# Lab 8 - Math 58 / 58b: bootstrap

*your name here*

*due Mar 31, 2020*

## Lab Goals

- bootstrap from a single sample (quantitative variable)
- describe the bootstrap distribution in words

## Getting started

The lab today creates a new type of sampling distribution for a statistic. The process by which we do statistical inference almost always relies on understanding how a statistic varies from one sample to another. And ideally, we'd know what that distribution looks like by taking many repeated samples from the population. But if we could do that, there wouldn't be any point to all of the statistics part (we'd just report the parameter of interest!).

Recall the previous ways we've created estimates for the distribution of a statistic:

- Using mathematical models
  - normal distribution
  - $\chi^2$ distribution
- Using simulation
  - flipping coin model (with p defined either with a null value or an alternative value)
  - permuting cards model (**always** assuming the null distribution is true)

In today's lab we create a distribution for a statistic using simulation techniques (called "bootstrapping" or "resampling") **without** assuming any particular structure (i.e., no null hypothesis assumption and no alternative hypothesis structure).

The key to bootstrapping is that we use the sample as a proxy for the population! To find the distribution of the statistic, take many repeated samples (with replacement) from the original sample. Generally (as long as the original sample is random and big enough), the bootstrap sampling distribution will be a good approximation of the true sampling distribution:

- same shape (i.e., symmetric, skewed, etc.)
- same spread (i.e., good estimate of the SE)
- not same center (i.e., centered at the observed statistic instead of the true parameter)

### Load packages

For the lab, we'll use the `infer` syntax.

```r
library(tidyverse)
library(infer)
```

### Structure of the lab

In order to understand how a bootstrap distribution is generated, we are going to follow a series of steps:

1. Calculate the statistic of interest (use `summarize()`).
2. Resample a sample of the same size from the original sample. Calculate the statistic of interest. Repeat a few times to get a sense for what you expect to happen.
3. Use `infer` to formalize step #2 for 1000 reps
4. Using words, describe the center, spread, and shape of the bootstrap sampling distribution.

Together, we will go through the process with the median statistic.

Look at the following applet (click all the options to see how things change!) to get a feel for how a bootstrap sampling distribution works: http://www.rossmanchance.com/applets/OneSample.html?population=bootstrap

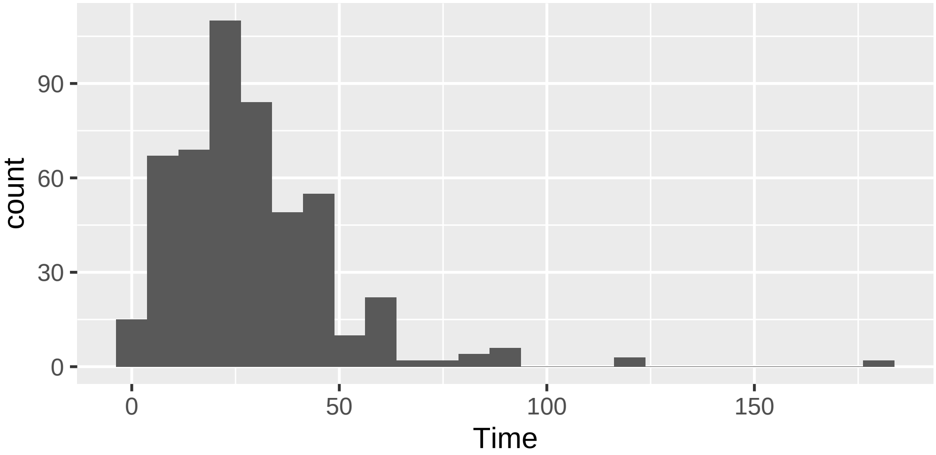
### Load the data

The data will will use comes from the American Housing Survey where people were asked the time (in minutes) and distance (in miles) that they typically traveled on their commute to work each day. The data is a sample (we will assume a random samples) of residents in Atlanta, GA.[1]

```
commute <- read.table("http://www.lock5stat.com/datasets/CommuteAtlanta.csv", sep=",", header = TRUE)
str(commute)
```

```
## 'data.frame':    500 obs. of  5 variables:
##  $ City    : Factor w/ 1 level "Atlanta": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age     : int  19 55 48 45 48 43 48 41 47 39 ...
##  $ Distance: int  10 45 12 4 15 33 15 4 25 1 ...
##  $ Time    : int  15 60 45 10 30 60 45 10 25 15 ...
##  $ Sex     : Factor w/ 2 levels "F","M": 2 2 2 1 1 2 2 1 2 1 ...
```

```
commute %>%
  ggplot() +
  geom_histogram(aes(x=Time), bins=25)  #feel free to change the number of bins
```



**The median! (Interest is in how the mean of the commute times varies from sample to sample.)**

As mentioned above, let's investigate the bootstrap distribution of the median time to work using the steps outlined.

1. Calculate the statistic of interest (use `summarize()`). Report the value of the statistic.

```
commute %>%
  summarize(med_time = median(Time))
```

```
##   med_time
## 1       25
```

The sample median commute time is 25 min. In the sample, half of people commute 25 min or less, half commute 25 min or more.

[1] Data and example taken from Statistics: Unlocking the Power of Data by Lock5, 2013.

2. Resample a sample of the same size from the original sample. Calculate the statistic of interest. Repeat a few times to get a sense for what you expect to happen. What do you expect to see (center, spread, shape) for the bootstrap distribution?

```
# don't set the seed here.  click on the green arrow a few times to see how things change

commute %>%
  sample_n(500, replace = TRUE) %>%
  summarize(med_time = median(Time))
```

```
##   med_time
## 1       25
```

The bootstrap median seems to be mostly 25 or 26. There is not a lot of variability in the values, so I would expect the full bootstrap distribution to be centered around 25 or 26 without a lot of variability.

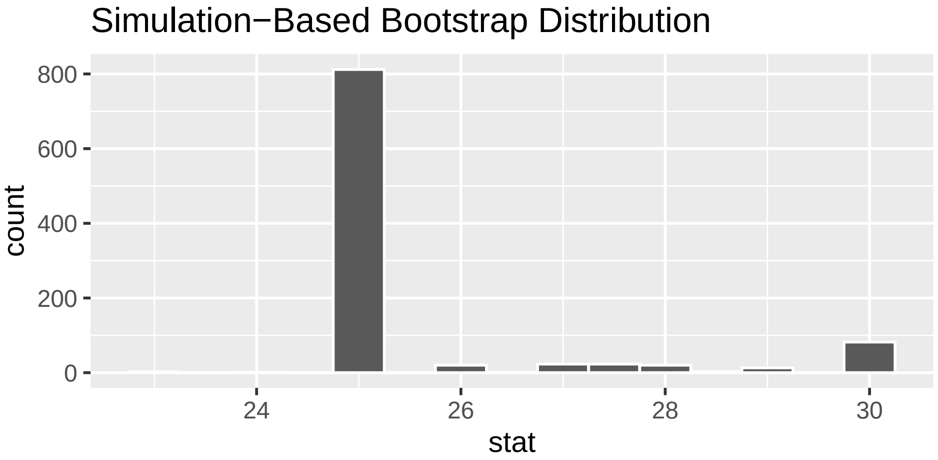
3. Use `infer` to formalize step #2 for 1000 reps

```
bs_median <- commute %>%
  specify(response = Time) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "median")

bs_median %>% head()
```

```
## # A tibble: 6 x 2
##   replicate  stat
##       <int> <dbl>
## 1         1    25
## 2         2    30
## 3         3    30
## 4         4    26
## 5         5    25
## 6         6    25
```

```
visualize(bs_median)
```



4. Using words, describe the center, spread, and shape of the bootstrap sampling distribution.

The bootstrap distribution is centered around 25 minutes, but can sometimes jump up to 30 minutes. The distribution is skewed right (the long tail is on the right side). The full range of values is only 5 minutes (in comparison to the range of commute times which have a min of 1 minute and a max o 181 minutes).

## To turn in

2. The mean! (Interest is in how the mean of the commute times varies from sample to sample.)
   (a) Calculate the statistic of interest (use `summarize()`). Report the value of the statistic.
   (b) Resample a sample of the same size from the original sample. Calculate the statistic of interest. Repeat a few times to get a sense for what you expect to happen. What do you expect to see (center, spread, shape) for the bootstrap distribution?
   (c) Use `infer` to formalize (b) for 1000 reps
   (d) Using words, describe the center, spread, and shape of the bootstrap sampling distribution.

3. The standard deviation (`sd()`)! (Interest is in how the standard deviation of the commute times varies from sample to sample.)
   (a) Calculate the statistic of interest (use `summarize()`). Report the value of the statistic.
   (b) Resample a sample of the same size from the original sample. Calculate the statistic of interest. Repeat a few times to get a sense for what you expect to happen. What do you expect to see (center, spread, shape) for the bootstrap distribution?
   (c) Use `infer` to formalize (b) for 1000 reps
   (d) Using words, describe the center, spread, and shape of the bootstrap sampling distribution.