STAT 535: Statistical Machine Learning **Autumn 2025**

Lecture 12: The Bootstrap

Instructor: Yen-Chi Chen

A useful reference: Chapter 23 of

- Van der Vaart, A. W. (2000). *Asymptotic statistics (Vol. 3)*. Cambridge university press.

12.1 Introduction

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 $X_1, \cdots, X_n \sim F$, a $1-\alpha$ confidence interval of population mean is

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

where $\bar{X}_n$ and $\widehat{\sigma}_n$ 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*.

12.2 The 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 $X_1, \cdots, X_n$. Let $\widehat{\theta}_n$ be our estimator for a unknown parameter of interest θ.

The bootstrap procedure works as follows. First, we *sample with replacement* from these n points, leading to a set of new observations denoted as $X_1^{*(1)}, \cdots, X_n^{*(1)}$. Again, we repeat the sample procedure again, generating a new sample from the original dataset $X_1, \cdots, X_n$ by sampling with replacement, leading to another new sets of observations $X_1^{*(2)}, \cdots, X_n^{*(2)}$. Now we keep repeating the same process of generating new sets of observations, after B rounds, we will obtain

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

So totally, we will have B sets of data points. Each set of the data points, say $X_1^{*(1)}, \cdots, X_n^{*(1)}$, 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]. For each bootstrap sample, we can compute the estimator again, leading to B bootstrap estimate $\widehat{\theta}_n^{*(1)}, \cdots, \widehat{\theta}_n^{*(B)}$.

With the bootstrap estimators $\widehat{\theta}_n^{*(1)}, \cdots, \widehat{\theta}_n^{*(B)}$, we can access the uncertainties via various measures.

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

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

  as an estimate of $\mathsf{Var}(\widehat{\theta}_n)$.

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

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

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

$$\left[t_{\alpha/2}, t_{1-\alpha/2}\right],$$

  where t_β is the β-percentile of the bootstrap values $\widehat{\theta}_n^{*(1)}, \cdots, \widehat{\theta}_n^{*(B)}$. Formally, let $\widehat{G}(t) = \frac{1}{N}\sum_{b=1}^{B} I(\widehat{\theta}_n^{*(B)} \leq t)$ be the EDF of the bootstrap values. The percentile $t_\beta = \widehat{G}^{-1}(\beta)$. This is commonly known as the bootstrap *percentile* confidence interval.

Example: inference of median. Suppose our parameter of interest θ is the median of the distribution. Let $\widehat{\theta}_n = \mathsf{median}\{X_1, \cdots, X_n\}$ be the sample median. Recall that the bootstrap generates B samples from sampling with replacement of the original data:

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

We then compute the median of each bootstrap sample:

$$\begin{aligned} \widehat{\theta}_n^{*(1)} &= \mathsf{median}\{X_1^{*(1)}, \cdots, X_n^{*(1)}\} \\ \widehat{\theta}_n^{*(2)} &= \mathsf{median}\{X_1^{*(2)}, \cdots, X_n^{*(2)}\} \\ &\vdots \\ \widehat{\theta}_n^{*(B)} &= \mathsf{median}\{X_1^{*(B)}, \cdots, X_n^{*(B)}\}. \end{aligned}$$

A $1-\alpha$ confidence interval of θ is then $\left[t_{\alpha/2}, t_{1-\alpha/2}\right]$, where t_β is the β-percentile of the bootstrap medians $\widehat{\theta}_n^{*(1)}, \cdots, \widehat{\theta}_n^{*(B)}$. As you can see, the bootstrap procedure bypass the needs of computing the standard errors of the original median estimator (sample median).

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

12.2.1 Other variants of bootstrap confidence intervals

There are many variants of the bootstrap confidence interval. The above percentile method is the most popular approach. Here are three also popular bootstrap confidence intervals.

- **Symmetric percentile confidence interval.** We compute the deviation of bootstrap estimate from the original estimator
$$D^{*(1)}, \cdots, D^{*(B)}, \qquad D^{*(b)} = |\widehat{\theta}_n^{*(1)} - \widehat{\theta}_n|.$$
Let s_β be the β-percentile of $D^{*(1)}, \cdots, D^{*(B)}$. The confidence interval is
$$\left[\widehat{\theta}_n - s_{1-\alpha}, \widehat{\theta}_n + s_{1-\alpha}\right].$$
Unlike percentile method, this confidence interval is symmetric around the original median estimator $\widehat{\theta}_n$.

- **Normal confidence interval with bootstrap variance estimator.** Since we can estimate the variance of $\widehat{\theta}_n$ by the bootstrap method $\widehat{\mathsf{Var}}_B(\widehat{\theta}_n)$, we can easily use this to construct a confidence interval using asymptotic normality:
$$\widehat{\theta}_n \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathsf{Var}}_B(\widehat{\theta}_n)}.$$
This method is generally the most straightforward but its empirical coverage (the actual chance of covering the true median) converges to nominal coverage slower than the percentile method.

- **Studentized confidence interval.** Suppose we have a non-bootstrap method to estimate the sample standard variance $\sigma^2 = \mathsf{Var}(\widehat{\theta}_n)$, denoted it as $\widehat{\sigma}^2$. For each bootstrap sample, we not only compute the bootstrap estimator $\widehat{\theta}_n^{*(B)}$, we also use the bootstrap sample to compute the bootstrap standard error $\widehat{\sigma}^{*(B)}$. With these two quantities, we can then compute the bootstrap t-percentile:
$$T^{*(b)} = \frac{\widehat{\theta}_n^{*(B)} - \widehat{\theta}_n}{\widehat{\sigma}^{*(b)}}$$
for each $b = 1, 2, \cdots, B$. Let η_β be the β-percentile of $T^{*(1)}, \cdots, T^{*(B)}$. We then construct the confidence interval using
$$\left[\widehat{\theta}_n - \widehat{\sigma} \cdot \eta_{1-\alpha/2}, \widehat{\theta}_n - \widehat{\sigma} \cdot \eta_{\alpha/2}\right].$$
This method is called studentized because our bootstrapping quantity $T^{*(1)}, \cdots, T^{*(B)}$ is like a student t-statistic. While this procedure is more involved, the studentized confidence interval generally has a coverage that converges to the nominal coverage quickest.

12.2.2 High level idea of how bootstrap works

Let $X_1, \cdots, X_n \sim F$. Recall that a statistic $S(X_1, \cdots, X_n)$ is a function of random variables so its distribution will depend on the CDF F and the sample size n. Thus, the distribution of the estimator $\widehat{\theta}_n$, denoted as $F_{\widehat{\theta}_n}$, will also be determined by the CDF F and sample size n. Namely, we may write the CDF of the estimator as
$$F_{\widehat{\theta}_n}(x) = \Psi(x; F, n), \tag{12.1}$$
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 $X_1, \cdots, X_n$, what is the distribution we are sampling from? Let $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x)$ 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 $\widehat{F}_n$, then Z has the following probability distribution:

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

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

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

What is this sample $Z_1, \cdots, Z_n$? This sample is a sample generated by *sampling with replacement from* $X_1, \cdots, X_n$. Namely, *a bootstrap sample is an IID random sample from the EDF* $\widehat{F}_n$ *of the original data.*

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

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

Because a bootstrap estimator, say $\widehat{\theta}_n^{*(1)}$, is the estimator based on $X_1^{*(1)}, \cdots, X_n^{*(1)}$. Its CDF, by equation (12.1), is

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

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

$$\Psi(x; \widehat{F}_n, n) = F_{\widehat{\theta}_n^{*(1)}}(x) = F_{\widehat{\theta}_n^{*(2)}}(x) = \cdots = F_{\widehat{\theta}_n^{*(B)}}(x).$$

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

$$\widehat{F}_n \approx F \Longrightarrow F_{\widehat{\theta}_n^{*(\ell)}}(x) = \Psi(x; \widehat{F}_n, n) \approx \Psi(x; F, n) = F_{\widehat{\theta}_n}(x).$$

This means:

> *The CDF of a bootstrap estimator,* $F_{\widehat{\theta}_n^{*(\ell)}}(x)$*, is approximating the CDF of the original estimator,* $F_{\widehat{\theta}_n}(x)$.

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

$$\mathsf{Var}\left(\widehat{\theta}_n^{*(\ell)} | X_1, \cdots, X_n\right) \approx \mathsf{Var}(\widehat{\theta}_n).^{2}$$

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

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

$$\widehat{\mathsf{Var}}_B(\widehat{\theta}_n) = \frac{1}{B-1}\sum_{\ell=1}^{B}\left(\widehat{\theta}_n^{*(\ell)} - \bar{\theta}_B^*\right)^2 \approx \mathsf{Var}\left(\widehat{\theta}_n^{*(\ell)} | X_1, \cdots, X_n\right).$$

Therefore,

$$\widehat{\mathsf{Var}}_B(\widehat{\theta}_n) \approx \mathsf{Var}\left(\widehat{\theta}_n^{*(\ell)} | X_1, \cdots, X_n\right) \approx \mathsf{Var}(\widehat{\theta}_n),$$

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

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.

12.2.3 Other variants of bootstrap

From the above argument, we see that as long as we are sampling from a distribution G_n that is converging toward the true underlying distribution F, the bootstrap method is expected to work. Therefore, there are two common variants of the bootstrap.

- **Parametric bootstrap.** We fit a parametric model of the underlying data-generating process. And then sample from the fitted model to create our bootstrap sample.
- **Smooth bootstrap.** The smooth bootstrap generates the bootstrap sample from a nonparametric density estimator. When we use the KDE with a Gaussian kernel, $\widehat{p}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$, sampling from $\widehat{p}_h(x)$ is equivalent to

  1. Sample with replacements to obtain $\widetilde{X}_1, \cdots, \widetilde{X}_n$.
  2. Add a Gaussian noise $N(0, h^2)$ to each observation, i.e., $X_i^* = \widetilde{X}_i + \epsilon_i$, where $\epsilon_i \sim N(0, h^2)$.

  The resulting bootstrap sample from the above procedure is called smooth bootstrap.

In the modern time, we may have many powerful generative models approximating our data generating process. We can also use these generative models for creating bootstrap sample as well.

12.2.4 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 $X_1, \cdots, X_n \sim \mathsf{Uni}[0, 1]$ and $\widehat{\theta}_n = \min\{X_1, \cdots, X_n\}$ be the minimum value of the sample. Then it is known that

$$n \cdot \widehat{\theta}_n \xrightarrow{D} \mathsf{Exp}(1).$$

♠ : Think about why it converges to exponential distribution.

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

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

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* $\widehat{\theta}_n$ in the bootstrap sample is

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

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 $\widehat{\theta}_n$. Namely,

$$P(\widehat{\theta}_n^* = \widehat{\theta}_n) = 1 - e^{-1}.$$

Thus, $\widehat{\theta}_n^*$ has a huge probability mass at the value $\widehat{\theta}_n$, meaning that the distribution of $\widehat{\theta}_n^*$ will not be close to an exponential distribution.

12.3 Statistical Functionals

To study how the bootstrap works, we first introduce the concept of *statistical functionals*.

What is a functional? A functional is just a function of a function. Namely, it is a 'function' such that the input is another function and the output is a number. Formally speaking, a functional is a mapping $T : \mathcal{F} \mapsto \mathbb{R}$, where $\mathcal{F}$ is a collection of functions. A statistical functional is a mapping T such that you input a distribution (CDF) and it returns a number.

This sounds very complicated but actually, we have encountered numerous statistical functionals. Here are some examples.

- **Mean of a distribution.** The mean of a distribution is a statistical functional

  $$\mu = T_{\mathsf{mean}}(F) = \int x dF(x).$$

  When F has a PDF $p(x)$, $dF(x) = p(x)dx$ so the mean functional reduces to the form that we are familiar with:

  $$\mu = T_{\mathsf{mean}}(F) = \int x dF(x) = \int x p(x) dx.$$

  When F is a distribution of discrete random variables, we define

  $$\int x dF(x) = \sum_x x P(x) \Longrightarrow \mu = T_{\mathsf{mean}}(F) = \sum_x x P(x),$$

  where $P(x)$ is the PMF of the distribution F.

  You may have noticed that if a random variable X has a CDF F, then

  $$\mathbb{E}(X) = \int x dF(x) = T_{\mathsf{mean}}(F).$$

  Therefore, for any function g,

  $$\mathbb{E}(g(X)) = \int g(x) dF(x).$$

  Using the function g, we introduce another functional T_ω such that

  $$T_\omega(F) = \int \omega(x) dF(x).$$

  Such a functional, T_ω, is called a *linear functional*.

- **Variance of a distribution.** The variance of a distribution is also a statistical functional. Let X be a random variable with CDF F. Then

$$\sigma^2 = T_{\mathsf{var}}(F) = \mathsf{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2.$$

- **Median of a distribution.** Using the concept of a statistical functional, median and any quantile can be easily defined. The median of a distribution F is a point θ_{med} such that $F(\theta_{\mathsf{med}}) = 0.5$. Thus,

$$T_{\mathsf{med}}(F) = F^{-1}(0.5).$$

  Note that when F is a CDF of a discrete random variable, F^{-1} may have multiple values. In this case, we define

$$F^{-1}(q) = \inf\{x : F(x) \geq q\}.$$

  Any quantile of a distribution can be represented in a similar way. For instance, the q-quantile ($0 < q < 1$) will be

$$T_{\mathsf{q}}(F) = F^{-1}(q).$$

  As a result, the interquartile range (IQR) is

$$T_{\mathsf{IQR}}(F) = F^{-1}(0.75) - F^{-1}(0.25).$$

Why do we want to use the form of statistical functionals? One answer is: it elegantly describes a population quantity that we may be interested in. Recall that the statistical model about how the data is generated is that we observe a random sample $X_1, \cdots, X_n$ IID from an unknown distribution F. Thus, the distribution F is our model for the population. Because the statistical functionals map F into some real numbers, they can be viewed as quantities describing the features of the population. The mean, variance, median, quantiles of F are numbers characterizing the population. Thus, using statistical functionals, we have a more rigorous way to define the concepts of population parameters.

In addition to the above advantage, there is a very powerful features of statistical functionals–they provide a simple estimator to these population quantities. Recall that the EDF $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)$ is a good estimator of F. Thus, if we want to estimate a population quantity $\theta = T_{\mathsf{target}}(F)$, we can use $T_{\mathsf{target}}(\widehat{F}_n) = \widehat{\theta}_n$ as our estimator. Actually, many estimators do follow this form. For instance, in the case of estimating the mean $\mu = T_{\mathsf{mean}}(F)$, we often use the sample mean $\bar{X}_n$ as our estimator. However, if you plug-in $\widehat{F}_n$ into the statistical functional:

$$T_{\mathsf{mean}}(\widehat{F}_n) = \int x d\widehat{F}_n(x) = \sum_{i=1}^n X_i \frac{1}{n} = \sum_{i=1}^n \frac{X_i}{n} = \bar{X}_n.$$

This implies that the estimator from the statistical functional is the same as sample mean! Note that we in the above calculation, we use the fact that $\widehat{F}_n(x)$ is a distribution with whose PMF puts equal probability $(1/n)$ at $X_1, \cdots, X_n$. The estimator formed via replacing F by $\widehat{F}_n$ is called a *plug-in* estimator.

Similarly, we may estimate the variance $\sigma^2 = T_{\mathsf{var}}(F)$ via

$$T_{\mathsf{var}}(\widehat{F}_n) = \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n-1}{n} S_n^2.$$

This estimator is very similar to the sample variance S_n^2 (they are asymptotically the same).

Using how we define the inverse of a CDF of a discrete random variable, we can define the estimator of median

$$T_{\mathsf{med}}(\widehat{F}_n) = \widehat{F}_n^{-1}(0.5)$$

and other quantiles of a distribution. And it turns out that this estimator is the sample median (and the corresponding sample quantiles)!

Therefore, the statistical functional provides an elegant way to define a population quantities as well as an estimator. And the plug-in estimator will be a good estimator if the statistical functional $T(\cdot)$ is 'smooth' with respect to the input function because we know that $\widehat{F}_n \to F$ in various ways so that the smoothness of T with respect the input will implies $T(\widehat{F}_n) \to T(F)$[3].

12.4 Bootstrap and Statistical Functionals

We have learned that the (empirical) bootstrap sample is a new random sample from the EDF $\widehat{F}_n$. The bootstrap sample forms another EDF called the bootstrap EDF, denoted as $\widehat{F}_n^*$. Namely, let $X_1^*, \cdots, X_n^*$ be a bootstrap sample. Then the bootstrap EDF is

$$\widehat{F}_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i^* \le x).$$

Here is how the statistical functionals and the bootstrap is connected. In estimating the parameter $\theta = T_{\mathsf{target}}(F)$, we often use a plug-in estimate from the EDF $\widehat{\theta}_n = T_{\mathsf{target}}(\widehat{F}_n)$ (just think of how we estimate the sample mean). In this case, the bootstrap estimator, the estimator using the bootstrap sample, will be

$$\widehat{\theta}_n^* = T_{\mathsf{target}}(\widehat{F}_n^*),$$

another plug-in estimator but now we are plugging in the bootstrap EDF $\widehat{F}_n^*$.

Consistency of bootstrap variance estimator. How do we use the bootstrap to estimate the variance and construct a confidence interval? We keep generating bootstrap samples from the EDF $\widehat{F}_n$ and obtain several realizations of $\widehat{\theta}_n^*$'s. Namely, we generate

$$\widehat{\theta}_n^{*(1)}, \cdots, \widehat{\theta}_n^{*(B)}$$

and use their sample variance, $\widehat{\mathsf{Var}}_B(\widehat{\theta}_n^*)$, as an estimator of $\mathsf{Var}(\widehat{\theta}_n)$. Note that $\widehat{\mathsf{Var}}_B(\widehat{\theta}_n^*)$ is

$$\widehat{\mathsf{Var}}_B(\widehat{\theta}_n^*) = \frac{1}{B-1}\sum_{\ell=1}^{N}\left(\widehat{\theta}_n^{*(\ell)} - \bar{\widehat{\theta}}_{n,B}^*\right), \quad \bar{\widehat{\theta}}_{n,B}^* = \frac{1}{B}\sum_{\ell=1}^{B}\widehat{\theta}_n^{*(\ell)}.$$

When B is large, the sample variance of the bootstrap estimators

$$\widehat{\mathsf{Var}}_B(\widehat{\theta}_n^*) \approx \mathsf{Var}(\widehat{\theta}_n^* | \widehat{F}_n). \tag{12.2}$$

Note that $\cdot|\widehat{F}_n$ means *conditioned* on $\widehat{F}_n$ being fixed. The reason why here it converges to this conditioned variance is because when we generate bootstrap samples, the original EDF $\widehat{F}_n$ is fixed (and we are generating from it). Thus, the variance is conditioned on $\widehat{F}_n$ being fixed.

To argue that the bootstrap variance $\widehat{\mathsf{Var}}_B(\widehat{\theta}_n^*)$ is a good estimate of the original variance, we need to argue

$$\widehat{\mathsf{Var}}_B(\widehat{\theta}_n^*) \approx \mathsf{Var}(\widehat{\theta}_n^* | \widehat{F}_n) \approx \mathsf{Var}(\widehat{\theta}_n).$$

[3] Note that here we ignore lots of technical details. The smoothness of a 'functional' is an advanced topic in mathematics called *functional analysis*: https://en.wikipedia.org/wiki/Functional_analysis. There are formal ways of defining continuity of functionals and even 'differentiation' of functionals; see, e.g., https://en.wikipedia.org/wiki/G%C3%A2teaux_derivative.

However, because of equation (12.2) and we can select B as large as we wish, so what really matters is

$$\mathsf{Var}(\widehat{\theta}_n^*|\widehat{F}_n) \approx \mathsf{Var}(\widehat{\theta}_n).$$

Or more formally,

$$\frac{\mathsf{Var}(\widehat{\theta}_n^*|\widehat{F}_n)}{\mathsf{Var}(\widehat{\theta}_n)} \approx 1 \tag{12.3}$$

(people generally use the ratio expression because both quantities often converge to 0 when the sample size $n \to \infty$).

Therefore, we conclude that

> *as long as we can show that equation (12.3) holds, the bootstrap variance is a good estimate of the variance of the estimator $\widehat{\theta}_n$.*

Because $\widehat{\theta}_n = T_{\mathsf{target}}(\widehat{F}_n)$ is a statistic (a function of our random sample $X_1, \cdots, X_n$), its distribution is completely determined by the distribution $X_1, \cdots, X_n$ are sampling from, which is F, and the sample size n. This implies that the variance of $\widehat{\theta}_n$ is determined by F and n as well. Therefore, we can write

$$\mathsf{Var}(\widehat{\theta}_n) = \mathsf{Var}(T_{\mathsf{target}}(\widehat{F}_n)) = \mathbb{V}_{n,\mathsf{target}}(F).$$

And it turns out that we often have

$$\mathbb{V}_{n,\mathsf{target}}(F) \approx \frac{1}{n}\mathbb{V}_{1,\mathsf{target}}(F) \equiv \frac{1}{n}\mathbb{V}_{\mathsf{target}}(F).$$

Note that here $\mathbb{V}_{n,\mathsf{target}}(\cdot), \mathbb{V}_{\mathsf{targe}}(\cdot)$ are both again statistical functionals!

Because the bootstrap estimator $\widehat{\theta}_n^* = T_{\mathsf{target}}(\widehat{F}_n^*)$, its conditional variance will be

$$\mathsf{Var}(\widehat{\theta}_n^*|\widehat{F}_n) = \mathsf{Var}(T_{\mathsf{target}}(\widehat{F}_n^*)|\widehat{F}_n) = \mathbb{V}_{n,\mathsf{target}}(\widehat{F}_n) \approx \frac{1}{n}\mathbb{V}_{\mathsf{target}}(\widehat{F}_n).$$

Thus, as long as

$$\mathbb{V}_{\mathsf{target}}(\widehat{F}_n) \approx \mathbb{V}_{\mathsf{target}}(F), \tag{12.4}$$

equation (12.3) holds. Namely, the bootstrap variance estimate will be a good estimator of the variance of the true estimator[4].

Validity of bootstrap confidence interval. How about the validity of the bootstrap confidence interval? Here is a derivation showing that the consistency of bootstrap variance estimator implies the validity of bootstrap confidence interval.

For the bootstrap confidence interval, a simple way is first show that

$$\sqrt{n}(\widehat{\theta}_n - \theta) = \sqrt{n}\left(T_{\mathsf{target}}(\widehat{F}_n) - T_{\mathsf{target}}(F)\right) \approx N(0, \mathbb{V}_{\mathsf{target}}(F)) \tag{12.5}$$

which implies

$$\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) = \sqrt{n}\left(T_{\mathsf{target}}(\widehat{F^*}_n) - T_{\mathsf{target}}(\widehat{F}_n)\right) \approx N(0, \mathbb{V}_{\mathsf{target}}(\widehat{F}_n)).$$

Thus, as long as the bootstrap variance converges, we also have the convergence of the entire distribution, implying the validity of a bootstrap confidence interval. Note that to formally prove this, we need to show the

[4] A more formal way is to show that it converges in probability.

convergence in terms of CDF of the difference. In more details, let $Z_n = \sqrt{n}(\widehat{\theta}_n - \theta)$ and $Z_n^* = \sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)$. We need to prove

$$\sup_t \left| P(Z_n^* \le t | \widehat{F}_n) - P(Z_n \le t) \right| \xrightarrow{P} 0.$$

Later we will show examples about this using the sample mean as a starting point.

Example: mean. We now consider a simple example: the mean of a distribution $T_{\mathsf{target}} = T_{\mathsf{mean}}$. The mean of a distribution has the form

$$\mu = T_{\mathsf{mean}}(F) = \int x dF(x).$$

The plug-in estimator is

$$\widehat{\mu}_n = T_{\mathsf{mean}}(\widehat{F}_n) = \int x d\widehat{F}_n(x) = \bar{X}_n$$

and the bootstrap estimator is

$$\widehat{\mu}_n^* = T_{\mathsf{mean}}(\widehat{F}_n^*) = \int x d\widehat{F}_n^*(x) = \bar{X}_n^*.$$

It is clear from the Central Limit Theorem that

$$\sqrt{n}(\widehat{\mu}_n - \mu) \approx N(0, n\mathsf{Var}(T_{\mathsf{mean}}(\widehat{F}_n)))$$

so equation (12.5) holds and

$$\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) \approx N(0, n\mathsf{Var}(T_{\mathsf{mean}}(\widehat{F}_n^*)|\widehat{F}_n)).$$

In this case, we know that

$$\mathsf{Var}(T_{\mathsf{mean}}(\widehat{F}_n)) = \mathsf{Var}(\bar{X}_n) = \frac{1}{n}\mathsf{Var}(X_i) \Longrightarrow \mathbb{V}_{\mathsf{mean}}(F) = \mathsf{Var}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}^2(X_i) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2.$$

Therefore, the bootstrap variance is

$$\mathsf{Var}(T_{\mathsf{mean}}(\widehat{F}_n^*)|\widehat{F}_n) = \frac{1}{n}\mathbb{V}_{\mathsf{mean}}(\widehat{F}_n) = \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2.$$

Because of the Law of Large Number,

$$\int x^2 d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} \mathbb{E}(X_i^2) = \int x^2 dP(x)$$

$$\int x d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}(X_i) = \int x dP(x).$$

Thus, [5]

$$\mathbb{V}_{\mathsf{mean}}(\widehat{F}_n) \xrightarrow{P} \mathbb{V}_{\mathsf{mean}}(F),$$

which shows that equation (12.4) holds and so is equation (12.3). Thus, the bootstrap variance estimator converges to the true variance estimator and we conclude that

$$\frac{\mathsf{Var}(T_{\mathsf{mean}}(\widehat{F}_n^*)|\widehat{F}_n)}{\mathsf{Var}(T_{\mathsf{mean}}(\widehat{F}_n))} \xrightarrow{P} 1.$$

As a result, the bootstrap variance estimator is consistent and the bootstrap confidence interval is also valid.

[5] Note that here we use the continuous mapping theorem: if f is a continuous function and random variable $A_n \xrightarrow{P} a_0$, then $f(A_n) \xrightarrow{P} f(a_0)$. Setting $f(x) = x^2$, we obtain the convergence of the second quantity.

12.5 Consistency of Bootstrap

The analysis in the previous section is a high-level sketch of the consistency of bootstrap. In this section, we provide a formal way to derive the consistency of bootstrap.

Let $Z_n = \sqrt{n}(\widehat{\theta}_n - \theta)$ and $Z_n^* = \sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)$. To formally prove the validity bootstrap, we need to prove that

$$\sup_t \left| P(Z_n^* \le t | \widehat{F}_n) - P(Z_n \le t) \right| \xrightarrow{P} 0. \tag{12.6}$$

The above bound is also known as the Kolomogrov distance between two random variables.

Although this seems to be hard to prove, there are two popular approaches to derive equation (12.6). The first approach is to show that Z_n has an asymptotic linear form and then apply Lindeberg-Feller central limit theorem (triangular arrays) to Z_n^* since Z_n^* is sampled from a 'random distribution function' $\widehat{F}_n$. The second approach is via the Berry-Esseen bound of the sample mean, which is our preferred route.

12.5.1 Lindeberg-Feller's central limit theorem

In the plug-in estimate of statistical functionals, we see that many estimators can be written as a sample mean problem. Therefore, here we consider a simple scenario that we observe univariate $X_1, \cdots, X_n$ and we are interested in estimating the population mean, i.e., $\theta = \mathbb{E}(X_1)$, using the sample mean $\widehat{\theta}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n$.

The conventional central limit theorem (CLT) shows that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2),$$

where $\sigma^2 = \mathsf{Var}(X_1)$. This result is NOT enough for bootstrap consistency because it assumes that observations $X_1, \cdots, X_n$ are sampled from a fixed CDF F, not a distribution function that can change with respect to the sample size n. In the bootstrap case, the bootstrap sample $X_1^*, \cdots, X_n^*$ are IID from the EDF $\widehat{F}_n$ given $X_1, \cdots, X_n$. Thus, the 'population' of the bootstrap sample changes with respect to n, so conventional CLT is not applicable.

To resolve this issue, we use the Lindeberg-Feller's CLT (also known as triangular array CLT).

Theorem 12.1 (Lindeberg-Feller) *For each $n = 1, 2, 3, \cdots$, let $W_n = (W_{n,1}, \cdots, W_{n,k_n})$ be a vector of independent elements with finite variance, i.e., $W_{n,1}, \cdots, W_{n,k_n}$ are independent from each other. Assume that*

- **Uniform integrability.** $\sum_{i=1}^{k_n} \mathbb{E}[W_{n,i}^2 I(|W_{n,i}| > \epsilon)] \to 0$ *for every* $\epsilon > 0$.
- **Finite variance.** $\sum_{i=1}^{k_n} \mathsf{Var}(W_{n,i}) \to \sigma^2$.

Then

$$\sum_{i=1}^{k_n} (W_{n,i} - \mathbb{E}(W_{n,i})) \xrightarrow{d} N(0, \sigma^2).$$

We consider the scenario where the original data $X_1, \cdots, X_n$ are fixed so that the bootstrap sample $X_1^*, \cdots, X_n^*$ are IID from $\widehat{F}_n$.

To use Theorem 12.1 in the bootstrap setting, each $W_{n,i} = \frac{1}{\sqrt{n}} X_i^*$ is sampled from $\widehat{F}_n$, so $\mathbb{E}_{\widehat{F}_n}[W_{n,i}] = \frac{1}{\sqrt{n}}\widehat{\theta}_n$. Note that the expectation $\mathbb{E}_{\widehat{F}_n}(\cdot)$ is with respect to the distribution $\widehat{F}_n$. Also, $k_n = n$. Under this setting, $\sum_{i=1}^{k_n} W_{n,i} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i^* = \sqrt{n}\widehat{\theta}_n^*$ and

$$\sum_{i=1}^{k_n} (W_{n,i} - \mathbb{E}_{\widehat{F}_n}(W_{n,i})) = \sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n).$$

Thus, the conclusion of Theorem 12.1 is applicable to the setting of the bootstrap.

Now we investigate the two conditions in Theorem 12.1. The first uniform integrability condition

$$\sum_{i=1}^{k_n} \mathbb{E}_{\widehat{F}_n}[W_{n,i}^2 I(|W_{n,i}| > \epsilon)] \to 0$$

becomes

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 I(|X_i| > \sqrt{n}\epsilon) \to 0,$$

When the true distribution F has a finite second moment, i.e., $\mathbb{E}(X_i^2) < \infty$, strong law of large numbers implies $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \overset{a.s.}{\to} \mathbb{E}(X_i^2) < \infty$, so $\frac{1}{n}\sum_{i=1}^{n} X_i^2 I(|X_i| > \sqrt{n}\epsilon) \overset{a.s.}{\to} 0$.

The finite variance condition becomes

$$\sum_{i=1}^{k_n} \mathsf{Var}(W_{n,i}) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{Var}_{\widehat{F}_n}(X_i^*) = \widehat{\sigma}_n^* \overset{P}{\to} \sigma^2,$$

where $\widehat{\sigma}_n^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ and $\sigma^2 = \mathsf{Var}(X_1)$.

As a result, we conclude that

$$\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \overset{d}{\to} N(0, \sigma^2);$$

namely, it converges to the same limit as the original estimator $\sqrt{n}(\widehat{\theta}_n - \theta) \overset{d}{\to} N(0, \sigma^2)$. Thus, the equation (12.6) holds for bootstrapping $\widehat{\theta}_n$.

A more general form of this approach can be found in Theorem 23.4 of [Van der Vaart (2000)].

12.5.2 Berry-Esseen bound

Consider again the simple scenario that we observe univariate $X_1, \cdots, X_n$ and we are interested in estimating the population mean, i.e., $\theta = \mathbb{E}(X_1)$.

Theorem 12.2 (Berry-Esseen bound) *Assume that* $\mathbb{E}(|X_1|^3) < \infty$. *Let* $Z \sim N(0,1)$ *and* $\theta = \mathbb{E}(X_1)$ *and* $\sigma^2 = \mathsf{Var}(X_1)$. *Then for any* n, *we have*

$$\sup_t \left| P\left(\sqrt{n}\left(\frac{\bar{X}_n - \theta}{\sigma}\right) < t\right) - P(Z < t) \right| \le C \frac{\mathbb{E}|X_1|^3}{\sigma^3 \sqrt{n}},$$

for a constant $C \ge \frac{\sqrt{10}+3}{6\sqrt{2\pi}}$.

The Berry-Esseen bound quantifies how fast the limiting distribution converges to a Gaussian and the result is uniform across different quantiles.

Finite Sample bound and the bootstrap. It is important to note that the Berry-Esseen bound is a *finite sample* bound, meaning that its result holds *for any* n (some finite sample bound holds when n is larger than some constant). So it is a much stronger result than the conventional central limit theorem. The finite sample bound is important in deriving the validity of the bootstrap (see the proof below).

The Berry-Esseen bound can be used to derive bounds like equation (12.6). Now consider very simple scenario that we are interested in estimating the population mean $\theta = \mathbb{E}(X_1)$ and we use the sample mean as the estimator $\widehat{\theta}_n$.

Theorem 12.3 *Suppose that we are considering the sample mean problem, i.e.,* $\theta = \mathbb{E}(X_1)$ *and* $\widehat{\theta}_n = \bar{X}_n$ *is the original sample mean estimator and* $\widehat{\theta}^*_n = \bar{X}^*_n$ *is the sample mean of the bootstrap sample. Assume that* $\mathbb{E}(|X_1|^3) < \infty$. *Let*

$$Z_n = \sqrt{n}(\widehat{\theta}_n - \theta), \qquad Z^*_n = \sqrt{n}(\widehat{\theta}^*_n - \widehat{\theta}_n).$$

Then

$$\sup_t \left| P(Z^*_n \le t | \widehat{F}_n) - P(Z_n \le t) \right| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Proof:

Let $\Psi_\sigma(t)$ be the CDF of $N(0, \sigma^2)$ and $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. We bound the difference using

$$\sup_t \left| P(Z^*_n \le t | \widehat{F}_n) - P(Z_n \le t) \right| \le \sup_t \left| P(Z^*_n \le t | \widehat{F}_n) - \Psi_{\widehat{\sigma}}(t) \right| + \sup_t |\Psi_{\widehat{\sigma}}(t)) - \Psi_\sigma(t)| + \sup_t |P(Z_n \le t) - \Psi_\sigma(t)|.$$

The Berry Esseen theorem implies that

$$\sup_t |P(Z_n \le t) - \Psi_\sigma(t)| = O_P\left(\frac{1}{\sqrt{n}}\right)$$

so the third quantity is bounded. Similarly, we can apply the Berry-Esseen bound to the first quantity by replacing $\mathbb{E}(\cdot)$ with the empirical version of it (sample average operation), which implies

$$\sup_t \left| P(Z^*_n \le t | \widehat{F}_n) - P(Z_n \le t) \right| \le C \frac{\frac{1}{n}\sum_{i=1}^n X_i^3}{\sigma^3 \sqrt{n}}.$$

Note that we can apply the Berry-Esseen theory to the bootstrap because this theory holds in *finite sample*! In the bootstrap world, the EDF is the population distribution generating our data, and that is why we replace the expectation $\mathbb{E}$ by the empirical version of it.

By strong law of large number, the probability that the right hand side is less than $2C\frac{\mathbb{E}|X_1|^3}{\sigma^3\sqrt{n}}$ is 1. Thus, we conclude that

$$\sup_t \left| P(Z^*_n \le t | \widehat{F}_n) - P(Z_n \le t) \right| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

For the second term, $\sup_t |\Psi_{\widehat{\sigma}}(t)) - \Psi_\sigma(t)|$, because $|\widehat{\sigma} - \sigma| = O_P\left(\frac{1}{\sqrt{n}}\right)$ so differentiating the CDF with respect to σ and take a uniform bound leads to

$$\sup_t |\Psi_{\widehat{\sigma}}(t)) - \Psi_\sigma(t)| = O_P\left(\frac{1}{\sqrt{n}}\right),$$

which completes the proof.

■

The Lindeberg-Feller central limit theorem approach requires a slightly less condition than the Berry-Esseen bound (we do not need third-moment but just need a bounded second moment). However, the Lindeberg-Feller approach will not give us a convergence rate while the Berry-Esseen approach gives us a convergence rate.

12.6 Delta Method

In this section, we will talk about a very useful technique in handling the convergence–the *delta method*. We start with an example of proving consistency theorem of some bootstrap estimates.

Example: inverse of mean. Assume we are interested in the inverse of the population mean. Namely, the statistical functional we will be using is

$$T_{\mathsf{inv}}(F) = \frac{1}{\int x dF(x)} = \lambda.$$

This statistical functional was implicitly used when we the MLE of the rate parameter of an exponential distribution. The plug-in estimator (as well as the MLE of estimating an exponential model) is

$$\widehat{\lambda}_n = T_{\mathsf{inv}}(\widehat{F}_n) = \frac{1}{\int x d\widehat{F}_n(x)} = \frac{1}{\bar{X}_n}.$$

The bootstrap estimator is

$$\widehat{\lambda}^*_n = T_{\mathsf{inv}}(\widehat{F}^*_n) = \frac{1}{\int x d\widehat{F}^*_n(x)} = \frac{1}{\bar{X}^*_n}.$$

In the lab session, we have shown that this estimator follows asymptotically a normal distribution. But how do we show this? and how do we compute the variance of the estimator $\widehat{\lambda}_n$? Here is how the delta method will help us.

The Delta Method

Assume that we have a sequence of random variables $Y_1, \cdots, Y_n \cdots$ such that

$$\sqrt{n}(Y_n - y_0) \xrightarrow{D} N(0, \sigma_Y^2) \tag{12.7}$$

for some constants y_0 and σ_Y^2. Note that this implies that $\mathsf{Var}(Y_n) = \sigma_Y^2$. If a function f is differentiable at y_0, then using the Taylor expansion,

$$\sqrt{n}\left(f(Y_n) - f(y_0)\right) \approx \sqrt{n} f'(y_0) \cdot (Y_n - y_0) = f'(y_0)\sqrt{n}\left(Y_n - y_0\right).$$

Notice that $f'(y_0)$ is just a constant. Thus, this implies

$$\sqrt{n}\left(f(Y_n) - f(y_0)\right) \xrightarrow{d} N(0, |f'(y_0)|^2\sigma_Y^2), \quad \mathsf{Var}(f(Y_n)) = \frac{1}{n}|f'(y_0)|^2\sigma_Y^2 + o(n^{-1}). \tag{12.8}$$

Now using equation (12.8) and identifying Y_n as $\bar{X}_n$ and $f(x)$ as $\frac{1}{x}$, we obtain

$$\sqrt{n}(\widehat{\lambda}_n - \lambda) = \sqrt{n}\left(\frac{1}{\bar{X}_n} - \frac{1}{\mathbb{E}(X_i)}\right) \approx -\frac{1}{\mathbb{E}^2(X_i)}\sqrt{n}\left(\bar{X}_n - \mathbb{E}(X_i)\right) \approx N\left(0, \underbrace{\frac{1}{\mathbb{E}^4(X_i)}\mathsf{Var}(X_i)}_{=\mathbb{V}_{\mathsf{inv}}(F)}\right).$$

Using the fact that $\mathbb{E}(X_i) = \int x dF(x)$ and $\mathsf{Var}(X_i) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$, we obtain

$$\sqrt{n}(\widehat{\lambda}_n - \lambda) \approx N(0, \mathbb{V}_{\mathsf{inv}}(F)),$$

where

$$\mathbb{V}_{\mathsf{inv}}(F) = \frac{\int x^2 dF(x) - \left(\int x dF(x)\right)^2}{\left(\int x dF(x)\right)^4}.$$

So equation (12.5) holds and

$$\sqrt{n}(\widehat{\lambda}^*_n - \widehat{\lambda}_n) \approx N(0, \mathbb{V}_{\mathsf{inv}}(\widehat{F}_n)),$$

where

$$\mathbb{V}_{\mathsf{inv}}(\widehat{F}_n) = \frac{\int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2}{\left(\int x d\widehat{F}_n(x)\right)^4}$$

is the corresponding bootstrap variance component.

As long as $\int x dF(x) \neq 0$, each component in $\mathbb{V}_{\mathsf{inv}}(\widehat{F}_n)$ is a natural estimator of the corresponding component in $\mathbb{V}_{\mathsf{inv}}(F)$. Therefore, we conclude

$$\mathbb{V}_{\mathsf{inv}}(\widehat{F}_n) \xrightarrow{P} \mathbb{V}_{\mathsf{inv}}(F),$$

which shows that equation (12.4) holds, implying that the bootstrap variance estimator is consistent:

$$\frac{\mathsf{Var}(T_{\mathsf{inv}}(\widehat{F}^*_n)|\widehat{F}_n)}{\mathsf{Var}(T_{\mathsf{inv}}(\widehat{F}_n))} \xrightarrow{P} 1$$

and moreover, the bootstrap confidence interval is also valid.

12.7 Influence Function

12.7.1 Linear Functional

In the above derivations, we see many examples of statistical functionals that are of the form

$$T_\omega(F) = \int \omega(x) dF(x),$$

where g is a function. As we have mentioned, this type of statistical functionals are called *linear* functionals.

Linear functionals has a feature that the estimators

$$T_\omega(\widehat{F}_n) = \int \omega(x) d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n g(X_i),$$

$$T_\omega(\widehat{F}^*_n) = \int \omega(x) d\widehat{F}^*_n(x) = \frac{1}{n}\sum_{i=1}^n g(X^*_i).$$

Moreover, a powerful feature of the linear functional is that for another CDF G, we always have

$$\begin{aligned} T_\omega(G) - T_\omega(F) &= \int \omega(x) dG(x) - T_\omega(F) \\ &= \int \omega(x) dG(x) - \int T_\omega(F) dG(x) \\ &= \int L_F(x) dG(x), \end{aligned}$$

where

$$L_F(x) = \omega(x) - T_\omega(F) \tag{12.9}$$

is called the *influence function* of the functional T_ω.

Theorem 12.4 *Suppose that T_ω is a linear functional with an influence function $L_F(x)$ define in equation (12.9) and $\int \omega^2(x) dF(x) < \infty$. Then*

$$\sqrt{n}\left(T_\omega(\widehat{F}_n) - T_\omega(F)\right) \xrightarrow{D} N\left(0, \mathbb{V}_\omega(F) = \int L_F^2(x) dF(x)\right)$$

and a consistent estimator of $\mathbb{V}_\omega(F)$ is $\mathbb{V}_\omega(\widehat{F}_n) = \frac{1}{n}\sum_{i=1}^n L_F^2(X_i)$.

As a result, the bootstrap always works for the linear functional whenever $T_{\omega^2}(F) < \infty$.

Proof:

It is easy to see that

$$T_\omega(\widehat{F}_n) - T_\omega(F) = \int L_F(x) d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n L_F(X_i).$$

Moreover,

$$\mathbb{E}(L_F(X_i)) = \int L_F(x) dF(x) = \int \left(\omega(x) - T_\omega(F)\right) dF(x) = T_\omega(F) - T_\omega(F) = 0.$$

Thus, by central limit theorem,

$$\sqrt{n}\left(T_\omega(\widehat{F}_n) - T_\omega(F)\right) \xrightarrow{D} N\left(0, \mathbb{V}_\omega(F) = \int L_F^2(x) dF(x)\right)$$

Note that for a linear functional T_ω, equation (12.5) always holds with

$$\mathbb{V}_\omega(F) = \int L_F^2(x) dF(x) = \int \left(\omega^2(x) - 2\omega(x) T_\omega(F) - T_\omega^2(F)\right) dF(x) = \int \omega^2(x) dF(x) = T_{\omega^2}(F). \tag{12.10}$$

Moreover,

$$\begin{aligned} \mathbb{V}_\omega(\widehat{F}_n) &= \int L_{\widehat{F}_n}^2(x) d\widehat{F}_n(x) = \int \left(\omega^2(x) - 2\omega(x) T_\omega(\widehat{F}_n) + T_\omega^2(\widehat{F}_n)\right) d\widehat{F}_n(x) \\ &= \int \omega^2(x) d\widehat{F}_n(x) - T_\omega^2(\widehat{F}_n). \end{aligned} \tag{12.11}$$

By Law of Large Number (and continuous mapping theorem),

$$T_\omega^2(\widehat{F}_n) \xrightarrow{P} T_\omega^2(F)$$

if $\mathbb{E}(|\omega(X_i)|) = T_{|\omega|} < \infty$. And

$$\int \omega^2(x) d\widehat{F}_n(x) = T_{\omega^2}(\widehat{F}_n) \xrightarrow{P} T_{\omega^2}(F)$$

if $\mathbb{E}(\omega(X_i)^2) = T_{\omega^2}(F) < \infty$. Therefore, we conclude that when $T_{\omega^2}(F) < \infty$,

$$\mathbb{V}_\omega(\widehat{F}_n) = \int \omega^2(x) d\widehat{F}_n(x) - T^2_\omega(\widehat{F}_n) \xrightarrow{P} \mathbb{V}_\omega(F) = \mathbb{V}_\omega(F),$$

implying that the equation (12.4) holds.

■

12.7.2 Non-linear Functional

Although the linear functional has so many beautiful properties, many statistical functionals are not linear. For instance, the median

$$T_{\mathsf{med}}(F) = F^{-1}(0.5)$$

is not a linear functional. Therefore, our results of linear functional cannot be directly applied to analyze the median.

Then how can we analyze the properties of non-linear statistical functionals? One way to proceed is to generalize the notion of influence function. And here is the formal definition of the influence function.

Let δ_x be a point mass at location x. The *influence function* of a (general) statistical function T_{target} is

$$L_F(x) = \lim_{\epsilon \to 0} \frac{T_{\mathsf{target}}((1-\epsilon)F + \epsilon\delta_x) - T_{\mathsf{target}}(F)}{\epsilon}. \tag{12.12}$$

Some of you may find equation (12.12) very familiar; it seems to be taking a derivative. And yes – it is a derivative of a functional with respect to a function. This type of derivative is called *Gâteaux derivative*[6], a type of derivative of functionals. You can check that applying equation(12.12) to a linear functional leads to an influence function as we defined previously.

A powerful feature of this generalized version of influence function is that when the statistical functional T_{target} is 'smooth[7]', equation (12.10) and (12.11) hold in the sense that

$$\mathbb{V}_{\mathsf{target}}(F) = \int L^2_F(x) dF(x), \quad \mathbb{V}_{\mathsf{target}}(\widehat{F}_n) = \int L^2_{\widehat{F}_n}(x) d\widehat{F}_n(x) \tag{12.13}$$

and, moreover, equation (12.5) holds. Note that $L_{\widehat{F}_n}(x)$ is defined via replacing F by $\widehat{F}_n$ in equation (12.12). That is, when the statistical functional T_{target} is smooth, we only need to verify

$$\int L^2_{\widehat{F}_n}(x) d\widehat{F}_n(x) \approx \int L^2_F(x) dF(x) \tag{12.14}$$

to argue the validity of bootstrap consistency.

Example: median. Why median follows a normal distribution? Here we will show this using the influence function. The influence function of the functional T_{med} is

$$L_F(x) = \frac{1}{2p(F^{-1}(0.5))},$$

[6] https://en.wikipedia.org/wiki/G%C3%A2teaux_derivative.

[7] More precisely, we need it to be Hadamard differentiable with respect to the L_∞ metric $d(F,G) = \sup_x |F(x) - G(x)|$; see https://en.wikipedia.org/wiki/Hadamard_derivative

where p is the PDF of F (you can verify it). Thus, equation (12.5) implies

$$\sqrt{n}\left(\underbrace{T_{\mathsf{med}}(\widehat{F}_n)}_{\text{sample median}} - \underbrace{T_{\mathsf{med}}(F)}_{\text{population median}}\right) \approx N\left(0, \frac{1}{4p^2(F^{-1}(0.5))}\right).$$

Note that $F^{-1}(0.5) = T_{\mathsf{med}}(F)$ is the median of F. So this shows not only the asymptotic normality of sample median but also its limiting variance, which is inversely related to the PDF at the median.

The influence function is also related to the robustness of an estimator[8] and plays a key role in the semi-parametric statistics[9]. You would encounter it several times if you want to pursue a career in statistics.

12.8 Functional Delta Method

The Berry-Esseen theory shows that we can use the bootstrap to a sample mean problem and the delta method further implies that the bootstrap is applicable to any statistical functional that can be written as a smoothed function of a sample mean.

However, in many scenarios such as the sample median, the regular delta method does not work but the analysis using influence function shows that the bootstrap is still applicable. This motivates us to generalize the regular delta method. It turns out that there is a much wider classes of statistical functionals that the bootstrap method works. Here we will introduce a new technique called *functional delta method* that generalizes the regular delta method to a much wider class of statistical functionals.

Before we formally explain the functional delta method, we first introduce the concept of *Hadamard differentiation*. Let $\mathbb{D}, \mathbb{F}$ be two normed spaces (you can think of function spaces) and let $\phi : \mathbb{D} \mapsto \mathbb{F}$ be a mapping. ϕ is said to be Hadamard differentiable at $\omega \in \mathbb{D}$ with a differentiation $\dot{\phi}_\omega$ if for any sequence $\eta_t \to \eta$ when $t \to 0$,

$$\lim_{t\to 0} \left\| \frac{\phi(\omega + t\cdot\eta_t) - \phi(\omega)}{t} - \dot{\phi}_\omega(\eta) \right\|_{\mathbb{F}} = 0,$$

where $\|\cdot\|_{\mathbb{F}}$ is the norm of the space $\mathbb{F}$.

The followings are two informal statements about the functional delta method, which are simplified from

- Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. New York: Springer.
- Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer.

Theorem 12.5 (Functional delta method) *Let $X_1, \cdots, X_n \sim F_0$ and $\widehat{F}_n$ be the EDF and $\theta(F_0)$ be the parameter of interest. If θ is Hadamard differentiable at F_0 with respect to the L_∞ norm, then*

$$\sqrt{n}(\theta(\widehat{F}_n) - \theta(F_0)) \xrightarrow{D} \dot{\theta}_{F_0}(\mathbb{B}),$$

where $\mathbb{B}$ is a Gaussian process defined over $\mathbb{R}^d$ with $\mathsf{Cov}(\mathbb{B}(x), \mathbb{B}(y)) = \sqrt{n}\mathsf{Cov}(\widehat{F}_n(x), \widehat{F}_n(y))$.

[8] https://en.wikipedia.org/wiki/Robust_statistics#Influence_function_and_sensitivity_curve
[9] https://en.wikipedia.org/wiki/Semiparametric_model

Theorem 12.5 shows that after rescaling, the difference between the estimator and the parameter of interest converges to a stable distribution (often is also a normal distribution). A formal statement of the above result can be found in Theorem 2.8 of Kosorok (2008). Moreover, we have the following result about the bootstrap.

Theorem 12.6 (Bootstrap Functional delta method) *Under the assumption of Theorem 12.5, then*

$$\sqrt{n}(\theta(\widehat{F}_n^*) - \theta(\widehat{F}_n)) \approx \dot{\theta}_{F_0}(\mathbb{B}),$$

where $\dot{\theta}_{F_0}(\mathbb{B})$ is the same as the one in Theorem 12.5.

Theorem 12.6 shows that as long as the functional delta method works, the bootstrap method also works. The formal statement of Theorem 12.6 can be found in Theorem 2.9 of Kosorok (2008). Formally, we say that the bootstrap method is consistent if the random variable $\sqrt{n}(\theta(\widehat{F}_n^*) - \theta(\widehat{F}_n))$ converges in the Kolmogorov distance to $\dot{\theta}_{F_0}(\mathbb{B})$, i.e.,

$$\sup_t \left| P\left(\sqrt{n}(\theta(\widehat{F}_n^*) - \theta(\widehat{F}_n)) < t | X_1, \cdots, X_n\right) - P\left(\dot{\theta}_{F_0}(\mathbb{B}) < t\right) \right| = o_P(1). \tag{12.15}$$

The Berry-Esseen bound is a common approach to establish the consistency of a bootstrap method.

Remark (higher-order accuracy). There are many ways of constructing a bootstrap confidence interval. So one may be wondering which method give the best confidence interval. Here is a common way of measuring how good a bootstrap confidence interval is using the idea of accuracy. In a bound like the one in equation (12.15), the $o_P(1)$ can often be explicitly written as $O_P(n^{-k})$. The Berry-Esseen bound gives an accuracy at rate $k = 1/2$ and the percentile method also give the same rate. However, the bootstrap t-percentile method may lead to $k = 1$. When the limit is a Gaussian distribution, a common approach of finding k is via the *Edgeworth expansion*: we try to expand the CDF of $\sqrt{n}(\theta(\widehat{F}_n^*) - \theta(\widehat{F}_n))$ around a normal CDF using a Taylor expansion-like method. See Chapter 23.3 of Van der Vaart (2000) for more details.

12.8.1 Bootstrap and empirical process

In density estimator and regression analysis, we have seen that our estimator may be a function. So it will be of great interest to study if the bootstrap can be applied to a function estimator.

In this case, we need to introduce the concept of empirical processes. Let $\mathcal{F} = \{f_t : t \in \mathbb{T}\}$ be a collection of functions such that $f_t : \mathcal{X} \mapsto \mathbb{R}$, where $\mathcal{X}$ is the support of the observations and $\mathbb{T}$ is the index set. Consider the sample mean type estimator:

$$\widehat{\mathbb{P}}_n(f_t) = \frac{1}{n}\sum_{i=1}^n f_t(X_i).$$

It is easy to see that this is an unbiased estimator of

$$\mathbb{P}(f_t) = \mathbb{E}(f_t(X_1)).$$

The scaled difference is called the *empirical process*:

$$\mathbb{G}_n(f_t) = \sqrt{n}(\widehat{\mathbb{P}}_n(f_t) - \mathbb{P}(f_t)).$$

When we vary this quantity over $t \in \mathbb{T}$, you can see that $\mathbb{G}_n(f_t) = G(t)$ is a random function of the argument $t \in \mathbb{T}$. Although this seems to be abstract, here are some examples of empirical processes.

Example: EDF. Consider $\mathbb{T} = \mathcal{X}$ and

$$f_t(x) = I(x \le t).$$

In this case,

$$\widehat{\mathbb{P}}_n(f_t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t) = \widehat{F}_n(t)$$

is the EDF.

Example: KDE. Consider $\mathbb{T} = \mathcal{X}$ and

$$f_t(x) = \frac{1}{h^d} K\left(\frac{x-t}{h}\right).$$

Then it is easy to see that

$$\widehat{\mathbb{P}}_n(f_t) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{X_i - t}{h}\right) = \widehat{p}_h(t),$$

which is the KDE.

Example: log-likelihood function. Consider $\mathbb{T} = \Theta$ and

$$f_\theta(x) = \ell(\theta|x),$$

where $\ell(\theta|x) = \log p_\theta(x)$ is a log-likelihood function. Then

$$\widehat{\mathbb{P}}_n(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta|X_i)$$

is the joint log-likelihood function.

Example: empirical risk. Consider a prediction problem where we want to predict Y based on X and we assume that our predictor $m_\theta : \mathcal{X} \to \mathcal{Y}$ is a parametric model with parameter θ. Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be the loss function. Then the empirical risk

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} L(m_\theta(X_i), Y_i) = \widehat{\mathbb{P}}_n(f_\theta),$$

where $f_\theta(x, y) = L(m_\theta(x), y)$.

When the estimator is a vector, we have central limit theorem showing that the estimator converges to a Gaussian vector after rescaling. A similar pattern occurs for the empirical process as well. We call $\mathcal{F} = \{f_t : t \in \mathbb{T}\}$ a Donsker class if

$$\mathbb{G}_n(f_t) \xrightarrow{D} \mathbb{B}(f_t),$$

where $\mathbb{B}(f_t)$ is a Gaussian process and the notation $\xrightarrow{D}$ stands for convergence in distribution (weak convergence) of a stochastic process under L_∞ norm. Informally, we have the following result, known as the uniform central limit theorem.

Theorem 12.7 (Uniform central limit theorem) *Let $\mathbb{G}_n^*(f_t) = \sqrt{n}(\widehat{\mathbb{P}}_n^*(f_t) - \widehat{\mathbb{P}}_n(f_t))$ be the bootstrap process. When $\mathcal{F}$ is a Donsker class,*

$$\mathbb{G}_n^*(f_t) \approx \mathbb{G}_n(f_t).$$

In the formal statement, the approximation sign $\approx$ will be replaced by weak convergence of a stochastic process (notice that $\mathbb{G}_n^*(f_t)$ is a process indexed by t) conditioned on the sample $X_1, \cdots, X_n$. The take away message is that when the function class is Donsker, the bootstrap method works.

Roughly speaking, most common parametric models is a Donsker class as described in the following proposition.

Proposition 12.8 (Example 19.7 in van der Vaart (2000)) *Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a collection of functions such that Θ is bounded subset of $\mathbb{R}^d$. Suppose that there exists a function m such that*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \le m(x)\|\theta_1 - \theta_2\|$$

for any $\theta_1, \theta_2 \in \Theta$ and $\mathbb{E}(|m(X)|^2) < \infty$. Then $\mathcal{F}$ is Donsker.

There are several ways to argue that a function class $\mathcal{F}$ is Donsker. Common approaches are based on bounding the uniform covering number or the bracketing number. The two textbooks mentioned in the above (Kosorok 2008 and van der Vaart & Wellner 1996) are good references. You will learn more about this in STAT 580 sequence.

The Donsker theory can be combined with the functional delta method. Roughly speaking, if the parameter of interest is a statistical functional such that $\theta(\mathbb{P}(\mathcal{F}))$, where $\mathbb{P}(\mathcal{F}) = \{\mathbb{P}(f_t) : t \in \mathbb{T}\}$ is a stochastic process, then

$$\sqrt{n}(\theta(\widehat{\mathbb{P}}_n^*(\mathcal{F})) - \theta(\widehat{\mathbb{P}}_n(\mathcal{F}))) \approx \sqrt{n}(\theta(\widehat{\mathbb{P}}_n(\mathcal{F})) - \theta(\mathbb{P}(\mathcal{F}))),$$

when θ is Hadamard differentiable at $\mathbb{P}(\mathcal{F})$.

12.9 Beyond Functional Delta Method and Donsker Class

The functional delta method along with the Donsker theory makes the bootstrap a widely applicable approach. However, there are cases where these methodologies cannot be applied. Here are two examples that the regular Donsker theory does not work.

High-dimensional models. Let $X_1, \cdots, X_n \in \mathbb{R}^d$ be IID. Suppose that we are interested in estimating the mean vector $\mu = \mathbb{E}(X_1)$. It is easy to see that the sample mean $\bar{X}_n$ has asymptotic normality when d is fixed and the covariance matrix $\Sigma = \mathbb{E}(X_1 X_1^T)$ is finite. However, when $d = d_n \to \infty$, we may not have the same asymptotic normality[10]. In particular, when $\log d \asymp n$ (regular high-dimensional model), there is no asymptotic normality of the sample mean so it is unclear how to use the bootstrap to construct a confidence set of μ.

KDE with decreasing h. In the case of KDE, if h is fixed, then the function class

$$\left\{K\left(\frac{\cdot - x}{h}\right) : x \in \mathcal{X}\right\}$$

is a Donsker class. However, if $h = h_n \to 0$, then the function class being considered

$$\left\{K\left(\frac{\cdot - x}{h}\right) : x \in \mathcal{X}, 1 > h > 0\right\}$$

is no longer a Donsker class. Note that we use 1 as the upper bound of h since asymptotically the bandwidth will be less than 1. It can be replaced by any finite upper bound.

[10] We may still have CLT with $d \to \infty, d^2/n \to 0$, see *Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p 2/n is large. I. Consistency. The Annals of Statistics, 1298-1309.*

12.9.1 Bootstrap and High-Dimensional Models

Let $X_1, \cdots, X_n \in \mathbb{R}^d$ be IID with $\mathbb{E}(X_i) = \mu$ and $\Sigma = \mathbb{E}(X_1 X_1^T)$. In the recent work of Chernozhukov and his collaborator, they showed that under good conditions (allowing $d = o(e^{n^{1/8}})$),

$$\sqrt{\frac{n}{\log d}} \|\bar{X}_n - \mu\|_{\max} \approx \sqrt{\frac{n}{\log d}} \|\bar{Z}_n - \mu\|_{\max},$$

where $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ is the mean of standard normal vector. The $\approx$ here is actually under the Kolomogrov distance.

One famous paper is

> Chernozhukov, V., Chetverikov, D., & Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4), 2309-2352.

Moreover, the above result can be applied to the bootstrap sample mean $\bar{X}_n^*$, which leads to the following "rectangle confidence set". Let

$$\widehat{t}_\alpha = \widehat{G}^{-1}(1-\alpha/2), \quad \widehat{G}(s) = P(\|\bar{X}_n^* - \bar{X}_n\|_{\max} \le s | X_1, \cdots, X_n).$$

Namely, $\widehat{t}_\alpha$ is the $1-\alpha$ quantile of the the bootstrap maximum $\|\bar{X}_n^* - \bar{X}_n\|_{\max}$. Define a confidence set as

$$C_n = \{\mu \in \mathbb{R}^d : \|\bar{X}_n - \mu\|_{\max} \le \widehat{t}_\alpha\}.$$

Note that C_n looks like a rectangle in the parameter space.

Theorem 12.9 *Assume conditions in Chernozhukov, Chetverikov, and Kato (2017), which allowed $d = o(e^{n^{1/8}})$. Then*

$$P(\mu \in C_n) \ge 1 - \alpha - \frac{c_0 \log d}{n^{1/8}},$$

where c_0 is a constant.

Covariance matrix. Using the delta method, this result applies to other smoothed functionals of sample mean or higher moments as well. One notable result is the use of this idea to constructing confidence sets of the population covariance matrix in high dimensions. Let $\widehat{\Sigma}_n$ be the sample covariance matrix and let $\widehat{\Sigma}_n^*$ be the bootstrap sample covariance matrix. Define

$$\widehat{\eta}_\alpha = \widehat{G}^{-1}(1-\alpha/2), \quad \widehat{G}(s) = P(\|\widehat{\Sigma}_n^* - \widehat{\Sigma}_n\|_{\max} \le s | X_1, \cdots, X_n)$$

and two matrices

$$U_\alpha = \widehat{\Sigma}_n + \widehat{\eta}_\alpha, \quad L_\alpha = \widehat{\Sigma}_n - \widehat{\eta}_\alpha.$$

Then you can show that

$$P(L_{\alpha,ij} \le \Sigma_{ij} \le U_{\alpha,ij}, \ \forall i, j) \ge 1 - \alpha - \frac{c_0 \log d}{n^{1/8}}$$

for some constant. See the following paper for more details

> Wasserman, L., Kolar, M., & Rinaldo, A. (2014). Berry-Esseen bounds for estimating undirected graphs. *Electronic Journal of Statistics*, 8(1), 1188-1224.

12.9.2 Bootstrap and the Supremum of an Empirical Processes

The convergence of a maximum of a Gaussian vector also implies that the supremum of a good empirical process can also be well-approximated by the supremum of a Gaussian process. This result works even for some non-Donsker class. In particular, it can be applied to the KDE.

Let $X_1, \cdots, X_n$ be IID from some distribution supported on $\mathcal{X}$ and $\widehat{p}_h(x) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$ be the KDE and $\widehat{p}_h^*(x) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{X_i^* - x}{h}\right)$ be the bootstrap KDE. Also let $p_h(x) = \mathbb{E}(\widehat{p}_h(x))$ be the expected version of the KDE. Under the conditions in the following paper

> Chernozhukov, V., Chetverikov, D., & Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5), 1787-1818

we have the bootstrap validity of the confidence band of the KDE (actually, their result is stronger than the following theorem).

Theorem 12.10 (Chernozhukov, Chetverikov, and Kato (2014)) *Assume conditions in Chernozhukov, Chetverikov, and Kato (2014). Then there exists a mean zero Gaussian process $\mathbb{B}$ defined over $\mathcal{X}$ such that*

$$\sup_t \left| P\left(\sqrt{\frac{nh^d}{\log n}}\|\widehat{p}_h - p_h\|_\infty < t\right) - P\left(\|\mathbb{B}\|_\infty < t\right)\right| = O_P\left(\left(\frac{\log n}{nh^d}\right)^{1/8}\right).$$

Moreover,

$$\sup_t \left| P\left(\sqrt{\frac{nh^d}{\log n}}\|\widehat{p}_h^* - \widehat{p}_h\|_\infty < t | X_1, \cdots, X_n\right) - P\left(\sqrt{\frac{nh^d}{\log n}}\|\widehat{p}_h - p_h\|_\infty < t\right)\right| = O_P\left(\left(\frac{\log n}{nh^d}\right)^{1/8}\right).$$

The conditions in Theorem 12.10 is quiet mild–actually, it is the same conditions as the rate of uniform convergence. This Theorem also implies a construction of a simultaneous confidence band. Let

$$\widehat{t}_\alpha = \widehat{G}(1-\alpha/2), \quad \widehat{G}(s) = P(\|\widehat{p}_h^* - \widehat{p}_h\|_\infty \le s | X_1, \cdots, X_n)$$

be the $1-\alpha$ quantile of the bootstrap supremum error $\|\widehat{p}_h^* - \widehat{p}_h\|_\infty$. We define

$$L_\alpha = \widehat{p}_h - \widehat{t}_\alpha, \quad U_\alpha = \widehat{p}_h + \widehat{t}_\alpha.$$

Then you can show that L_α, U_α can be used as a simultaneous confidence band and we have

$$P(L_\alpha(x) \le p_h(x) \le U_\alpha(x)\ \forall x) \ge 1 - \alpha + c_0\left(\frac{\log n}{nh^d}\right)^{1/8}$$

for some constant $c_0 > 0$.

Note that the above confidence band is simultaneous for p_h, not the true PDF p. So we have to undersmooth the KDE (i.e, choosing h to be at a fast rate to 0 than the optimal rate) to obtain a valid confidence band. One approach that can by pass this problem is via bootstrapping the *debiased estimator*. See the following paper

> Cheng, G., & Chen, Y. C. (2019). Nonparametric inference via bootstrapping the debiased estimator. *Electronic Journal of Statistics*, 13(1), 2194-2256.