МИНОБРНАУКИ РОССИИ

––––––––––––––––––––––––––––––––––––––––––––––––––––

Санкт-Петербургский государственный электротехнический

университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

––––––––––––––––––––––––––––––––––––––––––––

**АНАЛИЗ И ИНТЕРПРЕТАЦИЯ ДАННЫХ**

Учебно-методическое пособие

Санкт-Петербург

Издательство СПбГЭТУ «ЛЭТИ»

# 1. Знакомство с системой STATISTICA

**Общая структура системы *Statistica.*** *Statistica* представляет собой интегрированную систему статистического анализа и обработки данных. Система *Statistica* содержит многофункциональную систему для работы с данными, мощную графическую систему для визуализации данных и результатов статистического анализа, набор статистических модулей, в которых собраны группы логически связанных между собой статистических процедур обработки данных, специальный инструментарий для подготовки отчетов, встроенные языки, позволяющие расширять систему.

**Запуск системы *Statistica.*** *Statistica 10* представляет собой удобный математический пакет, не требующий установки. После запуска приложения, пользователю предоставляется возможность выбрать наиболее удобный на его взгляд и привычный вид панели инструментов.

**Начало работы в *Statistica.*** После запуска программы пользователю предлагается выбрать, откуда загружать исходные данные (открыть таблицу соответствующего типа).

**Справочная система *Statistica.*** Для получения справки о назначении тех или иных диалогов, меню, методов и других элементов пользовательского интерфейса предназначена справочная система *Statistica*. Вызвать ее можно одним из следующих способов:

–воспользоваться пунктом выпадающего меню Справка – Справка.

–контекстно-зависимый способ: для получения информации о конкретной команде меню необходимо выделить ее и нажать кнопку F1.

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

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

В новой версии *Statistica10*выбор модуля можно осуществлять посредством элементов управления на вкладках *Анализ* и *Добыча данных* (*DataMining).*

После запуска модуля на экране откроется основное окно системы *Statistica*. При запуске системы в него автоматически загружается последний файл, с которым в ней работали.

**Обработка данных в системе *Statistica.*** Статистический анализ данных в системе *Statistica* может быть разбит на следующие основные этапы:

1. ввод исходных данных в электронную таблицу системы *Statistica*;
2. предварительные преобразования данных перед непосредственным применением конкретного статистического метода;
3. визуализация данных при помощи того или иного типа графиков;
4. статистический анализ при помощи некоторого статистического метода;
5. подбор модели и задание необходимых параметров в статистических процедурах;
6. вывод численных, текстовых и графических результатов;
7. подготовка и печать отчетов;
8. анализ результатов;
9. автоматизация рутинных процессов обработки при помощи макрокоманд, языка *SCL* или *StatisticaBASIC*.

**Типы документов в системе *Statistica.*** *Statistica* использует пять основных типов Документов:

– рабочие книги;

– таблицы данных (мультимедийные таблицы);

– отчеты;

– графики;

– макросы (*STATISTICA Visual Basic*).

*Statistica* предоставляет возможность открывать и одновременно использовать несколько Таблиц данных. Кроме хранения данных,*Statistica* использует Таблицы для отображения численных результатов Анализов и других данных (например, мультимедийных объектов, скриптов, связей). Поскольку в *Statistica 6.0* больше не существует различий между [исходной Таблицей данных](Glossary.chm::/GlossaryTwo/I/InputSpreadsheet.htm) (*Spreadsheet*) (на основе которой *Statistica* проводит анализ) и [Таблицей результатов](Glossary.chm::/GlossaryTwo/O/OutputSpreadsheet.htm) (*Scrollsheet,* численные и текстовые результаты статистического анализа), то возможно напрямую использовать результаты одного Анализа в качестве исходных данных для другого Анализа.

Любая Таблица данных, открытая из файла, является исходной Таблицей данных. Можно открыть несколько Таблиц данных одновременно. Во избежание ошибок, Таблицу результатов (содержащую результаты Анализа) нельзя сразу же использовать в качестве исходных данных для другого анализа. Перед тем как использовать ее в Анализе, необходимо изменить ее представление.

В предыдущей версии системы *Statistica* таблица результатов представляла собой электронную таблицу с возможностью прокрутки *Scrollsheet*. В этих таблицах выводились результаты анализов, которые можно было сохранить с расширениями .scr. В данной версии результаты анализа выводятся также в таблицах результатов, но формат этих таблиц ничем не отличается от таблиц исходных данных. Вы можете использовать файл с расширением .scr в системе *Statistica*, при открытии файла .scr система автоматически переформатирует его в .sta.

Чтобы сохранять совместимость с программами *STATISTICA BASIC*, написанными в ранней версии *Statistica*, функции *STATISTICA BASIC*, использующие таблицы результатов (таблицы с расширение .scr), все еще имеются в языке *STATISTICA Visual Basic*. Когда вы записываете макрос или читаете справку SVB, термин таблица результатов (*scrollsheet*) подразумевает электронную таблицу. Основное принципиальное различие в назначении таблиц в старой версии: *Spreadsheet* – для задания исходных данных, а *Scrollsheet* – для вывода результатов.

**Создание файла с исходными данными.** Создание нового файла может быть осуществлено двумя способами:

– на вкладке *Главная* нажать значок *Создать*(в раскрывающемся списке можно выбрать требуемый объект данных;

– автоматическое создание объекта данных можно осуществлять при нажатии комбинации клавиш CTRL+N.

В системе *Statistica* строки таблиц или записи называются Наблюдениями – Cases (случайные величины), столбцы таблицы (поля)– Переменными – Variables (значения случайных величин). Например, таблица может состоять из 10 переменных, которые по умолчанию имеют имена var1, var2, … var10 и 10 пронумерованных случаев, которые не имеют имен.

**Основные операции над переменными.** Для задания имен переменных и их свойств можно использовать способы, описанные далее.

Существуют следующие способы вызова основных команд:

– кнопка Переменные на панели инструментов Данные;

– двойным кликом по существующей переменной в таблице или по нажатию кнопки Спецификации можно перейти в редактирование данной переменной.

Редактирование всех переменных централизованно можно осуществлять по нажатию кнопки «Все спецификации».

Все возможные действия с переменными и наблюдениями можно осуществлять посредством инструментов на вкладке Данные (табл. 1).

*Таблица 1*

**Описание инструментов вкладки Данные**

|  |  |  |
| --- | --- | --- |
| Переменные | Команда | Действие |
| Добавить… | Добавляет переменные (пустые строки) в электронную таблицу (ЭТ) |
| Переместить… | Перемещает переменные (столбцы) в ЭТ |
| Копировать… | Копирует переменные (столбцы) в ЭТ |
| Удалить… | Удаляет переменные (столбцы) в ЭТ |
| Спецификации | Основное диалоговое окно, в котором задаются спецификации текущей переменной. Здесь задается имя переменной, ее формат, длинное имя, при помощи которого может быть задана формула, или динамическая связь. |
| Все спецификации | Просмотр и редактирование спецификаций всех переменных в файле данных |
| Текстовые значения | Вызов диалогового окна Менеджера текстовых значений, в котором может быть установлено или отредактировано соответствие между текстовыми и численными значениями переменной при использовании соглашения о «двойной записи». В этом диалоговом окне могут быть заданы и метки для текстовых значений |
| Сдвинуть (Лаг) | Сдвиг значений переменной на один или несколько случаев вперед или назад |
| Преобразования | Операции с датами | Основные операции по работе с датами. Команда позволяет создать дату из значений двух или трех переменных и, наоборот, разбить дату на две или три переменные. Можно также перевести значения даты в числовые значения и наоборот |
| Пересчитать | Команда позволяет пересчитать значения переменных, которые связаны формулами, при изменении данных в ЭТ |
| Ранжировать | Ранжирование значений переменной |
| Перекодировать | Перекодировка значений переменной |
| Транспонировать | Транспонировать выделенный блок значений |

Процедура стандартизации подразумевает следующие шаги:

*Переменные.* Нажмите кнопку «Переменные», чтобы вызвать стандартный диалог «Выберите переменные», который используется для выбора переменных, значения которых будут стандартизованы.

*Наблюдения*. Нажмите кнопку «Наблюдения» для вызова диалога «Условия выбора наблюдений». По умолчанию при вычислении средних и стандартных отклонений учитываются все наблюдения. Однако можно использовать условия выбора наблюдений для вычисления среднего и стандартного отклонений на основе выбранного подмножества наблюдений в активной Таблице данных.

*Веса.* Нажмите кнопку «Веса», чтобы вызвать диалог «Веса» наблюдений Таблицы данных. По умолчанию при вычислении средних и стандартных отклонений используются значения выбранных наблюдений. Вы также можете выбрать весовую переменную, нажав кнопку Веса и указать переменную, значения которой будут рассматриваться как веса. Влияние значений каждого наблюдения будут пропорциональны значениям соответствующих весов.

*ОK*. Нажмите кнопку «OK» для выполнения процедуры стандартизации значений.

*Отмена.* Нажмите кнопку «Отмена», чтобы закрыть диалог, не изменяя данные.

Основные операции над наблюдениями приведены в табл.2.

*Таблица 2*

**Команды по работе с наблюдениями**

|  |  |
| --- | --- |
| Команда | Действие |
| Добавить… | Добавляет наблюдения (пустые строки) в электронную таблицу (ЭТ) |
| Переместить… | Перемещает наблюдения (столбцы) в ЭТ |
| Копировать… | Копирует наблюдения (столбцов) в ЭТ |
| Удалить… | Удаляет наблюдения (столбцов) в ЭТ |
| Диспетчер имен | Задает имена наблюдений |
| Состояние наблюдений | Команда позволяет выбрать подмножество наблюдений из файла данных для анализа. Условие выбора может быть активно и не активно. Статус условия отображается в строке состояния |

**Операции с блоками.** Выделение блока значений производится либо при нажатой правой кнопке мыши, либо при помощи комбинации клавиши SHIFT и стрелок управления курсором. Команды редактирования данных расположены в меню Edit (табл.3).

*Таблица 3*

**Команды редактирования данных**

|  |  |  |
| --- | --- | --- |
| Буфер обмена данных | Команда | Действие |
| Отменить | Отмена последнего действия |
| Вырезать | Вырезает значения выделенного блока и помещает его в буфер обмена |
| Копировать | Копирует блок вместе с именами случаев и переменных и с разделителями сетки |
| Копировать содержание | Копируется только содержание ячеек |
| Очистить | Удаление выделенного блока |
| Вставить | Вставка содержимого буфера обмена |
| Специальная вставка | Вставка содержимого буфера обмена вместе со связью |
| Захват экрана | Помещает содержимое части экрана в буфер обмена |
| Найти/  заменить | Найти | Найти значение |
| Заменить | Заменить значение |
| Заполнить | Заполнить случайными значениями | Заполнить блок случайными значениями |
| Заполнить/копировать вниз | Заполнить значение блока и копировать значения вниз. Копирует первую строку блока вниз |
| Заполнить/копировать вправо | -//- в столбцы |
| Стандартизировать | Стандартизация столбцов | Стандартизация значений в столбцах |
| Стандартизация строк | Стандартизация значений в строках |
|  | BlockStats/ Cols – Блоковые статистики/столбцы | Вычисление основных статистик для столбцов |
|  | BlockStats/ Rows – Блоковые статистики/строки | Вычисление основных статистик для строк |
|  | Links –Связи | Регулировка режимов динамической связи данных |

**Создание отчета.** В некоторых случаях полезно создавать файл регистрации всех результатов анализа, т.е. содержимого всех таблиц и/или графиков, которые появляются на экране. Это позволит просмотреть все результаты последовательности тестов, полученных в данной сессии. Такой файл называется *отчетом*. Для автоматического занесения информации (создания *автоотчета)* необходимо сделать следующее.

Для создания нового Отчета выполните следующие простые действия:

– выберите «Создать...» в меню Файл или нажмите кнопку «Создать» на панели инструментов;

– в диалоге «Создать новый Документ» выберите вкладку Отчет;

– в поле Положение можно выбрать создание нового Отчета как части новой Рабочей книги, или в отдельном окне. Сделайте необходимый выбор и нажмите ОК.

Отчеты являются идеальными контейнерами для Документов *Statistica*, так же как для Объектов из других программ. При добавлении элемента в Отчет не нужно открывать сам элемент. Кроме того, не нужно самому открывать Отчет; тем не менее, для доступа к описанным здесь опциям, необходимо открыть Отчет или другой Документ *Statistica* (например, Таблицу, График или Макрос). Для добавления активного Документа *Statistica* (кроме Рабочей книги) в Отчет используйте кнопку на панели инструментов на вкладке «Главная» «Добавить в Отчет». Когда Отчет является активным, в него можно добавлять Документы и Объекты, используя команды меню Вставка (или горячие клавиши).

**Графические возможности системы *Statistica.*** В системе *Statistica* для визуализации предназначены следующие две основные группы графиков, отличающиеся подмножеством данных, на основе которых эти графики строятся.

1. *Статистические графики* (StatsGraphs – Статистические графики), предназначенные для графического представления значений переменных (столбцов ЭТ) из текущего файла данных. Диалоговые окна этих графиков позволяют выбрать переменные, для которых необходимо построить графики. Они включают в себя большое количество специализированных статистических типов графиков, уже настроенных для статистической визуализации значений переменных. Для быстрого построения некоторых типов *Статистических графиков* существуют так называемые QuickStatsGraphs‑ Быстрые статистические графики. Они включают в себя наиболее употребительные категории и типы *Статистических графиков*. Быстрый доступ к ним осуществляется при помощи контекстного меню, вызываемого при помощи щелчка правой кнопки мыши и в любом столбце ЭТ, для которых строятся графики.
2. *Пользовательские графики* служат для визуализации значений из предварительного выделенного блока в ЭТ. При этом они не содержат ряда специализированных статистических типов графиков. Доступ к ним осуществляется при помощи контекстного меню, вызываемого при помощи щелчка правой кнопки мыши, кнопок на панели инструментов и обычного выпадающего меню. Эта группа графиков включает в себя CustomGraphs– Пользовательские графики и BlockStatsGraphs– Блоковые статистические графики.

Графики в системе хранятся в файлах с расширением \*.stg. После того как график построен, соответствующий графический документ может быть сохранен следующим образом:

– в специальном графическом формате \*.stg системы, который может быть открыт позже и использован в процессе статистического анализа;

– в растровом графическом формате (\*.bmp, \*pcx);

– в графическом формате Windows метафайла (\*.wmf).

Категории Пользовательских графиков.

Описание графиков представлено в табл.4.

*Таблица 4*

**Описание графиков**

|  |  |
| --- | --- |
| Категория  графика | Предназначение графиков этой категории и конкретные типы графиков, которые доступны в этой категории |
| Двумерные  графики | Для отображения на плоскости набора значений вида (Xi, Yi) i=1, 2, …, N.  Возможные типы графиков:  - диаграмма рассеяния (Scatterplot, двойной осью Y, Квантиль, Частоты, Основной, Простой, Диаграмма Вороного Voronoj);  – линейный график (LinePlot);  – диаграмма размаха (Pangeplot);  – столбчатые диаграммы (Bar);  – круговые диаграммы (Pie) |
| Трехмерные  последовательные графики | Для визуализации в пространстве набора последовательностей вида  *i=*1, 2*, …, N*. Каждая такая серия значений отображается на плоскости, после чего эти плоскости выстраиваются в пространстве.  Возможные типы графиков:  – диаграммы исходных данных;  – гистограммы двух переменных;  – диаграммы диапазонов;  – диаграммы размаха |
| Пользовательские графики в  пространстве | Для визуализации в пространстве набора последовательностей вида (*Xi, Yi, Zi*)*i=1, 2, …, N*.  Возможные типы графиков:  –объемная диаграмма рассеяния (Scatterplot);  –график поверхности (SurfacePlot);  –карта линий уровня (ContourPlot) |

*Окончание табл. 4*

|  |  |
| --- | --- |
| Категория  графика | Предназначение графиков этой категории и конкретные типы графиков, которые доступны в этой категории |
| Матричные  графики | Для визуализации связи между переменными в ЭТ. Матричный график состоит из отдельных кадров, в каждом из которых выводится свой конкретный график. На диагонали в этой таблице выводятся гистограммы, а в остальных кадрах этой матрицы – различные типы двумерных графиков:  –диаграммы рассеяния (Scatterplot (matrix));  –линейный график (Line (matrix)) |
| Пользовательские пиктографики | Для визуализации значений нескольких переменных из таблицы. График состоит из нескольких кадров, которые соответствуют случаям. В каждом кадре отображается свой пиктографик. Пиктографик состоит из нескольких элементов, каждый из которых отображает свою переменную.  Возможные типы пиктографиков:  –лица Чернова (ChernoffFaces, Диаграмма лиц Чернова);  –круги (Pies, Круговая диаграмма);  –звезды (Stars, Диаграмма звезды);  –лучи (SunRays, Диаграмма лучевая);  –многоугольники (Poligons, Полигоны);  –столбцы (Column, Столбчатая диаграмма);  –линии (Lines, Линейчатая диаграмма);  –профили (Profiles, Профильная диаграмма) |

**Блоковые статистические графики.** При помощи этих графиков можно построить специализированные статистические графики: различные виды гистограмм, график разброса значений от среднего и медианы, график на нормальной вероятностной бумаге для значений из предварительно выделенного блока в ЭТ. Эти графики доступны из контекстного меню.

**Статистические графики.** Эти графики можно найти на вкладке «Графика» в модуле «Все виды графиков».

Галерея графиков предлагает сотни типов графического представления данных. Они не зависят от выделенного блока или текущей позиции курсора; они обрабатывают данные непосредственно из текущего файла данных.

Основные категории статистических графиков подразделяются следующим образом:

1. Двухмерные графики (2DGrahps).
2. Трехмерные последовательные графики (3DSequentialGraphs)

Каждая группа графиков включает десятки типов конкретных графиков:

–Диаграммы исходных данных (столбцы, блоки, ленты, линии, всплески, дискретная карта линий уровня, поверхность, контуры).

Диаграммы исходных данных позволяют построить 3М-изображение последовательностей исходных данных выбранных переменных. Выбранные переменные представлены на оси *Y*, последовательные наблюдения – на оси *X*, а значения откладываются по оси *Z*. Трехмерные графики исходных данных используются для визуализации последовательностей значений. По своей сущности эти графики сходны с составными линейными графиками с тем лишь отличием, что для 3М-диаграмм исходных данных ленты, линии, параллелепипеды и другие трехмерные представления значений каждой переменной не перекрываются (как на двухмерном графике), а «раздвигаются» в трехмерной перспективе.

– *гистограммы двух переменных*.

– *трехмерные гистограммы* двух переменных используются для визуализации кросстабуляций значений двух переменных. Их можно рассматривать как сочетание двух простых гистограмм (т.е. гистограмм одной переменной), соединенных таким образом, что можно исследовать частоты совместного появления значений двух анализируемых переменных.

– *диаграммы диапазонов* (точечные диапазоны, граничные диапазоны, диапазоны ошибок, диапазоны двойных лент, летящие ящики, летящие блоки).

– *диаграммы размаха* (точечные диапазоны, граничные диапазоны, диапазоны ошибок, диапазоны двойных лент, летящие ящики, летящие блоки).

В основном диаграммы диапазонов и диаграммы размаха используются для одних и тех же задач (на диаграммах диапазонов все значения, определяющие диапазоны, не вычисляются по данным, а являются исходными значениями переменных). В основном эти графики используются для изображения диапазонов значений для отдельных элементов анализа (наблюдений, выборок и т. д.) или вариации значений в отдельных группах или выборках.

1. Графики в пространстве (3DXYZGraphs).

**Настройка графических установок.** Многие опции, которые управляют процессом построения и видом отображения, могут быть заданы при помощи инструментов на вкладке «Сервис» и подвкладке «Параметры». По умолчанию в окне «Параметры» открыта вкладка «Общие».

С помощью диалогового окна, появляющегося при выборе этой команды, можно настроить конфигурацию графического окна системы и сохранить эти параметры как параметры по умолчанию. С их помощью можно изменить практически все аспекты внешнего оформления графического окна (всего более 500 параметров). Вы можете сохранить измененную конфигурацию системы, нажав «Ok».

Также настройка графических параметров становится доступна после двойного нажатия на область отображения графика:

Нажатием на кнопку «Стили» (Styles) можно перейти в диалоговое окно редактирования стилей. Из данного окна есть возможность сохранять свои стили и подгружать их для других графиков.

Щелкнув по «Списку стилей» (Listofstyles…) можно перейти к списку стандартных и сохраненных пользователем стилей для той области графика (задний фон или Ширина линий), которая редактируется.

**Интерактивный графический анализ данных. Кисть, редактор данных графика.** Возможно, что в русскоязычной версии пакета возникнут проблемы с отображением графиков. В связи с этим приводим описание данного раздела для англоязычной версии пакета.

В системе существует удобный инструмент для интерактивного графического анализа данных. Это так называемая *Кисть*. С помощью *кисти* можно легко установить соответствие между точками на графике и их числовыми значениями. Например, вы можете выделить необходимые точки на графике при помощи мыши и далее применить одну из следующих операций: пометить их маркером, временно удалить их из анализа или вывести их метки, перейти в режим просмотра координат этих точек. *Кисть* является мощным средством для анализа зависимостей между переменными, изучения влияния выбросов и т. д.

**Основные статистики/таблицы.** Во вкладке «Анализ» и далее в «Основные статистики и таблицы» объединены процедуры, наиболее часто используемые на начальном этапе обработки данных, когда выясняется структура и зависимости между данными, производится их группировка.

Для использования этого блока требуется выбрать исходные данные.

**Основные элементы диалога в модуле Основные статистики/таблицы (*BASIC*Statistic/Tables).** При выборе в предлагаемом меню Descriptivestatistics– Описательные статистики пред вами откроется диалоговое окно, в котором можно:

– выбрать переменные для анализа (из уже открытого файла);

– вычислить различные описательные статистики;

– оценить близость распределения к нормальному закону;

– построить гистограммы;

– просматривать таблицы частот;

– сгруппировать данные, разбив различными способами на классы;

– удобно визуализировать и т.д.

В этом окне имеется несколько групп кнопок. Следует обратить особое внимание на следующие:

– подробные описательные статистики (Detaileddescriptivestatistics). Позволяет просмотреть всевозможные описательные статистики выбранных переменных;

– дополнительные настройки. Эта вкладка содержит и расширенный набор описательных статистик (позволяет вычислить отмеченные галочками дополнительные статистики).

Здесь можно задать дополнительные статистики:

– медиана и квартиль (Median&quartile);

– доверительные границы для среднего (Conf. Limitofmeans);

– альфа ошибка задает уровень значимости (Alphaerror).

**Структура директорий при работе с системой *Statistica.*** При работе с системой используются файлы трех типов:

– файлы данных с расширением.sta;

– файлы программ с расширением .stb;

– файлы отчетов с расширением .rtf.

Рекомендуемая структура директорий:

диск:\MStat\№\_группы\Фамилия\Lab\_№работы

Пример: c:\MStat\ETU\_9305\Liss\Lab\_1

## 2. Описание практических занятий

**Практическое занятие 1**

## ИЗУЧЕНИЕ ПРИЕМОВ РАБОТЫ С СИСТЕМОЙ STATISTICA

Цель работы: научиться создавать и редактировать таблицы, формировать отчеты и строить графики в системе *Statistica*.

*Задания*

1. Создать таблицу с заголовком «Социологический опрос» из 6 случаев и 4 переменных. Имена переменных задать следующие: Пол, Возраст, Образование, Доход. В качестве названий случаев задать ФИО, заполнить таблицу (табл.5).

*Таблица 5*

**Социологический опрос**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ФИО | Пол | Возраст | Образование | Доход |
| Петров П. | Муж | 21 | Среднее | 125 |
| Комарова И. | Жен | 23 | Высшее | 345 |
| Филова Р. | Жен | 40 | Нет | 50 |
| Статистов М. | Муж | 35 | Среднее | 399 |
| Ивлева Т. | Жен | 18 | Среднее | 136 |
| Часов И. | Муж | 20 | Высшее | 120 |

1. Создать файл отчета.
2. Сопоставить текстовым значениям Пол и Образование численные значения и метки.
3. Добавить новую переменную Доход\_20, значения которой равны Доход + премия в размере 20 % от Дохода (рассчитать по формуле).
4. Изучить блоковые операции и определить среднее значение Дохода.
5. Используя Статистические графики построить столбчатую диаграмму для Возраста и Дохода.
6. Изменить графические установки и сохранить их в отдельном файле, применить их к новому графику.
7. Изучить интерактивный графический анализ данных: кисть и редактор данных графика.
8. Отредактировать файл отчета и сохранить.

**Практическое занятие 2**

**Основы работы с языком STATISTICA Basic**

Цель работы: научиться создавать программы на языке *STATISTICA Visual Basic*, формировать и обрабатывать с их помощью статистические данные, а также пользоваться контекстной помощью.

**Вызов редактора *STATISTICA Visual Basic.*** Для вызова редактора *STATISTICA Visual Basic* необходимо выбрать опцию *STATISTICA Visual Basic* в меню Анализ.

После нажатия на кнопку "Создать…" появится окно, в котором можно задать свойства нового макроса:

– имя макроса;

– тип макроса;

– описание;

– используемый язык программирования.

После нажатия на кнопку "OK..." на экране появится окно редактора, в котором можно вводить текст программы.

Для просмотра справки о функциях, которые можно использовать в программе, необходимо в меню Правка открыть диспетчер функций.

**Примеры использования языка.** Следует иметь в виду наличие в системе достаточного количества примеров макросов. Они находятся в каталоге диск:\ProgramFiles(x86)\StatSoft\STAISTICA 10.

**Типы обработки данных.** В системе *Statistica 5* были два вида обработки данных Sequential и RandomAccess. По умолчанию использовался RandomAccess, это позволяло обрабатывать каждую ячейку отдельно. В противоположность этому Sequential позволял вычислять значения для всего столбца (переменной) одновременно. Сравните:

*Пример 1*

Sequential;

v1:=(v2+v3)/2;

*Пример 2*

RandomAccess;

sum := 0;

for i := 1 to NCases do

for j := 1 to NVars do

sum := sum + V(i,j));

WriteLn('sum=',sum);

Для RandomAccess возможно несколько способов обращения к данным:

–Data(i, j) – *i* номер случая, *j* номер переменной (предопределенный массив);

– V(i, j) - эквивалент Data(i, j);

– Vj(i) – *j* номер переменной, *i* номер случая.

– VARNAME(i) – VARNAME имя переменной, *i* номер случая.

Обращение к отдельной ячейке в Sequential вызывало сообщение об ошибке.

Для того чтобы лучше понять разницу между Sequential и RandomAccess рассмотрим пример. Обе программы находят максимальное по абсолютной величине значение переменной 1 для первых ста случаев, а затем делят значения переменной 1 первых ста случаев на найденное значение.

В Sequential программа читает данные дважды (последовательно по одному случаю): за первый проход находится максимальное по абсолютной величине значение переменной 1, после того как номер случая стал больше 100 программа возвращается (Jump) к случаю номер 1. И затем все значения переменной 1 первых ста случаев делятся на найденное значение – второй проход.

Заметьте, что переменная Mem(1) используется в качестве флага перехода ко второму проходу, а переменная Mem(2) используется для хранения найденного максимума.

Sequential; {specifies Sequential mode}

if Mem(1) = 0 then {if this is the first pass}

begin

if abs(v1) > Mem(2) then Mem(2) := abs(v1);

if CaseNo> 100 then

begin

Mem(1) := 1;

Jump(1);

end;

end

else {if this is the second pass}

begin

v1 := v1 / Mem(2);

if CaseNo> 100 then Stop;

end;

В RandomAccess все случаи доступны программе в предопределенном массиве данных Data(i,j). Программа выполняет действия аналогичные предыдущей, но за один проход.

RandomAccess; {specifiesRandomAccessmode}

maxval := -9999;

for i := 1 to 100 do

if abs(Data(i,1)) >maxval then

maxval := abs(Data(i,1));

for i := 1 to 100 do

Data(i,1) := Data(i,1)/maxval;

В системе *Statistica10* обработка данных производится только в режиме RandomAccess, режим Sequential не используется. Если необходимо произвести вычисления для всего столбца, нужно вызвать цикл *For* с использованием встроенной переменной LOOP\_CASE в качестве счетчика.

**Генерирование выборки, имеющей заданное распределение и расположенной в столбце таблицы.** Для генерирования выборки, имеющей заданное распределение, необходимо выделить столбец щелчком мыши и указать в спецификациях переменной в нижнем поле следующее выражение:

= Vраспределение (Rnd(1), параметры распределения),где V распределение– функция, обратная к функции распределения, Rnd(1) - генератор равномерно распределенных случайных чисел.

Например:

= VNormal (Rnd(1); 3; 0.4) позволяет получить в таблице выборку, имеющую нормальное распределение со средним равным 3 и дисперсией 0.4.

**Генерирование выборки, имеющей заданное распределение при написании программ на языке *STATISTICA Basic.*** Для генерирования выборки, имеющей заданное распределение, необходимо написать очень простую программу – команду по формату:

Variable = Vраспределение (Rnd(1), параметры распределения),где Vраспределение – функция, обратная к функции распределения, Rnd(1) –генератор равномерных случайных чисел, Variable ‑ некоторая переменная.

Пример использования:

Variable = VNormal (Rnd(1), 3, 0.4) позволяет получить в переменной Variable значение из выборки, имеющей нормальное распределение со средним значением, равным 3 и дисперсией 0.4.

Чтобы сгенерировать выборку в столбце таблицы, необходимо использовать цикл *For*со встроенной переменной LOOP\_CASE.

*Задания*

1. В табл. 5 из задания работы № 1 добавить новую переменную Доход\_С, значения которой равны Доход + Доход\_20.
2. Добавить новую переменную в файл данных социологического опроса Доход\_М. Необходимо найти максимальный доход и для каждого респондента определить, сколько процентов от максимального дохода составляет его доход.
3. Составить таблицу 1000 на 3, содержащую в качестве первой переменной значения , второй+случайная ошибка (VNormal (Rnd(1), 0., 0.005)), третьей –  на интервале , причем  изменяется с постоянным шагом.
4. Построить соответствующие графики.
5. Создать кнопку на панели Автозадач для быстрого вызова файла данных социологического опроса.
6. Создать кнопку на панели Автозадач для программ п.2 и п.3.
7. Сохранить описание панели Автозадач в файле lab2.atb.
8. Восстановить установки по умолчанию из файла default.atb.
9. Результаты оформить в форме отчета.

**Практическое занятие 3**

**Подготовка статистических данных для   
работы с пакетом STATISTICA**

*Задания*

Практические занятия проводятся на базе набора статистических данных: IRIS, APPEN, EEG (по номеру варианта).

**Подготовка данных.** Подготовка данных состоит из следующих этапов.

1. Запустить пакет *Statistica*.
2. В разделе меню Home выбрать Open -> Open Document.
3. Выбрать файл с данными (например, IRIS.txt).
4. Выбрать тип данных Delimited.
5. В появившемся окне в качестве разделителя указать пробел (Space). Необходимо также обозначить типы переменных, для этого нужно выбрать соответствующий столбец переменных и в разделе VariableOptions задать тип данных (DataType).
6. Сохранить рабочий файл в формате .sta (например, IRIS.sta).
7. Просмотреть данные, убедиться в правильности их импорта.
8. Выбрать элемент общего меню Data.
9. Используя опцию Subset (код выборки), которая находится в подразделе Manage, подготовить данные для дальнейшего анализа.

Разбить общий файл данных на группы файлов с одинаковым номером класса (например, 1,2). В окне CreateaSubset указать переменные, включенные в коды выборки (Variables: ALL). При выборе строк (cases) в окне SpreadsheetCaseSelectionConditions выбрать EnableSelectionConditions (использовать условный выбор), затем в разделе Includecases ввести выражение для отбора, например, для формирования выборки 1 класса необходимо ввести следующее условие:

V5 = 1.

Сохранить полученный файл. Аналогично подготовить файлы с данными остальных классов.

Подготовить файлы для обучения классификации и проверочные файлы. Файл обучения формируется из первой половины выборок по всем классам. Контрольный файл формируется из второй половины данных (рис.1).



*Рис. 1*

Аналогично подготовить файл для обучения и контроля на основе данных с четными и нечетными номерами.

Подготовить аналогичные файлы с использованием центрирования и нормировки данных (опция Standardize в подразделе Transformations раздела Data).

Сохранить подготовленные файлы.

1. Ознакомиться с остальными опциями меню Data:

– Merge –слитьфайлы;

– Variables;

– Cases;

– Sort;

– Filter/Recode -> Replace Missing Data (замена пропущенных данных).

**Предварительный анализ данных с использованием меню Graphs.**

1. Построить графики зависимостей значений признаков для различных классов данных с использованием опции 2D.
2. Построить линейные графики (LinePlots). Тип графика Multiple.

При нажатии правой кнопки мыши на элементах сетки внутри графика можно открыть GraphOptions для установки параметров графика.

1. Ознакомиться с опциями настройки графика.

**Предварительный анализ данных с использованием Basic Statistics and Tables/DescriptiveStatistics.**

1. В разделе Statistics выбрать Basic Statistics -> Descriptive Statistics

Выбрать Variables (все значения, кроме последнего)

Во вкладке Advanced установить параметры:

‑ ValidN–число строк;

‑ Mеan– математическое ожидание;

‑ Sum – сумма;

‑ Median – медиана;

‑ Standard Deviation– стандартное отклонение;

‑ Variance – дисперсия;

‑ 95 % confidence limits of mean– доверительный интервал для математического ожидания.

Для сохранения таблицы использовать Ctrl-C. Далее строятся таблицы частот (Frequency Tables), гистограммы (Histograms), которые можно найти во вкладке Normality.

1. Далее ознакомиться с опциями обработки "Box&Whisker Plot for All Variables", "Normal probability plots", Matrix (матрица парных корреляций).

**Практическое занятие 4**

**Изучение дискриминантного анализа  
 с помощью пакета программ Statistica**

Цель работы: ознакомиться с методами дискриминантного анализа на основе пакета *Statistica*.

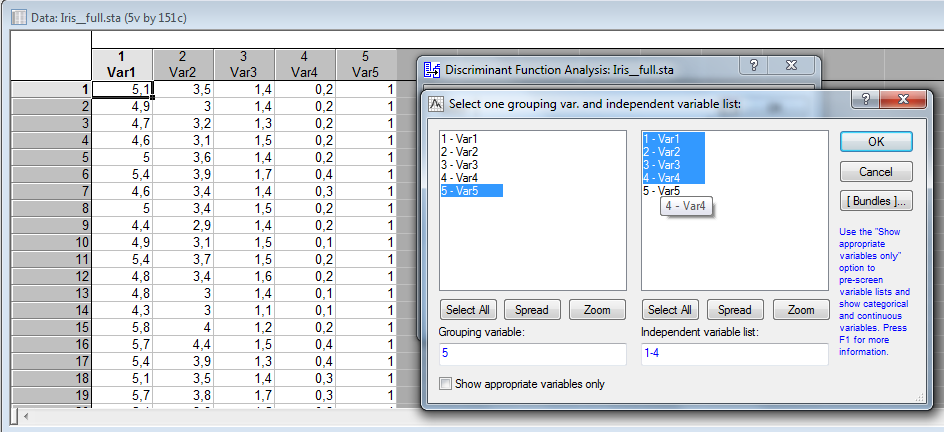
*Задания*

1. Провести исследование по построению дискриминантных функции для имеющихся экспериментальных данных: IRIS, APPEN, EEG (по номеру варианта).

В системе *Statistica* в панели Statistics (Анализ) ->MultivariateExploratoryTechniques (Многомерный анализ) выбрать пункт Discriminant (Дискриминантный анализ).

Выбрать переменные Variables, указать номер группирующей переменной и независимых переменных.

В окне появляется панель с единственной закладкой Quick (Быстрый). Выбрать Variables (Переменные), в первом окне в качестве группирующей переменной указать класс (как правило, последняя переменная), во втором окне выбрать все остальные переменные (рис. 2).



*Рис. 2*

Запустить дискриминантный анализ нажатием на клавишу ОК. В окне Discriminant Function Analysis (Результаты анализа дискриминантных функций) во вкладке Quick (Быстрый) нажать «Summary: VariablesinModel» (Переменные в модели).

Получены результаты дискриминантного анализа.

Просмотреть результаты дискриминантного анализа в окне DiscriminantFunctionAnalysis (Результаты анализа дискриминантных функций) во вкладке Classification (Классификация):

‑ Variablesinmodel (Переменные в модели);

‑ Classificationmatrix (Матрица классификации);

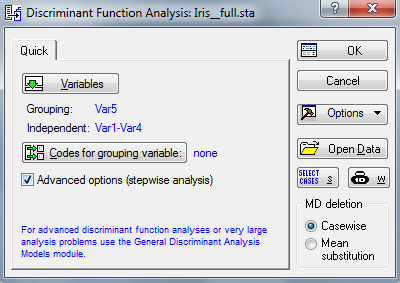
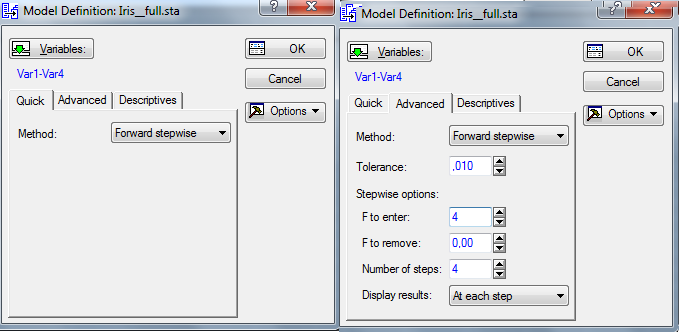
‑ Classificationcases (Классификация наблюдений);

‑ Classificationfunctions (Функции классификации).

Сохранить результаты для отчета.

1. Провести исследование методом дискриминантного анализа с включением переменных.

Запустить дискриминантный анализ аналогично пункту 1, выбрать переменные для анализа. Указать опцию «Advancedoption» (Дополнительные параметры) (рис. 3).



*Рис. 3*

В окне «ModelDefinition» (Определение модели) на вкладке «Quick» (Быстрый) выбрать метод «Forwardstepwise» (Пошаговый с включением). На вкладке «Advanced» (Дополнительно) установить *F* – включения равным 4.0, а также Displayresults – Ateachstep (Вывод результатов – На каждом шаге).

Нажать Next, тем самым перейдя к первому шагу.

Выполнить все шаги, вывести таблицы из п. 1 в отчет.

Сравнить результаты, полученные на последнем шаге с результатами из п. 1.

1. Провести исследование методом дискриминантного анализа с исключением переменных.

Выполнить действия, описанные в п. 2, со следующими изменениями.

Выбрать метод «Backwardstepwise» (Пошаговый с исключением). На вкладке «Advanced» (Дополнительно) установить F-исключения равными 3.9.

Выполнить все шаги, вывести таблицы из п. 1 в отчет.

Сравнить результаты, полученные на последнем шаге, с результатами из п. 1 и 2.

1. Разработка статистической модели данных.

Подготовить самостоятельно модельные примеры из многомерных нормальных векторов с заданными векторами математического ожидания и ковариационной матрицей (параметры модели выбираются по бригадам из задания на разработку статистической модели).

Провести эксперименты по классификации на модельных примерах аналогично описанным выше. Провести теоретический расчет ошибок классификации для модельных примеров и сравнить их с результатами, полученными по программе *Statistica*.

Варианты заданий приведены в табл. 6.

*Таблица 6*

**Варианты заданий**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Вариант | Размерность m | Объем выборки N по классу | μ1 | μ2 | μ3 | σ1 | σ2 | σ3 | Кол-во  классов M |
| 1 | 2 | 100 | 1 | 2 | - | 1 | 1 | - | 2 |
|  |  |  | 1 | -2 | - |  |  |  |  |
| 2 | 2 | 200 | -0,5 | 2 | - | 1,5 | 1 | - | 2 |
|  |  |  | -1,0 | 4 | - |  |  |  |  |
| 3 | 3 | 150 | 1 | 2,5 | 4 | 0,5 | 1 | 2 | 3 |
|  |  |  | 1 | 2,5 | 4 |  |  |  |  |
|  |  |  | 1 | 2,5 | 4 |  |  |  |  |
| 4 | 3 | 200 | -1 | 0 | 2 | 1 | 0,5 | 1 | 3 |
|  |  |  | -1 | 0 | 2 |  |  |  |  |
|  |  |  | -1 | 0 | 2 |  |  |  |  |
| 5 | 3 | 100 | 0 | 3 | 6 | 1,5 | 1 | 2 | 3 |
|  |  |  | 0 | 3 | 6 |  |  |  |  |
|  |  |  | 0 | 3 | 6 |  |  |  |  |
| 6 | 2 | 200 | 4 | 6 | - | 3 | 3 | - | 2 |
|  |  |  | 4 | 6 | - |  |  |  |  |
| 7 | 3 | 150 | 1 | 2 | - | 0.5 | 1 | 0.5 | 3 |
|  |  |  | 1 | 2 | - |  |  |  |  |
|  |  |  | 1 | 2 | - |  |  |  |  |
| 8 | 2 | 300 | 1.5 | 3.0 | -1,5 | 1 | 1 | - | 3 |
|  |  |  | 3.0 | 4.0 | -1,0 |  |  |  |  |

*Окончание табл. 6*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Вариант | Размерность m | Объем выборки N по классу | μ1 | μ2 | μ3 | σ1 | σ2 | σ3 | Кол-во  классов M |
| 9 | 3 | 500 | 0 | -1.5 | 3 | 1.5 | 2.0 | 0.75 | 3 |
|  |  |  | 0 | -1.5 | 3 |  |  |  |  |
|  |  |  | 0 | -1.5 | 3 |  |  |  |  |
| 10 | 2 | 250 | 1.5 | 3.0 | - | 1.0 | 1.2 | - | 2 |
|  |  |  | 1.0 | 2.7 | - |  |  |  |  |
| 11 | 3 | 150 | 4.0 | 5.2 | 2.2 | 0.3 | 0.5 | 0.6 | 3 |
|  |  |  | 4.5 | 4.9 | 3.9 |  |  |  |  |
|  |  |  | 3.7 | 4.1 | 3.8 |  |  |  |  |

Модель представляет собой набор многомерных векторов:

,

имеющих заданные вектора математических ожиданий и заданные ковариационные матрицы (одинаковые по классам). Компоненты векторов имеют нормальное распределение. Количество классов M = 2, 3.

**Практическое задание 5**

**Исследование методов кластер-анализа**

Цель работы: ознакомиться с методами кластер-анализа на основе пакета *Statistica*.

*Задания*

1. Ознакомиться с методическими указаниями "Кластер-Анализ. Работа с пакетом программ *Statistica*.
2. Исследования провести на имеющихся данных:

IRIS, APPEN, EEG, а также на модельных примерах в соответствии с индивидуальным заданием.

1. Запустить пакет *Statistica*. Войти в меню Statistic-Mult/Exploratory-Cluster.
2. Исследовать методы иерархической группировки – Joining (treeclustering) из меню Clustering Method.
   1. Выбрать переменные (группировочная переменная не участвует в обработке).
   2. Установить параметр Inputfile = Rawdata.
   3. Установить параметр Cluster = Cases (rows).
   4. Выбрать метод связывания Правило объединения.
   5. Выбрать метод измерения расстояния – Мера близости.
   6. Запустить процедуру кластеризации.
   7. Просмотреть результаты построения иерархического дерева выбрав пункт Горизонтальная дендрограмма или Вертикальная дендрограмма.
   8. Проанализировать, какие данные образуют кластеры (опция – Схема объединения).
   9. Исследовать процесс кластеризации при различных сочетаниях методов связывания и методов измерения расстояния.
3. Провести исследование иерархической кластеризации при использовании кластеризации по признакам.
   1. Выбрать в меню Cluster Analysis значение Cluster = Variables (columns).
   2. Провести кластеризацию аналогично п. 5.4 – 5.7.
   3. Определить наборы признаков наиболее и наименее связанные друг с другом.
   4. Выбрать наиболее информативный минимальный набор признаков и проверить его эффективность в режиме Дискриминантный анализ.
4. Выбратьметод K-means clustering изменю Clustering Method.
   1. Загрузить исходные данные.
   2. Выбрать переменные.
   3. Выполнить кластер-анализ.
   4. Проанализировать результаты кластеризации k-MeansClusteringResults (рис. 4).

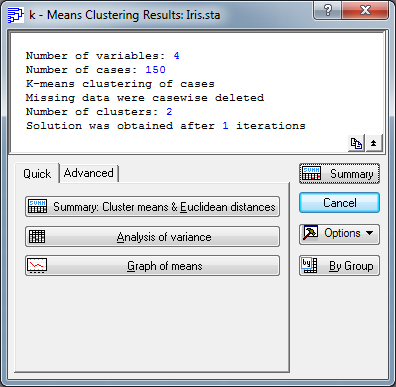
‑ анализ дисперсии (дисперсионный анализ);

‑ математические ожидания и эвклидовы расстояния между кластерами (средние кластеров и евклидовы расстояния);

‑ графики математических ожиданий по кластерам (график средних);

‑ дискриптивные статистики по кластерам (статистика для каждого кластера);

‑ содержание кластеров (элементы кластеров и расстояния).



*Рис. 4*

* 1. Провести кластеризацию методом К-средних по признакам.

1. Сохранить графики и таблицы результатов исследования.
2. Сделать сравнительные выводы по проведенным исследованиям.
3. Оформить результаты в виде отчета.

**Практическое занятие 6**

**Исследование методов факторного анализа**

Цель работы: ознакомиться с методами факторного анализа на основе пакета *Statistica*.

*Задания*

1. Распаковать файл FACT\_DOC.exe.
2. Ознакомиться с основами факторного анализа (файл BASIC.doc).
3. Ознакомиться с работой в пакете *Statistica* c разделом Фак­торный анализ (файл GUIDE.DOC) и примером использования (файл EXAMPLE.doc).
4. Провести исследование по построению факторных моделей для данных IRIS, APPEN, EEG.
5. Подготовить отчет по исследованию.

**Практическое занятие 7**

**Исследование методов многомерного шкалирования**

Цель работы: ознакомиться с методами многомерного шкалирования на основе пакета *Statistica*.

*Задания*

1. Ознакомиться с теорией многомерного шкалирования.

Исследования провести на имеющихся данных:

IRIS, APPEN, EEG (по номеру варианта).

1. Запустить пакет *Statistica*. Войти в меню CLUSTERANALISYS.

С помощью одного из методов кластер-анализа получить матрицу расстояний, задав следующие параметры для кластеризации:

– выбрать все вектора выборки (cases) (в случае выдачи ошибки о слишком большом числе векторов разделите выборку на 2 части, отобрав по половине векторов из каждого класса, работайте только с одной из частей);

– установить параметр Cluster = cases;

– установить параметр Input = RAWDATA;

– выбрать метод связывания;

– выбрать метод измерения расстояния;

– запустить процедуру кластеризации.

Далее сохранить матрицу расстояний.

1. Запустить модуль статистики MultidimensionalScaling:

– выбрать переменные (allvariables);

– ввести значение размерности для анализа;

– выбрать конфигурацию StandardGuttman-Lingoes;

– просмотреть данные в окне Parameter Estimation;

– в диалоговом окне Results проанализировать результаты многомерного шкалирования:

Final Configuration;

D-hat Values;

D-star Values;

Distance Matrix;

Summary;

– cохранить координаты заключительной конфигурации в стандартном файле данных;

– построить график заключительной конфигурации Graphsfinalconfiguration - 2D, 3D;

– построить диаграмму Шепарда зависимости D-hats от расстояний Graph D-hatvs.Distances;

– построить диаграмму Шепарда зависимости D-stars от расстояний Graph D-starvs.Distances.

1. Вернитесь к исходной выборке (неразделенной). Проделать п.2–3 для признаков (variables). То есть при формировании матрицы расстояний выбрать все переменные кроме целевой и установить параметр Cluster = variables. При проведении многомерного шкалирования выбрать все переменные (allvariables) и ввести значение размерности для анализа = 2.
2. Сохранить графики и таблицы результатов исследования.
3. Сделать сравнительные выводы по проведенным исследованиям. Сравнить результаты шкалирования выборки по признакам с результатами факторного анализа, полученными в предыдущей работе.
4. Оформить результаты в виде отчета.

**3. Используемые экспериментальные данные**

**Описание данных по измерению параметров цветов ириса.** Используются ботанические статистические данные из работы известного специалиста по математической статистике Р. А. Фишера, использованные им в работах по дискриминантному анализу, создателем которого он является. Эти данные используются в настоящее время достаточно широко как текстовые для исследования в области математической статистики.

Данные представляют собой результаты четырех типов измерений (в см) над 50 цветками от каждой из трех разновидностей ириса:

– ирис щетинистый (IRIS Setora) - обозначим класс 1 (Se);

– ирис разноцветный (IRIS Versicolor) - обозначим класс 2 (Ve);

– ирис вирджиника (IRIS Virginica) - обозначим класс 3 (Vi).

В качестве признаков используются измерения:

– X1 - длина чашелистика;

– X2 - ширина чашелистика;

– X3 - длина лепестка;

– X4 - ширина лепестка.

Результаты измерения приведены в файле IRIS.txt, где каждая колонка с 1‑4 соответствует признакам X1– X4, а колонка 5 определяет номер класса (признак X5).

**Описание медицинских статистических данных для диагностики типов заболевания аппендицит на основании данных клинических обследований.** Задача состоит в выработке решающих правил для диагностики трех типов острого аппендицита (1 – гангренозный, 2 – флегмонозный, 3– катаральный) на основе использования восьми симптомов (табл. 7).

*Таблица 7*

**Симптомы острого аппендицита, степени выраженности и их коды**

|  |  |  |
| --- | --- | --- |
| №  симптома | Симптомы острого аппендицита | Степени выраженности  симптомов и коды |
| 1 | Боли в правой подвздошной области | 1 – незначительные  2 - выраженные |
| 2 | Продолжительность болей в правой подвздошной области | 4 - до 12 ч.; 3 - 13-24 ч.  2 - 25-48 ч.; 1 - свыше 2-х суток |
| 2 | Частота пульса | 1 - до 80; 2 - 81-100;  3 - свыше 100 ударов в минуту |
| 4 | Лейкоциты крови | 1 - до 8; 2 - 8-14; 3 - свыше 14 тыс. |
| 5 | Изменение языка | 0 - не обложен;  1 - обложен |
| 6 | Симптом Щеткина-Блюмберга | 0 - отсутствует;  1 - выражен |
| 7 | Симптом Ровзинга | 0 - отсутствует;  1 - выражен |
| 8 | Защитное мышечное напряжение | 0 - отсутствует;  1 - выражен |

По данным клиники составлена матрица обучающей информации восьми симптомов для трех видов острого аппендицита и неподтвержденного диагноза. Девятым показателем является группировочный признак, т. е. группа, к которой относится данный индивидуум. Группа больных с гангренозным аппендицитом состоит из 28 наблюдений, ее группировочный признак 1. Больных с флегмонозным аппендицитом 25 человек, группировочный признак 2. Группа больных с катаральным аппендицитом содержит 26 наблюдений и ее группировочный признак 3. С неподтвержденным диагнозом острого аппендицита было 24 человека, группировочный признак 4. Массив обучающей информации содержится в файле appen.txt.

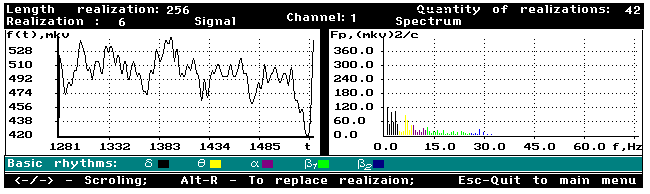
**Описание данных для диагностики типов артефактов при анализе электроэнцефалограмм человека.** Рассматривается задача автоматического выделения артефактов при анализе электорэнцефалограмм человека.

При регистрации и анализе фоновой и вызванной биоэлектрической активности мозга человека с помощью электорэнцефалографии существенные погрешности в оценке этих характеристик вносят сигналы внемозгового происхождения, так называемые «сигналы артефактов», связанных как с нарушением работы аппаратуры, так и с функционированием исследуемого организма. Проблема обнаружения и устранения артефактов является весьма сложной задачей в связи с тем, что источники возникновения последних очень разнообразны, а построение формальных моделей оказывается затруднительным. Это приводит к тому, что в большинстве цифровых ЭЭГ-систем используются интерактивные методы обнаружения и редактирования ЭЭГ, что требует весьма больших затрат времени.

С целью уменьшения влияния этих факторов необходимо разработать автоматические алгоритмы классификации артефактов.

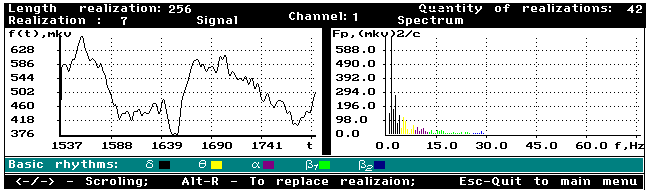
В основу первичного описания артефактов положено использование спектральных описаний реализаций ЭЭГ.

На этапе формирования эталонных описаний артефактов были использованы записи ЭЭГ для различных больных, из которых (из ЭЭГ) вырезались участки, содержащие артефакты. Далее по этим записям производился расчет спектра мощности по реализации длиной 256 точек. На рис.5 приведен пример ЭЭГ сигнала без артефакта и спектр Фурье этого сигнала.

****

*Рис. 5*

На рис. 6 приведен пример ЭЭГ сигнала с артефактом и спектр этого сигнала.

****

*Рис. 6*

В работе рассматриваются четыре типа артефактов:

‑ моргание глаз (класс 1);

‑ движение электродов (класс 2);

‑ плохой контакт (класс 3);

‑ пароксизмальная активность (класс 4).

Для удобства обработки спектры ЭЭГ-сигналов осреднялись в диапазонах так называемых ритмов ЭЭГ. Среднее значение энергии в диапазоне частот определяет значение соответствующего признака описания сигналов.

Характеристики используемых признаков:

|  |  |  |
| --- | --- | --- |
| Признак | Содержание | Диапазон частот в Гц |
| 1  2  3  4  5  6 | Ритм дельта  Ритм тета  Ритм альфа  Ритм бета1  Ритм бета 2  Класс артефакта | 0,5 – 4  4 – 8  8 – 12  12 – 25  25 – 45 |

Признаки 1–5 представляют собой вещественные числа, признак 6 – целое число, соответствующее номеру класса артефакта по 154 экспериментам.

Используемые данные распределяются по классам следующим образом:

|  |  |
| --- | --- |
| № класса | Номера выборок |
| 1  2  3  4 | 1 – 37  38 – 86  87 –120  121 – 154 |

# Список рекомендуемой литературы

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

Д. М. Клионский, В. В. Геппенер, А. В. Экало, Д. И. Каплун Модели и методы анализа и интерпретации данных и их программно-аппаратная реализация СПб.: СПбГЭТУ "ЛЭТИ", 2016.

Халафян А.А. Статистика 6. Статистический анализ данных. М.: Изд-во Бином-Пресс, 2007.

Содержание

1. [Знакомство с системой STATISTICA……………………………………….3](#_Toc528241679)

2. Описание практических занятий………………………………………...15

[Практическое занятие 1. 14](#_Toc528241703)

[Практическое занятие 2……………………………………………..………………**Ошибка! Закладка не определена.**](#_Toc528241705)

[Практическое занятие 3 **Ошибка! Закладка не определена.**](#_Toc528241709)

[Практическое занятие 4……………………………………………………………..**Ошибка! Закладка не определена.**](#_Toc528241710)

[Практическое занятие 5. **Ошибка! Закладка не определена.**](#_Toc528241711)

[Практическое занятие 6. **Ошибка! Закладка не определена.**](#_Toc528241712)

[Практическое занятие 7. **Ошибка! Закладка не определена.**](#_Toc528241713)

3. Используемые экспериментальные данные………………………...32

[Список литературы 33](#_Toc528241714)

Геппенер Владимир Владимирович

Клионский Дмитрий Михайлович

Мандрикова Оксана Викторовна

Будчан Дарья Сергеевна

**Анализ и интерпретация данных**

Учебно-методическое пособие

Редактор М.Б. Шишкова

––––––––––––––––––––––––––––––––––––––––––––––––––––––––––

Подписано в печать . Формат 60×84 1/16.

Бумага офсетная. Печать цифровая. Печ. л.

Гарнитура «TimesNewRoman». Тираж 49 экз. Заказ

––––––––––––––––––––––––––––––––––––––––––––––––––––––––––

Издательство СПбГЭТУ «ЛЭТИ»

197376, С.-Петербург, ул. Проф. Попова, 5