IE 6200 Final Project
Adult Census Income

Yufei Wang
12/6/2019

Introduction

The chosen dataset was extracted from the 1994 Census Bureau Database by Ronny Kohavi and Barry Becker[1]. It contains key characteristics of the United States' population in 1994, and the basic description can be found at UCI Machine Learning Repository[2]. Based on my desk research, census data are collected via various ways. In general, for every ten years, households will have the option of responding online, by mail, or by phone to answer the census questionnaire. The census process also includes special provisions to count people who are homeless and those in other types of living quarters, such as college dorms and ship [3]. The Census Bureau will expect that many people to complete the questionnaire based on formal instructions. Getting an accurate count is important because census numbers impact daily life in many ways. It shows how many people live in a given area that determines the number of representatives who will be in Congress as well as the numbers of electoral votes. It also shapes how federal tax dollars get allocated to certain areas to support local housing, transportation, and other services. [4]

The data collection method of chosen dataset includes a national sample of 57,000 housing units in 1,973 counties were personally interviewed. However, compare this method with the traditional census process, there are two inevitable biases: sampling bias and response bias. Sampling bias is introduced when some members of the intended population are less likely to be surveyed than others, and it can cause an overestimate or underestimate of the observations. Since all respondents were interviewed personally when they did the questionnaire, some people who live in remote areas were hardly chosen for this questionnaire. The way to remedy this trouble is to obey the Census Bureau`s instruction to conduct the questionnaire. Response bias comes from less-than-truthful responses. Respondents may not be willing to uncover the drug histories or any private medical histories. The way to reduce it is to ask more neutrally and generally worded questions and to make sure that the answer options are not leading.

According to UCI, data was extracted using the following rules: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)), which means individuals with:

- Age > 16
- Adjusted gross income > $100
- Final weighting > 1
- Working hours per week > 0

The extracted data set has 48,842 records and was randomly split into training and test set at ratio 2/3 and 1/3 respectively. According to UCI , the chosen dataset only contains the training data set, which includes 32,561 records and 15 attributes, and it will be analyzed in this report. In general, each record contains data of one respondent and each column describes one feature of that person. For the chosen dataset, ten variables are categorical variables and five variables are continuous variables as shown in **Table 1**

Table 1 Variable type and description summary

Variables	**R class**	**Data type**	**Descriptions**
education	factor	categorical	highest level of education that individual achieved
education_num	integer	categorical	highest level of education in a numerical form

income	factor	categorical	(target) whether or not the person makes more than $50,000 per annually income
marital_status	factor	categorical	marital status of the individual
native_country	factor	categorical	country of origin
occupation	factor	categorical	occupation of the individual
race	factor	categorical	the individual's ethnicity
relationship	factor	categorical	primary family relationship individually
sex	factor	categorical	gender
workclass	factor	categorical	type of employer the individual has
age	integer	continuous	age of the individual
capital_gain	integer	continuous	capital gains (from selling an asset such as a stock, bond for more than the original purchase price)
capital_loss	integer	continuous	capital losses (from selling an asset such as a stock, bond for less than the original purchase price)
fnlwgt	integer	continuous	number of people in the universe (U.S. population) that each individual in the survey can represent
hours_per_week	integer	continuous	hours worked per week

Note: fnlwgt is weighting value that ensures sample from this survey can represent the 'true status' of U.S. population. In general, a respondent gets a high weight if the proportion of that type of people in the survey is small compared to its size in the whole population. When applying machine learning method, this variable might be used as a weighted field to give extra importance to underrepresented groups [1].

From the **Table 1**, there are some interesting features that can be used as general guidance for pre-processing or further exploration:

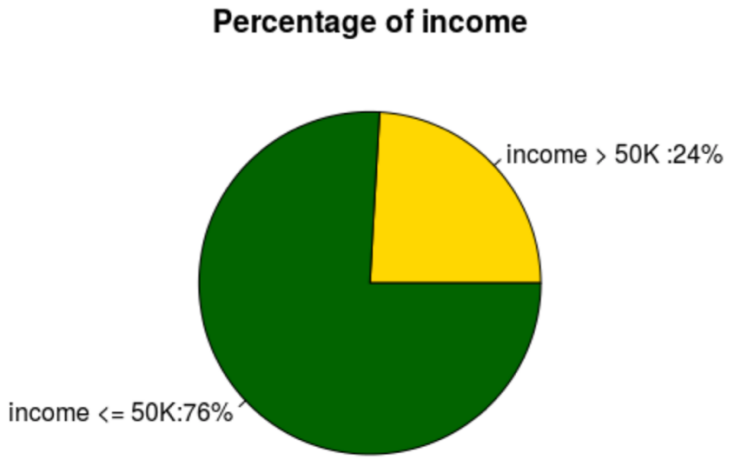
- education and education_num are same as they contain the same information in different forms.
- capital_gain and capital_loss can be considered to combine into one attribute as it can represent one feature. Capital is the term for the financial assets such as funds held in deposit accounts and funds obtained from special financing sources.

The target variable of the chosen dataset is income, since it described whether or not the person makes more than $50,000 per annually income. A better understanding of it is essential, because the statistic about income data can measure the economic well-belling of the nation and help the government to determine the distribution of food, health care, housing, and other assistance [5].

Exploratory Analysis and Data Visualization

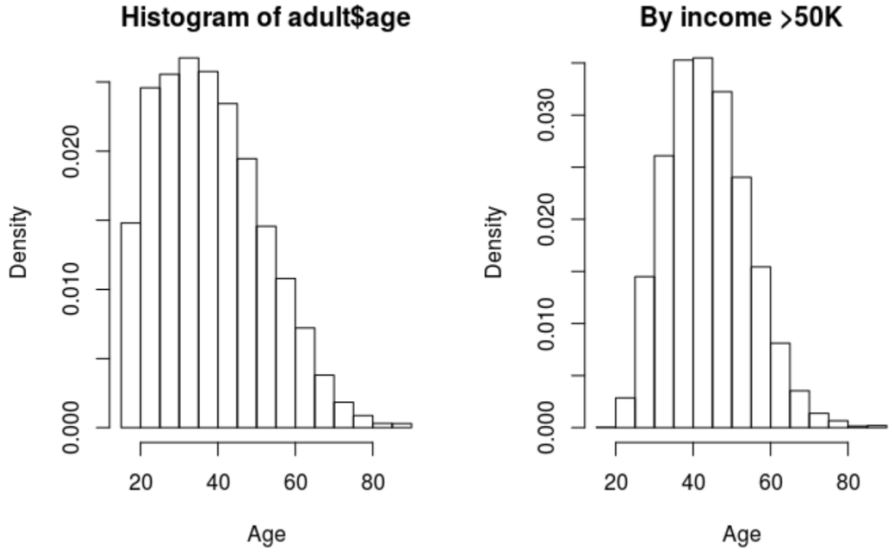
This dataset contains weighted census data extracted from the 1994 to 1995 current population surveys conducted by the US Census Bureau. The data includes 32,561 records, ten categorical variables, and five numerical variables. The income is the target variable in this report. For a better understanding of the important aspects of the chosen dataset, an exploratory analysis is conducted through the proportion distribution of high income **(Figure 1)**, the density distributions of key continuous variables filtered by income (>50K). **(Figure 2-Figure 6) ,** the proportion distribution of high income between male and female **(Figure 7)**, the boxplots of key categorical variables**(Figure 8-Figure10)**, and the correlated plot of all continuous variables**(Figure 11)**.

Figure 1 The pie chart of income



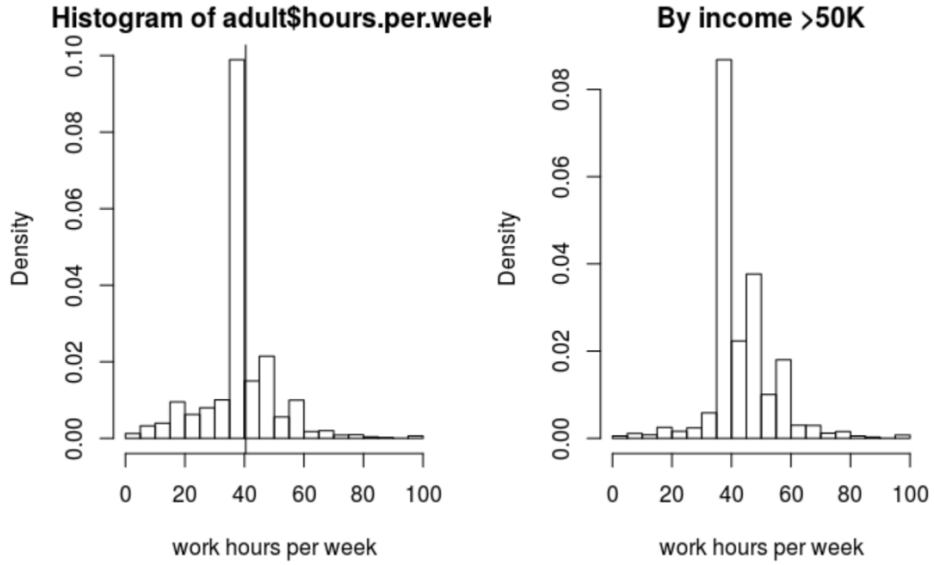
Based on **Figure 1,** 76% of respondents have the annual income that is less than 50,000 USD and 31% of respondents have the annual income that is more than 50,000 USD.

Figure 2 Age distribution & By income groups



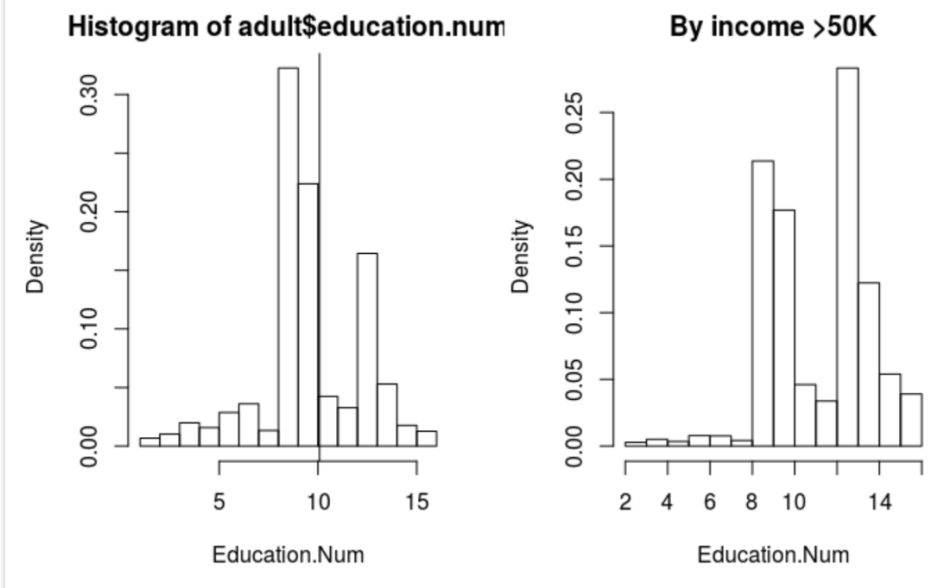
Based on the **Figure 2,** the age distribution is right-skewed due to the data extraction rule that age is larger than 16. It can be seen that people who aged from 35 to 60 are more likely to have a higher income than the average. In other word, it is very unlikely that someone age under 20 could earn more than 50,000 USD in 1994.

Figure 3 Hours per week distribution & By income groups



The average hours worked per week are around 40. Moreover, 47% of respondents work exactly 40 hours per week. It can be seen that people who worked above 40 hours per week are more likely to have a higher income than the rest. On the other hand, longer hours worked per week is not the causality for the higher annual income.

Figure 4 Education number distribution & By income groups



As mentioned in the introduction, education and education_num contain the same information. Higher education levels should result in better employment opportunities and consequently, higher income. From **Figure 4**, high-school graduate, some college and bachelors are the most common education levels, which accounted for 70% of survey responses. It can be seen that people who own the doctorate degree are the most likely to have a higher income than the rest degrees. People who own the prof-school and master's degree have the second and third likelihoods. Therefore, education levels indicate likelihood of higher incomes.

Figure 5 Capital gain distribution & By income groups

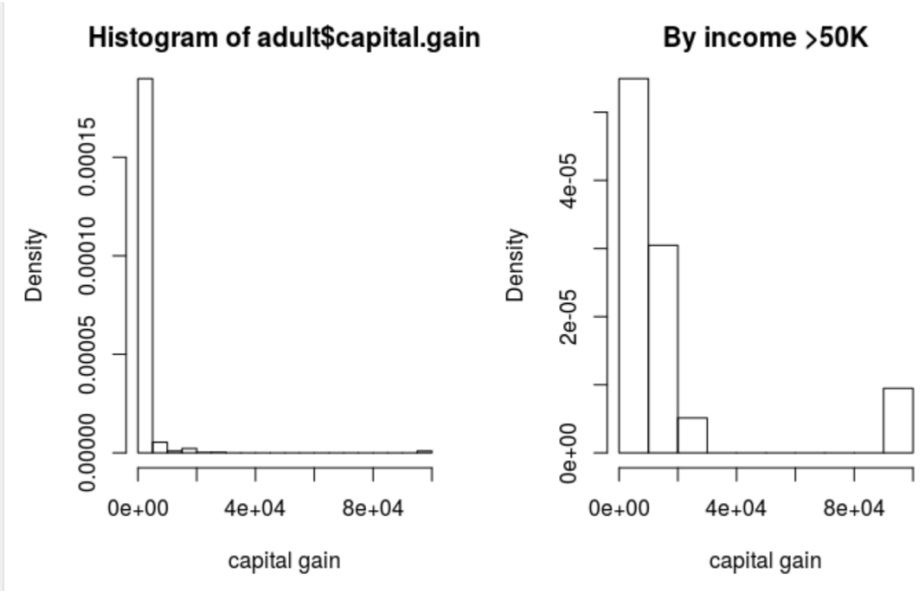
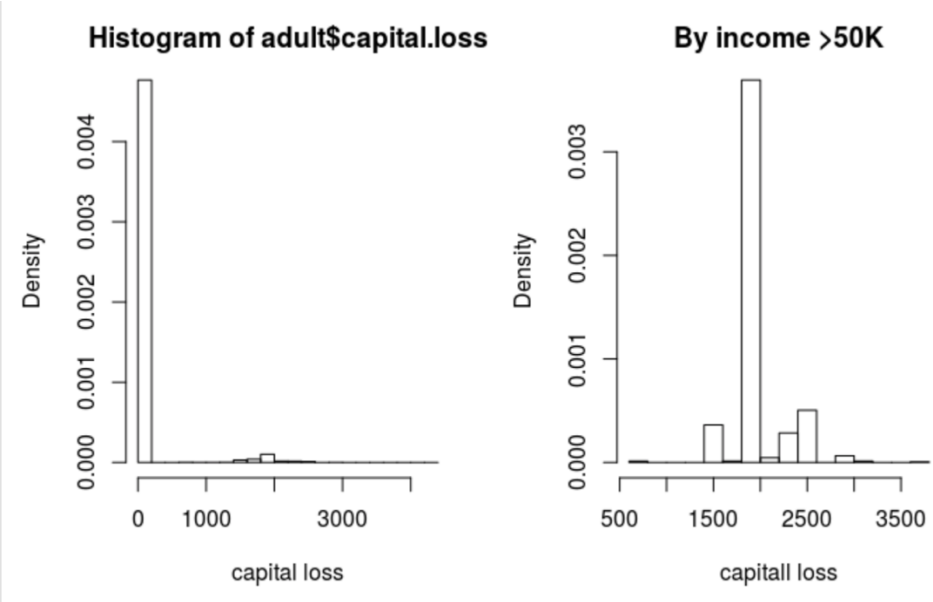


Figure 6 Capital loss distribution & By income groups



From the left parts of **Figure 5** and **Figure 6**, the capital gain and capital loss, the densest distributed parts are all equal to zero, which reflects the fact that at least 75% of people in this data set has no capital gain and capital loss. The distributions of these variables are all extremely right-skewed. For the higher income, people who have the zero capital gain have more likelihoods. People who have the 2,000 USD capital loss are more likely to have higher income

Figure 7 The proportion bar chart of high income between female and male

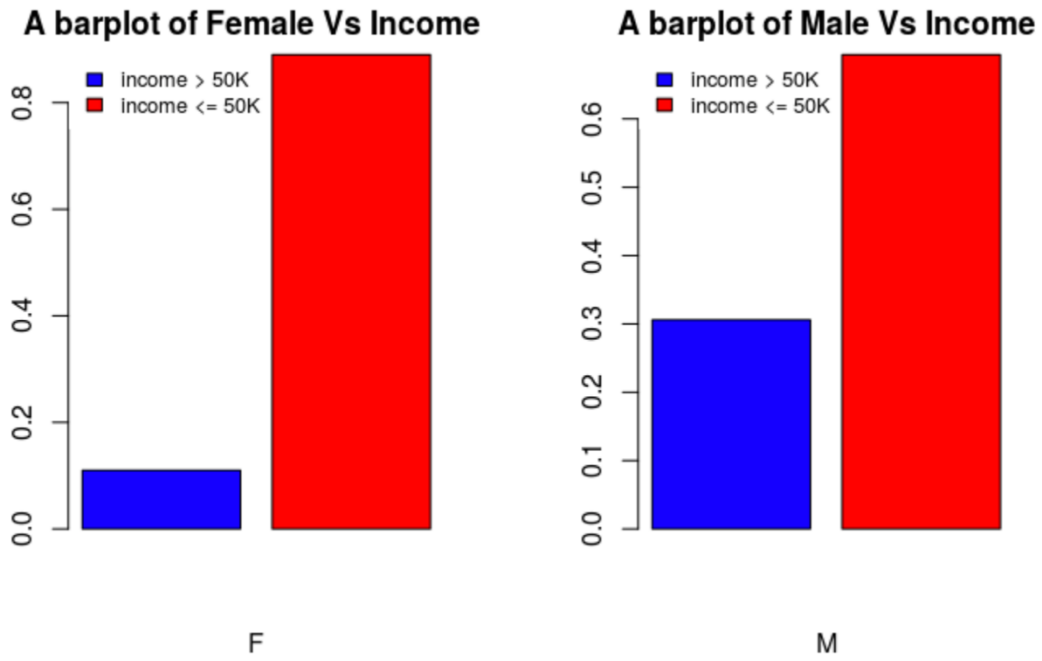


Figure 7 shows the proportion bar chart of high income between female and male. 11% of female respondents have higher income over 50,000 USD annually, and the proportion is 30.6% for male respondents, indicating that a larger proportion of males are in the higher income bracket. For this chosen dataset, 70% of the work class is private business, as an entrepreneur, a male has more advantages than a female in business world. Therefore, sex does play an important role in determining income levels.

Boxplots are a standardized way of displaying the distribution of data based on a five number summary (minimum, first quartile, median, third quartile, and maximum). Also, they can show outliers and what their values are. From **Figure 3,** longer hours worked per week seems not the causality for the higher annual income. More studies are focused on the distributions between categorical variables and the hours worked per week. From **Figure 8 to Figure 10,** the medians show 40 hours worked per week. The extreme values (100 hours worked per week), which seems unrealistic, exists in all three boxplots.

Figure 8 The boxplot of Work class and Hours worked per week

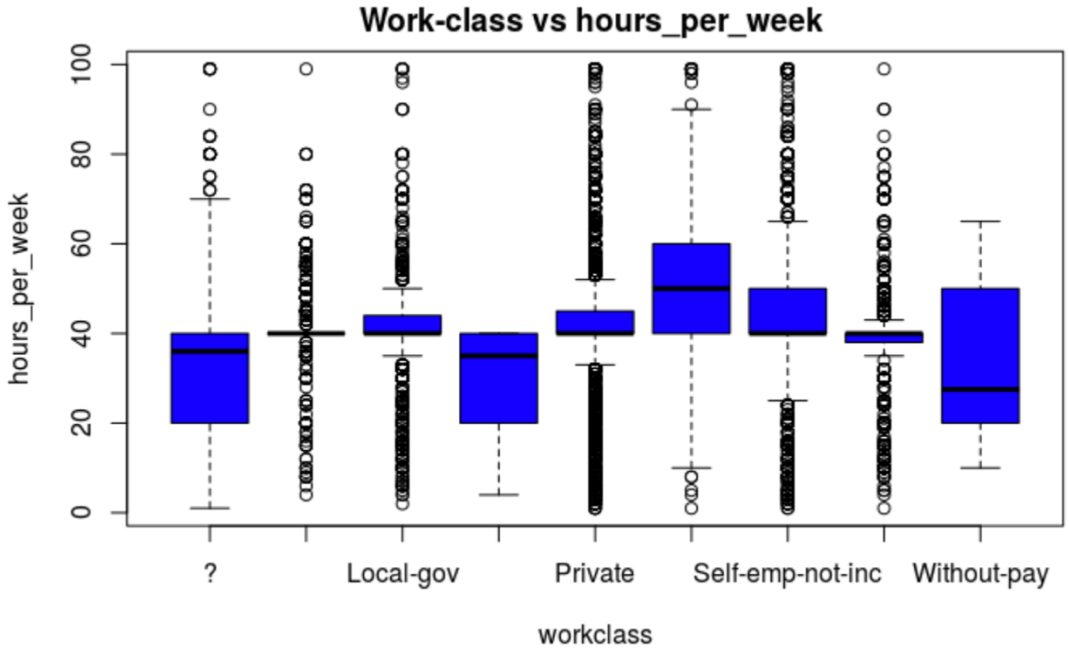


Figure 9 The boxplot of Education and Hours worked per week

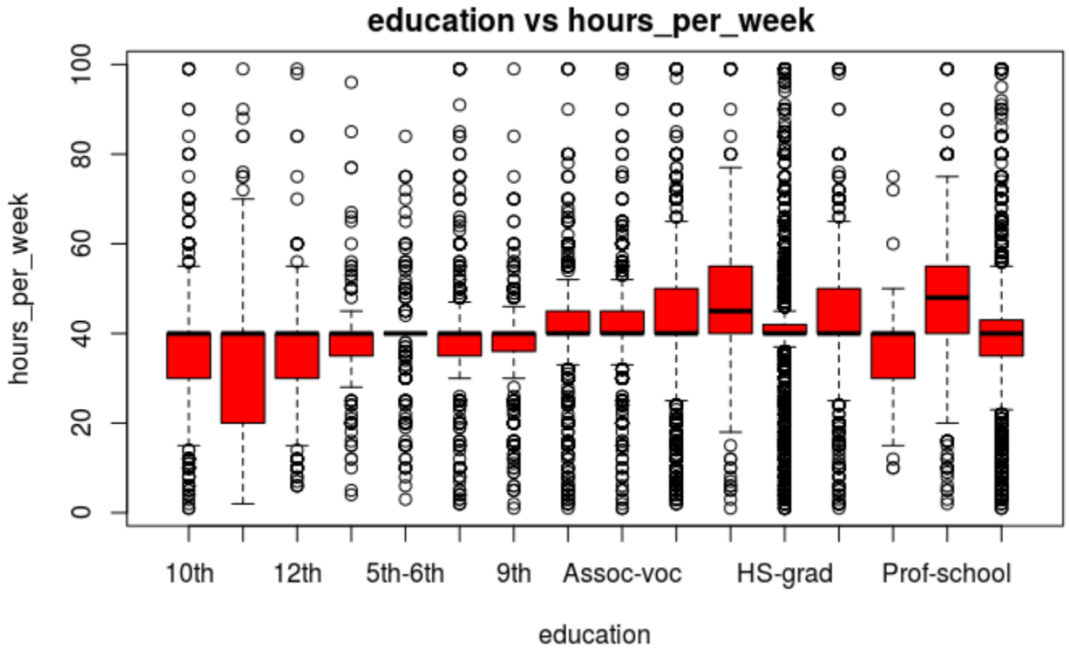


Figure 10 The boxplot of Occupation and Hours worked per week

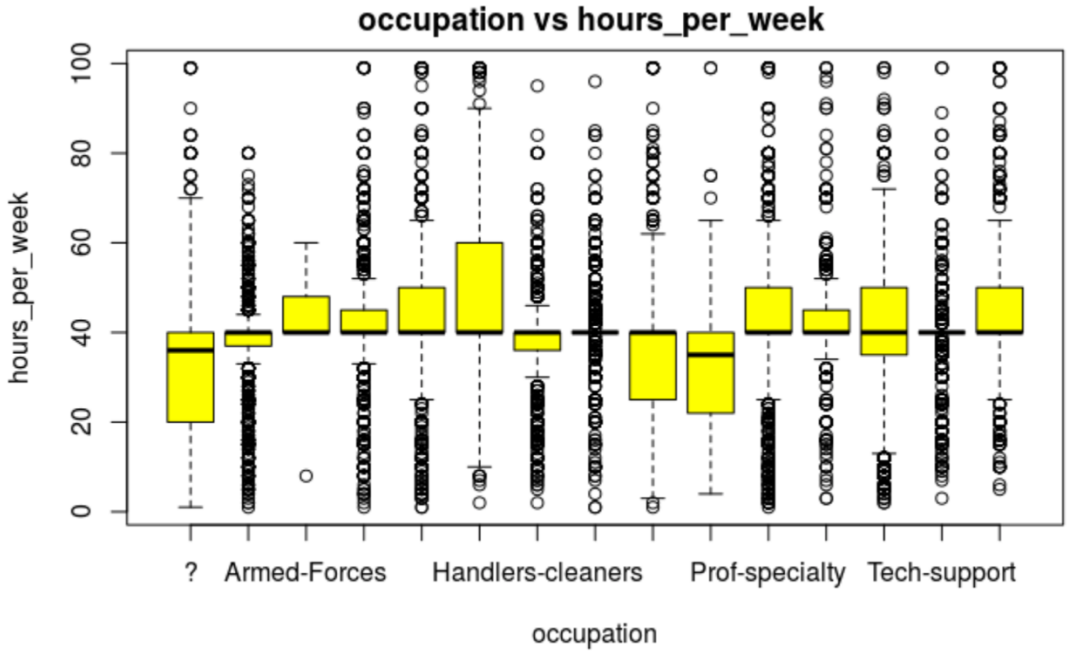
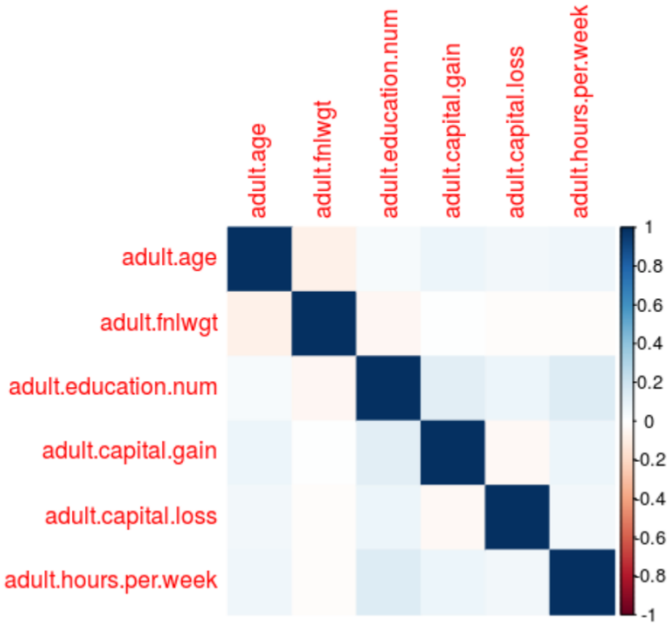


Figure 11 The correlated plot of all continuous variables



Correlation is used to describe the linear relationship between two continuous variables. This correlated plot measures the strength(qualitatively) and the direction of the linear relationships. As you can see from **Figure 1,** the correlation is 0 for each horizonal variable and

vertical variable. Therefore, every two continuous variables in this chosen dataset are said to be uncorrelated.

Statistical Analysis

1. **One sample t-test**
   - Traditional Statistical Tools

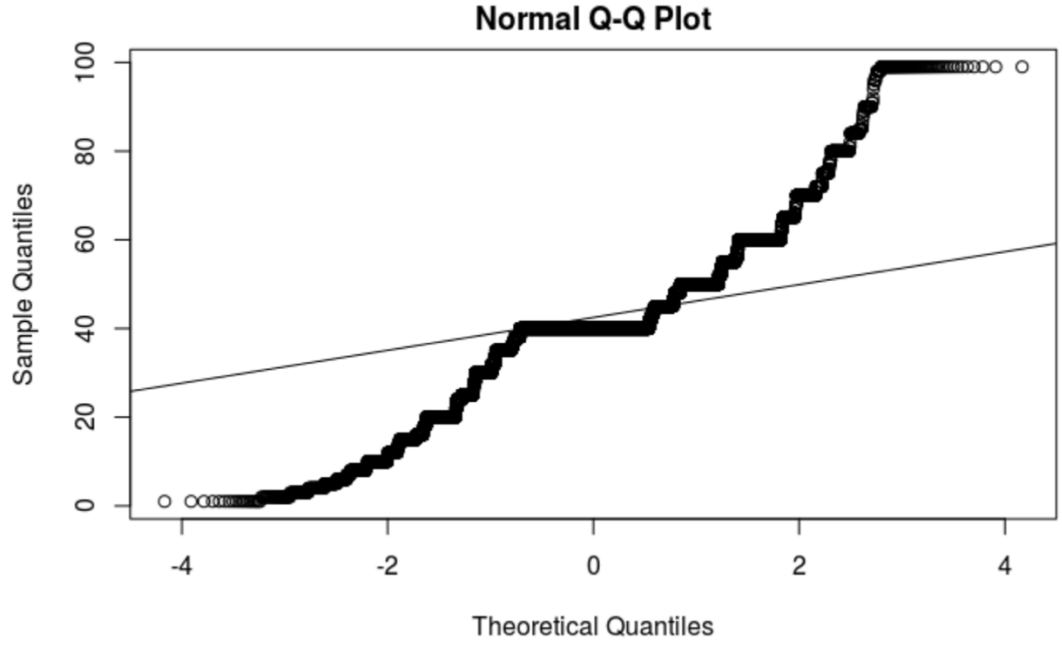
According to the Bureau of Labor Statistics, the average American works 44 hours per week[6]. The question of interest is to compare the sample mean hours worked per week with theoretical hours worked per week, and to make inference about the population mean. The statistical one sample t-test is used for this statistical test, because the true population variance is unknown.

- Conditions for use and checks for conditions

The sample is representative of the population- this chosen dataset is extracted from the 1994 to 1995 current population surveys conducted by the US Census Bureau.
One quantitative variable of interest-yes, the hours worked per week is a continuous variable, therefore it is quantitative.
We want to make inference about the population mean using the sample mean-yes
The population variance is unknown, so we estimate it using sample data-yes
The sample comes from a single population-yes

- The population data must be normally distributed-need to check
  1. To check this condition, look at the QQ plot of the sample data
  2. Particularly important if the size is less than 30—in this case, n=32561 but we should still check that the data look normal

Figure 12



This plot indicates that the points fall along the y=x line with the most points concentrated in the center of the line and fewer points towards the ends. There are more points concentrated in the middle and less towards the edges, but this is far from the straight line. The normality

assumption is not completely met, and we should be concerned about the reliability of p-value and the confidence intervals.

- Parameter: the population parameter we want to make inference to is the population mean (μ)
- Hypotheses
  - Two-sided
  - The null hypothesis, H_0: μ_0 =44
    The true population mean of hours worked per week is 44 hours
  - The alternative hypothesis, H_1: $\mu_0 \neq 44$
    The true population mean of hours worked per week is not 44 hours
- Sample Statistic
  The sample statistic is the sample mean hours worked per week: $\bar{x}_1$
- Test Statistic
  When we do not know the population variance and we have to estimate it with the sample mean, the reference distribution of the test statistics shifts to the t-distribution. The shape of the t-distribution depends on the sample size. As n approaches infinity, the t-distribution approaches the normal distribution.
- Distribution of the test statistic

$$t_{n-1} = \frac{(\bar{x}_1)-(\mu_0)}{\sqrt{\frac{s}{n}}} \sim t_{n-1}$$

- Result
  I did two-sided one sample t test with 95% confidence interval and put detailed R codes for t-test of a knitted R markdown file in Appendix.

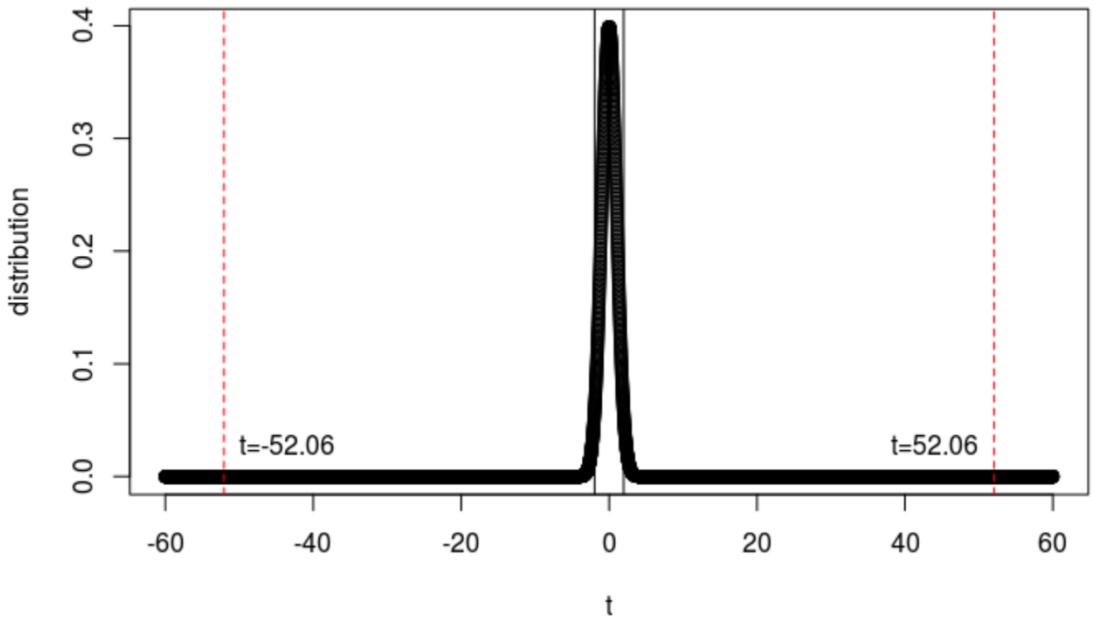
```
        One Sample t-test

data:  adult$hours.per.week
t = -52.063, df = 32560, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 44
95 percent confidence interval:
 40.30334 40.57158
sample estimates:
mean of x
 40.43746

[1] -52.06341
```

- P-value
  P-value <- $2.2*10^{-16} \approx 0$
- Confidence interval
  The lower bound is 40.3
  The upper bound is 40.57

Figure 13 Histogram of the sampling distribution



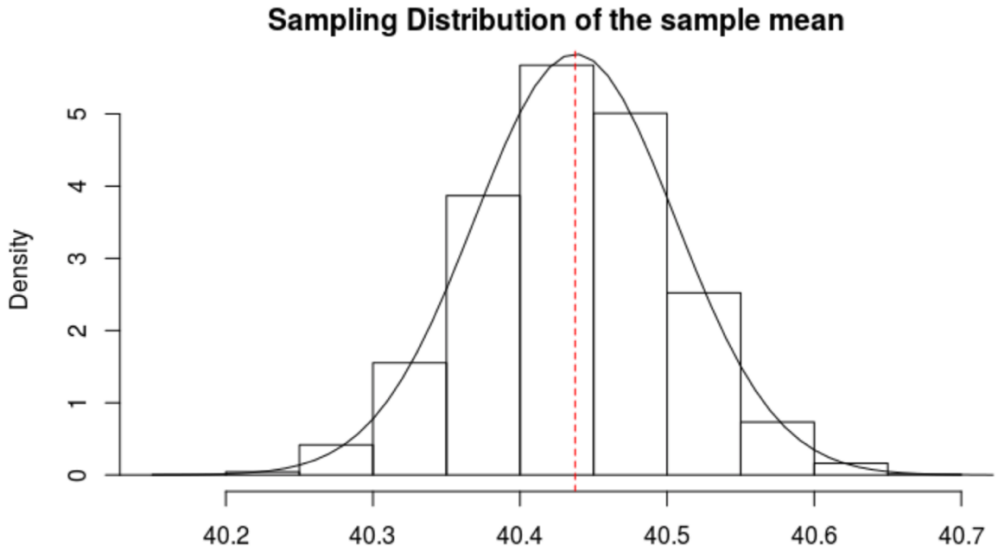
The curve represents the t distribution with the same degree of freedom of the original dataset. The two vertical red dashed lines represent the two test statistic values. The two black vertical lines represent the critical t values associated at the α =0.05 level and (n-1) degree of freedom.

- Interpretation

There is strong evidence to suggest that we reject the null hypothesis since P-value from the hypothesis test is smaller than 0.05 (p-value $\approx$ 0),which indicates that the true population mean for hours worked per week does not equal to 44 hours at the α =0.05 level. With 95% confidence, the true mean hours worked per week is between 40.30 hours and 40.57 hours, which suggests that the true mean hours worked per week is less than 44 hours.

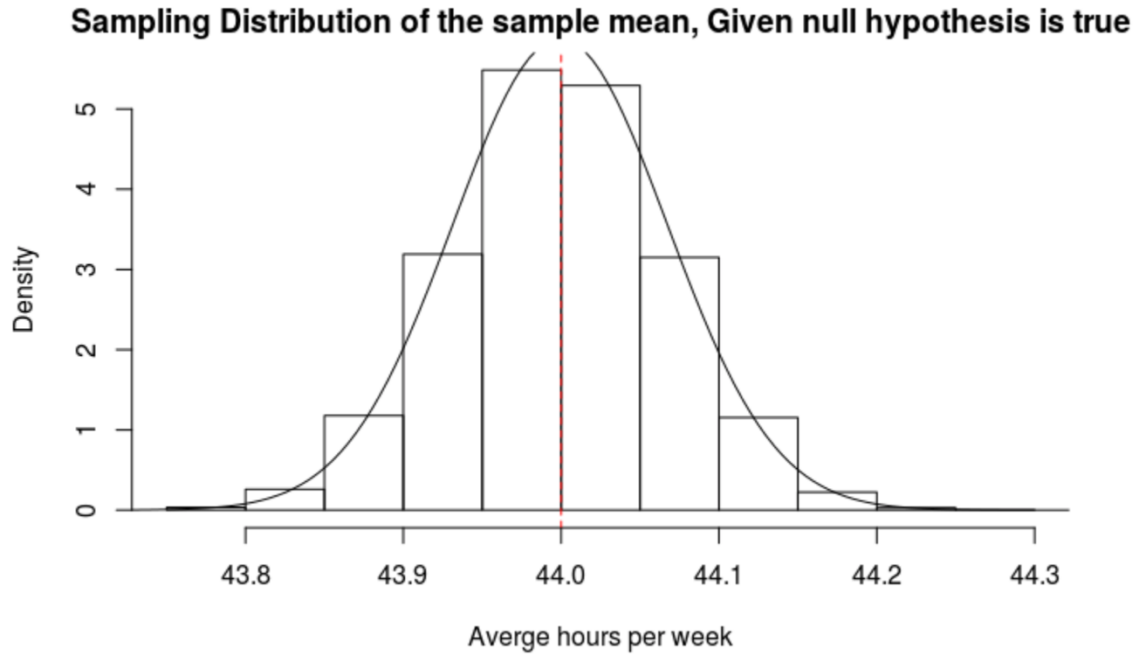
- Bootstrap Methods

Figure 14 Histogram of sampling distribution of the sample mean



Once we have the sampling distribution how do we find the p-value? We need to use the definition of the p-value that the p-value is the probability that we observed a test statistic as or more unusual than the one we observed, given the null hypothesis is true. If the null hypothesis is true, then the true mean hours worked per week us 44 hours. We need to shift the distribution so that it is true.

Figure 15 Histogram of the sample mean given that the null hypothesis is true



The lower bond and upper bond are not shown in Figure 15, because they are out of the ranges of x-axis (43.8 to 44.3)

- Result

I did the bootstrap simulations when the simulation times equal to 10000 and put detailed R codes for the bootstrap p-values and the bootstrap confidence interval of a knitted R markdown file in Appendix.

```
[1] 0
[1] 0
[1] 40.30223 40.57269
    2.5%    97.5%
40.30153 40.56724
```

- Bootstrap p-value: p-value =0
- Bootstrap CI
  The lower bond: 40.3
  The upper bond: 40.57

- Compare two methods

We got the same result from two methods. The bootstrap p value is less than 0.05, suggesting the result as the traditional method that the true population mean for hours worked per week does not equal to 44 hours at the α =0.05 level. With 95% confidence, the bootstrap confidence interval is between 40.30 hours and 40.57 hours, which suggests that the true mean hours worked per week is less than 44 hours.

2. One proportion z test

- Traditional Statistical Tools

According to the Median Wage report in United States, the proportion of people who earn more than 50,000 USD a year is around 31.5%[7]. The question of interest is about a categorical variable of interest with two categories. We are interested the true population proportion of white people who earn more than 50,000 USD a year. The statistical test is one sample test of proportion.

- Conditions for use and checks for conditions

For normal approximation, each sample size should be larger than 10. For this chosen dataset, the sample population of white people is 27816, and the sample population of white who earn more than 50,000 USD a year is 7117, therefore, the proportion from the sample population is 25.6%.

- Parameter
  The population parameter we want to make inference to is the population proportion of white people who earn more than 50,000 USD a year.
- Hypotheses
  - Two-sided
  - The null hypothesis, H_0: p_0 =0.315
    The true proportion of white people who earn more than 50,000 USD a year is 31.5%
  - The alternative hypothesis, H_1: p_0 ≠0.315
    The true proportion of white people who earn more than 50,000 USD a year is not 31.5%.
- Sample statistic

The sample statistic is p_1=0.256

- Test statistic

  For normal approximation, we use p_0 and $p_1$to find the test statistic z:

  $z = \frac{p_1 - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.256-0.315}{\sqrt{\frac{0.315(1-0.315)}{27816}}}$=-21.23

- Distribution of the test statistic

$$z = \frac{p_1 - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

  - Result

  I did two-sided one proportion z test with 95% confidence interval and put detailed R codes for t-test of a knitted R markdown file in Appendix

```
        Exact binomial test

data:  n2 and n1
number of successes = 7117, number of trials = 27816, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.315
95 percent confidence interval:
 0.2507399 0.2610310
sample estimates:
probability of success
             0.2558599

[1] -21.23384
```

- P-value

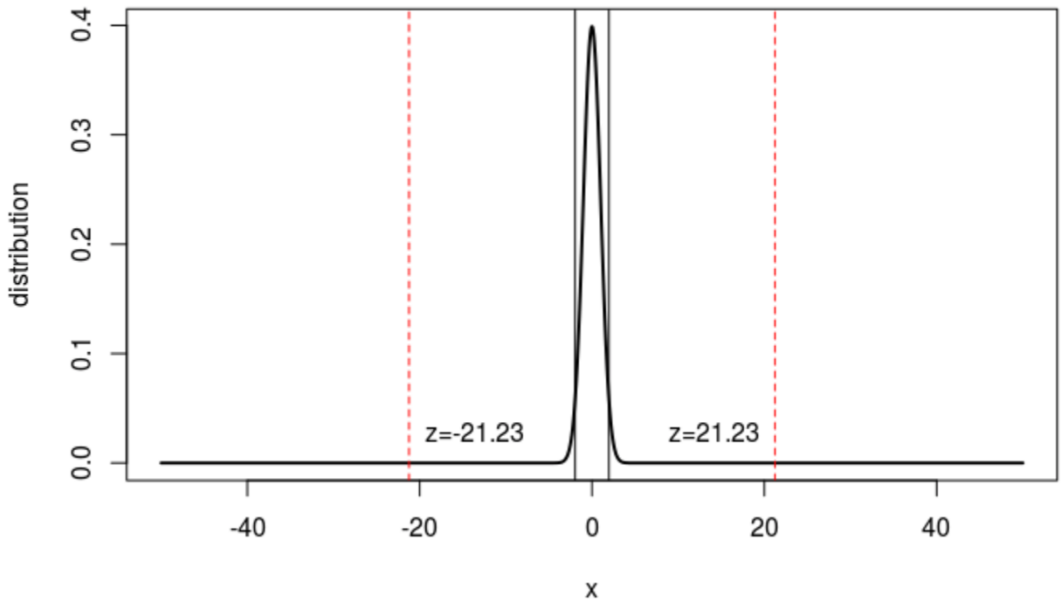
P-value <- $2.2*10^{-16} \approx 0$

- Confidence interval

The lower bound is 0.25

The upper bound is 0.261

Figure 16 Histogram of the sampling distribution



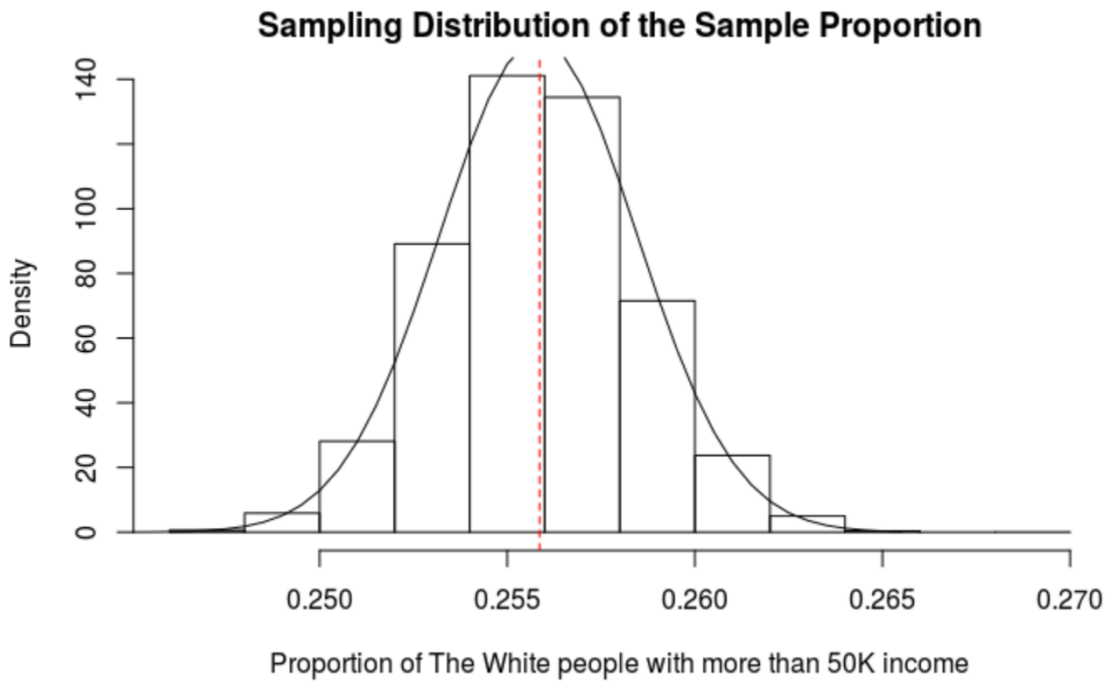
The curve represents the standard normal distribution when the mean equals to 0 and the standard deviation equals to 1. The two vertical red dashed lines represent the two test statistic values. The two black vertical lines represent the critical z values associated at the α =0.05 level

- Interpretation

Use the exact binomial methods for a one-sample test of proportion, there is a strong evidence (p-value $\approx$ 0) to reject the null hypothesis that the true proportion of white people who earn more than 50,000 USD a year is 31.5% at the α =0.05 level. With 95 % confidence, the true population proportion of white people who earn more than 50,000 USD is between 25% and 26.1%, which suggests that the true population proportion of white people who earn more than 50,000 USD is less than 31.5%

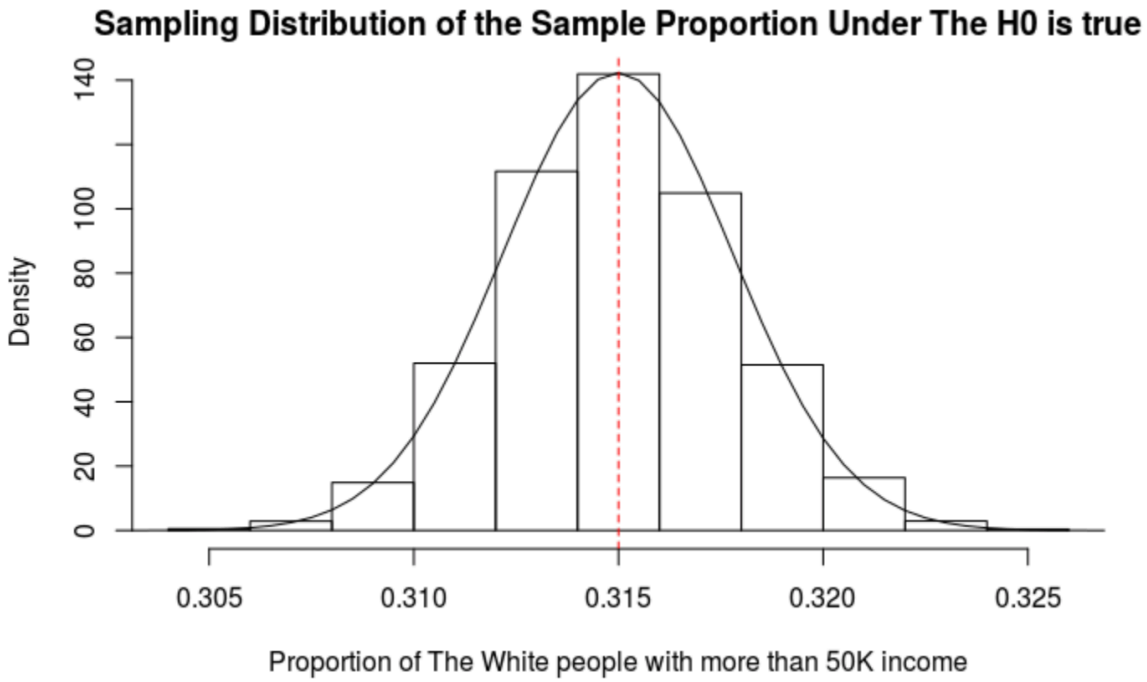
- Bootstrap Methods

Figure 17 Histogram of sampling distribution of the sample proportion



Once we have the sampling distribution how do we find the p-value? We need to use the definition of the p-value that the p-value is the probability that we observed a test statistic as or more unusual than the one we observed, given the null hypothesis is true. If the null hypothesis is true, then the true proportion of white people who earn more than 50,000 USD a year is 31.5%. We need to shift the distribution so that it is true.

Figure 18 Histogram of the sampling distribution of the sample proportion, given the null hypothesis is true



I did the bootstrap simulations when the simulation times equal to 10000 and put detailed R codes for the bootstrap p-values and the bootstrap confidence interval of a knitted R markdown file in Appendix.

```
[1] 0
[1] 0.00262252
[1] 0.2506149 0.2611050
     2.5%     97.5%
0.2507550 0.2610728
```

- Bootstrap p-value
  p-value=0
- Bootstrap Confidence interval
  Use this sampling distribution to find the 5th and the 95th percentiles
  The lower bond= 0.25
  The upper bond =0.261
- Compare two methods

We got the same result from two methods. The bootstrap p value is less than 0.05, suggesting the result as the traditional method that the true proportion of white people who earn more than 50,000 USD a year is different from 31.5% at the α =0.05 level. With 95 % confidence, the bootstrap confidence interval is between 25% and 26.1%, which suggests that the true population proportion of white people who earn more than 50,000 USD is less than 31.5%.

Two sample t-test for difference in means

- Traditional Statistical Tools

The question of interest is to compare the difference in means of capital loss for private workclass and local government work class, and to make inference about the true population difference in means. The two sample t-test is used for this statistical test.

- Conditions for use and checks for conditions
  - The sample is representative of the population- this chosen dataset is extracted from the 1994 to 1995 current population surveys conducted by the US Census Bureau.
  - Questions of interest has to do with the difference of means between two populations—yes, private workclass and local government workclass are independent, and the difference in the mean capital loss of each population
  - Two independent samples from two populations--yes
  - We want to make inference about the population difference in means using the sample difference in means—yes
  - The population data must be normally distributed-need to check
    1. To check this condition, look at the QQ plot of the sample data
    2. Particularly important if the size is less than 30—in this case, sample sizes are both larger than 30.

Figure 19

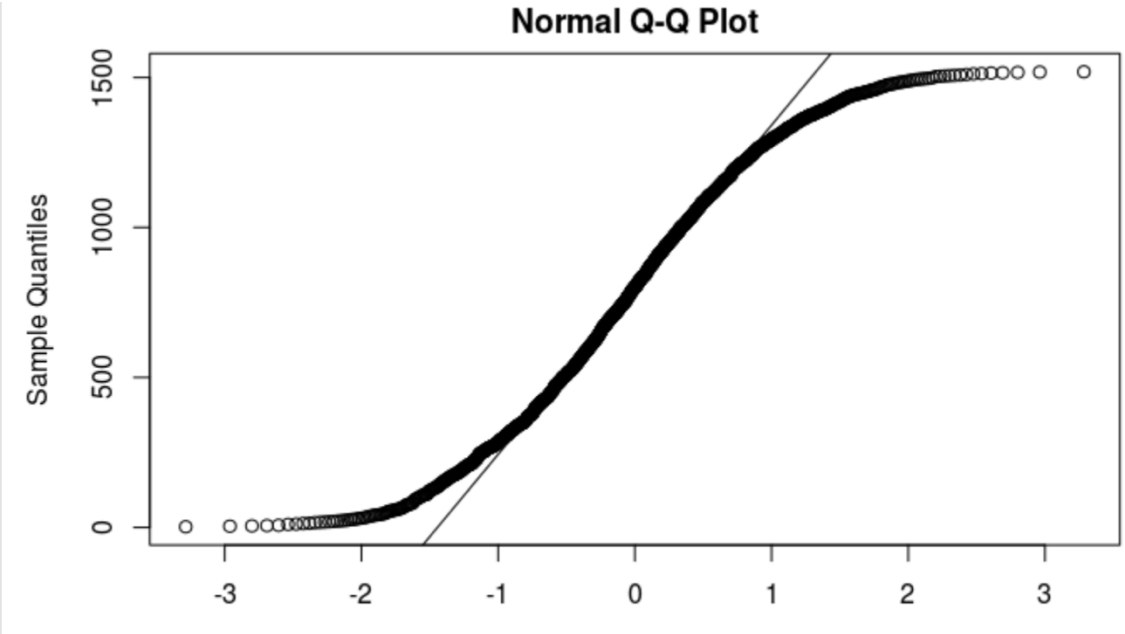
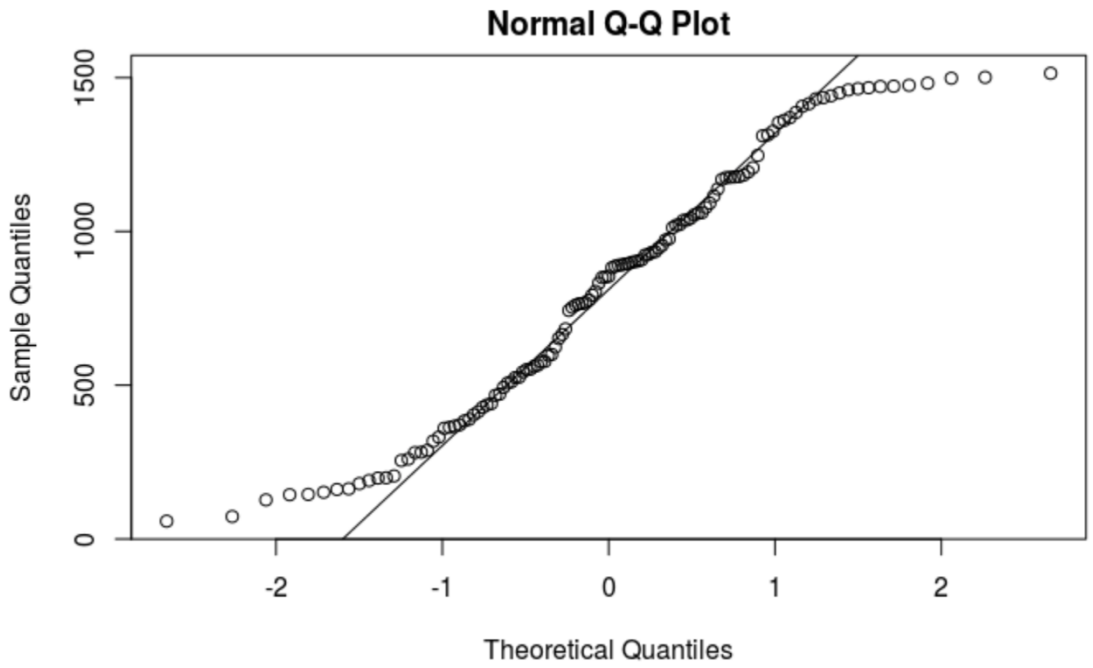


Figure 20



These plots indicate that all data points fall along the y=x line with the most points concentrated in the center of the line and fewer points towards the ends. There are more points concentrated in the middle and less towards the edges, and this is not far from the straight line. The normality assumption is met, and we should trust the reliability of p-value and the confidence intervals.

- Parameter
  - We are interested in the true population mean difference of capital loss for private business and non-private business: μ_1-μ_2
- Hypothesis
- The null hypothesis, H_0: μ_1-μ_2=0
  The true population mean of capital loss for private workclass is equal to the true population mean of capital loss for local government workclass
- The alternative hypothesis, H_1: μ_1-μ_2 ≠0
  The true population mean of capital loss for private workclass is different from the true population mean of capital loss for local government workclass
- Sample statistic
  $\bar{x}_1$-$\bar{x}_2$
- Test statistic

$$t_{\min\left(n_{p-1,n_{n-1,}}\right)}=\frac{(\bar{x}_p-\bar{x}_n)-(\mu_p-\mu_n)}{\sqrt{\frac{s_p^2}{n_p}+\frac{s_n^2}{n_n}}}$$

- Distribution of the test statistic

$$t_{\min\left(n_{p-1,n_{n-1,}}\right)}=\frac{(\bar{x}_p-\bar{x}_n)-(\mu_p-\mu_n)}{\sqrt{\frac{s_p^2}{n_p}+\frac{s_n^2}{n_n}}}\sim t_{\min\left(n_{p-1,n_{n-1,}}\right)}$$

- Result

  I did two-sided based on the alternative hypothesis with 95% confidence interval and put detailed R codes for t-test of a knitted R markdown file in Appendix.

```
        Two Sample t-test

data:  data1 and data2
t = -0.67934, df = 1107, p-value = 0.4971
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -107.38269   52.14826
sample estimates:
mean of x mean of y
 789.1151  816.7323

[1] -107.38269   52.14826
attr(,"conf.level")
[1] 0.95
[1] 0.4970636
```

  - P-value

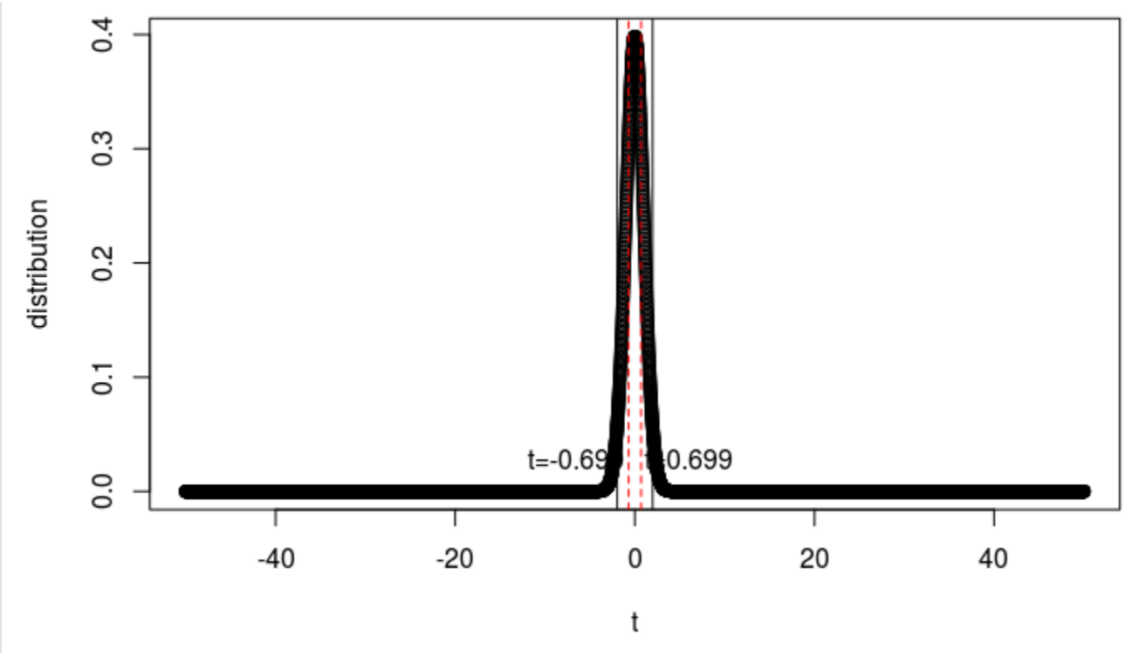
  p-value=0.4971

  - Confidence interval

  The lower bond=-107.38

  The upper bond=52.15

Figure 21 Histogram of the sampling distribution



The curve represents the t distribution with the same degree of freedom of the original dataset. The two vertical red dashed lines represent the two test statistic values. The two black

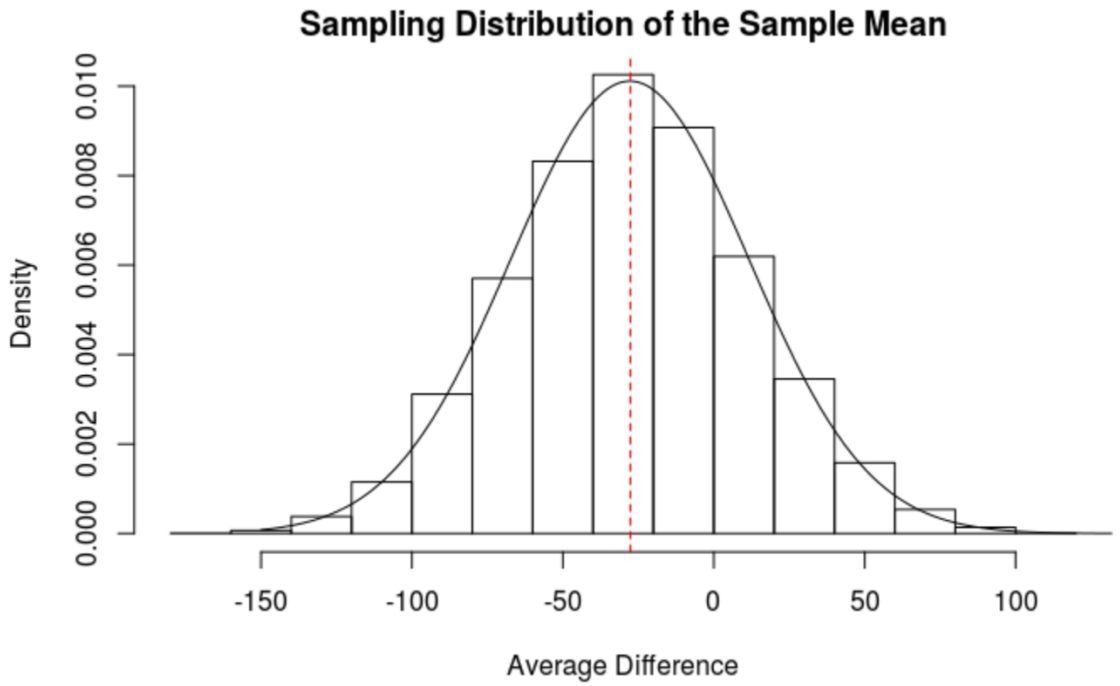
vertical lines represent the critical t values associated at the α =0.05 level and (n-1) degree of freedom.

- Interpretation

There is an evidence to suggest that we fail reject the null hypothesis since P-value from the hypothesis test is larger than 0.05 (p-value $= 0.4971$),which indicates that the true population mean of capital loss for private workclass is equal to the true population mean of capital loss for local government workclass at the α =0.05 level. With 95% confidence, the true population mean of capital loss is between -107.38 USD and 52.15 USD, which include the zero.

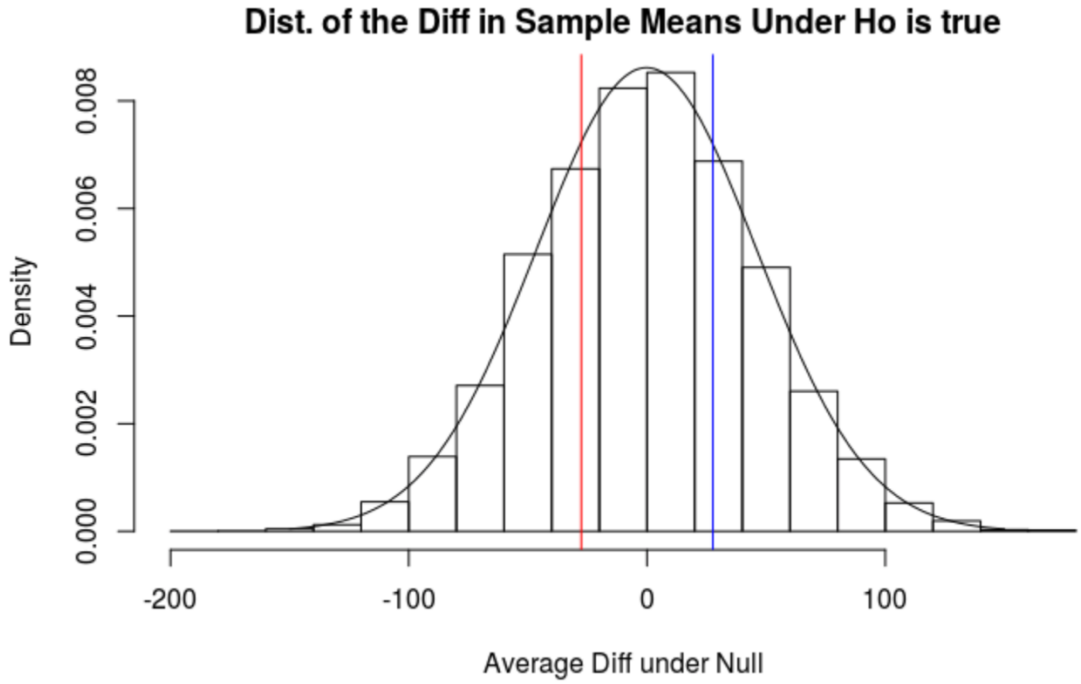
- Bootstrap Methods

Figure 22 Histogram of sampling distribution of the sample mean



Once we have the sampling distribution how do we find the p-value? We need to use the definition of the p-value that the p-value is the probability that we observed a test statistic as or more unusual than the one we observed, given the null hypothesis is true. If the null hypothesis is true, then the true mean difference between the two group is zero. We need to shift the distribution so that it is true, or we can create a randomization distribution by using the transform function in R. In general, we plotted the randomization distribution by simulating many samples assuming that the null hypothesis is true. Under the null there is no relationship between two variables. Firstly, we can create many samples as a treatment group that is shuffled. Next, we can compute the difference in groups for each of the samples. Finally, we create the histogram of the randomized statistics in **Figure 23**.

Figure 23 Histogram of the distribution of the difference in sample means, given the null hypothesis is true



The vertical blue lines and the vertical red lines represent the true sample mean difference of private business and non-private business. $\bar{x}_1$-$\bar{x}_2$=27.6

- Bootstrap p-value
  P=0.5497
- Bootstrap confidence interval

Use this sampling distribution to find the 5th and the 95th percentiles

The lower bond= -105.09

The upper bond =50.27

- Compare two methods

We got the same result from two methods. the bootstrap p-value is larger than 0.05, suggesting that the true population mean of capital loss for private workclass is equal to the true population mean of capital loss for local government workclass at the α =0.05 level. With 95% confidence, the bootstrap confidence interval is between -105.09 USD and 50.27USD, which include the zero. These values suggest that which suggests that the true mean of capital loss for private workclass is same as the true population mean of capital loss for local government workclass.

3. Two sample z test for difference in proportions

- Traditional Statistical Tools

The question of interest is to compare the difference of higher income proportion between private work class and local government work class.

- Conditions for use and checks for conditions

  Samples needs to be representative of the population—yes

  Categorical response variables with two categories—yes

  Two independent samples from two populations—yes

  For both populations, sizes are larger than 10---yes

- Parameter

We are interested in the difference between the true population proportion of higher income in private workclass and the true population proportion of higher income in local government workclass: $p_1 - p_2$

- Hypotheses
- The null hypothesis, H_0: $p_1 - p_2$=0

  There is no difference between the true population proportion of higher income in private workclass and the true population proportion of higher income in local government workclass.
- The alternative hypothesis, H_1: $p_1 - p_2 \neq 0$

  There is a difference between the true population proportion of higher income in private workclass and the true population proportion of higher income in local government workclass.
- Sample statistic

  $p_1 - p_2$
- Test statistic

$$z = \frac{(p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

- Distribution of the test statistic

$$z = \frac{(p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

- Result

  I did two-sided based on the alternative hypothesis with 95% confidence interval and put detailed R codes for t-test of a knitted R markdown file in Appendix.

```
        2-sample test for equality of proportions with continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 63.219, df = 1, p-value = 1.85e-15
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.09664040 -0.05559814
sample estimates:
   prop 1    prop 2
0.2186729 0.2947922

[1] -7.363747
```

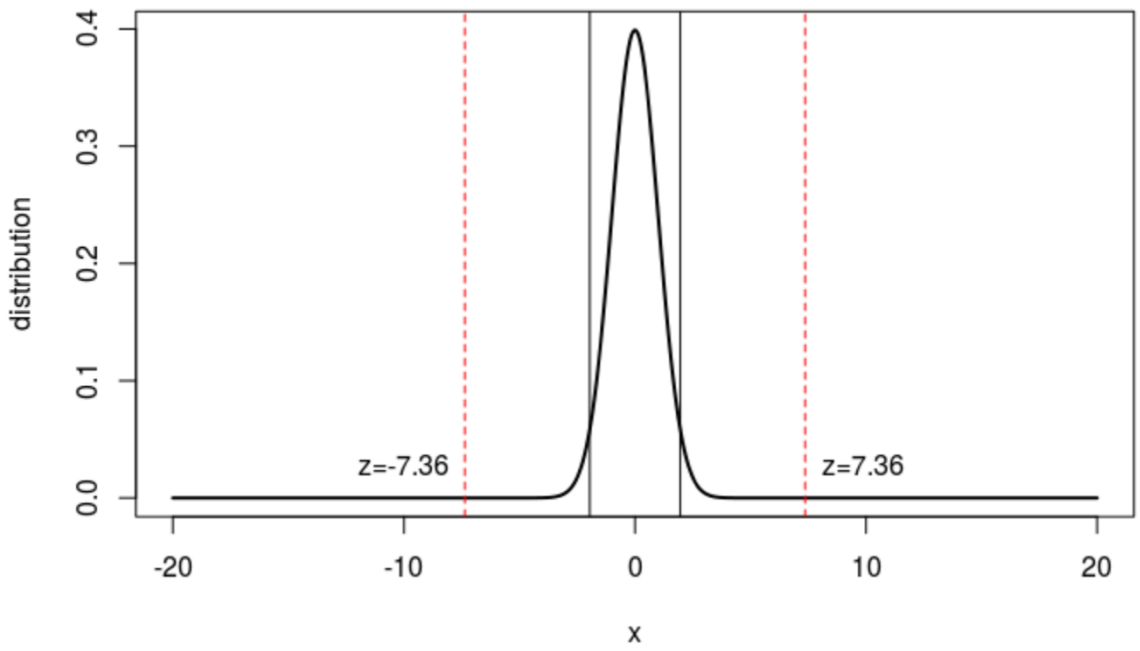
- P-value

  P-value <- $1.85*10^{-15} \approx 0$
- Confidence interval

  The lower bound is -0.0966

  The upper bound is -0.0556

Figure 24 Histogram of the sampling distribution



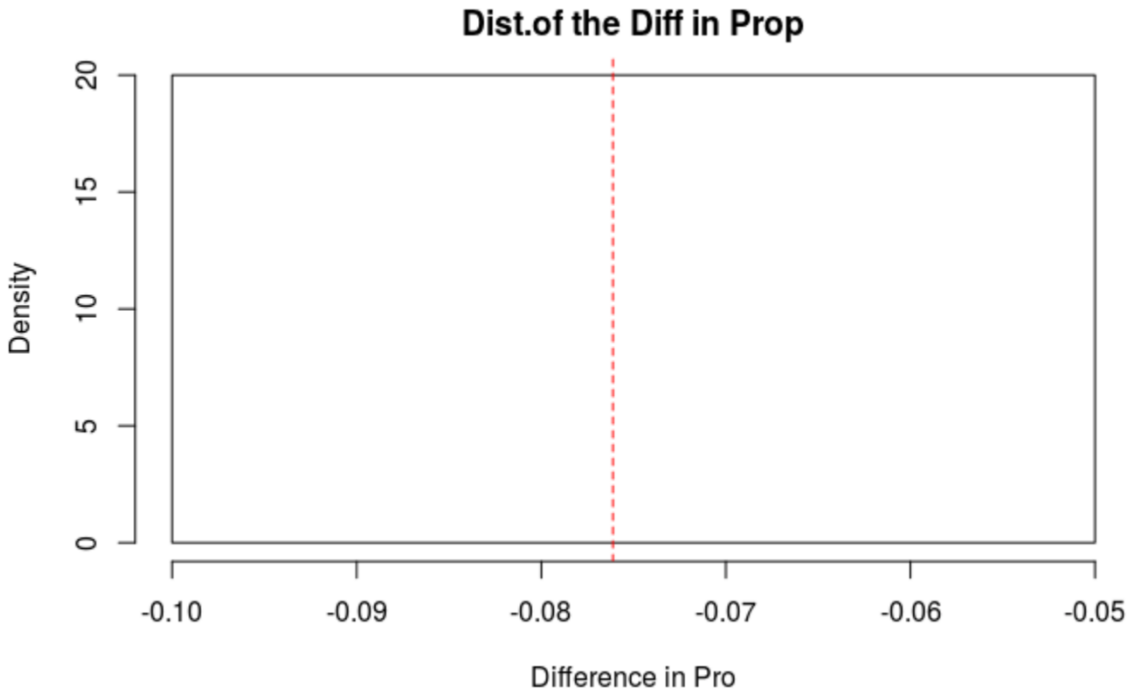
The curve represents the standard normal distribution when the mean equals to 0 and the standard deviation equals to 1. The two vertical red dashed lines represent the two test statistic values. The two black vertical lines represent the critical z values associated at the α =0.05 level.

- Interpretation

There is strong evidence to suggest that we reject the null hypothesis since P-value from the hypothesis test is smaller than 0.05 (p-value $\approx$ 0),which indicates that there is a difference between the true population proportion of higher income in private workclass and the true population proportion of higher income in local government workclass at the α =0.05 level. With 95% confidence, the difference of true population proportion of higher income is between negative 9.66% to negative 5.56%, indicating that the true population proportion of higher income in private workclass is less than the true population proportion of higher income in local government workclass.

- Bootstrap Methods

Figure 25 Histogram of sampling distribution of the difference in proportion



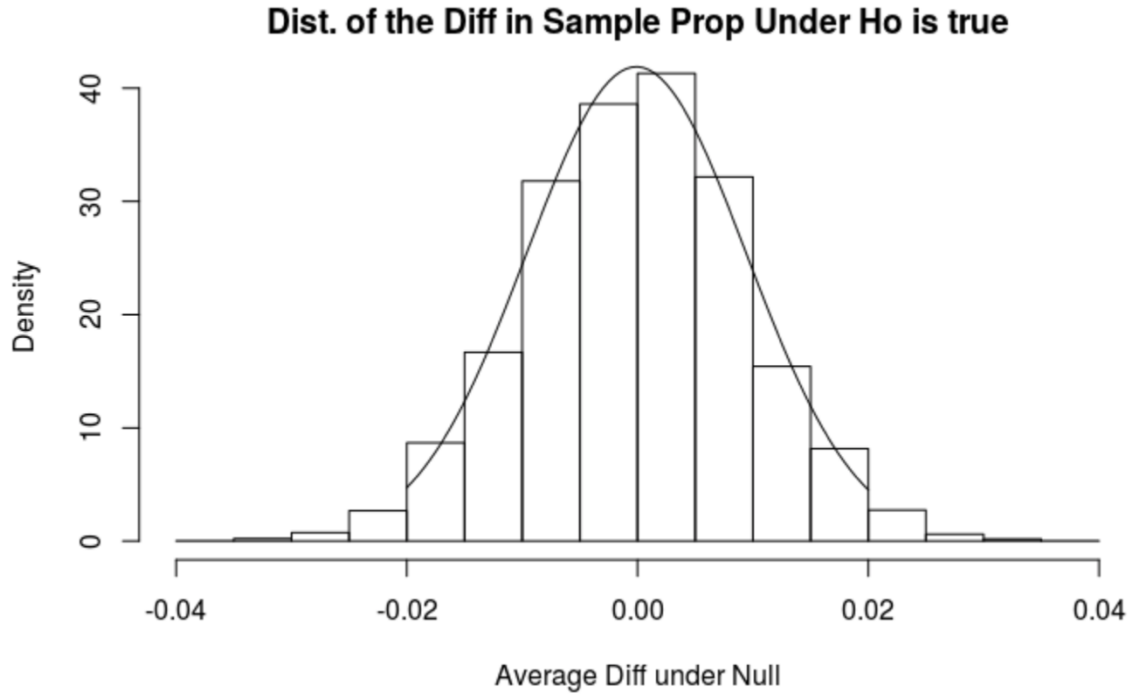
There is no distribution of the difference in proportion, since the output from 10000 simulations is the same number (-0.07612)

```
 [891] -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927
-0.07611927 -0.07611927 -0.07611927
 [901] -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927
-0.07611927 -0.07611927 -0.07611927
 [911] -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927
-0.07611927 -0.07611927 -0.07611927
 [921] -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927
-0.07611927 -0.07611927 -0.07611927
 [931] -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927
-0.07611927 -0.07611927 -0.07611927
 [941] -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927
-0.07611927 -0.07611927 -0.07611927
 [951] -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927
-0.07611927 -0.07611927 -0.07611927
 [961] -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927
-0.07611927 -0.07611927 -0.07611927
 [971] -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927
-0.07611927 -0.07611927 -0.07611927
 [981] -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927
-0.07611927 -0.07611927 -0.07611927
 [991] -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927 -0.07611927
-0.07611927 -0.07611927 -0.07611927
 [ reached getOption("max.print") -- omitted 9000 entries ]
      2.5%       97.5%
-0.07611927 -0.07611927
```

Once we have the sampling distribution how do we find the p-value? We need to use the definition of the p-value that the p-value is the probability that we observed a test statistic as or more unusual than the one we observed, given the null hypothesis is true. If the null hypothesis is true, then the true mean difference between the two group is zero. We need to shift the

distribution so that it is true, or we can create a randomization distribution by using the transform function in R. In general, we plotted the randomization distribution by simulating many samples assuming that the null hypothesis is true. Under the null there is no relationship between two variables. Firstly, we can create many samples as a treatment group that is shuffled. Next, we can compute the difference in groups for each of the samples. Finally, we create the histogram of the randomized statistics in **Figure 26**.

Figure 26 Histogram of the distribution of the difference in sample proportion, given the null hypothesis is true



- Bootstrap p-value=0
- Bootstrap CI
  The lower bond =the upper bond =-0.07612

- Compare two methods

We got the same result from two methods. The bootstrap p-value is 0, indicating that there is a difference between the true population proportion of higher income in private workclass and the true population proportion of higher income in local government workclass at the α =0.05 level. The bootstrap confidence interval gave one single value -7.76%, which is close to the mean of traditional confidence interval. As you can see, after the randomization, we can have more accurate result that the true population proportion of higher income in private workclass is less than the true population proportion of higher income in local government workclass.

- Chi-square goodness of fit
- Traditional statistical tools

We are interested that when people have different occupations, are proportions of them different from each other?

- Conditions for use and checks for conditions
- Single categorical variable with more than two variables—yes
- The expected count of each count is at least five---yes

From **Figure 27**, you can see that there are fifteen categorical variables in occupation column. The least count of Armed Forces is 9 which is larger than 5.

Figure 27

```
               ?      Adm-clerical      Armed-Forces      Craft-repair   Exec-managerial   Farming-fishing
            1843              3770                 9              4099              4066               994
Handlers-cleaners Machine-op-inspct     Other-service   Priv-house-serv    Prof-specialty   Protective-serv
            1370              2002              3295               149              4140               649
           Sales      Tech-support  Transport-moving
            3650               928              1597

               ?      Adm-clerical      Armed-Forces      Craft-repair   Exec-managerial   Farming-fishing
    0.0566014557      0.1157826848      0.0002764043      0.1258867971      0.1248733147      0.0305273180
Handlers-cleaners Machine-op-inspct     Other-service   Priv-house-serv    Prof-specialty   Protective-serv
    0.0420748749      0.0614845981      0.1011946808      0.0045760265      0.1271459722      0.0199318203
           Sales      Tech-support  Transport-moving
    0.1120972943      0.0285003532      0.0490464052
[1] 32561
[1] 0.06666667
[1] 2170.733
[1] 14884.19
[1] 0
```

- Parameter of interest

  We are interested in the true proportion of each categorical variable : p1-p15

- Hypotheses
- The null hypothesis, H_0: $p_1 = p_2 = p_3 = \cdots p_{15} = 0.067$

  The proportion of each occupation is the same and is equal to 0.067
- The alternative hypothesis, H_1: $Some\ p_i \neq 0.067$

  At least one of the proportion is not equal to 0.067

This test does not tell us which proportion is not equal to 0.067. All the test tells us about that at least one of the proportions is significantly different than 0.2. There are 32561 occupations in this chosen sample. If each of occupation type has the same frequency, then each occupation would have a count of 32561*0.067=2170. In other word, under the null hypothesis, the expected count is 2170.

- Sample statistic

In this example, we have 15 sample statistics: p1-p15

- Test statistic and Distribution

We need a statistic that compare our sample counts to the expected count. The statistic we will use is the Chi-square statistic where k is the number of cells in the table. So far, all of our statistics have had a normal or t reference distribution. Both of which are roughly symmetric and bell-shaped. However, the distribution of a chi-square statistic is skewed

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \sim \chi^2_{k-1}$$

- P-value=0
- There is no confidence interval in this test

Figure 28 The histogram of proportion sampling distribution

- Interpretation

  There is a strong evidence that we can reject the null hypothesis since the p value is less than 0.05, indicating that at least one of the proportions is not equal to 0.067 at the α=0.05 level.

Discussion

- **Summary of findings**

The data exploratory analysis enables people to understand more about the domain knowledge. As a result, gaining further knowledge about the area of expertise helps data analysis to discover interesting patterns with a reasonable explanation. For continuous variables, to some extent, age, hours worked per week, and capital support the discrimination of different income groups. It can be seen that people who aged from 35 to 60 are more likely to have a higher income than the average. People who worked above 40 hours per week have higher chances in higher income. For the higher income, people who have the zero capital gain have more likelihoods. For categorical variables, those who are males with higher degrees are more likely to have higher income. Because 70% of the work class of this chosen dataset is private workclass, as an entrepreneur, a male has more advantages than a female in business world. Therefore, sex does play an important role in determining income levels.

At the α =0.05 level, from the one sample t-testing, the true mean hours worked per week for the population is less than 44 hours. from the one proportion z test, the true population proportion of white people who earn more than 50,000 USD is less than 31.5%. from the two sample t-test for difference in means, the true population mean of capital loss for private workclass is equal to the true population mean of capital loss for local government workclass. From the two sample z test for difference in proportions, the true population proportion of higher income in private workclass is less than the true population proportion of higher income in local

government workclass. From the chi-square goodness of fit, at least one of the proportions is not equal to 0.067 for different occupations.

- **Implication of findings**

Combining with exploratory analytics and statistical analytics, we do have the comprehensive visons of the chosen dataset. Moreover, the income is the key categorical variable, and causal inferences between income and other variables can be drawn. The capital is not the causality for higher income, which is matched with the result from the two sample t-test for difference in means.

- **Limitations and Extensions**

As mention in the introduction, we discussed about the sampling bias and response bias in the design phase. However, those biases cannot be completely corrected after the completion of a study, thus we have to minimize their impact during the analysis phase. For this chosen dataset, The U.S. accounts for 90% in native_country attribute while each of other 40 countries account for less than 1%. Similarly, "White" is dominant in race feature with 85%. For work class , the private work class accounts for 70%. Those selected samples may have effects on the analysis.

Confounding bias is a systematic error in inference due to the influence of confounding variables. Confounding variables are extra variables that we do not account for in the study. For this chosen dataset, there are other occupations and work class that we did not account for, and those are the confounding variables.

Data cleaning is necessary. For this chosen dataset, we did not check the duplications or any missing values. There are unknown values with question marks in the dataset. Therefore, it is better to do more researches to figure the meanings out.

- **Further questions and steps**

One of the most effective methods that can be used to avoid sampling bias is simple random sampling, in which samples are chosen strictly by chance. We need to enlarge the sample sizes and share the survey via various methods such as email and website.

Several data pre-processing steps are considered and hopefully help improve the prediction performance in terms of accuracy, stability, speed and interpretation.

Besides minor data errors such as missing, duplicated, the main on-going concern will be about the representative of the data set and how one can improve it to achieve business goals.

Appendix

- References

[1] Uci. (2016, October 7). Adult Census Income. Retrieved from https://www.kaggle.com/uciml/adult-census-income.
[2] UCI Machine Learning Repository: Census Income Data Set. (2019). Retrieved 7 December 2019, from https://archive.ics.uci.edu/ml/datasets/census+income
[3] Why Are They Asking That? What Everyone Needs to Know About 2020 Census Questions – Population Reference Bureau. (2019). Retrieved 7 December 2019, from https://www.prb.org/why-are-they-asking-that-what-everyone-needs-to-know-about-2020-census-questions/
[4]Tingle, K. (2019). Why Should We Care About an Accurate Census? - CPPP Blog | Center for Public Policy Priorities. Retrieved 7 December 2019, from http://bettertexasblog.org/2018/06/why-should-we-care-about-an-accurate-census/
[5] Bureau, U. (2019). Why We Ask About... Income. Retrieved 6 December 2019, from https://www.census.gov/acs/www/about/why-we-ask-each-question/income/
[6] one sample t-test
Ward, M. (2019). A brief history of the 8-hour workday, which changed how Americans work. Retrieved 7 December 2019, from https://www.cnbc.com/2017/05/03/how-the-8-hour-workday-changed-how-americans-work.html
[7] one sample proportion test
(2019). Retrieved 7 December 2019, from https://www.quora.com/How-many-people-in-the-us-earn-less-than-50k-a-year

- The csv dataset :[1] Uci. (2016, October 7). Adult Census Income. Retrieved from https://www.kaggle.com/uciml/adult-census-income.
  The csv name uploaded on R studio : adult.csv

- R knitted codes