Технологии организации, обработки и хранения статистических данных

ФИО преподавателя: Митина О.А.

e-mail: alogmi@yandex.ru

1

Лекция

Работа с данными. Описательная статистика

Условия обучения

- По итогам изучения дисциплины проводится экзамен
- В течение семестра необходимо выполнить все практические работы

Что такое данные

Данные – это воспринимаемые человеком факты, события, сообщения, измеряемые характеристики, регистрирующие сигналы.

Шкалы

- Шкала наименований.
- Шкала порядка.
- Интервальная шкала.
- Шкала отношений.

Шкала наименований

- каждый член некоторого множества объектов должен быть отнесен лишь к одному классу объектов (или к собирательному классу **прочие объекты);**
- ни один из объектов не может быть отнесен одновременно к двум или большему числу классов.

Порядковая шкала (ранговая шкала)

строится на отношении тождества и порядка и позволяет устанавливать предпочтения между различными объектами.

Шкала отношений (пропорциональная шкала)

классифицирует объекты пропорционально степени выраженности измеряемого свойства. Есть **абсолютный нуль (0).**

Метрические шкалы – это шкалы, у которых есть единицы измерения (например, метр, м/с). К ним относится шкала отношений.

Неметрические шкалы – это шкалы, у которых нет единицы измерений. К ним относятся шкала наименований, порядковая и интервальная шкала.

Выбор данных для заданной шкалы

Пример

Таблица понятий, относящихся к определенной шкале

Шкала наименований	Порядковая шкала	Интервальная шкала	Шкала отношений
номера предприятий или отделов, цвет глаз, тип автомобиля	ранги специалистов, порядок мест победителей	результат по десятибалльной шкале, температура, календарь	время выполнения задания, рост, вес, скорость ветра

Выбор данных для заданной шкалы

Пример

Таблица номинальных переменных

Переменная	Категории
Наличие машины	Да Нет
Владение акциями	Прибыльные Стабильные Другие
Провайдер	Мегафон МТС Билайн Теле2

Числовые (скалярные) переменные

определяют числовые величины, измеряемые на некоторой *интервальной шкале* (относительной шкале) или *шкале отношений* (абсолютной шкале).

Переменная	Шкала
Температура	Интервальная шкала
Экзаменационная оценка	Интервальная шкала
Грегорианский календарь	Интервальная шкала
Вес	Шкала отношений
Возраст в годах	Шкала отношений
Зарплата в рублях	Шкала отношений

Выбор данных для заданной шкалы

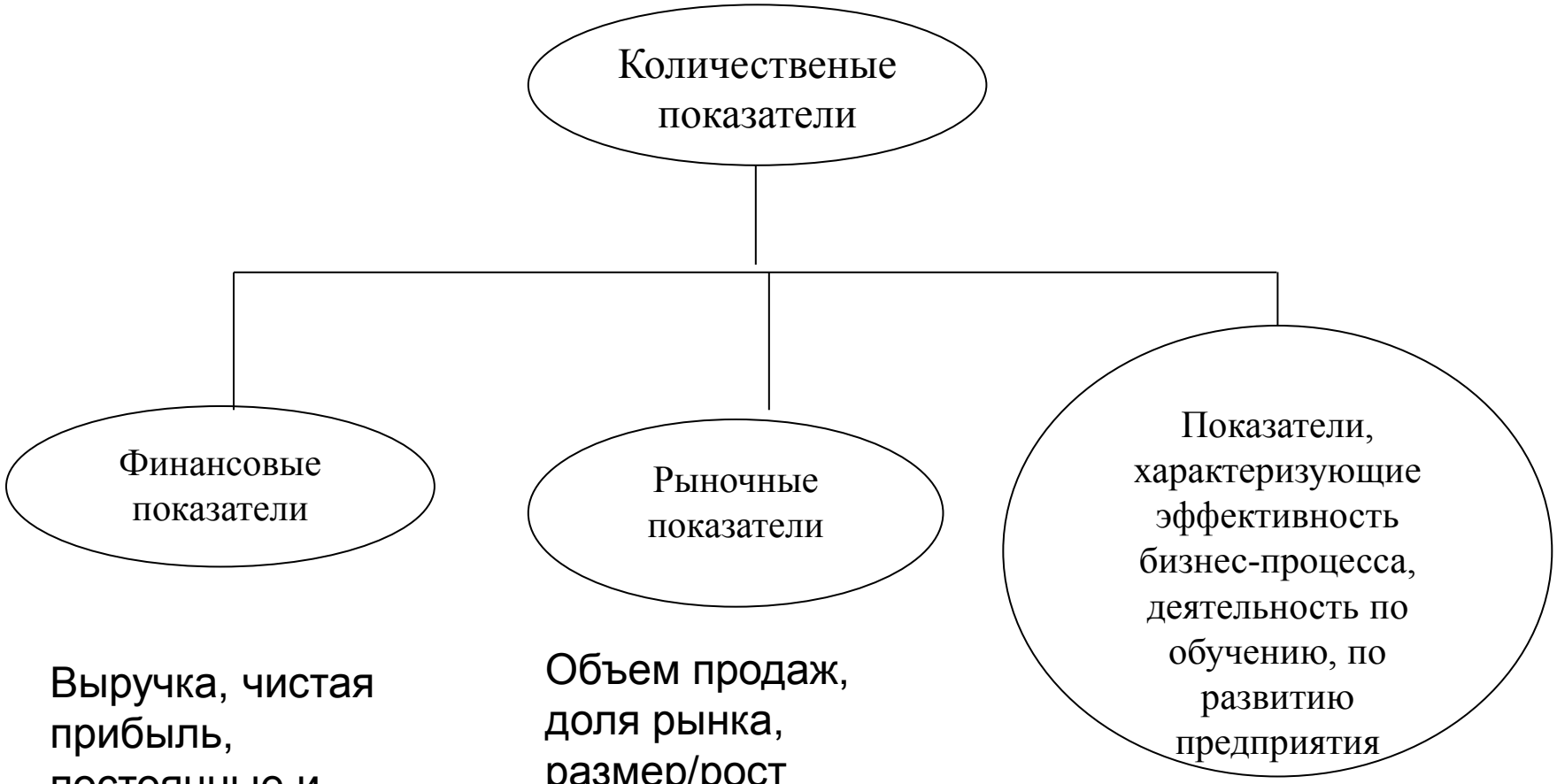
Шкалы	Центральная тенденция	Меры изменчивости	Связь
Шкала наименований	Количество объектов в классе, мода	Распределение процентных отношений	Сопряженность, коэффициент корреляции Чупрова
Порядковая шкала	Мода, медиана	Распределение процентных отношений, квантили	Коэффициенты корреляции Чупрова, Спирмена
Интервальная шкала, шкала отношений	Мода, медиана, среднее значение	Распределение процентных отношений, дисперсия, стандартное отклонение, коэффициент вариации	Коэффициенты корреляции Чупрова, Спирмена, Пирсона

Непрерывные и дискретные переменные

Дискретные данные являются значениями признака, общее число которых конечно или бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности.

Непрерывные данные – это данные, значения которых могут принимать какое угодно значение в некотором интервале.

Количественные данные

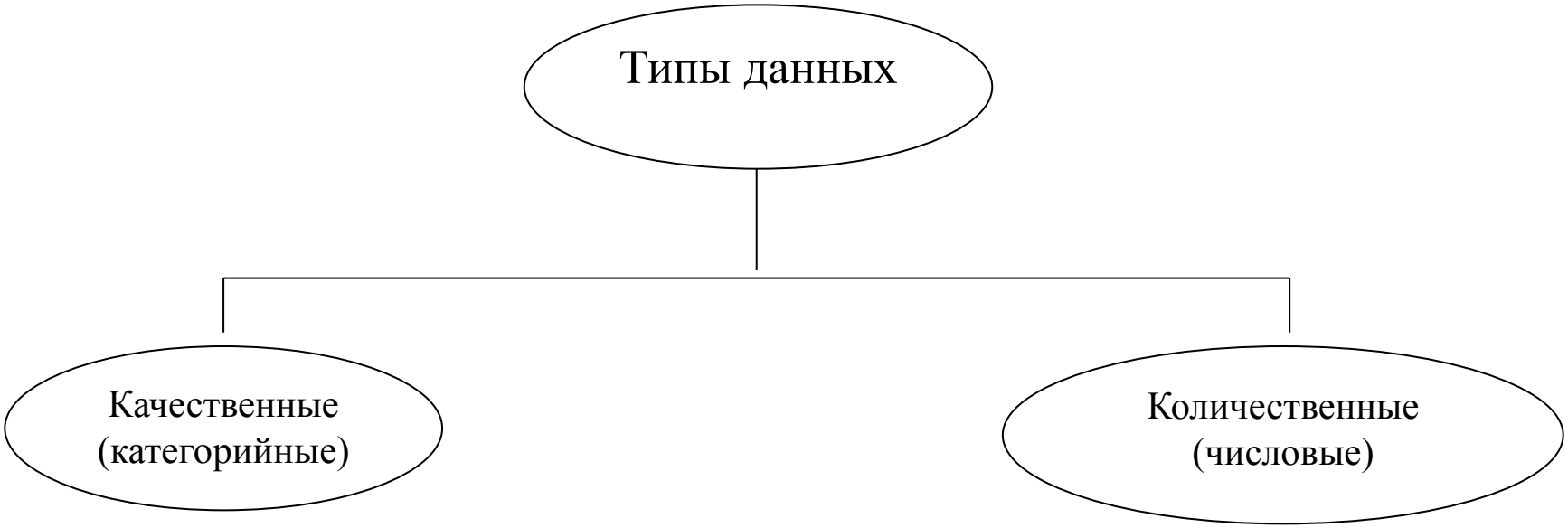


Выручка, чистая прибыль, постоянные и переменные издержки, показатели рентабельности

Объем продаж, доля рынка, размер/рост клиентской базы

Производительность труда, производственный цикл, время выполнения заказа, текучесть персонала

Типы данных



Данные, регистрирующие определенное качество, которое обладает объект или явление. Словесная форма

Порядковые (есть содержатель-ный смысл порядка)

Номинальные (нет содержательно интерпретируемого порядка

Значения переменных имеют содержательный смысл

дискретные

непрерывные)

Качественные и количественные данные

Пример 1

a) Количество телефонов:

b) Тип телефона:

c) Количество разговоров:

d) Продолжительность разговора:

e) Цвет телефона:

f) Оплата:

Качественные и количественные данные

Пример 2

2. Предположим, что от студентов, посещавших книжный магазин в студенческом городке на протяжении первой недели занятий, получена следующая информация:

   a) количество денег, потраченных на книги;

   b) количество приобретенных книг;

   c) количество времени, проведенного в магазине;

   d) академическая специализация студента;

   e) пол;

   f) владение персональным компьютером;

   g) количество курсов, посещаемых студентом в текущем семестре;

   h) покупал ли студент в книжном магазине какие-либо предметы одежды;

   i) способ оплаты покупки.

Определить, какие пункты опроса соответствуют категорийным переменным, а какие – числовым и по какой шкале измеряются.

Описательная статистика

- частоты;
- среднее значение;
- *мода*;
- *медиана;*
- *среднее квадратическое отклонение;*
- минимальное *и* максимальное значения *переменных;*
- *вариация;*
- *размах.*

Генеральная и выборочная совокупность

Под **генеральной совокупностью** понимается вся совокупность однотипных объектов, которые изучаются в данном исследовании.

Выборка *(выборочная совокупность)* - часть объектов из генеральной совокупности, отобранных для изучения, с тем чтобы сделать заключение о всей генеральной совокупности.

Вариация - колеблемость, изменяемость значения признака у отдельных единиц совокупности.

Объемом совокупности (выборочной или генеральной) называют число объектов этой совокупности.

Генеральная и выборочная совокупность

Ранжированный ряд - это ряд данных, расположенных в порядке возрастания (убывания) размера группировочного признака.

Статистическим рядом распределения называют упорядоченное распределение единиц совокупности на группы по изучаемому признаку.

Вариационный ряд распределения – ряд, построенный по количественному признаку.

Дискретный ряд распределения - это ряд, в котором варианты выражены целым числом.

Интервальный ряд распределения - это ряд, в котором значения признака заданы в виде интервала.

Генеральная и выборочная совокупность

Варианта - это отдельное значение варьируемого признака, которое он принимает в ряду распределения.

Пример. Пусть из генеральной совокупности извлечена выборка, причем x_1 наблюдалось n_1 раз, x_2 - n_2 раз,..., x_k – n_k раз и объем выборки:

$$\sum_{i=1}^{k} n_i = n$$

x_i - варианты, n_i - частоты, а их отношения к объему выборки (n) w_i - относительные частоты

Генеральная и выборочная совокупность

Накопленные частоты показывают, сколько единиц совокупности имеют значение признака не больше, чем рассматриваемое, и определяются последовательным суммированием частот, предшествующих данному значению признака.

Генеральная и выборочная совокупность

Пример

Из большой группы предприятий одной из отраслей промышленности случайным образом отобрано 30, по которым получены показатели основных фондов в млн. руб.:

3; 4; 2; 3; 3; 6; 5; 2; 4; 7; 5; 5; 3; 4; 3; 2; 6; 7; 5; 4; 3; 4; 5; 7; 6; 2; 3; 6; 6; 4.

Построим дискретный вариационный ряд.

Решение

Различные значения признака запишем в порядке возрастания и под каждым из них запишем соответствующие частоты. Получим дискретное статистическое распределение выборки:

x_i	2	3	4	5	6	7
n_i	4	7	6	5	5	3

Проверка: сумма всех частот должна быть равна объему выборки:

$n = 4 + 7 + 6 + 5 + 5 + 3 = 30$.

Найдем относительные частоты: $w_1 = \frac{4}{30} = 0{,}13$; $w_2 = \frac{7}{30} = 0{,}23$; $w_3 = \frac{6}{30} = 0{,}2$; $w_4 = \frac{5}{30} = 0{,}17$; $w_5 = \frac{5}{30} = 0{,}17$; $w_6 = \frac{3}{30} = 0{,}1$.

Составим таблицу распределения относительных частот:

x_i	2	3	4	5	6	7
w_i	0,13	0,23	0,2	0,17	0,17	0,1

Генеральная и выборочная совокупность

Пример

Выборочно обследовано 26 предприятий легкой промышленности по валовой продукции. Получены следующие результаты в млн. руб.:

15,0; 16,4; 17,8; 18,0; 18,4; 19,2; 19,8; 20,2; 20,6; 20,6; 20,6; 21,3; 21,4; 21,7; 22,0; 22,2; 22,3; 22,7; 23,0; 24,2; 24,2; 25,1; 25,3; 26,0; 26,5; 27,1.

Построим интервальное распределение выборки с началом $x_0 = 15$ и длиной частичного интервала $h = 2,5$.

Решение

Для составления интервального распределения составим таблицу. В первой строке расположим в порядке возрастания интервалы, длина каждого из которых $h = 2,5$. Во второй сроке запишем количество значений признака в выборке, попавших в этот интервал (т.е. сумму частот вариант, попавших в соответствующий интервал).

Частичный интервал	15-17,5	17,5-20	20-22,5	22,5-25	25-27,5
Частота интервала	2	5	10	4	5

Объем выборки $n = 2 + 5 + 10 + 4 + 5 = 26$.

Список литературы

- Тюрин Ю.Н. Анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров. – М.: МЦНМО, 2016. – 368 с.
- Мхитарян В.С. Анализ данных: учебник для академического бакалавриата / под ред. В.С. Мхитаряна. – М.: Изд. Юрайт, 2017 – 490 с.
- Хрусталёв Е.М. Агрегация данных в OLAP-кубах. http :// www . olap . ru /

Темы дисциплины

1 Работа с данными. Описательная статистика

2 Анализ данных. Бизнес-аналитика. Основные понятия и определения

3 Концепция хранилища данных. Понятие хранилища данных

4 Многомерная модель данных

5-6 Интеграция данных и бизнес-аналитика

7-8 Интеграция данных

9 Хранилища данных

10 Процессы информативной корпоративной фабрики

11 Базовые архитектуры корпоративной информационной фабрики

12 Технология OLAP и ее особенности

13 Понятие OLAP-куба. Операции над OLAP-кубами

14 Аналитические платформы. Инструменты бизнес-аналитики

15-16 Большие данные. Наука о данных

Спасибо за внимание!