US Census Income Analysis

Neha Bais

Graduate Student

BU MET College Computer Science Department

MET CS-544

Summer’20

Objective:

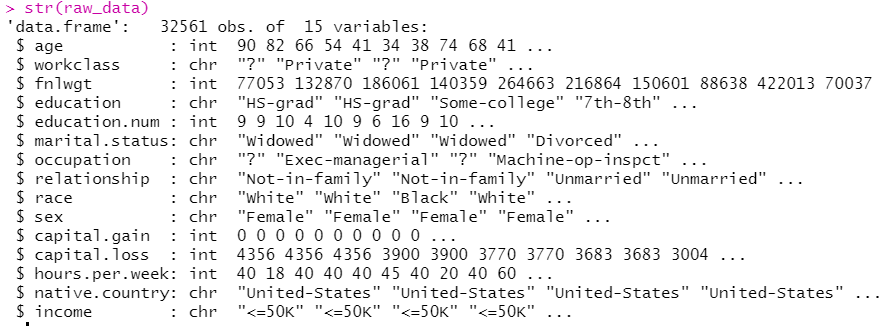
# The objective of this project is to visualize and understand various factors that affect an individual’s income as per the data collected from a survey conducted in 1994.

About Dataset:

# The dataset is extracted from Kaggle. <https://www.kaggle.com/uciml/adult-census-income/data>

# The data consists of 32561 observations and 15 rows. The data is pretty much clean except for some missing values in 3 columns (work class, occupation and native.country).

It consists of 9 categorical and 6 continuous variables.

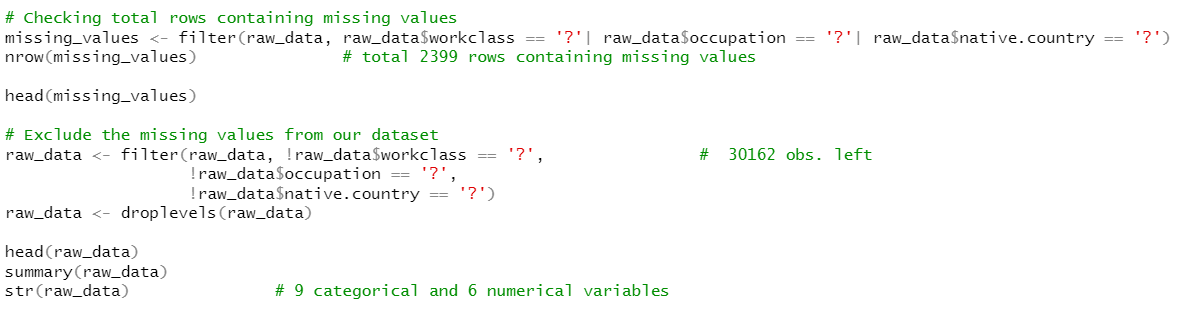


|  |  |  |
| --- | --- | --- |
| Age | Numerical | Age of the respondents of survey |
| Workclass | Categorical | Type of employer of the respondent |
| fnlwgt | Numerical | number of respondents that each row of the dataset represents. |
| education | Categorical | represents the level of education of the respondent |
| education.num | Numerical | represents the *education levels in numeric form*. |
| marital.status | Categorical | Married/Unmarried/Divorced |
| occupation | Categorical | represents the type of employment of the respondent |
| relationship | Categorical | represents the position in the family of the respondent |
| race | Categorical | represents the race of the respondent |
| sex | Categorical | Male/ Female |
| capital.gain | Numerical | represents the income gained by the respondent from  sources other than salary/wages. |
| capital.loss | Numerical | represents the income lost by the respondent from sources  other than salary/wages. |
| hours.per.week | Numerical | represents the hours worked per week by the respondent |
| native.country | Categorical | represents the native country of the respondent. |
| income | Categorical | contains yearly income of respondent (<=50K and >50K) |

Data Processing / Data Cleaning:

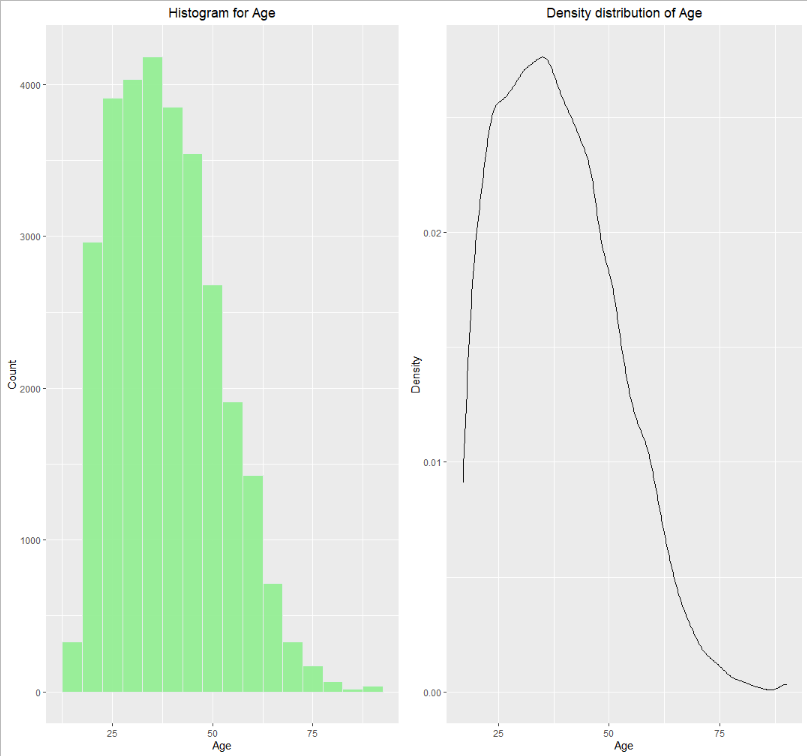
Identified the rows that contain the missing values in those 3 columns.

Total rows containing missing values = 2399



Analyzing Data:

As the first step, I visualized the continuous variable – Age, and their distribution across the data.

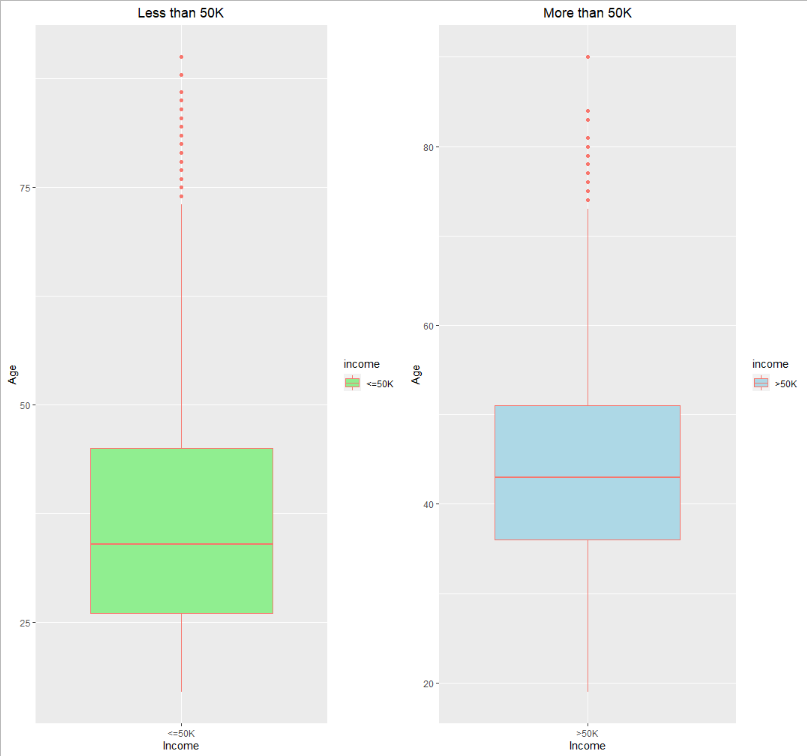


The ages range from 17 to 90 years old with the majority of entries between the ages of 25 and 50 years

The average age of individuals who responded to this survey is 39 years.

**Age vs Income**

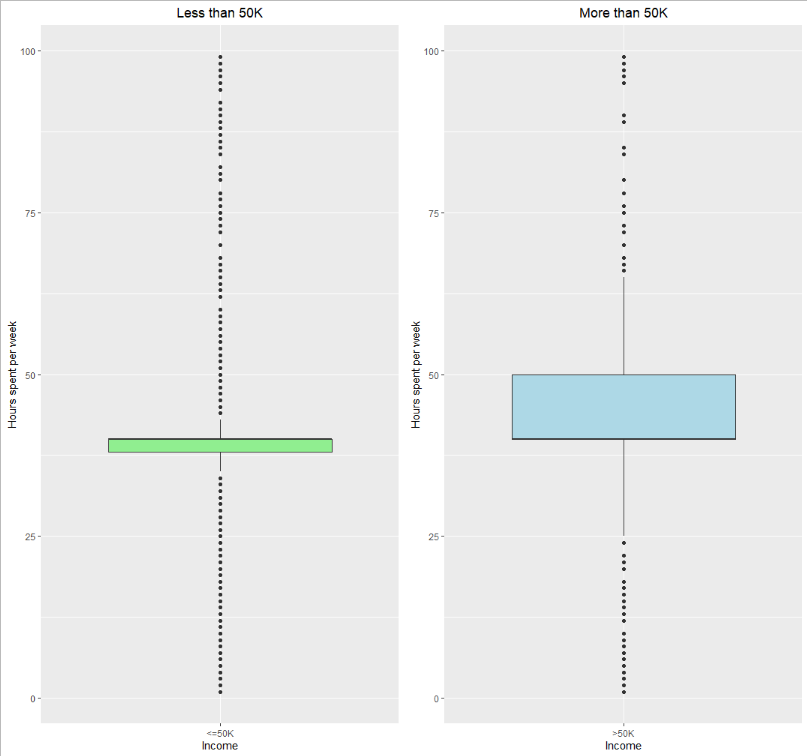
Age being continuous and Income as categorical variable, relationship between them is observed as below:



Median age of people earning >50K is greater than median age of people earning <50K

Older people are more likely to earn more compared to younger ones.

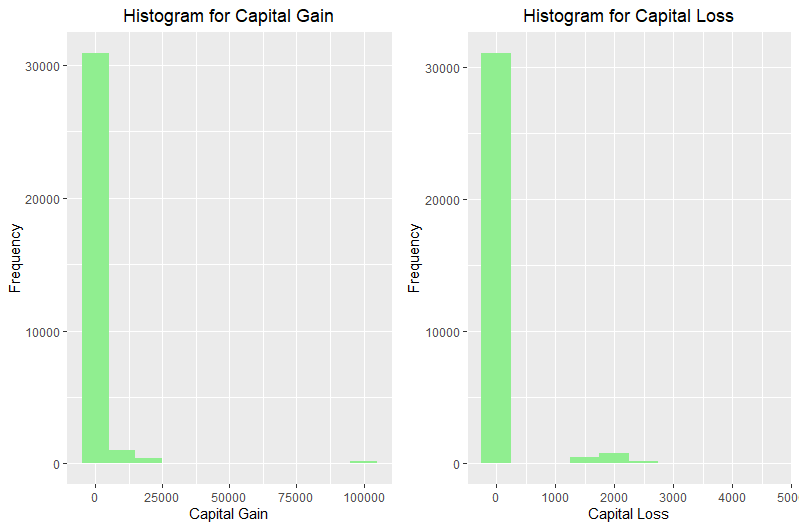
**Hours per week vs Income**



People who spend more time are likely to earn more.

The outliers in both the plots represent variations in both groups. Some jobs demand more working hours but pay less, while some pay higher even for less working hours.

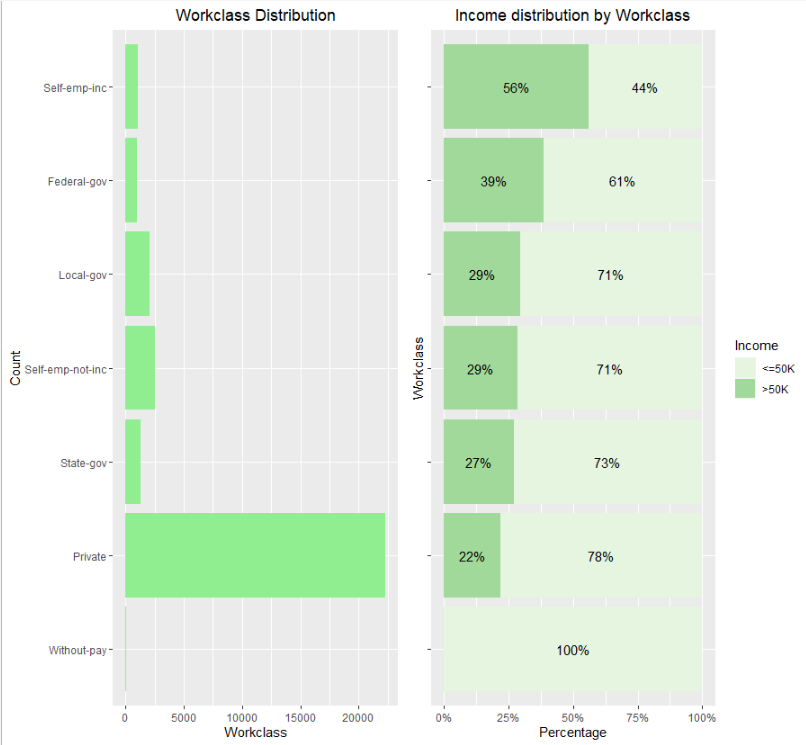
**Capital Gain and Capital Loss distribution:**



We can see that most of the individuals that responded to the survey have no income other than salary.

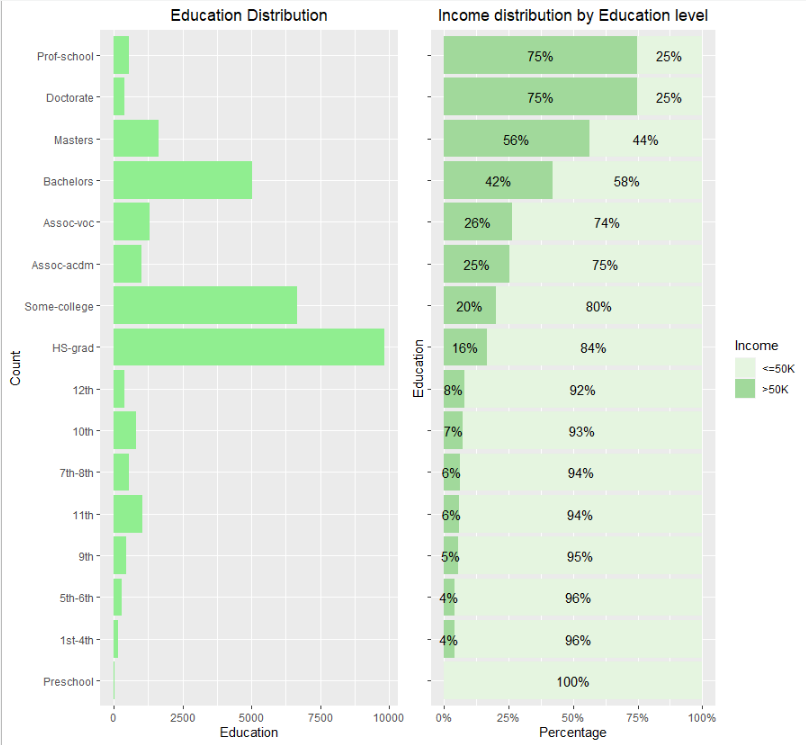
Very few of them have gained some profit or in loss through other income.

**Work class vs Income**



People who responded to this survey mostly work for Private sector. Federal govt employees and people who are self-employed have a higher percentage of earning >50K.

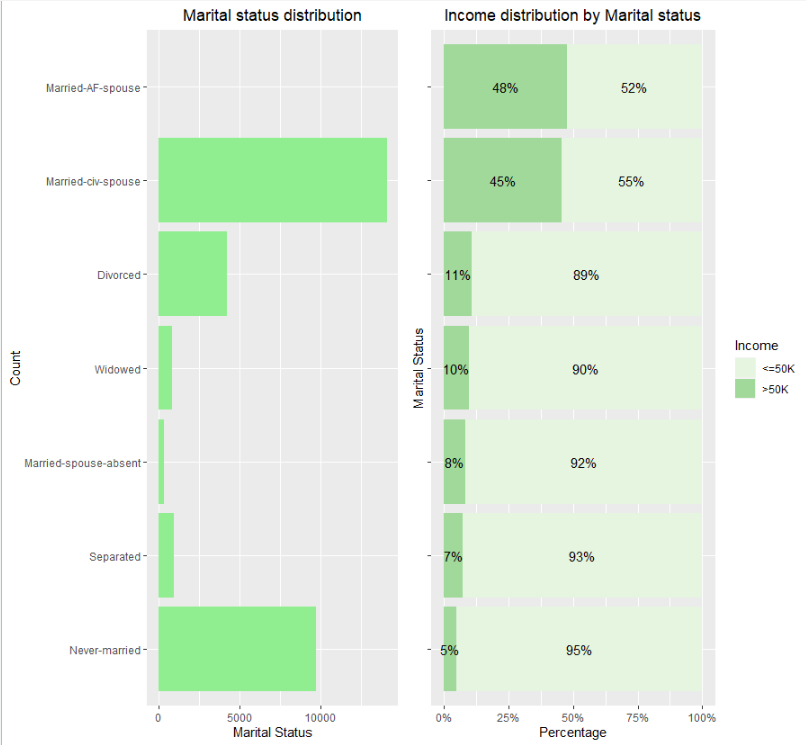
**Education vs Income**



As per the plot, it is clear that most of the respondents of this survey have completed their high school, gone to some college or have bachelor's degree.

People of Prof-school, and those who have doctorate and have done masters are more likely to earn > 50K.

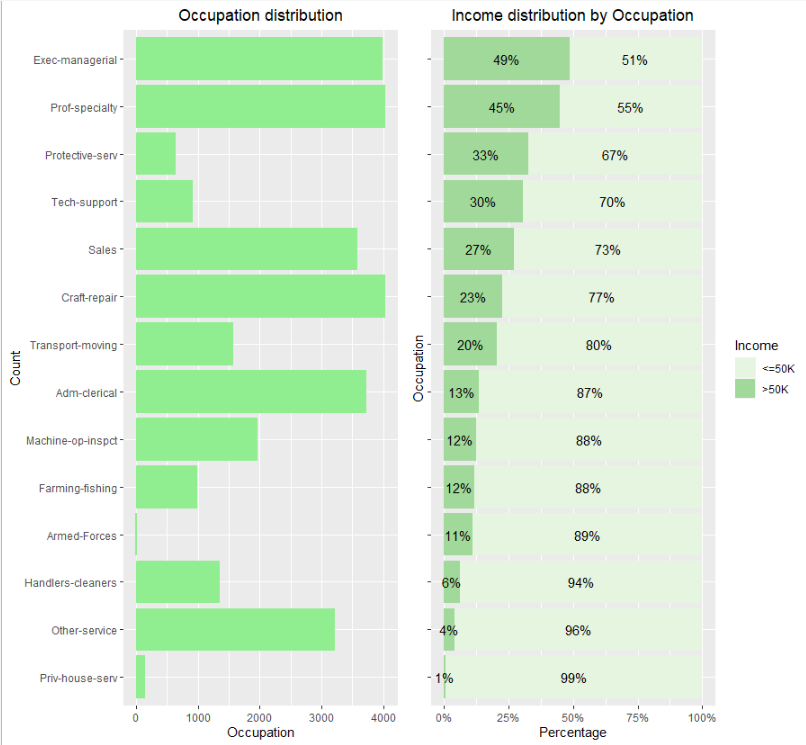
**Marital Status vs Income**



Most of the respondents are married civilians or Never married.

As per this survey, married people earn more as compared to not married ones.

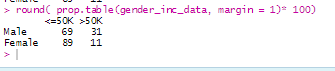
**Occupation vs Income**



As per the graph, people at a position of Executive management and professional specialization have a higher probability of earning more than 50K.

**Sex vs Income**





Created a contingency table to see what % of male and female earn more than 50K and less than 50K.

Out of total male population - 69% of male earn less than 50K & 31% male earn more than 50K.

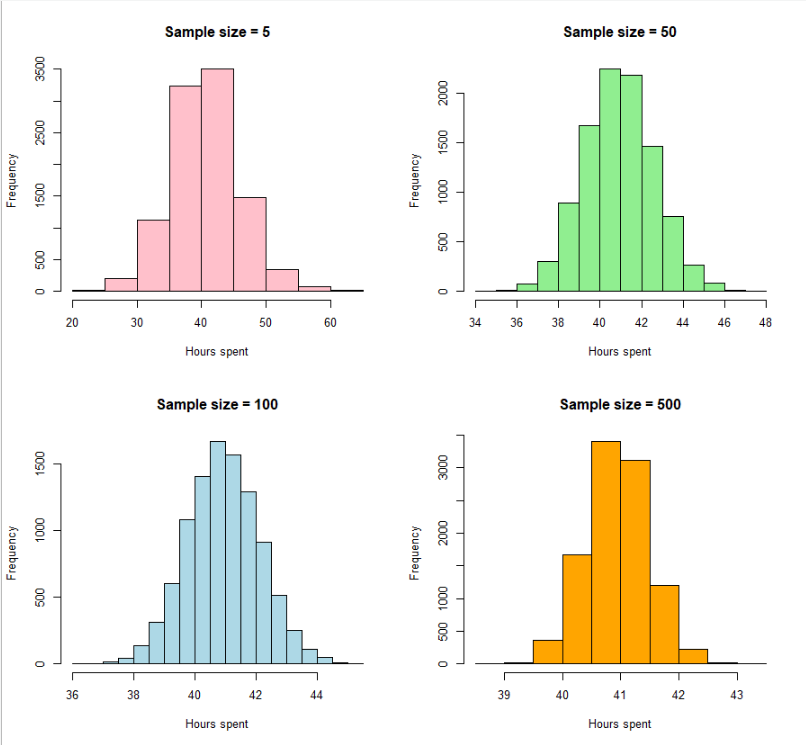
Out of total female population -89% female earn less than 50K & 11% female earn more than 50K.

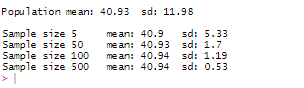
We can see that there is almost double the sample size of male as compared to female in the dataset.

% of male who earn more than 50K is much greater than the % of female who earn the same amount.

**Central Limit Theorem**

Applying central limit theorem on hours per week using 4 different sample sizes.

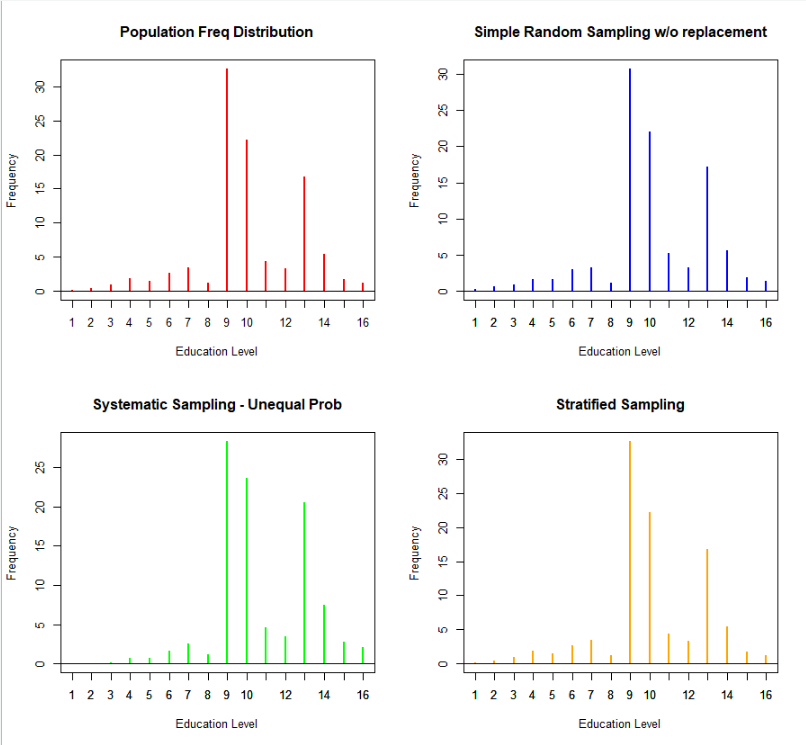


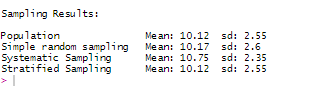


As the sample size increases, the distribution tends to be a normal distribution. Mean of samples same as population, and standard deviation reduces as sample size increases.

**Sampling**

Performed 3 sampling methods on education.num variable taking a sample of 3000.



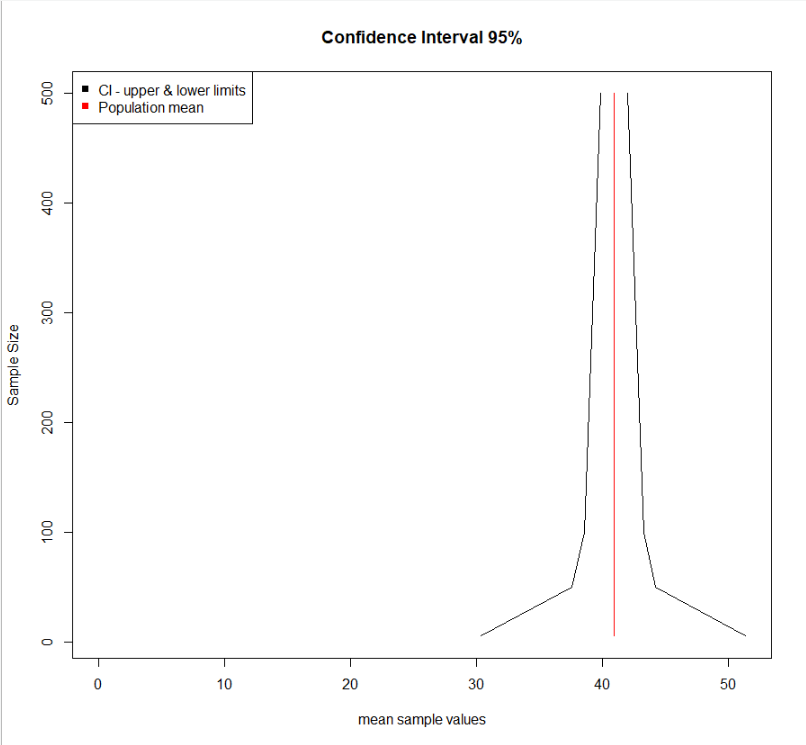


Stratified sampling gave a close approximation for our population.

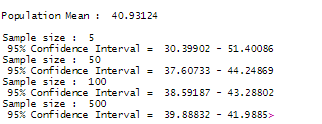
So, we can say that if we take a sample of population using stratified sampling, we can get a boundary close to our entire population and can predict population estimates.

**Confidence Interval**

Next I tried to find the lower and upper limit between which our population mean will fall using 95% confidence interval and 4 different sample sizes on hours.per.week variable.



Here are the lower and upper boundaries between which our population mean will fall for different sample sizes.



**Conclusion**

* Cleaned the dataset by dropping all the rows containing missing values.
* Explored various categorical and continuous variables.
* Tried to explore the relationship between different variables.
* Performed Central limit theorem for 4 different sample sizes.
* Performed 3 sampling methods using a sample of 3000 records and find out that stratified sampling can be used if we want to estimate the population data using some sample data.
* Lastly, find the lower and upper limits within which population mean can fall using 95% Confidence Interval for 4 different sample sizes.