

Непараметрические тесты, критерии согласия, бутстррап

План

- Критерий знаков, ранговые критерии
- Бутстррап
- Эмпирическая функция распределения
- Критерии согласия

Непараметрические критерии

Непараметрические критерии

- Проверка гипотезы – способ посчитать расстояние
- На прошлой неделе мы обсудили параметрические критерии
- Мы предполагали вид распределения и получали критерий
- Можно считать расстояния без предположений о виде распределения → **непараметрические критерии**
- Помогают тестировать гипотезы в ситуациях, когда распределение нестандартное

Непараметрические критерии

Параметрические критерии:

- Нулевая гипотеза формулируется о конкретных параметрах распределения
- Перед проведением теста на выборку накладываем предположения о распределении

Непараметрические критерии:

- Свойства распределения неизвестны
- Используют только информацию из выборки

Критерии знаков

Идея критерия

- Превратим выборку в нули и единицы
- Часть информации о выборке мы теряем
- Но зато всегда можем воспользоваться биномиальным распределением

Критерии знаков (одновыборочный)

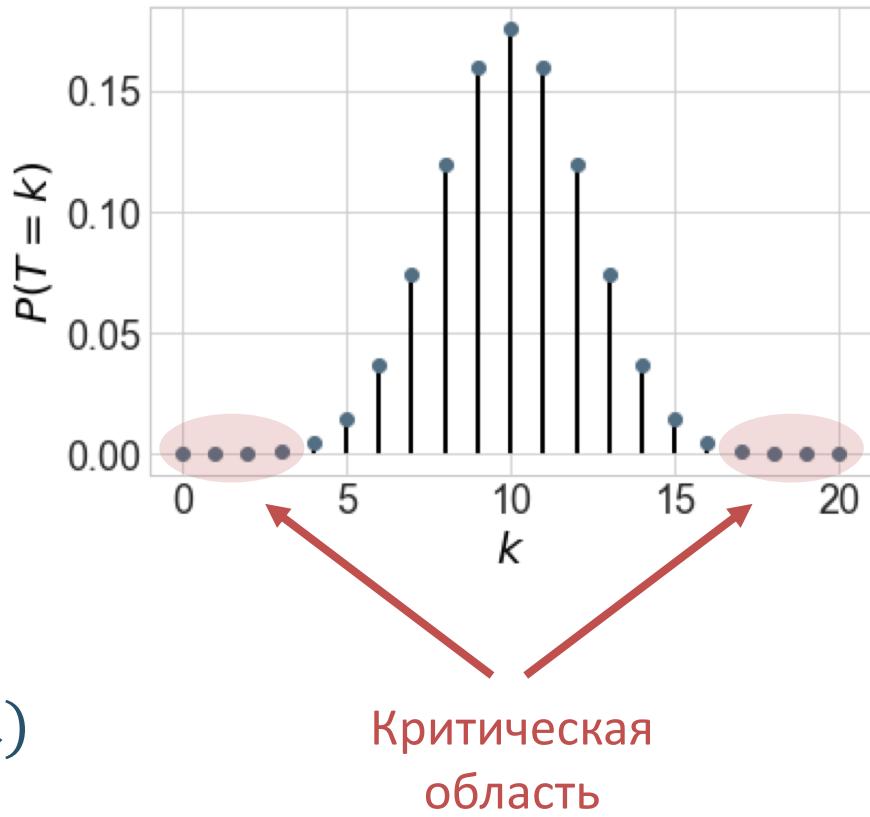
$X_1, \dots, X_n \sim iid$

$H_0: Med(X) = m_0$

$H_a: Med(X) \neq m_0$

Критерий для проверки:

$$T = \sum_{i=1}^n [X_i > m_0] \sim Bin(0.5, n)$$



Задача о попугаях

Пример: есть данные о говорливости попугаев (слов в день), правда ли, что в среднем попугай говорят больше 5 слов?

6, 4, 4, 7, 8

1, 0, 0, 1, 1

$$T_{\text{набл.}} = 3 \quad < \quad T_{\text{кр.}} = T_{1-\alpha} = 5$$

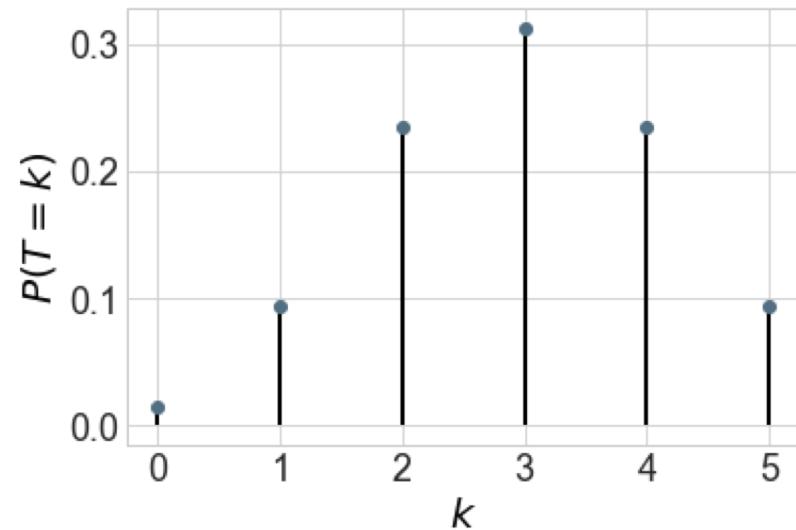
$$H_0: Med(X) = 5$$

$$H_a: Med(X) > 5$$

$$\alpha = 0.05$$



Гипотеза
не отвергается



Критерии знаков (двухвыборочный)

$X_1, \dots, X_n \sim iid$

$Y_1, \dots, Y_n \sim iid$

Выборки связанные

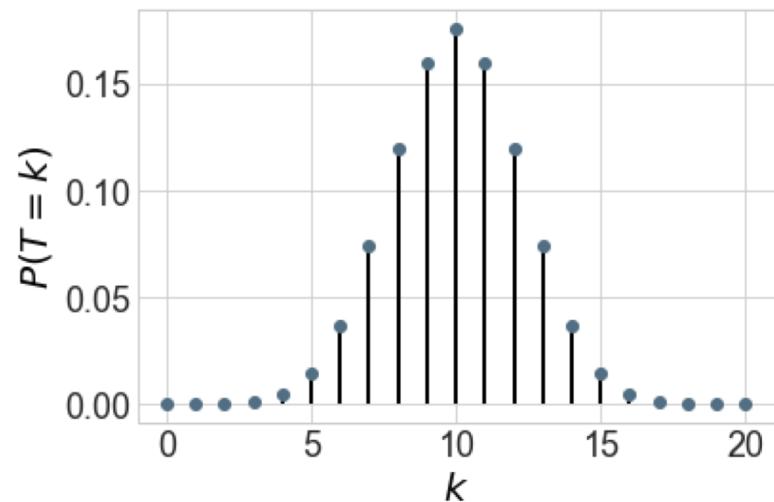
$H_0: \mathbb{P}(X > Y) = 0.5$

$H_a: \mathbb{P}(X > Y) \neq 0.5$

Критерий для проверки:

$$T = \sum_{i=1}^n [X_i > Y_i] \sim Bin(0.5, n)$$

! Снова превращаем выборки в нули и единицы, но немного другим способом



Задача про апелляцию

Пример: даны баллы студентов до и после апелляции.
Правда ли, что в среднем апелляция не повышает балл
за контрольную?

До: 48, 54, 67, 56, 55, 55, 90, 71, 72, 69

После: 47, 52, 60, 60 58, 60, 70, 81, 87, 60

0 0 0 1 1 1 0 1 1 0

$$T_{\text{набл.}} = 5$$

$$T_{\text{кр.}} = 8$$

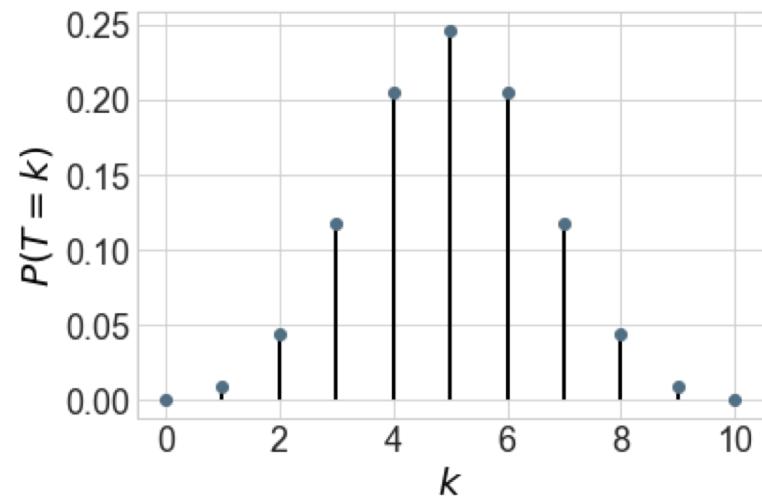
$$H_0: \mathbb{P}(X > Y) = 0.5$$

$$H_a: \mathbb{P}(X > Y) > 0.5$$

$$\alpha = 0.05$$



Гипотеза
не отвергается



Резюме

- Критерий знаков позволяет проверять гипотезу о равенстве медианы и отсутствии сдвига в связных выборках
- Критерий знаков игнорирует величину изменений и обращает внимание только на её направление
- Из-за этого происходит потеря части информации о выборке

Ранговые критерии

Идея критерия

- Критерии знаков превращают выборку в нули и единицы, из-за этого теряется информация
- Чтобы сохранить больше информации о выборке, можно превращать наблюдения в ранги

Ранг наблюдения

x_1, x_2, \dots, x_n – выборка

$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ Упорядочим по возрастанию

Правила выставления ранга:

1. Порядковый номер наблюдения – ранг
2. Если встречаются несколько одинаковых значений, им присваивается одинаковое значение ранга, равное среднему арифметическому их порядковых номеров

Критерии Уилкоксона (одновыборочный)

$X_1, \dots, X_n \sim iid$

$F_X(x)$ симметрична
относительно медианы

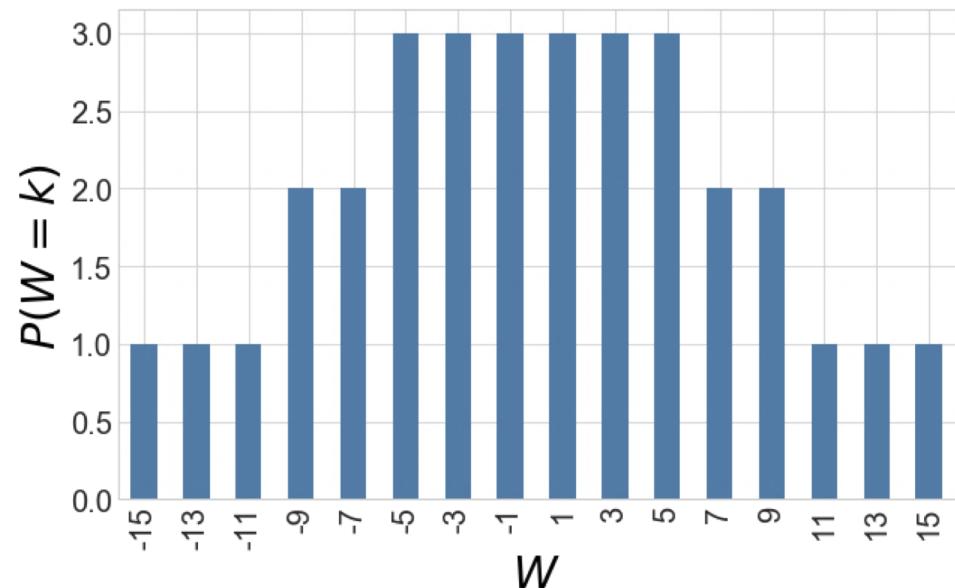
$H_0: Med(X) = m_0$

$H_a: Med(X) \neq m_0$



У статистики табличное
распределение

Распределение для $n = 5$



Критерий
для проверки:

$$W = \sum_{i=1}^n rank(|X_i - m_0|) \cdot sign(X_i - m_0)$$

Распределение статистики

$$W = \sum_{i=1}^n rank(|X_i - m_0|) \cdot sign(X_i - m_0)$$

Пусть в выборке пять наблюдений:

1	2	3	4	5
—	—	—	—	—
+	—	—	—	—
—	+	—	—	—
...
—	+	+	+	+
+	+	+	+	+

Ранги

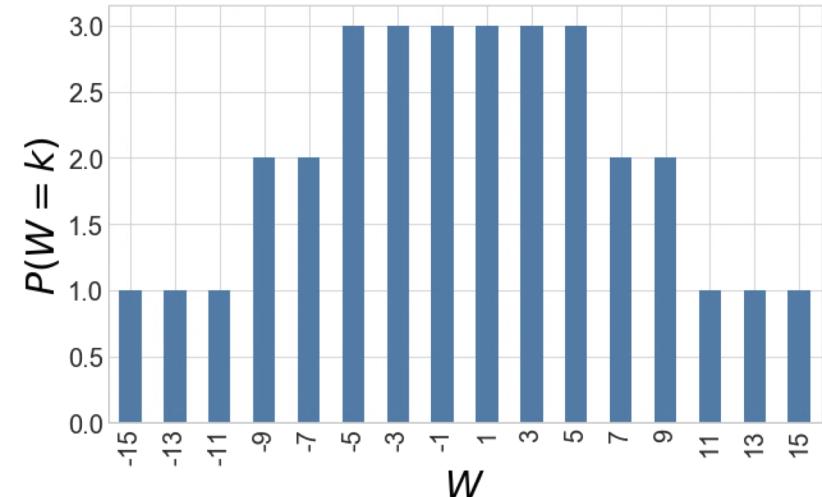
Знаки

Распределение статистики

$$W = \sum_{i=1}^n rank(|X_i - m_0|) \cdot sign(X_i - m_0)$$

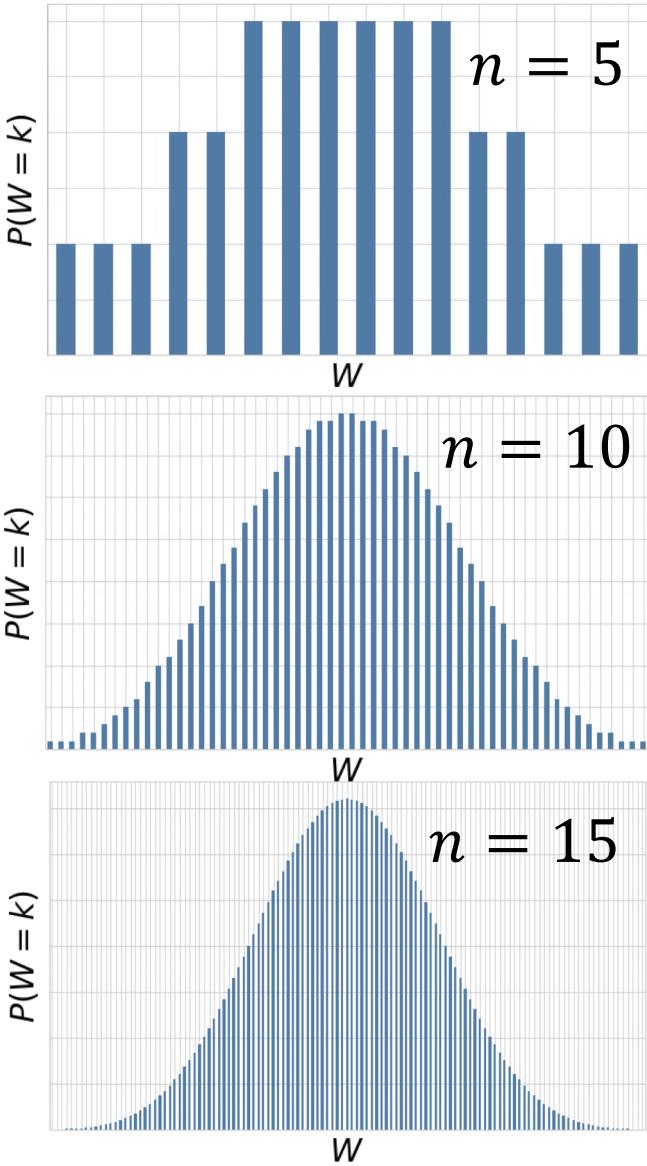
Пусть в выборке пять наблюдений:

1	2	3	4	5	
—	—	—	—	—	-15
+	—	—	—	—	-13
—	+	—	—	—	-11
...
—	+	+	+	+	13
+	+	+	+	+	15



Всего вариантов: 2^n

Апроксимация



При больших n пользуются нормальным приближением:

$$W \stackrel{asy}{\sim} N\left(0, \frac{n \cdot (n + 1) \cdot (2n + 1)}{6}\right)$$

Задача о попугаях

Пример: есть данные о говорливости попугаев (слов в день), правда ли, что в среднем попугай говорит больше 5 слов?

X_i	6	4	4	7	8	$m_0 = 5$
$ X_i - m_0 $	1	1	1	2	3	
$rank(X_i - m_0)$	2	2	2	4	5	
$sign(X_i - m_0)$	+	-	-	+	+	

$$\Rightarrow W_{\text{набл.}} = 7 \quad W_{\text{кр.}} = W_{1-\alpha} = 12$$



Гипотеза **не**
отвергается

$$W = \sum_{i=1}^n rank(|X_i - m_0|) \cdot sign(X_i - m_0)$$
$$\alpha = 0.05$$

Критерии Уилкоксона (двуихвыборочный)

$X_1, \dots, X_n \sim iid$

$Y_1, \dots, Y_n \sim iid$

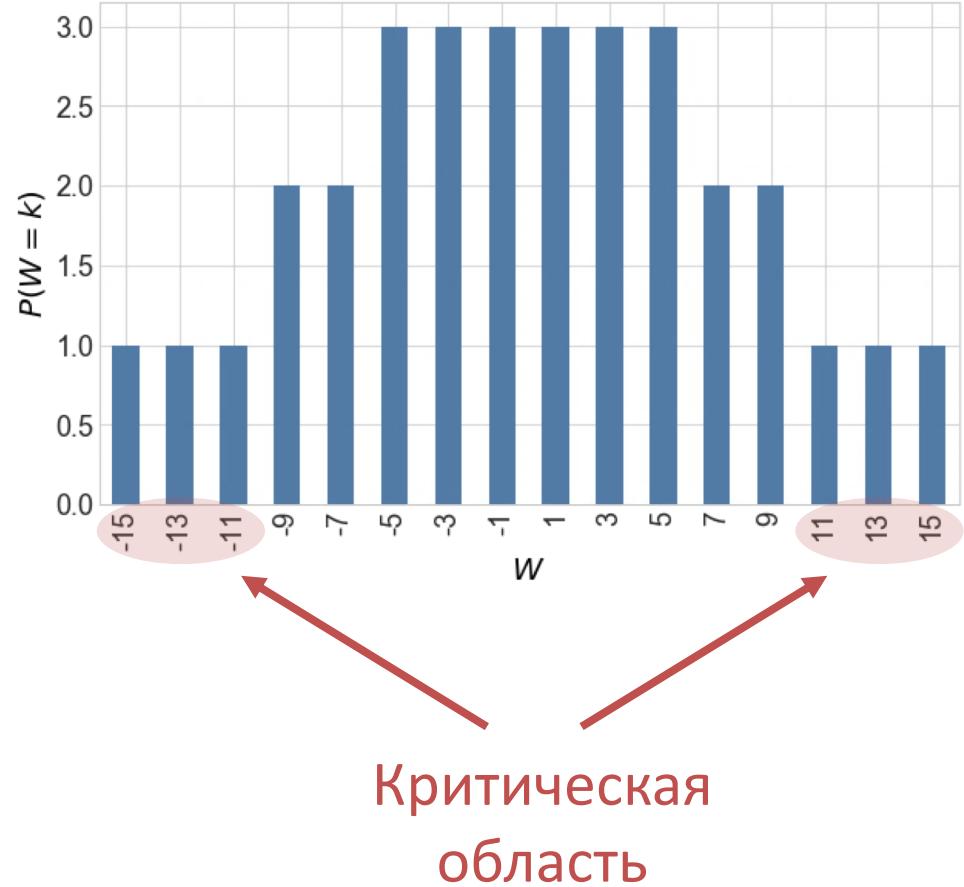
Выборки связанные

$H_0: Med(X - Y) = 0$

$H_a: Med(X - Y) \neq 0$

Критерий для проверки:

$$W = \sum_{i=1}^n rank(|X_i - Y_i|) \cdot sign(X_i - Y_i)$$



Критическая
область

Задача про апелляцию

Пример: даны баллы студентов до и после апелляции.
Правда ли, что в среднем апелляция не повышает балл
за контрольную?

До: 48, 54, 67, 56, 55, 55, 90, 71, 72, 69

После: 47, 52, 60, 60 58, 60, 70, 81, 87, 60

$ X_i - Y_i $	1	2	7	4	3	5	20	10	15	9
$rank(X_i - Y_i)$	1	2	6	4	3	5	10	8	9	7
$sign(X_i - Y_i)$	+	+	+	-	-	-	+	-	-	+

$$\Rightarrow W_{\text{набл.}} = -3 \quad W_{\text{кр.}} = W_{1-\alpha} = 33$$



Гипотеза
не отвергается

$$W = \sum_{i=1}^n rank(|X_i - Y_i|) \cdot sign(X_i - Y_i)$$

Критерии Манна-Уитни (двуихвыборочный)

$$X_1, \dots, X_{n_x} \sim iid$$

$$Y_1, \dots, Y_{n_y} \sim iid$$

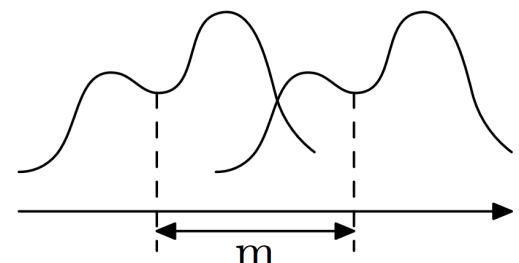
Распределения одинаковые по форме, различаются сдвигом

Выборки независимые

$$n_x \leq n_y$$

$$H_0: f_X(x) = f_Y(y)$$

$$H_a: f_X(x) = f_Y(y + m), m \neq 0$$



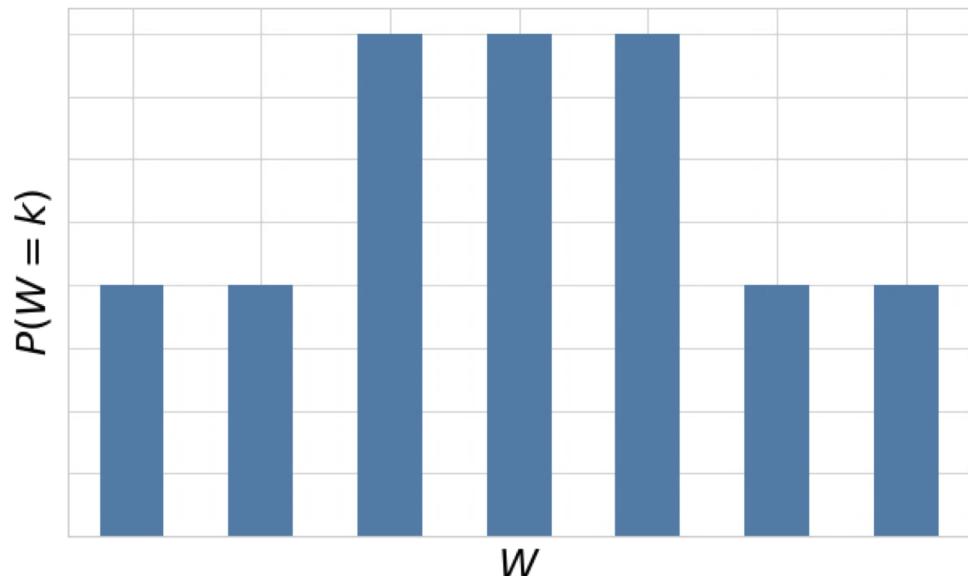
Объединим обе выборки в одну общую и посчитаем для всех чисел ранги

Критерий для проверки:

$$W = \sum_{i=1}^{n_x} rank(X_i)$$

Критерии Манна-Уитни (двуихвыборочный)

$rank(X)$	$rank(Y)$
$\{1, 2\}$	$\{3, 4, 5\}$
$\{1, 3\}$	$\{2, 4, 5\}$
$\{1, 4\}$	$\{2, 3, 5\}$
$\{1, 5\}$	$\{2, 3, 4\}$
$\{2, 3\}$	$\{1, 4, 5\}$
...	...



Всего $C_{n_x+n_y}^{n_x}$ вариантов

- Распределение статистики снова оказывается табличным
- Для больших объёмов выборки можно использовать нормальное приближение

Задача про кофе

Пример: две группы людей принимают препарат перед забегом на сто метров

Кофеин: 12, 10

Плацебо: 13, 11, 14

12, 10, 13, 11, 14

5 1 3 2 4

$$W_{\text{набл.}} = 5 + 1 = 6 \quad W_{\text{кр.}} = W_{1-\alpha} = 8$$

$$W = \sum_{i=1}^n rank(|X_i - m_0|) \cdot sign(X_i - m_0)$$

Задача про кофе

Пример: две группы людей принимают препарат перед забегом на сто метров

Кофеин: 12, 10

Плацебо: 13, 11, 14

12, 10, 13, 11, 14

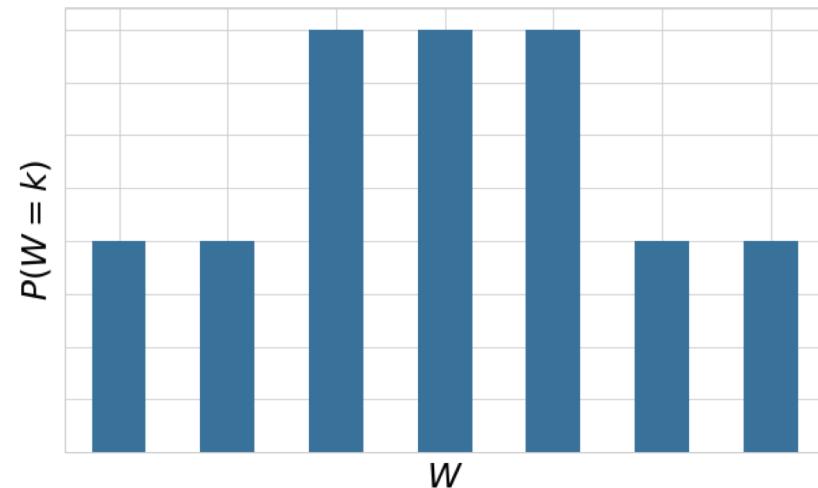
5 1 3 2 4

$$W_{\text{набл.}} = 5 + 1 = 6$$

$$W_{\text{кр.}} = W_{1-\alpha} = 8$$



Гипотеза,
что влияния
нет **не отвергается**



Резюме

- Ранговые критерии превращают выборку в ранги наблюдений и позволяют сохранить больше информации
- Чтобы сохранить информацию, приходится делать дополнительные предположения
- Предположений о законе распределения, по-прежнему, не требуется

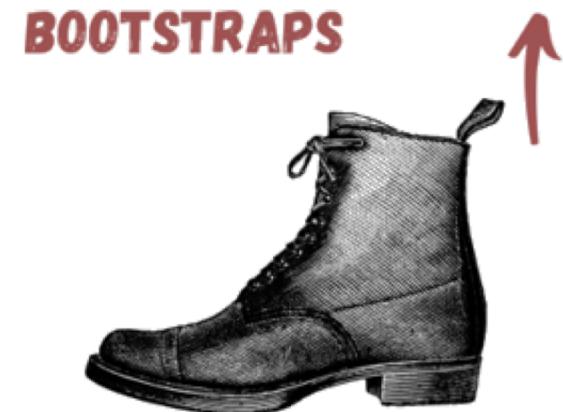
Бутстрап

Бутстррап

- Не для всех описательных статистик можно найти распределение в аналитическом виде (медиана, эксцесс, куртосис)
- Бутстррап помогает решить эту проблему

Бутстрэп

- “bootstrap” – кожаные петельки на задниках Ботинок
- “lift himself by his bootstraps” - “вытащить себя из болота за задники ботинок”, “выбиться в люди благодаря собственным усилиям”



- Предложен в 1977 году
~~бароном Мюнхгаузеном~~
Б. Эфроном из Стенфорда

Бутстррап

- **Идея метода:** имеющаяся выборка – это единственная информация об истинном распределении данных
- Давайте приблизим истинное распределение эмпирическим, то есть “сами себя вытащим”
- Предполагается, что бутстррап-распределение окажется похожим на реальное распределение

Пример

- У Саши есть выборка из двух наблюдений: 1 и 4
- Нужно построить по этой выборке бутстррап-распределение для статистики \bar{x}

Выборки с повторениями:

$$1,4 \Rightarrow \bar{x} = 2.5$$

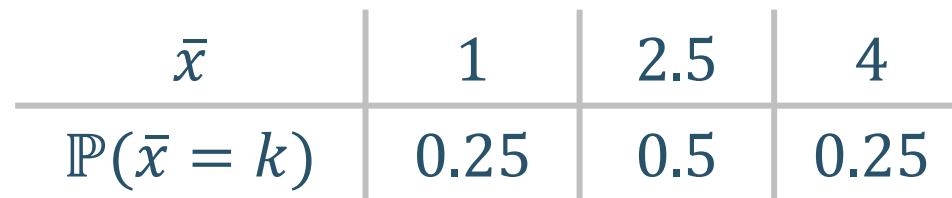
$$4,1 \Rightarrow \bar{x} = 2.5$$

$$1,1 \Rightarrow \bar{x} = 1$$

$$4,4 \Rightarrow \bar{x} = 4$$

Всего вариантов выборок:

$$n^n = 2^2 = 4$$



Пример

- У Саши есть выборка из двух наблюдений: 1 и 4
- Нужно построить по этой выборке бутстррап-распределение для статистики \bar{x}

Всего вариантов выборок:

$$n^n = 2^2 = 4$$

- Строить такие распределения для больших выборок дорого
- Задача слишком сложная, а точность излишняя
- **Выход:** симуляции

Схема бутстрата

- Извлечение выборок из генеральной совокупности – сэмплирование из неизвестного распределения $F(x)$
- У нас есть оценка для $F(x) - \hat{F}_n(x)$
- Сэмплировать из такого распределения – то же самое, что брать выборки с повторениями объёма n

Схема бутстрата

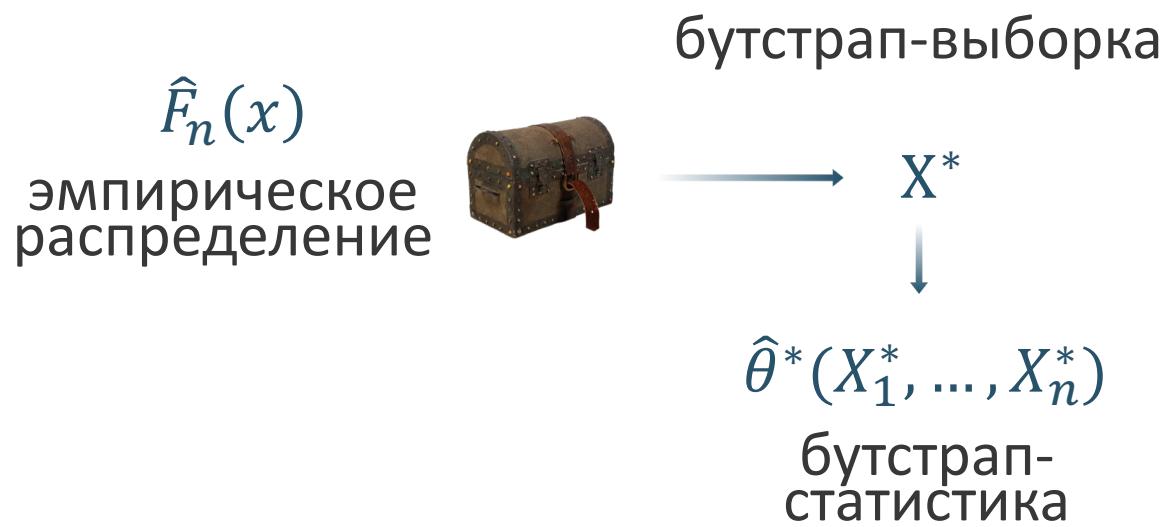


Схема бутстрата

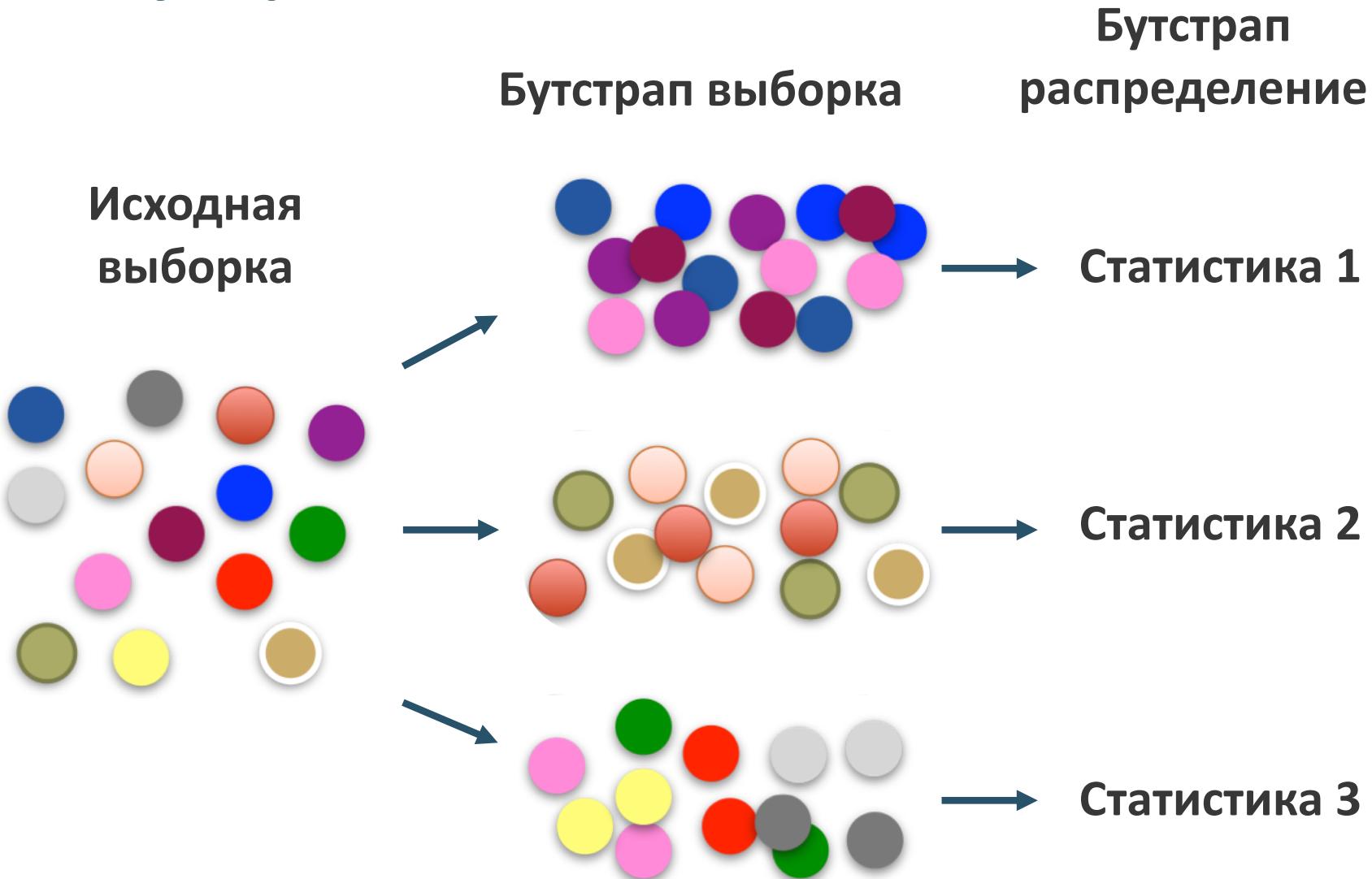


Схема бутстрата

- Генерируем B “псевдовыборок” с повторениями объёма n и оцениваем настоящее распределение “псевдоэмпирическим”
- По каждой выборке вычисляем интересующую нас статистику, получаем для неё бутстрат-распределение
- **Эфронов доверительный интервал:** находим выборочные квантили бутстрат-распределения
- В таком доверительном интервале возникает смещение \Rightarrow более сложные техники бутстрата

Доверительный интервал Эфона

Бутстранируем: оценку неизвестного параметра

Сэмплируем: x_1^*, \dots, x_n^*

Считаем: $\hat{\theta}^*$

Повторяем: B раз

Строим распределение: $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$

Интервал: $[\hat{\theta}_{\frac{\alpha}{2}}^*; \hat{\theta}_{1-\frac{\alpha}{2}}^*]$

- ! Если распределение выборки несимметрично, такой доверительный интервал усиливает смещение, присущее изначальной выборке

Бутстрап

Центрирование

- Чтобы не создавать искусственное смещение между $\hat{\theta}$ и θ , нужно бутстрартировать центрированную статистику
- Хотим бутстрартировать разность $\hat{\theta} - \theta$
- Бутстрарпируя $\hat{\theta}$, мы подсчитываем её каждый раз заново, но на бутстрарповских выборках, $\hat{\theta}^*$
- Бутстрарповским аналогом для θ будет $\hat{\theta}$, так как мы сэмплируем данные их эмпирического распределения
- Бутстрарповский аналог для $\hat{\theta} - \theta$ – это $\hat{\theta}^* - \hat{\theta}$

Доверительный интервал Холла

Бутстранируем: Отклонение оценки от истинного значения

Сэмплируем: x_1^*, \dots, x_n^*

Считаем: $\hat{q}_i^* = \hat{\theta}^* - \hat{\theta}$

Повторяем: B раз

Строим распределение: $\hat{q}_1^*, \dots, \hat{q}_B^*$

Интервал: $[\hat{\theta} - \hat{q}_{1-\frac{\alpha}{2}}^*; \hat{\theta} - \hat{q}_{\frac{\alpha}{2}}^*]$

t -процентильный доверительный интервал

Бутстранируем: t - статистику

Сэмплируем: x_1^*, \dots, x_n^*

Считаем: $t_i^* = \frac{(\hat{\theta}^* - \hat{\theta})}{se(\hat{\theta}^*)}$

Повторяем: B раз

Строим распределение: t_1^*, \dots, t_B^*

Интервал: $[\hat{\theta} - t_{1-\frac{\alpha}{2}}^* \cdot se(\hat{\theta}); \hat{\theta} + t_{\frac{\alpha}{2}}^* \cdot se(\hat{\theta})]$

❗ Если бутстрарировать $|\hat{\theta}^* - \hat{\theta}|$, можно получить симметричный интервал

Проверка гипотез при помощи бутстрата

$$H_0: \theta = \theta_0$$

$$H_a: \theta > \theta_0$$

t-статистика:

$$t = \frac{(\hat{\theta} - \theta_0)}{se(\hat{\theta})}$$

Бутстррап-аналог:

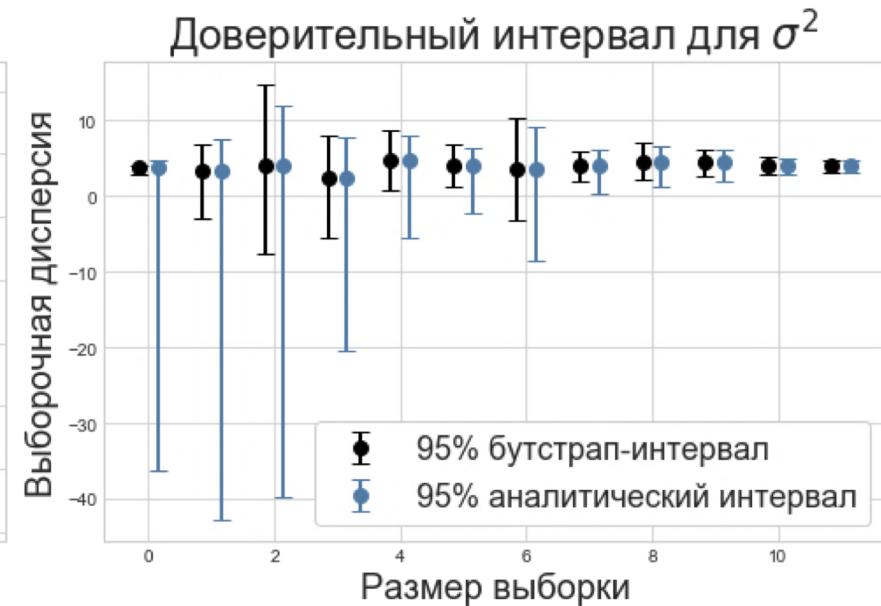
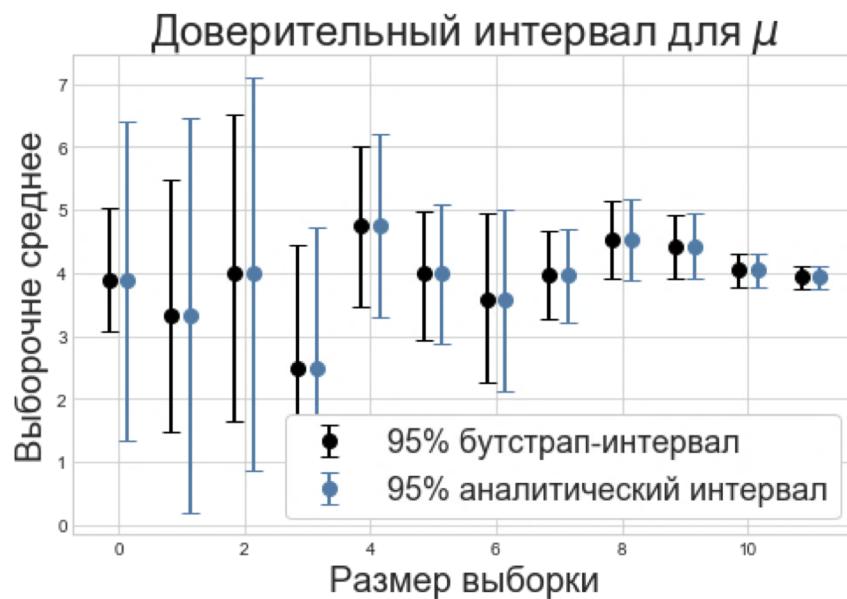
$$t^* = \frac{(\hat{\theta}^* - \hat{\theta})}{se(\hat{\theta}^*)}$$

- Гипотеза отвергается, если $t_{obs} > t_{1-\alpha}^*$
- По аналогии можно проверять гипотезы против других альтернатив
- Для более сложных гипотез есть специальные бутстрраповские алгоритмы проверки

► <http://quantile.ru/03/03-SA.pdf>

Проблемы бутстрата

- Чтобы бутстрат сработал, выборка должна быть репрезентативной



- Если исходная выборка маленькая, бутстратовский доверительный интервал будет уже аналитического, так как в выборке недостаточно “неопределенности”

Проблемы бутстрата

- Если в данных есть структура (регрессия, временные ряды), бутстррап нужно устроить так, чтобы учитывать её ⇒ разные виды бутстрата
- Бутстррап ненадёжно работает в хвостах распределения из-за маленького числа наблюдений: мы можем хорошо оценить медиану, но не 99% квантиль
- Если у распределения тяжёлые хвосты, бутстррап может работать некорректно и в средиземье

Сколько процентов выборки берем

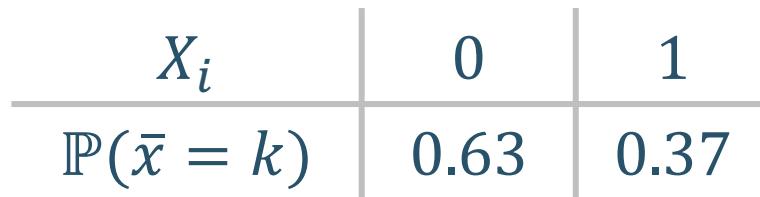
У Винни-Пуха есть 100 песенок (кричалок, вопелок, пыхтелок и сопелок). Каждый день он поёт одну равновероятно наугад. Одну и ту же песенку он может петь несколько раз. Сколько в среднем песенок не будут спеты ни разу за 100 дней?

- Вероятность конкретной песенки $\frac{1}{n}$
- В конкретный день не споёт эту песенку с вероятностью $1 - \frac{1}{n}$
- Не споёт песенку ни разу с вероятностью

$$\left(1 - \frac{1}{n}\right)^n \rightarrow e^{-1} = \frac{1}{e} \approx 0.37 \text{ при } n \rightarrow \infty$$

Пример

- Случайная величина $X_i = 1$, если песенка не была ни разу спета за n дней



- Случайная величина $Y = X_1 + \dots + X_n$ – число песенок, которое Винни-Пух ни разу не споёт за n дней

$$\mathbb{E}(Y) = \mathbb{E}(X_1 + \dots + X_n) = n \cdot \mathbb{E}(X_1) = 0.37 \cdot n$$

! В одной бутстрраповской выборке будет в среднем оказываться 63% наблюдений

Резюме

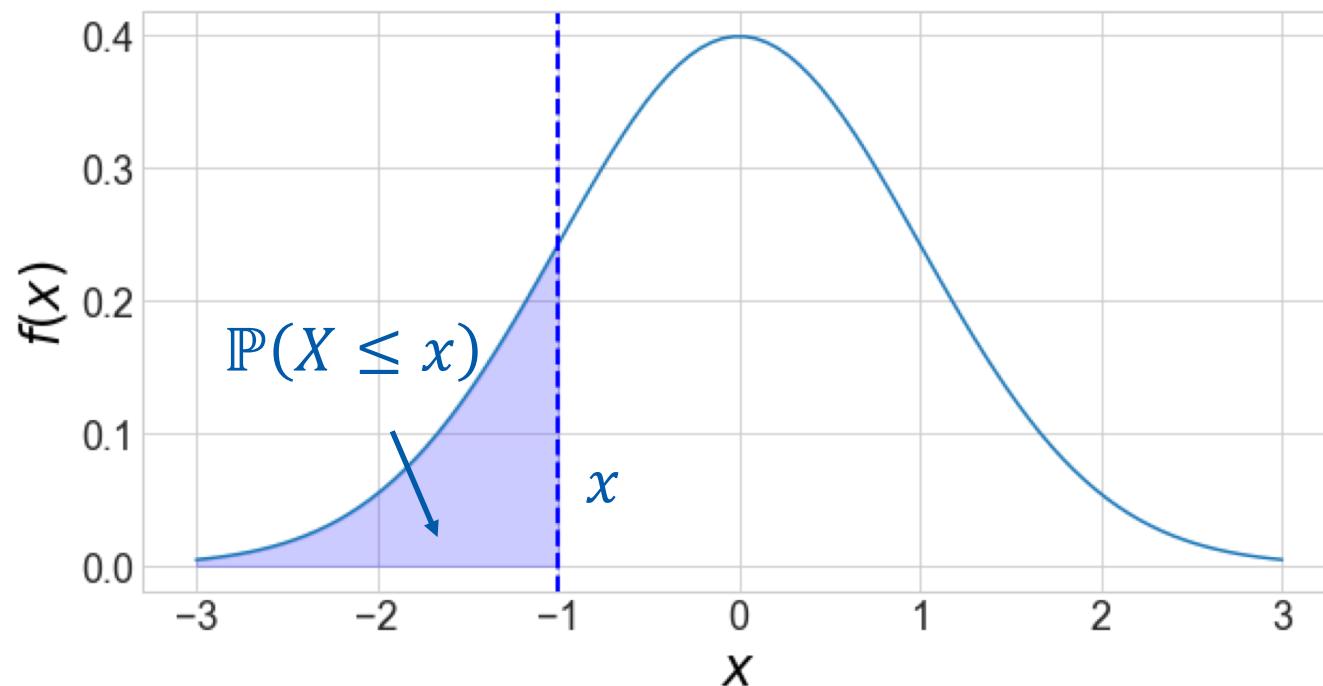
- Бутстреп – это метод получения критических значений статистики
- Процедура может требовать много времени для оценки
- При некоторых ограничениях бутстреп даёт состоятельные оценки, но не в общем случае
- Плохо работает для статистик, значение которых зависит от небольшого числа элементов выборки

Эмпирическая функция распределения

Функция распределения

Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

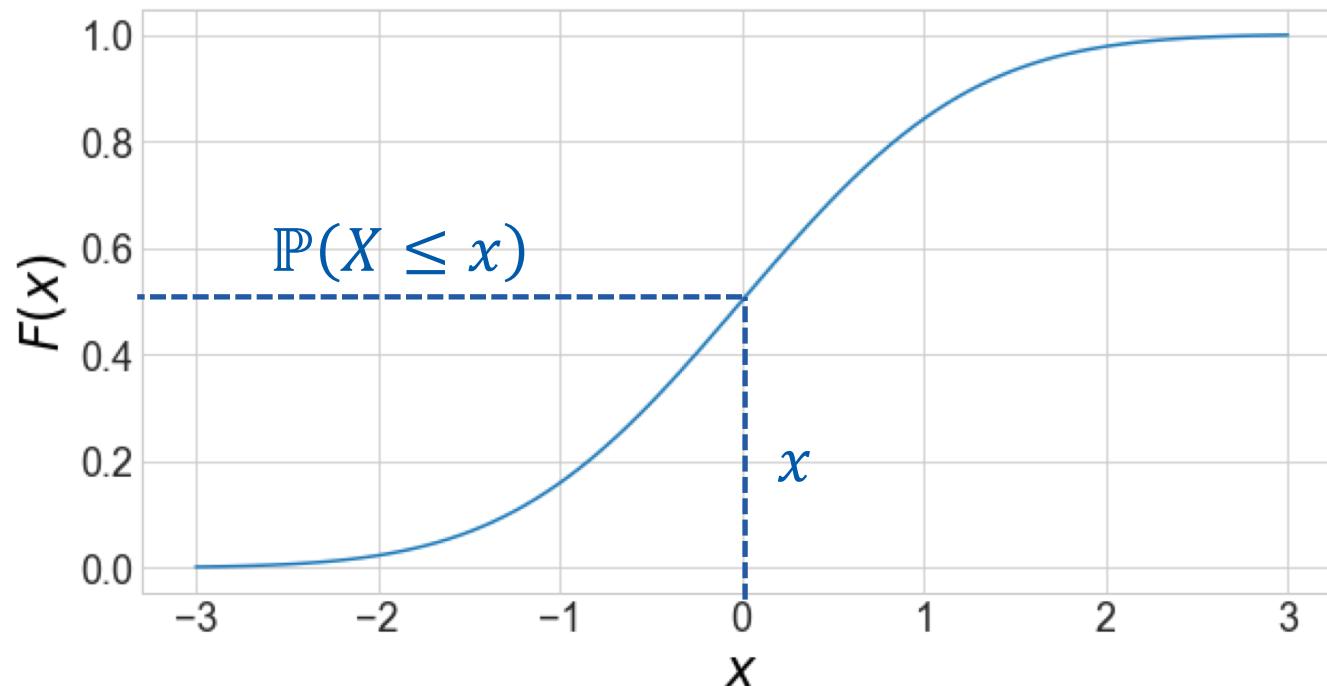
$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) \, dt, f(t) \text{ – плотность}$$



Функция распределения

Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) \, dt, f(t) \text{ – плотность}$$



Эмпирическая функция распределения

Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

$$F(x) = \mathbb{P}(X \leq x)$$

Эмпирическая функция распределения – функция, которая определяет для каждого x частоту события $X \leq x$, то есть

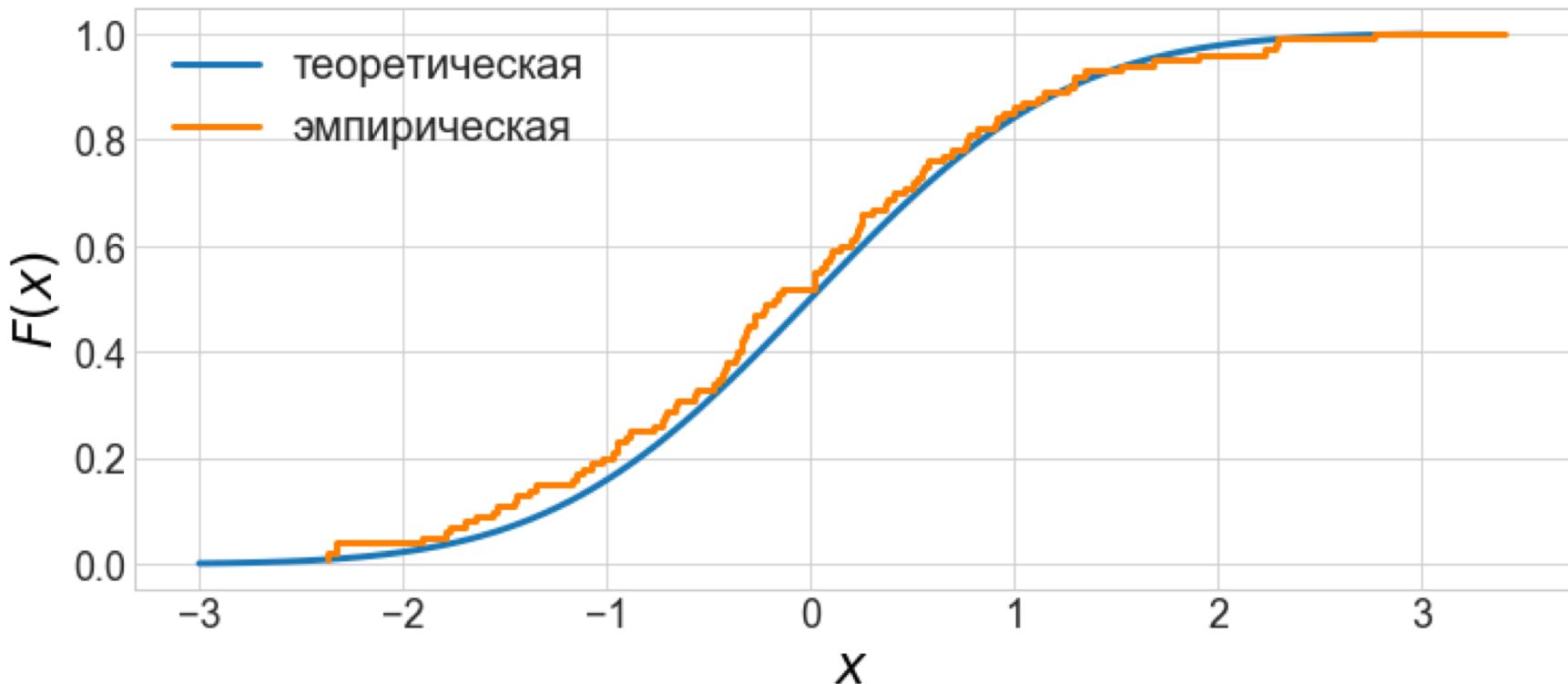
$$\hat{F}_n(x) = \widehat{\mathbb{P}}(X \leq x) = \frac{1}{n} \sum_{i=1}^n [X_i \leq x],$$

где $[]$ – индикаторная функция, то есть:

$$[X_i \leq x] = \begin{cases} 1, & X_i \leq x \\ 0, & \text{иначе} \end{cases}$$

Эмпирическая функция распределения

Чем больше выборка, тем чаще ступеньки и тем больше эмпирическая функция распределения похожа на теоретическую.



Эмпирическая функция распределения

Эмпирическая функция распределения – статистика, оцениваемая по данным



Какими свойствами
она обладает?

Свойства эмпирической функции распределения

1. Несмешённость: $\mathbb{E}(\hat{F}_n(x)) = F_X(x)$

Эмпирическая функция распределения

$$\mathbb{E}(\hat{F}_n(x)) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n [X_i \leq x]\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}([X_i \leq x]) =$$

$$\frac{1}{n} \sum_{i=1}^n (1 \cdot \mathbb{P}(X_i \leq x) + 0 \cdot \mathbb{P}(X_i > x)) = \frac{1}{n} \sum_{i=1}^n \mathbb{P}(X_i \leq x) =$$

$$\frac{1}{n} \sum_{i=1}^n F_X(x) = \frac{1}{n} \cdot n \cdot F_X(x) = F_X(x)$$

Свойства эмпирической функции распределения

1. Несмешённость: $\mathbb{E}(\hat{F}_n(x)) = F_X(x)$
2. Состоятельность: $\operatorname{plim}_{n \rightarrow \infty} \hat{F}_n(x) = F_X(x)$

Эмпирическая функция распределения

$$\hat{F}_n(x) = \widehat{\mathbb{P}}(X \leq x) = \frac{1}{n} \sum_{i=1}^n [X_i \leq x] =$$
$$= \frac{[X_1 \leq x] + \cdots + [X_n \leq x]}{n} \xrightarrow{p} F_X(x) \text{ при } n \rightarrow \infty$$

По ЗБЧ наша оценка состоятельна

Свойства эмпирической функции распределения

1. Несмешённость: $\mathbb{E}(\hat{F}_n(x)) = F_X(x)$

2. Состоятельность: $\operatorname{plim}_{n \rightarrow \infty} \hat{F}_n(x) = F_X(x)$

3. Асимптотическая нормальность:

$$\hat{F}_n(x) \stackrel{asy}{\sim} N\left(F_X(x), \frac{F_X(x) \cdot (1 - F_X(x))}{n}\right)$$

Свойства эмпирической функции распределения

1. Несмешённость: $\mathbb{E}(\hat{F}_n(x)) = F_X(x)$

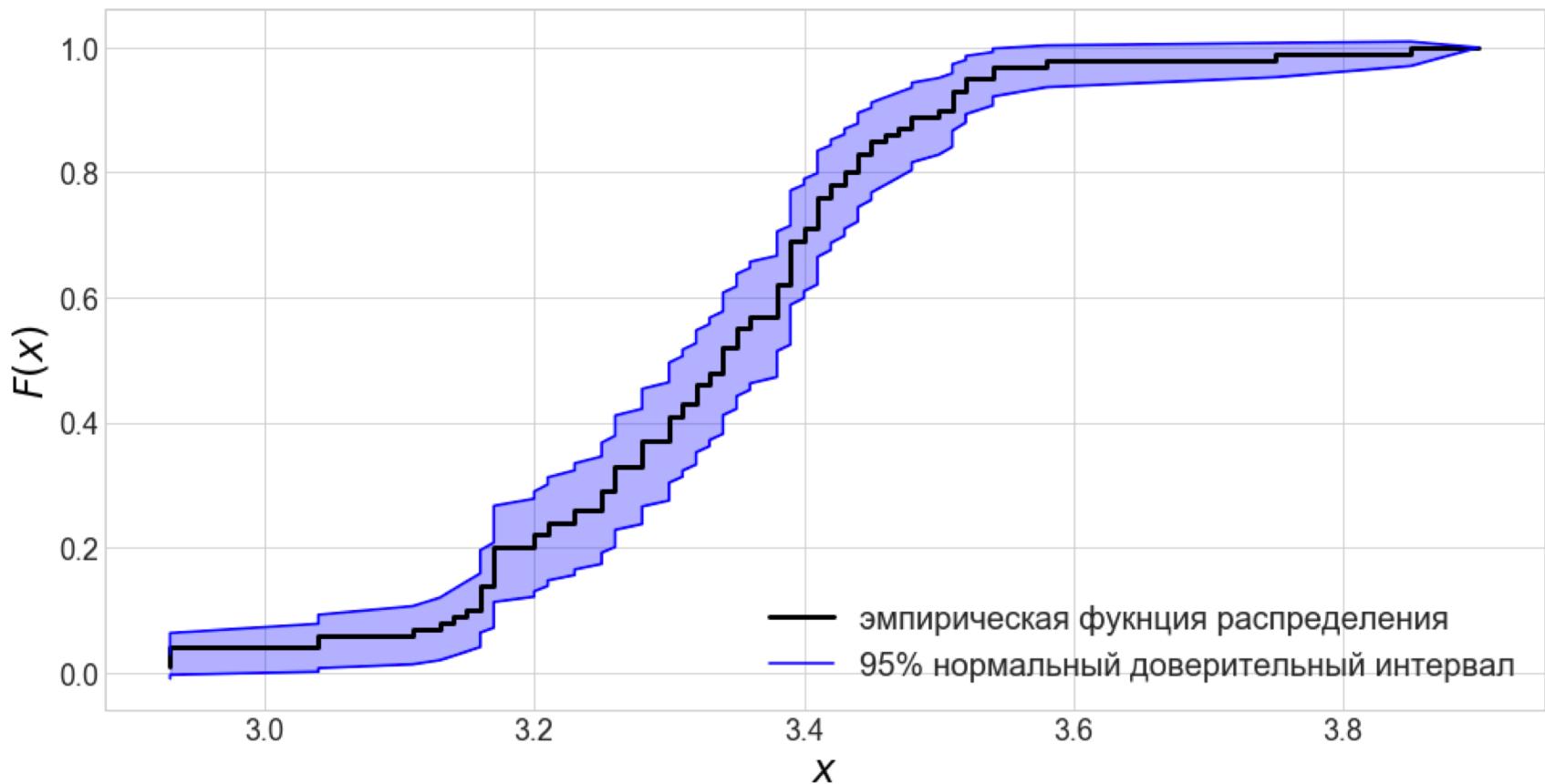
2. Состоятельность: $\operatorname{plim}_{n \rightarrow \infty} \hat{F}_n(x) = F_X(x)$

3. Асимптотическая нормальность:

$$\hat{F}_n(x) \stackrel{asy}{\sim} N\left(F_X(x), \frac{\hat{F}_n(x) \cdot (1 - \hat{F}_n(x))}{n}\right)$$

Нормальный доверительный интервал

ЦПТ помогает построить доверительный интервал для эмпирической функции распределения в каждой точке



Критерий Колмогорова

Критерии согласия

- **Критерии согласия** – критерий о виде неизвестного закона распределения

$$H_0: X \sim F_X(x)$$

H_a : гипотеза H_0 неверна

- Распределение случайной величины описывается её функцией распределения

$$H_0: F_X(x) = F_0(x)$$

$$H_a: F_X(x) \neq F_0(x)$$

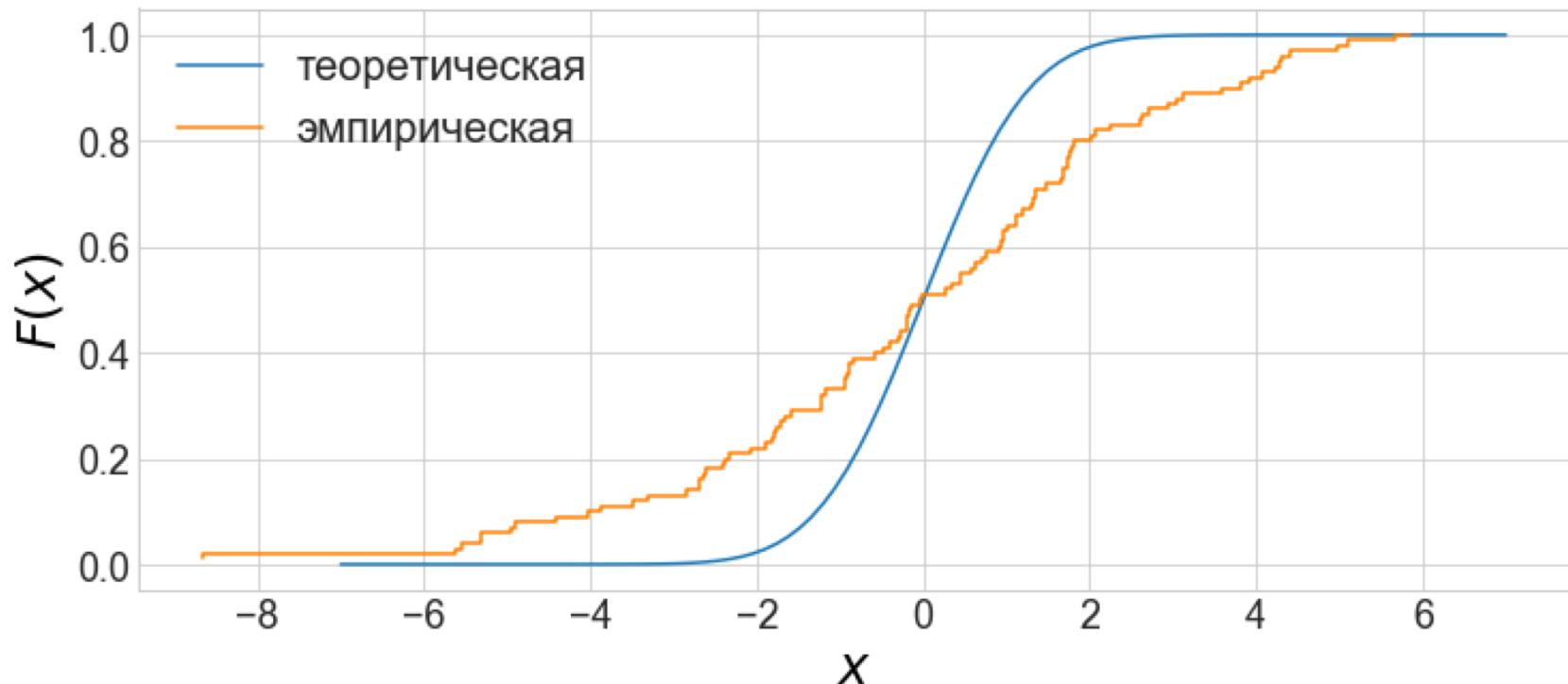
- Параметры неизвестного закона распределения мы фиксируем как нулевую гипотезу

Критерии согласия

- Критерии о неизвестных законах распределения строятся по аналогии с параметрическими
- У нас в распоряжении есть выборка, по ней мы оценили эмпирическую функцию распределения $\hat{F}_n(x)$
- Надо понять насколько она отличается от $F_0(x)$, то есть надо посчитать расстояние между функциями и узнать его распределение

Расстояние между функциями

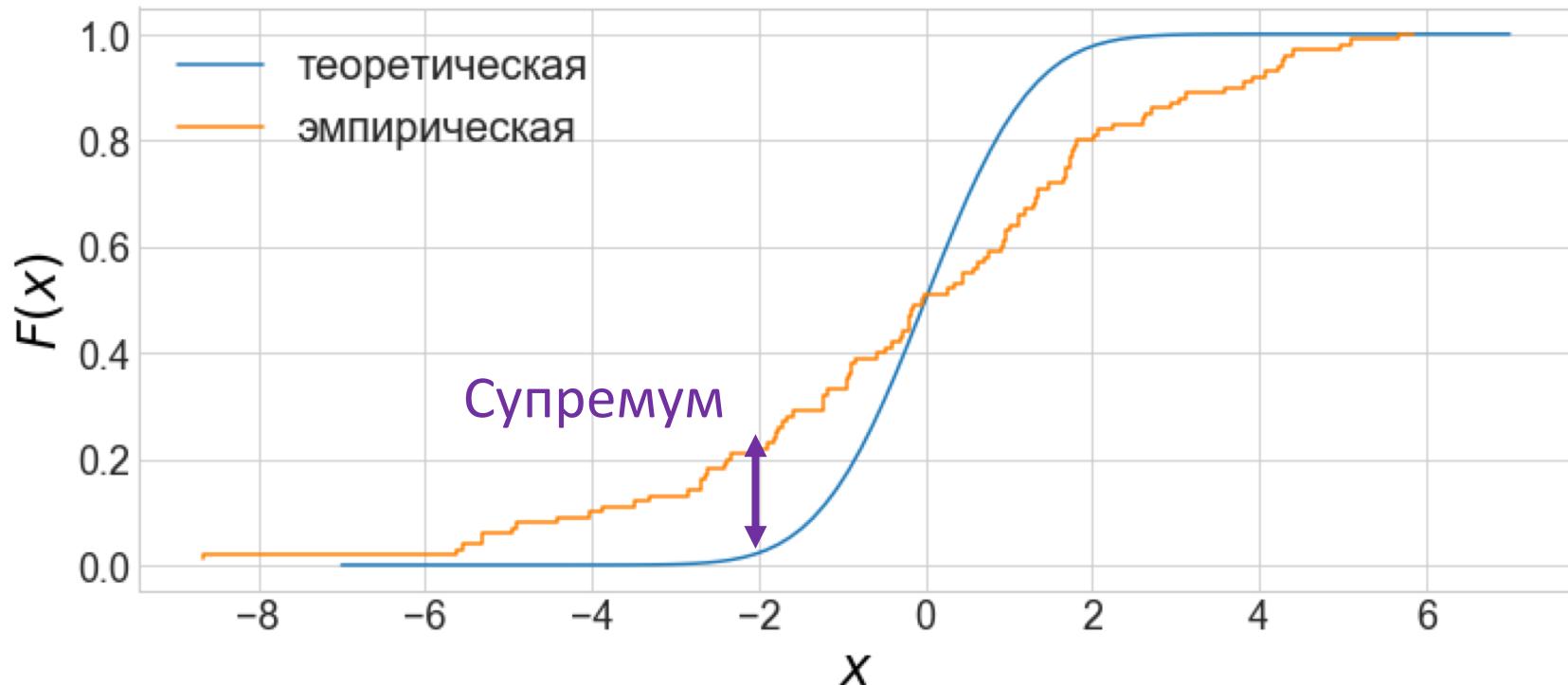
- Как измерить расстояние между двумя функциями?
- Разные способы сделать это \Rightarrow разные критерии



Критерий Колмогорова

- Можно найти между ними самое большое расстояние:

$$D_n(X_1, \dots, X_n) = \sup_x |F_0(x) - \hat{F}_n(x)|$$



Теорема Колмогорова

Статистика Колмогорова:

$$D_n(X_1, \dots, X_n) = \sup_x |F_0(x) - \hat{F}_n(x)|$$

При справедливости нулевой гипотезы, распределение статистики D_n одинаково для любых **непрерывных** распределений, при этом его функция распределения:

$$\lim_{n \rightarrow \infty} \mathbb{P}(\sqrt{n} \cdot D_n \leq z) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 z^2}$$



Распределение Колмогорова –
наш новый союзник, на её основе
мы можем построить критерий

Критерий Колмогорова

$X_1, \dots, X_n \sim iid F_X(x)$

$H_0: F_X(x) = F_0(x)$

$H_a: F_X(x) \neq F_0(x)$

Критерий для проверки:

$$K_n = \sqrt{n} \cdot \sup_x |F_0(x) - \hat{F}_n(x)|$$

$K_n \underset{H_0}{\overset{asy}{\sim}}$ распределение Колмогорова



$K_{\text{набл.}} > K_{1-\alpha}$



Критерий применяется только для непрерывных распределений



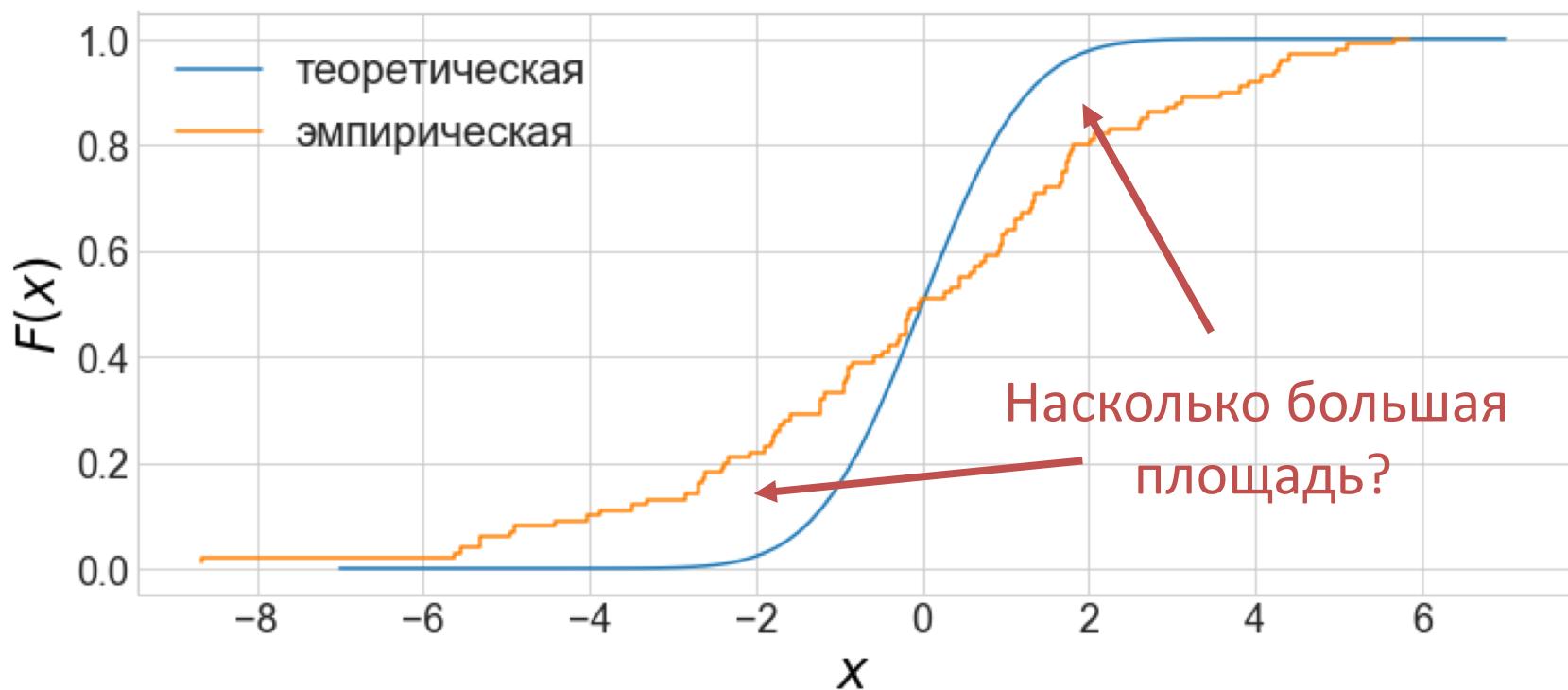
$K_{\text{набл.}} \leq K_{1-\alpha}$

- Критические значения рассчитываются по таблицам, составленным для распределения Колмогорова

Другие способы искать расстояние

$$\int_{-\infty}^{+\infty} \psi(F_0(x)) \cdot [F_0(x) - \hat{F}_n(x)]^2 \cdot f_0(x) dx$$

функция - вес
усредняем по распределению
чем больше разница,
тем выше штраф



Критерий Крамера-Мизеса

$$\int_{-\infty}^{+\infty} 1 \cdot [F_0(x) - \hat{F}_n(x)]^2 \cdot f_0(x) dx$$

- Чем больше разница между функциями, тем выше штраф
- Разница усредняется по всем возможным значениям случайной величины
- Хорошо улавливает разницу в области “типовых значений” случайной величины

Критерий Андерсона-Дарлинга

$$\int_{-\infty}^{+\infty} \frac{1}{F_0(x)(1 - F_0(x))} \cdot [F_0(x) - \hat{F}_n(x)]^2 \cdot f_0(x) dx$$

- При $F_0(x) \rightarrow 1$ или $F_0(x) \rightarrow 0$ дополнительный множитель оказывается очень большим
- Получается, что основное внимание уделяется разнице “на хвостах” распределения
- Для подобных критериев также известны предельные распределения, которыми можно пользоваться для проверки гипотез

Гипотезы об однородности выборок

$$X_1, \dots, X_{n_x} \sim iid F_X(x)$$

$$Y_1, \dots, Y_{n_y} \sim iid F_Y(x)$$

- Гипотезы о том, что выборки сделаны из одного и того же распределения

$$H_0: F_X(x) = F_Y(x)$$

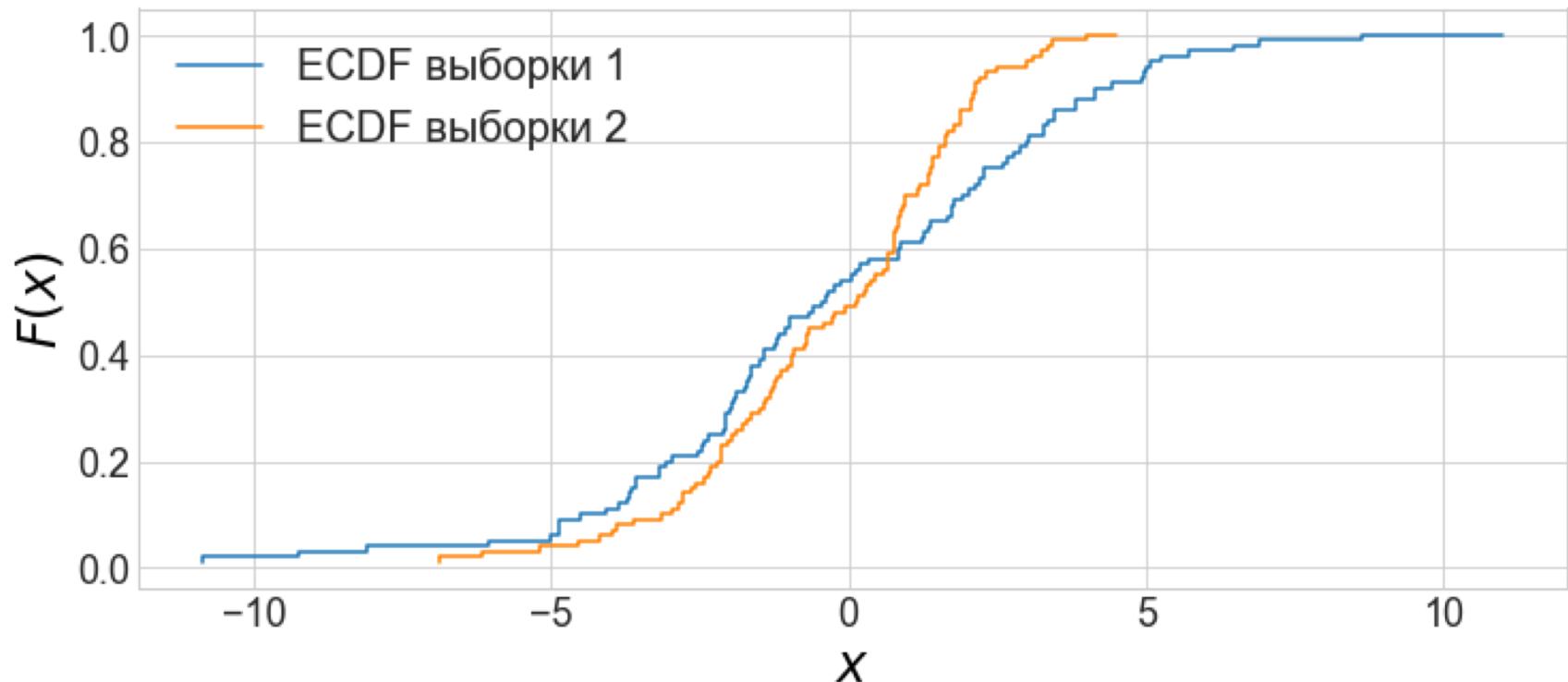
$$H_a: F_X(x) \neq F_Y(x)$$

- Для проверки этой гипотезы можно использовать ту же самую идею, но расстояние надо будет считать между эмпирическими функциями распределения

Критерий Колмогорова-Смирнова

Снова будем искать супремум, но уже между эмпирическими функциями

$$D_n = \sup_x | \hat{F}_X(x) - \hat{F}_Y(x) |$$



Критерий Колмогорова-Смирнова

Снова будем искать супремум, но уже между эмпирическими функциями

$$D_n = \sup_x | \hat{F}_X(x) - \hat{F}_Y(x) |$$

- Статистика

$$\sqrt{\frac{n_x \cdot n_y}{n_x + n_y}} \cdot D_n$$

при $n_x, n_y \rightarrow \infty$ имеет распределение Колмогорова

- Она используется для проверки гипотезы об однородности выборок

Резюме

- Чтобы проверять гипотезы о распределениях, нужно научиться считать между ними расстояния
- Критерий Колмогорова ищет расстояние как супремум и позволяет проверять гипотезы о распределении и однородности выборок
- Критерии Крамера-Мизеса и Андерсона-Дарлинга помогают специфицировать критерий либо для "средиземья" либо для "крайнеземья" (хвостов)

Резюме

- Эти критерии работают только для непрерывных распределений
- Неизвестные параметры распределения фиксируются в нулевой гипотезе
- Если мы оцениваем параметры по выборке, распределения Колмогорова не будет одинаковым для всех распределений
- Для различных распределений можно строить уточнения теста Колмогорова

Критерий Пирсона

Критерий Пирсона

- Критерий Колмогорова работает только для непрерывных распределений
- Для проверки гипотез о дискретных распределениях используется критерий Пирсона

$$H_0: F_X(x) = F_0(x)$$

$$H_a: F_X(x) \neq F_0(x)$$

- Критерий Пирсона считает расстояние между распределениями с помощью сверки теоретических и эмпирических частот между собой

Критерий Пирсона

- Выборка X_1, \dots, X_n из дискретного распределения
- Предполагаем, что случайная величина принимает s значений с какими-то вероятностями

X	z_1	z_2	\dots	z_s	- возможные значения
$\mathbb{P}(X = z)$	$p_1(\theta)$	$p_2(\theta)$	\dots	$p_s(\theta)$	- теоретические вероятности
$\#(X_i = z)$	v_1	v_2	\dots	v_s	- эмпирические частоты

- Нужно сравнить все эмпирические частоты с теоретическими
- $\hat{\theta}$ – состоятельная оценка параметров распределения
- k – их количество

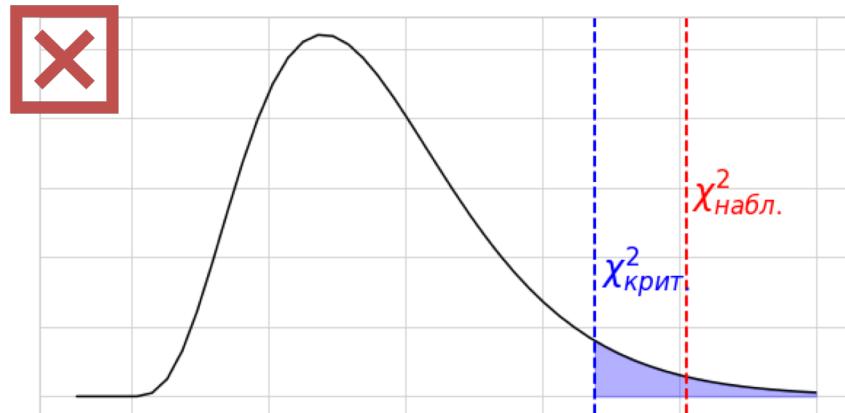
$$\sum_{j=1}^s \frac{(v_j - n \cdot p_j(\hat{\theta}))^2}{n \cdot p_j(\hat{\theta})} \stackrel{\text{asy}}{\underset{H_0}{\sim}} \chi_{s-k-1}^2$$

Критерий Пирсона

$X_1, \dots, X_n \sim iid F_X(x)$

$H_0: F_X(x) = F_0(x)$

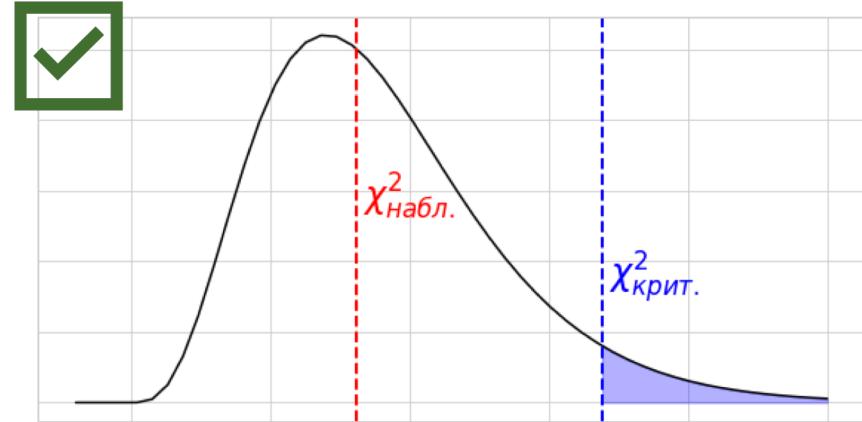
$H_a: F_X(x) \neq F_0(x)$



$$\chi^2_{\text{набл.}} > \chi^2_{s-k-1}(1 - \alpha)$$

Критерий для проверки:

$$\sum_{j=1}^s \frac{(v_j - n \cdot p_j(\hat{\theta}))^2}{n \cdot p_j(\hat{\theta})} \stackrel{asy}{\underset{H_0}{\sim}} \chi^2_{s-k-1}$$



$$\chi^2_{\text{набл.}} \leq \chi^2_{s-k-1}(1 - \alpha)$$

! Критерий обычно применяют для дискретных распределений

Пример:

- X – Количество сделок на фондовой бирже за квартал для пятисот инвесторов

вместе, т.к. их мало

X	0	1	2	3	4	5	6	7
$\mathbb{P}(X = z)$	0.37	0.37	0.18	0.06		0.02		
$\#(x_i = z)$	199	170	87	31	9	2	1	1

$$H_0: X \sim Poiss(\lambda)$$

$$\hat{\lambda} = \bar{x} \approx 1 - \text{состоятельная оценка}$$

$$\mathbb{P}(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

$$\mathbb{P}(X = 0) = \frac{e^{-1} 1^0}{0!} = 0.37$$

$$\mathbb{P}(X = 1) = \frac{e^{-1} 1^1}{1!} = 0.37$$

...

Пример:

- X – Количество сделок на фондовой бирже за квартал для пятисот инвесторов

вместе, т.к. их мало

X	0	1	2	3	4	5	6	7
$\mathbb{P}(X = z)$	0.37	0.37	0.18	0.06			0.02	
$\#(x_i = z)$	199	170	87	31	9	2	1	1

$$H_0: X \sim Poiss(\lambda) \quad \hat{\lambda} = \bar{x} \approx 1 - \text{состоятельная оценка}$$

$$\sum_{j=1}^s \frac{(v_j - n \cdot p_j(\hat{\theta}))^2}{n \cdot p_j(\hat{\theta})} = \frac{(199 - 500 \cdot 0.37)^2}{500 \cdot 0.37} + \dots = 3.85$$

$$\chi^2_{s-k-1}(1-\alpha) = \chi^2_{5-1-1}(1-0.05) = \chi^2_3(0.95) = 7.8$$



Гипотеза о распределении Пуассона **не отвергается**

Критерий Пирсона

- Критерий Пирсона не состоятелен против всех альтернатив, т.е. бывают распределения, которые он не может отличить друг от друга
- Можно попробовать использовать критерий Пирсона для непрерывных распределений
- Для этого нужно будет разбить все возможные значения непрерывной случайной величины на бины (как на гистограмме)
- Результат работы теста будет зависеть от числа выбранных бинов \Rightarrow лучше пользоваться для непрерывных распределений другими критериями

Гипотезы об однородности выборок

- Критерий Пирсона по аналогии с критерием Колмогорова можно использовать для проверки гипотез об однородности выборок

$$X_1, \dots, X_{n_x} \sim iid F_X(x) \quad Y_1, \dots, Y_{n_y} \sim iid F_Y(x)$$

Выборки независимые

$$H_0: F_X(x) = F_Y(x)$$

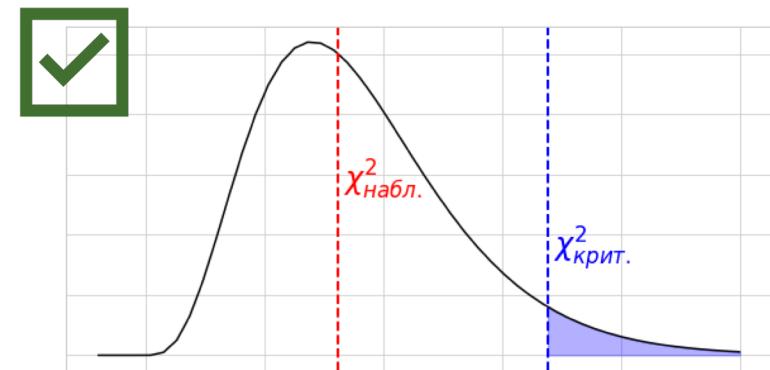
$$H_a: F_X(x) \neq F_Y(x)$$

Критерий Пирсона

- Нужно сравнивать эмпирические частоты между собой

X, Y	z_1	z_2	...	z_s
$\#(y_i = z)$	w_1	w_2	...	w_s
$\#(x_i = z)$	v_1	v_2	...	v_s

$$\sum_{j=1}^s \frac{\left(\frac{v_j}{n_x} - \frac{w_j}{n_y}\right)^2}{\frac{n_x + n_y}{n_x}} \stackrel{\text{asy}}{\underset{H_0}{\sim}} \chi^2_{s-1}$$



- Можно использовать критерий для непрерывных распределений, но тогда придется дробить данные на бины

Резюме

- Критерий Пирсона позволяет проверять гипотезы о виде распределения для дискретных случайных величин
- Его можно использовать и для непрерывных величин, но результат будет зависеть от числа выбранных групп