

УДК 004.02:004.6

Салмин А.А.

Анализ данных. Конспект лекций. – Самара.: ПГУТИ, 2013. - 72 с.

Рассматриваются вопросы анализа данных. Приводятся некоторые из основополагающих методик анализа данных, такие как: регрессионный анализ, корреляция, дисперсионный анализ и др. Отражены вопросы интеллектуального анализа данных, с помощью которого можно выявить ранее неизвестные, нетривиальные закономерности в данных.

Рецензент:

Тарасов В.Н. – д.т.н., профессор, зав. кафедрой «Программного обеспечения и управления в технических системах» ПГУТИ

Федеральное государственное образовательное бюджетное учреждение
высшего профессионального образования
«Поволжский государственный университет телекоммуникаций и
информатики»

© Салмин А.А., 2013

Содержание конспекта лекций

ВВЕДЕНИЕ	6
1. ВВЕДЕНИЕ В «АНАЛИЗ ДАННЫХ»	7
1.1. Работа с данными	7
1.2. Этапы решения задачи анализа данных и их взаимосвязи	8
2. РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ	14
2.1. Вероятность	14
2.2. Распределения вероятностей	15
2.3. Случайные переменные и случайные выборки данных	17
2.4. Нормальное распределение	17
2.5. Формула Байеса	18
3. СТАТИСТИКА ВЫВОДОВ	21
3.1. Доверительные интервалы	21
3.2. Проверка гипотез	22
3.2.1. Типы ошибок	23
3.2.2. Области принятия и непринятия	23
3.2.3. t-распределение	24
3.3. Применение непараметрического теста для парных данных	26
4. АНАЛИЗ ТАБЛИЧНЫХ ДАННЫХ	29
4.1. Сводные таблицы	29
4.2. Вычисление ожидаемого количества наблюдений	31
4.3. Статистика хи-квадрат Пирсона	32
5. ОСНОВЫ РЕГРЕССИОННОГО АНАЛИЗА	34
5.1. Понятие «регрессия»	34
5.2. Простая линейная взаимосвязь	34
5.2.1. Уравнение регрессии	34
5.2.2. Подгонка линии регрессии	36
5.2.3. Интерпретация параметров регрессии	38
5.3. Проверка модели регрессии	39
6. КОРРЕЛЯЦИЯ	42
6.1. Понятие «корреляции»	42
6.2. Матрица корреляции	43
6.3. Матрица точечных диаграмм корреляций	44
7. АППАРАТ МНОЖЕСТВЕННОЙ РЕГРЕССИИ	46
7.1. Уравнение множественной регрессии	46
7.2. Проверка допущений регрессии	48
7.3. Пошаговая регрессия	50
7.4. Логистическая регрессия	50
7.5. Нелинейная регрессия	51
8. ДИСПЕРСИОННЫЙ АНАЛИЗ	52
8.1. Однофакторный дисперсионный анализ	52
8.2. Однофакторный дисперсионный анализ и анализ регрессии	56

8.2. Двухфакторный дисперсионный анализ	57
9. КОГНИТИВНЫЙ АНАЛИЗ. ГРАФЫ	61
9.1. Когнитивный анализ	61
9.2. Методика когнитивного анализа сложных ситуаций	62
9.3. Регрессионно - когнитивный анализ	63
10. ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ	66
10.1. Системы аналитической обработки данных	66
10.1.1. CRM – технология	66
10.1.2. ERP – системы	68
10.1.3. OLAP – технология	68
10.2. Интеллектуальный анализ данных (Data Mining)	69
10.2.1. Этапы исследования данных с помощью методов Data Mining	70
10.2.2. Типы закономерностей	70
10.2.3. Методы Data Mining	71

ВВЕДЕНИЕ

Предлагаемый конспект лекций по дисциплине «Анализ данных» обеспечивает подготовку студентов к эффективному использованию современных компьютерных средств анализа данных. Предлагаются основные темы, посвященные формированию теоретических и практических навыков работы с пакетами прикладных программ для решения задач анализа и интерпретации данных для создания прогнозов ситуации и принятия управленческих решений. В рамках конспекта лекций по дисциплине рассматриваются различные способы создания, форматирования, описания базовых принципов работы с таблицами данных с целью их последующего анализа при помощи статистических и математических методов. Таким образом, у будущих специалистов осуществляется формирование основ теоретических знаний и практических навыков работы в области анализа данных и принятия управленческих решений.

Следует также отметить тот факт, что в качестве программного обеспечения для усвоения курса предлагается использовать продукт MS Excel, который располагает достаточными средствами анализа данных, такими как: пакет анализа, общими статистическими функциями мастера функций и т.д. Кроме того, предлагается дополнительно использовать подключаемый модуль StatPlus.

Дисциплина «Анализ данных» базируется на знании предметов «Информационные технологии», «Электронные таблицы», «Вероятность и статистика», изучаемых в образовательных учреждениях высшего образования.

Элементы курса «Анализ данных» используются при изучении курсов «Моделирование систем», «Проектирование информационных систем», «Надежность информационных систем».

Задача материала данного конспекта лекций в том, чтобы:

- предоставить студентам общие сведения о принципах обработки и анализа данных с целью получения из них новых сведений;
- показать методы, средства и технологии анализа данных;
- показать на примере регрессионного анализа принцип получения новых знаний из данных.

Знания и навыки, полученные в результате изучения данной дисциплины, могут быть применены:

1. при проведении анализа данных с целью получения статистической информации или прогноза ситуации;
2. для интерпретации полученных результатов в ходе анализа;
3. при формулировании технического задания при создании ИС силами профессиональных разработчиков.