STAT 331 Lab 12

Instructions

For this lab you will, again, be submitting an HTML markdown document. We will be working agin with the San Francisco City Salary Data set posted on www.kaggle.com. The data set is posted on our Poly Learn, but you may also just get it straight from Kaggle.

Importing the data involves a few special things:

- Make sure `header = TRUE`

- Include `na.strings = c("Not Provided", "Not provided", "")`

If you're using `read_csv()`, then the argument is just `na`

Exercises

1. Our variable of interest for this lab is going to be Total Pay. Plot a histogram of Total Pay and comment on its shape, center, and spread.

2. Suppose we're interested in making inference about the typical salary (Total Pay) of all San Francisco city employees and this is our representative sample. Is the mean a good statistic to use here to describe the typical value of salary? Why or why not?

3. Recall that one of the conditions of a one-sample t-test for a mean is that the population is normally distributed. Based on yor graph in (1), does this assumption seem reasonable? Why or why not?

4. Compute a 95% t-confidence interval for the mean Total Pay.

5. Use the bootstrap procedure to construct a 95% bootstrap confidence interval for the mean Total Pay. Compare this interval to your t-interval in (3). Be sure to explain any differences you see.

6. Since the distribution of Total Pay is so skewed, there may be other statistics that are better at describing the typical salary. Write your own function for each of the following statistics:

- Midhinge

- Trimmed Mean (this should take two arguments: the data and the percent to trim)

7. Perform the bootstrap procedure and produce graphs of the distributions of:

- Midhinge

- 5% Trimmed Mean

- 10% Trimmed Mean

- 25% Trimmed Mean

- Median

8. Compute 95% bootstrap confidence intervals for the statistics listed in (7) and compare to your intervals for the mean from above. How do they compare? Which would you use to describe the typical salary and why?