Extensions to the Multiplier Method for Inferring Population Size

by

Vivian Yun Meng

B.Sc. Applied Science, Queen's University, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Statistics)

The University Of British Columbia
(Vancouver)

August 2014

Abstract

Estimating population size is an important task for epidemiologists and ecologists alike, for purposes of resource planning and policy making. One method is the "multiplier method" which uses information about a binary trait to infer the size of a population. The first half of this thesis presents a likelihood-based estimator which generalizes the multiplier method to accommodate multiple traits as well as any number of categories (strata) in a trait. The asymptotic variance of this likelihood-based estimator is obtained through the Fisher Information and its behaviour with varying study designs is determined. The statistical advantage of using additional traits is most pronounced when the traits are uncorrelated and of low prevalence, and diminishes when the number of traits becomes large. The use of highly stratified traits however, does not appear to provide much advantage over using binary traits. Finally, a Bayesian implementation of this method is applied to both simulated data and real data pertaining to an injection-drug user population. The second half of this thesis is a first systematic approach to quantifying the uncertainty in marginal count data that is an essential component of the multiplier method. A migration model that captures the stochastic mechanism giving rise to uncertainty is proposed. The migration model is applied, in conjunction with the multi-trait multiplier method, to real-data from the British Columbia Centre for Disease Control.

Preface

This dissertation is original, unpublished, independent work by the author, V. Meng. Chapter 3, Appendix A, Appendix B, Appendix C, Appendix D and Appendix E contain work submitted to the Annals of Applied Statistics, titled "Inferring Population Size: Extending the Multiplier Method to Incorporate Multiple Traits with a Likelihood-Based Approach" (under review), co-authored with Prof. Paul Gustafson who provided guidance for the research which led to said journal submission.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank Prof. Paul Gustafson for his guidance in my research, and NSERC for financial support. Dr. Mark Gilbert and Travis Salway Hottes have been a major source of motivation for this research. I would also like to thank Dr. Lisa Johnston and H. Fisher Raymond for providing data that complements the theoretical analysis in Chapter 3. Lastly, thanks mom, dad, and Johnty for your continued support.

Chapter 1

Introduction

Population size estimation is a fundamental interest for ecologists and epidemiologists alike. This knowledge is used for resource allocation, program planning, and estimating disease incidence, amongst other important applications. Popular methods for estimating population size include: capture-recapture, "direct" survey, network scale-up, and the multiplier method. The capture-recapture method uses identified membership of a sampled individual across multiple lists to infer population size. It was originally developed for applications in ecology (Pollock et al., 1990), but applications for epidemiological purposes can be found in, as examples, Paz-Bailey et al. (2011) and Hook and Regal (1995). The "direct" survey method uses information on the size of a super-population which encompasses the target population together with the proportion of people belonging to the target population in the inference (Purcell et al., 2012; Lieb et al., 2011). The network scale-up method uses sampled individuals from a super-population whereby each respondent is asked to provide information on the size of his/her personal network and the number of his/her acquaintances belonging to the target population (Ezoe et al., 2012; Salganik et al., 2011). Lastly, the multiplier method (UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance, 2010; Johnston et al., 2013) uses knowledge of the exact count (marginal count) of people with a certain trait (N_t) in the target population and a sample proportion of this trait ($\hat{p}_t$) to infer the size of the target population, N, based on the simple relationship $p_t = N_t/N$. The World Health Organization has a publication that gives a good account of the

advantages and disadvantages of the various methods listed above (UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance, 2010). This document also points out the importance of considering data-availability and data-reliability in the selection of an appropriate estimation method.

The work documented in this dissertation is motivated by an application in collaboration with the British Columbia Centre for Disease Control (BCCDC), whereby estimating the size of a "hard-to-reach population" (Magnani et al., 2005) is of importance to public health. When inferring the size of hard-to-reach populations, extra constraints are imposed in the method selection. It is often noted (Fendrich et al., 1999; Colón et al., 2001; Delaney-Black et al., 2010) that members of the hard-to-reach population do not self-report on their behaviour readily which results in biased estimates; as a result, the reliability of direct survey and network-scale-up methods is compromised due to relying on self-reporting of membership in the hard-to-reach population in a survey of a super-population. As a second constraint, privacy regulations in Canada often result in data free of personal identifying information, which makes capture-recapture studies difficult to implement for human populations. In contrast, the multiplier method is unaffected by the two constraints above.

The highly popular multiplier method was, however, developed at a time when data were hard to come by. The method prescribes a way to estimate population size based on one binary trait only, whereas recently, data on the marginal counts and sample prevalences for *numerous* health related traits are available to public health agencies (Okal et al., 2013; Raymond et al., 2013; Johnston et al., 2013). Without a statistically sound prescription to incorporate data from multiple traits, researchers resort to constructing multiple estimates of N with the multiplier method, one from each trait, while using sample prevalences captured in a *single* survey. This attempt is statistically unsatisfactory because it ignores the correlation between multiple estimates derived from the same survey. Additionally, it does not result in a unified statistical conclusion about the size of the target population.

Furthermore, the importance of accounting for uncertainty in the marginal counts has rarely been discussed in literature. This uncertainty may be attributed to various mechanisms, yet previous attempts to capture uncertainty on marginal counts have resorted to using a simple parametric distribution based on the intuition

of the practitioner (Archibald et al., 2001; Johnston et al., 2013). In these instances, a scientific evaluation of the quality of the selected distribution is difficult when the rationale for selection is not formulated explicitly based on stochastic mechanisms.

These serious problems were encountered in collaborative work with the British Columbia Center for Disease Control (BCCDC), and motivated the methodologies presented in this thesis. These methodologies advance the multiplier method toward a statistically correct utilization of data on multiple traits based on a popular data collection scheme, and a tractable formulation of uncertainty in marginal counts; this is approached using a likelihood-based and model-based extension of the multiplier method. As an additional feature, the new methodology now allows categorical traits in the analysis without having to collapse them into binary categories as before. This thesis contains work that can impact the practice of epidemiology immediately and widely, given current interests in the epidemiology literature on estimating population with the multiplier method using multiple traits.

Chapter 2

Background

The original multiplier method developed out of the epidemiology literature, with little use of statistical language in its definition even in an authoritative reference document by the UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance (2010). The method has acquired many aliases, including the service multiplier method and the indirect method. The method relies on the following definition for a *trait proportion*:

$$p_t = \frac{N_t}{N} \tag{2.1}$$

where N, N_t, p_t are as defined in Chapter 1. If N_t and p_t are known exactly, a rearrangement of Equation 2.1 leads to the exact population size, with no need for estimation.

However, neither the exact proportion nor marginal count of people with a particular trait is known exactly in the majority of cases and so are inferred from data. A survey is conducted within the target population to obtain the sample trait prevalence. While the target population has no defined sampling frame, sampling designs exist, e.g. venue-based sampling (Muhib et al., 2001) or respondent-driven sampling (Heckathorn, 1997), to give estimators of prevalences that are unbiased under certain mathematical assumption and adjustments.

As for the marginal count for a trait, it is inferred from pre-existing records in public health agencies. These records provide a list, and hence a head-count, of people with the trait in question. The raw head-count may differ from the actual

marginal count of a trait in the target population if, for example, trait misidentification occurs during record-keeping (Archibald et al., 2001) or if the population "catchable" by the record keeping procedure is not the same as the target population in question, as noted by UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance (2010). However, the raw head-count is commonly accepted as an estimate of the marginal count without formal justification.

Nevertheless, a substitution of p_t and N_t with their estimates in Equation 2.1 gives an *estimate* of N. While the majority of studies consulted do not calculate uncertainty on the estimated population size (Raymond et al., 2013; Okal et al., 2013; Luan et al., 2005), the papers by Archibald et al. (2001) and Johnston et al. (2013) give two approaches to generating the confidence interval for the following estimator,

$$\hat{N} = \frac{\hat{N}_t}{\hat{p}_t}. \tag{2.2}$$

Archibald et al. (2001) treats $\hat{N}$ as a function of two random variables, and uses a Monte Carlo simulation to obtain the distribution of $\hat{N}$. Archibald assumed a Normal distribution for both $\hat{N}_t$, based on ad hoc *a priori* information, and $\hat{p}_t$, based on information from survey sampling.

Johnston et al. (2013) obtains the confidence interval for the estimator in Equation 2.2 by assuming its convergence to the Normal distribution. The variance is approximated by the Delta Method (Taylor series expansion),

$$Var(\hat{N}) \approx \frac{Var(\hat{N}_t)}{\mathbb{E}[\hat{p}_t]^2} + \frac{\mathbb{E}[\hat{N}_t]^2}{\mathbb{E}[\hat{p}_t]^4} Var(\hat{p}_t).$$

The 95% confidence interval is hence

$$95\%CI = \hat{N} \pm 1.96 Var(\hat{N}).$$

Both methods in (Archibald et al., 2001) and in (Johnston et al., 2013) deal with quantifying the statistical uncertainty from estimating population size with the multiplier method using a *single binary trait*. However, these formulas 1) do not extend naturally to estimation with multiple traits , and 2) lack a systematic way to specify the uncertainty in the inferred marginal count. The following subsections

highlight each of these problems in detail.

2.1 Problem with using multiple traits in the multiplier method

In recent studies, epidemiologists have reported access to information on a vast number of "traits" with which to carry out their estimation. Generally there is accepted intuition that using more information provides better estimates. Such belief has lead epidemiologists to synthesize k estimates of the same population size using information from k traits, i.e.

$$\begin{aligned}\hat{N}_1 &= \frac{\hat{N}_{t_1}}{\hat{p}_{t_1}},\\ \hat{N}_2 &= \frac{\hat{N}_{t_2}}{\hat{p}_{t_2}},\\ &\vdots\\ \hat{N}_k &= \frac{\hat{N}_{t_k}}{\hat{p}_{t_k}},\end{aligned}$$

see (Raymond et al., 2013) for example. Should estimators $\hat{N}_1, \cdots, \hat{N}_k$ be independent, standard procedures, e.g. inverse variance weighting, exist to combine them into a single best estimate of N.

However, these estimates are not independent in the studies involving multiple traits, due to a data collection scheme commonly adopted for practicality. In these studies, a single survey querying k traits is conducted, such that $\hat{p}_{t_1}, \cdots, \hat{p}_{t_k}$ are estimated from a single sample of the target population. Currently, no statistical method exists to account for dependency that arises from the particular design for surveying the population only once for multiple traits.

2.2 Problem with capturing uncertainties in marginal counts

The uncertainty in marginal counts may be due to several important factors, e.g. population migration ("mobility" in Saidel et al. (2010)), record under-counting/over-

counting due to various reasons, that occur in conjunction or in isolation. In the past, uncertainty in inferred marginal counts is often specified in an ad hoc manner. As an example, Johnston et al. (2013) treated the inferred marginal count as variable according to a single parameter Poisson distribution with mean value equal to the raw head-count, without further justification. Archibald et al. (2001) assumes inferred marginal count to be Normally distributed with mean value equal to the raw head-count, and lower and upper 95th percentile value determined based on "the extent of duplicates, risk-category misclassification together with [their] subjective assessment of the likely magnitude of uncertainty." In either case, it is difficult to support nor rebut the parametric distributions selected when they are chosen in an ad hoc manner. A systematic way to describing the mechanism may be the key toward better assessing model misspecification.

2.3 Thesis organization

This thesis makes two contributions to the statistical literature on the multiplier method. It firstly outlines a likelihood-based approach to address current methodological limitations due to the "one survey multiple traits" design in Chapter 3, and secondly presents the first systematic approach to capture uncertainty in marginal counts in Chapter 4. However, the scope of Chapter 4 is limited to addressing a single mechanism, population migration, as the source of uncertainty in inferred marginal counts, which is motivated by the needs of a collaboration in progress. This thesis concludes with remarks on limitations and future work in Chapter 5.

Chapter 3

Extended Multiplier Method for Multiple Traits

In this chapter a likelihood-based method is developed for estimating the size of a population with multi-trait data that is currently inadequately addressed by the single-trait multiplier method due to employing a particular data collection design. A likelihood-based estimator is presented in Section 3.1 to represent the design with which the data are obtained. Properties of this estimator are explored in Section 3.2 under varying study design parameters. In Section 3.3, considerations for Bayesian inference with the proposed likelihood are outlined. A real data example of Bayesian inference with the proposed likelihood is shown in Section 3.4. Finally, a discussion on the advantages and limitations of the method developed in this chapter is found in Section 3.5.

Table 3.1: Contingency table for two traits, binary categories

		X_2		
		absent (0)	present (1)	
X_1	absent (0)	n_{00}	n_{01}	
	present(1)	n_{10}	n_{11}	
				n

3.1 A Likelihood Model for Estimating N

3.1.1 The Reparameterized Multinomial Likelihood Model for Modelling Two Binary Traits

Data and Assumptions

In the case that information on two binary traits, X_1 and X_2, are collected for the purpose of inferring population size, the following likelihood is suitable for data resulting from a data collection scheme that results in

- a 2 by 2 contingency table (Table 3.1) that cross-classifies respondents of a sample survey based on two traits,
- marginal count $N_{1\cdot}$, the number of people in the target population with $X_1 = 1$,
- marginal count $N_{\cdot 1}$, the number of people in the target population with $X_2 = 2$.

It is assumed that no personal identifying information is available where the above data come from due to privacy regulations. Secondly, the target population is assumed *closed* without births, deaths, or migrations. The marginal counts are assumed to be *exact*, i.e. known without uncertainty – this restriction is relaxed in Chapter 4. The sample survey is obtained through simple random sampling without replacement (SRSWOR). For target populations without a sampling frame, the assumption of sampling through SRSWOR cannot be met, but is used here to simplify the theoretical development. Application of the method in violation of SRSWOR assumption is addressed in Section 3.4 and Section 3.5.

The above data may at first appear equivalent to multiple "lists" that can be analysed with capture-recapture methods but this is not so, as persons appearing in each data source may not be cross-identified due to privacy regulations – one of many reasons why such data ought not be analysed with capture-recapture (Pollock et al., 1990) or similar methods (Olkin et al., 1981) of analysis (see Appendix A for details).

A likelihood for the stochastic mechanism

The mechanism that gave rise to the data at hand is analogous to drawing marbles out of a bag where the total number of marbles in the bag, N, is unknown. Each marble in the bag may be either clear, red, sparkly, or red and sparkly. Additionally, we know exactly the number of marbles that are red, and the number of marbles that are sparkly in the bag. Finally, a *random sample* of n marbles are drawn without replacement, their traits observed and recorded in a contingency table.

For this process, a multinomial likelihood is appropriate for approximating the distribution of the sample frequencies of every trait combinations obtained via SRSWOR. The usual multinomial likelihood has parameters $p_{00}, p_{01}, p_{10}, p_{11}$ that determine the frequencies observed in Table 3.1. Note that only three of these parameters are free to vary in the support space because the parameter space is constrained by $\sum p_{ij} = 1$.

Additionally, knowledge of marginal counts $N_{1\cdot}$ and $N_{\cdot 1}$ presents extra information that further constrains the support space. Necessarily in the overall target population,

$$\frac{N_{1\cdot}}{N} = p_{10} + p_{11}, \tag{3.1}$$

$$\frac{N_{\cdot 1}}{N} = p_{01} + p_{11}. \tag{3.2}$$

The additional constraints on the support are incorporated by reparameterizing via $p_{10} = N_{1\cdot}/N - p_{11}$, and $p_{01} = N_{\cdot 1}/N - p_{11}$. The reparameterized likelihood of survey data is, given fixed marginal counts,

$$\mathscr{L}(p_{11}, N|data) =$$
$$\frac{(\sum n_{ij})!}{n_{00}!n_{10}!n_{01}!n_{11}!} p_{11}^{n_{11}} \left(\frac{N_{1\cdot}}{N} - p_{11}\right)^{n_{10}} \left(\frac{N_{\cdot 1}}{N} - p_{11}\right)^{n_{01}} \left(1 - \frac{N_{1\cdot} + N_{\cdot 1}}{N} + p_{11}\right)^{n_{00}}, \tag{3.3}$$

where the statistics n_{ij} = the number of people in the survey with status i for the first trait, and status j for the second trait.

As a consequence of the above reparameterization, the population size, N, has become a parameter in the likelihood (Eq.(3.3)). Estimation of N can be carried out via several standard techniques, e.g. maximum likelihood or Bayesian inference. As a side note, this reparameterized multinomial likelihood must not be confused with the multinomial models in capture-recapture inspired methodology – the stochastic mechanisms described by these methods are fundamentally different (Appendix A).

3.1.2 Generalizing the Reparameterization for k Traits

Let a set of k traits be denoted by $\{X_1, \ldots, X_k\}$. A trait X_i has l_i categories (strata), coded from 0 to $l_i - 1$. Throughout this thesis, a trait with more strata than another is referred to as being more "stratified". Let stratum 0 always represent the *absence* of a trait. As an example, if "positive test result for disease a" is a trait of interest, then stratum 0 necessarily corresponds to the absence of a positive result and stratum 1 to stratum $l_i - 1$ may be defined to represent a positive result diagnosed in different time periods.

As data, a k-dimensional contingency table is observed. The contingency table is populated by cross-classifying a sample of people from the target population based on the status of k traits. A set of marginal counts is also known to us. Let the observed marginal count be denoted by M_{is}, where $i \in \{1, \ldots, k\}$ indexes the trait, and $s \in \{1, \ldots, l_i - 1\}$ specifies the stratum/sub-type of this trait. Let $\boldsymbol{M}$ denote the collection of marginal counts observed, where the size of the collection, $|\boldsymbol{M}|$, is $[\sum_{i=1}^{k} (l_i - 1)]$.

In this case the development of the reparameterized likelihood follows a similar

procedure to that in the previous section. Beginning with the standard multinomial likelihood, the parameters of this multinomial likelihood are reparameterized based on $|\boldsymbol{M}|$ known marginal counts using equations similar to Eq.(3.1) and Eq.(3.2) to incorporate constraints on the support space. A systematic reparameterization scheme has been chosen, which rearranges each marginalization equation in the form

$$\frac{M_{is}}{N} = \sum_{j_1=0}^{l_1-1} \cdots \sum_{j_{i-1}=0}^{l_{i-1}-1} \sum_{j_{i+1}=0}^{l_{i+1}-1} \cdots \sum_{j_k=0}^{l_k-1} Pr\{X_1 = j_i, \ldots, X_{i-1} = j_{i-1}, X_i = s, X_{i+1} = j_{i+1}, \ldots, X_k = j_k\} \tag{3.4}$$

for the probability of having only 1 trait.

The exact expression of the reparameterized likelihood is found in Appendix C. Note that the resulting reparameterized likelihood will have $\{[(\prod_{i=1}^{k} l_i) - 1] - |\boldsymbol{M}| + 1\}$ free parameters. The first term, $[(\prod_{i=1}^{k} l_i) - 1]$, is the number of free parameters of a standard multinomial likelihood. The second term, $-|\boldsymbol{M}|$, is a reduction in free parameters equal to the number of observed marginal counts. The last term, $+1$, results from the inclusion of N as a parameter in the reparameterization process. With N being a parameter of the likelihood, its estimation may be carried out using standard likelihood-based methods.

3.2 Uncertainty about the Maximum Likelihood Estimator of N

The following section assesses the uncertainty in the proposed estimator (MLE) obtained from maximum likelihood theory. The variance of the MLE based on the reparameterized likelihood is approximated asymptotically from the Fisher Information, the derivation of which is documented in Appendix C.

An important outcome from this section is understanding how multiple traits may best be used to provide a desired level of precision for the population size estimate from a study-design perspective. Thus, the Fisher Information is used to

explore the behaviour of the estimation uncertainty, $SD(\hat{N}_{MLE})$, as important parameters and design factors are varied. In specific, the parameter N is investigated along with the following design factors: sample size, the number of traits used, and the level of trait stratification. The effects of trait prevalence and strength of trait association are also explored. Note that "trait prevalence" and "strength of trait association" are two functions of the cell probability parameters in the reparameterized multinomial likelihood. They are chosen for investigation instead of cell probability parameters due to better interpretability and better chance that *a priori* information exist about them. While they are not entirely controllable by the study designer, the use of *a priori* information about prevalence and association leads to better design selection in light of the results shown in the following sections. Hence, in this thesis, any reference to the set of design factors will also include prevalence and trait association.

3.2.1 The Effect of n and N on Estimation Uncertainty

The SWSWOR assumption on the survey data implies that $SD(\hat{N}_{MLE}) \propto 1/\sqrt{n}$, a well-known classical result. The derivation of the effect of N is, however, non-trivial and thus reserved for Appendix C. There it is shown that $SD(\hat{N}_{MLE}) \propto N$ when all other design parameters are kept fixed. An extension of this result with the Delta Method shows $SD(\log \hat{N}_{MLE})$ to be unaffected by the size of N as well. The proportionality between population size and estimation uncertainty allows the effect of design parameters to be studied on the scale of relative error ($SD(\hat{N}_{MLE})/N$), or equivalently, $SD(\log \hat{N}_{MLE})$, in the sections to come.

3.2.2 The Effect of Additional Traits of Given Prevalence and Degree of Association

In this section, designs with varying number of traits, prevalence and correlation are mapped to cell probabilities to obtain the Fisher Information. This exploration is restricted to binary traits of equi-prevalence and equal pairwise correlation (equi-correlation). The equi-prevalence, equi-correlation restriction is used to reduce the complexity of the design parameters under exploration, and may be extrapolated to represent general situations– this is justified Appendix B. Appendix B also pro-

vides guidance to benchmarking estimator precision from a study-design perspective using the equi-correlation, equi-prevalence assumption. Note that situations beyond binary stratification are explored in the next section.

Figure 3.1 provides a summary of the effect of increasing the number of binary traits, prevalence and correlation. One makes the observations that (a), increasing the number of traits has a beneficial but diminishing effect for lowering the uncertainty, (b), the addition of traits is most effective at lowering uncertainty when traits are not correlated, (c), using more prevalent traits is effective for lowering the uncertainty, and (d), when the prevalence is high ($\approx 10\%$), the impact of additional traits becomes negligible.

3.2.3 The Effect of Increased Stratification

Stratification, as previously mentioned, refers to the fineness of the categorization for a trait. When the stratification is beyond binary, it is challenging to map a given set of design factor values to a point in the parameter space because trait prevalence and trait association are typically defined for binary traits (indicators). Instead, a Monte Carlo, generative approach is taken to explore the parameter space directly for a given level of stratification. One will see shortly that a *distribution* for the estimation uncertainty results for a given set of design parameters based on this approach.

This Monte Carlo simulation is motivated by the situation where a subject-area expert may have some expectation of the strength of association and prevalence for a set of binary traits but is unsure if obtaining further detailed information (stratification) for each trait will provide useful information to lower estimation uncertainty, given there is no knowledge concerning what the stratification will look like. To provide practical advice to this hypothetical subject-area expert, test scenarios were generated randomly in a way that captures the variability in the estimation uncertainty assuming *a priori* any stratification is equally likely to happen.

Let us begin by assuming a set of two binary traits as the starting point for exploring traits with three strata. Take a population with two binary traits of given prevalence and association. This specification translates to some cell probability parameters when sampling and observing trait combinations with SRSWOR. When

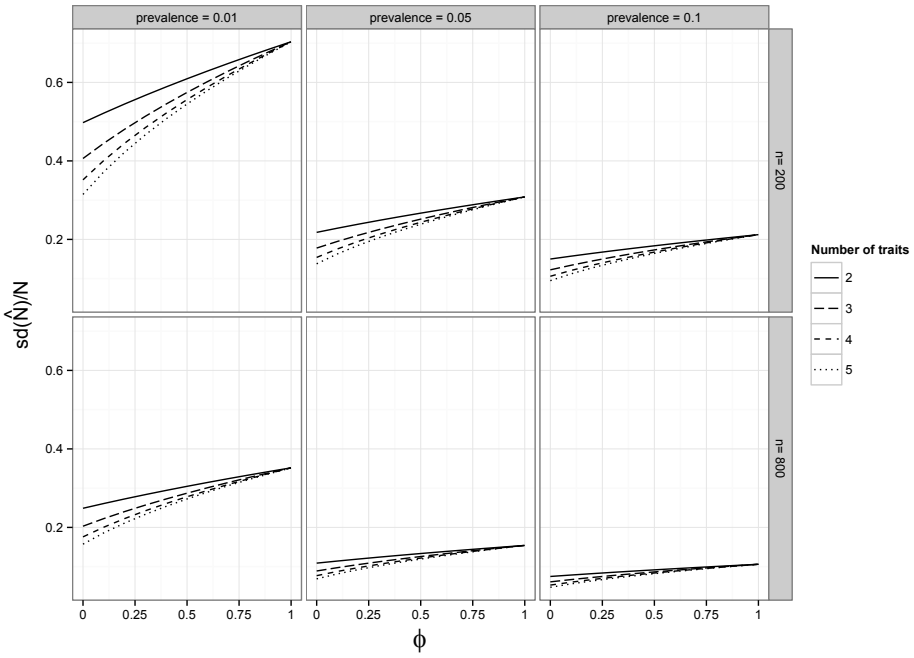

Figure 3.1: Examining the effect of increasing the number of binary traits with changing prevalence and association on the relative error of estimating N. The association between traits is measured by ϕ, ranging from 0 (no association) to 1 (complete association). One makes the observation that the relative error *decreases* with (1) increasing number of traits (2) increasing trait prevalence and (3) decreasing association between traits. The advantage of additional traits disappears as traits become completely associated, or as prevalence becomes high. One may further note that the relative error doubles as n is decreased by 4 folds, reflecting its inversely proportional relationship with $\sqrt{n}$.

the binary category for "trait present" is stratified into two mutually exclusive categories, each cell probability parameter, pre-stratification, is necessarily equivalent to the sum of a set of d parameters post-stratification (see Figure 3.2 for an illustration). One may see that the mapping from pre-stratification cell probabilities to post-stratification cell probabilities is not unique.

Thus in the Monte Carlo experiment, to generate one possible instance of stratification, a sample from the d-dimensional Dirichlet(1,...,1) distribution is drawn to

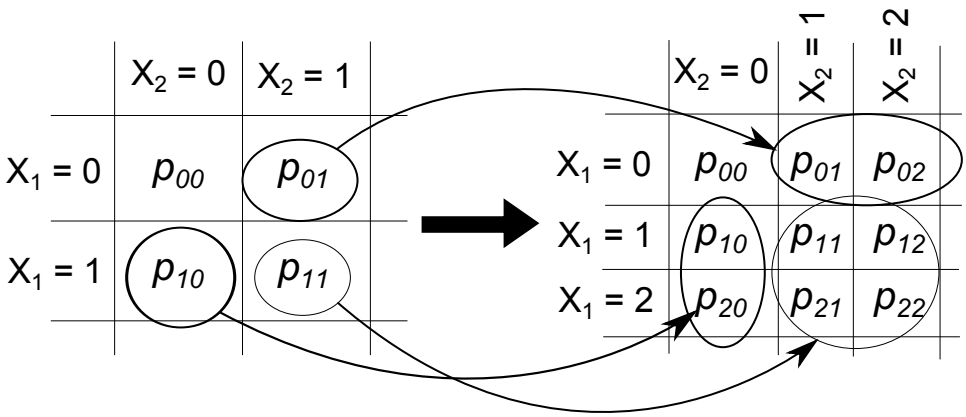

Figure 3.2: Showing how binary traits are further stratified in the simulation in Section 3.2.3.

provide the weights with which a cell probability, pre-stratification, is distributed into the post-stratified cells. Estimator error is then calculated using the post-stratified cell probabilities. This process is repeated 100 times for a given set of binary traits to characterize the variation in estimator error in the space of possible scenarios for stratification. A similar process is applied to k starting binary traits of equi-prevalence and equi-correlation up to $k = 4$ traits.

The result of this simulation study is shown in Figure 3.3, where each box-plot summarizes the estimation uncertainty of all randomly generated scenarios that correspond to the same starting k binary traits. Generally speaking, the reduction in estimation uncertainty from increased stratification is small compared with the reduction due to an increased number of traits, given an uninformative *a priori* belief about stratification. While there are a few special cases where stratification provides reduction comparable to that from additional traits, it is difficult for the subject-area expert to evaluate if a set of traits of interest might constitute such a special case.

3.3 Bayesian Inference

In recent years, both in the research of statistical methodologies and in subject-area applications, Bayesian inference has grown to be increasingly useful and popular. Bayesian estimation has several strengths for inferring population size with the

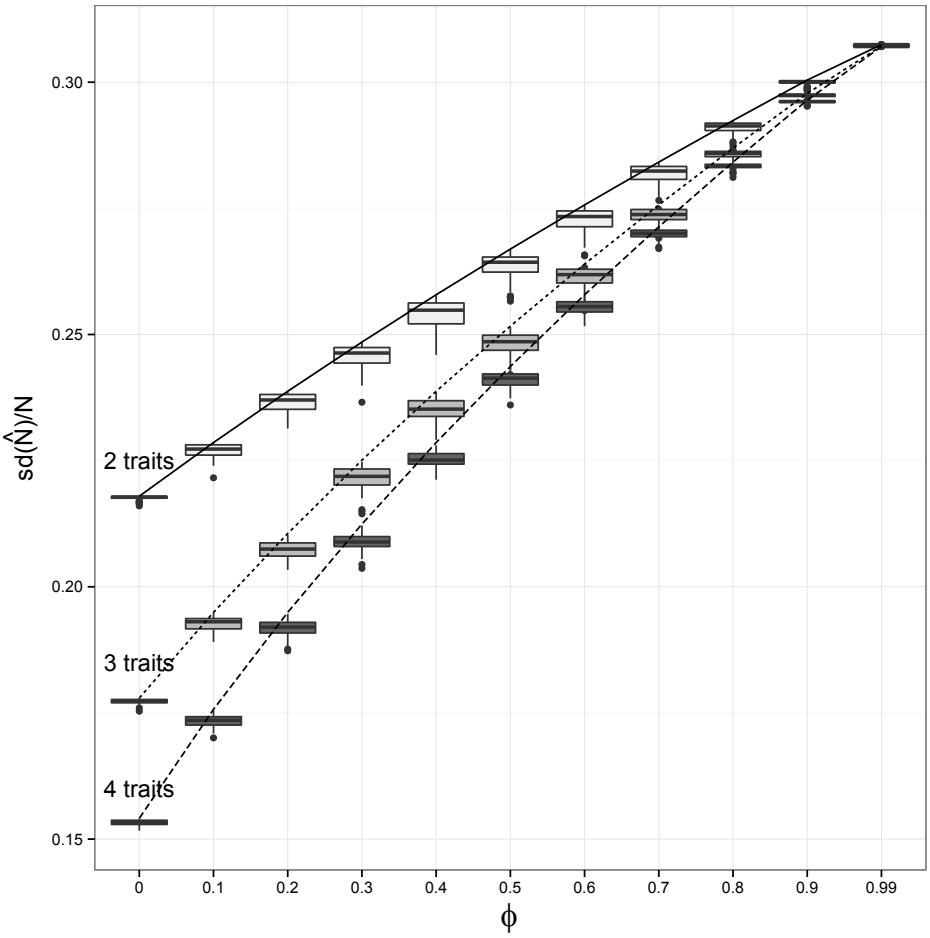

Figure 3.3: The effect of stratification: smooth curves show estimator uncertainty obtained with binary traits with varying strength of association. The boxplots show the empirical distributions of estimator uncertainty when binary traits are increased to three strata at varying strengths of association from $\phi = 0$ to $\phi = 1$ in increments of 0.1. Given k traits at some the level of association (along the x-axis), the vertical difference between the location of the "box" and the curve for shows the effect of stratification. Compare this with the vertical difference between curves, which marks the effect of increased number of traits. The comparison shows that stratification has a small effect for reducing estimation uncertainty relative to the effect of using more traits.

reparameterized multinomial model described in this chapter. Firstly, one is likely to observe a sparse contingency table, based on surveying the absence/presence of multiple traits in a population, should the trait prevalences be low. This sparsity poses a problem in MLE-based inference but not in Bayesian inference. Secondly, with limited computational expertise, Bayesian inference is easier to implement than the MLE for complex models are specified through standard distributions and simple deterministic functions due to availability of many MCMC software packages. Lastly, while the Bayes estimator in theory agrees asymptotically with the MLE as $n \to \infty$, in subject-area applications, a Bayes estimator may achieve a lower level of uncertainty compared with the MLE when the use of informative priors can be justified.

However, one recognizes the challenge to find an easy to implement, valid joint-prior for a large set of parameters in a constrained space. The following paragraphs outline some suggestions on an appropriate prior and its implementation. The proposed prior is evaluated with a simulation study and is shown to have good estimation properties (Appendix E).

The key requirement of a suitable prior is that it allows expert knowledge to enter the analysis in a natural way. The set of parameters in the reparameterized likelihood function is, according to Appendix C, $\boldsymbol{\eta} = \{\boldsymbol{\theta}, N\}$. Rather than defining a prior $\pi(\boldsymbol{\theta}, N)$ multivariately, factorizing the joint-prior into smaller modules results in a simpler implementation and is easier to interpret. Two factorizations are considered: $\pi(N) \times \pi(\boldsymbol{\theta}|N)$ and $\pi(\boldsymbol{\theta}) \times \pi(N|\boldsymbol{\theta})$. Substantive prior knowledge of $\pi(\boldsymbol{\theta})$ is *unlikely* to exist because this requires knowledge of how the traits are associated with each other and prior knowledge of every cell probability, and for this reason the use of $\pi(\boldsymbol{\theta}) \times \pi(N|\boldsymbol{\theta})$ is not pursued. On the other hand, knowledge about $\pi(N)$ is relatively easy to elicit from an expert in the form of perceived natural limits to population sizes. Thus the formulation $\pi(N) \times \pi(\boldsymbol{\theta}|N)$ is a useful choice for carrying out Bayesian inference.

In the case of data consisting of two binary traits ($k = 2$, $l_1 = l_2 = 2$), recall the notations in Section 3.1.1. One possible formulation of the prior with factorization

$\pi(p_{11}|N) \times \pi(N)$ is to let

$$N \sim Uniform(L,U), \tag{3.5}$$

$$p_{11}|N \sim Unif(0,b_N), \tag{3.6}$$

where $L \geq \max(N_{\cdot 1}, N_{1\cdot}, n)$, and U is the upper bound on the possible size of N reflecting the range of possible values *a priori*, $b_N = \min(N_{\cdot 1}/N, N_{1\cdot}/N)$ – the smallest value appearing in the set of marginal probabilities. Further details on the exact implementation of these priors can be found in Appendix D.

In the case of inference with three or more binary traits, once again, the notation $(\{\boldsymbol{\theta}, N\})$, as given by Appendix C, is adopted for the set of parameters. Let $N \sim Uniform(L,U)$ with L,U defined as in the case of two binary traits. For the parameter $\boldsymbol{\theta}$, a uniform prior prior is placed in the constrained parameter space given N, that is, $\pi(\theta|N) = 1/V_N$. The quantity V_N can be interpreted as the "volume" of the constraint space for $\boldsymbol{\theta}$ defined by the inequalities $0 < p_c < 1$; here p_c is a cell probability and a function of parameters $\boldsymbol{\theta}$ and N. Appendix D lists the details for implementing this prior for inferring N with three binary traits.

3.4 Application: San Francisco Injection-Drug User Study

A recent study using the multiplier method to infer the population size of San Francisco Injection-Drug Users (IDUs) in 2009 had been documented in Johnston et al. (2013). In this study, information about a number of traits, ranging from service usages to disease diagnoses, were used to construct several estimates of population size via the traditional multiplier method, without taking into account correlation between traits. Note that a single Respondent-Driven Sampling (RDS) survey was conducted that asked each respondent to self-report on all traits.

An illustration of the likelihood-based multi-trait multiplier method shall be made with with two binary traits – usage of a particular substance use treatment center (Walden House) and status of being a reported HIV case. Table 3.2 shows, for each trait, the marginal count, the RDS adjusted prevalence estimate, and population size estimate made with the Multiplier Method. No respondent indicated

Table 3.2: Marginal count and estimated prevalences of two traits in the IDU population in San Francisco in 2009. The table also lists two estimates on the size of IDU population obtained by applying the Multiplier Method to each trait.

Trait	Marginal Count	$\hat{p}_{RDS}$	$s.e.(\hat{p}_{RDS})$	$\hat{N}_{Multiplier}$	95% C.I.
Using Walden House	104	2%	0.008	5,200	1,003 - 9,398
Reported HIV cases	3,308	7.3%	0.017	45,315	24,576 - 66,057

to being a reported HIV case and also using Walden House, based on a personal communication with the authors of Johnston et al. (2013).

3.4.1 Obtaining a Single Estimate of the Size of IDU Population Using Two Traits

To estimate the population size of IDU population in San Francisco in 2009 (N) with two traits, a contingency table that cross-classifies the survey data according to usage of Walden House and reported HIV status is required, on top of the marginal counts appearing in Table 3.2. However, pseudo-data, synthesized based on the raw data with some adjustments, is used in place of a raw contingency table of the RDS sample, to account for one unmet requirement that the survey be carried out through SRSWOR. The RDS design differs from SRSWOR in having a selection bias and a reduced efficiency; hence, two adjustments are built in to the pseudo-data.

To adjust for bias in RDS sampling, unbiased RDS adjusted proportion estimates is used to construct the pseudo-data. Let the answer to a survey question be coded 1 for "yes" and 0 for "no", and let $\hat{p}_{ij_{RDS}}$ indicate the RDS adjusted estimate for the proportion of IDUs who answer i for using Walden House and answer j for being a reported HIV case. Given that no respondents in the survey indicated being *both* a reported HIV case and a user of Walden House, $\hat{p}_{00_{RDS}}$ must be zero – the rest of the RDS-adjusted proportion estimates are calculated using this fact together with the RDS-adjusted marginal prevalences in Table 3.2. Table 3.3 shows the RDS adjusted sample proportions as per calculation.

To adjust for the RDS sampling being less efficient that SRSWOR, the effective sample size (ESS) is used in constructing the pseudo-data. ESS represents the

Table 3.3: RDS-adjusted proportions in the contingency table for IDU's in San Francisco in 2009.

	Reported HIV case = no	Reported HIV case = yes
Uses Walden House = no	90.7%	2%
Uses Walden House = yes	7.3%	0%

Table 3.4: Pseudo-data: adjusted contingency table for estimating the size of IDU population in San Francisco in 2009 with the proposed method of analysis.

	Reported HIV case = no	Reported HIV case = yes
Uses Walden House = no	245	5
Uses Walden House = yes	20	0

sample size required of a SRSWOR survey to yield the same estimation uncertainty as the given RDS survey, and is defined (Kish, 1965) in the case of estimating a proportion as

$$ESS = \frac{n}{D_{eff}} = \frac{n}{(var(\hat{p})_{RDS}/var(\hat{p})_{SRSWOR})} = \frac{p(1-p)}{var(\hat{p})_{RDS}}. \tag{3.7}$$

The ESS is estimated by substituting the RDS-adjusted prevalence estimate and its standard error for each trait into Eq. (3.7). This results in two estimates of ESS, 234 and 306, that are similar in magnitude; an average value ($\hat{ESS} = 270$) is used to construct the pseudo-data.

Finally, the pseudo-data consists of "cell counts", $n^*_{ij} = ESS \times \hat{p}_{ij_{RDS}}$, rounded to the nearest whole number. The adjusted contingency table containing this pseudo-data is shown Table 3.4.

A Bayesian approach is taken to infer the size of an IDU population in San Francisco in 2009. Note that the presence of a 0 count in the pseudo-data poses no difficulty for the use of Bayesian inference. Following the recommendations in Section 3.3, a prior distribution where $N \sim Unif(3308, 1e10)$ and $p_{11}|N \sim Unif(0, 104/N)$ is chosen to reflect the lack of prior knowledge about N, given the marginal counts of 104 and 3308. The posterior distribution of $\log N$ is obtained via Markov Chain Monte Carlo (MCMC) in the software JAGS. To set-up MCMC, the

Table 3.5: Analysis of plausible scenarios with proposed method. Each row contains hypoethetical pseudo-data that fit the required ESS and estimated marginal prevalences (2% and 7.3%), followed by an estimated population size based on analysing this pseudo-data.

$n^*_{00}; n^*_{01}; n^*_{10}; n^*_{11}$	$\hat{N}_{Bayes}$	95% CI
245; 5; 20; 0	38966	[27059, 58561]
246; 4; 19; 1	40511	[27843, 61688]
247; 3; 18; 2	42206	[28795, 64631]
248; 2; 17; 3	44078	[29828, 68462]
249; 1; 16; 4	46180	[30898, 72729]
250; 0; 15; 5	48430	[32023, 77213]

model-file appearing in Appendix D, the marginal counts and the pseudo contingency table are supplied to the software. The MCMC is run for a total of 500,000 iterations and demonstrates good convergence. The population size is estimated with $\hat{N}_{Bayes}$ defined as the exponentiated posterior expectation of $\log N$. The 95% CI is obtained through exponentiating the HPD 95%CI for the posterior of $\log N$. Based on this analysis, the population size of IDU in San Francisco in 2009 was 38,966 people with a 95% CI of [27059, 58561].

3.4.2 Alternative Analysis with Marginal Prevalences Only

In the above application, extraneous information beyond what is readily available in the published study is required in order to complete the analysis. However, there is merit in considering the likelihood-based multi-trait multiplier method even if only marginal prevalences and marginal counts are accessible, i.e. the information contained in Table 3.2. To demonstrate, Table 3.5 lists, in each row, pseudo-data with an ESS of 270 that fit the RDS-adjusted marginal prevalences of 7.3% and 2% along with an estimate of N made by analysing that pseudo-data with the reparameterized likelihood method to combine multiple traits. One may contrast the estimates in Table 3.5 with that of the original study in Table 3.2; by attempting the analysis in a "what-if" approach, one may still obtain a sense of what combining multiple traits says about N that is otherwise difficult to replicate with intuition and marginal estimates alone.

3.5 Discussion

This chapter outlines a likelihood based method for estimating population size, N, based on extending the multiplier method to multiple traits. This is achieved through reparameterizing the multinomial likelihood to incorporate N as a parameter. The statistical advantage of using additional traits, of varying association and prevalence, is explored by comparing the asymptotic standard deviation of the estimator for N. This understanding of how different parameters affect the efficiency is useful to subject-area experts for obtaining the desired precision on N through a well-planned study design. One may find in Appendix B some suggestions regarding the approach to the design process. Figure 3.1 may also serve as a starting point for visually extrapolating design values that provide the required estimation precision.

In many applications, Bayesian inference presents important advantages over maximum likelihood methods, for example when the observed data table contains a 0 frequency or when prior knowledge exists. A joint prior outlined in this chapter emphasizes considering the prior knowledge on population size *marginally*. One may find in Appendix D details on how to obtain the posterior distribution in convenient MCMC software with the recommended prior. Appendix E documents a simulation study which verifies the performance of estimating population size using this recommended Bayesian approach.

An application of the multi-trait multiplier method is provided which estimates the population size of IDUs in San Francisco in 2009, based on data published in Johnston et al. (2013). The data requirement outlined in Section 3.1.1 is translated to this particular context. Note that, in practice, the survey data may require adjustments as in this application when it is not collected through SRSWOR.

As mentioned in the introduction, this chapter addresses the gap in methodology for combining multiple correlated estimates arising from the multiplier method. Certainty one may approach this problem in other ways, e.g. by minimizing the variance of linear combination of correlated estimators. A likelihood-based approach is chosen because not only does it accomplish the original task of combining multiple estimates with asymptotically optimal properties, it also provides an easy transition to the Bayesian framework. A likelihood approach further en-

ables simple model extensions to deal with problems that arise with real data. As an example, the concern of marginal counts being inexact was noted in (Johnston et al., 2013). In this situation, the reparameterized likelihood model can be easily extended to model population migration, or other mechanisms that causes uncertainty in the marginal counts; the subsequent chapter addresses this issue in depth.

Chapter 4

Uncertainty in Marginal Counts – Accounting for Migration

This chapter presents a systematic approach to accounting for uncertainty in marginal counts that result from population migration through the use of a migration model. The migration mechanism is motivated by an application of the multiplier method to data collected with the disease surveillance system employed by the British Columbia Center for Disease Control (BCCDC) and the Public Health Agency of Canada (PHAC). The model outlined in this chapter may be generalized to countries with a similar model of disease monitoring for extended applications.

Central to the problem of migration is the requirement of defining a "closed" target population as the target of inference in the multiplier method. A closed population (Lukacs, 2009), is one that is fixed in composition, with no migration, births nor deaths. However, the composition of human populations is constantly changing. Thus to achieve a truly closed population, one must consider the target population at a single instance in time (target time) in a certain well-defined region (target region). Thus, the two essential descriptors necessary to define a human target population are geography and time.

The previous chapter demonstrated inference under the assumption that the marginal counts are known exactly. As a reminder, a marginal count is the number of individual with a certain trait (*marked* individuals) in the target population. However, in reality, data on the exact number of marked individuals in the target

region at the exact target time is impossible to obtain, since collecting such data is not instantaneous.

During the time between enrolling a marked individual in the record and the target time, an individual is subject to migration. Should recorded individuals migrate out of the target region before the target time without notifying the record keeping agency, the records present an over-counting of actual marginal counts. Similarly, should marked individuals who lived outside the target region migrate into the region without being noted in the relevant record, the records produce an under-counting of actual marginal counts. However, this nuance has been overlooked in previous literature and no adjustment method has been proposed.

This problem was observed in a collaboration with the British Columbia Centre for Disease Control (BCCDC) to infer the population size of Men-who-have-sex-with-men (MSM) living in the Greater Vancouver Regional District (GVRD) at the end of year 2008. In this application, the disease surveillance database, whose data are collected over the years, appear initially as a source for tabulating "marginal counts" for traits related to disease diagnosis. However, as individuals diagnosed with diseases are not tracked over time in the Canadian disease surveillance system, the surveillance data does not provide the actual marginal count at the target time.

While the true marginal count for a trait is not observed, fuzzy information on the marginal count can be obtained if one is to make modelling assumptions. In the following sections a simple model is outlined, followed by details on how to integrate the migration model with the multiplier method via Bayesian inference. This is followed by an application of the proposed model and inference to real data. The chapter concludes with a discussion of the proposed method.

4.1 A Model for Migration

The proposed migration model contains two basic components, emigration and immigration. The unobserved marginal count is composed of individuals who stayed in the region of interest until the target time, J, after a being recorded at a earlier time j. The model simplifies the continuous time migration process into discrete time process (with time indexed incrementally by integers). A limitation of the proposed model is that it may be used with the single trait multiplier method, or

multi-trait multiplier method with only one marginal count affected by migration.

Let

- N_t be the marginal count, i.e. number of marked individuals in the target population, i.e. living in the target region at the target time,
- S_j be the count of marked individuals who stay in the target region at the target time after being listed in the record as residing in target region at time j,
- I_j be the count of marked individuals who immigrate to the target region by the target time after being listed in the record as residing *outside* target region at time j,
- D_j be the count of marked individuals enlisted in the record as residing in target region at time j,
- O_j be the count of marked individuals enlisted in the record as residing *outside* target region at time j,
- p_s be the probability that a marked individual residing in the target region at time j stays in the target region at time $j+1$,
- p_m be the probability that a marked individual residing outside the target region at time j move in to the target region at time $j+1$.

A model, $\mathscr{M}_m$ for N_t is the following:

$$N_t = \sum_j S_j + I_j \tag{4.1}$$

$$S_j \sim Binomial(p_s^{(J-j)}, D_j) \tag{4.2}$$

$$I_j \sim Binomial(1-(1-p_m)^{(J-j)}, O_j). \tag{4.3}$$

The above model makes the following assumptions about migration:

1. Migration is a discrete time process.
2. The probabilities p_s, p_m that dictate the chance of staying/moving do not vary with time.

3. The probability of moving more than once is negligible.

4. An individual's probability of being in the target population depends on a binary residency descriptor (in the target region or not) at the time he/she was enlisted in the records.

4.2 Combining the Migration Model with the Multiplier Method

To infer N with uncertainty in one of the marginal counts with the migration model $\mathscr{M}_m$ is to use the migration model as a prior for N_t in the likelihood-based multiplier method of Chapter 3 in a Bayesian inference, creating a two-level *combined* model. The parameters p_m and p_s in $\mathscr{M}_m$ are effectively hyper-parameters in the combined model.

To infer N in the case of two binary traits, the reparameterized likelihood for the survey data follows the parameterization based on the Section 3.1.1. Where N_{t1} is affected by migration, the posterior distribution of the combined model is

$$f(p_{11},N,N_{t1},p_m,p_s|\text{data}) \propto f(\text{data}|p_{11},N,N_{t1},p_m,p_s) \times f(p_{11},N,N_{t1},p_m,p_s). \tag{4.4}$$

An intuitive prior that allows a subject area expert to prioritize a-priori knowledge on N over p_{11}, can be obtained by factoring $f(p_{11},N,N_{t1},p_m,p_s)$ as $f(p_{11}|N,N_{t1},p_m,p_s) \times f(N|N_{t1},p_m,p_s) \times f(N_{t1}|p_m,p_s) \times f(p_m|p_s) \times f(p_s)$.

For example, one may choose the following:

- $p_s \sim Uniform(a,b)$, with a,b reflecting subject area knowledge
- $p_m|p_s = p_m \sim Uniform(c,d)$, with c,d reflecting subject area knowledge
- $N_{t1}|p_m,p_s$ is the model $\mathscr{M}_m$
- $N|N_{t1},p_m,p_s = N|N_{t1} \sim Uniform(L,U)$, with $L \geq \max(N_{\cdot 1},N_{1\cdot},n)$, and U an upper bound on the possible size of N
- $p_{11}|N,N_{t1},p_m,p_s = p_{11}|N,N_{t1} \sim Uniform(0,b_N)$, wtih $b_N = \min(N_{\cdot 1}/N,N_{1\cdot}/N)$

Note that the above example highlights some important observations about integrating uncertainty on marginal count with performing the Bayesian inference on N with the method in Chapter 3. Should N_{t1} be given, the variables N, p_{11} do not depend on the mechanism that determines N_{t1}. In this respect, the migration model can be thought of as "standing-alone". Therefore the prior on N, p_{11} can be formulated according to Section 3.3, without consideration of migration.

4.3 Application: Estimating the Size of MSM Population in GVRD

4.3.1 The Data

As a potential source of marginal counts, the British Columbia Centre for Disease Control (BCCDC) monitors new diagnoses of sexually transmitted infections such as HIV, syphilis, etc. Beginning in 2004, any first time diagnosis of either HIV or Syphilis, a reportable disease, must be reported to the BCCDC. The surveillance database also contains detailed demographic information, e.g. age, sex, residence, etc., on each patient. As a potential source for the trait prevalence, the BCCDC also has data from The ManCount Project (Moore et al., 2011), a survey undertaken to understand the sexual health of MSM in Vancouver, Canada. The survey included questions regarding time of an individual's first diagnosis of HIV and Syphilis, if any. This survey was conducted at the end of 2008 using Venue-Based Sampling (VBS). The survey was restricted to MSM who were age 19 or older at the time of the survey. The survey captures information regarding the sexual behaviour, disease diagnosis, disease testing behavior, together with basic demographic information (e.g. age, residence, etc.). A total of 1012 respondents from the Greater Vancouver Regional District (GVRD) were surveyed.

An appropriate definition of the target population is "the population of MSM above age 19 living in GVRD at the end of year 2008," based on the information contained in the data. Secondly, based on the above two sources, two traits of interest are identified as the following:

- Trait 1: First diagnosis of HIV between 2004 and 2008.

- Trait 2: First diagnosis of Syphilis in year 2008.

Based on the definition of the target population, a perfect marginal count for each trait needs to be tabulated at the end year 2008. However, the BCCDC surveillance database enrols people into the database at the time of their diagnosis without tracking the location of these individuals after enrolment in database. As such, the marginal count for trait 1 is suspected to be inaccurate as the oldest records are from 2004 giving those individuals plenty of time to migrate out of GVRD. Furthermore, other provincial health agencies also do not report emigration of individuals diagnosed in their provinces to BCCDC, such that anyone who was diagnosed in another province but moved to GVRD by 2008 would not be present in the BCCDC record. Epidemiologists agree that international migration does not seriously affect marginal count for trait 1 due to stringent immigration rules surrounding HIV infected individuals.

In order to model the effect of inter-provincial migration on the surveillance counts, further data are needed. At a minimum, data on the number of new HIV diagnosis within Canada each year is required. These data are readily available from the Public Health Agency of Canada (PHAC), in publicly accessible reports (Surveillance and Risk Assessment Division, Centre for Communicable Diseases and Infection Control, 2010). This report also provides new HIV diagnoses stratified by province. However, in the following model all outside-GVRD diagnoses are aggregated to reduce model complexity. Further notes about this modelling decision are given in the discussion.

4.3.2 Bayesian Inference

As in Section 4.2, the posterior distribution of model parameters is, $f(p_{11}, N, N_{t1}, p_m, p_s|\text{data}) \propto f(\text{data}|p_{11}, N, N_{t1}, p_m, p_s) \times f(p_{11}, N, N_{t1}, p_m, p_s)$. To implement the reparametrized likelihood with pre-defined distributions in popular MCMC software, e.g. JAGS,

Table 4.1: Selected hyper-parameter values for parameters relating to migration

	a	b	c	d
Set 1	0.9	1	0	0.06
Set 2	0.978	1	0	0.004

it may be written as,

$$\begin{aligned}
\mathbf{x} &\sim Multinomial(\mathbf{p}, n) \\
p_{10} &= N_{t1}/N - p_{11} \\
p_{01} &= N_{t2}/N - p_{11} \\
p_{00} &= 1 - p_{10} - p_{01} - p_{11}.
\end{aligned}$$

The prior distribution is specified according to Section 4.2, with migration hyper-parameters as in Table 4.1. Table 4.1 lists two sets of values. Hyper-parameter values in Set 1 reflects a conservative (wide) prior upon consulting epidemiologists at BCCDC. Hyper-parameters in Set 2 are based on inter-provincial migration statistics (BC Stats, 2009; Statistics Canada, Demography Division, 2013) for the overall BC population, with the mean migration probability matching the average migration proportion in 2004-2008 and upper cut-off extending to 1 for p_s and lower cut-off extending to 0 for p_m. The prior for $f(N|N_{t1})$ is capped at maximum of 270,000 which is roughly 30% of the adult male population in 2008 according to Statistics Canada.

Population size, N, is estimated based on the posterior distribution of its logarithm, i.e. $\log(N)$, a standard practice for estimating variables that are necessarily positive. The Bayes estimate, $\hat{N}_{\text{Bayes}}$, is defined as the exponentiated mean of the posterior distribution for $\log(N)$. A 95% credibility interval (CI) for N is obtained by exponentiating the end-points of the 95% HPD CI for $\log(N)$. Note that the posterior is obtained with 50,000 MCMC runs in JAGS. A complete model file for the model is found in Appendix F.

Table 4.2: Result from inferring size of MSM population with the combined model

	migration (parameter set 1)	migration (parameter set 2)
$\hat{N}_{\text{Bayes}}$	+22%	no change
95%CI $\hat{N}_{\text{Bayes}}$	+32%	no change
mean(N_{t1})	+27%	no change
length C.I.(N_{t1})	500 people	60 people

4.3.3 Result

Due to confidentiality, the exact estimate on the size of MSM population in 2008 cannot be disclosed in this thesis. Detailed results are planned for publication in an epidemiology journal in the near future. Nevertheless, the difference in key quantities between inference with migration versus without migration are provided in Table 4.2.

By using migration hyper-parameter set 1 (see Table 4.1), which is a conservative (wide) prior, migration resulted in a large amount of uncertainty on N_{t1}, with 95% CI range of 500 people. Both the estimate of N and its standard error increased by a large percentage (+22% and +32%). Using the second set of hyper-parameters, a tight prior, in the migration model resulted in a range for 95% CI for the unobserved N_{t1} of about 60 people. This uncertainty, however, had a negligible effect on the population size estimate and its standard error. Though not displayed due to confidentiality, both sets of hyper-parameters resulted in the confidence interval for $\hat{N}_{\text{Bayes}}$ after adjusting for migration overlapping with the CI without adjusting for migration, indicating no large inconsistency in estimating N without accounting for migration and versus accounting for migration.

4.4 Discussion

This chapter introduced a migration model in attempt to systematically capture the uncertainty in marginal count data due to migration. The model can be easily integrated with the multi-trait multiplier method of Chapter 3 under a Bayesian framework. The Bayesian analysis is demonstrated with an example to estimate the population size of MSM in GVRD in 2008 with data hosted by the BCCDC. In

this application extra data from outside the region are required, but these extra data are easily found using the data from Canada wide disease surveillance database.

Section 4.1 noted four important assumptions in the proposed migration model. Firstly, the proposed model describes migration in discrete time steps. While reducing the length of each time-step may increase the resemblance of the discrete time model to one of continuous time, there are disadvantages. As the time-interval decreases, the probabilities p_s and p_m may become too small for epidemiologists to formulate an appropriate prior for them. Furthermore, there is an advantage in using a time-interval based on a natural time division in human activities, e.g. monthly or yearly. In many organizations records are summarized periodically by the year, or by the month so that supplementary data may be easily acquired. In the exemplary application, by using time-steps by the year, inter-provincial migration statistics (BC Stats, 2009) can be used to derive a suitable prior for p_s and p_m. However, the need to explore trade off between model fit and ease of implementation due to varying time-interval length remains.

Secondly, the model assumes p_s and p_m as invariant in time. This allows a greater degree of parsimony. This assumption can be empirically validated in the case of the example application. One may obtain, again, from the inter-provincial migration statistics the proportion of people leaving and entering BC each year; this number is roughly the same over the period of 2004-2008 so there is no good reason to suspect the necessity of time-varying parameters. Furthermore, as the proportion of overall inter-provincial migration in Canada is small (BC Stats, 2009; Statistics Canada, Demography Division, 2013), the assumption regarding the chance of moving more than once as negligible is likely adequate.

Lastly, in this model, the only "covariate" that influences the probability of migration is the binary indicator of a person being in the target region or not at the time enlisted in the records. This is again a simplification of migration mechanism for 1) parsimony, and 2) the ease of generating prior information. In the context of the example application, data on migration into and out of BC exist, but these data stratified by the province from which an immigrant comes from are not publicly accessible. As such, a complicated model which incorporate finer geographical information may be difficult to implement due to the lack of prior information on p_s and p_m for small regions. On a similar note, other covariates (e.g. age,

ethnicity, etc.) may also have an important role in the probability of migration, but as such information is not well documented in migration records in Canada, prior knowledge on migration with respect to any combination of covariates would be extremely difficult to obtain.

In the application, two sets of hyper-parameter values were tested with results shown in Table 4.2. The choice of hyper-parameter values has a significant effect on how much uncertainty is place on the unobserved N_{t1}. Based on the two test cases, increased uncertainty on N_{t1} translates into uncertainty about $\hat{N}$, but when the uncertainty about N_{t1} is small, the amount of uncertainty surrounding $\hat{N}$ may not be affected at all. The two sets of scenarios tested represent very extreme cases; one likely ought to adjust the final hyper-parameter choice to be somewhere in between the values used in the two sets. According to the collaborators, the GVRD is attractive to MSM due to lifestyle and the quality of care for STI diagnosed individuals, such that an influx of MSM over time to GVRD is a reasonable expectation.

Finally, the chapter has not included any validation of the migration model, nor comparison with the ad-hoc method to account for uncertainty in marginal count. The method described here also is also suitable for when only one marginal count is affected by migration. These shortcomings will hopefully be addressed in future work.

Chapter 5

Conclusions

This thesis has highlighted two important improvements on the multiplier method. Firstly, a likelihood-based method is presented to incorporate information from multiple traits for a widely practised data collection scheme. This likelihood-based extension not only enables the multiplier method to use data from multiple traits simultaneously, but also allows the use of categorical traits should one desire. In addition, the exploration on the effect of study design on inference precision (Section 3.2 and Appendix B) enables the prediction of inference precision associated with choices made in study design, such that resources may be spent in a predictable way.

Secondly, Chapter 4 is the first body of work to account for uncertainty in the marginal count data based on an explicitly formulated migration model. An explicit specification of the migration model allows distributional parameters to be chosen or critiqued based on publicly available data, as in the application in Section 4.3. This model can be conveniently integrated into the likelihood-model presented in Chapter 3 under a Bayesian framework. Furthermore, an implementation of Bayesian inference is also outlined which may be carried out in freely available MCMC software.

As final remarks, this work is by no means a comprehensive solution that covers every problem with the multiplier method. The method in Chapter 3 is is tailored to the data collection scheme where a single survey queries multiple traits. While this particular data collection scheme is the only one employed at the moment,

when new data collection designs surface in the public health system, development of new statistical methods to infer population size will be necessary. Also, the migration model in Chapter 4 falls under what is completely model-based inference, where model misspecification could seriously undermine the estimation. Thus, work to examine the robustness of the migration model is prioritized for the immediate future. Future work may also include validating the practicality of the migration model, and comparison with previous methods to capturing uncertainty in the marginal counts.

Bibliography

Archibald, C., Jayaraman, G., Major, C., and Patrick, D. (2001). Estimating the size of hard-to-reach populations: a novel method using HIV testing data compared to other methods. *Aids* **15,** S41–S48.

BC Stats (2009). Migration review 2008. Technical report, BC Stats.

Carroll, R. J. and Lombard, F. (1985). A note on n estimators for the binomial distribution. *Journal of the American Statistical Association* **80,** 423–426.

Chao, A., Tsay, P., Lin, S. H., Shau, W. Y., and Chao, D. Y. (2001). The applications of capture-recapture models to epidemiological data. *Statistics in Medicine* **20,** 3123–3157.

Colón, H. M., Robles, R. R., and Sahai, H. (2001). The validity of drug use responses in a household survey in puerto rico: comparison of survey responses of cocaine and heroin use with hair tests. *International Journal of Epidemiology* **30,** 1042–1049.

Delaney-Black, V., Chiodo, L. M., Hannigan, J. H., Greenwald, M. K., Janisse, J., Patterson, G., Huestis, M. A., Ager, J., and Sokol, R. J. (2010). Just say i don't: lack of concordance between teen report and biological measures of drug use. *Pediatrics* **126,** 887–893.

Ezoe, S., Morooka, T., Noda, T., Sabin, M. L., and Koike, S. (2012). Population size estimation of men who have sex with men through the network scale-up method in Japan. *PLoS ONE* **7,** e31184.

Fendrich, M., Johnson, T. P., Sudman, S., Wislar, J. S., and Spiehler, V. (1999). Validity of drug use reporting in a high-risk community sample: a comparison of cocaine and heroin survey reports with hair tests. *American Journal of Epidemiology* **149,** 955–962.

Heckathorn, D. D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems* **44,** 174–199.

Hook, E. B. and Regal, R. R. (1995). Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews* **17,** 243–64.

Johnston, L. G., Prybylski, D., Raymond, H. F., Mirzazadeh, A., Manopaiboon, C., and McFarland, W. (2013). Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations: case studies from around the world. *Sexually Transmitted Diseases* **40,** 304–10.

Kish, L. (1965). *Survey sampling*. John Wiley and Sons.

Lieb, S., Fallon, S. J., Friedman, S. R., Thompson, D. R., Gates, G. J., Liberti, T. M., and Malow, R. M. (2011). Statewide estimation of racial/ethnic populations of men who have sex with men in the U.S. *Public Health Reports* **126,** 60–72.

Luan, R., Zeng, G., Zhang, D., Luo, L., Yuan, P., Liang, B., and Li, Y. (2005). A study on methods of estimating the population size of men who have sex with men in Southwest China. *European Journal of Epidemiology* **20,** 581–585.

Lukacs, P. (2009). Closed population capture-recapture models. In Cooch, E. G. and White, G. C., editors, *Program MARK: a gentle introduction*. http://www.phidot.org.

Magnani, R., Sabin, K., Saidel, T., and Heckathorn, D. (2005). Review of sampling hard-to-reach and hidden populations for hiv surveillance. *Aids* **19,** S67–S72.

Moore, D. M., Kanters, S., Michelow, W., Gustafson, R., Hogg, R. S., Kwag, M., Trussler, T., McGuire, M., Robert, W., Gilbert, M., et al. (2011). Implications for HIV prevention programs from a serobehavioural survey of men who have sex with men in Vancouver, British Columbia: the ManCount study. *Canadian Journal of Public Health* **103,** 142–46.

Muhib, F. B., Lin, L. S., Stueve, A., Miller, R. L., Ford, W. L., Johnson, W. D., Smith, P. J., for Youth Study Team, C. I. T., et al. (2001). A venue-based method for sampling hard-to-reach populations. *Public Health Reports* **116,** 216.

Okal, J., Geibel, S., Muraguri, N., Musyoki, H., Tun, W., Broz, D., Kuria, D., Kim, A., Oluoch, T., and Raymond, H. F. (2013). Estimates of the size of key populations at risk for HIV infection: men who have sex with men, female sex workers and injecting drug users in Nairobi, Kenya. *Sexually Transmitted Infections* **89,** 366–71.

Olkin, I., Petkau, A. J., and Zidek, J. V. (1981). A comparison of n estimators for the binomial distribution. *Journal of the American Statistical Association* **76,** 637–642.

Paz-Bailey, G., Jacobson, J. O., Guardado, M. E., Hernandez, F. M., Nieto, A. I., Estrada, M., and Creswell, J. (2011). How many men who have sex with men and female sex workers live in El Salvador? Using respondent-driven sampling and capture-recapture to estimate population sizes. *Sexually Transmitted Infections* **87,** 279–82.

Pollock, K. H., Nichols, J. D., Brownie, C., and Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs* pages 3–97.

Purcell, D. W., Johnson, C. H., Lansky, A., Prejean, J., Stein, R., Denning, P., Gau, Z., Weinstock, H., Su, J., and Crepaz, N. (2012). Estimating the population size of men who have sex with men in the United States to obtain HIV and syphilis rates. *The Open AIDS Journal* **6,** 98–107.

Raymond, H. F., Bereknyei, S., Berglas, N., Hunter, J., Ojeda, N., and McFarland, W. (2013). Estimating population size, HIV prevalence and HIV incidence among men who have sex with men: a case example of synthesising multiple empirical data sources and methods in San Francisco. *Sexually Transmitted Infections* **89,** 383–7.

Saidel, T., Loo, V., Salyuk, T., Emmanuel, F., Morineau, G., and Lyerla, R. (2010). Applying current methods in size estimation for high risk groups in the context of concentrated epidemics: lessons learned. *jHASE-Journal of HIV/AIDS Surveillance & Epidemiology* **2,**.

Salganik, M. J., Fazito, D., Bertoni, N., Abdo, A. H., Mello, M. B., and Bastos, F. I. (2011). Assessing network scale-up estimates for groups most at risk of HIV/AIDS: evidence from a multiple-method study of heavy drug users in Curitiba, Brazil. *American Journal of Epidemiology* **174,** 1190–6.

Statistics Canada, Demography Division (2013). Annual estimates of population for canada, provinces and territories, from july 1, 1971 to july 1, 2013. Technical report, Statistics Canada.

Surveillance and Risk Assessment Division, Centre for Communicable Diseases and Infection Control (2010). Hiv and aids in canada: Surveillance report to december 31, 2009. Technical report, Public Health Agency of Canada.

Sutherland, J. M. (2003). *Multi-list methods in closed populations with stratified or incomplete information*. PhD thesis, Simon Fraser University.

UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance (2010). Guidelines on estimating the size of populations most at risk to HIV. Technical report, World Health Organization.

Appendix A

On the Inappropriateness of Using Capture-Recapture Inspired Methodology to Analyse Data for the Multiplier Method

The following discussion is restricted to applying the multiplier method with at least two traits. Recall that in the multiplier method, one has data on the exact number of people in the target population with a trait for t traits, and a survey querying the status of all t traits in a sample of the target population. One may be inclined to draw parallels between the formation of a trait with "capturing and tagging" individuals in the target population which forms the basis of capture-recapture type inference. Should one take this view, data on t marginal counts gives rise to t "lists" in capture-recapture terminology, $L_{m_1}, \ldots, L_{m_t}$. However, unlike true capture-recapture, individuals cannot be matched at all across these first t lists, due to privacy regulations.

The survey component of data for multiplier method gives rise to another list, L_s, of individuals sampled from the target population. By design, survey respondents self-report on t traits for which marginal counts are available. This means only for members of L_s can their membership in $L_{m_1}, \ldots, L_{m_t}$ be ascertained. Thus

Table A.1: Comparison of observable statistics from a two trait multiplier method study to a true three list capture-recapture study. N_{000} is never observable even in true capture-recapture designs.

	Observable summary statistics for capture-recapture analysis							
	N_{000}	N_{001}	N_{010}	N_{011}	N_{100}	N_{101}	N_{110}	N_{111}
data from multiplier-method study	X	✓	X	✓	X	✓	X	✓
data from true capture-recapture study	X	✓	✓	✓	✓	✓	✓	✓

data for the multiplier method, when cast under the "capture-recapture" framework, produces $t+1$ lists with some unmatchable member. Table A.1 illustrates the critical information missing from casting a two trait multiplier method study in a three list capture-recapture analysis. While connections between capture-recapture and multiplier method likely exists, the following paragraphs emphasize the limitations of forcing capture-recapture analysis on multi-trait multiplier method data.

Capture-recapture methodologies for lists with unmatchable members are underdeveloped for application in epidemiology. Sutherland (2003) argues that classical models motivated by ecological applications for lists with unmatchable members are unsuitable given the mechanism in epidemiology. A method motivated by epidemiological applications is described in Sutherland (2003) for two-list capture-recapture. However without extension to multi-list this method cannot be applied to the problem at hand.

Furthermore, key assumptions in analysis tailored for capture-recapture experiments cannot be justified in the multiplier data. A fundamental flaw is the requirement that any person in the target population must have non-zero probability of being on any list (Chao et al., 2001). Say for example, that one of the traits is "positive diagnosis of HIV." Given the epidemiological practice of using confirmatory testing to eliminate false positives, people without the HIV virus are surely excluded from the list of people diagnosed with HIV. According to Chao et al. (2001), an epidemiologist must modify the definition of the target population to exclude those with zero probability of being jointly captured in all lists. How-

ever, in the example of using HIV diagnosis as one of the traits, this advice leads to exclusion of the *majority* of the intended target population as HIV prevalence is expected to be low - rendering the inference useless in practice. Furthermore, given the wide variety of traits that can be employed in the multiplier method, e.g., disease diagnosis, service usage, etc., it is difficult for epidemiologists to justify choosing any particular structure for dependencies between traits when employing either the ecological or log-linear models of capture-recapture. While Chao et al. (2001) note a third category of analysis in capture-recapture, the "sample coverage" approach, which bypasses this requirement, this method does rely on an untestable assumption.

A second type of analysis – known as "capture-recapture without recapture" (Carroll and Lombard, 1985; Olkin et al., 1981) – does not require cross-identification of subjects across lists. However, these methods makes a key assumption that the probability of members of the target population appearing on a list being the same across lists, which does not apply to the data at hand. Furthermore, this type of analysis assumes equal catchability of all members and having no uncatchable members, which, similar to classical capture-recapture. These assumptions are are unwarranted in the epidemiology data used in the multiplier method for the reasons given in the previous paragraph.

In contrast, the reparameterized likelihood multiplier method of Chapter 3 does not share the above concerns of capture-recapture due to having been truly motivated by the data at hand. An analogy of the reparameterized likelihood method is drawing marbles out of a bag where one observes each marble having up to t types of markings (traits), conditional on the number of each type of marking in the bag being known exactly. The only assumptions required, besides having a closed-population, is that the survey be a representative probability survey of the target population.

Appendix B

Bench-Marking with Equi-Correlation and Equi-Prevalence for the General Case

In Section 3.2, the relationship between estimation uncertainty and properties of the traits being used was explored under equi-prevalence and equi-correlation scenarios only. However, intuition suggests that the observed behaviour of estimation uncertainty may be extrapolated to apply more generally. In this section, this intuition is verified by showing that the performance in general falls between bench-marking numbers generated under the equi-correlation, equi-prevalence restriction. This is done via a Monte Carlo simulation.

The simulation is restricted to study designs with between 3 to 5 binary traits, each with 100 test cases. To generate a random case of k binary traits, $\boldsymbol{q}$, a k dimensional vector of marginal prevalences is sampled one dimension at a time from $Uniform(0, p_{max})$, and $\boldsymbol{\rho}$, representing $\binom{k}{2}$ correlation parameters is sampled from $Unif(0,1)$ one dimension at a time. This selected case is mapped to cell probabilities by using the cumulative distribution for a k-variate Normal distribution, $MVN_k(0,\Sigma)$, with a vector of cut-points which correspond to marginal trait preva-

lences $\boldsymbol{q}$. The diagonal elements of Σ are equal to 1, and off-diagonal elements of Σ are set as randomly selected correlations $\boldsymbol{\rho}$. The cut-points are the desired trait prevalences are specified through. The matrix Σ is confirmed to be positive-definite, otherwise $\boldsymbol{\rho}$ is re-drawn.

The estimation uncertainty for this randomly generated set of traits is calculated upon obtaining the contingency table. Three bench-marks for this uncertainty are investigated: b_l, b_m, b_u. Each benchmark is obtained from a set of k traits of equi-correlation and equi-prevalence for which the correlation is a function of $\boldsymbol{\rho}$, and prevalence is a function of $\boldsymbol{q}$. The first benchmark, b_l, uses $\max(\boldsymbol{q})$ and $\min(\boldsymbol{\rho})$ for prevalence and correlation. The second, b_m, uses $avg(\boldsymbol{q})$ and $avg(\boldsymbol{\rho})$ for prevalence and correlation, Lastly, b_u uses $\min(\boldsymbol{q})$ and $\max(\boldsymbol{\rho})$ for prevalence and correlation.

The result of the study shows that the estimator uncertainty arising from a general set of traits is always bounded by the maximum and minimum of the 3 benchmarks. The result from this simulation study is visualized in Figure B.1. In particular, b_m appear to be highly predictive of the actual estimator error. This property suggests only the *a priori* knowledge on the average prevalence and average correlation is required of an subject-area expert to produce good predictions of the study precision.

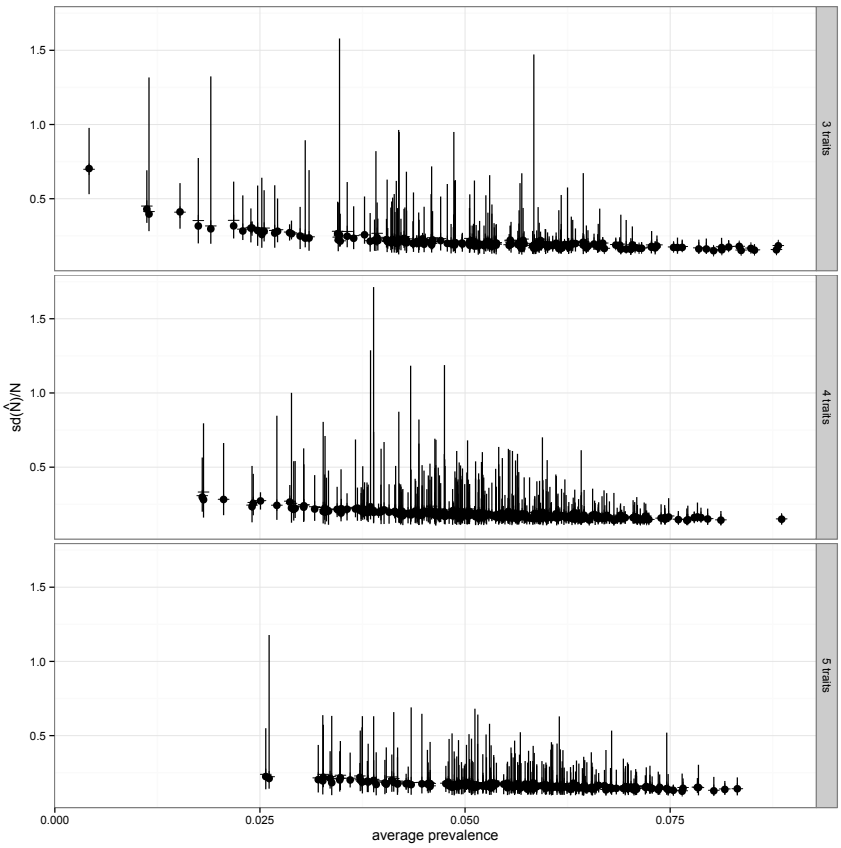

Figure B.1: The case of unequal association and unequal prevalence. Solid-circles in the plot represent the actual estimation uncertainty for each randomly selected test-scenario. Each circle is accompanied by a vertical line extending from the lower bench-mark (b_l) to the upper bench-mark (b_u). Each vertical line is accompanied by a tick-mark which locate the middle bench-mark (b_m). Each bench-mark is obtained with the assumption of equi-correlation (chosen to be a function of $\binom{K}{2}$ original correlations) and equi-prevalence (chosen to be a function of K original prevalences). One observes that the actual estimator error is always less than b_u. This suggests an *a priori* knowledge of the maximum plausible correlation and minimum plausible prevalence will lead to a conservative prediction of the estimation uncertainty. However, in some cases, this usage may be overly conservative. Instead, b_m appears highly predictive of the actual estimation uncertainty which suggests the use of average prevalence and average correlation in study design.

Appendix C

Deriving the Fisher Information

C.1 Notations

In this section some new notations are defined for describing k traits building on those appearing in Section 2.3 of the main body . Let $\boldsymbol{x}$ be a vector which denotes the statuses of k traits, $\boldsymbol{x} = (x_1, \ldots, x_k)$. Next, viewing the k traits as forming a multi-dimensional contingency table, let C denote the number of cells in this contingency table, $C = \prod_{i=1}^{K} l_i$. Intuitively, each cell in the contingency table corresponds to a unique combination of the status of k traits, i.e. there exist a 1-1 function g which maps $\boldsymbol{x}$ to cell index, c. In this thesis the following mapping is followed:

$$c = g(\boldsymbol{x}) = 1 + x_1 + l_1 x_2 + l_1 l_2 x_3 + \ldots + \left(\prod_{i=1}^{k-1} l_i \right) x_k \tag{C.1}$$

It is easy to see that cell indexing begins from 1, and that cell 1 corresponds to the situation where all k traits are of status 0 – an absence of every trait. Using g, a new random variable $Z = g(\boldsymbol{X})$ is defined, and it is interpreted as the multinomial variable indicating the cell in the k dimensional table a sampled individual falls under. Let $\boldsymbol{p} = (p_1, \ldots, p_C)$ be the cell probabilities that govern the frequencies in the k dimensional table, i.e., $p_c = Pr[Z = c]$. Lastly, define $\boldsymbol{x}(c) = g^{-1}(c)$ which returns the combination of the k statuses that correspond to cell c.

C.2 Defining a Reparameterization for the Cell Probabilities

The reparameterization of a particular cell probability p_c, $c \in \{1,\ldots,C\}$, depends on the combination of trait statuses given by $\boldsymbol{x}(c)$. Because this systematic approach to reparameterization relies on expressing the cells that correspond to 1 or less traits being present as a function of the rest of the probabilities, the following sets of cell indices are defined to help express this mapping:

- $\mathscr{S} = \{1,\ldots,C\}$: the complete set of possible index numbers
- $\mathscr{A} = \left\{ c : \sum_{i=1}^{k} \mathbb{I}[\boldsymbol{x}(c)_i > 0] = 1 \right\}$: the index of the cells that correspond to trait combinations where only one trait is non-zero
- $\mathscr{B} = \mathscr{S} \setminus (\mathscr{A} \cup \{1\})$: the index of the cells that correspond to trait combinations where more than one trait is non-zero

Secondly, the cell probabilities that do not need reparameterizing are defined as identity mapped to a vector $\boldsymbol{\theta}$. The elements of $\boldsymbol{\theta}$ are part of the parameters in the reparameterized likelihood.

Next the function h is defined, such that $p_c = h(\boldsymbol{\theta}, N, \boldsymbol{M}, c)$:

$$p_c = \begin{cases} \theta_c & \text{if } c \in \mathscr{B} \\ \sum_{i=1}^{k} \sum_{s=1}^{l_i - 1} \mathbb{I}[\boldsymbol{x}(c)_i = s] \left(\frac{M_{is}}{N} - \sum_{m \in \mathscr{B}} \theta_m \mathbb{I}[\boldsymbol{x}(m)_i = \boldsymbol{x}(c)_i] \right) & \text{if } c \in \mathscr{A} \\ 1 - \sum_{i=1}^{k} \sum_{s=1}^{l_i - 1} \frac{M_{is}}{N} + \sum_{m \in \mathscr{B}} \theta_m \left(\sum_{j=1}^{k} \mathbb{I}[\boldsymbol{x}(m)_j > 0] - 1 \right) & \text{if } c = 1 \end{cases} \tag{C.2}$$

Note that $\boldsymbol{\theta} = \{\theta_c : c \in \mathscr{B}\}$. This means the size of $\boldsymbol{\theta}$ is $|\mathscr{B}|$, and θ_c is named with indices that identify to which p_c it is identity mapped, which also identify the unique combination of k trait values it represents.

C.3 The Fisher Information

Using the multinomial likelihood approximation for simple random sampling without replacement from a large population, a likelihood model for the survey data is:

$\log f(Z_1,\ldots,Z_n) = \sum_{c=1}^{C}\left(\sum_{i=1}^{n}\mathbb{I}[Z_i=c]\right)\log p_c$. Here p_c is used for simplicity but it is understood to be a function of $\boldsymbol{\theta}$, N, $\boldsymbol{M}$ and c as in Equation (C.2). Note that $\boldsymbol{M}$ is known in this case.

To obtain the Fisher Information with respect to vector of parameters $\boldsymbol{\eta} = \{\boldsymbol{\theta}, N\}$ of length $|\mathscr{B}|+1$, let

$$A = \begin{bmatrix} \frac{\partial p_1}{\partial \theta_1} & \cdots & \frac{\partial p_1}{\partial N} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_C}{\partial \theta_1} & \cdots & \frac{\partial p_C}{\partial N} \end{bmatrix}.$$

As is standard for multinomial models, it is hereby stated without proof the form of the Fisher information matrix:

$$\begin{aligned} I &= \mathbb{E}_Z\left[A^t diag\left(\frac{\mathbb{I}[Z=1]}{p_1^2},\ldots,\frac{\mathbb{I}[Z=C]}{p_C^2}\right)A\right] \\ &= A^t diag\left(\frac{1}{p_1},\ldots,\frac{1}{p_C}\right)A \end{aligned}$$

C.4 Examining the Effect of Changing N on $Var(\hat{N}_{MLE})$

Assuming each person is sampled independently, the variance-covariance matrix $VCOV(\hat{\boldsymbol{\eta}}_{\text{MLE}}) \approx I_1^{-1}/n$, where I_1^{-1} is the inverse of the Fisher Information for one sample. The variance component of interest is given as $Var(\hat{N}_{\text{MLE}}) \approx (I_1^{-1})_{N,N}/n$.

In the last section, it was shown that $I = A^t diag\left(\frac{1}{p_1},\ldots,\frac{1}{p_C}\right)A$. In order to examine the effect of increasing N on $Var(\hat{N}_{\text{MLE}})$, *while keeping all other parameters the same*, an expression for I^{-1} that isolates out all appearances of N in the expression is desired. Let $q_{is} = M_{is}/N$ be the marginal prevalence of a trait, where $i \in \{1,\ldots,k\}, s \in \{1,\ldots,l_i-1\}$. Note that q_{is} is kept constant.

In the matrix of partial derivatives, A, according to the parameterization of p_c in Eqn. (C.2), $\frac{\partial p_c}{\partial \theta_l}$ does not contain N no matter which c or l. As for the elements

$\frac{\partial p_c}{\partial N}$ in the matrix A, the partial derivatives are:

$$\frac{\partial p_c}{\partial N} = \begin{cases} 0 & \text{if } c \in \mathscr{B} \\ -\sum_{i=1}^{k}\sum_{s=1}^{l_i-1} \mathbb{I}[\boldsymbol{x}(c)_i = s]\left(\frac{q_{is}}{N}\right) & \text{if } c \in \mathscr{A} \\ \sum_{i=1}^{k}\sum_{s=1}^{l_i-1} \frac{q_{is}}{N} & \text{if } c = 1 \end{cases}$$
$$\propto \frac{1}{N}$$

With the above knowledge I may be factored into the following matrix sub-blocks:

$$I = \begin{bmatrix} B_1 & \frac{1}{N}B_2 \\ \frac{1}{N}B_2^t & \frac{1}{N^2}B_3 \end{bmatrix}$$

where sub-block B_1 is of dimension $(|\mathscr{B}| \times |\mathscr{B}|)$, B_2 is of dimension $(|\mathscr{B}| \times 1)$ and B_3 of dimension (1×1).

Using the block inversion matrix formula, the element $(I^{-1})_{(N,N)} = (\frac{1}{N^2}B_3 - \frac{1}{N}B_2^t \times B_1 \times \frac{1}{N}B_2)^{-1} \propto N^2$. Hence asymptotically, $Var(\hat{N}_{MLE}) \propto N^2$, and $SD(\hat{N}_{\text{MLE}}) \propto N$. $\square$

Appendix D

Implementing Bayesian Inference

This appendix outlines the details regarding implementing Bayesian inference appearing in Section 3.3 of the main body with Markov Chain Monte Carlo (MCMC) in the software package OpenBUGS.

D.1 Model Specification for Inferring N with Information from Two Traits

In the case that inference is carried out with information from two traits, the specification of the reparameterized likelihood and priors can be done with built-in distribution functions in OpenBUGS. In this case the chain for $\log N$ is directly generated by OpenBUGS and inference from MCMC can be performed with the `coda()` package in R. Please see the following for the model file:

```
#uninformative (Uniform prior) #logN
#requires x, n, Marg1, Marg2
# U is the upper limit ...

model{
  logN <- log(N)
  x ~ dmulti(pi, n)
  pi[1] <- p00
  pi[2] <- p01
```

```
  pi[3] <- p10
  pi[4] <- p11
  p00 <- 1-p01-p10-p11
  p01 <- Marg2/N - p11
  p10 <- Marg1/N - p11
  p11 ~ dunif(0, min(Marg1/N, Marg2/N))
  N ~ dunif(max(n,Marg1,Marg2),U)
}
```

D.2 Model Specification for Inferring N with Information from Three Traits

In the main body, an argument was presented for preferring a prior of the form $\pi(N) \times \pi(\boldsymbol{\theta}|N)$. However, specifying a joint prior in a convenient MCMC software (e.g. OpenBUGS) of the form $\pi(N) \times \pi(\boldsymbol{\theta}|N)$ in the case of $K > 2$ traits is non-trivial. This is because the sampled $\boldsymbol{\theta}$ vector must satisfy marginal probabilities (constraints) given N, which will likely involve custom distribution functions and MCMC samplers. A simple work-around is to first obtain the posterior MCMC chains under a "convenience prior", $\pi(\boldsymbol{\theta}) \times \pi(N|\boldsymbol{\theta})$, then post-process these chains with importance sampling to represent the actual posterior distribution. This approach simplifies the sampling because the convenience prior involves a unconstrained multivariate distribution, $\pi(\boldsymbol{\theta})$, and a constrained univariate distribution, $\pi(N|\boldsymbol{\theta})$. Constraints on univariate distributions are readily implemented in most convenient MCMC software. The convenience prior chosen for this simulation consists of $\pi(\boldsymbol{\theta})$ taken to be the first $|\boldsymbol{\theta}|$ dimensions of a $Dirichlet(\boldsymbol{\alpha})$ distribution, and $\pi(N|\boldsymbol{\theta}) \sim Uniform(L_{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}})$ with $L_{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}}$ being the upper and lower limits as determined by $\boldsymbol{\theta}$, based on the constraints that $0 < p_c < 1$ for all p_c a function of $\boldsymbol{\theta}, N$ (i.e. Equation C.2).

Let $g'(N,\theta)$ be the posterior density under the convenience prior, and $g(N,\theta)$ be the posterior density up to a proportionality constant, i.e. without the normalization constant. Let $f'(N,\theta)$ be the actual posterior density, and $f(N,\theta)$ the actual posterior density up to a proportionality constant. An importance sampling estimate of the posterior expected value of a function of parameters $h(N,\boldsymbol{\theta})$, based on

a MCMC chain of length T, is $\mathbb{E}_{f'}[h(N,\boldsymbol{\theta})] \approx \tilde{h}_T$ where

$$\tilde{h}_T = \frac{\frac{1}{T}\sum_{t=1}^{T}\left[\frac{h(N^{(t)},\boldsymbol{\theta}^{(t)})f(N^{(t)},\boldsymbol{\theta}^{(t)})}{g(N^{(t)},\boldsymbol{\theta}^{(t)})}\right]}{\frac{1}{T}\sum_{i=1}^{T}\left[\frac{f(N^{(t)},\boldsymbol{\theta}^{(t)})}{g(N^{(t)},\boldsymbol{\theta}^{(t)})}\right]}$$

$$f(N^{(t)},\boldsymbol{\theta}^{(t)}) = \frac{1}{U-L}\frac{1}{V_{N^{(t)}}}\pi(data|p,N)$$

$$g(N^{(t)},\boldsymbol{\theta}^{(t)}) = \frac{\left(\prod_{j=1}^{|\boldsymbol{\theta}|}(\theta_j^{(t)})^{\alpha_j-1}\right)\left(1-\sum_{j=1}^{|\boldsymbol{\theta}|}\theta_j^{(t)}\right)^{\alpha_{|\boldsymbol{\theta}|+1}-1}}{B(\boldsymbol{\alpha})}\frac{1}{U_{\boldsymbol{\theta}^{(t)}}-L_{\boldsymbol{\theta}^{(t)}}}\pi(data|p,N)$$

The above re-weighting of the MCMC samples can be carried out completely in R, by using the R packages `geometry` and `rcdd` to obtain the approximate "volume" of the constraint space, V_N, for a given sample of N. To obtain the performance measures appearing in Table 2 in the main body with importance sampling, let $h_1(N,\boldsymbol{\theta}) = \log N$ and $h_2(N,\boldsymbol{\theta}) = (\log N)^2$ which gives the 1st and 2nd moment of $\log N$, and let $h_3(N,\boldsymbol{\theta}) = \mathbb{I}[\log N < t]$ for obtaining the equal tailed 95% credibility interval for $\log N$.

The model file for obtaining the posterior MCMC chain *under the convenience prior* is provided below:

```
#uninformative (Uniform prior) -> f(pi)f(N|pi)
#requires x, n, Marg1, Marg2, Marg3, alpha

model{
  logN <- log(N)
  x[1:8] ~ dmulti(pp[1:8], n)
  pp[1] <- 1- sum(pp[2:8])
  pp[2] <- Marg1/N - a-b-d
  pp[3] <- Marg2/N - a-c-d
```

```
pp[4] <- a
pp[5] <- Marg3/N - b-c-d
pp[6] <- b
pp[7] <- c
pp[8] <- d

theta[1:5]~ ddirich(alpha[1:5]) # let alpha be c(1,1,1,1,4)
a <- theta[1]
b <- theta[2]
c <- theta[3]
d <- theta[4]
U2 <- Marg1/(a+b+d)
U3 <- Marg2/(a+c+d)
U4 <- Marg3/(b+c+d)
U1 <- (Marg1+Marg2+Marg3)/(a+b+c+2*d)
L2 <- Marg1/(1+a+b+d)
L3 <- Marg2/(1+a+c+d)
L4 <- Marg3/(1+b+c+d)
L1 <- (Marg1+Marg2+Marg3)/(1+a+b+c+2*d)


#specify new prior for N|p
dummy <- 0
dummy ~ dloglik(phi)
phi <-  log(S)       #logLik of N
N ~ dflat()
S <- step(UpperLim- N)*step(N- LowerLim)*Y/Z
Z <- UpperLim-LowerLim


#define some vars
#check to make sure the upper limit > lower limit
Y <- step(UpperLim-LowerLim)
```

```
  LowerLim <- max(n,max(L1,max(L2,max(L3,L4))))
  UpperLim <- min(U1,max(U2,max(U3,U4)))

}
```

Appendix E

Properties of the Bayesian Estimator using Multiple Traits with the Multiplier Method – a Simulation Study

The purpose of this simulation study is twofold – to demonstrate that Bayesian inference, in terms of estimation uncertainty, performs similarly to the MLE with the use of an uninformative prior, and outperforms the MLE with the use of an informative prior. For each selected scenario, a random sample of size n from is drawn from a simulated population $B = 1000$ times. With *each* sample of size n, Bayesian inference is obtained from $\log N$ with $\hat{\log N}_{Bayes} = \mathbb{E}[\log N|Data]$ and $SD[\log N|Data]$ playing the role of standard error. The estimator error, $SD(\hat{\log N}_{Bayes})$, is numerically approximated with the standard deviation of B values of $\hat{\log N}_{Bayes}$.

In the summary of results, $\overline{(\hat{\log N}_{Bayes})} = 1/B\sum_{i=1}^{B}(\hat{\log N}_{Bayes})_i$ is provided to assess bias, $SD(\hat{\log N}_{Bayes})$ is listed alongside $SD(\hat{\log N}_{MLE})$ for comparison, $\overline{SD[\log N|Data]} = 1/B\sum_{i=1}^{B} SD[\log N|Data]_i$ is provided to assess the quality of using $SD[\log N|Data]$ as an substitute for estimator standard deviation- a leap of faith that is necessary in real applications. The coverage probability from equal-tailed 95% credibility interval is also presented as a indication of the performance of

finite-sample inference.

E.1 Simulation Results

Table E.1 summarises the performance of the Bayes estimator in select hypothetical populations. Comparing $\overline{\log \hat{N}}_{Bayes}$ with $\log N$ gives an approximation of bias. Note that in all cases the numerically obtained bias indicates a multiplicative bias of a few percent on the scale of N.

The standard deviation of the Bayes estimator is numerically approximated by $SD\left(\log \hat{N}_{Bayes}\right)$. A comparison of the standard deviation of the Bayes estimator that uses an uninformative prior with the theoretical asymptotic standard deviation of the MLE reveals close agreement. This gives us confidence in the results obtained from the theoretical exploration in previous sections. Furthermore, empirically, a gain in efficiency is observed with the use of informative priors, as seen in the top half and bottom half of Table E.1.

In practice, the posterior standard deviation of $log(N)$ plays the role of standard error. From Table E.1, the standard deviation of posterior density, $SD\left[\log N|Data\right]$, on average agrees with the standard deviation of the Bayes estimator. Furthermore, the equal-tailed 95% credibility interval provides good coverage probability. Overall, this simulation has demonstrated many good properties that support the use of Bayesian inference with the proposed reparameterized multinomial likelihood (Chapter 3) to infer population size.

Table E.1: Performance of Bayesian inference under select populations and prior distributions

Description of Population	Prior distribution	$\log N$	$\overline{\left(\log \hat{N}_{Bayes}\right)}$	$SD\left(\log \hat{N}_{MLE}\right)$	$SD\left(\log \hat{N}_{Bayes}\right)$	$\overline{SD\left[\log N \mid Data\right]}$	coverage prob.
$N = 20000$, $p = 0.05$, $n = 200$, $k = 2$, $\phi = 0$	uninformative: $\sim$Uniform(L, 1e6)	9.903	9.869	0.218	0.21	0.221	0.95
$N = 20000$, $p = 0.05$, $n = 200$, $k = 2$, $\phi = 0$	informative: $\sim$Uniform(L, 30000)	9.903	9.83	0.218	0.159	0.191	0.958
$N = 40000$, $p = 0.1$, $n = 500$, $k = 3$, $\phi = 0$	uninformative: $\sim$Uniform(L, 1e6)	10.597	10.588	0.077	0.075	0.079	0.96
$N = 40000$, $p = 0.1$, $n = 500$, $k = 3$, $\phi = 0$	informative: $\sim$Uniform(L, 50000)	10.597	10.584	0.077	0.069	0.076	0.964

Appendix F

JAGS Model File for Combined Modelling of Migration and Multiple Trait Multiplier Method

```
#in JAGS syntax

model{
  logN <- log(N)
  x ~ dmulti(pi, n)
  pi[1] <- p00
  pi[2] <- p01
  pi[3] <- p10
  pi[4] <- p11
  p00 <- 1-p01-p10-p11
  p01 <- N_t2/N - p11
  p10 <- N_t1/N - p11

  #prior
  p11 ~ dunif(0, min(N_t1/N, N_t2/N))
  N ~ dunif(max(n,N_t1,N_t2),270000)
  N_t1 <-  sum(S) + sum(I)
```

```
  S[1] ~ dbin(p_s^4, D[1])
  S[2] ~ dbin(p_s^3, D[2])
  S[3] ~ dbin(p_s^2,D[3])
  S[4] ~ dbin(p_s, D[4])
  S[5] <- D[5] #  S[j], where j = 5 represent year 2008

  I[1] ~ dbin( 1- (1-p_m)^4, O[1])
  I[2] ~ dbin( 1- (1-p_m)^3, O[2])
  I[3] ~ dbin( 1- (1-p_m)^2, O[3])
  I[4] ~ dbin( 1- (1-p_m), O[4])
  I[5] = 0

  ## hyper prior
  p_m ~ dunif(0,0.06)
  p_s ~ dunif(0.9,1)

}
```