Федеральное агентство по образованию

Государственное образовательное учреждение
высшего профессионального образования
«Самарский государственный аэрокосмический университет
имени академика С.П.Королева»

ПРОГРАММНЫЕ СТАТИСТИЧЕСКИЕ КОМПЛЕКСЫ.

Лабораторный практикум

Самара 2007 г.

Составитель: А.С.Кучеров

УДК 681.3.01

Программные статистические комплексы. Часть 1. STADIA и STATGRAPHICS: Лабораторный практикум/ Самар. гос. аэрокосм. ун-т; Сост. *А.С.Кучеров.*- Самара, 2005. 32 с

В настоящем лабораторном практикуме содержится краткое описание программных статистических комплексов STADIA и STATGRAPHICS, приводится общий порядок работы с ними, а также рекомендации по применению конкретных статистических методов.

Указания предназначены для студентов специальности 200503 «Стандартизация и сертификация», изучающих учебную дисциплину «Программные статистические комплексы», а также могут быть полезны для студентов других специальностей при изучении современных методов статистического анализа. Подготовлены кафедрой конструкции и проектирования летательных аппаратов.

Печатается по решению редакционно-издательского совета Самарского государственного аэрокосмического университета.

Рецензент – д.т.н., профессор Куренков В.И.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время на рынке компьютерного обеспечения имеются сотни различных программ, предназначенных для статистической обработки данных. Наиболее мощные из них (такие, как SAS), относящиеся к профессиональным программным статистическим комплексам, предназначены для обработки сверхбольших объемов данных и предъявляют весьма высокие требования как к подготовке пользователя, так и к характеристикам используемой вычислительной техники.

В учебном курсе «Программные статистические комплексы» студенты знакомятся с универсальными комплексами, которые сочетают в себе такие достоинства, как богатый набор реализуемых ими статистических процедур и достаточная простота их освоения. К числу таких комплексов относятся STADIA, STSTGRAPHICS и STATISTICA; с первыми двумя из них студенты знакомятся при выполнении лабораторных работ, описанных в настоящих методических указаниях.

Лабораторные работы выполняются студентами индивидуально. Для получения зачета студент должен уметь ответить на контрольные вопросы, изложенные в конце указаний к каждой работе.

1 ПРОГРАММНЫЙ СТАТИСТИЧЕСКИЙ КОМПЛЕКС STADIA

Программный статистический комплекс **STADIA** относится к универсальным ПСК. Ему доступны такие методы статистического анализа данных, как расчет показателей описательной статистики, дисперсионный и регрессионный анализ, методы так называемого разведочного анализа данных, анализ временных рядов и некоторые другие. В то же время возможности комплекса **STADIA** в части методов контроля качества довольно ограничены: хотя он и позволяет строить простейшие контрольные карты и диаграммы Парето, но такие методы, как планирование эксперимента и построение планов контроля, ему недоступны.

ПСК **STADIA** - отечественной разработки. Ведение диалога на русском языке и удобный интерфейс позволяют быстро научиться работе с комплексом. **STADIA** предъявляет весьма скромные требования к производительности персонального компьютера и требует для своей установки всего 4 Мбайт дополнительной памяти.

Файлы данных, создаваемых ПСК **STADIA,** имеют расширение **.std**; имеется возможность импорта данных, записанных в других форматах, в том числе с расширениями **.txt** и **.dbf.**

1.1 Интерфейс и состав комплекса

Под строкой заголовка окна ПСК **STADIA** находится линейка команд **главного меню,** часть из которых дублируется «горячими» функциональными клавишами.

Ниже размещается линейка кнопок **пиктографического меню** с всплывающими подсказками:

 - чтение данных из дискового файла и запись в файл;

 - вырезание, копирование данных в буфер обмена и вставка из него;

 - печать содержимого активной страницы;

 - настройка шрифта активной страницы.

Основную часть экрана занимает **электронная таблица,** которая предназначена для ввода, хранения и редактирования исходных данных. Столбцы таблицы соответствуют переменным, строки – значениям переменных (наблюдениям). Для работы с таблицей необходимо активизировать закладку **Dat** в нижней части окна.

Графопостроитель вызывается командой главного меню **График=F6** и позволяет строить графики исходных данных и визуализировать результаты выполненного анализа. Графики выводятся на графические страницы, которых может быть до 8. Каждая страница снабжается закладкой **Gr** с соответствующим номером.

Для выполнения вычислений служит **калькулятор,** который вызывается командой **Вычисл=F7.**

Блок преобразований, вызываемый командой **Преобр=F8**, позволяет производить различные преобразования над исходными данными и генерировать заданные последовательности чисел.

Блок статистики содержит большое количество статистических процедур; его вызов осуществляется командой главного меню **Статист=F9**.

Текстовый редактор служит для выдачи результатов выполненного статистического анализа и их редактирования. Работа с редактором становится возможной после активизации закладки **Rez.**

1.2 Порядок проведения статистического анализа

Для статистического анализа данных необходимо последовательно выполнить следующие действия.

1.Ввести данные в электронную таблицу с клавиатуры или из дискового файла.

2.Выбрать метод анализа, для чего необходимо задействовать блок статистики.

3.Выбрать переменную для анализа из бланка выбора (рис. 1), который появляется на экране после выбора метода анализа. Вид бланков выбора зависит от используемых методов анализа, но общие принципы их построения сохраняются.

Рисунок 1 – Бланк выбора

В левой части бланка имеется поле, содержащее полный перечень переменных, находящихся в таблице. Для статистической обработки переменной ее необходимо выделить щелчком левой клавиши мыши и нажать на изображение стрелки, направленной вправо – имя переменной появится в поле **Для**. Удаление переменной из этого поля осуществляется с помощью стрелки, направленной влево. Сделанный выбор переменных для анализа подтверждается нажатием кнопки **Утвердить.** Кнопка **Отменить** отменяет сделанный выбор.

4.Выполнить диалог, содержание которого зависит от реализуемого метода анализа.

5.Получить графики, иллюстрирующие полученные результаты.

Некоторые графики выводятся в обязательном порядке («по умолчанию»), для получения других необходимо дать положительный ответ на запрос системы.

Рисунок 2 – Меню выбора

Ряд графиков носит справочный характер, и после их выдачи на экран появляется меню **Посмотрите график** (рис. 2). График может быть оставлен на графической странице после нажатия в этом меню кнопки **Оставить.**

По окончании анализа его результаты и исходные данные можно удалить с экрана, выполнив команду главного меню **Файл \ Очистить.** Очистка электронной таблицы и очистка страницы текстового редактора выполняются независимо друг от друга. Страницы графического редактора закрываются с помощью кнопки .

2. ПРОГРАММНЫЙ СТАТИСТИЧЕСКИЙ КОМПЛЕКС STATGRAPHICS

2. 1 Общие сведения

Рассматриваемая версия **STATGRAPHICS PLUS** for Windows включает более 250 статистических и системных процедур, применяющихся в бизнесе, экономике, маркетинге, медицине, на производстве и других областях.

Пакет имеет модульную структуру. Каждой группе модулей соответствует свое меню. Базовую систему образуют следующие меню:

- Describe: статистические методы анализа по одной и множеству переменных, подбора распределений, табуляции и кросс-табуляции данных;

- Compare: методы сравнения двух и более выборок, процедуры одно- и многофакторного анализа;

- Relate: процедуры простого, полиномиального и множественного регрессионного анализа.

Из меню **Special** открываются дополнительные модули:

- Quality Control (Контроль качества): процедуры построения Парето-карт, **X** и **R –** карт;

- Experimental Design (Планирование эксперимента);

- Time-Series Analysis (Анализ временных рядов): описательные методы, процедуры сглаживания рядов, сезонной декомпозиции и прогнозирования;

- Multivariate Methods (Многомерные методы): факторный, кластерный, дискриминантный и канонический корреляционный анализ, анализ по методу главных компонент.

Система обеспечивает связь со всеми Windows-приложениями посредством технологий **OLE** и **DDE.** Кроме того, система может обмениваться данными с другими программными продуктами, использующими Lotus, dBASE, DBF, DIF, ASCII – файлы.

Интегрированная и интерактивная графика сопровождает выполнение каждой статистической процедуры. Все элементы графических изображений (цвета, масштабы, надписи и пр.) после выделения могут редактироваться с помощью контекстного меню. Выделив точку на графике, можно получить информацию, связанную с ней в таблице данных. Трехмерные изображения можно вращать и рассматривать с разных сторон.

Для создания собственных статистических проектов имеется специальное средство **Statfolio** – возможность сохранять в специальном файле выбранные методы анализа, параметры статистических процедур, виды графических отображений результатов анализа, табличные формы и пр. При загрузке в **Statfolio** новых данных они автоматически подвергаются заданной процедуре обработки.

В состав ПСК входит интеллектуальная экспертная система **StatAdvisor (Статконсультант),** интерпретирующая результаты анализа, определяющая значимые эффекты и выявляющая ошибки в проведенном анализе.

Для составления отчетов используется средство **StatGallery,** позволяющее располагать в одном окне или на одном листе до 9 различных фрагментов текста и графических иллюстраций.

2.2 Интерфейс пользователя

В строке заголовка после названия системы следует имя загруженного статистического проекта (StatFolio); при открытии нового окна проект безымянен (Untitled). Ниже следуют строки главного и пиктографического меню.

Наиболее часто употребляются следующие кнопки пиктографического меню:

 - чтение и запись файла StatFolio;

 - чтение и запись файла данных;

 - вырезание, копирование данных в буфер обмена и вставка из него.

Далее следует ряд кнопок, вызывающих различные методы статистического анализа.

Большую часть экрана занимает рабочая область, в которую выводятся таблицы с данными, графики, комментарии. В нижней части экрана расположен набор пиктограмм, связанных с окнами:

 - электронной таблицы;

 - StatAdvisor;

 - StatGallery;

 - блокнота для записи комментариев к проводимому статистическому анализу.

Чтобы воспользоваться тем или иным окном, необходимо развернуть его, нажав кнопку на соответствующей пиктограмме.

В электронной таблице столбцы соответствуют переменным, строки – значениям переменных. Щелкнув на заголовке нужного столбца и нажав правую кнопку мыши, можно вызвать контекстное меню. Выбор в нем команды **Modify Column (Изменить столбец)** открывает одноименное диалоговое окно (ДО), в поле которого **Name** задается имя переменной (обязательно латинскими буквами); расположенными ниже переключателями определяется тип данных.

Для преобразования переменных используется команда **«Generate Data»** из контекстного меню. В открывшемся ДО можно производить манипуляции над данными с помощью различных операторов.

2.3 Порядок проведения статистического анализа

После открытия (создания) файла данных из меню выбирается соответствующая процедура их обработки, затем задаются анализируемые переменные. В рабочем поле появляются общие данные об этих переменных: число наблюдений, диапазоны изменения и т.д. В верхней части поля имеется ряд управляющих кнопок, в том числе:

 - изменение входных данных для анализа;

 - задание табличных опций;

 - задание графических опций;

 - сохранение результатов анализа.

Окна, в которых отображаются табличные и графические результаты, можно раскрыть на все поле двумя щелчками мыши, щелчок правой кнопкой открывает доступ к меню задания новых параметров. В нижней части окна табличных результатов, после заголовка **The StatAdvisor**, содержатся комментарии к полученным результатам анализа и дальнейшие рекомендации.

Чтобы повторить весь проведенный анализ на новом массиве данных, не повторяя задание все опций, необходимо выполнить команду **File | Save StatFolio As** и задать имя; затем достаточно загрузить новый файл данных **(File | Open Data File)** и вызвать записанный статистический проект **(File | Open StatFolio ,** указать имя сохраненного проекта**).**

3 ОБЩИЕ СВЕДЕНИЯ О ПРОГРАММНОМ СТАТИСТИЧЕСКОМ КОМПЛЕКСЕ STATISTICA

Программный статистический комплекс (ПСК) **Statistica,** по богатству своих возможностей и удобству работы с ним, может по праву считаться лидером среди программ статистической обработки данных в среде Windows. В **Statistica** имеются разнообразные процедуры, обеспечивающие практически все возможные виды статистического анализа, и сотни типов графиков, предназначенных для визуализации данных.

Statistica совместима с другими Windows-приложениями, с которыми она может обмениваться данными. Наличие встроенного языка программирования позволяет наращивать систему, создавая собственные методы анализа.

Рабочее окно системы соответствует стандартам программ, работающих в среде Windows. При запуске системы автоматически загружается последний файл, с которым выполнялась работы. Данные содержатся в электронной таблице, столбцы которой соответствуют переменным (**Variables**), а строки – наблюдениям (**Cases**). Таблица является мультимедийной, то есть может содержать самые различные объекты: числа, текст, рисунки, клипы и т.д. (так называемые *Active-X* объекты).

Большинство групп команд главного меню (такие как **File, Edit** и другие) являются традиционными для программ, работающих под управлением Windows.

Группа команд **Statistics** открывает доступ к различным статистическим процедурам, сгруппированным в отдельные модули. Назовем основные из них:

- базовые статистики и таблицы;

- множественная регрессия;

- однокомпонентный дисперсионный анализ;

- непераметрические статистические методы;

- «подгонка» распределения;

развернутые линейные и нелинейные модели: анализ временных рядов, моделирование структурными уравнениями и другие;

многомерные методы: кластерный анализ, факторный анализ, метод главных компонент, дискриминантный анализ и т.д.;

промышленный анализ: карты контроля качества, планирование эксперимента, планы выборочного контроля;

вероятностный калькулятор.

Группа команд **Graphs** позволяет строить многочисленные графики.

Группа команд **Data** обеспечивает выполнение различных операций с данными: вставку, удаление, сортировку, импорт из внешних программ и т.д.

Группа команд **Workbook** управляет операциями с рабочей книгой, создаваемой в процессе сеанса в ПСК **Statistica**.

4 ЛАБОРАТОРНЫЕ РАБОТЫ

4. 1 Лабораторная работа №1. Описательная статистика

Цели выполняемой работы:
1) выполнить сравнительный анализ числовых характеристик заданной случайной величины с помощью табличного процессора **Excel** и ПСК **STADIA**;
2) получить случайную величину с нормированным нормальным законом распределения и проверить гипотезу о нормальности распределения;
3) выполнить построение простой регрессии.

4.1.1 Расчет показателей описательной статистики

К показателям описательной статистики относятся основные выборочные характеристики, ошибки их определения, доверительные интервалы.

Пусть имеется случайная выборка $x_1, x_2, \ldots, x_n$. Тогда выборочное среднее, выборочная дисперсия, стандартное отклонение рассчитываются по следующим формулам:

$$\hat{m} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad \hat{D} = \frac{\sum_{i=1}^{n} (x_i - \hat{m})^2}{n-1}, \qquad \hat{\sigma} = \sqrt{\hat{D}}. \tag{1}$$

Ошибки определения выборочного среднего и стандартного отклонения рассчитываются по формулам

$$E_m = \frac{\hat{\sigma}}{\sqrt{n}}, \qquad E_\sigma = \hat{\sigma}\sqrt{\frac{2}{4(n-1)+1}}. \tag{2}$$

Нижняя и верхняя границы доверительных интервалов для выборочного среднего с доверительной вероятностью β определяются зависимостями

$$I_{1m} = m - \frac{t_\beta \hat{\sigma}}{\sqrt{n}}, \qquad I_{2m} = m + \frac{t_\beta \hat{\sigma}}{\sqrt{n}}, \tag{3}$$

где t_β - статистика Стьюдента для уровня значимости $p = 1-\beta$ с $n-1$ степенью свободы.

Нижняя и верхняя границы доверительных интервалов для выборочной дисперсии определяются зависимостями

$$I_{1D} = \frac{(n-1)\hat{D}}{\chi_1^2}, \qquad I_{2D} = \frac{(n-1)\hat{D}}{\chi_2^2}, \tag{4}$$

где χ_1^2 - статистика хи-квадрат для уровня значимости $p = (1-\beta)/2$ с $n-1$ степенью свободы, χ_2^2 - статистика хи-квадрат для $p = 1-(1-\beta)/2$ с $n-1$ степенью свободы.

Задача. Вычислить показатели описательной статистики для заданной переменной с использованием табличного процессора **Excel** и ПСК **STADIA**.

Порядок выполнения.

1. Использование табличного процессора Excel. Ввести в столбец электронной таблицы значения переменной, заданные преподавателем. Используя формулы (1), (2), рассчитать выборочные характеристики переменной и ошибки их определения.

Используя формулы (3), (4), рассчитать границы доверительных интервалов при доверительной вероятности β=0,95.

Указание. Для вычисления статистики Стьюдента использовать функцию =СТЬЮДРАСПОБР с первым аргументом $p = 1 - \beta$ и вторым аргументом $n - 1$. (После вызова **Мастера функций** указанную функцию можно найти в категории **Полный алфавитный перечень**). Для вычисления статистик хи-квадрат использовать функцию =ХИ2ОБР вначале с аргументами $p = (1 - \beta)/2$ и $n - 1$, затем - с аргументами $p = 1 - (1 - \beta)/2$ и $n - 1$.

Записать полученные результаты: выборочное среднее и его ошибку, выборочную дисперсию, стандартное отклонение, границы доверительных интервалов.

2. Использование ПСК STADIA..Запустить ПСК **STADIA** и в первый столбец электронной таблицы **(x1)** ввести значения переменной, заданные преподавателем.

Находясь на закладке **Dat,** вызвать меню статистических методов и нажать в нем кнопку **1=Описательная статистика.** В бланке выбора переменных выбрать для анализа переменную **x1.** На запрос системы **Выдать дополнительную статистику?** ответить **Yes.** На запрос о записи результатов в матрицу данных – **No.**

Записать полученные результаты: выборочное среднее и его ошибку, выборочную дисперсию, стандартное отклонение, границы доверительных интервалов. Сравнить полученные результаты со значениями, рассчитанными с помощью табличного процессора Excel.

Произвести очистку экрана.

4.1.2 Получение случайной величины с нормальным законом распределения. Проверка распределения на нормальность

В статистическом анализе большую роль играют случайные величины (СВ) с нормальным законом распределения. Нормальная случайная величина X с математическим ожиданием m и среднеквадратическим отклонением σ может быть получена с помощью следующего преобразования

$$X = m + \sigma \cdot \gamma ,$$

где γ - так называемая нормированная нормальная случайная величина, у которой математическое ожидание равно нулю, а дисперсия – единице. Для получения величины γ используют следующее свойство: при сложении большого количества независимых случайных величин получается случайная величина, распределенная по нормальному закону (***Центральная предельная теорема***). Величина γ вычисляется по формуле

$$\gamma = \sum_{i=1}^{12} R_i - 6 ,$$

где R_i - случайная величина, распределенная с равномерной плотностью на интервале [0, 1]. На практике даже сумма трех-четырех независимых случайных величин уже мало отличается от случайной величины с нормальным законом распределения.

Убедимся в этом, выполнив следующее упражнение.

1.Получение значений случайной величины, распределенной с равномерной плотностью на интервале [0, 1]. Установить курсор в первую ячейку столбца **x1** и в головном меню **Преобр=F8** выполнить пункт **3=генератор чисел**. В поле **Всего чисел** ввести значение **80**, в поле **a=** ввести **0**, в поле **b= 1** и нажать кнопку **3=равномерное.** Повторить операцию для столбцов **x2, x3, x4.**

2.Получение значений нормированной случайной величины, распределенной с нормальной плотностью. Установить курсор в первую ячейку столбца **x5,** в головном меню **Преобр=F8** выполнить пункт **2=задаваемая функция.** В свободную строку блан-

ка формул ввести выражение (x1+x2+x3+x4-2). Щелчком мыши активизировать переключатель слева от поля формулы и нажать кнопку **Утвердить.** В столбце **x5** появятся значения случайной величины, распределенной с нормальной плотностью и имеющей характеристики $m = 0, \sigma = 1$.

3. Построение гистограммы и проверка на нормальность. С помощью пункта меню статистических методов **2=Гистограмма / нормальность** выполнить проверку на нормальность вначале для одной из случайных величин **x1…x4** (на выбор), затем – для величины **x5**.

Записать для каждой величины результаты проверки гипотезы о нормальности по каждому из трех критериев, указав соответствующие уровни значимости.

Произвести очистку экрана и закрыть графические окна.

4.1.3 Простая регрессия

Процедура построения простой регрессии позволяет исследовать зависимость между двумя величинами, подбирая наиболее подходящую модель.

В блоке статистики процедура находится в разделе **Регрессионный анализ** и вызывается нажатием кнопки **L=Простая регрессия / Тренд.** Бланк выбора переменных для регрессионного анализа имеет в правой части два поля: **Y –переменная**, для ввода зависимой переменной, и **X-переменные,** в которое вводятся независимые переменные (в случае простой регрессии такая переменная одна).

После выбора переменных для анализа появляется окно **Регрессия,** в котором производится выбор модели регрессии. Затем открывается окно **Интерполяция,** позволяющее осуществить интерполирование значения зависимой переменной для указанного в нем значения независимой переменной. Далее система предлагает построить график регрессии. Следующее окно (рис. 3) предоставляет дополнительные возможности анализа. В частности, для выполнения прогноза нужно указать число шагов (в поле **шагов=**) и величину шага прогнозирования (**шаг=**).

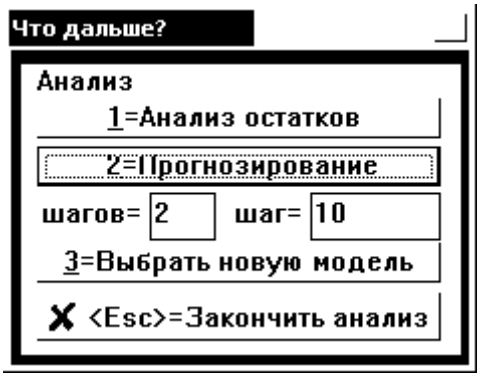


Рисунок 3 – Окно прогнозирования

Задача. Проанализировать тенденцию изменения средней урожайности зерновых культур в СССР в период с 1945 по 1989 г.г., используя данные, содержащиеся в файле **corn.std.**

Порядок выполнения. 1. Открыть файл **corn,** находящийся в папке **stadia \ dat,** для чего нажать клавишу **F3 ,** в левой части открывшегося диалогового окна **Чтение файла** выбрать нужный файл и дважды щелкнуть левой кнопкой мыши.

2. Проанализировать зависимость переменной **zerno**, содержащей значение средней урожайности (в центнерах с гектара) от переменной **data.** Для этого вызвать графопостроитель и выбрать тип графика **1=функциональный.** Выбрать для анализа все переменные. В полученной зависимости наблюдается линейно возрастающая тенденция, поэтому можно попытаться построить линейную регрессионную зависимость.

3. Вызвать процедуру построения простой регрессии и выбрать в качестве независимой переменной **data,** в качестве зависимой – **zerno.** Выбрать модель линейной регрессии и отменить выполнение интерполяции. Затребовать вывод графика регрессии и оставить его на экране. Выполнить построение прогноза на последующие 10 лет, сохранив график. В диалоговом окне **Что дальше?** выбрать опцию завершения анализа.

Записать: уравнение регрессии и уровень значимости гипотез о равенстве нулю ее параметров.

Завершить работу с программой.

Контрольные вопросы

1.Каким образом в ПСК STADIA осуществляется построение графиков?

2. Каким образом можно вызвать требующуюся статистическую процедуру?

3.Как осуществляется переход между окном текстового редактора, электронной таблицей и графическими окнами?

4.Изложите общий порядок проведения статистического анализа в ПСК STADIA.

5.Как с помощью ПСК STADIA рассчитать показатели описательной статистики?

6.Как в ПСК STADIA получить случайную величину с заданным законом распределения?

7. Как в ПСК STADIA рассчитать значение переменной по заданной формуле?

8. Как с помощью ПСК STADIA проверить, имеет ли случайная величина нормальное распределение?

9. Как в ПСК STADIA выполнить построение регрессии?

10. Как в ПСК STADIA выполнить прогноз?

4.2 Лабораторная работа № 2

Разведочные методы анализа

Цели выполняемой работы:

1) освоить методы разведочного анализа данных;

2) познакомиться с ПСК **Statgraphics:** изучить общий порядок работы и такие возможности, как создание статистического проекта и использование интерактивной графики.

Так называемые разведочные методы анализа (в английской терминологии – data mining, т.е. «добыча данных») используются на начальной стадии обработки многомерных данных, когда связи между ними не вполне ясны. К этой группе методов относят кластерный и дискриминантный анализ, методы многомерного шкалирования, главных компонент и некоторые другие.

4.2.1 Кластерный анализ

Кластерный анализ позволяет разбить множество объектов наблюдения на заданное (или неизвестное заранее) число классов.

Задача. Разбить множество литературных персонажей на кластеры, используя результаты оценки их черт характера (величина оценки соответствует степени присутствия данного качества).

Порядок работы.1.Запустить ПСК **STADIA.** Открыть файл данных **tales**.**std** («сказочные персонажи»).

2.Командой главного меню **Статист=F9** открыть ДО **Статистические методы** и в группе **Многомерные методы** выбрать **Q=Кластерный.** Внести в поле для анализа все числовые переменные (не вносить текстовые переменные **Характер** и **Персонаж**). Нажать кнопку **Утвердить.**

3.В ДО Исходные данные выбрать пункт **1=Переменные*Объекты,** в ДО **Метрика вычисления расстояний –** пункт **1=Эвклид.**

4. Из стратегий кластеризации (рис. 4) выбрать категорию **Объединяющая, 1=ближайшего соседа.** На запрос **Выдать таблицу расстояний?** ответить **Yes.** В ДО **Посмотрите график** нажать кнопку **Оставить.** Объединяющая стратегий кластеризации формирует кластеры, последовательно добавляя объекты, все более удаленные от центра кластера, положение которого пересчитывается по мере включения новых объектов):

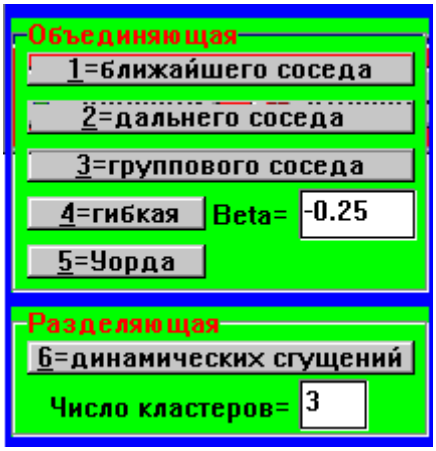


Рисунок 4 – ДО выбора стратегии

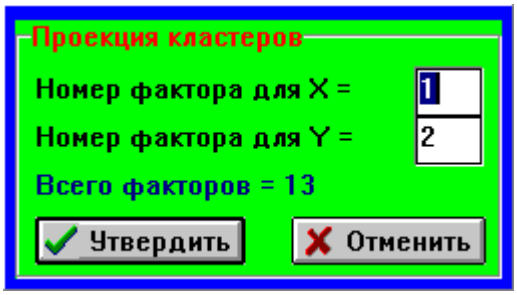


Рисунок 5 – ДО выбора проекции

5.По таблице результатов кластерного анализа можно проследить порядок включения в кластеры новых объектов.

К л а с т е р ы:

(список объектов) -> расстояние

(6,3) --> 1,682
(6,2,3) --> 2,325
(8,7) --> 2,599
(8,1,7) --> 3,002
(5,4) --> 3,034
(8,6,2,3,1,7) --> 3,08
(8,5,4,6,2,3,1,7) --> 3,813

Построенный график (дендрограмма) иллюстрирует этот процесс (расстояния откладываются по оси ординат).

5.Для более наглядного представления состава кластеров можно использовать другую стратегию кластеризации. Выполнить п.п. 2-3, но на следующем шаге выбрать категорию **Разделяющая,** задать **Число кластеров = 3** и нажать кнопку **6=динамических сгущений.**

На запрос **Выдать таблицу расстояний?** ответить **Yes.** В ДО **Проекция кластеров** (см. рис. 5) оставить все предложенные установки и нажать кнопку **Утвердить.** В ответ на запрос **Посмотрите график** нажать кнопку **Оставить.**

6.Анализируя таблицу результатов, можно видеть, что кластеры формируются по принципу близости объектов к центру кластера:

- к объекту 8 (Пьеро) – 1 (Айболит) и 7 (Мальвина);
- к объекту 6 (Буратино) – 2 (Кот) и 3 (Карлсон);
- к объекту 4 (Снежная королева) – 5 (Карабас-Барабас).

На графике в выбранной проекции это не всегда очевидно (так, объект 1 ближе к объекту 6, чем к 8 – центру своего кластера) – можно попытаться найти более удачную проекцию (например, задав **Номер фактора для X =1**, **Номер фактора для Y=5**) .

7.Очистить страницы результатов (**Rez**) и данных. (**Dat).**

4.2.2 Дискриминантный анализ

Этот метод анализа позволяет классифицировать новые объекты, отнеся их неким оптимальным образом к одному из ранее определенных классов.

Задача. Имеются результаты измерения содержания (в %) примесей ***x1*** и ***x2*** в 10 образцах сырья. Девять образцов классифицированы, т.е. отнесены к одному из трех классов (переменная ***x3***). Образец № 7 пока не классифицирован. Требуется отнести его к одному из классов.

Порядок работы. 1.Открыть файл данных **DA.**

2. Командой главного меню **Статист=F9** открыть ДО **Статистические методы** и в группе **Многомерные методы** выбрать **P=Дискриминантный.**

3. В полученной сводке **значимость=0 –** признак того, что полученная модель классификации объектов статистически значима. При этом образец № 7 с вероятностью 1 отнесен к третьему типу сырья.

Задание для самостоятельной работы. Классифицировать объект № 11, задав для него значения ***x1=5,и x2=5, x3=0****.*

4. Завершить работу программы.

4.2.3 Метод главных компонент

Этот метод служит для уменьшения размерности данных путем нахождения малого числа линейных комбинаций исходных переменных, объясняющих большую долю изменчивости (характеризуемой дисперсией), заключенной в исходных переменных.

Задача. Произвести сравнительную оценку автомобилей, характеризующихся несколькими различными параметрами: весом **(weight)**, числом цилиндров **(cylinders)**, ускорением **(accel.)**, объемом двигателя **(displace)**, мощностью в л.с. **(horsepower).**

Порядок работы. 1.Запустить программу **STATGRAPHICS.** Командой главного меню **File | Open | Open Data File…** открыть файл **Cardata.sf.** Здесь, помимо указанных параметров, приводятся название фирмы-изготовителя **(make)**, марка автомобиля **(model)** и ряд других данных.

2.Выполнить в главном меню команду **Special| Multivariate Methods| Principal Components…** В раскрывшемся диалоговом окне выбрать для анализа пять факторов, указанных в постановке задачи: вес **(weight)**, число цилиндров **(cylinders)**, ускорение **(accel.)**, объем двигателя **(displace)**, мощность **(horsepower). Выбор произвести в указанном порядке![1]** Нажать **<OK>.**

3.В ДО **Principal Component Analysis** появится таблица (табл. 1), содержащая собственные значения **(Eigenvalue)** главных компонент (ГК) в порядке убывания, процент дисперсии, приходящийся на каждую из них **(Percent of Variance)**, а также накопленный процент дисперсии **(Cumulative Percentage)**:

Таблица1 – Параметры главных компонент

Principal Components Analysis

Component Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	3,62417	72,483	72,483
2	1,04339	20,868	93,351
3	0,214423	4,288	97,640
4	0,0820629	1,641	99,281
5	0,035953	0,719	100,000

Можно видеть, что уже три первые главные компоненты объясняют 97,64% дисперсии исходных данных.

4.Для более детального анализа можно получить таблицу весов компонент, для чего нажать кнопку табличных опций в  верхней части окна и установить флажок **Component Weights** (Веса компонент).

Нажать **<OK>.**

Количество главных компонент, которые будут включены в полученную таблицу, можно задать, выполнив в контекстном меню команду **Analysis Options.** В открывшемся ДО (рис.6) в поле **Extract By** (Оставлять по…) положение переключателя **Minimum Eigenvalue** позволяет задать (в расположенном ниже одноименном поле) минимальное собственное значение главной компоненты, при котором она будет оставлена в таблице. По умолчанию это значение было принято равным 1, поэтому в данном случае программа оставит для дальнейшего анализа только две первые компоненты (см. табл. 1). Чтобы увеличить число анализируемых компонент до трех, необходимо установить переключатель в положение **Number of Components** (Число компонент) и ввести в одноименное поле число 3. Нажать **<OK>.**

[1] Примечание. При изменении порядка выбора параметров изменится графическая интерпретация результата.

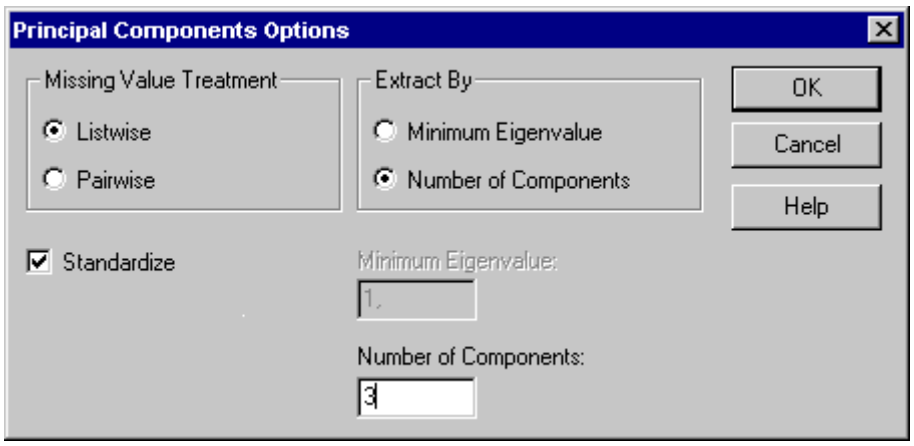


Рисунок 6 – ДО опций главных компонент

5.Из полученных результатов (табл. 2) следует, что в первой ГК близкие по значению удельные веса имеют все исходные параметры, кроме ускорения, во второй превалирует именно эта величина (accel.) . В третьей ГК наблюдается сочетание веса машины, мощности и количества цилиндров (знаки компонент не имеют значения), зато, по сравнению с первой ГК, малую роль играет объем двигателя (displace):

Таблица 2 – Веса главных компонент

Table of Component Weights

	Component 1	Component 2	Component 3
weight	0,484397	0,281143	0,426531
cylinders	0,489981	0,125914	-0,665775
accel	-0,178778	0,91435	0,130289
displace	0,507767	0,142972	-0,241578
horsepower	0,485273	-0,220516	0,547248

“Статистический советчик” StatAdvisor в качестве примера приводит зависимость первой ГК от исходных переменных:

```
0,484397*weight + 0,489981*cylinders - 0,178778*accel +
0,507767*displace + 0,485273*horsepower
```

6.Для получения графической иллюстрации нажать кнопку графических опций  и в открывшемся ДО установить флажки **Scree Plot** и **3D Scatterplot**. Первый график (рис. 7) дает изображение «факториальной осыпи», подтверждающее целесообразность использования для последующего анализа трех первых ГК.

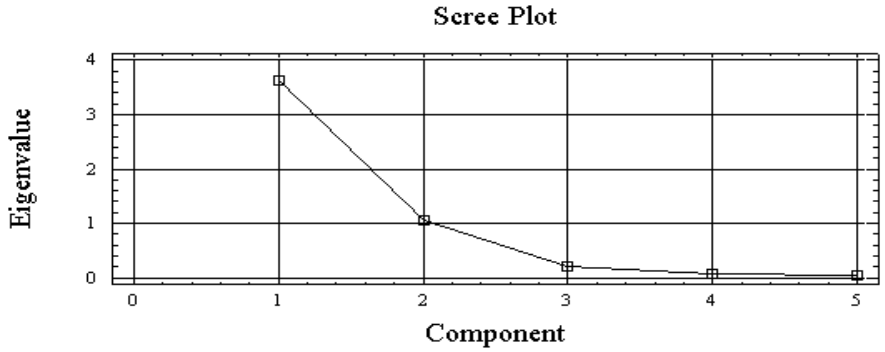


Рисунок 7 – «Факториальная осыпь»

Из второго графика (рис. 8) видно, что наиболее значительная группа машин (первая слева) обладает сравнительно небольшими значениями первой ГК (сочетания веса, количества цилиндров, мощности и объема двигателя). Зато в этой группе в широком диапазоне изменяются значения ускорения (вторая компонента). Значение третьей ГК также изменяется существенно – отсюда, сравнивая ее содержание с содержанием первой ГК, можно заключить, что малое значение первой ГК обусловлено влиянием объема двигателя, который в третьей компоненте играет малую роль.

Соответствующие выводы можно сделать и для других групп.

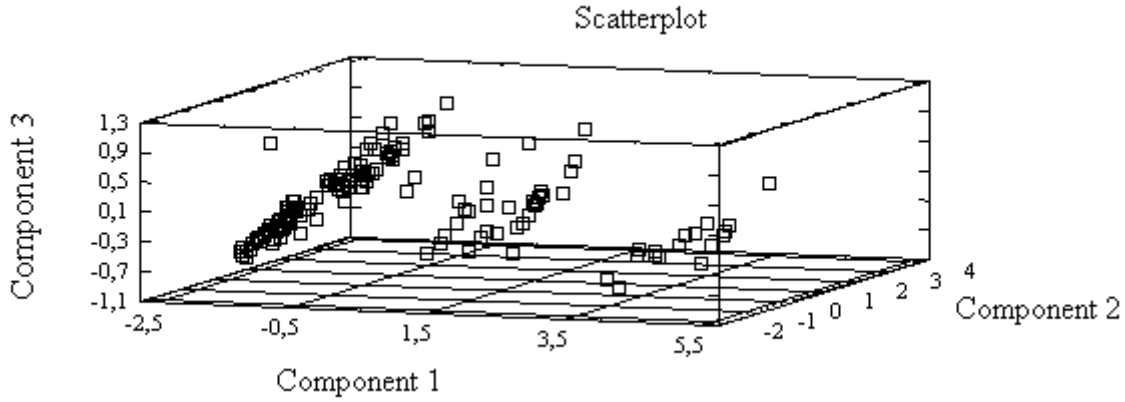


Рисунок 8 – Диаграмма рассеяния в осях главных компонент

7. Закрыть файл данных командой главного меню **File | Close | Close Data File.**

4.2.4 Дополнительные возможности ПСК Statgraphics

4.2.4.1 Статистический проект StatFolio

Если в сохраненный проект (**StatFolio**) ввести новые данные, с ними автоматически будут выполнены те же процедуры, что и с данными, для которых проект был первоначально разработан. Рассмотрим порядок использования проекта на примере процедуры множественной регрессии.

Создание StatFolio. Множественная регрессия зависимой переменной Y по независимым переменным $X_1, X_2, ..., X_n$ задается уравнением $Y = b_0 + \sum_{i=1}^{n} b_i X_i$, где $b_0, b_1, ..., b_n$ - коэффициенты регрессии.

Построим регрессию зависимой переменной Y от двух независимых по данным, приведенным в табл. 3

Таблица 3 - Исходные данные для построения регрессии

X_1	3	4	5	5	5	5	6	7	15	20
X_2	1.8	1.5	1.4	1.3	1.3	1.5	1.6	1.2	1.3	1.2
Y	2.1	2.8	3.2	4.5	4.8	4.9	5.5	6.5	12.1	15

1.Ввести в электронную таблицу значения переменных X_1, X_2, Y. Дать столбцам таблицы соответствующие названия.

2. В главном меню **Relate** выполнить пункт **Multiple Regression;** в диалоговом окне ввести в соответствующие поля зависимую переменную **(Dependent Variable)** и независимые переменные **(Independent Variables)**.

Уравнение полученной регрессии имеет вид $Y = 2{,}88148 + 0{,}718919X_1 - 1{,}51303X_2$. Уровень значимости модели (p-Value = 0,0000), процент объясняемой дисперсии (R-squared = 98%), средняя абсолютная ошибка (Mean absolute error = 0,503) говорят о высоком качестве полученной модели.

3. Вызвав диалоговое окно графических опций, построить график зависимости наблюдаемых значений переменной Y от ее значений, вычисленных по уравнению регрессии **(Observed versus Predicted)** – качество регрессии тем выше, чем меньше рассеяние точек относительно прямой.

Для оценки зависимости от каждой переменной в отдельности следует построить график **Interval Plots.** Щелкнув ПК мыши на графике и выполнив команду **Pane Options…,** в поле **Plot Versus** (Зависимость от…) открывшегося окна можно выбрать вид графика: Y в зависимости от X_1 или X_2.

4. Сохранить файл данных: выполнить команду главного меню **File** │**Save As** │**Save Data File As…,** в открывшемся ДО в поле **Имя файла** ввести ***Mydata*** и нажать кнопку **Сохранить.**

Сохранить проект (**StatFolio**): выполнить команду главного меню **File** │**Save As** │**Save StatFolio As…,** в открывшемся ДО в поле **Имя файла** ввести ***Myfolio*** и нажать кнопку **Сохранить.**

Использование StatFolio. 1.Открыть файл данных ABC. При этом программа выдаст предупреждение **Unrecognized variable: y** (Неопознанная переменная y), причиной чего является несовпадение имен переменных в файле данных с именами переменных в созданном проекте **StatFolio.** Нажать кнопку **ОК**. В текстовом и графическом полях рабочего окна также появится сообщение об ошибке, вызванной несовпадением имен.

2. Переименовать переменные: A→X1, B→X2, C→Y (программа будет выдавать предупреждения до тех пор, пока не будет переименована последняя переменная – в ответ следует нажимать кнопку **ОК**). После того, как все переменные получат имена, определенные в данном **StatFolio,** автоматически будет выполнено построение множественной регрессии и выданы результаты.

Не закрывать окно анализа.

4.2.4.2 Использование интерактивной графики

1.Идентификация точек графика по номеру строки таблицы. Развернуть (двойным щелчком) окно с графиком зависимости **Y** от **X2.** Щелкнуть на одной из точек графика. В поле с изображением бинокля, расположенном над графическим окном, появится номер соответствующей строки таблицы исходных данных; в таблице эта строка будет выделена цветом.

Можно выполнить и обратный поиск: идентифицировать точку графика, введя номер строки таблицы в 

поле и нажав на изображение бинокля.

2.Идентификация точек графика по значению переменной. Нажать расположенную над графическим  окном кнопку

В поле **Identify by:** (Идентифицировать по…) ввести имя переменной, по значению которой будет выполняться идентификация. При выделении на графике точки в поле появится ее координата. 

Возможен также поиск точки графика по ее координате, введенной в указанное поле.

Задание. Найти на графике точку, соответствующую пятой строке таблицы данных. Определить координату Y последней точки графика.

2. Ввод текста. Пользователь имеет возможность снабдить построенные графики собственными надписями. Для этого необходимо нажать кнопку  и в открывшемся окне ввести текст. Текст появится на поле графика.

Выделив текст, с помощью мыши его можно переместить в нужное место. Можно также с помощью контекстного меню задать опции текста, для чего выбрать пункт **Text options.** В открывшемся ДО кнопка **Fonts..** дает доступ к параметрам шрифта.

Задание. Введите 18 кеглем текст «My diagram», придав ему розовый цвет.

3. Удаление/добавление точек. Для этой операции используется кнопка 

Для удаления точки ее нужно предварительно выделить щелчком мыши либо ввести в поле **Row** с изображением бинокля номер соответствующего ряда таблицы данных. После удаления точки с графика результаты анализа будут автоматически обновлены, хотя вид графика не изменится. Вернуть точку на график можно повторным нажатием на кнопку 

*Задание. Удалите с графика зависимости **Y** от **X2** 1-2 точки и посмотрите, как изменятся параметры регрессии в текстовом окне.*

4. Закройте окно анализа, не сохраняя результатов.

Контрольные вопросы

1. Каково назначение метода кластерного анализа?
2. Какие стратегии кластеризации существуют?
3. Какой метод анализа позволяет отнести объект к одной из заданных групп?
4. В чем заключается назначение метода главных компонент?
5. Для чего служит график «факториальной осыпи» (Scree Plot)?
6. Какие выводы позволяет сделать диаграмма рассеяния (Scatterplot), построенная в осях главных компонент?
7. Запишите общий вид уравнения множественной регрессии.
8. Какое средство, входящее в состав ПСК Statgraphics, позволяет «автоматизировать» выполнение статистического анализа для измененных исходных данных?
9. Как в ПСК Statgraphics определить точное значение координаты точки графика?
10. Каким образом удалить / вернуть точку графика?

4.3 Лабораторная работа № 3.
Дисперсионный анализ и планирование эксперимента

4.3.1 Дисперсионный анализ

Метод дисперсионного анализа служит для исследования влияния одной или более качественных переменных (факторов) на зависимую количественную переменную (отклик).

Рассмотрим модель *однофакторного* дисперсионного анализа.

Пусть на значение случайной величины Y оказывает влияние некий фактор X, имеющий I различных уровней. Тогда по результатам измерений Y можно вычислить следующие оценки дисперсий: оценку межуровневой дисперсии SS_B, оценку внутриуровневой дисперсии SS_R и оценку полной дисперсия SS_T:

$$SS_B = \sum_{i=1}^{I} J_i \cdot (\overline{y}_i - \overline{y})^2, \quad SS_R = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \overline{y}_i)^2, \quad SS_D = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \overline{y})^2,$$

где J_i - количество измерений случайной величины Y при i- ом уровне фактора X $(i=1,2,...,I)$;

$\overline{y}_i = \dfrac{\sum_{j=1}^{J_i} y_{ij}}{J_i}$ - среднее значение, принятое величиной Y при i- ом уровне фактора X;

y_{ij} - j -ое измерение величины Y при i - ом уровне фактора;

$\overline{y} = \dfrac{\sum_{i=1}^{I}\sum_{j=1}^{J_i} y_{ij}}{n}$ - среднее значение величины Y по всей совокупности измерений;

$n = \sum_{i=1}^{I} J_i$ - общее количество измерений величины Y.

Средние значения оценок указанных дисперсий

$$MS_B = SS_B / \nu_B, \quad MS_R = SS_R / \nu_R, \quad MS_T = SS_T / \nu_T,$$

где

$$\nu_B = I - 1, \quad \nu_D = n - I, \quad \nu_D = n - 1$$

представляют собой соответствующие степени свободы.

Величина $F_0 = MS_B / MS_R$ есть так называемое F- отношение. Очевидно, что эта величина тем больше, чем сильнее средние значения измеряемой величины при различных уровнях действующего фактора X отличаются друг от друга. Рассчитав значение F, можно проверить так называемую *нулевую гипотезу H_0: фактор X статистически незначим (иначе говоря, различные уровни фактора X по своему влиянию на переменную Y в статистическом отношении не различаются).*

Если справедлива гипотеза H_0, то величина F_0 имеет *F- распределение* с ν_B и ν_R степенями свободы. Вероятность p того, что значение $F(\nu_B, \nu_R)$ превысит рассчитанное значение F_0, есть площадь под кривой плотности распределения F справа от значения F_0. В программах статистического анализа эта вероятность называется *p-уровнем* (*p-level*).

Если полученное значение p меньше заданной величины, называемой *уровнем значимости,* то гипотеза H_0 отвергается, и принимается *альтернативная гипотеза H_1: фактор X статистически значим.* Как правило, уровень значимости принимается равным 0,05.

По аналогии можно оценить влияние на количественную переменную двух и более факторов. Однако в этом случае необходимо иметь в виду, что, помимо действия каждого фактора по отдельности, может быть значимо их совместное влияние – так называемое взаимодействие.

Задача. Требуется исследовать влияние двух факторов - возраста и стажа работников - на производительность труда. В табл. 4 приведены значения средней часовой выработки (в натуральных единицах продукции) 60 работников. Здесь же дана кодировка градаций канализируемых факторов.

Таблица 4 – Исходные данные для дисперсионного анализа

Стаж	Код значения фактора «стаж»	Возраст и код значения фактора «возраст»		
		От 25 до 35 лет	От 35 до 45 лет	От 45 до 55 лет
		1	2	3
От 1 до 4 лет	1	19 20 20 20 22	19 20 20 23 25	18 19 20 21 23
От 7 до 4 лет	2	30 31 32 32 34	20 29 30 31 31	19 25 25 26 26
От 7 до 10 лет	3	35 35 39 40 41	36 40 41 42 45	24 24 24 25 25
Свыше 10 лет	4	40 40 41 41 42	28 31 35 36 40	20 24 25 31 32

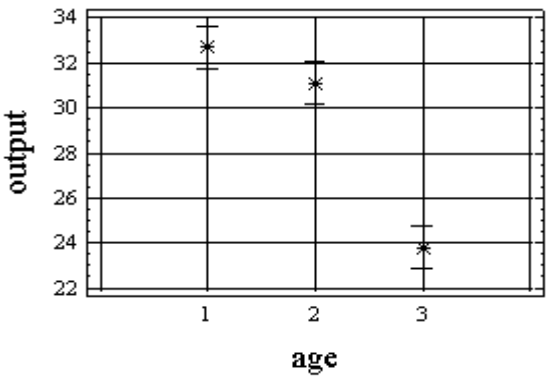
Порядок работы. 1.Открыть файл данных **plant.sf3.** Переменные названы: **output** (выработка), **age** (возраст), **record** (стаж).

2.В главном меню **Compare** выполнить пункт **Analysis of Variance| Multifactor ANOVA…,** в открывшемся ДО внести переменную **output** в поле **Dependent Variable:** (Зависимая переменная) , переменные **age** (возраст), **record** - в поле **Factors:** и нажать **<OK>.**

Нажать кнопку табличных опций**,** установить флажок **ANOVA Table** (Таблица дисперсионного анализа) и нажать **<OK>.** Будет выдана таблица дисперсионного анализа, из которой видно, что оба фактора статистически значимы (p-уровень равен 0,0000). **StatAdvisor** подтверждает вывод о статистической значимости с уровнем доверия 95% (строго говоря, при полученном значении p он близок к 1).

3. Для оценки совместного влияния обоих факторов щелчком правой клавиши (ПК) мыши на поле таблицы вызвать контекстное меню и выбрать пункт **Analysis Options.** Ввести порядок взаимодействия, равный 2, нажать **<OK>.** В таблицу будут добавлены оценки значимости совместного влияния возраста и стажа на производительность труда **(INTERACTIONS AB).** Можно видеть, что ***совместно действующие*** факторы также существенно влияют на производительность труда изучаемой генеральной совокупности работников.

4. Нажать кнопку графических опций и в открывшемся ДО установить флажок **Means Plot** (график средних).



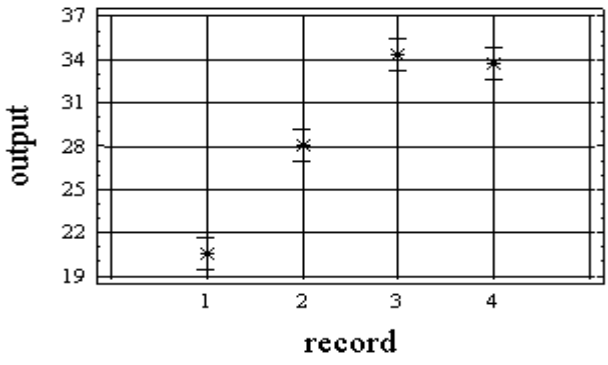


Рисунок 9 – Графики средних значений отклика

а) б)

На рис. 9-а) показаны значения средней производительности труда с очерченными 95% доверительными интервалами для трех обследуемых возрастных групп.

Выполнив команду контекстного меню (оно вызывается щелчком ПК мыши на поле графика) **Pane Options**, можно открыть ДО **Means Plot Options** (Опции графика средних).

В поле **Factor** можно переключиться на зависимость от другого фактора (стажа). Можно видеть, что производительность труда достигает максимума у работников со стажем от 7 до 10 лет, а затем снижается (см. рис.9-б).

5.Причину такой зависимости можно понять, если построить график взаимодействий. Для этого нажать кнопку графических опций и установить флажок **Interactions Plot**. Из полученного графика понятно, что сопутствующее увеличению стажа старение работников, начиная с определенного момента времени, ведет к снижению производительности (рис. 10).

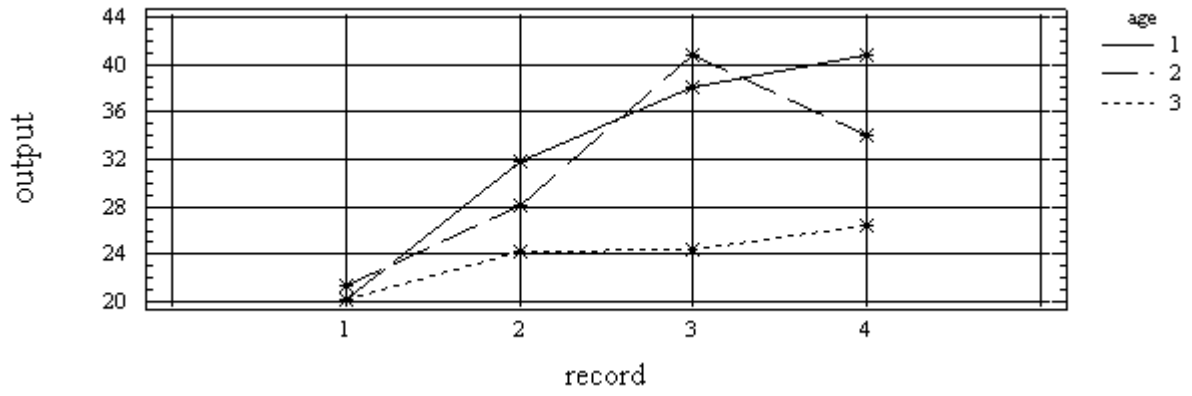


Рисунок 10 - График взаимодействия факторов

Переключиться на зависимость графика от другого фактора можно, выполнив в контекстном меню команду **Pane Options,** которая откроет окно **Interaction Plot Options.** После этого в поле **Plot on Axis** (График в осях) нужно сменить положение переключателя и нажать **<OK>**.

6. Закрыть окно анализа, подтвердив удаление его результатов (OK to delete this analysis?), ответив **Yes.**

4.3.2 Планирование эксперимента

Планирование эксперимента – это раздел математической статистики, изучающий рациональную организацию измерений и наблюдений. Поскольку при этом анализируется влияние ряда факторов на отклик, то в теории планирования экспериментов также важную роль играет дисперсионный анализ.

4.3.2.1 Разработка полного факторного плана.

Задача. Изучить влияние на дальность полета (**Distance**) бумажного самолетика следующих четырех факторов:

- конструкции крыла **Design** (сплошное -**straight** или Т-образное - **high-tee**);
- формы носовой части **Nose** (срезанный - **clip** или нет **- none**);
- типа бумаги **Paper** (ватманская - **construct** или писчая - **notebook**);
- формы крыла в плане **Wing** (прямое - **straight** или стреловидное - **bent**).

Порядок работы.1.Выполнить в основном меню команду **Special | Experimental Design | Create Design**– появится ДО **Create Design Options** для задания начальных параметров плана. Установить переключатель **Design Class** (Тип плана) в положение **Screening** (Просмотр на экране)**,** в поле **No. of Response Variables.** (Количество целевых переменных) установить **1**, в поле **No. of Experimental Factors** (Число экспериментальных факторов) – **4**. В поле **Comment** ввести **Plane design** (Конструкция самолета), нажать **<OK>.**

2.Ввести данные в ДО **Factor Definition Options** (Задание опций факторов)**.**
***(Важно:** Кнопку <OK> следует нажать только после того, как будут заданы значения для всех четырех факторов!)*

Для задания этих значений необходимо, устанавливая переключатель в поле **Factor** поочередно в положения **A,B,C,D**, заполнить поля согласно табл. 5.

Таблица 5 – Характеристики исследуемых факторов

Factor	**Name** (Имя)	**Low** (Нижний уровень)	**High** (Верхний уровень)	Переключатель **Continuous**
A	Design	straight	high-tee	снят
B	Nose	none	clip	снят
C	Paper	notebook	construct	снят
D	Wing	straight	bent	снят

В конце нажать **<OK>.**

Например, после задания значений для первого фактора ДО должно выглядеть следующим образом (рис. 11):

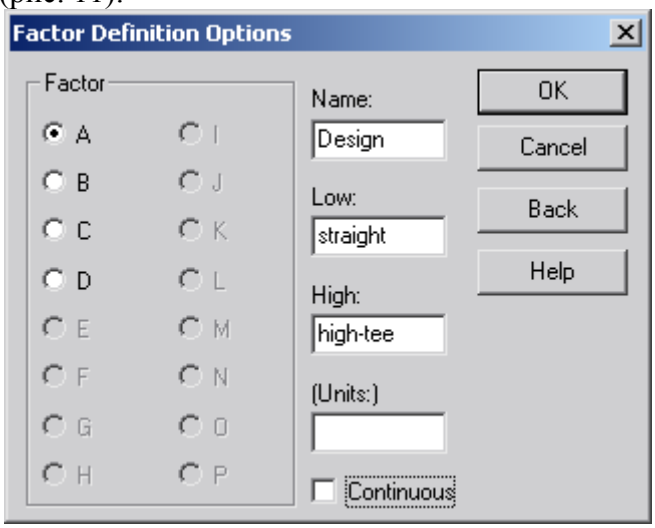


Рисунок 11 – Вид ДО задания характеристик фактора A

***(Замечание:** Если Вы поторопились нажать кнопку <OK>, не закончив определение всех факторов, можно нажать кнопку <Back> и продолжить процедуру).*

3.Заполнить ДО **Response Definition Options** (Задание опций функции отклика). Для этого ввести в поле **Name** название функции отклика **Distance,** в поле **[Units:]** (Единицы измерения) – значение **feet** (футы)**.** Нажать **<OK>.**

4.В ДО **Screening Design Selection** (Выбор типа плана) в поле выбора установить тип плана **Factorial 2^4** (Факторный 2^4), нажать **<OK>.**

5.В ДО **Screening Design Options** (Опции плана) снять флажок **Randomize** (Рандомизировать)**,** нажать **<OK>.** Откроется окно с первичной сводкой плана эксперимента. Командой **File**| **Save As**| **Save Design File As** сохранить план с именем **plane.**

6.Развернуть на весь экран электронную таблицу (из пиктограммы в нижней части окна программы) с именем **plane.sfx** и ввести значения функции отклика согласно табл.6. Сохранить план командой **File**| **Save** | **Save Design File.**

7.Для анализа полученных результатов эксперимента выполнить команду **Special** | **Experimental Design** | **Analyze Design...** В открывшемся ДО **Analyze Design** в левом поле выделить переменную **Distance** и, нажав кнопку со стрелкой, ввести ее в поле **Data** (Данные), нажать **<OK>.**

В окне результатов **Analysis Summary** система **StatAdvisor** указывает, что оценки главных эффектов и двухфакторных взаимодействий расположены в порядке убывания значений и советует использовать для анализа результатов **Парето-карты** из списка гра-

фических опций (третья слева кнопка вверху окна) и таблицу **ANOVA** из списка табличных опций (вторая слева кнопка вверху окна).

Таблица 6 – Исходные данные для анализа плана эксперимента

	Block	**Design**	**Nose**	**Paper**	**Wing**	**Distance**
1	1	straight	none	notebook	straight	6.25
2	1	high-tee	none	notebook	straight	15.5
3	1	straight	clip	notebook	straight	7.00
4	1	high-tee	clip	notebook	straight	16.5
5	1	straight	none	construct	straight	4.75
6	1	high-tee	none	construct	straight	5.50
7	1	straight	clip	construct	straight	4.50
8	1	high-tee	clip	construct	straight	6.00
9	1	straight	none	notebook	bent	7.00
10	1	high-tee	none	notebook	bent	10.00
11	1	straight	clip	notebook	bent	10.00
12	1	high-tee	clip	notebook	bent	16.00
13	1	straight	none	construct	bent	4.50
14	1	high-tee	none	construct	bent	6.00
15	1	straight	clip	construct	bent	4.50
16	1	high-tee	clip	construct	bent	5.75

8.Последовав совету, обнаружим, что статистически значимые эффекты имеют факторы **C** (тип бумаги), **A** (конструкция) и взаимодействие **AC** (их столбцы пересекают белую вертикальную линию, соответствующую тесту на значимость с уровнем доверия 95%):

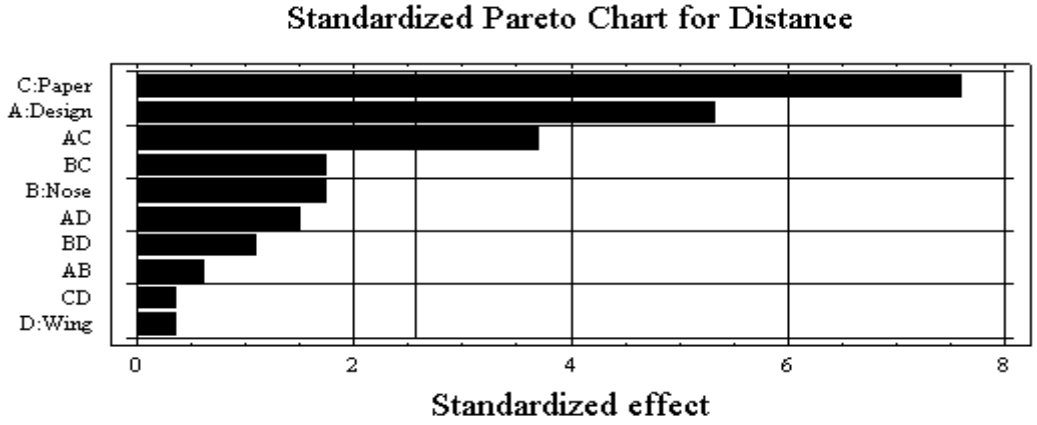


Рисунок 12 – Парето - карта

Из таблицы ANOVA также следует, что именно эти факторы и их взаимодействие имеют уровень значимости менее 0,05 (см. последний столбец таблицы, P-Value). При этом полученная модель объясняет 95,6384% дисперсии функции отклика (`R-squared = 95.6384 percent`).

9.С целью анализа степени влияния различных эффектов на функцию отклика можно также построить графики эффектов для нормального распределения вероятностей. Для этого нажать кнопку графических опций, установить флажок **Normal Probability Plots of Effects** и нажать **<OK>.**

Щелкнув на новом графике ПК, выполнить команду **Pane Options** и установить в ДО **Effects Normal Probability Plot Options** флажок **Label Effects** (Пометить эффекты), нажать **<OK>.** Отклонение параметров **A, C** и их взаимодействия **AC** от линии нормаль-

ного распределения является признаком их более сильного влияния на целевой параметр по сравнению с другими (рис. 13).

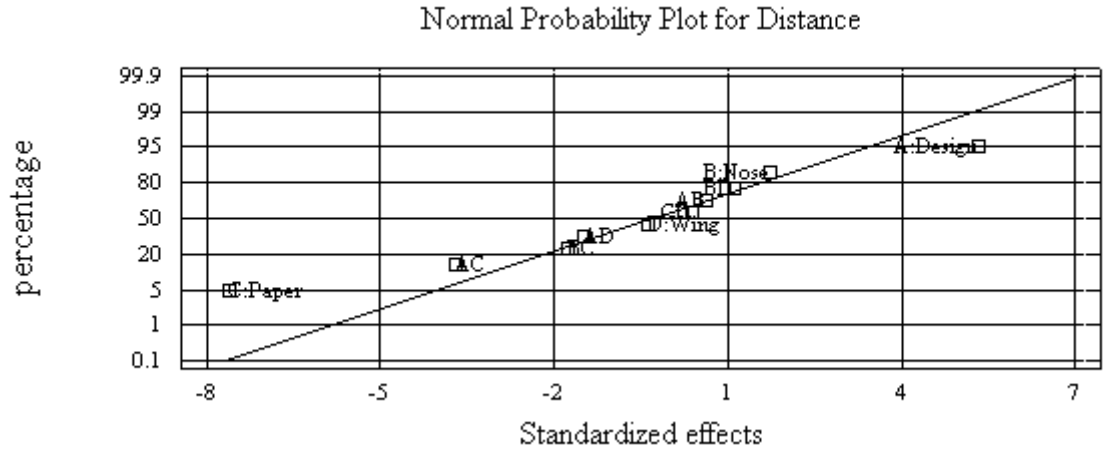


Рисунок 13 – График нормального распределения для значения отклика

10. Построить графики взаимодействий, для чего нажать кнопку графических опций, установить флажок **Interaction Plots** и нажать **<OK>**. Выполнив команду **Pane Options** (после нажатия ПК), в ДО **Interaction Plot Options** снять флажки незначимых факторов **Wing** и **Nose** и нажать **<OK>**. Оставшиеся графики показывают, что наибольшую дальность полета имеют самолетики с Т-образной конструкцией, сделанные из писчей бумаги (рис. 14).

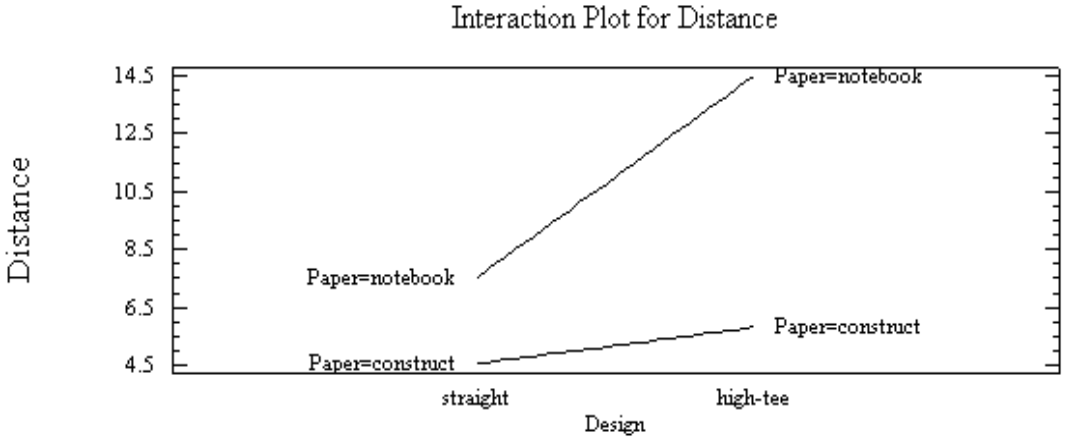


Рисунок 14 – График взаимодействия факторов

11.Закрыть окно анализа, не сохраняя его результатов.

4.3.3.2 Построение и анализ поверхности отклика.

Задача. Исследовать факторы, влияющие на износ выпускаемых пластиковых дисков: состав композиционного материала (содержание наполнителя в связующем) и расположение диска в форме. Предварительные исследования показали, что линейная модель неадекватна экспериментальным данным, поэтому решено использовать двухфакторный центральный композиционный план, состоящий из куба и звезды, ротатабельного типа. Первый фактор **(A)** назовем **Ratio** (Соотношение)**,** второй **(B) – Mold** (Форма)**.** Отклик назовем **Thickness** (Толщина).

Порядок работы. 1. Выполнить в основном меню команду **Special | Experimental Design | Create Design**. На запрос о сохранении текущего плана (Do you want to save the current experiment?) ответить **Нет.** В открывшемся ДО **Create Design Options** в поле **Design Class** (Тип плана) установить переключатель в положение **Response Surface** (Поверхность отклика). В поле **No. of Response Variables** (Количество переменных отклика) ввести **1**, в поле **No. of Experimental Factors** (Количество экспериментальных факторов) – **2**. В поле **Comment** ввести название плана **Disk wear experiment** (Эксперимент по износу дисков). Нажать **<OK>.**

2.Откроется ДО **Factor Definition Options**. Последовательно устанавливая переключатель **Factor** в положения **A** и **B**, ввести значения, указанные в табл. 7.

Таблица 7 – Характеристики исследуемых факторов

Factor	**Name**	**Low**	**High**
A	Ratio	0.75	0.9
B	Mold	0.75	0.9375

Нажать **<OK>**. В новом ДО **Response Definition Options** задать имя функции отклика **Thickness,** нажать **<OK>.**

3.В ДО **Response Surface Design Selection** в поле **Name** выбрать тип плана **Central composite design: 2^2+ star,** нажать **<OK>**.

4.В открывшемся окне опций плана **Composite Design Options** указан выбранный тип, количество необходимых экспериментов (Runs), число степеней свободы для ошибки (Error d.f.). Переключатель **Placement** установить в положение **Last** и снять флажок **Randomize** (рандомизировать), нажать **<OK>**. Будет выдана сводка плана.

5.Сохранить план с именем **disk** (команда главного меню **File| Save As| Save Design File As).**

6. Для определения порядка сбора экспериментального материала удобно получить рабочую таблицу. Для этого нажать кнопку табличных опций, установить флажок **Worksheet** и нажать **<OK>.**

7. После сбора экспериментальных данных необходимо дополнить файл с планом эксперимента значениями функции отклика. Для этого нужно с помощью команды главного меню **Window** открыть окно **disk.sfx** и заполнить последний столбец согласно табл. 8.

8. Выполнить команду **Special | Experimental Design | Analyze Design…** Ввести в поле **Data** переменную **Thickness** и нажать **<OK>.**

Таблица 8– Исходные данные для анализа плана эксперимента

disk.sfx

	BLOCK	Ratio	Mold	Thickness
1	1	0,75	0,75	7,3
2	1	0,9	0,75	7
3	1	0,75	0,9375	7,1
4	1	0,9	0,9375	8
5	1	0,718934	0,84375	7,6
6	1	0,931066	0,84375	7,4
7	1	0,825	0,711167	7,4
8	1	0,825	0,976333	7,9
9	1	0,825	0,84375	8,2
10	1	0,825	0,84375	8,3

9.Для определения адекватности полученной модели следует проанализировать таблицу дисперсионного анализа **ANOVA Table**, которая, после установки соответст-

вующего флажка в окне табличных опций и нажатия кнопки **<OK>,** выводится на экран под заголовком `Analysis of Variance for Thickness - Disk Wear Experiment`

Из нее следует, что статистически значимые эффекты (p<0,05) имеют два квадратичных члена: **AA** и **BB**.

10.Щелкнув ПК на таблице **ANOVA**, выполнить команду **Pane Options** и в ДО **Analysis of Variance Options** установить флажок **Include Lack-of-Fit Test** (Включить тест на неадекватность модели) и нажать **<OK>**. В таблицу будет добавлена следующая строка:

```
Lack-of-fit                0,199757      3      0,0665858      13,32      0,1955
```

Последнее значение – это уровень значимости для ***гипотезы о неадекватности модели.*** Так как этот уровень превышает 0,05, то гипотезу следует отвергнуть, то есть ***можно признать модель адекватной.***

11.Нажав кнопку графических опций, установить флажок **Normal Probability Plots of Effects** (Графики эффектов нормального распределения вероятностей) и нажать **<OK>.** Нажав ПК мыши, выполнить команду **Pane Options** и в открывшемся ДО установить флажок **Label Effects**. Значительное отклонение точек, соответствующих квадратичным членам **AA** и **BB,** от линии нормального распределения – признак статистически значимого влияния их на отклик.

12. Нажав кнопку графических опций, установить два флажка **Response Plots** (Графики отклика); нажать **<OK>.** Трехмерная диаграмма (рис. 15) позволяет судить о наличии максимума функции отклика:

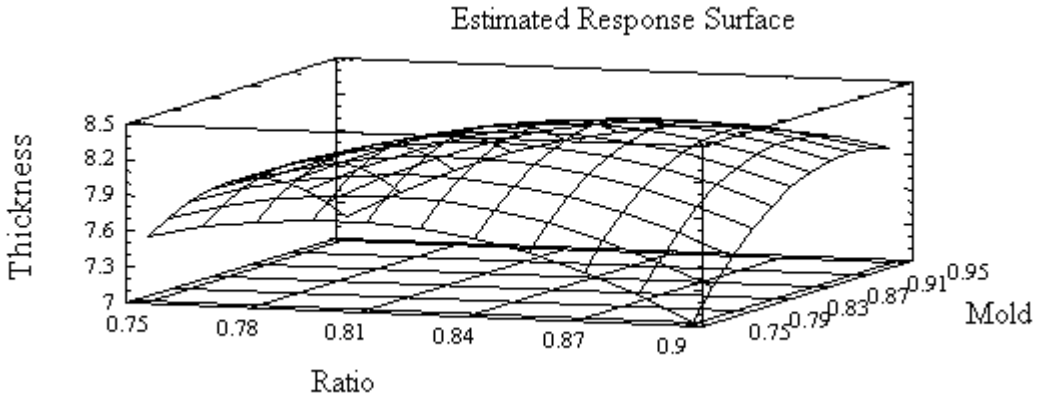


Рисунок 15 – Поверхность отклика

Если увеличить диаграмму до максимального размера (дважды щелкнув на ней), на расположенной выше панели инструментов активизируются две кнопки со стрелками, нажатие на которые вызывает вращение диаграммы в горизонтальной и вертикальной плоскостях. Для остановки вращения нужно щелкнуть на диаграмме.

Вторая, плоская, диаграмма – это линии равного уровня функции отклика, получаемые в результате сечения трехмерной поверхности плоскостями, параллельными плоскости факторов. Эта диаграмма позволяет более точно определить положение максимума.

13. Вызвав командой **Pane Options** ДО **Response Plot Options,** в поле **Contours** (Контуры) установить переключатель в положение **Painted Regions** (Окрашенные зоны) и нажать **<OK>.** Можно видеть, что область максимума соответствует значениям фактора **Ratio** в диапазоне 0,81...0,87 и значениям фактора **Mold** в диапазоне 0,84...0,91:

14. В завершение анализа можно получить уравнение регрессии функции отклика по экспериментальным факторам. Для этого нужно нажать кнопку табличных опций и установить флажок **Regression Coefficients. StatAdvisor** выдаст уравнение регрессии:

```
Thickness = -45.2189 + 91.0287*Ratio + 35.209*Mold - 76.6668*Ratio^2 +
42.6667*Ratio*Mold - 40.533*Mold^2
```

15. Закрыть окно анализа.

4.3.3.3 Разработка смесевого плана

Задача. Проанализировать влияние пропорций ингредиентов ракетного топлива на его эластичность. Определить состав топлива, при котором его эластичность оценивается показателем 3000.

Порядок работы. 1.Выполнить в основном меню команду **Special | Experimental Design | Create Design**– появится ДО **Create Design Options** для задания начальных параметров плана. Установить переключатель **Design Class** (Тип плана) в положение **Mixture** (Смесь). Задать число компонентов равным 3, число откликов – 1. В поле комментариев ввести запись **Rocket propellant study** (Изучение ракетного топлива). Нажать **<OK>.**

2.Откроется ДО **Component Definition Options**. Последовательно устанавливая переключатель **Component** в положения **A, B**, **C,** ввести значения, указанные в табл. 9.

Таблица 9 – Характеристики исследуемых факторов

Component	**Name**	**Low**	**High**
A	Binder	0.2	0.4
B	Oxidizer	0.4	0.6
C	Fuel	0.2	0.4

Примечание. Binder- связующее, oxidizer-окислитель, fuel – горючее.

Нажать **<OK>**.

3.В новом ДО **Response Definition Options** задать имя функции отклика **Elasticity,** нажать **<OK>.**

4.В окне **Mixture Design Selection** (Выбор смесевого плана) выбрать – **Simplex-Centroid** (Симплексный центроидный план), нажать **<OK>.**

В новом ДО задать тип модели **Special Cubic** (Специальная кубическая), установить флажок **Augment Design** (Расширенный план) и снять флажок **Randomize** (рандомизировать) (рис. 16).

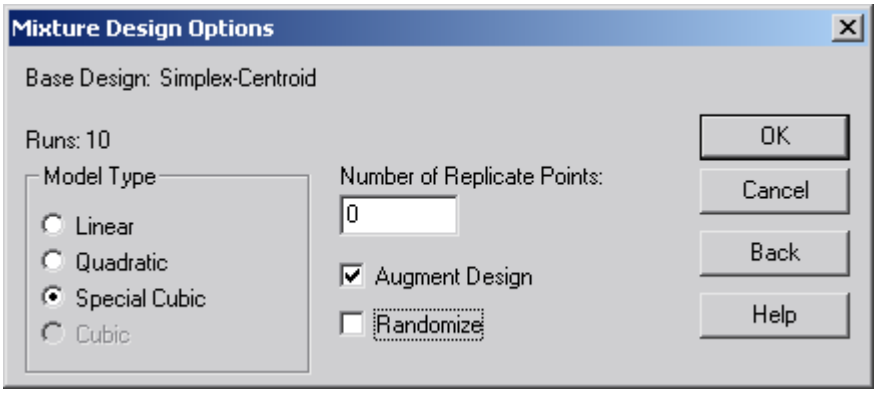


Рисунок 16 – ДО опций плана

Нажать **<OK>.** Будет выдана сводка выбранного плана. В ней, в частности, указывается, что общее число экспериментов по данному плану должно составлять 10.

5.Сохранить план с именем **rocket** (команда главного меню **File| Save As| Save Design File As).**

6.Раскрыть таблицу с названием **rocket.sfx** и ввести в нее экспериментальные значения функции отклика (табл. 10). Сохранить таблицу (**File| Save | Save Design File**).

7.Выполнить команду **Special | Experimental Design | Analyze Design.** Ввести в поле **Data** переменную **Elasticity** и нажать **<OK>.** В полученной сводке результатов анализа статистически значимые эффекты наблюдаются у квадратической и специальной кубической моделей (их p-уровни меньше 0,05). Наименьшее значение ошибки **(SE=10,2573)** присуще специальной кубической модели, она же объясняет наибольшее значение дис-

персии функции отклика **(R-squared=99,94, Adj. R-squared=99,82)**. Поэтому для дальнейшего анализа целесообразно выбрать именно эту модель.

8.Нажав кнопку табличных опций, установить флажки **Model Results** (Результаты модели), **ANOVA Table** (Таблица дисперсионного анализа) и нажать **<OK>.** Из полученных таблиц можно видеть, что все члены модели, кроме взаимодействия **AB,** имеют значимые эффекты, а модель в целом очень хорошо объясняет полученные результаты (p-уровень для нее составляет всего 0,0001).

Таблица 10 – Исходные данные для анализа плана

	BLOCK	Binder	Oxidezer	Fuel	Elasticity
1	1	0.4	0.4	0.2	2350
2	1	0.2	0.6	0.2	2450
3	1	0.2	0.4	0.4	2650
4	1	0.3	0.5	0.2	2400
5	1	0.3	0.4	0.3	2750
6	1	0.2	0.5	0.3	2950
7	1	0.266667	0.466667	0.266667	3000
8	1	0.333333	0.433333	0.233333	2690
9	1	0.233333	0.533333	0.233333	2770
10	1	0.233333	0.433333	0.333333	2980

9.Нажать кнопку графических опций, установить два флажка **Response Plots** и нажать кнопку **<OK>.** На полученном трехмерном графике (рис. 17) хорошо видна зона максимума функции отклика.

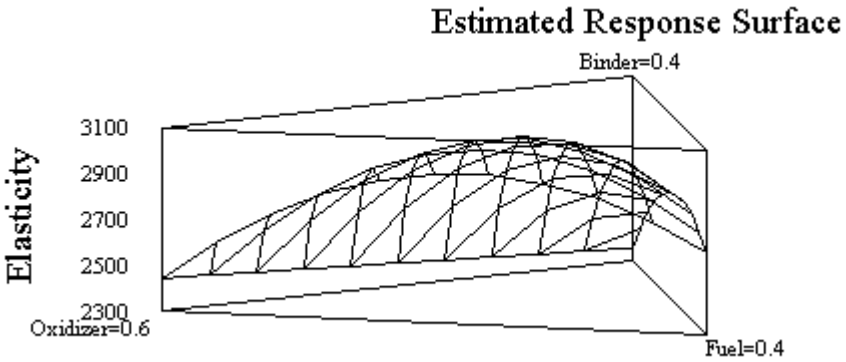


Рисунок 17 – Поверхность отклика

Для более точного определения положения максимума следует рассмотреть второй график (рис. 18). Вызвав командой **Pane Options** ДО **Mixture Response Plot Options,** в поле **Contours** установить переключатель в положение **Painted Regions** и нажать **<OK>.**

Можно видеть, что значение эластичности 3000 лежит вблизи доли связующего (Binder) 0,25, окислителя (Oxidizer) 0,45 и топлива (Fuel) 0,3:

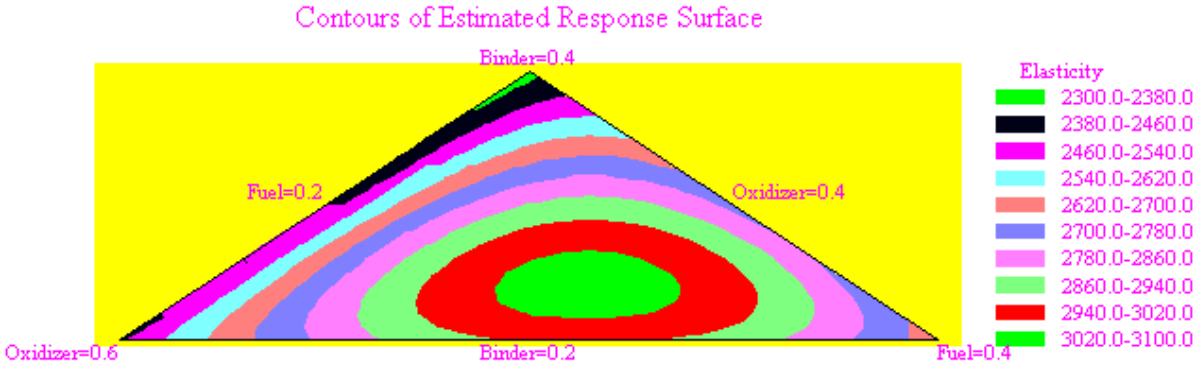


Рисунок 18 – Линии уровня функции отклика

10. В комментариях к таблице результатов (озаглавленной Special Cubic Model Fitting Results for elasticity) статистический советчик **StatAdvisor** приводит уравнение регрессии:

Elasticity = 2351.17*Binder + 2445.71*Oxidizer + 2652.98*Fuel -
6.24733*Binder*Oxidizer + 1008.3*Binder*Fuel + 1597.39*Oxidizer*Fuel +
6141.1*Binder*Oxidizer*Fuel

Чтобы использовать полученное уравнение в дальнейшем, необходимо оценить точность построенной модели. Для этого, нажав кнопку табличных опций и установив флажок **Predictions** (Предсказания), нужно нажать **<OK>.**

Из полученной таблицы следует, что значения показателя эластичности топлива, предсказанные моделью (Fitted Value), для каждого из 10 проведенных экспериментов весьма близки к реально наблюдавшимся (Observed Value). Доверительный интервал для среднего значения показателя эластичности накрывает эти наблюдавшиеся значения с вероятностью 0,95:

Lover 95% CL for Mean < Observed Value < Upper 95% CL for Mean

(наблюдаемое значение лежит между нижней и верхней границами 95% доверительного интервала).

11.Завершить работу программы.

Контрольные вопросы

1.В чем заключается задача дисперсионного анализа?

2.Какая гипотеза проверяется в ходе дисперсионного анализа?

3.Какая величина, полученная в результате дисперсионного анализа, позволяет заключить, значим ли исследуемый фактор?

4.Как в ПСК Statgraphics получить график взаимодействия факторов?

5.Какие выводы можно сделать, используя график взаимодействия?

6.Как, используя Парето-карту результатов анализа плана эксперимента, заключить, какие факторы значимо влияют на отклик?

7.Для чего используется график нормального распределения значения отклика?

8.В каких случаях следует использовать тип плана, который в ПСК Statgraphics именуется **Response Surface** (Поверхность отклика)?

9.В каких случаях следует использовать тип плана, который в ПСК Statgraphics именуется **Mixture** (Смесь)?

10.Как можно повернуть в пространстве изображение трехмерной диаграммы, построенной ПСК Statgraphics?

4.4 Лабораторная работа № 4. Выполнение статистического анализа в программном статистическом комплексе STATISTICA

Цель работы – получение начальных навыков работы с программным статистическим комплексом Statistica.

4.4.1 Создание рабочей книги и таблицы данных

Для создания нового файла необходимо выполнить команду **File** / **New**; открывается диалоговое окно (ДО) **Create New Document** с закладками:

- Spreadsheet (таблица данных);

- Report (отчет);

-Macro (SVB) Program (макропрограмма на встроенном в ПСК языке программирования **STATISTICA Visual Basic**);

- Workbook (рабочая книга).

Таблица данных и отчет о результатах анализа могут быть созданы как в составе рабочей книги (своего рода папки с документами), так и в виде отдельного окна. Для выбора одного из этих вариантов в ДО **Create New Document** имеется переключатель с соответствующими положениями **In a new Workbook** и **As a stand-alone Window.** При создании таблицы данных в полях выбора **Number of variables:** и **Number of cases:** можно задать необходимое количество переменных и наблюдений соответственно.

Для изменения числа переменных в уже существующей таблице удобнее всего воспользоваться контекстным меню, для чего щелкнуть на одном из заголовков ее столбцов правой клавишей мыши и в открывшемся контекстном меню выполнить:

- для удаления переменных – команду **Delete Variables…** и в открывшемся ДО в полях **From variable:** (От переменной:) и **To variable:** (До переменной:) указать границы диапазона удаляемых переменных;

- для добавления переменных – команду **Add Variables…** и в открывшемся ДО заполнить поля **How many:** (Сколько) и **After:** (После) – произойдет вставка заданного количества переменных после указанной.

В контекстном меню имеются и другие команды, в том числе перемещения переменных **- Move Variables…** и копирования **- Copy Variables…**

Аналогичные действия можно выполнять со значениями переменной, то есть результатами отдельных наблюдений (**Cases**), для чего контекстное меню вызывается щелчком правой клавиши мыши на заголовке одной из строк таблицы.

Под заголовком созданной таблицы имеется белое поле, в котором, после активизации его двойным щелчком мыши, можно записать заголовок таблицы (заголовок не следует путать с именем таблицы - имя задается при ее сохранении!).

Для изменения имени и других реквизитов переменной нужно дважды щелкнуть на заголовке соответствующего столбца и в открывшемся ДО **Variable n** (где **n** – номер переменной) задать необходимую информацию. В расположенном в нижней части ДО белом поле **Long name (label or formula with Functions):** (Длинные имена (метка или формула с функциями)) с помощью знаков операций и стандартных функций можно задавать формулы для расчета значения переменных. Список доступных функций открывается при нажатии на расположенную над полем кнопку **Functions.** В качестве аргументов функций и операндов могут использоваться числа и имена переменных, содержащихся в таблице (вместо имени переменной можно использовать букву **V (variable)** с номером соответствующего столбца), например: ввод в поле **Long name** выражения **= v1+v2** приведет к тому, что столбец соответствующей переменной будет заполнен суммой значений переменных **Var1** и **Var2.**

Упражнение 1. Создать таблицу для двух переменных, каждая из которых примет по 25 значений. Назвать таблицу «Корреляция», а переменные – X и Y. Добавить в таблицу переменную Z.

Для многих статистических процедур используются случайные числа, равномерно распределенные на отрезке [0; 1]. Для заполнения такими числами столбца таблицы необходимо выделить соответствующий столбец и в контекстном меню выполнить команду **Fill / Standardize Block ► Fill Random Values.**

Упражнение 2. Заполнить первый и второй столбцы случайными числами. В третьем столбце вычислить сумму переменных X и Y. На сообщение системы **Expression OK,** которое появится при вводе в поле **Long name** переменной Z корректного выражения, ответить «Да».

4.4.2 Порядок проведения статистического анализа

После заполнения таблицы данных или загрузки ее из файла в меню **Statistics** выбирается необходимый метод анализа. После этого могут последовательно открыться еще одно или более диалоговых окна, в которых пользователь должен указать одну из разновидностей выбранного метода. Окончательный выбор подтверждается нажатием кнопки **OK**.

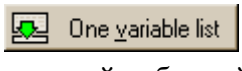
Во вновь открывшемся диалоговом окне необходимо нажать кнопку 

(Переменные). В ДО **Select the variables for analysis** (Выбрать переменные для анализа) выбор переменной подтверждается нажатием кнопки **OK**. Как правило, в этом же окне имеется кнопка 

, при нажатии на которую выдается сводка результатов проведенного анализа, само же диалоговое окно, управляющее анализом, автоматически сворачивается в пиктограмму, располагающуюся в нижней части рабочего окна. Для продолжения анализа эту пиктограмму нужно активизировать щелчком левой клавиши мыши. В диалоговом окне могут предлагаться различные дополнительные опции, для реализации которых необходимо включить переключатель (установить «флажок») у соответствующей надписи. Возврат к предыдущему диалоговому окну (что часто необходимо для изменения установленных ранее опций) происходит после нажатия кнопки **Cancel.**

Упражнение 3.

Вычислить показательные статистики для сгенерированных переменных X, Y, Z. Для этого в меню **Statistics** выбрать метод **Basic Statistics / Tables** (Базовые статистики / Таблицы), а затем его разновидность **Descriptive Statistics** (Описательные статистики). После получения сводки анализа продолжить его выполнение, запросив расчет ошибки выборочного среднего (**Std. err. of mean**) и доверительного интервала (**Conf. limits for means**). Для этого нужно перейти в диалоговом окне **Descriptive Statistics...** на закладку **Advanced** и активизировать переключатели у соответствующих надписей.

Вычислить коэффициенты корреляции между переменными X, Y, Z. Для этого в ДО **Descriptive Statistics...** (развернув его из пиктограммы) нажать кнопку **Cancel** и в ДО **Basic Statistics and Tables...** выбрать его пункт **Correlation matrices**. В открывшемся ДО нажать кнопку 

(Переменные в одном списке) и выбрать все три переменные. Над выданной сводной таблицей результатов можно видеть комментарий: **Marked correlations are significant at p<0.05000** (Выделенные корреляции значимы с p<0.05000).

4.4.3 Построение типовых графиков

Наиболее часто встречающиеся простейшие графики строятся с помощью меню **Graphs**. После выбора нужного типа графика в окне выбора переменных необходимо задать переменные, подлежащие анализу.

В число указанных графиков входят:

- Гистограммы (**Histograms...**). В поле **Intervals**, на закладке **Quick**, установив переключатель в положение **Categories:**, можно задать количество интервалов для построения гистограммы.

- Диаграммы рассеяния (**Scatterplots...**). Для их построения нужно задать по одной переменной в левой и правой части окна выбора. Если в поле **Regression bands** предварительно установить переключатель в положение **Confidence,** на графике будут показаны границы доверительного интервала; значение доверительной вероятности задается с помощью поля выбора, расположенного рядом с переключателем.

- Трехмерные графики (**Surface Plots…**). Требуют задания трех переменных в различных частях поля выбора.

- Трехмерные последовательные диаграммы (**3D Sequential Graphs**). Разновидность графика **Raw Data Plots...** (Графики исходных данных) позволяет изобразить на одном графике трехмерные гистограммы для любого числа переменных, выделенных в окне выбора.

- «Матричные графики» (**Matrix Plots...**). Позволяют отобразить связь между выбранными парами переменных в виде диаграмм рассеяния и соответствующих гистограмм.

Упражнение 4. Построить перечисленные типы графиков для переменных X, Y, Z.

Заметим, что по мере получения результатов анализа все они будут включаться в рабочую книгу (рис. 19). В левой части окна рабочей книги, получившей по умолчанию имя **Workbook1*** , находится *дерево (tree),* отображающее ее структуру. Щелкнув на пиктограмме того или иного объекта в дереве, получим его изображение в правой части окна.

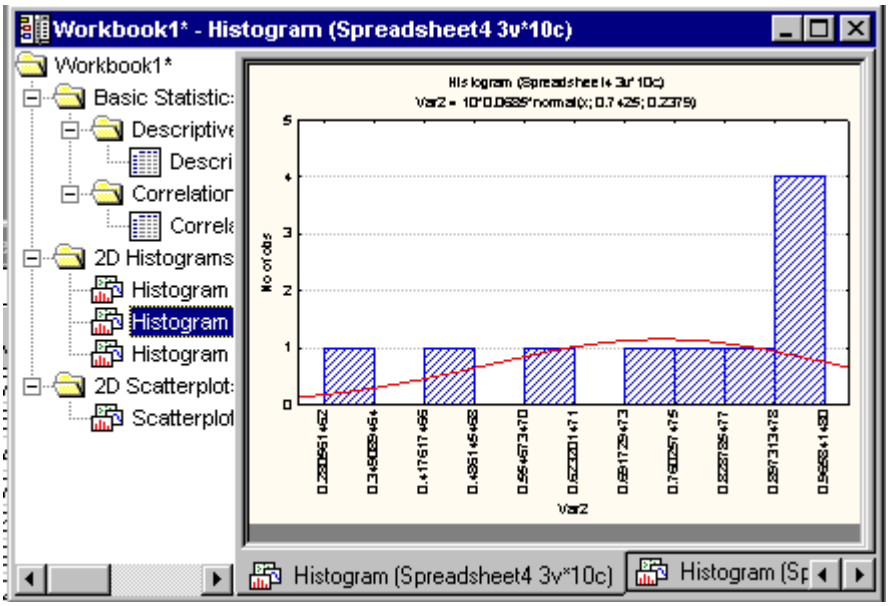


Рисунок 19 – Окно рабочей книги

Другой способ просмотреть необходимый объект заключается в использовании закладок в нижней части рабочей книги.

4.4.4 Вероятностный калькулятор

Этот модуль ПСК **Statistica** позволяет решать многие статистические задачи, заменяя собой различные таблицы распределений вероятностей. Чтобы воспользоваться калькулятором, нужно в меню **Statistics** выбрать метод **Basic Statistics / Tables** и далее **Probability Calculator**. Открывается диалоговое окно, изображенное на рис. 20.

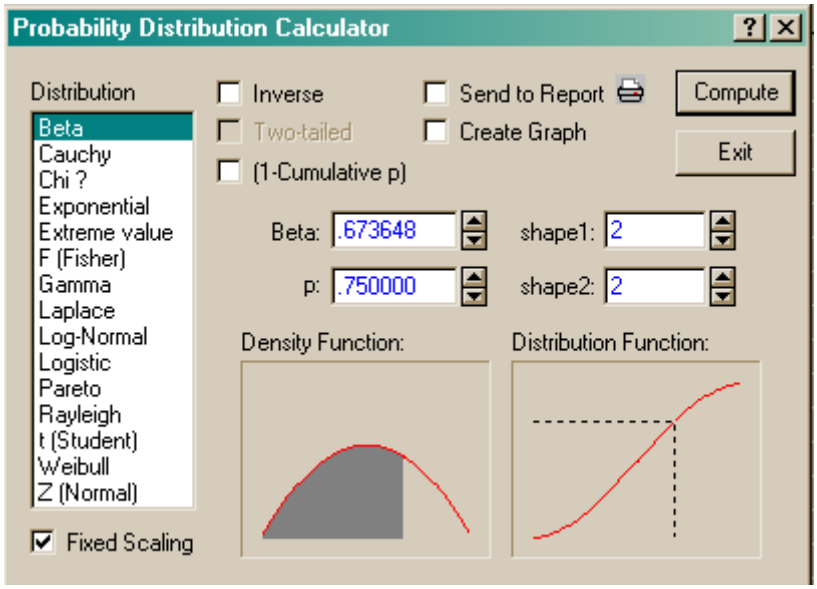


Рисунок 20 – Диалоговое окно вероятностного калькулятора

В левой части окна имеется поле **Distribution** со списком распределений. Правее находится общее для всех типов распределений поле, обозначенное **p:** (вероятность). Над ним расположено поле для аргумента функции распределения, название которого совпадает с обозначением выбранного распределения. При вводе в поле **p:** значения вероятности и нажатии кнопки **Compute** в поле аргумента появляется значение квантиля ***x*,** т.е. корня уравнения

$$F(x) = p,$$

где $F(x) = P\{X \le x\}$ - функция распределения, равная вероятности того, что случайная величина X, имеющая выбранный закон распределения, не превысит неслучайного значения ***x***.

Если, напротив, задать значение ***x***, то калькулятор вычислит соответствующее ему значение **p.**

В расположенных ниже полях изображаются графики плотности распределения (**Density Function:**) и функции распределения (**Distribution Function:**).

При активизации опции **Create Graph** на экран выводятся графики плотности распределения и функции распределения, а опции **Send to Report** – график включается в отчет о результатах анализа.

Содержание других опций, указанных в ДО, как правило, зависит от вида распределения, в частности:

- **Two-tailed** – двухстороннее (для симметричных распределений);
- **mean:** - значение среднего;
- **st. dev.:** - значение стандартного отклонения;
- **shape:** - коэффициент формы;
- **df.:** - число степеней свободы.

Упражнение 5. Значение диаметра вала распределено по нормальному закону. В партии деталей среднее значение диметра равно 151 мм, стандартное отклонение – 7 мм. Вычислить

вероятность того, что диаметр случайно выбранной детали отклонится от среднего значения не более чем на 5 мм.

Указание. Искомая вероятность равна $F(x_1) - F(x_2)$, *где* x_1 *– значение верхнего допуска размера,* x_2 *– значение нижнего допуска.*

После выполнения заданий закрыть окно рабочей книги, не сохраняя его. Закрыть таблицу данных; при этом программный комплекс выдает предупреждение (рис. 21).

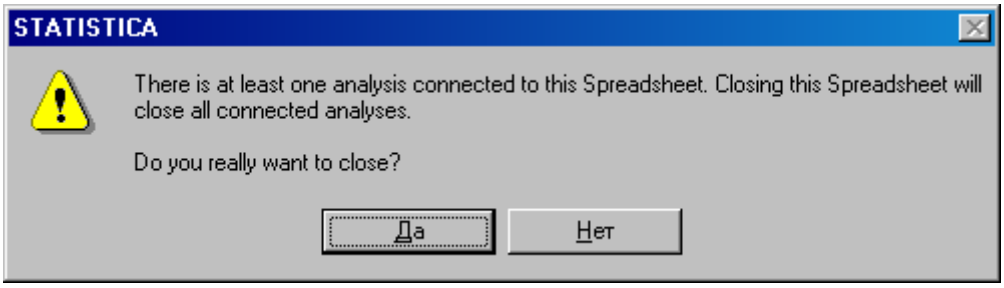


Рисунок 21 – Предупреждение о попытке закрыть таблицу данных

Оно означает, что данные, содержащиеся в таблице, используются в анализе, и ее закрытие приведет также к выходу из выполняемого анализа.

Ответив **Да**, подтвердить свое намерение закрыть таблицу.

Контрольные вопросы.

1.Какие объекты может содержать таблица данных программного статистического комплекса **STATISTICA**?

2.Что означают термины **Variables** и **Cases** и каким измерениям таблицы данных они соответствуют?

3.Как изменить размеры таблицы данных?

4.Каким образом определить реквизиты переменной?

5.Как можно получить выборку случайных чисел с равномерным законом распределения?

6.Изложите общий порядок выполнения статистического анализа.

7. Как вычислить показательные статистики для заданных переменных?

8.Как ПСК **STATISTICA** указывает, что корреляционная зависимость между переменными является статистически значимой?

9.Какая группа команд головного меню позволяет построить типовые графики для переменных, содержащихся в таблице данных?

10.Расскажите о возможностях «Вероятностного калькулятора». Как им воспользоваться?

4.5 Лабораторная работа №5
Анализ временных рядов

Цель работы:

1) научиться анализировать взаимную зависимость временных рядов;
2) ознакомиться с основными видами преобразований временных рядов и методами выделения регулярных составляющих;
3) ознакомиться с процедурой прогнозирования значений временного ряда.

4.5.1 Анализ распределенных лагов

Данный метод анализа временных рядов позволяет исследовать взаимосвязь между значениями временных рядов, сдвинутых относительно друг друга на определенный интервал – *лаг.*

Задача. Файл ***Teachers.sta*** содержит следующие данные за период с 1900 по 1980 г.г. с 10-летним интервалом: 1)количество учащихся средних школ (***Children***), 2)количество учителей (***Teachers***), 3)средняя зарплата школьного учителя (***Salary***). Резонно предположить, что численность учителей зависит от количества учеников, но изменение их числа происходит с некоторым запаздыванием. То же касается и зарплаты учителей. Необходимо исследовать зависимость между указанными временными рядами.

Порядок работы.1.Открыть файл ***Teachers.sta*** из папки ***Examples\Datasets***. *В* главном меню выполнить команду **Statistics/ Advanced Linear/Nonlinear Models ►Time Series/Forecasting**. В открывшемся ДО **Time Series Analysis: Teachers** нажать кнопку **Variables** , в открывшемся ДО выбора переменных нажать кнопку **Select All** (Выбрать все), нажать **OK**.

2.На закладке **Quick** Нажать кнопку **Distributed lags analysis** (Анализ распределенных лагов) в правой нижней части окна - откроется ДО с этим же именем. В качестве зависимой нужно выбрать переменную ***Teachers***, для чего выделить ее в окне. В качестве независимой выбрать переменную ***Children,*** для чего нажать кнопку **Independent variable** и в открывшемся ДО **Currently available variables and transformations** (Доступные текущие переменные и преобразования) выделить указанную переменную и нажать **OK**.

3.Установить в поле **Lag length** (Размер лага) значение **2**, чтобы анализировать 10- и 20-летние интервалы. Нажать кнопку **Summary: Distributed lags analysis** в верхней правой части окна. Результаты представлены в двух таблицах, переключение между которыми выполняется щелчком на соответствующей пиктограмме «дерева» (рис.22).

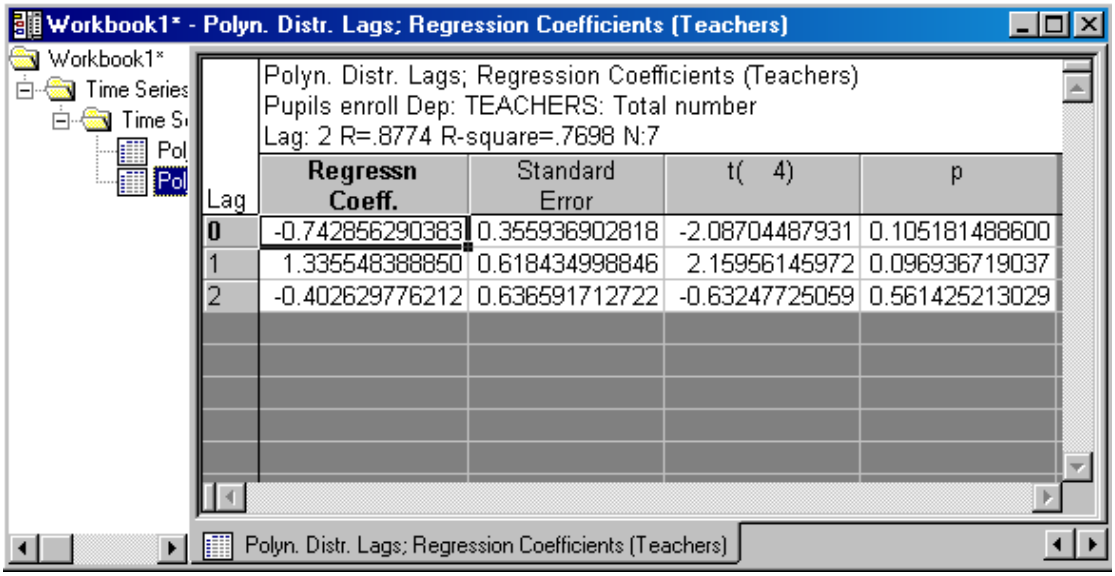


Рисунок 22 - Результаты анализа зависимости рядов

В заголовке окна ***Polyn. Distr. Lags; Regression Coefficients (Teachers)*** отмечается сильная корреляция между исследуемыми переменными **(R=0,8774)**. О существенном влиянии количества учеников на количество учителей с запаздыванием в 10 лет говорит приведенная в последнем столбце таблицы величина p-уровня для ***Lag = 1*** (она <0,1). Это значит, что следует отвергнуть гипотезу о равенстве 0 коэффициента регрессии переменной ***Teachers*** по переменной ***Pupils*** при указанном временном сдвиге.

Задание. Выполнить анализ зависимости зарплаты учителей ***(Salary)*** *от числа учеников* ***(Children).***

Закрыть окна **Workbook** и **Data.**

4.5.2 Визуализация и преобразования временного ряда

Задача. Провести анализ изменения складских запасов двух товаров за период с февраля по ноябрь 1991 г.

Порядок работы. 1. Задание переменных для анализа. Открыть файл ***Stocks.sta,*** находящийся в папке ***Examples\ Datasets.*** В главном меню выполнить команду **Statistics/ Advanced Linear/Nonlinear Models ►Time Series/Forecasting.** В открывшемся ДО **Time Series Analysis:Stocks.sta** нажать кнопку Variables, в ДО выбора переменных выбрать ***Stock1, Stock2*** (удерживая нажатой клавишу **Ctrl**) и нажать кнопку **OK.** В новом ДО **Time series analysis: Stocks** нажать кнопку **OK (transformations, autocorrelations, crosscorrelations, plots) –** то есть "Трансформации, корреляции, графики". Откроется ДО преобразования переменных **Transformations of Variables: Stocks** (рис. 23).

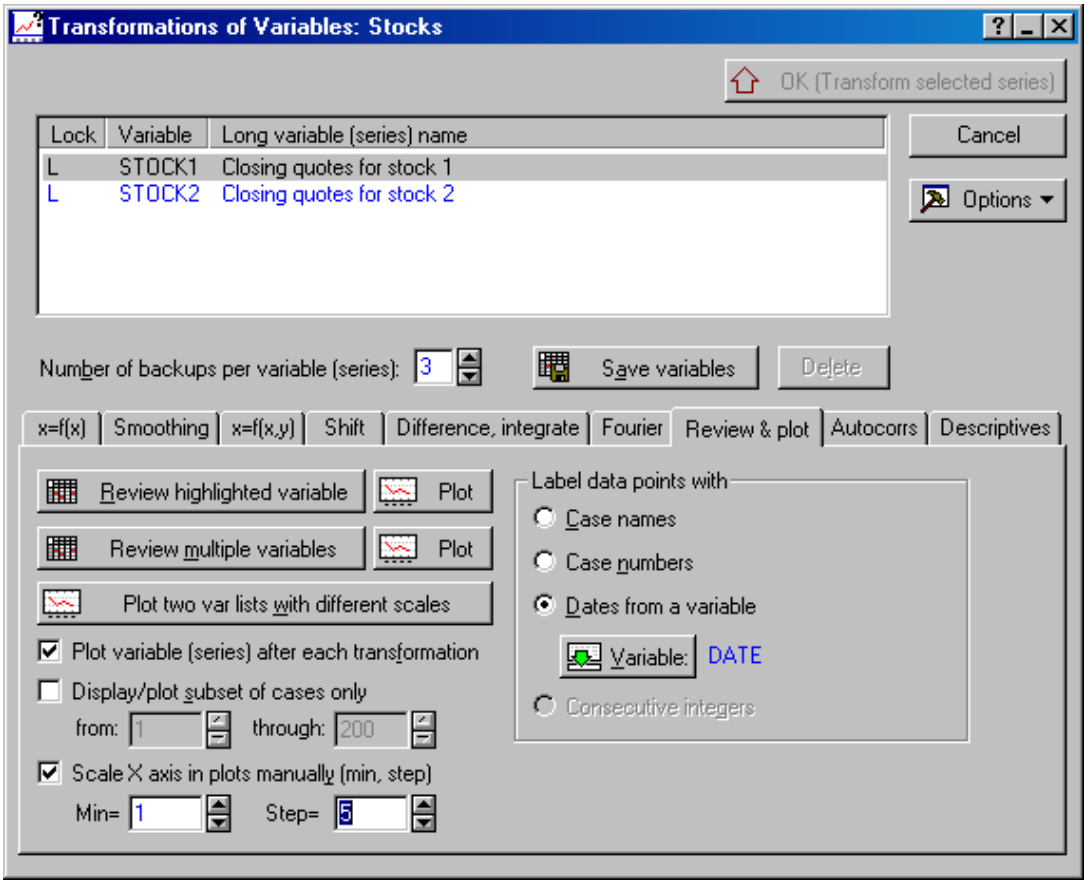


Рисунок 23 – Окно преобразования переменных

2.Построение графика временного ряда. Щелкнуть на закладке **Review & plot** (Просмотр и график)**. В** поле **Label data points with** (Маркировать точки), расположенном в правой нижней части окна, установить опцию **Dates from a variable -** это позволит маркировать ось абсцисс датами, содержащимися в переменной ***Date***. В открывшемся ДО **Select the variable with dates** выбрать переменную ***Date***, затем нажать **OK**.

Активизировать переключатель **Scale X axis in plots manually…** (Ручное шкалирование оси X), расположенный в левой нижней части окна, и ввести значения **Min=1** (начать анализ с первого дня) и **Step=5** (величина шага равна 5 рабочим дням в неделю). В окне задания переменных (в светлой части ДО) выделить переменную ***Stock1*** и нажать кнопку **Review highlighted variable** (Просмотр значений выделенной переменной) – это первая сверху кнопка под рядом закладок.

После указанных действий окно **Transformations of Variables: Stocks** будет, вероятно, свернуто программой в пиктограмму, которая расположится в нижней части рабочего поля (это зависит от настройки ПСК). Для продолжения работы необходимо развернуть окно щелчком на этой пиктограмме.

Для построения графика временного ряда нужно нажать кнопку **Plot** (График) рядом с кнопкой **Review highlighted variable.**Чтобы сравнить временные зависимости для обеих переменных, нужно нажать кнопку **Plot** рядом с кнопкой **Review multiple variables** (Просмотр значений многих переменных) – она расположена под кнопкой **Review highlighted variable.** В открывшемся ДО выделить переменные ***Stock1*** и ***Stock2*** (удерживая нажатой клавишу **Ctrl**)и нажать кнопку **OK** - будет получен график (ри24)

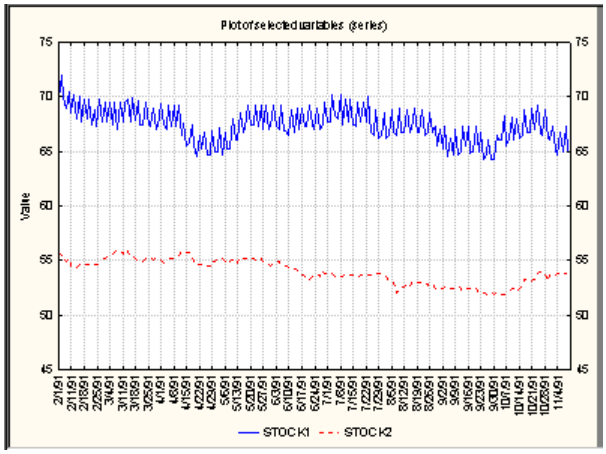


Рисунок 24 – Изображения двух рядов

3. Сглаживание временного ряда по методу скользящего среднего. Выделить в окне **Transformations of Variables: Stocks** (возможно, его снова придется развернуть из пиктограммы) переменную ***Stock1,*** на закладке **Smoothing** (Сглаживание) установить переключатель **N-pts. mov. averg.** (Сколь-зящее среднее по N точкам) и ввести значение ***N*=5**, после чего нажать кнопку **OK (transform selected series)** (Преобразовать выбранные ряды). Можно видеть, что полученный ряд является более «плавным», и тренд прослеживается более отчетливо (рис.25).

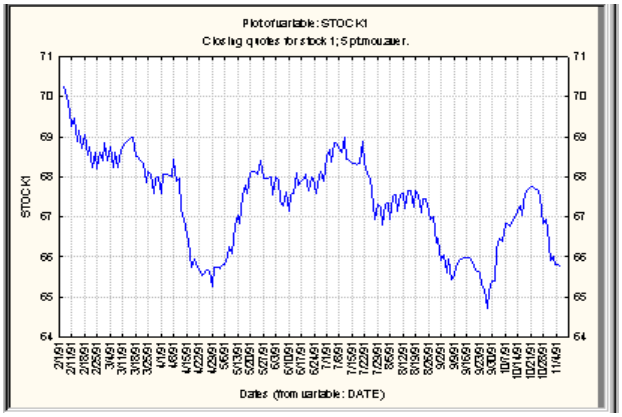


Рисунок 25 – Сглаженный ряд

Задание. Провести сглаживание по 10 точкам; сравнить результаты и сделать вывод.

4. Анализ автокорреляционной функции (АКФ). В ДО **Transformations of Variables: Stocks** выделить переменную ***Stock2*,** перейти на закладку **Review & plot** и нажать кнопку **Plot** (рядом с кнопкой **Review highlighted variable***). Из графика видна тенденция переменной к убыванию, т.е. тренд. Наличие тренда влияет на АКФ: действительно, можно ожидать, что в каждый момент времени значение переменной будет ближе к значению, «соседнему» по времени, чем к более удаленному.

Для последующего анализа тренд нужно удалить. Для этого, перейдя на закладку **x=f(x)**, установить переключатель в правой нижней части ДО в положение **Trend subtract x=x-(a+b*t))** (Вычесть тренд) и нажать кнопку **OK (Transform selected series)**. Появившийся график – это разность между значениями ряда и рассчитанным трендом. Нажав на закладке **Autocorrs** кнопку **Autocorrelations**, расположенную в левой нижней части окна, получим таблицу и график АКФ для интервалов времени 1, 2,..., 15 дней (рис.26). Как и следовало ожидать, АКФ монотонно убывает с увеличением временного интервала. Отметим, что требуемое количество временных интервалов задается в поле выбора **Number of lags** в нижней части ДО.

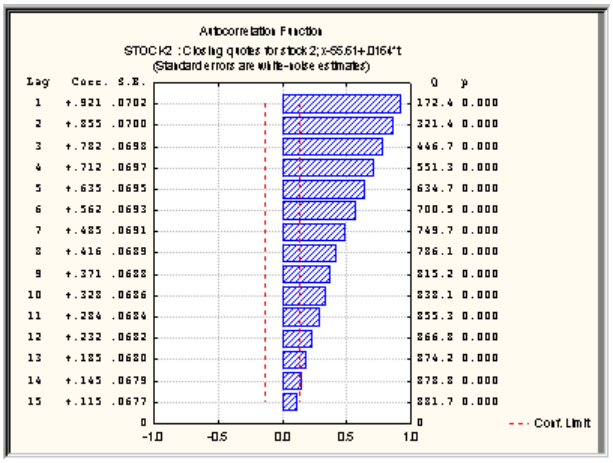


Рисунок 26 – График АКФ

5.Анализ частных автокорреляционных функций (ЧАКФ). Этот анализ является дополнительным средством исследования временного ряда. Нажав кнопку **Partial autocorrelations** (Частные автокорреляции)**,** расположенную ниже кнопки **Autocorrelations,** увидим, что значимой является только ЧАКФ для лага=1 (только ее столбец пересекает красную линию 95% доверительного интервала. Это означает, что значение процесса в текущий момент времени практически полностью определяется его значением в предыдущий момент. Такие процессы (известные как *«процессы без последействия»*) практически невозможно прогнозировать.

Задание. Построить график ЧАКФ для 20 временных интервалов. Сделать вывод.

Закрыть окна **Workbook** и **Data** без сохранения результатов анализа.

4.5.3 Процедура АПРС (ARIMA)

Процедура, называемая «автопроинтегрированное регрессионное среднее» (англоязычный термин – ARIMA) позволяет строить математическую модель временного ряда и на ее основе прогнозировать значения ряда в будущем.

Задача. Имеются данные об объемах авиаперевозок за 11 лет. Построить модель временного ряда и предсказать объем перевозок на последующие 12 месяцев.

Порядок работы. 1.Открыть файл данных ***Series_ G.sta*,** содержащий данные о количестве пассажиров за 1949-60 г.г. (в тысячах). Выполнить в главном меню команду **Statistics/ Advanced Linear/Nonlinear Models ► Time Series/Forecasting**. В открывшемся ДО нажать кнопку **Variables**, в открывшемся ДО выбрать единственную переменную и нажать **OK**. На закладке **Quick** нажать кнопку **ARIMA & autocorrelation functions** (ARIMA и автокорреляционные функции). Откроется ДО **Single Series ARIMA: Series_G** (ARIMA для одного ряда) (рис.31.)

2.Прежде чем задать параметры для работы процедуры ARIMA, необходимо выполнить идентификацию модели. Для этого выполняется анализ АКФ и ЧАКФ. Вначале выполняется визуальный анализ.

Перейти на закладку **Advanced** (Расширенные опции) и нажать кнопку **Other transformations & plots** (Другие преобразования и графики) внизу окна слева. В открывшемся ДО **Transformations of Variables: Series_G** нужно прежде всего ввести масштаб по оси X. Для этого перейти на закладку **Review & plot**, установить переключатель **Scale X-axis in plots manually (min, step)** (Масштабировать ось X вручную (минимум, шаг)), ввести значение **1** в поле **Min** и **12** в поле **Step** (12 месяцев в году). Чтобы маркировать ось X датами, установить в поле **Label**

data points with опцию **Case names** (Названия наблюдений). Для получения графика ряда нажать кнопку **Plot.**

3.Можно видеть (рис. 27), что амплитуда сезонных колебаний объемов перевозок имеет тенденцию к увеличению (так называемая мультипликативная сезонность) – это может повлиять на значения АКФ и ЧАКФ. Поэтому следует выполнить логарифмирование ряда. Для этого развернуть окно **Transformations of Variables: Series_G,** перейти на закладку **x=f(x)** и установить опцию ***Natural log (x=ln(x))*** в левой нижней части окна, после чего нажать кнопку **OK (Transform selected series)** (Преобразовать выбранные ряды)**.** На новом графике амплитуды колебаний практически одинаковы (рис. 28).

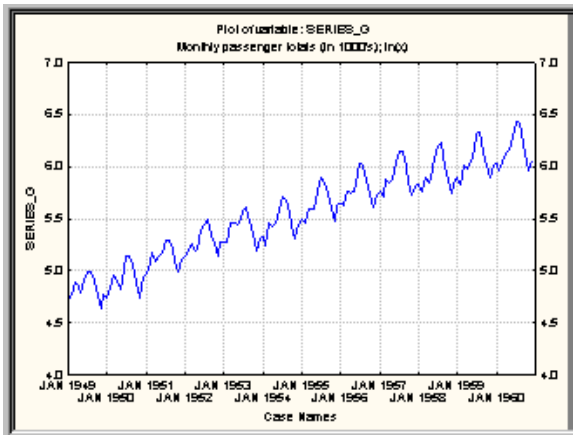


Рисунок 27 – Исходный ряд

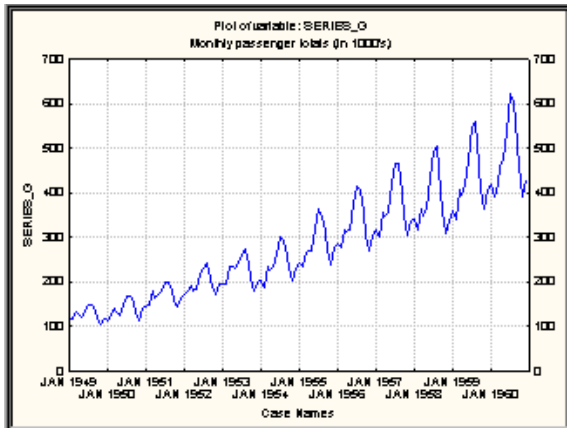


Рисунок 28 – Прологарифмированный ряд

4.Для вычисления АКФ на закладке **Autocorrs** окна **Transformations of Variables: Series_G** нужно сделать следующую установку: в поле **Number of lags** (Количество лагов) внизу слева сменить значение **15** на значение **25**. Затем нажать кнопку **Autocorrelations** в левой части окна. На полученном графике можно видеть, что значения АКФ весьма велики, особенно в диапазоне лагов от 1 до 12 (рис. 29).

5.Для устранения этой серийной зависимости прибегают к дифференцированию ряда. На закладке **Difference, integrate** установить слева внизу переключатель в положение **Differencing (x=x-x(lag))** , не меняя других установок (в частности, значение **lag** рядом с переключателем **Differencing…** должно быть задано равным **1**), и нажать кнопку **OK (Transform selected series)**.

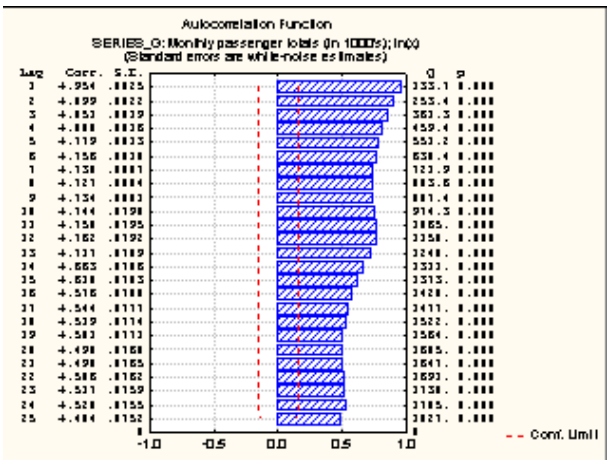


Рисунок 29 – АКФ исходного ряда

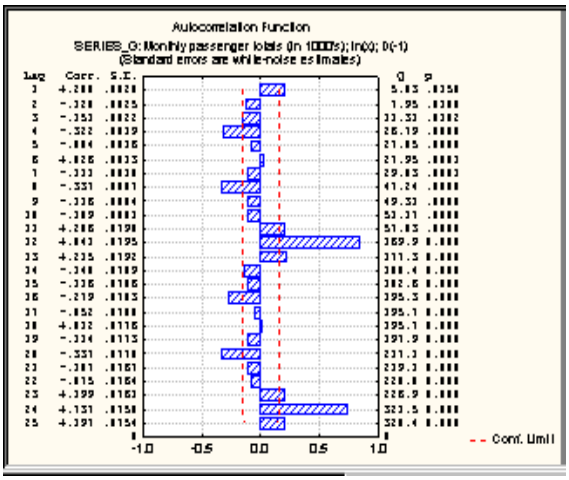


Рисунок 30 – АКФ продифференцированного ряда

Будет построен график продифференцированного ряда. Вернувшись в ДО **Transformations of Variables: Series_G** вновь построить АКФ (см. п. 4). Теперь большинство зависимостей исчезло (см. рис. 30).

6.Наряду с тем, после удаления короткопериодической зависимости проявилась (ранее скрытая) длиннопериодическая зависимость (с лагами, кратными 12). Это – сезонная зависимость. Для ее устранения нужно повторно выполнить п. 5, установив значение **lag=12**. На вновь построенном графике АКФ остались заметные зависимости для лагов 1 и 12, но величина последней уменьшилась примерно вдвое.

7. Поскольку выполненный анализ говорит о наиболее сильной зависимости ряда для лагов 1 и 12, то целесообразно применить модель ARIMA первого порядка (что соответствует лагу 1) с сезонным параметром 12.

Для выполнения процедуры ARIMA нажать в ДО **Transformations of Variables: Series_G** кнопку **Cancel** – произойдет возврат в ДО **Single Series ARIMA: Series_G**.

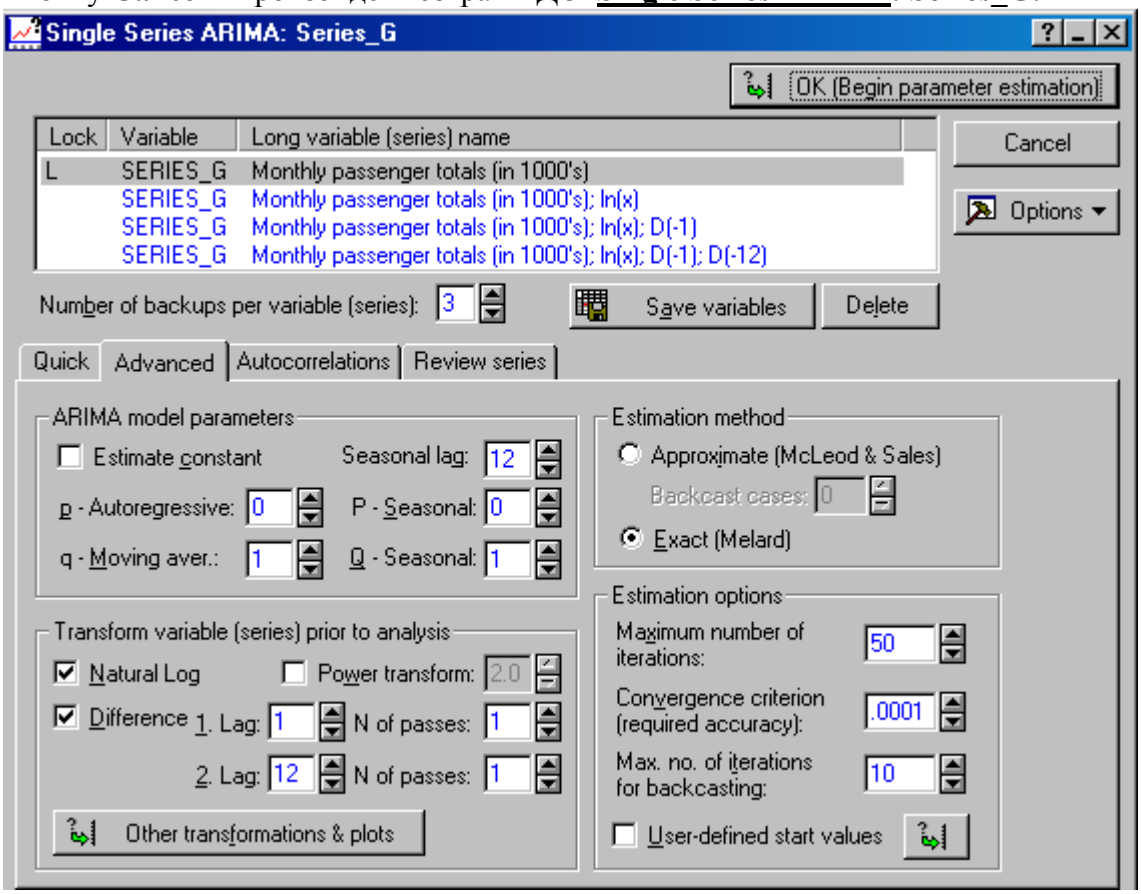


Рисунок 31 – Настройка опций для выполнения процедуры ARIMA

В окне выбора переменных для анализа выбрать исходную переменную ***Series_G Monthly passenger totals (in 1000's.)*** В поле **ARIMA model parameters** (Параметры модели ARIMA) установить значения **q-Moving aver.:** и **Q-Seasonal:** равными **1** (последнее значение даст **12** при умножении на установленное выше значение **Seasonal lag:**. В поле **Transform variable (series) prior to analysis** (Преобразовать переменную (ряд) перед проведением анализа) необходимо указать все описанные выше преобразования, которым подвергалась переменная в процессе предварительного анализа: установить флажки **Natural Log** и **Difference,** значения **1.Lag: 1; 2.Lag: 12.** В обоих полях **N of passes:** (Число шагов) установить значение, равное **1**.Для расчета параметров модели в поле **Estimation method** (Метод оценки) установить опцию **Exact (Melard)**.

После установки всех указанных опций окно диалога должно выглядеть, как показано на рис. 31. Нажать кнопку **OK (Begin parameter estimation) –** то есть "Начать оценивание параметров"- в верхнем правом углу окна**.**

8. Результаты анализа представлены в ДО **Single Series ARIMA Results: Series_G**. Рассчитанные параметры модели будут выданы после нажатия на закладке **Quick** кнопки **Summary: Parameter estimates**. Оба параметра (**q=0.401823** и **Qs=0.556937)** статистически значимы- на это указывают значения p-уровней. Для построения прогноза нужно вернуться в ДО **Single Series ARIMA Results: Series_G** и на закладке **Advanced** в поле **Forecasting** (Прогнозирование) установить соответствующие опции: **Number of cases:** (Число наблюдений), **Start at case:** (Начать с наблюдения), **Confidence level:** (Доверительный уровень).

В данном случае требуется построить прогноз на 12 месяцев, начиная со 145, доверительная вероятность (уровень) задается обычно равной 0,9. Поэтому необходимо задать указанные значения параметров и нажать кнопку **Forecast cases** (Предсказанные значения). Результаты предсказаний будут представлены в таблице. Для получения графика зависимости нажать кнопку **Plot series & forecasts** (Графики рядов и предсказаний). Полученный прогноз должен выглядеть, как на рис. 32, где предсказанные значения ВР даны красным цветом, 95% доверительные границы – зеленым.

9.Проверка качества модели. Нажав в ДО **Single Series ARIMA Results: Series_G** на закладке **Distribution of residuals** (Распределение остатков) кнопки **Histogram** и **Normal probability plot** (График нормального распре-деления)**,** получим сравнения распределения остатков модели с нормальным законом распределения. Наблюдается хорошее согласие. Наконец, на закладке **Autocorrelations** нажатие одноименной кнопки позволяет получить график АКФ остатков. Все значения АКФ, кроме одного, малы (не выходят за границы 95% доверительного интервала) – это признак того, что в данных практически «не осталось зависимости, которая не была бы объяснена построенной моделью».

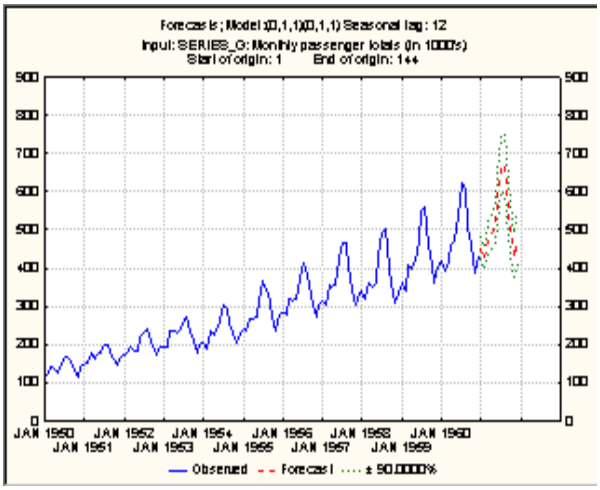


Рисунок 32 – Результат построения прогноза

9.Закрыть окна **Workbook** и **Data.**

Контрольные вопросы.

1. Что означает термин «лаг»?
2. Какие из величин, приведенных в таблице результатов анализа распределенных лагов, указывают, насколько сильно связаны значения временного ряда с заданным лагом?
3. Как построить график временного ряда?
4. Как можно выявить тренд на графике временного ряда?
5. Для чего прибегают к логарифмированию временного ряда; к дифференцированию временного ряда?
6. Как по графикам АКФ и ЧАКФ установить, для каких значений лага зависимость значений ряда является статистически значимой?
7. Какие установки нужно выполнить в поле **Transform variable (series) prior to analysis** (Преобразовать переменную (ряд) перед проведением анализа) прежде, чем строить прогноз значений ряда?
8.Как проверяется качество модели временного ряда, полученной с помощью процедуры ARIMA?

4.6 Лабораторная работа №6
Промышленный статистический анализ

Цель работы:
1) научиться строить причинно-следственные диаграммы;
2) усвоить порядок построения планов выборочного контроля;
3) ознакомиться с методами контроля качества Тагути.

4.6.1 Причинно-следственные диаграммы

Эти диаграммы (диаграммы Ишикавы, или «рыбья кость») используются для установления причинно-следственных связей между неисправностями (дефектами) и действующими факторами.

Задача. Проанализировать возможные причины, по которым не включается настольная лампа.

Порядок работы. 1. Запустить программу **STATISTICA.**

2. Создать таблицу с перечнем возможных неисправностей. Для этого выполнить команду главного меню **File/ New**… В открывшемся ДО **Create New Document** на закладке **Spreadsheet** (Таблица) необходимо задать значения **Number of variables:** (Число переменных) и **Number of cases:** (Число наблюдений). В данном случае в качестве переменных будут выступать группы действующих факторов, в качестве наблюдений – факторы. Зададим **Number of variables:=4, Number of cases:=3.** Переключатель **Placement** (Расположение) установить в положение **As a stand-alone window** (Как отдельное окно). Нажать кнопку **OK.**

Чтобы дать переменной название, необходимо щелкнуть на ее заголовке правой кнопкой мыши и в контекстном меню выполнить команду **Variable Specs…**В открывшемся ДО в поле **Name** ввести название. Назвать переменные: ***Энергия, Вилка-Шнур, Лампочка, Выключатель.*** Чтобы подогнать ширину столбца под его содержимое, необходимо навести курсор на правую границу заголовка (он при этом превратится в двунаправленную стрелку) и дважды щелкнуть левой кнопкой мыши. Заполнить таблицу (рис.33)

	1 Энергия	2 Вилка-Шнур	3 Лампочка	4 **Выключатель**
1	Отключение на линии	Вилка не вставлена в розетку	Отсутствует	Выключен
2	Отключились предохранители	Обрыв шнура	Перегорела	Сломан
3			Неплотно вкручена	Нет контакта

Рисунок 33- Таблица действующих факторов

Подогнать ширину столбцов под содержимое.

3. Выполнить команду главного меню **Statistics** и выбрать метод анализа **Industrial Statistics & Six Sigma ► Process Analysis.**

В открывшемся ДО выбрать метод анализа **Cause-effect (Ishikawa, Fishbone) diagrams** (Причинно-следственные диаграммы (Ишикава, «рыбья кость»))**.** Нажать кнопку **OK.**
В открывшемся диалоговом окне (рис. 34) можно задать атрибуты диаграммы. На закладке **Font sizes** задаются размеры шрифта. На закладке **Arrows** (Стрелки) можно задать размеры стрелок и угол их наклона к осевой линии.

Нажать кнопку **Variables** (Переменные).

3.В открывшемся ДО **Select variables for main causes ('Fishbones')** (Выбрать переменные для главных причин) в левом поле выделить (при нажатой клавише **Ctrl**) названия групп факторов, которые будут помещены над центральной линией диаграммы, в правом – названия групп факторов, которые будут помещены под центральной линией. Нажать кнопку **OK.**

4.Во вновь открывшемся ДО снова нажать **OK.** Будет получена диаграмма Ишикавы.

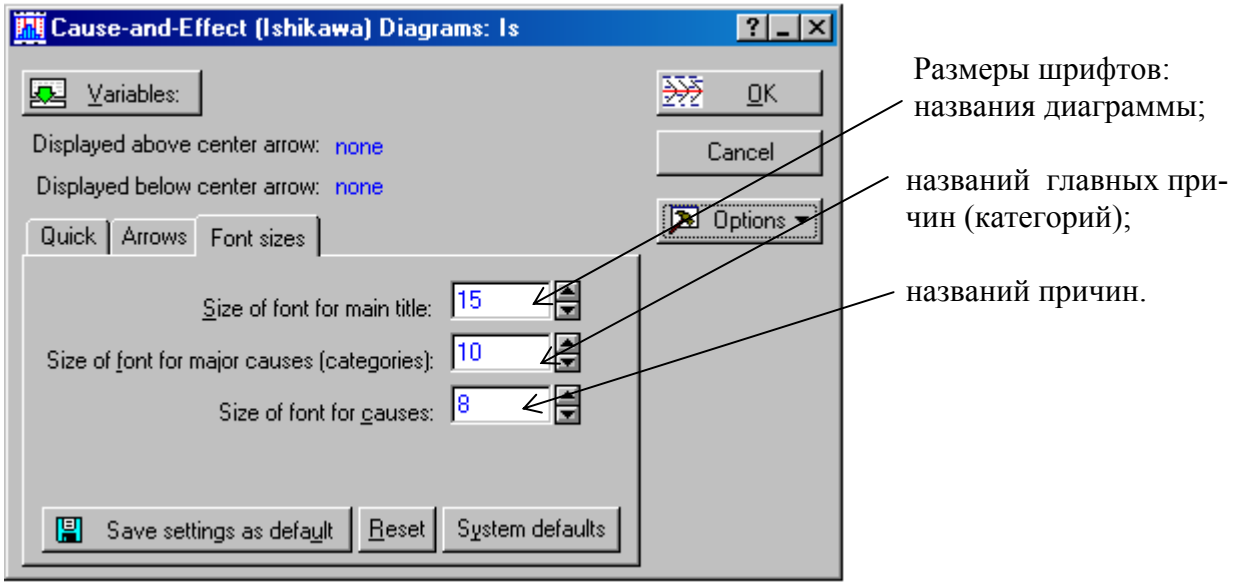


Рисунок 34 - Окно задания атрибутов диаграммы

5. После анализа результатов закрыть окна: **Workbook1…** (Рабочая книга…), **Report…** (Отчет…), **Data:** (Данные) - без сохранения полученных результатов.

4.6.2 Планы выборочного контроля

Задача, которая решается при построении планов выборочного контроля, заключается в том, чтобы определить объем выборки из партии деталей, которую необходимо испытать (проверить), чтобы принять или забраковать всю партию. Первое решение принимается, если справедливой будет признана гипотеза ***H_0: контролируемый параметр находится в пределах допуска,*** второе – если справедлива альтернативная гипотеза ***H_1: контролируемый параметр выходит за пределы допуска.***

Задача. Имеются результаты измерений диаметра поршневого кольца в трех случайных выборках, по 100 изделий в каждой. Номинальный диаметр кольца – 74 мм. Из результатов предшествующих измерений известно, что среднеквадратическое отклонение *(sigma)* равняется 0,01 мм. Требуется разработать план контроля, обеспечивающий браковку колец в случае отклонения диаметра от номинала на величину более 0,005 мм.

Порядок работы. 1.Запустить программу **STATISTICA.** Командой главного меню **File / Open …** открыть файл ***Pistons2.sta*** из папки ***Examples / Datasets.*** Выполнить команду **Statistics/ Industrial Statistics & Six Sigma►Process Analysis.** В ДО **Process Analysis Procedures: Pistons2.sta** выбрать тип плана **Sampling plans for means, proportions, & Poisson frequencies** (Планы для средних, отношений и пуассоновских частот)**.** Нажать кнопку **OK.**

2.В ДО **Sampling plans…** (рис. 35) на закладке **Advanced** установить следующие параметры: в поле **Hypothesized means for H0...** (Среднее, при котором принимается гипотеза H0) = **74**; в **Hypothesized means for H1...**(Среднее, при котором принимается H1) = **74,005** (или 73,995, если гипотеза H1 принимается как при отклонении на 0,005 вправо, так и влево - для этого в поле **Test criterion** должно быть установлено значение **Two tailed** – «с двумя хвостами»); в поле **Assumed sigma…** (Заданное среднеквадратическое отклонение) =**0,01. Alpha error…** (Вероятность ошибки 1 рода) =**0,05**; **Beta error…** (Вероятность ошибки 2 рода) =**0,10**. Нажать **OK**.

3.Построение плана с фиксированным объемом выборки. В ДО **Sampling Plans Results…** перейти на закладку **Fixed sampling plan** (План с фиксированным объемом выборки)**.** Рядом с кнопкой **Sample size** указан потребный объем выборки для заданного значения ошибки второго рода (***43)***. Нажав кнопку **Summary** в правой части окна, можно получить параметры плана контроля, в том числе нижнюю контрольную границу **Lover conf. limit (73,9970)** и верхнюю контрольную границу **Upper conf. limit (74,0030).**

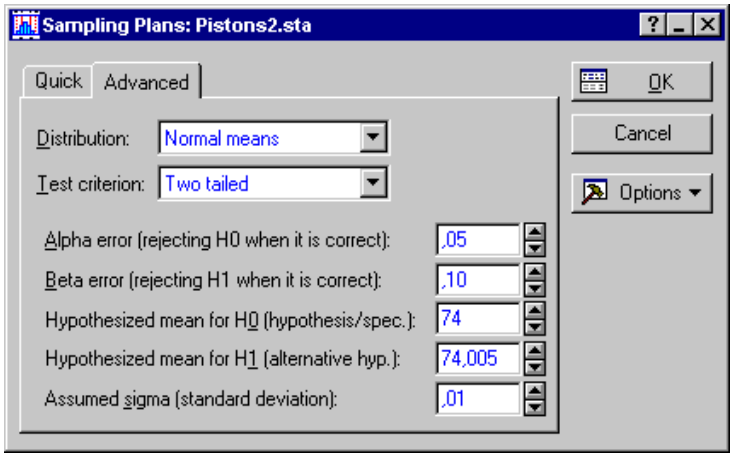


Рисунок 35- Окно задания параметров плана

4.Развернув ДО **Sampling Plans results…** (из пиктограммы в нижней части рабочего поля), нажать кнопку **Operating characteristics curve** (График операционной характеристики). Эта характеристика (рис. 36) дает мощность плана, т.е. вероятность того, что данное отклонение будет обнаружено по выборке данного объема. В ДО **Sampling Plans Results…**нажатием **Sample size:** можно задать другой объем выборки в поле **Specify Sample Size** и посмотреть, как изменится величина **Beta**.

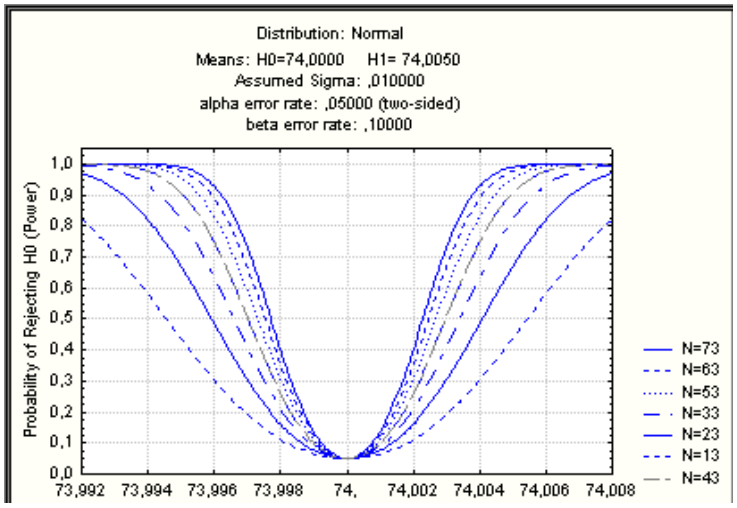


Рисунок 36- График операционной характеристики

Задание. Задавая различные объемы выборки, проанализировать, как изменяется мощность плана.

Построение последовательного плана. 1.В ДО **Sampling Plans Results…** перейти на закладку **Sequential sampling plan** (Последовательный план). Нажав кнопку **Variable containing data following current plan:** (Переменная, содержащая данные для текущего анализа), нужно выбрать переменную (в данном случае - №1), нажать **OK**. После нажатия кнопки **Summary of equivalent sequential sampling plan** (Сводка результатов для последовательного плана), получаем таблицу, в которой видно (в последнем столбце), что, начиная с 21 замера, предлагается партию принять (**accept**). Отметим, что слово **continue** в последнем столбце указывает на необходимость продолжать испытания, а слово **reject** означает рекомендацию отвергнуть партию изделий.

2.После возврата в ДО **Sampling Plans Results…** нажать **Plot equivalent sequential sampling plan** (График последовательного плана) – можно видеть изменение кумулятивной суммы и коридоры принятия решения (рис 37).

Задание. Построить план последовательного контроля для переменных Sample_2, Sample_3. Сделать выводы о возможности принятия партии изделий.

3. Чтобы создать пустой контрольный бланк для последующего заполнения вручную, необходимо при выборе переменных для анализа один раз нажать на кнопку **Variable containing data following current plan:** (рядом с ней появится значение **none -** нет) и нажать кнопку **Plot equivalent sequential sampling plan.**

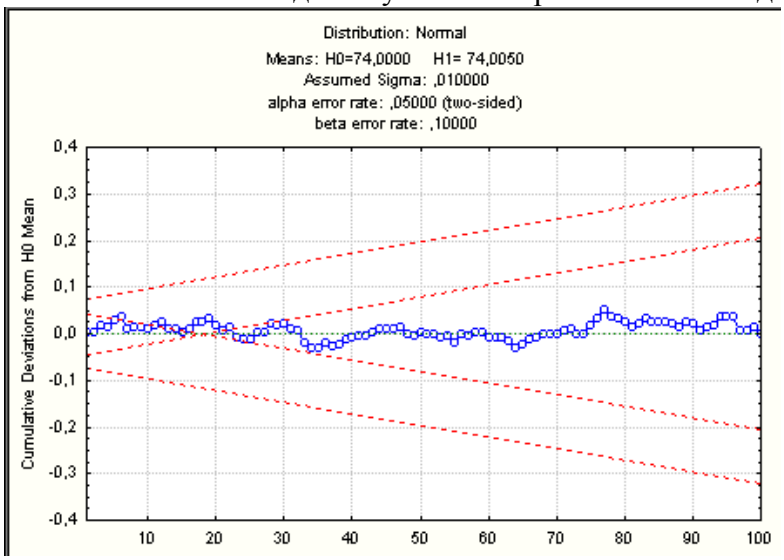


Рисунок 37 - График кумулятивной суммы

Следует отметить, что уменьшение значения *бетта* увеличивает размер выборки, необходимой для принятия партии, а уменьшение значения *альфа* – увеличивает размер выборки, необходимой для того, чтобы отвергнуть партию.

4. Закрыть все окна.

4.6.3 Методы обеспечения качества Тагути

Показателем качества по Тагути является, как известно, «Отношение сигнал / шум» (**Signal-to-Noise ratio,** или **S/N).** Под сигналом понимается управляемый фактор, под шумом – неуправляемый, влияние которого на изделие способствует снижению его качества. Для оценки этого отношения могут использоваться методы планирования экспериментов. Использование при планировании экспериментов специально разработанных Тагути ортогональных массивов позволяет минимизировать их число.

Задача. Для того, чтобы повысить качество полисиликоновых пластин, требуется уменьшить число дефектов поверхности и минимизировать вариацию толщины ее покрытия. Исследовалось влияние 6 факторов (описываемых далее) на указанные показатели качества.

Каждый фактор варьировался на трех уровнях, полученные данные занесены в файл ***Taguchi.sta*.** Выявить факторы, оказывающие наиболее сильное влияние на качество пластин.

Порядок работы. 1. Запустить программу **Statistica.** Открыть файл ***Taguchi.sta***. Используя в меню **Data** команду **All Variable Specs...** (Спецификации всех переменных), вызвать диалоговое окно **Variable Specifications Editor** (Редактор спецификаций переменных). Он позволяет просмотреть формат переменных. В поле **Long name…** приведено содержательное описание каждой используемой переменной:

- Независимые переменные (факторы):

- TEMPERATURE (температура при напылении).
- PRESSURE (давление при напылении).
- NITROGEN (расход азота).
- SILANE (расход силана).
- SETT_TIM (время осаждения).
- CLEANING (методы очистки).

- Зависимые переменные:

-S_DEF1…S_DEF9 - количество дефектов, обнаруженных на 9 различных участках пластин.
-THICK1…THICK9 – толщина 9 различных участков пластин.

- ***D_0_3 ... D_1001__ -*** переменные, содержащие данные о количестве образцов пластин (из 9, проверенных в каждом эксперименте), количество дефектов поверхности которых укладывается в соответствующий интервал: переменная ***D_0_3*** – количество пластин с числом де-

фектов от 0 до 3, ***D_4_30*** - количество пластин с числом дефектов от 4 до 30 и т.д. Выйти из окна редактора спецификаций, нажав кнопку **Cancel.**

2. Создать план эксперимента, для чего в меню **Statistics** выбрать модуль **Industrial Statistics & Six Sigma** и в нем – раздел **Experimental Design (DOE)** (Планирование эксперимента). В открывшемся ДО на закладке **Advanced** выбрать тип эксперимента **Taguchi robust design experiments (orthogonal arrays)** (План робастного эксперимента Тагути (ортогональные планы)). Нажать кнопку **OK.**

3. Во вновь открывшемся ДО **Design & Analysis of Taguchi Robust Design Experiments** перейти на закладку **Design experiment,** на которой можно выбрать ортогональный массив. Для данного случая подходит массив ***L18***: он предусматривает проведение 18 экспериментов для 8 факторов, из которых 7 варьируются на трех уровнях и один – на двух (в данном примере два фактора останутся неиспользованными).

Ортогональность массива означает, что его столбцы независимы – это облегчает оценку главных эффектов (то есть «чистых» факторов). Отметим, что существует возможность создания иных массивов помимо стандартных, предлагающихся при использовании данной процедуры.

Выбрав массив и нажав кнопку **OK**, в открывшемся ДО **Design of a Robust Design Experiment: Taguchi** на закладке **Display Design** (Показать план) в поле **Order of runs** (Порядок выполнения опытов) установить опцию **Standard order** (Стандартный порядок) и нажать кнопку **Summary: Display design** (Сводка плана эксперимента) - получим таблицу плана, фрагмент которой приводится на рис. 38.

Standard Run	F 1	F 2	F 3	F 4	F 5	F 6	F 7	F 8
1	1	1	1	1	1	1	1	1
2	1	1	2	2	2	2	2	2
3	1	1	3	3	3	3	3	3
4	1	2	1	1	2	2	3	3
5	1	2	2	2	3	3	1	1

Рисунок 38 – Фрагмент плана

В этой таблице указаны уровни факторов в каждом из планируемых экспериментов. Вернувшись в ДО **Design of a Robust Design Experiment,** на закладке **Alias structure** (Структура смешивания эффектов) нажать одноименную кнопку. Будет получена таблица взаимодействий факторов. Звездочка в ячейке указывает, что главный эффект, расположенный в данной строке, смешан со взаимодействием факторов, расположенных в столбце.

Alias Structure (Taguchi)
L18: 1 factor with 2 levels; 7 factors with 3 levels
* = partially or completely confounded

Effect	**1 2**	1 3	1 4	1 5	1 6	1 7	1 8	2 3	2 4	2 5	2 6	2 7	2 8	3 4	3 5	3 6	3 7	3 8	4 5	4 6	4 7	4 8	5 6	5 7	5 8	6 7	6 8	7 8
1														*	*			*		*	*		*	*			*	*
2														*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
3			*	*		*		*	*	*	*	*							*	*	*	*	*	*	*	*	*	*
4		*			*	*		*		*	*	*	*		*	*	*	*					*	*	*	*	*	*
5		*			*	*		*	*		*	*	*	*		*	*	*		*	*	*				*	*	*
6			*	*			*	*	*	*		*	*	*	*		*	*	*		*	*		*	*			*
7			*	*			*	*	*	*	*		*	*	*	*		*	*	*		*	*		*		*		
8		*			*	*		*	*	*	*	*		*	*	*	*		*	*	*		*	*		*		

Рисунок 39– Таблица взаимодействий факторов

4. Выполнить анализ результатов эксперимента, для чего в ДО **Design of a Robust Design Experiment** нажать кнопку **Cancel** – это позволит вернуться в окно предыдущего шага – и перейти на закладку **Analyze design.**

Прежде всего проанализируем количество дефектов на поверхности пластины. Естественно, что их желательно иметь как можно меньше – поэтому выберем в поле выбора **Problem type:** (Тип проблемы) вариант **Smaller-the-better** («Чем меньше – тем лучше»).

Выбрать переменную для анализа: с помощью кнопки **Variables** открыть окно выбора и в левом поле **Dependent variables:** (Зависимые переменные) выделить переменные ***S_def1 ... S_def9***, характеризующие дефекты поверхности. В поле **Independent vars (factors):** (Независи-

мые переменные) выбрать переменные ***2-Temperature.. 6-Sett_tim*** и переменную ***8-Cleaning.*** Таким образом, из 8 факторов, предусмотренных для этого типа массивов, не будут использованы переменные 1 и 7. Нажать кнопку **ОК** для подтверждения сделанного выбора переменных.

5.Для получения результатов в ДО **Design & Analysis of Taguchi Robust Design Experiment: Taguchi** нажать кнопку **ОК.** В открывшемся ДО **Analysis of a Robust Design Experiment** приведена сводка плана эксперимента: факторы, планируемое число экспериментов, независимые переменные, тип анализируемой проблемы, формула для определения параметра ETA, то есть **Signal-to-Noise ratio** («Отношения сигнал / шум»).

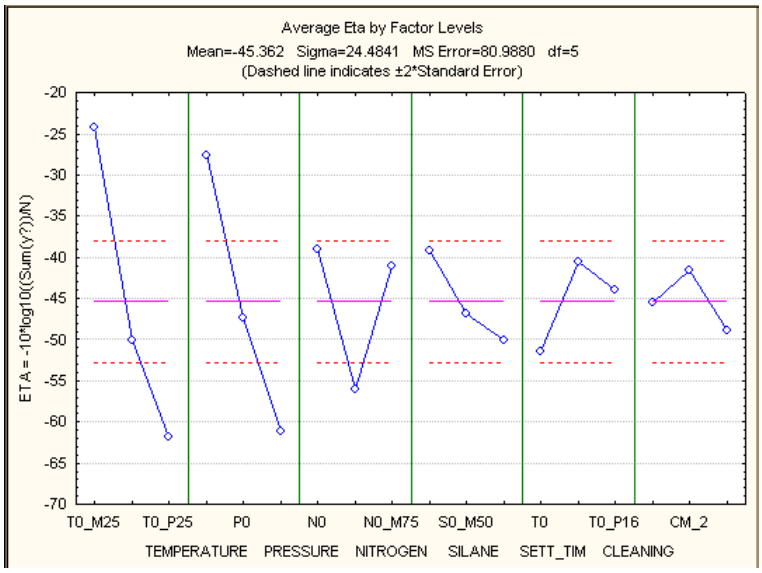


Рисунок 40 – Средние значения отношения S/N

6.Нажав кнопку **Marginal means** (Предельные средние), можно получить таблицу, в которой для каждого уровня всех факторов представлены средние по общему числу экспериментов отношения S/N («сигнал-шум»). Кнопка **Means plot** (Графики средних) в этом же окне позволяет получить графическую иллюстрацию (см. рис. 40)

Пунктиром показан интервал [-2σ;+2σ] для значения S/N. Из этого графика можно видеть, что наибольший разброс отношения S/N обусловлен фактором «Температура при напылении» (т.е. его влияние на качество максимально), наименьший – фактором «Методы очистки».

7. Дополнительные данные о рассматриваемой проблеме можно получить, выполнив интегральный анализ (**Accumulation analysis).**

Вернувшись в ДО **Design & Analysis of Taguchi Robust Design Experiments** (для чего нужно нажать кнопку **Cancel**), в поле выбора **Problem type:** выбрать вариант **Accumulation analysis** (Обобщающий анализ). С помощью кнопки **Variables** выбрать переменные ***D_0_3 ... D_1001*** в поле **Dependent variables:,** переменные ***2-Temperature.. 6-Sett_tim*** и переменную ***8-Cleaning*** – в поле **Independent vars (factors):.**

После нажатия кнопки **ОК** (последовательно в двух окнах) откроется ДО **Accumulation Analysis Results: Taguchi** со сводкой анализа. Нажав кнопку **Bar plot of cumulative proportions** (Столбчатая диаграмма накопленных частот), получим диаграмму, на которой для каждого фактора показано распределение частот попадания числа дефектов в каждый из интервалов, характеризуемых переменными ***D_0_3 ... D_1001.***

Так, при уровне фактора ***Temperature,*** равном ***TO_M25,*** были получены 54 пластины, в 34 из которых наблюдалось число дефектов поверхности, не превышающее трех. Это значение можно получить, если в таблице данных **Data: Taguchi…**[1] сложить числа, стоящие в ячейках на пересечении столбца 28 (переменная ***D_0_3***) и строк 1, 2, 3, 10, 11, 12 (соответствующих уровню ***TO_M25*****).** Доля таких пластин составляет 63%, что можно видеть по первому столбцу диаграммы.

Эта же информация может быть получена для каждого фактора на отдельном линейном графике. Для этого нужно в ДО **Accumulation Analysis Results: Taguchi** нажать кнопку **Line**

[1] Для вывода на экран таблицы нужно в главном меню выполнить команду **Window► Data.**

graph by factor, выбрать в открывшемся окне нужный фактор и нажать кнопку **OK.** Так, для фактора ***Temperature*** («Температура») получим следующий интегральный график (рис. 41).

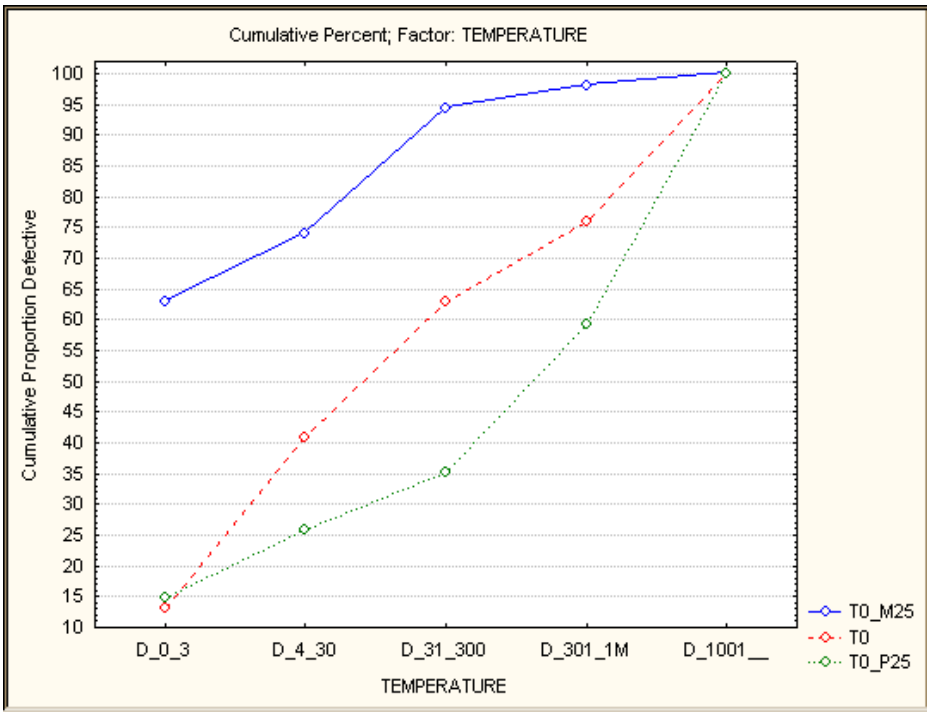


Рисунок 41 – Интегральный график для фактора “TEMPERATURE”

Из него следует, что при уровне температурного фактора, составляющем ***TO_M25,*** 63% образцов имеет не более трех дефектов поверхности, 75 – 63 = 12 процентов – от 4 до 30 и т.д. При уровне же фактора ***TO_P25*** число образцов с тремя и менее дефектами составляет лишь 15%, в то время как число образцов с более чем 301 дефектом составляет 45%.

Полученные результаты могут оказать существенную помощь в решении проблемы качества.

Задание: проанализировать, как влияет качество очистки поверхности (фактор Cleaning) на количество дефектов поверхности; при каком уровне фактора число дефектов на поверхности пластины максимально?

Контрольные вопросы

1. Какие исходные данные необходимы для построения плана выборочного контроля?
2. В чем преимущество последовательного плана перед планом с фиксированным объемом выборки?
3. Для чего служит график операционной характеристики?
4. Как, используя график последовательного плана, решить, можно ли принять данную партию изделий?
5. Какая величина является показателем качества в методах Тагути?
6. Для какой цели используются графики средних значений отношения S/N?
7. Что показывает интегральный график для данного исследуемого фактора?

4.7 Лабораторная работа №7
Построение карт контроля качества

Цель работы:
1) усвоить общий порядок построения карт контроля качества;
2) научиться оценивать качество измерительной системы;
3) научиться строить карты Парето, X-R карты и карты контроля по альтернативному признаку.

4.7.1 Оценка качества измерительной системы

Для оценки качества измерительной системы выполняется анализ *повторяемости (repeatability) и воспроизводимости (reproducibility).*

Задача. Оценить качество измерений, производимых пятью операторами, каждый из которых выполняет по три повторных измерения 8 образцов. Данные для анализа содержатся в файле ***Temperat.sta.***

Порядок работы. 1.Открыть файл ***Temperat.sta.*** Вызвать процедуру **Process Analysis** из раздела **Industrial Statistics & Six Sigma**. В ДО **Process Analysis Procedures** выполнить пункт **Gage repeatability & reproducibility** (Повторяемость и воспроизводимость измерительной системы) - откроется ДО **Repeatability & Reproducibility Analysis - Generate design**. На закладке **Generate design** (Генерировать план контроля) ввести в поле **Number of operators:** (Число операторов) значение **5** , в поле **Number of parts:** (Число образцов) – **8,** в поле **Number of trials:** (Число измерений) - **3**. Нажать кнопку **OK**.

2. В открывшемся ДО **Repeatability & Reproducibility Design: Temper..**, выбрав одну из закладок и нажав на ней кнопку **Save design**, можно сохранить план контроля в формате файла данных (закладка **Data files**) или в формате таблицы данных (закладка **R & R data sheets**). Выбрать формат **data file.**

Порядок измерений может быть выбран детерминированным (**Standard order**) или случайным (**Randomize trials**). Достоинство последнего: рандомизация устраняет систематическую ошибку, накапливающуюся по мере уставания операторов. Эти виды выбираются соответствующим переключателем. После нажатия кнопки **Summary: Display design** будет выдана таблица для сбора результата измерений. (В данном примере полученную таблицу заполнять не нужно, так как все необходимые для анализа данные уже содержатся в таблице загруженного файла).

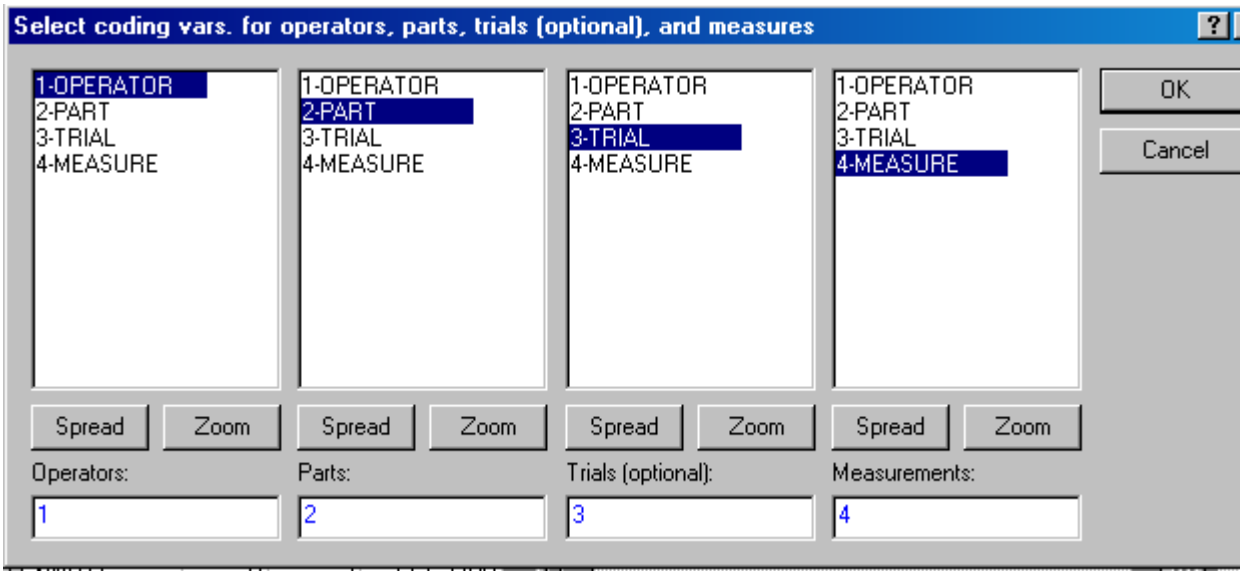


Рисунок 42 – Окно задания переменных для анализа

3.Нажав в ДО **Repeatability & Reproducibility Design** кнопку **Cancel**, вернуться в ДО **Repeatability & Reproducibility Analysis – Generate design: Temperat**, где перейти на

закладку **Analyze data file** и с помощью кнопки **Variables** открыть ДО выбора переменных. Задать переменные, как показано на рис. 42.

Нажатие кнопки **OK** приведет к возврату на закладку **Analyze data** . Нажать кнопку **Codes: (for operators, parts, trials)** и в открывшемся ДО выбрать все сочетания, нажав кнопку **Select All**. Тем самым указывается, что анализу подлежат результаты всех выполненных измерений. Нажав **OK** , вернуться на закладку **Analyze data file**. Нажать кнопку **OK**.

4. В открывшемся ДО результатов **Gage Repeatability & Reproducibility Results**: **Temperat** в текстовом поле представлена информация о количестве выполненных измерений и их статистических оценках. В частности, средний размер детали по результатам **120** измерений оценивается как **120,025** (параметр Mean), а среднеквадратическое отклонение – как **8,10851** (Std. Dev). В данном ДО имеется ряд опций для представления результатов анализа. Нажав кнопку **Repeatability & Reproducibility plot**, получим сводный график результатов (рис. 43).

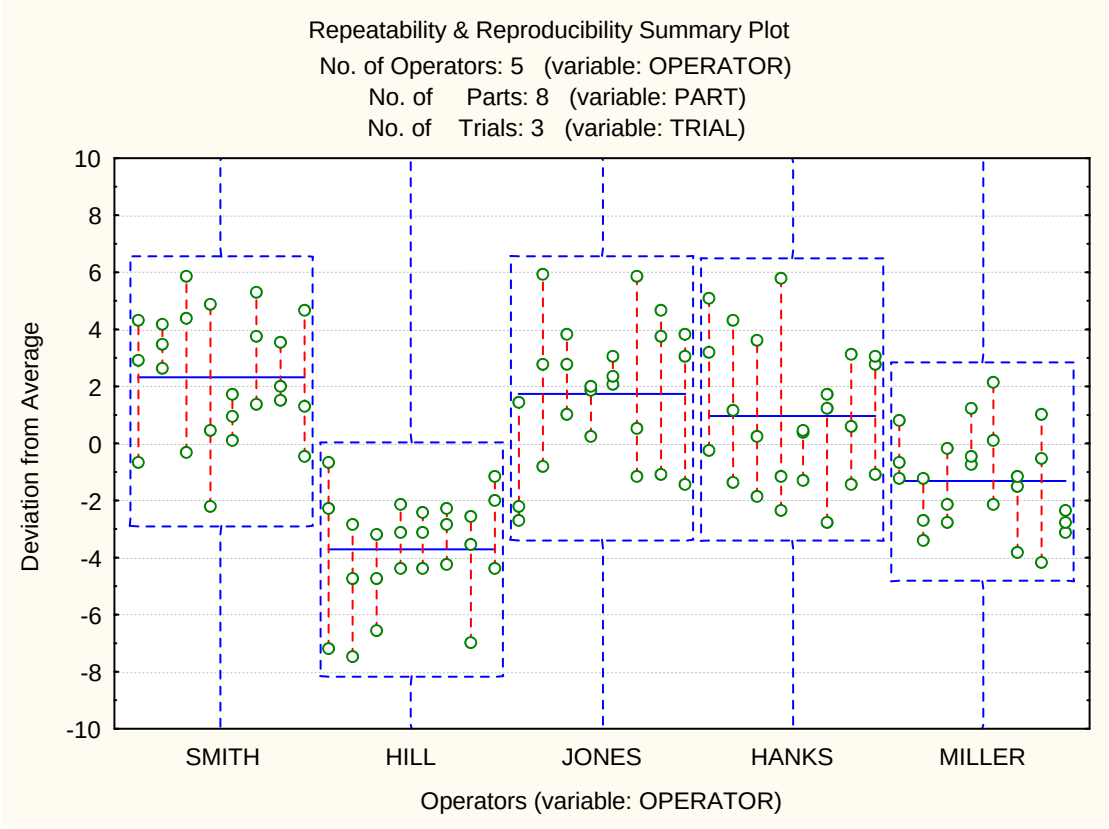


Рисунок 43 – Сводный график результатов анализа

Здесь среднее в каждом «ящике» (для данного оператора) изображено сплошной горизонтальной линией; можно видеть, что оператор Хилл явно занижает результаты. Высота «ящика» показывает разброс результатов (наименьший разброс - у Миллера).

Для идеального случая полностью повторяемых измерений результаты для каждого образца будут совпадать – тогда вертикальные линии выродятся в точки; для полностью воспроизводимых измерений все операторы дадут одинаковые результаты, поэтому по отношению к оси OX все «ящики» расположатся одинаково.

5. Завершить анализ без сохранения результатов.

4.7.2 Построение X- R карт

Задача. Построить карты контроля ширины (***Width)*** детали (пластиковой крышки) по результатам обработки 20 выборок, по 3 измерения в каждой.

Порядок работы.1.Открыть в папке ***Datasets*** файл ***Cover.sta.*** В главном меню **Statistics** выполнить команду **Industrial Statistics & Six Sigma ► Quality Control Charts.**

2.В ДО **Quality Control Charts: Cover.sta** на закладке **Quick** выбрать **X-bar & R chart for variables** (X- R карты для переменных) На закладке **Real-time** установить переключатель в положение **Auto-update...,** (Автоматически обновлять), нажать **OK.** В случае установки этой опции построенные карты будут автоматически обновляться при вводе в таблицу данных новых результатов измерений.

3. В ДО **Defining Variables for X-bar & R Chart: Cover.sta** (Задание переменных для КК) на закладке **Quick** нажать **Variables** (Переменные) и в открывшемся ДО выбрать в первом поле переменную ***Width***. Ее номер появится в поле **Measurements:** (Измерения:), нажать **OK**. (Заметим, что кнопки **Spread, Zoom** позволяют получить более подробную информацию о переменной).

4. У поля **Constant sample size** (Постоянный объем выборки) установить флажок и ввести значение **3.** Это же значение ввести в поле **Minimum number of observations per sample** (Минимальное число наблюдений в выборке). В результате 3 последовательных измерения будут соотноситься с одним и тем же образцом (второе значение не может быть больше первого; образцы, для которых число измерений меньше значения ***Minimum…,*** игнорируются). Нажав **OK**, получим карты и гистограммы распределения среднего и размаха (рис. 44).

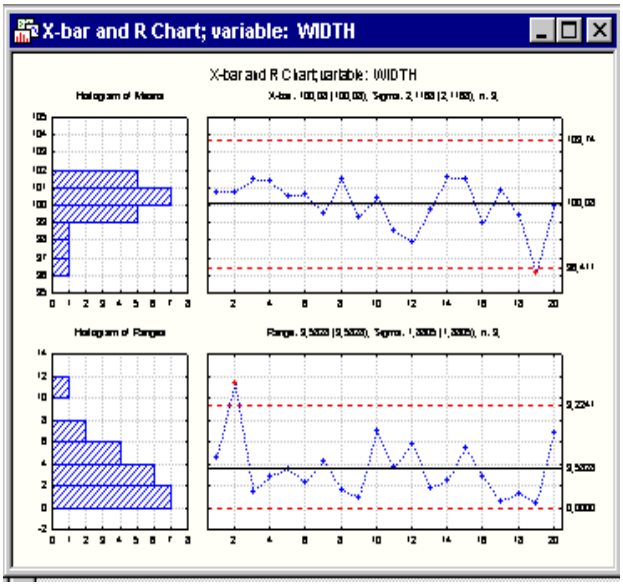


Рисунок 44 – X-R карты

5.Окно диалога **X-Bar/R:WIDTH:Cover.sta** позволяет проконтролировать параметры КК (рис. 45). Нажав на закладке **Charts** кнопку **Runs tests** (позиция 1), получим для каждой карты по таблице. (Если в рабочем окне видна только одна таблица, другую можно вызвать, используя команду главного меню **Windows**).Можно видеть, что все тесты (9 точек по одну сторону от среднего и другие) указывают на удовлетворительное качество процесса.

6. Для получения индексов качества в ДО **X-Bar/R:WIDTH:Cover.sta** на закладке **X-(MA…) specs** нажать кнопку **Process capability**; в ДО **Specifications for Capability Analysis:Cover.sta** убедиться, что параметр **Type** имеет значение **Nominal+-delta**, ввести значение номинала (**100**) в поле **Nominal:** и **10** - в поле **±Delta:**, нажать кнопку **OK (compute)**. Будут получены два окна с таблицами индексов (**Capability Index** и **Performance Index**).

Задание: оценить качество процесса, используя полученные таблицы.

7. Можно также получить гистограмму качества процесса, нажав в ДО **X-Bar/R:WIDTH:Cover.sta** на закладке **Charts** кнопку **Histogram** (поз. 2)**.** На полученной гистограмме будут представлены данные для всех образцов, а не для выборок, как это сделано на гистограммах, расположенных рядом с контрольными картами.

8. Нажав кнопку **Options** (поз.3) и выбрав закладку **Stats**, в поле **Include (display) capability indices in graph, Normal** (Отобразить показатели качества на графике) установить флажки **Cp, Cr, Cpl, Cpu, Cpk.** Можно также добавить границы спецификации, для чего в ДО **Options** на закладке **Layout** установить флажок **Show process specifications and specification limits in chart .**

Нажать кнопку **<OK>** - на изображении карт будут выведены соответствующие индексы.

Окно анализа не закрывать!

4.7.3 Использование технологии “Brushing”

Эта технология используется для выделения непосредственно на графике точек и последующей обработки данных (координат), по которым они построены.

Задача. Выяснить причины выбросов для исследуемого процесса (см. на КК выборки с номерами 2 и 19) и дополнить контрольные карты соответствующей информацией.

Порядок работы. 1.В ДО **X-Bar/R: WIDTH: Cover.sta** (рис.46) на закладке **Brushing** в поле **Include/exclude samples** (Включать, исключать выборки) установить переключатель **All out-of-control samples** (Все неконтролируемые выборки) – позиция 4. Прокрутив список выборок, отметим, что выборки с номерами 2 и 19 выделены цветом. Теперь для образцов, входящих в эти выборки, можно просмотреть результат каждого измерения или получить индивидуальные статистики.

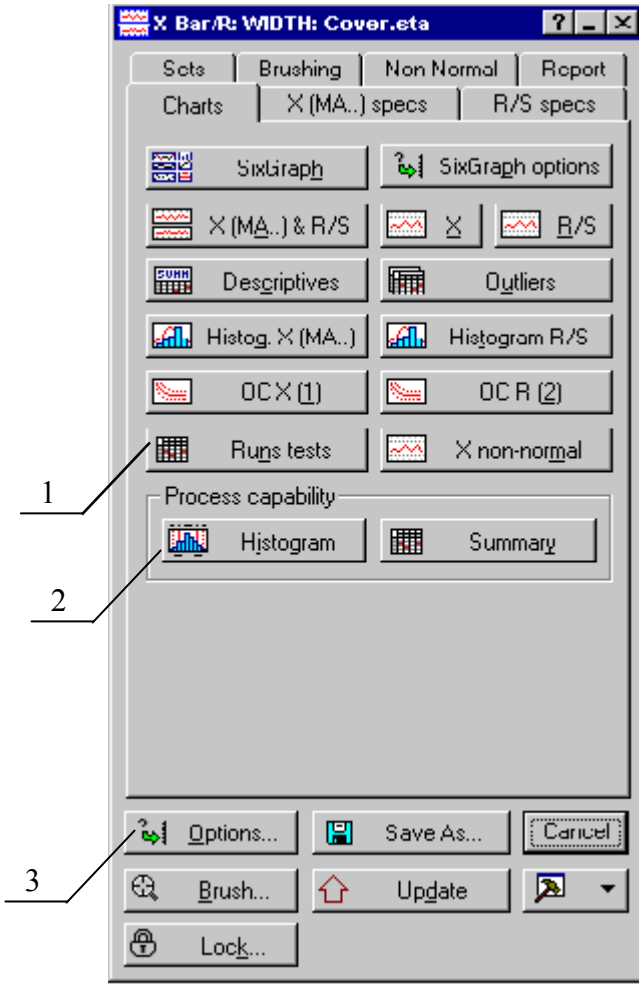


Рисунок 45 – ДО опций анализа

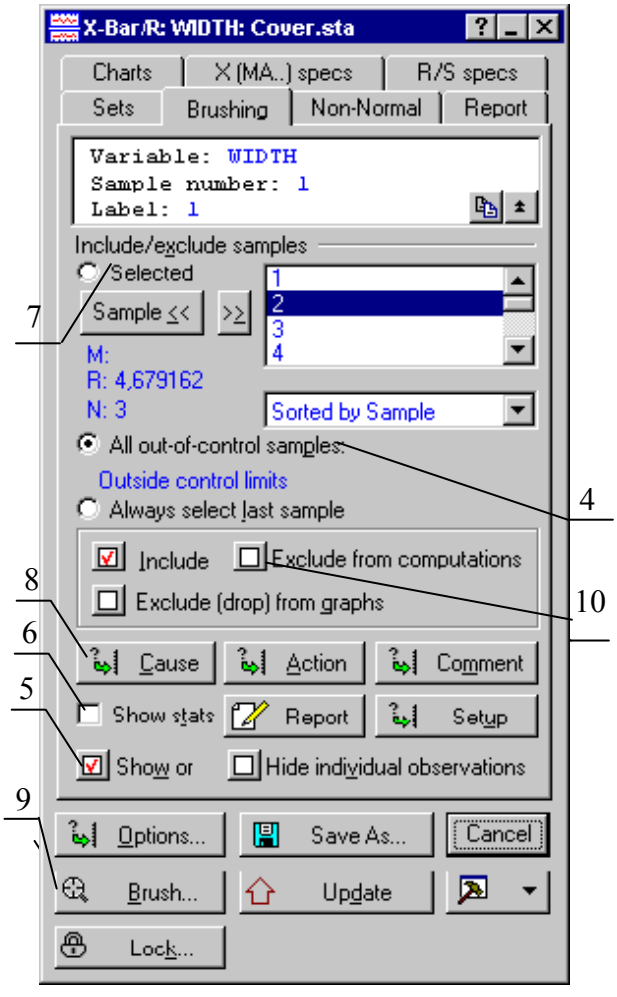


Рисунок 46 –ДО опций “Brushing”

2.Для этого установить переключатель **Show or/Hide individual observations** (Показывать или скрывать индивидуальные измерения) в первое положение (поз. 5). На X-карте будут отображены результаты измерений для каждого образца выборок 2 и 19. Анализируя X-карту, можно видеть большую вариацию измеряемого параметра для образца 2. Единичный выброс «вниз», вызвавший выход R-графика из границ, явился, похоже, результатом ошибки измерения. Зато для образца 19 выброс на X-карте представляется закономерным, вызванным вполне

определенным фактором. Чтобы снова убрать результаты индивидуальных измерений, установить переключатель в положение **Hide individual observations.**

3.Для получения значений индивидуальных статистик для выборки 19 выделить ее щелчком мыши в поле выбора и установить переключатель **Show Stats** (поз. 6)**.** Откроется окно **Sample 19***,* где, в частности, даны среднее значение (Mean) и разброс (R) для этой выборки. Далее сообщается (красным цветом), что образец не укладывается в контрольные границы на X-карте.

4.Пусть причины наблюдаемых отклонений установлены: для образца 2 это ошибка измерительного прибора, для образца 19 – неопытность оператора. Приняты соответствующие меры: прибор настроен, оператор направлен на переподготовку. Занесем эту информацию на график – для этого таблица ***Cover.sta*** содержит пустые переменные ***Causes (Причины)*** и ***Actions (Действия),*** которые и будут использованы. Информация, полученная при анализе причин разладки процесса, может быть закодирована в интерактивном режиме.

5.На закладке **Brushing** активизировать переключатель **Selected** (поз.7) и в списке выделить номер нужной выборки (2). Нажать кнопку **Cause** (поз.8) и в открывшемся ДО **Select vars with codes for causes and actions** (Выбрать переменные с кодами для причин и действий) выбрать в первом слева поле переменную ***Causes*** , во втором - ***Actions*** и нажать **OK**. В ДО **Assigning a Cause:** установить переключатель в поле **Assign cause to** (Указать причину) в положение **R or S (MR)** *(*поскольку образец 2 дает выброс только на R-карте) и нажать **Specify a new cause** (Определить спецификацию новой причины)**.** В ДО **Specify text labels for cause** нажать кнопку **New.** В окне **Edit text label** набрать в белом поле комментарий "Погрешность прибора" и нажать **OK** в двух последовательных окнах для возврата в ДО **Assigning a Cause:**, где также нажать **OK***)*. График будет снабжен соответствующим комментарием. По аналогии можно ввести комментарий ***Action.***

Задание. Создать комментарий для выборки 19. При выборе карты для комментария переключатель в поле ***Assign cause to*** *установить в положение* ***X-bar****…*

6. Нажать кнопку **Brush** (поз. 9) в ДО **X-Bar/R:WIDTH:Cover.sta.** Если навести курсор на любую точку графика, то, когда он превратится в инструмент «кисть» (лупа с перекрестием), появится рамка со статическими данными для выбранного образца. При этом откроется ДО, управляющее режимом "Brushing".

7. Закрыть ДО **Brushing commands**. На закладке **Brushing** в поле выбора выделить выборку номер 2*,* нажать клавишу **CTRL**, затем выделить выборку номер 19. Чтобы исключить выборку из рассмотрения, нужно установить переключатель **Exclude from computations** (Исключить из расчета), расположенный в средней части ДО (поз.10). В открывшемся ДО выбора переменных выделить переменную **Exclude** , нажать **OK**. Произойдет обновление графика с пересчетом контрольных пределов. (Если переключатель установить в положение **Exclude (drop) from graphs***,* точки будут удалены с графиков без пересчета пределов.) Для возврата исключенных переменных нужно установить переключатель в позицию **Include**.

8. Закрыть окно анализа.

4.7.4 Карты контроля по альтернативному признаку

Карты этого типа в программно-статистическом комплексе **Statictica** обозначаются термином **"Charts for attributes"** (карты для атрибутов).

Задача. Проверено 30 выборок, содержащих по 100 изделий. После испытаний в переменную ***Defects*** записали количество дефектных изделий в каждой выборке. Изобразить процент дефектов для различных образцов в виде **P-карты.**

Порядок работы.1.Открыть файл ***Bulbs.sta.*** В главном меню **Statistics** выполнить команду **Industrial Statistics & Six Sigma►Quality Control Charts.**

2.В ДО **Quality Control Charts:Bulbs** на закладке **Quick** выбрать **P chart for attributes.** На закладке **Real-time tab** установить переключатель в положение **Auto-update...,** нажать **OK .** В ДО **Defining variables for P (Attribute) chart** оставить переключатель в положении **Counts (divide measures by sample size to compute rates or p)** (Количество (делить число измерений на объем выборки для расчета отношений или частот).

3. Нажав на закладке **Quick** кнопку **Variables**, щелкнуть в первом окне на переменной ***Defects***, тем самым выбрав ее в качестве переменной ***Counts or proportions:*** (Количества или отношения), нажать **OK**. В поле **Constant sample size:** ввести объем выборки(**100**), нажать **OK** .

4.На полученной карте по оси абсцисс отложены номера выборок, по оси ординат – число дефектов в них. Здесь же приведены гистограммы числа дефектов. Можно видеть, что параметры выборок не выходят за контрольные пределы. Нажав на закладке **Charts** кнопку **Runs tests**, получим предупреждение ***«2 из 3»*** для образцов от 20 до 22 (2 из 3 расположенных подряд точек попадают в зону А или выходят за ее пределы), что является индикатором разладки процесса.

5.В стандартных КК Шухарта контрольные пределы рассчитываются обычно с использованием значения "3 Sigma". Для нормально распределенного случайного параметра это соответствует попаданию 99.73% образцов в интервал между нижним и верхним пределами (это соответствует значению p=0.0027). Можно при необходимости задать другие контрольные пределы.

Пусть процесс признается контролируемым, если дефектных узлов менее 10% от общего их количества. Зададим указанную величину верхнего контрольного предела. Для этого на закладке **Specs** нужно нажать кнопку **UCL:** , чтобы открыть ДО **Upper Control limit: Bulbs**. (Верхний контрольный предел). Установив опцию **Specific value** (Заданное значение), ввести в поле значение **0.1**, нажать **OK** . Затем нажать кнопку **Update** – КК будет обновлена с новыми пределами.

6. Закрыть окно анализа.

4.7.5 Карты Парето

Задача. Построить карту Парето для переменной ***No_defct***, содержащей распределение количества дефектов изделий по их типам.

Порядок работы. 1. Открыть файл ***Circuits.sta.*** В главном меню **Statistics** выполнить команду **Industrial Statistics & Six Sigma►Quality Control Charts.** На закладке **Quick** выбрать **Pareto chart analysis.** Задать режим автоматического обновления карты. В ДО **Defining Variables for Pareto Chart** задать переменную **1-NO_DEFCT** и формат данных, для чего в поле **Format of input data:** установить переключатель в положение **Aggregated counts…**, соответствующее распределению числа дефектов по типам. Нажать кнопку **OK.**

2. Построенная карта по умолчанию получила имя переменной (NO_DEFCT). Для того, чтобы дать карте желаемое имя, необходимо в ДО нажать кнопку **Options**, на закладке **Layout** в поле **Project header:** ввести новое название (например, "Распределение дефектов по типам") и установить переключатель в положение **Include project header line in titles** (Включать название проекта в заголовки). После нажатия кнопки **OK** карта получит заданное имя.

3. Чтобы получить распределение дефектов в %, необходимо на закладке **Quick** установить **Show % to label columns** (Показывать % над столбцами) и нажать кнопку **Update.**

Контрольные вопросы

1. С какой целью выполняется анализ повторяемости и воспроизводимости метода измерения?

2. Какие исходные данные необходимо задать для выполнения анализа повторяемости и воспроизводимости?

3. Как выглядит сводный график результатов анализа в случае полностью повторяемых результатов измерений? В случае полностью воспроизводимых?

4. Из какого раздела меню запускается процедура построения карт контроля качества?

5. Как задается опция автоматического обновления контрольных карт?

6. Каково назначение технологии “Brushing”?

7. Каким термином в ПСК Statistica обозначаются карты контроля по альтернативному признаку?

8. Какая информация отображается на картах Парето?

4.8 Лабораторная работа №8
Построение карт контроля качества. Дополнительные возможности

Цель работы:
1) изучить дополнительные возможности работы с X-R картами;
2) научиться строить контрольные карты для скользящего среднего и размаха, кумулятивной суммы;
3) научится строить карты для индивидуальных измерений.

4.8.1 Построение карт для различных партий изделий

Задача. В результате внедрения нового оборудования контролируемый параметр детали изменился, и процесс вышел за ранее рассчитанные контрольные пределы. Установив точку смещения процесса, разбить исходную партию образцов на две новые и проконтролировать для них состояние процесса по отдельности.

Порядок работы.1. Построить контрольные карты для переменной ***Length*** из файла ***Cover,*** относя к каждой выборке по три образца. Можно видеть, что, хотя размах процесса (по **R-карте**) находится под контролем, на **X-карте** результаты для многих выборок выходят за контрольные пределы (рис. 47).

2.В ДО **X-Bar/R:LENGTH:Cover**, на закладке **Sets,** нажать верхнюю кнопку **Make-a-new-Set Wizard** («Мастер» новой партии). В открывшемся ДО **Label (name) for the new set of samples? Cover:** (Метка (название) для новой партии) ввести название первой партии (***"Старое оборудование"***) и нажать кнопку **Next**. Убедиться, что в новом ДО **Compute the set from range or codes? Cover:** в поле **Compute statistics from** (Рассчитывать статистики, исходя из) выбрана опция **Range of consecutive samples** (Диапазона последовательных выборок) и снова нажать **Next.** В следующем открывшемся ДО **Specify the range of samples for computing the set stati…** (Задать диапазон выборок для расчета...) необходимо разбить все имеющиеся выборки на две партии («Старую» и «Новую»). Для этого ввести значения *1* и *10* в поля **From sample:** (С выборки) и **To sample:** (По выборку) соответственно и нажать **Next**. Другой способ: нажать кнопку **Select range by dragging (on graph),** чтобы ввести диапазон в режиме **brushing.**

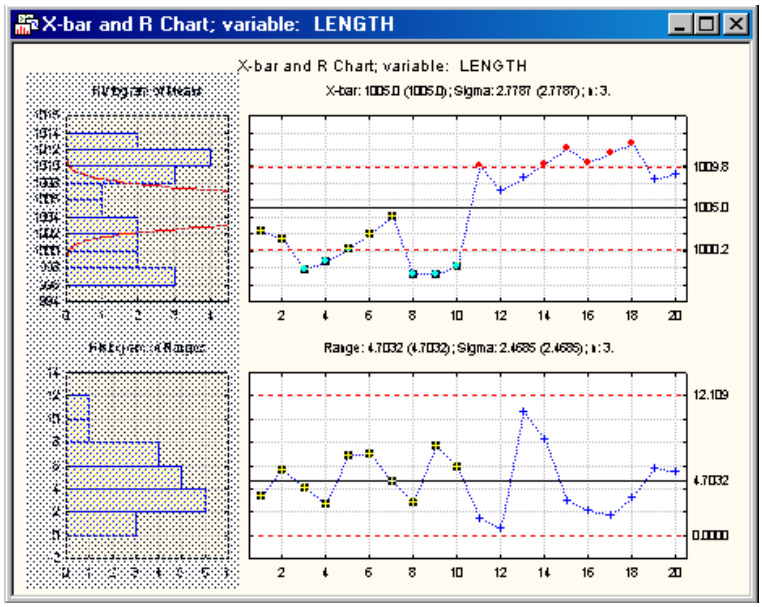


Рисунок 47 – Исходные контрольные карты

После этого нужно протащить указатель мыши при нажатой левой клавише по соответствующему диапазону оси абсцисс, который в результате выделится цветом, и в открывшемся ДО **Brushing Commands** нажать кнопку **OK**.

3. В следующем ДО **Apply the set statistics and specifications…** в поле **Apply statistics and specifications to** (Применить статистики и спецификации для) установить переключатель в положение **Range of consecutive samples** и нажать **Next.**

В новом ДО **Specify the application range for the set: Cover** (Задать применяемый диапазон...) снова задать тот же диапазон и нажать **Finish**.

4.Повторить п.п.2-3 для образцов с 11 по 20, задав в качестве названия новой партии ***«Новое оборудование».***

5.Нажать кнопку **Options**, выбрать в ДО **Options** закладку **Layout**, активизировать второй снизу переключатель **Identify sets of samples (with separate specs) in the chart** (Идентифицировать серии образцов с разными спецификациями на карте) и нажать **OK**. Будут построены две пары карт с разными средними и пределами. Можно видеть, что процесс действительно сместился после внедрения нового оборудования, но каждая партия в отдельности находится под контролем (рис. 48).

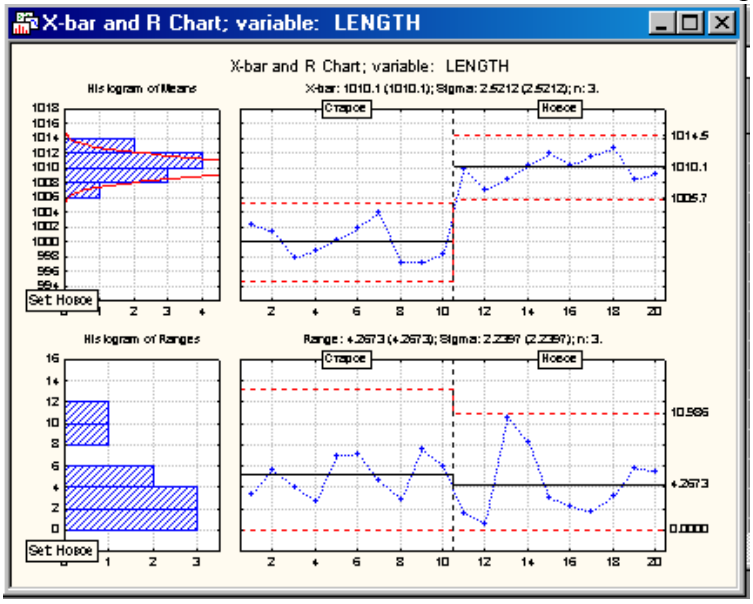


Рисунок 48 – Карты для разных партий изделий

6.Наконец, можно получить таблицу результатов для каждой новой партии. Для этого на закладке **Sets** нажать кнопку **Summary specifications for all sets** (четвертую снизу в диалоговом окне)**.** Данные будут соответствовать последней заданной партии (в исходной партии объектов не осталось - они распределились между двумя новыми партиями).

Чтобы просмотреть данные по другой партии (образцы с 1 по 10), нужно на закладке **Sets** в верхнем поле выбора с помощью кнопки **Set>>** задать партию «Новое оборудование» и нажать кнопку **Delete** (над кнопкой **Summary specifications for all sets**).

7. Закрыть окно анализа.

4.8.2 Карты контроля процесса на рабочем месте

Такие карты имеют упрощенный интерфейс, который позволяет оператору, контролирующему производственный процесс, вводить результаты новых измерений, но не допускает вводить программные изменения. При выходе процесса из-под контроля предусматривается автоматическая выдача предупреждающего сообщения.

Задача. Построить карты для контроля на рабочем месте размера детали, записываемого в переменную Size.

Порядок выполнения. 1.Открыть файл ***Pistons***. Поскольку при построении карт в таблицу данных будут вноситься изменения, то, во избежание порчи исходного файла, сохранить файл с именем ***Student***.

2. С помощью команды главного меню **Insert / Add Variables…**добавить две переменные после переменной ***Size*** (для этого в поле **How many:** открывшегося ДО ввести значение **2**, а поле **After:** - имя ***Size***). Назвать новые переменные ***Causes*** и ***Actions***, для чего использовать в контекстном меню команду **Variable Specs…** С помощью команды **Delete Variables…** удалить переменную ***Samples.***

3. С помощью команды главного меню **Insert / Add Cases…**добавить **10** «пустых» наблюдений после наблюдения **№ 125**.

4. Построить **X-R** карты для переменной ***Size***, задав количество образцов в каждой выборке равным **5.** Процесс находится под контролем.

5.Пусть заданы следующие номинальные характеристики процесса: среднее значение равно **74**, среднеквадратическое отклонение **0,01**. Задать эти значения, для чего на закладке **X (MA…) specs** нажать кнопку **Center** и ввести соответствующие значения в поля **Process mean:**

и **Sigma:** После нажатия кнопки **Update** карты будут перестроены с новыми контрольными границами.

6. Запрограммировать автоматическую выдачу сообщения о выходе процесса из-под контроля. Для этого на закладке **Brushing** нажать кнопку **Setup** (в пятом снизу ряду). В ДО **Causes, Actions, Comments, Data Brushing Setup...** нажать первую сверху кнопку **Variables containing Causes and Actions** (Переменные, содержащие причины и действия). В открывшемся диалоговом окне выделить в первом поле переменную **Causes**, во втором - **Actions**. Нажать кнопку **OK** последовательно в двух диалоговых окнах.

7. Нажать кнопку **Options** и в открывшемся ДО **Options** перейти на закладку **Alarm** (Тревога). Активизировать переключатель **Sample out of control** (Выборка вне контрольных пределов) и нажать кнопку **Define automatic action** (Задать действие, выполняемое автоматически), как показано на рис. 49.

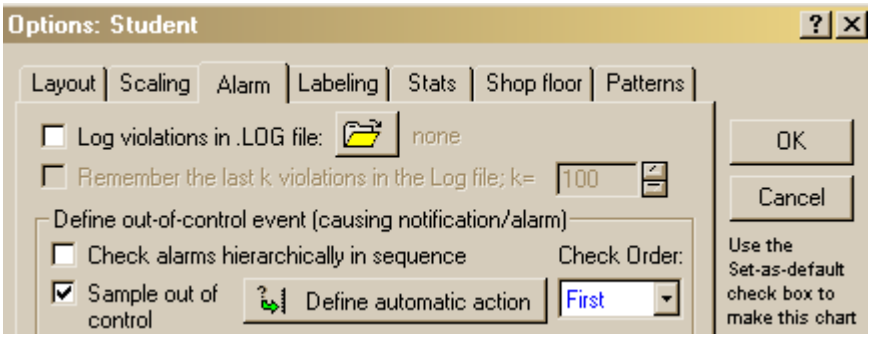


Рисунок 49 – Фрагмент окна задания опций

В открывшемся диалоговом окне активизировать переключатели **Automatically request input of cause** (Автоматически запрашивать ввод причины) и **Automatically request input of action** (Автоматически запрашивать ввод действия), как показано на рис. 50.

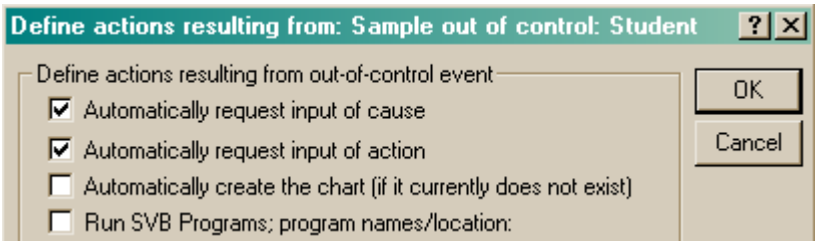


Рисунок 50 – Фрагмент окна задания «тревожных» действий

Отметим, что при активизации расположенного здесь же переключателя **Run / lunch other application…** (Исполнить / запустить другое приложение...) в случае выхода процесса из-под контроля будет автоматически запущена заданная программа – например, отправлено сообщение по E-mail и т.д.

Дважды нажать кнопку OK, чтобы вернуться в ДО **X-Bar/R: SIZE: Student.**

8. После проделанных подготовительных действий можно защитить все сделанные установки от несанкцинированного их изменения и снабдить оператора упрощенным интерфейсом. Пусть из всех полученных данных о процессе на картах нужно отражать только последние 25 результатов (как наиболее актуальные). Для этого, вызвав ДО **Options,** на закладке **Layout** активизировать переключатель **Plot subset of samples in control chart** (Отображать на КК подмножество измерений) и переключатель **plot last N samples only** (только последние N) и установить значение **25** (рис. 51).

9.В том же ДО опций перейти на закладку **Shop floor** (Цех). Не меняя других установок, ввести какое-либо слово (число) в поля **Password** (Пароль) и **Verify** (Подтвердить).

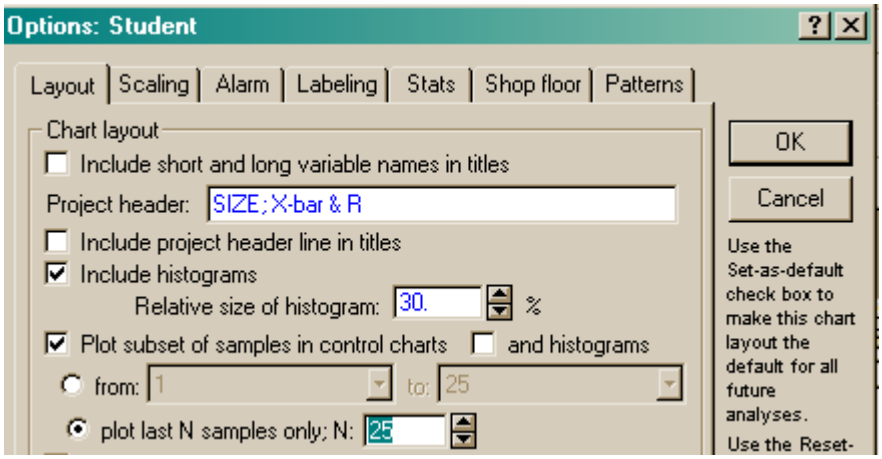


Рисунок 51 – Фрагмент окна задания атрибутов карты

Нажать кнопку **Lock** (Запереть). В ответ на предупреждение системы, что текущий анализ будет «заперт» нажать кнопку **ДА**. Окно задания опций закроется, а в окне **X-Bar/R: SIZE: Student**, управляющем анализом, многие опции будут дезактивированы (например, на закладке **X (MA…) specs** недоступны кнопки **Center:** и **Sigma**). Чтобы вновь сделать все опции активными, нужно нажать кнопку **Unlock** (Отпереть) в нижней части **X-Bar/R: SIZE: Student** и ввести пароль.

10. Пусть получены новые результаты. Ввести в строки 126-130 таблицы данных для переменной **Size** значения

74.001 74.012 73.992 74.002 74.001.

(Если таблица не видна на экране, нужно использовать команду главного меню **Windows**).

С вводом последнего значения карты будут автоматически обновлены. Процесс пока остается под контролем.

11. Ввести в строки 131-135 значения:

74.003 73.978 74.5 74 74.59

Так как процесс при этом выйдет из-под контроля, то будет запрошено определение комментариев. Вначале откроется окно **Assign a cause:**, в котором следует задать описание причины разладки процесса, а затем – окно **Assign an action:** для описания предполагаемого корректирующего действия. После ввода комментариев (см. п. 2.4.2 настоящих указаний) они отобразятся на картах.

4.8.3 Контрольные карты для скользящего среднего и скользящего размаха

Контрольные карты этого типа полезны для выявления восходящего или нисходящего тренда, или сдвига, контролируемого параметра, который желательно обнаружить на возможно более ранней стадии.

Задача. Построить КК для определения тренда в среднем или размахе средней ширины изделия, содержащихся в файле. **Cover.sta**.

Порядок работы. 1.Открыть файл. В главном меню **Statistics** выполнить команду **Industrial Statistics & Six Sigma►Quality Control Charts.** На закладке **Variables** выбрать вариант **MA X-bar & R chart for variables.** На закладке **Real-time** выбрать **Auto-update**, нажать **OK** .

2.В ДО **Defining Variables for MA X-bar and R Chart: Cover** нажать на закладке **Quick** кнопку **Variables** и выделить ***Width*** в левом поле (**Measurements:**), нажать **OK**. Так как выборки содержат по 3 образца, установить опцию **Constant sample size** и ввести значение **3**.

3.Далее нужно задать интервал для вычисления скользящего среднего (СС). Величина интервала определяет, сколько последовательных измерений используется для расчета СС. На-

пример, при значении интервала **=3** первое СС будет рассчитано для выборки, включающей образцы с 1 до 3, второе – со 2 по 4 и т.д. Чем больше интервал, тем более гладкой будет кривая. Задать в поле **Moving average span** (шаг скользящего среднего) значение **5**, нажать **ОК**.

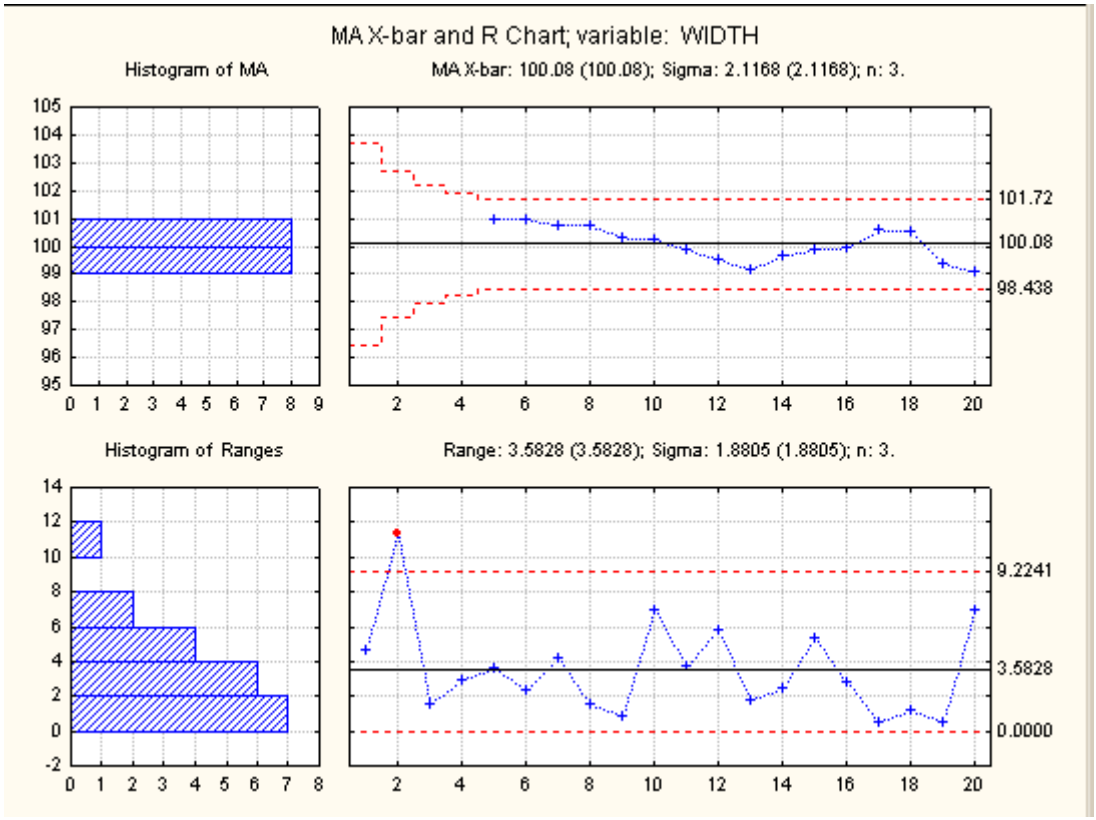


Рисунок 52 – Карты для скользящего среднего и скользящего размаха

Отметим, что значение среднего откладывается на карте, начиная с пятой точки. Можно видеть, что, хотя на графике СС нет точек вне контрольных пределов, но имеется два разных участка: вначале СС убывают, затем начинают возрастать. Можно установить причину убывающего тренда для первых 9 выборок, исключить ее и тем скорректировать процесс. (Заметим, однако, что любая интерпретация субъективна, иначе говоря, решение о корректировке процесса должно основываться не только на исследовании КК, но также на знании технической стороны процесса.)

4.8.4 Контрольные карты для индивидуальных измерений и скользящего размаха

Эти карты строятся в случае, когда измерения нельзя группировать, как это делается, например, для X-R карт, и приходится анализировать отдельные наблюдения с использованием скользящего размаха для последовательности наблюдений.

Задача. Проанализировать выход химического процесса; измерения количества синтезируемого вещества производятся с интервалом в один час и заносятся в переменную ***Var2*** файла ***Individ.sta.***.

Порядок работы. 1.Открыть указанный файл. Выполнить в меню **Statistics** команду **Industrial Statistics & Six Sigma►Quality Control Charts.** В ДО **Quality Control Charts: Cusum** на закладке **Quick** выбрать **Individuals & moving range** (Индивидуальные измерения и скользящий размах), на закладке **Real-time** установить опцию **Auto-update.** Нажать кнопку **OK** .

2. В ДО **Defining Variables for X and moving range** задать переменную ***Var2.*** Отметим, что здесь не требуется устанавливать размер выборки для построения каждой точки карты. Дважды нажать кнопку **OK.** На полученных картах (рис. 53) значение самой переменной ***Var2***

отображено для каждого измерения, а значение размаха (начиная со второго измерения) определено, по умолчанию, по двум соседним наблюдениям. Можно видеть, что процесс находится в контролируемом состоянии.

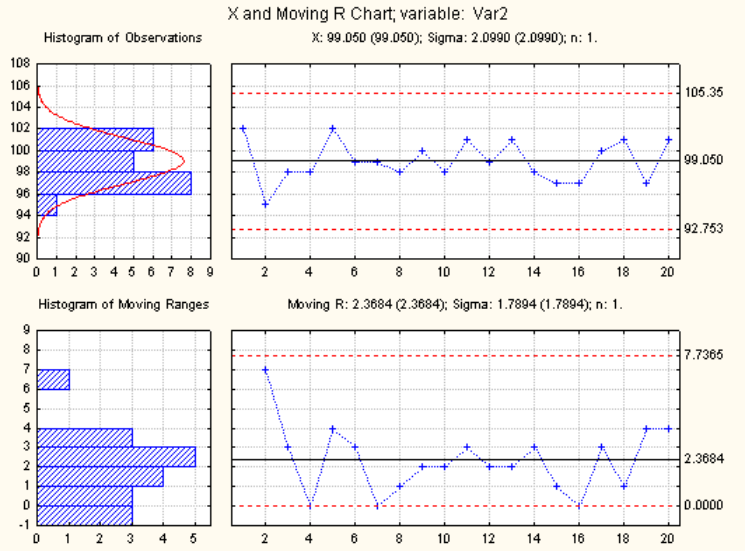


Рисунок 53 – Карты для индивидуальных измерений и скользящего размаха

В заключение отметим, что карты такого типа строятся в предположении, что результаты последовательных измерений некоррелированы. В противном случае следует использовать процедуру ARIMA для построения модели временного ряда и строить карты для остатков этого ряда после устранения систематической составляющей.

3. **Окно анализа не закрывать!**

4.8.5 Контрольные карты для кумулятивной суммы

Достоинством карт этого типа является возможность обнаружить малое отклонение среднего значения контролируемого параметра от его среднего значения, то есть дрейф. В частности, полезно строить данные карты совместно с картами для индивидуальных измерений и скользящего размаха, поскольку индивидуальные КК неспособны отражать малые изменения контролируемого параметра.

Задача. Построить карты для выхода анализируемого химического процесса.

Порядок работы. 1.В диалоговом окне, управляющем анализом, нажать кнопку **Cancel**. В окне задания переменных для анализа также нажать **Cancel.** Произойдет возврат в ДО выбора типа контрольных карт. На закладке **Variables** выбрать **CuSum chart for individuals**, на закладке **Real-time** установить опцию **Auto-update.** Нажать кнопку **OK** .

2.В ДО задания переменных выбрать ту е переменную ***Var2.*** Нажать **OK.** Можно видеть (рис. 54), что карта для скользящего среднего не изменилась. В то же время на верхней карте вместо значения контролируемого параметра откладываются накопленные отклонения его значений от среднего.

3. Пусть необходимо задать номинальное значение контролируемого параметра, равное 100. Для этого на закладке **X (MA..) specs** нужно нажать кнопку **Center**:. В открывшемся ДО **Chart center line** (Центральная линия карты) ввести в поле **Process mean** (Среднее процесса) значение **100** , нажать **OK.** Нажать кнопку **Update**. Обновленные КК показывают, что процесс по-прежнему находится под контролем.

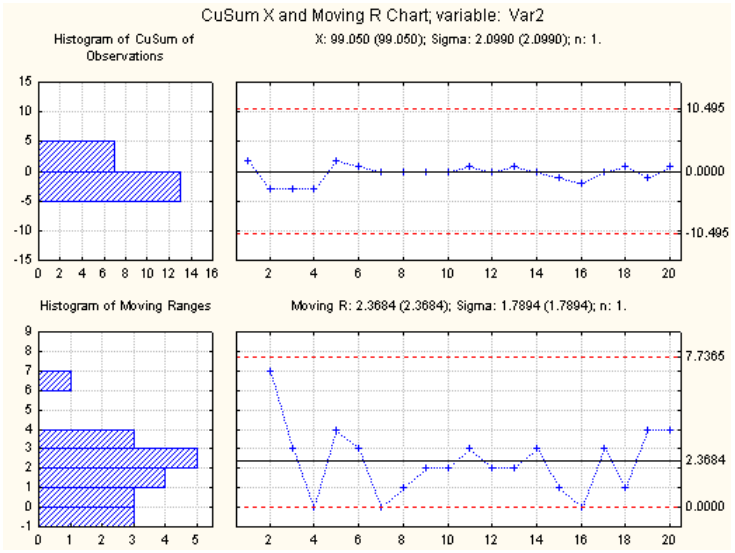


Рисунок 55 – Карты для кумулятивной суммы и скользящего размаха

4. Для расчета описательных статистик нажать на закладке **Charts** кнопку **Descriptives** . Будут выведены таблицы результатов для карты кумулятивной суммы и скользяще-го размаха.

Контрольные вопросы

1.В каких случаях следует применять прием построения контрольных карт для различных партий изделий?

2.Какие дополнительные настройки характерны для карт контроля на рабочем месте?

3.В каком случае целесообразно использовать контрольные карты для скользящего среднего и скользящего размаха?

4.Как вычисляется скользящее среднее?

5.Для какой цели применяется контрольная карта для индивидуальных измерений?

6.Какую процедуру статистического анализа следует применить в случае, когда результаты индивидуальных последовательных измерений коррелированы?

7.Какой тип контрольных карт целесообразно использовать для выявления малых отклонений контролируемого параметра от его среднего значения?

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1.А.Афифи, С.Эйзен. Статистический анализ. Подход с использованием ЭВМ.-М.: Мир, 1982.-488 с.

2.Кулаичев А.П. Методы и средства анализа данных в среде Windows. STADIA. Изд. 4-е.- М: Информатика и компьютеры, 2002. – 341 с.

3. Боровиков В.П. Программа STATISTICA для студентов и инженеров.-М.: КомпьютерПресс, 2001.-301 с.

4.Дюк В. Обработка данных на ПК в примерах – СПб: Питер, 1997. – 240 с.

Учебное издание

ПРОГРАММНЫЕ СТАТИСТИЧЕСКИЕ КОМПЛЕКСЫ.

Лабораторный практикум

Составитель: *Кучеров Александр Степанович*

Самарский государственный аэрокосмический университет
имени академика С.П.Королева
443086, Самара, Московское шоссе, 34