544 Final Report

Mandatory Tasks

1. Clean the dataset

(1) Split the 4 cases with corresponding date and compute the daily stats for each case:

Step1: Split each column and get the cumulative data of each case and its corresponding date

Step2: For each cumulative data, subtract the data last one day except the first day to get the daily stats of the case

Then get the following data lists for each case:

CA confirmed as **confirmed_CA**

CO confirmed as **confirmed_CO**

CA deaths as **deaths_CA**

CO deaths as **deaths_CO**

(2) Remove outliers by Tukey's rule:

Step1: Sorted the list of each case and compute the $Q1 = 25\%ile$ and $Q3 = \%75ile$

Step2: Compute the $IOR = Q3 - Q1$ for each list

Step3: Compute the $lower\ bound = Q1 - \alpha \cdot IOR$ and $upper\ bound = Q3 + \alpha \cdot IOR$ of outliers for each case.

CA confirmed:

lower bound of outlier = -9514.0

upper bound of outlier = 19854.0

CO confirmed:

lower bound of outlier = -1452.5

upper bound of outlier = 3063.5

CA deaths:

lower bound of outlier = -159.5

upper bound of outlier = 348.5

CO deaths:

lower bound of outlier = -24.5

upper bound of outlier = 43.5

Step4: For each list, keep the data in the above range and 0s (keep the corresponding date as well). Then report the number of outliers by using the length of the original list to subtract the length of the cleaned list.

CA confirmed:

number of outliers = 55

CO confirmed:
number of outliers = 47
CA deaths:
number of outliers = 51
CO deaths:
number of outliers = 47

(3) Check if there is any negative value in these 4 lists and replace it by 0, and only find negative values in CO deaths data list:

negative value and corresponding date:
-5, 2020-07-07
and
-8, 2020-12-20

Then replace these 2 values with 0s.

(4) Write the data list for each case with corresponding date:

We write the csv file for both data after removing outlier and original one as:

CA_confirmed_cleaned.csv
CA_confirmed_origin.csv
CO_confirmed_cleaned.csv
CO_confirmed_origin.csv
CA_deaths_cleaned.csv
CA_deaths_origin.csv
CO_deaths_cleaned.csv
CO_deaths_origin.csv

We also split the cumulative data for each case with corresponding date and write the csv files for convenience as:

CA_confirmed.csv
CO_confirmed.csv
CA_deaths.csv
CO_deaths.csv

2. Solve the required inferences for the COVID19 dataset

a. Predict COVID19 stats for each state using four prediction techniques:

AR:

(1) Load the cleaned data (no missing data) in the first 4 weeks of August .2020 for each case and get the 4 data list.

(2) For the dth day prediction, use the data list of days before it ($1st\ to\ (d-1)th$) to build the X and Y for Regression. X is a matrix with $(n-p)\times(p+1)$ size ($(i+1)th$ column is the ith data to $(n-p-1+i)th$ data in the list and append the constant column as first column), and Y is a matrix with $(n-p)\times 1$ size (from $(p+1)th$ data to nth data in

the list), where n is the number of data list and p is the parameter for AR. Then compute the weight matrix $\hat{\beta} = (X^T X)^{-1} X^T Y$ with $(p+1)\times 1$ size

(3) Compute each prediction as $\hat{y}_i = [1, y_{i-p}, ..., y_{i-1}] \bullet \hat{\beta}_i$. And combine these 7 predictions to $\hat{Y}$ and the last 7 days true values as Y. Then compute the error matrix $\varepsilon = \hat{Y} - Y$.

EWMA:

(1) Load the cleaned data (no missing data) in the first 4 weeks of August .2020 for each case and get the 4 data list.

(2) Compute prediction of each day in last week $\hat{y}_{t+1} = \alpha \cdot \sum_{i=1}^{t} (1-\alpha)^{i-1} \bullet y_{t+1-i}$ (where t from 21 to 27 and α is the parameter for EWMA) and get the $Y = [y_{22}, y_{23}, y_{24}, y_{25}, y_{26}, y_{27}, y_{28}]$ and $\hat{Y} = [\hat{y}_{22}, \hat{y}_{23}, \hat{y}_{24}, \hat{y}_{25}, \hat{y}_{26}, \hat{y}_{27}, \hat{y}_{28}]$

(3) Compute the error matrix $\varepsilon = \hat{Y} - Y$

Compute the MAPE and MSE:

MAPE = $\frac{1}{n} \cdot 100 \sum \left| \hat{\varepsilon}_i \right| / y_i$

(Note: when meet $y_i = 0$ just discard this data to compute the MAPE)

MSE = $\frac{1}{n} \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2$

Accuracy for 4 data lists:

<u>CA confirmed:</u>

CA confirmed for AR(3):
MAPE = 37.79898180994047%
MSE = 4049771.0908803567

CA confirmed for AR(5):
MAPE = 47.26586637005722%
MSE = 5840284.910336106

CA confirmed for EWMA(0.5):
MAPE = 30.25583357208969%
MSE = 3489073.2748513333

CA confirmed for EWMA(0.8):
MAPE = 29.96836941920888%
MSE = 4132008.6987732253

<u>CO confirmed:</u>

CO confirmed for AR(3):
MAPE = 37.89426668632648%
MSE = 14221.302456601095

CO confirmed for AR(5):
MAPE = 33.9249266588338%
MSE = 15693.24974268478

CO confirmed for EWMA(0.5):
MAPE = 36.1981098850506%
MSE = 14791.329900243762

CO confirmed for EWMA(0.8):
MAPE = 41.975399050127116%
MSE = 20175.36116651393

CA deaths:

CA deaths for AR(3):
MAPE = 46.12369849140459%
MSE = 1440.9920692434937

CA deaths for AR(5):
MAPE = 38.45012279981787%
MSE = 1351.9920878028438

CA deaths for EWMA(0.5):
MAPE = 56.24950744775232%
MSE = 2247.3163982473134

CA deaths for EWMA(0.8):
MAPE = 59.504970913058855%
MSE = 2516.5880185936544

CO deaths:

CO deaths for AR(3):
MAPE = 114.33275403504017%
MSE = 8.911501479770108

CO deaths for AR(5):
MAPE = 127.05100543984918%
MSE = 8.71841955158698

CO deaths for EWMA(0.5):
MAPE = 126.18158575561311%
MSE = 14.425718096480448

CO deaths for EWMA(0.8):
MAPE = 122.20158821725467%
MSE = 18.551278644491326

b. Analyze how the mean of monthly COVID19 stats has changed

In this part, we apply the Wald's test, Z-test, and t-test to check whether the mean of COVID19 deaths and cases are different for Feb'21 and March'21 in the two states (CA and CO). Here we use the death data for Mar'21 in CA from the original dataset since this part of data is treated as outliers after data-cleaning.

1. One-sample test

(1) Wald's Test

Let's set: H_0: $\widehat{\mu} = \mu_0$, that is, the means of COVID19 data(cases or deaths) are the same for Feb'21 and Mar'21. And W can be obtained by

$$W = \left|(\widehat{\mu} - \mu_0) / Se(\widehat{\mu})\right| = \left|(\widehat{\mu} - \mu_0) / \sqrt{Var(\widehat{\mu})}\right|.$$

Since we use MLE for Wald's test as the estimator and assume for Wald's estimator purposes that daily data is Poisson distributed, we can have

$$\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=0}^{n} X_i$$

Based on the article, the sample mean of daily values from Feb'21 is used as a guess for mean of daily values for March'21. Therefore, we can have

$\widehat{\mu} = \frac{1}{n}\sum_{i=0}^{n} X_i$ and $\mu_0 = \frac{1}{m}\sum_{j=0}^{n} Y_j$,

where X_i indicates the data (the number of daily deaths and cases) for March'21 from CA or CO and Y_j indicates the data for Feb'21 from CA or CO.

Therefore, we can have results for the one-sample Wald's Test:

For monthly cases difference in two states, based on $\alpha = 0.05$, $Z_{0.025} = 1.962$, we can have

1) $W_{CA} = 8.863 > 1.962$. We reject H_0, that is, under one-sample Wald's Test, the means of COVID19 cases are different for Feb21 and Mar21 in CA.
2) $W_{CO} = 3.439 > 1.962$. We reject H_0, that is, under one-sample Wald's Test, the means of COVID19 cases are different for Feb21 and Mar21 in CO.

For monthly deaths difference in two states, based on $\alpha = 0.05$, $Z_{0.025} = 1.962$, we can have

1) $W_{CA} = 5.196 > 1.962$. We reject H_0, that is, under one-sample Wald's Test, the means of COVID19 deaths are different for Feb21 and Mar21 in CA.
2) $W_{CO} = 4.953 > 1.962$. We reject H_0, that is, under one-sample Wald's Test, the means of COVID19 deaths are different for Feb21 and Mar21 in CO.

(2) Z-Test

Based on the same H_0, we can obtain Z from

$$Z = \left| (\overline{X} - \mu_0) / (\sigma/\sqrt{n}) \right|$$

Here we use the corrected sample standard deviation of the entire COVID19 dataset of each state as the true sigma value, that is, $\sigma = \sqrt{Var(D)}$, where D indicates the whole number of deaths or cases from CA or CO. And n is the number of the sample data, in this part, $n = the\ number\ of\ Mar'21\ from\ CA\ or\ CO$.

Therefore, we can have results for the one-sample Z-Test:

For monthly cases difference in two states, based on $\alpha = 0.05$, $Z_{0.025} = 1.962$, we can have

1) $Z_{CA} = 0.702 > 1.962$. We accept H_0, that is, under one-sample Z Test, the means of COVID19 cases are similar for Feb21 and Mar21 in CA.
2) $Z_{CO} = 1.116 > 1.962$. We accept H_0, that is, under one-sample Z Test, the means of COVID19 cases are similar for Feb21 and Mar21 in CO.

For monthly deaths difference in two states, based on $\alpha = 0.05$, $Z_{0.025} = 1.962$, we can have

1) $Z_{CA} = 2.891 > 1.962$. We reject H_0, that is, under one-sample Z Test, the means of COVID19 deaths are different for Feb21 and Mar21 in CA.
2) $Z_{CA} = 0.650 > 1.962$. We accept H_0, that is, under one-sample Z Test, the means of COVID19 deaths are similar for Feb21 and Mar21 in CO.

(3) T-test

$$T = \left| (\overline{X} - \mu_0) / (S/\sqrt{n}) \right|, \text{ where } S = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n} (X_i - \overline{X})}$$

Therefore, we can have results for the one-sample T-test:

For monthly cases difference in two states, based on $\alpha = 0.05$, and $t_{(n-1,\,0.025)} = t_{(30,\,0.025)} = 2.042$, we can have

1) $t_{CA} = 48.547 > 2.042$. We reject H_0, that is, under one-sample T-test, the means of COVID19 cases are different for Feb21 and Mar21 in CA.
2) $t_{CO} = 18.834 > 2.042$. We reject H_0, that is, under one-sample T-test, the means of COVID19 cases are different for Feb21 and Mar21 in CO.

For monthly cases difference in two states, based on $\alpha = 0.05$, and $t_{(n-1,\,0.025)} = t_{(30,\,0.025)} = 2.042$, we can have

1) $t_{CA} = 28.458 > 2.042$. We reject H_0, that is, under one-sample T-test, the means of COVID19 deaths are different for Feb21 and Mar21 in CA.
2) $t_{CO} = 27.126 > 2.042$. We reject H_0, that is, under one-sample T-test, the means of COVID19 deaths are different for Feb21 and Mar21 in CO.

2. **Two-sample test**

(1) Wald's Test

Let's set $H_0: \mu_a = \mu_b$, where μ_a indicates the mean number of COVID19 data (cases or deaths) for Feb'21 from CA or CO (sample A mean), while μ_b indicates the mean number of COVID19 data (same type as μ_a) for Mar'21 from the same state.

Therefore, we can obtain W value from the formula

$W = \left|(\widehat{\mu} - \mu_0) / Se(\mu_a - \mu_b)\right|$, where $\widehat{\mu} = \mu_a - \mu_b$, $\mu_0 = 0$, and

$Se(\mu_a - \mu_b) = \sqrt{Var(\mu_a) + Var(\mu_b)}$, since the data from Feb'21 and Mar'21 are unpaired data.

Therefore, we can have results for the two-sample Wald's Test:

For monthly cases difference in two states, based on $\alpha = 0.05$, $Z_{0.025} = 1.962$, we can have

1) $W_{CA} = 2.341 > 1.962$. We reject H_0, that is, under two-sample Wald's Test, the means of COVID19 cases are different for Feb21 and Mar21 in CA.
2) $W_{CO} = 2.442 > 1.962$. We reject H_0, that is, under one-sample Wald's Test, the means of COVID19 cases are different for Feb21 and Mar21 in CO.

For monthly deaths difference in two states, based on $\alpha = 0.05$, $Z_{0.025} = 1.962$, we can have

1) $W_{CA} = 2.444 > 1.962$. We reject H_0, that is, under two-sample Wald's Test, the means of COVID19 deaths are different for Feb21 and Mar21 in CA.
2) $W_{CO} = 2.154 > 1.962$. We reject H_0, that is, under two-sample Wald's Test, the means of COVID19 deaths are different for Feb21 and Mar21 in CO.

(2) Two-sample unpaired T-test

Let's set $H_0: \overline{X_a} = \overline{X_b}$ where $\overline{X_a}$ indicates the mean number of COVID19 data (cases or deaths) for Feb'21 from CA or CO (sample A mean), while $\overline{X_a}$ indicates the mean number of COVID19 data (same type as $\overline{X_a}$) for Mar'21 from the same state.

Therefore we can obtain T value from the formula:

$T = \left|(\overline{X} - \mu_0) / (S_a/\sqrt{n_a} + S_b/\sqrt{n_b})\right|$, where $\overline{X} = \overline{X_a} - \overline{X_b}$, $\mu_0 = 0$,and

$S_a = \sqrt{\frac{1}{n_a - 1}\sum_{i=0}^{n}(X_{ai} - \overline{X_a})}$ and $S_b = \sqrt{\frac{1}{n_b - 1}\sum_{i=0}^{n}(X_{bi} - \overline{X_b})}$.

Here we calculate:
$\alpha = 0.05, n = n_a + n_b - 2 = 28 + 31 - 2 = 57$, so we have $t_{(56,0.025)} = 2.003$.

Therefore, we can have results for the two-sample unpaired T-test:
For monthly cases difference in two states, based on $\alpha = 0.05$, $t_{(56,0.025)} = 2.003$, we can have

1) $t_{CA} = 12.206 > 2.042$. We reject H_0, that is, under two-sample T-test, the means of COVID19 cases are different for Feb21 and Mar21 in CA.
2) $t_{CO} = 13.026 > 2.042$. We reject H_0, that is, under two-sample T-test, the means of COVID19 cases are different for Feb21 and Mar21 in CO.

For monthly cases difference in two states, based on $\alpha = 0.05$, $t_{(56,0.025)} = 2.003$, we can have

1) $t_{CA} = 12.840 > 2.042$. We reject H_0, that is, under two-sample T-test, the means of COVID19 deaths are different for Feb21 and Mar21 in CA.
2) $t_{CO} = 11.302 > 2.042$. We reject H_0, that is, under two-sample T-test, the means of COVID19 deaths are different for Feb21 and Mar21 in CO.

c. Inference the equality of distributions in the two states

Inference the equality of distributions in the two states (distribution of daily #cases and daily #deaths) for the last three months of 2020 (Oct, Nov, Dec) of your dataset using K-S test and Permutation test.

1) K-S test, 1-sample

H_0: The distribution of confirmed cases in CO and Poisson/Geometric/Binomial distribution derived from CA through MME is the same.

Step 0: Obtain parameters of the poisson/Geometric/Binomial distribution from the CA dataset using MME, say $mean = \bar{X}$.
Step 1: $D(F_y, F_x) = max_{\backslash alpha}|F_x - F_y|$
Step2: if $D > 0.05$, reject H_0.
We use this strategy for both confirmed cases and death cases, along poisson/Geometric/Binomial distributions.

a) the distribution of daily cases
The mean and variance of CA are 6669.0645 and 19543804.0926.

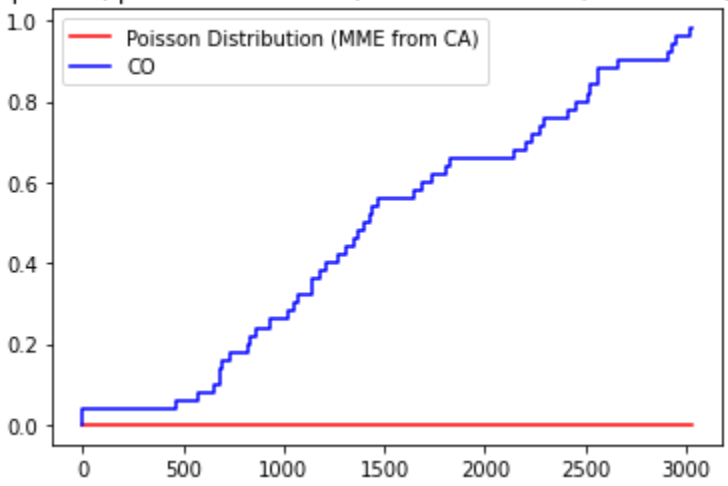
Poisson!
estimated μ = mean (MME from CA) = 6669.0645

H_0: The distribution of confirmed cases in CO and poisson distribution derived from CA through MME is the same.

The max difference $D(F_y, F_x)$= 1.0 is greater than threshold 0.05, reject H_0.

The Figure below shows the distribution:

K-S test, 1-sample test, poisson distribution, estimate from CA, test on CO, confirmed cases.



Geometric!

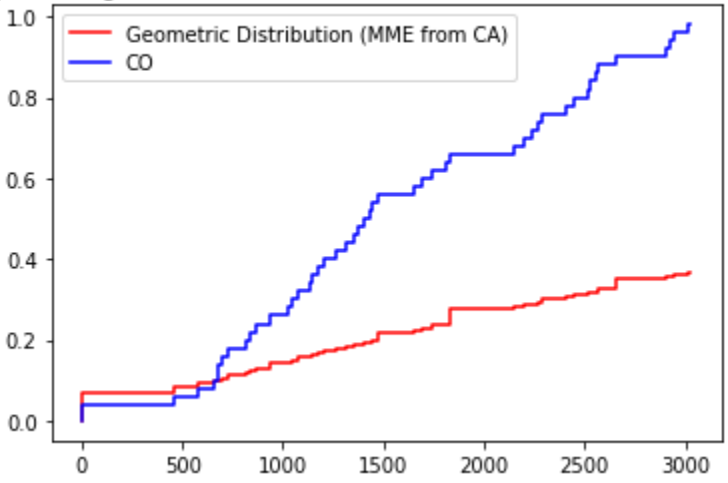
estimated p = 1 / mean (MME from CA) = 0.00014994606778529659

H_0: The distribution of confirmed cases in CO and geometric distribution derived from CA through MME is the same.

$D(F_y, F_x)$= 0.6350378108823924, which is greater than the threshold 0.05, we reject H_0.

The Figure below shows the distribution:

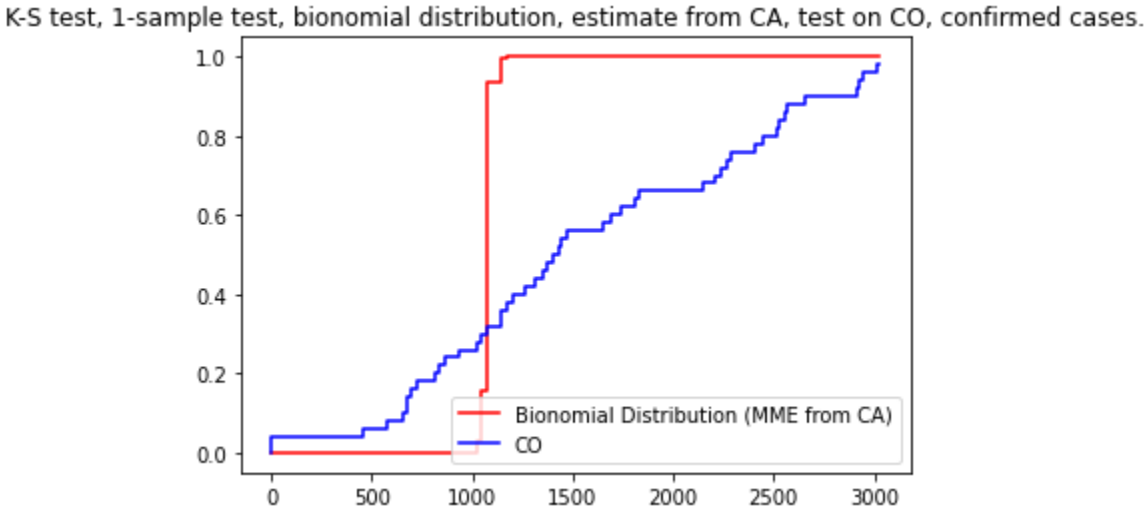
K-S test, 1-sample test, geometric distribution, estimate from CA, test on CO, confirmed cases.



Binomial!

estimated p = mean / n (MME from CA) = 0.362981794814621

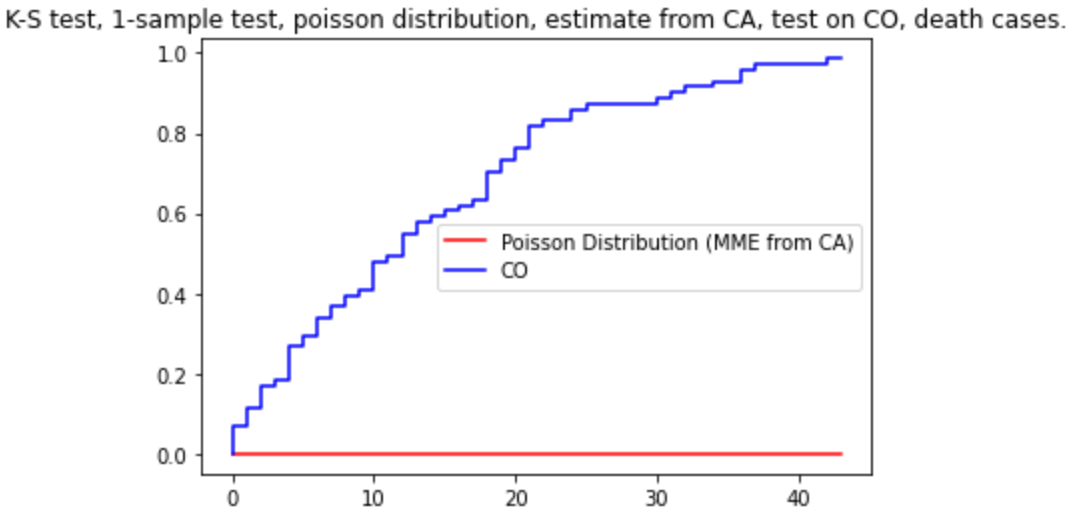
H_0: The distribution of confirmed cases in CO and binomial distribution derived from CA through MME is the same.
$D(F_y, F_x)$=0.6385931576623072, greater than threshold 0.05, we reject H_0.



b) distribution of daily #deaths

The mean and variance of CA are 86.41379310344827 and 5369.02417756639.

Poisson!
estimated μ = mean (MME from CA) = 86.41379310344827
H_0: The distribution of death cases in CO and poisson distribution derived from CA through MME is the same.
The max difference $D(F_y, F_x)$= 0.9999998210180034 is greater than threshold 0.05, we reject H_0. The Figure below shows the distribution:



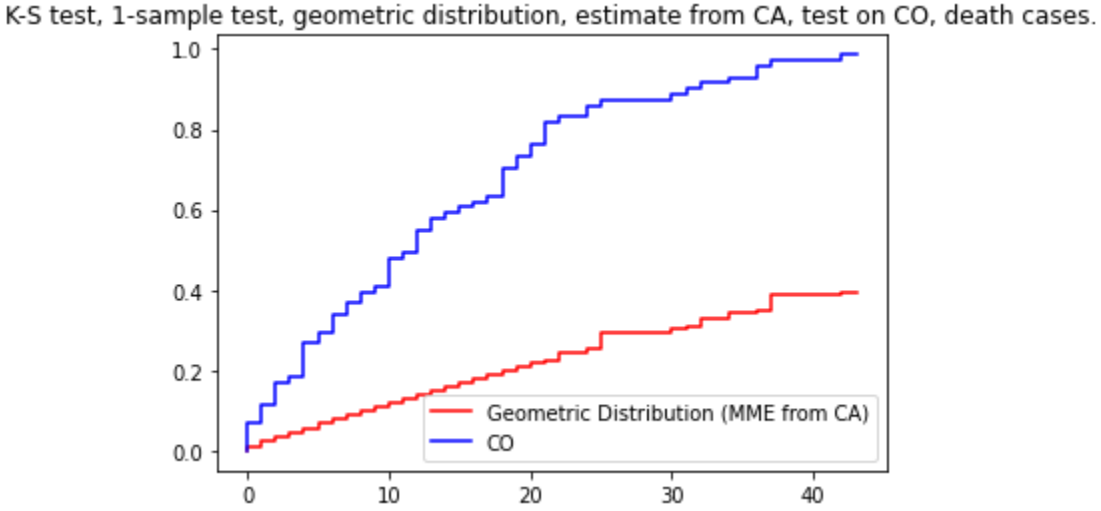
Geometric!

estimated p = 1 / mean (MME from CA) = 0.01157222665602554

H_0: The distribution of death cases in CO and geometric distribution derived from CA through MME is the same.

$D(F_y, F_x)$= 0.6219049570862651, which is greater than the threshold 0.05, we reject H_0.

The Figure below shows the distribution:

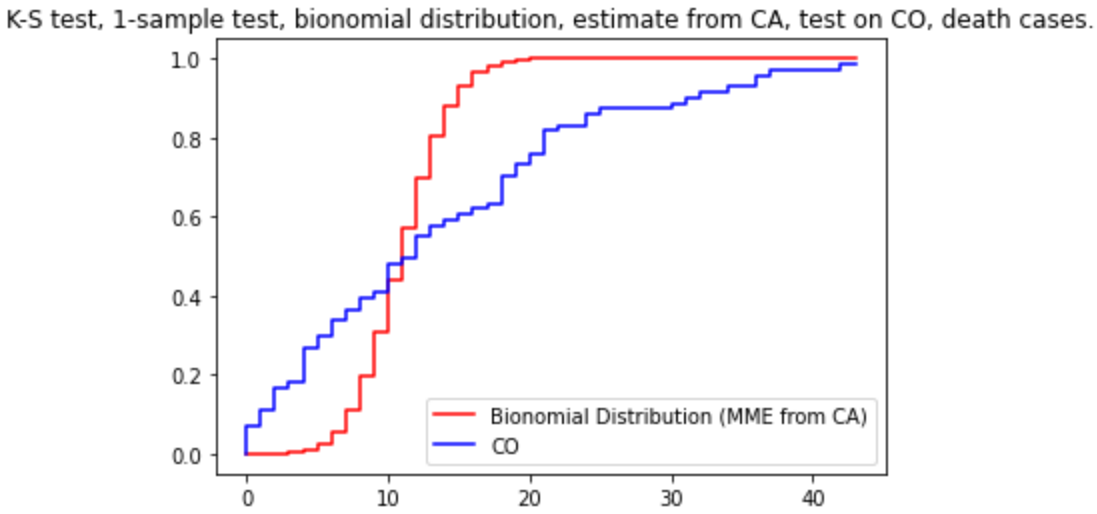


Binomial!

estimated p = mean / n (MME from CA) = 0.279656288360674

H_0: The distribution of death cases in CO and binomial distribution derived from CA through MME is the same.

$D(F_y, F_x)$= 0.34957451889394886, greater than threshold 0.05, we reject H_0. The figure shows as below:



2) K-S test, two-sample

K-S two-sample test is similar to the K-S one-sample test; the difference is that we compute the CDF of both two samples from their empirical CDF. And We use this strategy for both confirmed cases and death cases.

H_0: The distribution of confirmed cases in CA and CO is the same.

Step1: $D(F_y, F_x) = max_{\backslash alpha}|F_x - F_y|$

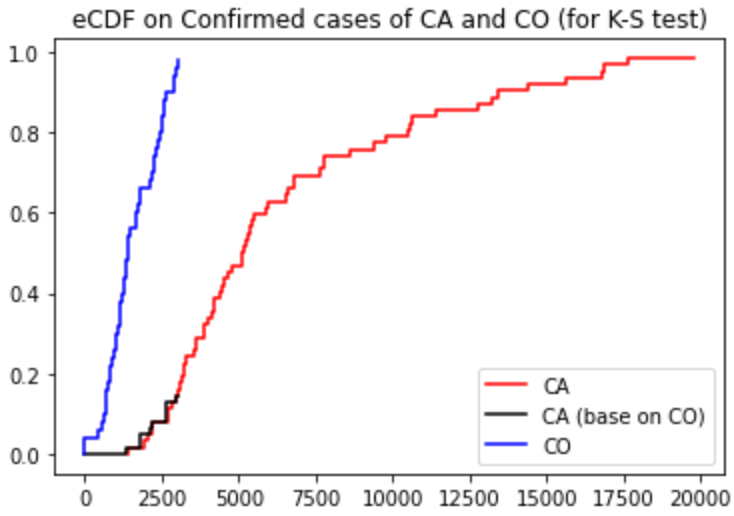
Step2: if $D > 0.05$, reject H_0.

c) distribution of daily #cases

H_0: The distribution of confirmed cases in CA and CO is the same.

estimated μ = mean (MME from CA) = 86.41379310344827

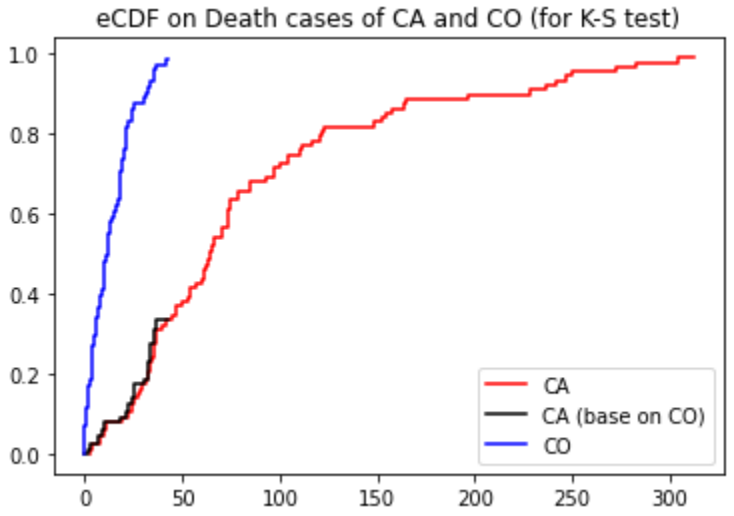
The max difference $D(F_y, F_x)$= 0.8548387096774194 is greater than threshold 0.05, we reject H_0. The Figure below shows the distribution:



d) distribution of daily #deaths

H_0: The distribution of death cases in CA and CO is the same.

The max difference $D(F_y, F_x)$= 0.7353084021369596 is greater than threshold 0.05, we reject H_0. The Figure below shows the distribution:



3) Permutation test:

H_0: The distribution of confirmed/death cases in CA and CO is the same.

Step0: Compute $T_{obs} = |D_1 - D_2|$

Step1: For each $i = 1\ to\ N!$, compute $T_i = |D_1^i - D_2^i|$

Step3: p-value = $sum\ I(I_i > T_{obs}) / N!$

Step4: if p-value <= 0.05, then reject H_0.

e) distribution of daily #cases

H_0: The distribution of confirmed cases in CA and CO is the same.

t_{obs} = 5085.804516129032

T_i(count_extreme)= 0

$p = 0.0 \leq 0.05$, so we reject H_0

f) distribution of daily #deaths

H_0: The distribution of death cases in CA and CO is the same.

t_{obs} = 72.75182127246235

T_i(count_extreme)= 0

$p = 0.0 \leq 0.05$, so we reject H_0

d. Analyze how the mean of monthly COVID19 stats has changed

Follow the instructions from slides about how to do Bayesian Inference.
Step 0: Get the prior distribution for p from MME.
Step 1: Observe D_5, and update the posterior distribution of p, based on the observe D_5 and prior distribution from step0.
Step 2: Iteratively repeat Step1, Observe D_6, and update the posterior distribution of p based on the observe $D_6$and prior distribution from Step1 (which is the posterior distribution obtained from step 1).
Step3: Iteratively repeat Step1, until observe week8's data.

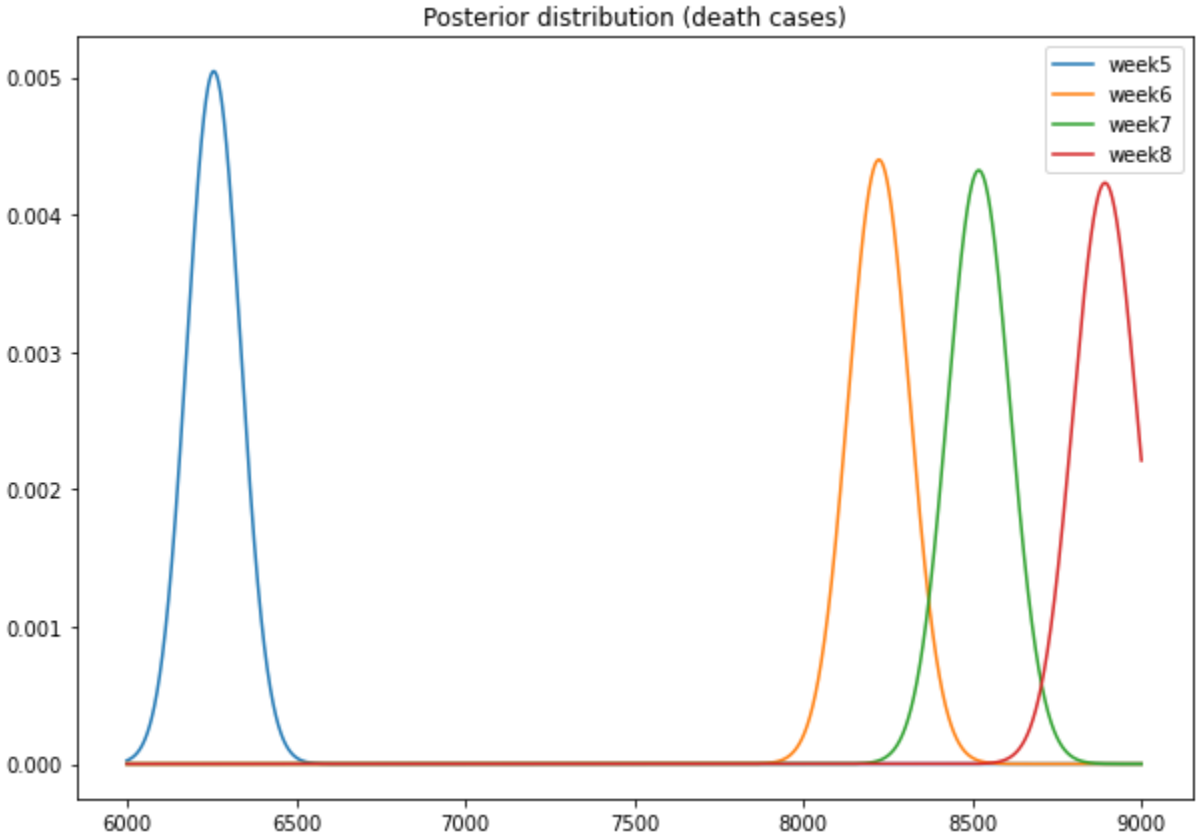
For MAP, just compute value when the posterior distribution's gradient equals to 0, which in our poisson distribution case, it equals to the mean (MME estimation of mu) of observed data D .

g) distribution of daily #cases

Combine daily confirmed from CA and CO states.
The mean and variance of CA are 3826.6428571428573 and 1308324.5153061224 respectively.
First Prior is the estimated μ of Poisson, which is the mean of the first four weeks:
μ= 3826.6428571428573

And the μ values from week 5 to week 8 are:
μ of poisson in posterior distribution (after week5) is: 6257.285714285715
μ of poisson in posterior distribution (after week6) is: 8223.57142857143
μ of poisson in posterior distribution (after week7) is: 8518.714285714286
μ of poisson in posterior distribution (after week8) is: 8891.82142857143

The figure below shows the posterior distribution of daily cases from week 5 to week 8, and we can obtain the four MAP values from the figure:



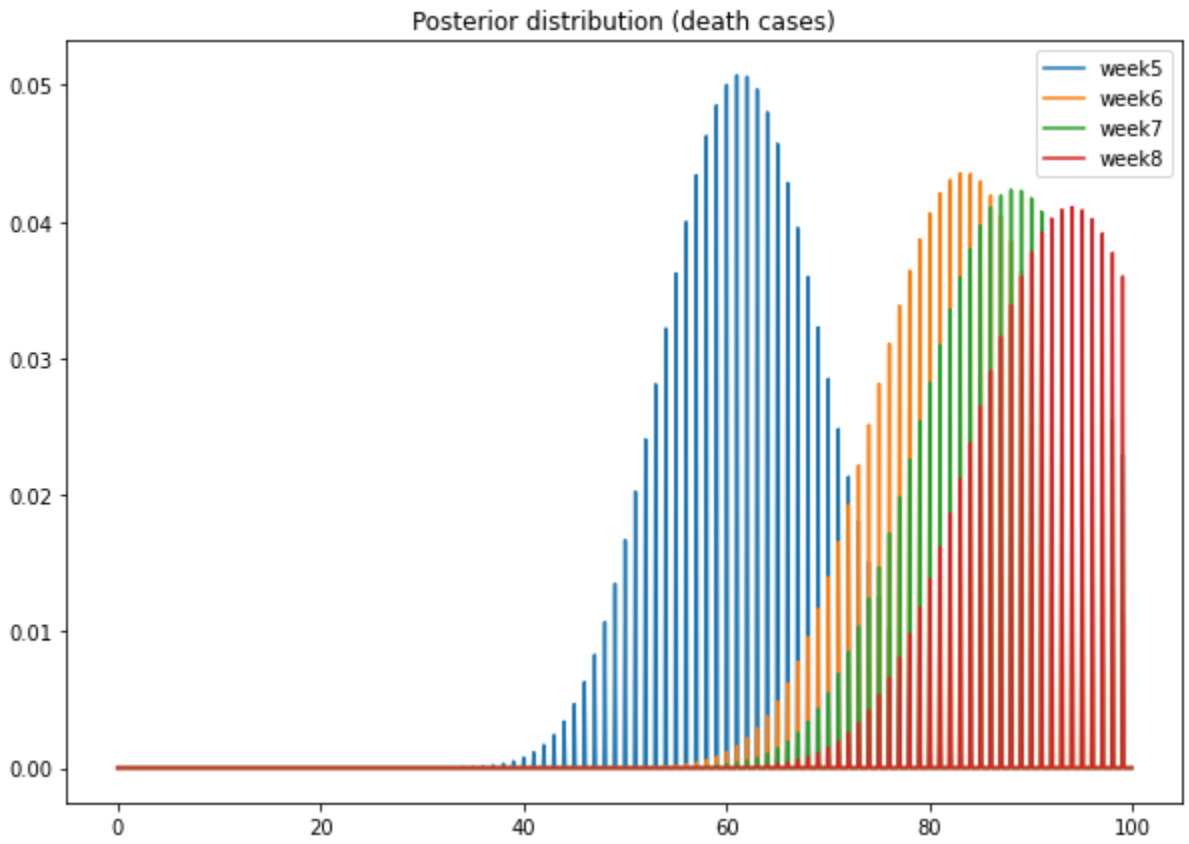
MAP for week 5: 6257.285714285715
MAP for week 6: 8223.57142857143
MAP for week 7: 8518.714285714286
MAP for week 8: 8891.82142857143

h) distribution of daily #deaths

Combine daily death from CA and CO states.
The mean and variance of CA are 71.03571428571429 and 1076.9630102040817 respectively.
First Prior is the estimated μ of Poisson, which is mean of first four weeks:
μ= 71.03571428571429

And the μ values from week 5 to week 8 are:
μ of poisson in posterior distribution (after week5) is: 61.857142857142854
μ of poisson in posterior distribution (after week6) is: 83.92857142857143
μ of poisson in posterior distribution (after week7) is: 88.85714285714286
μ of poisson in posterior distribution (after week8) is: 94.46428571428571

The figure below shows the posterior distribution of deaths from week 5 to week 8, and we can obtain the four MAP values from the figure:



MAP for week 5: 61.857142857142854
MAP for week 6: 83.92857142857143
MAP for week 7: 88.85714285714286
MAP for week 8: 94.46428571428571

3. Exploratory Task

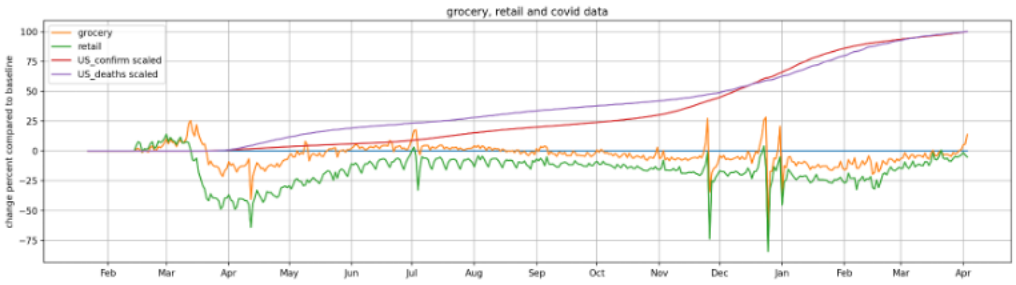
We selected dataset X to examine the change in the percentage of daily traffic from 2020 to 2021 for a number of location categories (e.g., retail and entertainment venues, grocery stores and pharmacies, parks, bus stops, workplaces, and residences) in each U.S. state. The data we choose is from the google mobility data U.S. section, collected from the google app on people's smartphones. We aim to examine the relationship between changes in human traffic at these locations over time and the number of confirmed cases and deaths per day in each state across the United States.

a. Mobility data in workplace and residents vs. Covid data

Mobility data in the workplace is inversely correlated with the data in residents and presents a weekly period repeat pattern. We can see that on Mar 20th, employees returned to the home as implementing the stay-at-home policy. And on Apr 20th and May 20th, part of the essential workers returned to the workplace as their employers got the permit, then data held level as the policy remained the same.

b. Mobility data in grocery and retail vs. Covid data

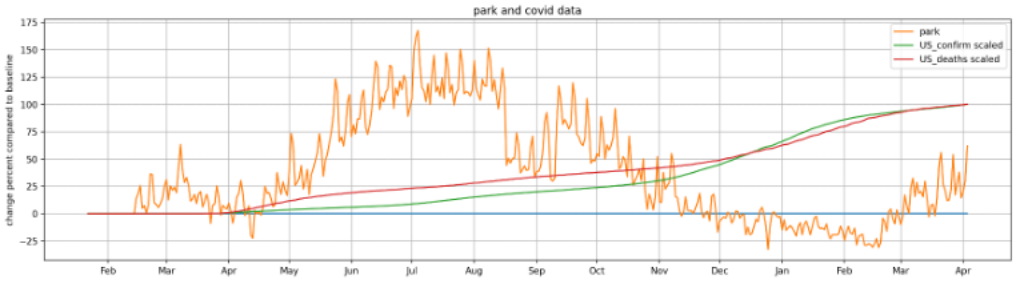
Mobility data for grocery and retail are shown in the graph below. Generally, grocery data is positively correlated with retail data, and retail data is consistently lower than grocery data, which means pandemics have a greater impact on retails than grocery. At the end of February, there is an increasing trend for grocery and retail, indicating people are hoarding goods at the beginning of the pandemic. Then, the dramatically decreased data on grocery and retails showing people have enough goods in their home and restricted their grocery and retail activities after the mandatory stay-at-home order issuing. Later on, people avoided unnecessary retail store visits, but grocery visits are around normal (0% change). Unsurprisingly, grocery and retail are affected by holidays. For instance, there are peaks around the days like Independence Day(Jul 4th), Thanksgiving(Nov 11th), Christmas(Dec 25th), New Year(Jan 1st, 2021). Interestingly, as the confirmed case number and death number accelerate their increasing speed after the Thanksgiving holiday, people keep a lower grocery activity and much lower retail activity. This trend is eased after vaccine acceptance increased around Feb 26th, after the Pfizer vaccine announcing a 98.8% effectiveness against death.



c. Mobility data in the park vs. Covid data

The graph below is the park data, which shows a steady increase during Apr, May, Jun, Jul 2020, doubling visits compared to the normal period. And then a big number of parks announced closure and quoted Covid19 concerns. We see a deep drop in mid Aug20, but it's still 50% more. Following Fall 2020, the visits decreased back to normal on Nov20, showing fewer visits on Dec20, Jan21, Feb21. Then as temperature resumed in Spring21, the visits inflated.

To evaluate whether park closure changes the daily confirmed number, We take 45 days' data before park closure and 45 day's data after park closure. Since we have enough sample size, we use the two-tail Z-test to compare two samples with ± 1.96 boundaries corresponding to its 95% confidence interval. However, the Z statistic,1.39, is smaller than 1.96, and we accept H0. We don’t have enough power to prove that the Covid data is significantly different before and after the park closure police. We can say that park closure police didn't help reduce daily confirmed cases as expected.



d. Mobility data overall vs. Covid data

To verify whether lowering public activity and reducing travel can affect the number of infections and the number of covid-19 deaths. We hope to verify the effectiveness of the stay-home policy for the prevention and control of the pandemic. For the number of the confirmed cases per day, the first thing we did was to calculate Pearson's correlation coefficients between covid-19 cases per day identified in each state and the percentage change in traffic per day in each public

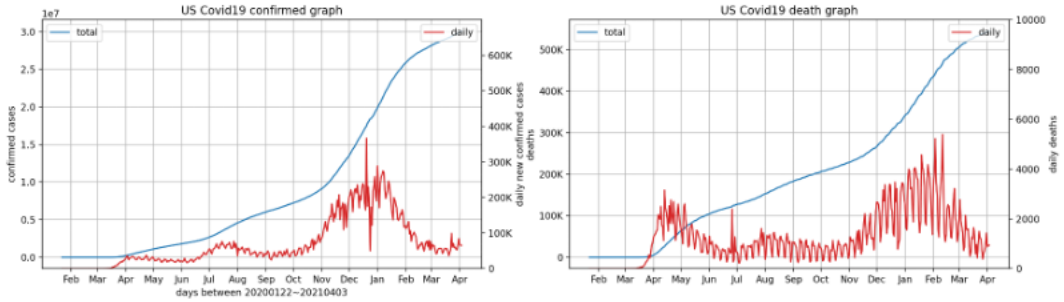
place. We found that most of these Pearson correlation coefficients are very small, close to 0. Some of them are even negatively correlated, which is against basic common sense. For instance, Pearson's correlation coefficient between transit stations' percent change from baseline and confirmed covid-19 cases number in AZ state is -0.034631. For the death cases number per day, we also calculated the Pearson correlation coefficient between death cases per day and the percentage change in traffic per day in each public place. The result is similar to the result of confirmed cases number per day. Based on this, we can conclude that there is no linear relationship between them.

Above non-significant results may be due to the small effect size. So, we separately select the 100 days with the highest percentage of activity at these locations based on baseline changes for each state and refer to it as the highest 100 days. Then, we find the number of Covid-19 confirmed cases per day and the number of Covid-19 death cases per day in these states corresponding to the highest 100 days. We denote these samples as sample x1 and sample x2, respectively. Similarly, we separately select the 100 days with the lowest percentage of activity at these locations based on baseline changes for each state and refer to them as the lowest 100 days. Then, we find the number of Covid-19 confirmed cases per day and Covid-19 death cases per day in these states corresponding to the lowest 100 days. We denote these samples as sample y1 and sample y2, respectively. Next, we do two paired sample T tests for x1 and y1, x2 and y2, respectively. We use park activity percentage change from baseline data as an example. For confirmed covid-19 cases per day, all states except HI state reject H0. For covid-19 death cases per day, all states except DC and NJ state reject H0. The tests show both the number of Covid confirmed cases per day and Covid death cases per day are significantly lower in the lowest 100 days compared to the highest 100 days. Therefore, we can say that human public activities do affect the number of covid-19 confirmed cases and deaths cases

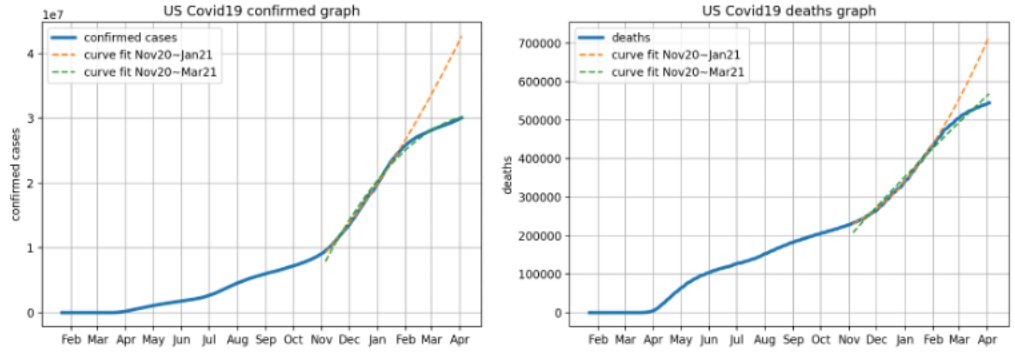
However, considering the 14-day incubation period for the Covid-19 virus, we also compare covid-19 data after 14 days of the traffic data dates. For the 100 days with the highest traffic, we select their corresponding daily confirmed diagnosis data and daily death data after 14 days. I denote these two datasets as sample x3 and sample x4. For the 100 days with the lowest traffic, we select their corresponding daily confirmed covid-19 data and daily death data after 14 days. I denote these two datasets as sample y3 and sample y4. Similarly, we do two paired sample T-tests for x3 and y3, x4 and y4, respectively. The result for daily confirmed cases remains the same while the results for daily death change to all the states rejecting the H0. Although simply applying 14 days delay could not wholly adjust for the incubation variance and death variance, the overall concordant trend of our analyses implies that social distance is critical in pandemic prevention.

e. Cumulative Covid data trend analysis

Below is the graph of the national dataset. We summed confirmed cases and deaths from 51 regions (50 states and Washing D.C.) to get the national data. The data is plotted in the following picture.



The Covid confirmed cases graph shows two phrases separated by two peaks. The bigger peak is around Dec 25th and the increase rate is consistently high from Nov 20th to late Jan 15th. The Covid death graph shows three phrases with three peaks. And the biggest peak around Feb 1st (Dec 1st – Feb 10th) corresponds to the confirmed case graph peak around Dec 25th (Nov 20th – Jan 15th). These can be easily understood that near Nov 20th, American enter the shopping season and large family gatherings. People gather around to celebrate Thanksgiving, Christmas and New Year, which sadly helped the virus to spread. The death data lagged the confirmed case data because of the clinical progression time. Near Jan 21st, the virus spread slowed down as the vaccine started to roll out and the end of the holiday. We have some easy ways to determine the trend has changed near Jan 21st. First, intuitively, if we fit a polyline based on data from Nov 20th – Jan 21st, we have a convex curve (infection rate is increasing). However, if we fit a polyline based on data from Nov 20th – Mar 21st, we have a concave curve (infection rate is decreasing). The second method is by using KS-test, we calculate the max difference between the trending line and data point. If the difference under the error margin (0.05 in this case), the data fits the trending line well, we accept H0. Otherwise, the data doesn't fit the trending line, therefore the trend has changed, we reject H0. Based on this method, we calculated the trending line (polyfit line) using data from Nov 20th – Jan 21st. KS-test statistic on Nov 20th – Jan 21st data is 0.0334. It's below the 0.05 error margin. But the KS-test statistic on Nov 20th – Feb 21st data is 0.1647. It's well above the 0.05 error margin. Thus we can more confidently say the trend has changed near mid Jan 21st. We apply the same method to US_deaths number, the conclusion is the same. After adding Feb 21st data, the max difference changed from 0.0164 to 0.0691.



In conclusion, we believe that changes in human traffic at these locations over time do have a strong influence on the number of confirmed cases and deaths per day in each state across the United States.