Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Уфимский государственный нефтяной технический университет»

Филиал ФГБОУ ВО УГНТУ в г. Салавате

Кафедра «Информационных технологий»

Математическое моделирование в задачах нефтегазовой отрасли

Салават
2019

Учебно-методическое пособие «Математическое моделирование в задачах нефтегазовой отрасли» для практических занятий и самостоятельной работы содержит теоретические основы и задачи статистики и метода аппроксимации эмпирических данных с использованием математического пакета Mathcad 2000, модуль “Планирование эксперимента” статистической графической системы STATGRAPHICS Plus for Windows.

В практических заданиях предлагается разработка плана полного факторного эксперимента, создание и анализ поверхности отклика, разработка и анализ экспериментального плана. Приведен список рекомендуемой литературы.

Разработано автором с целью оказания помощи магистрантам направления 09.04.01 «Информатика и вычислительная техника» в подготовке к практическим занятиям, выполнения самостоятельной и контрольных работ по дисциплине «Математическое моделирование в задачах нефтегазовой отрасли».

Публикуется в авторской редакции.

Составители: Ефимова Г.Ф., канд. физ.-мат. наук, доц. каф. ИнТех

Рецензенты: Ишмухаметова А.А., канд. физ.-мат. наук, доц. каф. ИнТех

Минлибаев М. Р., канд. физ.-мат. наук, доц. каф. ЭАПП

Содержание

ВВЕДЕНИЕ

Любой процесс есть комбинация материалов, методов, технологий, людей, оборудования, измерительных приборов и т. п., совместная работа которых обеспечивает производство товаров или выполнение определенных заданий.

Планирование эксперимента является научным подходом, позволяющим экспериментатору лучше разобраться в происходящих процессах, определить взаимосвязи между входными и выходными параметрами и сделать надежные выводы.

Планирование эксперимента - раздел математической статистики, изучающий рациональную организацию измерений и наблюдений.

Целью эксперимента служит либо оценка параметров распределения некоторой случайной функции, либо проверка некоторых гипотез о параметрах.

Исходя из цели, формулируется критерий оптимальности плана эксперимента, под которым понимается совокупность значений исследуемых переменных. Оптимальный план организации позволяет уменьшить количество опытов, сократив тем самым расходы на их проведение и временные затраты, уменьшить ошибку эксперимента, выработать четкие формализованные правила принятия решений на каждом этапе проведения эксперимента и получить многофакторные математические модели с желаемыми статистическими свойствами. Предлагаемое учебное пособие состоит из двух частей, не связанных между собой, однако взаимодополняющих друг друга в рамках изучаемого курса. В первой части рассматриваются основные задачи статистики и метода аппроксимации эмпирических данных с использованием математического пакета Mathcad 2000. Лабораторные работы этой части включают в себя: работу с выборками (числовые характеристики); построение гистограмм, полигонов частот; оценку функции распределения; аппроксимацию эмпирических данных методом наименьших квадратов,

Во второй части рассматривается модуль “Планирование эксперимента” статистической графической системы STATGRAPHICS Plus for Windows. Лабораторные работы этой части включают в себя разработку плана полного факторного эксперимента. создание и анализ поверхности отклика, разработку и анализ экспериментального плана.

Для выполнения практических работ, предлагаемых в учебном пособия, от читателя не требуется каких-либо начальных знаний и умений работы с математическим пакетом Mathcad и статистической графической системой Statgraphic.

ОБ ИСТОРИИ ПРИКЛАДНОЙ СТАТИСТИКИ

Типовые примеры раннего этапа применения статистических методов описаны в Ветхом Завете (см.. например, Книгу Чисел). С математической точки зрения, они сводились к подсчетам числа попаданий значений наблюдаемых признаков в определённой градации. В дальнейшем результаты стали представлять в виде таблиц и диаграмм, как это и сейчас делает Госкомстат РФ. Надо признать, что по сравнению с Ветхим Заветом есть прогресс - в Библии не было таблиц.

Сразу после возникновения теории вероятностей (Паскаль,Ферма,XVII век)вероятностные модели стали использоваться при обработке статических данных. Например, изучалась частота рождения мальчиков и девочек , было установлено отличие вероятностей рождения мальчиков от 0,5;анализировались причины того, что в парижских приютах эта вероятность не та, что в самом Париже и т.д.

В 1794 г.(по другим данным –в 1795)К.Гаусс разработал метод наименьших квадратов , один из наиболее популярных ныне статических методов, и применил его при расчете орбиты астероида Церера. В XIX веке заметный вклад а развитии практической статистики внес бельгиец Кетле, на основе анализа большого числа реальных данных показавший устойчивость относительных статических показателей , таких как доля самоубийств среди вех смертей. Интересно, что основные идеи статического приемочного контроля и сертификации продукции обсуждались академиком Буняковским и применялись в российской армии еще в середине 19 в. Статические методы управления качеством, сертификации и классификации продукции сейчас весьма актуальны.

Современный тп развития прикладной статистики можно отчитывать с1900 г;. когда англичанин К.Пирсон основал журнал "Biometrika". Первая треть XX в прошла под знаком параметрической статистики. Изучались методы, основанные на анализе данных из параметрических семейств распределений, описываемых кривыми семейства Пирсона. Наиболее популярным было нормальное (гауссово) распределение. Для проверки гипотез использовались критерии Пирсона. Стьюдента , Фишера. Были предложены метод максимального правдоподобия, дисперсионный анализ, сформулированы основные идеи **планирования эксперемента.**

Разработанную в первой трети XX в. теорию называют параметрической машинкой. поскольку ее основной объект изучения - это выборки из распределений, описываемых одним или небольшим числом параметров. Наиболее общим является семейство кривых Пирсона, задаваемых четырьмя параметрами Как правило, нельзя указать каких-либо веских причин, по которым конкретное распределение результатов наблюдений должно входить в то или иное параметрическое семейство. Исключения хорошо известны: если вероятностная модель предусматривает суммирование независимых случайных величин, то сумму естественно описывать нормальным распределением; если же в модели рассматривается произведение таких величин, то итог, видимо, приближается логарифмически нормальным распределением и т.д.

Управление качеством играет сейчас ключевую роль в область промышленного серийного производства. Так, во многих отраслях промышленности предприятия получают прибыль благодаря тому, что они научились сводить к нулю долю бракованных экземпляров в общем объеме производимой продукции.

Mathacad обладает такими мощными средствами для решения подобных задач, основанных на статистических механизмах контроля, что им стоит посвятить отдельную главу.

ИСПОЛЬЗУЕМЫЕ ИНСТРУМЕНТЫ MATHCAD

MathCAD имеет стандартный интерфейс *Windows*.

- Строка меню.
- Строка инструментов.
- Строка форматирования.
- Рабочая область.
- Строка состояния.
- Всплывающее или контекстное меню (нажимается правая кнопка мыши), содержание зависит от места вызова.
- Панель инструментов **Математика** и доступные из нее инструменты.

Среди особых элементов интерфейса следует отметить панель инструментов **Математика** (рис. 32)**.** Эта панель служит для доступа к панелям инструментов, обеспечивающих вставку математических вычислений или символов. При необходимости панели инструментов можно установить**:** ***View – Toolbars – v Resources***.

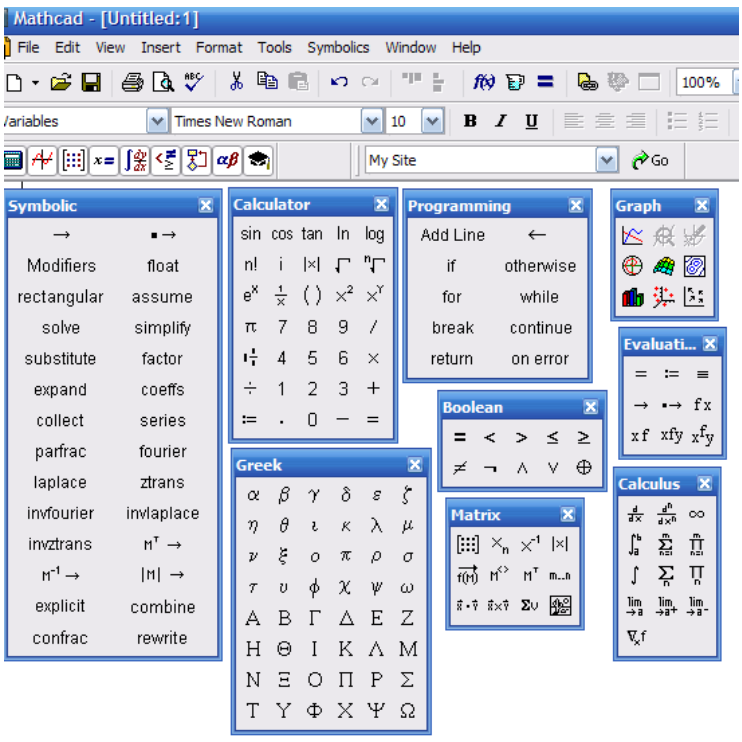


Рис. 32. Панель инструментов **Математика** и доступные из нее инструменты

- Панель ***Calculator*** служит для вставки основных математических операций.
- Панель ***Graph*** служит для вставки графика в документ.
- Панель ***Matrix*** служит для вставки матрицы, для работы с матрицами и матричными операциями.
- Панель ***Evaluation*** представляет операторы вычисления.
- Панель ***Calculus*** представляет операторы интегрирования, дифференцирования, суммирования, ..
- Панель ***Boolean*** представляет булевы операторы и предназначена для вставки логических или булевых операций.
- Панель ***Programming*** служит для программирования средствами *MathCad* .
- Панель ***Greek*** представляет греческие символы.
- Панель ***Symbolic*** служит для вставки символьных операторов.

СТАНДАРТНЫЕ ФУНКЦИИ MATHCAD

Для решения уравнения (1.1) в MathCAD служит функция root, реализующая описанный выше метод секущих. Если $F(x)$ – это полином, то вычислить все его корни можно также с помощью функции polyroots.

Встроенная функция root в зависимости от типа задачи может иметь либо два аргумента: root (f(x), x), либо четыре аргумента: root (f(x), x, a, b). Здесь $f(x)$ – скалярная функция, определяющая уравнение (1.1); x – скалярная переменная, относительно которой решается уравнение; a, b – границы интервала, внутри которого происходит поиск корня.

Первый тип функции root требует дополнительного задания *начального значения* переменной x. Для этого нужно просто предварительно присвоить x некоторое число. Поиск корня будет производиться вблизи этого числа. Таким образом, присвоение начального значения требует априорной информации о примерной локализации корня.

Рассмотрим решение уравнения $sin(x) = 0$, которое имеет бесконечное количество корней $x_N = N\,\pi$ ($N = 0, \pm1, \pm2, ...$). Для поиска корня средствами MathCAD требуется его предварительная локализация путем задания начального приближения, например, $x = 0.5$. MathCAD находит с заданной точностью только один корень $x_0 = 0$, лежащий наиболее близко к заданному начальному приближению. Если задать другое начальное значение, например, $x = 3$, то решением будет другой корень уравнения $x_1 = \pi$ и т.д.

На рис. 1.9 приведен пример вызова стандартной функции root с двумя аргументами для нахождения корней уравнения $sin(x) = 0$, график функции $f(x) = sin(x)$ и положение найденного корня.

$x := 0.5$

$f(x) := \sin(x)$

$s := \mathrm{root}(f(x), x)$

$s = -6.2 \times 10^{-7}$

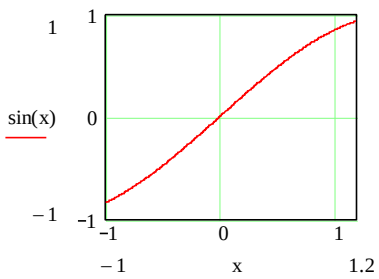


Рис. 1.9. Использование стандартной функции root для решения нелинейного уравнения $sin(x) = 0$

Если уравнение неразрешимо, то при попытке найти его корень будет выдано сообщение об ошибке. Кроме того, к ошибке или выдаче неправильного корня может привести и попытка применить метод секущих в области локального минимума или максимума $f(x)$. В этом случае секущая может иметь направление близкое к горизонтальному, и выводить точку следующего приближения далеко от предполагаемого положения корня. Для решения таких уравнений лучше применять встроенную функцию Minerr. Аналогичные проблемы могут возникнуть, если начальное приближение выбрано слишком далеко от настоящего решения или $f(x)$ имеет особенности типа бесконечности.

Иногда удобнее задавать не начальное приближение к корню, а интервал $[a, b]$, внутри которого корень заведомо находится. В этом случае следует использовать функцию root с четырьмя аргументами, а начальное значение x присваивать не нужно. Поиск корня осуществляется в промежутке между a и b.

При этом явный вид функции $f(x)$ может быть определен непосредственно в теле функции root. На рис. 1.10 приведен листинг программы с использованием этого варианта функции root.

$$x := \mathrm{root}(\sin(x), x, -1, 1)$$

$$x = 0 \quad \sin(x) = 0$$

Рис. 1.10. Поиск корня алгебраического уравнения в заданном интервале

Когда функция root имеет четыре аргумента, следует помнить о двух ее особенностях:

– внутри интервала $[a, b]$ не должно находиться более одного корня, иначе будет найден один из них, заранее неизвестно какой именно;

– значения $f(a)$ и $f(b)$ должны иметь разный знак, иначе будет выдано сообщение об ошибке.

Если уравнение не имеет действительных корней, но имеет мнимые, то их также можно найти. На рис. 1.11 приведен пример, в котором уравнение $x^2 + 1 = 0$, имеющее два чисто мнимых корня, решается два раза с разными начальными значениями.

Для решения этого уравнения второй вид функции root (с четырьмя аргументами) неприменим, поскольку f(x) является положительно определенной и указать интервал, на границах которого она имела бы разный знак, невозможно.

$$x := 0.5$$

$$\mathrm{root}\left(x^2 + 1, x\right) = -i$$

$$x := -0.5$$

$$\mathrm{root}\left(x^2 + 1, x\right) = i$$

Рис. 1.11. Поиск мнимого корня

Отметим, что $f(x)$ может быть функцией не одного, а любого количества аргументов. Эта возможность проиллюстрирована на рис. 1.12 на примере функции двух переменных $f(x, y) = = x^2 – y^2 + 3$. В самой функции root необходимо определить, относительно какого из аргументов следует решить уравнение. Затем уравнение $f(x, 0) = 0$ решается относительно переменной x, а потом другое уравнение – $f(1, y) = 0$ относительно переменной y.

$$f(x, y) := x^2 - y^2 + 3$$

$$x := 1$$

$$y := 0$$

$$\mathrm{root}(f(x, y), x) = -1.732i$$

$$\mathrm{root}(f(x, y), y) = 2$$

Рис. 1.12. Поиск корня уравнения, заданного функцией двух переменных

При численном решении уравнений относительно одной из переменных необходимо предварительно определить значения остальных переменных. Иначе попытка вычисления уравнения приведет к появлению ошибки «**This variable or function is not defined above**», в данном

случае говорящей о том, что другая переменная ранее не определена. Конечно, можно указать значения других переменных непосредственно внутри функции root.

Если функция $f(x)$ – полином, то все его корни можно определить, используя встроенную функцию polyroots(v), где v – вектор, составленный из коэффициентов полинома. Поскольку полином N-й степени имеет ровно N корней (некоторые из них могут быть кратными), вектор v должен состоять из N+1 элемента. Результатом действия функции polyroots является вектор, составленный из N корней рассматриваемого полинома. На рис. 1.13 приведен пример решения уравнения $f(x) = (x – 13)\,(x – 1)^3 = x^4 – 6x^3+12x^2 – 10x+3 = 0$.

Коэффициенты полинома записаны в виде вектора в первой строке примера. Первым в векторе должен идти свободный член полинома, вторым – коэффициент при x^1 и т.д. Последним, $N + 1$, элементом вектора

$$v := (3 \quad -10 \quad 12 \quad -6 \quad 1)^T$$

$$\text{polyroots}(v) = \begin{pmatrix} 0.992 \\ 1.004 + 7.177i \times 10^{-3} \\ 1.004 - 7.177i \times 10^{-3} \\ 3 \end{pmatrix}$$

Рис. 1.13. Поиск корня полинома

должен быть коэффициент при старшей степени x^N. Во второй строке показано действие функции polyroots. При этом численный метод вместо двух действительных единичных корней выдает одинаковые мнимые числа. Однако малая мнимая часть этих корней находится в пределах погрешности, определяемой константой TOL, и не должна вводить пользователей в заблуждение. Необходимо помнить, что корни полинома могут быть комплексными и ошибка вычислений может сказываться как на действительной, так и на комплексной части искомого корня.

В следующем примере, представленном на рис. 1.14, показано вычисление трех действительных корней полинома $f(x) = 6 – 7x + x^3$ с понижением порядка полинома. Здесь используется вариант функции root с двумя аргументами. Приведенный на рисунке график функции $f(x)$ показывает, что уравнение имеет три действительных корня. Задавая начальное приближение $z = –2$, находим один из корней полинома: $x_1 = –3$. Затем исходный полином делится на $(z – x_1)$ и отыскивается второй корень $x_2 = 1$. Далее функция root еще раз вызывается для нахождения корня полинома первого порядка, получаемого делением исходного полинома на $(z – x_1)$ и $(z – x_2)$. Для каждого из найденных корней производится проверка – вычисляется невязка уравнения.

$$a_1 := 6 \quad a_2 := -7 \quad a_3 := 0 \quad a_4 := 1$$

$$f(y) := a_1 + a_2 \cdot y + a_3 \cdot y^2 + a_4 \cdot y^3$$

$$z := -2 \quad x_1 := \text{root}(f(z), z) \quad x_1 = -3 \quad \left|f(x_1)\right| = 2.463 \times 10^{-5}$$

$$z := 0.5 \quad x_2 := \text{root}\left(\frac{f(z)}{z - x_1}, z\right) \quad x_2 = 1 \quad \left|f(x_2)\right| = 6.078 \times 10^{-4}$$

$$z := 3 \quad x_3 := \text{root}\left[\frac{f(z)}{(z - x_1)\cdot(z - x_2)}, z\right] \quad x_3 = 2 \quad \left|f(x_3)\right| = 1.904 \times 10^{-4}$$

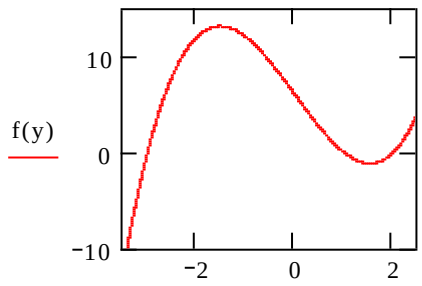


Рис. 1.14. Поиск корня полинома с понижением порядка

Практическая работа №1

Аппроксимация эмпирических данных методом наименьших квадратов

Необходимость решения несовместных систем достаточно часто возникает в практических расчетах, например, при анализе эмпирических данных.

В экономике рассматриваются связи между стоимостью продукции, объемом производства, ценой и прибылью. Несмотря на сложность этих связей, в определенных моделях они могут быть линейными. Пусть, например, выпуск t_1 экземпляров изделия обходится в сумму y_1 , выпуск t_1 экземпляров - в сумму; и т.д. Тогда производитель может оценить сумму издержек у на следующей неделе, предположив, что она линейно зависит от объема выпуска t: y = ct = d, и оценив значения коэффициентов c, d по уже имеющимся данным. Коэффициент c называется предельной стоимостью производства, а коэффициент d определяет накладные расходы. Предположив, что эмпирические данные (t1, y1), (t2, y2), …, (tn yn,,)подчиняются зависимости y = ct + d, получим относительно неизвестных c, d систему линейных алгебраических уравнений.

$$\begin{cases} ct_1 + d = y_1, \\ ct_2 + d = y_2, \\ \dots\dots \\ ct_n + d = y_n, \end{cases} \quad Ax = b, \quad A = \begin{pmatrix} t_1 & 1 \\ t_2 & 1 \\ \dots & \\ t_n & 1 \end{pmatrix}, \quad b = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \quad x = \begin{pmatrix} c \\ d \end{pmatrix},$$

в которой при n > 2 уравнений больше, чем неизвестных. Линейная система, число уравнений в которой больше числа неизвестных, называется нераспределенной.

Если линейное соотношение действительно справедливо и эмпирические данные (h, уд), (tz, yz), (tn, y") измерены точно, то полученная система совместна, ранг матрицы системы равен двум (число неизвестных) и значения коэффициентов линейной зависимости можно найти из первых двух уравнений системы. На практике такая ситуация невозможна - эмпирические данные по своей природе всегда содержат ошибку, а линейная модель лишь приближенно описывает реальные связи величин. Следовательно, система несовместна и ее нормальное обобщенное решение позволяет найти наилучшие приближенные знания коэффициентов линейной функции, поскольку в этом случае невязка минимальна Построенному таким Образом решению можно дать геометрическую интерпретацию. Поскольку,

$$\min_{x \in \mathbb{R}^n} |Ax - b| = = \min_{x \in \mathbb{R}^n} \sqrt{\sum_{i=1}^{n} (ct_i + d - y_i)^2},$$

то линейная зависимость y=ct+d-Это прямая на плоскости переменных y,t сумма квадратов расстояний до которых заданных эмпирических точек.(t_1,y_1)(t_2 y_2)…($t_n y_n$) минимальная. Нормальное обобщенное решение в этом случае является решением нормальной системы метода наименьших квадратов:

$$(A^T A)\begin{pmatrix} c \\ d \end{pmatrix} = (A^T b),$$

$$A^{<1>} = \begin{pmatrix} t_1 \\ \dots \\ t_n \end{pmatrix}, \quad A^{<2>} = \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}, \quad b = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}.$$

ЗАДАНИЕ

Найдите методом наименьших квадратов значения коэффициентов линейной зависимости y = ax + Ь по заданным эмпирическим данным. Используя найденную линейную зависимость, вычислите значение y в точке N + 0,55, где N - номер варианта.

Порядок выполнения задания

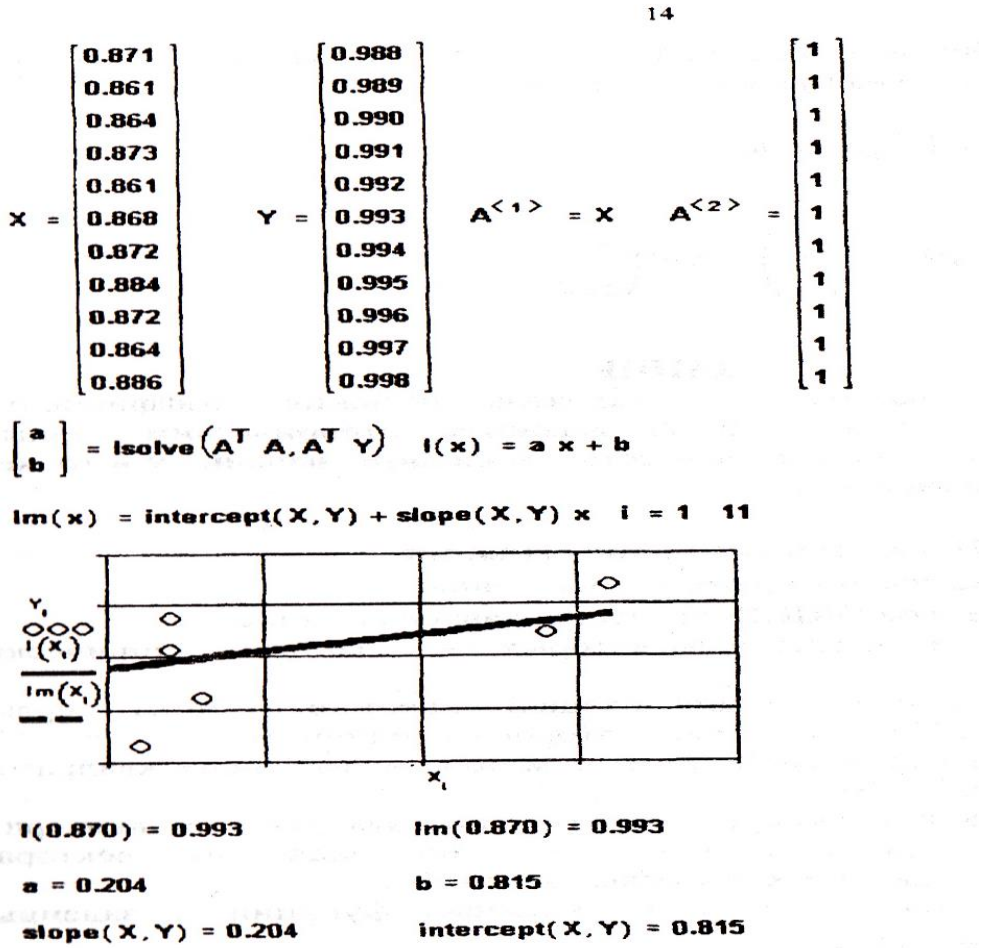
1. Установите автоматический режим вычислений.
2. Присвойте переменной ORIGIN значение, равное единице.
3. Введите векторы x, y, элементы которых - заданные эмпирические данные.
4. Определите матрицу A соответствующей линейной системы, первый столбец которой A“) = x, а элементы второго - единицы.
5. Найдите решение нормальной системы метода наименьших квадратов,используя функцию lsolve.
6. Вычислите аппроксимируюшую прямую, используя функции intersept(x,y) и 510pe(x,y), которые вычисляют по заданным векторам экспериментальных данных x, y коэффициенты Ь,
7. Изобразите графики полученных линейных функций и заданные экспериментальные точки.
8. Найдите значение y = ax + b B указанной точке x

Для изучения зависимости отнового числа бензина от чистоты катализатора (%) провели 11 измерений, приведенных ниже.

Октановое число	988	989	99 0	991	992	993	994	99 5	99 6	997	998
Чистота	87 1	86 1	86 4	87 3	86 1	868	87 2	884	87 2	86 4	886

Найдите коэффициент 3. b линейной зависимости y = ax + Ь октанового числа от чистоты катализатора. Вычислите значение октанового числа для чистоты

катализатора 87%. Фрагмент рабочего документ Mathcad, содержащий соответствующие вычисления и графики, приведен ниже



Указание. Функция lsolve(A, b) возвращает вектор x решения системы Ax b. найденного методом Гаусса с оценкой числа обусловленности. Здесь не используется явная формула решения нормальной обобщенной системы x $(A^TA)^{-1}A^Tb$, поскольку часто в задачах об аппроксимации эмпирических данных матрица получается плохо обусловленной* и при вычислении обратной к ней матрицы возникают большие погрешности округления. Функции intersept(x, y) и slope(x, y) возвращают значения коэффициентов Ь и а линейной. функции y = x + b, аппроксимируюшей экспериментальные данные, сохраненные в векторах x и y. Для того, чтобы построить графики, щелкните по свободному месту в рабочем документе и по кнопке декартова графика в панели графиков Введите в качестве аргумента имя x_i, а В качестве функции, через запятую, имена Yi, I(Xi), Im(Xi) Здесь Yi – экспериментальные точки, $I(X_I)$- линейная функция, вычисленная с помощью isolve, Im(Xi)линейная функция, вычисленная с помошью intersept и slope. На рис.3 представлен вид окон настройки графиков, изображенных в документе. Значения октанового числа, вычисленные для частоты катализатора 0,870 по обеим формулам, совпадают. Совпадают и приведенные в последних строках рабочего документа коэффициенты линейной функции, вычисленные обоими способами. Матрица **A** плохо обусловлена, если малые изменения ее элементов (например, округление) приводят к существенным

изменениям элементов матрицы A". Число“ обусловленности матрицы CandA) -› мера зависимости погрешностей вычисления A“l от погрешности элементов A. Например, можно определить число обусловленности как модуль отношения наибольшего собственного значения матрицы к ее наименьшему собственному .значению.“

ВАРИАНТЫ ЗАДАНИЙ

1.

x	1	1.1	1 2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
y	0.686	0.742	0.767	0.646	0.807	0.774	0.97	0.932	0.936	0.978	1.048

2.

x	2	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3
y	2.312	2.251	2.418	2.752	2.459	2.7	3.022	3.079	2.42	2.669	3.241

3.

x	3	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4
y	4.615	4.591	5.13	5.481	5.492	5.553	5.471	5.727	5.798	6.11	6.605

4.

x	4	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5
y	8.472	8.805	9.096	8.993	9.312	9.465	9.771	9.61	9.722	11.419	10.285

5.

x	5	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6
y	12.36	13.63	13.304	13.148	13.482	14.24	14.516	14.882	15.246	15.369	15.158

6.

x	6	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7
y	17.631	19.747	19.783	18.806	19.886	21.118	20.208	19.481	20.153	20.505	21.29

7.

x	7	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8
y	25.243	25.133	25.669	26.627	26.753	27.234	26.491	26.876	27.228	28.065	27.781

8.

x	8	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9
y	30.528	34.221	34.233	34.114	33.595	34.058	34.498	35.822	35.678	37.442	35.698

9.

x	9	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10
y	41.742	42.244	43.884	42.167	43.696	45.042	42.461	45.727	44.056	45.863	44.953

10.

x	10	10.1	10.2	10.3	10.4	10.5	10.6	10.7	10.8	10.9	11
y	49.758	51.924	50.083	52.376	53.413	54.966	52.771	54.115	55.476	55.686	56.196

11.

x	11	11.1	11.2	11.3	11.4	11.5	11.6	11.7	11.8	11.9	12
y	62.173	63.055	63 725	64.237	64.086	63.587	65.412	65.284	65.05	68.876	65.74

12.

x	12	12.1	12.2	12.3	12.4	12.5	12.6	12.7	12.8	12.9	13
y	71.167	74.264	72.658	74.507	76.649	75.517	75.708	76.359	79.316	77.373	77.698

13.

x	13	13.1	13.2	13.3	13.4	13.5	13.6	13.7	13.8	13.9	14
y	86.612	85.491	87.803	88.613	89.075	89.24	89.633	90.761	91.323	91.428	91.712

14.

x	14	14.1	14.2	14.3	14.4	14.5	14.6	14.7	14.8	14.9
y	99.811	100.31	99.492	102.61	103.20	104.36	104.73	105.16	104.65	105.58

Практическая работа № 2
Основные задачи статистики. Выборки. Гистограммы. Полигоны частот

Математическая статистика в основном занимается изучением случайных величин и случайных событий по результатам наблюдений. Ее главная задача извлечь максимум информации из эмпирических данных.

Важнейшими понятиями математической статистики являются генеральная совокупность и выборка.

Генеральная совокупность - это вероятностное пространство с определенной на нем случайной величиной § Функцию распределения этой случайной величины F§(x) часто называют теоретической функцией распределения, хотя более правильным представляется другой термин истинная функция распределения, в отличие от эмпирической (экспериментальной, приближенной) функции распределения, которая будет определена ниже. В результате проведения n экспериментов со случайной величиной получаем n выборочных значений x_i, i = 1, 2,…, n. Вся совокупность этих значений называется выборкой. **Выборка** - это, вообще говоря, случайный вектор: если в одной серии из n испытаний получена выборка x1, x2, ..., xn, то в другой серии будет получена, скорее всего, другая выборка x'1, x'2, …, x'n)

Mathcad обладает богатой библиотекой встроенных функций, предназначенных для решения основных задач статистики. Они собраны в разделе Statistics библиотеки встроенных функций Mathcad.

Эмпирические распределения и числовые характеристики

Выборка из генеральной совокупности является основным источником информации о случайной величине. По выборке оценивается класс распределений, к которому принадлежит распределение исследуемой случайной величины, устанавливаются интервалы, в которых лежат истинные значения параметров распределения, проверяются гипотезы об этой случайной величине и формулируются выводы о других ее свойствах.

Чтобы использовать аппарат математической статистики, нужно, прежде всего уметь находить некоторые числовые характеристики выборок и строить эмпирические распределения, с помощью которых в дальнейшем можно делать соответствующие выводы.

Рассмотрим некоторые правила предварительной обработки выборочных данных. Представленная ниже таблица выборки объема n=250 будет использоваться далее во всех вычислениях, а так же станет источника построения выборок для индивидуальных вариантов заданий.

145.61	143.206	145.267	140.485	133.143	150.435	148.794	155.564	171.918
158.087	159.851	158.622	159.156	156.73	139.557	150.691	142.444	156.967
148.181	143.556	142.769	144.834	155.58	147.552	150.895	162.618	142.945
150.019	161.076	158.926	120.991	128.429	152.06	143.842	138.023	150.99
157.708	153.059	150.113	142.355	145.909	143.262	148.678	160.181	151.805
155.133	157.398	149.837	152.788	151.622	154.285	145.248	143.045	180.482
147.135	137.201	157.594	146.073	137.964	139.631	149.807	150.32	152.649
154.915	152.383	143.155	133.852	164.113	159.715	138.44	151.437	166.972
146.797	129.688	135.888	136.747	144.829	150.621	144.042	146.693	155.391
152.186	154.05	138.441	138.949	138.966	145.927	136.867	121.596	162.762
157.911	151.429	139.937	140.73	141.22	152.777	145.978	163.02	136.219
153.803	154.377	167.603	143.527	155.51	165.465	131.784	163.079	139.511
154.591	139.478	137.579	154.241	130.834	148.761	154.132	164.656	137.711
146.154	154.763	151.862	151.96	155.206	158.229	159.314	158.972	152.601
143.066	154.656	148.493	141.368	171.144	137.64	133.062	153.865	135.711
145.891	158.742	144.311	140.903	141.323	160.971	139.771	137.484	156.247
142.623	155.409	156.641	155.196	151.459	149.488	153.16	152.488	148.294
145.475	152.937	151.507	140.659	157.925	157.163	160.438	158.11	156.17
147.549	149.142	156.848	157.911	153.578	147.887	148.445	151.36	158,639
169.584	150.688	155.646	155.572	168.911	164.788	127.059	156.623	145.593
145.263	150.889	143.012	153.472	141.25	169.001	122.741	158.702	171.791
160.849	161.757	140.286	134.241	154.64	164.744	161.654	142.365	155.094
154.96	141.977	143.729	144.466	146.54	145.355	152.509	146.266	147.269
162.895	151.941	170.865	134.377	150.79	154.205	166.274	156.198	132.828
136.274	173.96	157.332	149.975	141.54	139.826	133.692	139.462	161.159
159.455	157.597	139.385	145.867	166.069	150.237	146.685	145.436	153.969
154.961	149.211	150.83	154.224	142.28	148.655	135.371	152.018	166.807
140.923	157.864	148.745	138.823	157.239	152.912	141.182		

Первичная обработка данных состоит обычно в отыскании максимального X_{max} и минимального X_{min} значений выборки (в Mathcad они вычисляются соответственно функциями max(§) и min(§), а также размаха варьирования R = X_{max} - X_{min}. Для приведенной выше выборки эти величины равны: X_{max} = 180.482, x… = 120.991, R = 59.49.

Следующий этап первичной обработки - группировка и ее графическое представление. Группировка выборки объема n состоит в следующем. Промежуток [X_{min}, X_{max}] разбивают на m интервалов группировки (чаще всего одинаковой длины) и подсчитывают число n_1, выборочных значений, которые попали в J-й `интервал. Обычно выбирают m = 7 - 20. Теперь каждый интервал группировки Δj = (aj, bj) представлен своими левой a,- и правой Ьі границами и числом “3 элементов выборки, принадлежащих ему. Каждый интервал удобно представлять не двумя границами, а одним числом - срединным значением.

Наиболее наглядная форма графического представления группировкигистограмма.

Если $\delta_1, \delta_2, \ldots, \delta_n$ - длины интервалов группировки, а x], x2, хп – их середины и 11}- = {\,/11 - относительные частоты попадания наблюдений в 1-й интервал группировки, то можно построить график ступенчатой функции: $f(x) = h_j/\delta_j,\ x \in \Delta_j,\ j = 1, 2, \ldots, m.$

Этот график называется гистограммой. В Mathcad для построения гистограмм предназначена функция hist(Δ, §)

Очевидно, что величина интервала группировки существенно влияет на вид гистограммы. При малой их ширине в каждый интервал (попадает Ё, незначительное число наблюдений или даже не попадает ни одного, в результате гистограмма становится сильно "изрезанной" и плохо передает основные особенности изучаемого распределения. Другая крайность большие интервалы группировки; в этом случае скрадываются характерные J черты распределения.

Иная форма графического представления группированных данных полигон частот. Полигон частот - это ломаная линия, соединяющая точки с координатами (xi, hi), т.е. с абсциссами, равными серединам интервалов группировки, и ординатами,

равными соответствующим частотам. Можно также построить полигон накопленных частот – график.

Можно так же **построить полигон накопленных частот** график ломаной ,соединяющий точки с координатами $\left(b_j, \sum_{k=1}^{j} \frac{n_k}{n}\right)$ или $\left(b_j, \sum_{k=1}^{j} n_k\right)$ т.е. с абсциссами, равными правым границам интервалов группировки, и ординатами, равными соответствующим накопленным частотам или относительным накопленным частотам. Ниже приведен фрагмент рабочего документа Mathcad с вычислением X_{max}], x… и R = X_{max}] - x… для исследуемой выборки, а также с гистограммами и полигонами частот для различных интервалов группировки.

ORIGIN := 1 ξ := rnorm(250, 150, 10)

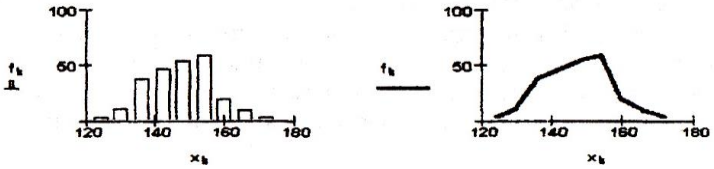
xmax := max(ξ) xmin := min(ξ) R := xmax – xmin

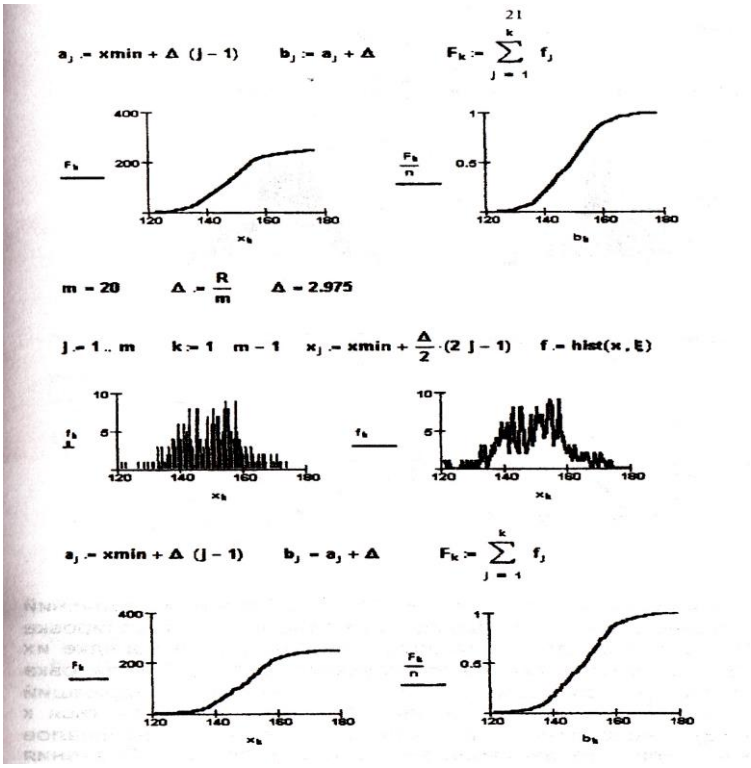
xmax = 180.482 xmin = 120.991 R = 59.49

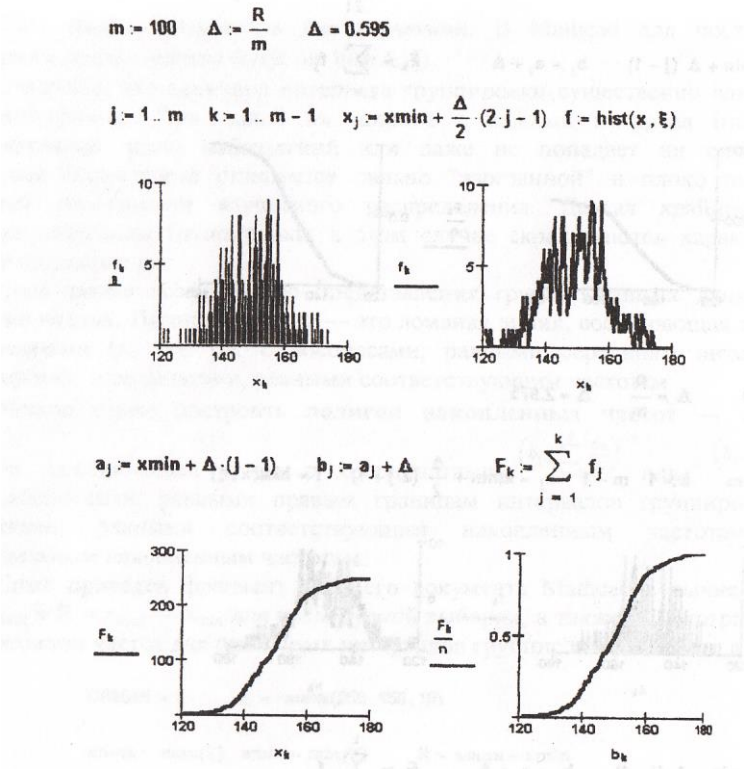
ξ := sort(ξ) n := 250

m := 10 $\Delta := \frac{R}{m}$ Δ = 5.949

$j := 1..m$ $k := 1..m-1$ $x_j := xmin + \frac{\Delta}{2}\cdot(2\cdot j - 1)$ f := hist(x, ξ)







Указание. В приведенном фрагменте 250 выборочных значении сохранены в массиве с именем &, Прежде чем приступать к группировке выборки, необходимо Упорядочить выборочные значения в порядке их возрастания. Эту операцию выполняет функция sort(§). Группировка производится с помощью функции hist(x,§), где x - массив, содержащий значения середин интервалов группировки. Прежде чем обратиться к функции hist(x,§), необходимо вычислить середины интервалов группировки и присвоить их значения элементам массива x. Значения функции hist(x, §,) - вектор, компоненты которого равны количеству элементов массива § которые попадают в интервал группировки, середина которого равна соответствующей компоненте массива x. На рис. 4 приведены окна настройки параметров изображения гистограмм.

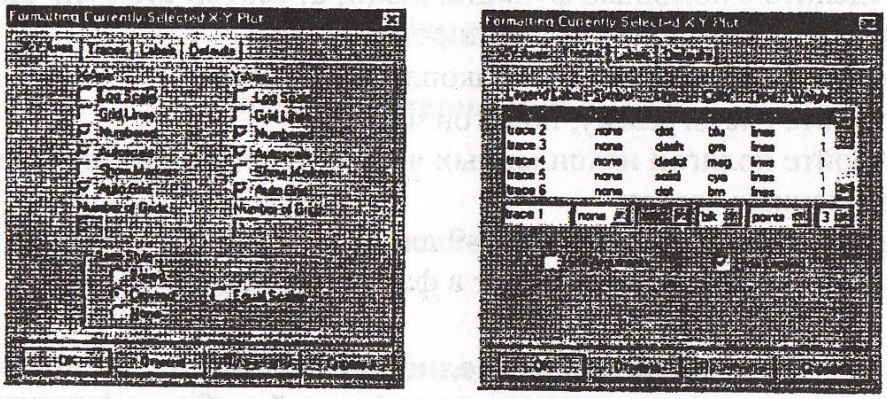


Рис. 4. Окна настройки графиков

При первичной обработке выборочных данных можно рекомендовать несколько общих правил:

1. Перед началом группировки следует упорядочить выборочные значения в порядке возрастания. Такая упорядоченная в порядке возрастания выборка называется вариационным рядом.
2. При выборе числа интервалов группировки следует ориентироваться на 10-20 интервалов.
3. Предпочтительнее использовать интервалы одинаковой длины.
4. При анализе охватывайте всю область данных.
5. Избегайте полуоткрытых промежутков.
6, Интервалы группировки не должны перекрываться.

ЗАДАНИЕ

Вычислите максимальное, минимальное значения и размах для заданной выборки. Выполните группировку для заданных значений т=10, 20, постройте соответствующие гистограммы, полигоны частот и полигоны накопленных частот. Выполните вычисления для 100 чисел из приведенной выборки, начиная с числа и, номер которого указан в таблице.

Порядок выполнения задания

1. Определите и введите вектор-столбец выборочных значений.
2. Упорядочите выборку в порядке возрастания выборочных значений.
3. Вычислите минимальное значение и размах для полученной выборки.
4. Определите число интервалов группировки и их длину.
5. Определите вектор-столбец, содержащий середины интервалов группировки.

Пример выполнения задания

Примерный вариант выполнения задания для всей выборки для m = 10, 20, 100 приведен выше.

ВАРИАНТЫ ЗАДАНИЙ

N	n	N	n	N	n	N	n	N	n
1	10	5	50	9	90	13	95	17	135
2	20	6	60	10	270	14	105	18	145
3	30	7	70	11	75	15	115	19	155
4	40	8	80	12	85	16	125	20	165

Практическая работа № 3

Числовые характеристики выборки.

Показатели положения. Среднее значение выборки вычисляется по формуле $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

В Mathcad для вычисления выборочного среднего значения выборки, сохраненной в матрице A, предназначена функция mean(A). Выборочной **кватилью** уровня p называется решение уравнения

$$F_n(x) = p,$$

f‘ где Fn(x)-выборочная функция распределения.

В частности, выборочная медиана есть решение уравнения Fn(x)=0.5, т.е. **выборочная медиана**-Это выборочная квантиль уровня 0.5. Выборочная медиана разбивает выборку пополам: слева и справа от нее оказывается Одинаковое число элементов выборки. Если число элементов выборки четно n = 2k, то выборочную медиану определяют по формуле $(x_k + x_{k+1})/2$, где X_k и x_{k+1}—k-е и (k+1)-е выборочные значения из вариационного ряда. При нечетном объеме выборки (n=2k + 1) в качестве значения медианы принимают величину x_{k+1} и (k+1)-е выборочные значения из вариационного ряда. При нечетном объеме выборки (n=2k+1) в качестве значения медианы принимают величину x_{k+1}

В Mathcad для вычисления выборочной медианы выборки, сохраненной в матрице A, предназначена функция median(A).

К показателям положения относятся минимальный и максимальный элементы выборки, а также верхняя и нижняя квантили (они ограничивают зону, в которой сосредоточены 50% элементов выборки).

Для вычисления минимального и максимального элементов выборки, размещенной в матрице A, в Mathcad предназначены соответственно функции min(A) и max(A).

Показатели разброса. К показателям разброса относятся дисперсия выборки (выборочная дисперсия), стандартное отклонение, размах выборки, межквантильный размах, коэффициент эксцесса (выборочный эксцесс). Выборочной дисперсией называется величина

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

Однако в статистике чаще в качестве выборочной дисперсии используется величина

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

Причина такого, на первый взгляд, неожиданного, способа вычисления дисперсии будет объяснена далее. В Mathcad для определения дисперсии выборки,

сохраненной в матрице A предназначена функция var(A), а величину S^2 можно вычислить по формуле $S^2 = \frac{n}{n-1}$ var(A).

Стандартное отклонение рассчитывается по формуле $\hat{\sigma} = \sqrt{s^2}$.

Размах выборки вычисляется по формуле $R = X_{max} - X_{min}$.

Межквантильный размах равен $X_{0.75} - X_{0.25}$, где $X_{0.75}$—75%-ная квантиль, решения уравнения $Fn(X_{0.75}) = 0.75$, $X_{0.25}$—25%-ная квантиль, решение уравнения $Fn(X_{0.25}) = 0.25$.

Выборочный эксцесс определяется следующим образом. Сначала отыскиваются величина **выборочного центрального момента 4-го порядка**

$$\hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4.$$

А затем по формуле $E = \hat{E} = \hat{\mu}_4 (s^2)^{-2} - 3$ вычисляется выборочный эксцесс.

Показатели асимметрии. На основании этих показателей изучают информацию о симметрии распределения выборочных данных около центра выборки. Сюда в первую очередь относится коэффициент асимметрии, которой вычисляется по формуле

$$\hat{a} = \frac{\hat{\mu}_3}{\hat{\sigma}^3},$$

Где $\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$-выборочный центральный момент 3-го порядка а $\hat{\sigma}$ стандартное отключение, формула для вычисления которого приведена выше.

ЗАДАНИЕ

Для выборки, сформированной в предыдущем задании, вычислите все описанные выше выборочные характеристики.

Порядок выполнения задания

1. Прочтите сохраненный ранее файл, содержащий выборку.
2. Вычислите максимальный и минимальный элементы и размах выборки.
3. Рассчитайте выборочное среднее.
4. Найдите медиану.
5. Вычислите выборочную дисперсию и стандартное отклонение.
6. Найдите выборочные моменты 3-го и 4-го порядков.
7. Вычислите выборочный эксцесс.
8. Определите коэффициент асимметрии

Пример выполнения задания

Ниже представлен фрагмент рабочего Документа Mathcad, содержащий вычисление характеристик выборочных Данных, приведенных в начале раздела.

$n := 250$

$xmax := max(\xi)$ $xmin := min(\xi)$ $R := xmax - xmin$

$xmax = 180.482$ $xmin = 120.991$ $R = 59.49$

$mean := mean(\xi)$ $s2 := \frac{n}{n-1} \cdot var(\xi)$ $\sigma := \sqrt{s2}$

$mean = 149.849$ $s2 = 98.174$ $\sigma = 9.908$

$$\mu 3 := \frac{1}{n} \cdot \sum_{i=1}^{n} (\xi_i - mean)^3 \qquad \mu 4 := \frac{1}{n} \cdot \sum_{i=1}^{n} (\xi_i - mean)^4$$

$median = median(\xi)$ $E = \frac{\mu 4}{s2^2} - 3$ $\alpha := \frac{\mu 3}{\sigma^3}$

$median = 150.69$ $E = 0.136$ $\alpha = -0.055$

Указание. В Mathcad нет встроенных функций для вычисления выборочных моментов. Для определения среднеквадратичного отклонения в Mathcad предназначена функция stdev(A) == $\sqrt{var(A)}$. Рассчитываемое с ее помощью значение среднеквадратичного отклонения отлично от определенного выше, поэтому среднеквадратичное отклонение следует вычислять как $\sqrt{s^2}$.

Практическая работа №4
Оценка функции распределения

Как уже упоминалось ранее, распределение случайной величины является ее паспортом", содержащим всю информацию о случайной величине.

Рассмотрим методы оценивания функции распределения F§(x) случайной величины, о которой известно, что она непрерывна.

Пусть x== { x., x2, ...,x_n } - совокупность выборочных значений случайной величины &, т.е. выборка из случайной величины &. Расположим наблюдения x,, x2, x_n

. . в порядке их возрастания. Обозначим новую упорядоченную последовательность - **вариационный ряд** – x1 ; x 2, x_n. , x1<x 2< x_n .По этому вариационному ряду построим следующую неубывающую ступенчатую функцию:

$$\widehat{F}_n(x) = \begin{cases} 0, & x \le x'_1, \\ \frac{k-1}{n}, & x'_{k-1} < x \le x'_k, \quad k = 1, 2, \ldots, n, \\ 1, & x > x'_n. \end{cases}$$

Из приведенной выше формулы видно, что функция F ’n(x) претерпевает и каждой точке вариационного ряда скачок, равный по величине 1/п. Если какая-нибудь точка вариационного ряда повторяется m раз (m точек вариационного ряда совпадают), то скачок функции в этой точке равен m/n.

Функция $\hat{F}'_n(x)$ называется эмпирической функцией распределения.

Замечание. Эмпирическая функция распределения $\hat{F}_n(x)$ зависит не только от x, но и от всей выборки $\hat{x}$. Чтобы обратить внимание на этот факт, будем обозначать эмпирическую функцию распределения через $\hat{F}'_n(x)$ Именно принимают за оценку теоретической функции распределения F(x). Остается выяснить, насколько хорошо эмпирическая функция распределения аппроксимирует теоретическую функцию распределения.

Если Fg(x) - теоретическая функция распределения, а $F_n(x)$ – эмпирическая функция распределения, построенная по заданной выборке значений случайной величины §, то в качестве меры расхождения теоретической и эмпирической функций распределения возьмем величину

$$D_n(\hat{x}) = \sup_x |F_n(x) - F_\xi(x)|.$$

Эта функция от выборочных значений x называется **статистикой.**

Колмогорова. Следует помнить, что $D_n(\hat{x})$ случайная величина и что ее распределение не зависит от неизвестной теоретической функции распределения $F_\xi(x)$, если она непрерывна. Более того, справедлива **теорема Колмогорова :**если функция распределения $F_\xi(x)$, случайные величины § непрерывна ,а $\hat{F}_n(x)$ -ее выборочная функция распределения , то при $n \to \infty$

$$P\left(\sup_x |\hat{F}_n(x) - F_\xi(x)| < \frac{z}{\sqrt{n}}\right) \to K(z) = \begin{cases} 0, & z \le 0, \\ \sum\limits_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}, & z > 0. \end{cases}$$

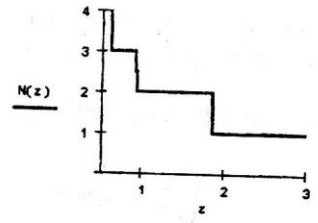
ФункцияK(z)представляет собой функциональный ряд, который следует протабулировать. Сразу обратим внимание на то, что этот ряд сходится абсолютно для всех Z>0 , но неравномерно на промежутке $[0, +\infty]$. Это означает, что для достижения заданной точности при вычисление K(z) число N членов в соответствующей частичной сумме зависит от z. Если e-требуемая точность вычисления К(г), то число N вычисляется по формуле:

$$N = \left[\frac{1}{z}\sqrt{\frac{1}{2}\ln\frac{1}{\varepsilon}}\right] + 1,$$

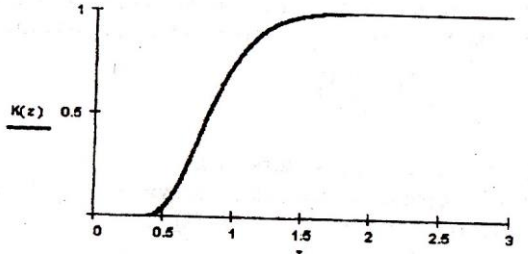
Где символом [] обозначена часть числа .

Ниже приведен фрагмент рабочего документа Mathcad, содержащий приближенное определение функции К(%)для e=0.001, N=3, и соответствующие графики.

$$\varepsilon := 0.001 \qquad N(z) := \mathrm{floor}\left(\frac{1}{z}\cdot\sqrt{\frac{1}{2}\cdot\ln\left(\frac{1}{\varepsilon}\right)}\right) + 1$$



$N := 3 \qquad K(z) := \begin{cases} 0 & \text{if } z \le 0 \\ \sum_{k=-N}^{N} (-1)^k \cdot \exp(-2 \cdot k^2 \cdot z^2) & \text{if } z > 0 \end{cases}$



Из приведенных в документе графиков видно, что для малых z величину К(z) можно положить равной нулю, а для z > 2 можно считать К(z) равной единице.

Зададимся вероятностью а такой, что событие, происходящее с вероятностью 1 - α представляется практически достоверным. Вычислим корень уравнения 1 - К(1)= а, тогда неравенство $\hat{F}_n(x) - \frac{z_\alpha}{\sqrt{n}} < F_\xi(x) < \hat{F}_n(x) + \frac{z_\alpha}{\sqrt{n}}$ выполняется для всех действительных х с вероятностью, близкой к l-α

Таким образом, в окрестности эмпирической функции распределения построен"коридор", в котором лежит истинная, теоретическая функция распределения $F_\xi(x)$ С ростом n "ширина" этого коридора стремится к нулю.

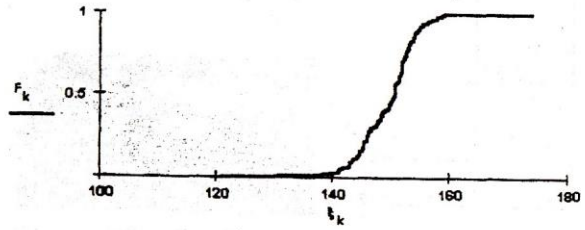
Вместо эмпирической функции распределения будем использовать функцию накопленных относительных частот, поскольку $\hat{F}_n(x) = F_k$ для $x \in (\xi_{k-1}, \xi_k]$ и значения функций совпадают вне промежутка $[x_{min}, x_{max}]$.

Ниже приведен фрагмент рабочего документа Mathcad с построением 95%-ного ”коридора" для функции распределения случайной величины по приведенной в лабораторной работе №2 выборке.

$m = 250 \qquad \Delta = \frac{R}{m} \qquad j = 1..m \qquad k = 1..m-1$

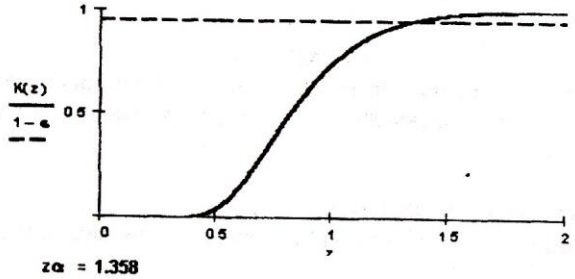
$x_j := xmin + \frac{\Delta}{2}\cdot(2\cdot j - 1) \qquad f = hist(x, \xi) \qquad \Delta = 0.238$

$a_j = xmin + \Delta\cdot(j-1) \qquad b_j = a_j + \Delta \qquad F_k = \sum_{j=1}^{k} \frac{f_j}{n}$

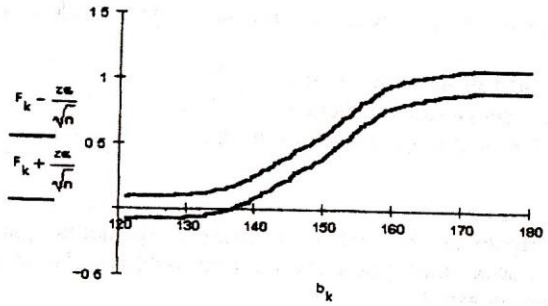


$\alpha := 0.05 \qquad K(z) := \begin{cases} 0 \text{ if } z \le 0 \\ \sum_{k=-3}^{3} (-1)^k \cdot \exp(-2\cdot k^2 \cdot z^2) \text{ if } z > 0 \end{cases}$

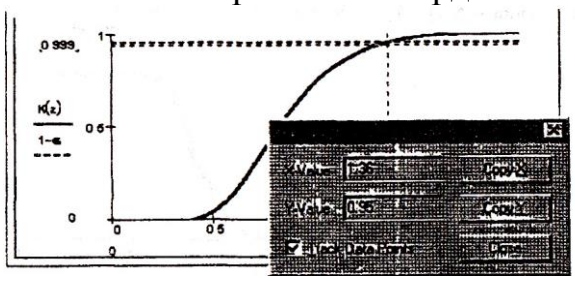
$p = 1 - \alpha$



$z\alpha = 1.358$



Указание. Как уже отмечалось выше, в качестве эмпирической функции распределения использована эмпирическая функция накопленных частот. Заметим, что Mathcad вместо графика ступенчатой функции строит ломаную линию, соединяя "ступеньки" вертикальными отрезками прямых. Корень уравнения 1 - K(z) = a проще всего найти графически, используя операцию Trace пункта Graph меню Format как точку пересечения графика K(z) и прямой y = 1 -.α Ниже приведен фрагмент окна Mathcad с окном отображения координат точки пересечения.



Для оценки плотности распределения случайной величины можно воспользоваться полигоном частот, который представлен выше. При не очень обременительных ограничениях доказано, что выборочная плотность вероятностей, т.е. полигон частот, с ростом объема выборки до бесконечности стремится к истинной, теоретической, плотности распределения исследуемой случайной величины.

ЗАДАНИЕ 1

Постройте для выборки, сформированной в лабораторной работе №2, 95%-ный ”коридор" для функции распределения исследуемой случайной величины.

Порядок выполнения задания

1. Прочитайте файл, сохраненный при выполнении лабораторной работы№2.
2. Определите статистику Колмогорова - функцию К(г) и постройте ее график.
3. Определите значение величины *a*.
4. Решите графически уравнение 1 - K(z) = α
5. Постройте "коридор" для теоретической функции распределения.

Пример выполнения задания

Пример построения 95%-ного "коридора" функции распределения для исследуемой во всех примерах этого раздела выборки 250 значении случайной величины приведен выше. При анализе статистических данных большую роль играет опыт и интуиция исследователя. В этой связи чрезвычайно полезными представляются следующие упражнения. Пользователь генерирует достаточно большую выборку значений случайной величины, имеющей известное непрерывное распределение ‹: известными параметрами. А затем произв0дит описанные выше вычисления, изменяя параметры задачи - объем выборки, количество интервалов группировки, доверительные вероятности и др., и сравнивает полученные оценки с известными теоретическими значениями. Здесь прежде всего полезно изучить равномерное и нормальное распределения. Приведенное ниже задание заключается в решении именно такой задачи – исследование выборки значений случайной величины с заданным распределением.

Напомним, что исследованная во всех примерах раздела выборка представляет собой сгенерированную функцией Mathcad топи выборку 250 значений случайной величины, имеющей нормальное распределение N(150, 10). Следовательно, внимательный читатель может не затрудняться ручным вводом выборки для индивидуального варианта задания, а просто аккуратно сгенерировать ее.

ЗАДАНИЕ 2

Сгенерируйте выборку объема п значений случайной величины с заданным непрерывным распределением и выполните полный предварительный ее анализ для числа интервалов группировки, равного целой части размаха и доверительной вероятности а. “Постройте графики плотности вероятностей и функции распределения и сравните их с полученными графиками соответствующих выборочных функций.

Порядок выполнения задания

1. Установите в меню Math режим Optimization.
2. Присвойте переменной *n* значение, равное 100.
3. Постройте для заданного распределения графики плотности вероятностей и функции распределения.
4. Найдите математическое ожидание, дисперсию, среднеквадратичное отклонение, медиану, моменты 3- и 4-го порядка, асимметрию и эксцесс заданного распределения.
5. Сгенерируйте выборку объема n значений случайной величины, имеющей заданное распределение.

6. Определите как функции переменной *n* и найдите выборочные значения среднего, среднеквадратичного отклонения, моментов 3- и 4-го порядка, асимметрии и эксцесса.

7. Постройте гистограмму, полигон частот, график накопленных относительных частот.

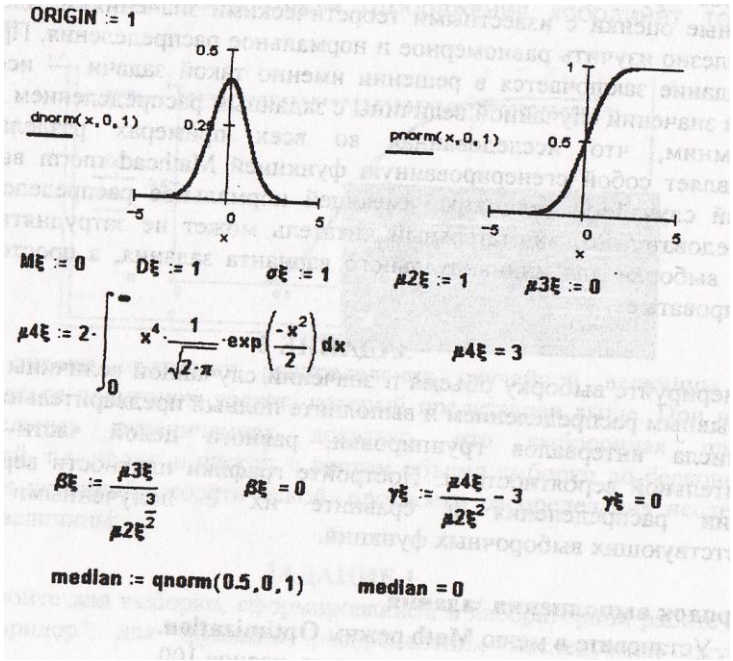
8. Постройте 95%-ный "коридор" для теоретической функции распределения и изобразите на этом же графике функцию заданного в условии распределения вероятностей.

9. Сравните вычисленные теоретические и выборочные значения параметров.

10. Выполните вычисления пп. 4-7 для n = 150, 200, 300, 500.

Пример выполнения задания

Ниже приведен пример выполнения задания для стандартного нормального распределения N=0,1



ТЕОРИЯ ПЛАНИРОВАНИЯ ЭКСПЕРИМЕНТА

Началом экспериментальных' исследований является сбор, изучение и анализ всех имеющихся данных об объекте. Априорная информация может быть скупой или обширной, но именно она является той базой, на которой строятся первые шаги исследования. Чем полнее знания об объекте, тем быстрее исследователь придет к окончательному решению поставленной задачи. В результате проведения предварительного, априорного этапа исследователь должен: составить полный список факторов, влияющих на изучаемое явление, исходя из того, что лучше, назвать несколько малозначащих факторов; задать ориентировочные пределы изменения факторов с учетом требований их совместимости; выбрать параметры оценки результатов экспериментов (критерии , оптимизации, функцию отклика и т.д.) в соответствии с поставленной задачей. Если список факторов окажется больше 5-7, необходимо выделить из них 3 – 5 наиболее значимых с помощью так называемых отсеивающих экспериментов. На этом завершается предварительный этап экспериментальных исследований.

В соответствии с идеей шагового поиска эксперимент проводится в несколько этапов. Число этапов и действия на каждом из них зависят от результатов предыдущего этапа и конечной цели исследования. Все многообразие конечных целей исследования можно обобщенно разделить на два типа: найти адекватное описание изучаемого явления ШШ найти значения факторов, при которых исследуемый процесс протекает наилучшим образом.

Планы экспериментов

Математический аппарат матирования экспериментов позволяет проводить активный эксперимент и получать только необходимую информацию отдельно о каждом факторе или сочетании факторов. В частности, это выражается В том, что коэффициенты регрессии, которые являются основными характеристиками каждого фактора, определяются независимо друг от друга. Управляемость процесса ,получения информации заключается в том, что в процессе исследований ставятся эксперименты не по всем возможным сочетаниям факторов, а только по сочетаниям (значениям факторов в каждом эксперименте),которые обеспечат получение нужной информации. Это, в первую очередь, резко сокращает количество опытов и облегчает обработку и анализ полученных результатов и, во-вторых, заставляет целенаправленно проводить исследования, четко обосновывая условия и количество экспериментов.

Если высказывается гипотеза о линейной зависимости исследуемого процесса, а прямую линию можно построить по двум точкам, то минимальное значение числа уровней факторов в экспериментах равно двум (-1 и +1). При количестве факторов равном и, количество экспериментов N будет равно 2”. План, построенный таким образом, получил название полного факторного эксперимента (ПФЭ). В ряде случаев бывает нужным определить не все коэффициенты в уравнении регрессии, а лишь часть из них. В этом случае ПФЭ даст избыточную информацию. В подобных ситуациях надо переходить к планам, представляющим собою части плана ПФЭ (половину, четверть и т.д.) и называемым дробными репликами ПФЭ или дробным факторным экспериментом (ДФЭ), и количество опытов N равно уже 2 в степени (n - k). Дробные реплики особенно удобны при большом числе факторов (n больше 5), так как там удается смешивать коэффициенты при факторах и, например, двойных

взаимодействиях с коэффициентами при тройных и более высоких взаимодействиях, которые обычно слабо влияют на процессы. В теории планирования экспериментов разработаны планы ДФЭ, учитывающие, какие коэффициенты в уравнении регрессии необходимо определить.

Планы ПФЭ и ДФЭ позволяют найти коэффициенты регрессии, если проксимируемая поверхность (исследуемый процесс) хорошо описывается полиномом без квадратичных членов. Если же исследуемый процесс носит нелинейный характер, то соответственно двух уровней значения факторов уже недостаточно. В этом случае используются так называемые ортогональные планы второго порядка.

В ортогональном плане второго порядка к ядру, представляющему собойплан ПФЭ, добавляются центральная точка (xi = 0, i = 1,2, .,n) и по две так называемые "звездные" точки для для каждого фактора (x; = ia). Величина а зависит от n, при n = 2, 3, 4, 5 величина ***a*** = 1,000; 1,215; 1,414; 1,515.

Таким образом, в случае трехфакторного эксперимента ортогональный план второго порядка будет состоять из 8 опытов по ПФЭ, где каждый фактор будет варьировать на уровне -1 и +1; 6 опытов, в которых каждый фактор берегся в кодированных значения на уровнях О и i1,215 и один эксперимент проводится, когда все факторы имеют центральное значение 0. Как для ПФЭ, ДФЭ, так и для ортогональных планов второго порядка и других методов В теории математического планирования экспериментов разработаны стандартные планы, соответствующие различным ситуациям и целям проведения исследований. Разработаны такие планы и для отсеивающих экспериментов. Эти планы обеспечивают получение информации достаточной только для того, чтобы сравнить между собой степень влияния каждого фактора на исследуемый процесс. Анализ полученных результатов позволит выбрать для последующих экспериментов только наиболее значащие факторы.

Оценка результатов экспериментов

В экспериментальной работе, предусматривающей получение статистических уравнений регрессии, надо учитывать, что единственным ограничением здесь является количество опытов (число степеней свободы) и, если это условие соблюдено, уравнение регрессии любого вида может быть получено для любого результата, вплоть до таблицы случайных чисел. Для снижения влияния этой концепции в статистике и теории планирования экспериментов используют ряд оценочных показателей, в основе которых лежит анализ дисперсий полученных результатов.

Оценка воспроизводимости опытов необходима для того, чтобы определить, в какой степени обычные для любых экспериментов расхождения результатов повторностей одного и того же опыта обусловлены статистически случайными явлениями или являются следствием некорректной постановки опытов и влиянием неучтенного фактора(ов). Эта оценка в общем случае проводится по критерию Фишера, а в случае, когда все эксперименты проводятся с одинаковым числом повторностей, используют критерий Кохрена. Статистическая значимость коэффициентов регрессии проверяется с помощью критерия Стьюдента. Адекватность (степень соответствия) полученного уравнения регрессии экспериментальным результатам проверяется на основании критерия Фишера.

СРЕДА ПАКЕТА "STATISTICA NEURAL NETWORKS"

STATISTICA — это интегрированная система анализа и управления данными. STATISTICA — это инструмент разработки пользовательских приложений в бизнесе, экономике, финансах, промышленности, медицине, страховании и других областях. STATISTICA легка в освоении и использовании.

Все аналитические инструменты, имеющиеся в системе, доступны пользователю и могут быть выбраны с помощью альтернативного пользовательского интерфейса. Пользователь может всесторонне автоматизировать свою работу, начиная с применения простых макросов для автоматизации рутинных действий вплоть до углубленных проектов, включающих в том числе интеграцию системы с другими приложениями или Интернет. Технология автоматизации позволяет даже неопытному пользователю настроить систему на свой проект.

Процедуры системы STATISTICA имеют высокую скорость и точность вычислений.

Гибкая и мощная технология доступа к данным позволяет эффективно работать как с таблицами данных на локальном диске, так и с удаленными хранилищами данных.

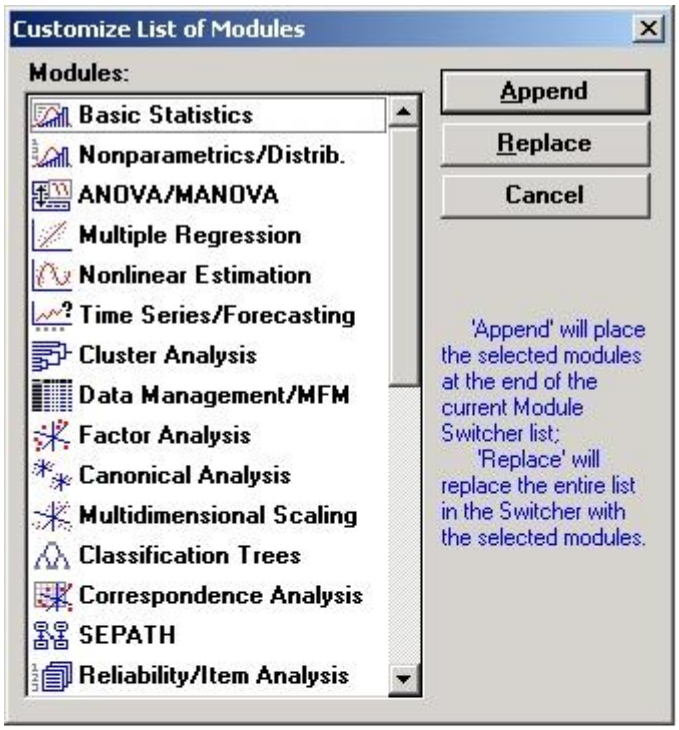
Система обладает следующими общепризнанными достоинствами:

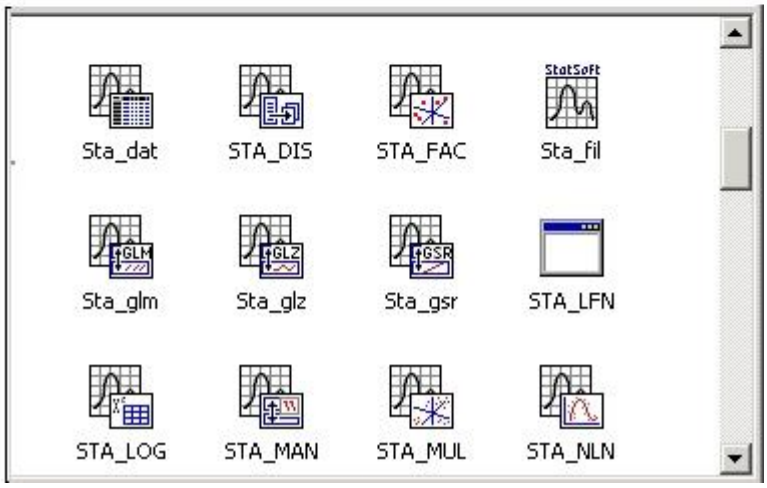
- содержит полный набор классических методов анализа данных: от основных методов статистики до продвинутых методов, что позволяет гибко организовать анализ;
- является средством построения приложений в конкретных областях;
- в комплект поставки входят специально подобранные примеры, позволяющие систематически осваивать методы анализа;
- отвечает всем стандартам Windows, что позволяет сделать анализ высокоинтерактивным;
- система может быть интегрирована в Интернет;
- поддерживает web-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 состоит из набора модулей, 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).

При исполнении модулей STATISTICA как самостоятельных приложений в любой момент времени в любом модуле имеется прямой доступ к «общим» ресурсам (таблицам данных, языкам BASIC и SCL, графическим процедурам).





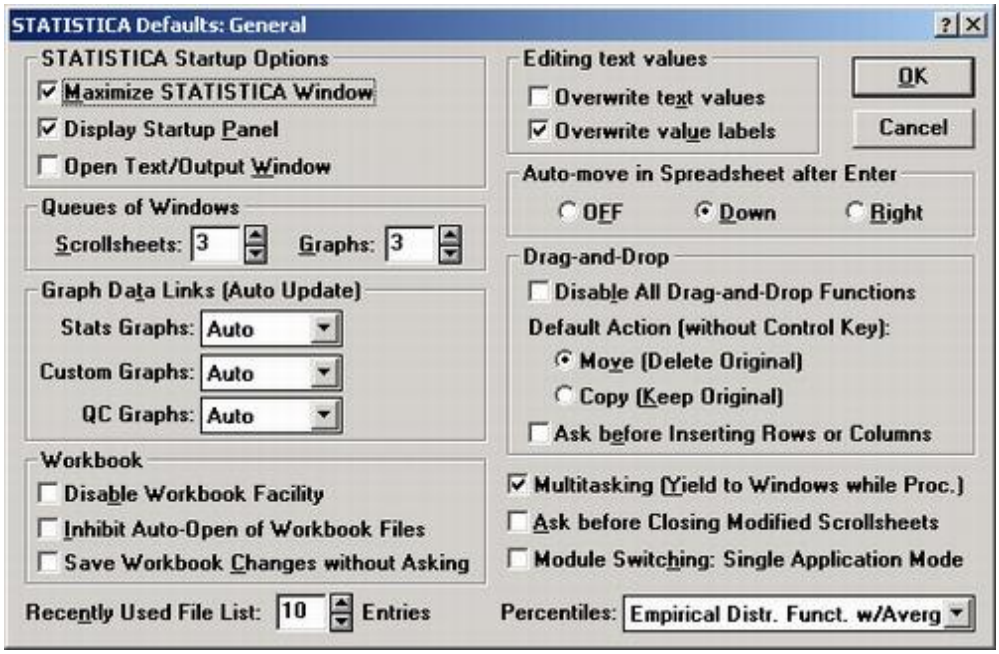
При инсталляции системы программа установки (**Setup**) создает на рабочем столе группу приложений под названием STATISTICA и помещает туда значки окна Переключатель модулей (пиктограмма STATISTICA — первая в группе, см. рис.), модуля Основные статистики и таблицы и некоторых других программ (**Help, Setup**). Пользователю может показаться более удобным запускать модули, щелкая по их значкам на рабочем столе (вместо того чтобы пользоваться окном Переключатель модулей); поэтому он, вероятно, захочет создать дополнительные пиктограммы для модулей помимо тех, которые будут автоматически созданы программой установки (**Setup**). Для того чтобы создать еще один значок в данной группе, следуйте стандартной процедуре **Windows** (выберите пункт Новый в меню **Файл** в окне **Диспетчер** программ (**Program Manage**r) и создайте новый программный элемент).

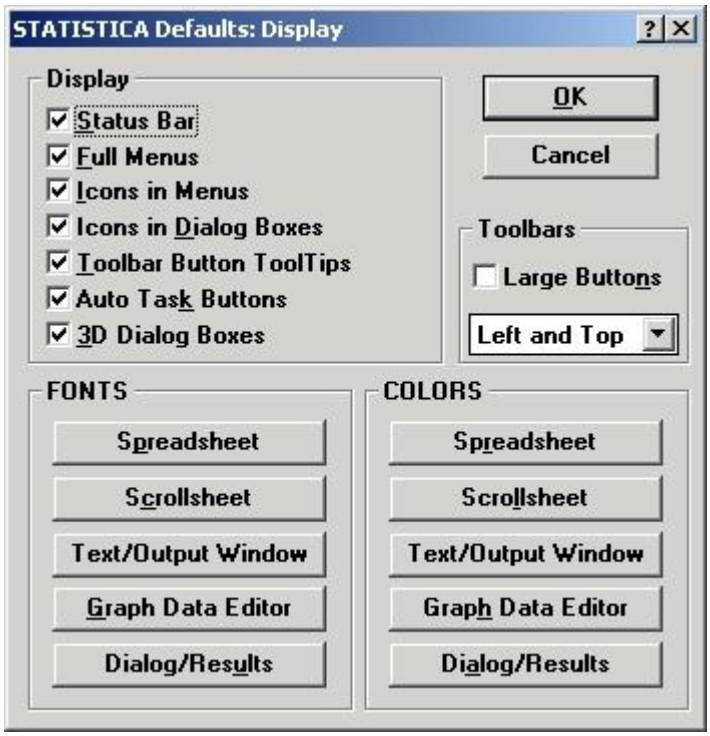
Настройка системы STATISTICA. В системе предусмотрена возможность настройки множества характеристик и интерфейса программы в соответствии с предпочтениями пользователя. Можно изменить, например, процесс запуска, а именно — отменить установленный по умолчанию полноэкранный режим, изменить вид стартовой 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).

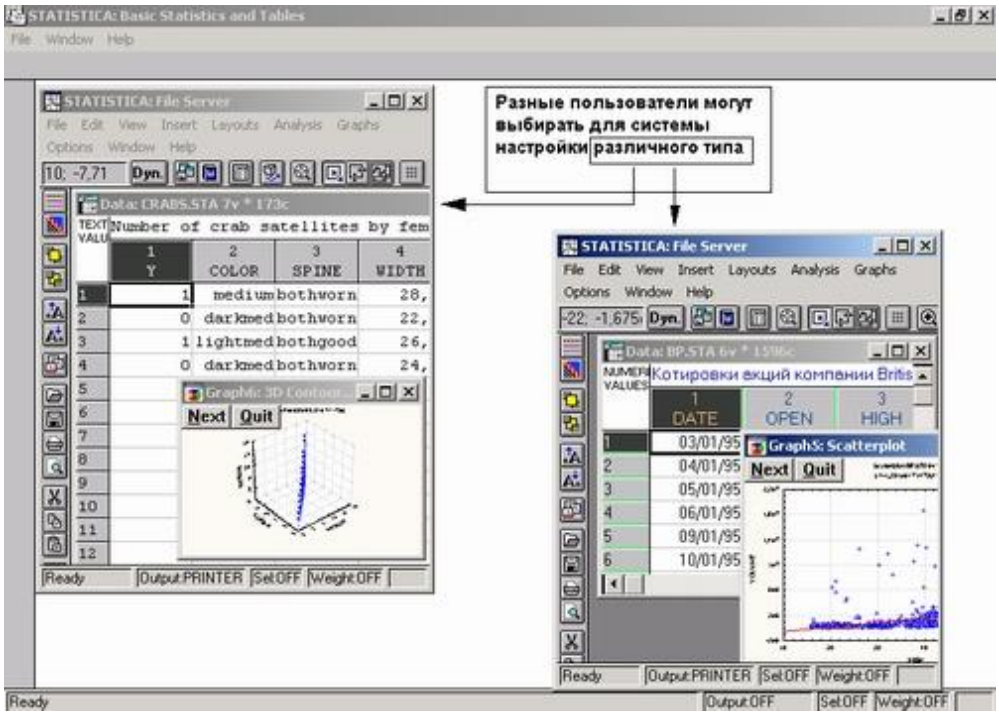
Каждый из этих параметров можно настроить в соответствующем окне, доступ к которому осуществляется через меню **Сервис**. На следующих рисунках показаны два примера таких окон.



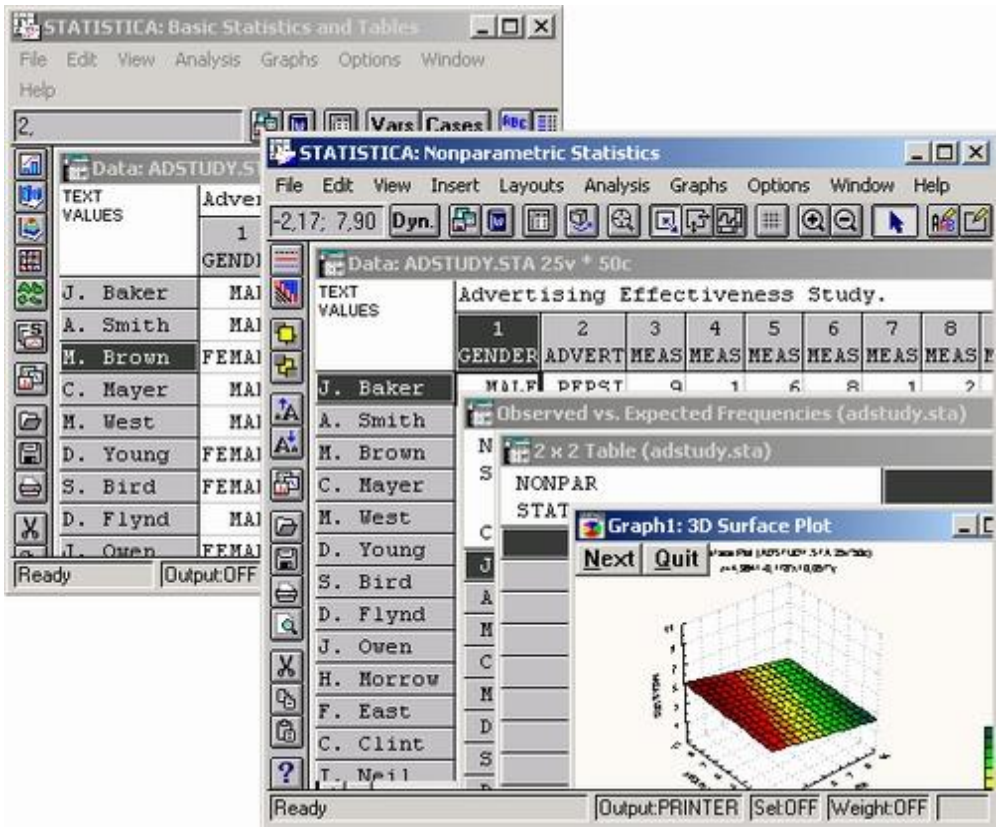


Все общие параметры могут быть настроены независимо от типа окна документа (например, таблица или график), которое активно в данный момент.

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.



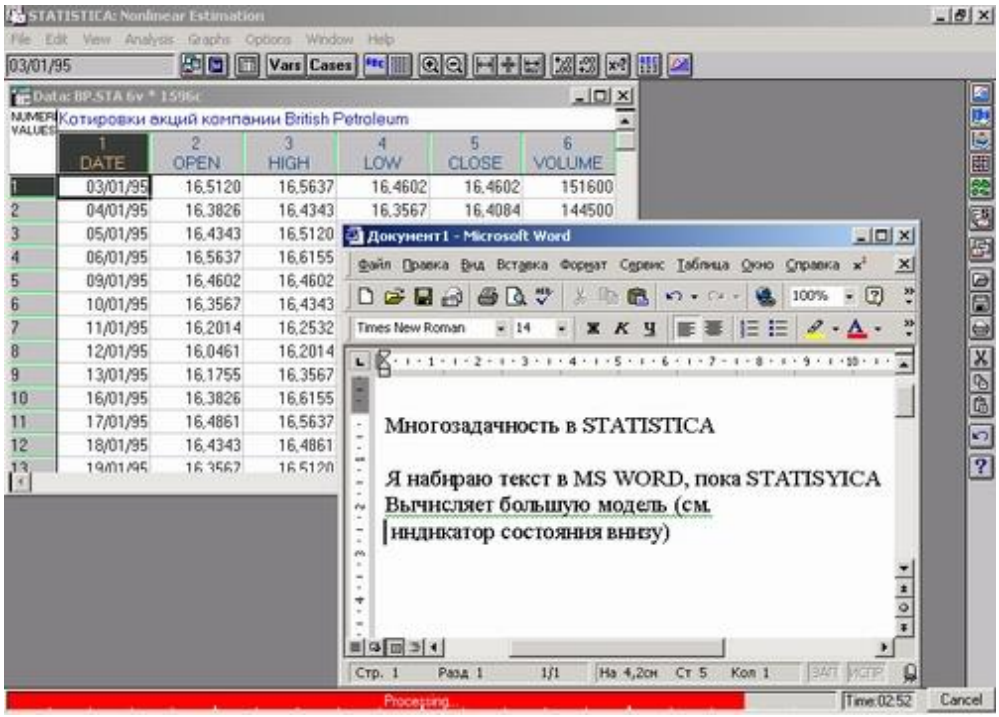
В зависимости от требований задачи и личных предпочтений (а также эстетических соображений) можно использовать разнообразные «режимы» и условия работы программы.



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 (**Working Directory)** (с помощью диалогового окна системы **Windows** Свойства программного элемента **(Program Item Properties**)).

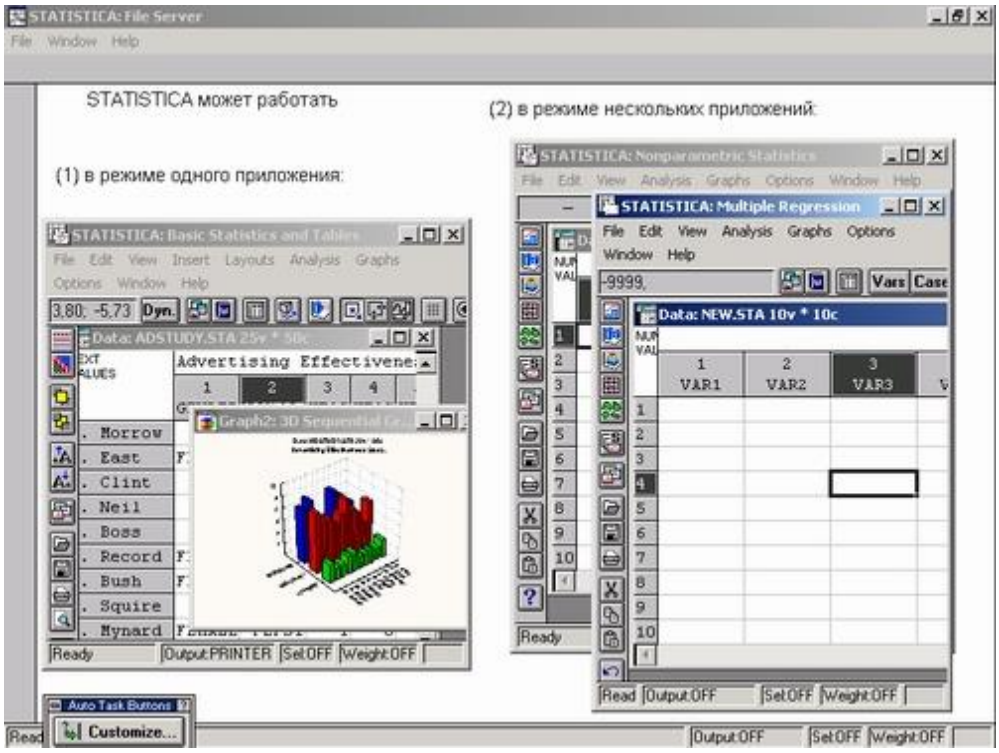
Многозадачность. STATISTICA поддерживает режим многозадачности (между своими модулями или другими приложениями).



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



Режим одного приложения. При выбранном режиме одного окна приложения переключение с одного модуля на другой будет происходить без открытия новых окон. Новый модуль всякий раз будет открываться в том же самом окне, заменяя предыдущий. Некоторые пользователи предпочтут именно такой «простой» режим работы, поскольку весь анализ будет происходить в одном окне приложения, а количество активных программ на рабочем столе будет минимальным.

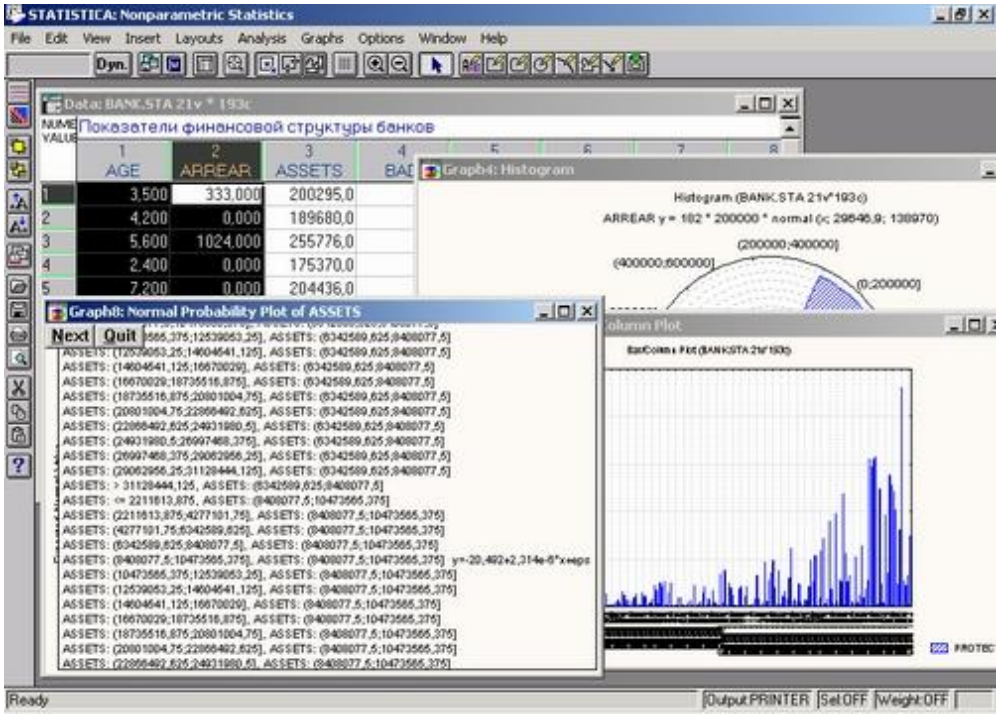
Примерно такого же эффекта можно достичь, нажимая кнопку **Закончить** и переключиться в диалоговом окне Переключатель модулей; при этом окно приложения текущего модуля закроется, но не будет заменено новым окном; вместо этого система откроет «следующее» окно приложения.

Режим нескольких приложений. Основное преимущество режима нескольких приложений — возможность параллельного выполнения различных процедур анализа (модули) в разных одновременно открытых окнах приложения. При этом можно переключаться между модулями, не закрывая предыдущие, и использовать все преимущества работы с независимыми очередями таблиц результатов и графиков для окон приложений разных модулей. Этот режим имеет очевидные преимущества для большинства задач анализа данных и дает возможность использовать различные методы анализа (и сравнивать полученные результаты).

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

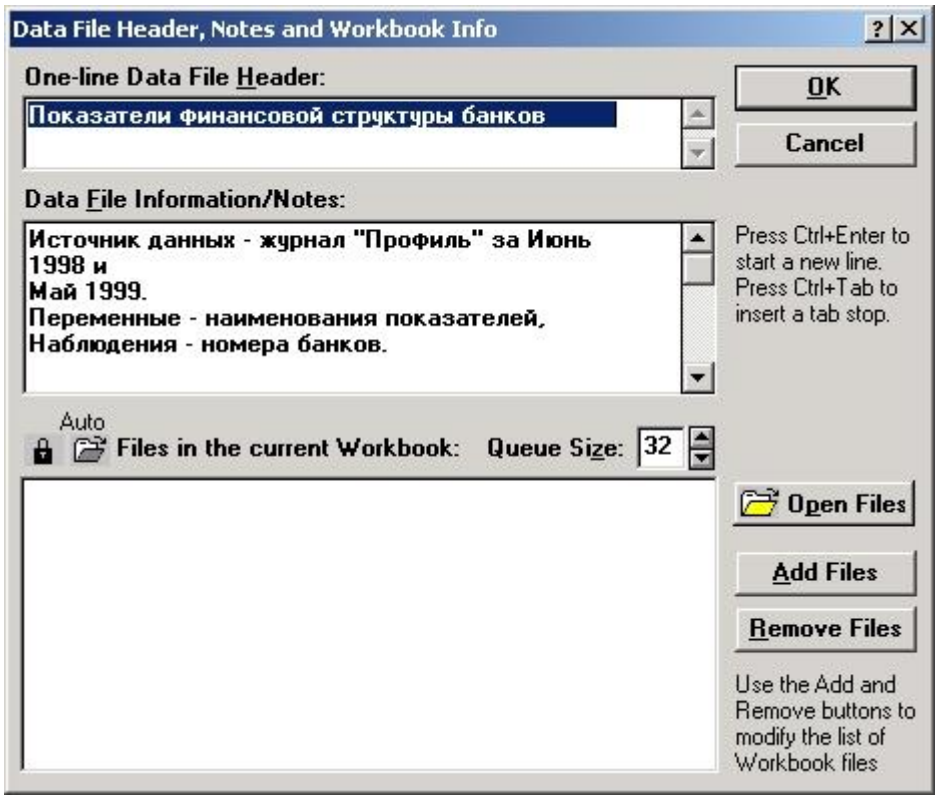
моделирующий анализ уже после того, как данные обработаны и получены основные результаты.

Количество выводимых окон также может быть настроено, чтобы не перегружать экран компьютера.



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.

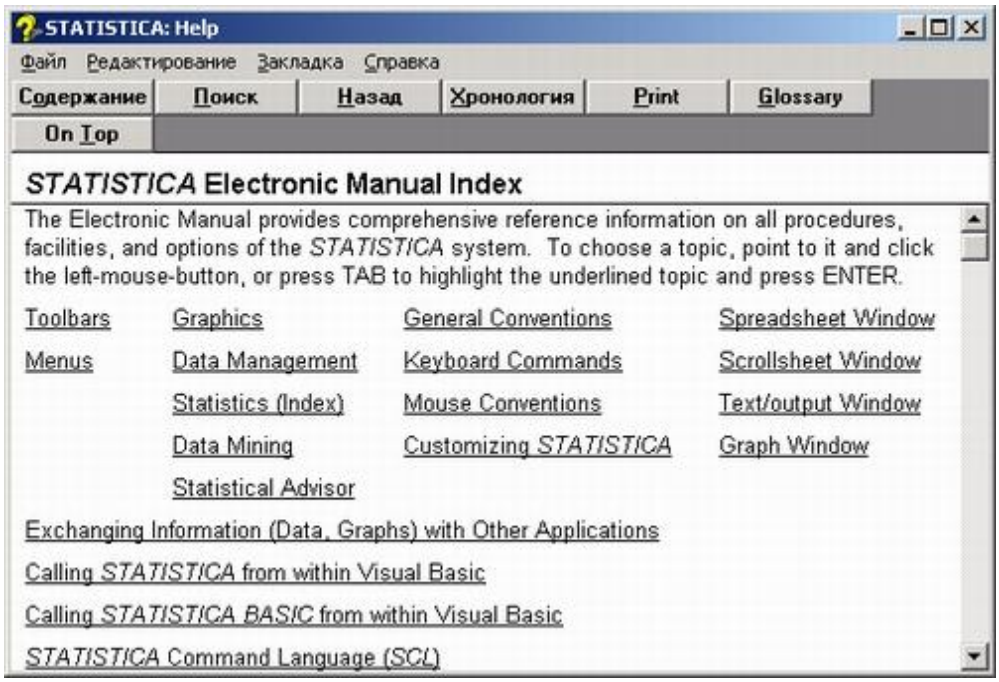
Какие возможности предоставляют рабочие книги. Рабочие книги помогают организовывать наборы файлов (например, таблиц результатов, графиков, текстовых/графических отчетов, пользовательских программ и т. д.), которые были созданы или использовались (например, просматривались) во время анализа набора данных. Рабочие книги хранят список всех файлов, использовавшихся с текущим набором данных.



Обновленный список этих файлов автоматически сохраняется с файлом данных. Если поставить пометку в поле **Авто** около имени файла, то он будет автоматически открываться с текущим набором данных.

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). Благодаря динамической организации Электронного руководства с помощью гиперссылок (и различным возможностям его настройки), как правило, быстрее использовать эту справочную систему, чем искать нужную информацию в напечатанном виде. Справку также можно вызвать двойным щелчком на поле сообщений строки состояния в нижней части окна приложения STATISTICA (в поле сообщений тоже отображаются краткие комментарии о функциях выпадающих меню или кнопках панели инструментов соответственно при выделении пункта меню или нажатии кнопки).

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).



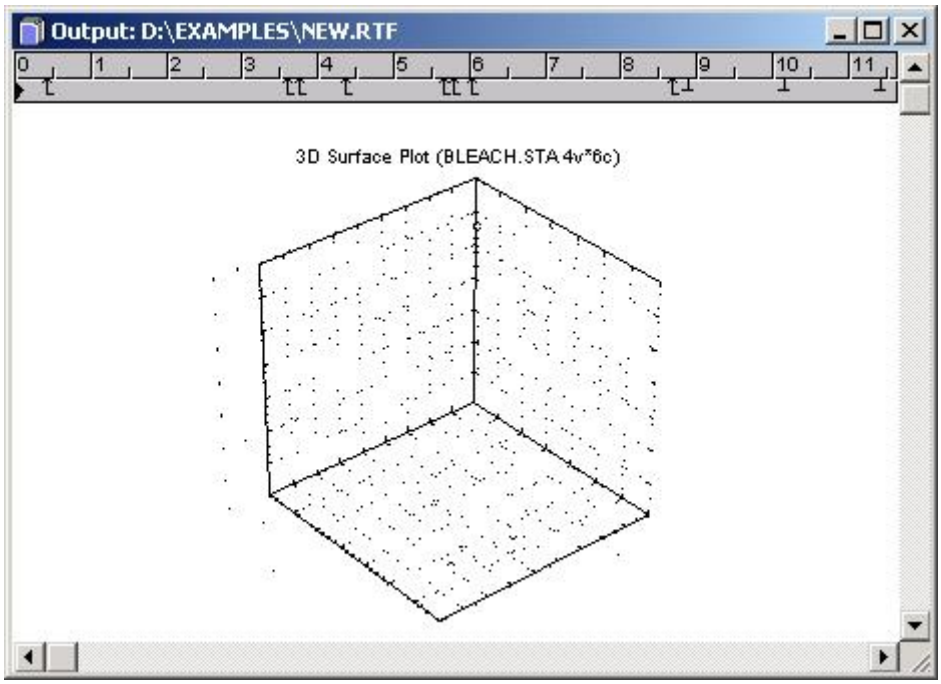
С помощью гиперссылок можно непосредственно перейти из раздела Статистический советник к подробному описанию соответствующих статистических методов и процедур в разделе Вводный обзор.

Приложения. Все рассмотренные возможности (доступные в любой момент работы с системой) могут служить весомой альтернативой или дополнением к обычному интерактивному пользовательскому интерфейсу, поскольку они позволяют автоматизировать рутинный процесс многократного выполнения одних и тех же, в том числе весьма сложных, задач. Например, макрокоманда (вызываемая щелчком мыши по кнопке на панели инструментов Кнопки автозадач или одним нажатием клавиши) может содержать длинный список переменных, часто используемый график, операцию внедрения и т. п.

Автоматические отчеты и автоматическая распечатка таблиц результатов. Независимо от того, происходит ли 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.

Автоматическая печать графиков. Режим автоматической печати всех возникающих на экране графиков особенно полезен как средство пакетной графической печати.



Как правило, печать графиков занимает довольно много времени. Поэтому имеет смысл воспользоваться этим режимом для распечатки последовательности («каскада») графиков, получающихся при применении определенных методов анализа (например, для зрительного представления конфигураций средних при исследовании связей высших порядков в дисперсионном анализе необходима длинная последовательность графиков, а для многомерных таблиц — каскад трехмерных гистограмм для двух переменных).

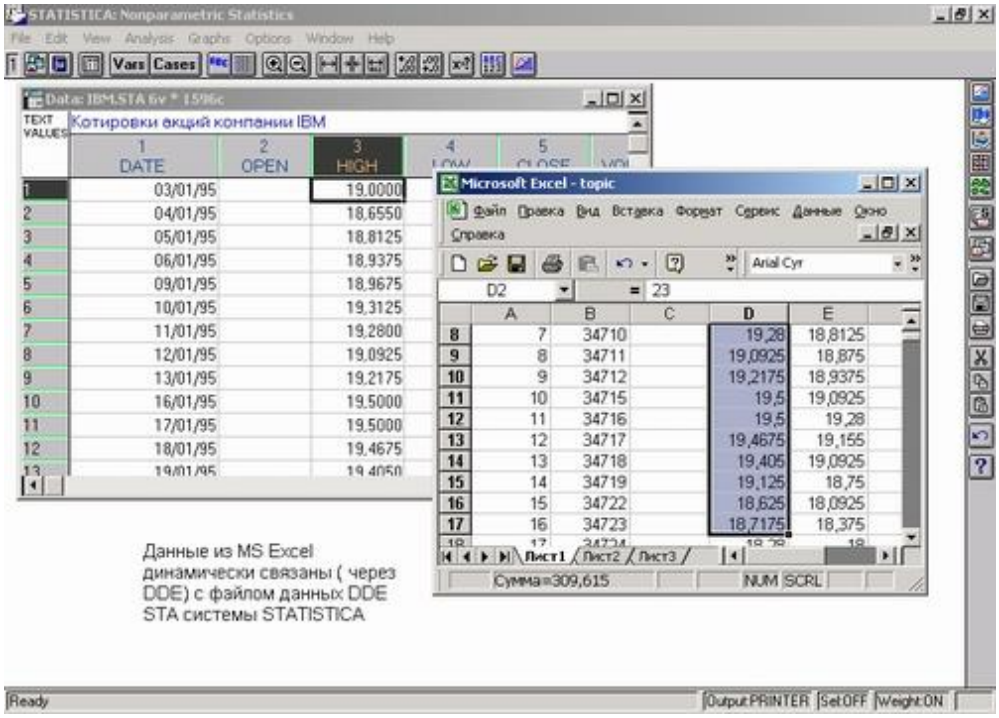
Однако гораздо эффективнее направить создаваемую последовательность графиков в Окно текста/вывода. В STATISTICA предусмотрена возможность пакетной печати всех ранее сохраненных графиков и таблиц результатов; для этого нужно выбрать пункт Печать файлов в выпадающем меню **Файл**.

Буфер обмена. Наиболее быстрый и во многих случаях наиболее простой способ получения данных из других приложений **Windows** (например, электронных таблиц) — это использование буфера обмена, который в STATISTICA поддерживает специальные форматы данных, создаваемые такими приложениями, как MS Excel или Lotus для Windows. Например, STATISTICA правильно интерпретирует форматированные (например, 1 000 000 или $10) и текстовые значения. Буфер обмена и преобразование файлов данных можно также использовать для экспорта данных из системы STATISTICA в другие форматы. При импорте и экспорте данных STATISTICA использует один и тот же набор форматов и типов данных.

Функции импорта файлов. Файлы данных из приложений **Windows** и других операционных систем также можно переводить в формат системы STATISTIC A с помощью 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:

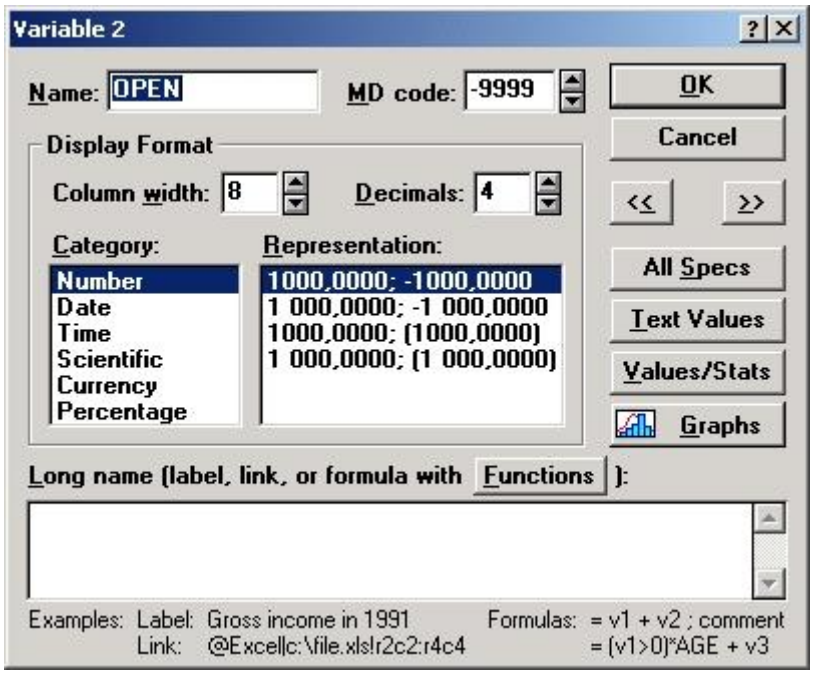
- он позволяет пользователю точно указать, как должен проводиться импорт (например, выбирать из файлов диапазоны значений, импортировать или не импортировать имена переменных, текстовые значения и имена наблюдений и указывать способ их интерпретации);
- он предоставляет пользователю доступ к типам данных,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

Связи DDE. STATISTICA поддерживает соглашения динамического обмена данными (DDE), что позволяет динамически связывать диапазон данных в таблице исходных данных с набором данных других приложений (Windows). Эта процедура на самом деле гораздо проще, чем она может показаться, и ее легко освоить, не имея технических знаний о механизме DDE, особенно при использовании команды Установить связь (вместо ввода описания связи). 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.

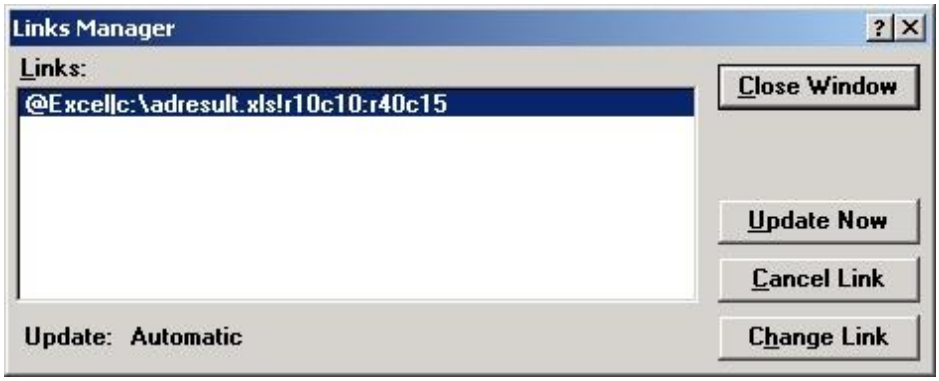


Обычно два файла динамически связываются в промышленных установках, когда к последовательному порту компьютера, на котором находится файл данных системы STATISTICA, подключено измерительное устройство (например, для ежечасного автоматического обновления определенных измерений).

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.

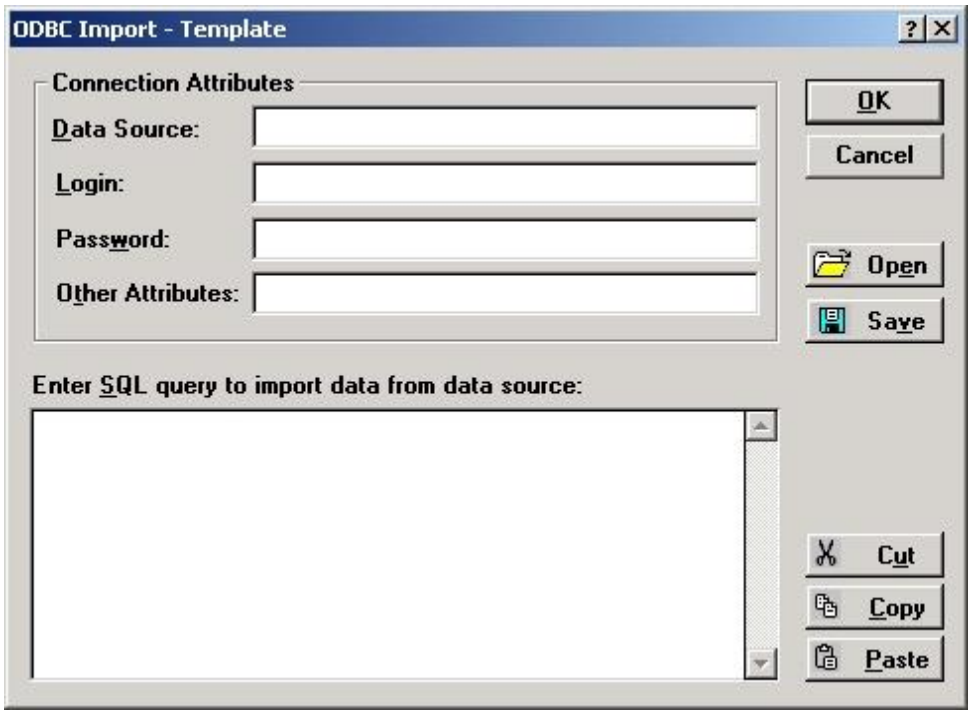


Если связь установлена, то можно управлять ею в диалоговом окне **Диспетчер связей** (вызывается с помощью команды **Связи**... выпадающего меню **Правка**).



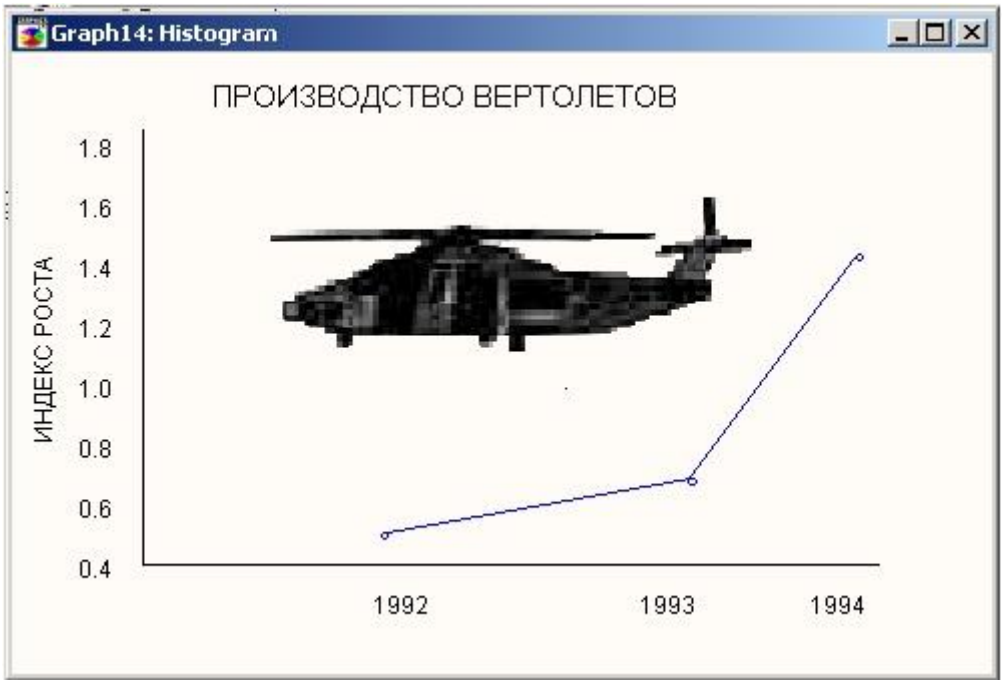
Форматы Дата и Время. В файлах данных системы (которые организованы как базы данных) формат отображения значений применяется ко всей переменной, а не к отдельным ячейкам (как в Excel). Поэтому значения, которые в Excel были отформатированы как даты, в файле системы STATISTICA будут отображаться как юлианские (целые) значения (например, 34092 вместо May 3, 1993), если для соответствующих переменных не установлен формат **Дата** или **Время**.

Поддерживает ли STATISTICA интерфейс ODBC? Да, для того чтобы реализовать эту возможность, существует список команд Импорт данных, который вызывается из выпадающего меню Файл любого модуля. Интерфейс ODBC STATISTICA включает возможности для объединения полей из нескольких таб лиц и предоставляет доступ к множеству файлов баз данных, включая форматы больших и персональных компьютеров (например, dBASE для Windows, Paradox, Sybase, Oracle, SAS и т. д.).

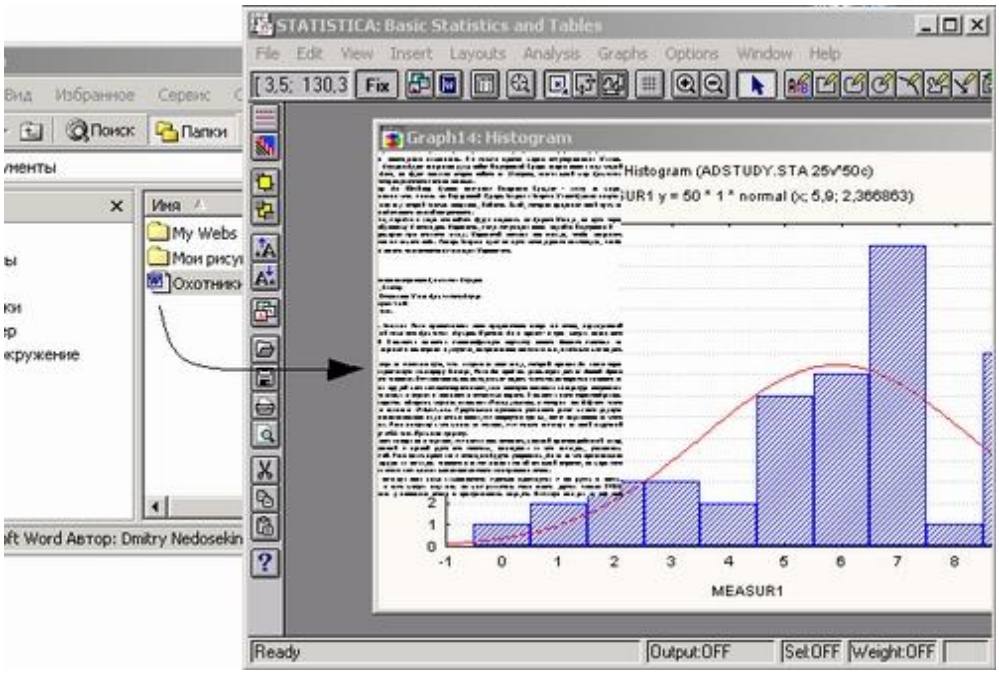


Импорт через ODBC можно автоматизировать с помощью функции ODBC/Шаблоны или программ на языке SCL.

Типы объектов. Если задан режим Новый объект, то тип создаваемого объекта может быть выбран из списка приложений Windows, которые поддерживают средства OLE. После выбора типа и нажатия кнопки **ОК** будет открыто окно соответствующего приложения для создания нового объекта. Если задан режим Объект из файла, то тип объекта для вставки также выбирается из списка приложений Windows, поддерживающих средства OLE; после выбора типа будут показаны все предварительно сохраненные файлы этого приложения. В режиме Картинка из файла можно вставить объект, несовместимый с методом OLE, но записанный в одном из графических форматов Windows: в формате метафайла (файл с расширением *.wmf) или растрового изображения (файл с расширением *bmp).



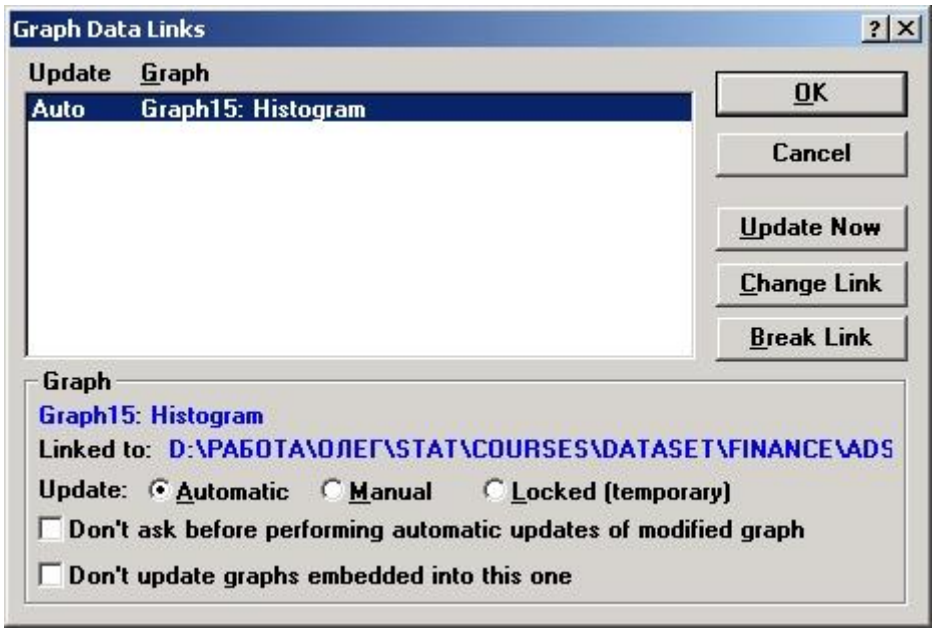
Связывание и внедрение. STATISTIC A поддерживает средства OLE (связывания и внедрения объектов) как в режиме клиента, так и в режиме сервера. Таким образом, возможна не только динамическая настройка графиков STATISTICA в других приложениях (режим сервера), но также внедрение и последующее преобразование OLE-совместимых объектов других приложений (например, графиков или таблиц) или собственных объектов в графики STATISTICA. Другими словами, помимо присоединения внешних элементов к графикам STATISTICA с помощью вставки можно обращаться непосредственно к объектам, содержащимся в файле на диске (например, перетащить их непосредственно из окна **Диспетчер файлов** или **Проводник** (Windows Explorer) и поместить на график STATISTICA).



STATISTICA поддерживает как связанные (то есть динамически присоединенные), так и внедренные (то есть статически «встроенные») объекты. При этом они могут быть расположены в любом файле, созданном приложениями Windows, включая файлы в собственном графическом формате STATISTICA (с расширением *.stg). Более того, STATISTICA одновременно может являться как клиентом, так и сервером в методе OLE, поддерживая при этом уникальную возможность создания вложенных составных документов (до четвертого порядка включительно), то есть документ STATISTICA с внедренным документом может быть, в свою очередь, внедрен в другой документ этой системы.

Заметим, что каждый из этих двух способов присоединения (связывание и внедрение) имеет свои преимущества и недостатки.

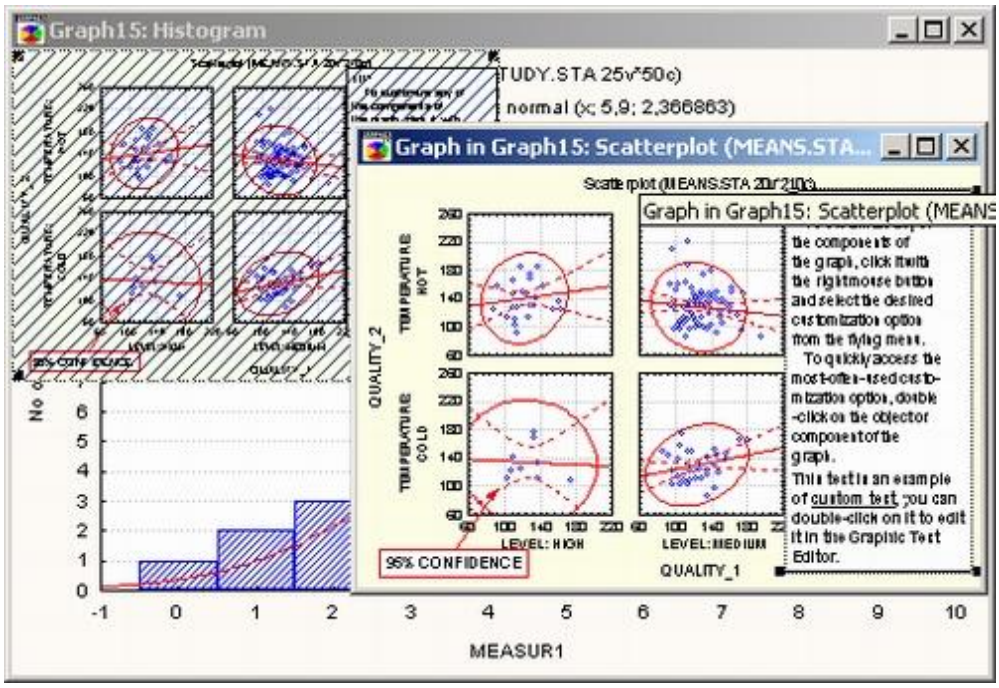
Связанные объекты. Графики со связанными объектами медленнее перерисовываются, поскольку при этом могут быть задействованы связи с внешними файлами. В то же время эти графики обновляются автоматически (статус связей может быть установлен в диалоговом окне Связи данных и графика, которое вызывается из графического меню Правка), а это позволяет легко создавать составные документы, которые включают именно «текущее» содержимое других файлов.



Внедренные объекты. Графики с внедренными объектами перерисовываются быстрее, чем со связанными объектами, поскольку здесь отсутствуют связи с обновляемыми внешними файлами. Если дважды щелкнуть на внедренном объекте, то будет вызвано приложение-сервер (то есть источник), в котором можно изменить данный объект. При этом обновить внедренный объект можно двумя способами: отредактировать его или заменить вручную.

В меню **Правка** можно настроить все параметры внешних объектов {связанных или внедренных), а также их связи с другими компонентами графика. Кроме того, щелкнув на объекте правой кнопкой мыши, можно выбрать нужные команды настройки из контекстного меню. Единственным исключением является способ присоединения объекта (связывание или внедрение), который определяется в момент подключения файла (после этого только связанный объект можно преобразовать во внедренный, но не наоборот (см. команду Преобразовать во внедренный из выпадающего меню **Правка**)).

Настройка связанных или внедренных объектов OLE. Объекты OLE-графиков STATISTICA могут быть отредактированы после двойного щелчка мышью на объекте; при этом приложение-источник будет открыто в режиме сервера OLE с готовым к редактированию объектом. Если этот объект является графиком STATISTICA, то в текущем модуле откроется новое графическое окно, что позволит системе одновременно выступать как в роли клиента, так и сервера.



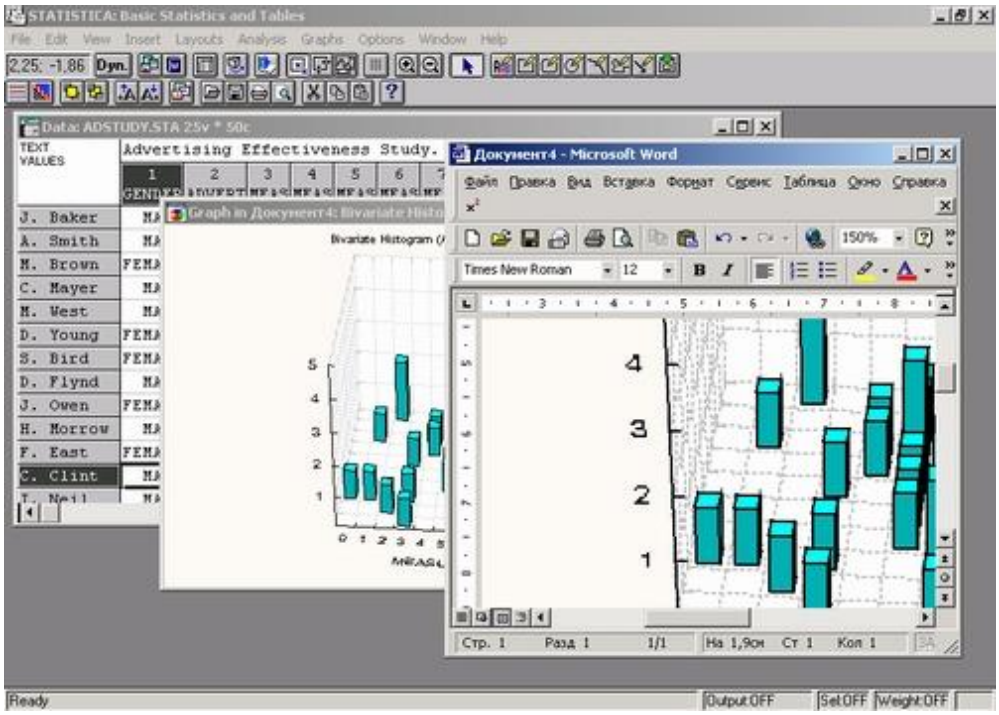
Когда редактирование завершено, можно применить любое из стандартных соглашений OLE для выхода из режима сервера и обновления графика в системе STATISTICA (используя команды Обновить, Обновить и вернуться к... и т. д. в выпадающем меню приложения Файл; эти команды доступны только в случае, если приложение запущено в режиме сервера).

Графические форматы Метафайл и Растровое изображение. Для вставки графического файла в приложения, не поддерживающие методы OLE, используются команды Сохранить метафайл или Сохранить растровое изображение (из выпадающего графического меню Файл). График в формате метафайла Windows будет записан в файл с расширением *.wmf, а в формате растрового изображения — с расширением *.bmp. Эти форматы, описанные в двух следующих параграфах, не позволяют полностью реализовать все возможности настройки графиков STATISTICA, но в то же время совместимы со всеми приложениями, поддерживающими графические форматы Windows.

Что такое метафайл Windows? Графический формат Метафайл — это один из стандартов для записи графических файлов (с расширением *.wmf) и их представления в буфере обмена Windows. Он содержит картинку в виде описаний и определений всех компонент графика и его атрибутов (например, элементов линий, их цветов и шаблонов, шаблонов заполнения, описания текста и его параметров).

По сравнению со стандартом растрового изображения (см. ниже) формат метафайла дает возможности более гибкой настройки OLE-несовместимых объектов в приложениях Windows.

Например, при открытии метафайла в программе Microsoft Draw можно «разложить» изображение графика, выделить и изменить отдельные линии, шаблоны заполнения или цвета, а также отредактировать текст и изменить его атрибуты.



Однако не все приложения Windows полностью поддерживают все возможности формата метафайла, доступные в системе STATISTICA. Некоторые параметры графиков, записанных системой STATISTICA в этом формате, могут измениться при их воспроизведении в других приложениях. Например, может исчезнуть поворот некоторых шрифтов. Поэтому по возможности используйте графический формат STATISTICA и методы 01£для работы с графиками в других приложениях, чтобы иметь доступ ко всем возможностям настройки самой STATISTICA.

Ограничения стандартного формата Метафайл Windows. Сложные графические изображения, создаваемые системой STATISTICA, могут оказаться слишком большими (по числу представленных точек данных) для записи в формате метафайла, который по умолчанию используется системой Windows для большинства операций по связыванию и внедрению графических объектов. В таких случаях нужно использовать растровое изображение. За дополнительной информацией обратитесь к Электронному руководству из диалогового окна Дополнительные параметры, которое вызывается из вкладки Графика диалогового окна Параметры страницы/вывода.

Что такое формат растрового изображения? Формат Растровое изображение — это второй стандартный графический формат системы Windows, который используется для представления графических файлов (с расширением *.bmp) и передачи изображения через буфер обмена (как и формат Метафайл). В этом формате не сохраняются никакие дополнительные данные или параметры, кроме изображения самой картинки.

В отличие от метафайла растровое изображение представляет собой «пассивное» поточечное отображение графического окна. Возможности настройки такого графика в других приложениях Windows очень ограничены. Обычно они включают только операции растяжения, сжатия, вырезания, вставки и рисования поверх графика. Как уже отмечалось выше, для работы с графиками в других приложениях удобнее использовать запись в графическом формате STATISTICA и методы OLE, чтобы иметь доступ ко всем возможностям настройки самой системы STATISTICA.

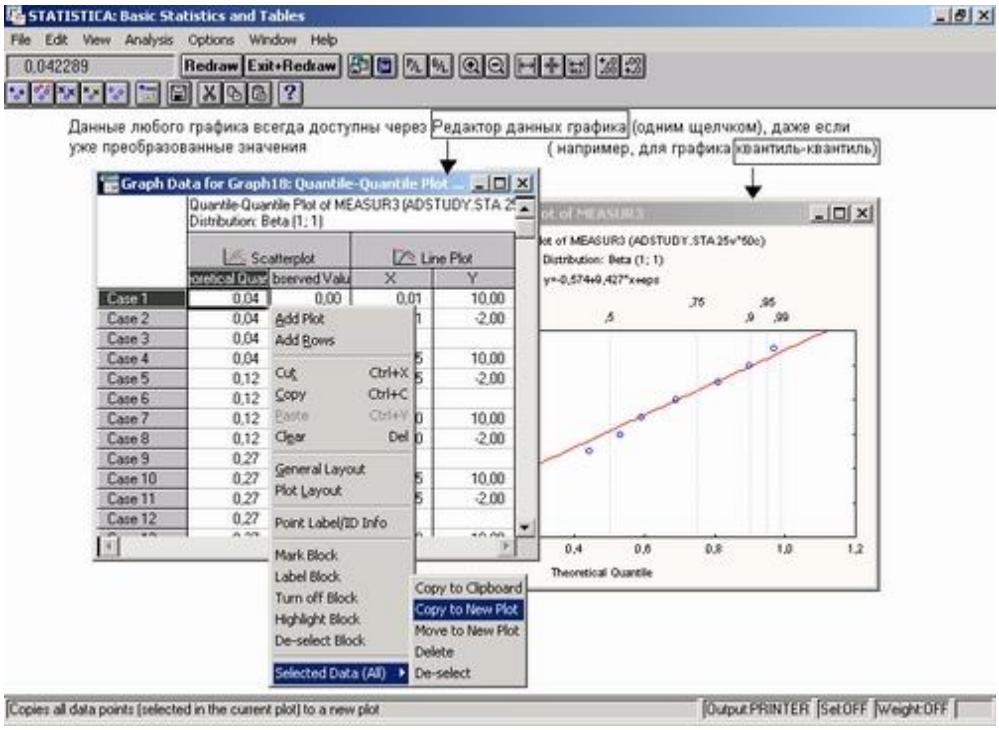
Что такое собственный графический формат STATISTICA? Графические файлы системы STATISTICA имеют расширение *stg. Их основное отличие от метафайлов и растровых изображений состоит в том, что они содержат не только картинку, но и всю информацию, необходимую для настройки графика и анализа данных. Здесь записаны все представленные на графике данные, их связи, уравнения подгонки, параметры внедренных объектов, связи графиков и рисунков и т. п. Записанные в таком формате графики можно впоследствии открыть в любом из модулей системы STATISTICA для продолжения настройки и анализа данных. Кроме того, их можно распечатать в пакетном режиме с помощью команды **Печать** файлов из выпадающего меню **Файл**. Графические файлы в собственном формате системы STATISTICA можно динамически связать с документами приложений Windows с помощью методов OLE.

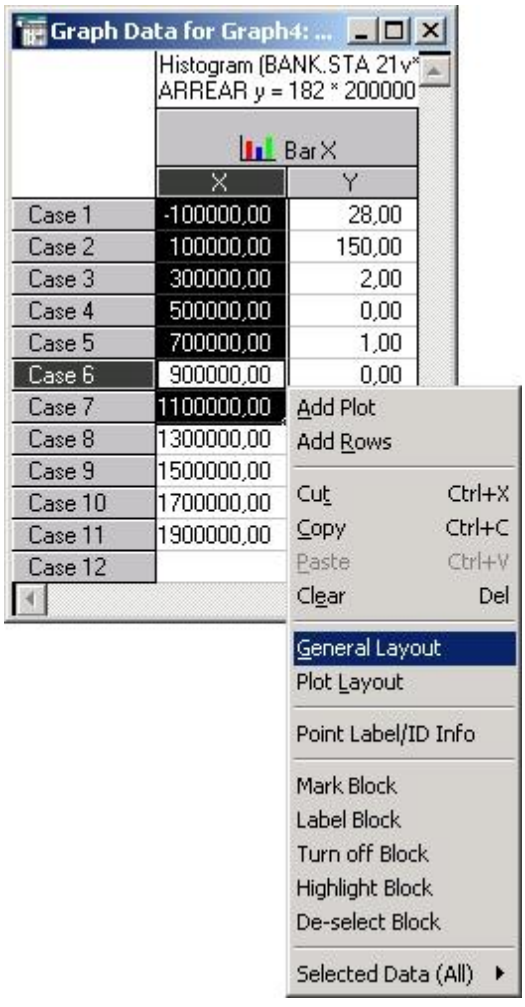
Экспорт через буфер обмена (вставка или специальная вставка методами OLE). Использование буфера обмена — это самый быстрый способ экспорта графика в другое приложение. При копировании в буфер обмена создается три графических представления объекта: в собственном формате STATISTICA, в формате метафайла Windows и в формате растрового изображения. Каждое из них может быть использовано в других приложениях.

Графики системы STATISTICA могут присутствовать в других приложениях (редакторах или электронных таблицах) как в качестве связанных, так и внедренных объектов. При использовании методов OLE они сохраняют свою связь с системой STATISTICA и, следовательно, могут интерактивно редактироваться в рамках других приложений.

Доступ ко всем данным графика. Данные, представленные на графиках системы, можно непосредственно просматривать и изменять независимо от их типа во встроенном Редакторе данных графика. Это могут быть исходные данные, части таблицы результатов или ряд рассчитанных значений (например, вероятностный график).

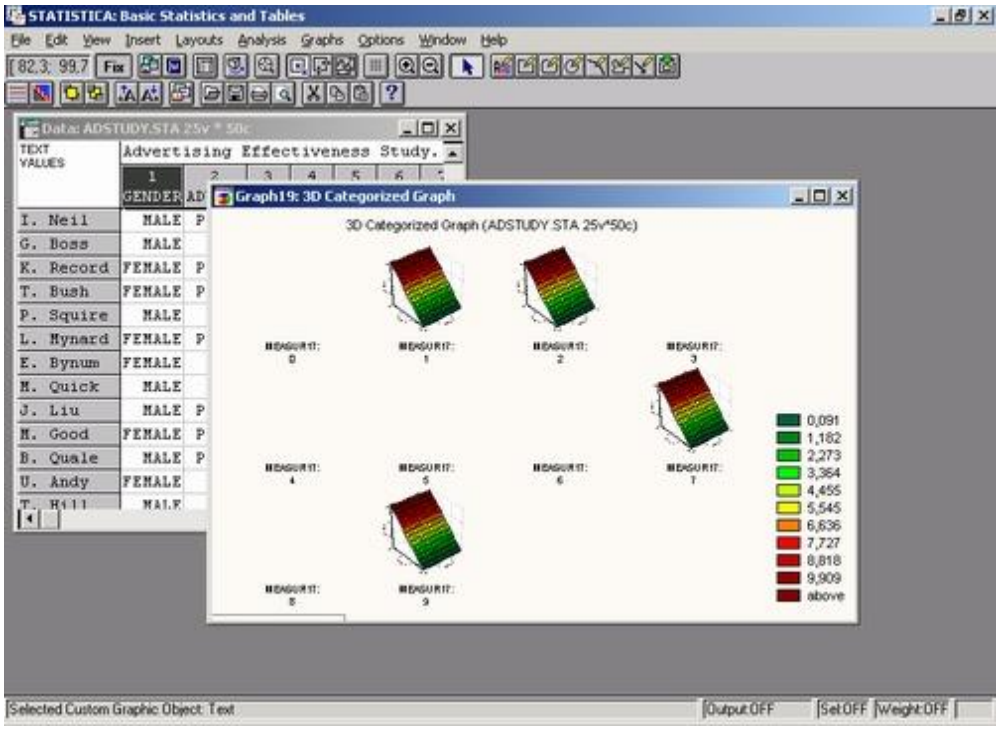
Для каждого графика создается связанное с ним «дочернее» окно Редактора, которое закрывается вместе со своим графическим окном. Редактор организован в виде групп столбцов, представляющих отдельные зависимости данного графика (см. следующий параграф).





Категоризованные графики. Для создания категоризованных графиков данные разбиваются на подгруппы. На одном изображении будет одновременно представлено несколько графиков, по одному для каждой из заданных подгрупп. Например, можно построить графики отдельно для субъектов мужского и женского пола, разделить

пациентов на группы женщин с высоким давлением, женщин с низким давлением, мужчин с высоким давлением, разделить товары по качеству, странам-производителям и т. п. Разбиение данных на однородные группы и исследование связей между этими группами — чрезвычайно важный прием анализа данных.



Категоризованные графики широко применяются в системе STATISTICA:

- Они доступны в большинстве диалоговых окон с результатами анализа (эти графики автоматически создаются в тех процедурах, где анализируются группы или подгруппы данных, например, при классификации, проверки t-критериев, в дисперсионном, дискриминантном и непараметрическом анализе).
- Эти типы графиков присутствуют в списке Быстрые статистические графики в контекстных меню всех таблиц исходных данных и таблиц результатов.
- Их можно вызвать из списка Статистические графики (в выпадающем меню , Графика), при построении которых предлагается большой выбор различных методов категоризации данных.

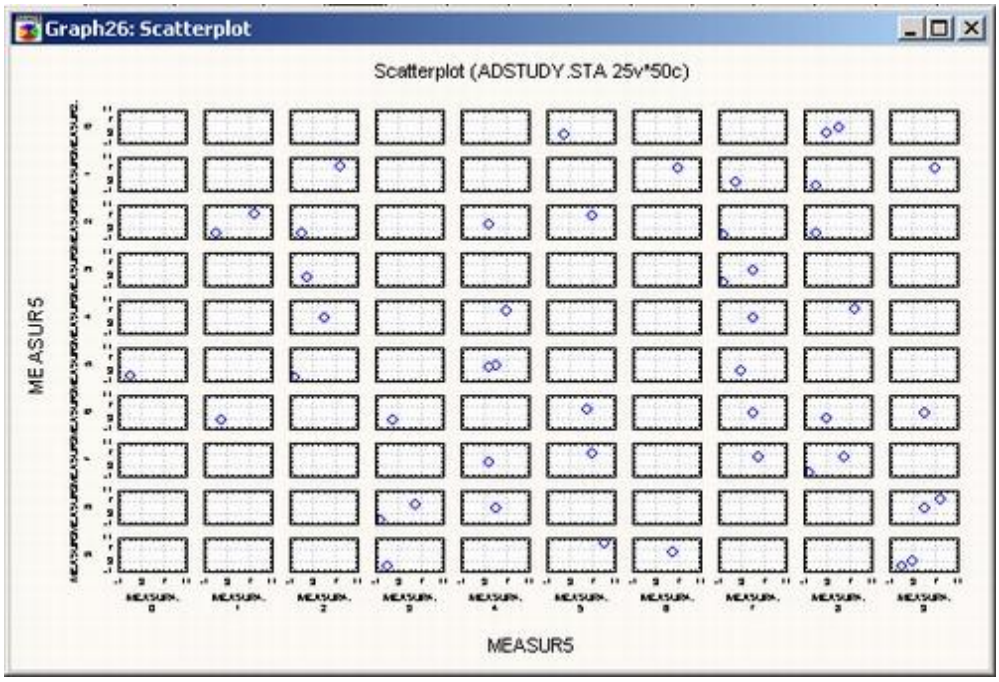
Методы категоризации, предлагаемые в системе STATISTICA, описаны в следующем пункте.

Каким образом задаются «категории» для категоризованных графиков? Итак, вначале нужно разбить данные на группы. При построении категоризованных графиков из диалоговых окон с результатами анализа подгруппы данных определяются автоматически (поскольку такое разделение является частью исследования данных). При построении статистических графиков предлагаются различные способы задания подгрупп по одной или двум группирующим переменным. Кроме того, разбиение на подгруппы может организовать сам пользователь, используя любые комбинации переменных из текущего набора данных.

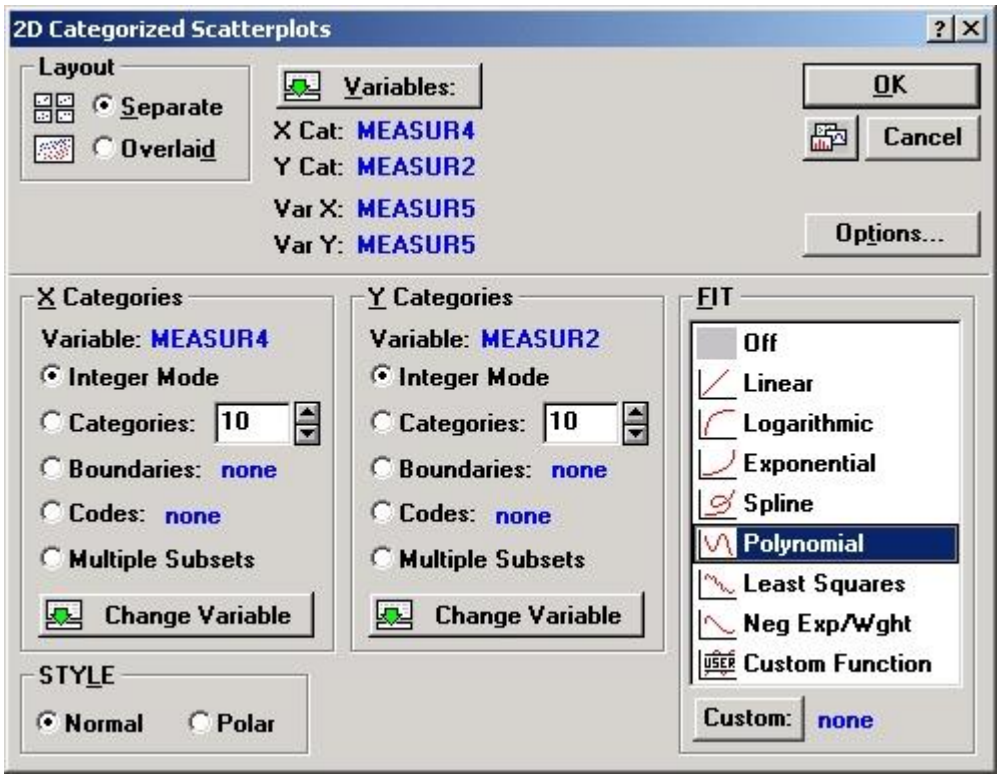
Существует несколько методов выделения категорий:

- по целым значениям группирующих переменных (**Целые числа**),
- разделением группирующих переменных на заданное число интервалов (**Категории**),
- разделением группирующих переменных на интервалы с заданными граничными значениями (**Границы**),
- с помощью задания конкретных значений (кодов) группирующих переменных (**Коды**),
- путем формирования сложных подгрупп (**Сложные подгруппы**); для этого пользователь может ввести условия выбора наблюдений практически неограниченной сложности и использовать значения любой переменной текущего файла данных, как показано ниже.

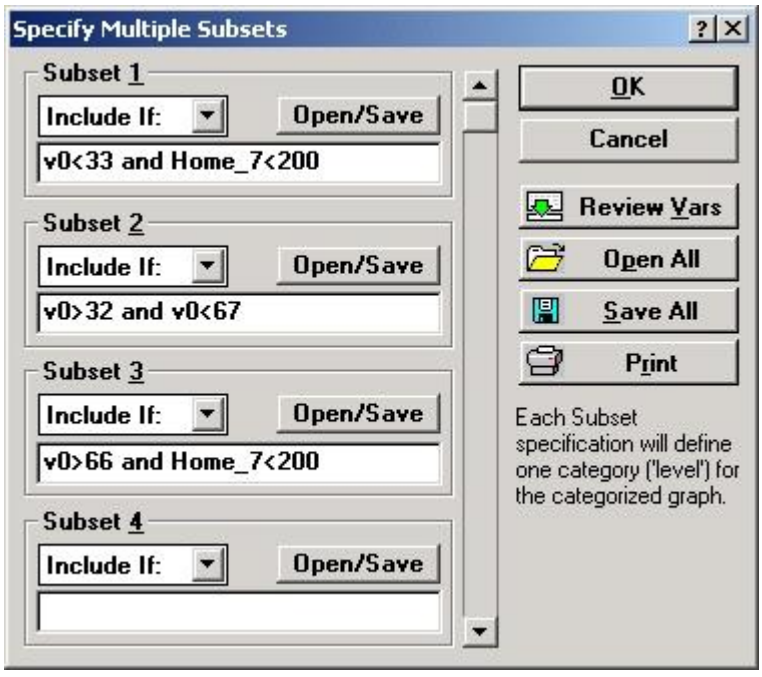
На следующем рисунке показан достаточно сложный график, категоризованный по двум признакам. При этом использован смешанный метод выделения подгрупп. Категоризация по двум признакам означает, что элементы графика располагаются как элементы двухвходовой таблицы, полученной после использования двух различных методов категоризации.



Две строки на приведенном графике представляют разделение на подгруппы по значениям переменной Measur5 . Три столбца графика представляют подгруппы, заданные специальным образом по номерам наблюдений (нулевая переменная) и значениям переменной Home_7. Ниже показано диалоговое окно, задавались параметры этого графика.

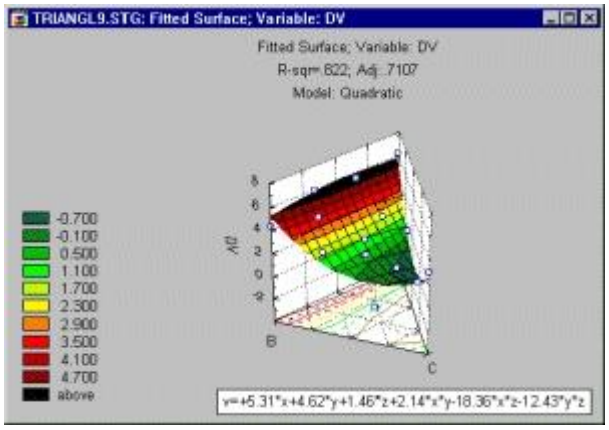


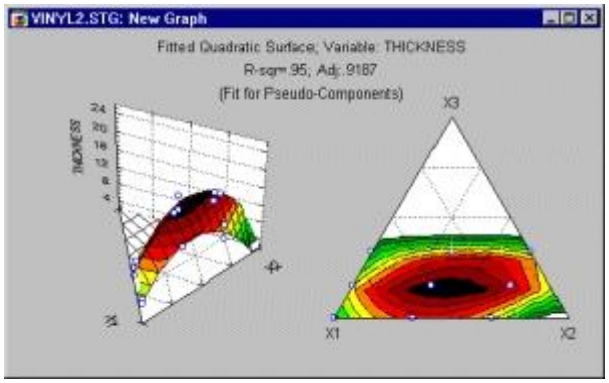
На каждом маленьком графике представлена зависимость меду переменнымиWork_1 и Work_2 (в качестве X и Y соответственно). Первая категоризация (Категории по X - «столбцы» графиков) проводится методом **Сложные подгруппы** в диалоговом окне, вызываемом кнопкой **Задать подгруппы:**



Второй класс (Категории по У или «строки» графиков) определяется группирующей переменной Home_2. Диапазон этой переменной разделен на два равных интервала. Для этого в диалоговом окне задания параметров графика в поле Категории введено значение 2 (при этом распределение переменной Home_2 разделено на две группы: наблюдения, для которых значения меньше либо равны 104,62, и наблюдения со значениями данной переменной, большими этого числа).

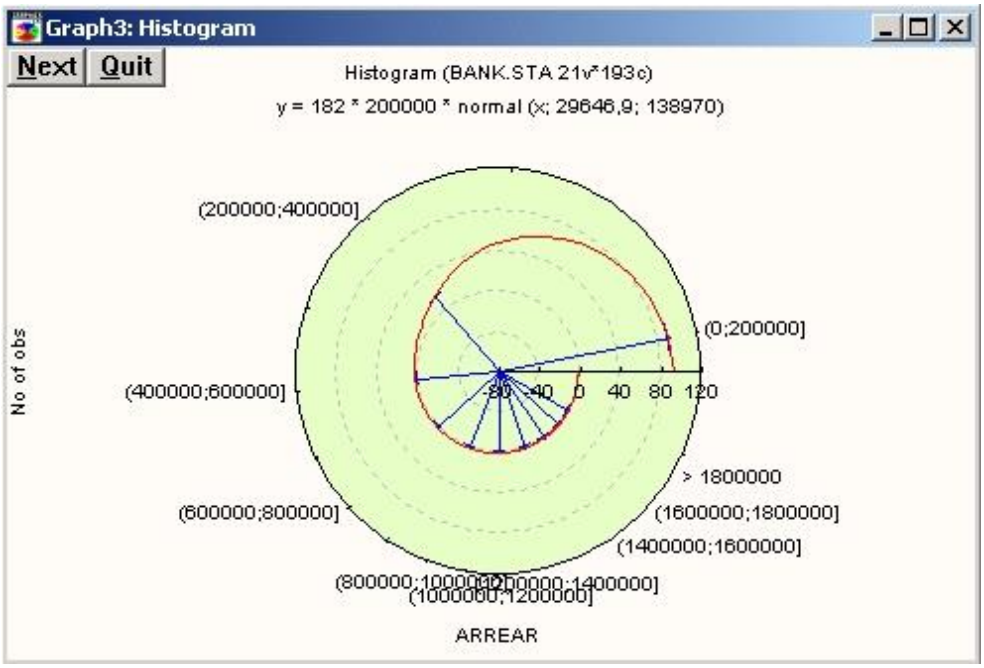
Тернарные графики поверхности и карты линий уровня. При выводе результатов анализа по составлению смесей в модуле Планирование эксперимента можно построить тернарные графики в виде трехмерных поверхностей или карт линий уровня.

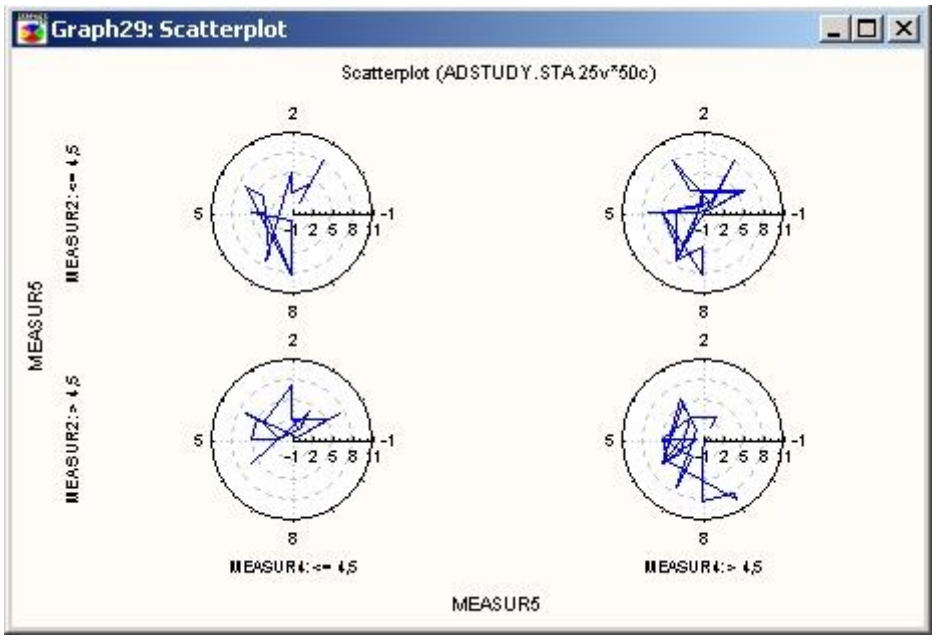




Тернарные графики можно построить из подменю **Статистические XYZ графики, Статистические категоризованные графики и Пользовательские графики** выпадающего меню **Графика.**

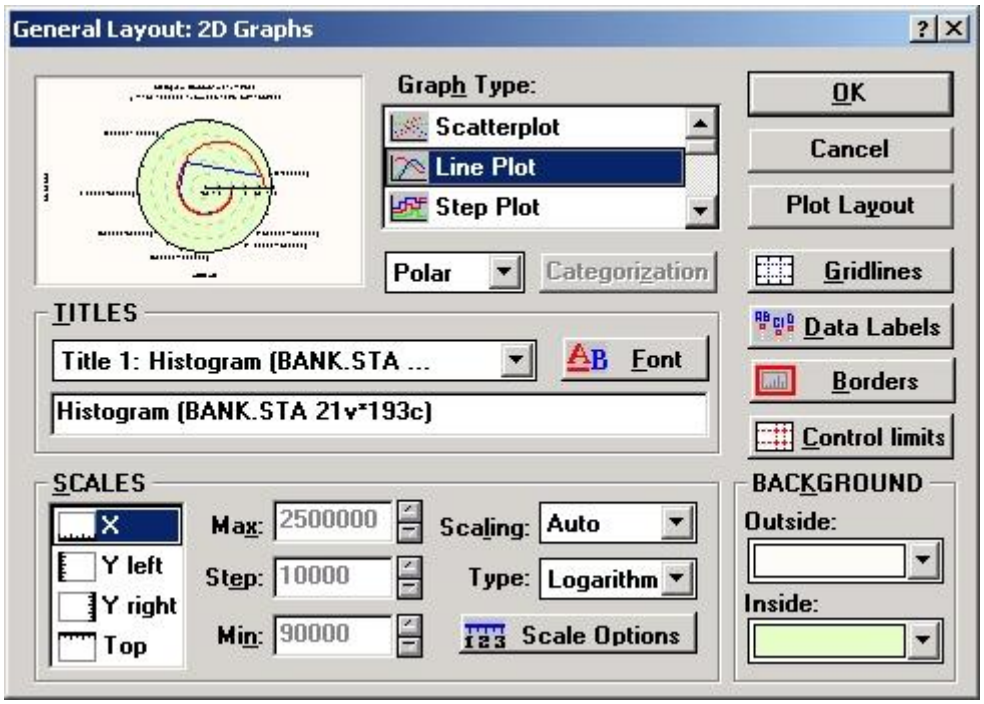
Графики в полярных координатах. Некоторые типы графиков можно построить в полярных координатах. К ним относятся графики рассеяния, линейные графики и последовательные вложенные графики из подменю Статистические 2М графики (оно вызывается из выпадающего меню Графика).





В полярных координатах можно построить и категоризованные графики.

Многие графики, нарисованные в обычной прямоугольной системе координат, можно представить в полярных координатах. Для этого нужно установить соответствующий переключатель в диалоговом окне Общая разметка в положение Полярные.

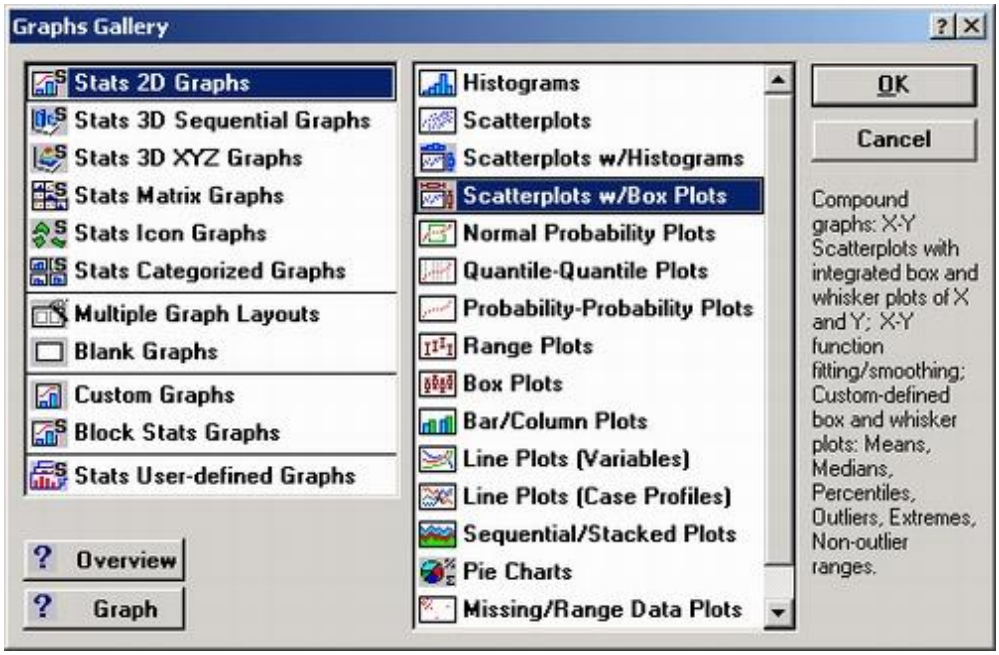


Как поместить на график системы STATISTICA графический объект из другого приложения? Для вставки любых графических объектов, совместимых с системой Windows, можно использовать все описанные выше операции вставки посредством буфера обмена (включая связывание и внедрение методами OLE). Эти операции можно совершать над растровыми объектами, метафайлами Windows, графиками в формате STATISTICA, а также любыми OLf-совместимыми объектами.

Как поместить текст на график STATISTICA (отчеты, таблицы и т. п.)? С помощью описанных выше операций с буфером обмена на графики STATISTICA можно поместить очень большой текстовый объект (например, отчет длиной несколько страниц). Этот текст редактируется и изменяется в окне Редактор текста графика системы STATISTICA или в соответствующем приложении, которое является сервером в методе OLE.

Все описанные в предыдущем разделе операции вставки и использования буфера обмена применимы к любым совместимым с Windows графическим объектам, а операции связывания и внедрения выполняются для всех объектов, поддерживающих методы OLE.

Галерея графиков STATISTICA. С помощью этой кнопки открывается диалоговое окно Галерея графиков STATISTICA. Эта кнопка присутствует в диалоговом окне каждого типа графиков.



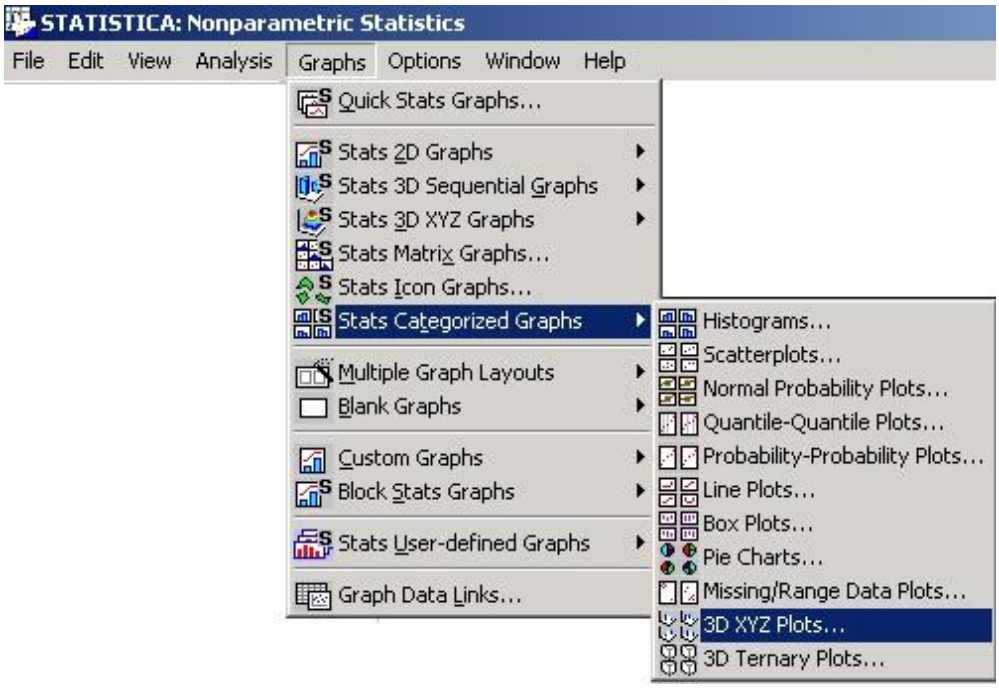
Отсюда быстро и легко вызываются все статистические и пользовательские графики, пустые графические окна и статистические графики пользователя. Для этого нужно выделить название нужного типа графика и дважды щелкнуть на нем (или нажать кнопку **ОК**).

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 (и быстрые статистические) графики.

Главное различие между двумя основными типами графиков заключается в источнике данных для отображения. Более подробно эти различия описаны в следующих разделах.

Пользовательские 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. Содержание этого диалогового окна зависит от выбранного типа пользовательского графика. 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 рамках основного типа). Вид графика также можно подобрать и после построения (с помощью диалоговых окон Общая разметка или Размещение графика, которые открываются при двойном щелчке мышью на области фона графического окна или при выборе соответствующей строки выпадающего меню Разметки).

Статистические графики. В отличие от пользовательских графиков, которые представляют собой средство наглядного отображения числовых данных любых таблиц (исходных данных пли результатов, см. выше), 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**.

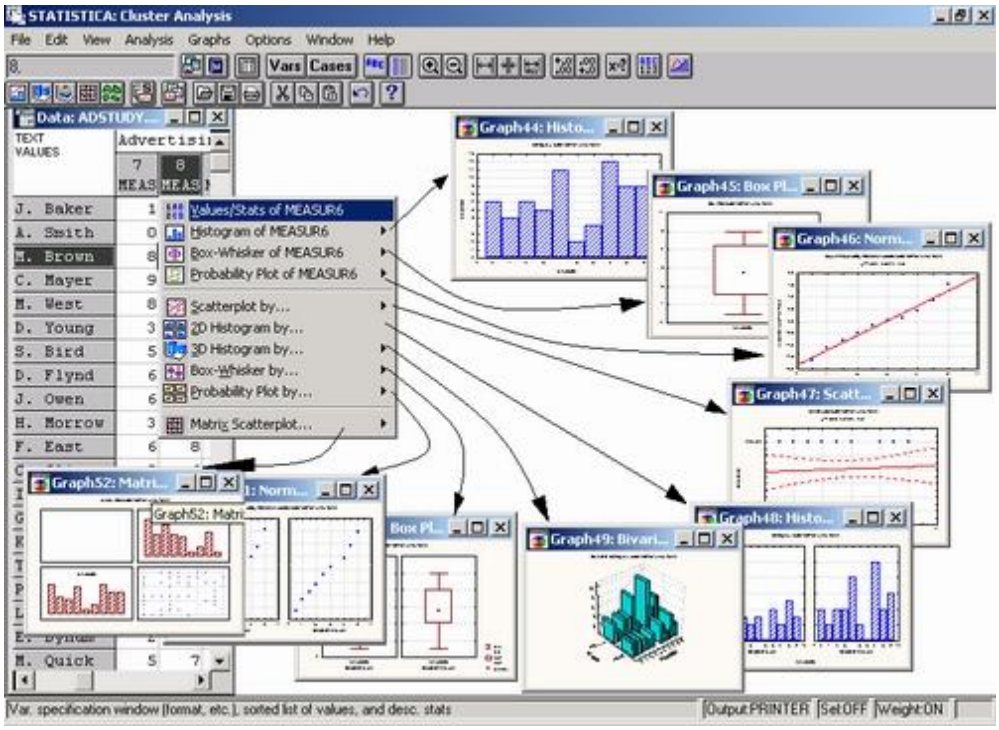


При построении таких графиков используются значения непосредственно из файла данных, которые не зависят от содержания текущей таблицы, выделения блоков и положения курсора. При этом предлагаются либо стандартные методы графического анализа исходных данных (различные графики разброса значений, гистограммы, 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.

Быстрые статистические графики. Наиболее широко используемые типы статистических графиков (вызываемых из меню **Графика**, см. предыдущий параграф) представлены в меню **Быстрые статистические графики**. Эти списки графиков не предоставляют такой широкий спектр возможностей, как меню **Статистические графики**, но в отличие от последних упрощают и ускоряют процедуру построения графика. Быстрые статистические графики:

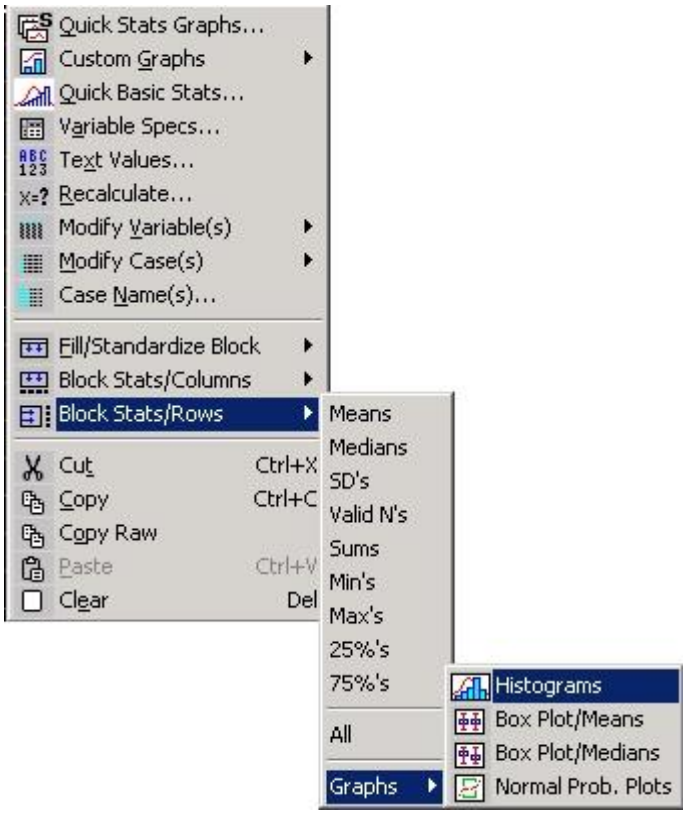
- вызываются из контекстных меню или с панели инструментов любой таблицы (обычно они не требуют обращения к выпадающим меню или диалоговым окнам),
- не требуют от пользователя выбора переменных (этот выбор определяется текущим положением курсора в таблице) и промежуточной настройки параметров (формат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

При выборе пункта Б**ыстрые статистические графики** (с помощью кнопки на панели инструментов из контекстного меню или из выпадающего меню **Графика**) появляется меню выбора статистического графика 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.

Блоковые статистические графики. Эти типы (пользовательских) графиков вызываются из пунктов контекстных меню **Статистики блока по столбцам** и **Статистики блока по строкам** или из диалогового окна **Галерея графиков.**



Любой из этих вариантов дает 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.

Другие специализированные графики. 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**). Как уже упоминалось ранее, специализированные графики, которые связаны не с конкретной таблицей результатов, а с определенным методом анализа данных (например, графики аппроксимирующих функций в модуле **Нелинейное оценивание** или средних в модуле **Дисперсионный анализ)**, вызываются непосредственно из диалогового окна с результатами анализа (то есть из окна, содержащего выходные параметры используемого метода обработки данных).

Настройка графика до и после его построения. Любые изменения параметров графика в STATISTICA осуществляются из активного графического окна (после отображения графика на экране). Как правило, сначала имеет смысл построить график, приняв значения параметров по умолчанию, а затем уже вносить различные изменения. Однако в тех редких случаях, когда построение графика занимает слишком много времени (при создании сложных составных графических изображений или обработке больших наборов данных), можно вмешаться в этот процесс, чтобы сделать необходимые настройки. Прервать рисование можно одним нажатием клавиши или щелчком мыши в любом месте экрана, а затем продолжить его после ввода необходимых изменений.

Предусмотрено два основных метода настройки графика — добавление и редактирование пользовательских графических объектов, изменение структурных элементов графика.

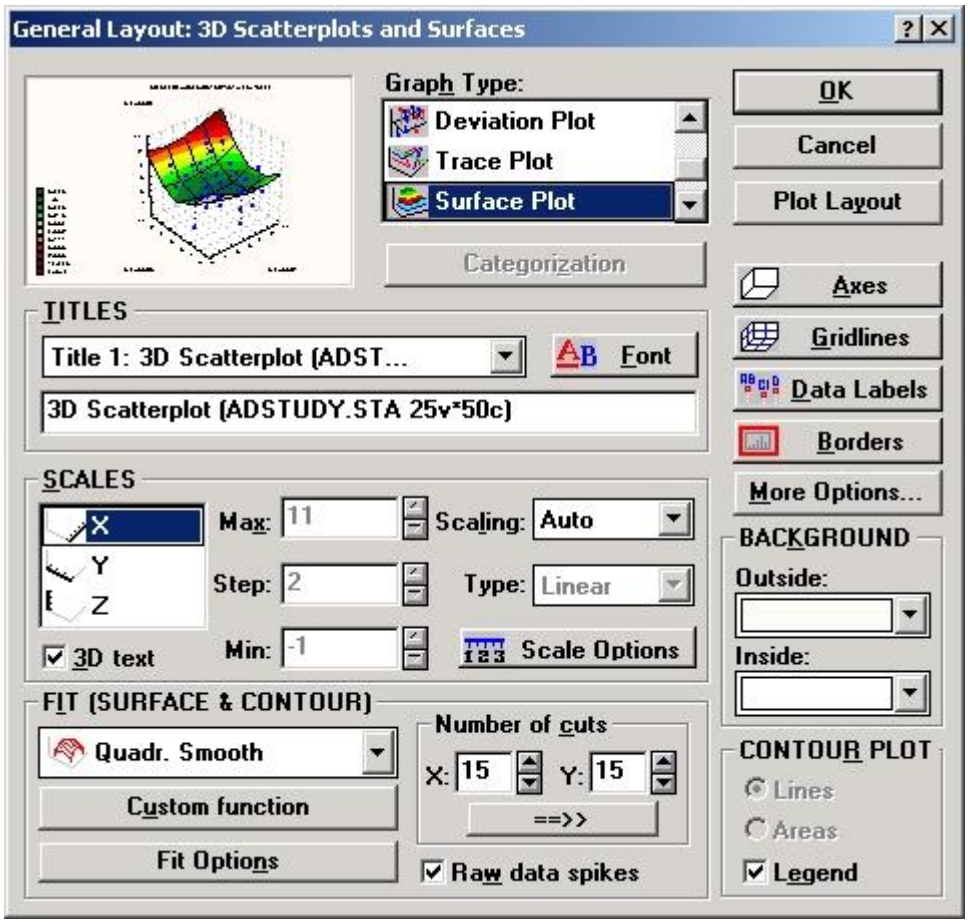
Применяются ли к различным типам графиков различные методы настройки?

Нет. Независимо от способа создания графика для его настройки и изменения можно использовать любые возможности, предусмотренные в системе STATISTICA. К любому графику можно добавить новый график, объединить его с другим графиком, поместить в него связанный или внедренный объект. Кроме того, график можно любым образом изменять, рисовать на нем и использовать различные методы подгонки функций. Эти же методы настройки доступны при работе с графиками, которые были предварительно сохранены и вызваны из дискового файла.

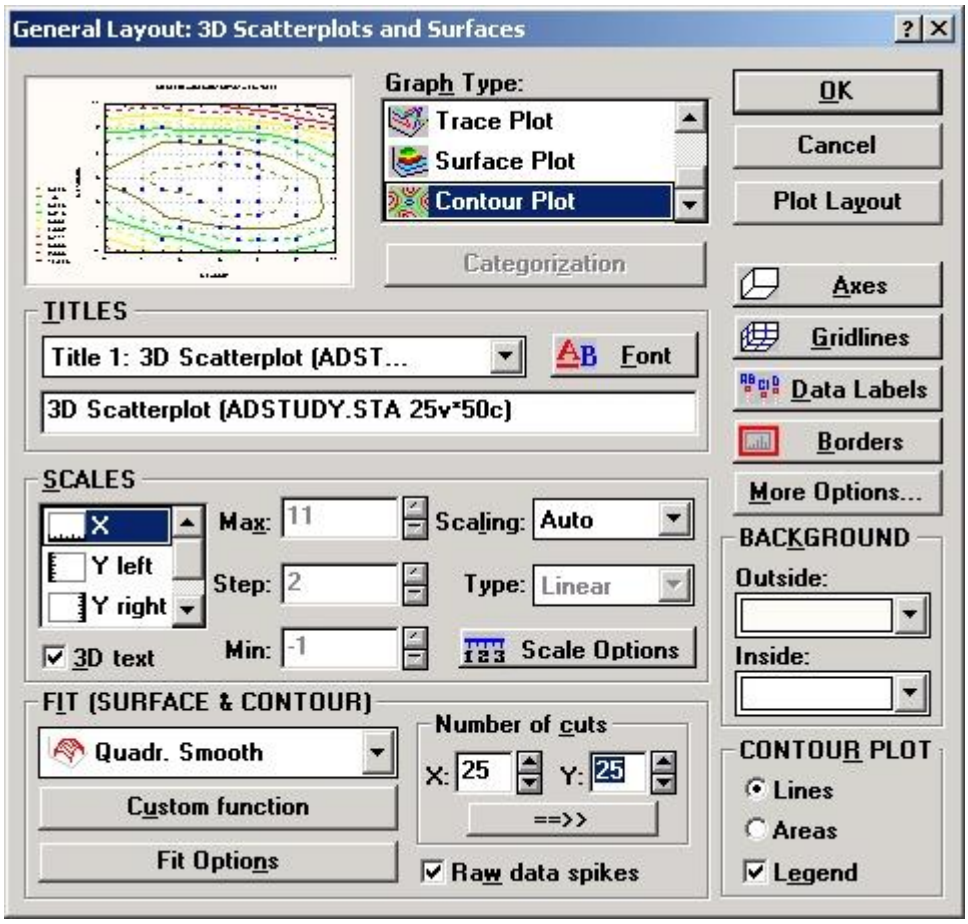
Настройка статистического графика до и после его построения. В разделе Как настроить график STATISTICA показано, что большинство возможностей настройки (сотни различных вариантов графического представления) доступны непосредственно после построения графика. Для этого достаточно щелкнуть на конкретном элементе графика или выбрать соответствующий пункт в диалоговых окнах **Общая разметка** или **Размещение графика**, которые вызываются из выпадающего меню **Разметки.**

В то же время отдельные параметры, которые определяют источник данных, нужно задать до построения графика, например, переменные, метод категоризации, значения меток, имена наблюдений, метки осей. В данном примере перед построением графика нужно выбрать переменные и метол, категоризации, а также при необходимости задать значения некоторых параметров с помощью кнопки **Параметры** (которая здесь не использована).

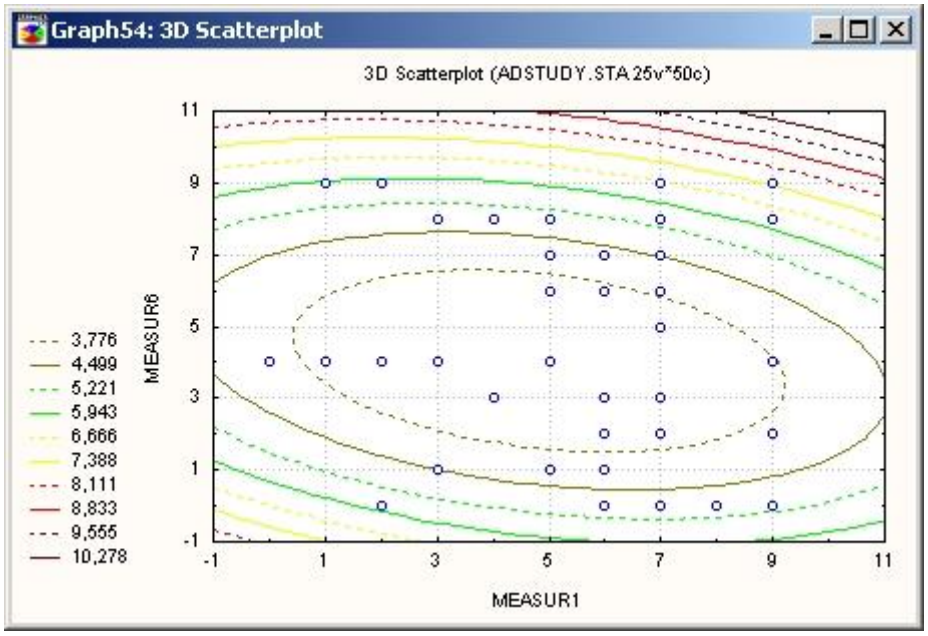
Теперь вернемся к нашему примеру. После построения графика при щелчке на любом месте фона графического окна появится диалоговое окно **Общая разметка**, в котором регулируются параметры общего расположения графика.



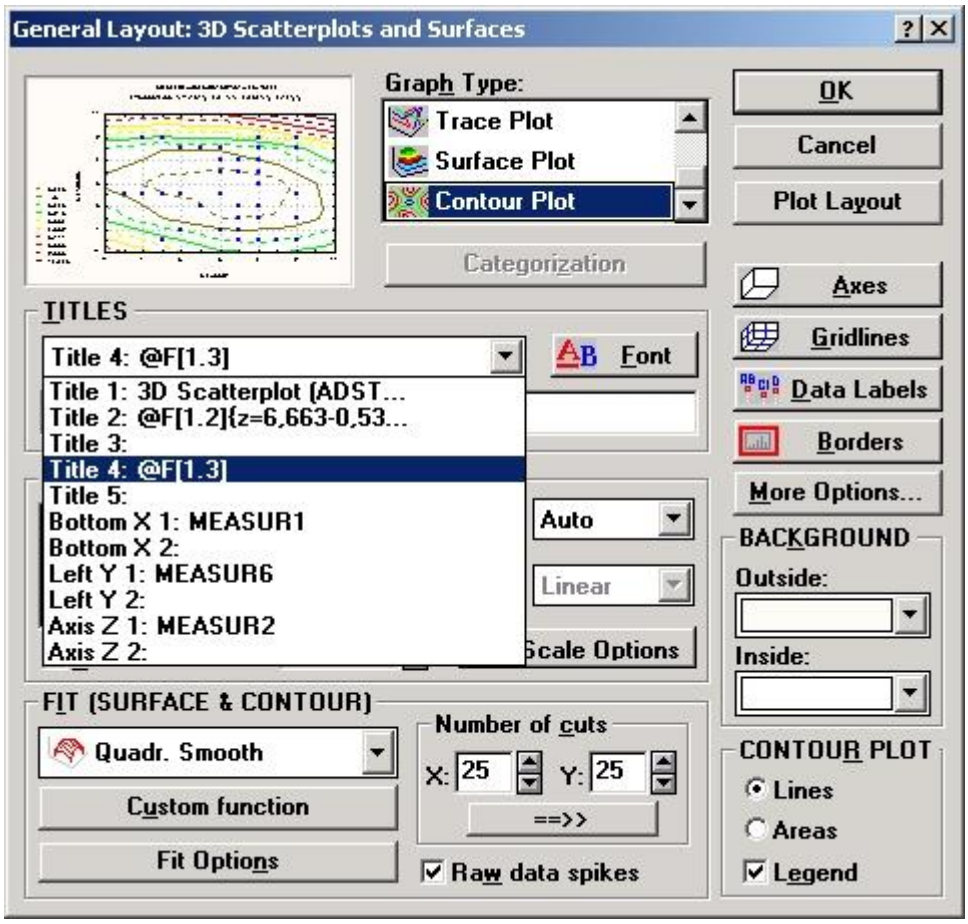
В этом окне можно изменить тип графика и задать построение карты линий у ровня (используйте для этого поле **Тип графика**). Кроме того, можно изменить параметр Число сечений с установленного по умолчанию со значением 15x15 на 25 x 25 (этот параметр определяет точность построения карты линий уровня):



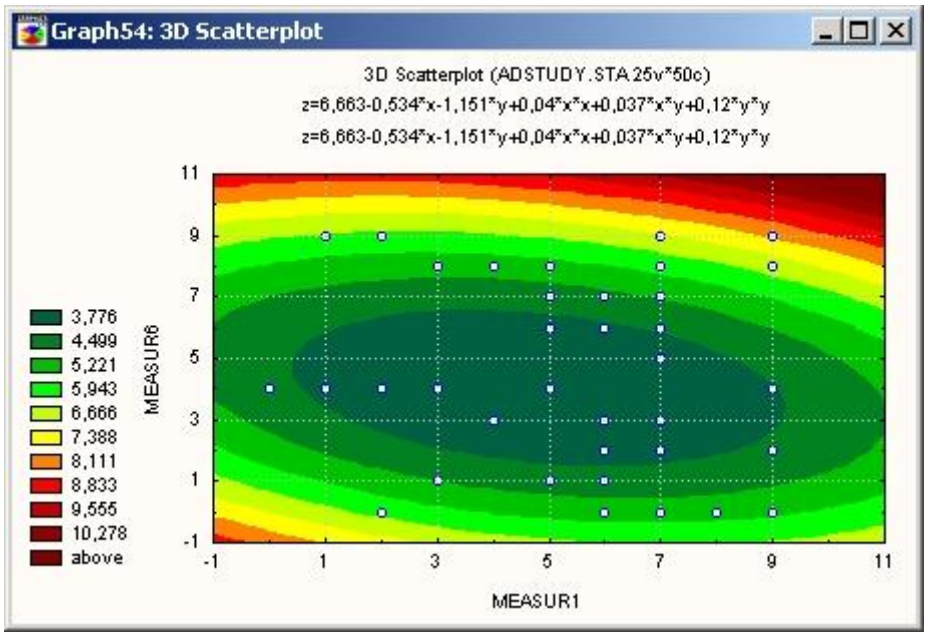
После внесения изменений нажмите **ОК**, и вы увидите новый график:



Снова вернемся к диалоговому окну **Общая разметка** и 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 (цифра 1 на месте первого параметра в квадратных скобках) для каждого из трех отдельных графиков (цифры 1,2иЗъ качестве вторых параметров):

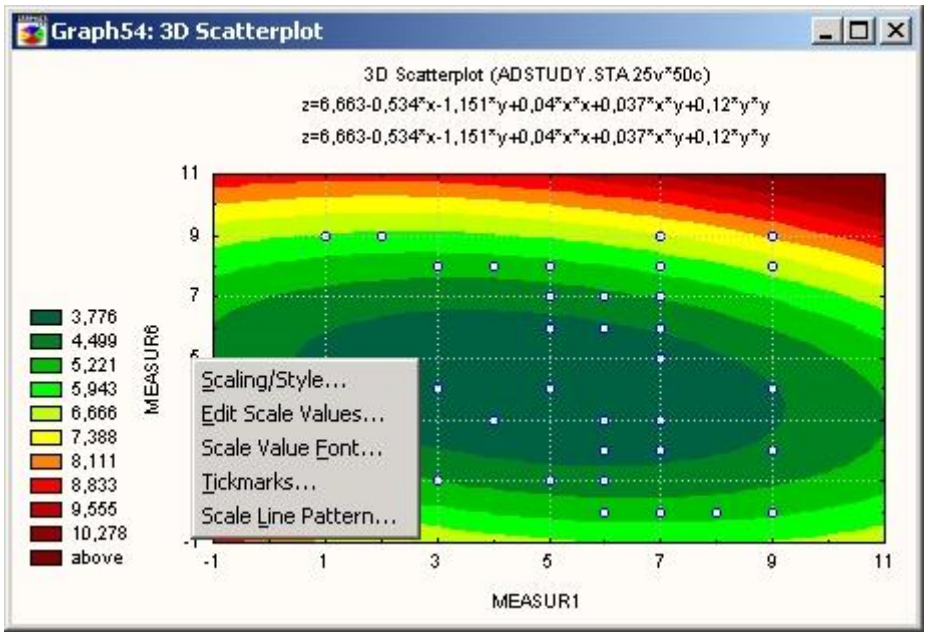


Для быстрейшего отображения и всестороннего форматирования уравнений функций лучше использовать диалоговое окно **Параметры**, которое вызывается из диалогового окна **Статистические графики.** Нажмите **ОК**, и вы увидите измененный график:

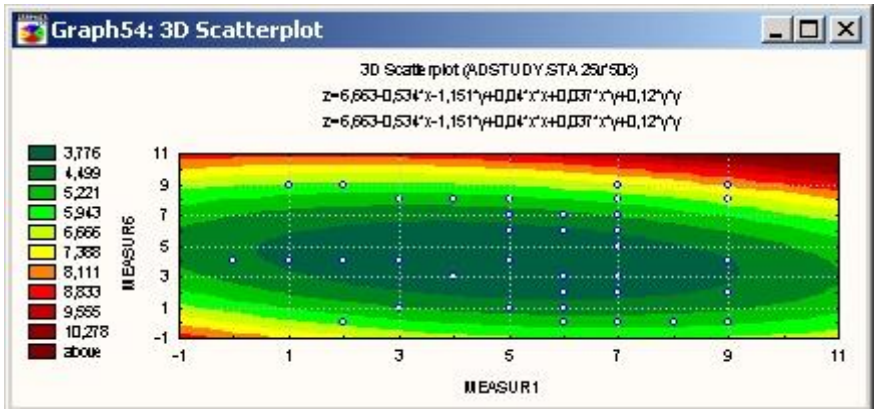


Теперь можно продолжить знакомство с различными способами настройки графика. Самый простой (и самый быстрый) способ изменения параметров какого-либо элемента — это двойной щелчок на нем кнопкой мыши. Кроме того, с помощью одного щелчка правой кнопкой мыши на данном объекте можно вызвать соответствующее ему контекстное меню.

Например, при щелчке правой кнопкой мыши на одной из осей графика появится показанное ниже контекстное меню, в котором предлагается выбор вариантов настройки для данной оси:

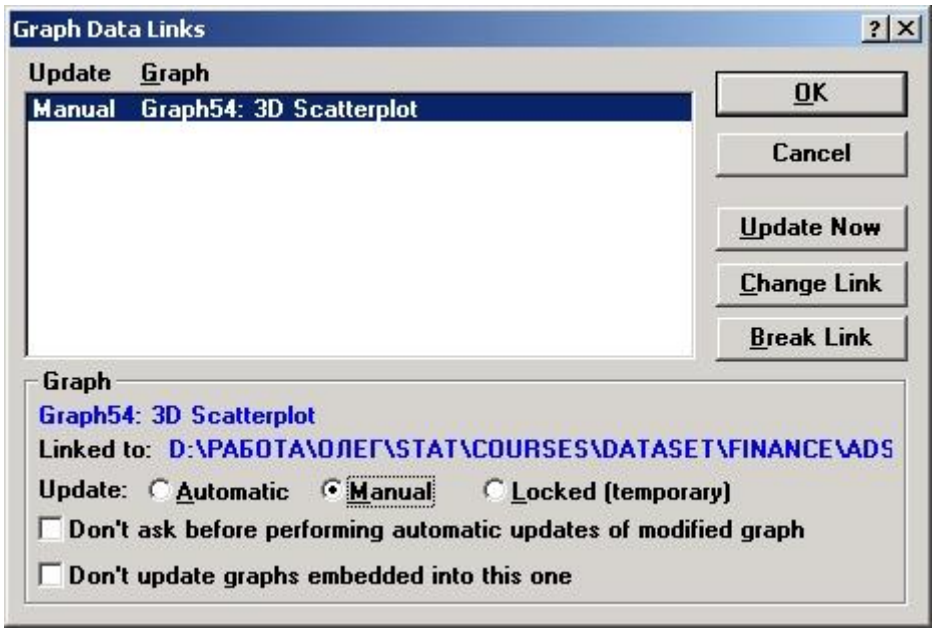


На показанном ниже графике с помощью кнопки панели инструментов подобраны другие пропорции графического окна, кроме того, изменен статус условных обозначений с фиксированного на перемещаемый, а их текст отредактирован, упорядочен и перемещен на другое место.



Могут ли графики автоматически обновляться при изменении файла данных?

Да, могут. Все графики сохраняют связи с таблицей исходных данных, по которым они построены. При этом, если обновление не происходит вручную и связи не отменены, график автоматически обновляется при изменении исходных данных. Для управления связями имеется специальное диалоговое окно **Связи данных и графика.** Оно вызывается из выпадающего меню **Графика**.



Здесь можно установить автоматический режим связи, когда график автоматически обновляется при изменении данных, по которым он построен. Можно также задать режим Вручную или временно заблокировать связь. Кроме того, можно установить режим Связь с текущим файлом данных и построить такой же график или серию графиков для других файлов данных. Способ связи можно глобально изменить с помощью команды выпадающего меню **Сервис**.

STATISTICA поддерживает и «вложенные» связи с другими приложениями. Например, можно установить связь графика с данными электронной таблицы Excel 5 путем динамического обмена данными (DDE). При нажатии клавиши F9 для пересчета таблицы Excel произойдет автоматическое обновление как данных этой таблицы, так и соответствующего им графика в системе STATISTICA. См. также два следующих пункта.

Графический формат STATISTICA. Графики и рисунки могут быть сохранены в графическом формате STATISTICA в файле с расширением *.stg. Для этого используются команды **Сохранить** и **Сохранить как..**. из выпадающего меню **Файл**. Именно этот формат рекомендуется для записи графического файла, если предполагается в дальнейшем снова открывать его в системе STATISTICA или присоединять к другим приложениям методами OLE. В отличие от других графических форматов формат STATISTICA хранит не только саму картинку, но и Редактор данных графика со всеми представленными на графике данными, все аналитические параметры (уравнения подгонки, эллипсы и пр.), а также другие параметры, позволяющие впоследствии продолжить анализ графических данных. Этот формат наиболее удобен при связывании или внедрении графика в другой график STATISTICA. Сохраненные в данном графическом формате файлы можно распечатать в пакетном режиме с помощью команды **Печать файлов** из выпадающего меню **Файл**.

Практическая работа № 5

ИЗУЧЕНИЕ ОСНОВНЫХ ПРИЕМОВ РАБОТЫ В СРЕДЕ ПАКЕТА "STATISTICA NEURAL NETWORKS"

Цель работы – освоение основных приемов работы в среде пакета «Statistica Neural Networks» (далее SNN) и приобретение навыков исследования нейросетей перцептронного типа.

1. ОСНОВНЫЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ И ОПИСАНИЕ ПРАКТИЧЕСКИХ ПРИЕМОВ РАБОТЫ

1.1. ВВЕДЕНИЕ

В этой лабораторной работе на примере классической задачи «исключающего или» будут описаны основные шаги по созданию и применению нейронных сетей в среде пакета *ST Neural Networks.*

Проделав эту лабораторную работу, вы будете знать, как:

- создать файл данных пакета *ST Neural Networks;*
- построить нейронную сеть в пакете *ST Neural Networks;*
- обучить нейронную сеть на множестве данных;
- запустить обученную нейронную сеть на исполнение;
- сохранить или открыть ранее сохраненный набор данных или экземпляр сети.

1.2. ЗАДАЧА «ИСКЛЮЧАЮЩЕГО ИЛИ»

Задача «исключающего ИЛИ» (XOR) была очень популярна среди пионеров в области нейронных сетей. Она формулируется очень просто: имея две бинарные входные переменные, каждая из которых может принимать значение «ноль» или «единица», распознать случаи, когда одна из этих переменных равна единице, а другая нулю.

Всего здесь возможны четыре типа наблюдений (таблица 1.1):

Таблица 1.1 – Набор данных функции «Исключающее или»

Вход 1	Вход 2	XOR
0	0	Нет
1	0	Да
0	1	Да
1	1	Нет

Из таблицы 1 видно, чем интересна задача «исключающего ИЛИ». На рисунке 1.1 соответствующие точки изображены на плоскости, где ось *X* соответствует переменной *Вход 1,* а ось Y - переменной *Вход 2.* Случаи «Да» обозначены кружками, а случаи «Нет» - крестиками.

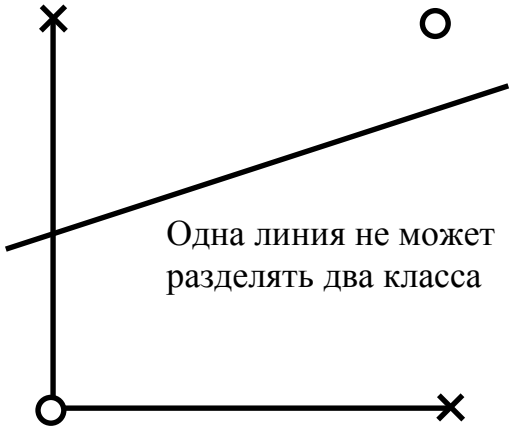


Рисунок 1.1 – Пример изображения «исключающего ИЛИ»

на плоскости

Хотя, задача кажется совсем простой, у нее есть одна особенность, представляющая трудность для многих методов решения задач анализа данных: она линейно неотделима. Иначе говоря, невозможно провести прямую линию на плоскости так, чтобы положительные случаи оказались по одну сторону от нее, а отрицательные - по другую.

Это значит, что линейные методы, широко применявшиеся на протяжении многих лет, неприменимы для решения такой задачи. В то же время, задача «исключающего ИЛИ» - это действительно фундаментальная задача: она является простейшим примером из целого класса наиболее часто встречающихся задач.

Нейроные сети способны решать линейно неотделимые задачи классификации, простейшим примером которых служит задача «исключающего ИЛИ».

1.3. СОЗДАНИЕ НАБОРА ДАННЫХ

Прежде чем двигаться дальше, давайте запустим *ST Neural Networks.* После того, как программа была инсталлирована, это делается обычным образом - путем выбора соответствующего пункта в списке программ или же двойным щелчком по пиктограмме *STATISTICA Neural Networks* на рабочем столе Windows.

Одно из главных свойств нейронных сетей состоит в том, что они способны учиться решать задачи на примерах. Вместо того, чтобы непосредственно задавать значения весов сети (хотя в системе *ST Neural Networks* это возможно), мы подаем на вход сети набор обучающих примеров, а затем с помощью того или иного алгоритма обучения - например, методом обратного распространения - веса сети корректируются таким образом, чтобы она училась понимать обучающие данные.

В пакете *ST Neural Networks,* обучающие данные хранятся в виде набора данных *(Data Set),* содержащего некоторое количество наблюдений, для каждого из которых заданы значения нескольких входных и выходных переменных. Как правило, данные берутся из какого-то внешнего источника (например, системы *STATISTICA* или электронной таблицы). Однако новый набор данных можно создать прямо в пакете *ST Neural Networks.* Для этого нужно проделать следующие действия.

1. Войти в диалоговое окно *Создать набор данных - Create Data Set* с помощью команды *Набор данных - Data Set...* из меню *Файл - Создать - File-New.*
2. Ввести значения числа *входных (Inputs)* и *выходных (Outputs)* переменных в будущем наборе данных. В задаче «исключающего ИЛИ» имеется две входные переменные и одна выходная.
3. Нажать кнопку *Создать - Create.*

При создании нового набора данных программа *ST Neural Networks* автоматически открывает окно *Редактор данных - Data Set Editor* (рисунок 1.2).

Основной элемент окна *Редактор данных - Data Set Editor* - это таблица, содержащая все записи (наблюдения - Cases) набора данных. Каждому наблюдению соответствует одна строка таблицы. В начальный момент таблица будет содержать всего одну строку, а значения всех переменных будут «неизвестны» (обозначены знаком вопроса). У входных переменных заголовок столбца черного цвета, у выходных - голубого; входы от выходов отделяются темной вертикальной линией. Добавление новых наблюдений и редактирование данных в уже имеющихся наблюдениях осуществляется обычным редактированием этой таблицы.

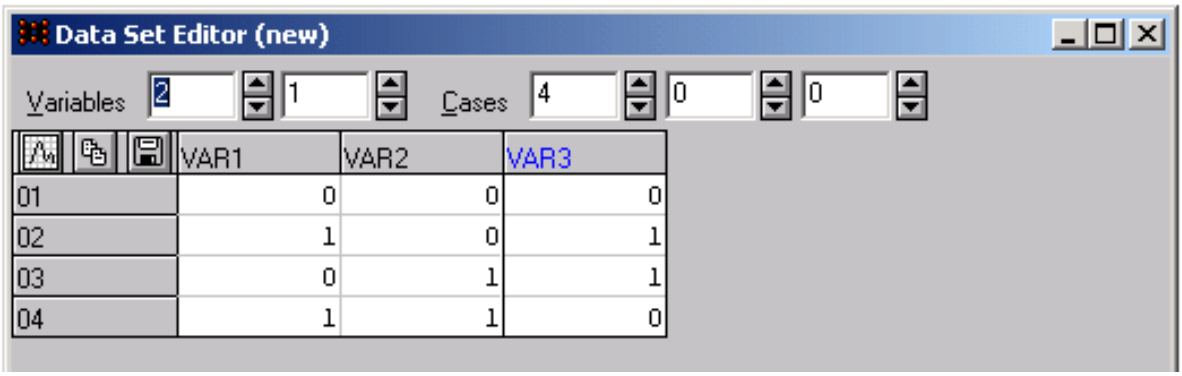


Рисунок 1.2 - Элемент окна Редактор данных - Data Set Editor

Данные, которые потребуются нам в задаче «исключающего ИЛИ», приведены на рисунке. Положительный результат классификации обозначается цифрой 1 в столбце выходной переменной, отрицательный результат - цифрой 0. Щелкните мышью в поле первой ячейки таблицы, введите «0» и нажмите ВВОД. Рамка выделения переместится на следующую ячейку. Введите нули в две оставшиеся клетки, затем действуйте в соответствии с приведенными далее инструкциями.

Добавление наблюдений

1. Выберите ячейку таблицы, щелкнув по ней мышью.
2. Нажмите клавишу СТРЕЛКА вниз. Всякий раз при попытке выйти вниз за границы таблицы, программа *ST Neural Networks* создает новое наблюдение.
3. Введите значения для второго наблюдения (1,0,1). Заполнив все клетки строки, нажмите клавишу ввод.
4. Нажатие клавиши ввод в последней ячейке рассматривается системой как попытка продвинуться за нижнюю границу таблицы, поэтому программа создаст строку третьего наблюдения.
5. Заполните данными оставшиеся наблюдения. Не нажимайте ввод в последней ячейке последнего наблюдения, потому что пятое наблюдение нам не нужно.

Удаление лишних наблюдений

Если вы случайно создали лишнее наблюдение, его можно удалить следующим образом:

1. Щелкните в средней части метки строки, соответствующей лишнему наблюдению, метки расположены в левой части таблицы. Вся строка станет выделенной.
2. Нажмите клавиши CTRL+X. Наблюдение будет удалено.

Изменение имен переменных и наблюдений

В пакете *ST Neural Networks* имеется возможность присваивать имена отдельным наблюдениям и переменным. Чтобы присвоить наблюдению имя, сделайте следующее (рисунок 1.3):

1. Дважды щелкните в средней части метки строки этого наблюдения; метки строк расположены в левой части таблицы. Появится текстовый курсор (серая вертикальная полоса).
2. Введите имя. В качестве метки строки по умолчанию берется ее номер. Не обращайте на него внимания - он отображается только в строках, которым не присвоено имя, и исчезнет сразу же, как только вы начнете вводить символы.
3. С помощью клавиш СТРЕЛКА ВЛЕВО и СТРЕЛКА ВПРАВО курсор можно передвигать по буквам имени, клавишами DELETE и BACKSPACE - удалять лишние символы, с помощью клавиш СТРЕЛКА ВВЕРХ И СТРЕЛКА ВНИЗ можно перейти к именам других наблюдений, клавиша ESCAPE прерывает редактирование.

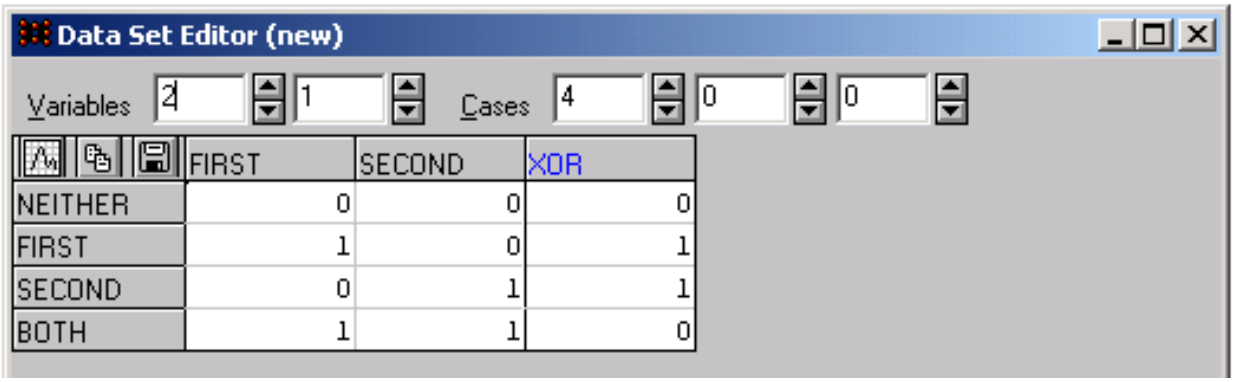


Рисунок 1.3 - Присваивание имен отдельным наблюдениям и переменным

Аналогично присваиваются имена переменным - для этого нужно отредактировать метки столбцов (вместо строк). Измените имена всех наблюдений и переменных, чтобы ваша таблица выглядела, как на нашем рисунке.

Другие возможности редактирования

Таблицы пакета *ST Neural Networks* предлагают большой набор средств, облегчающих создание наборов данных и последующую работу с ними. Приведем их краткое описание.

- *Перемещение активной ячейки.* Осуществляется клавишами: СТРЕЛКА ВЛЕВО, СТРЕЛКА ВПРАВО, СТРЕЛКА ВВЕРХ, СТРЕЛКА ВНИЗ, HOME, END, PAGE UP, PAGE DOWN.

• *Выделение диапазона ячеек*. Производится перетаскиванием указателя мыши или клавишами курсора при нажатой клавише SHIFT.

• *Копирование и вставка*. Чтобы скопировать выделенный диапазон ячеек в буфер обмена, нажмите CTRL+C, чтобы вставить содержимое буфера обмена в таблицу - нажмите CTRL+V. Можно копировать и вставлять целые строки и столбцы целиком. Возможен также обмен данными между пакетом *ST Neural Networks* и другими приложениями. Чтобы скопировать всю таблицу в буфер обмена, файл или систему STATISTICA, используйте кнопки пиктограмм в левой верхней части таблицы (кроме того, экспорт данных в файл и систему STAT1STICA можно осуществить соответственно с помощью клавиш CTRL+SHIFT+S или CTRL+Q).

• *Вставка*. В любом месте таблицы можно вставить новую строку или столбец. Поместите курсор мыши на линию, разделяющую метки двух соседних строк или столбцов (при этом курсор превратится в двустороннюю стрелку), и щелкните кнопкой - откроется полоса вставки. После нажатия клавиши INSERT будет вставлена новая строка/столбец.

• Чтобы назначить тип переменной - *Входная - Input, Выходная -Output, Входная/Выходная - Input/Output* или *Неучитываемая - Ignored*, выберите переменную, щелкнув на метке соответствующего столбца, затем нажмите правую кнопку мыши и выберите нужный тип из контекстного меню.

• Чтобы задать номинальную переменную (например, *Пол ={Муж, Жен}*), выберите переменную, щелкнув на метке соответствующего столбца, затем нажмите правую кнопку мыши и выберите команду *Определение - Definition...* из контекстного меню.

• Чтобы задать тип подмножества, *Обучающее - Training, Контрольное - Verification, Тестовое - Test* или *Неучитываемое - Ignored,* выбирайте наблюдения, щелкая на метках их строк, нажимайте правую кнопку мыши и выбирайте нужный тип из контекстного меню.

• Все перечисленные возможности доступны также через команды *Наблюдения - Cases...* и *Переменные - Variables...* меню *Правка - Edit.*

1.4. СОЗДАНИЕ НОВОЙ СЕТИ

Создать новую сеть в пакете *ST Neural Networks* можно либо средствами диалогового окна *Создать сеть - Create Network,* доступ к которому осуществляется через команду *Сеть... - Network...* меню *Файл-Создать -File-New*. Кроме того, можно создать сеть, пользуясь автоматическим конструктором сети (кнопка ). Диалоговое окно *Создать сеть -Create Network* показано на рисунке 1.4.

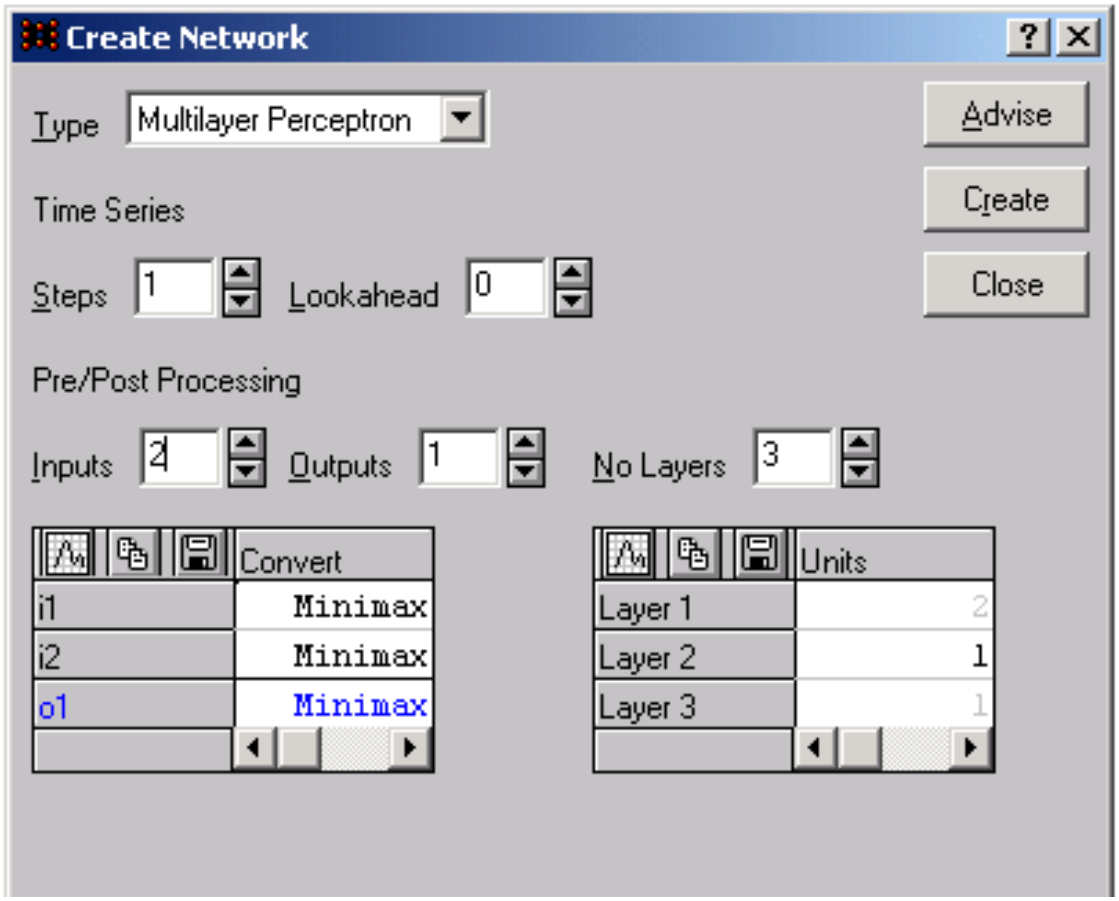


Рисунок 1.4 - Диалоговое окно Создать сеть -Create Network.

Первый взгляд на это окно может озадачить пользователя. Дело в том, что в пакете *ST Neural Networks* для конструирования сетей реализованы довольно сложные возможности, в том числе и мощные инструменты пре- и пост-процессирования, которое необходимо для преобразования информации в числовую форму (для использования в сети) и обратно. Для тех, кто не хочет вникать во все эти тонкости, в пакете *ST Neural Networks* имеется функция *Совет - Advise*, позволяющая автоматически сконфигурировать большинство характеристик сети по набору исходных данных.

Создание сети

1. Выберите тип сети из выпадающего списка *Тип - Type*. Сейчас нам нужен тип *Многослойный персептрон - Multilayer Perceptron*, который всегда предлагается по умолчанию.
2. Нажмите кнопку *Совет - Advise*. Программа *ST Neural Networks* установит параметры по умолчанию для пре/пост-процессирования и конфигурации сети, исходя из типа переменных, составляющих исходные данные.
3. Введите необходимые исправления в соответствия переменных и спецификации слоев сети (см. ниже).
4. Нажмите кнопку *Создать - Create*, и в результате будет создана новая сеть.

Задание режима пре/пост-процессирования

и параметров сети

Диалоговое окно *Создать сеть - Create Network* содержит две таблицы (рисунок 1.4): левая предназначена для пре/пост-процессирования переменных, а правая - собственно для задания параметров сети. Нажав кнопку *Совет - Advise,* вы сможете быть уверены, что пре/пост-процессирование *(Pre/Post Processing)* переменных будет произведено в соответствии с типом данных (в данном случае должно быть две входных и одна выходная переменная) и что число слоев в сети и элементов в каждом слое выбрано разумным образом. Обычно от пользователя требуется выполнить несколько действий.

1. Изменить, если потребуется, преобразующую функцию для пре/пост-процессирования. В данном случае вполне подойдет функция *Мини-макс - Minimax.*
2. Задать число слоев и скрытых элементов в сети. В пакете *ST NeuralNetworks* на экран также выдается число элементов во входном и выходном слоях. Однако два последних параметра полностью определяются числом входных и выходных переменных, и их нельзя менять (они отображаются серым цветом).

Для задачи «исключающего ИЛИ» нужна сеть с тремя слоями: входным слоем из двух элементов, промежуточным слоем из двух элементов и выходным слоем из одного элемента. Щелкните по ячейке, в которой указано число скрытых элементов (Layer 2), и задайте его равным двум.

Замечание. В этом диалоговом окне можно задать и некоторые другие параметры, в том числе: параметры временного ряда *(Time Series) Временное окно-Steps* и *Горизонт - Lookahead,* параметры преобразования и подстановки пропущенных значений при пре/пост-процессировании, ширину слоев сети.

Если вы точно следовали всем инструкциям, то у вас получится сеть, показанная на рисунке 1.5. Если же вы где-то ошиблись, то повторите все снова.

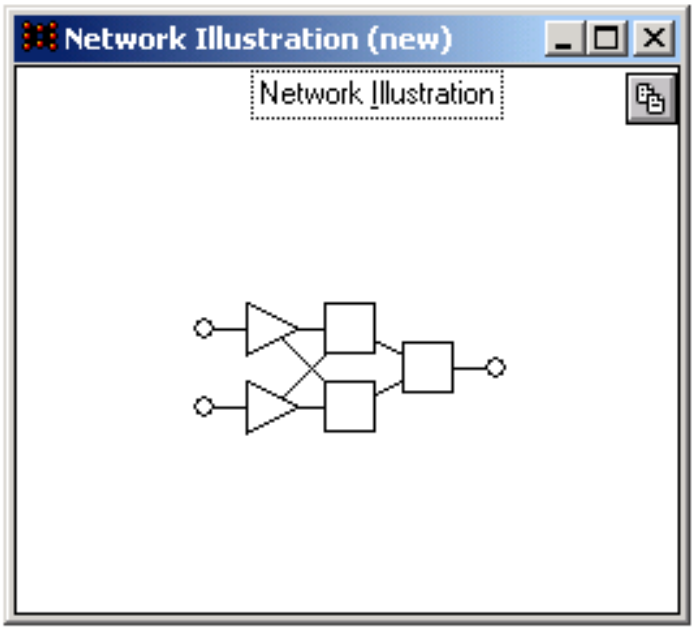


Рисунок 1.5 – Диалоговое окно, получившееся после создания сети

Сохранение набора данных и сети

После того, как мы потратили время и усилия на создание сети и набора данных, неплохо бы сохранить результаты нашей работы для дальнейшего использования. В *ST Neural Networks* сеть и набор данных сохраняются в разных файлах - для этого используются диалоговые окна *Сохранить сеть - Save Network* и *Сохранить набор данных - Save DataSet.*

Сохраним набор данных.

1. Откройте диалоговое окно *Сохранить набор данных - Save DataSet* с помощью команды *Набор данных - DataSet...* из меню *Файл-Сохранить как - File-Save as.*
2. Введите имя файла данных *Xor.sta* в поле *Имя файла - File Name.*
3. Нажмите кнопку *Сохранить - Save.*

Сеть сохраняется аналогичным образом с помощью окна *Сохранить сеть - Save Network;* в качестве стандартного расширения имени файла сети используются **.net* или **.bnt.*

Во время сеанса работы имеет смысл периодически сохранять сеть и набор данных; функции *Сеть... - Network...* и *Набор данных... - DataSet...* меню *Файл-Сохранить - File-Save* сохраняют текущее состояние сети и файла данных, не требуя повторного ввода имени файла. Данные можно сохранить также с помощью кнопки  .

1.5. ОБУЧЕНИЕ СЕТИ

Следующий шаг после задания набора данных и построения подходящей сети - это обучение.

В пакете *ST Neural Networks* реализованы основные алгоритмы обучения многослойных персептронов: методы обратного распространения, сопряженных градиентов и Левенберга-Маркара.

Суть метода обратного распространения

1. Алгоритм обратного распространения последовательно обучает сеть на данных из обучающего множества. На каждой итерации (они называются эпохами) все наблюдения из обучающего множества (в данном случае оно совпадает со всем набором данных) по очереди подаются на вход сети. Сеть обрабатывает их и выдает выходные значения.

2. Эти выходные значения сравниваются с целевыми выходными значениями, которые также содержатся в наборе исходных данных, и ошибка, то есть разность между желаемым и реальным выходом, используется для корректировки весов сети так, чтобы уменьшить эту ошибку.

3. Алгоритм должен находить компромисс между различными наблюдениями и менять веса таким образом, чтобы уменьшить суммарную ошибку на всем обучающем множестве; поскольку алгоритм обрабатывает наблюдения по одному, общая ошибка на отдельных шагах не обязательно будет убывать.

В пакете *ST Neural Networks* отслеживается общая ошибка сети - на графике, а также ее ошибки на отдельных наблюдениях - на гистограмме. Мы рекомендуем следить за ходом обучения сети, как минимум, по графику общей ошибки.

Чтобы обучить сеть в задаче «исключающего ИЛИ», следя при этом за работой алгоритма, действуйте так, как описано далее.

Обучение методом обратного распространения

1. Откройте окно *График ошибки обучения - TrainingError Graph* с помощью команды *График обучения... - Training Graph...* меню *Статистики - Statistics* или кнопки  (рисунок 1.6).

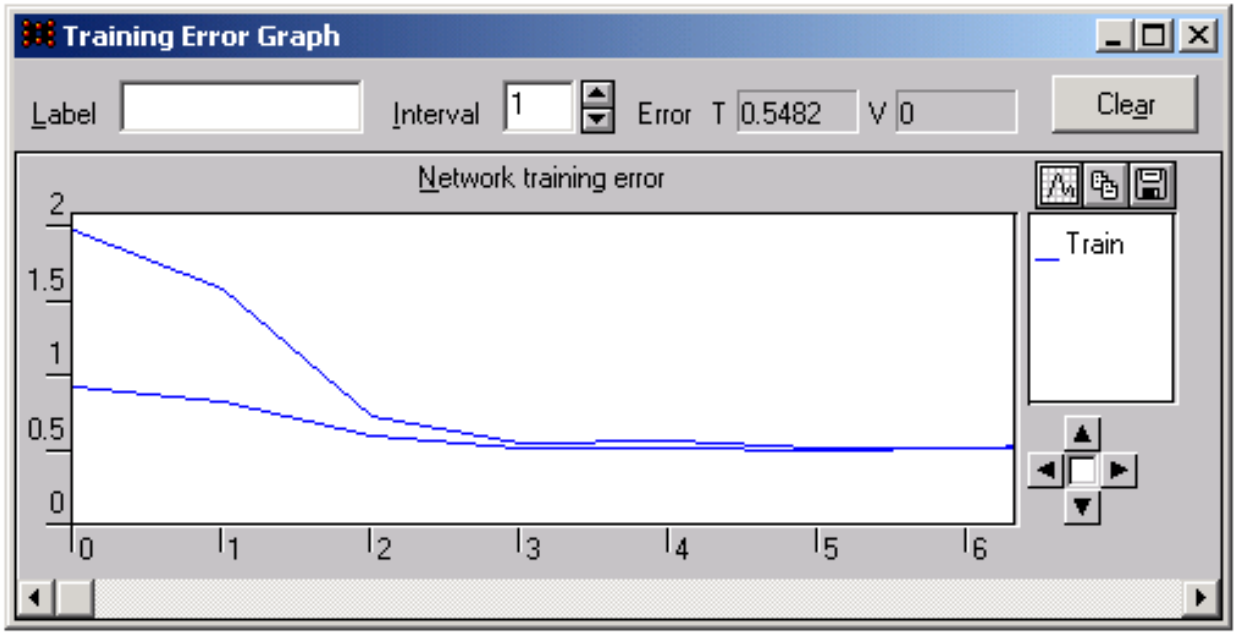


Рисунок 1.6 – Диалоговое окно График ошибки обучения –

Training Error Graph

2. Откройте диалоговое окно *Обратное распространение – Back Propagation* (справа) с помощью команды *Обратное распространение - Back Propagation...* меню *Обучение многослойного персептрона - Train-Multilayer Perceptrons* или кнопки  (рисунок 1.7).

Подвиньте окна так, чтобы они не пересекались и были удобно расположены.

3. Нажмите кнопку *Обучить - Train* в диалоговом окне *Обратное распространение - Back Propagation* - будет запущен алгоритм обучения. При этом на график будет выводиться ошибка (рисунок 1.6).

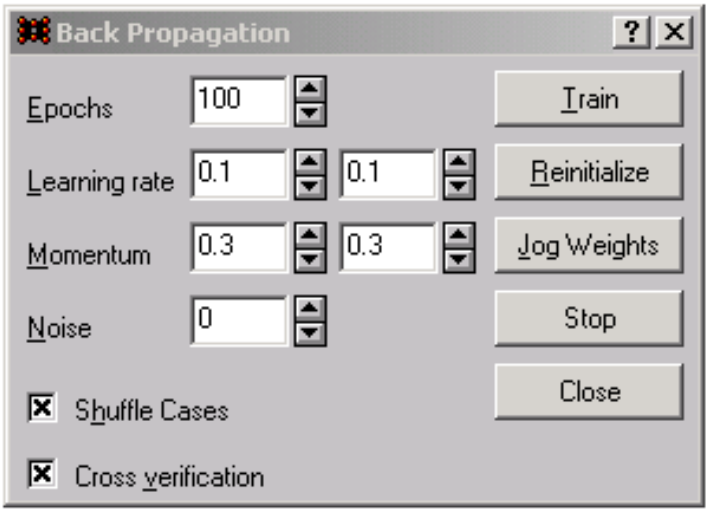


Рисунок 1.7 – Диалоговое окно Обратное распространение –

Back Propagation

4.Повторно нажимайте кнопку *Обучить - Train,* чтобы алгоритм переходил к очередным эпохам.

Не удивляйтесь, если поначалу ошибка не будет существенно уменьшаться. Хотя задача «исключающего ИЛИ» выглядит совсем простой, многослойному персептрону решить ее намного сложнее, чем многие реальные задачи, которые кажутся очень сложными. При тех параметрах по умолчанию, которые предложила нам программа *ST Neural Networks,* может потребоваться порядка тысячи итераций, прежде чем ошибка станет близка к нулю.

Оптимизация обучения

Режим работы алгоритма обратного распространения зависит от ряда параметров, и большинство из них собрано в диалоговом окне *Обратное распространение - Back Propagation* (рисунок 1.7). Для большинства практических задач хорошим начальным приближением будут значения, принимаемые по умолчанию, но при необходимости их можно изменить. В случае же с задачей «исключающего ИЛИ» эти настройки оказываются весьма неудачными.

Опишем кратко наиболее важные параметры и выберем их значения для задачи «исключающего или».

- *Эпохи - Epochs.* Задает число эпох обучения, которые проходятся при одном нажатии клавиши *Обучить - Train.* Значение по умолчанию 100 вполне приемлемо.
- *Скорость обучения - Learning rate.* При увеличении скорости обучения алгоритм работает быстрее, но в некоторых задачах это может привести к неустойчивости (особенно если данные зашумлены). Для задачи «исключающего или» подходит относительно высокая скорость обучения, например, 0,9.
- *Инерция - Momentum.* Этот параметр улучшает (ускоряет) обучение в ситуациях, когда ошибка мало меняется, а также придает алгоритму дополнительную устойчивость. Значение этого параметра всегда должно лежать в интервале [0;1) (т.е. быть больше или равно нулю и меньше единицы). Часто рекомендуется использовать высокую скорость обучения в сочетании с небольшим коэффициентом инерции и наоборот.
- *Перемешивать наблюдения - Shuffle Cases.* При использовании этой функции порядок, в котором наблюдения подаются на вход сети, меняется в каждой новой эпохе. Это добавляет в обучение некоторый шум, так что ошибка может испытывать небольшие колебания. Однако при этом меньше вероятность того, что алгоритм «застрянет», и общие показатели его работы обычно улучшаются.

При таких настройках параметров пакет *ST Neural Networks* решает задачу «исключающего ИЛИ» примерно за двести итераций.

Замечание. В пакете *ST Neural Networks* имеется возможность менять скорость обучения и/или коэффициент инерции от эпохи к эпохе, постепенно сдвигая их от начальных значений, заданных в полях в левой части диалогового окна *Обратное распространение - Back Propagation,* к их конечным значениям, заданным в правой части окна. Например, можно уменьшать скорость по ходу обучения. При задании начальных значений конечные значения по умолчанию устанавливаются такими же.

Выполнение повторных прогонов

Если вы хотите сравнить результаты работы алгоритма в разных вариантах, воспользуйтесь кнопкой *Переустановить - Reinitialize* диалогового окна *Обратное распространение - Back Propagation.* В результате веса сети вновь будут установлены случайным образом для начала следующего сеанса обучения. Если теперь после кнопки *Переустановить — Reinitialize* нажать кнопку *Обучить - Train,* на графике начнет рисоваться новая линия.

Если в результате таких действий график станет слишком «замусоренным», его можно очистить с помощью кнопки *Очистить - Clear* окна *График обучения - Training Graph.*

Совет. Чтобы сделать сравнение более наглядным, можно рисовать линии разными цветами. Введите значение в поле *Метка - Label* в окне

График обучения - Training Graph, и тогда следующая линия будет нарисована другим цветом, а указанная метка будет выведена справа от графика как условное обозначение.

Ошибки для отдельных наблюдений

В окне *График обучения - Training Graph* выводится суммарная ошибка сети. Но иногда бывает полезно проследить за тем, как алгоритм обучения воспринимает отдельные наблюдения.

В пакете *ST Neural Networks* это делается в окне *Ошибки наблюдений... -Case Errors,* которое открывается командой *Ошибки наблюдений... - Case Errors...* меню *Статистики - Statistics* или кнопкой  . Ошибки на отдельных наблюдениях выводятся в виде гистограммы, приведенной на рисунке 1.8 .

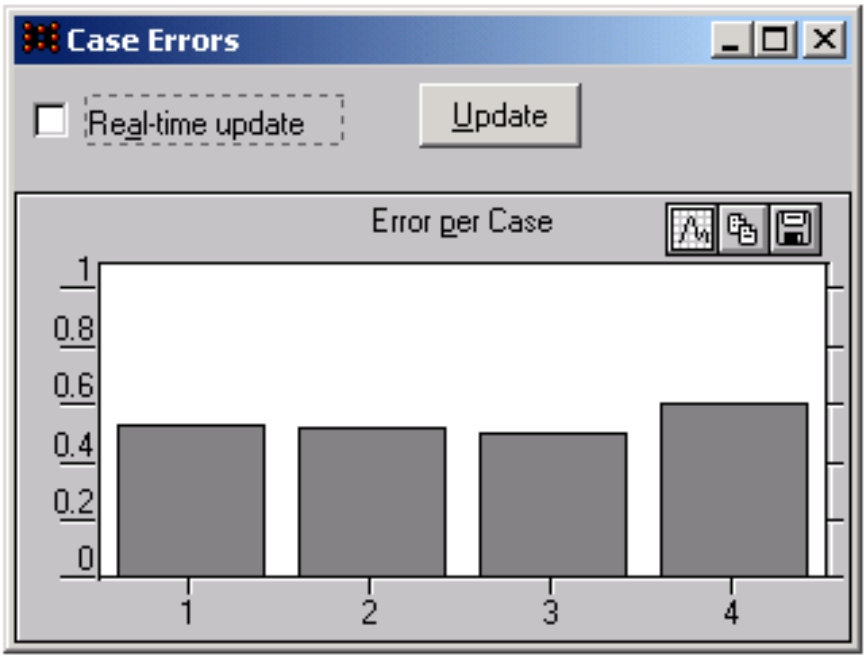


Рисунок 1.8 – Диалоговое окно Ошибки наблюдений... - Case Errors

В конце сеанса обучения ошибки пересчитываются. Имеется также возможность следить за тем, как они меняются в процессе обучения - для этого служит функция *Пересчитывать по ходу -Real-time update* окна *Ошибки наблюдений... - Case Errors;* ее нужно активизировать перед запуском алгоритма обратного распространения. Сделав это, вы сможете наблюдать, как алгоритм пытается искать компромисс между обучающими и мешающими наблюдениями.

1.6. ЗАПУСК НЕЙРОННОЙ СЕТИ

После того, как сеть обучена, ее можно запустить на исполнение. В пакете *ST Neural Networks* это можно сделать в нескольких вариантах:

- на текущем наборе данных - в целом или на отдельных наблюдениях;
- на другом наборе данных - в целом или на отдельных наблюдениях (такой набор данных уже может не содержать выходных значений и предназначаться исключительно для тестирования);
- на одном конкретном наблюдении, для которого значения переменных введены пользователем, а не взяты из какого-то файла данных;
- из другого приложения с помощью интерфейса прикладного программирования SNN API.

Рассмотрим вначале запуск сети на текущем наборе данных.

При запуске сети на текущем наборе данных возможны два варианта: либо обрабатывать отдельные наблюдения, либо все множество целиком. Во втором

варианте подсчитываются суммарные статистики - они будут подробно описаны в последующих разделах.

Обработка наблюдений по одному

Для обработки отдельных наблюдений из набора данных служит окно *Прогнать одно наблюдение - Run Single Case* (рисунок 1.9), доступ к которому осуществляется через пункт *Одно наблюдение -Single Case...* меню *Запуск - Run* или кнопкой  .

В поле *Номер наблюдения - Case No* задается номер наблюдения, подлежащего обработке. Чтобы обработать текущее наблюдение, нажмите кнопку *Запуск - Run*, а для обработки какого-либо другого наблюдения введите соответствующий номер в поле *Номер наблюдения - Case No* и нажмите клавишу ВВОД.

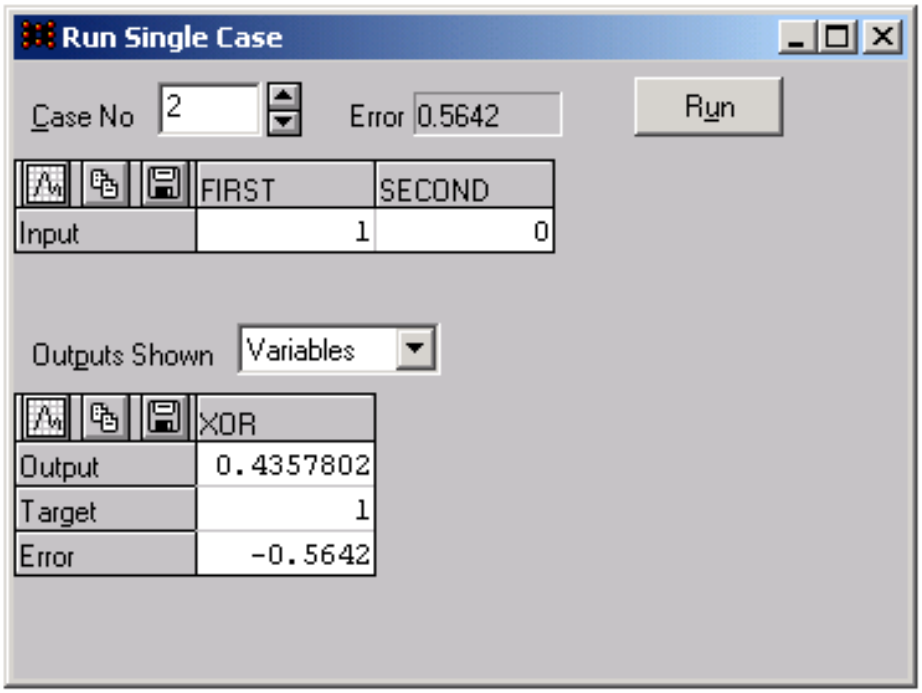


Рисунок 1.9 – Диалоговое окно Прогнать одно наблюдение

Run Single Case

Совет. Самый простой способ работы с этим окном такой: просчитать первый тестовый пример, нажав кнопку *Запуск - Run.* а затем, нажимая верхнюю стрелку на кнопке микропрокрутки, которая расположена справа от поля *Номер наблюдения - Case No,* последовательно обрабатывать другие наблюдения.

Значения входных переменных для текущего наблюдения отображаются в таблице, расположенной в верхней части окна, а выходные значения - в нижней таблице.

Помимо фактического выходного значения, которое выдает сеть, выводится также целевое значение и ошибка, т.е. разность между первым и вторым. Если для обучения сети использовался этот же набор данных, то фактическое значение должно быть близко к целевому. Кроме того, ошибка, соответствующая данному наблюдению, выводится отдельно в верхней части окна (поле *Ошибка — Error).*

В пакете *ST Neural Networks* предусмотрены различные форматы вывода (для этого служит выпадающий список *Показывать при выводе - Outputs Shown).* Сейчас мы использовали установленный по умолчанию вариант *Переменные - Variables.*

Прогон всего набора данных

Для тестирования сети на всем наборе данных служит окно *Прогнать набор данных - Run Data Set* (рисунок 1.10), доступ к которому осуществляется через пункт *Набор данных - Data Set...* меню *Запуск - Run* или кнопкой  . Нажмите кнопку *Запуск — Run,* чтобы протестировать сеть, и результаты будут выведены в таблицу в нижней части окна.

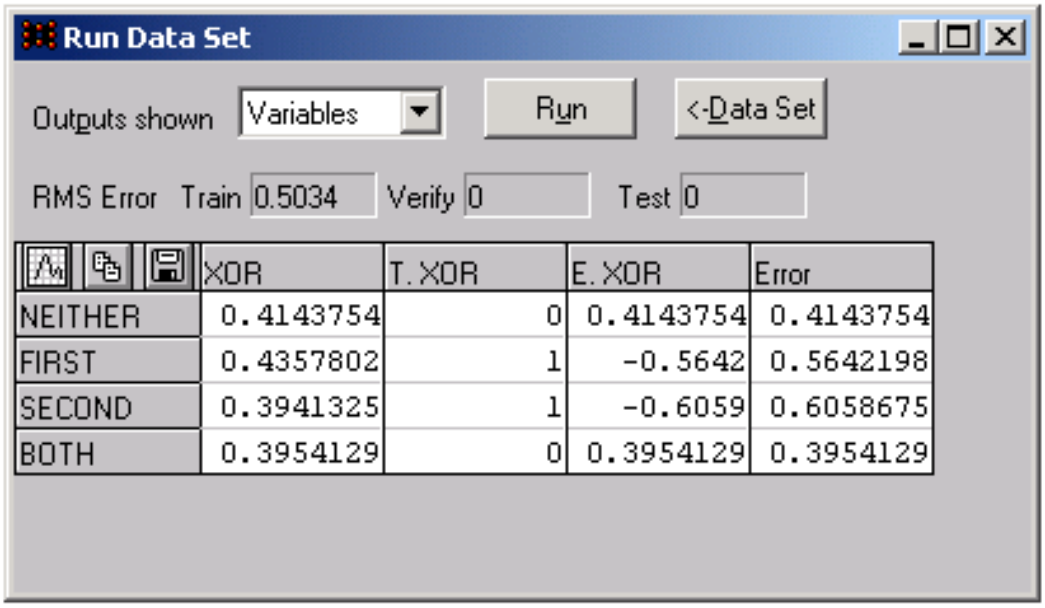


Рисунок 1.10 – Диалоговое окно Прогнать набор данных - Run Data Set

В таблице окна *Прогнать набор данных - Run Data Set* содержатся следующие значения (перечисленные слева направо): фактические выходы сети, целевые выходные

значения, ошибки, т.е. разности между первыми и вторыми, и суммарная ошибка по каждому наблюдению (в сети, построенной для задачи «исключающего или», значительная часть информации повторяется). Над таблицей выдается окончательная *среднеквадратичная ошибка (СКО) - RMS error* сети на этом наборе данных (эта же статистика выводится в окне *График обучения - Training Graph).* В нашем примере весь набор данных использовался как обучающее множество, поэтому отличной от нуля получилась только *СКО обучения - Train RMS Error.*

Тестирование на отдельном наблюдении

Иногда необходимо протестировать сеть на отдельном наблюдении, не принадлежащем никакому набору данных. Причины для этого могут быть такие:

- Обученная сеть используется для построения прогнозов на новых данных с неизвестными выходными значениями.
- Вы хотите поэкспериментировать с сетью, например, проверить чувствительность результата к малым изменениям в данных.

Тестирование заданных пользователем наблюдений проводится из окна *Прогнать отдельное наблюдение - Run One-off Case* (рисунок 10), доступ к которому осуществляется через пункт *Отдельное - One-off...* меню *Запуск -Run.*

Для этого нужно ввести входные значения в таблицу, расположенную в верхней части окна, и нажать кнопку *Запуск - Run,* результаты будут выведены в нижнюю таблицу (рисунок 1.11).

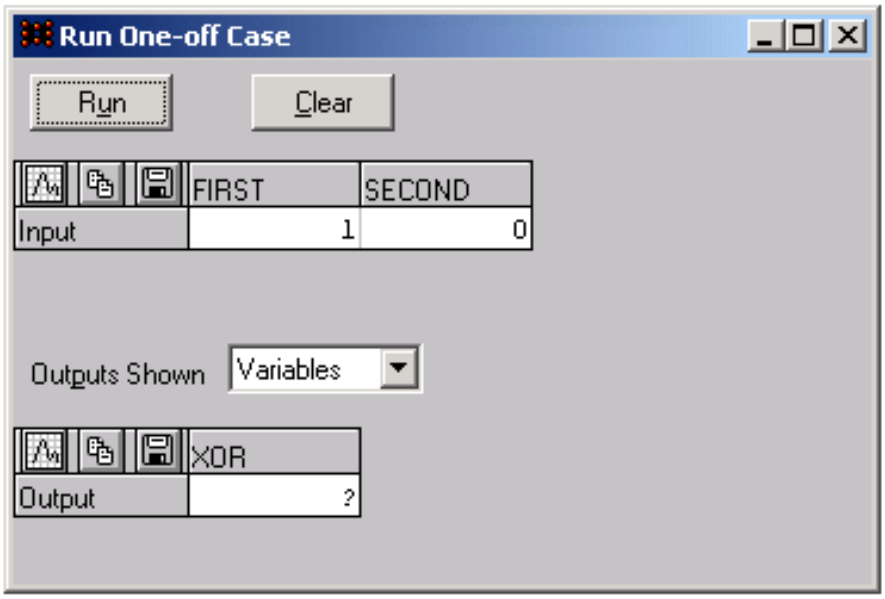


Рисунок 1.11 – Диалоговое окно Прогнать отдельное наблюдение - Run One-off Case

Попробуйте, например, немного менять значения исходных данных задачи «исключающего или», например, задать в какой-нибудь строке числа 0,1 и 0,9. Как правило, нейронные сети неплохо работают при наличии помех. При малых возмущениях исходных значений результаты будут близки к ожидаемым.

1.7. ПРОВЕДЕНИЕ КЛАССИФИКАЦИИ

Решение задач классификации — одна из наиболее важных областей применения нейронных сетей. В таких задачах входные данные представляют собой результаты измерений некоторых характеристик объекта. Цель состоит в том, чтобы определить, к какому из нескольких заданных классов принадлежит этот объект. Обычно классов бывает ровно два (или один, и наблюдение может либо принадлежать, либо не принадлежать ему). Задача «исключающего или» - пример задачи классификации с двумя классами. Наблюдение может принадлежать или не принадлежать классу Лог.

В пакете *ST Neural Networks* можно работать с так называемыми номинальными переменными (или атрибутами), то есть с переменными, которые могут принимать конечное число значений, представленных в виде строк текста. Простейший пример - переменная *Пол = {Муж, Жен},* это номинальная переменная с двумя возможными значениями (состояниями). В задаче «исключающего или» выходная переменная как раз должна быть номинальной переменной с двумя состояниям и: Лог = *{False, True}.*

В описываемой программе номинальными могут быть как входные, так и выходные переменные, и имеется много способов преобразования содержащейся в них нечисловой информации к виду, понятному нейронной сети, и, наоборот, способов интерпретировать числовой выход сети как номинальную переменную. Поддержка номинальных переменных - органическая часть системы пре/пост-процессирования пакета ST *Neural Networks.*

Проиллюстрируем сказанное, видоизменив пример XOR таким образом, чтобы выходом была номинальная переменная.

Задача XOR с номинальной выходной переменной

Начнем с того, что заново определим набор данных. Используем команду *Набор данных - Data Set...* меню *Файл-Создать - File-New* и определим две входные и одну выходную переменную.

В открывшемся окне *Редактор данных - Data Set Editor* (рисунок 1.12) зададим имена переменных *First, Second* и *Xor* (дважды щелкая на заголовках столбцов).

Чтобы сделать переменную *Xor* номинальной, выделите ее (щелчком на заголовке столбца), затем нажмите правую кнопку мыши и выберите в появившемся контекстном меню пункт *Определение - Definition...*. Откроется диалоговое окно *Определение переменной - Variable Definition*, в котором содержатся следующие сведения о переменной: ее имя в данный момент, число возможных номинальных значений (для числовых переменных - ноль) и сами номинальные значения.

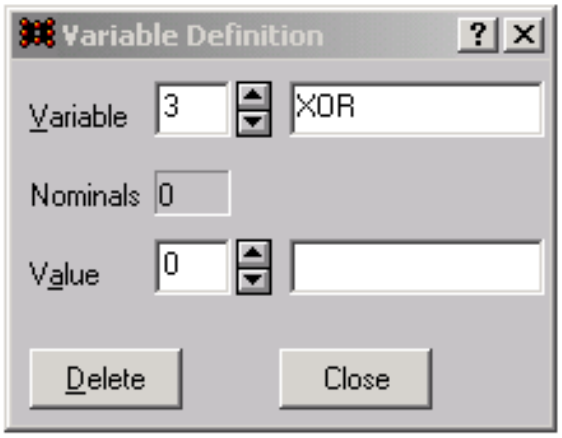


Рисунок 1.12 – Диалоговое окно Редактор данных - Data Set Editor

Нажмите верхнюю стрелку на кнопке микропрокрутки, расположенной справа от поля *Значение - Value: y* переменной появится два номинальных значения – *v1* и *v2* (номинальных значений должно быть не менее двух — единственное номинальное значение не имеет никакого смысла).

Поменяйте имя первого номинального значения с *v1* на *False (Ложь),* затем нажмите верхнюю стрелку на кнопке микропрокрутки и поменяйте имя второго номинального значения с *v2* на *True (Истина).*

Нажмите кнопку *Закрыть - Close* — номинальная переменная определена.

Теперь можно ввести данные наблюдений, чтобы набор данных выглядел как на рисунке 1.13. Вводить значения номинальной переменной можно разными способами: напрямую напечатать слова *(True* или *False),* ввести соответствующее порядковое значение (соответственно 1 или 2) или, выделив ячейку, щелкнуть правой кнопкой и выбрать значение из контекстного меню.

После того, как набор данных будет создан, создайте сеть с помощью команды *Сеть - Network...* меню *Файл-Создать - File-New.* Нажмите кнопку *Совет - Advise* - программа *ST Neural Networks* автоматически выберет параметры пре/пост-процессирования и архитектуру сети.

Задайте число скрытых элементов равным двум. Обучите сеть с помощью алгоритма обратного распространения.

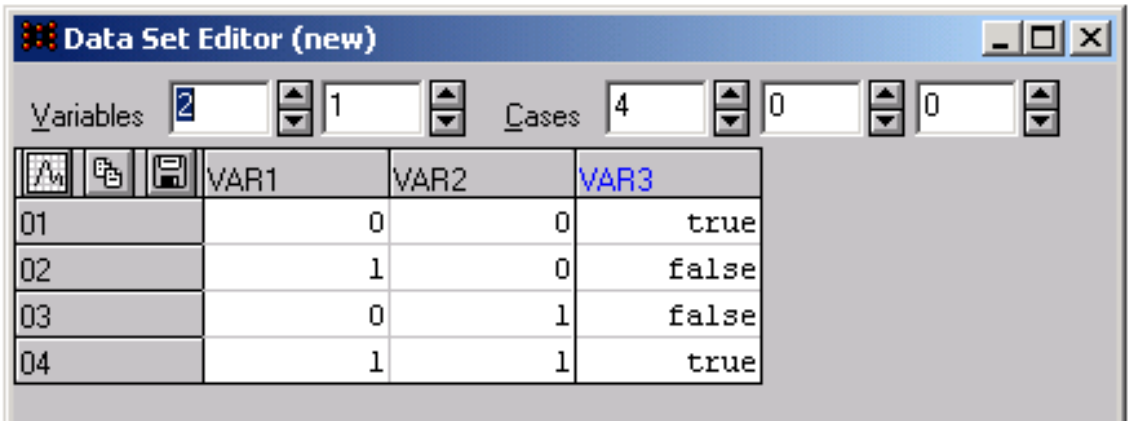


Рисунок 1.13 –Ввод данных наблюдений

Замечание. Функция преобразования *(Convert)* для выходной переменной будет изменена с *Минимакс - Minimax* на *Два значения - Two-State:* программа *ST Neural Networks* автоматически определила, что это двузначная номинальная выходная переменная, и соответствующим образом изменила режим препроцессирования. Такой способ - обычный при решении на нейронной сети задачи двузначной классификации (т.е. с двумя классами): двузначной переменной соответствует один выходной элемент, который будет выдавать значение 1 для одного из классов и 0 - для

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Основы компьютерного моделирования: химико-технологических процессов [Текст] : учеб. пособие для вузов / Т. Н. Гартман, Д. В. Клушин. - М. : "Академкнига", 2006. - 416 с. : рис. - (Учебное пособие для вузов). - Библиогр.: с. 413.
2. Моделирование в химической технологии и расчет реакторов [Текст] : учеб. пособие для студентов и слушателей Ин-та дополнительного профессионального образования / Н. А. Самойлов. - Уфа : Монография, 2005. - 224 с. : ил., табл. - (ГОУ ВПО УГНТУ. Ин-т доп. проф. образования). - Библиогр.: с.219
3. Математическое моделирование основных процессов химических производств [Текст] : учебное пособие / В. В. Кафаров, М. Б. Глебов. - М. : Высш. шк., 1991. - 400 с. : рис., табл. - Библиогр.: с. 365.
4. Математическое моделирование основных процессов химических производств [Текст] : учебное пособие / В. В. Кафаров, М. Б. Глебов. - М. : Высш. шк., 1991. - 400 с. : рис., табл. - Библиогр.: с. 365.
5. Методы кибернетики в химии и химической технологии [Текст] : учебное пособие / В. В. Кафаров . - М. : Химия, 1968. - 379 с.
6. Моделирование физико-химических процессов нефтепереработки и нефтехимии [Текст] / Ю. М. Жоров. - М. : Химия, 1978.