Examining racial discrimination in the US job market

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 black-sounding or white-sounding names and observing the impact on requests for interviews from employers.

Data

In the dataset provided, each row represents a resume. The 'race' column has two values, 'b' and 'w', 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.

Note that the 'b' and 'w' values in race are assigned randomly to the resumes.

Exercise

Perform a statistical analysis to establish whether race has a significant impact on the rate of callbacks for resumes.

Resources

- Experiment information and data source: http://www.povertyactionlab.org/evaluation/discrimination-job-market-united-states
- Scipy statistical methods: http://docs.scipy.org/doc/scipy/reference/stats.html

---

In [1]:

```
import pandas as pd
import numpy as np
from scipy import stats
```

In [2]:

```
data = pd.io.stata.read_stata('C:/us_job_market_discrimination.dta')
```

In [3]:

```
# number of callbacks for balck-sounding names
sum(data[data.race=='b'].call)
```

Out[3]:

```
157.0
```

Exercise

1. What test is appropriate for this problem? Does CLT apply?
2. What are the null and alternate hypotheses?
3. Compute margin of error, confidence interval, and p-value.
4. Discuss statistical significance.

You can include written notes in notebook cells using Markdown:

- In the control panel at the top, choose Cell > Cell Type > Markdown
- Markdown syntax: http://nestacms.com/docs/creating-content/markdown-cheat-sheet

---

In [4]:

```
data.head()
```

Out[4]:

	id	ad	education	ofjobs	yearsexp	honors	volunteer	military	empholes	occupspecific	...
0	b	1	4	2	6	0	0	0	1	17	...
1	b	1	3	3	6	0	1	1	0	316	...
2	b	1	4	1	6	0	0	0	0	19	...
3	b	1	3	4	6	0	1	0	1	313	...
4	b	1	3	3	22	0	0	0	0	313	...

5 rows × 65 columns

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

In [5]:

```
#To answer this question lets split the population in two groups and find
out the rate of calls per group

sample_white=data[data.race=='w']
sample_black=data[data.race=='b']
```

In [6]:

```
#Total Number of resume(job applied) per race:
w_res=len(sample_white.race)
b_res=len(sample_black.race)

#Number of interview request per race:
w_calls=sum(data[data.race=='w'].call)
b_calls=sum(data[data.race=='b'].call)

# Sample proportions
w_sample_prop = w_calls / w_res
b_sample_prop = b_calls / b_res
```

```
print ("Rate of receiving interview requests for White applicants = "+ str(
w_sample_prop))
print ("Rate of receivibg interview request for Black applicants = "+ str(b
_sample_prop))
```

```
Rate of receiving interview requests for White applicants = 0.0965092402464
Rate of receivibg interview request for Black applicants = 0.064476386037
```

Does CLT apply?

We can perform hypothesis test to compare between white and black sample proportions.

These two condition must be satisfied for CLT.

1. Observations should be independent. Black and White sounding names were randomly assigned to similar resumes so they represent a random sample and are independent as well.
2. Sample size should be large enough so that n >= 10 and np >= 10: We have a sample size of n=2435 independent observations per race. Since,we dont know have information about the actual total population , hence we can use pooled proportion.(combined total of sample proportion)

Since both the conditions are met , CLT can be applied.

2. What are the null and alternate hypotheses?

Null Hypothesis (Ho): w_sample_prop = b_sample_prop

Alternative Hypothesis (Ha): w_sample_prop ≠ b_sample_prop

Significance Level = .05

We are considering null hypotheses to be true meaning there is no different in the rate of calls received by these two groups.

we can calculate the pooled population as Ppooled= (w_calls+b_calls)/(w_res+b_res)

In [7]:

```
Ppooled= round((w_calls+b_calls)/(w_res+b_res),2)
Ppooled
```

Out[7]:

```
0.080000000000000002
```

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

In [8]:

```
import math
```

```
# first we will compute the standard deviation (σ) of the sampling
distribution.
# To calculate Std we use the null hypothesis as true.
# It means that mean = 0, and SE uses Ppooled as population proportion:

Ppooled_std=round((((Ppooled*(1-Ppooled))/b_res)+((Ppooled*(1-Ppooled))/w_r
es))**0.5,2)
Ppooled_std

# Zvalue
z_score = ((w_sample_prop- b_sample_prop)- 0)/Ppooled_std

Ppooled_std,z_score
print("Standard Dev = " + str(Ppooled_std))
print("z_score = " + str(z_score))
```

```
Standard Dev = 0.01
z_score = 3.20328542094
```

In [9]:

```
#margin of error
moe = round((z_score * Ppooled_std),2)
print("Margin of error is " + str(moe))
```

```
Margin of error is 0.03
```

In [10]:

```
#confidence interval
lb = (w_sample_prop - b_sample_prop)- moe #lower bound
ub = (w_sample_prop - b_sample_prop)+ moe #upper_bound

print ("Confidence Interval = " + str(lb) + "," + str(ub))
```

```
Confidence Interval = 0.00203285420945,0.0620328542094
```

There is a 95% chance that the true difference of white-sounding call back rates and black-sounding call back rates is between .002 and .062

4.Discuss statistical significance.

In [11]:

```
#Lets find out the p-value.
p_value=stats.norm.pdf(z_score) * 2 # two sided
print("p_value = " + str(p_value)+ " which is ≈ 0")
```

```
p_value = 0.00471828403844 which is ≈ 0
```

Since calculated p-value is way below the significance level of .05, that gives us good evidence to reject the null hypothesis that the rate of interview request received by black applicants is same as white population. This indicates that there exists a difference between the two races in which applicants with white-sounding names are favored over black-sounding names.