

Введение в математическую статистику. Доверительные интервалы. Бутстрэп

Леонид Иосипой

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

3 февраля 2021

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

Доверительные интервалы

Пусть, как обычно, имеется реализация выборки x_1, \dots, x_n из некоторого распределения F_θ 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 $\hat{\theta}_1$ и $\hat{\theta}_2$ определяют границы доверительного интервала для параметра θ с коэффициентом доверия $1 - \alpha$, если для выборки $\mathbf{X} = (X_1, \dots, X_n)$ из закона распределения F_θ при всех $\theta \in \Theta$ справедливо неравенство

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

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 $n \rightarrow \infty$, то есть выполняется

$$\lim_{n \rightarrow \infty} \mathbb{P}(\hat{\theta}_1(\mathbf{X}) < \theta < \hat{\theta}_2(\mathbf{X})) \geq 1 - \alpha,$$

то доверительный интервал называется **асимптотическим**.

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

С асимптотическим доверительным интервалом мы сталкивались, когда изучали метод Монте-Карло.

Доверительные интервалы

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

Например, для $X \sim \mathbf{B}_{1/2}$ равенство $\mathbb{P}(X < a) = 0.25$ невозможно при любом a , а 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, \dots, 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 θ 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:

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

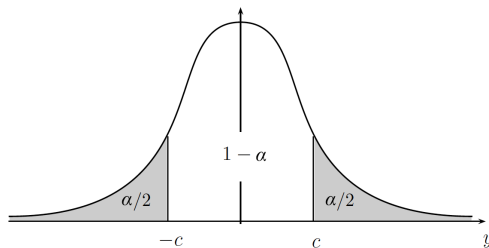
Следовательно, после стандартизации суммы мы получим стандартное нормальное распределение:

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

Доверительные интервалы

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

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



Доверительные интервалы

Разрешив затем неравенство внутри вероятности относительно θ , получим точный доверительный интервал:

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

Это можно записать и так:

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

Доверительные интервалы

Пусть $F(x)$ — функция распределения некоторого закона.
Число c_α называется **квантилью** уровня α , 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:

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

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 $n \rightarrow \infty$?

Доверительные интервалы

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

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

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

Доверительные интервалы

Задача

Пусть X_1, \dots, X_n — выборка из экспоненциального распределения Exp_θ с неизвестным параметром $\theta > 0$.

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

Доверительные интервалы

Решение. Вспомним центральную предельную теорему: для больших n

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

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

$$\mathbb{P}\left(a < \sqrt{n}(\theta\bar{X} - 1) < b\right) \rightarrow \mathbb{P}\left(a < Z < b\right) \quad \text{при } n \rightarrow \infty.$$

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

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

Доверительные интервалы

Разрешив относительно θ неравенство внутри вероятности, получим асимптотический доверительный интервал:

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

Доверительные интервалы

Построение точных доверительных интервалов:

1. Найти функцию $G(\mathbf{X}, \theta)$ с известным распределением, которое не зависит от неизвестного параметра θ .
Необходимо, чтобы функция $G(\mathbf{X}, \theta)$ была обратима по θ .
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 θ 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 θ . Необходимо, чтобы функция $G(\mathbf{X}, \theta)$ была обратима по θ .
2. Найти числа c_1 и c_2 — квантили распределения Z , для которых

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

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

Доверительные интервалы

Задача

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

Построить точный доверительный интервал для параметра θ^2 уровня доверия $1 - \alpha$.

Доверительные интервалы

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

Попробуйте разрешить неравенство относительно θ :

$$-c < \frac{\sqrt{n}(\bar{X} - \mu)}{\theta} < c.$$

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

Распределения, связанные с нормальным

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

Остался нерешённым вопрос: как построить точные доверительные интервалы для σ^2 при известном и при неизвестном μ , а также для μ при неизвестной σ^2 .

Распределения, связанные с нормальным

Для решения этих задач 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).

Распределение хи-квадрат

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

Распределением χ^2 (хи-квадрат) с k степенями свободы называется распределение случайной величины

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

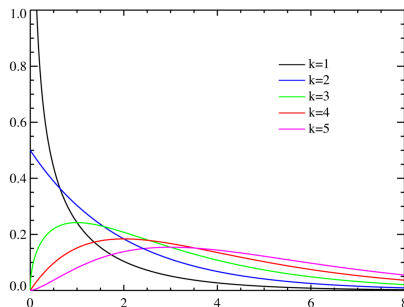
Обозначение: χ_k^2 или H_k .

Распределение хи-квадрат

Плотность распределения хи-квадрат с k степенями свободы:

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

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



Распределение хи-квадрат

Задача

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

Построить точный доверительный интервал для параметра θ^2 уровня доверия $1 - \alpha$.

Распределение хи-квадрат

Решение. В этой модели можно рассмотреть статистику:

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

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

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

Распределение хи-квадрат

Разрешив неравенство, получим

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

Обозначим оценку дисперсии σ^2 при известном μ через S_o^2 :

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

Тогда доверительный интервал можно записать так:

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

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

Распределение хи-квадрат

Задача

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

Построить точный доверительный интервал для параметра σ^2 уровня доверия $1 - \alpha$.

Распределение хи-квадрат

Решение. Обозначим несмещенную оценку дисперсии σ^2 при неизвестном μ через S^2 :

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

Рассмотрим статистику

$$\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2}.$$

Из некоторого общего факта (лемма Фишера) следует, что

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

Распределение хи-квадрат

Проводя все те же вычисления, что и в предыдущей задаче, 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 χ_{n-1}^2 .

Распределение Стьюдента

Английский статистик Госсет, публиковавший научные труды под псевдонимом Стьюдент, ввёл следующее распределение.

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

Распределением Стьюдента с k степенями свободы называется распределение случайной величины

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

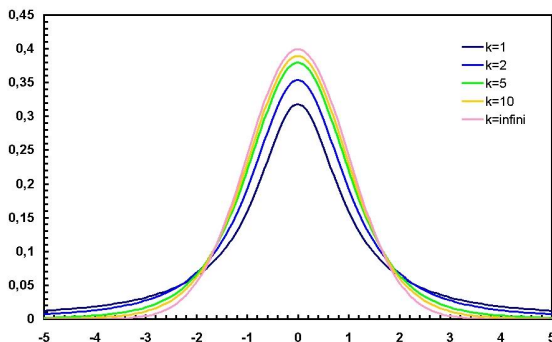
Обозначение: t_k или T_k .

Распределение Стьюдента

Плотность распределения Стьюдента с k степенями свободы:

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

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



Распределение Стьюдента

Задача

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

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

Распределение Стьюдента

В данной модели можно показать, что

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

где, напомним, S^2 — несмещенная оценка дисперсии,

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

Распределение Стьюдента

Поэтому

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

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

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

Доверительные интервалы в нормальной модели

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

- ▶ доверительный интервал для μ при известном σ^2 :

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

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

- ▶ доверительный интервал для μ при неизвестном σ^2 :

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

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

Доверительные интервалы в нормальной модели

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

- ▶ доверительный интервал для σ^2 при известном μ :

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

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

- ▶ доверительный интервал для σ^2 при неизвестном μ :

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

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

Спасибо за внимание!