Name: **Sakshi Patel**

Project Stage - III (Distributions and Hypothesis Testing)

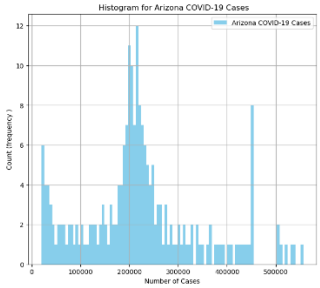
1. Introduction

The project focuses on the analysis of COVID-19 cases using statistical modeling and data correlation techniques. Key objectives include fitting distributions to COVID-19 cases, modeling Poisson distributions for case counts and deaths, and exploring correlations between demographic (enrichment) variables and COVID-19 outcomes. The data used for this analysis was provided in earlier phases of the project, covering different states in the U.S.

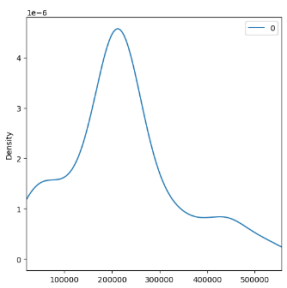
2. Task 1: Fitting a Distribution to COVID-19 Case Data

2.1. Data Overview

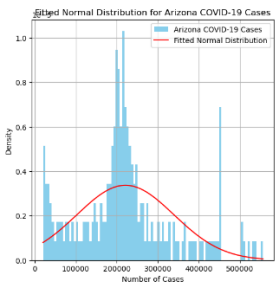
The analysis started by extracting COVID-19 case data for the state of Arizona from a larger dataset. Data was filtered from June 1, 2020, to January 3, 2021. A total of 217 days of data was analyzed, representing daily COVID-19 cases across multiple counties within the state.



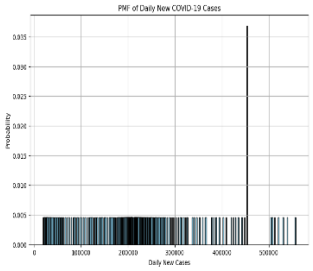
Histogram for Arizona COVID-19 Cases



Probability Density Plot



Fitted Normal Distribution for Arizona COVID-19 Cases



PMF of Daily New COVID-19 Cases

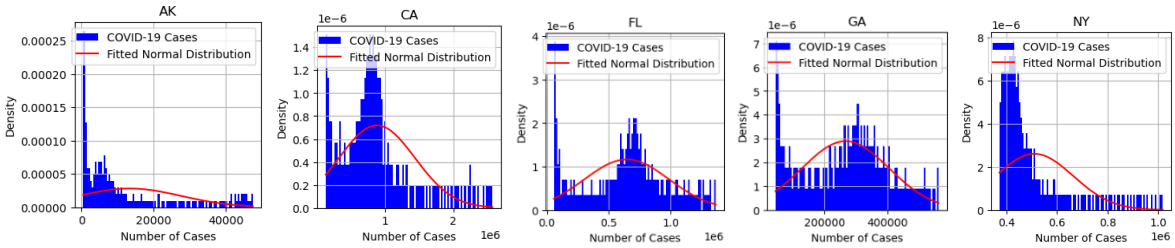
2.2. Distribution Selection and Statistics

A continuous, right-skewed distribution was selected for the COVID-19 cases. The distribution's mean, variance, skewness, and kurtosis were calculated:

- Mean: 221,013.68
- Variance: 13,997,220,198.95
- Skewness: 0.58 (right-skewed)
- Kurtosis: 0.27

A histogram and a probability mass function (PMF) were plotted to visualize the data, indicating the discrete nature of daily case counts. A fitted normal distribution was also overlaid on the histogram to further explore the distribution shape.

2.3. Comparison with Other States



AK CA FL GA NY

- CA and NY have more extreme data distributions with higher skewness and kurtosis, meaning they have more outliers and greater variability.
- FL and GA have more symmetric, less peaked distributions, suggesting more even spread across their data.
- AK has the smallest variability and skewness, indicating more uniform data with less outlier presence.
- States like California and Florida experienced large and fluctuating COVID-19 outbreaks, with periods of extreme case surges, as reflected in their high means, variability, and positive skewness.
- Arizona had a moderate outbreak, with some periods of increased cases, but it was more stable compared to highly impacted states like CA.
- Alaska had a much smaller outbreak, with more consistent and lower case numbers, indicated by its low mean and low variability.
- States like New York and Georgia also experienced notable outbreaks, with NY showing more extreme events, while GA had a more stable case count.

3. Task 2: Modeling Poisson Distribution for COVID-19 Cases and Deaths

3.1. Poisson Distribution for Case Counts

Poisson distributions were modeled for daily new COVID-19 cases per 100,000 people in Arizona and compared to five other states. Arizona, California, and Florida displayed wider, more variable distributions, indicating fluctuating case numbers. States like Alaska and Georgia had narrower distributions, implying more stable case patterns.

3.2. Poisson Distribution for Deaths

A similar Poisson distribution was modeled for COVID-19 deaths, showing that states like Alaska experienced fewer fluctuations in daily death counts, while Florida and Arizona had wider, more variable distributions.

3.3. Differences Between Poisson and Continuous Distribution Models

- **Poisson Distribution**:
    - **Discrete**: The Poisson distribution models the probability of a number of discrete events occurring in a fixed interval of time or space.
    - Example: Counting the number of daily new COVID-19 deaths per 100k population.
- **Continuous Distribution**:
    - **Continuous**: A right-skewed continuous distribution models data that takes on a range of continuous values, such as waiting times, income levels, or life spans.
    - Example: Modeling the time between COVID-19 deaths or the length of hospital stays.

2. Shape and Skewness:

- **Poisson Distribution**:
    - For smaller values of the mean (λ), the Poisson distribution is **right-skewed**, but it becomes more symmetric as λ increases.
    - The distribution shows distinct "steps" as it represents discrete probabilities for each possible count.
- **Continuous Distribution**:
    - Always skewed to the right (tail to the right), meaning that most data points are concentrated on the left with a long tail on the right.

4. Task 3: Correlation Between Enrichment Data and COVID-19 Outcomes

4.1. Hypothesis Formulation

Hypotheses:

1. **Age Groups and COVID-19 Outcomes**:

- **Hypothesis**: Counties with a larger proportion of children (5-9 years old and under 18 years) will have fewer COVID-19 deaths per 100k population.
- **Reasoning**: Younger populations are less likely to experience severe symptoms, which may contribute to lower mortality rates.

2. **Under 5 Years Population and COVID-19 Outcomes**:
    - **Hypothesis**: Counties with a higher percentage of the population under 5 years old will have fewer cases and deaths per 100k population.
    - **Reasoning**: Infants and young children are generally at lower risk of severe COVID-19 outcomes compared to older adults, though they may contribute to transmission.
3. **Female Population and COVID-19 Outcomes**:
    - **Hypothesis**: Counties with a higher proportion of females will have similar or slightly lower COVID-19 death rates compared to males.
    - **Reasoning**: Globally, men have shown a slightly higher COVID-19 death rate, though female populations may exhibit different outcomes based on cultural and socioeconomic factors.

4.2. Correlation Analysis

A correlation matrix was calculated for the enrichment variables against total COVID-19 cases and deaths. Significant correlations were found between variables such as race (American Indian and Alaska Native populations) and COVID-19 outcomes. The findings suggest that demographic factors play a critical role in the spread and impact of COVID-19 across different regions.