Aditya Pitchuka  
INFO 5502.002 (14676)  
Spring 2020

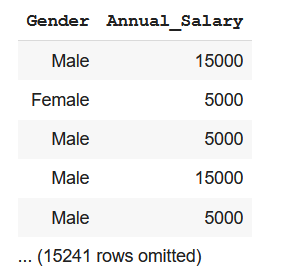
**Assignment – 4**

Data cleaning performed on the data set:

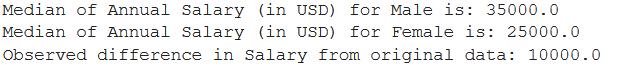
1. Re-labelled columns Q1 to *Gender* and Q9 to *Annual\_Salary* and excluding the survey questions.
2. Removing the values which are unclear in both *Gender* and *Annual\_Salary* like –

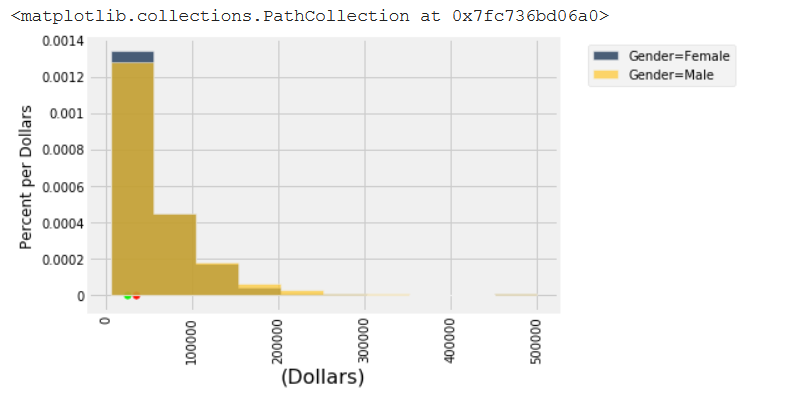
‘*nan*’, ‘*Prefer not to disclose*’ and ‘*I do not wish to disclose my approximate yearly compensation*’

1. Removing the salary intervals to exact values to clear calculations of median.



**Task 1 & 2**  
Calculating Median for respective genders and plotting a histogram for the given population to observe the distribution curve.





1 Distribution of Original Population with Median Salary of Male and Female with Statistic

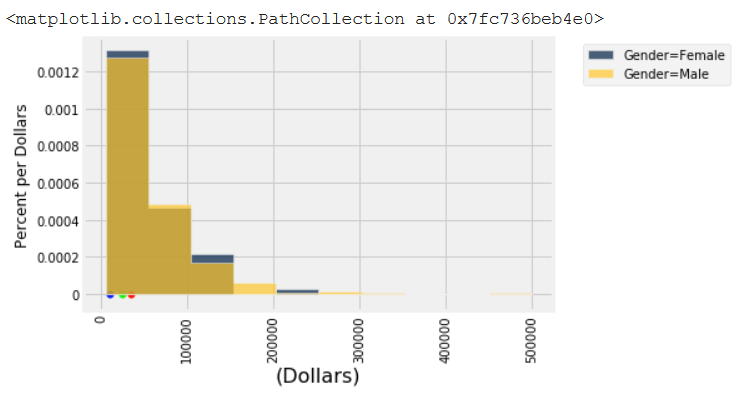
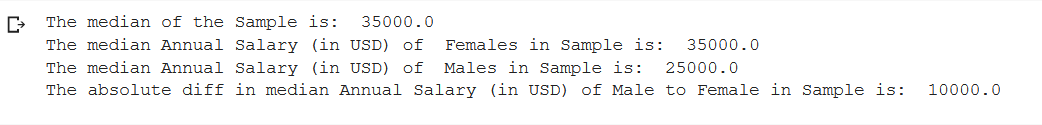
**Task 3**

First, we define the hypothesis, Null & Alternative based on the statistic.   
Null Hypothesis – There will not be significant difference between the Salary of male and female.

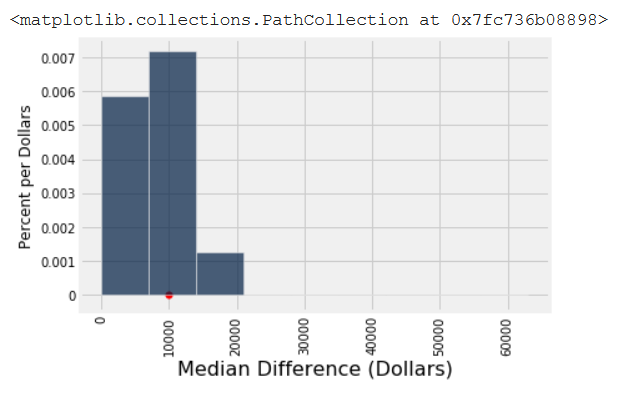
Alternative Hypothesis – There will be noticeable difference between salaries of males and females.

Statistic – Absolute Difference between Median of Male Salary and Female Salary.

We define a function that makes a sample of 500 which maintains the ratio of Male & Female with respect to the original population and returns the sample derived. We find the median of salary for male and female and repeat it many times to stimulate the empirical distribution.

For One Sample Run:  


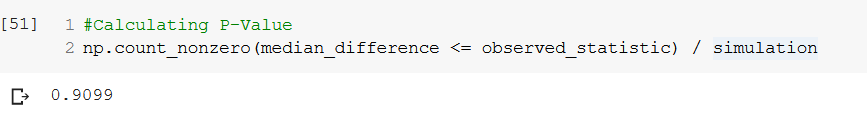
Simulation Results:



Histogram showing the median differences of Salary for 10000 samples ran against the actual salary difference

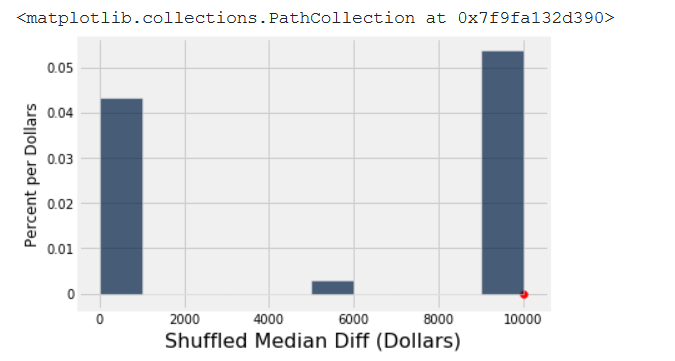
Calculating P-Value:

We calculate P-value by counting the number of times the difference between male and female salary median is less than or equal to the actual value of the population times the simulation ran.



A/B Testing:

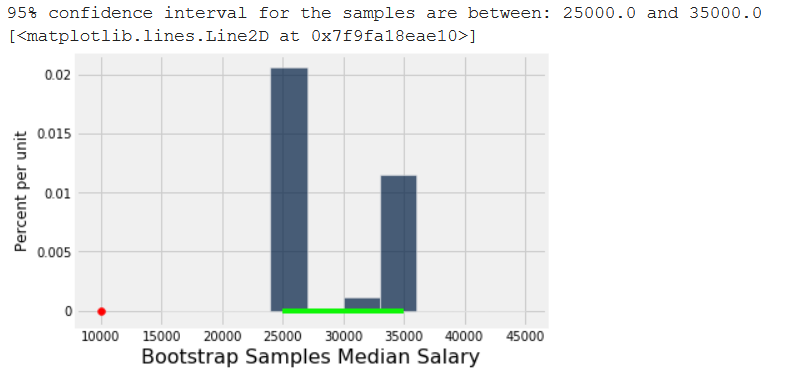
We shuffle the column values of gender and make samples and calculate the statistic needed. We simulate this multiple times and observe the statistic with original population.



Histogram containing the Median Difference of Salary of Male and Female with Shuffled values in the Gender Column

Bootstrap:

From the original sample made, we consider it the original population and create more and more samples from it and calculate the confidence interval for the median salary.



Bootstrap samples Median Salary with Confidence level

Confidence Interval:

For the samples we selected 95% confidence interval goes from 25000 to 35000 meaning we are 95% confident of making a prediction that the population median compensation will fall between 25000 and 35000.

Results & Conclusion:

Notice how the majority of distribution is not centred around 0. This does not make sense if null hypothesis is true, because under the null hypothesis the two groups should have roughly the same average. Therefore, the difference between the group averages should be around 0.

The observed difference in the original sample is about 10000, which is where the maximum differences are clustered. The observed value of the statistic and the predicted behaviour of the statistic under the null hypothesis are consistent.

The conclusion of the test is that the data favour the alternative over the null. The median compensation for males and females differs by at least 10000.