Statistics Final Project – Written Report

Robert Chen, Yuming Lu, Jerry Yao

December 21 2022

1 Topic & Outline

The prompt we chose is to analyze a real-world dataset. Using the CDC COVID-19 data, **the main problem we studied and basic assumptions are as follows: Given a population of n people, what is the probability of having $N(N = 1, 2, 3, ...)$ new COVID cases on day $k(k = 1, 2, 3, ...)$?** We assumed that the number of days before a person i gets infected is a random variable X_i, let $i = 1, 2, 3, ..., n$ and $X_1, X_2, ..., X_n$ be *i.i.d.*

The background and the outline are as follows: We are now in the post-pandemic era. Although there are still reported new cases and multiple COVID variants, the number of daily new COVID cases has become more stable. Supported by real-world data, the goal of our research is to gain more knowledge about the daily new COVID cases through statistical modeling and interference. We used several ideas studied in class, including likelihood function and confidence interval. In the rest of the report, we are going to present the basic model of the main problem and a lemma. In Section 3, we are going to apply our model to real-world data analysis. In Section 4, we are going to discuss our results. In Section 5, we are going to present potential future work.

2 Basic Model

Given the main problem in Section 1, we planned to use exponential distribution to model the number of days before a person gets infected. Recall that exponential distribution describes the time elapsed between two events. Admittedly, exponential distribution has memoryless property, which suggests that the number of those who are infected will not affect the number of new cases. Therefore, this basic model is not compatible with the highly infectious nature of Coronavirus, so it will be improved. But for now, let $X_i \sim \text{Exp}(\lambda)$. The probability density function of this exponential distribution is given by Equation (1).

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Exponential distribution describes continuous random variable. Assume a day starts at 0:00 and ends 24 hours later at 24:00. The probability that person X_i get infected on day j is given by Equation (2).

$$\mathbb{P}[k \leq X_i \leq k+1] = \mathbb{P}[X_i \leq k+1] - \mathbb{P}[X_i \leq k] = \mathrm{e}^{-\lambda(k)} - \mathrm{e}^{-\lambda(k+1)} \tag{2}$$

Let random variable Y denote the number of new COVID cases on the day k. We assume that, on the day k, the probability for k people to get infected can be modeled by binomial distribution. $Y \sim \mathrm{Bi}(n,p)$, the pdf is given by Equation (3), $\mathbb{P}[k \leq X_i \leq k+1] = p$. The relationship between $X\&Y$ is given by Figure 1.

$$\mathbb{P}[Y = N] = \binom{n}{N} \mathbb{P}[k \leq X_i \leq k+1]^N (1 - \mathbb{P}[k \leq X_i \leq k+1])^{n-N} \tag{3}$$

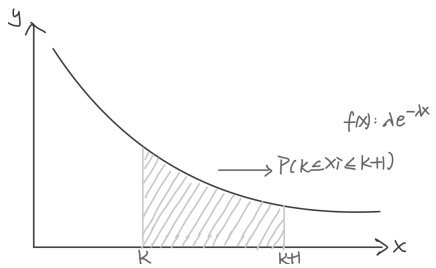

Figure 1: Relationship between $X\&Y$

The goal is to solve for $\hat{\lambda}_{MLE}$ in the statistic model. Observe Figure 1, we will be using a lemma (Lemma 1): For fixed n, find $\hat{P}_{MLE}$ given that $X_1, X_2, ..., X_N \sim \mathrm{Bi}(n,p)$. The likelihood function is as follows

$$L_n(p) = \prod_{i=1}^n f_{n,p}(X_i) = \prod_{i=1}^n \binom{n}{X_i} p^{X_i}(1-p)^{n-X_i}$$

Since n is fixed, $\prod_{i=1}^n \binom{n}{X_i}$ is a constant, so let $\prod_{i=1}^n \binom{n}{X_i} = C$, and let $S = X_1 + X_2 + ... + X_N$. The likelihood function can be written as:

$$L_n(p) = C * \prod_{i=1}^n p^{X_i}(1-p)^{n-X_i} = C * p^S (1-p)^{n(N-S)}$$

The log-likelihood function is as follows:

$$l_n(p) = l_n(C) + S l_n(p) + (nN - S) l_n(1-p)$$

Let $\frac{dl_n(p)}{dp} = 0$, The MLE estimator is given by Equation (4):

$$\hat{P}_{MLE} = \frac{S}{nN} = \frac{\bar{X}_N}{n} \tag{4}$$

Hence for given dataset $Y_1, Y_2, ..., Y_m \sim \text{Bi}(n, P(k \leq X_i \leq k+1))$ Recall that n is the total population. $Y_1, Y_2, ..., Y_m$ are the number of new cases on day k. The maximum likelihood estimator for $P(k \leq X_i \leq k+1) = \frac{1}{n}\bar{Y}_m$ by Equation (4). By Equation (1), Equation (5) can be proved:

$$e^{-\lambda_{MLE}k} - e^{-\lambda_{MLE}(k+1)} = \frac{1}{n}\bar{Y}_m \tag{5}$$

Equation (5) is hard to solve analytically. Instead, we drew the graph to calculate $\hat{\lambda}_{MLE}$ numerically. The graph of $e^{-\lambda k} - e^{-\lambda(k+1)} = \frac{1}{n}\bar{Y}_m$ is as follows:

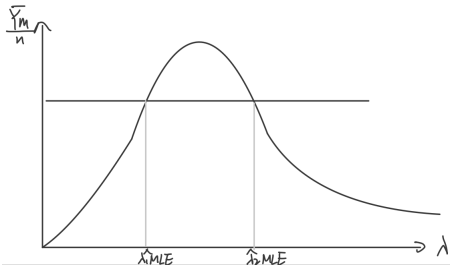

Figure 2: Graph of $e^{-\lambda k} - e^{-\lambda(k+1)} = \frac{1}{n}\bar{Y}_m$

The next goal is to calculate the confidence interval for λ. The first step is to calculate the confidence interval for $P(k \leq X_i \leq k+1)$. Let p denote $P(k \leq X_i \leq k+1)$. By Theorem 9.19 Wasserman, the $1-\alpha$ confidence interval for p is as follows:

$$C_p = (\hat{p}_{MLE} - z_{\frac{\alpha}{2}}\hat{s}_e, \hat{p}_{MLE} + z_{\frac{\alpha}{2}}\hat{s}_e) \tag{6}$$

In Equation (6), $\hat{s}_e = \sqrt{\frac{1}{I_n(\hat{p})}}$, and I_n is the Fisher Information of Binomial distribution. $I_n(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})}$, so that the $1-\alpha$ confidence interval for p is as follows:

$$C_p = (\frac{\bar{Y}_m}{n} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \frac{\bar{Y}_m}{n} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \tag{7}$$

Now we go backward to calculate the confidence interval for λ numerically. There might be two situations, and confidence intervals for λ in the two situations are shown in Figure 3 and Figure 4.

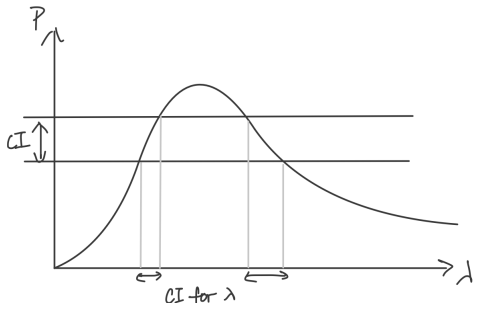

Figure 3: Confidence intervals for λ – Situation 1

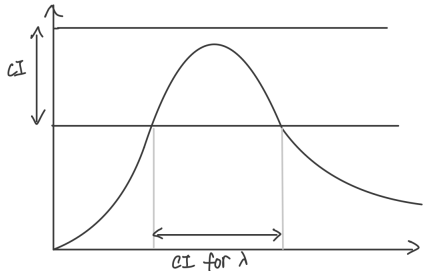

Figure 4: Confidence intervals for λ – Situation 2

3 Real World Data Analysis

Choose country $C_1, C_2, ..., C_m$, and assume they have a population of $p_1, p_2, ..., p_m$ separately. Choose a date as day 0. For this project, we chose Jan 1, 2022, roughly when COVID Omicron variant outbroke worldwide. On day k, each city should have new cases $c_1, c_2, ..., c_m$

However, according to our model, the total population should be the same. Therefore, we may scale $p_1, p_2, ..., p_m$ to an appropriate number n. Now, the number of new cases each day is $\frac{c_1}{p_1}n, \frac{c_2}{p_2}n, ..., \frac{c_m}{p_m}n$.

Now our new data:

$$Y_1 = \frac{c_1}{p_1}n, Y_2 = \frac{c_2}{p_2}n, ..., Y_m = \frac{c_m}{p_m}n$$

should fit into our model.

We will use the data from *ourworldindata.org*. We picked March 1, 2022 as the experiment date. Hence $k = 59$. All the daily new cases have already been scaled to cases per million people by the website., meaning we can use the data directly.

Hence the maximum likelihood estimator for p can be calculated:

$$\hat{p}_{MLE} = 3.025 \times 10^{-4}$$

The maximum likelihood estimator for λ is correspondingly:

$$\hat{\lambda}_{MLE} = 3.081 \times 10^{-4} or 9.699 \times 10^{-2}$$

A 95% confidence interval for p is:

$$CI = (2.684 \times 10^{-4}, 3.366 \times 10^{-4})$$

and the 95% confidence interval for λ is:

$$CI = (2.728 \times 10^{-4}, 3.435 \times 10^{-4}) or (9.481 \times 10^{-2}, 9.941 \times 10^{-2})$$

the code and the graph are as follows:

```
In [1]: import numpy as np
        import scipy.stats as stats
        import matplotlib.pyplot as plt
        import seaborn as sns

/Users/muddy_flesh/opt/anaconda3/lib/python3.9/site-packages/scipy/__init__.py:146: UserWarning: A NumPy version >=1.
16.5 and <1.23.0 is required for this version of SciPy (detected version 1.23.4
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"

In [2]: ## Calculating p_mle, the maximum likelihood estimator for
        ## Set 2022-01-01, roughly when Omicron bursted worldwide as day 0
        ## Choose 2022-03-01 as the experiment date, hence k = 59
        Cases59 = np.array([7251.30, 2289.81, 1759.26, 1498.12,1323.55,1310.15,1298.10,1230.20,1044.83,908.66,714.51,704.49,596
                  541.71,541.58,452.67,387.60,305.67,261.51,251.44,247.58,234.45,225.02,218.07,198.53,185.42,160.88,160.62,
                  158.78,135.86,129.76,121.66,114.07,107.97,102.06,98.54,83.89,77.44,73.50,71.57,68.11,63.11,51.30,49.77,
                  48.78,48.36,48.05,47.84,44.29,40.06,39.30,37.86,37.08,35.68,35.13,29.09,27.40,25.42,25.03,22.86,21.77,
                  21.16,21.11,20.85,19.66,19.23,17.84,17.37,17.34,17.21,16.95,16.21,11.65,11.13,8.18,7.53,7.40,6.94,5.87,
                  5.86,5.61,5.06,4.47,3.85,3.64,3.63,2.89,2.72,2.23,2.21,2.18,1.86,1.55,1.54,1.44])
        ## Data from https://ourworldindata.org/, is the new cases per million in each country.
        n = 1000000
        Ym_hat = Cases59.mean()
        p_mle = Ym_hat/n
        p_mle

Out[2]: 0.00030250336842105264

In [5]: ## Calculating a 95% Confidence interval for p
        se_hat = np.sqrt(p_mle*(1-p_mle)/n)
        an = p_mle - 1.96*se_hat
        bn = p_mle + 1.96*se_hat
        an, bn

Out[5]: (0.00026841898193856635, 0.00033658775490353894)
```

Figure 6: Solve for $\hat{p}_{MLE}$ & 95% confidence interval for p

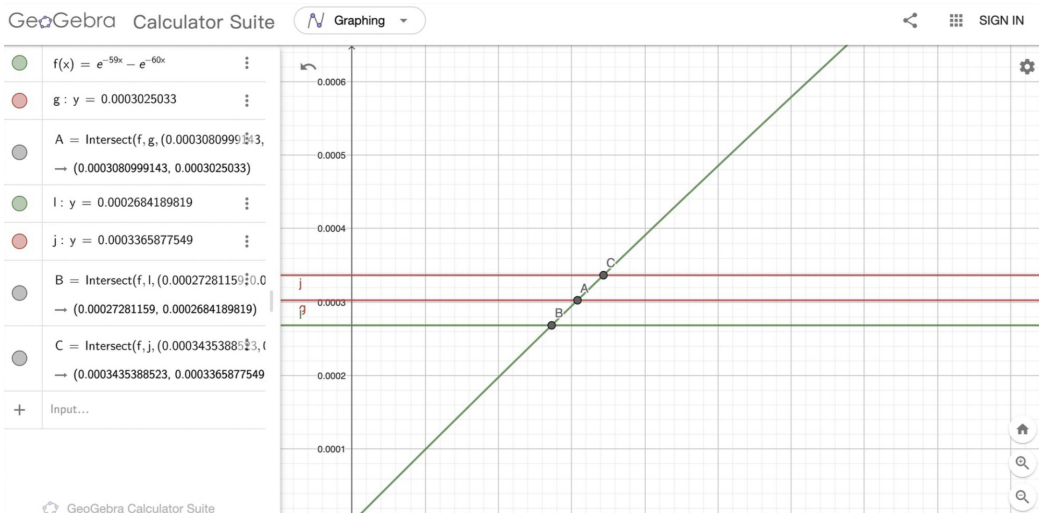

Figure 7: $\hat{\lambda}_{MLE}$ is point A; 95% confidence interval for $\hat{\lambda}_{MLE}$ between two lines that cross point B and C separately.

Now we shall check whether our model fits the result by choosing Dec 1, 2022 as the experiment date ($k = 334$). Now $X_1, X_2, ..., X_n \sim \text{Exp}(3.081 \times 10^{-4})$. Therefore,

$$P(k \leq X_i \leq k+1) = \int_{335}^{334} 3.081 \times 10^{-4} \times e^{-3.081 \times 10^{-4} x} dx = 2.7 \times 10^{-4}$$

. Then we compare the distribution with the new real-world distribution on day 334. As shown in Figure 8, although the real-world distribution has some outliers, our model distribution generally fits the real-world distribution.

```
In [26]: Cases334 = np.array([5437.55, 3141.22, 1985.47, 1830.15, 1353.61, 1022.60, 1021.24,953.61, 822.12,766.57,654.66,
         605.19,594.43,512.35,496.26,457.85,407.29,393.64,379.86,364.88,348.59,346.18,325.82,322.90,285.25,257.14,232.87,
         231.33,218.65,215.10,205.02,190.54,125.73,96.48,95.10,91.27,80.83,79.89,79.45,78.98,77.38,69.98,68.25,60.24,59.73,
         57.80,52.56,49.44,46.56,42.88,39.38,31.65,24.82,23.60,22.56,22.17,18.07,13.93,13.24,10.99,8.66,8.30,8.20,7.65,5.98,
         5.92,5.31,4.95,4.78,4.78,4.63,4.45,4.04,3.55,2.88,2.63,2.12,1.38,1.18,1.18,1.11,0.54,0.54,0.45,0.42,0.35,0.32,0.29,0.23
         0.23,0.19,0.16,0.13,0.07,0.07,
         ])

In [35]: x = np.random.binomial(n=10**6, p=2.7*10**(-4), size=600)

         sns.distplot(x, hist=True, kde=False, label="Model Distribution")
         sns.distplot(Cases334,hist=True, kde=False, label="Real worldDistribution")
         plt.legend()
         plt.show()
```

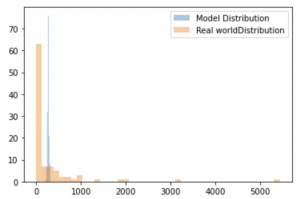

Figure 8: Model Distribution vs. Real-World Distribution

4 Result & Discussion

After deriving $\hat{\lambda}$, the first question one might ask is: is this parametric model able to predict the number of daily new cases for the future? The answer is yes. By plugging in $\hat{\lambda}$, we can derive the probability distribution function on any day in the future.

Moreover, we can see that the number of daily new cases is converging to 0 according to our model. This is because as k goes to ∞, the following equation holds:

$$\lim_{k\to\infty} P(k \le X_i \le k+1) = \lim_{k\to\infty} e^{-\lambda k} - e^{-\lambda(k+1)} = \lim_{k\to\infty} e^{-\lambda k}(1-e^{-\lambda}) = 0 \quad (8)$$

Therefore, $\forall N \le n$, the following equation holds:

$$\lim_{k\to\infty} P(Y_i = N) = \binom{n}{N} \lim_{k\to\infty} P(k \le X_i \le k+1)^N (1-P(k \le X_i \le k+1))^{n-N} = 0 \quad (9)$$

Therefore, $\lim_{k\to\infty} P(Y_i = 0) = 1$. It means that for large k, the probability of having 0 new cases on day k should be 1. This also shows that our model has fit in the real-world case. As the daily new cases are gradually decreasing, the probability of having new cases will eventually be 0.

5 Potential Directions for Future Work

We realized that it is arbitrary to say that $X_1, X_2, ..., X_n$ are parametrized by a fixed λ. Therefore, we should further introduce the classification of $X_(i)$ and multi-parametrization. Notice λ can be influenced by two variables – population density and population of elders & youngsters. Therefore, countries can be classified into four different groups, corresponding to the four quadrants in Figure 5:

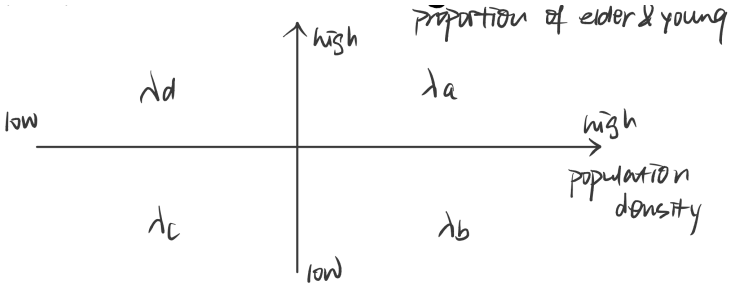

Figure 5: Classification of countries

Now our model is parametrized by four parameters, namely $\lambda_a, \lambda_b, \lambda_c, \lambda_d$. The dataset should be split into four groups:

1. $X_{a1}, X_{a2}, ..., X_{an} \sim \text{Exp}(\lambda_a)$
2. $X_{b1}, X_{b2}, ..., X_{bn} \sim \text{Exp}(\lambda_b)$
3. $X_{c1}, X_{c2}, ..., X_{cn} \sim \text{Exp}(\lambda_c)$
4. $X_{d1}, X_{d2}, ..., X_{dn} \sim \text{Exp}(\lambda_d)$

Using the four groups of data, we can derive the maximum likelihood estimator using the same method. However, the factors that can affect the classification of countries can be much more complicated. We might have infinite parameters, which means our model will no longer be a parametric model. In this case, we may consider λ as a function depending on whatever variables we are concerned about: $\lambda(\alpha_1, \alpha_2, ..., \alpha_n)$. In this case, we may still derive the maximum likelihood estimator $\hat{\lambda}_{MLE}$

6 Bibilography

1. Larry Wassermann, All of Statistics: A Concise Course in Statistical Inference (Springer Texts in Statistics), 2010.
2. "Covid-19 Data Explorer." Our World in Data, https://ourworldindata.org/explorers/coronavirus-data-explorer?tab=table.