AE 13: Bootstrapping for Confidence Intervals

Dav King-

2/22/22

```
library(tidyverse)
library(tidymodels)
```

```
manhattan <- read_csv("manhattan.csv")
```

Learning goals

- Understand how to draw a bootstrap sample and calculate a bootstrap statistic
- Use `infer` to obtain a bootstrap distribution
- Calculate a confidence interval from the bootstrap distribution
- Interpret a confidence interval in context of the data

Estimation

Point estimate

A point estimate is a single value computed from the sample data to serve as the "best guess", or estimate, for the population parameter.

Suppose we were interested in the population mean. What would be natural point estimate to use?

You would use the first option below, the sample mean.

Quantity	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Median	M	$\tilde{x}$
Proportion	p	$\hat{p}$

What is the downside to using point estimates?

Confidence intervals

A plausible range of values for the population parameter is an interval estimate. One type of interval estimate is known as a **confidence interval**.

- If we report a point estimate, we probably won't hit the exact population parameter.
- If we report a range of plausible values, we have a good chance at capturing the parameter.

Rent in Manhattan

On a given day in 2018, twenty one-bedroom apartments were randomly selected on Craigslist Manhattan from apartments listed as "by owner". The data are in the `manhattan` data frame. We will use this sample to conduct inference on the typical rent of 1 bedroom apartments in Manhattan.

Variability of sample statistics

- In order to construct a confidence interval we need to quantify the variability of our sample statistic.
- For example, if we want to construct a confidence interval for a population mean, we need to come up with a plausible range of values around our observed sample mean.
- This range will depend on how precise and how accurate our sample mean is as an estimate of the population mean.
- Quantifying this requires a measurement of how much we would expect the sample mean to vary from sample to sample.

Suppose you randomly sample 50 students and 5 of them are left handed. If you were to take another random sample of 50 students, how many would you expect to be left handed? Would you be surprised if only 3 of them were left handed? Would you be surprised if 40 of them were left handed?

Quantifying the variability of a sample statistic

We can quantify the variability of sample statistics using

1. **simulation**: via bootstrapping (today);
2. **theory**: via Central Limit Theorem (later in the course).

Bootstrapping

- The term **bootstrapping** comes from the phrase "pulling oneself up by one's bootstraps", to help oneself without the aid of others.
- In this case, we are estimating a population parameter, and we'll accomplish it using data from only from the given sample.
- This notion of saying something about a population parameter using only information from an observed sample is the crux of statistical inference, it is not limited to bootstrapping.
- Here is a cool animation of the bootstrapping process.

Bootstrapping scheme

1. **Take a bootstrap sample** - a random sample taken with replacement from the original sample, of the same size as the original sample.
2. **Calculate the bootstrap statistic** - a statistic such as mean, median, proportion, slope, etc. computed from the bootstrap samples.
3. **Repeat steps (1) and (2) many times to create a bootstrap distribution** - a distribution of bootstrap statistics.
4. **Calculate the bounds of the XX% confidence interval** as the middle XX% of the bootstrap distribution.

Part 1: Drawing a bootstrap sample

Let's start by using bootstrapping to estimate the **mean** rent of one-bedroom apartments in Manhattan.

Exercises

Part 1

What is the point estimate of the typical rent? Do you think this is the exact average rent for an apartment?

```
mean(manhattan$rent)
```

```
## [1] 2625.8
```

It is definitely not the exact average rent for an apartment - nobody would list a price at 80 cents, especially with such a nice round number so close to it.

Part 2: Bootstrap confidence interval

We will use the `infer` package, included as part of `tidymodels` to calculate a 95% confidence interval for the mean rent of one-bedroom apartments in Manhattan.

We start by setting a seed to sure our analysis is reproducible. We'll use 101121 to set our seed today but you can use any value you want on assignments unless we specify otherwise.

```
set.seed(101121)
```

Generating the bootstrap distribution

We can use R to take many bootstrap samples and generate a bootstrap distribution.

You can uncomment the lines and fill in the blanks to create the bootstrap distribution of sample means and save the results in the data frame `boot_dist`.

We will **100 reps** for the in-class activity. (You will use about 15,000 reps for assignments outside of class.)

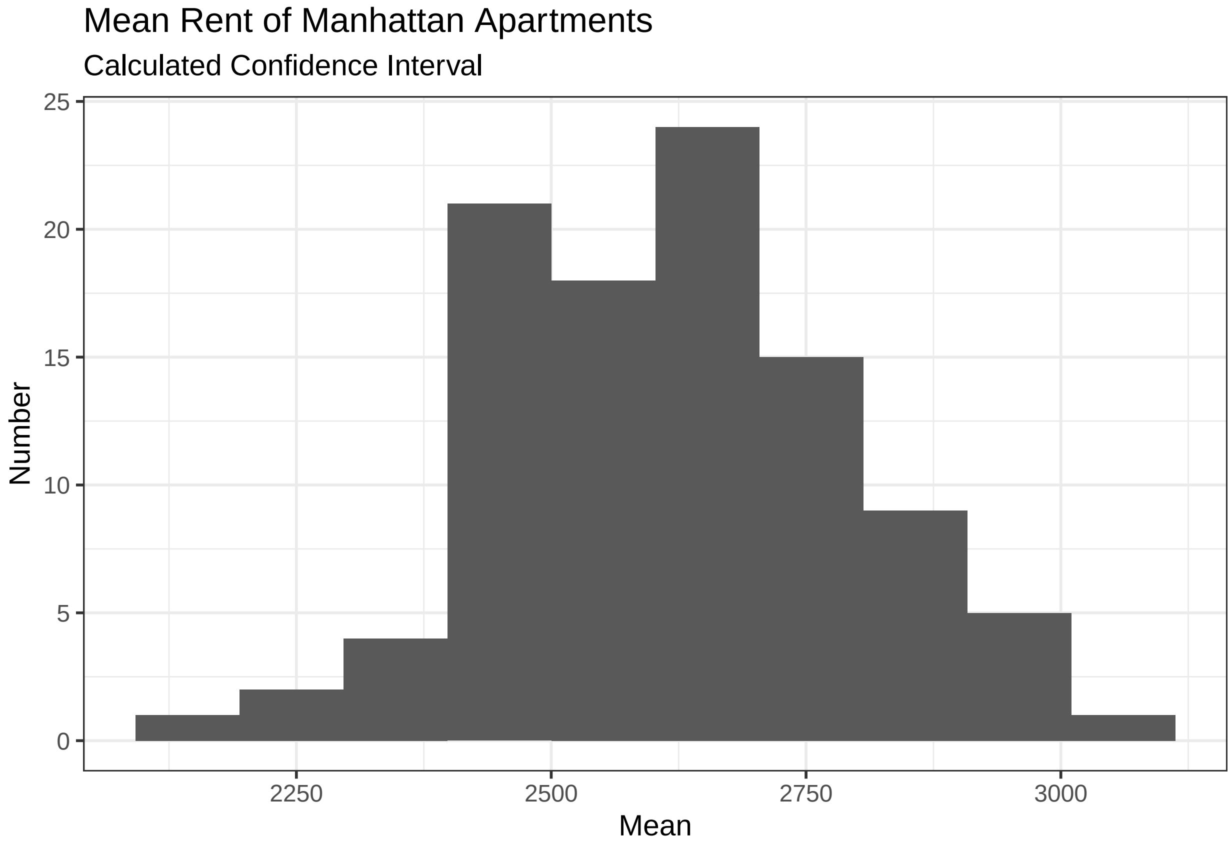
```
boot_dist <- manhattan %>%
  specify(response = rent) %>%
  generate(reps = 100, type = "bootstrap") %>%
  calculate(stat = "mean")
```

- How many rows are in `boot_dist`? 100
- What does each row represent? One bootstrapped sample mean
- What are the variables in `boot_dist`? What do they mean? `replicate`, which is just an ordinal number, and `stat`, which is the sample mean.

Visualize the bootstrap distribution

Visualize the bootstrap distribution using a histogram. Describe the shape, center, and spread of this distribution.

```
ggplot(boot_dist, aes(x = stat)) +
  geom_histogram(bins = 10) +
  labs(x = "Mean", y = "Number", title = "Mean Rent of Manhattan Apartments",
       subtitle = "Calculated Confidence Interval") +
  theme_bw()
```



Calculate the confidence interval

Uncomment the lines and fill in the blanks to construct the 95% bootstrap confidence interval for the mean rent of one-bedroom apartments in Manhattan.

```
boot_dist %>%
  summarize(lower = quantile(stat, 0.025),
            upper = quantile(stat, 0.975))
```

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 2295. 2947.
```

Interpret the interval

We can be 95% confident that the true mean rent of Manhattan one-bedroom apartments lies between $2292 and 2947/month.

Part 3: Changing the confidence level

- Modify the code used to calculate a 95% confidence interval to calculate a **90% confidence interval** for the mean rent of one-bedroom apartments in Manhattan. How does the width of this interval compare to the width of the 95% confidence interval?

```
boot_dist %>%
  summarize(lower = quantile(stat, 0.05),
            upper = quantile(stat, 0.95))
```

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 2385. 2916.
```

It is notably less wide than the 95% confidence interval.

- Now let's calculate a 99% confidence interval for the mean rent of one-bedroom apartments in Manhattan. How does the width of this interval compare to the width of the 95% confidence interval?

```
boot_dist %>%
  summarize(lower = quantile(stat, 0.005),
            upper = quantile(stat, 0.995))
```

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 2192. 3024.
```

It is notably wider than then 95% confidence interval.

- What is one advantage to using a 90% confidence interval instead of a 95% confidence interval to estimate a parameter? - What is one advantage to using a 99% confidence interval instead of a 95% confidence interval to estimate a parameter?

A 90% interval is more precise, but a 99% interval is more accurate than 95% (i.e., more likely to include the true value).

Part 4: Additional practice- on your own or in groups

Next, use bootstrapping to estimate the median rent for one-bedroom apartments in Manhattan.

- Generate the bootstrap distribution of the sample medians. Use 100 reps. Save the results in `boot_dist_median`. Why have I set a seed here again?

```
## add code
set.seed(101121)

boot_dist_median <- manhattan %>%
  specify(response = rent) %>%
  generate(reps = 100, type = "bootstrap") %>%
  calculate(stat = "median")
```

The seed is set to make this dataset reproducible and consistent across both our own work and the entire class's as a whole.

- Calculate a 92% confidence interval.

```
boot_dist_median %>%
  summarize(lower = quantile(stat, 0.04),
            upper = quantile(stat, 0.96))
```

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 2174.  2875
```

- Interpret the 92% confidence interval: We can be 92% confident that the median rent of all one-bedroom Manhattan apartments is between $2174 and 2875/month.

For next time:

The infer package is tough to learn (but once you do, you can do lots with it)! Here are two resources that I think you will find useful:

- Getting to Know Infer
- Full pipeline examples in infer