

В.П.БОРОВИКОВ

ПОПУЛЯРНОЕ ВВЕДЕНИЕ В СОВРЕМЕННЫЙ АНАЛИЗ ДАННЫХ В СИСТЕМЕ *STATISTICA*

*МЕТОДОЛОГИЯ И ТЕХНОЛОГИЯ
СОВРЕМЕННОГО АНАЛИЗА ДАННЫХ*

*Допущено УМО по образованию в области прикладной математики
и управления качеством в качестве учебного пособия
для студентов высших учебных заведений, обучающихся
по направлению подготовки 230400 «Прикладная математика»*

Москва
Горячая линия - Телеком
2013

УДК 004.9:519.25

ББК 32.973

Б83

Боровиков В.П.

Б83 Популярное введение в современный анализ данных в системе *STATISTICA*. Учебное пособие для вузов. – М.: Горячая линия – Телеком, 2013. – 288 с., ил.

ISBN 978-5-9912-0326-5.

Книга открывает широкому кругу читателей современный анализ данных в программе *STATISTICA*. *STATISTICA* (производитель StatSoft, USA) занимает лидирующее положение среди программ анализа данных и имеет сотни тысяч зарегистрированных пользователей в России и мире. На простых, ясных примерах популярно описаны современные методы анализа данных – визуальный анализ и графическое представление данных, описательные статистики, методы классификации и прогнозирования.

Классические методы дополнены самым современным инструментарием, включая нейронные сети и DataMining. Читатель знакомится с методами и компьютерными технологиями анализа данных и учится применять их на практике, *основной лейтмотив книги – теория неотделима от практики.*

Для широкого круга читателей, желающих познакомиться с современными методами и компьютерными технологиями анализа данных и их применением в различных областях: экономика, маркетинг, финансы, страхование, промышленность, телекоммуникации, медицина и др. Книга будет особенно полезна студентам и преподавателям вузов при проведении учебных и практических занятий.

ББК 32.973

Адрес издательства в Интернет www.TECHBOOK.RU

Учебное издание

Боровиков Владимир Павлович
Популярное введение в современный анализ данных
в системе *STATISTICA*
Учебное пособие для вузов

Подготовка оригинал-макета Н. В. Дмитриевой
 Обложка художника В. Г. Ситникова

Подписано в печать 15.03.13. Формат 70×100/16. Усл. изд. л. 24. Изд. № 8015

ISBN 978-5-9912-0326-5

© В. П. Боровиков, 2013

© Издательство «Горячая линия – Телеком», 2013

Оглавление

ВВЕДЕНИЕ. ПРИГЛАШЕНИЕ В СОВРЕМЕННЫЙ АНАЛИЗ ДАННЫХ НА КОМПЬЮТЕРЕ	5
ГЛАВА 1. ПЕРВЫЕ ШАГИ В <i>STATISTICA</i>	11
1.1. Запуск программы	11
1.2. Рабочее окно <i>STATISTICA</i> : классическое меню или Лента	12
1.3. Панели инструментов	14
1.4. Аналитические модули <i>STATISTICA</i>	16
1.5. Создание файла данных. Пример 1: результаты олимпийских чемпионов	18
1.6. Пример 2. Импорт газа и топлива в США.....	25
1.7. Вычисление дескриптивных статистик исходных данных.....	32
1.7.1. Некоторые сведения из элементарной статистики.....	32
1.7.2. Вычисление описательных статистик в <i>STATISTICA</i>	35
1.8. Корреляции: определения и вычисления	38
1.9. Простейшая визуализация: диаграммы рассеяния и гистограммы.....	39
1.9.1. Диаграмма рассеяния.....	39
1.9.2. Гистограмма	42
ГЛАВА 2. ВЕРОЯТНОСТНЫЙ КАЛЬКУЛЯТОР И ВЕРОЯТНОСТНЫЕ РАСПРЕДЕЛЕНИЯ	44
2.1. Вероятностный калькулятор	44
2.1.1. Нормальное распределение.....	46
2.1.2. Распределение хи-квадрат.....	52
2.1.3. t-распределение Стьюдента.....	54
2.1.4. Распределение Фишера.....	58
2.1.5. Логарифмически-нормальное распределение	60
2.2. Биномиальное распределение и игровые задачи	62
2.2.1. Задача о коровах.....	65
2.2.2. Задача шевалье де Мере	67
2.2.3. Измененная задача шевалье де Мере	68
2.2.4. Еще одна задача игрока	70
2.2.5. Задачи для самостоятельного решения	72
2.2.6. Генуэзская лотерея.....	72
2.3. Генерация случайных чисел в <i>STATISTICA</i>	74
ГЛАВА 3. ВИЗУАЛЬНЫЙ АНАЛИЗ ДАННЫХ	76
3.1. Двумерный визуальный анализ данных	76
3.1.1. Гистограммы.....	76
3.1.2. Диаграммы рассеяния	83
3.2. Трехмерный визуальный анализ данных	91
ГЛАВА 4. КЛАССИФИКАЦИЯ ДАННЫХ В <i>STATISTICA</i>	94
4.1. Обзор метода.....	94
4.2. Постановка задачи.....	94
4.3. Пример Фишера: классификация цветов ирисов.....	96
4.4. Обобщенный дискриминантный анализ.....	108
ГЛАВА 5. КЛАСТЕРИЗАЦИЯ: МОДУЛЬ КЛАСТЕРНЫЙ АНАЛИЗ	115
5.1. Обзор метода.....	118
5.2. Постановка задачи, обзор методов	120

5.3. Модуль Кластерный анализ – технология, пошаговый разбор примера	121
ГЛАВА 6. РЕГРЕССИОННЫЙ АНАЛИЗ В STATISTICA – МОДУЛЬ	
МНОЖЕСТВЕННАЯ РЕГРЕССИЯ	129
6.1. Описание модели	130
6.2. Метод решения	131
6.3. Технология регрессионного анализа в STATISTICA	136
6.4. Пошаговые примеры	143
6.5. Примеры использования средства кисть для анализа данных	150
6.6. Задачи для самостоятельного решения	154
ГЛАВА 7. АНАЛИЗ ВЫЖИВАЕМОСТИ В STATISTICA	158
7.1. Таблицы жизни	160
7.2. Оценки Каплана – Мейера	165
7.3. Сравнение выживаемости в группах	168
7.4. Регрессионные модели в анализе выживаемости	169
ГЛАВА 8. АВТОМАТИЗИРОВАННЫЕ НЕЙРОННЫЕ СЕТИ STATISTICA (SANN).....	172
8.1. Основные парадигмы нейронных сетей	173
8.2. Математические модели	174
8.3. Обучение и кросс-проверка	175
8.4. Модель Розентблатта	176
8.5. Пошаговый пример: прогнозирование временных рядов с помощью нейронных сетей	177
ГЛАВА 9. DATA MINING – ДОБЫЧА ДАННЫХ	187
9.1. Этапы работы в Data Mining	187
9.2. Меню STATISTICA Data Miner	189
9.3. Средства анализа STATISTICA Data Miner	192
9.4. Пример проекта в STATISTICA Data Miner	192
ГЛАВА 10. ПОПУЛЯРНОЕ ВВЕДЕНИЕ В ТЕОРИЮ ВЕРОЯТНОСТЕЙ.....	198
10.1. Формула полной вероятности	201
10.2. Формула Байеса	201
10.3. Классическое вероятностное рассуждение	204
10.4. Вероятностные модели в биологии	207
10.5. Вероятностные модели в телекоме	208
10.6. Выборочный контроль качества	211
10.7. Занимательные вероятностные задачи	213
10.8. Вероятностный подход к задачам классификации	218
ПРИЛОЖЕНИЕ 1. ЯЗЫК STATISTICA VISUAL BASIC	222
ПРИЛОЖЕНИЕ 2. ПОДКЛЮЧЕНИЕ К БАЗЕ ДАННЫХ	224
ПРИЛОЖЕНИЕ 3. ОПЕРАЦИИ СТЕКИНГ И АНСТЕКИНГ	233
ПРИЛОЖЕНИЕ 4. ГАЛЕРЕЯ ГРАФИКОВ STATISTICA	238
БИБЛИОГРАФИЧЕСКИЙ СПИСОК	285
ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ	286

Введение. Приглашение в современный анализ данных на компьютере

Эта книга посвящена анализу данных – мощному методу исследования окружающего мира, в котором мы существуем и в котором необходимо принимать осознанные решения. Если вы инженер, актуарий, маркетинг, маркетолог, аналитик, врач, то эта книга для вас.

В современном информационно организованном мире невозможно обойтись без всестороннего исследования данных и, следовательно, без системы, позволяющей провести этот анализ. *STATISTICA* является лидером среди программ статистической обработки данных в среде Windows.

На простых примерах, взятых из различных областей человеческой деятельности: экономика, бизнес, маркетинг, промышленность, телекоммуникации, медицина и др., мы показываем, как анализируются данные в системе *STATISTICA*, объединяющей в едином интерфейсе классические и современные методы анализа данных.

Лейтмотивом книги является соединение методологии и компьютерной технологии анализа данных.

Материал в книге расположен таким образом, что вы можете повторить все описанные действия вслед за нами на собственном компьютере. Упражнения и задачи для самостоятельной работы позволят углубить понимание предмета.

Наш опыт показывает, что без самостоятельной работы с данными вы не сможете овладеть анализом данных, так же как не сможете научиться плавать, не входя в воду. Собственно наша цель состоит в том, чтобы научить вас использовать программу *STATISTICA* в своих целях.

Исследование данных имеет свою последовательность: вначале данные нужно загрузить в систему из внешних баз данных. Далее необходимо провести чистку данных, удалить выбросы, заполнить пропуски, визуализировать данные, представить в удобном для исследования виде. Затем использовать разнообразные аналитические процедуры разведочного анализа (группировку, кластерный и дисперсионный анализ, регрессионные модели и др.) позволяющие найти закономерности и сформулировать разумные гипотезы о структуре данных. Именно эта последовательность действий реализована в системе *STATISTICA* в виде диалоговых окон с предопределенными настройками анализа.

STATISTICA позволяет провести эти действия в удобной графической среде, гибко настраиваемой по желанию пользователя.

Мы включили в книгу большое количество самых разнообразных примеров, чтобы пользователь повторил вслед за нами действия на компьютере. Именно в повторении действий и выполнении упражнений заключается лейтмотив книги.

Если у вас осталось смутное представление о науке статистике после института, не отчаивайтесь: примеры и упражнения подобраны таким образом, что доступны даже школьникам старших классов. Делайте вслед за нами, и вы научитесь решать задачи с помощью *STATISTICA* самостоятельно!

В самой науке статистика нет ничего сложного; следует напомнить, что первоначально большинство задач статистики возникло из игр: бросание костей, монет, карточных игр, рулетки. Близки к ним лотереи и разнообразные задачи на угадывание. Подобные задачи формулируются совершенно просто, «без всяких заумностей». Они доступны любому человеку со здравым мышлением.

Математика (гр. *mathêma* – знание, понимание) призвана объяснять закономерности, наблюдаемые на опыте, а не затемнять сознание сложными формулами и манипуляциями над числами.

Известная задача о том, стоит ли ставить на выпадение двух шестерок одновременно при бросании пары костей, возникла во Франции в конце XVII века из наблюдений за игрой. В *STATISTICA* эта задача может быть решена несколькими щелчками мыши. И мы покажем, как это сделать.

В системе *STATISTICA* есть замечательное средство – *вероятностный калькулятор*, пользоваться которым так же просто, как обычным калькулятором. Многие элементарные вероятностные задачи могут быть решены с помощью этого средства. Мы научим вас пользоваться вероятностным калькулятором, а также строить разнообразные статистические графики: гистограммы, диаграммы рассеяния, графики типа «ящики с усами», вычислять простейшие статистики: среднее, стандартное отклонение, корреляции, процентные точки и т. д. Мы научим вас также генерировать случайные последовательности в *STATISTICA*, например, последовательности, возникающие при бросании монет.

Мы научим вас решать простейшие игровые задачи, доводя их до численного результата с помощью *STATISTICA*.

Слышали ли вы когда-нибудь о Генуэзской лотерее или о задачах, предложенных Сэмюэлем Пеппайсом Ньютоном? Если нет, мы расскажем вам об этих задачах и покажем, как они решаются с помощью *STATISTICA*.

В этой книге популярно рассказывается о современном анализе данных, науке статистике и о системе *STATISTICA*, позволяющей проводить анализ данных на компьютере.

Разбираемые примеры сгруппированы в разделах:

- описательный анализ;
- визуальный анализ;
- разведочный анализ данных;
- оценивание зависимостей в данных;
- классификация – отнесение объекта к определенной группе – дискриминантный анализ, обобщенный дискриминантный анализ, деревья классификации;
- кластерный анализ.

Специальные темы, относящиеся к доказательной медицине, в частности, анализ выживаемости, собраны в отдельной главе. Анализ выживаемости представляет собой раздел современного анализа данных, объединяющий различные статистические процедуры для построения таблиц жизни, оценки функции выживания и др., наиболее интенсивно используемые в медицинских приложениях, биологии, а также при проведении актуарных расчетов.

Кратко опишем основные разделы.

В *описательном анализе* вычисляются самые общие дескриптивные статистики (среднее, стандартное отклонение, медиана и др.), позволяющие компактно описать данные. Эти статистики вычисляются как для всех данных, так и для группированных данных, например, для мужчин и женщин.

Очень важным этапом исследования является *визуализация данных*. Вначале данные нужно увидеть, потом сформулировать разумные гипотезы относительно их природы, уникальные графики *STATISTICA* позволяют это сделать.

Смысл нашего подхода к анализу данных состоит в том, чтобы получать всестороннее визуальное представление данных на всех этапах статистического исследования и на основе этого представления выбирать следующий шаг анализа. Визуализируя данные, вы выдвигаете гипотезы, которые невозможно было бы выдвинуть, имея только численное представление.

В *STATISTICA* имеются сотни типов графиков, предназначенных для визуализации, разведывательного анализа, графического представления результатов и выбора последующих направлений анализа. Такие уникальные графики, как лица Черного, диаграммы Вороного, матричные графики, позволяющие, например, визуализировать корреляционную матрицу, категоризированные, трассировочные и др. графики, а также большой выбор двумерных и трехмерных научных и деловых графиков и диаграмм становятся доступными для пользователя.

Кроме стандартных типов графиков в *STATISTICA* имеется большое количество специализированных статистических графиков: «ящиков с усами» с разнообразными опциями по выбору средней точки, граничных значений, подгонки распределений, определения выбросов, разнообразных гистограмм, графиков на нормальной вероятностной бумаге, графиков типа «вероятность-вероятность», «квантиль-квантиль» и т. д.

Примеры нескольких графиков приведены на рис. В.1–В.3.

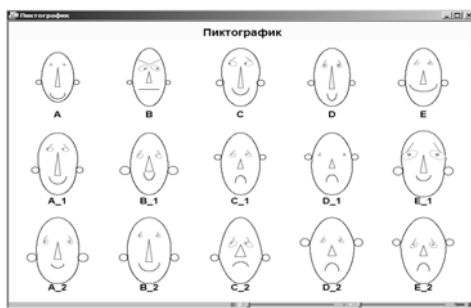


Рис. В.1. Лица Чернова – результаты допинг-контроля спортсменов

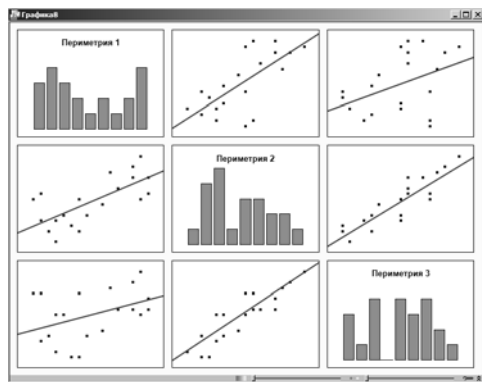


Рис. В.2. Визуализация корреляций

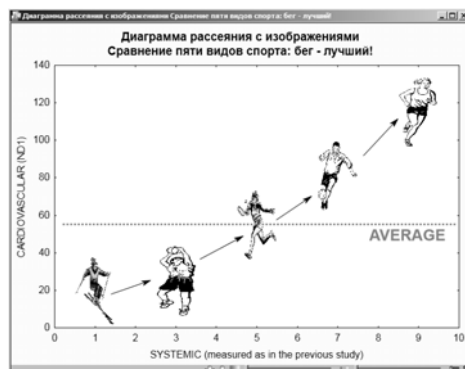


Рис. В.3. Диаграмма рассеяния с образами – предпочтения в видах спорта

Графики можно уменьшать, увеличивать, накладывать друг на друга, вращать, корректировать перспективу, применять средство «Рентген» в трехмерной графике, чтобы увидеть «очертания дальних гор на фоне ближних», определять собственную палитру цветов, добавлять пользовательский текст, рисунки, стрелки и т. д.

В последних версиях системы *STATISTICA* многие настройки можно осуществлять непосредственно в окне графика, не открывая дополнительных окон. Например, враще-