# Проверка статистических гипотез

Часть 4 Аббакумов Вадим Леонардович Версия 4

#### Раздел 3

Популярные статистические критерии

#### Гипотеза независимости

 Н<sub>0</sub>: Случайные величины X и Y независимы

■ Н₁: Случайные величины X и Y зависимы

## Комментарий

 На практике отвечаем на вопрос: переменная X влияет на переменную Y?

- Если неизвестно,
- X влияет на Y или
- Y влияет на X
- то статистический критерий не поможет!

Пример Бернарда Шоу

 Гибридизация нескольких методов распознавания образов

## Диаграмма рассеивания

Иногда пишут - диаграмма рассеяния

Пример – швейцарские банкноты.

#### Зависимость -1

- X в количественной шкале
- Y в количественной шкале

- Применяется коэффициент корреляции Пирсона
- Или Спирмена
- Иногда Кендалла

## Функциональная зависимость

#### Статистическая зависимость

- Обобщение функциональной зависимости.
- Одному и тому же значению х могут соответствовать разные значения у.
- Например, один и тот же товар (например, телефон) может продаваться в разных магазинах по разной цене, то есть одному и тому же товару соответствуют разные цены.

#### Статистическая зависимость

- Определение
   статистическая зависимость это функциональная
   зависимость СРЕДНЕГО значения переменной у от значения
   переменной х.
- Откуда появляется среднее значение? Проводятся эксперименты (или наблюдается явление) при одном и том же значении x, при этом регистрируются разные значения y, затем эти значения усредняются.
- На практике не всегда заметно, что одному и тому же значению переменной х может соответствовать много значений у, например когда повторные наблюдения при одном значении х не делались.

среднее значение переменной у равно натуральному логарифму значения х.

среднее значение переменной у равно натуральному логарифму значения х.

 Коэффициент корреляции как «градусник», измеряющий степень зависимости

Формула для коэффициента корреляции

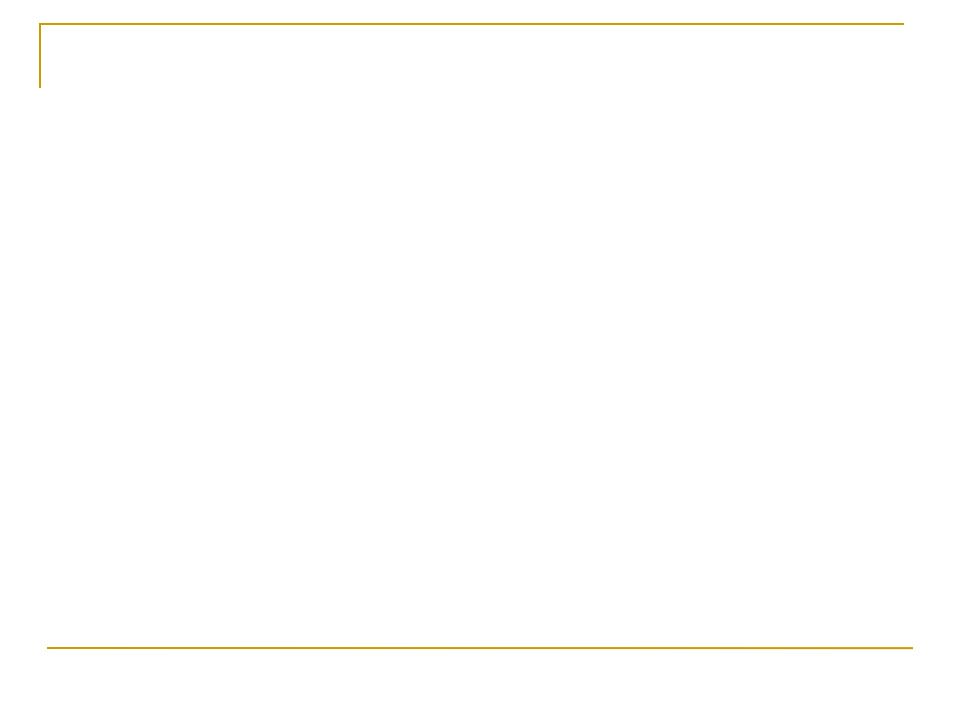
## Выбор коэффициента

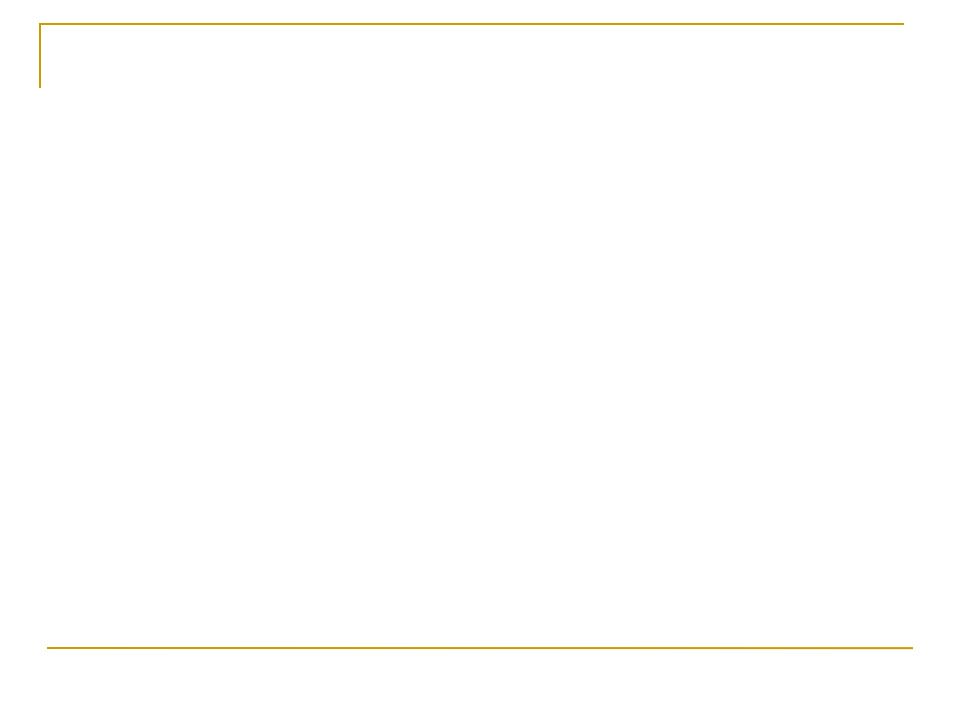
- Если распределение каждой переменной несущественно отличается от нормального, применяется коэффициент корреляции Пирсона
- В остальных случаях коэффициент корреляции Спирмена
- Вместо коэффициента корреляции
  Спирмена используют коэффициент корреляции Кендалла

Интервал значений коэффициента корреляции	Интерпретация
0 - 0,2	Очень слабая корреляция
0,2 - 0,5	Слабая корреляция
0,5-0,7	Средняя корреляция
0,7-0,9	Высокая корреляция
0,9 - 1	Очень высокая корреляция

Как проявляется зависимость на диаграмме рассеивания

 Проблемы и ошибки при использовании коэффициента корреляции





## Данные без выброса коэффициент корреляции равен -0.81

Добавлен выброс в точке (10,10).

Коэффициент корреляции упал до -0,55.

## Выброс сдвинут в точку (18,5, 18,5) Коэффициент равен 0

Выброс сдвинут в точку (53, 53). Корреляция равна +0,81

- Ложная корреляция

#### Зависимость -2

- X в количественной шкале
- Y в номинальной шкале

- Сравниваем средние или медианы в группах
- Или перекодируем количественную переменную, переводим ее в номинальную шкалу

#### Зависимость -3

- X в порядковой шкале
- Y в порядковой шкале

- Используем коэффициент корреляции Спирмена
- Или Кендалла

#### Spearman rho или Kendall-tau

Difference between Spearman and Kendall-Tau correlation test

https://stats.stackexchange.com/questions/309901/difference-between-spearman-and-kendall-tau-correlation-test?rq=1

Kendall Tau or Spearman's rho?

https://stats.stackexchange.com/questions/3943/kendall-tau-or-spearmans-rho#:~:text=Again %20somewhat%20philosophical%20answer%3B %20the,statistic%20for%20nonlinear %20correlation%20test.

#### Зависимость -4

- X в номинальной шкале
- Y в номинальной шкале

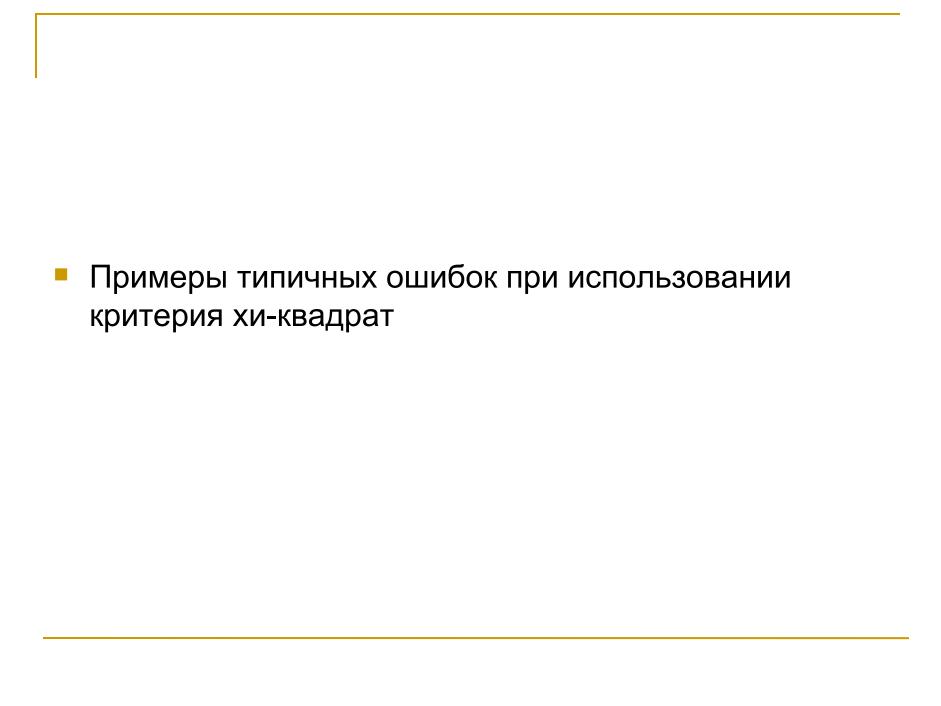
■ Таблица сопряженности и критерий х²

Критерий хи-квадрат

Формула для статистики

# Статистика хи-квадрат как коэффициент корреляции

- Коэффициент Пирсона
- Коэффициент Чупрова



При анализе подгрупп можно узнать много неожиданного...

- Действительно ли использование Internet связано с полом?
- Все опрошенные пользуются Интернетом.
  Тех из них, кто использует Интернет пять часов в месяц или меньше, отнесли к мало пользующимся, остальных к активным пользователям.

- sex = пол.
- Кодировка: "1" мужчина, "0" женщина.
- internet = использование Internet.
- Кодировка: "0" использует мало, "1" использует активно.
- Имеется 30 наблюдений (опрошенных).

 В результате изучения связи между покупкой модной одежды и семейным положением получены, среди прочих, следующие данные.

 Имеется 1000 наблюдений (опрошенных).

- Переменные.
- sex = пол.
- Кодировка: "1" мужчина, "0" женщина.
- marriage = семейное положение.
- Кодировка: "1" женат/замужем, "0" не женат/не замужем.
- fashion = покупка модной одежды.
- Кодировка: "0" покупает мало, "1" покупает много.

- Маркетолог проводит исследование для рекламного агентства, разрабатывающего рекламу для автомобилей стоимостью свыше 30 тысяч долларов.
- Он пытается проанализировать факторы, влияющие на владение дорогими автомобилями.

- Переменные.
- high\_edu = образование.
- Кодировка: "1" высшее образование, "0" нет высшего образования.
- expe\_car = наличие дорогого автомобиля.
- Кодировка: "0" дорогого автомобиля нет, "1" дорогой автомобиль есть.
- income = доход.
- Кодировка: "0" низкий доход, "1" высокий доход.
- Имеется 1000 наблюдений (опрошенных).

- Маркетолог, исследующий сферу туристических поездок за границу, предположил, что на желание путешествовать влияет возраст.
- Имеющиеся в его распоряжении данные содержат, среди прочего, следующую информацию.

- Переменные.
- desire = желание совершить путешествие за границу.
- Кодировка: "1" желание есть, "0" желания нет.
- sex = пол.
- Кодировка: "0" женщина, "1" мужчина.
- age = возраст.
- Кодировка: "0" –до 45 лет, "1" 45 лет или старше.
- Имеется 1000 наблюдений (опрошенных).

- Результаты анкетирования о проведении семейного досуга содержат, среди прочего, следующую информацию.
- Переменные.
- fastfood = частота посещения ресторанов быстрого питания.
- Кодировка: "1" часто, "0" редко.
- income = доход семьи.
- Кодировка: "1" высокий, "0" низкий.
- family = размер семьи.
- Кодировка: "1" большая семья, "0" малая семья.