Module 10
The Bootstrap

We begin by thinking about a sample from a Normal distribution. Let's assume that $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, and let's assume that we want to learn about the mean, μ. If this is the case, we can take the average $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$. This makes sense, as $E\left(\overline{X}\right) = \mu$, and $\text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$. And if I know σ^2, I can easily create a $100(1-\alpha)\%$ confidence interval for μ with

$$\overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution.

Now let's assume that σ^2 is not known. In this case, if you want to create a confidence interval for μ, you have to estimate σ^2 with

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2.$$

The hat over the sigma implies, of course, that it it sht estimated value of σ^2. If σ^2 is estimated this way, a $100(1-\alpha)\%$ confidence interval for μ is calculated as

$$\overline{X} \pm t_{n-1;1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}.$$

And if n is large enough, then it is fair to replace $t_{n-1;1-\frac{\alpha}{2}}$ with $z_{1-\frac{\alpha}{2}}$ and calculate the confidence interval for μ as

$$\overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}.$$

Now let's consider something else. Assume that $X_1, X_2, \ldots, X_n$ are i.i.d. and come from some distribution that is not normal. I will call this distribution f_X and assume that $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. If we want to learn more about the population mean, μ, then we can estimate it with $\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} X_j$. And if n is large enough, the central limit theorem "kicks in" and I can approximately say that $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. With this approximation in mind, a $100(1-\alpha)\%$ confidence interval for μ can be calculated just as above, with

$$\overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}.$$

But what if the Central Limit Theorem doesn't adequately approximate the distribution, or what if the sample size isn't large enough to justify its use? Or what would happen if we were interested in another parameter besides the population mean, μ. Say we were interested in some complicated parameter θ, where θ can be estimated from the dataset as $\hat{\theta} = g(\mathbf{X})$, where g is some complicated (perhaps nonlinear) function How would we estimate the variability of $\hat{\theta}$ if analytically calculating $\text{Var}\left(\hat{\theta}\right)$ were too difficult? How would we make any statistical inference on $\hat{\theta}$? How would we construct a confidence interval of θ using $\hat{\theta}$ if we don't know what distribution the data come from and we can't analytically calculate $\text{Var}\left(\hat{\theta}\right)$? This is what the bootstrap is used for.

In this module, we will talk about how the bootstrap can be used (i) to estimate the variability of a parameter estimator $\hat{\theta}$, (ii) construct a confidence interval for θ, and (iii) estimate the bias in $\hat{\theta}$. Recall that the bias of an estimate is calculated as

$$\text{bias}\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta.$$

Before any of this is done, though, we discuss the empirical distribution function.

1 The Empirical Distribution Function

Before details are provided on the actual bootstrap, it is important to remind you that the bootstrap was partially motivated by the fact that often times it is difficult (or unreasonable) to assume that the data came from from any distribution that can easily be parametrized (such as the normal, exponential, cauchy, etc.) If the distribution is not known, it is difficult to do any analytical calculations.

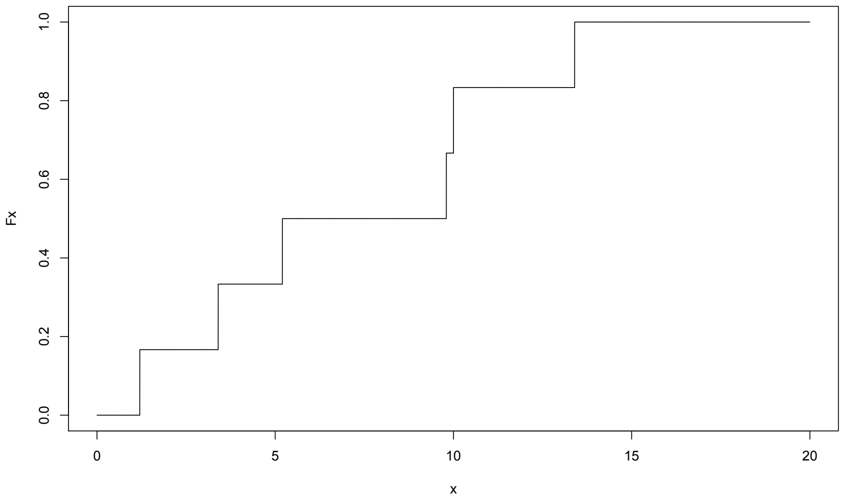
In cases such as this, the distribution function must first be estimated. The estimated distribution function is typically called the "empirical distribution function" and for a dataset of size n, $\{x_1, x_2, x_3, \ldots, x_n\}$, the empirical distribution function places a probability of $1/n$ at each observed value of x. To state this mathematically, the empirical distribution function is defined as

$$\hat{F}(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}(x_j \leq x) .,$$

where

$$\mathbf{1}(A) = \begin{cases} 1 & \text{if true} \\ 0 & \text{otherwise} \end{cases}$$

Below is a graph of how the empirical distribution function (e.d.f.) would look if $\mathbf{x} = (1.2, 3.4, 5.2, 9.8, 10, 13.4)$.



Although the principle and calculation of the e.d.f. is straight-forward (and may seem pointless to introduce), it is critical to understand that it is with this distribution function that many theoretical questions about the bootstrap are answered. People often ask: "How can the bootstrap really be proven?" and "What distribution are you really sampling from?" The answer is what you just read: In the bootstrap, you are continuously sampling from the empirical distribution function.

2 Estimating the Variability of $\hat{\theta}$

Recall that, for now, we are assuming that θ is some population parameter that is estimated from the data via some complicated function of the data. We'll write $\hat{\theta} = g(\mathbf{X})$. where $g()$ is some complicated function. Since calculating $\text{Var}(\hat{\theta})$ is hard to do analytically, we'll do it numerically via the bootstrap. The bootstrap estimates $\text{Var}(\hat{\theta})$ by first estimating the sampling distribution of $\hat{\theta}$. And this sampling distribution is estimated by generating B samples of size n from the original sample of $\mathbf{X}$. The details of this are given below.

Generating Sampling Distribution of $\hat{\theta}$.

```
For $i = 1 : B$ {

  1. Resample $\mathbf{X}$ with replacement to generate bootstrap sample $\mathbf{X}_i^* = \left(X_{i,1}^*, X_{i,2}^*, \ldots, X_{i,n}^*\right)$.
```

```
    2. Calculate an estimate of $\hat{\theta}$ based on this sample, $\hat{\theta}_i^* = g(\mathbf{x}_i^*)$.
}
```

After implementing this algorithm, you'll have B values of $\hat{\theta}$ and the distribution of these values should approximate the sampling distribution of $\hat{\theta}$. The variability of $\hat{\theta}$ can be estimated from this sample with

$$\widehat{\text{Var}}_{\text{bootstrap}}\left(\hat{\theta}\right) = \frac{1}{B-1}\sum_{j=1}^{B}\left(\hat{\theta}_i^* - \hat{\theta}(\cdot)\right)^2,$$

where $\hat{\theta}(\cdot) = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_i^*$. Below is a very simple example of how the bootstrap can be implemented to estimate the variability of the estimator of the median.

Example 1 Let $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)$ where $X_1 = 1, X_2 = 5, X_3 = 3, X_4 = 8$, and $X_5 = 7$. We will let $B = 3$. (We'll do the bootstrap three times). Let's say

$$\mathbf{X}_1^* = (5,3,1,3,8)\,, \quad \mathbf{X}_2^* = (8,8,1,3,7)\,, \quad \mathbf{X}_3^* = (5,1,8,7,7)\,.$$

The medians of these respective samples are 3, 7, 7, making $\widehat{\text{Var}}\left(\widehat{\text{median}}\right) = 5.3$.

3 Confidence Intervals for θ

As mentioned in the beginning of this module, constructing a $100(1-\alpha)\%$ confidence interval for a parameter θ may be difficult if n is small and/or the central limit theorem doesn't apply. A confidence interval can be constructed, however, using the bootstrap. With the bootstrap algorithm provided above, B values of $\hat{\theta}$ are calculated. Again, these values of $\hat{\theta}^*$ estimate the sampling distribution of $\hat{\theta}$. A $100(1-\alpha)\%$ confidence interval for θ can easily be calculated from these values of $\hat{\theta}$ by ordering them from smallest to largest. This would give

$$\hat{\theta}_{(1)}^*, \hat{\theta}_{(2)}^*, \ldots, \hat{\theta}_{(B)}^*,$$

Just take the middle % of numbers

where $\hat{\theta}_{(j)}^*$ is the j^{th} smallest value of $\hat{\theta}^*$. A $100(1-\alpha)\%$ confidence interval can then be calculated as

$$\left(\hat{\theta}_{B\times\frac{\alpha}{2}}^*, \hat{\theta}_{B\times(1-\frac{\alpha}{2})}^*\right).$$

Below is a simple example of how this is calculated. Assume $B = 20$ and the ordered sample of $\hat{\theta}^*$ is

$$\hat{\theta}_{(1)}^* = .9, \hat{\theta}_{(2)}^* = 1.2, \hat{\theta}_{(3)}^* = 1.24, \ldots, \hat{\theta}_{(18)}^* = 2.01, \hat{\theta}_{(19)}^* = 2.1, \hat{\theta}_{(20)}^* = 2.12.$$

In this case, a 90% confidence interval for θ would be (1.2, 2.1).

4 Bias of $\hat{\theta}$

The bias of an estimator $\hat{\theta}$, conceptually, captures how far you expect your estimator to be from the true value of θ. Mathematically, the bias of an estimator is defined as

$$\text{Bias}\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta.$$

Again, this number can be very difficult to calculate analytically if $E(\hat{\theta})$ is difficult to calculate. Recall that if $\hat{\theta} = g(\mathbf{X})$, then

$$E\left(\hat{\theta}\right) = \int g(\mathbf{x})f(\mathbf{x})d\mathbf{x}, \tag{1}$$

and the integral in (1) may not be trivial (or even possible) to do. The bootstrap can help estimate the bias in $\hat{\theta}$ by, again, considering the estimated sampling distribution of $\hat{\theta}$. The expected value of $\hat{\theta}$, $E\left(\hat{\theta}\right)$, can be estimated with $\hat{\theta}\left(\cdot\right) = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^*$, and θ can be estimated with $\hat{\theta}$. The bootstrapped estimate of the bias then becomes

$$\widehat{\text{Bias}}\left(\hat{\theta}\right) = \hat{\theta}^*(\cdot) - \hat{\theta}. \tag{2}$$

Consider the first example in this module. In this example, the sample median is $\hat{\theta} = 5$. The average of the medians based on the three bootstrapped samples is 5.6, making the estimated bias $5.67 - 5 = .67$.

There is yet another way (and in often cases "better") way to estimate the bias using the bootstrap. To illustrate how to calculate this "improved" estimate of the bias, consider the dataset $\mathbf{x} = (x_1, x_2, \ldots, x_8)$ and let us assume, again, we are interested in estimating the parameter θ where

$$\theta = \int_{\mathcal{R}} g(x)f(x).$$

Estimating the bias using the traditional bootstrap can be done via Equation (2). In Equation (2) observe, of course, that in calculating the bias we use the original estimate of θ, $\hat{\theta}$. In the alternative method proposed here, we replace this original estimate of θ with $\hat{\theta}_{\text{revised}}$. The technique of calculating this revised estimate is described below.

Assume we collect B bootstrap estimates as before.

$$\begin{aligned}
\mathbf{X}_1^* &= (x_1, x_7, x_6, x_6, x_5, x_8, x_2, x_4) \\
\mathbf{X}_2^* &= (x_5, x_1, x_8, x_8, x_2, x_7, x_6, x_1) \\
\vdots \quad & \quad \vdots \\
\mathbf{X}_B^* &= (x_2, x_2, x_1, x_8, x_5, x_4, x_7, x_1).
\end{aligned}$$

From these vectors, we create another series of vectors which capture the frequency at which each observation occurs in each sample. In this case, it would be

$$\begin{aligned}
\mathbf{P}_1^* &= \left(\frac{1}{8}, \frac{1}{8}, 0, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{2}{8}, \frac{1}{8}\right) \Leftarrow \text{frequency of observations in } \mathbf{x}_1. \\
\mathbf{P}_2^* &= \left(\frac{2}{8}, \frac{1}{8}, \frac{1}{8}, 0, 0, \frac{1}{8}, \frac{1}{8}, \frac{2}{8}\right) \Leftarrow \text{frequency of observations in } \mathbf{x}_2. \\
\vdots \quad & \quad \vdots \\
\mathbf{P}_B^* &= \left(\frac{2}{8}, \frac{2}{8}, \frac{1}{8}, 0, \frac{1}{8}, \frac{1}{8}, 0, \frac{1}{8}, \frac{1}{8}\right), \Leftarrow \text{frequency of observations in } \mathbf{x}_\text{B}.
\end{aligned}$$

and from these we would calculate the average frequency vector as $\overline{\mathbf{P}}^* = \frac{1}{B}\sum_{j=1}^{B}\mathbf{P}_j^*.$. The revised original estimate of θ is then

$$\hat{\theta}_{\text{revised}} = g\left(\overline{\mathbf{P}}^*\mathbf{x}^T\right).$$

So if $\overline{\mathbf{P}}^* = (0.177, 0.234, 0.040, 0.241, 0.126, 0.094, 0.009, 0.079)$,

$$\hat{\theta}_{\text{revised}} = g\left(.177 \cdot x_1 + .234 \cdot x_2 + \cdots + .079 \cdot x_8\right).$$

The revised bias estimate is then

$$\widehat{\text{Bias}}_{\text{revised}}\left(\hat{\theta}\right) = \hat{\theta}^*(\cdot) - \hat{\theta}_{\text{revised}}.$$

5 Parametric Bootstrap

The methods you have studied so far are actually referred to as the "nonparametric bootstrap." It is called *nonparametric* because, once again, no parametric form is assumed of the distribution from which the data come from. There is such a thing as the "parametric bootstrap," and this is typically used to calculate the estimation uncertainty of complicated statistics of data coming from a known disribution. An example is provided below.

- Example: Assume you have $X_1, X_2, \ldots, X_n \sim \text{Exp}(\lambda)$. and you wish to estimate the 75^{th} percentile and provide some confidence bounds for your estimate. In the parametric bootstrap, one would caclulate the maximum likelihood of λ, $\hat{\lambda}_{\text{MLE}}$, and then generate from the corresponding exponential distribution B times. The details of the algorithm are given below:

  ```
  For $i = 1 : B$ {
      1. Generate a sample of size $n$, $\mathbf{X}_i^* = \left(X_{1,i}^*, X_{2,i}^*, \ldots, X_{n,i}^*\right)' \sim \text{Exp}\left(\hat{\lambda}_{\text{MLE}}\right)$.
      2. Calculate an estimate of the parameter of interest, $\hat{\theta}_i^* = g\left(\mathbf{X}_i^*\right)$. (In this case, $\hat{\theta}_i^* = 75^{\text{th}}$ percentile of $X_{1,i}^*, X_{2,i}^*, \ldots, X_{n,i}^*$
  }
  ```

  With the B estimates of θ provided, $(\theta_1^*, \theta_2^*, \ldots, \theta_B^*)$ a confidence interval for θ could be constructed by simply ordering the values from smallest to largest and taking the appropriate percentiles.