# Coronavirus in the Southeast

STAT 3480 Consulting Project

Ishaan Dey & Tal Dunne

11/23/2020

---

## Executive Summary

# Introduction

## Project Description

The coronavirus pandemic has taken the United States by storm - it is important that we understand as much as possible about the disease and its effects. As such, it is crucial that we run data analytics and obtain predictions and inferences about existing information. Given the airborne transmission of COVID, it makes sense to operate on a regional level. The focus of this report is on the Southeastern United States, defined as Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, Missouri, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia. These thirteen states have implemented a variety of measures, including mask mandates or lack thereof. The intent of this report is to analyze the response to COVID in these areas. The main research focuses on the effects of the mask mandate on total cases per 1,000 people and the connection between a lower number of cases per 1,000 people and the death rate.

The CDC has already looked into the effect of the mask mandate in Kansas, where the government implemented the legislation with a county opt-out option. Of the 24 counties that implemented a mask mandate, all of them showed a decrease in COVID incidence. In contrast, the COVID incidence in the 81 counties without the mandate showed a continuing increase. As such, the researchers at the CDC came to the conclusion that wearing face masks in public helps to reduce the spread of the virus. Another study that has been conducted on the topic used the mask mandates in 15 states plus Washington D.C. to come to the conclusion that the states with mandates had a greater decline in daily COVID growth rates compared to the states that did not issue mandates.

## Data Summary & Discussion

From public CDC records, several variables of interest were obtained, with total cases and total deaths per state recorded as of October 27th, 2020. We extract the following features relevant to answer our research questions: `State`, `Mask.Mandate`, `Death.Rate.Pct`, and `Cases.Per.1000`. `Mask.mandate` is an indicator variable for if the state ever adopted state-wide mask mandate measures. `Death.Rate.Pct`, the percent of cases resulting in a fatality, is calculated as the total number of deaths by total number of cases. `Cases.Per.1000`, or rate of cases in the general population, was obtained by normalizing total case counts to 2019 Census estimates of state populations.

There are several limitations to this approach. Mask mandates are only one of the many possible public-health measures enacted by the states, therefore we expect several unmeasured confounders to impact the total case counts (as opposed to examining data at a per-county level within a state). Furthermore, case count reports are understood to underestimate actual counts, by as much as 40%, given large volumes of asymptomatic cases and scarcity of testing resources early in the pandemic.

Our data represents a random sample of 13 states, with each observation at the per-state level. For each of these states, we obtain 4 variables as defined above. Of the 13 states, 7 have adopted mask mandates, 6 have not. The average case rate is 29.59 cases per 1000 people, ranging from 12.67 in West Virginia to 39.74 in Louisiana. Figure 1 describes the univariate distribution of case rates. The average death rate in these same states is 2.03%, ranging from 1.27% in Tennessee to 3.18% in Louisiana. Figure 2 describes the univariate distribution of death rates in Southeast states.

# Methodology

## Mask Mandate

The first research question is as follows: "Do states with a mask mandate tend to have fewer total cases per 1,000 people?". In the following section, we will select the best methodology for analyzing this question.

The client has hypothesized that the states with masks mandates see a lower number of total cases per 1,000. Given this starting point, we can begin our analysis and select an appropriate test. We have two distinct groups of observations - states with mask mandates (total of 7) and states without mask mandates (total of 6). The goal is to test if the distribution of the number of cases per 1,000 people is significantly different between the two groups. We need to conduct a two-sample inference procedure, which means that we are comparing one variable between two populations of interest. In this case, the variable we test is the number of cases per 1,000 for each observation, and the two populations are the two sets of states - with or without the mask mandate, respectively.

With regards to the procedure of interest, the analysis will proceed using the Wilcoxon Rank-Sum Test. This test procedure allows us to compare two distributions in order to determine if there is a significant difference between them. In contrast to a number of other inference procedures, the Wilcoxon test is applied to the ranks of the observations, rather than the data itself. Ranks are a fairly straightforward concept - the smallest value is assigned a rank of 1, the second smallest is assigned a rank of 2, and so on. In using ranks instead of the actual data, the test accounts for distributions that may not follow the traditional bell-shaped curve.

The Wilcoxon Rank-Sum test is the best method to use in this instance for a number of reasons. First and foremost, the only requirement to use the test is that both distributions are continuous. This assumption is satisfied because the number of cases per 1,000 people is a continuous variable - it can be any number greater than zero, including decimals. There are no assumptions about the shape of the distribution for the Wilcoxon test, which is perfect for this instance given such small sample sizes. We can observe the shape of the two distributions in figure 3. The distribution for the mask mandate states especially is very skewed and heavy-tailed, which means it differs greatly from the standard bell-shaped curve. Given the analysis of interest, the continuous distributions, and the relative skewness of those distributions, the Wilcoxon Rank-Sum Test is the best methodology for our procedure.

The procedure itself revolves around the aforementioned ranks of observations. The hypothesis that the mask mandate decreases the number of cases indicates that we expect the observations of states with mask mandates to have lower ranks. As such, we expect to see that the population of states with mask mandates has smaller ranks than the population of states without mask mandates. From this starting point, we can proceed and calculate the test statistic. The test statistic, in this case, is just the sum of all ranks for the population of states with mask mandates. After assigning ranks from smallest to largest, we look at the 7 states which implemented a mask mandate and add up their respective ranks. This number, called the test statistic, is important for determining if the result of the test is significant or not.

The last step of the procedure is to run a permutation. A permutation is a method of running repeated trials on the data set, changing the observations each time in order to assess the validity of our conclusions. In the Wilcoxon Rank-Sum Test, we run a permutation by shuffling all the ranks of observations and assigning them to the respective groups of 7 (mask mandate) and 6 (no mask mandate). For the purposes of the permutation, it does not matter where the data get sorted - they will all end up in each group. On paper, this sounds like a complex and difficult task, but fortunately it is relatively standard and easy to run. There are a total of 1,716 possible permutations, or ways to sort the data into the groups. This is not too many and will not require a large amount of computation power. As such, we will find all of the permutations. For every permutation, we calculate the sum of ranks for the mask mandate group. We keep track of the number of permutations for which this sum is less than or equal to the test statistic. Lastly, we divide this number by 1,716 to find the proportion of tests that resulted in an outcome as extreme or more extreme as our original data set. This proportion is called the p-value.

A lot of information on the Wilcoxon Rank-Sum Test procedure was just given - the results of this test will be given in the Results section and analyzed in the Discussion section.

## Fatality

The second research question is as follows: "do states with a lower number of cases per 1,000 people also have a lower death rate?" In the following section, we will select the best methodology for analyzing this question.

The client has hypothesized that there is a relationship between a lower number of cases per 1,000 people and a lower death rate. Their reasoning is that a smaller amount of cases would put less strain on the medical resources for that state. This, in turn, would cause more time and effort to be devoted to each case, which would hopefully result in a lower death rate. From this starting point, we can evaluate if there is a relationship between the two metrics. The strength of this association is going to be measured by finding the Spearman's Rank Correlation.

Spearman's Rank Correlation is a procedure that helps to find nonlinear association. figure 4 shows a scatterplot of the number of cases per 1,000 people compared to the death rate. As you can see, the plot is sporadic and does not appear to follow a linear relationship. As such, we need to use a procedure that accounts for nonlinear relationships. Additionally, Spearman's Rank Correlation is the correct test to use because all of the other assumptions are met. We have a selection of paired observations - the death rate and number of cases per 1,000 people for each state. We can also assume that the error terms are independently and identically distributed around zero. This means that the error bars around each point have the same probability distribution, and that they are all independent - they have no connection to each other. There is actually another test methodology we could use here, called Kendall's Tau. However, due to the rank-based procedure conducted in section 3a, it made more sense to continue the rank-based theme of this report and use Spearman's Rank Correlation.

Spearman's Correlation is very similar to the traditional statistical notion of correlation. However, instead of directly finding the association between the two variables, we are going to find the association between the ranks of the two variables. As such, the first step is to assign ranks, smallest to largest, for both samples (death rate and number of cases). For more information about ranks, please see section 3a. The first step of the procedure is to calculate the test statistic, or Spearman's Correlation itself. This is a relatively straightforward calculation that uses the ranks of the observations. From there, we use a permutation based procedure to reshuffle the observations and calculate a p-value. This procedure is very similar to the one conducted in part 3a, so for more information, please see above. However, in this case, there are a total of 6,227,020,800 possible permutations, so the test will run using a random sample of 2000. This is standard in the statistical field and will produce coherent results while limiting computing time. It is important to note that we expect to see a positive correlation between the two variables, which is in accordance with the client's hypothesis. In the next section, the results of the Spearman's Correlation procedure will be given.

# Results

## Mask Mandate

In performing the *Wilcoxon Rank Sum Test*, we obtain the following results:

| Test-Statistic | p-value |
|---|---|
| $W = 14$ | 0.183 |

Given that our p-value of 0.183 is less than $\alpha$ of 0.05, we *fail to reject* $H_0$ that the true difference between our samples is 0.

## Fatality

We find that Spearman's Correlation $\rho$ is 0.302.

In performing the *Spearman's Correlation Permutation Test*, we obtain the following results:

| Test-Statistic | p-value |
|---|---|
| $\rho = 0.302$ | 0.158 |

Given that our p-value of 0.158 is less than $\alpha$ of 0.05, we *fail to reject* $H_0$ that the true correlation between the features is 0.

See Appendix A: Code for relevant code on all reported values.

# Discussion

## Mask Mandate

The data was aggregated and run through the Wilcoxon Rank-Sum test. The results of the procedure give a p-value of 0.183. This means that during repeated permutation trials, we have an 18.3% chance of seeing a result as extreme or more extreme than our original data set. Traditionally, we say that we have a statistically significant result when this percentage is 5% or less. As such, we do not have evidence that the result is significant. In real world terms, this means we do not have enough evidence to reject the notion that there is no difference between the number of cases in mask mandate states versus no mask mandate states. The evidence does not point in favor of the client's hypothesis.

However, this result should not be alarming. There are a number of reasons why we have failed to make a significant conclusion in favor of the hypothesis. The sample size for this test was very small - only 13 states in total. The research done by the CDC on the mask mandates in Kansas had a total sample size of 105 counties. The Kansas research showed a statistically significant decrease in COVID cases with mask mandates. However, the sample in that experiment is substantially larger than the sample used in our hypothesis testing. Due to the very small sample size, it makes sense that we would observe an insignificant result - further testing needs to be conducted using more data in order to accurately make a claim. Another potential limitation of this study is that we are taking a top-down approach to case numbers. Within each of the 13 southeast states, there are rural areas and urban areas. The urban areas, with higher population density, will naturally have a higher incidence of COVID cases. As such, it would make more sense to analyze these states at the county level. This way, we would potentially have a large enough sample size to make an accurate conclusion. Additionally, we could provide insights to specific local municipalities as to how they can improve their responses to the virus.

In summation, we cannot make the claim that a mask mandate reduces the number of cases per 1,000 people in the 13 southeastern states.

## Fatality

The data was aggregated and run through the Spearman's Rank Correlation procedure. The results of the test give a p-value of 0.154, or 15.4%. This indicates that we have a 15.4% chance of observing a result as extreme or more extreme than our original data set - again, not low enough to claim statistical significance. As such, it is important to understand why the hypothesis could not be verified.

As with the mask mandate test, the sample size is simply too small to make any accurate conclusions. Moving to the county level in these states would provide a better understanding of how the death rate relates to the number of cases per 1,000 people. One other thing that we do not consider is population demographics. For example, it is known that there is a large elderly population in Florida. If the high-risk categories were to be trusted, we would expect to see a higher number of deaths in this area. Disregarding both the county level information and the demographic information reduces the capacity of this test to make accurate conclusions. Additionally, one of the limitations of the Spearman's Rank Correlation is that it only accounts for a monotonic trend. That is, it only can find a trend in one direction, either positively or negatively. While that may or may not have impacted the results of this test, it is something to note as the true relationship between these variables may be parabolic in nature.

In short, we cannot make the claim that states with a lower number of cases per 1,000 people also have a lower death rate.

## Conclusion

The client at the CDC posed two research questions, one relating to the effectiveness of mask mandates and one relating to the association between case incidence and death rate. More specifically, these questions were posed for 13 states in the southeast. For the question of mask mandates, we do not observe significant results in favor of the notion that the states with a mask mandate observe a lower number of cases. For the question of association, we do not observe significant results in favor of the notion that there is a positive correlation between number of cases and death rate. It is important to note that we do not have evidence to outright reject these hypotheses, but rather we fail to reject their alternatives. Further research is required in order to continue to understand the statistics behind the coronavirus.

For future research, it is imperative that the sample size is increased instead of the state level, our research team recommends investigating trends at the county level. A large portion of the reason we were unable to verify the client's hypothesis is that the data set is too small. As such, the client's research questions are still incredibly valid, but they require further research to accurately make conclusive claims.

# Appendix

## A. Code

Load in data processing and statistical libraries

```
library(readr)
library(dplyr)
source("http://www4.stat.ncsu.edu/~lu/ST505/Rcode/functions-Ch5.R")
```

Load in raw data

```
covid_data <- readr::read_csv("covid-data.csv") %>% as_tibble()
covid_data$Mask.Mandate <- dplyr::if_else(covid_data$Mask.Mandate.Start.Date == 'None',
                                          true=0, false=1) %>% as.factor()
covid_data %>%
  dplyr::select(c(State, Mask.Mandate, Cases.Per.1000, Death.Rate.Pct)) %>%
  head()
```

```
## # A tibble: 6 x 4
##   State     Mask.Mandate Cases.Per.1000 Death.Rate.Pct
##   <chr>     <fct>                 <dbl>          <dbl>
## 1 Alabama   1                      38.0           1.55
## 2 Arkansas  1                      35.7           1.72
## 3 Florida   0                      36.1           2.13
## 4 Georgia   0                      33.3           2.22
## 5 Kentucky  1                      22.3           1.43
## 6 Louisiana 1                      39.7           3.18
```

### Data Exploration

**Case Rate** 5 Number Summary

```
fivenum(covid_data$Cases.Per.1000) %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.67   25.16   33.52   29.59   36.87   39.74
```

Code for Figure 1

```
hist(covid_data$Cases.Per.1000, xlab='Case Rate Per 1000', main='', ylab='No. of States')
```

**Death Rate** 5 Number Summary

```
fivenum(covid_data$Death.Rate.Pct) %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.270   1.600   1.900   2.034   2.220   3.180
```

Code for Figure 2

```
hist(covid_data$Death.Rate.Pct, xlab='Death Rate Per Case (%)', main='', ylab='No. of States')
```

```
covid_data$Mask.Mandate %>% recode(`1`="Mandate",`0`="No Mandate") %>% table()
```

**State Mask Mandates**

```
## .
## No Mandate    Mandate
##          6          7
```

**Mask Mandate**

Pull case rates for states with mandates, and those without.

```
mask.mandate <- covid_data %>%
  filter(Mask.Mandate == 1) %>% pull(Cases.Per.1000)
no.mask.mandate <- covid_data %>%
  filter(Mask.Mandate == 0) %>% pull(Cases.Per.1000)
```

Code for figure 3

```
par(mfrow=c(2,1))
breaks <- seq(10, 40, 2)
hist(mask.mandate, xlab='', main='Mask Mandated',
     ylab='No. of States', breaks = breaks)
hist(no.mask.mandate, xlab='Case Rate Per 1000', main='Mask Not Mandated',
     ylab='No. of States',breaks = breaks)
```

Perform Wilcoxon Rank Sum Test

```
wilcox.test(mask.mandate, no.mask.mandate, alternative="less")
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  mask.mandate and no.mask.mandate
## W = 14, p-value = 0.183
## alternative hypothesis: true location shift is less than 0
```

**Fatality**

Pull case rates and death rates

```
case.rate <- covid_data$Cases.Per.1000
death.rate <- covid_data$Death.Rate.Pct
```

Code for figure 4

```r
plot(case.rate, death.rate, main='Scatterplot of Death Rate (%) against Case Rates (per 1000) ',
     xlab='Case Rates (per 1000)', ylab = 'Death Rate (%)')
```

Identify Spearman's correlation, perform permutation test

```r
r.obs<-cor(rank(case.rate), rank(death.rate))

perm.r<-perm.approx.r(rank(case.rate),rank(death.rate),2000)

p.upper<-mean(perm.r >= r.obs)

paste("Spearman's Correlation:", round(r.obs,4)) %>% print()
```

```
## [1] "Spearman's Correlation: 0.3022"
```

```r
paste("p-value:", round(p.upper,3)) %>% print()
```

```
## [1] "p-value: 0.154"
```

## B. Figures

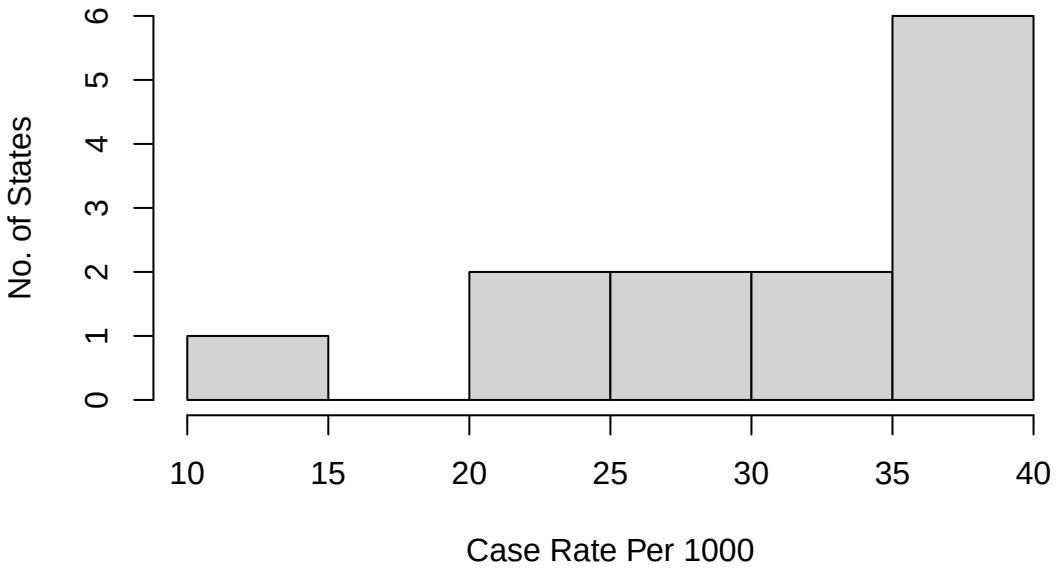


Figure 1: Univariate distribution of case rates.

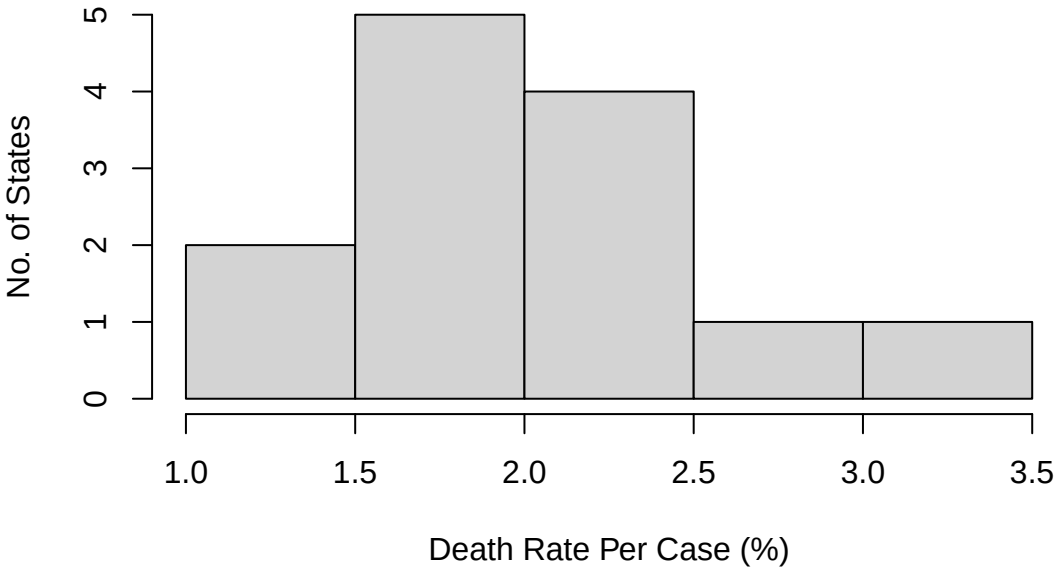


Figure 2: Univariate distribution of death rates.

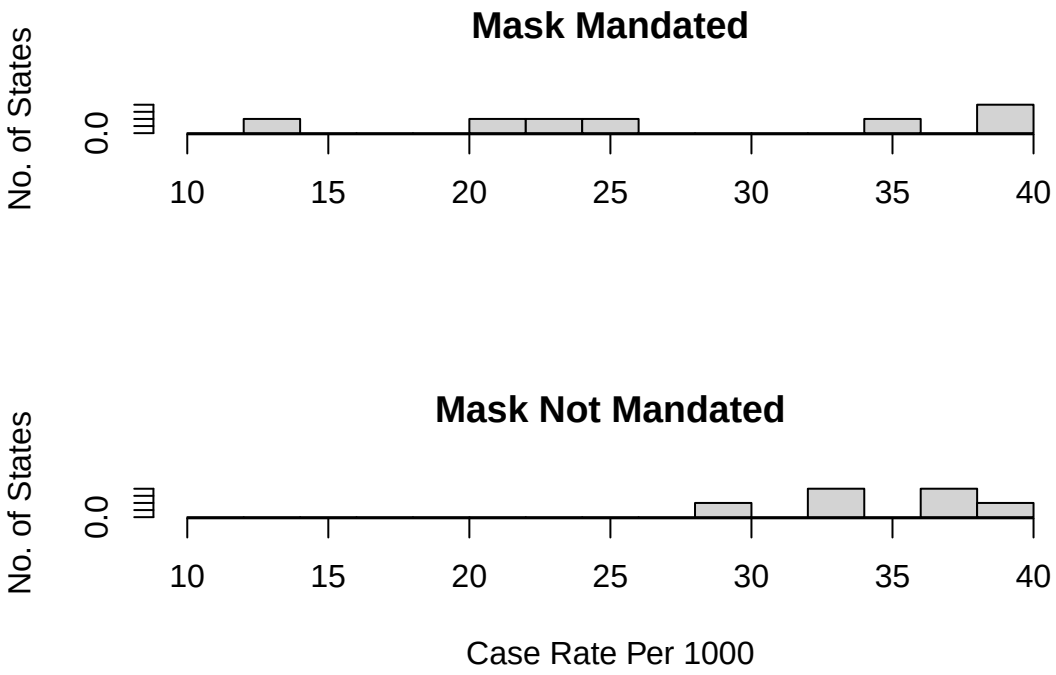


Figure 3: Distribution of Case Rates (per 1000), by State Mask Mandates.

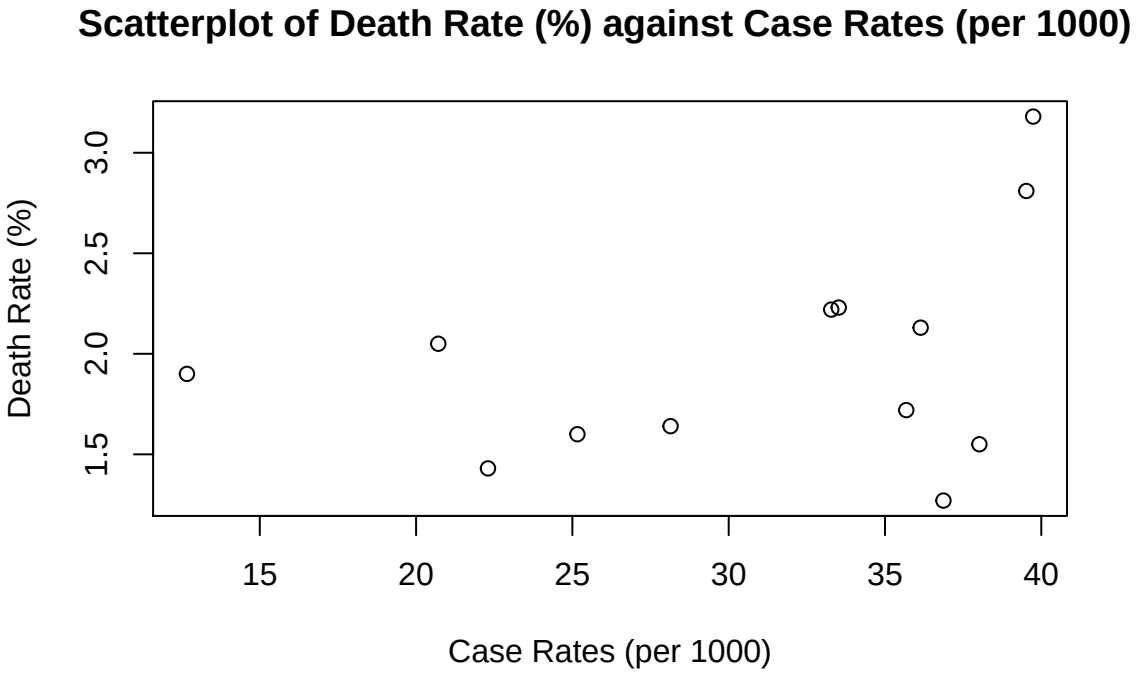


Figure 4: Scatterplot of Death Rate (%) against Case Rates (per 1000)

# References

Van Dyke ME, Rogers TM, Pevzner E, et al. Trends in County-Level COVID-19 Incidence in Counties With and Without a Mask Mandate — Kansas, June 1–August 23, 2020. MMWR Morb Mortal Wkly Rep. ePub: 20 November 2020. DOI: http://dx.doi.org/10.15585/mmwr.mm6947e2.

Lyu W, Wehby GL. Community Use of Face Masks And COVID-19: Evidence From A Natural Experiment Of State Mandates In The US. Health Affairs. Vol. 39. ePub: 16 June 2020. DOI: https://doi.org/10.1377/hlthaff.2020.00818.

Jenkins, Holman W. "It's Biden's Virus Now." The Wall Street Journal, 10 Nov. 2020.