How many are out there? A novel approach for open and closed systems

Zia Rehman
University of Central Florida

Find similar works at: http://stars.library.ucf.edu/etd

University of Central Florida Libraries http://library.ucf.edu

STARS Citation

Rehman, Zia, 'How many are out there? A novel approach for open and closed systems' (2014). Electronic Theses and Dissertations. 4569.
http://stars.library.ucf.edu/etd/4569

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of STARS. For more information, please contact lee.dotson@ucf.edu.
HOW MANY ARE OUT THERE?
A NOVEL APPROACH FOR OPEN AND CLOSED SYSTEMS

by

ZIA REHMAN

M.A. University of Louisville, 2001

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Modeling and Simulation (Statistics)
in the Institute for Simulation and Training
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Summer Term
2014

Major Professor: Peter Kincaid
ABSTRACT

We propose a ratio estimator to determine population estimates using capture-recapture sampling. It’s different than traditional approaches in the following ways:

(1) Ordering of recaptures: Currently data sets do not take into account the "ordering" of the recaptures, although this crucial information is available to them at no cost. It is being simply ignored in the estimation process. For example, a system with a small population will have a greater frequency of recaptures “early” than an identical system with a larger population. Our estimator accounts for ordering and we will demonstrate through simulations and comparison with existing estimators.

(2) Dependence of trials and cluster sampling: Our model explicitly considers trials to be dependent and improves existing literature which assumes independence. Thus cluster sampling is handled in a natural manner in our paper as opposed to adjustments made “after the fact” in the existing literature. This is demonstrated throughout with mathematical development and also simulated results using cluster sampling.

(3) Rate of convergence: The percentage sampled has an inverse relationship with population size, for a chosen degree of accuracy. We demonstrate this with simulated data. This makes the estimator very useful for larger populations.

(4) Asymptotic Attainment of Minimum Variance (Open Systems): We provide mathematical proof that our estimator attains asymptotic minimum variance (=population variance). We test this attainment through simulations. Additionally, a formula for population variance provides insights into the underlying system variance.
(5) **Full use of data and model applicability:** Information gathered on repeated recaptures of the same unit is used in the estimation process and thus our ratio estimator is equally efficient in “data utilization”. This is seen in the data section of the paper. Additionally, we show application to a wide range of problems. This is shown through examples although no real world data is used to support this claim.

(6) **Non-parametric:** It is a non parametric approach and this not only means that we avoid over-parameterization but also that model will not produce non zero birth/death rates in closed systems. The theoretical development of the model is non parametric and we support this claim through mathematical derivations.

(7) **Heterogeneity:** We allow for implicit heterogeneity when units being sampled are hard to identify. We provide the theoretical extension for heterogeneity but do not test this empirically.

(8) **Open and closed systems:** Simpler results are presented separately for closed systems. Open systems allows for all possibilities including births, deaths, exit, re-entry. The results are proved mathematically and tested through simulations.

(9) **Robustness to assumptions in open systems:** The estimator is shown to be robust under widely changing conditions in open systems. The results are tested through simulations.

Given the above properties we make the case that our estimator is not "MT estimator", such as the one suggested by Schnabel in 1938. This is spite of the fact that it is a ratio estimator.

The estimator is asymptotically unbiased and we show the confidence intervals for the underlying known population. We present computer simulated results that validate the theoretical results. We also compare the results with existing models.
We prove interesting asymptotic results for open systems: estimator variance converges almost surely to population variance; point estimator converges almost surely to population parameter.
ACKNOWLEDGMENTS

We are indebted to Dr. Sat Gupta (Mathematics and Statistics department, University of North Carolina at Greensboro) for providing technical feedback and extremely valuable suggestions on closed systems. The authors would like to thank Dr. Jan Rychtář (Mathematics and Statistics department, University of North Carolina at Greensboro) for providing modeling and simulation expertise on closed systems and that greatly facilitated the development of the paper. We are grateful to Dr. Mark Johnson (Statistics department, University of Central Florida) for providing the closed system problem. Indeed the early work in the paper was motivated by a solution to this problem. The authors are grateful to an anonymous scholar for very insightful modeling and statistical expertise.

The dissertation committee members, namely, Dr. Paul Wiegand, Dr. Nizam Uddin, Dr. Craig Finch and Dr. Peter Kincaid provided invaluable mathematical insights and suggestions that led to a polished and improved manuscript. We are very grateful for their contributions. Christina Toms (Biology department, University of Central Florida) provided ecological expertise and insights that were very useful. Thank you, Christina.
# TABLE OF CONTENTS

LITERATURE REVIEW ........................................................................................................... 1

The Challenge of Multiple Mark Releases ........................................................................ 1

Ordering of Trials .................................................................................................................. 2

Dependence of Trials (Over Dispersion) ............................................................................ 2

Over Parameterization ........................................................................................................ 3

Historical Development ...................................................................................................... 3

Illustration: Insurance ......................................................................................................... 5

Illustration: Epidemiology .................................................................................................. 6

NOTATION & DATA .............................................................................................................. 7

Schematic Flow: .................................................................................................................. 8

Capture History as an Ordered String ............................................................................... 8

Miscellaneous ..................................................................................................................... 9

CLOSED SYSTEM SAMPLING – DRAW BY DRAW WITH REPLACEMENT .................... 10

Estimator ............................................................................................................................. 10

Unbiased estimator .............................................................................................................. 10

Variance Estimation .......................................................................................................... 11

Covariance Estimation: $\text{cov}(x_i, x)$ ............................................................................. 12

Covariance Estimation: $\text{Cov}(\bar{x}, \bar{X})$ ........................................................................ 14
Estimation of variance of average captures................................................................. 36

Estimation of covariance of average marked units in system and captures ....................... 37

Population changes approach zero .................................................................................. 40

Open system as time continuous process ......................................................................... 41

Comparison of open versus closed systems ..................................................................... 42

Rate of convergence ........................................................................................................ 43

GENERAL DISCUSSION .................................................................................................... 46

OPEN SYSTEM SIMULATION UNDER CLUSTER SAMPLING ..................................... 47

   Robustness..................................................................................................................... 47

   Open System Rate of Convergence ............................................................................. 52

   Comparison of variance of the estimator: formula versus simulated ......................... 55

SAMPLING UNDER IMPLICIT HETEROGENEITY ......................................................... 57

   Explicit & implicit heterogeneity .................................................................................. 57

   Complete capture history ............................................................................................ 57

   Multinomial likelihood model ..................................................................................... 58

RESULTS .......................................................................................................................... 59

   Theorem 1: Finite population closed system ............................................................... 59

   Theorem 2: Finite population open system (estimator) .............................................. 60

   Theorem 3: Finite population open system (variance) ................................................ 61

CONCLUSION ................................................................................................................. 63
REFERENCES ................................................................................................................................. 64

APPENDIX 1: COMPARISON WITH EXISTING MODELS .......................................................... 67
LIST OF FIGURES

Figure 1: Estimation for a closed system (Population Size 1,000) .................................................. 18
Figure 2: Estimation for a closed system (Population Size 10,000) ................................................. 19
Figure 3: Estimation for a closed system (Population Size 100,000) ............................................... 19
Figure 4: Estimation for a closed system (Population Size 1,000,000) .......................................... 20
Figure 5: A Random Selection of (n, N) Involving 250 Trajectories ............................................ 48
Figure 6: Estimation for a Certain Trajectory (Population Size 1,000) ........................................ 49
Figure 7: Estimation for a Certain Trajectory (Population Size 10,000) ...................................... 50
Figure 8: Estimation for a Certain Trajectory (Population Size 100,000) .................................... 51
Figure 9: Estimation for a Certain Trajectory (Population Size 1,000,000) ................................ 52
Figure 10: Rate of Convergence for a Certain Trajectory (2,000 encounters) ............................ 53
Figure 11: Rate of Convergence for a Certain Trajectory (5,000 encounters) .............................. 54
Figure 12: Rate of Convergence for a Certain Trajectory (10,000 encounters) ......................... 55
LIST OF TABLES

Table 1: Data table under cluster sampling................................................................. 23
Table 2: Estimated standard deviation........................................................................ 56
LITERATURE REVIEW

The Challenge of Multiple Mark Releases

Mark release is a method that is widely used in fields such as ecology. The approach is based on capturing an animal, marking it and then releasing it back in the environment. The trials continue until the marked animals are recaptured and are tagged again. With this data set, it is possible to estimate the size of the population and its confidence interval. The basic premise is that number of recaptures reflects the size of the population since in a small closed system there would be more recaptures with the same number of trials.

There are different names to this approach such as capture-recapture, capture-mark-recapture, mark-recapture, sight-resight, mark-release-recapture, multiple systems estimation. The approach is also commonly used in epidemiology to determine the use of services such as learning disabilities, services for elderly or number of people infected with HIV, drug addicts etc.

Capture recapture methods are usually implemented using cluster trials – more than one unit is sampled at a time. To illustrate, suppose that series of \( k \) independent samples containing set of numbers \( y_1, y_2 \ldots y_k \) of units captured from a closed population (Schnabel, 1938). With the \( ith \) sample the number of \( x_i \) of marked units is observed then the animals are given new marks before release (note that \( x_1 = 0 \)). Then with total number of ways of choosing \( x_i \) out of \( X_i \) being \( \binom{X_i}{x_i} \)

\[
f(x_1, x_2 \ldots x_k) = \prod_{i=2}^{k} \binom{X_i}{x_i} \left( \frac{\tau - X_i}{y_i - x_i} \right)^{x_i \choose y_i}
\]
Where \( X_i \) is the number marked in the population when (just before) the \( ith \) sample is taken. The value of \( \tau \) that maximizes the probability above is the maximum likelihood estimate of the population size. The solution must be obtained iteratively.

While appealing in simplicity and asymptotic properties of the estimator such as minimum variance, the model is limited to closed systems and the estimator must be determined by trial and error.

**Ordering of Trials**

There is gap in the literature for a model that explicitly accounts for ordering of recaptures. For example suppose we label a capture for a unit as 0 and no recapture as 1. In this case, the two sampling histories such as 10001 and 11000 would be different in our case but treated identically in the current literature. This is a very relevant issue because early recaptures are a sign of a small population, all else being equal.

**Dependence of Trials (Over Dispersion)**

There is gap in the literature for a model that explicitly accounts for dependence of trials. The multinomial maximum likelihood approach exploits independence of trials and this becomes an issue when cluster sampling is encountered as in that case units caught within a cluster trial cannot be treated as independent. The “over dispersion” adjustments are to correct for this effect. Our approach explicitly deals with dependence of trials and hence such adjustments are not necessary.
**Over Parameterization**

After the pioneering paper by Jolly (1965) and Seber (1965), considerable amount of ecological literature has focused on extending the model to account for implicit heterogeneity or other modern sampling issues found in animal populations. One such paper Mc Clintock (2009) provides a similar structure by setting up a maximum likelihood functions and maximizing eit with respect to related parameters. They cite issues arising by making assumptions about constant survival probabilities and the complexity of the likelihood function. These issues are typical in such models and arise because each sampling nuance leads to increasing number of parameters being introduced and these in turn may vary on space/time – something that is very hard to incorporate. While statistically powerful, the drawback can be an overly parameterized model with too many parameters and unintended consequences such as positive birth/death rates in closed systems where this is impossible. Our approach is non parametric and provides abundance estimates along with confidence intervals.

Ecologists currently rely on model fitting techniques and do not depend on a single model. In this context, our model is intended to supplement existing literature and provide a choice to the researcher.

**Historical Development**

Petersen (1896) and Lincoln (1930) obtained a ratio-based estimator for a closed homogeneous population size based on data obtained in two samples. We present a model for both open and closed systems as well as heterogeneous population. The articles refer to the term “heterogeneity” to mean a number of situations: units cannot be identified accurately (such as male versus female), dependence in sampling (cluster sampling), the assumption that all units
have the same probability of capture. Schnabel (1938) generalized this result for multiple
samples. Jolly (1965) and Seber (1965) obtained a likelihood-based estimator for an open
homogeneous population at time t. All these latter papers could perform on more complex
situations but suffered from assumptions such as dependence and equal capture probabilities.

Pollock (1981) extended this work to allow observable heterogeneity. However these later
models continued to be based on maximum likelihood estimation and were therefore
parameterized. Our approach is specifically non parametric and thus carries less assumptions
such as dependence. Model selection is another example and we do not require this process due
to non-parametric modeling.

Crosby and Manly (1985) began to consider model selection, while Buckland et. al (1997) gave a
comprehensive discussion of model selection. Brownie and Pollock (1985) demonstrated a
unified likelihood-based approach for band-recovery models, while Self and Liang (1987)
produced asymptotic results under capture-recapture's non-standard conditions.

Apart from these standard approaches, several different methods have been utilized to obtain
population size estimates. For example, Cormack (1989) utilized log-linear models, Smith
(1988) and others used Bayesian methods, and Burnham and Overton (1978) employed the
jackknife. These are more intriguing but nonetheless parametric in approach. The Bayesian
approach requires specification of the prior and can be a practical problem.

Much of the pre-1990 capture-recapture literature is summarized by the review articles of Otis et
al (1978) for a closed population, and by Pollock et. al (1990), Seber (1982), Seber (1992) and
Lebreton et. al. (1992) for both open and closed systems. The 1990's and beyond has produced
substantial research which allows likelihood-based modeling of non-observable heterogeneity.
The issue of non-observable heterogeneity with respect to dependence has been addressed in our paper and consideration of the ordering of the sample is important in this respect.

In particular, Norris and Pollock (1996), Pledger (2000), and Dorazio and Royle (2003) performed maximum likelihood analysis under non-observable heterogeneity for a closed population, while Pledger et. al. (2003) and Pledger et. al. (2010) constructed such analysis for an open population.

Our approach is substantially different as its based on a simple ratio estimator. In this context, we can consider it as an extension to the work of Schnabel (1938) ratio estimator (for a closed population) and then amend it for an open population. Mathematical properties and asymptotic of the estimators are emphasized, with detailed simulations supporting their results. Also, we discuss some modifications for this estimator when there is non-observable heterogeneity.

Appendix 1 compares our approach with existing models.

**Illustration: Insurance**

Suppose we want to know how many college educated, male drivers, ages 24-28, with no claims in the last 5 years are out there in Guilford County, NC. Such an arbitrary and “custom” survey is of course not available. To find the answer, we use data from people obtaining auto insurance quotes. The first time a person gets a quote, the event is called “capture”. If the person does not buy insurance and returns again for a quote after some time has elapsed, they can be considered as a “recapture. Using this dataset, we can then determine the answer.
Illustration: Epidemiology

Incidence of disease: Suppose a person was tested negative for a disease (capture) and they come back randomly for a second test (recapture). Assuming that disease can reoccur, we can determine the total count of people “exposed”. The disease could be due to person’s lifestyle and thus the captures/recaptures are not random in the general population. However, it would still be random in the sub population of people with that lifestyle and thus the exposure will be measured with reference to “subset of people with this lifestyle”. For example blood cancer is not due to lifestyle and its measurement gives total cancer exposure in the population. Lung cancer caused by smoking is due to lifestyle and its incident will be given among smoking population.

Spread of infection: Suppose a person gets cured for a disease (capture) and get the disease again (recapture). This tells us how many people have the disease. The above comments about counts with reference to a sub population apply equally in this case. Thus randomization takes place within this subgroup.

In both cases, all counts are given in certain geography and thus it is useful information for doctors and administrators/hospitals as they will know the incidence or spread of disease in their geographical area.
NOTATION & DATA

We use mathematical “operator” notation instead of functional notation. Thus $E_x$ refers to $E(x)$ and $Var_x$ is same as $Var(x)$ . We use the abbreviation “(a.s)” as a short form of “almost surely”.

Next, all captures/recaptures are Bernoulli trials with each trial resulting in capture of a single unit. This is relaxed in cluster sampling situation (open systems) and we let the reader know accordingly. Thus each trial is a simple random sample of one unit.

$p_i = \text{Probability of catching a marked unit in } \text{ith recapture (computed at end of ith trial)}$;

$i = 1 \ldots n$

$X_i = \text{Random # of marked unit at end of ith recapture}$

$X_o = \text{Initial known # of marked units at time zero}$

$\tau = \text{Total population to be estimated}$

Then

$p_i = \frac{X_i}{\tau}$ with $X_i$ random and $\tau$ fixed. With $x_i = \text{bernoulli}(p_i)$ trials,

$E x_i = p_i; Var x_i = p_i(1 - p_i); E X = \bar{p}$

Using our definitions and non iid version of Strong Law of Large Numbers (SLLN),

$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \overset{\text{SLLN}}{\rightarrow} E x_1 + x_2 + \ldots + x_n = \frac{p_1 + p_2 + \ldots + p_n}{n} = \bar{p}$
Schematic Flow:

All values are reported at the end of the period. For example for the $i$th interval, the data $x_i$ is collected throughout the interval but we assume that values are reported at the end of the interval. This assumption is made out of mathematical convenience only.

The results presented are asymptotic. At the same time the populations are finite. Thus $n \rightarrow \infty$ means sampling all population units in the system.

Capture History as an Ordered String

One the key aspect of the model is to capture the sequence or ordering of the capture-recapture history. Assuming no implicit heterogeneity issues, we do not care about “unique” histories and their frequencies and instead focus on the collective sampling history. This is different than current literature that relies on unique capture history of each unit, regardless of heterogeneity. In the case of heterogeneity, our model also requires these capture histories of each unit to be recorded. Thus crucial (and more) information in the form of ordering of data string is used by our model than existing literature.

To illustrate, suppose we assign 0 to a capture and 1 to a recapture (first or subsequent). Then the entire sampling history of 20 encounters could be denoted by a single string such as:

$x_i : 000101010011000101$

$X_i : 1,2,3,3,4,5,5,6,7,7,8,8,8,9,10,11,11,12,12$
Note that ordering is relevant as $\sum X_i$ will change on a different ordering. Thus permutations on this string will produce different estimates of the population. We will show later that it also produces different rates of convergence and confidence intervals of the estimator.

**Miscellaneous**

The data set requirements under implicit heterogeneity (sampled units are indistinguishable) is discussed later in the paper.
CLOSED SYSTEM SAMPLING – DRAW BY DRAW WITH REPLACEMENT

Estimator

The probability of recapture is given by \( p_i = \frac{X_i}{\tau} \Rightarrow \tau = \frac{X_i}{p_i} \). If the experiment was done \( i = 1 \ldots n \) times, an estimator for population could be \( \tau = \frac{x}{\sum x_i} = \frac{\sum X_i}{\sum x_i} \). This follows since \( p = x \). Next, fix \( \tau \) and let the sample be denoted by \( \Omega \). We propose the estimator,

\[
\hat{\tau} = \frac{\sum X_i}{\sum x_i}
\]  
(1)

Consider a closed, fixed population system where the total population \( \tau \) is fixed and unknown. The estimator implies the average proportion assumption: the probability of catching any unit among the total population is the same as the average probability of catching a marked unit among marked population:

\[
\frac{\sum_{i=1}^{i=n} X_i}{\sum_{i=1}^{i=n} x_i} = \frac{x}{\sum x_i} = \frac{1}{N}
\]

Unbiased estimator

\[
\hat{\tau} \bigg| _{\Omega}^{(a.s)} \rightarrow \tau
\]  
(2)

\[
E \hat{\tau} \rightarrow \tau
\]  
(3)
Proof:

Due to Strong Law of Large Numbers (SLLN),

\[ \hat{\tau} \left| \Omega = \frac{\tilde{X}}{X} \rightarrow \frac{\tilde{X}}{\bar{x}} = \tau \]  

Due to Dominated Convergence Theorem (DCT) expectation \( \hat{\tau} \) will also converge almost surely as long as \( |\tau| < \infty \). Thus,

\[ E \left| \hat{\tau} \right| \Omega = E \hat{\tau} \rightarrow \tau \]

**Variance Estimation**

The variance of the estimator \( \hat{\tau} \) requires computation of \( \text{Var} \tilde{X}, \text{Var} \bar{x}, \text{Cov}(\tilde{X}, \bar{x}) \) and we do them in turn. These estimates are then combined into a single formula for estimator variance.

The number of marked units at the end of time \( i - 1 \) (beginning of time \( i \)) is \( X_{i-1} \). At the end of time \( i \) the marked population is updated due to \( x_i \) recaptures and the population at the end of time \( i \) is \( X_i \). If there is a recapture \( x_i = 1 \) else \( x_i = 0 \). Formally,
\[ X_i = X_{i-1} + 1 - x_i; i = 1 \ldots k; \]
\[ X_k = X_o + k - \sum_{i=1}^{k} x_i \]
\[ \sum_{k=1}^{n} X_k = nX_o + \frac{n(n+1)}{2} - \sum_{i=1}^{k} x_i \]
\[ \bar{X} = X_o + \frac{(n+1)}{2} - \frac{1}{n} \sum_{i=1}^{k} x_i \]
\[ \bar{X} = X_o + \frac{(n+1)}{2} - \frac{1}{n} [nx_i + (n-1)x_2 + \ldots + x_n] \]

\[
\begin{align*}
\text{Var} \bar{X} & = \frac{1}{n^2} \left\{ \sum_{i=1}^{n} (n-i+1)^2 p_i (1-p_i) + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} \text{cov} \left[ (n-i+1)x_i, (n-j+1)x_j \right] \right\} \\
E \bar{X} & = E \frac{1}{n} \sum x_i = \frac{1}{n} \sum Ex_i = \frac{1}{n} \sum p_i = \bar{p} \\
\text{Var} \bar{x} & = \frac{1}{n^2} \text{Var} \sum_{i=1}^{n} x_i = \frac{1}{n^2} \left\{ \sum_{i=1}^{n} p_i (1-p_i) + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} \text{cov}(x_i, x_j) \right\} 
\end{align*}
\]

**Covariance Estimation:** \( \text{cov}(x_i, x_j) \)

Assume \( i < j \). Then

\[
\begin{align*}
\text{cov}(x_i, x_j) & = Ex_i x_j - Ex_i Ex_j \\
& = P(x_i = 1, x_j = 1) - p_i p_j \\
& = P(x_j = 1 | x_i = 1)P(x_i = 1) - p_i p_j \\
& = p_{j-1} - p_j p_j \;
\end{align*}
\]

The second to last line follows because while \( \text{cov}(x_i, x_j) \) here one here one less trial is needed as we know population was unchanged at \( ith \) trial. Thus for \( i < j \) \( P(x_j = 1 | x_i = 1) = p_{j-1} \)
Note that since \( p_j \) are monotonically increasing, \( \text{cov}(x_i, x_j) \leq 0 \) with strict equality if and only if \( p_{j+1} = p_j \). In other words, successive captures of marked units are uncorrelated; and have negative covariance otherwise. Note that the mappings \( F_{ij} \) and \( F_{ji} \) yield identical values and thus the covariance function are symmetric:

\[
(i < j): F_{ij} : (x_i, x_j) \rightarrow \text{cov}(x_i, x_j) = p_i(p_j - p_i)
\]

\[
(i > j): F_{ji} : (x_j, x_i) \rightarrow \text{cov}(x_j, x_i) = p_j(p_j - p_i)
\]

Proceeding,

\[
\text{Var} \bar{X} = \frac{1}{n^2} \left( \sum_{i=1}^{n} (n-i+1)^2 p_i (1-p_i) + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} \text{cov}(x_i, x_j) [(n-i+1)(n-j+1)] \right)
\]

\[
\text{Var} \bar{X} = \frac{1}{n^2} \left( \sum_{i=1}^{n} (n-i+1)^2 p_i (1-p_i) + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} p_j(p_j - p_i) [(n-i+1)(n-j+1)] \right)
\]

(4)

\[
\text{Var} x = \frac{1}{n^2} \text{Var} \sum_{i=1}^{n} x_i = \frac{1}{n^2} \left( \sum_{i=1}^{n} p_i(1-p_i) + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} \text{cov}(x_i, x_j) \right)
\]

\[
\text{Var} x = \frac{1}{n^2} \text{Var} \sum_{i=1}^{n} x_i = \frac{1}{n^2} \left( \sum_{i=1}^{n} p_i(1-p_i) + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} p_j(p_j - p_i) \right)
\]

(5)
Covariance Estimation: \( \text{Cov}(\bar{X}, \bar{X}) \)

\[
X_i = X_{i-1} + 1 - x_i
\]

\[
X_i - X_{i-1} = 1 - x_i
\]

\[
\sum_{i=1}^{n} (X_i - X_{i-1}) = \sum_{i=1}^{n} (1 - x_i)
\]

\[
\frac{X_n - X_0}{n} = 1 - x
\]

\[
X_n = n(1 - x) + X_0
\]

\[
\bar{x} = 1 - \frac{X_n - X_0}{n}
\]

\[
X_k = k \left( 1 - \frac{\sum_{i=1}^{k} x_i}{k} \right) + X_0 ; k \leq n
\]

\[
\text{Cov}(\bar{X}, \bar{X}) = \text{Cov}(\bar{X}, 1 + \frac{(X_0 - X_n)}{n}) = -\text{Cov} \left( \frac{X_n}{n}, \sum_{k=1}^{n} X_k \right) = -\frac{1}{n^2} \text{Cov} \left( X_n, \sum_{k=1}^{n} X_k \right)
\]

\[
= -\frac{1}{n^2} \text{Cov} \left( n(1-x) + X_0, nX_0 + \sum_{k=1}^{n} k \left( 1 - \frac{\sum_{i=1}^{k} x_i}{k} \right) \right) = -\frac{1}{n^2} \text{Cov} \left( n\bar{x}, \sum_{k=1}^{n} x_1 + x_2 + ... x_k \right)
\]

\[
= -\frac{1}{n^2} \text{Cov} \left( (x_1 + x_2 + ... x_n), \sum_{k=1}^{n} x_1 + x_2 + ... x_k \right)
\]

\[
= -\frac{1}{n^2} \text{Cov} \left( (x_1 + x_2 + ... x_n), [nx_1 + (n-1)x_2 + ... + x_n] \right)
\]
\[ = - \frac{1}{n^2} \text{Cov} \left( \sum_{i=1}^{n} x_i, \sum_{i=1}^{n} (n-i+1)x_i \right) \]
\[ = - \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{j} \text{cov} \left( x_i, (n-j+1)x_j \right) \]
\[ = - \frac{1}{n^2} \left\{ \sum_{i=1}^{n} \text{Cov} \left( x_i, (n-i+1)x_i \right) + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} \text{Cov} \left( x_i, (n-j+1)x_j \right) \right\} \]
\[ = - \frac{1}{n^2} \left\{ \sum_{i=1}^{n} (n-i+1) \text{Cov} \left( x_i, x_i \right) + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} (n-j+1) \text{Cov} \left( x_i, x_j \right) \right\} \]
\[ \text{Cov} \left( \bar{x}, X \right) = - \frac{1}{n^2} \left\{ \sum_{i=1}^{n} (n-i+1) \text{Var} x_i + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} (n-j+1) p_i (p_{j-1} - p_j) \right\} \]

**Estimator Variance**

\[ \hat{\text{Var}} \tau = \text{Var} \left( \frac{\sum X_i}{\sum x_i} \right) = \text{var} \frac{\bar{X}}{\bar{x}} \]

Using first order Taylor approximation, where we choose to ignore \((\bar{X} - E \bar{X})^k; (\bar{x} - E \bar{x})^k; k \geq 2\) due to Strong Law of Large Numbers (SLLN),

\[ \hat{\tau} \rightarrow \frac{E \bar{X}}{E \bar{x}} + \frac{1}{E \bar{x}} (\bar{X} - E \bar{X}) - \frac{E \bar{X}}{E \bar{x}^2} (\bar{x} - E \bar{x}) \]

\[ \hat{\text{Var}} \tau \rightarrow \frac{1}{E \bar{x}} \text{Var} \frac{\bar{X} + (E \bar{X})^2}{E \bar{x}} \frac{\text{Var} x}{E \bar{x}^3} - 2 \frac{E \bar{X}}{E \bar{x}^3} \text{Cov} (\bar{X}, x) \]

\[ \hat{\text{Var}} \tau \rightarrow \frac{1}{p} \text{Var} \frac{\bar{X} + (E \bar{X})^2}{p} \frac{\text{Var} x}{p^3} - 2 \frac{E \bar{X}}{p^3} \text{Cov} (\bar{X}, x) \]
Var \hat{\tau} \rightarrow \frac{\text{Var} \bar{X} + \tau^2 \text{Var} \bar{x} - 2\tau \text{Cov}(\bar{X}, \bar{x})}{\left(\frac{\bar{p}}{p}\right)^2}

So far we have

\begin{align*}
\tau^2 \text{Var} \hat{\bar{x}} &= \frac{1}{n^2} \left\{ \sum_{i=1}^{n} \tau^2 p_i (1 - p_i) + 2 \sum_{j=2}^{n-1} \tau^2 \sum_{i=1}^{j-1} p_i (p_{j-1} - p_j) \right\} \\
-2\tau \text{Cov}(\hat{\bar{x}}, \hat{x}) &= \frac{1}{n^2} \left\{ \sum_{i=1}^{n} 2\tau (n - i + 1) p_i (1 - p_i) + 2 \sum_{j=2}^{n-1} \sum_{i=1}^{j-1} 2\tau (n - j + 1) p_i (p_{j-1} - p_j) \right\} \\
\text{Var} \hat{\bar{X}} &= \frac{1}{n^2} \left\{ \sum_{i=1}^{n} (n - i + 1)^2 p_i (1 - p_i) + 2 \sum_{j=2}^{n-1} \sum_{i=1}^{j-1} p_i (p_{j-1} - p_j) \left[ (n - i + 1)(n - j + 1) \right] \right\}
\end{align*}

Therefore,

\begin{align*}
\text{Var} \hat{\tau} &\rightarrow \\
\sum_{i=1}^{n} \left[ (n - i + 1)^2 + \tau^2 + 2\tau (n - i + 1) \right] p_i (1 - p_i) + 2 \sum_{j=2}^{n-1} \sum_{i=1}^{j-1} p_i (p_{j-1} - p_j) \left[ \tau^2 + (n - i + 1)(n - j + 1) + 2\tau (n - j + 1) \right] \\
&\quad \div \left( \sum_{i=1}^{n} p_i \right)^2
\end{align*}

\begin{align*}
\hat{p}_i &= \frac{\hat{X}_i}{\hat{\tau}}; \quad \hat{\tau} = \frac{\sum \hat{X}_i}{\sum \hat{x}_i}
\end{align*}

Confidence Interval

The estimated confidence interval for \( \tau \) based on estimated variance \( \text{Var} \hat{\tau} \)

\begin{align*}
\text{CI}(\tau) \rightarrow \frac{\bar{X}}{\bar{x}} \pm z_{\alpha/2} \sqrt{\text{Var} \hat{\tau}}
\end{align*}
Formula Insights

(1) The first portion of the variance formula $\hat{\tau}$ relates with variances of the terms

$$\hat{\tau} = \frac{\bar{X}}{x}$$

and the second portion with covariance. The presence of the term $\text{cov}(x_i, x_j) = p_i(p_{j-1} - p_j)$ prohibits symmetry in that it does not allow terms to be combined into a single double summation (or substitution of $i = j$ in the covariance part yields the variance part). The covariance portion of formula reflects dependence in trials as in this case $\text{cov}(x_i, x_j) \leq 0$.

(2) The information about the “ordering” of the recaptures is taken into account in the variance formula but not in the point estimate itself as that is just a ratio of two sums.
CLOSED SYSTEM SIMULATIONS

Simulation Approach for both Open and Closed Systems

All simulations in this paper were done using Monte Carlo-based approach. Multiple population and sample sizes were tested. Each population/sample size combination was tested 250 times. The mean of these 250 trials is used as the estimator’s mean value. We emphasize that some recaptures are needed to form an estimator and that requires a certain number of trials.

In the charts we use the convention that “N” represents true population size and “n” denotes number of trials.

![Graph: Estimation for a closed system (Population Size 1,000)](image)

Figure 1: Estimation for a closed system (Population Size 1,000)
Figure 2: Estimation for a closed system (Population Size 10,000)

Figure 3: Estimation for a closed system (Population Size 100,000)
Figure 4: Estimation for a closed system (Population Size 1,000,000)
OPEN SYSTEM UNDER CLUSTER SAMPLING

Random population problem: the definition of $\tau$

Define a sequence of $\tau_i = \frac{Y_i}{q_i}$ as random to denote the exact population at each discrete time period.

$Y_i =$ Total marked units present at the start of $ith$ trial. Thus $Y_i$ is no longer monotonic and may decrease or increase as units enter/leave.

$q_i =$ Probability of recapture and is a population parameter. It is also no longer monotonic and may decrease or increase as units enter/leave.

$s =$ Observed sample

$x_i = Ber(q_i); Ex_i = q_i; E x = q$

Define the random quantity $\tau$ as,

$$\tau = \sum \frac{Y_i}{q_i}$$

We can write this as

$$\tau = \sum \frac{Y_i}{q_i} = \sum \frac{1}{q_i} Y_i = \sum q_i \frac{1}{q_i} Y_i = \sum w_i Y_i = \sum w_i \tau_i$$

$$w_i = \frac{q_i}{\sum q_i}$$
With this formulation, $\tau$ is a weighted average of exact population values at discrete times. Note that if we make the additional assumption,

$$E\tau_i = \tau_M$$

Then,

$$E\tau = \sum w_i E\tau_i = \tau_M \sum w_i = \tau_M$$

We are interested in estimating random $\tau$

**Estimator**

Following our usual approach, define the estimator,

$$\hat{\tau} = \frac{\sum Y_i}{\sum x_i}$$

Note the following about the estimator:

1. Since due to non iid SLLN, $\bar{x} \rightarrow q$ we have $\hat{\tau} \big| s \rightarrow \tau(a.s)$. Thus due to conditional Dominating convergence Theorem (DCT) $E\left(\hat{\tau} \big| s\right) \rightarrow E\tau(a.s)$.

2. We can also observe the above result by noting $E\tau \big| s \rightarrow \frac{EY}{Ex}$, by Taylor series expansion. Further $\frac{EY}{Ex} = \frac{EY}{q} = \frac{EY}{q} = E\tau$. Thus $\hat{\tau} \rightarrow E\tau$ and estimator is approximately unbiased.
Cluster sampling

Throughout our development, we have assumed draw by draw with replacement. Cluster sampling situation arises when more than one unit is caught at a time, such as when the experimenter is using a net or a hook to catch fish. In a cluster sampling situation, a unit cannot be captured twice (recaptured) in a single trial since many units are caught at once and this eliminates the opportunity for a unit to be released and possibly recaptured.

The open system model lends itself to cluster sampling situation. We simply model each individual recapture as an immediate exit from the system. In other words, these individuals are no longer available for further recaptures. We continue to note these individual “exits” as if they were individual experiments. Finally, we allow all of them to return at once (entry) into the system before the start of the next cluster trial. A typical data setup is shown below as an example.

Example:

Suppose the end of $i$th trial is a cluster trial with 3 recaptures. Translated in terms of our “individualized model” the data would appear as:

<table>
<thead>
<tr>
<th>Individual trials within cluster trial, $i$</th>
<th>Singleton recaptures $x_i$</th>
<th>Cluster recaptured exits, $s_i$</th>
<th>Cluster recaptured re-entries, $k_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
The first two rows reflect the adjusted data to account for **first** two recaptures due to cluster sampling. The row reflects the **last** recaptured unit which is treated as a singleton trial with no adjustment. The ordering is important as after the last recapture, the experiment is repeated. Finally, the exited units re-enter the system at the end of the trial \( k_i = 2 \).

Note that clustering adjustments to the data are needed only when more than one unit is recaptured. Cases where several unmarked units are captured/and or simply one unit is recaptured do not require any data adjustment.

**Open system model under cluster sampling**

We change notation to avoid confusion with closed system case. The number of marked units at the end of time \( i-1 \) (beginning of time \( i \)) is \( Y_{i-1} \). At the end of time \( i \) the marked population is updated due to \( x_i = 0 \) recaptures, \( s_i \) marked units leaving the system (including deaths), \( k_i \) marked units re-enter the system due to cluster trials. The marked population at the end of time \( i \) is \( Y_i \). Formally,

\[
Y_i = Y_{i-1} + 1 - x_i - s_i + k_i; i = 1 \cdots n
Y_i = Y_{i-1} + 1 - (x_i + s_i - k_i)
\]

This is different than our standard model in two ways: \((x_i + s_i - k_i)\) replaces \( x_i \) and \( \tau_i \) is not fixed at each discrete trial \( i \).

**Net marked exits** \( s_i := \) Random marked exits – Random marked re-entries + Cluster marked exits
Cluster marked re-entries $k_i$ : = number of marked units re-entering the system as a result of cluster trials. The quantity, $k_i \geq 0$, is completely known once $s_i$ is known. Thus, $P(K_i = k_i|s_i) = 1$. We will thus treat $k_i$ as a constant whose values are known.

Also note the following:

1. We draw the distinction between “random” re-entries of marked units and “cluster re-entries” that were explained above. The cluster re-entries are non-random and arise due to model specification.

2. The exits $s_i$ are modeled as binomial trials with a given average probability of exit over all available marked units. We thus require that each trial is random.

3. The random variable $s_i$ is non-negative. This does not mean that we assume that there are more marked exits than random re-entries on each trial. If there are more marked exits than random re-entries, the case is explained below in detail.

4. Unmarked units can leave or enter the system including births as these are unmarked.

Excess random marked re-entries

Random re-entries are netted against marked exits and are therefore “handled” by the quantity $s_i \geq 0$. The limitation on non negative $s_i$ leads us to consider the case where “excess” random re-entries have to be considered. We do not specifically model these excess random re-entries. Instead we treat them as unmarked units when they re-enter the system. This has no affect on our estimation but the price we pay is that this marked data is lost and we have to start fresh. In other words, when these previously marked units are caught again, we remove the old marks and remark them.
Estimator Attains Asymptotic Minimum Variance

Assume that $\tau$ is random and we showed that:

$$\hat{E}\tau|s \to E\tau.$$  

With this formulation and denoting mean squared error with $MSE$, 

$$\text{MSE} \hat{\tau}|s = E(\hat{\tau} - \tau)^2|s \to \text{Var}(\hat{\tau} - \tau)|s \text{ (Estimator is asymptotically unbiased)}$$

$$E(\hat{\tau} - \tau)^2 \to E[\text{Var}(\hat{\tau} - \tau)|s] = \text{Var}(\hat{\tau} - \tau) - \text{Var}[E(\hat{\tau} - \tau)|s] = \text{Var}(\hat{\tau} - \tau)$$

$$E(\tau - \tau)^2 \to \text{Var} \hat{\tau} + \text{Var} \tau - 2\text{Cov}(\tau, \tau)$$

Now, since $\tau \to \tau$ we must have $E(\hat{\tau} - \tau)^2 \to 0$. Thus,

$$\text{Var} \hat{\tau} + \text{Var} \tau - 2\text{Cov}(\tau, \tau) \to 0$$

$$\hat{\text{Var}} \tau \to 2\text{Cov}(\tau, \tau) - \text{Var} \tau$$

Note that,

$$\text{Cov}(\tau, \tau) = \text{Cov} \left( \frac{\bar{Y}}{x}, \frac{\bar{Y}}{q} \right) = E \frac{\bar{Y}^2}{xq} - \left( E \frac{\bar{Y}}{x} \right) \left( E \frac{\bar{Y}}{q} \right) = E \frac{\bar{Y}^2}{xq} - \left( E \frac{\hat{\tau}}{xq} \right)^2$$

$$\text{Cov}(\tau, \tau) \to E \frac{\bar{Y}^2}{xq} - (E \tau)^2$$

Using Strong Law of Large Numbers (SLLN),
\[ \text{Cov}(\tau, \tau) \rightarrow E \frac{Y}{q} - (E\tau)^2 = (E\tau^2) - (E\tau)^2 = \text{var} \tau \]

\[ \hat{\text{var}} \text{ (a.s)} \]

\[ \text{Var} \tau \rightarrow 2\text{Var} \tau - \text{Var} \tau = \text{Var} \tau \]

Since the estimator variance cannot be less than the variance of the underlying population variance, we claim that our estimator attains (asymptotically) minimum variance.

**Asymptotic confidence interval in open system**

\[
\hat{\tau} - E\tau \sim N(0,1) \quad \sqrt{\text{Var} \tau}
\]

\[ \text{CI} (E\tau) \rightarrow \hat{\tau} \pm z_{\alpha/2} \sqrt{\text{Var} \tau} \]

Thus in the open system we provide confidence interval for \( E\tau \).

**Estimator variance: Derivation of Formula**

As suggested earlier the estimator variance is calculated conditioned on the sample \( s \). To keep notation simple, we drop the notation for sample \( s \).

\[ \hat{\text{Var}} \tau = \text{Var} \hat{Y} \]

Using First Order Taylor approximation and SLLN, we can use the same result in the closed system case,
\[ \text{Var } \tau \to \frac{\text{Var } \bar{X} + \tau^2 \text{Var } \bar{x} - 2\tau \text{Cov}(\bar{X}, \bar{x})}{\left( \frac{q}{g} \right)^2} \]

The remaining sections will focus on calculating the right hand side.

**Binomial distribution of** \( S_i \)

*We remind the reader that from here onward the pair \( \{s_i, Y_i\} \) can be a constant or a random variable. This is to be gleaned from context and where necessary, we will use \( S_i \) to emphasize that \( s_i \) is a constant.*

As stated earlier, we assume that each trial is random. This permits us to model \( s_i \) as a binomial random variable with the following observations:

1. The total population of units available (marked and unmarked) for exit (at end of trial) is \( Y_i \).
2. For each trial, we care about average net exit probabilities over all available units and thus we do not require specification of separate exit probabilities for each marked unit.
3. While the true mean (net exit) probability is unknown, it is estimated as a sample average \( \frac{s_i}{Y_i} \). It represents an average proportion over all marked units. Both quantities in the ratio are constants and observed for the trial. While these would change in a different experiment, we can always replace them by the new observed values. Thus the pair \( \{s_i, Y_i\} \) is fixed and known.*
With this development we postulate,

\[ S_i \sim Binomial \left( Y_i, \frac{s_i}{Y_i} \right) \]

\[ ES_i = s_i \]

\[ VarS_i = s_i \left( 1 - \frac{s_i}{Y_i} \right) \]

We can determine probabilities of other possible outcomes \( s_k = 1, 2 \ldots Y_i \):

\[ P \left( S_i = s_k \right) = B \left( s_k, Y_i, \frac{s_i}{Y_i} \right) \quad \text{Binomial probability of } S_i \text{ evaluated at } s_k. \]

**Estimation of covariance of marked exits**

First note a useful result. Without loss of generality assume \( i < j \)

\[ Y_i = Y_{i-1} + 1 - x_i - s_i + k_i; i = 1 \ldots n \]

\[ Cov\left( (Y_i - Y_{i-1}), (Y_j - Y_{j-1}) \right) = \text{cov}(1 - x_i - s_i, 1 - x_j - s_j) \]

\[ = \text{cov}(x_i, x_j) + 2 \text{cov}(x_i, s_j) + \text{cov}(s_i, s_j) \]

Proceeding with estimation,

\[ Cov(s_i, s_j) = Es_i s_j - Es_i Es_j = Es_i s_j - s_i s_j \]

\[ Es_i s_j = \sum_{s_k = 0}^{Y_i} \sum_{s_l = 0}^{Y_j} s_k s_l P \left( S_i = s_k, S_j = s_l \right) \]
\[ P(S_i = s_k, S_j = s_l) = P(S_j = s_l | S_i = s_k)P(S_i = s_k) \]
\[ = \left[ P(S_j = s_l, X_i = 0 | S_i = s_k) + P(S_j = s_l, X_i = 1 | S_i = s_k) \right] P(S_i = s_k) \]
\[ = \left[ P(S_j = s_l | S_i = s_k, X_i = 0)P(X_i = 0) + P(S_j = s_l | S_i = s_k, X_i = 1)P(X_i = 1) \right] P(S_i = s_k) \]
\[ = \left[ (1 - q_i)P(S_j = s_l | S_i = s_k, X_i = 0, k_i) + q_iP(S_j = s_l | S_i = s_k, X_i = 1, k_i) \right] P(S_i = s_k) \]

Consider restarting the experiment at \( i + 1 \) with a starting marked unit population of \( Y_i - s_k + 1 + k_i \) and additional new marked units \( Y_{j-i} \). Therefore,

\[ P(S_j = s_l | S_i = s_k, X_i = 0) = \text{Binomial} \left( s_i, Y_i - s_k + 1 + k_i + Y_{j-i}, \frac{s_j}{Y_i - s_k + 1 + k_i + Y_{j-i}} \right) \]

\[ P(S_j = s_l | S_i = s_k, X_i = 1) = \text{Binomial} \left( s_i, Y_i - s_k + k_i + Y_{j-i}, \frac{s_j}{Y_i - s_k + k_i + Y_{j-i}} \right) \]

\[ P(S_i = s_k) = \text{Binomial} \left( s_k, Y_i, \frac{s_i}{Y_i} \right) \]

\[
E_{s,s_j} = \sum_{s_i=0}^{Y_i} \sum_{s_k=0}^{Y_i} s_k s_j B \left( s_i, Y_i, \frac{s_i}{Y_i} \right) \left[ (1 - q_i)B \left( s_i, u_{ij}(s_k) + 1, \frac{s_j}{u_{ij}(s_k) + 1} \right) + q_iB \left( s_i, u_{ij}(s_k), \frac{s_j}{u_{ij}(s_k)} \right) \right]
\]

\[ u_{ij}(s_k) = Y_i - s_k + 1 + k_i + Y_{j-i} \]

\[
Q_{ij} = \text{Cov}(s_i, s_j) = \sum_{s_i=0}^{Y_i} \sum_{s_k=0}^{Y_i} s_k s_j B \left( s_k, Y_i, \frac{s_i}{Y_i} \right) \left[ (1 - q_i)B \left( s_i, u_{ij}(s_k) + 1, \frac{s_j}{u_{ij}(s_k) + 1} \right) + q_iB \left( s_i, u_{ij}(s_k), \frac{s_j}{u_{ij}(s_k)} \right) \right] - s_i s_j
\]

\[ u_{ij}(s_k) = Y_i - s_k + 1 + k_i + Y_{j-i} \]

Note that the RHS is a function of \( \{s_i, Y_i\} \) and these values are observed and known for the experiment. Next, a computationally efficient approximation,
\[ Q_j \to \sum_{s_i=0}^{y_j} \sum_{s_k=0}^{y_j} s_k s_j \left[ (1-q_j)A + q_j B \right] - s_i s_j \]

\[ C = \Phi \left( \frac{Y_i(s_k/s_i - 1)}{Y_i - s_i} \right) - \Phi \left( \frac{Y_i(1/s_i - 1)}{Y_i - s_i} \right) \]

\[ A = \Phi \left( \frac{\{u_y(s_k)+1\} s_i/s_j - 1\} \right) - \Phi \left( \frac{(s_i - 1)/s_j - 1\} \right) \]

\[ B = \Phi \left( \frac{\{u_y(s_k)-1\} s_i/s_j - 1\} \right) - \Phi \left( \frac{(s_i - 1)/s_j - 1\} \right) \]

Asymptotically, this approximation is just as accurate as the binomial formula due to the fact that marked units, \( Y_i \to \infty \)

**Estimation of covariance of marked exits and captures**

Without loss of generality assume \( i < j \)

\[ \text{cov}(x_i, s_j) = Ex_i s_j - Ex_i Es_j \]

\[ = \sum_{x_i=0}^{x_i=y_i} \sum_{s_k=0}^{s_k=y_i} s_k x_i P(X_i = x_i, S_j = s_k) - Ex_i Es_j \]

\[ = \sum_{x_i=0}^{x_i=y_i} s_k P(X_i = 1, S_j = s_k) - Ex_i Es_j \]

\[ = \sum_{x_i=0}^{x_i=y_i} s_k P(S_j = s_k | X_i = 1) P(X_i = 1) - Ex_i Es_j \]

Note that in the \( ith \) trial \( (i < j) \), we have no new marked units. Hence this trial did not increase the marked population \( s_i \). Effectively, we can delete the \( ith \) trial and when we count the trials the correct count would be:

\[ 1,2...i-1, mis \sin g...j-1, j \]

\[ \equiv 1,2,...,j-1 \]
Thus the two experiments are identical. Therefore,

\[ P\left( S_j = s_k | x_i = 1 \right) = P\left( S_{j-1} = s_k \right) \]

\[ P\left( S_{j-1} = s_k \right) = \text{Binomial} \left( s_k, Y_{j-1}, \frac{s_{j-1}}{Y_{j-1}} \right) \]

Proceeding,

\[ \sum_{s_k=0}^{s_{Y_{j-1}}} s_k P\left( S_j = s_k | x_i = 1 \right) P\left( X_i = 1 \right) - \text{Ex} E s_j = \sum_{s_k=0}^{s_{Y_{j-1}}} s_k P\left( S_{j-1} = s_k \right) q_i - q_i s_j \]

\[ \text{cov}(x_i, s_j) = C_{ij} = \sum_{s_k=0}^{s_{Y_{j-1}}} q_i s_k \left( \frac{Y_{j-1}}{s_k} \right) \left( \frac{s_{j-1}}{Y_{j-1}} \right)^{s_k} \left( 1 - s_{j-1} \frac{1}{Y_{j-1}} \right)^{Y_{j-1} - s_k} - q_i s_j \]

Note that above formula \( \text{cov}(x_i, s_j) \) remains the same even if \( i = j \) and we just replace the subscripts. A computationally efficient approximation,

\[ C_{ij} \rightarrow \sum_{s_k=0}^{s_{Y_{j-1}}} q_i s_k \left[ \Phi \left( \frac{Y_{j-1} \left( s_k / s_{j-1} - 1 \right)}{(Y_{j-1} - s_{j-1})} \right) - \Phi \left( \frac{Y_{j-1} \left( (s_k - 1) / s_{j-1} - 1 \right)}{(Y_{j-1} - s_{j-1})} \right) \right] - q_i s_j \]

Asymptotically, this approximation is just as accurate as the binomial formula due to the fact that marked units, \( Y_i \rightarrow \infty \)

**Estimation of covariance of marked captures**

Without loss of generality assume that \( i < j \). Then
\[ \text{cov}(x_i, x_j) = E(x_i x_j) - E(x_i)E(x_j) \]
\[ P(x_i = 1, x_j = 1) - q_i q_j \]
\[ P(x_j = 1| x_i = 1) P(x_i = 1) - q_i q_j \]
\[ P(x_j = 1| x_i = 1) q_i - q_i q_j \]
\[ q_i \left( P(x_j = 1| s_i = 0, x_i = 1) P(s_i = 0) + P(x_j = 1| s_i > 0, x_i = 1) P(s_i > 0) \right) - q_i q_j \]
\[ q_i \left( P(x_j = 1| s_i = 0, x_i = 1, k_i) P(s_i = 0) + P(x_j = 1| s_i > 0, x_i = 1, k_i) \left[ 1 - P(s_i = 0) \right] \right) - q_i q_j \]

We now focus on evaluating \( P(x_j = 1| s_i > 0, x_i = 1, k_i) , \)

\[ \ln \left( \frac{\tau_i}{\tau_{i-1}} \right) = \delta_i \]
\[ \tau_i = \tau_{i-1} \exp(\delta_i) \]

Now assume that \( \{s_i, k_i\} > 0 \) such that \( \sup_i \left| \frac{k_i - s_i}{\tau_{i-1}} \right| \to 0 . \) Then,

\[ Y_i = Y_{i-1} + 1 - x_i - s_i + k_i; i = 1 \cdots n; x_i = 1 \]
\[ \sup_i \left| \frac{Y_i}{\tau_{i-1}} - \frac{Y_{i-1}}{\tau_{i-1}} \right| \to 0 \]

Thus,

\[ Y_i \to Y_{i-1}; i \geq 1 \]
\[ \frac{Y_i}{\tau_{i-1}} \to \frac{Y_{i-1}}{\tau_{i-1}}; i \geq 1 \]

At the \textit{ith} trial, under the condition that \( x_i = 1 \), the sequence of \( q_k 's \) “updates” as follows:
\[
\frac{Y_{i-1}}{\tau_{i-1}}, \frac{Y_{j}}{\tau_{j}} \Rightarrow \frac{Y_{i-1}}{\tau_{i-1}} \exp(-\delta_i) \Rightarrow \frac{Y_{j}}{\tau_{j}} \Rightarrow \frac{Y_{i-1}}{\tau_{i-1}}, \frac{Y_{j}}{\tau_{j}}, \frac{Y_{i-1}}{\tau_{i-1}}, \frac{Y_{j}}{\tau_{j}}.
\]

Using Taylor expansion,

\[
\exp(-\delta_i) = 1 - \delta_i + \frac{\delta_i^2}{2} - \frac{\delta_i^3}{6} + ..
\]

\[
\Rightarrow \frac{Y_{i-1}}{\tau_{i-1}}, \frac{Y_{j}}{\tau_{j}} - \delta_i \frac{Y_{i-1}}{\tau_{i-1}} + \frac{\delta_i^2}{2} \frac{Y_{i-1}}{\tau_{i-1}} - \frac{\delta_i^3}{6} \frac{Y_{i-1}}{\tau_{i-1}} + .... \frac{Y_{j}}{\tau_{j}} \Rightarrow \frac{Y_{i-1}}{\tau_{i-1}}, \frac{Y_{j}}{\tau_{j}}.
\]

Thus no change is seen in the \(i\)th trial as long as we assume \(|\delta| < 1\) such that \(\sup_i |\delta_i| Y_{i-1}/\tau_{j-1}| \to 0\).

With this assumption,

\[
P(x_j = 1|s_j > 0, x_j = 1) \to q_{j-1}
\]

\[
P(x_j = 1|s_j = 0, x_j = 1) \to q_{j-1}
\]

Thus,

\[
\text{cov}(x_i, x_j) \to q_i q_{j-1} \{P(s_i = 0)+1-P(s_i = 0)\}-q_i q_j
\]

\[
\text{cov}(x_i, x_j) \to q_i (q_{j-1} - q_j)
\]

Finally \(i < j\),

\[
\text{Cov}\{\left(Y_i - Y_{i-1}\right), \left(Y_j - Y_{j-1}\right)\} \to q_i (q_{j-1} - q_j) + 2C_{ij} + Q_{ij}
\]

Going forward we will also need
\[
\begin{align*}
\text{Var}(Y_i - Y_{i-1}) &= \text{Var}(1 - x_i - s_i) = \text{Var}(x_i + s_i) \\
\text{Var}(x_i + s_i) &= \text{Var}x_i + \text{Vars}_i + 2\text{Cov}(x_i, s_i) \\
\text{Var}(Y_i - Y_{i-1}) &= q_i(1 - q_i) + Y_i \frac{s_i}{Y_i} \left(1 - \frac{s_i}{Y_i}\right) + 2C_{ii} \\
\text{Var}(x_i + s_i) &= q_i(1 - q_i) + s_i \left(1 - \frac{s_i}{Y_i}\right) + 2C_{ii}
\end{align*}
\]

Estimation of variance of average marked units in system

Since,

\[
Y_i - Y_o = \sum_{j=1}^{i}(Y_j - Y_{j-1})
\]

\[
\sum_{i=1}^{n}(Y_i - Y_o) = \sum_{i=1}^{n} \sum_{j=1}^{i}(Y_j - Y_{j-1})
\]

\[
\text{Var}\sum_{i=1}^{n}(Y_i - Y_o) = \text{Var}\sum_{i=1}^{n} \sum_{j=1}^{i}(Y_j - Y_{j-1})
\]

\[
\text{Var}\sum_{i=1}^{n} Y_i = \text{Var}\sum_{i=1}^{n} \sum_{j=1}^{i}(Y_j - Y_{j-1})
\]

Since we have finite sums, we can switch order of summation. Note that \((Y_j - Y_{j-1})\) can be both negative and positive, so \(n\) is large but still finite,

\[
\text{Var}\sum_{i=1}^{n} Y_i = \text{Var}\sum_{j=1}^{n} \sum_{i=j}^{n}(Y_j - Y_{j-1}) = \text{Var}\sum_{j=1}^{n}(Y_j - Y_{j-1})(n - j + 1)
\]
\[ \text{Var} \sum_{i=1}^{n} Y_i \rightarrow \]
\[ \left( \sum_{i=1}^{n} (n-i+1)^2 \left\{ q_i(1-q_i) + s_i \left( 1 - \frac{s_i}{Y_i} \right) + 2C_{ii} \right\} \right) \]
\[ \text{Var} Y = \text{Var} \left( \frac{\sum_{j=1}^{n} Y_j}{n} \right) = \frac{\text{Var} \sum_{j=1}^{n} Y_j}{n^2} \rightarrow \]
\[ \left( \sum_{i=1}^{n} (n-i+1)^2 \left\{ q_i(1-q_i) + s_i \left( 1 - \frac{s_i}{Y_i} \right) + 2C_{ii} \right\} \right) \]
\[ = \frac{\sum_{j=1}^{n} (n-j+1)(n-i+1) \left\{ q_i(q_{j-1} - q_j) + 2C_{ij} + Q_{ij} \right\}}{n^2} \]

**Estimation of variance of average captures**

\[ \text{Var} x = \text{Var} \left( \frac{\sum_{i=1}^{n} x_i}{n} \right) = \frac{1}{n^2} \text{Var} \sum_{i=1}^{n} x_i = \frac{1}{n^2} \left\{ \sum_{i=1}^{n} \text{var} x_i + 2 \sum_{j=1}^{n} \sum_{i=1}^{j-1} \text{cov}(x_i, x_j) \right\} \]
\[ \text{Var} x \rightarrow \frac{1}{n^2} \left\{ \sum_{i=1}^{n} q_i(1-q_i) + 2 \sum_{j=1}^{n} \sum_{i=1}^{j-1} q_i(q_{j-1} - q_j) \right\} \]
\[ \tau^2 \text{Var} x \rightarrow \frac{1}{n^2} \left\{ \sum_{i=1}^{n} \tau^2 q_i(1-q_i) + 2 \sum_{j=1}^{n} \sum_{i=1}^{j-1} \tau^2 q_i(q_{j-1} - q_j) \right\} \]
Estimation of covariance of average marked units in system and captures

\[ Y_i = Y_{i-1} + 1 - x_i - s_i + k_i \]
\[ Y_i - Y_{i-1} = 1 - x_i - s_i + k_i \]
\[ \frac{Y_n - Y_0}{n} = 1 - \bar{x} - \bar{s} + \bar{k} \]
\[ Y_n = n(1 - x - s + k) + Y_0 \]
\[ \bar{x} + \bar{s} - \bar{k} = 1 - \frac{Y_n - Y_0}{n} \]

\[ Y_k = k \left( \frac{1}{\bar{k}} \sum_{i=1}^{k} (x_i + s_i - k_i) \right) + Y_0; k \leq n \]

\[ \text{Cov}\left( x, Y \right) = \text{Cov}\left( 1 + \frac{(Y_n - Y_0)}{n} - s + k, Y \right) = -\text{Cov}\left( \frac{Y_n}{n} + s, \sum_{k=1}^{n} Y_k \right) = -\frac{1}{n^2} \text{Cov}\left( Y_n + \sum_{k=1}^{n} s_k, \sum_{k=1}^{n} Y_k \right) \]

\[ = -\frac{1}{n^2} \text{Cov}\left( n(1 - \bar{x} - \bar{s} + \bar{k}) + Y_0 + n s, \sum_{k=1}^{n} k \left( \frac{1}{\bar{k}} \sum_{i=1}^{k} (x_i + s_i - k_i) \right) \right) \]

\[ = -\frac{1}{n^2} \text{Cov}\left( n \sum_{k=1}^{n} (x_1 + s_1) + (x_2 + s_2) + \ldots (x_k + s_k) \right) \]
\[ = -\frac{1}{n^2} \text{Cov}\left( x_1 + x_2 + \ldots x_n, \sum_{k=1}^{n} (x_1 + s_1) + (x_2 + s_2) + \ldots (x_k + s_k) \right) \]
\[ = -\frac{1}{n^2} \text{Cov}\left( x_1 + x_2 + \ldots x_n \left[ n(x_1 + s_1) + (n-1)(x_2 + s_2) + \ldots + (x_n + s_n) \right] \right) \]
\[- \frac{1}{n^2} \text{Cov} \left\{ \sum_{i=1}^{n} x_i, \sum_{i=1}^{n} (n-i+1)(x_i + s_j) \right\} \]
\[- \frac{1}{n^2} \sum_{j=1}^{i} \sum_{i=1}^{n} \text{cov} \left\{ x_i, (n-j+1)(x_j + s_j) \right\} \]
\[- \frac{1}{n^2} \left\{ \sum_{i=1}^{n} \text{Cov} \left[ x_i, (n-i+1)(x_i + s_j) \right] + 2 \sum_{j=1}^{n} \sum_{i=1}^{i-1} \text{Cov} \left[ x_i, (n-j+1)(x_j + s_j) \right] \right\} \]
\[- \frac{1}{n^2} \left\{ \sum_{i=1}^{n} (n-i+1) \text{Cov} \left[ x_i, (x_i + s_j) \right] + 2 \sum_{j=1}^{n} \sum_{i=1}^{i-1} (n-j+1) \text{Cov} \left[ x_j, (x_j + s_j) \right] \right\} \]
\[ \text{Cov} \left( \bar{x}, \bar{Y} \right) = - \frac{1}{n^2} \left\{ \sum_{i=1}^{n} (n-i+1) \left[ \text{Var} (x_i) + \text{cov} (x_i, s_j) \right] + 2 \sum_{j=1}^{n} \sum_{i=1}^{i-1} (n-j+1) \left[ \text{cov} (x_i, x_j) + \text{cov} (x_i, s_j) \right] \right\} \]

Note the earlier result,

\[ \text{cov} (x_i, x_j) + \text{cov} (x_i, s_j) \rightarrow \left[ q_i (q_{j-1} - q_j) + C_{ij} \right] \]
\[ \text{Var} (x_i) = q_i (1-q_i) \]
\[ \text{cov} (x_i, s_j) = C_{ii} \]
\[ \text{Cov} \left( \bar{x}, \bar{Y} \right) \rightarrow - \frac{1}{n^2} \left\{ \sum_{i=1}^{n} (n-i+1) \left[ q_i (1-q_i) + C_{ii} \right] + 2 \sum_{j=1}^{n} \sum_{i=1}^{i-1} (n-j+1) \left[ q_i (q_{j-1} - q_j) + C_{ij} \right] \right\} \]
\[ -2 \tau \text{Cov} (\bar{Y}, \bar{x}) \rightarrow \]
\[ \frac{1}{n^2} \left\{ \sum_{i=1}^{n} 2 \tau (n-i+1) \left[ q_i (1-q_i) + C_{ii} \right] + 2 \sum_{j=1}^{n} \sum_{i=1}^{i-1} 2 \tau (n-j+1) \left[ q_i (q_{j-1} - q_j) + C_{ij} \right] \right\} \]

Further,

\[ \hat{\tau}^{(x,y)} \rightarrow \frac{\text{Var} \bar{Y} + \tau^2 \text{Var} \bar{x} - 2 \tau \text{Cov} (\bar{Y}, \bar{x})}{\left( \hat{q} \right)^2} \]
\[-2\tau \text{Cov}(\bar{Y}, x) \rightarrow \]

\[ \frac{1}{n^2} \left\{ \sum_{i=1}^{n} 2\tau(n-i+1) \left[q_i(1-q_i) + C_{ii} \right] + 2 \sum_{j=1}^{n-1} \sum_{i=1}^{j} 2\tau(n-j+1) \left[q_i(q_{j-1} - q_j) + C_{ij} \right] \right\} \]

\[ \tau^2 \text{Var} \bar{x} \rightarrow \frac{1}{n^2} \left\{ \sum_{i=1}^{n} \tau^2 q_i(1-q_i) + 2 \sum_{j=1}^{n-1} \sum_{i=1}^{j} \tau^2 q_i(q_{j-1} - q_j) \right\} \]

\[ \text{Var} \bar{Y} \rightarrow \]

\[ \left[ \frac{\sum_{i=1}^{n} (n-i+1)^2 \left[q_i(1-q_i) + s_i \left(1 - \frac{s_i}{Y_i} \right) + 2C_{ii} \right] + 2 \sum_{j=1}^{n-1} \sum_{i=1}^{j} (n-j+1)(n-i+1) \left[q_i(q_{j-1} - q_j) + 2C_{ij} + Q_{ij} \right]}{n^2} \right] \]

\[ \text{Var} \tau \rightarrow \]

\[ \left[ \frac{\sum_{i=1}^{n} \tau^2 q_i(1-q_i) + (n-i+1)^2 \left[q_i(1-q_i) + s_i \left(1 - \frac{s_i}{Y_i} \right) + 2C_{ii} \right] + 2\tau(n-i+1) \left[q_i(1-q_i) + C_{ii} \right]}{n^2 \left(\tau^2 \right)} \right] \]

\[ + \frac{2 \sum_{j=1}^{n-1} \sum_{i=1}^{j} \tau^2 q_i(q_{j-1} - q_j) + (n-j+1)(n-i+1) \left[q_i(q_{j-1} - q_j) + 2C_{ij} + Q_{ij} \right] + 2\tau(n-j+1) \left[q_i(q_{j-1} - q_j) + C_{ij} \right]}{n^2 \left(\tau^2 \right)} \]

\[ \text{Var} \tau \rightarrow \text{Var} \tau \rightarrow \]
\[
\sum_{i=1}^{n} \tau^2 q_i (1-q_i) + (n-i+1)^2 \left\{ q_i (1-q_i) + s_i \left( 1 - \frac{s_i}{Y_i} \right) + 2C_i \right\} + 2\tau(n-i+1)\left[ q_i (1-q_i) + C_i \right] \\
\left( \sum_{i=1}^{n} q_i \right)^2
\]

\[
2 \sum_{j=1}^{n} \sum_{i=1}^{j-1} \tau^2 q_i (q_{i,j} - q_j) + (n-j+1)(n-i+1)\left\{ q_i (q_{i,j} - q_j) + 2C_{i,j} + Q_{i,j} \right\} + 2\tau(n-j+1)\left[ q_i (q_{i,j} - q_j) + C_{i,j} \right] \\
\left( \sum_{i=1}^{n} q_i \right)^2
\]

Population changes approach zero

In the last proof of \( \hat{\text{Var}} \) we have made these assumptions: \( \sup_i \left| \delta_i \frac{Y_{i-1}}{\tau_{i-1}} \right| < \varepsilon_1 \) for any given \( \varepsilon_1 > 0 \)

& \( \sup_i \left| k_i - s_i \right| \frac{1}{\tau_{i-1}} < \varepsilon_2 \) for any given \( \varepsilon_1, \varepsilon_2 > 0 \). These conditions were needed under the constraint

\( Y_i - Y_{i-1} = k_i - s_i \). We show in this section that there is a single equivalent assumption covering both these cases. Start by making only the assumption \( \sup_i \left| k_i - s_i \right| \frac{1}{\tau_{i-1}} < \varepsilon_2 \).

\[
\frac{k_i - s_i}{\tau_{i-1}} = \frac{k_i - s_i}{\tau_{i-1}} = \frac{Y_i - Y_{i-1}}{\tau_{i-1}} = \frac{Y_i}{\tau_{i-1}} - \frac{Y_{i-1}}{\tau_{i-1}} < \varepsilon_2
\]

\[
\frac{\tau_i Y_i}{\tau_{i-1}} - \frac{Y_{i-1}}{\tau_{i-1}} < \varepsilon_2
\]

\[
|q_i \exp(\delta_i) - q_{i-1}| < \varepsilon_2
\]

The above are iff statements and thus the condition is equivalent to

\( \Leftrightarrow |q_i \exp(\delta_i) - q_{i-1}| < \varepsilon_2 \)
Now make the second assumption \( \sup_i \left| \delta_i \frac{Y_i}{\tau_i} \right| < \varepsilon_1 \)

\[
\iff \left| \delta_i \frac{\tau_i}{\tau_{i-1}} \frac{Y_i}{\tau_i} \right| = |\delta_i||q_i\exp(\delta_i)| < \varepsilon_1
\]

Let \( \sup_i |\delta_i| = \delta < 1 \). Now choose \( 2\varepsilon = \min(\varepsilon_1, \delta \varepsilon_2) \) such that,

\[
|\delta_i q_i \exp(\delta_i)| = \varepsilon \left( < 2\varepsilon \right) \tag{6}
\]

\[
|\delta_i q_i \exp(\delta_i) - \delta_i q_{i-1}| < 2\varepsilon \tag{7}
\]

A stronger second condition is,

\[
|\delta_i q_i \exp(\delta_i)| + |\delta_i q_{i-1}| < 2\varepsilon
\]

If the above is true then second condition is true also. But the above condition is equivalent to

\[
|\delta_i q_{i-1}| < \varepsilon. \text{ So a single condition } |\delta_i q_{i-1}| < \varepsilon \text{ is needed. Equivalently, } \sup_i |\delta_i q_{i-1}| \to 0. \text{ But}
\]

\[
0 < q_{i-1} \leq 1, \sup_i |\delta_i q_{i-1}| = \sup_i \left( |\delta_i| q_{i-1} \right) \text{. So a stronger condition is, } \sup_i |\delta_i| \to 0. \text{ The almost sure statement has been introduced to avoid the pathological case where a large change in population takes place with measure 0 (in time interval } \to 0 \text{). We accept this as the overall condition and equivalent to stating that population changes approach zero at the start of each trial.}
\]

**Open system as time continuous process**

The above “condition” may seem strange. In fact it’s not a condition and rather results from mischaracterization of a continuous process (open system) into a discrete model. For example, as
noted above, setting $\delta_j = 0$ in the open system formula for $Var \tau$ will not improve the estimation in any way and would lead to a closed system! Note that in continuous time, we would naturally have $\text{Sup}_{i} |\delta_i^{(a,s)}| \to 0$. In conclusion we require $\text{Sup}_{i} |\delta_i|$ being close to 0 but not exactly equal to zero in the discrete open system case.

**Comparison of open versus closed systems**

The formulas for estimators are identical under both open and closed systems but under open system it estimates the weighted average populations. Thus the difference is in interpretation only.

When terms involving $s_i$ are zero the formulas for estimator variances are identical:

$$Var \tau \to$$

$$\sum_{i=1}^{n} p_i (1-p_i) \left[ \tau^2 + 2\tau(n-i+1) + (n-i+1)^2 \right] + 2 \sum_{j=1}^{n-1} p_i (p_{j+1} - p_j) \left[ \tau^2 + 2\tau(n-j+1) + (n-j+1)(n-j+1) \right]$$

$$\left( \sum_{i=1}^{n} p_i \right)^2$$

$$\text{Var} \tau \to \text{Var} \tau \to$$

$$\sum_{i=1}^{n} q_i (1-q_i) \left[ \tau^2 + 2\tau(n-i+1) + (n-i+1)^2 \right] + 2 \sum_{j=1}^{n-1} q_i (q_{j+1} - q_j) \left[ \tau^2 + 2\tau(n-j+1) + (n-j+1)(n-j+1) \right]$$

$$\left( \sum_{i=1}^{n} q_i \right)^2$$

The identical appearance is misleading. The reason is that in a closed system the “saturation parameter” $p_i$ is non-decreasing and this causes $p_i (p_{j+1} - p_j) \leq 0$. Secondly $\tau$ is fixed in closed
system while random in open system. Thirdly \( \left( \sum_{i=1}^{n} p_i \right)^2 \) is non-decreasing in closed systems it is not in open system. In fact for open systems \( \text{Var} \tau \to \text{Var} \tau \) and due to the fact that \( p_i \to 1 \) we have \( \text{Var} \tau \to 0 \) for closed systems.

**Rate of convergence**

The order of convergence depends on the asymptotic assumptions made regarding the estimator and its variance. In both cases, closed form results are hard to obtain analytically and simulation results are used to show the order of convergence. We start with some analytical insights.

There are two asymptotic assumptions made and these two are discussed separately.

**Estimator: \( x \to q \)**

Note that \( x \) is random and,

\[
\text{cov}(x_i, x_j) = q_i (q_{j-1} - q_j)
\]

\[
\text{Var} x_i = q_i (1 - q_i)
\]

\[
\text{Var} x = \frac{1}{n^2} \left[ \sum_{i=1}^{n} q_i (1 - q_i) + 2 \sum_{j=1}^{n} \sum_{i=1}^{j-1} q_i (q_{j-1} - q_i) \right]
\]

The double sum is true for \( i < j \). We can switch the order of summation in which case \( i > j \) as long as we replace with \( q_j (q_{j-1} - q_i) \). Next, note that due to CLT (non-iid case),
\[ \text{DIST} \]
\[ x \rightarrow N\left(q, \text{Var} x\right) \]

The two-sided confidence interval for \( q \)

\[ x \pm z_{a/2} \sqrt{\text{Var} x} \]

Thus for \( x \rightarrow q \) as long as the sample size is large enough so that

\[ \sqrt{\sum_{i=1}^{n} q_i (1-q_i) + 2 \sum_{j=1}^{n} \sum_{i=1}^{j-1} q_i (q_{j-1} - q_j)} \rightarrow 0(a.s) \]

**Estimator variance:** \( (\bar{Y} - E \bar{Y})^k \rightarrow 0; k \geq 2 \)

First note that \( (\bar{Y} - E \bar{Y})^k \rightarrow 0(a.s) \) if \( (\bar{Y} - E \bar{Y}) \rightarrow 0(a.s) \) and the latter follows from SLLN.

Following the argument as above, \( (\bar{Y} - E \bar{Y}) \rightarrow 0(a.s) \) iff

\[ \sqrt{\text{Var} \bar{Y}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (n-i+1)^2 \left[ q_i (1-q_i) + s_i \left( 1 - \frac{s_i}{\bar{Y}_i} \right) + 2C_i \right] + 2 \sum_{j=1}^{n} \sum_{i=1}^{j-1} (n-j+1)(n-i+1) \left[ q_i (q_{j-1} - q_j) + 2C_{ij} + Q_{ij} \right] \rightarrow 0(a.s) } \]

Thus, \( (\bar{Y} - E \bar{Y})^2 \rightarrow 0(a.s) \) iff
\[ Var \bar{Y} = \]
\[
\sum_{i=1}^{n} \frac{(n-i+1)^2 \left( q_i (1-q_i) + s_i \left(1 - \frac{s_i}{Y_i} \right) + 2C_i \right)}{n^2} + 2 \sum_{j=1}^{n} \sum_{i=1}^{j-1} \frac{(n-j+1)(n-i+1) \left( q_i (q_{i+1} - q_j) + 2C_{ij} + Q_{ij} \right)}{n^2} \rightarrow 0 (a.s)
\]
GENERAL DISCUSSION

Randomness: We require that each trial is random. This does not mean that the entire population completely randomizes itself before each trial. Rather, it simply means that units caught are not due to fixed trajectories or predictable outcomes.

Average probability of capture: We make use of the fact that the average probability of catching a unit is \( \frac{Y_i}{\tau_i} \). This does not mean that on each sampling occasion we require the same probability of capture for all units. We do not track individual units and hence all of them are indistinguishable and only the average capture probability is needed.

Population changes approach zero: The analytical assumption \( \sup \delta_i \rightarrow 0 \) requires small changes to the total population. The netting of marked exits with marked re-entries helps keep \( s_i \) small. However, in order for cluster re-entries \( k_i \) to be small, the cluster sizes will need to be small. Overall, this assumption is practically unattainable and the robustness is tested in simulation trials.

Zero Excess marked re-entries in cluster trials: The excess marked random re-entries can lead to a loss of marked data. However, for many cluster trials we hypothesize this to be minimal,

\[ s_i = \text{Marked cluster exits} + \text{Random marked exits} - \text{Marked re-entries} \geq 0 \]

Our hypothesis is based on the fact that generally the density of marked units is inside the system and not outside. Secondly, cluster trials help keep \( s_i \) non-negative.
OPEN SYSTEM SIMULATION UNDER CLUSTER SAMPLING

Many systems are open and truly closed systems are harder to find. The latter being the case with small land locked lakes/ponds, artificial human experiments. This distinction is very important for the researcher. For example open system results require tracking of $s_i$ (marked units leaving system) to estimate variance accurately. This is a limitation of the model, although some existing models have similar tracking requirements. The exclusion of this will result in understatement of true variance and confidence interval. Secondly, the variance result may surprise the researcher since in the open system the variance never reduces to zero but rather approaches population variance!

The interpretation of the estimators in both systems is different as well. In open systems we obtain weighted averages of populations at different discrete times while in closed system we are trying estimate fixed but unknown population.

There are several reasons for doing a computer based simulation: Firstly, it validates the theoretical development. Secondly we provide results where assumptions are less than ideal, such as small $n$. This will provide an idea of the robustness of the assumptions made.

**Robustness**

For any given sampling period, the trajectory could involve a growth or decay scenario. The change in population dependent on the existing population size and thus the population changes were dependent on previous trials. The exists and re-entries of marked units were modeled using binomial distribution. It is possible to perform robustness tests under different conditions until and note conditions when the estimation “breaks” down. This was not done in the results below.
Figure 5: A Random Selection of (n, N) Involving 250 Trajectories
Figure 6: Estimation for a Certain Trajectory (Population Size 1,000)
Figure 7: Estimation for a Certain Trajectory (Population Size 10,000)
Figure 8: Estimation for a Certain Trajectory (Population Size 100,000)
Open System Rate of Convergence

Each bubble in the chart below represents an estimate of the population. Larger encounters improve the population estimate.
Figure 10: Rate of Convergence for a Certain Trajectory (2,000 encounters)
Figure 11: Rate of Convergence for a Certain Trajectory (5,000 encounters)

N = 1e+05, num_clusters = 500
Comparison of variance of the estimator: formula versus simulated

The purpose of the simulation was to check the validity and accuracy of the asymptotic open system variance formula. We display below percentage deviance between the true (simulated) versus formula variance of the estimator.

We used the binomial formulae and ran out of computational power. To perform more simulations, we suggest using the normal approximation and that will yield more rows in the table.
Table 2: Estimated standard deviation

(\% of standard deviation of estimated \tau)

<table>
<thead>
<tr>
<th>Clusters/N</th>
<th>10,000</th>
<th>100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>385.1</td>
<td>8198.3</td>
</tr>
<tr>
<td>20</td>
<td>198.1</td>
<td>784.7</td>
</tr>
<tr>
<td>50</td>
<td>100.1</td>
<td>268.6</td>
</tr>
<tr>
<td>100</td>
<td>107.8</td>
<td>146.7</td>
</tr>
</tbody>
</table>
Explicit & implicit heterogeneity

In most practical sampling situations the researcher gathers data on many different types of units at once. Such a heterogeneous dataset involves different types of units with varying probability of capture/recapture. This of course is not an issue for our model as long as we can identify different types of individuals and separate our datasets accordingly.

The “implicit” effect arises when the population is heterogeneous but the researcher cannot identify the various classes of units that generated the dataset. In this situation our standard model needs to be augmented and we must first develop an approach to identify these classes.

Complete capture history

A row vector comprising of complete capture and recapture history of each unit is recorded. Note that in many instances there will be missing values as the same unit is not captured in that trial. Thus a unit with more frequent recaptures will likely have a higher probability of recapture and thus possibly belong to a different class, at least from a recapture probability viewpoint. Next, we group similar capture histories and index them by $m$. The order of captures is irrelevant and we simply care about total count of captures/recaptures. Our goal will be to use this extended dataset to identify the classes of individuals, thereby converting our implicit heterogeneity problem into explicit. Our standard model then applies individually to each explicit class. However, we will now need to “rebuild” the ordered data set for this class and know the capture-recapture sequence of each class separately.
An example is shown below. Typically, several such strings would be recorded, one for each unique string:

<table>
<thead>
<tr>
<th>Ordered unique string</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000100010001110101010011100111</td>
<td>24</td>
</tr>
</tbody>
</table>

**Multinomial likelihood model**

Let there be $m$ distinct patterns (capture histories) and there are a total of $n$ independent singleton trials. The cell probabilities $p_i, i = 1 \ldots m$ are such that $\sum_{i=1}^{m} p_i = 1$. We note the frequency of patterns, $x_i, i = 1 \ldots m$ and the likelihood function of the observed set $\{x_1, x_2 \ldots x_m\}$ is given by,

$$L(p_1 \ldots p_m) = \frac{n!}{x_1! \ldots x_m!} \prod_{i=1}^{m} p_i^{x_i}$$

Using standard maximum likelihood estimation the parameters $p_1 \ldots p_m$ and their respective confidence intervals are obtained. Classes with overlapping confidence intervals are combined and the process continued until classes with distinct capture histories are identified. This classification is based on distinct probability of observed capture history.
RESULTS

Theorem 1: Finite population closed system

Suppose a design based draw by draw with replacement approach with fixed but unknown \( \tau \).

Then an approximate unbiased point estimate is given by:

\[
\hat{\tau} = \frac{\sum X_i}{\sum X_i}
\]

This estimator also converges almost surely to \( \tau \). Further,

\[
\hat{\text{Var}} \tau \rightarrow \\
\sum_{i=1}^{n} \left[ (n-i+1)^2 + \tau^2 + 2\tau(n-i+1) \right] p_i (1-p_i) + 2 \sum_{j=1}^{n-1} \sum_{i=1}^{j} p_i (p_{j-1} - p_j) \left[ \tau^2 + (n-i+1)(n-j+1) + 2\tau(n-j+1) \right] \\
\left( \sum_{i=1}^{n} p_i \right)^2
\]

\[
\hat{p}_i = \frac{X_i}{\hat{\tau}} . \text{The estimated confidence interval for } \tau
\]

\[
\frac{\hat{X}}{n} \pm z_{\alpha/2} \sqrt{\text{Var} \tau}
\]
Theorem 2: Finite population open system (estimator)

Let the true population be random \( \tau_i \), \( \tau = \sum w_i \tau_i ; w_i = \frac{q_i}{\sum q_i} \). Let,

\[
Y_i = \text{Total marked units present at the end of } ith \text{ trial}
\]

Then an approximate unbiased estimator of \( \tau \) is given by:

\[
\hat{\tau} = \frac{\sum Y_i}{\sum x_i}
\]

This estimator also converges almost surely to \( \tau \).
Theorem 3: Finite population open system (variance)

Let \( \delta_i = \ln \left( \frac{\tau_i}{\tau_{i-1}} \right) \). Assume that \( \sup_i \delta_i \to 0 \). Then an asymptotically minimum variance (=population variance) is attained by the estimator,

\[
\hat{\text{Var}} \tau \to \text{Var} \tau \to
\]

\[
\frac{\sum_{i=1}^n \tau^2 q_i (1-q_i) + (n-i+1)^2 \left( q_i (1-q_i) + s_i \left( 1 - \frac{s_i}{Y_i} \right) + 2C_u \right) + 2\tau(n-i+1) \left[ q_i (1-q_i) + C_u \right]}{\left( \sum_{i=1}^n q_i \right)^2}
\]

\[
+ 2 \sum_{j=1}^n \sum_{i=1}^{i-1} \tau^2 q_i (q_{i-1} - q_i) + (n-j+1)(n-i+1) \left( q_i (q_{i-1} - q_i) + 2C_u + Q_y \right) + 2\tau(n-j+1) \left[ q_i (q_{i-1} - q_i) + C_y \right]
\]

\[
\left( \sum_{i=1}^n q_i \right)^2
\]

\( s_i = \# \) of units marked units leaving at the end of \( i \)th trial

\( k_i = \# \) of units marked units re-entering due to cluster sampling at the end of \( i \)th trial

\[
C_{ij} \to \sum_{s_k=0}^{s_k=Y_{i-1}} q_i s_k \left[ \Phi \left( \frac{Y_{j-1} \left( s_k / s_{j-1} - 1 \right)}{Y_{j-1} - s_{j-1}} \right) - \Phi \left( \frac{Y_{j-1} \left( (s_k - 1) / s_{j-1} - 1 \right)}{Y_{j-1} - s_{j-1}} \right) \right] - q_i s_j
\]
\[ Q_{ij} \rightarrow \sum_{i=0}^{y} \sum_{s=0}^{s_i} s_i s_j \left[ (1 - q_i)A + q_i B \right] - s_j \]

\[ C = \Phi \left( \frac{Y_i s_j - 1}{(Y_i - s_i)} \right) - \Phi \left( \frac{Y_i (s_j - 1) / s_i - 1}{(Y_i - s_i)} \right) \]

\[ A = \Phi \left( \frac{u_i (s_k) + 1}{u_i (s_k) + 1 - s_i} \right) - \Phi \left( \frac{u_i (s_k) + 1}{u_i (s_k) + 1 - s_i} \right) \]

\[ B = \Phi \left( \frac{u_i (s_k) (s_i / s_j - 1)}{u_i (s_k) - s_j} \right) - \Phi \left( \frac{u_i (s_k) (s_i / s_j - 1)}{u_i (s_k) - s_j} \right) \]

\[ u_i (s_k) = Y_i - s_k + 1 + k_i + Y_{j-i} \]

\[ q_i = \frac{Y_i}{\tau} \sum_{i=1}^{n} x_i \rightarrow \sum_{i=1}^{n} q_i \]

A \((1 - \alpha)\) confidence interval of \( E\tau \) is:

\[ CI (E\tau) \rightarrow \tau \pm z_{a/2} \sqrt{\text{Var}\tau} \]
CONCLUSION

We present a novel approach to the capture recapture problem that is different than usual maximum likelihood based models. It has the advantage that it considers ordering, dependence, minimum variance, non parametric and accounts for exits and entries. It has an impressive rate of convergence.

Future research could be done to make this model continuous in time. We could obtain explicit formulae for population estimators and its variance as a function of time. We also suggest different estimation models and estimators that could measure mortality and other parameters of interest to biologists.

Perhaps the test of a good model is how closely it mirrors reality and in that sense we have shown that our model works well. Too many assumptions and parameters to estimate often leads to loss of “reality” and we propose using specialized separate models to measure mortality and other quantities of interest. Perhaps the current maximum likelihood approach makes too many unrealistic assumptions.

Model fitting is useful in parametric models, and it is not necessary for a non parametric model. The true test is the performance of the model in real life situations. Everything else is secondary to this objective.
REFERENCES


# APPENDIX 1: COMPARISON WITH EXISTING MODELS

<table>
<thead>
<tr>
<th>ASSUMPTION</th>
<th>PROBLEM</th>
<th>SOLUTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each individual has equal probability of being captured</td>
<td>This assumption examined by Otis et al. 1987. Sources of variation in individual capture attributed to: <strong>Behavior</strong>: directional change in probability of sighting after initial capture. Possible from frequent interactions with boats and proximity to human activities. Possible response to our presence, but this will be the same presence for all groups <strong>Time</strong>: temporally dependent variables (survey conditions, season, time of day)</td>
<td>None Our approach: no such assumption as we use the average probability of capture in the model.</td>
</tr>
</tbody>
</table>
| Heterogeneity | Carothers (1973). Fletcher et al also says this and suggests that the uniqueness of any given individual will produce differences in probability of capture | Minimize heterogeneity concerns by stratified sampling: no regions were preferentially covered or confounded with time of day.  
Drummer 1999 (elec PDF)- discusses solutions to distribution effort btw intensive small area for cryptic animals vs. larger areas but with lower detection rates. Concludes that larger area should be covered at expense of lower probabilities of sighting (assumes search intensity is linearly related to detectability). Including covariates such as weather or sampling effort may help to partially alleviate concerns (pg 75)  
Our approach: Stratify data using maximum likelihood before running model. |
<table>
<thead>
<tr>
<th>Ordering</th>
<th>No ordering of data is taken into account and this information is lost.</th>
<th>Our approach: Takes ordering into account to improve estimates and confidence intervals.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emigration</td>
<td>Pollock et al. (1990) cited temporary emigration as a serious source of bias</td>
<td>Burnham (1993) has shown how to account for random emigration. But when emigration is permanent, it’s not distinguishable from mortality</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fletcher et al. 2002 (elec PDF) also accounts for local emigration- uses multi-state models</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pradel et al. 1997 (elec PDF) accounts for transients and residents. Assumes residents behave similarly and that movement out of region by transients is relatively synchronous for known period of time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Our approach: No such assumption.</td>
</tr>
<tr>
<td>Dependence (Cluster sampling)</td>
<td>Probability of seeing one individual may be highly related to observance of another (e.g., killer whales, bottlenose dolphins). Consequently, sightings and missed observations may be clustered (cites Cowan and Malec, 1986) Estimates can be improved when non-independence of individuals in population is take into account (Wilson et al. 1999)</td>
<td>Adaptive sampling suggested for improving precision of spatially clumped data (Thompson and Seber 1996, Seber 1999) Increase sampling effort in areas where more individuals are present= allows for more non-zero samples to be recorded. Trade-off between this and inadequate coverage</td>
</tr>
</tbody>
</table>