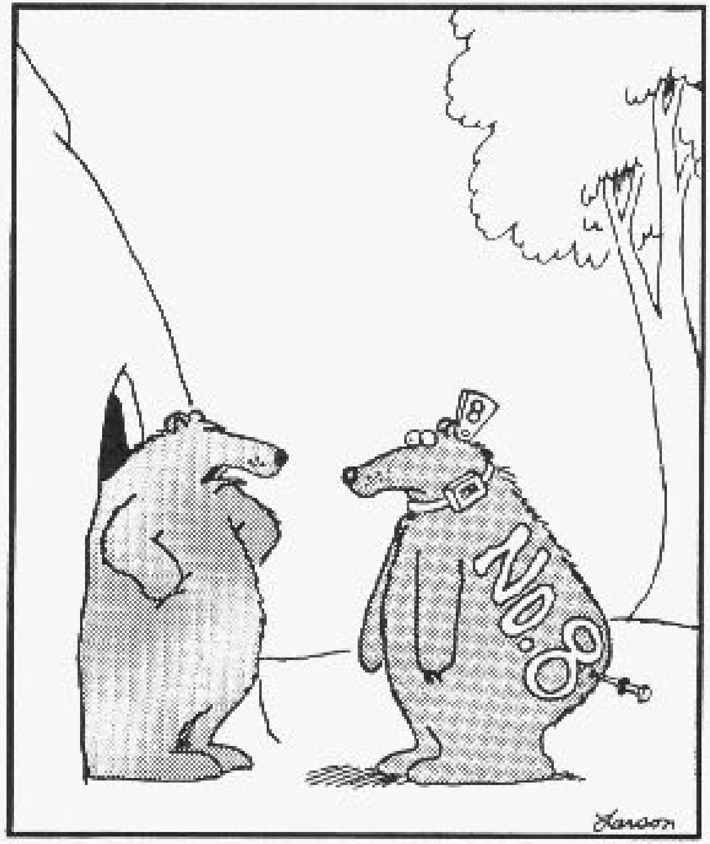
Estimating population size with maximum likelihood



"Late again! ... This better be good!"

$$L(q_i|N, H_i) = \binom{N}{y_i} q_1^{x_{11}} q_2^{x_{10}} q_3^{x_{01}} q_4^{x_{00}}$$

Estimating population size from two captures – maximum likelihood

- We can express mark-recapture estimators in terms of **capture histories**
  - Requires individual marks if there are more than 2 capture periods
- We can model the probability of capturing an individual as an **encounter probability**
- We can use capture histories to estimate population size using **maximum likelihood** methods
- Advantages:
  - Wide range of different mark/recapture designs can be accommodated
  - We can test hypotheses about the mark/recapture study

Encounter probability

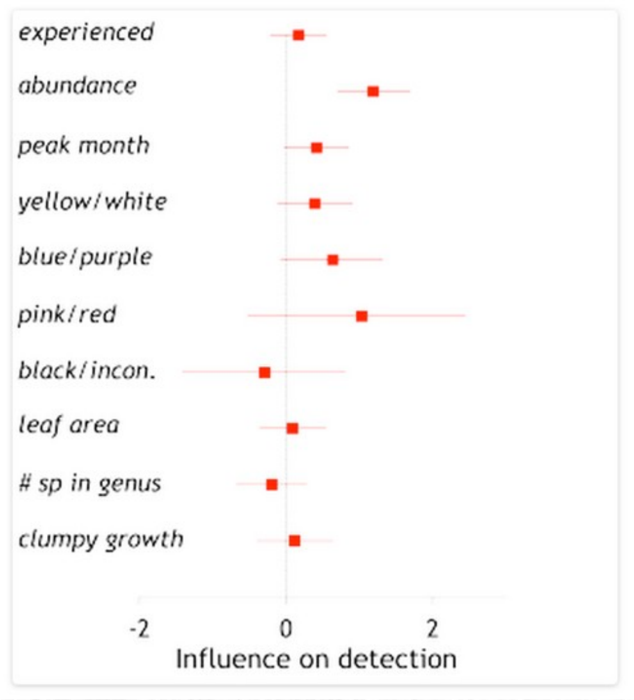
- In mark-recapture work, we know not every individual is caught in every capture period (some are never caught)
- Encounter probability = p = the probability of encountering an individual during a single capture period
- Encounter probability ranges from 0 to 1
  - $p = 1 \rightarrow$ every individual is observed (if $p = 1$ we are doing a census)
  - $0 < p < 1 \rightarrow$ some degree of uncertainty about how many individuals are present, random variation in numbers counted at each capture period
  - The closer p is to 1, the more certainty we have about population size
- Low encounter probability comes from several sources, some of which we can minimize

Encounter probability for a botanist

- Plants don't move and hide
- But, annual plants aren't always visible due to phenology
- Small plants may be perfectly detectable when you're near them, but detectability may drop off rapidly with distance
- Defining individuals may be difficult

Encounter probability at a distance

From Garrard et al. 2012



Defining individuals – sometimes it matters, sometimes not

Matters for population genetics

May not matter for population dynamics

Encounter probability for a zoologist

- We have all the problems botanists do (aside from defining individuals), plus animals move and hide
- Low encounter probability is caused by:

- Camouflage, hiding
- Nocturnal species
- Large home ranges, low density
- Rarity

- Trap saturation
- Environmental conditions
- Inexperienced workers

Example: bird surveys

- For large birds (raptors) usually done by sight
  - What if it's in a different portion of its home range than you?
  - What if it's behind something?
  - What if it's camouflaged?
- For songbirds, usually done by sound
  - Males do most of the singing, how many females and subadults?
  - What if they don't sing while you're there?

Art Wolfe's Vanishing Act pictures

The LP estimator and encounter probability

- Turns out, the LP estimator is a special case of a general model based on encounter probability
- Expressing the estimator in terms of encounter probability allows us to use methods to asses:
  - Variation in encounter probability over time
  - Trap responses (trap happiness, trap shyness)
- We need to re-express LP capture data as capture histories

Capture histories

- Capture histories are the sets of captures (1) or non-captures (0) for individuals in the population during the capture periods
- Require individually identifiable marks
- For a two-period study, the four possible outcomes are:
- These are often shortened to 10, 11, 01, 00
- At the end of the study, we can count how many 10, 11, and 01 there were
- Since these are all the possible histories, the sum of their frequencies will equal N – but the frequency of 00 is unknown
- So, estimating N means **estimating how many 00's there are**

Mark	Recapture
1	0
1	1
0	1
0	0

Relating capture histories to LP

Total caught (and marked) on first occasion (M in LP)

$$n_1 = x_{10} + x_{11}$$

Total caught on the second occasion (c in LP)

$$n_2 = x_{01} + x_{11}$$

Total caught on both occasions (r in LP)

$$m_2 = x_{11}$$

t1	t2	Freq.
1	0	X_{10} = 50
1	1	X_{11} = 50
0	1	X_{01} = 150
0	0	X_{00} = ??

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

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

M = n_1 = 100

r = m_2 = 50

c = n_2 = 200

The payoff...

- The proportion of animals that are marked recaptures in the second is the encounter probability:

  $p = m_2/n_2$

- We can thus re-express the LP estimator as

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

- LP can be expressed as estimator that uses capture histories, and encounter probabilities

Translated LP estimate

$M = n_1 = 100$

$r = m_2 = 50$

$c = n_2 = 200$

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

$p = m_2/n_2 = 0.25$

$$\hat{N} = \frac{100 \times 200}{50} = \frac{100}{0.25} = 400$$

Total captures was 50 + 50 + 150 = 250

Can calculate never captured as 400 – 250 = 150, but this value not estimated directly by the model

Maximum likelihood estimates of population size

- Four possible capture histories: (11, 10, 01, 00)
- Two unknowns:
  - The probability of each capture history
  - How many of history 00 there are
- With four possible outcomes (not two) we need the multinomial distribution (not binomial) to be our likelihood function

The multinomial probability distribution

- Four possible outcomes: 11, 10, 01, 00
- The data will be frequencies of each outcome: x_{11}, x_{10}, x_{01}, x_{00}
- Frequencies of 11, 10, and 01 are known from the data, but frequency of 00 is unknown
- Multinomial is structured like the binomial
  - Probability of each outcome = probability part
  - Number of different ways to get it = counting part

$$p(H_i | N, q_i) = \binom{N}{y_i} q_1^{x_{11}} q_2^{x_{10}} q_3^{x_{01}} q_4^{x_{00}}$$

The multinomial probability distribution and likelihood function

History	Frequency	Probability
11	x_{11}	q_1
10	x_{10}	q_2
01	x_{01}	q_3
00	x_{00}	q_4

$$p(H_i|N, q_i) = \binom{N}{y_i} q_1^{x_{11}} q_2^{x_{10}} q_3^{x_{01}} q_4^{x_{00}}$$

$$L(q_i|N, H_i) = \binom{N}{y_i} q_1^{x_{11}} q_2^{x_{10}} q_3^{x_{01}} q_4^{x_{00}}$$

Using encounter probability (p) to calculate probabilities of histories (q)

So....

$q_1 = p \times p$

$q_2 = p \times (1-p)$

$q_3 = (1-p) \times p$

$q_4 = (1-p) \times (1-p)$

History	Frequency	Probability
11	x_{11}	$p \times p$
10	x_{10}	$p \times (1-p)$
01	x_{01}	$(1-p) \times p$
00	x_{00}	$(1-p) \times (1-p)$

...need to estimate p

Estimating population size – summing frequencies

History	Frequency
11	x_{11}
10	x_{10}
01	x_{01}
00	x_{00}

These are all the possible histories, which means that N is equal to the sum of the frequencies

We know x_{11}, x_{10}, x_{01} from the data

If we can estimate x_{00} we can estimate N

Basic setup

History	Freq		Parameter	MLE		Probability of history	
11	50		p	0.5		0.25	
10	50		f(00)	50		0.25	
01	150					0.25	
00						0.25	

We will be estimating the two values under MLE:

p and f(00) (aka x_{00})

Starting values for p and f(00) are entered (any values, but better if close to final estimates)

Probability of histories are based on the current values of p and 1-p

Calculating the log-likelihood

- Now to convert the multinomial probability distribution into a log-likelihood function
- The “probability part” is easy

$$\log\left(q_1^{x_{11}} q_2^{x_{10}} q_3^{x_{01}} q_4^{x_{00}}\right) =$$

$$x_{11}\log\left(q_1\right) + x_{10}\log\left(q_2\right) + x_{01}\log\left(q_3\right) + x_{00}\log\left(q_4\right)$$

  - For history 00, just use the initial guess for f(00) as the frequency x_{00}
  - p is part of every q, so all depend on one of parameters we are estimating – no unneeded terms

The counting part is trickier...

- We can simplify this $\binom{N}{y_i} = \frac{N!}{x_{11}!\,x_{10}!\,x_{01}!\,x_{00}!}$

$$\ln\left(\frac{N!}{x_{11}!\,x_{10}!\,x_{01}!\,x_{00}!}\right)$$

Drop unneeded terms

$$\ln(N!) - \ln(x_{11}!\,x_{10}!\,x_{01}!\,x_{00}!)$$

$$\ln((M_{t+1} + x_{00})!) - [\ln(x_{11}!) + \ln(x_{10}!) + \ln(x_{01}!) + \ln(x_{00}!)]$$

$$\ln((M_{t+1} + x_{00})!) - \ln(x_{00}!)$$ *To this*

Problem: big numbers and computers

- Computers have limits on the number of digits they can store
- Exceeding the limit can lead to an error message (if you’re lucky), or a silently incorrect answer (if you’re unlucky)
- We need to calculate factorials, which can be very big
- Excel can do factorials up to 170 – any N over 170 will exceed that limit
- Fortunately, we don’t actually need the factorial, we need the log of the factorial

The gammaln() function

- In Excel, you can calculate the log of a factorial by using the gammaln() function

  $\ln(x_{00}!) = \text{gammaln}(x_{00} + 1)$

  $\ln((M_{t+1} + x_{00})!) = \text{gammaln}(M_{t+1} + x_{00} + 1)$

- We now have the second part needed for the log-likelihood in a form we can use in Excel

  $\text{gammaln}(M_{t+1} + x_{00} + 1) - \text{gammaln}(x_{00} + 1)$

Putting it together...

History	Freq		Parameter	MLE		Probability of history		
11	50		p	0.5		0.25		
10	50		f(00)	50		0.25		
01	150					0.25		
00						0.25		
Mt+1	Mult coef	Portion L	Full LnLik					
250	1266.43	-415.888	850.54					

Sum of the two parts

Multinomial coefficient (counting part)

Probability part (sum of frequencies x ln(probabilities))

Numeric solutions – Solver

- With LP, we calculated an **analytical** solution
  - Plugged in numbers to a formula, got an estimate
- With ML, it's common to use **numeric** solutions
  - The likelihood functions often can't be solved for parameters of interest
  - Instead, use a sophisticated form of trial and error to find estimates of p and f(00) that maximize the log-likelihood
  - Solutions found to a fixed (specified) level of precision
- The tool that Excel uses to do this is called Solver

What will happen...

2...by changing the estimates

History	Freq		Parameter	MLE		Probability of history	
11	50		p	0.5		0.25	
10	50		f(00)	50		0.25	
01	150					0.25	
00						0.25	
Mt+1	Mult coef	Portion L	Full LnLik				
250	1266.43	-415.888	850.54				

1. We will tell Solver to make this as big as possible...

The probability of history column depends on p

The mult coef cell depends on f(00)

Full LnLik depends on both, so as p and f(00) are varied Full LnLik will change – Solver stops when changes in p and f(00) no longer increase Full LnLik

And the results...

History	Freq		Parameter	MLE		Probability of history
11	50		p	0.33519		0.11236
10	50		f(00)	197.501		0.22284
01	150					0.22284
00						0.44197
Mt+1	Mult coef	Portion L	Full LnLik			
250	1437.87	-570.829	867.042			

Number of animals never captured, f(00) = 197.501

Encounter probability is 0.335

Probability of never being seen (00) = 0.442

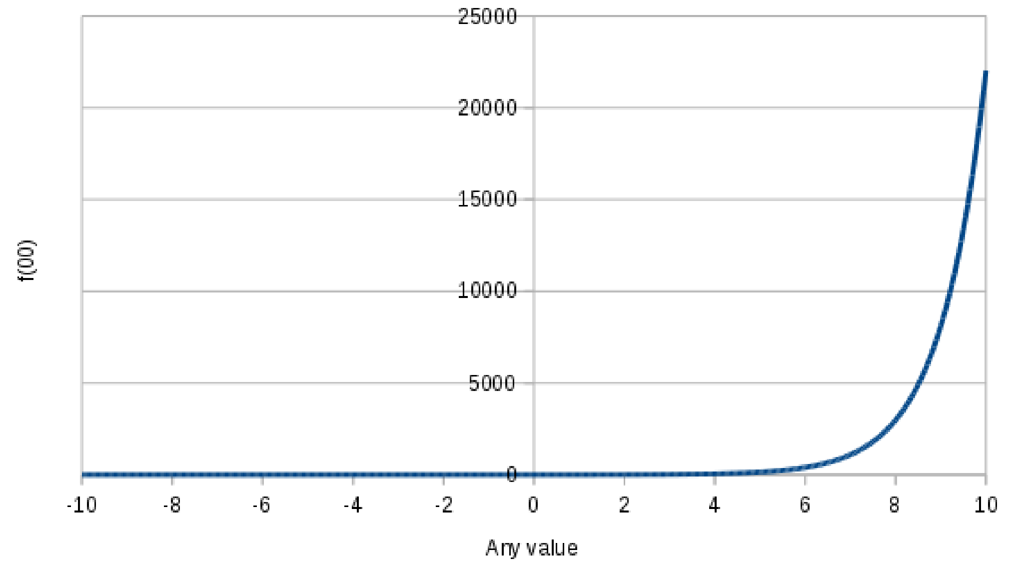
Population size is M_{t+1} + f(00) = 447.5

How did ML compare to LP?

- LP
  - Estimate = 400
  - Never seen = 150
  - Encounter prob. = 0.25
- ML
  - Estimate = 447.5
  - Never seen = 197.5
  - Encounter prob. = 0.335
- They aren’t the same
- ML is better
  - Both p and f(00) are unknown, and the estimate depends on both
  - LP doesn’t estimate f(00), and can be biased at low sample sizes

The log link

- The inverse function for the natural log is exp (it means, raise the base e to the power n)
- A negative exponent, e^{-n}, is just $1/e^{n}$
- So, exp(any number) is positive, negative numbers become increasing close to 0
- Having Solver change (any number), and use exp(any number) for f(00) keeps the estimates of f(00) over 0



Extensions, complications

- This is easy to extend to more capture events
- With more capture events we can ask questions like
  - Are initial capture probabilities different from recapture probabilities (trap happy, trap shy)?
  - Do capture probabilities change over time?
- We will look at these in lab next week