Asifur Rahman

May 19, 2022

MTH 412

James A. Livsey

Analysis of the difference in suicide rates between men and women in 2015.

Abstract

Do males tend to commit suicide more often then the opposite sex? And does the GDP of a country effect the difference in suicide rates for both genders? The data indicates that the mean suicides per 100k for males is in general greater than females. The difference in suicide rates for the age groups 5-14, 15-24, 25-34, 35-54, 55-75, and 75+ are .038, 8.054, 12.363, 14.759, 15.749, and 26.296 respectively. This indicates that the difference increases as the age increases. Using a bootstrap analysis I concluded that the result for the 5-14 age group is statistically insignificant since the confidence interval is (-0.4716149, 0.4223888) and it contains 0 in the interval and the P-value obtained from the t test is .8713 which is too high to reject the null hypothesis that there is no difference in mean suicides per 100k population for men and women. Charting the difference against the GDP per capita for each of the 62 countries in the data set indicates a negative relationship with a slope of –87.73, however, there is also a correlation of -.05 which indicates a weak negative relationship between the GDP and differences in means.

Introduction

Suicide have been an issue in every country for years, and it has only become an increasingly more pressing issue. However, in recent years mental health has become less of a taboo topic and more openly talked about. With this change comes more research into the topic and more solutions to the problem. Even though mental health has become a much more talked about issue, one thing still remains stigmatized - male mental health. There is and always has been a major stigma around male mental health, males are less likely to seek therapy and less likely to open up about their issues. This is partly due to both societal and cultural stigmatizations along with toxic masculinity. Males are expected to be more “manly” and therefore less “emotional”. This brings me to my question, do males tend to commit suicide more often then the opposite sex?. A follow up question I would also like to ask is, does the GDP of a country effect the difference in suicide rates for both genders?

Data

The data I will be looking at are the suicide rates from the world health organization(WHO). The dataset contains suicide information for 62 countries from 1985 to 2015, however, for the purposes of this analysis I will only look at the information in 2015. The dataset includes variables representing total suicides reported, sample population for each country, and suicides per 100,000 population. Since each country has different total population, in order to get a more consistent result I will use suicides per 100,000 population. The data is divided by the sex, and each sex is divided based on 6 age groups(5-14,15-24, 25-34, 35-54, 55-75, 75+). Finally, the data set also includes the GDP per capita for each country.

Methods

For the purposes of this analysis, I will be looking at the data based on each age group. For each age group I will subtract the mean male suicide per 100k population with the female for all of the countries. This will give us the observed mean difference in suicides per 100k population for each age group. A positive and greater difference will indicate higher suicides rates for men.

The null hypothesis for this analysis is that men and women generally have the same suicides per 100k population, while the alternative hypothesis is that men have higher suicides than women.

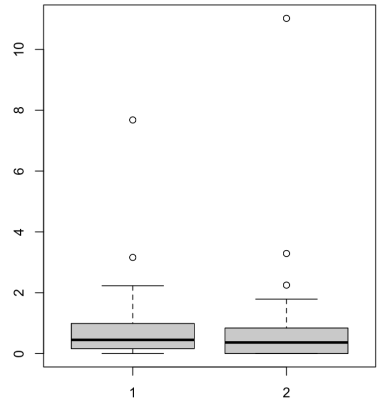
I will use the bootstrap t method to obtain a 95% confidence interval for the difference in means in the data.The Bootstrap method resamples from the original sample to approximate what the result would be if we took many samples from the population, since the original sample only approximates the test statistic of the population. The bootstrap distribution of a test statistic approximates the sampling distribution of the test statistic based on many resamplings. The confidence interval obtained from the bootstrap distribution indicates that 95% of the time the actual test statistic (in this case the difference in mean) is included in the interval. We use the t distribution since the population variance is unknown for this data set and therefore we cannot assume the standard deviation.

We use the t distribution again for the Bootstrap T test for the hypothesis test. If the population variance were known I would've opted for a permutation test since that is a more accurate indicator, however, for the purposes of this analysis the Bootstrap method will suffice. A low enough P-value indicated by the hypothesis test will allow us to reject the null hypothesis.

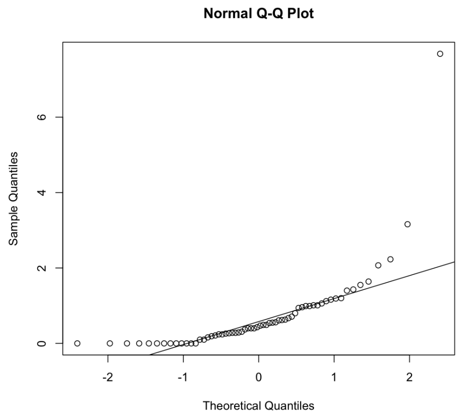
Finally to check to see if the GDP affects the difference between mean suicide rates for men and women, I make a linear regression model with the difference against the GDP. A linear regression model will show the correlation between the two variables, or how strong the relation is, and it will show how the difference changes as the GDP increases.

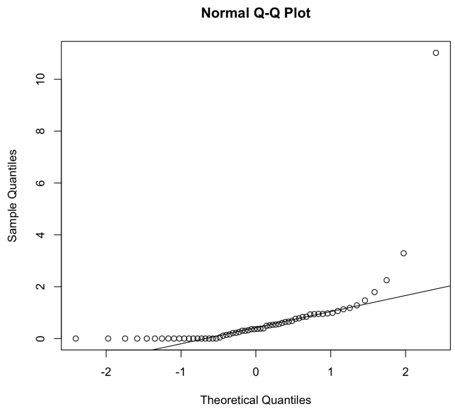
Exploratory Data Analysis

For the first age group, 5-14, the male and female boxplots for suicide per 100k population can be shown below. The suicide rates for this group are fairly low, so the means are very close to each other. Left side represents the male group and the right side represents the female group.

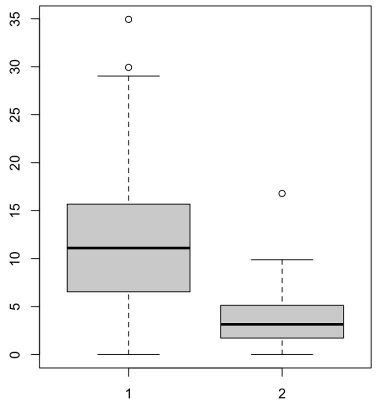


`	If we look at the normal q-q plot for this group, we can see that the data is approximately right skewed. With the female plot being slightly lower and more skewed. However the amount of suicides for this group is very small and the two groups are still very similar due to this. The male plot is shown on the left, and the female plot is shown on the right.

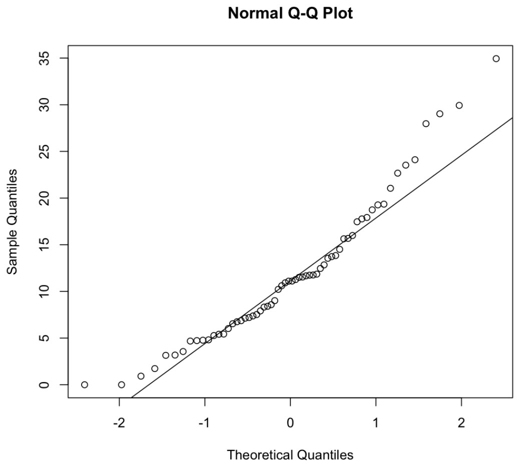


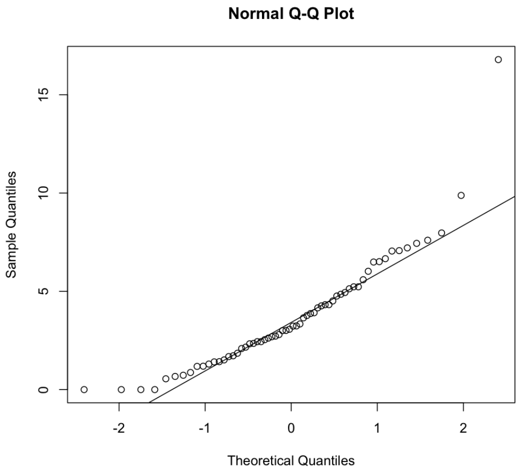


We notice a major difference when we look at the second group, ages 15 to 24. The difference in means are larger than the first age group and there are more suicides reported for both sexes. The mean for the male demographic is much larger than that of the female demographic. The maximum for the female demographic is also much lower, and the spread is lower as well.

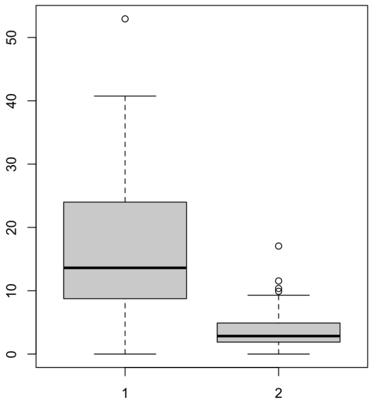


When looking at the Normal Q-Q plots for this age group, men on the left and women on the right. One will notice that the data is still approximately right skewed, and this seems to be the case for all of the age groups.

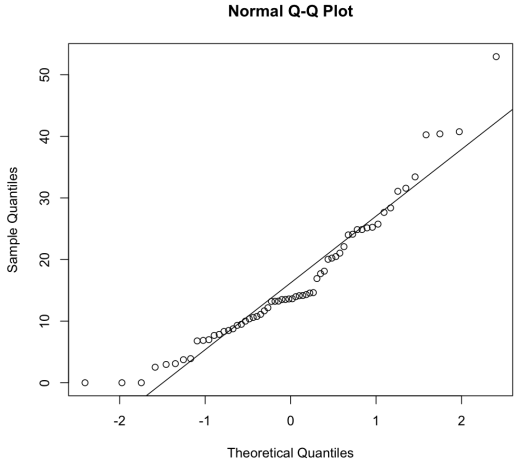


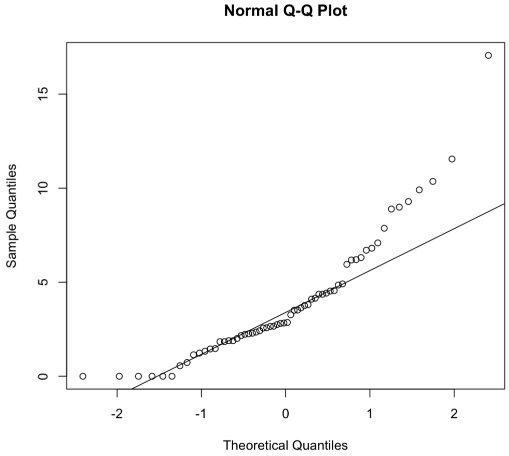


For the third age group, age 25 to 34, the box plots again show that the male demographic has a higher average(mean) suicide rate then women. While the mean suicide rate for both demographic seems to have increased from the previous age groups.

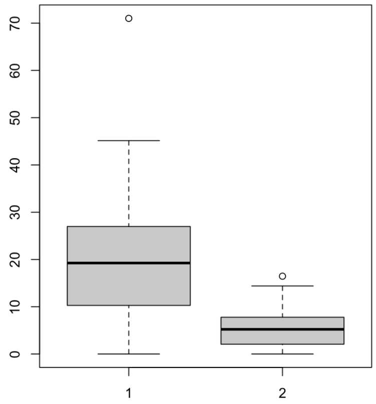


The normal q-q plot again shows a right skewed distribution for both demographics. While the male plot shows a more close relationship with a normal distribution.

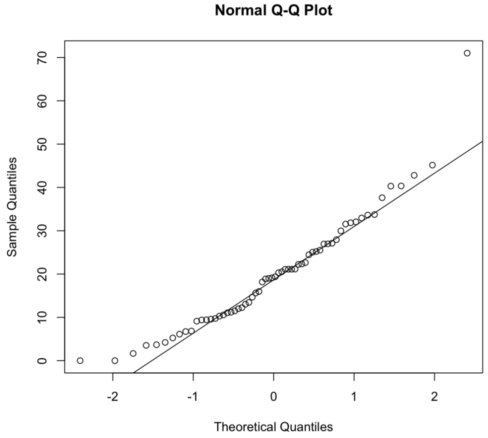


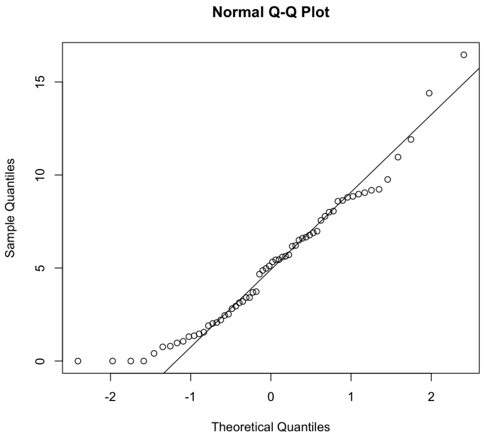


For ages 35-54, the mean suicide rate is greater for males which is a reoccuring theme for all of these groups. The female demographic has a lot less spread in the data and maximum suicides per 100k reported by a country for the male demographic is much higher than that of the female demographic

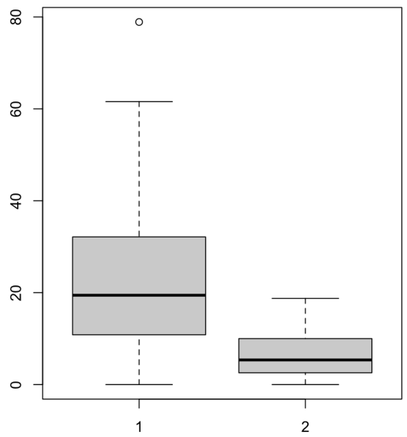


The Q-Q plot below again shows a right skew, however this age group follows the normal distribution very closely.

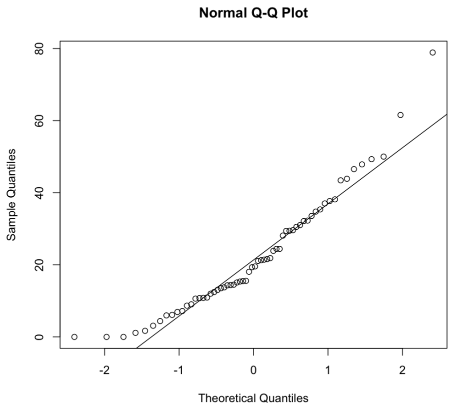


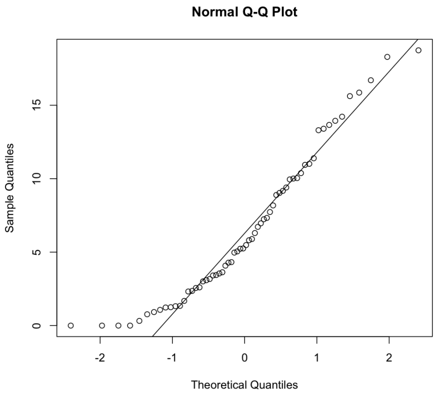


The penultimate group, ages 55-74, again shows a greater mean and spread for the male demographic than the female demographic.

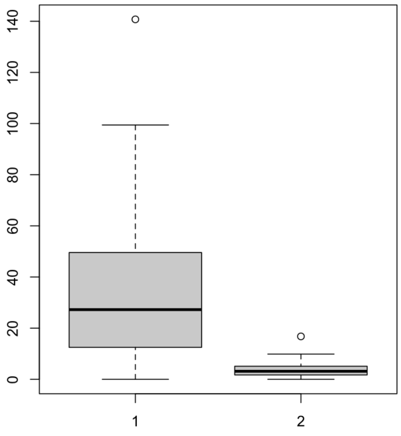


The normal Q-Q plot shows that this age group also follows the normal distribution fairly closely.





Finally, the last age group 75+ has the highest average suicide right among all of the age groups for both demographics, however, the male demographic still has greater average suicides per 100k population. Much larger than the female demographic.



The normal Q-Q plot follows a normal distribution for the female demographic much more than the male demographic. The male demographic has a more right skewed distribution.

N = 62	5-14 year	15-24 year	25-34 year	35-54 year	55-74 year	75+ year
Mean S/100k	.724	.11.815	16.293	19.935	22.326	34.115

Male						
Mean S/100k Female	.686	3.731	3.930	5.176	6.577	7.819
Observed Difference in Means	.038	8.054	12.363	14.759	15.749	26.296

From the above table we can see how the mean suicides per 100k population for both demographics increase from age group to age group, with 5-14 having below 1 suicide per 100k. The observed difference in means also increases as the age group increases. It is worth noting that the age groups from 35 to 75 represent larger intervals of ages than the previous three. Which may affect the amount of suicides reported.

Results

The results for the bootstrap t confidence interval for a the 95% confidence interval is shown below:

N_x= 62 N_y= 62	5-14 year	15-24 year	25-34 year	35-54 year	55-74 year	75+ year
SE	0.2324613	1.03285	1.455489	1.708487	2.148616	3.929227
95% Confidence Interval	-0.4716149, 0.4223888	6.139131, 10.282922	9.601045, 15.443339	11.61908 ,18.53771	11.67852 ,20.42401	19.10511, 35.06223

The 95% confidence interval for the age group, 5-14 years, contains a 0 in the interval, which means that the data for this group is not statistically significant.

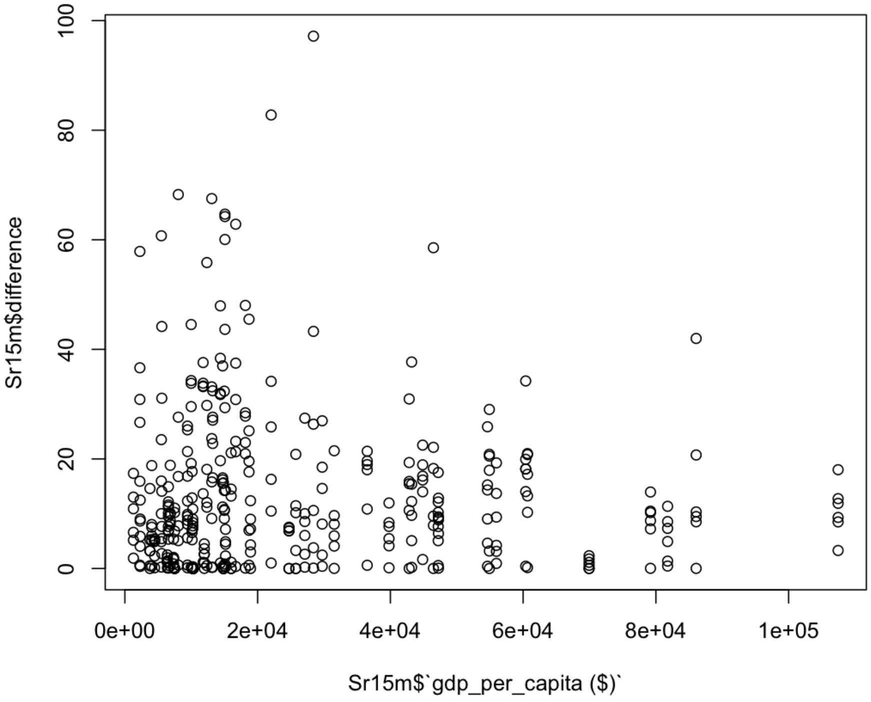
The 95% confidence interval for the age group, 15-24 is (6.139131,10.282922) with a standard error of 1.03285. The 95% confidence interval for the age group, 25-34 is (9.601045,15.443339) with a standard error of 1.455489. The 95% confidence interval for the age group, 35-54 is (11.61908 ,18.53771) with a standard error of 1.708487. The 95% confidence interval for the age group, 55-74 is (11.67852 ,20.42401) with a standard error of 2.148616. Finally, The 95% confidence interval for the age group, 75+ years is (19.10511, 35.06223) with a standard error of 3.929227. Again, a 95% confidence interval means that the true test statistic, in this case the difference in means, falls within the interval in 95% of the resamplings done. It is worth noting that as the age increases so does the confidence interval and the standard error.

To see if the results are statistically significant let's look at the hypothesis t test which uses the bootstrap t method to find the p-value. If the p-value is small enough that means the likelihood of getting the results I got randomly is very small, which means it is statistically significant.

N_x= 62 N_y= 62	5-14 year	15-24 year	25-34 year	35-54 year	55-74 year	75+ year
t	-0.16236	-7.8268	-8.4938	-8.6383	-7.3297	-6.6924
P-value	0.8713	2.089e-11	1.896e-12	1.139e-12	2.553e-10	4.565e-09
Significance	Not statistically significant	Statistically significant	Statistically significant	Statistically significant	Statistically significant	Statistically significant

The p-value for the 5-14 age group is the highest out of all of the age groups with a p-value of .8713. Making this result not statistically significant, since there is too much of a random chance of getting this specific data. However, for the rest of the age groups the p-value is very small with the smallest being 75+ years with a p-value of 4.565e-09. Which means all of these age groups have p-value that indicate the data has a very low chance of occurring randomly.

When plotting a linear regression model of the difference of means between men and women against the change in GDP through the 62 countries the slope of the model was -87.73 indicating a negative relationship. Which means the difference decreases as the GDP in the country increases.



The scatter plot above indicates a generally negative relationship between the GDP per capita and the difference in means. However, the data itself is very far apart and spread out.

When checking the correlation coefficient we get -.05 which indicates a weak negative correlation.

Conclusion

Overall, the data indicates that on average the suicides per 100k population is larger for men then women. The data also indicates the difference grows as the age increases. The highest difference being for the 75+ age group. The 5-14 age group has generally very low reported suicides for most countries and the average is below 1. This is the reason why the data for this age group is statistically insignificant and therefore we cannot reject the null hypothesis for this group. However, for every other age group the p-value obtained is small enough so that we can reject the null hypothesis and conclude that the data is statistically significant.

When looking at the regression model we can conclude that the difference between the mean suicide rates between male and female decrease as the GDP per capita increases. However, the correlation coefficient is -.05 indicating a relatively weak correlation which can be seen by how spread out the data is.

R Code

```
library(readr)
library(dplyr)

#Data Frame of suicide rates
suicide_rates<- read_csv("Downloads/master 2.csv")

Sr15 <- filter(suicide_rates, year == 2015) #Suicide Rates in 2015
Sr15m <- filter(suicide_rates, sex == 'male' & year == 2015) #Male suicide rates in 2015
Sr15f <- filter(suicide_rates, sex == 'female' & year == 2015) #Female suicide rates in 2015

#Age groups Male
Sr15m_1524 <- filter(Sr15m, age == '15-24 years')
Sr15m_2534 <- filter(Sr15m, age == '25-34 years')
Sr15m_3554 <- filter(Sr15m, age == '35-54 years')
Sr15m_5574 <- filter(Sr15m, age == '55-74 years')
Sr15m_75 <- filter(Sr15m, age == '75+ years')
Sr15m_514 <- filter(Sr15m, age == '5-14 years')

#Age groups Female
Sr15f_1524 <- filter(Sr15f, age == '15-24 years')
Sr15f_2534 <- filter(Sr15f, age == '25-34 years')
Sr15f_3554 <- filter(Sr15f, age == '35-54 years')
Sr15f_5574 <- filter(Sr15f, age == '55-74 years')
Sr15f_75 <- filter(Sr15f, age == '75+ years')
Sr15f_514 <- filter(Sr15f, age == '5-14 years')

#Means
mean(Sr15f_514$`suicides/100k pop`)
mean(Sr15f_1524$`suicides/100k pop`)
mean(Sr15f_2534$`suicides/100k pop`)
mean(Sr15f_3554$`suicides/100k pop`)
mean(Sr15f_5574$`suicides/100k pop`)
mean(Sr15f_75$`suicides/100k pop`)

mean(Sr15m_514$`suicides/100k pop`)
mean(Sr15m_1524$`suicides/100k pop`)
mean(Sr15m_2534$`suicides/100k pop`)
mean(Sr15m_3554$`suicides/100k pop`)
mean(Sr15m_5574$`suicides/100k pop`)
mean(Sr15m_75$`suicides/100k pop`)

#QQ norm plots for females
qqnorm(Sr15f_514$`suicides/100k pop`)
qqline(Sr15f_514$`suicides/100k pop`)
qqnorm(Sr15f_1524$`suicides/100k pop`)
qqline(Sr15f_1524$`suicides/100k pop`)
qqnorm(Sr15f_2534$`suicides/100k pop`)
qqline(Sr15f_2534$`suicides/100k pop`)
qqnorm(Sr15f_3554$`suicides/100k pop`)
```

```
qqline(Sr15f_3554$`suicides/100k pop`)
qqnorm(Sr15f_5574$`suicides/100k pop`)
qqline(Sr15f_5574$`suicides/100k pop`)
qqnorm(Sr15f_75$`suicides/100k pop`)
qqline(Sr15f_75$`suicides/100k pop`)

#QQ norm Plots for males
qqnorm(Sr15m_514$`suicides/100k pop`)
qqline(Sr15m_514$`suicides/100k pop`)
qqnorm(Sr15m_1524$`suicides/100k pop`)
qqline(Sr15m_1524$`suicides/100k pop`)
qqnorm(Sr15m_2534$`suicides/100k pop`)
qqline(Sr15m_2534$`suicides/100k pop`)
qqnorm(Sr15m_3554$`suicides/100k pop`)
qqline(Sr15m_3554$`suicides/100k pop`)
qqnorm(Sr15m_5574$`suicides/100k pop`)
qqline(Sr15m_5574$`suicides/100k pop`)
qqnorm(Sr15m_75$`suicides/100k pop`)
qqline(Sr15m_75$`suicides/100k pop`)

#Box plots for each age group
boxplot(Sr15m_514$`suicides/100k pop`,Sr15f_514$`suicides/100k pop`)
boxplot(Sr15m_1524$`suicides/100k pop`,Sr15f_1524$`suicides/100k pop`)
boxplot(Sr15m_2534$`suicides/100k pop`,Sr15f_2534$`suicides/100k pop`)
boxplot(Sr15m_3554$`suicides/100k pop`,Sr15f_3554$`suicides/100k pop`)
boxplot(Sr15m_5574$`suicides/100k pop`,Sr15f_5574$`suicides/100k pop`)
boxplot(Sr15m_75$`suicides/100k pop`,Sr15f_1524$`suicides/100k pop`)

#Difference in means for each age group
thetahat514 <- mean(Sr15m_514$`suicides/100k pop`) - mean(Sr15f_514$`suicides/100k pop`)
thetahat1524 <- mean(Sr15m_1524$`suicides/100k pop`) - mean(Sr15f_1524$`suicides/100k pop`)
thetahat2534 <- mean(Sr15m_2534$`suicides/100k pop`) - mean(Sr15f_2534$`suicides/100k pop`)
thetahat3554 <- mean(Sr15m_3554$`suicides/100k pop`) - mean(Sr15f_3554$`suicides/100k pop`)
thetahat5574 <- mean(Sr15m_5574$`suicides/100k pop`) - mean(Sr15f_5574$`suicides/100k pop`)
thetahat75 <- mean(Sr15m_75$`suicides/100k pop`) - mean(Sr15f_75$`suicides/100k pop`)


#Bootstrap resampling for each age group
nx <- length(Sr15m_514$`suicides/100k pop`)
ny <- length(Sr15f_514$`suicides/100k pop`)
SE <- sqrt(var(Sr15m_514$`suicides/100k pop`)/nx + var(Sr15f_514$`suicides/100k pop`)/ny)
N <- 10000
Tstar <- numeric(N)
for (i in 1:N)
{
  bootx <- sample(Sr15m_514$`suicides/100k pop`, nx, replace = TRUE)
  booty <- sample(Sr15f_514$`suicides/100k pop`, ny, replace = TRUE)
  Tstar[i] <- (mean(bootx) - mean(booty) - thetahat514) /
    sqrt(var(bootx)/nx + var(booty)/ny)
}
thetahat514 - quantile(Tstar, c(.975, .025)) * SE
SE

t.test(Sr15m_514$`suicides/100k pop`, Sr15f_514$`suicides/100k pop`)$conf


nx <- length(Sr15m_1524$`suicides/100k pop`)
ny <- length(Sr15f_1524$`suicides/100k pop`)
SE <- sqrt(var(Sr15m_1524$`suicides/100k pop`)/nx + var(Sr15f_1524$`suicides/100k pop`)/ny)
N <- 10000
Tstar <- numeric(N)
for (i in 1:N)
{
  bootx <- sample(Sr15m_1524$`suicides/100k pop`, nx, replace = TRUE)
  booty <- sample(Sr15f_1524$`suicides/100k pop`, ny, replace = TRUE)
  Tstar[i] <- (mean(bootx) - mean(booty) - thetahat1524) /
    sqrt(var(bootx)/nx + var(booty)/ny)
}
thetahat1524 - quantile(Tstar, c(.975, .025)) * SE
SE


nx <- length(Sr15m_2534$`suicides/100k pop`)
ny <- length(Sr15f_2534$`suicides/100k pop`)
SE <- sqrt(var(Sr15m_2534$`suicides/100k pop`)/nx + var(Sr15f_2534$`suicides/100k pop`)/ny)
N <- 10000
Tstar <- numeric(N)
for (i in 1:N)
{
  bootx <- sample(Sr15m_2534$`suicides/100k pop`, nx, replace = TRUE)
  booty <- sample(Sr15f_2534$`suicides/100k pop`, ny, replace = TRUE)
  Tstar[i] <- (mean(bootx) - mean(booty) - thetahat2534) /
    sqrt(var(bootx)/nx + var(booty)/ny)
}
thetahat2534 - quantile(Tstar, c(.975, .025)) * SE
SE

nx <- length(Sr15m_3554$`suicides/100k pop`)
ny <- length(Sr15f_3554$`suicides/100k pop`)
SE <- sqrt(var(Sr15m_3554$`suicides/100k pop`)/nx + var(Sr15f_3554$`suicides/100k pop`)/ny)
N <- 10000
Tstar <- numeric(N)
for (i in 1:N)
{
  bootx <- sample(Sr15m_3554$`suicides/100k pop`, nx, replace = TRUE)
  booty <- sample(Sr15f_3554$`suicides/100k pop`, ny, replace = TRUE)
  Tstar[i] <- (mean(bootx) - mean(booty) - thetahat3554) /
    sqrt(var(bootx)/nx + var(booty)/ny)
```

```
}
thetahat3554 - quantile(Tstar, c(.975, .025)) * SE
SE

nx <- length(Sr15m_5574$`suicides/100k pop`)
ny <- length(Sr15f_5574$`suicides/100k pop`)
SE <- sqrt(var(Sr15m_5574$`suicides/100k pop`)/nx + var(Sr15f_5574$`suicides/100k pop`)/ny)
N <- 10000
Tstar <- numeric(N)
for (i in 1:N)
{
  bootx <- sample(Sr15m_5574$`suicides/100k pop`, nx, replace = TRUE)
  booty <- sample(Sr15f_5574$`suicides/100k pop`, ny, replace = TRUE)
  Tstar[i] <- (mean(bootx) - mean(booty) - thetahat5574) /
    sqrt(var(bootx)/nx + var(booty)/ny)
}
thetahat5574 - quantile(Tstar, c(.975, .025)) * SE
SE

nx <- length(Sr15m_75$`suicides/100k pop`)
ny <- length(Sr15f_75$`suicides/100k pop`)
SE <- sqrt(var(Sr15m_75$`suicides/100k pop`)/nx + var(Sr15f_75$`suicides/100k pop`)/ny)
N <- 10000
Tstar <- numeric(N)
for (i in 1:N)
{
  bootx <- sample(Sr15m_75$`suicides/100k pop`, nx, replace = TRUE)
  booty <- sample(Sr15f_75$`suicides/100k pop`, ny, replace = TRUE)
  Tstar[i] <- (mean(bootx) - mean(booty) - thetahat75) /
    sqrt(var(bootx)/nx + var(booty)/ny)
}
thetahat75 - quantile(Tstar, c(.975, .025)) * SE
SE

#Bootstrap t test for each age group
Sr15_514 <- filter(Sr15, age == "5-14 years")
t.test(`suicides/100k pop` ~ sex, data = Sr15_514)

Sr15_1524 <- filter(Sr15, age == "15-24 years")
t.test(`suicides/100k pop` ~ sex, data = Sr15_1524)

Sr15_2534 <- filter(Sr15, age == "25-34 years")
t.test(`suicides/100k pop` ~ sex, data = Sr15_2534)

Sr15_3554 <- filter(Sr15, age == "35-54 years")
t.test(`suicides/100k pop` ~ sex, data = Sr15_3554)

Sr15_5574 <- filter(Sr15, age == "55-74 years")
t.test(`suicides/100k pop` ~ sex, data = Sr15_5574)

Sr15_75 <- filter(Sr15, age == "75+ years")
```

```
t.test(`suicides/100k pop` ~ sex, data = Sr15_75)


#Creating a seperate column for difference in means for creating a regression model
agegroup <- list("5-14 years","15-24 years","25-34 years","35-54 years","55-74 years", "75+
years")

diffmeans <- list()

x <- 0
for( i in Sr15f$country){
  x <- x + 1
  if(x == 7){
    x <- 1
    print(agegroup[x])
    tempf <- filter(Sr15f, country == i, age == agegroup[x] )
    tempm <- filter(Sr15m, country == i, age == agegroup[x] )
    diff <- abs(tempm$`suicides/100k pop`-tempf$`suicides/100k pop`)
    diffmeans <- c(diffmeans,diff)
  }
  else if(x != 7){
    print(agegroup[x])
    tempf <- filter(Sr15f, country == i, age == agegroup[x] )
    tempm <- filter(Sr15m, country == i, age == agegroup[x] )
    diff <- abs(tempm$`suicides/100k pop`-tempf$`suicides/100k pop`)
    diffmeans <- c(diffmeans,diff)
  }
}

Sr15m$difference = diffmeans
#Regression model
Sr15m.lm <- lm(Sr15m$`gdp_per_capita ($)` ~ unlist(Sr15m$difference))
Sr15m.lm

unlist(Sr15m$difference)
plot(Sr15m$`gdp_per_capita ($)`,Sr15m$difference)
cor(unlist(Sr15m$difference),Sr15m$`gdp_per_capita ($)`)
```