Bootstrap Methods

Typically we use (asymptotic) 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:

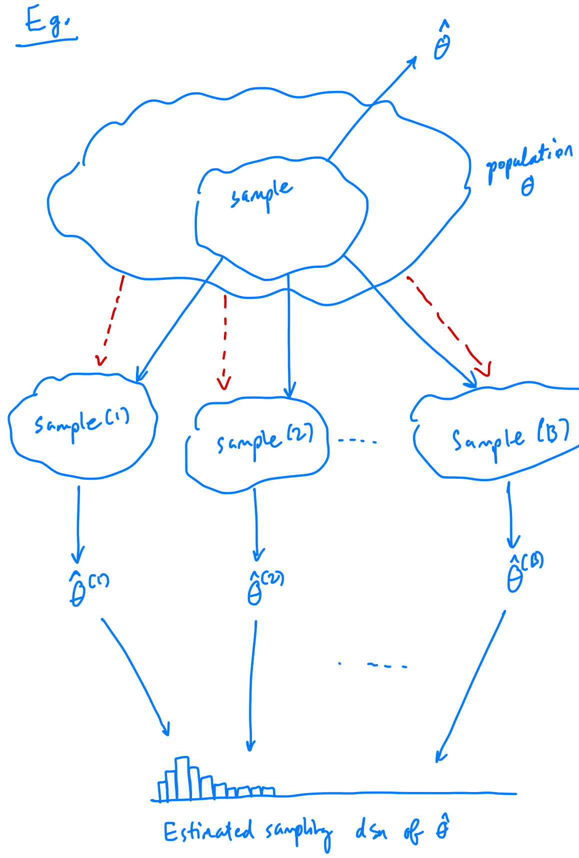
what if the sampling is impossible to obtain or asymptotic theory doesn't hold?

Basic idea of bootstrapping:

Use the data to approximate the sampling distribution of the statistic.

How? Estimate the sampling distribution by creating a large # of datasets that we might have seen and compute the statistic on each of these data sets.

Eg.



Goals:

estimate bias, se, CI's when

1. There is doubt about whether distributional assumptions are met.
2. There is doubt about whether asymptotic results are valid.
3. Theory to derive dsn is too hard.

In reality, we only have a sample, need to make sample$^{(1)}$, ..., sample$^{(B)}$

"Bootstrap World" where the data analyst knows everything.

idea: treat the sample $Y_1, \dots, Y_n$ as the population.

E.g., we are interested in the variance of an estimator

- In "bootstrap world" we can calculate the exact variance b/c we have access to the "population"
- In practice, estimate variance by repeatedly sampling from the pseudo-population.

Real world

True population $Y_1, Y_2, \dots$ w/ dsn F_θ

True pop. parameter θ

↓ sample

$Y_1, \dots, Y_n \Rightarrow \hat{\theta}(Y_1, \dots, Y_n)$ is estimator

$$MSE = E_F\left[(\hat{\theta} - \theta)^2\right]$$

If we don't have access to F (we don't), we can't take this expectation.

If we had access to the population (we don't!) could estimate MSE w/

$$\hat{MSE} = \frac{1}{rep} \sum_{i=1}^{rep} (\hat{\theta}_i - \theta)^2$$

Bootstrap world

$Y_1, \dots, Y_n$ is population

$\hat{\theta}$ is true value of parameter

Since we have access to the population,

$$\hat{MSE}_{Boot} = \frac{1}{Boot\ Rep} \sum_{i=1}^{Boot\ Rep} (\hat{\theta}_i - \hat{\theta})^2$$

We hope $\hat{MSE}_{Boot} \approx MSE$

1 Nonparametric Bootstrap

Let $Y_1, \ldots, Y_n \sim F$ with pdf $f(y)$. Recall, the empirical cdf is defined as

$\underline{Y} = (Y_1, \ldots, Y_n)$

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(Y_i \le y) \quad y \in \mathbb{R}.$$

↑ MLE of F and as $n \to \infty$, $F_n \to F$

Theoretical: Sample $\underline{Y} \sim F$, use $Y_1, \ldots, Y_n$ to compute F_n

Bootstrap: Sample $\underline{Y}^* \sim F_n$, use $Y_1^*, \ldots, Y_n^*$ to compute F_n^*

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

How many possible bootstrap samples? n^n

Are $Y_1^*, \ldots, Y_n^*$ independent?

$$P(Y_1^* = a, Y_2^* = b) = \frac{\sum_{i=1}^{n} \mathbb{I}(Y_i^* = a)}{n} \cdot \frac{\sum_{i=1}^{n} \mathbb{I}(Y_2^* = b)}{n} = P(Y_1^* = a)\,P(Y_2^* = b) \Rightarrow \text{yes}$$

Do we always want this?

No! More later

```r
# 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
```

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

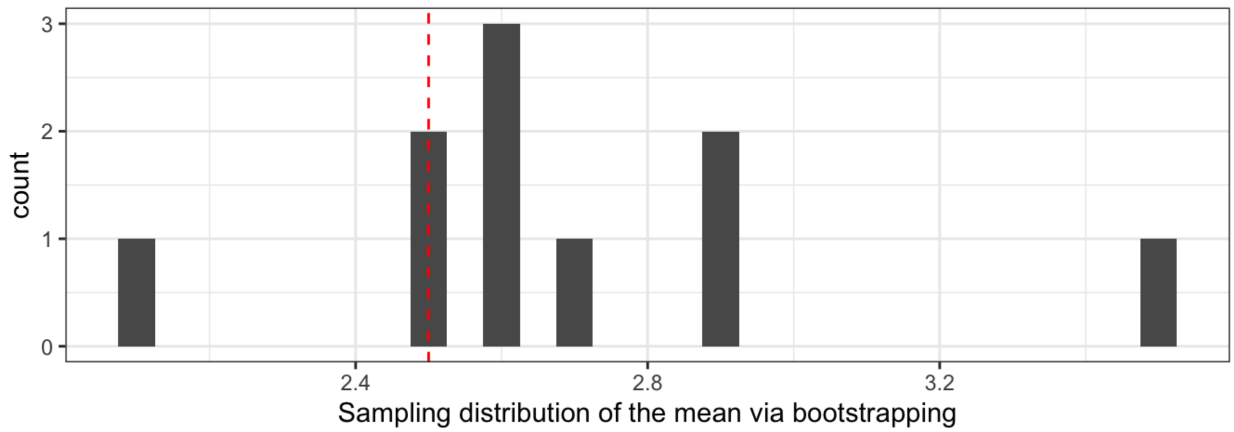
```r
# compare mean of the sample to the means of the bootstrap samples
mean(x)
```

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

```r
colMeans(x_star)
```

```
##  [1] 2.9 2.7 2.6 2.1 2.9 2.6 3.5 2.5 2.5 2.6
```

```r
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 of NP bootstrap for iid data.

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

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

For $b = 1, \ldots, B$

① Sample $y^{*(b)} = (y_1^{*(b)}, \ldots, y_n^{*(b)})$ by sampling w/ replacement from the sample data (i.e. sample from F_n)

② Compute $\hat{\theta}^{*(b)} = \hat{\theta}(y^{*(b)})$

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

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

- estimate the sampling dsn of $\hat{\theta}$ (histogram of $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$).
- estimate the SE of $\hat{\theta}$ (st. dev. of $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$).
- estimate a CI (many ways).

etc.

1.2 Justification for iid data

Suppose $Y_1, \ldots, Y_n$ are iid with $\mathrm{E}Y_1 = \mu \in \mathbb{R}$, $\mathrm{Var}(Y_1) = \sigma^2 \in (0, \infty)$. Let's approximate the distribution of $T_n = \sqrt{n}(\bar{Y}_n - \mu)$ via the bootstrap.

Theorem: If $Y_1, Y_2, \ldots$ are iid with $\mathrm{Var}(Y_1) = \sigma^2 \in (0, \infty)$, then $\sup_{y \in \mathbb{R}} |P(T_n \le y) - P_*(T_n^* \le y)| \equiv \Delta_n \to 0$ as $n \to \infty$ almost surely (a.s).

Given $\underline{Y} = \{Y_1, \ldots, Y_n\}$ draw $Y_1^*, \ldots, Y_n^*$ bootstrap sample. Then,

bootstrap probability. $\rightarrow P_*(Y_i^* = Y_i) = P(Y_i^* = Y_i \mid \underline{Y}) = \frac{1}{n}$ $1 \le i \le n$

The bootstrap version of T_n is $T_n^* = \sqrt{n}(\bar{Y}_n^* - E_* Y_i^*) = \sqrt{n}(\bar{Y}_n^* - \bar{Y}_n)$

where $E_*[Y_i^*] = E[Y_i^* \mid \underline{Y}] = \sum_{i=1}^n \frac{1}{n} Y_i = \bar{Y}_n$ also $E_*(\bar{Y}_n^*) = E_*\left(\frac{1}{n}\sum_{i=1}^n Y_i^*\right) = \frac{1}{n}\sum_{i=1}^n E_* Y_i^* = \bar{Y}_n$

bootstrap expected value

exists b/c dsn of $Y_1^*, \ldots, Y_n^*$ exist, but hard to compute directly b/c n^n bootstrap samples $\Rightarrow$ use simulation to estimate

Also, $P_*(T_n^* \le y) = P(T_n^* \le y \mid \underline{Y})$ approximates $P(T_n \le y)$, $y \in \mathbb{R}$ (Theorem).

The proof of this theorem requires two facts:

i. (Berry-Esseen Lemma) Let $Y_1, \ldots, Y_n$ be independent with $\mathrm{E}Y_i = 0$ and $\mathrm{E}|Y_i|^3 < \infty$ for $i = 1, \ldots, n$. Let $\sigma_n^2 = n\mathrm{Var}(\bar{Y}_n) = n^{-1}\sum_{i=1}^n \mathrm{E}Y_i^2 > 0$. Then,

$$\sup_{y \in \mathbb{R}} \left| P\left(\frac{\sqrt{n}\bar{Y}_n}{\sigma_n} \le y\right) - \Phi(y) \right| = \sup_{x \in \mathbb{R}} \left| P\left(\sqrt{n}\bar{Y}_n \le x\right) - \Phi\left(\frac{x}{\sigma_n}\right) \right| \le \frac{2.75}{n^{3/2}\sigma_n^3} \sum_{i=1}^n \mathrm{E}|Y_i|^3.$$

M-Z

ii. (Marcinkiewicz-Zygmund SLLN) Let X_i be a sequence of iid random variables with $\mathrm{E}|X_i|^p < \infty$ for $p \in (0, 2)$. Then, for $S_n = \sum_{i=1}^n X_i$,

$$\frac{1}{n^{1/p}}(S_n - nc) \to 0 \text{ as } n \to \infty \text{ almost surely } (*)$$

for any $c \in \mathbb{R}$ if $p \in (0, 1)$ and for $c = \mathrm{E}X_1$ if $p \in [1, 2)$. If $(*)$ holds for some $c \in \mathbb{R}$, then $\mathrm{E}|X_1|^p < \infty$.

Specifically, we will use that if $\{Y_i\}$ are iid w/ $EY_1^2 < \infty$, then

$$\frac{1}{n^{3/2}} \sum_{i=1}^n |Y_i|^3 \to 0 \text{ as } n \to \infty \text{ a.s.}$$

letting $X_i = |Y_i|^3$ because $E|X_1|^p = E|Y_1|^{3p} < \infty$ for $p = 2/3$ we may take $c = 0$.

Proof:

Write $\sup_{y\in\mathbb{R}} |P(T_n \le y) - P_*(T_n^* \le y)| \le \underbrace{\sup_{y\in\mathbb{R}} |P(T_n \le y) - \Phi(y/\sigma)|}_{\tilde{\Delta}_n} + \underbrace{\sup_{y\in\mathbb{R}} |P_*(T_n^* \le y) - \Phi(y/\sigma)|}_{\Delta_{n1}}$

$\tilde{\Delta}_n \to 0$ by CLT since $Y_1, \ldots, Y_n$ iid, $EY_1^2 < \infty$.

Note that

$$\sigma_{n*}^2 \equiv n \mathrm{Var}_*(\bar{Y}_n^*) = n \mathrm{Var}_*\left(\frac{1}{n}\sum_{i=1}^n Y_i^*\right) = \frac{n}{n^2}\sum_{i=1}^n \mathrm{Var}_* Y_i^* = \mathrm{Var}_* Y_1^*$$

$$= E_*(Y_1^*)^2 - [E_* Y_1^*]^2 \quad \text{where} \quad Y_1^* = \begin{cases} Y_1 & \text{w.p. } \frac{1}{n} \\ \vdots & \\ Y_n & \text{w.p. } \frac{1}{n} \end{cases}$$

$$= \frac{1}{n}\sum_{i=1}^n Y_i^2 - [\bar{Y}_n]^2$$

So, $\sigma_{n*}^2 \to EY_1^2 - (EY_1)^2 = \sigma^2$ as $n \to \infty$ w.p. by SLLN since $EY_1^2 < \infty$.

By the Berry Esseen Lemma on $T_n^* = \sqrt{n}(\bar{Y}_n^* - E_* Y_1^*)$ and $|a-b| \le 2\max\{|a|,|b|\}$

$\Rightarrow |a-b|^3 \le 8\max\{|a|^3, |b|^3\}$

$\le 8(|a|^3 + |b|^3)$

$$\sup_{y\in\mathbb{R}} \left|P_*(T_n^* \le y) - \Phi\left(\frac{y}{\sigma_{n*}}\right)\right| \overset{\text{Berry Esseen}}{\le} \frac{2.75}{n^{3/2}\sigma_{n*}^3}\, n\, E_*|Y_1^* - E_* Y_1^*|^3$$

($T_n^* = \sqrt{n}(\bar{Y}_n^* - E_* Y_1^*)$)

1.3 Properties of Estimators

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

1.3.1 Standard Error

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

1.3.2 Bias

Recall $\text{bias}(\hat{\theta}) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta$. We can get a **bootstrap** estimate of the bias:

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

1.4 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,

Guidelines for B –

Best approach –

Your Turn

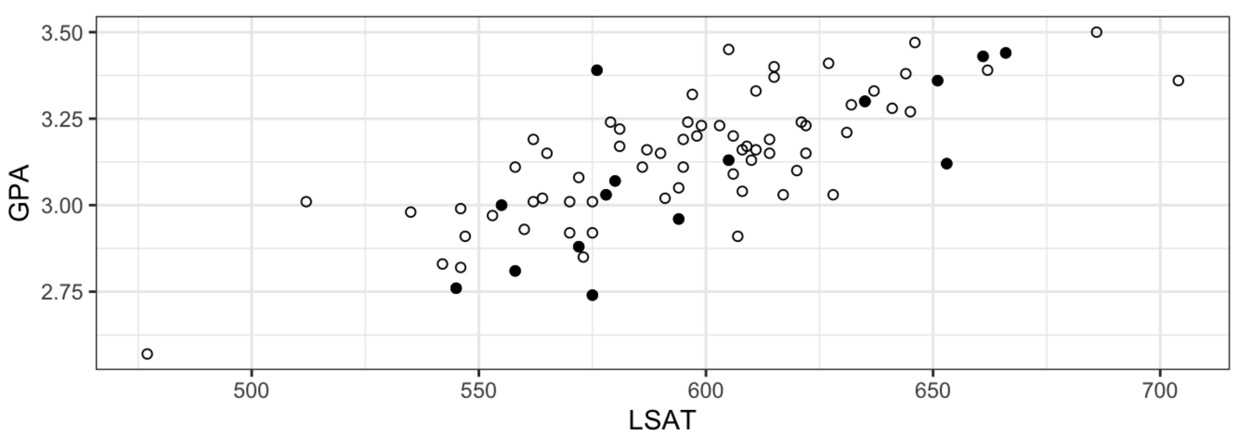
In this example, we explore bootstrapping in the rare case where we know the values for the entire population. If you have all the data from the population, you don't need to bootstrap (or really, inference). It is useful to learn about bootstrapping by comparing to the truth in this example.

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)
```

```
##   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)
```



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(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 $\hat{\theta}$

1.5 Bootstrap CIs

We will look at five different ways to create confidence intervals using the boostrap and discuss which to use when.

1. Percentile Bootstrap CI
2. Basic Bootstrap CI
3. Standard Normal Bootstrap CI
4. Bootstrap t
5. Accelerated Bias-Corrected (BCa)

Key ideas:

1.5.1 Percentile Bootstrap CI

Let $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)}$ be bootstrap replicates and let $\hat{\theta}_{\alpha/2}$ be the $\alpha/2$ quantile of $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)}$.

Then, the $100(1-\alpha)\%$ Percentile Bootstrap CI for θ is

In `R`, if `bootstrap.reps = c(`$\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)}$`)`, the percentile CI is

```
quantile(bootstrap.reps, c(alpha/2, 1 - alpha/2))
```

Assumptions/usage

1.5.2 Basic Bootstrap CI

The $100(1-\alpha)\%$ Basic Bootstrap CI for θ is

Assumptions/usage

1.5.3 Bootstrap t CI (Studentized Bootstrap)

Even if the distribution of $\hat{\theta}$ is Normal and $\hat{\theta}$ is unbiased for θ, the Normal distribution is not exactly correct for z.

Additionally, the distribution of $\hat{se}(\hat{\theta})$ is unknown.

$\Rightarrow$ The bootstrap t interval does not use a Student t distribution as the reference distribuion, instead we estimate the distribution of a "t type" statistic by resampling.

The $100(1-\alpha)\%$ Boostrap t CI is

Overview

To estimate the "t style distribution" for $\hat{\theta}$,

Assumptions/usage

1.5.4 BCa CIs

Modified version of percentile intervals that adjusts for bias of estimator and skewness of the sampling distribution.

This method automatically selects a transformation so that the normality assumption holds.

Idea:

The BCa method uses bootstrapping to estimate the bias and skewness then modifies which percentiles are chosen to get the appropriate confidence limits for a given data set.

In summary,

Your Turn

We will consider a telephone repair example from Hesterberg (2014). `Verizon` has repair times, with two groups, CLEC and ILEC, customers of the "Competitive" and "Incumbent" local exchange carrier.

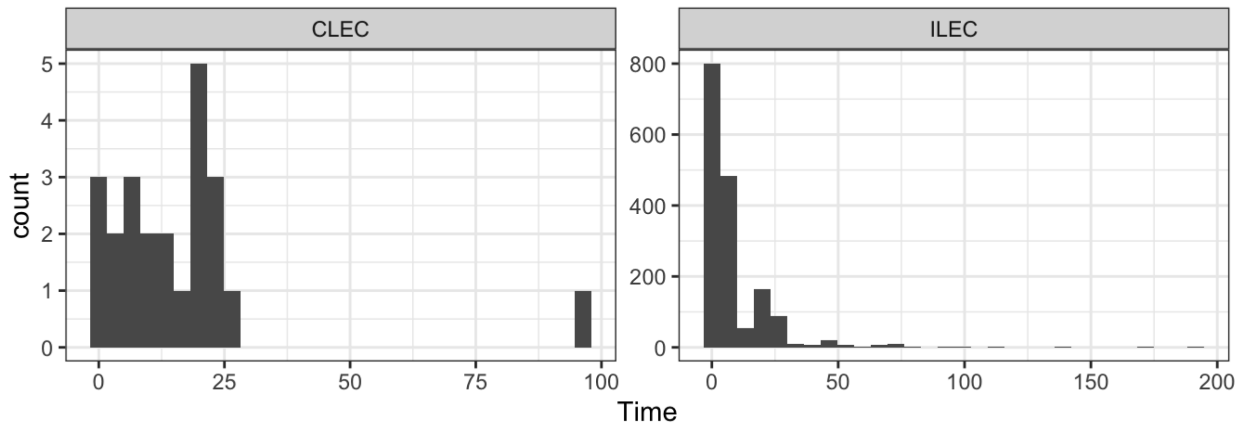
```
library(resample) # package containing the data

data(Verizon)
head(Verizon)
```

```
##    Time Group
## 1 17.50  ILEC
## 2  2.40  ILEC
## 3  0.00  ILEC
## 4  0.65  ILEC
## 5 22.23  ILEC
## 6  1.20  ILEC
```

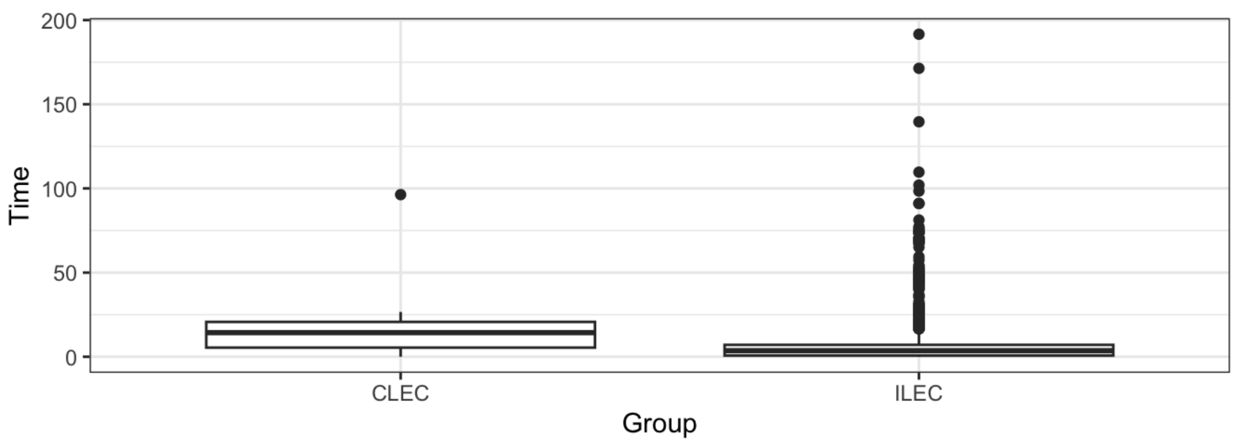
```
Verizon |>
  group_by(Group) |>
  summarize(mean = mean(Time), sd = sd(Time), min = min(Time), max =
        max(Time)) |>
  kable()
```

Group	mean	sd	min	max
CLEC	16.509130	19.50358	0	96.32
ILEC	8.411611	14.69004	0	191.60

```
ggplot(Verizon) +
  geom_histogram(aes(Time)) +
  facet_wrap(.~Group, scales = "free")
```




```
ggplot(Verizon) +
  geom_boxplot(aes(Group, Time))
```

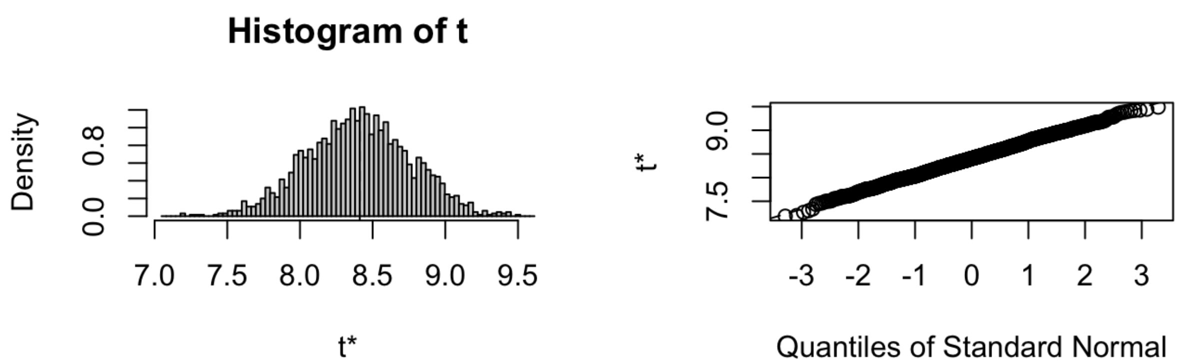


1.6 Bootstrapping CIs

There are many bootstrapping packages in `R`, we will use the `boot` package. The function `boot` generates R resamples of the data and computes the desired statistic(s) for each sample. This function requires 3 arguments:

1. data = the data from the original sample (data.frame or matrix).
2. statistic = a function to compute the statistic from the data where the first argument is the data and the second argument is the indices of the obervations in the boostrap sample.
3. R = the number of bootstrap replicates.

```r
library(boot) # package containing the bootstrap function

mean_func <- function(x, idx) {
  mean(x[idx])
}

ilec_times <- Verizon[Verizon$Group == "ILEC",]$Time
boot.ilec <- boot(ilec_times, mean_func, 2000)

plot(boot.ilec)
```



If we want to get Bootstrap CIs, we can use the `boot.ci` function to generate the different nonparametric bootstrap confidence intervals.

```r
boot.ci(boot.ilec, conf = .95, type = c("perc", "basic", "bca"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.ilec, conf = 0.95, type = c("perc",
"basic",
##     "bca"))
##
## Intervals :
## Level      Basic              Percentile            BCa
## 95%   ( 7.733,  9.110 )   ( 7.714,  9.091 )   ( 7.755,  9.125 )
## Calculations and Intervals on Original Scale
```

```r
## we can do some of these on our own
## percentile
quantile(boot.ilec$t, c(.025, .975))
```

```
##     2.5%    97.5%
## 7.714075 9.084725
```

```r
## basic
2*mean(ilec_times) - quantile(boot.ilec$t, c(.975, .025))
```

```
##    97.5%     2.5%
## 7.738496 9.109147
```

To get the studentized bootstrap CI, we need our statistic function to also return the variance of $\hat{\theta}$.

```r
mean_var_func <- function(x, idx) {
  c(mean(x[idx]), var(x[idx])/length(idx))
}

boot.ilec_2 <- boot(ilec_times, mean_var_func, 2000)
boot.ci(boot.ilec_2, conf = .95, type = "stud")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.ilec_2, conf = 0.95, type = "stud")
##
## Intervals :
## Level    Studentized
## 95%   ( 7.728,  9.183 )
## Calculations and Intervals on Original Scale
```

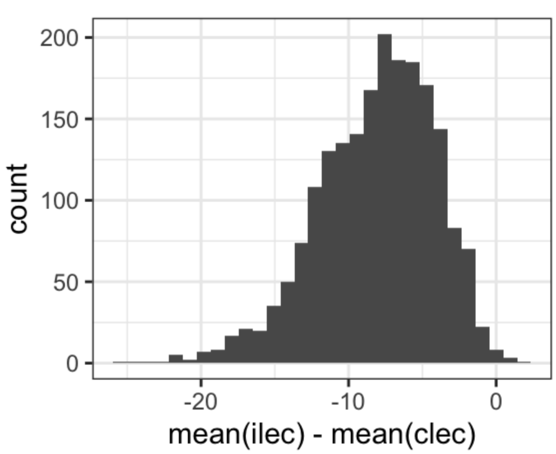
Which CI should we use?

1.7 Bootstrapping for the difference of two means

Given iid draws of size n and m from two populations, to compare the means of the two groups using the bootstrap,

The function `two.boot` in the `simpleboot` package is used to bootstrap the difference between univariate statistics. Use the bootstrap to compute the shape, bias, and bootstrap sample error for the samples from the `Verizon` data set of CLEC and ILEC customers.

```
library(simpleboot)

clec_times <- Verizon[Verizon$Group == "CLEC",]$Time
diff_means.boot <- two.boot(ilec_times, clec_times, "mean", R = 2000)

ggplot() +
  geom_histogram(aes(diff_means.boot$t)) +
  xlab("mean(ilec) - mean(clec)")
```



```r
# Your turn: estimate the bias and se of the sampling distribution
```

Which confidence intervals should we use?

```r
# Your turn: get the chosen CI using boot.ci
```

Is there evidence that

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 < 0$$

is rejected?