Статистика для специалистов по бизнес-информатике
Введение

Д-р Вольфганг Бёме

Октябрь 2010

Содержание

- **1. Введение**
- 2. Дескриптивный одномерный анализ распределений
- 3. Дескриптивный корреляционный и регрессионный анализ
- 4. Дескриптивный анализ временных рядов
- 5. Относительные показатели / Индексы
- 6. Введение в избранные многомерные методы анализа

- Статистическое программное обеспечение: SPSS

Некоторые мнения о статистике

1. Не доверяй никакой статистике, которую сам не подделал! – Черчилль
2. Статистика – это наука о систематической лжи. – Клаус Брунштайн, *1937; немецкий информатик
3. Если голова в сауне, а ноги в холодильнике, статистики говорят о приятной средней температуре. – Франц Йозеф Штраус
4. Статистика – как бикини: то, что она открывает, многообещающе, а то, что скрывает, – существенно! – Аарон Левенштейн, нью-йоркский профессор делового администрирования
5. Некоторые (политики, менеджеры и др.) используют статистику, как пьяный уличный фонарь: чтобы держаться, а не для освещения!
6. Статистика – это метод, позволяющий выражать неточные данные с точностью до сотых долей процента.
7. Если охотник промахнулся мимо зайца слева и справа, то в среднем заяц мёртв.

1. Введение – Содержание

- **1.1 Задачи статистики**
- 1.2 Основные понятия
- 1.3 Этапы статистического исследования

Задачи и области статистики

Разрабатывать и применять методы и процедуры для сбора, обработки и анализа данных и интерпретировать результаты, полученные в ходе их применения

Области:

- Дескриптивная статистика (описательная статистика)
- Индуктивная статистика (выводная статистика)
- (Теория вероятностей)

Дескриптивная статистика

- Полный сбор данных об исследуемом объекте
- Обработка собранного массива данных; наглядная агрегированная форма (таблицы, графики...)
- Анализ обработанных данных (вычисление показателей, выявление закономерностей...)
- Предметная интерпретация результатов анализа

Индуктивная статистика

Ситуация

- Сбор данных об исследуемом объекте является неполным (выборочное исследование)
- Информация об исследуемом объекте представлена не полностью, а только для подмножества генеральной совокупности

Цель:

По данным о свойствах в подмножестве необходимо делать выводы о свойствах в вышестоящей генеральной совокупности.

Проблема:

Вывод сопряжён с риском ошибки.

Риск ошибки может быть количественно оценён при определённых условиях с помощью знаний из теории вероятностей.

Индуктивная статистика

Разделы индуктивной статистики:

- Разработка и применение методов выборки
- Разработка и применение методов оценки неизвестных статистических параметров
- Разработка и применение методов проверки гипотез
- Разработка и применение методов прогнозирования

Теория вероятностей

Термин «стохастика» происходит из греческого и означает приблизительно «искусство угадывания». Математическая стохастика занимается описанием и изучением случайных экспериментов.
Случайный эксперимент:

- Существует несколько возможных результатов (исходов) эксперимента.
- Результат при проведении эксперимента не может быть предсказан с уверенностью.
- Эксперимент воспроизводим в одинаковых условиях.

Задача:

Количественно оценить степень уверенности в том, что определённый исход случайного эксперимента наступит (вероятность).

1. Введение – Содержание

- 1.1 Задачи статистики
- **1.2 Основные понятия**
- 1.3 Этапы статистического исследования

1.2. Основные понятия

- **1.2.1 Статистическая единица, статистическая совокупность**
- 1.2.2 Статистические признаки, значения признаков
- 1.2.3 Измерение, измеренные значения
- 1.2.4 Статистические шкалы
- 1.2.5 Преобразование измеренных значений

Статистическая единица, статистическая совокупность

Статистическая единица – наименьший элемент в статистике. Статистическая единица является носителем информации или свойств, представляющих интерес для статистического исследования.

Обозначение: Z_i

Синонимы: носитель признака, единица наблюдения, элемент

Конечное множество $Z = \{Z_i, i = 1, 2, ..., n\}$ хорошо различимых, содержательно, пространственно и временно одинаково ограниченных статистических единиц Z_i называется **статистической совокупностью** Z объёма n.

Синонимы: статистическая масса, генеральная совокупность







1.2. Основные понятия

- 1.2.1 Статистическая единица, статистическая совокупность
- **1.2.2 Статистические признаки, значения признаков**
- 1.2.3 Измерение, измеренные значения
- 1.2.4 Статистические шкалы
- 1.2.5 Преобразование измеренных значений

Статистические признаки, значения признаков

Свойство статистической единицы Z_i, которое является основой или предметом статистического исследования, называется **статистическим признаком**.

Обозначение: X_i, Y_i, Z_i *bzw.* X, Y, Z

Синонимы: переменная, свойство

Высказывание о признаке или о свойстве статистической единицы Z_i называется **значением признака**.

Обозначение: x_i, y_i, z_i *bzw.* x, y, z

Синонимы: реализация, данные, наблюдение







1.2 Основные понятия

- 1.2.1 Статистическая единица, статистическая совокупность
- 1.2.2 Статистические признаки, значения признаков
- **1.2.3 Измерение, измеренные значения**
- 1.2.4 Статистические шкалы
- 1.2.5 Преобразование измеренных значений

Измерение, измеренные значения

Под **измерением** понимается присвоение чисел (измеренных значений) носителям признаков по определённому правилу.

Признак называется **непрерывным**, если между любыми двумя измеренными значениями всегда существует ещё одно измеренное значение. Если это не так, признак называется **дискретным**.

Истинное значение измерения – точное числовое значение, которое присваивается носителю признака идеальным измерительным прибором с бесконечной точностью. **Наблюдаемое значение измерения** – округлённое числовое значение, которое присваивается носителю признака реальным измерительным прибором с неточностями и погрешностями.

1.2. Основные понятия

- 1.2.1 Статистическая единица, статистическая совокупность
- 1.2.2 Статистические признаки, значения признаков
- 1.2.3 Измерение, измеренные значения
- **1.2.4 Статистические шкалы**
- 1.2.5 Преобразование измеренных значений

Статистические шкалы

Множество $M = \{m_j, j = 1, 2, ..., m\}$ всех m теоретически возможных или всех m эмпирически наблюдённых, хорошо различимых значений признака x_i статистического признака X, которые могут быть или были получены у n носителей признаков Z_i статистической совокупности $Z = \{Z_i, i = 1, 2, ..., n\}$, называется **множеством состояний M** статистического признака X.

Замечание: Совокупность наблюдённых значений признака x_i статистического признака может быть меньше множества состояний статистического признака X.

важно для понятия шкалы

Статистические шкалы

Отображение значений признака изучаемого признака на знаковое или числовое множество, сохраняющее отношения, называется **статистической шкалой**.

Отображение, сохраняющее отношения, на числовое множество называется **статистической измерительной шкалой**

При этом должно выполняться:

- Носителям признаков, одинаковым по признаку, присваивается одно и то же число
- Носителям признаков, различающимся по признаку, присваиваются разные числа







Статистические шкалы

Замечания:

- Число или знак (некоторые шкалы не требуют числового представления) $\Rightarrow$ различие между качественными и количественными признаками
- Качественным признакам могут присваиваться числовые значения.
- Если числа, то можно различать дискретные и непрерывные признаки
- Обработка признаков значительно упрощается благодаря числовым значениям

Статистические шкалы

Отношения между числами статистической шкалы:
(математические операции возможны только при числовом представлении!!)

- Различие [сравнение на тождественность] (допустимые математические операции: $=, \neq$)
- Порядок [сравнение по величине] (допустимые математические отношения: $<, >$)
- Расстояние [соседние числа имеют равные расстояния] (допустимые математические операции: $+, -$)
- Отношение [соседние числа имеют равное соотношение] (допустимые математические операции: $*, /$)

Уровни шкал:

- Номинальная шкала: различие
- Порядковая шкала: различие, порядок
- Интервальная шкала: различие, порядок, расстояние
- Шкала отношений: различие, порядок, расстояние, нулевая точка

Статистические шкалы – Номинальная шкала

Статистическая шкала, с помощью которой можно выражать лишь равенство или различие значений признаков, называется **номинальной шкалой**

Пример: Пол (мужской, женский)

статистическая совокупность: группа лиц
статистическая единица: отдельный человек
Признак: пол
Значение признака: мужской, женский или одно из двух
Множество состояний: мужской, женский

для математических операций необходимо перекодирование:
0 – мужской, 1 – женский

Статистические шкалы – Порядковая шкала

Статистическая шкала, с помощью которой можно выражать как равенство или различие, так и ранговый порядок значений признаков, называется **порядковой шкалой**

Пример: Выставление оценок: отлично, хорошо, ... или 1, 2, ...

статистическая совокупность: группа лиц
статистическая единица: отдельный человек
Признак: оценка по предмету «Статистика»
Значение признака: отлично, хорошо, ... или 1, 2, ...
Множество состояний: отлично, хорошо, ... или 1, 2, ...

Статистические шкалы – Метрические шкалы

Статистическая шкала, которая с помощью вещественных чисел может выражать как равенство или различие, так и ранговый порядок, а также измеримые и счётные различия (расстояние, кратность) для значений признаков, называется **метрической шкалой**

Различия метрических значений признаков:

(a) по виду операций

- Признак, измеренный по интервальной шкале $(+, -)$
- Признак, измеренный по шкале отношений $(*, /)$

(b) по характеру множества состояний

- Дискретный признак
- Непрерывный признак

Статистические шкалы – Примеры

Примеры (по математическим операциям):

- Интервальная шкала: шкала температуры в °C
  Разность возможна, но 20°C не вдвое теплее, чем 10°C
- Шкала отношений: процентные ставки
  Отношения можно вычислять,
  напр., 4 процента – это вдвое больше, чем 2 процента

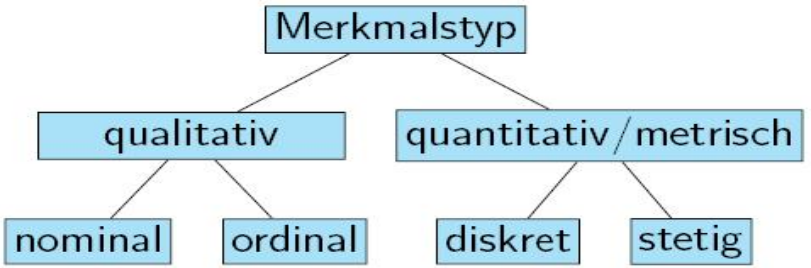
Статистические шкалы – Примеры

Примеры (по множеству состояний):

- Количество детей в семье
  Множество состояний: конечное, целочисленное
- Дискретный признак: цена книги в евро и центах
  Множество состояний: конечное, часто аппроксимируется как непрерывный признак.
- Непрерывный признак: процентные ставки
  В интервале потенциально возможно любое вещественное число.
  Множество состояний: бесконечно много значений в интервале.

Статистические шкалы

Klassifikation von Merkmalen: **Merkmalstypen**



Статистические шкалы

Замечания:

- Номинальные и порядковые шкалы не обязательно отображаются на числовые множества (Качественные признаки – Количественные признаки)
- Номинальные и порядковые шкалы также называются категориальными шкалами
- Интервальные шкалы и шкалы отношений также называются кардинальными или метрическими шкалами
- Признак с только двумя возможными значениями: дихотомический признак
- В SPSS различаются следующие шкалы: номинальная, порядковая и метрическая (шкала)

Статистические шкалы – простые примеры

Признак	Значение	
Семейное положение	женат/замужем, холост/не замужем, ...	
Возраст	$t \geq 0$, $t \in R$	
Возраст	$t \geq 0$, $t \in N$	
Возраст	0–10, 11–20, ...	
Цвет	красный, жёлтый, ...	
Цвет	длина волны в(μm)	
Оценка за экзамен	$0, 1, \ldots, 15$ (баллов)	
Результат экзамена	сдан, не сдан	







1.2. Основные понятия

- 1.2.1 Статистическая единица, статистическая совокупность
- 1.2.2 Статистические признаки, значения признаков
- 1.2.3 Измерение, измеренные значения
- 1.2.4 Статистические шкалы
- **1.2.5 Преобразование измеренных значений**

Преобразование измеренных значений

Определение: Преобразование – это однозначное отображение переменных (Каждому аргументу x функция f ставит в соответствие ровно одно значение y или $f(x)$:

$$f : x \longrightarrow y = f(x)$$

Уровень шкалы и допустимое преобразование:

Уровень шкалы	**Тип преобразования**	**Свойство преобразования**
Номинальная шкала	взаимно однозначное	$x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$
Порядковая шкала	строго монотонное	$x_1 < x_2 \Rightarrow f(x_1) < f(x_2)$
Интервальная шкала	положительно линейное	$\lvert x_1 - x_2\rvert = \lvert x_3 - x_4\rvert \Rightarrow$ $\lvert f(x_1) - f(x_2)\rvert = \lvert f(x_3) - f(x_4)\rvert$
Шкала отношений	растяжение	$x_1 = 0 \Rightarrow f(x_1) = 0$

Преобразование измеренных значений – Иллюстрация

1. Взаимно однозначное преобразование
2. Строго монотонное преобразование
3. Положительно линейное преобразование
4. Растяжение

1. Введение – Содержание

- 1.1 Задачи статистики
- 1.2 Основные понятия
- **1.3 Этапы статистического исследования**

Этапы статистического исследования

Пример:

Beispiel EU Vergleich.xls

1.3 Этапы статистического исследования

Содержание

- **1.3.1 Планирование исследования**
- 1.3.2 Сбор данных
- 1.3.3 Обработка данных
- 1.3.4 Анализ данных

Планирование исследования

Задачи

1. Определение статистической единицы и разграничение статистической совокупности
2. Определение признаков наблюдения
3. Формулировка исследовательских гипотез
4. Выбор методов сбора исходных данных
5. Выбор методов обработки и анализа данных
6. Вычислительное обеспечение сбора, хранения, обработки и анализа данных
7. Организационное, финансовое и кадровое обеспечение статистического исследования

Пример: Планирование исследования

Некоторые избранные постановки вопросов

1. Формулировка некоторых высказываний о состоянии ЕС (совокупность рассматриваемых государств-членов ЕС)
   - Средние значения
   - Отклонения от среднего
   - Положение отдельных стран относительно среднего или относительно друг друга
2. Взаимосвязи между собранными данными
3. Дополнительные желательные данные

1.3 Этапы статистического исследования

Содержание

- 1.3.1 Планирование исследования
- **1.3.2 Сбор данных**
- 1.3.3 Обработка данных
- 1.3.4 Анализ данных

Сбор данных

Процесс определения и регистрации значений одного или нескольких статистических признаков у отдельных единиц определённой статистической совокупности.

Результат сбора данных: первичный список

$$X_{(n,k)} = \begin{pmatrix} x_{\mathbf{11}} & x_{\mathbf{12}} & \dots & x_{\mathbf{1}k} \\ x_{\mathbf{21}} & x_{\mathbf{22}} & & x_{\mathbf{2}k} \\ \vdots & \vdots & \ddots & \\ x_{n\mathbf{1}} & x_{n\mathbf{2}} & & x_{nk} \end{pmatrix}$$

где

x_{ij} — значение j-го признака у i-го носителя признака

j — индекс признака $(j = 1, 2, \dots, k)$: $X_{\mathbf{1}}, X_{\mathbf{2}}, \dots, X_k$

i — индекс носителя признака $(i = 1, 2, \dots, n)$

Пример: Первичный список

$$X_{(n,k)} = \begin{pmatrix} 6861 & 73.0 & 6 & 21765.0 & 44.9 & 1.0 & 310.0 \\ 5150 & 72.5 & 7 & 28245.0 & 60.4 & 1.0 & 332.0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 3018 & 74.6 & 7 & 12201.0 & 37.1 & 4.2 & 104.0 \end{pmatrix}$$

где

x_{ij} — значение j-го признака у i-го носителя признака

j — индекс признака $(j = 1, 2, \ldots, 8)$: СТРАНА, ЭНЕРГИЯ, ..., ГАЗЕТА

i — индекс носителя признака $(i = B,\ DK,\ \ldots,\ E)$

1.3 Этапы статистического исследования

Содержание

- 1.3.1 Планирование исследования
- 1.3.2 Сбор данных
- **1.3.3 Обработка данных**
- 1.3.4 Анализ данных





Обработка данных

1. Проверка и корректировка собранных данных (полнота; правдоподобность; ....)
2. Преобразование данных
3. Упорядочение, обобщение и представление массива данных (элементарные статистические обработки, численное и/или графическое представление массива данных)

$\Rightarrow$ Импорт и обработка данных в SPSS

1.3 Этапы статистического исследования

Содержание

- 1.3.1 Планирование исследования
- 1.3.2 Сбор данных
- 1.3.3 Обработка данных
- **1.3.4 Анализ данных**







Анализ данных

Применение статистического метода, адекватного аналитической задаче и состоянию данных, к собранным и обработанным данным с целью получения новых знаний.

Проблемно-ориентированная оценка и использование результатов анализа для принятия решений.