

Lecture 5: Bootstrap

Instructor: Yen-Chi Chen

Question 1: error of sample median? We start with a simple example: what is the error of *sample median*? Like sample mean is an estimate of the mean of population, the sample median is an estimate of the median of population. Because it is an estimator, we can define the bias, variance, and mean square error (MSE) of sample median. But what are these quantities?

Question 2: confidence interval of sample median? Moreover, how can we construct a confidence interval for the population median? We know that given a random sample $X_1, \dots, X_n \sim F$, a $1-\alpha$ confidence interval of population mean is

$$\bar{X}_n \pm z_{1-\alpha/2} \cdot \frac{\hat{\sigma}_n}{\sqrt{n}},$$

where \bar{X}_n and $\hat{\sigma}_n$ are the sample mean and sample standard deviation. Can we do the same thing (construct a confidence interval) for the median?

In this lecture, we will address these problems for median and many other statistics using the well-known approach: the *bootstrap*.

5.1 Empirical Bootstrap

Here is how we can estimate the error of sample median and construct the corresponding confidence interval. Assume we are given the data points X_1, \dots, X_n . Let $M_n = \text{median}\{X_1, \dots, X_n\}$. First, we *sample with replacement* from these n points, leading to a set of new observations denoted as $X_1^{*(1)}, \dots, X_n^{*(1)}$. Again, we repeat the sample procedure again, generating a new sample from the original dataset X_1, \dots, X_n by sampling with replacement, leading to another new sets of observations $X_1^{*(2)}, \dots, X_n^{*(2)}$. Now we keep repeating the same process of generating new sets of observations, after B rounds, we will obtain

$$\begin{array}{ccc} X_1^{*(1)}, \dots, X_n^{*(1)} \\ X_1^{*(2)}, \dots, X_n^{*(2)} \\ \vdots & \vdots & \vdots \\ X_1^{*(B)}, \dots, X_n^{*(B)}. \end{array}$$

So totally, we will have B sets of data points. Each set of the data points, say $X_1^{*(1)}, \dots, X_n^{*(1)}$, is called a bootstrap sample. This sampling approach—sample with replacement from the original dataset—is called the *empirical bootstrap*, invented by Bradley Efron (sometimes this approach is also called *Efron's bootstrap* or *nonparametric bootstrap*)¹.

Now for each set of data, we then compute their sample median. This leads to B sample medians, called

¹For more details, check the wikipedia: [https://en.wikipedia.org/wiki/Bootstrapping_\(statistics\)](https://en.wikipedia.org/wiki/Bootstrapping_(statistics))

bootstrap medians:

$$\begin{aligned} M_n^{*(1)} &= \text{median}\{X_1^{*(1)}, \dots, X_n^{*(1)}\} \\ M_n^{*(2)} &= \text{median}\{X_1^{*(2)}, \dots, X_n^{*(2)}\} \\ &\vdots \\ M_n^{*(B)} &= \text{median}\{X_1^{*(B)}, \dots, X_n^{*(B)}\}. \end{aligned}$$

Now here are some real cool things.

- **Bootstrap estimate of the variance.** We will use the sample variance of $M_n^{*(1)}, \dots, M_n^{*(B)}$ as an estimate of the variance of sample median M_n . Namely, we will use

$$\widehat{\text{Var}}_B(M_n) = \frac{1}{B-1} \sum_{\ell=1}^B \left(M_n^{*(\ell)} - \bar{M}_B^* \right)^2, \quad \bar{M}_B^* = \frac{1}{B} \sum_{\ell=1}^B M_n^{*(\ell)},$$

as an estimate of $\text{Var}(M_n)$.

- **Bootstrap estimate of the MSE.** Moreover, we can estimate the MSE by

$$\widehat{\text{MSE}}(M_n) = \frac{1}{B} \sum_{\ell=1}^B \left(M_n^{*(\ell)} - M_n \right)^2.$$

- **Bootstrap confidence interval.** In addition, we can construct a $1 - \alpha$ confidence interval of the population median via

$$M_n \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\text{Var}}_B(M_n)}.$$

Well... this sounds a bit weird—we generate new data points by sampling from the existing data points. However, under some conditions, this approach does work! And here is a brief explanation on why this approach works.

Let $X_1, \dots, X_n \sim F$. Recall from Lecture 1, a statistic $S(X_1, \dots, X_n)$ is a function of random variables so its distribution will depend on the CDF F and the sample size n . Thus, the distribution of median M_n , denoted as F_{M_n} , will also be determined by the CDF F and sample size n . Namely, we may write the CDF of median as

$$F_{M_n}(x) = \Psi(x; F, n), \tag{5.1}$$

where Ψ is some complicated function that depends on CDF of each observation F and the sample size n .

When we sample with replace from X_1, \dots, X_n , what is the distribution we are sampling from? Let $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x)$ be the EDF of these data points. The EDF is a step functions that jumps at each data point. We know that for a discrete random variable, each jump point in its CDF corresponds to the possible value of this random variable and the size of the jump is the probability of selecting that value.

Therefore, if we generate a random variable Z from \hat{F}_n , then Z has the following probability distribution:

$$P(Z = X_i) = \frac{1}{n}, \quad \text{for each } i = 1, 2, \dots, n.$$

If we generated IID $Z_1, \dots, Z_n \sim \hat{F}_n$, then the distribution of each Z_ℓ is

$$P(Z_\ell = X_i) = \frac{1}{n}, \quad \text{for each } i = 1, 2, \dots, n, \text{ and for all } \ell = 1, \dots, n.$$

What is this sample Z_1, \dots, Z_n ? This sample is a sample generated by *sampling with replacement* from X_1, \dots, X_n .

Recall that each set of the bootstrap sample, say $X_1^{*(1)}, \dots, X_n^{*(1)}$, is obtained via sampling with replacement from X_1, \dots, X_n . Thus, each set of the bootstrap sample is an IID sample from \hat{F}_n . Namely,

$$\begin{aligned} X_1^{*(1)}, \dots, X_n^{*(1)} &\sim \hat{F}_n \\ X_1^{*(2)}, \dots, X_n^{*(2)} &\sim \hat{F}_n \\ &\vdots \\ X_1^{*(B)}, \dots, X_n^{*(B)} &\sim \hat{F}_n. \end{aligned}$$

Because a bootstrap median, say $M_n^{*(1)}$, is the sample median of $X_1^{*(1)}, \dots, X_n^{*(1)}$. Its CDF, by equation (5.1), is

$$F_{M_n^{*(1)}}(x) = \Psi(x; \hat{F}_n, n).$$

And because each of the bootstrap sample are all from the distribution \hat{F}_n , we will have

$$\Psi(x; \hat{F}_n, n) = F_{M_n^{*(1)}}(x) = F_{M_n^{*(2)}}(x) = \dots = F_{M_n^{*(B)}}(x).$$

We know that \hat{F}_n is very similar to F when the sample size is large. Thus, as long as Ψ is smooth (smoothly changing) with respect to F , $\Psi(x; \hat{F}_n, n)$ will also be very similar to $\Psi(x; F, n)$, i.e.,

$$\hat{F}_n \approx F \implies F_{M_n^{*(\ell)}}(x) = \Psi(x; \hat{F}_n, n) \approx \Psi(x; F, n) = F_{M_n}(x).$$

This means:

The CDF of a bootstrap median, $F_{M_n^{(\ell)}}(x)$, is approximating the CDF of the true median, $F_{M_n}(x)$.*

This has many implications. For an example, when two CDFs are similar, their variances will be similar as well, i.e.,

$$\text{Var}\left(M_n^{*(\ell)} | X_1, \dots, X_n\right) \approx \text{Var}(M_n).^2$$

Now the bootstrap variance estimate $\widehat{\text{Var}}_B(M_n)$ is just a sample variance of $M^{*(\ell)}$. When B is large, the sample variance is about the same as the population variance, implying

$$\widehat{\text{Var}}_B(M_n) = \frac{1}{B-1} \sum_{\ell=1}^B \left(M_n^{*(\ell)} - \bar{M}_B^*\right)^2 \approx \text{Var}\left(M_n^{*(\ell)} | X_1, \dots, X_n\right).$$

Therefore,

$$\widehat{\text{Var}}_B(M_n) \approx \text{Var}\left(M_n^{*(\ell)} | X_1, \dots, X_n\right) \approx \text{Var}(M_n),$$

which explains why the bootstrap variance is a good estimate of the true variance of the median.

Generalization to other statistics. The bootstrap can be applied to many other statistics such as sample quantiles, interquartile range, skewness (related to $\mathbb{E}(X^3)$), kurtosis (related to $\mathbb{E}(X^4)$), ...etc. The theory basically follows from the same idea.

² The reason why in the left-hand-side, the variance is conditioned on X_1, \dots, X_n is because when we compute the bootstrap estimate, the original observations X_1, \dots, X_n are fixed.

Failure of the bootstrap. However, the bootstrap may fail for some statistics. One example is the minimum value of a distribution. Here is an illustration why the bootstrap fails. Let $X_1, \dots, X_n \sim \text{Uni}[0, 1]$ and $M_n = \min\{X_1, \dots, X_n\}$ be the minimum value of the sample. Then it is known that

$$n \cdot M_n \xrightarrow{D} \text{Exp}(1).$$

♠: Think about why it converges to exponential distribution.

Thus, M_n has a continuous distribution. Assume we generate a bootstrap sample X_1^*, \dots, X_n^* from the original observations. Now let $M_n^* = \min\{X_1^*, \dots, X_n^*\}$ be the minimum value of a bootstrap sample. Because each X_ℓ^* has an equal probability ($\frac{1}{n}$) of selecting each of X_1, \dots, X_n , this implies

$$P(X_\ell^* = M_n) = \frac{1}{n}.$$

Namely, for each observation in the bootstrap sample, we have a probability of $1/n$ selecting the minimum value of the original sample. Thus, the probability that we *do not select* M_n in the bootstrap sample is

$$P(\text{none of } X_1^*, \dots, X_n^* \text{ select } M_n) = \left(1 - \frac{1}{n}\right)^n \approx e^{-1}.$$

This implies that with a probability $1 - e^{-1}$, one of the observation in the bootstrap sample will select the minimum value of the original sample M_n . Namely,

$$P(M_n^* = M_n) = 1 - e^{-1}.$$

Thus, M_n^* has a huge probability mass at the value M_n , meaning that the distribution of M_n^* will not be close to an exponential distribution.

5.2 Parametric Bootstrap

When we assume the data is from a parametric model (e.g., from Normal distribution, exponential distribution, ...etc), we can use the parametric bootstrap to access the uncertainty (variance, mean square errors, confidence intervals) of the estimated parameter. Here is an illustration using the *variance* of a normal distribution.

Example: normal distribution. Let $X_1, \dots, X_n \sim N(0, \sigma^2)$, where σ^2 is an unknown number. A natural way to estimate σ^2 is via the sample variance $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Because the sample variance is an estimator, it is a random quantity. How do we estimate the variance of the sample variance? How do we estimate the MSE of the sample variance? How do we construct a $1 - \alpha$ confidence interval for σ^2 ³?

Here is what we are going to do. Because we know that the sample variance is a good estimator of σ^2 , we can use it to replace σ^2 , leading to a new distribution $N(0, S_n^2)$. We know to sample from this new distribution, so we just generate bootstrap samples from this distribution. Assume we generate B sets of samples:

$$\begin{aligned} X_1^{*(1)}, \dots, X_n^{*(1)} &\sim N(0, S_n^2) \\ X_1^{*(2)}, \dots, X_n^{*(2)} &\sim N(0, S_n^2) \\ &\vdots \\ X_1^{*(B)}, \dots, X_n^{*(B)} &\sim N(0, S_n^2). \end{aligned}$$

³Some of you might have learned an approach via inverting the χ^2 distribution. That is a viable approach as well.

To estimate the variability of S_n^2 , we use the sample variance of each bootstrap sample. Let $S_n^{2*(1)}, \dots, S_n^{2*(B)}$ be the sample variance of each bootstrap sample ($S_n^{2*(\ell)}$ is the sample variance of $X_1^{*(\ell)}, \dots, X_n^{*(\ell)}$).

We then use

$$\widehat{\text{Var}}_B(S_n^2) = \frac{1}{B-1} \sum_{\ell=1}^B \left(S_n^{2*(\ell)} - \bar{S}_B^{2*} \right)^2, \quad \bar{S}_B^{2*} = \frac{1}{B} \sum_{\ell=1}^B S_n^{2*(\ell)},$$

as an estimator of the variance of the original sample variance, i.e., $\text{Var}(S_n^2)$. Similarly, the MSE can be estimated by

$$\widehat{\text{MSE}}_B(S_n^2) = \frac{1}{B} \sum_{\ell=1}^B \left(S_n^{2*(\ell)} - S_n^2 \right)^2.$$

And a confidence interval of σ^2 can be constructed using

$$S_n^2 \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\text{Var}}_B(S_n^2)}.$$

This approach, sampling from the distribution formed by plugging the estimated parameters, is called *parametric bootstrap*.

Example: exponential distribution. The similar approach applies to many other models. For example, if we assume the data $X_1, \dots, X_n \sim \text{Exp}(\lambda)$, where λ is an unknown quantity. And we estimate λ by an estimator such as the MLE $\hat{\lambda}_n = \frac{1}{\bar{X}_n}$. To assess the quality of $\hat{\lambda}_n$, say the its MSE, we first generate B bootstrap samples:

$$\begin{aligned} X_1^{*(1)}, \dots, X_n^{*(1)} &\sim \text{Exp}(\hat{\lambda}_n) \\ X_1^{*(2)}, \dots, X_n^{*(2)} &\sim \text{Exp}(\hat{\lambda}_n) \\ &\vdots \\ X_1^{*(B)}, \dots, X_n^{*(B)} &\sim \text{Exp}(\hat{\lambda}_n). \end{aligned}$$

Then using each sample, we obtain a bootstrap estimate of λ :

$$\hat{\lambda}_n^{*(1)}, \dots, \hat{\lambda}_n^{*(B)}, \text{ where } \hat{\lambda}_n^{*(\ell)} = \frac{1}{\bar{X}_n^{*(\ell)}} = \frac{1}{\frac{X_1^{*(\ell)} + \dots + X_n^{*(\ell)}}{n}}.$$

Then the MSE of $\hat{\lambda}_n$ can be estimated by

$$\widehat{\text{MSE}}_B(\hat{\lambda}_n) = \frac{1}{B} \sum_{\ell=1}^B \left(\hat{\lambda}_n^{*(\ell)} - \hat{\lambda}_n \right)^2.$$

5.3 ♦ :Remark on the Bootstrap

There are some variants of bootstrap such as the Jackknife⁴ (leave one observation out each time) and subsampling (only subsample m out of n sample).

Moreover, when the data are dependent such as the time series dataset or spatial dataset, the bootstrap can also be applied; in these case, one will use the block bootstrap⁵ or spatial bootstrap.

⁴https://en.wikipedia.org/wiki/Jackknife_resampling

⁵[https://en.wikipedia.org/wiki/Bootstrapping_\(statistics\)#Block_bootstrap](https://en.wikipedia.org/wiki/Bootstrapping_(statistics)#Block_bootstrap)