WILD 502 - Estimating Abundance for Closed Populations with Mark-Recapture Methods

Reading: Chapter 14 of WNC book (especially sections 14.1 & 14.2)

- Estimating *N* is much more difficult than you might initially expect
- A variety of methods can be used:
  - Census – assume count all animals in the population
  - Sample plots – assume count all animals on plots
  - Transect methods – estimate detection probability as a function of distance from a line transect or point to animals
  - Capture-recapture – estimate capture probability – our focus
- Or, can abandon estimation and use an index – often done, seldom tested
- Essentially comes down to dealing with counting animals and relating the count to the number in the population somehow.

Basic concept or model that is often relevant when estimating N:

Expected Count = N x p, where N = pop'n size & p = prob of detection

Estimator: $\hat{N} = \frac{C}{\hat{p}}$

That is, you go out and count animals and want to relate that count *(C)* to *N*.
Thus, can see that rigorous estimation of *p* is key to good estimation of *N*.

- Census – assume $p = 1$ for the entire study area

A closed population has no births, deaths, immigration, or emigration during the study period. Thus, *N* (a state variable) really is a logical focus for a population study during such a period! And rate variables such as survival are not of interest with closure. (We'll see how to work with both closed and open periods soon).

Background:

Ball and urn studies:

Reach in to an urn with 100 white balls and remove a sample ($n_1 = 30$). Mark the balls in the sample and replace them. Take another random sample ($n_2 = 36$) and count the number of marked ($m_2 = 10$) and unmarked balls ($u_2 = 26$).

We expect the proportion of marked balls to total balls in the population to be the same as the proportion of marked and unmarked balls observed in our sample.

$$\frac{n_1}{\hat{N}} = \frac{m_2}{n_2} \quad \text{OR} \quad \frac{30}{\hat{N}} = \frac{10}{36},$$

$$\hat{N} = \frac{n_1 \cdot n_2}{m_2}$$

This is the **Lincoln-Petersen Estimator** for 2 trapping occasions (see pg 290 of *WNC*)

$$\hat{N} = \frac{30 \cdot 36}{10} = 108$$

You can re-arrange this equation to see that $\hat{N} = \frac{C}{\hat{p}}$.

Consider n_1 to be the count statistic and estimate $\hat{p}_1 = {}^{n_1}/_{\hat{N}} = {}^{m_2}/_{n_2}$

$$\hat{N} = \frac{n_1}{\hat{p}_1}, \text{ or } \hat{N} = \frac{n_1}{\frac{m_2}{n_2}} = \frac{n_1 \cdot n_2}{m_2}$$

Probability distribution for 2 occasions:

$$P(n_1, n_2, m_2 \mid N, p_1, p_2) = \frac{N!}{m_2!(n_1 - m_2)!(n_2 - m_2)!(N - r)!}(p_1 p_2)^{m2}(p_1 q_2)^{n_1 - m_2}(q_1 p_2)^{n_2 - m_2}(q_1 q_2)^{N-r}$$

Where $q_i = 1\text{-}p_i$, $r = n_1 + n_2 - m_2$ = number of unique animals captured in study, & $N - r$ *is* the number of animals never caught.

Closed-form MLEs:

- $\hat{p}_1 = {}^{m_2}/_{n_2} = n_1 / \hat{N}$
- $\hat{p}_2 = {}^{m_2}/_{n_1} = n_2 / \hat{N}$

Bias-adjusted estimator: $\hat{N} = \frac{(n_1 + 1)(n_2 + 1)}{m_2 + 1} - 1$; unconditionally unbiased $n_1 + n_2 \geq N$.

$$\widehat{var}(\hat{N}) = \frac{(n_1+1)(n_2+1)(n_1-m_2)(n_2-m_2)}{(m_2+1)^2(m_2+2)^2}$$

Example: *N = 250, p_1 = 0.6, p_2 = 0.3.*
Expected frequency for each encounter history

Encounter history	Expected frequency	Probability
11	45	$Np_1p_2 = 250 \cdot 0.6 \cdot 0.3$
10	105	$Np_1(1-p_2) = 250 \cdot 0.6 \cdot (1-0.3)$
01	30	$N(1-p_1)p_2 = 250 \cdot (1-0.6) \cdot 0.3$
00 (not seen)	70	$N(1-p_1)(1-p_2) = 250 \cdot (1-0.6) \cdot (1-0.3)$

- $n_1=150$
- $n_2=75$
- $m_2=45$
- $\hat{p}_1 = m_2 / n_2 = 45/75 = 0.6$ or ... $\hat{p}_1 = n_1 / \hat{N} = 150/250 = 0.6$
- $\hat{p}_2 = m_2 / n_1 = 45/150 = 0.3$ or ... $\hat{p}_2 = n_2 / \hat{N} = 75/250 = 0.3$

Assumptions:

1) Population is closed to additions (birth & immigration) and to losses (death & emigration)
2) Marks are not lost or overlooked or misread by researchers
3) All animals are equally likely to be captured in each sample (this is the assumption that we will almost always need to relax). We'll relax it to be *capture probabilities are appropriately modeled.*

For most real biological applications, we need to:

1) Consider possible heterogeneity in capture probability,
2) Carefully consider the concept of closure, and
3) Use more than 2 occasions.

Parameters of interest:

1) N – population size
2) D – population density (possible though not our focus)
3) *p, c* – capture & recapture probabilities

Statistics:

1) n_j - # captured on the j^{th} occasion, j = 1, 2, …, t.
2) n. – total # of captures in the study = $\sum_{j=1}^{n} n_j$
3) u_j - # of new animals captured on the j^{th} occasion, j = 1, 2, …, t.
4) f_j – capture frequencies = # individuals captured exactly *j* times in *t* days of trapping.
5) M_{t+1} –# of different individuals caught in entire study.
6) M_j – number marked animals in the population at the time of the j^{th} sample, Note that $M_0 = 0$.
7) $M_.$ – Sum of the M_j, not including $M_{t+1.}$
8) $m_{j\,-\,\#}$ of marked animals captured in the j^{th} sample.
9) m. – sum of the $m_{j.}$
10) t – the number of trapping occasions in the study.
11) $\{X_\omega\}$ = the set of possible encounter histories in the study.

The probability distribution for the set of possible encounter histories under constant capture probability is:

$$P(\{X_\omega\} \mid N, p) = \frac{N!}{\left[\prod_\omega X_\omega !\right] \cdot (N - M_{t+1})!} \cdot p^{n.} \cdot (1-p)^{t \cdot N - n.}$$

The MLE for constant capture probability is:

$$L(N, p \mid \underline{X}) = \frac{N!}{\left[\prod_\omega X_\omega !\right] \cdot (N - M_{t+1})!} \cdot p^{n.} \cdot (1-p)^{t \cdot N - n.}$$

To help you see what's going on with the multinomial coefficient, consider :
3 animals with 11 EH,
2 animals with 01 EH, &
3 animals with 10 EH.
Thus, t = 2, $n_1 = 6$, $n_2 = 5$, and n. = 11.

$$L(N, p \mid X) = \frac{N!}{3! \cdot 2! \cdot 3! (N-8)!} \cdot p^{11} \cdot (1-p)^{t \cdot N - 11}$$

Substituting values of *N* and *p* into the equation allows estimation of the most likely estimates for these 2 parameters given the data set and the model. Of course, this model is simple and makes the restrictive assumptions that $p = c$ and is constant across occasions & animals. As we'll discuss in class, much of the work that has been done for this class of models has focused on easing this assumption and modeling variation in *p*.

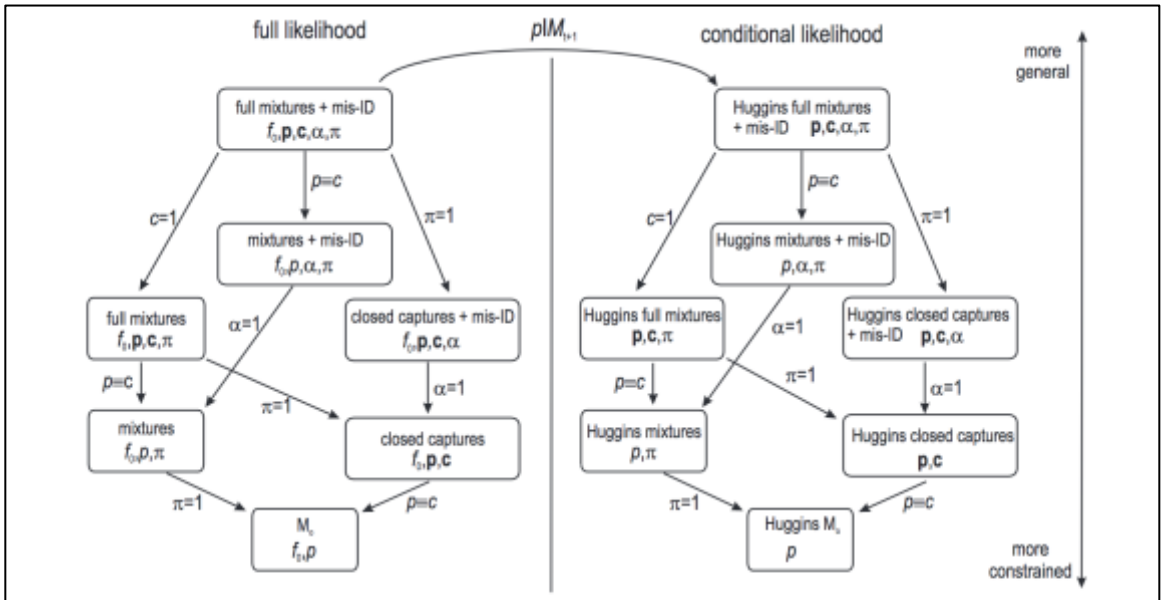
Basically, 3 broad sources of variation are considered possible:

1) temporal variation in capture probability ('t'),
2) behavioral responses to trapping ('b'): $p \neq c$, and
3) heterogeneity in capture probabilities for different individuals ('h').

By considering all possible combinations of these 3 factors, 8 models are obtained: M(0), M(t), M(b), M(h), M(tb), M(th), M(bh), and M(tbh). MARK also has a variety of other models that use a conditional likelihood approach in which N is not actually considered as a parameter. This is done by conditioning on only animals encountered (r). This can be handy because it lets us bring individual covariates to bear.

NOTE: Individual covariates are typically not used in closed models though some approaches do allow their use (see pages 300-302 in *WNC* book). This is because, we are now estimating capture probability for all the animals in the population but don't have covariate values for the animals that were never captured. We can use group-level covariates, which can be useful. And, if time permits we will discuss models that allow use of individual covariates.

The figure below is from chapter 14 of *CW*, which was written by Paul Lukacs, and shows the various models that exist. We'll focus on the *full likelihood models for p and c*. We'll discuss the ideas of mixtures and mis-identification if time allows later in the semester.



The first and most important split is between the models with abundance (or, rather, f_0) *in* the likelihood (Otis *et al.* 1978) and those with abundance *conditioned out* of the likelihood (Huggins 1989). We refer to the former as 'full likelihood' models, and the latter as either 'conditional likelihood' or 'Huggins' models. This is a major division that results in the two types of models not being comparable

Likelihoods for some of the basic models:
To give you a better feel for how the parameters are estimated in the various models, here are the likelihood equations for M(t), M(b), and M(tb).

M(t) (t+1 parameters):

$$L(N, p_j \mid \underline{X}) = \frac{N!}{\left[\prod_{\omega} X_{\omega}!\right] \cdot (N - M_{t+1})!} \cdot \prod_{j=1}^{t} p_j^{\,n_j} \cdot (1 - p_j)^{N - n_j}$$

If $t=2$, the MLE for N for M(t) is: $\hat{N} = \frac{n_1}{\frac{m_2}{n_2}} = \frac{n_1 \cdot n_2}{m_2}$, i.e., the Lincoln-Petersen estimator

M(b) (3 parameters):

As noted on page 299 of your textbook, in the behavioral response model, capture probability can vary as a result of previous capture such that $p \neq c$, and the response can be trap happiness ($c > p$) or trap shyness ($p > c$).

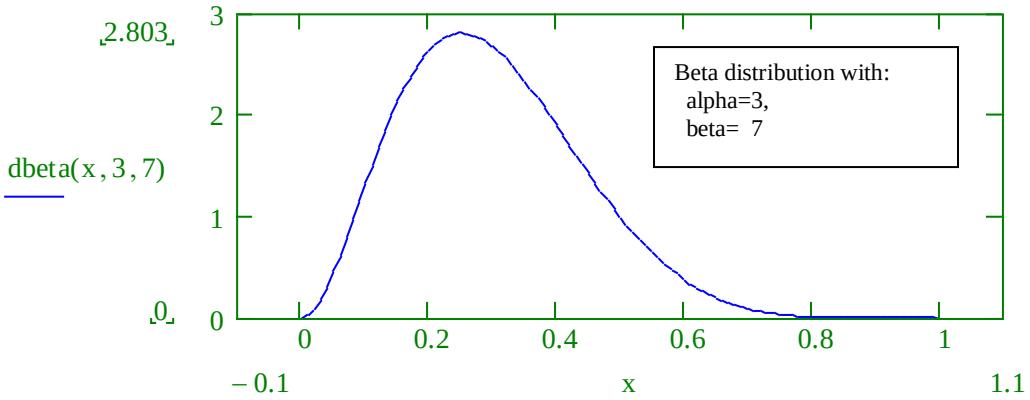
$$L(N, p, c \mid \underline{X}) = \frac{N!}{\left[\prod_{\omega} X_{\omega}!\right] \cdot (N - M_{t+1})!} \cdot p^{M_{t+1}} \cdot (1 - p)^{t \cdot N - M_{t+1} - M_{\cdot}} \cdot c^{m_{\cdot}} \cdot (1 - c)^{M_{\cdot} - m_{\cdot}}$$

A noteworthy property of this model is that the recapture information is used only in the estimation of c (a nuisance parameter). That is, recaptures don't provide information with respect to estimation of p or N. Thus, once M(b) has been identified as appropriate, the estimation is done as if the study were a removal study. Thus, it is critical in this type of study that "depletion" of unmarked animals is achieved, i.e., the number of new animals captured on each occasion should decrease with each successive occasion. If this does not occur sufficiently, the model can fail to produce results. Valid estimates are obtained if:

$$\sum_{j=1}^{t} (t + 1 - 2 \cdot j) \cdot (n_j - m_j) > 0$$

Failure can readily occur if you have temporal changes in p such that few individuals are caught early in the study and many new individuals are caught late in the study. Individual heterogeneity in p causes underestimation of N with M(b): the magnitude of bias depends upon the number of animals that are essentially uncatchable.

M(h) (N+1 parameters):
As estimation is not possible under the general formulation where every animal has its own p ($p_1, p_2, \ldots, p_N = \{p_i\}$), it is useful to think of $\{p_i\}$ as a random sample (size of sample is N) coming from some probability distribution F(p), where each p_i falls between 0 & 1. A maximum likelihood estimator is possible if the particular family of distributions of which F(p) is a member is specified. One possibility is the Beta distribution (the beta distribution is specified by 2 parameters (alpha and beta) and hence, the name. The beta distribution is a continuous distribution with all of its non-zero probability between 0 and 1.



This is an appealing idea that is useful for conceptualizing how the p's might be distributed in the population, but has not proved useful in practice. A variety of alternatives exist (See figure above from chapter 14 of *CW* by P. Lukacs) including mixture models, as well as several approaches available in Program CAPTURE, which can be called from within Program MARK.

Note: Models containing heterogeneity are probably appropriate for many studies. Thus, it is worth understanding the concept and knowing some of the ways you can obtain estimates for M(h) models.

M(bh) (2N + 1 parameters- each animal has its own *p* and a *c*):
The estimator of N for this model is based on the first-capture data (similar to what's done with M(b), i.e., a removal estimator). But, with heterogeneity more must be done; we need a *generalized* removal estimator. With heterogeneity, you expect the average probability of first capture to decrease with each occasion & you expect the most rapid decrease over the first few occasions. On the first occasion, no animal has been caught before so the average p for $j=1$ is high relative to p_6. Why? Because by occasion 6 most of the animals that are easy to catch have already been caught! So, the trick to estimation is to see if you can fit the data using only a few average p's. The idea is to fit a sequence of increasingly general models to the average value for the values of p_j and to use the simplest model that fits.

Step 1: Evaluate the fit of using a single average p for all ocassions

$$\bar{p}_1 = \bar{p}_2 = \bar{p}_3 = \ldots \bar{p}_t \text{, this is the same as M(b)}$$

Step 2: If the model on the previous step doesn't fit, generalize & evaluate again.

$$\overline{p}_1 \neq \overline{p}_2 = \overline{p}_3 = ...\overline{p}_t$$

Step 3: Generalize more as needed:

$$\overline{p}_1 \neq \overline{p}_2 \neq \overline{p}_3 = ...\overline{p}_t$$

Etc. Etc.

Now that you're not trying to fit a *p* for every animal in the population, you CAN build these models in MARK. This is very useful because for models in MARK we can compare models with our typical AIC procedures. In lab this week, you'll see how to build each of the models described in the steps above.

Okay, we won't go over the other models in any more detail. But, I hope this introduction to the models along with readings has given you a decent idea of the key concepts of estimation for these models. Also, realize that there is a solid body of literature available on this topic. Finally, if you are working with this model type in your research, you will need to become more familiar with some of the more complex models and the attributes of their estimates. In lab, you will become familiar with how to: (1) produce estimates for all models except M(tbh) and (2) conduct simulations that evaluate estimator performance under different sampling conditions.

As I think you can surmise, it is best if you can be justified in using simpler models. Thus, you need to consider ways of having as few factors as possible affect the capture probabilities in your studies. In terms of heterogeneity, with the models we are reviewing ,you are limited in your use of individual covariates. That is, you can only use covariates that can be entered using the design matrix, e.g., covariates that apply to a group of animals, e.g., trapping effort on each occasion, sex, size class, etc. Breaking the data up by group covariates, e.g., sex, size class, etc., is a method that can be used to reduce heterogeneity within each group. Of course, it also means that you will need to have adequate data for estimation for each group! And, you can work to learn about other approaches such as Huggins Models listed under "Conditional Likelihood" on the right side of the conceptual diagram of Closed Models in Chapter 14 of CW.

Over-Parameterized Closed Captures Models

As nicely explained in section 14.3.1 of the CW Chapter by P. Lukacs, when a model is specified in Program MARK with unidentifiable *p* estimates, the estimates of *N* are just M_{t+1}. Conceptually, using a model with time-varying capture probabilities, here's why. The estimate of population size is basically,

$$\hat{N} = \frac{M_{t+1}}{1-(1-p_1)\cdot(1-p_2)\cdot...(1-p_t)}$$

where the *numerator* is the number of individuals captured and the *denominator* is the probability that an animal was captured during the study (1 minus all the probabilities of *not* being captured!). That is, the probability of not being initially captured on the first occasion is 1-p_1, not being captured on the second occasion is 1-p_2, and so on to 1-p_t. The product of these terms is the probability of never being captured during the study. Thus, 1 minus this product is the probability of being captured at least once during the study. Therefore, the denominator is a "correction" to inflate the numerator.

When a closed-captures model is over-parameterized, the last term in the denominator becomes zero because p_t is estimated as 1 (for reasons I won't go into). This makes the product portion of the denominator = 0 and the entire denominator = 1. This results in $\widehat{N}$ = M_{t+1}. When you use closed-captures modeling, you should always know what the value of M_{t+1} is and be sure that your results for $\widehat{N}$ have not collapsed to the value of M_{t+1} because of over-parameterization. The value of M_{t+1} is in the full output for each model and so very easy to check.

Note: Program MARK actually parameterizes the likelihood in terms of the number of individuals never caught, f_0, such that $\widehat{f_0} = \widehat{N} - M_{t+1}$, where the notation f_0 is a reference to the **f**requency, or count, of animals observed **0** times. Thus, MARK provides estimates of f_0 in the real parameters list, and provides estimates of N under the derived parameters list. MARK uses a log link for f_0, which keeps values of f_0 between 0 and infinity. N is estimated as M_{t+1} + $\widehat{f_0}$, and so the variance for $\widehat{N}$ is the variance for $\widehat{f_0}$ as M_{t+1} is known.

An additional problem with closed capture models in Program MARK is that often the number of parameters is not correctly computed because values of $\hat{N}$ close to M_{t+1} are assumed to not be estimated – causing an error in the algorithm to determine the number of parameters actually estimated, which causes errors in the resulting AIC value. So, as always pay attention to parameter counts and adjust counts as needed.