Examining Racial Discrimination in the US Job Market

Pamela Augustine

BACKGROUND

Racial discrimination continues to be pervasive in cultures throughout the world. Researchers examined the level of racial discrimination in the United States labor market by randomly assigning identical résumés to black-sounding or white-sounding names and observing the impact on requests for interviews from employers.

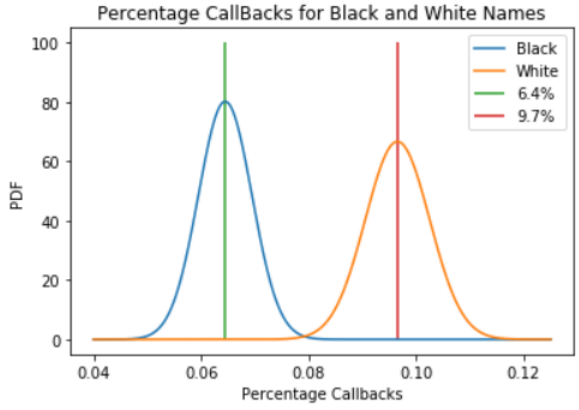
THE DATA

Each row in the dataset analyzed represents a resume. The race column has ‘b’ and ‘w’ values indicating black-sounding and white-sounding. The column 'call' has two values, 1 and 0, indicating whether the resume received a call from employers or not.

The 'b' and 'w' values in race are assigned randomly to the resumes when presented to the employer.

TEST APPROPRIATE FOR THIS PROBLEM

This is a binary response type of problem (1,0) which makes it a Bernoulli distribution or binomial distribution. However, testing the difference between the "percentage callbacks" for each race will follow a normal distribution in which CLT can be applied. Two sample t-test is appropriate to use in comparing these two percentages. In this sample % callback for black sounding names is 6.4% and 9.7% for white sounding names.



NULL & ALTERNATE HYPOTHESES

Two hypotheses will be tested for this problem.

Ho: There is no difference between black and white resumes/ There is no significant difference between "percentage callbacks" for black and white sounding resumes.

H1: There is difference between black and white resumes/ There is significant difference between "percentage callbacks" for black and white sounding resumes.

Two sampling t test was used by calculating the following parameters

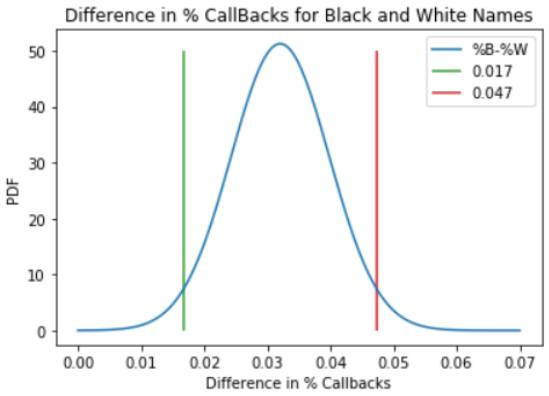
$t_{P1-P2} = \frac{(P1-P2))-d}{SE}$ where P1 is 6.4%, P2 is 9.7% , d is difference of P1-P2, d=0 for Ho

$SE = \sqrt[2]{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ where SE is standard error, s_1 is black % callback standard deviation, s_2 is white % callback standard deviation n_1 and n_2 are their respective sample sizes.

$DF = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$ where DF is degrees of freedom.

The calculated t_{P1-P2} and DF are used to calculate p value using t distribution table. Two tail distribution was used in solving p value. Summary of results are in the table below.

P1-P2	0.032
Assumed significance level	0.05
SE	0.0078
t_{P1-P2}	-4.11
DF	4714
p value	3.93E-5
Margin of error	0.015



Calculated p value<0.05 which means Ho will be rejected and H1 will be accepted. H1: There is significant difference between % callbacks for black and white sounding resume. Using 95% confidence level as threshold in all calculations in above table, we can say that There is 95% confidence that confidence interval of “difference in % callbacks in black and white sounding resume” is from 0.017 to 0.047.

RACE/NAME IS NOT THE MOST IMPORTANT FACTOR

The analysis I did does not mean that race/name is the most important factor in callback success. It only means that race/name is a factor that affect callback success. Correlation between different features and callback success must be analyzed and ranked to test which feature is most important in callback success.

IPython NOTEBOOK SOLUTIONS

All solutions can be viewed in IPython Notebook in my github below.

https://github.com/DrAugustine1981/Springboard/blob/master/racial_discrimination_eda.ipynb