SAPIENZA
UNIVERSITÀ DI ROMA

Department of Methods and Models for Territory,
Economics and Finance

Fisher's noncentral hypergeometric distribution and population size estimation problems

PhD candidate: Veronica Ballerini
Supervisor: Prof. Brunero Liseo

XXXIII PhD. Cycle

Fisher's noncentral hypergeometric distribution and population size estimation problems

Veronica Ballerini

Abstract

Fisher's noncentral hypergeometric distribution (FNCH) describes a biased urn experiment with independent draws of differently coloured balls where each colour is associated with a different weight (Fisher (1935), Fog (2008a)). FNCH potentially suits many official statistics problems. However, such distribution has been underemployed in the statistical literature mainly because of the computational burden given by its probability mass function. Indeed, as the number of draws and the number of different categories in the population increases, any method involving evaluating the likelihood is practically unfeasible. In the first part of this work, we present a methodology to estimate the posterior distribution of the population size, exploiting both the possibility of including extra-experimental information and the computational efficiency of MCMC and ABC methods. The second part devotes particular attention to overcoverage, i.e., the possibility that one or more data sources erroneously include some out-of-scope units. After a critical review of the most recent literature, we present an alternative modelisation of the latent erroneous counts in a capture-recapture framework, simultaneously addressing overcoverage and undercoverage problems. We show the utility of FNCH in this context, both in the posterior sampling process and in the elicitation of prior distributions. We rely on the PCI assumption of Zhang (2019) to include non-negligible prior information. Finally, we address model selection, which is not trivial in the framework of log-linear models when there are a few (or even zero) degrees of freedom.

Acknowledgements

First and foremost, I thank my supervisor, Prof. Brunero Liseo, who showed me the kind of researcher I aspire to become. Without his exceptional guidance and his trust during all these years, this work would not exist.
My deepest gratitude also goes to Prof. Li-Chun Zhang. Since we first met in 2017, our insightful discussions have always made me question my beliefs and helped me sharpen my reasoning. Our collaboration was supported by the Department of Mathematics of the University of Oslo that hosted me as a visiting PhD student. I extend my gratitude to Prof. Sven Ove Samuelsen, Prof. Ingrid Kristine Glad, and Prof. Arnoldo Frigessi, among others, for their availability.
Back to Rome, I would like to express my sincere gratitude to Prof. Andrea Tancredi, another reference point in my academic life, whose suggestions have proved crucial more than once.
I thank Dr Davide Di Cecco for his advice and support in providing me with important material during these years and for being the "clique separator" between Stefano De Santis and me. I extend my gratitude to Stefano for the provision of the data that motivate part of this work.
Thanks to Dr Rosario Barone, who has become a role model and dear friend since the beginning of my PhD. He taught me a lot.
I thank my colleagues for the exchange of views that enhanced this work. Special thanks to Roberta Di Stefano and Chiara Ferrante; our mutual daily support during the last year made work lighter and life easier.
My gratitude goes to the people of MEMOTEF, the stimulating environment where I grew up academically and humanly. I thank the Head of the Department, Prof. Giorgio Alleva, who has always done everything possible to meet PhD students' needs.
Finally, I would like to sincerely thank the reviewers for their careful comments on this work and valuable suggestions.

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

John Tukey

Contents

Introduction

In recent years, the National Statistics Institutes (NSIs) have been more and more interested in producing statistics using administrative data only. The idea of replacing censuses with the integration of data from multiple sources is attractive due to the former's several shortcomings. First, there exists a well-known and unavoidable trade-off between timeliness and accuracy; second, the response rate becomes lower year after year. Third, censuses imply huge costs. In these terms, a data integration approach would potentially produce an enormous gain.
Since October 2018, the Italian Statistical Institute (Istat) has shifted from a census-based statistics paradigm to a register-based one, initiating the "permanent census of the Population and Housing" (see Istat (2018)). Unlike the decennial census in use until then, the survey involves only a sample of households; simultaneously, it provides detailed information about the whole population yearly, thanks to data integration from statistical sample surveys and administrative sources. In this sense, sample surveys *support* the information collected by the administrative registers and not the other way around. Nevertheless, since the aims of who collect data and who use them differ, several methodological issues emerge; thus, more uncertainty and the risk of biasedness naturally arise. Istat itself is paying attention to the issues caused by the "paradigm shift" in official statistics; see Filipponi et al. (2017), and Chiariello and Tuoto (2018), among others.

In line with NSIs interests, academia is producing a vast literature. One of the most recent examples is "Analysis of Integrated Data", edited by Zhang and Chambers (2019). The contributes therein deal with statistical uncertainty and inference issues that arise when a dataset is created via multiple sources' integration. Among the key issues, there is *undercoverage*. A list is subject to undercoverage when it enumerates only a

subset of the population of interest (or *target population*). In the case of a single list, such a problem is intrinsic and challenging to treat. However, it is pervasive, especially when the target population is elusive, e.g. non-resident inhabitants in a big city, homeless people, irregular migrants. In the case of multiple lists, undercoverage occurs if the union of the lists is still a subset of the target population.
Literature has primarily treated this issue in such multiple sources case. The predominant models are the *capture-recapture* models, so called because they were born in the ecological field. In such models, the capture of specimens belonging to a target animal population is registered, and the animal marked; the registration is repeated for several capture occasions. Hence, it is possible to infer the entire multiple recapture history of any specimens from its unique mark anytime. A milestone is Fienberg (1972). In his work, Fienberg uses a contingency table to describe the capture histories of the registered units; the table is *incomplete* since the cell referring to those units that have never been captured is unobserved. The use of log-linear models allows the estimation of the missing cell count and, consequently, the total population size. Since 1972, a vast literature dealing with population size estimation in a capture-recapture framework has been developing, shifting focus from wildlife to human populations; see Böhning et al. (2018) for the most recent developments in capture-recapture models for social sciences. When dealing with humans, we must devote special attention to the possibility that one or more data sources erroneously include some out-of-scope units, i.e., the *overcoverage* problem. The overcoverage issue, which has become relevant only with the increase in interest of the NSIs in the production of statistics through data integration, does not boast literature as detailed as that of undercoverage. The approach which has spread the most is latent class modelling: see Di Cecco (2019), and Di Cecco et al. (2020), among others.

This work aims to make a methodological contribution to the open questions in the population size estimation field. On the one hand, we are interested in the size estimation of population sub-groups for whom data are scarce in the case of i) a single list is available or ii) there is more than one list, but we lack unique identifiers. On the other hand, we aim to contribute to the estimation problem in the presence of out-of-target units.
We investigate the application of an underused probability distribution,

i.e. the Fisher's noncentral hypergeometric, to these scopes. Such distribution applies to a biased urn problem: some coloured balls have been independently drawn from an urn, and the probability of observing that sample depends not only on the total number of balls of each colour but also on the relative odds, or *weights*, of the colours. Fisher's noncentral hypergeometric distribution has a high potential in official statistics. Assume that a sample survey partially enumerates a heterogeneous population and that the coverage probabilities vary among the different sub-groups; this is equivalent to observing different coloured balls drawn according to their different weights. Moreover, assume to cross-classify some sources affected by overcoverage: the cell's counts of the contingency table will be the sum of target units and erroneous enumerations. Under certain specifications, the full conditional distribution of the erroneous enumerations is Fisher's hypergeometric, whether the lists differently cover the target and non-target units.

This work, entirely embracing a Bayesian approach, is organised as follows.
Chapter 1 introduces Fisher's noncentral hypergeometric distribution in comparison with its "twin", i.e. the Wallenius' noncentral hypergeometric. For decades, the two distributions had been homonymous, leading to confusion that - together with the complexity of their probability masses - slowed down their spread. In 2008, Agner Fog clarified the misunderstanding, also giving methods for sampling from these distributions. We highlight the primary analogies and differences between the two, both under the formal aspect and their uses, and introduce part of the notation we will use throughout the work.

After a brief overview of the approaches to heterogeneity in population size estimation problems, chapter 2 applies Fisher's noncentral hypergeometric to the context of the official statistics. Therein we present a methodology involving such underused distribution in the size estimation of the population's sub-groups for whom data are scarce. Firstly, we describe the problem for the univariate case; afterwards, we extend the model to the multivariate one, proposing two different ways to overcome the computational complexity of Fisher's mass. We conclude the chapter with the case study that motivated this work.

Starting from chapter 3, the considered framework becomes that of multi-

way contingency tables, i.e. we have more than one sample, and we can infer the "capture history" of a unit from its unique identifier.
In particular, chapter 3, which is a modified version of Ballerini (2020), contains an overview of the literature on capture-recapture models with a specific focus on the population size estimation problem in the presence of out-of-scope units.

In light of the considerations that emerged reviewing the literature, chapter 4 presents a model to estimate the population size that includes an alternative way to treat the overcoverage. There, Fisher's noncentral hypergeometric distribution comes up again: we rely on it in the latent erroneous enumerations' sampling process. The chapter results from a collaboration with Prof. Li-Chun Zhang[1] during a visiting period at the University of Oslo in 2019.

In the conclusions, we will summarise the main results of this work and express some considerations concerning the directions of further research in the field of population size estimation.

[1]Statistics Norway, University of Oslo and University of Southampton

Part I

The use of Fisher's noncentral hypergeometric distribution for official statistics

Chapter 1

The noncentral hypergeometric distributions

In 2008, Agner Fog clarified the distinction between two distributions, both known in the literature as "the" noncentral hypergeometric distribution (see Fog (2008a) and Fog (2008b)). He firstly solved the nomenclature issue, attributing to each of them a "patronymic": Wallenius' and Fisher's, after the persons who first proposed them.
Indeed, Professor R. A. Fisher first described Fisher's noncentral hypergeometric distribution (FNCH) in 1935, in "The Logic of Inductive Inference", published in the Journal of the Royal Statistical Society (Fisher (1935)); however, he did not name it. Thirty years later, W. L. Harkness give the distribution the name of "extended hypergeometric" (Harkness (1965)), but the term "extended" has been barely used in the literature, as asserted by Fog (2008b). Instead, the most prevalent name attributed to Fisher's is "noncentral hypergeometric distribution".
Wallenius' noncentral hypergeometric distribution (WNCH) owes its name to K. T. Wallenius, who first described it in the univariate case in his PhD thesis in 1963, and named it "noncentral hypergeometric distribution" (see Wallenius (1963)). J. Chesson extended the distribution to the multivariate case (see Chesson (1976)), and preserved the name given by Wallenius. Therefore, it was natural that some confusion surrounded the terminology of the two distributions.

In this work, we adopt the solution proposed by Fog (2008a) identifying the two noncentral hypergeometrics with their attributes.
In §1.1, we describe the two distributions, highlighting the primary analo-

gies and differences; we will refer to Fog (2008a) throughout the section. In §1.2, we provide some examples to let the differences between the two be more evident to the reader and highlight potential uses of the noncentral hypergeometric distributions.

1.1 Definitions

Wallenius' noncentral hypergeometric

Assume an urn of size N contains M_1 balls of color 1 and M_2 balls of color 2. In the univariate case, Wallenius' noncentral hypergeometric distribution describes a situation in which the balls are drawn without replacement until n balls are sampled, and the probability to sample X_1 balls of colour 1, and X_2 balls of colour 2 depends on some weights w_1^W, w_2^W. It is said to describe a biased urn experiment since the weight associated with each colour can be seen as the probability to retain a ball of that colour when drawn (as suggested by Chesson (1976)). The probability of collecting a ball of colour 1 at the mth draw is then equal to the weighted proportion of balls of colour 1 still in the urn, i.e.

$$\frac{(M_1 - X_{1,m-1})w_1^W}{(M_1 - X_{1,m-1})w_1^W + (M_2 - X_{2,m-1})w_2^W} \ , \tag{1.1}$$

where $X_{c,m-1}, c = 1, 2$, is the number of balls of color c sampled in the first $m - 1$ draws.
Hence, the univariate WNCH is built on five parameters:

$$X_1 | X_1 + X_2 = n \sim \text{WNCH}(M_1, M_2, n, w_1^W, w_2^W) \ . \tag{1.2}$$

However, since w_1^W and w_2^W are defined up to a constant, we may state that the distribution is defined by their ratio. The probability mass function, derived by Wallenius (1963), is

$$P(X_1 = x_1 | X_1 + X_2 = n) = \binom{M_1}{x_1}\binom{M_2}{x_2} \int_0^1 \prod_{c=1,2} (1 - t^{w_c^W/d})^{x_c} \, \mathrm{d}t \tag{1.3}$$

where $d = (M_1 - x_1)w_1^W + (M_2 - x_2)w_2^W$ and $x_2 = n - x_1$.
The extension to the multivariate case is straightforward (see Chesson (1976)). The probability to observe $\boldsymbol{X} = \{X_c\}$ balls of colours $\{c\}, c =$

$1, 2, ..., C$, given that the total number of sampled balls is n, is equal to

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, ..., X_C = x_C | \sum_{c=1}^{C} X_c = n) = \\ = \prod_{c=1}^{C} \binom{M_c}{x_c} \int_0^1 \prod_{c=1}^{C} (1 - t^{w_c^W/d})^{x_c} \, \mathrm{d}t \end{aligned} \tag{1.4}$$

where $\boldsymbol{M} = \{M_c\}$ is the total number of balls of each colour contained in the urn, and $\boldsymbol{w}^W = \{w_c^W\}$ is the vector of respective weights. As in the univariate case, d is a weighted sum of the balls still in the urn after n have been sampled, i.e.

$$d = \sum_{c=1}^{C} (M_c - x_c) w_c^W \ . \tag{1.5}$$

The lack of a closed form for the probability mass function makes complex to evaluate it, even in the univariate case.

Fisher's noncentral hypergeometric

Instead, the univariate Fisher's noncentral hypergeometric distribution describes an urn experiment when the balls are drawn independently, without replacement, and the sample size n is observed only at the end of the experiment. It is the conditional distribution of two independent Binomial distributions given their sum (Harkness (1965)):

$$\begin{aligned} X_1 &\sim \mathrm{Binom}(M_1, \zeta_1) \\ X_2 &\sim \mathrm{Binom}(M_2, \zeta_2) \end{aligned} \tag{1.6}$$

$$X_1 | X_1 + X_2 = n \sim \mathrm{FNCH}(M_1, M_2, n, w_1^F, w_2^F) \tag{1.7}$$

with probability mass function

$$P(X_1 = x_1 | X_1 + X_2 = n) = \frac{\binom{M_1}{x_1}\binom{M_2}{x_2} w_1^{F\,x_1} w_2^{F\,x_2}}{\sum\limits_{(z_1,z_2)\in\mathcal{Z}} \binom{M_1}{z_1}\binom{M_2}{z_2} w_1^{F\,z_1} w_2^{F\,z_2}} \ . \tag{1.8}$$

where $\mathcal{Z} = \{(x_1, x_2) \in \mathbb{Z}^2 : x_1 + x_2 = n \ \cap \ 0 \leq x_c \leq M_c, c = 1, 2\}$.
Here the weights are the odds:

$$w_c^F = \frac{\zeta_c}{1 - \zeta_c} \ , \quad c = 1, 2 \ . \tag{1.9}$$

(see Appendix A for all steps). The notation in (1.7) highlights the similarity to Wallenius' distribution. Since $M_2 = N - M_1$ and $x_2 = n - x_1$, the formulation (1.7) is equivalent to:

$$X_1 | n \sim \text{FNCH}(M_1, N, n, w_1^F, w_2^F) \ . \tag{1.10}$$

We will interchangeably use the two parameterisations throughout this work. Moreover, (1.7) is also equal to

$$X_1 | X_1 + X2 = n \sim \text{FNCH}(M_1, M_2, n, k \cdot w_1^F, k \cdot w_2^F) \ , \quad k \in \mathbb{R}^+ \tag{1.11}$$

i.e., the odds ratio $w = w_1^F / w_2^F$ defines the noncentral hypergeometric distributions rather than the weights themselves (see Appendix A) - this is also valid for WNCH.
In the multivariate case, we have

$$\boldsymbol{X} | \sum_{c=1}^{C} X_c = n \sim \text{FNCH}(\boldsymbol{M}, n, \boldsymbol{w}^F) \tag{1.12}$$

and the probability mass function is

$$P(X_1 = x_1, X_2 = x_2, ..., X_C = x_C | \sum_{c=1}^{C} X_c = n) = \frac{\prod_{c=1}^{C} \binom{M_c}{x_c} {w_c^F}^{x_c}}{\sum_{\boldsymbol{z} \in \mathcal{Z}} \prod_{c=1}^{C} \binom{M_c}{z_c} {w_c^F}^{z_c}} \tag{1.13}$$

where $\mathcal{Z} = \{\boldsymbol{x} \in \mathbb{Z}^c : \sum_{c=1}^{C} x_c = n \cap \ 0 \leq x_c \leq M_c, \forall c\}$.

In the next section, we propose some examples of interest to help clarify the differences between the two hypergeometric distributions and suggest their potential applications in the socio-economic and official statistics' fields.

1.2 Examples

The subsidy

The central government allocates funding for households at poverty risk. In a first scenario, assume that families with more than two children are

favoured, i.e. the probability to obtain the subsidy after the request is higher for those families with more than two underage siblings. Moreover, assume that the allocated funding allows subsidising one hundred thousand households. Therefore, the number of subsidised families with two or less than two children and those with more than two, given that their sum is one hundred thousand, will follow WNCH distribution of parameters i) M_1 = the total number of families with more than two children; ii) M_2 = the total number of families with two or less than two children; iii) n = one hundred thousand; iv) w_1^W = the probability of obtaining the subsidy for a family with more than two children once asked for it; v) w_2^W = the probability of receiving the subsidy for a family with two or less than two children once asked for it.

Now assume the government does not decide the amount of funding in advance. Instead, households can request the subsidy for a year, and all the requests are automatically approved. Let us assume that families with more than two children apply for the contribution with a higher probability. After a year, the government observes the total number of subsidised households. In this case, the number of subsidised families with more than two children will follow FNCH distribution with parameters i-iii) same as the previous case; iv) w_1^F = the odds ratio of families with more than two children asking for the subsidy; v) w_2^F = the odds ratio of families with two or less than two children asking for the subsidy.

The real estate market

In a city, there exist two neighbourhoods, i.e. the Centre (C) and the Periphery (P). For the sake of simplicity, assume that the number of empty houses in the two neighbourhoods is the same, i.e. $M_C = M_P$; the properties are identical and have the same price.
As a first option, assume that the monthly goal of the only agency in the city is to sell $n = 20$ properties. Since the real estate agent's commission is higher when selling houses in the Centre, they will accept appointments of the buyers interested in such kind of houses[1] more likely. The number of houses sold in the Centre will follow WNCH distribution.

As a second option, assume that properties' sales are open for a month.

[1]We are assuming that the customer will buy the house with probability equal to 1, after visiting it.

Since the Centre has more services, it is more likely that buyers will be interested in purchasing houses there. At the end of the month, the real estate agency observes the total number of houses sold is twenty, i.e. $n = 20$. Hence, the number of sold properties in the Centre will follow FNCH distribution.
In both cases the parameters of the distributions will be M_C, $M_P = M_C$, $n = 20$, and w_C^*, w_P^* such that $w_C^* > w_P^*$, where * stands for either W or F.

The prestigious M.Sc. program

The Department of Statistics of a prestigious University opens the enrolment to its M.Sc. program, which is known to be very challenging. For this reason, students may apply only if during their Bachelor they have registered an average grade of either A or B.
Firstly, assume that the program can host a limited number of students, e.g. 20; the Department wants to guarantee a certain level of heterogeneity within them, including students with grades A and B. However, it gives the "priority" to the best performing students, preferring those with A. The number of students of the two groups enrolled on the program will follow WNCH distribution of parameters i) M_A = total number of applying students with an average grade of A; ii) M_B = total number of applying students with an average grade of B; iii) $n = 20$; iv-v) w_A^W and w_B^W such that $w_A^W > w_B^W$.

As a second option, assume that the program does not contemplate a limited number of students. However, since the course is very challenging, a sort of self-selection occurs: the odds ratio of students with grade A applying for the program is higher than that of students with average grade B. At the end of the enrolment period, the Department observe the total number of those enrolled, e.g. 20. The number of best-performing students enrolled in the M.Sc program will follow FNCH distribution with M_A and M_B defined as in the Wallenius case; $n = 20$; w_A^F and w_B^F such that their ratio is greater than 1.

The surveys

The Global Statistical Institute (GSI) collects two samples of individuals for two different scopes:

- it wants to study the association between individuals' hair colour and their income;
- it wants to infer which eyes colour is dominant in the world.

In the first case, assume that the GSI knows that those with dark hair are twice more common than those with light hair thanks to some extra-experimental information, such as previous censuses. To have a significant number of light-haired people in a sample of size n, the interviewer validates their participation in the survey twice more probably than people with dark hair. Once collected n interviews, the number of dark hair individuals in the sample will follow WNCH distribution with parameters i) M_D = total number of dark-haired people; ii) M_L = total number of light-haired people; iii) n = sample size; iv) w_D^W and v) w_L^W such that $w_D^W = 2w_L^W$.

In the second case, the GSI knows nothing in advance and wants to collect as many interviews as possible. It assigns the task to its office in North Africa, where it is five times more likely to find a person with brown or black eyes than one with blue or green eyes. At the end of the period, the GSI observes that the interviewer has collected n interviews; the number of people with brown eyes in the sample will follow FNCH distribution of parameters i) M_{BB} = total number of people with brown or black eyes in the world; ii) M_{BG} = total number of people with blue or green eyes in the world; iii) n = sample size; iv) w_{BB}^F and v) w_{BG}^F such that $w_{BB}^F = 5w_{BG}^F$.

This last case will be the critical issue of the next chapter.

Chapter 2

Fisher's noncentral hypergeometric distribution for the size estimation of population's subgroups

This chapter presents a methodology based on Fisher's noncentral hypergeometric distribution that aims at the size estimation of a *heterogeneous* population. In this framework, we say that a population is *homogeneous* if the probability of being registered in a list is the same for any individual i, $i = 1, ..., N$ belonging to the target population of (unknown) size N; otherwise, the population is said to be heterogeneous. We allow such probability to vary across different "capture occasions" [1].

Capture probabilities may vary among individuals due to some measurable attributes, e.g. sex or age, or given unmeasurable characteristics (Johnson et al. (1986)). There is a vast literature on estimating population size in the presence of heterogeneity when the source of such heterogeneity is unknown. Chronologically speaking, one of the first approaches in modelling heterogeneity in the capture-recapture context is the random-effects one, introduced by Darroch et al. (1993) and Agresti (1994); for Bayesian versions, see Fienberg et al. (1999) and Basu and Ebrahimi (2001). Based on the Rasch model, such an approach overcome the traditional log-linear models' inability to consider the depen-

[1] This situation is that described by model M_{th} in Otis et al. (1978).

dence given by capture heterogeneity properly. A different and more recent approach is the latent classes modelling: see Bartolucci et al. (2004), Di Cecco (2019) and Di Cecco et al. (2020)[2]. For a nonparametric approach to the latent class models, see Johndrow et al. (2016) and Manrique-Vallier (2016).

In the official statistics' field, there are plenty of specific applications of capture-recapture models in the presence of heterogeneity. Chiariello and Tuoto (2018) succeed in estimating the size of the hidden criminal population in Italy during 2006-2014, accounting for heterogeneity by including a set of available individual covariates. Kaskasamkul and Böhning (2018) estimate the size of the illegal immigrant population in the Netherlands, allowing for heterogeneity. The just cited Manrique-Vallier (2016) proves how the estimates of the number of killings during the Kosovo war in 1999 improve once we account for the possibility that capture probabilities vary among individuals. Such difference in capture probabilities might be due to different "weights" each group of the target population has in the various capture occasions; in other words, "the presence of capture heterogeneity is equivalent to bias in the sampling process" (Johndrow et al. (2016)). The noncentral hypergeometric (NCH) distributions described in the previous chapter arise naturally in such situations. However, to our knowledge, they have not been used in official statistics yet.

Such distributions have been underemployed in the statistical literature mainly because of the computational complexity given by their probability mass functions. Nowadays, modern computational tools allow for exploiting such distributions, suitable for various contexts; their main (recent) applications are in natural sciences - genomics, genetics, physics (e.g. Lodato et al. (2018), Barrett et al. (2019)). Another obstacle to the extensive use of such distributions has been the rooted confusion concerning the existence of two different NCH distributions, now known as Wallenius' and Fisher's NCH (see Fog (2008a)). As clarified in chapter 1, the major difference between the two lies in the draws' dependence structure: if a ball's draw affects another's, i.e. the balls compete among them, then the draws follow Wallenius' distribution; yet, independent

[2]The latent classes approach can nimbly adapt to the overcoverage problem; we discuss it in detail in Chapter 3

draws follow Fisher's distribution. For this reason, Wallenius' distribution fits perfectly, for instance, in a context of preference or ranking data - see the very recent Grazian et al. (2019). On the other hand, Fisher's distribution potentially suits the population size estimation problems; indeed, we aim to give Fisher's NCH a new guise, using it in the context of the official statistics.
Suppose a list partially enumerates a population's subgroups; the objective is to estimate the total number of individuals belonging to those subpopulations. Likely, the different groups, or categories, have not the same weight in the same capture occasion; however, experts may express their opinion about the presence of such different groups in terms of relative odds.

This chapter examines how to infer a target population size when observing at least one sample, assuming it to be Fisher's NCH distributed. Here the source of heterogeneity is known, and the attributes are available; however, there is a lack of units' unique labels. Hence, even when multiple lists are available, the use of log-linear models is not feasible.

In this work, the Bayesian methodology allows us to estimate the posterior distribution of the population size, exploiting both the possibility of including extra-experimental information and the computational efficiency of MCMC and ABC methods when dealing with distributions as complex as FNCH.
We first present different scenarios concerning the univariate FNCH, and then we extend the framework to the multivariate case. The model's applicability to the case in which only one list is available makes it suitable to estimate elusive target populations, e.g. non-resident inhabitants in a city, homeless people, irregular migrants, unemployed people.
Precisely the last example is the object of the case study in §2.4, which is the result of an in-progress collaboration with Stefano De Santis (Istat).
The following sections describe how to infer Fisher's NCH parameters with a Bayesian approach, both in the univariate (§2.1) and multivariate (§2.2) cases. For the multivariate case, we propose two methods, which are compared in §2.3.

2.1 The univariate case

A list enumerates n units belonging to a population of unknown size N. Assume that the population comprises only two groups of sizes M_1 and M_2, and whose number of units captured by the list is X_1 and X_2, respectively. Hence, $M_1 + M_2 = N$ and $X_1 + X_2 = n$. Exactly as in the example of the coloured balls presented in chapter 1, assume X_1 and X_2 to follow a Binomial distribution, as in (1.6). Therefore, the distribution of X_1 conditional to the sum $X_1 + X_2 = n$ is univariate Fisher's noncentral hypergeometric, with probability mass function described by equation (1.8). The weight parameters, or odds, are defined by equation (1.9), and the odds ratio can be seen as a measure of exposure to the register of one population group over the other. With a little abuse of notation, and since we never refer to Wallenius' distribution in this chapter, we will write w instead of w^F to indicate the Fisher's weight parameter.

2.1.1 Prior setting

It is reasonable to assume that experts, e.g. who collected the data, can give their opinion about the odds ratio value. Hence, w may be fixed, or we may subjectively elicit a prior distribution for it. The subjective elicitation is a debated issue since the attribute "subjective" is often perceived as including personal beliefs in a negative sense. Instead, making the elicitation process a rational way to incorporate experts' knowledge and take advantage of their experience. For a deep and detailed discussion about the probabilities' elicitation process, see Berger (1985) and O'Hagan et al. (2006).

Even fixing the value of w, if everything else is unknown, we can only estimate the relative size of the two groups in the population; we show it from the empirical point of view in Appendix B. Nevertheless, we may have some prior information on one of the groups: such a situation is prevalent when dealing with administrative data. Indeed, consider a sample of resident (group 1) and non-resident (group 2) persons living in a city; assuming M_1 known will be a proper assumption. Yet, if we have a sample of self-employed individuals, depending on their working condition, they may have (group 1) or have not (group 2) a VAT number; again, we may reasonably assume M_1 to be known.

The information may be a point or an interval estimate, and we can either fix M_1 or elicit a prior distribution for it.
Hence, to estimate N, we assume the following hierarchical model:

$$\begin{aligned} X_1|X_1+X_2=n &\sim \text{FNCH}(M_1, N-M_1, n, w_1, w_2) \\ N|M_1 &\sim \text{Unif}(M_1+x_2, N^{\text{upper}}) \end{aligned} \tag{2.1}$$

where $N^{\text{upper}} \in \mathbb{Z}$ is a large value. According to the extra-experimental information, $N|M_1$ may have any suitable alternative distribution.
If M_1 is unknown, we can specify, e.g.,

$$M_1 \sim \text{Pois}(\lambda_1) \tag{2.2}$$

or

$$M_1 \sim \text{Unif}(a_1, b_1) \tag{2.3}$$

or assume any other suitable distribution. In the next paragraphs, we test the results' sensitivity to the prior specification and provide results for both (2.2) and (2.3).

If also w_1, w_2 are unknown, we include an informative prior distribution for their ratio, i.e. w:

$$w \sim \text{Unif}(a_w, b_w) \tag{2.4}$$

where the hyperparameters a_w and b_w are chosen such that the associated density matches some quantiles that are subjectively estimated (see Berger (1985)). Also for w, we test how different priors impact the results in the following paragraphs. In the case of strong prior information, we could also assume

$$w \sim \text{Normal}(\mu_w, \sigma_w^2)\ . \tag{2.5}$$

We denote with $\pi(N, M_1, w)$ the joint prior distribution, which can be factorised into $\pi(N|M_1)\pi(M_1)\pi(w)$ assuming that the odds ratio for the two groups of being included in the sample is independent on the groups' sizes:

$$\frac{\zeta_1/(1-\zeta_1)}{\zeta_2/(1-\zeta_2)} \perp N, M_1\ . \tag{2.6}$$

2.1.2 Posterior computation

The joint posterior distribution will be

$$\begin{aligned}\pi(M_1, N, w|x_1, n) &\propto L(x_1|M_1, N, n, w)\pi(M_1, N, w) \\ &\propto L(x_1|M_1, N, n, w)\pi(N|M_1)\pi(M_1)\pi(w) \ ;\end{aligned} \tag{2.7}$$

we compute it using a Metropolis-within-Gibbs algorithm. The following boxes show the algorithms in the cases of

- both M_1 and w known: in this case, we run a Metropolis-Hastings since we only estimate N (Algorithm 1);
- both M_1 and w unknown (Algorithm 3).

We may also assume the "mixed" situations, for which the posterior computation derives directly from the two described above; see Algorithm 2.

Algorithm 1: Metropolis-Hastings, known M_1 and w

```
Choose initial value N^(0) ;
for t ← 1 to T do
    draw N* from a proposal distribution q_t(N*|N^(t-1)) ;
    compute the acceptance ratio
      γ_N = min(1; π(N*|M_1,w,n,x_1)π(N*) / π(N^(t-1)|M_1,w,n,x_1)π(N^(t-1)) · q_t(N^(t-1)|N*) / q_t(N*|N^(t-1))) ;
    draw u ~ Unif(0,1) ;
    if γ_N > u then
        set N^t = N*;
    else
        set N^t = N^(t-1)
    end
end
```

Algorithm 2: Metropolis-within-Gibbs, unknown M_1

1 Choose initial values $N^{(0)}$, M_1^0 ;
2 **for** $t \leftarrow 1$ **to** T **do**
3 draw M_1^* from a proposal distribution $q_t(M_1^*|M_1^{t-1})$, e.g. $M_1^* \sim \text{Pois}(M_1^{t-1})$;
4 compute the acceptance ratio
5 $\gamma_{M_1} = \min\left(1; \frac{\pi(M_1^*|N^{t-1},w,n,x_1)}{\pi(M_1^{t-1}|N^{t-1},w,n,x_1)} \frac{q_t(M_1^{t-1}|M_1^*)}{q_t(M_1^*|M_1^{t-1})}\right)$;
6 draw $u \sim \text{Unif}(0,1)$;
7 **if** $\gamma_{M_1} > u$ **then**
8 set $M_1^t = M_1^*$;
9 **else**
10 set $M_1^t = M_1^{t-1}$
11 **end**
12 draw N^* from a proposal distribution $q_t(N^*|N^{t-1})$, e.g. $N^* \sim \text{Pois}(N^{t-1})$;
13 compute the acceptance ratio $\gamma_N = \min\left(1; \frac{\pi(N^*|M_1^t,w,n,x_1)}{\pi(N^{t-1}|M_1^t,w,n,x_1)} \frac{q_t(N^{t-1}|N^*)}{q_t(N^*|N^{t-1})}\right)$;
14 repeat lines 6-11 for N using γ_N
15 **end**

Algorithm 3: Metropolis-within-Gibbs, unknown M_1 and w

1 Choose initial values $N^{(0)}$, $M_1^{(0)}$ and $w^{(0)}$;
2 **for** $t \leftarrow 1$ **to** T **do**
3 draw M_1^* from a proposal distribution $q_t(M_1^*|M_1^{t-1})$, e.g. $M_1^* \sim \text{Pois}(M_1^{t-1})$;
4 compute the acceptance ratio
5 $\gamma_{M_1} = \min\left(1; \frac{\pi(M_1^*|N^{t-1},w^{t-1},n,x_1)}{\pi(M_1^{t-1}|N^{t-1},w^{t-1},n,x_1)} \frac{q_t(M_1^{t-1}|M_1^*)}{q_t(M_1^*|M_1^{t-1})}\right)$;
6 draw $u \sim \text{Unif}(0,1)$;
7 **if** $\gamma_{M_1} > u$ **then**
8 set $M_1^t = M_1^*$;
9 **else**
10 set $M_1^t = M_1^{t-1}$
11 **end**
12 draw N^* from a proposal distribution $q_t(N^*|N^{t-1})$, e.g. $N^* \sim \text{Pois}(N^{t-1})$;
13 compute the acceptance ratio $\gamma_N = \min\left(1; \frac{\pi(N^*|M_1^t,w^{t-1},n,x_1)}{\pi(N^{t-1}|M_1^t,w^{t-1},n,x_1)} \frac{q_t(N^{t-1}|N^*)}{q_t(N^*|N^{t-1})}\right)$;
14 repeat lines 6-11 for N using γ_N ;
15 draw w^* from a proposal distribution $q_t(w^*|w^{t-1})$, e.g. $w^* \sim \text{Norm}(w^{t-1}, s_w^2)$;
16 compute the acceptance ratio $\gamma_w = \min\left(1; \frac{\pi(w^*|N^t,M_1^t,n,x_1)}{\pi(w^{t-1}|N^t,M_1^t,n,x_1)}\right)$;
17 repeat lines 6-11 for w using γ_w
18 **end**

2.1.3 Sensitivity analysis: the posterior distribution of N under the different specifications of M_1 with fixed w

Here we show the sensitivity of N to the different specifications of M_1 (as in (2.2) and (2.3)). We simulate 200 observed vectors (x_1, x_2) and test the model for three population sizes, i.e. $N = 1000$ (purple in figures), $N = 10000$ (blue) and $N = 100000$ (green). Figures 2.1, 2.2 and 2.3 show the distribution of the posterior means of N estimated on the 200 samples. Table 2.1 shows the mean, the standard deviation and the 95% Highest Posterior Density interval of the posterior means of N obtained via simulation.

The posterior of N strongly reflects M_1's prior uncertainty; the wider the prior, the greater the standard deviation associated with the posterior and the larger the Highest Posterior Density interval for the population size. It appears more evident as the population size increases.

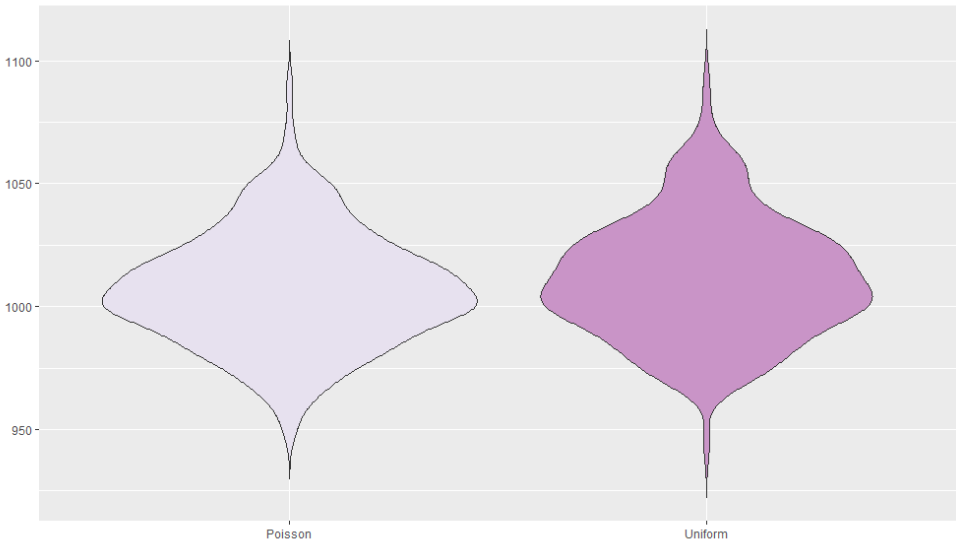


Figure 2.1: Posterior mean of N for different specification of the M_1 prior, 200 samples. True value of $N = 1000$

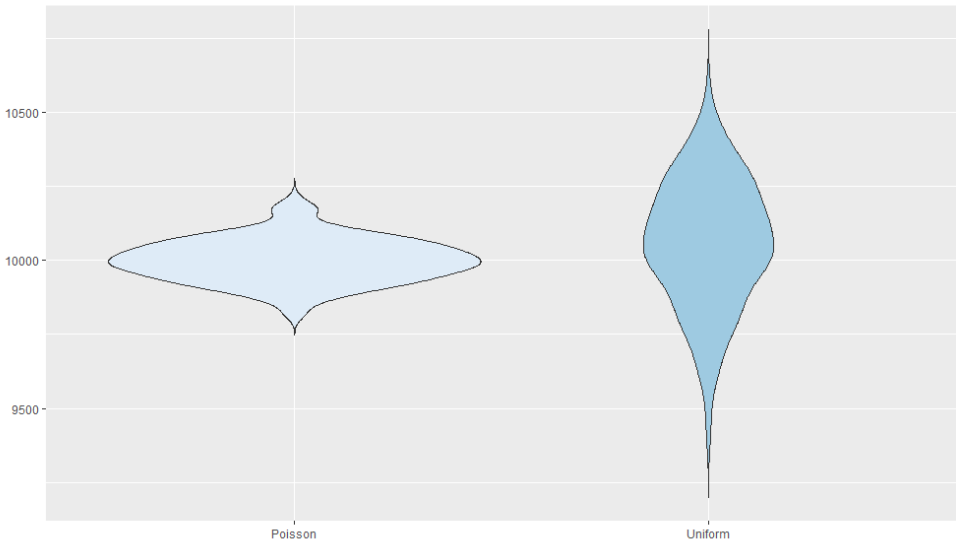


Figure 2.2: Posterior mean of N for different specification of the M_1 prior, 200 samples. True value of $N = 10000$

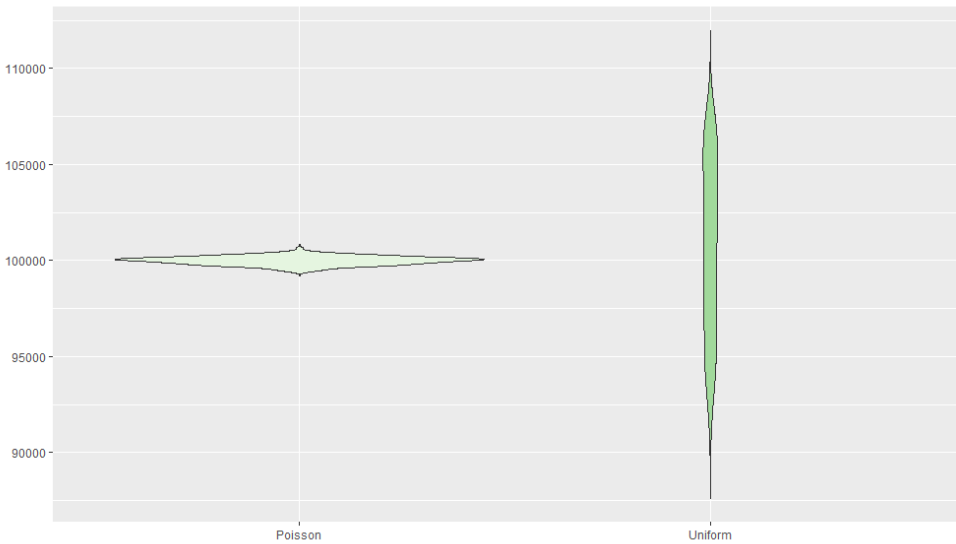


Figure 2.3: Posterior mean of N for different specification of the M_1 prior, 200 samples. True value of $N = 100000$

Prior on M_1	True value of N	Mean	Sd	HPD
	1000	1006.823	23.009	$[971.74; 1057.226]$
Pois(m_1^*)	10000	10000.53	72.833	$[9881.172; 10174.839]$
	100000	99995.29	223.763	$[99543.02; 100412.41]$
	1000	1011.091	24.381	$[965.873; 1061.187]$
Unif($m_1^* \pm 20\% m_1^*$)	10000	10049.64	212.101	$[9602.222; 10419.035]$
	100000	100616.2	4493.005	$[93533.26; 107774.07]$

Table 2.1: Sensitivity of the posterior mean of N to different prior specifications of M_1. Mean, standard deviation and 95% Highest Posterior Density interval for the posterior mean of N, estimated on 200 samples for $N = 1000, 10000$ and 100000, with $m_1^* = 266, 2655, 26551$ being M_1 true values in the respective cases.

Prior on w	True value of N	Mean	Sd	HPD
	1000	1006.764	23.163	$[969.348; 1056.271]$
Unif($w^* \pm 20\% w^*$)	10000	10013.47	74.727	$[9878.628; 10183.178]$
	100000	100120.3	532.9	$[99177.34; 101224.76]$
	1000	1015.297	24.397	$[971.147; 1060.923]$
Unif($w^* \pm 50\% w^*$)	10000	10105.36	126.209	$[9799.559; 10314.797]$
	100000	101128	2522.769	$[96321.36; 105173.56]$

Table 2.2: Sensitivity of the posterior mean of N to different prior specifications of w. Mean, standard deviation and 95% Highest Posterior Density interval for the posterior mean of N, estimated on 200 samples for $N = 1000, 10000$ and 100000. $w^* = 0.188$.

2.1.4 Sensitivity analysis: the posterior distribution of N under different specifications of w with fixed M_1

Now we study how N is sensitive to the prior of w getting larger, with fixed M_1; in particular, we set a Uniform prior for w, centred on the true value, and observe the impact of widening its support. We test it for $N = 1000$ (Figure 2.4), $N = 10000$ (Figure 2.5), and $N = 100000$ (Figure 2.6).
When $N = 1000$, there is no significant difference between the two posterior means' distributions; the standard deviation associated with the widest distribution is only about 1.2 points bigger; see Table 2.2. However, as the population size increases the prior uncertainty of w highly affects the posterior estimates of N; the difference in the standard deviations is about 51.5 for $N = 10000$ and just under 2000 for $N = 100000$.

Figure 2.4: Posterior mean of N, 200 samples, M_1 fixed. True value of $N = 1000$.

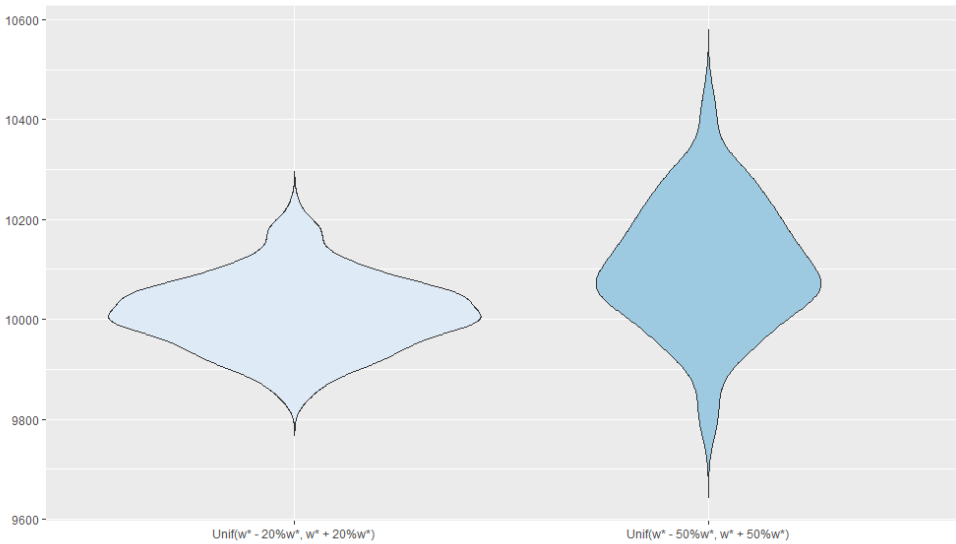


Figure 2.5: Posterior mean of N, 200 samples, M_1 fixed. True value of $N = 10000$.

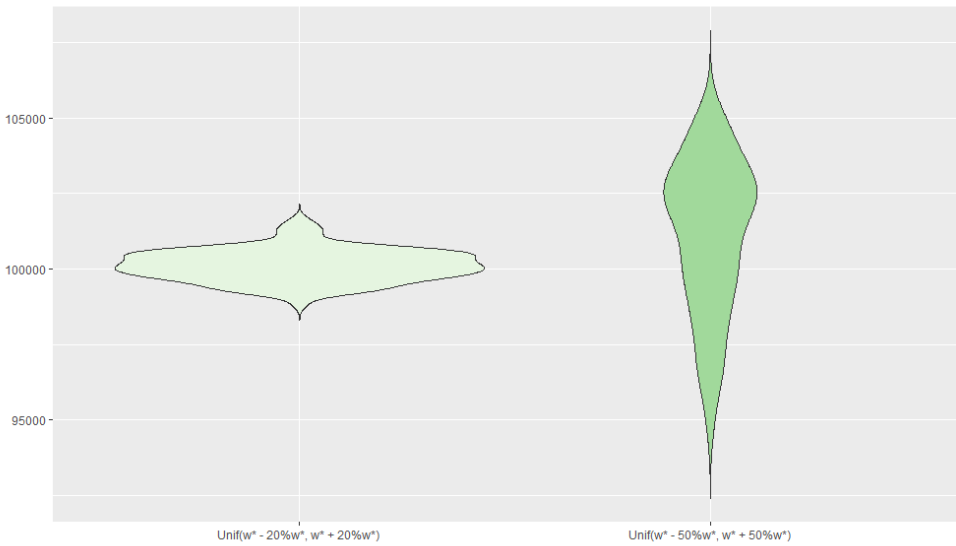


Figure 2.6: Posterior mean of N, 200 samples, M_1 fixed. True value of $N = 100000$.

2.1.5 Sensitivity analysis: the posterior distributions of N and w under different specifications of w and M_1

This paragraph shows how the posteriors of N and w change as the prior interval specified for w becomes wider and under the two different prior specifications of M_1.
Table 2.3 summarises the posterior mean of N in the separate cases; we report only the smallest and the largest population sizes considered so far. For $N = 1000$, the standard deviation does not vary significantly with the change of the priors; however, introducing more uncertainty leads to an increasing bias of small size (up to 2%) for the posterior mean (see also Figures 2.7 and 2.8). Yet, when the population size is larger, the standard deviation of the posterior mean of N increases with the increase in prior uncertainty of both w and M_1 (see Figures 2.9 and 2.10). In the most uncertain case, the width of the 95% Highest Posterior Density interval amounts to the 20% of the true value.
Table 2.4 summarises the posterior mean of w for each case. As the prior on w widens, the posterior mean suffers a slight upward bias. Nevertheless, the Highest Posterior Density intervals always include the true value w^*. The prior choice for M_1 seems not to affect the posterior of w (see Figures 2.11,2.12,2.13 and 2.14).

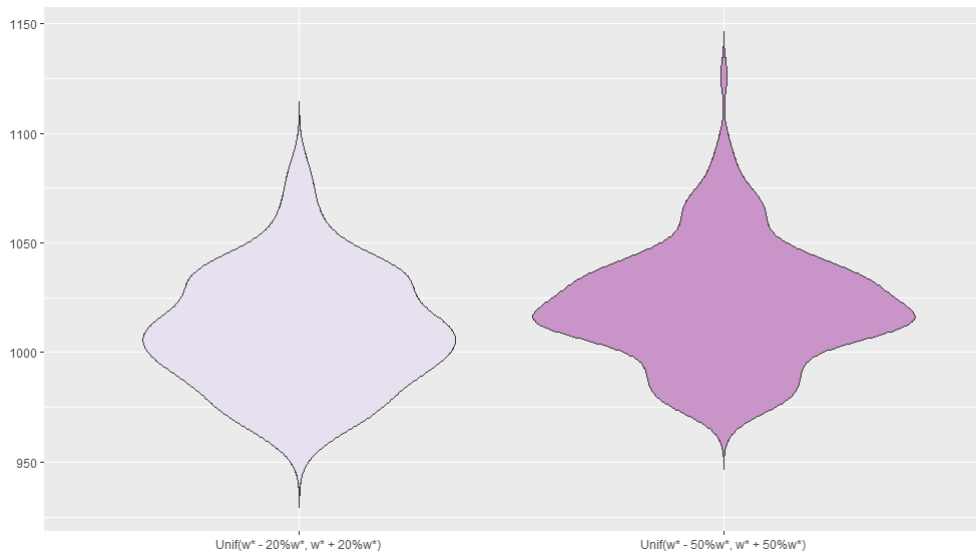


Figure 2.7: Posterior mean of N, 200 samples, with Uniform prior on M_1. True value of $N = 1000$.

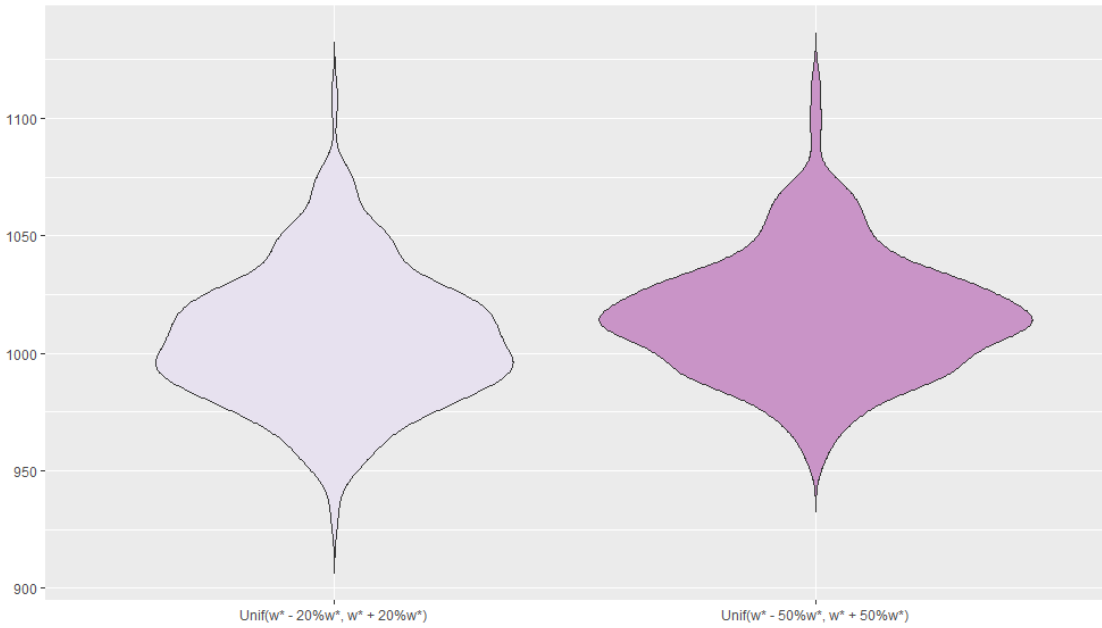


Figure 2.8: Posterior mean of N, 200 samples, with Poisson prior on M_1. True value of $N = 1000$.

$N = 1000$	Prior on M_1	Mean	Sd	HPD
Unif($w^* \pm 20\% w^*$)	Pois(m_1^*)	1006.5	28.242	[952.503; 1060.003]
	Unif($m_1^* \pm 20\% m_1^*$)	1010.479	26.857	[959.571; 1058.614]
Unif($w^* \pm 50\% w^*$)	Pois(m_1^*)	1016.306	26.05	[973.654; 1072.973]
	Unif($m_1^* \pm 20\% m_1^*$)	1021.495	26.609	[974.742; 1072.473]
$N = 100000$				
Unif($w^* \pm 20\% w^*$)	Pois(m_1^*)	100002.9	230.338	[99529.47; 100434.93]
	Unif($m_1^* \pm 20\% m_1^*$)	100368.4	4580.214	[91837.33; 107422.06]
Unif($w^* \pm 50\% w^*$)	Pois(m_1^*)	101052	2543.402	[95828.24; 105080.86]
	Unif($m_1^* \pm 20\% m_1^*$)	101293.4	5750.152	[92120.33; 112927.97]

Table 2.3: Sensitivity of the posterior mean of N to different prior specifications of w and M_1 (as in the previous tables). Mean, standard deviation and 95% Highest Posterior Density interval for the posterior mean of N, estimated on 200 samples for $N = 1000, 100000$. $m_1^* = 266, 26551$ for the respective sizes of N, and $w^* = 0.188$

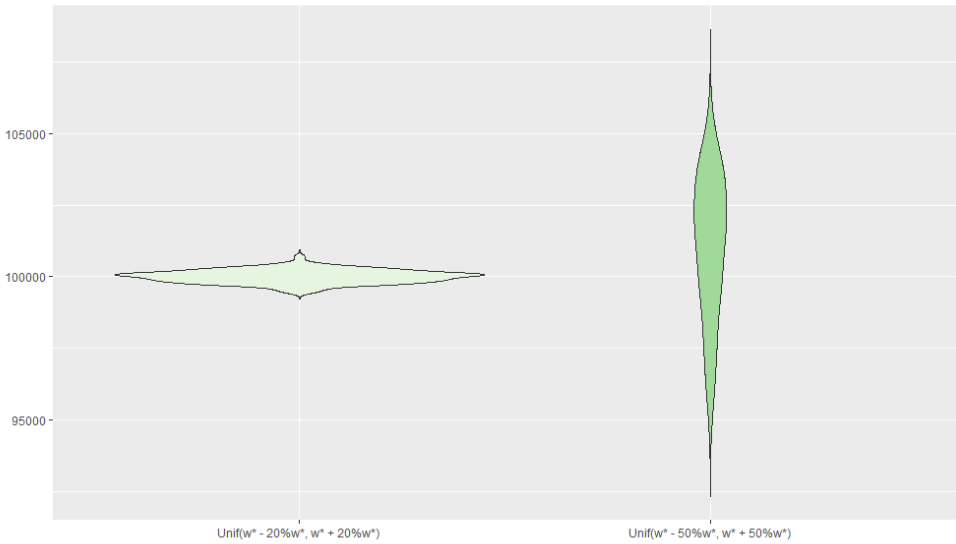


Figure 2.9: Posterior mean of N, 200 samples, with Uniform prior on M_1. True value of $N = 100000$.

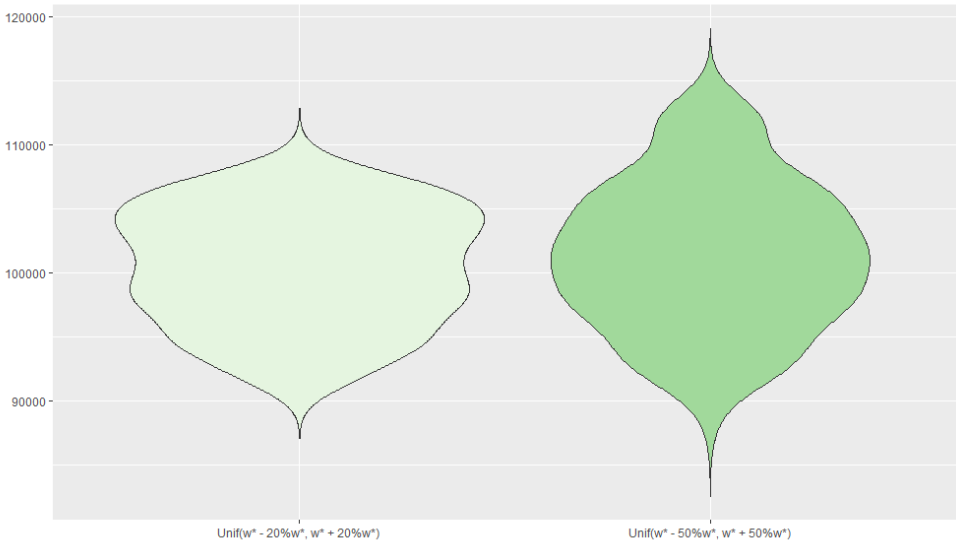


Figure 2.10: Posterior mean of N, 200 samples, with Poisson prior on M_1. True value of $N = 100000$.

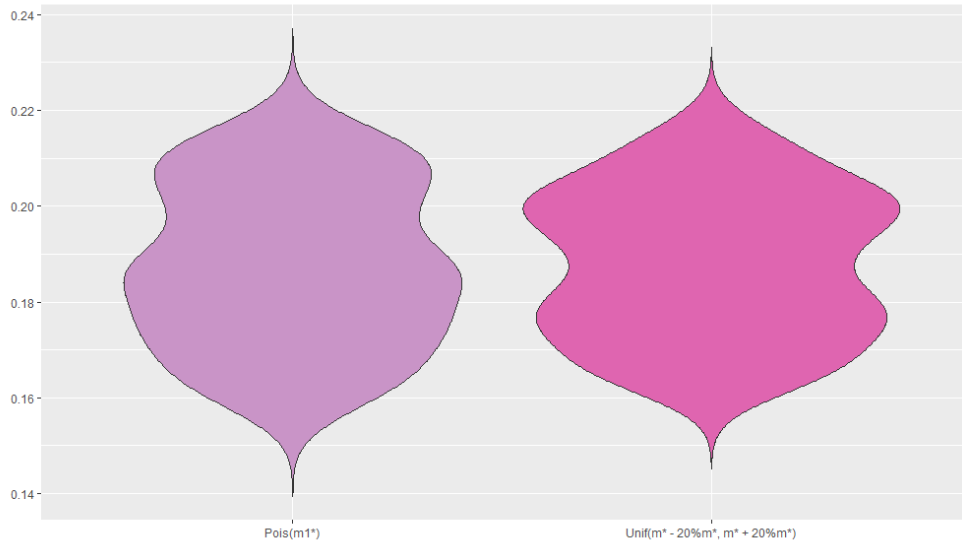


Figure 2.11: Posterior mean of w, 200 samples, with M_1 Poisson (left) or Uniform (right). $w \sim \text{Unif}(w^* \pm 20\% w^*)$, with $w^* = 0.188$. $N = 1000$

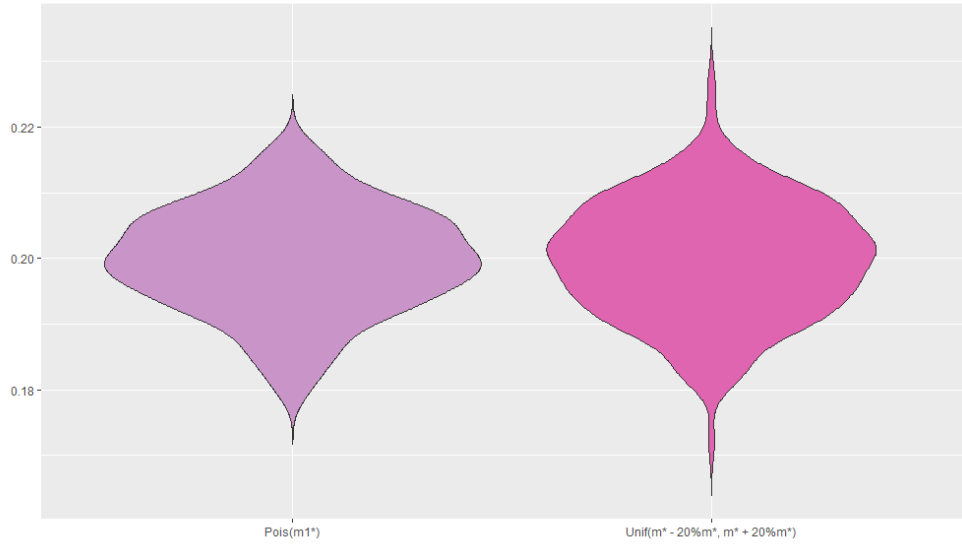


Figure 2.12: Posterior mean of w, 200 samples, with M_1 Poisson (left) or Uniform (right). $w \sim \text{Unif}(w^* \pm 50\% w^*)$, with $w^* = 0.188$. $N = 1000$

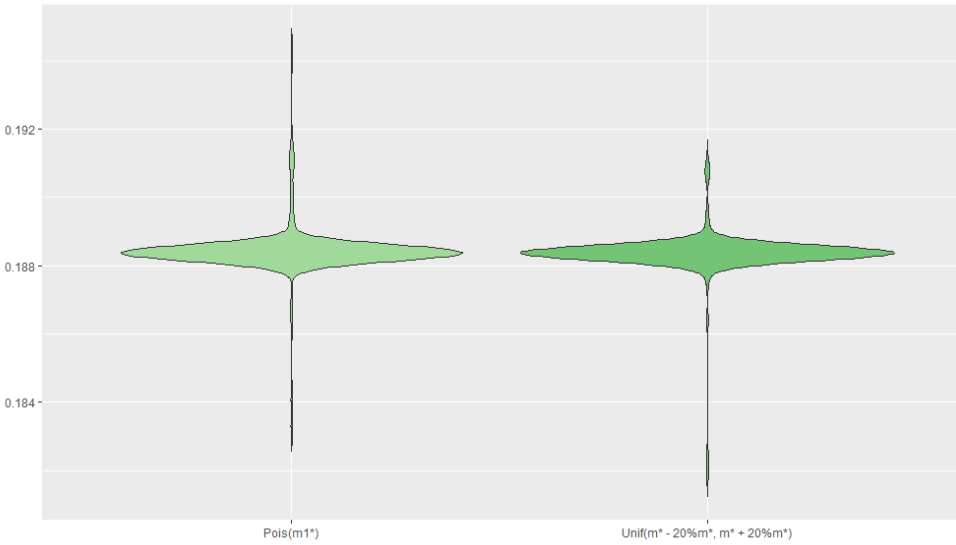


Figure 2.13: Posterior mean of w, 200 samples, with M_1 Poisson (left) or Uniform (right). $w \sim \text{Unif}(w^* \pm 20\% w^*)$, with $w^* = 0.188$. $N = 100000$

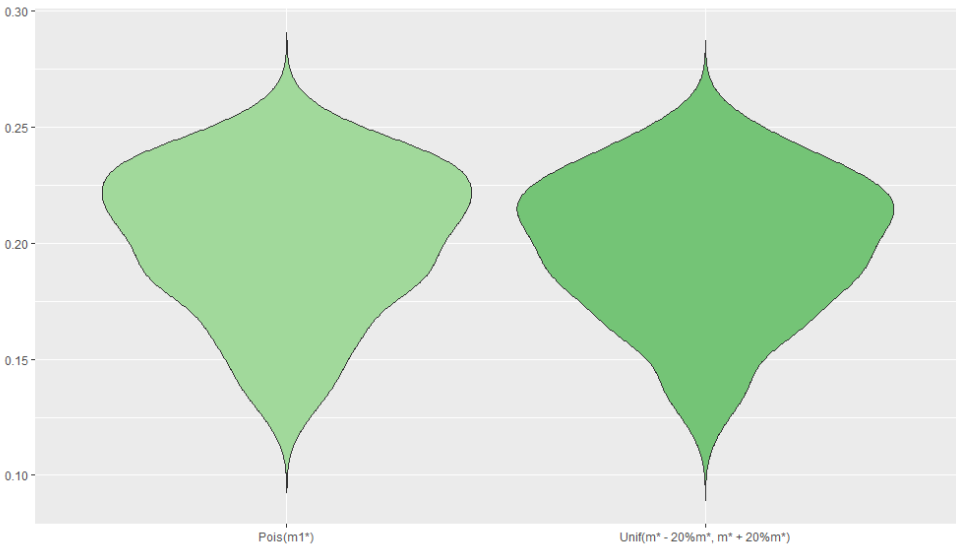


Figure 2.14: Posterior mean of w, 200 samples, with M_1 Poisson (left) or Uniform (right). $w \sim \text{Unif}(w^* \pm 50\% w^*)$, with $w^* = 0.188$. $N = 100000$

$N = 1000$	Prior on $M1$	Mean	Sd	HPD
Unif($w^* \pm 20\% w^*$)	Pois(m_1^*)	0.187	0.017	$[0.159; 0.216]$
	Unif($m_1^* \pm 20\% m_1^*$)	0.188	0.016	$[0.162; 0.213]$
Unif($w^* \pm 50\% w^*$)	Pois(m_1^*)	0.2	0.008	$[0.185; 0.216]$
	Unif($m_1^* \pm 20\% m_1^*$)	0.2	0.009	$[0.181; 0.214]$
$N = 100000$				
Unif($w^* \pm 20\% w^*$)	Pois(m_1^*)	0.188	0.001	$[0.188; 0.19]$
	Unif($m_1^* \pm 20\% m_1^*$)	0.188	0.001	$[0.188; 0.189]$
Unif($w^* \pm 50\% w^*$)	Pois(m_1^*)	0.202	0.032	$[0.136; 0.251]$
	Unif($m_1^* \pm 20\% m_1^*$)	0.199	0.031	$[0.138; 0.252]$

Table 2.4: Sensitivity of the posterior mean of w to different prior specifications of w and M_1 (as in the previous tables). Mean, standard deviation and 95% Highest Posterior Density interval for the posterior mean of w, estimated on 200 samples for $N = 1000, 100000$. $m_1^* = 266, 26551$ for the respective sizes of N, and $w^* = 0.188$

2.1.6 Multiple lists

So far, we have addressed the case where a single list is available, which is perhaps the case of most interest. However, the findings discussed in the previous sections are also applicable when there is more than one list, but we lack unique identifiers. Such a situation lies outside the capture-recapture framework, and popular models like the log-linear ones are not feasible. In this paragraph, we briefly discuss the multiple lists' case.

Assume K lists of size n_k partially enumerate a population composed of two subgroups, which are differently exposed to each list $k, k = 1, 2, ..., K$. Denote with $\boldsymbol{X}_k \in \mathbb{R}^2$ the bi-dimensional vector whose elements are the two subgroups' number of units observed at the kth occasion, i.e.

$$\boldsymbol{X}_k := (X_{k,1}, X_{k,2}) \tag{2.8}$$

for each k. Clearly,

$$X_{k,1} + X_{k,2} = n_k \tag{2.9}$$

We assume

$$X_{k,1}|X_{k,1} + X_{k,2} = n_k \sim \text{FNCH}(M_1, M_2, n_k, w_k) \tag{2.10}$$

where w_k is the relative exposure of subgroup 1 with respect to subgroup 2 at occasion k. We also admit $w_k = w$ for some or all k.
Hence, the likelihood will be

$$L(M_1, M_2, w|\boldsymbol{x}_k) = \prod_{k=1}^{K} \text{FNCH}(\boldsymbol{x}_k; M_1, M_2, n_k, w_k) \tag{2.11}$$

The prior setting is quite similar to the single list case. Algorithm 1, 2 and 3 (as well as the ones we will describe in the next section) nimbly adapt to the multiple lists context.

To check the results' sensitivity to an increase in information, we simulate 100 samples with $K = 1$, $K = 3$ and $K = 10$ and compare the results in terms of variation of summary statistics of the empirical posterior of N. Figure 2.15 clearly shows how the posterior uncertainty of N decreases as K increases. As expected, when more information is included in the model, the HPD intervals shrink; see Table 2.5.

K	Mean	Sd	HPD
1	9995.58	72.634	$[9878.219; 10174.692]$
3	10004.85	43.33	$[9914.891; 10070.752]$
10	9999.989	26.154	$[9953.16; 10046.67]$

Table 2.5: Summaries of the posterior mean of N to different K. Mean, standard deviation and 95% Highest Posterior Density interval for the posterior mean of N, estimated on 100 samples. Poisson prior on M_1 and fixed w.

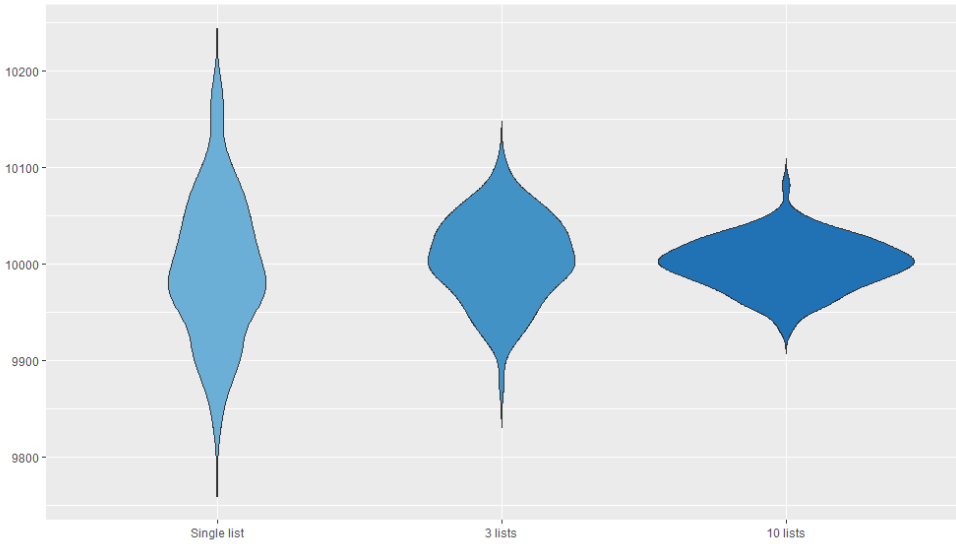


Figure 2.15: Posterior mean of N, 100 samples, for $K = 1, 3, 10$, with Poisson prior on M_1 and w fixed. $N = 10000$

2.2 The multivariate case

2.2.1 Prior setting

Let $\boldsymbol{X} = (X_1, ..., X_c)$ be the vector of the number of units belonging to C different groups, or categories, registered in a list of size n. For each group c, $c = 1, ..., C$, we assume:

$$X_c \sim \text{Binom}(M_c, \zeta_c) \tag{2.12}$$

We indicate with w_c the ratio $\dfrac{\zeta_c}{1-\zeta_c}$, $\forall c$. Note that $\sum\limits_c M_c = N$. Hence, we can say that $\boldsymbol{X}$ is distributed as a multivariate Fisher's NCH with parameters $\boldsymbol{M} = (M_1, ..., M_C)$, n and $\boldsymbol{w} = (w_1, ..., w_C)$ and probability mass function

$$P(\boldsymbol{X} = \boldsymbol{x} | \sum_{c=1}^{C} X_c = n) = \frac{\prod\limits_{c=1}^{C} \binom{M_c}{x_c} {w_c}^{x_c}}{\sum\limits_{\boldsymbol{z} \in \mathcal{Z}} \prod\limits_{c=1}^{C} \binom{M_c}{z_c} {w_c}^{z_c}} \tag{2.13}$$

where

$$\mathcal{Z} = \{\boldsymbol{x} \in \mathbb{Z}^c : \sum_{c=1}^{C} x_c = n \ \cap \ 0 \leq x_c \leq M_c, \forall c\} \tag{2.14}$$

As in the univariate case, we need to introduce some prior information on at least one of the M_c's; for convenience, we refer to such parameter as M_1. The hierarchical model in the multivariate case will be:

$$\begin{aligned} &\boldsymbol{X} | \sum_c X_c = n \sim \text{mvFNCH}(\boldsymbol{M}, n, \boldsymbol{w}) \\ &M_1, ..., M_c \overset{ind}{\sim} f(m_c) \end{aligned} \tag{2.15}$$

where $f(\cdot)$ stands for any suitable distribution. The vector $\boldsymbol{w}$ can be either known or unknown. For simplicity, we fix $\boldsymbol{w}$ in this section, but the extension to the case of unknown weights is straightforward and similar to the univariate case.

As the number of draws and the number of different categories in the population increases, the set $\mathcal{Z}$ (as defined by (2.14)) enlarges; thus, it takes longer to evaluate the probability mass function. For n sufficiently

large, the computational capacity of popular software like `R` may be insufficient, making any method involving the evaluation of the likelihood of such a multivariate variable $\boldsymbol{X}$ practically unfeasible. In the following paragraphs, we propose different methods for estimating $\boldsymbol{M}$.

2.2.2 Posterior computation: ABC method

To avoid evaluating the likelihood, we take inspiration from Grazian et al. (2019). In that paper, the authors use an ABC (Approximate Bayesian Computation) method to estimate the weights of a Wallenius' distribution.
Originally developed by Pritchard et al. (1999), ABC methods have spread enormously for the last two decades thanks to their flexibility. Such methods replace the evaluation of the likelihood with the simulation of a synthetic data set $\boldsymbol{z}$ and the computation of a summary statistics $\eta(\boldsymbol{z})$; then, $\eta(\boldsymbol{z})$ is compared to $\eta(\boldsymbol{x})$, namely the statistics relative to the observed data, on the base of some metric $\rho(\eta(\boldsymbol{x}), \eta(\boldsymbol{z}))$. There exist several reviews of such methods; see Sisson et al. (2018) among others. For the sake of clarity, we describe the most basic ABC algorithm for a generic parameter θ below, i.e. the ABC rejection, or "Vanilla ABC" (as referred to in Clarté et al. (2020)); see Algorithm 4.

Algorithm 4: ABC rejection

```
for t ← 1 to T do
    repeat
        draw θ* from its prior distribution π(θ) ;
        simulate z ~ L(θ|z)
    until ρ(η(x), η(z)) < ε;
    θ^t = θ*
end
```

In Grazian et al. (2019), the use of ABC works around the problem linked to the complexity of Wallenius' NCH density, allowing for sampling from an approximation of the posterior $\pi(\boldsymbol{w}|\boldsymbol{x})$. Although our scope is different, i.e. the target parameter is $\boldsymbol{M}$ instead of $\boldsymbol{w}$, we can use the approach of Grazian et al. (2019) in a different context.
As a summary statistics for data drawn from WNCH distribution, Grazian et al. (2019) propose the arithmetic mean of the observed and simulated

frequency vectors, i.e.

$$\eta(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{x_{k,1}}{n_k}, ..., \frac{x_{k,C}}{n_k}\right) \text{ and } \eta(\boldsymbol{z}) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{z_{k,1}}{n_k}, ..., \frac{z_{k,C}}{n_k}\right) \tag{2.16}$$

where $\boldsymbol{z_k}$ is a vector of counts randomly drawn from a multivariate Wallenius distribution at the kth occasion. Then, to compare the two statistics, Grazian et al. (2019) employ the "distance in variation" (see Brémaud (2013)) metric

$$\rho(\eta(\boldsymbol{x}), \eta(\boldsymbol{z})) = \frac{1}{2}\sum_{c=1}^{C} |\eta(\boldsymbol{x}) - \eta(\boldsymbol{z})| \ . \tag{2.17}$$

In the case of fixed $\boldsymbol{w}$, we strictly follow Grazian et al. (2019). The following figures show the distribution of the posterior means of the parameters, computed over 100 samples simulating a single list case, implying $K = 1$ in (2.16). We fix $N = 10000$ and simulate $\boldsymbol{M} = (M_1, M_2, M_3)$ and the respective ζ_c parameters.
Figure 2.16 shows how well Algorithm 5 estimates the posterior of N (summarised by the expected value). Figure 2.17 shows the posterior of M_1, M_2, M_3 and their relative sizes, again summarised by the expected values.

Algorithm 5: ABC rejection for population subgroups' size estimation

```
for t ← 1 to T do
    repeat
        draw M* from the joint prior distribution π(M*) ;
        simulate z ∼ mvFNCH(M*, n, w) ;
        compute ρ(η(x), η(z))
    until ρ(η(x), η(z)) < ε;
    M^t = M* ;
end
```

As shown in Table 2.6, the estimates obtained via ABC rejection are close to the real values, always included in the HPD intervals. However, there is a trade-off between precision and computational speed: as ε gets smaller, the time required by the algorithm increases. Such a scenario worsen as the dimension of the parameters' vector increases. In the next paragraph, we propose an alternative method that turns out to be more effective.

Parameter's true value	Mean	Sd	HPD
$N = 10000$	10039.7	50.11	$[9954.128; 10134.172]$
$M_1 = 4841$	4844.085	3.093	$[4838.795; 4849.506]$
$M_2 = 3701$	3725.654	41.345	$[3664.424; 3819.45]$
$M_3 = 1458$	1469.914	15.907	$[1425.489; 1488.672]$

Table 2.6: Mean, standard deviation and 95% Highest Posterior Density interval of the posterior mean of N and $\boldsymbol{M}$, estimated on 100 samples via ABC. Poisson prior on M_1 and fixed w.

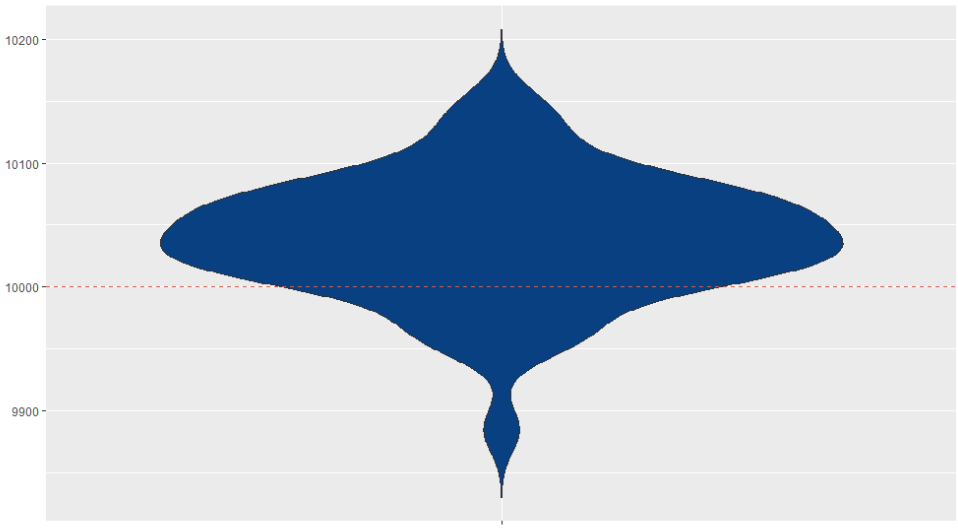
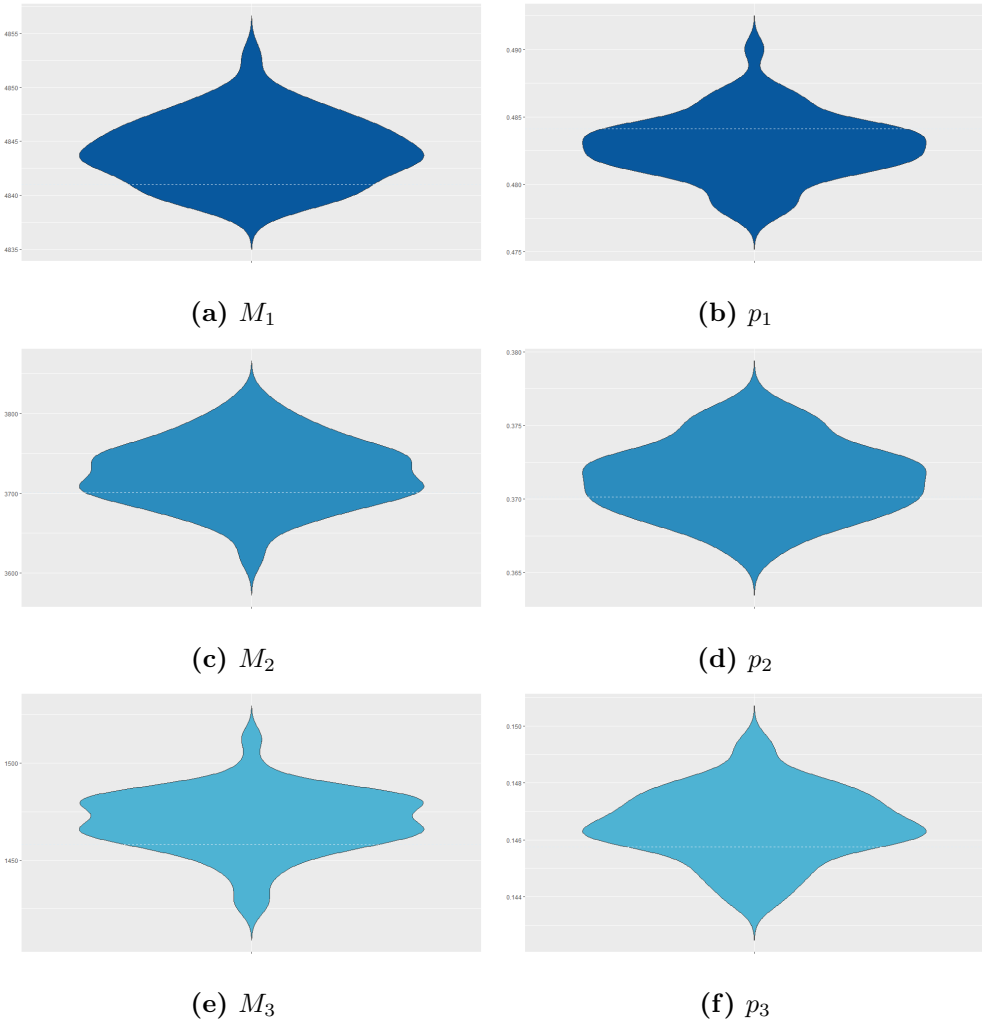


Figure 2.16: Posterior mean of N simulated via ABC, 100 samples. $|\boldsymbol{M}| = 3$



(a) M_1 **(b)** p_1

(c) M_2 **(d)** p_2

(e) M_3 **(f)** p_3

Figure 2.17: Posterior means of $\boldsymbol{M} = (M_1, M_2, M_3)$ (left) and $\boldsymbol{p} = (p_1, p_2, p_3)$ (right) simulated via ABC, 100 samples.

2.2.3 Posterior computation: MCMC method

As an alternative to the algorithm proposed in the previous section, we present a simple Metropolis-within-Gibbs.
At each iteration t, we first simulate M_1^* via Metropolis-Hastings. The acceptance ratio will be

$$\gamma_{M_1} = \min\left(1; \frac{\text{mvFNCH}(\boldsymbol{x}|M_1^*, M_2^{t-1}, ..., M_C^{t-1}, n)\pi(M_1^*)}{\text{mvFNCH}(\boldsymbol{x}|M_1^{t-1}, M_2^{t-1}, ..., M_C^{t-1}, n)\pi(M_1^{t-1})} \times \frac{q_t(M_1^{t-1}|M_1^*)}{q_t(M_1^*|M_1^{t-1})}\right) \tag{2.18}$$

which still involves the evaluation of the multivariate likelihood.
However, we can write the probability mass function of $\boldsymbol{X}$ as

$$\begin{aligned} P(\boldsymbol{X}|\sum_{c=1}^{C} X_c = n) &= P(X_1, X_2, ..., X_C|\sum_{c=1}^{C} X_c = n) \\ &= P(X_1, X_{c'}|\boldsymbol{X}_{-(1,c')}, \sum_{c=1}^{C} X_c = n) \\ &\times P(\boldsymbol{X}_{-(1,c')}|\sum_{c=1}^{C} X_c = n) \\ &= P(X_1, X_{c'}|X_1 + X_{c'} = n - \sum_{c,-(1,c')} X_c) \\ &\times P(\boldsymbol{X}_{-(1,c')}|\sum_{c=1}^{C} X_c = n) \end{aligned} \tag{2.19}$$

where c' can be any $c \neq 1$.

$$P(X_1, X_{c'}|X_1 + X_{c'} = n - \sum_{c,-(1,c')} X_c) \tag{2.20}$$

is the probability mass function of a univariate Fisher's NCH. The ratio in (2.18) then becomes

$$\frac{\text{FNCH}(x_1, x_{c'}|M_1^*, M_{c'}^{t-1}, n_{1c'})f(\boldsymbol{x}_{-(1,c')}|\boldsymbol{M}_{-(1,c')}^{t-1}, n)\pi(M_1^*)}{\text{FNCH}(x_1, x_{c'}|M_1^{t-1}, M_{c'}^{t-1}, n_{1c'})f(\boldsymbol{x}_{-(1,c')}|\boldsymbol{M}_{-(1,c')}^{t-1}, n)\pi(M_1^{t-1})} \times \frac{q_t(M_1^{t-1}|M_1^*)}{q_t(M_1^*|M_1^{t-1})} \tag{2.21}$$

where $n_{1c'} = x_1 + x_{c'}$. We sample the remaining M_c, $c \neq 1$ in the same fashion, always setting $M_{c'} = M_1$; the acceptance ratio γ_{M_c} will be the minimum between 1 and

$$\frac{\text{FNCH}(x_c, x_1 | M_c^*, M_1^t, n_{c1})\pi(M_c^*)}{\text{FNCH}(x_c, x_1 | M_c^{t-1}, M_1^t, n_{c1})\pi(M_c^{t-1})} \frac{q_t(M_c^{t-1}|M_c^*)}{q_t(M_c^*|M_c^{t-1})} \tag{2.22}$$

See Algorithm 6 below for the complete procedure.

Algorithm 6: MCMC

Choose initial values $\boldsymbol{M}^{(0)}$, N^0 ;
for $t \leftarrow 1$ **to** T **do**
  draw M_1^* from a proposal distribution $q_t(M_1^*|M_1^{t-1})$;
  compute the acceptance ratio $\gamma_{M_1} =$
   $\min\left(1; \frac{\text{FNCH}(x_1,x_2|M_1^*,M_2^{t-1},x_1+x_2,w_{12})\pi(M_1^*)}{\text{FNCH}(x_1,x_2|M_1^{t-1},M_2^{t-1},x_1+x_2,w_{12})\pi(M_1^{t-1})} \frac{q_t(M_1^{t-1}|M_1^*)}{q_t(M_1^*|M_1^{t-1})}\right)$
   where $w_{12} = w_1/w_2$;
  draw $u \sim \text{Unif}(0,1)$;
  if $u < \gamma_{M_1}$ **then**
    set $M_1^t = M_1^*$
  else
    set $M_1^t = M_1^{t-1}$
  end
  for $c \leftarrow 2$ **to** C **do**
    draw M_c^* from a proposal distribution $q_t(M_c^*|M_c^{t-1})$;
    compute the acceptance ratio $\gamma_{M_c} =$
     $\min\left(1; \frac{\text{FNCH}\left(x_1,x_c|M_1^t,M_c^*,x_1+x_c,w_{1c}\right)\pi(M_c^*)}{\text{FNCH}\left(x_1,x_c|M_1^t,M_c^{t-1},x_1+x_c,w_{1c}\right)\pi(M_c^{t-1})} \frac{q_t(M_c^{t-1}|M_c^*)}{q_t(M_c^*|M_c^{t-1})}\right)$
     where $w_{1c} = w_1/w_c$;
    draw $u \sim \text{Unif}(0,1)$;
    set $M_c^t = \begin{cases} M_c^* & \text{if } u < \gamma_{N_{M_c}} \\ M_c^{t-1} & \text{otherwise} \end{cases}$;
  end
  set $N^t = \sum_c^C M_c$
end

Figure 2.18 and 2.19 show the distribution of the posterior means of the parameters of interest, estimated over 200 samples, for the three-dimensional case. Table 2.7 presents the Highest Posterior Density intervals for such summaries of the posterior distributions. The results are pretty good.
We will test the method's efficacy compared to the ABC one in higher-dimensional problems in the next section.

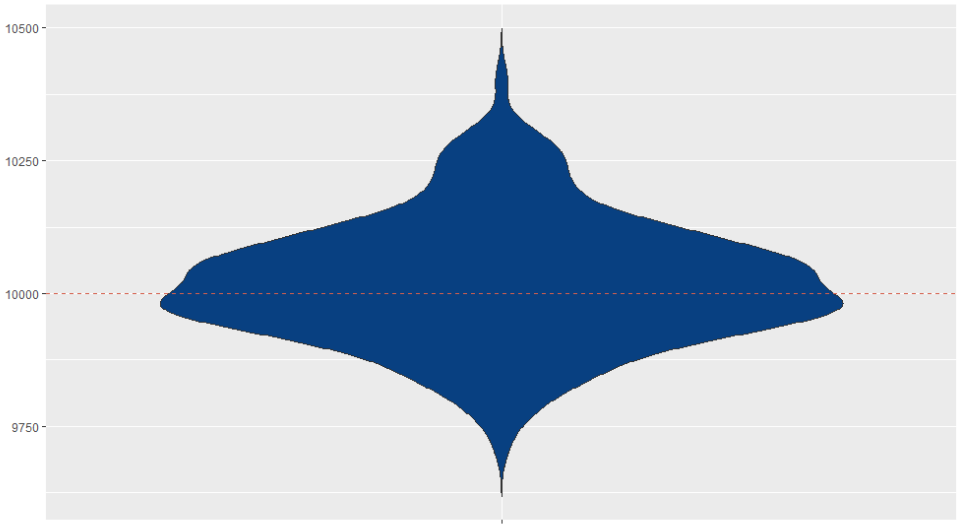
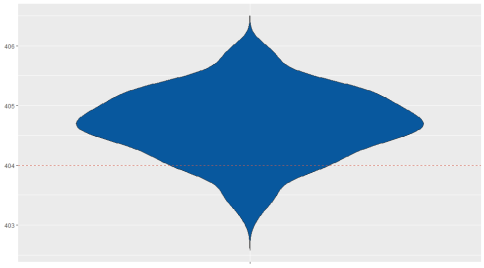


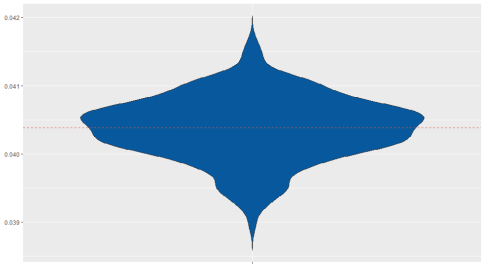
Figure 2.18: Posterior mean of N simulated via MCMC, 100 samples.

Parameter's true value	Mean	Sd	HPD
$N = 10000$	10021.72	116.98	$[9807.55; 10272.85]$
$M_1 = 404$	404.688	0.571	$[403.432; 405.764]$
$M_2 = 4899$	4904.414	33.758	$[4854.575; 4995.473]$
$M_3 = 4697$	4712.617	88.368	$[4561.827; 4893.473]$

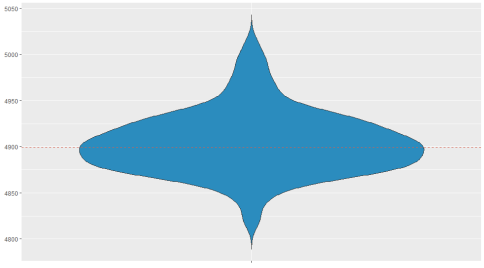
Table 2.7: Mean, standard deviation and 95% Highest Posterior Density interval of the posterior mean of N and $\boldsymbol{M}$, estimated on 100 samples via MCMC. Poisson prior on M_1 and fixed w.



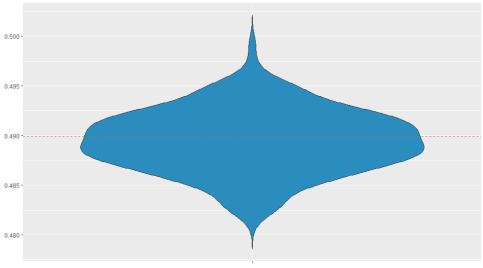
(a) M_1



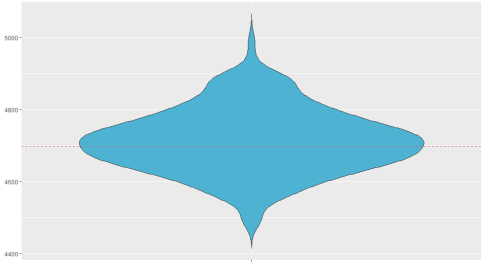
(b) p_1



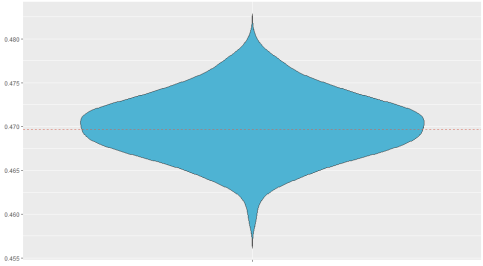
(c) M_2



(d) p_2



(e) M_3



(f) p_3

Figure 2.19: Posterior means of $\boldsymbol{M}$ (left) and $\boldsymbol{p}$ (right) simulated via MCMC, 100 samples.

2.3 Methods comparison: simulation studies

This section compares the results of two simulation studies aiming to estimate the total population size N in the presence of 5 subgroups. For the same true value of $N = 10000$, we implement both the methodologies described in the previous sections.
Figures 2.20, 2.21 and 2.22 clearly show a better ability of the MCMC approach in centring the parameters' true values. However, considering the whole distributions, we observe that the ABC approach's wider intervals always include the true values; see Tables 2.8, 2.9 and 2.10.

	N
MCMC	0.960
ABC	1.000

Table 2.8: The 95% Highest posterior density intervals for N include the true value, frequencies over 100 samples.

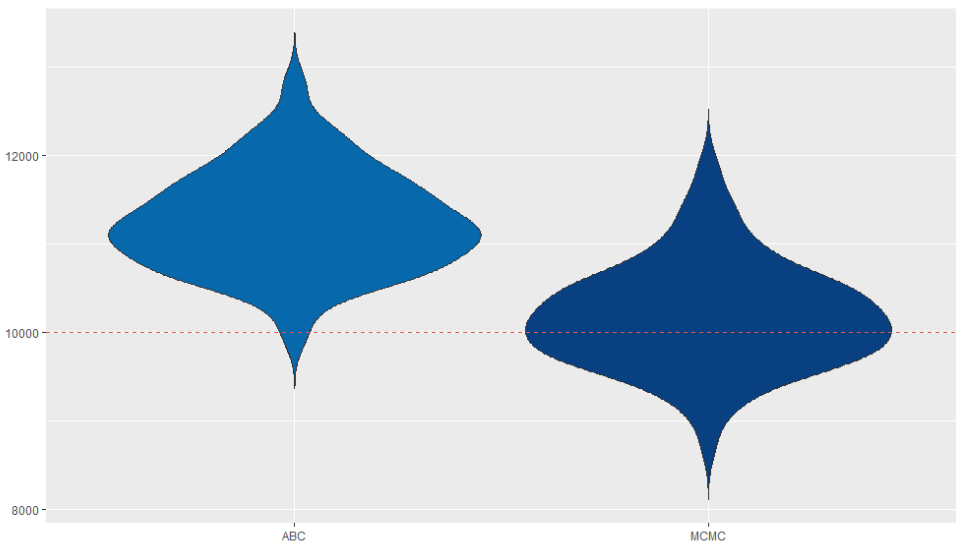
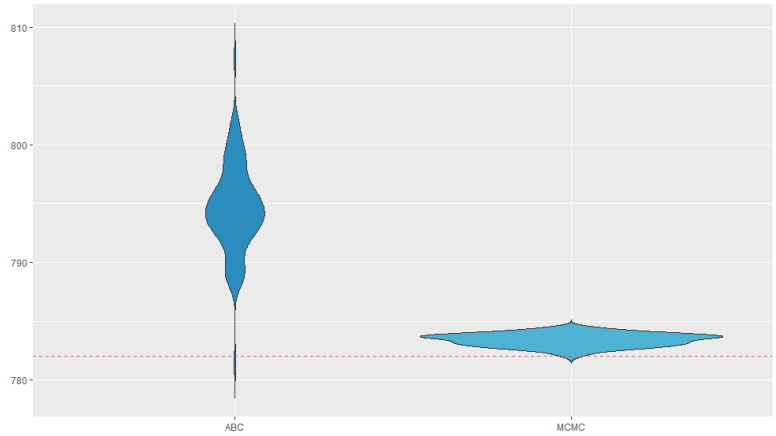


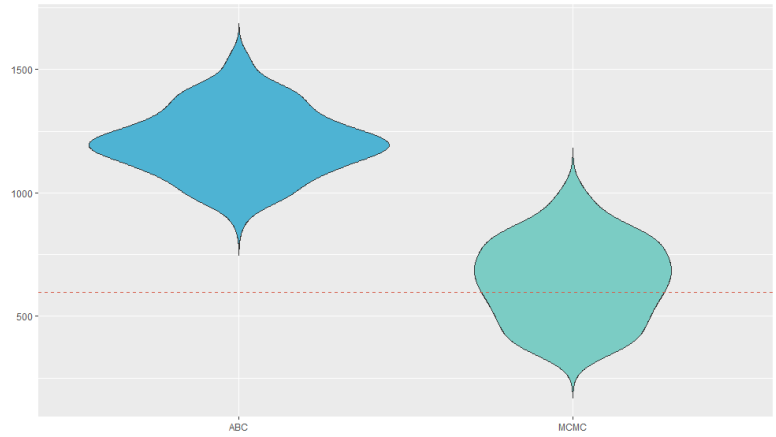
Figure 2.20: Posterior means of N simulated via MCMC (left) and ABC (right), 100 samples

	M_1	M_2	M_3	M_4	M_5
MCMC	1.000	0.990	0.960	0.990	0.960
ABC	1.000	1.000	1.000	1.000	1.000

Table 2.9: The 95% Highest posterior density intervals for $\boldsymbol{M}$ include the true values, frequencies over 100 samples.

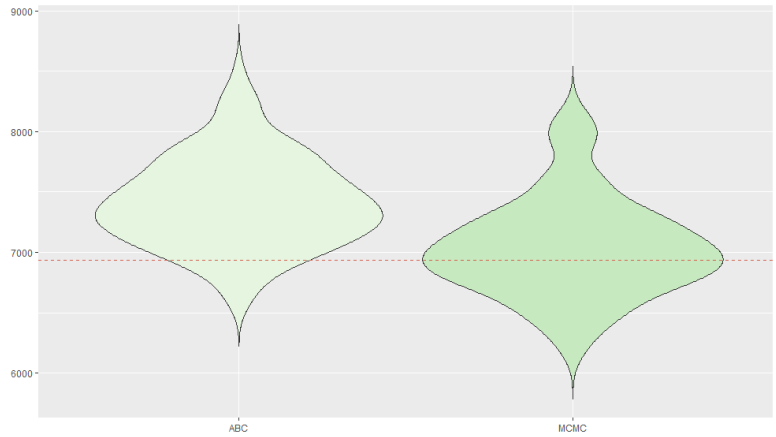


(a) M_1

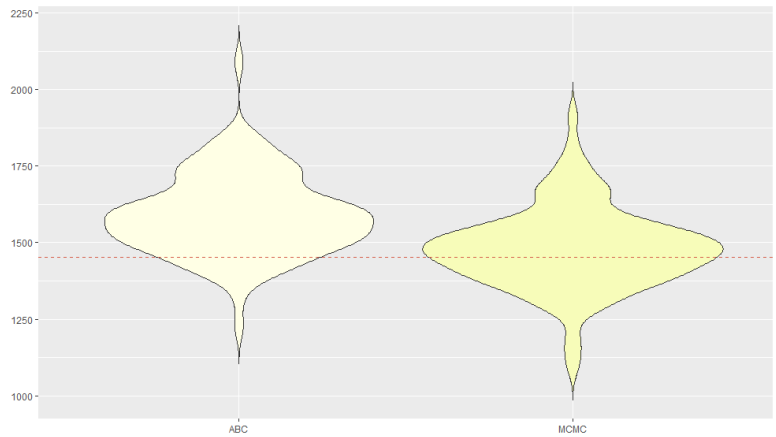


(b) M_2

(c) M_3



(d) M_4

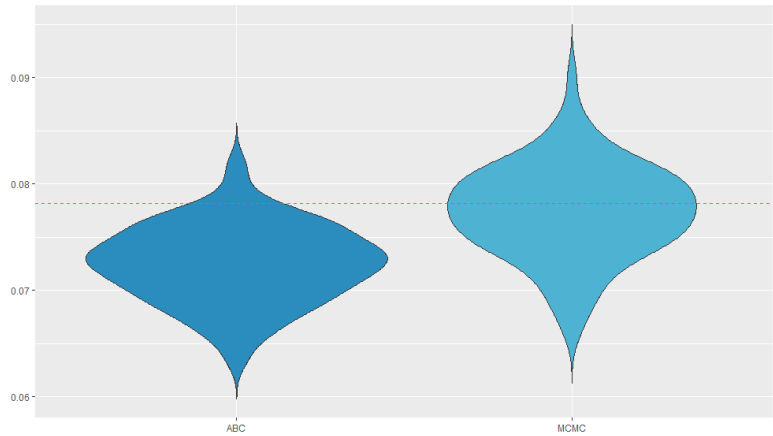


(e) M_5

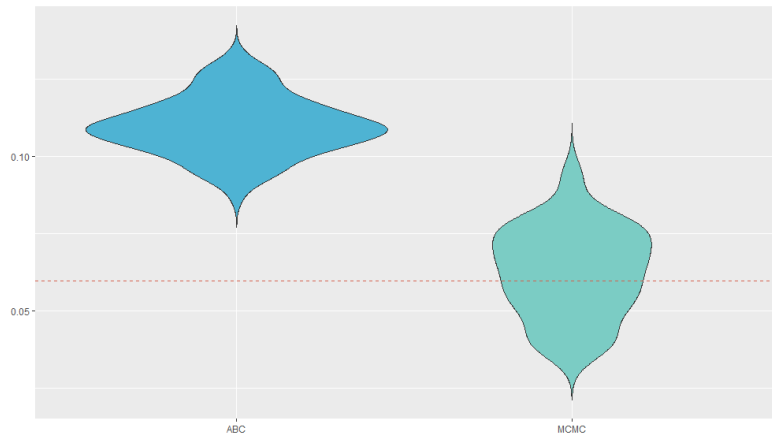
Figure 2.21: Posterior means of $\boldsymbol{M}$ simulated via MCMC (left) and ABC (right), 100 samples.

	p_1	p_2	p_3	p_4	p_5
MCMC	0.920	1.000	0.980	1.000	1.000
ABC	1.000	1.000	1.000	1.000	1.000

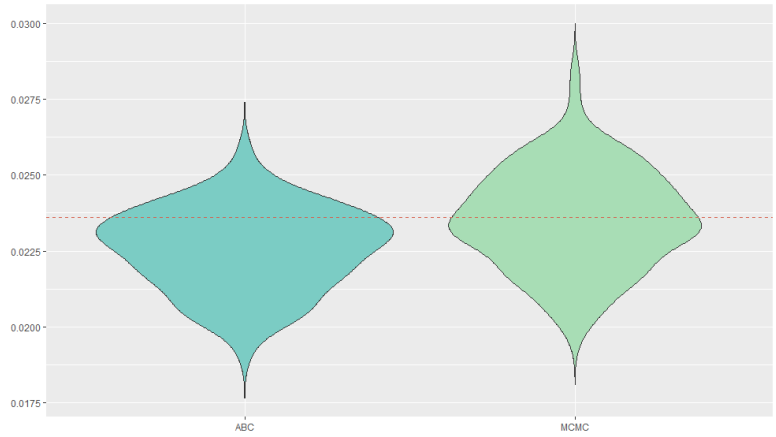
Table 2.10: The 95% Highest posterior density intervals for $\boldsymbol{p}$ include the true values, frequencies over 100 samples.



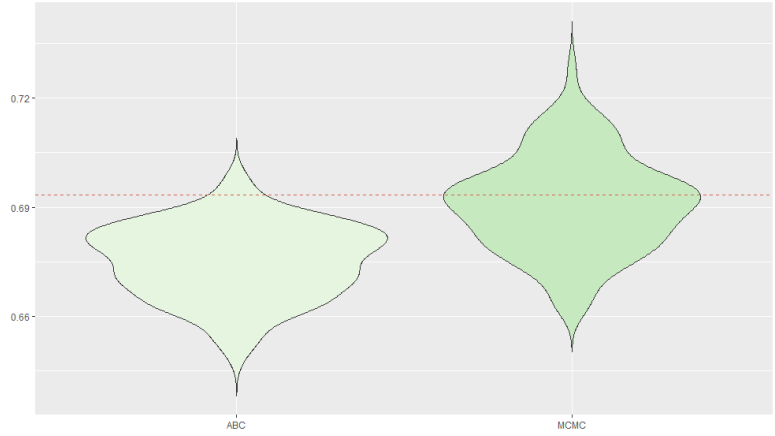
(a) p_1



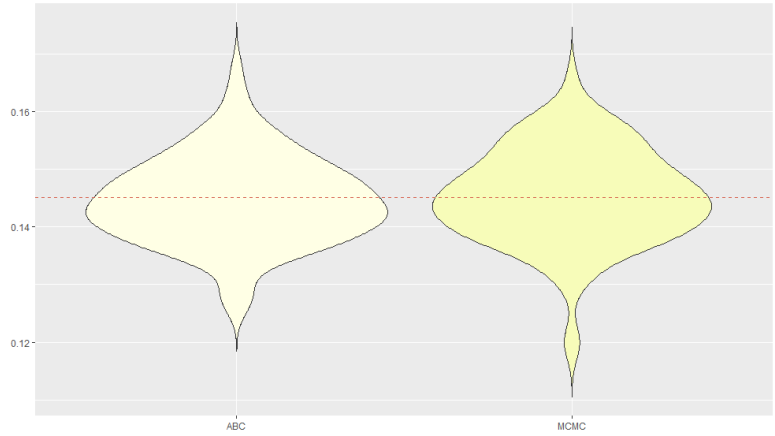
(b) p_2



(c) p_3



(d) p_4



(e) p_5

Figure 2.22: Posterior means of $\boldsymbol{p}$ simulated via MCMC (left) and ABC (right), 100 samples.

2.4 A case study: graduated job seekers in Italy

A critical task for the institutions is to guarantee an efficient transition into the labour market for the youngsters who concluded their education. Adopting appropriate policies favouring employment is crucial for many reasons, which concern the individual and the collective spheres. Indeed, it is instrumental in paying back the education, whose investment is private (made by the individuals/households themselves) and public (since the education system burdens the public expenditure). Furthermore, adequate employment policies support the pension system's intergenerational contract's stability and boost growth. A final (not less crucial) point concerns that human work contributes to society's development; employment is thus a primary objective for the welfare systems.
It is straightforward that implementing efficient policies requires knowledge about the entity of the phenomenon of unemployment. Eurostat[3] identifies *unemployed* persons as individuals aged 15 to 74 years who are not employed, currently available for work and actively seeking employment. This definition includes those previously employed and those who have never been employed and seek their first job. It is crucial to distinguish unemployed from *inactive* persons who are not employed or actively seeking work. The latter category also includes children, full-time students, pensioners and housewives (men).

In this section, we aim at estimating the size of a subpopulation of graduates, namely those who are still seeking a job after one year from graduation. Our final goal is to have a yearly estimate, exploiting the Interuniversity Consortium "Almalaurea" annual survey.

Every year the Italian National Statistics Institute provides data on the population size of that year's graduates[4]. Before the academic year 2012/2013, the collection of such data was survey-based; since then, it has come from the Student National Register of the Ministry of Education, University and Research (MIUR). Hence, the degree of accuracy is high.

[3]EU labour force survey - methodology
[4]available at I.stat

Variables	Value	Notes
Degree's classification	Laurea Triennale	the first level of the higher education system (corresponding to a Bachelor Degree)
	Laurea Magistrale	the second level of the higher education system (corresponding to the Master Degree)
	Laurea Magistrale a ciclo unico	a program that contemplates a 5/6 years course e.g. Law, Architecture, Medicine
	Laurea Vecchio Ordinamento	programs in effect before the Bologna process, 1999
Degree's achievement date	dd-mm-yy	
University that released the degree	The whole set of Italian Universities	
Date of the first job contract	dd-mm-yy	

Table 2.11: Variables in the available dataset

Thanks to the collaboration with Stefano De Santis[5], we can access a microdata sample collecting information about individuals who graduated in 2011. The dataset includes some variables about the achieved degree and the individuals' employment contract history from their first contract - that can date from before 2011 - to 2016. Hence, the available data allow detecting the number of graduates who were still unemployed one year after their graduation. Table 2.11 shows the variables of interest included in the dataset. We decide to exclude from the sample those who achieved a Bachelor degree (*Laurea Triennale*) since we are interested in the employment level of those who had completely concluded their education. Master's graduates may decide to continue their education; however, we assume such units' incidence to be negligible. We also exclude the units that were employed when they graduated. After the cleaning procedures, the sample contains $n_I = 3798$ individuals, $x_{I,u} = 1372$ of them who were still unemployed one year after graduation. The units who entered the labor market within a year are $x_{I,e} = n_I - x_{I,u} = 2426$. We can assume

$$\begin{aligned} X_{I,u} &\sim \text{Binom}(M_u, \zeta_{I,u}) \\ X_{I,e} &\sim \text{Binom}(M_e, \zeta_{I,e}) \end{aligned} \tag{2.23}$$

where M_u, M_e are the total number of people who graduated in 2011 and who were still unemployed or were employed one year after their graduation, respectively. Hence,

$$X_{I,u}|n_I \sim \text{FNCH}(M_u, N - M_u, n_I, w_I) \tag{2.24}$$

with $w_I = \dfrac{\zeta_{I,u}/(1-\zeta_{I,u})}{\zeta_{I,e}/(1-\zeta_{I,e})}$ being the relative weight of the unemployed persons in the Istat sample. In this case, the urn is not biased, i.e. we know that the sample has been randomly selected, and the probability of inclusion of the unemployed is the same as the employed ones. Therefore, $w_I = 1$ and FNCH is a hypergeometric distribution; nevertheless, we can still use the algorithms described in §2.1. Concerning N, although the high level of accuracy of the Istat estimate, we prefer considering the intrinsic uncertainty linked to any estimate; for this reason, we elicit a prior distribution for it, that is centred on the value estimated by the Istat, i.e. a Poisson($\lambda_N = 130067$). Finally, we consider M_u approximately Normal, centred on the value of M_u obtained via numerical approximation,

[5]Istat

σ_{M_u}	N	M_u
5000	130054(353.031)	47001.88(976.818)
1000	130065.2(371.175)	47025.09(1011.288)
15000	130088.6(365.825)	47022.19(972.446)

Table 2.12: Estimated posterior mean and standard deviation (in parenthesis) of N and M_1, for different values of $\sigma^2_{M_u}$. Data source: Istat.

assuming N fixed and equal to λ_N:

$$M_u \sim \mathrm{N}(\mu_{M_U} = 46998, \sigma^2_{M_u}) \ . \tag{2.25}$$

We tested the sensitivity of the results for different values of $\sigma_{M_u} = 5000$, $10000, 15000$; the results are robust (see Table 2.12).
Figure 2.23 shows the empirical posterior distributions of N and M_1 simulated via MCMC. As we expect, the posteriors are centred on the prior means. They give us a measure of uncertainty about the total number of people who graduated in 2011 (who were unemployed for at least one year). The results suggest that about 36% of the 2011 graduates had not found a job within the year. We can see such an estimate as an upper bound for M_u: indeed, it could include the inactive people, that include who continued their studies, and those who were employed without a regular contract.

Suppose we had information about the employment condition of a sample of recently graduated people for each year. In that case, we could estimate the time series of the number of those who have found it difficult to enter the labour market. Yet, our sample is part of the census survey that used to take place on a ten-year basis[6]. As an alternative, we consider exploiting the information collected every year by the Graduates' Employment Status Survey (GESS) by AlmaLaurea [7].
AlmaLaurea is an interuniversity consortium that yearly conducts surveys on a sample of people who graduated the year before. The GESS collects information on the employment condition of the respondents via CAWI (Computer-Assisted Web Interview) and CATI (Computer-Assisted Telephone Interview) methodologies. The data is integrated

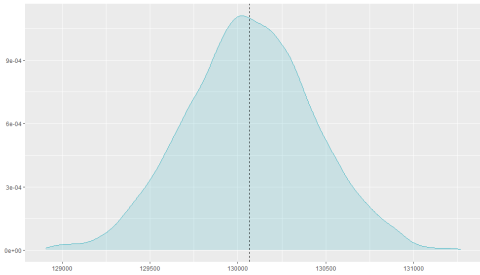
[6]As discussed in the Introduction, since 2018, Istat has started the "permanent census of the Population and Housing" (see Istat (2018)).
[7]AlmaLaurea - Consulta i dati

with the universities' administrative archives involved in the investigation (such as gender, date of birth, information on the course attended, etc.).
The response rate for the 2012 survey is 85%[8], for a total sample size of $n_A = 75443$. However, we expect the propensity to participate in the survey for purely statistical purposes to be different between those employed and those who have not found a job yet. Likely, the unemployed would be less enticed to fill a questionnaire about their employment condition.
The observed number of unemployed after one year from their graduation[9] is $x_{A,u} = 22853$, and the number of employed $n_A - x_{A,u} = 52590$. To estimate the unemployed's exposure in the Almalaurea survey, we exploit the estimates previously obtained using the Istat sample. We elicit prior distributions for N and M_u that are centred on the posterior means estimated in the "first step"; in this "second step", we aim at estimating w_A, i.e. the relative weight of the unemployed in the Almalaurea survey. We opt for a wide prior for w_A, i.e.

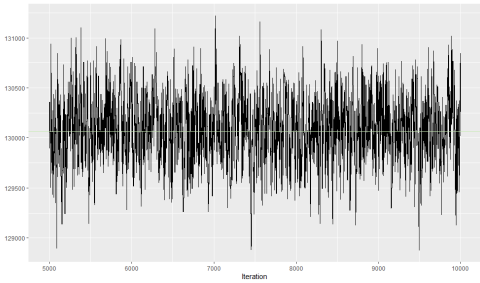
$$w_A \sim \text{Unif}\left(\frac{1}{10}, 10\right) \; ; \tag{2.26}$$

it means that we consider the possibility that the employed are up to ten times more exposed than the unemployed, and vice-versa. The posterior of w_A produced via MCMC (Algorithm 3) is shown in Figure 2.24 (a) and summarised in Table 2.13. From Figure 2.24 (b), it emerges a quite strong autocorrelation that is expected when all parameters are unknown; however, the range of values that w_A covers is very narrow. The results seem to confirm our initial guess; the estimated posterior mean of w_A suggests that the employed are about 1.8 times more exposed to the Almalaurea questionnaire than the unemployed.

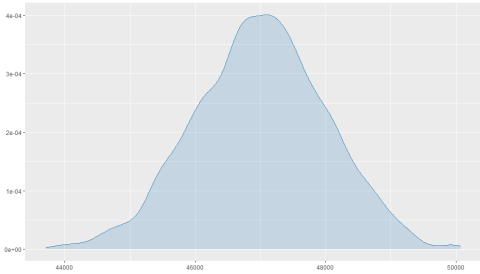
[8]Excluding those who achieved the Bachelor degree.
[9]precisely, who have never been employed for the year after their graduation.



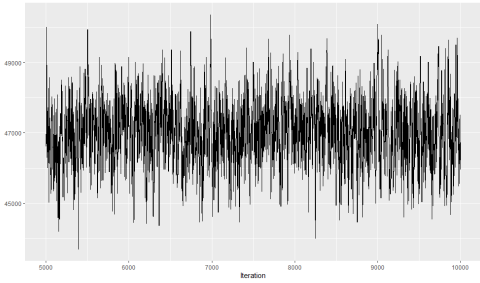
(a) Total graduated ($\sigma_{M_u} = 5000$)



(b) Total graduated ($\sigma_{M_u} = 5000$)

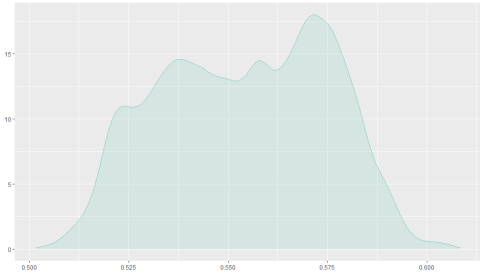


(c) Still unemployed

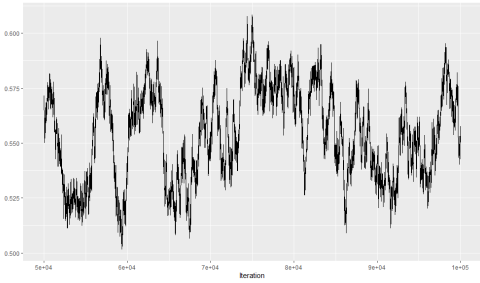


(d) Still unemployed

Figure 2.23: Posterior of N and $\boldsymbol{M_1}$. Data source: Istat



(a) Relative weight of the unemployed



(b) Relative weight of the unemployed

Figure 2.24: Posterior of w_A. Data source: Almalaurea

	Posterior mean	Posterior sd	95%HPDi
N	130062.2	341.066	$[129393; 130723]$
M_u	46871.22	459.734	$[46181; 47634]$
w_A	0.554	0.021	$[0.518; 0.590]$

Table 2.13: Estimates for the posterior mean, standard deviation and Highest Posterior Interval of the parameters of interest. Almalaurea, GESS 2012.

Assume that the propensity to fill the Almalaurea questionnaire does not significantly change once we control for the employment status. In this case, we can use the obtained results to make inference for the following years. Indeed, the 2011 cohort can be seen as a "training sample": it makes possible the yearly estimation of the number of recently graduated unemployed people avoiding further surveys or needing only targeted post enumeration surveys. As an example, we show the results for the 2012 cohort.

We use the Istat's estimate of the number of graduates in 2012[10] to elicit the prior for N^{2012}:

$$N^{2012} \sim \text{Pois}(\lambda_{N^{2012}} = 127161) \ . \tag{2.27}$$

For the exposure of the unemployed to the 2013 survey[11], i.e. w_A^{2012}, we exploit the information obtained in the second step of our analysis, and set

$$w_A^{2012} \sim \text{N}(\bar{w_A}, (10 \ \text{sd}_{w_A})^2) \tag{2.28}$$

where $\bar{w_A}$ is the estimated posterior mean of w_A and sd_{w_A} is the standard deviation. For M_u^{2012}, we opt again for a Normal distribution centred on the value obtained via numerical approximation assuming $N^{2012} = \lambda_{N^{2012}}$ and w_A^{2012} set equal to the posterior mean's estimate of w_A ($w_A = 0.554$, as in Table 2.13). The results are shown in Table 2.14.

The mean estimate for the number of 2012 graduates who were still unemployed after one year is 66733.7, about the 52.5% of the total graduates. This result appears dramatic and worthy of attention. We wonder if such an increase in unemployment among recent graduates is plausible.

[10] We still excludes those who achieved the Bachelor degree.
[11] referred to 2012 graduates

	Posterior mean	Posterior sd	95%HPDi
N^{2012}	127146.4	339.016	$[126443; 127784]$
M_u^{2012}	66733.68	1517.433	$[63914; 69290]$
w_A^{2012}	0.572	0.047	$[0.499; 0.664]$

Table 2.14: Estimates for the posterior mean, standard deviation and Highest Posterior Interval of the parameters of interest. Data source: Almalaurea, GESS 2013.

Indeed, once we considered the Italian economic situation in 2012, the number seems reliable. The last quarter of 2011 saw the beginning of the sovereign public debt crisis, which led to the then Prime Minister Silvio Berlusconi's resignation and the installation of a technical government headed by Mario Monti. During 2012 the economic activity slowed down drastically, and the unemployment rate increased. At the end of 2012, the Italian GDP was still 8% points lower than five years before, and the level of investment more than 20% lower (see, for example, Busetti and Cova (2013)).
Therefore, the two estimated proportions of still unemployed graduates likely arise in the context of that time, characterised by increasing unemployment.

2.5 Discussion

In this chapter, we addressed the estimation of the size of a heterogeneous population when a single list is available, or we have multiple lists, but we lack unique identifiers. We presented a model relying on the underemployed Fisher's noncentral hypergeometric distribution and faced the issues arising from the computational burden of its probability mass function, especially in the multivariate context.
Indeed, in §2.2 we presented two methods for estimating the population size in the presence of multiple subgroups. On the one hand, the "ABC method" bypasses evaluating the likelihood; it results being a valid solution even though not computationally efficient. Moreover, it is very approximate, as its name suggests. On the other hand, the "MCMC method" is exceptionally performing in the parameters' estimation. Nevertheless, even though we lightened the computation by sampling one

subgroup's size at the time, it becomes onerous as N increases.

The simulation studies performed in §2.1 for the single list case show how the parameters' posterior strongly reflects prior uncertainty; this is typical of models with little information. Indeed, the estimated posterior intervals shrink when more lists are considered and more information is injected into the model. The Bayesian approach allows us to formalise such uncertainty and give credible intervals for the population size when other methods appear unfeasible. The case study we presented in §2.4 gives the motivation to the whole work. Nowadays, data integration is a crucial task; an advantage of our model is the ability to extract information from one (or a few) data sources and integrate it in a multisource environment.

Part II

Capture-recapture in the presence of overcoverage

Chapter 3

Multisource population size estimation in the presence of out-of-scope units: an overview

This chapter aims to review the influential and most recent literature about the problem of population size estimation via multiple lists and can be seen as preparatory to chapter 4. §3.1 introduces the basic notation we will use throughout this chapter and the next. §3.2 presents the most recent literature about capture-recapture models, which follows Fienberg's line, from its milestone Fienberg (1972) to the Bayesian version of log-linear models. In §3.3, we will present an example of interest. §3.4 extends the models previously introduced to the overcoverage issue, and §3.5 compares them with two examples simulating different scenarios. The conclusions follow.

3.1 Notation

We are interested in estimating the unknown size N of a closed population U. Assume K lists partially enumerating U are available. Let $\delta_{ik} = 1$ if unit i, $i \in \mathbb{Z}$, is enumerated in the list k, $k = 1, ..., K$, and 0 otherwise. Define

$$\boldsymbol{\delta}_i = (\delta_{i1}, ..., \delta_{iK}) \tag{3.1}$$

	$\delta_{i1} = 1$		$\delta_{i1} = 0$	
	$\delta_{i2} = 1$	$\delta_{i2} = 0$	$\delta_{i2} = 1$	$\delta_{i2} = 0$
$\delta_{i3} = 1$	x_{123}	x_{13}	x_{23}	x_3
$\delta_{i3} = 0$	x_{12}	x_1	x_2	?

Table 3.1: Three lists observed contingency table

i.e. $\boldsymbol{\delta}_i$ is the *capture vector* (or *capture history*) of unit i. If $\boldsymbol{\delta}_i = \mathbf{0}$, it means that the unit i has never been captured.
Let ω_i be a position index indicating which $\delta_{ik} = 1$ in $\boldsymbol{\delta}_i$. For instance, assume $K = 3$; if $\boldsymbol{\delta}_i = (1, 0, 1)$, ω_i will be equal to the set $\{13\}$. Let x_ω be the number of individuals whose capture vectors are summarised by the same ω_i; e.g., x_1 is the number of units captured only by list 1 or, equivalently, whose $\delta_{ik} = 1$ for $k = 1$ and 0 otherwise. The counts can be summarised in an incomplete contingency table, as shown in Table 3.1.

We might need to refer to the number of individuals captured *at least* by the first list, i.e. those individuals whose $\delta_{ik} = 1$ for $k = 1$ and either 0 or 1 for $k = 2, 3$; we indicate it with x_{1+}, $\{1+\}$ being a shortcut to indicate $\{1\} \cup \{12\} \cup \{13\} \cup \{123\}$. Therefore, we denote with $x_{\omega+}$ the *marginal* count and with x_ω the *cross-classified* one (as in Zhang (2019)).
Now assume that any list k may also enumerate some units that do not belong to the target population U. Let us order the lists such that the first K' do not enumerate *out-of-scope* units. Hence, let $A = \{1, 2, ..., K'\}$ be the lists' set only enumerating units belonging to U and $B = \{K' + 1, K' + 2..., K\}$ be the one also including some $i \notin U$. It follows that the observed number of units captured by some $k \in B$ is equal to $x_\omega = y_\omega + r_\omega$, where y_ω represents the number of *in-scope* units and r_ω the number of the out-of-scope ones. Both y_ω and r_ω are unobserved. If ω indicates at least one $\delta_k = 1, k \in A$, then $r_\omega = 0$ and $x_\omega = y_\omega$.
We will denote the cross-classified *error rate*, i.e. the probability of being erroneously enumerated given that one belongs to the cell indexed by ω, with ξ_ω; similarly, $\tau_\omega = 1 - \xi_\omega$ will be the cross-classified *hit rate*.

Finally, let us indicate with $y_{\mathbf{0}}$ the unknown number of in-scope units not captured by any list; thus, the unknown size of the population U

		$\delta_{A_1}(i)=1$		$\delta_{A_1}(i)=0$	
		$\delta_{A_2}(i)=1$	$\delta_{A_2}(i)=0$	$\delta_{A_2}(i)=1$	$\delta_{A_2}(i)=0$
$\delta_{A_3}(i)=1$	$\delta_U(i)=1$	y_{123}	y_{13}	y_{23}	y_3
	$\delta_U(i)=0$	r_{123}	r_{13}	r_{23}	r_3
$\delta_{A_3}(i)=0$	$\delta_U(i)=1$	y_{12}	y_1	y_2	y_0
	$\delta_U(i)=0$	r_{12}	r_1	r_2	?

Table 3.2: Latent structure of three lists contingency table where all sources are affected by overcoverage

will be $N = \sum_{\omega} y_\omega$. Table 3.2 helps clarifying the notation; it shows a three-way incomplete contingency table where all sources are affected by overcoverage.

3.2 Capture-recapture

Capture-recapture models were born in ecology at the end of the 19th century when the need for accurate tools able to estimate the number of specimens belonging to a target animal population became stronger. Indeed, C. G. J. Petersen[1] and F. C. Lincoln[2] are the ones who can be considered the "fathers" of capture-recapture methods; a marine biologist and an ornithologist, respectively. Starting from the popular Lincoln-Petersen estimator, capture-recapture models have evolved and found one of their primary applications in social sciences. In the following paragraph, we briefly review the most recent models suitable for official statistics' needs.

[1]see Petersen (1985)
[2]see Lincoln (1930)

3.2.1 Log-linear models' setup

Log-linear models have been the most popular representation for count data so far. The way these models work is self-explanatory; considering the observed counts as random variables' realisations, we may express the natural logarithm of their expected values as a linear function of a set of unknown parameters.
Assume we can classify y units belonging to a particular population in a contingency table according to K characterising factors. Also, assume that each factor has different levels, $l_1 = 1, ..., c_1, l_2 = 1, ..., c_2$ up to $l_K = 1, ..., c_K$ respectively.
Let $Y_{l_1 l_2 ... l_K}$ be the random variable indicating the number of counts for the cell corresponding to level $l_1 l_2 \dots l_K$. In case of independence among factors, i.e. $P(\delta_{i12...K} = 1) = P(\delta_{i1+} = 1)P(\delta_{i2+} = 1) \cdot ... \cdot P(\delta_{1K+} = 1)$, its expected value is

$$\lambda_{l_1 l_2 ... l_K} = \frac{y_{l_1+}}{y} \frac{y_{l_2+}}{y} \cdot ... \cdot \frac{y_{l_K+}}{y} y \tag{3.2}$$

where $\lambda_{l_1 l_2 ... l_K} = \mathbb{E}(Y_{l_1 l_2 ... l_K})$ and $y = \sum_{l_1} \sum_{l_2} ... \sum_{l_K} y_{l_1 l_2 ... l_K}$ is the total number of counts.
Analogously to the analysis of the variance, Fienberg (1970) expresses the natural logarithm of such expected value as

$$\log(\lambda_{l_1 l_2 ... l_K}) = \phi + \beta_{l_1} + \beta_{l_2} + ... + \beta_{l_K} \tag{3.3}$$

where the β's represent deviations from the grand mean, i.e. ϕ. However, if the factors' independence assumption does not hold, we need to introduce additional parameters, i.e. the interaction terms. In the case of three factors, the so-called *saturated model* will include three two-factor interaction terms and one three-factor:

$$\log(\lambda_{l_1 l_2 l_3}) = \phi + \beta_{l_1} + \beta_{l_2} + \beta_{l_3} + \beta_{l_1 l_2} + \beta_{l_1 l_3} + \beta_{l_2 l_3} + \beta_{l_1 l_2 l_3} \tag{3.4}$$

Any model which does not include all the interaction terms is said to be *unsaturated*. Generalizing,

$$\log(\lambda_{l_1 ... l_K}) = \phi + \sum_k \beta_{l_k} + \sum_k \sum_{j>k} \beta_{l_k l_j} + ... + \beta_{l_1 ... l_K} \ . \tag{3.5}$$

The overparameterization of the model emerges clearly, asking for a constraint which will allow for the model's identification. One possibility is

the *sum-to-zero* constraint:

$$\sum_{l_k=1}^{c_k} \beta_{l_k} = \sum_{l_1=1}^{c_1} \beta_{l_1 l_k} = ... = 0 \quad \forall k \ . \tag{3.6}$$

Another option is the *corner point* constraint, which we will use throughout this work:

$$\beta_{(l_k=1)} = 0, ..., \beta_{(l_k=1)...l_j} = 0 \quad \forall k \neq j \ . \tag{3.7}$$

Fienberg (1972) applied log-linear models to capture-recapture data for the first time. Here, the K factors are the capture occasions, or lists. For each list, only two levels l_k are possible: captured ($l_k = 1$) or missed ($l_k = 0$). As a result, the observed units can be classified in an incomplete contingency table of dimension 2^K. By incomplete we mean that it will presents a missing cell, the one corresponding to $\{l_1 = 0 \ ... \ l_K = 0\}$, by construction. In this framework, three assumptions play a crucial role. First, the population is assumed to be closed (Fienberg (1972)). Second, the probability of being captured in one or more lists is the same for any individual i belonging to the target population; in other words, there is *capture homogeneity*. Another crucial assumption is that the units have unique labelling: one can infer the entire multiple recapture history of any observed individual from its label anytime. In the conclusions, we will see how literature has addressed the deviations from the last two assumptions.
To our knowledge, the entire capture-recapture literature on log-linear models has only focused on the hierarchical ones, i.e. those models where the higher-order relatives of a zero term are constrained to be zero as well (see Fienberg (1972)). For instance, assume $K = 3$; the saturated model results being

$$\begin{aligned}\log(\lambda_{l_1=1 \ l_2=1 \ l_3=1}) = \phi + \beta_{l_1=1} + \beta_{l_2=1} + \beta_{l_3=1} + \beta_{l_1=l_2=1} + \beta_{l_1=l_3=1} + \\ + \beta_{l_2=l_3=1} + \beta_{l_1=l_2=l_3=1}\end{aligned} \tag{3.8}$$

Since the levels for each list are only 0 and 1, with a little abuse of notation, we replace the subscript l_k with its index k when $l_k = 1$, and we omit it when $l_k = 0$. The result is in line with the notation described in §3.1. Therefore, we can rewrite the equation above as

$$\log(\lambda_{123}) = \phi + \beta_1 + \beta_2 + \beta_3 + \beta_{12} + \beta_{13} + \beta_{23} + \beta_{123} \tag{3.9}$$

Model specification	Highest order interactions
$\phi + \beta_1 + \beta_2 + \beta_3 + \beta_{12} + \beta_{13} + \beta_{23} + \beta_{123}$	[123]
$\phi + \beta_1 + \beta_2 + \beta_3 + \beta_{12} + \beta_{13} + \beta_{23}$	[12][13][23]
$\phi + \beta_1 + \beta_2 + \beta_3 + \beta_{12} + \beta_{13}$	[12][13]
$\phi + \beta_1 + \beta_2 + \beta_3 + \beta_{12} + \beta_{23}$	[12][23]
$\phi + \beta_1 + \beta_2 + \beta_3 + \beta_{13} + \beta_{23}$	[13][23]
$\phi + \beta_1 + \beta_2 + \beta_3 + \beta_{12}$	[12]
$\phi + \beta_1 + \beta_2 + \beta_3 + \beta_{13}$	[13]
$\phi + \beta_1 + \beta_2 + \beta_3 + \beta_{23}$	[23]
$\phi + \beta_1 + \beta_2 + \beta_3$	[1][2][3]

Table 3.3: Different model specifications for $\log(\lambda_{123})$

Generalising,

$$\log(\lambda_\omega) = \phi + \sum_{\nu \in \Omega(\omega)} \beta_\nu \tag{3.10}$$

where $\Omega(\omega)$ is the set of all non-empty subsets of ω; equivalently,

$$\log(\lambda_\omega) = \phi + \boldsymbol{d}_\omega{}^T \boldsymbol{\beta} \tag{3.11}$$

where $\phi \in \mathbb{R}$ is the grand mean and $\boldsymbol{d}_\omega$ is the design vector that indicates which elements of the regression parameters vector $\boldsymbol{\beta}$ apply to the cell indexed by ω. $\boldsymbol{\beta}$ is the vector of the factors' effects and interaction terms:

$$\boldsymbol{\beta} = (\beta_1, ..., \beta_K, ..., \beta_{k_1 k_2}, ..., \beta_{k_1 k_2 k_3}, ..., \beta_{1...K})^T \tag{3.12}$$

Table 3.3 shows all possible model specifications for the number of counts y_{123} when $K = 3$.

Fienberg's approach is to estimate the most parsimonious log-linear model, restricted to the incomplete table, and use it to predict the count of the missing cell. To give an insight into the effectiveness of such an estimation procedure, we briefly describe the main steps; see Fienberg (1972) for the details.

Let $\{y_\omega\}$ be Multinomial with parameters N, p_ω, where p_ω is function of some parameters $\boldsymbol{\zeta}$, i.e. $p_\omega = p_\omega(\boldsymbol{\zeta})$, and let $L(N; \boldsymbol{\zeta})$ be the relative likelihood function:

$$L(N; \boldsymbol{\zeta}) = \frac{N!}{y_0!}(1 - \sum_\omega p_\omega(\boldsymbol{\zeta}))^{y_0} \prod_\omega \frac{p_\omega(\boldsymbol{\zeta})^{y_\omega}}{y_\omega!} \tag{3.13}$$

As shown in Sanathanan (1972), the likelihood can be factorised and expressed as

$$L(N;\boldsymbol{\zeta}) = L_1(N;\sum_{\omega} p_\omega(\boldsymbol{\zeta}))L_2(\boldsymbol{\zeta}) \ , \tag{3.14}$$

where

$$L_1(N;\sum_{\omega} p_\omega(\boldsymbol{\zeta})) = \frac{N!}{(\sum_{\omega} y_\omega)!y_0!}(1-\sum_{\omega} p_\omega(\boldsymbol{\zeta}))^{y_0}(\sum_{\omega} p_\omega(\boldsymbol{\zeta}))^{\sum_{\omega} y_\omega} \tag{3.15}$$

and

$$L_2(\boldsymbol{\zeta}) = (\sum_{\omega} y_\omega)!\prod_{\nu}\frac{p_\nu(\boldsymbol{\zeta})^{y_\nu}}{y_\nu!(\sum_{\omega} p_\omega(\boldsymbol{\zeta}))^{y_\nu}} \tag{3.16}$$

Maximizing $L(N;\zeta)$, we obtain the unrestricted estimates of N, $\hat{N}_U$, and ζ, $\hat{\zeta}$. Yet, maximizing $L_1(N,\sum_{\omega} p_\omega(\hat{\zeta}_C))$ where $\hat{\zeta}_C$ is the MLE of L_2, we obtain a conditional estimate of N, $\hat{N}_C$. Sanathanan (1972) proves that $(\hat{N}_U,\hat{\zeta}_U)$ and $(\hat{N}_C,\hat{\zeta}_C)$ are both consistent estimators; hence, Fienberg (1972) suggests to use $(\hat{N}_C,\hat{\zeta}_C)$ to assess the appropriateness of a given model.
The maximum likelihood estimator for N_C is

$$\hat{N}_C = \left[\frac{\sum_{\omega} y_\omega}{\sum_{\omega} p_\omega}\right] \ , \tag{3.17}$$

where [.] stands for the the closest integer. After some algebraic manipulation, we get

$$\hat{N}_C = \sum_{\omega} y_\omega + \hat{\lambda}_0 \tag{3.18}$$

where, for any K,

$$\hat{\lambda}_0 = \frac{\hat{\Lambda}_{odd}}{\hat{\Lambda}_{even}} \ , \tag{3.19}$$

$\hat{\Lambda}_{odd}$ ($\hat{\Lambda}_{even}$) being the product of all $\hat{\lambda}_\omega$ whose ω has an odd (even) number of elements. $\hat{\lambda}_\omega$'s are the MLE's obtained from the incomplete contingency table, given by setting the expected values of the marginal totals corresponding to the highest order interaction terms in the model equal to their observed value (see Fienberg (1972)). It is possible to compute a confidence interval relying either on the asymptotic normality

of the estimator $\hat{N}$ Bishop et al. (1975), or the profile likelihood of $\hat{N}$ (see Cormack (1992)). We can assess the appropriateness of a given model using either Pearson's Chi-squared or the likelihood ratio statistics. Indeed, it is feasible to use any information criterion, such as the Akaike or the Bayesian, as well. However, notice that for small K and a large number of model's parameters, the available degrees of freedom may be very few (0 in case of $K = 2$ and independence model). Moreover, as proved by Regal and Hook (1991), more than one specification can fit the data perfectly, even with the same number of parameters, giving very different confidence intervals for the population size. Zhang (2019) proposes a model selection criterion based on a so-called *latent likelihood ratio* that may help to select a model in cases of zero degrees of freedom. Another limitation of Fienberg's approach is the difficulty of including any information on the population's size that may be available a priori.

3.2.2 Decomposable graphical models

Intending to overcome the limitations mentioned above of Fienberg (1972), Madigan and York (1997) presented the Bayesian approach to population size estimation problem, which has deeply influenced the following literature. This approach is hierarchical log-linear models based, but it focuses only on a subset of such models, namely the so-called *decomposable graphical models*. A statistical model is said to be *graphical* if it embodies a set of conditional independence relationships, which can be summarised by a graph. For the sake of clarity, we briefly introduce the basic terminology of graph theory, mainly relying on Madigan and York (1995).

Define a graph as a pair $G = (V, E)$, with V being a finite set of vertices and E being the set of edges, i.e. a subset of $V \times V$ ordered pairs of distinct vertices. In practice, the vertices represent the model's variables and the edges the dependence relations among them. A graphical model consists of a statistical model describing the conditional independence relationships among variables via a graph. Two variables may be just correlated, or there might be a causal relation between them. In the former case, both $(V_j, V_{j'}), (V_{j'}, V_j) \in E, \forall\ j, j'$ and we represent such edge as a straight line; V_j and $V_{j'}$ are said to be *neighbours*, and the resulting is a graph so-called *undirected*. Instead, if $(V_j, V_{j'}) \in E$ but $(V_{j'}, V_j) \notin E$, the edge is a directed arrow, and the graph is *directed*. When an edge

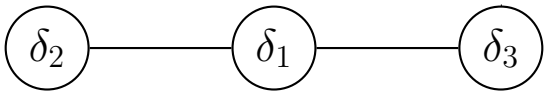


Figure 3.1

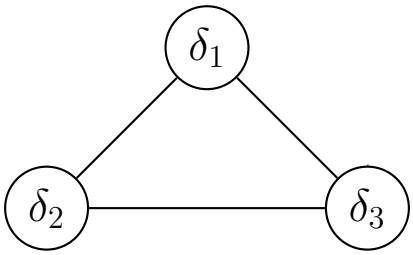


Figure 3.2

joins all pairs of vertices, the graph is *complete*. A complete subset of the vertex set not contained in any other complete subset is a maximal *clique* C. In an undirected graph, $G = \cup_{h=1}^{H} C_h$, $(C_1, ..., C_H)$ is a *perfect ordering* of the cliques when the vertices of each clique C_h also contained in previous cliques are all members of one previous clique only; the sets $S_h = C_h \cap (\cup_g^{h-1} C_g)$ are called *clique separators*.
Figure 3.1 shows an undirected graph composed of two cliques, $C_1 = \{\delta_2, \delta_1\}$ and $C_2 = \{\delta_1, \delta_3\}$, whose ordering is perfect and where δ_1 is a separator. Now focus on undirected graphs. Define a *path* as a sequence $V_0, ..., V_n$ of distinct vertices such that (V_i, V_{i-1}) are neighbours for all $i = 1, ..., n$. If V_0 and V_n coincide, that path is said to be a *cycle*. An undirected graph is chordal when it contains no cycles of four or more vertices without a chord, i.e. two non-consecutive vertices that are neighbours. Only if a graph is chordal, it admits a perfect ordering of its cliques. An undirected chordal graph represents a decomposable model.

Let us go back to log-linear models. Let δ_k be the variable indicating a unit's presence or absence in each of the K lists. Allowing for all the pairwise interactions $[12]...[(K-1)K]$ the resulting graph would be a cycle, and there is no way to exclude the Kth-order interaction $[1...K]$. Figure 3.2 shows this concept in the case $K = 3$. It results that decomposable graphical models are only a subset of the log-linear models. Although restrictive, if we can represent a log-linear model as a decomposable graph, its analysis results much more tractable from a computa-

tional point of view. Indeed, Dawid and Lauritzen (1993) show that if a model is decomposable, the joint distribution can be factorised into a product of conditional distributions. Following the notation in Di Cecco (2019):

$$p_G = \prod_{h=1}^{H} p_{C_h} \left(\prod_{g=2}^{H} p_{S_g} \right)^{-1} = p_{C_1} \prod_{h=2}^{H} \frac{p_{C_h}}{p_{S_h}} = p_{C_1} \prod_{h=2}^{H} p_{C_h|S_h} \tag{3.20}$$

where p_{C_h} (p_{S_g}) is the marginal distribution over the variables included in the hth clique (g^{th} separator), and $p_{C_h|S_h}$ is the conditional distribution of the hth clique given the relative separator. This way, solving both the maximisation and the integration tasks in closed form becomes viable; see Dawid and Lauritzen (1993). Following the Bayesian approach, it is possible to set a prior distribution on cells probabilities conjugate with multinomial sampling, as proved by Dawid and Lauritzen (1993). Notably, a Dirichlet marginal distribution on the probability of each clique C_h, ρ_{C_h} must be placed following the perfect ordering of the cliques. Such prior distribution is the so-called *hyper-Dirichlet*. To account for model uncertainty, Madigan and York (1997) suggests to average all posterior distributions of N conditional on different models m weighted by their posterior model probabilities to obtain an unconditional posterior distribution. On model averaging for graphical models, see Madigan and Raftery (1994). An illustration of the model in Madigan and York (1997) follows.

Indicate with M the class of possible models for the cell probabilities of the contingency table, indexed by $\mathcal{M} = \{1, 2, ..., s\}$. Define $\boldsymbol{p}(m)$ as being the vector of cell probabilities for each model $m \in \mathcal{M}$.
Let $y|N, \beta, \mathcal{M} = m$ be Multinomial with parameters $(N, \boldsymbol{p}(m))$, where

- the prior on N might be
  - $\pi(N) \propto \frac{1}{N}$, i.e. Jeffreys prior;
  - $\pi(N) \propto 2^{-\log^*(N)}$, with $\log^*(N)$ is the sum of the positive terms in $\{\log_2(N), \log_2\{\log_2(N)\}, ...\}$, i.e. the Rissanen's prior;
  - $N \sim \text{Poisson}(\lambda)$ and $\lambda \sim \text{Gamma}(a, b)$ if needed;
- the prior on M is Uniform, and

- $\boldsymbol{p}(m) \sim$ hyper-Dirichlet, as stated before, with $p_C(m) \sim \text{Dirichlet}(\rho)$ for each clique. As non-informative choice, ρ is set equal to 1 or $\frac{1}{2}$.

The **R** package **dga** (Johndrow et al. (2015)) implements the model. It is easy to compute all posterior distributions of N conditional on different model M when K is low within the decomposable models. However, as K increases, the number of parameters grows exponentially; thus, the calculation becomes cumbersome and the model averaging impracticable. Madigan and York (1995) suggest using Markov Chain Monte Carlo Model Composition to approximate the average of the posterior distributions under each model. On the other hand, Madigan and Raftery (1994) propose to average over a small set of models, facilitating the communication of model uncertainty (Madigan et al. (1994)). Green (1995) suggests an entirely different approach that introduces the *reversible jump MCMC*. The RJMCMC is a sampler able to move ("jump") across different dimensions' parameters spaces, thus exploring different dimensions' models in a single chain. Dellaportas and Forster (1999) and King and Brooks (2001) apply the RJMCMC to the log-linear models, going beyond the class of decomposable graphical ones.
In the next paragraph, we will review Bayesian hierarchical log-linear models for capture-recapture and give an insight into the RJMCMC.

3.2.3 Bayesian log-linear models and the Reversible Jump sampler

Dealing with the decomposable class of graphical log-linear models from a Bayesian perspective requires a prior specification for the cell probabilities. However, specifying a prior for the model parameters implies the possibility of going beyond the decomposable model and taking into account the broader class of log-linear ones.
Dellaportas and Forster (1999) and King and Brooks (2001) are the first references for a detailed specification of a fully Bayesian log-linear model. Nowadays, Overstall and King (2014b) is a popular approach; mainly based on Forster (2010), it is the original theoretical support of the **R** package **conting**[3] (see also Overstall and King (2014a)). Notice that these works mainly deal with general log-linear models; they can be easily adapted to capture-recapture problems though. Below we outline a

[3]available from the CRAN at https://cran.r-project.org/web/packages/conting/index.html

model common to the most recent literature on Bayesian log-linear models but plugging it in the particular framework of incomplete contingency tables. We mainly refer to Overstall and King (2014a), dwelling on the differences when not negligible.
Following the notation of the previous sections, let y_ω be the number of population units count in cell ω. Moreover,

$$y_\omega|\phi, \boldsymbol{\beta}, m \sim \text{Poisson}(\lambda_\omega) \tag{3.21}$$

$$\log(\lambda_\omega) = \phi + \boldsymbol{d}_\omega^T \boldsymbol{\beta} \ . \tag{3.22}$$

Then, introducing the model indicator m as in §3.2.2 we obtain

$$\log(\lambda_\omega) = \phi + \boldsymbol{d}_{m,\omega}^T \boldsymbol{\beta}_m \ . \tag{3.23}$$

In compact form,

$$\log(\boldsymbol{\lambda}) = (\mathbb{1}_{2^K}, \boldsymbol{d}_m)\boldsymbol{\theta}_m \tag{3.24}$$

where $\boldsymbol{d}_m$ is a the matrix whose rows are given by $\boldsymbol{d}_{m,\omega}$, and $\boldsymbol{\theta}_m = (\phi, \boldsymbol{\beta}_m)^T$. As an alternative model for the data counts, we can also consider

$$\boldsymbol{y}|N, \boldsymbol{\beta}, m \sim \text{Multinomial}(N, \boldsymbol{p}) \tag{3.25}$$

with $\boldsymbol{p}$ being the vector of p_ω's, and $p_\omega = \frac{\lambda_\omega}{\sum_\omega \lambda_\omega}$. Whatever the model specification, we need to specify the joint prior

$$\pi(\phi, \boldsymbol{\beta}_m, m) = \pi(\phi, \boldsymbol{\beta}_m|m)\pi(m) \tag{3.26}$$

For the first factor, Overstall and King (2014a) follows Sabanés Bové and Held (2011) using the hyper-g prior, i.e. decomposing the joint prior on ϕ and $\boldsymbol{\beta}_m$ as

$$\pi(\phi, \boldsymbol{\beta}_m|m) = \pi(\phi)\pi(\boldsymbol{\beta}_m|m) \tag{3.27}$$

with $\pi(\phi) \propto 1$, and

$$\boldsymbol{\beta}_m|\sigma^2, m \sim N(\boldsymbol{0}, \boldsymbol{S}) \tag{3.28}$$

$$\boldsymbol{S} = \sigma^2 n(\boldsymbol{d}_m^T \boldsymbol{d}_m)^{-1} \tag{3.29}$$

$$\sigma^2 \sim \text{Inverse Gamma}\left(\frac{a}{2}, \frac{b}{2}\right) \tag{3.30}$$

Additionally, a Uniform prior is set over the model space, i.e.

$$\pi(m) = \frac{1}{|\mathcal{M}|} = \frac{1}{s} \tag{3.31}$$

Overstall et al. (2014) proves that under the prior specifications depicted above and choosing the Jeffreys prior for N, the joint posterior for $\boldsymbol{\beta}_m$, m and y_0 are identical under Poisson or Multinomial models.
We report the presence of sensible differences to the other references exclusively concerning the specification of the variance-covariance matrix Σ.
We refer the reader to Dellaportas and Forster (1999) and King and Brooks (2001) for more details.
The joint posterior distribution results being

$$\pi(y_0, \phi, \boldsymbol{\beta}_m, m|\boldsymbol{y}) \propto \pi(\boldsymbol{y}, y_0|\phi, \boldsymbol{\beta}_m, m)\pi(\boldsymbol{\beta}_m|\sigma^2, m)\pi(\sigma^2)\pi(m) \tag{3.32}$$

Updating y_0 and σ^2 from their full conditional distributions is straightforward; yet, to simulate from the full conditional distribution of the other parameters, Overstall and King (2014a) implement an RJMCMC algorithm.

Let m be the current model at iteration t; denote $\boldsymbol{\theta}_{\boldsymbol{m}}^{(\boldsymbol{t})}$ the current parameter vector.

1. Propose a move from model m to model $m' \in \mathcal{M}$ with probability $\pi_{m,m'}$. Typical jumps are those to models with parameters' dimension close to that of model m; in the case of log-linear models, this is equivalent to limit the moves to models with one more or one less interaction term. A move to model m itself is also allowed. If $m' = m$, the algorithm turns up being a Metropolis-Hastings, otherwise step 2 follows;

2. generate a vector of *innovation variables* $\boldsymbol{u}_{m,m'}$ from a proposal distribution $q_{m,m'}(u)$;

3. apply a mapping function T to $(\boldsymbol{\theta}_{\boldsymbol{m}}, \boldsymbol{u}_{m,m'})$ to obtain $\boldsymbol{\theta}_{m'}$;

4. set $\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\theta}_{m'} & \text{with probability } \gamma \\ \boldsymbol{\theta}_m & \text{with probability } 1-\gamma \end{cases}$
   where
   $$\gamma = \min\left\{1; \frac{\pi(\boldsymbol{y}, y_0|\boldsymbol{\theta}_{m'}, m')\pi_{m',m}q_{m',m}(u_{m',m})}{\pi(\boldsymbol{y}, y_0|\boldsymbol{\theta}_m^{(t)}, m)\pi_{m,m'}q_{m,m'}(u_{m,m'})}\left|\frac{\partial T(\boldsymbol{\theta}_m^{(t)}, u_{m,m'})}{\partial(\boldsymbol{\theta}_m^{(t)}, u_{m,m'})}\right|\right\} \tag{3.33}$$

		$\delta_{i1} = 1$		$\delta_{i1} = 0$	
		$\delta_{i2} = 1$	$\delta_{i2} = 0$	$\delta_{i2} = 1$	$\delta_{i2} = 0$
$\delta_{i3} = 1$	$\delta_{i4} = 1$	27	32	42	123
	$\delta_{i4} = 0$	18	31	106	306
$\delta_{i3} = 1$	$\delta_{i4} = 1$	181	217	228	936
	$\delta_{i4} = 0$	177	845	1131	?

Table 3.4: Number of casualties during the conflict in Kosovo, March-June 1999. Ball et al. 2002

See Green (1995) or Robert and Casella (2004) for more details. The RJMCMC can adapt to the restricted class of decomposable models; see King and Brooks (2001).

Once simulated from the posterior distribution, it is possible to assess model adequacy via information criteria or the computation of the Bayesian p-value (see Gelman et al. (2004)), as in Overstall et al. (2014).

3.3 A comparing example: killings in Kosovo

We compare the methods outlined in the previous section using the dataset reported in Ball et al. (2002) about killings in Kosovo during March-June 1999. Four different sources have documented a total of 4400 deaths, i.e. the interviews conducted by the American Bar Association/Central and East European Law Initiative (1), the exhumation reports produced on behalf of the International Criminal Tribunal for Former Yugoslavia (2), the Human Rights Watch (3), and the Organization for Security and Cooperation in Europe (4). Table 3.4 summarises the number of casualties recorded by the four sources.
A decade after the conflict, the Humanitarian Law Center (HLC) has published a near-exhaustive list of victims (Center (2014)) for the whole period 1998-2000. Manrique-Vallier (2016) uses these data to compute the total number of casualties for the period considered by Ball et al.

	Model	$\hat{N}$	95% CI
Cormack 1992	[124][23][34]	10356	(9002 12122)
Madigan and York 1997	–	11257	(9352 14318)
Overstall and King 2014	–	12672	(9888 15728)

Table 3.5: Killings in Kosovo: results

(2002), giving us a point of reference for the "true" N, which is $N_{HLC} = 10401$.
Ball et al. (2002) estimates $\hat{N}$ for all possible hierarchical log-linear models, computing the confidence interval according to the profile likelihood method of Cormack (1992). According to the adjusted Pearson Chi-square statistic, the best model has one three-factor and two two-factor interaction terms, as shown in Table 3.5. The 95% confidence interval contains N_{HLC}. To obtain comparative estimates of the total number of casualties, we use the **R** packages **dga** and **conting** implementing Madigan and York (1997) and Overstall and King (2014b) methods, respectively. We remind that, according to the former, $\hat{N}$ represents the posterior mean of the unconditional posterior distribution of N; for the latter, $\hat{N}$ is the posterior mean of a distribution sampled via a reversible jump algorithm. Table 3.5 shows the results. We decided to use the default priors when implementing these models, which are noninformative priors. The credible intervals contain N_{HLC} as well, but they are much wider than the confidence interval obtained with Cormack (1992) method since they incorporate the uncertainty linked to the model.

3.4 Dealing with out-of-scope units

The overcoverage issue has become relevant only recently, with the increase of the interest of the NSIs in the production of statistics through data integration. Such a problem naturally arises since the aims of who collect the data and who use them differ. Out-of-scope units cannot be ignored; the estimate would result strongly biased otherwise. In the following paragraphs, we show how the models discussed in the previous section have since been extended to deal with the presence of latent out-of-scope units in the lists.

		$\delta_{i1}=1$	$\delta_{i1}=0$
$\delta_{i2}=1$	$i \in U$	y_{12}	y_2
	$i \notin U$	r_{12}	r_2
$\delta_{i2}=0$	$i \in U$	y_1	y_0
	$i \notin U$	r_1	

Table 3.6

3.4.1 Log-linear models

In the log-linear models' framework Zhang (2015) addresses the over-coverage issue directly modelling the probability of being erroneously classified, i.e. the *error rate*. The author introduces two alternatives to deal with out-of-scope units when $K = 2$, the first of which relies on the conditional independence assumption (CIA) at the base of the standard log-linear models; we briefly discuss the details.
Consider the latent structure of a contingency table of two lists, both affected by overcoverage, as in Table 3.6. Assume the cell counts to be Multinomial with parameters N^*, defined as the sum of N and the captured out-of-target units, and $\boldsymbol{p}_{\delta_U \omega}$:

$$\boldsymbol{p}_{\delta_U \omega} = \begin{cases} \boldsymbol{p}_{1\omega} & \text{if } i \in U \\ \boldsymbol{p}_{0\omega} & \text{otherwise} \end{cases} \tag{3.34}$$

We defined in §3.1 the cross-classified error rates as the probability that a unit does not belong to the target population given that it has been captured by the set of lists summarised by ω; we write it

$$\xi_\omega = P(i \notin U|\omega) \tag{3.35}$$

The error rates $\{\xi_\omega\}$ can be defined as functions of $\boldsymbol{p}_{U\omega}$, i.e.

$$\xi_{12} = \frac{p_{0\{12\}}}{p_{+\{12\}}} \tag{3.36}$$

$$\xi_1 = \frac{p_{0\{1\}}}{p_{+\{1\}}} \tag{3.37}$$

$$\xi_2 = \frac{p_{0\{2\}}}{p_{+\{2\}}} \tag{3.38}$$

Since we only observe x_1, x_2, x_{12}, the vector $\boldsymbol{p}_{\delta_U\omega}$ can not be estimated without further assumptions. Hence, assume that a coverage survey S exclusively affected by undercoverage is available and that its inclusion probability is equal to τ_S; also, let y_S be the number of units listed in S, and $y_{S\omega}$ the number of units captured by both S and the set of lists indexed by ω. Then, Zhang (2015) introduces a system of moment equations to model the observations as a function of the error rates:

$$\begin{cases} \mathbb{E}(y_{S12}|\boldsymbol{x}) = x_{12}(1-\xi_{12})\tau_S \\ \mathbb{E}(y_{S1}|\boldsymbol{x}) = x_1(1-\xi_1)\tau_S \\ \mathbb{E}(y_{S2}|\boldsymbol{x}) = x_2(1-\xi_2)\tau_S \\ \mathbb{E}(y_{S0}|\boldsymbol{x}) = (\mathbb{E}(N|\boldsymbol{x}) - x_{12}(1-\xi_{12}) - x_1(1-\xi_1) - x_2(1-\xi_2))\tau_S \end{cases} \tag{3.39}$$

The system is underidentified due to the presence of four parameters in the first three equations; the additional unknown in the fourth equation, $\mathbb{E}(N|\boldsymbol{x})$, can be derived given the estimates of the others. The idea is to impose a constraint on the ξ_ω's to make the system identifiable. Let us define a log-linear model for the $p_{\delta_U\omega}$'s; the largest nonsaturated model will be

$$\log(p_{U\omega}) = \beta + \beta_U + \beta_1 + \beta_2 + \beta_{U1} + \beta_{U2} + \beta_{12} \tag{3.40}$$

Now consider the logit of ξ_{12}; after some algebra, it results

$$\text{logit}(\xi_{12}) = \text{logit}(\xi_1) + \text{logit}(\xi_2) + p_{1\{\mathbf{0}\}} \quad ; \tag{3.41}$$

the model above is said to be *incidental* since it introduces a constraint between the error rate and the population size; thus, it can not be considered. To overcome this issue, Zhang (2015) sets a log-linear model for a transformation of $p_{U\omega}$, namely $q_{U\omega} = \frac{\boldsymbol{p}_{U\omega}}{1-p_{1\{\mathbf{0}\}}}$. Again, we can express the error rate as a function of the $q_{U\omega}$'s, i.e. $\xi_\omega = \frac{q_{0\{\omega\}}}{p_{+\{\omega\}}}$ and $\text{logit}(\xi_{12})$ becomes

$$\text{logit}(\xi_{12}) = \text{logit}(\xi_1) + \text{logit}(\xi_2) \tag{3.42}$$

which is not an incidental model and makes the system (3.39) identifiable. For small ξ_ω, $\text{logit}(\xi_\omega) \approx \log(\xi_\omega)$; hence,

$$\xi_{12} = \xi_1\xi_2 \quad . \tag{3.43}$$

However, in the case of good data quality and low error rates, it is reasonable to assume that the domain $\{12\}$ is much larger than both $\{1\}$ and $\{2\}$. Moreover, the error rate among the units in $\{12\}$ must be much lower than that in $\{1+\}$ and $\{+2\}$. Therefore, as an alternative to the previous model, Zhang (2015) suggests expressing ξ_{12} as the product of the marginal error rates; in other words, we can assume

$$P(i \notin U|\omega = \{12\}) = P(i \notin U|\omega+ = \{1+\})P(i \notin U|\omega+ = \{2+\}) \quad , \tag{3.44}$$

or

$$\xi_{12} = \xi_{1+}\xi_{2+} \quad . \tag{3.45}$$

Contrarily to identity (3.43), the condition above can not be derived from a standard log-linear model; thus, it does not rely on the concept of conditional independence; Zhang (2015) calls that in (3.45) the *pseudo conditional independence* (PCI) assumption. Zhang (2019) extends this concept to the case of $K \geq 2$.
The idea is to define a general log-linear model for the marginal $\xi_{\omega+}$

$$\log(\xi_{\omega+}) = \sum_{\nu \in \Omega(\omega)} \log \psi_{\nu+} \tag{3.46}$$

such that each unsaturated model corresponds to a different specification of the PCI assumption; e.g. the model including none of the interaction terms corresponds to the mutual PCI between the marginal list domains. ξ_ω is estimated via ML; we refer the reader to Zhang (2019) for the details.

3.4.2 Decomposable graphical models

Di Cecco (2019) extends the decomposable model described in §3.2.2, introducing a latent class (LC) approach developed both from a frequentist and a Bayesian perspective - the latter also proposed in Di Cecco et al. (2020).
There is a vast literature on the use of LC models in the capture-recapture framework, particularly addressing the heterogeneity problem; among others, see Agresti (1994), Bartolucci and Forcina (2001) and Bartolucci and Pennoni (2007). The first dealing with erroneous enumeration using LC models was Biemer et al. (2001a), followed by Biemer et al. (2001b) and Biemer et al. (2004). Such a strand of literature has led to frame the identifiability problems arising in this context, highlighting that at least

four lists are needed to estimate any LC model that includes interactions among lists; see Brown et al. (2004) and Biemer (2011) for further details. Here we describe the approach by Di Cecco (2019) mainly because of its computational advantages.
The class of model considered can be expressed as

$$p_\omega = \sum_{\delta_U} p_{\delta_U} p_{\omega|\delta_U} \tag{3.47}$$

Restrict the interest to decomposable models only. Since the latent variable δ_U interacts with all other variables, 3.20 can be written as

$$p_G = p_{\delta_U} p_{C_1|\delta_U} \prod_{h=2}^{g} p_{C_h|S_h} \ . \tag{3.48}$$

Such likelihood function may be maximised via the EM algorithm (see Di Cecco (2019) for a detailed description) or used to compute population size's posterior distribution in a fully Bayesian analysis. In the latter case, the prior specification is similar to that described in §3.2.2; it is sufficient to add a Beta-prior to p_{δ_U}. We can use MCMC methods to sample from the posterior distribution; in particular, a Gibbs sampler is appropriate in the case of Jeffreys prior on N; otherwise, we need a Metropolis-within-Gibbs.

3.4.3 Bayesian log-linear models

The last approach to overcoverage we analyse is that in Overstall et al. (2014), also adopted in Overstall and King (2014a) as an extension of the basic model. Assume that $J < K$ lists include units that are not part of the target population. For those cells ω affected by overcoverage, y_ω can be seen as the true value of a *left-censored* cell count since only its upper bound is observed, i.e. x_ω. Let $\boldsymbol{y}$ indicate the vector of counts such that $x_\omega = y_\omega$; let $\boldsymbol{x}^c$ and $\boldsymbol{y}^c$ be the vectors of observed counts and number of population units respectively such that $y_\omega < x_\omega$ (c stands for *censored*). Recall the joint posterior introduced in §3.2.3; now it becomes

$$\begin{aligned} \pi(y_0, \boldsymbol{y}^c, \phi, \beta_m, \sigma^2, m|\boldsymbol{y}, \boldsymbol{x}^c) \propto & \pi(\boldsymbol{y}, y_0, \boldsymbol{y}^c|\phi, \boldsymbol{\beta}_m, m)\times \\ & \times \pi(\boldsymbol{x}^c|\boldsymbol{y}^c)\pi(\boldsymbol{\beta}_m|\sigma^2, m)\pi(\sigma^2)\pi(m) \end{aligned} \tag{3.49}$$

The additional step to include in the algorithm relative to the model described in §3.2.3 is the sampling of the latent true count for those cells

affected by overcoverage;

$$y^c_\omega | \phi, \boldsymbol{\beta}_m, x^c_\omega, m \sim \text{Truncated Poisson}(\lambda_\omega, x^c_\omega) \tag{3.50}$$

3.5 Comparing examples with simulated data

As for the general capture-recapture setting, we may want to compare the methods described in the previous section. Nevertheless, while the three models introduced in the general framework apply to the same context and aim at the same objective, the models proposed for facing the overcoverage issue differ in their motivations. E.g. the idea behind the strand of LC models for capture-recapture is to identify the different behaviours of different (sub)populations captured on the same occasions, to estimate the size of the population of interest reliably.
In Overstall et al. (2014), the cross-classified overcount can be seen as a noise, a measurement error deriving from no specific underlying behaviour, yet Zhang (2019) models the erroneous enumerations relying on the goodness of data. Moreover, Zhang (2019) models a situation in which all the sources are affected by overcoverage but the enumeration survey. In contrast, Overstall et al. (2014) define a maximum number of cell counts with erroneous enumeration depending on the total number of observations to preserve the identifiability of the model. For the same reason, it is possible to estimate an LC model with less than four lists only if local independence is assumed.

Having this premise in mind, in the following paragraphs, we present two different simulated scenarios. For each of them, we compare the models described in §3.4 to see how different assumptions impact the results. The first scenario simulates the situation in which the presence of erroneous enumerations affects the sources homogeneously. The second scenario is the (typical) case in which a post-enumeration survey is conducted to assess the goodness of administrative lists in covering the target population and the data quality of such lists is pretty good.
We follow the frequentist approach in the estimation of the decomposable graphical models (see Di Cecco (2019)). We obtain the estimates for Overstall et al. (2014) model using the **R** package **conting** (see Overstall and King (2014a)).

			$\delta_{i1}=1$		$\delta_{i1}=0$	
			$\delta_{i2}=1$	$\delta_{i2}=0$	$\delta_{i2}=1$	$\delta_{i2}=0$
$\delta_{i3}=1$	$\delta_{i4}=1$	$i \in U$	25	40	22	446
		$i \notin U$	74	30	25	55
	$\delta_{i4}=0$	$i \in U$	148	245	134	2697
		$i \notin U$	200	81	67	148
$\delta_{i3}=0$	$\delta_{i4}=1$	$i \in U$	164	270	148	2981
		$i \notin U$	148	60	49	110
	$\delta_{i4}=0$	$i \in U$	992	1636	898	18034
		$i \notin U$	403	164	134	

Table 3.7: Data from scenario 1

3.5.1 Scenario 1: capturing two groups

First, we generate the contingency table in Table 3.7 from a decomposable graphical LC model $[\delta_U 12][\delta_U 3][\delta_U 4]$. We set the coverage rates of lists $k = 1, 2, 3, 4$ between 9% and 15%, yet their marginal error rates are equal to 0.25, 0.3, 0.15 and 0.12 respectively; the target population size N amounts to 28880, and the number of units captured by the four sources is equal to 30927.

We compare the estimates obtained via the EM algorithm for capture-recapture LC models by Di Cecco (2019) using the true data model with those obtained using Zhang (2019) algorithm and the **R** package **conting**. Table 3.8 shows the best results in terms of AIC or Bayesian p-value.

For Zhang (2019), we indicate which of the sources is preferred as being the error-free source, whereas for Overstall et al. (2014) the maximal model.

Model	$\hat{N}$	$\sum_{\omega} \hat{y}_{\omega}$
Di Cecco 2019 - $[\delta_U 12][\delta_U 3][\delta_U 4]$	28943	10903
Zhang 2019; error free list: $k = 4$	19238	9599
Overstall, King et al. 2014 - $[12][13][14][A_2 3][24][34]$	28428	27925
True values	28880	10846

Table 3.8: Results from scenario 1

In this scenario, the PCI based model from Zhang (2019) performs poorly in terms of y_0 estimation, whatever the choice of the error-free list. However, it is right in detecting the amount of out-of-scope units in the sample; according to the information criterion, such model best performs assuming list 4 to be the overcoverage-free source, which is, in fact, the list with the lowest error rate. On the other hand, for the Bayesian log-linear model by Overstall et al. (2014) the population size estimate improves as the number of interaction terms included in the maximal model increases, regardless of which of the cells are censored. Note that to include at least all the two-way interactions, the number of censored cells can not be more than three; the best performing model allows for the censoring of cells (x_1, x_2).

3.5.2 Scenario 2: post-enumeration survey and accurate administrative data

Table 3.9 summarises the capture histories of 40945 units during four capture occasions. $k = 1$ is an error-free source with a high population coverage rate, equal to 0.83. In contrast, the others are affected by the presence of erroneous enumerations, with marginal error rates set to be between 10% and 15%, with a total of 6288 out-of-target units. Since a post-enumeration survey should capture target units uniformly, we assume list 1 to be independent of the other sources. Indeed, lists $k = 2, 3, 4$ target captures are dependent on each other, i.e. the target counts in Table 3.9 are generated from the log-linear model [1][234]. Counts of out-of-scope units are added such that the error rates domains {23+}, {24+} and {34+} are smaller than {2+}, {3+} and {4+} and larger than {234+}; moreover, cross-classified error rates are much big-

			$\delta_{i1}=1$		$\delta_{i1}=0$	
			$\delta_{i2}=1$	$\delta_{i2}=0$	$\delta_{i2}=1$	$\delta_{i2}=0$
$\delta_{i3}=1$	$\delta_{i4}=1$	$i \in U$	768	5474	164	1145
		$i \notin U$	0	0	6	380
	$\delta_{i4}=0$	$i \in U$	3660	3703	711	721
		$i \notin U$	0	0	172	2259
$\delta_{i3}=0$	$\delta_{i4}=1$	$i \in U$	4563	4951	843	1006
		$i \notin U$	0	0	403	1696
	$\delta_{i4}=0$	$i \in U$	4112	1659	834	343
		$i \notin U$	0	0	1372	

Table 3.9: Data from scenario 2

ger than their respective marginal ones. We estimate the population size implementing the three models described in §3.4. Table 3.10 shows the results.

Zhang (2019) algorithm correctly detects list 1 as the error-free source, although it overestimates the number of out-of-target units. Nevertheless, it performs well in the estimation of y_0. On the other hand, Overstall et al. (2014) overestimate both the number of out-of-target units and y_0, despite it recognizes the true data model; in this case, it allows for the censoring of cells $(x_2, x_3, x_4, x_{23}, x_{24}, x_{34})$. Concerning the LC model, we tested the EM algorithm in Di Cecco (2019), allowing for the interaction between the latent variable and list 1; results are far from the true values. Hence, we split the estimation procedure into two steps. Firstly the EM for LC in capture-recapture is implemented considering the lists affected by erroneous enumeration only. Given the identifiability problems previously discussed, we are constrained to the local independence model, i.e.

Model	$\hat{N}$	$\sum_{\omega} \hat{y}_{\omega}$
Di Cecco 2019 - local independence	37291	36876
Zhang 2019 - error free list: $k = 1$	33345	33089
Overstall, King et al. 2014; [1][234]:	35258	33807
True values	34657	34314

Table 3.10: Results from scenario 2

$k = 2, 3, 4$ are set to be independent given the latent variable. Then, we use the vector $\hat{\boldsymbol{y}}$ to get an estimate of y_0 fitting a log-linear model. Here, the main issue consists of the impossibility of comparing (or average) different models and allowing for the manifest variables' interaction; this leads to underestimating the number of out-of-scope units.

3.6 Further topics

This chapter reviewed the primary and most recent literature dealing with the population size estimation problem, with an insight into the overcoverage issue. Nevertheless, we do not claim to be exhaustive. There exist methodological issues other than overcoverage that have not been covered in this paper and deserve attention. In real applications, the assumptions at the basis of log-linear models introduced in §3.2.1 may not hold. There might be heterogeneity in the population, i.e. the capture probabilities vary among individuals; it may occur that the captured units are not uniquely labelled across the multiple lists.

The latter case, i.e. the lack of unique labelling, implies a non-exact match of the units observed in multiple lists, hence a potential overestimation of the population of interest; in this framework, linkage uncertainty must be considered. Di Consiglio et al. (2019) reviews some of the approaches considering such uncertainty in the population size estimation procedure, from natural extensions of Fienberg (1972) model (see Di Consiglio and Tuoto (2018)), to fully Bayesian models like that in Tancredi and Liseo (2011). Unique labelling may miss within the sources as well, implying the presence of duplicates in some lists; see Tancredi et al. (2019) for an approach to this kind of issue.

The former case includes scenarios in which heterogeneity is either due to some measurable attributes, e.g. sex, age, or given by some unmeasurable characteristics. If the source of heterogeneity is known and the attributes are recorded in the available data, a convenient strategy is to stratify the population to obtain different homogeneous groups. Otherwise, literature offers a variety of methods to address the issue. Among them, we cite Fienberg et al. (1999), which encompasses the log-linear models and the Rasch model (Rasch (1960)) in a fully Bayesian hierarchical framework. The latent variable approach described in §3.4.2 can nimbly adapt to the heterogeneity problem; actually, in Di Cecco (2019) and Di Cecco et al. (2020) the mixture model is allowed to have more than two components. A nonparametric Bayesian approach dealing with population heterogeneity in capture-recapture experiments can be found in Manrique-Vallier (2016). The underlying idea is that, in the case of heterogeneous population and in the absence of covariates allowing to stratify the sample appropriately, it is convenient to assume that the population may be partitioned into an unknown number of homogeneous strata within which the independence model holds. In this case, the generating mechanism of the capture vectors consists of a general capture-recapture Multinomial model in which the probability mass function of each capture vector is a Dirichlet-process mixture of product-Bernoulli distributions. Indeed, letting the weights of the mixture be generated from a stick-breaking process allows to avoid the specification of the number of mixture components in advance; the identification of the number of homogeneous groups within the heterogeneous population occurs in an unsupervised way. Manrique-Vallier (2016) applies the nonparametric latent class model to the data killings in Kosovo shown in §1.2, and it performs very well. The point estimate of N is incredibly close to the count NHLC. See Table 3.6 to compare Manrique-Vallier (2016) results to the real count of casualties and to the best model in Ball et al. (2002). We refer the reader to Manrique-Vallier (2016) for the model's details. From the same author, and specifically for casualties estimation in capture-recapture experiments, see Manrique-Vallier et al. (2019). Yet, for a review of most of the literature on heterogeneous animal populations in capture-recapture models, see Gimenez et al. (2018).

	$\hat{N}$	95% CI
HLC count	10401	–
Manrique-Vallier 2016	10442	(9020 13637)
Ball et al. 2002 $[A_1\ A_2\ A_4][A_2\ A_3][A_3\ A_4]$	10356	(9002 12122)

Table 3.11: Killings in Kosovo: Manrique-Vallier 2016 results

Chapter 4

Log-linear models in the presence of out-of-scope units

In the previous chapter, we revised the primary and most recent approaches to population size estimation in a capture-recapture framework in the presence of out-of-scope units. It emerges that the best model's choice depends on the situation addressed since the discussed approaches differ in their motivations. In particular, the described methods seem to differ for the erroneous enumerations' reference set. For instance, consider the methods by Zhang (2019) and Di Cecco (2019). The former explicitly defines the reference set for the out-of-scope units, which consists of the union of the lists; in other words, the population of the out-of-scope units is a subset of the lists' universe. The latter does not explicitly define a reference set; the presence of out-of-scope units is also allowed in the unobserved cell, underlying the existence of two populations with different capture probabilities.
In this chapter, we assume an open erroneous enumerations' set a priori, but we constrain this set to the lists' universe once we observe the contingency table. The Bayesian approach allows us to formalise this kind of assumption setting a Poisson prior on the latent erroneous enumerations; a posteriori, their distribution will be conditioned on the observed data.

The framework we consider in this chapter is the current one for many of the National Statistics Institutes in the developed countries. Indeed, they are experiencing the "shift" from a census-based statistics paradigm to a register-based one, where the data quality is good, and lists' error rates are low, as depicted by Zhang (2015). We propose a flexible model,

	$\delta_{i1} = 1$		$\delta_{i1} = 0$	
	$\delta_{i2} = 1$	$\delta_{i2} = 0$	$\delta_{i2} = 1$	$\delta_{i2} = 0$
$\delta_{i3} = 1$	33630	24324	84175	621654
$\delta_{i3} = 0$	181495	332999	544792	?

Table 4.1: Resident individuals in Lazio, Italy, simulated data

especially suitable to situations where strong prior information is available. Moreover, we address the problem of model selection, which is not trivial in the framework of log-linear models for capture-recapture when we have a few (or even zero) degrees of freedom, as highlighted in Zhang (2019). To give an intuition about the importance of such an issue, assume that for the first time, the Istat aims to estimate Italian regions' population size via administrative data only. Imagine having three sources that partially enumerate the population of the region of, e.g., Lazio: an enumeration survey (1), a list from the health system (2), and the tax register (3). Assume Table 4.1 summarises the captured individuals. Suppose we fit a log-linear regression on the observed contingency table, depending on the model's specification. In that case, we will obtain very different results: the estimated total population varies from slightly more than one million people to more than ten million. Whether there is the likely presence of out-of-scope units, it does not matter how precisely the model fits the data.

In §4.1, we present our alternative modelisations of the erroneous counts in a Bayesian log-linear framework. We follow the notation introduced in §3.1. After the prior setting in §4.2, §4.3 describes the scenario of good data quality and low lists' error rates, as depicted by Zhang (2015) and extended by Zhang (2019); there, we introduce the Pseudo Conditional Independence assumption in a Bayesian framework. §4.4 is devoted to the computation of the joint posterior distribution of the parameters of interest; there, we show how Fisher's noncentral hypergeometric distribution, the main object of Part I of this work, enters the sampling process. §4.5 suggests the well-known method by Chib and Jeliazkov (2001) as a

model selection method for log-linear models' in the presence of latent variables. Finally, a simulation study is presented in §4.6.

4.1 Out-of-scope units in a Bayesian log-linear models framework

Each observed cross-classified count x_ω can be seen as a realisation of a random variable

$$X_\omega = Y_\omega + R_\omega \ , \tag{4.1}$$

where Y_ω represents the latent count of target units associated with the cell indexed by ω, and R_ω the relative out-of-scope units' count. Concerning the former, we may specify:

$$Y_\omega \sim \text{Pois}(\lambda_\omega) \tag{4.2}$$

independently for all ω, where $\log(\lambda_\omega) = \phi + \mathbf{d}'_{m,\omega}\boldsymbol{\beta}_m$, ϕ being the grand mean, $\boldsymbol{\beta}_m$ the coefficients' vector for model m, and $\mathbf{d}_{m,\omega}$ the design vector that indicates which elements of $\boldsymbol{\beta}_m$ apply to the cell indexed by ω (as specified in §3.2.1). The log-linear models considered are the hierarchical ones, the minimal being the independence model (no interaction terms) and the maximal the saturated. As an alternative specification for the *in-target* units, we could have assumed

$$\boldsymbol{Y} \sim \text{Multinom}(N, \boldsymbol{p} := \{p_\omega\}) \tag{4.3}$$

where $p_\omega = \frac{\exp\{\mathbf{d}'_{m,\omega}\boldsymbol{\beta}_m\}}{\sum_\omega \exp\{\mathbf{d}'_{m,\omega}\boldsymbol{\beta}_m\}}$. The two specifications are equivalent under strict assumptions (see Overstall and King (2014b)).

Now let us focus on the *erroneous enumeration* problem. In §3.1 we define $A = \{1, 2, ..., K'\}$ the lists' set only enumerating units belonging to the target population U and $B = \{K'+1, K'+2..., K\}$ the set of lists also including some units $i \notin U$. Therefore, we assume

$$R_\omega \begin{cases} = 0 & \forall\omega \text{ s.t. } \delta_k = 1 \text{ for at least one } k \in A \\ \sim \text{Pois}(\mu_\omega) & \text{otherwise} \end{cases} \tag{4.4}$$

where μ is set to be equal to $x_\omega\xi_\omega$, with $\xi_\omega = P(i \notin U|\omega = \{\boldsymbol{k}\})$ being the marginal error rate. Alternatively, we may specify

$$R_\omega \sim \text{Binom}(x_\omega, \xi_\omega) \ . \tag{4.5}$$

The main advantage of the above specification is its computational convenience.

4.2 Prior specification

We need to define $\pi(\boldsymbol{\theta}_m, \boldsymbol{\xi})$, i.e. the joint prior distribution of the in-target and out-of-target counts' parameters. It is plausible to assume

$$\pi(\boldsymbol{\theta}_m, \boldsymbol{\xi}) = \pi(\boldsymbol{\theta}_m)\pi(\boldsymbol{\xi}) \tag{4.6}$$

i.e. the parameters related to the target population are a priori independent of the erroneous enumerations' ones.
Concerning the log-linear model, under the Poisson specification Overstall and King (2014a) use the generalized hyper-g prior by Sabanés Bové and Held (2011) for $(\boldsymbol{\theta}_m)$, i.e. $\pi(\boldsymbol{\theta}_m|m) = \pi(\phi)\pi(\boldsymbol{\beta}_m|m)$ where $\pi(\phi) \propto 1$ and

$$\boldsymbol{\beta}_m|\sigma^2, m \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_m = 2^K \sigma_\beta^2 (\mathbf{d}_m' \mathbf{d}_m)^{-1}) \tag{4.7}$$

$$\sigma_\beta^2 \sim \mathrm{InvGa}\left(\frac{h_1}{2}, \frac{h_2}{2}\right) \tag{4.8}$$

with h_1, h_2 fixed. We may interpret such prior distribution as the posterior distribution from a locally uniform prior and an imaginary sample where $\frac{1}{\sigma_\beta^2}$ is the size of the "prior sample", i.e. the prior contains $\frac{1}{\sigma_\beta^2}$ as much information as the data $\boldsymbol{y}$ (Sabanés Bové and Held (2011)). According to Sabanés Bové and Held (2011), ϕ parameterises the average linear predictor in each model, thus using an improper flat prior seems appropriate.
The specification of the prior distribution of $\boldsymbol{\beta}_m$ looks coherent with our purposes; however, we might need to introduce some extra-experimental prior information about ϕ. One of the most common situations is where the consulted experts can express their beliefs in terms of "confidence" or "credible" intervals $[q_1, q_2]$. We can define ϕ

$$\phi \sim \mathrm{N}(\varphi, \sigma_\phi^2) \tag{4.9}$$

and specify

$$\begin{cases} q_1 = \varphi - z_{1-\alpha}\sigma_N \\ q_2 = \varphi + z_{1-\alpha}\sigma_N \end{cases} \tag{4.10}$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard Normal distribution. We obtain the hyperparameters φ and σ_N after simple algebra for fixed α. Hence, assuming independence between ϕ and $\boldsymbol{\beta}_m$,

$$\boldsymbol{\theta}_m := (\phi, \boldsymbol{\beta}_m) \sim \mathrm{N}((\varphi, \mathbf{0})', \boldsymbol{\Sigma}_m^{\phi}) \tag{4.11}$$

where $\boldsymbol{\Sigma}_m^{\phi}$ is the variance-covariance matrix whose first row and first column elements are 0's, except for the first element, which is equal to σ_ϕ^2; the remaining minor is equal to $\boldsymbol{\Sigma}_m$.

Concerning the erroneous enumerations, since x_ω is observed, we only need to introduce some information on ξ_ω, which represents a probability; thus, we assume

$$\xi_\omega \sim \mathrm{Beta}(a_\omega, b_\omega)\ . \tag{4.12}$$

4.3 Erroneous enumerations' parameters elicitation

It is reasonable to assume that experts may have prior beliefs on the error rate of each list k, namely on the probability to be out-of-scope given that the units have been listed in the kth source.
Hence, the experts would express their guess in terms of confidence or credible intervals $[q_1; q_2]$ for *marginal* error rates $\{\xi_{\omega+}\}$ rather than for *cross-classified* ones[1], $\{\xi_\omega\}$, where the cardinality of ω is equal to 1 and ω can only be equal to $\{K'+1\}, \{K'+2\}, \ldots, \{K\}$. We may either follow the Normal approximation method mentioned in §4.2 or rely on the error rates' quantile function and proceed via numerical approximation. We briefly describe both methods below.

- Normal approximation

$$\begin{cases} q_1 = \mathbb{E}(\xi_{\omega+}) - z_{1-\alpha}\sqrt{\mathbb{V}(\xi_{\omega+})} \\ q_2 = \mathbb{E}(\xi_{\omega+}) + z_{1-\alpha}\sqrt{\mathbb{V}(\xi_{\omega+})} \end{cases} \tag{4.13}$$

[1] as defined in chapter 3

where

$$\begin{cases} \mathbb{E}(\xi_{\omega+}) = \dfrac{a_{\omega+}}{a_{\omega+} + b_{\omega+}} \\ \mathbb{V}(\xi_{\omega+}) = \dfrac{a_{\omega+} b_{\omega+}}{(a_{\omega+} + b_{\omega+} + 1)(a_{\omega+} + b_{\omega+})^2} = \dfrac{\mathbb{E}(\xi_{\omega+})(1 - \mathbb{E}(\xi_{\omega+}))}{a_{\omega+} + b_{\omega+} + 1} \end{cases} . \tag{4.14}$$

After some algebra and for fixed α, we obtain the hyperparameters $a_{\omega+}$ and $b_{\omega+}$

$$\begin{cases} a_{\omega+} = \dfrac{\mathbb{E}(\xi_{\omega+})}{1 - \mathbb{E}(\xi_{\omega+})} \left(\dfrac{\mathbb{E}(\xi_{\omega+})(1 - \mathbb{E}(\xi_{\omega+}))^2}{\mathbb{V}(\xi_{\omega+})} + \mathbb{E}(\xi_{\omega+}) - 1 \right) \\ b_{\omega+} = \dfrac{\mathbb{E}(\xi_{\omega+})(1 - \mathbb{E}(\xi_{\omega+}))^2}{\mathbb{V}(\xi_{\omega+})} + \mathbb{E}(\xi_{\omega+}) - 1 \end{cases} . \tag{4.15}$$

Such a method is straightforward to implement; the main shortcoming is that we are constrained to symmetric intervals.

- Quantile function and numerical approximation
  Berger (1985) suggests estimating some quantiles of the prior distribution subjectively and choose the parameters to obtain a density matching these quantiles. Let q_v, $v \in [0, 1]$ be the v-quantile of the $\xi_{\omega+}$ distribution, i.e. a point such that $\xi_{\omega+}$ has a probability v of being less than or equal to q_v:

$$P(\xi_{\omega+} \leq q_v) = \int_0^{q_v} f(\xi_{\omega+}; a_{\omega+}, b_{\omega+}) \, \mathrm{d}\xi_{\omega+} = v \ . \tag{4.16}$$

  The inverse of such cumulative distribution function is the quantile function, $F^{-1}(v)$. Assuming to have at least two quantiles, q_{v1} and q_{v2}, we solve

$$\begin{cases} F^{-1}(v1) - q_{v1} = 0 \\ F^{-1}(v2) - q_{v2} = 0 \end{cases} \tag{4.17}$$

  for $a_{\omega+}$ and $b_{\omega+}$ via Newton-Raphson algorithm, and obtain the hyperparameters.

The Pseudo Conditional Independence assumption

For those $\xi_{\omega+}$'s s.t. $|\omega| > 1$, we refer to the Pseudo Conditional Independence assumption by Zhang (2015) and Zhang (2019) discussed in §3.4.1 and we include it in this Bayesian framework.

Recall that, in the case of two lists, the PCI is defined as

$$\xi_{12} = \xi_{1+}\xi_{2+} \ , \tag{4.18}$$

whose meaning is that the probability of a unit to be an erroneous enumeration given that both the sources have captured it is much lower than the probability to be out-of-target given that a single list captures it. Such an assumption is a natural way to model the error rates structure in the considered framework, i.e. when the data quality is high, and the error rates are low.
If the error rates are fixed in the case of $K \geq 3$, it is sufficient to introduce the PCI as in Zhang (2019). However, whether the error rates have a prior distribution, a modification might be convenient. Hence, we introduce a constraint on the first moment of the marginal error rate rather than on its entire distribution, i.e.

$$\mathbb{E}(\xi_{\omega+}) = \prod_{\nu \in \Omega(\omega)} \mathbb{E}(\xi_{\nu+}) \tag{4.19}$$

where $\Omega(\omega)$ is the set of all non-empty subsets of ω. To uniquely identify the two hyperparameters $a_{\omega+}$ and $b_{\omega+}$, we need another equation. One possibility is to constrain higher moments, or the variance of $\xi_{\omega+}$[2]. Then, we use the equations in the system (4.15) to obtain the hyperparameters for $\xi_{\omega+}$ when $|\omega| > 1$.

Once defined the marginal error rates' distributions, we can finally derive the relative cross-sectionals'; see Appendix C for the cross-sectional error rates derivation.

[2] $\mathbb{V}(\xi_{\omega+})$ can be arbitrarily chosen. To have enough variability avoiding wide jumps in the MCMC, a rule of thumb might be fixing $\mathbb{V}(\xi_{\omega+}) = \left(\frac{\mathbb{E}(\xi_{\omega+})}{3}\right)^2$

4.4 Posterior computation

We aim to estimate the joint posterior distribution of $\boldsymbol{\theta}_m, \boldsymbol{\xi}, y_0$ and $\boldsymbol{r}$ (or, equivalently, the latent $\boldsymbol{y}$):

$$\pi(\boldsymbol{\theta}_m, \boldsymbol{\xi}, y_0, \boldsymbol{r}|\boldsymbol{x}) \propto f(y_0, \boldsymbol{x}|\boldsymbol{\theta}_m, \boldsymbol{\xi}, \boldsymbol{r})\pi(\boldsymbol{r}|\boldsymbol{x}, \boldsymbol{\xi})\pi(\boldsymbol{\theta}_m)\pi(\boldsymbol{\xi}) \tag{4.20}$$

Given such distribution's intractability, we generate an MCMC sample using a Metropolis-within-Gibbs; see Algorithm 7.

Under the Poisson specification, we sample the latent erroneous counts using Fisher's noncentral hypergeometric distribution, which is the object of chapters 1 and 2. The motivation lies in the proportionality of r_ω's full conditional to FNCH; the results we obtained from a mathematical-computational point of view are shown in Appendix D.
From an applicative perspective, the use of FNCH in the posterior sampling process is a crucial point. Indeed, sampling from such distribution allows for the possibility of expressing prior beliefs about erroneous enumerations in the lists in terms of relative odds. For instance, assume to know that the in-target units' capture probability weights about twice the out-of-target ones in the kth source; namely

$$\frac{P(\omega+ = \{k+\}|i \notin U)/(1 - P(\omega+ = \{k+\}|i \notin U))}{P(\omega+ = \{k+\}|i \in U)/(1 - P(\omega+ = \{k+\}|i \in U))} \simeq \frac{1}{2}\ . \tag{4.21}$$

In the absence of further information, there exist infinite solutions to the equation 4.21. However, the expression above reminds us of FNCH weight parameter, namely

$$w = \frac{\zeta_r/(1-\zeta_r)}{\zeta_y/(1-\zeta_y)}\ ;$$

see chapter 1 for further details. Hence, eliciting a prior for w we may identify the latent erroneous enumerations via MCMC.
For the sake of completeness, we point out that assuming the Binomial specification mentioned in §4.1 for R_ω, the conjugacy makes the ξ_ω sampling step faster and the modification of Algorithm 7 straightforward. However, introducing information about the erroneous enumerations in terms of relative odds is no longer possible.

Algorithm 7: Log-linear model in the presence of erroneous enumerations

Choose initial values $\boldsymbol{\theta}^0$, $\boldsymbol{\xi}^0$, $\boldsymbol{r}^0$, y_0^0 and ${\sigma_\beta^2}^0$;
for $t \leftarrow 1$ **to** T **do**
  set $\tilde{N} = N^{t-1}$;
  for $\omega = \{\boldsymbol{k}\}$ *s.t.* $\boldsymbol{k} \in B$ **do**
    draw ξ_ω^* from a proposal distribution $q_{\xi,t}(\xi_\omega^* | \xi_\omega^{t-1})$;
    compute the acceptance ratio
    $\gamma_{\xi_\omega} = \min\left(1; \frac{\pi(\xi_\omega^* | r_\omega^{t-1}, y_\omega^{t-1}) q_t(\xi_\omega^{t-1} | \xi_\omega^*)}{\pi(\xi_\omega^{t-1} | r_\omega^{t-1}, y_\omega^{t-1}) q_t(\xi_\omega^* | \xi_\omega^{t-1})}\right)$;
    draw $u \sim \text{Unif}(0,1)$;
    if $u < \gamma_{\xi_\omega}$ **then**
      set $\xi_\omega^t = \xi_\omega^*$
    else
      $\xi_\omega^t = \xi_\omega^{t-1}$
    end
    set $\mu_\omega^t = x_\omega \xi_\omega^t$;
    set $w_\omega = \frac{\mu_\omega^t / M}{\lambda_\omega^{t-1} / N^{t-1}} \frac{1 - (\lambda_\omega^{t-1} / N^{t-1})}{1 - (\mu_\omega^t / M)}$;
    draw $r_\omega^t \sim \text{FNCH}\left(M, N^{t-1}, x_\omega, w_\omega\right)$;
    set $y_\omega^t = x_\omega - r_\omega^t$;
    set $\tilde{N}$ equal to the sum of the latest y_ω
  end
  draw $y_0^t \sim \text{Pois}(\exp\{\phi^{t-1}\})$;
  draw ${\sigma_\beta^2}^t \sim$
  $\text{InvGamma}\left(\frac{a + l_m}{2}, \frac{d + 2^{-K} {\boldsymbol{\beta}_m'}^{t-1} \boldsymbol{X}_m' \boldsymbol{X}_m {\boldsymbol{\beta}_m}^{t-1}}{2}\right)$
  where l_m is the dimension of $\boldsymbol{\beta}_m$;
  draw $\boldsymbol{\theta}_m^*$ from a proposal distribution $q_{\boldsymbol{\theta},t}(\boldsymbol{\theta}_m^* | \boldsymbol{\theta}_m^{t-1})$;
  compute the acceptance ratio
  $\gamma_{\boldsymbol{\theta}_\omega} = \min\left(1; \frac{\pi(\boldsymbol{\theta}_m^* | \boldsymbol{y}^t) q_{\boldsymbol{\theta},t}(\boldsymbol{\theta}_m^{t-1} | \boldsymbol{\theta}_m^*)}{\pi(\boldsymbol{\theta}_m^{t-1} | \boldsymbol{y}^t) q_{\boldsymbol{\theta},t}(\boldsymbol{\theta}_m^* | \boldsymbol{\theta}_m^{t-1})}\right)$;
  draw $u \sim \text{Unif}(0,1)$;
  if $u < \gamma_{\boldsymbol{\theta}_\omega}$ **then**
    $\boldsymbol{\theta}_\omega^t = \boldsymbol{\theta}_\omega^*$
  else
    $\boldsymbol{\theta}_\omega^t = \boldsymbol{\theta}_\omega^{t-1}$
  end
  ;
end

4.5 Model selection

To select the "best" among the possible models, according to their log-linear specification, we proceed to the models' comparison using the Bayes factor. The Bayes factor is the ratio of the marginal distributions under the alternative models 0 and 1:

$$BF = \frac{\mathrm{m}_m(\boldsymbol{y})}{\mathrm{m}_{m'}(\boldsymbol{y})} \tag{4.22}$$

with $\mathrm{m}_m(.)$ and $\mathrm{m}_{m'}(.)$, being the normalizing constant of the mth and m'^{th} model posterior distributions respectively, i.e.

$$\mathrm{m}_m(\boldsymbol{y}) = \int_{\Psi} f(\boldsymbol{y}|\boldsymbol{\psi_m})\pi(\boldsymbol{\psi_m})\mathrm{d}\boldsymbol{\psi_m} \tag{4.23}$$

where $\boldsymbol{\psi}_m$ here is the vector of all m^{th} model's parameters. This ratio is an indicator of the relative evidence of one model against the other model. According to Kass and Raftery (1995), a value of Bayes factor greater than 3.2 (or smaller than 1/3.2) is a substantial evidence in favor of model m against model m' (or the other way around)[3].
Except for the models in which the parameters' distribution is conjugate, computing the normalising constant of a posterior is often challenging. Chib (1995) introduced a new approach to compute the marginal density of $\boldsymbol{y}$ under a particular model m when the posterior distribution can only be approximated via MCMC, particularly by the output of a Gibbs sampler. Exploiting the fact that $m(\boldsymbol{y})$ can be written as

$$m(\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})}{\pi(\boldsymbol{\psi}|\boldsymbol{y})} \tag{4.24}$$

for any value of $\boldsymbol{\psi}$, $m(\boldsymbol{y})$ is computed evaluating all the quantities above at a certain high posterior density value $\boldsymbol{\psi}^*$; i.e., in this work,

$$m(\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_m^*, \boldsymbol{\xi}^*)\pi(\boldsymbol{\theta}_m^*, \boldsymbol{\xi}^*)}{\pi(\boldsymbol{\theta}_m^*, \boldsymbol{\xi}^*|\boldsymbol{y})} \ . \tag{4.25}$$

[3]We refer to the following scale of interpretation of the Bayes Factor's value:

- $\frac{1}{3.2} < BF < 3.2$: not worth than a bare mention;
- $3.2 < BF < 10$ or $\frac{1}{10} < BF < \frac{1}{3.2}$: substantial evidence in favour of one of the models;
- $BF < \frac{1}{10}$ or $BF > 10$: strong evidence in favour of one of the model.

We need to estimate $\pi(\boldsymbol{\theta}_m^*, \boldsymbol{\xi}^*|\boldsymbol{y})$ in order to compute $m(\boldsymbol{y})$. By the law of total probability, our posterior will be

$$\pi(\boldsymbol{\theta}_m^*, \boldsymbol{\xi}^*|\boldsymbol{y}) = \pi(\boldsymbol{\theta}_m^*|\boldsymbol{y})\pi(\boldsymbol{\xi}^*|\boldsymbol{\theta}_m^*, \boldsymbol{y}) \tag{4.26}$$

where

$$\pi(\boldsymbol{\theta}_m^*|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}_m^*|\boldsymbol{y}^*, \boldsymbol{y}^l, y_0)\pi(\boldsymbol{y}^l, y_0|\boldsymbol{y}^*)\ \mathrm{d}\boldsymbol{y}^l\ \mathrm{d}y_0 \tag{4.27}$$

and

$$\pi(\boldsymbol{\xi}^*|\boldsymbol{y}, \boldsymbol{\theta}_m^*) = \int \pi(\boldsymbol{\xi}^*|\boldsymbol{y}^*, \boldsymbol{y}^l, y_0, \boldsymbol{\theta}_m^*)\pi(\boldsymbol{y}^l, y_0, \boldsymbol{\theta}_m^*|\boldsymbol{y}^*)\ \mathrm{d}\boldsymbol{y}^l\ \mathrm{d}y_0\ . \tag{4.28}$$

Under the Binomial specification for $\boldsymbol{R}$, the normalising constant of the posterior distribution of the error rates $\boldsymbol{\xi}$ is known due to conjugacy. In such a case, we can obtain an estimate of $\pi(\boldsymbol{\xi}^*|\boldsymbol{y}, \boldsymbol{\theta}_m^*)$ by taking the average of the full conditional density evaluated at $(\boldsymbol{\theta}_m^*)$ for G iterations, i.e.

$$\hat{\pi}(\boldsymbol{\xi}^*|\boldsymbol{y}, \boldsymbol{\theta}_m^*) = \frac{1}{G}\sum_{g=1}^{G} \pi(\boldsymbol{\xi}^*|\boldsymbol{y}^{(g)}, \boldsymbol{\theta}_m^*) \tag{4.29}$$

However, we ignore the normalising constant of $\pi(\boldsymbol{\theta}_m|\boldsymbol{y})$, making the approach in Chib (1995) not feasible. Nonetheless, Chib and Jeliazkov (2001) suggest how to estimate the posterior in such cases, setting

$$\hat{\pi}(\boldsymbol{\theta}_m^*|\boldsymbol{y}) = \frac{\frac{1}{G}\sum_{g=1}^{G} \gamma(\boldsymbol{\theta}_m^{(g)}, \boldsymbol{\theta}_m^*|\boldsymbol{y}^{(g)})q(\boldsymbol{\theta}_m^*|\boldsymbol{\theta}_m^{(g)})}{\frac{1}{J}\sum_{j=1}^{J} \gamma(\boldsymbol{\theta}_m^*, \boldsymbol{\theta}_m^{(j)}|y^{(j)})} \tag{4.30}$$

where $\{\boldsymbol{\theta}_m^{(g)}\}$ are the sampled draws from the posterior, and $\{\boldsymbol{\theta}_m^{(j)}\}$ are additional draws from a proposal distribution $q(\boldsymbol{\theta}_m|\boldsymbol{\theta}_m^*)$ and

$$\gamma(\boldsymbol{\theta}_m^{(g)}, \boldsymbol{\theta}_m^*|\boldsymbol{y}^{(g)}) = \min\left\{1, \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_m^*, \boldsymbol{y}^{(g)})\pi(\boldsymbol{\theta}_m^*)}{f(\boldsymbol{y}|\boldsymbol{\theta}^{(g)}, \boldsymbol{y}^{(g)})\pi(\boldsymbol{\theta}^{(g)})}\frac{q(\boldsymbol{\theta}^{(g)}|\boldsymbol{\theta}_m^*)}{q(\boldsymbol{\theta}_m^*|\boldsymbol{\theta}^{(g)})}\right\}\ . \tag{4.31}$$

Under the Poisson specification for $\boldsymbol{R}$, we estimate $\pi(\boldsymbol{\xi}^*|\boldsymbol{y}, \boldsymbol{\theta}_m^*)$ in a similar fashion. Now we can compute $m(\boldsymbol{y})$ for each model and then obtain the Bayes factor.

4.6 Simulation studies

4.6.1 Model [12][13][23]

We simulate 100 three-lists complete contingency tables from fixed $\boldsymbol{\theta}$ coefficients, i.e. $\boldsymbol{\theta} = (\phi, \beta_1, \beta_2, \beta_3, \beta_{12}, \beta_{13}, \beta_{23})' = (6.80, 1.62, 0.90, 0.80, -0.75, -0.86, -0.30)'$. Then, we partition the lists' universe as follows: $A = \{1\}$ and $B = \{2, 3\}$. For each sample, we simulate out-of-target counts for the cells indexed by $\omega = (\{2\}, \{3\}, \{23\})$ from fixed marginal error rates, namely $\xi_{1+} = 0.1$ and $\xi_{2+} = 0.2$, deriving the cross-sectional error rates' prior hyperparameters with the quantile method mentioned in §4.3. We run the MCMC[4] described in the previous section for all samples and, for each sample, for all model specifications, discarding the saturated model. For convenience, we use the Binomial specification for $\{R_\omega\}$.
All models manage to center the posterior distributions of the error rates on their true values thanks to the strong prior information, and give the estimates in Table 4.2. Table 4.3 summarises the posteriors of the model's coefficients for each log-linear model's specification.

Marginal rates		Cross-classified rates	
ξ_{2+}	0.101 (0.010)	ξ_2	0.409 (0.045)
ξ_{3+}	0.203 (0.006)	ξ_3	0.632 (0.018)
ξ_{23+}	0.019 (0.004)	ξ_{23}	0.037 (0.008)

Table 4.2: Estimates for marginal and cross-classified error rates.

Table 4.4 shows the mean and standard deviation of the Mean Square Error computed for each model and each sample. The simulation study provides the smallest MSE (on average and for every sample) with the lowest standard deviation for the true model. Table 4.5 shows for each model how often the true population size is included in different Highest Posterior Density intervals. As the interval shrinks, all models lose their ability to simulate values of N so near the true value; the true model still performs well.

[4]number of iterations: 50000; burnin: 25000

Model	ϕ	β_1	β_2	β_3	β_{12}	β_{13}	β_{23}
[1][2][3]	7.089	0.547	0.110	-0.069			
	(0.035)	(0.030)	(0.019)	(0.022)			
[12][3]	7.577	0.809	0.398	-0.032	-0.382		
	(0.049)	(0.048)	(0.053)	(0.023)	(0.057)		
[13][2]	7.508	0.886	0.160	0.323		-0.531	
	(0.076)	(0.076)	(0.020)	(0.078)		(0.081)	
[1][23]	7.865	0.533	0.039	-0.145			0.127
	(0.034)	(0.029)	(0.030)	(0.029)			(0.039)
[12][13]	7.031	1.438	0.638	0.514	-0.622	-0.722	
	(0.089)	(0.091)	(0.055)	(0.080)	(0.060)	(0.083)	
[12][23]	7.600	0.798	0.368	-0.055	-0.370		0.037
	(0.052)	(0.048)	(0.067)	(0.031)	(0.060)		(0.044)
[13][23]	7.515	0.882	0.154	0.313		-0.528	0.011
	(0.076)	(0.075)	(0.032)	(0.082)		(0.080)	(0.042)
[12][13][23]	6.726	1.672	0.943	0.819	-0.790	-0.874	-0.305
	(0.099)	(0.097)	(0.074)	(0.091)	(0.066)	(0.086)	(0.051)

Table 4.3: Estimates for $\boldsymbol{\theta}$. True model: [12][13][23].

Model	Mean	Sd	$\text{MSE}_{[12][13][23]} < \text{MSE}_m$
[1][2][3]	615397.632	31162.066	1
[12][3]	413046.688	28388.089	1
[13][2]	290839.211	22171.050	1
[1][23]	685122.540	35859.447	1
[12][13]	63269.146	12018.815	1
[12][23]	428186.170	30910.644	1
[13][23]	294395.017	21996.929	1
[12][13][23]	1275.036	706.625	-

Table 4.4: Mean Square Errors. True model: [12][13][23]

Model	$N^* \in \text{HPD}_{95\%}$	$N^* \in \text{HPD}_{80\%}$	$N^* \in \text{HPD}_{50\%}$
[1][2][3]	0	0	0
[12][3]	0	0	0
[13][2]	0	0	0
[1][23]	0	0	0
[12][13]	1	1	0.56
[12][23]	0	0	0
[13][23]	0	0	0
[12][13][23]	1	1	0.71

Table 4.5: The true population size lies in the Highest Posterior Density interval, frequencies. True model: [12][13][23]

[12][13][23]	$(-\infty, 0.1)$	$[1, 3.2)$	$[3.2, 10)$	$[10, \infty)$
[1][2][3]	0	0	0	1
[12][3]	0	0	0	1
[13][2]	0	0	0	1
[1][23]	0	0	0	1
[12][13]	0.04	0	0.01	0.94
[12][23]	0	0	0	1
[13][23]	0	0	0	1

Table 4.6: Interval values of the Bayes Factor. True model [12][13][23] against the others

Table 4.6 shows how often the Bayes Factor computed using Chib (1995) and Chib and Jeliazkov (2001) method favours the true model against the others. The true model is not favoured in very few cases, particularly against model [12][13]: likely, it is because the interaction coefficient β_{23} is close to zero (equal to -0.3).

4.6.2 Model [13][2]

The scope of this second study is to test the performance of the model selection method in the presence of fewer interaction terms. Indeed, to allow for more interactions generally let the model fit better the data; we need to verify that the true model can be recognised.
We run another simulation using 100 complete contingency tables with $K = 3$ from a vector $\boldsymbol{\theta} = (\phi, \beta_1, \beta_2, \beta_3, \beta_{13})' = (6.80, 1.62, 0.90, 0.80, -0.86)'$. Hence, now our true model is [13][2]. As before, we partitioned the lists' universe in $A = \{1\}$ and $B = \{2, 3\}$ and we simulated out-of-target counts for the cells indexed by $\omega = (\{2\}, \{3\}, \{23\})$ from $\xi_{1+} = 0.1$ and $\xi_{2+} = 0.2$. Again, we derive the cross-sectional error rates' prior hyperparameters with the quantile method mentioned in §4.3.
The strong prior information leads again to unbiased estimates for the error rates. Table 4.7 summarises the posteriors of the model coefficients for each log-linear model's specification. Table 4.8 shows the mean and standard deviation of the Mean Square Error computed for each sample. On average, the lowest MSE is registered for the model [13][23]; this is quite expected since the presence of another interaction term allows for more flexibility, and the model generally fits the data better.
Table 4.9 shows for each model how often the true population size has been included in different Highest Posterior Density intervals. As the interval shrinks, all models perform worse; however, the true model can almost always include the true value of N in the 80% HPD.
Finally, we verify how often the Bayes Factor computed using the method described in §4.5 favours the true model against the others; see 4.10. The rows do not always sum up to one; the computational capacity of `R` sometimes does not allow for the exact evaluation of the ratio, generally due to too small values of the normalising constant $\mathrm{m}_{m'}(\boldsymbol{y})$ that make the denominator of the Bayes Factor go to 0. Since we cannot observe the exact value of the BF, we do not include these cases in the table.
The true model is favoured at least in two-third of the cases against any model, even against the most competing [12][23].

Even without the high accuracy shown in the study of the previous section, there is evidence that the Bayes Factor computed with Chib and Jeliazkov (2001) method can recognise the true model.

Model	ϕ	β_1	β_2	β_3	β_{12}	β_{13}	β_{23}
[1][2][3]	7.273	1.083	0.868	0.112			
	(0.036)	(0.029)	(0.017)	(0.018)			
[12][3]	7.398	0.941	0.721	0.102	0.179		
	(0.052)	(0.053)	(0.058)	(0.019)	(0.061)		
[13][2]	6.738	1.669	0.918	0.826		-0.886	
	(0.098)	(0.097)	(0.020)	(0.099)		(0.101)	
[1][23]	7.346	1.072	0.782	0.002			0.148
	(0.035)	(0.029)	(0.026)	(0.030)			(0.035)
[12][13]	6.651	1.770	1.006	0.850	-0.107	-0.910	
	(0.109)	(0.111)	(0.061)	(0.098)	(0.065)	(0.100)	
[12][23]	7.556	0.863	0.537	-0.056	0.257		0.207
	(0.056)	(0.052)	(0.068)	(0.033)	(0.062)		(0.040)
[13][23]	6.754	1.665	0.902	0.802		-0.883	0.028
	(0.098)	(0.096)	(0.030)	(0.103)		(0.100)	(0.041)
[12][13][23]	6.646	1.773	1.011	0.854	-0.109	-0.911	-0.005
	(0.114)	(0.112)	(0.076)	(0.104)	(0.069)	(0.099)	(0.047)

Table 4.7: Estimates for $\boldsymbol{\theta}$. True model: [13][2].

Model	Mean	Sd	$\text{MSE}_{[13][2]} < \text{MSE}_M$
[1][2][3]	650148.354	39014.298	1
[12][3]	647463.983	38629.172	1
[13][2]	6046.777	1619.855	–
[1][23]	709191.809	42142.492	1
[12][13]	8115.541	1517.174	0.94
[12][23]	761324.722	44876.640	1
[13][23]	5720.108	1559.573	0.47
[12][13][23]	7381.883	1499.072	0.81

Table 4.8: Mean square errors. True model: [13][2].

Model	$N^* \in \text{HPD}_{95\%}$	$N^* \in \text{HPD}_{80\%}$	$N^* \in \text{HPD}_{50\%}$
[1][2][3]	1	0.82	0.01
[12][3]	0.62	0	0
[13][2]	1	0.99	0.08
[1][23]	0.98	0.05	0
[12][13]	1	0.73	0
[12][23]	0	0	0
[13][23]	1	0.99	0.09
[12][13][23]	1	0.65	0

Table 4.9: The true population size lies in the Highest Posterior Density interval, frequencies. True model: [13][2].

[13][2]	$(-\infty, 0.1)$	$[1, 3.2)$	$[3.2, 10)$	$[10, \infty)$
[1][2][3]	0.02	0	0	0.66
[12][3]	0.05	0	0	0.80
[13][2]	0	0	0	0.75
[1][23]	0.28	0	0.03	0.68
[12][13]	0.02	0	0	0.85
[12][23]	0.36	0.01	0.01	0.61
[13][23]	0.19	0	0.01	0.79

Table 4.10: Relative frequency of Bayes Factor interval values. True model: [13][2].

4.7 Discussion

In this chapter, we proposed an alternative model for population size estimation in the presence of out-of-scope units. Our proposal is an "alternative" to the models discussed in chapter 3 because it addresses a precise context for which the introduction of strong prior information is needed. The elicitation of error rates' prior distributions relying on the Pseudo Conditional Independence assumption is only one of the possible ways we may walk to include such information; the good performance of the model is independent of how we elicit the hyperparameters.

The use of Fisher's noncentral hypergeometric distribution in the posterior sampling process is a key point that deserves attention. In the context of official statistics, it is common to express information in relative terms; thus, allowing for the possibility to express prior beliefs about erroneous enumerations in the lists in terms of relative odds can be crucial.

Another aspect of this chapter we would like to stress is the model selection approach. The simulation studies showed how the method by Chib and Jeliazkov (2001) could perfectly fit this context, and it confirms the goodness of the model set. It can test many other models aiming at the population size estimation, and it results more adaptable than other approaches, such as the Reversible Jump used in Overstall and King (2014a).

Conclusions

Nowadays, official statistics faces a dichotomic situation that we may summarise in the following way. On the one hand, the high technological development makes available a large amount of data. This situation gives the official statistics many opportunities that often translate into complex challenges. Indeed, data integration from multiple sources is a crucial concept for all the National Statistics Institutes; it usually involves methodological issues, such as handling erroneous enumerations. On the other hand, and entirely in contrast with the "data full" context, some populations, or groups, are elusive, making it difficult to estimate their size.

Our work places itself in this context, giving a contribution to such methodological issues. We devoted the first part to estimating a heterogeneous population's size when a single list is available or we have multiple lists, but we lack unique identifiers. Thanks to their ability to extract information from one or a few data sources, the methods presented therein are particularly suitable when dealing with elusive populations, such as the recent graduates trying to enter the labour market, as in the case study presented in §2.4.
The second part of the work deals with the erroneous enumerations problem in the multisource context. After a critical review of the main and most recent literature about population size estimation, we propose an alternative model that addresses both overcoverage and undercoverage problems.

The model presented in chapter 2 leaves some open questions. Indeed, both the estimation methods suggested therein have pros and cons; one aspect that will deserve further research is the applicability of other (more efficient) ABC methods to enhance the implementation's speed. To ap-

proach the multivariate problem using component-wise ABC steps as proposed by Clarté et al. (2020) seems a feasible solution.
At the same time, we may enrich the model proposed in chapter 4 as well. Including extra-experimental information is a real need; here, we have mainly focused on including such information about the erroneous enumerations. More attention will be devoted to the target units' part, refining the log-linear coefficients' prior elicitation process.

Appendix A

FNCH distribution

Assume

$$\begin{aligned} X_1 &\sim \text{Binom}(M_1, \zeta_1) \\ X_2 &\sim \text{Binom}(M_2, \zeta_2) \end{aligned} \tag{A.1}$$

Then, conditional on the sum $X_1 + X_2 = n$, the probability mass function of X_1 will be:

$$\begin{aligned} P(X_1 = x_1 | X_1 + X_2 = n) &= \frac{P(X_1 = x_1 \cap X_1 + X_2 = n)}{P(X_1 + X_2 = n)} \\ &= \frac{P(X_1 = x_1) P(X_2 = n - x_1)}{P(X_1 + X_2 = n)} \\ &= \frac{\binom{M_1}{x_1}\binom{M_2}{n-x_1}\zeta_1^{x_1}(1-\zeta)^{M_1-x_1}\zeta_2^{n-x_1}(1-\zeta_2)^{M_2-(n-x_1)}}{\sum_{z_1=0}^{n}\binom{M_1}{z_1}\binom{M_2}{n-z_1}\zeta_1^{z_1}(1-\zeta)^{M_1-z_1}\zeta_2^{n-z_1}(1-\zeta_2)^{M_2-(n-z_1)}} \\ &= \frac{\binom{M_1}{x_1}\binom{M_2}{n-x_1}\left(\frac{\zeta_1}{1-\zeta_1}\right)^{x_1}(1-\zeta_1)^{M_1}\left(\frac{\zeta_2}{1-\zeta_2}\right)^{n-x_1}(1-\zeta_2)^{M_2}}{\sum_{z_1=0}^{n}\binom{M_1}{z_1}\binom{M_2}{n-z_1}\left(\frac{\zeta_1}{1-\zeta_1}\right)^{z_1}(1-\zeta_1)^{M_1}\left(\frac{\zeta_2}{1-\zeta_2}\right)^{n-z_1}(1-\zeta_2)^{M_2}} \\ &= \frac{(1-\zeta_1)^{M_1}(1-\zeta_2)^{M_2}\binom{M_1}{x_1}\binom{M_2}{n-x_1}\left(\frac{\zeta_1}{1-\zeta_1}\right)^{x_1}\left(\frac{\zeta_2}{1-\zeta_2}\right)^{n-x_1}}{(1-\zeta_1)^{M_1}(1-\zeta_2)^{M_2}\sum_{z_1=0}^{n}\binom{M_1}{z_1}\binom{M_2}{n-z_1}\left(\frac{\zeta_1}{1-\zeta_1}\right)^{z_1}\left(\frac{\zeta_2}{1-\zeta_2}\right)^{n-z_1}} \end{aligned}$$

Setting $w_1 = \dfrac{\zeta_1}{1-\zeta_1}$, $w_2 = \dfrac{\zeta_2}{1-\zeta_2}$;

$$P(X_1 = x_1 | X_1 + X_2 = n) = \frac{\binom{M_1}{x_1}\binom{M_2}{n-x_1} w_1^{x_1} w_2^{n-x_1}}{\sum\limits_{z_1=0}^{n} \binom{M_1}{z_1}\binom{M_2}{n-z_1} w_1^{z_1} w_2^{n-z_1}} \tag{A.2}$$

$$= \frac{w_2^n \binom{M_1}{x_1}\binom{M_2}{n-x_1} \left(\dfrac{w_1}{w_2}\right)^{x_1}}{w_2^n \sum\limits_{z_1=0}^{n} \binom{M_1}{z_1}\binom{M_2}{n-z_1} \left(\dfrac{w_1}{w_2}\right)^{z_1}} \quad .$$

Finally, indicating the odds ratio as $w = \dfrac{w_1}{w_2}$,

$$= \frac{\binom{M_1}{x_1}\binom{N-M_1}{n-x_1} w^{x_1}}{\sum\limits_{z_1=0}^{n} \binom{M_1}{z_1}\binom{N-M_1}{n-z_1} w^{z_1}} \quad .$$

Writing A.2 as

$$\frac{\prod\limits_{c=1,2} \binom{M_c}{x_c} w_c^{x_c}}{\sum\limits_{\boldsymbol{z}\in\mathcal{Z}} \prod\limits_{c=1,2} \binom{M_c}{z_c} w_c^{z_c}} \ , \tag{A.3}$$

$\mathcal{Z} = \{(x_1, x_2) \in \mathbb{Z}^2 : x_1 + x_2 = n\}$, the extension to the multivariate case is straightforward.

Appendix B

Estimating the proportion of sub-groups in a population

Whether $\frac{M_c}{x_c}, c = 1, ..., C$, increases for all c,

$$\text{FNCH}(\boldsymbol{x}|\boldsymbol{M}, n, \boldsymbol{w}) \rightarrow \text{FNCH}(\boldsymbol{x}|k \cdot \boldsymbol{M}, n, \boldsymbol{w}) \tag{B.1}$$

where $k \in \mathbb{Z}$ is any constant. Indeed, consider the probability mass function of FNCH:

$$P(\boldsymbol{X} = \boldsymbol{x}| \sum_{c=1}^{C} X_c = n) = \frac{\prod_{c=1}^{C} \binom{M_c}{x_c} {w_c}^{x_c}}{\sum_{\boldsymbol{z} \in \mathcal{Z}} \prod_{c=1}^{C} \binom{M_c}{z_c} {w_c}^{z_c}} = \frac{\prod_{c=1}^{C} \frac{M_c!}{(M_c - x_c)! x_c!} {w_c}^{x_c}}{\sum_{\boldsymbol{z} \in \mathcal{Z}} \prod_{c=1}^{C} \frac{M_c!}{(M_c - z_c)! z_c!} {w_c}^{z_c}} \tag{B.2}$$

For $M_c \gg x_c$, $M_c!$ and $(M_c - x_c)!$ cancel out.

Figure B.1a shows how the curves overlap almost exactly when $M_1 = 10n$, $M_2 = 20n$. As n increases, the overlap worsens (B.1b-B.1c); however, it is never completely lacking, not even for $M_1 = n$ (B.1d).

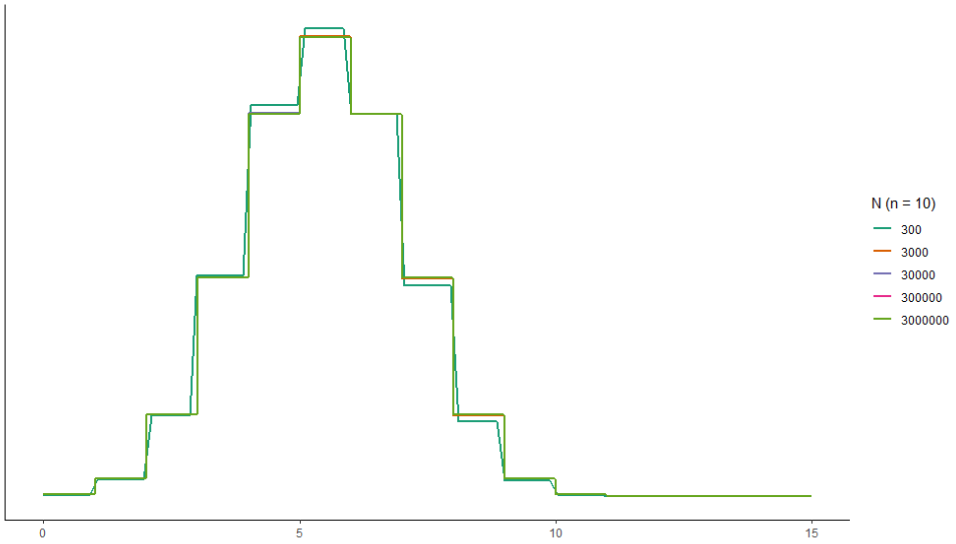
Figure B.2 shows how the curves become smoother and the overlap more

precise for higher N and n, under the condition of $M_c \gg n$, $\forall c$. Figure B.2c presents the only case where the overlap fails, i.e. $n > M_1$.

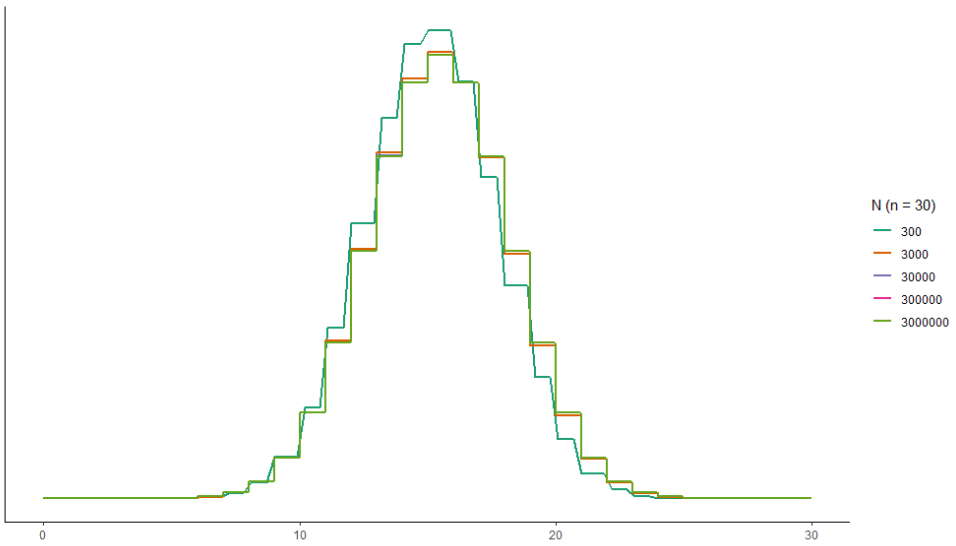
Hence, whether we do not include information on at least one M_c, we would be only able to estimate the subgroups' proportions within the population.
As an example, we simulate (x_1, x_2), i.e. the observed counts of two subgroups, from a population of size $N = 10000$. Table B.1 shows how to assume different sizes of N a priori affects minimally the estimate of the proportions of the two groups in the population.

$N \sim \text{Pois}(\lambda_N)$	M_1/N
$\lambda_N = 10000$	0.553
$\lambda_N = 100000$	0.584
$\lambda_N = 1000000$	0.581

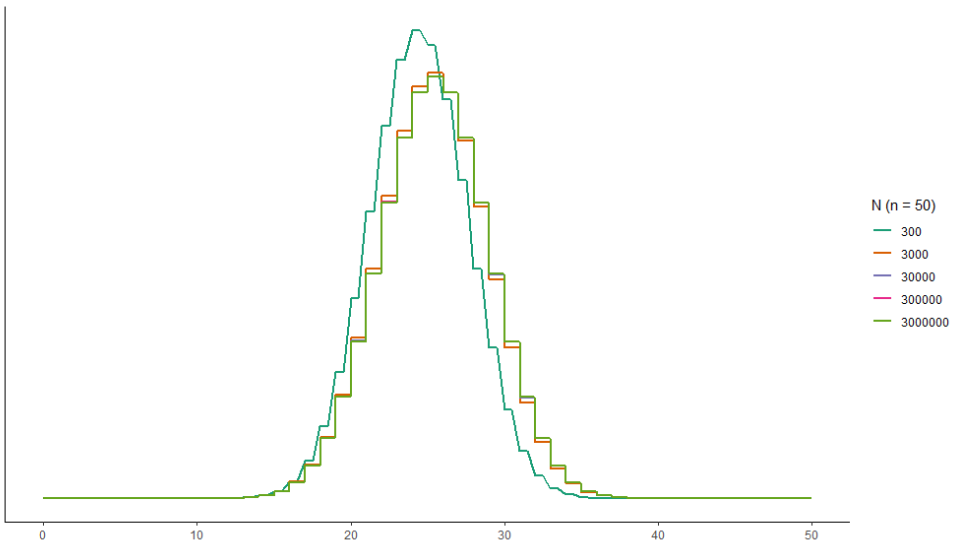
Table B.1: Estimated posterior mean of the proportion of the first subgroup within a population of size $N = 10000$. True value for $M_1/N = 0.567$.


N (n = 10)
300
3000
30000
300000
3000000
0
5
10
15

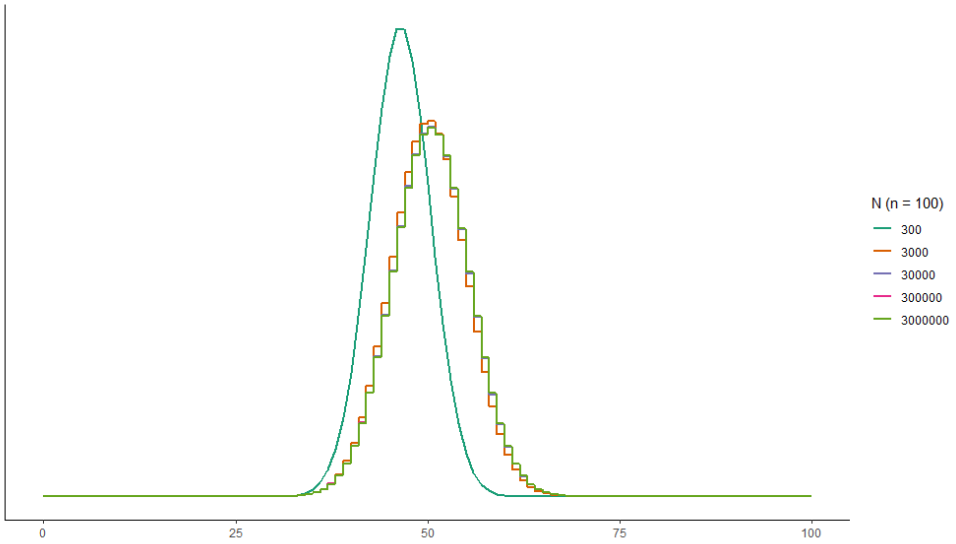
(a) $n = 10$


N (n = 30)
300
3000
30000
300000
3000000
0
10
20
30

(b) $n = 30$

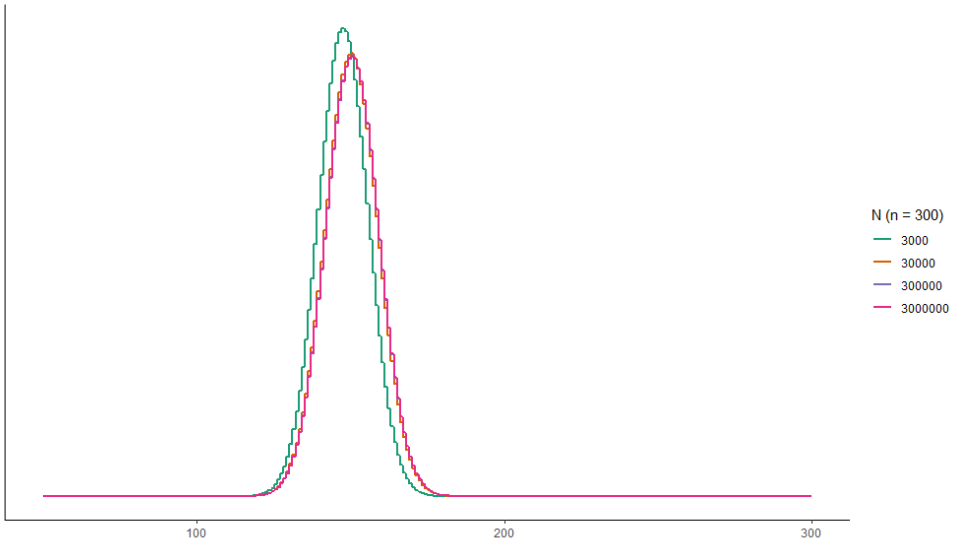


(c) $n = 50$

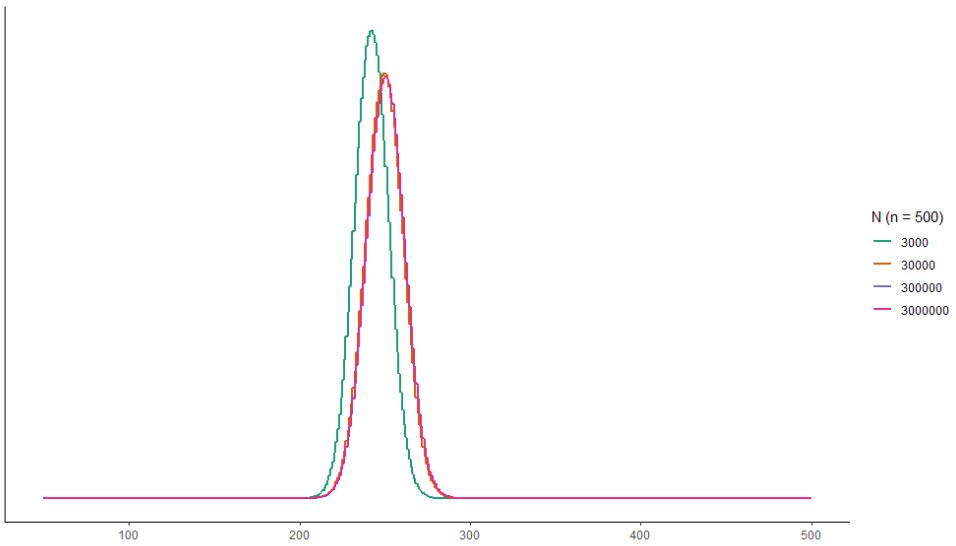


(d) $n = 100$

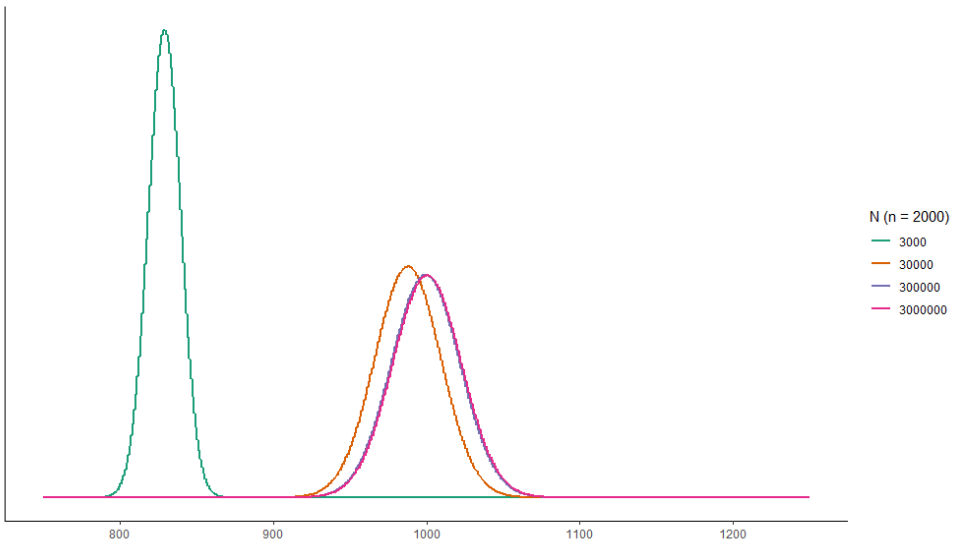
Figure B.1: Probability mass of a univariate FNCH with $M_2 = 2M_1$ with $M_1 + M_2 = 300, 3000, 30000, 300000, 3000000$, $w = 2$, and different n.


N (n = 300)
3000
30000
300000
3000000
100
200
300

(a) $n = 300$


N (n = 500)
3000
30000
300000
3000000
100
200
300
400
500

(b) $n = 500$



(c) $n = 20$

Figure B.2: Probability mass of a univariate FNCH with $M_2 = 2M_1$ with $M_1 + M_2 = 3000, 30000, 300000, 3000000$, $w = 2$, and different n.

Appendix C

From marginal to cross-sectional error rates

The marginal domain $\omega+$ is the union of a J-dimensional partition $(\{\nu_1\}, \{\nu_2\}, \ldots, \{\nu_J\})$ of the cross-sectional domains:

$$\omega+ = \cup_{j=1}^{J} \nu_j \tag{C.1}$$

e.g. in case of $K = 3$,

$$\{3+\} = \{123\} \cup \{13\} \cup \{23\} \cup \{3\} \tag{C.2}$$

Now, according to the definition of $\xi_{\omega+}$,

$$\begin{aligned} \xi_{\omega+} = P(i \notin U|\omega+) &= \frac{P(\omega+|i \notin U)P(i \notin U)}{P(\omega+)} \\ &= \frac{P(\cup_j\{\nu_j\}|i \notin U)P(i \notin U)}{P(\cup_j\{\nu_j\})} = \frac{\sum_j P(\{\nu_j\}|i \notin U)P(i \notin U)}{\sum_j P(\{\nu_j\})} \end{aligned} \tag{C.3}$$

where

$$P(\{\nu_j\}) = \frac{x_{\nu_j}}{\sum_{\omega \neq \mathbf{0}} x_\omega} \tag{C.4}$$

and

$$\begin{aligned} P(\{\nu_j\}|i \notin U) &= \frac{P(i \notin U|\{\nu_j\})P(\{\nu_j\})}{P(i \notin U)} \\ &= \xi_{\nu_j} \frac{x_{\nu_j}}{\sum_{\omega \neq \mathbf{0}} x_\omega} \frac{1}{P(i \notin U)} \end{aligned} \tag{C.5}$$

Hence

$$\xi_{\omega+} = \frac{\sum_j \xi_{\nu_j} \frac{x_{\nu_j}}{\sum_{\omega\neq\mathbf{0}} x_\omega} \frac{P(i \notin U)}{P(i \notin U)}}{\sum_j \frac{x_{\nu_j}}{\sum_{\omega\neq\mathbf{0}} x_\omega}} = \frac{\sum_j \xi_{\nu_j} \frac{x_{\nu_j}}{\sum_{\omega\neq\mathbf{0}} x_\omega}}{\sum_j \frac{x_{\nu_j}}{\sum_{\omega\neq\mathbf{0}} x_\omega}} = \frac{\sum_j \xi_{\nu_j} x_{\nu_j}}{\sum_j x_{\nu_j}} = \frac{\sum_j \xi_{\nu_j} x_{\nu_j}}{x_{\omega+}} \tag{C.6}$$

and

$$\xi_{\nu_j} = \frac{x_{\omega+}}{x_{\nu_j}} \xi_{\omega+} - \frac{\sum_{j^-} \xi_{\nu_{j^-}} x_{\nu_{j^-}}}{x_{\nu_j}} \tag{C.7}$$

where $\{j^-\}$ exclude j.
E.g., in the case of $K = 3$, $A = \{1\}$ and $B = \{2, 3\}$:

$$\begin{cases} \xi_2 = \frac{x_{2+}}{x_2} \xi_{2+} - \frac{x_{23}}{x_2} \xi_{23} \\ \xi_3 = \frac{x_{3+}}{x_3} \xi_{3+} - \frac{x_{23}}{x_3} \xi_{23} \\ \xi_{23} = \frac{x_{23+}}{x_{23}} \xi_{23+} \end{cases} \tag{C.8}$$

$$\begin{cases} \xi_2 &= \frac{x_{2+}}{x_2} \xi_{2+} - \frac{x_{23}}{x_2} \frac{x_{23+}}{x_{23}} \xi_{23+} \\ \xi_3 &= \frac{x_{3+}}{x_3} \xi_{3+} - \frac{x_{23}}{x_3} \frac{x_{23+}}{x_{23}} \xi_{23+} \end{cases} \tag{C.9}$$

For the PCI:

$$\begin{cases} \xi_2 &= \frac{x_{2+}}{x_2} \xi_{2+} - \frac{x_{23}}{x_2} \frac{x_{23+}}{x_{23}} \xi_{2+} \xi_{3+} \\ \xi_3 &= \frac{x_{3+}}{x_3} \xi_{3+} - \frac{x_{23}}{x_3} \frac{x_{23+}}{x_{23}} \xi_{2+} \xi_{3+} \end{cases} \tag{C.10}$$

hence the cross-classified error rates are fully defined in terms of the marginal ones.
Finally, taking the mean and variance of the expression above, we can elicitate the hyperparameters for ξ_{ν_j}.

Appendix D

Sampling the erroneous counts from FNCH

Each observed cross-classified count x_ω can be seen as a realisation of a random variable

$$X_\omega = Y_\omega + R_\omega \ , \tag{D.1}$$

where Y_ω represents the latent count of target units associated with the cell indexed by ω, and R_ω the relative out-of-scope units' count. We specify:

$$\begin{aligned} Y_\omega &\sim \text{Pois}(\lambda_\omega) \\ R_\omega &\sim \text{Pois}(\mu_\omega := x_\omega \xi_\omega) \end{aligned} \tag{D.2}$$

independently for all ω. Denote with y_0 the unobserved count, i.e. the number of units belonging to the target population captured by none of the lists.
We aim to estimate the joint posterior distribution of $\boldsymbol{\lambda}$[1], $\boldsymbol{\xi}, y_0$ and $\boldsymbol{r}$ (or, equivalently, the latent $\boldsymbol{y}$):

$$\pi(\boldsymbol{\lambda}, \boldsymbol{\xi}, y_0, \boldsymbol{r}|\boldsymbol{x}) \propto f(y_0, \boldsymbol{x}|\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{r})\pi(\boldsymbol{r}|\boldsymbol{x}, \boldsymbol{\xi})\pi(\boldsymbol{\lambda})\pi(\boldsymbol{\xi}) \tag{D.3}$$

In (D.3),

$$\pi(\boldsymbol{r}|\boldsymbol{x}, \boldsymbol{\xi}) = \prod_{\omega^O} \pi(r_\omega|x_\omega, \xi_\omega) \tag{D.4}$$

[1]for the sake of simplicity we directly use $\boldsymbol{\lambda}$ rather than $\boldsymbol{\theta}$

where ω^O indicates the cells affected by overcoverage. Each element in the product in (D.4) can be written as

$$\pi(r_\omega|x_\omega,\xi_\omega) = \frac{\pi(r_\omega, x_\omega|\xi_\omega)}{\pi(x_\omega|\xi_\omega)} \tag{D.5}$$

Since $x_\omega = y_\omega + r_\omega$ by definition, we write:

$$\begin{aligned}
\frac{\pi(r_\omega, x_\omega|\xi_\omega)}{\pi(x_\omega|\xi_\omega)} &= \frac{\pi(r_\omega|\xi_\omega)\pi(y_\omega = x_\omega - r_\omega|\xi_\omega)}{\pi(x_\omega|\xi_\omega)} \\
&= \frac{\dfrac{\mathrm{e}^{-x_\omega\xi_\omega}(x_\omega\xi_\omega)^{r_\omega}}{r_\omega!}\dfrac{\mathrm{e}^{-\lambda_\omega}(\lambda_\omega)^{x_\omega - r_\omega}}{(x_\omega - r_\omega)!}}{\dfrac{\mathrm{e}^{-(x_\omega\xi_\omega+\lambda_\omega)}(x_\omega\xi_\omega+\lambda_\omega)^{x_\omega}}{x_\omega!}} \qquad \text{(D.6)} \\
&= \frac{\dfrac{(x_\omega\xi_\omega)^{r_\omega}}{r_\omega!}\dfrac{\lambda_\omega^{x_\omega - r_\omega}}{(x_\omega - r_\omega)!}}{\dfrac{(x_\omega\xi_\omega+\lambda_\omega)^{x_\omega}}{x_\omega!}} \\
&= \frac{x_\omega!}{(x_\omega - r_\omega)!r_\omega!}\left(\frac{x_\omega\xi_\omega}{x_\omega\xi_\omega+\lambda_\omega}\right)^{r_\omega}\left(\frac{\lambda_\omega}{x_\omega\xi_\omega+\lambda_\omega}\right)^{x_\omega - r_\omega} \qquad \text{(D.7)}
\end{aligned}$$

which is the probability mass function of a $\text{Binom}\left(x_\omega, \dfrac{x_\omega\xi_\omega}{x_\omega\xi_\omega+\lambda_\omega}\right)$.

Hence,

$$\begin{aligned}
\pi(\boldsymbol{\lambda}, \boldsymbol{\xi}, y_0, \boldsymbol{r}|\boldsymbol{x}) \propto \frac{\mathrm{e}^{-\lambda_0}\lambda_0^{y_0}}{y_0!}\prod_\omega \frac{\mathrm{e}^{-(\lambda_\omega + x_\omega\xi_\omega)}(\lambda_\omega + x_\omega\xi_\omega)^{x_\omega}}{x_\omega!} \\
\times \prod_{\omega^O}\frac{x_\omega!}{(x_\omega - r_\omega)!r_\omega!}\left(\frac{x_\omega\xi_\omega}{x_\omega\xi_\omega+\lambda_\omega}\right)^{r_\omega}\left(\frac{\lambda_\omega}{x_\omega\xi_\omega+\lambda_\omega}\right)^{x_\omega - r_\omega}\pi(\boldsymbol{\lambda})\pi(\boldsymbol{\xi})
\end{aligned} \tag{D.8}$$

In the Gibbs sampler, we sample r_ω from its full conditional. To derive it, we must cancel out all terms in the posterior (D.8) that do not depend on r_ω. It will be proportional to:

$$\begin{aligned}
&\propto \frac{1}{(x_\omega - y_\omega)!r_\omega!}\left(\frac{x_\omega\xi_\omega}{x_\omega\xi_\omega+\lambda_\omega}\right)^{r_\omega}\left(\frac{\lambda_\omega}{x_\omega\xi_\omega+\lambda_\omega}\right)^{-r_\omega} \\
&\propto \frac{1}{(x_\omega - y_\omega)!r_\omega!}\left(\frac{(x_\omega\xi_\omega)(x_\omega\xi_\omega+\lambda_\omega)}{(x_\omega\xi_\omega+\lambda_\omega)\lambda_\omega}\right)^{r_\omega} \qquad \text{(D.9)} \\
&\propto \frac{1}{(x_\omega - y_\omega)!r_\omega!}\left(\frac{x_\omega\xi_\omega}{\lambda_\omega}\right)^{r_\omega}
\end{aligned}$$

The expression above is different from (D.7), thus using

$$\text{Binom}\left(x_\omega, \frac{x_\omega \xi_\omega}{x_\omega \xi_\omega + \lambda_\omega}\right)$$

to sample r^t would not be a proper choice.
The ratio in D.9 can be seen as the relative weight of the erroneous counts with respect to the in-target units in the cell indexed by ω. Therefore, we consider a univariate FNCH distribution of parameters M_1, $M_2 = N$, $n = x_\omega$ and $\boldsymbol{w} = (w_1, w_1)$, where

$$\begin{aligned} w_R &= \frac{x_\omega \xi_\omega / M}{1 - x_\omega \xi_\omega / M} = \frac{x_\omega \xi_\omega}{M - x_\omega \xi_\omega} \\ w_Y &= \frac{\lambda_\omega / N}{1 - \lambda_\omega / N} = \frac{\lambda_\omega}{N - \lambda_\omega} \end{aligned} \tag{D.10}$$

The probability mass is proportional to:

$$\text{FNCH}(r_\omega, y_\omega | M, N, x_\omega, w_\omega) \propto \frac{M!}{(M - r_\omega)! r_\omega!} \frac{N!}{(N - y_\omega)! y_\omega!} w_R^{r_\omega} w_Y^{y_\omega} \tag{D.11}$$

As $M \gg r_\omega$, the above becomes:

$$\frac{N!}{(N - y_\omega)! y_\omega! r_\omega!} w_R^{r_\omega} w_Y^{y_\omega} \tag{D.12}$$

and using interchangeably y_ω, $x_\omega - r_\omega$:

$$\begin{aligned} &= \frac{N!}{(N - y_\omega)!(x_\omega - r_\omega)! r_\omega!} \left(\frac{x_\omega \xi_\omega}{M - x_\omega \xi_\omega}\right)^{r_\omega} \left(\frac{\lambda_\omega}{N - \lambda_\omega}\right)^{x_\omega - r_\omega} = \\ &= \frac{N!}{(N - y_\omega)!(x_\omega - r_\omega)! r_\omega!} \left(\frac{\frac{x_\omega \xi_\omega}{M - x_\omega \xi_\omega}}{\frac{\lambda_\omega}{N - \lambda_\omega}}\right)^{r_\omega} \left(\frac{\lambda_\omega}{N - \lambda_\omega}\right)^{x_\omega} = \\ &= \frac{N!}{(N - y_\omega)!(x_\omega - r_\omega)! r_\omega!} \left(\frac{x_\omega \xi_\omega}{\lambda_\omega}\right)^{r_\omega} \left(\frac{N - \lambda_\omega}{M - x_\omega \xi_\omega}\right)^{x_\omega - y_\omega} \left(\frac{\lambda_\omega}{N - \lambda_\omega}\right)^{x_\omega} \end{aligned} \tag{D.13}$$

which is proportional to (D.9). Hence, introducing for convenience a fixed and arbitrarily large M_1, we can use FNCH distribution to sample r_ω at each iteration t.

Alternatively, we can simply divide and multiply (D.9) by $x_\omega^{r_\omega}$:

$$\begin{aligned} &\frac{1}{r_\omega!(x_\omega - r_\omega)!} \frac{x_\omega^{r_\omega}}{x_\omega^{r_\omega}} \left(x_\omega \xi_\omega\right)^{r_\omega} \lambda_\omega^{-r_\omega} \\ &= \frac{1}{r_\omega!(x_\omega - r_\omega)!} \left(\frac{x_\omega \xi_\omega}{x_\omega}\right)^{r_\omega} \left(\frac{\lambda_\omega}{x_\omega}\right)^{-r_\omega} \end{aligned} \tag{D.14}$$

By definition, $X_\omega = R_\omega + Y_\omega$, and $\mathbb{E}(Y_\omega) = \lambda_\omega$, $\mathbb{E}(R_\omega) = x_\omega \xi_\omega$. Hence, we can see (D.14) as an approximation of the kernel of a Binomial distribution of parameters x_ω and ξ_ω. Such a result gives us another option for sampling r_ω^t. In practise, the two sampling distributions give equivalent results.

Bibliography

A. Agresti. Simple capture-recapture models permitting unequal catchability and variable sampling effort, 1994.

P. Ball, W. Betts, F. Scheuren, J. Dudukovic, and J. Asher. Killings and refugee flow in Kosovo – March-June 1999. *Report to ICTY*, 2002.

V. Ballerini. Capture-recapture models for official statistics in presence of out-of-scope units: an overview. *Annali del Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza 2020*, pages 15–32, 2020.

R. D. H. Barrett, S. Laurent, R. Mallarino, S. P. Pfeifer, C. C. Y. Xu, M. Foll, K. Wakamatsu, J. S. Duke-Cohan, J. D. Jensen, and H. E. Hoekstra. Linking a mutation to survival in wild mice. *Science*, 363 (6426):499–504, 2019.

F. Bartolucci and A. Forcina. Analysis of capture-recapture data with a Rasch-type model allowing for conditional dependence and multidimensionality. *Biometrics*, 57(3):714–719, 2001.

F. Bartolucci and F. Pennoni. A class of latent Markov models for capture–recapture data allowing for time, heterogeneity, and behavior effects. *Biometrics*, 63(2):568–578, 2007.

F. Bartolucci, A. Mira, and L. Scaccia. Answering two biological questions with a latent class model via MCMC applied to capture-recapture data. In *Applied Bayesian statistical studies in biology and medicine*, pages 7–23. Springer, 2004.

S. Basu and N. Ebrahimi. Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, 88(1):269–279, 2001.

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis. Second Edition.* Springer series in statistics. Springer-Verlag, 1985.

P. P. Biemer. *Latent class analysis of survey error*, volume 571. John Wiley & Sons, 2011.

P. P. Biemer, G. G. Brown, D. H. Judson, and C. Wiesen. Triple system estimation with erroneous enumerations in the administrative records list, 2001a.

P. P. Biemer, H. Woltmann, D. Raglin, and J. Hill. Enumeration accuracy in a population census: an evaluation using latent class analysis. *Journal of Official Statistics*, 17(1):129, 2001b.

P. P. Biemer, G. G. Brown, and D. H. Judson. Latent class models for evaluating the accuracy of census counts, 2004.

Y. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis - Theory and Applications.* Interdisciplinary Statistics Series. Cambridge, Massachusetts: The MIT Press, 1975.

G. G. Brown, P. P. Biemer, and D. H. Judson. Estimating erroneous enumeration in the US decennial census using four lists, 2004.

P. Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media, 2013.

F. Busetti and P. Cova. L'impatto macroeconomico della crisi del debito sovrano: Un'analisi controfattuale per l'economia italiana (the macroeconomic impact of the sovereign debt crisis: A counterfactual analysis for the Italian economy). *Bank of Italy Occasional Paper*, (201), 2013.

D. Böhning, P. G. M. van der Heijden, and J. Bunge, editors. *Capture-Recapture Methods for the Social and Medical Sciences.* Interdisciplinary Statistics Series. Chapman and Hall/CRC, 2018.

H. L. Center. The Kosovo memory book project 1998-2000: List of killed, missing and disappeared 1998-2000, 2014. URL `http://www.kosovskaknjigapamcenja.org/db/kkp_en/index.html`.

J. Chesson. A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation. *Journal of Applied Probability*, pages 795–797, 1976.

V. Chiariello and T. Tuoto. Estimation of criminal populations using administrative registers in the presence of linkage errors. 2018. URL `https://www.istat.it/it/files//2018/11/Chiariello_original-paper.pdf`.

S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96 (453):270–281, 2001.

G. Clarté, C. P. Robert, R. Ryder, and J. Stoehr. Component-wise approximate bayesian computation via gibbs-like steps. *Biometrika*, asaa090, 2020.

R. M. Cormack. Interval estimation for mark-recapture studies of closed population. *Biometrics*, 48(2):567–576, 1992.

J. N. Darroch, S. E. Fienberg, G. F. Glonek, and B. W. Junker. A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88(423):1137–1148, 1993.

A. P. Dawid and S. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 21: 1272–1317, 1993.

P. Dellaportas and J. J. Forster. Markov Chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 86(3):615–633, 1999.

D. Di Cecco. *Estimating Population Size in Multiple Record System with Uncertainty of State Identification*, chapter 8. Chapman and Hall/CRC, 2019.

D. Di Cecco, M. Di Zio, and B. Liseo. Bayesian latent class models for capture–recapture in the presence of missing data. *Biometrical Journal*, 2020.

L. Di Consiglio and T. Tuoto. Population size estimation and linkage errors: the multiple lists case. *Journal of Official Statistics*, 34(4): 889—-908, 2018.

L. Di Consiglio, T. Tuoto, and L.-C. Zhang. *Capture-Recapture Methods in the Presence of Linkage Errors*, chapter 2. Chapman and Hall/CRC, 2019.

S. E. Fienberg. The analysis of multidimensional contingency tables. *Ecology*, 51(3):419–433, 1970.

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

S. E. Fienberg, M. S. Johnson, and B. W. Junker. Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162(3):383–405, 1999.

D. Filipponi, U. Guarnera, and R. Varriale. Integration of administrative sources and survey data through Hidden Markov models for the production of labour statistics. 2017. URL `https://www.istat.it/it/files/2018/11/Filipponi{_}original-paper.pdf`.

R. A. Fisher. The logic of inductive inference. *Journal of the royal statistical society*, 98(1):39–82, 1935.

A. Fog. Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. *Communications in Statistics—Simulation and Computation*, 37(2):241–257, 2008a.

A. Fog. Calculation methods for Wallenius' noncentral hypergeometric distribution. *Communications in Statistics—Simulation and Computation®*, 37(2):258–273, 2008b.

J. J. Forster. Bayesian inference for Poisson and Multinomial log-linear models, 2010. Working Paper M09/11 University of Southampton, Southampton Statistical Sciences Research Institute.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, editors. *Bayesian Data Analysis, 2nd ed.* Chapman Hall: Boca Raton, 2004.

O. Gimenez, E. Cam, and J.-M. Gaillard. Individual heterogeneity and capture–recapture models: what, why and how? *Oikos*, 127(5):664–686, 2018.

C. Grazian, F. Leisen, and B. Liseo. Modelling preference data with the Wallenius distribution. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):541–558, 2019.

P. J. Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

W. L. Harkness. Properties of the extended hypergeometric distribution. *The Annals of Mathematical Statistics*, 36(3):938–945, 1965.

Istat. Censimento permanente della popolazione e delle abitazioni, 2018. URL `https://www.istat.it/it/censimenti-permanenti/popolazione-e-abitazioni`.

J. Johndrow, K. Lum, and P. Ball. `dga`: capture-recapture estimation using Bayesian model averaging, 2015. URL `https://CRAN.R-project.org/package=dga`.

J. E. Johndrow, K. Lum, and D. Manrique-Vallier. Estimating the observable population size from biased samples: a new approach to population estimation with capture heterogeneity. *arXiv preprint arXiv:1606.02235*, 2016.

D. H. Johnson, K. P. Burnham, and J. D. Nichols. The role of heterogeneity in animal population dynamics, 1986.

P. Kaskasamkul and D. Böhning. *Population size estimation for one-inflated count data based upon the geometric distribution*, chapter 14. Chapman and Hall/CRC, 2018.

R. E. Kass and A. E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

R. King and S. P. Brooks. On the Bayesian analysis of population size. *Biometrika*, 88(2):317–336, 2001.

F. C. Lincoln. *Calculating waterfowl abundance on the basis of banding returns*. Number 118. US Department of Agriculture, 1930.

M. A. Lodato, R. E. Rodin, C. L. Bohrson, M. E. Coulter, A. R. Barton, M. Kwon, M. A. Sherman, C. M. Vitzthum, L. J. Luquette, C. N. Yandava, et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*, 359(6375):555–559, 2018.

D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89:1535–1546, 1994.

D. Madigan and J. C. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63(2):215–232, 1995.

D. Madigan and J. C. York. Bayesian methods for estimation of the size of a closed population. *Biometrika*, 84(1):19–31, 1997.

D. Madigan, A. E. Raftery, J. C. York, and J. M. Bradshaw. *Strategies for Graphical Model Selection*, pages 91–100. New York: Springer-Verlag, 1994.

D. Manrique-Vallier. Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, 72(4):1246–1254, 2016.

D. Manrique-Vallier, P. Ball, and M. Sadinle. Capture-recapture for casualty estimation and beyond: Recent advances and research directions. 2019.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

D. L. Otis, K. P. Burnham, G. C. White, and D. R. Anderson. Statistical inference from capture data on closed animal populations. *Wildlife monographs*, (62):3–135, 1978.

A. M. Overstall and R. King. `conting`: An `R` package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software*, 58(7):1–27, 2014a.

A. M. Overstall and R. King. A default prior distribution for contingency tables with dependent factor levels. *Statistical Methodology*, 16:90–99, 2014b.

A. M. Overstall, R. King, S. M. Bird, S. J. Hutchinson, and G. Hay. Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. *Statistics in Medicine*, 33(9):1564–1579, 2014.

C. G. J. Petersen. The yearly immigration of young plaice into the Limfjord from the German sea. *Report of the Danish Biological Station*, 6:5–84, 1985.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution*, 16(12): 1791–1798, 1999.

G. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Interdisciplinary Statistics Series. 1960.

R. R. Regal and E. B. Hook. The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine*, 10: 717–721, 1991.

C. P. Robert and G. Casella, editors. *Monte Carlo Statistical Methods*. Springer Text in Statistics. Springer, 2004.

D. Sabanés Bové and L. Held. Hyper-g priors for Generalized Linear Models. *Bayesian Analysis*, 6(3):387–410, 2011.

L. Sanathanan. Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, 43(1):142–152, 1972.

S. Sisson, Y. Fan, and M. Beaumont. Overview of ABC. *Handbook of approximate Bayesian computation*, pages 3–54, 2018.

A. Tancredi and B. Liseo. A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5(2B):1553–1585, 2011.

A. Tancredi, R. Steorts, and B. Liseo. A unified framework for deduplication and population size estimation. *Bayesian Analysis*, TBA (TBA):1–26, 2019.

K. T. Wallenius. Biased sampling; the noncentral hypergeometric probability distribution. Technical report, Stanford University CA Applied Mathematics and Statistics Labs, 1963.

L.-C. Zhang. On modelling register coverage errors. *Journal of Official Statistics*, 31(3):381–396, 2015.

L.-C. Zhang. *Log-linear Models of Erroneous List Data*, chapter 9. Chapman and Hall/CRC, 2019.

L.-C. Zhang and R. L. Chambers, editors. *Analysis of Integrated Data*. Statistics in the Social and Behavioral Sciences Series. Chapman and Hall/CRC, 2019.