Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Jiandong Chen Date

Sensitivity and Uncertainty Analysis for Two-Stream Capture-Recapture in Epidemiological Surveillance

By

Jiandong Chen

Degree to be awarded: Master of Science in Public Health

Department of Biostatistics and Bioinformatics

_____

Robert H. Lyles, PhD

Committee Chair

_____

Lance A. Waller, PhD

Committee Member

Sensitivity and Uncertainty Analysis for Two-Stream Capture-Recapture in Epidemiological Surveillance

By

Jiandong Chen

Bachelor of Science
China Jiliang University
2018

Thesis Committee Chair: Robert H. Lyles, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2020

Abstract

Sensitivity and Uncertainty Analysis for Two-Stream Capture-Recapture in Epidemiological Surveillance

By Jiandong Chen

The capture-recapture approach is a well-studied paradigm for estimating wildlife population sizes, based on tag and release strategies. Statistical methods associated with this approach are also used in epidemiological studies to estimate total numbers (N) of cases or deaths from multiple registries. Using simulated data and DRS data on death obtained from a Population Change Survey conducted by the National Statistical Office in Malawi between 1970 to 1972 as examples, sensitivity and uncertainty analyses are proposed and incorporated to provide a more defensible picture of variability in estimates of homogeneous population size when the assumption of list independence fails in the two-capture scenario. In this report, maximum likelihood estimators (MLEs) for population size (N) and the variance of these MLEs are formulated upon fixing the values of key non-identifiable parameters. A discussion is made of the placement of the Lincoln-Petersen (LP) estimate and the estimator of Chao (1987) on the proposed sensitivity analysis plots, and the proposed uncertainty analyses are demonstrated and evaluated through simulations based on two prior assumptions for a key parameter upon which estimation hinges. Some features of the proposed MLEs are also highlighted in this report.

Sensitivity and Uncertainty Analysis for Two-Stream Capture-Recapture in Epidemiological Surveillance

By

Jiandong Chen

Bachelor of Science
China Jiliang University
2018

Thesis Committee Chair: Robert H. Lyles, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2020

Acknowledgments

I could not have accomplished the steps toward my graduation without the love and support I receive from my parents and family. I am indebted to the friendships I have made prior to and since joining Emory. It is through these relationships that I gained a greater appreciation of my role as an individual and team member in the field of public health.

I am grateful to Robert H. Lyles for the direction and guidance he has given me both as an instructor and mentor. His patience, clarity, and enthusiasm for his work inspired me to pursue more challenging, rewarding goals. Working with him was a joy and a privilege, and the source of so much of my progress as a graduate student.

I am grateful to Min Zhu, my undergraduate research mentor at China Jiliang University, for allowing me to contribute to research which truly captured my interests and eased my hesitation in committing to study statistics. It was through her tutelage that I came to appreciate the wide-ranging influence and applicability of statistics.

To the faculty and staff of the Emory Rollins School of Public Health Department of Biostatistics and Bioinformatics, thank you for your time, lectures, assistance, and support.

Table of Contents

1. Introduction 1
2. Methods 3
   2.1 Preliminaries 3
   2.2 Maximum likelihood estimators (MLEs) and variance based on known $\boldsymbol{\psi = p2|1}$ 5
   2.3 Maximum likelihood estimators (MLEs) and variance based on known $\boldsymbol{\phi = p2|1p2|1}$ 6
   2.4 Clarification for the formulas of MLEs based on $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ 7
   2.5 Sensitivity Analysis 8
   2.6 Uncertainty Analysis 9
3. Results 10
4. Discussion 12
5. Tables and Figures 14

Tables 14
   Table 1 14
   Table 2 14
   Table 3 15

Figure 17
   Figure 1 17

Reference 19

1. Introduction

Capture-recapture methods offer well known ways to estimate the size of populations. The classic capture-recapture paradigm is based on tag and release procedures to estimate the number of animals in a specific area [1-3]. In recent years, more and more researchers have begun to apply this idea in other areas such as social science or epidemiology for quantifying unique or vulnerable human population [4-6]. A Dual-Record System (DRS) is a special type of data-structure obtained by capture-recapture experiments to estimate a total number of cases in recent epidemiology studies [7-9]. Fundamentally, this is a missing data problem caused by the unobserved number of units that are not captured in all T surveillance efforts (where T=2 in the case of a DRS). The overall objective is to use the observed capture data, together with assumptions and sometimes augmented by covariate information, to make a defensible estimate of the total population size.

Most common uses of capture-recapture methods based on DRS are typically thought to rely on certain attributes: (i) the population will not change (i.e., is closed) during the DRS process; (ii) all individuals have the same probability during the second of the two capture periods; (iii) capture (or not) in the second period is independent of capture in the first [8, 10]. However, such an independence assumption is often violated in the real world. In other words, an individual who is captured the first time might be less (or more) likely to be captured the second time. Therefore, consideration of a set of conceivable capture-recapture models labeled M_{tbh} [11, 12] has led to a vast literature. In this notation, the *b* subscript makes an attempt to account for behavioral characteristics of individual units that may affect their likelihood of recapture, the subscript *t* refers to the notion that the *T* different capture efforts may have varying overall success in identifying

population members, and the *h* subscript incorporates the opinion that individual units may have their own unique profile of capture probabilities. The well-known Lincoln-Petersen (LP) estimator [1, 2, 13] is based on a special case of M_{tbh} when T=2 and one can ignore the *b* component and *h* component (we refer to this as the "LP condition"). It is the best known and most classic estimator in the two capture case, perhaps rivaled only by a bias-corrected alternative proposed by Chapman (1951) [14].

Chao, Pan and Chiang (2008) shed light on further relaxation of the LP condition and extended the LP method to the case of two populations [15]. Chao et al. (2000, 2008) also published some inference procedures to address capture-recapture problems with unequal catchability and when time and behavioral response affect capture probabilities [16, 17]. Ayhan (2000) provided an estimator improving the underestimation by further dividing the cells of the table from DRS [18]. Along with the development of statistics, more new ideas were used in capture-recapture problems. One of the appealing ones is the Bayesian approach. It was pioneered by Castledine (1981), Smith (1991) and later, by George and Robert (1992) [19-21]. Wang et al. (2007) concluded that the Bayesian approach can provide more accurate estimates of population size than the maximum likelihood estimation (MLE) for small samples [22]. However, it is known that the estimators from Bayesian approaches are sensitive to the priors and the process is more complicated than deriving MLEs. Maximum likelihood estimation in the capture-recapture scenario was developed by Zippin (1956) and Darroch (1958) [23, 24], among others. Two other common approaches are the estimation based on the Poisson model which was proposed by Fienberg (1972) and Cormack (1985, 1989) [25-27] and the

estimation inspired by stochastic processes developed by Godambe (1985) and Lloyd (1987) [28, 29].

In the current article, our focus is on the "two catch" (or two surveillance stream) setting, which is common in epidemiologic and demographic settings. We begin with a section to derive formulae for maximum likelihood estimators (MLEs) for population size (N) and the variance of MLEs based on observed counts and specified values of key parameters under a natural multinomial model for cell counts. This involves a focus on formulating sensitivity analysis plots by using the data of the three observed counts (n_{11}, n_{10}, n_{01}) from real data, providing insight relevant to the LP estimator, an estimator due to Chao (1987) [16, 17] applied to the two-catch case, and our MLEs by charting their locations on the plot. In addition, inspired by Bayesian analytic approaches used in the capture-recapture problem by Chatterjee and Mukherjee (2016) [8], we propose an accessible approach to uncertainty analyses for unknown parameters to provide epidemiologists and demographers with a more realistic and robust assessment of variability in the estimate of N.

2. Methods

2.1 Preliminaries

This thesis will focus on the T=2 case, where the population-level capture recapture experience is drawn from a multinomal distribution. That is, (N_{11}, N_{10}, N_{01}, N_{00}) ~ *Multinomial*(N, p_{11}, p_{10}, p_{01}, p_{00}) where the (0=no, 1=yes) supscripts i and j in p_{ij} denote the first and second surveillance streams separately. The problem is that the cell count of N_{00} is unobserved and cannot be estimated unless an assumption is applied. Therefore, a

new multinomial model based on observed cell counts is introduced, $(N_{11}, N_{10}, N_{01}|N_c = n_c) \sim Multinomial(n_c, p^*_{11}, p^*_{10}, p^*_{01})$, where $N_c = N_{10} + N_{01} + N_{11}, n_c = n_{10} + n_{01} + n_{11}, p_c = p_{10} + p_{01} + p_{11}$ and $p^*_{ij} = \frac{p_{ij}}{p_c}$. In addition, we define $p_1 = p_{11}+p_{10}$, $p_2 = p_{11}+p_{01}$, $p_{2|1} = \frac{p_{11}}{p_1}$ and $p_{2|\bar{1}} = \frac{p_{01}}{1-p_1}$. Here, p_1 and p_2 represent the marginal probabilities that identification happens in the first and second surveillance streams. Also, $p_{2|1}$ is the conditional probability that identification happens in the second surveillance stream given that identification happens in the first surveillance stream. Similarily, $p_{2|\bar{1}}$ is the conditional probabiltity that the second surveillance stream identifies given that the first does not.

Suppose epidemiologists have prior guesses of the parameter $\psi = p_{2|\bar{1}}$ or $\phi = \frac{p_{2|1}}{p_{2|\bar{1}}}$, Maximum likelihood estimators can be derived based on treating either of these parameters as known. Here, the parameter ϕ is clearly a measure of the population-level dependeucy between the first and second surveillance streams. To be more detailed, ϕ can be regarded as the capture relative risk. Note that $\phi > 1$ represents a "trap happy" case (identification in the first surveillance stream will be prone to the identification in the second stream overall), $\phi < 1$ represents a "trap averse" case (identification in the first surveillance stream will reduce the identification in the second stream overall), and $\phi = 1$ yields the LP condition. Under the LP conditions, it is well known (e.g. Seber 1982 [10]) that one derives the same MLE for N under a hypergeometric or a multinomial model for the population-level data; this simple closed form estimator is the Lincoln-Petersen estimator:

$$\widehat{N}_{LP} = \frac{n_{1\cdot} n_{\cdot 1}}{n_{11}} \tag{1}$$

Where $n_{1\cdot} = n_{11} + n_{10}$ and $n_{\cdot 1} = n_{11} + n_{01}$. The same corresponding multivariate delta-method variance also applies under either model (e.g. Seber 1982 [10]):

$$\widehat{N}_{LP} = \frac{n_{1\cdot} n_{\cdot 1} n_{10} n_{01}}{{n_{11}}^3} \tag{2}$$

In this regard, one of the most prominent alternative estimators is the Chao estimator. It is an estimator for the T-catch case that exists a lower bound for N under certain mathematical conditions, and these conditions indicate that the estimator is prepared for the case of "large" T and "small" capture probabilities (p_i). In addition, the Chao estimator also assumes that capture probabilities are the same for a given unit at every capture event but vary arbitrarily across units. The Chao estimator is given by

$$\widehat{N}_{Chao} = n_c + \frac{(N_{10} + N_{01})^2}{2N_{11}} \tag{3}$$

Where n_c is the total number of units captured at least once; Chao (1987) also gave an approximate variance for the Chao estimator [17].

2.2 Maximum likelihood estimators (MLEs) and variance based on known $\boldsymbol{\psi = p_{2|\bar{1}}}$

Based on the multinomial model proposed above, the likelihood for the observed data can be expressed by:

$$L = {p_{11}^*}^{n_{11}} {p_{10}^*}^{n_{10}} {p_{01}^*}^{n_{01}}$$

It follows that

$$\ln(L) = n_{11} \ln\left(\frac{p_{2|1} p_1}{p_1 + \psi(1 - p_1)}\right) + n_{10} \ln\left(\frac{(1 - p_{2|1}) p_1}{p_1 + \psi(1 - p_1)}\right) + n_{01} \ln\left(\frac{\psi(1 - p_1)}{p_1 + \psi(1 - p_1)}\right)$$

If we take the derivatives with respect to p_1 and $p_{2|1}$, MLEs for $p_{2|1}$ and p_1 can be expressed as $\frac{n_{11}}{n_{11}+n_{10}}$ and $\frac{\psi(n_{11}+n_{10})}{\psi(n_{11}+n_{10})+n_{01}}$ respectively. Furthermore, the MLE for the population size (N) can be calculated as:

$$\widehat{N} = \frac{n_c}{\hat{p}_c} = \frac{n_{11}+n_{10}+n_{01}}{\hat{p}_{11}+\hat{p}_{10}+\hat{p}_{01}} = \frac{n_{11}+n_{10}+n_{01}}{\hat{p}_{2|1}\hat{p}_1+(1-\hat{p}_{2|1})\hat{p}_1+\psi(1-\hat{p}_1)} = n_{11}+n_{10}+\frac{n_{01}}{\psi} \quad (4)$$

In order to calculate the variance of the MLE, we apply the multivariate delta method:

$$Var(\widehat{N}) = \widehat{\boldsymbol{D}}'\boldsymbol{Var}\begin{pmatrix}\boldsymbol{n_{11}}\\ \boldsymbol{n_{10}}\\ \boldsymbol{n_{01}}\end{pmatrix}\widehat{\boldsymbol{D}} \ \boldsymbol{,where} \quad \widehat{\boldsymbol{D}}' = \begin{pmatrix}\frac{\partial\widehat{\mathrm{N}}}{\partial \mathrm{n}_{11}} & \frac{\partial\widehat{\mathrm{N}}}{\partial \mathrm{n}_{10}} & \frac{\partial\widehat{\mathrm{N}}}{\partial \mathrm{n}_{01}}\end{pmatrix} = \begin{pmatrix}1 & 1 & \frac{1}{\psi}\end{pmatrix}$$

$$\textbf{and } \boldsymbol{Var}\begin{pmatrix}\boldsymbol{n_{11}}\\ \boldsymbol{n_{10}}\\ \boldsymbol{n_{01}}\end{pmatrix} = \begin{pmatrix}\widehat{N}\hat{P}_{11}(1-\hat{P}_{11}) & -\widehat{N}\hat{P}_{11}\hat{P}_{10} & -\widehat{N}\hat{P}_{11}\hat{P}_{01}\\ -\widehat{N}\hat{P}_{11}\hat{P}_{10} & \widehat{N}\hat{P}_{10}(1-\hat{P}_{10}) & -\widehat{N}\hat{P}_{10}\hat{P}_{01}\\ -\widehat{N}\hat{P}_{11}\hat{P}_{01} & -\widehat{N}\hat{P}_{01}\hat{P}_{10} & \widehat{N}\hat{P}_{01}(1-\hat{P}_{01})\end{pmatrix}$$

and the variance-covariance matrix takes the usual multinomial form. After algebraic simplification, the final result shows that the variance of MLE has a closed form:

$$Var(\widehat{N}) = \frac{(1-\psi)n_{01}}{\psi^2} \quad (5)$$

2.3 Maximum likelihood estimators (MLEs) and variance based on known $\boldsymbol{\phi} = \frac{\mathbf{p_{2|1}}}{\mathbf{p_{2|\bar{1}}}}$

The same idea is applied for constructing the likelihood in this case, with the only difference being that ψ is replaced by $\frac{\mathrm{p}_{2|1}}{\phi}$ here. The MLEs for $p_{2|1}$ and p_1 can be expressed as $\frac{n_{11}}{n_{11}+n_{10}}$ and $\frac{n_{11}}{n_{11}+\phi n_{01}}$ respectively. The final result shows that the MLE for N is equal to

$$\widehat{N} = \frac{(n_{11}+\phi n_{01})(n_{11}+n_{10})}{n_{11}} \quad (6)$$

The process of calculating the variance of the MLE is more tedious this time. In this case, $\widehat{\boldsymbol{D}}'$ can be written as follows:

$$\widehat{\boldsymbol{D}}' = \left(1 - \frac{\phi n_{01} n_{10}}{n_{11}^2} \quad 1 + \frac{\phi n_{01}}{n_{11}} \quad \phi + \frac{\phi n_{10}}{n_{11}}\right)$$

Further, based on the properties of matrix operations, $\widehat{\boldsymbol{D}}'$ can be re-written as

$$\widehat{\boldsymbol{D}}' = \boldsymbol{A}' + \boldsymbol{B}', \boldsymbol{where} \quad \boldsymbol{A}' = (\mathbf{1} \quad \mathbf{1} \quad \phi) \ \boldsymbol{and} \quad \boldsymbol{B}' = \left(\frac{-\phi n_{01} n_{10}}{n_{11}^2} \quad \frac{\phi n_{01}}{n_{11}} \quad \frac{\phi n_{10}}{n_{11}}\right)$$

Subsequently, $Var(\widehat{N})$ can be simplified as

$$Var(\widehat{N}) = \widehat{\boldsymbol{D}}' \boldsymbol{Var} \begin{pmatrix} \boldsymbol{n_{11}} \\ \boldsymbol{n_{10}} \\ \boldsymbol{n_{01}} \end{pmatrix} \widehat{\boldsymbol{D}} = (\boldsymbol{A}' + \boldsymbol{B}') \boldsymbol{Var} \begin{pmatrix} \boldsymbol{n_{11}} \\ \boldsymbol{n_{10}} \\ \boldsymbol{n_{01}} \end{pmatrix} (\boldsymbol{A} + \boldsymbol{B})$$

$$= \boldsymbol{A}' \boldsymbol{Var} \begin{pmatrix} \boldsymbol{n_{11}} \\ \boldsymbol{n_{10}} \\ \boldsymbol{n_{01}} \end{pmatrix} \boldsymbol{A} + \boldsymbol{A}' \boldsymbol{Var} \begin{pmatrix} \boldsymbol{n_{11}} \\ \boldsymbol{n_{10}} \\ \boldsymbol{n_{01}} \end{pmatrix} \boldsymbol{B} + \boldsymbol{B}' \boldsymbol{Var} \begin{pmatrix} \boldsymbol{n_{11}} \\ \boldsymbol{n_{10}} \\ \boldsymbol{n_{01}} \end{pmatrix} \boldsymbol{A} + \boldsymbol{B}' \boldsymbol{Var} \begin{pmatrix} \boldsymbol{n_{11}} \\ \boldsymbol{n_{10}} \\ \boldsymbol{n_{01}} \end{pmatrix} \boldsymbol{B}$$

$$= \boldsymbol{A}' \boldsymbol{Var} \begin{pmatrix} \boldsymbol{n_{11}} \\ \boldsymbol{n_{10}} \\ \boldsymbol{n_{01}} \end{pmatrix} \boldsymbol{A} + \mathbf{2} \boldsymbol{A}' \boldsymbol{Var} \begin{pmatrix} \boldsymbol{n_{11}} \\ \boldsymbol{n_{10}} \\ \boldsymbol{n_{01}} \end{pmatrix} \boldsymbol{B} + \boldsymbol{B}' \boldsymbol{Var} \begin{pmatrix} \boldsymbol{n_{11}} \\ \boldsymbol{n_{10}} \\ \boldsymbol{n_{01}} \end{pmatrix} \boldsymbol{B}$$

After some considerable algebra and efforts to simplify, we arrive at the following variance for the MLE in the case of ϕ known:

$$Var(\widehat{N}) = \frac{\widehat{N} \hat{P}_{10} \hat{P}_{01} \phi^2}{\hat{P}_{11}^2} \left(\frac{\hat{P}_{11}^2}{\phi}\left(1 - \frac{1}{\hat{P}_{10}} + \frac{\phi}{\hat{P}_{10}}\right) + \hat{P}_{11}\left(\frac{\hat{P}_{10}}{\phi} + \hat{P}_{01} + 2\left(1 - \frac{1}{\phi}\right)\right) + \right.$$

$$\left. \hat{P}_{10} \hat{P}_{01}\left(1 + \frac{1}{\hat{P}_{11}}\right) + \hat{P}_{10} + \hat{P}_{01}\right) \tag{7}$$

2.4 Clarification for the formulas of MLEs based on ψ and ϕ

We note that the value of the maximized log-likelihood is identical regardless of the value of ψ or ϕ assumed in sections 2.2 and 2.3. That is, there is no

information in the observed data alone to identify these parameters, and any value of N greater than or equal to n_c is in fact equally consistent with the observed data. Also note, for example, that the famous Lincoln-Petersen (LP) estimator is exactly equal to eqn. (6) upon taking $\phi = 1$. Similarly, the delta method variance that we derived in eqn. (7) is equivalent to the well-known variance of the LP estimator when $\phi = 1$.

2.5 Sensitivity Analysis

A publicly available set of DRS data on migration, death and birth obtained from a Population Change Survey conducted by the National Statistical Office in Malawi between 1970 to 1972 is reported by Greenfield [30]. Papers by Nour (1982) and Chatterjee and Mukherjee (2016) also used this data for illustration [8, 31]. Table 1 shows data on death records from Lilongwe and other urban areas chosen to plot the MLEs for N with the corresponding error bars (± 1.96 standard errors) against the parameters ψ and ϕ (see eqns. (4-7)). The plots represent a sensitivity analysis to reflect how the estimates of total population and their variability change with ψ and ϕ. In addition, the LP and Chao (1987) estimates are also located on the plots for reference. However, as noted previously, the specific value of ϕ cannot be inferred from the data alone. In this regard, only the LP estimate might be specifically justified, since only $\phi = 1$ could be accurately targeted by design in practice (for example, if one surveillance stream could be implemented as a simple random sample of the population that is collected without reference to the other, potentially non-random, surveillance effort).

2.6 Uncertainty Analysis

In this section, distributions are assumed for ψ and ϕ, then simulations are used to evaluate the corresponding realistic and robust assessment of variability in the estimate of N. Note that these are akin to Bayesian prior distributions, except that there is no information in the observed data to update them unless assumptions are made. As an initial proof of concept, suppose $p_{2|1}$ follows a Jeffreys Beta(0.5, 0.5) prior distribution, and assume the LP conditions so that $\psi = p_{2|1}$. It is then easily shown that the conjugate beta posterior distribution for ψ is as follows:

$$\psi \sim \boldsymbol{Beta}(n_{11} + 0.5, n_{10} + 0.5) .$$

100,000 ψs were randomly generated from this distribution, where the 100,000 results of $\widehat{N}$ and its $Var(\widehat{N})$ can be acquired by eqn. (4) and eqn. (5) according to the two Table 1 datasets. We use the mean of the estimated $Var(\widehat{N})$ to represent the within variance, and the sample variance of the $\widehat{N}$ to represent the between variance. Then, we calculate a total variance (B+W) based on Rubin's approach [32] in the context of multiple imputation, which is the sum of the within (W) variance and between (B) variance. The simulation is designed to assess the validity of this approach to estimating the actual variance of N under the LP conditions. The mean of the $\widehat{N}$ values is also reported, and compared with the LP estimator and its typical delta method-based variance (e.g., Seber [10]).

In a second set of simulation studies, we temporarily set true population size (N) to 500, 1,000, 5,000 and 10,000, the probability of being found in the first capture (p_1) to 0.1 and the probability of being found in the second capture (p_2) to 0.25. Based on the information above and assuming the LP conditions, initial values of p_{11}, p_{10}, p_{01} and

p_{00} can be calculated. Subsequently, 500 multinomial datasets are generated with cell probabilities (p_{11}, p_{10}, p_{01} and p_{00}). Subsequently, only the three observable cell counts n_{11}, n_{10} and n_{01} are used. We then randomly generate a "true" value for ϕ from a uniform "prior" distribution, as follows:

$$\phi \sim \mathbf{U}(0.75, 1.25)$$

Note that this ϕ is centered at 1 under this uniform distribution, which is the case of the LP conditions. After drawing the "true" value of ϕ from the uniform distribution, a corresponding new value of p_c (the probability of being caught) is calculated by

$$p_c^* = \frac{P_{11}}{P_{11}^2 + P_{11}P_{10} + \phi P_{01}(P_{11} + P_{10})}$$

The "new" corresponding true N becomes the original n_c divided by the p_c^*. The purpose here is to mimic reality, where the epidemiologist does not know ϕ but is willing to specify a distribution in order to acknowledge uncertainty in ϕ. Then, 100,000 random ϕ values were drawn from the same uniform distribution and utilized with the original 3 cell counts (n_{11}, n_{10}, n_{01}) to calculate the MLE for N assuming ϕ equals to the value drawn. The variance estimate is again B+W, and we assess coverage of a 95% CI for N as:

$$\widehat{\bar{N}} \pm 1.96 * sqrt(\mathrm{B} + \mathrm{W})$$

3. Results

As Figure 1 shows, $\widehat{N}$ and the variance of $\widehat{N}$ are monotonically increasing as ϕ becomes larger but decreasing as ψ becomes larger. It makes sense because

ϕ is equal to $\frac{p_{2|1}}{\psi}$. When ϕ becomes larger, ψ will be smaller. The LP estimate is always

located where $\phi = 1$. One explanation based on probability is that when $p_{2|1} = p_{2|\bar{1}}$ at the population level, it is a sufficient LP condition. In addition, the plot also shows that the Chao estimate projects a value of $\phi > 1$, which in the case of this example is approximately 1.1 times the LP estimate. The Chao (1987) estimator as applied to the two-catch case always estimates a larger population size (sometimes dramatically so) compared with the LP estimator.

Table 2 summarizes some key output of the uncertainly analysis with a beta posterior distribution for ϕ. Not only is the mean of the MLEs very close to the LP estimator, but also the variance based on Rubin's approach matches the variance of the LP estimator based on eqn. (2). This conclusion can be seen as a validation that applying Rubin's approach to get the total variance in this scenario is reasonable.

Table 3 shows the uncertainty analysis simulation results based on a uniform distribution. The true N used to generated two-capture datasets is 500 in Table 3A, 1000 in Table 3B, 5,000 in Table 3C and 10,000 in Table 3D. All tables arise from 500 two-capture dataets with 100,000 random ϕ values. Originally it seemed reasonable to expect the $SE_{Rubin}(\widehat{N}_{MLE})$ should be close to the $SD(N)$. However, it is not the case. Instead, the square root of $Var_{between}(\widehat{N}_{MLE})$ (equivalently, the square root of B) is close to the $SD(N)$. From my perspective it is sensible, because the between variance contributes the most of variability of true N during the simulation process. In addition, the difference between $SD(\overline{\widehat{N}_{MLE}})$ and $SD(\widehat{N}_{LP})$ is small. The reason is likely that the uniform distribution that was assumed here is centered at 1, which is the case of the LP condition, leading to a very close result. In a addition, convergence of the proposed CIs centered around $\overline{\widehat{N}_{MLE}}$ is 93.60% in Table 3A, 95.60% in Table 3B, 97.20% in Table 3C and

96.60% in Table 3D, which corresponds to near-nominal (95%) coverage in each case. All convergences look reasonable, which again serves as a proof of concept for the application of the Rubin-type approach to variance estimation in this context.

4. Discussion

In this article focused on the standard closed population single-recapture scenario, we first express our idea that the capture-recapture problem can be solved through the maximum likelihood estimators (MLEs) based on assumed values of the key parameters ψ and ϕ. One advantage of this method is that it illustrates the fact that the maximized log-likelihood value is identical for any admissible value of ψ and ϕ that one chooses to specify. Therefore, as long as N is greater or equal to the number of distinct units observed by the two streams, it can be seen to be as consistent with the observed data as any other. We note that the parameters ψ and ϕ are measures of the population-level dependency between capture events. As a result, this method can be applied in the scenario where the LP estimator is not suitable, although we are generally limited to sensitivity and uncertainty analyses because the true value of ϕ is not identifiable based on the observed data. In this sense, we still argue that the LP estimator remain central, because only the population-level state of nature $\phi = 1$ can specifically be targeted by design or defended epidemiologically. No statistical information to identify ϕ in fact exists without assumptions best judged by those with a true understanding of the operating characteristics of the two streams, and we suggest that any reported estimate for N should be accompanied by a clear discussion of its implications about ψ and ϕ. From our perspective, a reliable point estimate for N in the two capture surveillance setting is produced only when the epidemiologist can project a reasonable guess for ψ or ϕ. It

might be difficult to get the best one in real life, but it is relatively reasonable for epidemiologists to be able to give a range for our parameters, which leads to our sensitivity and uncertainty analysis. If a range for our parameter ϕ is given, it becomes easy to insert the minimum and maximum values of ϕ into eqn. (6) to get the estimate range for N.

While we emphasize the need for caution in using it in practice with only two surveillance streams, we recognize the historical significance of the Chao (1987) estimator. It is important to note that this estimator was derived under mathematical conditions quite different from the likely reality in two-stream surveillance. Unfortunately and as a result, the notion that $\widehat{N}_{Chao}$ can be defended as a general lower bound for N in the capture-recapture case can cause misguidance. If investigators are quite sure ϕ is greater than one at the population level, it is actually $\widehat{N}_{LP}$ that serves as a lower bound instead of $\widehat{N}_{Chao}$. Similarly, if the LP condition is defensible at the population level, it follows that $\widehat{N}_{Chao}$ is always biased upward because it necessarily projects a value of $\phi > 1$. The problem will be more pronounced when $\phi < 1$, as in that case the Chao estimator is not only an upper (rather than lower) bound, but also is severely biased upward.

With regard to the uncertainty analysis, the approach we have proposed is somewhat similar in spirit to the Bayesian method presented by Chatterjee and Mukherjee [8]. The advantage of our approach, we believe, is its relative clarity and ease of implementation by the practicing epidemiologist. While we suspect that our approach may more accurately reflect the true uncertainty in the estimate of N over a specified

"prior" distribution for ϕ, we leave a more thorough investigation of this question for future research.

5. Tables and Figures

Table 1. Motivating data on death records in Greenfield's paper [30]

A. Lilongwe's death record					B. Other urban areas' death record			
	Found in 1st capture					Found in 1st capture		
Found in 2nd capture	Yes	No	Total		Found in 2nd capture	Yes	No	Total
Yes	n_{11} = 192	n_{01} = 24	$n_{.1}$ = 216		Yes	n_{11} = 1645	n_{01} = 805	$n_{.1}$ = 1861
No	n_{10} = 132	n_{00} =?	?		No	n_{10} = 315	n_{00} =?	?
Total	$n_{1.}$ = 324	?	N = ?		Total	$n_{1.}$ = 1960	?	N = ?

Table 2. Uncertainty analysis results with beta posterior distribution for ψ by using the motivating data on death records in Greenfield's paper

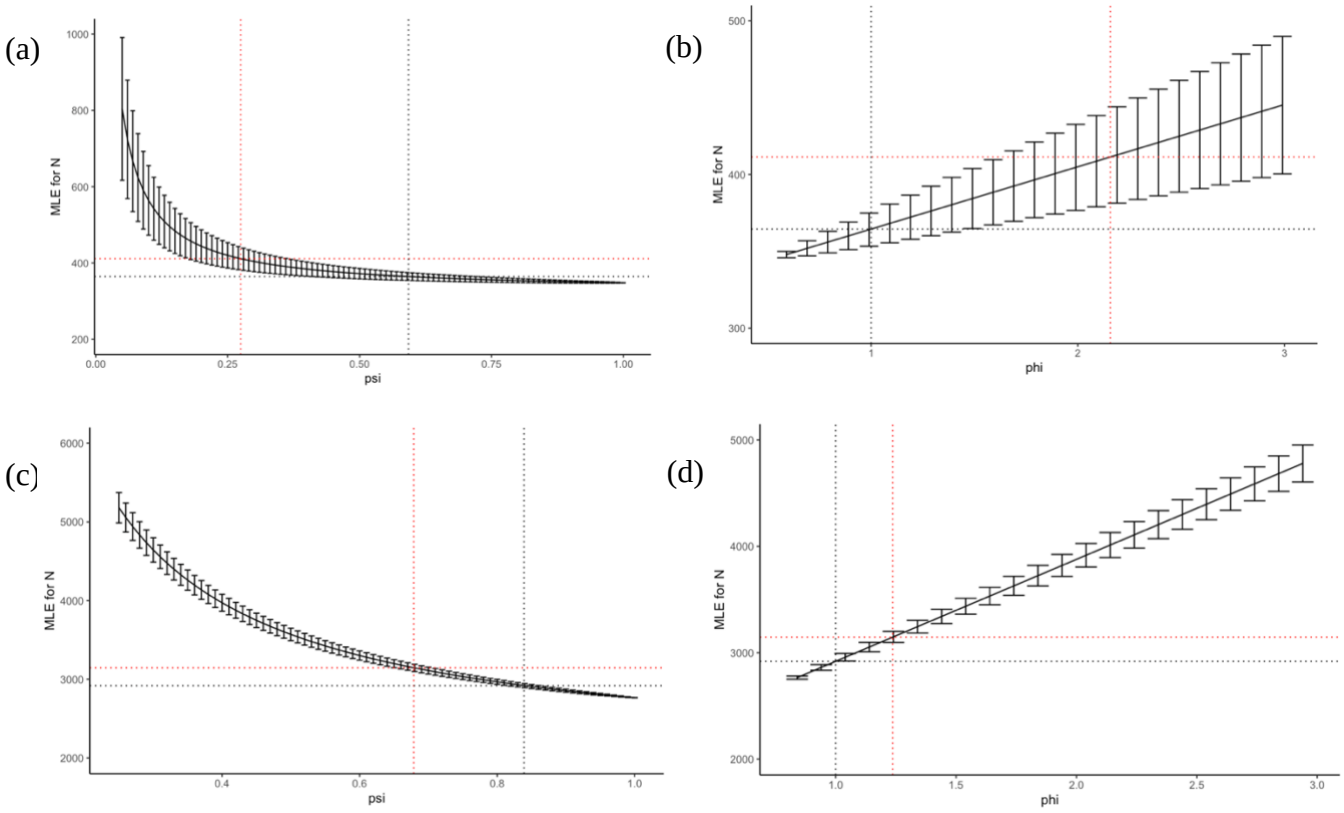
A. Lilongwe's death record			B. Other urban areas' death record	
$Var_{within}(\widehat{N}_{MLE})$	28.234		$Var_{within}(\widehat{N}_{MLE})$	184.167
$Var_{between}(\widehat{N}_{MLE})$	3.524		$Var_{between}(\widehat{N}_{MLE})$	89.794
$Var_{Rubin}(\widehat{N}_{MLE})$	31.758		$Var_{Rubin}(\widehat{N}_{MLE})$	273.961
$SE_{Rubin}(\widehat{N}_{MLE})$	5.635		$SE_{Rubin}(\widehat{N}_{MLE})$	16.552
$\widehat{N}_{LP}$	364.500		$\widehat{N}_{LP}$	2919.150
$SE(\widehat{N}_{LP})$	5.597		$SE(\widehat{N}_{LP})$	16.539
$\overline{\widehat{N}_{MLE}}$	364.602		$\overline{\widehat{N}_{MLE}}$	2919.430

Table 3. Uncertainty analysis results with uniform posterior distribution for ϕ

A. Original true N=500			B. Original true N=1000	
$Var_{within}(\widehat{N}_{MLE})$	26967.95		$Var_{within}(\widehat{N}_{MLE})$	35724.91
$SE_{within}(\widehat{N}_{MLE})$	164.22		$SE_{within}(\widehat{N}_{MLE})$	189.01
$Var_{between}(\widehat{N}_{MLE})$	5260.64		$Var_{between}(\widehat{N}_{MLE})$	18922.17
$SE_{between}(\widehat{N}_{MLE})$	72.53		$SE_{between}(\widehat{N}_{MLE})$	137.56
$Var_{Rubin}(\widehat{N}_{MLE})$	32228.60		$Var_{Rubin}(\widehat{N}_{MLE})$	54647.08
$SE_{Rubin}(\widehat{N}_{MLE})$	151.60		$SE_{Rubin}(\widehat{N}_{MLE})$	225.17
N	500.63		N	1006.77
$SD(N)$	63.80		$SD(N)$	132.00
$\widehat{N}_{LP}$	526.72		$\widehat{N}_{LP}$	1034.96
$SD(\widehat{N}_{LP})$	160.18		$SD(\widehat{N}_{LP})$	184.42
$SE(\widehat{N}_{LP})$	133.11		$SE(\widehat{N}_{LP})$	177.88
$\overline{\widehat{N}_{MLE}}$	526.71		$\overline{\widehat{N}_{MLE}}$	1034.97
$SD(\overline{\widehat{N}_{MLE}})$	160.18		$SD(\overline{\widehat{N}_{MLE}})$	184.41
$Convergence\ of\ \widehat{N}_{MLE}$	93.60%		$Convergence\ of\ \widehat{N}_{MLE}$	95.60%
A. Original true N=5000			B. Original true N=10000	
$Var_{within}(\widehat{N}_{MLE})$	145624.74		$Var_{within}(\widehat{N}_{MLE})$	277519.28
$SE_{within}(\widehat{N}_{MLE})$	381,61		$SE_{within}(\widehat{N}_{MLE})$	526.80
$Var_{between}(\widehat{N}_{MLE})$	432449.48		$Var_{between}(\widehat{N}_{MLE})$	1687734.71
$SE_{between}(\widehat{N}_{MLE})$	657.61		$SE_{between}(\widehat{N}_{MLE})$	1299.13
$Var_{Rubin}(\widehat{N}_{MLE})$	578074.22		$Var_{Rubin}(\widehat{N}_{MLE})$	1965254.00
$SE_{Rubin}(\widehat{N}_{MLE})$	756.78		$SE_{Rubin}(\widehat{N}_{MLE})$	1399.04
N	5018.80		N	10028.82
$SD(N)$	659.40		$SD(N)$	1322.99

$\widehat{N}_{LP}$	5038.26		$\widehat{N}_{LP}$	9986.90
$SD(\widehat{N}_{LP})$	382.47		$SD(\widehat{N}_{LP})$	531.98
$SE(\widehat{N}_{LP})$	374.19		$SE(\widehat{N}_{LP})$	518.98
$\overline{\widehat{N}_{MLE}}$	5038.31		$\overline{\widehat{N}_{MLE}}$	9986.98
$SD(\overline{\widehat{N}_{MLE}})$	382.53		$SD(\overline{\widehat{N}_{MLE}})$	531.71
$Convergence\ of\ \widehat{N}_{MLE}$	97.2%		$Convergence\ of\ \widehat{N}_{MLE}$	96.6%

Figure 1. a) The MLE for N in eqn. (1) as a function of the assumed value $\psi = p_{2|\bar{1}}$, based on observed data in Table 1A. b) The MLE for N in eqn. (3) as a function of the assumed value $\phi = p_{2|1}/p_{2|\bar{1}}$, based on observed data in Table 1A. c) The MLE for N in eqn. (1) as a function of the assumed value $\psi = p_{2|\bar{1}}$, based on observed data in Table 1B. d) The MLE for N in eqn. (3) as a function of the assumed value $\phi = p_{2|1}/p_{2|\bar{1}}$, based on observed data in Table 1B.



* Error bars indicate ±1.96 times estimated standard errors

** Dashed lines drawn to indicate estimates of N (365 and 411) and ψ (0.593 and 0.275) for (a), estimates of N (365 and 411) and ϕ (1 and 2.157) for (b), estimates of N (2919 and 3146) and ψ (0.839 and 0.679) for (c), and estimates of N (2919

and 3146) and ϕ (1 and 1.237) for (d), corresponding to the LP and Chao (1987) estimators, $\frac{(n_{11}+n_{10})(n_{11}+n_{01})}{n_{11}+0.5}$ and $n_{11}+n_{10}+n_{01}+\frac{(n_{10}+n_{01})^2}{2n_{11}}$ respectively.

Reference

1. Petersen, C.G.J., *The yearly immigration of young plaice into the Limfjord from the German Sea, ect.* Report of the Danish Biological Station for 1985, 1986. **6**: p. 1-48.
2. Lincoln, F.C. and U.S.D.o. Agriculture, *Calculating Waterfowl Abundance on the Basis of Banding Returns*. 1930: U.S. Government Printing Office.
3. Schnabel, Z.E., *The Estimation of the Total Fish Population of a Lake.* The American Mathematical Monthly, 1938. **45**(6): p. 348-352.
4. Krebs, C.J., C.L. KREBS, and P.Z.C.J. Krebs, *Ecological Methodology*. 1999: Benjamin/Cummings.
5. Zucchini, D.L.B.S.T.B.W., et al., *Estimating Animal Abundance: Closed Populations*. 2002: Springer.
6. Bohning, D., P.G.M. van der Heijden, and J. Bunge, *Capture-Recapture Methods for the Social and Medical Sciences*. 2017: CRC Press.
7. Chatterjee, K. and D. Mukherjee, *An improved integrated likelihood population size estimation in Dual-record System.* Statistics & Probability Letters, 2016. **110**: p. 146-154.
8. Chatterjee, K. and D. Mukherjee, *On the estimation of homogeneous population size from a complex dual-record system.* Journal of Statistical Computation and Simulation, 2016. **86**(17): p. 3562-3581.
9. Stephen, C., *Capture-recapture methods in epidemiological studies.* Infect Control Hosp Epidemiol, 1996. **17**(4): p. 262-6.

10. Seber, G.A.F., *The estimation of animal abundance and related parameters*. 1982: Charles Griffin.

11. Otis, D.L., et al., *Statistical inference from capture data on closed animal populations.* Wildlife Monographs, 1978(62): p. 3-135.

12. Evans, M.A., D.G. Bonett, and L.L. McDonald, *A general theory for modeling capture-recapture data from a closed population.* Biometrics, 1994. **50**(2): p. 396-405.

13. Sekar, C.C. and W.E. Deming, *On a Method of Estimating Birth and Death Rates and the Extent of Registration.* Journal of the American Statistical Association, 1949. **44**(245): p. 101-115.

14. Chapman, D.G., *Some properties of the hypergeometric distribution with applications to zoological sample censuses*. 1951: University of California Press.

15. Chao, A., H.Y. Pan, and S.C. Chiang, *The Petersen-Lincoln estimator and its extension to estimate the size of a shared population.* Biom J, 2008. **50**(6): p. 957-70.

16. Chao, A., W. Chu, and C.H. Hsu, *Capture-Recapture When Time and Behavioral Response Affect Capture Probabilities.* Biometrics, 2000. **56**: p. 427-33.

17. Chao, A., *Estimating the Population Size for Capture-Recapture Data with Unequal Catchability.* Biometrics, 1987. **43**(4): p. 783-791.

18. Ayhan, H.O., *Estimators of vital events in dual-record systems.* Journal of Applied Statistics, 2000. **27**(2): p. 157-169.

19. Castledine, B.J., *A Bayesian Analysis of Multiple-Recapture Sampling for a Closed Population.* Biometrika, 1981. **68**(1): p. 197-210.

20. Smith, P.J., *Bayesian Analyses for a Multiple Capture-Recapture Model.* Biometrika, 1991. **78**(2): p. 399-407.

21. George, E.I., *Capture—recapture estimation via Gibbs sampling.* Biometrika, 1992. **79**(4): p. 677-683.

22. Wang, X., C.Z. He, and D. Sun, *Bayesian population estimation for small sample capture-recapture data using noninformative priors.* Journal of Statistical Planning and Inference, 2007. **137**(4): p. 1099-1118.

23. Zippin, C., *An Evaluation of the Removal Method of Estimating Animal Populations.* Biometrics, 1956. **12**(2): p. 163-189.

24. Darroch, J.N., *THE MULTIPLE-RECAPTURE CENSUS: I. ESTIMATION OF A CLOSED POPULATION.* Biometrika, 1958. **45**(3-4): p. 343-359.

25. Fienberg, S.E., *The multiple recapture census for closed populations and incomplete 2k contingency tables.* Biometrika, 1972. **59**(3): p. 591-603.

26. Cormack, R.M., *Log-Linear Models for Capture-Recapture.* Biometrics, 1989. **45**(2): p. 395-413.

27. Cormack, R., *Examples of the Use Of Glim to Analyse Capture-Recapture Studies*. 1985. p. 243-273.

28. Godambe, V.P., *The foundations of finite sample estimation in stochastic processes.* Biometrika, 1985. **72**(2): p. 419-428.

29. Lloyd, C.J., *Optimal martingale estimating equations in a stochastic process.* Statistics & Probability Letters, 1987. **5**(6): p. 381-387.

30. Greenfield, C.C., *On the Estimation of a Missing Cell in a 2 × 2 Contingency Table.* Journal of the Royal Statistical Society. Series A (General), 1975. **138**(1): p. 51-61.

31. Nour, E.-S., *On the Estimation of the Total Number of Vital Events with Data from Dual Collection Systems.* Journal of the Royal Statistical Society. Series A (General), 1982. **145**(1): p. 106-116.

32. Little, R.J.A. and D.B. Rubin, *Statistical Analysis with Missing Data*. 2002: Wiley.