, включающих аналитическое обобщение статистических данных. Они вызываются из диалогового окна

Галерея графиков, которое открывается с помощью одноименной кнопки панели инструментов или из выпадающего меню *Графика* (рис. 6.3).

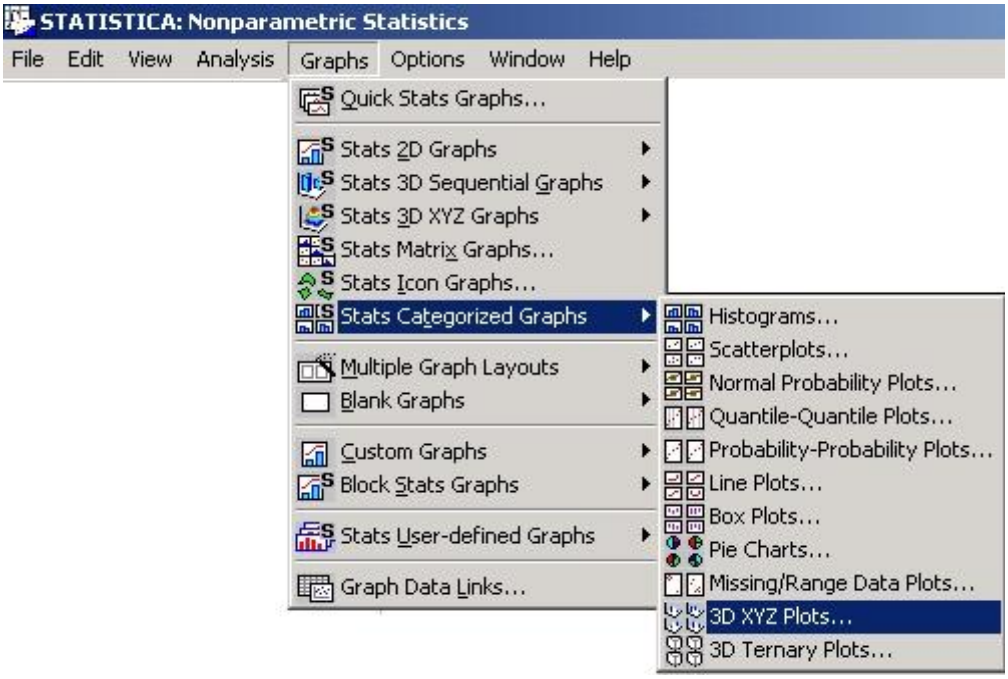


Рис. 6.3. Панель инструментов из меню «Графика»

При построении таких графиков используются значения непосредственно из файла данных, которые не зависят от содержания текущей таблицы, выделения блоков и положения курсора. При этом предлагаются либо стандартные методы графического анализа исходных данных (различные графики разброса значений, гистограммы, графики средних значений, например, медиан), либо стандартные аналитические методы исследований (графики нормальной плотности распределения, вероятностные графики с исключенным трендом или графики доверительных интервалов линий регрессии). При построении статистических графиков программа учитывает условия выбора и веса наблюдений.

Наиболее широко используемые типы статистических графиков представлены в меню *Быстрые статистические графики*. Эти списки графиков упрощают и ускоряют процедуру построения графика. Быстрые статистические графики:

- вызываются из контекстных меню или с панели инструментов любой таблицы (обычно они не требуют обращения к выпадающим меню или диалоговым окнам);
- не требуют от пользователя выбора переменных (этот выбор определяется текущим положением курсора в таблице) и промежуточной настройки параметров (формат соответствующих графиков определяется по умолчанию).

При выборе пункта *Быстрые статистические графики* появляется меню выбора статистического графика (рис. 6.4) для текущей переменной таблицы, то есть той, на которую в настоящий момент указывает курсор. Если курсор не указывает ни на одну из переменных, то перед построением любого графика из меню *Быстрые статистические графики* будет предложено выбрать переменную из списка. При создании таких графиков система STATISTICA учитывает текущие условия выбора и веса наблюдений.

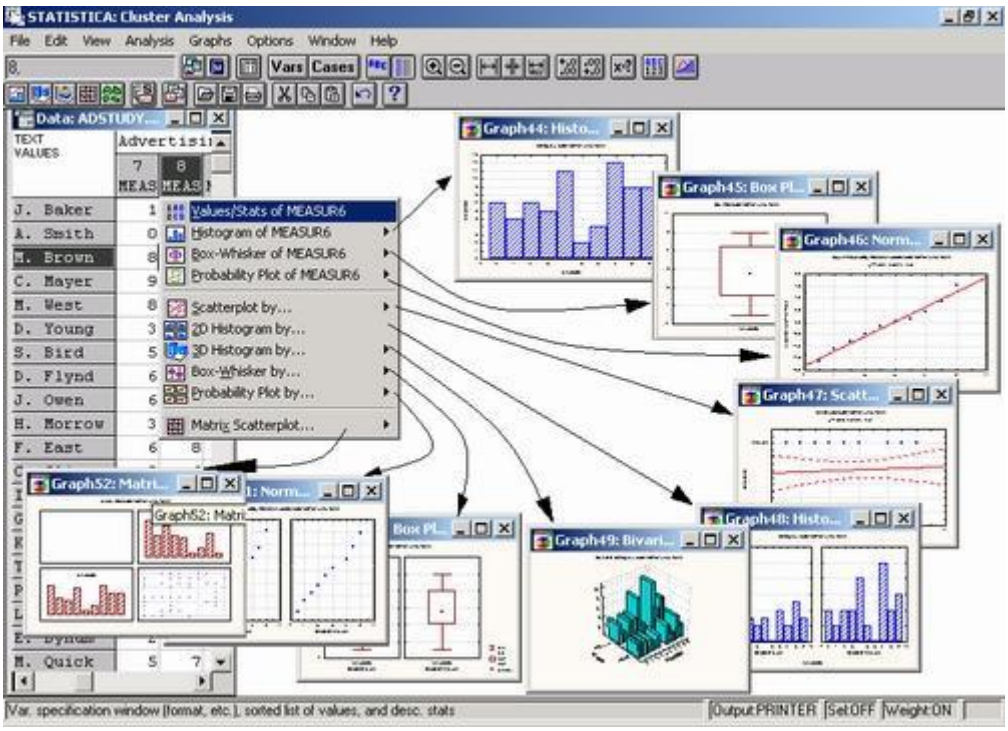


Рис. 6.4. Выбор статистического графика

Блоковые статистические графики дают возможность построить итоговый статистический график для выделенного блока, чтобы сравнить значения в строках (*Статистики блока по строкам*) или в столбцах таблицы (*Статистики блока по столбцам*). Данный тип графиков похож на те пользовательские графики, на которых отображаются данные текущего блока таблицы.

Помимо стандартного набора быстрых статистических графиков, некоторые таблицы позволяют строить и более специализированные статистические графики (например, временные последовательности в модуле *Временные ряды*, пиктографики регрессионных остатков, а также контурные графики в модуле *Кластерный анализ*). Специализированные графики, которые связаны не с конкретной таблицей результатов, а с определенным методом анализа данных (например, графики аппроксимирующих функций в модуле *Нелинейное оценивание* или сред-

них в модуле *Дисперсионный анализ*), вызываются непосредственно из диалогового окна с результатами анализа (то есть из окна, содержащего выходные параметры используемого метода обработки данных).

Перед построением графика нужно выбрать переменные и метод категоризации, а также при необходимости задать значения некоторых параметров с помощью кнопки *Параметры*.

После построения графика при щелчке на любом месте фона графического окна появится диалоговое окно *Общая разметка*, в котором регулируются параметры общего расположения графика.

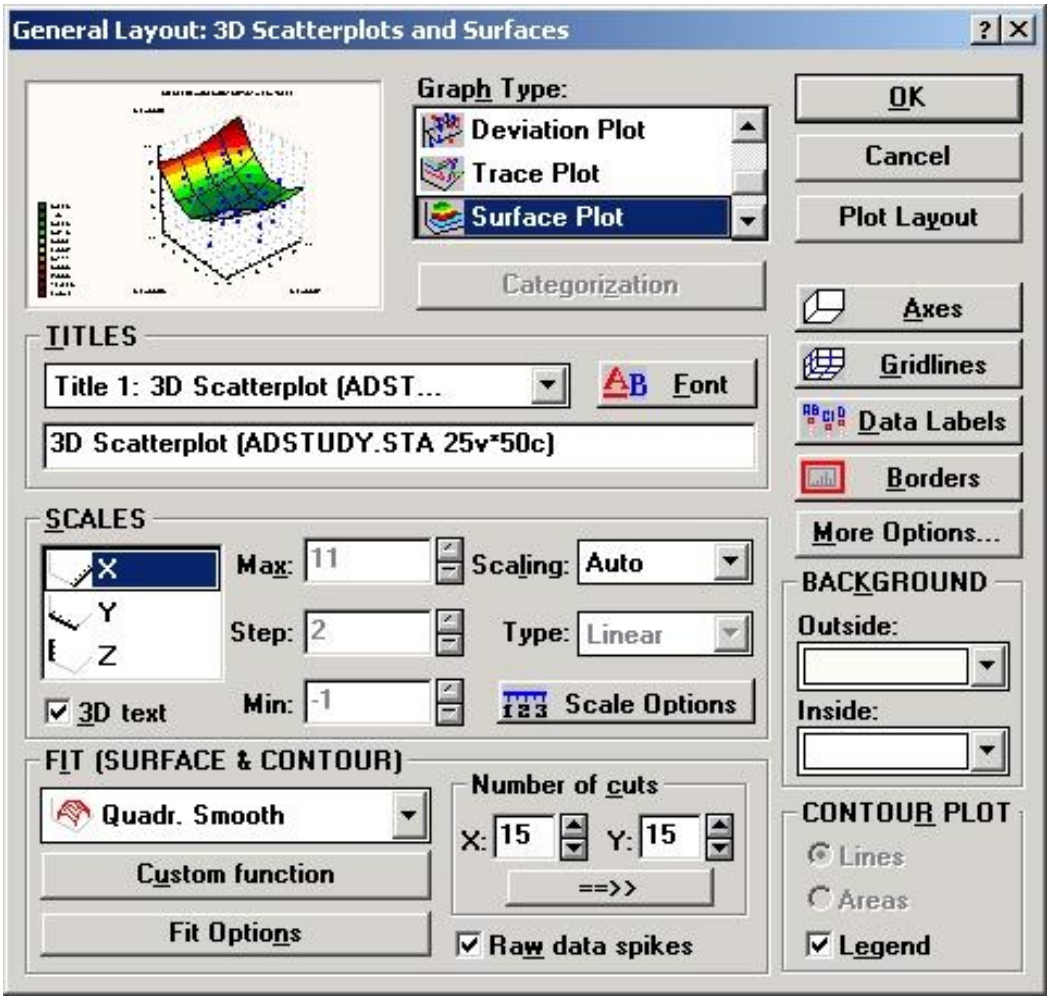


Рис. 6.5. Окно параметров расположения графика

В этом окне можно изменить тип графика и задать построение карты линий, используя для этого поле *Тип графика* (см. рис 6.6). Кроме того, можно изменить параметр *Число сечений* с установленного по умолчанию значения 15 × 15 на 25 × 25 (этот параметр определяет точность построения карты линий уровня).

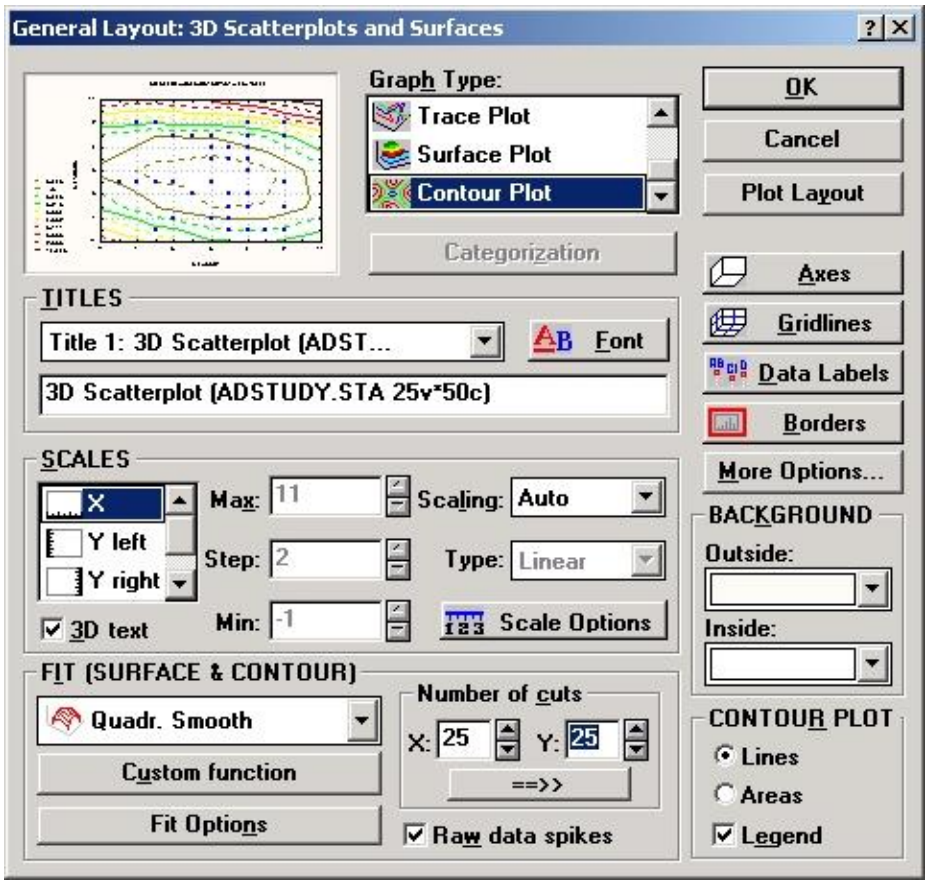


Рис. 6.6. Окно изменения вида графика

После внесения изменений нажмите *ОК*, и вы увидите новый график (рис. 6.7).

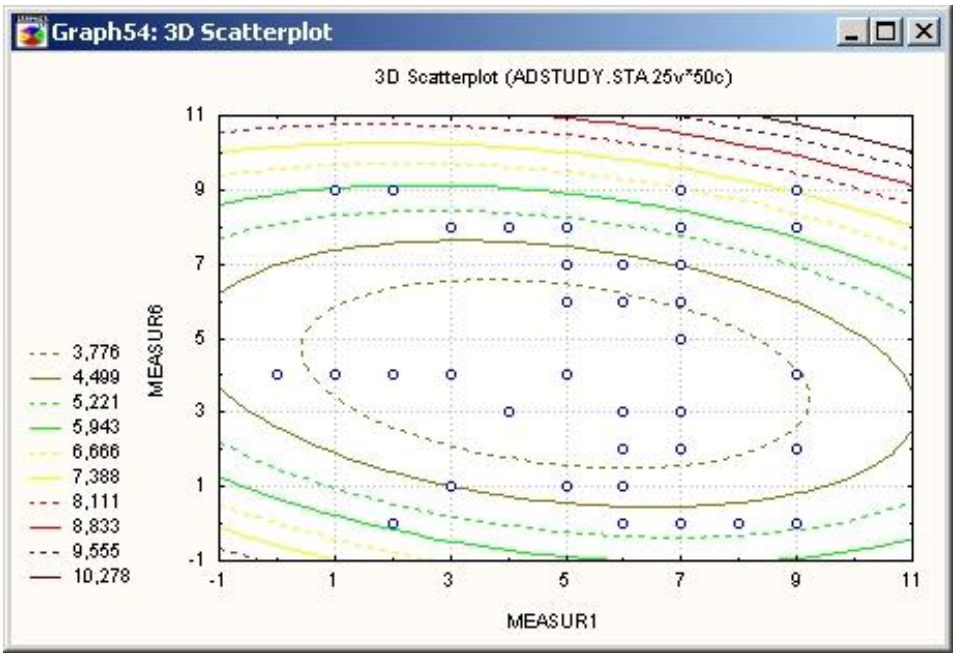


Рис. 6.7. Вид статистического графика

В диалоговом окне *Общая разметка* выберем для типа контурной линии значение *Зона*. Кроме того, в первые три строки заголовка графика поместим управляющие символы @F[1, 1], @F[1, 2] и @F[1, 3], чтобы записать там уравнения аппроксимирующей квадратичной функции для первой зависимости для каждого из трех отдельных графиков (цифры 1, 2 и 3 в качестве вторых параметров).

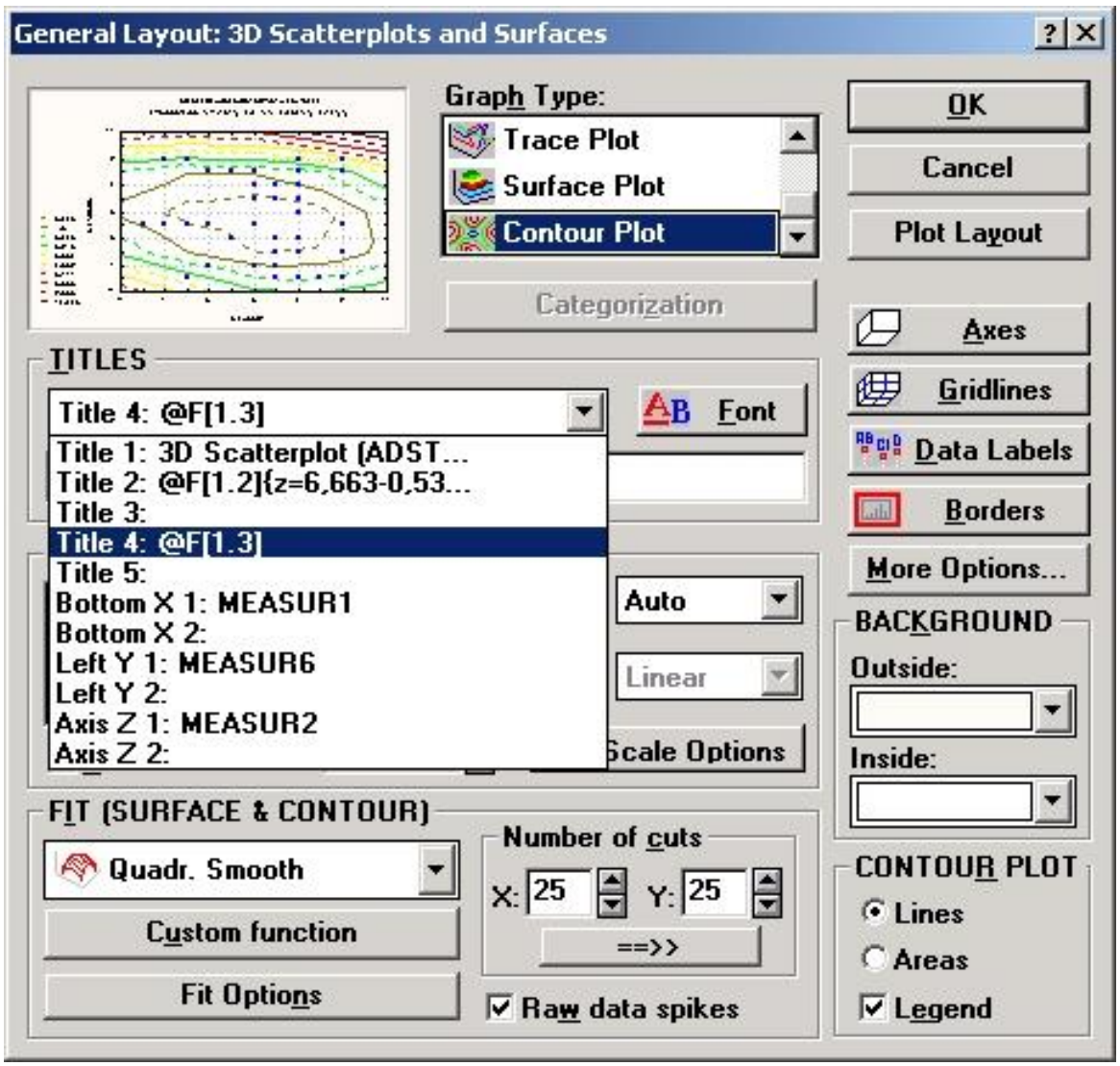


Рис. 6.8. Окно параметров графика

Для быстрейшего отображения и всестороннего форматирования уравнений функций лучше использовать диалоговое окно *Параметры*, которое вызывается из диалогового окна *Статистические графики*. Нажмите *ОК*, и вы увидите измененный график (рис. 6.9).

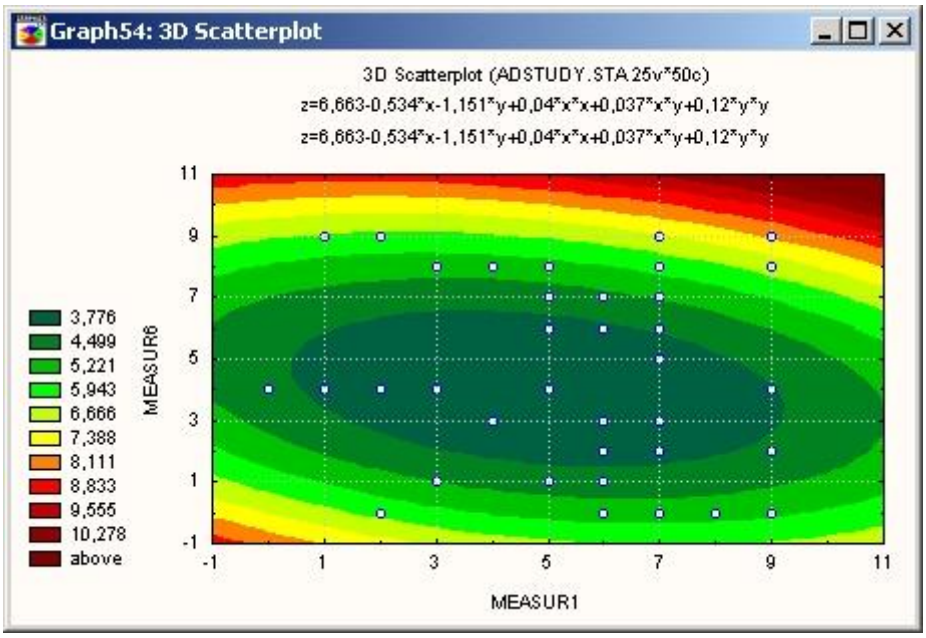


Рис. 6.9. Измененный вид статистического графика

Графики и рисунки сохраняются в графическом формате STATISTICA в файле с расширением *.stg.

6.1.3. Настройка системы STATISTICA

В системе предусмотрена возможность настройки множества характеристик и интерфейса программы в соответствии с предпочтениями пользователя. Можно изменить, например, процесс запуска, а именно отменить установленный по умолчанию полноэкранный режим, изменить вид стартовой панели, панели инструментов, таблиц с данными и другие параметры.

Настройка общих параметров системы.

Настройку общих параметров системы изменить в любой момент работы с программой. Эти параметры определяют:

- общие аспекты поведения программы (максимизация окна STATISTICA при запуске, Рабочие книги, инструмент *Перетащить* и отпустить − *Drag-and-Drop*, автоматические связи между графиками и данными, многозадачный режим и т. д.);
- режим вывода (например, автоматическая распечатка таблиц или графиков, форматы отчетов, буферизация и т. д.);
- общий вид окна приложения (значки, панели инструментов и т. д.);
- вид окон документов (цвета, шрифты).

Каждый из этих параметров можно настроить в соответствующем окне, доступ к которому осуществляется через меню *Сервис*. На рис. 6.10 и 6.11 показаны два примера таких окон.

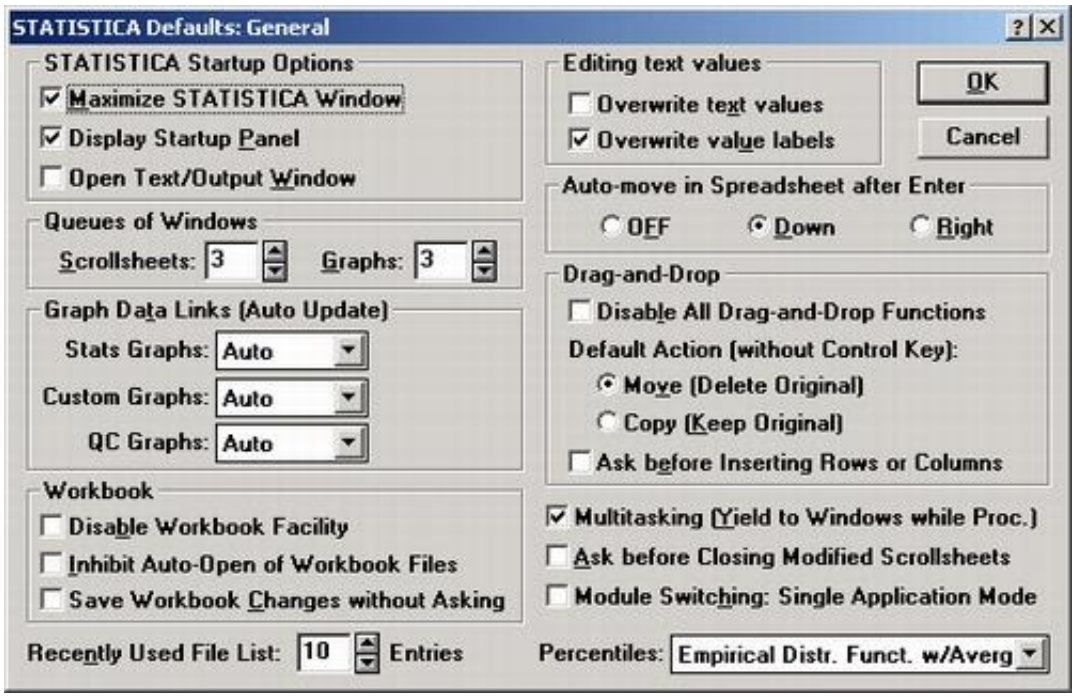


Рис. 6.10. Основное окно настройки параметров системы

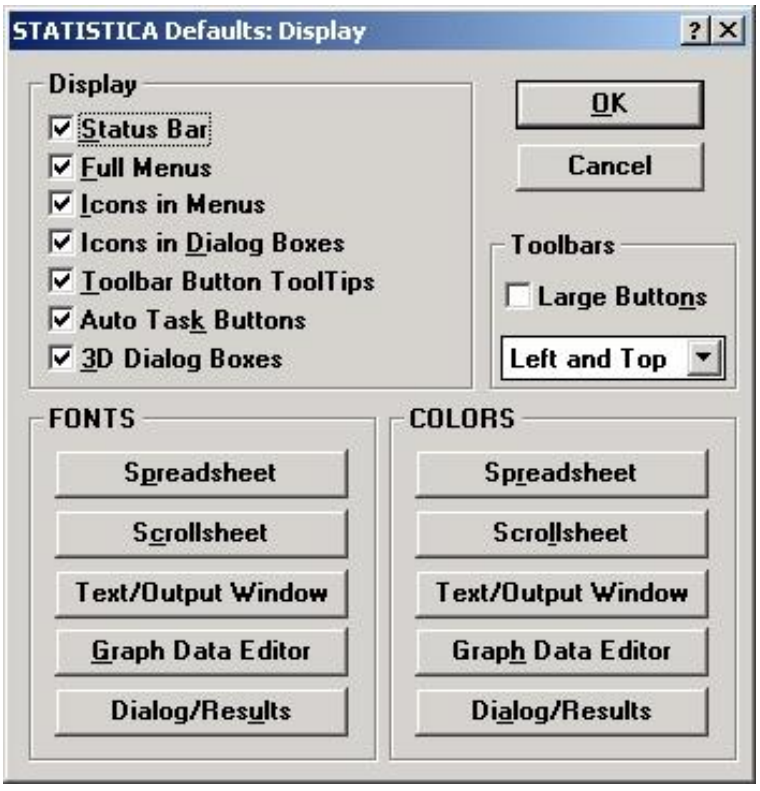


Рис. 6.11. Дисплей настройки

Все общие параметры могут быть настроены независимо от типа окна документа (например, таблица или график), которое активно в данный момент.

Настройка пользовательского интерфейса.

При работе с системой STATISTICA имеется возможность настройки пользовательского интерфейса программы таким образом, чтобы он стал более «продуманным» с точки зрения потребностей конкретного пользователя.

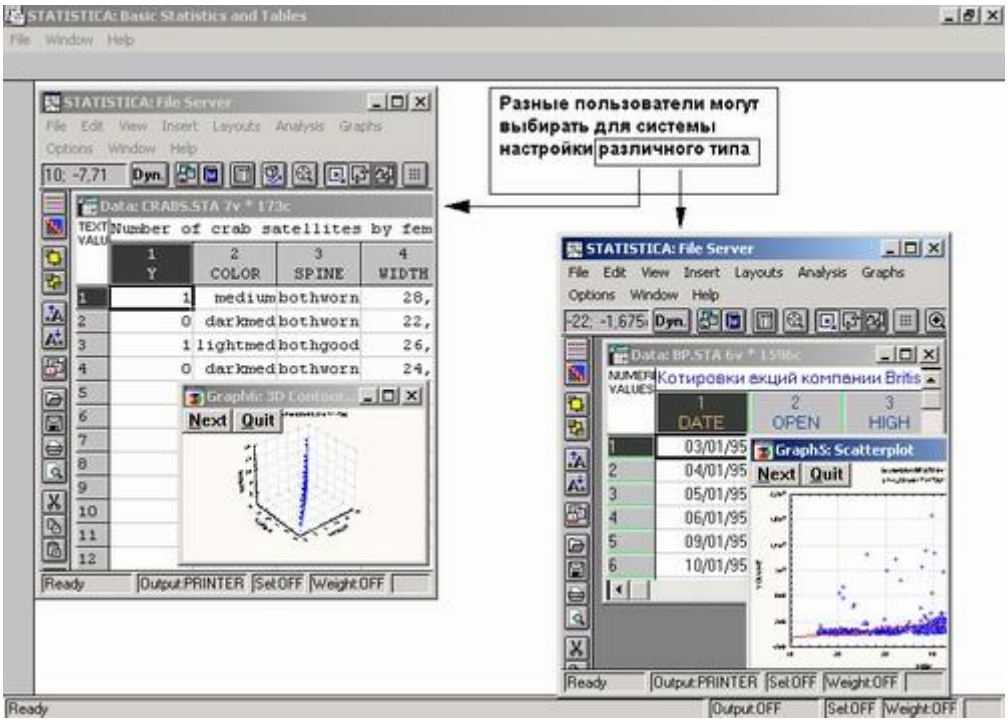


Рис. 6.12. Пример разных интерфейсов в системе STATISTICA

В зависимости от требований задачи и личных предпочтений (а также эстетических соображений) можно использовать разнообразные «режимы» и условия работы программы.

Поддержка нескольких различных конфигураций системы STATISTICA.

До внесения специальных изменений STATISTICA будет хранить все текущие настройки и параметры по умолчанию.

То обстоятельство, что сведения о конфигурации системы хранятся в той же папке, из которой вызывается программа STATISTICA, позволяет иметь в своем распоряжении различные варианты конфигурации программы для разных проектов или видов работ. Например, можно вызывать программу из разных папок на диске, каждая из которых содержит определенный связный набор документов, и для каждой из этих папок система может быть сконфигурирована со своими настройками вывода, параметрами графиков по умолчанию и т. д. Можно создать несколько значков STATISTICA в разных группах приложений на рабочем столе Windows (каждая из которых соответствует определенному

проекту или виду работ) и задать для них различные значения в поле *Рабочая директория* (с помощью диалогового окна системы Windows *Свойства программного элемента* (Program Item Properties).

Многозадачность.

STATISTICA поддерживает режим многозадачности (между своими модулями или другими приложениями).

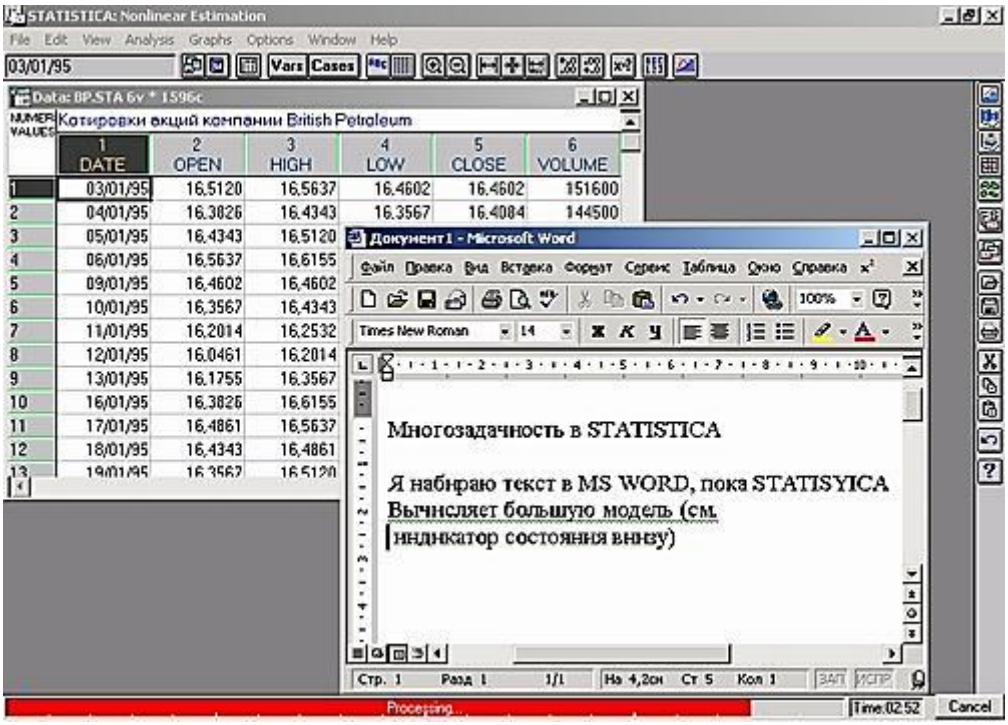


Рис. 6.13. Многооконный режим системы

При обработке очень больших объемов информации или выполнении сложных процедур анализа можно переключиться в другой модуль STATISTICA (или другое приложение Windows), используя возможность вести процесс обработки данных в фоновом режиме.

Работа в одном окне приложения STATISTICA.

Один из вариантов глобальной системной настройки пакета STATISTICA позволяет пользователю задать режим, в котором по умолчанию будет работать программа: в одном окне приложения или же как набор приложений (каждое в своем окне). Одним из непосредственных следствий этого выбора будет то, в каком режиме будет работать окно *Переключатель модулей*: при двойном щелчке на имени модуля в этом окне выбранный модуль будет открываться либо вместо уже открытого, либо для него будет открываться новое окно приложения, причем предыдущее окно останется открытым.

Выбор того или другого режима работы производится в поле *Переключение модулей*: режим одного приложения в диалоговом окне *Пара-*

метры по умолчанию: общие настройки (вызывается из меню *Сервис*). Если это поле отмечено, STATISTICA будет работать в режиме одного приложения.

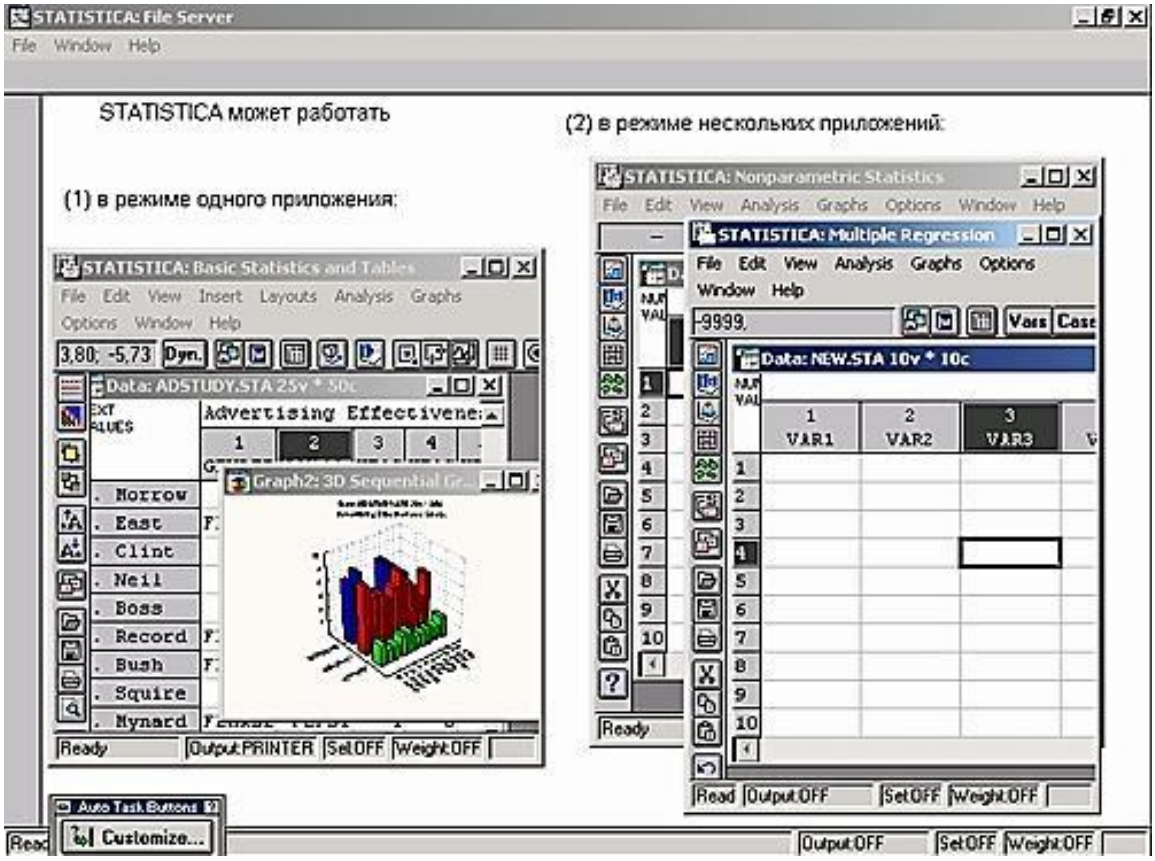


Рис. 6.14. Режимы окон в системе STATISTICA

Интерактивный анализ данных в STATISTICA.

Система не требует, чтобы пользователь еще до проведения анализа указал всю информацию, которую следует вывести на экран. Ведь анализ даже простого плана может породить большое число таблиц результатов и просто необозримое количество графиков, поэтому при проведении реального анализа, до изучения основных результатов, трудно представить, какие графики или таблицы следует анализировать в первую очередь. Именно поэтому STATISTICA предоставляет пользователю возможность выбрать определенные типы вывода и интерактивно провести последовательные сравнения и моделирующий анализ уже после того, как данные обработаны и получены основные результаты. Количество выводимых окон также может быть настроено, чтобы не перегружать экран компьютера.

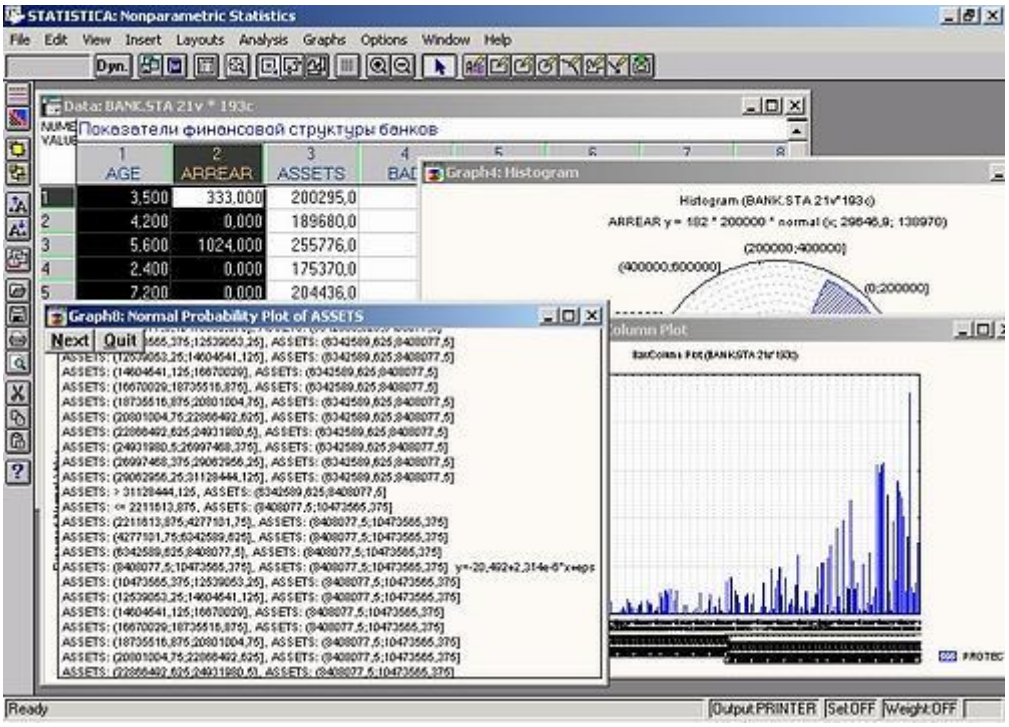


Рис. 6.15. Пример работы с данными в системе STATISTICA

Гибкие вычислительные процедуры STATISTICA и широкий выбор методов графического представления данных любого типа открывают перед пользователем безграничные возможности проведения разведочного анализа и проверки статистических гипотез.

Справочная система и интерактивное (электронное) руководство.

Чтобы получить дополнительную информацию о некоторых функциях системы, нажмите клавишу справки (F1), когда выделена соответствующая команда или пункт меню. STATISTICA содержит *Электронное руководство* − справочную информацию по всем процедурам и функциям программы, доступную в контекстно-зависимом режиме при нажатии клавиши F1 или кнопки справки в строке заголовка всех диалоговых окон (справочник содержит свыше 10 мегабайт документации в сжатом виде). Справку также можно вызвать двойным щелчком на поле сообщений строки состояния в нижней части окна приложения STATISTICA.

Статистический советник.

Статистический советник представляет собой интерактивную справочную систему. После выбора пункта *Советник* из выпадающего меню (*Справка*) программа задаст вам несложные вопросы о характере решаемой проблемы и типе исходных данных, а затем предложит список наиболее подходящих процедур (и объяснит, где их найти в системе STATISTICA).

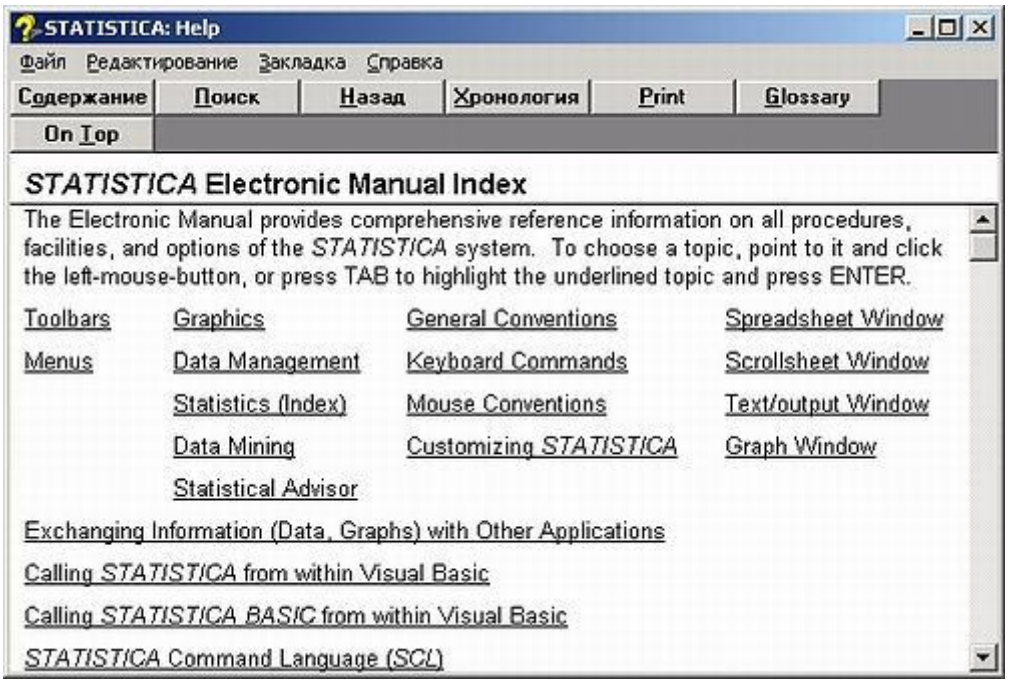


Рис. 6.16. Интерактивная справочная система

С помощью гиперссылок можно непосредственно перейти из раздела *Статистический советник* к подробному описанию соответствующих статистических методов и процедур в разделе *Вводный обзор*.

Автоматические отчеты и автоматическая распечатка таблиц результатов.

Независимо от того, происходит обработка в пакетном режиме или интерактивно запрашивается пользователем, может быть выбран режим вывода *Автоотчет*. Этот режим позволяет автоматически, без каких-либо действий со стороны пользователя распечатывать (или направлять в окно отчета или в файл) содержание всех окон вывода, которые получаются в процессе анализа.

Режим автоматического вывода каждой строящейся на экране таблицы результатов и/или графика может оказаться полезным не только для создания полного отчета о результатах анализа, но и при разведочном анализе данных, когда возникает необходимость вернуться к предыдущему шагу и просмотреть результаты, полученные на ранних этапах обработки данных. Для этого всю выходную информацию (таблицы результатов и графики) можно направить во временное *Окно текста/вывода* с прокруткой и уже затем в случае необходимости сохранить ее, распечатать или скопировать в файл текстового редактора.

Функции импорта файлов.

Файлы данных из приложений Windows и других операционных систем также можно переводить в формат системы STATISTICA с по-

мощью функций импорта файлов, которые включают доступ ко всем базам данных (через поддержку метода ODBC), а также возможности импорта форматированных текстовых файлов и текстовых файлов свободного формата (ASCII). Импорт файлов без использования буфера обмена имеет свои преимущества:

- он позволяет пользователю точно указать, как должен проводиться импорт (например, выбирать из файлов диапазоны значений, импортировать или не импортировать имена переменных, текстовые значения и имена наблюдений и указывать способ их интерпретации);
- он предоставляет пользователю доступ к типам данных, которые недоступны (или труднодоступны) при операциях с буфером обмена (например, длинные метки значений или специальные коды пропущенных данных).

STATISTICA поддерживает соглашения динамического обмена данными (DDE), что позволяет динамически связывать диапазон данных в таблице исходных данных с набором данных других приложений (Windows). Связи DDE (динамического обмена данными) можно установить между файлом-источником (сервером), например электронной таблицей MS Excel, и файлом данных системы STATISTICA (файлом-клиентом), так что при вынесении изменений в файл-источник данные в соответствующей части таблицы исходных данных STATISTICA (файле-клиенте) будут автоматически обновляться.

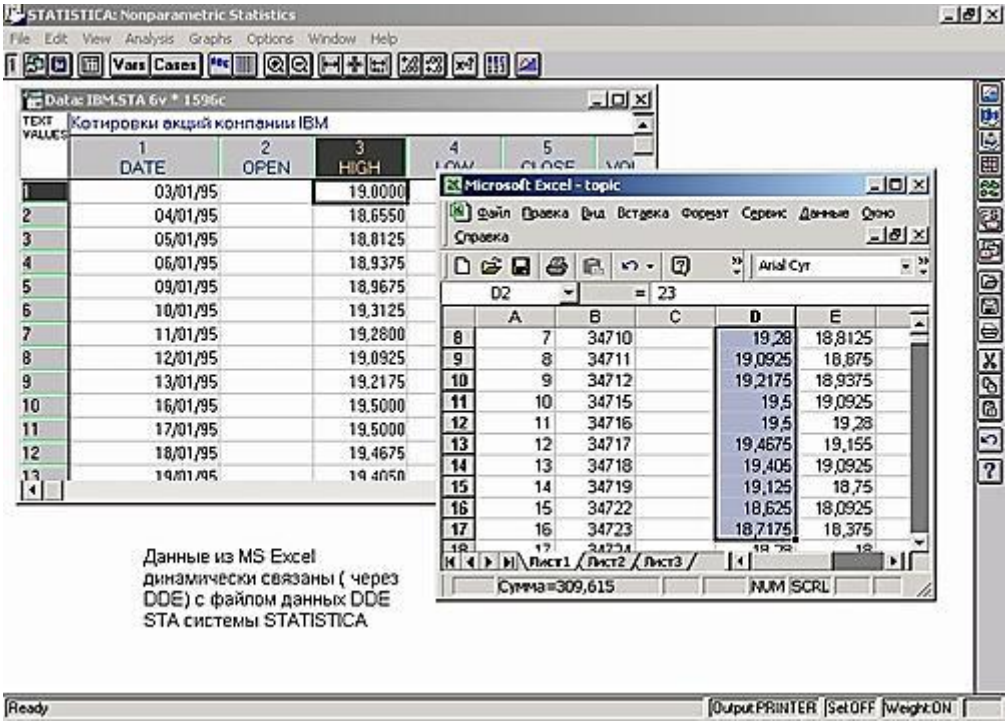


Рис. 6.17. Динамический обмен данными между STATISTICA и Excel

Обычно два файла динамически связываются в промышленных установках, когда к последовательному порту компьютера, на котором находится файл данных системы STATISTICA, подключено измерительное устройство, например, для ежечасного автоматического обновления определенных измерений.

Связи DDE можно установить с помощью команды *Установить связь* выпадающего меню *Правка таблицы* исходных данных или введя определение связи в поле *Длинное имя* (метка, формула, связь) диалогового окна спецификаций переменной.

6.1.4. Командный язык STATISTICA

STATISTICA содержит два встроенных языка программирования: BASIC и SCL (командный язык). Оба языка предназначены для работы в среде STATISTICA и содержат встроенные операции для обращения к таблицам исходных данных, таблицам результатов и графическим функциям.

Язык STATISTICA BASIC представляет собой простой и одновременно достаточно мощный язык программирования. С его помощью можно создать широкий спектр приложений, начиная от простых программ преобразования данных и заканчивая сложными пользовательскими процедурами комплексного анализа и вывода информации.

Этот язык программирования пригоден для решения больших вычислительных задач, поскольку обрабатываемые массивы данных могут иметь до 8 измерений и нет ограничений на размеры массивов. Таким образом, пользователь может использоваться всю доступную память и создавать процедуры, включающие операции с большими многомерными матрицами.

Встроенный язык BASIC доступен в любой момент анализа вместе с интегрированной средой, которая позволяет писать, редактировать, проверять, отлаживать (предварительно прогонять) и выполнять программы. Как обычный язык программирования, поддерживает циклические операции и условные переходы, функции и подпрограммы, а также работу с динамическими библиотеками (DLL). В то же время он «понимает» структуру файлов данных системы STATISTICA и позволяет организовать интерактивную обработку данных в среде самой системы с помощью пользовательских диалоговых окон. С помощью этого языка пользователь может создавать свои собственные сложные программы анализа данных, одновременно используя готовые алгоритмы расчетов и построения графиков, предусмотренные в системе STATISTICA.

Командный язык SCL (STATISTICA Command Language) предназначен для организации пакетной обработки данных и создания собственных при-

ложений на основе процедур, содержащихся в системе STATISTICA. Для того чтобы пользователь мог при этом реализовать собственные алгоритмы расчетов, предусмотрена возможность интеграции языков BASIC и SCL.

Программы, написанные на встроенных языках системы STATISTICA, доступны в любом модуле системы и на любом этапе анализа данных, при этом их можно вызывать и выполнять как с помощью кнопок автозадач, так и непосредственно из окна редактирования. Пользователь также имеет возможность создавать собственные библиотеки функций и подпрограмм и таким образом значительно расширять предлагаемый набор процедур обработки данных и представления результатов.

STATISTICA может работать в «истинном» пакетном режиме как система, управляемая командами, с помощью встроенного языка управления приложениями, доступного в любом модуле системы из выпадающего меню *Анализ*. Можно ввести последовательность команд для выполнения определенных действий, а затем сколько угодно раз исполнять ее в пакетном режиме.

Возможен и другой способ действий − использование диалогового окна *Мастер команд* для быстрого выбора и ввода требуемого списка команд (рис. 6.18).

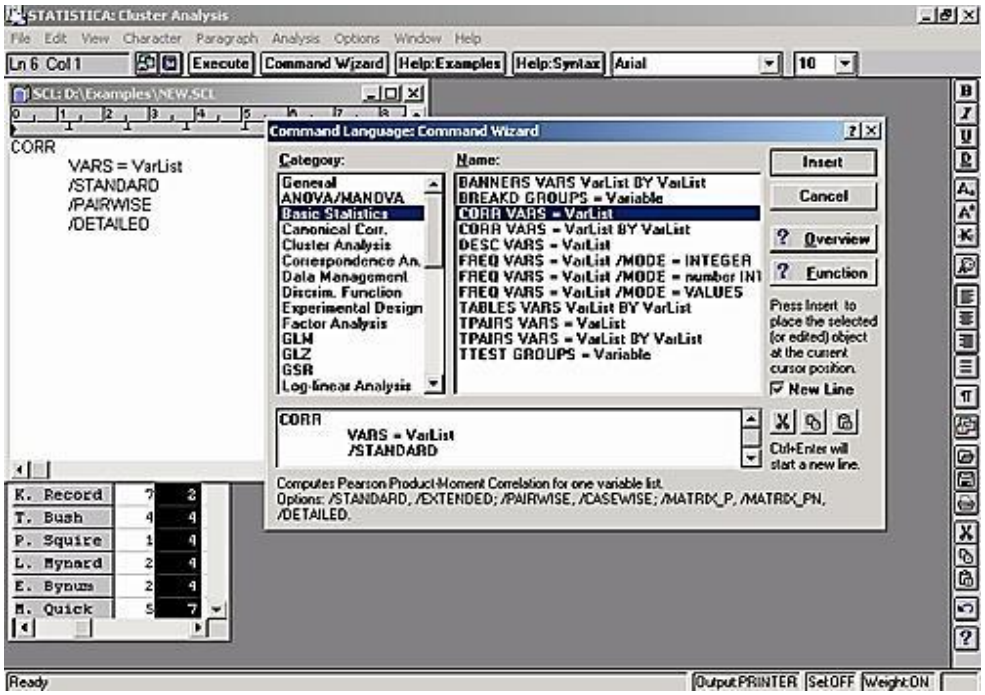


Рис. 6.18. Окно «Мастер команд»

Для написания и отладки «пакетов» команд используется интегрированная среда языка SCL. Она включает текстовый редактор, совмещенный с окном *Мастер команд*, систему помощи по синтаксису языка с примерами и интегрированные средства проверки правильности программ.

Несмотря на то что в командном языке SCL не заложен в непосредственном виде специальный пользовательский интерактивный интерфейс, для этих целей можно использовать программы на языке BASIC, вызываемые из SCL-программ, например, для создания диалоговых окон, позволяющих выбирать переменные, файлы данных и т. п. в ходе выполнения программы.

Командный язык содержит специальный *Исполняемый модуль*, позволяющий разрабатывать приложения «под ключ», которые вызываются двойным щелчком на значке соответствующего «пользовательского приложения» на рабочем столе Windows.

Эта возможность позволяет экономить время пользователя, когда многократно повторяется одна и та же процедура или последовательность процедур анализа, а также дает возможность использовать SCL-программы пользователями, которые не знакомы с соглашениями системы STATISTICA.

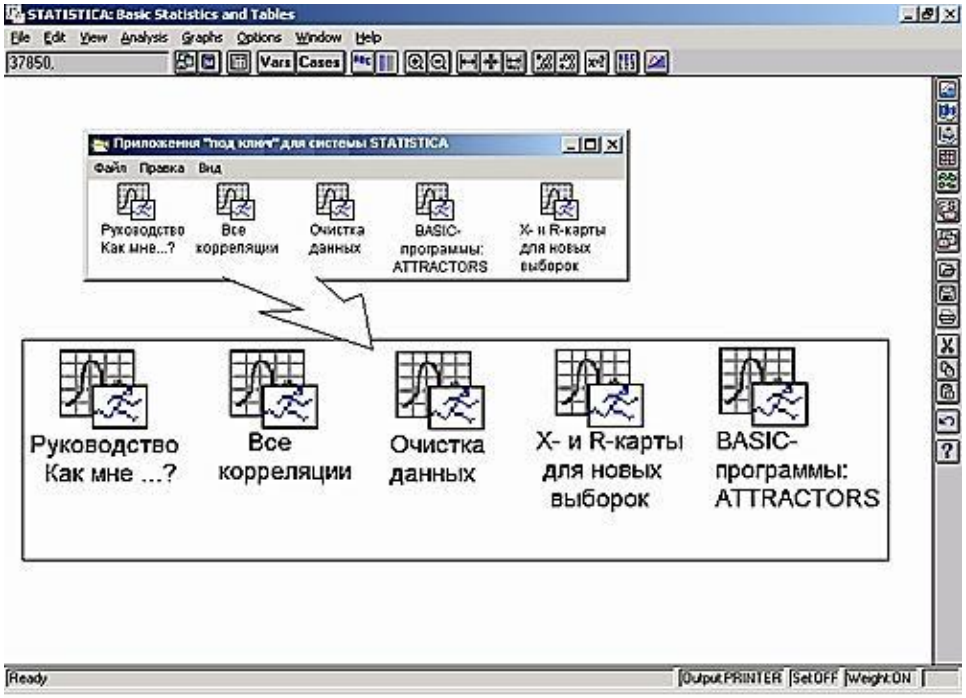
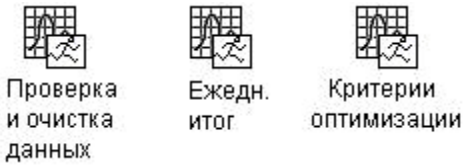


Рис. 6.19. Приложения, создаваемые на рабочем столе

Чтобы создать такое приложение «под ключ», сначала нужно написать саму SCL-программу и сохранить ее обычным образом (например, в файле Program 1.sct). Затем в окне *Диспетчер программ* системы Windows нужно создать пиктограмму для исполняемого модуля с именем Sta_run.exe (оно находится в папке STATISTICA на диске).

В поле команд нужно задать имя SCL-программы, подлежащей исполнению (например, d:\data\program1.scl'). Теперь при щелчке мышью на этом значке будет начинаться выполнение программы (в данном случае Program1.se!). Описанным способом можно создать любое количество пользовательских приложений, а с помощью окна *Диспетчер программ* дать им содержательные имена, соответствующие тем задачам анализа данных, которые эти приложения выполняют.



6.1.5. Построение корреляционной матрицы на примере анализа показателей деятельности нефтехимического предприятия

Проведем графический анализ полученной матрицы. В качестве исходных данных будем использовать данные из файла *data.sta*.

Имеется система переменных $Y_1 ... Y_3$, $X_4 ... X_{17}$. Рассматриваются следующие показатели:

Y 1 – производительность труда;
Y 2 – индекс снижения себестоимости продукции;
Y 3 – рентабельность;
X 4 – трудоемкость единицы продукции;
X 5 – удельный вес рабочих в составе ППП;
X 6 – удельный вес покупных изделий;
X 7 – коэффициент сменности оборудования;
X 8 – премии и вознаграждения на одного работника;
X 9 – удельный вес потерь от брака;
X 10 – фондоотдача;
X 11 – среднегодовая численность ППП;
X 12 – среднегодовая стоимость ОПФ;
X 13 – среднегодовой фонд заработной платы;
X 14 – фондовооруженность труда;
X 15 – оборачиваемость нормированных оборотных средств;
X 16 – оборачиваемость ненормированных оборотных средств;
X 17 – непроизводственные расходы.

В модуле *Основные статистики* легко можно вычислить и проанализировать корреляционную матрицу выбранных вами переменных. Для начала проведем корреляционный анализ переменных *Y*1, *X*4, *X*5 и *X*6.

- Запустите программу STATISTICA. Переключитесь в модуль *Основные статистики*. Нажмите кнопку *Open data* (Открыть файл данных) и откройте файл *data.sta*.

• В стартовой панели модуля *Основные статистики* выберите пункт *Correlation matrices* (Корреляционные матрицы). Дважды щелкните по ней либо высветите и нажмите кнопку *OK*. На экране появится окно *Pearson Product-Moment Correlation* (Корреляция Пирсона) (рис. 6.20).

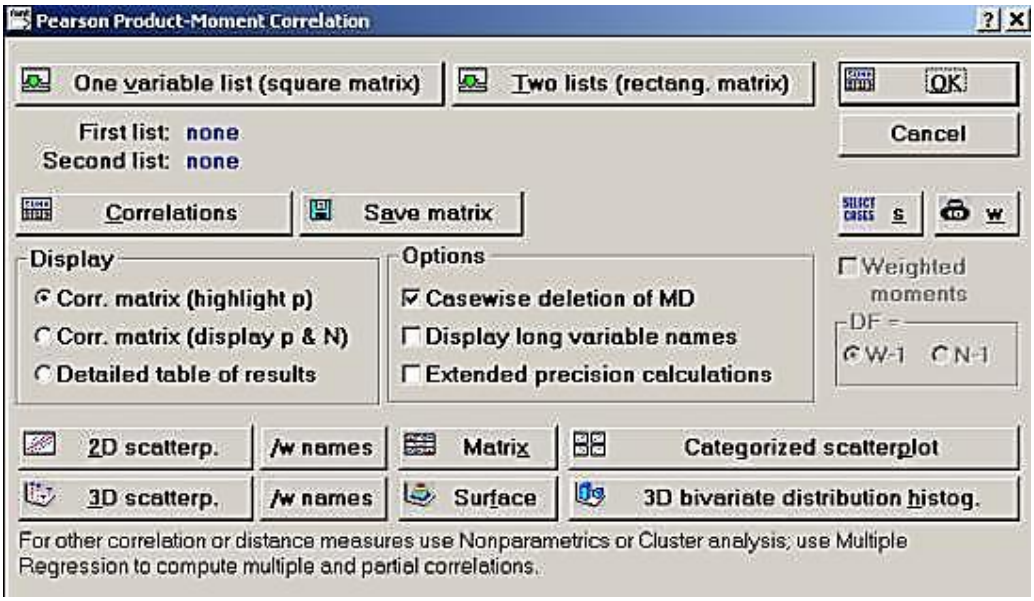


Рис. 6.20. Окно корреляции Пирсона

• Нажмите на кнопку *Two lists* (Два списка). После чего откроется окно выбора переменных. Выберите переменные, как показано на рис. 6.21. Таким образом, мы определили два списка переменных: *X*4–*X*6 – *First variables list* (Первый список переменных) и *Y*1 – *Second variables list* (Второй список переменных). Мы хотим подсчитать корреляции между переменной *Y*1 и переменными *X*4−*X*6. Щелкните по кнопке *OK* для подтверждения вашего выбора и возврата к окну *Pearson Product-Moment Correlation.*

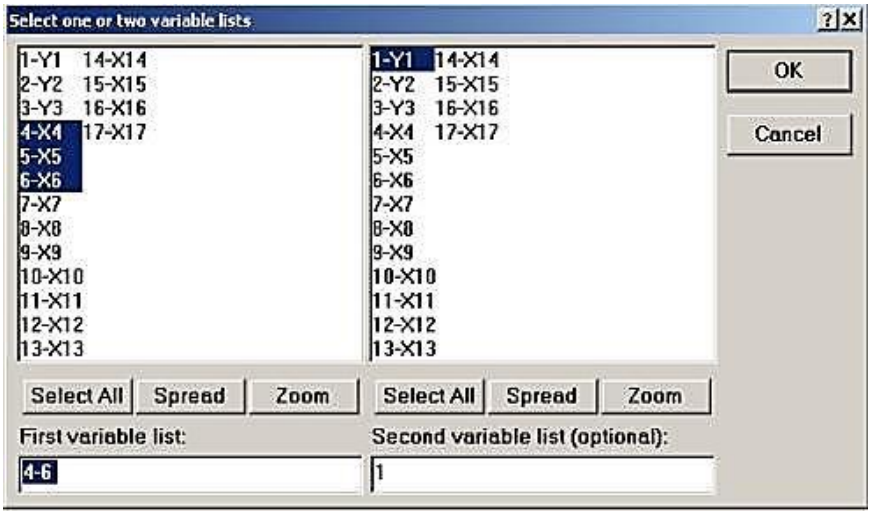


Рис. 6.21. Выбор списка переменных для нахождения корреляции

• В окне *Pearson Product-Moment Correlation* нажмите кнопку *OK*. На экране вы увидите корреляционную матрицу (рис. 6.22).

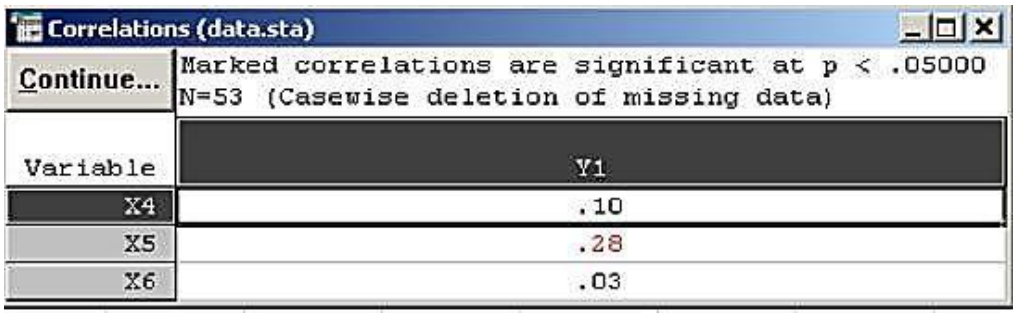


Рис. 6.22. Корреляционная матрица

В этой матрице имеется только один столбец, так как во втором списке мы выбрали только одну переменную. В столбце даны коэффициенты корреляции между переменной *Y*1 и *X*4–*X*6. В нашей корреляционной матрице красным цветом автоматически выделены коэффициенты для уровня $p < 0{,}05$. Именно на эти коэффициенты следует обратить наибольшее внимание. Грубо говоря, зависимость между переменными с выделенными красным цветом коэффициентами корреляции наиболее значимая. В нашем случае переменная *Y*1 наиболее зависима от переменной *X*5. Коэффициент корреляции между этими переменными равен 0,28. Так как 0,28 > 0, то мы можем считать, что при возрастании переменной *X*5 переменная *Y*1 также возрастает. Рассмотрим эти переменные более внимательно. Полезно просмотреть зависимость между переменными *Y*1 и *X*5 графически.

• В окне корреляционной матрицы нажмите на кнопку *Continue* (Продолжить). После чего вы вернетесь в окно *Pearson Product-Moment Correlation*. Нажмите на кнопку *Two lists* (Два списка). После чего откроется окно выбора переменных (в данном случае для графика). Выберите переменные *X*5 и *Y*1 (рис. 6.23) и нажмите кнопку *OK*.

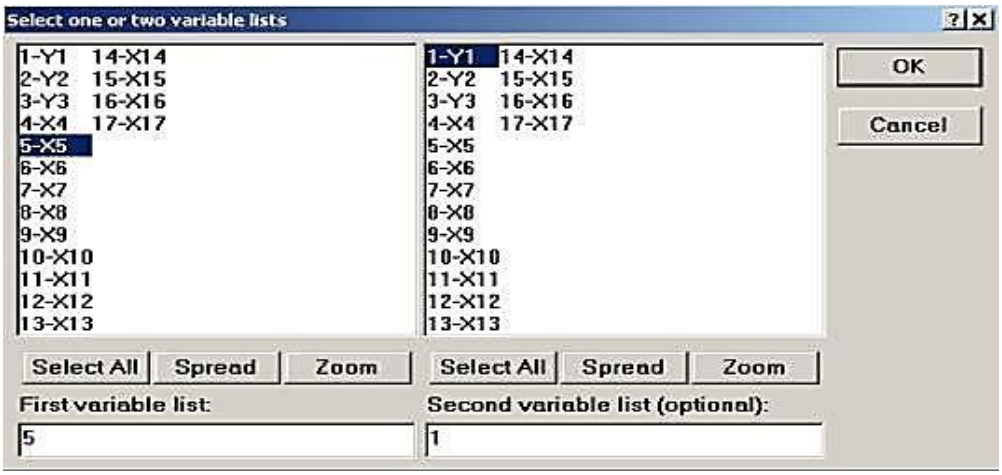


Рис. 6.23. Окно выбора переменных для построения графической зависимости

• В окне *Pearson Product-Moment Correlation* нажмите кнопку *2D-scatterplot* (2D-диаграмма рассеяния). После этого появится окно диаграммы рассеяния (рис. 6.24) для выбранных переменных.

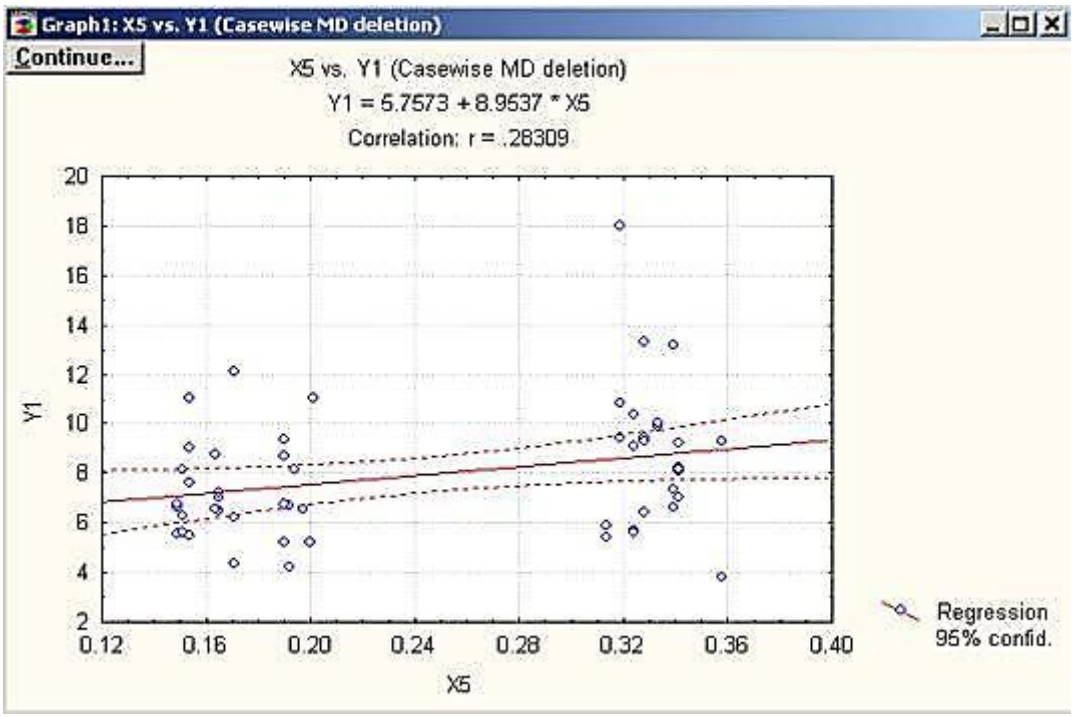


Рис. 6.24. Диаграмма рассеяния для выбранных переменных

• Из графика (рис. 6.24) отчетливо видно, что зависимость не является линейной – прямая очень плохо «ложится» на данные. На графике дана лучшая прямая. Как бы мы ни меняли коэффициент наклона, подгонка будет только хуже. STATISTICA предлагает возможности, которые позволяют провести углубленное рассмотрение данных.

• Выберите средство *Кисть*, щелкнув по кнопке на инструментальной панели вверху. Перед вами справа появится панель *Brushing* (Кисть). На панели *Brushing* сделайте установки, как показано на рис. 6.25.

Рис. 6.25. Настройки инструмента «Кисть»

• Войдите в график (просто щелкните по любой точке в его пространстве, сделав тем самым график активным) и отметьте лассо точки, которые, с вашей точки зрения, наиболее сильно отклоняются от прямой на графике. Мы выделяем точки с помощью лассо (обводя их карандашом, как бы захватывая лассо). Ранее была отмечена опция *Lasso* (Лассо) на панели инструментов *Кисть*. Выбрав опцию *Point* (Точка), мы удаляли бы точки последовательно одну за другой. Выберите, например, точки над прямой, как показано на рис. 6.26.

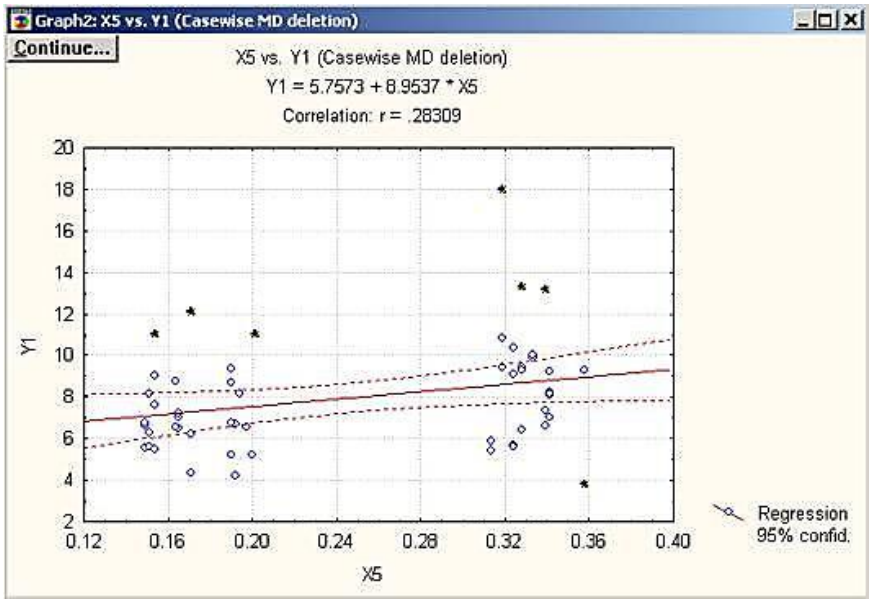


Рис. 6.26. Удаление точек, наиболее отклонившихся от прямой

• Щелкните далее на кнопку *Update* (Обновить) на панели *Brushing*. Вы увидите следующий график (рис. 6.27).

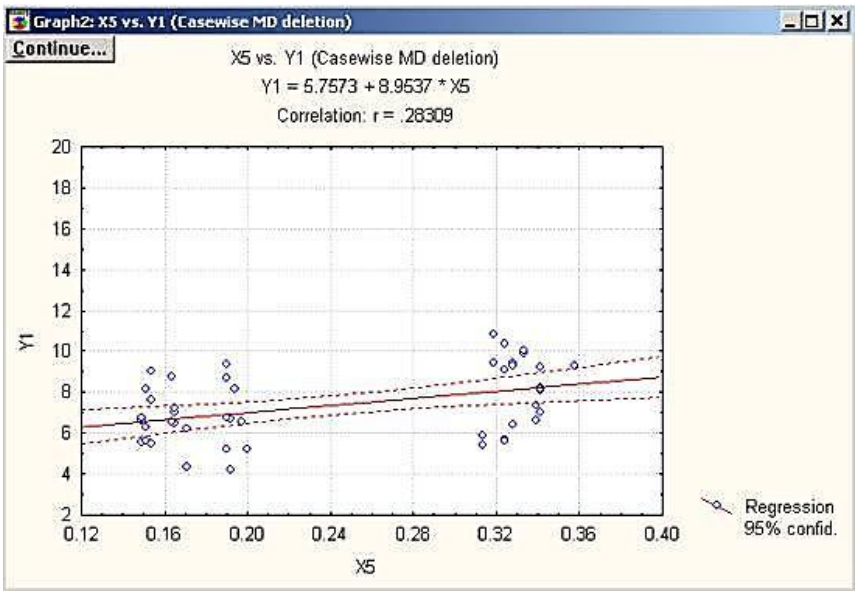


Рис. 6.27. Обновленная диаграмма рассеяния

Теперь данные лучше ложатся на прямую. Вы можете продолжить исследование. Вполне может оказаться, что в исключительных случаях имеется некоторая закономерность. Безусловно, эти закономерности стоит исследовать дополнительно. Вы легко можете определить, какие случаи были удалены вами. Для этого можно воспользоваться кнопкой *Label* (Метка).

- Щелкните на кнопку *De-select All* (Отменить выбор всех) вверху панели *Brush*. Пометьте опцию *Label* на панели *Brush* и вновь захватите лассо нужные точки. Далее нажмите кнопку *Update*, и вы увидите на экране график, в котором рядом с выделенными точками появились имена случаев, к которым они относятся (рис. 6.28).

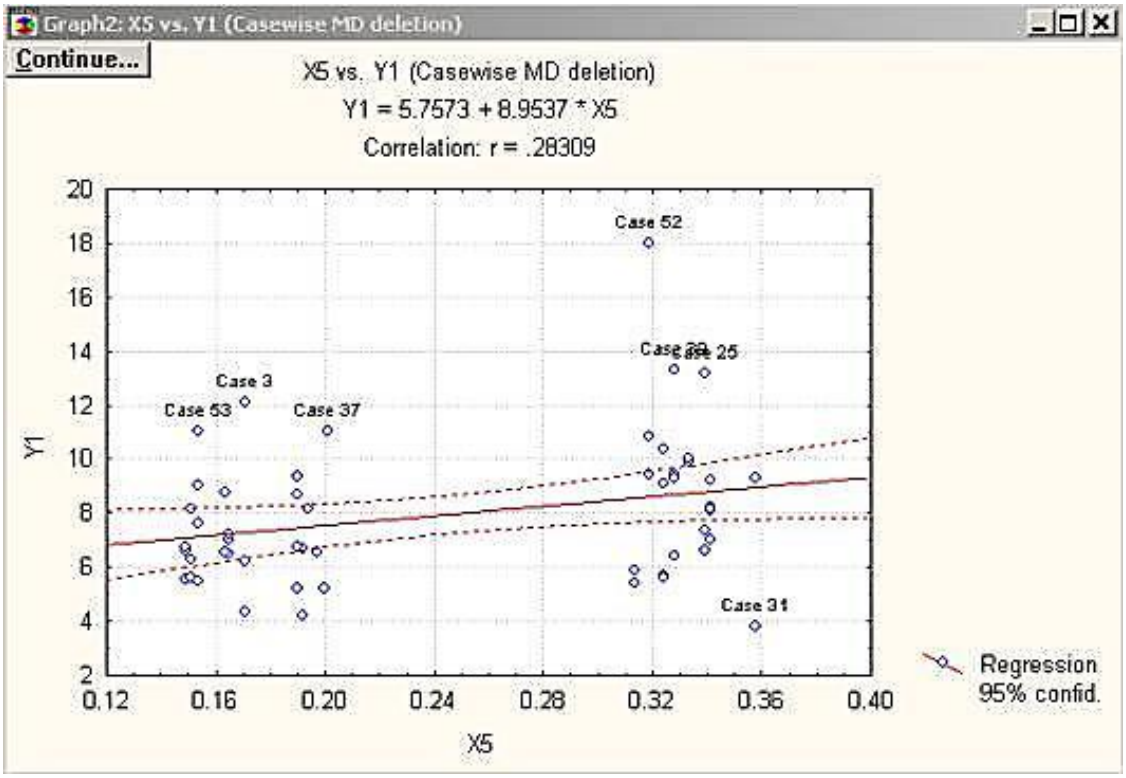


Рис. 6.28. Диаграмма рассеяния с отметкой удаленных случаев

Эти случаи как раз и требуют дополнительного исследования. Например, исключение их из рассмотрения может привести к значительному изменению исследуемого коэффициента корреляции. В том случае, если в корреляционной матрице имеются несколько высвеченных коэффициентов корреляции, то далее вам следует рассмотреть данные с другими высвеченными коэффициентами корреляции, построить графики зависимости, поработать с инструментом *Кисть*. Коэффициенты корреляции хорошо подходят для описания линейных связей и плохо, если зависимость между переменными не линейная. Вы можете про-

смотреть корреляционную матрицу «графически» с помощью кнопки *Matrix* (Матричный график) в окне *Pearson Product-Moment Correlation.*

- Вернитесь в окно *Pearson Product-Moment Correlation* нажатием на кнопку *Continue*, расположенную на панели графика.
- Щелкните по кнопке *One variable list* (Один список переменных). Выберите переменные *Y*1, *X*4–*X*6, как показано на рис. 6.29, а затем нажмите кнопку *OK*.

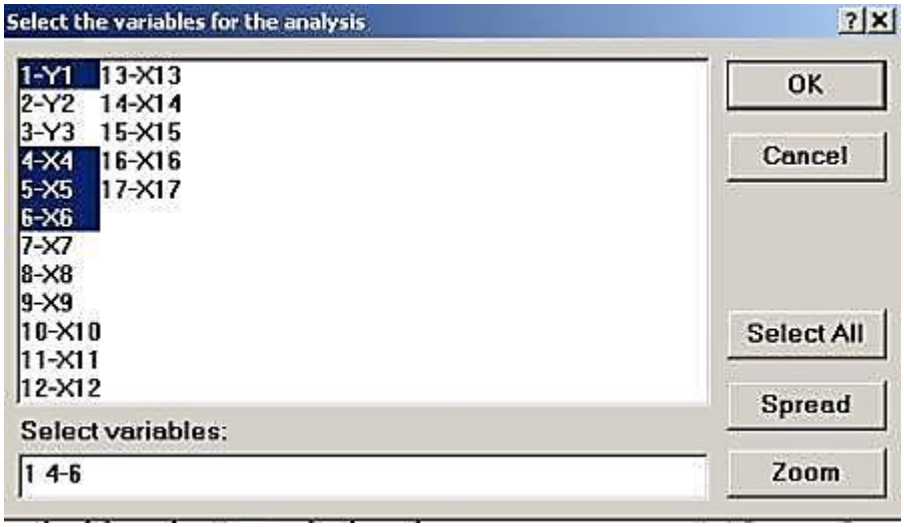


Рис. 6.29. Выбор данных для построения графической корреляционной матрицы

- Далее в окне *Pearson Product-Moment Correlation* нажмите кнопку *Matrix*. После этого откроется окно выбора переменных для построения графиков (рис. 6.30).

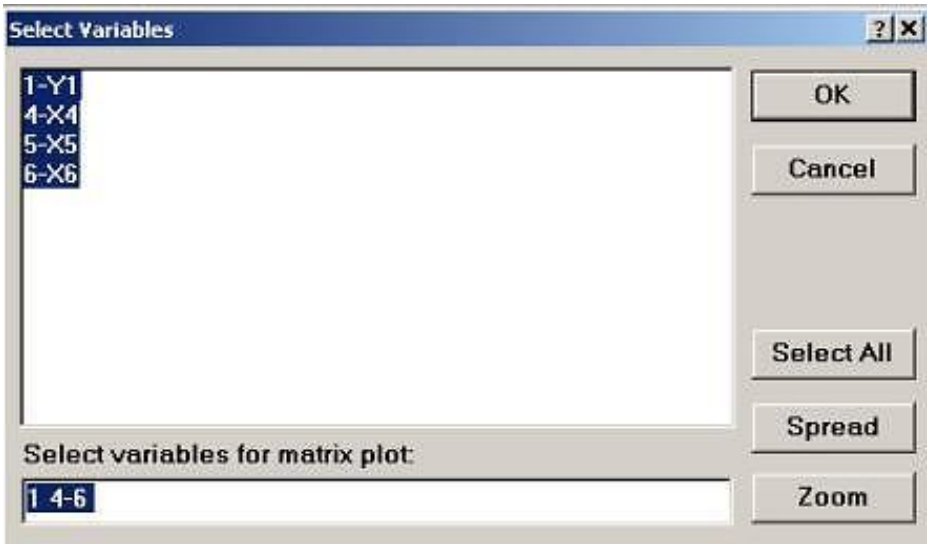


Рис. 6.30. Выбранные данные для построения графиков

• В окне выбора переменных выберите все переменные нажатием на кнопку *Select All* (Выбрать все). Подтвердите свой выбор, нажав кнопку *OK*. На экране появится корреляционная матрица в графическом виде, позволяющая оценить линейные связи визуально (рис. 6.31).

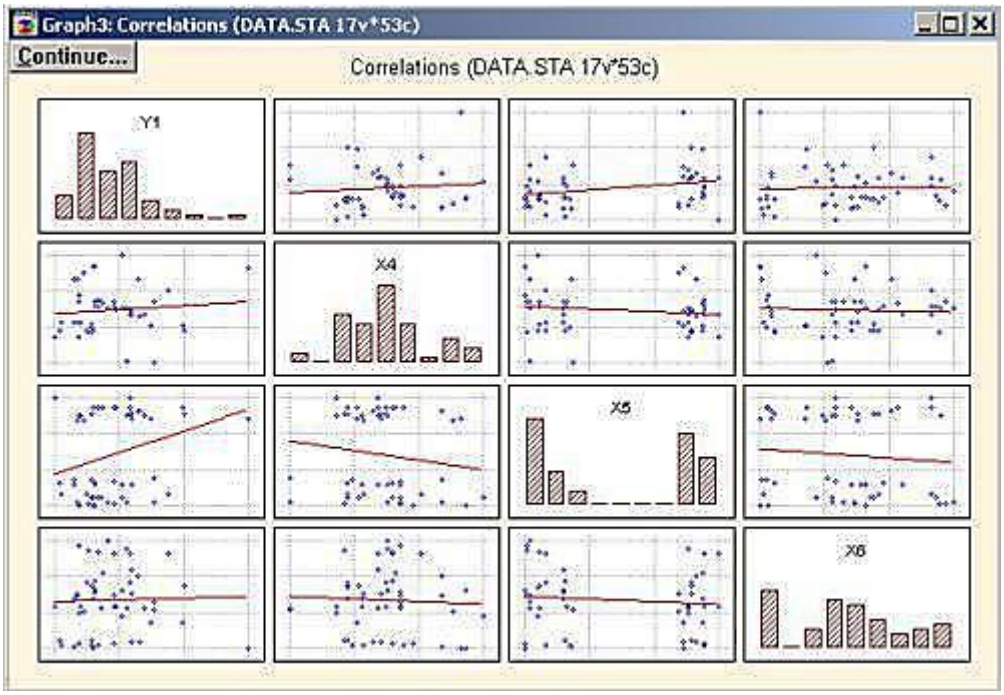


Рис. 6.31. Корреляционная матрица в графическом виде

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Васильев Ф.П. Численные методы решения экстремальных задач / Ф.П. Васильев. − Москва : Наука, 2008.

2. Сеа Ж. Оптимизация. Теория и алгоритмы / Ж. Сеа. − Москва : Мир, 2011.

3. Лесин В. В. Основы методов оптимизации / Лесин В.В., Ю.П. Лисовец. – Москва : Изд-во МАИ, 2009.

4. Бочкарев В.В. Оптимизация химико-технологических процессов : учебное пособие / В.В. Бочкарев. – Томск : Изд-во ТПУ, 2014. – 264 с.

5. Возможности STATISTICA. – Режим доступа : http://www.statsoft.ru/#tab-STATISTICA-link.

6. Руководство пользователя STATISTICA 5.1. – Режим доступа : http://www.exponenta.ru/soft/Statist/stat5_1/1/1.asp

7. Харченко М.А. Корреляционный анализ : учебное пособие для вузов / М.А. Харченко. – Воронеж : ВГУ, 2008. – 31 с

8. Елисеева И.И. Статистика : учебник / И.И. Елисеева. – Москва : Крокус, 2008.

9. Орлов А.И. Прикладная статистика / А.И. Орлов. – Москва : Экзамен, 2004.

Учебное издание

ДОПОЛНИТЕЛЬНЫЕ ГЛАВЫ МАТЕМАТИКИ. СТАТИСТИЧЕСКИЙ АНАЛИЗ

Учебное пособие

Составители

КРИВЦОВА Надежда Игоревна
МОЙЗЕС Ольга Ефимовна

Корректура *С.Н. Карапотин*
Компьютерная верстка *В.П. Аршинова*
Дизайн обложки *Т.В. Буланова*

Подписано к печати 30.12.2015. Формат 60x84/16. Бумага «Снегурочка».
Печать XEROX. Усл. печ. л. 4,94. Уч.-изд. л. 4,47.
Заказ 541-15. Тираж 100 экз.

Издательство
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ