677 final project

Shicheng Wang

5/1/2020

1.Abstract

This paper focuses on the novel coronavirus situation of global transmission, and introduces multiple data sets about novel coronavirus. Valuable visualization analysis is made at the world level, the American level, the state level and the patient level respectively. In addition, this paper makes a further exploration, comparing the patterns of virus spread in California and New Jersey, and comparing the distribution of patient growth in the two states, so as to infer the impact of different government response measures on the health status of people in the region.

2.Introduction

2.1 Background

Starting in November 2019, covid-19 has gradually spread around the world, endangering the lives of many people. In many countries, including China and the United States, governments are starting to advise citizens to stay at home, keep a social distance, and keep an eye on health and safety. School closures, social shutdowns and other phenomena are what we do not want to see. However, there are still some regional governments that have failed to raise security awareness, leading to serious consequences. The purpose of this study is to analyze the global covid-19 situation and compare the differences in the spread of the virus in different states.

2.2 Data sources

All data included in this research has been collected from multiple resources. Datasets used for EDA part are from CDC and John Hopkins University medical center. The datasets used for distribution analysis and ABtest are from Kaggle.

2.3 Reaserch question

The novel coronavirus pneumonia is studied in this paper, and the impact of different initiatives of New Jersey and California governments on the growth rate of patients is also studied.

2.4 Data Overview

EDA data overview

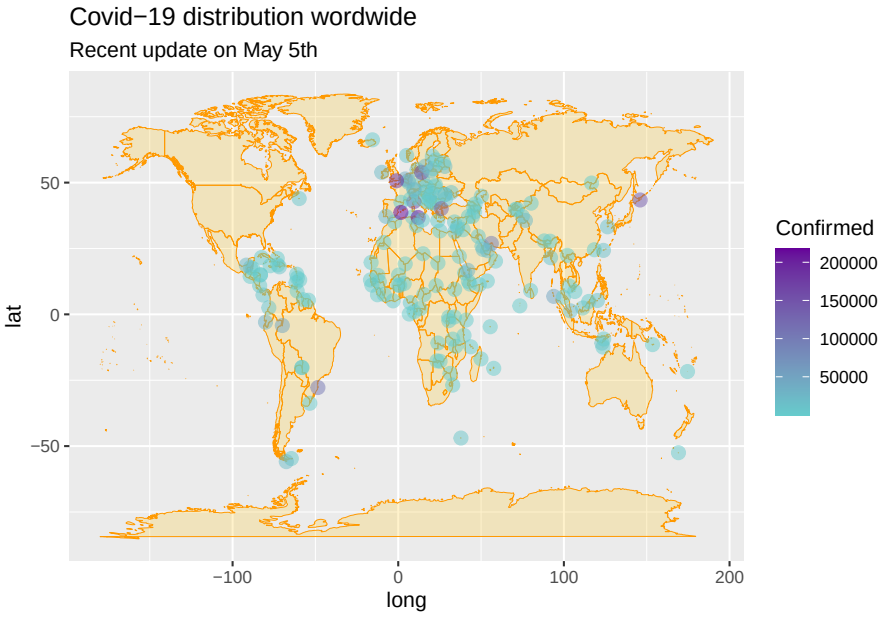
region	long	lat	Confirmed	Deaths	Recovered
Afghanistan	74.891312	37.23164	2894	90	397
Albania	20.063965	42.54727	803	31	543
Algeria	8.576563	36.93721	4648	465	1998
Andorra	1.706055	42.50332	750	45	499
Angola	23.966503	-10.87178	35	2	11
Argentina	-64.549164	-54.71621	4887	260	1442

Analysis data overview

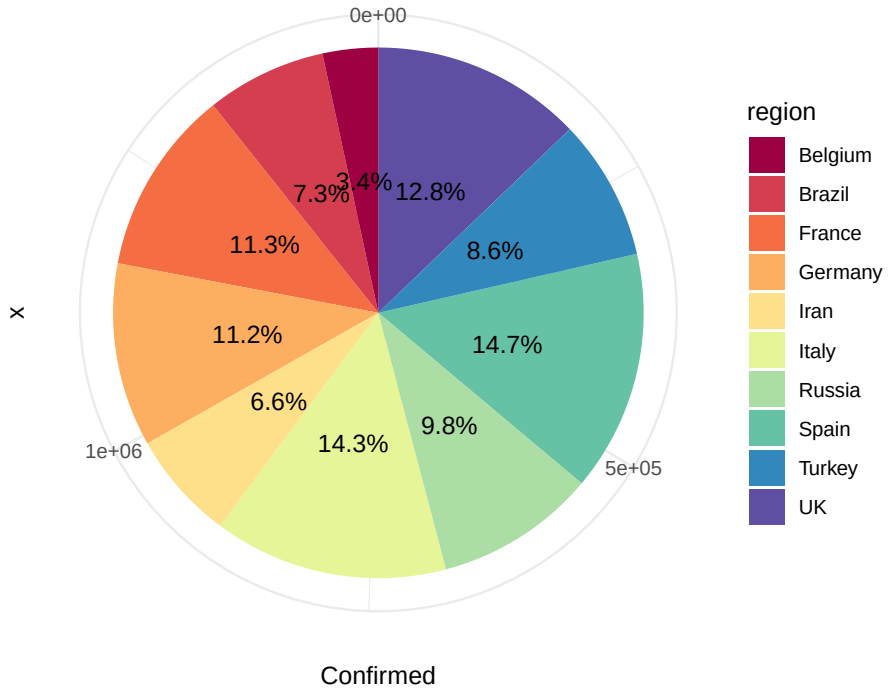
region	long	lat	Confirmed	Deaths	Recovered
Afghanistan	74.891312	37.23164	2894	90	397
Albania	20.063965	42.54727	803	31	543
Algeria	8.576563	36.93721	4648	465	1998
Andorra	1.706055	42.50332	750	45	499
Angola	23.966503	-10.87178	35	2	11
Argentina	-64.549164	-54.71621	4887	260	1442

3.Exploratory data analysis

3.1 Worldwide analysis

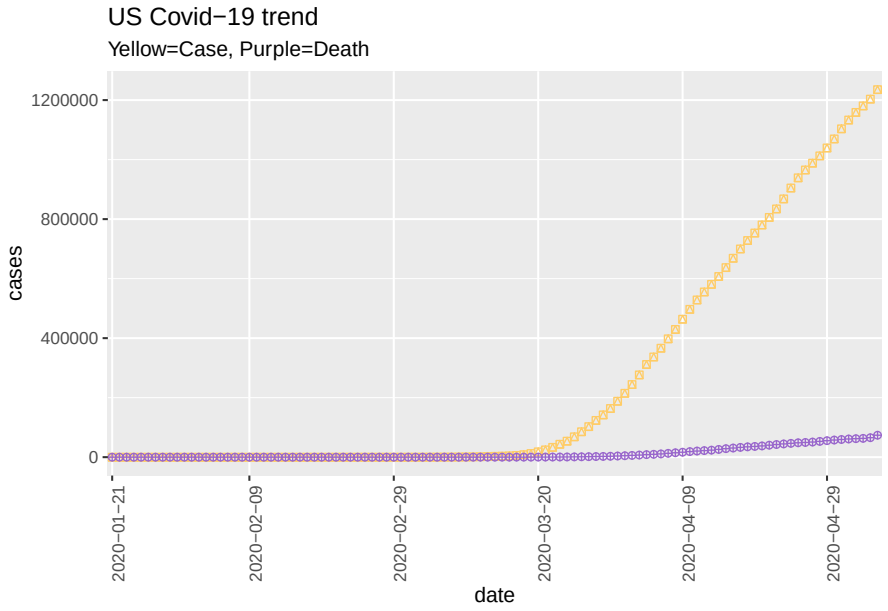


- This picture shows the total number and distribution of cases in various countries around the world. We can see that Europe, America and Asia have serious epidemic areas.

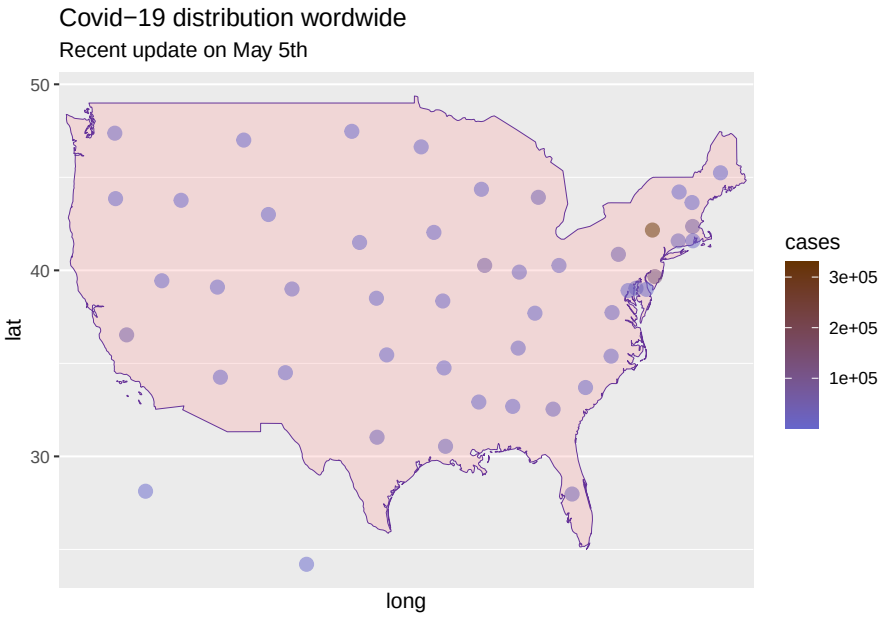


- The pie chart shows the proportion of countries with severe global epidemics (the United States is excluded from the data set and will be analyzed later). We know that America and Europe are still the areas with serious epidemic.

3.2 US-wide analysis

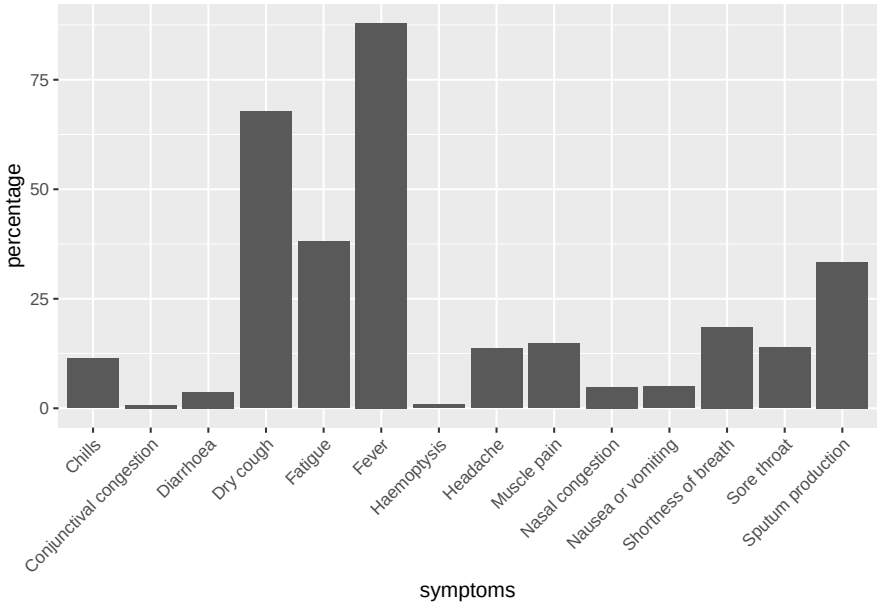


- This chart shows the daily growth of the United States from January 21 to the beginning of May, which is in line with the exponential distribution, while the growth rate does not show a pattern to stop.

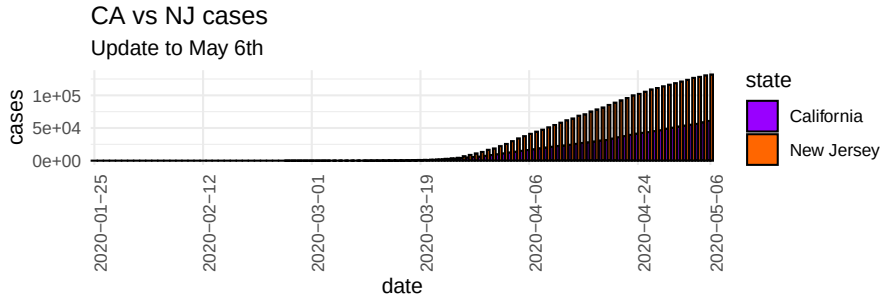


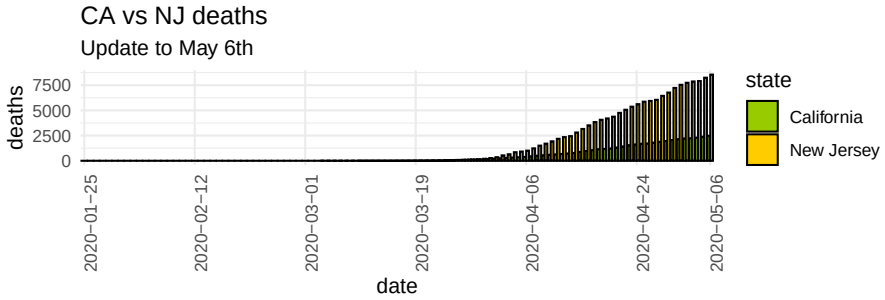
- The eastern part of the United States is the worst, followed by the west coast.

3.3 Patient scope analysis



3.4 CA & NJ analysis





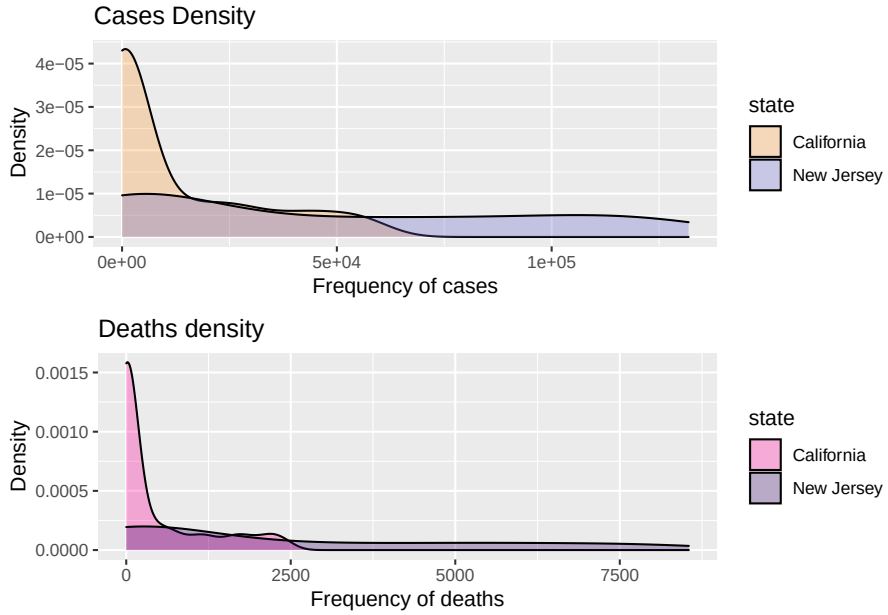
4.AB Test

4.1 Overview

A/B test is the shorthand for a simple controlled experiment. As the name implies, two versions (A and B) of a single variable are compared, which are identical except for one variation that might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, this becomes more complex. This test can give an obvious overview of state comparasion.

Aimming at my research question, the hypothesis could be the following: * Null Hypothesis: New York and California have the same growth pattern. * Alternative Hypothesis: New York and California have different growth pattern.

4.2 Compare distribution



It seems CA has more density at the lower level while the NJ has a high death rate.

construct initial ABtest

state	all	cov
ca	54698	1150
nj	26254	5099

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(1150, 5099) out of c(54698, 26254)
## X-squared = 7467.4, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1781556 -0.1682313
## sample estimates:
##     prop 1     prop 2
## 0.02102453 0.19421802
```

The p-value is less than 0.05, so we can reject the null hypothesis.

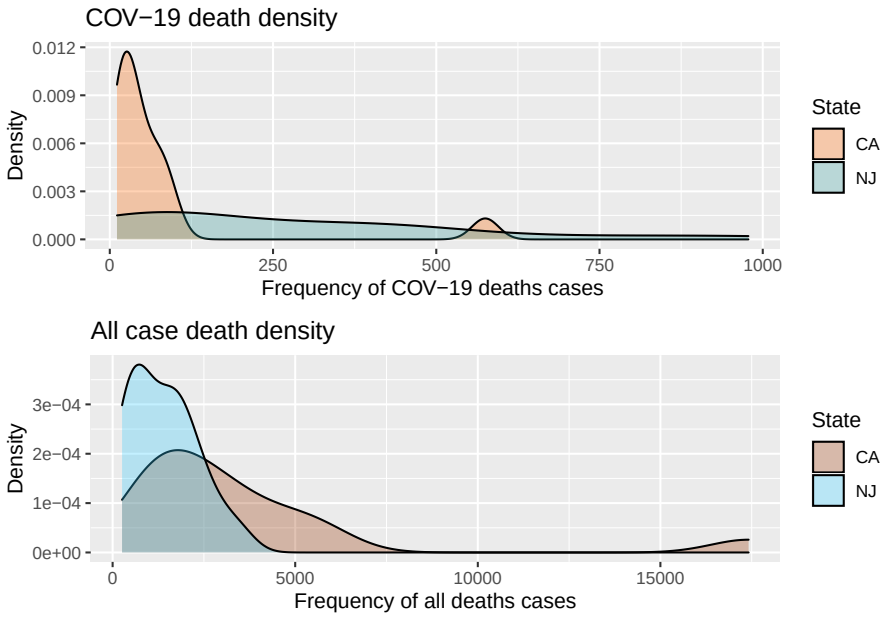
But one cannot directly conclude that A and B have dissimilar death rates. Here true underlying behavior is not known as we are trying to test the hypothesis by carrying out the experiment over a sample.

Bayesian ABtest

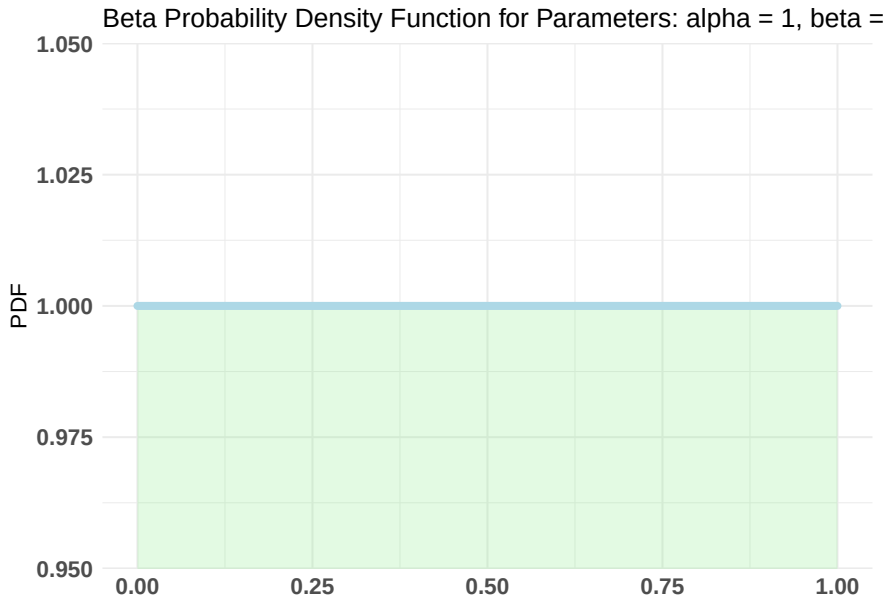
Bayesian statistics in A/B testing is mainly based on past or prior knowledge of similar experiment and the present data. The past knowledge is known as prior also prior probability distribution (Wiki) is combined with current experiment data to make a conclusion on the test at hand.

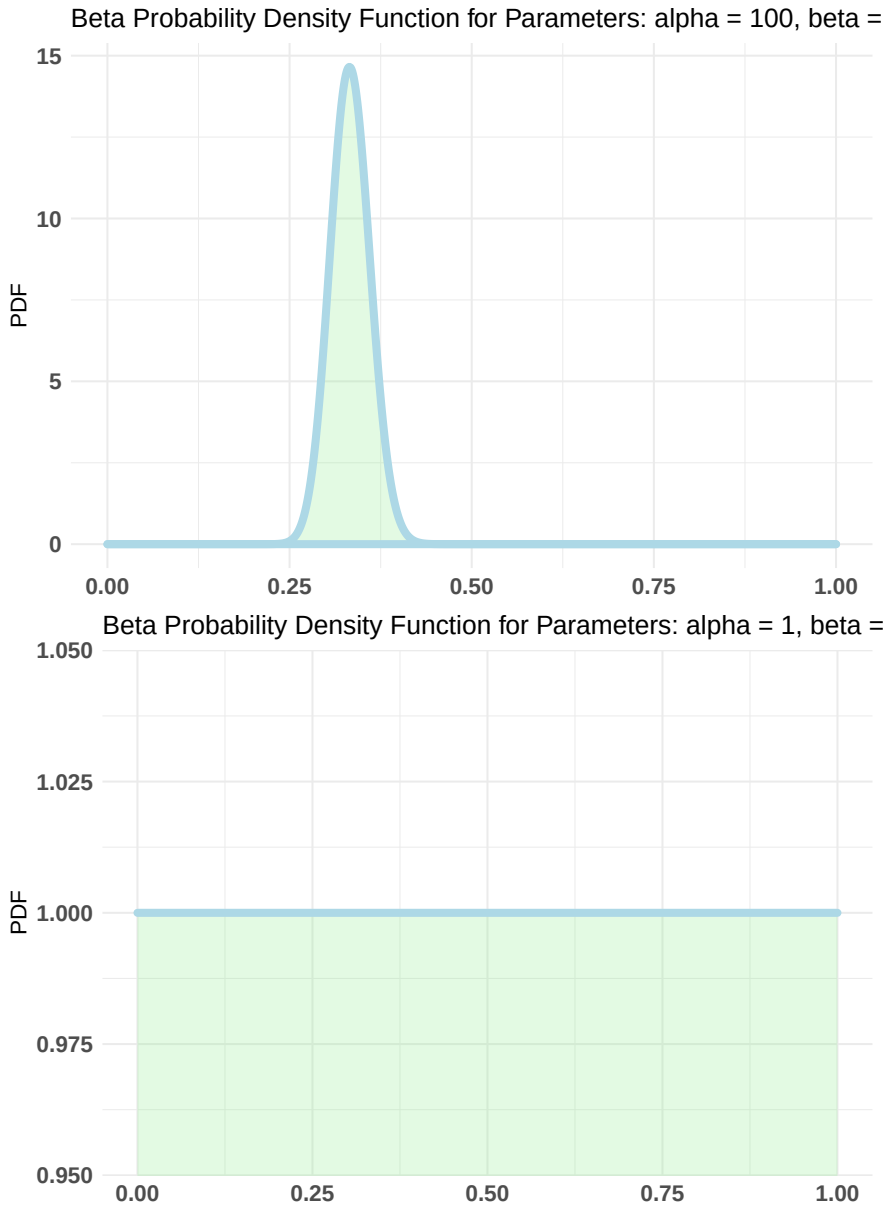
In this method, we model the metric for each variant. We have prior knowledge about the conversion rate for A which has a certain range of values based on the historical data. After observing data from both variants, we estimate the most likely values or the new evidence for each variant.

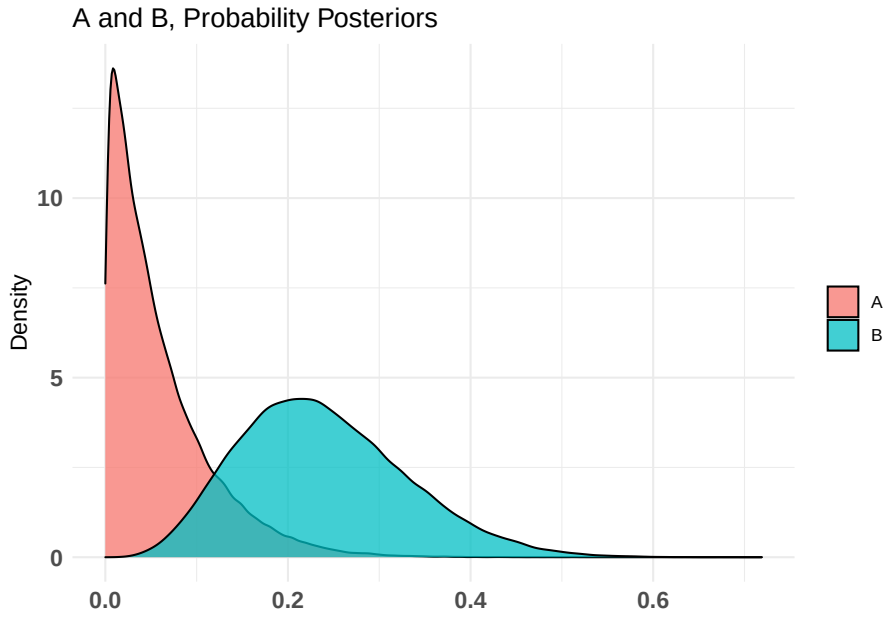
To better understanding the prior distribution, checking the following two plots.

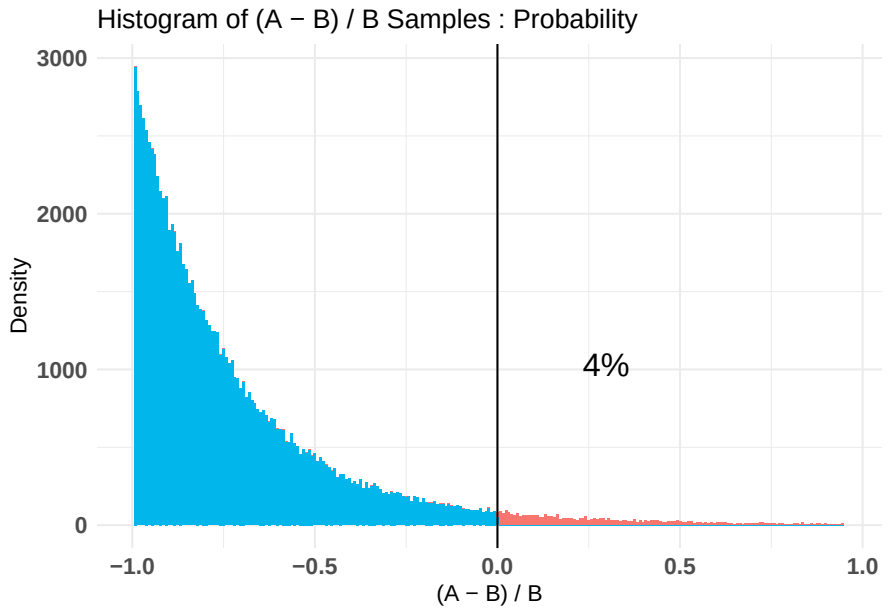


According to two density plot, i could assume the distribution could follow part of binomial distribution.




Beta Probability Density Function for Parameters: alpha = 100, beta =
PDF
15
10
5
0
0.00
0.25
0.50
0.75
1.00
Beta Probability Density Function for Parameters: alpha = 1, beta =
PDF
1.050
1.025
1.000
0.975
0.950
0.00
0.25
0.50
0.75
1.00

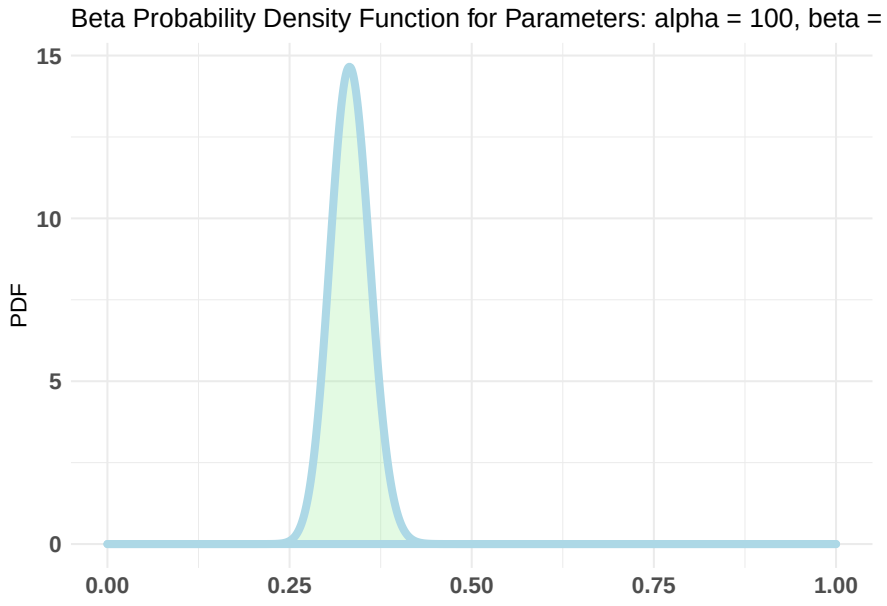


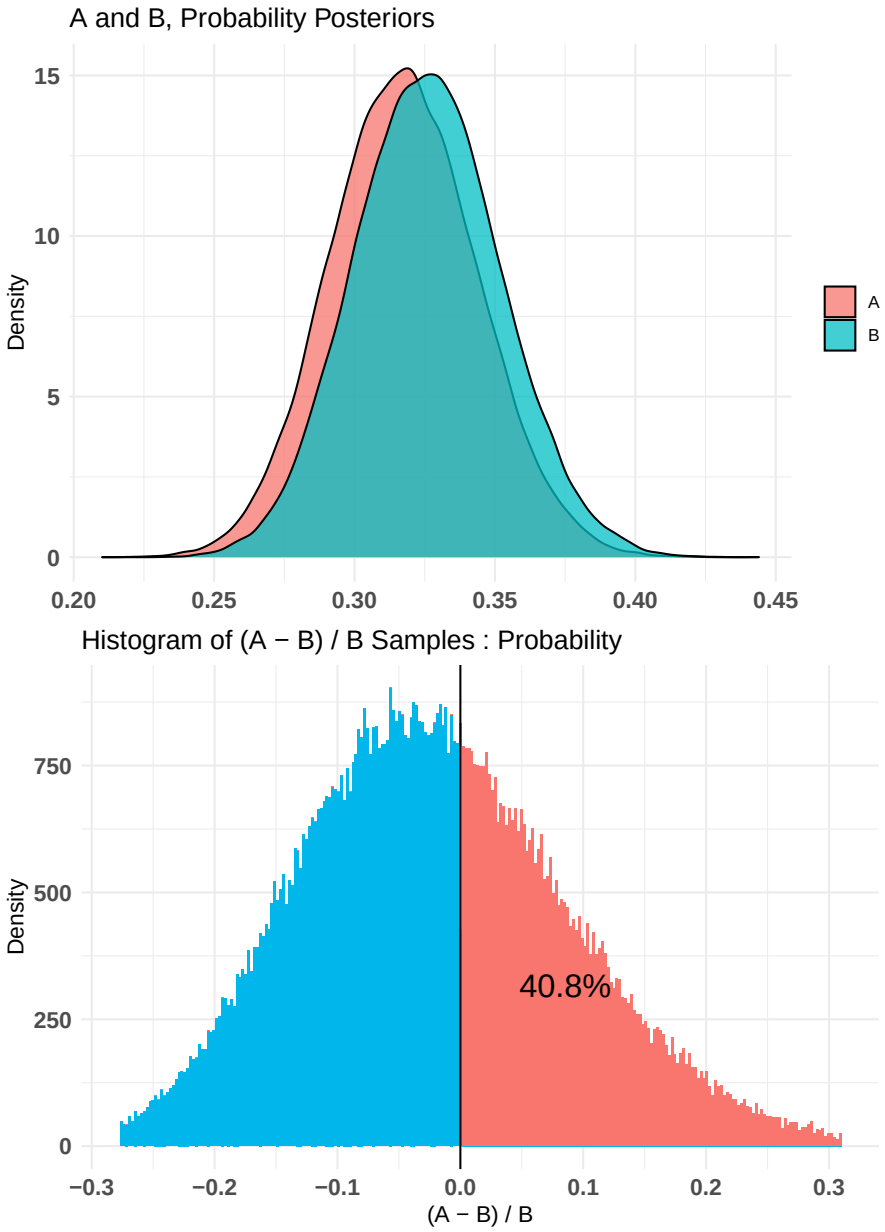



```
## Quantiles of posteriors for A and B:
##
## $Probability
## $Probability$A
##           0%          25%          50%          75%         100%
## 3.072942e-07 1.787531e-02 4.232931e-02 8.282348e-02 6.360226e-01
##
## $Probability$B
##        0%       25%       50%       75%      100%
## 0.0183487 0.1719686 0.2296765 0.2959092 0.7188890
##
##
## ------------------------------------------
##
## P(A > B) by (0)%:
```

```
##
## $Probability
## [1] 0.04037
##
## -----------------------------------------
##
## Credible Interval on (A - B) / B for interval length(s) (0.9) :
##
## $Probability
##          5%         95%
## -0.98650679 -0.07902672
##
## -----------------------------------------
##
## Posterior Expected Loss for choosing B over A:
##
## $Probability
## [1] 51.07426
```

Under the assumption that the statistics follow the Beta(1,1) prior distribution, we can see the summary of the statistics and density distribution in A & B class. The conclusion could be make that New Jersey grew faster than California in all respects in early May. The government needs to strengthen prevention.






```
## Quantiles of posteriors for A and B:
##
## $Probability
## $Probability$A
##        0%       25%       50%       75%      100%
## 0.2101849 0.2995499 0.3171644 0.3350064 0.4439486
##
## $Probability$B
##        0%       25%       50%       75%      100%
## 0.2197870 0.3081022 0.3257672 0.3435036 0.4438761
##
##
## ------------------------------------------
##
## P(A > B) by (0)%:
```

```
##
## $Probability
## [1] 0.40778
##
## ------------------------------------------
##
## Credible Interval on (A - B) / B for interval length(s) (0.9) :
##
## $Probability
##         5%        95%
## -0.1949853  0.1772941
##
## ------------------------------------------
##
## Posterior Expected Loss for choosing B over A:
##
## $Probability
## [1] 0.06586604
```

Under the assumption that the statistics follow the Beta(100,200) prior distribution, we can see there are relative high proportion of overlap of both two states. We can assume Null hypothesis true in some way.

5.Discussion

There could be several disadvantages during the analysis.

- The datasets are collected from different resourses which could lower the accuracy of this analysis.
- The Disadvantages of Using A/B Test: The A/B test considers the sample data of the whole population at a certain time point. Therefore, the test is limited to a certain point in time. Second, the sample data may not tell us the true mortality rate of the original population.
- The Disadvantages of Using Bayesian A/B Test: The prior probability is artificially speculated and may not be accurate In addition, the running time of Bayesian algorithm is too long.

6.Reference

- https://github.com/CSSEGISandData/COVID-19/tree/master/archived_data
- https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Place-of-Deat/uggs-hy5q
- https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku
- https://www.csdn.net/
- https://www.kaggle.com/