University of Manitoba

The Relationship Between Mask Usage and COVID-19:

Exploring the Relationship between Mask Usage and COVID-19 Infections and Deaths in the United States

Author: Adrien Dinzey

Course: STAT 3380 A01 – Introduction to Nonparametric Statistics

Instructor: Dr. Zenaida Mateo

Date: January 11, 2022

With the world-wide outbreak of the COVID-19 the world was required to adapt and change. As such many governments and health authorities have decided on appropriate health measures to circumscribe the spread of the virus and to ensure the safety of all. Of all measures put in place, one of the simplest also happens to be one of the most controversial. That measure is the simple action of wearing a 3-ply mask or face covering. As citizens, we have been informed by professionals and authorities on the matter that it is an effective health measure to prevent COVID-19 infections, and therefore help lower the amount of COVID-19 related deaths. However, many would argue and say that masks have no effect on the transmission of the COVID-19 virus. This research paper will be using several Nonparametric Statistical methods to examine the relationship between mask usage and COVID-19 infections and deaths.

INTRODUCTION

Wearing masks in public to prevent virus transmission is a very new idea to the public. While there have been many pandemics in the past, such as the Spanish Flu, the idea of government mandated mask usage is a big change for many people. As such, there are many who would oppose it, for various reasons. According to a study by the J Am Med Inform Association, of the tweets that contain opinions about wearing facial masks to prevent the spread of COVID-19, about 10% are “anti-mask” or opposed to the idea entirely (He, L., He, C. et al., 2021). While there are many arguments as to why biologically and physically these opinions are unfounded, perhaps the simplest argument is a data-based and evidence-based argument.

That leads to the purpose of this paper. Using data involving mask usage as well as COVID-19 data, nonparametric statistical methods can be applied to discover the relationship, if any. Overall, this paper will discover if there is a relationship between mask usage and COVID-19 infections and deaths, and if there is, then it will determine if that relationship has a downwards trend. Additionally, this paper will format the findings in a way that is easily understandable to all those who are curious, as these results should be understood by all citizens regardless of education.

METHOD

First off, the data must be obtained. The New York Times has been actively keeping track of total COVID-19 Infections and Deaths by County in the United States. Additionally, they have kept track of mask usage by county. To do that, they polled citizens and asked them to rank on an ordinal scale how often they wear their mask in public ("Covid in the U.S.: Latest Map and Case Count," 2022). The actual data is publicly available through the Google Cloud Public Dataset Program, a marketplace of datasets that are hosted by Google BigQuery for integration to any project. This data was retrieved through the Google BigQuery platform by using simple SQL queries and then exporting the data into a .csv file stored locally for later use. Additionally, one concern of comparing COVID-19 infections with mask usage by county is that the population size in counties varies heavily. As such, the infections and deaths should be normalized for proper comparison. To do this, census data was retrieved from the United States Census Bureau (2021) so that the infections and deaths reported by The New York Times could be transformed into proportions with respect to population. The census data was downloaded from the official United States Census Bureau website as a .csv file.

After obtaining the data, there are four different tables of data, while only one combined table is desired. Additionally, many of the separate tables have columns that are not required for the purpose of this paper. To accomplish the task of combining these tables, a simple ETL (Extract, Transform, Load) script was written to extract the required columns from each table, then transform it so the data is normalized (eg. use state abbreviations instead of state names) and finally load it into one table that can be worked with. One important transformation was the calculation of the Mask Usage value. The data from the Mask Usage poll was taken and for each county a weighted average was calculated based on the reported mask usage (0 - Never, 1- Rarely, 2 – Sometimes, 3- Frequently, 4 – Always). For example, if 100% of the population said they always use their mask, the value would be 1 and if 100% of the population said they never used their mask then that value would be 0.

Once the data processing is complete, the remaining .csv file has relevant data ready to work

with. This table contains a row for 2967 different rows (one row per county) containing the 6 following columns:

- "FIPS Code" - The Federal Information Processing Standard code of the County.
- "County" - The name of the County.
- "State" - The name of the State in which the County exists.
- "Mask Usage" - A decimal number on the range [0,1] where 0 means that on average the county never uses a mask and 1 means that on average the county always uses a mask.
- "Infection Proportion" - Proportion of the total county population that has had a confirmed positive COVID-19 infection.
- "Death Proportion" - Proportion of deaths with respect to the total county population due to COVID-19.

With this data at the ready, it is very easy to perform statistical analysis.

Now that the data is ready, the statistical methods must be selected. Going back to the objectives of this paper, to determine if there is a relationship between mask usage and COVID-19 infections and deaths the counties can be separated into two different populations. The first is the high-mask usage population and the other is the low-mask usage population. We will say "low-mask usage" is anything below what we ranked as "Frequently" in the original data. If a county on average frequently used masks, then it would be given the value 3*(1)/4=0.75. So, any county with a mask usage value less than 0.75 is considered a low mask-usage county and any county above 0.75 is considered a high mask usage county.

This allows for the Mann-Whitney-Wilcoxon Test to be conducted. To conduct such a test, certain assumptions must be true. The two populations are independent, as the cases were reported December 19, 2021, when each county would have already had its own COVID-19 outbreak and the outbreak would be handled according to local guidelines as well as the guidelines of the state and United States government. Additionally, since we are dealing with proportions, the variable of interest is a continuous random variable, and the measurement is the ratio scale. This also means both populations'

distributions differ only with respect to location, if at all. Therefore, this test is valid in this scenario.

Next, to further help determine if such a relationship exists at all, Point-Biserial Correlation Analysis was performed. For some considerations, the y values in this case are the proportions. Since there are almost 3000 different populations, and therefore almost 3000 different sets of proportions (one proportion for infections, one for deaths) according to the Central Limit Theorem these proportions are approximately normally distributed. Additionally, the proportions are continuous random variables and are measured on the ratio scale. As for the other variable, let this be "If the county has a high mask usage". This means, if the mask usage is below 0.75 then this variable will be "low" and if it is 0.75 or above it will be "high". This makes the second variable of interest a dichotomous variable. This will help determine if the presence of the "If the county has a high mask usage" influences COVID-19 infection and death proportions.

Finally, if such a relationship exists then the next objective is to determine if higher mask usage corresponds with lower COVID-19 infection and death proportions. After ordering the rows of data in ascending order with respect to mask usage, Cox Stuart Trend Analysis was performed on both COVID-19 infection proportions and death proportions.

DISCUSSION OF RESULTS

First the data needed to be explored. As such, boxplots of the reported infection proportions and death proportions were generated.

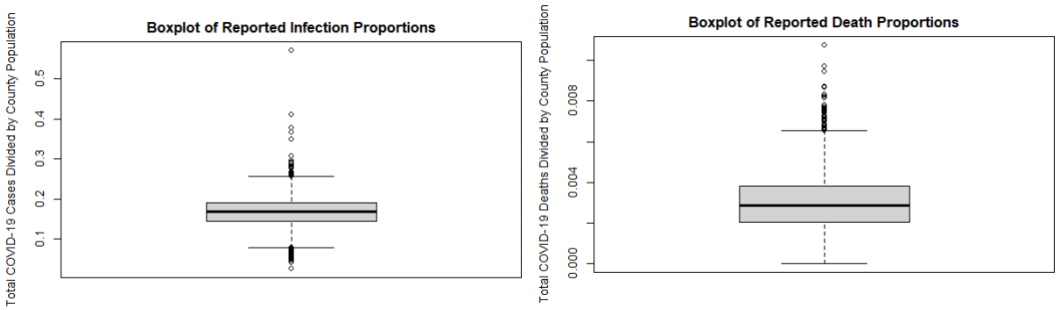

Figure 1: Initial Boxplots of the Proportions

While a trained eye can see that the proportions appear to be normally distributed around the mean, (as stated by central limit theorem) there are a lot of outliers. Removing them yields the following plots.

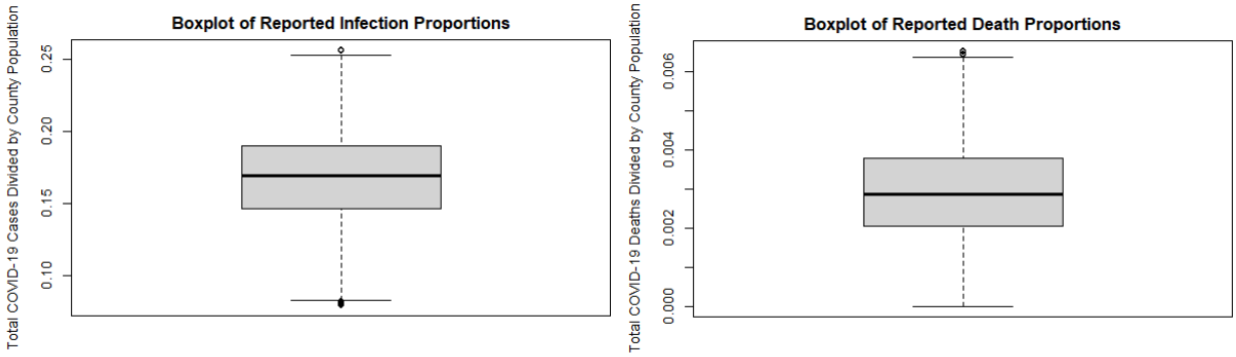

Figure 2: Cleaned Boxplots of the Proportions

Again, the distributions appear to both be normal. The five number summary is as follows

```
[1] "Infection Proportion Five Number Summary"
[1] "Min                First-Quartile    Median            Third-Quartile    Max                "
[1] "0.0261631213741326 0.145632621605356 0.169191630903325 0.190263621868794 0.571651233153021"
[1] ""
[1] "Death Proportion Five Number Summary"
[1] "Min                First-Quartile    Median            Third-Quartile    Max                "
[1] "0.0000000000000000 0.00202080346213624 0.00286578577555527 0.00382483104485087 0.0107671601615074"
```

Figure 3: Five Number Summary

Next the cleaned data is then put into two populations according to mask usage. The Mann-Whitney-Wilcoxon test resulted in extremely small p-values for both tests, so small that the computer could not accurately calculate it. All it could report was that the p-values were less than 2.2E-16. Therefore, in both tests we must reject the null hypotheses and conclude that there is a significant difference in infection proportion and death proportion between the two populations.

```
        Wilcoxon rank sum test with continuity correction

data:  test.df$pop1 and test.df$pop2
W = 1287777, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Figure 4: Mann-Whitney-Wilcoxon results for COVID-19 Infections

```
        Wilcoxon rank sum test with continuity correction

data:  test.df2$pop1 and test.df2$pop2
W = 1105949, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Figure 5: Mann-Whitney-Wilcoxon results for COVID-19 Death proportions

Next, a Point-Biserial Correlation Analysis was conducted to further determine if there is a relationship to be examined between mask usage and COVID-19 infections and deaths. The resulting process yielded weak, negative correlation coefficients. The correlation was about twice as strong for the relationship between mask usage and COVID-19 infections than the correlation between mask usage and COVID-19 deaths. We see it was a correlation coefficient of about -0.3382 for mask usage and infection proportions, and a correlation coefficient of about -0.1637 for mask usage and death proportions.

```
        Pearson's product-moment correlation

data:  MaskData$MaskGroup and MaskData$`Infection Proportion`
t = -19.568, df = 2964, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3697259 -0.3059693
sample estimates:
       cor
-0.3382356
```

Figure 6: Point-Biserial Correlation Analysis on Mask Usage and COVID-19 Infection Proportions

```
        Pearson's product-moment correlation

data:  MaskData$MaskGroup and MaskData$`Death Proportion`
t = -9.0367, df = 2964, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1985657 -0.1285112
sample estimates:
       cor
-0.1637449
```

Figure 7: Point-Biserial Correlation Analysis on Mask Usage and COVID-19 Death Proportions

This can further examine this result with a scatterplot of the proportions against the two mask levels.

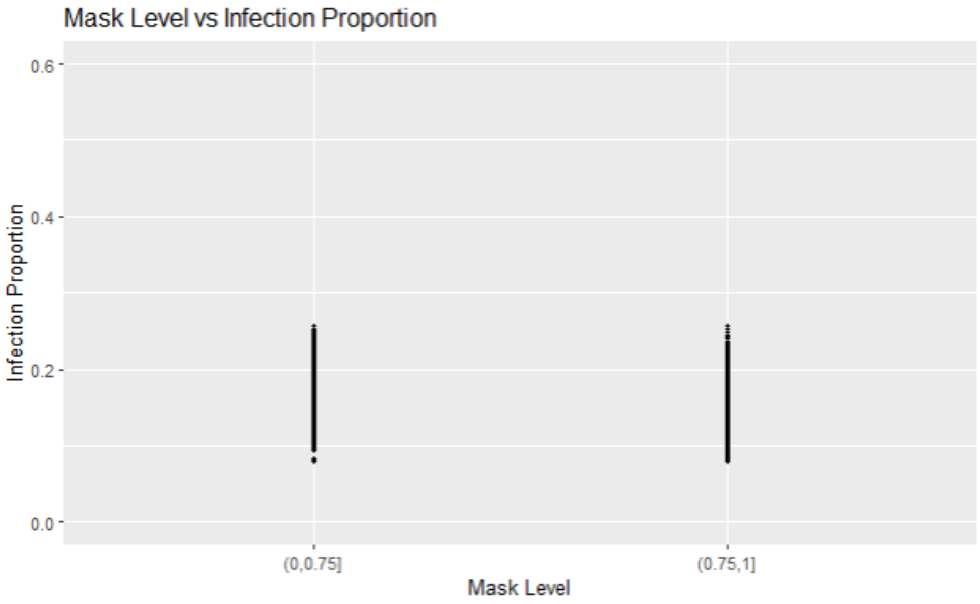

Figure 8: Scatterplot of the Two Populations vs COVID-19 Infection Proportions

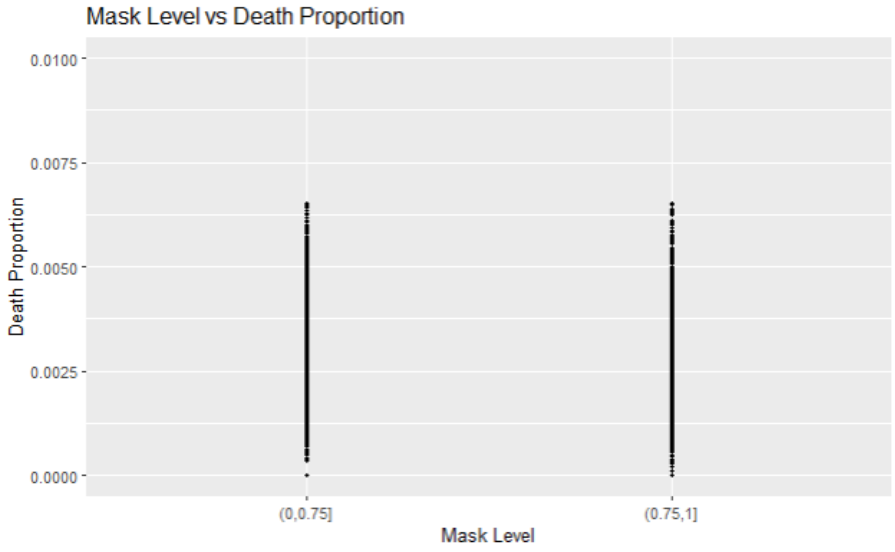

Figure 9: Scatterplot of the Two Populations vs COVID-19 Death Proportions

We see that although there is a clear decrease in infections and deaths, the graphs appear to show little to no trend. That is likely due to the density of the data, while it appears there are many points spread up and down the graph, there are many points that are closer to zero in the high mask usage population so on average, the infections and deaths do in fact decrease in the high mask usage population. The reason this is not very apparent is that there are around 1400 points in each population to be plotted, so there is a lot of overlap. Considering the correlation coefficients, we can conclude that there is in fact a negative correlation between mask usage and COVID-19 infection and death proportions

Finally, the Cox-Stuart Trend Analysis was performed to determine specifically if there is a downwards trend in the data. After ordering the data in ascending order with respect to mask usage values (the values on the range [0,1]), int the results of the one-sided test, P-Values were again so small that we must reject the null hypothesis and conclude that there exists a downwards trend in the data.

```
        Cox Stuart test

data:  sorted_data$`Infection Proportion`
statistic = 414, n = 1422, p-value < 2.2e-16
alternative hypothesis: decreasing trend
```

Figure 10: Cox-Stuart Trend Analysis on COVID-19 Infection Proportions

```
        Cox Stuart test

data:  sorted_data$`Death Proportion`
statistic = 554, n = 1422, p-value < 2.2e-16
alternative hypothesis: decreasing trend
```

Figure 11: Cox-Stuart Trend Analysis on COVID-19 Death Proportions

Graphing a scatterplot of the mask usage values against the COVID-19 infection and death proportions and assigning a line of best fit shows the results of the last two tests very clearly.

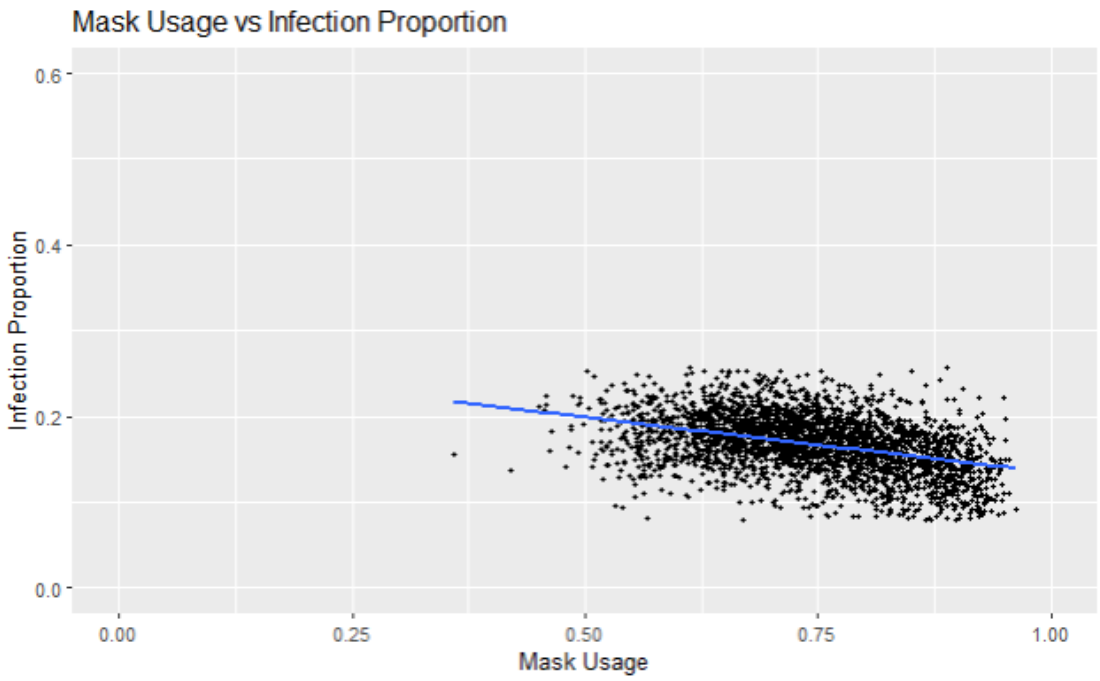

Figure 12: Mask Usage vs Infection Proportion Scatterplot

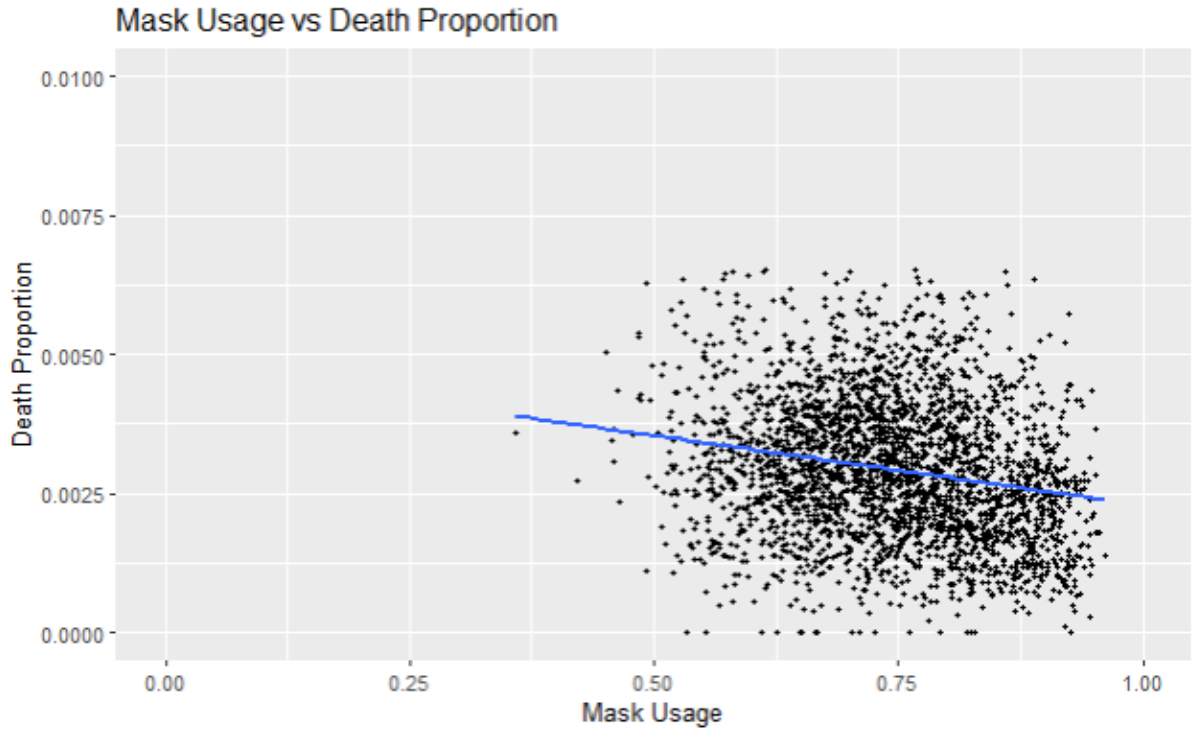

Figure 13: Mask Usage vs Death Proportion Scatterplot

SUMMARY OF RESULTS

In summary, the results clearly show overall that as mask usage increases, the proportion of COVID-19 infections and deaths decreases. While there is variation present, the number of cases where there is a steady decrease overshadows that variation. The first and third test results reported that this decreasing trend exists with extremely high confidence. The second test analyzed exactly how strong that relationship is, which was reflected in the scatterplot. It showed that mask usage at the very least has an impact on lowering COVID-19 cases and deaths because of less cases.

CHALLENGES/FUTURE PLAN

The main challenge with this is the variation between location of the population. This study only considers one preventative measure being put into place, and that is mask usage. This study does not consider other measures being enacted such as social capacity limits, social distancing guidelines and sanitization rules. Not only do these likely have a large affect on COVID-19 infections, but they can also differ greatly from state to state. In the future I would like to further aggregate data to examine by state and try to consider the difference in all preventative measures, not just mask usage.

REFERENCES

Google. (2022, January 10). BigQuery public datasets. Google Cloud. https://cloud.google.com/bigquery/public-data

He, L., He, C., Reynolds, T. L., Bai, Q., Huang, Y., Li, C., Zheng, K., & Chen, Y. (2021). Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic. *Journal of the American Medical Informatics Association : JAMIA, 28(7)*, 1564–1573. https://doi.org/10.1093/jamia/ocab047

United States Census Bureau. (2021, October 8). County Population Totals: 2010–2019. Census.Gov. https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html

Covid in the U.S.: Latest Map and Case Count. (2022, January 11). The New York Times. https://www.nytimes.com/interactive/2021/us/covid-cases.html

APPENDIX

Code to generate box plots (Figures 1&2):

```r
boxplot(MaskData$`Infection Proportion`,main="Boxplot of Reported Infection Proportions",
        ylab="Total COVID-19 Cases Divided by County Population")
outliers1 <- boxplot(MaskData$`Infection Proportion`, plot=FALSE)$out

boxplot(MaskData$`Death Proportion`,main="Boxplot of Reported Death Proportions",
        ylab="Total COVID-19 Deaths Divided by County Population")
outliers2 <- boxplot(MaskData$`Death Proportion`, plot=FALSE)$out
```

After removing outliers we get this.

```r
PrunedData1 <- MaskData
PrunedData1 <- PrunedData1[-which(PrunedData1$'Infection Proportion' %in% outliers1),]
boxplot(PrunedData1$`Infection Proportion`,main="Boxplot of Reported Infection Proportions",
        ylab="Total COVID-19 Cases Divided by County Population")

PrunedData2 <- PrunedData1
PrunedData2 <- PrunedData2[-which(PrunedData2$'Death Proportion' %in% outliers2),]
boxplot(PrunedData2$`Death Proportion`,main="Boxplot of Reported Death Proportions",
        ylab="Total COVID-19 Deaths Divided by County Population")
```

Code to generate five number summary (Figure 3):

```r
print("Infection Proportion Five Number Summary")
print("Min               First-Quartile    Median            Third-Quartile    Max               ")
print(paste(min(MaskData$`Infection Proportion`),
            quantile(MaskData$`Infection Proportion`,0.25),
            median(MaskData$`Infection Proportion`),
            quantile(MaskData$`Infection Proportion`,0.75),
            max(MaskData$`Infection Proportion`)))
print("")
print("Death Proportion Five Number Summary")
print("Min               First-Quartile    Median            Third-Quartile    Max               ")
print(paste(format(round(min(MaskData$`Death Proportion`),16),nsmall=16),
            quantile(MaskData$`Death Proportion`,0.25),
            median(MaskData$`Death Proportion`),
            quantile(MaskData$`Death Proportion`,0.75),
            max(MaskData$`Death Proportion`)))
```

Code to split data into two populations:

```r
MaskData<-PrunedData2
# Split into two different tables and ensure the column names have not changed
MaskData1 <- split(MaskData,MaskData$MaskLevel)
lowMaskPop<-data.frame(MaskData1[1])
highMaskPop<-data.frame(MaskData1[2])
names(lowMaskPop)<-names(df)
names(highMaskPop)<-names(df)

rows<-sample(nrow(lowMaskPop))
lowMaskPop<- lowMaskPop[rows,]
diff<-length(lowMaskPop$County)-length(highMaskPop$County)

lowMaskPop<-head(lowMaskPop,-diff)
#ignore some random counties so that the two populations have an equal number of observations
plotData <- MaskData
names(plotData)<-c("FIPS_CODE","COUNTY","STATE","MASK_USAGE","INFECTION_PROP","DEATH_PROP","MASK_LEVEL")
```

Code to run Mann-Whitney-Wilcoxon Test with infection proportions (Figure 4):

```r

test.df <- data.frame(lowMaskPop$'Infection Proportion',highMaskPop$'Infection Proportion')
colNames<-c("pop1","pop2")
names(test.df)<-colNames
testResults<-wilcox.test(test.df$pop1,test.df$pop2)

print(testResults)

```

Code to run Mann-Whitney-Wilcoxon Test with death proportions (Figure 5):

```r
test.df2<- data.frame(lowMaskPop$'Death Proportion',highMaskPop$'Death Proportion')
colNames<-c("pop1","pop2")
names(test.df2)<-colNames
testResults2<-wilcox.test(test.df2$pop1,test.df2$pop2)
print(testResults2)
```

Code to run Point-Biserial Correlation Analysis with infection proportions (Figure 6):

```r
# First we will add a binary column where 0 means the low group and 1 means the high group.
MaskData$MaskGroup <- 1
for(x in c(1:length(MaskData$`Infection Proportion`))){
  if(MaskData[x,4]<0.75){
    MaskData[x,8]<-0
  }
}

# Now we perform the test
corResults1<-cor.test(MaskData$MaskGroup,MaskData$`Infection Proportion`)
print(corResults1)


```

Code to run Point-Biserial Correlation Analysis with death proportions (Figure 7):

```r
corResults2<-cor.test(MaskData$MaskGroup,MaskData$`Death Proportion`)
print(corResults2)
```

Code to generate scatterplots of the two groups vs infection and death proportions (Figure 8&9):

```r
ggplot(plotData, aes(x=MASK_LEVEL,y=INFECTION_PROP)) +
  geom_point(size=1)+
  geom_smooth(method=lm,se=FALSE) +
  scale_x_discrete(name="Mask Level")+
  scale_y_continuous(name="Infection Proportion",limits = c(0,0.6)) +
  ggtitle("Mask Level vs Infection Proportion")
ggplot(plotData, aes(x=MASK_LEVEL,y=DEATH_PROP)) +
  geom_point(size=1)+
  geom_smooth(method=lm,se=FALSE) +
  scale_x_discrete(name="Mask Level")+
  scale_y_continuous(name="Death Proportion",limits = c(0,0.01)) +
  ggtitle("Mask Level vs Death Proportion")
```

Code to sort data and run the Cox-Stuart Trend Analysis on infection proportions (Figure 10):

```r
sorted_data <- MaskData[order(MaskData$'Mask Usage'),]
cox.stuart.test(sorted_data$`Infection Proportion`,alternative = "left.sided")
```

Code to run the Cox-Stuart Trend Analysis on death proportions (Figure 11):

```r
cox.stuart.test(sorted_data$`Death Proportion`,alternative = "left.sided")
```

Code to generate the scatterplot of Mask Usage vs Infection Proportions and Mask Usage vs Death Proportions (Figure 12&13):

```r

ggplot(plotData, aes(x=MASK_USAGE,y=INFECTION_PROP)) +
  geom_point(size=1)+
  geom_smooth(method=lm,se=FALSE) +
  scale_x_continuous(name="Mask Usage",limits = c(0,1))+
  scale_y_continuous(name="Infection Proportion",limits = c(0,0.6)) +
  ggtitle("Mask Usage vs Infection Proportion")

ggplot(plotData, aes(x=MASK_USAGE,y=DEATH_PROP)) +
  geom_point(size=1)+
  geom_smooth(method=lm,se=FALSE) +
  scale_x_continuous(name="Mask Usage",limits = c(0,1))+
  scale_y_continuous(name="Death Proportion",limits = c(0,0.01)) +
  ggtitle("Mask Usage vs Death Proportion")
```