Exploring detectability and occurrence with a simple example

EEB 5894-003 "Detection, Occurrence, and Abundance"

Lab: Tuesday, September 8, 2015

Occupancy is defined as a probability: the probability that a given spatial unit is occupied by a species[1]. Occupancy does not have to account for imperfect detection, but you should specify whether your occupancy estimates correct for imperfect detection or not. When occupancy does not account for imperfect detection, we call that *naïve occupancy.* Calculating naïve occupancy is often one of the first things we do when exploring a new occurrence data set.

[1] Occupancy assumptions are sometimes relaxed in order to define occupancy as the probability that a given spatial unit is used by a species. We will discuss these assumptions later

As we have now discussed in class, our innate knowledge of true occupancy increases when we return to a site and replicate our survey. By the addition of just 1 more survey, assuming that the species has not left the area if it truly occurs there[2], we suddenly have much more information on true occurrence, but also on detectability.

[2] This is called the 'closure' assumption. And yes, it's important.

Exploring occurrence patterns

Load in the data file "YRWA_DETECTION.CSV" into R, you can find it on HuskyCT. This file shows the outcomes of point-count surveys for Yellow-rumped Warblers (*Setophaga coronata*) at 368 different points in post-fire landscapes of the Sierra Nevada of California. You'll notice that the data is organized as a JxK matrix, with J rows indicating sites, and K columns indicating different survey events[3]. As you can see, each point was surveyed 3 times, and a warbler was either detected (1) or not-detected (0).

[3] In this case, all 3 surveys were done on the same day. How long between repeat surveys is an aspect of occupancy study design

- *What is the naïve occupancy rate for each of the three visits?*

Hopefully you discovered that the naïve occupancy rate for each visit varies. Combining the information across all 3 surveys, we can improve our estimate of naïve occupancy.

- *What is the estimate of naïve occupancy after combining the detections from all three visits per site?*

This guess, combining data from all three surveys, is definitely closer to truth than data from just one single survey. But to improve our estimate of true occupancy beyond this point, we need to start thinking about detectability[4].

[4] While detectability can be caused by many different survey- or site-specific characteristics, we are assuming for now that all sites and survey events were approximately the same, and thus detectability is a constant probability across all individual surveys.

Exploring detectability

The easiest way to start learning about detectability is to explore the pattern or frequency with which different *detection histories* were observed. A detection history is the series of detections or non-detections observed at a single site. We often write detection histories in curly brackets; for example {0 1 1} represents three visits, the first visit was a non-detection, followed by two detections.

- *Create a table that shows the proportion of sites that have 0, 1, 2, or 3 detections.*
- *Restrict the previous table to show the proportion of sites that 1, 2, or 3 detections.*

WHAT DO THESE FREQUENCIES TELL US ABOUT THE LIKELY TRUE PROBABILITY OF DETECTION?

WHY IS IT DIFFICULT TO USE {0,0,0} DETECTION HISTORIES TO ESTIMATE DETECTION PROBABILITY?

Throwing darts at p

The frequency of detection histories can tell us a lot about the likely probability of detection even without doing any modeling. Specifically, we can focus on just the sites that have at least 1 detection, because at the very least we know those sites are all occupied, and that 0's represent false negatives[5]. While any true probability of detection where $0 < p < 1$ may result in capture histories of {1 1 1}, {0 1 1}, and {0 0 1}, the frequency with which these capture histories are observed from a large sample of sites will depend greatly on the value of p. Thus, we can simulate the *expected frequencies* of detection histories given a chosen value for p by randomly drawing a large number of times (e.g., 10,000) from a binomial distribution[6].

- *What detection history frequencies would we expect if $p = 0.95$?*
- *What detection history frequencies would we expect if $p = 0.05$?*
- *What detection history frequencies would we expect if $p = 0.5$?*

COMPARE THE EXPECTED AND OBSERVED DETECTION HISTORY FREQUENCIES. WHERE DO WE THINK TRUE p LIES?

- *Using the same randomized process to get expected frequencies as above, use a trial and error method to roughly narrow in (to 1 decimal places) on what we think the true probability of detection is for Yellow-rumped Warblers in this dataset[7].*

[5] Assuming closure

[6] In R, you can draw randomly from a binomial distribution using the command `rbinom()`. It takes 3 arguments, the number of random draws, the number of trials per draw, and the probability of success of a single draw.

[7] *This is a very approximate method, so don't fret over it too much.*

Trial and error is good, but let's try one step better. Notice that as close as you get to our observed frequencies, our observations do not meet our expectations. Let's quantify this difference, and then try to minimize it. The value of p that minimizes the difference between expected frequencies and observed frequencies is arguably our best estimate of p (for now!).

Let's define this difference, δ, as the sum of the absolute value of the difference between expected and observed frequencies for capture histories with 1, 2, or 3 detections. In other words:

$$\delta = \sum_{i=1}^{3} |expected_i - observed_i|$$

- *For a sequence of 1000 potential values of p between 0.001 and 1, use a* `for`*-loop (or your own preferred method) to calculate the value of δ for the entire range of p, and save this as a vector.*
- *Plot δ versus p. Which value of p minimizes δ?*

Let's call this computational best estimate of *p:* $\hat{p}$.

Putting $\hat{p}$ into occurrence

So now we think we know something about p derived from the observed frequencies of capture histories. Terrific! The next step is to use our estimate of p to estimate how many of our {0 0 0} capture histories were probably truly occupied.

- *Given $\hat{p}$, what percentage of the {0 0 0} capture histories are likely false negatives, i.e., arising from imperfect detection of an individual that was truly present on all three occasions?*
- *Given how many sites we know were occupied (because they had at least 1 detection), plus the number of sites we expect to be occupied despite not detecting the species, what is our new best estimate of occupancy?*

THIS IS UNDOUBTEDLY AN IMPROVEMENT FROM OUR ORIGINAL OCCUPANCY ESTIMATE, BUT THERE IS STILL A MAJOR PROBLEM IN HOW WE ESTIMATED p. DO YOU KNOW WHAT IT IS?

One final exploration

As you realized above, our estimate of $\hat{p}$ is too high[8]. Our failure to learn from {0 0 0} histories has cost us, and the cost is that we we think we have done a better job at detecting the species than we have.

[8] We call this bias 'optimistic' because the truth is worse than we think: p is truly lower than we have estimated

We can correct for past mistakes by re-calculating δ as the difference between the observed and expected for all capture histories, including those with 0 detections, not just those with 1 detection.

- *It should be easy to repeat the steps above and calculate δ', which includes frequencies of 0-detection histories as well. After doing this, plot p versus δ' and calculate our new best estimate of $\hat{p}$.*

How different is this new estimate of $\hat{p}$ from the previous one? They should be radically different.

What is going on?

What assumption did we just make that's wrong?

If you can figure out the wrong assumption, you'll start to realize that this problem cannot be solved simply by optimizing 1 parameter (p). Rather, we have to optimize 2 parameters: p and the true occupancy rate, which we will call ψ.