CSE 544 - Probability and Statistics for Data Science

Assignment 6:

Names	SBU-ID	% of effort
Manish Reddy Vadala	114190006	25
Ranjith Reddy Bommidi	114241300	25
Venkata Ravi Teja Takkella	113219890	25
Angira Katyayan	113168055	25

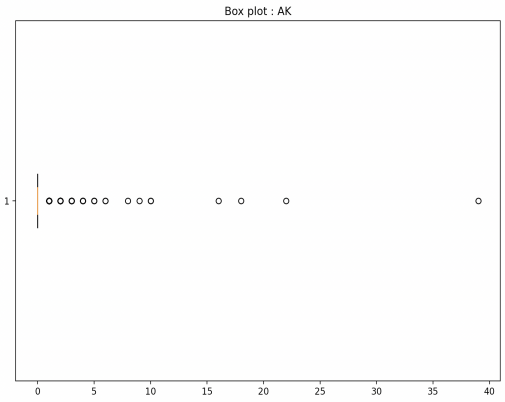
Mandatory Tasks :

The states assigned to our team are Alaska and Alabama. The mandatory tasks that need to be completed will be mentioned below.

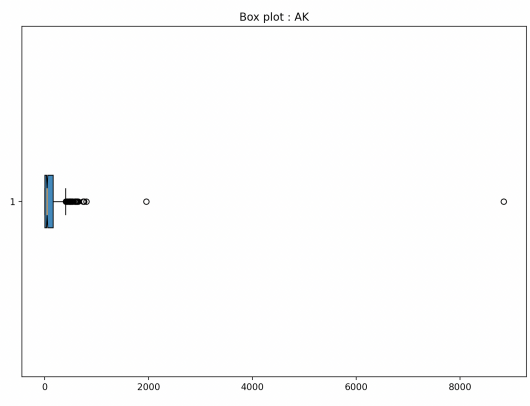
Task 1 : Data Pre-Processing :

The given dataset consists of an incremental sum of “Confirmed cases” and “Deaths” in the two mentioned two states, Alaska(AK) and Alabama(AL). This data ranges from 2020-01-22 to 2021-04-03. We are asked to work upon daily increases in deaths and confirmed cases in our mandatory tasks, the dataset needs to be converted to daily increase in cases and deaths. The steps followed are :

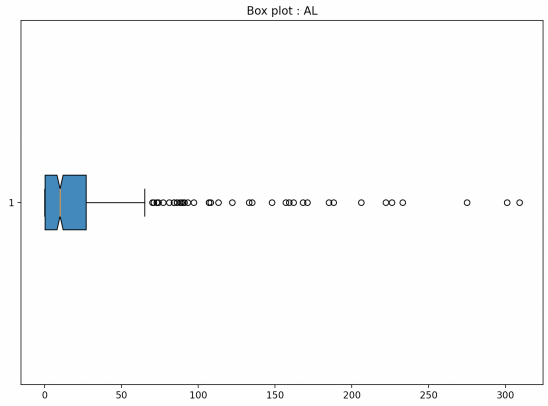
a) Changed the initial given dataset from cumulative increase to daily increase of confirmed cases and deaths.
b) There are some exceptions in the data, where the cumulative sum of deaths is decreased at a few points which is not a valid scenario (Ex: The total number of deaths till tomorrow are less than what we have today). To solve this, we have made the daily cases or deaths as 0 depending on the column.
c) Checked for missing values and finally the sanitation is complete.
d) Separated the initial given csv to separate state csvs for better access.
e) Used Tukey’s rule to detect the outliers for both the states.
1) The algorithm which I followed : We are detecting a datapoint as an outlier if both the values of confirmed cases and deaths are shown as outliers among their sets.
2) This is to maintain the data integrity as a whole; one column depicted as an outlier is not a correct way to remove the entire datapoint.
3) On top of that, one of our states, Alaska, has very few deaths in total and the whole death stats have been wiped out except 0s. So our algorithm above helps a lot to mark them as one if they are outliers in both the confirmed cases and deaths.
4) We used the default parameters for Tukey’s rule as discussed in our class.
5) The number of outliers found for **Alaska** are : **18**
The box plot for daily **deaths** look like :



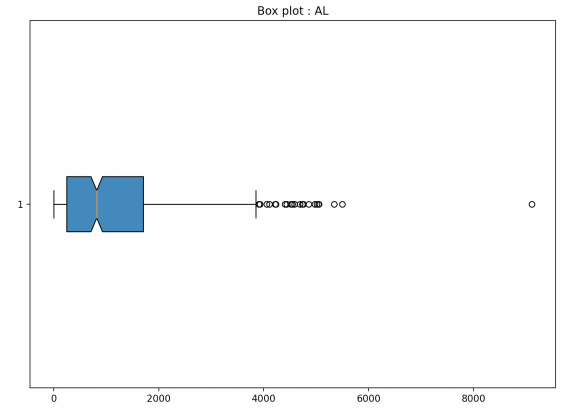
The box plot for daily **cases** look like :



6) The number of outliers found for **Alabama** are : **10**
   The box plot for daily **deaths** look like :



The box plot for daily **cases** look like :



f) Our entire Preprocessing is done at this point. We have removed the outliers and placed them in separate files namely “AK.csv” and “AL.csv”. We still have maintained the original state data in “AK_original.csv” and “AL_original.csv” if required for any other mandatory parts below.

Task 2a : Auto-Regression,EWMA:

In this task we have to take data for the first 3 weeks of August and predict the number of fatalities and cases for the 4th week of August.We need to find this for both task Alaska(AK) and Alabama(AL). Our preprocessed data has two files for both states that have the number of cases and fatalities of COVID.

$$Y = [\ y_1,\ y_2,\ y_3,\ y_4, \ldots\ldots\ldots\ldots y_{19},\ y_{20},\ y_{21}\ ]$$

i) **AR(3)**

For this time series data we need to fit an AR(3) model.

$$y_{t+1/t} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 y_{t-2}$$

This is a multiple linear Regression case.
So our train points will be:
Train set = [
{(y4),(y1,y2,y3)},
{(y5),(y2,y3,y4)},
{(y6),(y3,y4,y5)},
…
{(y20),(y17,y18,y19)},
{(y21),(y18,y19,y20)}
]

For t data points we will have t-p training data for AR(p)

We obtain weights $\{\beta_0, \beta_1, \beta_2, \beta_3\}$ by solving

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Where X = [
[1,y1,y2,y3],
[1,y2,y3,y4],
[1,y3,y4,y5],
…..
[1,y18,y19,y20],
]
Y = [y4,y5,y6…….y20,y21]

For state Alaska(AK):

Alaska(AK)	Cases	Deaths
MSE	2239.63	2.52
MAPE	178.36	1125.67

For state Alabama(AL):

Alabama(AL)	Cases	Deaths
MSE	1220112.37	158.5
MAPE	319.68	95.92

i) **AR(5)**

For this time series data we need to fit an AR(3) model.

$$y_{t+1/t} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 y_{t-2} + \beta_4 y_{t-3} + \beta_5 y_{t-4}$$

This is a multiple linear Regression case.
So our train points will be:
Train set = [
{(y6),(y1,y2,y3,y4,y5)},
{(y7),(y2,y3,y4,y5,y6)},
{(y8),(y3,y4,y5,y6,y7)},
…
{(y20),(y15,y16,y17,y18,y19)},
{(y21),(y16,y17,y18,y19,y20)}
]

For t data points we will have t-p training data for AR(p)

We obtain weights $\{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ by solving

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

For state Alaska(AK):

Alaska(AL)	Cases	Deaths
MSE	749.96	2.49
MAPE	34.85	476.2

For state Alabama(AL):

Alabama(AL)	Cases	Deaths
MSE	865348.38	338.16
MAPE	139.64	289.63

EWMA:

Here we are give time series data

$$Y = [y_t, y_{t-1}, y_{t-2}, \ldots\ldots\ldots\ldots y_1]$$

We need to predict for y_{t+1} which is given by:

$$\hat{y}_{t+1/t} = \alpha(y_t + (1 - \alpha)y_{t-1} + (1 - \alpha)^2 y_{t-2} + \ldots\ldots\ldots)$$

iii) **EWMA(0.5)**

$$\hat{y}_{t+1/t} = \alpha(y_t + (1 - \alpha)y_{t-1} + (1 - \alpha)^2 y_{t-2} + \ldots\ldots\ldots)$$

Given, $\alpha = 0.5$

$$\hat{y}_{22/21} = 0.5(y_{21} + (0.5)y_{20} + (0.25y_{19} + \ldots\ldots\ldots)$$

We compute for y_{22} and use them for predicting next point that is for day 23/Aug

$$\hat{y}_{23/22} = 0.5(y_{22} + (0.5)y_{21} + (0.5)^2 y_{20} + \ldots\ldots\ldots)$$

Similarly, we compute for next days using previous obtained values

$$\hat{y}_{24/23} = 0.5(y_{23} + (0.5)y_{22} + (0.5)^2 y_{21} + \ldots\ldots\ldots)$$

$$\hat{y}_{25/24} = 0.5(y_{24} + (0.5)y_{23} + (0.5)^2 y_{22} + \ldots\ldots\ldots)$$

$$\hat{y}_{26/25} = 0.5(y_{25} + (0.5)y_{24} + (0.5)^2 y_{23} + \ldots\ldots\ldots)$$

$$\hat{y}_{27/26} = 0.5(y_{26} + (0.5)y_{25} + (0.5)^2 y_{24} + \ldots\ldots\ldots)$$

$$\hat{y}_{28/27} = 0.5(y_{27} + (0.5)y_{26} + (0.5)^2 y_{25} + \ldots\ldots\ldots)$$

For state Alaska(AL):

Alaska(AL)	Cases	Deaths
MSE	632.45	2.05
MAPE	28.02	221.8

For state Alabama(AL):

Alabama(AL)	Cases	Deaths
MSE	676149.48	116.02
MAPE	96.28	51.9

iv) **EWMA(0.8)**

$$\hat{y}_{t+1/t} = \alpha(y_t + (1 - \alpha)y_{t-1} + (1 - \alpha)^2 y_{t-2} + \ldots\ldots\ldots)$$

Given, $\alpha = 0.8$

$$\hat{y}_{22/21} = 0.8(y_{21} + (0.2)y_{20} + (0.2)^2 y_{19} + \ldots\ldots\ldots)$$

We compute for y_{22} and use them for predicting next point that is for day 23/Aug

$$\hat{y}_{23/22} = 0.8(y_{22} + (0.2)y_{21} + (0.2)^2 y_{20} + \ldots\ldots\ldots)$$

Similarly, we compute for next days using previous calculated values

$$\hat{y}_{24/23} = 0.8(y_{23} + (0.2)y_{22} + (0.2)^2 y_{21} + \ldots\ldots\ldots)$$

$$\hat{y}_{25/24} = 0.8(y_{24} + (0.2)y_{23} + (0.2)^2 y_{22} + \ldots\ldots\ldots)$$

$$\hat{y}_{26/25} = 0.8(y_{25} + (0.2)y_{24} + (0.2)^2 y_{23} + \ldots\ldots\ldots)$$

$$\hat{y}_{27/26} = 0.8(y_{26} + (0.2)y_{25} + (0.2)^2 y_{24} + \ldots\ldots\ldots)$$

$$\hat{y}_{28/27} = 0.8(y_{27} + (0.2)y_{26} + (0.2)^2 y_{25} + \ldots\ldots\ldots)$$

Obtained results for the 4th week data is show in table below:

For state Alaska(AL):

Alaska(AL)	Cases	Deaths
MSE	632.89	1.81
MAPE	28.26	130.92

Obtained results for the 4th week data is show in table below:

For state Alabama(AL):

Alabama(AL)	Cases	Deaths
MSE	990160.63	117.95
MAPE	156.97	51.75

Task 2b : Walds, Z and T tests :

In this step, we were asked to compare the mean of monthly COVID19 stats have changed between Feb 2021 and Mar 2021. We used the original data for both these months without removing any outliers as it is mentioned in the question that Feb should have 28 days and March should have 31 days.

Walds One Sample Test:

We are told to use MLE as the estimator and also given the estimator purposes that the daily data is Poisson distributed. For this we initially calculated the Feb month MLE estimate of a poisson distribution and used that as the ground truth for the March data.

Null hypothesis: Sample mean of values of Feb'21 and Mar'21 are same.

Ground Truth: MLE Estimate of the given poisson distribution of Feb'21 data.

Alternative hypothesis: Sample mean of values of Feb'21 and Mar'21 are not same.

Applicability: Since we are talking about the mean estimates, using CLT we can say that mean estimate follows Normal distribution as the number

of observations increase which are around 31 for Mar'21 right now. And also as discussed in class we observations > 30 can be kind of considered as the case where CLT starts to work. So we can say that Wald's Test is applicable over here.

Parameter: We are asked to take α= 0.05. So $Z_{\alpha/2}$ = 1.95996

Using this information the results obtained are :

For State **Alaska** :
- Confirmed Cases :
  - W = 9.551290967418497.
  - So we reject the null hypothesis.
- Deaths :
  - W = 2.2121101063122923.
  - So we reject the null hypothesis.

For State **Alabama** :
- Confirmed Cases :
  - W = 101.11281155316361.
  - So we reject the null hypothesis.
- Deaths :
  - W = 74.56539047812959.
  - So we reject the null hypothesis.

Walds Two Sample Test:

We can now take individual mean estimates and do a subtraction to find the test statistics.

Null hypothesis: Sample mean of values of Feb'21 and Mar'21 are same.

Ground Truth: 0, since it is based on the Null hypothesis; both the means are same.

Alternative hypothesis: Sample mean of values of Feb'21 and Mar'21 are not same.

Applicability: Since we are talking about the mean estimates, using CLT we can say that mean estimate follows Normal distribution as the number

of observations increase which are around 31 for Mar'21 right now and 28 for Feb'21. And also as discussed in class we observations > 30 can be kind of considered as the case where CLT starts to work. So we can say that Wald's two sample tests are applicable over here.

Parameter: We are asked to take α= 0.05. So $Z_{\alpha/2}$ = 1.95996

Using this information the results obtained are :

For State **Alaska** :

Confirmed Cases :

W = 6.8470238158309344.

So we reject the null hypothesis.

Deaths :

W = 1.3576637193125367.

So we accept the null hypothesis.

For State **Alabama** :

Confirmed Cases :

W = 59.799155629485476.

So we reject the null hypothesis.

Deaths :

W = 32.052331967199464.

So we reject the null hypothesis.

Z One Sample Test:

Here we are taking the ground truth as the Feb'21 data. It's also mentioned to take the sigma value as the corrected sample standard deviation of the entire COVID state dataset.

Null hypothesis: Sample mean of values of Feb'21 and Mar'21 are same.

Ground Truth: Mean estimate of Feb'21 data.

Alternative hypothesis: Sample mean of values of Feb'21 and Mar'21 are not same.

Applicability: Since we are talking about the mean estimates, using CLT we can say that mean estimate follows Normal distribution as the number of observations increase which are around 31 for Mar'21 right now and Feb'21. And also as discussed in class we observations > 30 can be kind of considered as the case where CLT starts to work. It's mentioned to take the true sigma value based on the entire dataset. So we can say that the Z Test is applicable over here with the assumption that the true sigma value calculated is accurate enough.

Parameter: We are asked to take α= 0.05. So $Z_{\alpha/2}$ = 1.95996

Using this information the results obtained are :

For State **Alaska** :

Confirmed Cases :

Z = 0.24698057876588833.

So we accept the null hypothesis.

Deaths :

Z =0.660696947292419.

So we accept the null hypothesis.

For State **Alabama** :

Confirmed Cases :

Z = 2.1226194574410107.

So we reject the null hypothesis.

Deaths :

Z = 7.770455284874159.

So we reject the null hypothesis.

T One Sample Test:

Here we are taking the ground truth as the Feb'21 data. The standard deviation used in the denominator is again the corrected variance of the sample taken which is Mar'21 in this case.

Null hypothesis: Sample mean of values of Feb'21 and Mar'21 are same.

Ground Truth: Mean estimate of Feb'21 data.

Alternative hypothesis: Sample mean of values of Feb'21 and Mar'21 are not same.

Applicability: For T-test to be applicable, the initial assumption is that the sample data should be normally distributed. But given that data is Poisson distributed when explained about Wald's test. T-test is not applicable over here.

Parameter: We are asked to take α= 0.05. So $t_{n-1,\alpha/2}$ = 2.04 where n = 31 (Mar'21 data)

Using this information the results obtained are :

For State **Alaska** :

Confirmed Cases :

T = 0.7831585710643173.

So we accept the null hypothesis.

Deaths :

T =0.9493073099766663.

So we accept the null hypothesis.

For State **Alabama** :

Confirmed Cases :

T = 3.2282830386573385.

So we reject the null hypothesis.

Deaths :

T = 17.985457625870445.

So we reject the null hypothesis.

T Two Sample Test:

Here we are taking the difference of sample means of Feb'21 and Mar'21 months with their corrected standard deviation used when required.

Null hypothesis: Sample mean of values of Feb'21 and Mar'21 are same.

Ground Truth: 0, with the assumption that ground truth is true.

Alternative hypothesis: Sample mean of values of Feb'21 and Mar'21 are not same.

Applicability: For T-test to be applicable, the initial assumption is that the sample data should be normally distributed. But given that data is Poisson distributed when explained about Wald's test. T-test is not applicable over here.

Parameter: We are asked to take α= 0.05. So $t_{n-1,\alpha/2}$ = 2.00 where n = 59 (Mar'21 data)

Using this information the results obtained are :

For State **Alaska** :

Confirmed Cases :

T = 0.5899786558902169.

So we accept the null hypothesis.

Deaths :

T =0.4687488992805854.

So we accept the null hypothesis.

For State **Alabama** :

Confirmed Cases :

T = 2.7009842718772865.

So we reject the null hypothesis.

Deaths :

T = 3.632512268176839.

So we reject the null hypothesis.

Task 2c : KS Test and Permutation Test :

The given dataset consists of an incremental sum of “Confirmed cases” and “Deaths” in the two mentioned two states, Alaska(AK) and Alabama(AL). This data ranges from 2020-10-01 to 2020-12-31. As done in the preprocessing we are taking the daily increases in deaths and confirmed cases in our mandatory tasks. The test we have performed are :

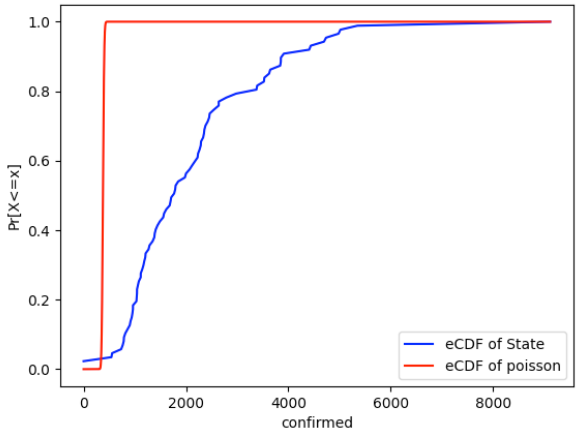
a) 1-sample KS test with Poisson: Generally we perform this test by assuming the distribution of the data as Poisson. But since we have data of two different states, we first find the parameters of our poisson distribution using state1 data and do the test on state2 data from the learned parameters. Assumptions are none.

   H_0: Both samples from state1 and state2 for confirmed cases follow the same distribution.

   i) Learned Parameters from state1 data are $\lambda = 378.38$

   ii) After finding Parameters and in the process of doing the ks test, we got the dmax value as 0.965 which is very high.

iii) We have taken the critical value as 0.05. Since $dmax > 0.05$we reject our Null hypothesis.

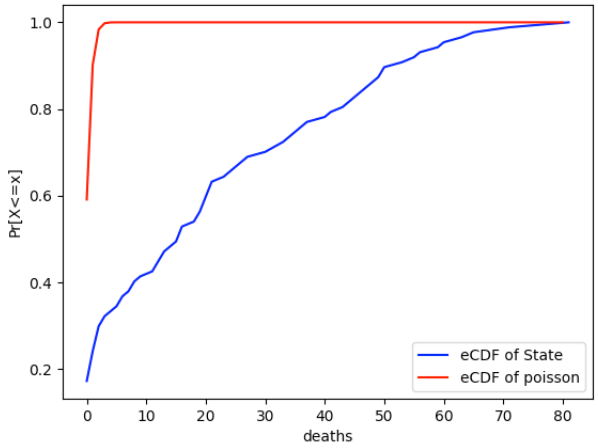


H_0: Both samples from state1 and state2 for deaths follow the same distribution.

iv) Learned Parameters from state1 data are $\lambda = 0.53$

v) After finding Parameters and in the process of doing the ks test, we got the dmax value as 0.68 which is very high.

vi) We have taken the critical value as 0.05. Since $dmax > 0.05$we reject our Null hypothesis.



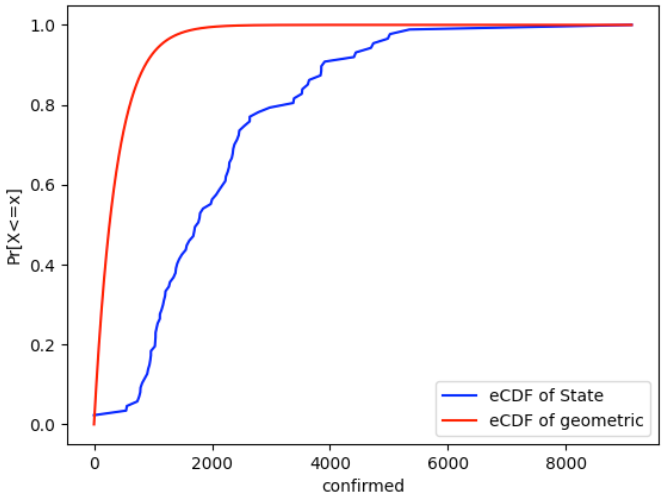
b) 1-sample KS test with Geometric: Generally we perform this test by assuming the distribution of the data as Geometric. But since we have data of two different states, we first find the parameters of our geometric distribution using state1 data and do the test on state2 data from the learned parameters. Assumptions are none.

H_0: Both samples from state1 and state2 for confirmed cases follow the same distribution.

i) Learned Parameters from state1 data are probability $p = 0.00264$

ii) After finding Parameters and in the process of doing the ks test, we got the dmax value as 0.799 which is very high.

iii) We have taken the critical value as 0.05. Since $dmax > 0.05$we reject our Null hypothesis.

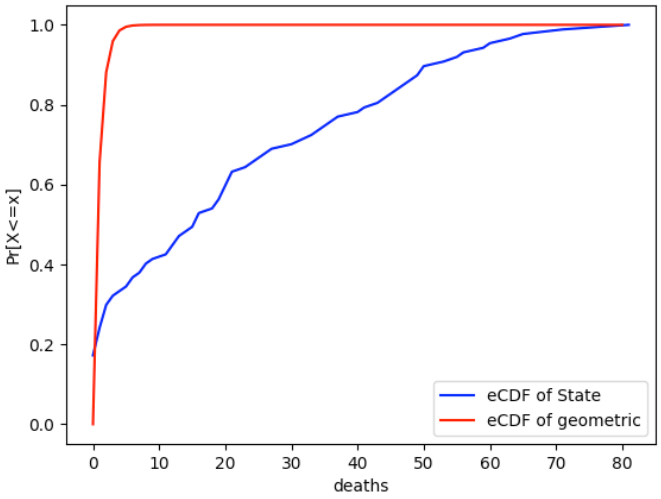


H_0: Both samples from state1 and state2 for deaths follow the same distribution.

iv) Learned Parameters from state1 data are probability $p = 0.655$

v) After finding Parameters and in the process of doing the ks test, we got the dmax values as 0.653 which is very high.

vi) We have taken the critical value as 0.05. Since $dmax > 0.05$we reject our Null hypothesis.



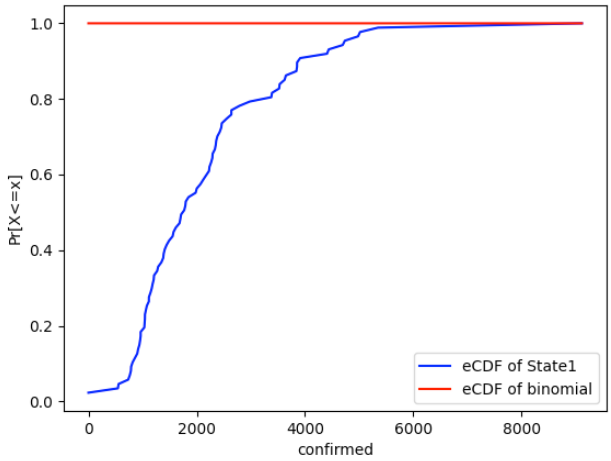
c) 1-sample KS test with Binomial: Generally we perform this test by assuming the distribution of the data as Binomial. But since we have data of two different states, we first find the parameters of our Binomial distribution using state1 data and do the test on state2 data from the learned parameters. Assumptions are none.

H_0: Both samples from state1 and state2 for confirmed cases follow the same distribution.

i) Learned Parameters from state1 data are probability $n = -0.143 \,\&\, p = -2645.62$

ii) We got negative values for Parameters which is not right. In the process of doing the ks test, we got the dmax value as 0.977.

iii) We have taken the critical value as 0.05. Since $dmax > 0.05$we reject our Null hypothesis.

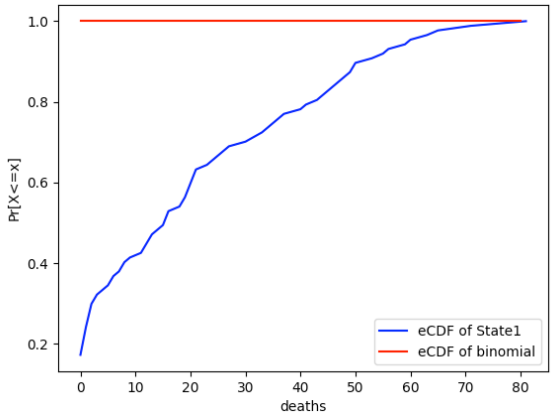


H_0: Both samples from state1 and state2 for deaths follow the same distribution.

iv) Learned Parameters from state1 data are probability $n = -0.41 \& p = -1.279$

v) We got negative values for Parameters which is not right. In the process of doing the ks test, we got the dmax values as 0.8275.

vi) We have taken the critical value as 0.05. Since $dmax > 0.05$we reject our Null hypothesis.



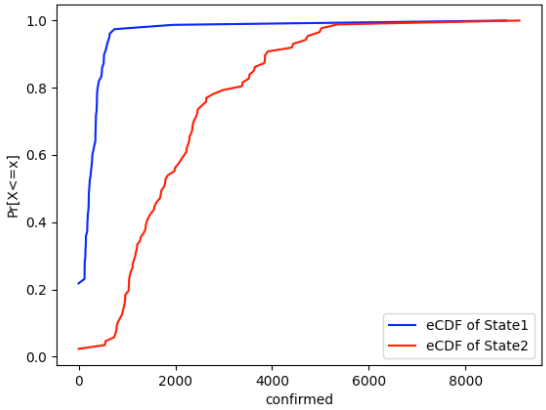
Inference: As we can see we got the negative parameters for n & p, which is not at all right. We cannot represent our data with binomial distribution at all.

d) 2-sample KS test: We have to find the cdfs of two datasets and find the dmax(maximum difference between the cdf graphs). Assumptions are none.

H_0: Both samples from state1 and state2 for confirmed cases follow the same distribution.

i) After Computing the CDFs, we got the dmax value as 0.91 which is very high.

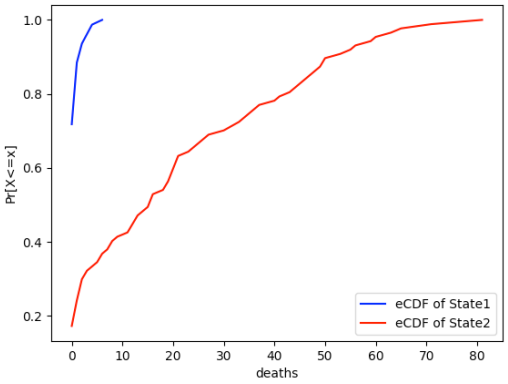
ii) We have taken the critical value as 0.05. Since $dmax > 0.05$we reject our Null hypothesis.



H_0: Both samples from state1 and state2 for deaths follow the same distribution.

iii) After Computing the CDFs, we got the dmax value as 0.65 which is very high.

iv) We have taken the critical value as 0.05. Since $dmax > 0.05$we reject our Null hypothesis.



Task 2d : Bayesian Inference :

For this task, we were asked to sum up the daily deaths and cases from both the states initially. And after that follow the procedure mentioned in the question. After calculating the posteriors and MAPs, the equations look like:

From the first four weeks of data we got :

$$\beta = \lambda_{MME} = \sum x_i/n$$

Given that the prior is Exponentially distributed.

$$prior = \lambda e^{-\lambda x}$$

And the likelihood is based on the Poisson distribution

$$likelihood = \pi \lambda^{x_i} e^{-\lambda}/x_i!$$

$$likelihood = \lambda^{\sum x_i} e^{-n\lambda} / \pi x_i!$$

The final posterior function looks proportional to:

$$posterior \sim \lambda^{\sum x_i} e^{-\lambda(n+1/\beta)}$$

If you look at this equation, this looks like a Gamma distribution with parameters:

$\alpha` = \sum x_i$ and $\beta` = n + 1/\beta$.

To get the constant, we can integrate this whole function from 0 to infinity since for a poisson distribution, the parameter should be greater than 0. And finally take the inverse of it. The constant which we obtain is,

$\beta`^{\alpha`}/\Gamma(\alpha`)$ This is nothing but the constant which we obtain in the pdf equation of a Gamma distribution.

And for further inclusions of weeks, the posterior distribution changes proportionally too. For example if we include the 6th week data as the likelihood to the already calculated posterior of 5th week's data; which will be the prior right now. The posterior distribution looks like:

$$posterior \sim \lambda^{\sum x_i + \sum x_j} e^{-\lambda(n+m+1/\beta)}$$

This still follows the gamma distribution with updated parameters :

$\alpha` = \sum x_i + \sum x_j$ and $\beta` = n + m + 1/\beta$.

So the posterior will always be a Gamma distribution when the data is Poisson distributed with the initial prior given as exponential distribution.

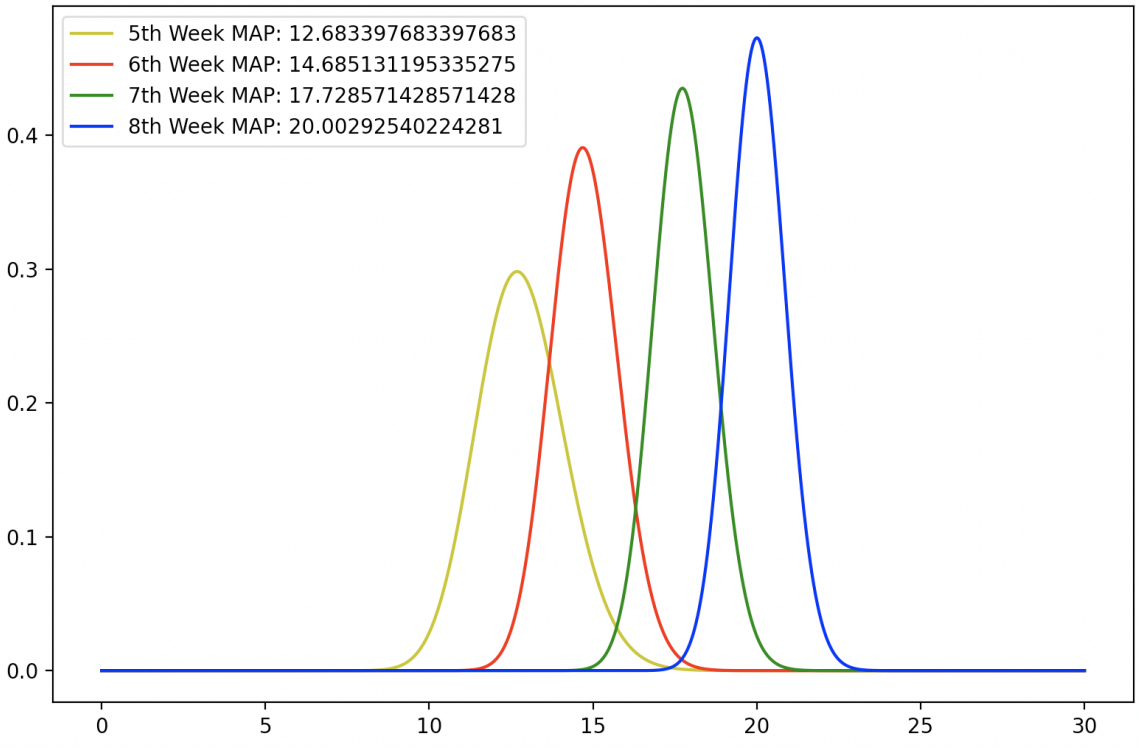
For calculating the MAP, we differentiated the equation above w.r.t λ and equate it to 0. So the MAP after including the 5th week's data is:

$$\lambda_{MAP} = \sum x_i/(n + 1/\beta)$$

And the MAP after including 6th week's data is :

$$\lambda_{MAP} = \sum x_i + \sum x_j/(n + m + 1/\beta)$$

So using all these equations, we plotted the posteriors along with the MAP values onto a chart. You can see the MAP values also mentioned on the plot below for each corresponding week inclusion.



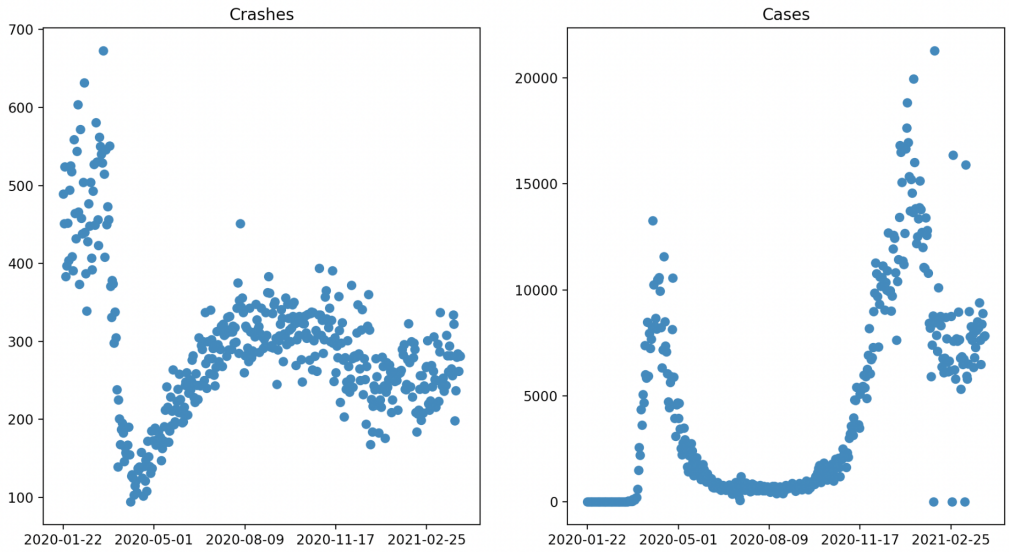
If you see, the plot is turning to be narrower as the new number of samples have been added, which is as expected for a posterior.

Using X Dataset :

Task X0 : Preprocessing:

We did a preprocessing on the entire X dataset initially. The initial dataset contains all the occurrences of the crashes. So we need to seperate the each date and make a map out of the number of occurrences in the dataset. Once done with this we merged the NY state data from the US-All dataset to the corresponding dates that we got from the X dataset.

a) **Initial Analysis (Task_X5)** :
After plotting the values onto a scatter plot, we wanted to show the general trends and show how the change in data of one plot affects the other.



If we check the trend over here, we can say as the number of cases tend to increase, the number of crashes are decreasing and the same works with the other cases where the number of cases decrease. So based on the plots, we can clearly say that there's a proper negative correlation between the number of crashes and number of cases per day.

Task X1 : Chi Square Test for Independence :

The X dataset we have has two columns Crashes and Cases the crashes are from the X dataset and the cases are from US dataset. We would like to know the dependency between these two datasets, in the time frame of before and after **March 13th 2020**. Why March 13th?

After checking the news, we found out that there's a **state emergency** announced on March 13th in New York state. And so, we want to see if that announcement has any change in the number of crashes recorded.

The test goes as follows :

H_0: Both samples from X dataset and US dataset for confirmed cases are independent before March 13th.

- After performing the chi-square test, we got the chi squared statistic value as 1556.10 .
- And even with the degrees of freedom = 29. From the table lookup p value is almost 0.
- Since p < 0.05, we reject our null hypothesis saying that they aren't independent.

H_0: Both samples from X dataset and US dataset for confirmed cases are independent after March 13th.

- After performing the chi-square test, we got the chi squared statistic value as 30928.89.

- And even with the degrees of freedom = 30. From the table lookup p value is almost 0.
- Since p < 0.05, we reject our null hypothesis saying that they aren't independent.

Inference : Given two datasets having a very different range of numerical values, we mostly cannot show whether they are dependent or independent. But based on the hypothesis above, we can obtain the information on how dependent the datasets are instead

So we can see that the statistical value is pretty high after March 13th compared to previous. This clearly explains that the rise in confirmed cases in NY really raised an alarm among the people resulting in their reduced movement. This also shows that the crashes are pretty highly dependent on the number of cases increment over here. Whereas before March 13th, the level of COVID issue might not be far spread among the people and hence the movement is still normal as before.

Task X2 : Pearson's Correlation Test:

We would like to know the Correlation between these two datasets Crashes and Cases, in the time frame of **2020/05/01 to 2020/06/31** using Pearson's Correlation Test.

Why this period again ?

This is the time in New York state, where the number of cases have increased a lot. So we wanted to see if there's really any correlation between the number of crashes recorded and confirmed cases increasing.

The test goes as follows :

H_0: Both samples from X dataset and US dataset for confirmed cases are negatively correlated in the given time frame.

- After performing the pearson correlation test, we got the S_{xy} value as -0.66.
- And since the value is < -0.5 we can say that the samples are negatively correlated.
- Hence we accept our null hypothesis.

Inference:

So we can clearly see that, there's a proper negative correlation between the number of crashes recorded and number of cases per day. This shows that people started to think and really believed that COVID is a real issue and this is the time period where they wanted to stay safe and stayed in.

Now let's check for the time period **2021/01/01 - 2021/01/31**. We wanted to check whether the negative correlation exists throughout the year or people got accustomed to COVID and started moving out which resulted in more crashes.

H_0: Both samples from X dataset and US dataset for confirmed cases are correlated in the time frame of 2021/01/01 - 2021/01/31.

- After performing the pearson correlation test, we got the S_{xy} value as 0.064.
- Even if our value is near to -0.5 but since the value is > -0.5 we say that the samples are not negatively correlated.
- Hence we reject our null hypothesis.

Inference :

Based on this test, we can see that the datasets are not correlated. We can clearly conclude now that people have gotten used to COVID, and hence the people started moving out a bit more compared to before even when the cases are increasing and this is explained by rejecting the null hypothesis based on the statistics above.

Task X3 : Linear Regression:

In this task we are predicting the number of motor vehicle crashes using the number of covid cases from the last 7 days. So, for this task we have taken data from day **13/March/2020 to 13/April/2020**.

Why do you think this will work ?

Since the state emergency on Mar 13th, people have stopped moving a lot and hence the crashes decreased too. But they slowly got accustomed to it and started going out for their daily tasks. We wanted to see if this accustomization follows any linear dependency and wanted to use Linear Regression for the same.

We took the input dataset as the number of cases for the last three days and the label is the number of crashes seen today.

So our data will be

$$Y = [\ y_1,\ y_2,\ y_3,\ y_4, \ldots\ldots\ldots\ldots y_{n-1,}\ y_n\ ]$$
$$C = [\ c_1,\ c_2,\ c_3,\ c_4, \ldots\ldots\ldots\ldots c_{n-1,}\ c_n\ ]$$

y_1 cases on 13/March/2020
y_i cases on i'th day
y_n cases as on 13/Apr/2020

c_i= crashes on ith day

Training set = [
{(c4),(y1,y2,y3)},
{(c5),(y2,y3,y4)},
{(c6),(y3,y4,y5)},
…
{(c_n),($y_{n-1}, y_{n-2}, y_{n-3}$)}
]

Here are we are predicting $c_i$motor vehicle crashes on th i+1 day using formula:

$$c_{i+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 y_{t-2}$$

Hence solving this multiple linear Regression Problem by finding weights(i.e $\hat{\beta}$) using:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

So using this we have calculated for the next 7 days and calculated MSE and MAPE
MSE = 1280.687
MAPE = 24.839%

Inference :

Using the MAPE value which is around 24%, we can't exactly infer that it's an excellent predictor. But we can obtain some loose insights from this regression. Based on the internet, a MAPE value < 20% is considered to be

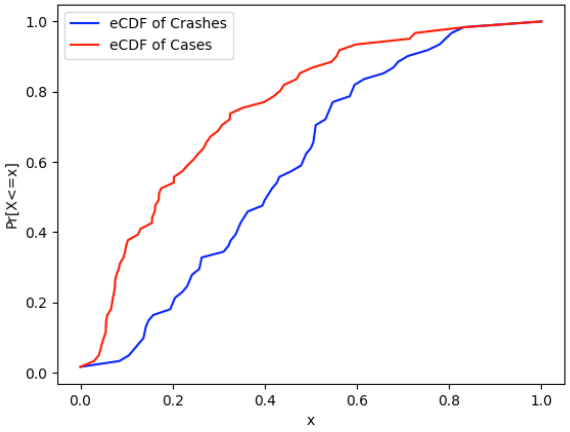
good. We are pretty close, and can infer saying that two datasets are a bit linearly dependent on each other after March 13th 2020.

Task X4 : KS Test:

We would like to see if these both datasets follow the same distribution, in the same frame of 2020/05/01 to 2020/06/31 that is used in the Pearson's correlation first hypothesis. For this we normalize the dataset values so that we can get a better understanding of the plots.

H_0: Both samples from X dataset and US dataset for confirmed cases follow the same distribution in the time period of 2020/05/01 to 2020/06/31.

- While performing the ks test, after Computing the CDFs, we got the dmax value as 0.36.
- We have taken the critical value as 0.05. Since $dmax > 0.05$, We reject our Null hypothesis.



Inference :

Based on the above hypothesis, we can say that the distributions are different. But if you see, there's a clear gap between the curves of the ecdfs. This brings out the inference where the gap is the time taken by the people to acknowledge the rise of cases. It does take a few days for people to realise the actual depth of the issue before practically applying it. So, we can bring out this inference using this hypothesis.