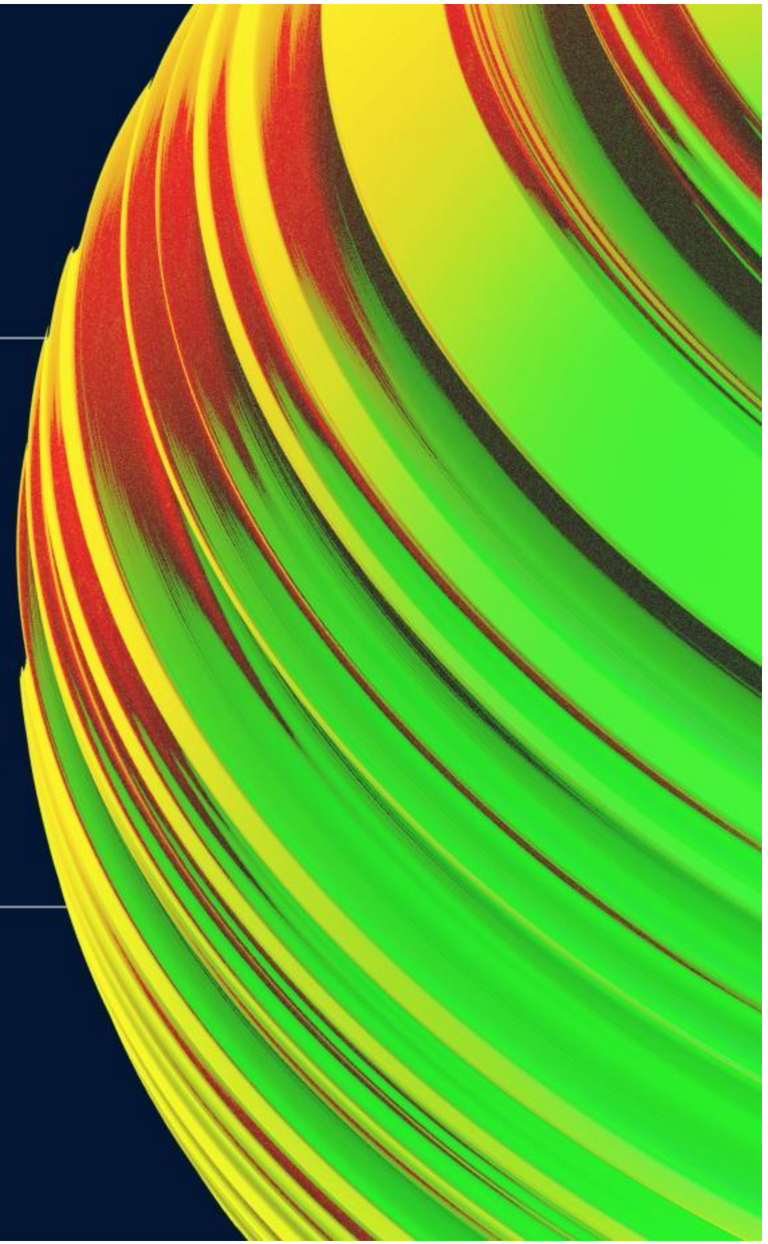
УНИВЕРСИТЕТ ИННОПОЛИС

# Непараметрическая статистика.
## KS-test. Bootstrap

Воробьёва Мария

- maria.vorobyova.ser@gmail.com
- @SparrowMaria



# Определение

Непараметрическая статистика — это раздел статистики, который не предполагает конкретную форму распределения для анализируемых данных.

В отличие от параметрической статистики, которая основывается на определенных допущениях о распределении (например, нормальном распределении), непараметрические методы более гибкие и применимы к данным, которые не соответствуют этим допущениям.

# Основные характеристики непараметрической статистики

**Отсутствие предположений о распределении:**

- Непараметрические методы не требуют предположений о форме распределения данных (например, нормального распределения), что делает их более универсальными и гибкими.

**Применение к малым выборкам:**

- Эти методы часто используются при анализе данных небольших объемов, где параметрические методы могут быть неэффективны.

**Использование рангов и медиан:**

- Непараметрические методы часто работают с ранговыми данными или медианами вместо средних значений, что делает их менее чувствительными к выбросам и асимметрии распределения.

# Примеры непараметрических статистических тестов

**Критерий знаков (Sign test):**

Используется для **проверки гипотез о медиане** одного выборочного набора данных или о различиях медиан двух связанных выборок

**Критерий Манна-Уитни (Mann-Whitney U test):**

Альтернатива t-тесту для независимых выборок. Сравнивает **два независимых набора данных.** Для оценки различий между двумя выборками по признаку, измеренному в количественной или порядковой шкале. U-критерий является ранговым, поэтому он инвариантен по отношению к любому монотонному преобразованию шкалы измерения

**Критерий Вилкоксона (Wilcoxon signed-rank test):**

непараметрический статистический критерий, применяемый для оценки **различий между двумя зависимыми выборками**, взятыми из закона распределения, отличного от нормального, либо измеренными с использованием порядковой шкалы.

# Примеры непараметрических статистических тестов

**Критерий Крускала-Уоллиса (Kruskal-Wallis test):**

Используется для определения наличия статистически значимой разницы между медианами трех или более независимых групп. Альтернатива однофакторному дисперсионному анализу (ANOVA) для сравнения более двух независимых групп

**Критерий Колмогорова-Смирнова (Kolmogorov-Smirnov test):**

Используется для сравнения эмпирического распределения с теоретическим или для сравнения двух эмпирических распределений.

# Преимущества и недостатки

**Преимущества**:

- Меньшая чувствительность к выбросам и несимметричным распределениям.
- Универсальность и возможность применения к различным типам данных.
- Эффективность при работе с небольшими выборками

**Недостатки**:

- Меньшая мощность тестов по сравнению с параметрическими методами, если данные действительно следуют предполагаемому распределению.
- Возможность потери информации, так как часто используется ранжирование данных.

# Преимущества и недостатки

**Недостатки**:

- Меньшая мощность тестов по сравнению с параметрическими методами, если данные действительно следуют предполагаемому распределению.
- Возможность потери информации, так как часто используется ранжирование данных.

напоминание:

Мощность статистического теста — это вероятность того, что тест правильно отвергнет нулевую гипотезу, когда альтернативная гипотеза верна. Другими словами, мощность теста показывает его способность обнаруживать реальные эффекты или различия в данных.

# Практика

https://colab.research.google.com/drive/1D3--CYmejjacbrJ0qZrylSzos8MrdS0k?usp=sharing

# Колмогоров-Смирнов тест (K-S тест)

Колмогоров-Смирнов тест (K-S тест) используется для сравнения двух распределений. Он определяет, насколько однообразно распределены данные в двух выборках или насколько одна выборка соответствует теоретическому распределению.

**Нулевая гипотеза ($H_0$):** Два выборочных распределения не отличаются (они происходят из одной и той же генеральной совокупности) или выборка соответствует теоретическому распределению.

**Альтернативная гипотеза ($H_1$):** Два выборочных распределения различаются (они происходят из разных генеральных совокупностей) или выборка не соответствует теоретическому распределению.

# Колмогоров-Смирнов тест (K-S тест)

**Статистика**: Расстояние между эмпирическими распределениями двух выборок или между эмпирическим распределением и теоретическим.

**Интерпретация**: Чем больше статистика теста, тем больше различие между распределениями.

Формально, статистика теста вычисляется как максимальное отклонение между двумя кумулятивными распределениями:

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|$$

где:

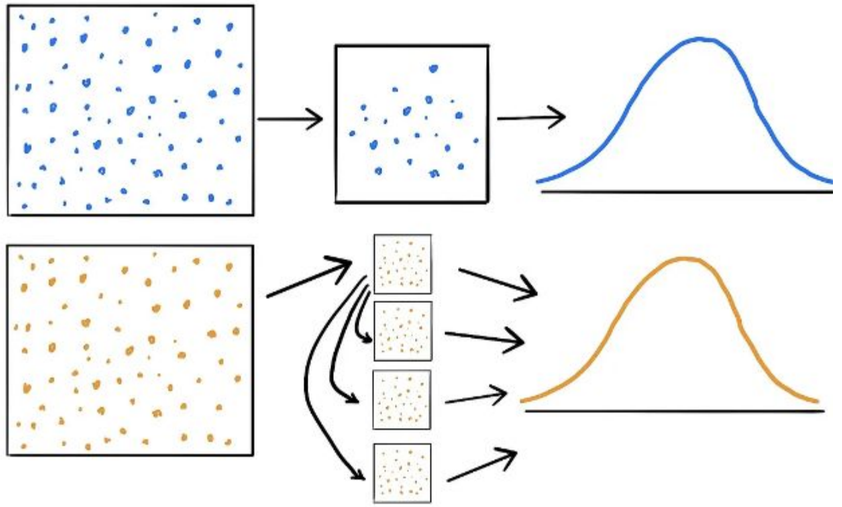
- $F_n(x)$ и $G_m(x)$ — эмпирические функции распределения двух выборок.
- $\sup$ обозначает верхнюю грань (супремум).

# Практика

https://colab.research.google.com/drive/14_EOpEIGwg9NYEIG9Rml0jumAWDoCrje?usp=sharing

# Bootstrap

Метод статистического анализа, который использует повторное выборочное извлечение из данных с заменой (т.е. один и тот же элемент может быть выбран более одного раза) для оценки распределения статистики выборки.



Этот метод полезен для оценки надежности оценок и создания доверительных интервалов, особенно когда распределение данных неизвестно или сложное.

# Bootstrap

Пусть имеются два наблюдения $(x_1, y_1) = (1,1), (x_2, y_2) = (2,3)$

Предположим, что нам необходимо оценить параметр в регрессии $y$ на $x$

$$y_i = \theta x_i + \varepsilon_i$$

Оценка параметра, методом наименьших квадратов будет равна

$$\hat{\theta} = \frac{x_1 y_1 + x_2 y_2}{x_1^2 + x_2^2} = \frac{1 \cdot 1 + 2 \cdot 3}{1^2 + 2^2} = \frac{7}{5}$$

# Bootstrap

Функция распределения будет

$$(x, y)' = \begin{cases} (1,1)', p = 1/2 \\ (2,3)', p = 1/2 \end{cases}$$

Данные из двух наблюдений будут распределены так:

$$(x_1, y_1)', (x_2, y_2)' = \begin{cases} (1,1)', (1,1)', p = 1/4 \\ (1,1)', (2,3)', p = 1/4 \\ (2,3)', (1,1)', p = 1/4 \\ (2,3)', (2,3)', p = 1/4 \end{cases}$$

Распределение Bootstrap

# Bootstrap

$$(x_1, y_1)', (x_2, y_2)' = \begin{cases} (1,1)', (1,1)', p = 1/4 \\ (1,1)', (2,3)', p = 1/4 \\ (2,3)', (1,1)', p = 1/4 \\ (2,3)', (2,3)', p = 1/4 \end{cases}$$

Распределение Bootstrap

Далее можем найти распределение МНК-оценки

$$\hat{\theta}_2^* = \begin{cases} 1, \ p = 1/4 \\ 7/5, \ p = 1/2 \\ 3/2, \ p = 1/4 \end{cases}$$

# Доверительный Интервал Bootstrap

Пусть дана выборка $(z_1, z_2, \ldots, z_n)$ из генеральной совокупности и требуется **оценить параметр** $\theta$. Необходимо выбрать количество $B$ псевдовыборок, которые будут формироваться из элементов исходной выборки. Для каждой из псевдовыборок $(z_1^*, z_2^*, \ldots, z_n^*)_b$, $b = 1,2,...,B$ вычисляется псевдостатистика $\hat{\theta}_b^*$.

Псевдостатистики $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$ **сортируются** от меньшей к большей.

Квантилями $q_{\alpha_1}^*, q_{1-\alpha_2}^*$ принимаются значения $\hat{\theta}_{[B\alpha_1]}^*, \hat{\theta}_{[B(1-\alpha_2)+1]}^*$ с их помощью строится **доверительный интервал**.

# Список рекомендуемой литературы


Ивченко Г. И., Медведев Ю. И., Введение в математическую статистику https://disk.yandex.ru/i/waXgDQWDh_rgTA

М. Б. Лагутин Наглядная математическая статистика http://iosipoi.com/teachingfiles/stat/Lagutin.pdf

Бородин А. Н., Элементарный курс теории вероятностей и математической статистики
https://disk.yandex.ru/i/Ubk5YLMk_PJjYw

Боровков А. А., Математическая статистика https://disk.yandex.ru/i/212K-4gWWwjQzA

Larry A. Wasserman All of Statistics: A Concise Course in Statistical Inference
https://egrcc.github.io/docs/math/all-of-statistics.pdf

Натан А. А., Горбачев О. Г., Гуз С. А., Математическая статистика https://disk.yandex.ru/i/gtKNf7r9uTNluw

Ушаков В. Г., конспекты лекций по математической статистике (ВМК МГУ, https://disk.yandex.ru/i/yx8zyo-oLIjwkQ

Zhou Fan (Stanford University) https://web.stanford.edu/class/archive/stats/stats200/stats200.1172/lectures.html
Philippe Rigollet (MIT) https://ocw.mit.edu/courses/18-650-statistics-for-applications-fall-2016/pages/lecture-slides/
Larry Wasserman (Carnegie Mellon University) https://www.stat.cmu.edu/~larry/=stat705/

УНИВЕРСИТЕТ
ИННОПОЛИС

# Спасибо!