Impact of COVID-19 on Mobility Change in United States

Data 102: Data, Inference, and Decisions, Spring 2021

Jiaqi Meng, Jui-Chen Pu, Qiping Zhang, Zhijian Deng

1. Data Overview

The Covid-19 has not only triggered one of the largest world-wide pandemic emergencies in recent history, but also significantly changed how people live as more governments imposed mandatory lock-down to curb the spread of the virus. Since mass-gathering is prohibited and work-from-home became a new normal, places such as restaurants, offices and metro stations are either closed or experiencing lower traffics. The dataset analyzed in this project is the *Covid-19 Community Mobility* dataset provided by Google with regional focus on the United States. The dataset is in time-series format with changes in mobility compared with baseline. The earliest entry started at around February 15th of 2020, the time right before the imminent pandemic, and ended at around April 17th, 2021 with small variation per state. Based on the official report, entries with lower accuracy or failure in achieving statistically significant were excluded. There are 6 main categories that represent 6 different avenues in the dataset: Retail & recreation, Grocery & pharmacy, Parks, Transit stations, Workplaces and Residential, each of which shows how visitors spend time on the categorical place compared with baseline, which is a median value from 5-week period from January 3rd to February 20th, 2020.

2. EDA

The data provided originally were separated into 2 files, each representing a year. By stacking over the 2 datasets into a single dataframe, it is more convenient to perform the later analysis. To better understand the time-series variations of each categorical avenue, we selected Alameda County of California as the starting point and visualized them via Matplotlib. Under the sub-region-1 being California, there are in total 57 counties under the state in the dataset. Figure 1 in the appendix shows the time-series variation of mobility change for each place in Alameda. As expected, people stay in residential areas with more time compared with baseline in the pandemic as lock-down and remote-working were imposed. The average stay inside of living areas is around 17.2, which is far higher than the baseline. From the graph, we could also see the gradual decrease in the length of stay at home as time passes, which may be explained by the loosened lock-down order when the

disease is controlled. For other places, such as pharmacy, grocery stores, transit, retail and recreation, mobility has drastically decreased while visit to parks seems to have a cyclical pattern with significant decline at the beginning but climbed back quickly above the baseline level. It makes sense that as social distancing is required, places such as national parks are where people could comply with the policy while achieving fitness or recreation goals. It's also worth noting the sudden rally in visit of pharmacy and grocery stores at the early stage of pandemic (the significant spike at the beginning of the time series plot), which climbed to 39 above the baseline. The highest change happened on March 16th, 2020, which is 3 days before the enactment of the official Stay-Home-Order in California. Therefore, the sudden rise could potentially be explained by the enormous panic caused by the mandate while people rushed into stores to accumulate supply of necessities.

Out of all the percent change, the distribution of that in grocery and pharmacy stores has a roughly normal distribution, which is shown in Figure 2. The density graph centers at 0 and spreads into the 2 directions relatively evenly with standard deviation being around 14.9. Figure 3 shows the distribution at the state level where the majority of them display the bell shape distribution. The detailed distribution could be found in the jupyter notebook. As low supply of food and necessity is the main issue, understanding the rise or fall in mobility to grocery stores in each state is essential to determine logistic distribution of supply. With that said, states with negative percent change may relatively need less resources than before compared to states with large positive percent change. Therefore, such observation prompts us to use Bayesian inference to understand the posterior distribution of each state.

What's more, the distribution of the change in the workplace from baseline also resembles a normal distribution. By inspecting the histogram, we start to consider why it follows such a pattern. Because we believe that the change in workplace can directly represent the status of the economy, we start to wonder if the pandemic immediately harms the economy, which then leads to a decrease in workplace. Therefore, we consider using multiple hypothesis testing to test whether the drop in workplace is more significant in March 2020, when the pandemic first started in the United States.

3. Research Questions

3.1 Question 1:

Is the economy in March 2020, when the pandemic first started, worse than the rest of the year 2020?

3.2 Question 2:

By observing the distribution of data reflecting fluctuation for grocery and pharmacy needs in each state in the U.S, what does the posterior distribution of each state look like which could help determine resource allocation?

4. Technique 1

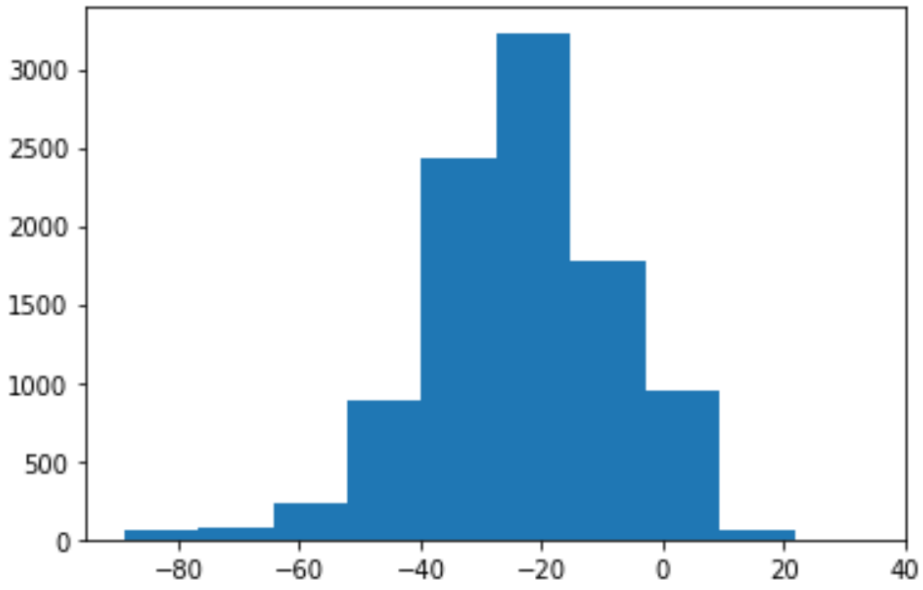
4.1. Method

H0s: The economy on date 2020/3/x is not different than that of the rest of the date in the year 2020. (i.e. The pandemic does not immediately influence the economy)
H1s: The economy on date 2020/3/x is worse than other dates in the year 2020.
Significance: 0.1

We first choose the column “workplaces_percent_change_from_baseline” as a test statistic that reflects the economic situation at each date.

We choose multiple hypothesis testing to answer question 2 is because simply taking the sum of “workplaces_percent_change_from_baseline” for each date to produce our statistic is meaningless and cannot reflect the true situation. Doing hypothesis testing for each date is more representative and precise.

Then we bootstrap 10000 samples from workplaces_percent_change_from_baseline at every date of year 2020.
We could visualize the distribution of workplaces_percent_change_from_baseline using a histogram.



Then, for each date starting from 2020/3/1 to 2020/3/31, we can calculate a corresponding p value from this distribution.

We use Bonferroni Correction and Benjamini-Hochberg procedure to make decisions for each date from 2020/3/1 to 2020/3/31. The Bonferroni Correction is simply using a significance level of alpha/len(p_vals) for each hypothesis test. The Benjamini-Hochberg procedure is the procedure of finding a largest k such that the kth sorted p-value is smaller than k * alpha / len(p_vals), and reject all H_i for i <= k.

4.2. Result

p_vals = array([0.0208, 0.0419, 0.0419, 0.0419, 0.0419, 0.0584, 0.0584, 0.0584,
0.083 , 0.0917, 0.0917, 0.0978, 0.1037, 0.1442, 0.1442, 0.4096,
0.5263, 0.6142, 0.7173, 0.7173, 0.7589, 0.7768, 0.813 , 0.8445,
0.8757, 0.9035, 0.9091, 0.9091, 0.9146, 0.9189, 0.9233])

bonferroni(p_vals, 0.1) = array([False, False, False, False, False, False, False, False, False, False,
False, False, False, False, False, False, False, False, False, False, False, False, False, False,
False, False, False, False, False, False])

benjamini_hochberg(p_vals, 0.1) = array([False, False, False, False, False, False, False, False, False,
False, False, False, False, False, False, False, False, False, False, False, False, False, False,
False, False, False, False, False, False])

Under 10% significance, we do not reject any of our null hypotheses in both correction procedures. We cannot conclude that the economy in March 2020 is worse than the economy of other months in the year 2020.
The Bonferroni Correction controls the family-wise error rate (FWER) with the proof attached below.

$$\mathrm{FWER} = P\left\{\bigcup_{i=1}^{m_0}\left(p_i \le \frac{\alpha}{m}\right)\right\} \le \sum_{i=1}^{m_0}\left\{P\left(p_i \le \frac{\alpha}{m}\right)\right\} = m_0\frac{\alpha}{m} \le m\frac{\alpha}{m} = \alpha.$$

The Benjamini Hochberg procedure controls the false discovery rate (FDR).

4.3. Discussion

All hypotheses survived in both correction procedures. This indicates none of these workplace percent changes in these days is significant when compared to other dates in the year of 2020. In general, we can conclude that the economy in March 2020 is not significantly worse than other months in 2020. If there is raw data of the absolute value of the workplace for each day and we conduct multiple hypothesis tests on these data, our conclusion could be more plausible.

5. Technique 2

5.1. Method

Before diving into the analysis, there are several assumptions that we made. We assumed that the visits to grocery and pharmacy stores indicate the demand for necessities supply such as food. We also assumed that the distribution of the true mean of percent change in grocery and pharmacy in the US, treated as the prior distribution, is normally distributed provided by the evidence as shown in Figure 2. In addition, the observed sample distribution of percent change for each state, or the likelihood, is also normally distributed as shown in Figure 3 where the mean is distributed based on the prior, although there are few extreme exceptions such as Hawaii. This means that the posterior distribution for each state is also normal by assumption made for the prior and likelihood. In the computation, we also assume that the likelihood of all states have the same standard deviation. The figure is determined by calculating the average of the aggregate standard deviations of each state's observations.

To calculate the posterior distribution, the pymc3 package is used to set up the model and generate samples from a pymc3 model of posterior distribution represented by the graphic model shown in Figure 4. The default algorithm used in pymc3 is the Markov-Chain-Monte-Carlo method (MCMC) and due to the long processing time, we limited the target acceptance rate to 80% and only generated 200 samples each chain. Limited number of states were processed because of the time constraint and selected

randomly, including Rhode Island, Georgia, Massachusetts, Nevada, Hawaii, California and New York.

5.2. Result

The posterior distribution generated by MCMC is displayed in Figure 5. Since both the prior and the likelihood distributions are assumed to be normal, it is reasonable for the posterior distribution to also have a normally-distributed shape. With only 400 samples in aggregate for each state, the distribution doesn't look like a perfectly Gaussian distribution but is relatively uneven. For the selected states, the majority of them have a negative mean, representing a decline in visit to grocery and pharmacy stores in the time span. Meanwhile, states such as Georgia have a positive mean, although the value is really close to 0 in absolute terms. This means that on average, the Georgia government may be more cautious on keeping the supply of food and necessity on par or even more than the pre-pandemic level as the demand or visit to stores doesn't drop significantly. In contrast, other states with a negative mean such as California experience a drop in visits. This may indicate that the state government needs to keep the supply at the same level as before to prevent shortage, and extra supply may not be necessary based on the posterior distribution.

5.3. Discussion

Computational-wise, the sample size may not be large enough to infer the posterior distribution accurately as we only use 200 samples for each chain with a relatively lower acceptance rate as 80%. Moreover, only a selected number of states were chosen to generate their respective posterior distribution while there are forty more states awaiting to be considered.

Some states such as Hawaii don't have a normal-like distribution of their percent change in visits to pharmacy and grocery stores. Figure 3 shows the distribution of Hawaii for reference and it has a bimodal distribution with heavy left tails. This may be explained by the low sample sizes of observations for Hawaii, which only has 2140 entries while a larger state such as California has 23,820 observations. However, such deviation may suggest that the distribution of the likelihood is not properly defined by the normal distribution.

6. Conclusion

From our tests in question 1, we can conclude that although people might think that the pandemic immediately affects the economy, such impact is not significant. This is also reflected in our analysis in question 2, which suggests that visits to grocery and pharmacy stores display an interesting distribution and seem to be unaffected by the lock-down order, although there are temporary rises and falls in the frequency of visits. The posterior distribution of percent change in grocery and pharmacy visits are also normally distributed if assumptions made above hold, where the degree of shifting from 0 differs. Such method and inference could be used by the government to make better decisions in allocating necessary resources under Covid-19, the time when over-panic and shortage of goods are likely to happen. States with positive percent change in visits to the stores should proportionally increase the supply to prevent from shortage while those with mean in the negative zone should at least keep the goods supply in tandem with the pre-pandemic level. However, further study to demonstrate the relationship between store visits and demand may be required. In this study, the association between visits and demand is assumed to be strongly related and positive, meaning that increase in store visits indicate a rise in demand of the necessity. This is a relatively strong assumption and study of the association between the two is essential so that the result in this study could be valid.

7. Appendix

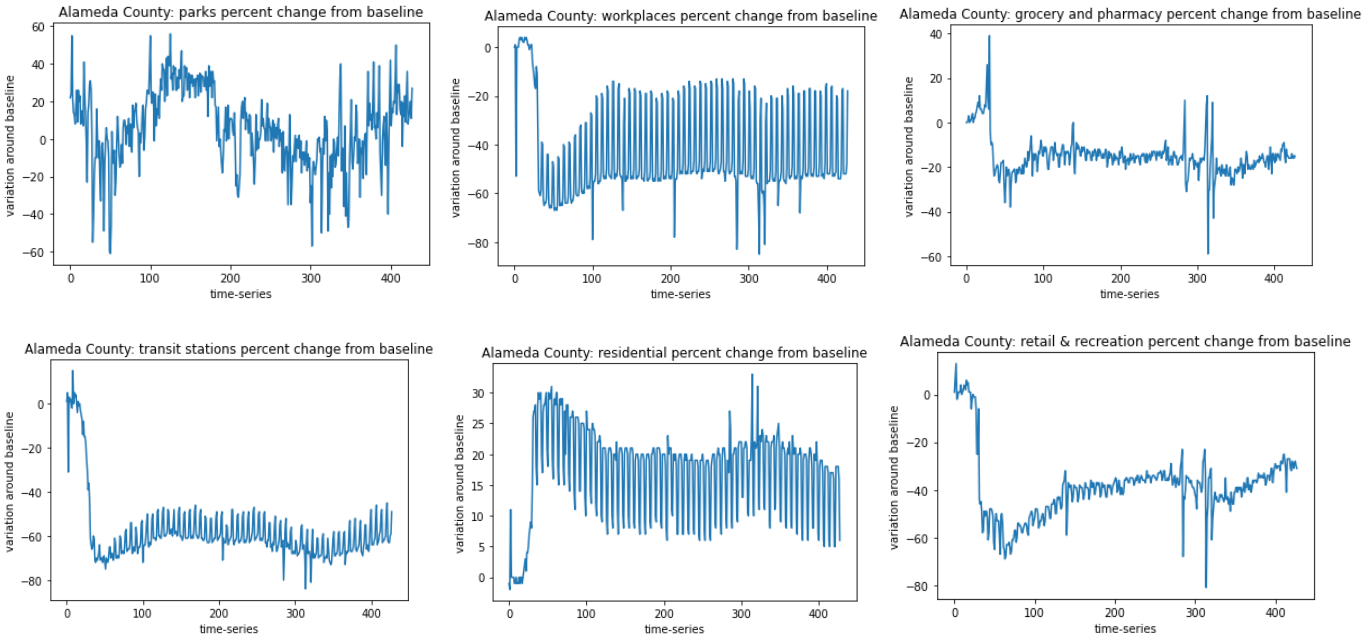


Figure 1: time-series plot of changes from baseline of each categorical place for Alameda County, California. The x-axis represents the time sequence where 0 denotes February 15th, 2020 and the sequence ends on April 17th, 2021. The horizontal axis shows the change in mobility compared with the baseline where 0 implies no change.

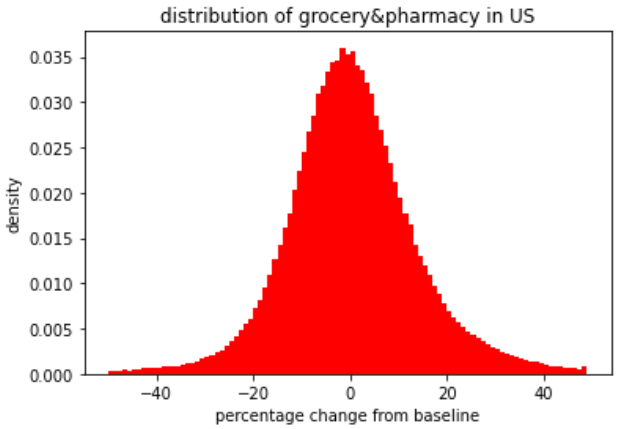


Figure 2: distribution of percent change from baseline in visit to grocery & pharmacy stores in the US.

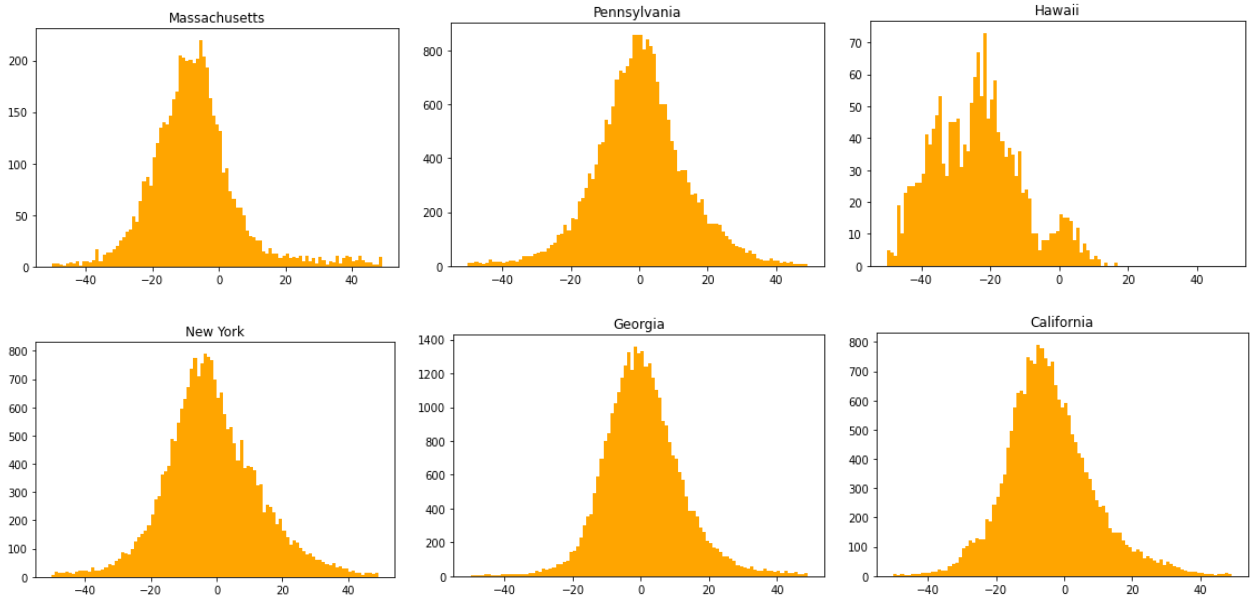


Figure 3: distribution of percent change from baseline in visit to grocery & pharmacy stores in various states.

Figure 4: the graphic model for the bayesian inference. μ represents the unknown mean of the percent change in the whole US and the shaded circle represents the observed percent change at the state level.

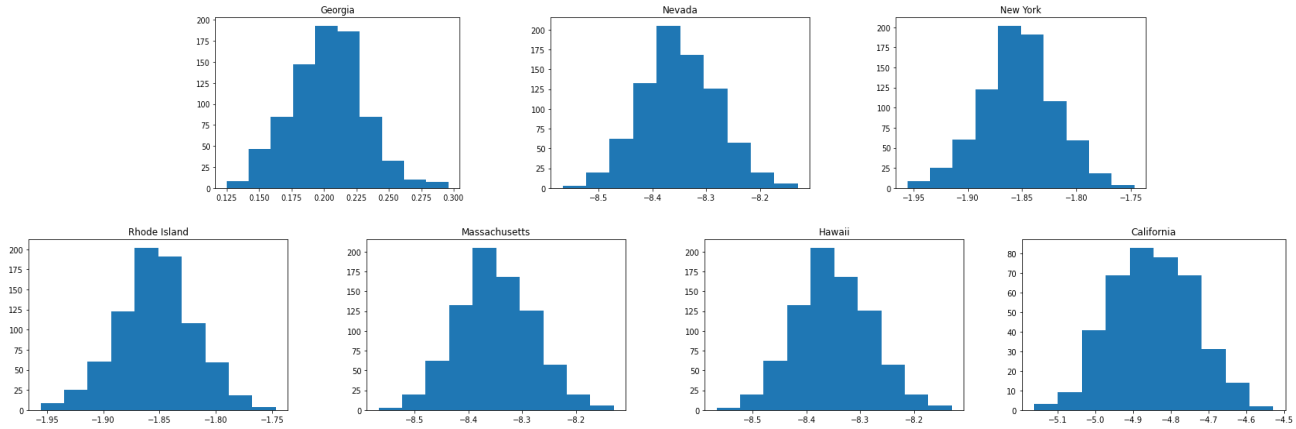


Figure 5: the posterior distribution by MCMC for selected states with sample size = 200 and acceptance rate = 0.8 for each chain.