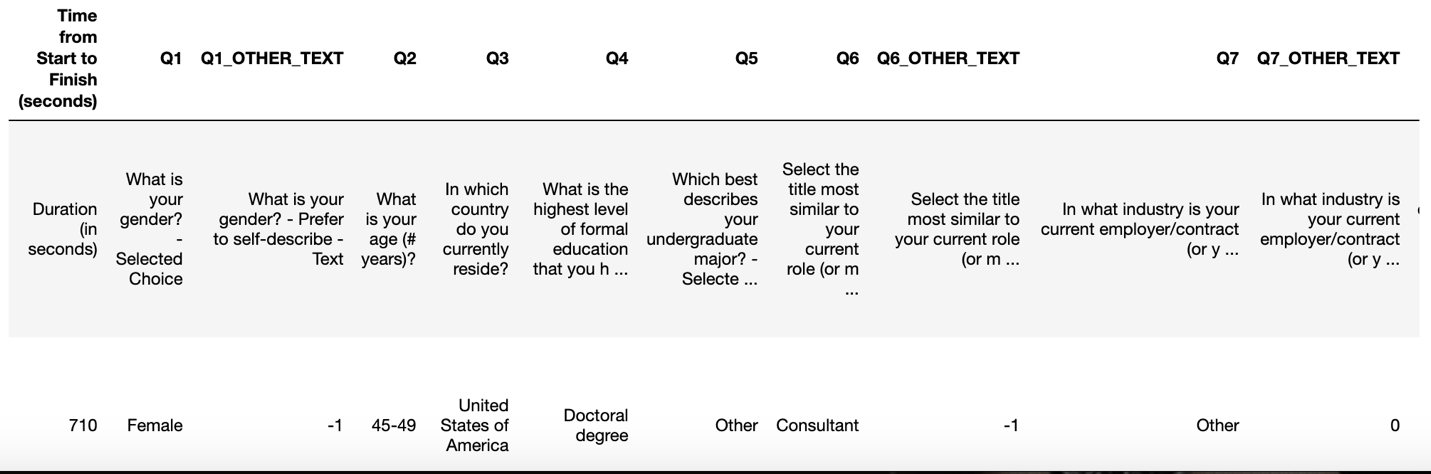
Sampling and Testing Hypothesis

# Data cleaning and EDA

The data set provided



It had a lot of variables and my analysis was only on gender and compensation

## Data cleaning 1: Remove columns not considered in the analysis

The only features retained are 'Gender','GenderDescription','Compensation','Above17','CurrentRole'. But later found columns 'GenderDescription' and 'Above17' did not contain any additional information that could aid in data analysis. Hence those were also removed.

## Data cleaning 2: Remove rows with missing values

Rows were removed if, for column Compensation was missing value.

## Data cleaning 3: Remove outliers

Rows were removed if, for column Compensation had value =500000+ as this could skew our results.

## Data cleaning 4: Remove inconsistent Data

Current Role column was examined to remove rows with value “not employed” and “student” as compensation or earnings by “not employed” and “student” were not consistant.

## Data cleaning 5: Remove partial data

Column Compensation had data where compensation was not revealed. This will not help in this analysis and those rows were removed from the data set. Same with gender column having value ‘Prefer not to say’ or 'Prefer to self-describe'

## Data cleaning 6: Convert compensation to integer

The range sting was converted to integer by first, slicing the range, convert to integer and finding the average of range.

## EDA

Gender: Counter({'Male': 10703, 'Female': 1903})

Compensation: Counter({5000.0: 2389, 15000.0: 1650, 25000.0: 1230, 35000.0: 1021, 45000.0: 911, 55000.0: 888, 112500.0: 820, 65000.0: 714, 75000.0: 662, 95000.0: 548, 137500.0: 522, 85000.0: 494, 175000.0: 446, 225000.0: 166, 275000.0: 74, 350000.0: 48, 450000.0: 23})

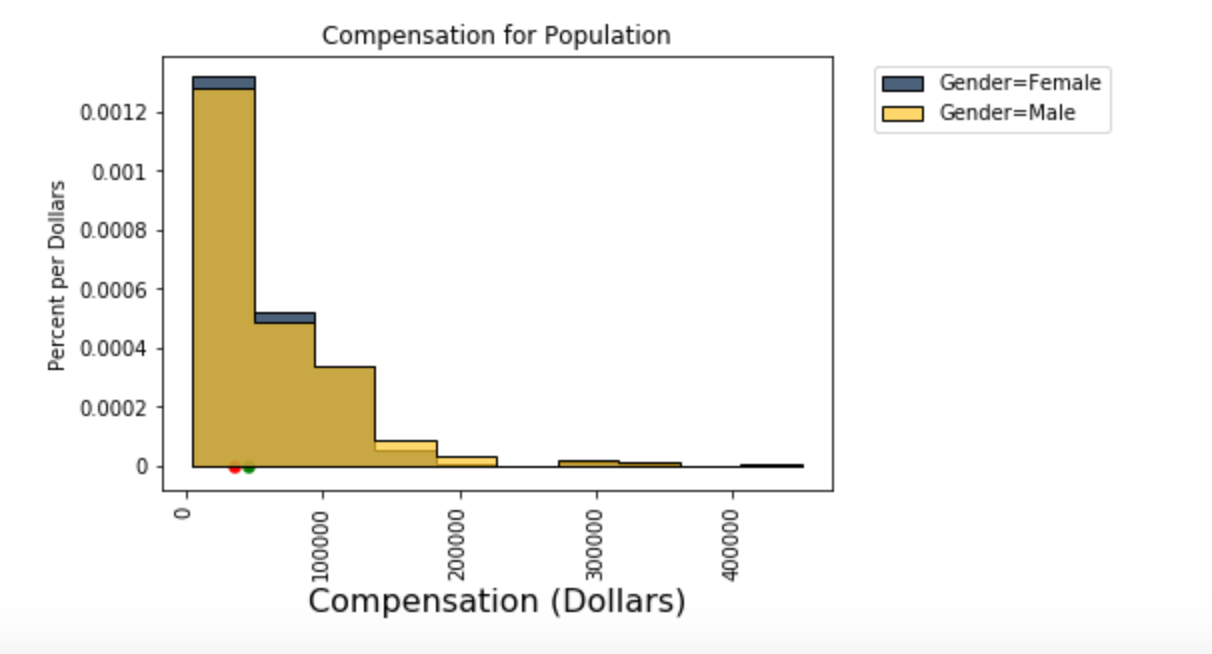
The median Compensation of the population is: 45000.0

The median Compensation of the Females in population is: 35000.0

The median Compensation of the Males in population is: 45000.0

The difference between median Compensation of the Females to Males in population is: -10000.0

The following graph was obtained for the compensation distribution of the population



# Test statistic

Since in the above figure there is a difference in median compensation based on gender, we select that as test statistics and analyze if this difference is due to chance or not. The absolute value of Female Median Compensation -Male Median compensation, is considered as our test statistics. And we aim to see if the difference in median compensation is centered around zero supporting notion that there is no statistically significant difference between male and femal compensation.

# Hypothesis

Based on the above the following are the hypostasizes of this study

* Null Hypothesis: Median Income of Male and female are same
* Alternative Hypothesis: Median Income of Male and Female differ considerably
* Test Statistics: We are going to simulate the median compensation of a random sample drawn from the population of survey responses. The absolute value of Female Median Compensation -Male Median compensation, is considered as our test statistics

# Random sampling

A random sample of size 500 is chosen from the population and the following histogram is plotted.

To ensure fairness, need to maintain the male to female ratio of the population in the sample

Sample Statistics

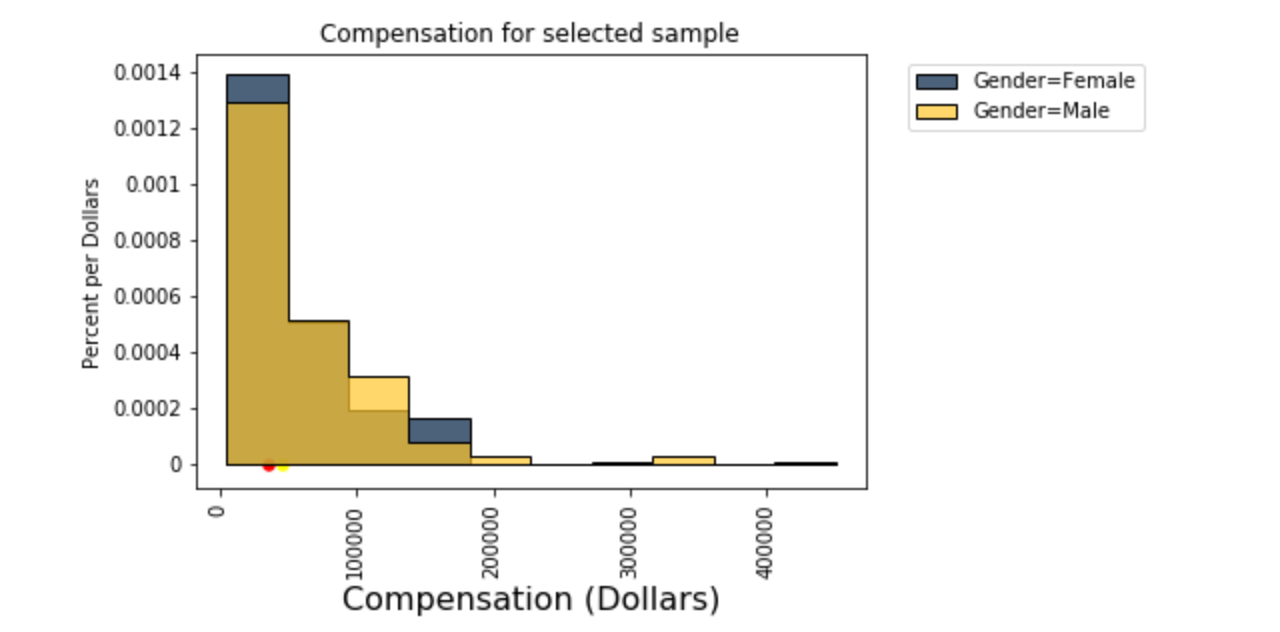
The median Compensation of the Sample is: 45000.0

The median Compensation of the Females in Sample is: 35000.0

The median Compensation of the Males in Sample is: 45000.0

The Difference in median Compensation of the Females to males in Sample is: -10000.0

A visible difference is evident from the histogram. To study the variability



To study the variability of the difference in median the empirical difference between group medians are calculated on 5000 samples of size 500 all taken from the population survey responses. The below given histogram is obtained.



The median difference is at 10,000.

The prediction under null hypothesis using shuffled Gender also is consistent with the above observation. Here, Predicting the Statistic Under the Null Hypothesis (no difference in compensation) no difference implies male or female does not make a difference in compensation and hence shuffling gender randomly. If there was no difference the distribution of difference after random shuffling of male and female should be around zero. But the following histogram contradicts that.

Hence the histogram reveals that there is a difference in median compensation based on gender and that is approximately around 10000. A random sampling distribution gives 75% results consistent with this



The empirical P-value is 1, meaning that, of the 5,000 permuted samples, almost all resulted in a difference of 10000.This is only an approximation. The exact chance of getting a difference in that range is not 1,but it is so close to one.

# Bootstrap

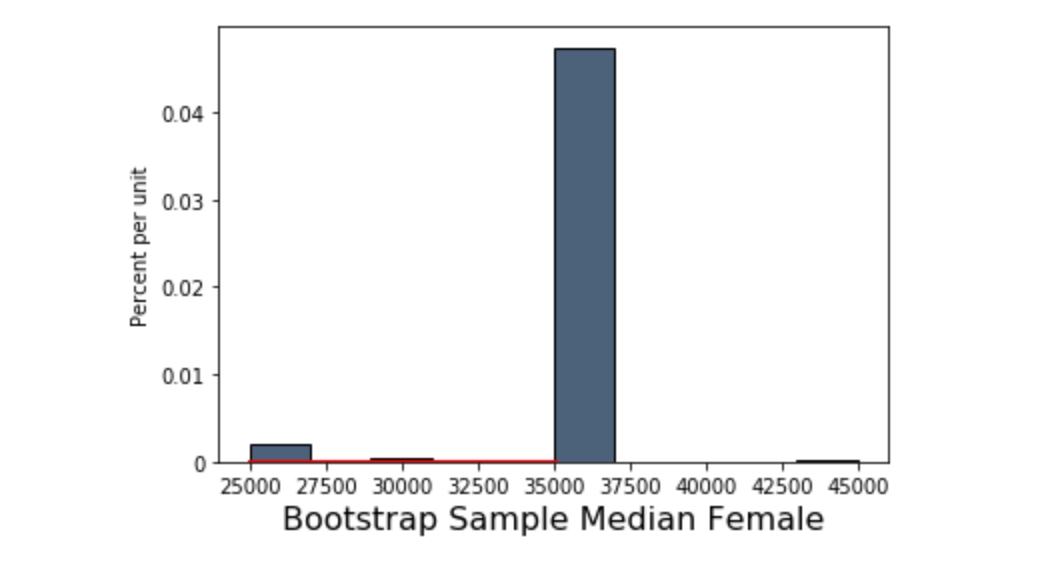
A random sample is bootstrapped 5000 times and the following are observed.

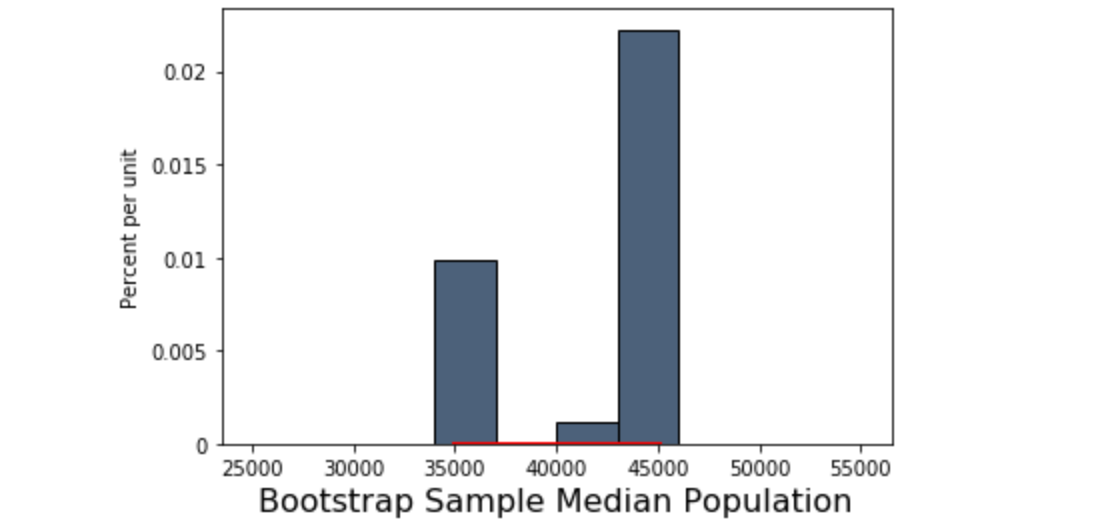
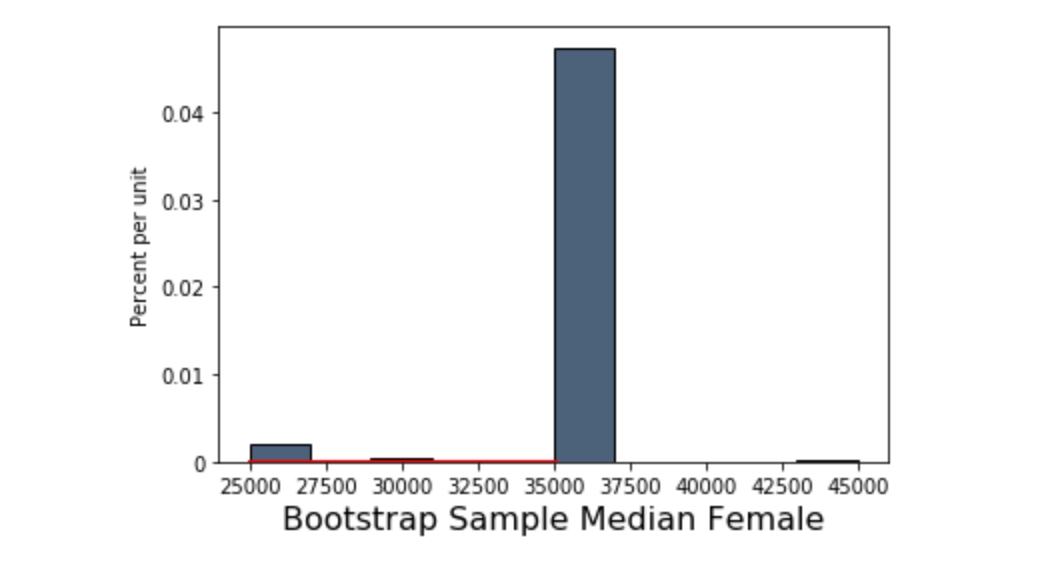
95%confidence interval for population [35000. 45000.]

95%confidence interval for female population [25000. 35000.]

95%confidence interval for male population [35000. 45000.]

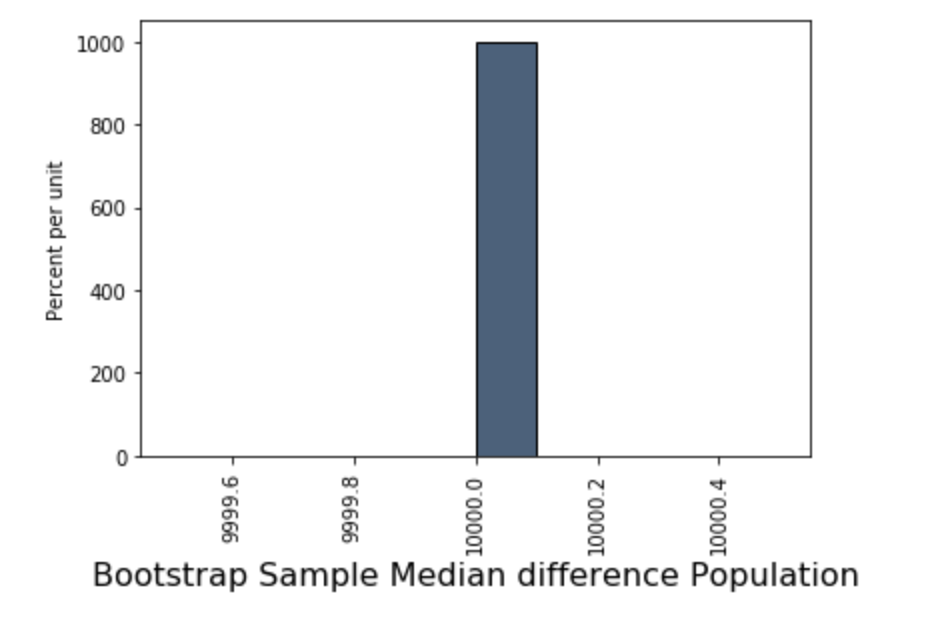
Bootstrapping the median produces the following histograms





These histograms are consistent with the observations we derived till now. That is population and male median range is almost same and is slightly more than female median compensation range by a total of 10000.

Now let us bootstrap for median difference and find if bootstrapping the difference in median provide us consistent results.



A histogram which is consistent with the behavior established so far.

99%confidence interval for population Median difference [10000. 10000.]

Also confidence interval is almost 100% sure about the median difference being 10000.

# Confidential Intervals

For median compensation

95%confidence interval for population [35000. 45000.]

95%confidence interval for female population [25000. 35000.]

95%confidence interval for male population [35000. 45000.]

For Test Statistics : Absolute value difference in Median

99%confidence interval for population Median difference [10000. 10000.]

1. P-values

The 99% confidence interval for the median difference in the population goes from about 10000.0 to about 10000.1. The interval doesn't contain 0. So, we reject the null hypothesis and state we are 99% confident the median difference is around10000 within 10000.1 and 10000.2

# Interpretation of Results

All of the studies conducted above gives a consistent result that the median compensation differs across gender and the difference is around 10000, with a confidence of 99%, at the least.