Mathematical & Computational Modeling of Infectious Diseases **Jason Hunter**
Homework 4 **2025**

Instructions:

- Code for Homework 4

---

1. The goal of this problem is to try out some of the methods we developed in class to estimate R_0 or R_t from data. You'll also have a chance to refresh yourself on confidence intervals. **What we know about Bison/Ralphie Unexplained Hiccups disease:**

   - BRUH disease affects bison like Ralphie.
   - It is non-fatal, and does not affect mortality.
   - Diagnosed via sporadic symptoms — mostly hiccups and bad breath.
   - There are 100,000 bison in the herd
   - Typical Bison lifespan in this herd is 100 weeks.
   - Typical infection lasts 2 weeks, and a separate study found duration of infection exponentially distributed.

   Weekly Incidence Data

   - Weekly new case counts were recorded for 10 years, which you can find on Canvas as **all_weeks.csv**.
   - Ecologists believe they are identifying only 10% of cases due to lack of funds.
   - This 10% ascertainment is an approximation — varies from week to week.

   Prevalence and Seroprevalence Studies

   - Ted Turner paid for a prevalence study to be done. A team of researchers went out into the field at night dressed in bison disguises, and subjected 1000 bison to tickling — a decent way to see if they have hiccups. Only 7 had hiccups.
   - The estate of Buffalo Bill paid for a seroprevalence study to be done. They took blood samples from 1000 randomly chosen bison and found that 517 had BRUH antibodies.

   a. Estimate R_0 by examining the period of exponential growth (Method 1, Week 9). Be sure to show your work and plots as relevant. In the process, look up the 95% confidence interval associated with estimating a slope from data points, and use the slope's confidence interval to provide a confidence interval for your R_0 estimate.

   Solution: We can estimate R_0 from epidemic growth data. Under the SIR model with natural death rate μ, we have the equation:

$$\dot{I} = \beta\frac{SI}{N} - (\gamma + \mu)I$$

   At the beginning of an epidemic, when $S \approx N$, this can be simplified to:

$$\dot{I} \approx (\beta - (\gamma + \mu))I \to I(t) \approx I(0)e^{(\beta-(\gamma+\mu))t} = I(0)e^{(R_0-1)(\gamma+\mu)t}$$

So exponential growth can be expected at the rate:

$$(R_0 - 1)(\gamma + \mu)$$

From the data, we can estimate the exponential growth rate by plotting $log[I(t)]$, and getting the slope of the line. From the slope, which is given by:

$$m = (R_0 - 1)(\gamma + \mu)$$

R_0 can be solved for as:

$$R_0 \approx \frac{\hat{m}}{\gamma + \mu} + 1$$

In code, this can be done by performing linear regression on 'log(I)' vs 'time', and then getting the slope and confidence intervals from the result.

We can use the data from 'all_weeks.csv', and values of $\frac{1}{2}$ for γ, and $\frac{1}{100}$ for μ as parameters.

Using a sliding window approach, maximizing R^2 of $\log I$ vs week, the best window selected is from weeks 5 - 14.

Using an OLS fit of $\log I$ on week over weeks 5 - 14 gives $\hat{m} = 0.4551 \text{ wk}^{-1}$ with 95% confidence interval $(0.4199, 0.4904)$.

Estimating R_0 with these values results in the following:

$$R_0 \approx \frac{\hat{m}}{\gamma + \mu} + 1 = \frac{0.4551}{0.51} + 1 = 1.892,$$
$$95\% \text{ CI: } \left(\frac{0.4199}{0.51} + 1, \ \frac{0.4904}{0.51} + 1 \right) = (1.823, \ 1.962).$$

So the estimate for R_0 is approximately 1.89 with a 95% confidence interval of $(1.82, \ 1.96)$.

The plots are shown on the following page:

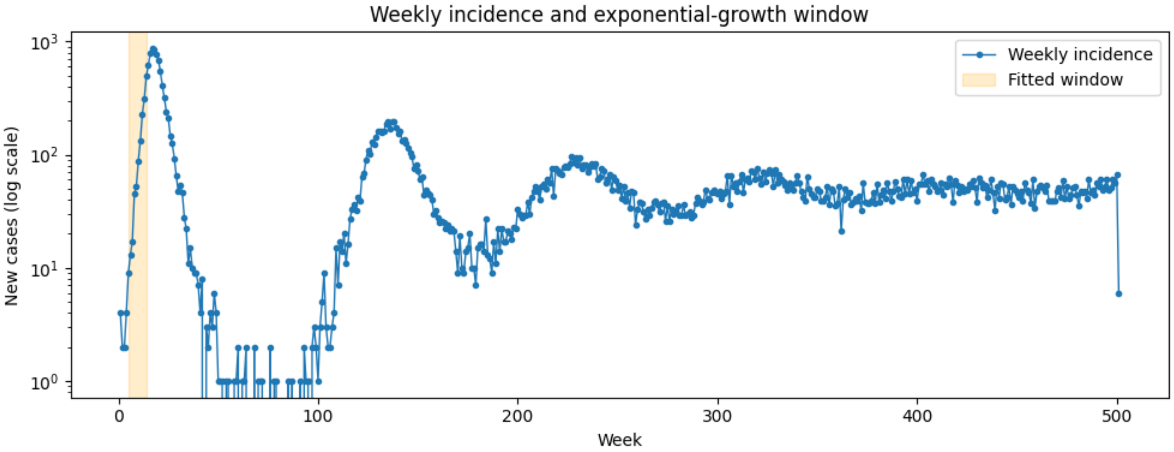

Figure 1: Weekly incidence on a log scale with fitted window (weeks 5–14) shaded.

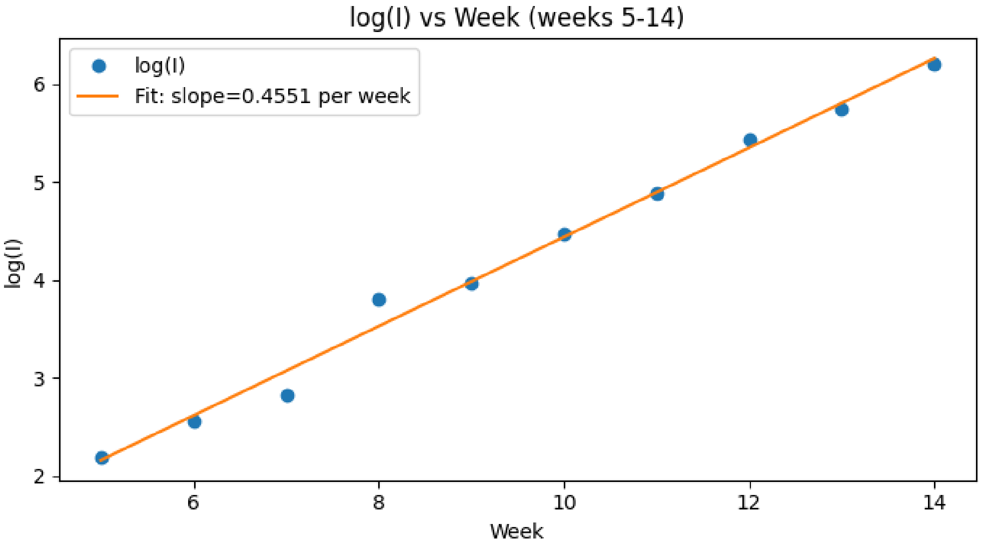

Figure 2: $\log I$ vs week with OLS fit; slope $\hat{m} = 0.4551$ per week.

b. Estimate R_0 by utilizing the prevalence *or* seroprevalence data. (Method 2 or 4, Week 9). Be sure to show your work and plots as relevant. Write down (or look up) the 95% confidence interval for the prevalence/seroprevalence estimate, and use it to provide a confidence interval for R_0.

Solution: Using the seroprevalence data, we can estimate R_0 using the formula:

$$R_0 \approx \frac{1}{1-\hat{\pi}}$$

where $\hat{\pi}$ is the estimated seroprevalence.

From the seroprevalence study, we have 517 positives out of 1000 samples, so the estimated seroprevalence is:

$$\text{seroprevalence} = \frac{517}{1000} = 0.517$$

From the notes, at steady state we expect:

$$S_{eq} = \frac{1}{R_0}$$

From this, a model based estimate of R_0 would be:

$$R_0 = \frac{1}{1-\text{seroprevalence}} = \frac{1}{1-0.517} = 2.066$$

To get a 95% confidence interval for the seroprevalence estimate, we can estimate the standard error for a proportion:

$$SE = \sqrt{\frac{0.517(1-0.517)}{1000}} = 0.0158$$

Using this, we can get a 95% confidence interval for the seroprevalence:

$$0.517 \pm 1.96 \times 0.0158 = (0.486,\ 0.548)$$

From here, we can get a confidence interval for R_0:

$$R_0 = \frac{1}{1-0.548}, \frac{1}{1-0.486} = (2.202,\ 1.946)$$

So the estimate for R_0 from the seroprevalence data is approximately 2.07 with a 95% confidence interval of $(1.95,\ 2.20)$.

This is shown in the figure below:

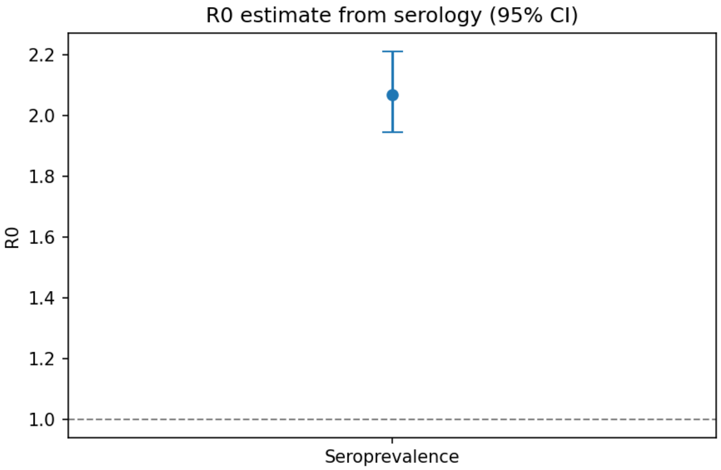

Figure 3: Estimate of R_0 from seroprevalence data with 95% confidence interval.

c. (Grad / EC) Estimate R_0 a third way from the same data.

d. Compare your estimates, the uncertainty associated with each, and discuss what might cause them to be different.

Solution: The two different estimates of R_0 are relatively close to each other.

The first estimate from the exponential growth method was about 1.89, with a 95% CI of $(1.82,\ 1.96)$.

The second estimate from the given information about seroprevalence ended up being about 2.07, with a 95% CI of $(1.95,\ 2.20)$.

The estimates are pretty close, there is a small overlap between the two confidence intervals upper and lower bounds.
The reason these are different is best explained by their varying data sources.
For the first method, incidence data was used, which requires accurate reporting of new cases. It was specified that only about 10% of cases are actually being identified, so the incidence data is fairly unrepresentative of the true number of cases.
True cases are likely to be modeled more accurately by the seroprevalence data, which looks at the number of bison who have antibodies.

e. (EC for all) Estimate R_t using Method 5.

2. The goal of this problem is to get some simple practice with sensitivity and specificity, and get a little more familiar with confidence intervals too.

   Suppose we've got a diagnostic with sensitivity 0.90 and specificity 0.98.

   a. Maria Lara conducts a prevalence study with the above diagnostic. She samples 100 people and gets 39 positives. What is your estimate of the prevalence after correcting for the sensitivity and specificity?

   Solution:
   If we let θ be the true prevalence. The number of true positives should be:

   $$\theta \times se + (1-\theta)(1-sp) = \phi$$

   Plugging in what we have:

   $$\begin{aligned} 0.90(\theta) + (1-\theta)(0.02) &= 0.39 \\ 0.90(\theta) + 0.02 - 0.02(\theta) &= 0.39 \\ 0.88(\theta) &= 0.37 \\ \theta &= 0.4205 \end{aligned}$$

   Thus, the corrected estimate of prevalence is approximately 42.05%.

   b. Write down a 95% confidence interval for your corrected estimate.

   Solution:
   To get a 95% confidence interval for the corrected prevalence estimate, we first need to get a CI for the given prevalence estimate.
   Using the standard error for a proportion:

   $$SE = \sqrt{\frac{0.39(1-0.39)}{100}} = 0.0487$$

   Using this, we can get a CI for the given prevalence:

   $$0.39 \pm 1.96 \times 0.0487 = (0.294,\ 0.486)$$

   From here, we can get another CI for the corrected prevalence by solving for θ at the bounds of the given prevalence CI:

$$\begin{aligned}\text{Lower bound: } & 0.90(\theta) + (1-\theta)(0.02) = 0.294 \\ & 0.88(\theta) = 0.274 \\ & \theta = 0.3114 \\ \text{Upper bound: } & 0.90(\theta) + (1-\theta)(0.02) = 0.486 \\ & 0.88(\theta) = 0.466 \\ & \theta = 0.5295\end{aligned}$$

So the 95% confidence interval for the corrected prevalence estimate is approximately $(31.14\%, 52.95\%)$.

c. Trying to be helpful, Burt Q. Losis conducts a second prevalence study in the same population and finds 18 positives out of 50 samples. Again estimate the prevalence and a 95% confidence interval.

Solution:
Using the exact same steps from before:

$$\begin{aligned}0.90(\theta) + (1-\theta)(0.02) &= 0.36 \\ 0.88(\theta) &= 0.34 \\ \theta &= 0.3864\end{aligned}$$

Thus, the corrected estimate of prevalence is approximately 38.64%.
To get a 95% confidence interval for the corrected prevalence estimate, we first need to get a confidence interval for the given prevalence estimate.
Using the standard error for a proportion:

$$SE = \sqrt{\frac{0.36(1-0.36)}{50}} = 0.0679$$

Using this, we can get a 95% CI for the given prevalence:

$$0.36 \pm 1.96 \times 0.0679 = (0.227, 0.493)$$

From here, we can get a confidence interval for the corrected prevalence by solving for θ at the bounds of the given prevalence CI:

$$\begin{aligned}\text{Lower bound: } & 0.90(\theta) + (1-\theta)(0.02) = 0.227\\ & 0.88(\theta) = 0.207\\ & \theta = 0.2352\\ \text{Upper bound: } & 0.90(\theta) + (1-\theta)(0.02) = 0.493\\ & 0.88(\theta) = 0.473\\ & \theta = 0.5375\end{aligned}$$

So the 95% CI for the corrected prevalence estimate is approximately $(23.52\%, 53.75\%)$.

d. Pool Burt's and Maria's data to get a third estimate of prevalence, and update your 95% confidence interval. How are your three estimates related? And, how are the widths of the three confidence intervals related?

Solution:
Combining the data, there are a total of 57 positives out of 150 samples.
Using this and plugging in as before:

$$\begin{aligned}0.90(\theta) + (1-\theta)(0.02) &= 0.38\\ 0.88(\theta) &= 0.36\\ \theta &= 0.4091\end{aligned}$$

The correct estimate when combining the data is about 40.91%.
Solving for the 95% confidence interval:

$$SE = \sqrt{\frac{0.38(1-0.38)}{150}} = 0.0396$$

So:

$$0.38 \pm 1.96 \times 0.0396 = (0.302,\ 0.458)$$

Now solving for θ at these bounds:

$$\begin{aligned}
\text{Lower bound: } & 0.90(\theta) + (1-\theta)(0.02) = 0.302 \\
& 0.88(\theta) = 0.282 \\
& \theta = 0.3205 \\
\text{Upper bound: } & 0.90(\theta) + (1-\theta)(0.02) = 0.458 \\
& 0.88(\theta) = 0.438 \\
& \theta = 0.4977
\end{aligned}$$

For this new estimate, the CI is $(32.05\%, 49.77\%)$.
The three estimates are 42.05%, 38.64%, and 40.91%.
The widths of these three estimate's confidence intervals are 21.81%, 30.23%, and 17.72%, respectively.
The main thing I notice is that the width of the CI for the combined data is the tightest. As more data is collected, the CI should be narrower.

e. (Grad / EC) You test yourself. Positive! What is your best guess of the probability that you are *actually* positive?

3. The goal of this problem is to learn about how sensitivity and specificity arise from calibration data, i.e. from positive and negative controls. For this problem, you will need to read in three .csv files to access the data they contain:

   - **HW4_Q3_neg.csv**: The assay values associated with a set of negative controls.
   - **HW4_Q3_pos.csv**: The assay values associated with a set of positive controls.
   - **HW4_Q3_data.csv**: The assay values associated with your prevalence study in the population.

   a. Read in the data and produce a tall, skinny plot with three columns of data: the negative controls (red), the positive controls (black), and the data from the field (blue). Use jitter and transparency ("alpha") to allow us to see the distributions of the data.

   Solution: Plot is on the next page.

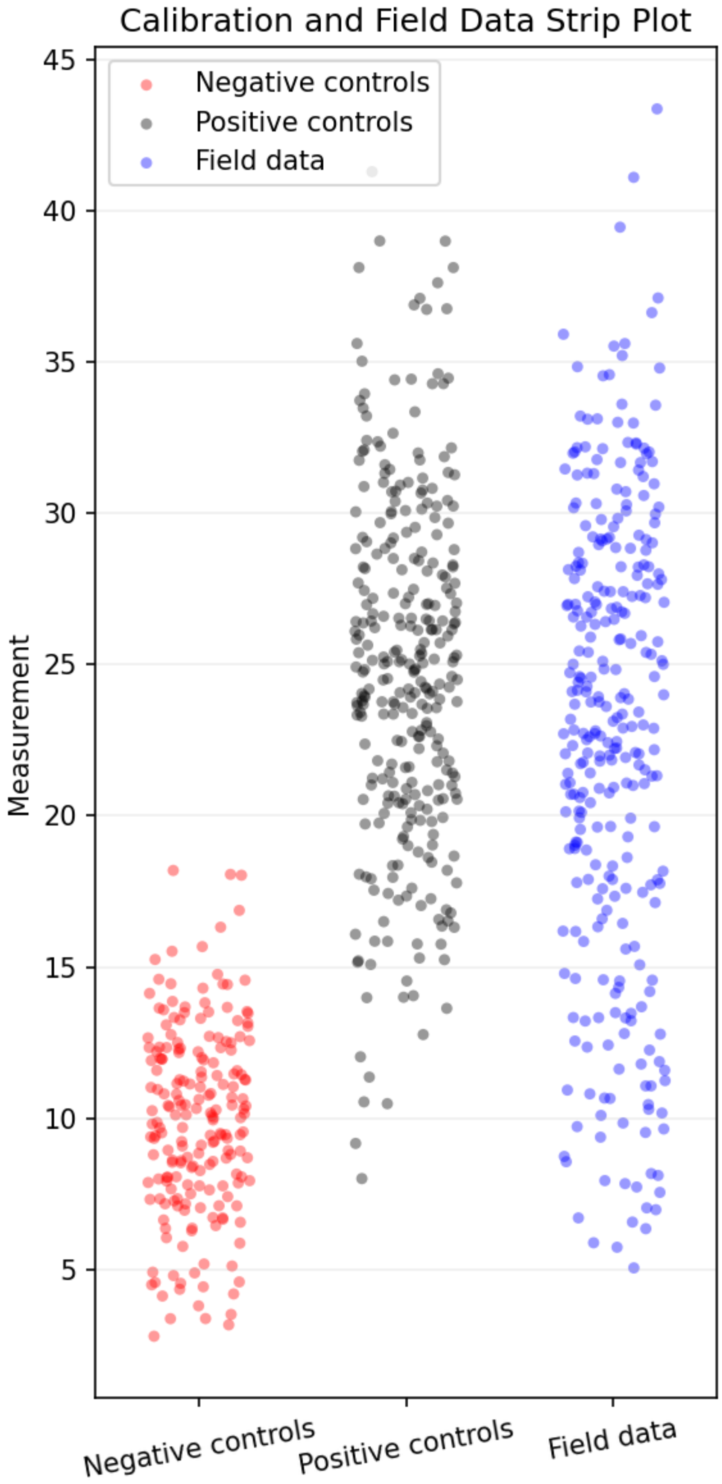

Figure 4: Distribution of the negative, positive, and field data assay values

b. Consider a cutoff c such that any assay values above c are to be called positive and any assay values below c are to be called negative. Then write four functions: $se(c)$, $sp(c)$, and $\hat{\phi}(c)$ and $\hat{\theta}(c)$. They should correspond to the sensitivity, the specificity, the raw prevalence in the field data, and the corrected prevalence in the field data. What value of c corresponds to the "Youden" choice?

Solution: The value of c that maximizes the Youden index is the Youden choice.
Calculating this gives $c = 14.9197$. Visually, this is close to what I would have guessed myself.

c. (Grad / EC) By sweeping over various choices of c, plot a receiver operator curve, and place a point at the Youden choice. Create a second plot showing how $\hat{\theta}(c)$ varies, and again, place a point at the Youden choice.

d. Write 3-4 sentences reflecting on how the conclusions of a study might be affected by how one decides to choose the cutoff at which positives and negatives are called.

Solution: The choice of cutoff has a significant effect on the sensitivity and specificity calculated. If the cutoff is too low, more false positives will occur, which will lower the specificity. Conversely, when the cutoff is set too high, there are more false negative, and sensitivity decreases. Without a proper cutoff, the study's results may be inaccurate and misleading.