MATH 254 - Statistical Modeling and Applications - HW4

Tural Sadigov

Problems

1. Suppose you are studying the how much time (on average) people in New York state spend on their daily commute. Suppose you conducted a survey and collected 100 commute times. Now you want to construct a 95% classical confidence interval for the true average commute time. In order to use the classical method, we need to have some idea about the true sampling distribution of $\bar{X}$.

   - Briefly explain how to construct the true sampling distribution of $\bar{X}$?
   - Do you think, in practice, we can (or we should) construct the true sampling distribution of $\bar{X}$? Explain.
   - Briefly explain how to construct a bootstrap sampling distribution of $\bar{X}$.
   - What information can we infer from the bootstrap sampling distribution of $\bar{X}$?

2. Consider the data set contains information about a random sample of commuters from Atlanta. Import Atlanta data to R Studio, using the code below:

```r
library(tidyverse)
url = "https://raw.githubusercontent.com/turalsadigov/MATH_254/main/data/CommuteAtlan
commute_atlanta <- read_csv(url)

commute_atlanta %>%
  head()
```

```
# A tibble: 6 x 5
  City      Age Distance  Time Sex
  <chr>   <dbl>    <dbl> <dbl> <chr>
1 Atlanta    19       10    15 M
```

```
2 Atlanta    55        45    60 M
3 Atlanta    48        12    45 M
4 Atlanta    45         4    10 F
5 Atlanta    48        15    30 F
6 Atlanta    43        33    60 M
```

We are interested in true correlation coefficient, ρ, between all Atlanta commuters' ages and the distance they travel daily. We have a sample of 500 such commuters.

- Find sample correlation coefficient that approximates ρ. How do you interpret this statistic?
- Create the bootstrap distribution of sample correlation coefficient of Age and Distance variables using 3000 bootstrap samples. Make sure you label the axes, have descriptive title, smoothed histogram on it.
- What is the center of the Bootstrap distribution? How does it compare to the sample correlation coefficient you found in part a?
- How does the Bootstrap distribution help you in the context of the problem, i.e., why did we create it in the first place?

3. Consider Atlanta commute data above.
   - Construct Bootstrap Percentile 90% Confidence Interval for true correlation coefficient. Interpret it.
   - Construct Bootstrap Percentile 95% Confidence Interval for true correlation coefficient. Interpret it.
   - Construct Bootstrap Percentile 99% Confidence Interval for true correlation coefficient. Interpret it.
   - Compare the width of the intervals you found above, and explain why one interval is narrower from the other?
   - If your confidence interval is wide, does it increase or decrease your confidence in the interval capturing true correlation coefficient?
   - If your confidence interval is wide, does it increase or decrease the accuracy of the interval estimating true correlation coefficient?

4. Review the Mice data example we worked on from Efron & Tibshirani. R code is on Blackboard. Let's say that we are interested in understanding if there is difference between true variances of two populations that these two groups sampled from. One way to check this is to create Confidence intervals for $\frac{\sigma_T^2}{\sigma_C^2}$.
   - Create 95% CI for $\frac{\sigma_T^2}{\sigma_C^2}$.

- Interpret the interval.
- Is it still plausible that two populations have the same variance?