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.

Мы рассмотрим параметрический и непараметрический бутстрэп. Начнем с параметрического.

Бутстрэп

Параметрический бутстрэп

Идея 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}. Поэтому можно генерировать новые выборки из $F_{\widehat{\theta}}$.

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

Бутстрэп

Пример

Допустим, мы построили какую-то оценку $\widehat{\theta}$ неизвестного параметра θ. Ни один из методов построения оценок, которые мы изучали, не гарантирует несмещенность.

Попытаемся исправить смещенность оценки с помощью параметрического бутстрэпа.

Бутстрэп

Это можно сделать следующим образом:

- 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

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

Бутстрэп

Бутстрэп имеет несколько неоспоримых плюсов — он 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, а в случае, если оценка $\widehat{\theta}$ значимо промахнулась мимо θ_0, рискуем неправильно изменить оценку.

Бутстрэп

Как строить доверительные интервалы с помощью бутстрэпа?

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

Бутстрэп

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

- возьмем несколько выборок из $F_{\widehat{\theta}}$ и построим на их основе другие 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

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

Бутстрэп

Непараметрический бутстрэп

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.

Бутстрэп

Теоретически это можно обосновать с помощью понятия эмпирической функции распределения.

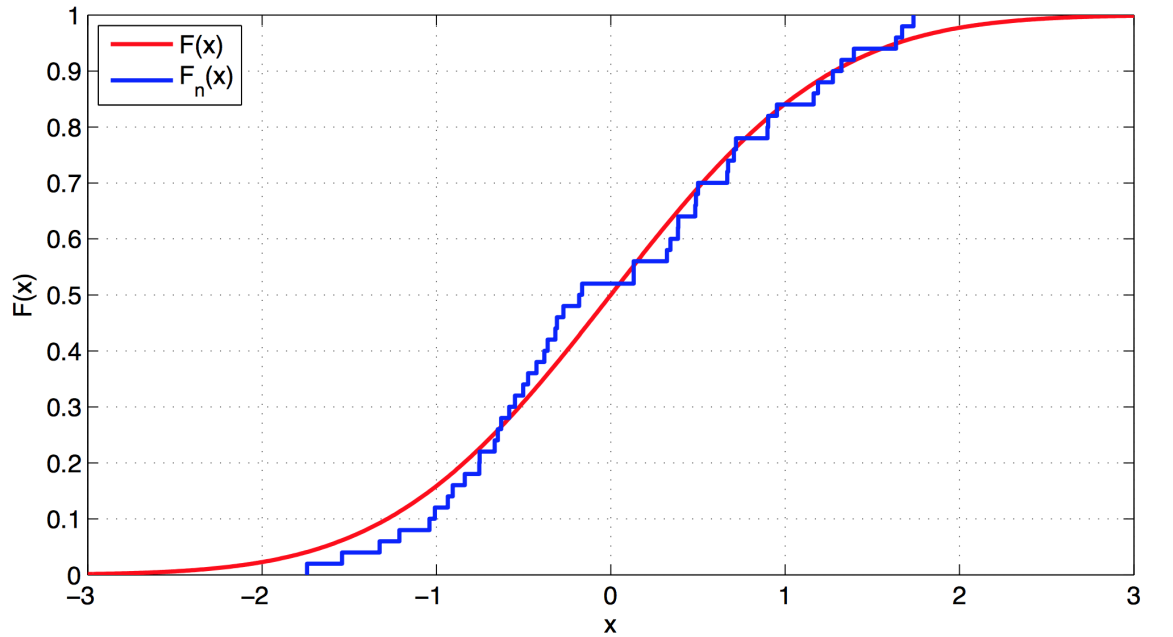
Эмпирическая функция распределения $\widehat{F}_n(u)$ определяется формулой

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

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

Бутстрэп

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



Бутстрэп

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.