Springboard Project 8.3 (Examine Racial Discrimination)

Using U.S. Job Market Discrimination data

1. Which test is appropriate for this problem? Does CLT apply?

I chose the Chi-square test for independence here. The CLT applies because expected cell counts are all greater than 4, per my calculations using row and column totals and the total number of observations.

2. What are the null and alternate hypotheses?

The null hypothesis is that there is no relationship between whether a person's name sounds black or white and whether or not that person receives a call from employers (i.e., the two events are independent). The alternative is that receiving a call from employers depends on whether a person's name sounds black or white.

3. Compute margin of error, confidence interval, and p-value.

To find the margin of error, I wrote code to count the number of observations in each cell of a 2x2 contingency table, pulling from relevant columns in the initial data frame. Next, I used the table values to find expected probabilities of each outcome (receiving a call or not receiving one) before using those probabilities to find the standard error for the confidence interval and before finding the interval per se.

To find the p-value (for the Chi-square test), I wrote a function to create the contingency table at hand and subsequently performed the Chi-square test.

4. Write a story describing the statistical significance in the context of the original problem.

Since the p-value (5e-5) is less than .05, we reject the null hypothesis and conclude that there is a significant difference between the probabilities of a white-sounding name and a black-sounding name getting a call from an employer.

5. Does your analysis mean that race/name is the most important factor in callback success? Why or why not? If not, how would you amend your analysis?

It does not mean that the race associated with one's name is necessarily the most important factor because I haven’t compared this variable’s significance to that of other variables in the data set. I would amend my analysis by conducting further exploratory steps with the other variables and performing a logistic regression and/or using tree-based analyses.