Теория и практика применения бутстрэпа

Москва
2023

Содержание

1. Введение

1.1. История метода бутстрэпа

Словом "бутстрэп" обозначают некоторое семейство или множество статистических методов или алгоритмов, предназначенных для нахождения оценок параметров распределения и доверительных интервалов. Эти методы основаны на генерации большого числа выборок из исходной, имеющейся в распоряжении исследователя выборки с возвращениями с использованием датчика (псевдо-) случайных чисел. Таким образом, из-за объемности вычислений реализация бутстрэпа предполагается на компьютере.

Одной из первых работ, в которой была изложена суть метода бутстрэпа, стала статья Брэдли Эфрона «Bootstrap methods: another look at the jackknife» (1979 г.), вдохновленная более ранней работой о методе Jackknife. Позже разрабатывалась теория для нахождения более точных оценок дисперсий параметров, повышения устойчивости метода на выборках небольших размеров, с включением байесовского подхода и т.д.

1.2. Классическая теория построения доверительных интервалов

Пусть имеется выборка $\mathbf{X} = (X_1, ..., X_n)$ и известно, что она была сгенерирована из нормального распределения $\mathcal{N}(\mu, \sigma^2)$ с неизвестным математическим ожиданием μ и известной дисперсией σ^2. В таком случае известным результатом является построение следующего доверительного интервала:

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

где $\mathbb{P}\left(\xi \leqslant z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(\xi \leqslant -z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$, $\xi \sim \mathcal{N}(0, 1)$.

Кроме того, похожий результат может быть получен в случае неизвестного параметра σ^2. В таком случае используется несмещенная оценка дисперсии

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

и получается, что

$$n \cdot \frac{\bar{X} - \mu}{\hat{\sigma}} \sim t_{n-1}$$

Тогда точный доверительный интервал приобретает вид

$$\mathbb{P}\left(\bar{X} - t_{1-\frac{\alpha}{2};n-1}\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2};n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha$$

При большом числе наблюдений стандартизированная случайная величина $n \cdot \frac{\bar{X}-\mu}{\hat{\sigma}}$, согласно центральной предельной теореме, по распределению сходится к стандартному нормальному распределению вне зависимости от вида закона распределения X, т. е.

$$n \cdot \frac{\bar{X} - \mu}{\hat{\sigma}} \xrightarrow[n \to \infty]{dist} \mathcal{N}(0, 1)$$

Поэтому может быть получен следующий асимптотический доверительный интервал:

$$\mathbb{P}\left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}\right) \approx 1 - \alpha.$$

Итак, было показано, что в случае построения доверительного интервала для параметра среднего в случае известного закона распределения имеется хорошо разработанная теория и могут быть получены точные или асимптотические формулы. Кроме того, такие результаты широко известны в случае построения доверительных интервалов для разницы математических ожиданий, дисперсий и отношения дисперсий.

Однако возникает вопрос, какую схему действий предпринять в случае, если закон распределения, порождающий данные, неизвестен, а также если стоит задача построить доверительный интервал для такой не менее важной характеристики, как, например, медиана распределения, асимптотический закон распределения которой не является широко известным.

По-прежнему можно построить $\hat{\theta}$ как точечную оценку неизвестного параметра θ. Но теоретических знаний о распределении $\hat{\theta}$ у нас может не быть, либо же эта теория недоступна исследователю. В таком случае можно использовать бутстрэп для построения доверительного интервала. Однако прежде стоит разобраться с тем, каким образом можно создавать новые выборки на основе исходной.

1.3. Построение псевдовыборок

Существует два распространенных метода построения случайной выборки (пример взят из Конспекта).

- Выборка без возвращения (without replacement, simple random sampling):

Предположим, мы поочереди берем 10 карт наугад из колоды из 52 карт, не возвращая ни одну из карт обратно в колоду между взятиями. Это называется выборкой без возвращению или простой случайной выборкой. При таком методе в нашей выборке из 10 карт не будет дубликатов карт. То есть, в данном подходе, если мы захотим создать выборку, размер которой совпадает с исходной, будет получена сама исходная выборка, что ограничивает возможности в исследовании распределения данных.

- Выборка с возвращением (with replacement):

Теперь предположим, что мы берем 10 карт наугад из колоды, но после каждого взятия мы кладем карту обратно в колоду и перемешиваем карты. Это называется выборкой с возвращением. При использовании этого метода выборка из 10 карт может иметь дубликаты. Возможно и такое, что была вытянута шестерка червей все 10 раз. Такая процедура позволяет создавать отлиные от исходной новые выборки, размер которых при этом совпадает с размером исходной выборки.

1.4. Суть метода бустрэпа

На основе имеющейся выборки может быть рассчитана выборочная или эмпирическая функция распределения $\hat{F}(x)$, которая для каждого x показывает долю наблюдений в выборке, не превосходящих x:

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} \mathbb{I}\{X_i \leqslant x\}}{n}$$

Известно, что эта функция обладает рядом "хороших" свойств: состоятельность, эффективность, асимптотическая нормальность, поэтому является "хорошей" оценкой для истинной функции распределения $F(x)$, порождающей данные.

Итак, основной идеей бутстрэпа в данном случае будет генерация n_{boot} бутстрэпированных выборок $\mathbf{X}_1^*, ..., \mathbf{X}_{n_{boot}}^*$ с повторениями из исходной выборки $\mathbf{X}$ и подсчет оценки неизвестного параметра распределения $\hat{\theta}_j^*$ для j-ой бутстрэпированный выборки, $j \in \{1, ..., n_{boot}\}$, для получения распределения $\hat{\theta}$.

2. Теория основных версий бутстрэпа

2.1. Перцентильный бутстрэп

Одной из основных и наиболее простых техник бустрэпа является перцентильный бутстрэп (Percentile Bootstrap). Его алгоритм следующий:

1. Из исходной выборки $\mathbf{X}$ генерируется n_{boot} бутстрэпированных выборок $\mathbf{X}_1^*$, $\mathbf{X}_2^*$, $\ldots$, $\mathbf{X}_{n_{boot}}^*$;
2. Для каждой бутстрэпированный выборки оценивается параметр $\hat{\theta}_j^*$, $1 \leqslant j \leqslant n_{boot}$, затем оцененные параметры сортируются по возрастанию: $\hat{\theta}_{(1)}^* \leqslant \hat{\theta}_{(2)}^* \leqslant \ldots \leqslant \hat{\theta}_{(n_{boot})}^*$, и собирается вектор $\hat{\theta}^* = \left(\hat{\theta}_{(1)}^*, \hat{\theta}_{(2)}^*, \ldots, \hat{\theta}_{(n_{boot})}^*\right)$;
3. Оцененным $(1-\alpha)$-процентным доверительным интервалом истинного параметра θ будет отрезок от $\frac{\alpha}{2}$ квантиля до $1-\frac{\alpha}{2}$ квантиля вектора $\hat{\theta}^*$: $\widehat{CI}_\theta = \left[\hat{\theta}_{\frac{\alpha}{2}}^*; \hat{\theta}_{1-\frac{\alpha}{2}}^*\right]$.

Скорость сходимости: $|CI_\theta - \widehat{CI}_\theta| = O\left(\frac{1}{\sqrt{n}}\right)$.

Оцененный таким образом доверительный интервал будет занижать вероятность накрытия.

Кроме того, если выборочное распределение было асимметричным, то это также приведет к асимметрию доверительного интервала, полученного с помощью перцентильного бутстрэпа.

2.2. Обратный перцентильнй бутстрэп

Модификацией первого метода можно считать обратный перцентильный бутстрэп (Reverse Percentile Bootstrap). Его аглоритм:

1. Из исходной выборки $\mathbf{X}$ генерируется n_{boot} бутстрэпированных выборок $\mathbf{X}_1^*$, $\mathbf{X}_2^*$, $\ldots$, $\mathbf{X}_{n_{boot}}^*$;
2. Для каждой бутстрэпированный выборки оценивается параметр $\hat{\theta}_j^*$, $1 \leqslant j \leqslant n_{boot}$, далее считается отклонение полученной оценки от выборочной оценки $d_j^* = \hat{\theta}_j^* - \hat{\theta}$, полученные разницы сортируются по возрастанию: $d_{(1)}^* \leqslant d_{(2)}^* \leqslant \ldots \leqslant d_{(n_{boot})}^*$, и собирается вектор $d^* = \left(d_{(1)}^*, d_{(2)}^*, \ldots, d_{(n_{boot})}^*\right)$;
3. Оцененным $(1-\alpha)$-процентным доверительным интервалом разницы $d = \hat{\theta} - \theta$ будет отрезок от $\frac{\alpha}{2}$ квантиля до $1 - \frac{\alpha}{2}$ квантиля вектора d^*: $\widehat{CI}_d = \left[d_{\frac{\alpha}{2}}^*; d_{1-\frac{\alpha}{2}}^*\right]$. Тогда для получения оценки доверительного интервала θ необходимо выполнить ряд преобразований:

$$d_{\frac{\alpha}{2}}^* \leqslant d \leqslant d_{1-\frac{\alpha}{2}}^* \Leftrightarrow d_{\frac{\alpha}{2}}^* \leqslant \hat{\theta} - \theta \leqslant d_{1-\frac{\alpha}{2}}^* \Leftrightarrow \hat{\theta} - d_{1-\frac{\alpha}{2}}^* \leqslant \theta \leqslant \hat{\theta} - d_{\frac{\alpha}{2}}^*$$

   Таким образом, $\widehat{CI_\theta} = \left[\hat{\theta} - d_{1-\frac{\alpha}{2}}^*; \hat{\theta} - d_{\frac{\alpha}{2}}^*\right]$.

Скорость сходимости обратного перцентильного бутстрэпа совпадает со скоростью сходимости обычного перцентильного бутстрэпа: $|CI_\theta - \widehat{CI}_\theta| = O\left(\frac{1}{\sqrt{n}}\right)$.

2.3. Стьюдентизированный бутстрэп (бутстрэп t-статистики)

Далее рассмотрим метод, который позволит рассчитывать стандартную ошибку оценки, и учитывать ее влияние на распределение. Алгоритм стьюдентизированного бутстрэпа:

1. Из исходной выборки $\mathbf{X}$ генерируется n_{boot} бутстрэпированных выборок $\mathbf{X}_1^*$, $\mathbf{X}_2^*$, $\ldots$, $\mathbf{X}_{n_{boot}}^*$;
2. Для каждой бутстрэпированный выборки оценивается параметр $\hat{\theta}_j^*$, $1 \leqslant j \leqslant n_{boot}$, далее считается стандартная ошибка полученной оценки. В случае построения доверительного интервала для математического ожидания можно воспользоваться формулой

$$\widehat{se}\left(\hat{\theta}_j^*\right) = \sqrt{\frac{\sum_{i=1}^{n}(X_i^* - \bar{X}^*)^2}{n \cdot (n-1)}}$$

   где $\bar{X}^*$ – среднее по бутстрэпированный выборке. Однако точная формула стандартной ошибки, например, медианы либо других выборочных характеристик может быть неизвестна.