Chapter 11

The Bootstrap

This chapter covers the following topics:

- What is the Bootstrap?
- Why Does it Work?
- Examples of the Bootstrap.

11.1 Introduction

Most of this volume is devoted to parametric inference. In this chapter we depart from the parametric framework and discuss a nonparametric technique called *the bootstrap*. The bootstrap is a method for estimating the variance of an estimator and for finding approximate confidence intervals for parameters. Although the method is nonparametric, it can be used for inference about parameters in parametric and nonparametric models which is why we include it in this volume.

11.2 A More General Notion of "Parameter"

We begin by broadening what we mean by a parameter. Let us begin with a few examples.

1. Let $X_1, \ldots, X_n \sim P$ where $P \in (P_\theta : \theta \in \Theta)$. Let $\widehat{\theta}_n$ be the maximum likelihood estimator of θ. We would like to estimate the variance of $\widehat{\theta}_n$ and we want a $1 - \alpha$ confidence interval for θ.

2. Let $X_1, \ldots, X_n \sim P$ and let $\theta = T(P)$ denote the mean of P. Hence, $\theta = \mathbb{E}[X_i] = \int x dP(x)$. Let $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Again, we would like to estimate the variance of $\widehat{\theta}_n$ and we want a $1-\alpha$ confidence interval for θ.

3. Let $X_1, \ldots, X_n \sim P$ and let $\theta = T(P)$ denote the median of P. Hence, $\mathbb{P}(X_i \leq \theta) = \mathbb{P}(X_i > \theta) = 1/2$. Let $\widehat{\theta}_n$ denote the sample median. Yet again, we would like to estimate the variance of $\widehat{\theta}_n$ and we want a $1-\alpha$ confidence interval for θ.

In the first example, θ denotes the parameter of a parametic model. In the second and third example, we are in a nonparametric situation; in these cases we think of a "parameter" as a function of the distribution P and we write $\theta = T(P)$. The bootstrap can be used in both the parametric and nonparametric settings.

Let P_n be the *empirical distribution*. This is the discrete distribution that puts mass $1/n$ at each datapoint X_i. Hence,

$$P_n(A) = \frac{1}{n}\sum_{i=1}^n I(X_i \in A). \tag{11.1}$$

In the nonparametric case, we will estimate the parameter $\theta = T(P)$ by $\widehat{\theta}_n = T(P_n)$ which is called the *plug-in estimator.* For example, when $\theta = T(P) = \int x dP(x)$ is the mean, the plug-in estmator is

$$\widehat{\theta}_n = T(P_n) = \int x dP_n(x) = \frac{1}{n}\sum_{i=1} X_i \tag{11.2}$$

which is the sample mean.

A sample of size n drawn from P_n is called a *bootstrap* sample, denoted by

$$X_1^*, \ldots, X_n^* \sim P_n.$$

Bootstrap samples play an important role in what follows. Note that drawing an iid sample $X_1^*, \ldots, X_n^*$ from P_n is equivalent to drawing n observations, with replacement, from the original data $\{X_1, \ldots, X_n\}$. Thus, bootstrap sampling is often described as "resampling the data." This can be a bit confusing and we think it is much clearer to think of a bootstrap sample $X_1^*, \ldots, X_n^*$ as n draws from the empirical distribution P_n.

11.3 The Bootstrap

Now we give the bootstrap algorithms for estimating the variance of $\widehat{\theta}_n$ and for constructing confidence intervals. The explanation of why (and when) the bootstrap gives valid estimates, is deferred until Section 11.5. Let $\widehat{\theta}_n = g(X_1, \ldots, X_n)$ denotes some estimator.

Bootstrap Variance Estimator

1. Draw a bootstrap sample $X_1^*, \ldots, X_n^* \sim P_n$. Compute $\widehat{\theta}_n^* = g(X_1^*, \ldots, X_n^*)$.
2. Repeat the previous step, B times, yielding estimators $\widehat{\theta}_{n,1}^*, \ldots, \widehat{\theta}_{n,B}^*$.
3. Compute:
$$\widehat{s} = \sqrt{\frac{1}{B} \sum_{j=1}^{B} (\widehat{\theta}_{n,j}^* - \overline{\theta})^2}$$
where $\overline{\theta} = \frac{1}{B} \sum_{j=1}^{B} \widehat{\theta}_{n,j}^*$.
4. Output $\widehat{s}$.

The next theorem states that $\widehat{s}^2$ approximates $\mathrm{Var}(\widehat{\theta}_n)$. The are two sources of error in this apprixmation. The first is due to the fact that n is finite and the second is due to the fact that B is finite. However, we can make B as large as we like. (In practice, it usually suffices to take $B = 10,000$.) So we ignore the error due to finite B.

Theorem 138. Under appropriate regularity conditions, $\frac{s^2}{\mathrm{Var}(\widehat{\theta}_n)} \xrightarrow{P} 1$ as $n \to \infty$.

Now we describe the confidence interval algorithm.

Bootstrap Confidence Interval

1. Draw a bootstrap sample $X_1^*, \ldots, X_n^* \sim P_n$. Compute $\widehat{\theta}_n^* = g(X_1^*, \ldots, X_n^*)$.
2. Repeat the previous step, B times, yielding estimators $\widehat{\theta}_{n,1}^*, \ldots, \widehat{\theta}_{n,B}^*$.
3. Let
$$\widehat{F}(t) = \frac{1}{B} \sum_{j=1}^{B} I\Big(\sqrt{n}(\widehat{\theta}_{n,j}^* - \widehat{\theta}_n\Big) \leq t).$$
4. Let
$$C_n = \left[\widehat{\theta}_n - \frac{t_{1-\alpha/2}}{\sqrt{n}},\ \widehat{\theta}_n - \frac{t_{\alpha/2}}{\sqrt{n}}\right]$$
where $t_{\alpha/2} = \widehat{F}^{-1}(\alpha/2)$ and $t_{1-\alpha/2} = \widehat{F}^{-1}(1-\alpha/2)$.
5. Output C_n.

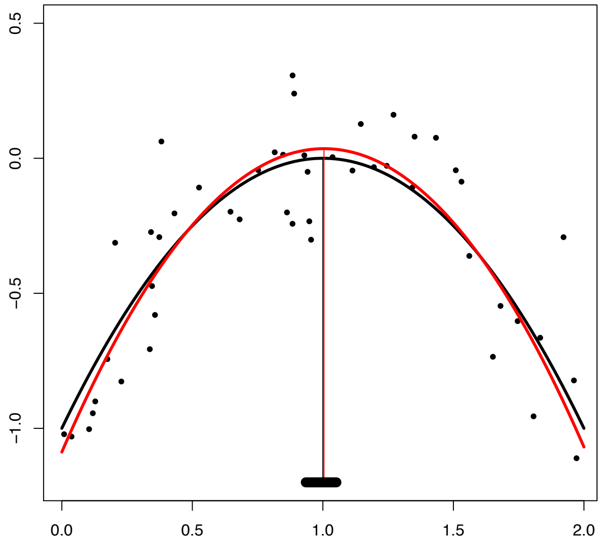

Figure 11.1: 50 points drawn from the model $Y_i = -1 + 2X_i - X_i^2 + \epsilon_i$ where $X_i \sim \text{Uniform}(0,2)$ and $\epsilon_i \sim N(0, .2^2)$. In this case, the maximum of the polynomail occurs at $\theta = 1$. The true and estimated curves are shown in the figure. At the bottom of the plot we show the 95 percent boostrap confidence interval based on $B = 1,000$.

Theorem 139. Under appropriate regularity conditions,

$$\mathbb{P}(\theta \in C_n) = 1 - \alpha - O\left(\frac{1}{\sqrt{n}}\right).$$

as $n \to \infty$.

11.4 Examples

Example 140. Consider the polynomial regression model $Y = g(X) + \epsilon$ where $X, Y \in \mathbb{R}$ and $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ we can estimate $\beta = (\beta_0, \beta_1, \beta_2)$ with the least squares estimator $\widehat{\beta}$. Suppose that $g(x)$ is concave and we are interested in the location at which $g(x)$ is maximized. It is easy to see that the maximum occurs at $x = \theta$ where $\theta = -(1/2)\beta_1/\beta_2$. A point estimate of θ is $\widehat{\theta} = -(1/2)\widehat{\beta}_1/\widehat{\beta}_2$. Now we use the bootstrap to get a confidence interval for θ. Figure 11.1 shows 50 points drawn from the above model with $\beta_0 = -1$, $\beta_1 = 2$, $\beta_2 = -1$. The X_i's were sample uniformly on $[0, 2]$ and we took $\epsilon_i \sim N(0, .2^2)$. In this case, $\theta = 1$. The true and estimated curves are shown in the figure. At the bottom of the plot we show the 95 percent boostrap confidence interval based on $B = 1,000$.

Example 141. Let $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n) \sim P$ where $X_i \in \mathbb{R}, Y_i \in \mathbb{R}, Z_i \in \mathbb{R}^d$. The partial correlation of X and Y given Z is

$$\theta = -\frac{\Omega_{12}}{\sqrt{\Omega_{11}\Omega_{22}}}$$

where $\Omega = \Sigma^{-1}$ and Σ is the covariance matrix of $W = (X, Y, Z)^T$. The partial correlation measures the linear dependence between X and Y after removing the effect of Z. For illustration, suppose we generate the data as follows: we take $Z \sim N(0, 1)$, $X = 10Z + \epsilon$ and $Y = 10Z + \delta$ where $\epsilon, \delta \sim N(0, 1)$. The correlation between X and Y is very large. But the partial correlation is 0. We generated $n = 100$ data points from this model. The sample correlation was 0.99. However, the estimate partial correaltion was -0.16 which is much closer to 0. The 95 percent bootstrap confidence interval is [-.33,.02] which includes the true value, namely, 0.

11.5 Why Does the Bootstrap Work?

To explain why the bootstrap works, let us begin with a heuristic. Let

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}_n) \leq t)$$

and let

$$\widehat{F}_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}_n) \leq t | X_1, \ldots, X_n).$$

be the bootstrap approximation to F_n. We do not know F_n be we do know $\widehat{F}_n$ in the sense that it depends only on the observed data. Usually, F_n will be close to some limiting distribution L. Similarly, $\widehat{F}_n$ will be close to some limiting distribution $\widehat{L}$. Moreover, L and $\widehat{L}$ will be close which implies that F_n and $\widehat{F}_n$ are close. In practice, we usually approximate $\widehat{F}_n$ by its Monte Carlo version

$$\overline{F}(t) = \frac{1}{B} \sum_{j=1}^{B} I(\sqrt{n}(\widehat{\theta}_j^* - \widehat{\theta}_j) \leq t).$$

But $\overline{F}$ is close to $\widehat{F}_n$ as long as we take B large. See Figure 11.2.

Now we will give more detail in a simple, special case. Suppose that $X_1, \ldots, X_n \sim P$ where X_i has mean μ and variance σ^2. Suppose we want to construct a confidence interval for μ.

Let $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and define

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\mu}_n - \mu) \leq t). \tag{11.3}$$

We do not know the cdf F. But, for the moment, that an oracle gave us F. For any $0 < \beta < 1$, define $z_\beta = F^{-1}(\beta)$. Define the *oracle confidence interval*

$$A_n = \left[\widehat{\mu}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}},\ \widehat{\mu}_n - \frac{z_{\alpha/2}}{\sqrt{n}}\right]. \tag{11.4}$$

We claim that B_n is a $1-\alpha$ confidence interval. To see this, note that the probability that A_n traps μ is

$$\begin{aligned}\mathbb{P}(\mu \in A_n) &= \mathbb{P}\left(\widehat{\mu}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}} \le \mu \le \widehat{\mu}_n - \frac{z_{\alpha/2}}{\sqrt{n}}\right)\\ &= \mathbb{P}\left(z_{\alpha/2} \le \sqrt{n}(\widehat{\mu}_n - \mu) \le z_{1-\alpha/2}\right)\\ &= F_n(z_{1-\alpha/2}) - F_n(z_{\alpha/2}) = \left(1 - \frac{\alpha}{2}\right) - \frac{\alpha}{2} = 1-\alpha.\end{aligned}$$

Unfortunately, we do not know F but we can estimate it. The bootstrap estimate if F is

$$\widehat{F}_n(t) = \mathbb{P}\Big(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) \le t\ \Big|\ X_1, \ldots, X_n\Big)$$

where $\widehat{\mu}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$ and $X_1^*, \ldots, X_n^* \sim P_n$. The data $X_1, \ldots, X_n$ are treated as fixed during the bootstrap which is why we write $\widehat{F}_n$ as a conditional distribution.

Note that when we do the bootstrap algorithm, we are just approximating $\widehat{F}_n(t)$ by

$$\overline{F}(t) = \frac{1}{B}\sum_{j=1}^{B} I(\sqrt{n}(\widehat{\mu}_{n,j}^* - \widehat{\mu}_n) \le t).$$

But

$$\sup_t |\overline{F}(t) - \widehat{F}_n(t)| \to 0$$

almost surely, as $B \to \infty$. Since we can take B as large as we want, we can ignore the approximation error and just assume we know $\widehat{F}_n(t)$. For any $0 < \beta < 1$, define $t_\beta = \widehat{F}^{-1}(\beta)$. The bootstrap confidence interval is

$$C_n = \left[\widehat{\mu}_n - \frac{t_{1-\alpha/2}}{\sqrt{n}},\ \widehat{\mu}_n - \frac{t_{\alpha/2}}{\sqrt{n}}\right]. \tag{11.5}$$

This is the same as the oracle confidence interval except that we have used $t_{\alpha/2}$ and $t_{1-\alpha/2}$ in place of $z_{\alpha/2}$ and $z_{1-\alpha/2}$. To show that $t_{\alpha/2} \approx z_{\alpha/2}$ and $t_{1-\alpha/2} \approx z_{1-\alpha/2}$, we need to show that $\widehat{F}_n(t)$ approximates $F_n(t)$.

Theorem 142 (Bootstrap Theorem). Suppose that $\mu_3 = \mathbb{E}|X_i|^3 < \infty$. Then,

$$\sup_t |\widehat{F}_n(t) - F_n(t)| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

$$\begin{tikzcd}
F_n \arrow[r, "O(1/\sqrt{n})"] \arrow[d, dashed] & L \arrow[d, "O_P(1/\sqrt{n})"] \\
\widehat{F}_n \arrow[r, "O_P(1/\sqrt{n})"'] \arrow[d, "O(1/\sqrt{B})"] & \widehat{L} \\
\overline{F} &
\end{tikzcd}$$

Figure 11.2: The distribution $F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \leq t)$ is close to some limit distribution L. Similarly, the bootstrap distribution $\widehat{F}_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \leq t | X_1, \ldots, X_n)$ is close to some limit distribution $\widehat{L}$. Since $\widehat{L}$ and L are close, it follows that F_n and $\widehat{F}_n$ are close. In practice, we approximate $\widehat{F}_n$ with its Monte Carlo version $\overline{F}$ which we can make as close to $\widehat{F}_n$ as we like by taking B large.

To prove this result, let us recall that Berry-Esseen Theorem from Chapter 2. For convenience, we repeat the theorem here.

Theorem 143 (Berry-Esseen Theorem)**.** Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Let $\mu_3 = \mathbb{E}[|X_i - \mu|^3] < \infty$. Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ be the sample mean and let Φ be the cdf of a $N(0,1)$ random variable. Let $Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$. Then

$$\sup_z \Big| \mathbb{P}(Z_n \le z) - \Phi(z) \Big| \le \frac{33}{4} \frac{\mu_3}{\sqrt{n}}. \tag{11.6}$$

Proof of the Bootstrap Theorem. Let $\Phi_\sigma(t)$ denote the cdf of a Normal with mean 0 and variance σ^2. Let $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2$. Thus, $\widehat{\sigma}^2 = \mathrm{Var}(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n)|X_1, \ldots, X_n)$. Now, by the triangle inequality,

$$\begin{aligned}\sup_t |\widehat{F}_n(t) - F_n(t)| &\le \sup_t |F_n(t) - \Phi_\sigma(t)| + \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| + \sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)| \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

Let $Z \sim N(0,1)$. Then, $\sigma Z \sim N(0, \sigma^2)$ and from the Berry-Esseen theorem,

$$\begin{aligned}\mathrm{I} = \sup_t |F_n(t) - \Phi_\sigma(t)| &= \sup_t \left| \mathbb{P}\left(\sqrt{n}(\widehat{\mu}_n - \mu) \le t\right) - \mathbb{P}\left(\sigma Z \le t\right) \right| \\ &= \sup_t \left| \mathbb{P}\left(\frac{\sqrt{n}(\widehat{\mu}_n - \mu)}{\sigma} \le \frac{t}{\sigma}\right) - \mathbb{P}\left(Z \le \frac{t}{\sigma}\right) \right| \le \frac{33}{4}\frac{\mu_3}{\sqrt{n}}.\end{aligned}$$

Using the same argument on the third term, we have that

$$\mathrm{III} = \sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)| \le \frac{33}{4}\frac{\widehat{\mu}_3}{\sqrt{n}}$$

where $\widehat{\mu}_3 = \frac{1}{n}\sum_{i=1} |X_i - \widehat{\mu}_n|^3$ is the empirical third moment. By the strong law of large numbers, $\widehat{\mu}_3$ converges almost surely to μ_3. So, almost surely, for all large n, $\widehat{\mu}_3 \le 2\mu_3$ and so $\mathrm{III} \le \frac{33}{4}\frac{2\mu_3}{\sqrt{n}}$. From the fact that $\widehat{\sigma} - \sigma = O_P(\sqrt{1/n})$ it may be shown that $\mathrm{II} = \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| = O_P(\sqrt{1/n})$. (This may be seen by Taylor expanding $\Phi_{\widehat{\sigma}}(t)$ around σ.) This completes the proof. □

We have shown that $\sup_t |\widehat{F}_n(t) - F_n(t)| = O_P\left(\frac{1}{\sqrt{n}}\right)$. From this, it may be shown that, for each $0 < \beta < 1$, $t_\beta - z_\beta = O_P\left(\frac{1}{\sqrt{n}}\right)$. From this, one can prove Theorem 139.

So far we have focused on the mean. Similar theorems may be proved for more general parameters. The details are complex so we will not discuss them here. We give a little more information in the appendix. For a thorough treatment, we refer the reader to Chapter 23 of van der Vaart (1998).

11.6 A Few Remarks About the Bootstrap

Here are some random remarks about the bootstrap:

1. The bootstrap is nonparametric but it does require some assumptions. You can't assume it is always valid. (See the appendix.)

2. The bootstrap is an asymptotic method. Thus the coverage of the confidence interval is $1 - \alpha + r_n$ where, typically, $r_n = C/\sqrt{n}$.

3. There is a related method called the jackknife where the standard error is estimated by leaving out one observation at a time. However, the bootstrap is valid under weaker conditions than the jackknife. See Shao and Tu (1995).

4. Another way to construct a bootstrap confidence interval is to set $C = [a, b]$ where a is the $\alpha/2$ quantile of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ and b is the $1-\alpha/2$ quantile. This is called the percentile interval. This interval seems very intuitive but does not have the theoretical support for the interval B_n. However, in practice, the percentile interval and B_n are often quite similar.

5. There are many cases where the bootstrap is not formally justified. This is especially true with discrete structures like trees and graphs. Nonethless, the bootstrap can be used in an informal way to get some intuition of the variability of the procedure. But keep in mind that the formal guarantees may not apply in these cases. For example, see Holmes (2003) for a discussion of the bootstrap applied to phylogenetic tres.

6. There is a method related to the bootstrap called subsampling. In this case, we draw samples of size $m < n$ without replacement. Subsampling produces valid confidence intervals under weaker conditions than the bootstrap. See Politis, Romano and Wolf (1999).

7. There are many modifications of the bootstrap that lead to more accurate confidence intervals; see Efron (1996).

8. There is also a parametric bootstrap. If $\{p(x : \theta) :\ \theta \in \Theta\}$ is a parametric model and $\widehat{\theta}$ is an estimator, such as the maximum likelihood estimator, we sample $X_1^*, \ldots, X_n^*$ from $p(x; \widehat{\theta})$ instead of sampling from P_n.

11.7 The High-Dimensional Bootstrap

Now we consider the bootstrap in high-dimensions. Let $X_1, \ldots, X_n \in \mathbb{R}^d$ where d may be larger than n. In fact, we allow the dimension $d = d_n$ to increase with n. We will assume that the distribution of X_i is sub-Gaussian, although this is stronger than needed. This means that $\mathbb{E}(e^{t^T X}) \leq e^{c||t||^2}$ for some $c > 0$.

Let $\mu = \mathbb{E}[X_i] \in \mathbb{R}^d$. Here is a bootstrap algorithm for constructing a confidence set for μ.

High Dimensional Bootstrap

1. Draw a bootstrap sample $X_1^*, \ldots, X_n^* \sim P_n$. Compute $\widehat{\mu}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$.
2. Repeat the previous step, B times, yielding estimators $\widehat{\mu}_{n,1}^*, \ldots, \widehat{\mu}_{n,B}^*$.
3. Let
$$\widehat{F}_n(t) = \frac{1}{B}\sum_{j=1}^B I(\sqrt{n}||\widehat{\mu}_{n,j}^* - \widehat{\mu}_n||_\infty \leq t).$$
4. Let
$$C_n = \left\{ a \in \mathbb{R}^d : \ ||a - \widehat{\mu}_n||_\infty \leq \frac{t_\alpha}{\sqrt{n}} \right\}$$
where $t_\alpha = \widehat{F}^{-1}(1-\alpha)$.
5. Output C_n.

Theorem 144 (Chernozhukov, Chetverikov and Kato, 2014)**.** Suppose that $d = o(e^{n^{1/8}})$. Then
$$\mathbb{P}(\mu \in C_n) \geq 1 - \alpha - \frac{c \log d}{n^{1/8}}$$
for some $c > 0$.

Under the stated conditions, the same result applies to higher-order moments. If $\theta = g(\mu)$ for some function g then we can get a confidence set for θ by applying g to C_n. We call this the *projected confidence set*. That is, if we define $A_n = \{g(\mu) : \ \mu \in C_n\}$ then it follows that
$$\mathbb{P}(\theta \in A_n) \geq 1 - \alpha - \frac{c \log d}{n^{1/8}}.$$

Alternatively, we can apply the bootstrap to $\sqrt{n}(g(\widehat{\mu}) - g(\mu))$. However, we do not automatically get the same coverage guarantee that the projected set has.

Example 145. Let us consider constructing a confidence set for a high-dimensional covariance matrix. Let $X_1, \ldots, X_n \in \mathbb{R}^k$ be a random sample and let $\Sigma = \mathrm{Var}(X)$ which is a $k \times k$ matrix. There are $d = O(k^2)$ parameters here. Let $\widehat{\Sigma} = (1/n)\sum_{i=1}^n (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T$. Also, let $\sigma = \mathrm{vec}(\Sigma)$ and $\widehat{\sigma} = \mathrm{vec}(\widehat{\Sigma})$, where vec takes a matrix and converts it into a vector by stacking the columns. We can then apply the bootstrap algorithm above to $\sqrt{n}(\widehat{\sigma} - \sigma)$

to get the bootstrap quantile t_α. Let $\ell_n = \widehat{\sigma} - t_\alpha/\sqrt{n}$ and $u_n = \widehat{\sigma} + t_\alpha/\sqrt{n}$. We can then unstack ℓ_n and u_n into $k \times k$ matrices L_n and U_n. It then follows that

$$\mathbb{P}(L_n \leq \Sigma \leq U_n) \geq 1 - \alpha - \frac{c \log d}{n^{1/8}}$$

where $A \leq B$ means that $A_{jk} \leq B_{jk}$ for all (j,k).

11.8 Subsampling

11.9 Finite Sample Methods

11.9.1 The Permutation Test

In this section we discuss a nonparametric hypothesis testing method. The test is not based on the bootstrap but we include it here because it is similar in spirit to the bootstrap. Let

$$X_1, \ldots, X_n \sim F, \qquad Y_1, \ldots, Y_m \sim G$$

be two independent samples and suppose we want to test the hypothesis

$$H_0 : F = G \quad \text{versus} \quad H_1 : F \neq G. \tag{11.7}$$

The permutation test gives an exact (nonasymptotic), nonparametric method for testing this hypothesis. Let $Z = (X, Y)$ where $X = (X_1, \ldots, X_n)^T$ and $Y = (Y_1, \ldots, Y_m)^T$. Define a vector W of length $N = n + m$ that indicates which group Z_i is from. Thus, $W_i = 1$ if $i \leq n$ and $W_i = 2$ if $i > n$. The data look like this:

$(X,Y)^T$	X_1	$\ldots$	X_n	Y_1	$\ldots$	Y_m
Z	Z_1	$\ldots$	Z_n	Z_{n+1}	$\ldots$	Z_{n+m}
W	1	$\ldots$	1	2	$\ldots$	2

Let $T = T(Z, W)$ be any test statistic. For example, consider $T = |\overline{X} - \overline{Y}|$. We can write T as a function of Z and W as follows. Define $X(Z, W) = \{Z_i : \; W_i = 1\}$ and $Y(Z, W) = \{Z_i : \; W_i = 2\}$ and then $T = |\overline{X} - \overline{Y}| = |\overline{X(Z, W)} - \overline{Y(Z, W)}|$.

Let $T^* = T(Z, W^*)$ where W^* denotes a random permutation of W. Define the permutation p-value

$$p = \mathbb{P}(T^* > t) \tag{11.8}$$

where $t = T(Z, W)$ is the observed value of the test statistic. This p-value defines an exact test. The steps of the algorithm are as follows:

Permutation Test

1. Compute $t = T(Z, W)$.
2. Repeat B times: form a random permutation W^* of W and compute $T^* = T(Z, W^*)$. Denote the values by $T_1^*, \ldots, T_B^*$.
3. Compute the p-value

$$p = \frac{1}{B}\sum_{j=1}^{B} I(T_j^* > t). \tag{11.9}$$

Theorem 146. Suppose we reject $H_0 : F = G$ whenever $p < \alpha$. If H_0 is true then $\mathbb{P}(\text{rejecting } H_0) \leq \alpha$.

Proof. xxxx □

The test is called exact since the probability of falsely rejecting the null hypothesis is less than or equal to α. There is no large sample approximation here.

Remark: *There is a bootstrap hypothesis test that is similar to the permutation test. The advantage of the bootstrap test is that it is more general than the permutation test. The disadvantage is that it is an approximate test, not an exact test. The bootstrap p-value based on a statistic $T = T(X)$ is*

$$p = \mathbb{P}_{F_0}(T^* > t) \tag{11.10}$$

where $t = T(X)$, $T^* = T(X^*)$ and X^* is drawn from the null distribution F_0. If the null hypothesis does not completely specify a distribution F_0 then we compute $p = \mathbb{P}_{\widehat{F}_0}(T^* > t)$ where $\widehat{F}_0$ is an estimate F under the restriction that $F \in \mathcal{F}_0$ where $\mathcal{F}_0$ is the set of distributions consistent with the null hypothesis. However, this is an approximate test while the permutation test is exact.

Example 147. Gretton et al (2008) developed a two sample test based on reproducing kernel Hilbert spaces. The test statistic is

$$T = \frac{1}{n^2}\sum_{i,j=1}^{n} K(X_i, X_j) - \frac{2}{nm}\sum_{i,j=1}^{n} K(X_i, Y_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} K(Y_i, Y_j)$$

where K is a symmetric kernel. Suppose we take $K = K_h(x, y) = e^{-||x-y||^2/(2h^2)}$ to be the Gaussian kernel. Rather than choosing a bandwidth h we can simply define the test statistic to be the maximum over all bandwidths:

$$T = \sup_{h>0}\left(\frac{1}{n^2}\sum_{i,j=1}^{n} K_h(X_i, X_j) - \frac{2}{nm}\sum_{i,j=1}^{n} K_h(X_i, Y_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} K_h(Y_i, Y_j)\right).$$

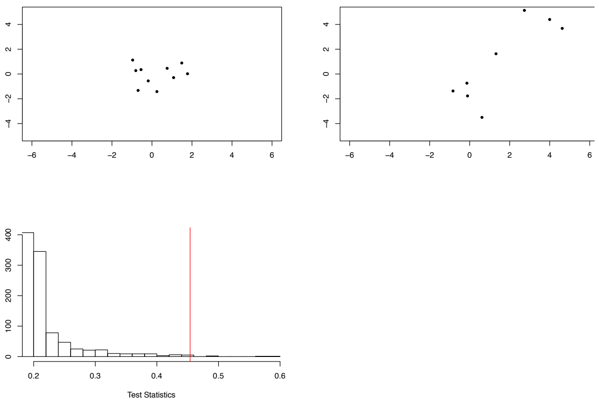

Figure 11.3: Top left: $X_1, \ldots, X_n$. Top right: $Y_1, \ldots, Y_m$. Bottom left: values of the test statistic from 1,000 permutations.

It would be difficult to find a useful expression for the distribution of the test statistic T under the null hypothesis $H_0 : F = G$. However, we can compute the p-value easily using the permutation test. Figure 11.3 shows an example. The top left plot shows $n = 10$ observations from F and the top right plot shows $n = 10$ observations from G. (We took F to be bivariate normal and G to be a mixture of two normals.) The test statistic is 0.45 and the p-value, based on $B = 1,000$ is 0.006 suggesting that we should reject H_0. The bottom left shows a histogram of the values of T from the 1,000 permutations. The vertical line is the observed value of T. The p-value is the fraction of statistics greater than T.

11.9.2 Confidence Rectangles for Quantiles

11.9.3 Confidence Rectangles for Means

11.9.4 Conformal Methods

11.10 Summary

The bootstrap provides nonparametric standard errors and confidence intervals. To draw a bootstrap sample we draw n observations $X_1^*, \ldots, X_n^*$ from the empirical distribution P_n. This is equivalent to drawing n observations with replacement from the original daa $X_1, \ldots, X_n$. We then compute the estimator $\widehat{\theta}^* = g(X_1^*, \ldots, X_n^*)$. If we repeat this whole process B times we get $\widehat{\theta}_1^*, \ldots, \theta_B^*$. The standard deviation of these values approximates the stanard error of $\widehat{\theta}_n = g(X_1, \ldots, X_n)$.

11.11 Bibliographic Remarks

Further details on statistical functionals can be found in [51], [13], [52], [23] and [59]. The jackknife was invented by [47] and [58]. The bootstrap was invented by [20]. There are several books on these topics including [22], [13], [29] and [52]. Also, see Section 3.6 of [60].

Appendix

More on Plug-in Estimators. Let $\theta = T(P)$. The plug-in estimator of θ is $\widehat{\theta}_n = T(P_n)$ where P_n is the empirical distribution that puts mass $1/n$ at each X_i. For example, suppose that $T(P) = \int x\, dP(x)$ is the mean. Then $T(P_n) = \int x\, dP_n(x) = n^{-1}\sum_{i=1}^n X_i$ since itegrating with respect to P_n corresponds to summing over the discrete measure with mass $1/n$ at X_i.

As another example, suppose that $\theta = T(P)$ is the variance of X. Let μ denote the mean. Then

$$\theta = \mathbb{E}(X-\mu)^2 = \int (x-\mu)^2 dP(x) = \int x^2 dP(x) - \left[\int x dP(x)\right]^2.$$

Thus, the plug-in estimator is

$$\widehat{\theta}_n = \int x^2 dP_n(x) - \left[\int x dP_n(x)\right]^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left[\frac{1}{n}\sum_{i=1}^n X_i\right]^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

For one more example, let θ be the α quantile of X. Here it is convenient to work with the cdf $F_n(x) = P(X \leq x)$. Thus $\theta = T(P) = T(F) = F^{-1}(\alpha)$ where $F^{-1}(y) = \inf_x\{F_n(x) \geq y\}$. The empirical cdf is $F_n(x) = n^{-1}\sum_{i=1}^n I(X_i \leq x)$ and $\widehat{\theta}_n = T(F_n) = \inf_x\{F_n(x) \geq \alpha\}$. In other words, $\widehat{\theta}_n$ is just the corresponding sample quantile.

Hadamard Differentiability. The key condition needed for the bootstrap is Hadamard differentiability. Let $\mathcal{P}$ denote all distributions on the real line and let $\mathcal{D}$ denote the linear space generated by $\mathcal{P}$. Write $T((1-\epsilon)P + \epsilon Q) = T(P + \epsilon D)$ where $D = Q - P \in \mathcal{D}$. The

Gateaux derivative, Gateaux derivative is defined by

$$L_P(D) = \lim_{\epsilon \downarrow 0} \left| \frac{T(P + \epsilon D) - T(P)}{\epsilon} - L_F(D) \right| \to 0. \tag{11.11}$$

Thus $T(P + \epsilon D) \approx \epsilon L_P(D) + o(\epsilon)$ and the error term $o(\epsilon)$ goes to 0 as $\epsilon \to 0$. Hadamard differentiability requires that this error term be small uniformly over compact sets. Equip $\mathcal{D}$ with a metric d. T is Hadamard differentiable at P if there exists a linear functional L_P on $\mathcal{D}$ such that for any $\epsilon_n \to 0$ and $\{D, D_1, D_2, \ldots\} \subset \mathcal{D}$ such that $d(D_n, D) \to 0$ and $P + \epsilon_n D_n \in \mathcal{P}$,

$$\lim_{n \to \infty} \left(\frac{T(P + \epsilon_n D_n) - T(P)}{\epsilon_n} - L_P(D_n) \right) = 0. \tag{11.12}$$

Let $d(P, Q) = \sup_x |P((-\infty, x]) - Q((-\infty, x])|$. Sufficient conditions for bootstrap validity are: T is Hadamard differentiable with respect to d and $0 < \int L_P^2(\delta_x - P) dP(x) < \infty$ where δ_x denotes a point mass at x.