

STA 104: Take Home Project

Aman Singh & Peter Mueller

amasin@ucdavis.edu

pfmueeller@ucdavis.edu

UC Davis

Instructor: Amy T. Kim

2/28/2021

Question 1: K or more groups

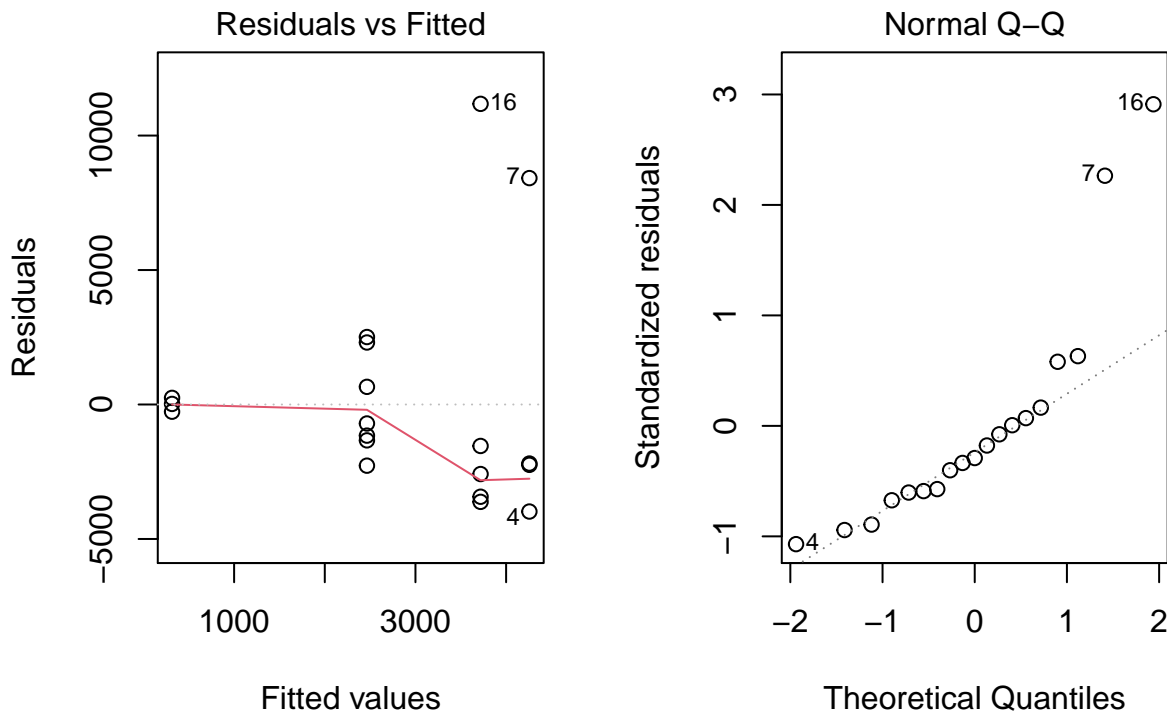
I Introduction

Covid-19 has vastly changed the way of how the world works. Millions of people in America are all unable to do their work properly and have been forced to work at home. We are currently analyzing a data set called CovidA, which lists all the months of states as well as their monthly deaths. We chose to instead subset the dataset and work with 4 states. We chose California, New York City, Florida, and Alabama. We chose California because that is where my partner and I are located and we would like to know more about California compared to other states. We chose New York City because NYC is a typically very popular spot for visitors and we were curious on how the analysis will work with a heavily populated city with Corona. Both of those states listed are also heavily Democratic. We chose Alabama because they are an republican state that is widely known for not believing in the Coronavirus. We also chose Florida largely because it is one of the only states where nobody ever believed in the Coronavirus. All the bars, indoor dining and tourism is rampant over there with the state not caring if anyone wears a mask. We are curious on how Florida and Alabama is doing in terms of the virus compared to states that are very progressive and strict on guidelines such as California and NYC. We also picked these states mainly because we really wanted to see if the Democratic way or the Republican way ultimately ends up in the result or not. Since we want to compare multiple groups, we will be using the K- sample method which is very useful in analyzing groups. We will be using multiple tests such as the Shapiro - Wilks Test and the Levene Test to determine whether to use the Permutation Test or the Kruskal-Wallis Test.

II Summary of Data

Table 1: Summary Statistics

| | California | Florida | Alabama | New York City |
|--------------------|------------|------------|------------|---------------|
| Group Mean | 4256.750 | 2465.00000 | 315.666667 | 3718.200 |
| Group SD | 5673.232 | 1863.53750 | 257.663217 | 6302.866 |
| Rank Mean | 11.750 | 11.42857 | 4.666667 | 9.800 |
| Sample Size | 4.000 | 7.00000 | 3.00000 | 5.000 |



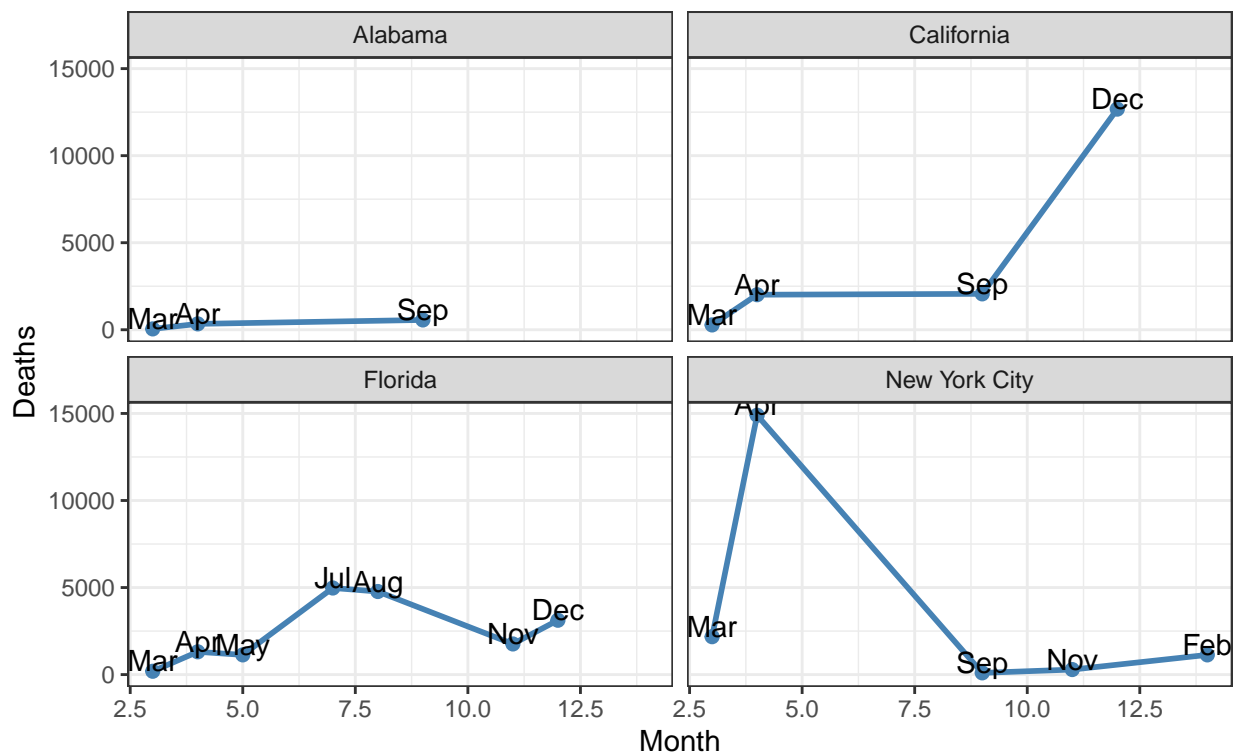
We see from the Shapiro- Wilks test based on the hypothesis that the variable is normally distributed in the dataset. We see that since the p-value is really low, we reject the null hypothesis and confirm that that the

variable is not normally distributed. We also use the Levene test to test if K samples have equal variances and find that we fail to reject the null hypothesis thus concluding that no violations of any assumptions have happened.

In this forthcoming graph, I purposely ended up changing a dataset column where I made a replica of the subsetting graph that we had from the CovidA dataset. I then changed the months to where month 1 was January of 2020 and month 14 was February of 2021. This helped in viewing the graphs as there was no data for every month and when I first did the graph, specifically for NYC, there was no data of February of 2020 but there was of February 2021. This confused GGplot and instead of skipping February of 2020, it displayed the plot of February 2021 instead, which changes the whole graph. I also added another column in the dataset which has the abbreviation names for the months for aesthetically pleasing graphs using GGplot. This graph really helps visualize the whole dataset and understand what we will be doing in this analysis.

Monthly Covid Deaths compared to States

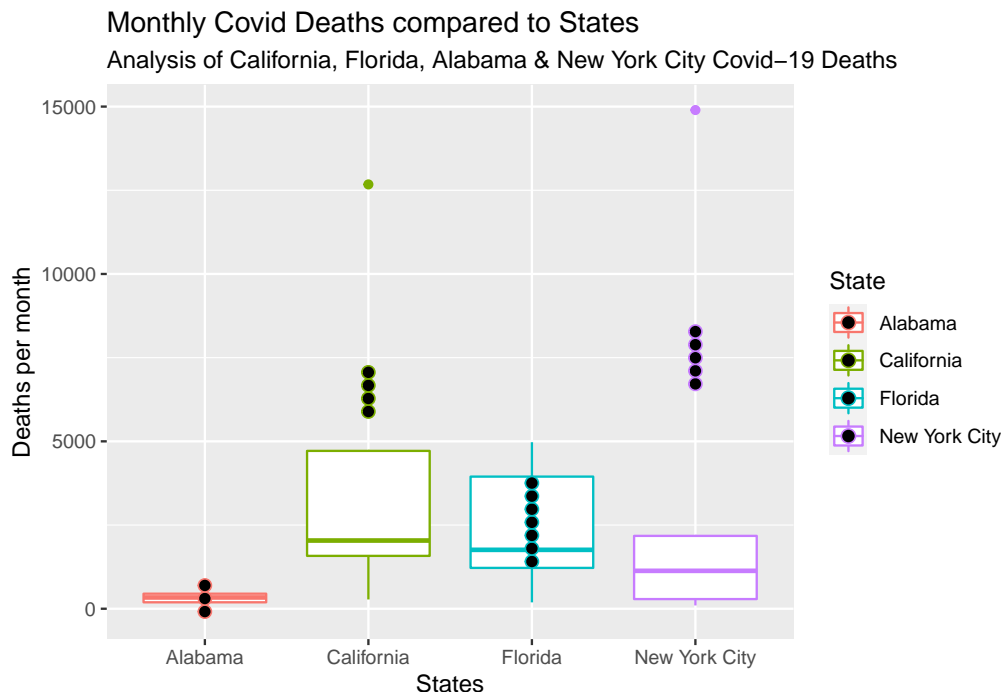
Analysis of California, Florida, Alabama & New York City Covid-19 Deaths



We see here that surprisingly, there were a lot of deaths in California, despite it being a much more Democratic and very progressive state. While they had it mostly low, the sudden increase in deaths was due to the fact that it was December, and people were probably visiting their friends and loved ones regardless which led to an increase in deaths. Florida, in this case, is very surprising how they have managed to keep their deaths down despite the fact that everything is open there and nobody believes in the virus. However, there have been multiple stories of Florida faking their death count as they want to remain open and not face criticism. Infact, during the beginning of the pandemic and data scientist was fired because she refused to fib the numbers of the death count. Alabama is not too surprising because it's not a popular spot to travel to and is not known for tourism. However, it is a shock that not too many deaths happened there as it is a republican state and indoor dining has been going on regardless of the pandemic. I would not be surprised if Alabama might be lying in its death count as well. NYC it seems like had the most death earlier on in 2020, which makes sense as a lot of people still weren't the processing the fact of how serious the pandemic was and so people were still traveling and doing work as well as flights being very active which shows in the graph as to why NYC had too many deaths in April of 2020. People also weren't walking around with masks and were enjoying their life as well. New York City is also known for a lot of consulting jobs as well which is why it

makes me think that led to can increase in cases as consulting companies require its employees to travel to meet clients which leads to more deaths and further rampant spreading.

Lastly we also graph the distribution of deaths in order to analyze which test we would be using, the F-test or the Kruskal-Wallis Test.



Looking at this dataset, we see that there are outliers in multiple groups as well as skewed distributions as well. This leads us to believe to use the Kruskal - Wallis Test in order to understand the results a bit more. We also use Kruskal - Wallis as it has more power than a permutation test when exposed to outliers, or a skewed distribution in 1 or more of the groups. We will be using the regular Kruskal - Wallis Test as the Large Sample Approximation requires the sample size to be $n \geq 30$ which in this case the dataset has only 19.

III Analysis

Since we have seen that we will be using the Kruskal - Wallis Test, we will be using ranks instead of the actual X_{ij} values. We will then make an Confidence Interval to see the difference in ranks. Our null hypothesis in using the Kruskal - Wallis Test is:

$$H_0 : F_1(x) = F_2(x) = F_3(x) = F_4(x)$$

$$H_A : F_i(x) \geq F_j(x) \text{ or } F_i(x) \leq F_j(x) \text{ for some } i \neq j$$

We than perform the test and find that using the KW test with 4000 permutations gives an p-value of .3395.

##Dont know if we want to include this yet We also perform an Permutation Test and find that the p-value while setting the seed presented .7025. This also showed an similar result as the KW test.

IV Interpretation

Since the P-value was larger than α , we conclude that we fail to reject the null hypothesis thus meaning that all the groups are similar and that none of any 2 groups are statistically significant from each other. Since we fail to reject the null hypothesis we find that there is no point in finding out which groups are significantly different from each other which results in us ending our test.

ⓧ Conclusion

We conclude thus that all the 4 groups are all similar with none of them being statistically significant. This comes to us as shocking as Democrats being California & NYC have been enforcing very serious covid restrictions. I also expected the Republican states being Alabama and Florida groups to differ. However, we should also not take this conclusion as an definite answer as there are many other confounding variables in play such as the fact that the population size in California and NYC combined is almost 60 million people. On the other side, Alabama and Florida have a combined of only 25 million. Thus it is more likely to result in more deaths in California and NYC compared to Alabama and Florida. There are also many other confounding variables that we should account for in our analysis such as weather, amount of people traveling between states, age of the population, and even how all the data was collected and arranged. There has been a lot of fake data spread around such as mentioned earlier with the data scientist. We might never be able to get an accurate depiction on whether the states are statistically significant or not due to misconception of data.

Question 2: Tests for Independence

I Introduction

In this section of our report, we test the given data set (`CovidB.csv`) to see if there are any dependencies between any of the variables in the given data set. Our data set consists of the following variables: Year (the year in which the death occurred), Month (the month in which the death occurred), Sex (whether the deceased was male or female), Age Group (the age of the deceased at the time of their death), and Death (the number of deaths within each of the previous categories). We will test for independence between each of these variables using the techniques learned in Lecture, such as the Parametric χ^2 test and Permutation test for independence.

For our project, we have filtered the given data into a smaller subset that concentrates around the 2020 holiday season; particularly October - December of 2020. This will give us insight as to whether or not the holiday season has an impact on the number of deaths.

II Summary of Data Below we have the first 4 lines of our altered data set:

```
## # A tibble: 4 x 5
##   Year Month Sex   Age_Group   Death
##   <dbl> <dbl> <chr> <chr>      <dbl>
## 1  2020    10 Male  0-17 years     6
## 2  2020    10 Male  18-29 years   41
## 3  2020    10 Male  45-54 years  667
## 4  2020    10 Male  55-64 years 1689
```

Next, we will find the average number of deaths for each variable, along with its standard deviation. We will start with the variable Months:

```
##   Month   Death
## 1    10 471.500
## 2    11 908.875
## 3    12 1745.250
```

```
##   Month   Death
## 1    10 611.002
## 2    11 1188.628
## 3    12 2290.832
```

As we can see above, the average number of deaths increases as we progress into the holiday season. This time period is also known for colder weather and is typically the Flu season. These factors may also play a role in the increased number of deaths.

Next, we have the variable Age Group:

```
##   Age_Group   Death
## 1  0-17 years  10.33333
## 2 18-29 years  75.50000
## 3 45-54 years 1104.83333
## 4 55-64 years 2976.83333
```

```
##   Age_Group   Death
## 1  0-17 years  3.204164
## 2 18-29 years 42.378060
## 3 45-54 years 761.362310
## 4 55-64 years 1926.509841
```


When looking at the groups that are most affected by Covid-19, we notice that as the age range of the subject increases, so does the death count.

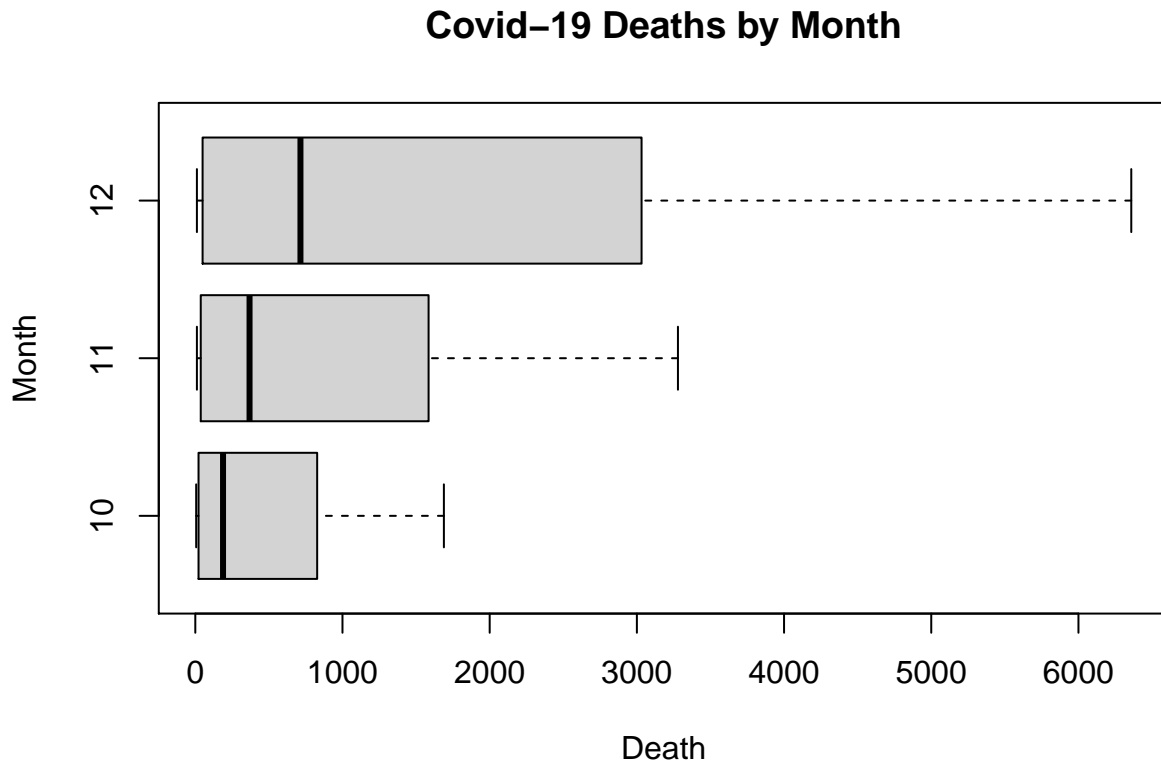
Finally, we have the variable **Sex**:

```
##      Sex      Death
## 1 Female  751.5833
## 2  Male 1332.1667
```

```
##      Sex      Death
## 1 Female 1091.165
## 2  Male 1926.396
```

As we can see above, there are far more male deaths than female deaths, almost by a factor of 2. This can be interpreted very widely, so we will conduct further analysis later in the report.

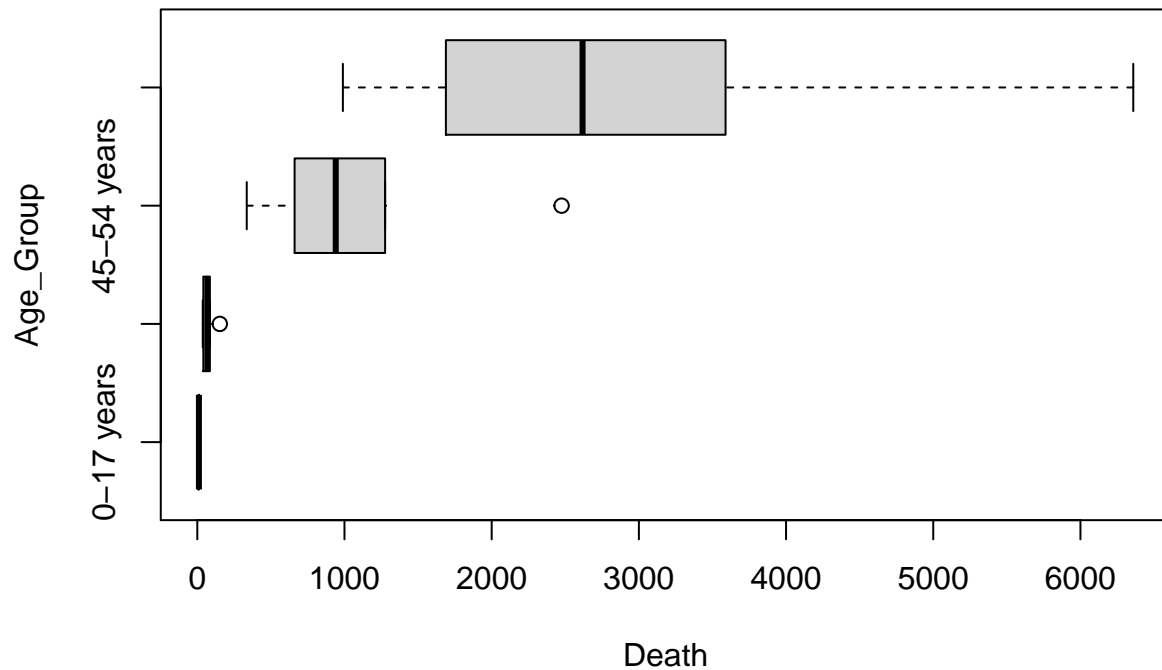
Below we have a box-plot for each month in our selected range:



As we can see from the above plot, the deaths increase as we get closer to the holiday season.

Below we have a box-plot for each age group in our selected range:

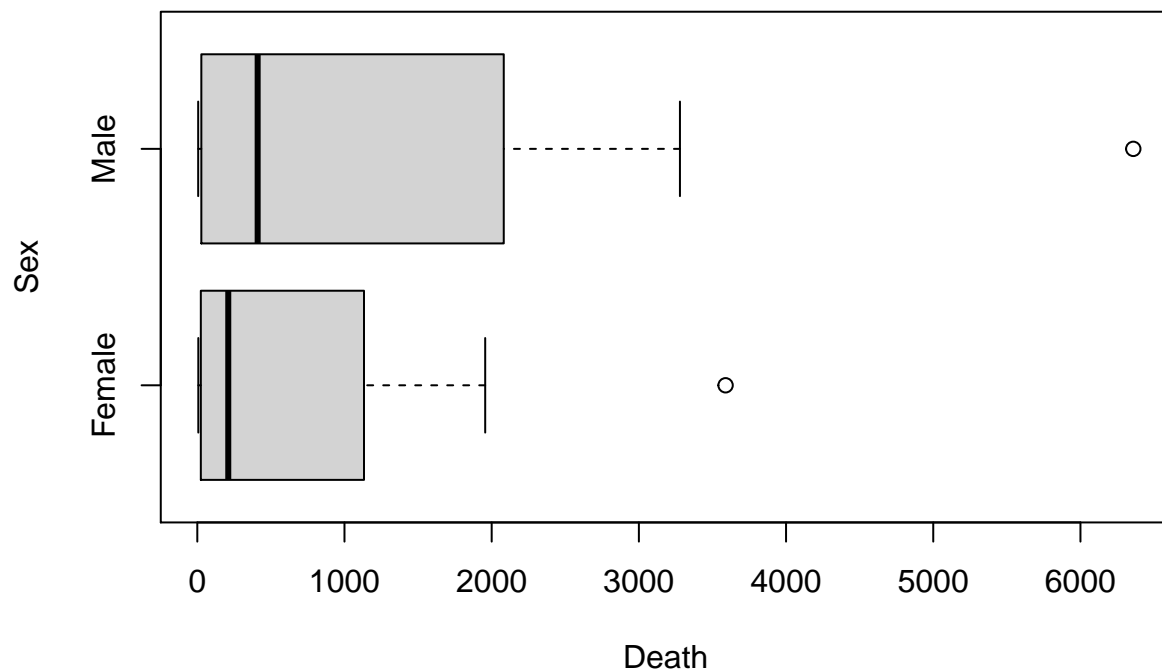
Covid-19 Deaths by Age Group



As we can see above, there is a large difference in deaths between age groups. This makes sense, as those who are older have weaker immune systems, and are therefore less able to fight off the virus once infected.

Below we have a box-plot for each sex:

Covid-19 Deaths by Sex



Here we see that males account for a much larger portion of those who die from the Covid-19 virus.

III Analysis

In this section of the report we will be analyzing our data to answer the following question: Are Sex and Age Group independent in relation to Covid-19 deaths? We will use two tests to help determine the answer. The first test we will use is the Parametric χ^2 test and the second is the Permutation test for independence.

We will start our analysis by looking at the contingency table formed from our filtered data:

```
##           Age_Group
## Sex      0-17 years 18-29 years 45-54 years 55-64 years
## Female      29      183      2273      6534
## Male       33      270      4356      11327
```

Now that we have our table, we will form a null and alternative hypothesis:

H_o : The sex and age group of a patient regarding Covid-19 deaths is independent.

H_A : The sex and age group of a patient regarding Covid-19 deaths is dependent.

Now we can calculate our test statistic, χ^2 :

```
## [1] 17.91532
```

Here we can see our χ^2 value above is 17.91532.

Next, we will proceed with obtaining a p-value by using the Permutation test:

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

As we can see above, our permutation based p-value is 1.

Next, we find the Tukey-Inspired cutoff value:

```
## [1] 3.520652
```

Lastly, we have our observed values:

```
##           0-17 Years 18-29 Years 45-54 Years 55-64 Years
## Males vs. Females  1.757583   1.936158   -3.520652    2.67481
```

IV Interpretation

V Conclusion

Code Appendix

```

# cuttingoffcode
library(knitr)
opts_chunk$set(tidy.opts = list(width.cutoff = 70), tidy = TRUE)
# importing dataset
library(readr)
CovidA <- read_csv("CovidA.csv")
# subsetting data
CovidA_sub = subset(CovidA, CovidA$State == "California" | CovidA$State ==
  "New York City" | CovidA$State == "Florida" | CovidA$State == "Alabama")
# summary of data with mean,sd, rank,sample size
CovidA_sub$Rank = rank(CovidA_sub$Death, ties = "average")
Group.order = aggregate(Death ~ State, data = CovidA_sub, mean)$State
Xi = aggregate(Death ~ State, data = CovidA_sub, mean)$Death
si = aggregate(Death ~ State, data = CovidA_sub, sd)$Death
Ri = aggregate(Rank ~ State, data = CovidA_sub, mean)$Rank
ni = aggregate(Death ~ State, data = CovidA_sub, length)$Death
results = rbind(Xi, si, Ri, ni)
rownames(results) = c("Group Mean", "Group SD", "Rank Mean", "Sample Size")
colnames(results) = as.character(Group.order)
SR.2 = var(CovidA_sub$Rank)
# Anova Test
Ano = aov(Death ~ State, data = CovidA_sub)
par(mfrow = c(1, 2))
plot(Ano, 1)
plot(Ano, 2)
# shapiro test and levene test
shapiro.test(CovidA_sub$Death)
library(lawstat)
levene.test(CovidA_sub$Death, as.factor(CovidA_sub$State))
library(ggplot2)
library(dplyr)
# line graph converting months as to not get confused by year so month
# 1 is January 2020 and month 14 is February 2021
CovidA_sub_graph = CovidA_sub
CovidA_sub_graph$Month[19] = 14
CovidA_sub_graph$month.name = month.abb[CovidA_sub$Month]
# ggplot graph with facetwrap
ggplot(data = CovidA_sub_graph, aes(Month, Death)) + geom_line(color = "steelblue",
  size = 1) + geom_point(color = "steelblue", size = 2) + labs(title = "Monthly Covid Deaths compared
  subtitle = "Analysis of California, Florida, Alabama & New York City Covid-19 Deaths",
  y = "Deaths", x = "Month") + facet_wrap(~State) + theme_bw() + geom_text(aes(label = month.name),
  hjust = 0.5, vjust = 0)
# Boxplot
ggplot(CovidA_sub, aes(x = Death, y = State, color = State)) + geom_boxplot() +
  coord_flip() + geom_dotplot(binaxis = "y", stackdir = "center", dotsize = 0.8) +
  labs(title = "Monthly Covid Deaths compared to States", subtitle = "Analysis of California, Florida
  y = "States", x = "Deaths per month")
# KW Test
set.seed(1)
N = nrow(CovidA_sub)

```

```

KW.OBS = 1/SR.2 * sum(ni * (Ri - (N + 1)/2)^2) #Note, this assumes you calculate ni and Ri above
R = 4000
many.perms.KW = sapply(1:R, function(i) {
  permuted.data = CovidA_sub #So we don't overwrite the original data
  permuted.data$Group = sample(permuted.data$State, nrow(permuted.data),
    replace = FALSE) #Permuting the groups
  SR.2 = var(permuted.data$Rank)
  ni = aggregate(Rank ~ Group, data = permuted.data, length)$Rank
  Ri = aggregate(Rank ~ Group, data = permuted.data, mean)$Rank
  KW.i = 1/SR.2 * sum(ni * (Ri - (N + 1)/2)^2)
  return(KW.i)
})
p.value = mean(many.perms.KW > KW.OBS)
# End of 1st part of Project code Permutation Test
set.seed(2)
R = 4000
many.perms = sapply(1:R, function(i) {
  permuted.data = CovidA_sub #So we don't overwrite the original data
  permuted.data$State = sample(permuted.data$State, nrow(permuted.data),
    replace = FALSE) #Permuting the groups
  Fi = summary(lm(Death ~ State, data = permuted.data))$fstatistic["value"]
  return(Fi)
})
F.OBS = summary(lm(Death ~ State, data = CovidA_sub))$fstatistic["value"]
mean(many.perms >= F.OBS)
# End of 1st part of Project code

library(readr)
CovidB <- read_csv("CovidB.csv")

Covid_B_sub = subset(CovidB, CovidB$Month == 10 & CovidB$Age_Group == "0-17 years" |
  CovidB$Month == 10 & CovidB$Age_Group == "18-29 years" | CovidB$Month ==
  10 & CovidB$Age_Group == "45-54 years" | CovidB$Month == 10 & CovidB$Age_Group ==
  "55-64 years" | CovidB$Month == 11 & CovidB$Age_Group == "0-17 years" |
  CovidB$Month == 11 & CovidB$Age_Group == "18-29 years" | CovidB$Month ==
  11 & CovidB$Age_Group == "45-54 years" | CovidB$Month == 11 & CovidB$Age_Group ==
  "55-64 years" | CovidB$Month == 12 & CovidB$Age_Group == "0-17 years" |
  CovidB$Month == 12 & CovidB$Age_Group == "18-29 years" | CovidB$Month ==
  12 & CovidB$Age_Group == "45-54 years" | CovidB$Month == 12 & CovidB$Age_Group ==
  "55-64 years")
head(Covid_B_sub, 4)

average_death_by_month = aggregate(Death ~ Month, data = Covid_B_sub, mean)
average_death_by_month

sd_death_by_month = aggregate(Death ~ Month, data = Covid_B_sub, sd)
sd_death_by_month

average_death_by_age = aggregate(Death ~ Age_Group, data = Covid_B_sub,
  mean)
average_death_by_age

```

```

sd_death_by_age = aggregate(Death ~ Age_Group, data = Covid_B_sub, sd)
sd_death_by_age

average_death_by_sex = aggregate(Death ~ Sex, data = Covid_B_sub, mean)
average_death_by_sex

sd_death_by_sex = aggregate(Death ~ Sex, data = Covid_B_sub, sd)
sd_death_by_sex

library(ggplot2)
boxplot(Death ~ Month, data = Covid_B_sub, main = "Covid-19 Deaths by Month",
        horizontal = TRUE)

boxplot(Death ~ Age_Group, data = Covid_B_sub, main = "Covid-19 Deaths by Age Group",
        horizontal = TRUE)

boxplot(Death ~ Sex, data = Covid_B_sub, main = "Covid-19 Deaths by Sex",
        horizontal = TRUE)

contingency_table = xtabs(Death ~ Sex + Age_Group, data = Covid_B_sub)
contingency_table

ni. = rowSums(contingency_table)
n.j = colSums(contingency_table)

the.test = chisq.test(contingency_table, correct = FALSE)
eij = the.test$expected
chi.sq.obs = as.numeric(the.test$statistic)

chi.sq.obs

set.seed(3223)
R = 4000
r.perms = sapply(1:R, function(i) {
  perm.data = Covid_B_sub
  perm.data$Age_Group = sample(perm.data$Age_Group, nrow(perm.data),
                              replace = FALSE)
  chi.sq.i = chisq.test(contingency_table, correct = FALSE)$stat
  return(chi.sq.i)
})
perm.pval = mean(r.perms >= chi.sq.obs)
perm.pval

n = sum(contingency_table)
ni. = rowSums(contingency_table)
n.j = colSums(contingency_table)
all.pjG1 = contingency_table[1, ]/ni.[1]
all.pjG2 = contingency_table[2, ]/ni.[2]
all.pbar = n.j/n
all.Zij = c(all.pjG1 - all.pjG2)/sqrt(all.pbar * (1 - all.pbar) * (1/ni.[1] +
  1/ni.[2]))

r.perms.cutoff = sapply(1:R, function(i) {

```

```

perm.data = Covid_B_sub
perm.data$Sex = sample(perm.data$Sex, nrow(perm.data), replace = FALSE)
row.sum = rowSums(contingency_table)
col.sum = colSums(contingency_table)
all.pji = contingency_table[1, ]/row.sum[1]
all.pji. = contingency_table[2, ]/row.sum[2]
all.pbar = col.sum/sum(row.sum)
all.Zij = c(all.pji - all.pji.)/sqrt(all.pbar * (1 - all.pbar) * (1/row.sum[1] +
  1/row.sum[2]))
Q.r = max(abs(all.Zij))
return(Q.r)
})
alpha = 0.05
cutoff.q = as.numeric(quantile(r.perms.cutoff, (1 - alpha)))
cutoff.q
all.Zij_2 = matrix(all.Zij, nrow = 1)
colnames(all.Zij_2) = c("0-17 Years", "18-29 Years", "45-54 Years", "55-64 Years")
rownames(all.Zij_2) = c("Males vs. Females")
all.Zij_2

```