Stage III (Distributions and Hypothesis Testing)

Priyanka Budavi

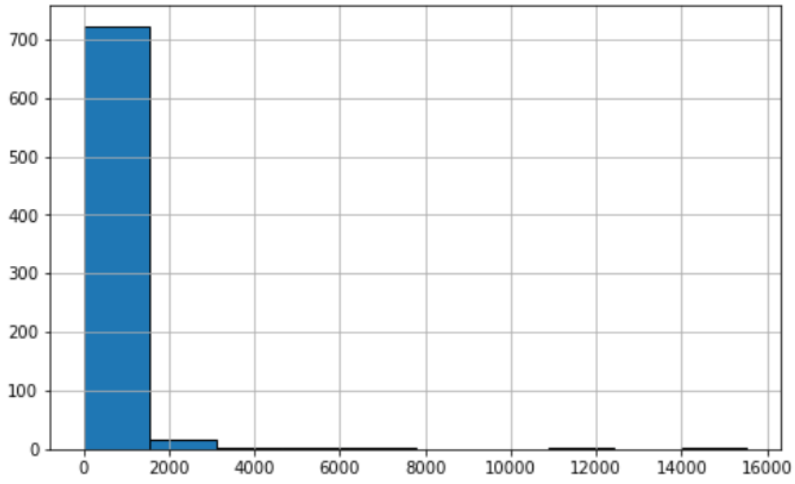
Task1: Use the state data to fit distribution to the number of COVID-19 new cases.

- For this task, I used the data generated from stage 1 and normalized the cases and deaths column for Alabama state. Later, compared moments of distribution with the other states.
- The statistics of Alabama state:

Mean	Variance	Kurtosis	Skewness
340.1542	814610.5234935433	152.4980006931335	10.725900604575097

Inference:

- The mean of the Alabama state for new cases per day is 340.1542787860785 which means per day minimum 340 cases were rising.
- The skewness is positive hence the tail of the distribution is longer towards the right-hand side of the curve.
- The kurtosis is positive and tells distribution is peaked and possesses thick tails.

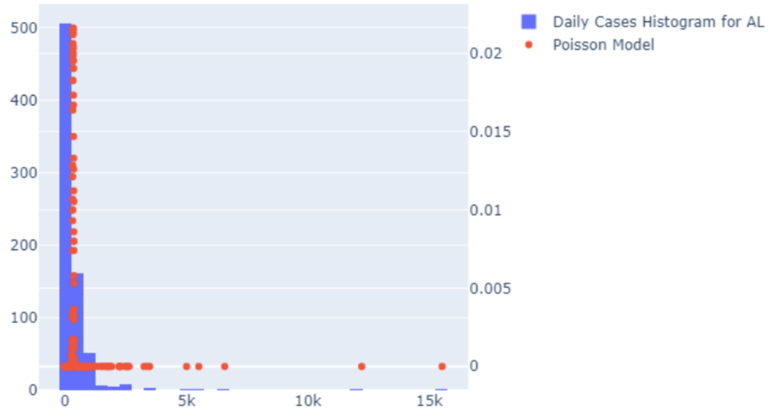


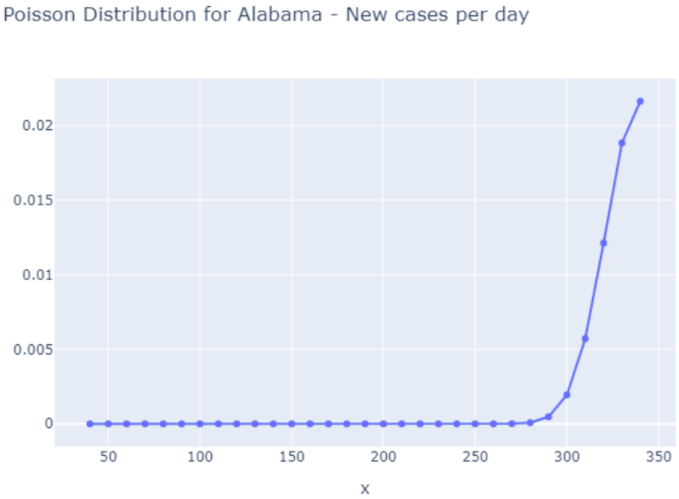
From the above plot we observe:

1. The graphs show the count of cases per day in the state of Alabama
2. The data is skewed at the right which means it’s a positive skew.
3. The data has a peak initially and then falls.
4. Thus, the data is discrete

Since the data is discrete, poission distribution is the best fit for this data.

The below is the Poisson model for Alabama state:





Inference

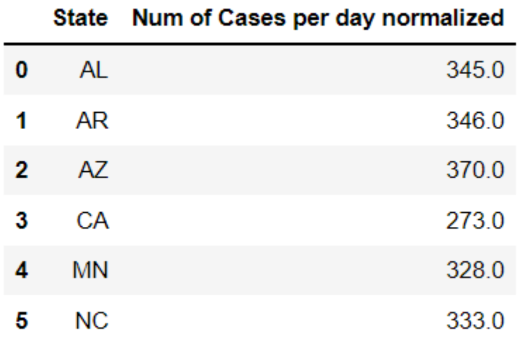
- I set the range from 40 to 350 and these are my interval where am calculating the probability mass distribution.
- Using the mean obtained from the above dataset I found the mean and plotted a graph with x axis being my interval and y axis being probability.
- From the graph we observe that initially the cases were low hence the probability in that Alabama state is zero but after few intervals the cases rise and hence the probability of a person contracting covid is high.

Compare the distribution and its statistics to 5 other states of your choosing. Describe if the distributions look different and what does that imply.

- For this task, I filtered the states that I am interested in and calculated the moments of distribution in those states. Before that, I normalized the cases and deaths column for the dataset.

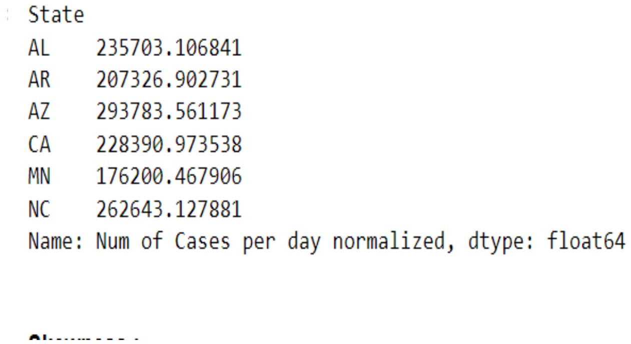
- Below are Moments of distribution for all the states

Mean of all states:



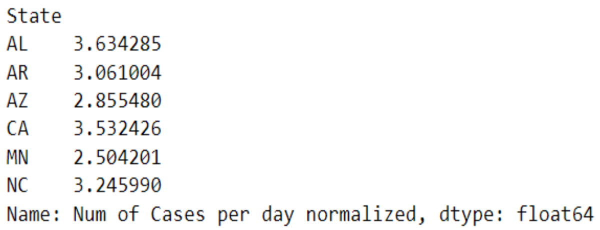
	State	Num of Cases per day normalized
0	AL	345.0
1	AR	346.0
2	AZ	370.0
3	CA	273.0
4	MN	328.0
5	NC	333.0

Variance of all states



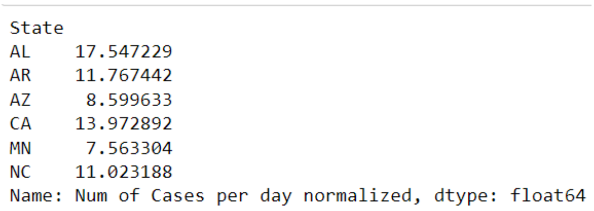
```
State
AL    235703.106841
AR    207326.902731
AZ    293783.561173
CA    228390.973538
MN    176200.467906
NC    262643.127881
Name: Num of Cases per day normalized, dtype: float64
```

Skewness of all states



```
State
AL    3.634285
AR    3.061004
AZ    2.855480
CA    3.532426
MN    2.504201
NC    3.245990
Name: Num of Cases per day normalized, dtype: float64
```

Kurtosis of all states



```
State
AL    17.547229
AR    11.767442
AZ     8.599633
CA    13.972892
MN     7.563304
NC    11.023188
Name: Num of Cases per day normalized, dtype: float64
```

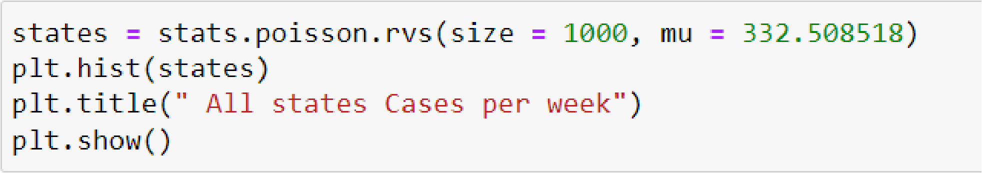
Inference:

- The above table shows the statistics of the distribution of the data of all the states.
- In kurtosis, Alabama is highly peaked, and Minnesota is low.
- The data is positively skewed for all the states and the skewness is observed towards right which is same for all states.

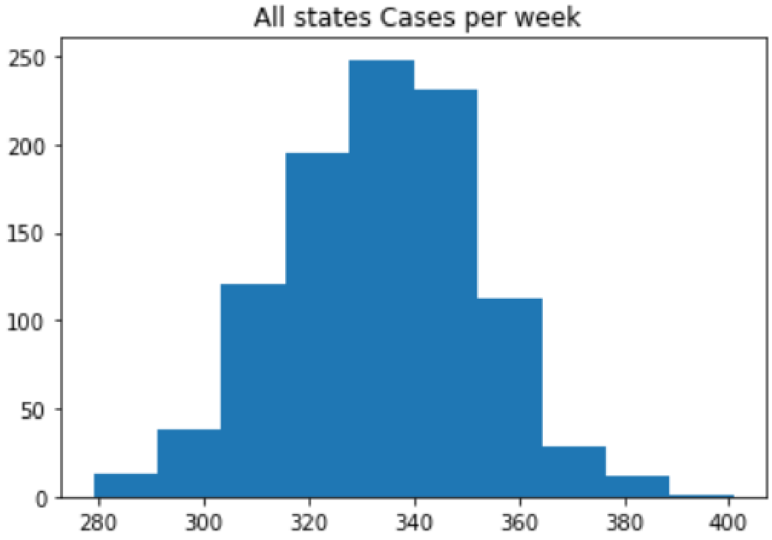
Task 2: Model Distribution for new cases and deaths

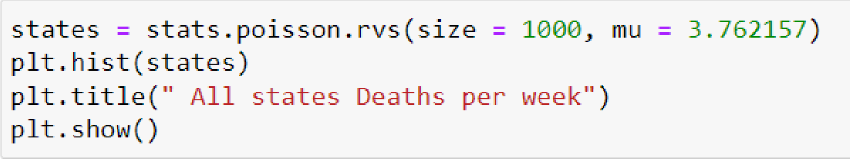
- In this task I used the normalized cases and deaths to plot the distribution with x axis being the random interval against the probability. I used a combined dataset to show the cases and deaths distribution for all the states in a single graph.

- The below plot is the plot of all new cases and deaths with the mean of the columns.

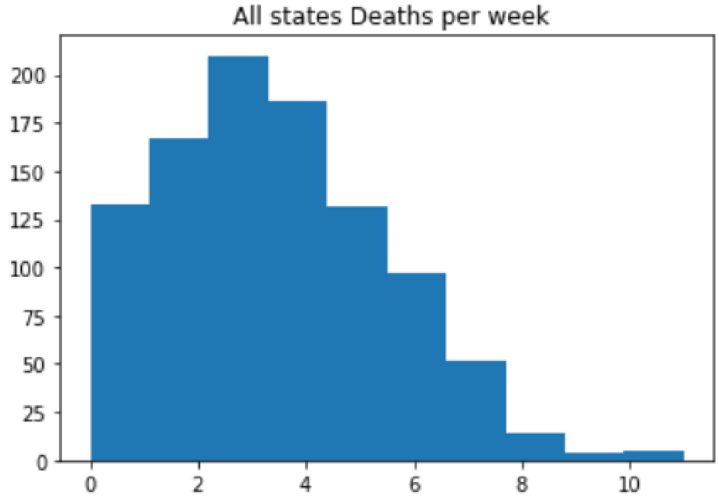


```
states = stats.poisson.rvs(size = 1000, mu = 332.508518)
plt.hist(states)
plt.title(" All states Cases per week")
plt.show()
```

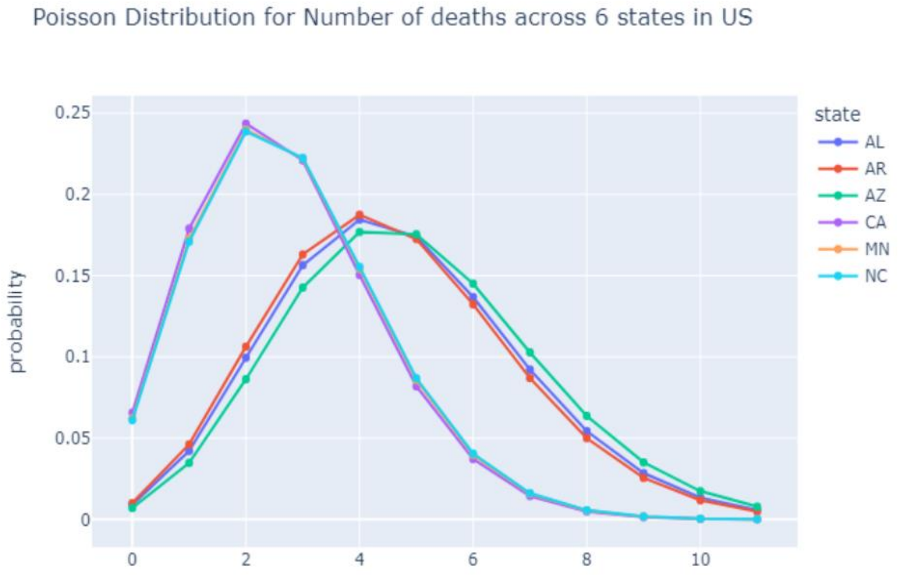




```
states = stats.poisson.rvs(size = 1000, mu = 3.762157)
plt.hist(states)
plt.title(" All states Deaths per week")
plt.show()
```



Poission Distribution for the deaths:



Inference:

- We observe that all the states were having same deaths rates during a given interval
- The whole dataset we observe the probability of death rate in all states is in the range 0.20 - 0.25. This is a weekly dataset and hence the death rates seem to decrease at the start and end of a particular interval.
- The reason for the decrease could be reduced number of cases. Thus, we compare that as the number of cases rose, the death rates also increased and vice versa.

Task 3: Perform correlation between Enrichment data variables and COVID-19 cases to observe any patterns.

- For this task I used the previously generated covid and hospital dataset merged file.
- The merged dataset, the columns of new cases and new deaths were normalized by a scale of 1000000.

Hypothesis between Enrichment data and number of cases to be compared against states.

- We see that as the normalized cases increase, the death cases also increased. This is because the two variables have a positive correlation and thus if one variable increases the other increases as well.

- With cases and adult_icu_bed_covid_utilization variables the correlation is positive so as the cases increased the beds used also increased

- With cases and critical_staffing_shortage_today_yes, the relation is negative. This means that with increase in cases, there might be a possibility that the staff member are also affected in the hospital and hence there is shortage of staff.

And same goes with the Deaths. Thus, I conclude that we are trying to show how the variables are correlated with each other and how its dependent or effects the other variable.