Suicide Rates Overview - 1985 to 2016

Aakash Atnoorkar, Sanghamitra Shanmugam, Danish Bhat

November 26, 2019

Overview

To analyse the suicides rates overview from 1985 to 2016 using the dataset found in kaggle by finding meaningful insights using inferential statistics along with descriptive statistics. Our main objective is to determine whether the gdp of the country is a strong factor in determining the suicide rates of that country, whether any particular gender is more prone to committing suicide also to determine the pattern of suicide rates across all countries over the years.

Data Description

The raw data is obtained from Kaggle(https://www.kaggle.com/). The data is titled "Suicide Rates Overview 1985 to 2016" which compares socio-economic info with suicide rates by year and country. The data is collected from 30 countries from 1985 to 2016 in which there are 8 columns which can be analyzed to find information and correlation to suicide rates among different cohorts globally, across the socio-economic spectrum. The columns are country, year, sex, age, number of suicides, population, suicides/100k pop, country-year, HDI for year, gdp for year($), gdp per capita (in dollars), and generation. The data comprises of 30 countries from across different continents of the world and includes all age groups from 5-75+ years.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
library(fitdistrplus)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: survival

## Loading required package: npsurv

## Loading required package: lsei
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(devtools)
```

```
## Loading required package: usethis
```

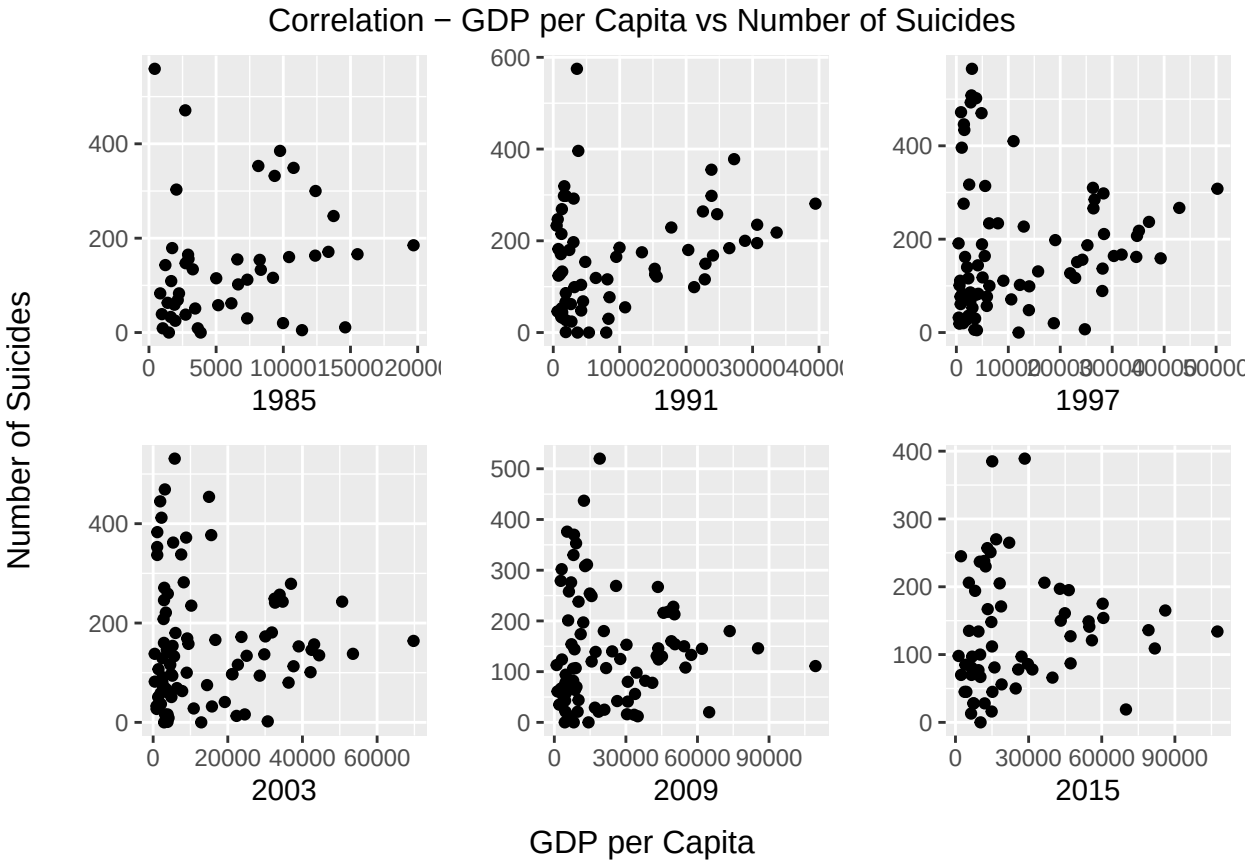
Correlation - GDP per Capita vs Number of Suicides

From the data available, we are trying to find out the correlation between the GDP (per Capita) of the country and number of suicides. As GDP per Capita is measure of a country's economic output, this plot can provide us with meaningful information whether the factors such as employment, overall economical condition of the country determine the suicides happening in the country.

GDP per Capita is considered instead of only GDP because it tells us how prosperous a country feels to each of its citizens. Similarly, as the population is different for each country, instead of total number of suicides it is better to consider suides per 100k of population.

```
#Importing the dataset
suicides <- read.csv("~/IE 6200 - Prob Stats/Final Project/Suicide Rates Overview 1985-2016.csv", na.st
#Cleaning the dataset
names(suicides)[names(suicides) == "ï..Country"] <- "Country"


for(i in seq(from = 1985, to = 2015, by = 6)) {
 nam <- paste("gdpAndSuicides_", i, sep = "")
 temp_df <- suicides %>%
  dplyr::select(Year, Country, gdp_per_capita...., suicides.100k.pop) %>%
  filter(Year == i) %>%
  group_by(Country, gdp_per_capita....) %>%
  summarise(Total_suicides_per100K = round(sum(suicides.100k.pop, na.rm = T)))
 assign(nam, temp_df)
}

par(mfrow=c(3,2))
#Plot all correlation plots below
#Correlation plot for year 1985
plot_1985 <- ggplot(data = gdpAndSuicides_1985, mapping = aes(x = gdp_per_capita...., y = Total_suicides
  labs(x = "1985", y = "")
```

```r
#Correlation plot for year 1991
plot_1991 <- ggplot(data = gdpAndSuicides_1991, mapping = aes(x = gdp_per_capita...., y = Total_suicides
  labs(x = "1991", y = "")

#Correlation plot for year 1997
plot_1997 <- ggplot(data = gdpAndSuicides_1997, mapping = aes(x = gdp_per_capita...., y = Total_suicides
  labs(x = "1997", y = "")

#Correlation plot for year 2003
plot_2003 <- ggplot(data = gdpAndSuicides_2003, mapping = aes(x = gdp_per_capita...., y = Total_suicides
  labs(x = "2003", y = "")

#Correlation plot for year 2009
plot_2009 <- ggplot(data = gdpAndSuicides_2009, mapping = aes(x = gdp_per_capita...., y = Total_suicides
  labs(x = "2009", y = "")

#Correlation plot for year 2015
plot_2015 <- ggplot(data = gdpAndSuicides_2015, mapping = aes(x = gdp_per_capita...., y = Total_suicides
  labs(x = "2015", y = "")

grid.arrange(plot_1985,plot_1991,plot_1997, plot_2003, plot_2009, plot_2015, nrow = 2, ncol = 3, top="C
```



The correlation between GDP (per Capita) & number of suicides have been plotted for the years 1985, 1991, 1997, 2003, 2009, 2015. The reason behind choosing these years is that the pattern is almost similar during other years and putting a gap of 6 years between the years provides good visuals and progress. We can see

from the above correlation plots that,

1. Most of the countries has low GDP(per Capita) and the number gows down as the GDP(per Capita) value increases
2. We can also find from the plots that the country with maximum number of suicides committed falls in low GDP(per Capita) area
3. As the GDP(per Capita) increases the number slowly decreases but remains constant at some point. Even the country with maximum GDP(per Capita), the number of suicides have been around the average number of suicides.
4. Though we cannot say that economic situation of a country can be a factor towards suicides, there can be multiple other components which contribute to the number of suicides in a country.

Hypothesis to test the sample countries have number of suicides less than the mean number of suicides

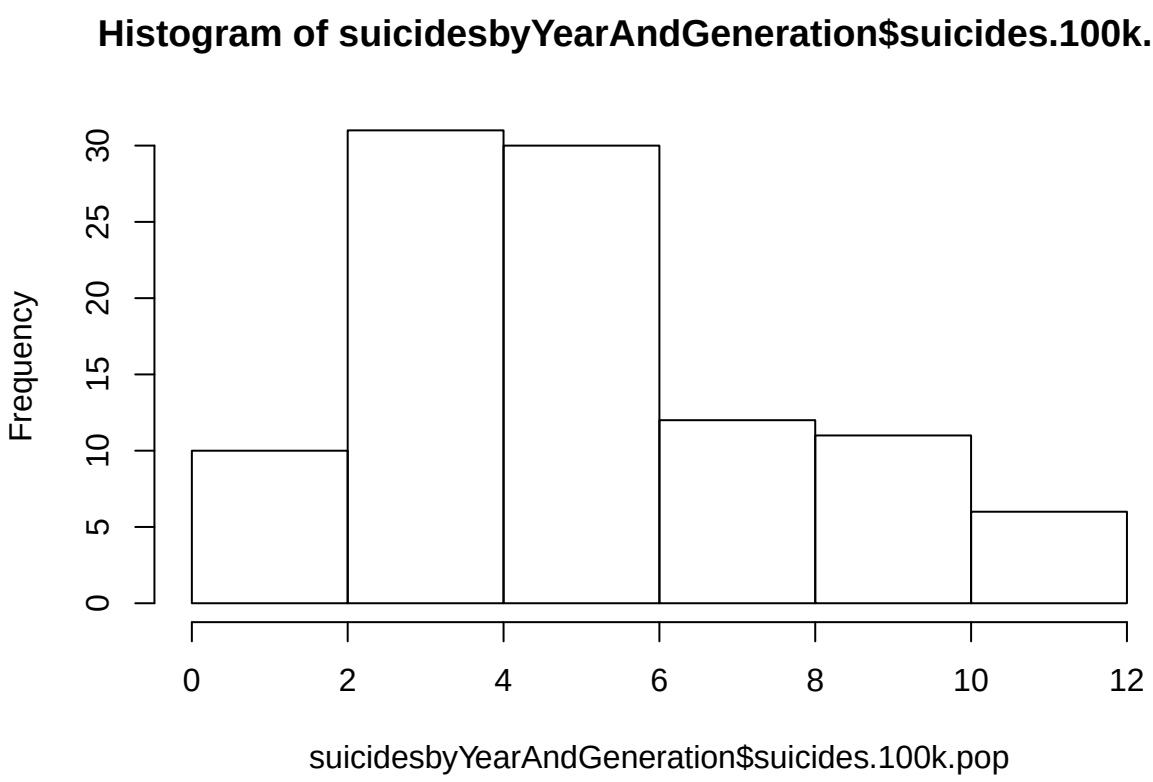
Based on the data, a fairly normal data has been obtained by filtering out the dataset. For this hypothesis, suicides from the year 2007 & committed by Generation X people have been considered. The people who have age between 15 to 24 are considered as Generation X.

Hypothesis **H0:** Mean number of suicides are equal to population mean **H1:** Mean number of suicides are **less than** population mean

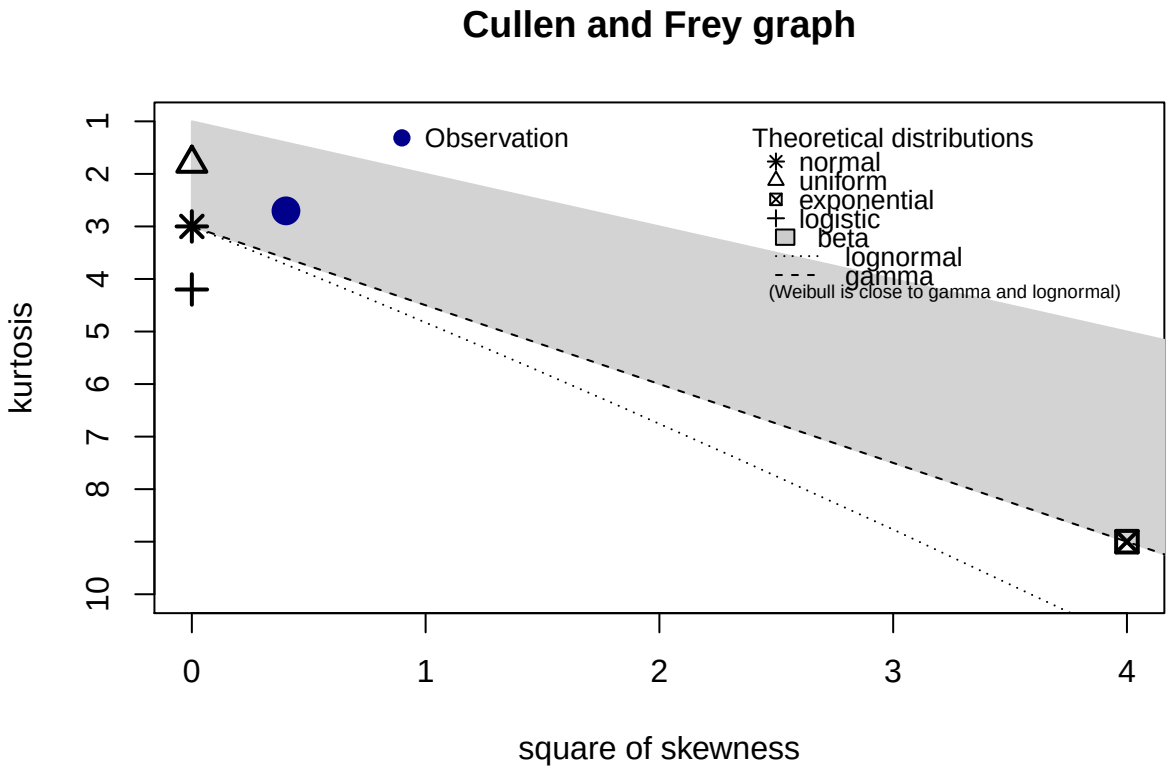
We are considering a random sample from the population with sample size as 15. As the sample size is less than 30, we will perform t test for this hypothesis.

We want to know if the mean number of suicides committed are less than the actual mean as the lesser the suicides the better it is. Hence, it will be a one-tailed test (lower tailed).

```
suicides <- read.csv("~/IE 6200 - Prob Stats/Final Project/Suicide Rates Overview 1985-2016.csv", na.st
#Cleaning the dataset
names(suicides)[names(suicides) == "ï..Country"] <- "Country"
#We got 100 observations as population
suicidesbyYearAndGeneration <- suicides %>%
  filter(Year == 2007 & generation == "Generation X") %>%
  filter(suicides.100k.pop > 0 & suicides.100k.pop < 12)

#Checking the normality of the data
hist(suicidesbyYearAndGeneration$suicides.100k.pop)
```




```
descdist(suicidesbyYearAndGeneration$suicides.100k.pop)
```




```
## summary statistics
## ------
## min:  0.29   max:  11.95
## median:  4.6
## mean:  4.9416
## estimated sd:  2.828801
## estimated skewness:  0.634541
## estimated kurtosis:  2.70382
```

```r
sampleSuicides <- suicidesbyYearAndGeneration[sample(nrow(suicidesbyYearAndGeneration),20),]

#Mean number of suicides for year and Generation X
mu <- mean(suicidesbyYearAndGeneration$suicides.100k.pop)
xBar <- mean(sampleSuicides$suicides.100k.pop)

#H0: Mean number of suicides are equal to population mean
#H1: Mean number is less than population mean

Tcalc <- t.test(sampleSuicides$suicides.100k.pop, mu=mu, alternative = "less", conf.level = 0.95)
Tcalc
```

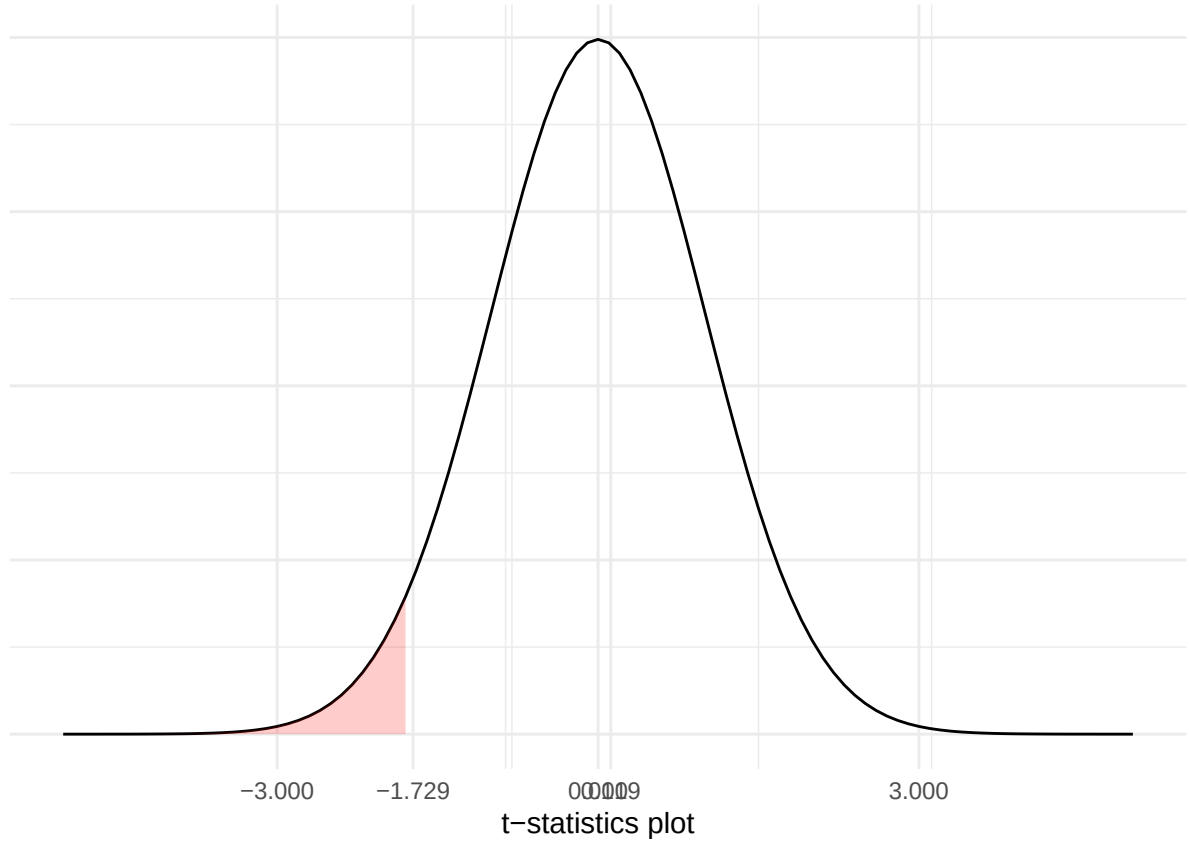
```
##
##  One Sample t-test
##
## data:  sampleSuicides$suicides.100k.pop
```

```
## t = 0.11877, df = 19, p-value = 0.5466
## alternative hypothesis: true mean is less than 4.9416
## 95 percent confidence interval:
##      -Inf 6.293607
## sample estimates:
## mean of x
##    5.0285
```

```r
sd(sampleSuicides$suicides.100k.pop)
```

```
## [1] 3.272004
```

```r
#get t-value for 5 percent error
tvalue <- round(qt(c(.05), df=19), 3)

m = 0
std = 1

funcShaded <- function(x, lower_bound) {
  y = dnorm(x, mean = m, sd = std)
  y[x > lower_bound] <- NA
  return(y)
}

ggplot(data.frame(x = c(m - (5*std), m + (5*std))), aes(x = x)) +
  stat_function(fun = dnorm, args = list(mean = m, sd = std)) +
  stat_function(fun = funcShaded, args = list(lower_bound = tvalue), geom = "area", fill = "red", alpha
  scale_x_continuous(breaks = c(m - (3*std), m + (3*std), 0, tvalue, round(as.numeric(Tcalc["statistic"]
  theme_minimal()+ theme(axis.text.y = element_blank()) + labs(x = "t-statistics plot", y = "")
```




```
##Thus we fail to reject the null
```

After calculation of t-value and plotting for the 5% error, we get the information that the t-calc does not fall in the rejection region and hence we **failed to reject** the hypothesis and say that the average number of suicides happened are actually equal to mean number of suicides.

Hypothesis to test the difference in mean number of suicides between Generation X & Boomers

For the year 2015, we are finding out whether the difference in the mean number of suicides for 15-24 & 25-34 is same or not.

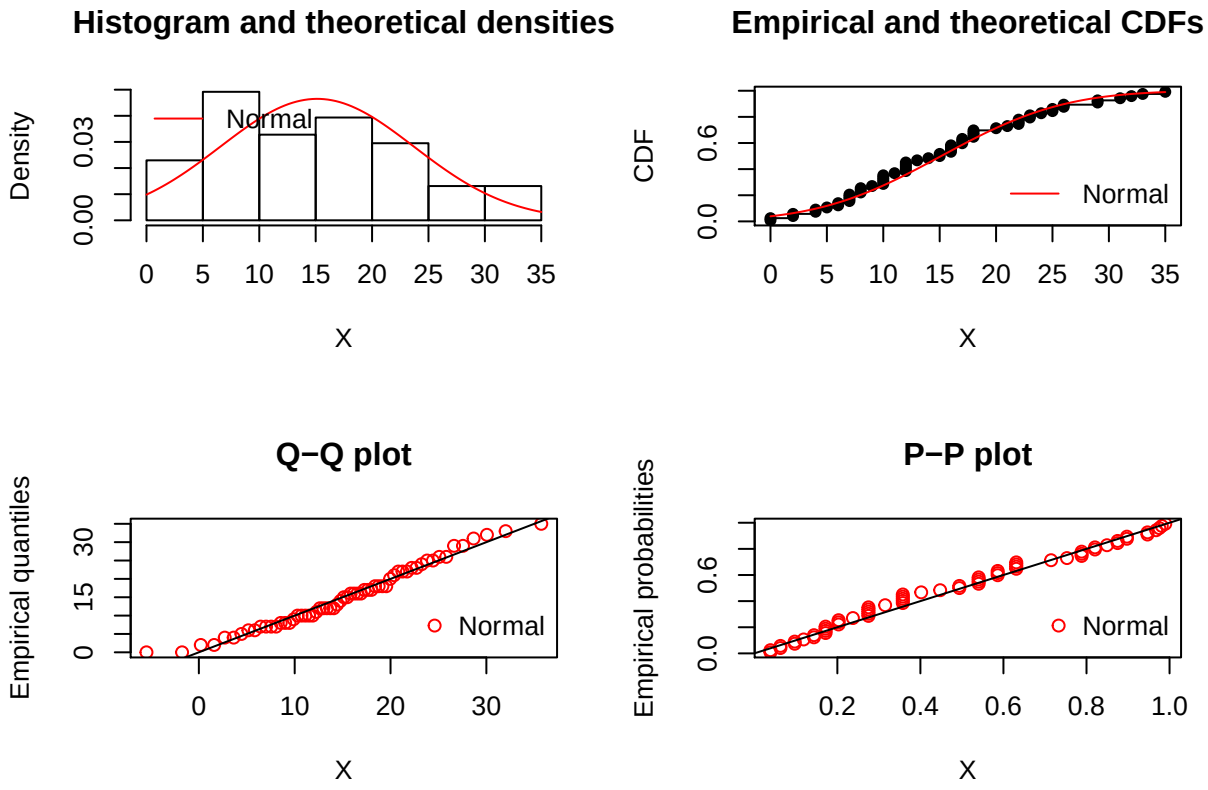
We are considering the data from the year 2015 for Generation X & Boomers. By taking sample size of 15 from Generation X data & 17 from Boomers, we are plotting.

```
suicides <- read.csv("~/IE 6200 - Prob Stats/Final Project/Suicide Rates Overview 1985-2016.csv", na.st
#Cleaning the dataset
names(suicides)[names(suicides) == "ï..Country"] <- "Country"
#Difference between the mean of no of suicides for 15-24 & 25-34
#H0: is that its same
#h1: is not same

#Getting data for Generation X
worldSuicides2015_GenerationX <- suicides %>%
  filter(Year == 2015, Age %in% c("15-24 years")) %>%
```

```r
  group_by(Country,Age) %>%
  summarise(Total = round(sum(suicides.100k.pop))) %>%
  filter(Total < 40)

#Checking normality of data for Generation X
worldSuicides2015_GenerationX_norm <- fitdist(worldSuicides2015_GenerationX$Total, distr = "norm")
summary(worldSuicides2015_GenerationX_norm)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 15.131148  1.0988439
## sd    8.582245  0.7769999
## Loglikelihood:  -217.6867   AIC:  439.3734   BIC:  443.5951
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

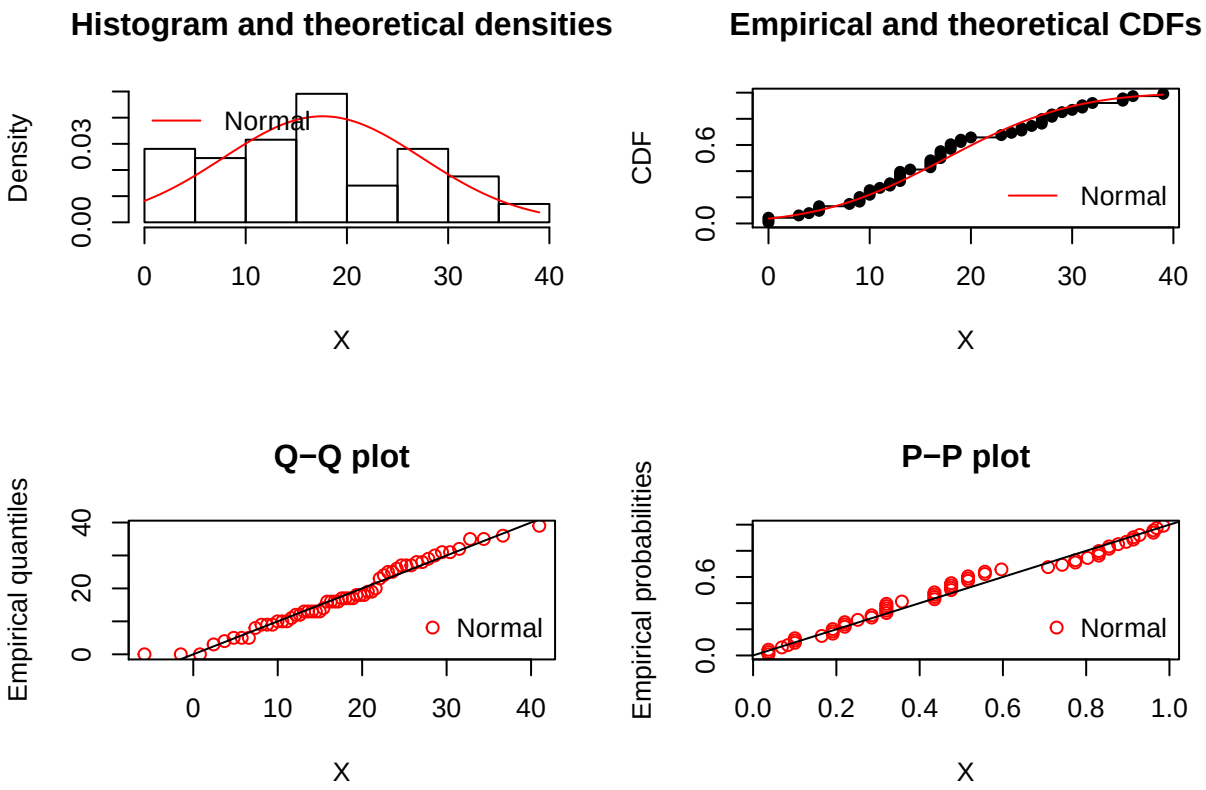
```r
par(mfrow=c(2,2))
plot.legend <- c("Normal")
denscomp(list(worldSuicides2015_GenerationX_norm), legendtext = plot.legend, xlab = 'X', xlegend = 'top
cdfcomp (list(worldSuicides2015_GenerationX_norm), legendtext = plot.legend, xlab = 'X')
qqcomp (list(worldSuicides2015_GenerationX_norm), legendtext = plot.legend, xlab = 'X')
ppcomp (list(worldSuicides2015_GenerationX_norm), legendtext = plot.legend, xlab = 'X')
```



```r
#Getting data for boomers
worldSuicides2015_Boomers <- suicides %>%
  filter(Year == 2015, Age %in% c("25-34 years")) %>%
  group_by(Country,Age) %>%
  summarise(Total = round(sum(suicides.100k.pop))) %>%
  filter(Total < 40)

#Checking normality of data for Boomers
worldSuicides2015_Boomers_norm <- fitdist(worldSuicides2015_Boomers$Total, distr = "norm")
summary(worldSuicides2015_Boomers_norm)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 17.596491  1.3036539
## sd    9.842371  0.9218225
## Loglikelihood:  -211.2212   AIC:  426.4424   BIC:  430.5285
## Correlation matrix:
##              mean           sd
## mean 1.000000e+00 4.269429e-09
## sd   4.269429e-09 1.000000e+00
```

```r
par(mfrow=c(2,2))
plot.legend <- c("Normal")
denscomp(list(worldSuicides2015_Boomers_norm), legendtext = plot.legend, xlab = 'X', xlegend = 'topleft
cdfcomp (list(worldSuicides2015_Boomers_norm), legendtext = plot.legend, xlab = 'X')
qqcomp (list(worldSuicides2015_Boomers_norm), legendtext = plot.legend, xlab = 'X')
ppcomp (list(worldSuicides2015_Boomers_norm), legendtext = plot.legend, xlab = 'X')
```

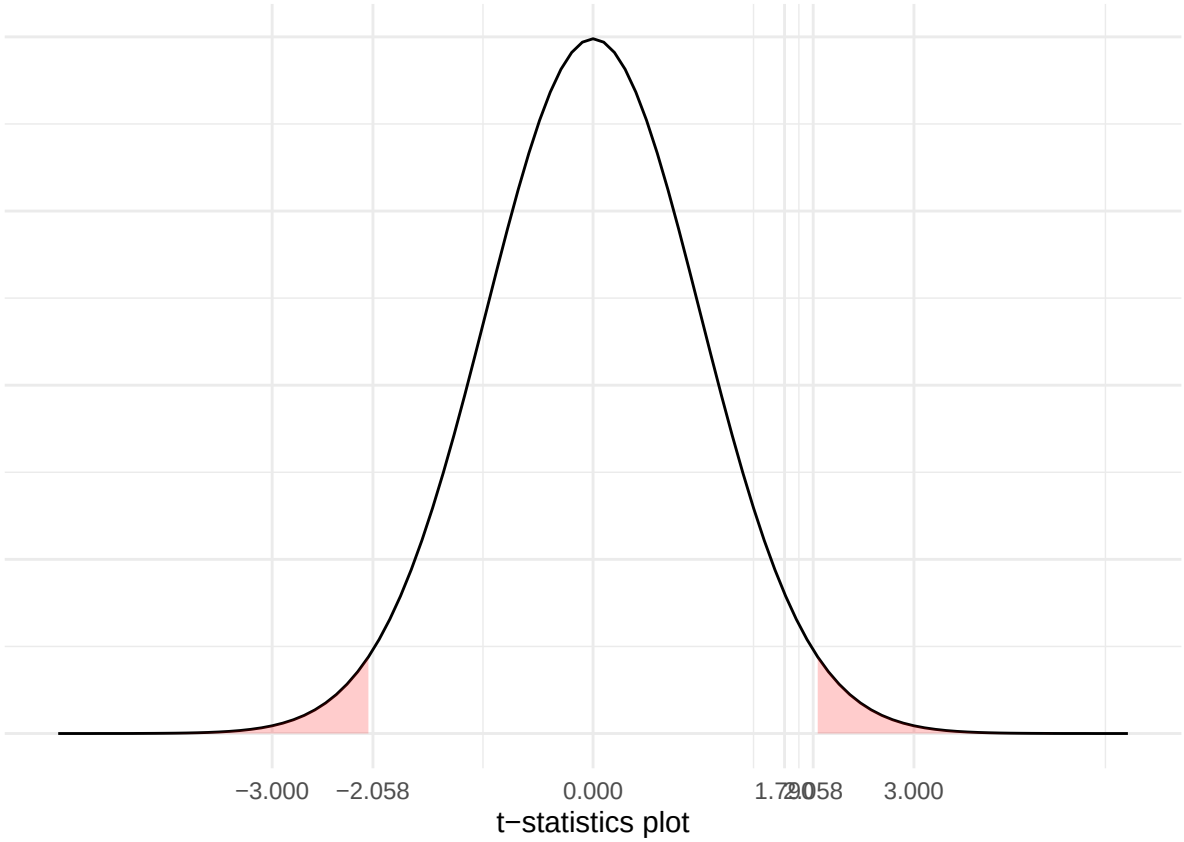



```
n1 <- 15
xbar1 <- mean(worldSuicides2015_GenerationX$Total)
sample_n1 <- worldSuicides2015_GenerationX[sample(nrow(worldSuicides2015_GenerationX),n1),]

n2 <- 17
xbar2 <- mean(worldSuicides2015_Boomers$Total)
sample_n2 <- worldSuicides2015_Boomers[sample(nrow(worldSuicides2015_Boomers),n2),]

Tcalc <- t.test(y = sample_n1$Total,x  = sample_n2$Total)
Tcalc
```

```
##
##  Welch Two Sample t-test
##
## data:  sample_n2$Total and sample_n1$Total
## t = 1.79, df = 25.4, p-value = 0.08538
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.874483 12.560757
## sample estimates:
## mean of x mean of y
##  19.17647  13.33333
```

```
tvalue <- round(qt(c(.025), df=as.numeric(Tcalc["parameter"])),3)
```

```
m = 0
std = 1

funcShaded <- function(x, lower_bound, upper_bound) {
  y = dnorm(x, mean = m, sd = std)
  y[x > lower_bound & x < upper_bound] <- NA
  return(y)
}

ggplot(data.frame(x = c(m - (5*std), m + (5*std))), aes(x = x)) +
  stat_function(fun = dnorm, args = list(mean = m, sd = std)) +
  stat_function(fun = funcShaded, args = list(lower_bound = tvalue, upper_bound = -tvalue), geom = "area
  scale_x_continuous(breaks = c(m - (3*std), m + (3*std), 0, tvalue, -tvalue, round(as.numeric(Tcalc["st
  theme_minimal()+ theme(axis.text.y = element_blank()) + labs(x = "t-statistics plot", y = "")
```



As seen from the t-statistic curve, the T calc falls in the acceptance region and thus we say that we **failed to reject** the hypothesis and the mean number of suicides between Generation X and Boomers are essentially the same.

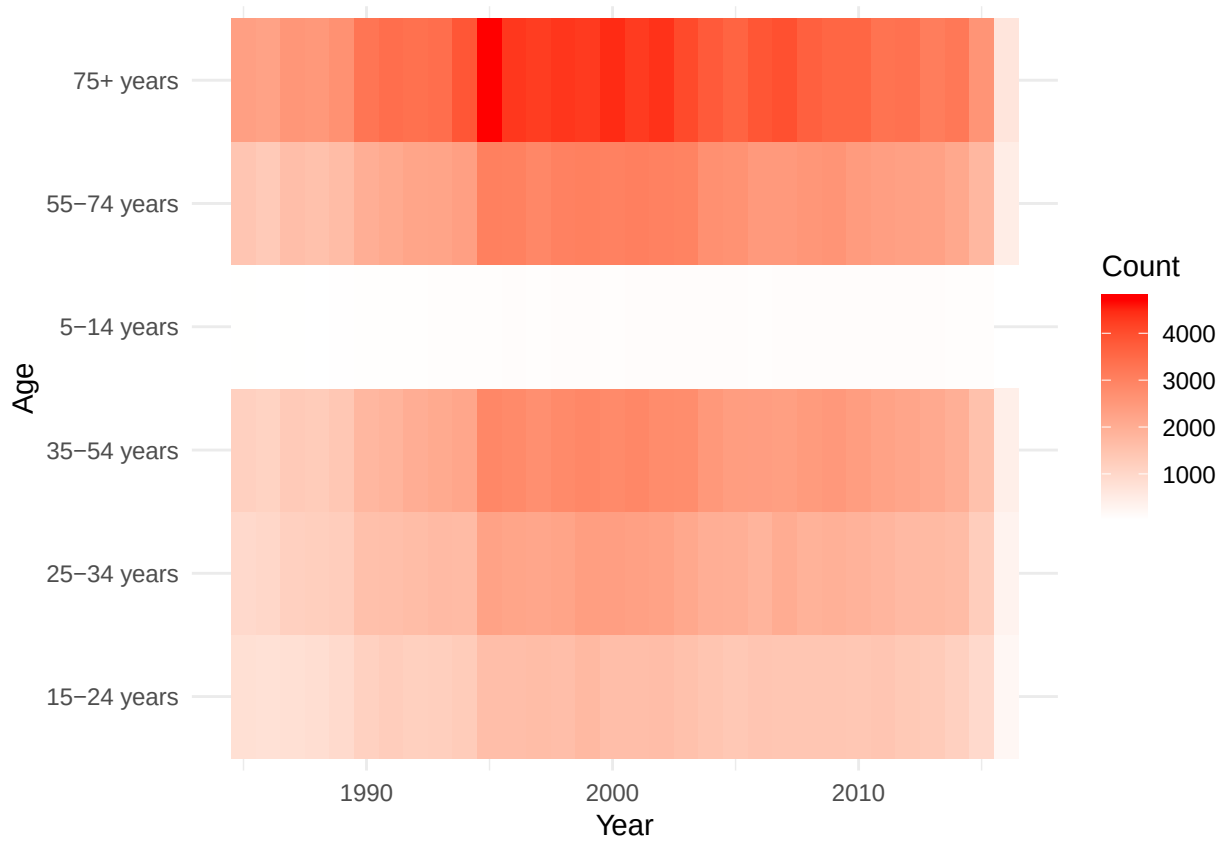
Joint Probability for top 20 Countries over the years

```
library(reshape2)
```

```
##
```

```
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

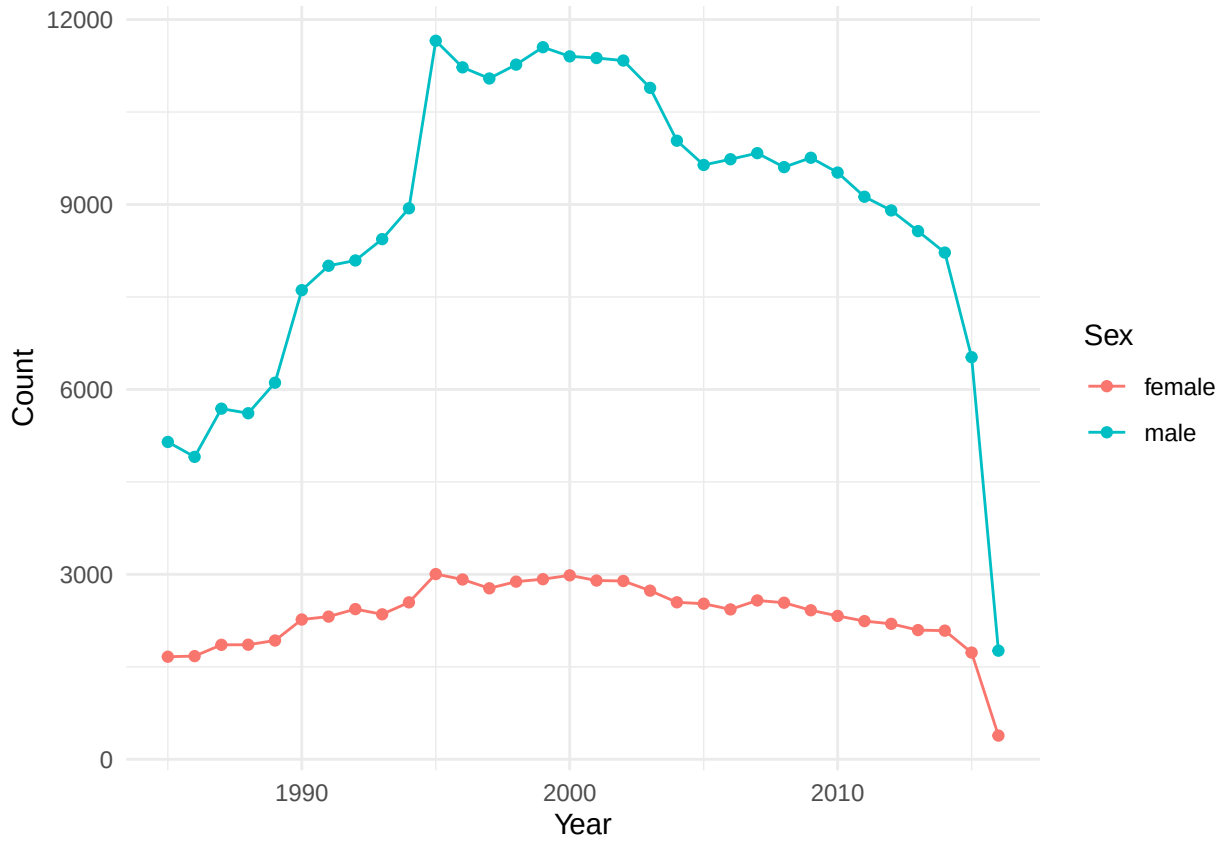
```
suicides <- read.csv("~/IE 6200 - Prob Stats/Final Project/Suicide Rates Overview 1985-2016.csv", na.st
#Cleaning the dataset
names(suicides)[names(suicides) == "ï..Country"] <- "Country"


suicidesByCountry.Age <- suicides %>%
  group_by(Year,Age) %>%
  summarise(Count = round(sum(suicides.100k.pop, na.rm = T)))

recasted.suicides <- reshape2::recast(suicidesByCountry.Age, formula = suicidesByCountry.Age$Year ~ sui
rownames(recasted.suicides) <- recasted.suicides$`suicidesByCountry.Age$Year`
recasted.suicides <- subset(recasted.suicides, select = -c(1))

joint_prob <- round(recasted.suicides/sum(recasted.suicides),3)


ggplot(suicidesByCountry.Age, mapping = aes(x = Year, y = Age, fill = Count)) + geom_tile(stat="identity
  theme_minimal()
```



Suicide rate for Male & Female over the years

We are interested in finding the pattern and number of suicides committed by male & Female over the years. We are taking help of a time-seris plot where separate lines will represent the suicide rate pattern through the years.

```
suicides <- read.csv("~/IE 6200 - Prob Stats/Final Project/Suicide Rates Overview 1985-2016.csv", na.st
#Cleaning the dataset
names(suicides)[names(suicides) == "Ã¯..Country"] <- "Country"
suicidesBy.Year.Sex <- suicides %>%
  group_by(Year, Sex) %>%
  filter(Year != 2019) %>%
  summarise(Count = round(sum(suicides.100k.pop)))

ggplot(suicidesBy.Year.Sex, mapping = aes(x = Year, y = Count, color = Sex)) + geom_line() + geom_point
```



Rates of suicide per 100 000 population are shown in the figure.

Trends: 1. There is a significant gap in number of suicides committed by men & women across the years. 2. Number of suicides went on rising from 1985 to 1990 and then a sudden rise continued till in the year 1995 which then remained constant till 2003-2004. 3. The numbers have been gradually decreasing since 2005 and we see a steep curve near the years 2014-15

Conclusion: The recent trend shows a decline in number of suicides, so the focus should be towards identifying the contributing factors in the current approach towards depression diagnosis and treatment, suicide prevention, and postsuicide attemptâ “care most responsible for this decline. Our aim is to stimulate a fruitful conversation towards focusing and investigations of the larger social, contextual, policy, and treatment trends.

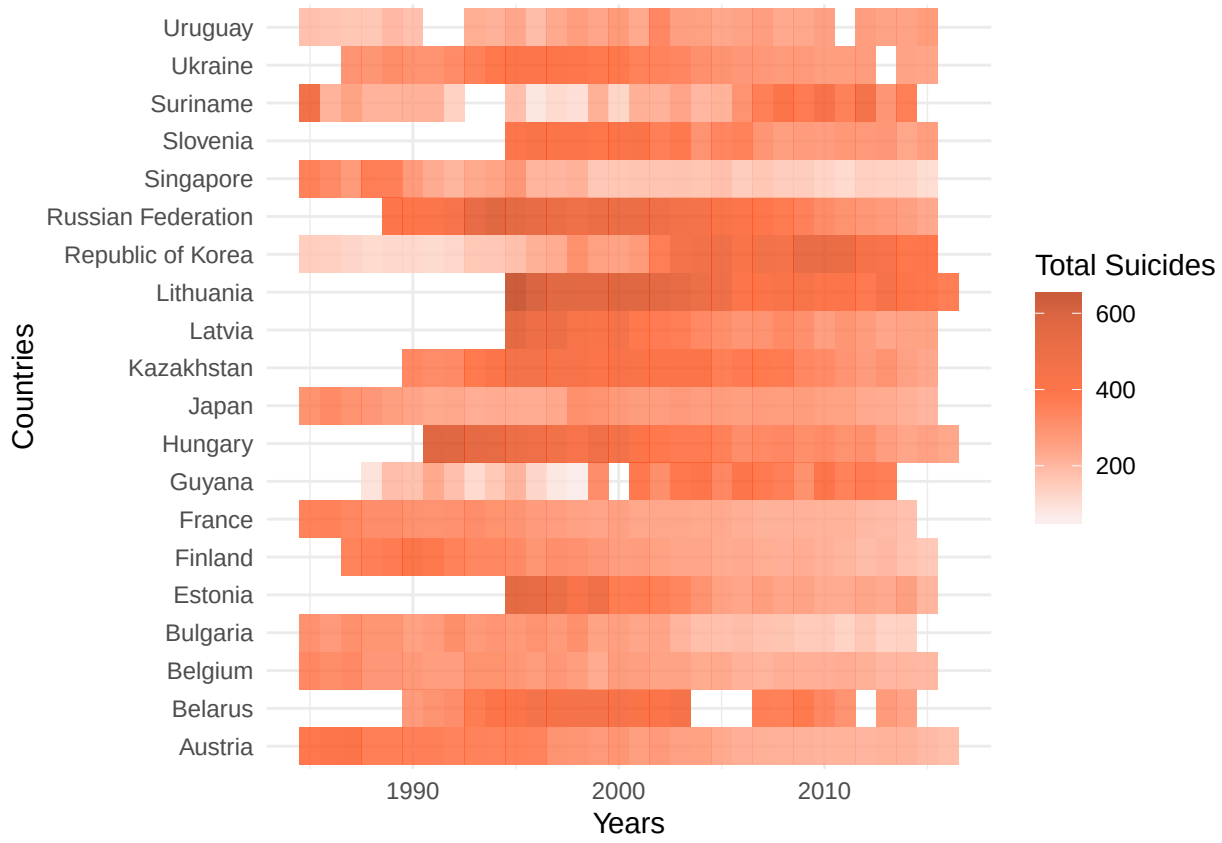
Top 20 Countries with most Suicides By Year

In this graph, we want to know how the suicides have happened over the years. For this, we have selected top 20 countries where the maximum number of suicides happened and plotted map where the dark colour represents more number of suicides.

```
suicides <- read.csv("~/IE 6200 - Prob Stats/Final Project/Suicide Rates Overview 1985-2016.csv", na.st
#Cleaning the dataset
names(suicides)[names(suicides) == "ï..Country"] <- "Country"
library("ggsci")
TopCountrySuicides <- suicides %>%
  group_by(Country) %>%
  summarise(Total_suicides = round(sum(suicides.100k.pop))) %>%
  top_n(20) %>%
  arrange(desc(Total_suicides))
```

```
## Selecting by Total_suicides
```

```
CountriesYearSuicides <- suicides %>%
  filter(Country %in% TopCountrySuicides$Country) %>%
  group_by(Country, Year) %>%
  summarise(Total_suicides = round(sum(suicides.100k.pop)))

library(ggplot2)
ggplot(CountriesYearSuicides, mapping = aes(y = CountriesYearSuicides$Country, x = CountriesYearSuicide
```



Observations: 1. Over the years, all the countries seem to have a reduced suicide rates. 2. Some countries such as Hungary, Lithuania and Russian Federation has maximum number of suicides and these also show a negative trend in suicides, which is a positive sign. 3. Some countries such as Guyanaas well as Republic of Korea show increase in number of suicides.

Hypothesis testing for the Confidence interval for the mean number of suicides committed by males in the year 1988 with 95% confidence

Considering the data of the number of suicides all over the world in the year 1988, we can make the observations that the number of suicides committed by males across the world in the year 1988 follows a normal distribution.

The qq plot,pp plot, Cullen and frey graph together clearly hold up that the male number of suicides follows a normal distribution.

Here, we are trying to find the confidence interval for the mean number of suicides committed by men in that year.

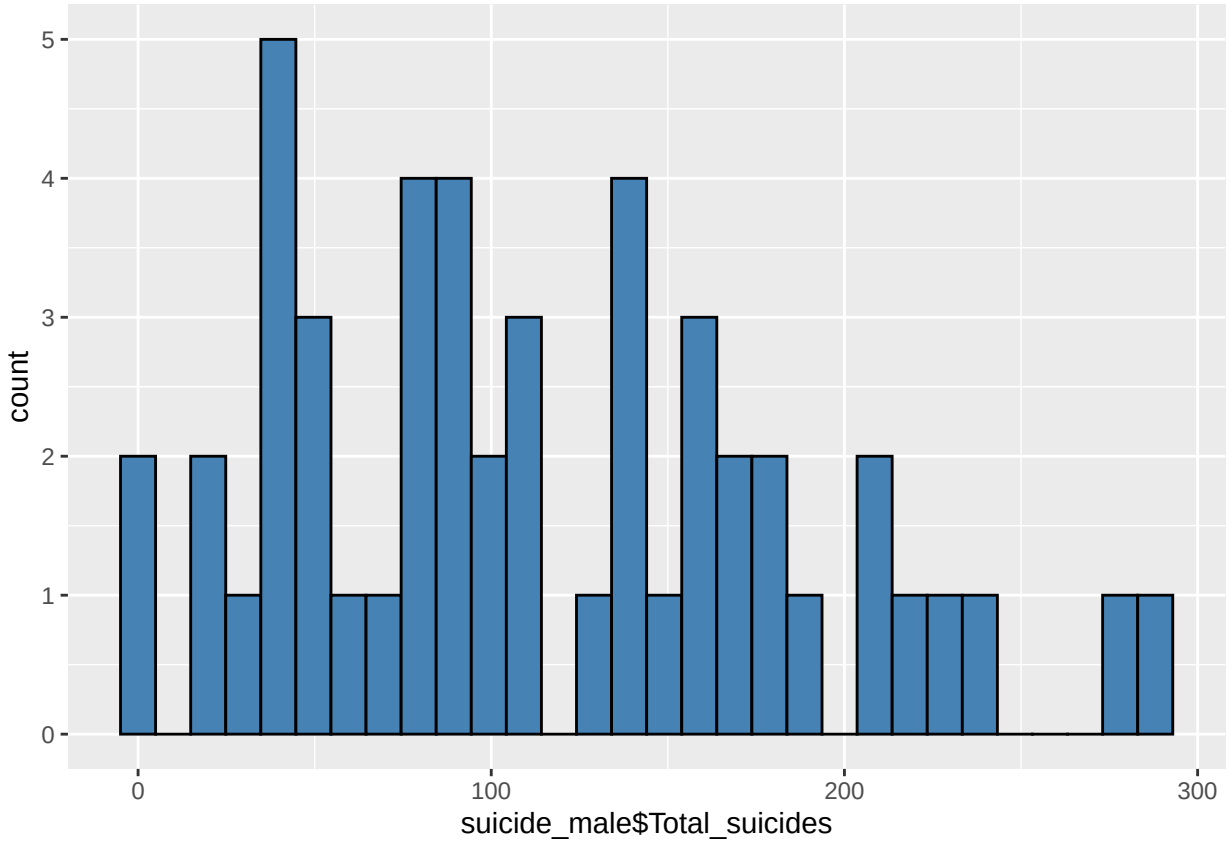
Stating the Hypothesis below :

Null Hypothesis **Ho : Sample mean = population mean** Alternate Hypothesis **H1 : Sample mean != population mean**

To conduct the test, we use t.test() function, and for one sample, the inputs given are the vectors for the number of suicdes using the sample() function. The outputs returned are the t-value, p-value, alternative hypothesis statement, 95% confidence interval value and the mean of the sample vectors.

```
#Importing the dataset
suicides <- read.csv("~/IE 6200 - Prob Stats/Final Project/Suicide Rates Overview 1985-2016.csv", na.st
#Cleaning the dataset
names(suicides)[names(suicides) == "ï..Country"] <- "country"

#Question
#CI of mean number of male suicides in the year 1988
#Number of suicides committed by male in the year 1988
suicides_1988_male <- filter(suicides, suicides$Year == 1988 & suicides$Sex == 'male')
suicide_male <- suicides_1988_male %>%
  dplyr::select(country, suicides.100k.pop)%>%
  group_by(country) %>%
  summarise(Total_suicides = round(sum(suicides.100k.pop)))

#Plotting histogram
ggplot(suicide_male, aes(x= suicide_male$Total_suicides)) + geom_histogram(color = "black", fill = 'ste
```

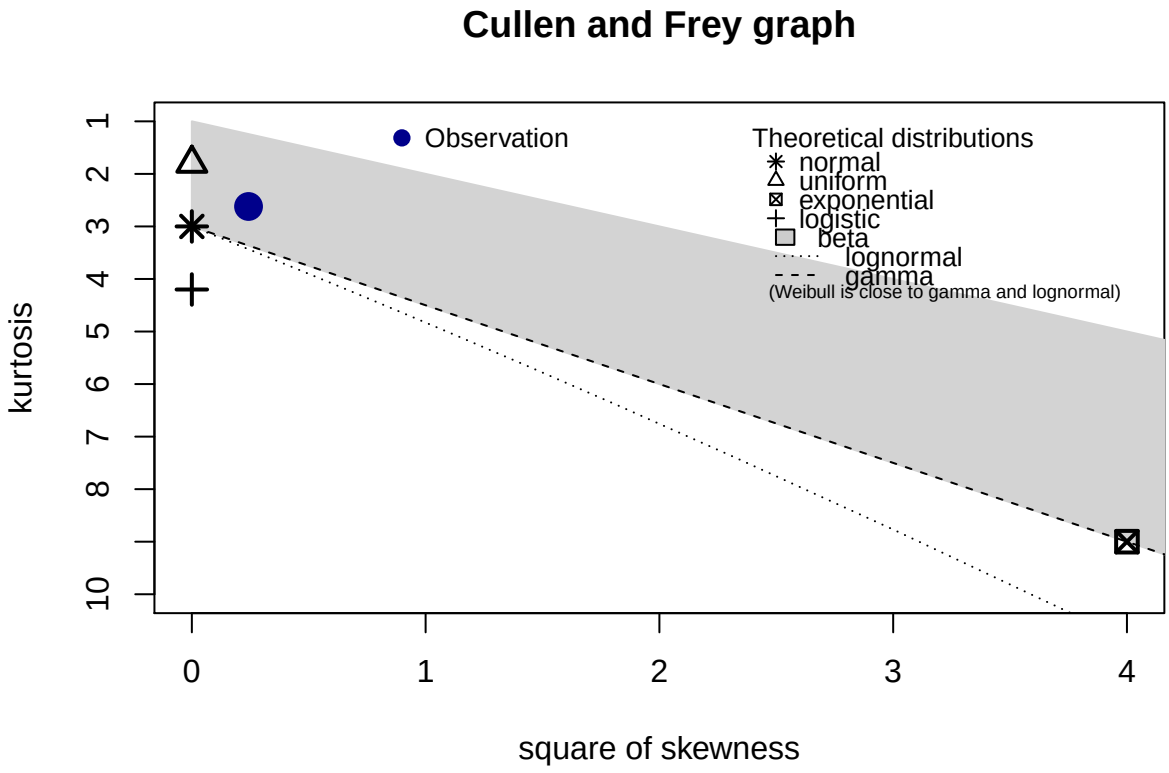
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
fitdist_n <- fitdist(suicide_male$Total_suicides, "norm")
summary(fitdist_n)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 114.67347  10.100952
## sd    70.70675   7.142459
## Loglikelihood:  -278.1965   AIC:  560.393   BIC:  564.1766
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

```
#The Cullen and Frey graph
descdist(suicide_male$Total_suicides)
```

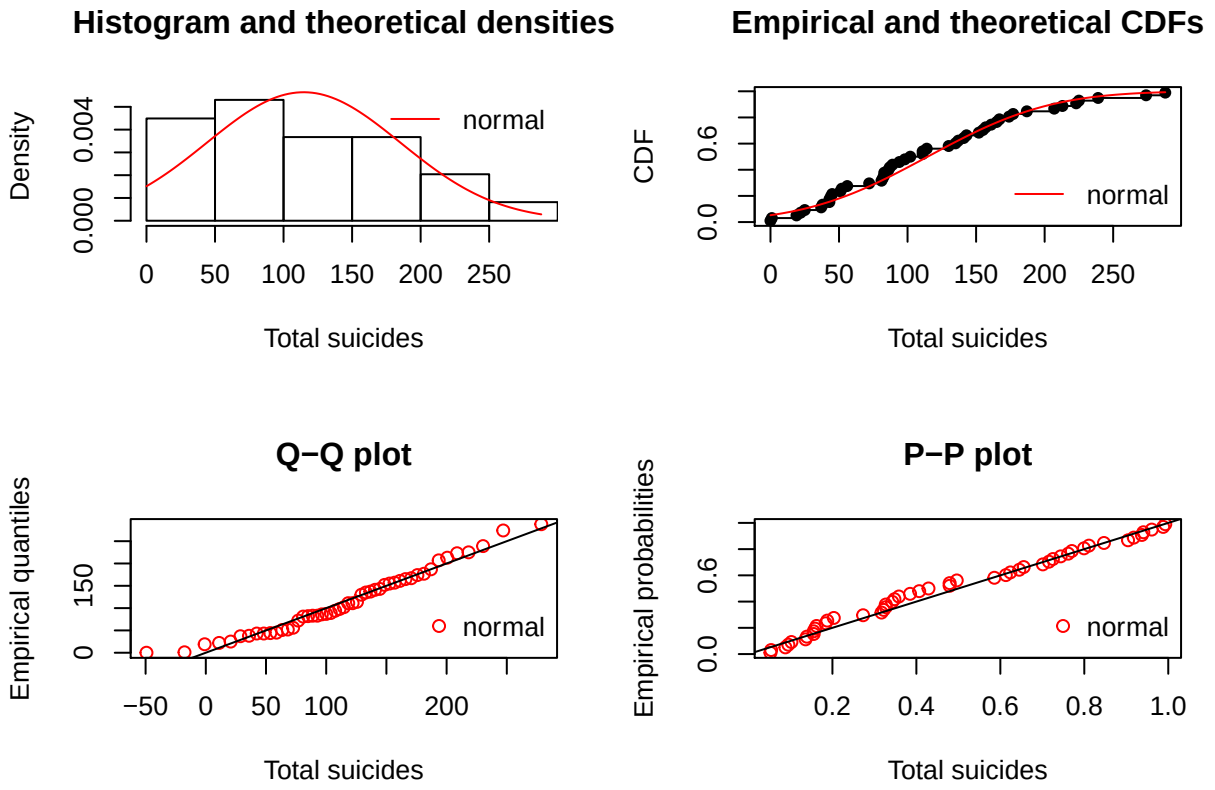



```
## summary statistics
## ------
## min:  0   max:  288
## median:  102
## mean:  114.6735
## estimated sd:  71.43948
## estimated skewness:  0.4925116
## estimated kurtosis:  2.620386
```

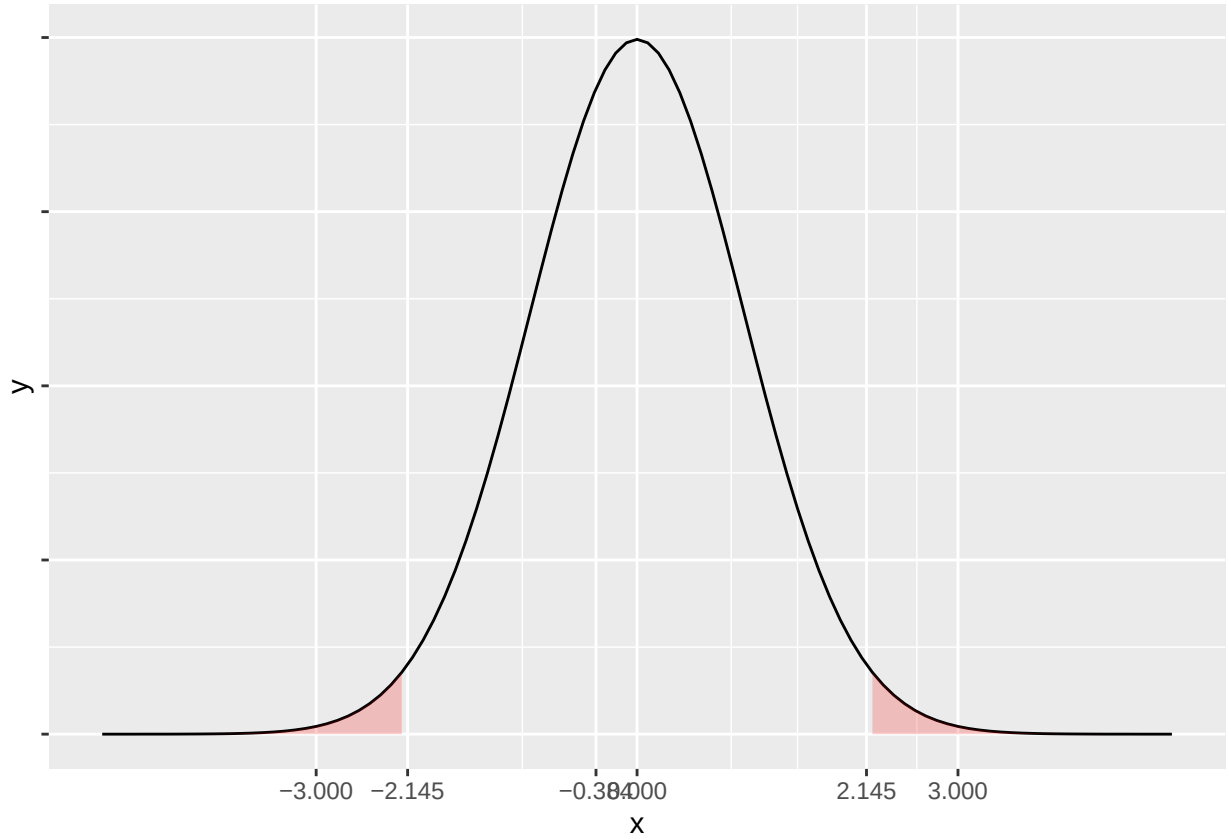
```
par(mfrow=c(2,2))
plot.legend <- c("normal")
fit_n <- fitdist(suicide_male$Total_suicides, "norm")
summary(fit_n)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 114.67347  10.100952
## sd    70.70675   7.142459
## Loglikelihood:  -278.1965   AIC:  560.393   BIC:  564.1766
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

```
denscomp(list(fit_n), legendtext = plot.legend, xlab = 'Total suicides')
cdfcomp (list(fit_n), legendtext = plot.legend, xlab = 'Total suicides')
qqcomp (list(fit_n), legendtext = plot.legend, xlab = 'Total suicides')
ppcomp (list(fit_n), legendtext = plot.legend, xlab = 'Total suicides')
```




```
#Performing t test to test the hypothesis
pop_mean <- mean(suicide_male$Total_suicides)
suicide_male_sample <- sample_n(suicide_male, 15)
Tcalc <- stats::t.test(suicide_male_sample$Total_suicides, mu = pop_mean, alternative = "two.sided")
tvalue <- round(qt(0.025 , 14), 3)

m = 0
std = 1

funcShaded <- function(x, lower_bound, upper_bound) {
  y = dnorm(x, mean = m, sd = std)
  y[x > lower_bound & x < upper_bound] <- NA
  return(y)
}

ggplot(data.frame(x = c(m - (5*std), m + (5*std))), aes(x = x)) +
  stat_function(fun = dnorm, args = list(mean = m, sd = std)) +
  stat_function(fun = funcShaded, args = list(lower_bound = tvalue, upper_bound = -tvalue),
                geom = "area", fill = "red", alpha = .2) +
  scale_x_continuous(breaks = c(m - (3*std), m + (3*std), m, tvalue, -tvalue, round(as.numeric(Tcalc["s
```



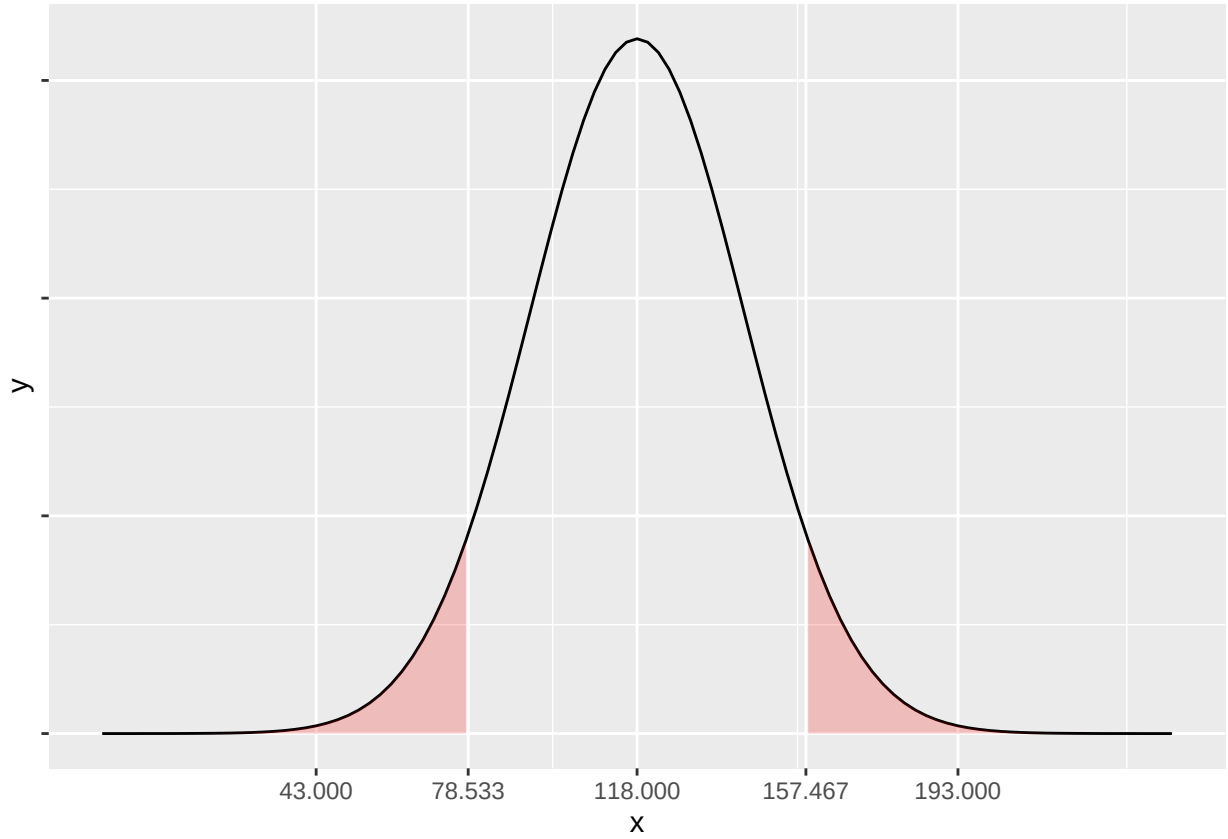
As the p-value > 0.05 and by plotting my graph we can see that Tcalc does not fall in the rejection region. Hence, we fail to reject the null hypothesis and conclude that there is no significant difference between the sample mean of number of suicides by male and population mean of number of suicides by male.

The Confidence Interval for the mean number of suicides committed by male in 1988 with 95% confidence.

Now we tried to find out the confidence interval for the mean number of suicides committed by male n 1988 with 95% confidence. Using the sample_n() function, we took out 15 samples from the male suicide population of 1988. Using the formula and the qt() function, we can find out the lower limit and the upper limit of the confidence interval with 95% confidence.

```
#Importing the dataset
suicides <- read.csv("~/IE 6200 - Prob Stats/Final Project/Suicide Rates Overview 1985-2016.csv", na.st
#Cleaning the dataset
names(suicides)[names(suicides) == "ï..Country"] <- "country"

#Question
#CI of mean number of male suicides in the year 1988
#Number of suicides committed by male in the year 1988
suicides_1988_male <- filter(suicides, suicides$Year == 1988 & suicides$Sex == 'male')
suicide_male <- suicides_1988_male %>%
  dplyr::select(country, suicides.100k.pop)%>%
  group_by(country) %>%
  summarise(Total_suicides = round(sum(suicides.100k.pop)))
```

```r
suicide_male_sample <- sample_n(suicide_male, 15)
sample_mean <- round(mean(suicide_male_sample$Total_suicides))
sample_sd <- sd(suicide_male_sample$Total_suicides)
n <- length(suicide_male_sample$Total_suicides)
e <- qt(0.975, n-1) * (sample_sd/sqrt(n))
lower_limit <- round(sample_mean - e, 3)
upper_limit <- round(sample_mean + e, 3)

m = sample_mean
std = 25

funcShaded <- function(x, lower_bound, upper_bound) {
  y = dnorm(x, mean = m, sd = std)
  y[x > lower_bound & x < upper_bound] <- NA
  return(y)
}

ggplot(data.frame(x = c(m - (5*std), m + (5*std))), aes(x = x)) +
  stat_function(fun = dnorm, args = list(mean = m, sd = std)) +
  stat_function(fun = funcShaded, args = list(lower_bound = lower_limit, upper_bound = upper_limit),
                geom = "area", fill = "red", alpha = .2) +
  scale_x_continuous(breaks = c(m - (3*std), m + (3*std), m, lower_limit, upper_limit)) + theme(axis.te
```



For one such sample, we get the confidence interval of mean as **59.86064 < mean < 139.87270** with the value of error alpha being 5%.

The Confidence interval for the difference in proportion of the number of suicides committed by the adults aged 75 and above in the male and female suicide population.

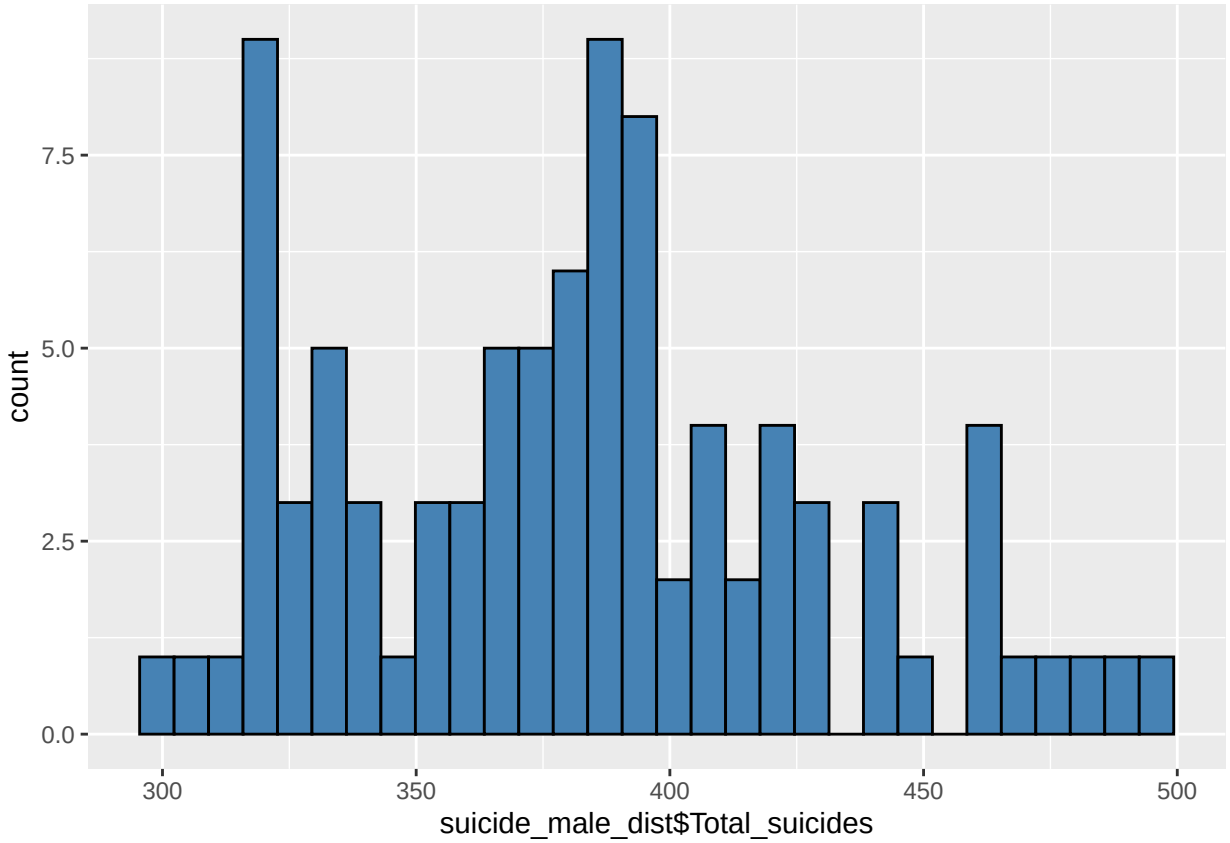
Now lets consider two populations of the number of suicides committed by male and the number of suicides committed by female in the years 2000 to 2016.

Using the pp plot, qq plot, Cullen and frey graph it is clearly evident that both the populations follow a normal distribution.

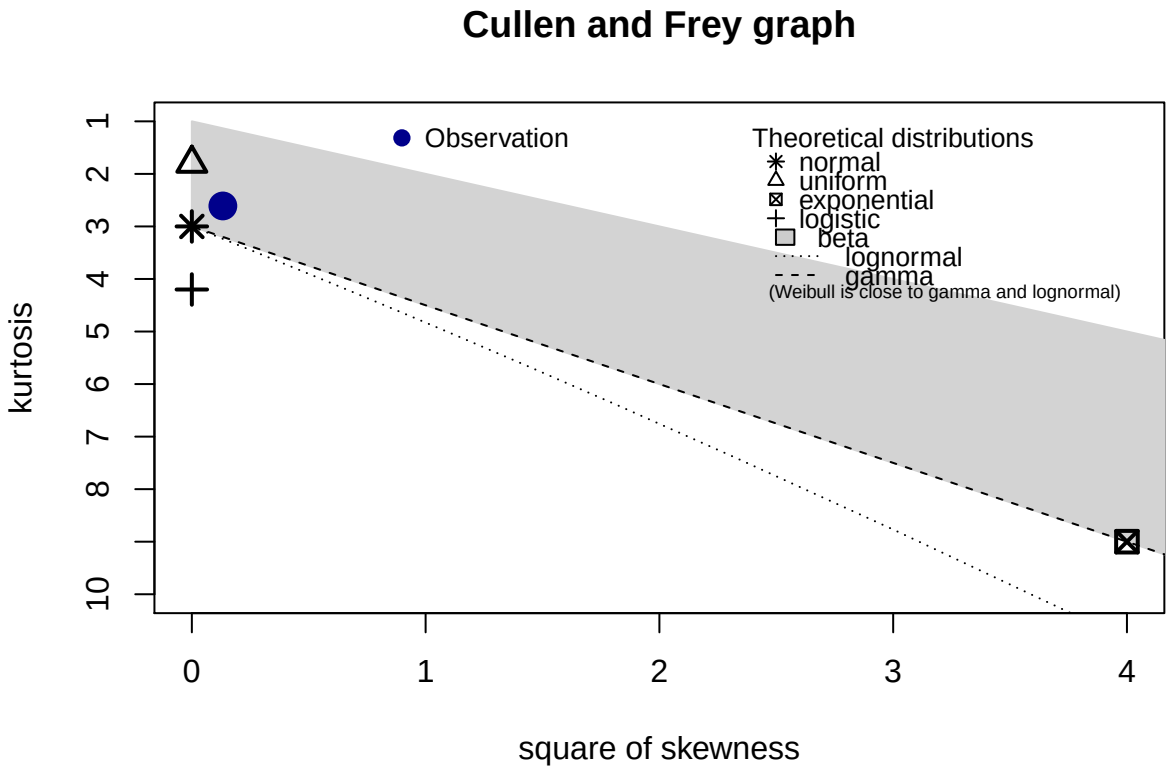
In order to calculate lower and upper limits for the difference between two independent proportions, say the proportion of 75+ males who committed suicide among the whole population of number of male suicides and the proportion of 75+ females who committed suicide among the whole population of number of female suicides.

```
#Importing the dataset
suicides <- read.csv("~/IE 6200 - Prob Stats/Final Project/Suicide Rates Overview 1985-2016.csv", na.st
#Cleaning the dataset
names(suicides)[names(suicides) == "ï..Country"] <- "country"
#Difference in proportion 75+ years between male and female suicides in year 2007
#Male suicides in year 2000 to 2016
suicides_2000s_male <- filter(suicides, suicides$Year >= 2000 & suicides$Year <= 2016 & suicides$Sex ==
#Female suicides in year 2000 to 2016
suicides_2000s_female <- filter(suicides, suicides$Year >= 2000 & suicides$Year <= 2016 & suicides$Sex


#Making male data fit into a normal distribution
suicide_male_dist <- suicides_2000s_male %>%
  dplyr::select(country, Age, suicides.100k.pop) %>%
  group_by(country, Age) %>%
  summarise(Total_suicides = round(sum(suicides.100k.pop)))%>%
  filter(Total_suicides > 300 & Total_suicides < 500)

#Plotting the histogram
ggplot(suicide_male_dist, aes(x= suicide_male_dist$Total_suicides)) + geom_histogram(color = "black", fi
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
descdist(suicide_male_dist$Total_suicides)
```

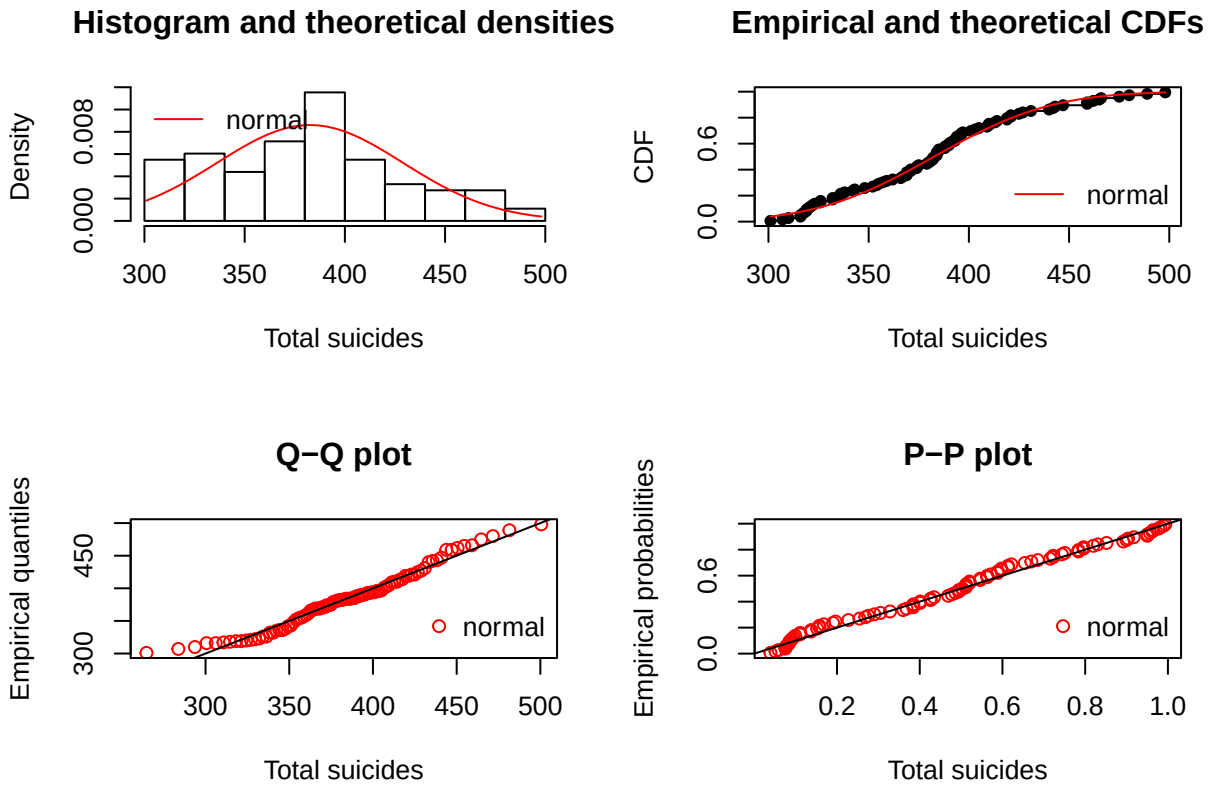



```
## summary statistics
## ------
## min:  301   max:  498
## median:  383
## mean:  382.6264
## estimated sd:  46.62657
## estimated skewness:  0.3643303
## estimated kurtosis:  2.610331
```

```
par(mfrow=c(2,2))
plot.legend <- c("normal")
fit_norm_m <- fitdist(suicide_male_dist$Total_suicides, "norm")
summary(fit_norm_m)
```

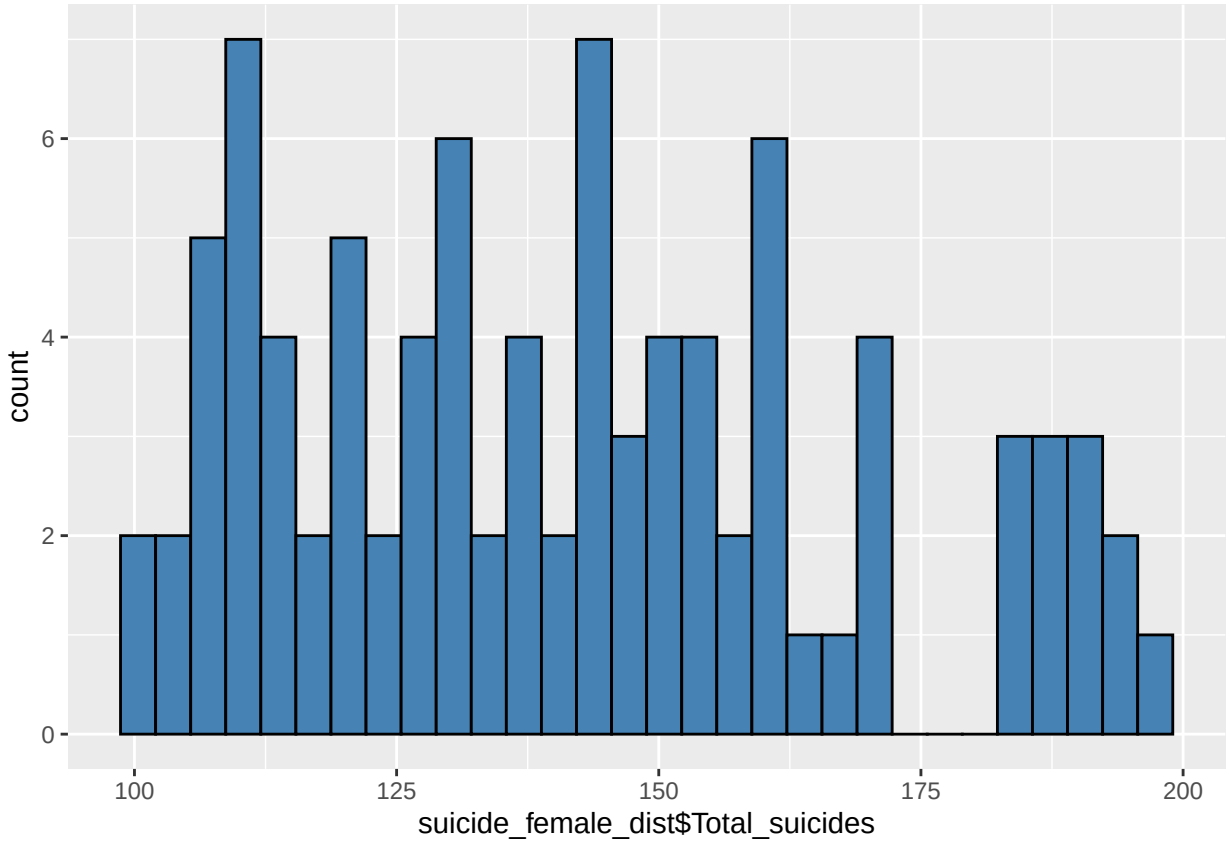
```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 382.62637   4.860862
## sd    46.36967   3.437148
## Loglikelihood:  -478.2582   AIC:  960.5163   BIC:  965.538
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

```
denscomp(list(fit_norm_m), legendtext = plot.legend, xlab = 'Total suicides', xlegend = 'topleft')
cdfcomp (list(fit_norm_m), legendtext = plot.legend, xlab = 'Total suicides')
qqcomp (list(fit_norm_m), legendtext = plot.legend, xlab = 'Total suicides')
ppcomp (list(fit_norm_m), legendtext = plot.legend, xlab = 'Total suicides')
```




```
#Making female data into normal distribution
suicide_female_dist <- suicides_2000s_female %>%
  dplyr::select(country, Age, suicides.100k.pop) %>%
  group_by(country, Age) %>%
  summarise(Total_suicides = round(sum(suicides.100k.pop)))%>%
  filter(Total_suicides>100 & Total_suicides < 200)

#Plotting the histogram
ggplot(suicide_female_dist, aes(x= suicide_female_dist$Total_suicides)) + geom_histogram(color = "black
```

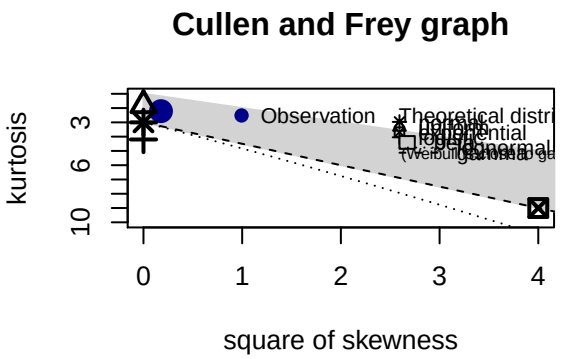
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
descdist(suicide_female_dist$Total_suicides)
```

```
## summary statistics
## ------
## min:  101   max:  198
## median:  139
## mean:  141.6374
## estimated sd:  26.64854
## estimated skewness:  0.4211451
## estimated kurtosis:  2.214501
```

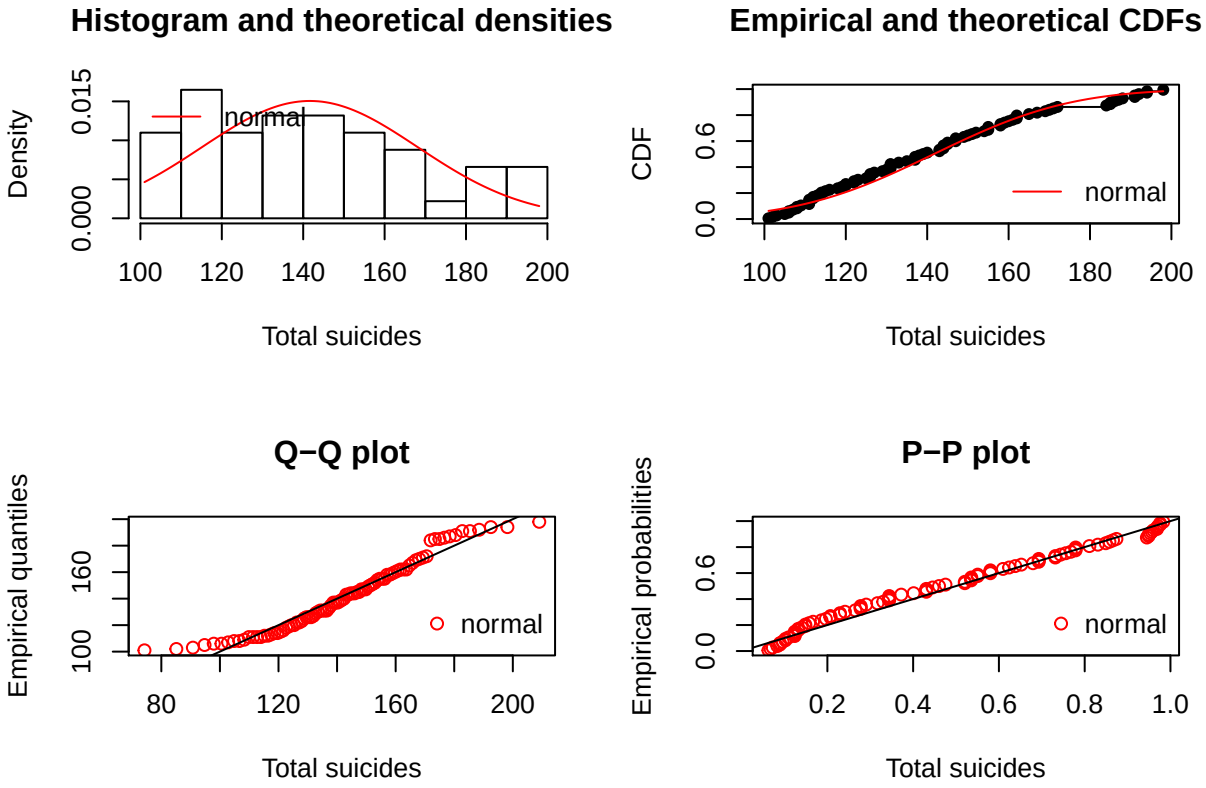
```
fit_norm_f <- fitdist(suicide_female_dist$Total_suicides, "norm")
par(mfrow=c(2,2))
```




```
summary(fit_norm_f)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 141.63736   2.778135
## sd    26.50172   1.964438
## Loglikelihood:  -427.3495   AIC:  858.6989   BIC:  863.7207
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

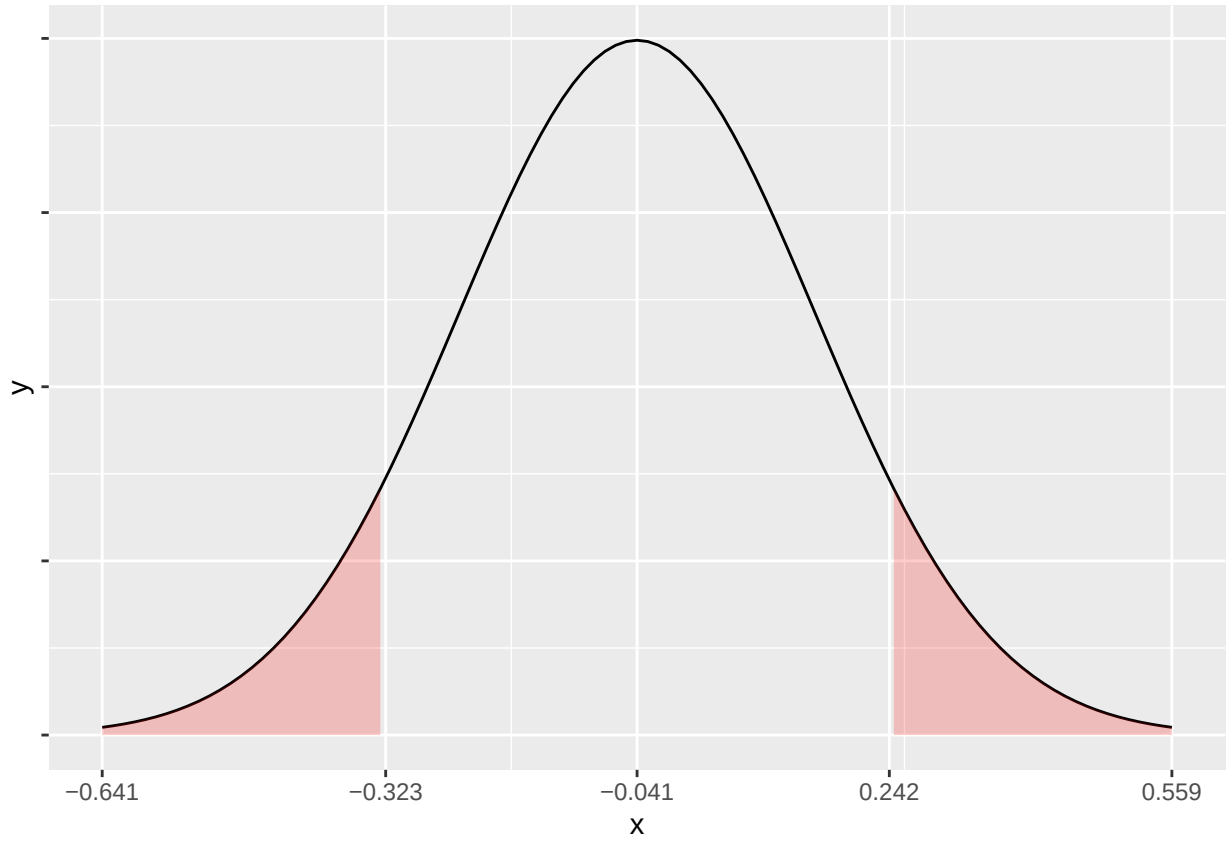
```
denscomp(list(fit_norm_f), legendtext = plot.legend, xlab = 'Total suicides', xlegend = 'topleft')
cdfcomp (list(fit_norm_f), legendtext = plot.legend, xlab = 'Total suicides')
qqcomp (list(fit_norm_f), legendtext = plot.legend, xlab = 'Total suicides')
ppcomp (list(fit_norm_f), legendtext = plot.legend, xlab = 'Total suicides')
```




```
#p1 = Total number of female suicides age 75+/Total number of female suicides
sample1 <- sample(nrow(suicide_female_dist), size = 15)
sample_data_female <- suicide_female_dist[sample1,]
plus_75_female <- sample_data_female %>%
  filter(Age == '75+ years')
p1 <- sum(plus_75_female$Total_suicides)/sum(sample_data_female$Total_suicides)


#p2 = Total number of male suicides age 75+/Total number of male suicides
sample2 <- sample(nrow(suicide_male_dist), size = 15)
sample_data_male <- suicide_male_dist[sample2,]
plus_75_male <- sample_data_male %>%
  filter(Age == '75+ years')
p2 <- sum(plus_75_male$Total_suicides)/sum(sample_data_male$Total_suicides)

#Calculating the degrees of freedom
df = nrow(sample_data_male) + nrow(sample_data_female) - 2
q1 <- 1 - p1
q2 <- 1 - p2
error <- qt(0.975, df = df) * sqrt(((p1*q1)/nrow(sample_data_female)) + ((p2*q2)/nrow(sample_data_male))
left <- round((p1 - p2) - error,3)
right <- round((p1 - p2) + error, 3)
(p1*q1)/nrow(sample_data_female)
```

```
## [1] 0.008602076
```

```r
p <- round(p1-p2,3)



m = p
std = 0.2

funcShaded <- function(x, lower_bound, upper_bound) {
  y = dnorm(x, mean = m, sd = std)
  y[x > lower_bound & x < upper_bound] <- NA
  return(y)
}

ggplot(data.frame(x = c(m - (3*std), m + (3*std))), aes(x = x)) +
  stat_function(fun = dnorm, args = list(mean = m, sd = std)) +
  stat_function(fun = funcShaded, args = list(lower_bound = left, upper_bound = right),
                geom = "area", fill = "red", alpha = .2) +
  scale_x_continuous(breaks = c(m - (3*std), m + (3*std), m, left, right)) + theme(axis.text.y = elemen
```



Calculating the upper and lower limit of the difference of proportion between two populations, we get the lower limit as **-0.3370747 < p1 - p2 < 0.2266442** The difference of proportion for the number of males aged 75 and above who committed suicide and the number of females aged 75 and above who committed suicide are in the above range with 95% confidence.

The Confidence interval of difference of means between the male number of suicides and female number of suicides in the years 2000 to 2016.

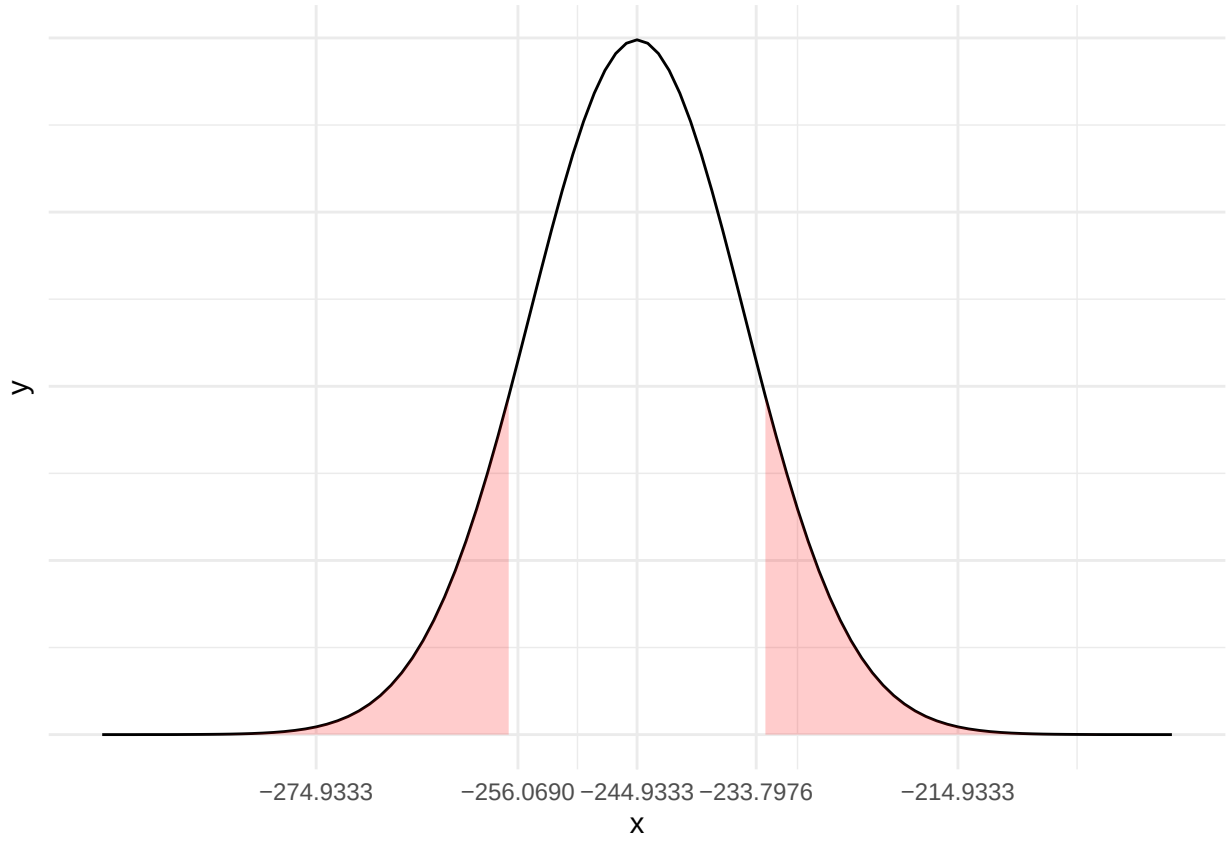
In order to calculate the difference of means between the male suicides and the female suicides in the year 2000 to 2016, we seperate the data into two populations - male and female. We filter the data with certain values to make the population fit into a normal distribution.

In order to calculate the difference of means with unknown and unequal variances, we first take out the sample of each population of size 15 and perform the t test to find the t value and fix that into the formula to find the error.

Using the formula, we add the error and subtract the error to get the upper limit and lower limit of the confidence interval respectively.

```
#Importing the dataset
suicides <- read.csv("~/IE 6200 - Prob Stats/Final Project/Suicide Rates Overview 1985-2016.csv", na.st
#Cleaning the dataset
names(suicides)[names(suicides) == "ï..Country"] <- "country"

#Male suicides in year 2000 to 2016
suicides_2000s_male <- filter(suicides, suicides$Year >= 2000 & suicides$Year <= 2016 & suicides$Sex ==

#Female suicides in year 2000 to 2016
suicides_2000s_female <- filter(suicides, suicides$Year >= 2000 & suicides$Year <= 2016 & suicides$Sex =


#Making male data fit into a normal distribution
suicide_male_dist <- suicides_2000s_male %>%
  dplyr::select(country, Age, suicides.100k.pop) %>%
  group_by(country, Age) %>%
  summarise(Total_suicides = round(sum(suicides.100k.pop)))%>%
  filter(Total_suicides > 300 & Total_suicides < 500)

#Making female data fit into the normal distribution
suicide_female_dist <- suicides_2000s_female %>%
  dplyr::select(country, Age, suicides.100k.pop) %>%
  group_by(country, Age) %>%
  summarise(Total_suicides = round(sum(suicides.100k.pop)))%>%
  filter(Total_suicides>100 & Total_suicides < 200)

sample_suicide_female <- sample(nrow(suicide_female_dist), size = 15)
sample_suicide_male <- sample(nrow(suicide_male_dist), size = 15)
ssf_data <- suicide_female_dist[sample_suicide_female,]
ssm_data <- suicide_male_dist[sample_suicide_male,]
ssf_mean <- mean(ssf_data$Total_suicides)
ssm_mean <- mean(ssm_data$Total_suicides)
ssf_sd<- sd(sample_suicide_female)
ssm_sd<- sd(sample_suicide_male)
err <- qt(0.925, 14) * sqrt((ssf_sd^2/15)+(ssm_sd^2/15))
mean_diff <- ssf_mean - ssm_mean
lower_lim <- mean_diff - err
upper_lim <- mean_diff + err


m = mean_diff
```

```r
std = 10

funcShaded <- function(x, lower_bound, upper_bound) {
  y = dnorm(x, mean = m, sd = std)
  y[x > lower_bound & x < upper_bound] <- NA
  return(y)
}

ggplot(data.frame(x = c(m - (5*std), m + (5*std))), aes(x = x)) +
  stat_function(fun = dnorm, args = list(mean = m, sd = std)) +
  stat_function(fun = funcShaded, args = list(lower_bound = lower_lim, upper_bound = upper_lim),
                geom = "area", fill = "red", alpha = .2) +
  scale_x_continuous(breaks = c(m - (3*std), m + (3*std), m, lower_lim, upper_lim))+
  theme_minimal()+ theme(axis.text.y = element_blank())
```



The above plot is plotted using ggplot and taking the output of one of the samples for the difference of the means.

The confidence interval for the difference of the means between the male suicides and the female suicides in the year 2000s to 2016 is **-247.3818< mean1 - mean2 < -218.0849**

The difference of means of the female number of suicides and the male number of suicides is in the above range with 95% confidence where error alpha is 0.05.

Since the upper limit and the lower limit are both negative, clearly the mean number of suicides committed by male is more than mean number of suicides committed by female. We can conclude that men are more prone to suicides than women.

Range of number of suicides of countries

Here the method of analysis is choosing nine countries with low, moderate and high rates of suicide and plotting a boxplot for each of these countries. The high suicide countries include Hungary, Lithuania, Russian Federation; the moderate countries include Croatia, Norway and Spain; the low suicide rate countries include Greece, Mexico and Paraguay. Plotting the Box plot of the nine countries, it can be seen that all countries in the high suicide range (Lithuania, Russian Federation and Hungary), the degree of dispersion(spread) and skewness in the data is very high compared to countries having moderate suicide rate countries like Croatia and Spain. It can be seen from the data that countries which have high degree of dispersion like Hungary, Lithuania, Russian Federation, Mexico and Greece do not have a GDP per capita.

```
#Importing the dataset
suicides <- read.csv("~/IE 6200 - Prob Stats/Final Project/Suicide Rates Overview 1985-2016.csv", na.st
#Cleaning the dataset
names(suicides)[names(suicides) == "ï..Country"] <- "country"

top3Csuicide <- suicides %>%
  filter(country %in% c("Hungary","Lithuania","Russian Federation", "Oman", "United Arab Emirates", "
  dplyr::select(country, suicides.100k.pop)

ggplot(top3Csuicide, mapping = aes(x = top3Csuicide$country, y = top3Csuicide$suicides.100k.pop)) + geo
```