FISH 621
Estimation of Fish Abundance:
3: Simple Mark-Recapture

Dr. Curry Cunningham: cjcunningham@alaska.edu

Course Schedule

Course Calendar

	Tuesday	Thursday	Friday
Week of	2:00-3:30pm	2:00-3:30pm	11:45 am to 2:45 pm
1/10/22	Lecture 1	Lecture 2	Lab 1
1/17/22	Lecture 3	Lecture 4	Lab 2
1/24/22	AMSS: No Class		No lab
1/31/22	SSC Meeting: No Class		No lab
2/7/22	Lecture 5	Lecture 6	Lab 3
2/14/22	Lecture 7	Lecture 8	Lab 4
2/21/22	Lecture 9	Lecture 10	Lab 5
2/28/22	Alaska AFS: No Class		No Lab
3/7/22	SPRING BREAK		
3/14/22	Lecture 11	Lecture 12	Lab 6
3/21/22	Lecture 13	Lecture 14	Lab 7
3/28/22	Lecture 15	Lecture 16	Lab 8
4/4/22	SSC Meeting: No Class		No lab
4/11/22	Lecture 17	Lecture 18	Lab 9
4/18/22	Lecture 19	Lecture 20	Lab 10
4/25/22	Final Exam (Take home)		

Office Hours:

4:00-5:00 ***Mondays*** or by appointment

Lecture	Topic	Lab	Topic
1	Sampling Theory 1		
2	Sampling Theory 2	1	Sampling Theory
3	Simple Mark-Recapture: Petersen		
4	Simple Mark-Recapture: Uncertainty	2	Simple Mark-Recapture
5	Removal, Catch-Effort Estimators		
6	Change-in-Ratio Estimator	3	Removals and Ratios
7	Schnabel: Multi Release-recapture		
8	Jolly-Seber: Open Populations	4	Advance Mark-Recapture
9	Intro to Stan and Bayesian Analysis		
10	Bayesian Mark-Recapture Estimators	5	Bayesian MR
11	CPUE: GLM(M)s		
12	CPUE: GAMs	6	CPUE Analysis
13	Spatiotemporal Models 1		
14	Spatiotemporal Models 2	7	Spatiotemporal Models
15	Line Transect Models		
16	Distance Sampling	8	Transect and Distance
17	Spatial Mark-recapture		
18	CWT, Darroch Implementation (Hilborn)	9	Spatial MR Considerations
19	Heterogeneity		
20	Synthesis	10	Advanced Models

Grading and Evaluation

- Computer Labs (30%)
  - Attendance-based (unless pre-arranged)
  - Encourage group work and discussion
- Homework Assignments (40%)
  - Turned in on ***Canvas***

Correct Version!

Name	Assign Date	Due Date
Homework 1	Thursday, January 27	Friday, February 11
Homework 2	Thursday, February 17	Friday, March 4
Homework 3	Thursday, March 10	Friday, March 25
Homework 4	Thursday, March 31	Friday, April 15

- Final (take home) Exam (30%)
  - Multiple questions, based on analysis

Simple Random Sampling *with Replacement*

- Extending our example of SRS
  - We can include the potential that of our population of sampling units N , our sample n,
    - May contain rare ***repeat*** observations of the same sampling unit
- Leveraging our hypothetical population of $N = 1{,}000$ fish in a lake
  - Each with a length attribute (mm)
    - Under SRS ***with replacement*** an individual may be measured multiple times within our samples y_i
- Within our Alaska Coastal Plain caribou example
  - We might select transects to fly among the $N = 286$ at random
    - But survey the same transect ***on multiple occasions*** if selected again during survey design
- Why sample with replacement?
  - Practical advantage in some situations, if difficult to determine if individual has already been sampled
    - Hook and line sampling (i.e. linear graphite sampling)
  - For a given sample size n simple random sampling ***with replacement*** is inherently
    - ***Less*** efficient than simple random sampling ***without*** replacement

Simple Random Sampling *with Replacement*

- Sample mean of n observations
  - $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$
    - Note: if a unit is selected more than once, its y-value is utilized more than once in the estimator
- The variance of $\bar{y}_n$ (likely unknown)
  - $var(\bar{y}_n) = \frac{1}{nN}\sum_{i=1}^{N}(y_i - \mu)^2 = \frac{N-1}{nN}\sigma^2$
    - Population mean: $\mu = \frac{1}{N}(y_1 + y_2 + \cdots + y_N) = \frac{1}{N}\sum_{i=1}^{N} y_i$
    - Finite-population variance: $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$
- The variance of the sample mean under SRS ***without replacement*** is ***lower***
  - Since it is $(N-n)/(N-1)$ times that of the the sample mean
    - Of all the observations when the sampling is ***with replacement***
  - ***Without*** replacement
    - $var(\bar{y}) = \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}$
  - ***With*** replacement
    - $var(\bar{y}_n) = \frac{N-1}{nN}\sigma^2 = \left(\frac{N-1}{N}\right)\frac{\sigma^2}{n}$

Simple Random Sampling *with Replacement*

- Unbiased estimate of the variance of $\bar{y}_n$ ***with replacement*** (known)
  - $\widehat{var}(\bar{y}_n) = \frac{s^2}{n}$
    - Sample variance is: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$
- Compared with the unbiased estimate of the variance
  - From SRS ***without*** replacement
    - $\widehat{var}(\bar{y}) = \left(\frac{N-n}{N}\right)\frac{s^2}{n}$
    - Where $\left(\frac{N-n}{N}\right)$ will always be <1
  - Therefore $\widehat{var}(\bar{y})$ ***without replacement*** << $\widehat{var}(\bar{y}_n)$ ***with replacement***

Simple Random Sampling *with Replacement*

- The number of ***distinct*** sampling units contained within a sample
  - Termed the ***effective sample size*** ν
  - Can be used to calculate the sample mean of the ***distinct*** observations
    - $\bar{y}_\nu = \frac{1}{\nu}\sum_{i=1}^{\nu} y_i$
- The estimator $\bar{y}_\nu$ is an ***unbiased estimator*** of the sample mean
  - The variance of $\bar{y}_\nu$ can be shown to be ***less than*** that of $\bar{y}_n$ (the sample ***with*** replacement)
    - However, the ***variance*** will ***never be as small*** as a sample of size n ***without*** replacement

Simple Random Sampling *with Replacement:* Example

- In a SRS with replacement
  - Of nominal sample size $n = 5$
    - Has the following y-values: 2, 4, 0, 4, 5
  - Let's assume that we know one unit was ***sampled twice*** with the value: $y_i = 4$
- The ***estimate of the population mean***, based on the sample mean of five observations is
  - $\bar{y}_n = \frac{2+4+0+4+5}{5} = 3.0$
- The estimate based only on the four ***distinct*** units is
  - $\bar{y}_v = \frac{2+4+0+5}{5} = 2.75$
    - We should expect this to be a ***more precise estimate*** of the true population mean

Theory of Mark-recapture

- Objective
  - To estimate the abundance $\widehat{N}$ of a population
    - Leveraging information from individuals observed on ***multiple*** occasions
- Practice
  - Catch and mark some number individuals
  - Catch some number of individuals during second event
    - Record the number that are marked (i.e. recaptures)
    - Record the number that are unmarked (i.e. new captures)
  - From the number of "recaptures", relative to the number of unmarked individuals
    - Learn about the total population size

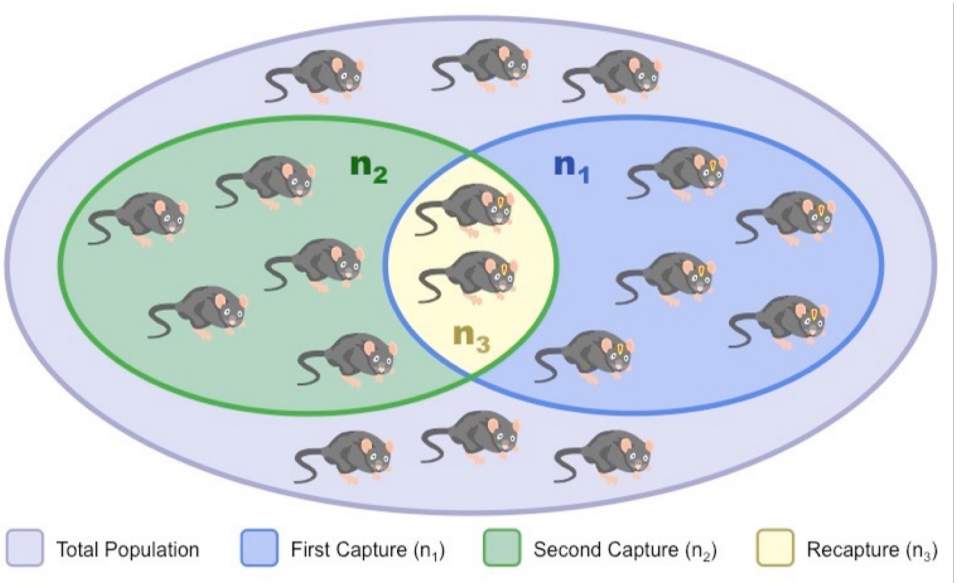


Figure: https://ib.bioninja.com.au/options/option-c-ecology-and-conser/c5-population-ecology/population-sampling.html

Simple Mark-recapture

- "Petersen estimator" or "Lincoln Index" or "Lincoln-Petersen Estimator"
  - Single-release, single-recapture
    - Seber (1982) Ch 3.
- Estimating the number of animals N
  - In a closed population

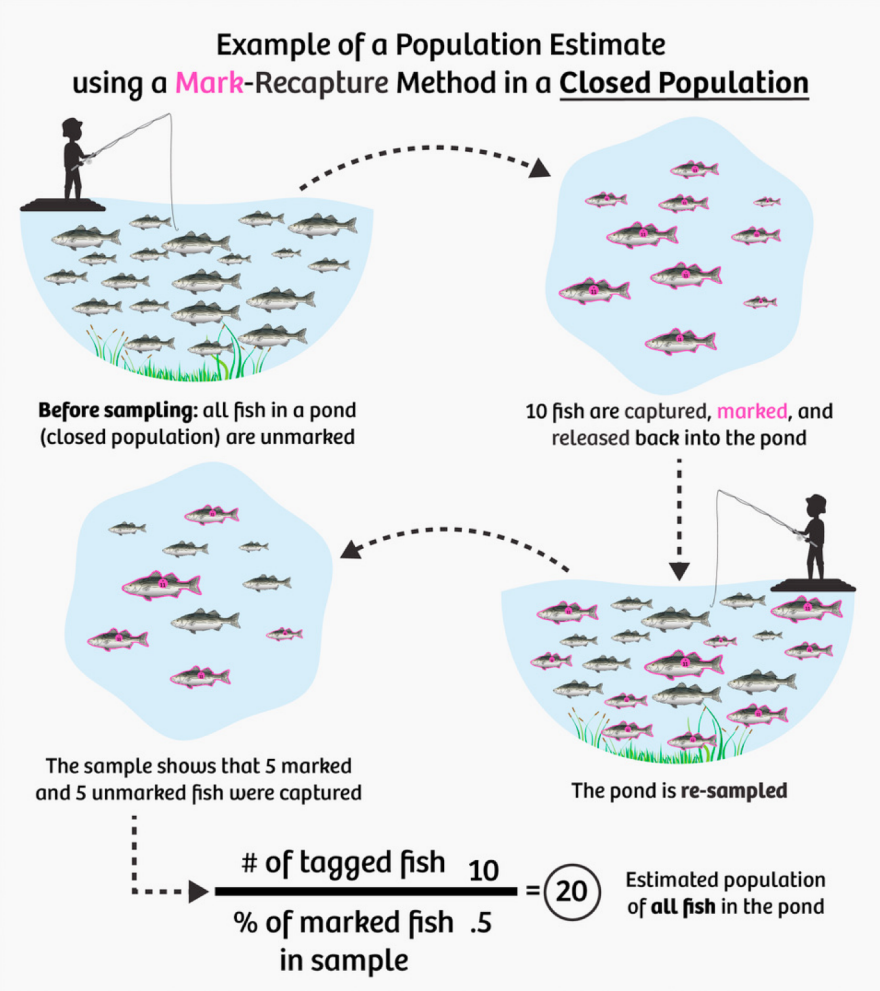


Figure: https://fishbio.com/field-notes/population-dynamics/using-mark-recapture-estimate-population-size

Simple Mark-recapture

- "Petersen estimator"
- Experimental design
  - Catch n_1 animals at time t_1
    - And mark them
  - Catch n_2 animals at time t_2
    - Of which some number m_2 are marked
- Notation
  - $p_1 = n_1/N$
    - Probability of ***capture*** (or proportion caught) at the ***1st time period***
    - Equivalent to ***probability of being marked***
  - $p_2 = n_2/N$
    - Probability of capture at the ***2nd time period***

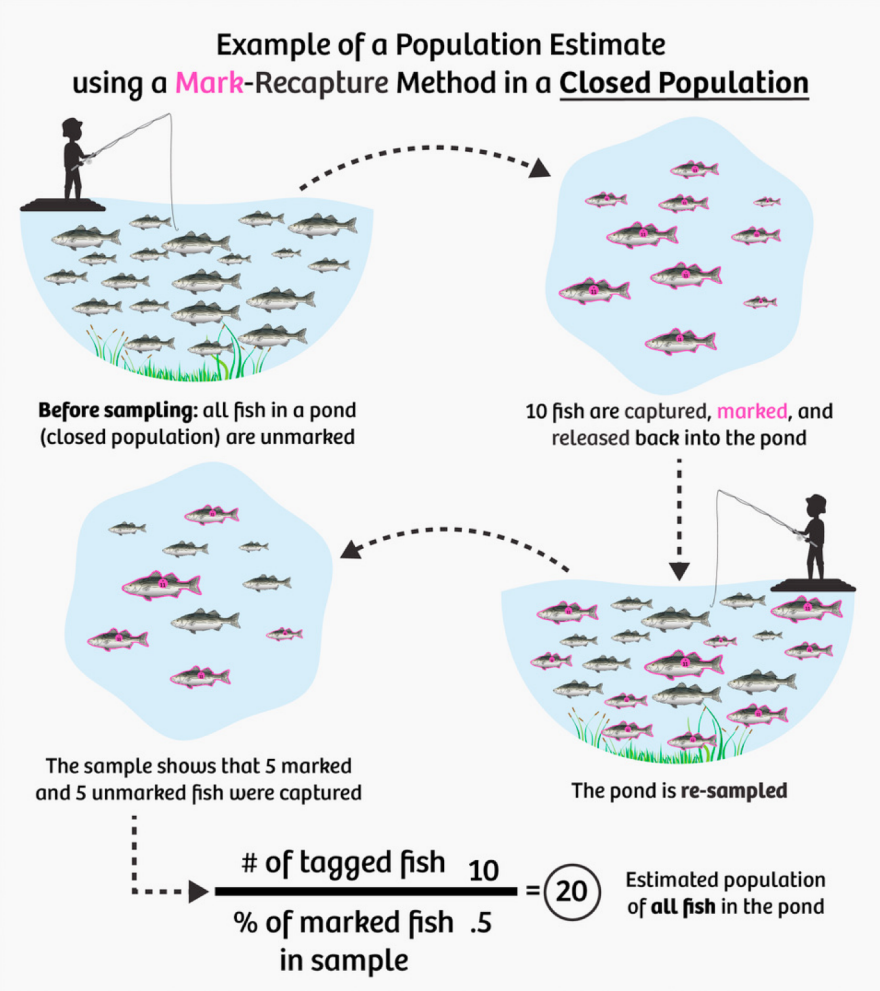
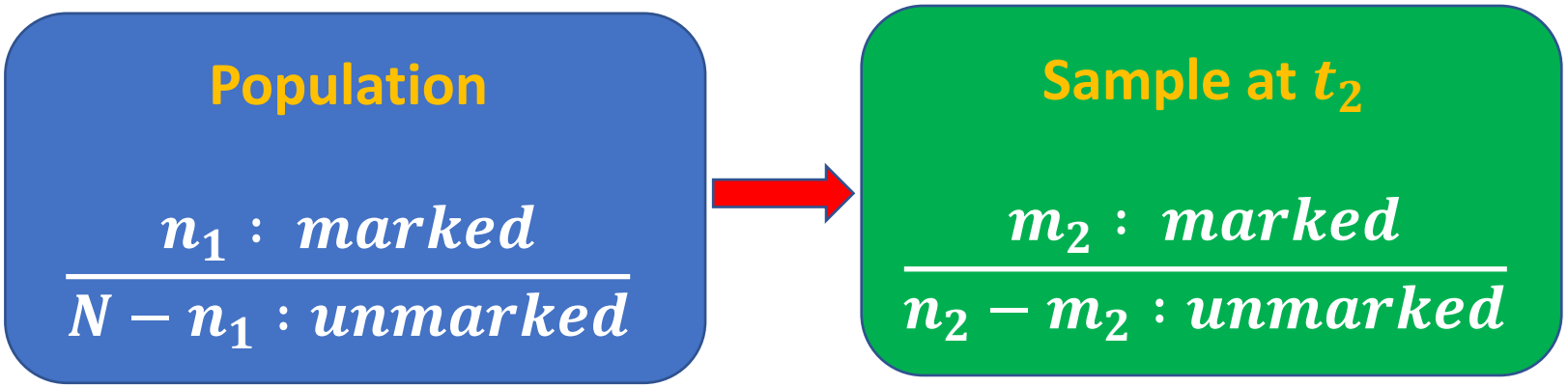


Figure: https://fishbio.com/field-notes/population-dynamics/using-mark-recapture-estimate-population-size

Intuitive Relationship



- $\frac{n_1}{N} = \frac{m_2}{n_2}$
  - The ratio of the number marked individuals n_1 to the total (unknown) population size N
    - Is equal to the ratio of the number of marked individuals m_2 (recaptures) to total individuals caught in t_2
- Petersen Estimator
  - $\widehat{N} = \frac{n_1 n_2}{m_2} = \frac{n_2}{m_2/n_1} = \frac{n_2}{\hat{p}_2} = \frac{\#\ sampled}{est.\ prob.\ of\ being\ sampled}$

Mark-recapture Assumptions

- The population is ***closed***
  - So that N is constant
    - no immigration into or emigration out of the population
- All animals have the same probability of being caught
  - In the first sample
- Marking does not affect the catchability of an animal
- There must be no mortality between the mark (t_1) and recapture (t_2) times
- The second sample is a SRS
  - i.e. each of the $\binom{N}{n_2}$ possible samples
    - has an equal chance of being chosen
- Animals do not lose their marks in the time between samples
  - I.e. no tag loss
- All marks are reported on recovery in the second samples

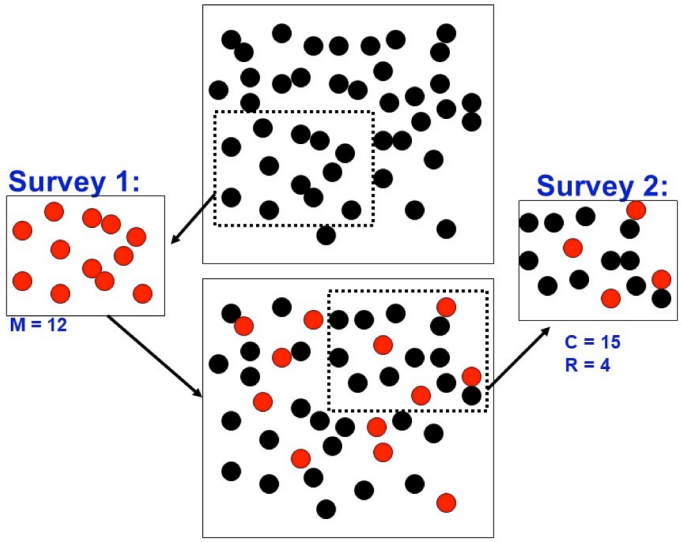


Figure: https://mathbooks.unl.edu/Contemporary/sec-1-1-stat.html

Theory

- Values n_1 and n_2 are fixed by design
  - Therefore m_2 is a random variable
    - When assumptions or simple mark-recapture are satisfied, then the distribution of m_2, given n_1 and n_2, is the geometric distribution
- Geometric distribution
  - $f(m_2|n_1, n_2, N) = \binom{n_1}{m_2}\binom{N-n_1}{n_2-m_2}\Big/\binom{N}{n_2}$
    - Where, "x choose y" is $\binom{x}{y} = \frac{x!}{y!(x-y)!}$
  - Expected value (***unbiased***)
    - $E(m_2|n_1, n_2, N) = \frac{n_1 n_2}{N} = \mu$
  - Variance
    - $var(m_2|n_1, n_2, N) = \frac{n_1 n_2}{N}\left(1 - \frac{n_1}{N}\right)\left(\frac{N-N_2}{N-1}\right)$

Variance of the Hypergeometric

- Variance
  - $var(m_2|n_1, n_2, N) = \frac{n_1 n_2}{N}\left(1 - \frac{n_1}{N}\right)\left(\frac{N-N_2}{N-1}\right)$
    - $= E(m_2|n_1, n_2, N)\left(\frac{N}{N-1}\right)\left(1 - \frac{n_1}{N}\right)\left(1 - \frac{n_2}{N}\right)$
- An approximately ***unbiased*** variance estimator
  - Based only on our estimated or known quantities is
    - $\widehat{var}(m_2|n_1, n_2, N) = m_2 \frac{\widehat{N}}{\widehat{N}-1}\left(1 - \frac{n_1}{\widehat{N}}\right)\left(1 - \frac{n_2}{\widehat{N}}\right)$

Hypergeometric Distribution

- It can be shown that $\widehat{N}$ is the maximum likelihood estimator (MLE)
  - i.e. $f\left(m_2|n_1, n_2, \widehat{N}\right)$ has higher probability than any other value of N
- However, it is ***biased***!
  - Consequence of random variable m_2 appearing in the denominator of
    - $\widehat{N} = \frac{n_1 n_2}{m_2} = \frac{n_2}{m_2/n_1}$
  - A ***small*** value of m_2 will have a ***disproportionately greater*** effect on the abundance estimator than a large value

Chapman Estimator

- Chapman (1951) recommended the alternative estimator
  - Called the ”Chapman estimator”
    - $N^* = \frac{(n_1+1)(n_2+1)}{(m_2+1)} - 1$
- The Chapman estimator is unbiased if $n_1 + n_2 > N$
  - And has small bias otherwise, given by
    - $E(N^*) - N = -Ne^{-(n_1+1)(n_2+1)/N} = -Nb$

Chapman Estimator cont.

- Chapman bias
  - $E(N^*) - N = -Ne^{-(n_1+1)(n_2+1)/N} = -Nb$
- Robson and Regier (1964) noted that
  - If the expected number of recaptures $\mu > 4$
    - Then $b < 2\%$
  - Conversely, if $m_2 > 7$
    - Then there is more than a 95% chance that $\mu > 4$
- Using a Poisson approximation Chapman showed that
  - $var(N^*) \approx N^2 \left(\frac{1}{\mu} + \frac{2}{\mu^2} + \frac{6}{\mu^3}\right)$
  - But, in practice we use the variance estimator
    - $\widehat{var}(N^*) = v^* = \frac{(n_1+1)(n_2+1)(n_1-m_2)(n_2-m_2)}{(m_2+1)^2(m_2+2)}$

"Almost equal to"

Chapman Estimator cont.

- Chapman variance estimator
  - $\widehat{var}(N^*) = v^* = \frac{(n_1+1)(n_2+1)(n_1-m_2)(n_2-m_2)}{(m_2+1)^2(m_2+2)}$
  - This is unbiased if $n_1 + n_2 > N$
    - And has small bias of the order $\mu^2 e^{-\mu}$ otherwise
- Relative variation in the Chapman estimator is measured by the expected coefficient of variation (CV)
  - $CV^* = CV(N^*) = \sqrt{\widehat{var}(N^*)}/N^*$
    - $\approx \frac{1}{\sqrt{\mu}} = \sqrt{N/(n_1 n_2)}$
  - Note that the CV^* ***decreases*** as N ***decreases***, n_1 ***increases***, and/or n_2 ***increases***

Bailey's Binomial Model

- If the experimental design is modified so that sampling is ***with replacement***
  - Then the ***binomial model*** is the appropriate statistical distribution
- $f(m_2|n_1, n_2, N) = \binom{n_2}{m_2} {p_1}^{m_2}(1-p_1)^{n_2-m_2}$
  - Where the probability of being marked is: $p_1 = \frac{n_1}{N}$
    - And $\binom{n_2}{m_2}$ is "n_2 choose m_2"

Bailey's Binomial Model cont.

- $f(m_2|n_1, n_2, N) = \binom{n_2}{m_2} {p_1}^{m_2}(1-p_1)^{n_2-m_2}$
- In this case the number of captures and recaptures
  - During the second time period
    - May include ***repeat captures*** of the same individual
  - If attention were ***restricted*** to ***unique*** individuals
    - Then the hypergeometric would be the correct distribution
- Under Bailey's model
  - The ***expect value*** of m_2 is
    - $E(m_2|n_1, n_2, N) = n_2 p_1 = n_2 \frac{n_1}{N} = \mu$, which is ***unbiased***
  - The ***variance*** of m_2 is
    - $var(m_2|n_1, n_2, N) = n_2 p_1 (1-p_1) = E(m_2|n_1, n_2, N)(1-p_1)$

Bailey's Binomial Model cont.

- It can be shown that $\widehat{N}$ is once again the ***maximum likelihood estimator*** (MLE)
  - But is biased!
- Bailey (1951, 1952) recommended the alternative estimator, the "***Bailey estimator***"
  - $\widehat{N}_1 = \frac{n_1(n_2+1)}{(m_2+1)}$
    - Which has small bias of order $e^{-\mu}$
- The ***variance*** of the Bailey estimator is
  - $\widehat{var}\left(\widehat{N}_1\right) = v_1 = \frac{{n_1}^2(n_2+1)(n_2-m_2)}{(m_2+1)^2(m_2+2)}$
    - With small bias

Bailey's Binomial Model cont.

- Summary
  - Assumption
    - Closed population
  - Most effective with sample sizes smaller than 20
  - Allows individuals to be counted multiple times

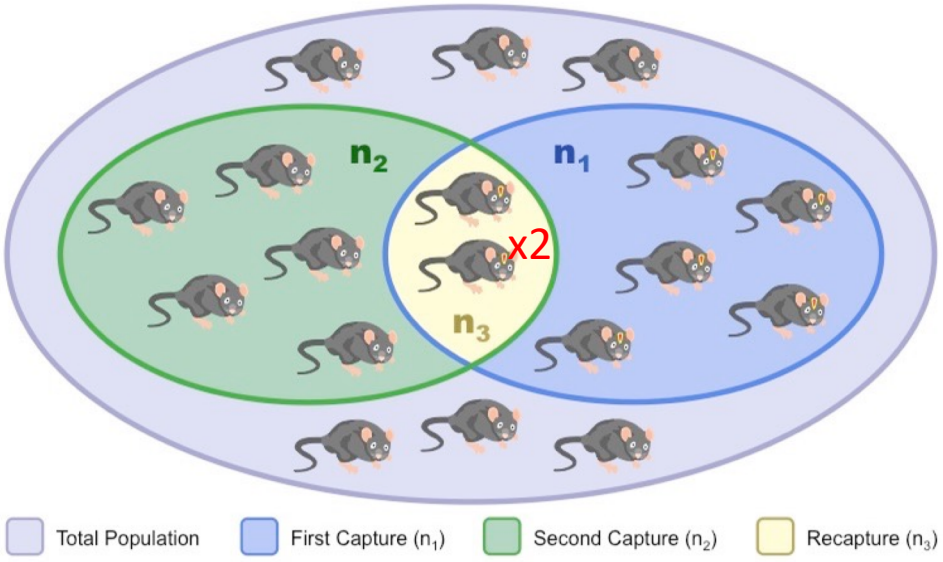


Figure: https://ib.bioninja.com.au/options/option-c-ecology-and-conser/c5-population-ecology/population-sampling.html

Binomial Approximation

- Sometimes the binomial is used as an approximation to the hypergeometric
  - When the sampling fraction $p_2 = n_2/N$ is small
    - Say, less than 5%
  - Although, I see no practical reason for doing so in most contexts
- A more useful (and commonly encountered) situation
  - Is when ***non-random sampling*** occurs
    - Systematic, opportunistic, stratified
- If (1) ***uniform mixing*** of marked and unmarked animals occurs between the two sampling periods
  - And (2) all animals have the ***same probability of capture*** p_2 for all subareas occupied by the population
    - Then the probability of being marked is still $p_1 = n_1/N$
    - And the binomial model still applies

Example: Petersen

- 500 individuals (n_1) are captured at time t_1
  - These individuals are marked and released
- 500 individuals (n_2) are captured at time t_2
  - Of which $m_2 = 100$ of those individuals are marked
- Petersen estimator
  - $\widehat{N} = \frac{n_1 n_2}{m_2} = \frac{500*500}{100} = 2{,}500$
  - $\widehat{N} = \frac{n_2}{m_2/n_1} = \frac{500}{100/500} = 2{,}500$

Example: Chapman

- 500 individuals (n_1) are captured at time t_1
  - These individuals are marked and released
- 500 individuals (n_2) are captured at time t_2
  - Of which $m_2 = 100$ of those individuals are marked
- Chapman estimator
  - $N^* = \frac{(n_1+1)(n_2+1)}{(m_2+1)} - 1 = \frac{(500+1)(500+1)}{(100+1)} - 1 = 2{,}484.2$
  - Variance of the estimator
    - $\widehat{var}(N^*) = v^* = \frac{(n_1+1)(n_2+1)(n_1-m_2)(n_2-m_2)}{(m_2+1)^2(m_2+2)}$
    - $\widehat{var}(N^*) = v^* = \frac{(500+1)(500+1)(500-100)(500-100)}{(100+1)^2(100+2)} = 38{,}596.91$
  - Relative error
    - $CV^* = CV(N^*) = \sqrt{\widehat{var}(N^*)}/N^* = 0.0791 = 7.91\%$

Example: Bailey

- 500 individuals (n_1) are captured at time t_1
  - These individuals are marked and released
- 500 individuals (n_2) are captured at time t_2 (***with replacement***)
  - Of which $m_2 = 100$ of those individuals are marked
    - ***May include resampling of marked individuals!***
- Bailey estimator
  - $\widehat{N}_1 = \frac{n_1(n_2+1)}{(m_2+1)} = \frac{500(500+1)}{(100+1)} = 2{,}480.2$
  - The ***variance*** of the Bailey estimator is
    - $\widehat{var}(\widehat{N}_1) = v_1 = \frac{{n_1}^2(n_2+1)(n_2-m_2)}{(m_2+1)^2(m_2+2)}$
    - $\widehat{var}(\widehat{N}_1) = v_1 = \frac{500^2(500+1)(500-100)}{(100+1)^2(100+2)} = 48{,}149.8$