# IAC 621 Project

## Stage III (Distributions and Hypothesis Testing)

Reetika Sarkar

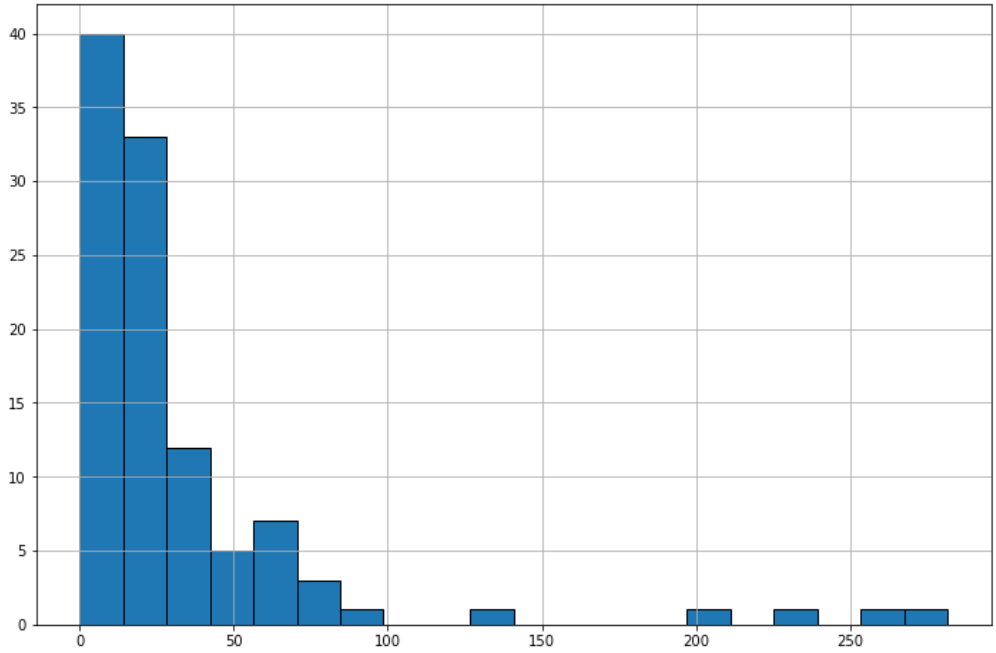
**Task 1: Use state data to fit a distribution to COVID-19 new cases.**

- The distribution statistics of new COVID-19 cases in North Carolina (NC) are:

| Minimum | Maximum | Mean | Standard deviation | Skewness | Kurtosis |
|---|---|---|---|---|---|
| **0** | 281.68 | 31.72 | 47.75 | 3.52 | 13.81 |

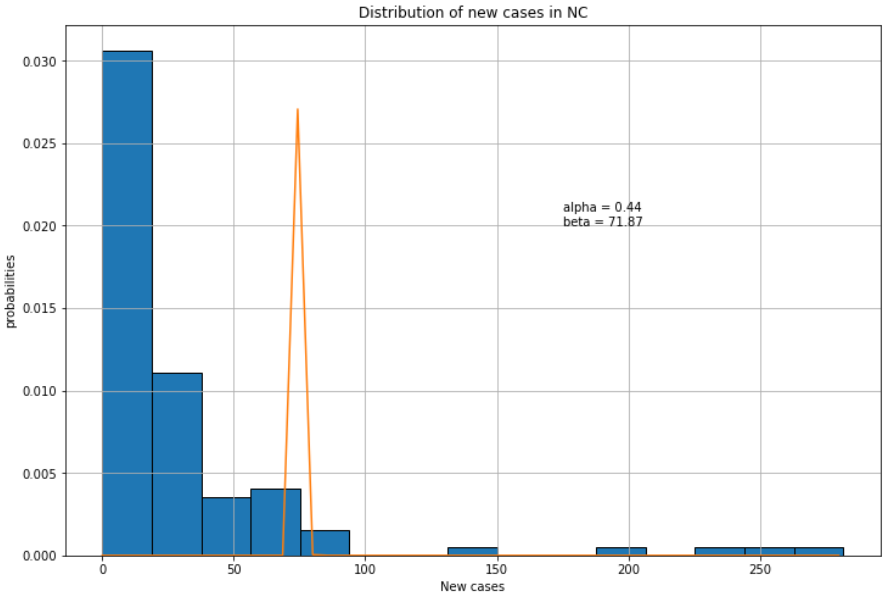
  The data is skewed to the right and has a tall peak (compared to a standard normal distribution). The mean number of new cases in the state are about 32.

- The graph of the new cases is shown below.




  The histogram is skewed to the right, the data points are all positive. It resembles that of a Gamma distribution or Poisson distribution (although the given data is discrete indicating Gamma might not be appropriate). The x-axis represents the counts of new COVID-19 cases while the y-axis represents the frequencies. The distribution appears to be **unimodal** with maximum number of new case counts at 0.

- The distribution of new cases was fitted to a Gamma distribution (using `gamma.pdf()` in `scipy.stats.distributions`) with the parameter estimates computed using Method of Moments. The graph is shown below.

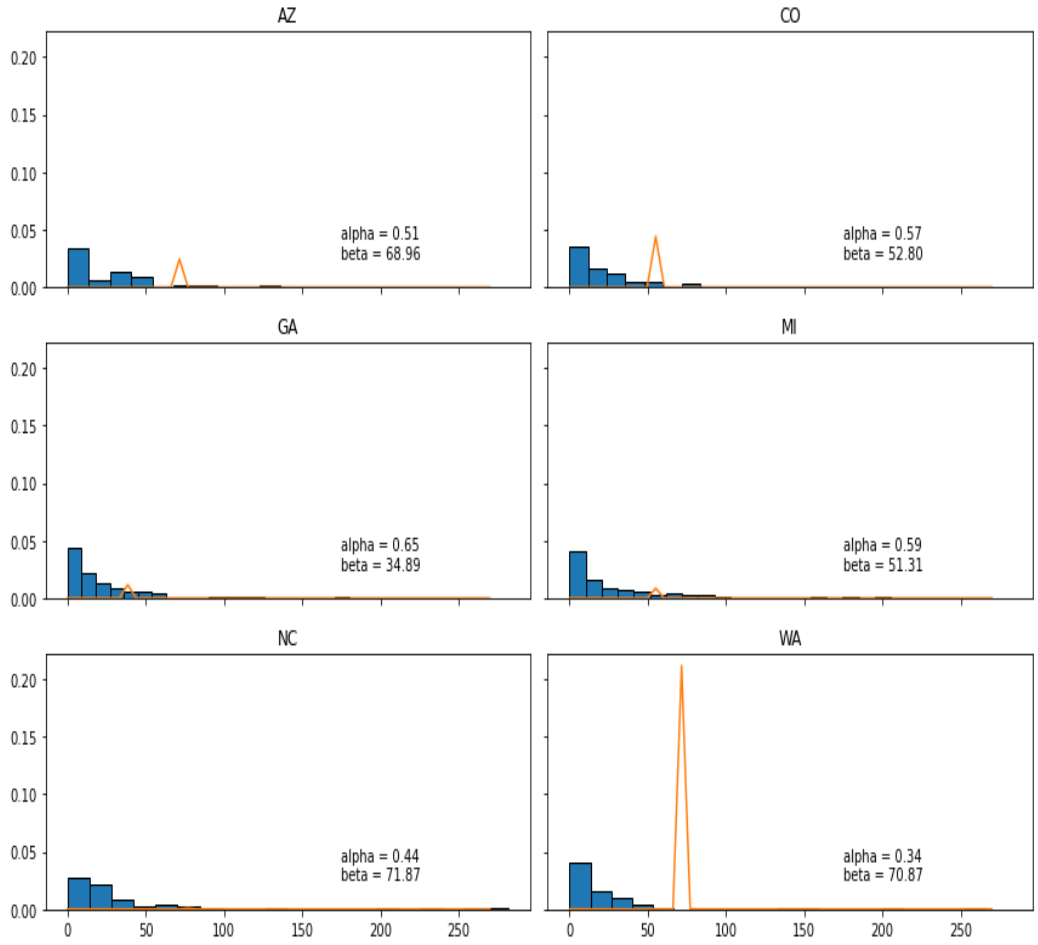



- The plot for the distribution of new cases in NC was made over the range of minimum and maximum counts. The fitted Gamma distribution with parameter estimates obtained using Method of Moments as $\alpha$=0.44 and $\beta$=71.87 **does not** fit the data well. It could be due to the data being discrete while the theoretical distribution (Gamma) is continuous in nature. The peak for Gamma distribution is much to the right of the data values.

- Next, comparisons in the distributions of new cases and deaths were performed between NC and five other states namely, Arizona (AZ), Colorado (CO), Georgia (GA), Michigan (MI) and Washington (WA).

  The fitting was done using Gamma distribution with parameters for each state estimated using Method of Moments. The results are shown below:

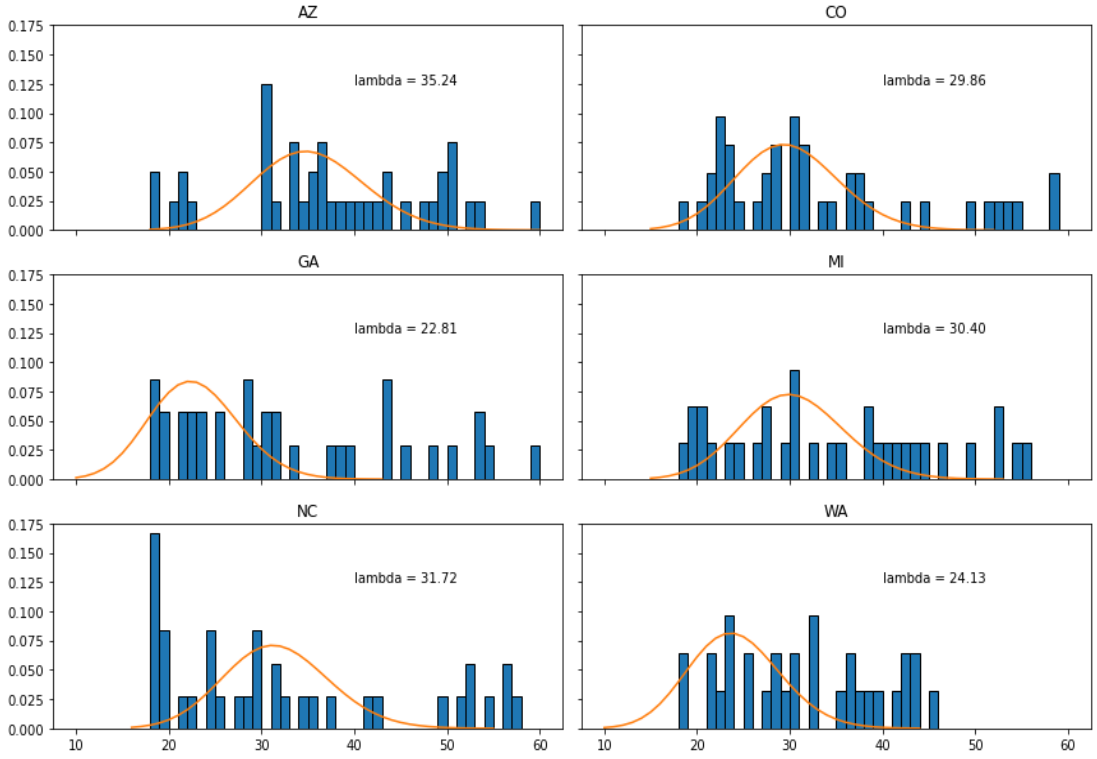
| State | Minimum | Maximum | Mean | Standard deviation | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| NC | 0 | 281.68 | 31.72 | 47.75 | 3.52 | 13.81 |
| AZ | 0 | 272.34 | 35.24 | 49.30 | 2.93 | 9.74 |
| CO | 0 | 238.23 | 29.86 | 39.71 | 3.11 | 11.70 |
| GA | 0 | 180.15 | 22.81 | 28.21 | 2.83 | 10.55 |
| MI | 0 | 205.42 | 30.40 | 39.49 | 2.53 | 7.59 |
| WA | 0 | 266.20 | 24.13 | 41.35 | 4.06 | 17.90 |




- Some observations from the plots and table above:
    - Maximum number of new cases recorded in AZ, CO and WA are comparable to that of NC. GA and MI have lower maximum values.
    - Barring Arizona, all other states have mean number of new cases lower than NC.
    - The standard deviation of new cases is higher in NC than that in CO, GA, MI and WA while AZ has a higher standard deviation than NC.
    - All the state datasets are skewed to the right with NC and WA having higher level of skewness than others.
    - NC and WA also have the highest levels of kurtosis.
    - From the plots above it appears that Gamma distribution is not appropriate to fit the counts of new COVID-19 cases. All states have peaks much further away from the actual peak in the data.
    - These estimates could be improved somewhat by using Maximum Likelihood estimates but they are not likely to show a significant improvement in fit owing to the inconsistencies in the nature of actual data (discrete counts) versus the distribution used to fit (continuous). Poisson distribution which models counts for rare events seems to be an appropriate model to fit the given data.

**Task 2: Model a Poisson distribution on new COVID-19 cases and deaths for all states. Describe how it is different from model worked first.**

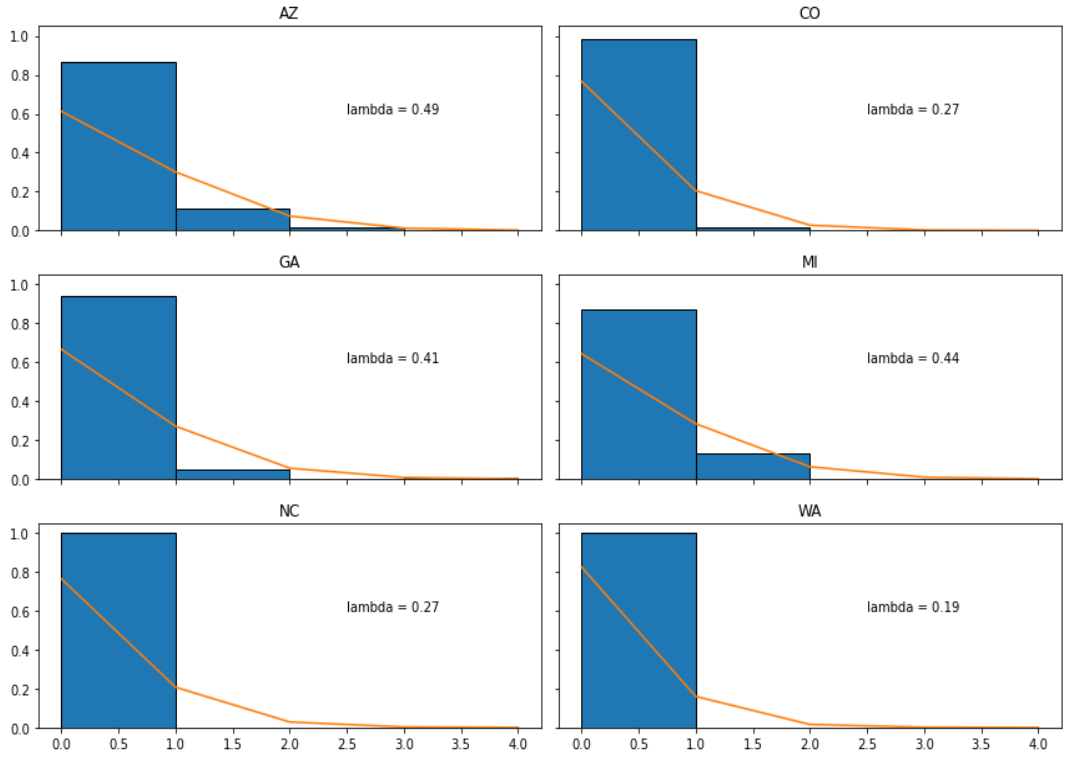
- For fitting Poisson distribution, we need to first compute the parameter estimates ($\lambda$ values). These estimates are computed using Method of Moments, i.e., the sample mean for each state, which can then be used to fit the theoretical probability distributions. The theoretical probabilities can be computed using `poisson.pmf` in `scipy.stats.poisson`.
- Shown below are the fitted distribution plots for new COVID-19 cases in the chosen states.




**Observations**:

- The graphs above show that Poisson model fits the distribution of new COVID-19 cases better than Gamma distribution.
- The peaks for theoretical probabilities for all states resemble those of the actual sample observations.
- The peaks of the Poisson probabilities are however, not the same as the peaks in the histogram of data points. This could be due to high variance in the new cases counts (as can be seen from the summary of descriptive statistics of the chosen states). Poisson distribution fits well when mean and variance are very close to each other which is not true for the new COVID-19 cases.

- The fitted distribution plots for COVID-19 deaths in the chosen states are:




**Observations**:

  - The distribution plots for deaths are fit well by the Poisson distribution. The peaks in the data are aligned with the theoretical probabilities for all the selected states.
  - NC and CO have the same values of $\lambda$ meaning that the mean number of deaths are same in these states.

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

- Correlation coefficients between COVID-19 data (new cases and deaths) and the chosen enrichment datasets (presidential election results, gubernatorial election results and senate election results were computed using the corr() function.
- The election results data considers the difference in votes won by Democrats versus Republicans. A negative difference indicates that Republicans won the majority in the state, while a positive difference indicates that Democrats won the majority.
- The correlation coefficients are computed based on state averages. Although the presidential election results have data for all states, data for senate elections is available for 34 states and data for governor elections is available for 11 states.
- After computing the COVID-19 aggregates of new cases and deaths, the suitable data files pertaining to the enrichment datasets are read. COVID-19 counts are normalized per 100,000 persons in the state. The election results are normalized per 100 persons in the state.

- The correlation coefficients between the chosen enrichment data variables and the number of new COVID-19 cases and deaths are tabulated below:

| Enrichment variable | New cases | Deaths |
|---|---|---|
| Presidential election | -0.566 | -0.341 |
| Gubernatorial election | 0.031 | 0.135 |
| Senatorial election | 0.205 | -0.007 |

- Interpretations:

  A) Presidential election results 2020

  - Correlation coefficient between presidential votes in the states and new COVID-19 cases in the US is **-0.566**. It means that there is a negative linear relationship between difference in votes between Democrat and Republican party, and the incidence of COVID-19 cases. A lower or negative difference (meaning Republicans got more votes) is associated with increase in new cases.
  - Correlation coefficient between presidential votes in the states and COVID-19 deaths in the US is **-0.341**. It means that there is a negative linear relationship between difference in votes between Democrat and Republican party, and COVID-19 deaths. A lower or negative difference (meaning Republicans got more votes) is associated with increase in deaths.

  B) Gubernatorial election results 2020

  - Correlation coefficient between gubernatorial votes in the states and new COVID-19 cases in the US is **0.031**. It means that there is a very weak but positive linear relationship between difference in votes between Democrat and Republican party, and the incidence of COVID-19 cases. A lower or negative difference (meaning Republicans got more votes) is associated with decrease in new cases.
  - Correlation coefficient between gubernatorial votes in the states and COVID-19 deaths in the US is **0.135**. It means that there is a weak but positive linear relationship between difference in votes between Democrat and Republican party, and COVID-19 deaths. A lower or negative difference (meaning Republicans got more votes) is associated with decrease in deaths.

    NOTE: There were very few states with gubernatorial election results available (11 states).

  C) Senate election results 2020

  - Correlation coefficient between senatorial votes in the states and new COVID-19 cases in the US is **0.205**. It means that there is a weak but positive linear relationship between difference in votes between Democrat and Republican party, and the incidence of COVID-19 cases. A lower or negative difference (meaning Republicans got more votes) is associated with decrease in new cases.

- Correlation coefficient between senatorial votes in the states and COVID-19 deaths in the US is -**0.007**. It means that there is a weak but positive linear relationship between difference in votes between Democrat and Republican party, and COVID-19 deaths. A lower or negative difference (meaning Republicans got more votes) is associated with decrease in deaths.

  NOTE: There were 34 states with senatorial election results available out of all the states in the US.

**Task 4: Formulate hypothesis between Enrichment data and number of cases to be compared against states. Choose three different variables to compare against.**

- Is political affiliation to Republican party in presidential elections 2020 associated with an increase in the number of new cases? (Political affiliation to Republican party is indicated by a negative difference in votes between Democrat and Republican parties)
- Is political affiliation to Republican party in gubernatorial elections 2020 associated with an increase in the number of new cases?
- Is political affiliation to Republican party in senatorial elections 2020 associated with an increase in the number of new cases?