PHSX815_Project3:
Rakghoul Plague Spread Model Parameter Fitting

Gene Stejskal

April 10, 2023

1 Introduction

In this project, the group of scientists who originally worked to cure the Rakghoul plague (projects 1 and 2), are continuing their research into the plague. For this project, the group is applying live samples of the plague to artificially created human cells, and is collecting data hourly on the number of newly infected cells. From this data, they want to develop a plague spread model, which may prove beneficial in modelling and creating future cures. The model is nearly finished, but there is one last parameter, which will give them vital information on how fast the plague can spread, which they are trying to estimate from the collected data using minimization techniques.

This report is organized as follows: Sec. 2 discusses the experimental set up and how the data is collected, and a description of the computer simulation and minimization techniques used is shown in Sec. 3, with an analysis of the estimated parameter as well as it's associated error in Sec. 4. Finally, conclusions are presented in Sec. 5.

2 Discussion of the experiment and data collection

As mentioned above, the goal of this project is to, first, collect plague spread data from samples of artificially created human cells; then develop a model, which will allow us to predict the possible spread of the plague in an actual human population. From the gathered data, we want to use minimization methods to estimate a fitting parameter of our model.

Given four cultures of human cells, one to act as the control (no plague introduced), and the other three inoculated with rakghoul plague bacteria, we will collect data hourly on the number of newly infected cells. The three inoculated samples, will be watched for different amounts of time; batch 1, for 100 hours, batch 2, for 5500 hours, and batch 3 for 10000 hours. This will allow us to get multiple estimations of model parameter with varying errors. We expect that the more data is collected, the better our fitting parameter will be (less error).

3 Code and Experimental Simulation

The bones of our model are similar to past experiments. We are using three separate codes. The first being a random class which contains all of the code used to generate the distribution. Second, there is the data generation code, which uses the random class to generate a list of data according to our chosen distribution. And, lastly, there is the analysis code, which collects and reads our data, uses

the data to estimate the value of the fitting parameter and its error interval, then plots the data in a histogram.

From previous experiments with the plague, the plague spread data was found to be best described by a Log-normal distribution, which was chosen because the log-normal distribution is often used to model epidemiological data such as virus spread. The "log-normal" distribution, X, takes a random variable, R, generated according to a normal distribution, and transforms it using an exponential function:

$$X = e^{\mu + R\sigma}$$

where μ and σ are parameters related to the actual mean and standard deviation of the log-normal distribution, mux and $sigx$. I, however, want the distribution to take the variables mux and $sigx$, so transformations are needed to convert from μ and σ to mux and $sigx$. These are the following:

$$\mu = \ln\left(\frac{mux^2}{\sqrt{mux^2 + sigx^2}}\right) \tag{1}$$

$$\sigma^2 = \ln\left(1 + \frac{sigx^2}{mux^2}\right) \tag{2}$$

This allows use to have our distribution $X = X(mux, sigx)$. For this model, we fixed the average of our samples, $mux = 50$. This way we could threat the model as one dimension, and only focus on estimating the standard deviation, $sigx$. The data collected was generated using a distribution with a true standard deviation of $sigx = 10$. To estimate the value of our parameter given the data, which was collected, the negative log of the likelihood, shown in equation 3, was minimized with respect to our parameter of interest, $sigx$:

$$\frac{d(-ln(L(sigx))}{dsigx} = 0 \tag{3}$$

such that the likelihood, $L(sigx)$, is given by:

$$L(sigx) = \prod_{i=1}^{N} P(x_i|sigx) \tag{4}$$

Here, N is the total number of measurements, and P is the probability density function (PDF) for the log-normal distribution given by:

$$P(x_i|sigx) = \frac{1}{x_i\sqrt{2\pi\sigma^2}} exp\left(-\frac{(lnx_i - \mu)^2}{2\sigma^2}\right) \tag{5}$$

Finding the value of $sigx$, which minimizes $-ln(L(sigx))$, will be a good estimate for the true value, $sigx_{true}$.

The rest of the code is devoted to determining the upper and lower error bounds of our estimations. To find these bounds, the log likelihood ratio is examined. The upper and lower bounds of the estimated parameter are the roots of:

$$-\frac{1}{2} \leq ln\left(\frac{L(sigx)}{L(sigx_{best})}\right) \tag{6}$$

such that, $sigx_{low} \leq sigx_{best} \leq sigx_{high}$. These values are the bounds of the FWHM, which is width or difference between the log likelihood at 1/2 of the maximum. From equation 6, it is also clear that at $sigx = sigx_{best}$, the function is at a max, since $ln(1) = 0 \geq -0.5$.

The next section will discuss and analyze the outputs of the experiment as well as the corresponding figures.

4 Output Analysis

After gathering the three sets of data, the following parameter estimations were gathered. For $N = 100$ measurements (data collected for 100 hours), the value of $sigx$ is estimated to be 9.738, with error interval of [9.077,10.471]. For $N = 5,500$ measurements (data collected for 5,500 hours), the value of $sigx$ is estimated to be 9.847 with an error interval of [9.753,9.94]. For $N = 10,000$ measurements (data collected for 10,000 hours). the value of $sigx$ is estimated to be 9.9083 with an error interval of [9.837,9.979].

From this, it's seen that the error intervals are decreasing as the number of measurements increases. This makes sense as standard error is proportional to $1/\sqrt{N}$, which give us errors of 0.974, 0.132, 0.09908; these errors match up with the corresponding error intervals. Again, it's seen that these errors also decrease as the number of measurements increase.

These values of $sigx$ also seem to be converging on approximately 9.9, with rapidly decreasing error. Looking at Fig. 1-3, it can be seen that as the number of measurements increases, the shape of the distribution also becomes more resolved, and the value of the standard deviation becomes more apparent. In Fig.1, where N=100, it is not clear what the shape of the distribution is, other than a rough idea of the average value being around 50. For $N = 5,500$ and $N = 10,000$, seen in Fig 2-3, the shape is much more resolved allowing for a better picture of the model, and of what the value of the standard deviation is.

5 Conclusion

To conclude, in this experiment, three cultures of artificially created human cells were inoculated with the rakghoul plague, and the number of newly infected cells was recorded every hour for N=100, 5500, and 10000 hours. This data was then used to estimate the most likely value of a fitting parameter for the plague spread model. The parameter in question, is approximately the standard deviation of a log-normal distribution, which was used as the model for our data. From the data and subsequent analyse, values of $sigx$ were found to be converging around a value of 9.9 (the true value is 10) as the number of measurements increased. As the number of measurements increased, it is also seen that the error interval on the value is shrinking, meaning that the more data that is taken, the less error is involved in the parameter estimation, and the better the model can be fitted to the data.

References

[Rogan] `https://github.com/crogan?tab=repositories`

[Stejskal] `https://github.com/Stayskul`

[Wikipedia] `https://en.wikipedia.org/wiki/Log-normal_distribution#Properties`

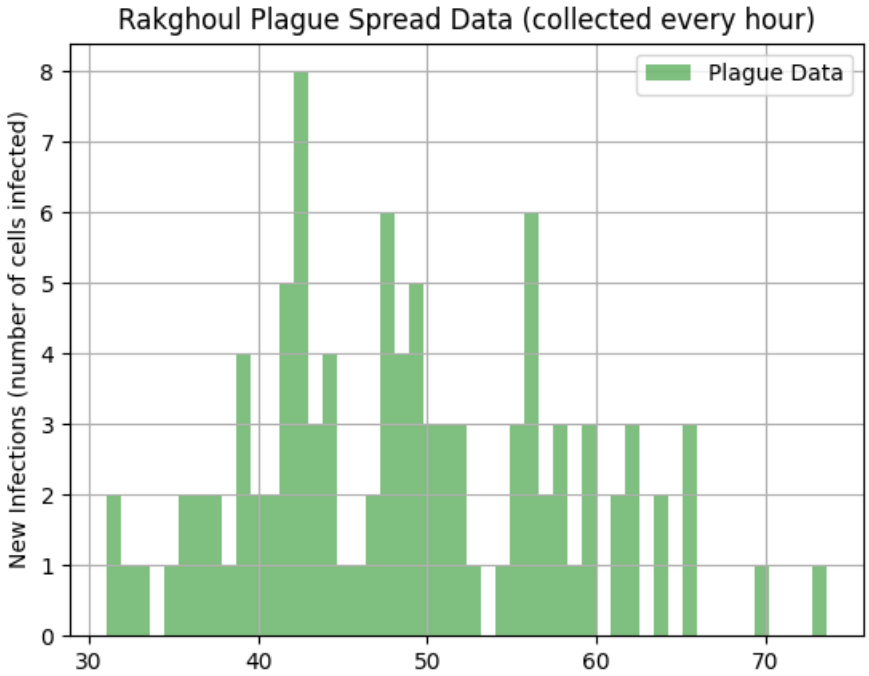

Figure 1: Histogram showing the number of newly infect cells per hour for N=100 hours.

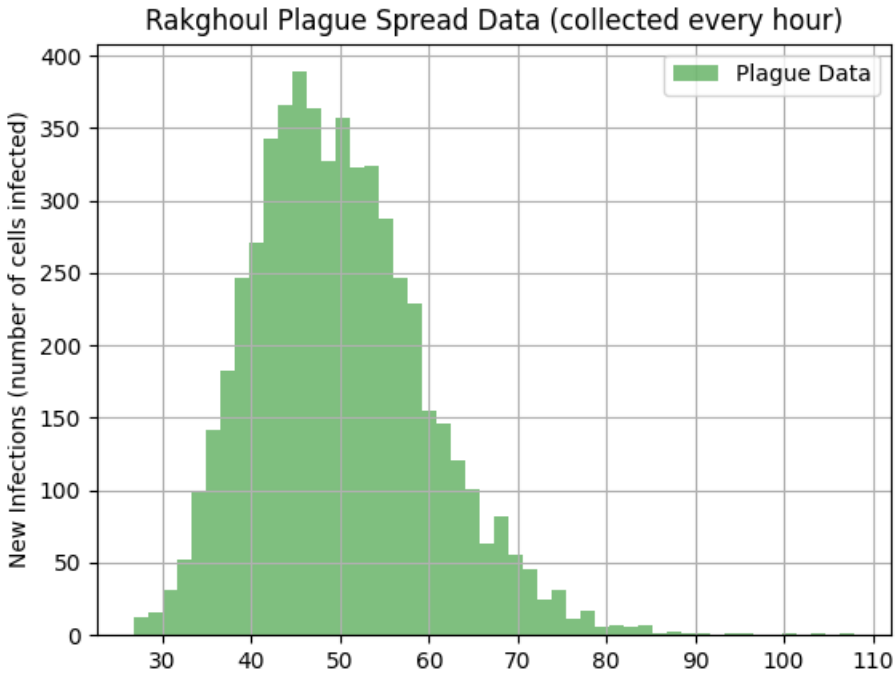

Figure 2: Histogram showing the number of newly infect cells per hour for N=5,500 hours.

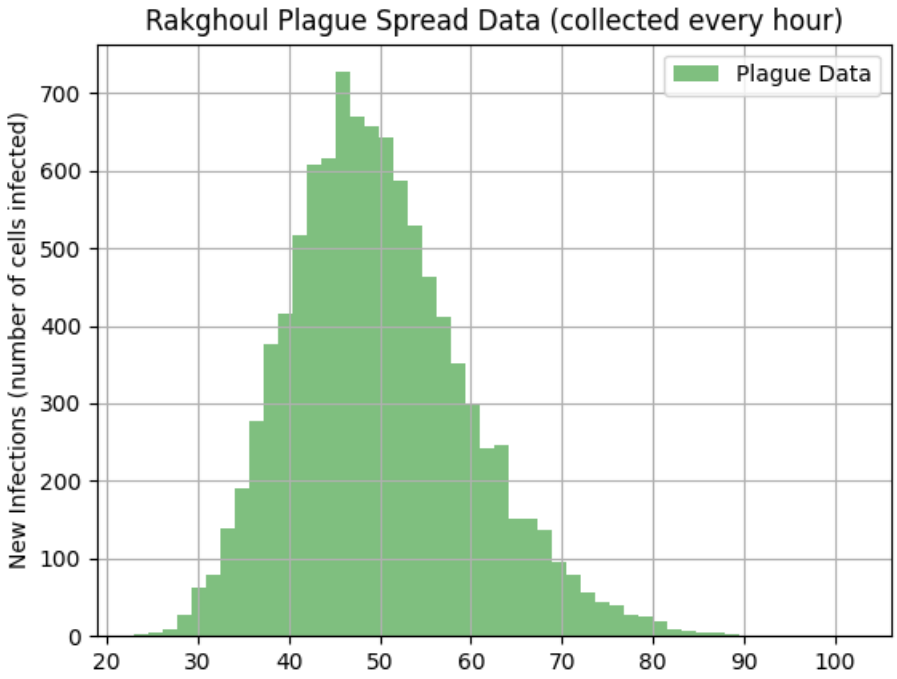

Figure 3: Histogram showing the number of newly infect cells per hour for N=10,000 hours.