STA 104: Take Home Project
Looking into States that are being affected from Covid-19

Aman Singh & Peter Mueller
amasin@ucdavis.edu
pfmueller@ucdavis.edu
UC Davis
Instructor: Amy T. Kim

2/28/2021

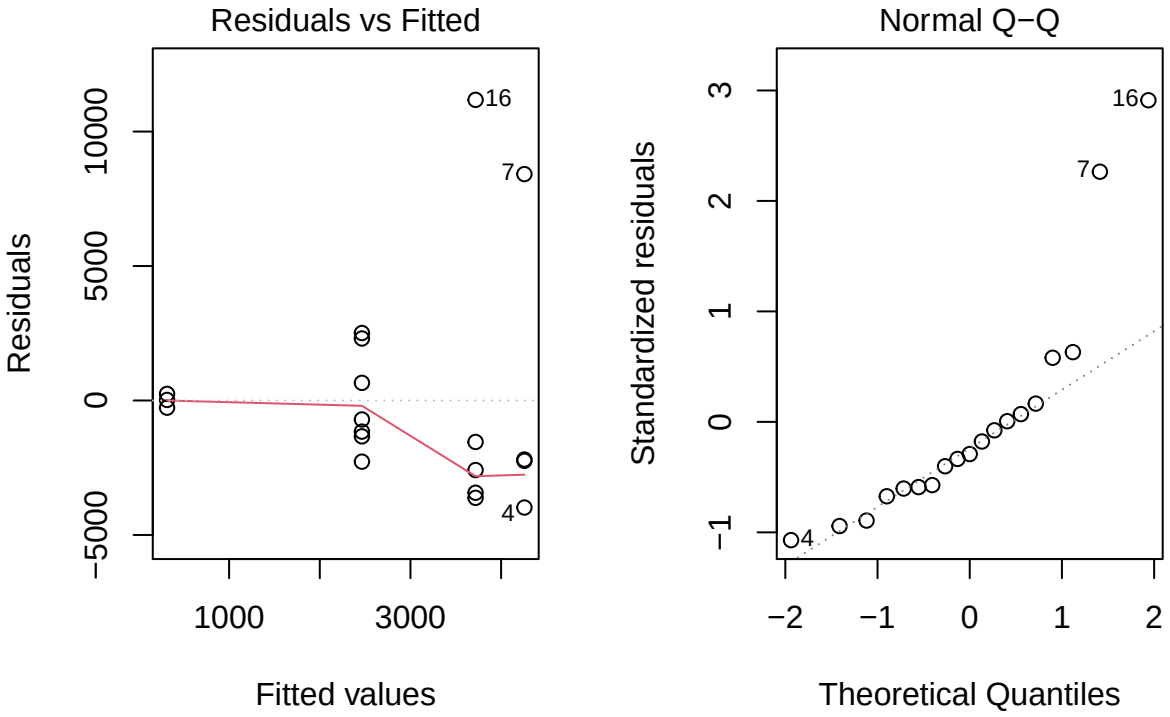
(I) Introduction

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

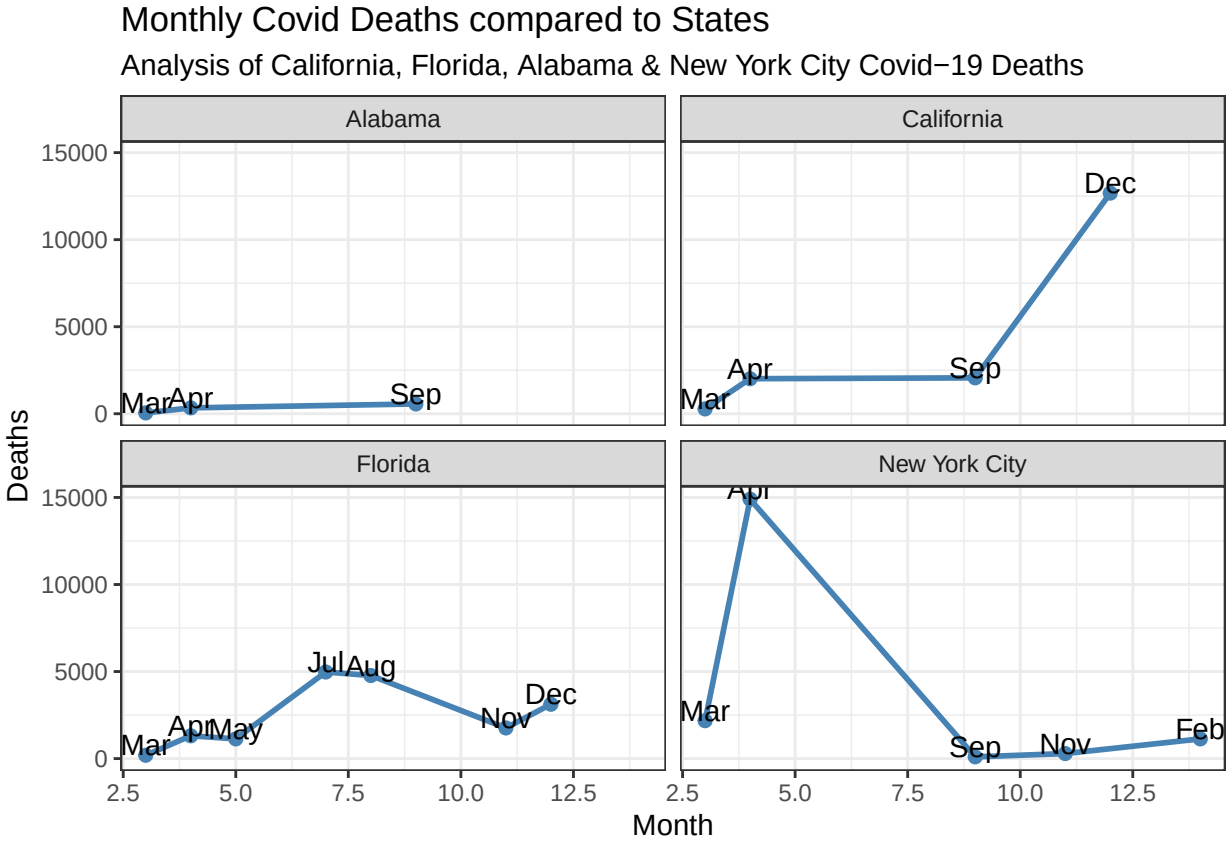
	California	Florida	Alabama	New York City
Group Mean	4256.750	2465.00000	315.666667	3718.200
Group SD	5673.232	1863.53750	257.663217	6302.866
Rank Mean	11.750	11.42857	4.666667	9.800
Sample Size	4.000	7.00000	3.00000	5.000



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.

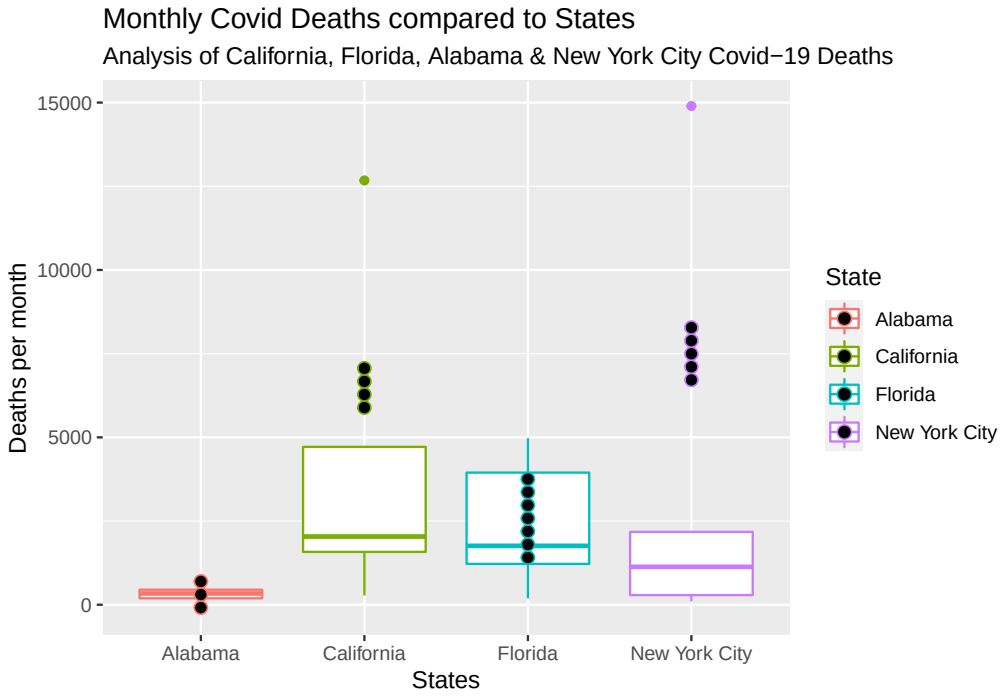
In this forthcoming graph, I purposely ended up changing a dataset column where I made a replica of the subsetted graph that we had from the CovidA dataset. I then changed the months to where month 1 was January of 2020 and month 14 was February of 2021. This helped in viewing the graphs as there was 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.



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:

$$H_0 : F_1(x) = F_2(x) = F_3(x) = F_4(x)$$
$$H_A : F_i(x) \geq F_j(x) \text{ or } F_i(x) \leq F_j(x) \text{ for some i} \neq \text{j}$$

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.

(V) 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 as well as the inaccuracy.

VI

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)
```

```r
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
```