Altaf Barelvi

I collaborated with Ben Aoki-Sherwood

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.

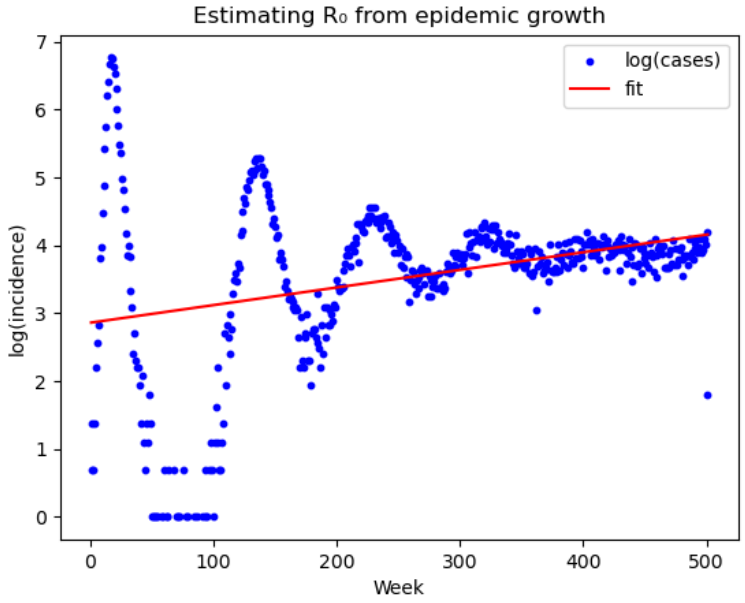

Figure 1: Exponential growth fit for estimating R_0.

Answer:

$$\text{CI}_{\text{low}} = \text{slope} - t_{\text{crit}} \times \text{SE}_{\text{slope}}$$
$$\text{CI}_{\text{high}} = \text{slope} + t_{\text{crit}} \times \text{SE}_{\text{slope}}$$

$$(\text{CI}_{\text{low}},\ \text{CI}_{\text{high}}) = \textbf{(0.0019,\ 0.0033)}$$

$$R_0^{95\%\,\text{CI}} = \left[1 + \frac{m_{\text{low}}}{\gamma + \mu},\ 1 + \frac{m_{\text{high}}}{\gamma + \mu}\right]$$

$$R_0^{95\%\,\text{CI}} = \left[1 + \frac{\text{CI}_{\text{low}}}{\gamma + \mu},\ 1 + \frac{\text{CI}_{\text{high}}}{\gamma + \mu}\right] = \textbf{(1.004,\ 1.007)}$$

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.

Answer:

$$n_{\text{total}} = 1000, \quad n_{\text{seropos}} = 517$$

$$\hat{P} = \frac{n_{\text{seropos}}}{n_{\text{total}}} = \frac{517}{1000} = 0.517$$

$$z = 1.96, \quad SE = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} = \sqrt{\frac{0.517(1-0.517)}{1000}} = 0.016$$

$$P_{\text{low}} = \hat{P} - z \times SE = 0.517 - 1.96(0.016) = 0.485$$

$$P_{\text{high}} = \hat{P} + z \times SE = 0.517 + 1.96(0.016) = 0.549$$

Thus, the 95% CI for seroprevalence is: $(0.485,\ 0.549)$

$$R_0^{95\%\,\text{CI}} = \left[\frac{1}{1-P_{\text{low}}},\ \frac{1}{1-P_{\text{high}}}\right] = \left[\frac{1}{1-0.485},\ \frac{1}{1-0.549}\right] = (1.94,\ 2.22)$$

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

Answer: The equilibrium prevalence study found 7 infected bison out of 1000, giving

$$i_{\text{eq}} = \frac{7}{1000} = 0.007.$$

$$\gamma = \frac{1}{2},$$

$$\mu = \frac{1}{100}.$$

The full equilibrium prevalence formula from Section 10.2 is

$$R_0 = \frac{1}{1 - i_{\text{eq}}\left(\frac{\gamma}{\mu} + 1\right)}.$$

$$R_0 = \frac{1}{1 - 0.007 \cdot 51} = 1.55.$$

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

Answer: The three methods give different R_0 estimates because they rely on different parts of the data and assumptions. The exponential-growth method gives an R_0 just above 1 with a very tight CI. The seroprevalence method gives a higher estimate (1.94-2.22) because it reflects how many bison were ever infected rather than short-term trends. The equilibrium–prevalence method gives an intermediate value (1.55) These differences arise because each method captures a different part of the epidemic and makes different simplifications.

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?

   Answer:

$$\hat{\theta} = \frac{\frac{n_{\text{pos}}}{n} - (1 - Sp)}{Se + Sp - 1}$$

$$\hat{\theta} = \frac{\frac{39}{100} - (1 - 0.98)}{0.9 + 0.98 - 1}$$

$$\hat{\theta} = 0.420$$

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

   Answer:

$$SE = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = \sqrt{\frac{0.420(1-0.420)}{100}}$$

$$SE = 0.0494 \qquad Z = 1.96$$

$$CI = \left[\hat{\theta} - Z \cdot SE,\ \hat{\theta} + Z \cdot SE\right]$$

$$CI = [0.324,\ 0.517]$$

   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.

   Answer:

$$\hat{\theta} = \frac{\frac{18}{50} - (1 - 0.98)}{0.9 + 0.98 - 1}$$

$$\hat{\theta} = 0.386$$

$$SE = \sqrt{\frac{0.386(1-0.386)}{50}}$$

$$SE = 0.0687$$

$$CI = [0.2514,\ 0.5213]$$

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?

Answer:

$$n_{\text{pos}} = 57 \qquad n = 150$$

$$\hat{\theta} = \frac{\frac{57}{150} - (1-0.98)}{0.9 + 0.98 - 1}$$

$$\hat{\theta} = 0.409$$

$$SE = 0.0695$$

$$CI = [0.273,\ 0.545]$$

The three estimates are related because we dont know which subset of the population pool burt collected overlapped with maria's collection. This would also bring into question which of the positive people from both pool overlapped. The confidence interval on the lower end are tailed towards pool burt's confidence interval. Both studies have similar confidence interval on the higher end and this is reflected by a similar number for the combined study.

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

Answer:

$$\text{PPV} = \frac{\text{Se}\,\theta}{\text{Se}\,\theta + (1-\text{Sp})(1-\theta)}.$$

Using sensitivity $\text{Se} = 0.90$, specificity $\text{Sp} = 0.98$, and the estimated prevalence $\theta = 0.409$,

$$\text{PPV} = \frac{0.90 \cdot 0.409}{0.90 \cdot 0.409 + 0.02 \cdot (1 - 0.409)} \approx 0.969.$$

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:

    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.

    Answer:

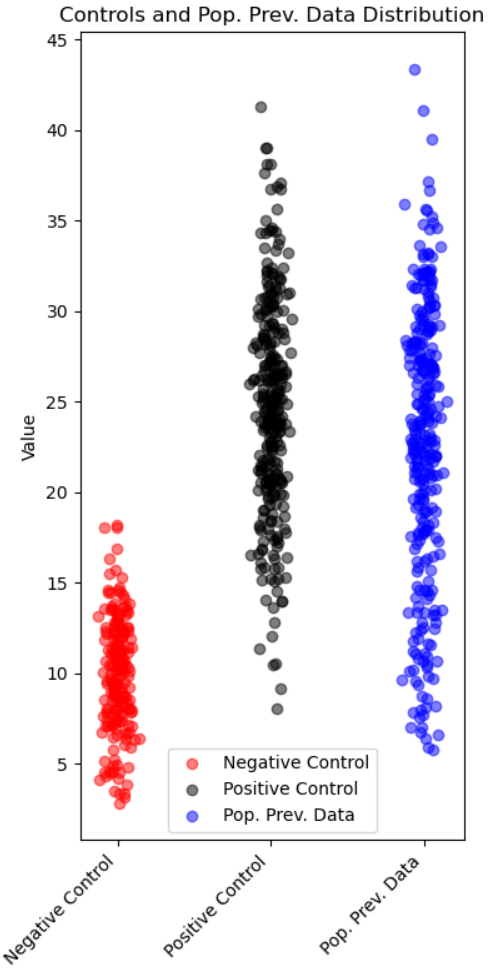

Figure 2: Distribution of assay values for three conditions: Negative control, positive control, and population prevalence data

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?

Answer: The optimal Youden cutoff was $c^* = 14.83$, yielding a sensitivity of $se = 0.96$, specificity of $sp = 0.96$, and a corrected prevalence estimate of $\hat{\theta} = 0.836$

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.

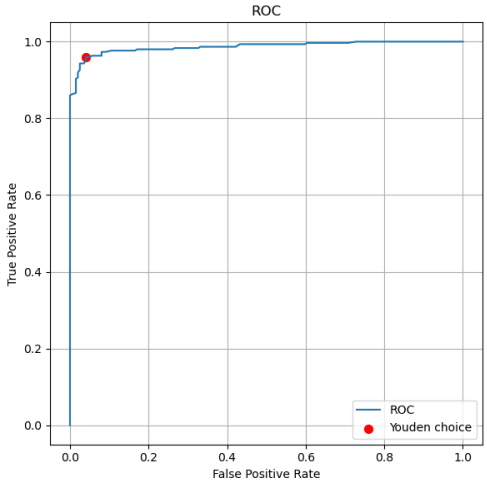

Figure 3: ROC curve

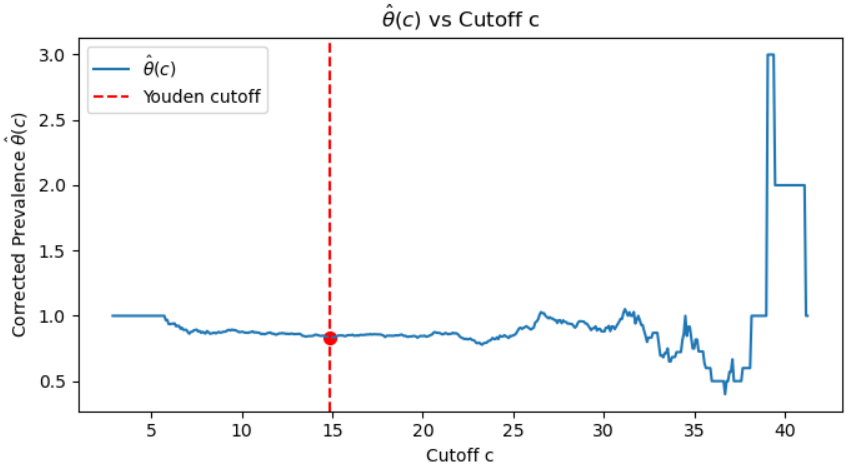

Figure 4: $\hat{\theta}$ with C cutoff

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.

Answer: The cutoff you choose affects how many results are called positive or negative. If the cutoff is too low, you'll get more positives, but many of them might be false. If it's too high, you might miss real cases. Because of this, the conclusions of a study can change a lot depending on where that cutoff is set.