

# Chapter 8: Bootstrapping

Typically in statistics, we use **theory** to derive the sampling distribution of a statistic. From the sampling distribution, we can obtain the variance, construct confidence intervals, perform hypothesis tests, and more.

**Challenge:**

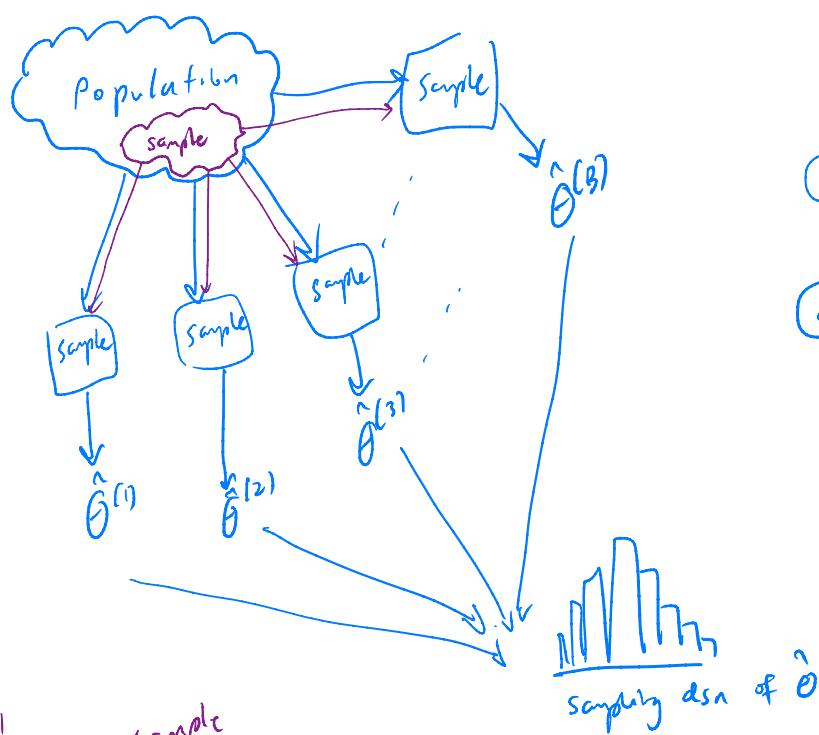
What if the sampling distribution is impossible to obtain or asymptotic theory doesn't hold?

**Basic idea of bootstrapping:**

- Use the data to estimate the sampling distribution of the statistic
- Estimate the sampling distribution by creating a large number of datasets that we might have seen and compute the statistic on each of these datasets.

"pull yourself up by your bootstraps"

E.g.



In reality, we only have a sample and we need to use it to make sample<sup>(1)</sup>, ..., sample<sup>(B)</sup>

## Goals of bootstrapping

- estimate bias, se, and CIs when
- ① There is doubt about whether distributional assumptions are met.
  - ② There is doubt about whether asymptotic results are valid
  - ③ No theory to derive the dsn of the test statistic is too hard.

# 1 Nonparametric Bootstrap

Let  $X_1, \dots, X_n \sim F$  with pdf  $f(x)$ . Recall, the cdf is defined as

$$F(x) = \int_{-\infty}^x f(t) dt = P(X \leq x)$$

**Definition 1.1** The *empirical cdf* is a function which estimates the cdf using observed data,

$$\hat{F}(x) = E_n(x) = \text{proportion of sample points that fall in } (\infty, x].$$

*depends on the data*

In practice, this leads to the following function. Let  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$  be the order statistics of the sample. Then,

*= sample in order*

$$F_n(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{i}{n} & X_{(i)} \leq x < X_{(i+1)}; \quad i = 1, \dots, n-1 \\ 1 & x \geq X_{(n)} \end{cases}$$

$F_n(x)$  is an estimator of  $F$   
cdf  
 and as  $n \uparrow \infty$ ,  $F_n \rightarrow F$ .

Theoretical: Sample  $X \sim F$ , use  $x_1, \dots, x_n$  to compute  $F_n$

Bootstrap: Sample  $X^* \sim F_n$ , use  $x_1^*, \dots, x_n^*$  to compute  $F_n^*$

**Example 1.1** Let  $x = 2, 2, 1, 1, 5, 4, 4, 3, 1, 2$  be an observed sample. Find  $F_n(x)$ .

$$\text{sorted} = 1, 1, 1, 2, 2, 2, 3, 4, 4, 5 \quad n=10$$

$$F_n(x) = \begin{cases} 0 & x < 1 \\ 3/10 & 1 \leq x < 2 \\ 6/10 & 2 \leq x < 3 \\ 7/10 & 3 \leq x < 4 \\ 9/10 & 4 \leq x < 5 \\ 1 & x \geq 5 \end{cases}$$

There is an easy way to sample from  $F_n$  without calculating it.<sup>2</sup>

The idea behind the bootstrap is to sample many data sets from  $F_n(x)$ , which can be achieved by resampling from the data **with replacement**.

```
# observed data
x <- c(2, 2, 1, 1, 5, 4, 4, 3, 1, 2)  $\downarrow x = (x_1, \dots, x_n)$ 

# create 10 bootstrap samples
x_star <- matrix(NA, nrow = length(x), ncol = 10)
for(i in 1:10) {
  x_star[, i] <- sample(x, length(x), replace = TRUE)  $\downarrow$  key part of the bootstrap
}
x_star  $\uparrow$  sample from  $F_n(x)$ .
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]     5    2    4    4    1    2    2    1    5    1
## [2,]     4    5    1    1    1    2    1    1    4    2
## [3,]     4    2    5    1    2    2    1    4    4    3
## [4,]     4    5    1    3    2    4    4    4    3    1
## [5,]     4    1    2    1    1    1    5    2    1    1
## [6,]     4    2    2    2    4    4    3    2    1    2
## [7,]     1    5    4    4    1    2    1    2    1    4
## [8,]     3    1    1    1    4    1    4    1    4    2
## [9,]     1    4    4    2    2    1    4    3    2    1
## [10,]    4    1    2    3    4    5    5    5    2    4
```

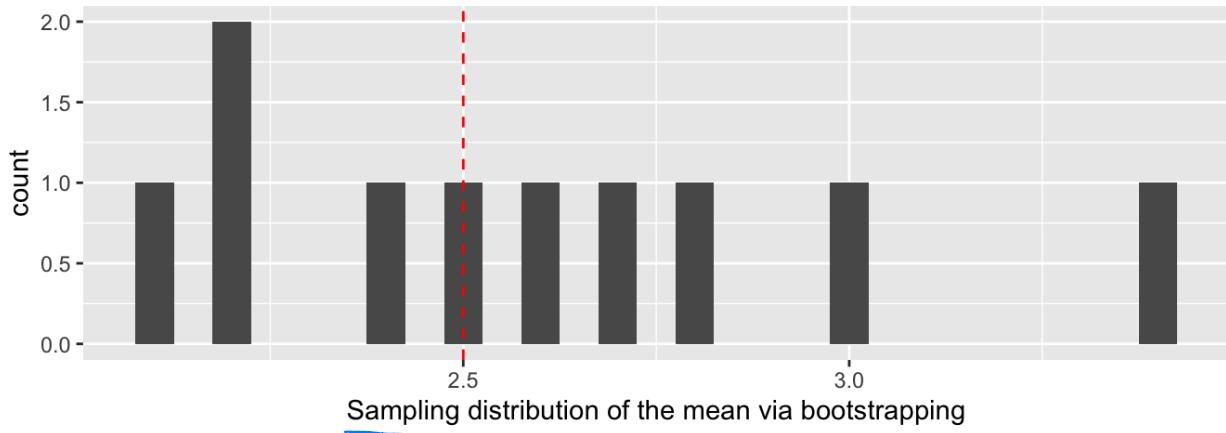
$\uparrow$   
 $x^*$

# compare mean of the same to the means of the bootstrap samples  
 $\text{mean}(x)$

$\# [1] 2.5 \leftarrow \bar{x}$

$\text{colMeans}(x_{\text{star}})$

```
## [1] 3.4 2.8 2.6 2.2 2.2 2.4 3.0 2.5 2.7 2.1
 $\uparrow$   $\bar{x}^*$ 
ggplot() +
  geom_histogram(aes(colMeans(x_star)), binwidth = .05) +
  geom_vline(aes(xintercept = mean(x)), lty = 2, colour = "red") +
  xlab("Sampling distribution of the mean via bootstrapping")
```



## 1.1 Algorithm

**Goal:** estimate the sampling distribution of a statistic based on observed data  $x_1, \dots, x_n$ .

Let  $\theta$  be the parameter of interest and  $\hat{\theta}$  be an estimator of  $\theta$ . Then,

For  $b=1, \dots, B \leftarrow \# \text{ bootstrap samples}$

① sample  $x^{*(b)} = (x_1^{*(b)}, \dots, x_n^{*(b)})$  by sampling w/ replacement from the observed data  
(i.e. sample from  $F_n$ )

②  $\hat{\theta}^{(b)} = \hat{\theta}(x^{*(b)})$   
estimate of  $\theta$  based on  $b^{\text{th}}$  bootstrap sample.

Using  $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)}$ , we can

- estimate the sampling dsn of the statistic  $\hat{\theta}$   
↳ make a histogram of  $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)}$

- Estimate the standard error of  $\hat{\theta}$   
↳ compute the st. deviation of  $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)}$

- Estimate a CI  
↳ we'll cover multiple methods.

- Estimate many other things.

## 1.2 Properties of Estimators

We can use the bootstrap to estimate different properties of estimators.

### 1.2.1 Standard Error

Recall  $se(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$ . We can get a **bootstrap** estimate of the standard error:

$$\hat{se}(\hat{\theta}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^{(b)} - \bar{\hat{\theta}}^*)^2}$$

where  $\bar{\hat{\theta}}^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{(b)}$

### 1.2.2 Bias

Recall  $bias(\hat{\theta}) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta$ .

#### Example 1.2

$$\begin{aligned} E[\hat{\sigma}^2] &= E\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right] = (1 - \frac{1}{n}) \sigma^2 \\ \Rightarrow bias[\hat{\sigma}^2] &= E[\hat{\sigma}^2] - \sigma^2 = (1 - \frac{1}{n}) \sigma^2 - \sigma^2 = -\frac{1}{n} \sigma^2 \\ \Rightarrow \text{we use } s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, E[s^2] = \sigma^2 \text{ (unbiased)} \end{aligned}$$

We can get a **bootstrap** estimate of the bias:

$$\hat{bias}(\hat{\theta}) = \bar{\hat{\theta}}^* - \hat{\theta} = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}^{(b)} - \hat{\theta})$$

Computed from bs samples      based on original data

If  $\hat{bias}(\hat{\theta}) > 0$ , the  $\hat{\theta}$  overestimates  $\theta$ , on average.

Overall, we seek statistics with small se and small bias.

but there typically is a bias/variance trade off as bias ↓, se ↑.

### 1.3 Sample Size and # Bootstrap Samples

$$n = \text{sample size} \quad \& \quad B = \# \text{ bootstrap samples}$$

If  $n$  is too small, or sample isn't representative of the population,

*then bootstrap results will be poor no matter how large  $B$  is.*

Guidelines for  $B$  –

$$B \approx 1000 \text{ for } \text{se} \text{ & bias}$$

$$B \approx 2000 \text{ for CI's (depends on } \alpha: \text{ small } \alpha \Rightarrow \uparrow B)$$

Best approach –

*Repeat bootstrap twice w/ different seeds.*

*If estimate are very different,  $\uparrow B$ .*

## Your Turn

In this example, we explore bootstrapping in the rare case where we know the values for the entire population. If you have all the data from the population, you don't need to bootstrap (or really, inference). It is useful to learn about bootstrapping by comparing to the truth in this example.

In the package `bootstrap` is contained the average LSAT and GPA for admission to the population of 82 USA Law schools (an old data set – there are now over 200 law schools). This package also contains a random sample of size  $n = 15$  from this dataset.

```
library(bootstrap)
```

```
head(law)
```

*↑ random sample of size n=15.*

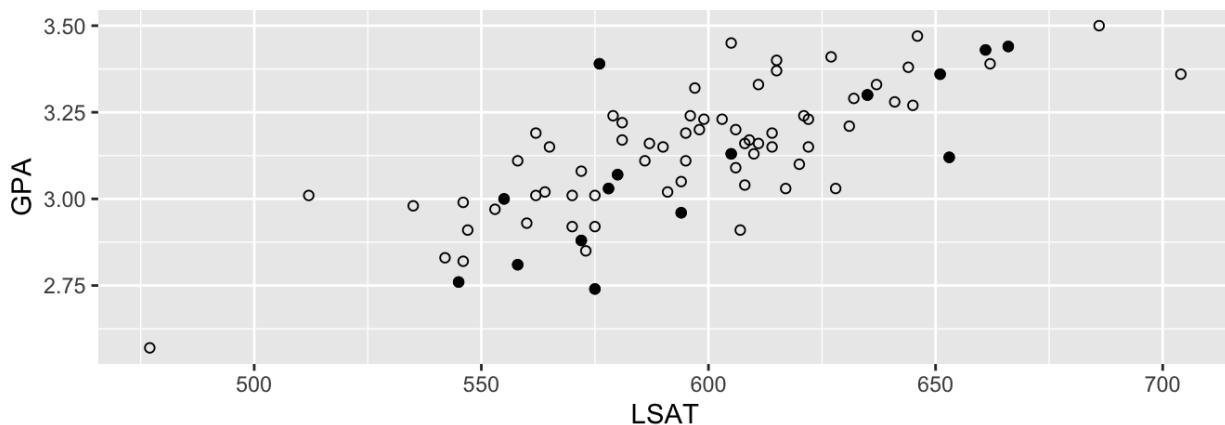
```
##   LSAT   GPA
## 1 576 3.39
## 2 635 3.30
## 3 558 2.81
## 4 578 3.03
## 5 666 3.44
## 6 580 3.07
```

```
ggplot() +
```

```
  geom_point(aes(LSAT, GPA), data = law) +
```

```
  geom_point(aes(LSAT, GPA), data = law82, pch = 1)
```

*↑ Full data set.*



$$\text{Recall } \hat{\rho} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

*= Correlation.*

We will estimate the correlation  $\theta = \rho(\text{LSAT}, \text{GPA})$  between these two variables and use a bootstrap to estimate the sample distribution of  $\hat{\theta}$ .

```
# sample correlation
cor(law$LSAT, law$GPA)
```

```
## [1] 0.7763745
```

```
# population correlation
cor(law82$LSAT, law82$GPA)
```

```
## [1] 0.7599979
```

```
# set up the bootstrap
B <- 200
n <- nrow(law)
r <- numeric(B) # storage for replicates

for(b in B) {
  ## Your Turn: Do the bootstrap!
}
```

*we know this because we have the population!*

1. Plot the sample distribution of  $\hat{\theta}$ . Add vertical lines for the true value  $\theta$  and the sample estimate  $\hat{\theta}$ .
2. Estimate  $sd(\hat{\theta})$ .
3. Estimate the bias of  $\hat{\theta}$