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:

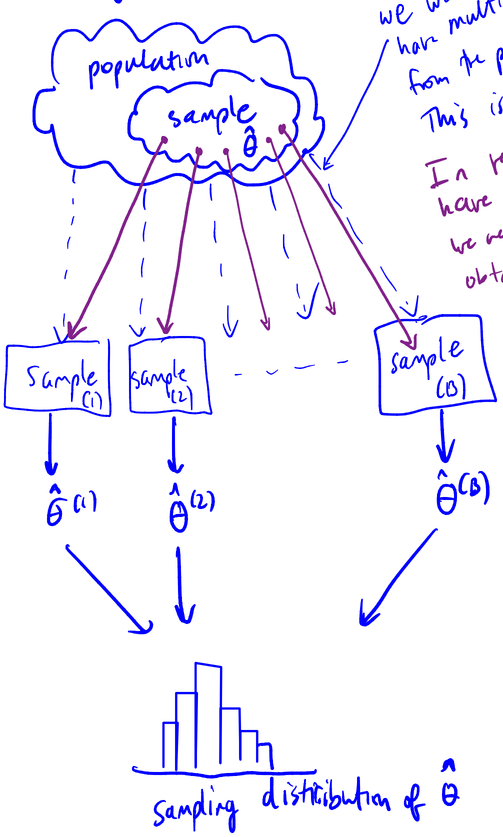
what if the sampling distribution is impossible to obtain or asymptotic theory doesn't hold?

Basic idea of bootstrapping:

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

"Pull ourselves up by our bootstraps"

Eg:



we would love to have multiple samples from the population. This is unrealistic

In reality we only have one sample and we need to use it to obtain sample$^{(1)}$, ..., sample$^{(B)}$.

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. the theory to derive the dsn of the test statistic is too hard.

Do not make distributional assumption.

1 Nonparametric Bootstrap

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

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

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

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

"depends on the data"

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

"min$(x_1, \ldots, x_n)$" max$(x_1, \ldots, x_n)$.

sample in order of value.

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

ECDF $F_n(x)$ is an estimator of cdf F

and as $n \uparrow \infty$, $F_n \to F$

Theoretical: Sample $X \sim F$, use $X_1, \ldots, X_n$ to compute F_n

Bootstrap: Sample $X^* \sim F_n$, use $X_1^*, \ldots, X_n^*$ to compute $\boxed{F_n^*}$

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

order statistic

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

ecdf in r.

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

– There is an easy way to sample from F_n if we've calculated.

$\ast$ – There is an easy way to sample from F_n without calculating F_n.

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**.

$x = \{x_1, \dots, x_{10}\}$.

```
# observed data
x <- c(2, 2, 1, 1, 5, 4, 4, 3, 1, 2)
```

key part of the bootstrap.

data

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)
}
x_star
```

sample of size n 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
```

$X^{*(1)}$ $X^{*(2)}$ $X^{*(10)}$.

```
# compare mean of the same to the means of the bootstrap samples
mean(x)
```

original sample.

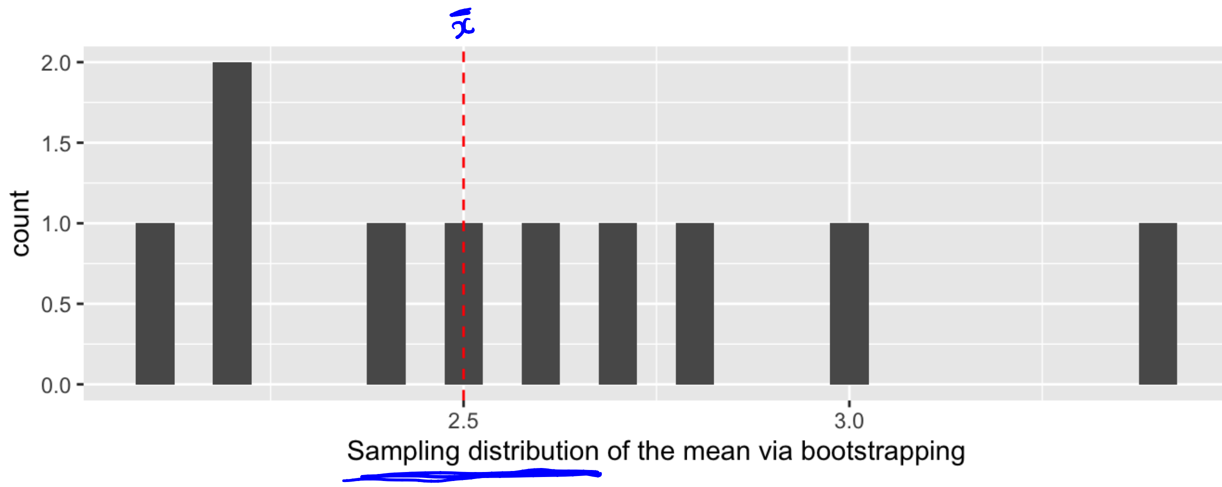
```
## [1] 2.5
```

```
colMeans(x_star)
```

$\bar{x}^{*(i)}$.

```
##  [1] 3.4 2.8 2.6 2.2 2.2 2.4 3.0 2.5 2.7 2.1
```

```
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 $\hat{\theta}$ based on observed data $x_1, \ldots, x_n$. one sample of size n

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

For $b = 1, \ldots, \boxed{B}$ # Bootstrap samples

1. sample $x^{*(b)} = (x_1^{*(b)}, \ldots, x_n^{*(b)})$ by sampling with replacement from the observed data (i.e. sample from F_n).
2. $\hat{\theta}^{(b)} = \hat{\theta}(x^{*(b)})$

estimate of θ based on b^{th} bootstrap sample.

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

- estimate the sampling distribution of the statistic $\hat{\theta}$
  - ↳ make a histogram of $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$.
- estimate the se of $\hat{\theta}$
  - ↳ compute the sd of $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$

SE = st dev of sampling distribution.

- estimate the CI
  - ↳ we'll cover multiple ways.
- estimate the bias of $\hat{\theta}$

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:

$$\hat{se}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}^{(b)} - \bar{\hat{\theta}}\right)^2} \leftarrow \text{sample st. dev. of } \hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$$

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

1.2.2 Bias

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

Example 1.2

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \left(1 - \frac{1}{n}\right)\sigma^2$$

$$Bias(\hat{\sigma}^2) = E[\hat{\sigma}^2] - \sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2.$$

$\Rightarrow$ we use $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $E(s^2) = \sigma^2$ (unbiased).

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

$$\hat{bias}(\hat{\theta}) = \underbrace{\bar{\hat{\theta}}}_{\text{computed from bs samples}} - \underbrace{\hat{\theta}}_{\text{computed from original data.}} = \frac{1}{B}\sum_{b=1}^{B}\left(\underbrace{\hat{\theta}^{(b)}}_{\text{bootstrap estimates of }\theta} - \underbrace{\hat{\theta}}_{\text{estimate of }\theta\text{ from original sample}}\right)$$

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

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

but there is typically is a bias/variance tradeoff as bias $\downarrow$, se $\uparrow$

1.3 Sample Size and # Bootstrap Samples

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

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

the bootstrap results will be poor no matter how large B is.

Guidelines for B –

$B \approx 1000$ for se & bias

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

Best approach –

Repeat bootstrap twice w/ different seeds

If estimates 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.

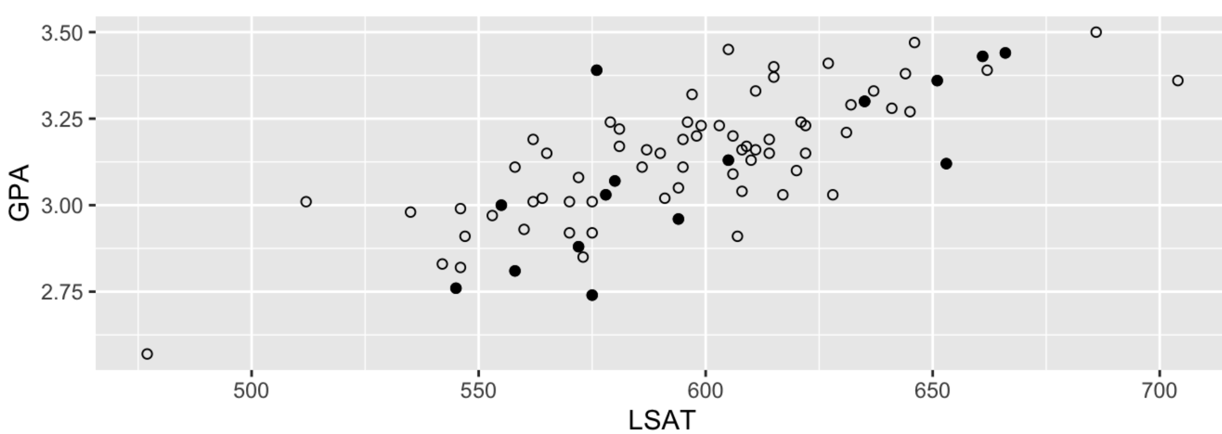
```
library(bootstrap)

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)
```

population



correlation

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

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 $\hat{\theta}$.

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

$\hat{\rho} =$

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

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

$\rho =$

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

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

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

① Sampling from F_n
② Computing $\hat{\rho}$ on bootstrap sample

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