

Основы интеллектуального анализа данных

Лекторы:

Цеховая Татьяна Вячеславовна,

к. ф.-м. н., доцент

Пашук Павел Александрович,

ст. преподаватель

Основы интеллектуального анализа данных

Лекции – **26 часов (10+16)**

Лабораторные занятия - **28 часов**

Отчетность - **зачет**

Мероприятия по текущему контролю знаний:

- **контрольные работы (40 %)**
- **отчеты по лабораторным работам (60 %)**

Основы интеллектуального анализа данных

Консультации:

ауд. 546, кафедра теории вероятностей и
математической статистики

Срочные вопросы и предложения:

Tsekhavaya@mail.ru

Литература (раздел 1 ОС):

1. *Айвазян, С. А.* Прикладная статистика и основы эконометрики / Айвазян С. А., Мхитарян В. С. – М.: Изд. объединение «ЮНИТИ», 1998.
2. *Ефимова, М. Р.* Общая теория статистики: Учебник / Ефимова М. Р., Петрова Е. В., Румянцев В. Н. – 2-е изд., М.: ИНФРА, 2006.
3. *Шмойлова, Р. А.* Теория статистики: Учебник / под ред. Шмойловой Р. А. – М.: Финансы и статистика, 2002.
4. *Кобзарь, А. И.* Прикладная математическая статистика / М.: ФИЗМАТЛИТ, 2006.

АКТУАЛЬНОСТЬ

В последние годы стремительное развитие получила область ***Data Science / Data Mining*** (в отечественной литературе наиболее используемая аналогия — ***интеллектуальный анализ данных***), направленная на поиск и разработку методов извлечения из имеющихся данных знаний, позволяющих принимать на их основе конкретные, в высокой степени обоснованные, практически полезные управленческие решения.

ТЕРМИНОЛОГИЯ

- Data Science;
- Data Mining;
- Big Data;
- Machine Learning;
- Deep Learning;
- Statistical Analysis and Data Mining;
- Predictive Analytics and Data Mining;
- Data Science and Data Mining;
- Discovery Driven Data Mining;
- Knowledge Discovery in Databases и др.

ПРИМЕРЫ ПРИМЕНЕНИЯ

Бизнес:

- 1) Розничная торговля;
- 2) Сфера развлечений;
- 3) Маркетинг, страхование, работа с персоналом.

Медицина

Государственное управление

РАЗДЕЛ 1. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

ТЕМА 1. Введение в описательную статистику

ТЕМА 1. ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Статистика — наука, изучающая количественную сторону массовых явлений в неразрывной связи с их качественной стороной.

Предметом статистики служат массовые общественные явления и процессы, а также складывающиеся в них количественные закономерности.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Статистическая закономерность
выражается в последовательности,
регулярности, повторяемости события с
достаточно высокой степенью вероятности,
если условия, порождающие это явление,
не изменяются или изменяются
незначительно.

ВВЕДЕНИЕ

Статистический метод включает :

- научно организованный сбор первичной статистической информации;
- сводка, обработка и группировка статистической информации;
- обобщение и интерпретация статистической информации.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Составные части статистики:

Описательная статистика – позволяет осуществить удобное представление данных для последующего анализа в виде частотных распределений, графических изображений и различных характеристик;

Математическая статистика – теория принятия статистических решений, позволяющая с помощью специальных методов обработки данных дать их правильную интерпретацию.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Этапы статистического исследования:

- Сбор первичного статистического материала.
- Предварительная обработка данных, подсчет групповых и общих итогов.
- Расчет и интерпретация обобщающих статистических показателей.
- Моделирование взаимосвязей между процессами и явлениями.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Основные категории статистики:

1. Статистическая совокупность.
2. Единица статистической совокупности.
3. Признак.
4. Статистический показатель.
5. Система статистических показателей.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

1. Статистическая совокупность – множество однокачественных (однородных) хотя бы по одному какому-либо признаку явлений.

- *генеральная совокупность*
- *выборочная совокупность*

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Генеральную совокупность образуют все элементы, анализируемые с некоторой точки зрения.

Выборочной совокупностью (**выборкой**) называют совокупность отобранных для изучения объектов из генеральной совокупности.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

2. Единица статистической

совокупности – это первичный элемент статистической совокупности, являющийся носителем признаков, подлежащих регистрации.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

3. Признак – это качественная особенность единиц совокупности.

По характеру отображения свойств единиц совокупности признаки делятся на:

- ***Качественные.***
- ***Количественные.***

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Качественные признаки делятся на:

- альтернативные,
- атрибутивные,
- порядковые.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Количественные признаки делятся на:

дискретные - могут принимать лишь отдельные значения из некоторого ряда чисел.

непрерывные - могут принимать любые значения в определенном интервале.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

4. **Статистический показатель** — понятие, которое отображает количественные характеристики соотношения признаков.
5. **Система статистических показателей** — совокупность статистических показателей, отображающая взаимосвязи, объективно существующие между изучаемыми явлениями.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Основные этапы статистической обработки данных

Этап 1. *Предварительный анализ
исследуемого процесса.*

Определяются:

- основная цель исследования;
- предмет статистического исследования;
- перечень статистических показателей;
- формализованная постановка задачи.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Этап 2. *Составление детального плана сбора исходной статистической информации.*

Этап 3. *Сбор исходных статистических данных.*

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Этап 4. *Первичная статистическая обработка исходных данных.*

- отображение текстовых переменных в номинальную или порядковую шкалу;
- анализ резко выделяющихся наблюдений;
- восстановление пропущенных наблюдений;
- проверка статистической независимости исходных данных.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Этап 5. *Составление детального плана вычислительного анализа собранного материала.*

Этап 6. *Вычислительная реализация.*

Этап 7. *Подведение итогов исследования.*

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Статистическое наблюдение – это массовое, планомерное, научно-организованное наблюдение за явлениями и процессами, заключающееся в регистрации необходимых признаков у каждой единицы изучаемой совокупности.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Формы статистического наблюдения:

- Отчетность;
- Специально организованное статистическое наблюдение;
- Регистры.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Виды статистического наблюдения:

- по ***охвату единиц совокупности***:
сплошное и несплошное;
- по ***времени регистрации***:
текущее, периодическое, единовременное;
- по ***источнику сведений***: документальное, опрос.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Требования к собираемым данным:

- **Достоверность** состоит в достаточных знаниях статистика, наличии программы наблюдения, статистического инструментария.
- **Сопоставимость** означает, что характеристики отдельных элементов можно легко обобщить на объект в целом.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

- *Планомерность статистического наблюдения.*
- *Полнота и массовость* – должно быть охвачено достаточно большое число случаев проявления изучаемого процесса для того, чтобы получить правдивые статистические данные.
- *Систематичность* – наблюдения проходят или систематически, или регулярно, или непрерывно.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Способы организации выборки:

1. Простой случайный отбор.
2. Простой отбор с помощью регулярной, но несущественной для изучаемого вопроса процедуры.
3. Стратифицированный (расслоенный) отбор.
4. Серийный (гнездовой) отбор .
5. Комбинированный (ступенчатый) отбор.
6. Последовательный (активный) отбор.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Точность статистического наблюдения – степень соответствия величины какого-либо показателя, определенного по материалам статистического наблюдения, действительной его величине.

Расхождение между расчетным и действительным значениями называется ошибкой наблюдения.

ВВЕДЕНИЕ В ОПИСАТЕЛЬНУЮ СТАТИСТИКУ

Различают следующие ошибки наблюдения:

- ❖ ошибки *регистрации*;
- ❖ ошибки *репрезентативности*.

Для выявления ошибок осуществляется
счетный и логический контроль.