Stacey Wheeler
DATA516_A4
Due: November 4, 2021

A4: Common Analysis

Visualization Explanation

My visualization shows the number of confirmed daily COVID-19 cases in Oakland, Michigan from February 1, 2020 through October 15, 2021 (Number of Daily Cases on left axis). It also shows the rate of change in daily cases over this time frame (Rate of Change of Daily Cases on right axis). The case count data used in this visualization was downloaded from the Center for Systems Science and Engineering at Johns Hopkins University. The confirmed case line is color coded according to when a mask mandate was in place: blue indicates that masking was enforced in the county, and red indicates that masking was not enforced or it is unknown. The masking data by county by day was acquired from the CDC. The rate of change of daily cases is shown with the plum line. A 95% confidence interval for the rate of change is also displayed with horizontal gray lines, which I calculated as 3 x std_dev(rate_of_cases).

Before making the plot, some steps were taken to prepare the data. Both datasets were filtered to include only information about Oakland, Michigan. Then I used the daily case count to calculate a 7 day moving average and the cumulative cases. I did the same processing for the dataset on daily COVID-19 deaths, and then I merged the two dataframes on the date. Using the county population I calculated the susceptible population for each day by subtracting the cumulative deaths and cumulative cases from the total population.

Some values of interest were calculated for each day using the following formulas:

Rate of Infection = (Cases moving average) / (Susceptible population)
Rate of Daily Infection = (Cases) / (Susceptible population)
Rate of Cases = (Daily difference in cases moving average) / (Total population)
Rate of Daily Cases = (Daily difference in cases) / (Total population)

The graph shows the cases 7 day moving average with the blue/red line and the ‘Rate of Cases’ with the plum. To understand the effectiveness of the masking policy in this county, we can observe what happened when the mandate was first enforced and when it was lifted. In April 2020, the county first put in place a masking policy that seems to have kept transmission at bay for approximately five months. October 2, 2020 was the first time the county lifted the mask mandate, and perhaps due to expected increase in transmission through the winter season it was immediately put back in place on October 5, 2020. Even though masks were required throughout the winter, the amount of gathering and travel that took place during the holidays is a confounding variable that caused huge spikes in transmission, as evidenced by the peaks in the blue lines as well as how many times the plum line goes outside the gray bars.

The peak in April 2021 may be due to spring break activities, but immediately following this peak there is a steep decline to the lowest daily case average since the pandemic began. Finally the mask mandate was lifted on June 22, 2021, and the cases began steadily rising again until the end of the time series on October 15, 2021. The last red portion of the graph is the strongest evidence for the efficacy of masking because even after vaccines became widely available, infections returned as soon as the mask mandate was lifted. Summer 2021 also coincided with the contagious Delta variant, so the new strain combined with no masking led to increasing infections.

Reflection

The collaborative aspect of this assignment encouraged me to think about challenges and limitations with the data that I would not have considered otherwise. I do not think I would have considered how complicated it is to figure out how much of the population is susceptible at any given time. My classmates encouraged me to think about transmission rates of different variants, recovery times, likelihood of reinfection, breakthrough cases after vaccination, and a myriad of other unknown variables like self-isolation or non-compliance with masking that we have no way to measure. I chose to use a relatively simple formula to find the susceptible population since so many of these other factors are impossible to know (and likely balance each other out when taken altogether). My estimate for susceptible population is likely an underestimate of the actual number of susceptible people because once a person has recovered from the virus they return to the susceptible population; however for this purpose we will assume it is a close enough approximation due to the number of individuals who are also in the susceptible population but had very limited exposure due to self-imposed quarantine.

We also had meaningful discussions about the significance of the daily case count vs daily case rate, and what we learn when the rate of cases exceeds the 95% confidence interval. For example, we may expect the daily case count to increase, but an increase in the daily case rate is concerning because it indicates that infected individuals are infecting others more quickly Ultimately we determined that there are so many unknown factors and confounding variables (e.g. holiday season or new variants) that make it almost impossible to determine precisely how effective masks were for a given county.

Attributions

I collaborated on this assignment with Apoorv Sharma, Anushna Prakash, and Preston Stringham. Apoorv and I discussed the formulas that we used to calculate the fields of interest and how to present the information on a visualization. Code that I shared with Apoorv includes the implementation of matplotlib to create a visualization that has two y-axes and a shared x-axis using the axes.twinx method. Anushna helped me understand how the data do and do not prove efficacy of masking, and she proposed the option to extrapolate the initial transmission rates as though masking were never put in place to get an idea of what might have happened if our counties never had any mandate. Preston discussed with me what the question was asking us to address and different ways I could use my visualization to answer the question and what kinds of factors I should consider.