Group Member: Zhe Zhou 112878243

Yanming Zhang 113275452

Weihao Yin 111682764

CSE 544 Project

Mandatory Tasks

1. Data Cleaning

- Clean Dataset
    - Cases Dataset:

      For cases dataset, there is no NA value. “dropna( )” function will not have effect in this case. And we have deleted the values that new case and new death that are smaller than zero.
    - Vaccinations Dataset

      For vaccinations dataset, no NA values in vaccinations dataset. “dropna( )” function will have no effect.
- Outlier Detection

  We apply the Tuckey’s rule to detect outliers and remove them.
    - Cases Dataset:

      Original number of samples of ME is 845, after removing the outliers the number of samples of ME is 706;

      Original number of samples of KY is 845, after removing the outliers the number

of samples of KY is 780.

Hence, after applying the Tukey's rule some outliers were being removed.

- Vaccinations Dataset

  Original number of samples of ME is 519, after removing the outliers the number of samples of ME is 519;

  Original number of samples of KY is 519, after removing the outliers the number of samples of KY is 519.

  Hence, there is no outliers in vaccination dataset.

2. Inference

a. Check whether the mean of COVID19 deaths and #cases are different for Feb'21 and March'21 in the two states.

1) Use one-sample tests for Wald's, Z-test, and t-test by computing the sample mean of daily values from Feb'21 and using that as a guess for mean of daily values for March'21

For Maine,

For Wald's test,

p-value for #cases is: 2.7367411481139397e-42

p-value for deaths is: 1.2996996017585634e-34

For Z-test,

p-value for #cases is: 0.02278449748858519

p-value for deaths is: 7.77068417273477e-06

For t-test,

p-value for #cases is: 0.03008236592589231

p-value for deaths is: 0.00010313145304921799

Conclusion: The p-values from three one-sample tests are all smaller than 0.05, so we're able to reject the null hypothesis. Therefore, we can conclude that there is significant difference in #cases and deaths between Feb'21 and Mar'21 in state ME.

For Kentucky,

For Wald's test,

p-value for #cases is: 0

p-value for deaths is: 1.7126041810925073e-149

For Z-test,

p-value for #cases is: 2.3241918476171816e-92

p-value for deaths is: 2.2869615940908e-92

For t-test,

p-value for #cases is: 3.9626245802498845e-19

p-value for deaths is: 3.958318685750323e-19

Conclusion: The p-values from three one-sample tests are all smaller than 0.05, so we're able to reject the null hypothesis. Therefore, we can conclude that there is significant difference in #cases and deaths between Feb'21 and Mar'21 in state KY.

2) Repeat with the two-sample version of Wald's and two-sample unpaired t-test

For Maine,

For Wald's test,

p-value for #cases is: 8.150897407215005e-23

p-value for deaths is: 2.415376728041971e-09

For t-test,

p-value for #cases is: 0.0554286170650609

p-value for deaths is: 0.06287889888015301

Conclusion: In Wald's test, p-values for both #cases and deaths are very small (both are smaller than 0.05), so we're able to reject the null hypothesis for them. However, in t-test, p-values for them are larger than we obtained in Wald's test. Both of them are greater than 0.05, so we cannot reject the null hypothesis. Therefore, we can conclude that there is significant difference in #cases and deaths between Feb'21 and Mar'21 in state ME.

For Kentucky,

For Wald's test,

p-value for #cases is: 0

p-value for deaths is: 1.2307580015011096e-42

For t-test,

p-value for #cases is: 5.683681248038381e-10

p-value for deaths is: 2.085950070933702e-14

Conclusion: In Wald's test, p-values for both #cases and deaths are very small (both are smaller than 0.05), so we're able to reject the null hypothesis for them. However, in t-test, p-values for them are larger than we obtained in Wald's test, but the results remain the same. For both of them, p-values are smaller than 0.05, so we

can also reject the null hypothesis. Therefore, we can conclude that there is significant difference in #cases and deaths between Feb'21 and Mar'21 in state KY.

b. Inference the equality of distributions between the two states (distribution of daily #cases and daily #deaths) for the last three months of 2021 (Oct, Nov, Dec) of your dataset using K-S test and Permutation test.

1) 1-sample test, try Poisson, Geometric, and Binomial.

For Maine,

For Poisson,

The K-S test statistics for #cases is: 0.9891304347826086

The K-S test statistics for deaths is: 0.6222816414424707

For Geometric,

The K-S test statistics for #cases is: 0.49352761072665247

The K-S test statistics for deaths is: 0.28319427662454166

For Binomial,

The values of MME is: n = -2.3164326322989957 and p = -830.5358886579016

For Kentucky,

For Poisson,

The K-S test statistics for #cases is: 0.9205299896393511

The K-S test statistics for deaths is: 0.9362244016772192

For Geometric,

The K-S test statistics for #cases is: 0.6100887263327393

The K-S test statistics for deaths is: 0.8121693700070038

For Binomial,

The value of MME is: n = -4.490745239131433 and p = -150.85856694809493

Conclusion: The threshold is 0.05 for K-S test. All K-S test statistics we obtained are greater than 0.05, so we're able to reject the null hypothesis. Therefore, the distributions of #cases and deaths between two states are different. For binomial distribution, the MME we obtained is negative, it implies that the data doesn't follow the binomial distribution.

2) 2-sample test

The 2-sample K-S test statistics for #cases is: 0.6739130434782608

The 2-sample K-S test statistics for deaths is: 0.5108695652173914

Conclusion: The threshold is 0.05 for K-S test. All K-S test statistics we obtained are greater than 0.05, so we're able to reject the null hypothesis. Therefore, the distributions of #cases and deaths between two states are different.

3) permutation test

The p-value of the permutation for #cases is: 0.0

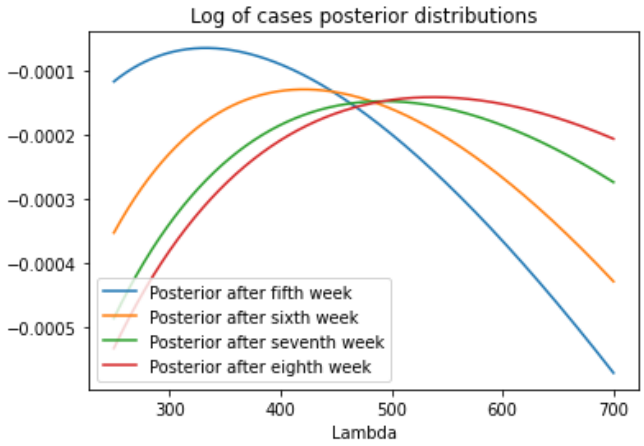
The p-value of the permutation for deaths is: 0.0

Conclusion: The threshold is 0.05 for permutation test. The p-values we obtained are smaller than 0.05 (actually 0), so we're able to reject the null hypothesis. Therefore, the distributions of #cases and deaths between two states are different.

c. Bayesian inference on daily stats

1) Result of cases data:

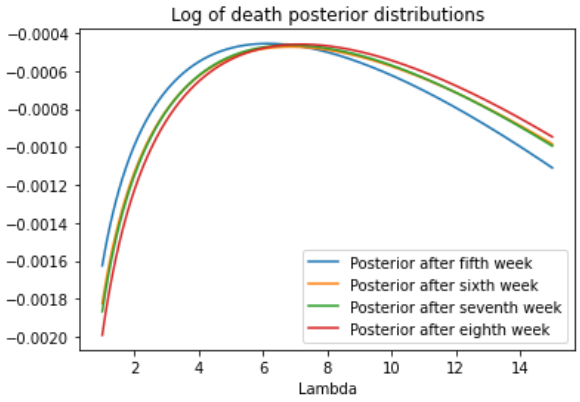
To prevent underflow, we take logarithm in calculation for this graph.



MAP of fifth week posterior: 332.70
MAP of sixth week posterior: 421.10
MAP of seventh week posterior: 492.90
MAP of eighth week posterior: 537.30

2) Result of death data:

To prevent underflow, we take logarithm in calculation for this graph.



MAP of fifth week posterior: 6.15
MAP of sixth week posterior: 6.85
MAP of seventh week posterior: 6.85
MAP of eighth week posterior: 7.17

d. Predict #vaccines administered for the fourth week in May 2021 using data from the first three weeks of May 2021.

1) AR(3)

MAPE for Maine prediction: 289.67%
MAPE for Kentucky prediction: 65.84%
MSE for Maine prediction: 66057162.20
MSE for Kentucky prediction: 87481232.80

2) AR(5)

MAPE for Maine prediction: 61.33%
MAPE for Kentucky prediction: 60.87%
MSE for Maine prediction: 5410097.04
MSE for Kentucky prediction: 68522710.74

3) EWMA with alpha = 0.5

MAPE for Maine prediction: 267.95%
MAPE for Kentucky prediction: 58.94%
MSE for Maine prediction: 33018646.79
MSE for Kentucky prediction: 58781695.62

4) EWMA with alpha = 0.8

MAPE for Maine prediction: 220.44%
MAPE for Kentucky prediction: 55.18%
MSE for Maine prediction: 34148660.11
MSE for Kentucky prediction: 58942450.69

e. Use the paired T-test to determine the equality of means of the #vaccines administered between the two states for the months of September 2021 and November 2021.

   The p-value of the paired T-test for the #vaccines administered between the two states is: 2.2622217870322283e-07

   Conclusion: Since the p-value is very close to 0, it's smaller than 0.05. So, we're able to reject the null hypothesis, and we can conclude that there is a significant difference between #vaccines administered in two states.

Exploratory Tasks

The Data we used in this part is about the price of bitcoin, we think there is an association between the bitcoin's price and the pandemic. Hence, we propose three new inferences about this dataset.

Inference 1

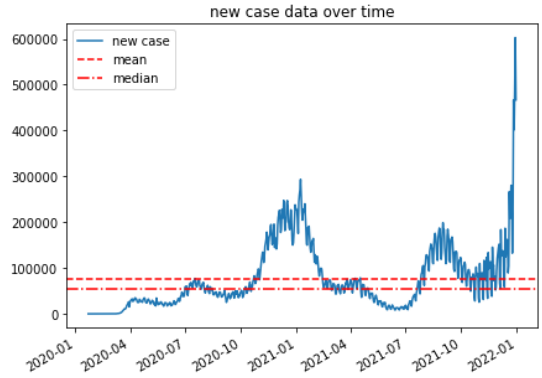
a. <u>Check if new case/death and bitcoin price are linear correlated</u>

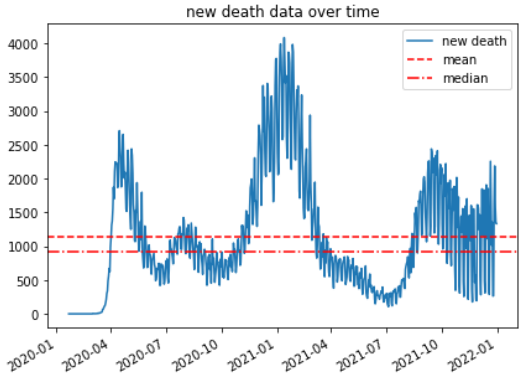
   Common beliefs suggest that COVID19 may have impact on the bitcoin price but not the other way around. So, to investigate the impact of cases data towards bitcoin dataset, the Pearson correlation test can check for possible linear relationship between to Dataset.

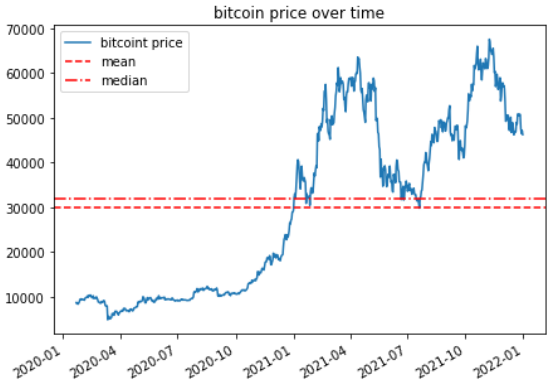
1) H0: new case/death and bitcoin price are NOT linearly correlated;

   H1: new case/death and bitcoin price are linearly correlated

2) Compute S(x,y)







The correlation between new case and bitcoin Price is 0.3045951530637549, the correlation between new death and bitcoin Price is 0.12849715144651486

Conclusion: Since the absolute value of both correlations are smaller than 0.5. We accept the original assumption. Therefore, new case/death and bitcoin price are NOT linearly correlated

b. <u>Use Chi-square test to check the independence between cases dataset and bitcoin price</u>

Use the mean to divide the case data into low number of new cases high number of new cases. Perform the same method on bitcoin price column and the price data is divided into high price and low price.

H0: new case and bitcoin price are independent;

H1: new case and bitcoin price are NOT independent

degree of freedom is 1

Q_obs value is:65.95674900752013

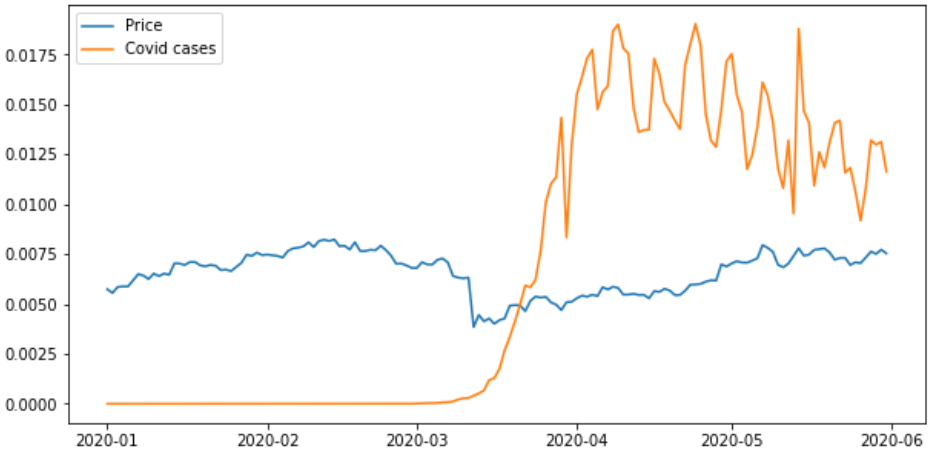
After table lookup the Chi square value for df=1, alpha=0.05 is 3.841

Conclusion: Since the Q_obs is smaller than Chi square value, we accept H0. i.e. the bitcoin price and new cases are independent

Inference 2

a. <u>Use Permutation test to check if covid cases are different after price falling</u>

We can see from the graph below, there is sudden falling of price:



We want to examine that whether the cases before March 12 is different from after March 12.

H0: Cases(before 03-12) equal Cases(after 03-12);

H1: Cases(before 03-12) not equal Cases(after 03-12)

P-value for the permutations test is: 0.00

Conclusion: Since the p-value is smaller than 0.05, we reject null hypothesis. Therefore, we can conclude that the cases before March 12 is different from after March 12.

Comment on inferences 1 & 2

Above, we used Chi-square test, Z-test, and Pearson correlation test. Among these three, the Z test is the best tool for investigating the impact of COVID.

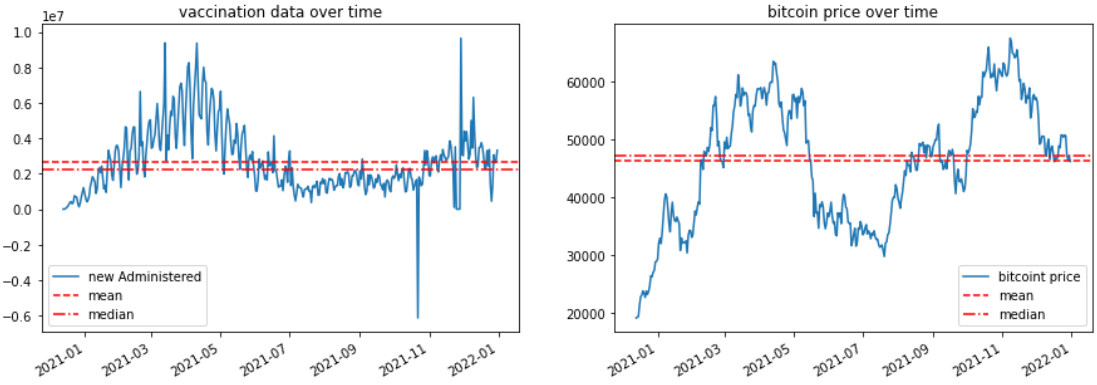
I. The Pearson correlation test only concerns the linear dependence between two datasets. And possible non-linear dependency might not be captured.

II. The Chi-square test divides the bitcoin price into two sets, so for the new cases number/ new deaths number. However, this clear cutoff highly depends on the cutoff value. And it might not suit every data distribution (multilevel etc.).

III. Z-test only divides new cases number/ new deaths number into two sets, and it can capture non-linear dependencies. To sum up, Z-test is the best inference tool overall. And we conclude that the COVID has an impact on the bitcoin price.

Inference 3

a. Check if vaccination number reached a certain level would have effect on bitcoin price



By observing the plot of vaccination data over time, we can see that after July 2021, the new vaccination number tends to a constant value. This might be a sign for vaccination number reaches a certain level. Hence, we want to check if this has an impact on bitcoin price.

1. H0: vaccination level has NO impact on bitcoin price; H1: vaccination level has impact on bitcoin price
2. Perform Z-test

Z is:-4.99057598377566, absolute value of Z is 4.99057598377566, p-value is:6.019951377389532e-07

since |Z|>1.96, reject H0. The vaccination level has an impact on bitcoin.