

# Bootstrap

Snehal M. Shekatkar

School of Computing and Data Sciences,  
FLAME University, Pune, India 412115

We have seen that a confidence interval associated with an estimator for quantity  $\theta$  can be constructed relatively easily if the sampling distribution of  $\theta$  is normal, and if the variance of the sampling distribution can be estimated (Probably it's a good time to remind yourself that the *standard error* 'se' is the square-root of the variance of the sampling distribution, and that the normal-based confidence interval is  $\hat{\theta}_n \pm z_{\alpha/2} \hat{se}_n$ ). In some cases, even if the sampling distribution is not normal, say because the sample size is too small, but if the IID sample  $X_1, X_2, \dots, X_n$  has normal distribution, then we can use workarounds like using *Student's t-distribution*, and still construct a proper  $1 - \alpha$  confidence interval.

Unfortunately, estimating the variance of sampling distribution is not always easy. This could be because the quantity  $\theta$  is a complex function of the CDF, or because it is a nonlinear statistical functional.

First consider the case where the quantity to be estimated is somewhat complex. The **skewness** of a distribution is defined as:

$$\kappa_3 = \frac{1}{\sigma^3} \int (x - \mu)^3 p(x) dx \quad (1)$$

The skewness measures asymmetry of a distribution. For example, the normal distribution being completely symmetric has skewness equal to zero. The plug-in estimator for skewness is:

$$\hat{\kappa}_3 = \frac{1}{n\hat{\sigma}^3} \sum_{i=1}^n (X_i - \bar{X}_n)^3 \quad (2)$$

Once we estimate  $\kappa_3$  using this estimator, if we decide to construct  $1 - \alpha$  confidence interval, we will need to find out the variance of  $\hat{\kappa}_3$ . However, because the formula involves  $\hat{\sigma}_3$  in the denominator, we cannot use the formula for the variance of the linear combination. Thus, it is not immediately clear how can we find  $\hat{\kappa}_3$ , and consequently the standard error.

Now consider the case for which the quantity to be estimated is a nonlinear statistical functional of the CDF. As you already know, the **median** of a distribution is defined as the value  $x$  such that  $F(x) = 0.5$ . In other words, the median is the value such that the probability that the random variable takes value less than median is exactly 0.5. Consider a linear combination  $H(x) = aF(x) + (1 - a)G(x)$  of two CDFs  $F$  and  $G$  where  $a \in [0, 1]$ . It is easy to verify that  $H(x)$  is a valid CDF. It turns out that in general:

$$T(H) = T(aF + (1 - a)G) \neq aT(F) + (1 - a)T(G) \quad (3)$$

and so median is not a linear statistical functional. Now suppose we want to estimate the median  $M$  from the IID sample  $X_1, X_2, \dots, X_n$ . Consider the following estimator:

$$\hat{M} = \begin{cases} Y_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{1}{2}(Y_{\frac{n}{2}} + Y_{\frac{n}{2}+1}) & \text{otherwise} \end{cases} \quad (4)$$

Here  $Y_j$  represent the sample  $X_i$  sorted in ascending order. It is not at all clear how we can compute  $\mathbb{V}(\widehat{M})$ .

These two examples show that we need a method to find confidence intervals for quantities of interest that works when direct computation of the variance is not possible. A method of choice in such cases is **bootstrap**. Actually, this method is a particular method in the large class of *resampling methods*. The idea behind bootstrap is to try to estimate the sampling distribution of the quantity of interest. The variance of the sampling distribution is then estimated by computing the variance of this estimated sampling distribution. The method was introduced by Bradley Efron in 1979.

Bootstrapping to estimate the variance of the estimator  $\widehat{\theta} = T(F)$  consists of two distinct steps:

1. First, replace  $\mathbb{V}_F(\widehat{\theta}_n)$  by  $\mathbb{V}_{\widehat{F}_n}(\widehat{\theta}_n)$ . Here  $\mathbb{V}_F$  denotes the variance assuming that the underlying CDF is  $F$ , whereas  $\mathbb{V}_{\widehat{F}_n}(\widehat{\theta}_n)$  is the variance computed assuming that the underlying distribution is  $\widehat{F}_n$
2. Now estimate  $\mathbb{V}_{\widehat{F}_n}(\widehat{\theta}_n)$  as follows. Draw many (say  $N$ ) samples, each of size  $n$  from  $\widehat{F}_n$ . Then compute a *bootstrap replication*  $\widehat{\theta}_n^*$  of  $\widehat{\theta}_n$  for each of these samples:

$$\widehat{\theta}_{n,j}^* = g(X_{1,j}^*, X_{2,j}^*, \dots, X_{n,j}^*)$$

Here  $X_{1,j}^*, X_{2,j}^*, \dots, X_{n,j}^*$  is  $j^{th}$  sample drawn from  $\widehat{F}_n$ , and  $g$  is the function that converts the sample into the value of the estimator. How do we draw this sample? Notice that each sample point  $X_i$  contributes weight  $1/n$  to  $\widehat{F}_n$ . Hence, it should be drawn with probability  $1/n$ . In other words, a bootstrap sample is simply obtained by uniformly randomly drawing one value at a time from the original sample with replacement!

Finally, compute the sample variance of  $\widehat{\theta}_n^*$  values as an estimate of  $\mathbb{V}_{\widehat{F}_n}(\widehat{\theta}_n)$ :

$$\mathbb{V}_{\text{boot}}(\widehat{\theta}_n) = \frac{1}{N} \sum_{j=1}^N \left( \widehat{\theta}_{n,j}^* - \frac{1}{N} \sum_{i=1}^N \widehat{\theta}_{n,i}^* \right)^2 \quad (5)$$

Notice that  $\mathbb{V}_{\widehat{F}_n}(\widehat{\theta}_n)$  can be estimated with arbitrarily large accuracy since the total number of bootstrap samples  $N$  can be arbitrarily large. However, the first step depends on the sample size  $n$ . Hence the justification of replacing  $\mathbb{V}_F(\widehat{\theta}_n)$  by  $\mathbb{V}_{\widehat{F}_n}(\widehat{\theta}_n)$  depends on the sample size  $n$ , and unless it is large, we can't decrease the size of the confidence interval no matter how many bootstrap replications we perform.

Once the variance of the sampling is estimated in this fashion, the estimated standard error is simply the square-root of the variance:

$$\widehat{se}_{\text{boot}} = \sqrt{\mathbb{V}_{\text{boot}}(\widehat{\theta}_n)} \quad (6)$$

Now, if again assume that  $\widehat{\theta}_n$  has normal distribution, the *bootstrap confidence interval* is given by:

$$\widehat{\theta}_n \pm z_{\alpha/2} \widehat{se}_{\text{boot}}$$

Even when the distribution of  $\widehat{\theta}_n$  is not normal, bootstrapping provides ways to construct a  $1 - \alpha$  confidence interval, however, here we will not discuss such cases.

Below is a python code to find a 95% bootstrap confidence interval for the median of the 'sepal widths' data from the 'iris' dataset.

---

```
import numpy as np
import matplotlib.pyplot as plt

# Load the data
```

```

x = np.loadtxt('iris_sepal_widths.dat')
n = len(x)

# Estimate the median
median_hat = np.median(x)
print(f"median_hat = {median_hat}")

N = 1000 # number of bootstrap samples to use
bootstrap_medians = []
for _ in range(N):
    # Draw a bootstrap sample
    y = np.random.choice(x, replace=True, size=n)
    bootstrap_medians.append(np.median(y))

# Compute the bootstrap standard deviation
sigma_boot = np.std(bootstrap_medians, ddof=1)
# Print the bootstrap 95% confidence interval
print("confidence interval:", round(median_hat-1.96*sigma_boot, 3), ",",
      round(median_hat+1.96*sigma_boot, 3))

```

---

On my computer, this code produces the sample median 3 and 95% confidence interval (2.94, 3.06) with 1000 bootstrap replications.

## Failure of bootstrap

At this point, we should take note of a very important fact about the bootstrap. As we saw above, the idea of the bootstrap is to estimate the sampling distribution of some quantity  $\theta$  (which could be median, mean, or anything that needs to be estimated). The hope is that as the sample size  $n \rightarrow \infty$ , the estimated distribution will converge to the true distribution. However, this may not always happen. Here is an example (probably the most famous one). Let  $X_1, X_2, \dots, X_n \sim U(0, \theta)$  where  $U(0, \theta)$  denotes the uniform distribution over the interval  $[0, \theta]$ . Now consider the following estimator for  $\theta$ :

$$\hat{\theta}_n = \max\{X_1, X_2, \dots, X_n\}$$

That is, we simply pick the maximum value from the sample as an estimate of the maximum possible value of the random variable. Consider using bootstrap to estimate sampling distribution of  $\hat{\theta}_n$ . It is easy to see that the true sampling distribution of  $\hat{\theta}_n$  is continuous given that the original distribution  $U(0, \theta)$  is continuous. We then expect that as  $n \rightarrow \infty$ , we should get a continuous distribution even when we use bootstrap to estimate it. However, as the following argument shows, this is not true.

As before, let  $\hat{\theta}_n^*$  denote a bootstrap replication of  $\hat{\theta}_n$ . Let's find out the probability that  $\hat{\theta}_n^*$  is equal to  $\hat{\theta}_n$ :

$$\mathbb{P}(\hat{\theta}_n^* = \hat{\theta}_n) = 1 - \mathbb{P}(\hat{\theta}_n^* \neq \hat{\theta}_n) = 1 - \mathbb{P}(X_1^* \neq \hat{\theta}_n, X_2^* \neq \hat{\theta}_n, \dots, X_n^* \neq \hat{\theta}_n) = 1 - \left(1 - \frac{1}{n}\right)^n$$

Hence,

$$\lim_{n \rightarrow \infty} \mathbb{P}(\hat{\theta}_n^* = \hat{\theta}_n) = 1 - \frac{1}{e} \approx 0.632$$

However, since the distribution of  $\hat{\theta}_n$  is continuous, this probability should have been zero (Recall that the probability density is different from probability), and that doesn't happen here. Thus, the estimated sampling distribution does not converge to the true sampling distribution even in the limit, and thus bootstrap fails. When you wish to apply bootstrap for a particular problem, it is best to make sure that the procedure will not fail in this fashion. You may want to have a look at this reference to understand more about such issues.