Project Stage II Report

By Amantii Samson, Alejandro Penaloza,
Peter Yuan, Ritu Joshi and Quinn Tjin-a-soe

Goals: The goal of Stage II is to develop formal hypothesis tests for the intuitions you had in Stage I and utilize statistical modeling to prove/disprove them.

Task 1:

- *Compare the weekly statistics (mean, median, mode) for the number of new cases and deaths across the US. You are calculating the mean (rounded to integer value) number of new cases and per week and then calculating (mean, median, mode) for all weeks taken together*.

- *Compare the data against other countries of the world (Brazil, Indonesia, Pakistan, Nigeria, Bangladesh)*
  We have chosen 5 countries where the overall country population is closer to the population of the US. Various preprocessing steps were taken to prepare the country wide data to compare with the US.
  - ➔ Data was only collected from Jan 22,2020 to Sep 15, 2020. Additional data was removed from the confirmed cases and deaths dataset.
  - ➔ NaN were replaced with 0's.
  - ➔ Number of new cases were calculated for each day by taking a difference of numbers from the previous day.
  - ➔ Cases were normalized by population of the country and then multiplied by 1,000,000 to get the count per 1,000,000 population.
  - ➔ Mean, median and mode are calculated for confirmed cases and deaths both for the selected 5 countries.
  - ➔ The comparison of mean median and mode among these 6 countries shows that….
- *Plot daily trends (cases and deaths) of the US and compare other countries. Utilize aggregate, normalized by population, difference in cases (new cases), and log normalized values*.
  - ○ Using the normalized data computed in the previous stage, the daily trends for the cases and deaths of the US compared to other countries was plotted

    - An aggregation function for the means of each column was also utilized on cases and deaths of the other countries along with plots
    -
- *Identify peak week of the cases and deaths in US and other countries.*

Task 2 - Member task

Ritu Joshi

- ***Fit a distribution to the number of COVID-19 cases of a state.***
    - ***Graphically plot the distribution and describe the distribution statistics.***
    - ***Describe why the distribution was choosen and its statistics in the report and the notebook.***

- I have chosen a poisson distribution for my data for the following reasons:
    - The dataset is discrete and we know how many times an event has occurred over a certain time period.
    - Dataset is random and independently collected. There is no bias associated with this data.
- Distribution Statistics after normalizing by population

Mean : 62.7899

Variance : 2385.010

Skewness : 0.1838

Kurtosis : -0.6943

Standard deviation - 48.8365

The dataset has huge variance. The data has very little positive skewness which explains that there is a small tail to the right of the data distribution. Negative kurtosis means that the data has light tail than a normal distribution. It also means that the data is not too peaked.

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

- Based on the calculation of correlation coefficient between the enrichment dataset and COVID19 dataset, we see a strong/high correlation between Number of ICU beds and deaths, Adults ICU Beds and deaths.
- This correlation does not mean that if a county/state has more ICU beds then there are more deaths. There are other factors too which can be combined with correlation information to give us some information. For example, we see that Number ICU Beds and deaths have high correlation, from that we can see that population and Number ICU beds also have high correlation. It means that a county with high population may have a higher number of ICU beds and because this is a high population county, the number of deaths are also higher.

- ***Formulate hypothesis between Enrichment data and number of cases to be compared against states. Choose 3 different variables to compare against***

- Does higher higher confirmed cases lead to higher bed utilization?
- Does higher deaths mean higher ventilator usage?
- Does a higher Number of Staffed beds mean higher confirmed cases?

<u>Amantii Samson</u>

- ***Fit a distribution to the number of COVID-19 cases of a state.***
  - Poisson distribution selected because there are discrete values within the dataset.
- ***Perform correlation between Enrichment data variables and COVID-19 cases to observe any patterns.***
  - Expect to see a relatively high correlation between the number of employed people and the number of COVID cases because areas with larger groups of people are expected to spread the virus easier
  - Industry is expected to have a relatively moderate correlation with an increase in the number of COVID cases because certain industries have large groups of people working together while other industries do not. Also, some industries may not be very popular within certain states such as factories or farming/agriculture.
  - Total quarterly wage is expected to have a relatively low or moderate correlation with the number of COVID cases because states/areas with higher quarterly wages could either indicate that an area is more densely populated or that the wages of workers within the area are higher than the national average. Therefore, total quarterly wages could either be moderately correlated in the case of a more densely populated area, or it could be minimally correlated in the case of a less densely populated area with higher wages.

- ***Formulate hypothesis between Enrichment data and number of cases to be compared against states. Choose 3 different variables to compare against***
  - Does the type of industry of employment indicate an increase in the number of COVID cases?
  - Looking at the employment numbers for the first three months of 2020, does a higher employment rate indicate a higher number of COVID cases?
  - Does a higher total quarterly wage indicate higher number of cases?

Peter Yuan

- ***Fit a distribution to the number of COVID-19 cases of a state.***
  - ***Graphically plot the distribution and describe the distribution statistics.***
  - ***Describe why the distribution was choosen and its statistics in the report and the notebook.***

- I have chosen a poisson distribution for my data for the following reasons:
  - The dataset is discrete
  - From the histogram, the data spread seems like a poisson distribution when lambda is 1.
- Distribution Statistics after normalizing by 10000 population which means that how many new cases per 10000 population

  Mean : 1

  Variance : 1.47

  Skewness : 1.2

  Kurtosis : 1.96

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

- I have computed all the selected variables with case number. We could not obvious see any pattern from the result. Some variables have a positive correlation and some have a negative correlation, the highest is only 0.3. It seems not that high correlated.

- ***Formulate hypothesis between Enrichment data and number of cases to be compared against states. Choose 3 different variables to compare against***

- Does higher percentage of service occupation lead to higher confirmed cases?
- Does a higher percentage of family income from 10000 to 14999 lead to higher confirmed cases?
- Does a higher percentage of no health insurance coverage lead to higher confirmed cases?