Math 300 Lesson 23 Notes

Bootstrap Introduction

YOUR NAME HERE

July, 2022

Contents

Objectives

1. Explain the concept of a bootstrap distribution using proper terminology and notation.
2. Use `R`, to find the bootstrap distribution of the sample statistic.

Reading

[Chapter 8 - 8.2](https://moderndive.com/8-confidence-intervals.html

Lesson

Work through the learning checks LC 8.1 - LC 8.2. Complete code as necessary.

- The bootstrap samples with replacement. Each bootstrap sample has the same size as the original sample. If you sampled 50 data points in the original sample, sample 50 in the bootstrap sample.
- The key idea of the bootstrap is that it gives an estimate of the standard error of a point estimate.
- The bootstrap assumes that original sample is representative of the population.

Libraries

```
library(tidyverse)
library(moderndive)
```

Re-create Ideas from Reading

Here is the original data.

```
head(pennies_sample)
```

```
## # A tibble: 6 x 2
##      ID  year
##   <int> <dbl>
## 1     1  2002
## 2     2  1986
## 3     3  2017
## 4     4  1988
## 5     5  2008
## 6     6  1983
```

Find the mean minting year of our sample and visualize the distribution of the minting years.

```
# Complete the code and remove the comment symbol
#pennies_sample %>%
#  summarize(mean_year = mean(________))
```

```
# Complete the code and remove the comment symbol
# ggplot(_________, aes(x = ______)) +
#   geom_histogram(binwidth = 10, color = "white") +
#   theme_classic()
```

In our sample, the average year of minting was 1995.44. It is not an integer since it is an average.

- What is the population of interest?
- What is the population parameter?
- What is the point estimate?

Now if we sampled with replacement from this sample, we would get a bootstrap sample. In the reading, this was done 35 times.

```
head(pennies_resamples)
```

```
## # A tibble: 6 x 3
## # Groups:   name [1]
##   replicate name     year
##       <int> <chr>   <dbl>
## 1         1 Arianna  1988
## 2         1 Arianna  2002
## 3         1 Arianna  2015
## 4         1 Arianna  1998
## 5         1 Arianna  1979
## 6         1 Arianna  1971
```

- Explain this tibble.
- Is it tidy?

- How was the data obtained?
- What is a bootstrap sample in this tibble?

```
# Complete the code and remove the comment symbol
# pennies_resamples %>%
#   filter(replicate==______) %>%
#   ggplot(aes(x = _______)) +
#   geom_histogram(binwidth = 10, color = "white") +
#   theme_classic()
```

- Compare the above histogram with the following histogram. What is the difference?

```
# Complete the code and remove the comment symbol
# pennies_resamples %>%
#   group_by(_______) %>%
#   summarize(mean_year = mean(________)) %>%
#   ggplot(aes(x = _________)) +
#   geom_histogram(binwidth = 1, color = "white", boundary = 1990) +
#   theme_classic()
```

Now let's use the computer to find the bootstrap distribution. (Objective 2)

```
head(pennies_sample)
```

```
## # A tibble: 6 x 2
##      ID  year
##   <int> <dbl>
## 1     1  2002
## 2     2  1986
## 3     3  2017
## 4     4  1988
## 5     5  2008
## 6     6  1983
```

```
# Complete the code and remove the comment symbol
#Set a seed for reproducibility of results.
set.seed(84337)
# Samples of size 50 repeated 1000 times. This is the bootstrap distribution of the mean
# virtual_resampled_means <- pennies_sample %>%
#   rep_sample_n(size = _____, replace = TRUE, reps = 1000) %>%
#   group_by(________) %>%
#   summarize(mean_year = mean(________))
```

```
#head(virtual_resampled_means)
```

```
# Complete the code and remove the comment symbol
# The bootstrap distribution
# ggplot(virtual_resampled_means, aes(x = ___________)) +
#   geom_histogram(binwidth = 1, color = "white", boundary = 1990) +
#   labs(x = "sample mean") +
#   theme_classic()
```

A couple of things to note is that it looks normal, by the CLT, and it is centered on the original sample mean.

```
# Complete the code and remove the comment symbol
# virtual_resampled_means %>%
#   summarize(mean_of_means = mean(___________))
```

LC 8.1 (Objective 1)

(LC 8.1) What is the chief difference between a bootstrap distribution and a sampling distribution?

Solution:

LC 8.2 (Objective 1)

(LC 8.2) Looking at the bootstrap distribution for the sample mean in Figure 8.14, between what two values would you say *most* values lie?

Solution:

Documenting software

- File creation date: 2022-07-06
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1
- `moderndive` package version: 0.5.4