Suicide Rates Among the Elderly

Tarsus Lam 3032759277

8/17/20

```r
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
```

Part 1

1. Are the elderly more susceptible to suicide due to poor mental health? This type of question is causative/etiologic since we are wondering whether being of old age influences suicide rate.

2. The target population is the suicide rates for all age groups in each country. In order to determine whether the suicide rates are higher for the elderly, we must compare them to the suicide rates of other age groups. We are including multiple countries because there might be other factors in a single country that would increase suicide rate other than old age.

3. The sampling frame is the suicide rates for all people of age groups 10 and above in each country. We do not include children under the age of 10 because their suicide rate is too low to show enough variance (close to zero). As a result, we're only comfortable generalizing the findings for age groups 10 and above. We cannot conclude that that the elderly are more susceptible to poor mental health compared to children under 10 years old.

4. Url: https://www.kaggle.com/twinkle0705/mental-health-and-suicide-rates/

Accessed on July 14, 2020, the source of my dataset is from Kaggle, a website hosting a data science community with public datasets. The data was published by Twinkle Khanna and titled as “Mental Health and Suicide Rates”. It contains the suicide rates of age groups in different countries. I scrolled down to the “Crude suicide rates.csv” and clicked the download icon.

5.

```
suicide_data <- read_csv("suicide_rates.csv")
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   Sex = col_character(),
##   `80_above` = col_double(),
##   `70to79` = col_double(),
##   `60to69` = col_double(),
##   `50to59` = col_double(),
##   `40to49` = col_double(),
##   `30to39` = col_double(),
##   `20to29` = col_double(),
##   `10to19` = col_double()
## )
```

6.

```r
dim(suicide_data)
```

```
## [1] 549  10
```

```r
names(suicide_data)
```

```
##  [1] "Country"  "Sex"      "80_above" "70to79"   "60to69"   "50to59"
##  [7] "40to49"   "30to39"   "20to29"   "10to19"
```

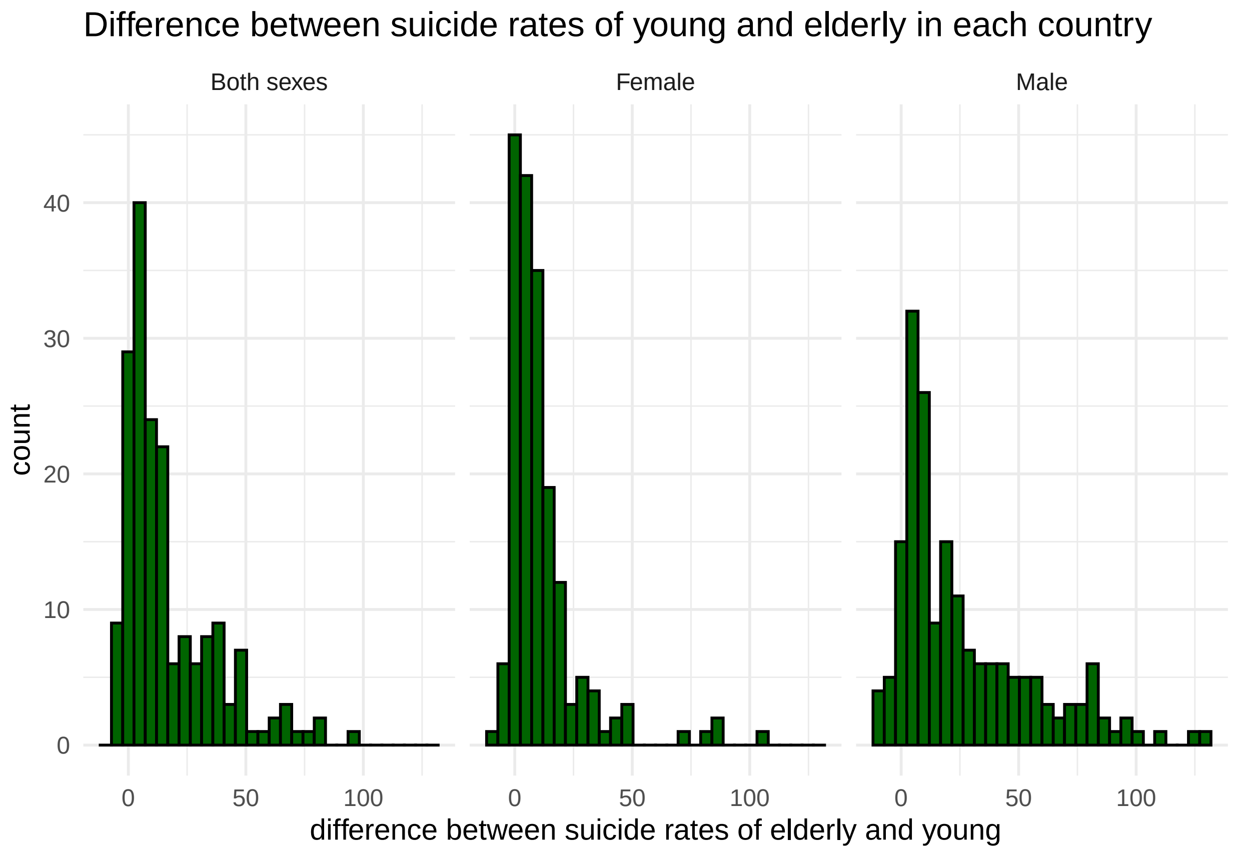
```r
head(suicide_data)
```

```
## # A tibble: 6 x 10
##   Country Sex   `80_above` `70to79` `60to69` `50to59` `40to49` `30to39`
##   <chr>   <chr>      <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 Afghan~ Both~       42       11        5.5      5.6      6.6      9.2
## 2 Afghan~ Male        70.4     20.9      9.8      9.3     10.5     15.1
## 3 Afghan~ Fema~       20.1      2.3      1.4      1.6      2.3      2.7
## 4 Albania Both~       16.3      8.3      6        7.8      9.1      6.1
## 5 Albania Male        23.2     11.9      8.1     11.4     13.5      8.8
## 6 Albania Fema~       10.9      4.9      3.9      4.4      5        3.4
## # ... with 2 more variables: `20to29` <dbl>, `10to19` <dbl>
```

7.

```
suicide_data_renamed <- rename(suicide_data, above_eighty = "80_above",
                               seventies = "70to79", sixties = "60to69",
                               fifties = "50to59", fourties = "40to49",
                               thirties = "30to39", twenties = "20to29",
                               teen = "10to19")

elderly_data <- mutate(suicide_data_renamed, elderly = (above_eighty + seventies + sixties)/3,
                       young = (fifties + fourties + thirties + twenties + teen)/5,
                       difference = elderly - young)

elderly_plot <- ggplot(elderly_data, aes(x = difference)) +
  geom_histogram(color = "black", fill = "darkgreen") +
  facet_wrap(~Sex) +
  theme_minimal() +
  labs(x = "difference between suicide rates of elderly and young",
       title = "Difference between suicide rates of young and elderly in each country")
elderly_plot
```

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



I first renamed the columns to contain no spaces and numerical values so they can be compatible and included in a ggplot function. Then, I added three new columns: classifying ages 60 and above as “elderly”, classifying ages 10-59 years old as “young” and calculating the difference between the two as “difference”. The columns were aggregated by taking the mean of the suicide rates. If the difference is a positive number, then the elderly suicide rate is greater than the young suicide rate in that country. The subsets are divided by gender.

Since the gender can affect suicide rates too, it is best to look at the graphs separately.

The statistical concept I used was the histogram of the difference between the suicide rates of the young and the old suicide. Since the histograms are skewed right and the peaks are at a positive value. Most of the elderly's suicide rates are greater than the young, revealing that the elderly is more susceptible to poor mental health.

Part 2

2. Let X be the event an individual is 80+ years of age

Let Y be the event an individual is 10-19 years old

Let Z be the event an individual commits suicide

I will calculate the difference in suicide rates between the oldest age group (80+ years of age) and the youngest (10-19 years old) based on the suicide dataset of all countries in 2016.

$P(Z|X) - P(Z|Y)$

I am planning to calculate conditional probabilities. We are trying to find the probability of an individual committing suicide **given** that the individual is 80+ years of age and the probability of an individual committing suicide **given** that the individual is 10-19 years old. The probability individuals commit suicide varies/depends on the age group they are in.

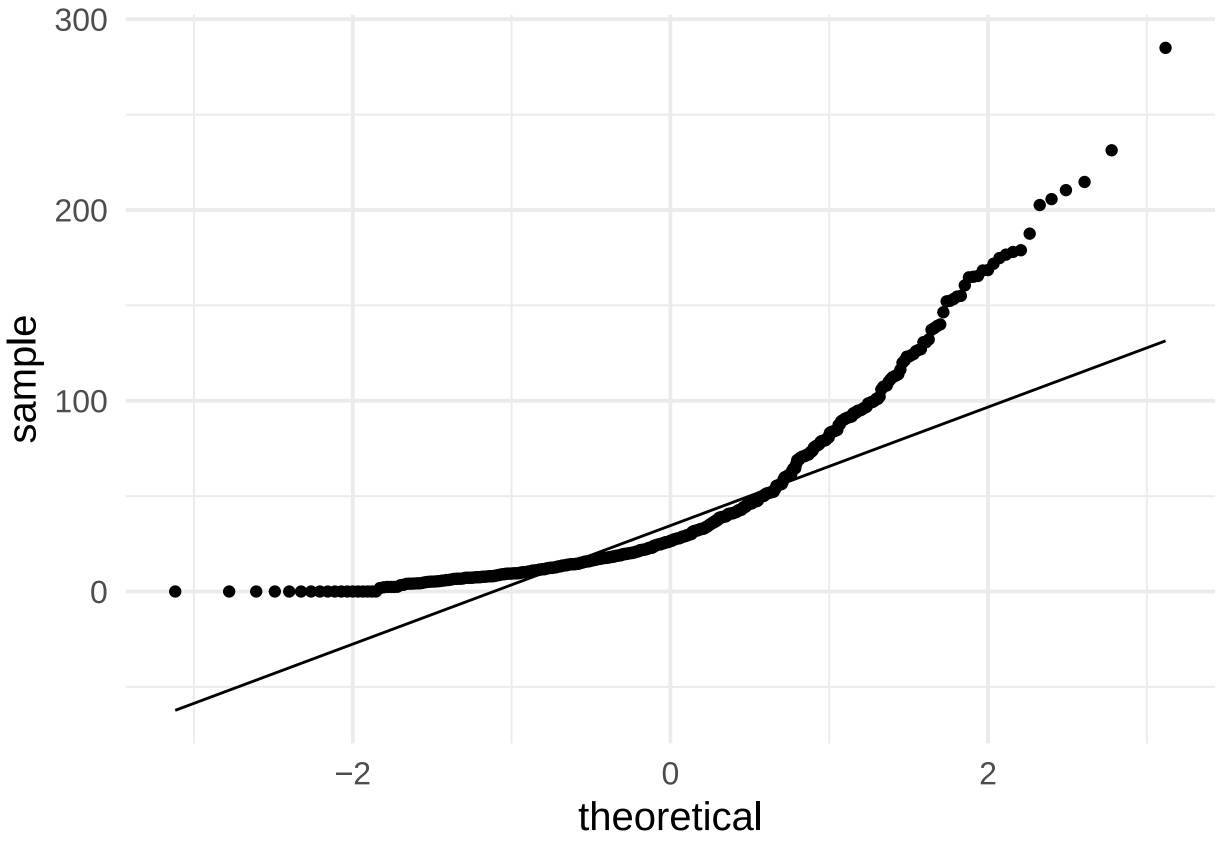
3. I am estimating the difference between suicide rates (not the number of suicides). Suicide rate is a continuous variable because it can take on any value at any given range.

It is most appropriate to use the *Normal Distribution* since suicide rate is a continuous variable, unlike Poisson and Binomial Distributions.

The data contains all suicide rates in 2016 by country. Each suicide rate is independent of each other, X can take on any value between 0 and 100, and the data is sufficiently large containing the suicide rate of each country. Therefore, the data will tend to a normal distribution.

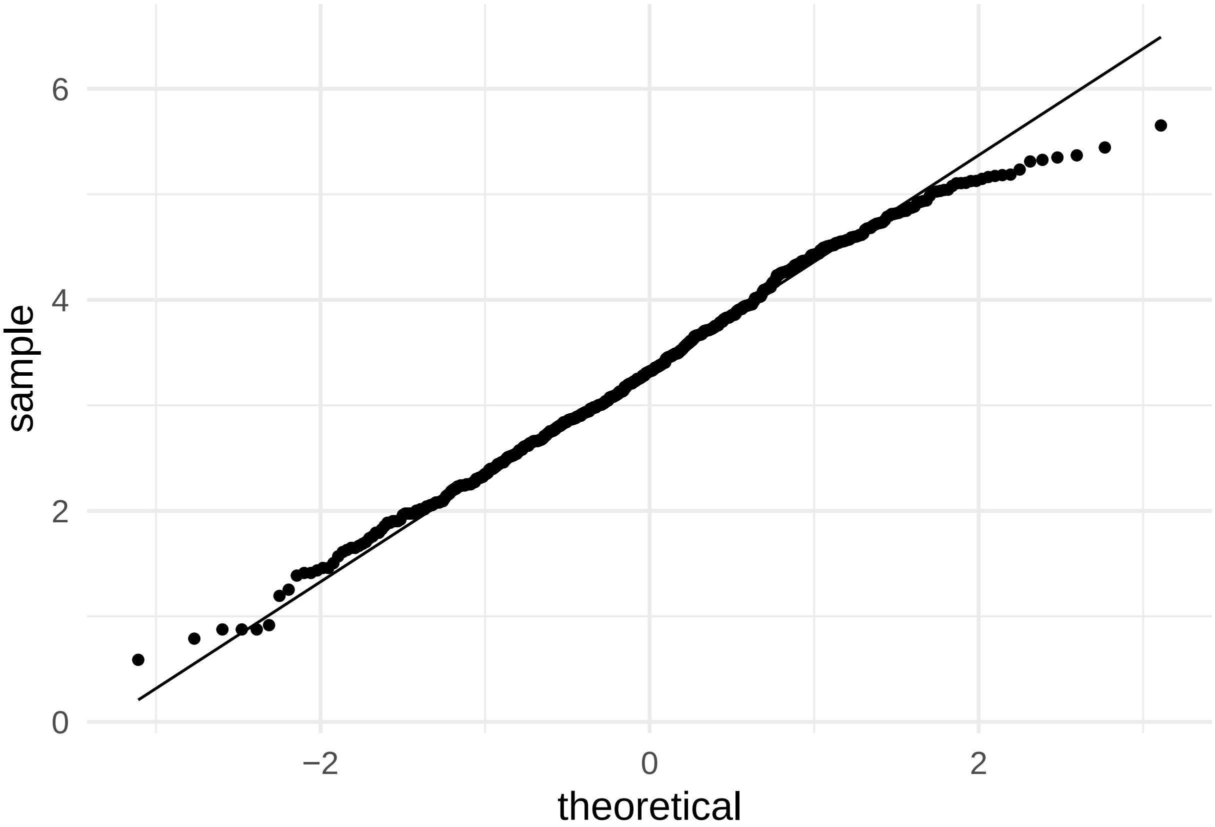
4. I will examine the Q-Q plot of the suicide rates of individuals 80+ years of age in all countries.

```
ggplot(suicide_data_renamed, aes(sample = above_eighty)) + stat_qq() + stat_qq_line() +
theme_minimal(base_size = 15)
```



```
ggplot(suicide_data_renamed, aes(sample = log(above_eighty))) + stat_qq() + stat_qq_line() +
theme_minimal(base_size = 15)
```

```
## Warning: Removed 18 rows containing non-finite values (stat_qq).

## Warning: Removed 18 rows containing non-finite values (stat_qq_line).
```



Q-Q plots help us examine the Normality of the distribution of a variable, in this case, the distribution of suicide rates for individuals 80+ years of age. The more data points that lie on the straight lie, the more confident we can say that the variable (suicide rates) is Normally Distributed.

The first QQ plot shows that the suicide rate distribution is skewed right since the two tails both increasingly slope upwards. In order to resolve this, I took the log of the suicide rates. Consequently, almost all the data points know lie on the straight line. However, the tail ends still deviate (more on the right), showing "light" tails.

Nonetheless, we can conclude that the distribution of the log of the suicide rates for those 80+ years of age resembles a Normal Distribution, but not the initial unchanged suicide rate values.

Part 3

5a. I will be forming an Analysis Of Variance (ANOVA) test on my data, followed by the appropriate Tukey's HSD test.

5b. There are three assumptions required to use the ANOVA test:

1) The samples are simple random samples drawn independently (8 SRS's, one for each of the 8 age group populations).

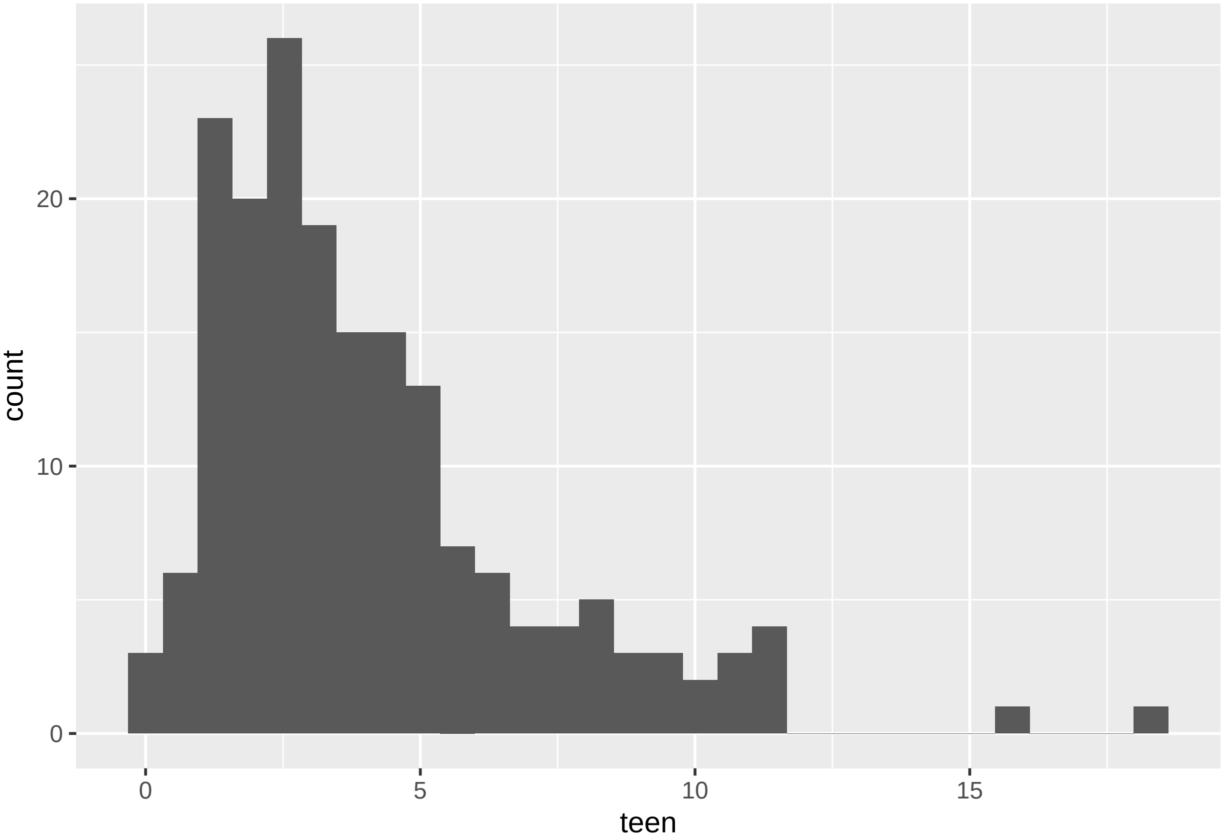
This is satisfied because the samples of an age group are samples drawn from multiple different countries. More importantly, each age group contain data from different people. Therefore, a high/low suicide rate for an age group will not influence the suicide rate of another age group.

2) Each of the 8 age group populations has a normal distribution with an unknown population mean.

I graphed out the histogram of each the age group samples like the one below. This is only somewhat satisfied since most of the populations are skewed right. However, this assumption is not that necessary since the ANOVA test is robust to non-Normality. The population means are unknown because it was not given and since the data does not contain the suicide rates of all countries.

```
# Only use "Both sexes" from each country
both_suicide <- filter(suicide_data_renamed, Sex == "Both sexes")

# Make histogram for suicide rates of teens
ggplot(both_suicide, aes(x = teen)) +
  geom_histogram()
```

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



3) All the age group populations have the same standard deviation, whose value is unknown.

I used dplyr functions (shown below) to calculate the sample SD's and see if the largest sample SD divided by the smallest sample SD is less than 2. This is not satisfied but is not necessary for ANOVA sine the sample sizes are exactly the same. The sample standard deviations are unknown because it was not given and since the data does not contain the suicide rates of all countries.

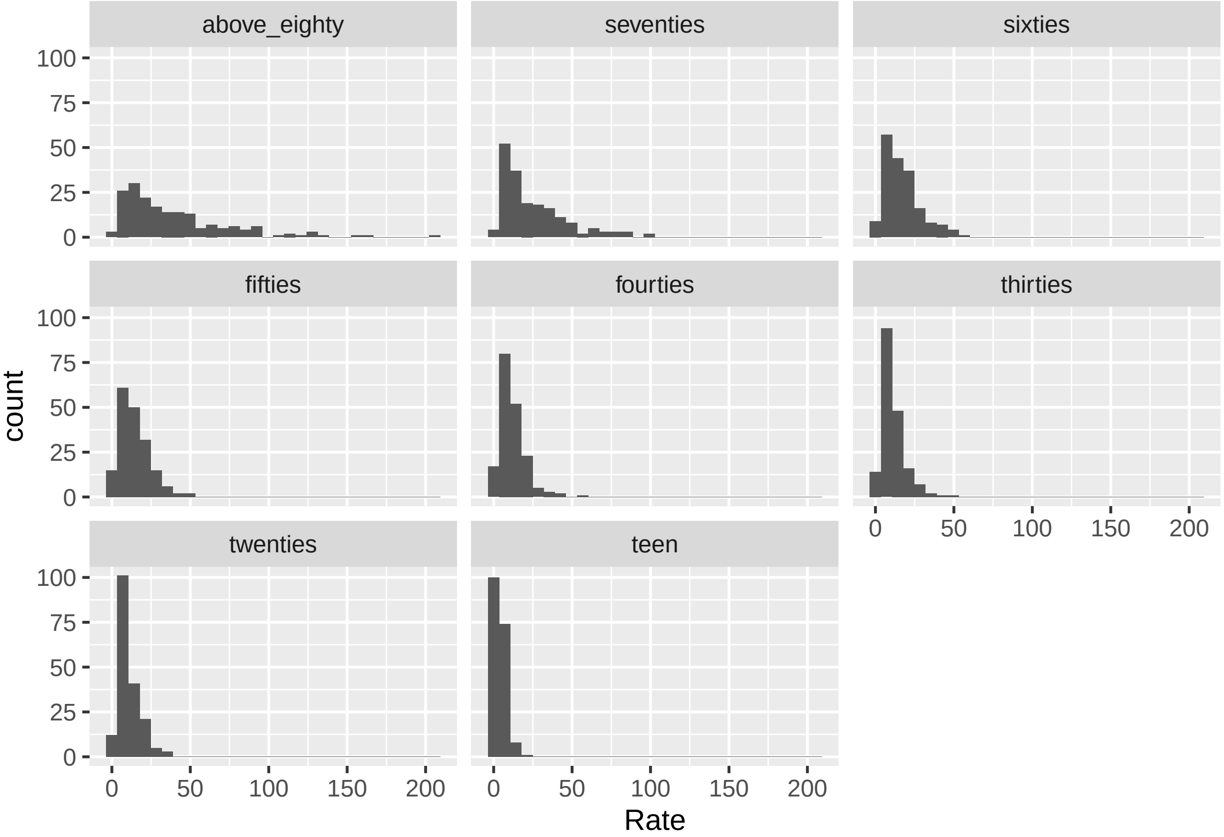
```
# Stack age group columns to correct format so that each suicide rate is labelled
# with respective age group
stacked <- data.frame(both_suicide[1], stack(both_suicide[3:ncol(both_suicide)]))

# Rename columns to informative names 'Age' and 'Rate'
stacked_renamed <- rename(stacked, Age = "ind", Rate = "values")

# Obtain standard deviations of all age group samples
stacked_renamed %>% group_by(Age) %>% summarize(sample_sd = sd(Rate))
```

```
## # A tibble: 8 x 2
##   Age          sample_sd
##   <fct>            <dbl>
## 1 above_eighty     34.5
## 2 seventies        20.9
## 3 sixties          11.4
## 4 fifties           9.74
## 5 fourties          8.41
## 6 thirties          7.73
## 7 twenties          6.94
## 8 teen              2.98
```

5c. The question is whether the elderly is more susceptible to suicide. The ANOVA test will compare suicide rate means of each age group and see if the elderly group(s) are significantly higher. The ANOVA test will also show that there is no significant difference between the suicide rates of most non-elderly groups.

```r
# Make faceted histogram for suicide rates of each age group
ggplot(stacked_renamed, aes(x = Rate)) +
  geom_histogram() +
  facet_wrap(~Age)
```

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



We see in the histograms how most of the graphs are roughly normally distributed, suitable for ANOVA testing, and how the age group “above_eightly” contains higher suicide rates than the other age groups. We want to see if this difference is substantial.

5d. The null hypothesis is that the mean difference in suicide rates of elderly age groups (above_eighty and seventies) is the same as the other age groups.

The alternative hypothesis is that the mean difference in suicide rates of elderly age groups (above_eighty and seventies) is greater than the other age groups.

6.

```
# Only use "Both sexes" from each country
both_suicide <- filter(suicide_data_renamed, Sex == "Both sexes")

# Stack age group columns to correct format for ANOVA
stacked <- data.frame(both_suicide[1], stack(both_suicide[3:ncol(both_suicide)]))

# Rename columns to informative names 'Age' and 'Rate'
stacked_renamed <- rename(stacked, Age = "ind", Rate = "values")

# Perform ANOVA test on relationship between age groups and suicide rates
anova <- aov(Rate ~ Age, stacked_renamed)

# Use Tukey HSD to determine the which group(s) are different from the others
TukeyHSD(anova, conf.level = 0.99)
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = Rate ~ Age, data = stacked_renamed)
##
## $Age
##                              diff        lwr         upr     p adj
## seventies-above_eighty -14.631148 -20.528646  -8.7336491 0.0000000
## sixties-above_eighty   -22.481967 -28.379466 -16.5844687 0.0000000
## fifties-above_eighty   -24.969945 -30.867444 -19.0724469 0.0000000
## fourties-above_eighty  -27.397268 -33.294766 -21.4997693 0.0000000
## thirties-above_eighty  -28.635519 -34.533018 -22.7380206 0.0000000
## twenties-above_eighty  -29.056284 -34.953783 -23.1587857 0.0000000
## teen-above_eighty      -35.463934 -41.361433 -29.5664359 0.0000000
## sixties-seventies       -7.850820 -13.748318  -1.9533212 0.0000758
## fifties-seventies      -10.338798 -16.236296  -4.4412993 0.0000000
## fourties-seventies     -12.766120 -18.663619  -6.8686217 0.0000000
## thirties-seventies     -14.004372 -19.901870  -8.1068731 0.0000000
## twenties-seventies     -14.425137 -20.322635  -8.5276381 0.0000000
## teen-seventies         -20.832787 -26.730285 -14.9352884 0.0000000
## fifties-sixties         -2.487978  -8.385477   3.4095203 0.8125293
## fourties-sixties        -4.915301 -10.812799   0.9821979 0.0646417
## thirties-sixties        -6.153552 -12.051050  -0.2560534 0.0057397
## twenties-sixties        -6.574317 -12.471815  -0.6768185 0.0021755
## teen-sixties           -12.981967 -18.879466  -7.0844687 0.0000000
## fourties-fifties        -2.427322  -8.324821   3.4701761 0.8312985
## thirties-fifties        -3.665574  -9.563072   2.2319247 0.3543360
## twenties-fifties        -4.086339  -9.983837   1.8111597 0.2188480
## teen-fifties           -10.493989 -16.391488  -4.5964906 0.0000000
## thirties-fourties       -1.238251  -7.135750   4.6592471 0.9956849
## twenties-fourties       -1.659016  -7.556515   4.2384821 0.9753318
## teen-fourties           -8.066667 -13.964165  -2.1691682 0.0000405
## twenties-thirties       -0.420765  -6.318264   5.4767335 0.9999968
## teen-thirties           -6.828415 -12.725914  -0.9309168 0.0011708
## teen-twenties           -6.407650 -12.305149  -0.5101518 0.0032220
```

7. The Tukey HSD analysis has "differences" which shows that the mean difference in suicide rates for the older age group in the comparison is greater if negative. When looking at the 'above_eighty' and 'seventies' (and most of the 'sixties') comparisons, the differences and confidence levels are all negative and the p-values are significantly low (< 0.05 and < 0.01). These significant values are shown below.

```r
# Produce table with differences, confidence level and p-values for elderly comparisons
comparisons <- c("seventies-above_eighty", "sixties-above_eighty", "fifties-above_eighty",
                 "fourties-above_eighty", "thirties-above_eighty", "twenties-above_eighty",
                 "teen-above_eighty", "sixties-seventies", "fifties-seventies",
                 "fourties-seventies", "thirties-seventies", "twenties-seventies",
                 "teen-seventies", "thirties-sixties", "twenties-sixties",
                 "teen-sixties")

difference <- c(-14.631148, -22.481967, -24.969945, -27.397268, -28.635519, -29.056284,
                -35.463934, -7.850820, -10.338798, -12.766120, -14.004372, -14.425137,
                -20.832787, -6.153552, -6.574317, -12.981967)

conf_interval <- c("(-20.528646, -8.7336491)", "(-28.379466, -16.5844687)",
                   "(-30.867444, -19.0724469)", "(-33.294766, -21.4997693)",
                   "(-34.533018 -22.7380206)", "(-34.953783, -23.1587857)",
                   "(-41.361433, -29.5664359)", "(-13.748318, -1.9533212)",
                   "(-16.236296, -4.4412993)", "(-18.663619, -6.8686217)",
                   "(-19.901870, -8.1068731)", "(-20.322635, -8.5276381)",
                   "(-26.730285, -14.9352884)", "(-12.051050, -0.2560534)",
                   "(-12.471815, -0.6768185)", "(-18.879466, -7.0844687)")

p_value <- c(0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000,
             0.0000758, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0057397,
             0.0021755, 0.0000000)
tibble("Age_Group_Comparison", comparisons,
       "Mean_Difference_In_Suicide_Rates", difference,
       "Confidence_Interval", conf_interval,
       "Adjusted_p-value", p_value)
```

```
## # A tibble: 16 x 8
##    `"Age_Group_Com~ comparisons `"Mean_Differen~ difference
##    <chr>            <chr>       <chr>                 <dbl>
##  1 Age_Group_Compa~ seventies-~ Mean_Difference~     -14.6
##  2 Age_Group_Compa~ sixties-ab~ Mean_Difference~     -22.5
##  3 Age_Group_Compa~ fifties-ab~ Mean_Difference~     -25.0
##  4 Age_Group_Compa~ fourties-a~ Mean_Difference~     -27.4
##  5 Age_Group_Compa~ thirties-a~ Mean_Difference~     -28.6
##  6 Age_Group_Compa~ twenties-a~ Mean_Difference~     -29.1
##  7 Age_Group_Compa~ teen-above~ Mean_Difference~     -35.5
##  8 Age_Group_Compa~ sixties-se~ Mean_Difference~      -7.85
##  9 Age_Group_Compa~ fifties-se~ Mean_Difference~     -10.3
## 10 Age_Group_Compa~ fourties-s~ Mean_Difference~     -12.8
## 11 Age_Group_Compa~ thirties-s~ Mean_Difference~     -14.0
## 12 Age_Group_Compa~ twenties-s~ Mean_Difference~     -14.4
## 13 Age_Group_Compa~ teen-seven~ Mean_Difference~     -20.8
## 14 Age_Group_Compa~ thirties-s~ Mean_Difference~      -6.15
## 15 Age_Group_Compa~ twenties-s~ Mean_Difference~      -6.57
## 16 Age_Group_Compa~ teen-sixti~ Mean_Difference~     -13.0
## # ... with 4 more variables: `"Confidence_Interval"` <chr>,
## #   conf_interval <chr>, `"Adjusted_p-value"` <chr>, p_value <dbl>
```

8. The statistical test and results show that the elderly are more susceptible to suicide rates since the p-values of every elderly comparison is less than < 0.01. We reject the null hypothesis and accept the alternative, concluding that the mean difference in suicide rates of elderly age groups (above_eighty and seventies) is greater than the other age groups. The strength of the testing is really high since the p-value are really low ($= 0.00000$) even with Bonferroni Correction. The findings can only be generalized with the limited countries in our dataset and is only applicable to the year the data was collected.