Прикладная статистика. Доверительные интервалы и Бутстрэп

Леонид Иосипой

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

24 ноября 2020

- Доверительные интервалы
- Распределения, связанные с нормальным
- Доверительные интервалы в нормальной модели
- Немного о Бутстрэпе

Доверительные интервалы

Пусть, как обычно, имеется реализация выборки $x_1, \ldots, x_n$ 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 $\widehat{\theta}_1$ и $\widehat{\theta}_2$ 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

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

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).

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 $X_1, \ldots, X_n$ — 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 θ уровня доверия $1 - \alpha$.

Доверительные интервалы

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:

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

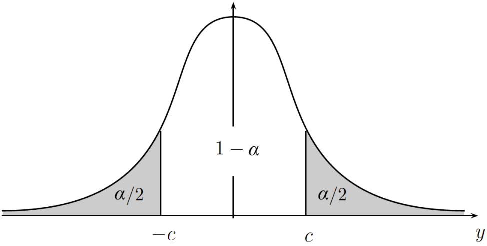
Следовательно, после стандартизации суммы мы получим стандартное нормальное распределение:

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

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 θ, 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_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:

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

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 Exp_θ с неизвестным параметром $\theta > 0$.

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

Доверительные интервалы

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 θ 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 $G(\mathbf{X}, \theta)$, распределение которой 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 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
$$\mathbb{P}(c_1 < G(\mathbf{X}, \theta) < c_2) \to \mathbb{P}(c_1 < Z < c_2) = 1 - \alpha.$$
3. Разрешив неравенство $c_1 < G(\mathbf{X}, \theta) < c_2$ относительно θ получить асимптотический доверительный интервал.

Доверительные интервалы

Задача. Пусть $X_1, \ldots, X_n$ — 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 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:

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

- 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.

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

Распределения, связанные с нормальным

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

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

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

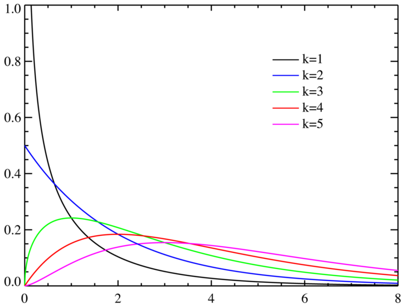
Обозначение: χ_k^2 или H_k.

Распределение хи-квадрат

Плотность распределения хи-квадрат с k степенями свободы:

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

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



Распределение хи-квадрат

Задача. Пусть $X_1, \ldots, X_n$ — 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 уровня доверия $1 - \alpha$.

Распределение хи-квадрат

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

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

Распределение хи-квадрат

Задача. Пусть $x_1, \ldots, x_n$ — реализация выборки 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 уровня доверия $1 - \alpha$.

Распределение хи-квадрат

В данной задаче можно уже рассмотреть статистику:

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

где $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})$ — несмещенная оценка дисперсии.

Из некоторого общего факта (лемма Фишера) следует, что

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

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^2_{n-1}.

Распределение Стьюдента

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

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

Распределением Стьюдента называется распределение случайной величины

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

Обозначение: T_k.

Распределение Стьюдента

Задача. Пусть $X_1, \ldots, X_n$ — 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

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

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 $c_{\alpha/2} = -c_{1-\alpha/2}$).

Распределение Стьюдента

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 μ при известном σ^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 σ^2 при известном μ:

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

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

Доверительные интервалы в нормальной модели

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

- доверительный интервал для μ при неизвестном σ^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}.

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

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

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

Немного о Бутстрэпе

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

Немного о Бутстрэпе

Пример. Построив оценку для параметра каким-либо изученным методом, мы никак не гарантируем, что полученная оценка будет несмещенной.

Но если мы часто повторяем эксперимент, нам может быть важно снизить смещение. Поэтому мы бы хотели научиться исправлять или снижать смещенность оценки.

Немного о Бутстрэпе

Идея метода бутстрэп заключается в том, что если оценка $\widehat{\theta}$ близка к настоящему параметру θ_0, то распределение $F_{\widehat{\theta}}$ будет похоже на F_{θ_0}

Мы здесь предполагаем, что семейство распределений F_θ непрерывно зависит от параметра.

Немного о Бутстрэпе

Таким образом, мы можем:

- сгенерировать выборку $Y_1, \ldots, Y_n$ из $F_{\widehat{\theta}}$, и подсчитать по ней $\widehat{\theta}(Y_1, \ldots, Y_n)$;
- «оценить» смещение $\mathbb{E}\left[\widehat{\theta}(X_1, \ldots, X_n)\right] - \theta_0$ с помощью $\widehat{\theta}(X_1, \ldots, X_n) - \widehat{\theta}(Y_1, \ldots, Y_n)$;
- посчитать «поправленную» оценку

$$2\widehat{\theta}(X_1, \ldots, X_n) - \widehat{\theta}(Y_1, \ldots, Y_n).$$

Немного о Бутстрэпе

Использовать метод бутстрэпа можно не только для оценивания смещения, но и для оценивания любых других параметров распределения/математических ожиданий.

Немного о Бутстрэпе

Этот метод имеет несколько неоспоримых плюсов — он прост в использовании и не требует сложных вычислений, применим даже к весьма громоздким моделям.

С другой стороны, мы не можем явным образом оценить его погрешность, а в случае, если оценка $\widehat{\theta}$ значимо промахнулась мимо θ_0, рискуем неправильно изменить оценку.

Немного о Бутстрэпе

Задача. Пусть имеется реализация выборки $x_1, \ldots, x_n$ из равномерного распределения на $[0, \theta]$.

Допустим мы оценили θ с помощью $2\overline{x}$. А затем берем новую выборку из равномерного распределения на $[0, 2\overline{x}]$ и оцениваем с помощью ее среднего параметр θ.

Какую дисперсию будет иметь эта новая оценка?

Немного о Бутстрэпе

Существует и несколько методов построения доверительных интервалов с помощью бутстрэпа.

Мы рассмотрим наиболее простой доверительный интервал — pivotal интервал.

Немного о Бутстрэпе

Идея: рассмотрим оценку $\widehat{\theta}$ параметра θ_0.

- возьмем несколько выборок из $F_{\widehat{\theta}}$ и построим на основе них оценки $\widehat{\theta}_1, \ldots, \widehat{\theta}_m$;
- 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;
- тогда нашим интервалом будет

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

Немного о Бутстрэпе

Очень часто бутстрэп используется в непараметрической постановке.

Это означает, что у нас нет никакого «семейства» распределений F_θ, а есть только реализация выборки $x_1, \ldots, x_n$ из некоторого неизвестного распределения F.

Немного о Бутстрэпе

Очень часто бутстрэп используется в непараметрической постановке.

Это означает, что у нас нет никакого «семейства» распределений F_θ, а есть только реализация выборки $x_1, \ldots, x_n$ из некоторого неизвестного распределения F.

В этом случае бустрэп-выборки генерируются с помощью выбора с возвращением.

Немного о Бутстрэпе

Теоретически это можно обосновать с помощью понятия эмпирической функции распределения.

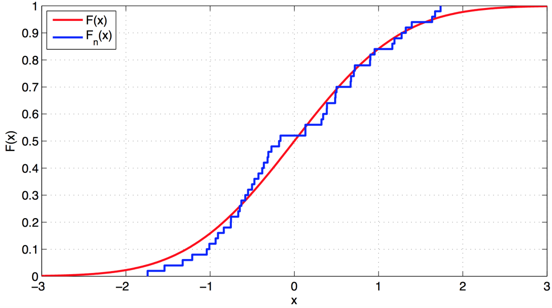
Эмпирическая функция распределения $\widehat{F}_n(u)$ определяется формулой

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

где $\mathbf{I}_{\{X_i \le u\}}$ — индикатор события $\{X_i \le u\}$.

Немного о Бутстрэпе

График $\widehat{F}_n(x)$ представляет собой ступенчатую функцию, растущую скачками высоты $1/n$. Скачки происходят в точках с координатами $X_1, \ldots, X_n$.



Немного о Бутстрэпе

Известно, что эмпирическая функция распределения является очень хорошим приближением для истинной функции распределения.

Следовательно, чтобы сгенерировать бустрэп-выборку, можно использовать закон, соответствующий эмпирической функции распределении.

А это и будет выбором с возвращением.

Спасибо за внимание!