Michigan Tech

MA 5701 Group Project

COVID-19 and Pandemic Impacts

By: Ezequiel Carrillo, Evan Neibauer, and Samantha Hair

MA 5701: Statistical Methods

Dr. Xiao Zhang

Dec 12, 2021

Intro

Given the current global situation, the study of relationships between variables pertinent to a pandemic is more relevant than ever. Particularly in the U.S, as the COVID-19 crisis has progressed, large amounts of data from individual states has been collected, by the *Covid Tracking Project* (covidactnow.org). The site has received data on an ongoing basis, as states continue to report COVID-19 information. The data has been collected from hospitals, vaccination sites, and testing sites for each state, therefore, a sampling unit is already in a format that is fit for analysis. Since this data has simply been collected, with no controlled variables, this is an observational study. The target population is all 50 states, with a sampling unit being a single state. For the analysis, a random selection of 20 states will be used as the sample population. The variables that will be useful in the study will be overall risk level, deaths from covid, cases of covid, hospital bed capacity, and the ratio of completed vaccines. The categorical variable, overall risk level, will be measured by the following levels: 0 = Low, 1 = Medium, 2 = High, 3 = Critical, 4 = Unknown, 5 = Extreme. The levels have been calculated as a result of case density, test positive ratio, and infection rate. The three discrete numerical variables are deaths, cases, and hospital bed capacity. Deaths will be measured in the number of people who are suspected or confirmed to have passed from covid. Cases will be measured in the cumulative number of suspected or confirmed cases. Hospital bed capacity will be measured in the number of current staffed acute bed capacity. The continuous variable, vaccinations completed ratio, is the ratio of the sampling unit's population that has been vaccinated. Analysis will be computed for us to understand the reasons that COVID-19 may be creating risk, and the outcomes that have been a result of COVID-19.

Methods

The sampling unit in our dataset is a given U.S State. Within each sampling unit we have several quantitative variables that describe the vaccination rate, total deaths, hospital bed capacity, etc., for a given state. The nature of our study is observational, as the data has already been collected for us by the *Covid Tracking Project* (covidactnow.org). The variables that will be useful in the study will be overall risk level (levels 0-5), deaths from covid (number of people), cases of covid (cumulative number of cases), hospital bed capacity (current staffed beds), and the ratio of completed vaccines (ratio of sampling units' population that has been vaccinated). Our observed data is housed and continuously updated by *Covid Tracking Project*. In order to clean the data, we excluded U.S territories from the data set, since our study is limited to the scope of U.S' States. Furthermore, we decided to limit our sample size to 20, which we deemed to be modest and reasonable. We implemented our random sampling scheme in RStudio, where we randomly selected 20 states using the sample() function to generate a random sample of 20 numbers from 1 to 50, without replacement. The random sample has been determined using a set.seed() for others to reproduce the study. Each random number represents the index for a state in the dataset. The sample data set was then indexed from the original dataset by using the random sample.

Results

Analysis #1: Simple Linear Regression on Deaths due to Covid-19 and Hospital Bed Capacity

The first analysis done on this data was a simple linear regression between deaths from Covid, and hospital bed capacity. Our independent variable will be the hospital bed capacity, and the response variable will be deaths due to Covid. The hypothesis test was:

$$H0: \beta 1 = 0 \text{ against } HA: \beta 1 \neq 0$$

In plain words, does the hospital bed capacity play a role in determining deaths due to covid. For our significance value, we will be using α = .05 When checking for a relationship between the two variables, both deaths and hospital bed capacity have been divided by population in order to create ratios. First, we had to check our assumptions. A Q-Q plot was generated to determine whether or not the variables were distributed normally or not (Figure 1). Based on the Q-Q plot of the residuals of the model, we determined that both variables do follow a normal distribution. Homoscedasticity was assumed based on the plot of residuals against the fitted values shown in Figure 2. In Figure 3, we can view the scatter plot between Covid deaths and hospital bed capacity, and we see that a linear model would be appropriate to use. Since all assumptions were met, no additional transformations are required. In our scatterplot, we see that the plot has a correlation coefficient of .4657. This means that we have a weak positive correlation between the two variables. When we review the summary output of our model (Figure 4), we see that a simple linear model had a p-value of .1275. Since we have a p-value that is greater than or significant value, we do not reject the null hypothesis. The hospital bed capacity does not play a significant role in determining deaths from covid.

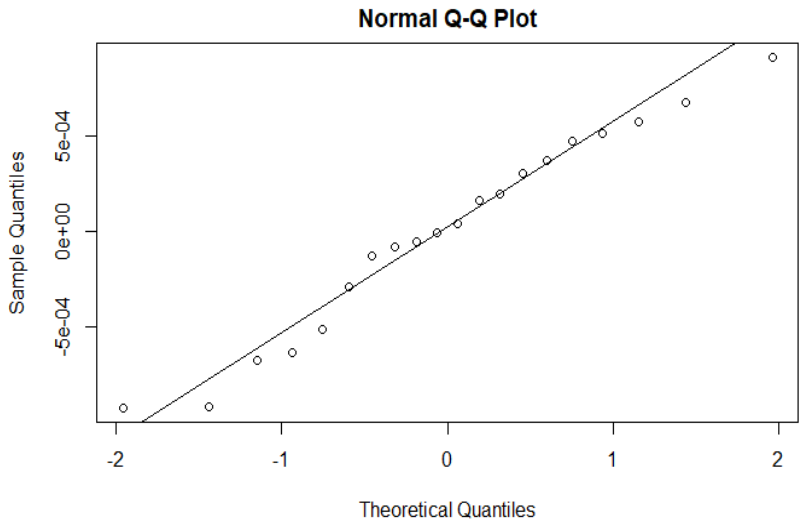


Figure 1. Residual Q-Q plot on the simple linear regression model.

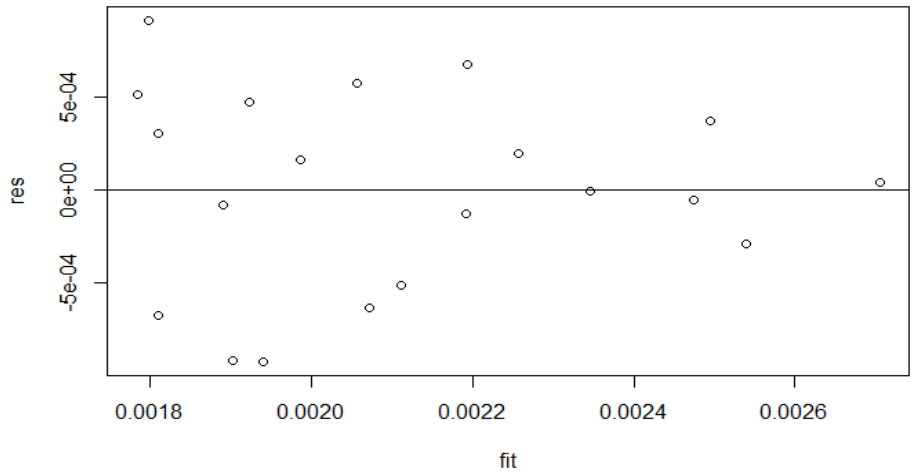
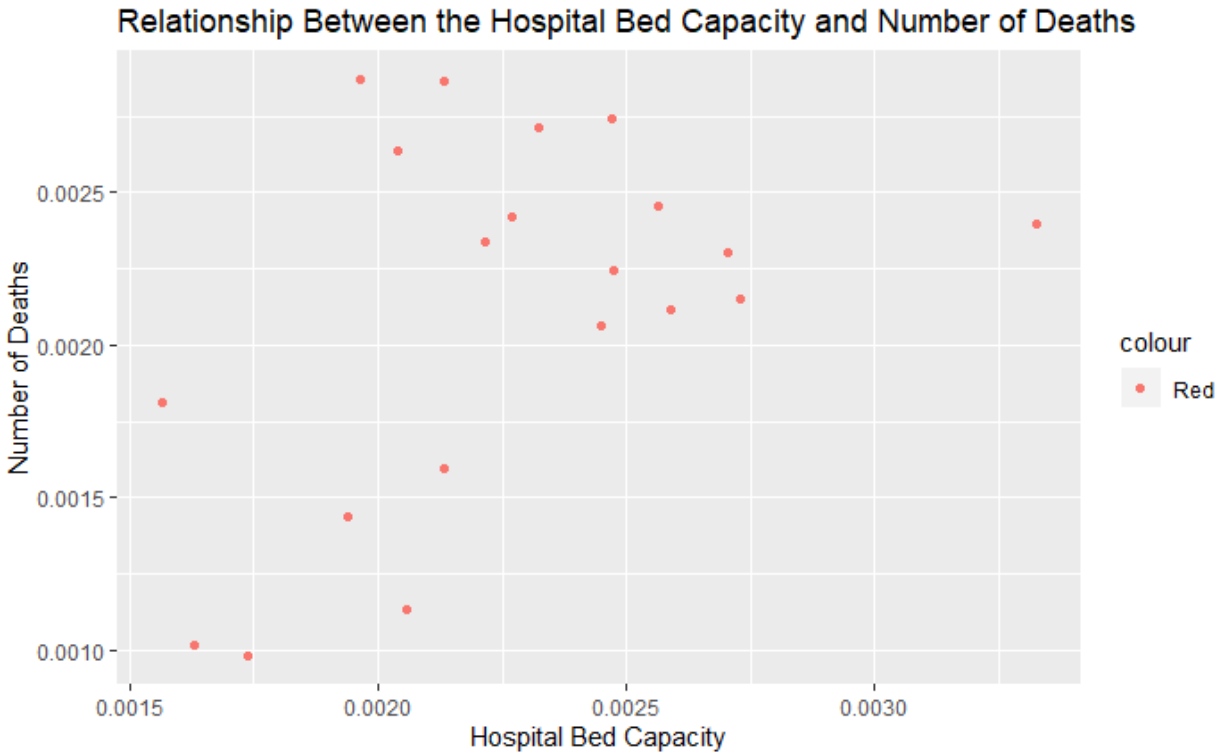


Figure 2. Residuals against fitted values to determine homoscedasticity.



"correlation coefficient between actuals.hospitalBeds.capacity and actuals.deaths:"
[1] 0.4657321

Figure 3. Scatterplot between hospital bed capacity and deaths from Covid.

```
# Check p-value for hypothesis test
summary(lm.model)
```

```
##
## Call:
## lm(formula = (actuals.deaths/population) ~ (actuals.hospitalBeds.capacity/population))
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -9.231e-04 -3.477e-04  1.478e-05  3.987e-04  9.143e-04
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               1.708e-03  2.269e-04   7.529 8.25e-07
## actuals.hospitalBeds.capacity             3.763e-08  1.801e-08   2.090    0.052
## actuals.hospitalBeds.capacity:population -8.773e-16  5.258e-16  -1.669    0.114
##
## (Intercept)                              ***
## actuals.hospitalBeds.capacity            .
## actuals.hospitalBeds.capacity:population
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0005608 on 17 degrees of freedom
## Multiple R-squared:  0.2152, Adjusted R-squared:  0.1229
## F-statistic: 2.331 on 2 and 17 DF,  p-value: 0.1275
```

Figure 4. Summary data from the simple linear regression model for number of Covid deaths and hospital bed capacity.

Analysis #2: Pooled t-test for Vaccination Ratios in Risk Levels 2 and 3

We decided to conduct a pooled t-test for vaccination ratios between risk levels 2 and 3. We chose Risk Levels 2 and 3 because the vast majority of data was in those two categories. Figure 5 is a box plot of vaccination rates by risk level for the randomly selected states. Risk level 1 and 4 only have one data point and that is why we chose to exclude them. The hypothesis was:

$$H0;\ \mu 1 = \mu 2 \text{ against } HA;\ \mu 1 \neq \mu 2$$

That is, Ho: The vaccination ratios are the same between states with risk level 2, and states with risk level 3. Our alternative hypothesis was that Ha: the vaccination ratios are significantly different between states with risk level 2, and states with risk level 3. The normality assumption is met by the vaccination ratio Q-Q plot shown in Figure 6, as well as by the Shapiro-Wilk test. Homoscedasticity can be assumed by the box plot generated for the vaccination rates for risk levels 2 and 3 in Figure 9. Figure 7 shows the results from the Shapiro-Wilk normality test done on the vaccination ratios. With a p-value of .899 (Figure 7) at a significance level of .05, we can retain Ho: Vaccination ratios are normally distributed.

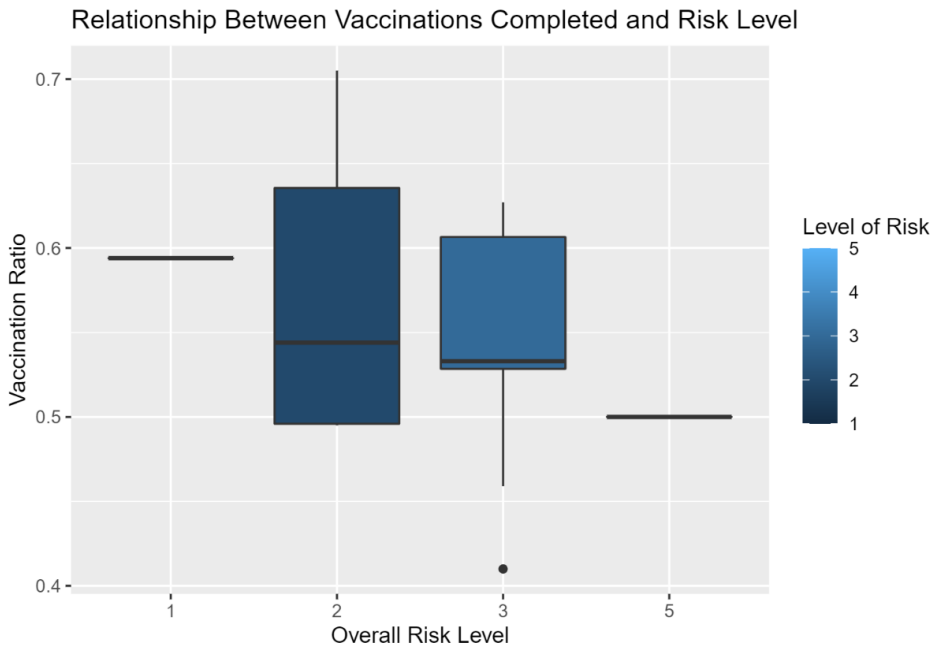


Figure 5. Box plot of the relationship between vaccinations completed and risk level.

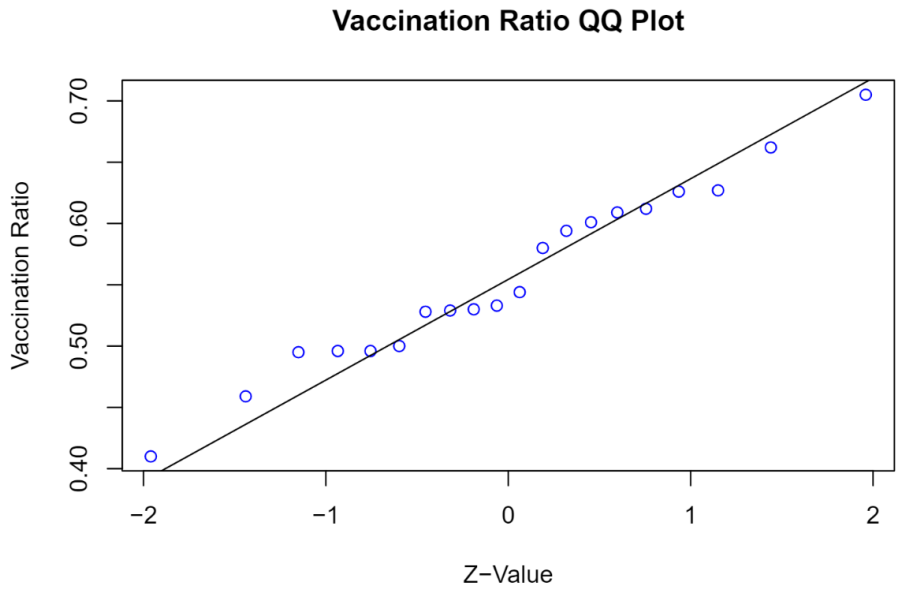


Figure 6. Q-Q plot for vaccination ratios.

```
## [1] "The metrics.vaccinationsCompletedRatio variable is normally distributed, indicated by the QQ pl
shapiro.test(metrics.vaccinationsCompletedRatio)

##
##  Shapiro-Wilk normality test
##
## data:  metrics.vaccinationsCompletedRatio
## W = 0.97757, p-value = 0.899
```

Figure 7. Shapiro-Wilk normality test results for vaccination ratio.

A crucial step is to determine if the variances are equal or not. The reason we need to check for equal variances is because the pooled t-test assumes equal variances, so an F-test is appropriate here to confirm that we satisfy this assumption. Our null hypothesis was Ho: The variances between risk level 2 and 3 are equal, and the alternative hypothesis was Ha: There is a difference between the variances in risk level 2 and 3. For this, we conducted an F-test to compare the variances (Figure 8). Because the p-value is .5091 and the significance level is .05, we failed to reject the null hypothesis. There is evidence that the variances do not equal each other, or in other words, there is no significant difference between the two population variances at the 95% confidence level.

```
        F test to compare two variances

data:  data$metrics.vaccinationsCompletedRatio by data$riskLevels.overall
F = 1.5601, num df = 6, denom df = 10, p-value = 0.5091
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3831149 8.5201789
sample estimates:
ratio of variances
          1.560094
```

Figure 8. F-test summary comparing risk level 2 and risk level 3 variance.

Finally, we ran a two tailed pooled t-test to determine if there is any difference between the mean vaccination rates in risk level 2 and risk level 3. Our null hypothesis was Ho: There is no difference in vaccination rates between risk level 2 and risk level 3, and the alternative hypothesis was Ha: There is a difference in vaccination rates between risk level 2 and risk level 3. Based on the p-value of .5298 from the pooled t-test, we fail to reject the null hypothesis that there is no difference in mean vaccination rates between risk level 2 and risk level 3.

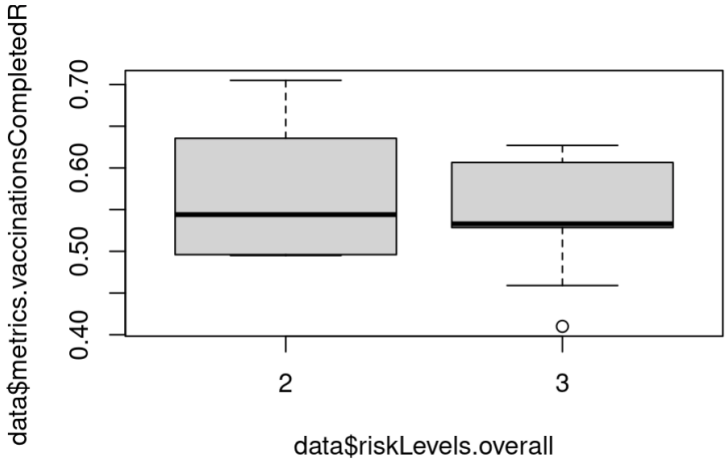


Figure 9. Boxplots of vaccination rates for risk level 2 and risk level 3.

```
##  Two Sample t-test
##
## data:  data$metrics.vaccinationsCompletedRatio by data$riskLevels.overall
## t = 0.64219, df = 16, p-value = 0.5298
## alternative hypothesis: true difference in means between group 2 and group 3 is not equal to 0
## 95 percent confidence interval:
##  -0.05474711  0.10233152
## sample estimates:
## mean in group 2 mean in group 3
##       0.5724286       0.5486364
```

Figure 10. Summary data from the pooled t-test conducted for vaccination ratio means of risk level 2 and risk level 3.

We also conducted a non parametric test on the vaccination ratios for all risk levels that were present in our dataset. The Kruskall Wallis test was most appropriate here because it does not assume normality, nor does it perform hypothesis testing on a parameter such as the mean. Instead, the test is performed on the medians of vaccination ratios for each risk level. So, our null hypothesis was Ho:There is no difference between the median vaccination ratios for each risk level against Ha: There is a difference between the median vaccination ratios for each risk level.

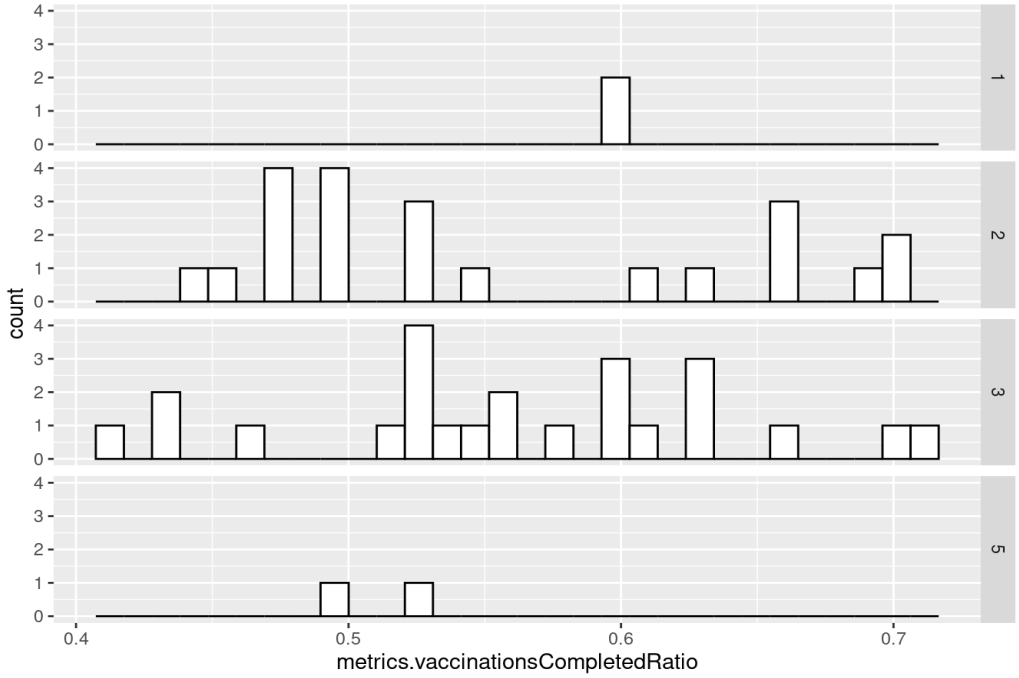


Figure 11. Plots of vaccination ratio frequencies per risk level

By the Kruskall Willis test, we were able to retain the null hypothesis given a p-value of 0.6934. Therefore, we were unable to find any significant differences between the vaccination ratios of all risk levels at the time of our analysis (this dataset is continuously being updated). There were also no significant differences for states with risk levels of 2 & 3, which were the most prevalent in our sample (as noted by the results of our pooled t-test).

```
##
##  Kruskal-Wallis rank sum test
##
## data:  metrics.vaccinationsCompletedRatio by riskLevels.overall
## Kruskal-Wallis chi-squared = 1.4519, df = 3, p-value = 0.6934
```

Figure 10. Summary data from the kruskal wallis test conducted for vaccination ratio means of between all risk levels.