Anushna Prakash
DATA 512 - Human-Centered Data Science

A5 - Extension Plan

I. Motivation

The datafication of the pandemic has allowed the public access to more health-related data at a granular level than ever before. The county that I was assigned, Orange County, California, had quite a high number of cases given the high population of the county. The lockdowns and mask mandates enforced did prevent people from leaving their homes and the spread of other diseases like the flu, and likely even other kinds of death. However, at the same time, with hospitals being overloaded and sometimes even needing to turn away patients, I wonder how many deaths there were in excess of what is typical to see in Orange County as this may provide evidence that there were actually more COVID-related deaths than is officially being reported, since the healthcare industry may not want to use tests on already deceased persons. Essentially, this question will help answer if we can be certain that there were more deaths due to the pandemic than what the "official" count may suggest, or if the pandemic reduced other causes of death such as from accidents.

Another angle here is that COVID added a lot of uncertainty to the world. This had many adverse effects on people's lives, and I am curious to see how this affected fertility. Perhaps during these stressful times, fewer people were confident in the future and their ability to support a child, or perhaps quarantine gave people the opportunity to move forward with their family planning because of the flexibility that working from home provided. I would like to inspect the pandemic and its effects on the lives of people in Orange County, CA from the angle of effect on mortality and fertility.

II. Research Questions & Hypotheses

My hypothesis related to mortality is that all-cause mortality was higher during the pandemic than pre-pandemic, indicating that there were more deaths due to COVID-19 than were being reported. My hypothesis relating to fertility is that there were fewer births during the pandemic compared to pre-pandemic due to the stress and uncertainty it caused.

III. Data Used

In addition to the Johns Hopkins data regarding daily case counts, I plan to use their daily death counts. Additionally, the California Health & Human Services (CHHS) department has an open data set regarding mortality by county aggregated by month. The data is available for prior years, and also through September 2021. The deaths data set includes other counties within California and tracks different causes of death. There is all-cause death, which will include COVID-19 deaths, but there are also accidents (unintentional injury), diabetes, heart disease, and intentional self-harm (suicide). This will help provide additional context around what kind of deaths are being seen in excess during the pandemic. For example, if accidental injury-related deaths went down during the pandemic, then it may be possible that quarantine conditions did help keep people safe to a certain extent, even excluding COVID-19 deaths. However, if suicide-related deaths are shown to be significantly higher during the pandemic compared to prior years, then perhaps this is an indicator that mental anguish was exacerbated.

The same source from California Health & Human Services tracks the number of births on a monthly basis and has data until September 2021. It does track where the births take place, including in a hospital, at home, or in a freestanding clinic, but I am less interested in where births took place over time and moreso in the total number of births, which is still included in the data set.

The Johns Hopkins COVID data is available under the creative commons license. This states that I can freely use and adapt the data as needed, as long as I provide attribution, indicate if any changes were made to the data, and link to the license. I also cannot place any additional restrictions on the legal terms or technological measures that restrict others from being able to do anything to the license permits.

The data from CHHS is available under these terms of use: Your reuse, publication, and/or distribution of the Content on the CHHS Oen Data Portal requires attribution of credit to the CHHS department or office providing such Content and a citation to the webpage and date of publication of the material cited. Preferred citation language will be provided in the metadata of each data table. Any analyses, interpretations, or conclusions reached by your use or analysis of the Content shall be credited to the author and not to the State or Content Source Organizations. If you modify the Content for your own purposes in any way, you may not claim the data is "official government data" and must clearly indicate that the data and/or data table has been modified.

Given that this is government-created and government sponsored data, I do not think there are any serious ethical considerations when using this data. The data is HIPAA-compliant as it is de-identified, and the data is aggregated anyway which prevents my ability to identify specific individuals based on the content in the data. Since it is government data, I am confident that it was obtained in a legal manner and that patients and hospitals understand that this data will be shared with the government and for public use.

IV. Unknowns and Dependencies

In order to answer the question relating to how fertility was affected by the pandemic, I need to assume that there was nothing concurrently happening with individuals' health that might affect fertility related to causes outside of the COVID-19 pandemic. Unfortunately, I am unable to find any evidence of this at the time, so it is difficult to say if my analysis will be confounded from this perspective. I am also aware that fertility in the U.S. has been declining in general, so it is possible that any declines during the pandemic are simply extending a preexisting trend rather than being attributable to the pandemic. I realize that this is a limitation of my analysis and will caveat it appropriately.

As for mortality, I think it is difficult to tell exactly if I find excess deaths for non pandemic-related causes, that this isn't simply a coincidence and that it is indeed excess deaths due to COVID in disguise. For example, if heart disease-related deaths increase substantially during the pandemic, it is difficult to make an argument that these heart disease-related deaths are actually excess COVID-related deaths. However, I think that this can be mitigated by keeping the view of deaths at a high level rather than looking at specific disease-related causes of death. I do not want to try and make a case that a specific cause of death is actually all COVID-related, rather I would like to see if there is evidence that the pandemic reduced overall deaths from non COVID causes or not.

V. Methodology

I plan to combine the county-level data for Orange County from the Johns Hopkins data set, of both cases and deaths, and aggregate this to the monthly level instead of daily level. This way it will match the granularity of the mortality and fertility data sets. When all of the data sets are combined, I intend to have each row be a different month, and have a column for the monthly new COVID cases, monthly COVID-related deaths, monthly total births, and monthly deaths by various causes (all-cause, accidents, and suicides). I will exclude COVID-19 deaths from all-cause deaths so that I do not capture the effect of the pandemic in the mortality data. I will use data from three years pre-pandemic so that there is a comparison available of pre and post-pandemic trends. Depending on the analysis, I will mark when the pandemic "started" differently. In the case of mortality, I think it is fair to claim that the pandemic began in February 2020, which is most of when the data becomes available and the first confirmed cases are tracked. Each subsequent month after until the end of the data set will be considered "pandemic".

For fertility, this process will be a bit different since there is a lag between when family planning begins and when children are born. In this case, I will mark the pandemic as starting 9 months after February, so November 2020. Each month after will also be considered the "pandemic". This way, I only capture the time in which people are actively able to make family planning decisions rather than attributing fewer births early in the pandemic as being related to "uncertainty", when people were not aware of the pandemic at the time. Consequently, there will be less data to compare to since there will be fewer months of data left to analyze (less than a year), so it is possible that there will not be anything significant that comes from this analysis. In that case, I will suggest that it will take more time to fully understand the impacts of the pandemic on fertility since I do not have data on how many people are pregnant, just the number of births.

After combining the data, I will do statistical t-tests to see if the monthly average number of deaths and births can be considered significantly different pre-pandemic and during the pandemic. I will be able to do this analysis at the overall level, and also at different kinds of deaths (accidents and suicides) to see if there was an effect. Similarly with fertility, I will be able to evaluate if there was a significant difference in the average number of monthly births before the pandemic and after. Furthermore, I will be able to evaluate not just significant statistical differences, but also practical differences. If there is a significant difference that is ultimately quite small as a result of the statistical tests, then I will be able to make a case that the pandemic actually had very little impact on mortality or fertility.

To present the findings, I will show the averages as time series and mark when the pandemic "starts", and visually show what would be considered a "significant" departure in trend. This way it is clear to the audience that my results do or do not show evidence supporting a major difference in the number of deaths.

VI. Timeline to Completion

Collecting and joining the data together should be a short affair, however exploring the data, seeing what the quality of it is, if there are any missing values or inconsistencies will take a bit longer. This will also help inform me if I need to adjust the scope of my project up or down. I will spend roughly five days accomplishing these tasks before moving on to manipulating and doing calculations.

Performing the statistical analyses should be very quick, but analyzing and validating the results and ensuring that my interpretations are correct will take a bit longer. I will allocate three days for this to the mortality topic, and two days to the fertility topic. In total, this will be another roughly five days to analyze the output of my tests and run additional tests if they come up.

I will out of town during Thanksgiving so I will allocate some days of leeway during this time. Upon return, I will be focused on putting together the presentation. I will have extra time to do other analyses if something interesting comes up to have extra material for the paper. I will also use this time to refine some interesting visualizations for the presentation. This should take roughly five days. Then I will take another few days to put together the presentation and practice. And, since I am a slow writer, the remaining time will be dedicated to writing up the results formally to submit the final paper.

- Data gathering & EDA: November 12th - 16th
- Statistical tests: November 17th - 21st
- Thanksgiving: November 23rd - 27th
- Visualizations: November 28th - December 2nd
- Presentation: December 3rd - 6th
- Final paper: December 7th - 16th