Mathematical & Computational Modeling of Infectious Diseases **D.B. Larremore**
Homework 4 **2025**

Instructions:

- Please turn in a single PDF file.
- Please share a link to your code, but do not attach the actual code.
- Handwritten math (scanned and included in a PDF) is fine, but please watch out for 10MB+ file sizes!

Name: Ben Aoki-Sherwood

Code: `https://github.com/aoki-sherwoodb/CSCI-5897-InfectiousDiseases`

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.

   From the population parameters provided, we estimate the death rate as $\mu = 1/100$ weeks^{-1}, and the recovery rate as $\gamma = 1/2$ weeks^{-1}. After inspecting the incidence data, I

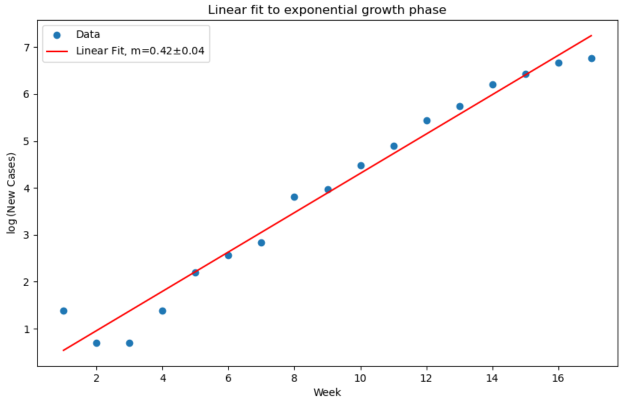

Figure 1: Linear fit to log-transformed incidence data over the exponential growth period (weeks 1-17). The slope estimate is 0.42 weeks^{-1} with a 95% confidence interval of (0.38, 0.46) weeks^{-1}.

estimate the exponential growth period as the period between week 0 and week 17, when the maximum case count occurs. The linear fit to the log-transformed incidence data over this period is shown below.

The confidence interval for a least-squares fit of a slope from points is calculated as $\hat{m} \pm t_{\alpha/2,n-2} \cdot SE(\hat{m})$**, where** $t_{\alpha/2,n-2}$ **is the critical value from the t-distribution with** $n-2$ **degrees of freedom, and** $SE(\hat{m})$ **is the standard error of the slope estimate. Plugging in** $\hat{m} = 0.42$ **weeks**$^{-1}$**,** $t_{0.025,14} = 2.13$**, and** $SE(\hat{m}) = 0.018$**, we find a 95% confidence interval of (0.38, 0.46) weeks**$^{-1}$ **for the slope.**

Thus we estimate R_0 **as**

$$\begin{aligned} R_0 &= 1 + \frac{\hat{m}}{\mu + \gamma} \\ &= 1 + \frac{0.42}{0.01 + 0.5} \\ &= 1.82, \end{aligned}$$

with a 95% **confidence interval of** $(1.75, 1.90)$**.**

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.

Based on Ted Turner's prevalence study, we estimate the equilibrium incidence as $i_{eq} = 7/1000 = 0.007$.

Plugging in the same values for μ and γ yields the estimate

$$R_0 = \frac{1}{1 - i_{eq}(\gamma/\mu + 1)} = \frac{1}{1 - 0.007(50 + 1)} \approx 1.56.$$

We can compute a confidence interval for the prevalence estimate using the Clopper-Pearson method for a binomial proportion [1]. This yields a 95% confidence interval for the prevalence of $(0.0028, 0.0144)$, which translates to a 95% confidence interval for R_0 of $(1.17, 3.74)$.

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

I already used the prevalence method, so I will use the seroprevalence data. This method yields an approximation of

$$R_0 = \frac{1}{1 - \textbf{seroprevalence}} = \frac{1}{1 - 0.517} \approx 2.07.$$

Applying the same binomial confidence interval method as before results in a 95% confidence interval for R_0 of $(1.94, 2.21)$.

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

My estimates of R_0 were 1.82 (1.75, 1.90) from the exponential growth method, 1.56 (1.17, 3.74) from the prevalence method, and 2.07 (1.94, 2.21) from the seroprevalence method. All of the R_0 estimates are greater than 1, but the estimates from the prevalence and seroprevalence methods differ by about 33%. The uncertainty for the prevalence-based estimate is much larger than the other two because the small number of positives (7 of 1000). The seroprevalence-based estimate has smaller uncertainty because of the larger number of positives, but it may be biased by the sensitivity and specificity of the test used, and may miss individuals who lost antibodies due to seroreversion.

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

[1] https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

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?

   We estimate prevalence as

   $$\hat{\theta} = \frac{n^+/n - (1 - sp)}{se + sp - 1} = \frac{0.39 - 0.02}{0.90 + 0.98 - 1} \approx 0.420.$$

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

   We again use the Clopper-Pearson method to compute a binomial confidence interval for the proportion $\hat{\phi} = 0.39$, which yields the interval $(0.294, 0.493)$. Propagating this interval through the prevalence-correction formula yields a 95% confidence interval for prevalence of $(0.311, 0.517)$.

   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.

   The 95% confidence interval for the positive proportion is $(0.229, 0.508)$ and the corrected prevalence point estimate is

   $$\hat{\theta} = \frac{0.36 - 0.02}{0.90 + 0.98 - 1} \approx 0.386.$$

   Thus the 95% confidence interval for prevalence is $(0.238, 0.555)$.

   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?

   The pooled data has $39 + 18 = 57$ positives out of $100 + 50 = 150$ samples, yielding a positive proportion of 0.38. This translates to a corrected prevalence estimate of

   $$\hat{\theta} = \frac{0.38 - 0.02}{0.90 + 0.98 - 1} \approx 0.409,$$

   with a confidence interval of $(0.321, 0.503)$.

   The pooled prevalence estimate lies between the two individual estimates, and the confidence interval is narrower than either of the individual intervals because of the larger

pooled sample size. Similarly, the width of the confidence interval from Maria's data is smaller than that from Burt's data because of her larger sample size.

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

Given a positive test, we can compute the positive predictive value (PPV) to determine the probability of a true positive. Computing the PPV requires an estimate of the prevalence, so we will use the pooled estimate of $\hat{\theta} = 0.409$ from the previous part. Using Bayes' rule, we have that

$$\begin{aligned} PPV = P(TP|\text{pos. test}) &= \frac{P(\text{pos. test}|TP)P(TP)}{P(\text{pos. test})} \\ &= \frac{se \cdot \hat{\theta}}{se \cdot \hat{\theta} + (1 - sp)(1 - \hat{\theta})} \\ &= \frac{0.90 \cdot 0.409}{0.90 \cdot 0.409 + 0.02 \cdot (1 - 0.409)} \\ &= 0.969 \end{aligned}$$

is the probability that I am actually positive given a positive test result.

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.

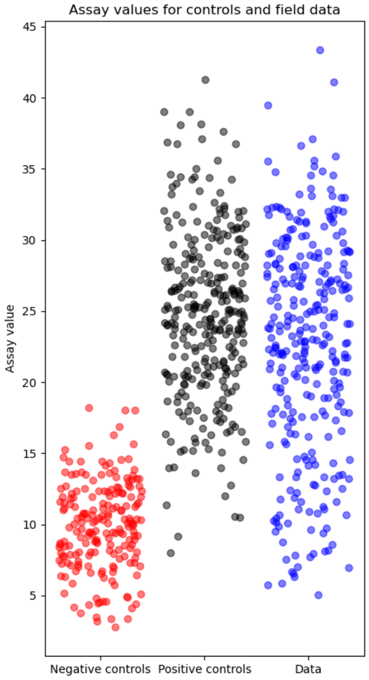

Figure 2: Assay values for negative controls (red), positive controls (black), and field data (blue).

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?

See code repository for functions. For a cutoff value c, define $TP(c)$ as the number of positive controls with assay values above c and $TN(c)$ as the number of negative controls with assay values below c.

Define $FN(c)$ = total positive controls $- TP(c)$ and $FP(c)$ = total negative controls $- TN(c)$.

Then the sensitivity and specificity functions are defined as

$$se(c) = \frac{TP(C)}{TP(C) + FN(C)}$$

and

$$sp(c) = \frac{TN(c)}{TN(c) + FP(c)}.$$

Next, define $n^+(c)$ as the number of field data assay values above c and n as the total number of field data points.

Then the raw prevalence function is defined as $\hat{\phi}(c) = \frac{n^+(c)}{n}$ and the corrected prevalence function is defined as

$$\hat{\theta}(c) = \frac{\hat{\phi}(c) - (1 - sp(c))}{se(c) + sp(c) - 1}.$$

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.

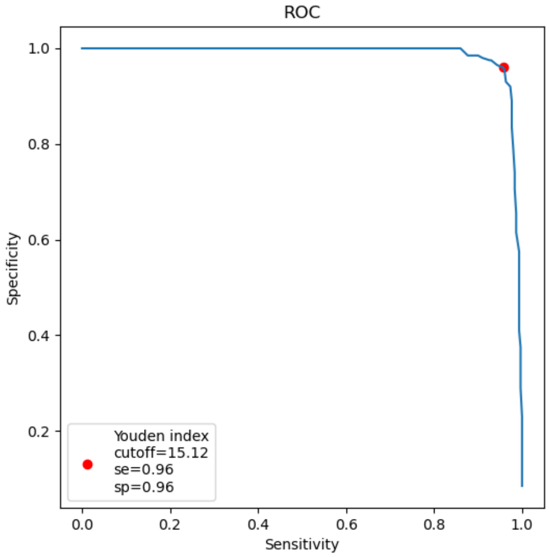

Figure 3: ROC curve for the assay data showing specificity vs. sensitivity at various cutoff values. The Youden choice is marked with a red dot at (0.96, 0.96).

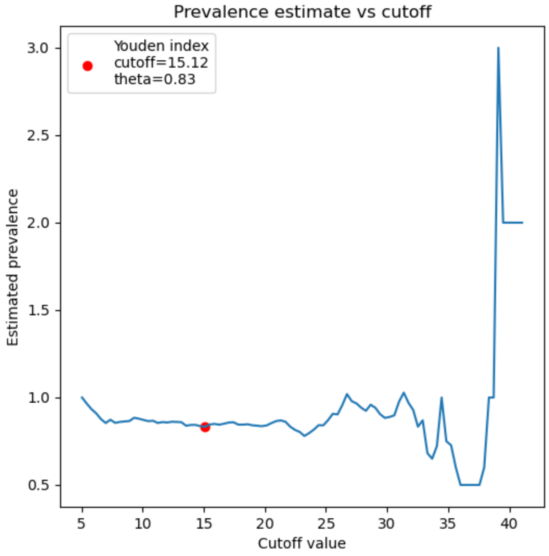

Figure 4: Corrected prevalence estimate vs. assay cutoff value. The Youden choice is marked with a red point at (15.09, 0.83).

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.

The ROC curve shows that there is a tradeoff between sensitivity and specificity based on cutoff choice; as the cutoff increases, specificity decreases while sensitivity increases. This tradeoff affects the corrected prevalence estimate, which varies significantly with cutoff choice. While the prevalence estimate at the Youden choice of $c \approx 15$ is high (0.83), choosing a higher cutoff $c \approx 36$ results in a much lower prevalence estimate of 0.49.