Analyze the Income Gap between Female and Male Employees

October 25, 2019

Introduction

This assignment is developed based on the three datasets collected from Kaggle.com at: https://www.kaggle.com/theoviel/kagglers-gender-pay-gap-salary-prediction. You may refer the website for explanation, cleaning, visualization, and basic analysis of the datasets. The three datasets: surveySchema.csv, freeFormResponses.csv, multipleChoiceResponses.csv.

Tasks

Complete the following tasks:

1. Calculate the median income of male employees and the median income of female employee in the population. (look the set of all employees in the datasets as the population). (1 point)
2. Draw an overlaid graph to show the histograms of the incomes of female and male employees in the population. (You create one histogram for male, and another histogram for female, but the two histograms will be displayed in the same graph with different colors). (1 point)
3. Use the random sampling, empirical distribution, sample comparison, bootstrap, hypothesis testing as well as A/B testing we discussed in the class to analyze the income gap between female and male employees.
   - Select a sample from the population. Make sure your sample include 500 employees selected from the population, and consider how to ensure the sampling strategy is fair since the datasets include overwhelmed male employees than female employees (1 point).
   - Define test statistic, null hypothesis and alternative hypothesis (1 point).
   - Draw the income histogram for the sample, calculate the median income of the sample, and draw a red dot and a yellow dot of the female median income and male median income of the population, respectively, in the histogram (1 point).

- Draw the histogram of the test statistic of the sample, and draw a red dot to show the corresponding test statistic of the population (e.g. the difference of the median incomes between female and male employees) in the diagram (1 point).
- Write a procedure to use bootstrap to produce at least 5000 samples (1 point).
- Draw the histogram of the test statistic of the bootstrap samples (1 point).
- Define confidence interval and P-value to validate the hypothesis you defined (2 points).

4. Submit Python code, the writing for explaining the data cleaning procedure, defining the test statistic, hypothesises, random sampling, bootstrap, confidential intervals, P-vales, as well as interpretation of your results, and all outputs described above.