# Введение в математическую статистику. Доверительные интервалы. Бутстрэп.

Леонид Иосипой

Программа «Математика для анализа данных»
Центр непрерывного образования, ВШЭ

16 июня 2021

- Повторение
- Доверительные интервалы
- Распределения, связанные с нормальным
- Доверительные интервалы в нормальной модели
- Бутстрэп

# Повторение

1. Оценивание параметров распределения.
   Метод максимального правдоподобия.

   Допустим, что у нас есть реализация выборки из некоторого распределения, известного с точностью до одного или нескольких параметров.

   **Метод максимального правдоподобия:** чтобы оценить неизвестные параметры модели, нам необходимо найти максимум функции правдоподобия (то есть найти частные производные по всем параметрам и приравнять их к нулю).

# Повторение

1. Оценивание параметров распределения.
   Метод максимального правдоподобия.

   Пусть дана реализация выборки $x_1, \dots, x_n$ из некоторого распределения с неизвестным (многомерным) параметром

   $$\theta \in \Theta \subset \mathbb{R}^d.$$

   Введем величину:

   $$p(u, \theta) = \begin{cases} \mathbb{P}_\theta(X = u) & \text{в дискретном случае,} \\ f_\theta(u) & \text{в непрерывном случае.} \end{cases}$$

   Функцией правдоподобия называется величина:

   $$L(\theta) = p(x_1, \theta) \cdot \ldots \cdot p(x_n, \theta).$$

# Повторение

1. Оценивание параметров распределения.
   Метод максимального правдоподобия.

   В общем случае $L(\theta)$ характеризует вероятность получить реализацию $x_1, \ldots, x_n$ выборки при заданном $\theta$.

   Представляется разумным в качестве оценки параметра $\theta$ взять наиболее правдоподобное значение, которое получается при максимизации функции $L(\theta)$.

# Повторение

1. Оценивание параметров распределения.
   Метод максимального правдоподобия.

   При некоторых условиях на регулярность модели оценки максимального правдоподобия являются:

   - Возможно смещёнными
   - Состоятельными
   - Асимптотически эффективными

     Это означает, что дисперсия при $n \to \infty$ является наименьшей возможной среди многих других оценок.

# Повторение

2. Оценивание характеристик распределения.
   Метод Монте-Карло.

   В большинстве случаев характеристика распределения величины $X$ может быть записана как $\mathbb{E}[g(X)]$, где $g : \mathbb{R} \to \mathbb{R}$ — некоторая (известная) функция.

   Если функция $g$ не зависит от других характеристик $X$, то оценить $\mathbb{E}[g(X)]$ можно с помощью оценки Монте-Карло:

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i).$$

   Эта оценка является несмещенной и состоятельной.

# Повторение

2. Оценивание характеристик распределения.
   Метод Монте-Карло.

   В более сложных случаях, когда $g$ зависит от других характеристик $X$, можно воспользоваться:

   **Plug-in principle 1:** если оценка некоторой характеристики требует знания каких-то других характеристик, то можно попробовать подставить в оценку вместо неизвестных характеристик их оценки.

   **Plug-in principle 2:** если необходимо оценить какую-то функцию от нескольких неизвестных характеристик, то можно подставить оценки характеристик в эту функцию.

# Повторение

2. Оценивание характеристик распределения.
   Метод Монте-Карло.

   При этом, естественно, нет никаких гарантий, что полученная оценка будет несмещенной и состоятельной.

   Несмещенная оценка дисперсии:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2.$$

# Доверительные интервалы

Пусть нам дана реализация выборки $x_1, \ldots, x_n$ из некоторого распределения $F_\theta$ с неизвестным параметром

$$\theta \in \Theta \subset \mathbb{R}.$$

До сих пор мы занимались «точечным оцениванием» неизвестного параметра — находили оценку, способную в некотором смысле заменить параметр.

Существует другой подход к оцениванию, при котором мы указываем интервал, накрывающий параметр с заданной наперед вероятностью. Такой подход называется «интервальным оцениванием».

# Доверительные интервалы

Пусть $\alpha \in (0, 1)$. Две оценки $\widehat{\theta}_1$ и $\widehat{\theta}_2$ определяют границы доверительного интервала для параметра $\theta$ с коэффициентом доверия $1 - \alpha$, если для выборки $\mathbf{X} = (X_1, \ldots, X_n)$ из закона распределения $F_\theta$ при всех $\theta \in \Theta$ справедливо неравенство

$$\mathbb{P}\Big(\widehat{\theta}_1(\mathbf{X}) < \theta < \widehat{\theta}_2(\mathbf{X})\Big) \geq 1 - \alpha.$$

Как правило, длина доверительного интервала возрастает при увеличении коэффициента доверия $1 - \alpha$ и стремится к нулю с ростом размера выборки $n$.

# Доверительные интервалы

Если вероятность в левой части неравенства в пределе не превосходит $1 - \alpha$ при $n \to \infty$, то есть выполняется

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{\theta}_1(\mathbf{X}) < \theta < \widehat{\theta}_2(\mathbf{X})\right) \geq 1 - \alpha,$$

то доверительный интервал называется асимптотическим.

Асимптотические доверительные интервалы возникают тогда, когда мы пользуемся предельными теоремами (например, центральной предельной теоремой).

# Доверительные интервалы

Неравенство «$\geq 1 - \alpha$» обычно соответствует дискретным распределениям, когда нельзя добиться равенства.

Например, для $X \sim \mathrm{Ber}(1/2)$ равенство $\mathbb{P}(X < a) = 0.25$ невозможно при любом $a$, но неравенство имеет смысл:

$$\mathbb{P}(X < a) \geq 0.25 \quad \text{для } a > 0.$$

Если вероятность доверительному интервалу накрыть параметр равна $1 - \alpha$, интервал называют точным доверительным интервалом.

# Доверительные интервалы

Прежде чем рассматривать какие-то способы построения доверительных интервалов, разберем два примера и затем попробуем извлечь из этих примеров некоторую общую философию доверительных интервалов.

# Доверительные интервалы

## Задача

Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения $\mathcal{N}(\theta, \sigma^2)$ с неизвестным параметром $\theta \in \mathbb{R}$ и известным параметром $\sigma^2 > 0$.

Построить точный доверительный интервал для параметра $\theta$ уровня доверия $1 - \alpha$.

# Доверительные интервалы

**Решение.** Будем пользоваться фактом, что нормальное распределение устойчиво по суммированию:

если

- $X_1 \sim \mathcal{N}(a_1, \sigma_1^2)$,
- $X_2 \sim \mathcal{N}(a_2, \sigma_2^2)$,
- $X_1$ и $X_2$ независимы,

то

$$X_1 + X_2 \sim \mathcal{N}(a_1 + a_2, \sigma_1^2 + \sigma_2^2).$$

# Доверительные интервалы

Поэтому распределение суммы элементов выборки нормально:

$$n\overline{X} = X_1 + \ldots + X_n \sim \mathcal{N}(n\theta, n\sigma^2).$$

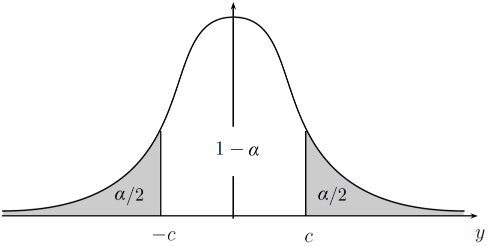
Следовательно, после стандартизации суммы мы получим стандартное нормальное распределение:

$$\frac{n\overline{X} - n\theta}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}(\overline{X} - \theta)}{\sigma} \sim \mathcal{N}(0, 1).$$

# Доверительные интервалы

По заданному $\alpha > 0$ найдём число $c$ такое, что

$$\mathbb{P}\left(-c < \frac{\sqrt{n}(\overline{X} - \theta)}{\sigma} < c\right) = 1 - \alpha.$$



# Доверительные интервалы

Разрешив затем неравенство внутри вероятности относительно $\theta$, получим точный доверительный интервал:

$$\mathbb{P}\left(\overline{X} - \frac{c\sigma}{\sqrt{n}} < \theta < \overline{X} + \frac{c\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Это можно записать и так:

$$\theta \in \left(\overline{X} - \frac{c\sigma}{\sqrt{n}}, \overline{X} + \frac{c\sigma}{\sqrt{n}}\right) \quad \text{с вероятностью } 1 - \alpha.$$

# Доверительные интервалы

Пусть $F(x)$ — функция распределения некоторого закона.
Число $c_\alpha$ называется квантилью уровня $\alpha$, если $F(c_\alpha) = \alpha$.

Если функция $F$ строго монотонна, квантиль определяется единственным образом.

# Доверительные интервалы

Итак, искомый точный доверительный для нормального распределения имеет вид:

$$\mathbb{P}\left(\overline{X} - \frac{c_{1-\alpha/2}\sigma}{\sqrt{n}} < \theta < \overline{X} + \frac{c_{1-\alpha/2}\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

где мы использовали тот факт, что $c_{\alpha/2} = -c_{1-\alpha/2}$.

- Какова середина полученного доверительного интервала?
- Какова его длина?
- Что происходит с его границами при $n \to \infty$?

# Доверительные интервалы

- Зачем мы брали симметричные квантили?
- Какой будет длина, например, у такого доверительного интервала?

$$\mathbb{P}\left(\overline{X} - \frac{c_{1-\alpha/3}\sigma}{\sqrt{n}} < \theta < \overline{X} + \frac{c_{1-2\alpha/3}\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

- Какой из двух доверительных интервалов одного уровня доверия и разной длины следует предпочесть?

# Доверительные интервалы

## Задача

Пусть $X_1, \ldots, X_n$ — выборка из экспоненциального распределения $\mathrm{Exp}(\theta)$ с неизвестным параметром $\theta > 0$.

Построить асимптотически точный доверительный интервал для параметра $\theta$ уровня доверия $1 - \alpha$.

# Доверительные интервалы

**Решение.** Вспомним центральную предельную теорему: для больших $n$

$$\frac{n\overline{X} - \mathbb{E}[n\overline{X}]}{\sqrt{\mathrm{Var}(n\overline{X})}} = \frac{\sqrt{n}(\overline{X} - 1/\theta)}{1/\theta} = \sqrt{n}(\theta\overline{X} - 1) \quad \approx \quad \mathcal{N}(0, 1).$$

Следовательно, можем записать, что для произвольных $a < b$

$$\mathbb{P}\Big(a < \sqrt{n}(\theta\overline{X} - 1) < b\Big) \to \mathbb{P}\Big(a < Z < b\Big) \quad \text{при } n \to \infty.$$

# Доверительные интервалы

Возьмём, как в прошлой задаче, следующие квантили стандартного нормального распределения:

$$a = c_{\alpha/2} = -c_{1-\alpha/2}, \quad b = c_{1-\alpha/2},$$

и получим

$$\mathbb{P}\left(-c_{1-\alpha/2} < \sqrt{n}(\theta\overline{X} - 1) < c_{1-\alpha/2}\right) \to 1 - \alpha \quad \text{при } n \to \infty.$$

# Доверительные интервалы

Разрешив относительно $\theta$ неравенство внутри вероятности, получим асимптотический доверительный интервал:

$$\mathbb{P}\left(\frac{1}{\overline{X}} - \frac{c_{1-\alpha/2}}{\overline{X}\sqrt{n}} < \theta < \frac{1}{\overline{X}} + \frac{c_{1-\alpha/2}}{\overline{X}\sqrt{n}}\right) \to 1 - \alpha \quad \text{при } n \to \infty.$$

# Доверительные интервалы

## Построение точных доверительных интервалов:

1. Найти функцию $G(\mathbf{X}, \theta)$ с известным распределением, которое не зависит от неизвестного параметра $\theta$. Необходимо, чтобы функция $G(\mathbf{X}, \theta)$ была обратима по $\theta$.
2. Найти числа $c_1$ и $c_2$ — квантили распределения, для которых
$$\mathbb{P}(c_1 < G(\mathbf{X}, \theta) < c_2) = 1 - \alpha.$$
3. Разрешив неравенство $c_1 < G(\mathbf{X}, \theta) < c_2$ относительно $\theta$ получить точный доверительный интервал.

# Доверительные интервалы

## Построение асимптотических доверительных интервалов:

1. Найти функцию $G(\mathbf{X}, \theta)$, которая бы сходилась к известной случайной величине $Z$, не зависящей от неизвестного параметра $\theta$. Необходимо, чтобы функция $G(\mathbf{X}, \theta)$ была обратима по $\theta$.

2. Найти числа $c_1$ и $c_2$ — квантили распределения $Z$, для которых

$$\mathbb{P}(c_1 < G(\mathbf{X}, \theta) < c_2) \to \mathbb{P}(c_1 < Z < c_2) = 1 - \alpha.$$

3. Разрешив неравенство $c_1 < G(\mathbf{X}, \theta) < c_2$ относительно $\theta$ получить асимптотический доверительный интервал.

# Доверительные интервалы

## Задача

Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения $\mathcal{N}(\mu, \sigma^2)$. Мы построили точный доверительный интервал для среднего $\mu \in \mathbb{R}$ при известной дисперсии $\sigma^2 > 0$.

Построим оставшиеся точные доверительные интервалы: для среднего $\mu$ при неизвестной дисперсии $\sigma^2$, а также для дисперсии $\sigma^2$ при известном и при неизвестном среднем $\mu$.

# Доверительные интервалы

Можно ли, пользуясь схемой построения доверительного интервала для среднего нормального распределения, построить точный доверительный интервал для дисперсии?

Попробуйте разрешить неравенство относительно $\sigma$:

$$-c < \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} < c.$$

- Чем плох интервал бесконечной длины?
- А получился ли интервал бесконечной длины?

# Распределения, связанные с нормальным

Для решения этой задачи требуется отыскать такие функции от выборки и неизвестных параметров, распределения которых не зависят от этих параметров.

Особый интерес к нормальному распределению связан, разумеется, с центральной предельной теоремой: почти всё в этом мире нормально (или близко к нормальному).

# Распределения, связанные с нормальным

Пусть $X_1, \ldots, X_k$ независимы и имеют стандартное нормальное распределение $\mathcal{N}(0, 1)$.

Распределением хи-квадрат с $k$ степенями свободы называется распределение случайной величины

$$Y = X_1^2 + \ldots + X_k^2.$$

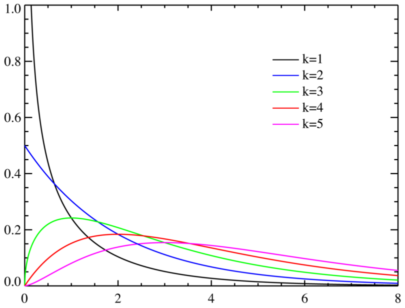
Обозначение: $\chi_k^2$.

# Распределения, связанные с нормальным

Плотность распределения хи-квадрат с $k$ степенями свободы:

$$f(u) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)}\, u^{k/2-1} e^{-u/2}, & u > 0, \\ 0, & u \le 0, \end{cases}$$

где $\Gamma(u)$ — гамма-функция Эйлера (специальная функция).



# Распределения, связанные с нормальным

Пусть $X_0, X_1, \ldots, X_k$ независимы и имеют стандартное нормальное распределение $\mathcal{N}(0,1)$.

Распределением Стьюдента с $k$ степенями свободы называется распределение случайной величины

$$Y = \frac{X_0}{\sqrt{\frac{X_1^2 + \ldots + X_k^2}{k}}}.$$

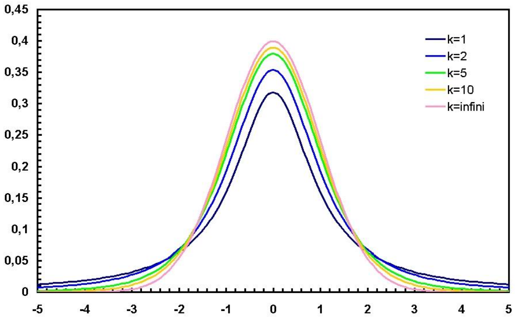
Обозначение: $t_k$.

# Распределения, связанные с нормальным

Плотность распределения Стьюдента с $k$ степенями свободы:

$$f(u) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{u^2}{k}\right)^{-\frac{k+1}{2}},$$

где $\Gamma(u)$ — гамма-функция Эйлера (специальная функция).



# Распределения, связанные с нормальным

## Теорема (Лемма Фишера)

*Пусть* $X_1, \dots, X_n$ *— выборка из нормального распределения* $\mathcal{N}(\mu, \sigma^2)$ *с некоторыми* $\mu \in \mathbb{R}$ *и* $\sigma^2 > 0$. *Обозначим*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

*Тогда случайные величины* $\overline{X}$ *и* $S^2$ *независимы и*

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

# Доверительные интервалы в нормальной модели

Вернемся к задаче построения доверительных интервалов в нормальной модели. Рассмотрим все возможные случаи.

- Доверительный интервал для $\sigma^2$ при известном $\mu$.

  В этой модели можно рассмотреть статистику:

$$\sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2.$$

  Пусть $c_{\alpha/2}$ и $c_{1-\alpha/2}$ будут квантилями $\chi_n^2$. Тогда

$$\mathbb{P}\left( c_{\alpha/2} < \sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2 < c_{1-\alpha/2} \right) = 1 - \alpha.$$

# Доверительные интервалы в нормальной модели

Разрешив неравенство, получим

$$\mathbb{P}\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{c_{1-\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{c_{\alpha/2}}\right) = 1-\alpha.$$

Обозначим оценку дисперсии $\sigma^2$ при известном $\mu$ через $S_o^2$:

$$S_o^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2.$$

Тогда доверительный интервал можно записать так:

$$\mathbb{P}\left(\frac{nS_o^2}{c_{1-\alpha/2}} < \sigma^2 < \frac{nS_o^2}{c_{\alpha/2}}\right) = 1-\alpha.$$

# Доверительные интервалы в нормальной модели

- Доверительный интервал для $\sigma^2$ при неизвестном $\mu$.

  В этой модели можно рассмотреть статистику:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

  Проводя все те же вычисления, что и в предыдущем пункте, мы получим:

$$\mathbb{P}\left(\frac{(n-1)S^2}{c_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{c_{\alpha/2}}\right) = 1-\alpha,$$

  где $c_{\alpha/2}$ и $c_{1-\alpha/2}$ уже квантили распределения $\chi^2_{n-1}$.

# Доверительные интервалы в нормальной модели

- Доверительный интервал для $\mu$ при неизвестном $\sigma^2$.

  В этой модели можно рассмотреть статистику:

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t_{n-1}.$$

  Действительно, по лемме Фишера

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S} = \frac{\frac{\sqrt{n}}{\sigma}(\overline{X}-\mu)}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}} = \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} \sim t_{n-1}.$$

# Доверительные интервалы в нормальной модели

Поэтому

$$\mathbb{P}\left(-c_{1-\alpha/2} < \frac{\sqrt{n}(\overline{X}-\mu)}{S} < c_{1-\alpha/2}\right) = 1-\alpha,$$

где $c_{1-\alpha/2}$ — квантиль $t_{n-1}$ (так как распределение Стьюдента симметрично, то $c_{\alpha/2} = -c_{1-\alpha/2}$).

Разрешив неравенство, получим

$$\mathbb{P}\left(\overline{X} - \frac{c_{1-\alpha/2}S}{\sqrt{n}} < \mu < \overline{X} + \frac{c_{1-\alpha/2}S}{\sqrt{n}}\right) = 1-\alpha.$$

# Доверительные интервалы в нормальной модели

**Резюме.** Пусть $X_1, \ldots, X_n$ — выборка из $\mathcal{N}(\mu, \sigma^2)$.

- доверительный интервал для $\mu$ при известном $\sigma^2$:

$$\mathbb{P}\left(\overline{X} - \frac{c_{1-\alpha/2}\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{c_{1-\alpha/2}\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

  где $c_{1-\alpha/2}$ — квантиль распределения $\mathcal{N}(0, 1)$.

- доверительный интервал для $\mu$ при неизвестном $\sigma^2$:

$$\mathbb{P}\left(\overline{X} - \frac{c_{1-\alpha/2}S}{\sqrt{n}} < \mu < \overline{X} + \frac{c_{1-\alpha/2}S}{\sqrt{n}}\right) = 1 - \alpha,$$

  где $c_{1-\alpha/2}$ — квантиль распределения $t_{n-1}$.

# Доверительные интервалы в нормальной модели

- доверительный интервал для $\sigma^2$ при известном $\mu$:

$$\mathbb{P}\left(\frac{nS_o^2}{c_{1-\alpha/2}} < \sigma^2 < \frac{nS_o^2}{c_{\alpha/2}}\right) = 1 - \alpha,$$

  где $c_{\alpha/2}$ и $c_{1-\alpha/2}$ — квантили распределения $\chi^2_n$.

- доверительный интервал для $\sigma^2$ при неизвестном $\mu$:

$$\mathbb{P}\left(\frac{(n-1)S^2}{c_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{c_{\alpha/2}}\right) = 1 - \alpha,$$

  где $c_{\alpha/2}$ и $c_{1-\alpha/2}$ — квантили распределения $\chi^2_{n-1}$.

# Бутстрэп

Бутстрэп — это набор практических методов, который основан на многократной генерации выборок на базе одной имеющейся выборки.

Бутстрэп используется для оценки каких-то параметров распределений, построения доверительных интервалов и т.д.

Рассмотрим параметрический и непараметрический бутстрэп.

# Бутстрэп

## Параметрический бутстрэп:

- Делается предположение, что данные получены из некоторого параметрического семейства $F_\theta$.
- Новые выборки генерируются из закона $F_{\widehat{\theta}}$, где $\widehat{\theta}$ — некоторая оценка неизвестного параметра $\theta$.
- Если семейство распределений $F_\theta$ непрерывно зависит от параметра и оценка $\widehat{\theta}$ не сильно уклонилась от истинного значения, то $F_{\widehat{\theta}}$ будет близко к закону, из которого получена выборка.
- Новые выборки используем для оценки того, что нужно.

# Бутстрэп

**Непараметрический бутстрэп:**

- Никакого предположения относительно семейства распределений $F_\theta$ не делается.
- Новые выборки генерируются с помощью выбора с возвращением из исходной выборки.
- У этой идеи есть теоретическое подспорье: мы тем самым генерируем новую выборку из эмпирической функции распределения, которая является хорошим приближением истинной функции распределения.
- Новые выборки используем для оценки того, что нужно.

# Бутстрэп

Теперь подробнее о теоретическом обосновании работы непараметрического бутстрэпа.

Эмпирическая функция распределения $\widehat{F}_n(u)$ определяется формулой

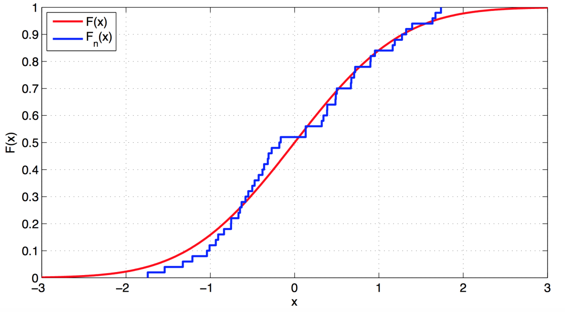
$$\widehat{F}_n(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\{x_i \le u\}},$$

где $\mathbf{I}_{\{x_i \le u\}}$ — индикатор события $\{x_i \le u\}$, то есть

$$\mathbf{I}_{\{x_i \le u\}} = \begin{cases} 1, & \text{если } x_i \le u, \\ 0, & \text{если } x_i > u. \end{cases}$$

# Бутстрэп

График $\widehat{F}_n(x)$ представляет собой ступенчатую функцию, растущую скачками высоты $1/n$. Скачки происходят в точках реализации выборки $x_1, \ldots, x_n$.



# Бутстрэп

Эмпирическая функция распределения является хорошим приближением для истинной функции распределения.

**Теорема (Гливенко-Кантелли)**

*Пусть $F$ — функция распределения элементов выборки. Тогда с вероятностью* 1

$$\sup_{u\in\mathbb{R}} \left| F(u) - \widehat{F}_n(u) \right| \to 0 \quad \textit{при } n \to \infty.$$

Следовательно, чтобы сгенерировать бустрэп-выборку, можно использовать закон, соответствующий эмпирической функции распределении. А это и будет выбором с возвращением.

# Бутстрэп

**Как строить доверительные интервалы с помощью бутстрэпа?**

Существует и несколько методов построения доверительных интервалов. Наиболее простой из них — pivotal интервал.

Его идея заключается в том, чтобы посчитать некоторую характеристику, доверительный интервал для которой мы хотим построить, много раз на основе бутрэп-выборок и затем «отрезать» выборочные квантили.

# Бутстрэп

**Пример.** Допустим мы хотим построить доверительный интервал для некоторого параметра $\theta$ параметрического распределения $F_\theta$ на основе одной выборки $x_1, \ldots, x_n$.

- Сгенерируем $m$ новых бутстрэп-выборок.
  (Для параметрического бутстрэпа построим оценку $\widehat{\theta}$ параметра $\theta$ и будем генерировать из $F_{\widehat{\theta}}$, а для непараметрического бутстрэпа будем выбирать с возвращением из $x_1, \ldots, x_n$.)
  На их основе мы посчитаем $m$ новых оценок $\widehat{\theta}_1, \ldots, \widehat{\theta}_m$.
- Упорядочим $\widehat{\theta}_i$ и выберем те из них, $\widehat{\theta}_-$ и $\widehat{\theta}_+$, которые стоят на местах $[(\alpha/2)m]$ и $[(1-\alpha/2)m]$ по возрастанию;
- В качестве доверительного интервала возьмем

$$\left(\widehat{\theta}_-\,,\,\widehat{\theta}_+\right).$$

# Бутстрэп

**Плюсы и минусы бустрэпа.**

Бутстрэп прост в использовании, не требует сложных вычислений и применим даже к весьма громоздким моделям.

С другой стороны, мы не можем явным образом оценить его погрешность. В случае, если оценка $\widehat{\theta}$ значимо промахнулась мимо истинного значения $\theta$ или эмпирическая функция распределения $\widehat{F}_n$ сильно отличается от истинной $F$, мы рискуем сильно ошибиться в выводах.

Спасибо за внимание!