Modelling COVID-19

Physics 132 – Spring 2020 Remote Instruction

Objectives:

By the end of this lab, you should be able to…

Complete a full analysis chain on a set of data including linearization, fitting, and Monte Carlo error propagation. You will then use your result to build a mathematical model which you will analyze to describe the effects of various parameters.

The Experiment

This lab works a bit differently from all your previous labs. In this lab, you will be taking publicly available data and performing your own analysis on it. While this change is definitely being done due to the current pandemic, this process is also something that is commonly done in science. The National Science Foundation (NSF) strives to make all data from research that it funds public as does the National Institutes of Health (NIH): the logic being that since the public paid for it the public should see what it paid for). Some of the data available from NIH can be found in this database.

The Data

The data for this lab are the number of confirmed COVID-19 cases in the United States. The data come from two different sources, the World Health Organization (WHO) and the US Centers for Disease Control (CDC) as compiled by the Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The data are provided in an Excel spreadsheet with three columns: date, WHO count, and CDC count.

The Analysis

Different Sources Yield Different Numbers

You will notice that at some point the number of counts from the WHO and the CDC are different.

1. I want you to speculate what may be the cause of this difference. Do some research. See if you can figure out why they are different. Be sure to cite any sources that you use.

   We want to see what you find and figure out. We suspect that, frankly, you will learn more than we might as you are more interested and more expert-like in the life-sciences than we are! Basically, any coherent argument justified by research from reliable sources (CDC, WHO, journals, etc.) is acceptable.

2. Think back to the first lab on the library, do you think that these are systematic or statistical uncertainties or some combination thereof? Justify your response.

3. In your model, how will you define the number of cases each day? How will you define the uncertainty on that number? Justify your choice.

   There are multiple solutions to this question. Just like in a real analysis, there is no single right answer. Just make sure you have a clear, consistent method for how you will define the number of cases each day and how you will define the uncertainty on that number.

Plot the Data

4. Make a plot of the number of cases per day, however you defined it.

Your plot should have error bars consistent with their definition from before. Also, both axes should be clearly labelled. Also, *on the figure itself* there should be a short description of the method of calculating the number of cases. An important good practice for scientific figures is that they should stand alone: all the information someone needs to understand the plot should be in the figure. The reasons for this are two-fold. First, there is so much scientific literature nowadays that a single person cannot possibly read all of it. Therefore, people tend to skim, which involves looking at the figures first (and possibly only!). The second reason for stand-alone figures is that people often take figures from the literature and put them in their own talks for comparison etc. and you want your figure to be interpretable under such circumstances.

5. What general shape is your data (linear, logarithmic, …)? A one-word answer here is sufficient.

Linearize the Data

Hopefully, your answer to question 5 was exponential! The data follow a shape that is
$$N = Ca^{Kt}$$
Where N is the number of cases, t is the time in days. Meanwhile, C, K, and a are constants. If you took calculus, you will recall that $y = e^x$ with $e = 2.71 \dots$ has some nice properties, namely that the derivative of e^x is e^x:
$$\frac{dy}{dx}(e^x) = e^x$$
For this reason, scientists tend to use e as the base for our exponentials. Meaning we will model our data as
$$N = Ce^{Kt}$$
Where C and K are constants to be determined from the data.

6. If we are modeling our data as $N = Ce^{Kt}$, what variables should we plot on each axis to ensure that we get a line?

7. The uncertainties. There is actually a formula to determine the uncertainty in your new y-variable:
$$\sigma_{y_{new}} = \frac{\sigma_{y_{old}}}{y}$$
Use this formula to determine the uncertainty in your new y-variable and make the linearized plot. Again, make sure that your plot can stand alone: all axis labels should be clear etc.

Building a Model

The linearized data clearly has several different sections representing different phases of the epidemic. This is another reason for linearizing data: you can see effects which are hard to observe in the raw plot you made for question 4. Towards the end, there is a linear section.

8. We are going to fit this linear section to build our model. As always in science there is a bit of a judgment call here: where do you begin fitting? There are multiple answers, just make a call and justify it.

9. Fit your linear section using the LINEST() function you learned in the last lab. What is the slope and intercept, including errors, of your fit?
   Hint: LINEST() can behave weird with dates. To solve this problem, add another column that is "the number of days since the beginning of the fitted region." I.e. the first day will be 0, the next 1, etc. Use these numbers, instead of your dates, for your x-values and the y-values and uncertainties from the previous part.

10. Add a trend-*line* to the linearized plot. The easiest way to do that is to use the slope and intercept from part 9, add a column to your spreadsheet with the values of your fit for each day and then add that column to your graph as a line.

11. Now, convert back. You have a fit $y = mt + b$. Undo your linearization procedure to get $N = Ce^{Kt}$. How do m and b map to C and K?

12. Interpret these numbers. What does C represent? What about K?
    Hint: For K, if t goes up by $1/K$ days, how much does N increase?

13. What are the units of K?
    Hint: In general, the quantity in an exponent, *must* be unit-less.

Use the Model to Predict the Future

The data ends on 25 March as that was when this lab was put together. We did not have data past that date at that time. However, you are now, comparatively, in the future.

14. Use your model to predict the number of cases *today*. Be sure to specify what today is so your TA knows.

15. Determine the uncertainty on your prediction; build a spreadsheet to run the Monte Carlo error propagation technique from the first lab. Remember to use the spreadsheet function RAND() to generate a random number and function NORM.INV() to generate a value normally distributed around its average value.

16. Compare your prediction with [the current CDC value](#) and the [WHO](#). For the WHO, you will need to click today's *Situation Report* and scroll through the pdf to the United States in Table 2. Remember, you should not hit it exactly. If the measured value is within the error bars, that is a great prediction.

17. Predict the number of cases next week with uncertainties. Again, write down what day you are looking at.

Understanding Your Model and What Exponential Growth Means

The quantity K can be thought of as the number of people per day you come into contact with E, times the probability of transmitting the disease p: $K = E \cdot p$. As you know, we are all now encouraged to practice physical distancing. Say you reduce the probability of transmission, and therefore K by *only 5%*, i.e. multiply K by 0.95:

18. If we had done this from the first day of your fit, how many fewer cases would there have been today? The central value is sufficient here, i.e. no need for uncertainties.

19. Given that the death rate is about 2.3%[1], how many fewer people would have died? Again, the central value is sufficient.

20. Say we all took more measures and decreased K from your measured value by 10% now. How many new cases would there be next week in comparison to your prediction from question 17?
    Hint: You are re-starting your prediction from today's numbers!

Applying your Model to Epidemiology

The constant K you measured is related to a quantity in epidemiology, called the basic reproduction number R_0 (which you may have heard about in the news), is the number of people that a given person is likely to infect. A good diagram of this is shown below for Ebola with $R_0 = 2$ and the 2002 SARS outbreak with $R_0 = 4$. A $R_0 > 1$ will continue to spread (the number of people continues to grow), meanwhile $R_0 < 1$ results in the number of cases declining. Specifically, to a first approximation (meaning K only appears to the first power),

$$R_0 = 1 + K(\tau_E + \tau_I)$$

where τ_E is how long you are infected, not showing symptoms, and NOT able to infect others, while τ_I is how long you are infected, not showing symptoms, but able to infect others (the idea being that once you are infected, you probably don't interact with many people and therefore don't interact with people). The current estimate for the sum of these two parameters for COVID-19 is $\tau_E + \tau_I = 5.1^{+0.7}_{-0.6}$ days[2].

21. Estimate R_0. Create a spreadsheet to propagate the uncertainties. You can assume the uncertainties on $\tau_E + \tau_I$ are symmetric at 0.7.

22. Compare to currently published results. Be sure to cite your source.

---

[1] Onder G, Rezza G, Brusaferro S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. JAMA. Published online March 23, 2020. https://doi:10.1001/jama.2020.4683.
[2] Lauer SA, Grantz KH, Bi Q, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med. 2020; [Epub ahead of print 10 March 2020]. doi: https://doi.org/10.7326/M20-0504.
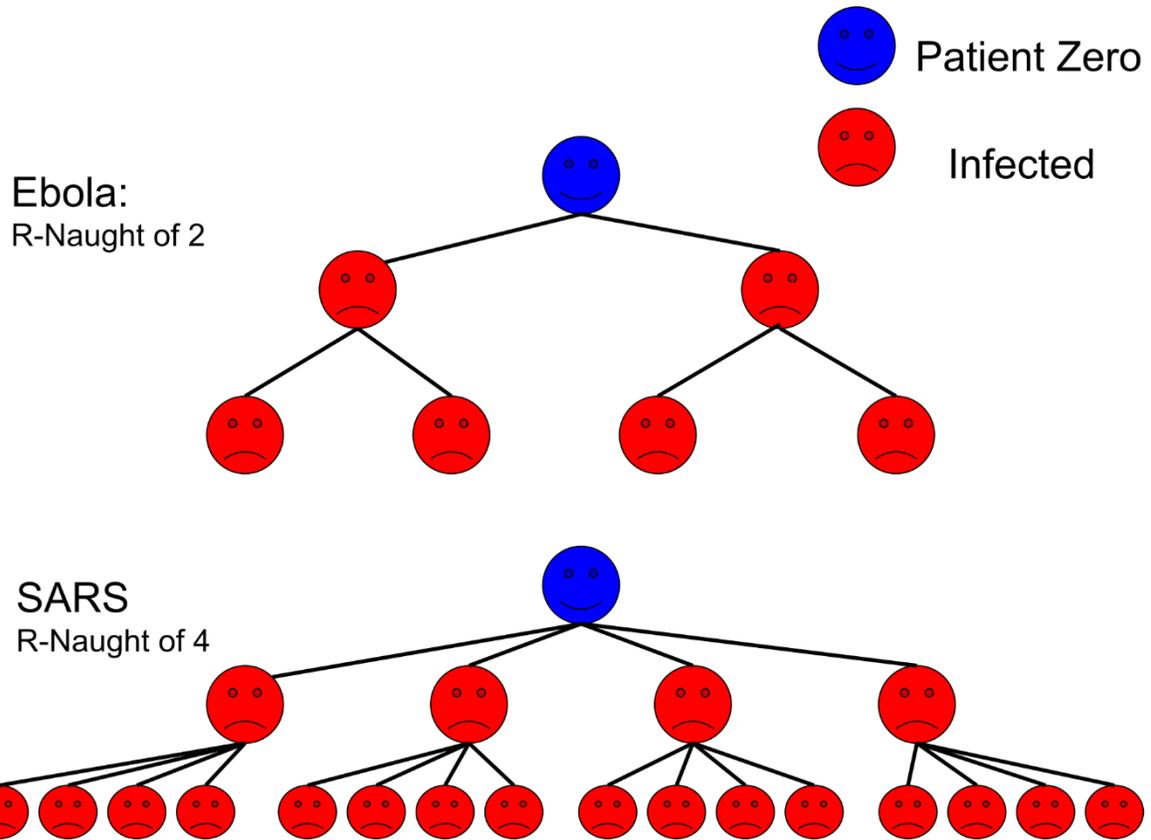
Figure 1 "R Naught Ebola and Flu Diagram" KieraCampbell. CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0). https://upload.wikimedia.org/wikipedia/commons/f/f4/R_Naught_Ebola_and_Flu_Diagram.svg.