Multi-List Methods in Closed Populations with Stratified or Incomplete Information

by

Jason Murray Sutherland

M.Sc. (Statistics), Simon Fraser University, 1997

B.A. (Mathematics), University of British Columbia, 1992

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in the Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY

April 2003

APPROVAL

Name: Jason Murray Sutherland

Degree: Doctor of Philosophy

Title of thesis: Multi-List Methods in Closed Populations with Stratified or Incomplete Information

Examining Committee: Dr. Michael Stephens
Chair

Dr. Carl J. Schwarz
Senior Supervisor

Dr. Richard Lockhart

Dr. Tim B. Swartz

Dr. Rick Routledge
Internal External Examiner

Dr. Louis-Paul Rivest
External Examiner
Université Laval

Date Approved: _____

Abstract

Capture-recapture, or multi-list methods, are used by investigators to estimate the unknown size of a target population whose size cannot be reasonably enumerated. This thesis presents three new methods to estimate population sizes when lists are only partially available or where there is incomplete information available regarding individuals on lists. These methods can assist with population estimation problems occurring in technological, ecological and biological sciences, as well as in epidemiological and public health settings.

First, stratification of lists has often been used to reduce the biases caused by heterogeneity in the probability of list membership among members of a target population. A method is developed to deal with cases when not all lists are active in all strata. Using a log-linear modelling framework, list dependencies and differential probabilities of ascertainment are incorporated. The methodology uses an EM algorithm and is applied to three examples; estimating the number of people with diabetes, the number of people who misuse drugs, and the number of forest fires that occurred in recent history.

Second, a key assumption in multi-list methods is that individuals have a unique "tag" that allows them to be matched across all lists. A multi-list methodology that relaxes the assumption of a single tag common to all lists is developed. The proposed methodology is akin to tag-loss methods, in that it uses supplementary information to improve estimates. Population parameter estimates are found using estimating functions. An example illustrates its application to estimating the prevalence of diabetes and a simulation study investigates conditions under which the methodology is robust to different list and population sizes.

Third, existing multi-list methods assume that an individual's "tag" uniquely identifies an individual. A methodology for multi-list methods when not all list members can be matched across lists because of missing tags or partially missing tags is developed. Estimating functions are used to derive estimates of population size, capture probabilities and rates of tag loss. An example is used to illustrate the methodology and how to select among competing models. A simulation study compares our results against methods based on complete information.

Acknowledgements

I wish to acknowledge the support of various individuals and organizations during my studies. In particular, my senior supervisor, Dr. Carl J. Schwarz, provided continual encouragement and mentorship that guided my academic development. His advice and support on academic, professional and personal issues is recognized and appreciated. Carl's dedication to his students and his profession are to be envied.

I am indebted to the Department of Statistics and Actuarial Science of Simon Fraser University. With the support of faculty and staff, the Department provided the freedom and flexibility to continue full-time employment while completing their doctoral program.

Two institutions provided the organizational support necessary to complete my doctoral degree. I first drew inspiration to complete my doctoral work from Health Canada, where the calibre and dedication of employees is admired. The Canadian Institute for Health Information (CIHI) is an organization where education and personal development among employees is valued and supported. The support of both institutions has been invaluable.

I would like to recognize the special contributions of the following people: Sheri Adams (and family), Ian Affleck, Daniel Benoit, Norman Cameron, Mike Chambers, Ann (Rob) Jolly, Andre Lalonde, Bianca Lang, Linda (Steve) Liberatore, Jason Loeppky, Brett Merkley, Ashok Modha, Louise Ogilvie, Michael Stephens, Jill Strachan, Donald Sutherland, Nancy White, Lisa Whiteside and Ping Yan.

Many friends, colleagues and fellow students provided ongoing encouragement to complete this thesis. With their support, and that of my parents and family, successfully completing this degree was possible.

Contents

Approval Page ii
Abstract iii
Acknowledgements v
List of Tables ix
List of Figures xii
1 Introduction 1
1.1 Background 1
1.2 Closed Models 3
1.3 Sampling Protocol, Data Representation and Assumptions . . 3
1.3.1 Protocol 3
1.3.2 Data 4
1.3.3 Assumptions 5
1.4 Capture-Recapture Methods for Homogeneous Animals 6
1.4.1 Multinomial Model 6
1.4.2 Poisson Model 10
1.5 Capture-Recapture Methods with Heterogeneity among Animals 11
1.5.1 The Otis Suite of Models 11
1.5.2 Using Stratification 14
1.5.3 Individual-Based Heterogeneity 16
1.6 Multi-List Methods 24
1.6.1 Multi-List and Capture-Recapture: Key Differences . 25
1.6.2 Log-linear Modelling 26
1.6.3 Heterogeneity 28

1.7 Common Methodological Issues . . . 30
1.7.1 Bias Reduction in Small Samples . . . 30
1.7.2 Estimates of Precision . . . 31
1.7.3 Model Selection . . . 32
1.7.4 Model Uncertainty and Estimates of Precision . . . 34
1.7.5 Bayesian Methods . . . 35
1.8 Other Violations of Assumptions . . . 36
1.9 Outline of Thesis . . . 37
1.9.1 Chapter 2 . . . 37
1.9.2 Chapter 3 . . . 38
1.9.3 Chapter 4 . . . 38
2 Incomplete and Partially Stratified Lists . . . 39
2.1 Notation . . . 40
2.1.1 Parameters . . . 40
2.1.2 Statistic . . . 41
2.2 Model Development . . . 41
2.3 Examples . . . 47
2.3.1 Auckland Diabetes Prevalence Study . . . 47
2.3.2 Scottish Drug Use Prevalence . . . 50
2.3.3 Forest Fire Incidence . . . 53
2.4 Simulation Study . . . 57
2.4.1 Effects of List Interactions . . . 58
2.4.2 Small Sample Size Effects . . . 63
2.5 Discussion . . . 64
3 Population Estimation with Incomplete Lists . . . 66
3.1 Notation . . . 67
3.1.1 Parameters . . . 67
3.1.2 Statistics . . . 67
3.2 Model Development . . . 69
3.3 Example . . . 73
3.4 Simulation . . . 78

3.4.1 List Independence 78
3.4.2 List Dependence 82
3.5 Discussion 83
4 Estimation with Unmatchable List Members 85
4.1 Notation 87
4.1.1 Parameters 87
4.1.2 Statistics 87
4.2 Model Development 89
4.3 Example 95
4.4 Simulation Study 98
4.5 Discussion 100
5 Conclusions and Future Work 103
5.1 Summary 103
5.2 Future Work 106
5.3 An Apparent Duality 107
Bibliography 109

List of Tables

1.1 2^2 contingency table for two list example. 5

2.1 Partially stratified example; three lists, two strata. 41
2.2 Original data for Auckland Diabetes Study. Lists: General practioners records (G), Pharmacy records (P), Outpatient records (O) and Inpatient discharge records (D). 48
2.3 Auckland Diabetes Study lists. Lists: General practitioners records (G), Pharmacy records (P), Outpatient records (O) and Inpatient discharge records (D). 48
2.4 Simulated cell counts for Auckland Diabetes Prevalence Study data. . 49
2.5 Model fitting summary of Auckland Diabetes Study data 49
2.6 Original data for Aberdeen City opiate/benzodiazepine misuse. List 1: Combined data from the Substance Misuse Service and GP returns to the Scottish Drug Misuse Database. List 2: Couselling service. List 3: Needle/syringe exchange. List 4: Combined data from the Police and Social work department. 51
2.7 Summary of which lists were operating in which strata. Drug misuse in Aberdeen City, Scotland example. 52
2.8 Simulated data for Aberdeen City opiate/benzodiazepine misuse. . . . 52
2.9 Model fitting summary of drug misuse in Aberdeen City, Scotland . . 53
2.10 Observed Fires of Plot 12. 54
2.11 Cell counts for Plot 12. 55
2.12 Summary of model fitting process of Plot 12. 56

2.13 Multi-list setting for the simulation study. 58
2.14 Independent lists, 1,000 samples. Model was $\beta_1 = -1, \beta_2 = -0.8, \beta_3 = -0.5, \beta_4 = -0.3$ 59
2.15 Independent lists, stratum effect, 1,000 samples. Model was $\beta_1 = -1, \beta_2 = -0.8, \beta_3 = -0.5, \beta_4 = -0.3, \lambda = -0.5$ 60
2.16 Dependent lists, stratum effect, 1,000 samples. Model was $\beta_1 = -1, \beta_2 = -0.8, \beta_3 = -0.5, \beta_4 = -0.3, \lambda = -0.5, \beta_{13} = \beta_{23} = \beta_{34} = 0.2$ 61
2.17 Dependent lists, stratum effect, 1,000 samples. Model was $\beta_1 = -1, \beta_2 = -0.8, \beta_3 = -0.5, \beta_4 = -0.3, \lambda = -0.5, \beta_{13} = \beta_{23} = \beta_{34} = 0.2, \lambda\beta_{1,2} = \lambda\beta_{2,1} = 0.8$ 62
2.18 Simulation Results, $n = 300$. True value of $\delta = 0.1244$. No corrections to cell counts. A model with independent tree effects was fit. 64
2.19 Simulation Results, $n = 300$. True value of $\delta = 0.1244$. $\frac{2}{17}$ added to each cell before model fitting. A model with independent tree effects was fit. 65

3.1 Four list, two tag example. 68
3.2 Four list example of simulated Diabetes data. 74
3.3 Statistics for the 4-list Diabetes example 76
3.4 Results of estimating population size of the Auckland Diabetes Study. 76
3.5 Selected results of model fitting to Auckland Diabetes Study data. *Unable to fit model because the number of parameters exceeds the number of data points. 77
3.6 Population estimates, 1,000 simulations of 4 list, 2-tag example, equal list size. Theoretical population, $N = 60{,}000$. Model fit, $\mathbf{\Phi} = \{N, p_1, p_2, p_3, p_4\}$, which corresponds to model used to generate data. 79
3.7 Population estimates, 1,000 simulations of list size, 4 list, 2-tag example, unequal list size. Model fit, $\mathbf{\Phi} = \{N, p_1, p_2, p_3, p_4\}$, which corresponds to model used to generate data. 80

3.8 Population estimates, 1,000 simulations of list size, 3 list, 2-tag example, unequal list size. Model fit, $\mathbf{\Phi} = \{N, p_1, p_2, p_3, p_4\}$, which corresponds to model used to generate data. 81
3.9 Population estimates, 1,000 simulations of list size, 4 list, 2-tag example, list dependence between lists 1 and 4 and between lists 2 and 4. 82

4.1 Observed statistics and expected values for unmatchable list members under model $\phi = (N, p_1, p_2, \theta_{A_1} = \theta_{A_2}, \theta_{B_1} = \theta_{B_2})$. 92
4.2 Comparison of population estimates and estimated standard errors. $Y_{AB}^{11} = 166, Y_{AB}^{10} = 1405, Y_{AB}^{01} = 2170$ and $Y_{\cdot B}^{1\cdot} = 192$. 97
4.3 Results from two list, two tag simulation study of performance of estimating equations and approximately unbiased Petersen under several scenarios of tag loss. 1,000 replicates, $N = 40,000$ and model fit is $\phi = \{N, p_1, p_2, \theta_{A_1} = \theta_{A_2}, \theta_{B_1} = \theta_{B_2}\}$. 99
4.4 Results from three list, two tag simulation study of performance of estimating equations and log-linear models under several scenarios of tag loss when model $\phi = \{N, p_1, p_2, \theta_{A_1} = \theta_{A_2} = \theta_{A_3}, \theta_{B_1} = \theta_{B_2} = \theta_{B_3}\}$ .1,000 replicates and $N = 40,000$. 101

List of Figures

3.1 Graphical representation of Table 3.1. Four list, two tag example. . . 69
3.2 Graphical representation of lists of Auckland Diabetes Data. 75

Chapter 1

Introduction

1.1 Background

Capture-recapture, or multi-list methods, are used by investigators to estimate the unknown size of a target population whose size cannot be reasonably enumerated. This thesis presents methods to estimate population sizes when lists are only partially available or where there is incomplete information available regarding individuals on lists. These methods can assist with population estimation problems occurring in technological, ecological and biological sciences, as well as in epidemiological and public health settings.

In the simplest experiment, the two sample or two list problem, population size is estimated using the Petersen estimator. The method proceeds by sampling n_1 items from a population of unknown size, N, and applying a 'tag' to each. A second random sample of size n_2 is drawn from the population, hence the name capture-recapture. The number of items drawn in the second sample that appeared in the first sample (i.e. 'tagged') are enumerated, m_2. The Petersen estimator is simply the initial sample size 'inflated' by the fraction of marked items occurring in the second sample,

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

The Petersen estimator performs poorly when $E[m_2]$ is small. In fact, the estimator has infinite bias because there is a non-zero probability that $m_2 = 0$. Estimators

with better small sample size properties have been developed. Chapman (1951) suggested a correction factor to the Petersen estimator to reduce bias:

$$\widehat{N} = \frac{(n_1+1)(n_2+1)}{m_2+1} - 1,$$

which is approximately unbiased for $n_1 + n_2 < N$ and unbiased for $n_1 + n_2 \geq N$. Seber (1970) and Wittes (1972) proposed a widely-adopted approximately unbiased estimator of the variance of Chapman's estimator:

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

The Petersen estimator makes strong assumptions regarding the characteristics of the population which often are not realistic. They are: the population is closed, the samples are independent from one another, there is complete mixing of the population between sampling occasions, and individuals are equally likely to be captured on each occasion. If these assumptions are violated, population estimates may be severely biased.

This simple experiment can be generalized in a number of ways with differing objectives. The major division is between open models and closed models. Models that incorporate the dynamics of changing population due to births, deaths, immigration or emigration are known as 'open' models. Closed models assume that the population size does not vary over the study period. The closed population assumption is often reasonable provided that the study is completed in a short period of time.

As this thesis is concerned with closed populations, open models will not be reviewed further. For further discussion on this subject, Seber (1982, 1986 and 1992), Manly and McDonald (1996) and Schwarz and Seber (1999) provide comprehensive reviews of the capture-recapture literature focusing on animal abundance issues. Meanwhile, Buckland, Goudie and Borchers (2000) and Seber (2001) suggest future directions for capture-recapture research.

1.2 Closed Models

Due to the Petersen estimator's strong assumptions, more flexible and realistic methods have been developed based on more than two sampling occasions (multiple recaptures).

The closed population field has self-divided into two distinct sub-fields. In the first, referred to as capture-recapture studies, and most commonly applied in wildlife settings, there are a series of samples of the population taken over time. At each sample, captured individuals are tagged and returned to the population. The key feature is the time ordering of the samples.

In the second sub-field, referred to as multi-list studies, and most commonly applied to human populations, there are a series of lists compiled. Here, 'capture' and 'recapture' refer to being a member of a list. For example, consider estimating the number of individuals with diabetes based on pharmacy records and hospitalization records. Two lists are compiled, one based on pharmacy visits (list of length n_1), the other based on hospitalizations (list of length n_2). Each subject can appear only once on each list. The number of people on both lists, m_2, is assessed by matching names, the 'tag', across lists.

The major difference between capture-recapture and multi-list methods is that there is no time ordering associated with composing lists. 'Time effects' in capture-recapture models are associated with 'list effects' in multi-list models.

Despite the differences between capture-recapture and multi-list methods, many methods developed in either situation are applicable to both.

1.3 Sampling Protocol, Data Representation and Assumptions

1.3.1 Protocol

Closed capture-recapture experiments first proceed by capturing a portion of the target population, n_1. In a fisheries example described in Schwarz and Arnason (1996),

salmon are captured with electrofishing methods. At the time of first capture, members of the population are 'tagged' and released back to the general population.

On each of $k = 1, \ldots, K$ sampling occasions, individuals are sampled from the population. At each sample, individuals are either captured with a tag or without a tag. If a tag is not present (not previously captured), a tag is affixed and the individual is released back into the population. If a tag is present (previously captured), the 'tag' information is recorded and the individual is released. This process is repeated for each of K sampling occasions.

In closed population multi-list studies, individuals are not physically tagged, but rather typically come with unique identifiers (*e.g.* health insurance numbers) that serve as "tags". The identifying tags are used to match individuals across K lists, which act as sampling occasions.

1.3.2 Data

In both capture-recapture and multi-list studies, a "capture-history" vector $\omega = \{\omega_1, \ldots, \omega_K\}$ can be created for each individual. The indicator variables $\omega_1, \ldots, \omega_K$ identify whether the subject was "captured" on each of the K lists or sampling occasions with

$$\omega_k = \begin{cases} 1 & \text{list } k \text{ recorded the individual;} \\ 0 & \text{list } k \text{ did not record the individual.} \end{cases}$$

The history vector $\omega = \{0, \ldots, 0\}$ identifies animals or individuals not seen at any time or on any list. A 2^K multi-way contingency table can be created by summing over individuals with like capture histories, such that n_ω is the number of individuals with history ω. For example, in a two-list example shown in Table 1.1, n_{01} is the number of individuals captured on list 2 only, n_{11} is the number of individuals captured on both lists, while n_{00} is the number of individuals unobserved on either list.

Define $L = 2^K$, the number of capture histories. Then, $L - 1 = 2^K - 1$ is the number of observable capture histories *i.e.*, excluding $\omega = \{0, \ldots, 0\}$.

		List 2		
		Present	Absent	Total
List 1	Present	n_{11}	n_{10}	$n_{1.}$
	Absent	n_{01}	n_{00}	
	Total	$n_{.1}$		

Table 1.1: 2^2 contingency table for two list example.

1.3.3 Assumptions

As the International Working Group for Disease Monitoring and Forecasting (IWGDMF) (1995a) note, capture-recapture analyses for closed models typically assume the following:

- The population is closed during the study;
- Each individual has the same probability of being included in a sample or on a list (equal catchability);
- Capturing an individual does not affect its probability of subsequent capture;
- The simpler models often assume that samples are independent but this assumption is relaxed somewhat in multi-list studies;
- Tags are not lost between sampling occasions;
- Tags are read correctly;
- There are no 'matching' errors across lists;
- Tags are not missed when examining individuals.

The following section develops models under these assumptions. Violations of these assumptions and their impact on population estimates are discussed in a later section.

1.4 Capture-Recapture Methods for Homogeneous Animals

1.4.1 Multinomial Model

Two Sampling Occasions

Borrowing from the notation of Borchers, Buckland and Zucchini (2002), consider the sequence of two captures. The probability of capture in the first sample is

$$P_{1u} = \left(\begin{array}{c} U_1 \\ u_1 \end{array} \right) p_1^{u_1} (1-p_1)^{U_1 - u_1},$$

where $U_1 = N$ is the population size, $u_1 = n_1$ is the number caught in the first sample and p_1 is the probability of capture in the first sample.

The probability of capture in the second sampling occasion is partitioned between those previously captured and those not previously captured.

First, consider those available to be re-captured in the second sample, M_2 (equal to the first sample size, n_1). Let m_2 be the number of individuals captured in the first sample observed in the second sample. The probability of observing m_2 marked individuals in the second sample is

$$P_{2m} = \left(\begin{array}{c} M_2 \\ m_2 \end{array} \right) p_2^{m_2} (1-p_2)^{M_2 - m_2}.$$

Second, consider the unmarked individuals observed in the second sample. Let U_2 be the number of unmarked individuals available for capture in the second sample (equal to $N - M_2$). The number of unmarked in the second sample is u_2. The probability of capturing U_2 unmarked individuals in the second sample is

$$P_{2u} = \left(\begin{array}{c} U_2 \\ u_2 \end{array} \right) p_2^{u_2} (1-p_2)^{U_2 - u_2}.$$

Considered together, the probability of observing (m_2, u_2), marked and unmarked individuals, in the second sample is the product of two binomial likelihoods, or

$$\left(\begin{array}{c} M_2 \\ m_2 \end{array} \right) p_2^{m_2} (1-p_2)^{M_2 - m_2} \times \left(\begin{array}{c} U_2 \\ u_2 \end{array} \right) p_2^{u_2} (1-p_2)^{U_2 - u_2}.$$

The full likelihood for the two sampling occasions can then be written as the product of its marked and unmarked probabilities, or

$$\begin{aligned} L_2 &= \prod_{s=1}^{2} P_{sm} \times P_{su} \\ &= \prod_{s=1}^{2} \left(\begin{array}{c} M_s \\ m_s \end{array} \right) p_s^{m_s} (1-p_s)^{M_s - m_s} \left(\begin{array}{c} U_s \\ u_s \end{array} \right) p_s^{u_s} (1-p_s)^{U_s - u_s}, \end{aligned}$$

where $P_{1m} \equiv 1$ to facilitate generalization of this notational approach.

The likelihood L_2 can also be written as a multinomial function (Darroch, 1958). In this case, the likelihood is written

$$L(N, \mathbf{p}) = \frac{N!}{n_{11}! n_{10}! n_{01}! n_{00}!} p_{11}^{n_{11}} p_{10}^{n_{10}} p_{01}^{n_{01}} p_{00}^{n_{00}},$$

where $n_{00} = N - n_{11} - n_{10} - n_{01}$ and $p_{00} = 1 - p_{11} - p_{10} - p_{01}$.

For N to be estimable, it is necessary to constrain the number of parameters. Sanathanan (1972) suggests writing p_ω (capture probability for capture history ω) as some known function of parameters, $p_\omega(\beta)$, where $\beta = (\beta_1, \ldots, \beta_k)$ and $k \leq 2^K - 1$, the number of observable capture histories.

Written in this manner, the likelihood is

$$L(N, \beta) = \frac{N!}{n_{11}! n_{10}! n_{01}! n_{00}!} p_{11}(\beta)^{n_{11}} p_{10}(\beta)^{n_{10}} p_{01}(\beta)^{n_{01}} p_{00}(\beta)^{n_{00}}.$$

For example, in the simple case where capture probabilities are independent across samples, $p_{11}(\beta) = \beta_1\beta_2$, $p_{10}(\beta) = \beta_1(1-\beta_2)$, $p_{01}(\beta) = (1-\beta_1)\beta_2$ and $p_{00}(\beta) = 1 - \beta_1 - \beta_2 + \beta_1\beta_2$.

The number of parameters (N, β_1, β_2) is constrained to less than or equal to the number of observed cells in the multi-way contingency table.

Sanathanan (1972) provides two estimates for N, an unconditional and a conditional estimate. Although details regarding each are summarized in Sanathanan (1972) and Feinberg's (1972) review of her paper, we summarize the main points of each method.

Unconditional likelihood The unconditional likelihood, L_U, estimates N, p_1 and p_2 simultaneously.

Conditional likelihood The conditional likelihood approach factors the likelihood by capture status. The first factor is the probability of being seen anywhere or at any time during the study. The second factor is the probability of capture history ω given the probability of being seen anywhere or at any time during the study. Estimates of all parameters, except N, are based on maximizing the second factor which does not have N in the likelihood. The estimates of the parameters are then 'plugged' into the first factor to estimate N.

Sananthanan (1972) found that under certain regularity conditions both the unconditional MLE and conditional MLE have the same asymptotic normal distribution.

Maximizing the conditional likelihood, we find that

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

and the asymptotic variance is $\widehat{Var}(\widehat{N}) = \frac{n_1 n_2^2 (n_1 - m_2)}{m_2^3}$. Normal-based asymptotic confidence regions can be obtained for $\widehat{N}$ and are discussed in a later section.

Confidence intervals for $\widehat{N}$ can be constructed in a similar manner for Chapman's modified estimator. Borchers, Buckland and Zucchini (2002) show an example of using the likelihood function to obtain profile likelihood estimates for N.

Parametric or non-parametric bootstrap confidence intervals can also be obtained. In the parametric case, resampling occurs from the multinomial distribution based on the observed counts and the estimated parameters. In the non-parametric case, resampling occurs with replacement from the capture histories, ω, with equal probability for each individual history. Buckland and Garthwaite (1991) note that bootstrap methods "provide a robust and general approach to estimating variance and confidence intervals in mark-recapture" that are superior to asymptotic likelihood-based inference.

Multiple Sampling Occasions

The likelihood for two sampling occasions, L_2, can readily be generalized to K occasions. The likelihood for K occasions is

$$\begin{aligned} L_K &= \prod_{k=1}^{K} P_{km} \times P_{ku} \\ &= \prod_{k=1}^{K} \left(\begin{array}{c} M_k \\ m_k \end{array} \right) p_k^{m_k} (1-p_k)^{M_k - m_k} \left(\begin{array}{c} U_k \\ u_k \end{array} \right) p_k^{u_k} (1-p_k)^{U_k - u_k}. \end{aligned}$$

Then, the generalization of the multinomial theorem can be applied, and the likelihood is rewritten as

$$L_K = \frac{N!}{\prod_\omega n_\omega!} \prod_\omega p_\omega^{n_\omega},$$

the product over all observable capture histories.

Unconditional and Conditional Maximization Likelihood Estimation

As Sanathanan (1972) writes, there are two types of maximum likelihood equations, the unconditional MLE and the conditional MLE. The unconditional MLE is the usual one, in the sense that all parameters are estimated simultaneously. The unconditional likelihood function is

$$L(N, \beta) = \frac{N!}{\prod_\omega n_\omega!} \prod_\omega (p_\omega(\beta))^{n_\omega}.$$

Define the following

$$\begin{aligned} \prod_\omega p_\omega^{n_\omega} &= \text{product over all histories, including } \omega_{0,\ldots,0} \\ \prod_{\omega:\omega\neq 0} p_\omega^{n_\omega} &= \text{product over all observable histories, excluding } \omega_{0,\ldots,0} \\ N &= \sum_\omega n_\omega, \quad (n_{0,\ldots,0} \text{ is included}) \\ n &= \sum_{\omega:\omega\neq 0} n_\omega, \quad \text{(observed sample size)}. \end{aligned}$$

Then, the conditional likelihood can be written as

$$\begin{aligned} L_1(N,\beta) &= \frac{N!}{(N-\sum_{\omega:\omega\neq 0} n_\omega)!(\sum_{\omega:\omega\neq 0} n_\omega)!}\left[\sum_{\omega:\omega\neq 0} p_\omega(\beta)\right]^{\sum_{\omega:\omega\neq 0} n_\omega} \\ &\quad \left[1-\sum_{\omega:\omega\neq 0} p_\omega(\beta)\right]^{N-\sum_{\omega:\omega\neq 0} n_\omega} \\ L_2(\beta) &= \frac{(\sum_{\omega:\omega\neq 0} n_\omega)!}{\prod_{\omega:\omega\neq 0} n_\omega!}\frac{\prod_{\omega:\omega\neq 0}(p_\omega(\beta))^{n_\omega}}{(1-\sum_{\omega:\omega\neq 0} p_\omega(\beta))^{\sum_{\omega:\omega\neq 0} n_\omega}}. \end{aligned}$$

The conditional maximization proceeds by maximizing $L_2(\beta)$, then using the parameter estimates to maximize $L_1(N,\beta)$. $L_1(N,\beta)$ yields $\widehat{N} = \frac{\sum_\omega n_\omega}{1-\sum_\omega \widehat{p}_\omega}$ as the conditional estimator for N, where the $\widehat{p}_\omega$'s are obtained by maximizing $L_2(\beta)$. Sanathanan (1972) has shown that the two are asymptotically equivalent. Generally, the conditional likelihood is preferred when there are several capture occasions due to its ease of computation.

1.4.2 Poisson Model

The Poisson capture-recapture model was proposed independently by Cormack (1979) and Jolly (1979). They suggested modelling the multi-way contingency table cell counts as independent Poisson variables. As Huakau (2001) noted, the likelihood can be written as

$$L(N,\mathbf{p}) = \prod_{\omega:\omega\neq 0} \frac{(Np_\omega)^{n_\omega} exp(-Np_\omega)}{n_\omega!},$$

where n_ω is the same as in the multinomial model and p_ω is the capture probability for capture history ω.

For N to be estimable, the number of parameters cannot exceed the number of observed cells in the multi-way contingency table. Following Sanathanan's (1972) parameterization, the p_ω are again modelled as $p_\omega(\beta)$. The likelihood for the Poisson capture-recapture model can be written

$$L(N,\beta) = \prod_{\omega:\omega\neq 0} \frac{(Np_\omega(\beta))^{n_\omega} exp(-Np_\omega(\beta))}{n_\omega!}.$$

The estimates derived from the multinomial model and the Poisson model are identical.

Sandland and Cormack (1984) found that the asymptotic variance of N under the Poisson model is the same as that using the multinomial model plus an additional term N. Cormack and Jupp (1991) showed that the asymptotic variances for β are the same under both approaches.

The multinomial and Poisson models can be used interchangeably provided that N is large (Poisson and multinomial variances can be shown to be equivalent as $N \to \infty$).

1.5 Capture-Recapture Methods with Heterogeneity among Animals

In the methodology discussed in the earlier sections, we assumed that all animals were equally catchable at every sampling time. Unfortunately, these assumptions do not always hold. As Chao (2001) shows, it is important to incorporate sources of heterogeneity into estimates of capture probability, as population estimates may otherwise be severely biased.

1.5.1 The Otis Suite of Models

Pollock (1974) considered closed population models that relaxed the assumption of equal probability of capture. These models were generalized by Otis et al. (1978) into 8 model types to accommodate three types of heterogeneity in capture probability. Using the notation of Borchers, Buckland and Zucchini (2002), the models are denoted M_m, where m refers to the source of capture probability heterogeneity. The models are:

M_t Capture probabilities vary by sampling occasion only. All individuals in the population are equally catchable at any occasion.

M_b Capture probabilities vary by response to capture (behavioral responses to capture.)

M_h Individuals in the population have heterogeneous capture probabilities that remain constant over the study period.

Note that there are two types of heterogeneity, the first is due to external sources, such as sampling occasion effect (time) and behavioral response to capture. The second type of heterogeneity is intrinsic to animals. Models that incorporate intrinsic heterogeneity will be discussed in later sections.

The full complement of models is: $M_0, M_t, M_b, M_h, M_{tb}, M_{th}, M_{bh}$ and M_{tbh}, where the models M_{tb}, M_{th}, M_{bh} and M_{tbh} combine the multiple sources of heterogeneity. Model M_0 represents homogeneous capture probabilities, for which all individuals in the population are equally catchable.

As Lee and Chao (1994) note, the capture probability for the i^{th} individual and j^{th} sampling occasion for models M_0, M_t, M_b, M_{tb} can be written:

- Model M_0: $p_{ij} = p$;
- Model M_t: $p_{ij} = p\alpha_j$, where α is the unknown time effect of the j^{th} sample;
- Model M_b: $p_{ij} = p$ for the first capture in any sample, $p_{ij} = \widetilde{p}$ for any recapture;
- Model M_{tb}: $p_{ij} = pe_j$ for the first capture in any sample, $p_{ij} = be_j^*$ for any recapture, where e_j is the known relative time effect of the j^{th} sampling occasion on capture probability (variable catch-effort model).

Model M_0

The likelihood model for M_0 is written

$$L(N,p) = \frac{N!}{\prod_{\omega:\omega\neq 0} n_\omega!(N-\sum_{\omega:\omega\neq 0} n_\omega)!}\left\{\prod_{\omega:\omega\neq 0}\left[\prod_{k=1}^{K} p^{\omega_k}(1-p)^{1-\omega_k}\right]^{n_\omega}\right\} \times\left[\prod_{k=1}^{K}(1-p)\right]^{N-\sum_{\omega:\omega\neq 0} n_\omega},$$

where p is constant at all sampling occasions for each of the L distinct capture histories, ω.

Model M_t

The likelihood model for M_t is written

$$L(N,p,\alpha) = \frac{N!}{\prod_{\omega:\omega\neq 0} n_\omega!(N-\sum_{\omega:\omega\neq 0} n_\omega)!}\left\{\prod_{\omega:\omega\neq 0}\left[\prod_{k=1}^{K}(p\alpha_k)^{\omega_k}(1-p\alpha_k)^{1-\omega_k}\right]^{n_\omega}\right\} \times\left[\prod_{k=1}^{K}(1-p\alpha_k)\right]^{N-\sum_{\omega:\omega\neq 0} n_\omega},$$

where p is constant across all samples, and α_k is the effect due to capture on occasions ω_k. Lloyd (1994) writes that Darroch (1958) found that including time variation does not negatively affect variance estimation of $\widehat{N}$, thus model M_0 should never be preferred over M_t.

Model M_b

The likelihood model for the behavioral effects model, M_b, is

$$L(N,p,\widetilde{p}) = \frac{N!}{\prod_{\omega:\omega\neq 0} n_\omega!(N-\sum_{\omega:\omega\neq 0} n_\omega)!} \times\prod_{\omega:\omega\neq 0}\left[\prod_{k=1}^{k'}p^{\omega_k}(1-p)^{1-\omega_k}\prod_{k=k'+1}^{K}\widetilde{p}^{\omega_k}(1-\widetilde{p})^{1-\omega_k}\right]^{n_\omega} \times\left[\prod_{k=1}^{K}(1-p)\right]^{N-\sum_{\omega:\omega\neq 0} n_\omega},$$

where the capture probabilities have been partitioned into two sections, prior to first capture and after first capture. k' denotes the period of first capture, $k' \in \{1,\ldots,K\}$. Also, p is the capture probability for any first capture and $\widetilde{p}$ is the capture probability for any subsequent capture. In abundance studies, $\widetilde{p}$ is often a nuisance parameter. Note that M_b is equivalent to the removal model of Zippen (1956).

Model M_{tb}

The likelihood model for time and behavioral effects model, M_{tb}, is

$$\begin{aligned} L(N, p, \widetilde{p}, \alpha, \widetilde{\alpha}) &= \frac{N!}{\prod_{\omega:\omega\neq 0} n_\omega!(N - \sum_{\omega:\omega\neq 0} n_\omega)!} \\ &\quad \times \prod_{\omega:\omega\neq 0} \left[\prod_{k=1}^{k'} (p\alpha_k)^{\omega_k}(1 - p\alpha_k)^{1-\omega_k} \prod_{k=k'+1}^{K} (\widetilde{p}\widetilde{\alpha}_k)^{\omega_k}(1 - \widetilde{p}\widetilde{\alpha}_k)^{1-\omega_k} \right]^{n_\omega} \\ &\quad \times \left[\prod_{k=1}^{K} (1 - p) \right]^{N - \sum_{\omega:\omega\neq 0} n_\omega}, \end{aligned}$$

where the probabilities are again partitioned according to the first capture in time k'. Again, p is the capture probability for any first capture, α_ω is the effect due to capture in occasion ω_k, $\widetilde{p}$ is the capture probability for any subsequent capture and $\widetilde{\alpha}$ is the effect due to capture on occasion ω_k having already been previously captured.

Estimation

Although closed form solutions exist for these models, numerical methods are widely available as to make their application unnecessary.

Chao, Chu and Hsu (2000) have also developed a maximum quasi-likelihood estimator (MQLE) for the model M_{tb} and its sub-models M_t, M_b. These models have estimated variances that are asymptotically normal, consistent and are the same as the MLEs. In simulations studies, Chao, Chu and Hsu (2000) found that the results of the unconditional likelihood, the conditional likelihood and the quasi-likelihood are comparable.

1.5.2 Using Stratification

As Plante, Rivest and Tremblay (1998) note, heterogeneous capture probabilities among animals result in estimates of population size being negatively biased. Stratification is one strategy used to minimize the impact of intrinsic heterogeneity. Heterogeneity can be due to the characteristics of the population (say, large fish are easier to catch than small fish), or it may be due to variability in the conditions under which

animals are captured (*e.g.*, catch effort that varies or capture probabilities that are related to abundance). Stratification can be classified into two broad categories; according to whether individuals are grouped by unchanging characteristics determined at release (such as sex), or characteristics that change over time, such as movement among geographical strata.

Alho (1990) suggests dividing the population into subgroups based on homogeneous probabilities of capture. Stratification is based on covariate information. Alho (1993) extended the method to incorporate continuous covariates when studying the under-count in the U.S. census. The method assumes that an individual's capture probability is related to their covariates through a logistic model, written

$$\text{logit}(p_{ij}) = \mathbf{X}_{ji}^T \boldsymbol{\beta},$$

where p_{ij} is the capture probability of individual i in period j, $\mathbf{X}$ is the matrix of covariates and $\boldsymbol{\beta}$ is the vector of parameters. Population size is then estimated with a modified Horvitz-Thompson estimator

$$\widehat{N} = \sum_{i=1}^{n} \frac{1}{\widehat{p}_i},$$

where $\widehat{p}_i$ is the probability of animal i being observed. As Alho (1993) notes, the logistic modelling approach is capable of incorporating observable heterogeneity only.

Darroch (1961) was the first to consider stratification in the two-sample Petersen estimator. Little new was developed (except Seber, 1982) until the generalization developed by Plante, Rivest and Tremblay (1998). The population is stratified at the time of tagging and at the time of recovery.

Plante, Rivest and Tremblay (1998) temporally stratify to reduce heterogeneity caused by catching conditions (such as catch-effort or river flow). Define t_i to be the probability of tagging in stratum i, r_j be the the probability of recovery in stratum j and N_{ij} to be the number of individuals in the population in tagging stratum i and recovery stratum j. The total population size $N = \sum_{ij} N_{ij}$. Recovery probabilities (r_j) are modelled as

$$\log \frac{1 - r_j}{r_j} = \sum_{k=1}^{p} x_{jk} \beta_k,$$

where $X_j = (x_{j1}, \ldots, x_{jp})^T$ are known covariates and $\boldsymbol{\beta}$ is a vector of parameters. Define $\mu_{ij} = t_i N_{ij} r_j$ to be the expected number of animals tagged in stratum i and recovered in stratum j, and $\psi_j = r_j \sum_{i=1} N_{ij}(1-t_i)$ is the expected number of animals tagged in all strata and uncaptured in stratum j. Define m_{ij} as the number of animals tagged in tagging stratum i and captured in recovery stratum j, $a_i = n_i - \sum_{l=1}^{t} m_{il}$, the number of captured animals tagged in tagging stratum i but not seen again, and $b_j = n_i - \sum_{l=1}^{s} m_{li}$, the number of uncaptured animals in recovery stratum j.

Assuming the observed counts follow independent Poisson distributions, a conditional likelihood is proportional to $L(N, \beta) = L_1(N, p(\beta))L_P(\beta)$, where

$$\begin{aligned} L_1(N, p(\beta)) &= \frac{N!}{n!(N-n)!} p(\beta)^n \left[1 - p(\beta)\right]^{N-n} \\ L_P(\beta) &\propto \prod_{ij} \mu^{m_{ij}} e^{-\mu_{ij}} \prod_{j=1}^{t} \psi_j^{b_j} e^{-\psi_j} \prod_i \left[\left(\sum_{l=1}^{t} e^{X_l^T \beta} \mu_{il} \right)^{a_i} exp\left(- \sum_{l=1}^{t} e^{X_l^T \beta} \mu_{il} \right) \right]. \end{aligned}$$

Two methods are suggested for confidence regions for $\widehat{N}$, a conditional estimator and a profile likelihood. Plante, Rivest and Tremblay (1998) use temporal stratification to reduce the bias of estimates and increase the precision in studies of smolt populations.

Little work has been done on extending the methodology for stratification to reduce heterogeneity in closed populations except Schwarz and Ganter (1995), who develop models for K sampling occasions when estimating the number of geese moving among staging areas.

1.5.3 Individual-Based Heterogeneity

A significant problem with models that deal with individual-level intrinsic heterogeneity (any model with h in the subscript) is that the likelihood cannot be evaluated without catching the whole population, so full likelihood-based inference is not possible.

Ignoring heterogeneity of individual-level capture probabilities can lead to significantly negatively biased estimates of population size. For example, consider a

population in which a specific sub-population is uncatchable. Population size estimates will not be capable of incorporating this sub-population leading to negatively biased estimates.

To incorporate individual-level heterogeneity, different approaches have been tried. The classification of Otis et al. (1978) considered the following models.

- Model M_h: $p_{ij} = p_i$;
- Model M_{bh}: $p_{ij} = p_i$, for the first capture in any sample, $p_{ij} = b_i$ for any recapture;
- Model M_{th}: $p_{ij} = p_i \alpha_j$;
- Model M_{tbh}: $p_{ij} = p_i e_j$ for the first capture in any sample, $p_{ij} = b_i e_j^*$ for any recapture, where e_j is the known relative time effect of the j^{th} sampling occasion on capture probability.

Burnham (1972, PhD Thesis) modelled individual-based capture heterogeneity with a beta distribution. He found that the likelihood function for N was very flat and had correspondingly poor asymptotic properties. Since this attempt, other methods have been proposed to model heterogeneous capture probabilities.

Jackknife Estimators

One of the first methods proposed to incorporate heterogeneous capture probabilities was a jackknife estimator (Burnham and Overton, 1978 and 1979). The method assumes that capture probabilities are constant for each individual across capture occasions, but vary between individuals.

The approach assumes that observed capture probabilities are samples from some distribution in the interval (0,1). Let f_i equal the number of individuals captured exactly i times (the sum of the indices of $\omega = i$), $i = 1, \ldots, K$. Then $N = \sum_i^K f_i + f_0$. The jackknife estimator is written

$$N_{Jk} = \sum_{i=1}^{K} a_{ik} f_i,$$

where subscript J indicates the jackknife estimator, k represents the order of the estimator and a_{ik} is some known constant. The $k = 1, \ldots, 5$ order jackknife estimators, N_{Jk} are given in Burnham and Overton (1979). Later, Pollock and Otto (1983) compared the properties of several jackknife estimators under different population assumptions.

Originally, developed as a method to reduce bias, the jackknife was an early attempt to incorporate heterogeneous capture probabilities. Lee and Chao (1994) note that the jackknife estimator severely overestimates population size if the number of captured individuals is relatively large. Currently, the jackknife estimator is not as popular as competing methods.

Moment-based Estimators

As an improvement to the jackknife estimator, Chao (1987) proposed a moment-based estimator for N. Let f_j equal the number of individuals captured exactly j times, and $N = \sum_j^K f_j + f_0$. From Burnham and Overton (1978), Chao writes the unconditional distribution of $f_0, f_1, \ldots, f_K$ as

$$p(f_0, f_1, \ldots, f_K) = \left(\begin{array}{c} N \\ f_0, f_1, \cdots, f_K \end{array} \right) \prod_{j=1}^{K} \left[\int_0^1 \left(\begin{array}{c} K \\ j \end{array} \right) p^j (1-p)^{K-j} dF(p) \right]^{f_j},$$

where $F(p)$ represent the distribution of capture probabilities.

It follows that

$$E(f_j) = \int_0^1 \left(\begin{array}{c} K \\ j \end{array} \right) p^j (1-p)^{K-j} dF(p)$$

for $j = 1, \ldots, K$. Chao (1987) suggests that for K large and p small, a reasonable approximation is

$$E(f_j) \approx \int_0^1 \left[\frac{(Kp)^j e^{-Kp}}{j!}, \right] dF(p) \quad j = 1, \ldots, K.$$

Chao (1987) estimates

$$\begin{aligned} E(f_0) &\approx N \int_0^1 e^{-Kp} dF(p) \\ &\approx [E(f_1)] \int_0^K u^{-1} dG(u), \end{aligned}$$

while the rth moment of G is given by

$$\begin{aligned}\mu_r &= \int_0^K u^r dG(u) \\ &\approx (r+1)! \frac{E[f_{r+1}]}{E[f_1]}.\end{aligned}$$

Replacing $E[f_j]$ by f_j, Chao (1987) finds the moment estimator to be

$$m_r = (r+1)! f_{r+1}/f_1.$$

Using this result and Jensen's inequality, Chao finds a lower bound on the population estimator to be

$$\widehat{N} = \sum_j^K f_j + f_1^2/(2f_2).$$

Chao (1987) also noted a 'correction factor' for $\widehat{N}$ based on the first two moments.

An advantage is that this estimator is simple to use. However, model selection is difficult due to the absence of the likelihood-based goodness-of-fit statistics. Also, more recent advances, such as model averaging cannot be easily applied.

Horvitz-Thompson Estimator

Huggins (1989, 1991) showed that all 8 models (in the Otis suite of models) could be modelled using a conditional likelihood approach. By conditioning on observed individuals, a logit function using covariate information (such as age or weight) models capture probabilities.

Let p_{ij} represent individual-level capture probabilities, (capture at occasion $j = 1, \ldots, K$ for individual $i = 1, \ldots, N$). Then, the conditional likelihood (Huggins, 1991) is proportional to

$$\prod_{i=1}^n \prod_{j=1}^K \frac{p_{ij}^{x_{ij}}(1-p_{ij})^{(1-x_{ij})}}{1-\prod_{j=1}^K (1-p_{ij}^*)},$$

where $x_{ij} = 1$ if individual i is captured on occasion j, 0 otherwise. To represent past capture history, we say that $z_{ij} = 1$ if individual i has been captured before occasion j, 0 otherwise. Then, p_{ij}^* is p_{ij} evaluated at $z_{ij} = 0$.

In an example of model M_h, the logit link is related to the covariate information, written

$$\text{logit}(p_{ij}) = \log(\frac{p_{ij}}{1-p_{ij}}) = \beta_0 + \beta_{age} + \beta_{weight} + \cdots ,$$

Population size is then estimated using a modified Horvitz-Thompson estimator (Cochran, 1977). First, the probability that an individual i is captured at least once is estimated,

$$p_i(\beta) = 1 - \prod_{j=1}^{K}(1 - p_{ij}^*).$$

Then, population size is estimated using a modified Horvitz-Thompson estimator which assumes that $p_i(\beta)$ are known exactly. The estimator is written

$$\widehat{N} = \sum_{i}^{n} \frac{1}{\widehat{p_i}(\beta)}.$$

The drawback with this approach is that it assumes that captured and uncaptured animals have the same relationship between catchability and covariates. Advantages of this approach are that model selection can be AIC-based, and profile likelihoods can be used to generate confidence intervals.

Mixture models

Norris and Pollock (1996) and Pledger (2000) have used mixture models to incorporate individual-level capture probability heterogeneity. In the simplest case, the population is a mixture of animals from 2 groups. Some proportion of the population, α, is easy to capture, with probability of capture p_e, while the remainder, $1-\alpha$, is the proportion that is difficult to capture, with probability of capture p_h. In this case, p, the average probability of capture is modelled as $p = \alpha p_e + (1-\alpha)p_h$.

More formally, mixture models assume that $p_1, p_2, \ldots, p_N$ originate from a common distribution, F. Norris and Pollock (1996) start by assuming a multinomial-based likelihood

$$L_(N, F) = \frac{N!}{\prod_{\omega:\omega\neq 0} n_\omega! (N - \sum_{\omega:\omega\neq 0} n_\omega)!} \left[\prod_{\omega:\omega\neq 0} (P_{\omega,F})^{n_\omega} \right] (P_{\omega=0,F})^{N-\sum_\omega n_\omega},$$

where $P_{\sum \omega_i=0,F}$ is the probability of not being observed.

For model M_h, Norris and Pollock (1996) model the distribution of individual capture probabilities as

$$P_{\omega=0,F} = \int_0^1 p^{\sum \omega_i}(1-p)^{K-\sum \omega_i} dF(p).$$

Norris and Pollock (1996) show that the likelihood for model M_h is proportional to

$$T_1 = \left(\begin{array}{c} N \\ \sum_\omega n_\omega \end{array} \right) \left[\prod_{j=1}^{N} \int_0^1 \left(\begin{array}{c} K \\ i \end{array} \right) p^i (1-p)^{K-i} dF(p) \right],$$

where i is the number of times individual j is observed. Norris and Pollock (1996) also show the likelihood for the model M_{bh}, which uses only the time until first capture for each individual.

The biggest advantage with this approach is that all 8 models previously identified fit within the likelihood framework. Profile likelihoods for parameters can be obtained, and AIC-based model selection and model averaging can be applied.

Using a mixture of two groups has been shown to represent the effects of heterogeneity well, the approach does not say that there are exactly two groups but rather that such a model is a good representation. The biggest disadvantage is that in practice, it is difficult to partition the population into more than two groups due to lack of sample points (two populations require at least two sampling occasions).

Sample Coverage Methods

The sample coverage approach models the degree of overlap between samples. Chao and Tsay (1998) write "the estimated sample coverage for the general dependent models is the average (over all available samples) of the fraction of animals captured more than once." Chao and Tsay (1998) and Tsay and Chao (2001) note that the methods work best when there are at least three lists and a substantial degree of overlap between lists.

The sample coverage C is defined as the proportion of the total individual effects

that are associated with the captured individuals. This is written

$$C = \frac{\sum_{i=1}^{N} p_i \times 1[i^{th} \text{ individual is captured}]}{\sum_{i=1}^{N} p_i}.$$

If all p_i's are equal, as they are in models M_0, then C is the proportion of distinct individuals observed, or $C = \frac{D}{N}$. In the case of these models, $\widehat{N} = \frac{D}{\widehat{C}}$, where $\widehat{C}$ is the estimate of C.

In the case where capture probabilities vary by individual, the case is more complicated. Chao and Lee (1992a) and Chao, Lee and Jeng (1992b) propose the following two estimators for model M_h,

$$\begin{aligned} \widehat{N} &= \frac{D}{\widehat{C}} + \frac{f_1}{\widehat{C}}\widehat{\gamma}^2, \\ \overline{N} &= \frac{D}{\overline{C}} + \frac{f_1}{\overline{C}}\overline{\gamma}^2, \end{aligned}$$

where f_1 is the number of individuals captured once and $\widehat{C}, \overline{C}$ are estimators of $E[C]$, defined as

$$\begin{aligned} \widehat{C} &= 1 - \frac{f_1}{\sum_{k=1}^{K} k f_k} \\ \overline{C} &= 1 - \frac{f_1 - \frac{2f_2}{K-1}}{\sum_{k=1}^{K} k f_k}. \end{aligned}$$

The coefficient of variation, $\widehat{\gamma}^2$, is estimated by

$$\widehat{\gamma}^2 = max\left\{ \frac{\widehat{N}_0 \sum_k k(k-1) f_k}{2 \sum \sum_{j<k} n_j n_k} - 1, 0 \right\},$$

where $\widehat{N}_0 = D/\widehat{C}$. $\overline{\gamma}^2$ is obtained by replacing $\widehat{N}_0$ with $\overline{N}_0 = D/\overline{C}$. There are similar derivations for all 8 models in Lee and Chao (1994).

Simulation studies (Lee and Chao, 1994) showed that these estimators do not outperform the jackknife or maximum likelihood estimators under moderate heterogeneous capture probabilities. However, when mean capture probabilities are small, the bias is less than that of the jackknife's estimator.

Although conceptually easy to understand, this method does not lend itself to likelihood-based model-selection tools. Automated software (CARE, for CApture-REcapture) for sample coverage methods for all model types is available from the author.

Martingale Methods

Yip (1991) uses a method of moments for martingales to derive an estimate of population size with some success. Let m_j be the number of marked individuals in the j^{th} sample, u_j be the number of unmarked individuals in the j^{th} sample and let $M_j = \sum_{j=1}^{K} u_j$, the total number of unmarked individuals in the first $j-1$ sampling occasions.

Using the notation of Yip (1991), let $\mathcal{F}_j$ represent the information generated by the process up to sampling occasion j. Yip (1991) starts with

$$\frac{E[u_j]}{N - m_j} = \frac{E[m_j]}{m_j}$$

and defines the martingale difference with respect to $\mathcal{F}_{j-1}$ as

$$D_j = (N - M_j)m_j - M_j u_j.$$

Conditioning on $\mathcal{F}_{j-1}$, m_j and u_j are assumed to be independently binomially distributed. The zero-mean martingale is then

$$\sum_j D_j = \sum_j \left\{(N - M_j)m_j - M_j u_j\right\}.$$

Equating the zero-mean martingale to its observed mean, Yip (1991) proposes the following estimator of population size

$$\widehat{N} = \frac{\sum_{j=1}^{K} M_j(u_j + m_j)}{\sum_{j=1}^{K} m_j},$$

which is the same as the Schnabel (1938) estimator, $\widehat{N} = \frac{\sum_j^N M_j n_j}{\sum_j^N m_j}$. The estimated standard error is

$$\widehat{se}(\widehat{N}) = \left[\frac{1}{\widehat{N}} \sum_{j=1}^{K} n_j(\widehat{N} - n_j)M_j(\widehat{N} - M_j)\right]^{\frac{1}{2}} / \sum_{j=1}^{K} m_j.$$

The same author also suggests the following estimator

$$\widehat{N} = \frac{\sum_{j=1}^{K} W_{j-1} M_j n_j}{\sum_{j=1}^{K} W_{j-1} m_j},$$

where W_{j-1} are weights. If $W_{j-1} = 1$, we again have the Schnabel (1938) estimator. Optimal weights (those that minimize the asymptotic confidence interval) are $W^*_{j-1} = \frac{1}{(1-p_i)(N-M_i)}$, where p_i is defined as the capture probability of any individual on the i^{th} capture occasion.

Lloyd and Yip (1991) suggested modelling heterogeneous capture probabilities with a beta(α, β) distribution and solving a system of martingales set equal to 0. Solved iteratively, Lloyd and Yip (1991) showed that the estimator was reasonable for moderate sample sizes, but performed poorly when there was a high degree of heterogeneity.

Lloyd (1992) suggests limiting the parameter which models heterogeneity, effectively saying that extreme heterogeneity cannot be observed beyond some limit. Then, if the bound is exceeded, Lloyd (1992) uses the estimator of Lloyd and Yip (1991).

The martingale method compares favorably to the jackknife estimator, though the variance estimates do not perform well compared to bootstrap standard errors. Lloyd (1994) found that martingale estimators for time or behavioral response are fully efficient when compared to maximum likelihood. Martingale estimators have not yet proven to be as popular as likelihood-based methods due to their relative computational complexity and lack of model selection statistics.

1.6 Multi-List Methods

Multi-list methods are similar to capture-recapture methods, but multi-list methods are most commonly used with human population studies. In multi-list methods, "capture" corresponds to being observed on a list. It is commonplace for lists to be compiled from administrative sources, such as hospital admissions or treatment centers.

Multi-list population estimation, or multi-list estimation, has become commonplace in fields such as the epidemiological and health sciences and census under-count

adjustments. Specific examples include estimating alcohol related problems (Corrao et al., 2000), fetal alcohol syndrome (Egeland, Perham and Hook, 1995), homelessness (Fisher et al., 1994) and prevalence of Down Syndrome (Orton, Rickard and Miller, 2001; Huether and Gummere, 1982; Hook and Regal, 1982; Roecker and Huenther, 1983).

Statistical capture-recapture methodology tailored to human population estimation studies was examined by Hook and Regal (1992, 1995 and 2000). Meanwhile, the paired papers of the International Working Group for Disease Monitoring and Forecasting (IWGDMF) (1995a,b) greatly facilitated application of the methods in the epidemiological and health sciences.

1.6.1 Multi-List and Capture-Recapture: Key Differences

In both types of studies, the primary objective is estimating the size of some unknown population. Meanwhile, capture probabilities, p_k, are often nuisance parameters.

An important difference between capture-recapture and multi-list methods is that there is no time ordering of lists. In some cases, the lack of time ordering means that certain models from the animal case are not applicable. For example, in model M_b, the behavioral effect corresponds to a change in capture probability after the first capture. Because there is no time ordering to the lists, this model is typically not of interest.

However, in capture-recapture, it is common that results of successive capture occasions are assumed independent. In multi-list methods, there are often list interactions. For example, being observed on list 1 may mean that an individual is more likely to be observed on list 2. As Schwarz and Seber (1999) note, list dependence is the norm in multi-list methods. These effects have no direct correspondence in standard capture-recapture models except for "behavioral response" in capture-recapture settings. However, the lack of time ordering leads to more complex behavior in the case of multi-lists.

Since most statistical software programs have log-linear modelling routines, log-linear models have become the method of choice for modelling multi-list data. Also,

Cormack (1989) has shown how to equate model parameters to capture probabilities. As a result, log-linear modelling is often usefully applied to animal capture-recapture studies.

1.6.2 Log-linear Modelling

The first to apply log-linear models to capture-recapture data was Feinberg (1972), who drew a parallel between multi-list data and cell counts in a multi-way contingency table. Later, Darroch et al. (1993) and Agresti (1994) drew parallels between the log-linear model and the Rasch model used in educational testing. Historical developments of log-linear models for capture-recapture analysis are summarized in Feinberg (2000).

The Likelihood

In log-linear modelling, each of the K lists are treated as a dimension of a multi-way contingency table, with L unique capture histories, represented by ω. Let y_ω be the cell count corresponding to capture history ω, and μ_ω be the corresponding expected count. In vector form, they are written ($\mathbf{Y}$, $\boldsymbol{\mu}$).

Often a natural logarithm is taken as a "link" function relating μ_ω to linear parameters. Often, these parameters are 'list' effects, but may also correspond to covariates. The model is written

$$\begin{aligned} \mathbf{Y} &\sim Poisson(\boldsymbol{\mu}) \\ \log(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\Phi}, \end{aligned}$$

where $\mathbf{X}$ is a design matrix including list effect and interaction between lists and $\boldsymbol{\Phi}$ is the vector of parameters.

As noted, list dependence is very common in multi-list studies. An effective solution to incorporating list dependencies into the log-linear model is to include effects for list interactions. For example, if we find being on list 1 and list 3 makes an individual less likely to be captured on other lists, a three factor interaction effect representing list 1 and 3 interaction is included in the linear parameters.

Using the Poisson model framework, the likelihood is written

$$L(N, \mathbf{\Phi}) = \prod_{\omega:\omega\neq 0} \frac{\mu_\omega^{y_\omega} e^{-\mu_\omega}}{y_\omega!}.$$

Standard likelihood results follow from the likelihood function, such as the score function and the Fisher information.

Estimating Population Size, N

Inference for the unobserved cell, $\omega = \{0, \ldots, 0\}$, is drawn from the fitted model, $\widehat{n}_{\omega=\{0,\ldots,0\}} = e^{\widehat{\beta}_0}$. Note that $\widehat{\beta}_0$ represents the intercept of the linear component of the model. Population size is estimated as the sum of the observed and unobserved cell counts, $\widehat{N} = \sum_{\omega}^{L-1} n_\omega + \widehat{n}_{\omega=\{0,\ldots,0\}}$.

Estimating Capture Probabilities, p

Although in multi-list studies, capture probabilities are rarely of importance, Cormack (1989) equates model parameters to capture probabilities by solving the system of equations to determine the capture probabilities. For illustration, in the two-list case, we have

$$\begin{aligned} Np_1(1-p_2) &= e^{\beta_0+\beta_1} \\ N(1-p_1)p_2 &= e^{\beta_0+\beta_2} \\ Np_1p_2 &= e^{\beta_0+\beta_1+\beta_2}, \end{aligned}$$

where β_1 and β_2 represent effects of list 1 and 2, respectively, on capture probability.

The capture probabilities are $\widehat{p}_i = \frac{1}{1+e^{-\widehat{\beta}_i}}$, and estimated population size is $\widehat{N} = \frac{e^{\widehat{\beta}_0}}{(1-\widehat{p}_1)(1-\widehat{p}_2)}$. In cases where there are list interactions, it is not always possible to work out an explicit solution for capture probabilities.

The Poisson score function is $\sum_{\omega:\omega\neq 0} y_\omega(n_\omega - \mu_\omega)$. Thus, when the model is saturated, maximizing the Poisson likelihood amounts to solving $n_\omega = \mu_\omega$. Then there are list interactions, the capture probabilities can be expressed as functions of the log-linear parameters. When the model is not saturated, there are no closed form expressions for the parameter estimates, with log-linear or capture probabilities.

1.6.3 Heterogeneity

As we have seen in the capture-recapture section, heterogeneous capture probabilities result in negatively biased estimates of population size. A variety of techniques has been proposed to model individual level heterogeneity of capture probability in capture-recapture experiments, including the jackknife, estimating functions, sample coverage and martingales. In multi-list studies, the question of heterogeneity has been less studied.

Using the log-linear model framework described above, Evans, Bonett and McDonald (1994) stratify the population into s homogeneous sub-populations to reduce heterogeneity. First, $\mathbf{F}_h$ is defined as the vector of length t_h observable capture histories for stratum $h = 1, \ldots, s$. Within each stratum, it is assumed that the $(t_h + 1)$ vector of counts is multinomial, including the unobserved cell count. Then, the $\mathbf{F}_h$ are 'stacked' into a single $(\sum t_h 1)$ vector $\mathbf{F}$ which is modelled using a log-linear relationship

$$\mathbf{F} = exp\,(\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ is an $st \times sq$ design matrix, $\boldsymbol{\beta}$ is a $sq \times 1$ vector of unknown parameters and $\boldsymbol{\epsilon}$ is residual variation. Estimates are found using the observed data and extrapolated to the unobserved cells. In an analysis of capture-recapture data, Evans, Bonett and McDonald (1994) found the method detects heterogeneity of capture probabilities while not finding evidence of temporal or behavioral variation in capture probabilities.

Modelling census under-count in three lists, Darroch et al. (1993) use the Rasch model to incorporate individual-level heterogeneity. The Rasch model (Rasch, 1980) has been used extensively in an educational context to model test responses. Darroch et al. (1993) assume that individual heterogeneity is captured using a logistic function, with additive, fixed list effects and random individual-level effects. For a three-list example, capture probabilities are written

$$\log p(\omega) = \beta_0 + \beta_1\omega_1 + \beta_2\omega_2 + \beta_3\omega_3 + \gamma(c(\omega)),$$

where $\gamma(c(\omega))$ identifies the individual-level heterogeneity assuming the additive logit model. This is known as the quasi-symmetric model. A partial quasi-symmetry model relaxes the assumption of equal second order interactions.

The key difference between Evans, Bonett and McDonald (1994) and Darroch et al. (1993) is that the former include variables to account for the heterogeneity, while in the latter case, the variable related to heterogeneity is unobserved.

Hook and Regal (1993) developed methodology for the two-list case which takes into account non-independence of lists, ∇, and variable catchability, Δ. Using notation different from capture-recapture studies, let A_B, A_C and A_{BC} denote the number of cases on list B, C and B *and* C jointly. First, split lists B and C into 'subgroups' 1 and 2 based on covariate information. Let $A_{1,B}$ represent the number of cases ascertained by source B in subgroup 1, etc.

As seen previously, the maximum likelihood estimates of population size for each subgroup, when considered separately, are

$$\begin{aligned} \widehat{N}_1 &= \frac{A_{1,B}A_{1,C}}{A_{1,BC}} \\ \widehat{N}_2 &= \frac{A_{2,B}A_{2,C}}{A_{2,BC}}. \end{aligned}$$

Define p_1 and p_2 to be the proportion of the affected population N in each of the two subgroups. Let $b_1 = \frac{A_{1,B}}{N}$, an estimate of the probability of ascertainment of cases by list B in subgroup 1. Similarly, $(bc)_1 = \frac{A_{1,BC}}{N}$ estimates the probability of ascertainment of cases by list B and C in subgroup 1, and $c_1 = \frac{A_{1,C}}{N}$ estimates the probability of ascertainment of cases by list C in subgroup 1.

Also, define $r_1 = N\frac{A_{1,BC}}{A_{1,B}A_{1,C}}$, the source dependency in subgroup 1. And let $\Delta b = \frac{A_{2,B}}{N} - \frac{A_{1,B}}{N}$, the difference between the proportions ascertained by list B in subgroup 1 and in subgroup 2. Then, Hook and Regal (1993) define list dependency and variable catchability as

$$\begin{aligned} \nabla &= (p_1 b_1 c_1)\Delta r_1 + (p_2 b_2 c_2)\Delta r_2 \\ \Delta &= p_1 p_2 \Delta b \Delta c, \end{aligned}$$

where $\Delta r_1 = r_1 - 1$.

The relative accuracy, a, is then equal to

$$a = \frac{p_1 b_1 c_1 + p_2 b_2 c_2 - \Delta}{p_1 b_1 c_1 + p_2 b_2 c_2 + \nabla}.$$

Hook and Regal (1993), summarize the impact on estimated population size based on non-independence of lists and variability of capture using the above relationship. For example, if $\Delta b > 0$ and $\Delta c > 0$, then $\Delta > 0$, and if $\Delta r_1 \geq 0$ and $\Delta r_2 \geq 0$, then $\nabla \geq 0$, which results in an overestimate of N.

Hook and Regal (1993) illustrate the benefits from 'pooling' lists to minimize list dependency effects and variability of capture. The above methods are not widely adopted since simpler methods for multiple lists (next sections) have been shown to be robust to departures from independence of lists and capture variability.

It is expected that methods for modelling heterogeneity discussed in capture-recapture will find their way into multi-list methods as the more sophisticated approaches become incorporated into standard software applications.

1.7 Common Methodological Issues

1.7.1 Bias Reduction in Small Samples

Since Chapman's (1951) improvement to the Petersen estimator for small sample sizes, it has been known that capture-recapture estimators are sensitive to small cell counts. In the K-list case, Evans and Bonett (1994) propose adding 0.5^{K-1} to the observed frequency of each cell.

According to Evans, Bonett and McDonald (1994), the log-linear model with Evans and Bonett's (1994) proposed adjustment to cell counts can be written as:

$$(\mathbf{F} + \mathbf{Z}) = exp\,(\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\epsilon},$$

where, $\mathbf{Z}$ is a $(t-1)\times 1$ vector with each element equal to 0.5^{K-1}, where K is the number of sampling occasions or lists. The remaining parameters are previously specified in the section on stratification to reduce heterogeneity. Results of this adjustment have shown that population estimates are less biased and have smaller MSE.

More recently, Rivest and Levesque (2001) have suggested modifying cell counts to reduce first order bias. This approach was shown to reduce bias and MSE of population size estimates and improve on Evans and Bonett's (1994) frequency adjustments

for some models described by Otis et al. (1978).

1.7.2 Estimates of Precision

Wald Intervals

It is often desirable to include a estimate of the precision of the estimated population size. Based on the asymptotic properties of the likelihood estimators for generalized linear models, Wald-type confidence intervals can be derived such that the confidence interval is $\widehat{N} \pm z_{\alpha/2}\widehat{se}(\widehat{N})$.

Coull and Agresti (1999) show that in the presence of severe heterogeneity, likelihoods may be flat and estimates may be unstable, leading to wide confidence intervals.

Several authors have noted that the distribution of estimates may be highly skewed and have suggested transformations before finding Wald-based intervals. Specifically, Sprott (1981) suggests that $\widehat{N}^{1/3}$ is more symmetrical. Also, Borchers, Buckland and Zucchini (2002) illustrate the use of a log-transform.

Profile Likelihood Methods

Seber (1992) and Cormack (1992) note that profile likelihoods are generally preferred to Wald-based intervals. As Cormack (1992) notes, $\widehat{N}$ in the simple Petersen two sample case, is estimated by a ratio which includes, in its denominator, a small random component (m_2), which results in a skewed distribution of $\widehat{N}$. At the extreme, m_2 can observe values of 0, resulting in infinite bias.

Percentile-Based Methods

Buckland and Garthwaite (1991) advocate the use of the bootstrap for confidence intervals. Buckland (1984), Norris and Pollock (1996) and Chao, Chu and Hsu (2002) also recommend the bootstrap for variance estimators and confidence intervals. Buckland and Garthwaite (1991) propose that the bootstrap should be preferred over the jackknife estimator because a robust method for determining confidence intervals has

not been developed (as in the bootstrap), and the pseudovalues are assumed normally distributed. Buckland and Garthwaite (1991) also suggest that the jackknife confidence intervals may be more susceptible to variability when list sizes are small.

In their text, Borchers, Buckland and Zucchini (2002) discuss two alternative bootstrap methods; a parametric bootstrap based on multinomial sampling, the other nonparametric based on sampling capture histories. In multi-list studies reviewed, neither method appears to be superior to the other. The limiting factor is computational intensity.

1.7.3 Model Selection

Classical model selection uses a series of likelihood ratio tests to select the 'best' fitting model. As Feinberg (1972) notes, the χ^2 and G^2 statistics have asymptotic chi-square distributions under the null hypothesis that the fitted model is correct. Also, Evans, Bonett and McDonald (1994b) suggest the likelihood ratio and Wald test to aid model selection.

However, this approach ignores biological considerations underlying competing models. Cormack (1989) suggests that "Selection of an appropriate model to represent the data should be governed by biological considerations, but is aided by consideration of the residual deviance or Pearson χ^2".

A concern with strictly adhering to likelihood ratio tests is that model selection in the presence of small cell counts tends to indicate a simpler model than that which is realistic of the underlying dynamics (Agresti, 1994). This view is supported by Cormack (1989), who suggests that model selection methods may be unreliable in the presence of small cell counts. Agresti (1994) and Coull and Agresti (1999) note that the consequences of a simpler model are to have smaller standard errors and narrower confidence intervals than otherwise justified.

AIC

Lebreton et al. (1992) propose an information theoretic approach to model selection using Akaike's Information Criterion (AIC),

$$AIC = -2\log(L(\widehat{N}, \widehat{\beta})) + 2p,$$

where $L(\widehat{N}, \widehat{\beta})$ is the likelihood evaluated at the MLE with p parameters. Some authors prefer other criteria for model selection, the BIC, AIC_c or the $CAIC$ (discussed in the capture-recapture modelling context by Burnham, White and Anderson, 1995). These statistics are defined as

$$\begin{aligned} BIC &= -2\log(L(\widehat{N}, \widehat{\beta})) + p\log(n) \\ AIC_c &= AIC + \frac{2p(p+1)}{n-p-1} \\ CAIC &= -2\log(L(\widehat{N}, \widehat{\beta})) + p(\log(n)+1), \end{aligned}$$

where n is the number of observations. The Bayes Information Criteria (BIC) adjusts for the number of observations. The AIC_c adjusts for small sample sizes, while $CAIC$ is the Consistent AIC, which penalizes having more parameters more heavily than the BIC.

Application of the AIC is straightforward. For each model under consideration, the AIC statistic is computed. Models are ranked according to their AIC. The model with the smallest AIC is chosen as the 'best' model. The BIC, AIC_c and the $CAIC$ are applied in the same manner.

Using simulated datasets from a closed population, Stanley and Burnham (1998) report that using the AIC for model selection and model averaging results in the lowest RMSE relative to the AIC_c and $CAIC$.

Burnham and Anderson (1992) suggest the AIC statistic for model selection in open populations, and Huggins (1991) for closed populations. Burnham, White and Anderson (1995) compared information theoretic approaches to likelihood ratio tests and found that the AIC performed favorably.

1.7.4 Model Uncertainty and Estimates of Precision

Estimates of precision of $\widehat{N}$ are usually computed conditional upon the model selected. The impact of mis-specifying the model is inaccurate estimates of precision (and population size). As a result, model selection is an area that is increasingly attracting attention, as evidenced by the recent text of Burnham and Anderson (1998).

Buckland, Burnham and Augustin (1997) propose model weighting to adjust for model uncertainty. There are two popular methods of determining model weights. In the first, the difference in AIC from the best model is used to estimate the weight directly (Burnham and Anderson, 1998). The second approach is bootstrap-based. A series of bootstrap samples are generated by re-sampling from the population with a nonparametric bootstrap. For each sample, apply the model selection procedure, selecting from one of $1, \ldots, K$ distinct models using either model selection criteria, AIC or BIC. The weight for each model, w_k, is the proportion of times that each model was selected as the best approximating model ($\sum_{k=1}^{K} \omega_k = 1$).

Once the weights are determined, population size is estimated as

$$\widehat{N} = \sum_{k}^{K} w_k \widehat{N}_k.$$

Buckland, Burnham and Augustin (1997) also propose a variance estimator for this method

$$var(\widehat{N}) = \sum_{k}^{K} w_k^2 \left(var(\widehat{N}_k|\beta_k) + \beta_k^2 \right),$$

where β_k is the model mis-specification bias, estimated as $\widehat{\beta}_k = \widehat{N}_k - \widehat{N}$. $\widehat{var}(\widehat{N}|\beta_k)$, the variance of the estimated population size given model k, is estimated in the regular manner.

Buckland, Burnham and Augustin (1997) note that "Raftery (1996) and Kass and Raftery (1995) use Bayes factors to incorporate model selection in a Bayesian context, but the method can be sensitive to the choice of priors." Madigan and York (1997) demonstrate that a Bayesian approach to model averaging outperforms frequentist approaches.

1.7.5 Bayesian Methods

Castledine (1981) first applied Bayesian analyses to capture-recapture data. Castledine (1981) assumed that capture probabilities are the same at each sampling occasion and that capture probabilities are independently and identically distributed with a known distribution. Beta priors are assumed for two similar models,

$$\begin{aligned} model_1 &: p_i = p, p \sim beta(a,b); \\ model_2 &: p_i \sim beta(a,b) \text{ independently.} \end{aligned}$$

Using the likelihood based on product of binomials, the posteriors are approximated (Smith, 1991) for both models. They are

$$\begin{aligned} model_1: \ \pi(N,p|data) &\propto \left(\begin{array}{c} N \\ \sum_{\omega:\{\sum \omega<2\}} n_\omega \end{array} \right) p^{\sum n_\omega + a - 1}(1-p)^{NK - \sum n_\omega + b - 1}\pi(N); \\ model_2: \ \pi(N,p|data) &\propto \left(\begin{array}{c} N \\ \sum_{\omega:\{\sum \omega<2\}} n_\omega \end{array} \right) \prod_i^K p_i^{n_i + a - 1}(1-p_i)^{N - n_i + b - 1}\pi(N). \end{aligned}$$

Integration over p (Castledine, 1981) finds

$$\begin{aligned} model_1: \ \pi(N|data) &\propto \frac{N!}{(N - \sum_{\omega:\{\sum \omega<2\}} n_\omega)!} \frac{(NK - \sum n_\omega + b - 1)!}{(NK + a + b - 1)!}\pi(N); \\ model_2: \ \pi(N|data) &\propto \frac{N!}{(N - \sum_{\omega:\{\sum \omega<2\}} n_\omega)!} \prod_i^K \frac{(NK - n_i + b - 1)!}{(NK + a + b - 1)!}\pi(N). \end{aligned}$$

Smith (1991) determined the exact posterior distribution for model M_t. He also found that when the number of sampling occasions is large and the number of fraction recaptured is small, different priors have a large impact on estimates of population size. Basu and Ebrahimi (2001) and Tardella (2002) extend the Bayesian capture-recapture model to include heterogeneity of capture probabilities. The Bayesian methods produce similar estimates to frequentist approaches, however, their complexity has limited their wider application.

George and Robert (1992) introduce the use of Gibbs sampling to avoid the numerical approximation methods of the posterior and the assumptions required to calculate the exact posterior. George and Robert (1992) and Basu and Ebrahimi (2001)

demonstrate how Gibbs sampling can be used to calculate the Bayesian estimator of N.

York et al. (1995) propose a Bayesian approach to model uncertainty. They suggest averaging over all models within the class of models. That is

$$pr(N|data) = \sum_{k=1}^{K} pr(N|M_{(k)}, data) \times pr(M_{(k)}|data),$$

where $M_{(k)}$ are the models under consideration. Madigan and York (1995 and 1997) show that the coverage of model averaging outperforms frequentist approaches in several capture-recapture applications. King and Brooks (2001) extend the work of Madigan and York (1997) to model-averaged estimates of population size by calculating posterior model probabilities.

Model averaging is becoming increasingly adopted in applications of capture-recapture and multi-list methods (Stanley and Burnham, 1998 and Huakau, 2001). Because the limitations of the approach are computational, it is likely that model averaging to incorporate model uncertainty will continue to gain favor.

1.8 Other Violations of Assumptions

Capture-recapture methods reviewed assume that tags are not lost, erroneously recorded or mismatched across lists. Recently, methods have been developed that allow these assumptions to be relaxed.

Rajwani and Schwarz (1997) propose a tag-loss adjustment for the simple Petersen experiment. In their approach, another survey estimates the number of overlooked tags in the second sample. Using this number, the population estimate is suitably adjusted for missed tags. Using the same information on the number of missed tags, Rajwani and Schwarz (1997) also propose an estimator for the variance of N.

Seber, Huakau and Simmons (2000) examine the effect of list errors on estimates of population size. Seber, Huakau and Simmons (2000) partition 'tags' into two components. In the South Auckland Diabetes Study analysis of Huakau (2001), first name, surname, age, date of birth, sex, street name and suburb name were available

for each individual. Tag A was defined to be first name, surname and age, while Tag B was defined to be date of birth sex, street name and suburb name. Seber, Huakau and Simmons (2000) show that if error rates are high on a single tag (A or B alone), population size is overestimated.

To adjust the population estimate for mismatches, their estimator is

$$N^* = \frac{(n_1+1)(n_2+1)}{(m_T+1)}\left\{1 - \frac{m_A m_B}{m_T(m_{AB}+1)}\right\} - 1,$$

where m_A is the number of individuals matching on tag A only, m_B on tag B only, m_T are those matching on tag A only plus tag B only and plus matching on both tag A and B. Seber, Huakau and Simmons (2000) also derive a variance estimator for this estimate. Huakau (2001) applies this estimate in a model selection context to derive a model-averaged estimate of population size. Lee et al. (2001) and Lee (2002) have extended this approach to more than two lists when the lists are not necessarily independent.

1.9 Outline of Thesis

This thesis uses log-linear modeling to derive estimates of population size in the multi-list context when the usual assumptions regarding list coverage do not apply. Chapters 2, 3 and 4 are extended versions of papers submitted for publication. As a result, there is some repetition of the introductory material.

1.9.1 Chapter 2

Stratification of lists has often been used to reduce the biases caused by heterogeneity in the probability of list membership among members of a target population. In this chapter, a method is developed to deal with cases when not all lists are active in all strata.

This chapter builds upon the work of Hook and Regal (1993) and presents a generalized approach for more than two lists. Using a log-linear modelling framework, list dependencies and differential probabilities of ascertainment are incorporated.

This methodology uses an EM algorithm and is applied three examples; estimating the number of people with diabetes, the number of people who misuse drugs, and the number of forest fires that occurred in recent history.

1.9.2 Chapter 3

A key assumption in multi-list methods is that individuals have a unique "tag" that allows them to be matched across all lists. This paper develops a multi-list methodology that relaxes the assumption of a single tag common to all lists.

The proposed methodology is akin to tag-loss methods (Rajwani and Schwarz, 1997) and methods for incorporating tag mismatches (Seber, Huakau and Simmons, 2000; Huakau, 2001; Lee et al., 2001; and Lee 2002), in that it uses supplementary information to improve estimates.

Population parameter estimates are found using estimating functions. An example illustrates its application to estimating the prevalence of diabetes and a simulation study investigates conditions under which the methodology is robust to different list and population sizes.

1.9.3 Chapter 4

Existing multi-list methods assume that an individual's "tag" uniquely identifies an individual. This chapter relaxes this assumption, and presents a methodology for multi-list methods when not all list members can be matched across lists because of missing tags or partially missing tags.

Estimating functions are used to derive estimates of population size, capture probabilities and rates of tag loss. An example is used to illustrate the methodology and how to select among competing models. A simulation study compares our results against methods based on complete information.

Chapter 2

Incomplete and Partially Stratified Lists

Multi-list capture-recapture methods are commonly used to estimate the size of elusive populations. Stratification has often been used to reduce the biases caused by heterogeneity in the probability of list membership among members of the population. We develop a method to deal with cases when not all lists are active in all strata.

Hook and Regal (1993) considered a two-list problem with stratification in the context of estimating the number of individuals affected with Huntington's disease, but their approach required each list to cover all strata. This analysis builds upon the work of Hook and Regal (1993) and presents a generalized approach for more than two lists when not all lists cover all strata. We allow for list dependencies and differential probabilities of ascertainment in each list using a log-linear modelling framework described by Otis et al. (1978) and Evans, Bonett and McDonald (1994b).

Log-linear models have become a standard statistical tool to analyze capture-recapture data in multi-dimensional contingency table contexts since being introduced by Feinberg (1972). Cormack (1989) applied log-linear models to a wider scope of problems and linked ecological or biological parameters to capture histories. An historical development of the applications of the log-linear model to epidemiological problems is given by the IWGDMF (1995a, 1995b), while a detailed history of ecological applications of capture-recapture methodology is given by Seber (1982, 1992)

and Schwarz and Seber (1999). Log-linear models are now used for a broad spectrum of problems, including the estimation of census under-count (Darroch et al., 1993) and animal population size (Agresti, 1994). Our method also uses log-linear models applied to contingency tables with unobservable cells that result when not all lists operate in all strata.

We begin with the notation to be used in the paper and then develop our model using an EM algorithm. Finally, we applied our method to three examples; estimating the number of people with diabetes, the number of people who misuse drugs and also the number of forest fires that have taken place in recent history.

2.1 Notation

Suppose the closed population of interest is stratified into I strata. There are K lists which sample from the population. A "capture-history" vector ω can be created for each subject in a stratum with components $\omega_1, \ldots, \omega_K$. Indicator variables $\omega_1, \ldots, \omega_K$ identify whether the subject was recorded on each of the K lists in stratum k. Elements of ω are written

$$\omega_k = \begin{cases} 1 & \text{list } k \text{ recorded the subject;} \\ 0 & \text{list } k \text{ did not record the subject;} \\ \cdot & \text{list } k \text{ was not present to record the subject.} \end{cases}$$

Each subject appears in one and only one stratum.

2.1.1 Parameters

Define the following

$\mu_{i,\{\omega_1,\ldots,\omega_K\}}$	Expected cell count in stratum i with history $\omega_1, \ldots, \omega_K$;
λ_i	effect of stratum i;
β_0	intercept;
β_m	effect of list m;
β_{mn}	interaction effect between lists m and n;
$\lambda\beta_{i,m}$	effect of list m in stratum i;

	List 1	List 2	List 3
Stratum 1	operating	not operating	operating
Stratum 2	operating	operating	operating

Table 2.1: Partially stratified example; three lists, two strata.

The notation is extended to higher order interactions in the usual way. Define the vector $\mathbf{\Phi}$ be the parameter set.

2.1.2 Statistic

Define the statistics

$Y_{i,\{\omega_1,\ldots,\omega_K\}}$ Observed cell count in stratum i with history $\omega_1, \ldots, \omega_K$.

Define $\mathbf{Y}$ to be the vector of observable statistics.

2.2 Model Development

We illustrate model development by referring to a simple example of a population stratified into 2 strata, where 3 lists were available to identify subjects within the population. Among the three lists, one list had coverage for only one stratum, while the other two lists had complete coverage of the population. The representation of lists and strata is summarized in tabular format in Table 2.1.

Because list 2 was not operating in stratum 1, not all history vectors are observable. In this example, only the statistics $Y_{1,\{0,\cdot,1\}}, Y_{1,\{1,\cdot,0\}}$ and $Y_{1,\{1,\cdot,1\}}$ are observable in stratum 1.

A contingency table could be constructed for each stratum, cross-classifying elements in the population by their membership on the various lists. In this case, the data from stratum 1 results in a 2×2 contingency table while the data from stratum 2 results in a 2×2×2 contingency table with each table having one unobservable cell.

One approach to estimating the overall population size is to fit separate log-linear models in each stratum and simply total the estimated population sizes over the

strata. In the above example, only the simple Petersen model could be used in stratum 1 (using lists 1 and 3), while in stratum 2, a more complex log-linear model could be fit allowing for 2-factor interactions among the lists. These interaction terms represent dependencies between lists.

The major difficulty in fitting log-linear models to all strata simultaneously is that the structure of the contingency table may differ among each strata. For example, stratum 1 is represented by a 2^2 table, while stratum 2 is represented by a 2^3 table.

At the same time, we would like to use the information on "list interactions" from stratum 2 to improve the estimate in stratum 1. In doing so, we need to assume that the parameters not estimable from the observed data in a particular stratum are common across strata.

We start by assuming the existence of an underlying "complete" (unobservable) $I \times 2^K$ contingency table with counts $\mathbf{Z}$. The statistics $\mathbf{Y}$ are related to the unobservable counts $\mathbf{Z}$ through various linear combinations. For example, $Y_{1,\{1\cdot 1\}} = Z_{1,\{101\}} + Z_{1,\{111\}}$.

We model the counts $Z_{i,\{\omega\}}$ as Poisson random variables with means $\widetilde{\mu}_{i,\{\omega\}}$, *i.e.,*

$$\begin{aligned} \mathbf{Z} &\sim \text{Poisson}(\widetilde{\mu}) \\ \log(\widetilde{\mu}) &= \mathbf{X}\widetilde{\mathbf{\Phi}}, \end{aligned}$$

where $\mathbf{X}$ is the usual design matrix that relates the parameter set $\widetilde{\mathbf{\Phi}}$ (which contains stratum, list, list interactions and stratum $\times$ list effects) to the underlying mean $\widetilde{\mu}$.

In our simple 2-stratum example, the parameterization $\mathbf{X}\widetilde{\mathbf{\Phi}}$ of the design matrix

for a main effects only model (strata and list effects) is written

$$\log\begin{pmatrix} \widetilde{\mu}_{1,\{0,0,1\}} \\ \widetilde{\mu}_{1,\{0,1,0\}} \\ \widetilde{\mu}_{1,\{0,1,1\}} \\ \widetilde{\mu}_{1,\{1,0,0\}} \\ \widetilde{\mu}_{1,\{1,0,1\}} \\ \widetilde{\mu}_{1,\{1,1,0\}} \\ \widetilde{\mu}_{1,\{1,1,1\}} \\ \widetilde{\mu}_{2,\{0,0,1\}} \\ \widetilde{\mu}_{2,\{0,1,0\}} \\ \widetilde{\mu}_{2,\{0,1,1\}} \\ \widetilde{\mu}_{2,\{1,0,0\}} \\ \widetilde{\mu}_{2,\{1,0,1\}} \\ \widetilde{\mu}_{2,\{1,1,0\}} \\ \widetilde{\mu}_{2,\{1,1,1\}} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \lambda_1 \end{bmatrix}.$$

The observed cells have means μ that are linear combinations of $\widetilde{\mu}$. The relationship between μ and $\widetilde{\mu}$ is written

$$\mu = \mathbf{T}\widetilde{\mu}$$

for some matrix $\mathbf{T}$. For example $\mu_{1,\{1\cdot 1\}} = \widetilde{\mu}_{1,\{101\}} + \widetilde{\mu}_{1,\{111\}}$.

In our example, the observable means are written as linear combinations $\mu = \mathbf{T}\widetilde{\mu}$

$$\begin{bmatrix} \mu_{1,\{0,\cdot,1\}} \\ \mu_{1,\{1,\cdot,0\}} \\ \mu_{1,\{1,\cdot,1\}} \\ \mu_{2,\{0,0,1\}} \\ \mu_{2,\{0,1,0\}} \\ \mu_{2,\{0,1,1\}} \\ \mu_{2,\{1,0,0\}} \\ \mu_{2,\{1,0,1\}} \\ \mu_{2,\{1,1,0\}} \\ \mu_{2,\{1,1,1\}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \widetilde{\mu}_{1,\{0,0,1\}} \\ \widetilde{\mu}_{1,\{0,1,0\}} \\ \widetilde{\mu}_{1,\{0,1,1\}} \\ \widetilde{\mu}_{1,\{1,0,0\}} \\ \widetilde{\mu}_{1,\{1,0,1\}} \\ \widetilde{\mu}_{1,\{1,1,0\}} \\ \widetilde{\mu}_{1,\{1,1,1\}} \\ \widetilde{\mu}_{2,\{0,0,1\}} \\ \widetilde{\mu}_{2,\{0,1,0\}} \\ \widetilde{\mu}_{2,\{0,1,1\}} \\ \widetilde{\mu}_{2,\{1,0,0\}} \\ \widetilde{\mu}_{2,\{1,0,1\}} \\ \widetilde{\mu}_{2,\{1,1,0\}} \\ \widetilde{\mu}_{2,\{1,1,1\}} \end{bmatrix}.$$

Note that $\widetilde{\mu}_{1,\{0,1,0\}}$ plays no part in estimating the vector μ as the corresponding column of $\mathbf{T}$ has all entries equal to 0.

Finally, from the properties of the Poisson distribution, we now model the counts in each cell of the observed contingency table as independent Poisson random variables with means $\mu_{i,\{\omega_1,\ldots,\omega_K\}}$, where $\mu_{i,\{\omega_1,\ldots,\omega_K\}}$ is a function of parameters for the list, strata, and interaction effects. The likelihood function is

$$L\left(\widetilde{\mathbf{\Phi}};\mathbf{y}\right) = \prod_{i=1}^{I}\left[\prod\cdots\prod_{\substack{\omega_1,\ldots,\omega_K\in\{1,0,\cdot\} \\ \forall\{\omega_1,\ldots,\omega_K\}\in Y_i}} \frac{e^{-\mu_{i,\{\omega_1,\ldots,\omega_K\}}}{\mu_{i,\{\omega_1,\ldots,\omega_K\}}}^{y_{i,\{\omega_1,\ldots,\omega_K\}}}}{y_{i,\{\omega_1,\ldots,\omega_K\}}!}\right],$$

where $\{\omega_1,\ldots,\omega_K\} \in Y_i$ defines observable cells in stratum i, $\mu = \mathbf{T}\widetilde{\mu}$ and $\log(\widetilde{\mu}) = \mathbf{X}\widetilde{\mathbf{\Phi}}$.

The above model can be fit in several ways. First, the likelihood function can be maximized directly. Although standard log-linear routines can be used, this approach is computationally difficult since the model is no longer log-linear in the parameters.

Second, the estimate can be calculated from the system of estimating equations. The parameter estimates are solutions of $\mathbf{D}'\mathbf{V}(\mathbf{Y} - E[\mathbf{Y}]) = 0$, where $\mathbf{D}$ is the matrix of partial derivatives of μ with respect to $\mathbf{\Phi}$, where $\mu = \mathbf{T}e^{\mathbf{X}\mathbf{\Phi}}$. The working covariance $\mathbf{V}$ assumes Poisson counts. This leads to the same score equations as the maximum likelihood approach above.

Third, the maximum likelihood estimate can be obtained using an Expectation-Maximization (EM) algorithm approach (Dempster, Laird and Rubin, 1977). The major advantage of this approach is that the maximization step is easy to implement with standard log-linear software, although there is some minor programming involved in finding the expectation.

Using notation similar to that of McLachlan and Krishnan (1997), who provided an example of the EM algorithm applied to frequency count data, the conditional expected log-likelihood is

$$E_{\widetilde{\mathbf{\Phi}}}\left[\ell\left(\mathbf{\Phi}\right) \mid \mathbf{y}\right] \propto \sum_{i=1}^{I}\left[\sum_{\omega_1,\ldots,\omega_K \in \{1,0\}} \cdots \sum - \widetilde{\mu}_{i,\{\omega_1,\ldots,\omega_K\}} + z_{i,\{\omega_1,\ldots,\omega_K\}} \log(\widetilde{\mu}_{i,\{\omega_1,\ldots,\omega_K\}})\right],$$

where $\{\omega_1, \ldots, \omega_K\} \in \{1, 0\}$ defines the complete data observable cells in stratum i and $z_{i,\{\omega_1,\ldots,\omega_K\}} = E\left(Z_{i,\{\omega_1,\ldots,\omega_K\}} | \mathbf{Y}, \widehat{\mathbf{\Phi}}\right)$.

Because our underlying model is Poisson, the conditional estimates of the unobservable data were imputed using multinomial (or binomial) probabilities. Referring to our example,

$$Z_{1,\{1,1,1\}} \sim Bin(Y_{1,\{1,\cdot,1\}}, \frac{\widetilde{\mu}_{1,\{1,1,1\}}}{\widetilde{\mu}_{1,\{1,1,1\}} + \widetilde{\mu}_{1,\{1,0,1\}}}),$$

and $z_{1,\{1,1,1\}}$ is found by $Y_{1,\{1,\cdot,1\}} \times \frac{\widehat{\widetilde{\mu}}_{1,\{1,1,1\}}}{\widehat{\widetilde{\mu}}_{1,\{1,1,1\}} + \widehat{\widetilde{\mu}}_{1,\{1,0,1\}}}$.

The maximization step updates the parameter estimates corresponding to $\widetilde{\mu}$ using standard maximization methods for log-linear models. The cycles are repeated until the conditional expected log-likelihood is maximized.

An estimate of the standard error of the parameter estimates can be obtained by (1) calculating the asymptotic covariance matrix of $\mathbf{\Phi}$ directly using the inverse of the information matrix computed on the basis of the observed data log-likelihood,

or (2) using the usual estimating equations variance formula, or (3) using methods for finding standard errors from the EM algorithm, or (4) using bootstrap sampling methodology (Efron and Tibshirani, 1993).

The bootstrap methodology may be more suitable when cell sizes are small. To derive bootstrap samples, we sample from all capture histories with multinomial probabilities until the sum of the cell counts equals the original sample size (the count in each stratum is not necessarily the same). The sampling probabilities are set according to the observed counts in each stratum and capture history.

The overall population size is estimated by estimating the counts in the unobserved cells of the multi-way contingency table. More care is required in our context of partial stratification as the missing cell may be composed of several underlying cells. In our example, in stratum 1, the missing cell is $\mu_{1,\{0,\cdot,0\}}$, and $\mu_{1,\{0,\cdot,0\}} = \widetilde{\mu}_{1,\{0,1,0\}} + \widetilde{\mu}_{1,\{0,0,0\}}$. The estimates of the unobserved counts are then summed with observed counts to derive the estimate of population size.

Many possible models can be fit to the data. The most complex would allow for complete list by stratum interaction among the parameters and list interactions - this would be equivalent to fitting separate models to each stratum. The simplest may have all β parameters common across all strata - this would imply that the strata can be "ignored."

A goodness-of-fit statistic can be computed using a Pearson-type statistic:

$$\mathbf{X}^2 = \sum_{observed\ cells} \frac{(O-E)^2}{E},$$

which is compared to a chi-square distribution with degrees of freedom equal to the number of cells less one.

Model selection can be performed by calculating the Akaike information criterion (AIC) statistic (Akaike, 1973) where $AIC = -2\log(L(\widehat{N}, \widehat{\mathbf{\Phi}})) + 2p$, where $L(\widehat{N}, \widehat{\mathbf{\Phi}})$ is the maximum likelihood with p parameters. The model with the lowest AIC among the competing models is usually preferred, though additional insight into model selection is usually gained by considering the sources of the data. If several models have similar AIC, a weighted average of the MLE's can be computed. Buckland,

Burnham and Augustin (1997) show how standard errors can also be adjusted for model uncertainty using model averaging.

The greatest problem in all approaches is that the presence of missing cells makes it difficult to determine if all parameters are identifiable. The matrix $\mathbf{X}$ provides some assistance in determining whether model parameters are non-identifiable. If $\mathbf{X}$ is of full rank, parameters should be identifiable. Another problem can arise in model fitting if initial estimates are poorly chosen. The numerical optimization procedure may converge to a local maximum. The choice of initial values did not prevent convergence in the examples studied.

2.3 Examples

2.3.1 Auckland Diabetes Prevalence Study

We applied the methodology to data available from the Auckland Diabetes Study data. Huakau (2001) described four lists, whose information was to used to estimate the prevalence of diabetes in Auckland, New Zealand. The 4 lists were: general practioners records (G), pharmacy records (P), outpatient records (O) and inpatient discharge records (D). Huakau (2001) provided counts of exact matches for capture histories of all 4 lists by gender. The original data is available in Table 2.2.

Because all cell counts were available, we assumed that only some information was available to the study. We assumed that males were only present on lists P and O, while females were only present on lists G, O and D, as summarized in Table 2.3. The cells counts of the simulated data are available in Table 2.4.

A summary of the model selection process is presented in Table 2.5. The final model selected according to the AIC statistic had list effects and a stratum effect. The final model selected was $\widehat{\mathbf{\Phi}} = [G][P][O][D][\lambda]$, which gives a population estimate of $\widehat{N}_{\widehat{\mathbf{\Phi}}} = 22,813$. The estimated standard error is 344. As Agresti (1994) and Coull and Agresti (1999) note, a small variance is due in part to the very parsimonious model chosen, and the fact that model uncertainty is not incorporated into this standard error.

List G	List P	List O	List D	Males	Females
0	0	0	1	652	752
0	0	1	0	4865	4377
0	0	1	1	794	840
0	1	0	0	253	251
0	1	0	1	18	43
0	1	1	0	234	224
0	1	1	1	62	66
1	0	0	0	260	255
1	0	0	1	26	32
1	0	1	0	221	231
1	0	1	1	67	61
1	1	0	0	19	15
1	1	0	1	0	2
1	1	1	0	32	40
1	1	1	1	4	11

Table 2.2: Original data for Auckland Diabetes Study. Lists: General practioners records (G), Pharmacy records (P), Outpatient records (O) and Inpatient discharge records (D).

	Stratum	
List	Male	Female
List G	not	operating
List P	operating	not
List O	operating	operating
List D	not	operating

Table 2.3: Auckland Diabetes Study lists. Lists: General practitioners records (G), Pharmacy records (P), Outpatient records (O) and Inpatient discharge records (D).

Lists				
G	P	O	D	**Y**
Male				
·	0	1	·	5,947
·	1	0	·	290
·	1	1	·	332
Female				
0	·	0	1	795
0	·	1	0	4,601
0	·	1	1	906
1	·	0	0	270
1	·	0	1	45
1	·	1	0	271
1	·	1	1	72

Table 2.4: Simulated cell counts for Auckland Diabetes Prevalence Study data.

Model	**Φ**	Number of Parameters	AIC Statistic	Goodness-of-fit Statistic	$\widehat{N}$	$\widehat{se}(\widehat{N})$
(a)	$[G][PO][OD][\lambda]$	8	100.62	5.17	23,528	679
(b)	$[G][PO][OD]$	7	113.80	20.17	23,120	630
(c)	$[G][D][PO][\lambda]$	7	99.95	6.39	22,836	393
(d)	$[G][P][O][D][\lambda]$	6	97.96	6.40	22,813	344
(e)	$[G][P][O][D]$	5	111.59	21.89	22,576	308

Table 2.5: Model fitting summary of Auckland Diabetes Study data

Our final model has independent list effects, and a multiplicative effect of gender ($\widehat{\lambda}_{Female} = -0.07$), suggesting females were less in number than males, *i.e.* 93% of male abundance.

The Petersen estimate (Male) was 11,748 ($\widehat{se} = 426$), while the log-linear estimate (Female) was 12,462 ($\widehat{se} = 916$). The best fitting model to this stratum was the one with equal second-order interactions. Together, they yield an estimate of 24,210 ($\widehat{se} = 1,010$).

The same model, $\mathbf{\Phi} = [G][P][O][D][\lambda]$, fit to the complete $2 \times (2^4 - 1)$ contingency tables, gives $\widehat{N}_{log\text{-}linear} = 22,724$ ($\widehat{se} = 235$).

In an analysis using exact matches across lists by gender, Huakau (2001) employed tag-loss methods which used model averaging to account for model uncertainty. Using AIC criteria for model selection, $\mathbf{\Phi} = [GPO][GOD]$ was the most heavily weighted model, with $\widehat{N}_{\mathbf{\Phi}} = 26,610$. However, using BIC, $\mathbf{\Phi} = [GP][OD]$ was the most heavily weighted model; $\widehat{N}_{\mathbf{\Phi}} = 24,443$.

We were not surprised our models differed since summing over 'underlying' cell counts resulted in our model being unable to detect some multi-list interaction effects. Also, our analysis did not incorporate tag loss, which results in overestimating the population. Model uncertainty was not taken into account for $\widehat{se}(\widehat{N})$, and our estimated standard error (344) is likely underestimated because no adjustment was made for model uncertainty.

2.3.2 Scottish Drug Use Prevalence

Hay (2000) sought to quantify the prevalence of individuals misusing drugs in Scotland. We apply our methodology to their data collected in the Grampian Health Board. We focused our analysis on the area of Aberdeen City.

In the original study, there were four strata, corresponding to combinations of gender and two age groups. There were four lists; list 1 was compiled from the Substance Misuse Service and physician submissions to the Scottish Drug Misuse Database, list 2 was compiled from a counselling service, list 3 was compiled from a needle/syringe exchange, while list 4 was compiled from the Police and Social work

				Males		Females	
List 1	List 2	List 3	List 4	15-24	25-54	15-24	25-54
0	0	0	1	138	82	31	33
0	0	1	0	93	110	50	18
0	0	1	1	36	29	14	6
0	1	0	0	19	18	17	10
0	1	0	1	5	2	2	2
0	1	1	0	4	8	3	3
0	1	1	1	5	3	0	1
1	0	0	0	94	88	42	19
1	0	0	1	25	13	3	9
1	0	1	0	21	10	12	5
1	0	1	1	8	6	7	2
1	1	0	0	1	3	6	0
1	1	0	1	2	1	1	0
1	1	1	0	0	1	3	2
1	1	1	1	1	1	1	0

Table 2.6: Original data for Aberdeen City opiate/benzodiazepine misuse. List 1: Combined data from the Substance Misuse Service and GP returns to the Scottish Drug Misuse Database. List 2: Couselling service. List 3: Needle/syringe exchange. List 4: Combined data from the Police and Social work department.

department. The original data set is given in Table 2.6.

Because all data were available, we simulated not having all strata observable on each list. A summary of which lists were operating in which strata is shown in Table 2.7 and the simulated counts are in Table 2.8.

A summary of the model fitting procedure is found in Table 2.9. The selected model was model (f) with a population estimate of $\widehat{N}_{\widehat{\mathbf{\Phi}}} = 1,362$ ($\widehat{se}(\widehat{N}_{\widehat{\mathbf{\Phi}}}) = 92.3$), but model (e) also had substantial support.

Ordinarily, one would sum the estimates from the four strata. The Petersen estimate for $\widehat{N}_{Males,\ 15\text{-}24} = 598.6$ ($\widehat{se} = 109$), while $\widehat{N}_{Females,\ 15\text{-}24} = 349.8$ ($\widehat{se} = 68$). The best fitting log-linear model for Males, 25-54 was $\mathbf{\Phi} = [1][23][24][34]$, and $\widehat{N}_{Males,\ 25\text{-}54} = 945.5$ ($\widehat{se} = 664$). The best fitting log-linear model for Females, 25-54 was $\mathbf{\Phi} = [1][2][3][4]$, and $\widehat{N}_{Females,\ 25\text{-}54} = 181.1$ ($\widehat{se} = 119$). The overall population estimate is then $2,075$ ($\widehat{se} = 687$). The models are quite different in the different strata,

	Males		Females	
List	Age 15-24	Age 25-54	Age 15-24	Age 25-54
1	not	not	operating	operating
2	operating	operating	not	not
3	not	operating	not	operating
4	operating	operating	operating	operating

Table 2.7: Summary of which lists were operating in which strata. Drug misuse in Aberdeen City, Scotland example.

Lists				
1	2	3	4	**Y**
Male, 15-24				
·	0	·	1	207
·	1	·	0	24
·	1	·	1	13
Male, 25-54				
·	0	0	1	95
·	0	1	0	120
·	0	1	1	35
·	1	0	0	21
·	1	0	1	3
·	1	1	0	9
·	1	1	1	4
Female, 15-24				
0	·	·	1	47
1	·	·	0	63
1	·	·	1	12
Female, 25-54				
0	·	0	1	35
0	·	1	0	21
0	·	1	1	7
1	·	0	0	19
1	·	0	1	0
1	·	1	0	7
1	·	1	1	2

Table 2.8: Simulated data for Aberdeen City opiate/benzodiazepine misuse.

Model	$\mathbf{\Phi}$	p	AIC Statistic	$\mathbf{X}^2$	$\widehat{N}$	$\widehat{se}(\widehat{N})$
(a)	$[13][14][23][24][34][\lambda_1][\lambda_2][\lambda_3][\lambda\beta_{1,4}][\lambda\beta_{2,3}]$	15	216.29	82.02	1,105	103.8
(b)	$[13][14][23][24][34][\lambda_1][\lambda_2][\lambda_3][\lambda\beta_{2,3}]$	14	220.19	87.81	1,144	115.2
(c)	$[13][14][23][24][34][\lambda_1][\lambda_2][\lambda_3]$	13	222.18	90.38	1,229	140.9
(d)	$[13][14][24][34][\lambda_1][\lambda_2][\lambda_3]$	12	217.82	88.93	1,284	135.8
(e)	$[13][14][24][\lambda_1][\lambda_2][\lambda_3]$	11	216.50	90.21	1,470	114.9
(f)	$[2][13][14][\lambda_1][\lambda_2][\lambda_3]$	10	215.75	95.89	1,362	92.3
(g)	$[1][2][3][4][\lambda_1][\lambda_2][\lambda_3]$	8	221.37	106.43	1,478	95.2
(h)	$[1][2][3][4]$	5	412.40	310.93	1,386	54.3

Table 2.9: Model fitting summary of drug misuse in Aberdeen City, Scotland

ranging from independent lists in Females, 25-54, and complex 2-factor interactions for Males, 25-54.

Using the entire data, the author (Hay, 2000) estimated $\widehat{N} = 2,396$ ($\widehat{se}$ not provided), with a 95% confidence interval of (2,149-2,832). The model had two second-order list interaction effects. The author also fit a separate model to each stratum, obtaining $\widehat{N} = 2,519$ with a 95% confidence interval of (2,048-3,200).

Our model and population estimate differs substantially from Hay's (2000) because no observable statistics captured the relevant interaction terms that were present in the full dataset. A contributing factor is the large number of small cell counts (< 6) in the original data (see Table 2.6).

2.3.3 Forest Fire Incidence

Lastly, we applied our methodology to estimate the incidence of forest fires, where trees functioned as lists. This example provided an opportunity to observe how well our methodology performs with small cell counts.

Forest fire incidence was based on physical examination of tree cores. The first fire a tree experiences that damages the cambium initiates the fire recording ability of a tree. Subsequent fires can leave distinctive marks in the radial growth patterns of impacted trees. These rings were dated dendrochronologically to determine the exact

Stratum	Tree A	Tree B	Tree C
	Year Fire Observed		
1629-1688	1645, 1667, 1676, 1688	not operating	not operating
1689-1739	1697, 1706, 1721, 1732	not operating	1710, 1714, 1721, 1739
1740-1898	1741, 1751, 1756, 1771, 1780, 1794, 1799, 1829, 1840	1751, 1794, 1844, 1869, 1898	1794, 1829, 1844, 1869

Table 2.10: Observed Fires of Plot 12.

year of fire occurrence. Swetnam (1993) noted that there was excellent agreement between known dates of fire and tree ring data in a study of giant sequoias, although not every fire is recorded on each tree.

A plot is a small area in a forested region where several trees have been sampled in close physical proximity. Trees in a Plot are assumed to have experienced similar fire histories, though not all trees record all fires.

Although each of the K trees in the Plot recordS fire history information, the intervals over which fire history data was recorded are generally not the same length, nor do they have the same start or endpoints. The time dimension was stratified into I strata according to the number of trees alive and recording fires.

This analysis focussed on one plot from a larger study, conducted by Heyerdahl (1997). There were three trees recording fire information in Plot 12, with three distinct recording intervals. The first interval is 1629 to 1688, 60 years duration. The second interval is 1689 to 1739, 51 years duration. The third interval is 1740 to 1898, 159 years duration. Table 2.10 lists the observed fires by tree and stratum while the history vector representation of the observed fires is presented in Table 2.11.

Our contingency table is very sparse. Depending on the model being fit, sparse tables often lead to numerical and theoretical difficulties; it may be difficult to maximize the likelihood function, to obtain parameter estimates and to obtain unbiased goodness-of-fit statistics (Agresti, 1990).

Recording Tree			
A	B	C	**Y**
1629-1688			
1	·	·	4
1689-1739			
0	·	1	3
1	·	0	3
1	·	1	1
1740-1898			
0	0	1	0
0	1	0	1
0	1	1	2
1	0	0	6
1	0	1	1
1	1	0	1
1	1	1	1

Table 2.11: Cell counts for Plot 12.

Potential remedies to our sparse data were to add a small constant to cells or to pool over strata or lists. Because there didn't appear to be an ecological interpretation to pooling cells, we followed the bias reduction technique of Evans and Bonnett (1994a), who added 0.5^{k-1} to each cell, where k is the dimension of the contingency tables. In our case with multiple contingency tables of different sizes, a reasonable approach may have been to divide the total value of 2 among the total potential cells from the complete contingency tables and then subtract the total added from the final estimate.

A variety of models were fit to assess the presence of interval and tree effects. Forward selection was used to add additional terms while fitting successive models because the estimates tended to be numerically unstable when fit to sparse data. More general models, with strata, list and interaction effects, were over-parameterized and had many unidentifiable parameters.

A summary of the model fitting process is given in Table 2.12. There was very strong evidence that Model (a), where the probability of detecting fires is the same

Model	$\mathbf{\Phi}$	Number of Parameters	AIC	Goodness-of-fit Statistic	$\widehat{N}$	$\widehat{se}(\widehat{N})$
(a)	$[\beta_0]$	1	54.19	38.45	26.74	0.87
(b)	$[A][B][C]$	4	51.56	17.42	33.59	8.45
(c)	$[AB = AC = BC]$	5	52.37	16.51	36.30	49.45
(d)	$[AB = BC][AC]$	6	53.19	14.40	31.92	41.72

Table 2.12: Summary of model fitting process of Plot 12.

for each tree was not tenable. Model (b) was a reasonable representation and was the selected model, though Model (c) was a very close competitor. Model (c) is the model of equal second order interaction effects (quasi-symmetry model), while model (d) is a partial quasi-symmetric model.

Because of some trees were closely spaced in a Plot, there was reason to suspect strata and list interaction effects. We did not find this to be the case in our example, as more complex models did not lead to substantially better fits. We did not find it surprising that a simple model provided the best fit given the sparse data of Plot 12. The parameter estimates of model (b) are

$$\widehat{\mathbf{\Phi}} = \begin{bmatrix} \beta_0 \\ \beta_A \\ \beta_B \\ \beta_C \end{bmatrix} = \begin{bmatrix} -3.52 \\ 0.03 \\ -0.85 \\ -0.81 \end{bmatrix}.$$

The estimated number of fires was 33.59 ($\widehat{se} = 6.72$) in the 270 year recording period. Dividing by the length of the strata led to an estimated incidence of $\widehat{\delta} = 0.1244$ fires per year ($\widehat{se} = 0.0249$).

Because cell counts were small, we used a non-parametric bootstrap to estimate the standard error (200 replications). However, the simple non-parametric bootstrap had two limitations which may result in an underestimate of the standard error. First, if a stratum had only a single tree operating, all bootstrap samples would have this single tree with the observed number of fires. Second, if a stratum had a tree operating that failed to record any fires, then all bootstrap samples would also have this tree

failing to record any fires. A parametric bootstrap may be more appropriate in this case.

There was evidence in Table 2.12 to suggest that models (b)-(d) described the data almost equally well. Though not completed for this example, a model averaging approach (Hook and Regal, 1997 or Buckland, Burnham and Augustin, 1997) could be used to account for model uncertainty.

The goodness-of-fit test indicated that some lack of fit or overdispersion may be occurring (but the sparse data may make the statistic unreliable). Based on the goodness-of-fit statistic for Model (b), the overdispersion factor was estimated to be about $\frac{17.42}{7} \approx 2.5$. In other words, the estimated standard error perhaps should be increased by about 50%, however, the results from our simulation experiment (not yet discussed) seem to indicate that no adjustment was required.

For comparative purposes, if we were to estimate each stratum separately, we find: the observed minimum value of 4 observed fires in stratum 1, an estimate of 9 fires in stratum 2 (Petersen estimate), an estimate of 14.81 fires in stratum 3 (list effects log-linear model) which gives an overall estimate of 27.81 fires for the recording period, or an incidence rate of 0.1030 fires per year, or 17.2% lower than the estimate using the EM algorithm. It is not surprising that our estimate is larger since we assumed that parameters not estimable in a particular stratum were common across all strata, which is particularly true for stratum 1, where a single tree is operating.

2.4 Simulation Study

To assess our methodology, we generated simulated data with different list dependencies for various population sizes. We used 4 lists, each with two strata. Not all lists operated in all strata, as shown in Table 2.13. The traditional approach would have been to estimate population size in stratum 1 using a Petersen estimator, while a log-linear model would have been fit to counts of capture histories of stratum 2.

Four different studies were done. In each study, a model $\mathbf{\Phi}$ was specified as the true model for a given N. Cell counts derived from this model were used as the 'true' cell counts in all $2^4 - 1$ cells of both strata. Records were then re-sampled from the

	Stratum	
	1	2
List 1	not	operating
List 2	operating	not
List 3	operating	operating
List 4	not	operating

Table 2.13: Multi-list setting for the simulation study.

'true' generated dataset until the original sample size was equalled. This process was repeated 1,000 times for each simulation study.

For each simulation study, the best fitting model using the EM algorithm was fit to the 'true' data. This same model was fit to each simulated dataset. We also fit a model to the complete underlying data *i.e.*, no partially stratified lists. For each stratum, the best-fitting $2^4 - 1$ log-linear model was fit to the underlying data.

2.4.1 Effects of List Interactions

In the first study, lists were assumed to operate independently with the same capture probability for each stratum. The results in Table 2.14 show that the estimates derived from the EM algorithm were unbiased. The Petersen and log-linear model also gave good estimates of the population size in their respective strata. Our estimate has a smaller standard deviation than that of the sum. The standard deviation of the sum of the estimates is smaller than expected since the two estimates are negatively correlated through shared list 3.

In the second study, a strata effect was included. As shown in Table 2.15, our methodology's results were unbiased, as were the sum of the Petersen estimator and log-linear estimate, though the latter has a larger standard deviation.

In the third study, a list dependency effect was included. Table 2.16 shows that the EM algorithm estimate is slightly positively biased ($\approx 1\%$). However, the Petersen estimator was negatively biased ($\approx 6\%$), while the log-linear model was slightly positively biased ($\approx 1.5\%$). The estimate derived from the sum of the Petersen and

	Population Size N		
	10,000	20,000	40,000
EM algorithm			
Mean of estimates	10,000	20,010	40,000
std dev.	113	164	227
Petersen, stratum 1 (lists 2 and 3)			
Mean of estimates	4,993	9,992	20,000
std dev.	138	203	273
Log-linear, stratum 2 (lists 1, 3 and 4)			
Mean of estimates	5,000	10,010	20,000
std dev.	80	114	156
Sum of Petersen and Log-linear			
Mean of sum of estimates	9,993	20,000	40,000
std dev.	138	204	279
All $2^4 - 1$ counts available for each strata			
Mean of estimates in stratum 1	4,999	9,996	19,990
Mean of estimates in stratum 2	5,000	10,000	20,000
Mean of sum of strata estimates	9,999	20,000	40,000
std dev.	47	63	89

Table 2.14: Independent lists, 1,000 samples. Model was $\beta_1 = -1, \beta_2 = -0.8, \beta_3 = -0.5, \beta_4 = -0.3$

	Population Size N		
	10,000	20,000	40,000
EM algorithm			
Mean of estimates	10,020	20,010	40,020
std dev.	134	187	274
Petersen, stratum 1 (lists 2 and 3)			
Mean of estimates	6,232	12,450	24,910
std dev.	155	213	332
Log-linear, stratum 2 (lists 1, 3 and 4)			
Mean of estimates	3,783	7,550	15,100
std dev.	84	110	148
Sum of Petersen and Log-linear			
Mean of sum of estimates	10,020	20,000	40,010
std dev.	161	221	325
All $2^4 - 1$ counts available for each strata			
Mean of estimates in stratum 1	6,227	12,450	24,900
Mean of estimates in stratum 2	3,780	7,550	15,100
Mean of sum of strata estimates	10,010	20,000	40,000
std dev.	45	64	91

Table 2.15: Independent lists, stratum effect, 1,000 samples. Model was $\beta_1 = -1, \beta_2 = -0.8, \beta_3 = -0.5, \beta_4 = -0.3, \lambda = -0.5$

	Population Size N		
	10,000	20,000	40,000
EM algorithm			
Mean of estimates	10,110	20,200	40,350
std dev.	254	378	529
Petersen, stratum 1 (lists 2 and 3)			
Mean of estimates	5,781	11,560	23,280
std dev.	121	163	239
Log-linear, stratum 2 (lists 1, 3 and 4)			
Mean of estimates	3,618	7,223	14,450
std dev.	77	103	148
Sum of Petersen and Log-linear			
Mean of sum of estimates	9,399	18,780	37,740
std dev.	124	171	242
All $2^4 - 1$ counts available for each strata			
Mean of estimates in stratum 1	6,226	12,450	24,910
Mean of estimates in stratum 2	3,780	7,551	15,100
Mean of sum of strata estimates	10,010	20,010	40,000
std dev.	66	95	133

Table 2.16: Dependent lists, stratum effect, 1,000 samples. Model was $\beta_1 = -1, \beta_2 = -0.8, \beta_3 = -0.5, \beta_4 = -0.3, \lambda = -0.5, \beta_{13} = \beta_{23} = \beta_{34} = 0.2$

log-linear models was negatively biased ($\approx 6\%$). The standard deviation of the population estimates using the EM algorithm was roughly double that of the sum of the Petersen and log-linear model, although clearly, the EM algorithm would be preferred in this situation (the mean square error (MSE) of the 1,000 simulations using the EM algorithm is considerably smaller than that of the sum of the estimates.)

In the fourth study, in addition to the previous model, list $\times$ stratum interaction effects were included. The results in Table 2.17 show that the Petersen and log-linear models were both positively biased, resulting in a positively biased estimate of population size ($\approx 8\%$). The estimates using the EM algorithm were very close to the true population size. The MSE of the EM algorithm is considerably smaller than that of the sum of the estimates.

The simulation studies have shown that when list effects are independent, the EM algorithm offers only a modest improvement in precision compared to the sum

	Population Size N		
	10,000	20,000	40,000
EM algorithm			
Mean of estimates	10,040	20,020	40,010
std dev.	307	435	588
Petersen, stratum 1 (lists 2 and 3)			
Mean of estimates	6,797	13,600	27,200
std dev.	150	219	301
Log-linear, stratum 2 (lists 1, 3 and 4)			
Mean of estimates	4,174	8,333	16,680
std dev.	127	176	257
Sum of Petersen and Log-linear			
Mean of sum of estimates	10,970	21,940	43,880
std dev.	179	258	370
All $2^4 - 1$ counts available for each strata			
Mean of estimates in stratum 1	6,263	12,520	25,250
Mean of estimates in stratum 2	3,743	7,478	15,050
Mean of sum of strata estimates	10,010	20,000	40,010
std dev.	75	110	152

Table 2.17: Dependent lists, stratum effect, 1,000 samples. Model was $\beta_1 = -1, \beta_2 = -0.8, \beta_3 = -0.5, \beta_4 = -0.3, \lambda = -0.5, \beta_{13} = \beta_{23} = \beta_{34} = 0.2, \lambda\beta_{1,2} = \lambda\beta_{2,1} = 0.8$

of estimates from each of the strata. However, the Petersen and log-linear model provide poor estimates when there are list dependencies that cannot be modelled in each strata. In these cases, the EM algorithm is an improvement due to its much smaller mean squared error.

2.4.2 Small Sample Size Effects

As cell counts were very small in our forest fire example, a simulation study was conducted to evaluate the performance of the estimator with small cell counts.

The same list structure as in the forest fire example (Table 2.11) was used. We assumed that the estimated incidence of fires, $\widehat{\delta} = 0.1244$ per year, was the "true" incidence rate and assumed that the trees operated independently.

Common capture probabilities, p, as listed in Table 2.18 were used. Based on the observed counts, the independent lists model was fit (this was shown to be the best fitting model in the example) to the data. This process was repeated $n = 300$ times.

Theoretically, the estimates must have infinite bias, because there is a small, but non-zero probability that in some cases, no fires would have been recorded, by chance, on multiple trees. This is analogous to the case of no marks being recovered in a simple Petersen experiment. In practice, these cases are non-informative. Following Otis et al. (1978) such replicates were discarded from the simulations.

Even with discarding such "poor data", some residual, but small bias seemed to be present, and Table 2.18 suggests that this bias increased as the recording probabilities decreased, similar to the findings of Evans and Bonett (1994a). Recording probabilities lower than 0.15 led to many cells having counts of 0, resulting in an inability to fit the model.

As expected, the standard deviation of the estimates also increased with decreased recording probability. Comparing to the example, our bootstrap estimated standard error ($\widehat{se}(\widehat{\delta}) = 0.0249$) appears to be a reasonable estimate of precision even with the sparse data. Of course, uncertainty induced from the model selection process was not been accounted for.

Low recording probabilities can lead to sparse data and cause numerical problems

	Event Recording Probability of Each Tree					
	0.95	0.90	0.85	0.80	0.75	0.70
Mean $\widehat{\delta}$	0.1222	0.1222	0.1225	0.1226	0.1233	0.1236
Std Dev.	0.0213	0.0216	0.0219	0.0220	0.0228	0.0235
	0.55	0.50	0.45	0.40	0.35	0.30
Mean $\widehat{\delta}$	0.1242	0.1246	0.1257	0.1240	0.1305	0.1331
Std Dev.	0.0261	0.0281	0.0307	0.0403	0.0509	0.0609

Table 2.18: Simulation Results, $n = 300$. True value of $\delta = 0.1244$. No corrections to cell counts. A model with independent tree effects was fit.

estimating δ. We also investigated the approach of Evans and Bonett (1994a) and added a small constant, $\frac{2}{17}$, to each cell before model fitting. The denominator of this amount was chosen based on 17 cells in the three complete contingency tables. After model fitting, we adjusted the estimated number of fires by the actual amount added to the 11 observed cells, or $\frac{22}{17}$.

From Table 2.19, we find that the sign of the difference between the estimate and the known population size changed as the recording probability increased. This result is similar to the findings of Evans and Bonett (1994) when very low capture probabilities were considered. The results of Table 2.19 indicate that the bias reduction approach of Evans and Bonett (1994a) appears to be appropriate.

2.5 Discussion

Hook and Regal (1993) demonstrated that capture-recapture population estimation could be improved by using stratified lists when available. This paper builds upon their approach by generalizing to multiple overlapping lists even when some lists may not be operating in all strata. In these situations, the observed likelihood can be evaluated, but it may be complex and cannot be maximized with traditional log-linear software. The EM algorithm may be preferred in these situations because it allows standard model fitting software to be used for the M-step and the E-step is relatively straightforward.

	Event Recording Probability of Each Tree				
	0.95	0.90	0.85	0.80	0.75
Mean $\widehat{\delta}$	0.1225	0.1229	0.1232	0.1235	0.1243
Std Dev.	0.0213	0.0216	0.0219	0.0220	0.0228
	0.50	0.45	0.40	0.35	0.30
Mean $\widehat{\delta}$	0.1260	0.1268	0.1273	0.1295	0.1299
Std Dev.	0.0280	0.0302	0.0362	0.0446	0.0540
	0.25	0.20	0.15	0.10	0.05
Mean $\widehat{\delta}$	0.1290	0.1312	0.1265	0.1009	0.0452
Std Dev.	0.0656	0.0810	0.0877	0.0755	0.0349

Table 2.19: Simulation Results, $n = 300$. True value of $\delta = 0.1244$. $\frac{2}{17}$ added to each cell before model fitting. A model with independent tree effects was fit.

In our methodology, we assumed that there were no lost tags or missed matches, both potential sources of bias. Our model may not work well in this situation, though the Petersen and log-linear model are also expected to perform poorly. Existing methods for tag mismatch and tag loss could be extended to partially stratified lists.

The population estimation technique in this paper benefits from "sharing" information among strata and will be best applied when there are multiple overlapping lists and when the underlying model is common over strata.

The methodology may not be an improvement if there are 'hidden' list dependencies on unobserved lists since a key assumption is that parameters that are not estimable in a particular strata are equal across strata. In this case, all methods will produce poor estimates of overall population size. Also, the method will not work well when the underlying relationship among lists are quite different in the different strata.

Chapter 3

Population Estimation with Incomplete Lists

Multi-list methods have become a common application of capture-recapture methodology to estimate the size of human populations, having been successfully applied to estimating prevalence of diabetes, human immunodeficiency virus (HIV) and drug abuse. A key assumption in multi-list methods is that individuals have a unique "tag" that allows them to be matched across all lists. In some cases, this may not be true. For example, a subset of the lists may use health insurance number to cross-match, while another subset of lists may use date of birth, while only a few lists may have both keys. This paper develops multi-list methodology that relaxes the assumption of a single tag common to all lists.

There are other capture-recapture methods that address difficulties matching individuals across lists; these include corrections for tag loss and adjustments for tag mismatches across lists. Tag loss adjustments were discussed by Seber (1982) and Seber and Felton (1981) and more recently tag errors by Schwarz and Stobo (1999a). Huakau (2001), Lee (2002) and Lee et al. (2001) have advanced methods in the area of tag mismatches. Our methodology assumes that there is no tag loss and that no errors occur in matching, but does not assume that all tags are available on each list.

Similar to many capture-recapture methods reviewed by Schwarz and Seber (1999b), our approach assumes that the population is closed. We also assume that each tag is

sufficient to identify an individual.

Under these assumptions, the estimates are found using estimating functions. An example illustrates the application of estimating functions to estimating the prevalence of diabetes. The results of the model fitting are compared against other analyses of the same data. The last section is a simulation that investigates the performance of the methodology.

3.1 Notation

3.1.1 Parameters

Let

$$\begin{aligned} N &= \text{Population size;} \\ \beta_0 &= \text{Intercept;} \\ \beta_k &= \text{Effect of List } k,\ k = 1, \ldots, K; \\ \beta_{jk} &= \text{Interaction effect between List } j \text{ and List } k. \end{aligned}$$

The notation above is extended to higher order interactions in a similar fashion. The vector of parameters is written $\mathbf{\Phi}$.

3.1.2 Statistics

To help illustrate the definition and construction of the statistics, it will be helpful to consider the small example of four lists in Table 3.1. Although list 3 has both Tag A and Tag B, lists 1 and 2 have Tag A only, and list 4 has Tag B only.

Given the structure of available tags in Table 3.1, a population member could be matched across all lists; matching Tag A on lists 1, 2 and 3, and matching Tag B on lists 3 and 4. It is not possible to match individuals on lists 1 and 4 who are not also present on list 3.

Multi-list methods commonly begin with the $2^K - 1$ contingency table containing the observed counts of individuals on various combinations of lists. Let the vector

List	Tags Available for Matching	
List 1	Tag A	
List 2	Tag A	
List 3	Tag A	Tag B
List 4		Tag B

Table 3.1: Four list, two tag example.

$\omega = \{\omega_1, \ldots, \omega_K\}$ represent the "capture-history" of each individual. We define ω_k as

$$\omega_k = \begin{cases} 1 & \text{if individual } i \text{ is known to be present on List } k; \\ 0 & \text{if individual } i \text{ is known not to be present on List } k; \\ \cdot & \text{unknown whether individual } i \text{ is present/absent on List } k. \end{cases}$$

The count of individuals with the same capture history ω is the statistic Y_ω. For example, Y_{1111} is the count of individuals present on lists 1, 2, 3 and 4.

However, not all history vectors can be observed. In Table 3.1, the statistics Y_{1101} and Y_{1100} are not observable because records on lists 1 and 2 cannot be matched with list 4 without list 3 acting as a "link". Let the vector $\mathbf{Y}$ of length L represent the observable statistics.

The set of observable statistics is sometimes difficult to determine. The process is facilitated by drawing a graph of the list and tag structure. Vertices (lists) are connected by an edge if they share a tag. An 'internal' vertex is one that is connected to more than one vertex, while a 'leaf' is a vertex that is connected to a single vertex. The graph of Table 3.1 is illustrated in Figure 3.1.

The observable capture histories are found by letting the components of ω corresponding to internal vertices take the value 0 or 1 (but not all 0, which corresponds to unobserved on any list). Then, for each set of labels applied to the internal vertices, each component of ω corresponding to a list that is a leaf may take the value of 0 or 1 if the vertex to which it is attached has the value 1. If an internal vertex has the value 0, the component of ω corresponding to a single leaf may take the value 0 without affecting the value of other vertices, or take the value 1 but then all other components of ω corresponding to all other vertices in the graph must simultaneously

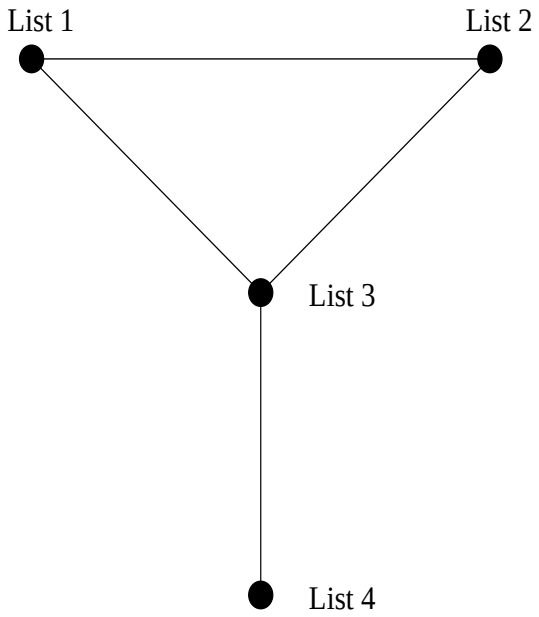

Figure 3.1: Graphical representation of Table 3.1. Four list, two tag example.

take the value '·'. All observable histories must have a 1 on at least one list.

For example, from the graph in Figure 3.1, components of ω corresponding to lists 1, 2 and 3 may take all combinations of $(0, 1)$, but not all can be 0 simultaneously. Whenever the component of ω for list (vertex) 3 takes the value 1, then the component of ω for list 4 may take the value of 0 or 1; if the component of ω for list 3 takes the value 0, then the component of ω for list 4 can only take the value 1 or '·'.

Using the preceding algorithm, there are 12 observable statistics in our example comprising $\mathbf{Y}$. They are $\omega \in \{\{100\cdot\}, \{010\cdot\}, \{110\cdot\}, \{1010\}, \{1011\}, \{0110\}, \{0111\}, \{0010\}, \{0011\}, \{1110\}, \{1111\}, \{\cdot\cdot 01\}\}$.

3.2 Model Development

In a closed, multi-list setting, population size is often estimated using the well developed log-linear modelling framework (Feinberg, 1972 and Cormack, 1989). In a two-list setting, the simple Petersen estimator is used.

In our example, we would be able to apply a log-linear model only to counts derived from capture histories based on matches across lists 1, 2 and 3 using Tag A. However, the log-linear model ignores information provided by unmatched records on List 4 with Tag B, statistic $Y_{\cdot\cdot 01}$.

Alternatively, the Petersen estimate could be formed based on counts from lists 3 and 4 using Tag B. In this example, $n_{List\ 3} = Y_{1111} + Y_{1110} + Y_{0111} + Y_{1011} + Y_{1010} + Y_{0110} + Y_{0011} + Y_{0010}$, while $n_{List\ 4} = Y_{1111} + Y_{0111} + Y_{0111} + Y_{0011} + Y_{\cdot\cdot 01}$ and $m_{Lists\ 3\ and\ 4} = Y_{1111} + Y_{0111} + Y_{1011} + Y_{0011}$, but this ignores information provided by unmatched records on List 1, 2 and 3, statistics $Y_{100\cdot}, Y_{010\cdot}$ and $Y_{110\cdot}$.

Our model starts with a vector $\mathbf{Z}$ of underlying 'complete' counts, representing the (unobservable) counts derived as though all 'tags' were available on each list. In our example, lists 1, 2, 3 and 4 would have had both Tag A and B. The same notation is used to denote capture histories for statistics Z_ω.

As is the case with log-linear models, assume that

$$\begin{aligned} \mathbf{Z} &\sim \text{Poisson}(\mu_{\mathbf{Z}}) \\ \log(\mu_{\mathbf{Z}}) &= \mathbf{X}\mathbf{\Phi}, \end{aligned}$$

where the design matrix $\mathbf{X}$ corresponds to the full $2^K - 1$ cells counts, $\mathbf{Z}$, and the vector of parameters, $\mathbf{\Phi}$, represent list and list interaction effects.

The observable statistics $\mathbf{Y}$ can be related to the vector $\mathbf{Z}$ through a matrix $\mathbf{T}$,

$$\mathbf{Y} = \mathbf{T}\mathbf{Z},$$

where $\mathbf{T}$ is defined as an $L \times \left(2^K - 1\right)$ matrix of indicator variables representing linear combinations of elements of $\mathbf{Z}$.

The matrix $\mathbf{T}$ is derived from the history vector of the statistics $\mathbf{Y}$. First, note that a statistic Y_ω, $\omega_k = \cdot$ for any k, is the sum of some Z_ω, Z_ω determined by replacing the unobservable component(s) of ω with 0 and 1. For example, $Y_{110\cdot} = Z_{1101} + Z_{1100}$.

Next, sort elements of $\mathbf{Z}$ by the 'decimal equivalent' of the binary history vector ω, *e.g.* Z_{1101} would be in the 13^{th} position. The row of $\mathbf{T}$ corresponding to Y_ω has non-zero entries in columns corresponding to the decimal equivalent of ω.

For example, for the statistic Y_{0010}, the corresponding row of $\mathbf{T}$ will have a 1 in column 2. In the case of $Y_{\cdot\cdot 01}$, we note that $Y_{\cdot\cdot 01} = Z_{1101} + Z_{1001} + Z_{0101} + Z_{0001}$. Then, non-zero entries of the row of $\mathbf{T}$ corresponding to $Y_{\cdot\cdot 01}$ are in columns 13, 9, 5 and 1.

In our example from Table 3.1, $\mathbf{Y} = \mathbf{TZ}$ is written

$$\begin{bmatrix} Y_{1111} \\ Y_{1110} \\ Y_{1011} \\ Y_{0111} \\ Y_{110\cdot} \\ Y_{1010} \\ Y_{0110} \\ Y_{0011} \\ Y_{100\cdot} \\ Y_{010\cdot} \\ Y_{0010} \\ Y_{\cdot\cdot 01} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} Z_{0001} \\ Z_{0010} \\ Z_{0011} \\ Z_{0100} \\ Z_{0101} \\ Z_{0110} \\ Z_{0111} \\ Z_{1000} \\ Z_{1001} \\ Z_{1010} \\ Z_{1011} \\ Z_{1100} \\ Z_{1101} \\ Z_{1110} \\ Z_{1111} \end{bmatrix}.$$

A likelihood-based approach would be difficult to develop because of the possibility of double-counting individuals on lists where there is incomplete information. For example, the count Z_{1101} appears in both $Y_{110\cdot}$ and $Y_{\cdot\cdot 01}$, as $Y_{110\cdot} = Z_{1101} + Z_{1100}$ and $Y_{\cdot\cdot 01} = Z_{1101} + Z_{1001} + Z_{0101} + Z_{0001}$. We propose employing a system of estimating functions (Godambe, 1960) to estimate the population size.

Estimates of the vector of parameters are obtained as numerical solutions of

$$\mathbf{D}'\mathbf{V}^{-1}(\mathbf{Y} - \mathrm{E}[\mathbf{Y}]) = \mathbf{0},$$

where $\mathrm{E}[\mathbf{Y}]$ is the vector of first moments derived from $\mathrm{E}[\mathbf{Y}] = \mathbf{T}\mathrm{E}[\mathbf{Z}] = \mathbf{T}\mu_{\mathbf{Z}}$.

$\mathbf{D}$ is a rectangular matrix of first partial derivatives of $\mu_{\mathbf{Y}}$ with respect to the parameter set (i.e. $\mathbf{D} = \mathbf{T}\frac{\partial \mu_{\mathbf{Z}}}{\partial \mathbf{\Phi}}$), while $\mathbf{V}$ is the working covariance matrix of $\mathbf{Y}$. This

is equivalent to treating the statistics of $\mathbf{Y}$ as independent Poisson random variables for the working covariance. Parameter estimates are known to be consistent estimators even if $\mathbf{V}$ is not the true covariance matrix of $\mathbf{Y}$ (Liang and Zeger, 1986).

The covariance of $\widehat{\mathbf{\Phi}}$ is consistently estimated by

$$\widehat{Var}\left(\widehat{\mathbf{\Phi}}\right) = \left[\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right]^{-1}\left[\mathbf{D}'\mathbf{V}^{-1}\widehat{V}(\mathbf{Y})\mathbf{V}^{-1}\mathbf{D}\right]\left[\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right]^{-1},$$

where $\mathbf{D}$ and $\mathbf{V}$ are previously defined. $V(\mathbf{Y})$ is based on the "true" distribution of the statistics of $\mathbf{Y}$.

There are several approaches to estimating $Var\left(\widehat{\mathbf{\Phi}}\right)$. The model-based covariance estimator assumes that $V(\mathbf{Y}) = \mathbf{V}$, and $Var\left(\widehat{\mathbf{\Phi}}\right)$ simplifies to $\left[\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right]^{-1}$. The model-based approach to estimating the covariance will likely mis-specify the form of the variance of $\mathbf{Y}$. If the statistics are correlated the covariance estimates will not reflect the true $Var\left(\widehat{\mathbf{\Phi}}\right)$. However, it's simple application makes it an attractive alternative estimator.

The empirical, or robust covariance estimator (Liang and Zeger, 1986) estimates the covariance $\widehat{V}(\mathbf{Y})$ using cross-products of residuals for each record, $\left[\mathbf{Y}_i - \mu(\widehat{\mathbf{\Phi}})\right]\left[\mathbf{Y}_i - \mu(\widehat{\mathbf{\Phi}})\right]^T$, where $\mathbf{Y}_i$ is a vector of indicator variables for each observed individual and $\mu(\widehat{\mathbf{\Phi}})$ is the expectation. The expression $\mathbf{D}'\mathbf{V}^{-1}\widehat{V}(\mathbf{Y})\mathbf{V}^{-1}\mathbf{D}$ is then summed over all observed records.

The non-parametric bootstrap standard error (Efron and Tibshirani, 1993) can also be used to estimate $Var\left(\widehat{\mathbf{\Phi}}\right)$. To establish a bootstrap sample, $\mathbf{Y}^*$, we re-sample records to create a 'new' sample of the same size as the original data. For each $\mathbf{Y}^*$, we estimate the parameters $\widehat{\mathbf{\Phi}}^*$ using the methodology described.

Many models can be fit with the log-linear framework provided that the number of parameters determining $\mathbf{\Phi}$ does not exceed L. For instance, we can model equal capture probabilities for all lists ($\mathbf{\Phi} = \{\beta_1 = \beta_2 = \beta_3 = \beta_4\}$), corresponding to M_0 (reviewed by Chao, 2001). Also, we can fit equal second order interaction effects, $\{\beta_{12} = \beta_{13} = \beta_{14} = \beta_{23} = \beta_{24} = \beta_{34}\}$, the quasi-symmetric model, and compare it to that of M_t, different capture probabilities for each list, $\mathbf{\Phi} = \{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4\}$.

The Akaike information criterion (AIC) and the deviance (G^2) are well established for log-linear model selection. Because we do not employ a likelihood function, we use

alternative approaches. A Pearson statistic is asymptotically chi-square, although as Cormack (1989) notes, capture history data is often sparse and applying the statistic may not be appropriate if this is the case. We also compute the QIC statistic (Pan, 2001), a statistic analogous to the AIC for estimating functions that adjusts for the number of parameters in the model.

We can relate list effect parameters to capture probabilities if there are no list interaction effects as outlined by Cormack (1989). For example, in the four-list example of Table 3.1, we equate

$$\begin{aligned} e^{\beta_0+\beta_1+\beta_2+\beta_3+\beta_4} &= Np_1p_2p_3p_4 \\ e^{\beta_0+\beta_1+\beta_2+\beta_3} &= Np_1p_2p_3(1-p_4) \\ e^{\beta_0+\beta_1+\beta_3+\beta_4} &= Np_1(1-p_2)p_3p_4 \\ e^{\beta_0+\beta_2+\beta_3+\beta_4} &= N(1-p_1)p_2p_3p_4 \\ e^{\beta_0+\beta_3+\beta_4} &= N(1-p_1)(1-p_2)p_3p_4, \end{aligned}$$

and then solve to yield $\widehat{p}_4 = \frac{1}{1+e^{-\widehat{\beta}_4}}$ and $\widehat{N} = \frac{e^{\widehat{\beta}_0}}{(1-\widehat{p}_1)(1-\widehat{p}_2)(1-\widehat{p}_3)(1-\widehat{p}_4)}$.

If there are only list effects, the properties of the multinomial distribution can be used to evaluate $V(\mathbf{Y})$ at $\widehat{\mathbf{\Phi}}$. Using our example, to determine $var(Y_{110\cdot})$, we write

$$\begin{aligned} var(Y_{110\cdot}) &= var(Y_{1101} + Y_{1100}) \\ &= N \times p_1p_2(1-p_3)p_4 \times (1 - p_1p_2(1-p_3)p_4) \\ &\quad +N \times p_1p_2(1-p_3)(1-p_4) \times (1 - p_1p_2(1-p_3)(1-p_4)) \\ &\quad -2 \times N \times p_1p_2(1-p_3)p_4 \times p_1p_2(1-p_3)(1-p_4), \end{aligned}$$

while covariance terms can be factored in a similar manner. In situations where there are several lists which give rise to statistics with undetermined capture histories, such as $Y_{\cdot\cdot 01}$, it quickly becomes laborious to define all terms in this manner.

3.3 Example

We used as an example the data available from the Auckland Diabetes Study data. Huakau (2001) describes the characteristics of the lists whose information was to used

List	Tags Available for Matching		
List G	Tag A	Tag B	Tag C
List P	Tag A		
List O		Tag B	
List D			Tag C

Table 3.2: Four list example of simulated Diabetes data.

to estimate the prevalence of diabetes. In this example, four lists were available; 1,276 general practioners records (List G), 1,297 pharmacy records (List P), 12,792 outpatient records (List O) and 3,436 inpatient discharge records (List D). In the analyses of Seber, Huakau and Simmons (2001), Lee (2001 and 2002) and Huakau (2001), five 'tags' were split such that Tag A consisted of first name, surname and age, while Tag B consisted of date of birth, gender and address.

Since Huakau (2001) wrote "retention probabilities are high for tag A but low for tag B," we proceeded using the information provided by matching on Tag A only. This step is consistent with our assumption of a closed population and matching without error.

To develop an example for our methodology, we simulated a new list structure whereby Lists P, O and D do not share a tag, but each shares one with List G as shown in Table 3.2. Figure 3.2 shows the graph of the list structure in Table 3.2.

The set of observable statistics was derived by enumerating observable capture histories, $\omega = \{$ List G, List P, List O, List D $\}$. Note that list G is an internal vertex, while lists P, O and D are leaves. First, consider the case of the internal vertex equal to 1. This leads to the observable statistics $\omega = \{1, x, y, z\}; x, y, z \in \{0, 1\}$. Secondly, let the internal vertex equal 0 and each leaf, in turn, is set equal to 1, leading to the observable histories $\omega = \{0, 1, \cdot, \cdot\}, \{0, \cdot, 1, \cdot\}, \{0, \cdot, \cdot, 1\}$. There are 11 observable capture histories.

To establish our simulated dataset from the Diabetes Study, we aggregated over appropriate cells of $\mathbf{Z}$. For instance, $Y_{01\cdot\cdot} = Z_{0111} + Z_{0110} + Z_{0101} + Z_{0100}$. The set

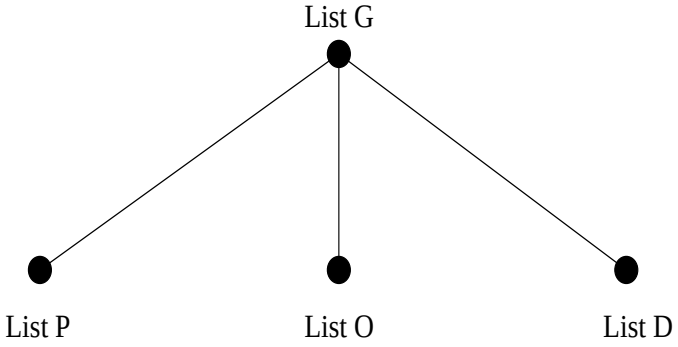

Figure 3.2: Graphical representation of lists of Auckland Diabetes Data.

of observable statistics, $\mathbf{Y}$, is shown in Table 3.3. For this reduced list structure, we could not use a log-linear model on all 4 lists, nor could we could use a log-linear model on any subset of 3 lists. The approximately unbiased Petersen estimates formed using list pairs G-P, G-O and G-D are shown in Table 3.4. The Petersen estimator provided widely varying point estimates of the population size.

We applied the methodology developed above. The results of our model fitting are in Table 3.5. We selected model $\mathbf{\Phi}^{(13)}$ because its QIC statistic is the smallest, although there is little difference between $\mathbf{\Phi}^{(12)}$ and $\mathbf{\Phi}^{(13)}$. At this point, a model averaging approach, such as discussed by Hook and Regal (1997) and Buckland et al. (1997), could be used to account for model uncertainty in the estimate of N.

The estimated population size was $\widehat{N}_{\mathbf{\Phi}^{(13)}} = 45,853$. This estimate was substantially larger than the pairwise Petersen estimates. We calculated $\widehat{V}(Y)$ using three methods. The model-based, the robust estimator and the bootstrap estimate of the standard error were substantially larger than those from the Petersen estimates.

Because all $2^4 - 1$ cells of the contingency table were available, we fit log-linear models to the counts of $\mathbf{Z}$ for comparison (Table 3.4). Using the same model as $\mathbf{\Phi}^{(13)}$, the population estimate based on the complete data was 57,179, which is substantially larger than our estimate.

In several models, $\widehat{N}_{\mathbf{\Phi}}$ differed considerably from the equivalent log-linear model fit to all 15 cells, as shown in Table 3.5. There appeared to be interaction effects

Z	Count	**Y**	Count
Z_{0001}	1,685	$Y_{01\cdot\cdot} = Z_{0111} + Z_{0110} + Z_{0101} + Z_{0100}$	1,183
Z_{0010}	10,393	$Y_{0\cdot1\cdot} = Z_{0111} + Z_{0110} + Z_{0011} + Z_{0010}$	12,265
Z_{0011}	1,450	$Y_{0\cdot\cdot1} = Z_{0111} + Z_{0101} + Z_{0011} + Z_{0001}$	3,276
Z_{0100}	713	Y_{1000}	654
Z_{0101}	48	Y_{1001}	51
Z_{0110}	329	Y_{1010}	366
Z_{0111}	93	Y_{1011}	91
Z_{1000}	654	Y_{1100}	40
Z_{1001}	51	Y_{1101}	4
Z_{1010}	366	Y_{1110}	56
Z_{1011}	91	Y_{1111}	14
Z_{1100}	40		
Z_{1101}	4		
Z_{1110}	56		
Z_{1111}	14		

Table 3.3: Statistics for the 4-list Diabetes example

Estimation Method	$\widehat{N}$	$\widehat{se}\left(\widehat{N}\right)$
Approx. Unbiased Petersen		
List G and P	14,412	1,201
List G and O	30,940	1,007
List G and D	27,260	1,935
Estimating Functions, model $\mathbf{\Phi}^{(13)}$	45,853	
Model Based		4,530
Robust		4,343
Bootstrap		4,203
All Information Available		
$\mathbf{\Phi}^{(13)}$, $2^4 - 1$ log-linear model, both tags	57,179	3,459

Table 3.4: Results of estimating population size of the Auckland Diabetes Study.

Model	QIC	Pearson Statistic	$\widehat{N}_{\mathbf{\Phi}}$	$\widehat{N}_{ll}$
$\mathbf{\Phi}^{(1)} = [GP = GO = GD = PO = PD = OD][GPO][GPD][POD]$	115.8	145.9	63,986	44,604
$\mathbf{\Phi}^{(2)} = [GP = GO = GD = PO = PD = OD][GPO = GPD = POD]$	33.0	37.2	29,081	44,142
$\mathbf{\Phi}^{(3)} = [GP = GO = GD = PO = PD = OD]$	33.0	37.2	61,032	46,831
$\mathbf{\Phi}^{(4)} = [GP][GO][GD][PO = PD = OD]$	10.7	10.2	80,812	45,934
$\mathbf{\Phi}^{(5)} = [GP][GO][GD]$	64.9	66.7	38,082	31,547
$\mathbf{\Phi}^{(6)} = [GP][GO = GD]$	64.9	66.7	43,854	31,615
$\mathbf{\Phi}^{(7)} = [P][GO = GD]$	126.8	101.8	25,070	31,212
$\mathbf{\Phi}^{(8)} = [D][GP][GO]$	64.9	66.7	29,574	31,460
$\mathbf{\Phi}^{(9)} = [O][GP][GD]$	64.9	66.7	32,891	31,805
$\mathbf{\Phi}^{(10)} = [G][P][OD]$	75.4	100.9	31,680	34,979
$\mathbf{\Phi}^{(11)} = [O][D][GP]$	67.1	72.7	32,332	31,724
$\mathbf{\Phi}^{(12)} = [GP][GO = GD = PO = PD][OD]$	7.1	6.9	46,396	55,455
$\mathbf{\Phi}^{(13)} = [GP = OD][GO = GD = PO = PD]$	7.0	7.2	45,853	57,179
$\mathbf{\Phi}^{(14)} = [G][P][O][D]$	118.7	150.4	30,912	31,396
$\mathbf{\Phi}^{(15)} = [G = P][O][D]$	118.9	150.2	30,918	31,396
$\mathbf{\Phi}^{(16)} = [GPD][O]$	-*	-*	-*	64,055

Table 3.5: Selected results of model fitting to Auckland Diabetes Study data. *Unable to fit model because the number of parameters exceeds the number of data points.

between Lists P, O and D which our model cannot detect because no observable statistics captured this interaction. The log-linear model fit to the entire dataset was able to detect the effects.

Huakau (2001) showed, using log-linear modelling, that the best fitting model to the full dataset was [GPD][O], from which $\widehat{N}_{log-linear[GPD][O]} = 64,055$. We could not fit this model to the statistic $\mathbf{Y}$ because the number of parameters exceeded L.

3.4 Simulation

We investigated whether our proposed method using estimating functions produced reasonable estimates in more general settings than our example. We simulated two list structures, one with 4 lists and one with 3 lists.

3.4.1 List Independence

In our first study we had four lists available; on lists 1 and 2, only Tag A was available, on list 3 Tag A and B were both available, while on list 4, only Tag B was available (Figure 1). For each list, records were sampled without replacement from a population of fixed size with equal probability of capture on each list, *i.e.* the true model was $\beta = \{N, p_1 = p_2 = p_3 = p_4\}$. This procedure was repeated 1,000 times. The results are shown in Table 3.6.

The model fit to the count data was $\beta = \{N, p_1, p_2, p_3, p_4\}$, which accommodated different capture probabilities across lists. The population estimates were compared to the approximately unbiased Petersen and log-linear model (main effects model was fit, consistent with the method of sampling). A $2^4 - 1$ log-linear model was also fit as though all tags were available on all lists. This estimate served as a 'gold-standard' against which the other estimates were compared.

In the first study (Table 3.6), we drew four lists of equal size from a population of 60,000. Not surprisingly, the approximately unbiased Petersen estimate based on Tag B on lists 3 and 4 has worse precision in comparison to the log-linear model where three lists were available (Tag A on lists 1, 2 and 3). The mean standard error from

	List 1 = List 2 = List 3		
	2,000	4,000	6,000
Unbiased Petersen (List 3 and 4)			
Mean $\widehat{N}$	59,311	59,732	59,942
Mean $\widehat{se}(\widehat{N})$	6885	3394	2198
Log-Linear Model (Lists 1, 2 and 3)			
Mean $\widehat{N}$	60,258	59,925	60,019
Std dev $\widehat{N}$	4152	2033	1270
Estimating Functions (all lists)			
Mean $\widehat{N}$	60,147	59,926	60,023
Mean $\widehat{se}$ (model based)	3560	1708	1106
Mean $\widehat{se}$ (robust covariance)	3542	1677	1061
All Information Available ($2^4 - 1$ log-linear model, both tags)			
Mean $\widehat{N}$	60,086	59,944	60,001
Std dev $\widehat{N}$	2850	1354	861

Table 3.6: Population estimates, 1,000 simulations of 4 list, 2-tag example, equal list size. Theoretical population, $N = 60{,}000$. Model fit, $\mathbf{\Phi} = \{N, p_1, p_2, p_3, p_4\}$, which corresponds to model used to generate data.

List 1 =1,000, List 2 = 3,000, List 3=4,000, List 4 = 2,000			
	$N = 40{,}000$	$N = 60{,}000$	$N = 80{,}000$
Unbiased Petersen (List 3 and 4)			
Mean $\widehat{N}$	39,906	59,780	79,446
Mean $\widehat{se}(\widehat{N})$	2595	4876	7551
Log-Linear Model (Lists 1, 2 and 3)			
Mean $\widehat{N}$	40,083	60,248	80,277
Std dev $\widehat{N}$	1682	3175	5104
Estimating Functions (all lists)			
Mean $\widehat{N}$	40,077	60,186	80,185
Mean $\widehat{se}(\widehat{N})$ (model based)	1423	2696	4202
Mean $\widehat{se}(\widehat{N})$ (robust covariance)	1428	2675	4184
All Information Available ($2^4 - 1$ log-linear model, both tags)			
Mean $\widehat{N}$	40,061	60,137	80,172
Std dev $\widehat{N}$	1280	2317	3574

Table 3.7: Population estimates, 1,000 simulations of list size, 4 list, 2-tag example, unequal list size. Model fit, $\mathbf{\Phi} = \{N, p_1, p_2, p_3, p_4\}$, which corresponds to model used to generate data.

our method lies between those from estimates based on subsets of the lists and those derived from the "full" table.

In the second study (Table 3.7), we varied the list size. Again, estimates are unbiased in all methods. Precision of estimates is poorer than if the complete data were available, but our method is more precise than either unbiased Petersen or the log-linear model.

In the third study (Table 3.8), we simulated a different three lists structure. List 1 had only Tag A, list 2 had Tag A and Tag B, while list 3 had Tag B only. The unbiased Petersen estimate was computed for lists 1 and 2 based on Tag A, and a second unbiased Petersen estimate was computed for lists 2 and 3 based on Tag B. For comparison, we computed the population estimate when all information was present (Tag A and B for all 3 lists) using a $2^3 - 1$ log-linear model. Our findings are similar to the previous results.

List 1 =1,000, List 2 = 4,000, List 3=6,000			
	$N = 40{,}000$	$N = 60{,}000$	$N = 80{,}000$
Unbiased Petersen (List 1 and 2)			
Mean $\widehat{N}$	39,732	59,422	78,755
Mean $\widehat{se}(\widehat{N})$	3675	6840	10495
Unbiased Petersen (List 2 and 3)			
Mean $\widehat{N}$	40,022	60,103	79,983
Mean $\widehat{se}(\widehat{N})$	1429	2754	4321
Estimating Functions (all lists)			
Mean $\widehat{N}$	40,082	60,245	80,239
Mean $\widehat{se}(\widehat{N})$ (model based)	1352	2589	4057
Mean $\widehat{se}(\widehat{N})$ (robust covariance)	1326	2568	4039
All Information Available ($2^3 - 1$ log-linear model, both tags)			
Mean $\widehat{N}$	40,076	60,170	80,219
Std dev $\widehat{N}$	1226	2420	3607

Table 3.8: Population estimates, 1,000 simulations of list size, 3 list, 2-tag example, unequal list size. Model fit, $\mathbf{\Phi} = \{N, p_1, p_2, p_3, p_4\}$, which corresponds to model used to generate data.

	$N = 49{,}404$	$N = 74{,}106$	$N = 98{,}809$
Unbiased Petersen (List 3 and 4)			
Mean $\widehat{N}$	57,851	87,438	116,536
Mean $\widehat{se}(\widehat{N})$	4431	5515	6378
Log-Linear Model (Lists 1, 2 and 3)			
Best model $\mathbf{\Phi}_{fit} = \{\beta_1, \beta_2, \beta_3, \beta_4, \beta_{12}\}$			
Mean $\widehat{N}$	51,501	77,128	103,233
Mean $\widehat{se}(\widehat{N})$	2897	3456	3848
Estimating Functions (all lists)			
Mean $\widehat{N}$	50,487	75,498	100,759
Mean $\widehat{se}(\widehat{N})$ (model based)	2282	2765	3184
Mean $\widehat{se}(\widehat{N})$ (robust covariance)	2264	2744	3161
All Information Available ($2^4 - 1$ log-linear model, both tags).			
Best model $\mathbf{\Phi}_{true} = \{\beta_1, \beta_2, \beta_3, \beta_4, \beta_{14}, \beta_{24}\}$			
Mean $\widehat{N}$	49,361	74,036	99,001
Std dev $\widehat{N}$	2060	2342	2789

Table 3.9: Population estimates, 1,000 simulations of list size, 4 list, 2-tag example, list dependence between lists 1 and 4 and between lists 2 and 4.

3.4.2 List Dependence

From the list structure of Figure 1, we simulated cell counts from a $2^4 - 1$ contingency table with list dependencies between list 1 and 4; and between list 2 and 4, *i.e.*, $\mathbf{\Phi} = \{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_{14}, \beta_{24}\}$. The best fitting log-linear model to our list structure was $\mathbf{\Phi}_{fit} = \{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_{12}\}$. Results are shown in Table 3.9.

The unbiased Petersen provided a poor fit to the data. This is not surprising because the estimate is based on list 3 and 4, and is not able to incorporate interaction effects between lists 1 and 4 and between lists 2 and 4 into the estimate *i.e.,* the 'true' model had effects β_{14}, β_{24} that could not be detected.

Although the log-linear model provided a better fit to the data than the Petersen, the estimating functions were an improved fit (approximately $2 - 4\%$). The estimated

standard errors of the estimating functions were more precise (20%) than the log-linear model. The robust covariance was negligibly more precise than the model based estimated standard error.

Dependencies between lists 1 and 4 and between lists 2 and 4 in the true model resulted in an interaction effect in the model fit to the reduced data *i.e.*, the fitted model included interaction between Lists 1 and 2. This interaction term fit to the model suggests that 'unobservable' list dependencies can result in interaction effects appearing in our models.

In general, we found that the population estimates derived from the estimating functions were more precise that those of the log-linear model or Petersen estimate derived on subsets of lists. This result was not surprising given that the estimating functions used more information that either other method on their own. The estimated standard error based on the robust covariance estimator was consistently smaller than that which used the model based estimator, but the difference was very small ($< 2\%$) in the cases examined.

3.5 Discussion

This paper developed a multi-list method which incorporates lists when not all lists are required to have the same tags. This situation is likely to appear in epidemiological settings when lists being used for matching only have subsets of tags available. An example demonstrates that it is easily applied to real data.

A simulation study showed that it has consistently better precision than log-linear models or the simple Petersen on subsets of lists. However, not surprisingly, the methodology is less precise than if all data were available.

The simulation illustrated a danger with multi-list modelling; missing lists can lead to 'misleading' models as a result of interaction effects not being properly detected. If this happens, population estimates can be severely compromised. This is evident with the Petersen estimator, where list interaction effects cannot be modelled, although this is also true for the log-linear estimate and the new estimator.

At first glance, the EM algorithm (Dempster et al, 1977) would appear to be

another method that could be suitable for this problem. Using this approach, a likelihood could be developed assuming Poisson counts for $\mathbf{Z}$. Then, Z_ω with missing values would be estimated by conditioning on the observed count Y_ω. Calculating the expectations of the missing Z_ω is difficult because the unobserved Z_ω appear in more than one observed statistic. For example, from Table 1, in the expectation step, Z_{1101} appears in $Y_{110\cdot}$ and $Y_{\cdot\cdot 01}$, and $Z_{1101}|Y_{110\cdot} \sim Bin(Y_{110\cdot}, \frac{Z_{1101}}{Z_{1101}+Z_{1100}})$ and $Z_{1101}|Y_{\cdot\cdot 01} \sim Bin(Y_{\cdot\cdot 01}, \frac{Z_{1101}}{Z_{1101}+Z_{1001}+Z_{0101}+Z_{0001}})$. It is unclear how to implement the E-step under these constraints.

In the development of our model, we assumed that we were able to match records across lists without error. Violation of this assumption may positively bias population estimates. We also assumed that there were no missing tags, so we did not correct for tag loss (Seber, 1982, Seber and Felton, 1981), transcription errors or misread tags (Seber, Huakau and Simmons, 2000), each a potential source of bias.

The closed population assumption is common in multi-list methods. If lists are compiled over long periods, this assumption may not be viable; leading to biased estimates of population size. In this case, methods for open populations should be investigated.

Chapter 4

Estimation with Unmatchable List Members

This paper presents a methodology to estimate the population size for capture-recapture/multi-list methods when not all subjects can be matched across lists because of missing tags. These methods do not make the key assumption of existing methods, namely that sufficient information is present for all members to match across lists. This problem is motivated by a problem in an urban setting where administrative lists contain birthdates and initials which are used to match individuals across lists, but for some records, not all fields are present.

Capture-recapture/multi-list methods are applied to epidemiological population estimation problems. IWGDMF (1995a; 1995b) and Schwarz and Seber (1999) provide a comprehensive review. In the estimation of disease prevalence, lists are often compiled from administrative databases. Each individual in the population is identifed by a sequence of "tags" where, for example, the first "tag" is last name, and the second "tag" is address. These tags are used to match individuals across lists. In the two-list case, the Petersen estimator is used to estimate population size. In multi-list cases, log-linear models are used to model contingency table counts with one unobserveable cell.

If records have missing tags (tag loss), estimates of population size may be substantially biased (Seber and Felton, 1981; Seber, 1982; Seber, Huakau and Simmons,

2000). Seber and Felton (1981) and Seber (1982) suggest double tagging subjects to reduce bias due to tag loss. Seber and Felton (1981) and Seber (1982) provide an adjustment for population estimation methods that incorporates information based on the prevalence of missing tags. Adjustments made to allow for tag loss are known as 'tag-loss' models. The method developed assumes that all tags must be present to uniquely identify an individual

Seber, Huakau and Simmons (2000) estimate population size by considering the number of mismatched tags estimated from a subset of more comprehensive tags. Tag mismatch differs from tag loss in that tag mismatch refers to errors made matching across tags. Ding and Feinberg (1994) address tag mismatches by modeling matching probabilities using re-match information. Jaro (1989, 1995) develops a method of assigning weights to pairs of records to combat matching errors. Weights for each tag are assigned based on the probability of a match. Pairs of observations whose probability of matching exceeds a set threshhold are considered a match, while if the probability is below the threshhold, the match is rejected (Jaro, 1989; 1995).

The methods in this chapter use estimating functions to derive estimates of population size, capture probabilities and rates of tag loss. Different models can be built within the estimating functions framework. An example illustrates the methodology and how to select among competing models. A simulation study compares our estimator with the Petersen estimator and log-linear models based on complete data only.

4.1 Notation

4.1.1 Parameters

Define the following parameters:

$$\begin{aligned} N &= \text{Population size;} \\ p_i &= \text{Probability of being captured on List } i,\ i = 1, 2; \\ \theta_{A_i} &= \text{Probability of having tag } A \text{ present at time of capture on List } i,\ i = 1, 2; \\ \theta_{B_i} &= \text{Probability of having tag } B \text{ present at time of capture on List } i,\ i = 1, 2. \end{aligned}$$

4.1.2 Statistics

Each population member can be assigned a capture history determined by the presence or absence on the lists. This history is qualified by the presence or absence of tags at the time of each capture. Let Y be the count of records with the same capture history. For each statistic Y, let the subscript denote the tags present or absent at capture. For example,

$$Y = \begin{cases} Y_A & \text{Tag A present at capture;} \\ Y_{\cdot} & \text{Tag A not present at capture.} \end{cases}$$

If tags are not all present at capture, a complete capture history cannot be established. Let superscripts of each statistic identify the capture history, *i.e.* Y^{ω}. We write ω in vector form $\omega = (\omega_1, \ldots, \omega_l)$, where l is the number of lists. Each ω_i is defined as

$$\omega_i = \begin{cases} 1 & \text{captured on list } i; \\ 0 & \text{not captured on list } i; \\ \cdot & \text{undetermined whether captured on list } i. \end{cases}$$

The period indicates a capture with at least one missing tag, where a complete capture history cannot be established.

A two-list, two tag example will clarify notation. A record matching across two lists on both tags is a member of statistic Y_{AB}^{11}. A record observed on List 1 with only tag B present is a member of $Y_{\cdot B}^{1\cdot}$.

The vector $\mathbf{Y}$ is the complete set of statistics. For the two list example, $\mathbf{Y}$ has elements

$$
\begin{aligned}
Y_{AB}^{11} &= \text{Number of records with both tags present observed on both lists;} \\
Y_{AB}^{10} &= \text{Number of records with both tags present, observed on List 1 only;} \\
Y_{AB}^{01} &= \text{Number of records with both tags present, observed on List 2 only;} \\
Y_{A\cdot}^{1\cdot} &= \text{Number of records with only tag A present on List 1;} \\
Y_{A\cdot}^{\cdot 1} &= \text{Number of records with only tag A present on List 2;} \\
Y_{\cdot B}^{1\cdot} &= \text{Number of records with only tag B present on List 1;} \\
Y_{\cdot B}^{\cdot 1} &= \text{Number of records with only tag B present on List 2;} \\
Y_{\cdot\cdot}^{1\cdot} &= \text{Number of records with no tags present on List 1;} \\
Y_{\cdot\cdot}^{\cdot 1} &= \text{Number of records with no tags present on List 2;}
\end{aligned}
$$

It is often not possible to observe all tags on all records; administrative databases often have fields which are incomplete. However, we since we have assumed that all tags must be present to uniquely identify an individual. Consequently, records without both tags cannot be matched across lists and it is possible that an individual can be present on multiple lists without being fully matched across lists. For example, the statistics $Y_{A\cdot}^{11}$ and $Y_{A\cdot}^{10}$ cannot be observed.

To illustrate apportioning records with missing tags to statistics of $\mathbf{Y}$, consider a two list, two tag example.

$$
List\ 1 = \begin{bmatrix} R & 5 \\ C & 1 \\ R & - \\ M & 4 \\ - & - \\ - & 9 \end{bmatrix}, \quad List\ 2 = \begin{bmatrix} M & 4 \\ - & 1 \\ - & - \\ - & - \\ R & 2 \\ J & - \end{bmatrix},
$$

where "$-$" indicates a tag is not present. The statistic $\mathbf{Y}$ is:

$$\mathbf{Y} = \begin{bmatrix} Y_{AB}^{11} \\ Y_{AB}^{10} \\ Y_{AB}^{01} \\ Y_{A\cdot}^{1\cdot} \\ Y_{A\cdot}^{\cdot 1} \\ Y_{\cdot B}^{1\cdot} \\ Y_{\cdot B}^{\cdot 1} \\ Y_{\cdot\cdot}^{1\cdot} \\ Y_{\cdot\cdot}^{\cdot 1} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 2 \end{bmatrix}, \text{ corresponding to } \begin{bmatrix} & M & 4 & \\ R & 5\ , & C & 1 \\ & R & 2 & \\ & R & - & \\ & J & - & \\ & - & 9 & \\ & - & 1 & \\ & - & - & \\ - & -\ , & - & - \end{bmatrix}.$$

4.2 Model Development

In developing our model for the two-list case, we assume that the population is closed, lists are independent, individuals can only be observed once on any list, and observed tags are matched without error. In contrast to Seber, Huakau and Simmons (2000), we consider a single tag as being insufficient to uniquely identify an individual.

Based on these assumptions, methods for modelling tag mismatches (Ding and Feinberg, 1994 and Seber, Huakau and Simmons, 2000) are not directly applicable. Also, we cannot adapt tag loss methods developed by Seber and Felton (1981) or Seber (1982) because these methods assume each individual is given both tags at the time of first capture.

As noted by Seber, Huakau and Simmons (2000), there is often no time ordering of lists in epidemiological settings and they consider the closure assumption reasonable provided that lists are compiled in a reasonably short period of time.

We assume that tag loss is independent between individuals and tag type, although we allow tag loss to vary by tag type or list. If tag loss is not independent by tag type, we may observe more records without both tags. In this situation, Ding and Feinberg (1994) remark that not recognizing matches impacts population estimation by underestimating the number of individuals on both lists.

One option for estimating the population size is to use the Petersen estimator based

on records containing complete information only. Note that the Petersen estimate based on all records on both lists, *i.e.* $n_1 = Y_{AB}^{11} + Y_{AB}^{10} + Y_{A\cdot}^{1\cdot} + Y_{\cdot B}^{1\cdot} + Y_{\cdot\cdot}^{1\cdot}$ and $n_2 = Y_{AB}^{11} + Y_{AB}^{01} + Y_{A\cdot}^{\cdot 1} + Y_{\cdot B}^{\cdot 1} + Y_{\cdot\cdot}^{\cdot 1}$, and $m = Y_{AB}^{11}$, $\widehat{N}_{Petersen,\, all} = \frac{n_1 n_2}{m}$ is not consistent. However, the estimator derived from the subset of records with all tags observed,

$$\widehat{N}_{Petersen,\, complete} = \frac{(Y_{AB}^{11} + Y_{AB}^{10}) \times (Y_{AB}^{11} + Y_{AB}^{01})}{Y_{AB}^{11}},$$

is consistent.

The inability to rule out double counting some individuals makes it problematic to consider a multinomial-based likelihood as is done by Seber (1982) and Seber, Huakau and Simmons (2000), where they provide an adjustment factor to the denominator of the Petersen estimator. Seber and Felton (1981) use the same approach and present an analogous adjustment to the approximately unbiased Petersen estimator. We can consider a partial likelihood of $\mathbf{Y}$ by conditioning on the sum of records with both

tags present:

$$
\begin{aligned}
& f\left(Y_{AB}^{11}|Y_{AB}^{11}+Y_{AB}^{10}, Y_{AB}^{11}+Y_{AB}^{01}\right) \\
& \times f\left(Y_{A\cdot}^{1\cdot}, Y_{A\cdot}^{\cdot 1}, Y_{\cdot B}^{1\cdot}, Y_{\cdot B}^{\cdot 1}, Y_{\cdot\cdot}^{1\cdot}, Y_{\cdot\cdot}^{\cdot 1}|Y_{AB}^{11}, Y_{AB}^{11}+Y_{AB}^{10}, Y_{AB}^{11}+Y_{AB}^{01}\right) \\
= \; & f\left(Y_{AB}^{11}|Y_{AB}^{11}+Y_{AB}^{10}, Y_{AB}^{11}+Y_{AB}^{01}\right) \times f\left(Y_{A\cdot}^{1\cdot}, Y_{\cdot B}^{1\cdot}, Y_{\cdot\cdot}^{1\cdot}|Y_{AB}^{11}, Y_{AB}^{11}+Y_{AB}^{10}\right) \\
& \times f\left(Y_{A\cdot}^{\cdot 1}, Y_{\cdot B}^{\cdot 1}, Y_{\cdot\cdot}^{\cdot 1}|Y_{AB}^{11}, Y_{AB}^{11}+Y_{AB}^{01}\right) \\
= \; & \frac{\begin{pmatrix} Y_{AB}^{11}+Y_{AB}^{10} \\ Y_{AB}^{11} \end{pmatrix} \begin{pmatrix} N-Y_{AB}^{11}-Y_{AB}^{10} \\ Y_{AB}^{01} \end{pmatrix}}{\begin{pmatrix} N \\ Y_{AB}^{11}+Y_{AB}^{01} \end{pmatrix}} \\
& \times \begin{pmatrix} N-Y_{AB}^{11}-Y_{AB}^{10} \\ Y_{A\cdot}^{1\cdot}, Y_{\cdot B}^{1\cdot}, Y_{\cdot\cdot}^{1\cdot}, (N-Y_{AB}^{11}-Y_{AB}^{10}-Y_{A\cdot}^{1\cdot}-Y_{\cdot B}^{1\cdot}-Y_{\cdot\cdot}^{1\cdot}) \end{pmatrix} \left(p_1\theta_{A_1}(1-\theta_{B_1})\right)^{Y_{A\cdot}^{1\cdot}} \\
& \times \left(p_1\theta_{B_1}(1-\theta_{A_1})\right)^{Y_{\cdot B}^{1\cdot}} \left(p_1(1-\theta_{A_1})(1-\theta_{B_1})\right)^{Y_{\cdot\cdot}^{1\cdot}} \\
& \times \left(1-p_1(1-\theta_{A_1}\theta_{B_1})\right)^{N-Y_{AB}^{11}-Y_{AB}^{10}-Y_{A\cdot}^{1\cdot}-Y_{\cdot B}^{1\cdot}-Y_{\cdot\cdot}^{1\cdot}} \\
& \times \begin{pmatrix} N-Y_{AB}^{11}-Y_{AB}^{01} \\ Y_{A\cdot}^{\cdot 1}, Y_{\cdot B}^{\cdot 1}, Y_{\cdot\cdot}^{\cdot 1}, (N-Y_{AB}^{11}-Y_{AB}^{01}-Y_{A\cdot}^{\cdot 1}-Y_{\cdot B}^{\cdot 1}-Y_{\cdot\cdot}^{\cdot 1}) \end{pmatrix} \left(p_2\theta_{A_2}(1-\theta_{B_2})\right)^{Y_{A\cdot}^{\cdot 1}} \\
& \times \left(p_2\theta_{B_2}(1-\theta_{A_2})\right)^{Y_{A\cdot}^{\cdot 1}} \left(p_2(1-\theta_{A_2})(1-\theta_{B_2})\right)^{Y_{\cdot B}^{\cdot 1}} \\
& \times \left(1-p_2(1-\theta_{A_2}\theta_{B_2})\right)^{N-Y_{AB}^{11}-Y_{AB}^{01}-Y_{A\cdot}^{\cdot 1}-Y_{\cdot B}^{\cdot 1}-Y_{\cdot\cdot}^{\cdot 1}}
\end{aligned}
$$

We use the approach of Seber (1982) and model Y_{AB}^{11} using the hypergeometric probability distribution. We could use the binomial approximation with similar results.

Alternatively, one could use a system of estimating functions (Godambe, 1960; Liang and Zeger, 1986) based on the first moment of $\mathbf{Y}$. Estimating functions have been used in capture-recapture settings in different contexts for some time. Yip (1991) uses them to estimate population size in a closed population; Wang (1999) applies them to model growth parameters; and Schwarz, Andrews and Link (1999) apply them to the stratified Petersen estimator. In our application, estimates of the vector of parameters are obtained as solutions of

$$\mathbf{D}'\mathbf{V}^{-1}(\mathbf{Y}-\mu) = \mathbf{0},$$

Observed Statistic	Expected Value
Y_{AB}^{11}	$Np_1p_2\theta_A^2\theta_B^2$
Y_{AB}^{10}	$N\left[p_1\theta_A\theta_B(1-p_2)+p_1\theta_A\theta_Bp_2(1-\theta_A\theta_B)\right]$
Y_{AB}^{01}	$N\left[(1-p_1)\theta_A\theta_Bp_2+p_1(1-\theta_A\theta_B)p_2\theta_A\theta_B\right]$
$Y_{A\cdot}^{1\cdot}$	$Np_1\theta_A(1-\theta_B)$
$Y_{A\cdot}^{\cdot 1}$	$Np_2\theta_A(1-\theta_B)$
$Y_{\cdot B}^{1\cdot}$	$Np_1(1-\theta_A)\theta_B$
$Y_{\cdot B}^{\cdot 1}$	$Np_2(1-\theta_A)\theta_B$
$Y_{\cdot\cdot}^{1\cdot}$	$Np_1(1-\theta_A)(1-\theta_B)$
$Y_{\cdot\cdot}^{\cdot 1}$	$Np_2(1-\theta_A)(1-\theta_B)$

Table 4.1: Observed statistics and expected values for unmatchable list members under model $\phi=(N,p_1,p_2,\theta_{A_1}=\theta_{A_2},\theta_{B_1}=\theta_{B_2})$.

where μ is the vector of expected values of $\mathbf{Y}$ (see Table 4.1), $\mathbf{D}$ is a rectangular matrix of first partial derivatives of μ with respect to the parameter set ϕ, while $\mathbf{V}$ is a working covariance matrix of $\mathbf{Y}$ whose elements are derived based on assuming independent Poisson random variables for components of $\mathbf{Y}$. Although the parameter estimates cannot be solved explicitly, the estimates are consistent even if $\mathbf{V}$ is not the true covariance matrix of $\mathbf{Y}$ (Liang and Zeger, 1986).

The covariance matrix of the estimated parameters is

$$Var\left(\widehat{\phi}\right)=\left[\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right]^{-1}\left[\mathbf{D}'\mathbf{V}^{-1}V(\mathbf{Y})\mathbf{V}^{-1}\mathbf{D}\right]\left[\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right]^{-1},$$

where $\mathbf{D}$ and $\mathbf{V}$ are previously defined, and $V(\mathbf{Y})$ is the true covariance matrix of $\mathbf{Y}$.

The model-based (naive) covariance estimator assumes that $V(\mathbf{Y})=\mathbf{V}$, thus $Var\left(\widehat{\phi}\right)$ simplifies to $\left[\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right]^{-1}$. By assuming that $V(\mathbf{Y})=\mathbf{V}$, it is possible to misspecify the form of the variance of $\mathbf{Y}$. If the model or variance is misspecified and the model-based estimator is used, the estimated covariance will not be a consistent estimator of $Var\left(\widehat{\phi}\right)$.

We follow Liang and Zeger's (1986) approach of not explicitly defining the second

moment of $\mathbf{Y}$ and estimate $V(\mathbf{Y})$ by the matrix of residuals,

$$\widehat{V}(\mathbf{Y}) = \sum_{i}^{records} \mathbf{D}'\mathbf{V}^{-1}\left[\mathbf{Y}_i - \mu_i(\widehat{\phi})\right]\left[\mathbf{Y}_i - \mu_i(\widehat{\phi})\right]'\mathbf{V}^{-1}\mathbf{D},$$

where $\mathbf{Y}_i$ is an indicator vector for each individual. This is the empirical or robust covariance estimator. Because $\widehat{\phi} - \phi$ is asymptotically normal, with variance consistently estimated by $\widehat{Var}\left(\widehat{\phi}\right)$ (Prentice and Zhao, 1991), Wald-type confidence regions can be constructed for $\widehat{\phi}$.

To avoid misspecifying $V(\mathbf{Y})$ or relying on asymptotic results, we also calculated the non-parametric bootstrap standard error of $\widehat{\phi}$ (Efron and Tibshirani, 1993). Multinomial sampling based on the number of original observations in each statistic of $\mathbf{Y}$ relative to the total over all statistics Y generates bootstrap statistics $\widehat{\mathbf{Y}}^*$. For each of 1,000 bootstrap samples, $\widehat{\mathbf{Y}}^*$, we estimate ϕ^*. The non-parametric bootstrap standard error and confidence region are based on the estimates ϕ^*.

Different models can be fit to the observed data. For example, for any list, the tag loss rate could be equal over tags, *i.e.* $\theta_{A_1} = \theta_{B_1}$. The same methodology can be used to fit the model after the appropriate substitution in μ and $\mathbf{D}$. For different model types, we use the same assumptions for the working covariance matrix and the robust estimator, while $\mathbf{D}$ is now differentiated with respect to newly defined parameter sets.

Two approaches for model selection are used. First, we use the ordered Pearson-type goodness-of-fit statistics from the bootstrap samples (Efron and Tibshirani, 1993). Assuming that a better fitting model corresponds to a smaller statistic, we establish a p-value for the observed goodness-of-fit statistic. Secondly, we compute AIC statistics for estimating functions as developed by Pan (2001), the QIC. The AIC and QIC assume that the elements of $\mathbf{Y}$ are uncorrelated. Correlation among statistics may be result in over-dispersion in the counts. In our example, the number of records without both tags is low so we think it is unlikely that there is significant correlation.

The estimating functions framework is readily generalizeable to more than two lists

or more than two tags. For example, to estimate disease prevalence, names and birthdates may be available from three administrative sources. Under the assumptions previously outlined, an individual observed on all three lists is a member of cell Y_{AB}^{111}. The statistic $\mathbf{Y}$ now has 16 components. The components $Y_{AB}^{111}, Y_{AB}^{110}, Y_{AB}^{101}, Y_{AB}^{011}, Y_{AB}^{100}, Y_{AB}^{010}$, and Y_{AB}^{001} represent those records with both tags present, while $Y_{A\cdot}^{1\cdot\cdot}, Y_{A\cdot}^{\cdot 1\cdot}, Y_{A\cdot}^{\cdot\cdot 1}, Y_{\cdot B}^{1\cdot\cdot}, Y_{\cdot B}^{\cdot 1\cdot}$, $Y_{\cdot B}^{\cdot\cdot 1}, Y_{\cdot\cdot}^{1\cdot\cdot}, Y_{\cdot\cdot}^{\cdot 1\cdot}$, and $Y_{\cdot\cdot}^{\cdot\cdot 1}$ are counts of records with missing tags. We can no longer use the Petersen estimator for comparison. Traditionally, a log-linear model (Cormack, 1989) would be fit to the $2 \times 2 \times 2$ contingency table with one missing cell using the 7 components of $\mathbf{Y}$ with all tags present.

To illustrate the method to generate μ for a more complicated case, consider a three tag (Tag A, B and C), two list example. A two step is considered. First, records are partitioned into statistics Y based on tags present at capture. Then, μ associated with each Y are generated.

For those records with all tags present, we find possible statistics are: $Y_{ABC}^{01}, Y_{ABC}^{10}$, Y_{ABC}^{11}. We assumed that all tags were required to match records across lists, so that records with fewer than three tags cannot be matched across lists. For those records with any two tags present, the statistics are: $Y_{AB\cdot}^{1\cdot}, Y_{A\cdot C}^{1\cdot}, Y_{\cdot BC}^{1\cdot}$ (list 1) and $Y_{AB\cdot}^{\cdot 1}, Y_{A\cdot C}^{\cdot 1}, Y_{\cdot BC}^{\cdot 1}$ (list 2). The statistics for records with a single tag are: $Y_{A\cdot\cdot}^{1\cdot}, Y_{\cdot B\cdot}^{1\cdot}, Y_{\cdot\cdot C}^{1\cdot}$ (list 1) and $Y_{A\cdot\cdot}^{\cdot 1}, Y_{\cdot B\cdot}^{\cdot 1}, Y_{\cdot\cdot C}^{\cdot 1}$ (list 2). Lastly, statistics with no tags present at capture on either list are: $Y_{\cdots}^{1\cdot}, Y_{\cdots}^{\cdot 1}$.

The second step is generating the expectations associated with each statistic Y. This step is generally done by inspection. In our two-list, three-tag case we find

$$\begin{aligned}
E[Y_{ABC}^{11}] &= Np_1\theta_{A_1}\theta_{B_1}\theta_{C_1}p_2\theta_{A_2}\theta_{B_2}\theta_{C_2} \\
E[Y_{AB\cdot}^{\cdot 1}] &= Np_2\theta_{A_2}\theta_{B_2}(1-\theta_{C_2}) \\
E[Y_{A\cdot\cdot}^{1\cdot}] &= Np_1\theta_{A_1}(1-\theta_{B_1})(1-\theta_{C_1}) \\
E[Y_{\cdots}^{1\cdot}] &= Np_1(1-\theta_{A_1})(1-\theta_{B_1})(1-\theta_{C_1}).
\end{aligned}$$

4.3 Example

This example is based on a study to estimate the size of a population of health care users in a large urban setting. The population is transient and census figures are believed to be unreliable. The first list is derived from visits to several drop-in centres where respondents are asked for their initials and birthdate. The second list is a compilation of administrative lists in the health care region. We are unfortunately unable to provide original data due to confidentiality restrictions. Based on our analysis of the original data, artificial data is created.

For each list, there are two tags. The first tag is birthdate (year, month and day), and the second tag is the first and last initial. Records have both initials or neither. All records on List 2 have both tags. The (simulated) statistics are

$$\mathbf{Y} = \begin{bmatrix} Y_{AB}^{11} \\ Y_{AB}^{10} \\ Y_{AB}^{01} \\ Y_{A\cdot}^{1\cdot} \\ Y_{A\cdot}^{\cdot 1} \\ Y_{\cdot B}^{1\cdot} \\ Y_{\cdot B}^{\cdot 1} \\ Y_{\cdot\cdot}^{1\cdot} \\ Y_{\cdot\cdot}^{\cdot 1} \end{bmatrix} = \begin{bmatrix} 166 \\ 1405 \\ 2170 \\ 0 \\ 0 \\ 192 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The Petersen estimate using all n_1, n_2 records is $\widehat{N}_{Petersen,\ all} = 24,810$. Excluding records with missing tags, $\widehat{N}_{Petersen,\ complete} = 22,108$. The approximately unbiased Petersen estimate using complete records is $21,998$, with an estimated standard error of $1,530.9$.

We fit two models to the observed data, $\phi' = \{N, p_1, p_2, \theta_{A_1} = \theta_{A_2}, \theta_{B_1} = \theta_{B_2}\}$ and $\phi'' = \{N, p_1, p_2, \theta_{A_1} = \theta_{A_2} = \theta_{B_1} = \theta_{B_2}\}$. Model $\phi''' = \{N, p_1, p_2, \theta_{A_1}, \theta_{B_1}, \theta_{A_2}, \theta_{B_2}\}$ is not fit because several statistics of $\mathbf{Y}$ have zero counts.

The system of estimating functions fails to converge. Agresti (1990) suggests adding a small constant to each cell in the log-linear modeling context to avoid non-convergence; we added $\frac{1}{10}$ to each statistic of $\mathbf{Y}$ and found that the system of equations

converged. The population estimate is not sensitive to the constant added. For example, we separately added $\frac{1}{10}$ to $Y_{\cdot\cdot}^{1\cdot}$ and $Y_{\cdot\cdot}^{\cdot 1}$ only, and found that $\widehat{N}$ increased very slightly (0.04%), while the remaining estimates were unchanged to 4 decimal places.

The Pearson goodness-of-fit statistics and QIC statistics were calculated for each model. The $1,000$ goodness-of-fit statistics for each of models $\phi^{'}$ and $\phi^{''}$ were calculated based on the bootstrap samples. 35% and 42% of the ordered goodness-of-fit statistics were less than the Pearson statistic from models $\phi^{'}$ and $\phi^{''}$, respectively. The small difference between 35% and 42% is unexpected, as model $\phi^{''}$ is less flexible than $\phi^{'}$, recalling $\phi^{'}=\{N, p_1, p_2, \theta_{A_1} = \theta_{A_2}, \theta_{B_1} = \theta_{B_2}\}$ and
$\phi^{''}=\{N, p_1, p_2, \theta_{A_1} = \theta_{A_2} = \theta_{B_1} = \theta_{B_2}\}$.

The QIC statistic suggests that $\phi^{'}$ is a better fit, with $QIC_{\phi^{'}} = 358.7$ and $QIC_{\phi^{''}} = 603.2$. In our example, the goodness-of-fit and QIC statistics may not be reliable because both are adversely affected by small expected counts in several cells. We decided to proceed with model $\phi^{'}$ as it seems to be the most plausible given that we observe no individuals on either list without Tag B. The parameter estimates are:

$$\widehat{\phi^{'}} = \begin{bmatrix} \widehat{N} \\ \widehat{\theta}_A \\ \widehat{\theta}_B \\ \widehat{p}_1 \\ \widehat{p}_2 \end{bmatrix} = \begin{bmatrix} 22,159 \\ 0.9531 \\ 0.9999 \\ 0.0792 \\ 0.1059 \end{bmatrix}.$$

We note that $\widehat{\theta}_B \neq 1$ due to the small constant added to the statistics of $\mathbf{Y}$ to ensure convergence.

Using a model based approach, the estimated standard error of $\widehat{N}$ is $1,561.7$. When we calculated $\widehat{Var}\left(\widehat{\phi^{'}}\right)$ using the robust estimator based on the residuals, we estimate the standard error of $\widehat{N}$ to be $1,535.4$. The non-parametric bootstrap

Estimation Method	$\widehat{N}$	$\widehat{se}(\widehat{N})$
Approx. Unbiased Petersen (all records)	24684	1728.7
Approx. Unbiased Petersen (complete records)	21998	1530.9
Partial Likelihood Function	22098	1562.8
Estimating Functions	22159	
Model Based *se*		1561.7
Robust *se*		1534.5
Non-parametric Bootstrap *se*		1550.3

Table 4.2: Comparison of population estimates and estimated standard errors. $Y_{AB}^{11} = 166, Y_{AB}^{10} = 1405, Y_{AB}^{01} = 2170$ and $Y_{\cdot B}^{1\cdot} = 192$.

standard errors and confidence regions are:

$$\widehat{se}(\widehat{\phi}') = \begin{bmatrix} 1550.3 \\ 0.0033 \\ 0.0003 \\ 0.0058 \\ 0.0078 \end{bmatrix}, 95\% \text{ c.i. for } \widehat{\phi}' = \begin{bmatrix} 19,552, 25,531 \\ 0.9464, 0.9595 \\ 0.9989, 0.9999 \\ 0.0678, 0.0906 \\ 0.0907, 0.1212 \end{bmatrix}.$$

For comparative purposes, we maximized the partial likelihood function. As above, we added $\frac{1}{10}$ to each statistic to obtain convergence and fit the same model, ϕ'. The parameter estimates are $\widehat{N} = 22,098, \widehat{p}_1 = 0.0188, \widehat{p}_2 = 0.0015, \widehat{\theta}_A = 0.5000$ and $\widehat{\theta}_B = 0.9508$. The estimates of p_1, p_2, θ_A, and θ_B are poor and are reflective of the statistics upon which they are based having counts of 0.

As shown in Table 4.2, the estimates of $\widehat{N}$ and estimated standard error are very similar among competing methods except for the approximately unbiased Petersen estimate based on all records. Unfortunately, we do not have access to the original study's model or population estimate for comparative purposes.

4.4 Simulation Study

A simulation study was used to determine whether the estimating functions approach to population estimation is an improvement to existing methodologies in circumstances wider than our example. We do so by determining how well the estimators perform while simulating differing rates of tag loss on lists of differing lengths.

We created a 'theoretical' population of pre-determined size, each individual having two tags. Two tags are required to uniquely identify an individual. Two independent lists of equal sizes are sampled without replacement from the population. Then, for each record, tags are removed at random with a specified probability. A model is fit to the vector of statistics $\mathbf{Y}$ and summary statistics computed. Scenarios are constructed by varying the rate of tag loss by list or tag type and repeated 1,000 times.

In the first scenario, we simulate equal tag loss rates across tags and lists, *i.e.* $\theta_{A_1} = \theta_{A_2} = \theta_{B_1} = \theta_{B_2}$. In the second, we vary tag loss rates by tag, such that $\theta_{A_1} = \theta_{A_2}, \theta_{B_1} = \theta_{B_2}$, and the third has tag loss varying by list, $\theta_{A_1} = \theta_{B_1}, \theta_{A_2} = \theta_{B_2}$. To these simulated theoretical populations, we fit the model $\phi = \{N, p_1, p_2, \theta_{A_1} = \theta_{A_2}, \theta_{B_1} = \theta_{B_2}\}$. Lists have 2,000 records (5% of the population) or 4,000 records. Our comparator is the approximately unbiased Petersen estimator using records with complete information only.

In the first portion of Table 4.3, the mechanism of tag loss is $\theta_{A_1} = \theta_{A_2} = \theta_{B_1} = \theta_{B_2}$, while the model fit is $\theta_{A_1} = \theta_{A_2}, \theta_{B_1} = \theta_{B_2}$. The estimating functions estimate the true population size well, although the distribution of the estimates of $\widehat{N}$ is slightly right skewed. Estimates are similar to those using the approximately unbiased Petersen estimator.

When we simulate tag loss by tag type ($\theta_{A_1} = \theta_{A_2}, \theta_{B_1} = \theta_{B_2}$), the population estimates derived from the estimating functions are very similar to the approximately unbiased Petersen. This is not surprising given that the model fit is $\phi = \{N, p_1, p_2, \theta_{A_1} = \theta_{A_2}, \theta_{B_1} = \theta_{B_2}\}$. The median $\widehat{N}$ is close to the true population size for both list sizes.

Varying the rate of tag loss between lists, but not by tag ($\theta_{A_1} = \theta_{B_1}$ and $\theta_{A_2} = \theta_{B_2}$),

Tag Retention Rate				Estimating Functions			Unbiased Petersen	
				Median $\widehat{N}$	Mean $\widehat{N}$	sd $\widehat{N}$	Mean $\widehat{N}$	sd $\widehat{N}$
θ_{A_1}	θ_{B_1}	θ_{A_2}	θ_{B_2}	Two lists of 2,000 records				
.95	.95	.95	.95	40145.7	40568.7	4440.5	40106.5	4332.5
.90	.90	.90	.90	39822.9	40266.5	4796.5	39694.4	4651.9
.85	.85	.85	.85	40182.8	40817.9	5674.4	40095.1	5501.3
.80	.80	.80	.80	40252.4	41191.0	6567.0	40260.6	6453.2
.95	.90	.95	.90	39970.3	40333.4	4568.8	39822.3	4446.9
.95	.85	.95	.85	39723.6	40329.6	4972.0	39751.8	4821.1
.95	.80	.95	.80	39963.1	40466.0	5101.6	39807.2	4927.2
.95	.75	.95	.75	39985.6	40646.1	5516.2	39886.0	5298.4
.95	.95	.90	.90	40328.5	40630.6	4649.0	40005.6	4510.2
.95	.95	.85	.85	40586.9	41053.2	4863.7	40012.0	4658.0
.95	.95	.80	.80	41508.1	41975.6	5332.1	40179.4	4997.1
.95	.95	.75	.75	42370.5	43128.8	6126.0	40179.6	5562.5
θ_{A_1}	θ_{B_1}	θ_{A_2}	θ_{B_2}	Two lists of 4,000 records				
.95	.95	.95	.95	40090.3	40161.6	2254.7	40058.6	2242.2
.90	.90	.90	.90	39940.5	40125.4	2498.6	39995.0	2481.0
.85	.85	.85	.85	39904.5	40046.5	2702.7	39879.5	2678.0
.80	.80	.80	.80	40004.6	40247.1	3321.8	40027.2	3282.6
.95	.90	.95	.90	39964.6	40042.2	2345.2	39927.2	2330.6
.95	.85	.95	.85	39941.3	40113.6	2525.9	39982.4	2507.8
.95	.80	.95	.80	39992.2	40154.2	2614.6	40003.5	2593.4
.95	.75	.95	.75	39895.7	40146.9	2844.3	39973.7	2817.8
.95	.95	.90	.90	40198.9	40241.4	2195.0	40030.2	2175.8
.95	.95	.85	.85	40451.0	40567.4	2375.8	40019.7	2330.4
.95	.95	.80	.80	40993.0	41105.2	2541.6	39940.9	2447.3
.95	.95	.75	.75	41770.6	42019.1	2772.0	39877.8	2589.4

Table 4.3: Results from two list, two tag simulation study of performance of estimating equations and approximately unbiased Petersen under several scenarios of tag loss. 1,000 replicates, $N = 40,000$ and model fit is $\phi = \{N, p_1, p_2, \theta_{A_1} = \theta_{A_2}, \theta_{B_1} = \theta_{B_2}\}$.

has a negative impact on the ability of the estimating functions to estimate population size. This result is not surprising because the model fit is not correct.

In the two-list, two-tag scenario of Table 4.3, we found that if the model is not incorrectly specified, our estimator appears equally precise when compared to the approximately unbiased Petersen estimator for both list sizes sampled. This is unexpected in situations of high rates of tags loss, where we thought that the precision of the unbiased Petersen would suffer with fewer complete records.

We also examine whether the estimating function estimate population size well with more than two lists. We drew three lists from the same two tag population and randomly removed tags from records. As our comparator, we fit a log-linear model to the counts of the subset of statistics of $\mathbf{Y}$ with all tags present. Since the lists were drawn independently, only list effects are included in the log-linear model (no interaction effects).

In the three-list, two-tag example of Table 4.4, our estimator is equally precise as the log-linear model fit to the complete records. There appears to be little skewness in the estimate of population size in the three list case. To simplify output, the median estimate is not provided. Unexpectedly, we found that even if the model is incorrectly specified, there appears to be little bias in the estimate of $\widehat{N}$ when compared to the log-linear model. We had the benefit of determining the characteristics of the sampled lists; so the 'correct' log-linear model could be fit. If however, the incorrect log-linear model were fit, we would expect population estimates to be less biased.

4.5 Discussion

Previous investigators have shown that considerable bias can be introduced into population estimates if tag loss and mismatch errors are not addressed. The difficulty with existing tag loss models is that they require records to match uniquely on remaining tags (Seber and Felton, 1981; Seber, 1982; Seber, Huakau and Simmons, 2000). There are potentially many epidemiological population estimation applications where it is difficult to meet these criteria when drawing upon administrative sources.

Our method relaxes the assumption of having to match records uniquely across

Tag Retention Rate						Estimating Functions		Log-Linear Models	
						Mean $\widehat{N}$	sd $\widehat{N}$	Mean $\widehat{N}$	sd $\widehat{N}$
θ_{A_1}	θ_{B_1}	θ_{A_2}	θ_{B_2}	θ_{A_3}	θ_{B_3}	Three lists of 2,000 records			
.95	.95	.95	.95	.95	.95	40036.3	2438.9	40073.4	2436.3
.90	.90	.90	.90	.90	.90	40279.2	2744.9	40346.7	2740.2
.85	.85	.85	.85	.85	.85	40160.9	3174.1	40253.5	3167.0
.80	.80	.80	.80	.80	.80	40170.3	3467.7	40287.2	3457.8
.95	.90	.95	.90	.95	.90	40037.0	2483.5	40075.9	2480.1
.95	.85	.95	.85	.95	.85	40089.9	2788.2	40142.0	2783.3
.95	.80	.95	.80	.95	.80	40230.7	2983.0	40305.1	2976.3
.95	.75	.95	.75	.95	.75	40277.3	3164.0	40382.8	3155.3
.95	.95	.90	.90	.90	.90	40125.6	2624.6	40138.1	2617.9
.95	.95	.90	.90	.85	.85	40138.2	2771.6	40065.9	2755.3
.95	.95	.90	.90	.80	.80	40397.8	2827.9	40144.2	2871.7
.95	.95	.90	.90	.75	.75	40653.1	3104.1	40317.7	3213.7
θ_{A_1}	θ_{B_1}	θ_{A_2}	θ_{B_2}	θ_{A_3}	θ_{B_3}	Three lists of 4,000 records			
.95	.95	.95	.95	.95	.95	39985.4	1147.7	40136.0	1143.4
.90	.90	.90	.90	.90	.90	39796.4	1275.6	40075.4	1265.5
.85	.85	.85	.85	.85	.85	39606.7	1501.0	39994.5	1485.7
.80	.80	.80	.80	.80	.80	39609.6	1637.3	40088.9	1617.5
.95	.90	.95	.90	.95	.90	39837.6	1218.9	40031.4	1212.0
.95	.85	.95	.85	.95	.85	39739.9	1295.9	39993.7	1285.8
.95	.80	.95	.80	.95	.80	39679.9	1398.0	40008.1	1385.4
.95	.75	.95	.75	.95	.75	39599.2	1494.0	40016.0	1478.0
.95	.95	.90	.90	.90	.90	39791.0	1261.1	40031.4	1212.0
.95	.95	.90	.90	.85	.85	39855.3	1308.9	39993.7	1285.8
.95	.95	.90	.90	.80	.80	39998.5	1417.0	40008.1	1385.4
.95	.95	.90	.90	.75	.75	40189.8	1428.8	40016.0	1478.0

Table 4.4: Results from three list, two tag simulation study of performance of estimating equations and log-linear models under several scenarios of tag loss when model $\phi = \{N, p_1, p_2, \theta_{A_1} = \theta_{A_2} = \theta_{A_3}, \theta_{B_1} = \theta_{B_2} = \theta_{B_3}\}$ .1,000 replicates and $N = 40,000$.

lists when tags are missing. However, we assume that when matching records with all tags present that records are matched without error.

Although multi-list methods assume a closed population, lists are compiled over time. If the period is too long, the closed population assumption may not be tenable and options for open populations examined.

The estimating function framework is flexible, as it allows different models to be investigated. This approach to estimating population size produces estimates that have a degree of precision competitive with existing methodologies.

Surprisingly, in the two-list case, our approach appears to have no noticeable improvement over the approximately unbiased Petersen estimator based on records with all tags. Also, our estimator did not out-perform the log-linear model estimator when the latter was based on complete records only. It appears that there is little contribution to improving the estimate of $\widehat{N}$ from the additional cells of $\mathbf{Y}$.

Advantages that our approach holds are that estimates of parameters, other than $\widehat{N}$, may also be easily obtained, such as capture probabilities and tag loss rates. Also, θ appears to be robust to modelling different mechanisms of tag loss.

The next steps would include assessing the estimating functions' ability to estimate population size with small sample sizes or higher tag loss rates. Also, comparisons against the log-linear model will be investigated when the model is misspecified.

Chapter 5

Conclusions and Future Work

5.1 Summary

In Chapter 2, the work of Hook and Regal (1993) with stratified lists was generalized to multiple over-lapping lists even when some lists may not have been operating in all strata. Cell counts are modelled as independent Poisson counts. Then, in strata in which not all lists operated, cell counts were treated as though part of the history vector was unobservable. In other words, $Y_{1,\{1,\cdot,1\}} = Z_{1,\{1,1,1\}} + Z_{1,\{1,0,1\}}$, where $Y_{1,\{1,\cdot,1\}}$ is the observed count based on matching across lists 1 and 3 in stratum 1, while $Z_{1,\{1,1,1\}} + Z_{1,\{1,0,1\}}$ are, respectively, the unobservable counts on lists 1, 2 and 3, and lists 1 and 3 only in stratum 1. Then, conditioning on the observed cell counts, the expected cell counts are estimated using a binomial (or multinomial) distribution. An EM algorithm maximizes the conditional expected log-likelihood and estimates the model parameters.

Our first example is a re-analysis of the Auckland Diabetes Study data. Although the full data was available, we simulated missing lists in some strata to assess our method. The AIC statistic was used for model selection. Our estimate of 22,813 varied considerably from Huakau's (2001), whose estimate was 26,886. The difference between estimators is not surprising considering we 'collapsed' our data, resulting in an inability to detect some complex model effects. Huakau's (2001) final model according to the AIC statistic found two three-factor interactions, [GPD][POD], where we were

only able to determine main effects, [G][P][O][D].

The analysis of the forest fire data introduced small cell count problems. Our original models would not converge because of marginal counts of 0 , although when we added a small constant, based on the methods of Evans and Bonett (1994), the models converged rapidly. Our estimated standard error was likely an underestimate for two reasons. First, in strata in which a single list was recording fires, each bootstrap sample was the same for that list. Secondly, if a list failed to record any fires in a stratum, then that list "observes" no fires in each re-sample of the data. We also detected some over-dispersion in the model.

As discussed in the introduction, an estimate of precision is conditional upon the model selected. It would be appropriate to use model selection criteria of the EM algorithm and apply model averaging to account for model uncertainty of estimates. The framework has been laid by other authors, but it remains to be seems whether estimates generated using our method are sensitive to the model selected.

The method developed based on the EM algorithm benefits from "sharing" information among strata. The method is best applied when there are multiple, overlapping lists and when the underlying model is common over strata. Our method will fail if there are 'hidden' list dependencies on unobserved lists in some strata that our model cannot detect. In this case, however, all methods will produce poor estimates of overall population size.

In Chapter 3, a multi-list method was developed which relaxed the assumption that all lists are required to have the same tags. In other words, not all history vectors of the $2^K - 1$ contingency table can be observed. Graph theory was used to develop a method for determining which lists share a common 'tag' and, hence, which capture histories are observable. We assume that observed counts are independent Poisson counts. Then, we relate the observed counts to the unobserved counts through a sum of exponential terms. Because of the potential of double-counting individuals, an estimating functions approach is favored over likelihood methods. Several methods for estimating the variance of estimates are shown.

Again, the Auckland Diabetes Study data is used for an example. We simulate cell counts by summing over unobserved cells. For model selection, we use the QIC

statistic, the estimating functions analogue to the AIC. Our estimate varied considerably from the Petersen estimate from two lists on one tag, and the log-linear estimate based on the remaining tag on three lists. The difference between methods of estimating diabetes prevalence is not surprising because our model is able to detect more complex list interactions that either the Petersen or the log-linear model.

Using the estimating functions, precision of the estimate of population size is still based on the selected model. We have not seen in the literature that the QIC has been used for model selection in model averaging to account for model uncertainty. It would be useful to determine whether our estimate was sensitive to the model selected and whether the QIC statistic impacts our estimate of population size through model averaging.

The methodology developed in this chapter is shown to have better precision than log-linear models or the simple Petersen on subsets of lists when only subsets of lists are available. The method is best applied when the underlying model is common across tags. Simulation study shows that missing lists can lead to 'misleading' models as a result of interaction effects not being properly detected. Consequently, inappropriate models are fit to the data which lead to poor estimates of population size. However, this is also true for the other methods which only used a subset of tags available for matching.

In Chapter 4, the assumption that having to match records uniquely across lists is relaxed due to missing tags. However, we maintain the assumption that when matching records with all tags present, records are matched without error. Because records without all tags cannot be match across lists, it is possible for individuals to be present on multiple lists without be matched across lists. The simplest approach for this type of data is to simply use existing capture-recapture methods for individuals matched on all tags only. In this case, individuals with less than all tags are discarded and provide no further information. However, there must be information available from the 'discarded' cases.

Because we cannot rule out double counting, a multinomial likelihood is not appropriate. However, a partial likelihood that conditions on the sum of individuals with all tags present is shown to work well. An estimating function framework that

models population size, probability of capture and rates of tag loss is also employed. The statistics are based on presence/absence of tags of capture and observed capture history.

The estimating function approach permits a flexible framework to incorporate different mechanisms of tag-loss and is generalizable to any number of lists.

In our example, we had several cells with counts of 0. To ensure convergence, we added a small constant to cell counts. Model selection used the QIC statistic. Our estimated population size was very similar to the Petersen estimator based on complete tags.

Based on simulated data, our estimating equations approach offers no noticeable improvement over the Petersen estimator based on records with all tags. Also, the method did not improve precision of estimates of the log-linear model based on complete records only. Surprisingly, there appears to be little contribution from the individuals with partial information to improving the estimate of $\widehat{N}$. However, this approach is advantageous in that estimates of parameters, other than $\widehat{N}$, are easily obtained. This includes capture probabilities and tag loss rates. Our mechanism for modelling different mechanisms of tag loss, θ, appears to be robust to different rates of tag loss. It would be interesting to consider whether the method improves over existing methods when sample sizes were smaller or tag loss rates were extreme. And as in the previous chapters, it would be instructive to determine whether incorporating model averaging for model uncertainty improves estimates of precision.

5.2 Future Work

In Chapter 2 and Chapter 3, we assumed that individuals could be matched across lists without error. It seems reasonable to consider that this assumption could be relaxed to include adjustments for tag mismatches if more than a single tag were available on each list, as per Huakau (2001) and Lee (2002) and Lee et al (2001). Administrative lists often have more than a single tag, incorporating tag mismatches is not without applications. In this manner methods for partially stratified and incomplete lists would incorporate possible mismatches across individuals.

Also, in Chapter 2 and Chapter 3, our estimates of precision conditioned on the model being correct. Recent work is showing that adjusting estimates of population size for model uncertainty is prudent. Next steps would assess whether model averaging has a large impact on our developed methods.

5.3 An Apparent Duality

While developing the methodology for partially stratified and incomplete lists, we considered whether there was an element of 'duality' involved such that individuals (as represented by their tags) could be considered strata. However, an example shows this apparent duality is an illusion.

Consider the following three list example, written

$$\begin{bmatrix} & \text{Tag A} & \text{Tag B} & \text{Tag C} \\ \text{List 1} & \surd & \surd & \\ \text{List 2} & \surd & \surd & \surd \\ \text{List 3} & & \surd & \surd \end{bmatrix}.$$

In this example, Tag A and Tag B are present on list 1, Tag A, B and C are present on list 2 and on list 3, tag B and C are present. Now, consider 'switching' the strata for tags, written

$$\begin{bmatrix} & \text{Stratum 1} & \text{Stratum 2} & \text{Stratum 3} \\ \text{List A} & \surd & \surd & \\ \text{List B} & \surd & \surd & \surd \\ \text{List C} & & \surd & \surd \end{bmatrix},$$

where stratum 1 represents all individuals with Tag A and B, stratum 2 represents all individuals with Tag A, B and C and stratum 3, Tag B and C.

In the second arrangement of the information, list A is a compilation of all tag A's. However, it is possible that there is double-counting for some Tag A's, as an individual

required both Tag A and B to be uniquely identified in the original arrangement. Since double-counting violates our multi-list assumptions, our methods cannot be applied to estimating population size.

Bibliography

Alho, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, **46**, 623-635.

Alho, J.M., Mulry, M.H., Wurdeman, K. and Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, **88**, 1130-1136.

Agresti, A. (1990). Categorical data analysis. New York, John Wiley & Sons, Inc.

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, **50**, 494-500.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, 2^{nd} International Symposium on Information Theory (Petrov, S.N. and Csaki, F. eds.), Akademiai Kiado, Budapest, 267-281.

Basu, S. and Ebrahimi, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, **88**, 269-279.

Brownie, C., Hines, J.E., Nichols, J.D., Pollock, K.H. and Hestbeck, J.B. (1993). Capture-recapture studies for multiple strata including non-Markovian transitions. *Biometrics*, **49**, 1173-1187.

Borchers, D.L, Buckland, S.T. and Zucchini, W. (2002). *Statistics for Biology and Health - Estimating Animal Abundance Closed Populations*. Great Britain: Springer.

Buckland, S.T. (1980). A modified analysis of the Jolly-Seber capture-recapture model. *Biometrics*, **36**, 419-435.

Buckland, S.T. (1984). Monte Carlo confidence intervals. *Biometrics*, **40**, 811-817.

Buckland, S.T. and Garthwaite, P.H. (1991). Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics*, **47**, 255-268.

Buckland, S.T., Goudie, I.B.J. and Borchers, D.L. (2000). Wildlife population assessment: past developments and future directions. *Biometrics*, **56**, 1-12.

Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997). Model selection: an integral part of inference. *Biometrics*, **53**, 603-618.

Burnham, K.P. (1978). Unpublished PhD dissertion. Oregon State University, Department of Statistics.

Burnham, K.P. and Overton, W.S. (1978) Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, **65**, 625-633.

Burnham, K.P. and Overton, W.S. (1979). Robust estimation of population size when capture probabiltities vary among animals. *Ecology*, **60**, 927-936.

Burnham, K.P. and Anderson, D.R. (1992). Data-based selection of an appropriate biological model: the key to modern data analysis. In *Wildlife 2001: Populations*, D.R. McCullough and R.H. Barrett (eds.), London: Elsevier Science Publishers.

Burnham, K.P., White, G.C. and Anderson, D.R. (1995). Model selection strategy in the analysis of capture-recapture data. *Biometrics*, **51**, 888-898.

Burnham, K.P. and Anderson, D.R. (1998). *Model selection and inference: a practical information-theoretical approach*. New York: Spring-Verlag.

Castledine, B.J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, **67**, 197-210.

Chapman, D.G. (1951). Some properties of the hypergeometric distribution with applications to zoological censuses. *Univ. Calif. Public. Statist.*, **1**, 131-160.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783-791.

Chao, A., Lee, S.-M. (1992a). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, **87**, 210-217.

Chao, A., Lee, S.-M., Jeng, S.-L. (1992b). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, **48**, 201-216.

Chao, A. and Tsay, P.K. (1998). A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistician*, **93**, 283-293.

Chao, A., Chu, W. and Hsu, C.-H. (2000). Capture-recapture when time and behavioral response affect capture probabilities. *Biometrics*, **56**, 427-433.

Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological and Environmental Statistics*, **6**, 158-175.

Chao, A. (2001b). Population size estimation based on estimating functions for closed capture-recapture models. *Journal of Statistical Planning and Inference*, **92**, 213-232.

Cochran, W.G. (1977). Sampling Techniques, 3^{rd} ed. John Wiley and Sons, New York, New York.

Cormack, R.M. (1979). Models for capture-recapture. In Cormack, R.M., Patil, G.P. and Robson, D.S., editors, *Sampling Biological Populations*. International Co-operative Publishing House, Fairland.

Cormack, R.M. (1989). Log-linear models for capture-recapture. *Biometrics*, **45**, 395-413.

Cormack, R.M. and Jupp, P.E. (1991). Inference for Poisson and multinomial models for capture-recapture experiments. *Biometrika*, **78**, 911-916.

Cormack, R.M. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics*, **48**, 567-576.

Corrao, G., Bagnardi, V., Vittadini, G. and Favilli, S. (2000). Capture-recapture methods to size alcohol related problems in a population. *Journal of Epidemiology and Community Health*, **54**, 603-610.

Coull, B.A. and Agresti (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, **55**, 294-301.

Darroch, J.N. (1958). The multiple recapture census: I. Estimation of a closed population *Biometrika*, **45**, 343-359.

Darroch, J.N. (1961). The two-sample capture-recapture census when taagging and sampling are stratified. *Biometrika*, **48**, 241-260.

Darroch, J.N., Feinberg, S.E., Glonek, G.F.V. and Junker, B.W. (1993). A three-sample multiple-recapture approach to census population estiamtion with heterogeneous catchability. *Journal of the American Statistical Association*, **88**, 1137-1148.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Ding, Y. Feinberg, S.E. (1994). Dual system estimation of census undercount in the presence of matching error. *Survey Methodology*, **20**, 149-158.

Efron, B., Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, Chapman and Hall.

Egeland, G.M., Perham-Hester, K.A. and Hook, E.B. (1995). Use of capture-recapture analyses in fetal alcohol syndrome surveillance in Alaska. *American Journal of Epidemiology*, **141**, 335-341.

Evans, M.A. and Bonett, D.G. (1994). Bias reduction for multiple-recapture estimators of closed population size. *Biometrics*, **50**, 388-395.

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

Feinberg, S.E. (1972). Multiple-recapture census for closed population and incomplete contingency tables. *Biometrika*, **59**, 591-603.

Feinberg, S.E., Johnson, M.S. and Junker, B.W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society, Series A*, **162**, 383-405.

Feinberg, S.E. (2000). Contingency tables and log-linear models: basic results and new developemnts. *Journal of the American Statistical Association*, **95**, 643-647.

Fisher, N., Turner, S.W., Pugh, R. and Taylor, C. (1994). Estimating numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis. *British Medical Journal*, **308**, 27-30.

George, E.I. and Robert, C.P. (1992). Capture-recapture estimation via Gibbs sampling. *Biometrika*, **79**, 677-683.

Godambe V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, **31**, 1208-1212

Hay, G. (2000). Capture-recapture estimates of drug misuse in urban and non-urban settings in the north east of Scotland. *Addiction*, **95**, 1795-1803.

Heyerdahl, E. (1997). Spatial and temporal variation in historical fire regimes of the Blue Mountains, Oregon and Washington: the influence of climate. University of Washington, Seattle, Ph.D. dissertation.

Hook, E.B. and Regal, R.R. (1982). Validity of Bernoulli census, log-linear, and truncated binomial models for correcting for underestimates in prevalence studies. *American Journal of Epidemiology*, **116**, 168-176.

Hook, E.B. and Regal, R.R. (1992). The value of capture-recapture methods even for apparent exhaustive surveys. *American Journal of Epidemiology*, **135**, 1060-1067.

Hook, E.B., Regal, R.R. (1993). Effect of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates of prevalence. *American Journal of Epidemiology*, **137**, 1148-1166.

Hook, E.B. and Regal, R.R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *American Epidemiologis Reviews*, **17**, 243-264.

Hook, E.B., Regal, R.R. (1997). Validity of methods for model selection, weighting for model uncertainty, and small sample adjustments in capture-recapture estimation. *American Journal of Epidemiology*, **145**, 1138-1144.

Hook, E.B. and Regal, R.R. (2000). Accuracy of alternative approaches to capture-recapture estimates of disease frequency: Internal validity analysis of data from five sources. *American Journal of Epidemiology*, **152**, 771-779.

Huakau, J.T. (2001). PhD dissertion. The University of Aukland. New Zealand.

Huether, C.A. and Gummere, G.R. (1982). Influence of demographic factors on annual Down's syndrome in Ohio, 1970-1979, and the United States, 1920-1979. *American Journal of Epidemiology*, **115**, 846-860.

Huggins, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, **76**, 133-140.

Huggins, R.M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, **47**, 725-732.

International Working Group for Disease Monitoring and Forecasting (IWGDMF) (1995a). Capture-recapture and multiple-record systems estimation I. History and theoretical development. *American Journal of Epidemiology*, **142**, 1047-1058.

International Working Group for Disease Monitoring and Forecasting (IWGDMF) (1995b). Capture-recapture and multiple-record systems estimation II. Applications in human diseases. *American Journal of Epidemiology*, **142**, 1059-1068.

Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal ofthe American Statistical Association*, **84**, 414-420.

Jaro, M.A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, **14**, 491-498.

Jolly, G.M. (1979). A unified approach to mark-recapture stochastic models exemplified by a constant survival rate model. In Cormack, R.M., Patil, G.P. and Robson, D.S., editors, *Sampling Biological Populations*. International Co-operative Publishing House, Fairland.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

King, R. and Brooks, S.P. (2001). On the Bayesian analysis of population size. *Biometrika*, **88**, 317-336.

Lebreton, J.D., Burnham, K.P., Clobert, J. and Anderson, D.R. (1992). Modeling survival and testing biological hypotheses using marked animals- A unified approach with case studies. *Ecological Monographs*, **62**, 67-118.

Lee, A. J., Seber, G.A.F., Holden, J.K., Huakau, J.T. (2001). Capture-recapture, epidemiology, and list mismatches: several lists. *Biometrics*, **57**, 707-713.

Lee, A. J. (2002). Effects of list errors on the estimation of population size. *Biometrics*, **58**, 185-191.

Lee, S.-M. and Chao, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics*, **50**, 88-97.

Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Lloyd, C.J. (1992). Modified martingale estimation for recapture experiments with heterogeneous capture probabilities. *Biometrika*, **79**, 833-836.

Lloyd, C.J. (1994). Efficiency of martingale methods in recapture studies. *Biometrika*, **81**, 305-315.

Lloyd, C.J. and Yip, P. (1991). A unification of inference for capture-recapture studies through martingale estimating functions, in *Estimating Equations*, ed. V.P. Godambe, Oxford: Claredon Press.

Madigan, D. and York, J. (1995). Bayesian graphical methods for discrete data. *International Statistical Review*, **64**, 215-232.

Madigan, D. and York, J. (1997). Bayesian methods for estiating the size of a closed population. *Biometrika*, **84**, 19-31.

Manly, B.F.J. and McDonald, L.L. (1996). Sampling wildlife populations. *Chance*, **9**, 9-19.

McLachlan, G.J., Krishnan, T. (1997). *The EM algorithm and extensions*. New York, John Wiley & Sons, Inc.

Norris, J.L. and Pollock, K.H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, **52**, 639-649.

Orton, H., Rickard, R. and Miller, L. (2001). Using active medical record review and capture-recapture methods to investigate the prevalence of Down Syndrome among live-born in Colorado. *Teratology*, **64**, S14-S19.

Otis, D.L., Burnham, K.P., White, G.C. and Anderson, D.R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, **62**, 1-135.

Pan, W. (2001). Akaike's information criteria in generalized estimating equations. *Biometrics*, **57**, 120-125.

Plante, N., Rivest, L.-P., Tremblay, G. (1998). Stratified capture-recapture estimation of the size of a closed population. *Biometrics*, **54**, 47-60.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, **56**, 434-442.

Pollock, K.H. (1974). The assumption of equal catchability of animals in tag-recapture experiements. Unpublished PhD dissertion. Cornell University, Biometrics Unit. United States.

Pollock, K.H. (1991). Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future. *Journal of the American Statistical Society*, **86**, 225-238.

Prentice, R.L. and Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, **47**, 825-839.

Raftery, A.E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, **83**, 251-266.

Rajwani, K. N. and Schwarz, C. J. (1997). Adjusting for missing tags in salmon escapement surveys. *Canadian Journal of Fisheries and Aquatic Sciences*, **54**, 800-808.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Rivest, L.-P. and Levesque, T. (2001). Improved log-linear model estimators of abundance in capture-recapture experiments. *Canadian Journal of Statistics*, **29**, 555-572.

Sanathanan, L. (1972). Models and estimation methods in visual scanning experiments. *Technometrics*, **14**, 813-829.

Sandland, R.L. and Cormack, R.M. (1984). Statistical inference for Poisson and multinomial models for capture-recapture experiments. *Biometrika*, **71**, 27-33.

Schnabel, Z.E. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, **45**, 348-352.

Schwarz, C.J. and Ganter, B. (1995). Estimating the movement among staging areas of the barnacle goose (*Branta leucopis*). *Journal of Applied Statistics*, **22**, 711-724.

Schwarz, C.J. and Arnason, A.N. (1996). A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics*, **52**, 860-873.

Schwarz, C.J., Andrews, M., and Link, M.R. (1999). The stratified Petersen estimator with a known number of unread tags. *Biometrics*, **55**, 1014-1021.

Schwarz, C.J. and Seber, G.A.F. (1999). A review of estimating animal abundance III. *Statistical Science*, **14**, 427-456.

Schwarz, C.J. and Stobo, W.T. (1999). Estimation and effects of tag-misread rates in capture-recapture studies. *Canadian Journal of Fisheries and Aquatic Sciences*, **56**, 551-559.

Seber, G.A.F. (1970). The effects of trap response on tag-recapture estimates. *Biometrika*, **26**, 13-22.

Seber, G.A.F., Felton, R. (1981). Tag loss and the Petersen mark-recapture experiment. *Biometrika*, **68**, 211-219.

Seber, G.A.F. (1982). *The Estimation of Animal Abundance (2nd ed.)*. London: Griffin.

Seber, G.A.F. (1986). A review of estimating animal abundance. *Biometrics*, **42**, 267-292.

Seber, G.A.F. (1992). A review of estimating animal abundance II. *International Statistical Review*, **60**, 129-166.

Seber, G.A.F., Huakau, J.T. and Simmons, D. (2000). Capture-recapture, epidemiology, and list mismatches: two lists. *Biometrics*, **56**, 1227-1232.

Seber, G.A.F. (2001). Some new directions in estimating animal population parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, **6**, 140-151.

Smith, E.P. and van Belle, G. (1984). Nonparametric estimation of species rishness. *Biometrics*, **40**, 119-129.

Smith, P.J. (1988). Bayesian methods for multiple capture-recapture surveys. *Biometrics*, **44**, 1177-1189.

Smith, P.J. (1991). Bayesian analyses for a multiple capture-recapture model. *Biometrika*, **78**, 399-407.

Sprott, D.A. (1981). Maximum likelihood applied to capture-recapture model. *Biometrics*, **37**, 371-375.

Stanley, T.R. and Burnham, K.P. (1998). Information theoretic model selection and model averaging for closed-population capture-recapture. *Biometrical*

Journal, **40**, 475-494.

Swetnam, T.W. (1993). Fire history and climate change in giant sequoia groves. *Science*, **262**, 885-889.

Tardella, L. (2002). A new Bayesian method for nonparametric capture-recapture models in the presence of heterogeniety. *Biometrika*, **89**, 807-817.

Tsay, P.K. and Chao, A. (2001). Population size estimation for capture-recapture models with applications to epidemiological data. *Journal of Applied Statistics*, **28**, 25-36.

Underhill, L.G. (1990). Bayesian estimation of the size of a closed population. *Ring*, **13**, 235-254.

Wang, Y.-G. (1999). Estimating equations for parameters in stochastic growth models from tag-recapture data. *Biometrics*, **55**, 900-903.

Wittes, J.T. (1972). On the bias and estimated variance of Chapman's two-sample capture-recapture population estimate. *Biometrics*, **28**, 592-597.

Yip, P. (1989). An inference procedure for a capture and recapture experiment with time-depedent capture probabilities. *Biometrics*, **45**, 471-479.

Yip, P. (1991). A martingale estimating equation for a capture-recapture experiment in discrete time. *Biometrics*, **47**, 1081-1088.

York, J., Madigan, D., Heuch, I. and Lie, R.T. (1995). Estimating a proportion of birth defects by double sampling: A Bayesian approach incorporating covariates and model uncertainty. *Applied Statistics* , **44**, 227-242.

Zippin, C. (1956). An evaluation of the removal method of estimating animal populations. *Biometrics*, **12**, 163-169.