Is this "population" free of infection?

Estimating CI's on proportions

- How confident can I be in my estimate? (e.g., 0 of 10 vs. 0 of 30)
- How different is the estimate of prevalence in two species, populations, times, … (quick and dirty)
- Skip the "simple" normal approximations
  - will always be a little wrong, sometimes nonsensical)
  - with modern stats packages, there is no need to resort to such a bad approximation

Estimating CI's on proportions

Use Wilson score interval (w/o continuity correction)…

$$CL = \frac{1}{1+\frac{1}{n}z^2}\left[\hat{p} + \frac{1}{2n}z^2 \pm \sqrt{\frac{1}{n}\hat{p}(1-\hat{p}) + \frac{1}{4n^2}z^2}\right]$$

where $z = 1 - \alpha/2 = 1.96$

it's ugly, but it works well

`binom.confint()` function in `binom` package

http://vassarstats.net/prop1.html

Adjusting prevalence estimates for imperfect tests

$$\phi_{True} = \frac{\phi_{Apparent} + \text{specificity} - 1}{\text{sensitivity} + \text{specificity} - 1}$$

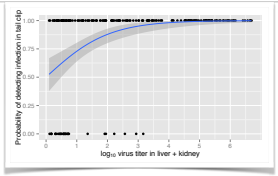
$$CL_{Adjusted} = \frac{CL_{Apparent} + \text{specificity} - 1}{\text{sensitivity} + \text{specificity} - 1}$$

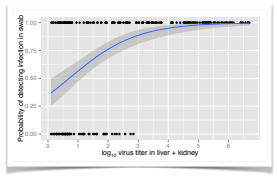
`epi.prev()` function in `epiR` package

Rogan W, Gladen B (1978). Estimating prevalence from results of a screening test. American Journal of Epidemiology 107: 71 - 76.

Detection varies with titer!

- We treat infections as binary (at least for microparasites)
- Virus titers vary by orders of magnitude
- The P(detect ranavirus) increases with titer





Take care in interpreting prevalence data

Just a snapshot in time

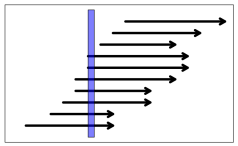
High incidence ≠ lots of disease

at least some individuals of many species are tolerant of RV

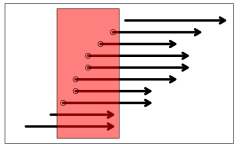
Low incidence ≠ lack of disease or impact

if individuals die or recover quickly, they will not be sampled and so will not be part of prevalence estimate

Take care in interpreting prevalence data

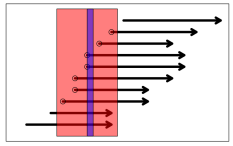


Prevalence — the proportion infected (or diseased) *at some time point*



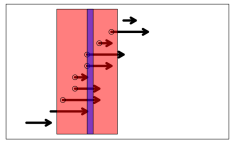
Incidence — the rate of *new* infections (or occurrence of disease) *over an interval*

Take care in interpreting prevalence data



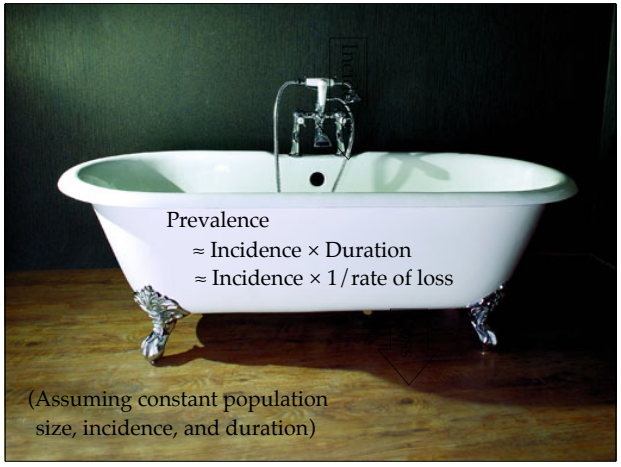
Scenario A:
Long-lasting infections (e.g., long time course, low mortality & recovery)

Prevalence = 7/10
Incidence = 7 (or 7/8 at risk)



Scenario B:
Short infections (e.g., rapid recovery)

Prevalence = 4/10
Incidence = 7 (or 7/8 at risk)



Take care in interpreting prevalence data

Combining prevalence with other data is usually more informative:

- Are there dead or dying animals?
- P(disease) *often* increases with intensity of infection
  - low prevalence of high intensity infections is more consistent with a die-off than low intensity infections
- Susceptibility of the species of interest
  - low prevalence in a very susceptible species would be interpreted differently than similar prevalence in a very tolerant species
- Timing/phenology
  - low prevalence in young larvae could mean low susceptibility/transmission OR very early in an epidemic

Comparing prevalence: Chi-square tests

	Pop A	Pop B	*Total*
Infected	10	20	*30*
Not infected	25	25	*50*
Total	*35*	*45*	*80*

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

- Can accommodate multiple groups (e.g., ponds, species, whatnot)
- Simple to calculate (even by hand)
- Requires that expected count in all cells be ≥5 which may be difficult with low sample sizes and/or low (or very high) prevalence

Comparing prevalence: Chi-square tests

	Pop A	Pop B	*Total*
Infected	10	20	*30*
Not infected	25	25	*50*
Total	*35*	*45*	*80*

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

If there is no difference between the two populations, we would expect the proportion infected to be the same in both: **30/80=0.375**

Of the 35 sampled in Pop A, we expect **35 × 0.375 = 13.125** infections. Similarly we would expect **45 × 0.375 = 16.875** infected in Pop B.

The expected number of uninfected in each pond is calculated similarly:
35 × (50/80) = 35 × 0.625 = 21.875 uninfected in population A, and
45 × (50/80) = 45 × 0.625 = 28.125

Comparing prevalence: Chi-square tests

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

	Pop A	Pop B
Infected	(10-13.125)²/13.125	(20-16.875)²/16.875
Not infected	(25-21.875)²/21.875	(25-28.125)²/28.125

	Pop A	Pop B
Infected	0.7440476	0.5787037
Not infected	0.4464286	0.3472222

Sum = 2.116402

Compare to Chi-square distribution with (rows-1) (columns-1) = (2-1)(2-1) = 1 d.f.
so ***P* = 0.1457**

Note: with 2x2 table, a correction is usually applied by stats packages

Comparing prevalence: Margins & test options

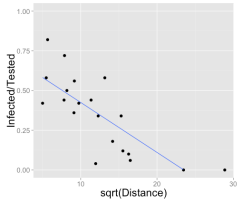
	Pop A	Pop B	*Total*
Infected	10	20	*30*
Not infected	25	25	*50*
Total	*35*	*45*	*80*

`chisq.test()` function in R stats
NOTE: when `simulate.p.value=TRUE` assumes both R & C fixed
`fisher.test()` function in R stats
`GTest()` function in R package `DescTools` or
`G.test()` function in R package `RVAideMemoire`
`barnardw.test()` function in R package `Barnard`

Experimental Design	What is fixed?	Large sample	Small sample
Model I	Total sample size, *N*	Chi-square G-test	G-test with Yates correction
Model II	Either row totals (*R*) or column totals (*C*)	Chi-square G-test Barnard's test	G-test with Yates correction Barnard's test
Model III	Both row totals (*R*) & column totals (*C*)	Chi-square Fisher's exact	Fisher's exact

Comparing/modeling prevalence: logistic regression

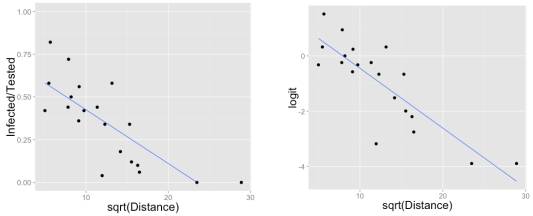
Accommodates one many
categorical (e.g., pond, species) or
continuous predictors (e.g., pond size, salinity)
Models the probability of some binary outcome (i.e., infection, death) in a pond (or individual)



Comparing/modeling prevalence: logistic regression

The logit transform of this probability is a linear function of the predictors

$$\text{logit}(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i + \cdots + \beta_n x_i$$



Comparing/modeling prevalence: logistic regression

$$\text{logit}(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i + \cdots + \beta_n x_i$$

We can recover the probability by simple back-transformations

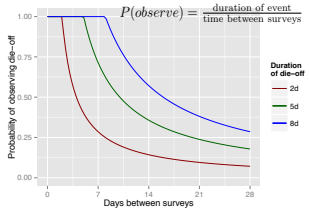
$$\exp(\text{logit}(p_i)) = \left(\frac{p_i}{1-p_i}\right) = e^{\beta_0+\beta_1 x_i+\cdots+\beta_n x_i}$$

$$p_i = \frac{e^{\beta_0+\beta_1 x_i+\cdots+\beta_n x_i}}{1+e^{\beta_0+\beta_1 x_i+\cdots+\beta_n x_i}} = \frac{1}{e^{-(\beta_0+\beta_1 x_i+\cdots+\beta_n x_i)}}$$

Can make statements about how the probability or odds of infection (or death) change with the predictor

Be careful about the units!

Detecting die-offs or other temporary events



General advice

- Remember that P=0.05 is not a magic threshold for what does/does not matter!
- Present effect sizes (change in prevalence between populations or with some predictor) to give a sense of biological importance
- Provide confidence intervals to give an idea of certainty in the estimate

General advice

- Graph your data in a way that
  - Honestly illustrates effects and confidence
    - Include zero and one when graphing prevalence
    - Show confidence intervals or confidence envelopes (logistic regression)
  - Allows the raw data can be recovered for future (e.g., meta) analyses
    - e.g., if you show prevalence as points on a graph, provide sample sizes
- Provide context: prevalence is only part of the story