**Does Mask-Wearing Have an Effect on Coronavirus-19**

**Rate of Infection and Death?**

**Introduction**

Coronavirus-19 (Covid19) is a global pandemic currently causing illness and death at rates not seen in a century. The infectious offender in this particular coronavirus is SARS-CoV-2 a type of severe acute respiratory syndrome (Dighe, 2020). In the first quarter of 2020, New Zealand and South Korea were very responsive early on in shutting down the virus before it became widespread by using two different approaches. In New Zealand, the response included early action before the virus became widespread, transparency in communication with the public, and monitoring and adjusting the response as necessary (Jamieson, 2020). In South Korea, the government responded early as in New Zealand, but their approach was a little different; they employed a virus-cluster based approach and intensive contact tracing to isolate the spread and stamp out the virus at its sources (Dighe, 2020)

These two countries provide good models for what to do before the virus becomes widespread. In the United States and many other countries, we are beyond that point. While it is cheaper to mitigate spread than respond to rampant infection (Jamieson, 2020), we must look at what can be done when we are past the point where shutdowns may be effective and contract tracing is overwhelming and nearly impossible. There are many social behavior suggestions that seem effective: social distancing, wearing masks, handwashing, and avoiding indoor gatherings that are advised by our governors. This data study will focus on mask-wearing to see if it affects the rate of infection and death by examining data of confirmed cases and death over a period of four months of Covid-19 data tracking from June 1st to November 30th of 2020 (Thomas Hale, 2020) and a survey of mask-wearing levels of compliance that was conducted from July 2nd to the 14th of 2020 (Dynata, 2020) .

**The Data**

The main datasets used are from a survey by the New York Times (NYT) and Dynata on mask wearing in the United States by county\* (Dynata, 2020) and a subset of the United States broken down by state from the Oxford Covid Data Tracking Project (Thomas Hale, 2020). I also relied on two U.S. population datasets. All of these are in the category of observational data. Variables are as follows:

Dynata and NYT survey:

* COUNTYFP: 4-digit FIPS code for the county in the United States

Answers to survey question of how often the respondent wears a mask when in public where they may be within 6 feet of someone

* NEVER: never wear a mask in public
* RARELY: rarely wear a mask in public
* SOMETIMES: sometimes wear a mask in public
* FREQUENTLY: frequently wear a mask in public
* ALWAYS: always wear a mask in public

\*https://www.nytimes.com/interactive/2020/07/17/upshot/coronavirus-face-mask-map.html for graphics

Covid-19 Government Response Tracker (OxCGRT):

* regionname: state
* date: date of data recording for each day beginning on January 1, 2020 (ending November 30, 2020 for this project)
* confirmedcases: cumulative report of confirmed by state and date
* confirmeddeaths: cumulative values by the state and date which uses data from Johns Hopkins University Covid Tracker
* economicsupportindexfordisplay: a 7-day smoothing average 1 to 100 scale of level of economic support from government
* containmenthealthindexfordisplay which is a 7-day smoothing average 1 to 100 scale and containment and health measures level taken by the government

U.S. Census Bureau: [Index of /programs-surveys/popest/datasets/2010-2019/counties/totals (census.gov)](https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals/)

popestimate2019: estimated 2019 population based on 2010 census which includes state and county FIPS codes

World Population Review: <https://worldpopulationreview.com/states> taken from Census 2017 estimate and estimated for 2018

pop2018: estimated 2018 population based on 2010 census by state total.

The Oxford Data set obviously suffers from the fact that we must trust that all states are reporting their cases and deaths honestly. There have been suggestions that there is deliberate underreporting by some states whose republican governors are seeking to keep the Trump Administration happy. We got shocking confirmation of this when the Florida Department of Law Enforcement broke into the home of a data scientist in Florida and seized her computer and equipment. (Ceballos, 2020). Rebekah Jones, a whistleblower claiming number manipulation in Florida cases and deaths, was fired from her job earlier this year and had been running an alternative Covid-19 dashboard on her own since her firing and claims the state is using a bogus hacking charge to stop her reporting. (Ceballos, 2020) This is not in court yet, but the newspaper story openly states that guns were drawn in the family’s home over a flimsy hacking charge which in itself is outrageous. The underreporting in some states will certainly need to be taken into consideration in reaching a conclusion about mask wearing results.

The New York Times survey has possible bias by respondents. The design of the survey posed a single question: “*How often do you wear a mask in public when you expect to be within six feet of another person?*” with each respondent choosing a rating of never, rarely, sometimes, frequently, or always. There may be bias associated with survey answers especially on this highly politicized issue: 1) The New York times name may influence answers 2) People may answer falsely "never" to "own the lib's" or "always" as wishful thinking of compliance to signal virtues and belief in scientists' suggestions. The composition of the categories are not clear demarcations, so we must assume for purposes of this analysis that they are roughly accurate reporting. To assist in measuring the two treatments clearly, consideration will be given only to those states who score below the 1st quartile as one group and those who score above the 3rd quartile in the second group.

The two census datasets are estimated population based on the Census Bureau 2010 census taken in the United States. I used two different sets because one was by county and the other totaled by state. Because they are both estimates relatively close in time, I felt comfortable using two different sources both based on the 2010 census numbers. They were used for weighting, and it was just easier. Therefore, the data here may be biased due to my not wanting to spend a lot of time manipulating estimated data just for the weighting given time constraints.

**DATA CLEANING:**

In order to use the mask-use dataset, I had to split the single column FIPS codes in state and county, preserving just the state code and then mutating them to a regionname column in order to join with the Covid-19 dataset (containing state name) . Otherwise, all other data cleaning was routine, i.e. changing column names to all lower case and removing special characters.

**STUDY QUESTIONS:**

1. I plan to look at a regression model for the Covid-19 data, although since it is common knowledge that spread of the virus is exponential, this would require logarithmic conversion. I will look at the R-squared value to see how well the model explains what is going on before proceeding with this plan. I will look at the residuals and the summary function in R to see where the quartiles for the mask survey dataset, as well as histograms and qqplots to view normality.

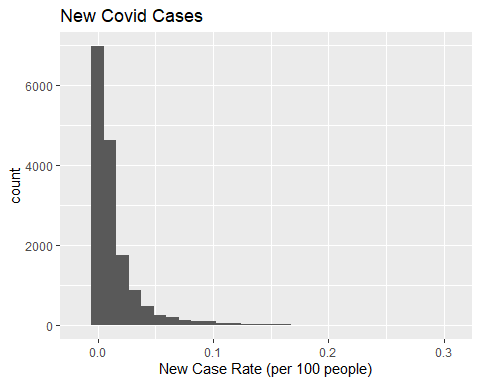
The Covid-19 data was weighted by population to get comparisons of new variables newdeaths and newcases per 100 and finally into newcases\_rate and newdeaths\_rate per person for each day by joining the Covid and population datasets. Additionally, Puerto Rico and DC were removed as data is not reported for them.

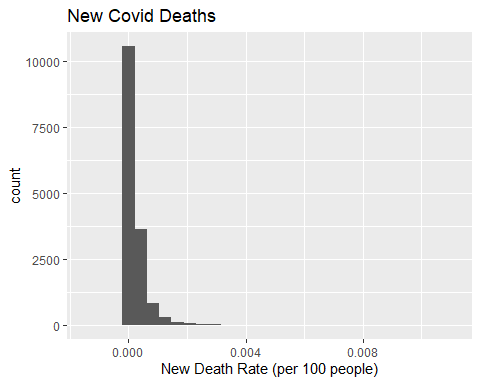
1. For the joined Covid-19 and mask survey data, I plan to look at two groups: those who always wear masks and those who never wear masks. I will use t-testing for difference in means if normality is determined or Wilcoxin-Mann-Whitney non-parametric test for significant shift in the two distributions if normality is questionable. If Wilcoxin-Mann-Whitney is chosen, I will use the Chi-Square test for independence between the two groups first to make sure the independence requirement is confirmed.

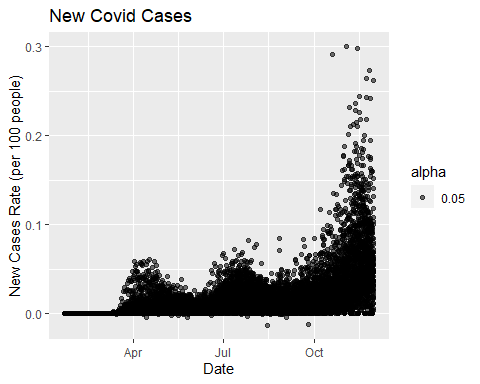
**RESULTS**:

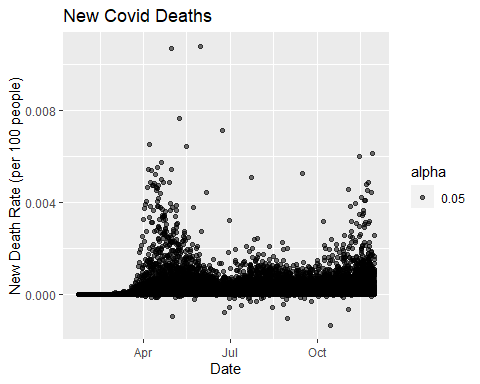
Covid19 dataset:

The histograms for cases and deaths are both highly skewed to the left.

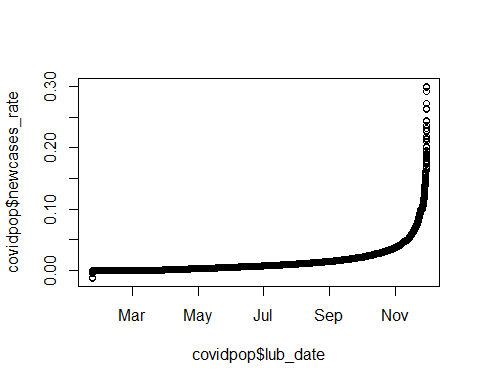


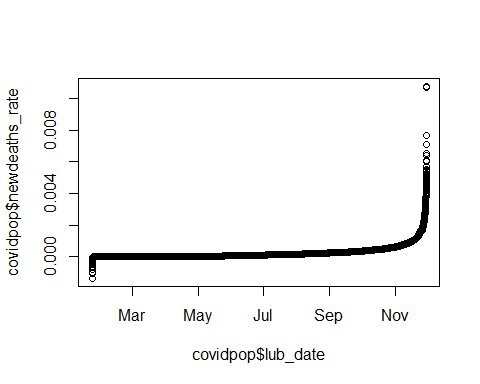






The new cases do not appear to be linear to the dates in the times series whereas the deaths may be, by viewing the scatterplot; however, R^2 = .03 indicates only 3% of y is explained by x. This is not surprising as the virus tracking has been described as exponential in growth rate. Normality is questionable because of the highly skewed histograms shown above, as well as the shape of the qqplots shown below.





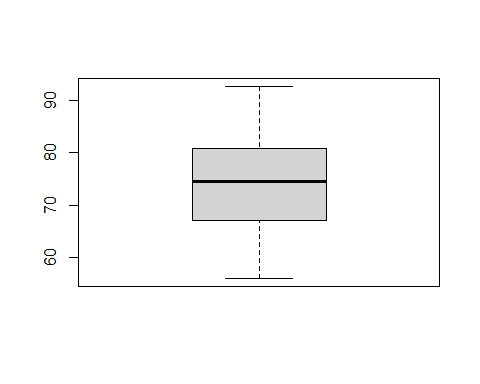
Mask-wearing and Covid-19 joined datasets: The question is does mask use affect rate of cases or rate of deaths due to Covid-19? Hypothesis testing is as follows:

H\_null: probability of new covid cases and deaths is not associated with mask usage, i.e. p\_highcompliance = p\_lowcompliance for new cases and new deaths.

H\_alpha: probability of covid cases and deaths is associated with mask usage, i.e. p\_highcompliance is not equal to p\_lowcompliance

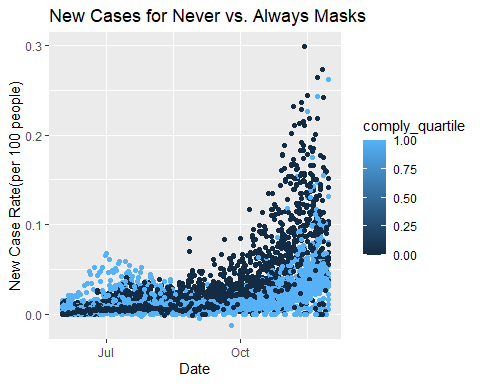
*Alpha is set at 0.05*.

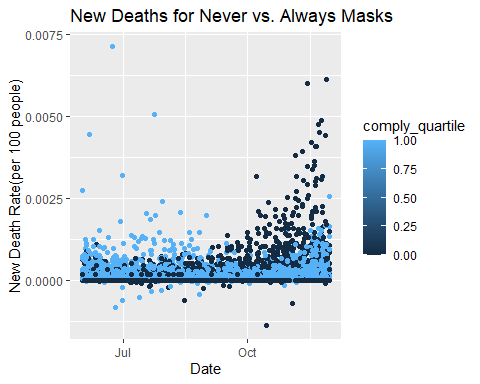
First, view the boxplot for quartile cutoffs. These will be used to filter the states into two groups, those with a state score variable st\_score >= 82.29 and st\_score <= 67.05 for a clear demarcation between never/always scale of compliance.

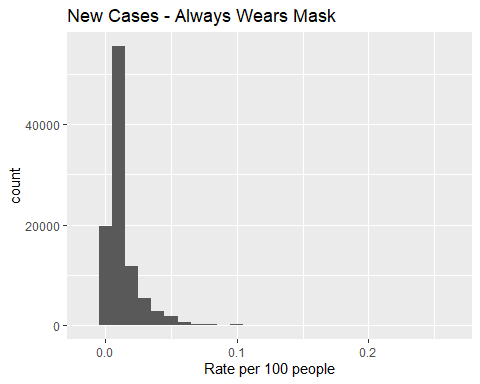


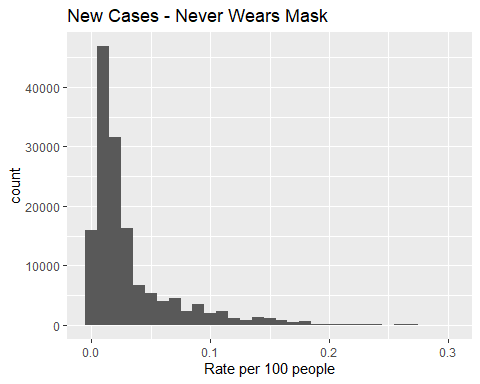
Because we are trying to determine if masks affect the rate of infection and death, we are essentially looking at 2 treatments (level of mask wearing) and comparing to see if there is other than random difference that occurs between the 2 treatments. We need to see that our observational population pool consists of the entire population of the United States. Assumptions are that beginning in June 2020 when New York was better through the initial wave and states began to ease off shutdowns and other extreme measures, so the period studied is June 1 thru November 30, 2020. Mask wearing was more commonly recommended beginning in April, so by June, should be settled behavior. Further, the survey (dataset) is a sampling of the whole U.S. population which was done by the New York Times at the beginning of July, and this analysis is assuming steady rates of compliance throughout. There may be bias associated with survey answers especially on this highly politicized issue: 1) The New York times name may influence answers 2) People may lie “never” to “own the lib’s” or “always” as wishful thinking of our compliance to signal virtues and belief in scientists’ suggestions. The composition of the categories are not clear demarcations, so we will assume for purposes of this analysis that they are roughly accurate reporting. To assist in measuring the two treatments clearly, consideration will be given only to those states who score below the 1st quartile as one group and those who score above the 3rd quartile in the second.

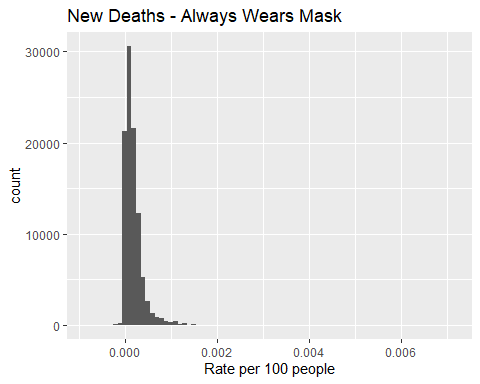
After separating into 2 groups “always” and “never” wears a mask, we revisit the scatterplots for Covid-19 new cases and new deaths. It’s fairly easy to see that there appears to be evidence for mask-wearing, but we will go on to test the hypothesis.

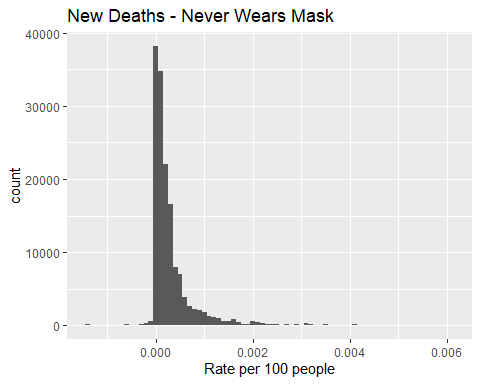








")



The histograms above show the counts of new cases and new deaths for a range of calculated rates per 100 people per day. These are also highly skewed to the left.

The t-test does not seem appropriate because of the large skew and original qqplots for the Covid-19 dataset; we can try a chisquare for association of treatment type(low compliance in mask wearing and high compliance in mask wearing) with a 2 way table and success defined as no daily increase in cases or deaths as follows:

Always Masks Never Masks

Success count success count success

Failure count failure count failure

In R, new cases returns p < 4.244xe-15 < 0.05; this indicates very strong evidence of independence. Also in R, new deaths returns p <.022e-16<0.05 ; these values provide very strong evidence of independence between the groups. This allows us to proceed to the Wilcoxon-Mann-Whitney Test where the results between the two groups result in p < 2.2e-16<0.05 for new cases and the same p < 2.2e-16<0.05 for new deaths. For new cases and new deaths, according to this test, the population of people who always wear masks versus the people who never wear masks shows very strong evidence of shift in distribution.

**CONCLUSION**

Because the Chi-Square test for independence is satisfied with p < 4.244xe-15 < 0.05 for new cases and p < .022e-16 < 0.05 for new deaths, the Wilcoxon-Mann-Whitney test shows that there is very strong evidence ( p < 2.2e-16<0.05 for new cases and deaths), that the distribution of mask wearing population and non-mask wearing population show significant shift from each other in the number of new deaths and the number of new cases: the null hypothesis is rejected.

Looking at the color scatterplots earlier for each group, it very strongly suggests that mask wearing may lower the rate of new Covid-19 cases and deaths. The implications of this are that mask-wearing must be de-politicized. Too many people are believing false information, and they are increasing their risk of illness and death to themselves and others. We have been failed by our leaders in the government, although the scientists and medical community have performed extraordinarily despite facing threats beyond just a deadly virus. I am hopeful that the new administration can reverse the course we are on through example and education. After we get out of this, schools must concentrate on critical thinking skills so that we choose our sources of information more carefully.

# Bibliography

Ceballos, M. E. (2020, 12 7). Did COVID data whistleblower hack Florida’s emergency alert system? Police raid home. *The Miami Herald*.

Dighe, A. e. (2020). Response to COVID-19 in South Korea and. *BMC Medicine*, 1-12.

Dynata, T. N. (2020, July). *Estimates from The New York Times, based on roughly 250,000 interviews conducted by Dynata from July 2 to July 14.* Retrieved from https://github.com: https://github.com/nytimes/covid-19-data/tree/master/mask-use

Jamieson, T. (2020). “Go Hard, Go Early”: Preliminary Lessons From New Zealand’s Response. *American Review of Public Administration*, 598-605.

Thomas Hale, S. W. (2020). *https://github.com*. Retrieved from Oxford COVID-19 Government Response Tracker. Blavatnik School of Government.: https://github.com/OxCGRT/covid-policy-tracker/blob/master/data/OxCGRT\_latest\_withnotes.csv