UNIVERSITY OF SOUTHAMPTON

On capture-recapture with validation information

by

Carla Azevedo

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

FACULTY OF SOCIAL SCIENCES

May 2019

UNIVERSITY OF SOUTHAMPTON

Abstract

FACULTY OF SOCIAL SCIENCES

Doctor of Philosophy

On capture-recapture with validation information

by Carla Azevedo

This work shows how capture-recapture modelling can be performed in the presence of a validation set, a sample that includes all the counts, in particular, zero counts which are not observed in typical capture-recapture settings. We start with the simple homogeneous case for estimation of the Binomial and the Poisson distribution using the EM algorithm. A flexible non-parametric mixture model approach allowing for heterogeneity of the data by means of a nested EM algorithm using validation information was used to allow for more components in the target population. The estimate for the total population size can be obtained by jointly fitting a zero-truncated distribution to the truncated data and an untruncated distribution of the same class to the untruncated data by means of the EM algorithm. Simulation studies demonstrated the value of including validation information into the modelling to estimate the total size of the population. This was also done following a ratio regression approach which is explained in detail along this work.

For illustration of the major ideas of these applications, these methods were applied to public health problem scenarios related with Salmonella infection in poultry, Bowel Cancer and transmittable diseases: Brucellosis and Syphilis. A community study on the number of Heroin users in Bangkok was also considered. The main goal of the present study is to adjust the undercount of disease/drug use occurrence in the UK farms/people during a period of study. Three models were considered for the last approach which seemed relevant for the data situation. However, situations of zero-inflated counts were also debated in the case the first ratio is particularly lower than the other ratios indicating potential presence of zero-inflation. This work also introduces simulation studies which help to understand the role of the validation sample in the estimation process showing that we can rely more confidently on the estimate for the population size using that additional information.

Declaration of Authorship

I, Carla Filipa Sampaio Azevedo, declare that the thesis entitled "On capture-recapture with validation information" and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published before submission in [23].

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Dankmar Böhning, for his guidance and support throughout. It was a great pleasure and honour to be his student. I owe him the possibility of studying in the University of Southampton which is undoubtedly the best achievement of my career. When I arrived in this city and University, young and inexperienced, I found not only a supervisor but also a friend who always gave me the best advice, patience and help for everything I needed.

This work was supported by a studentship funding from the University of Southampton vice-chancellor scholarship and the Animal and Plant Health Agency (APHA), UK. Here, I acknowledge my second supervisor Dr. Mark Arnold from the APHA for the financial support, availability, topic discussions and help provided during this period.

The support and continued encouragement of my family has been invaluable. A sincere and warm thank you to my close family, in special to my mother for all the sacrifices she passed in her life to give me the best education she could and my younger brother who is my confidant and best friend. A big and special thank you to my godmother who is always there for me and understands me better than anyone else.

I am also very delighted to get to know such amazing people during my studies. A special thank you to Mehmet Cihan who was always by my side during this period. Your support and your company was the best I could ever ask for. I also would like to thank all my friends who believed in me and my capabilities and who I could always count on. You know who you are.

Without you all, I would not be the person I am today and this adventure would never be the same. I will always be endlessly grateful.

Contents

List of Figures

List of Tables

Dedicated to my mother...

Chapter 1

Introduction

1.1 Introduction

Capture-recapture methods are an important and very useful tool to estimate the global size of a target population of interest when it cannot be completely observed. Estimating the size N of a specific population is of crucial importance in many areas such as social, biological and medical sciences. For example, for ecological purposes it is relevant to estimate the size of a wildlife population. In medicine, it is essential to estimate the quantity of people with a specific disease when a screening test is not totally accurate and we may often get false negatives.

Frequently, in real applications, due to a deficient identification-registration mechanism, only a portion of the population is observed - the positive counts and we might need to predict the number of unobserved identifications. Therefore, our interest is to determine the size N of a potentially elusive population in which zero-counts are missing.

Let us assume that the members of the population are identified at m observational occasions where m is considered fixed in this work. For each member i, the count X_i of identifications for a generic unit returns a count in $0, 1, ..., m$ and $i = 1, ..., N$. It is assumed that X_i is available if unit i has been identified for at least one occasion that is if $X_i > 0$. In that case X_i is observed; let $X_1, ..., X_n$ denote the observed counts with n representing the total number of recorded individuals. We assume w.l.o.g. that $X_{n+1} = ... = X_N = 0$. Hence, units $n + 1$ to N remain unobserved.

Very common examples are screening tests applied to human populations to detect a specific disease in its early stage when it is easier to treat and cure. Just one application of a test might have low sensitivity and as we know, any screening test cannot be 100% accurate. Thus, usually, people with a negative test result are not further assessed, so it remains unknown which disease status they actually have. In other words, we want to investigate how many false negatives we have adopting the described procedure.

Using the same notation as above, let us assume for example we are analysing a clinical disease whose status can be measured in m levels of severity, where the levels x denote the number of times the screening test is positive (number of captures):

TABLE 1.1: Frequency of each status of a certain disease.

x	0	1	2	...	m
f_x	?	f_1	f_2	...	f_m

Here, f_x represents the number of individuals captured exactly x times during the screening test. If the test is negative at all m times, the true status of the person is unknown. The true status of a person is always unknown until a final and conclusive test is performed. Hence, we intend to estimate f_0 using the zero-truncated distribution to estimate the total size of the diseased population N.

This is just a simple generic example of an application of capture-recapture methodology. However, this methodology can also be applied in other areas such as epidemiology, ecology or social sciences. In particular, it is a popular analysis to estimate animal abundance in the ecological field. Other good examples are applications in computer engineering to estimate the number of errors in a computer software, to estimate the number of scrapie infected sheep population in Great Britain and to investigate the number of illegal immigrants living in the Netherlands coming from some Middle East countries. See Böhning [18], [16] and Van der Heijden [87] for some of these applications.

Let us now see a practical example: from 1984 onwards, at the St Vincents Hospital in Australia, about 50000 subjects were screened for bowel cancer which is a medical procedure to detect blood in the bowel motion. This screening procedure was based on a sequence of binary diagnostic tests, self-administered on 6 successive days. On each of these 6 occasions, the absence or the presence of blood in faeces was recorded. Post-verification of the results of the tests was done by physical examination, sigmoidoscopy

and colonoscopy, performed only if at least one of the six test was positive. People with all six tests negative were not further assessed.

The frequency distribution of the number of positive tests is reported in the table bellow:

TABLE 1.2: Bowel cancer data positive sample.

x	0	1	2	3	4	5	6	Total
f_x	?	37	22	25	29	34	45	192

We can verify from the positive sample for the Bowel Cancer data that there were 37 individuals with one test positive, 22 individuals with two tests positive and so on.

In capture-recapture applications, we can deal with closed or open population models. A closed population is a population kept constant (at least approximately) during the study period, without any births, deaths or migration. It is certain that in the major part of the cases in the real life, most part of the populations are open. However, if the study period is short or we are studying small areas, then the assumption of considering a closed population will have a minor impact.

In the case of the Bowel Cancer data example and similar cases, each capture happens in a fixed period of time and it is assumed that each individual has equal probability of being captured during the study period.

Let f_x be the frequency of units with count $X = x$. The associated population density function can be described by a probability density function $p_x(\theta)$ which denotes the probability of exactly x identifications for a generic unit where $p_x(\theta) \geq 0$ and $\sum_{x=0}^{\infty} p_x(\theta) = 1$.

In general, the Poisson or the Binomial distributions are frequently used to model the observed counts. However, under these homogeneous models, it is assumed that the individuals of the population have the same probability to be captured which in unlikely to happen in real life situations. This leads very often to an underestimation of the true population size.

In practice, a population is naturally formed by different individuals/sub-populations. Thus, unobserved information translated in unobserved heterogeneity should be taken into account to model the data. Failure in doing this may lead to a serious underestimation of the population size. Heterogeneity is directly associated with the over-dispersion of the data, occurring when the variance predicted by the model is smaller than the variance of the data itself.

Situations of heterogeneity in the population can be detected by means of the ratio plot which works like a diagnostic device for the presence of a particular distribution [16]. We can then extend this theory to a regression approach which will consider the neighbour ratios of frequency counts and fit a proper model to the data. Finally, we use the model to derive an estimate for the frequency of hidden counts, f_0, projecting the model backwards. This approach is explained in detail in Chapetr 6.

Another approach to allow for heterogeneity is to use mixture models. Mixture models allow a flexible approach in modelling heterogeneity. The estimate of N can be achieved by means of the EM algorithm by fitting jointly a zero-truncated distribution and an untruncated distribution of the same class to the truncated and untruncated data respectively. We consider a flexible non-parametric mixture model approach allowing heterogeneity of in the data by means of a nested EM algorithm using a secondary sample also called validation information.

Eventually, another sub-sample of the target population can be available. In this secondary sample, usually smaller in size, we do observe zero counts which means that there are no hidden cases. This sample is called validation sample.

Still considering the example above, let us imagine that another sample of people was chosen and assessed to repeat the same test. The results are shown in Table 1.3.

TABLE 1.3: Frequency of each status of a certain disease.

y	0	1	2	...	m
g_y	g_0	g_1	g_2	...	g_m

The structure of the validation sample is similar to the structure of the positive sample. However, it is important to emphasize that here all the counts are observed and, in particular, we have information on g_0 which is unknown in the positive sample. Again, g_y is the frequency of counts exactly equal to y.

In the case of the Bowel Cancer screening test, a sample of 122 patients with confirmed cancer status were screened a second time using the same screening procedure. The corresponding frequency distribution is shown in following table and it represents the validation sample for this study:

TABLE 1.4: Bowel cancer data validation sample.

x	0	1	2	3	4	5	6	Total
f_x	22	8	12	16	21	12	31	122

The first introduction of capture-recapture modelling using validation information can be found in Böhning [16] and it was mentioned as an extension of a ratio regression approach which will be discussed in detail in Chapter 6.

We used simulation studies to evaluate the performance of validation information in the modelling. We conclude that the use of a validation sample not only substantially increases the estimation precision but, also, it reduces the bias significantly.

We are interested in applying this theory to public health problem scenarios which will be introduced in Chapter 2.

1.2 Basic Assumptions

When the total size of a population is difficult to achieve, capture-recapture methods are frequently applied to get an estimation of the unknown population size. As mentioned, a deficient identification system leads to a deficient count of the population when each individual is repeatedly sampled. It is important to highlight the set of assumptions which validate this capture-recapture study:

- during the observational period, the population of interest is closed, which means that there is no changes in its size during the study, i.e. the total number of the population is constant;
- the target population is well-defined;
- the count of identifications is available if each member has been identified at least once. All the other counts (non-recorded) remain unobserved;
- all the members of the population are independent of each other;
- all the observational occasions are independent of each other.

1.3 Objectives of the study

1.3.1 Main objective

This study focuses on the development of methodology to include validation information into capture-recapture modelling to increase the accuracy and efficiency of the final estimate for the total population size.

1.3.2 Secondary objectives

To achieve the main goal of this study, we define next the secondary targets that need to be achieved.

- To explore the limitations of homogeneous zero-truncated Binomial and Poisson modelling using just the positive sample and the positive and validation sample.
- To investigate the Binomial ratio plot in order to examine the occurrence of heterogeneity in the data.
- To incorporate validation information into ratio regression modelling.
- To extend the theory of mixture models to allow for two or more components of a Binomial mixture model including validation information through the EM algorithm.
- To investigate the advantage of having a validation sample in estimating the total population size through simulation studies and in the case of zero-inflated data.

1.4 Thesis outline

The thesis is composed by seven chapters. The first chapter is the Introduction and it establishes the context and importance of the topic. It is also in this chapter that we define the research objectives, state the basic assumptions made throughout the thesis and indicate the outline of this work. Finally, we end this chapter listing the notation that will be used in the following chapters.

Chapter 2 exploits what a validation sample is, its role in a capture-recapture study and how it can be obtained. It also describes the data sets that will be used to illustrate the practical use of the methods of the thesis. There are five different data sets: Salmonella data, Bowel cancer data, Syphilis in Izmir - Turkey data, Brucellosis in Izmir - Turkey data and, finally, Heroin users in Bangkok - Thailand data. The context of the data collection and the purpose of the study is presented in this chapter.

In Chapter 3, we give a review of the relevant literature on the topic explaining the background and key terms that are going to be used in the remaining chapters. Thus, we will review some examples where we can use capture-recapture methods and briefly introduce some insights on the methods used in the thesis (Chapters 4, 5 and 6).

Chapter 4 starts with the estimation of the parameters for homogeneous Binomial and Poisson modelling, which is done using maximum likelihood through the EM algorithm. Evaluation of model performance under homogeneity is carried out using only the positive sample and both samples. In order to model the heterogeneity of the data, a Binomial finite mixture using validation information is presented in Chapter 5. Insights are given about the link between the conditional and the unconditional likelihood estimation and we explore the occurrence of a profile likelihood as a natural path leaded by the unconditional likelihood. Once again, simulation scenarios are set to compare the results obtained using validation information. Zero-inflated data is also explored at the end of the chapter. The ratio plot is explored and this graphical method and results are applied to the datasets from Chapter 2.

In Chapter 6, ratio regression is proposed as an alternative to model data heterogeneity. Several models are presented as an extension of this modelling technique in order to include the validation sample. Simulation studies, which include zero-inflated data, closes this chapter. Those studies provide a practical guidance to model choice based on selection criteria. This chapter ends by laying down the main findings of the thesis and a full discussion exploring, critically, the analysis in Chapters 4, 5 and 6.

Chapter 7 is the final chapter and it draws upon the entire thesis, tying up a summary of the main conclusions and the next steps for future research.

1.5 Notation

In this short table, we present the essential notation used throughout the thesis.

Simbol	**Definition**
N	Total size of a population
m	Maximum number of recaptures in the population
X_i	Random variable representing the number of counts of identifications for unit i during the study period
n	Number of observed identifications
f_x	Frequency of individuals captured exactly x times out of m
p_x	Probability of exactly x identifications
q	Mixing density function
L^C	Conditional Likelihood
L^U	Unconditional Likelihood
N_1	Total size of the validation sample

Chapter 2

Validation information in capture-recapture studies

2.1 Validation information in capture-recapture

In traditional capture-recapture studies, we have access only to a (truncated at zero) positive sample. Therefore, it is assumed we are performing well when we model the unobserved part of the population, i.e., the information of this (incomplete) sample is enough to achieve an estimate of the total number of zero-counts that is close to the true value of zero identifications.

Sometimes, another sub-sample of the target population is available. This further sample is referred to as the validation sample. As already mentioned, zero counts are observed in this further set. The structure of the validation sample is similar to the structure of the positive sample, even if the first is normally smaller in size. Since we have information on g_0 while the frequency of zero counts is unknown in the positive sample, one can use this information to provide a better estimate for the unobserved counts/population size.

The first introduction to capture-recapture modelling with validation information can be found in Böhning [16] and it can be considered as an extension to the ratio regression approach which will be detailed in Chapter 6.

A validation sample can occur in a very natural way in a capture-recapture study. For example, a situation where a screening test was performed and the repeated tests

positive were counted. Right after this, the repeated tests are performed again for all those individuals which were confirmed positive. This procedure can be generalized to a sampling plan. This is the case of the Salmonella, the Syphilis and the Brucellosis datasets presented below. The Bowel Cancer data was already introduced in Chapter 1 is also an example and it is explored here in detail.

Another situation where a validation sample might occur is in a case with two sources where at least one source is a count and the second source is binary. By conditioning on the second source, a validation sample is received. This can be checked for the Heroin users case study introduced in this chapter.

2.2 Datasets

We will introduce some data which will be used to illustrate the theory in the next chapters. These are composed by a positive sample and a validation sample and, therefore, are appropriated to investigate the proposed approach for population size estimation. It can be seen in the following datasets how a validation sample was achieved for each case and even how to construct one from available data (see Heroin users dataset below).

2.2.1 Salmonella data

This project was developed as a joint work with the Animal and Plant Health Agency (APHA) in the UK and entails data related to Salmonella in commercial egg-laying flocks.

Human salmonellosis is a major public health concern in Europe, and in particular in the UK, with the majority of cases in recent years being caused by Salmonella strains Salmonella Enteritidis and Salmonella Typhimurium. The most common source of infection is thought to be through the consumption of contaminated eggs produced by infected laying hens, see Gillespie *et al.* [44] and Arnold *et al.* [7].

To assess the current prevalence of infected commercial egg-laying flocks, a European Union wide baseline survey of Salmonella infection was carried out between October 2004 and September 2005. The results of that survey were used as a basis for setting flock prevalence reduction targets for Salmonella national control programmes in each

member state of the European Union. The target was set to 10% reduction per annum for the UK and each member state in the prevalence of Salmonella; for details see Arnold *et al.* [10]. As a part of the baseline survey in the UK, a randomized sample of 454 commercial layer flock holdings was tested for Salmonella.

It is known there are a considerable number of unreported cases in surveys because Salmonella infection in poultry is not associated with clinical signs. To give an instance, a hen with Salmonella infection can lay non-contaminated eggs during large periods of time and eventually lay a contaminated one, resulting in a low probability that an apparently healthy chicken can lay contaminated eggs, see Arnold *et al.* [8].

In order to be able to monitor the progress of control measures for Salmonella, showing that, in fact, there is a reduction over time in the UK egg-laying farms, it is important to be able to obtain an accurate estimate of the initial prevalence at the time of the EU baseline survey. Therefore, it is important to appropriately adjust the potential under-count of disease occurrence. The main goal of the study is to provide an estimate, as accurate as possible, of the number of undetected cases, i.e., the number of farms which had Salmonella infected poultry but for which the result in the survey was negative.

In total, 454 holdings were sampled in the survey. From those, 53 tested positive for Salmonella in one or more samples of the survey using a method we will denote as the EU baseline survey method. Briefly, this consists in sampling 5 faeces samples, each composing a representative mix of litter from 1/5th of the poultry house, and 2 dust samples collected from around the poultry house, which would then be cultured for Salmonella. The EU baseline survey therefore consists of a total of 7 tests, so each farm could have 0,1,...,7 positives as Table 2.1 shows:

TABLE 2.1: Salmonella data positive sample.

x	0	1	2	3	4	5	6	7	Total
f_x	?	17	9	5	6	5	5	6	53

Table 2.1 shows the frequency distribution of tested farms by the number of positive tests out of the 7 Salmonella detection tests. We have that 17 farms had one positive test while 9 farms had two and so on.

The EU baseline survey data reported a prevalence of 11.7% for Salmonella. The sampling method used in the survey is known to be not 100% sensitive since it was developed

to be a cost-effective method [10]. In order to get comparable results, this method was implemented by all the member states of the EU. After analysing the data using Bayesian methods, Arnold *et al.* [10], indicates a prevalence of 18% (95% credibility interval 12-25%) for holdings infected with Salmonella which is much higher than the prevalence rate reported in the survey.

Biosecurity and hygiene practises are designed to prevent the spread of the Salmonella infection and they have been improved on commercial farms since the peak of human salmonellosis in the mid-1990s, e.g. through vaccination programmes, see Snow *et al.* [80]. It is fair to say that these practises are not implemented in all the farms exactly the same way, so that the probability that a test is positive (or negative) for Salmonella might vary across farms. Therefore, farms in which biosecurity and hygiene proceedings are more effective and taken more carefully are more likely to be successful in reducing the infection and less likely to have a positive test recorded. This aspect can be translated into heterogeneity among the farms.

The same method used in the survey was repeated in 21 out of 53 of the infected farms which provided a validation sample reported in Table 2.2. In fact, two more methods were applied to these 21 farms: an APHA in-house method that involved collecting 10 dust and 10 faecal samples from around the poultry house and another method, the "National Control Programme method", that involved single samples of pooled faeces and dust, each representing material from all of the poultry house. However, these two methods could not be applied to all the 21 sampled farms, so our analysis considers only the results from the EU method. A detailed study on the results and power of detection for each method can be found in Arnold *et al.* [9].

TABLE 2.2: Salmonella data validation sample.

y	0	1	2	3	4	5	6	7	Total
g_y	3	1	3	2	3	3	4	2	21

Again, it is important to highlight that we know $g_0 = 3$. It means that the test failed in 3 out of the 21 farms where Salmonella infection was known to be present which allows to induce a low sensitivity of the test, when this refers to the probability of a positive test given that the farm is infected.

2.2.2 Bowel cancer data

Screening for bowel cancer in human populations is a medical examination used to detect early cases of the disease and treat people timely since this disease can develop without early symptoms and grow on the inside wall of the bowel for many years before spreading to other parts of the body. A test called Faecal Occult Blood Test (FOBT) is used to detect small amounts of blood in the bowel motion, a sign of potential presence of bowel cancer, because it is simple and non-invasive.

From 1984 onwards, about 50000 subjects were screened for bowel cancer at the St Vincents Hospital, Sydney, Australia. More details about this data are discussed in Lloyd and Frommer [61], [62], [63], [23]. The screening procedure was based on a sequence of binary diagnostic tests, self-administered on 6 successive days. On each of these 6 occasions, the absence or the presence of blood in faeces was recorded. Post-verification of the results of the tests was done by physical examination, sigmoidoscopy and colonoscopy, performed only if at least one of the six test was positive. People with all six tests negative were not further assessed.

The frequency distribution of the number of positive tests is reported in Table 2.3.

TABLE 2.3: Bowel cancer data positive sample.

x	0	1	2	3	4	5	6	Total
f_x	?	37	22	25	29	34	45	192

Lloyd and Frommer [61], [62] mentioned that a sample of 122 patients with confirmed cancer status were screened again using the same screening procedure. The corresponding frequency distribution is shown in Table 2.4 and it represents the validation sample for this study.

TABLE 2.4: Bowel cancer data validation sample.

x	0	1	2	3	4	5	6	Total
f_x	22	8	12	16	21	12	31	122

2.2.3 Syphilis in Izmir - Turkey

The next surveillance data is related to Syphilis in Izmir, Turkey, diagnosed in 2003 and collected between 21 January 2003 and 25 March 2005. Specific details about this data

can be found in Durusoy [37] and Köse *et al.* [55]. The main goal of the study was to assess the completeness of the surveillance system for syphilis and other transmittable diseases in Izmir and quantify the under-notification of disease occurrence. This was done by collapsing the multiple laboratory identifications and notifications into one category and applying capture-recapture methodology. Köse *et al.* [55] consider only Syphilis and estimate the number of undercounts of the disease by following an extended Lincoln-Peterson approach for multiple identifications in one source.

The entire province of Izmir (nine urban and 19 rural districts) with approximately 3.5 million inhabitants formed the target case study population. Cases were identified by one of the two university hospitals, or one of the other six public hospitals. The probability of having cases identified elsewhere in Izmir was almost non-existent since all main medical facilities were covered in the study. The Izmir Provinical Health Directorate was notified of cases, providing a hospital notification list. Thus, 133 serology laboratories participated in the study with cases frequently identified by multiple laboratories.

The positive data sample for this case study is achieved matching the results from the hospitals and the laboratories. For each case it was determined if it was identified by the hospital and how often it was identified by the laboratories. This can be seen in the first row of Table 4.15.

Table 4.15 shows the frequencies of syphilis cases by hospital notifications and the count of laboratory identifications. For instance, 73 people with syphilis were identified by the hospital and the laboratory once. Notice that there were no cases identified by the hospital and the laboratories six times.

The validation data sample is achieved by the results of the serology laboratories applied on a sub-sample of the positive sample. The results are presented in the second row of Table 4.15. This serology test failed in 18 of the identified cases by the hospitals and the laboratories.

TABLE 2.5: Positive and validation sample of Syphilis data.

		Laboratories							
		0	1	2	3	4	5	6	Total
Hospital	0	?	73	52	17	6	1	0	149
	1	18	25	22	10	9	1	1	86

2.2.4 Brucellosis in Izmir - Turkey

As mentioned in the last section, Durusoy [37] had access to data related to other transmittable diseases from the same surveillance sources. Another infectious disease was brucellosis. Following the explanation given above, Table 2.6 shows the positive sample (first row in Table 2.6) and the validation sample (second row in Table 2.6) for the data collected for this disease.

TABLE 2.6: Positive and validation sample of Brucellosis data.

		Laboratories							
		0	1	2	3	4	5	6+	Total
Hospital	0	?	57	15	14	10	4	7	107
	1	68	26	14	7	4	1	6	126

2.2.5 Heroin users in Bangkok - Thailand

The following data reports the number of heroin users in Bangkok metropolitan region who contacted treatment centres (private and public health centres) during the year 2001 to treat drug dependence. The data was provided by the surveillance system of the Office of the Narcotics Control Board (ONCB) of the Ministry of Public Health in Thailand. More details about this data set can be seen in Lerdsuwansri [57].

TABLE 2.7: Count distribution of Heroin user contacts.

		2nd half year						
		0	1	2	3	4	5	6
1st half year	0	?	1401	369	98	23	1	1
	1	1736	315	129	50	26	1	0
	2	445	137	105	53	20	4	0
	3	164	89	75	49	30	1	2
	4	47	25	48	34	8	0	0
	5	5	7	8	2	3	0	0
	6	1	0	1	1	0	0	0
	8	0	0	0	1	0	0	0

This data can be split into those with and without contact in the first half year. These define the positive and the validation data sets for this case study as following.

The positive data is shown in Table 2.8 and it relates to the frequencies of the treatment episodes per drug addict who were contacting treatment centre only during the second

half of the year but not during the first half which corresponds to the first row in Table 2.8:

TABLE 2.8: Heroin users in Bangkok positive sample.

x	0	1	2	3	4	5	6	Total
f_x	?	1401	369	98	23	1	1	1893

It can be seen that 1401 drug users contacted a treatment centre one time, 369 heroin users contacted twice and so on in a maximum of 6 times. The size of this set of heroin users was $n = 1893$ during the period of the study.

The validation sample for this case shows the number of heroin users who contacted the treatment centre not just during the second half of the year but also during the first half. The results are shown in Table 2.9. These are marginalised over the count of contacts during the first half.

TABLE 2.9: Heroin users in Bangkok validation sample.

x	0	1	2	3	4	5	6	Total
g_x	2398	573	366	190	87	6	2	3622

This is an interesting data set as it shows how to construct a validation set from capture-recapture data which were not specifically designed for a validation study.

What is important to emphasize at this point is that the validation sample may help to check whether the model is "correct" for the unobserved part of the population which is not possible if we observe the positive sample. Only simulation studies will show that the use of a validation sample may help derive a better population size estimate with less bias and more precision.

Chapter 3

Review of Capture-Recapture Methods

The purpose of this chapter is to review the background on Capture-Recapture studies. It starts by explaining the problem of having zero-truncated count data and follows on how to estimate the probability of a zero count where we aim to estimate the total population size, the major issue in Capture-Recapture studies. After that, we describe two types of data set structure for capture-recapture data and give some examples of estimators that are frequently used to estimate the population size, the Horvitz-Thompson and the Turing estimator. These estimators will be used throughout this dissertation. Some other estimators are also presented as part of the literature review on this topic. We proceed with some practical examples where Capture-Recapture methods have been applied. Finally, we introduce the concept of validation information in a capture-recapture context and end this chapter with an introduction to some methodology to be analysed in the next chapters.

3.1 The zero-truncated count data problem

A very common procedure to estimate the total size of a population of interest and describe its features is by adopting a census approach. However, this identification/registration mechanism has some limitations since it may be impractical to reach all the individuals of the target population. Good examples are wildlife populations, human

populations with epidemic diseases or homeless people in a certain area [50], [53], [12], [86]. Another limitation is that this registration process is often dependent on the individual willingness to take part in the study. The registration will not be complete or it might fail if the individuals do not cooperate or are not committed with it. For instance, in medical studies, it is common to conduct screening tests to detect the presence or absence of a certain condition. If the patients fail to participate or the test is not totally accurate, it will lead to a portion of the population being not observed or misclassified and these individuals are referred to as zero-counts. The situation described forces the question about the total size of the population and how to deal with zero-truncated data - data with unobserved zeros.

We say that under-reporting happens when there is a failure in reporting data, i.e., the quantity of data reported is less than the real amount. In human populations, this is exemplified in the work undertaken by Merli [67] about the under-reporting of births and infant deaths in rural China. Approximately three decades after the implementation of the "One Child Policy" in 1979 in China, it was made an evaluation of the Chinese demographic data [67]. It reported that Chinese birth and infant mortality statistics suffer from severe under-reporting. The causes of this failure were not investigated, however, one of the many reasons pointed out for this issue was the flaws of the registration process.

Another example is the under-reporting rates of domestic violence and sexual abuse complaints on women. Frequent reasons for women not to report these abuses include fear/anxiety of not being believed by others, insecurity and fear of getting into trouble after the indictment against the partner. These factors lead to a sharp under-reporting collection of data [39].

Also, actions that damage biodiversity like illegal trade in wild animals and plants are under-reported, as well as the record of the number of illegal immigrants in the Netherlands [43], [1].

As it can be interpreted by looking at the examples above, a part of the population cannot be reached through registration or identification processes. Capture-recapture studies appear as a tool to do inference on the real size of the elusive population of interest.

A natural way to proceed with the modelling is to describe the population density function by a parametric probability density function $p_x(\theta)$ which denotes the probability of exactly x identifications for a generic unit, where $p_x(\theta) \geq 0$ and $\sum_{x=0}^{\infty} p_x(\theta) = 1$.

For example, we can consider the simple case of a Binomial distribution. Then, we have that the probability distribution is as follows:

$$p_x(\theta) = P(X = x) = \binom{m}{x} \theta^x (1-\theta)^{m-x} \tag{3.1}$$

where X is a Binomial random variable, $x = 0, ..., m$ and $p_x(\theta) = 0$ for $x > m$.

Naturally, we have that $p_0(\theta)$ is the probability of zero-counts (unobserved units) in the population. In the Binomial case, this is equal to $p_0(\theta) = (1-\theta)^m$. The probability that an individual is observed is $1 - p_0$ and the total size of the population N can be described by:

$$N = N(1 - p_0) + Np_0 \tag{3.2}$$

Taking into account that $N(1-p_0) = \mathbb{E}(n) \simeq n$ corresponds to the observed part of the population, we can rewrite the equation as:

$$N = n + Np_0 \tag{3.3}$$

where $n = \sum_{x=1}^{m} f_x = f_1 + ... + f_m$ corresponds to the sum of all the observed units.

Note that the population size can also be simply described by:

$$N = f_0 + n \tag{3.4}$$

where f_0 is the frequency of zero-counts.

The Horvitz-Thompson estimator follows:

$$\hat{N} = \frac{n}{1 - p_0} \tag{3.5}$$

$\hat{N}$ is a moment and maximum likelihood estimator as n is binomial with known event probability p_0 and unknown size parameter N. Hence, focus is now on p_0.

3.2 Estimating the probability of a zero count

Let $P(X = x)$ be the probability that a variable X takes the value x, $x \geq 0$, then $P(X = x|X > 0)$ is the probability of observing $X = x$ given that $X > 0$ and $P(X > 0)$ represents the probability of $X > 0$. We can write $P(X = x|X > 0)$ as follows:

$$P(X = x|X > 0) = \frac{P(X = x)}{P(X > 0)} = \frac{P(X = x)}{1 - P(X = 0)} \tag{3.6}$$

which can be written as

$$p_x^+ = \frac{p_x}{1 - p_0} \tag{3.7}$$

When $x > 0$ and p_x represents the discrete mass probability function $p_x = P(X = x)$.

Let us consider then, as an example, the Binomial probability distribution for a Binomial random variable X:

$$p_x(\theta) = P(X = x) = \binom{m}{x} \theta^x (1 - \theta)^{m-x} \tag{3.8}$$

$x = 0, 1, ..., m$ and $p_x = 0$ for $x > m$.

The zero-truncated Binomial considering that $p_0 = (1 - \theta)^m$ is as follows:

$$p_x = \frac{\binom{m}{x} \theta^x (1 - \theta)^{(m-x)}}{1 - (1 - \theta)^m} \tag{3.9}$$

$x = 1, 2, 3, ..., m$.

The next step is to derive an estimate $\hat{\theta}$ for θ and use $\hat{\theta}$ in $p_0(\hat{\theta}) = (1 - \hat{\theta})^m$ to estimate N, where p_0 is the probability of a zero count distribution. An estimator $\hat{\theta}$ for θ can be obtained by fitting a zero-truncated Binomial distribution using, for example, the EM

algorithm. The EM algorithm is a popular approach to estimate a parameter of interest by means of maximum likelihood estimation when data is not complete, i.e., through the EM algorithm we can obtain an estimate $\hat{\theta}$ and consequently $\hat{p_0}$. From here, we can straightly achieve an estimate for N, $\hat{N} = \frac{n}{1-\hat{p_0}}$ as seen in 3.5.

Another example is the zero-truncated Poisson distribution where, following the same steps as shown above for the Binomial distribution, we have:

$$p_x = \frac{exp(-\lambda)\lambda^x}{x!(1 - exp(-\lambda))} \tag{3.10}$$

where $p_0 = exp(-\lambda)$ and $1 - p_0 = (1 - exp(-\lambda))$.

3.3 Structure of Capture-Recapture data

Typically, capture-recapture data are a result of a history of trappings such as live-trapping in order to access the total size of an elusive population. During the trapping or registration process, each individual is identified multiple times over the period of the study.

Let us set the number of captures in m. The counts for each individual is a sequence of zeros and ones formed in a matrix X, where X_{ij} is 1 if the individual i has been identified in the j^{th} occasion and 0 if not.

$$X_{ij} = \begin{cases} 1, & \text{if individual } i \text{ is observed on occasion } j \\ 0, & \text{otherwise} \end{cases}$$

Notice that X_{ij} is only observed if $\sum_j X_{ij} > 0$.

The capture-recapture scenario can be expressed like in Table 3.1:

TABLE 3.1: Capture-Recapture history.

Individual	Occasion j				$X_i = \sum_{j=i}^{m} X_{ij}$
i	1	2	...	m	
1	$X_{1,1}$	$X_{1,2}$	...	$X_{1,m}$	X_1
2	$X_{2,1}$	$X_{2,2}$	...	$X_{2,m}$	X_2
3	$X_{3,1}$	$X_{3,2}$	...	$X_{3,m}$	X_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$X_{n,1}$	$X_{n,2}$	...	$X_{n,m}$	X_n
$n+1$	$X_{n+1,1}$	$X_{n+1,2}$	...	$X_{n+1,m}$	X_{n+1}
$n+2$	$X_{n+2,1}$	$X_{n+2,2}$	...	$X_{n+2,m}$	X_{n+2}
$n+3$	$X_{n+3,1}$	$X_{n+3,2}$	...	$X_{n+3,m}$	X_{n+3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N-1$	$X_{N-1,1}$	$X_{N-1,2}$	...	$X_{N-1,m}$	X_{N-1}
N	$X_{N,1}$	$X_{N,2}$	...	$X_{N,m}$	X_N

The sum X_i refers to the number of times individual i has been identified and takes values from 0 to m, i.e., each entry $X_{i,j}$ in the sampling occasions determine the reword sequence of times that the individual was observed out of m occasions. For example, if the number of sampling occasions is $m = 3$, and we get a sequence 101 for a given individual, this means that individual was observed at the first and last occasion. As it is shown in Table 3.1, the population is divided into two parts: a truncated and an untruncated part. The first is the sub-population that we observe $X_1, X_2, ..., X_n$, while the second is composed by the individuals $X_{n+1}, X_{n+2}, ..., X_N$ that we do not observe as, without loss of generality, we have assumed that $X_i > 0, \forall i = 1, ..., n$ and $X_i = 0, \forall i = (n+1), ..., N$. Capture-recapture methods attempt to estimate the number of missed individuals in order to have access to the total population size N.

Consequently, we are able to rewrite the information given in Table 3.1 using the frequency table in Table 3.2:

TABLE 3.2: Frequency distribution of the count of identifications per unit.

x	0	1	2	...	m
f_x	f_0	f_1	f_2	...	f_m

where f_x represents the number of individuals observed exactly x times during a study period with $x = 0, 1, 2, ..., m$. Therefore, the total number of observed individuals is:

$$f_1 + f_2 + f_3 + ... + f_m = \sum_{x=1}^{m} f_x = n \tag{3.11}$$

and the total number of identifications is $\sum_{x=0}^{m} x f_x$ where f_0 is missing, but that does not matter as it is multiplied by zero.

We can then write the total size N of the population as follows:

$$N = f_0 + f_1 + f_2 + ... + f_m = f_0 + \sum_{x=1}^{m} f_m = f_0 + n \tag{3.12}$$

A good starting point to model the observed frequency distribution for the sum X_i is to use the Binomial distribution $B(m, \theta)$; in this case, the number of recaptures m is fixed and finite and we assume that each identification occasion is independent of each other, with a homogeneous capture probability θ. If m is not specified in the study, the Poisson distribution $P(\lambda)$ where λ denotes the mean for each individual during the study period, would be a more appropriate way to fit the observed distribution. This is detailed by using the following benchmark data example.

Example 1: A capture-recapture history data for 38 deer mice with 6 different trapping occasions analysed by Amstrup *et al.* [4] is shown in Table 3.3. The duration of the study was short so it is reasonable to assume that this is a closed population [4]. The variable X_i denotes the number of times that each deer mice ($i = 1$) was identified over the 6 trapping occasions. We can clearly see from the table that the first deer mice was identified in all the occasions of the sequence is (1,1,1,1,1,1). Consequently, X_1 is equal to 6 for that animal. The second deer mice ($i = 2$) was identified in the first, fourth, fifth and sixth occasions (1,0,0,1,1,1). Note that a deer mice which was not identified in any occasion is represented by (0,0,0,0,0,0) and it corresponds to the unobserved part of the population.

We are interested in estimating the number of deer mice that were not identified in this study, hence the frequency of X_0.

TABLE 3.3: Capture-Recapture history of 38 deer mice with 6 trapping occasions.

Unit	Occasion						X_i
i	1	2	3	4	5	6	
1	1	1	1	1	1	1	6
2	1	0	0	1	1	1	4
3	1	1	0	0	1	1	4
4	1	1	0	1	1	1	5
5	1	1	1	1	1	1	6
6	1	1	0	1	1	1	5
7	1	1	1	1	1	0	5
8	1	1	1	0	0	1	4
9	1	1	1	1	1	1	6
1	0	1	1	0	1	1	1
11	1	1	0	1	1	1	5
12	1	1	1	0	1	1	5
13	1	1	1	1	1	1	6
14	1	0	1	1	1	0	4
15	1	0	0	1	0	0	2
16	0	1	0	0	1	0	2
17	0	1	1	0	0	1	3
18	0	1	0	0	0	1	2
19	0	1	0	1	0	1	3
20	0	1	1	0	1	0	3
21	0	1	0	1	0	1	3
22	0	1	0	0	0	1	2
23	0	1	0	0	1	1	3
24	0	0	1	0	0	0	1
25	0	0	1	1	1	1	4
26	0	0	1	0	1	1	3
27	0	0	1	1	1	1	4
28	0	0	1	0	1	0	2
29	0	0	1	0	0	0	1
30	0	0	0	1	0	0	1
31	0	0	0	1	1	1	3
32	0	0	0	1	1	0	2
33	0	0	0	0	1	0	1
34	0	0	0	0	1	0	1
35	0	0	0	0	1	0	1
36	0	0	0	0	0	1	1
37	0	0	0	0	0	1	1
38	0	0	0	0	0	1	1

Denoting by f_x the frequency of units identified exactly x times, $x = 1, 2, ..., 6$, we can rearrange the table using the frequency distribution of deer mice counts. See Table 3.4.

TABLE 3.4: Frequency distribution of counts for 38 deer mice.

x	0	1	2	3	4	5	6	Total
f_x	?	9	6	7	6	6	4	37

It is important to emphasize that f_0 is unknown and it needs to be estimated. Since we have a fixed number of recaptures, $m = 6$, we could proceed with the modelling using the Binomial distribution, by estimating p_0 and getting the total size of the population using the Horvitz-Thompson estimator $N = \frac{n}{1-\hat{p_0}}$.

There are two types of data structures. One originated by repeated count data and another originated by different sources of data. This is an example of data obtained from different sources since we are dealing with 6 different trapping occasions, and each occasion could be considered as a source. For more details about this dataset, please see [23]. Normally, sources of identified individuals are summarized in frequency/contigency tables; in the case of two sources, we may have:

TABLE 3.5: Capture-Recapture history with two sources.

	Source 2		
Source 1	Yes	No	Total
Yes	f_{11}	f_{10}	n_1
No	f_{01}	f_{00}	?
Total	n_2	?	N

where f_{00} denotes the frequency of unobserved units, f_{10} is the frequency of units identified only by the first source, f_{01} is the frequency of units identified only by the second source and f_{11} denotes the frequency of units identified by both.

The total population size N is given by the sum of the individuals observed by any of both sources plus the individuals that remain unobserved. Then, $N = f_{00}+f_{01}+f_{10}+f_{11}$. Rewriting this table in the form of a table of frequencies as above, we get:

TABLE 3.6: Frequency distribution counts two sources.

x	0	1	2
f_x	f_{00}	$f_{10} + f_{01}$	f_{11}

For three or more sources, the capture-recapture history can be written following the same principles. Let's see an example for three sources:

TABLE 3.7: Capture-Recapture history with three sources.

Sources			Frequency count
Source x	Source y	Source z	f_{xyz}
0	0	0	f_{000}
1	0	0	f_{100}
0	1	0	f_{010}
0	0	1	f_{001}
1	1	0	f_{110}
1	0	1	f_{101}
0	1	1	f_{011}
1	1	1	f_{111}

which can be summarized by the frequency table:

TABLE 3.8: Frequency distribution counts three sources.

x	0	1	2	3
f_x	f_0	f_1	f_2	f_3

In this case, the population size N is:

$$N = f_0 + f_1 + f_2 + f_3 = f_{000} + (f_{100} + f_{010} + f_{001}) + (f_{110} + f_{101} + f_{011}) + f_{111} \quad (3.13)$$

where $f_{000} = f_0$ corresponds to the unobserved part of the population. Notice that $f_1 + f_2 + f_3 = n$, the observed part of the population and $N = f_0 + n$.

There are many examples in the literature of frequency data obtained as observations from different sources. We are going to analyse two more examples. To have access to more datasets, see [23].

Example 2: The number of new HIV diagnoses in French children under 13 years old was estimated by Heraud-Bousquet [76] between January 2003 and December 2006. Data was collected from three sources: the EPF/ANRS-French Perinatal Cohort, the DOVIH-Mandatory HIV case reporting and the LaboVIH - Laboratory surveillance of HIV testing activity. The data are reported in Table 3.9.

TABLE 3.9: Capture-Recapture history with three sources for the HIV diagnoses in children under 13 years old in France.

Sources			Frequency count
DOVIH	LaboVIH	EPF	f_{xyz}
0	0	0	?
1	0	0	20
0	1	0	36
0	0	1	64
1	1	0	31
1	0	1	22
0	1	1	15
1	1	1	28

There were $n = 216$ observed diagnosed cases and the corresponding frequency distribution is reported in Table 3.10.

TABLE 3.10: Frequency distribution counts three sources.

x	0	1	2	3
f_x	?	120	68	28

Example 3: Another example of data with different sources is a data set analysed by Fienberg [40] and Wittes [91]. Five different sources: obstetric records, hospital records, the Department of Public Health, the department of Mental Health and special schools were used to estimate the number of children with Down's syndrome who were born between 1955 and 1959, based on the empirical evidence that 537 children were diagnosed during that period. The data is shown in Table 3.11.

TABLE 3.11: Frequency distribution counts (Down's syndrome data).

x	0	1	2	3	4	5
f_x	?	248	188	81	18	2

The number of children who were not identified by any of these sources (but in fact had the disease) is unknown.

Another type of data that occurs very frequently in Capture-Recapture studies is repeated count data. It occurs when each individual is repeatedly identified by the same identification process/mechanism any time during the study. The individuals are identified between 1 to m times with m being the largest observed count. This leads to the frequency of counts f_1, f_2, f_3,..., f_m. In particular, this is a special case where data is originated by the same source.

Two examples of this type of data are given.

Example 4: The occurrence of Scrapie in sheep in the holdings of Great Britain is monitored in the Compulsory Scrapie Flocks Scheme (CSFS) which was established in 2004; it is also monitored by an abattoir survey, stock survey and the statutory reporting of clinical cases. For more details see [24]. The frequency distribution of the count of Scrapie cases within each holding for the year 2005 is as follows:

TABLE 3.12: Frequency distribution counts.

x	0	1	2	3	4	5	6	7	8
f_x	?	84	15	7	5	2	1	2	2

A total of $n = 118$ holdings was observed. The aim here is to estimate the completeness of the surveillance system or to estimate the undercount of Scrapie by the surveillance system, that is, the number of holdings that have diseased sheep but which has not been observed.

Example 5: Oremus [73] estimated the size of a small community of Spinner dolphins around Moorea Island (Tahiti). Observations were done within an 8-month observational period. The following frequencies were reported. In total $n = 52$ different Spinner dolphins have been observed by Oremus *et al* [73].

TABLE 3.13: Frequency distribution counts.

x	0	1	2	3
f_x	?	42	7	2

In all of these example data sets, the count distribution could be modelled by a Binomial model, since the total number of captures m is fixed and known.

Let's observe now, how N could be estimated for all the presented cases by using two well known estimators in the next section.

3.4 Simple estimators for the total population size N

Estimating the total size of a population of interest is the main goal of a capture-recapture study. However, this is not a straightforward process.

Eventually, in the presence of a validation sample, one can raise the question: why not use $\frac{g_0}{N_1}$ where N_1 is the total size of the validation sample as an estimate for $\frac{f_0}{N}$. This is motivated on $\frac{g_0}{N_1} \simeq \frac{f_0}{f_0+n}$ from which the solution $\hat{f}_0 = n\frac{g_0}{N_1-g_0}$ can be found. This

non-parametric estimate is possible. However, it only uses g_0 (and N_1) but neither the full distribution of the validation sample nor the positive distribution of the positive sample. This estimate may then suffer from instability and lack of efficiency.

Recalling the definition of p_0 in capture-recapture studies, we have that p_0 is the probability of zero-counts, that is unobserved units in the population. Therefore, the probability that an individual is observed is $1 - p_0$ and the total size of the population N can be derived using the Horvitz-Thompson estimator $\hat{N} = \frac{n}{1-p_0}$ as reported before.

However other estimators can be used as well. For more informations about other estimators, the interested reader is referred to [23], [25], [5], [6], [45], [29].

Another estimator used in the thesis to estimate the total size of the population of interest was the Good-Turing estimator [45] which will be presented here.

Let us recall f_x that denotes the number of individuals identified exactly x times, while m is the largest observed count. The total number of identifications is given by:

$$\sum_{x=1}^{m} x f_x = S \tag{3.14}$$

The Good-Turing estimator is defined in the context of homogeneous Binomial distributions. Then, considering an homogeneous Binomial with parameter θ, we have:

$$p_0 = (1-\theta)^m = \left[\frac{m(1-\theta)^{(m-1)}\theta}{m\theta}\right]^{\frac{m}{m-1}} = \left(\frac{p_1}{E(X)}\right)^{\frac{m}{m-1}} = \left(\frac{E(f_1)/N}{E(S)/N}\right)^{\frac{m}{m-1}} = \left(\frac{E(f_1)}{E(S)}\right)^{\frac{m}{m-1}} \tag{3.15}$$

where $p_1 = m(1-\theta)^{m-1}\theta$. Replacing the expected values in the right-hand side by the corresponding observed quantities, we obtain:

$$\hat{p_0} = \left(\frac{f_1}{S}\right)^{\frac{m}{m-1}} \tag{3.16}$$

If we plug $\hat{p}_0$ into the Horvitz-Thompson estimator, the Good-Turing estimator is:

$$\hat{N}_{GT} = \frac{n}{1-(f_1/S)^{\frac{m}{m-1}}} \tag{3.17}$$

When $m \to \infty$:

$$\hat{N}_{GT} = \frac{n}{1 - f_1/S} \tag{3.18}$$

The variance for Turing estimation [57] is derived by:

$$\widehat{Var}(\hat{N}_{GT}) = \frac{n\frac{f_1}{S}}{(1+\frac{f_1}{S})^2} + \frac{n^2}{(1+\frac{f_1}{S})^4}\left[\frac{f_1(1-\frac{f_1}{\hat{N}})}{S^2} + \frac{f_1^2}{S^3}\right] \tag{3.19}$$

The Good-Turing estimator has the advantage of being easy to calculate, as the estimate is obtained in a straightforward way, with no need for an iterative procedure.

3.5 Examples of applications of Capture-Recapture data modelling

Applications of capture-recapture methods are countless. As mentioned in the previous sections, this methodology can be used to solve problems or challenges in many areas of research.

In this section, we provide some illustrations of capture-recapture studies which allow us to demonstrate the versatility and flexibility of these methods and have a rough idea of the areas of application.

The first example comes from a capture-recapture study originated by repeated count data in medical sciences. McKendrick [66] analysed data related to a cholera epidemic in an Indian village. More information about this data set is provided in [66] and [23]. Cholera is an infectious disease caused by a bacteria on the small intestine.

The table below shows the frequency distribution of Cholera cases per household:

TABLE 3.14: Frequency distribution counts of cholera in an Indian village.

x	0	1	2	3	4
f_x	?	32	16	6	1

As we can see from the table above, there were 32 households with exactly one cholera case ($f_1 = 32$), 16 households with exactly two cases ($f_2 = 16$), 6 households with exactly three cases ($f_3 = 6$) and one household with exactly 4 cases ($f_4 = 1$).

In summary, 55 households, $n = 55$, were identified with at least one cholera case with a maximum of 4 cases. It is known that this cholera epidemic had spread by many more households which were never identified in the study. It is then of interest to estimate how many households were affected by the epidemic but were not registered in this study, i.e., how many households did not refer any case but having the infection.

The next example is related with the number of illegal immigrants in four cities of the Netherlands. Data for this study was collected from police records. More details about this data can be found in Van der Heijden *et al.* [86], [84] and [85]. This data was analysed using a truncated Poisson regression model. The study is centred on the illegal immigrants who cannot be deported out of the country definitively once they are apprehended by the police. This is the situation that normally happens when there is no cooperation between the organization of deportation and the country of the deported immigrants.

The problem that arises from this situation of disagreement between the deportation organization and the country is that, most likely, those illegal immigrants will not leave the country where they were apprehended when they are asked to do so. Hence, they can be caught in this situation more than once.

The following table shows the frequency distribution of the apprehension counts of the illegal immigrants during the study period:

TABLE 3.15: Frequency distribution of illegal immigrants apprehension in four cities in the Netherlands.

x	0	1	2	3	4	5	6
f_x	?	1645	183	37	13	1	1

William Shakespeare is considered the greatest writer in English literature. Spevack [81] collected data on how many words Shakespeare used in his works. Efron and Thisted [38] tried to estimate how many words did Shakespeare know but not use. We can see in the next table part of the frequency distribution of different words:

TABLE 3.16: Frequency distribution of the words used by Shakespeare (only first 3 counts).

x	0	1	2	3
f_x	?	14376	4343	2292

For more information about this data and/or consultation of the data set, please see [81]. Shakespear's number of known words comprise a total of 884647 words. Spevack's study revealed that Shakespeare knew about 31500 different words, while Efron and Thisted estimated that he knew at least 35000 more words but he did not use all of them.

A closer look to the table of the frequencies shows that exactly 14376 words were used just once, 4343 words were used exactly twice and so on.

As we can see, capture-recapture methods are applied not only to medical sciences (first example), as it can be largely used in social sciences, literature studies (second and third examples respectively). More examples of applications of capture-recapture studies will be presented throughout the thesis. The next section presents another four examples with a particular feature.

3.6 Examples of applications of capture-recapture data modelling with inflation

A closer look to the data sets presented in the last section reveals that often, we may observe an accumulation of units for the first count. This situation is not rare.

Let's illustrate this situation using the study on the prevalence of domestic violence in the Netherlands by Van der Heijden *et al.* [85]. In this study, referring to year 2009,

capture-recapture methodology was applied to estimate the total population size of the offenders. The total number of observed indicted people was 17662 ($n = 17662$). There are 15169 culprits identified exactly once in a domestic violence incident, 1957 exactly twice and so forth. From the data, we can notice that the observed data may be in the form of a one-inflated distribution. It seems that a portion of the culprits captured for the first time changed their behaviour and have not been registered a second time.

TABLE 3.17: Frequency of domestic violence culprits by incident.

x	1	2	3	4	5	6
f_x	15169	1957	393	99	28	16

The reported data for this situation is evidently suffering from one-inflation. From the analysis of the table, it seems that many culprits changed their behaviours after being captured once which generated an increase in observed one counts.

However, it might happen that we have inflation of zeros for the positive sample since f_0 is unknown and no further information is given about unobserved units.

Zero-inflation arises when we are in the presence of a large quantity of zeros with respect to other counts. In capture-recapture studies, the presence of an excessive number of zeros tends to violate the underlying homogeneous distributional assumption questioning the validity of the inference for f_0 and, consequently, for the total population size.

It is very common to face count data with many zeros when working in agriculture, econometrics, manufacturing, species abundance and medicine. See Ridoutet *et al.* [77] for more details. Therefore, methodology has to be developed when we have evidence of zero-inflation occurrence.

Two examples that motivate the application of zero-inflated models for regression analysis are presented to illustrate the importance of an analysis when zero-inflation may be present.

The analysis of dental caries indices has been approached using zero-inflated count regression models over the past years. This happens because children are having less caries experiences due to an improvement in their oral health which leads to low or even zero counts. Böhning *et al.* [19], evaluated several programmes for reducing caries from a dental epidemiological study in an urban area of Belo Horizonte. For this study, the presence of an excessive quantity of zeros, would violate the usual Poisson distribution

mean-variance relationship. This study contemplated only school-children with 7 years of age from schools with similar backgrounds. The main goal of the study was to compare four methods to prevent dental caries. The interventions' scheme was as follows: school 1 - oral health education, school 2 - all four methods together, school 3 - control group, school 4 - enrichment of the school diet with bran, school 5 - mouthwash with 0.2% sodium fluoride (NaF) solution and finally, school 6 - oral hygiene. These six treatments were randomized to the six schools and all children of a given school received the same treatment. In total, 797 children were examined both before and after the study period. Results of the number of children per DMFT Index (Decayed, missing and filled teeth index) are available bellow. The DMFT Index is an indicator that measures the dental status of a person.

TABLE 3.18: Frequency of school-children per DMFT index in the beginning and end of the study period.

$DMFTIndexlabels$	0	1	2	3	4	5	6	7	8
$DMFTbeg$	172	73	96	80	95	83	85	65	48
$DMFTend$	231	163	140	116	70	55	22	0	0

In this study we have confirmation of zero-inflation if a Poisson model is considered. It is clear also that the DMFT index substantially improved which explains the increasing weight of zeros at the end of the study. More details about the data can be found in Böhning *et al.* [19].

Another example of a zero-inflated situation is given in the paper of Min and Agresti [70] for a pharmaceutical study. The original data was not used in the paper due to the companies' confidentiality, however, despite some values were modified, the basic structure of the data was kept. Zero-inflation is one of those characteristics. The study consisted of comparing two treatments for a particular disease in terms of the number of episodes of a certain side effect. In total, 118 patients ($n = 118$) were evaluated with 59 random patients receiving treatment A and the remaining receiving treatment B. At each of six visits, the number of side effects was measured. From the observations, around 83% were identified as zeros. The table below shows the frequency of side effects for treatments A and B.

Once again, it is clear the existence of zero-inflated data occurrence in this study when a homogeneous Poisson distribution is considered with treatment A showing signs of being more effective in controlling the disease.

TABLE 3.19: Side effect frequencies in treatment A and B.

Treatment	0	1	2	3	4	5	6
A	312	30	11	0	1	0	0
B	278	39	20	6	7	2	2

The question is how to deal with situations of excessive zero counts in order to estimate the total population size? In fact, if there is no information about the data collection, it might be impossible to identify zero inflated cases with only a positive sample. Consequently, it is impossible to check the performance of the adopted model. This situation changes if we have access to a validation sample as it will be explained in the next section and explored along this work.

3.7 Review of important methods used in the thesis

In this section, we will describe some methodology that will be used in the next chapters.

Since the main goal of applying a capture-recapture method is to model the count probability distribution, several estimators can be used according to the study purpose. In this dissertation will be presented two important estimators for the population size N.

The understanding of this chapter is important for the understanding of more complex topics presented in chapter 4, 5 and 6.

3.7.1 Estimators overview

As aforementioned, the Horvitz-Thompson and the Good-Turing estimators are going to be used to estimate the total size of the population in next three chapters (4, 5 and 6). Therefore, we start with a brief introduction to both.

3.7.1.1 The Horvitz-Thompson estimator

The Horvitz-Thompson estimator, named by Daniel Horvitz and Donovan Thompson in 1952 [51], is an estimator for the total size of a population of interest.

Let N be the total population size and $1 - p_0$ the probability that an individual is identified by a certain registration mechanism. Then, p_0 will be the probability that an individual is not identified. If X is the random variable which describes if the individual is identified or not, we have a dummy variable assuming the values:

$$X_i = \begin{cases} 1, & \text{if the i individual was identified} \\ 0, & \text{otherwise} \end{cases} \tag{3.20}$$

Notice that the total number of observed units n is defined by:

$$n = \sum_{i=1}^{N} X_i \tag{3.21}$$

As previously shown, N can be written as:

$$N = N(1 - p_0) + Np_0 \tag{3.22}$$

Notice that Np_0 corresponds to the part of the population which is not observed as opposite for $N(1 - p_0)$, which is the expected value for n. Therefore, N can be written as follows:

$$N = n + f_0 = \mathbb{E}(n) + \mathbb{E}(nf_0) = \mathbb{E}(n) + Np_0 \tag{3.23}$$

where $\sum_{x=1}^{m} f_1 + ... + f_m = n$. Solving for N, we obtain:

$$N = \frac{\mathbb{E}(n)}{1 - p_0} \tag{3.24}$$

which substituting the expected value by the observed one leads to the Horvitz-Thompson estimator:

$$\hat{N} = \frac{n}{1 - p_0} \tag{3.25}$$

This can be shown to be the maximum likelihood estimator for the population size N. For more on this topic, see Bishop *et al.* [13] and Van der Heijden *et al.* [84].

3.7.1.2 The Good-Turing estimator

The Good-Turing estimator was developed by Alan Turing and published in 1953 by I. J. Good [45]. It was initially defined to estimate the number of species in a wildlife context and then applied to human populations as well.

Again, let N be the total population size and X the number of times an individual was observed during a study period. Recalling the notation used before to define the positive sample, let f_x be the number of individuals identified exactly x times out of m recaptures. It was already shown in section 3.4 how to obtain the Good-Turing estimator for a Binomial distribution:

$$\hat{N}_{Turing} = \frac{n}{1 - (f_1/S)^{\frac{m}{m-1}}} \tag{3.26}$$

where n represents the sum of the number of observed individuals, f_1 represents the number of individuals observed in the first occasion of the study and, $S = \sum_{i=1}^{m} x f_x$. It should be mentioned that there are no iterations to get the final estimate for N using this method. So, the calculation of the value for $\hat{N}$ should be achieved directly from the formula in an easy and direct way.

Notice that for the Poisson case:

$$p_0 = \exp(-\lambda) = \frac{\exp(-\lambda)\lambda}{\lambda} = \frac{\hat{p_1}}{E(X)} = \frac{\frac{f_1}{N}}{\frac{S}{N}} = \frac{f_1}{S} \tag{3.27}$$

and the Good-Turing estimator for the Poisson case is:

$$\hat{N}_{Turing} = \frac{n}{1 - (f_1/S)} \tag{3.28}$$

In fact when m becomes large in equation (3.26), the Good-Turing estimator for the Binomial distribution goes to the expression of the Good-Turing estimator for the Poisson distribution.

3.7.2 An introduction to the Expectation-Maximization (EM) algorithm

The Expectation-Maximization (EM) algorithm is an iterative algorithm for parameter estimation formalised first by Dempster *et al.* in 1977 [32]. This algorithm leads to maximum likelihood estimation in the presence of unobserved variables, i.e. the data is incomplete. The EM algorithm is particularly useful when the complete data log-likelihood is easy to maximize, whereas the incomplete data likelihood does not have a closed form solution. Examples of the application of the EM algorithm can be found in McLachlan *et al.* [65].

The EM algorithm alternates between two steps: the E and the M step. Since the maximum likelihood computation for the complete data is easier to solve than the maximum likelihood estimation associated with the incomplete data, we proceed by using the EM algorithm with the complete data likelihood.

In the E-step of the algorithm, due to lack of observed data, we replace the unobserved data by its expectation conditional on the observed data and the current parameter estimates. In the M-step, we update the values of the parameters by maximizing the likelihood based on the available and the imputed data. The two steps are iterated until convergence of the algorithm.

3.7.2.1 Maximum Likelihood Estimation using the EM algorithm

Let X be a random vector with probability density function $f(x;\theta)$ where θ is an unknown parameter vector. Using the log-likelihood of X, $\log L(\theta;x)$, we want to get a maximum likelihood estimate $\hat{\theta}$ for θ.

To get an estimate $\hat{\theta}$ of θ, let's consider an initial value for θ, $\theta^{(0)}$, and start the first iteration of the EM algorithm based on the complete data log-likelihood.

The E-step takes $\theta^{(0)}$ and calculates the conditional expectation of the complete log-likelihood given the observed data X and the current parameter estimates $\hat{\theta}^0$:

$$Q(\theta, \theta^{(0)}) = E_{\theta^{(0)}}(\log L(\theta)|x) \tag{3.29}$$

In the M-step, $Q(\theta, \theta^{(0)})$ is maximized with respect to θ to obtain an updated estimate $\theta^{(1)}$:

$$Q(\theta^{(1)}, \theta^{(0)}) \geq Q(\theta, \theta^{(0)}) \tag{3.30}$$

We begin the second iteration with a new updated estimate for θ, $\theta^{(1)}$, and the process is repeated until convergence.

To sum up, the algorithm performs as follows:

E-step: Calculate

$$Q(\theta, \theta^{(r)}) = E_{\theta^{(r)}} \log L(\theta)|x \tag{3.31}$$

M-step: Take $\theta^{(r+1)}$ such that

$$Q(\theta^{(r+1)}, \theta^{(r)}) \geq Q(\theta, \theta^{(r)}) \tag{3.32}$$

where r corresponds to the iteration index. For more specific details, please see [65], [57].

3.7.2.2 Maximum Likelihood Estimation for truncated Binomial/Poisson distributions

We will detail the EM algorithm for truncated count data using the Binomial and the Poisson distributions. These distributions will be discussed in more detail in Chapter 4.

The main goal is to estimate the total size N of a population of interest when some of the information is missing.

Let X be the total number of times that a unit was identified over a study period where $X \sim B(n, \theta)$. Hence, X follows a Binomial distribution with probability density function:

$$f(x;\theta) = \binom{m}{x} \theta^x (1-\theta)^{(m-x)} \tag{3.33}$$

$x = 0, ..., m$.

As we know, in capture-recapture studies we are dealing with units which were not identified during a given period and this is the reason why N needs to be estimated. Then, the observed data includes only non-zero units.

The probability density function in (3.33) becomes zero-truncated Binomial considering that $f(0;\theta) = (1-\theta)^m$ as follows:

$$f(x;\theta) = \frac{f(x;\theta)}{1 - f(0;\theta)} = \frac{\binom{m}{x} \theta^x (1-\theta)^{(m-x)}}{1 - (1-\theta)^m} \tag{3.34}$$

$x = 1, 2, 3, ..., m$.

Notice that m is the fixed number of sampling occasions.

Let f_x be the number of units identified exactly x times and $n = f_1 + f_2 + ... + f_m$ the total number of observed counts.

The incomplete observed likelihood function in this case is given by:

$$L(\theta) = \prod_{i=1}^{m} \left[\frac{\binom{m}{i} \theta^i (1-\theta)^{(m-i)}}{1 - (1-\theta)^m} \right]^{f_i} \tag{3.35}$$

Based on the observed counts, we may get an estimate $\hat{\theta}$ for θ in order to achieve an estimate $\hat{N}$ for N.

This is equivalent to estimate the unknown parameter θ by the value for which the likelihood function $L(\theta; x)$ is maximised. This approach is called Maximum-Likelihood Estimation (MLE) [22].

To find the MLE we need to maximize $L(\theta; x)$ with respect to θ. Maximizing the likelihood function is equivalent to maximize the log-likelihood function since the natural logarithmic function is an increasing function and normally has a much simpler form which makes it easier to differentiate.

Let's then consider the observed zero-truncated Binomial log-likelihood function:

$$\log L(\theta) = \sum_{i=1}^{m} f_i \log \left(\frac{\binom{m}{i} \theta^i (1-\theta)^{m-i}}{1-(1-\theta)^m} \right) \tag{3.36}$$

The maximum likelihood estimate $\hat{\theta}$ can be achieved by numerical methods simply computing $\frac{dl(\theta)}{d\theta} = 0$ where $l(\theta)$ represents the log-likelihood function.

Note that the term $\sum_{i=1}^{m} f_i \log \binom{m}{i}$ is a simple fixed constant which does not affect the MLE so it can be omitted in the calculations and it does not appear in the expression of the first derivative of the log-likelihood function:

$$\frac{dl(\theta)}{d\theta} = \frac{\sum_{i=1}^{m} i f_i}{\theta} - \frac{\sum_{i=1}^{m} f_i (m-i)}{1-\theta} + n \frac{m(1-\theta)^{m-1}}{1-(1-\theta)^m} \tag{3.37}$$

Solving $\frac{dl(\theta)}{d\theta} = 0$ to find the MLE does not lead to a closed-form solution.

Let's now define the complete data log-likelihood (unconditional likelihood):

$$\log L(\theta) = \sum_{i=0}^{m} f_i \log \left(\binom{m}{i} \theta^i (1-\theta)^{m-i} \right) \tag{3.38}$$

The E-step takes:

$$Q(\theta) = E(\log L(\theta)|f_1, f_2, ..., f_m; \theta) \tag{3.39}$$

Replacing $L(\theta)$ by its form as in 3.39 we get:

$$Q(\theta) = \sum_{i=1}^{m} f_i \log \binom{m}{i} \theta^i (1-\theta)^{m-i} + \hat{f}_0 \log(1-\theta)^m \tag{3.40}$$

Now, the expectation value of f_0 given θ, $\hat{f}_0 = E(f_0|f_1, f_2, ..., f_m; \theta)$, is found:

$$\hat{f}_0 = N(1-\theta)^m \underbrace{=}_{1} (n + \hat{f}_0)(1-\theta)^m \underbrace{=}_{2} \frac{n(1-\theta)^m}{1-(1-\theta)^m} \tag{3.41}$$

Notice that equalities (1) and (2) in (3.41) are justified by the following equalities:

$$E(f_0) = Np_0 = (n + E(f_0))p_0 \tag{3.42}$$

and

$$E(f_0)(1-p_0) = np_0 \Rightarrow E(f_0) = n\frac{p_0}{1-p_0} = n\frac{(1-\theta)^m}{1-(1-\theta)^m} \tag{3.43}$$

Replacing this new expression of f_0 in 3.40, the unconditional likelihood assumes the form:

$$Q(\theta) = \frac{n(1-\theta)^m}{1-(1-\theta)^m} \log((1-\theta)^m) + \sum_{i=1}^{m} \log\left(\binom{m}{i} \theta^i (1-\theta)^{m-i}\right) \tag{3.44}$$

$$= \hat{f}_0 \log((1-\theta)^m) + \sum_{i=1}^{m} \log\left(\binom{m}{i} \theta^i (1-\theta)^{m-i}\right) \tag{3.45}$$

which does have a closed form solution:

$$\hat{\theta} = \frac{\sum_{j=0}^{m} j f_j}{m(n + \hat{f}_0)} \tag{3.46}$$

Hence, the EM algorithm maximizes then the observed log-likelihood function and the computation executed by the algorithm is as shown:

Conditions: choose an initial value for $\hat{\theta}^{(0)}$ and set $r = 0$

E-step: Calculate

$$\hat{f}_0^{(r)} = \frac{n(1-\theta^{(r)})^m}{1-(1-\theta^{(r)})^m} \tag{3.47}$$

M-step: Update the estimate of θ:

$$\hat{\theta}^{(r+1)} = \frac{\sum_{j=1}^{m} j f_j}{m\left(n + \hat{f}_0{}^{(r+1)}\right)} = \frac{S}{m\left(n + \hat{f}_0{}^{(r+1)}\right)} \tag{3.48}$$

Set $r = r + 1$ and alternate between the E and M steps until the estimate for θ, $\hat{\theta}$, converges, i.e., $|\hat{\theta}^{(r+1)} - \hat{\theta}^{(r)}|$ is smaller than a chosen tolerance threshold.

For the Poisson case, the procedure is similar. Let's suppose now that $Y \sim P(\lambda)$ with a probability density function:

$$f(y;\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} \tag{3.49}$$

Therefore, the probability function for a zero-truncated Poisson distribution stays:

$$f(y;\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!(1-\exp(-\lambda))} \tag{3.50}$$

for $y = 1, 2, 3, \ldots$.

The observed likelihood function, for the Poisson case, is as follows:

$$L(\lambda) = \prod_{y=1}^{m} \left(\frac{\exp(-\lambda)\lambda^y}{y!(1-\exp(-\lambda))}\right)^{f_y} \tag{3.51}$$

The MLE is achieved by computing the first derivative of the expected unobserved log-likelihood function:

$$\hat{\lambda} = \frac{\sum_{j=1}^{m} j f_j}{\hat{f}_0 + n} \tag{3.52}$$

Computing $E(f_0|f_1, f_2, ..., f_m; \lambda)$ which is the same as $\hat{f}_0$, we get the estimator for f_0:

$$\hat{f}_0 = \frac{n \exp(-\lambda)}{1 - \exp(-\lambda)} \tag{3.53}$$

Similarly, the EM algorithm toggles between steps (3.52) and (3.53). For more details of the Poisson MLE steps, please see [19], [83], [33].

3.7.3 The ratio plot

Usually, we start the analysis by visualising the graph of the observed frequency distribution since it is a very simple and quick approach that may provide valuable insights on the next step of the study.

A graphical method was proposed by Dubey [36] and Ord [72] to inspect closeness to discrete distributions such as the Binomial distribution, the Poisson distribution and the Pascal distribution. The plot was further investigated by Hoaglin [48].

Böhning *et al.* [17] developed a graphical device - the ratio plot - to check for a homogeneous Poisson model in the context of frequency of frequencies distribution.

3.7.3.1 Zero-truncated power series distribution

Let

$$\eta(\theta) = \sum_{x=0}^{\infty} \alpha_x \theta^x \tag{3.54}$$

be a power series function. We can rewrite (3.54) as $\sum_{x=0}^{\infty} \frac{\alpha_x \theta^x}{\eta(\theta)} = 1$. Therefore, $p_x(\theta) = \frac{\alpha_x \theta^x}{\eta(\theta)}$ defines a power series distribution for $x = 0, 1, 2, ...$, $\theta > 0$ and $\alpha_x > 0$. Notice that $\eta(\theta) = \sum_{x=0}^{\infty} \alpha_x \theta^x$ is the normalizing constant.

It can be shown that the coefficient α_x defines the specific member of a power series. For example, if $\alpha_x = \frac{1}{x!}$, it defines a Poisson distribution and if $\alpha_x = \binom{m}{x}$ for $x = 0, ..., m$, it defines the binomial distribution (for $\alpha_x = 0$ and $x > m$). We are going now to introduce the ratio plot which is a member of the power series distributions. For more details, see [2].

In capture-recapture studies, we deal with zero-truncated distributions. The ratio plot is simply a plot of the ratio of neighbour frequencies associated with a coefficient related to a chosen distribution versus the number of captures/counts.

$$r_x = a_x \frac{p_{x+1}}{p_x} \tag{3.55}$$

where $x = 0, 1, ..., m$, p_x denotes the probability that the unit is identified exactly x times during the period of the study and is a power series distribution.

$a_x = \frac{\alpha_x}{\alpha_{x+1}}$ is chosen so that $r_x = 1$ under the power series.

This can be extended to focus on distributions where the first ratio is unknown as f_0 is unknown. The ratio plot for an untruncated Poisson distribution is obtained by fixing $a_x = (x + 1)$ as $\frac{\alpha_x}{\alpha_{x+1}} = \frac{(x+1)!}{x!}$, then obtaining:

$$r_x = a_x \frac{p_{x+1}}{p_x} = (x + 1)\frac{p_{x+1}}{p_x} \tag{3.56}$$

which is the expression for the ratio plot in the Poisson case.

The ratio plot is the graph of points (x, r_x). In capture-recapture studies, we do not observe zero counts, so f_0, the frequency of zeros, remains unknown. On the other hand, we do observe the sample frequencies $f_1, f_2, ..., f_m$. Hence, we need to consider the ratio plot for the zero-truncated probabilities:

$$p^+(x) = \frac{p_x}{1 - p_0}. \tag{3.57}$$

The ratio for the zero truncated probability is defined by:

$$r_x = \frac{(x + 1)p_{x+1}/(1 - p_0)}{p_x/(1 - p_0)} \tag{3.58}$$

$x = 1, ..., m$, which is identical to that refering to the untruncated distribution as it holds:

$$r_x = \underbrace{\frac{(x+1)p_{x+1}}{p_x}}_{untruncated} = \underbrace{\frac{(x+1)p_{x+1}/(1-p_0)}{p_x/(1-p_0)}}_{zero-truncated} \tag{3.59}$$

Replacing p_x by the correspondent frequency of counts, we get:

$$\hat{r}_x = (x+1)\frac{f_{x+1}}{f_x} \tag{3.60}$$

which can be used for checking the zero-truncated distribution and the untruncated count distribution.

3.7.3.2 The Binomial distribution

As previously described, the ratio plot is an important tool for exploring the observed count distribution. Let us consider the Binomial probability distribution which is given by:

$$p_x(\theta) = P(X = x) = \binom{m}{x} \theta^x (1-\theta)^{m-x} \tag{3.61}$$

$x = 0, 1, ..., m$.

The main idea is to consider ratios of the observed frequencies to estimate ratios of neighbouring count probabilities. To illustrate this idea, still working with the Binomial distribution, let us consider the ratios as below:

$$\frac{p_{x+1}}{p_x} = \frac{\binom{m}{x+1} \theta^{x+1}(1-\theta)^{m-x-1}}{\binom{m}{x} \theta^x (1-\theta)^{m-x}} = \frac{m-x}{x+1} \ \frac{\theta}{1-\theta} \tag{3.62}$$

Using the non-negative coefficients $a_x = \frac{x+1}{m-x}$, we can write the ratio plot terms as follows:

$$r_x = a_x \frac{p_{x+1}}{p_x} = \underbrace{\frac{x+1}{m-x}}_{a_x} \frac{p_{x+1}}{p_x} = \frac{\theta}{1-\theta}. \tag{3.63}$$

The result is a constant, the odds for the event, which is independent of x. Note that r_x does not change whether we consider the truncated or the untruncated distributions as shown in (3.59). As we have mentioned before, the coefficients a_x directly depend on the chosen base distribution. In this situation, the base is represented by the homogeneous Binomial distribution which is obtained by assuming the absence of unobserved heterogeneity.

In the case of the Binomial distribution, the ratio r_x is constant over x as we can see in 3.1 for $N = 100$, $N = 1000$, $N = 10000$ and $N = 100000$ with sampling error naturally visible for small population sizes. Since the quantity p_x is unknown, a non-parametric estimation f_x/N of r_x is given by:

$$\hat{r}_x = a_x \frac{f_{x+1}/N}{f_x/N} = a_x \frac{f_{x+1}}{f_x} \tag{3.64}$$

where f_x is the observed frequency of counts x.

3.7.3.3 Application to real data

The ratio plot has been frequently used in literature to explore the frequency distribution of real data and their similarity to homogeneous distributions. Three examples are given here. The first entails literature data from the work of Shakespeare, other two come from medicine and entail a colorectal polyps study, and an illicit drug user research. For more information about this data and the methodology used here, please see [23].

1. **Shakespeare's data**

Efron and Thisted [38] took the complete work of Shakespeare to get a prediction for the number of words he did know but never used.

Data on the work of Shakespeare have been collected previouly by Spevack in 1963 [81]. Following Spevack, the total number of words known by Shakespeare sum up a total of

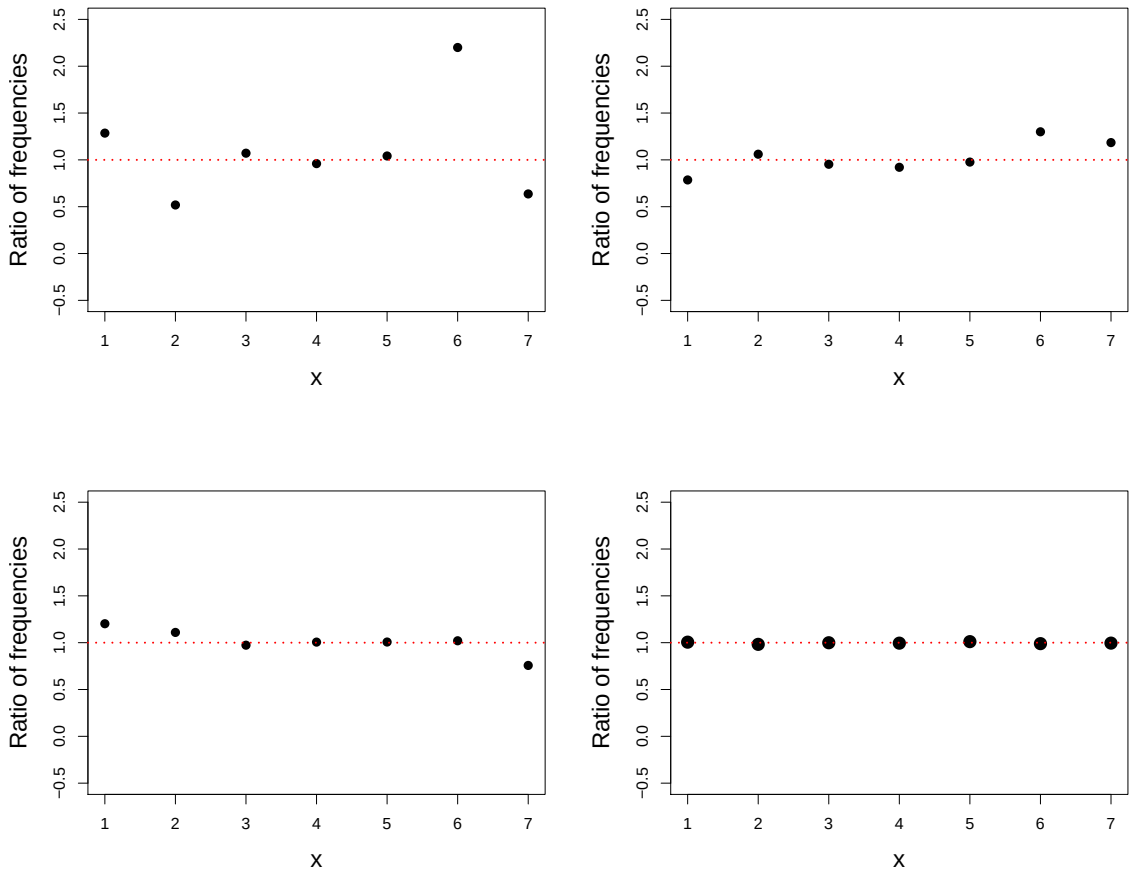

FIGURE 3.1: Ratio plot and regression line for 100, 1000, 10000 and 100000 simulated data (clockwise) from a Binomial distribution with $\theta = 0.5$ and $m = 7$.

884647. Out of these, 14376 ($=f_1$) types appearing just once, 4343 ($=f_2$) types appeared only twice and so on, where f_x denotes the number of words appearing exactly x times. Notice that m is not fixed as it is unkown.

Table 3.20 shows the observed frequency distribution for words count based on the work of Shakespeare (only the first 50 counts):

The problem consists in replying to the question "how many words did Shakespeare know but did not use ever in his writings?". Efron and Thisted estimated that he knew at least 35000 more different words but did not use them in his writings [38].

Alfó *et al.* [2] focused in methods specifically designed to estimating the population size, in particular, using ratio regression methodology. For the presented case, the Poisson and the Geometric distributions were analysed.

TABLE 3.20: Frequency distribution f_x of the words used by Shakespeare exactly x times.

x	1	2	3	4	5	6	7	8	9	10
f_x	14376	4343	2292	1463	1043	837	638	519	430	364
x	11	12	13	14	15	16	17	18	19	20
f_x	305	259	242	223	187	181	179	130	127	128
x	21	22	23	24	25	26	27	28	29	30
f_x	104	105	99	112	93	74	83	76	72	63
x	31	32	33	34	35	36	37	38	39	40
f_x	73	47	56	69	63	45	34	49	45	52
x	41	42	43	44	45	46	47	48	49	50
f_x	49	41	30	35	37	21	41	30	28	19

The ratio plot for this data set using the Poisson or the Geometric distribution shows no evidence that these are the right distributions to use for the total number of words Shakespeare knew but did not use. However, for the Poisson case, the ratio plot gives evidence of a straight line pattern. See Figure 3.2 (left panel). Used many times as a diagnostic device, the ratio plot leads in this case to a deviation from these two simple models since no horizontal line pattern is observed in both ratio plots.

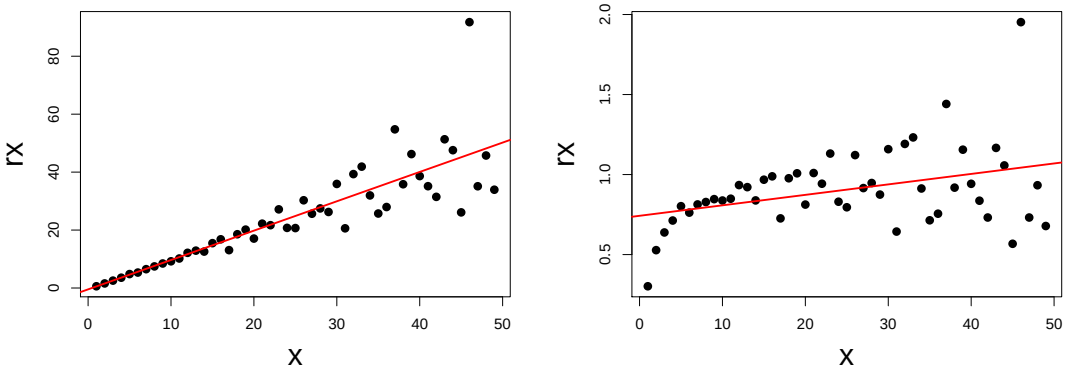

FIGURE 3.2: Ratio plot for the Shakespeare data set line for the Poisson (left panel) and the Geometric distribution (right panel).

Bunge and Sernaker [26] used the empirical probability generating function to estimate the parameters of the data distribution however using different models to choose the best estimate of the population size by comparing the goodness of fit from the different models obtained.

2. **Colorectal polyps:**

The second example where the ratio plot was applied in capture-recapture count data was performed by Maruotti *et al.* [64], [5]. The authors studied the use of the Conway-Maxwell-Poisson (CMP) distribution to estimate the unknown population size for zero-truncated count data in the case of under and over-dispersion. Using the CMP distribution as a base distribution, the ratio plot allowed them to provide insights on the uncertainty of the population size estimates for three different data sets.

One of the data sets relates to colorectal cancer, one of the most regular cancer in terms of mortality. In 1990, the Arizona Cancer Center recruited individuals with previous history of colorectal adenomatous polyps to determine if a wheat bran fibre can prevent the recurrence of those polyps. The individuals in the study were randomly allocated to either: low fibre or high fibre diet.

Despite colonoscopy is being seen as an effective tool to screen for this particular cancer, it is known that any screening test or study for diagnosis a medical condition is totally accurate and can lead to a misclassification during the process and consequently, to an undercount of the number of individuals who actually have cancer. Hence, there could be people false-negatively diagnosed during the colonoscopy and an unknown number of zero-counts in the sample.

The total population size for each group is known. The low fibre treatment group was composed by 584 individuals and the high fibre treatment group was composed by 722.

Following this approach, an estimate for the undercount of the non-zero frequencies is presented using ratio regression. For more details, please see [64], [5].

3. **Needle exchange programme in Scotland (1997)**

The number of individuals who visited a Scottish needle exchange during 1997 was reported in a study by Hay and Smit [47]. The data was collected during a research programme on drug misuse prevalence in Scotland. For each individual accessing the service, a unique identifier number was assigned. Therefore, the number of individuals who had contacted the service was recorded, see Table 3.21:

TABLE 3.21: Frequency distribution of the individual episode count in the needle exchanging program in Scotland (first 10 episodes).

x	0	1	2	3	4	5	6	7	8	9	10
f_x	?	175	85	50	47	37	38	32	16	17	17

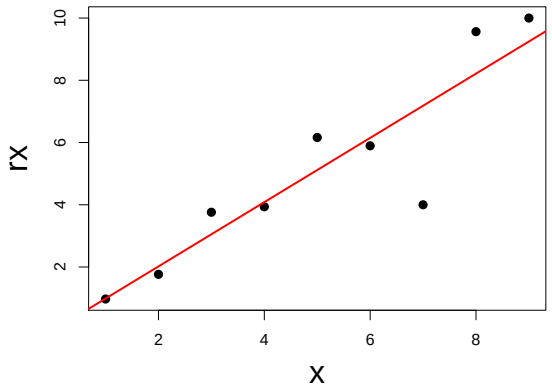

FIGURE 3.3: Ratio plot of episode count in Scottish needle exchange program, 1997.

Let us now observe the ratio plot using the Poisson distribution as reference, $a_x = (x+1)$:

In this case, as the graph 3.3 shows, the ratio plot for the Poisson distribution leads to a fitted regression line that adjusts very well to the ratios' distribution. Notice that the regression line is passing very close to the origin. Böhning *et al.* [17] investigated this effect in the ratio plot to estimate N using an exponential distribution and a generalized Turing estimator. For more information on this approach, please see [17].

Chapter 4

Estimation Under Homogeneity

4.1 The EM Algorithm

Let us suppose that we are interested in studying the problem of missing zero counts in a target population of interest. Therefore, the main goal is to estimate the total population size by estimating the number of missing individuals.

In order to estimate N, simple models using the Binomial or the Poisson distribution are considered as well as an extension of these models considering validation information. Finally, more complex and flexible modelling tools using finite mixtures models are discussed in the next chapter.

We focus here on the estimation of N under homogeneity and this will be done using the EM algorithm. The EM algorithm for the Binomial and the Poisson distribution was already introduced in section 3.7.2, and it represents a tool for maximum likelihood estimation.

For a sample of a population size N, let us consider a sample of positive counts $x_1, x_2, ..., x_N$, from a distribution which can be modelled using a mixture probability density function:

$$p_x = p(x, Q) = \int_0^\infty f(x|\theta)q(\theta)d\theta \tag{4.1}$$

where $q(\theta)$ represents an unspecified mixing density and $f(x|\theta)$ a kernel which needs to be specified. For example, the mixture kernel for the Binomial family will be $f(x|\theta) = \binom{m}{x}\theta^x(1-\theta)^{(m-x)}, \theta \in (0,1)$.

Let us now recall basic information on the current context:

- $x_i = 0$ is unobserved;
- $n = \sum_{i=1}^{m} f_i$ is observed, where f_i is the frequency of counts with value $x = i$ and m is the largest observed count;
- f_0 is unknown; therefore the total size of the population N is unknown.

Let us now denote by N the sum of all frequencies, $N = f_0 + f_1 + ... + f_m$, so that $n = N - f_0$, where f_0 is unobserved.

We may define the following likelihoods for capture-recapture modelling when both the positive and the validation samples of size n and N_1, respectively, are available:

- **Conditional Likelihood**:

$$L^C(p_0,...,p_m) = \underbrace{\binom{n}{f_1 f_2 ... f_m}\prod_{j=1}^{m}\left(\frac{p_j}{1-p_0}\right)^{f_j}}_{Positive} \times \underbrace{\binom{N_1}{g_0 g_1 ... g_m}\prod_{j=0}^{m} p_j^{g_j}}_{Validation}$$

This likelihood is a product of a truncated (observed, incomplete) likelihood referring to the positive sample and an untruncated (complete) likelihood referring to the validation sample which is completely observed.

- **Unconditional Likelihood**:

$$L^U(p_0,...,p_m;N) = \binom{N}{f_0 f_1 ... f_m}\prod_{j=0}^{m} p_j^{f_j} \times \binom{N_1}{g_0 g_1 ... g_m}\prod_{j=0}^{m} p_j^{g_j}$$

where $N_1 = g_0 + ... + g_m$ is known while N is unknown, since f_0 is unknown.

The maximization of the unconditional likelihood would imply the maximization over the unknown parameter f_0. This could be done considering a profile likelihood which can be constructed as $L(N) = sup_Q L(Q,N)$ maximizing $L(Q,N)$ in Q for a fixed N.

This profile likelihood can be then evaluated at $N = n, n+1, n+2, ..., n+M$ where M is an upper bound to be determined [57] [74].

Now, let's observe the following relation [3]:

$$\binom{n}{f_1 f_2 ... f_m}\binom{N}{n} = \binom{N}{f_0 f_1 ... f_m} \tag{4.2}$$

so that:

$$L^U = L^C \times p_0^{f_0}(1-p_0)\binom{N}{n} = L^C \times B(N, Q) \tag{4.3}$$

Notice that the unconditional likelihood depends on a large extent on the conditional likelihood which the EM algorithm is designed to maximize. In fact, the unconditional likelihood is a product of the conditional likelihood with a simple Binomial likelihood with parameter N [18]. This relationship is particularly valid for finite mixtures of Binomials with size parameter m and $p_x = \sum_{l=1}^{k} \theta_l^x (1-\theta_l)^{m-x} q_l$.

Sanathanan [79] showed that the two approaches are asymptotically equivalent. See [60] and [49] for more insights. Lerdsuwansri [57] proposed an unconditional maximum likelihood approach using a profile mixture likelihood for estimating the size of a closed population.

In this work, we will consider only the conditional likelihood as it does not depend on N and we believe it contains most of the information. The EM algorithm is defined to maximize the conditional likelihood above; it may be sketched as follows:

E-STEP: Given an estimate $\hat{Q}$ of the mixing distribution Q, compute a new estimate $\hat{N}$ by

$$\hat{N} = \frac{n}{1 - \int p(0|\theta)\hat{Q}(d\theta)} \tag{4.4}$$

M-STEP: Given an estimate $\hat{N}$ of N compute a new estimate of Q by maximizing

$$p_0^{\hat{f}_0+g_0} p_1^{f_1+g_1} ... p_m^{f_m+g_m} \tag{4.5}$$

in Q, where $p_j = \int p(j|\theta)Q(d\theta)$ is the mixture model.

The mixing distribution Q is estimated in the M-step. This new estimate will then be used to construct a new estimate of f_0.

For the time being, we considered the simple case of a homogeneous Binomial (when Q puts a unit mass at θ). The theory for finite mixtures will be introduced in the next chapter.

4.1.1 Application of the EM Algorithm to the case studies

Let us consider the simple homogeneous Binomial distribution

$$f(x|\theta) = \binom{m}{x} \theta^x (1-\theta)^{(m-x)} \tag{4.6}$$

and describe how to apply the EM algorithm with validation information.

The first step we choose an initial value for θ.

The incomplete observed likelihood to be maximized by the EM algorithm is as follows:

$$L^C(\theta) = \prod_{j=1}^{m} \left(\frac{\binom{m}{j}\theta^j(1-\theta)^{m-j}}{1-(1-\theta)^m} \right)^{f_j} \times \prod_{j=0}^{m} \left(\binom{m}{j}\theta^j(1-\theta)^{m-j} \right)^{g_j} \tag{4.7}$$

And the unconditional likelihood is yield as:

$$L^U(\theta) = \prod_{j=0}^{m} \left(\binom{m}{j}\theta^j(1-\theta)^{m-j} \right)^{f_j^*} \tag{4.8}$$

where $f_j^* = f_j + g_j$.

Taking $n = f_1 + f_2 + ... + f_m$ and $N_1 = g_0 + g_1 + ... + g_m$, the algorithm used in this case is described as following:

Choose some initial value $\hat{\theta}$ and set $\hat{p_0} = (1-\hat{\theta})^m$.

E-STEP: $\hat{N} = \frac{n}{1-\hat{p_0}} \Rightarrow \hat{f}_0 = N - n = n\frac{\hat{p_0}}{1-\hat{p_0}}$.

M-STEP: $\hat{\theta} = \frac{\sum_{j=0}^{m} f_j^* j}{m\sum_{j=0}^{m} f_j^*}$, where $f_j^* = f_j + g_j$.

This process is repeated until the difference $|\hat{\theta}^{(new)} - \hat{\theta}^{(old)}|$ is smaller than a chosen tolerance threshold, in our cases this threshold was chosen to be 0.0001. Another criteria could be considered, for instance, the difference between the new log L^C and the one calculated in the step before smaller than a chosen tolerance value as before, $|\log L^C_{(new)} - \log L^C_{(old)}|$.

The complete likelihood can be maximized calculating the first derivative and solving the equation $\frac{\delta l(\theta)}{\delta \theta} = 0$, where $l(\theta)$ represents the log-likelihood. This leads to the estimating equation $\hat{\theta} = \frac{\sum_{j=0}^{m} f_j^* j}{m \sum_{j=0}^{m} f_j^*}$.

The EM algorithm could also be applied for the Poisson distribution.

For the Poisson case, the incomplete (observed) likelihood to be maximized by the EM algorithm is given by:

$$L(\lambda) = \prod_{j=1}^{m} \left(\frac{\frac{\lambda^j exp(-\lambda)}{j!}}{1 - exp(-\lambda)} \right)^{f_j} \times \prod_{j=0}^{m} \left(\frac{\lambda^j exp(-\lambda)}{j!} \right)^{g_j} \tag{4.9}$$

while the unconditional likelihood is defined as:

$$L(\lambda) = \prod_{j=0}^{m} \left(\frac{\lambda^j exp(-\lambda)}{j!} \right)^{f_j^*} \tag{4.10}$$

The algorithm is now as follows:

Choose some initial value $\hat{\lambda}$ and consider $\hat{p}_0 = \exp(-\hat{\lambda})$.

E-STEP:$\hat{f}_0 = \frac{n}{exp(\lambda)-1} \Rightarrow \hat{N} = \hat{f}_0 + n$

M-STEP: $\hat{\lambda} = \frac{\sum_{j=0}^{m} f_j^* j}{\sum_{j=0}^{m} f_j^*}$ where $f_j^* = f_j + g_j$

The same considerations for the E and the M steps of the EM algorithm can be taken for the Poisson distribution. More details about this algorithm can be found in [22].

Bellow, we discuss the application of these algorithms to the case studies presented in 2.2 using both the Binomial and the Poisson distributions.

4.1.1.1 Salmonella Data

Let us recall the Salmonella data frequency distribution of tested farms by the number of positive tests out of 7 Salmonella detection tests for the positive and the validation sample:

TABLE 4.1: Positive and validation sample for Salmonella data.

x	0	1	2	3	4	5	6	7	Total
Positive sample	?	17	9	5	6	5	5	6	53
Validation sample	3	1	3	2	3	3	4	2	21

By adopting a Binomial distribution, we obtain for the Salmonella data set the population size equal to 54 farms in total.

The EM algorithm was also applied using the Poisson distribution as mentioned before.

The results using the Binomial distribution do not differ substantially from the results using the Poisson distribution. We obtained also 1 estimated unreported farm, therefore a population of 54 farms in total.

The EM algorithm using both the Binomial distribution and the Poisson distribution seem to indicate that just 54 farms have Salmonella infection in their eggs laying flocks. In both cases, just a few steps of the algorithm are enough to converge.

The question arises if the validation sample is in fact important in the estimation of the population size. We should explore the sensitivity of reported estimates to the validation sample. In other worlds, we would like to measure the impact of the secondary sample on the observed results.

We re-run the EM algorithm without the validation sample, i.e., using just the positive sample to compare the results we achieved before by using both the positive and the validation sample. The same distributions, Binomial and Poisson, were considered.

We obtained just 2 unreported farms using the Binomial distribution and 3 using the Poisson distribution, that is between 2 to 3 unreported cases in the total sample, as reported in Table 4.2:

TABLE 4.2: Salmonella data: population size estimates.

Distribution	f_0	N
Binomial with validation information	1	54
Binomial without validation information	2	55
Poisson with validation information	1	54
Poisson without validation information	3	56

The difference between these and the previous results using the validation sample appears to be minor in this case. We should notice though that we worked with simple homogeneous models which did not allow for the presence of heterogeneity. Ignoring heterogeneity can lead to seriously underestimate the true population size.

Moreover, the fit was not adequate to provide a good estimate of the distribution due to a lack of flexibility, and the benefit of having a validation sample available is not fully exploited to check if the model is reliable.

Thus, the variability between farms with respect to, for example biosecurity issues, may play an important role.

4.1.1.2 Bowel Cancer data

For the Bowel Cancer data, the frequency distribution of the number of positive tests is reported bellow for the positive and the validation sample:

TABLE 4.3: Positive and validation sample for the Bowel Cancer data.

x	0	1	2	3	4	5	6	Total
Positive Sample	?	37	22	25	29	34	45	192
Validation Sample	22	8	12	16	21	12	31	122

The results for the Bowel Cancer data are reported in Table 4.4. It can be observed that adding the validation sample into the modelling does not have any major impact on the population size estimate considering both the Binomial and the Poisson distributions.

Again, it seems that heterogeneity related to individual features is not caught by validation information.

Only one individual is estimated using the Binomial distribution and 5 individuals using the Poisson distribution (6 if we consider both the positive and the validation samples).

TABLE 4.4: Bowel Cancer data: population size estimates.

Distribution	f_0	N
Binomial with validation information	1	193
Binomial without validation information	1	193
Poisson with validation information	6	198
Poisson without validation information	5	197

4.1.1.3 Brucellosis data

Next table recalls the positive sample (first row) and the validation sample (second row) for the data collected for the Brucellosis disease data:

TABLE 4.5: Positive and validation sample for the Brucellosis data.

x	0	1	2	3	4	5	6	Total
Positive Sample	?	57	15	14	10	4	7	107
Validation Sample	68	26	14	7	4	1	6	126

Looking at the results for the Brucellosis data set, it is clear that there is a substantial difference between using both the validation sample and the positive sample or this last one alone. In fact, using the validation information with the Binomial model, we get an estimate three times higher than the estimate we achieve using only the positive sample (21 versus 7 individuals with Brucellosis).

When using the Poisson distribution, the difference is doubled - 29 individuals when we consider the validation sample in the modelling versus 14 using the positive sample only, see Table 4.6:

TABLE 4.6: Brucellosis data: population size estimates.

Distribution	f_0	N
Binomial with validation information	21	128
Binomial without validation information	7	114
Poisson with validation information	29	136
Poisson without validation information	14	121

In this case, modelling the validation sample helps catching some unobserved heterogeneity, at least partially, we should therefore investigate if these results are more reliable than the results achieved using the positive sample only.

4.1.1.4 Heroin users data

For the Heroin users data, the positive and validation are shown bellow:

TABLE 4.7: Positive and validation sample for the Heroin users data.

x	0	1	2	3	4	5	6	Total
Positive Sample	?	1401	369	98	23	1	1	1893
Validation sample	2398	573	366	190	87	6	2	3622

The results obtained for the Heroin drug users data (see Table 4.8) also point out some contrast between using only the positive sample and both samples:

TABLE 4.8: Heroin data: population size estimates.

Distribution	f_0	N
Binomial with validation information	1212	3105
Binomial without validation information	533	2426
Poisson with validation information	1364	3257
Poisson without validation information	672	2565

The conclusions we derive by observing the results for the Brucellosis data can be applied here as well. Again, dealing with simple homogeneous distributions, the estimates obtained using the validation sample exhibit huge differences when compared to those achieved by using the positive sample.

This behaviour may point out unobserved heterogeneity in the data that the validation sample may help to identify assuring a better modelling and, consequently, a better population size estimate.

Surely, in those situations where there are notable differences between the two estimates, the validation sample holds more information and it can be used to correct the estimate for the population size.

4.1.1.5 Syphilis data

The positive and validation sample can be checked again in the next table:

TABLE 4.9: Positive and validation sample for the Syphilis data.

x	0	1	2	3	4	5	6	Total
Positive Sample	?	73	52	17	6	1	0	149
Validation sample	18	25	22	10	9	1	1	86

Once again, there is a situation where there are no differences in using the validation sample. Table 4.10 shows the estimates for the Syphilis dataset.

We obtained an estimate of 22 individuals infected with Syphilis using the Binomial distribution and 22 without the validation sample. Considering the Poisson distribution, the estimate rises to 33 using validation information and 32 otherwise.

TABLE 4.10: Syphilis data: population size estimates.

Distribution	f_0	N
Binomial with validation information	23	172
Binomial without validation information	22	171
Poisson with validation information	33	182
Poisson without validation information	32	181

In cases like the Salmonella, the Bowel Cancer and the Syphilis data, the untruncated distribution we obtain by looking at the validation sample is coherent with the truncated distribution, conditioned on model choice.

It is necessary to underline that the use of a validation sample allows us to avoid unreliable estimate for the total population size.

4.1.2 Simulation study

At this point, a question turns up, whether even in the case of homogeneity, there is any profit in using the validation information. In situations where we get a very different estimate for f_0 using the validation set, how can we check for such estimates reliability. In the following simulation study we analyse if there is a gain in efficiency using the validation sample.

We considered several $N = \{25, 50, 100, 500, 1000\}$ varying population size and fixed $\theta = \{0.15, 0.20, 0.25\}$. The data was generated randomly from a Binomial distribution and zeros were dropped; We considered $M = 1000$ different simulated sample replications.

The validation sample - 25 observations - was randomly generated each time from a Binomial distribution using the same parameters; This secondary sample was constant in all the study. It represents 100%, 50%, 25%, 5% and 2.5% of the population size when compared.

Let us now consider $X \sim B(m, p)$. A non-parametric estimate for p_0 could be obtained by the validation sample:

$$\hat{p}_0^{NP} = \frac{g_0}{N_1} \tag{4.11}$$

where g_0 is the frequency of zeros from the validation sample and N_1 is the size of the validation sample, $N_1 = \sum_{i=0}^{m} g_i$.

We can then replace this estimate in the Horvitz-Thompson estimator expression to achieve an estimate for N, the total size of the population:

$$\hat{N}^{NP} = \frac{n}{(1 - \hat{p}_0^{NP})} \tag{4.12}$$

Another choice is given by the Good-Turing estimator which is easy to calculate since it is obtained in a direct way with no need for an iterative procedure.

Assuming $p_0 = P(X = 0)$ and $p_1 = P(X = 1)$, we have:

$$\left(\frac{p_1}{E(X)}\right)^{\frac{m}{m-1}} = \left[\frac{m(1-\theta)^{m-1}\theta}{m\theta}\right]^{\frac{m}{m-1}} = (1-\theta)^m = p_0 \tag{4.13}$$

Replacing the expression $\left(\frac{p_1}{E(X)}\right)^{\frac{m}{m-1}}$ by the observed quantities for the positive and the validation sample, we obtain the following estimator:

$$\hat{p}_0^T = \left[\frac{f_1 + g_1}{\sum_{j=1}^{m}(f_j + g_j)j}\right]^{\frac{m}{m-1}} \tag{4.14}$$

Again, using the Horvitz-Thompson estimator we easily get an estimate for the total population size:

$$\hat{N}^P = \frac{n}{1 - \hat{p}_0^T} \tag{4.15}$$

We run the EM algorithm using both the positive and the validation samples or using only the positive sample. The estimation for θ and for the total size of the population N was registered for each replication. The population size N was also estimated using

either just the positive sample or both the positive and the validation sample by means of the non-parametric estimator $\hat{N}^{NP} = \frac{n}{(1-\hat{p}_0^{NP})}$ where $\hat{p}_0^{NP} = \frac{g_0}{N_1}$ and the Good-Turing estimator $\hat{N}^{P} = \frac{n}{(1-\hat{p}_0^{T})}$ where $\hat{p}_0^{T} = \left[\frac{f_1+g_1}{\sum_{j=1}^{m}(f_j+g_j)j}\right]^{\frac{m}{m-1}}$.

The positive sample was increased to 5000 replications and all the process repeated to analyse if the results were stable and to eliminate random error effects.

Table 4.11 reports the results for $\theta = 0.15$ and the associated graphical analysis. The results for the data generated 5000 times can be found between brackets in Tables 4.11 - 4.14. The results for $\theta = 0.20$ and $\theta = 0.25$ are reported in Appendix A for the EM algorithm simulation study. However, conclusions about these two parts of the study can be found here.

TABLE 4.11: Simulation study: $\theta = 0.15$ results for $M = 1000$ ($M = 5000$) samples.

N	Mean with validation	Mean without validation
25	0.1506 (0.1493)	0.1463 (0.1470)
50	0.1493 (0.1493)	0.1485 (0.1489)
100	0.1491 (0.1497)	0.1491 (0.1498)
500	0.1502 (0.1499)	0.1502 (0.1500)
1000	0.1500 (0.1499)	0.1501 (0.1499)

We can conclude from Tables 4.11, A.1 and A.5 that the mean estimated values for the parameter θ are very close to each other, specially for a big positive sample size (500 and 1000). In fact, differences between the values is negligible. The values become close as we increase N as it is suggested from the graphical analysis (see left panel of Figure 4.1 and 4.2).

We can also see from the Tables 4.12, A.2 and A.6 that the variance without using validation information is always higher than the variance using the validation sample. It is evident that the blue line is always above the red line as we can see in the Figures 4.1 and 4.2 (right panel). This means that even a small amount of validation data makes difference in the estimation of θ. Nonetheless, when we increase the size of the positive sample for 500 or 1000, the validation sample acts basically as residual information.

TABLE 4.12: Simulation study: $\theta = 0.15$ estimated variance for $M = 1000$ ($M = 5000$) samples using the EM algorithm with and without the validation sample, the Good-Turing estimator (GT) and the non-parametric estimator (NP).

N	EM alg. with validation	EM alg. without validation	GT	NP
25	0.0005 (0.0005)	0.0019 (0.0018)	1.45 (1.30)	1.57 (1.02)
50	0.0004 (0.0004)	0.0009 (0.0009)	1.72 (1.68)	2.08 (1.76)
100	0.0003 (0.0003)	0.0004 (0.0004)	1.42 (1.51)	2.27 (1.98)
500	8.1979×10^{-05} (7.8639×10^{-05})	9.2990×10^{-05} (8.6905×10^{-05})	1.74 (1.67)	2.90 (2.68)
1000	4.2494×10^{-05} (4.0760×10^{-05})	4.4637×10^{-05} (4.3092×10^{-05})	1.86 (1.85)	3.15 (3.27)

Tables 4.13, A.3 and A.7 show that major differences in terms of efficiency can be found for relatively small positive sample sizes of 25, 50 and 100. Overall, we can conclude that we have a gain in efficiency in working with the validation sample instead of working just with the positive sample.

TABLE 4.13: Ratio (Variance with validation / Variance without validation) for $\theta = 0.15$ and $M = 1000$ ($M = 5000$) samples obtained by the EM algorithm.

N	Ratio of variances
25	0.2604 (0.2955)
50	0.4530 (0.4466)
100	0.6402 (0.6349)
500	0.8816 (0.9049)
1000	0.9520 (0.9459)

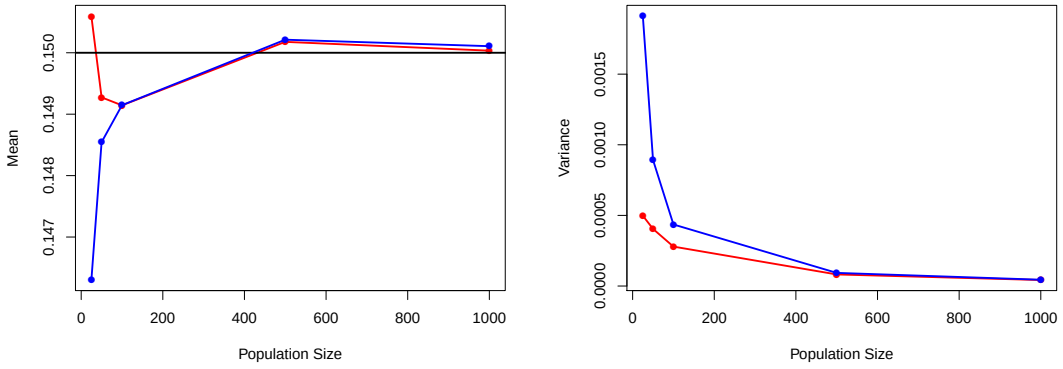

FIGURE 4.1: Simulation study: $\theta = 0.15$ and $M = 1000$ samples; Left panel: mean estimates for θ with varying N, using validation information (red) or not (blue); The true value is the black solid line; Right panel: corresponding variance values.

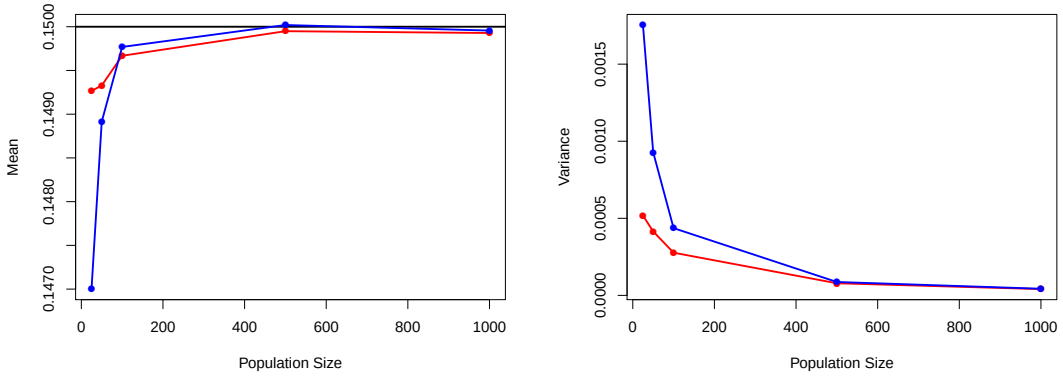

FIGURE 4.2: Simulation study: $\theta = 0.15$ and $M = 5000$ samples; Left panel: mean estimates for θ with varying N, using validation information (red) or not (blue); The true value is the black solid line; Right panel: corresponding variance values.

Let us take a look now at the estimates for N. The values obtained with $M = 5000$ samples are reported within brackets in Tables 4.11, 4.12 and 4.13. Table 4.13 also shows the results obtained for N using the Good-Turing estimator and the nonparametric estimator as described above.

TABLE 4.14: Simulation study: $\theta = 0.15$ and $M = 1000$ ($M = 5000$) samples: mean estimates for N with and without validation information obtained by the EM algorithm, the Good-Turing estimator (GT) and the non-parametric estimator (NP).

N	EM alg. without validation	EM alg. with validation	GT	NP
25	27.44 (27.14)	25.48 (25.48)	25.93 (25.32)	26.72 (26.41)
50	51.79 (51.59)	50.80 (50.62)	50.21 (50.17)	50.04 (49.89)
100	101.18 (101.39)	100.70 (100.92)	100.08 (99.87)	100.29 (100.04)
500	501.05 (500.95)	500.96 (500.97)	500.92 (500.84)	500.97 (500.21)
1000	1001.76 (1001.23)	1001.99 (1001.28)	1000.83 (1000.59)	1001.21(1000.07)

As expected, under homogeneity, we cannot find substantive differences using validation or not. However, it seems fair to say that the estimation of N seems to be relatively more accurate in the presence of a validation sample when we have a relatively small positive sample. Thus, in those cases, the bias seems to be smaller using validation information.

However, heterogeneity may play an important role in the estimation process and we should consider a validation sample to validate our results and a potential model departure from homogeneity.

Consequently, we will proceed with methodology which incorporates the validation sample in the study and focus on data heterogeneity.

4.2 The Ratio Plot

Another approach for the estimation of N under homogeneity is using a graphical device: the ratio plot. The main idea of this approach is to consider ratios of observed frequencies to estimate ratios of neighbouring count probabilities. To illustrate this idea, let us consider the Binomial distribution, and the corresponding ratios:

$$\frac{p_{x+1}}{p_x} = \frac{\binom{m}{x+1}\theta^{x+1}(1-\theta)^{m-x-1}}{\binom{m}{x}\theta^{x}(1-\theta)^{m-x}} = \frac{m-x}{x+1}\,\frac{\theta}{1-\theta} \tag{4.16}$$

Using the non-negative coefficients $a_x = \frac{x+1}{m-x}$, we can reparameterize the ratios multiplying by the inverse of their coefficients as follows:

$$r_x = \underbrace{\frac{x+1}{m-x}}_{a_x}\frac{p_{x+1}}{p_x} = a_x\frac{p_{x+1}}{p_x} = \frac{\theta}{1-\theta}. \tag{4.17}$$

The result is a constant, the odds for the event, independent of x [16]. Note that r_x does not change whether we consider the truncated or the untruncated distributions since it just depends on the parameter θ. In addiction, we emphasize that the coefficients a_x directly depend on the chosen reference distribution. In this situation, the reference is represented by the homogeneous Binomial distribution with which any mixture model of Binomial kernel reduces whenever unobserved heterogeneity is not present.

Since the quantity p_x is unknown, a non-parametric estimate f_x/N of r_x is given by:

$$r_x = a_x\frac{f_{x+1}/N}{f_x/N} = a_x\frac{f_{x+1}}{f_x} \tag{4.18}$$

where f_x is the observed frequency of units with exactly x captures, and the unknown N cancels out. Hence, r_x is a natural estimator for r_x.

The ratio plot works as a diagnostic device for the Binomial distribution [16] and depends directly on the coefficients a_x:

$$x \to r_x = a_x \frac{f_{x+1}}{f_x}. \tag{4.19}$$

Note that the coefficients a_x have large influence in the interpretation of the observed ratio plot; these coefficients change according to the reference distribution we are working with. Under the Binomial, we can expect that the ratio plot shows at least approximately an horizontal line (see Figure 3.1).

The general concept and construction of the ratio plot was already introduced in section 3.7.3. For more detailed information on this topic, please see [16], [21], [15].

Let us now consider the ratio plot for the positive sample together with the validation sample for all the case studies we have presented so far:

4.2.1 Salmonella data

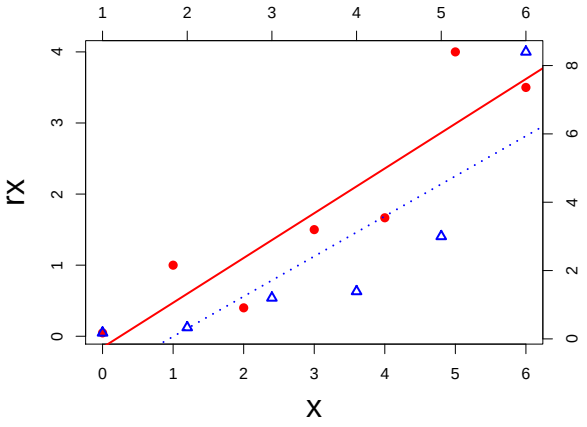

FIGURE 4.3: Salmonella data: ratio plot and estimated lines for the validation sample (red solid points) and for the positive sample (blue, empty triangles).

The graph in Figure 4.3 shows no evidence of a horizontal line pattern, whether we consider the validation or the positive sample. Instead, it shows substantial departures from the standard Binomial distribution as we can see by the monotone increasing trend. This violation of the Binomial assumption might be seen as supporting evidence for unobserved population heterogeneity translated in the figure by the non-zero slope. For instance, in the case of the Salmonella data, it could be that different farms have

different risks for a positive test result, for example related to biosecurity issues of farm factors [7].

A closer analysis of the ratio plot shows that there is something in common between the plots for the positive and the validation sample to be explored and that we can use to improve the inference on f_0. In fact, the regression lines are almost parallel for this case, which show evidence that both samples refer to distributions with similar shapes.

The fit in the case of a standard homogeneous Binomial distribution does not seem to be acceptable to the observed, zero-truncated distribution. A chi-square goodness-of-fit test confirms that we can reject the null hypothesis that the data follows a standard Binomial distribution.

For the Salmonella case study, the statistics used was $\chi^2 = \sum_{x=1}^{m-1}(\log \hat{r}_x - \log \hat{\hat{r}})^2 / \widehat{var}(\log \hat{r}_x)$, where $\widehat{var}(\log \hat{r}_x) = \frac{1}{f_{x+1}} + \frac{1}{f_x}$ and $\hat{\hat{r}} = \frac{\sum_{x=1}^{m-1} a_x f_{x+1}}{\sum_{x=1}^{m-1} f_x}$, see [16]. We found the value $\chi^2 = 46.03$ for the positive sample with 5 degrees of freedom and $\chi^2 = 5.26$ for the validation sample with 6 degrees of freedom. We can definitively reject that the data are consistent with a Binomial distribution at a significance level of $\alpha = 0.05$.

The ratio plots suggests a regression model taking advantage of the straight line pattern to determine an estimate of f_0. Namely, as $\log(r_x) = \alpha + \beta x + \epsilon_x$, an estimate of f_0 can be found using $\log\left(a_0 \frac{f_1}{f_0}\right) = \hat{\alpha} + \hat{\beta} \times 0$, or, $\hat{f}_0 = a_0 f_1 \exp(-\hat{\alpha})$.

4.2.2 Bowel Cancer data

If we consider the ratio plot for the Bowel Cancer data in Figure 4.4 for the positive and validation samples, we get no evidence of a horizontal line pattern as in the case of the Salmonella data ratio plot.

In fact, the validation sample ratios follow an increasing trend, forced mostly by the last ratio which is substantially bigger than the remaining ones for this sample. This trend is also followed by the ratios from the positive sample, even if in a smoother way. This behaviour may suggest the presence of unobserved heterogeneity in the data. Since we are dealing with human subjects here, it comes naturally that each individual reaction to the same situation may explain different probabilities of having a positive test result. It is also important to point out that any screening test is not 100% accurate which has

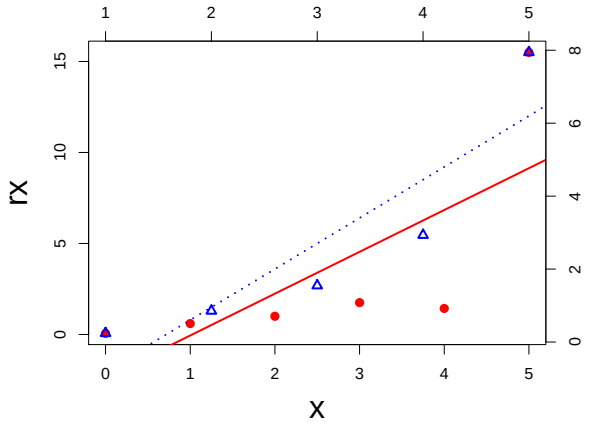

FIGURE 4.4: Bowel Cancer data: ratio plot and estimated lines for the validation sample (red solid points) and for the positive sample (blue, empty triangles).

a large influence in the data test results [30]. We can then conclude that we are not in the presence of a standard homogeneous Binomial distribution.

The analysis of the fit shown in the ratio plot for both samples, also suggests that the regression lines are almost parallel. This indicates that both samples follow distributions with similar shapes as was also the case for the Salmonella data in 4.2.1.

A chi-square goodness-of-fit test was performed for this study and it corroborates that we can reject the null hypothesis that the data follows a standard Binomial distribution.

For the Bowel Cancer case study, the statistics used was $\chi^2 = \sum_{x=1}^{m-1}(\log \hat{r}_x - \log \hat{\hat{r}})^2 / \widehat{var}(\log \hat{r}_x)$ as before, where $\widehat{var}(\log \hat{r}_x) = \frac{1}{f_{x+1}} + \frac{1}{f_x}$, $m = 6$ and we used the estimate $\hat{\hat{r}} = \frac{\sum_{x=1}^{m-1} a_x f_{x+1}}{\sum_{x=1}^{m-1} f_x}$. We recorded the value $\chi^2 = 127.91$ for the positive sample with 4 degrees of freedom and $\chi^2 = 58.22$ for the validation sample with 5 degrees of freedom. We can reject the null hypothesis that the data are consistent with a Binomial distribution at a significance level of $\alpha = 0.05$.

4.2.3 Brucellosis data

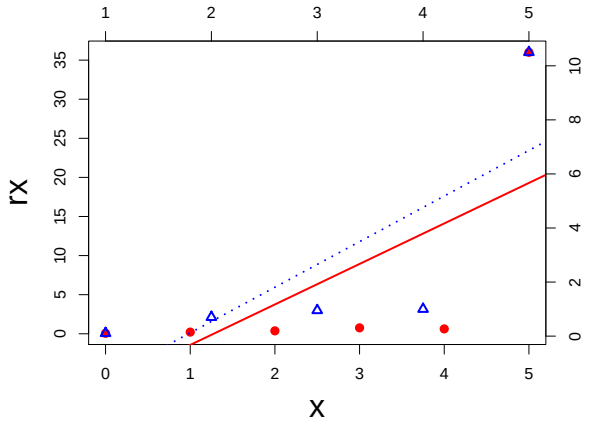

FIGURE 4.5: Brucellosis data: ratio plot and estimated lines for the validation sample (red, solid points) and for the positive sample (blue, empty triangles).

Looking at the ratio plot in Figure 4.5 for the positive and the validation samples from the Brucellosis data, it is confirmed that both do not seem to follow a standard Binomial distribution due to an increasing trend pattern.

Actually, the Brucellosis data behaves in a way which is quite similar to the Salmonella and the Bowel Cancer data. The regression lines look parallel in the graph once again, which suggests that a model of the type $\log(r_x) = \alpha + \beta x + \epsilon_x$ would perform well in this data set.

The chi-square goodness-of-fit test confirms that we can reject the null hypothesis that the data follows a standard Binomial distribution with a value $\chi^2 = 66.95$ for the positive sample with 4 degrees of freedom and $\chi^2 = 39.02$ for the validation sample with 5 degrees of freedom. We can reject that the data are consistent with a Binomial distribution at a significance level $\alpha = 0.05$.

4.2.4 Heroin users data

The graph in Figure 4.6 does not support the hypothesis that the positive sample and/or the validation sample follow a homogeneous Binomial distribution; in fact, we can observe a positive increasing pattern for the estimated lines in the graph.

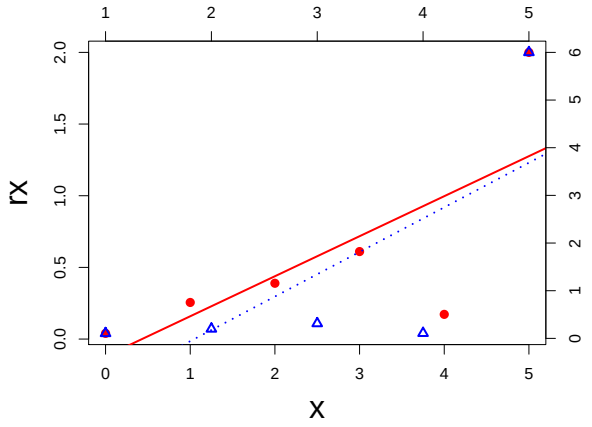

FIGURE 4.6: Heroin users data: ratio plot and estimated lines for the validation sample (red, solid points) and for the positive sample (blue, empty triangles).

The positive slope of the regression lines are a sign of the presence of population heterogeneity which, in this situation, can be explained by a failure in the registration system to contact less prone heroin users.

In this case study, the ratio plot suggests that interaction (non-parallelism between the regression lines) between the positive sample and the validation sample distributions may play an important role, suggesting to consider a regression model taking advantage of both to determine an estimate of f_0.

Specifically, as $\log(r_x) = \alpha + \beta x + \delta S + \lambda(x \times S)$ an estimate of f_0 can be found from the fitted model $\log(a_0 \frac{f_1}{f_0}) = \hat{\alpha} + \hat{\beta} \times 0 + \hat{\delta} \times 1 + \hat{\lambda} \times 0$, or, $\hat{f}_0 = a_0 \frac{f_1}{\exp(-\hat{\alpha}-\hat{\delta})}$ where S represents a dummy variable taking the value 1 if x belongs to the positive sample, else 0. The interaction term used is suggested by the non-parallelism between the regression line of the positive sample and both samples.

The chi-square goodness-of-fit test asserts that we can reject the null hypothesis the data follows a standard Binomial distribution, with a value $\chi^2 = 51.16$ for the positive sample and 4 degrees of freedom and $\chi^2 = 49.92$ for the validation sample and 5 degrees of freedom. We can reject the hypothesis that the data are consistent with a Binomial distribution at a significance level $\alpha = 0.05$.

4.2.5 Syphilis data

For this study, we re-arranged the Syphilis data for the positive and validation samples, since $f_6 = 0$ in the original positive sample data set and that would mean we would have a zero ratio in one sample due to a different number of recaptures m. This process facilitates the ratio regression study and does not cause any significant major loss of information.

The Syphilis data is then truncated at $m = 5$:

TABLE 4.15: Positive and validation sample of Syphilis data.

		Laboratories					
		0	1	2	3	4	5+
Hospital	0	?	73	52	17	6	1
	1	18	25	22	10	9	2

When $m = 5$ it is considered now the sum of counts when $m = 5$ and $m = 6$. The counts for $m = 5$ changes only for the validation sample. The ratio plot for the Syphilis data shows that any of the samples fit a standard Binomial distribution since both reveal a rising trend as stated in Figure 4.7.

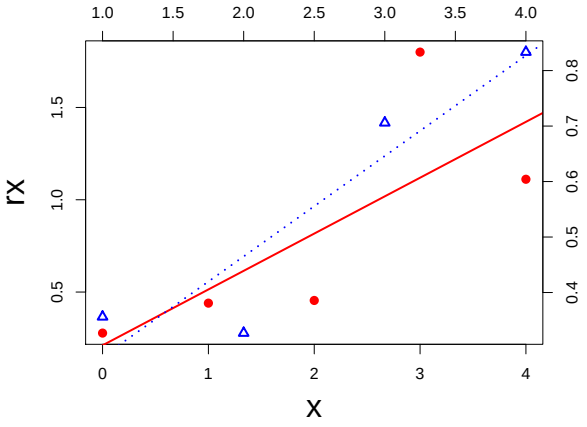

FIGURE 4.7: Syphilis data: ratio plot and estimated lines for the validation sample (red, solid points) and for the positive sample (blue, empty triangles).

In this case, the presence of interaction (i.e., the non-parallelism between the two regression lines) between the two regression lines is clear and suggests two different distributions for the positive and the validation sample. Having these aspects in mind, an estimate for f_0 can be achieved by a regression model of the type $\log(r_x) = \alpha + \beta x + \delta S$ where $S = 1$ for units in the positive sample and $S = 0$ else.

The chi-square goodness-of-fit test suggests that we can reject the null hypothesis that the data follows a standard Binomial distribution with a value $\chi^2 = 177.45$ for the positive sample and 3 degrees of freedom and $\chi^2 = 63.57$ for the validation sample and 4 degrees of freedom. According to these results, we can reject the hypothesis that the data are consistent with a Binomial distribution at a significance level $\alpha = 0.05$.

4.3 Discussion

The EM algorithm is one of the most popular algorithms to derive ML estimates in problems that involve incomplete data like capture-recapture applications. See [34], [31], [88], [90] for deep insights into the EM algorithm and applications. The main objective is to estimate the unseen part of a population of interest since some units are not observed in the population registration/identification and so data are incomplete. The algorithm was developed using the Binomial and the Poisson distribution. Analysis of the modelling process using the positive sample and incorporating the validation sample were discussed and comparisons illustrated to evaluate the role of the validation information. The ratio plot for the binomial distribution was also used as a diagnostic device to check if data heterogeneity is present and to infer about the distributions of the positive and the validation data.

The case studies presented in Chapter 2 were used to highlight the methodology. The Salmonella data, the Bowel Cancer data and lastly, the Syphilis data, do not seem to take advantage of the available validation information since differences between using validation or not are negligible. However, even under homogeneity, the estimates can lead to an under- or overestimation of the true population size since natural occurring variability is surely present in real data applications. In contrast, for the Brucellosis data and the Heroin users data the validation sample leads to more accurate estimates. We can observe huge differences when using validation validation or only the positive sample. Observing the results for these two cases, we may guess that heterogeneity is playing an important role and cannot be ignored.

Simulation studies showed that when the validation sample is embodied into the modelling process, even under homogeneity conditions, we achieve more accurate estimates

for f_0 with less bias than when only the positive sample is used. Finally, it was demonstrated that the Good-Turing estimator can be used in this context.

Overall, the use of validation information is important to confirm if the model is performing well for the unobserved part of the population which is not guaranteed if we have access only to a positive sample as well as to achieve a final estimate with less bias but still more precise. We follow up with a more flexible methodology to allow modelling heterogeneity.

Chapter 5

Mixture Models

5.1 Finite mixture models

Specific assumptions such as unit independence and parameter homogeneity are required by the simple Binomial model (as well as the simple Poisson model). Mixture models allow to relax these assumptions offering a more flexible approach in modelling heterogeneity. Also, mixture models are frequently used to handle situations when parameters vary across the population due to heterogeneity. For extensive details, see [18], [22], [71], [35] and [75].

Let $f(x, \theta)$ denote a homogeneous, parametric Binomial density function, parametrised by θ. A finite mixture of Binomial distributions is given by:

$$f(x, Q) = \sum_{j=1}^{k} f(x|\theta_j) q_j \tag{5.1}$$

It represents the distribution of X mariginalised over some discrete unobserved variable Z with distribution Q. Q is the mixing distribution and gives non-negative weights q_j to θ_j [18]. Notice that the weights are non-negative and sum to one, $\sum_{j=1}^{k} q_j = 1$.

The number of components in the mixture is given by k. If $k = 1$, we are in the presence of a homogeneous model which is the case we analysed in Chapter 4. If $k > 1$, each component k has the same density but different θ values.

$f(x, Q)$ is a mixture of component densities θ_j, when the component membership described by the latent variable Z is ignored[57], [59]. The mixing distribution Q can be seen as the heterogeneity distribution of the parameter θ of the population.

In order to model discrete counting frequencies, as it is the case presented, the probability mass function $f(x;\theta)$ will be set as the Binomial distribution.

Recalling for the Binomial distribution, we have the following form for $f(x;Q)$:

$$f(x, Q) = \sum_{j=1}^{k} q_j \binom{m}{x} \theta_j^x (1-\theta_j)^{m-x} \tag{5.2}$$

where X represents the number of times a unit was identified over the study period and $\sum_{j=1}^{k} q_j(1-\theta_j)^m$ is the probability of observing a zero-count p_0, and the mixing distribution Q has the form $Q = \begin{pmatrix} \theta_1\theta_2...\theta_k \\ q_1q_2...q_k \end{pmatrix}$ depending on the number of components k.

The number of components k is also an unknown parameter which needs to be estimated. In practice, the EM algorithm is used for fixed k, then k is increased from 1 to the maximum component size of the non-parametric maximum likelihood estimate (NPMLE). More details about the NPMLE and the EM algorithm can be found in [14] and [20].

If Q is available, we can estimate N by means of the Horvitz-Thompson estimator as $\frac{n}{(1-f(0,Q))}$ [51]. Therefore, we need to estimate Q and this will be done by maximum likelihood.

The EM algorithm for finite mixtures follows the scheme bellow [22]:

E-STEP: Given some initial value for the mixing distribution: $Q^0 = \begin{pmatrix} \theta_1^0 & \theta_2^0 & ... & \theta_k^0 \\ q_1^0 & q_2^0 & ... & q_k^0 \end{pmatrix}$, compute an updated estimate $\hat{N}$ by

$$\hat{N}^{j+1} = \frac{n}{1 - f(0, Q^j)} \tag{5.3}$$

and, further $\hat{f}_0^{\,j+1} = \hat{N}^{j+1} - n$.

M-STEP: Use the complete frequency table $\hat{f}_0^{\,j+1}$, f_1,..., f_m, to compute a new maximum likelihood estimator Q^{j+1}. Set $j = j + 1$ and go back to the E-step.

As previously shown, the EM algorithm toggles between step E, where missing data are replaced by their expected values conditional on the observed data and the current maximum likelihood parameter estimates, and step M, where the likelihood function is maximized, until the likelihood converges.

5.1.1 The EM algorithm for mixtures of Binomials

Bellow, we report the EM algorithm for mixtures of Binomial distributions with validation information.

Choose an initial value for the mixing distribution $Q^0 = \begin{pmatrix} \theta_1^0 & \theta_2^0 & ... & \theta_k^0 \\ q_1^0 & q_2^0 & ... & q_k^0 \end{pmatrix}$ and note again that $n = f_1 + f_2 + ... + f_m$ and $N_1 = g_0 + g_1 + ... + g_m$.

Let $\hat{p_0} = \sum_{l=1}^{k}(1-\theta_l)^m q_l$ and $e_{il} = \frac{f(i,\theta_l^j)q_l^j}{f(i,Q^j)}$.

E-STEP: $\hat{f}_0 = n\frac{\hat{p_0}}{1-\hat{p_0}} \Rightarrow \hat{N} = \hat{f}_0 + n$

M-STEP: $q_l^{(j+1)} = \frac{1}{\hat{N}^{j+1}} \sum_{i=0}^{m} f_i e_{il}^j$

$$\hat{\theta}_l^{(j+1)} = \frac{\sum_{j=0}^{m} f_j^* j e_{il}^j}{\sum_{j=0}^{m} m f_j^* e_{il}^j} \text{ where } f_j^* = f_j + g_j$$

Here, the unobserved latent variable indicator Z has generic element z_{il} denoting the component l each individual i belongs to. Therefore, it is replaced by e_{il} which is the posterior probability that a specific observation belongs to the l-th component. Now, the conditional likelihood is given by:

$$L(\theta) = \prod_{j=1}^{m} \left(\sum_{l=1}^{k} \frac{\binom{m}{j} \theta_l^j (1-\theta_l)^{m-j} q_l}{1 - \sum_{j=1}^{k} q_j (1-\theta_j)^m} \right)^{f_j} \times \prod_{j=0}^{m} \left(\sum_{l=1}^{k} \binom{m}{j} \theta_l^j (1-\theta_l)^{m-j} q_l \right)^{g_j} \quad (5.4)$$

which is maximized using the EM algorithm above. For completeness, the unconditional likelihood for this case is:

$$L(\theta) = \prod_{j=0}^{m} \left(\sum_{l=1}^{k} \binom{m}{j} \theta_l^j (1-\theta_l)^{m-j} q_l \right)^{f_j+g_j} \tag{5.5}$$

5.2 Model selection criteria

The question arises, which model (number of components) should we select? McLachlan [42] reviews several criteria on this topic in the context of finite mixture models. Usually, the criteria are built up in a way that the log-likelihood is penalized by a function of model complexity; for this reason, they differ in the way model complexity is measured [22].

The AIC (Akaike's Information Criterion) and the BIC (Bayesian Information Criterion) are two among those criteria that are frequently applied to choose the most appropriate model.

The AIC criterion is defined by:

$$AIC_\alpha = -2L(\hat{Q}_k) + \alpha(2k-1) \tag{5.6}$$

where k is the number of components of the finite mixture model, $L(\hat{Q}_k)$ represents the maximum log-likelihood of the model and $\alpha = 2$ is a model penalization parameter for the model complexity. Another value for α could be considered as well [69], leading to a different criterion.

The BIC criterion is defined by:

$$BIC = -2L(\hat{Q}_k) + (2k-1)log(n) \tag{5.7}$$

where n is the observed sample size.

The lower the AIC, or the lower the BIC, the better the model when comparing between different models.

The BIC penalizes complex models more strongly than the AIC [52]. However, when working with mixture models, BIC is deemed to be a better model criterion than AIC to choose the number of components. In fact, the BIC does not (asymptotically) overestimate the number of components [54], [56]. Moreover, since the BIC penalises model complexity more heavily than the AIC, it should be taken as the most appropriated criteria to choose the right model [42], [28].

In addition, it has been pointed out that AIC tends to overestimate the number of components,asymptotically [57], [89]. For these reasons, in the next section, when AIC and BIC criteria disagree on the choice for the best model, the BIC criterion will prevail.

5.3 Application of the finite mixtures estimator to the case studies

Finite mixtures have been applied to the case studies presented in section 2.2. The results for each data set follow. The code, developed in the environment *R*, can be found in Appendix B.

5.3.1 Salmonella Data

Table 5.1 displays the results of applying mixtures with 2, 3 and 4 Binomial components to the Salmonella data.

TABLE 5.1: Estimate for f_0 and for the population size N of the Salmonella case study using a Binomial mixture model of 2 components as described, the first row using the positive sample only and the second row using the positive and the validation sample.

K	Model	$\hat{f}_0$	$\hat{N}$
2	Positive	9	62
	Positive and Validation	7	60
3	Positive	20	73
	Positive and Validation	9	62
4	Positive	16	69
	Positive and Validation	10	63

Notice that since $m = 2$, the number of components of the mixture model that can be identified is not greater than 4 [82]. See also [41] for more details about identifiability.

Details on fit for the mixture models $K = 2, 3, 4$ are shown in the next table:

TABLE 5.2: Salmonella data: Model fit assessment.

Model	k	$\hat{\theta}_j$	$\hat{q}_j$	log-likelihood	AIC	BIC
Pos	1	0.4548	-	-146.05	294.10	296.07
Pos-Val	1	0.4807	-	-217.37	221.37	223.34
Pos	2	5.4665 1.2445	0.3380 0.6620	-98.7447	203.4895	209.4004
Pos-Val	2	5.4690 1.4565	0.4058 0.5942	-144.6408	295.2816	301.1925
Pos	3	6.4930 3.9974 0.7808	0.1305 0.2536 0.6159	-96.9100	203.82	213.6715
Pos-Val	3	4.2377 6.3843 1.1601	0.3047 0.1728 0.5225	-143.1297	296.2594	306.1109
Pos	4	0.8390 6.5644 4.2229 1.1620	0.4661 0.1248 0.2589 0.1502	-96.8821	207.7642	221.5562
Pos-Val	4	4.2429 6.3781 1.1182 3.8451	0.2678 0.1729 0.5222 0.0371	-143.1255	300.251	314.043

According to AIC and BIC it is clear that the mixture model with 2 components considering the positive and the validation information is the best option. The validation information seems not to impact model fit and this suggests to consider a model with just the positive sample.

5.3.2 Bowel Cancer Data

Table 5.3 shows the results of applying the EM algorithm with mixtures of 2 and 3 Binomial components to the bowel cancer data. Given that, in this case, we have $m = 6$, the highest number of components that can be readily identified is $K = 3$.

TABLE 5.3: Bowel cancer data: estimates for f_0 and for the population size N. Binomial mixture model with $K = 2, 3$ components.

K	Model	$\hat{f}_0$	$\hat{N}$
2	Positive	13	205
	Positive and Validation	21	213
3	Positive	34	226
	Positive and Validation	42	234

Assessment of model fit for the mixture models presented above with two and three components are shown in Table 5.4:

TABLE 5.4: Bowel Cancer data: model fit assessment.

Model	k	$\hat{\theta}_j$	$\hat{q}_j$	log-likelihood	AIC	BIC
Pos	1	0.6162	-	-774.9904	1551.981	1554.654
Pos-Val	1	0.5955	-	-1178.382	2358.764	2361.437
Pos	2	5.0788 1.6848	0.5270 0.4730	-342.9233	691.8466	701.6191
Pos-Val	2	1.3257 4.9431	0.4406 0.5594	-589.4105	1170.821	1180.593
Pos	3	3.8096 0.8529 5.6461	0.3679 0.3796 0.2525	-338.578	687.156	703.4435
Pos-Val	3	0.5634 3.5190 5.6987	0.3206 0.4134 0.2660	-571.0765	1152.153	1168.44

When we use the validation information, the best choice is the mixture model with 3 components, following both the AIC and BIC criteria, while these two criteria do not agree when we are using the positive sample only. Then, following the BIC criteria, the best model to use is that with only the positive sample.

5.3.3 Brucellosis Data

Table 5.5 shows the results of estimating a mixture of Binomial components to the Brucellosis data set with $K = 2, 3$ components. Also in this case, since $m = 6$, the maximum number of components that can be identified is $K = 3$.

TABLE 5.5: Brucellosis data: estimates for f_0 and for the population size N. Binomial mixture model with $K = 2, 3$ components.

K	Model	$\hat{f}_0$	$\hat{N}$
2	Positive	87	195
	Positive and Validation	106	214
3	Positive	206	314
	Positive and Validation	126	234

Assessment of model fit is obtained in the next table:

TABLE 5.6: Brucellosis data: Model fit assessment.

Model	k	$\hat{\theta}_j$	$\hat{q}_j$	log-likelihood	AIC	BIC
Pos	1	0.3350	-	-219.9097	441.8194	444.4922
Pos-Val	1	0.2597	-	-423.3088	848.6176	851.2904
Pos	2	3.9992	0.1808	-154.6407	315.2814	323.3278
		0.5711	0.8192			
Pos-Val	2	4.0093	0.1629	-334.5262	675.0524	683.0988
		0.4954	0.8371			
Pos	3	5.9845	0.0205	-149.7969	309.5938	323.0045
		3.0101	0.1354			
		0.2460	0.8441			
Pos-Val	3	0.3635	0.7825	-322.7791	655.5582	668.9689
		5.9286	0.0365			
		2.8027	0.1810			

In the case we use only the positive sample or the positive and validation samples together, a model with 3 Binomial components is the best option for these data.

5.3.4 Heroin users Data

Also in this case, $m = 6$ and we may identify $K \leq 3$ components. Table 5.7 shows the results of applying the EM algorithm with mixtures of 2 and 3 Binomial components to Heroin users data.

TABLE 5.7: Heroin users data: estimates for f_0 and for the population size N. Binomial mixture model with $K = 2, 3$ components.

K	Model	$\hat{f}_0$	$\hat{N}$
2	Positive	3600	5493
	Positive and Validation	3755	5648
3	Positive	3601	5494
	Positive and Validation	3648	5541

Assessment of model fit is reported in the next table:

TABLE 5.8: Heroin data: model fit assessment.

Model	k	$\hat{\theta}_j$	$\hat{q}_j$	log-likelihood	AIC	BIC
Pos	1	0.1743	-	-4723.725	9449.45	9454.996
Pos-Val	1	0.1115	-	-8571.805	17145.61	17151.16
Pos	2	0.3095 1.4263	0.8638 0.1362	-1433.614	2873.228	2889.866
Pos-Val	2	0.1847 1.4601	0.7441 0.2559	-5411.459	10828.92	10845.56
Pos	3	0.3111 0.2992 1.4262	0.7433 0.1205 0.1362	-1433.614	2877.228	2904.958
Pos-Val	3	0.3914 0.1898 1.6007	0.1531 0.6357 0.2112	-5406.54	10823.08	10850.81

When using the positive sample, only we should consider 2 Binomial components since the AIC and BIC both suggest this specific choice. In the case we use the validation information, the model selection criteria disagree with AIC suggesting 3 and the BIC a 2 components Binomial mixture model.

5.3.5 Syphilis Data

Also in this case, $m = 6$, and a model with $K \leq 3$ can be identified. Table 5.9 shows the results of a mixture model with Binomial kernel with $K = 2, 3$ components to the Syphilis data.

TABLE 5.9: Syphilis data: estimates for f_0 and for the population size N. Binomial mixture model with $K = 2, 3$ components.

K	Model	$\hat{f}_0$	$\hat{N}$
2	Positive	49	198
	Positive and Validation	43	192
3	Positive	49	198
	Positive and Validation	43	192

Assessment of model fit is reported in Table 5.10:

TABLE 5.10: Syphilis data: model fit assessment.

Model	k	$\hat{\theta}_j$	$\hat{q}_j$	log-likelihood	AIC	BIC
Pos	1	0.2499	-	-311.6429	625.2858	628.2897
Pos-Val	1	0.2858	-	-429.9715	861.943	864.9469
Pos	2	2.0413 1.1285	0.1854 0.8146	-168.4226	-342.8452	-351.857
Pos-Val	2	2.2432 1.1121	0.2918 0.7082	-314.9213	-635.8426	-644.8544
Pos	3	1.1551 1.9942 1.1026	0.2397 0.2048 0.5555	-168.4299	-346.8598	-361.8795
Pos-Val	3	1.1114 2.2433 1.1130	0.3971 0.2918 0.3111	-314.9212	-639.8424	-654.8621

According to the values of the AIC and the BIC criteria, the best choice for the Syphilis data is a 2 component mixture model regardless we consider the positive sample or both samples. In this specific case study, mixtures with 2 and 3 components give exactly the same results for the estimate of f_0 and consequently, N.

5.4 Finite mixtures: simulation study

We generated 1000 samples of size $N = 100$ (positive samples) and considered the validation samples with size $N_1 = 0.10N$. Data was generated from a Binomial distribution assuming $X \sim \sum_{j=1}^{2} p_j f(x|\theta_j)$, $x = 0, 1, ..., m$. We estimated a mixture model with 2 component Binomial.

The design for this simulation study is:

- $N = 100$ is the true positive sample size.
- $N_1 = 0.10N$.
- $K = 2$.
- $p_1 = p_2 = 0.5$.
- $\theta_1 = 0.2$ and $\theta_2 = 0.7$.

- $m = 7$.
- $M = 1000$ samples.

The estimate for f_0 and, consequently for N, was obtained using the mixture model just for the positive sample, using both samples (validation included) and also using a non-parametric estimate for f_0. This non-parametric estimate is based on the estimator: $\hat{N}^{NP} = \frac{n}{1-\hat{p}_0^{NP}}$ where $\hat{p}_0^{NP} = \frac{g_0}{N_1}$.

TABLE 5.11: Simulation study: distribution of estimates of the population size $N_1 = 0.1N$, $K = 2$ finite mixture.

	Positive	Pos-Val	Non-parametric estimator
Mean	102.43249	98.56646	106.45000
SD	16.801423	8.487229	16.196006

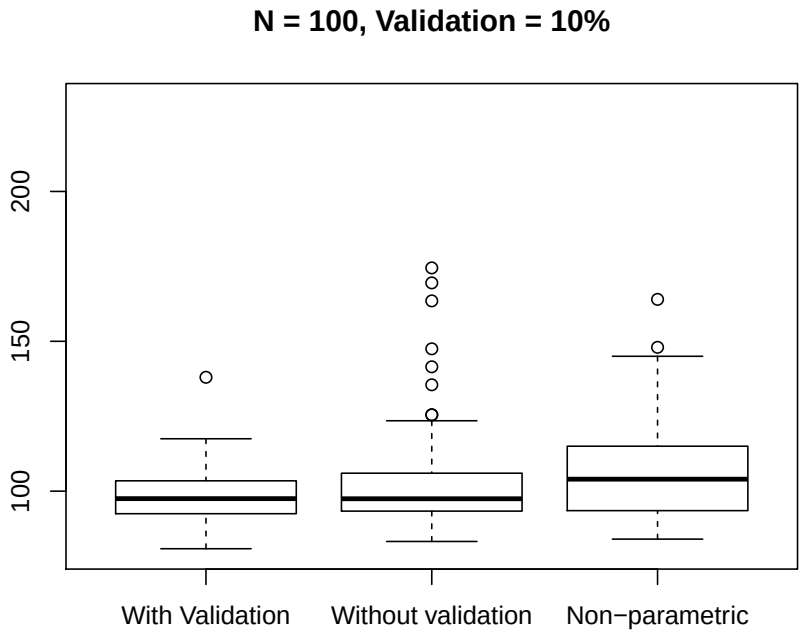

FIGURE 5.1: Boxplot of the results for the model with 10% validation (left), with just the positive sample (middle) and using the non-parametric estimator (right).

The non-parametric estimator overestimates the population size, with a (limited) bias when compared to the model using both samples. It is also the model with the highest variability which may lead to spurious estimates for N when the estimate is near the boundary. See [56], [46] for more details about the boundary problem in mixture models.

The two components mixture model estimated on the positive and the validation sample performs better in terms of both accuracy and precision when compared to other models. As Figure 5.1 shows, the model without validation may sometimes substantially overestimate the population size multiple times as it can be derived looking at the outliers in the boxplot for that model. This is a case where the modelling seems to suffer from the existence of some influential points which considerably raise the variance of the model [56]. This is also the case for the non-parametric estimator. Therefore, the mixture model using validation information outperforms the others as it is more robust in the most part of the situations.

The simulation study was after repeated now considering a validation sample with size 50% of the size of the positive sample $n = 100$. The results are reported bellow:

TABLE 5.12: Simulation study: distribution of estimates of the population size $N_1 = 0.5n$, $K = 2$ mixture components.

	Positive	Pos-Val	Non-parametric estimator
Mean	103.0408	100.4915	103.5900
SD	18.341070	7.490376	8.561713

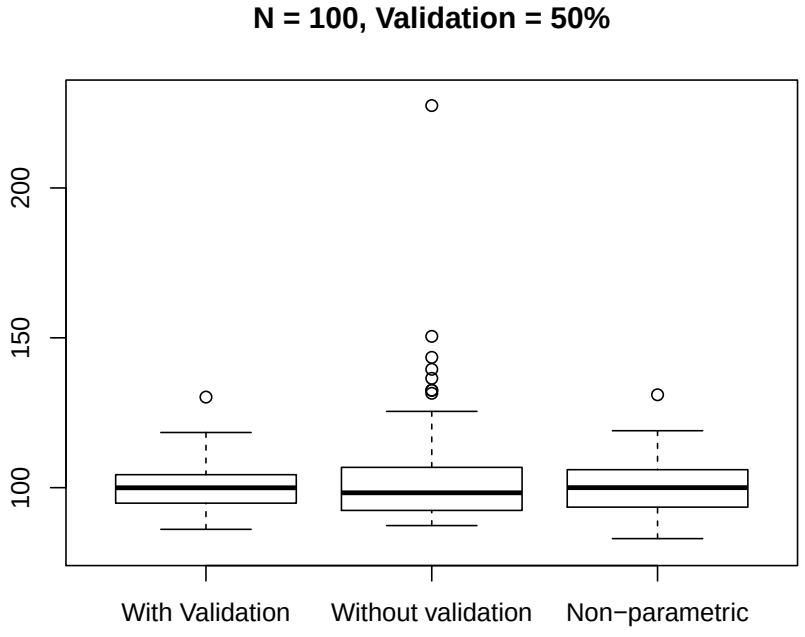

FIGURE 5.2: Boxplot of the results for the model with 50% validation (left), with just the positive sample (middle) and using the non-parametric estimator (right).

The results we achieved here allow us to conclude that using a validation sample brings mainly two important benefits: it makes the estimator more precise by substantially decreasing the bias, and increasing the confidence in the final estimate.

5.5 Simulation study: inflated data

Once the model is fixed, it can happen that data has a frequency of zeros much larger than the expected from the standard distributional assumptions. Therefore, due to this disproportional number of zeros, the model can suffer from lack of fit for the unobserved part of the population which is crucial for an accurate estimation of the true population size N.

We performed a simulation study of a Binomial with 50% zero-inflated data: there were taken $\pi N = 0.5N$ zeros and $\pi N = 0.5N$ sampling data from a homogeneous binomial distribution. The estimates for N were obtained using the positive sample only, the positive and the validation samples, the Good-Turing and the non-parametric estimator. All these estimates were then compared.

Recalling the estimates for N were obtained by a parametric estimate - the Good-Turing estimator: $\hat{N} = \frac{n}{1-\hat{p}_0^P}$, where $\hat{p}_0^T = \left[\frac{f_1+g_1}{\sum_{j=1}^{m}(f_j+g_j)j}\right]^{\frac{m}{m-1}}$ and a non-parametric estimator: $\hat{N} = \frac{n}{1-\hat{p}_0^{NP}}$, where $\hat{p}_0^{NP} = \frac{g_0}{N_1}$.

The design of the simulation study follows:

- $N = 100$, $N = 500$, $N = 1000$ and $N = 2000$ are the positive sample sizes considered in the study with $f_0 = 50$ $f_0 = 250$, $f_0 = 500$ and $f_0 = 1000$ expected zeros, respectively.
- $N_1 = N$.
- πN extra zeros and πN samples from a homogeneous binomial with parameter $\theta = 0.5$ and size parameter πN; $\pi = 0.5$.
- $m = 7$.
- $M = 1000$ samples.

We implemented a mixture model as described for the Binomial distribution with 2 components, with $m = 7$. The number of replications is $M = 1000$. Notice that the expected value for f_0 is 50% of N, i.e., $\mathbb{E}(f_0) = 0.5N$.

TABLE 5.13: Simulation study on inflated data finite mixture model with 2 components: distribution of estimates for N.

N	Mean - SD	Positive	Pos-Val	Parametric	Non-Parametric
100	Mean	50.634	100.099	50.814	104.008
	SD	0.704	0.333	1.142	0.971
500	Mean	251.956	499.911	251.458	501.892
	SD	1.138	0.568	0.811	0.783
1000	Mean	504.235	999.964	503.520	999.643
	SD	0.704	0.572	0.680	0.592
2000	Mean	1007.477	2000.018	1007.272	2000.575
	SD	1.447	0.417	0.827	0.502

Undoubtedly, the two components mixture model using both samples (positive and validation) performs the best in terms of bias and precision. The non-parametric estimator also captures the zero-inflation implemented in the data through the information given by the validation sample as expected once $\hat{p_0}$ depends on g_0 which gives that information.

The model with validation information offers the smallest bias compared to the other estimators. Clearly, the parametric estimator underestimates the true population size sharply.

Despite the fact that the parametric estimator takes into account the validation sample, there is no special correction for f_0 since it does not incorporate the zero-inflation part of the validation sample given by g_0.

Once again, the results achieved by this simulation study show that incorporating the validation information into the modelling process is essential to get a reliable estimate for the population size N with a low variance.

5.6 Discussion and conclusions

Finite mixture models represent a flexible class of models to account for heterogeneity. In this work, finite mixtures of Binomials were introduced and the parameters were estimated using the ML approach through the EM algorithm already introduced in Chapter

3 and 4. Adding information through a validation sample, which contains complete information about the data is, in this context, a novelty. We may consider either: the conditional or the unconditional likelihood. It was shown that the unconditional likelihood depends directly on N, the population size which we do not know access since we are dealing with a traditional capture-recapture problem. However, this could be done by constructing a profile mixture likelihood. It was also illustrated how the conditional likelihood is connected to the unconditional likelihood and that the first is more easily maximized. Therefore, our approach is based only on the conditional likelihood.

We used AIC and BIC to select the best model, i.e., the model with the most appropriate number of components. Several other selection criteria could be used for this purpose, for example, the likelihood ratio test.

Based on these model selection criteria, a Binomial mixture model of 2 components is the most appropriate model to estimate the total number of unreported farms for the Salmonella dataset. For the Bowel Cancer data and Brucellosis data, 3 components seem to be the most appropriate choice, whilst 2 components are needed for the Heroin users data and for the Syphilis data.

The Salmonella data was supplied by the Animal and Plant Health Agency in the UK. It is related to Salmonella infection in poultry recorded by a EU survey occurred between October 2004 and September 2005. The survey reported a prevalence of Salmonella in 11.7% of the 454 commercial layer flock holdings. Using a 2 components Binomial mixture model, we estimate 9 out of 454 farms using only the positive sample and 7 out of 454 farms using both samples, positive and validation. Arnold *et al.* [10] suggest, using a Bayesian approach on positive data, that the prevalence is 18% (82 infected farms out of 454 holdings). The results obtained with mixture models indicate a lower prevalence (13.7% using the positive sample and 13.2% using both) than the one achieved in Arnold *et al.* [10] work.

Lloyd and Frommer [61] proposed a beta heterogeneity model for the probability of a diseased individual testing positive on any single screening test and estimated the false negative fraction of the population illustrated on the Bowel Cancer data. They found an estimate $\hat{f}_0 = 61$ individuals who were disease positive but were not identified during the screening test. When considering the validation sample in the study, this value dropped to 56 individuals by fitting the beta heterogeneity model with observations from

both samples and considering a dummy variable S which is 1 if the chosen individual comes from the secondary sample, 0 else. When using a mixture model of 3 Binomial components, the estimate is still lower: 34 individuals using the positive sample and 42 using the positive and the validation.

Köse *et al.* [55] used an extension of the Lincoln-Petersen estimator to estimate the completeness of the surveillance system for Brucellosis and Syphilis. It was pointed out that there were 100 individuals with Brucellosis who were not identified by any laboratory. Considering a 3 component Binomial mixture model an estimate of 206 individuals with the disease using the positive sample were obtained and 126 individuals when validation sample takes part in the model process. Actually, the last estimate is quite close to the estimate performed by Köse *et al.*.

For the Syphilis case, Köse *et al.* [55] produced an estimate $\hat{f}_0 = 47$ who were not identified by any laboratory. This is again very close to the one obtained by the 2 component mixture model. The model estimates 49 disease positive individuals using the positive sample only and 43 using the positive and validation sample.

For the Heroin users data, a mixture model of 2 Binomial components provides an estimate $\hat{f}_0 = 3600$ of 3600 users using the positive sample and $\hat{f}_0 = 3755$ using validation. Lerdsuwansri [57] achieved an estimate of 4808 heroin users using a Non-Parametric Maximum Likelihood Estimator (NPMLe). Other estimators were used and very different estimates were achieved. For details, see [57].

Simulation studies on mixture models were designed and the results proved that the hypothesis raised in chapter 4 are verified also under heterogeneity: incorporating the validation sample, increases also in this case, efficiency, and produces more accurate estimates for the true population size.

The possibility that the positive sample might suffer from zero-inflation was also discussed. In practise, if only the positive sample is available, it remains unknown whether the model is performing well in the estimation process. In this specific case, the validation information is vital to correctly estimate N.

Chapter 6

Ratio Regression

6.1 Introduction

The main goal of any capture-recapture study is to estimate the total size of an elusive target population N when we do not have complete access to information about all the individuals that belong to the population.

The size $N = n + f_0$ is unknown since f_0, the frequency of units that we not capture any time during the study period, is unknown and this causes a reduction in the observable sample, where $n = \sum_{i=1}^{m} f_i$.

A natural way to proceed is to achieve an estimate for f_0 by means of an estimate for p_0. Using the Horvitz-Thompson estimator $\hat{N} = \frac{n}{1-p_0}$, one can see that if we get an estimate $\hat{p_0}$ for p_0, we easily reach an estimate $\hat{N}$ for N. Hence, we need to estimate p_0.

To find an estimate for p_0, we may look for a model $p_x = p_x(\theta)$ to find an estimate $\hat{\theta}$ for θ so that $\hat{p_0} = p_0(\hat{\theta})$.

In all the available datasets, described in chapter 2.2, we deal with a fixed number of sampling occasions, for example $m = 7$ for Salmonella data or $m = 6$ for the Brucellosis data. Therefore, a Binomial distribution seems to be a natural, good starting point to be considered. In addition, we are working with situations of presence/absence test for Salmonella infection, Bowel cancer, Syphilis and Brucellosis or a situation of presence/absence of a contact for heroin users and a treatment centre, consequently, the Binomial distribution seems to be the most appropriate to consider.

Let us consider the Binomial probability distribution:

$$p_x(\theta) = P(X = x) = \binom{m}{x} \theta^x (1-\theta)^{m-x} \tag{6.1}$$

$x = 0, 1, ..., m$.

Here θ represents the probability that a unit is identified at a trapping occasion, for example, the probability a Salmonella test is positive for an affected holding or the probability a heroin user is contacting the treatment centre.

We have to derive an estimate $\hat{\theta}$ for θ and use $\hat{\theta}$ in $p_0(\hat{\theta}) = (1-\hat{\theta})^m$ to estimate N, where p_0 is the probability of a zero count. An estimate for θ is usually obtained fitting a zero-truncated distribution to the available data usually through the Expectation-Maximization algorithm, see chapter 4.

However, as we are working with a simple homogeneous model, the fit may not be adequate to provide a good estimate of the distribution due to a lack of flexibility. Also, the benefit of having a validation sample is neglected, for example using this further sample to check whether the model is correct also for the unobserved part of the population. The variability associated with the assumption of homogeneity may not be appropriate in all empirical studies as unobserved heterogeneity may play an important role that should be accounted for. Ignoring heterogeneity can lead to underestimate the true population size [29], [17].

The Binomial model may not be flexible enough to provide a good fit, see section 4 for more details. As seen in Chapter 4, also through the analysis of the ratio plot, the Binomial model may not be suitable for the presented datasets.

A summary on the ratio plot based on the Salmonella data can be found in Azevedo *et al.* [11]. Here, all the datasets from chapter 2 will be used to illustrate the theory.

6.2 Ratio Regression

Let us consider a heterogeneous Binomial model with a marginal distribution given by:

$$p_x = \int_0^1 \binom{m}{x} \theta^x (1-\theta)^{m-x} h(\theta) d\theta \tag{6.2}$$

where $h(\theta)$ denotes a mixing distribution that controls departures from the homogeneous Binomial model. Notice that if $h(\theta)$ is a 1 point distribution putting the mass at θ, p_x defines a Binomial distribution with parameter θ [21].

Under general conditions, Böhning [21] proves that $R_x = a_x \frac{p_{x+1}}{p_x}$ is monotone increasing if p_x refers to a mixture model of the type of (6.2). This leads naturally to consider a model with response r_x once the marginal distribution for the Binomial distribution satisfies the monotonicity property. Specifically, the ratio plot for binomial mixtures is monotone non-decreasing.

Let us assume that R_x can be linked to a known set of predictor functions $z_0(x), ..., z_p(x)$, so that the following model is defined:

$$g(R_x) = \beta' \mathbf{z(x)} \tag{6.3}$$

where $x = 0, ..., m-1$ and g is a monotone link-function. The link-function is essentially used to guarantee that the predicted ratios remain positive, i.e., $\hat{r_x} > 0, x = 0, ..., m$. If we fit a simple straight line to the ratios r_x, this can lead to a non-feasible estimate for the ratios since we can get a negative intercept estimate as we can observe in Figures 4.3, 4.4, 4.5, 4.6 and 4.7. The choice of an appropriate link function avoids this problem; it is also shown in Böhning [21] and [15] that any regression model with the form (6.3) corresponds to a proper count distribution.

We are going to use the logarithmic function as link function. The logarithmic function is an increasing function for $z(x) \geq 1$, thus, the suggested for the regression lines is given by $\log(R_x) = \beta_0 + \beta_1 x$ with $z_0(x) = 1$ and $z_1(x) = x$; therefore, the ratios are obtained applying the inverse of the link function on both sides of the model equation: $\hat{r}_x = \exp(\hat{\beta}_0 + \hat{\beta}_1 x)$.

The estimation of the parameters β may be based on the (conditional) likelihood function:

$$L(\beta) = \prod_{x=1}^{m} \left(\frac{p_x}{1-p_0} \right)^{f_x} \tag{6.4}$$

where p_x is a function of $R_x = g^{-1}(\beta' z(x))$. However, we follow a different approach to find the estimates of β.

In detail, the scheme of this approach is in first place to generate the ratio plot by plotting x against the estimates of R_x, $r_x = a_x \frac{f_{x+1}}{f_x}$ and analyse the graph carefully. After an appropriate analysis of the ratio plot, we choose the link function g, and fit the model:

$$g(r_x) = \beta' z(x) + \varepsilon_x \tag{6.5}$$

where ε_x is such that $E(\varepsilon_x) = 0$, $cov(\varepsilon_x) = \Sigma$, while $\beta = (\beta_0, ..., \beta_p)'$ represents a $(p+1)$-dimensional vector of unknown fixed parameters, associated to the vector of regression functions $z(x) = (z_0(x), ..., z_p(x))'$. Now we can fit the model (6.5) by generalised least squares that for a Gaussian assumption on R_x implies maximising (6.4).

The first concern is to estimate Σ, see Rocchetti *et al.* [78], using the following tridiagonal matrix:

$$\begin{pmatrix}
\frac{1}{f_1}+\frac{1}{f_2} & \frac{-1}{f_2} & 0 & \cdots & 0 & \cdots & 0 \\
\frac{-1}{f_2} & \frac{1}{f_2}+\frac{1}{f_3} & \frac{-1}{f_3} & 0 & \cdots & \cdots & 0 \\
0 & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots \\
\vdots & & & \ddots & & & \\
0 & 0\cdots & \frac{-1}{f_i} & \frac{1}{f_i}+\frac{1}{f_{i+1}} & \frac{-1}{f_{i+1}} & 0\cdots & 0 \\
\vdots & & & & \ddots & & \\
0 & \cdots & & & 0 & \frac{-1}{f_{m-1}} & \frac{1}{f_{m-1}}+\frac{1}{f_m}
\end{pmatrix}. \tag{6.6}$$

It is possible to drop the off-diagonal terms of the matrix with a little loss of statistical precision for our purposes. For details, see Rocchetti *et al.* [78], Meurant [68] and Anan *et al.* [5]. Thus, we will get an estimate $\hat{\Sigma}$ of Σ determined just by the diagonal elements of the matrix above. Now, $\hat{\Sigma}$ is a diagonal matrix that contains the estimated variances given by $\frac{1}{f_i} + \frac{1}{f_{i+1}}$. The generalized least-squares estimate of β is known to be:

$$\hat{\beta} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}Y \tag{6.7}$$

where Y has elements $g(\hat{r}_x)$ and X has rows $z_0(x), ..., z_p(x)$, $x = 1, ..., m-1$, since no observation is available at $x = 0$. Note that the estimated covariance matrix of $\hat{\beta}$ is immediately available as $cov(\hat{\beta}) = (X'\hat{\Sigma}^{-1}X)^{-1}$.

In the current case, since the link function is the logarithmic function, we have:

$$Y = \begin{pmatrix} log(\hat{r}_1) \\ \dots \\ log(\hat{r}_{m-1}) \end{pmatrix} \text{ and } X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \dots & \dots \\ 1 & m-1 \end{pmatrix} \tag{6.8}$$

A regression-based estimator can be derived for the zero-count frequency as follows:

$$g(\hat{r}_0) = \hat{\beta}'z(0) \Longrightarrow \hat{r}_0 = g^{-1}(\hat{\beta}'z(0)) \tag{6.9}$$

We can now project the recurrence relation $r_x = a_x f_{x+1}/f_x$ onto $x = 0$ to obtain an estimate of f_0:

$$\hat{f}_0 = a_0 f_1/\hat{r}_0 = a_0 f_1/g^{-1}(\hat{\beta}'z(0)) \tag{6.10}$$

The population size is then obtained as the sum of the estimated number of unrecorded individuals and the size of the observed sample:

$$\hat{N}_{reg} = n + \hat{f}_0. \tag{6.11}$$

We see that the estimate for f_0 in the regression model directly depends on f_1, see (6.10). In case f_1 suffers from one-inflation, it might be better to base the estimate of f_0 on the entire distribution. Hence, the f_0 using the Horvitz-Thompson estimator could be more appropriate. The Horvitz-Thompson estimator for f_0 can be calculated as follows: $\hat{f}_0^{\,HT} = n\frac{p_0}{1-p_0}$. More information on how to achieve this result can be found in 3.7.1.1. An estimate for p_0 can be obtained as follows. We are able to estimate the probability

mass at 0 using the fitted values $\hat{r}_x = g^{-1}(\hat{\beta}' z(x))$, for R_x, $x = 0, ..., m-1$, according to the following result by Böhning [21]:

Theorem 6.1. *Let $R_x > 0$ be given for $x = 0, ..., m-1$, and let a_x, $x = 0, ..., m-1$, be known positive coefficients. Then, there exists a unique probability distribution $p_0, ..., p_m > 0$ such that:*

$$p_{x+1} = R_x \frac{p_x}{a_x}, \ \forall x = 0, ..., m-1 \tag{6.12}$$

Furthermore, we have that:

$$p_0 = \left[1 + R_0/a_0 + (R_0/a_0)(R_1/a_1) + ... + \prod_{x=0}^{m-1} R_x/a_x \right]^{-1} \tag{6.13}$$

We apply this result using estimates $\hat{r}_x$ for R_x. This result shows that any valid regression model leads to a proper probability distribution. Notice that the probability density function only depends on the model. This characteristic allows a flexible regression modelling.

Using conditioning moment techniques, it is possible to estimate the variance of $\hat{f}_0$, as shown in Böhning [21] for the Binomial case:

$$Var(\hat{f}_0) = \frac{1}{m^2} f_1 \exp(-\hat{\beta}_0)^2 (f_1 Var(\hat{\beta}_0) + 1 - f_1/(n + \hat{f}_0)). \tag{6.14}$$

An estimate for $Var(\hat{\beta}_0)$ is available from the result for $cov(\hat{\beta})$ discussed above. Thus, we provide the asymptotic 95% prediction interval for f_0 which is given by

$$\left(\hat{f}_0 - 1.96\sqrt{Var(\hat{f}_0)}, \hat{f}_0 + 1.96\sqrt{Var(\hat{f}_0)} \right) \tag{6.15}$$

Hence, a follow-up prediction interval for N also follows as

$$\left(n + \hat{f}_0 - 1.96\sqrt{Var(\hat{f}_0)}, n + \hat{f}_0 + 1.96\sqrt{Var(\hat{f}_0)} \right) \tag{6.16}$$

Until here, the presented approach covers just the analysis of the positive sample. An interesting extension that is based on incorporating the validation sample into the modelling is shown in section 6.3.

6.3 Ratio regression with validation information

The ratio regression approach can be extended to incorporate the information coming from the validation sample. Considering our data, this can be done as follows for each of the case studies:

6.3.1 Application to Salmonella data

For the Salmonella case study, let us consider the regression model as suggested by the ratio plot:

$$log(r_x) = \alpha + \beta x + \delta S + \epsilon_x \tag{6.17}$$

where S represents a dummy variable defined as $S = 1$ if x is from the positive sample and 0 otherwise. With this approach we allow a regression line for the two samples having the same slope but different intercepts on the log scale as Figure 6.1 shows. For the positive sample, we have the regression model: $\log(r_x) = (\alpha+\delta)+\beta x+\epsilon_x$ while for the validation sample: $\log(r_x) = \alpha + \beta x + \epsilon_x$. The model: $\log(\hat{r}_x) = -2.21 + 0.70x - 0.12S$ holds. The resulting estimate $\hat{f}_0 = f_1 \exp(-\hat{\alpha} - \hat{\delta}) = 25$ represents the frequency of undetected farms. Here, f_1 is the frequency of ones from the positive sample.

Note that if $\delta = 0$ both lines become identical and we allow for a single straight line regression model as Figure 6.2 shows.

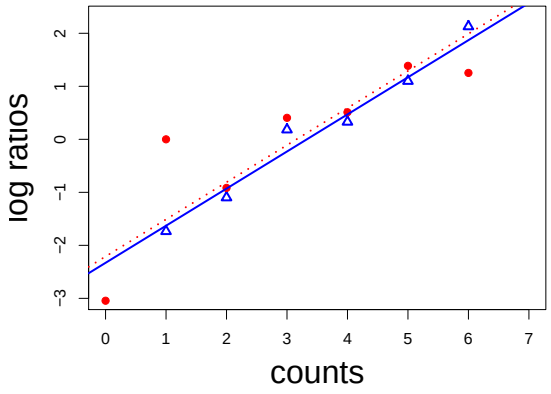

FIGURE 6.1: Salmonella data: parallel lines regression model $-2.21 + 0.70x - 0.12S$.

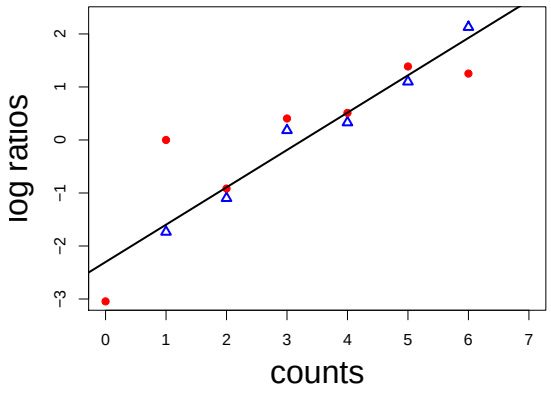

FIGURE 6.2: Salmonella data: single line regression model $-2.30 + 0.70x$.

The use of a validation sample increases the efficiency of the estimate based on the positive sample only as well as it guarantees that our model provides a reasonable final estimate, see Böhning [21]. We can also consider a model with interaction between the variable S and count x. The presence of a significant interaction between the two samples would represent that counts are not independent of the sample we consider. In case of interaction, the model becomes identical to fitting two separate lines and the benefit of the validation sample diminishes, see Figure 6.3. The model $log(\hat{r}_x) = -1.85 + 0.60x - 0.63S + 0.15(S \times x)$ holds.

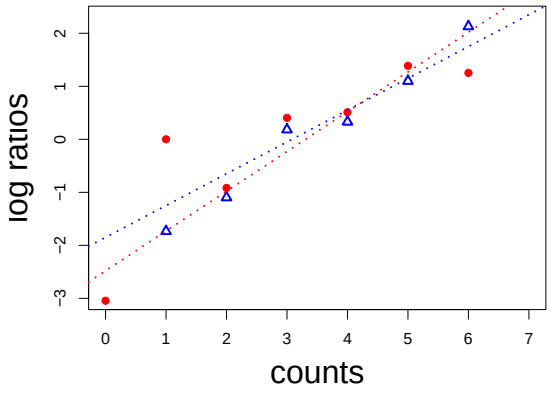

FIGURE 6.3: Salmonella data: separate lines regression model $-1.85+0.60x-0.63S+0.15(S\times x)$.

A zero-inflated model was also considered as it appears we have a large quantity of zeros in addition to those predicted by the non-inflated models. This can be seen in the ratio plot by the first ratio that is much lower than the other frequency ratios. A detailed analysis on this topic is presented in section 6.5.

We conducted simulations based on these models and the results show evidence that using the validation sample not only decreases the bias in our estimation, but also leads to more accuracy in the estimation of the population size.

A vast number of choices for regression models are possible once we consider a convenient link function for the frequency ratios. These are just three examples that seemed appropriate to explore.

The three models (single line, parallel lines, separate lines) were applied to the salmonella data and the results are presented in Table 6.1. Note that $n = 53$ for the positive sample and the coefficients a_x were set considering the Binomial distribution as the reference in our analysis.

TABLE 6.1: Salmonella data: estimates of the population size N based on different ratio regression models. The p-value refers to the last coefficient of the respective model.

Application	$\hat{f}_0$	PI for f_0	$\hat{N}$	PI for N	p-value	AIC	BIC
RR Positive	29	(1.01,56.63)	82	(54.02,109.64)	0.000		
Model 1	24	(3.65,44.90)	77	(56.65,97.90)	0.000	20.53	22.22
Model 2	25	(1.49,48.35)	78	(54.49,101.35)	0.660	22.26	24.52
Model 3	29	(5.98,51.68)	82	(58.98,104.68)	0.316	22.73	25.55

We obtained 29 undetected farms using just the positive sample. Model 3 provides exactly the same results as expected. The interaction term is not significant in model 3. The simple regression model (model 1) and the parallel lines model (model 2) produce a very similar result. Model 1 indicates 24 undetected farms while model 2 suggests 25 undetected farms. Table 6.1 includes the estimates for the coefficients of each model as well as prediction intervals for each estimate. As model 2 has a non-significant term for S, we conclude that model 1 is most suitable in our case and the estimate is $\hat{f}_0 = 24$. When comparing models fit to the same data, the smaller the AIC or BIC, the better the fit [27]. In this case, AIC and BIC criteria corroborate that model 1 is the most appropriate in our case study.

The results shown in this section are included in Azevedo *et al.* [11] of "Capture-Recapture Methods for the Social and Medical Sciences" book. In fact, the number of undetected farms/individuals may be much greater than the results we obtained using the various methods discussed in this work. However, a positive detection probability is assumed by the ratio regression approach. If this does not occur, a lower bound for the estimation of unreported farms/misclassified individuals was determined which it is necessary to discuss with the responsible authorities for these health concerns.

6.3.2 Application to Bowel Cancer data

For the Bowel Cancer data, we can point out the same that was said for the Salmonella data. In Figure 6.4 it is illustrated the regression model $-2.15 + 0.86x$:

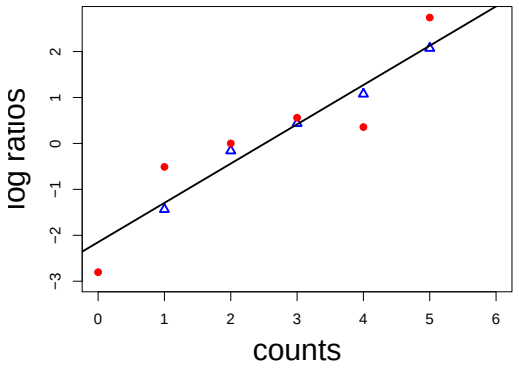

FIGURE 6.4: Bowel Cancer data: single line regression model $-2.15 + 0.86x$.

The resulting estimate $\hat{f}_0 = f_1 \exp(-\hat{\alpha}) = 53$ denotes the number of individuals for which the Bowel Cancer screening test failed in identifying the disease. This result turns out to be the most significant estimate that we could find with the ratio regression for this study.

A parallel lines model $\log(\hat{r}_x) = -2.10 + 0.86x - 0.10S$ was also estimated for this case study and it is presented below in Figure 6.5:

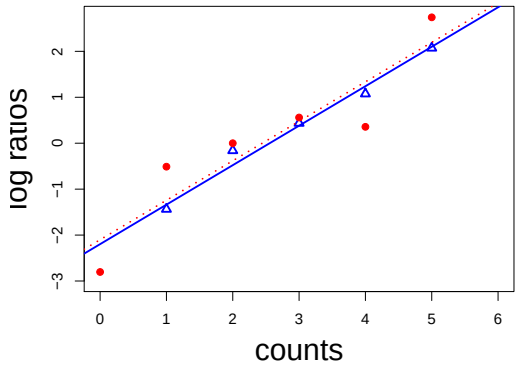

FIGURE 6.5: Bowel Cancer data: parallel lines regression model $-2.10+0.86x-0.10S$.

As it can be easily observed, the intercept of the two lines is almost the same falling in the case of a single straight line as above in Figure 6.4.

Finally, the two separate lines model $log(\hat{r}_x) = -2.21 + 0.90x - 0.11S - 0.07(S \times x)$ shows that in this case, the positive and the validation sample are coherent with the same model.

The results of the three models for the Bowel Cancer data are reported in Table 6.2. Here, $n = 192$ for the positive sample, while the coefficients refer to the Binomial distribution as reference model.

TABLE 6.2: Bowel Cancer data: estimates of the population size N based on different ratio regression models. The p-value refers to the last coefficient of the respective model.

Application	$\hat{f}_0$	PI for f_0	$\hat{N}$	PI for N	p-value	AIC	BIC
RR Positive	51	(34.90,66.02)	243	(226.90,258.02)	0.001		
Model 1	53	(26.56,79.35)	245	(218.56,271.35)	0.000	17.86	19.06
Model 2	56	(28.54,82.47)	248	(220.54,274.47)	0.751	19.71	21.31
Model 3	51	(13.66,87.25)	243	(205.66,279.25)	0.733	21.52	23.51

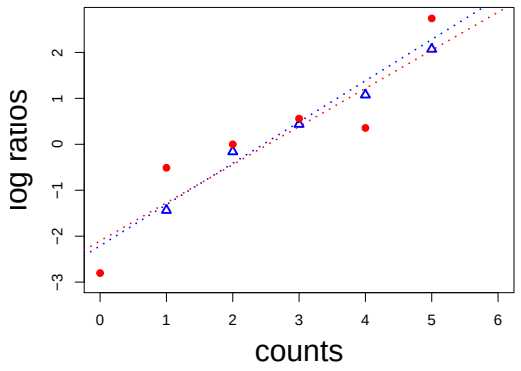

FIGURE 6.6: Bowel Cancer data: separate lines regression model $-2.21 + 0.90x - 0.11S - 0.07(S \times x)$.

As it was already mentioned, using the two separate model regression lines, guides us to the same result as using the positive sample only: $\hat{f}_0 = 51$ individuals not identified by the screening test for the bowel cancer.

The estimate for f_0 using model 3 has the largest prediction interval and the last term - the interaction term between S and x- is not significant. The last term in model 2, associated to the dummy variable S, is also non-significant with the best estimate for f_0 achieved by the straight line model with $\hat{f}_0 = 53$ and $\hat{N} = 245$, having the shortest prediction interval.

6.3.3 Application to the Brucellosis data

The results we obtain by estimating the ratio regression model on the Brucellosis data, do not differ from the results for the two previous cases (Salmonella and Bowel Cancer data). In fact, the straight line regression model $\log(\hat{r}_x) = -2.78 + 0.93x$ produces the best estimate of $\hat{f}_0 = 153$ misclassified individuals with Brucellosis disease considering the AIC and BIC criteria. The model is illustrated in the following figure:

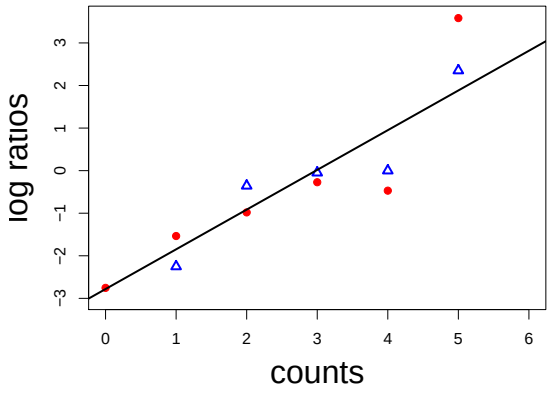

FIGURE 6.7: Brucellosis data: single line regression model $-2.78 + 0.93x$.

By estimating a parallel lines model $\log(\hat{r}_x) = -2.74 + 0.97x - 0.20S$ we have no substantial gain as shown in Figure 6.8:

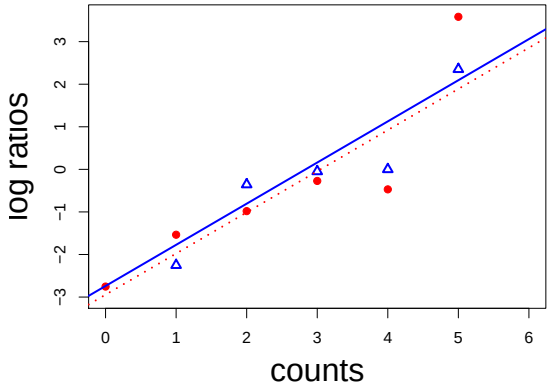

FIGURE 6.8: Brucellosis data: parallel lines regression model $-2.74 + 0.97x - 0.20S$.

If we fit a separate lines regression model, we obtain two regression lines that basically overlap, highlighting that we are using essentially the positive sample. Figure 6.9 illustrates the results.

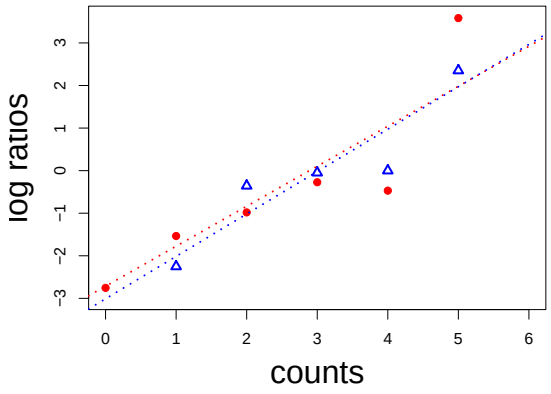

FIGURE 6.9: Brucellosis data: separate lines regression model $-2.71+0.94x-0.30S-0.06(S\times x)$.

Conclusions for model 3 $\log(\hat{r}_x) = -2.71+0.94x-0.30S-0.06(S\times x)$ can be confirmed by Table 6.3 where we get the same estimate for f_0 by the ratio regression using the positive sample and model 3, but with a lower validity. Note that $n = 107$. Model 1 turns out to be the most appropriate model to estimate f_0 with all terms significant and the best values achieved for AIC and BIC criteria; With $\hat{f}_0 = 153$ we obtain a population size of 260 individuals.

TABLE 6.3: Brucellosis data: estimates of the population size N based on different ratio regression models. The p-value refers to the last coefficient of the respective model.

Application	$\hat{f}_0$	PI for f_0	$\hat{N}$	PI for N	p-value	AIC	BIC
RR Positive	193	(47.90,337.63)	300	(154.90,444.63)	0.019		
Model 1	153	(89.08,217.55)	260	(196.08,324.55)	0.000	23.25	24.45
Model 2	180	(116.48,244.40)	287	(223.48,351.40)	0.565	24.77	26.36
Model 3	193	(125.80,259.76)	300	(232.80,366.76)	0.830	26.69	28.68

6.3.4 Application to the Heroin users data

In the case of the Heroin users data, the straight line model $\log(\hat{r}_x) = -0.42 - 0.09x$ seems not to provide an appropriate fit to the data, as shown in Figure 6.10:

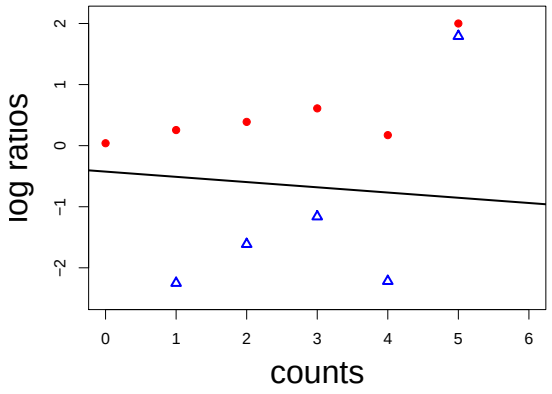

FIGURE 6.10: Heroin users data: single line regression model $-0.42 - 0.09x$.

However, with the parallel lines model $\log(\hat{r}_x) = 0.01 + 0.23x - 2.38S$, the situation changes substantially. This model seems to perform quite well with both the positive and validation sample datasets:

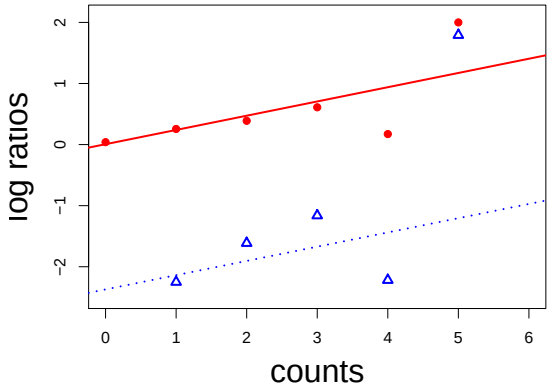

FIGURE 6.11: Heroin users data: parallel lines regression model $0.01 + 0.23x - 2.38S$.

Figure 6.12 reports the two separate lines model $\log(\hat{r}_x) = 0.05 + 0.18x - 2.87S + 0.40(S \times x)$ for the Heroin users data; Also in this case, the fit seems appropriate, with a significant interaction term.

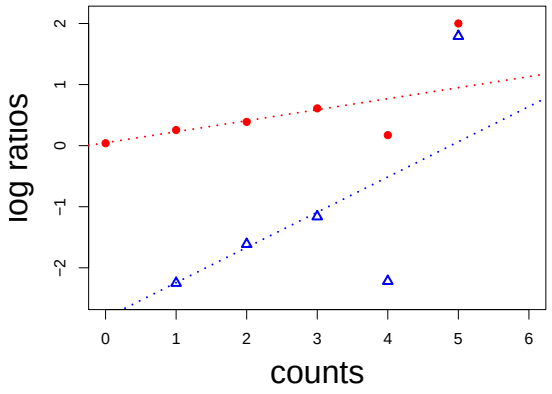

FIGURE 6.12: Heroin users data: separate lines regression model $0.05+0.18x-2.87S+0.40(S \times x)$.

We can also suspect that we might be in the presence of zero-inflated data that is suggested by the negative intercept of the positive sample regression line.

Table 6.4 presents the results for all the models we have estimated on this dataset:

TABLE 6.4: Heroin users data: estimates of the population size N based on different ratio regression models. The p-value refers to the last coefficient of the respective model.

Application	$\hat{f}_0$	PI for f_0	$\hat{N}$	PI for N	p-value	AIC	BIC
RR Positive	3919	(2945.26,4891.89)	5812	(4838.26,6784.89)	0.013		
Model 1	357	(344.11,370.12)	2250	(2237.11,2263.12)	0.830	56.05	57.25
Model 2	2501	(2479.66,2522.69)	4384	(4372.66,4415.69)	0.000	14.08	15.67
Model 3	3919	(3897.28,3939.93)	5812	(5790.28,5832.93)	0.006	3.62	5.61

Both models 2 and 3 have the last term of the model significant with model 3 performing better in terms of AIC and BIC values and with a slightly shorter prediction interval than the one for model 2. In this case, using a validation sample will not bring an extra advantage in the estimation of f_0 since we get exactly the same estimate if we use only the positive sample.

For this case, an estimate of $\hat{f}_0 = 3919$ for the number of heroin users that were not registered by the treatment centre is quite high, specially if we consider that $n = 1893$. This can be an indication for this data suffering from zero-inflation. This conclusion is corroborated by comparing the results with only the positive sample and model 3 using both samples.

6.3.5 Application to the Syphilis data

Finally, we applied the same ratio regression approach to the syphilis data. Figure 6.13 shows the single line regression model $\log(\hat{r}_x) = -1.35 - 0.33x$ for this case study:

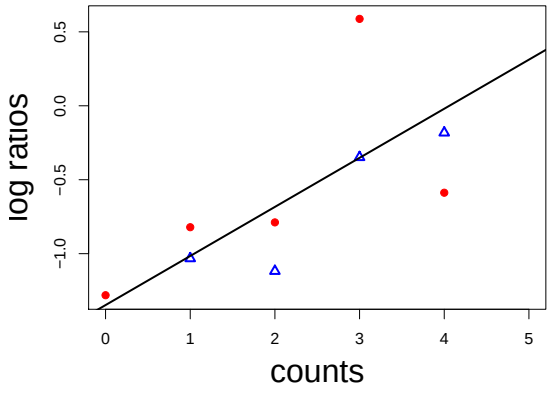

FIGURE 6.13: Syphilis data: single line regression model $-1.35-0.33x$ for the syphilis data set.

The estimated parallel lines regression model $\log(\hat{r}_x) = -1.20+0.35x-0.31S$ is reported in Figure 6.14:

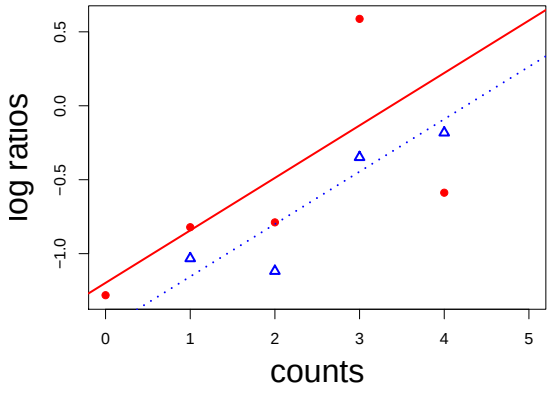

FIGURE 6.14: Syphilis data: parallel lines regression model $-1.20 + 0.35x - 0.31S$.

Lastly, as it is shown in Figure 6.15, the two separate lines regression model $\log(\hat{r}_x) = -1.30 + 0.44x - 0.01S - 0.21(S \times x)$ clearly points for a strong impact of the interaction $S \times x$.

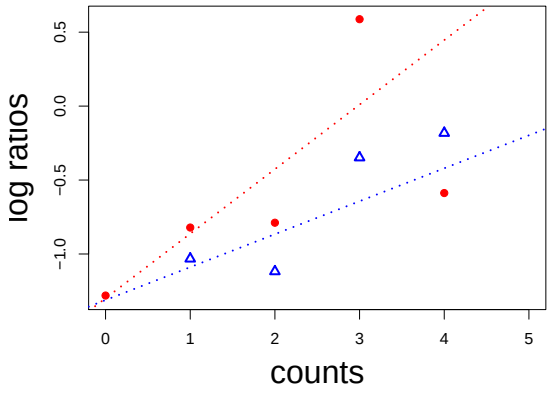

FIGURE 6.15: Separate lines regression mode $-1.30 + 0.44x - 0.01S - 0.21(S \times x)$ for the syphilis data set.

TABLE 6.5: Syphilis data: estimates of the population size N based on different ratio regression models. The p-value refers to the last coefficient of the respective model.

Application	$\hat{f}_0$	PI for f_0	$\hat{N}$	PI for N	p-value	AIC	BIC
RR Positive	54	(18.09,90.32)	203	(167.09,239.32)	0.299		
Model 1	56	(32.14,80.23)	205	(181.14,229.23)	0.032	14.40	14.99
Model 2	66	(15.45,116.73)	215	(164.42,265.73)	0.198	13.71	14.50
Model 3	54	(25.91,82.49)	203	(174.91,231.49)	0.422	14.42	15.41

The results for the three models are displayed in Table 6.5. Here, $n = 149$. Based on a first look at the table and the values for the AIC and BIC criteria, model 2 gives a more trustworthy estimate for f_0. However, we can see that the term for S is non-significant and that this model is not substantially different from model 1. In fact, the AIC and BIC for this model are slightly bigger than for model 2. Nevertheless, the shortest prediction interval associated with model 1 estimations makes this a preferred option to model 2. Therefore, the estimate $\hat{f}_0$ produced by model 1 is more reliable than the estimate we got using model 2.

6.4 Simulation study

A question arises about the real benefit in using the validation sample in the ratio regression modelling. A natural way to assess this question is to investigate the performance of each model above in the presence and absence of a validation sample through a simulation study. We are interested in simulating data with a close behaviour to the datasets

we have discussed. We have generated 1000 samples for positive samples in which all the 0 units were considered as missing values and discarded. Another 1000 sample replications were generated for validation samples, each one is paired with a positive sample. Note that all the samples will have a fixed number of 7 sampling occasions.

The population size N for the positive samples varied among 25, 50, 100, 500 and 1000, as well as for the validation sample size N_1. We calculated the population size N using only the positive sample and incorporating the information from the validation sample. The Horvitz-Thompson estimate was also considered in the study for comparison purposes.

6.4.1 Single Line Model Simulation Study

We present in this section the results for the simulation study based on the single line model. We set $\alpha = -2$ and $\beta = 0.6$ and construct the model $log(r_x) = \alpha + \beta x = -2 + 0.6x$. After that, we can easily find the ratios $r_x = \exp[\alpha + \beta x]$. Using (6.12) from the Theorem 6.1, we find p_0 and using the relation $p_x = \frac{r_x}{a_x} p_{x-1}$ for $x = 1, ..., 7$, we derive the probabilities $p_1, ..., p_7$, which determine the count distribution $P(X = x) = p_x$ for $x = 0, 1, ..., 7$.

In table 6.6, we report the estimator derived by the ratio regression ($\hat{N}$ Positive), for the positive sample, the Horvitz-Thompson ($\hat{N}$ HT) derived by the positive sample and the single line estimate ($\hat{N}$ SLM).

TABLE 6.6: Mean and variance for positive sample estimators from a population size $N = 25$, $N = 50$, $N = 100$ and $N = 1000$.

		$\hat{N}$ Positive	$\hat{N}$ HT	$\hat{N}$ SLM
$N = 25$	Validation sample size: 25	.	.	.
	Mean	51.37	51.35	50.42
	Variance	74.30	85.04	42.22
$N = 50$	Validation sample size: 25	.	.	.
	Mean	101.40	101.21	100.71
	Variance	132.83	144.76	94.94
	Validation sample size: 50	.	.	.
	Mean	101.40	101.15	100.41
	Variance	130.64	135.43	79.25
$N = 100$	Validation sample size: 25	.	.	.
	Mean	501.12	500.99	500.79
	Variance	578.15	613.67	536.32
	Validation sample size: 50	.	.	.
	Mean	501.30	501.14	501.19
	Variance	609.73	642.24	534.70
	Validation sample size: 100	.	.	.
	Mean	502.40	502.32	501.46
	Variance	593.41	629.47	462.94
$N = 1000$	Validation sample size: 25	.	.	.
	Mean	1000.43	1000.29	1000.11
	Variance	1198.88	1254.86	1146.99
	Validation sample size: 50	.	.	.
	Mean	1001.35	1001.10	1001.10
	Variance	1105.18	1163.92	1021.33
	Validation sample size: 100	.	.	.
	Mean	1002.38	1002.17	1001.66
	Variance	1247.03	1326.19	1087.35
	Validation sample size: 1000	.	.	.
	Mean	1001.79	1000.64	1000.15
	Variance	1045.35	1093.94	709.89

The estimates for N using the single regression model in the presence of a validation sample is always more accurate than using the ratio regression approach with only the positive sample. It can also be stated that the estimates given by the Horvitz-Thompson approach is consistently closer to the true value than the estimation using only the positive sample, even if it is associated to a larger variance. Also, it is shown that the variance using the model incorporating the validation information is smaller than the other two, revealing that the main differences are in terms of efficiency. The gain in efficiency is clear with a validation sample.

6.4.2 Parallel Lines Model Simulation Study

The results for the simulation study based on a parallel lines model are illustrated in this section. We set $\alpha = -2$, $\beta = 0.7$ and $\delta = -0.5$ and construct the model $log(r_x) = \alpha+\beta x+\delta S = -2+0.7x-0.5S$. This time the ratios are defined by $r_x = \exp[\alpha+\beta x+\delta S]$, with $S = 1$ for units in the positive sample, 0 else. Once more, we find p_0 using equation (6.12) and further probabilities by the recurrence $p_x = \frac{r_x}{a_x} p_{x-1}$ for $x = 1, ..., 7$. The count distribution $P(X = x) = p_x$ for $x = 0, 1, ..., 7$ is used to simulate the observed population estimators of the populaion size.

In the table below, we report the estimator derived by the ratio regression ($\hat{N}$ Positive), for the positive sample, the Horvitz-Thompson ($\hat{N}$ HT) derived by the positive sample and the parallel lines estimate ($\hat{N}$ PLM).

The estimate of N is more accurate and calculated with more efficiency using the validation sample and modelling both samples distributions using a parallel lines model, independently of the proportion of the validation sample available. In fact, once again, the Horvitz-Thompson estimator performs better when we use the positive sample only, even if the efficiency declines strongly.

TABLE 6.7: Mean and variance for positive sample estimators from a population size $N = 25$, $N = 50$, $N = 100$ and $N = 1000$.

		$\hat{N}$ Positive	$\hat{N}$ HT	$\hat{N}$ PLM
$N = 25$	Validation sample size: 25	.	.	.
	Mean	53.07	52.86	48.88
	Variance	320.65	382.01	154.00
$N = 50$	Validation sample size: 25	.	.	.
	Mean	105.96	105.00	102.66
	Variance	691.12	737.01	409.89
	Validation sample size: 50	.	.	.
	Mean	102.60	101.67	100.15
	Variance	470.65	523.74	301.15
$N = 100$	Validation sample size: 25	.	.	.
	Mean	502.38	501.28	500.44
	Variance	2187.31	2342.45	1937.02
	Validation sample size: 50	.	.	.
	Mean	504.41	503.48	502.61
	Variance	2373.50	2540.94	1978.10
	Validation sample size: 100	.	.	.
	Mean	503.03	502.19	501.71
	Variance	2065.83	2210.27	1622.29
$N = 1000$	Validation sample size: 25	.	.	.
	Mean	1001.84	1000.61	1000.06
	Variance	4340.12	4659.68	4114.06
	Validation sample size: 50	.	.	.
	Mean	1003.02	1002.07	1001.67
	Variance	4227.55	4564.76	3833.54
	Validation sample size: 100	.	.	.
	Mean	1002.33	1001.84	1001.56
	Variance	4666.76	4988.44	3987.99
	Validation sample size: 1000	.	.	.
	Mean	1004.63	1003.36	1000.60
	Variance	4311.05	4584.05	2864.56

6.4.3 Separate Lines Model Simulation Study

We report the results for the simulation study based on the separate lines model. We set $\alpha = -2$, $\beta = 0.6$, $\delta = -0.5$ and $\lambda = 0.1$ and construct the model $log(r_x) = \alpha + \beta x + \delta S + \lambda(S \times x) = -2 + 0.6x - 0.5S + 0.1(S \times x)$. The ratios $r_x = \exp[\alpha + \beta x + \delta S + \lambda(S \times x)]$ are found, where $S = 1$ if the unit belongs to the positive sample, 0 otherwise. Using the relation $p_x = \frac{r_x}{a_x} p_{x-1}$ for $x = 1, ..., 7$ we find all the probabilities $p_1, ..., p_7$ while p_0 is calculated using equation (6.12). The results are shown in the next tables, where the estimators based on the positive sample regression ($\hat{N}$ Positive), Horvitz-Thompson

based on the p_0 derived from the positive sample ($\hat{N}$ HT) and the separate lines model ($\hat{N}$ SepLM) are reported.

TABLE 6.8: Mean and variance for positive sample estimators from a population size $N = 25$, $N = 50$, $N = 100$ and $N = 1000$.

		$\hat{N}$ Positive	$\hat{N}$ HT	$\hat{N}$ SepLM
$N = 25$	Validation sample size: 25	.	.	.
	Mean	53.13	53.03	53.13
	Variance	282.21	332.82	282.21
$N = 50$	Validation sample size: 25	.	.	.
	Mean	102.34	101.40	102.34
	Variance	490.54	532.39	490.54
	Validation sample size: 50	.	.	.
	Mean	104.33	103.33	104.33
	Variance	540.39	579.53	540.39
$N = 100$	Validation sample size: 25	.	.	.
	Mean	501.71	500.55	501.71
	Variance	1985.23	2124.98	1985.23
	Validation sample size: 50	.	.	.
	Mean	501.75	500.86	501.75
	Variance	2382.47	2537.48	2382.47
	Validation sample size: 100	.	.	.
	Mean	503.03	501.87	503.05
	Variance	2226.05	2369.86	2226.05
$N = 1000$	Validation sample size: 25	.	.	.
	Mean	1004.59	1003.43	1004.59
	Variance	4695.25	4996.80	4695.25
	Validation sample size: 50	.	.	.
	Mean	1001.35	1000.14	1001.35
	Variance	4195.39	4466.94	4195.39
	Validation sample size: 100	.	.	.
	Mean	1004.50	1003.55	1004.50
	Variance	4310.63	4665.36	4310.63
	Validation sample size: 1000	.	.	.
	Mean	1002.71	1001.68	1002.71
	Variance	4138.85	4465.80	4138.85

For the separate lines regression model, the estimate for N we obtain using the validation sample is exactly the same we get using only the positive sample, which makes the use of validation information useless in this situation. Despite the Horvitz-Thompson estimator gives a closer estimate for the true value, it loses in terms of efficiency, which therefore makes the use of only the positive sample more reliable in this kind of situations since the associated variance is consistently smaller.

6.5 The inflated model

The previous modelling approaches do not account for any zero-inflation. It is possible that data has a number of non-observed cases much higher than expected which would lead to a first ratio that is potentially much lower than the others. To account for zero-inflation, at least approximately, we suggest to model the ratio by a quadratic form of this kind: $log(R_x) = \alpha + \beta x + \delta S + \lambda x^2$. This model may allow if necessary a bend in the straight line corresponding to the positive sample and, at the same time, to take advantage of the available validation sample.

The question arises, if this kind of approach may perform well on the previous datasets. As it turns out, the quadratic term is not significant in any of the cases and the corresponding model is not the best model to consider to any of the datasets than the one reported in section 6.3.

TABLE 6.9: Estimates of the population size N for the zero-inflated model according to the each model equation.

Data	$\hat{f}_0$	PI for f_0	$\hat{N}$	PI for N	p-value x^2	AIC	BIC
Salmonella	33	(-8.36,73.63)	86	(44.64,126.63)	0.437	23.34	26.16
Bowel Cancer	65	(24.67,106.20)	257	(216.67,298.20)	0.692	21.45	23.44
Brucellosis	174	(104.38,243.12)	281	(211.38,350.12)	0.889	26.74	28.73
Heroin users	11030	(10468.88,11592.01)	12923	(12361.88,13485.01)	0.562	15.51	17.50
Syphilis	63	(36.07,89.86)	211	(184.07,237.86)	0.775	13.91	14.90

For the Salmonella data, a total of 33 undetected farms were obtained employing the model equation $log(r_x) = -2.47 + 0.94x - 0.13S - 0.04x^2$ as Table 6.9 shows. In other words, a population size of 86 farms. In fact, the best model for the Salmonella data is the single line model. AIC and BIC criteria support that statement, since the values are bigger for this model than for the other three discussed models of Table 6.1.

For the Bowel Cancer data, the estimated model equation is $log(r_x) = -2.25 + 1.02x - 0.11S - 0.03x^2$. According to such equation, we get an estimate $\hat{f}_0 = 65$ which means that according to the model, 65 individuals were misclassified during the screening test. However, the single line model has lower values of AIC and BIC criteria and a shorter predictive interval so this is not a more appropriate model to consider for this data set.

The estimated regression equation for the Brucellosis data follows: $log(r_x) = -2.72 + 0.92x - 0.19S - 0.01x^2$. Again, the model applied to this data does not produce a better fit to the data or a better estimate for f_0, based on the AIC and BIC criteria as well as

on the prediction interval. The best model for this data is still model 1, the single line model.

For the Heroin users data, the estimated equation is $log(r_x) = 0.03 + 0.14x - 2.34S + 0.03x^2$. The model has the quadratic term non-significant and a much larger AIC, BIC and prediction interval. The model we get the best results for the Heroin users data is model 3, the separate lines regression model despite the suspicion of zero-inflation based on the first ratio of the ratio plot as discussed.

Finally, for the Syphilis dataset, we get the following estimated regression equation: $log(r_x) = -1.16 + 0.27x - 0.30S - 0.03x^2$. By using the inflated model, we achieved $\hat{f}_0 = 63$ individuals estimated as not been identified with Syphilis by any treatment centre. Comparing the results of this model with the results shown in Table 6.5 we can state that model 1 (the single line model) performs better in this data set using the same arguments as before: lower AIC and BIC and shorter confidence interval.

We conducted simulations that show that the estimation of N using the quadratic model, with the validation sample incorporated, may in some cases produce results that are substantially better in terms of precision, and bias.

6.6 Simulation study on zero-inflated data

There is no indication that our data suffers zero-inflation, but it could actually happen and we may know only in the case we have a validation sample. We performed a simulation study entailing a Binomial zero-inflated distribution and obtained the estimates for f_0 using different models after the analysis of the ratio plot. The simulation study covered the following situations:

- Case 1: Positive sample size: 100 (50 zeros); Validation sample size: 100 (50 zeros).
- Case 2: Positive sample size: 500 (250 zeros); Validation sample size: 500 (250 zeros).
- Case 3: Positive sample size: 1000 (500 zeros); Validation sample size: 1000 (500 zeros).

- Case 4: Positive sample size: 2000 (1000 zeros); Validation sample size: 2000 (1000 zeros).

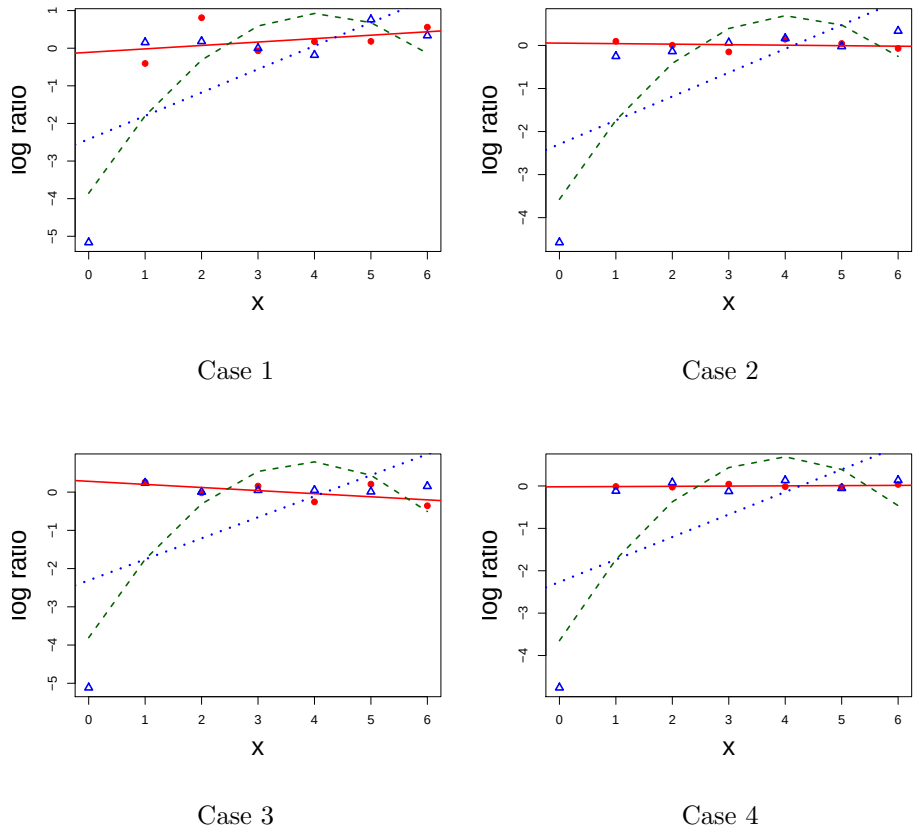

FIGURE 6.16: Ratio plot of the averaged frequencies for the positive samples (solid points) and for the validation samples (empty triangles) with corresponding regression lines.

The estimate for f_0 was achieved using the positive sample (Positive), a single line model (SLM) $log(r_x) = \alpha + \beta x$, a single quadratic model (SQM) $log(r_x) = \alpha + \beta x + \lambda x^2$, and a zero-inflation model (Zero-inflation model) $log(r_x) = \alpha + \beta x + \rho x$ where ρ is a dummy variable for the zero values. The true value for f_0 is between brackets in Table 6.10.

As we observe from Table 6.10, the zero-inflated model is always much closer to the true value in all the analysed situations. The results using just the positive sample are too low to be considered useful. Despite it does not appear in the table, the Horvitz-Thompson estimate was also calculated and the same values for f_0 were achieved. This can be expected, since we are working with just the positive sample. The single line model and the single quadratic model are not promising in situations where we have this type of data.

TABLE 6.10: Simulation study: estimates of f_0 from the simulation study of a zero-inflated data from a Binomial distribution.

	Positive	SLM	SQM	Zero-inflation model
Case 1 (50)	0	1	6	50
Case 2 (250)	2	11	50	181
Case 3 (500)	3	13	80	500
Case 4 (1000)	8	40	205	933

6.7 Discussion and conclusions

The ratio regression approach was discussed, and the ratio regression approach for the positive sample has been extended to include information from the validation sample, defined as a further untruncated sample including also zero counts which are not observed in conventional capture-recapture settings. Including a validation sample helps reduce bias and increase efficiency. Simulation studies corroborated the role of the validation sample in the estimation process showing that we can rely on the estimate for the population size with more confidence. The identical model might be used for positive and validation sample, or a partly congruent model such as the parallel lines model, or two separate models such as the separate lines model. In the latter case, there is no gain in bias, but only in efficiency. Modelling was also considered to account for a first ratio that is maybe lower than the others.

The Salmonella data used to illustrate the theory was provided by the Animal and Plant Health Agency and it is related with an important public health concern: Salmonella infection in poultry. The objective was to adjust the undercount of disease occurrence in UK farms during the period of the EU baseline survey which took place between October 2004 and September 2005. The work focuses essentially in the development of methodology to include validation information into the capture-recapture modelling to increase the accuracy and efficiency of the final estimate for the number of unrecorded cases.

Other data that was used in this study: the Bowel Cancer data, the Brucellosis and Syphilis data and lastly, the Heroin drug users data.

Using the ratio regression approach there are numerous ways to select an appropriate model. We have focused here on the Wald-statistic selecting significant coefficients and

model selection criteria were also used, such as AIC and BIC. Another way would be the likelihood ratio statistic.

In the case of the Salmonella data, on the basis of these criteria, the single line model considering only the positive counts variable seem to be the most appropriate to explore. This is also the case for the Bowel Cancer data, the Brucellosis data and the Syphilis data. However, in the case we consider the Heroin users data set, the most beneficial model to use is the separate lines model which is the same as using the positive sample only.

The EU survey reported a prevalence of Salmonella of 11.7% (53 infected farms out of 454 holdings), however, Arnold *et al.* [10] indicated a prevalence of 18% after analysing the positive data using Bayesian methods thus giving a much more alarming result. The results of this work help to confirm that the prevalence was in fact higher than 11.7%. According to the results of the most significant model (single line model), obtained by a ratio regression approach incorporating the validation sample, we report a prevalence of 17%, with a 95% confidence interval (9.83% - 27.89%).

Many studies using the Bowel Cancer data can be found in the literature. See, for example, Alfö *et al.* [2]. Another important study was performed by Lloyd and Frommer [61]. They estimated the false negative fraction of the population for the Bowel Cancer screening test by proposing a beta-binomial model to describe individual testing. They found an estimate $\hat{f}_0 = 61$ for the number of individuals who were disease positive but were not identified by the screening test, that is, 24% of the total of the individuals who were diagnosed, while when considering the validation sample, this proportion decreases to 22%. Using the ratio regression approach, it was estimated a value of 24 false negative individuals using the best model (single line model) following the mentioned selection criteria. These two results differ substancially. Other approaches can be studied as well; for instance, Böhning *et al.* [21] estimated $\hat{f}_0 = 21$ using a first-order fractional polynomial ratio regression approach with power $p = 0$ which is a very close result to the one we achieved in this case.

Köse *et al.* [55] used an extension of the Lincoln-Petersen estimator to estimate the completeness of the surveillance system for Syphilis and other transmittable diseases. Truncating multiple identifications larger than two and using a truncated Poisson model, they estimated a total population size of 282 individuals with Syphilis disease, with a 95%

confidence interval (265 - 300). Note that the population size for this study also includes the sum of the observed individuals for the validation sample. Therefore, the estimate $\hat{f}_0 = 47$ gives the number of individuals who were not identified by any laboratory. This is a close estimate to the one obtained by the ratio regression approach using the Single Line Model where $\hat{f}_0 = 56$.

The Brucellosis data was also analysed by Köse *et al.* [55]. Following the same approach used for the Syphilis data, the estimate was $\hat{f}_0 = 100$ individuals who were not diagnosed with the disease by any of the laboratories. This result, however, turned out to be very different from the result we reached using the ratio regression with the Single Line regression $\hat{f}_0 = 153$ which is, under perspective, the best model for this case study.

Lerdsuwansri [57] considered several estimators on the Heroin users data to find the total number of Heroin users in Bangkok, such as the Zelterman estimator and a Non-Parametric Maximum Likelihood estimator (NPMLe). For more details, please see [57] and Lerdsuwansri *et al.* [58]. Using a ratio regression approach, the separate lines regression model revealed to be the best choice for this data set, with $\hat{f}_0 = 3919$ and $\hat{N} = 5812$.

We see the most important aspect of the use of validation information lie in the fact that more trust can be developed in the model to estimate the unobserved part of the population. The ratio regression approach allows us to fit a flexible model to the data without losing identifiability.

The main findings and conclusions follow for each dataset used to illustrate the theory of the thesis follow. The two first rows show the results under homogeneity and the third and forth rows show the results for the finite mixture model and the ratio regression model, respectively.

Using the estimate for N and the estimated parameters obtained by each mixture model, confidence intervals were determined by simulating 1000 replications of 1000 estimates of N and the 25^{th} and the 975^{th} order statistics were taken to build the confidence intervals shown in tables below between brackets.

- Salmonella data.

TABLE 6.11: Salmonella data: estimates of N using only the positive sample (second column) and both samples (third column) with respective confidence intervals.

Model	Positive	Positive and Validation
Binomial distribution	55 (23.60,86.81)	54 (30.11-80.40)
Poisson distribution	56 (23.42,87.23)	54 (28.82,81.10)
Binomial finite mixture model (2 components)	62 (36.80,88.22)	60 (39.51,81.29)
Ratio regression - Single Line Model	82 (54.02,109.64)	77 (56.65,97.90)

Under homogeneity, there is basically any difference neither between using validation or the positive sample only nor between using the Binomial or the Poisson distribution. However, when heterogeneity is included in the modelling, there is a considerable difference between modelling data using a Binomial finite mixture model with 2 components or the ratio regression model using a single line model, and the fact we consider the validation sample since the confidence interval is narrower.

- Bowel Cancer data.

TABLE 6.12: Bowel Cancer data: estimates of N using only the positive sample (second column) and both samples (third column) with respective confidence intervals.

Model	Positive	Positive and Validation
Binomial distribution	193 (150.71,235.9)	193 (151.20,236.71)
Poisson distribution	197 (154.21,241.37)	198 (156.32,239.31)
Binomial finite mixture model (3 components)	226 (211.41,242.03)	234 (209.01,261.05)
Ratio regression - Single Line Model	243 (226.90,258.02)	245 (218.56,271.35)

In the case of the Bowel Cancer data, under homogeneity, we obtain the same estimates using both samples performing the EM algorithm with the Binomial distribution and the Poisson distribution. This finding does not occur with the results obtained by the finite mixtures and the ratio regression approach for for both samples.

- Brucellosis data.

TABLE 6.13: Brucellosis data: estimates of N using only the positive sample (second column) and using the both samples (third column) with respective confidence intervals.

Model	Positive	Positive and Validation
Binomial distribution	114 (49.08,177.35)	128 (76.40,179.33)
Poisson distribution	121 (61.02,179.80)	136 (82.81,188.74)
Binomial finite mixture model (3 components)	314 (178.29,449.30)	234 (192.81,275.63)
Ratio regression - Single Line Model	300 (154.90,444.63)	260 (196.08,324.55)

In the Brucellosis case, we observe clear differences between estimates using the positive sample only or the positive and the validation sample, even in the case of homogeneous models. The same can be seen for finite mixtures and ratio regression models. Notice that confidence intervals for the cases we use both samples are substantially narrower than when we use the positive sample only.

- Heroin users data.

TABLE 6.14: Heroin users data: estimates of N using only the positive sample (second column) and using both samples (third column) with respective confidence intervals.

Model	Positive	Positive and Validation
Binomial distribution	2426 (1338.75,3515.47)	3105 (2214.01,3996.92)
Poisson distribution	2565 (1506.37,3622.16)	3257 (2401.02,4114.40)
Binomial finite mixture model (2 components)	5493 (4508.22,6479.14)	5648 (5691.11,5934.01)
Ratio regression - Separate Lines Model	5812 (4838.26,6784.89)	5812 (5790.28,5832.93)

The ratio regression approach determined the best model for this case study is that based on using the positive sample only, see section 6.3.4. Using a finite mixture Binomial model with 2 components with and without the validation sample also produce very close estimates, however using validation information, we obtain a shorter confidence interval.

- Syphilis data.

TABLE 6.15: Syphilis data: estimates of N using only the positive sample (second column) and using both samples (third column) with respective confidence intervals.

Model	Positive	Positive and Validation
Binomial distribution	171 (130.31,212.91)	172 (133.22,210.39)
Poisson distribution	181 (132.01,227.84)	182 (132.90,230.11)
Binomial finite mixture model (2 components)	198 (167.20,230.34)	192 (174.79,210.19)
Ratio regression - Single Line Model	203 (167.09,239.32)	205 (181.14,229.23)

For the Syphilis data, using an homogeneous approach, there are no difference in the estimates using the positive sample or both samples which are basically the same. Under heteroheneity modelling, the Binomial mixture model and the ratio regression model also produce a close estimate for the total population size.

In fact, as it was aforementioned, it was shown that more trust is developed for estimates of the number of missing observations using models with validation information

incorporated. For these models, the bias is smaller and the precision of the model increases. However, it cannot be stated that the mixture models approach is a better or worse choice to estimate the total size of a target population. The number of missing observations can in fact be greater than the ones obtained by the discussed methods, however, lower bounds for the estimates of the number of unreported units can be found in this work.

Chapter 7

Concluding remarks

7.1 Conclusions and Future Work

Capture-recapture methods attempt to estimate the total size of a population of interest which is only partially observed. The typical capture-recapture structure involves a positive data sample where the frequency of the individuals of the population are counted for each occasion they have been identified during the period of study. However, due to errors in registration/identification processes, a part of the population is not observed. We focus on the estimation of such hidden part of the population.

Sometimes, tests, studies or procedures are repeated and we have complete access to another sample, which is called validation sample. Notice that there is no hidden information in this secondary sample.

When using only a positive sample, it is impossible to evaluate if our model is performing well to estimate the number of missing individuals, and therefore, the total size of the population. However, if we take advantage of the validation sample during the estimation process, we might be able to have more confidence in our final estimate.

This work focusses on developing methodology to incorporate validation information in standard capture-recapture methods.

The EM algorithm, a popular iterative algorithm for Maximum Likelihood estimation of parameters of interest was initially discussed using simple homogeneous Binomial and Poisson distributions. For an attempt to explore the role of the validation sample

into the modelling, there were considered models using the positive sample and models considering both samples (the positive and validation samples). It was shown that, even when the final estimate using validation does not seem to benefit from having this other sample, the estimate is significantly better in terms of decreasing the variance and consequently, the confidence in the obtained result is boosted.

An extension of the Good-Turing estimator using validation information was also introduced and, also in this case, simulation results show a strong difference between using or not using validation information. In fact, it is clear that even under homogeneity (which is the case), there is a clear benefit for the estimate accuracy in considering the validation sample.

Notice that this approach requires basic assumptions to be valid: there is individual homogeneity and independence among all the members of the population and between study period occasions for each member. However, these assumptions are unlikely to be met. Since each individual is unique and there are several factors that imply variability in the data, individual heterogeneity may play the most important role in the estimation process and cannot be ignored. Two paths were followed in this work to model heterogeneity while still incorporating the validation sample: finite mixture models and ratio regression.

These two approaches allow a flexible modelling of data considering heterogeneity; however, they cannot be compared in theoretical terms starting by the fact that mixture models were used with the classical discrete Binomial distribution whilst ratio regression is a continuous approach derived by ratios of frequencies of neighbour counts.

We discussed mixtures of Binomials for the positive and the positive-validation samples. The maximum number of components was defined taking into account the identifiability of the model which is closely linked to the number of sampling occasions. The number of components was chosen by model selection criteria, such as the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC). The EM algorithm tackled the conditional likelihood which contains the most part of the information, but using a profile mixture likelihood is also possible. It could happen that the data suffers from zero-inflation: a number of non-observed cases much higher than expected. Here, it is important to highlight that validation is vital in situations of this type. Without

using validation, as in traditional capture-recapture problems, there is no opportunity to develop a model which performs well for the hidden part.

The ratio plot works as a diagnostic device for the Binomial distribution. Using the ratio plot one can infer about the distribution of the positive and the validation samples. It is also a tool that allows us to identify individual heterogeneity in the data. To model heterogeneity, a ratio regression approach was investigated through three different models which incorporate validation information: the single line regression model, the parallel lines regression model and the separate lines regression model. The last performs exactly the same as considering the positive sample only. The best model was selected, once again, using the BIC and the AIC criteria. Zero-inflated models were also discussed in case the frequency of missing individuals causes a ratio which is lower than expected by looking at the others obtained by the positive sample. Simulation studies suggest that the use of the validation sample in this approach brings many advantages for the modelling. The final estimate is more precise and the bias is meaningly smaller.

Future work includes using the profile mixture likelihood with validation information and compare the results obtained using the conditional likelihood as well as trying different distributions on the specific datasets. It would be very interesting to develop other ratio regression models to include validation.

Appendix A

Supplementary material to the simulation study in section 4.1.2

This appendix contains the results for $\theta = 0.20, 0.25$ for the simulation study that was performed to investigate the effect of having a validation sample in capture-recapture studies under homogeneity in section 4.1.2 of chapter 4. Comments on the procedure and the results can be found in that section.

- **Results for** $\theta = 0.20$:

TABLE A.1: Simulation study: $\theta = 0.20$ results, $M = 1000 (M = 5000)$ samples.

N	Mean with validation	Mean without validation
25	0.1983 (0.1995)	0.1978 (0.1990)
50	0.2000 (0.1993)	0.1994 (0.1998)
100	0.2002 (0.1997)	0.2007 (0.2000)
500	0.2001 (0.1998)	0.2003 (0.1998)
1000	0.19981 (0.2000)	0.1999 (0.2000)

TABLE A.2: Simulation study: $\theta = 0.20$ estimated variance, $M = 1000(M = 5000)$ samples.

N	Variance with validation	Variance without validation
25	0.0006 (0.0006)	0.0017 (0.0016)
50	0.0005 (0.0004)	0.0009 (0.0009)
100	0.0003 (0.0003)	0.0005 (0.0004)
500	7.5754×10^{-05} (8.0636×10^{-05})	8.6187×10^{-05} (8.7849×10^{-05})
1000	4.3079×10^{-05} (4.1003×10^{-05})	4.6145×10^{-05} (4.2691×10^{-05})

TABLE A.3: Ratio (Variance with validation / Variance without validation) for $\theta = 0.20$, $M = 1000(M = 5000)$ samples.

N	Ratio of variances
25	0.3664 (0.3726)
50	0.5315 (0.5094)
100	0.6975 (0.6717)
500	0.8789 (0.9179)
1000	0.9336 (0.9604)

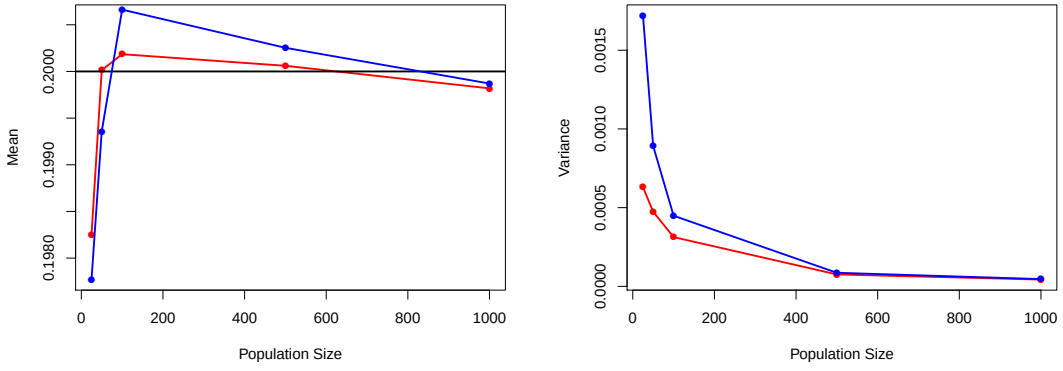

FIGURE A.1: Simulation study: $\theta = 0.20$. Left panel: mean estimates for θ with varying N, using validation information (red) or not (blue). The true value is the black solid line. Right panel: corresponding variance values. $M = 1000$ samples.

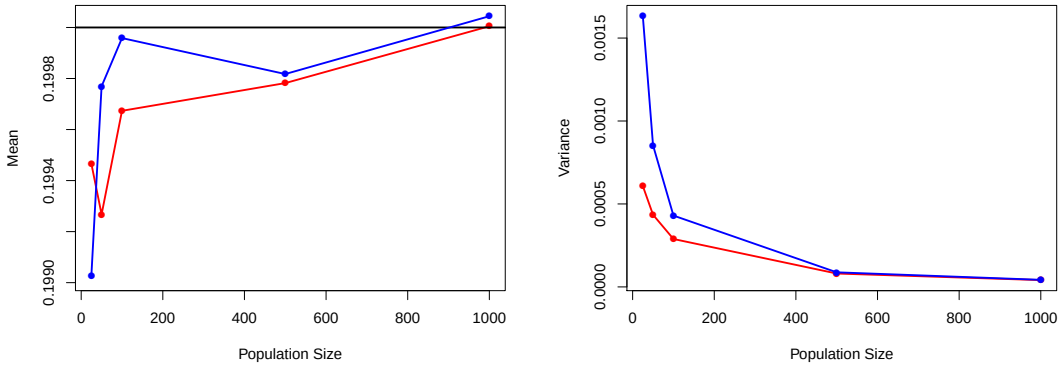

FIGURE A.2: Simulation study: $\theta = 0.20$. Left panel: mean estimates for θ with varying N, using validation information (red) or not (blue). The true value is the black solid line. Right panel: corresponding variance values. $M = 5000$ samples.

TABLE A.4: Simulation study: $\theta = 0.20$ mean estimates for N with (right column) and without validation information (left column). $M = 1000 (M = 5000)$ samples.

N	$\hat{N}$ without using Validation	$\hat{N}$ using Validation
25	25.90 (25.67)	25.41 (25.2449)
50	50.70 (50.66)	50.29 (50.40)
100	100.53 (100.76)	100.45 (100.62)
500	500.07 (500.82)	500.23 (500.81)
1000	999.60 (1000.40)	999.67 (1001.47)

- **Results for** $\theta = 0.25$:

TABLE A.5: Simulation study: $\theta = 0.25$ results, $M = 1000 (M = 5000)$ samples.

N	Mean with validation	Mean without validation
25	0.2466 (0.2443)	0.2497 (0.2467)
50	0.2481 (0.2474)	0.2397 (0.2491)
100	0.2487 (0.2486)	0.2499 (0.2496)
500	0.2499 (0.2494)	0.2504 (0.2497)
1000	0.2494 (0.2497)	0.2496 (0.2499)

TABLE A.6: Simulation study: $\theta = 0.25$ estimated variance, $M = 1000(M = 5000)$ samples.

N	Variance with validation	Variance without validation
25	0.0008 (0.0009)	0.0017 (0.0017)
50	0.0005 (0.0005)	0.0009 (0.0008)
100	0.0003 (0.0003)	0.0004 (0.0004)
500	8.1895×10^{-05} (7.9020×10^{-05})	8.7128×10^{-05} (8.3200×10^{-05})
1000	3.7650×10^{-05} (3.9175×10^{-05})	3.9145×10^{-05} (4.0016×10^{-05})

TABLE A.7: Ratio (Variance with validation / Variance without validation) for $\theta = 0.25$, $M = 1000(M = 5000)$ samples.

N	Ratio of variances
25	0.4344 (0.5082)
50	0.5710 (0.6256)
100	0.7655 (0.7924)
500	0.9400 (0.9498)
1000	0.9618 (0.9790)

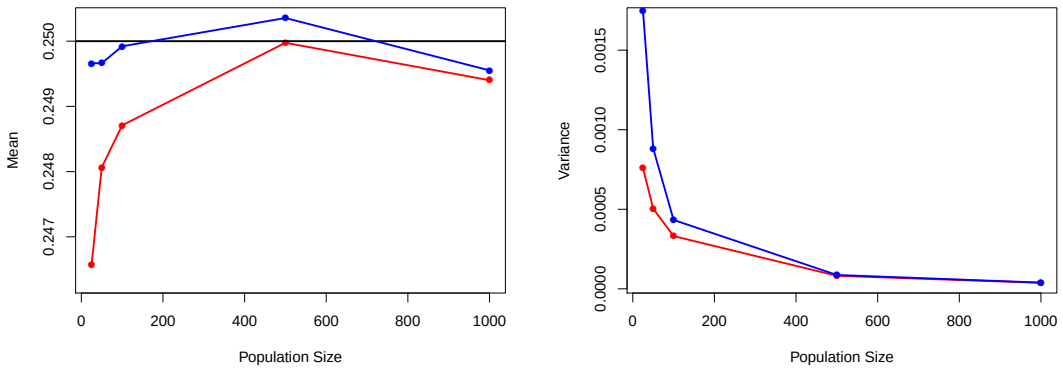

FIGURE A.3: Simulation study: $\theta = 0.25$. Left panel: mean estimates for θ with varying N, using validation information (red) or not (blue). The true value is the black solid line. Right panel: corresponding variance values. $M = 1000$ samples.

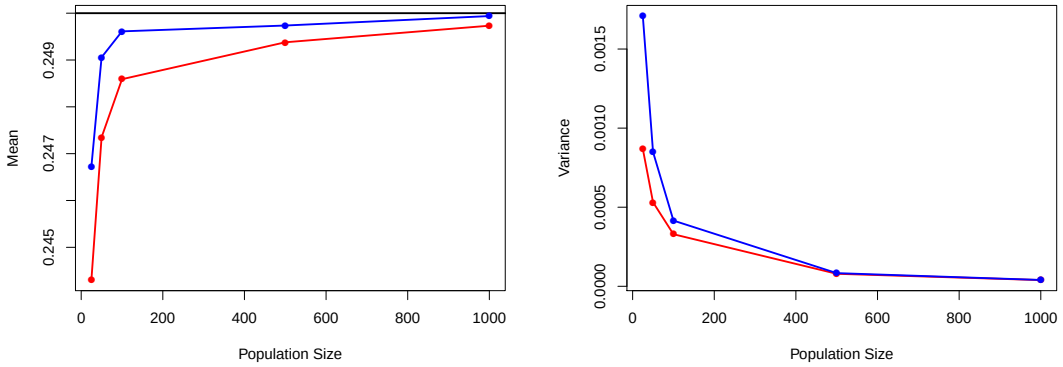

FIGURE A.4: Simulation study: $\theta = 0.25$. Left panel: mean estimates for θ with varying N, using validation information (red) or not (blue). The true value is the black solid line. Right panel: corresponding variance values. $M = 5000$ samples.

TABLE A.8: Simulation study: $\theta = 0.25$ mean estimates for N with (right column) and without validation information (left column). $M = 1000 (M = 5000)$ samples.

N	$\hat{N}$ without using Validation	$\hat{N}$ using Validation
25	25.29 (25.41)	25.27 (25.41)
50	50.44 (50.41)	50.40 (50.43)
100	100.23 (100.40)	100.34 (100.49)
500	499.84 (500.59)	500.09 (500.84)
1000	1000.82 (1000.25)	1001.02 (1000.53)

Appendix B

R code for finite mixture models

The code developed for finite mixture models with Binomial kernel was a joint work with Dr. Antonello Maruotti.

The code to perform a Binomial mixture model with the positive sample as discussed in chapter 5 follows:

```
cr.binom <- function(sample,K=2,m) # define the number of components K
{
  if(missing(m))
    stop("Number of maximum recaptures is missing")
#set.seed(st)
#
# Initialization
#
t.sample <- table(sample)[-1]

weights = matrix(runif((m+1)*K),ncol=K)
weights = weights/apply(weights,1,sum)
prior <- NULL
lambda <- lambda.sample <- NULL
f.sample.num <- f.sample <- matrix(NA,nrow = m+1,ncol=K)
f <- matrix(NA,nrow = m+1,ncol=K)
for (k in 1:K)
{
prior[k] <- (sum(t.sample*weights[-1,k]))/(sum(t.sample))
lambda[k] <- sum(t.sample*(1:m)*weights[-1,k])/sum(t.sample*weights[-1,k])
f.sample.num[,k] <- dbinom(0:m,m,lambda[k]/m)*prior[k]
}
loglike.old <- -Inf
f0 <- 0
```

```
loglike <- 0
dif <- Inf
iter <- 0
#
# E-step
#
while (dif > 10^-4)
{
iter <- iter+1
print(iter)
N <- sum(t.sample)/(1-sum(f.sample.num[1,]))
f0 <- round(N-sum(t.sample))

#
# M-step
#
for (k in 1:K)
{
lambda[k] <- sum(c(f0,t.sample)*(0:m)*weights[,k])/sum(c(f0,t.sample)*weights[,k])
f.sample.num[,k] <- dbinom(0:m,m,lambda[k]/m)*prior[k]
}
weights <- diag(1/apply(f.sample.num,1,sum))%*%f.sample.num

prior <- apply(matrix(rep(c(f0,t.sample),K),ncol=K)*weights,2,sum)/sum(c(f0,t.sample))
loglike <- sum(log(apply(f.sample.num[-1,]/(1-sum(f.sample.num[1,])),1,sum))*t.sample)
dif <- (loglike-loglike.old)
print(dif)
print(N)
loglike.old <- loglike
}
return(list(N=N,f0=f0,lambda=lambda,prior=prior,loglike=loglike))
}
```

The code to perform a Binomial mixture model incorporating the validation sample follows:

```
cond.like.binom <- function(sample,validation,K=2,m)
{
  if(missing(m))
    stop("Number of maximum recaptures is missing")
#set.seed(st)
#
# Initialization
#
t.sample <- table(sample)[-1]
t.validation <- table(validation)
```

```
weights = matrix(runif((m+1)*K),ncol=K)  #posterior weights
weights = weights/apply(weights,1,sum)
prior <- NULL #qs
lambda <- lambda.sample <- lambda.validation <- NULL
f.sample.num <- f.sample <- matrix(NA,nrow = m+1,ncol=K)
f <- f.validation <- matrix(NA,nrow = m+1,ncol=K)
for (k in 1:K)
{
prior[k] <- (sum(t.sample*weights[-1,k])+sum(t.validation*weights[,k]))
/(sum(t.sample)+sum(t.validation))
lambda.sample[k] <- sum(t.sample*(1:m)*weights[-1,k])
/sum(t.sample*weights[-1,k])
lambda.validation[k] <- sum(t.validation*(0:m)*weights[,k])
/sum(t.validation*weights[,k])
lambda[k] <- (lambda.sample[k]*sum(t.sample)+lambda.validation[k]*sum(t.validation))
/(sum(t.sample)+sum(t.validation))
f.sample.num[,k] <- dbinom(0:m,m,lambda[k]/m)*prior[k]
}
loglike.old <- -Inf
f0 <- 0
loglike <- 0
dif <- Inf
iter <- 0
#
# E-step
#
while (dif > 10^-4)
{
iter <- iter+1
print(iter)
N <- sum(t.sample)/(1-sum(f.sample.num[1,]))
f0 <- round(N-sum(t.sample))

#
# M-step
#
for (k in 1:K)
{
lambda.sample[k] <- sum(c(f0,t.sample)*(0:m)*weights[,k])
/sum(c(f0,t.sample)*weights[,k])
lambda.validation[k] <- sum(t.validation*(0:m)*weights[,k])
/sum(t.validation*weights[,k])
lambda[k] <- (lambda.sample[k]*sum(c(f0,t.sample))+
lambda.validation[k]*sum(t.validation))/(sum(c(f0,t.sample))+sum(t.validation))

#f[,k] <- dbinom(0:m,m,lambda[k]/m)
f.sample.num[,k] <- dbinom(0:m,m,lambda[k]/m)*prior[k]
```

```
#f.sample.den[,k] <- 1-dbinom(0,m,lambda[k]/m)*prior[k]
#f.sample[,k] <- f.sample.num[,k]/f.sample.den[,k]
f.validation[,k] <- dbinom(0:m,m,lambda[k]/m)*prior[k]
}
weights <- diag(1/apply(f.sample.num,1,sum))%*%f.sample.num

prior <- apply(matrix(rep(c(f0,t.sample)+t.validation,K),ncol=K)*weights,2,sum)
/sum(c(f0,t.sample)+t.validation)
loglike <- sum(log(apply(f.sample.num[-1,]/
(1-sum(f.sample.num[1,])),1,sum))*t.sample)+sum(log(apply(f.validation,1,sum))*t.validation)
dif <- (loglike-loglike.old)
print(dif)
print(N)
loglike.old <- loglike
}
return(list(N=N,f0=f0,lambda=lambda,prior=prior,loglike=loglike))
}
```

The results are achieved as in the next example for the Bowel Cancer data:

```
#Results Bowel Cancer:

positives=c(0,rep(1,37),rep(2,22),rep(3,25), rep(4,29),rep(5,34), rep(6,45))
validation=c(rep(0,22),rep(1,8),rep(2,12),rep(3,16),rep(4,21),rep(5,12),rep(6,31))

#With Validation:
mod1=cond.like.binom(positives,validation,m=6)

#Without Validation:
mod2=cr.binom(positives,m=6)
```

The results are shown in detail in section 5.3.

Bibliography

[1] E. Alacs and A. Georges. Wildlife across our borders: a review of the illegal trade in Australia. *Australian Journal of Forensic Sciences*, 40(2):147–160, 2008.

[2] M. Alfö, D. Böhning, and I. Rocchetti. Ratio regression and capture-recapture. In D. Böhning, P. G. van der Heijden, and J. Bunge (editors), editors, *Capture-Recapture Methods for the Social and Medical Sciences*, chapter 2, pages 19–37. CRC Press, 2018.

[3] D. Alunni Fegatelli and L. Tardella. Improved inference on capture-recapture models with behavioural effects. *Statistical Methods and Applications*, 22(1):45–66, 2013.

[4] S. C. Amstrup, T. L. McDonald, and B. F. Manly. *Handbook of capture-recapture analysis*. Princeton University Press, 2010.

[5] O. Anan, D. Böhning, and A. Maruotti. Population size estimation and heterogeneity in capture–recapture data: a linear regression estimator based on the Conway–Maxwell–Poisson distribution. *Statistical Methods & Applications*, 26(1):49–79, 2017.

[6] O. Anan, D. Böhning, and A. Maruotti. Uncertainty estimation in heterogeneous capture–recapture count data. *Journal of Statistical Computation and Simulation*, 87(10):2094–2114, 2017.

[7] M. Arnold, J. Carrique-Mas, and R. Davies. Sensitivity of environmental sampling methods for detecting Salmonella Enteritidis in commercial laying flocks relative to the within-flock prevalence. *Epidemiology and Infection*, 138(03):330–339, 2010.

[8] M. Arnold, F. Martelli, I. McLaren, and R. Davies. Estimation of the Rate of Egg Contamination from Salmonella-Infected Chickens. *Zoonoses and Public Health*, 61(1):18–27, 2014.

[9] M. Arnold, F. Martelli, I. McLaren, and R. Davies. Estimation of the sensitivity of environmental sampling for detection of Salmonella in commercial layer flocks post-introduction of national control programmes. *Epidemiology and Infection*, 142(05):1061–1069, 2014.

[10] M. Arnold, C. Papadopoulou, R. Davies, J. Carrique-Mas, S. Evans, and L. Hoinville. Estimation of Salmonella prevalence in UK egg-laying holdings. *Preventive Veterinary Medicine*, 94(3):306–309, 2010.

[11] C. Azevedo, D. Böhning, and M. Arnold. A ratio regression approach to estimate the size of the salmonella infected flock population using validation information. In D. Böhning, P. G. van der Heijden, and J. Bunge (editors), editors, *Capture-Recapture Methods for the Social and Medical Sciences*, chapter 5, pages 59–77. CRC Press, 2018.

[12] M. Begon. *Investigating animal abundance: capture-recapture for biologists.* Edward Arnold (Publishers) Ltd., 1979.

[13] Y. M. Bishop, S. E. Fienberg, P. W. Holland, R. J. Light, and F. Mosteller. Book review: Discrete multivariate analysis: Theory and practice. *Applied Psychological Measurement*, 1(2):297–306, 1977.

[14] D. Böhning. The EM algorithm with gradient function update for discrete mixtures with known (fixed) number of components. *Statistics and Computing*, 13(3):257–265, 2003.

[15] D. Böhning. Power series mixtures and the ratio plot with applications to zero-truncated count distribution modelling. *Metron*, 73(2):201–216, 2015.

[16] D. Böhning. Ratio plot and ratio regression with applications to social and medical sciences. *Statistical Science*, 31(2):205–218, 2016.

[17] D. Böhning, M. F. Baksh, R. Lerdsuwansri, and J. Gallagher. Use of the ratio plot in capture–recapture estimation. *Journal of Computational and Graphical Statistics*, 22(1):135–155, 2013.

[18] D. Böhning, E. Dietz, R. Kuhnert, and D. Schön. Mixture models for capture-recapture count data. *Statistical Methods and Applications*, 14(1):29–43, 2005.

[19] D. Böhning, E. Dietz, P. Schlattmann, L. Mendonca, and U. Kirchner. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(2):195–209, 1999.

[20] D. Böhning and R. Kuhnert. Equivalence of truncated count mixture distributions and mixtures of truncated count distributions. *Biometrics*, 62(4):1207–1215, 2006.

[21] D. Böhning, I. Rocchetti, M. Alfó, and H. Holling. A flexible ratio regression approach for zero-truncated capture–recapture counts. *Biometrics*, 72(3):697–706, 2016.

[22] D. Böhning and D. Schön. Nonparametric maximum likelihood estimation of population size based on the counting distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(4):721–737, 2005.

[23] D. Böhning, P. G. van der Heijden, and J. Bunge (editors). *Capture-Recapture Methods for the Social and Medical Sciences*. CRC Press, 2018.

[24] D. Böhning and V. J. D. R. Vilas. Estimating the hidden number of scrapie affected holdings in Great Britain using a simple, truncated count model allowing for heterogeneity. *Journal of Agricultural, Biological, and Environmental Statistics*, 13(1):1, 2008.

[25] S. Brittain and D. Böhning. Estimators in capture-recapture studies with two sources. *AStA Advances in Statistical Analysis*, 93(1):23–47, 2009.

[26] J. Bunge and S. Sernaker. Estimating the population size via the empirical probability generating function. In D. Böhning, P. G. van der Heijden, and J. Bunge (editors), editors, *Capture-Recapture Methods for the Social and Medical Sciences*, chapter 9, pages 107–118. CRC Press, 2018.

[27] K. P. Burnham and D. R. Anderson. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304, 2004.

[28] A. C. Cameron and P. K. Trivedi. *Microeconometrics: methods and applications*. Cambridge University Press, 2005.

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

[30] J. J. Deeks. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ*, 323(7305):157–162, 2001.

[31] F. Dellaert. The expectation maximization algorithm. Technical report, Georgia Institute of Technology, 2002.

[32] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, (Series B (Methodological)):1–38, 1977.

[33] E. Dietz and D. Böhning. On estimation of the poisson parameter in zero-modified poisson models. *Computational Statistics & Data Analysis*, 34(4):441–459, 2000.

[34] C. B. Do and S. Batzoglou. What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897, 2008.

[35] R. M. Dorazio and J. Andrew Royle. Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, 59(2):351–364, 2003.

[36] S. D. Dubey. Compound Pascal distributions. *Annals of the Institute of Statistical Mathematics*, 18(1):357–365, 1966.

[37] R. Durusoy and A. O. Karababa. Completeness of hepatitis, brucellosis, syphilis, measles and HIV/AIDS surveillance in Izmir, Turkey. *BMC Public Health*, 10(1):71, 2010.

[38] B. Efron and R. Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[39] M. Ellsberg, L. Heise, R. Pena, S. Agurto, and A. Winkvist. Researching domestic violence against women: methodological and ethical considerations. *Studies in Family Planning*, 32(1):1–16, 2001.

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

[41] É. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non-parametric hidden Markov models and applications. *Statistics and Computing*, 26(1-2):61–71, 2016.

[42] D. Geoffrey McLachlan. Finite mixture models. Hoboken, 2000.

[43] E. I. George. Capture-recapture estimation via Gibbs sampling. *Biometrika*, 79(4):677–683, 1992.

[44] I. Gillespie, S. O'brien, G. Adak, L. Ward, and H. Smith. Foodborne general outbreaks of Salmonella Enteritidis phage type 4 infection, England and Wales, 1992–2002: where are the risks? *Epidemiology and Infection*, 133(05):795–801, 2005.

[45] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[46] I. R. Harris. The estimated frequency of zero for a mixed Poisson distribution. *Statistics & probability letters*, 12(5):371–372, 1991.

[47] G. Hay and F. Smit. Estimating the number of drug injectors from needle exchange data. *Addiction Research & Theory*, 11(4):235–243, 2003.

[48] D. C. Hoaglin. A poissonness plot. *The American Statistician*, 34(3):146–149, 1980.

[49] H. Holzmann, A. Munk, and W. Zucchini. On identifiability in capture–recapture models. *Biometrics*, 62(3):934–936, 2006.

[50] E. B. Hook and R. R. Regal. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews*, 17(2):243–264, 1995.

[51] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[52] Z. John Lu. The elements of statistical learning: data mining, inference, and prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(3):693–694, 2010.

[53] K. U. Karanth. Estimating tiger Panthera tigris populations from camera-trap data using capture-recapture models. *Biological Conservation*, 71(3):333–338, 1995.

[54] C. Keribin. Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 49–66, 2000.

[55] T. Köse, M. Orman, F. Ikiz, M. Baksh, J. Gallagher, and D. Böhning. Extending the Lincoln-Petersen estimator for multiple identifications in one source. *Statistics in Medicine*, 33(24):4237–4249, 2014.

[56] R. Kuhnert, V. J. Del Rio Vilas, J. Gallagher, and D. Böhning. A Bagging-Based Correction for the Mixture Model Estimator of Population Size. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(6):993–1005, 2008.

[57] R. Lerdsuwansri. Generalisation of the Lincoln-Petersen approach to non-binary source variables. PhD Dissertation, Reading. January 2012.

[58] R. Lerdsuwansri and D. Böhning. Extending the lincoln-petersen estimator when both sources are counts. In D. Böhning, P. G. van der Heijden, and J. Bunge (editors), editors, *Capture-Recapture Methods for the Social and Medical Sciences*, chapter 23, pages 343–363. CRC Press, 2018.

[59] B. G. Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–163. JSTOR, 1995.

[60] W. A. Link. Nonidentifiability of population size from capture–recapture data with heterogeneous detection probabilities. *Biometrics*, 59(4):1123–1130, 2003.

[61] C. J. Lloyd and D. Frommer. Estimating the false negative fraction for a multiple screening test for bowel cancer when negatives are not verified. *Australian and New Zealand Journal of Statistics*, 46(4):531–542, 2004.

[62] C. J. Lloyd and D. J. Frommer. Regression-based estimation of the false negative fraction when multiple negatives are unverified. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(4):619–631, 2004.

[63] C. J. Lloyd and D. J. Frommer. An application of multinomial logistic regression to estimating performance of a multiple-screening test with incomplete verification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(1):89–102, 2008.

[64] A. Maruotti and O. Anan. The conway-maxwell-poisson distribution and capture-recapture count data. In D. Böhning, P. G. van der Heijden, and J. Bunge (editors), editors, *Capture-Recapture Methods for the Social and Medical Sciences*, chapter 3, pages 37–52. CRC Press, 2018.

[65] G. J. McLachlan, T. Krishnan, and S. K. Ng. The EM algorithm. Technical report, Papers/Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE), 2004.

[66] X. L. Meng. Missing data: dial M for??? *Journal of the American Statistical Association*, 95(452):1325–1330, 2000.

[67] M. G. Merli. Underreporting of births and infant deaths in rural China: Evidence from field research in one county of northern China. *The China Quarterly*, 155:637–655, 1998.

[68] G. Meurant. A review on the inverse of symmetric tridiagonal and block tridiagonal matrices. *SIAM Journal on Matrix Analysis and Applications*, 13(3):707–728, 1992.

[69] M. Miloslavsky and M. J. van der Laan. Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Computational Statistics & Data Analysis*, 41(3-4):413–428, 2003.

[70] Y. Min and A. Agresti. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5(1):1–19, 2005.

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

[72] J. Ord. Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society. Series A (General)*, pages 232–238, 1967.

[73] M. Oremus, M. M. Poole, D. Steel, and C. S. Baker. Isolation and interchange among insular spinner dolphin communities in the south pacific revealed by individual identification and genetic diversity. *Marine Ecology Progress Series*, 336:275–289, 2007.

[74] Y. Pawitan. *In all likelihood: statistical modelling and inference using likelihood.* Oxford University Press, 2001.

[75] S. Pledger. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56(2):434–442, 2000.

[76] J. Poorolajal, Y. Mohammadi, and F. Farzinara. Using the capture-recapture method to estimate the human immunodeficiency virus-positive population. *Epidemiology and Health*, 39, 2017.

[77] M. Ridout, C. G. Demétrio, and J. Hinde. Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference*, volume 19, pages 179–192, 1998.

[78] I. Rocchetti, J. Bunge, and D. Böhning. Population size estimation based upon ratios of recapture probabilities. *The Annals of Applied Statistics*, 5(2B):1512–1533, 2011.

[79] L. Sanathanan. Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, pages 142–152, 1972.

[80] L. Snow, R. Davies, K. Christiansen, J. Carrique-Mas, A. Wales, J. O'connor, A. Cook, and S. Evans. Survey of the prevalence of Salmonella species on commercial laying farms in the United Kingdom. *The Veterinary Record*, 161(14):471–476, 2007.

[81] M. Spevack. A complete and systematic concordance to the works of Shakespeare. 1-6, 1970.

[82] H. Teicher. Identifiability of finite mixtures. *The annals of Mathematical statistics*, pages 1265–1269, 1963.

[83] B. W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):290–295, 1976.

[84] P. G. van der Heijden, R. Bustami, M. J. Cruyff, G. Engbersen, and H. C. Van Houwelingen. Point and interval estimation of the population size using the truncated Poisson regression model. *Statistical Modelling*, 3(4):305–322, 2003.

[85] P. G. van der Heijden, M. Cruyff, and D. Böhning. Capture-recapture to estimate criminal populations. In *Encyclopedia of criminology and criminal justice*, pages 267–276. Springer, 2014.

[86] P. G. van der Heijden, M. Cruyff, and H. C. Van Houwelingen. Estimating the size of a criminal population from police records using the truncated Poisson regression model. *Statistica Neerlandica*, 57(3):289–304, 2003.

[87] P. G. Van der Heijden, J. Whittaker, M. Cruyff, B. Bakker, R. Van der Vliet, et al. People born in the Middle East but residing in the Netherlands: Invariant

population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics*, 6(3):831–852, 2012.

[88] P. C. Van Deusen. An EM algorithm for capture-recapture estimation. *Environmental and Ecological Statistics*, 9(2):151–165, 2002.

[89] E.-J. Wagenmakers and S. Farrell. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1):192–196, 2004.

[90] G. Ward, T. Hastie, S. Barry, J. Elith, and J. R. Leathwick. Presence-only data and the EM algorithm. *Biometrics*, 65(2):554–563, 2009.

[91] J. T. Wittes, T. Colton, and V. W. Sidel. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *Journal of Chronic Diseases*, 27(1):25–36, 1974.