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:

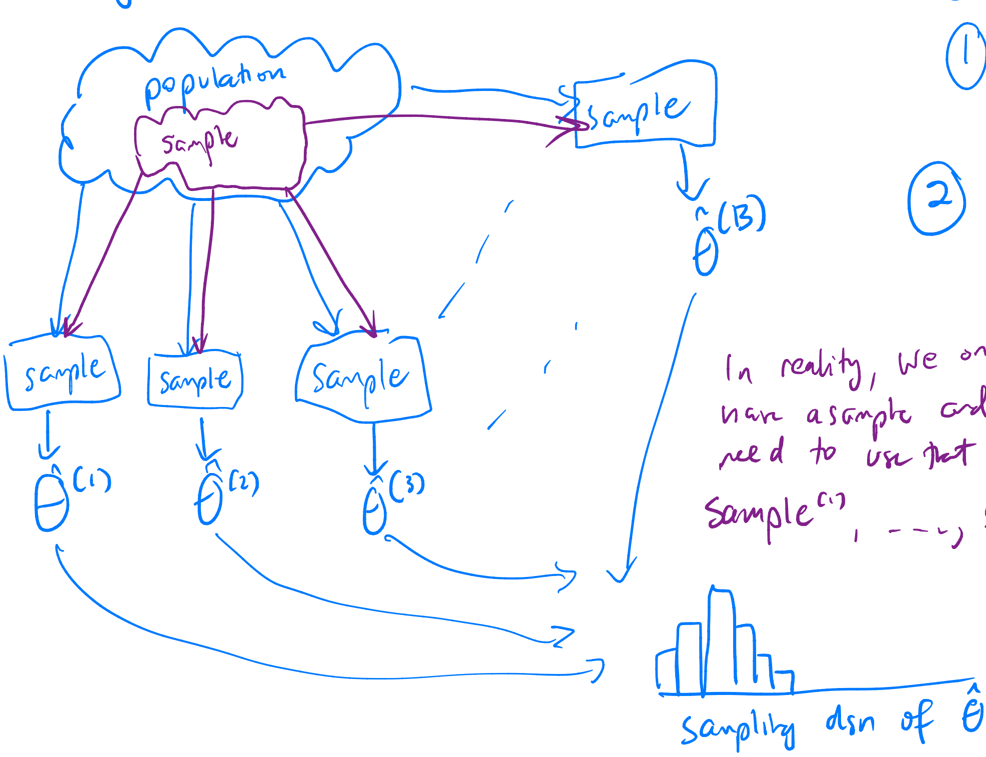
What if the sampling distribution is impossible to obtain or asymptotic theory doesn't hold?

Basic idea of bootstrapping:

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

"pull yourself up by your bootstraps"

Eg.



In reality, we only have a sample and we need to use that to make 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. When the theory to derive the dsn of the test statistic is too hard.

not making distribution assumptions

1 Nonparametric Bootstrap

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

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

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

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

"depends on data sample of size n"

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

Sample in order

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

$\min(X_1, \ldots, X_n)$

$\max(X_1, \ldots, X_n)$

$F_n(x)$ is an estimator of $F(x)$
ecdf — cdf
and as $n \uparrow \infty$, $F_n(x) \to F$

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 F_n^*

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

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

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

There is an easier way to sample from F_n without calculating it.

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**.

easy way to sample from F_n without calculating it.

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

# 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
```

key part of the bootstrap

sample data

sampling 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
```

$x^* \sim F_n$

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

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

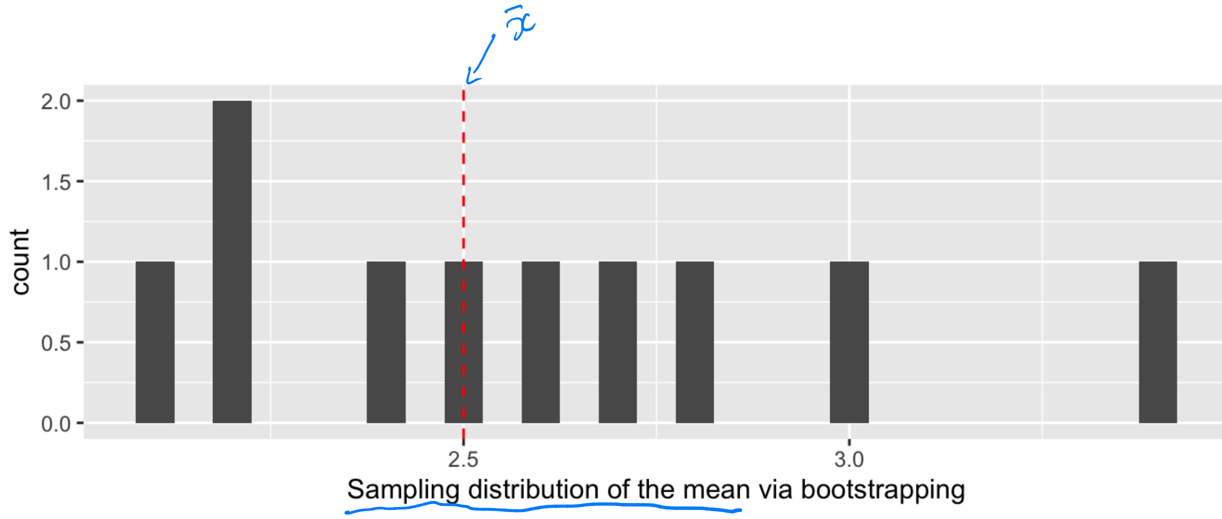
$\leftarrow \bar{x}$

```
colMeans(x_star)
```

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

$\bar{x}^*$

```
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 $x_1, \ldots, x_n$. ← # of observations.

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

For $b = 1, \ldots, B$ ← # 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 the ecdf $F_n(x)$).

2. $\hat{\theta}^{(b)} = \hat{\theta}(x^{*(b)})$

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

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

- estimate the sampling dsn of the statistic $\hat{\theta}$
  - ↳ make a histogram of $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$
- estimate the s.e. of $\hat{\theta}$
  - ↳ compute the st. deviation of $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$
- 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:

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

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

1.2.2 Bias

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

Example 1.2

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

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

$\Rightarrow$ 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:

$$\hat{\text{bias}}(\hat{\theta}) = \overline{\hat{\theta}}^{*} - \hat{\theta} = \frac{1}{B} \sum_{i=1}^{B} \left(\hat{\theta}^{(b)} - \hat{\theta} \right)$$

computed from bs samples

based on original data

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

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

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

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,

the bootstrap results will be poor no matter what B we choose.

Guidelines for B –

$B \approx 1000$ for se & bias

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

Best approach –

Repeat bootstrap twice w/ different seed

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.

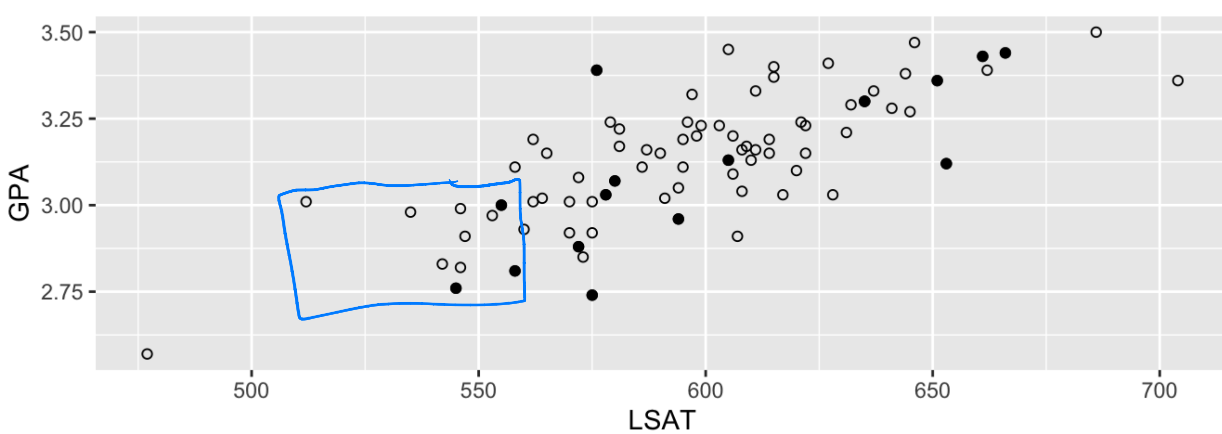
```
library(bootstrap)

head(law)
```

↖ sample of 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

$= \frac{E((LSAT - E LSAT)(GPA - E GPA))}{\text{normalized}}$

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

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

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

1. Plot the sample distribution of $\hat{\theta}$. Add vertical lines for the true value θ and the sample estimate $\hat{\theta}$.

2. Estimate $sd(\hat{\theta})$.

3. Estimate the bias of $\hat{\theta}$