IE6200 Final Project
Suicide Data in US between 1985 and 2015

Mei-Chun Hung UNID001084539

2019/12/06

Table of Contents

1. Introduction

According to American Foundation for Suicide Prevention (2017), suicide to the 10th leading cause of death in the US. There are more than 1.4 million suicide attempts per year in the United States. Unfortunately, there are 129 suicide per day on average. The suicide is happening every day. The purpose of this study is to understand the data of suicide in the United States between 1985 and 2015. The study is based on the dataset, Suicide Rate Overview 1985 to 2016 (Rusty 2018). There are twelve columns in the dataset. The columns are 'country', 'year', 'sex', 'age', 'suicides numbers', 'population', 'suicides number per 100k people', 'country-year', 'HDI for year', 'gdp' , 'gdp per capita', and 'generation'. This study is focusing on the data of United States. There are two variables in gender, 'female' and 'male', and six variables in age, '5-14 years', '15-24 years', '25-34 years', '35-54 years', '55-74 years', and '75+ years'. The data includes the number of suicides in each gender and each age ranges from 1985 and 2015. Therefore, there are 372 samples in this study. The main goal of this study is to understand the relationship between the numbers of suicide between gender, and numbers of suicide between age ranges between 1985 and 2015. The visualization of the datasets is in the exploratory analysis. Followed by the statistical analysis. There are five sample test is using in this study, one sample t-test with both traditional statistical tools and bootstrap methods, one sample test of proportion with both traditional statistical tools and bootstrap methods, two sample t-test for difference in means with both traditional statistical tools and bootstrap methods, two sample test for difference in proportions with traditional statistical tools and bootstrap methods, and chi-square goodness of fit and chi-square test of association. The detailed of tests will be discussed in the later sections. One sample t-test shows that the average number of suicides is different from 47085 which is mentioned by American Foundation for Suicide Prevention. One sample test of proportion shows that the proportion of age group "25-34 years old" is greater than $\frac{1}{6}$. Two sample test in difference in mean shows that there is difference between the average number of suicides per 100k people in female and the average number of suicides per 100k people in male. Two sample test in difference in proportion shows that there is difference between the suicides proportion of female and the suicides proportion of male. The chi-square goodness of fit test shows that there is at least one of the proportions in age group is not $\frac{1}{6}$. And the chi-square association between two categorical variables shows that the number of suicides is associate with gender. The results are showing that there are some relationship about number of suicides between gender and age group. The detailed will be discussed in the later sections.

2. Exploratory Analysis

This section is focusing on the visualization of the data. The number of suicides, year, age group, and gender will be represented. The relationship will be shown in the following graphs.

PART I. Boxplot

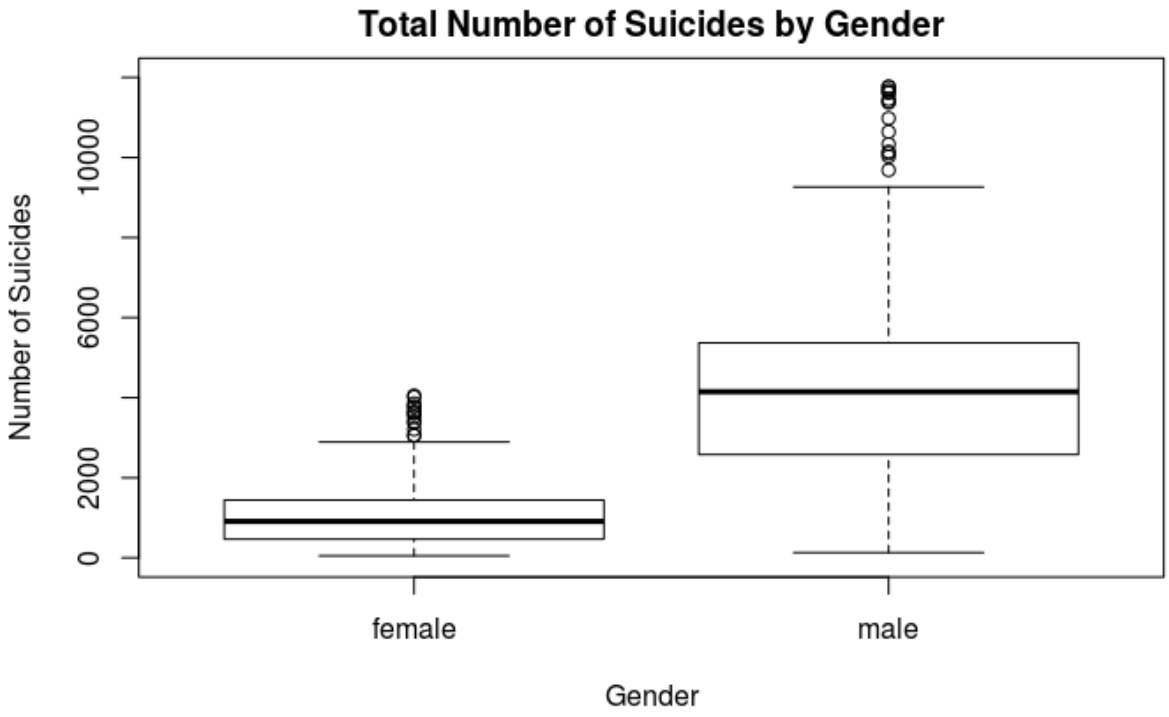

Figure 2.1 Total Number of Suicides by Gender

The boxplot above shows the summaries of female and male. Comparing to the information from female, mean of male is higher than mean of female. Male has wider range of the number of suicides. Both female and male have some outliers of the data. The maximum number of suicide of female is about 5000, and the maximum number of suicide of male is about 12000. From this boxplot, we can conclude that male has higher number of suicides in this dataset.

PART II Line Graph

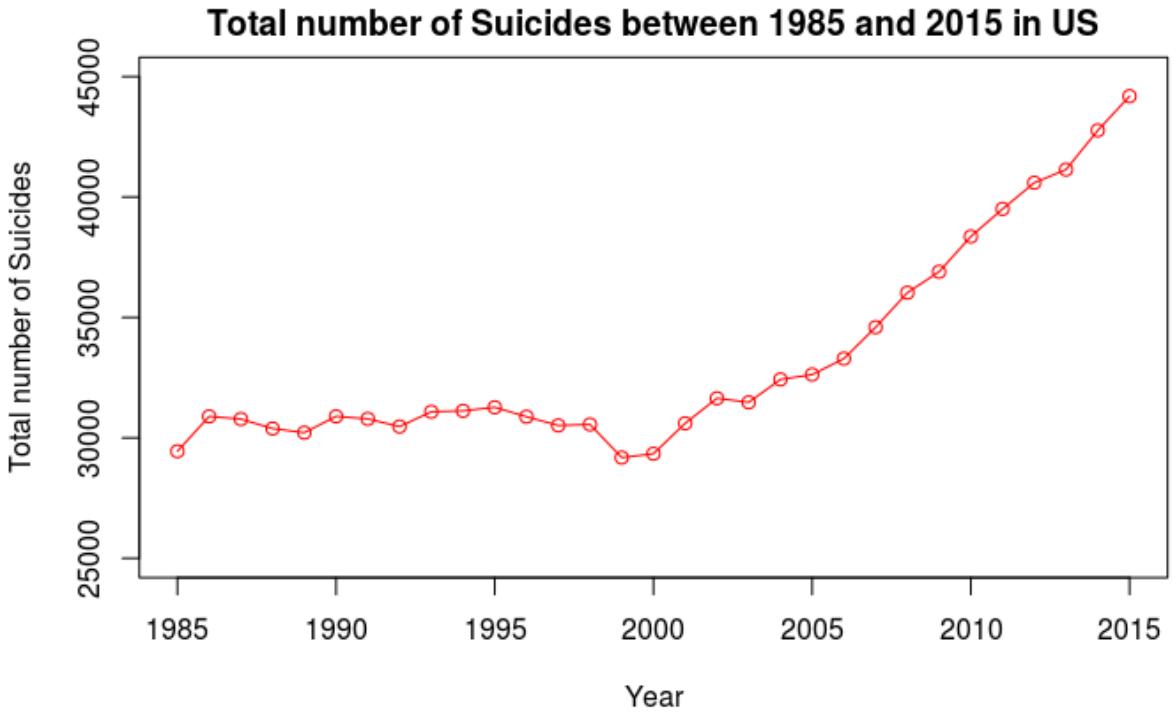

Figure 2.2 Total number of suicides between 1985 and 2015 in US

The figure 2.2 above clearly shows that the total number of suicides is pretty steady from 1985 to 2000. Starting from 2000, the total number of suicides increase every year. The total number of suicides increase from 30000 to 450000 in this 31 year.

PART III Dot Graph

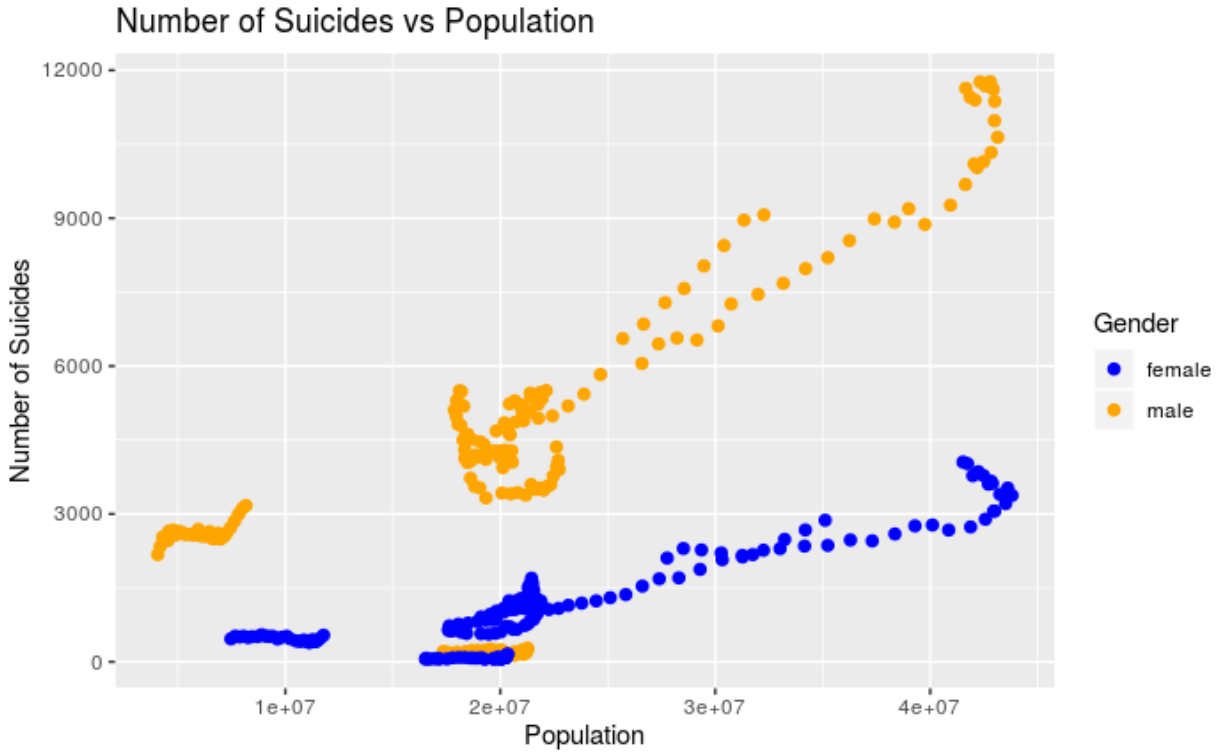

Figure 2.3 Number of Suicides vs Population by Gender

The dot plot above shows the relationship between number of suicides and population based on gender. The yellow represents male and the blue represents female on the graph. The dots are distributed similarly that the populations are focusing on the 20000000 range. The number of suicides of female is lower than the number of suicides of male when they are in the same population range. The number of suicides increase as the population increase for both genders.

PART IV Histogram in three categories

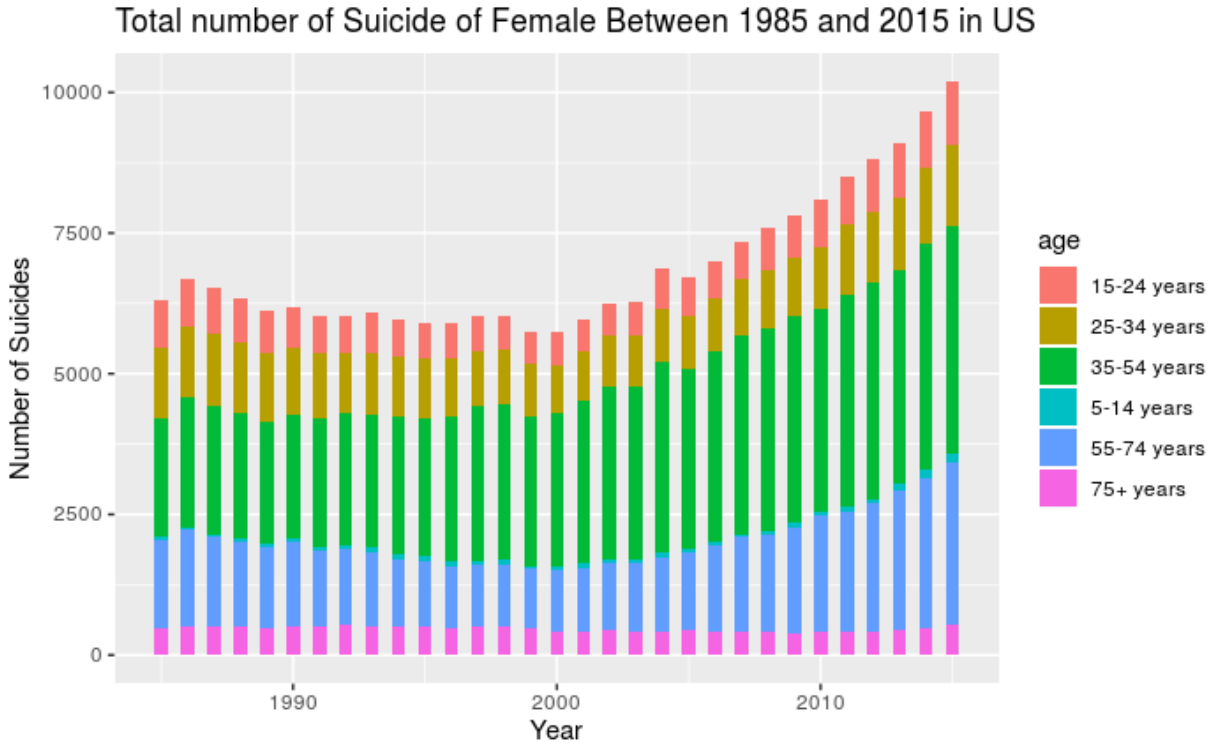

Figure 2.4 Total number of Suicide of Female Between 1985 and 2015 in US

The figure above shows that the number of suicides in each year for each age group of female. The number of suicides increases starting from 2000 which match the result for line graph. The age group “35-54 years old” is always having the largest amount of suicides in 1985 to 2015, and it seems increasing every year. The youngest and oldest age group do not change the overall amount of suicides in 1985 to 2015. 2015 is the first year reaching 10000 suicides for female.

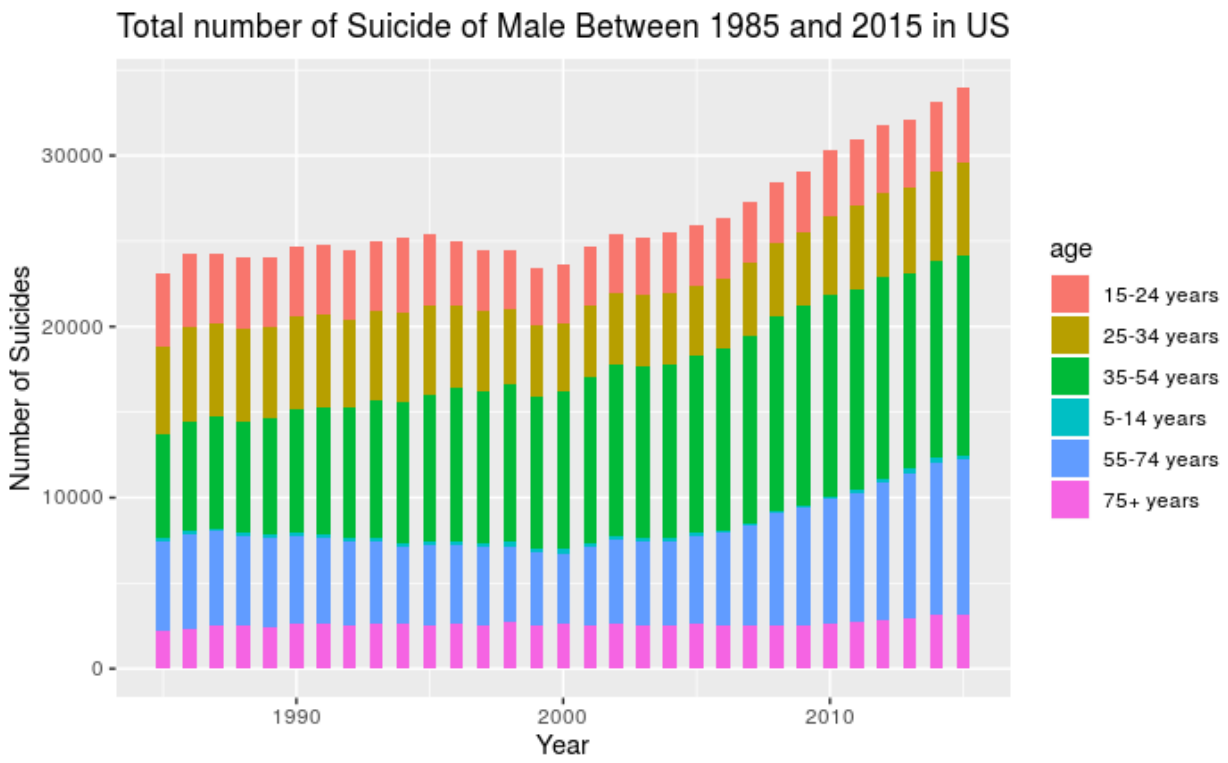

Figure 2.5 Total number of Suicide of Male Between 1985 and 2015 in US

The figure above shows that the number of suicides in each year for each age group of male. There are more than 20000 suicides starting from 1985 for male. The age group “35-54 years old” and “55-74 years old” male are the main age group of the total number of suicides. The total number reached 30000 first time in 2010, and still keep increasing every year. The youngest and oldest age group do not increase the total number overall in this 31 years.

PART V Histogram

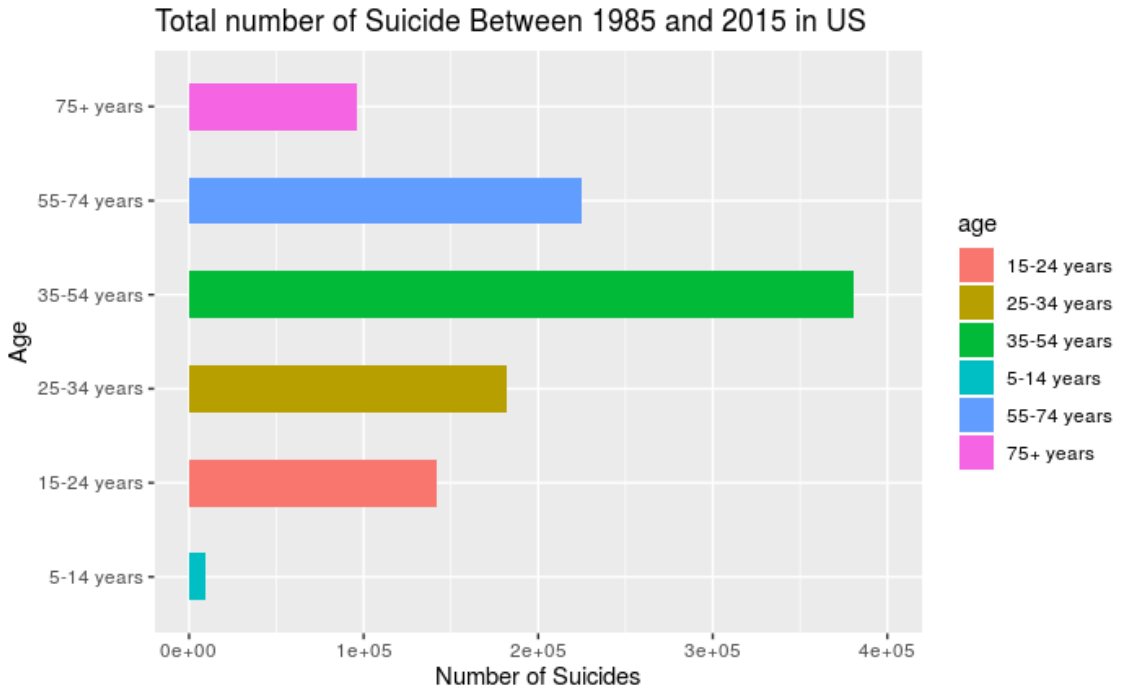

Figure 2.6 Total number of Suicide Between 1985 and 2015 in US

The figure above shows the total number of suicides for each age range in 1985 to 2015. The highest age group is “35-54 years old” which is same as the results of figure 2.4 and figure 2.5 above. The total number of suicides is greater than 350000 and less than 400000 for “35-34 years of” from 1985 and 2015 in US.

3. Statistical Analysis

This section is going to discuss the five tests using in the study, and the results found according to the dataset.

3.1 One Sample t-Test

According to the American Foundation for Suicide Prevention, there are 47085 suicides per year in United States. With the one sample t-Test, the sum of number of suicides of each year from 1985 to 2015 is calculated in the table. There are 31 samples in the table, each sample is including the number of suicides for each gender and each age group of the year.

Question: Is the average number of suicides per year is 47085 between 1985 and 2015?

3.1.1 Conditions for use of One Sample t-Test

The number of suicides is counted each individual by the World Health Organization. The sample is representative of the population because the amount of data is large. There is only one quantitative variable of interest, number of suicides. The purpose is to make inference about the population mean using the sample mean. The population variance is unknown, so we estimate it using sample data. The sample is from a single population. The following QQ plot will show the population data is normally distributed.

3.1.2 QQ Plot

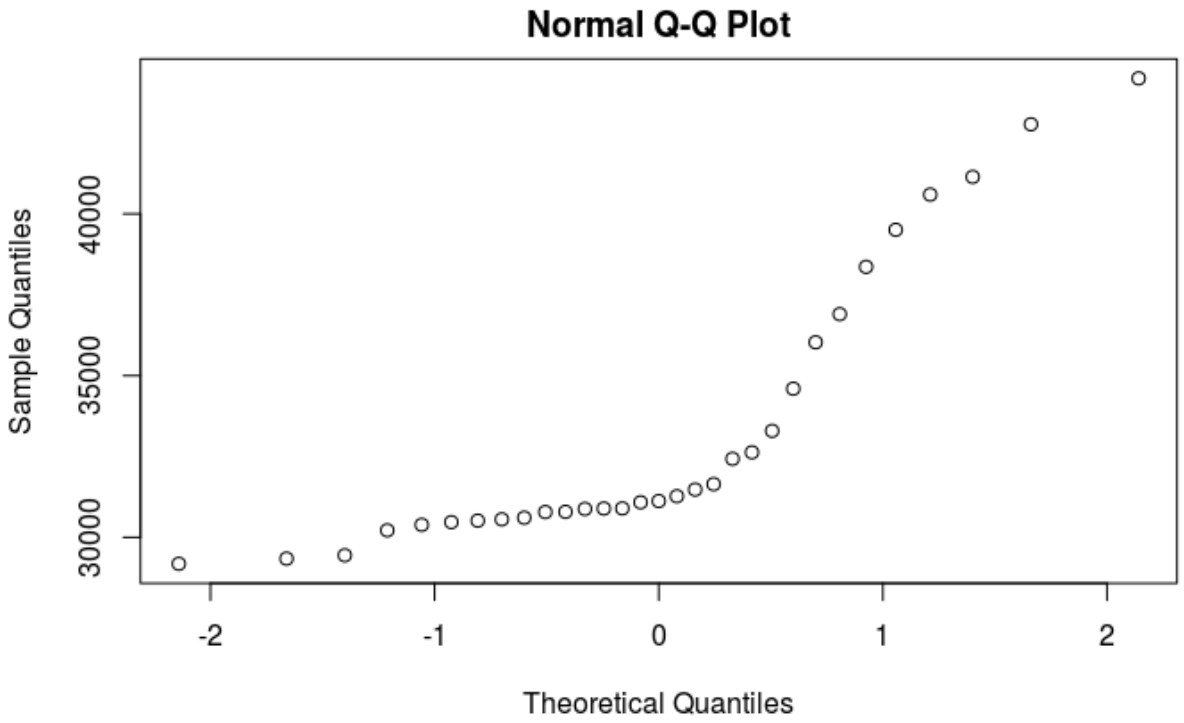

Figure 1 Normal QQ Plot

In the figure 1, there are 31 sample points on the QQ Plot which the shape of the plot is not perfect x=y line. However, it is a pretty good plot since there are only 31 sample points. There are more than two-third of the sample points in the -1 and 1 theoretical quantiles, therefore we can conclude that it is normally distributed.

3.1.3 Parameter

The test is interested in the population parameter we want to make inference to is μ.

3.1.4 Hypotheses

- $H_0 : \mu = 47085$
  - The true mean number of suicides for each year from 1985 to 2015 is 47085
- $H_A : \mu \neq 47085$
  - The true mean number of suicides for each year from 1985 to 2015 is different than 47085

3.1.5 Sample Statistic

The sample statistic is the sample mean number of suicides $\overline{x}$ which is the mean of the number of suicides of 31 sample years.

3.1.6 Test Statistic

The test statistic is $t_{n-1} = \frac{\overline{x}-\mu_0}{\frac{s}{\sqrt{n}}}$, where $\overline{x}$ = 33355.26, $\mu_0 = 47085$, s = 4505.852, and t_{30} = -17.7535

3.1.7 Distribution of the Test Statistic

$$t_{n-1} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

The p-value is 1.837489e-17 which is really close to 0.

3.1.8 Confidence Interval

With the 95% confidence interval, the lower bound of the one sample t-test is 31775.86 and upper bound is 34934.66. This shows that with 95% confidence, the true population mean is between 31775.86 and 34934.66 suicides each year between 1985 and 2015 in the United States. The 47085 suicides is not included in the confidence interval in this case.

3.1.9 Comparison with R build-in t-test

From the R build-in t-test, t = -17.754, df = 30, p-value < 2.2e-16 which is really close to 0, and 95% confidence interval is between 31775.86 and 34934.66. The results are same as we calculated above.

3.1.10 Interpretation

There is very strong evidence (p-value ≈ 0) to suggest that the true mean number of suicides for each year between 1985 and 2015 is different than 47085. We reject the null hypothesis that the true mean number of suicides for each year between 1985 and 2015 is 47085 at the $\alpha = 0.05$ level. With 95% confidence, the true mean number of suicides is between 31775.86 and 34934.66 people which suggests that the true mean number of suicides is less than 47085 people.

3.1.11 Bootstrap Approach

The sampling distribution of the sample mean which was simulated 1000 times.

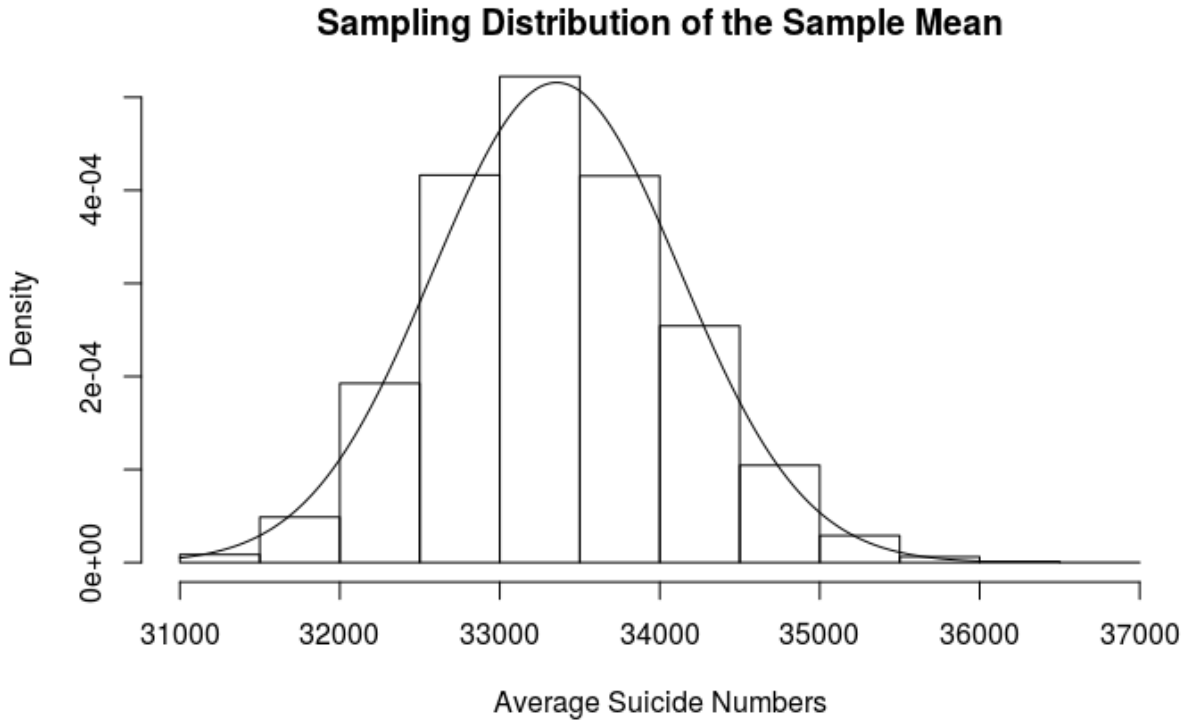

Figure 2 Sampling Distribution of the Sample Mean

The figure 2 shows the average number of suicides is about 33500 and is normally distributed about 1000 times simulations.

Shifting the sample so that the null hypothesis is true which also was simulated 1000 times

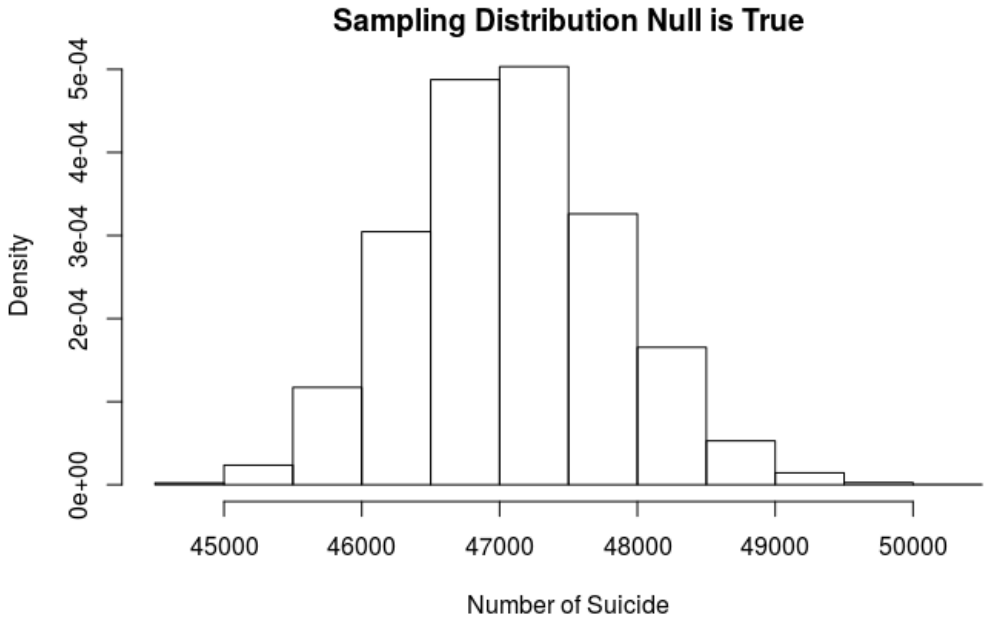

Figure 3 Sampling Distribution where Null Hypothesis is True

The figure 3 shows the average number of suicides is 47085 which is the null hypothesis, and it is normally distributed with 1000 simulations. The extreme value at upper end and extreme value at lower end is not shown because they are both out of the range of graph.

The p-value from bootstrap is 0 according to R.
The p-value from test statistics is 1.837489e-17 which is really close to 0.
The p-value from t-test is < 2.2e-16 which is really close to 0.
All three methods give similar results of p-value in this case.

The 95% confidence interval from bootstrap approach is between 31841.03 and 34869.49 people.
The 95% confidence interval from 5^{th} and 95^{th} quantile is between 31959.48 and 34897.26 people.
The 95% confidence interval from traditional t-test is between 31775.86 and 34934.66 people.
All three methods give similar results of 95% confidence interval in this case.

In both traditional method and bootstraps approach, we reject the null hypothesis which the true mean number of suicides for each year between 1985 and 2015 is 47085 at the $\alpha = 0.05$ level.

3.2 One Sample Test of Proportion

There are six age groups in the dataset, “5-14 years old”, “15-24 years old”, “25-34 years old”, “35-54 years old”, “55-74 years old”, “75+ years old”. With the one sample test of proportion, the sum of number of suicides of each age group is calculated in the table. The proportion of each age group is the sum of suicides in each age group divided by the total number of suicides between 1985 and 2015 in the United States.

Question: What proportion of number of suicides is the age group “25-34 years old”?

3.2.1 Conditions for use of One Sample Test of Proportion

No requirements for using Exact Binomial Test in R. The requirement for Normal approximation is met by $n\hat{p} = 182047 \geq 10$ and $n(1-\hat{p}) = 851966 \geq 10$

3.2.2 Parameter

The population parameter we want to make inference to is the population proportion or the age group “25-34 years old” in all sample sets.

3.2.3 Hypotheses

- $H_0 : p_{24\text{-}35years} = \frac{1}{6}$
  - The true proportion of age group “25-34 years old” is $\frac{1}{6}$
- $H_A : p_{24\text{-}35years} > \frac{1}{6}$
  - The true proportion of age group “25-34 years old” is greater than $\frac{1}{6}$

3.2.4 Sample Statistic

The sample statistic is $\hat{p} = \frac{182047}{1034013} = 0.1760587$, where there are 182047 suicides in age group “25-34 years old” and 1034013 suicides in total from 1985 to 2015 in the United States.

3.2.5 Test Statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.1760587 - \frac{1}{6}}{\sqrt{\frac{\frac{1}{6}(\frac{5}{6})}{1034013}}} = 25.62652$$

3.2.6 Distribution of the test statistic

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

The p-value from Exact Binomial Test is < 2.2e-16 which is really close to 0.
The p-value from Normal Approximation is 3.8636e-14 which is really close to 0.

3.2.7 Confidence Interval

With 95% confidence interval, the lower bound is 0.1754428 and the upper bound is 1.0000000 from Exact Binomial Test where the alternative hypothesis is one-sided.
With 95% confidence interval, the lower bound is 0.1760585 and the upper bound is 1.0000000 from Normal Approximation with z = 25.62652.
The 95% confidence interval is similar with both methods, the 0.16666 is not included in the confidence interval in this case.

3.2.8 Interpretation

There is very strong evidence (p-value≈0) that the true proportion of number of suicides of age group "25-34 years old" is greater than $\frac{1}{6}$. We can reject the null hypothesis that the true proportion of number of suicides of age group "25-34 years old" is equal to $\frac{1}{6}$ at the α = 0.05 level. With 95% confidence, the true proportion of number of suicides of age group "25-34 years old" is between 0.1754428 and 1.

3.2.9 Bootstrap Approach

The sampling distribution of the sample proportion which was simulated 1000 times.

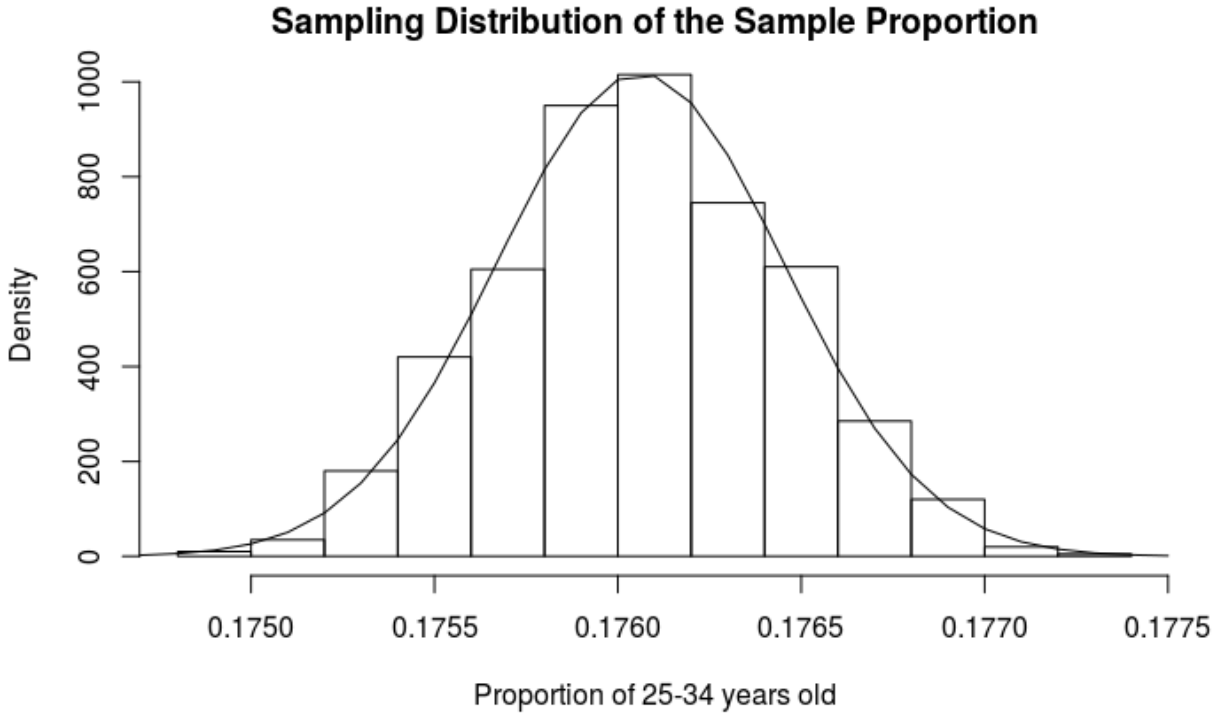

Figure 4 Sampling Distribution of the Sample Proportion

The figure 2 shows the proportion of the age group "25-34 years old" is about 0.1760 and is normally distributed about 1000 times simulations.

The 95% confidence interval from bootstrap approach is 0.1754165 and 1.00000.
The 95% confidence interval from Exact Binomial Test is 0.1754428 and 1.00000.
The 95% confidence interval from Normal Approximation is 0.1754444 and 1.0000.
All three methods give similar results of 95% confidence interval in this case.

Shifting the sample so that the null hypothesis is true which also was simulated 1000 times.

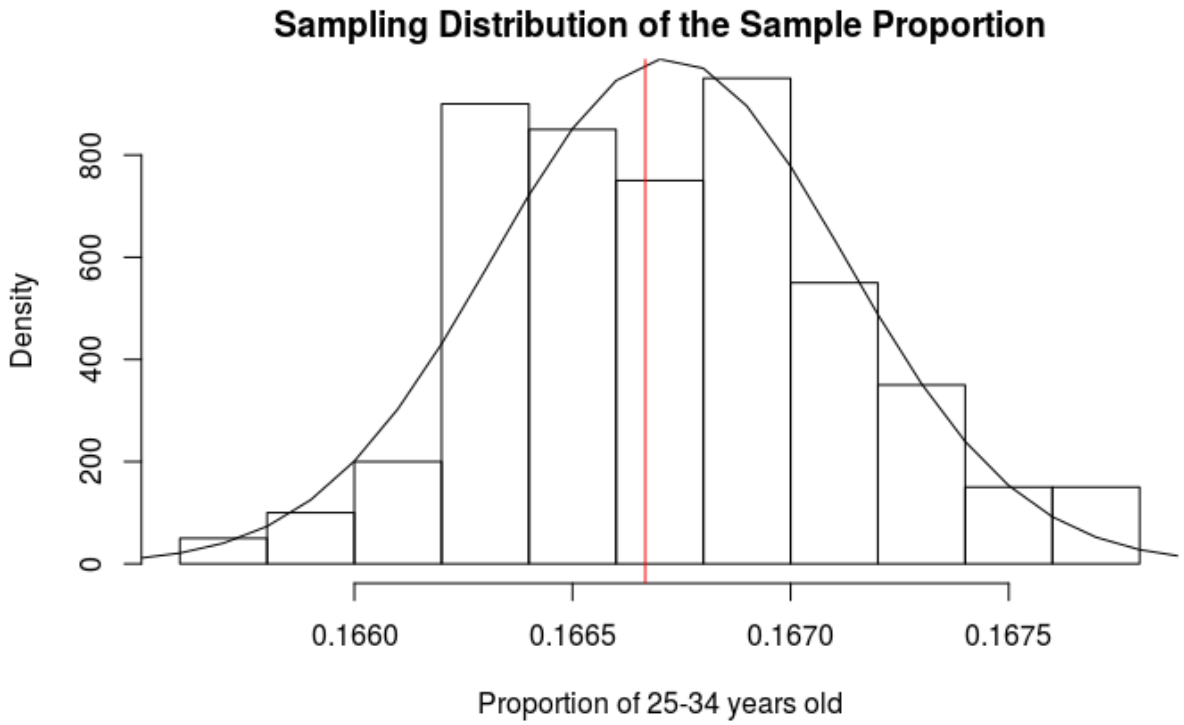

Figure 5 Sampling Distribution of the Sample Proportion under H_0: $p = \frac{1}{6}$

The figure 5 shows that the proportion of age group "25-34 years old" is about 0.16666 where the null hypothesis is true. It is normally distributed by 1000 times simulations.

The p-value from bootstrap approach is 0.
The p-value from Exact Binomial Test is 4.81915e-143 which is very close to 0.
The p-value from Normal Approximation is 3.8636e-145 which is very close to 0.
All three methods give similar results of p-value in this case.

In both traditional method and bootstraps approach, we can reject the null hypothesis that the true proportion of number of suicides of age group "25-34 years old" is equal to $\frac{1}{6}$ at the $\alpha = 0.05$ level.

3.3 Two Sample t-Test for Difference in Means

There are two independent populations, female and male, in the dataset. The total number of suicides and the total number of populations are both different between two populations. Therefore, the number of suicides per 100k people is used for both populations to be less bias based on the total number of populations.

Question: Is there any difference in average number of suicides per 100k people between female and male?

3.3.1 Conditions for use of One Sample t-Test

The sample is representative of the population because the dataset is collected from World Health Organization. The question of interest is with the difference of means between two populations, the female and male populations and the difference in the mean number of suicides per 100k people for each population. There are two independent samples from two populations. Sample size is greater than 30. The population data is normally distributed, show in the following section.

3.3.2 QQ Plot

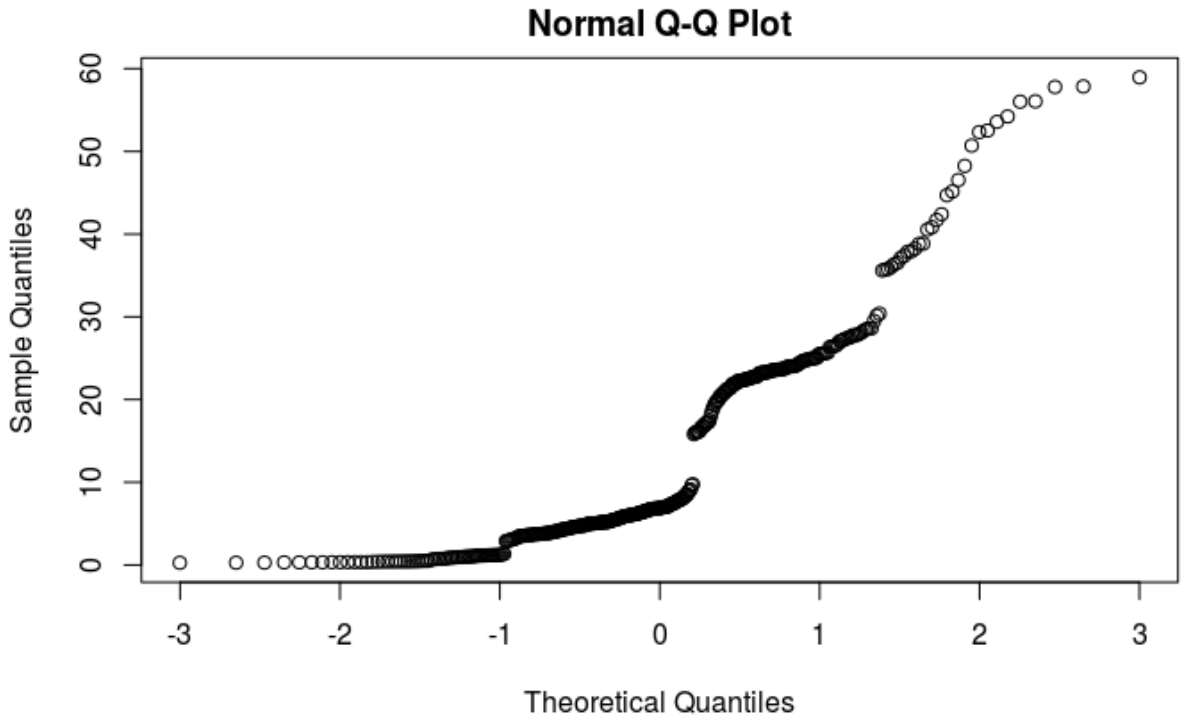

Figure 6 QQ Plot for All People

The figure 6 shows that every individual in the QQ Plot. Although there are some gaps between, overall most of the samples are located between -1 and 1. And the line looks pretty good to show that it is normally distributed.

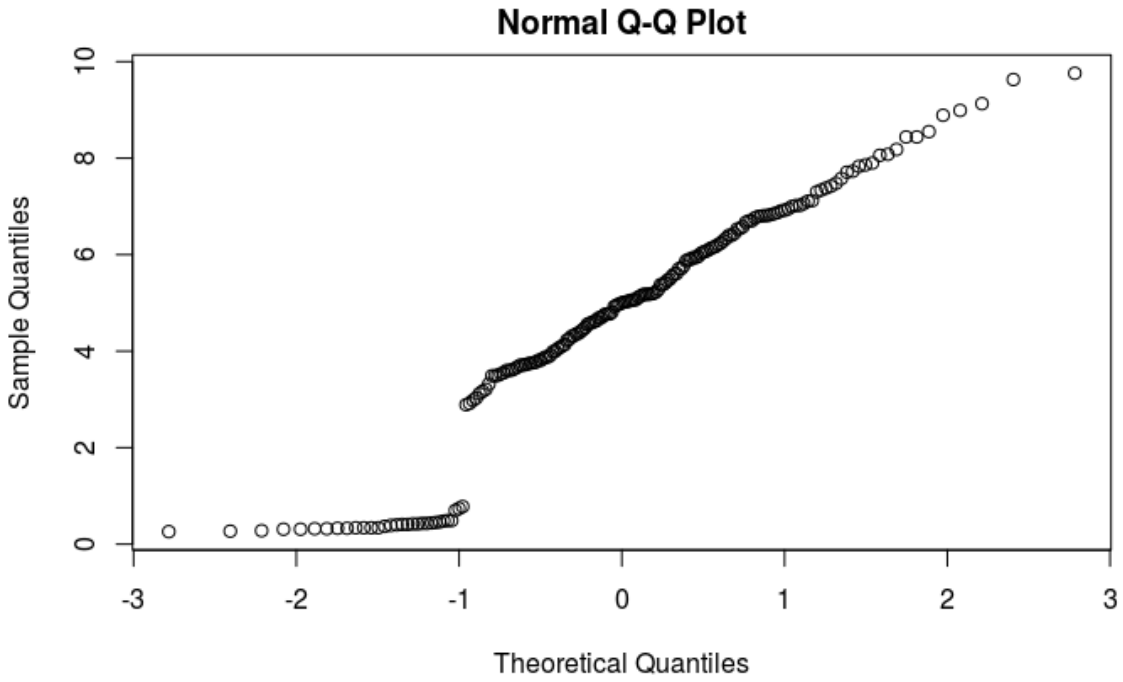

Figure 7 QQ Plot for Female

The figure 7 shows all the sample of female in the dataset. The line is about x = y which shows that female is normally distributed.

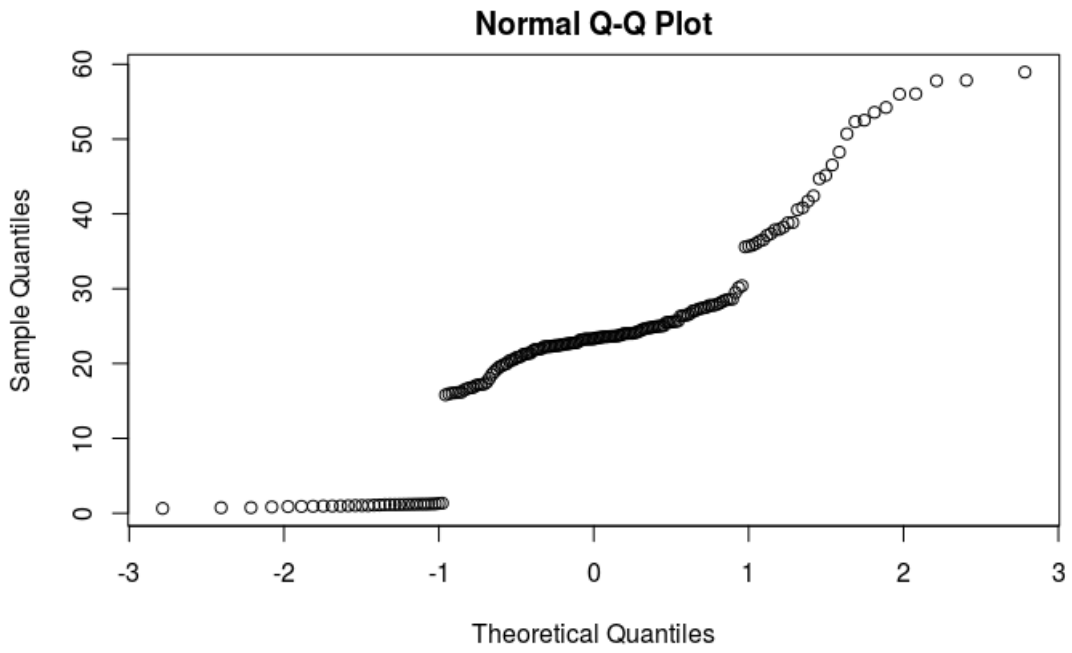

Figure 8 QQ Plot for Male

The figure 8 shows the sample of male in the dataset. There are two gaps at -1 and 1 which shows that there is a jump for the number of suicides in that number range. The line is good enough to show that male it is normally distributed.

3.3.3 Parameter

The study is interested in the true population mean difference in number of suicides per 100k people between female and male, $\mu_f - \mu_m$

3.3.4 Hypotheses

- $H_0 = \mu_f - \mu_m = 0$
  - The true population mean number of suicides per 100k female is equal to the true population mean number of suicides per 100k male.
- $H_A = \mu_f - \mu_m \neq 0$
  - The true population mean number of suicides per 100k female is different than the true population mean number of suicides per 100k male.

3.3.5 Sample Statistic

$$\bar{x}_f - \bar{x}_m$$

The sample statistic of the study is difference in mean number of suicides per 100k people for two sample data, referring to people who is female and male.

3.3.6 Test Statistic

$$t_{\min(n_f-1,n_m-1)} = \frac{(\overline{x}_f - \overline{x}_n) - (\mu_f - \mu_m)}{\sqrt{\frac{s_f^2}{n_f} + \frac{s_m^2}{n_m}}} = \frac{(4.6763 - 22.963) - 0}{\sqrt{\frac{5.6937}{186} + \frac{177.2037}{186}}} = -18.4413$$

3.3.7 Distribution of Test-Statistic

$$t_{\min(n_f-1,n_m-1)} = \frac{(\overline{x}_f - \overline{x}_m) - (\mu_f - \mu_m)}{\sqrt{\frac{s_f^2}{n_f} + \frac{s_m^2}{n_m}}} \sim t_{\min(n_f-1,n_m-1)}$$

The p-value is 8.980561e-44 which is very close to 0.

3.3.8 Confidence Interval

With 95% confidence, the lower bound of confidence interval is -20.24317 and the upper bound is -16.33048 in difference mean number of suicides per 100k people between female and male. The 0 difference in not included in the confidence interval in this case.

3.3.9 Comparison with R build-in t-test

From the R build-in t-test, t = -18.441 df = 196.88, p-value < 2.2e-16 which is really close to 0, and 95% confidence interval is between -20.24240 and -16.33126. The results are similar to what we calculated above.

3.3.10 Interpretation

There is very strong evidence (p-value ≈ 0) to suggest that the true population mean number of suicides per 100k people for female is different than male. We reject the null hypothesis that there is no difference in mean number of suicides per 100k people between female and male at the $\alpha = 0.05$ level. With 95% confidence, the true difference in mean number of suicides per 100k people between female and male is between -20.24317 and -1633048. The zero difference in mean stated in null hypothesis is not in the 95% confidence interval which is consistent with the rejection of the null hypothesis in this case. The values of confidence interval shows that the female has lower number of suicides per 100k people than male.

3.3.11 Bootstrap Approach

The sampling distribution with 1000 simulations for female, male and difference in mean. Shuffling the sample values since there is no difference with null is true.

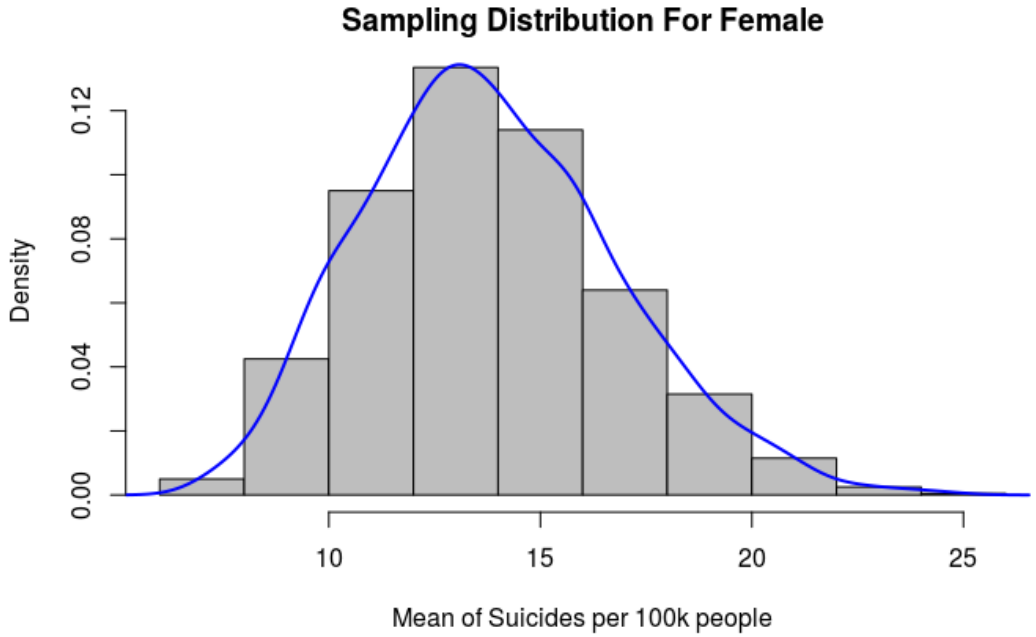

Figure 9 Sampling Distribution for Female

The figure 9 shows the average number of suicides per 100k female which was simulated 1000 times. It is normally distributed.

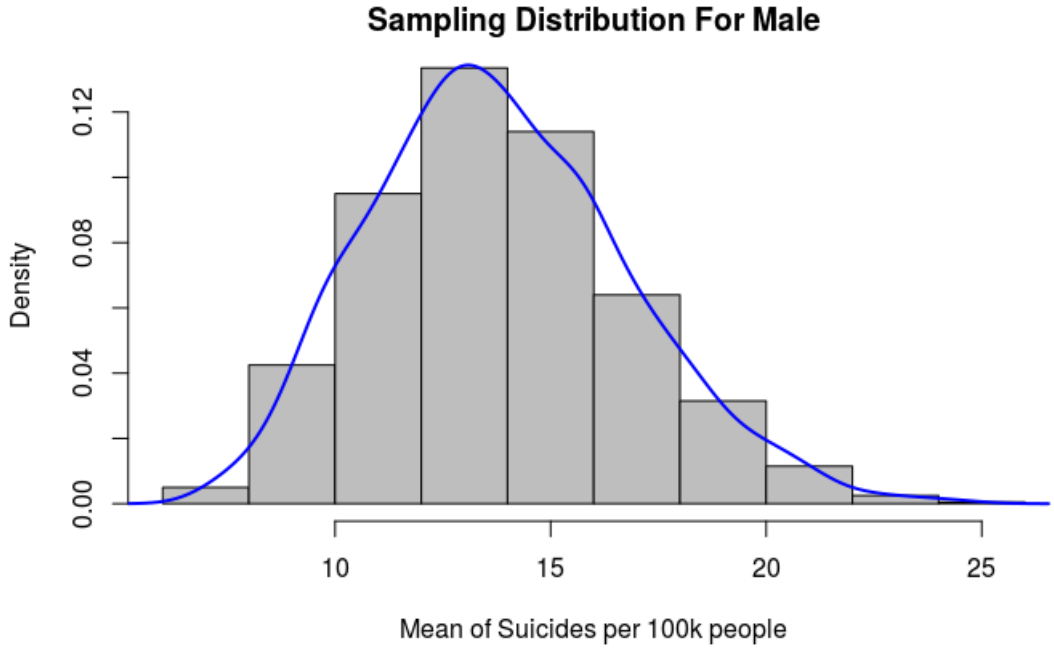

Figure 10 Sampling Distribution for Male

The figure 10 shows the average number of suicides per 100k male which was simulated 1000 times. It is normally distributed.

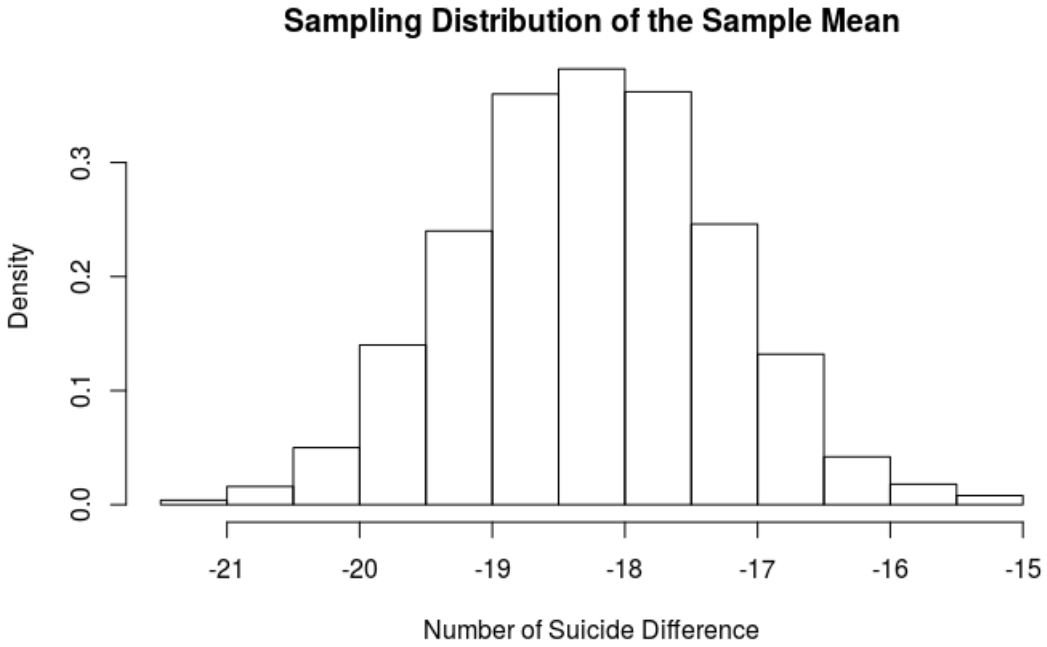

Figure 11 Sampling Distribution of the Sample Mean

The figure 11 shows the number of suicides difference in mean between female and male. The distribution is normal after 1000 simulations.

The 95% confidence interval from bootstrap approach is -20.14702 and -16.35057.
The 95% confidence interval from test statistic is -20.24317 and -16.33048.
The 95% confidence interval from t-test is -20.24240 and -16.33126.
All three methods have similar results in this case.

Shifting the sample so that the null hypothesis is true which also with 1000 simulations. And shuffling the sample values since the null is true.

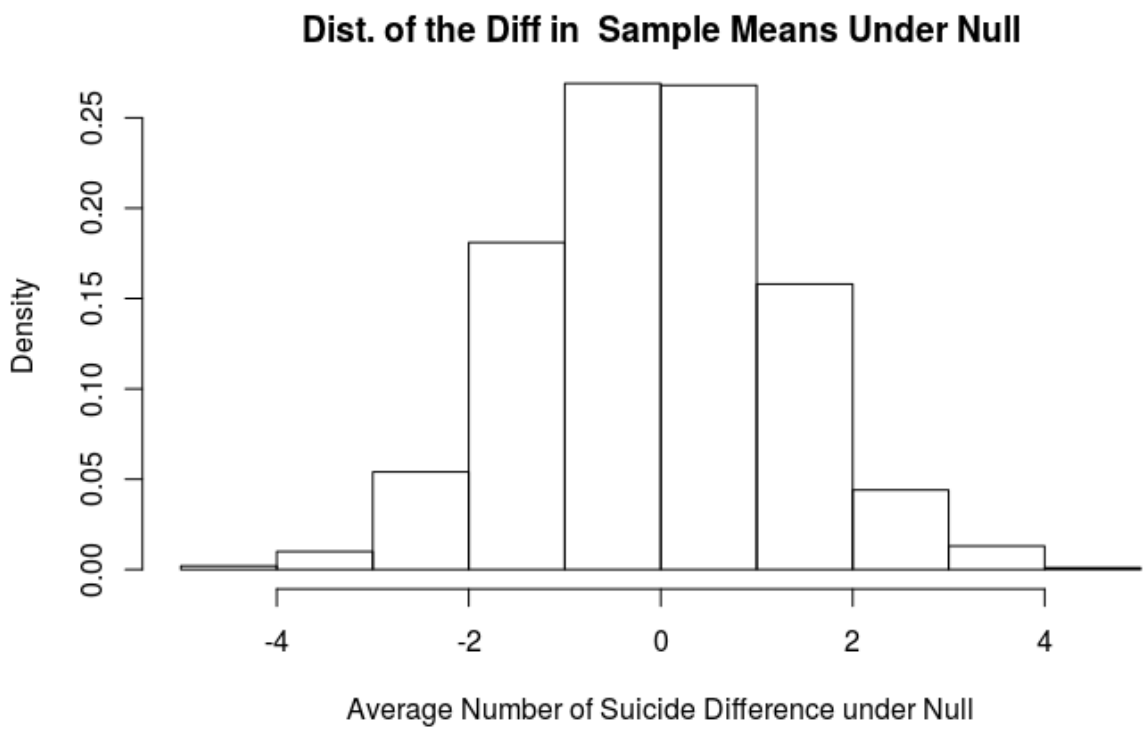

Figure 12 Distribution of the Difference in Sample Means under Null is True

The figure 12 shows that the difference in mean is 0 when null hypothesis is true. The positive/negative difference in mean are both out of the figure range, therefore not showing on the graph.

The p-value from bootstrap is 0.
The p-value from test statistic is 8.980561e-44 which is very close to 0.
The p-value from t-test is 9.082045e-45 which is really close to 0.
All three methods show similar results of p-value = 0 in this case.

In both traditional method and bootstraps approach, we can reject the null hypothesis that there is no difference in mean number of suicides per 100k people between female and male at the $\alpha = 0.05$ level.

3.4 Two Sample Test for Difference in Proportions

This test is interested in the difference in two independent proportions. The proportion of female is the number of suicides of female divided by the number of female population. The proportion of male is the number of suicides of male divided by the number of male population. There are some parts of calculation that the total of populations are adjusted because it is easier for R to run graph.

Question: Is there any difference between the proportion of female suicides and the proportion of male suicides?

3.4.1 Conditions for use of Two Sample Test for Difference in Proportions

The sample is representative of the population. The categorical response variable with two categories which suicide or not is 1 and 0 in this case. There are 2 independent samples, suicide and not suicide, from two populations, female and male. $np = 213797 \geq 10$ and $n(1-p) = 820216 \geq 10$ in this case.

3.4.2 Parameter

The study is interested in the difference between the true population proportion of female suicides and the true proportion of male suicides. $p_f - p_m$

3.4.3 Hypotheses

- $H_0 : p_f - p_m = 0$
  - There is no difference between the true population proportion of female suicides and the true population proportion of male suicides.
- $H_A : p_f - p_m \neq 0$
  - There is difference between the true population proportion of female suicides and the true population proportion of male suicides.

3.4.4 Sample Statistic

$$\widehat{p_f} - \widehat{p_m}$$

3.4.5 Test Statistic

$$z = \frac{(\widehat{p_f} - \widehat{p_m}) - (p_f - p_m)}{\sqrt{\frac{\widehat{p_f}(1-\widehat{p_f})}{n_f} + \frac{\widehat{p_m}(1-\widehat{p_m})}{n_m}}} = 610.5068$$

The p-value is 0 in this case.

3.4.6 Confidence Interval

With 95% confidence, the lower bound of confidence interval is -0.0001566887 and the upper bound is -0.0001556879. The 0 difference is not included in the confidence interval in this case.

3.4.7 Bootstrap and Randomization Approach

The sampling distribution with 1000 simulations for proportion of female and male. Shuffling the sample values since there is no difference when the null is true.

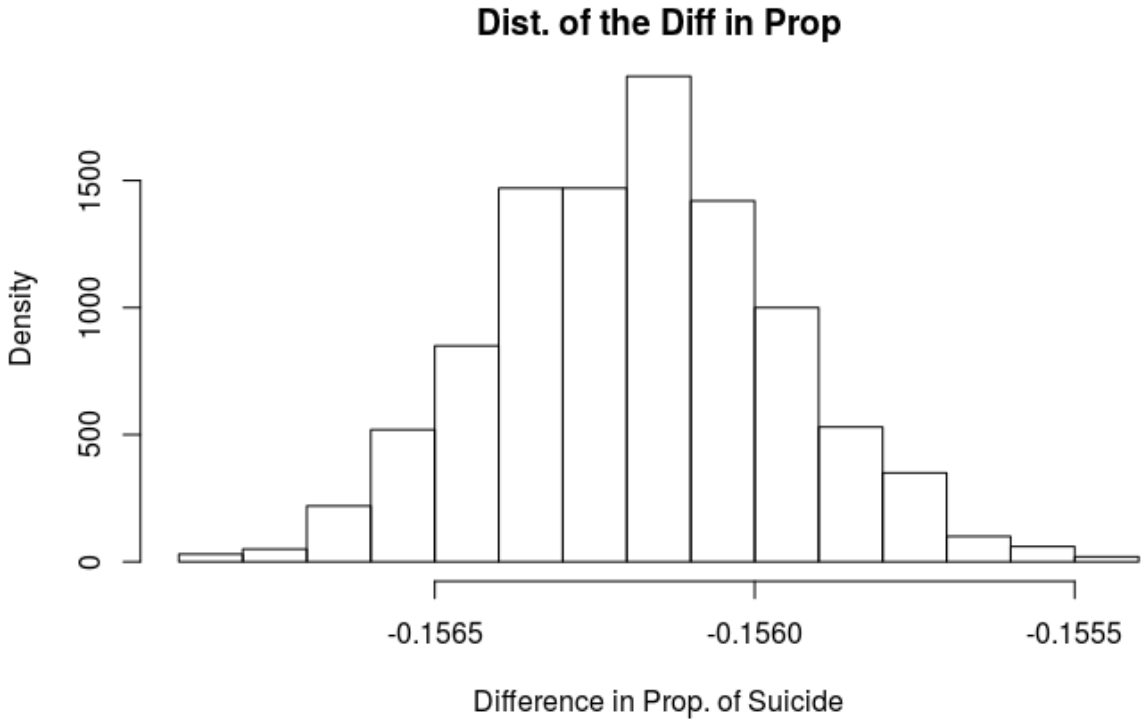

Figure 13 Distribution of the Difference in Proportion

The figure 13 shows the distribution of difference in proportion between female and male. The number shows on the graph is adjusted by 1000. All the number need to divide by 1000 or times by 0.001 to get the real results.

The 95% confidence interval from bootstrap is -0.0001561873 and -0.0001551832.
The 95% confidence interval from test statistic is -0.0001566887 and -0.0001556879.
The 95% confidence interval from Normal Approximation is -0.0001566887 and -0.0001556859.
All three methods show similar results of 95% confidence interval in this case.

Shifting the sample so that the null hypothesis is true which also with 100 simulations. And shuffling the sample values since the null is true.

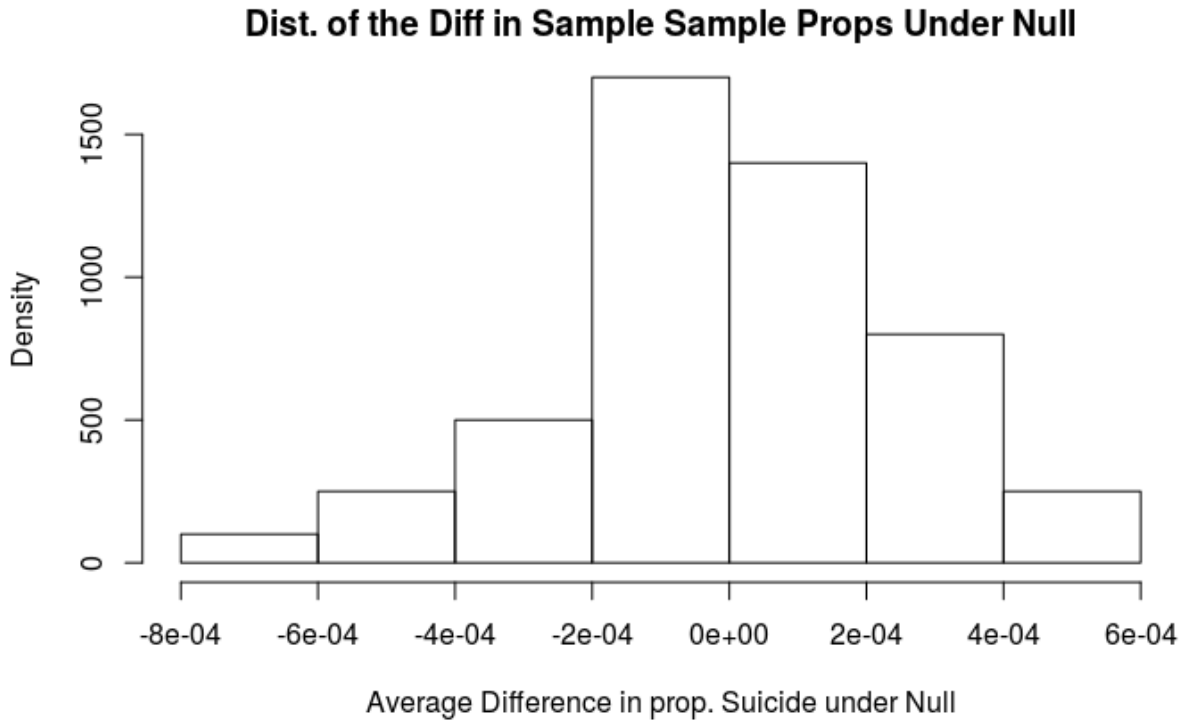

Figure 14 Distribution of the Difference in Sample Proportion Under Null is True

The figure 14 is the distribution of the difference in proportion under null is True. The mean of difference is 0 on the graph. The graph only simulated 100 times because the limitation of R.

The p-value from bootstrap is 0.
The p-value from test statistic is 0.
Both methods show the result of p-value is 0.

3.4.8 Interpretation

Using randomization methods, there is very strong evidence (p-value = 0) to suggest that there is a difference between the true proportion of female suicides and male suicides. We reject the null hypothesis that the true proportion of female suicides is equal to the true proportion of male suicides at the $\alpha = 0.05$ level. Using confidence intervals created by the bootstrap method, we can say that 95% confidence that the true population proportion difference is between -0.0001561873 and -0.0001551832. The 0 difference in proportion is not in the 95% confidence interval is agree with our rejection of null hypothesis.

3.5 Chi-Square Test: Goodness of Fit Test

There is a categorical column with six independent age groups in the dataset. Each age group has the total number of suicides. The proportion of each group is the number of suicides of each group divided by the total number of suicides.

3.5.1 Conditions for use

There is single categorical variable with six categories, “5-14 years old”, “15-24 years old”, “25-34 years old”, “35-54 years old”, “55-74 years old”, and “75+ years old”. The expected count of each count is a lot more than 5 in this case.

3.5.2 Parameter

Let’s set the “5-14 years old” as “A”, “15-24 years old” as “B”, “25-34 years old” as “C”, “35-54 years old” as “D”, “55-74 years old” as “E”, and “75+ years old” as “F”.
The study is interested in the true p_A, p_B, p_C, p_D, p_E, p_F

3.5.3 Hypotheses

- H_0 : $p_A = p_B = p_C = p_D = p_E = p_F = \frac{1}{6} = 0.16666$
  - The proportion of each age group is the same and is equal to 0.16666
- H_A : Some $p_i \neq 0$
  - At least one of the proportion is not equal to 0.16666

3.5.4 Sample Statistics

There are 6 sample statistics in this case.

$$\widehat{p_A}, \widehat{p_B}, \widehat{p_C}, \widehat{p_D}, \widehat{p_E}, \widehat{p_F}$$

3.5.5 Test Statistic and Distribution

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

The chi-square = 463456 from R. The number is huge because there are some huge number of (Observe – Expected).
The p-value of chi-square is 0.

3.5.6 Confidence Interval

There is no confidence interval for a goodness of fit test.

3.5.7 Randomization Approach

Create a new data and assume the null is true.

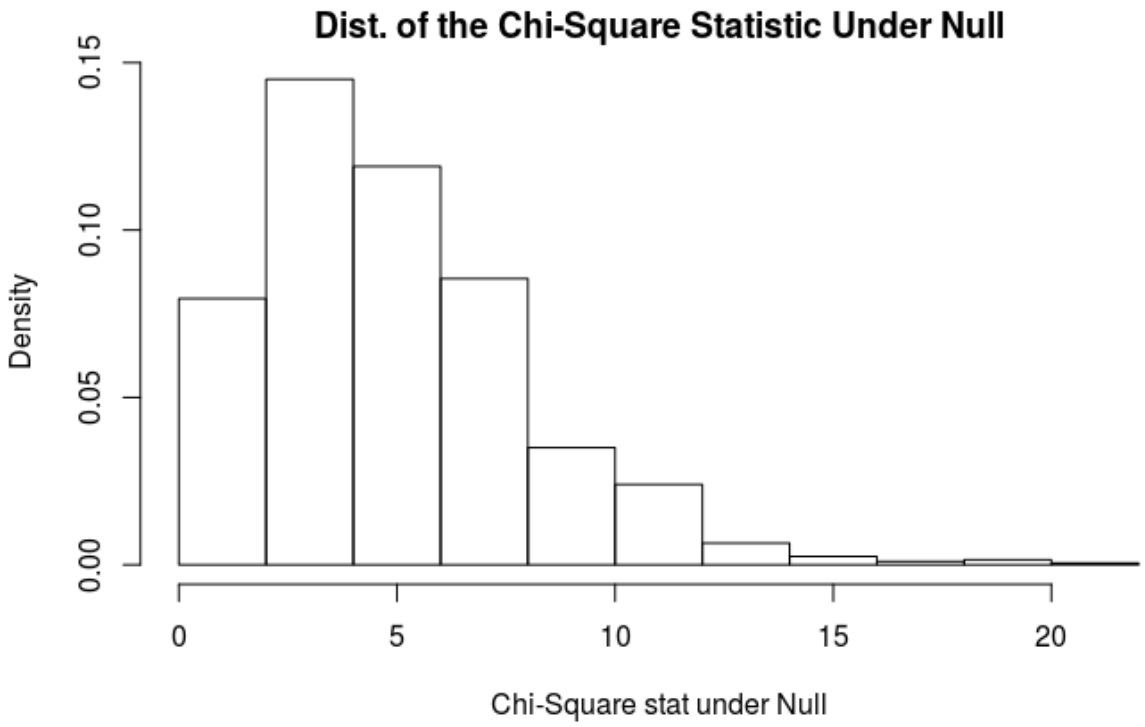

Figure 15 Distribution of the Chi-Square Statistic under Null is True

The figure 15 shows the distribution of chi-square under null hypothesis is true. The graph is skewed as expected and the numbers stop at 0.

The randomization p-value is 0.

3.5.8 Interpretation

There is very strong evidence (p-value = 0) that there is at least one proportion of age group is different than 0.16666. We reject the null hypothesis that the proportions of the age group are all equal to 0.16666 at the $\alpha = 0.05$ level.

3.6 Chi-Square Test: Association between Two categorical Variables

There are two categorical variables. First is the gender, female and male. Second is the age group, “5-14 years old”, “15-24 years old”, “25-34 years old”, “35-54 years old”, “55-74 years old”, “and “75+ years old”. They are separated by 2 X 6 rows in the table. Each gender has the number of suicides for each age group.

3.6.1 Conditions for use

There are two categorical variables and at least one quantitative variable in the dataset. Each group in the categorical variables is independent.

3.6.2 Parameter

There is no parameter of interest. It is testing the association between two categorical variables.

3.6.3 Hypotheses

- H_0 : The number of suicides is not associated with gender
- H_A : The number of suicides is associated with gender

3.6.4 Sample Statistics

	Gender		
Responses	Female	Male	SUM
5-14 years old	2447	6476	8923
15-24 years old	22903	118776	141679
25-34 years old	33796	148251	182047
35-54 years old	91572	289345	380917
55-74 years old	48591	176179	224770
75+ years old	14488	81189	95677
SUM	213797	820216	1034013

	Gender		
Responses	Female	Male	SUM
5-14 years old	0.002366508	0.006262977	0.008629485
15-24 years old	0.022149625	0.114868962	0.137018587
25-34 years old	0.032684309	0.143374406	0.176058715
35-54 years old	0.088559815	0.279827236	0.368387051
55-74 years old	0.046992639	0.170383738	0.217376377
75+ years old	0.014011429	0.078518355	0.092529784
SUM	0.206764325	0.793235675	1.000000000

3.6.5 Test Statistic and Distribution

$$\text{Expect Count} = \frac{(Row\ Total)(Column\ Total)}{(Sample\ Size)}$$

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

The chi-square is 7035.894 according to R in this case.

The p-value is 0 according to R in this case.

3.6.6 Confidence Interval

There is no confidence interval for the association test.

3.6.7 Interpretation

There is very strong evidence (p-value = 0) to suggest that there is an association between gender and number of suicides. Using a chi-square test of association, we find a chi-squared test statistic of 7035.894 with 5 degree of freedom resulting in p-value = 0. We reject the null hypothesis that there is no association between gender and age group on number of suicides at the $\alpha = 0.05$ level.

4. Discussion

4.1 Summary

In summary, this study is about the number of suicides between 1985 and 2015 in United States. The study is focusing on the relationship between the suicides and different genders or different age groups. First, from one sample t-tests, the result shows that there is difference from the average number of suicide according to the American Foundation for Suicide Prevention in 2017. The average of number of suicides between 1985 and 2015 is less than the population mean. Secondly, the proportion of age group "25-34 years old" is greater than 0.166666 which is 1/6 from one sample test of proportion. Third, when separate the gender to male and female, and see them as two different independent populations, the study found that the mean of female and the mean of male is different according to the two sample t-test for difference in mean. Fourth, the suicide proportion of population for female is different from the suicide proportion of population for male. It is the results found in two sample test for difference in proportion. There is bootstrap methods using in all one sample and two sample tests to enhance the accuracy of the results. Moreover, to understand if the number of suicides are happened evenly in each age group, the study use chi-square goodness of fit test to check. The result shows that there is at least one age group is not 0.166666 in proportion. The proportion of age group can be greater or less than 0.166666. Finally, chi-square association test is presented in the last section. The test shows that the number of suicides is associated with gender which shows the same results from figure 2.4 and figure 2.5. From figure 2.4 and figure 2.5, the number of suicides from male is greater than the number of suicides for female every year between 1985 and 2015.

4.2 Implication

The number of suicides is increasing every year. There are many people who died by suicide each year which is the 10^{th} causes of death in the United States according to American Foundation for Suicide Prevention. From this study, we learned that the number of suicides is increasing. And there is relationship between the gender and age for the suicides. It is a good dataset and study for further work about suicide prevention.

4.3 Limitations

There are some limitations about this study. For example, there is no location recorded in the dataset. We do not know where the person is from and located. And there is no ethnic information about the people who were recorded, the proportion of races may be very different between 1985 and 2015 in the United States. Furthermore, there is questions about the accuracy from the further past years.

4.4 Further Questions and Next Steps

To do extensions from this study, the researcher can collect more recent data to compare this dataset because this dataset only contains from 1985 to 2015. Moreover, there are 100 other countries data in the original dataset. The further study can be the comparison between different countries. Furthermore, the researcher can collect data from same years but different sources because there is some limitation about the data of number of suicides. The data many not be accurate enough because not every suicide will be recorded as suicide due to some morally reason or the reason of death is simply just unknown. The larger amount of data, and diversity of the data source is better for study.

5. APPENDIX

DATA Source: Suicide Rates Overview 1985 to 2016
From: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

R code

```
##title: "IE6200 Final_Hung"
##output: html_document

read.csv("/cloud/project/SuicideData.csv")
data <- read.csv("SuicideData.csv")
head(data)
summary(data)

##Boxplot for female and male
boxplot(suicides_no~sex, data = data.US, xlab = "Gender", ylab = "Number of Suicides",
      ylim = c(0, 12000), main = "Total Number of Suicides by Gender")


##ggplot with total number vs population
library(ggplot2)
ggplot(data.US, aes(x = data.US$population, y = data.US$suicides_no, color = data.US$sex))+
 geom_point(size = 2) +
 ggtitle("Number of Suicides vs Population")+
 scale_color_manual(values = c("Blue","Orange"))+
 ylab("Number of Suicides")+
 xlab("Population")+
 guides(color=guide_legend(title="Gender"))


##line plot with total number of each year
plot(UStotal$Year, UStotal$TotalSuicide, main = "Total number of Suicides between 1985
and 2015 in US",
    xlab = "Year", ylab = "Total number of Suicides", ylim = c(25000,45000), type = "o", col =
"red")


##ggplot with total number for female
library(tidyverse)
filter1 <- data.US %>% filter(sex=="female")
```

```
library(ggplot2)

ggplot(data=filter1, aes(x = year, y = suicides_no, fill = age))+
  geom_bar(stat="identity", width = 0.5)+
  xlab("Year")+
  ylab("Number of Suicides")+
  ggtitle("Total number of Suicide of Female Between 1985 and 2015 in US")

##ggplot with total number for male
library(tidyverse)
filter2 <- data.US %>% filter(sex=="male")
library(ggplot2)

ggplot(data=filter2, aes(x = year, y = suicides_no, fill = age))+
  geom_bar(stat="identity", width = 0.5)+
  xlab("Year")+
  ylab("Number of Suicides")+
  ggtitle("Total number of Suicide of Male Between 1985 and 2015 in US")

##ggplot with total number vs group age
library(ggplot2)

ggplot(data=data.US, aes(x = factor(age, level = c("5-14 years", "15-24 years", "25-34
years", "35-54 years", "55-74 years", "75+ years")), y = suicides_no, fill = age))+
  geom_bar(stat="identity", width = 0.5)+
  xlab("Age")+
  ylab("Number of Suicides")+
  coord_flip()+
  ylim(0,400000)+
  ggtitle("Total number of Suicide Between 1985 and 2015 in US")


##Subset data where country = United States
data.US <- subset(data, data$country=="United States")
##Calculate the total suicide each year from all age ranges
us1985 <- sum(data.US$suicides_no[data.US$year=="1985"])
us1986 <- sum(data.US$suicides_no[data.US$year=="1986"])
us1987 <- sum(data.US$suicides_no[data.US$year=="1987"])
us1988 <- sum(data.US$suicides_no[data.US$year=="1988"])
us1989 <- sum(data.US$suicides_no[data.US$year=="1989"])
```

```
us1990 <- sum(data.US$suicides_no[data.US$year=="1990"])
us1991 <- sum(data.US$suicides_no[data.US$year=="1991"])
us1992 <- sum(data.US$suicides_no[data.US$year=="1992"])
us1993 <- sum(data.US$suicides_no[data.US$year=="1993"])
us1994 <- sum(data.US$suicides_no[data.US$year=="1994"])
us1995 <- sum(data.US$suicides_no[data.US$year=="1995"])
us1996 <- sum(data.US$suicides_no[data.US$year=="1996"])
us1997 <- sum(data.US$suicides_no[data.US$year=="1997"])
us1998 <- sum(data.US$suicides_no[data.US$year=="1998"])
us1999 <- sum(data.US$suicides_no[data.US$year=="1999"])
us2000 <- sum(data.US$suicides_no[data.US$year=="2000"])
us2001 <- sum(data.US$suicides_no[data.US$year=="2001"])
us2002 <- sum(data.US$suicides_no[data.US$year=="2002"])
us2003 <- sum(data.US$suicides_no[data.US$year=="2003"])
us2004 <- sum(data.US$suicides_no[data.US$year=="2004"])
us2005 <- sum(data.US$suicides_no[data.US$year=="2005"])
us2006 <- sum(data.US$suicides_no[data.US$year=="2006"])
us2007 <- sum(data.US$suicides_no[data.US$year=="2007"])
us2008 <- sum(data.US$suicides_no[data.US$year=="2008"])
us2009 <- sum(data.US$suicides_no[data.US$year=="2009"])
us2010 <- sum(data.US$suicides_no[data.US$year=="2010"])
us2011 <- sum(data.US$suicides_no[data.US$year=="2011"])
us2012 <- sum(data.US$suicides_no[data.US$year=="2012"])
us2013 <- sum(data.US$suicides_no[data.US$year=="2013"])
us2014 <- sum(data.US$suicides_no[data.US$year=="2014"])
us2015 <- sum(data.US$suicides_no[data.US$year=="2015"])


##Create a data frame to plot QQ plot
UStotal <- data.frame("Year" = 1985:2015, TotalSuicide = c(us1985, us1986, us1987,
                                        us1988, us1989, us1990,
                                        us1991, us1992, us1993,
                                        us1994, us1995, us1996,
                                        us1997, us1998, us1999,
                                        us2000, us2001, us2002,
                                        us2003, us2004, us2005,
                                        us2006, us2007, us2008,
                                        us2009, us2010, us2011,
                                        us2012, us2013, us2014, us2015))
UStotal
```

ONE SAMPLE t-TEST - Traditional

##Null Hypothesis: mu = 47085 (The true mean suicide number per year is 47085)
##Alternate Hypothesis: mu != 47085 (The true mean suicide number per year is different than 47085)

```
##QQ plot
qqnorm(UStotal$TotalSuicide)

##sample mean
x_bar1 <- mean(UStotal$TotalSuicide)
##mu_0
mu_01 <- 47085
##sample SD
s1 <- sd(UStotal$TotalSuicide)
##number of years
n1 <- 31
##t-test test statistic
t1 <- (x_bar1 - mu_01)/(s1/sqrt(n1))
##two-sided p-value
two_sided_t_p1 <- pt(q = t1, df = n1-1)*2
two_sided_t_p1

##CI
lower1 <- x_bar1 + (qt(0.025, n1-1)*(s1/sqrt(n1)))
upper1 <- x_bar1 + (qt(0.975, n1-1)*(s1/sqrt(n1)))
lower1
upper1

t.test(UStotal$TotalSuicide, alternative = "two.sided", mu = 47085)
```

ONE SAMPLE t-Test - Bootstrap Approach

```
##Simulations 100 times
set.seed(111)
num_sims1 <- 10000
##A vector to store my results
results1 <- rep(NA, num_sims1)
```

```
##A loop for completing the simulation
for(i in 1:num_sims1){
  results1[i] <- mean(sample(x = UStotal$TotalSuicide, size = n1, replace = TRUE))
}
##Finally plot the results
hist(results1, freq = FALSE, main = "Sampling Distribution of the Sample Mean",
    xlab = "Average Suicide Numbers", ylab = "Density")
##Estimate a normal curve over it
lines(x = seq(31000,36000, 1), dnorm(seq(31000,36000, 1), mean = x_bar1,
                          sd = sd(UStotal$TotalSuicide)/sqrt(n1)))

##Shift the sample so that the null hypothesis is true
set.seed(111)
time_given_H0_true1 <- UStotal$TotalSuicide - mean(UStotal$TotalSuicide) + mu_01
##Simulations
num_sims2 <- 10000
##A Vector to store my results
results_given_h0_true1 <- rep(NA, num_sims2)
##A loop for completing the simulation
for(i in 1:num_sims2){
  results_given_h0_true1[i] <- mean(sample(x = time_given_H0_true1, size = n1, replace =
TRUE))
}
#Finally Plot the results
hist(results_given_h0_true1, freq = FALSE, main = "Sampling Distribution Null is True",
    xlab = "Number of Suicide", ylab = "Density")

set.seed(111)
##More extreme on lower end
low_end_extreme <- mean(results_given_h0_true1)+(mean(results_given_h0_true1)-
x_bar1)
##Counts of values more extreme than the test statistic in our original sample, given H0 is
true
##Two sided given the alternate hypothesis
count_of_more_extreme_lower <- sum(results_given_h0_true1 >= low_end_extreme)
count_of_more_extreme_upper <- sum(results_given_h0_true1 <= x_bar1)

bootsrtap_p1 <- (count_of_more_extreme_lower +
count_of_more_extreme_upper)/num_sims2
bootsrtap_p1
```

```
two_sided_t_p1

##CI of bootstrap
##Need the standard error which is the sd of the results
bootstrap_SE_x_bar1 <- sd(results1)
##An estimate is to use the formula statisitc +/- 2*SE
c(x_bar1 - 2*bootstrap_SE_x_bar1, x_bar1 + 2*bootstrap_SE_x_bar1)

##Find 5th and 95th quantiles to determine the bounds
c(quantile(results1, c(0.025, 0.975)))

##Compare t-methods
c(x_bar1 + (qt(0.025, n1-1)*(s1/sqrt(n1))), x_bar1+qt(0.975, n1-1)*(s1/sqrt(n1)))

### ONE SAMPLE PROPORTION TEST - Traditional

totalUS_R1 <- sum(data.US$suicides_no[data.US$age=="5-14 years"])
totalUS_R2 <- sum(data.US$suicides_no[data.US$age=="15-24 years"])
totalUS_R3 <- sum(data.US$suicides_no[data.US$age=="25-34 years"])
totalUS_R4 <- sum(data.US$suicides_no[data.US$age=="35-54 years"])
totalUS_R5 <- sum(data.US$suicides_no[data.US$age=="55-74 years"])
totalUS_R6 <- sum(data.US$suicides_no[data.US$age=="75+ years"])
totalUS <- sum(data.US$suicides_no)

z1 <- ((totalUS_R3/totalUS)-(1/6)) / sqrt(((1/6)*(5/6))/totalUS)
z1

binom.test(x = totalUS_R3, n = totalUS, p=(1/6), alternative = "greater")

pnorm(z1, lower.tail = FALSE)

binom.test(x = totalUS_R3, n = totalUS, p=(1/6), alternative = "greater")$conf.int

##normal approx - greater
c((totalUS_R3/totalUS) - 1.64*(((totalUS_R3/totalUS)*(1-
(totalUS_R3/totalUS)))/totalUS),1)

binom.test(x = totalUS_R3, n = totalUS, p=(1/6), alternative = "two.sided")$conf.int

##normal approx - two.sided
```

```
c((totalUS_R3/totalUS) - 1.64*(((totalUS_R3/totalUS)*(1-(totalUS_R3/totalUS)))/totalUS),
  (totalUS_R3/totalUS) + 1.64*(((totalUS_R3/totalUS)*(1-(totalUS_R3/totalUS)))/totalUS))

## One proportion for 101 countries
total_R1 <- sum(data$suicides_no[data$age=="5-14 years"])
total_R2 <- sum(data$suicides_no[data$age=="15-24 years"])
total_R3 <- sum(data$suicides_no[data$age=="25-34 years"])
total_R4 <- sum(data$suicides_no[data$age=="35-54 years"])
total_R5 <- sum(data$suicides_no[data$age=="55-74 years"])
total_R6 <- sum(data$suicides_no[data$age=="75+ years"])
total <- sum(data$suicides_no)

##binom test for 101 countries - one sided
binom.test(x = total_R3, n = total, p=(1/6), alternative = "greater")

##binom test for 101 countries - two sided
binom.test(x = total_R3, n = total, p=(1/6), alternative = "two.sided")

### ONE SAMPLE TEST of PROPORTION - Boostrap Approach

Range3 <- factor(rep(c("25-34 years", "other"), c(totalUS_R3, (totalUS - totalUS_R3))))
table(Range3)

##Easier to use for bootstrapping
Range3_ <- rep(c(1, 0), c(totalUS_R3, (totalUS - totalUS_R3)))

set.seed(111)
##Simulations
num_sims3 <- 1000
##A vector to store results
results2 <- rep(NA, num_sims3)
##A loop for completing the simulations
for(i in 1:num_sims3){
  results2[i] <- mean(sample(x = Range3_, size = totalUS, replace = TRUE))
}
##Finally plot
hist(results2, freq = FALSE, main = "Sampling Distribution of the Sample Proportion",
    xlab = "Proportion of 25-34 years old", ylab = "Density")
##Estimate a normal curve over it
```

```
lines(x = seq(0.1745, 0.1775, 0.0001), dnorm(seq(0.1745, 0.1775, 0.0001), mean =
mean(results2),
                                sd = sd(results2)))

##CI of bootstrap
c(quantile(results2, c(0.05, 1)))

binom.test(x = totalUS_R3, n = totalUS, p=(1/6), alternative = "greater")$conf.int

c((totalUS_R3/totalUS) - 1.64*sqrt(((totalUS_R3/totalUS)*(1-
(totalUS_R3/totalUS)))/totalUS),1)

Range3_2 <- rep(c(1, 0), c(totalUS/6, (totalUS*5/6)))

##Under Null is TRUE
set.seed(111)
##Simulations
num_sims4 <- 100
##A vector to store results
results3 <- rep(NA, num_sims4)
##A loop for completing the simulations
for(i in 1:num_sims4){
  results3[i] <- mean(sample(x = Range3_2, size = totalUS, replace = TRUE))
}
##Finally plot
hist(results3, freq = FALSE, main = "Sampling Distribution of the Sample Proportion",
    xlab = "Proportion of 25-34 years old", ylab = "Density")
##Estimate a normal curve over it
lines(x = seq(0.1655, 0.1690, 0.0001), dnorm(seq(0.1655, 0.1690, 0.0001), mean =
mean(results3),
                                sd = sd(results3)))
abline(v = (1/6), col="red")

propR3 <- totalUS_R3/totalUS
##Count extreme
count_of_more_extreme_uppertail <- sum(results3 >= propR3)
bootstrap_p2 <- count_of_more_extreme_uppertail/num_sims4
bootstrap_p2

##Exact Binomial p-value
```

```
binom.test(x = totalUS_R3, n = totalUS, p=(1/6), alternative = "greater")$p.value

##Normal Approx
pnorm(z1, lower.tail = FALSE)

### TWO SAMPLE t-Test - Traditional

##two populations are female and male, with numbers of suicide per 100k people
##install.packages("tidyverse")
##library(tidyverse)
##data.US  %>% select(2:7)

##QQ plot of suicide number of male and female in different age ranges
qqnorm(data.US$suicides.100k.pop)

##QQ Plot of suicide number of female in different age ranges
qqnorm(data.US$suicides.100k.pop[data.US$sex=="female"])

##QQ Plot of suicide number of male in different age ranges
qqnorm(data.US$suicides.100k.pop[data.US$sex=="male"])

##sample means
x_bar_f <- mean(data.US$suicides.100k.pop[data.US$sex=="female"])
x_bar_m <- mean(data.US$suicides.100k.pop[data.US$sex=="male"])
##null hypothesized population mean difference between the two groups
mu_02 <- 0
##sample variances
s_f_sq <- sd(data.US$suicides.100k.pop[data.US$sex=="female"])**2
s_m_sq <- sd(data.US$suicides.100k.pop[data.US$sex=="male"])**2
##sample size
n_f <- length(data.US$suicides.100k.pop[data.US$sex=="female"])
n_m <- length(data.US$suicides.100k.pop[data.US$sex=="male"])
##t-test test statistic
t2 <- (x_bar_f - x_bar_m - mu_02)/sqrt((s_f_sq/n_f)+(s_m_sq/n_m))
##one sided upper p-value
one_sided_diff_t_p <- pt(q=t2, df=(min(n_f, n_m)-1), lower.tail = TRUE)*2
one_sided_diff_t_p

##lower bound
(x_bar_f - x_bar_m) + (qt(0.025, min(n_f, n_m)-1)*sqrt((s_f_sq/n_f)+(s_m_sq/n_m)))
```

```
##upper bound
(x_bar_f - x_bar_m) + (qt(0.975, min(n_f, n_m)-1)*sqrt((s_f_sq/n_f)+(s_m_sq/n_m)))

##use t-test function in R
t.test(data.US$suicides.100k.pop[data.US$sex=="female"],data.US$suicides.100k.pop[data.
US$sex=="male"])

### TWO SAMPLE Diff in Mean - Bootstrap Approach

##install.packages('mosaic')
require(mosaic)
set.seed(111)
x_hist1 <- data.US %>%
  filter(data.US$sex== "female")
x_hist1 <- do(1000) * mean(sample(data.US$suicides.100k.pop,20))
h1 <- hist(x_hist1$mean, main = "Sampling Distribution For Female", xlab = "Mean of
Suicides per 100k people",
         prob = T, col = "grey")
lines(density(x_hist1$mean), col = "blue", lwd = 2)
h1

require(mosaic)
set.seed(111)
x_hist0 <- data.US %>%
  filter(data.US$sex=="male")
x_hist0 <- do(1000) * mean(sample(data.US$suicides.100k.pop,20))
h0 <- hist(x_hist1$mean, main = "Sampling Distribution For Male", xlab = "Mean of Suicides
per 100k people",
         prob = T, col = "grey")
lines(density(x_hist0$mean), col = "blue", lwd = 2)
h0

num_sims5 <- 1000
##A vector to store results
results4 <- rep(NA, num_sims5)
##A loop for completing the simulations
for(i in 1:num_sims5){
  mean_female <- mean(sample(x = data.US$suicides.100k.pop[data.US$sex=="female"],
size = n_f, replace = TRUE))
```

```
  mean_male <- mean(sample(x = data.US$suicides.100k.pop[data.US$sex=="male"], size =
n_m, replace = TRUE))
  results4[i] <- mean_female - mean_male
}
##Finally plot
hist(results4, freq = FALSE, main = "Sampling Distribution of the Sample Mean",
    xlab = "Number of Suicide Difference", ylab = "Density")

##Bootstrap one-sided CI
c(quantile(results4, c(0.025, 0.975)))

##T-method
t.test(data.US$suicides.100k.pop[data.US$sex=="female"],
      data.US$suicides.100k.pop[data.US$sex=="male"])$conf.int

set.seed(111)
num_sims6 <- 1000
##A vector to store results
results_given_h0_true2 <- rep(NA, num_sims6)
##A loop for completing the simulation
for(i in 1:num_sims6){
  shuffled_groups <- transform(data.US, sex = sample(sex))
  mean_female2 <-
mean(shuffled_groups$suicides.100k.pop[shuffled_groups$sex=="female"])
  mean_male2 <- mean(shuffled_groups$suicides.100k.pop[shuffled_groups$sex=="male"])
  results_given_h0_true2[i] <- mean_female2 - mean_male2
}
##Finally plot
hist(results_given_h0_true2, freq = FALSE, main = "Dist. of the Diff in  Sample Means Under
Null", xlab  =  "Average Number of Suicide Difference under Null", ylab = "Density")
diff_in_samplemean <- mean(data.US$suicides.100k.pop[data.US$sex=="female"]) -
mean(data.US$suicides.100k.pop[data.US$sex=="male"])
diff_in_samplemean
##abline(v=diff_in_samplemean, col = "blue")
##abline(v=abs(diff_in_samplemean), col = "red")

##Count of values more extreme
##Two sided given
count_of_more_extreme_lower2 <- sum(results_given_h0_true2 <= diff_in_samplemean)
```

```
count_of_more_extreme_upper2 <- sum(results_given_h0_true2 >=
abs(diff_in_samplemean))

boostrap_p2 <- (count_of_more_extreme_lower2 +
count_of_more_extreme_upper2)/num_sims6
boostrap_p2

t.test(data.US$suicides.100k.pop[data.US$sex=="female"],
      data.US$suicides.100k.pop[data.US$sex=="male"])$p.value

### TWO SAMPLE DIFF IN PROPORTION - Traditional

##suicides in two different popultaions: Female and Male
dataUS_male <- subset(data.US, data.US$sex=="male")
dataUS_female <- subset(data.US, data.US$sex=="female")

##data2 <- aggregate(suicides.100k.pop ~ year+sex, data=data.US, FUN=sum)
##data3 <- aggregate(suicides_no~ year, data=dataUS_female, FUN=sum)
##data4 <- aggregate(population~ year, data=dataUS_female, FUN=sum)

##sample props
p_hat_f2 <- sum(dataUS_female$suicides_no)/sum(dataUS_female$population)
p_hat_m2 <- sum(dataUS_male$suicides_no)/sum(dataUS_male$population)
#null hypothesized population pro difference between the two group
p_0 <- 0
#sample size
n_f2 <- sum(dataUS_female$population)
n_m2 <- sum(dataUS_male$population)
#sample variances
den_p_f2 <- (p_hat_f2*(1-p_hat_f2))/n_f2
den_p_m2 <- (p_hat_m2*(1-p_hat_m2))/n_m2
#z-test test statistic
z <- (p_hat_m2 - p_hat_f2 - p_0)/sqrt(den_p_f2 + den_p_m2)
#two sided p-value
two_sided_diff_prop_p<- pnorm(q=z, lower.tail = FALSE)*2
two_sided_diff_prop_p

##lower bound
(p_hat_f2-p_hat_m2)+(qnorm(0.025)*sqrt(den_p_f2+den_p_m2))
```

```
##upper bound
(p_hat_f2-p_hat_m2)+(qnorm(0.975)*sqrt(den_p_f2+den_p_m2))

### TWO SAMPLE Diff in PROPORTION - Bootstrap and Randomization Approach

f1 <- sum(dataUS_female$suicides_no)
f1
m_1 <- sum(dataUS_male$suicides_no)
ff <- as.integer(sum(dataUS_female$population)/1000)
ff
mx <- as.integer(sum(dataUS_male$population)/1000)

fake_f <- rep(c(1,0), c(f1, (ff-f1)))
fake_m <- rep(c(1,0), c(m_1, (mx-m_1)))

#Make the Data
set.seed(111)
num_sims7 <- 1000
##A vector to store results
results5 <- rep(NA, num_sims7)
##A loop for completing the simulation
for(i in 1:num_sims7){
  prop_female <- mean(sample(x=fake_f, size = ff, replace = TRUE))
  prop_male <- mean(sample(x=fake_m, size = mx, replace = TRUE))
  results5[i] <- prop_female - prop_male
}
##Finally Plot
hist(results5, freq=FALSE, main = "Dist. of the Diff in Prop",
    xlab = "Difference in Prop. of Suicide", ylab = "Density")

##Boostrap
c(quantile(results5, c(0.025, 0.975)))

##Normal Approximation
c((p_hat_f2-p_hat_m2)+(qnorm(0.025)*sqrt(den_p_f2+den_p_m2)),
  (p_hat_f2-p_hat_m2)+(qnorm(0.975)*sqrt(den_p_f2+den_p_m2)))

#Make Data
df_combined <- data.frame("population" = c(fake_f, fake_m),
                          "sex" = rep(c("female", "male"), c(ff, mx)))
```

```
summary(df_combined)

mean(df_combined$population[df_combined$sex=="female"]) == (p_hat_f2*1000)

mean(df_combined$population[df_combined$sex=="male"]) == (p_hat_m2*1000)

set.seed(111)
##A vector to store results
results_given_h0_true3 <- rep(NA, num_sims7)
##A loop for completing the simulation
for(i in 1:num_sims7){
  shuffled_groups2 <- transform(df_combined, sex=sample(sex))
  prop_f2 <- mean(shuffled_groups2$population[shuffled_groups2$sex=="female"])
  prop_m2 <- mean(shuffled_groups2$population[shuffled_groups2$sex=="male"])
  results_given_h0_true3[i] <- prop_f2 - prop_m2
}
##Finally plot
hist(results_given_h0_true3, freq = FALSE, main = "Dist. of the Diff in Sample Sample Props
Under Null",
     xlab = "Average Difference in prop. Suicide under Null", ylab = "Density")

##Count of values more extreme than the test statistic
#two sided given the alternate hypothesis
diff_in_sampleprop <- p_hat_f2 - p_hat_m2

count_of_more_extreme_lower3 <- sum(results_given_h0_true3/1000 <=
diff_in_sampleprop)
count_of_more_extreme_upper3 <- sum(results_given_h0_true3/1000 >= -
diff_in_sampleprop)

bootsrtap_p3 <-
(count_of_more_extreme_lower3+count_of_more_extreme_upper3)/num_sims7

### Chi-Square Goodnees of Fit Test

##Null Hypothesis: p_r1 = p_r2 = p_r3 = p_r4 = p_r5 = p_r6 = 1/6
##The proportion fo each age group is the same and is equal to 1/6
##Alternate Hypothesis: some p_i != 1/6
##At least one of the proportions is not equal to 1/6
```

```
##chi-square with 6 different age ranges
totalR1 <- sum(data.US$suicides_no[data.US$age=="5-14 years"])
totalR2 <- sum(data.US$suicides_no[data.US$age=="15-24 years"])
totalR3 <- sum(data.US$suicides_no[data.US$age=="25-34 years"])
totalR4 <- sum(data.US$suicides_no[data.US$age=="35-54 years"])
totalR5 <- sum(data.US$suicides_no[data.US$age=="55-74 years"])
totalR6 <- sum(data.US$suicides_no[data.US$age=="75+ years"])
totalR <- sum(data.US$suicides_no)
avgR <- totalR/6

suicide_age <- matrix(c(totalR1, totalR2, totalR3, totalR4, totalR5, totalR6), ncol=6,
byrow=TRUE)
colnames(suicide_age) <- c("5-14 years", "15-24 years", "25-34 years", "35-54 years",
                           "55-74 years", "75+ years")
suicide_age

prop.table(suicide_age)

##test statistics
chi <- (((totalR1-avgR)^2)/avgR)+(((totalR2-avgR)^2)/avgR)+(((totalR3-avgR)^2)/avgR)+
  (((totalR4-avgR)^2)/avgR)+(((totalR5-avgR)^2)/avgR)+(((totalR6-avgR)^2)/avgR)

##p-value
pchisq(chi, df=6-1, lower.tail=FALSE)

avgR
totalR/6

sol_under_H0 <- rep(c("5-14 years", "15-24 years", "25-34 years", "35-54 years",
                      "55-74 years", "75+ years"), avgR)
table(sol_under_H0)

set.seed(111)
num_sims <- 1000
##A vector to store my results
chisq_stats_under_H0 <- rep(NA, num_sims)
##A loop for completing the simulation
for(i in 1:num_sims){
  new_sample <- sample(sol_under_H0, totalR, replace=T)
  chisq_stats_under_H0[i] <- sum(((table(new_sample)-avgR)^2)/avgR)
```

```
}

hist(chisq_stats_under_H0, freq=FALSE, main = "Dist. of the Chi-Square Statistic Under
Null",
    xlab = "Chi-Square stat under Null", ylab = "Density")
abline(v=sum(((suicide_age - avgR)^2)/avgR), col="red")

sum(chisq_stats_under_H0 >= sum(((suicide_age)-avgR)^2)/avgR)/ num_sims

## Chi-square association between Two Categorical Variable


##Null: The number of suicide is not associated with gender
##Alternate: The number of suicide is associated with gender

male_R1 <- sum(dataUS_male$suicides_no[dataUS_male$age=="5-14 years"])
male_R2 <- sum(dataUS_male$suicides_no[dataUS_male$age=="15-24 years"])
male_R3 <- sum(dataUS_male$suicides_no[dataUS_male$age=="25-34 years"])
male_R4 <- sum(dataUS_male$suicides_no[dataUS_male$age=="35-54 years"])
male_R5 <- sum(dataUS_male$suicides_no[dataUS_male$age=="55-74 years"])
male_R6 <- sum(dataUS_male$suicides_no[dataUS_male$age=="75+ years"])
female_R1 <- sum(dataUS_female$suicides_no[dataUS_female$age=="5-14 years"])
female_R2 <- sum(dataUS_female$suicides_no[dataUS_female$age=="15-24 years"])
female_R3 <- sum(dataUS_female$suicides_no[dataUS_female$age=="25-34 years"])
female_R4 <- sum(dataUS_female$suicides_no[dataUS_female$age=="35-54 years"])
female_R5 <- sum(dataUS_female$suicides_no[dataUS_female$age=="55-74 years"])
female_R6 <- sum(dataUS_female$suicides_no[dataUS_female$age=="75+ years"])
f_m_age <- matrix(c(female_R1, male_R1, female_R2, male_R2, female_R3, male_R3,
female_R4, male_R4, female_R5, male_R5, female_R6, male_R6), ncol=2, byrow=TRUE)
colnames(f_m_age) <- c("Female", "Male")
rownames(f_m_age) <- c("5-14 years", "15-24 years", "25-34 years", "35-54 years", "55-74
years", "75+ years")
f_m_age

ad <- addmargins(f_m_age)
ad

addmargins((prop.table(f_m_age)))

n2 <- addmargins(f_m_age)[7,3]
```

```
ex_r1_m <- (ad[1,3]*ad[7,2]/n2)
ex_r1_f <- (ad[1,3]*ad[7,1]/n2)
ex_r2_m <- (ad[2,3]*ad[7,2]/n2)
ex_r2_f <- (ad[2,3]*ad[7,1]/n2)
ex_r3_m <- (ad[3,3]*ad[7,2]/n2)
ex_r3_f <- (ad[3,3]*ad[7,1]/n2)
ex_r4_m <- (ad[4,3]*ad[7,2]/n2)
ex_r4_f <- (ad[4,3]*ad[7,1]/n2)
ex_r5_m <- (ad[5,3]*ad[7,2]/n2)
ex_r5_f <- (ad[5,3]*ad[7,1]/n2)
ex_r6_m <- (ad[6,3]*ad[7,2]/n2)
ex_r6_f <- (ad[6,3]*ad[7,1]/n2)

chi_sq_stat <- sum(
  ((ad[1,1]-ex_r1_f)^2)/ex_r1_f,
  ((ad[2,1]-ex_r2_f)^2)/ex_r2_f,
  ((ad[3,1]-ex_r3_f)^2)/ex_r3_f,
  ((ad[4,1]-ex_r4_f)^2)/ex_r4_f,
  ((ad[5,1]-ex_r5_f)^2)/ex_r5_f,
  ((ad[6,1]-ex_r6_f)^2)/ex_r6_f,
  ((ad[1,2]-ex_r1_m)^2)/ex_r1_m,
  ((ad[2,2]-ex_r2_m)^2)/ex_r2_m,
  ((ad[3,2]-ex_r3_m)^2)/ex_r3_m,
  ((ad[4,2]-ex_r4_m)^2)/ex_r4_m,
  ((ad[5,2]-ex_r5_m)^2)/ex_r5_m,
  ((ad[6,2]-ex_r6_m)^2)/ex_r6_m
)
chi_sq_stat

pchisq(chi_sq_stat, df = (6-1)*(2-1), lower.tail = FALSE)
```

Resource

Rusty. “Suicide Rates Overview 1985 to 2016.” *Kaggle*, Rusty, 1 Dec. 2018, https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016.

Suicide Statistics. (2019, April 16). Retrieved from https://afsp.org/about-suicide/suicide-statistics/.