Thesis

Erofili Grapsa

December, 2010

University of Southampton

Faculty of Social and Human Sciences

School of Mathematics

Bayesian analysis for categorical survey data

by

Erofili Grapsa

Thesis for the degree of Doctor of Philosophy

December, 2010

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

SCHOOL OF MATHEMATICS

Doctor of Philosophy

BAYESIAN ANALYSIS FOR CATEGORICAL SURVEY DATA

Erofili Grapsa

In this thesis, we develop Bayesian methodology for univariate and multivariate categorical survey data. The Multinomial model is used and the following problems are addressed. Limited information about the design variables leads us to model the unknown design variables taking into account the sampling scheme. Random effects are incorporated in the model to deal with the effect of sampling design, that produces the Multinomial GLMM and issues such as model comparison and model averaging are also discussed. The methodology is applied in a true dataset and estimates for population counts are obtained.

Contents

List of Figures

List of Tables

Author's Declaration

I, Erofili Grapsa, declare that the thesis entitled

Bayesian Analysis for Categorical Survey Data

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;

Acknowledgements

I would like to thank my supervisor Professor Jon Forster for his help with completing this thesis and my friends Carlo, Vasthi, Renato, Alinne, Vasiliki for their support during these three years of my PhD.

Chapter 1

Introduction

1.1 Motivation

Sample surveys are an essential tool for obtaining information on populations. Bayesian statistical methodology is increasingly being used to obtain inferences and associated measures of uncertainty in complex practical problems. However, for categorical survey data Bayesian analysis is unexplored and there is a need for a unified approach for univariate and multivariate cases. Our goal is to provide methodology to analyse finite population quantities coming from categorical responses under different survey designs. Therefore, we combine Bayesian and survey sampling theory to a unified theory that is developed for categorical data. Little (2004); Little and Zheng (2007) have discussed several Bayesian models for continuous variables, that we extend, modify and apply to categorical variables assuming limited information about the sampling scheme.

In this Chapter we introduce basic concepts of Bayesian and survey sampling theory and describe how they are combined to give Bayesian inference for surveys.

1.2 Bayesian Theory

1.2.1 Bayes' Theorem

Let $\boldsymbol{y} = (y_1, ..., y_n)$ be a vector of observations with *sampling density (likelihood)* given by $f(\boldsymbol{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of unknown *model parameters* (Forster and O'Hagan, 2004; Gelman et al., 2004). The set of possible values of $\boldsymbol{\theta}$ is the parameter space Θ. The likelihood together with the *prior distribution* $f(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ give the joint probability mass or density function for y and θ as

$$f(\boldsymbol{\theta}, \boldsymbol{y}) = f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})$$

Inference on the parameter is based on the *posterior density* which is computed via *Bayes' theorem*

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{\theta}, \boldsymbol{y})}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})}{f(\boldsymbol{y})}$$

where

$$f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$$

is called *marginal likelihood.* The factor $f(\boldsymbol{y})$ can then be omitted since it does not depend on $\boldsymbol{\theta}$ and can be considered as a constant. This yields

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})$$

1.2.2 Inference and Prediction

Forecasting or predicting a future value of an observation $\tilde{y}$ is based on the *predictive distribution* (Forster and O'Hagan, 2004; Gelman et al., 2004), whose density is given by

$$f(\tilde{y}|\boldsymbol{y}) = \int f(\tilde{y}|\boldsymbol{y}, \boldsymbol{\theta})f(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

Since, $\boldsymbol{y}$, $\tilde{y}$ are typically assumed conditionally independent given $\boldsymbol{\theta}$ we usually have $f(\tilde{y}|\boldsymbol{y},\boldsymbol{\theta}) = f(\tilde{y}|\boldsymbol{\theta})$ and the predictive density becomes

$$f(\tilde{y}|\boldsymbol{y}) = \int f(\tilde{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$$

Often, we are interested in some feature of the posterior distribution which takes the form of the posterior expectation of some function $g(\boldsymbol{\theta})$, such as

- $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$, the posterior expectation of $\boldsymbol{\theta}$.
- $g(\boldsymbol{\theta}) = \boldsymbol{\theta}^r$ for some $r > 1$, higher order posterior moments of $\boldsymbol{\theta}$.
- $g(\boldsymbol{\theta}) = I[\boldsymbol{\theta} \in A]$ (the indicator function of a set A), the posterior probability that $\boldsymbol{\theta}$ lies in the set A

or we might be interested in marginal distributions of the parameters,

$$f(\theta_1|\boldsymbol{y}) = \int f(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_p$$

Other features of interest may be quantiles and we can always plot the posterior density to get a more general idea about it. As it is often hard to solve the normalising integrals (the denominator in Bayes' theorem), simulation techniques are commonly used to draw samples from the posterior density. Once we obtain a sample $\boldsymbol{\theta}^1, \ldots., \boldsymbol{\theta}^T$ we can approximate any quantity of the posterior distribution using the simulated draws. For example, suppose θ is one-dimensional, we can compute

$$E(g(\theta)|\boldsymbol{y}) \approx \frac{1}{T}\sum_{j=1}^{T} g(\theta^j)$$

1.2.3 Hierarchical models

As mentioned previously, the joint probability distribution of the data $\boldsymbol{y}$ depends on the parameters $\boldsymbol{\theta}$. The distribution of $\boldsymbol{\theta}$ can also depend on other parameters, often called hyperparameters $\boldsymbol{\phi}$ as $p(\boldsymbol{\theta}|\boldsymbol{\phi})$. Then a prior distribution $p(\boldsymbol{\phi})$ is required for $\boldsymbol{\phi}$. This process could go further defining *hyper-hyperparameters* and these models are called *hierarchical models* (Forster and O'Hagan, 2004; Gelman et al., 2004). The joint distribution of $\boldsymbol{y}$, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can then be written as

$$f(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi}) = f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\boldsymbol{\phi})f(\boldsymbol{\phi})$$

The distribution of the parameters at any level depends on the parameters at the higher level and conditional on these parameters, is independent of parameters at levels above that. The joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is

$$f(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\boldsymbol{\phi})f(\boldsymbol{\phi})$$

Sometimes, we might be interested in the hyperparameters themselves. For example inference about $\boldsymbol{\phi}$ is made through its marginal posterior distribution

$$f(\boldsymbol{\phi}|\boldsymbol{y}) = \int f(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})d\boldsymbol{\theta} \propto f(\boldsymbol{\phi})f(\boldsymbol{y}|\boldsymbol{\phi})$$

1.2.4 Model comparison

Sometimes, we have several models which we wish to compare in order to choose one among them. This is a problem addressed in Chapter 5, where there are different potential models for the contingency table of interest. Thus, we wish to select one model and find a way to deal with the uncertainty about the choice between alternative models.

To deal with this uncertainty, we consider all potential models and assign a prior probability $f(m)$ to each one of them. Assume there are M alternative models and each model m consists of a likelihood $f(y|\boldsymbol{\theta}_m, m)$ and a prior distribution for $\boldsymbol{\theta}_m$. The joint distribution under model m is

$$f(\boldsymbol{y}, m, \boldsymbol{\theta}_m) = f(\boldsymbol{y}|\boldsymbol{\theta}_m, m)f(\boldsymbol{\theta}_m)f(m)$$

from which we can obtain the posterior distribution

$$\begin{aligned}
f(m, \boldsymbol{\theta}_m|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_m, m)f(\boldsymbol{\theta}_m|m)f(m)}{f(\boldsymbol{y})} \\
&= \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_m, m)f(\boldsymbol{\theta}_m, m)}{f(\boldsymbol{y}|m)} \times \frac{f(m)f(\boldsymbol{y}|m)}{f(\boldsymbol{y})}
\end{aligned}$$

where $f(\boldsymbol{y}|m)$ is the marginal distribution of the data under the model m and is called the *marginal likelihood* for model m. It is given by

$$f(\boldsymbol{y}|m) = \int f(\boldsymbol{y}, \boldsymbol{\theta}_m|m)\mathrm{d}\boldsymbol{\theta}_m = \int f(\boldsymbol{y}|\boldsymbol{\theta}_m, m)f(\boldsymbol{\theta}_m|m)\mathrm{d}\boldsymbol{\theta}_m$$

We also see that

$$\frac{f(\boldsymbol{y}|\boldsymbol{\theta}_m, m)f(\boldsymbol{\theta}_m|m)}{f(\boldsymbol{y}|m)} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_m, m)f(\boldsymbol{\theta}_m|m)}{\int f(\boldsymbol{y}|\boldsymbol{\theta}_m, m)f(\boldsymbol{\theta}_m|m)\mathrm{d}\boldsymbol{\theta}_m} = f(\boldsymbol{\theta}_m|\boldsymbol{y}, m)$$

and

$$\frac{f(m)f(\boldsymbol{y}|m)}{f(\boldsymbol{y})} = \frac{f(m)f(\boldsymbol{y}|m)}{\sum_{m=1}^{M} f(m)f(\boldsymbol{y}|m)} = f(m|\boldsymbol{y})$$

The last equation gives the marginal posterior probability of model m which we need to calculate in the presence of model uncertainty. Posterior inference requires evaluation of the marginal likelihood $f(\boldsymbol{y}|m)$ for each model as well as of the posterior distribution $f(\boldsymbol{\theta}_m|\boldsymbol{y}, m)$ of the parameters $\boldsymbol{\theta}_m$ of each model m.

If we want to compare just two models we can use posterior odds to express how one model is better comparing to another. For example, if the two models have prior probabilities $f(1) = p_1$ and $f(2) = p_2 = 1 - p_1$, then the posterior odds in favour of model 1 is

$$\frac{f(1|\boldsymbol{y})}{1 - f(1|\boldsymbol{y})} = \frac{f(1|\boldsymbol{y})}{f(2|\boldsymbol{y})} = \frac{f(1)f(\boldsymbol{y}|1)}{f(2)f(\boldsymbol{y}|2)} = \frac{p_1}{1 - p_1}\frac{f(\boldsymbol{y}|1)}{f(\boldsymbol{y}|2)}$$

The ratio of the marginal likelihoods is called *Bayes factor* and it updates the prior odds to posterior odds after observing data $\boldsymbol{y}$.

In the presence of uncertainty it is sometimes advisable to average over a set of plausible models instead choosing one of them. This method is called *model averaging* and it fully integrates uncertainty in inference rather than condition on the “best” model. It is particularly useful for prediction or estimation in the presence of uncertainty. Assume that Q is a quantity of interest and its posterior distribution is

$$f(Q|\boldsymbol{y}) = \sum_{m=1}^{M} f(Q|m, \boldsymbol{y}) f(m|\boldsymbol{y}) \tag{1.1}$$

with $m = 1, ..., M$ all the models considered. Then, the posterior mean and variance of the quantity of interest are

$$E(Q|\boldsymbol{y}) = \sum_{m=1}^{M} \hat{Q}_m f(m|\boldsymbol{y}) \tag{1.2}$$

$$\mathrm{Var}(Q|\boldsymbol{y}) = \sum_{m=1}^{M} (\mathrm{Var}(Q|\boldsymbol{y}, m) + \hat{Q}_m^2) f(m|\boldsymbol{y}) - E(Q|\boldsymbol{y})^2 \tag{1.3}$$

where $\hat{Q}_m = E(Q|\boldsymbol{y}, m)$ (Hoeting et al., 1999). This method provides better average predictive ability than using any single model and is used in Chapter 5 and 6 in order to obtain more accurate estimates than under one model.

1.3 Generalised Linear Models

Generalised linear models (GLMs) are widely used for regression analysis and extend linear models to describe non-normal responses, like binary and count data. GLMs are used in this thesis, since we wish to make inference and prediction about categorical responses. The other important feature is that the mean is not a linear combination of the parameters but some monotonic function of the mean is.

In general, we can say that a GLM is defined by three points:

1. A probability density for y belonging to the exponential family

2. A linear predictor $\eta = \boldsymbol{X\beta}$

3. A link function g that $E(y) = \mu = g^{-1}(\eta)$

Conditional on the θ, y are independent with probability density function (pdf)

$$f(y;\theta,\xi) = \exp\left(\frac{y\theta - b(\theta)}{\alpha(\xi)} + c(y;\xi)\right) \tag{1.4}$$

where b and c are known functions and θ is the location parameter while ξ the dispersion parameter. For example, the Normal distribution has $\theta = \mu$, $\xi = \sigma^2$, $\alpha(\xi) = \xi$, $\beta(\theta) = \theta^2/2$ and $c(y;\xi) = -1/2(y^2/\sigma^2 + log(2\pi\sigma^2))$. Many other distributions belong to the exponential family like Poisson, Binomial, Beta, Gamma, Multinomial, etc. Common link functions are the log, logit, probit and $\log - \log$ link.

GLMs are models that assume that some monotonic function of the mean is a linear combination of the unknown fixed parameters. When some of the parameters are random variables and the function of the mean consists of fixed and random terms, the model is called *generalised linear mixed model* (GLMM). These random terms are usually assumed to have a Normal distribution with zero mean. They are usually used when units are nested within groups in order to express the group effect and they are called *random effects.* In surveys, the presence of groups in the population suggests stratified or cluster sampling.

Assume that observations are nested within groups and let $j = 1, ..., M$ denote the groups and $i = 1, ..., n_j$ the units within groups. Let y_{ij} be the value of the response variable for unit i in group j and $E(y_{ij}|\boldsymbol{u}_j) = \mu_{ij}$ the conditional mean of y_{ij} that is connected to the parameters through the link function

$$g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{u}_j$$

where $\boldsymbol{x}_{ij}$ is the $p \times 1$ column vector of explanatory variables for that observation, $\boldsymbol{z}_{ij}$ is the $q \times 1$ design vector for the random effects and $\boldsymbol{\beta}$ and $\boldsymbol{u}_j$ are the $p \times 1$ and $q \times 1$ vector of fixed regression parameters and random effects respectively. Usually, $\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. The likelihood of this model is

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \xi) = \prod_{j=1}^{M} \prod_{i=1}^{n_j} \exp\left(\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\alpha_{ij}(\xi)} + c(y_{ij}, \xi) \right)$$

Classical inference for GLMMs requires integrating out the random effects

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}, \xi) = \int f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \xi) f(\boldsymbol{u}|\boldsymbol{\Sigma}) d\boldsymbol{u}$$

and then maximising this new likelihood which is called *integrated likelihood* or *marginal likelihood.* Calculating the maximum likelihood estimators cannot be done analytically and needs numerical solutions.

1.3.1 Bayesian inference for GLMMs

Bayesian inference for GLMMs means that the model is built and analysed hierarchically as mentioned in Section 1.2.3, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{u})$ and $\boldsymbol{\phi} = \boldsymbol{\Sigma}$. In this thesis we examine cases where $\alpha(\xi) = 1$, such as the Multinomial model. Hence, for a Bayesian approach we need to specify a joint prior distribution for the model parameters $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{\Sigma}$.

$$f(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}) = f(\boldsymbol{\beta}) f(\boldsymbol{u}|\boldsymbol{\Sigma}) f(\boldsymbol{\Sigma})$$

where $f(\boldsymbol{u}|\boldsymbol{\Sigma})$ is the already specified $N(\mathbf{0}, \boldsymbol{\Sigma})$. The joint posterior distribution of the parameters is then given by

$$f(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}) f(\boldsymbol{\beta}) f(\boldsymbol{u}|\boldsymbol{\Sigma}) f(\boldsymbol{\Sigma})$$

1.4 Simulation methods

Bayesian applications require extensive computation which is important in order to calculate summaries of the target posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$.

Sometimes, the prior distribution together with the likelihood are of a convenient form and then it is possible to get results by straightforward computations. Nevertheless, it is often that the posterior distribution is very complicated, especially when it is high dimensional. An important tool to summarise posterior quantities is Markov Chain Monte Carlo simulation (MCMC). This is a set of methods to draw sequentially a sample of dependent observations from the normalised density

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})/\int f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$$

The approach is based on constructing the the Markov chain in a way that $\boldsymbol{\theta}^i$ can be considered, at least approximately, a sample from $f(\boldsymbol{\theta})$. This is feasible due to the Markov chain theory result that under appropriate conditions, the distribution of $\boldsymbol{\theta}^i$ converges to the invariant or stationary distribution of that chain.

1.4.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm was described by Hastings (1970) generalising the algorithm of Metropolis et al. (1953). It can draw samples from any probability distribution $f(\boldsymbol{\theta})$ requiring only that a function proportional to the density can be calculated. The normalization factor is often very hard to compute, so the ability to generate a sample without knowing this constant of proportionality is a major virtue of the algorithm. Suppose that the current state of the chain is $\boldsymbol{\theta}^i$. Then a proposal $\boldsymbol{\theta}^*$ is generated from a proposal density $q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)$. With probability

$$\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*) = min\left\{\frac{f(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)}{f(\boldsymbol{\theta}^i)q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)}, 1\right\}$$

the proposal is accepted and the next value of chain $\boldsymbol{\theta}^{i+1}$ is set to $\boldsymbol{\theta}^*$, and with probability $1 - \alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)$ the proposal is rejected and the next value of the chain is set to the current value $\boldsymbol{\theta}^i$. An important point here is that the ratio $f(\boldsymbol{\theta}^*)/f(\boldsymbol{\theta}^i)$ can be replaced by the equivalent ratio of the unnormalised densities.

A common version of Metropolis-Hastings algorithm is to choose $q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)$ to be such that $\boldsymbol{\theta}^* = \boldsymbol{\theta}^i + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is random and its distribution does not depend on $\boldsymbol{\theta}^i$. This is called *random walk* algorithm and if the distribution of $\boldsymbol{\epsilon}$ is symmetric about $\mathbf{0}$ then $q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)$ and the acceptance probability becomes

$$\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*) = min\left\{\frac{f(\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}^i)}, 1\right\}$$

It is common that the distribution of $\boldsymbol{\epsilon}$ is Multivariate Normal with mean $\mathbf{0}$.

1.4.2 The Gibbs sampler

The Gibbs sampler obtains a sample from the joint posterior distribution $f(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J)$ by successively and repeatedly simulating from the conditional distributions of each component given the other components. Hence, there are J steps in iteration t and at each iteration each $\boldsymbol{\theta}_j^{(t)}$ is sampled from

$$f(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j}^{(t-1)}, y)$$

where $\boldsymbol{\theta}_{-j}^{(t-1)}$ represents all the components of $\boldsymbol{\theta}$, except $\boldsymbol{\theta}_j$, at their current values

$$\boldsymbol{\theta}_{-j}^{(t-1)} = (\boldsymbol{\theta}_1^{(t-1)}, ..., \boldsymbol{\theta}_{j-1}^{(t-1)}, \boldsymbol{\theta}_{j+1}^{(t-1)}, ..., \boldsymbol{\theta}_J^{(t-1)})$$

Gibbs sampling can be implemented either in univariate blocks where $\boldsymbol{\theta}_j$ contains just a single component, or in multivariate blocks that contain more than one components. In any case, Gibbs sampler algorithm has the following form:

- Initialise with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{(0)}, ..., \boldsymbol{\theta}_J^{(0)})$
- Simulate $\boldsymbol{\theta}_1^{(1)}$ from the conditional distribution $f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(0)}, ..., \boldsymbol{\theta}_J^{(0)})$

- Simulate $\boldsymbol{\theta}_2^{(1)}$ from the conditional distribution $f(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_3^{(0)}..., \boldsymbol{\theta}_J^{(0)})$
- ...
- Simulate $\boldsymbol{\theta}_J^{(1)}$ from the conditional distribution $f(\boldsymbol{\theta}_J|\boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_2^{(1)}..., \boldsymbol{\theta}_{J-1}^{(1)})$
- Iterate this procedure

Over the next sections of this thesis, a Gibbs sampler is used to sample from the posterior distributions. When some of the required conditional distributions are analytically intractable, we use MCMC algorithms that simulate certain blocks using Metropolis-Hastings updates and the standard conditional distributions directly. These methods are often called *Metropolis-within-Gibbs* algorithms.

1.4.3 Convergence

Various problems might appear when simulating from the posterior distribution of the parameters. First, the simulations may not be enough to describe the target distribution or even when convergence is achieved, the early iterations are still not a representative sample from this distribution. Then, the within sequence correlation of the draws can cause inefficiency in estimation. Moreover, dependence between components of $\boldsymbol{\theta}$ can also create problems when drawing from the posterior distribution. There are ways of dealing with these problems, like discarding the early simulations, *thinning* the sequences, using multiple sequences or tuning the sampler, see Gelman et al. (2004). A method to reduce dependence is *reparameterising* and/or incorporating in the same updating block all components of $\boldsymbol{\theta}$ with high posterior correlation, methods that are both used in this thesis.

In order to apply the above methods we need firstly to diagnose if convergence is achieved or not. There are several suggested ways to do this with the most popular being an informal approach which involves inspecting the time series plot or *trace* plot of components of $\boldsymbol{\theta}$. Through this plot we can often decide on the burn-in number of simulations, whether convergence is achieved or a larger number of draws is required and it is the way we choose to check convergence in this thesis.

1.5 Survey sampling theory

The term *sampling* refers to the process of selecting a sample from a population and may also include the derivation of estimates and inference for the population. *Survey sampling* refers to sampling from *finite* populations. Different sampling methods have been developed with the aim to provide unbiased, efficient and robust estimates of the quantities of interest (Foreman, 1991; Lohr, 1999). Probability sampling is sampling where every unit in the population has a known probability of selection. The main probability sampling designs are: *simple random sampling*, *stratified sampling* and *cluster sampling*. Probability sampling can also be distinguished in *equal probability sampling* and *unequal probability sampling*, with units sampled respectively with equal or unequal probabilities. One method of unequal probability sampling is sampling with *probability proportional to size*, a sampling design which presents particular difficulties and is discussed in Chapter 3 of this thesis. In the following sections, we make inference for different sampling designs for finite populations. Although one may be interested in many different population quantities, the main target when developing theory for sample surveys is usually estimating population means or totals.

Simple random sampling (SRS)

It is the simplest form of probability sample and the foundation for more complex designs. Each unit is chosen randomly and each subset of n units has the same probability of being chosen in the sample as any other subset of n units.

Stratified sampling

Stratification is the process of grouping members of the population into relatively homogeneous subgroups (strata) before sampling. The strata should be mutually exclusive (every element in the population must be assigned to only one stratum) and collectively exhaustive (no population element can be excluded). Then a simple random sample is selected from each stratum. Since the strata are homogeneous, elements in the same stratum are more similar than other elements in the population and variance within stratum is lower than variance in the whole population. This means that stratification often increases precision.

Sometimes there is a desirable stratification variable the distribution of which could be known from previous surveys, but the sampling frame does not include information on it and the design of a stratified sample is not possible. In this case, an SRS sample can be taken, then create the post-strata by classifying the sampled elements according to the stratification variable. This procedure is called *post-stratification.* Problems appear when the sample size is small and results in very small or even zero stratum sizes. Thus, the calculation of the stratified estimator becomes hard and its variance can be infinite or undefined (Little, 2004). In this case, the classical approach is to re-sample until the desirable stratum sizes are obtained.

Cluster sampling

Cluster sampling is a technique used when natural groupings are evident in a population, like schools, households, dwellings and a sample of them is selected. Then, only units from the selected groups are included in the sample.

The technique works best when most of the variation in the population is within the groups and not between. In a single stage cluster sampling, all the units from each of the selected clusters are used, while in two stage cluster sampling a random sample of the units within the selected clusters is taken. The clusters are the primary sampling units (PSU) and the units within the clusters the secondary sampling units (SSU).

Weights

Sampling weights are used to correct for known discrepancies between the sample and the population. These are caused by imperfections in the sample like unequal selection probabilities, non-coverage of the population and non-response. Usually the construction of weights starts with a *base weight* for each sampled unit to correct for unequal probabilities of selection and it is exactly the inverse of the probability of selection p_i. Thus,

$$w_i = 1/p_i$$

For multistage designs the *weights* reflect the probabilities of selection in each stage. Suppose that two stage cluster sampling is performed. If there are M clusters in the population from which m are sampled, then clusters are selected first with probability:

$$p_j = m/M$$

and then units within clusters are selected with probabilities

$$p_{i|j} = n_j/N_j$$

where n_j is the number of sampled units within cluster j and N_j the total number of units in cluster j. Finally, the overall probability of selection of a unit within a cluster is

$$p_{ij} = p_j p_{i|j}$$

and the weight for this unit is the reciprocal of the probability

$$w_{ij} = 1/p_{ij}$$

1.6 Bayesian Inference for surveys

Suppose we have a population that consists of N units and after sampling there are two parts: the sampled n units and the non-sampled $N - n$ units. Let Y_S denote the sampled part and $Y_{\bar{S}}$ the non-sampled part. A model for the survey outcome Y is required, which is then used to predict the non-sampled values of the population and hence finite population quantities Q. We begin from the joint prior distribution $p(Y)$ for all the population values. Then, inference for finite population quantities $Q(Y)$ is based on the posterior predictive distribution $P(Y_{\bar{S}}|Y_S)$ of the non-sampled values, given the sampled. The specification of the joint prior distribution is done through a parametric model $p(Y|\theta)$ (where θ is an unknown hyperparameter) combined with a prior for θ

$$p(Y) = \int p(Y|\theta)p(\theta)\mathrm{d}\theta$$

The posterior predictive distribution of $Y_{\bar{S}}$ is then

$$p(Y_{\bar{S}}|Y_S) \propto \int p(Y_{\bar{S}}|Y_S, \theta)p(\theta|Y_S)\mathrm{d}\theta$$

where $p(\theta|Y_S)$ is the posterior distribution of the parameters. This posterior produces the posterior distribution $p(Q|Y_S)$ for any finite population quantity.

A brief example is given here from Forster and O'Hagan (2004) to describe finite population inference. Suppose we have y_i, $i = 1, ..., N$, independently and identically distributed with common distribution $f(y|\theta)$ and a prior distribution for parameters θ, $f(\theta)$. Then, the joint distribution of y_i is

$$f(y_1, ..., y_N) = \int \prod_{i=1}^{N} f(y_i|\theta)f(\theta)\mathrm{d}\theta$$

and conditioning on the observed y_i, $i = 1, ..., n$, gives

$$f(y_{n+1}, ..., y_N|y_1, ...y_n) = \int \prod_{i=n+1}^{N} f(y_i|\theta)f(\theta|y_1, ..., y_n)\mathrm{d}\theta$$

which is the posterior predictive distribution of $N-n$ new observations. Now, if we assume that $f(y|\theta,\tau)$ is $N(\theta,\tau)$ and the prior distribution of θ is $N(\mu,\omega)$, then the posterior of θ is $N(\mu_1,\omega_1)$ where $\mu_1=(n\omega\bar{y}+\tau\mu)/(n\omega+\tau)$, $\omega_1=\tau\omega/(n\omega+\tau)$. The predictive distribution of each unobserved y_i is then $N(\mu_1,\tau+\omega_1)$. For inference about a quantity like $Q=\sum_i^N y_i$ we need the posterior predictive distribution of the sum of $N-n$ unobserved y_is, which is $N((N-n)\mu_1,(N-n)\tau+(N-n)^2\omega_1)$. Finally, the posterior distribution of t is Normal with mean

$$E(t|y_1,...,y_n)=n\bar{y}+(N-n)\mu_1$$

and with variance

$$var(t|y_1,...,y_n)=(N-n)\tau+(N-n)^2\omega_1$$

In general, inferences about any population quantity of interest Q are obtained by first conditioning on the parameters $\boldsymbol{\theta}$ and then averaging over posterior of $\boldsymbol{\theta}$. Hence, the posterior predictive mean is

$$E(Q|Y_S)=E(E(Q|Y_S,\boldsymbol{\theta})|Y_S) \tag{1.5}$$

and the posterior predictive variance is

$$Var(Q|Y_S)=E(Var(Q|Y_S,\boldsymbol{\theta})|Y_S)+Var(E(Q|Y_S,\boldsymbol{\theta})|Y_S) \tag{1.6}$$

1.7 The Problem

In this thesis we are mainly interested in univariate and multivariate categorical responses (contingency tables). The majority of examples in the literature describe regression models where explanatory variables are known for the non-sampled cases or estimation of group totals/means where groups sizes in the population are also available to the data analyst. Our approach to the estimation of finite population quantities is from a different point of view.

We make the distinction between two different type of statisticians, the *data analysts* and the *survey statisticians*, see Breidt and Opsomer (2007). The survey statisticians are the people who organise the surveys, collect the data and thus have access to design variable values. On the other hand, data analysts are the individuals who are interested to analyse the data and usually have limited information about the sampling design. Moreover, confidentiality issues could restrict access to complete information which is available to survey statisticians. We focus on addressing the problem of inference about population totals or means from the *data analyst* point of view. Therefore, design variables and other explanatory variables are not recorded for the non-sampled cases.

Since the design variables are not recorded for non-sampled units, one cannot use regression models to predict for them. Usually, the information given in the documentation of the survey includes the type of the sampling, which are the design variables, the number of units in the population N and the number of the population strata/clusters. Details such as the population stratum/cluster size is rarely given although it is essential for a finite population analysis. In Bayesian inference for surveys, design variable values are usually assumed known Gelman (2007); Little (2004) but there are cases where the problem is addressed. Gelman (2007) mentions that “in some cases the cell populations are unknown and must be estimated” and Little and Zheng (2007) discuss a model for probability proportional to size when size variable is unknown for non-sampled cases.

Thus, we first address the problem of not knowing the design variable values which we model and then predict for the non-sampled part. We emphasise on the analysis of polytomous variables and contingency tables, where finite population methodology is limited. This is evident from the literature review, where one finds many references on Bayesian models for continuous finite population responses but no examples are found for discrete responses. In this thesis, polytomous variables and contingency tables are modelled using the Multinomial model and random effects are added to account for the design variable effect.

Several issues appear, such as high dimensionality and parameter non-identifiability problems that affect convergence of the chains. In addition, the absence of design variable values for non-sampled cases makes prediction through regression models impossible. Hence, we limit inference in obtaining point estimates for population cell totals. Finally, we deal with the model selection problem by calculating posterior model probabilities, compare plausible models and calculate model-averaged posterior distributions.

The software used in this thesis to sample, run MCMC algorithms, calculate posterior summaries is R Development Core Team (2011) together with various R packages like MCMCpack, (Martin et al., 2010) and Survey Sampling, (Till and Matei, 2009).

1.8 The Data

We use the dataset of the Health Education Population Survey in Scotland in 2002 obtained from the Economic and Social Data Service (`http://www.esds.ac.uk`) which is a national data archiving in operation since January 2003. The service is a jointly-funded initiative sponsored by the Economic and Social Research Council (ESRC) and the Joint Information Systems Committee (JISC) . The Health Education Population Survey (HEPS) monitored health-related knowledge, attitudes, behaviour and behavioural motivations amongst adults (aged 16-74) in mainland Scotland. The survey ran from 1996 to 2007 with a nationally representative annual sample of around 1800 individuals and a response rate around 70%. Further information about the survey can be found in `http://www.healthscotland.com`.

We use the 2002 dataset collected in two waves and contained questions about eneral health, diseases, nutrition, physical activity, alcohol, smoking, work-life balance, etc. The dataset is particularly suitable for the analysis in this thesis since it consists of several categorical variables and also variables that can be used as stratification or cluster variables. In order to be able to check final results validity, we handle the sample taken from the survey as the *finite population* for which we wish to make inference. This population consists of $N = 1742$ units from which we resample under different sampling designs. The following table shows some of the variables included in this dataset. Some variables are used as stratification variables, such as *age in categories*, *sex*, *social status* and some others are suitable for cluster sampling, for example *area* and *district*. We use *health status* as the response variable Y in Chapter 4. In Chapter 5 and 6 contingency tables are constructed using *health status* and various of the other variables.

Table 1.1: Variables in the dataset

Variable	Description
Health status	1=very good, 2=good, 3=fair, 4=poor, 5=very poor
Longstanding illness	1=yes, 2=no
Lifestyle	1=very healthy, 2=fairly healthy, 3=fairly unhealthy, 4=very unhealthy
Social grade	4 categories
Marital status	1=single, 2=maried/living as couple, 3=widowed/divorced/separated
Sex	1=male, 2=female
Age category	1=16-24, 2=25-34, 3=35-44, 4=45-54, 5=55-64, 6=65-74
Smoker	0=no, 1=yes
Alcohol consumption	0=no, 1=yes
Exercise	0=no, 1=yes
Diet with fruits & vegetables	0=no, 1=yes
Area	140 neibourhouds
District	43 districts
Health region	7 health regions

Chapter 2

Previous work

This chapter is divided in two main parts: first, discussion on survey sampling theory and already implemented Bayesian models for finite population quantities and second, description of the existing methodology for multivariate categorical responses. Both parts are important for the analysis conducted in this thesis, since we implement models for categorical survey data and in particular for responses with more than two categories and contingency tables.

2.1 Design based inference for surveys

We start by describing design based inference for surveys and the problems that this methodology presents. Assume a population with N units, $Y = (y_1, ..., y_N)$ where y_i is a set of survey variables for unit i and let be $I = (I_1, ..., I_N)$ the set of *inclusion indicator variables* where

$$I_i = \begin{cases} 1 & \text{if unit } i \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases}$$

In *design-based inference* or *randomisation theory*, y_is are considered to be fixed and any probabilities used arise from the probabilities of selecting units to be in the sample. It is a non-parametric approach since no assumptions are made about the distribution of y_i. For example, if we have a simple random sample of size n then the sample mean is

$$\bar{y} = \sum_{i \in S} \frac{y_i}{n} = \sum_{i=1}^{N} I_i \frac{y_i}{n}$$

and I_is are the only random variables here and are identically distributed Bernoulli random variables with

$$\pi_i = P(I_i = 1) = \frac{n}{N}$$

Now, suppose we want to estimate the finite population mean $\bar{Y}$ from a stratified random sample. The population is divided into J strata and N_j is the *known* population count in stratum j. Assume also $y_{ij}, j = 1, ..., J; i = 1, ..., n$ are the set of sampled Y in stratum j and n_j the sum of sampled units in stratum j. The quantity of interest is

$$Q = \bar{Y} = \sum_{j=1}^{J} P_j \bar{Y}_j$$

where $P_j = N_j/N$ is the proportion of the population in stratum j. The usual estimator of $\bar{Y}$ is the stratified mean

$$\hat{q} = \bar{y}_{st} = \sum_{j=1}^{J} P_j \bar{y}_j$$

where $\bar{y}_j$ is the sample mean in stratum j. This estimator is also a weighted mean of the sampled units, where units in stratum j are weighted by the inverse of their selection probability $\pi_j = n_j/N_j$. The estimated variance of the stratified mean is

$$\hat{v}_{st} = \sum_{j=1}^{J} P_j^2 s_j^2 (1/n_j - 1/N_j)$$

where s_j^2 is the sample variance in stratum j.

Next, we consider inference for a cluster sample where sampling is performed in two stages. The primary sampling units (PSU) are the clusters while the secondary sampling units (SSU) are the elements observed within the clusters. Let M denote the number of clusters in the population with N_j number if units within j cluster, $N = \sum_j N_j$ the total number of units in the population and m, n_j, n the corresponding sample quantities. Then, the estimator of the population total in cluster j is

$$\hat{t}_j = \sum_i \frac{N_j}{n_j} y_{ij} = N_j \bar{y}_j$$

and an unbiased estimator of the population total is

$$\hat{t} = \frac{M}{m} \sum_j \hat{t}_j = \frac{M}{m} \sum_j N_j \bar{y}_j$$

An unbiased estimator of the variance of $\hat{t}$ is

$$\hat{V}(\hat{t}) = M^2 \left(1 - \frac{m}{M}\right) \frac{s_t^2}{m} + \frac{M}{m} \sum_j \left(1 - \frac{N_j}{n_j}\right) N_j^2 \frac{s_j^2}{n_j}$$

where

$$s_t^2 = \frac{\sum_j (\hat{t}_j - \frac{\hat{t}}{M})^2}{m-1}$$

and

$$s_j^2 = \frac{\sum_i (y_{ij} - \bar{y}_j)^2}{n_j - 1}$$

All the above discussion is about sampling with equal probabilities but sometimes it is useful to sample with unequal probabilities. When sampling with unequal probabilities, we deliberately vary the probabilities of selection of the PSU's and compensate by providing suitable weights in the estimation. One common way is sampling with probability proportional to size (PPS), in which the probability that a particular sampling unit will be selected in the sample is proportional to the population size of that sampling unit. Then a large sampling unit has a greater chance of being in the sample than a small one. Moreover, two ways of sampling with unequal probabilities exist, sampling with replacement or without replacement (Lohr, 1999). Sampling with replacement ensures that the probabilities of selection do not alter when a unit is drawn but it is less efficient than sampling without replacement. Although sampling without replacement is more complicated because the probability of a unit selected is different for the first unit chosen than for the second and subsequent units, it is the method we use to sample with PPS in this thesis. In practice, there are many methods of sampling with unequal probabilities and without replacement, some easier and some harder to implement, see Hanif and Brewer (1980), but we are particularly interested in PPS sampling methods.

Let $\pi_i = P(i \text{ unit selected first})$ be the inclusion probability of unit i in without replacement sampling, where $0 < \pi_i \leq 1$ and

$$\sum_{i=1}^{N} \pi_i = n$$

Then, π_{ij} is the probability that units i and j are both in the sample and if we define the *average probability* that a unit i will be selected on one of the draws as π/n, we get the *Horvitz-Thompson estimator* (HT) Horvitz and Thompson (1952) of the population total as

$$\hat{t}_{HT} = \frac{1}{n} \sum_i \frac{\hat{t}_i}{\pi_i/n} = \sum_i \frac{\hat{t}_i}{\pi_i} \tag{2.1}$$

This estimator is unbiased and one unbiased estimator of its variance is

$$\hat{V}(\hat{t}_{HT}) = \sum_{i \in S} (1 - \pi_i) \frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i \in S} \sum_{k \in S k \neq i} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{\hat{t}_i}{\pi_i} \frac{\hat{t}_k}{\pi_k} + \sum_{i \in S} \frac{\hat{V}(\hat{t}_i)}{\pi_i} \tag{2.2}$$

which unfortunately can result in a negative estimate of the variance, Lohr (1999). In general, calculating the estimator of the variance of $\hat{t}_{HT}$ is troublesome and the use of the with-replacement variance is an alternative. Another problem here is that in order to use the HT estimator for $n > 1$ the inclusion probabilities must be known for every PSU, that is also hard to implement when $n > 2$ and the population is large.

2.2 Bayesian inference for surveys

Bayesian inference for surveys was briefly introduced in Section 1.6 and now we focus more on existing Bayesian models for continuous responses as often considered by Little (2004), Little and Zheng (2007), Gelman (2007). Little (2004) argues in favour of Bayesian modelling that offers great advantages such as:

- It provides a unified approach to survey inference which in large samples and uninformative prior distributions can give results similar to design based inference.
- It can handle for complex design features, when known in advance.
- It provides better inference for small sample problems where frequentist solutions are not available.
- It can incorporate prior information when available.
- It satisfies the likelihood principle and can outperform the design based inference if the model is well specified.

Before we start the discussion on Bayesian models for finite population quantities, we need to define what *ignorable* design is. As already defined $I = (I_1, ..., I_N)$ are the inclusion indicators for N population units and $Y = (Y_1, ..., Y_N)$ the variable of interest. Bayesian inference for a quantity $Q = Q(Y)$ should be based on the joint distribution of I and Y. However, inference can be based on the distribution of Y alone when the sampling mechanism is *non-informative* or *ignorable.* This is the case with probability sampling where the distribution of I given Y does not depend on the values of I (Little, 2004; Gelman et al., 2004; Rubin, 1983). Hence, if $Z = (Z_1, ..., Z_N)$ is the set of the design variables, then the probability of inclusion in the sample $\Pr(I_i = 1|Y, Z)$ is not dependent on Y and

$$\Pr(I_i = 1|Y, Z) = \Pr(I_i = 1|Z)$$

In this thesis, we assume probability sampling and the probability of a unit included in the sample depends on the design variables Z and not on Y. Then inference can be made based on the distribution of Y alone.

2.2.1 Models for stratified and cluster samples

Assume again the population is divided in M strata and a simple random sample is taken within stratum j for $j = 1, ..., M$. The variable of interest Y is continuous and a common model for *continuous* outcomes assumes that y_{ij} (value of Y for unit i in stratum j) is normal with mean μ_j and variance σ_j^2, Little (2004). A Bayesian model with noninformative prior distribution, that Little (2004) mentions, is

$$p(y_{ij}|z_{ij} = j, \mu_j, \sigma_j^2) \overset{iid}{\sim} N(\mu_j, \sigma_j^2)$$

with

$$p(\mu_j, \log \sigma_j^2) = const.$$

Recall the notation for the sampled and non-sampled part of Y as Y_S for the sampled and $Y_{\bar{S}}$ for the unsampled. With known variances, the posterior distribution of the population mean $\bar{Y}$ given Y_S, I and σ_j^2 is normal with

$$E(\bar{Y}|Y_S, I, \sigma_j^2) = \bar{y}_{st} = \sum_{j=1}^{M} P_j \bar{y}_j$$

$$Var(\bar{Y}|Y_S, I, \sigma_j^2) = v_{st} = \sum_{j=1}^{M} P_j^2 \sigma_j^2 (1/n_j - 1/N_j)$$

Hence, the posterior mean is the stratified mean from design based inference and if σ_j^2s are replaced by estimates s_j^2, the posterior variance equals the design-based variance, a substitution that according to Little (2004) is justified asymptotically. Note that the $P_j = N_j/N$ are assumed known, that means the size variable N_j for the population strata is known. Also, the factors $(1 - n_j/N_j)$ are finite population corrections that emerge automatically in Bayesian analysis. Stratification weights can help when the population proportions P_j are not known and the estimator can be written

$$\hat{\bar{y}}_{st} = \sum_{j=1}^{M} P_j \bar{y}_j = \sum_{j=1}^{M} w_j n_j \bar{y}_j / \sum_{j=1}^{J} w_j n_j$$

and the model-based approach replaces $\bar{y}_j$ by prediction $\hat{\mu}_j$ from the model.

In general, it is advisable to take into account stratum effects when stratification is present since strata construction is usually based on characteristics likely to be related to the survey outcome. Gelman (2007) argues that analysis should incorporate all variables affecting selection or nonresponse. Little (2004) claims that modelling the differences across groups is important for a well-specified model. Again, the proportions of groups in the population P_js are assumed to be known, otherwise a "supplemental model is needed to allow estimation of these proportions from the sample" (Little, 2004).

A special case where Bayesian modelling can improve inference is when post-stratification is applied. Sometimes, there is a desirable stratification variable that the sampling frame does not include information on but its distribution is known from previous surveys or census. Now, $P_j = N_j/N$ is the proportion of the population in post-stratum j. Then, a random sample of size n is taken from the population and n_j of the N_j units in post-stratum j are included in the sample. The estimator here has the same form as the stratified estimator, but inference now changes by the fact that n_js are now random depending on the sampling distribution. Thus, there is a non-zero probability that $n_j = 0$ for some j that makes $\bar{y}$ undefined. In this case the variance of the estimator is also undefined or maybe infinite. A Bayesian model with random effects can improve inference by allowing for borrowing strength for the prediction in small post-strata. It is evident that adding random effects in the model increases robustness, even in a stratified sample, especially when there are many strata and small samples.

When natural clusters exist in the population it is common in surveys to sample clusters first and then units within clusters. To incorporate this feature in the Bayesian analysis, random effects are introduced in the model. Assume the population consists of M clusters, like geographical areas and we sample a number m of these clusters. Next, we select a simple random sample of n_j units in each sample cluster j. The sampling mechanism is ignorable if we condition on cluster information but the Bayesian model needs to account for within-cluster correlation. A normal model that does this is

$$y_{ij}|\theta_j, \sigma^2 \sim N(\theta_j, \sigma^2)$$

$$\theta_j|\mu, \tau^2 \sim N(\mu, \tau^2)$$
$$p(\mu, \tau^2, \sigma^2) = const.$$

Similar models with random effects are used in this thesis but for a multivariate categorical response variable. Therefore, we use a GLMM version of the above models in Chapters 3 and 4.

2.2.2 Models for PPS samples

The Horvitz-Thompson estimator applies the idea of weighting the units more generally but design based inference becomes more troublesome. One of the major disadvantages of using this estimator is the complications when calculating its variance estimator, as mentioned in Section 2.1. The other disadvantage is that the HT estimator can have a high variance when an outlier in the sample has a low selection probability and so it receives a large weight. Little and Zheng (2003) consider alternatives to the HT estimator that assume a smoothly-varying relationship between y_j and the inclusion probability π_j using penalized splines. Their method is for PPS sampling and continuous outcomes as following

$$y_j = f(\pi_j, \beta) + \epsilon_m, \ \epsilon_j \overset{iid}{\sim} N(0, \pi_j^{2k}\sigma^2)$$

where π_j is the selection probability for unit j, k takes values $0, 1/2$ or 1 to model error heteroskedasticity and the function f is a p-spline written as a linear combination of truncated polynomials. They simulated different artificial populations and calculate the root mean squared error of point estimates in order to compare between different estimates. They conclude that p-spline model based estimators are generally more efficient than HT estimators.

More recently, Little and Zheng (2007) examined the case where the size measure Z is not recorded for the non-sampled units. While this information is essential to implement PPS sampling, it is usually not available in public use data files. Here the full Bayesian approach requires a supplemental model for the design variables in order to predict their values for the non-selected cases. Little and Zheng (2007) conclude by showing how this can be done through a Bayesian Bootstrap (BB) model for the size variable, modified to account for PPS sampling. We next describe how the BB model works and the theory behind it.

Assume PPS sampling where the selection probabilities π_j (or the size variable z_j) are only available for sampled j but the total number of non-sampled cases $M - m$ is known. In the BB model, predictions of the sizes $\tilde{z}_j$ for non-sampled cases have to be drawn from the posterior distribution given the data and that these units are not selected. Then, Little and Zheng (2007) use the same penalized spline model as before to draw y_j from the posterior distribution of y given the drawn value of z_j. The problem appears during the first part, since sample design becomes informative when sizes are unknown for the non-sampled units. The posterior distribution of the sizes given non-selection is related to the posterior distribution of sizes given selection as

$$\begin{aligned} p(z|data, i=0) &= cp(z|data, i=1)p(i=0|z, data)/p(i=1|z, data) \\ &= cp(z|data, i=1)(1-\pi(z))/\pi(z) \end{aligned}$$

where c is a normalising constant. Then, this predictive distribution is described through a BB model. Let $\{x_1, ..., x_K\}$ be the set of distinct sizes for the sampled units and v_k the number of sampled clusters with size x_k, $\sum_{k=1}^{K} v_k = v$. Assume that these counts are multinomial with probabilities $\{\rho_1, ..., \rho_K\}$ which are assigned a Dirichlet$(0, ..., 0)$ prior distribution.

$$(v_1, ..., v_K|\rho_1, ..., \rho_K, i = 1) \sim \text{Multinomial}(v; (\rho_1, ..., \rho_K)p(\rho_1, ..., \rho_K) \propto \prod_{k=1}^{K} \rho_k^{-1}$$

This model makes the assumption that only the selection probabilities that arise are those seen in the sampled clusters. Little and Zheng (2007) claim that this assumption does not seriously impact the resulting inferences. The posterior distribution of $\{\rho_1, ..., \rho_K\}$ is Dirichlet$(v_1, ..., v_K)$

$$p(\rho_1, ..., \rho_K) \propto \prod_{k=1}^{K} \rho_k^{v_k - 1}$$

If v_k^* is the number of non-sampled clusters with size x_k, $\sum_{k=1}^{K} v_t^* = M - m$, then the posterior predictive distribution of these counts is Multinomial

$$(v_1^*, ..., v_K^*) \sim \text{Multinomial}(M - m; (\rho_1^*, ..., \rho_K^*))$$

where $\rho_k^* = c\rho_k(1 - \pi_k)/\pi_k$ where $\pi_k = mx_k/M\bar{x}$ is the selection probability for units with size x_k and c is chosen so that $\sum_{k=1}^{K} \rho_k^* = 1$. Thus, the model suggests drawing values ρ_k and then drawing predicted counts v_k^* using the last two equations.

We compare this BB model with a parametric model suggested for PPS sampling in Section 3.6 and test both models through different ways of sampling with PPS.

2.2.3 Regression models for surveys

Many references in the literature review about modelling for surveys seem to focus on regression modelling and how regression estimates can approximate design-based estimates. Often, sampling weights are used in modelling in a way that is still controversial, as Pfeffermann (1993) comments while trying to answer questions about the use of sampling weights in modelling. For example, the estimator of the regression coefficient can be written as

$$\hat{\boldsymbol{\beta}}_w = (\boldsymbol{X}_s^T \boldsymbol{W}_s \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s^T \boldsymbol{W}_s \boldsymbol{Y}_s = \left(\sum_{i \in S} w_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \sum_{i \in S} w_i \boldsymbol{x}_i y_i$$

where $w_i = 1/\pi_i$, $\boldsymbol{x}_i$ is the vector of covariates for unit i, s denotes the sample, $\boldsymbol{W}_s = diag(w_1, ..., w_n)$, $\boldsymbol{Y}_s = (y_1, ..., y_n)^T$. Pfeffermann (1993) concludes that weights can help to protect against nonignorable sampling and model misspecification depending on the survey design and the form of the available data. See Pfeffermann (1993) for a review of several approaches for including the weights in the modelling procedure.

In this Section we describe briefly some Bayesian regression models, as they are not the main interest in this thesis. Note that in all these models covariates X are known for all the population units. Gelman (2007) reviews hierarchical regression together with post-stratification as a strategy for correcting for differences between sample and population. The goal is to estimate the population total or mean or the coefficients of a regression model for survey data. He focuses on the relation between the model for survey response and the corresponding weighted-average estimate. Also, Gelman (2007) aims to have a model based procedure for constructing weights or create methodology for regression modelling that gives efficient and approximately unbiased estimates.

The notation used is y, z for variables that are observed in the sample only and X for variables that are observed in the sample and are known in the population. There are variables X whose joint distribution in the population is known and an outcome y whose population distribution we are interested in estimating. Gelman (2007) assumes X to be discrete and labels all the possible categories of X as post-stratification cells j with population sizes N_j and sample sizes n_j. The population size is $N = \sum_{j=1}^{J} N_j$ and the sample size is $n = \sum_{j=1}^{J} n_j$. He assumes again that the population size N_j of each class j is known and these categories include all the cross-classifications of the predictors X. In the case where they are unknown, they have to be estimated usually from previous surveys. Then the population mean of any response can be written as a sum over post-strata

$$\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j}$$

and the estimate is

$$\hat{\theta} = \frac{\sum_{j=1}^{J} N_j \hat{\theta}_j}{\sum_{j=1}^{J} N_j} \tag{2.3}$$

where θ_j the population mean in group j and θ_j the sample mean in group j.

Gelman (2007) distinguishes between *unit weights* w_i, $i = 1, ..., n$ and *cell weights* $W_j = n_j w_i$ for units i within cell j. Then the weighted average is defined as

$$\bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{j=1}^{J} W_j \bar{y}_j}{\sum_{j=1}^{J} W_j}$$

Gelman (2007) then uses regression modelling to connect weighting and post-stratification by applying the idea to work with the post-stratified estimator which under certain conditions can be reinterpreted as a weighted average. A regression model that includes information about the post-stratification cells without including all the interactions is the following $y \sim N(X\beta, \Sigma_y)$ with a prior distribution on β of the form $N(0, \Sigma_\beta)$. Also, X is the $n \times k$ matrix of predictors in the data and X_{pop} the $J \times k$ matrix of predictors for the J post-stratification cells. The vector of post-stratum populations is $N_{pop} = (N_1, ..., N_J)$ with $N = \sum_{j=1}^{J} N_j$. The estimated vector of regression coefficients is then $\hat{\beta} = (X^T \Sigma^{-1} X + \Sigma_\beta^{-1})^{-1} X^T \Sigma^{-1} y$ and the post-stratified estimator of the population is

$$\hat{\theta} = \frac{1}{N} \sum_{j=1}^{J} N_{pop}^T X_{pop} (X^T \Sigma^{-1} X + \Sigma_\beta^{-1})^{-1} X^T \Sigma^{-1} y$$

The vector of unit weights (renormalised to sum to 1) is written

$$w = \Big(\frac{n}{N}(N_{pop})^T X_{pop}(X^T\Sigma^{-1}X + \Sigma_\beta^{-1})^{-1}X^T\Sigma^{-1}\Big)^T$$

which is a vector of n that takes at most J distinct values. The vector of J possible unit weights (corresponding to units in each of the J post-strata) is

$$w_{pop} = \Big(\frac{n}{N}(N_{pop})^T X_{pop}(X^T\Sigma^{-1}X + \Sigma_\beta^{-1})X_{pop}^T\Sigma^{-1}\Big)^T$$

Then, using an example of an exchangeable normal model for the J cell means, writing the posterior means of the cell means θ_k as a linear combination of the cell means $\bar{y}_k$ and using some appropriate approximations, Gelman (2007) manages to express the units weights as weighted average of the full post-stratification unit weight $\frac{N_j/N}{n_j/n}$ and the completely smoothed weight of 1.

2.3 Inference for polytomous variables

2.3.1 Introduction

In this thesis, we are mainly interested in analysing categorical responses with many categories, where the Multinomial model is suitable. The Multinomial model is often encountered in social statistics problems, market surveys, transportation and travel behaviour modelling, spatial or longitudinal data, health services surveys, etc. The variable of interest has more that two categories and several explanatory variables may affect the response variable. It also common that the subjects are observed within clusters or are repeatedly measured. In this case, observations from the same cluster are usually correlated and a mixed effects regression model is necessary. There could also be individual-specific covariates, group-specific and/or even choice-specific covariates. A Multinomial model can also be used in contingency table analysis, as described in Chapter 5. We start by describing the Multinomial distribution in general and then the Multinomial model as a GLMM.

Consider random variables $y_1,^* ..., y_n^*$ that may take one of several discrete values called *categories* and indexed $1, 2, ..., C$ with probabilities p_k for $k = 1, ..., C$ and $\sum_k p_k = 1$. The likelihood of the model is then

$$f(\boldsymbol{y}^*|\boldsymbol{p}) = \prod_{i=1}^{n}\prod_{k=1}^{C} p_k^{I[y_i^*=k]}$$

Let y_k be the number of y_i^*s that fall in category k and $\sum_k y_k = n$. When we only observe the y_ks the likelihood becomes

$$f(\boldsymbol{y}|\boldsymbol{p}) = \binom{n}{y_1, y_2, ..., y_C}\prod_{k=1}^{C} p_k^{y_k}$$

and is known as the *Multinomial distribution.*

2.3.2 The Multinomial logit model

The Multinomial distribution belongs to the exponential family of distributions, hence to construct a GLMM we need to define the link function and the linear predictor. Following the notation of Section 1.3, let j denote the groups with $j = 1, ..., M$ and i denotes the units nested within groups with $i = 1, ..., n_j$. The response variable y_{ij} can take $k = 1, ..., C$ discrete values with probabilities $\boldsymbol{p}_{ij} = (p_{ijk}, ..., p_{ikC})$. Moreover,

$$\sum_k y_{ijk} = 1, \sum_k p_{ijk} = 1 \quad \text{and} \quad p_{ijk} \geq 0$$

The likelihood of this model is

$$f(\boldsymbol{y}|\boldsymbol{p}) = \prod_{j=1}^{M}\prod_{i=1}^{n_j}\prod_{k=1}^{C} p_{ijk}^{y_{ijk}} \tag{2.4}$$

and

$$\boldsymbol{y}_{ij} \sim \text{Multinomial}(\boldsymbol{p}_{ij}; 1)$$

If y_{ij} takes $k = 1, ..., C$ discrete values with probabilities $\boldsymbol{p}_j = (p_{jk}, ..., p_{kC})$, which means that all units within groups have same probability of falling into a category, then the likelihood is

$$f(\boldsymbol{y}|\boldsymbol{p}) \propto \prod_{j=1}^{M}\prod_{k=1}^{C} p_{jk}^{y_{jk}} \tag{2.5}$$

where y_{jk} is the number of units in group j that fall into category k and $\sum_i \sum_k y_{ijk} = n_j$ the number of units in group j. Hence, in this case

$$\boldsymbol{y}_j \sim \text{Multinomial}(\boldsymbol{p}_j; n_j)$$

Now consider the probabilities

$$p_{ijk} = \Pr(y_{ij} = k)$$

and models where these probabilities depend on a vector $\boldsymbol{x}_{ij}$ of fixed covariates and group-specified random effects $\boldsymbol{u}_j$. To create the Multinomial logit model we nominate one category as the baseline or reference category calculate log-odds for all other categories relative to that one. Then, let the log-odds be a linear function of the predictors and/or the random effects. For the following analysis we use the first category as the baseline category. Hence, we have

$$\eta_{ijk} \equiv \log \frac{p_{ijk}}{p_{ij1}} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k + \boldsymbol{z}_{ij}^T \boldsymbol{u}_{jk}$$

The Multinomial logit model presents an extra difficulty when comparing it with the general description of GLMMs in Section 1.3 because this is a multivariate model and parameters $\boldsymbol{\beta}_k$ and $\boldsymbol{u}_{jk}$ depend on k. Thus, $\boldsymbol{x}_{ij}$ is the $p \times 1$ covariate vector and $\boldsymbol{z}_{ij}$ the design vector for the q random effects. Correspondingly, $\boldsymbol{\beta}_k$ is $p \times 1$ vector of unknown fixed regression parameters and $\boldsymbol{u}_{jk}$ is an $q \times 1$ vector of unknown random effects for the group j. Writing the same in vector form we get

$$\boldsymbol{\eta}_{ij} = \left(\log \frac{p_{ij2}}{p_{ij1}}, ..., \log \frac{p_{ijk}}{p_{ij1}}\right) = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{u}_j \tag{2.6}$$

where $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_2^T, ..., \boldsymbol{\beta}_{C-1}^T)$ is the vector of coefficients for every category, $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}$ are model matrices for the fixed and random effects and $\boldsymbol{u}_j^T =$

$(\boldsymbol{u}_{j2}^T, ..., \boldsymbol{u}_{jC-1}^T)$ the random effects. The fixed effects matrix for each unit is a $(C-1) \times p(C-1)$ matrix with non-zero elements

$$\boldsymbol{X}_{ij} = \begin{pmatrix} 1 & \boldsymbol{x}_{ij}^T & & & & & & \\ & & 1 & \boldsymbol{x}_{ij}^T & & & & \\ & & & & \ddots & & & \\ & & & & & & 1 & \boldsymbol{x}_{ij}^T \end{pmatrix}$$

The Multinomial logit model may also be written in terms of the original probabilities p_{ijk} rather than the log-odds as

$$\begin{aligned} p_{ijk} &= \frac{\exp(\eta_{ijk})}{1 + \sum_{l=2}^{C} \exp(\eta_{ijl})} \quad \text{for } k = 2, 3, ..., C \\ p_{ij1} &= \frac{1}{1 + \sum_{l=2}^{C} \exp(\eta_{ijl})} \end{aligned}$$

The likelihood of the model is then

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}) = \prod_{j=1}^{M} \prod_{i=1}^{n_j} \prod_{j=2}^{C} \left(\frac{\exp(\eta_{ijk})}{1 + \sum\limits_{j=2}^{C} \exp(\eta_{ijk})} \right)^{y_{ijk}} \tag{2.7}$$

To complete the GLMM description, we assume $\boldsymbol{u}_j$ are independent multivariate normal with covariance matrix $\boldsymbol{\Sigma}$. In Chapter 3 we deal with a simpler version of the multinomial random effects model that has only a fixed and a random intercept since we do not include any covariates in the model.

Classical inference for this model is made through the *integrated likelihood*, where random effects are integrated out

$$\begin{aligned} f(y|\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \int f(y, \boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) d\boldsymbol{u} \\ &= \int f(y|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}) f(\boldsymbol{u}|\boldsymbol{\Sigma}) d\boldsymbol{u} \end{aligned}$$

Although many applications of the Multinomial logit model with random effects appear in the classical literature, most of them implement this model by approximating the likelihood with the equivalent likelihoods of Poisson model. The connection between Multinomial and Poisson random variables is based on the fact that Multinomial distribution can be derived from a set of independent Poisson random variables conditionally on their total being fixed McCullach and Nelder (1989). Thus, the model is transformed to a Poisson log-linear model which is easier to work with. Also, a variety of other combinations of methods are used, like penalized quasi-likelihood (PQL) method that was introduced by Breslow and Clayton (1993) for the estimation of the fixed and random effects and maximum likelihood (ML) or restricted maximum likelihood (REML) for the estimation of the variance of random effects Mollina et al. (2007). Hedeker (2003) uses a full maximum marginal likelihood together with multi-dimensional quadrature to numerically integrate over the random effects and an iterative Fisher scoring algorithm to solve the likelihood equations. Hartzel et al. (2001) also present a general approach for logit random effects modelling. Their maximum likelihood estimation uses adaptive Gauss-Hermite quadrature within a quasi-Newton maximization algorithm. When this is computationally infeasible, they apply a Monte Carlo EM algorithm and they also compare the pseudo-likelihood with a semi-parametric approach.

2.3.3 Bayesian Multinomial logit model

On the other hand, Bayesian approach seems more efficient when dealing with random effects and easier to implement. Assigning a prior distribution to the parameters, fixed and random, and using MCMC techniques we can simulate directly from their posterior distributions. However, examples in the literature on the Bayesian Multinomial model with random effects are limited. Kazembe and Namangale (2007) model child co-morbidity of fever, diarrhoea and pneumonia in Malawi with a Multinomial logit model with random effects. The data are clustered within two geographical levels, subdistricts and districts. The response variable Y_{ijk} is the *sickness status* and π_{ijk} the probability of multiple morbidity of the above diseases, with

$j = 1, ..., n_i$ defining the j child in area i, $i = 1, ..., I$ and k the various combinations of co-morbidity. Assuming

$$Y_{ijk} \sim \text{Multinomial}(\pi_{ijk}, 1)$$

and adding some covariates x_{ij}, the probability of co-morbidity is modelled

$$\pi_{ijk} = \frac{\exp(\eta_{ijk})}{1 + \sum_{l=1}^{C} \exp(\eta_{ijl})}$$

with

$$\eta_{ijk} = x_{ij}^T \beta_k + s_{ik}$$

The random effects s_{ik} correspond to spatial effects and are modelled using conditional autoregressive (CAR) models, where $i = 1, ..., I$ the areas and $k = 1, ..., C$ the multinomial categories. They are district or subdistrict specific factors and are separated into spatially structured variation and unstructured multinomial heterogeneity, $s_{ik} = \theta_{ik} + \phi_{ik}$. Moreover, distinguishing between subdistrict and district levels produces

$$\eta_{hijk} = x_{hij}^T \beta_k + s_{hik} + d_{hk}$$

where i refers to subdistrict and h to district and both terms can also be split in spatially structured variation and unstructured heterogeneity. For the spatially structured random effects a CAR prior distribution was assigned,

$$\theta_i | \{\theta_l; l \sim i\} \sim N\left(\frac{1}{m_i} \sum_{l \sim i} \theta_l, \frac{\sigma_\theta^2}{m_i}\right)$$

that assumes the mean of each area θ_i conditional on the neighbouring areas, has a normal distribution with mean equal to the average of neighbouring areas and variance inversely proportional to the number of neighbours m_i and where $l \sim i$ denotes areas l and i are neighbouring. Then, σ_θ^2 is further assigned a non-informative Inverse Gamma prior distribution with hyperparameters $a = b = 0.001$. The unstructured heterogeneity is given an exchangeable normal prior $\phi_i \sim N(0, \sigma_\phi^2)$ and σ_ϕ^2 an Inverse Gamma hyperprior. Finally, the fixed regression coefficients have diffuse priors $p(\beta_k) \propto \textit{constant}$.

Kazembe and Namangale (2007) consider several models with various combinations of random effects and compare them using the *deviance information criterion* (DIC) which is an information criterior for model comparison proposed by Spiegelhalter et al. (2002). Also, a sensitivity analysis is performed in order to check the choice of hyperprior distributions. We notice that univariate prior distributions are used for the random effects across different categories. In this thesis, we adopt multivariate normal distributions for the group specific random effects $\boldsymbol{u}_j^T = (u_{j1}, ..., u_{jC})$

$$\boldsymbol{u}_j \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

for group j. This means that $\boldsymbol{u}_j$ are category dependent and helps accounting for correlation between categories.

2.4 Inference for contingency tables

In this Section we describe briefly existing methods for analysing contingency tables from surveys and the way the Multinomial model can also be used in this case. Let Y_1 and Y_2 denote two categorical variables with R and C categories respectively. A table with R rows for categories of Y_1 and C columns for categories of Y_2 gives the frequency counts of outcomes for a sample and is called *contingency table*. Let p_{ij} denote the probability that (Y_1, Y_2) occur in cell of row i and column j that defines the joint distribution of Y_1 and Y_2. There are various models describing cell counts in contingency tables, like Poisson sampling model, Multinomial sampling and product Multinomial sampling model (Agresti, 2002). Usually, testing if the two variables are independent or not is one of the important questions. Classical inference for contingency tables consists of chi-square tests for independence, like *Pearson's chi-square test* and *likelihood chi-square test* which are asymptotically equivalent. Assume we have a table with M cells with O_i the observed count in cell i and E_{ij} the expected count in the same cell, then the Pearson's chi-square test is

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

and the Likelihood chi-square test is

$$G^2 = 2 \sum_i O_i \ln \left(\frac{O_i}{E_i} \right)$$

Both approximately follow a chi-square distribution with 1 degree of freedom under the hypothesis of independence. This is an asymptotic approximation the equivalency of which is questioned for small samples.

The table of probabilities from two cross-classifying variables can be displayed as

Table 2.1: Contingency table example

		Y_2				
		1	2	$\cdots$	C	
Y_1	1	p_{11}	p_{12}	$\cdots$	p_{1C}	p_{1+}
	2	p_{21}	p_{22}	$\cdots$	p_{2C}	p_{2+}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	R	p_{R1}	p_{R2}	$\cdots$	p_{RC}	p_{R+}
		p_{+1}	p_{+2}	$\cdots$	p_{+C}	1

where $p_{i+} = \sum_{j=1}^{C} p_{ij}$, $p_{+j} = \sum_{i=1}^{R} p_{ij}$. The observed count in cell (i, j) is $y_i j$.

Both chi-square statistic tests are highly influenced by the survey design. Especially when cluster sampling is performed, the within clusters correlation might have an effect on the p-value of these tests. Also, their distribution is not anymore a x_1^2 distribution since the sampling is not anymore multinomial. More examples are given in Lohr (1999) about the effect of ignoring the sampling design in inference for contingency tables. One solution is to take into account any weights given in the sample (Lohr, 1999; Clogg and Eliason, 1987). Sampling weights can be used to estimate cell counts or cell proportions as

$$\hat{p}_{ij} = \frac{\sum_{l \in S} w_l y_{ijl}}{\sum_{l \in S} w_l}$$

where

$$y_{ijl} = \begin{cases} 1 & \text{if unit } l \text{ falls in cell } (ij) \\ 0 & \text{otherwise} \end{cases}$$

and w_l the weight for unit l. A new table can be created replacing p_{ij} in Table 2.1 with the estimated $\hat{p}_{ij}$. Significant differences when comparing the odds ratios $p_{ij}p_{kl}/p_{il}p_{kj}$ with the estimated odds ratios $\hat{p}_{ij}\hat{p}_{kl}/\hat{p}_{il}\hat{p}_{kj}$ might mean that other factors affect the relation between these two variables. Lohr (1999) also comments on the influence of stratification and clustering in hypothesis tests and confidence intervals. The conclusion is that ignoring stratification results in conservative tests and large confidence intervals. One case where stratification presents no problems is when the strata are the categories of one of the cross-classified variables. On the other hand, clustering has the opposite effect and ignoring it in chi-square tests can be critical, since it tends to yield significant associations between the cross-classifying variables when they are not. Lohr (1999) then suggests some corrections to chi-square tests when clustering is present, such as *Wald tests*, *Bonferroni tests*, or correcting the test statistics X^2 and G^2 by matching their moments to chi-square distribution moments. Rao and Scott (1981) also examine the effect of stratification and clustering on chi-squared test statistic for goodness of fit and independence, provide their asymptotical distributions and corrections, suitable for two way contingency tables.

Finally, *log-linear* models have been suggested to analyse relationships between two or more classification variables. Let μ_{ij} denote the expected frequency in cell (ij) and n the number of units. If the two cross-classified variables are independent, then

$$\mu_{ij} = np_{ij} = np_{i+}p_{+j}$$

and taking logarithms produces

$$\log(\mu_{ij}) = \log(n) + \log(p_{i+}) + \log(p_{+j})$$

which can then be written

$$\log(\mu_{ij}) = \beta + \beta_i^r + \beta_j^c$$

where β_i^r refers to the row effect and β_j^c to the column effect. Also, for identifiability reasons we typically impose constraints such as $\beta_1^r = \beta_1^c = 0$. This model is known as *log-linear model* and a more complicated version is the saturated model including interactions:

$$\log(\mu_{ij}) = \beta + \beta_i^r + \beta_j^c + \beta_{ij}^{rc}$$

The latter model implies dependence between the two variables. Again for identifiability it is common to constrain $\beta_{i1}^{rc} = \beta_{1j}^{rc} = 0$. In this type of models, β_i^r and β_j^c can be thought of as coefficients of dummy variables and β_{ij}^{rc} as the coefficient of the product of dummy variables for β_i^r and β_j^c. The number of parameters in the model is $1 + (R-1) + (C-1) + (R-1)(C-1) = RC$. Classical tests for independence check if the β_{ij}^{rc} parameters equal zero.

In order to analyse contingency tables with Multinomial models, we need to condition on the sum of the cell counts n. Then, the Poisson log-linear model becomes Multinomial for the cell probabilities. The saturated model is

$$p_{ij} = \frac{\exp(\beta + \beta_i^r + \beta_j^c + \beta_{ij}^{rc})}{\sum_l \sum_m \exp(\beta + \beta_l^r + \beta_m^c + \beta_{lm}^{rc})}$$

which assumes $p_{ij} \geq 0$ and $\sum_i \sum_j p_{ij} = 1$ and β parameter actually cancels here. It corresponds the total sample size which is random in Poisson model but fixed here. The Bayesian version of this Multinomial model for contingency tables is examined analytically in Chapter 5 under survey design influence.

2.5 Conclusions

Through this Chapter we described first how design based inference for surveys is made for different sampling designs and its weaknesses. Next, we discussed Bayesian inference for surveys for continuous variables of interest. It is evident that the models have to incorporate information on the sampling design in order to be well specified. Then, Bayesian models can provide estimates that are design consistent. Nevertheless, we see from the above review that Bayesian modelling for surveys has been restricted to continuous response variables. Also, the design variables are almost always assumed known for all the population units. For example, the population cluster sizes are assumed to be known for sampled and non-sampled clusters with only exemption the work of Little and Zheng (2007). This is particularly important when predicting for the non-sampled cases.

To summarise the above discussion, we make the following comments. Regression models assume the covariates X known for all population units which allows for prediction for non-sampled cases. Little and Zheng (2007) introduced the only example where the design variable (size variable) is not recorded for the non-sampled cases in PPS sampling and suggested a Bayesian Bootstrap model for the size variable. Moreover, the use of sampling weights in modelling is not explicit but there are suggested models (Little, 2004) where posterior estimates correspond to weighted averages.

We introduced the Multinomial model which is used in this thesis to model polytomous responses and contingency tables coming from surveys. Classical methods for Multinomial data and contingency tables may not be efficient in complex survey designs. Problems appear when the sample size is small and there are zeros in the table and when the sampling design effect is not considered.

These conclusions reveal that there is a need for Bayesian inference for categorical response outcomes, in particular when information about survey design variables is limited or non-available. In these cases regression models are inadequate because prediction is not possible. We also have to note that our goal is to use sampling weights when possible but not in modelling, where their use is controversial.

Chapter 3

Models for the size variable

3.1 Introduction

We mentioned in the previous Chapter that design variables are known to the survey statisticians but not to the data analysts. Therefore, as data analysts we only have the sampled values of the design variables. Since, the design variables are typically variables used for stratified or cluster sampling, their values in the sample are the stratum or cluster indicators. We introduce the term *size variable* to express the stratum or cluster sizes in the population and in the sample. The problem of not knowing the population values of the design variables equals to the problem of not knowing the group (strata or clusters) sizes in the population.

To summarise, we address the problem of the unknown sizes of the groups in the population during survey data analysis. As in *finite* population problems the data analysts need to know the population size, they also need to know the population group sizes. For example, in Section 2.2.1 we described a model that Little (2004) suggested for stratified sampling and that gives the following posterior expectation of the population mean $\bar{Y}$

$$E(\bar{Y}|Y_S, I, \sigma_j^2) = \bar{y}_{st} = \sum_{j=1}^{M} P_j \bar{y}_j$$

where $P_j = N_j/N$ are assumed known. In our analysis, we drop this assumption and try to find methods to calculate or model the population group sizes N_j. These quantities are then used in Chapter 4 and 5 to make inference for the population counts of a variable of interest Y with many categories.

At this point we need to clarify the assumptions we make in our analysis. We assume the population size and the number of groups in the population are known. This has to be taken into account when modelling and predicting the group sizes. We examine three sampling designs, stratified random sampling, cluster sampling with SRS and cluster sampling with PPS. In our dataset, possible stratification variables are the variables that present some kind of categories, like *sex*, *age*, *social grade*, etc. Potential cluster variables are the geographical variables such as the *area*, *district* or *health region.*

During the following analysis, we use two different notations for the sizes. This is important when discussing different sampling schemes, such as stratified and cluster sampling. First, we introduce the population size N and sample size n, quantities that are known and also n_j is the sample size in group j. In a stratified sample, units from all strata are included in the sample and so there are not any unsampled strata. Then, we use the notation N_j for the size of the group j in the population and N_js are calculated deterministically. In a cluster sample, some of the clusters are included in the sample, this means there are non-sampled clusters. Then, we define the *size variable* Z and z_j denotes the size for cluster j. In the latter case, Z is a variable and is modelled appropriately. Also, we use the subscript S when we want to define sampled units or groups and $\bar{S}$ for the non-sampled.

In the following analysis, many different samples were obtained for every sampling design but the results presented are all coming from a particular sample. This is decided for the sake of producing numerical results suitable for comparison between methods and designs. However, inference is similar when another sample is chosen and the choice of sample does not affect the inference procedure, only the numbers deriving from it.

3.2 Stratified random sampling

As described in Section 2.2 during stratified sampling, units in stratum j are weighted by the inverse of their selection probability, usually scaled to sum up to the total sample size

$$w_j = (N_j/n_j)(n/N) \tag{3.1}$$

for $j = 1, ..., M$, where M the total number of strata in the population. During stratified random sampling, all strata in the population are sampled and hence there is no need to account for any non-sampled strata. Using the dataset presented in Table 1.1, *age* or *sex* can be stratification variables. Often, survey statisticians create new stratification variables by cross-classifying units according to both *age* and *sex*. The new strata coming from these two variables have $2 \times 6 = 12$ categories with labels given from the combinations of the original strata labels. Then, a random sample is taken from each new stratum with some method of allocation and sampling weights are constructed. Choosing amongst methods to obtain stratum sample sizes is not our goal and we simply use proportional allocation (Lohr, 1999). Proportional allocation means that the size of the sample for each stratum is taken in proportion to the stratum size in the population.

Next, the data analysts obtain the dataset and can use Equation (3.1) to calculate directly the strata sizes by solving it with respect to N_j

$$N_j = w_j n_j N/n \tag{3.2}$$

In this case, the data analyst does not have to model the sizes, since all strata are included in the sample and population sizes can be calculated easily. Strata size N_j for stratum j can then be used in predicting a response variable Y within this stratum, as presented in Chapter 4.

3.3 One stage cluster sampling with SRS

During this sampling procedure, the survey statistician decides to sample a certain number of clusters as the primary sampling units (PSU) and then sample all the units within selected clusters. Cluster sampling with SRS is usually performed when the sizes do not differ significantly in the population. As data analysts, we do not have information about the sizes of the non-sampled clusters and need to model the size variable Z. However, we have information about the type of sampling design and which the cluster variable is.

Let z_j denote the observed cluster sizes, for $j = 1, ..., m$ where m denotes the number of sampled clusters from a total number M. If we know that cluster sampling with SRS was performed, we can safely conclude that cluster sizes do not vary significantly. Moreover, there are no clusters with size 0 in the population and so we have to adjust our model of choice to be *zero-truncated.*

For the dataset described in Section 1.8, we observe that potential variables to be used for cluster sampling are *area* and *district.* Both represent geographical variables and Table 3.1 gives population summaries for both of them. We see that the population mean and variance are almost equal for *area* that suggests a Poisson model is appropriate. On the other hand, the variance of *district* is much larger than its mean and a Negative Binomial model is assumed for this variable.

Table 3.1: Summary statistics for *area*, *district*

Statistic	*area*	*district*
Mean	12.44	40.51
Variance	12.23	1407.49
Min	3.00	9.00
Max	20.00	200.00

3.3.1 Zero-truncated Poisson model

Area is the cluster variable assumed to be used for SRS cluster sampling. A zero-truncated Poisson model is suggested for the *area* sizes or *size variable* Z. First, we describe the zero-truncated Poisson model. The zero-truncated Poisson distribution has probability mass function, corresponding to the untruncated distribution, defined by

$$f(z|\lambda) = \frac{f(z|\lambda)}{\Pr(z>0)} \quad \text{where} \quad z = 1, 2, ...$$

where

$$\Pr(z>0) = 1 - \Pr(z=0) = 1 - e^{-\lambda}$$

which leads to

$$f(z|\lambda) = \frac{e^{-\lambda}\lambda^z}{z!}\frac{1}{1-e^{-\lambda}} = \frac{\lambda^z}{z!(e^{\lambda}-1)}$$

Now, let z_j be the size for area j for $j = 1, ..., m$. Then,

$$f(z_j;\lambda) = \frac{\lambda^{z_j}}{z_j!(e^{\lambda}-1)}$$

with likelihood

$$f(\boldsymbol{z}|\lambda) = \prod_{j=1}^{m} \frac{\lambda^{z_j}}{z_j!(e^{\lambda}-1)} \tag{3.3}$$

We assign a $Gamma(\alpha,\beta)$ prior distribution for λ and its posterior distribution becomes

$$\begin{aligned} f(\lambda|\,\boldsymbol{z}) &\propto \frac{\lambda^{\sum z_j}\lambda^{\alpha-1}e^{-\lambda\beta}}{(e^{\lambda}-1)^m} \\ &= \frac{\lambda^{\sum z_j+\alpha-1}}{e^{\lambda\beta}(e^{\lambda}-1)^m} \end{aligned} \tag{3.4}$$

A Metropolis-Hastings (M-H) algorithm is used to simulate from the posterior distribution of λ. Then, we want to get simulations from the posterior predictive distribution $f(\tilde{\boldsymbol{z}}|\boldsymbol{z})$ of new sizes $\tilde{\boldsymbol{z}}$ and take into account the fixed known population size which makes the sum of the non-sampled sizes to be fixed and equal to $N-n$. Hence, after obtaining draws from $f(\lambda|\boldsymbol{z})$, we use again a M-H algorithm with the following steps:

1. Start from a random vector of $\tilde{\boldsymbol{z}} = (\tilde{z}_1, ..., \tilde{z}_{M-m})$ with $\sum\limits_{j=1}^{M-m} z_j = N-n$
2. Choose t (even number) elements from this vector
3. Increase $t/2$ of the chosen elements by step=s and decrease the remaining $t/2$ elements by s
4. Set the new vector as the proposed vector $\tilde{\boldsymbol{z}}^{can}$
5. Calculate rate
$$\alpha = \frac{\prod\limits_{j=1}^{M-m} f(\tilde{z}_j^{can}|\lambda)}{\prod\limits_{j=1}^{M-m} f(\tilde{z}_j|\lambda)}$$
where $f(\boldsymbol{z}|\lambda)$ is Equation 3.3.
6. Draw $u \sim \text{Unif}(0,1)$
7. If $u < \alpha$ then $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{z}}^{can}$ else go to (2).

The numbers s, t are chosen to produce an acceptance rate between 20% and 30%.

Example

As mentioned above, *area* is used for one stage cluster sampling with SRS. A sample of $m = 30$ areas from the $M = 140$ in the population is given to us for analysis. The sampled dataset includes $n = 357$ observations of the $N = 1742$ population units. Size variable z has sample mean 11.9 and sample variance 13.33. Different prior distributions were tried here by modifying the hyperparameters for Gamma distribution. Our wish is to represent our weak prior knowledge about the parameter and see how different prior distributions affect inference about it (Gelman, 2006; Gelman et al., 2008). In particular, Gamma(0.001,0.001), Gamma(1/2,0.001) and Gamma(1,0.001) were tested, the last two corresponding to Jeffrey's prior and positive Uniform distribution. All different produced similar results, shown in Table 3.2 that suggests the data is enough to make the choice of prior distribution negligible during inference.

Table 3.2: Posterior inference for λ under different prior distributions

Prior	Mean	s.e.
Gamma(0.001,0.001)	11.93571	0.63076
Gamma(1/2,0.001)	11.91546	0.63038
Gamma(1,0.001)	11.89298	0.62960

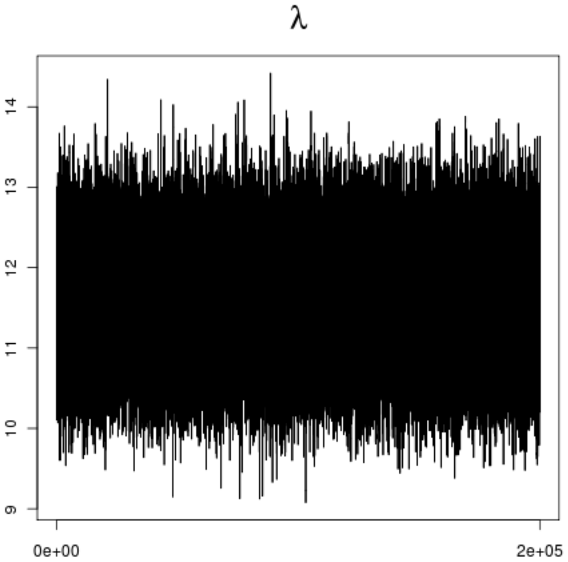

Figure 3.1: Trace plot for λ in the zero-truncated Poisson model

As mentioned before, we want to have an acceptance rate between 20% and

30% and after having run different simulations and tuned the proposal distribution, s and t, we choose a Normal proposal with variance equal to 1, $s = 10$ and $t = 1$. The positive Uniform is selected as the prior distribution to make inference. A number of 100000 draws gives satisfying updating of all the elements of the $\tilde{\boldsymbol{z}}$ vector. Convergence diagnosis is made through the trace plots observation and calculation of Gelman and Rubin's convergence diagnostic which is equal to 1 (Gelman and Rubin, 1992; Brooks and Gelman, 1997). Figure 3.1 shows the trace plot for λ and Table 3.3 summarises λ and $\tilde{\boldsymbol{z}}$ posterior distribution.

Table 3.3: Posterior inference for λ and non-sampled $\tilde{\boldsymbol{z}}$ in the zero-truncated Poisson model

	Mean	s.e.	95%C.I.
λ	11.89	0.629	(10.37, 12.81)
$\tilde{\boldsymbol{z}}$	12.69	3.551	(6, 21)

The histogram of the posterior predictive distribution for $\tilde{\boldsymbol{z}}$, together with the true mean (black vertical line) of the non-sampled cluster sizes is given in Figure 3.2.

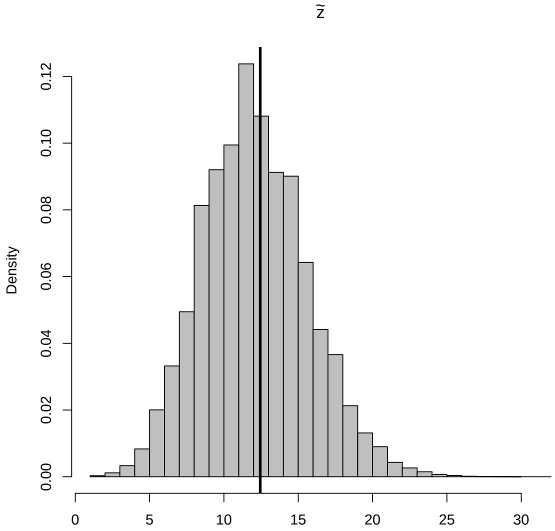

Figure 3.2: Posterior predictive density for non-sampled $\tilde{z}$ and true mean of the non-sampled cluster sizes

3.4 Two-stage cluster sampling with SRS

During this sampling procedure, a number of units within selected clusters is sampled in a second stage sampling. This means that the sampled cluster size n_j for cluster j is not equal to the population cluster size N_j. Therefore, we have first to calculate the population cluster sizes for sampled j in order to model the sizes. We use the sampling weights to get the approximate size of sampled clusters and then modelling becomes similar to the previous section. Weights reflect the probabilities of selection in each stage and during the first stage clusters are selected with probability m/M and the units within cluster j are sampled with probability n_j/N_j. Thus, the weight corresponding to units within cluster j is

$$w_j = w_1 w_2 = \frac{M}{m}\frac{N_j}{n_j}$$

and

$$N_j = w_j n_j \frac{m}{M}$$

where M, m, w_j and n_j are known. Having obtained the population sizes of the sampled clusters, we model the size variable with a zero-truncated Poisson model as described in previous sections.

3.5 One stage cluster sampling with PPS

Cluster sampling with probability proportional to size is common when the cluster sizes vary significantly and we want to include larger clusters in the sample. Thus, cluster j is selected with probability proportional to its size z_j. The sampling design becomes informative and we need to include this in the modelling process by specifying a model both for the survey data and the inclusion indicators. The following analysis is about Poisson sampling which is a kind of PPS sampling.

Poisson Sampling

Poisson sampling is a sampling process where each element of the population is subjected to an independent Bernoulli trial which determines whether the element becomes part of the sample. The term *element* here corresponds to clusters and each has a different probability of being included in the sample. We consider a Negative Binomial model for the cluster sizes in the population. This model is more realistic when it comes to PPS sampling where the sizes are considered to differ significantly. Since the variance usually exceeds the mean, the Negative Binomial model with one more parameter than the Poisson can be used to adjust the variance independently of the mean. Let z_j^* denote the population values for $j = 1, ..., M$ and z_j denote the sampled values.

3.5.1 Zero-Truncated Negative Binomial model.

We use the variable *district* which is ideal for PPS sampling. It is again a geographical variable but the sizes of the districts vary significantly compared to areas. Poisson sampling is applied as described above and the a number of $m = 9$ districts are sampled from the $M = 43$ in the population. Again, one sample is chosen to present the results and inference for every sample taken is similar to the one descibed here. We note that the largest size district ($N_j = 200$) is always sampled and very few small size districts are included in the sample. Summary statistics for the population district sizes, the sampled and non-sampled are given in Table 3.4, where one can see the differences between them. The sampled cluster sizes have a mean of 80.55 and a variance of 3443.52 that make the Poisson model non-suitable as a model for the population. One can also notice that the mean and variance of non-observed sizes is 29.91 and 403.47 respectively that supports the assumption that distributions of sampled and non-sampled sizes differ. Histograms of both distributions in Figures 3.3 and 3.4 confirm this as well. We observe that histogram of sampled sizes shows larger mean and variance comparing to unsampled.

Table 3.4: Summary statistics for *district* sizes after PPS sampling

Statistic	Population	Sampled	Non-sampled
Mean	40.51	80.55	29.91
Variance	1407.49	3443.52	403.47
Min	9.00	26.00	9.00
Max	200.00	200.00	86.00

The population size variable has a zero-truncated Negative Binomial distribution

$$f(z; p, \xi) = \frac{f(z)}{P(z > 0)}$$

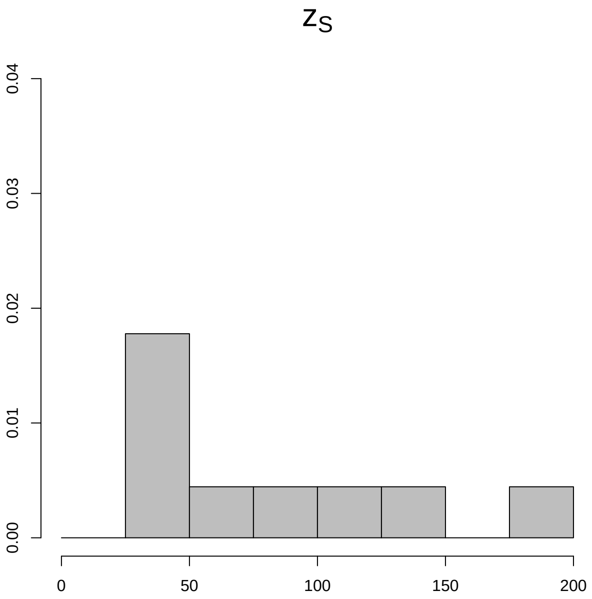

Figure 3.3: Histogram of sampled sizes in one stage PPS sampling

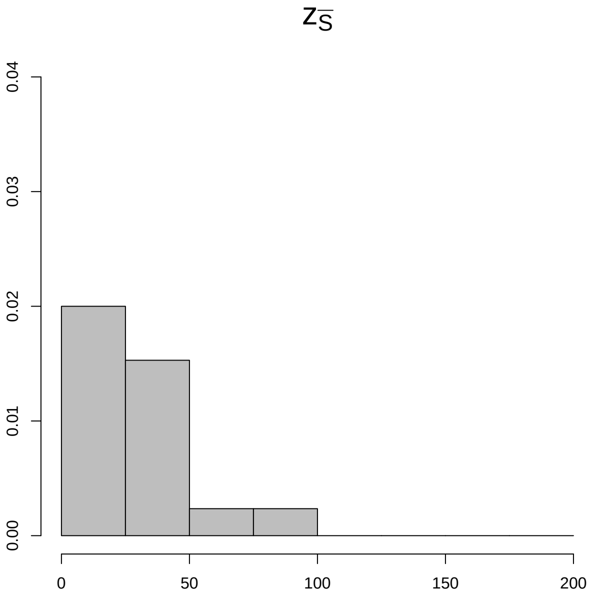

Figure 3.4: Histogram of non-sampled sizes in one stage PPS sampling

where $f(z)$ denotes the non-truncated density. Since

$$\Pr(z>0)=1-\Pr(z=0)=1-(1-p)^{\xi}$$

the probability mass function is

$$f(z;p,\xi)=\binom{z+\xi-1}{\xi-1}\frac{p^{z}(1-p)^{\xi}}{1-(1-p)^{\xi}} \tag{3.5}$$

Then sample

$$I_j \sim Bernoulli(\phi z_j^{*})$$

and observe

$$z_{j\in S} \text{ for } I_j=1$$

Sampling with PPS ensures that larger clusters are more likely to be included in the sample than smaller clusters. The probabilities used in the Bernoulli trials $\pi_j=\phi z_j$ have to satisfy $0<\pi_j\leq 1$ and $\sum_{j=1}^{M}\pi_j=m$, see Hanif and Brewer (1980) and Lohr (1999). These probabilities are calculated using the relation $\pi_j=mz_j/N$ where $N=\sum_{j=1}^{M}z_j$ that yields that $z_{max}\leq N/m$. This generally holds for every type of PPS sampling. In our case, we use $\phi=1/z_{max}\approx m/N$. Otherwise, z_{max} would get inclusion probability more than one, that is by definition impossible. Therefore, ϕ is set equal to the inverse of the maximum of the sizes, $\phi=1/200=0.005$. It is known during the sampling procedure but not available to the data analyst and so we need to estimate it. The number of sampled clusters remains random during Poisson sampling, and so after sampling we have $m=9$ sampled clusters from $M=43$. Also, the number of selected units is $n=725$ that leaves a total number of $N-n=1017$ non-selected units. The likelihood is

$$\begin{aligned}\mathcal{L}(p,\xi,\phi|z)&=\int f(z|p,\xi,\phi)\mathrm{d}z_{\bar{S}}=\int f(z_S|p,\xi,\phi)f(z_{\bar{S}}|p,\xi,\phi)dz_{\bar{S}}\\
&=\prod_{j\in S}\left(\phi z_j\frac{\Gamma(z_j+\xi)}{\Gamma(\xi)z_j!}\frac{p^{z_j}(1-p)^{\xi}}{1-(1-p)^{\xi}}\right)\sum_{j\in\bar{S}}\prod_{j\in\bar{S}}\left((1-\phi z_j)\frac{\Gamma(z_j+\xi)}{\Gamma(\xi)z_j!}\frac{p^{z_j}(1-p)^{\xi}}{1-(1-p)^{\xi}}\right)\\
&=\prod_{j\in S}\left(\phi z_j\frac{\Gamma(z_j+\xi)}{\Gamma(\xi)z_j!}\frac{p^{z_j}(1-p)^{\xi}}{1-(1-p)^{\xi}}\right)\prod_{j\in\bar{S}}\sum_{j\in\bar{S}}\left((1-\phi z_j)\frac{\Gamma(z_j+\xi)}{\Gamma(\xi)z_j!}\frac{p^{z_j}(1-p)^{\xi}}{1-(1-p)^{\xi}}\right)\end{aligned}$$

$$= \prod_{j\in S}\left(\phi z_j \frac{\Gamma(z_j+\xi)}{\Gamma(\xi)z_j!}\frac{p^{z_j}(1-p)^{\xi}}{1-(1-p)^{\xi}}\right)\prod_{j\in\bar{S}}\sum_{j\in\bar{S}}\left(\frac{\Gamma(z_j+\xi)}{\Gamma(\xi)z_j!}\frac{p^{z_j}(1-p)^{\xi}}{1-(1-p)^{\xi}}\right)$$

$$-\sum_{j\in\bar{S}}\left(\phi z_j \frac{\Gamma(z_j+\xi)}{\Gamma(\xi)z_j!}\frac{p^{z_j}(1-p)^{\xi}}{1-(1-p)^{\xi}}\right)$$

$$= \prod_{j\in S}\left(\frac{\Gamma(z_j+\xi)}{\Gamma(\xi)\Gamma(z_j)}\right)\phi^m \frac{p^{m\bar{z}}(1-p)^{m\xi}}{(1-(1-p)^{\xi})^m}\prod_{j\in\bar{S}}\left(1-\phi\frac{\xi p}{(1-p)(1-(1-p)^{\xi})}\right)$$

$$= \prod_{j\in S}\left(\frac{\Gamma(z_j+\xi)}{\Gamma(\xi)\Gamma(z_j)}\right)\phi^m \frac{p^{m\bar{z}}(1-p)^{m\xi}}{(1-(1-p)^{\xi})^m}\left(1-\phi\frac{\xi p}{(1-p)(1-(1-p)^{\xi})}\right)^{M-m}$$

where we sum out the non-sampled values in the second line and $0 < p < 1$, $\xi > 0$, $0 < \phi < 1$. Suitable prior distributions for the parameters are

$$\phi \sim Beta(c_1, d_1)$$

$$p \sim Beta(c_2, d_2)$$

and

$$\xi \sim Exp(t)$$

the hyperparameters of which are discussed in the following example. The joint posterior distribution is

$$f(p,\phi,\xi|\boldsymbol{z}) \propto \prod_{S}\left(\frac{\Gamma(z_j+\xi)}{\Gamma(\xi)\Gamma(z_j)}\right)\phi^m \frac{p^{m\bar{z}}(1-p)^{m\xi}}{(1-(1-p)^{\xi})^m}\left(1-\phi\frac{\xi p}{(1-p)(1-(1-p)^{\xi})}\right)^{M-m}$$
$$\phi^{c_1-1}(1-\phi)^{d_1-1}p^{c_2-1}(1-p)^{d_2-1}e^{-t\xi}$$

Let

$$A = \left(1-\phi\frac{\xi p}{(1-p)(1-(1-p)^{\xi})}\right)^{M-m}$$

Then the full conditional distributions are

$$f(\phi|\boldsymbol{z},\xi,p) \propto A\phi^m\phi^{c_1-1}(1-\phi)^{d_1-1} = A\phi^{m+c_1-1}(1-\phi)^{d_1-1} \tag{3.6}$$

$$f(p|\boldsymbol{z},\xi,\phi) \propto A\frac{p^{m\bar{z}}(1-p)^{m\xi}}{(1-(1-p)^{\xi})^m}p^{c_2-1}(1-p)^{d_2-1}$$
$$= A\frac{p^{m\bar{z}+c_2-1}(1-p)^{m\xi+d_2-1}}{(1-(1-p)^{\xi})^m} \tag{3.7}$$

$$f(\xi|\boldsymbol{z},p,\phi) \propto A\frac{(1-p)^{m\xi}e^{-t\xi}}{(1-(1-p)^{\xi})^m}\prod_{j=1}^{m}\left(\frac{\Gamma(z_j+\xi)}{\Gamma(\xi)\Gamma(z_j)}\right) \tag{3.8}$$

Example

Again the choice of hyperparameters is made through testing different prior distributions and perfoming sensitivity analysis. For p and ξ posterior inference is robust to different prior distributions, therefore non-informative priors are selected. For ϕ we want to express weak prior information and include the information described in page 57. Assuming we do not know whether the largest cluster is sampled or not, we try different Beta and check how robust inference is. We test Beta(1,1) (equivalent to Uniform(0,1)) and other Beta that give high probability to numbers less than the observed $1/z_{max}$. This makes sense because otherwise the largest sampled cluster would have inclusion probability more than 1. Figure 3.5 shows different prior distributions for ϕ together with the posterior distributions. It is evident that posterior inference is robust under suitable prior distributions that are vague enough. The 95% credible interval (see Table 3.5) includes the true value under all prior distributions. Also, the posterior density is consistent with the assumption for ϕ that it must be under $1/z_{max} = 0.005$.

Table 3.5: Posterior inference for ϕ under different prior distributions

Prior	Mean	s.e.	95%C.I.
Beta(1,1)	0.0028	0.0016	(0.0010, 0.0075)
Beta(1,50)	0.0028	0.0015	(0.0010, 0.0070)
Beta(1,100)	0.0027	0.0014	(0.0010, 0.0065)
Beta(1,200)	0.0025	0.0012	(0.0010, 0.0060)
Beta(0.1,1)	0.0023	0.0014	(0.0009, 0.0053)
Beta(0.1,20)	0.0023	0.0012	(0.0009, 0.0056)
Beta(0.1,50)	0.0023	0.0011	(0.0009, 0.0053)
Beta(0.1,100)	0.0022	0.0011	(0.0009, 0.0052)

For p and ξ we choose ($c_2 = 1$, $d_2 = 1$) and $t = 0.001$ respectively to produce non-informative prior distributions and express weak prior knowledge. Again, we use Normal proposals to update for each M-H step and reject the negative

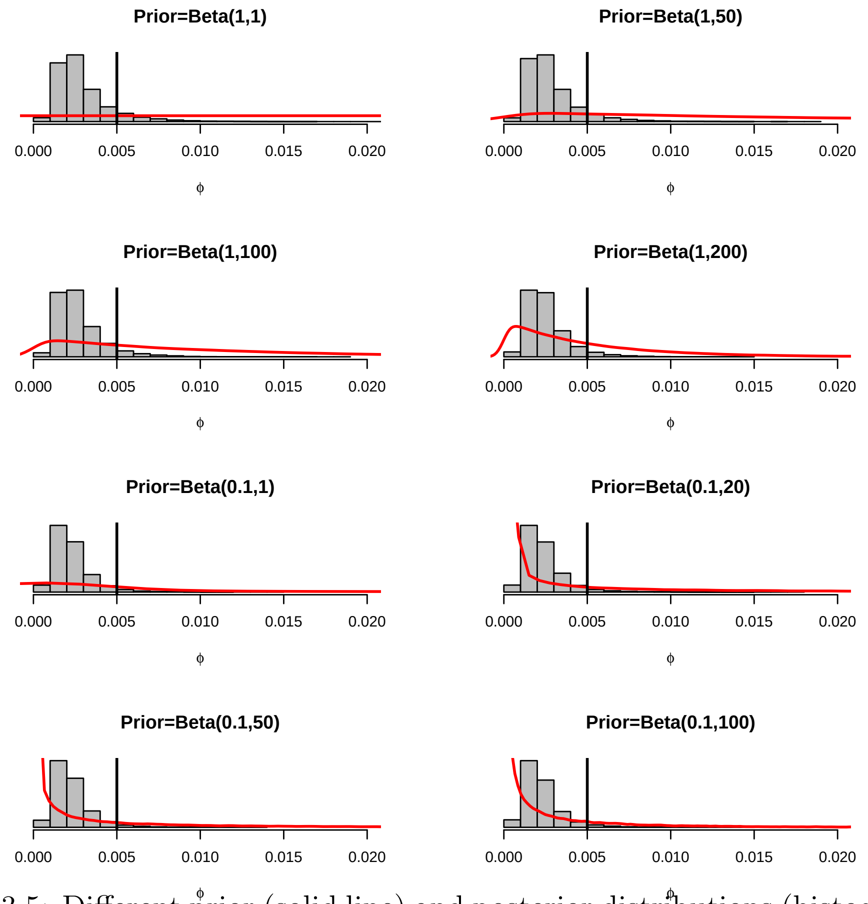

Figure 3.5: Different prior (solid line) and posterior distributions (histogram) plotted together with the true value for ϕ (black vertical line)

values. After running the simulations we summarise posterior inference in Table 3.6 and Figure 3.6 shows convergence of the three parameter chains.

Table 3.6: Summary of posterior inference for parameters in PPS sampling

	Mean	s.e.	95%C.I.
ϕ	0.0035	0.0028	(0.0021, 0.0132)
p	0.9639	0.0157	(0.9265, 0.9867)
ξ	1.7849	1.2335	(0.1382, 4.7559)

To check the validity of our model we can make predictive inference for new population sizes. To do this, we use the draws from the posterior distributions

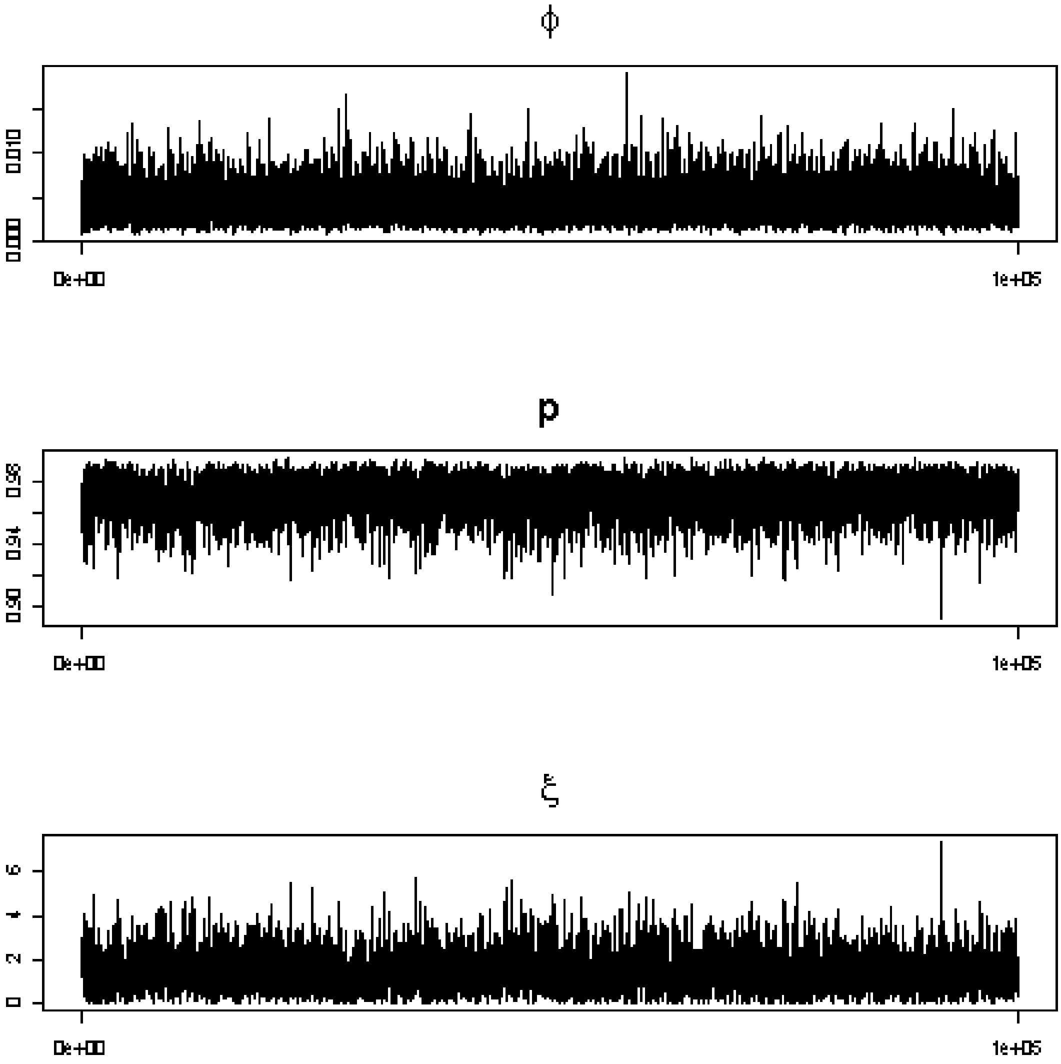

Figure 3.6: Trace plots for parameters ϕ, p, ξ in PPS sampling

of the parameters and then use Equation 3.5 to draw new population sizes. We implement the algorithm used in Section 3.3.1 and Figure 3.7 shows the histogram of the actual district sizes in the population and the posterior predictive distribution.

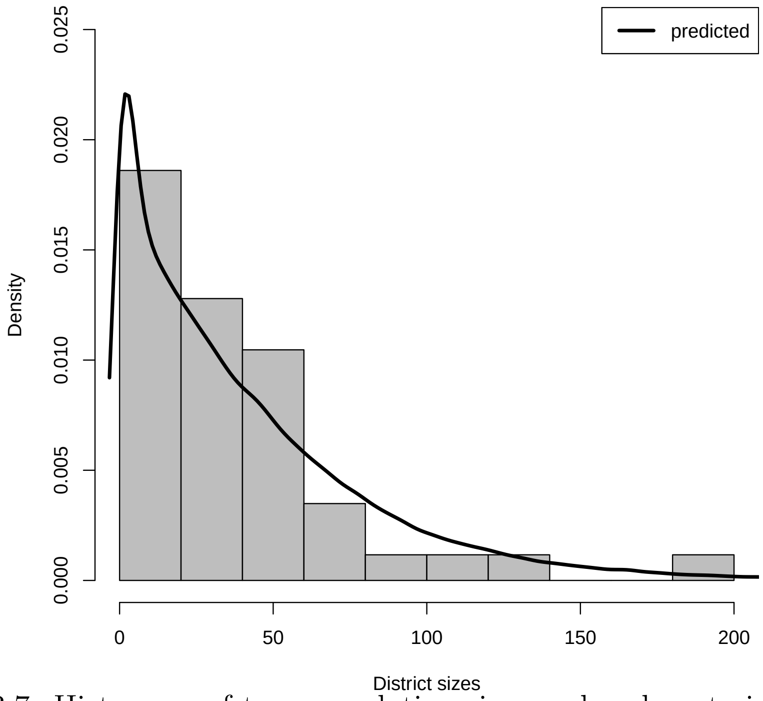

Figure 3.7: Histogram of true population sizes and and posterior predictive distribution

3.6 Comparing with the Bayesian Bootstrap model

We use the Bayesian Bootstrap model implemented by Little and Zheng (2007) and described in Section 2.2.2 to draw same number of samples from the posterior predictive distribution of the non-sampled clusters. We run the same number of simulations for the BB model and in Figures 3.8 and 3.9 we present graphical posterior predictive checks for the non-sampled cluster sizes using 7 different samples for illustrative purposes. We see there that the Negative Binomial model seems to predict better the non-sampled sizes than the BB model. Also, posterior predictive mean and variance for the Negative Binomial model are closer to the true values than the BB ones, as shown in Table 3.7. Moreover, the BB model produces $\tilde{\boldsymbol{z}}_{\bar{S}}$ that do not sum up to the total number of non-sampled units.

Table 3.7: Predictive inference for non-sampled sizes for the Neg.Bin. model, BB and true values

Estimate	Neg.Bin.	BB	True
Mean	29.91	57.08	29.91
Variance	517.02	1224.53	403.47

The advantage of the BB model is that there is no assumption about the type of sampling used to achieve PPS sampling while during our analysis we assumed Poisson sampling. We would like to test how strong this assumption is and if our model can be used in other types of PPS sampling or when the particular type of PPS sampling is unknown. Generally, sampling with unequal probabilities is quite complicated, especially when one wants to sample more than one primary sampling unit or to sample without replacement. Classical inference assumes knowing the inclusion probability for each PSU, which means finding the probability of each pair of PSU being in the sample and then the overall probability that the ith PSU would be in the sample (Lohr, 1999). This procedure becomes troublesome for large populations and sample sizes more than 2. Hanif and Brewer (1980) review and compare different methods of unequal probabilities sampling. General criteria are: the limitation in samples of size= 2, applicability, simplicity in selection and variance, efficiency of HT estimator, etc. However, since the conditions that the inclusion probabilities must satisfy are the same for all sampling schemes we can assume that $\phi \leq 1/z_{max}$ for any type of PPS sampling.

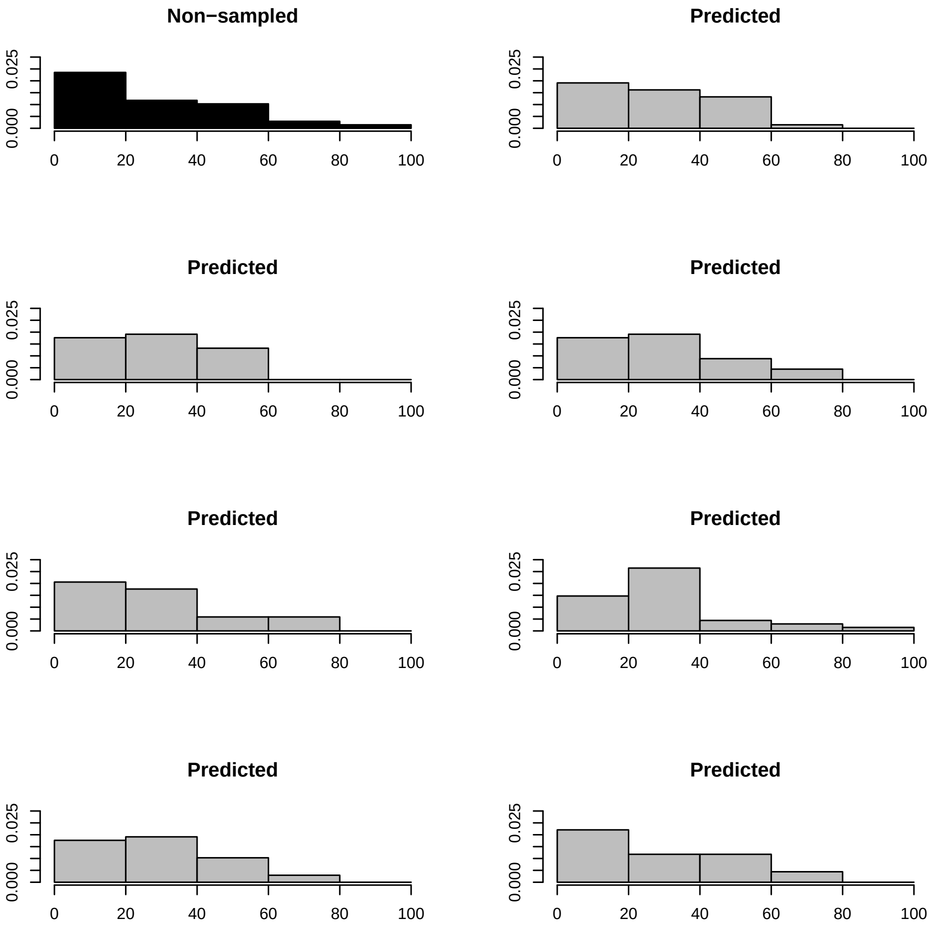

Figure 3.8: Posterior predictive histograms for Neg.Bin. Model for 7 new samples of non-sampled sizes-Poisson sampling

In order to test our model performance in other sampling designs than Poisson, we choose the *systematic sampling* with unequal probabilities and the sampling Brewer (1975) suggested. Our goal is to give evidence of our model robustness in different PPS samplings when comparing with the non-parametric BB. We use the R package Survey Sampling (Till and Matei, 2009) to draw these samples. Table 3.8 gives the mean and variance of the true non-observed sizes and predicted non-observed sizes for the 2 different

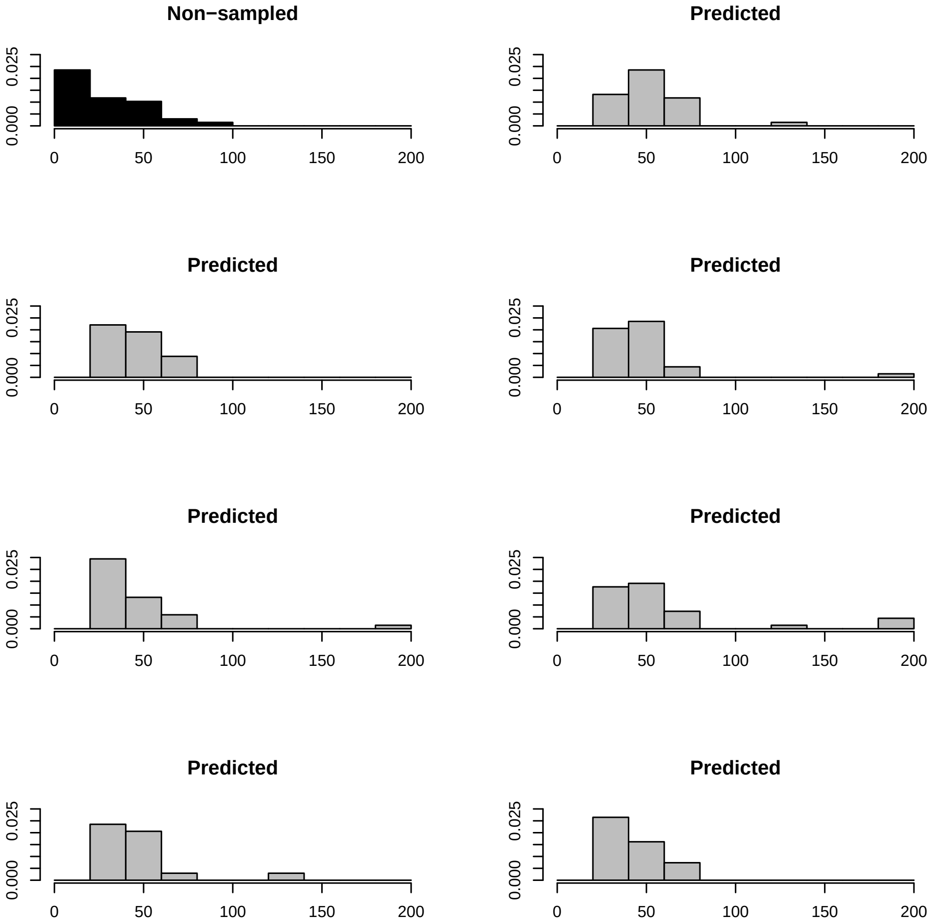

Figure 3.9: Posterior predictive histograms for BB Model for 7 new samples of non-sampled sizes-Poisson sampling

using the Negative Binomial (NB) model and the BB model. It is evident that the NB model performs better than the BB model when it comes to prediction of the non-sampled sizes independently of the type of PPS sampling. We note that the variance of the predicted $\tilde{z}_{\bar{S}}$ when applying the BB model is higher than the true one. This suggests that large clusters are frequently sampled when sampling from the posterior predictive distribution. The BB model fails to account for the fact that non-sampled clusters usually vary less than the sampled and have smaller variance.

Table 3.8: BB and Neg.Bin. model predictive inference for different PPS samples

		Systematic	Brewer
True	Mean	32.85	31.67
	Variance	584.67	633.28
Neg.Bin.	Mean	32.85	31.67
	Variance	541.19	543.84
BB	Mean	27.21	40.65
	Variance	1226.64	1874.96

3.7 Discussion

In this Chapter we suggested models for the size of the design variables when no information is available for the non-sampled part. This is usually the case when we use datasets for analysis and have no access to details about the survey. Therefore, the sizes of the various strata or clusters in the population used to construct the survey are unknown to us. However, data analysts are not completely disconnected with the survey statisticians and some information about the survey design is usually easy to obtain, such as the type of design, the design variables or the stages of sampling. This information is used and explored wherever possible in this thesis.

In Section 3.2, we examined how the sizes of different strata can be calculated when the sampling method is stratified sampling. As there are no unsampled strata and the sampling weights are given, calculating the stratum sizes is straightforward.

Section 3.3 and 3.4 are about cluster sampling with simple random sampling (SRS). Cluster sampling complicates analysis, since there are non-sampled clusters for which we have to predict the sizes. We suggested a zero-truncated Poisson model for data where is no evidence of overdispersion. We underline the fact that we took into account the finite number of population units when predicting new cluster sizes.

In Section 3.5, we discussed about probability proportional to size sampling (PPS) and in particular about Poisson sampling. In this type of sampling large clusters tend to be included in the sample more often than small clusters and so the distribution of sampled and non-sampled clusters vary significantly. Cluster sampling with PPS presents higher degree of difficulty since we need to account for the sampling process. We proposed a model for the population sizes that also accounts for PPS sampling. We implemented a Negative Binomial model that accounts for this type of sampling and predicts non-observed cluster sizes efficiently. Then, we also compared our model with the non-parametric Bayesian Bootstrap model and concluded in favour of the Negative Binomial model. Finally, we tested both models in two other types of PPS sampling to check how strong the assumption of Poisson sampling is. The conclusion is that the Negative Binomial model works better than the BB in any kind of PPS sampling.

The methodology suggested in this Chapter is useful when the size variable is the variable of interest in itself, but also when groups sizes are important as part of inference for another finite population quantity of interest. In the following analysis, we use inference derived in this Chapter in order to model polytomous variables and contingency tables.

Chapter 4

Modelling the polytomous response

4.1 Introduction

In this Chapter we are interested in modelling the main response variable and then use results from the previous chapter to make inference about a categorical response during different sampling designs. Inference about population quantities like totals is examined here which can be called *descriptive* inference. We briefly describe finite population inference for univariate categorical responses with the following example taken from Little and Raghunathan (2008). Suppose we have a binary response variable Y_i that

$$Y_i = \begin{cases} 1 & \text{if something is present in } i\text{th unit} \\ 0 & \text{otherwise} \end{cases}$$

and the quantity of interest is the the proportion $Q = \sum_{i=1}^{N} Y_i/N$ where N is the population size. A simple random sample of size n is taken and since $Y_i|\theta \sim Bernoulli(\theta)$ then $y^* = \sum_{i=1}^{n} Y_i$ is a sufficient statistic that has a

Binomial distribution with

$$f(y^*|\theta) = \binom{n}{y^*}\theta^{y^*}(1-\theta)^{n-y^*}$$

Hence,

$$Q = \sum_{i=1}^{N} Y_i/N = (y^* + \sum_{i=n+1}^{N} Y_i)/N$$

and assigning $p(\theta) = 1$ we get that

$$\theta|y^* \sim Beta(y^* + 1, n - y^* + 1)$$

and

$$\left(\sum_{i=n+1}^{N} Y_i|\theta, y^*\right) \sim Bin(N-n, \theta)$$

Finally, to get a point estimate

$$\begin{aligned} E(Q|y^*) &= E(E(Q|y^*, \theta)|y^*) \\ &= E\left[\left(y^* + \sum_{i=n+1}^{N} E(Y_i|y^*, \theta)\right)/N|y^*\right] \\ &= [y^* + (N-n)E(\theta|y^*)]/N \end{aligned}$$

and the posterior variance is

$$\begin{aligned} Var(Q|y^*) &= E(Var(Q|y^*, \theta)) + Var(E(Q|y^*, \theta)) \\ &= \frac{1}{N^2}E[(N-n)\theta(1-\theta)|y^*] + Var(y^* + (N-n)\theta|y^*) \end{aligned}$$

In the following analysis, we extend the previous example to multivariate responses and try to obtain posterior means and variances of population counts in various categories. We assume the population is divided in groups (strata or clusters), the presence of which affects the response variable. Therefore, we want to include this effect in the modelling procedure. In our model, there are no covariates since their values are not available for the non-sampled elements and this makes prediction impossible. However, we introduce random effects corresponding to the groups as described in the following Sections.

4.2 Model description

To provide a general description for the model we discuss in this Chapter, we use in this Section the general term "group" when referring to design variables like strata or clusters. In following Sections where we describe several examples, we make the distinction between stratum and cluster. A robust model for a population with groups should reflect the variation of the variable of interest Y between them. A model that assigns different group means and/or variances would be suitable for a sample from this population. Moreover, the introduction of random effects helps when there are many and small groups to borrow strength between them. Thus, a random effects model represents differences between groups in terms of the proportions of Y in the population. Therefore, we assume multivariate random effects with respect to the categories of the response variable.

We are interested in estimating the cell counts in each category of the response variable $\boldsymbol{Q}_j = (Q_{j1}, ..., Q_{jC})$ for $j = 1, ..., M$ groups and C categories and the final population counts in each category $\boldsymbol{Q} = (Q_1, ..., Q_C)$.

Assume the response variable is y_{jk} is the number of units observed to take the kth possible category, $k = 1, ..., C$, $j = 1, ..., M$ and n_j the number of units in group j. The vector of probabilities in group j is

$$\boldsymbol{p}_j = (p_{j1}, p_{j2}, ..., p_{jC})$$

and the likelihood of the model is

$$f(\boldsymbol{y}|\boldsymbol{p}) \propto \prod_{j=1}^{M} \prod_{k=1}^{C} p_{jk}^{y_{jk}} \tag{4.1}$$

Thus,

$$\boldsymbol{y}_j \sim \text{Multinomial}(\boldsymbol{p}_j; n_j)$$

As we suppose no individual or group specified covariates for the model, we have only the fixed and the random intercept included. After choosing the first category as the baseline category, we get the log-odds

$$\eta_{jk} \equiv \log \frac{p_{jk}}{p_{j1}} = \mu_k + u_{jk}$$

or

$$\boldsymbol{\eta}_j = \boldsymbol{\mu} + \boldsymbol{u}_j$$

where $\boldsymbol{\mu} = (\mu_2, ..., \mu_C)$ and $\boldsymbol{u}_j = (u_{j2}, ..., u_{jC})$ for stratum j. Also, we can write the model in matrix-form

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\mu} + \boldsymbol{u}$$

where

$$\boldsymbol{X} = \mathbf{1}_M \otimes \boldsymbol{I}_{C-1}$$

$\mathbf{1}_M$ is the $M-$vector of ones, $\otimes$ denotes the Kronecker product and $\boldsymbol{u}^T = (\boldsymbol{u}_1^T, ..., \boldsymbol{u}_M^T)$. We continue assuming that

$$\boldsymbol{u}_j \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

and if we define $\boldsymbol{\Sigma}^* = \boldsymbol{I}_M \otimes \boldsymbol{\Sigma}$ then we can write $\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^*)$. The likelihood of the model becomes

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{u}, \boldsymbol{\Sigma}) &= \prod_{j=1}^{M} \prod_{k=2}^{C} \left(\frac{\exp(\eta_{jk})}{1 + \sum\limits_{l=2}^{C} \exp(\eta_{lk})} \right)^{y_{jk}} \\ &= \prod_{j=1}^{M} \left(\frac{\prod\limits_{k=2}^{C} \exp(\eta_{jk})^{y_{jk}}}{(1 + \sum\limits_{l=2}^{C} \exp(\eta_{lk}))^{n_j}} \right) \end{aligned} \tag{4.2}$$

The prior distribution for $\boldsymbol{\mu}$ is multivariate Normal

$$\boldsymbol{\mu} \sim N(\boldsymbol{a}_0, \boldsymbol{D}_0)$$

and for $\boldsymbol{\Sigma}$ an Inverse Wishart distribution is assigned

$$\boldsymbol{\Sigma} \sim IW(d, \boldsymbol{S}_0)$$

where $d > C$ the degrees of freedom and $\boldsymbol{S}_0$ the inverse scale matrix is positive definite. To sample from the Inverse Wishart distribution we used the R package MCMCpack, (Martin et al., 2010). The full conditional distributions are

$$\begin{aligned} f(\boldsymbol{\mu}|\boldsymbol{y},\boldsymbol{u},\Sigma) &= f(\boldsymbol{y}|\boldsymbol{\eta})f(\boldsymbol{\mu}|a_0,\boldsymbol{D}_0) \\ &\propto \prod_{j=1}^{M}\left(\frac{\prod\limits_{k=2}^{C}\exp(\eta_{jk})^{y_{jk}}}{(1+\sum\limits_{l=2}^{C}\exp(\eta_{lk}))^{n_j}}\right)\exp(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{a}_0)^T\,\boldsymbol{D}_0^{-1}\,(\boldsymbol{\mu}-\boldsymbol{a}_0)) \end{aligned} \tag{4.3}$$

$$f(\boldsymbol{u}_j|\boldsymbol{y},\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto f(\boldsymbol{y}|\boldsymbol{\eta})f(\boldsymbol{u}_j|\boldsymbol{\Sigma}) \propto \frac{\prod\limits_{k=2}^{C}\exp(\eta_{jk})^{y_{jk}}}{(1+\sum\limits_{l=2}^{C}\exp(\eta_{lk}))^{n_j}}\exp(-\frac{1}{2}\boldsymbol{u}_j^T\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{u}_j) \tag{4.4}$$

Finally,

$$\boldsymbol{\Sigma}|\boldsymbol{y},\boldsymbol{u},\boldsymbol{\mu} \sim IW(M+d,\boldsymbol{S}_0+\boldsymbol{u}^T\boldsymbol{u})$$

where $\boldsymbol{u}$ is the matrix $M \times (C-1)$ with $\boldsymbol{u}_j$ in row j.

4.3 Examples under different sampling designs

4.3.1 Stratified random sampling

As described in Section 3.2 stratified random sampling assumes sampling a number of units within each stratum. This sampling design is ignorable if we condition on the stratum variable. We assume that the sampling process is the same as in Section 3.2, the strata are created by cross-classifying *age* and *sex* and the number of selected units within strata is defined by proportional allocation. We choose to sample half of the individuals in each population stratum. The final sample is of size $n = 871$. The stratum sizes can be calculated straightforward using the sampling weights. The response variable is the *health status* and we are interested in estimating the number of individuals that fall into each category of health status for every stratum and the total number of individuals in each of the 5 categories of *health status*. Table 4.1 gives the distribution of *health status* in each stratum that is created from *age* and *sex*.

Table 4.1: Distribution of health status in age-sex strata

Strata	Health status					Totals
	1	2	3	4	5	
1	49	61	18	12	3	143
2	85	74	26	7	3	195
3	64	76	17	4	1	162
4	57	68	40	19	3	187
5	52	56	51	23	7	189
6	53	51	12	5	1	122
7	55	55	21	4	0	135
8	35	67	18	4	0	124
9	28	42	38	14	5	127
10	39	71	17	9	0	136
11	47	49	26	7	3	132
12	32	36	19	3	0	90
Totals	596	706	303	111	26	1742

Following the general notation of Section 2.3.2, we have $p = 1$, $q = 1$, $C = 5$, $M = 12$. We give hyperparameters the following values $\boldsymbol{a}_0 = \mathbf{0}$, $\boldsymbol{C}_0 = \text{diag}(10^5, ..., 10^5)$, $d = 6$, $\boldsymbol{S}_0 = diag(1, ..., 1)$ to produce diffuse prior distributions and this way express our weak prior knowledge. Sensitivity analysis showed that $d = 6$, $\boldsymbol{S}_0 = \text{diag}(1, ..., 1)$ as hyperparameters for the variance prior, result in desirable convergence. Different scale matrices $\boldsymbol{S}_0$, such as $\boldsymbol{S}_0 = \text{diag}(10, ..., 10)$, $\boldsymbol{S}_0 = \text{diag}(10^2, ..., 10^2)$, and $\boldsymbol{S}_0 = \text{diag}(10^3, ..., 10^3)$

were tried and rejected as they worsened convergence. Althought the possibility of using different prior distributions from the Inverse Wishart distribution (Gelman, 2006; Gelman et al., 2008), this is not examined here. Since it is not the purpose of this thesis, we are satisfied when the chain shows evidence of convergence in the trace plot. Simulation from the posterior of $\mathbf{\Sigma}$ is straightforward while for the rest we need to use Metropolis-Hastings algorithm since they are analytically intractable. Trace plots for μ_4 and μ_5 are given in a) of Figure 4.1. There is some evidence that the chains are not stable, especially for μ_5 that corresponds to the less populated category.

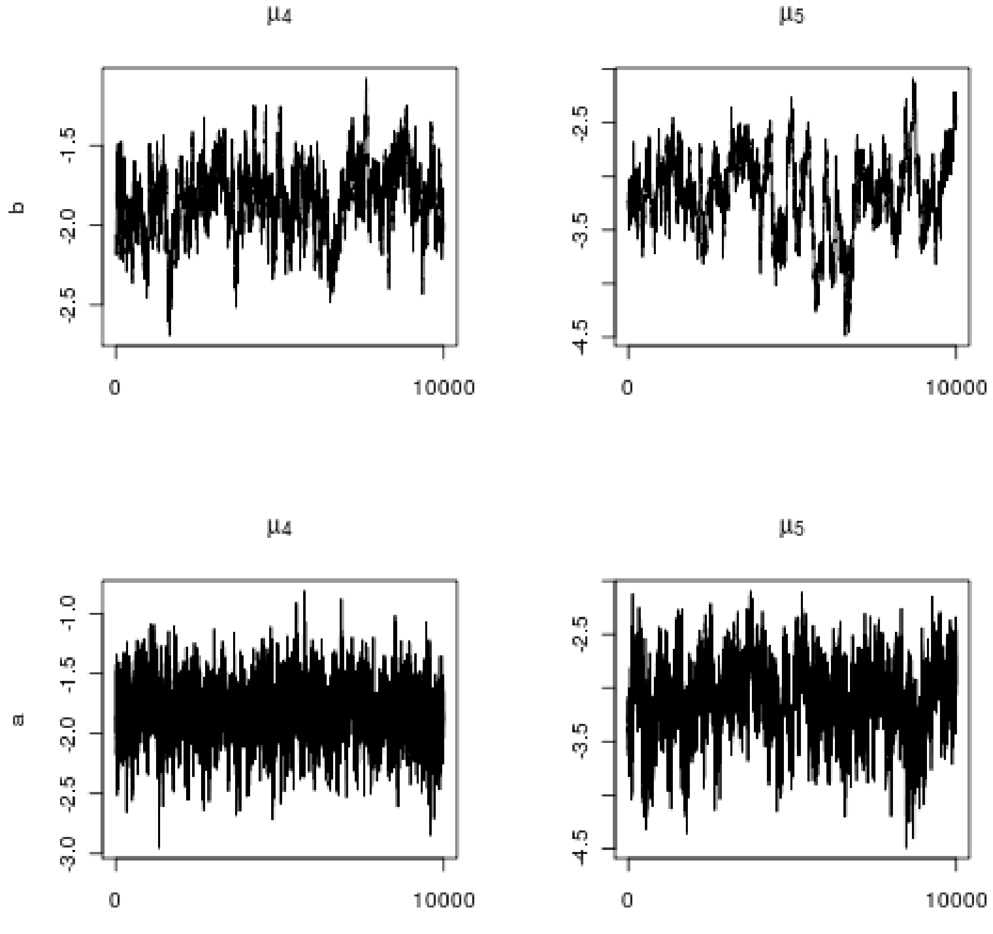

Figure 4.1: Trace plots for μ_4 and μ_5 for a stratified sample when a) non-reparameterisation b) hierarchical centering

The same problem gets worse for cells with even lower frequencies, as during cluster sampling in Section 4.3.2. This is a common problem appearing in discrete data models with random effects and it happens mainly due to the existence of high correlation in the posterior surface or weak identifiability of some model parameters that make convergence slow (Gelfand et al., 1996). There are methods to improve efficiency of MCMC techniques like reparameterisations, orthogonalisation or data expansion. The parameter identifiability problem appears when there are group specified covariates. Then, a *hierarchical centering* reparameterisation can improve convergence. It uses the fact that multilevel models contain a linear predictor consisting of variables with associated fixed effects and zero mean random effects, see Browne et al. (2009). The covariate here is constant within clusters associated with the random effects and the random effects can be centred around it. The mean of the new random effects will be a function of the original cluster-level predictors and fixed effects. However, hierarchical centering does not work well when the random effect variance is small and Gelfand et al. (1996) show this empirically for normal responses.

In our case, we can centre the random effects around the fixed intercept to simplify the algorithm. Thus, instead of considering the parameters $\boldsymbol{u}_j$, where $\boldsymbol{u}_j \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, we consider $\boldsymbol{\eta}_j|\boldsymbol{\mu}$ where $\boldsymbol{\eta}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Now, $\boldsymbol{\eta}_j$ are centred about $\boldsymbol{\mu}$ and the model can be written as

$$\boldsymbol{y}_j \sim Multinomial(\boldsymbol{p}_j; n_j)$$

$$\boldsymbol{\eta}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} \sim N(\boldsymbol{a}_0, \boldsymbol{D}_0)$$

and

$$\boldsymbol{\Sigma} \sim IW(d, \boldsymbol{S}_0)$$

The advantage is that the full conditional of $\boldsymbol{\mu}$ is now multivariate normal which is easier to simulate from using Gibbs sampler and so one less Metropolis step is implemented.

$$\boldsymbol{\mu}|\boldsymbol{\eta}, \boldsymbol{\Sigma} \sim N(\boldsymbol{a}_1, \boldsymbol{D}_1)$$

where

$$\boldsymbol{a}_1 = \boldsymbol{D}_1(\boldsymbol{D}_0^{-1}\boldsymbol{a}_0 + \boldsymbol{X}^T\boldsymbol{\Sigma}^{*-1}\boldsymbol{\eta})$$

and

$$\boldsymbol{D}_1 = (\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{\Sigma}^{*-1}\boldsymbol{X})^{-1}$$

Also,

$$f(\boldsymbol{\eta}_j|\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto f(\boldsymbol{y}_j|\boldsymbol{\eta}_j)f(\boldsymbol{\eta}_j|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and

$$\boldsymbol{\Sigma}|\boldsymbol{u} \sim IW(M + d, \boldsymbol{S}_0 + (\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\mu})^T(\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\mu}))$$

Since the full conditional distribution for $\boldsymbol{\eta}_j$ remains analytically intractable, Gibbs sampler cannot be used and a Metropolis step is added here to simulate from it. Hierarchical centering is applied in this Chapter when simulating from the full conditionals of a Multinomial logit model in order to improve convergence.

The new trace plots for the same parameters after applying hierarchical centering are given in *b*) of Figure 4.1, where we see that convergence is significantly improved. Moreover, the algorithm runs faster which saves substantial computational time. It allows us to run more simulations in less time and achieve desirable convergence as shown in Figures 4.2, 4.4 and 4.5.

The next step is to actually obtain new $\tilde{\boldsymbol{y}}$ for the non-sampled values. Assuming we obtained T simulations from the posterior distribution of the parameters $f(\boldsymbol{\mu}, \boldsymbol{u}, \boldsymbol{\Sigma}|\boldsymbol{y})$, we can produce T simulations from the posterior predictive of $\tilde{\boldsymbol{y}}_j$ with the following steps:

1. Calculate $\eta_{jk}^t = \mu_k^t + u_{jk}^t$ for $t = 1, ..., T$ simulations for $\boldsymbol{\mu}$ and $\boldsymbol{u}_j$.

2. Draw from the posterior distribution $f(\boldsymbol{p}_j|\boldsymbol{y}_j)$ by calculating

$$\begin{aligned} p^t_{jk} &= \frac{\exp(\eta^t_{jk})}{1+\sum_{l=2}^{C}\exp(\eta^t_{jl})} \quad \text{for } k = 2,3,...,C \\ p^t_{j1} &= \frac{1}{1+\sum_{l=2}^{C}\exp(\eta^t_{jl})} \quad k = 1 \end{aligned} \tag{4.5}$$

   for $t = 1, ..., T$ obtained draws. Finally, we have T random matrices of $M \times C$ containing the draws for $\boldsymbol{p}_j$ for $j = 1, ..., M$.

3. Obtain the stratum sizes N_j as described in Section 3.2 and calculate the $N_j - n_j$ sizes of non-sampled units for every stratum.

4. Get T simulations of new $\tilde{\boldsymbol{y}}^t_j$ from

$$\tilde{\boldsymbol{y}}^t_j \sim Multinom(\boldsymbol{p}^t_j; N_j - n_j)$$

Table 4.2: Posterior inference for $\boldsymbol{\mu}$

	Mean	s.e.	95%C.I..
$\boldsymbol{\mu}_2$	0.2050	1.1440	(-0.0772, 0.4898)
$\boldsymbol{\mu}_3$	-0.8015	0.2009	(-1.2053, -0.4131)
$\boldsymbol{\mu}_4$	-1.7968	0.2231	(-2.2503, -1.3720)
$\boldsymbol{\mu}_5$	-3.0901	0.3277	(-3.7731, -2.4899)

Table 4.3: Posterior inference for $\boldsymbol{\Sigma}$

	Mean	s.e.	95%C.I..
$\boldsymbol{\Sigma}_{11}$	0.1690	0.0864	(0.0671, 0.3878)
$\boldsymbol{\Sigma}_{22}$	0.3429	0.1958	(0.1163, 0.8426)
$\boldsymbol{\Sigma}_{33}$	0.2977	0.1924	(0.0912, 0.7982)
$\boldsymbol{\Sigma}_{44}$	0.3943	0.3040	(0.0973, 1.1846)

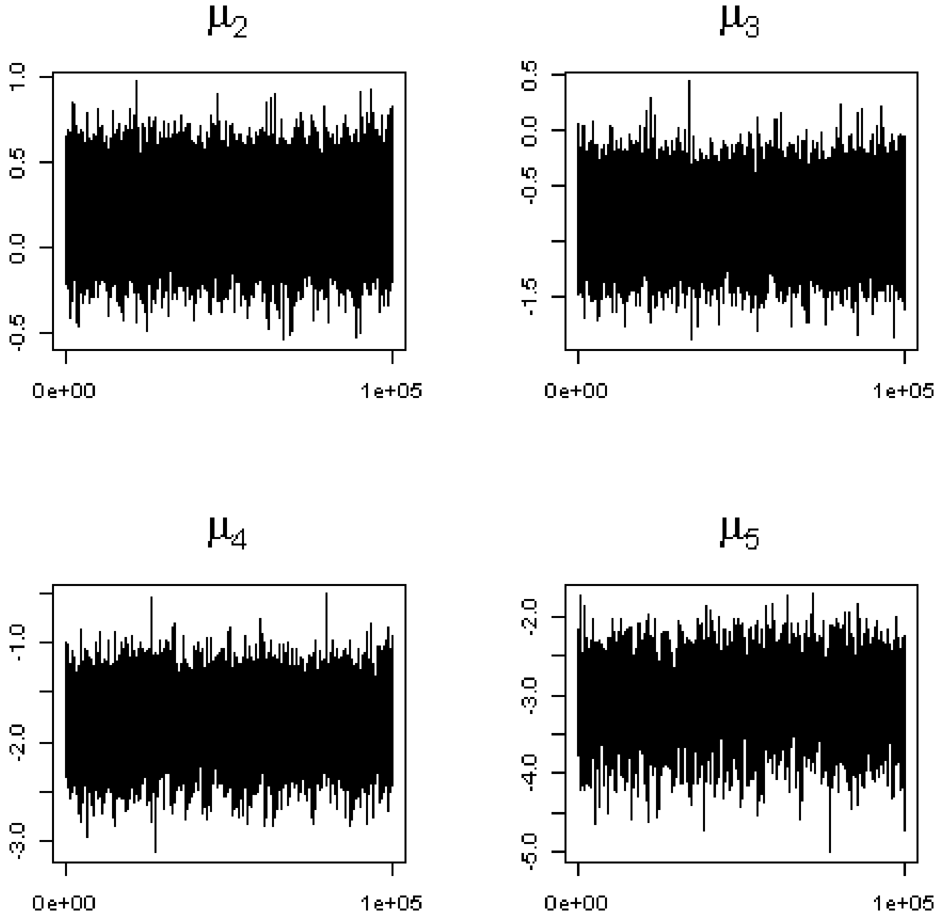

Figure 4.2: Trace plots for $\boldsymbol{\mu}$ under stratified sampling

As mentioned before, the quantity of interest $\boldsymbol{Q} = (Q_1, Q_2, ..., Q_5)$ is the vector of the population counts in each category which consists of the sum of the sampled units belonging to each category in all strata plus the sum of non-sampled units in all strata. It can be written as

$$\boldsymbol{Q} = \boldsymbol{Q}_S + \boldsymbol{Q}_{\bar{S}}$$

where S denotes the sampled part, $\bar{S}$ the non sampled. Also,

$$\boldsymbol{Q}_S = \sum_{j=1}^{M} \sum_{i \in S} y_{ij}$$

and

$$\boldsymbol{Q}_{\bar{S}} = \sum_{j=1}^{M} \sum_{i \in \bar{S}} y_{ij}$$

Hence, to get the posterior mean of $\boldsymbol{Q}$

$$\begin{aligned}
E(\boldsymbol{Q}|y_S) &= E[E(\boldsymbol{Q}|y_S, \boldsymbol{p})|y_S] \\
&= E\left[E\left(\sum_{j=1}^{M}\left(\sum_{i\in S} y_{ij} + \sum_{i\in\bar{S}} y_{ij}\right)|y_S, \boldsymbol{p}_j\right)|y_S\right] \\
&= E\left[\left(\sum_{j=1}^{M}\sum_{i\in S} y_{ij} + \sum_{j=1}^{M} E(\tilde{\boldsymbol{y}}_j|y_S, \boldsymbol{p}_j)\right)|y_S\right] \\
&= \sum_{j=1}^{M} \boldsymbol{y}_j + \sum_{j=1}^{M}(N_j - n_j)E(\boldsymbol{p}_j|y_S) \qquad (4.6)
\end{aligned}$$

Calculating the variance is a bit more complicated

$$\begin{aligned}
Var(\boldsymbol{Q}|y_s) &= Var(\boldsymbol{Q}_{\bar{S}}|y_S) \\
&= Var\left(E\left(\sum_{j=1}^{M}\tilde{\boldsymbol{y}}_j|y_S, \boldsymbol{p}_j\right)\right) + E\left(Var\left(\sum_{j=1}^{M}\tilde{\boldsymbol{y}}_j|y_S, \boldsymbol{p}_j\right)\right) \\
&= \sum_{j=1}^{M} Var(E(\boldsymbol{Q}_j|y_S, \boldsymbol{p}_j)) + \sum_{j=1}^{M} E(Var(\boldsymbol{Q}_j|y_S, \boldsymbol{p}_j)) \\
&= \sum_{j=1}^{M} Var((N_j - n_j)\boldsymbol{p}_j|y_S) \\
&+ \sum_{j=1}^{M} E((N_j - n_j)(diag(\boldsymbol{p}_j) - \boldsymbol{p}_j\boldsymbol{p}_j^T)|y_S) \\
&= \sum_{j=1}^{M} (N_j - n_j)^2 Var(\boldsymbol{p}_j|y_S) \\
&+ \sum_{j=1}^{M} (N_j - n_j)\left[E(diag(\boldsymbol{p}_j)|y_S) - E(\boldsymbol{p}_j\boldsymbol{p}_j^T|y_S)\right] \qquad (4.7)
\end{aligned}$$

Posterior totals and their standard deviation are given in Table 4.4, together with classical estimators for stratified sampling. The classical formulae used

Table 4.4: Estimates of health status responses in age-sex strata

Strata	Health status				
	1	2	3	4	5
1	58	50	19	11	5
2	77	80	25	9	4
3	62	76	16	7	2
4	61	72	35	15	4
5	56	60	52	15	5
6	51	51	13	5	2
7	55	56	18	5	1
8	37	71	11	5	1
9	27	47	37	11	5
10	38	71	18	8	1
11	47	53	20	9	2
12	27	36	20	6	1
Posterior Totals	596	723	284	107	33
Posterior st.dev.	13.93	14.44	10.79	6.99	3.95
Classical Estimators	598	724	286	106	32
Std error	19.50	20.26	15.11	9.89	5.54
True values	596	706	303	111	26

to calculate the stratified count estimates and estimates of their variance are:

$$\hat{\boldsymbol{q}} = \frac{N_j}{n_j}\boldsymbol{y}_j = N_j\boldsymbol{p}_j$$

and

$$\widehat{Var}(\hat{\boldsymbol{q}}) = \sum_j (N_j^2 - N_j n_j)(diag(\boldsymbol{p}_j) - \boldsymbol{p}_j\boldsymbol{p}_j^T)$$

where $\boldsymbol{y}_j$ and $\boldsymbol{p}_j$ are the observed counts and observed probabilities for stratum j. We see that Bayesian and frequentist estimators for population counts are close to the true values. Bayesian estimators have smaller variance in four out of five cases that suggests estimators closer to the true values. Bayesian approach has also the advantage of providing the whole posterior distribution of the population counts as plotted in Figure 4.3 together with the true values.

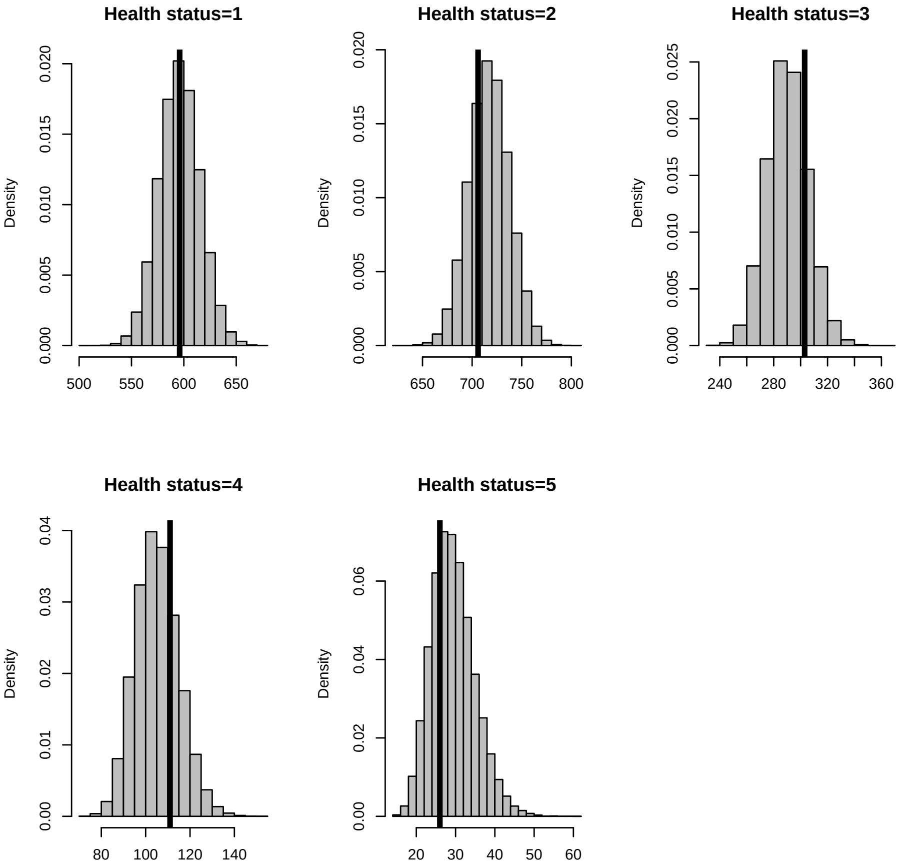

Figure 4.3: Posterior densities of population counts under stratified sampling

4.3.2 Cluster sampling with SRS.

In the following example, variable *area* is the cluster variable and a sample of $m = 30$ out of $M = 140$ areas is taken. According to one stage cluster sampling process, all the units within selected clusters are sampled. A total number of 366 units belonging to the 30 sampled clusters is included in the sample. Total number of units and clusters in the population is assumed to be known. Table 4.5 gives the distribution of the health status in 15 of the sampled clusters, where we observe many zero counts for category 5. Let Z denote the size variable corresponding to the cluster sizes in the population and Y denote the response variable *health status* as previously.

Table 4.5: Distribution of health status in sampled clusters

Clusters	Health status					Totals
	1	2	3	4	5	
12	5	11	2	0	0	18
17	2	4	0	2	1	9
23	3	3	1	2	0	9
40	0	3	2	1	0	6
44	5	5	1	0	0	11
46	5	4	2	3	0	14
48	2	8	0	0	0	10
55	5	9	1	0	0	15
60	4	4	2	0	1	11
64	6	7	1	0	0	14
68	8	2	3	2	0	15
70	1	7	0	0	0	8
73	3	5	2	1	0	11
80	5	5	1	1	0	12
82	1	9	5	1	0	16
Totals	126	156	65	24	6	377

Cluster sampling makes inference more complicated than stratified sampling since there are non-sampled clusters in the population for which we need to predict. We adopt random effects for the clusters as it allows the information from the sampled clusters to be used to predict for the non-sampled. In addition, we use the zero-truncated Poisson model described in Chapter 3 to model the cluster sizes as we assume again that cluster indicators are not given for the non-sampled units. As discussed in Section 3.3, the zero-truncated Poisson model is suitable for variables that do not present overdispersion, like *area*. The model for *area* is

$$f(\boldsymbol{z}|\lambda) = \prod_{j=1}^{m} \frac{\lambda^{z_j}}{z_j!(e^{\lambda} - 1)}$$

where z_j be the size for area j for $j = 1, ..., m$. Inference for *area* is made exactly as in Section 3.3.

The model for Y is the same Multinomial logit model as in the previous section. The only difference that not all the clusters are sampled, thus we have $M - m$ non-sampled clusters. We actually have two different models which we combine for inference, one for Z and one the response Y. Again we use a Metropolis-within-Gibbs algorithm to draw from the joint posterior distribution $f(\boldsymbol{\mu}, \boldsymbol{u}, \Sigma|\boldsymbol{y})$. Posterior inference for these parameters is given below, while for the sizes we take the results directly from Section 3.3.1. Convergence for the parameters is achieved by using a hierarchical centering parameterisation similarly as before.

Table 4.6: Posterior inference for $\boldsymbol{\mu}$ in cluster sampling with SRS

	Mean	s.e.	95%C.I.
$\boldsymbol{\mu}_2$	0.1064	0.1611	(-0.2084, 0.4244)
$\boldsymbol{\mu}_3$	-0.9042	0.2021	(-1.3145, -0.5200)
$\boldsymbol{\mu}_4$	-2.2073	0.3269	(-2.9030, -1.6065)
$\boldsymbol{\mu}_5$	-3.8374	0.6088	(-5.1374, -2.8030)

To make inference for the population counts for all clusters we need to obtain

Table 4.7: Posterior inference for $\mathbf{\Sigma}$ in cluster sampling with SRS

	Mean	s.e.	95%C.I..
$\mathbf{\Sigma}_{11}$	0.3015	0.1613	(0.1015, 0.7128)
$\mathbf{\Sigma}_{22}$	0.3610	0.2372	(0.1019, 0.9909)
$\mathbf{\Sigma}_{33}$	0.5769	0.5057	(0.1159, 1.9920)
$\mathbf{\Sigma}_{44}$	0.6096	0.8470	(0.1090, 2.6111)

samples from the posterior predictive distribution of new $\tilde{\boldsymbol{y}}_j$ for $j = m + 1, ..., M$. Hence, we first must draw new sizes $\tilde{z}$ for the unobserved clusters and then use these draws to get counts. To obtain a draw from the posterior predictive distribution of new data $\tilde{\boldsymbol{y}}_j$, we perform the following steps:

- Draw $(\boldsymbol{\mu}^t, \mathbf{\Sigma}^t)$ from their posterior distribution for $t = 1, ..., T$ number of simulations.
- Draw $M - m$ new $\tilde{\boldsymbol{u}}_j^t$ vectors as $\boldsymbol{u}_j \sim N(\mathbf{0}, \mathbf{\Sigma}^t)$.
- Draw $\tilde{\boldsymbol{p}}_j^t$ for the unsampled clusters as

$$\begin{aligned} \tilde{p}_{jk}^t &= \frac{\exp(\mu_k^t + \tilde{u}_{jk}^t)}{1 + \sum_{l=2}^{C} \exp(\mu_l^t + \tilde{u}_{jl}^t)} && \text{for } k = 2, ..., C \\ \tilde{p}_{j1}^t &= \frac{1}{1 + \sum_{l=2}^{C} \exp(\mu_l^t + \tilde{u}_{jl}^t)} && \text{for } k = 1 \end{aligned} \tag{4.8}$$

  for $t = 1, ..., T$ and $j = m + 1, ..., M$.
- Draw new $\tilde{z}_j^t$ from their posterior predictive for cluster $j = m+1, ..., M$ as in Section 3.3.
- Draw $\tilde{\boldsymbol{y}}_j^t$ from their posterior predictive distribution

$$\tilde{\boldsymbol{y}}_j^t \sim Multinom(\tilde{\boldsymbol{p}}_j^t; \tilde{z}_j^t)$$

To obtain the posterior mean and variance for the population counts $\boldsymbol{Q}$ we use again the conditional mean and variance formulae. For one stage cluster sampling, the non-sampled part $\boldsymbol{Q}_{\bar{S}}$ consists of the all the units within non selected clusters:

$$\boldsymbol{Q}_{\bar{S}} = \sum_{j=m+1}^{M} \sum_{i \in \bar{S}} y_{ij}$$

Therefore,

$$\begin{aligned}
E(\boldsymbol{Q}|y_S) &= E\left[\left(\sum_{j=1}^{m} \boldsymbol{y}_j + E(\sum_{j=m+1}^{M} \tilde{\boldsymbol{y}}_j | y_S, \tilde{\boldsymbol{p}}_j)\right)|y_S\right] \\
&= \sum_{j=1}^{m} \boldsymbol{y}_j + \sum_{j=m+1}^{M} E[(E(\tilde{\boldsymbol{y}}_j|y_S, \tilde{\boldsymbol{p}}_j))|y_S] \\
&= \sum_{j=1}^{m} \boldsymbol{y}_j + \sum_{j=m+1}^{M} E(\tilde{\boldsymbol{p}}_j\, \tilde{z}_j|y_S) \\
&= \sum_{j=1}^{m} \boldsymbol{y}_j + \sum_{j=m+1}^{M} E(\tilde{z}_j|z_S)E(\tilde{\boldsymbol{p}}_j|y_S) \qquad (4.9)
\end{aligned}$$

The posterior variance consists only of the variance of the counts in non-selected clusters:

$$\begin{aligned}
Var(\hat{Q}) = Var(\boldsymbol{Q}|y_S) &= \sum_{j=m+1}^{M} Var(\tilde{\boldsymbol{y}}_j|y_s) + 2\sum_{l<h} Cov(\tilde{\boldsymbol{y}}_l, \tilde{\boldsymbol{y}}_h|y_S) \\
&= \sum_{j=m+1}^{M} [E(Var(\tilde{\boldsymbol{y}}_j|\tilde{\boldsymbol{p}}_j, y_S)|y_S) + Var(E(\tilde{\boldsymbol{y}}_j|\tilde{\boldsymbol{p}}_j, y_S)|y_S)] \\
&+ 2\sum_{l<h} Cov[(E(\tilde{\boldsymbol{y}}_l|\tilde{\boldsymbol{p}}_l, y_S), E(\tilde{\boldsymbol{y}}_h|\tilde{\boldsymbol{p}}_h, y_S))|y_s] \\
&= \sum_{j=m+1}^{M} \left[E(\tilde{z}_j(diag(\tilde{\boldsymbol{p}}_j) - \tilde{\boldsymbol{p}}_j\tilde{\boldsymbol{p}}_j^T)|y_S) + Var(\tilde{z}_j\tilde{\boldsymbol{p}}_j|y_S)\right] \\
&+ 2\sum_{l<h} Cov[(\tilde{z}_l\tilde{\boldsymbol{p}}_l, \tilde{z}_h\tilde{\boldsymbol{p}}_h)|y_S] \qquad (4.10)
\end{aligned}$$

which can be calculated using the simulations drawn for $\tilde{z}_j$ and $\tilde{\boldsymbol{p}}_j$.

Classical estimators are also calculated using the following formulae taken from Lohr (1999)

$$\hat{\boldsymbol{q}} = M/m \sum_j \boldsymbol{y}_j$$

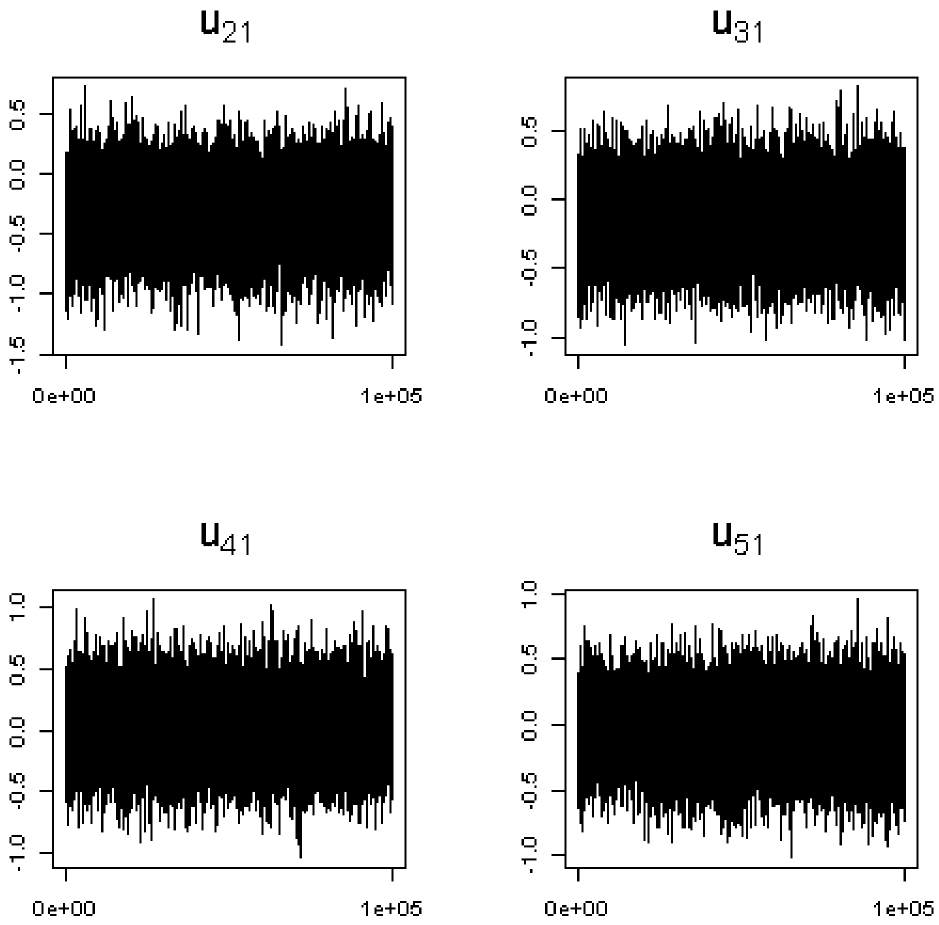

Figure 4.4: Trace plots for $\boldsymbol{u}_1$ under cluster sampling with SRS

and

$$se(\hat{\boldsymbol{q}}) = \sqrt{\left(1 - \frac{m}{M}\right)\frac{s_t^2}{m}}$$

where

$$s_t^2 = \frac{1}{m-1}\sum_j \left(\boldsymbol{y}_j - \frac{\hat{\boldsymbol{q}}}{M}\right)^2$$

and are presented in Table 4.8. We observe that Bayesian estimators for health status 1-4 haver smaller standard deviation than classical estimators, the difference is significantly higher for the first two categories. Thus, we can argue that Bayesian estimators tend to be closer to the true values than frequentist estimators.

Table 4.8: Estimated population counts for *health status* in cluster sampling with SRS

	Category				
	1	2	3	4	5
Posterior totals	623	715	289	95	20
Posterior st. dev.	30.28	33.59	19.98	10.84	4.61
Classical estim.	540	600	232	76	16
Std Error	86.11	95.68	36.99	12.12	3.55
True values	596	706	303	111	26

4.3.3 Two stage cluster sampling with PPS

In this section we assume two stage cluster sampling where the PSU are selected with PPS sampling and SSU with SRS. The cluster variable is *district* and the first stage of sampling is already implemented in Section 3.5 through Poisson sampling. Then, we assume that $n_j = N_j/2$ units within district j are selected and that the second stage sampling fraction $f_2 = 1/2$ is given to the data analyst. If it is not, it is impossible to calculate N_j for sampled j and use them to model for the non-sampled. The reason is that weights here are the inverse of $p_i = \phi \frac{N_j}{N} \frac{n_j}{N_j}$ for i unit belonging to cluster j, where N_j cancels out. Nevertheless, if f_2 is known we can calculate $N_j = n_j/f_2$ for sampled cluster j and use this information to model the size variable. From the first stage we have 9 selected clusters out of 43 in the population, as described in Section 3.5. During the second stage, with the assumed $f_2 = 1/2$, 362 individuals out of the 725 belonging to the 9 sampled clusters are included in the sample.

The model for the population counts of *health status* remains the same and inferences for $\boldsymbol{\mu}$, $\boldsymbol{u}_j$ and $\boldsymbol{\Sigma}$ are obtained as in previous sections. Moreover, the model for district sizes is the zero-truncated Negative Binomial model described in Section 3.5 that is suitable for PPS sampling. Posterior inference about the multinomial model parameters is given in Tables 4.9, 4.10, while for the size variable Z results are taken directly from Section 3.5.

Table 4.9: Posterior inference for $\boldsymbol{\mu}$ in two stage cluster sampling with PPS

	Mean	s.e.	95%C.I..
$\boldsymbol{\mu}_2$	0.3442	0.2202	(-0.0892, 0.7809)
$\boldsymbol{\mu}_3$	-0.6170	0.2540	(-1.1217, -0.1193)
$\boldsymbol{\mu}_4$	-1.8152	0.3459	(-2.5148, -1.1551)
$\boldsymbol{\mu}_5$	-3.0806	0.7242	(-4.8418, -1.9708)

Table 4.10: Posterior inference for $\boldsymbol{\Sigma}$ in two stage cluster sampling with PPS

	Mean	s.e.	95%C.I..
$\boldsymbol{\Sigma}_{11}$	0.2534	0.1658	(0.0813, 0.6796)
$\boldsymbol{\Sigma}_{22}$	0.2974	0.2092	(0.0877, 0.8417)
$\boldsymbol{\Sigma}_{33}$	0.4221	0.3793	(0.0977, 1.3928)
$\boldsymbol{\Sigma}_{44}$	1.4349	1.9534	(0.1488, 6.5364)

When it comes to prediction for non-sampled cases, we distinguish between non-sampled units within selected clusters and completely non-sampled clusters. In particular, we have m sampled clusters with sizes N_j for $j = 1, ..., m$, from which n_j units are selected. Moreover, there are $M - m$ non-sampled clusters, each with z_j size, for $j = m + 1, ..., M$. The random variable Z is used to denote non-sampled sizes. Therefore, a combination of steps used in the two previous sections for drawing new data $\tilde{\boldsymbol{y}}$ is applied as following:

- Draw $(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$ from their posterior distribution for $t = 1, ..., T$ number of simulations.

- Draw $M - m$ new $\tilde{\boldsymbol{u}}_j^t$ vectors as $\boldsymbol{u}_j \sim N(\mathbf{0}, \boldsymbol{\Sigma}^t)$.

- Draw from the posterior distribution of $\boldsymbol{p}_j$ given parameteres and data using

$$p^t_{jk} = \frac{\exp(\eta^t_{jk})}{1 + \sum_{l=2}^{C} \exp(\eta^t_{jl})} \quad \text{for } k = 2, 3, ..., C$$

$$p^t_{j1} = \frac{1}{1 + \sum_{l=2}^{C} \exp(\eta^t_{jl})} \quad k = 1$$

for $j = 1, ..., m$ sampled clusters and $t = 1, ..., T$ obtained draws.

- Draw $\tilde{\boldsymbol{p}}^t_j$ for $j = m + 1, ..., M$ unsampled clusters by using

$$\tilde{p}^t_{jk} = \frac{\exp(\mu^t_k + \tilde{u}^t_{jk})}{1 + \sum_{l=2}^{C} \exp(\mu^t_l + \tilde{u}^t_{jl})} \quad \text{for } k = 2, ..., C$$

$$\tilde{p}^t_{j1} = \frac{1}{1 + \sum_{l=2}^{C} \exp(\mu^t_l + \tilde{u}^t_{jl})} \quad \text{for } k = 1$$

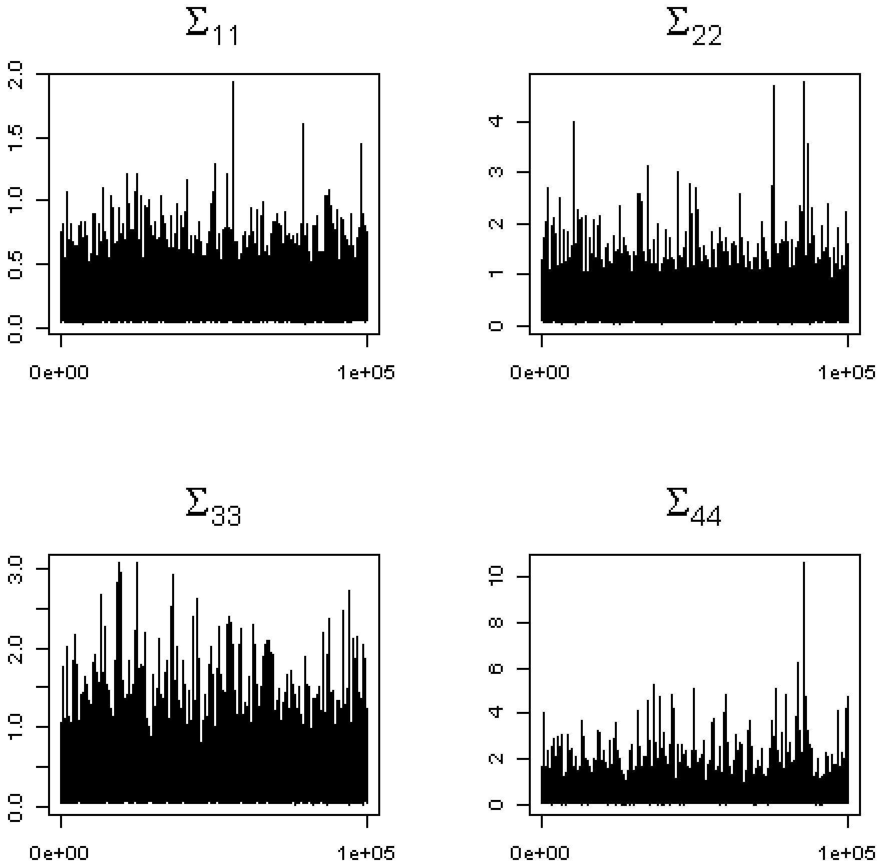

Figure 4.5: Trace plots for $\boldsymbol{\Sigma}$ under cluster sampling with PPS

for for $j = m+1, ..., M$ sampled clusters and $t = 1, ..., T$ obtained draws.

- Use $N_j - n_j$ as the number of non-selected units in sampled cluster j for $j = 1, ..., m$ and draw new $\tilde{z}_j^t$ from their posterior predictive for cluster $j = m+1, ..., M$ as in Section 3.5.

- Draw $\tilde{\boldsymbol{y}}_j^t$ from their posterior predictive distribution

$$\tilde{\boldsymbol{y}}_k^t \sim Multinom(\boldsymbol{p}_j^t; N_j - n_j) \qquad \text{for } j = 1, ..., m$$

  and from

$$\tilde{\boldsymbol{y}}_j^t \sim Multinom(\tilde{\boldsymbol{p}}_j^t; \tilde{z}_j^t) \qquad \text{for } j = m+1, ..., M$$

There is an extra part of uncertainty that affects calculation of the posterior variance of the population counts since it is added in the total variance. The population counts for *health status* can be decomposed again as

$$\boldsymbol{Q} = \boldsymbol{Q}_S + \boldsymbol{Q}_{\bar{S}}$$

where now

$$\boldsymbol{Q}_{\bar{S}} = \sum_{j=1}^{m} \sum_{i \in \bar{S}} y_{ij} + \sum_{j=m+1}^{M} \sum_{i \in \bar{S}} y_{ij}$$

and their posterior mean and variance are

$$\begin{aligned} E(\boldsymbol{Q}|y_S) &= E\left[\left(\sum_{j=1}^{m} \boldsymbol{y}_j + E\left(\sum_{j=1}^{m} \tilde{\boldsymbol{y}}_j|y_S, \boldsymbol{p}_j\right) + E\left(\sum_{j=m+1}^{M} \tilde{\boldsymbol{y}}_j|y_S, \tilde{\boldsymbol{p}}_j\right)\right)|y_S\right] \\ &= \sum_{j=1}^{m} \boldsymbol{y}_j + \sum_{j=1}^{m} (N_j - n_j)E(\boldsymbol{p}_j|y_S) + \sum_{j=m+1}^{M} E(\tilde{\boldsymbol{p}}_j\, \tilde{z}_j|y_S) \\ &= \sum_{j=1}^{m} \boldsymbol{y}_j + \sum_{j=1}^{m} (N_j - n_j)E(\boldsymbol{p}_j|y_S) + \sum_{j=m+1}^{M} E(\tilde{z}_j|z_S)E(\tilde{\boldsymbol{p}}_j|y_S) \end{aligned} \tag{4.11}$$

The posterior variance is the variance of the non-sampled units in selected clusters plus the variance of the non-sampled counts

$$\begin{aligned}
Var(\boldsymbol{Q}|y_S) &= \sum_{j=1}^{M} Var(\tilde{\boldsymbol{y}}_j|y_s) + 2\sum_{l<h} Cov(\tilde{\boldsymbol{y}}_l, \tilde{\boldsymbol{y}}_h|y_S) \\
&= \sum_{j=1}^{m} Var(E(\tilde{\boldsymbol{y}}_j|y_S, \boldsymbol{p}_j)) + \sum_{j=1}^{m} E(Var(\tilde{\boldsymbol{y}}_j|y_S, \boldsymbol{p}_j)) \\
&+ \sum_{j=m+1}^{M} Var(E(\tilde{\boldsymbol{y}}_j|\tilde{\boldsymbol{p}}_j, y_S)|y_S) + \sum_{j=m+1}^{M} E(Var(\tilde{\boldsymbol{y}}_j|\tilde{\boldsymbol{p}}_j, y_S)|y_S) \\
&+ 2\sum_{l<h} Cov[(E(\tilde{\boldsymbol{y}}_l|\tilde{\boldsymbol{p}}_l, y_S), E(\tilde{\boldsymbol{y}}_h|\tilde{\boldsymbol{p}}_h, y_S))|y_S] \\
&= \sum_{j=1}^{m} (N_j - n_j)^2 Var(\boldsymbol{p}_j|y_S) + \sum_{j=1}^{m} (N_j - n_j) E((diag(\boldsymbol{p}_j) - \boldsymbol{p}_j \boldsymbol{p}_j^T)|y_S) \\
&+ \sum_{j=m+1}^{M} Var(\tilde{z}_j \tilde{\boldsymbol{p}}_j|y_S) + \sum_{j=m+1}^{M} E(\tilde{z}_j(diag(\tilde{\boldsymbol{p}}_j) - \tilde{\boldsymbol{p}}_j \tilde{\boldsymbol{p}}_j^T)|y_S) \\
&+ 2\sum_{l<h} Cov[(\tilde{z}_l \tilde{\boldsymbol{p}}_l, (N_h - n_h)\boldsymbol{p}_h)|y_S]
\end{aligned} \tag{4.12}$$

which can be calculated using the simulations $\tilde{z}_j$, $\boldsymbol{p}_j$ and $\tilde{\boldsymbol{p}}_j$. Table 4.11 gives the Bayesian estimates for the population counts in each category and the classical HT estimators from Equation 2.1. The variance of HT estimators (see Equation 2.2) is not calculated due to the complexity of the existing formulae for sample size larger than 2. At this point, Bayesian inference also provides a method to obtain posterior variances and the chance to visualise the whole posterior distribution of the population counts.

Table 4.11: Estimated population counts for *health status* in two stage cluster sampling with PPS

	Category				
	1	2	3	4	5
Estimators	554	744	303	101	42
Std error	44.82	59.61	29.70	13.79	10.51
Classical estim.	566	761	307	109	43
True	596	706	303	111	26

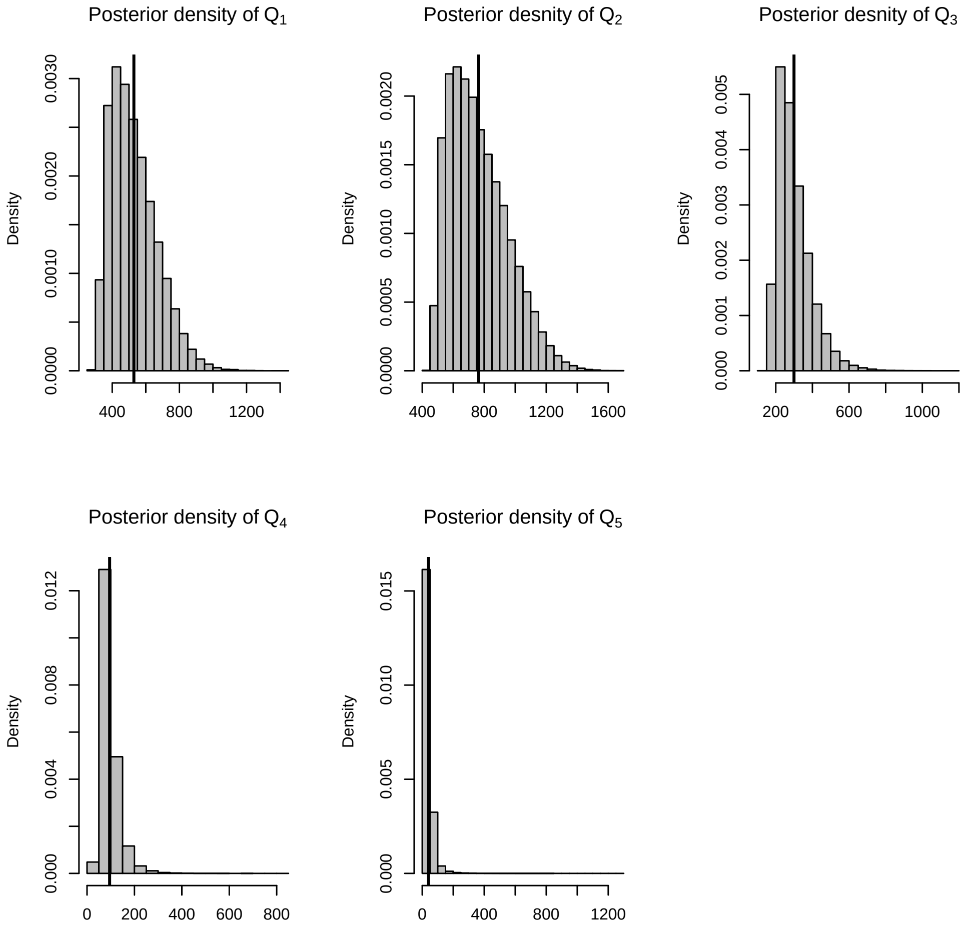

Figure 4.6: Posterior densities for population counts in two stage PPS sampling

4.4 Discussion

In this Chapter we presented a unified approach for multivariate categorical data coming from finite populations. The Bayesian Multinomial model together with modelling the size variable from Chapter 3, can provide estimates close to the true values for population counts of a variable with many outcomes. We assumed multivariate random effects that depend on the category, modelled with multivariate Normal distributions. Moreover, we provided formulae for calculating the posterior mean and variance of population counts in different sampling designs.

Classical estimates for each sampling design were calculated too and compared with the Bayesian results. The Bayesian approach as developed in this Chapter offers advantages such as

- Posterior variance can always be calculated, no matter the sampling design is or the sample size.
- It generally yields point estimates with smaller posterior variance than the classical approach.
- It provides a simple way to model categorical outcomes that can be extended in more complicated models for contingency tables and generate better estimates by accounting for model uncertainty (see Chapter 5).

Chapter 5

Modelling contingency tables

5.1 Introduction

In this Chapter we use the previously described Multinomial model with random effects to develop inference for contingency tables under sampling design based on strata or clusters. Analysts are usually interested in contingency tables with two or more dimensions and the relations between the variables that construct the tables. Our approach provides a method for prediction alternative to regression when the design variables are not available for non-sampled units.

As discussed in Section 2.4, sampling design may have a significant effect on classical methods of analysing contingency tables. We mentioned in Chapter 2 when discussing classical methods for contingency tables that under particular designs chi-square tests are not valid. When cluster sampling is performed, the within-cluster correlation might have an effect on the p-value of these tests. Chi-square tests under cluster sampling tend also to produce significant associations between the cross-classifying variables when they are not. This means that other factors affect the relation between these two variables. On the other hand, ignoring stratification can give conservative tests and large confidence intervals. One case where stratification presents no problems is when the strata are the categories of one of the cross-classified variables. However, if there are many strata, tables with a large number of cells are created that are hard to analyse. Another weak point of classical methods is that chi-squared tests are affected by a small sample size that produces many zeros in the table.

For these reasons we see there is a need for Bayesian analysis of contingency tables that is robust under various sampling schemes. The Multinomial model with random effects is used and compared to classical methods of analysing cross-classified data. One of the important aspects in this Chapter is comparing between models and choosing the most suitable or averaging over several models to obtain better estimates. Before we start describing the models used in this Chapter, we need to explain how we approximate the marginal likelihood of a model which is essential for model comparison and model averaging in later Sections.

5.2 Model comparison for GLMMs

As already mentioned in Section 1.2.4, the marginal likelihood of a model is used to evaluate its posterior model probability and the Bayes factor between two models. If we have $l = 1, ..., L$ potential models, we need to evaluate

$$f(\boldsymbol{y}|l) = \int f(\boldsymbol{y}|\boldsymbol{\theta}_l, l) f(\boldsymbol{\theta}_l) d\boldsymbol{\theta}_l \tag{5.1}$$

for each model. In the following analysis, we focus on describing a general methodology to approximate integrals, called *bridge sampling* and drop the subscript l. Bridge sampling is a method of Monte Carlo integration that was first proposed by Meng and Wong (1996) for approximating the ratio of normalising constants. We use and explain the method as suggested by Overstall and Forster (2010) and Overstall (2009).

Suppose we want to approximate the following general integral

$$I = \int_{\Theta} g(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

which is the normalising constant of a distribution $\pi(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) / \int g(\boldsymbol{\theta}) d\boldsymbol{\theta}$ and if $g(\boldsymbol{\theta}) = f_l(\boldsymbol{y}|\boldsymbol{\theta}_l) f_l(\boldsymbol{\theta}_l)$ then $I = f_l(\boldsymbol{y})$, the marginal likelihood for model l. Now suppose that $h(\boldsymbol{\theta})$ is a probability density function and $\gamma(\boldsymbol{\theta})$ is a function for which the following expectations are non-zero and finite. Since

$$\frac{\int \gamma(\boldsymbol{\theta}) g(\boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \gamma(\boldsymbol{\theta}) g(\boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta}} = 1$$

and

$$g(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \int g(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

then

$$\begin{aligned} \frac{\int \gamma(\boldsymbol{\theta}) g(\boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \gamma(\boldsymbol{\theta}) h(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \cdot I d\boldsymbol{\theta}} &= 1 \Leftrightarrow \\ \frac{\int \gamma(\boldsymbol{\theta}) g(\boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \gamma(\boldsymbol{\theta}) h(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{E_h[\gamma(\boldsymbol{\theta}) g(\boldsymbol{\theta})]}{E_\pi[\gamma(\boldsymbol{\theta}) h(\boldsymbol{\theta})]} &= I \end{aligned}$$

We approximate the nominator and denominator using Monte Carlo and so, the *bridge sampling* approximation to I is

$$\hat{I} = \frac{\frac{1}{n_h} \sum_{i=1}^{n_h} \gamma(\boldsymbol{\theta}_i^h) g(\boldsymbol{\theta}_i^h)}{\frac{1}{n_\pi} \sum_{i=1}^{n_\pi} \gamma(\boldsymbol{\theta}_i^\pi) h(\boldsymbol{\theta}_i^\pi)}$$

where $\{\boldsymbol{\theta}_1^h, ..., \boldsymbol{\theta}_{n_h}^h\}$ and $\{\boldsymbol{\theta}_1^\pi, ..., \boldsymbol{\theta}_{n_h}^\pi\}$ are samples generated from $h(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ respectively and n_h, n_π the sample sizes.

Meng and Wong (1996) show that the optimal $\gamma(\boldsymbol{\theta})$ with respect to minimising the variance of the approximation is

$$\gamma_o(\boldsymbol{\theta}) = (n_\pi g(\boldsymbol{\theta}) + n_h I h(\boldsymbol{\theta}))^{-1}$$

We see that that the optimal $\gamma(\boldsymbol{\theta})$ depends on the unknown I and to solve this Meng and Wong (1996) suggest starting from an initial value and then iterating the following scheme until convergence

$$\hat{I}^{(t+1)} = \frac{\frac{1}{n_h} \sum_{i=1}^{n_h} \frac{l_{hi}}{n_\pi l_{hi} + n_h \hat{I}^{(t)}}}{\frac{1}{n_\pi} \sum_{i=1}^{n_\pi} \frac{1}{n_\pi l_{\pi i} + n_h \hat{I}^{(t)}}} \qquad (5.2)$$

where $l_{ki} = g(\boldsymbol{\theta}_i^k)/h(\boldsymbol{\theta}_i^k)$ for $k = h, \pi$.

To perform bridge sampling we have to define an initial value $\hat{I}^{(0)}$, the probability distribution h and the allocation of the sample sizes $\frac{n_h}{n_h+n_\pi}$. Overstall (2009) suggests that using any of $\hat{I}^{(0)} = 0$ or $\hat{I}^{(0)} = \infty$ seems sensible as the iterative scheme (5.2) converges fast. Choosing a suitable h is more complicated and in general, it is required that h mimics π as closely as possible. Here, we use directly the approach Overstall (2009) concludes as best. This approach uses *Warp bridge sampling* (Meng and Shilling, 2002), where $h \equiv N(\mathbf{0}, \boldsymbol{I}_k)$ (or $h \equiv t_\nu(\mathbf{0}, \boldsymbol{I}_k)$) and π is transformed or "warped" to $\tilde{\pi}$ so that its properties match those of h.

To implement this, suppose $\boldsymbol{\theta} \sim \pi$, where the location and spread of π are $\boldsymbol{\mu}$ and $\boldsymbol{W} = \boldsymbol{S}\boldsymbol{S}^T$. We warp π to $\tilde{\pi}$ using the stochastic transformation

$$b\boldsymbol{S}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})$$

where b is Bernoulli($\frac{1}{2}$) on the sample space $\{-1, 1\}$. The probability density function of $\tilde{\pi}$ is now

$$\begin{aligned}
\tilde{\pi}(\boldsymbol{\theta}) &= \frac{1}{2}|\boldsymbol{S}|[\pi(\boldsymbol{\mu} - \boldsymbol{S}\boldsymbol{\theta}) + \pi(\boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{\theta})] \\
&= \frac{\frac{1}{2}|\boldsymbol{S}|[g(\boldsymbol{\mu} - \boldsymbol{S}\boldsymbol{\theta}) + g(\boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{\theta})]}{\int g(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&= \frac{\tilde{g}(\boldsymbol{\theta})}{\int g(\boldsymbol{\theta})d\boldsymbol{\theta}}
\end{aligned}$$

If $\{\boldsymbol{\theta}_1^\pi, ..., \boldsymbol{\theta}_{n_\pi}^\pi\}$ and $\{\boldsymbol{\theta}_1^h, ..., \boldsymbol{\theta}_{n_h}^h\}$ are samples from π and h respectively then the *Warp III bridge sampling* is obtained by iterating (5.2) until convergence, where

$$l_{hi} = |\boldsymbol{S}|\frac{g(\boldsymbol{\mu} - \boldsymbol{S}\boldsymbol{\theta}_i^h) + g(\boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{\theta}_i^h)}{2h(\boldsymbol{\theta}_i^h)} \tag{5.3}$$

and

$$l_{\pi i} = |\boldsymbol{S}|\frac{g(\boldsymbol{\theta}_i^\pi) + g(2\boldsymbol{\mu} - \boldsymbol{\theta}_i^\pi)}{2h(\boldsymbol{S}^{-1}(\boldsymbol{\theta}_i^\pi - \boldsymbol{\mu}))} \tag{5.4}$$

While there are methods to computationally find the optimum $\boldsymbol{\mu}$ and $\boldsymbol{S}$, they get hard to implement for high dimensional problems. Overstall (2009) proposes to take $\boldsymbol{\mu}$ to be the mean or mode and $\boldsymbol{W}$ to be the variance of π. Overstall (2009) also examines two strategies that investigate how to allocate $\{\boldsymbol{\theta}_1^{\pi}, ..., \boldsymbol{\theta}_N^{\pi}\}$ in order to find $\boldsymbol{\mu}$ and $\boldsymbol{S}$ and to use in the bridge sampler, the proportion strategy and the split strategy. The preferred strategy was the split strategy and this is the one we adopt. Finally, we use $N_h = N_\pi$ and $n_h = n_\pi = \frac{1}{2}N_h = \frac{1}{2}N_\pi$.

To summarise, we use the bridge sampling algorithm he suggested best approximates I with respect to minimising the mean squared error. The algorithm is:

1. Generate a sample $\{\boldsymbol{\theta}_1^{\pi}, ..., \boldsymbol{\theta}_{N_\pi}^{\pi}\}$ of size N_π from the target distribution π and a sample $\{\boldsymbol{\theta}_1^{h}, ..., \boldsymbol{\theta}_{N_h}^{h}\}$ of size N_h from $h \equiv N(0, \boldsymbol{I}_k)$.

2. Let $n_\pi = \frac{1}{2}N_\pi$ and $n_h = \frac{1}{2}N_h$.

3. Let $\boldsymbol{\mu}$ and $\boldsymbol{W} = \boldsymbol{S}\boldsymbol{S}^T$ be the sample mean and variance of $\{\boldsymbol{\theta}_1^{\pi}, ..., \boldsymbol{\theta}_{n_\pi}^{\pi}\}$.

4. Compute l_{hi} using (5.3) for $i = n_h + 1, ..., N_h$ and $l_{\pi i}$ using (5.4) for $i = n_\pi + 1, ..., N_\pi$

5. Let $\hat{I}^1$ be the final value of the following converged iterative scheme
$$\hat{I}^{(t+1)} = \frac{\frac{1}{n_h}\sum_{i=n_h+1}^{N_h} \frac{l_{hi}}{n_\pi l_{hi} + n_h \hat{I}^{(t)}}}{\frac{1}{n_\pi}\sum_{i=n_\pi+1}^{N_\pi} \frac{1}{n_\pi l_{\pi i} + n_h \hat{I}^{(t)}}}$$

6. Let $\boldsymbol{\mu}$ and $\boldsymbol{W} = \boldsymbol{S}\boldsymbol{S}^T$ be the sample mean and variance of $\{\boldsymbol{\theta}_{n_\pi+1}^{\pi}, ..., \boldsymbol{\theta}_{N_\pi}^{\pi}\}$.

7. Compute l_{hi} using (5.3) for $i = 1, ..., n_h$ and $l_{\pi i}$ using (5.4) for $i = 1, ..., n_\pi$

8. Let $\hat{I}^2$ be the final value of the following converged iterative scheme
$$\hat{I}^{(t+1)} = \frac{\frac{1}{n_h}\sum_{i=1}^{n_h} \frac{l_{hi}}{n_\pi l_{hi} + n_h \hat{I}^{(t)}}}{\frac{1}{n_\pi}\sum_{i=1}^{n_\pi} \frac{1}{n_\pi l_{\pi i} + n_h \hat{I}^{(t)}}}$$

9. Let $\hat{I} = \frac{1}{2}(\hat{I}^1 + \hat{I}^2)$.

5.3 Two way contingency tables

5.3.1 Model description

For a two way contingency table, let $i = 1, ..., R$ and $j = 1, ..., C$ denote the categories for the two variables constructing the table. Let k be the group indicator for $k = 1, ..., M$ and y_{kl} be the response l in cluster k that can take one of the discrete values of the "row" variable indexed by $i = 1, ..., R$ and one of the discrete values of "column" variable indexed by $j = 1, ..., C$, with probabilities $\boldsymbol{p}_k = (p_{11k}, p_{12k}, ..., p_{21k}, ..., p_{RCk})$ in cluster k. The likelihood of the model is written

$$f(\boldsymbol{y}|\boldsymbol{p}) \propto \prod_{k=1}^{m}\prod_{i=1}^{R}\prod_{j=1}^{C} p_{ijk}^{y_{ijk}} \tag{5.5}$$

where y_{ijk} is the number of units in cluster k that fall into category i of "row" variable and j of "column" variable. Also, $\sum_i \sum_j p_{ijk} = 1$ and $\sum_i \sum_j y_{ijk} = n_k$ which is the sample size of cluster k. The baseline cell is the first cell, where both variable indexes take value 1. Now, β_i^r and β_j^c can be thought of as coefficients of dummy variables for the last $R-1$ categories of the row-variable and the $C-1$ categories of the column variable in the table. Also, $\beta_1^r = \beta_1^c = 0$. Hence, we have

$$\boldsymbol{\beta}^r = (\beta_2^r, \beta_3^r)$$

and

$$\boldsymbol{\beta}^c = (\beta_2^c, \beta_3^c, \beta_4^c, \beta_5^c)$$

Assuming independence in the table, the number of parameters in the model is $R + C - 2$ and the log-odds can be written as function of the parameters:

$$\eta_{ijk} \equiv \log \frac{p_{ijk}}{p_{11k}} = \text{logit}(p_{ijk}) = \beta_i^r + \beta_j^c + u_{ijk}$$

where $\boldsymbol{\beta}^r$ and $\boldsymbol{\beta}^c$ are the fixed parameters and are constant across clusters and cluster effects $\boldsymbol{u}_k$ vary between clusters. All parameters depend on (i, j). To generalise we can write the model in a vector form

$$\boldsymbol{\eta}_k = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{u}_k$$

where assuming independence

$$\boldsymbol{x} = \left[\begin{array}{cc|cccc}
1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{array}\right]$$

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^r \\ \boldsymbol{\beta}^c \end{pmatrix}$$

and $\boldsymbol{u}_k$ a vector of length $RC - 1$ corresponding to the number of cells with $\boldsymbol{u}_k \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. In general, matrix $\boldsymbol{x}$ is $(RC - 1) \times (R + C - 2)$ when no interactions are included in the model.

Including the interactions between the two variables produces $(R-1)(C-1)$ extra terms β^{rc}_{ij} that are the coefficients of the product of the dummy variables for β^r_i, β^c_j. Also here we have $\beta^{rc}_{1j} = \beta^{rc}_{i1} = 0$ and the number of parameters in the model is now $(R-1)+(C-1)+(R-1)(C-1) = RC-1$. The fixed parameters are now

$$\boldsymbol{\beta}^r = (\beta^r_2, \beta^r_3)$$

$$\boldsymbol{\beta}^c = (\beta^c_2, \beta^c_3, \beta^c_4, \beta^c_5)$$

$$\boldsymbol{\beta}^{rc} = (\beta^{rc}_{22}, \beta^{rc}_{23}, \beta^{rc}_{24}, \beta^{rc}_{25}, \beta^{rc}_{32}, \beta^{rc}_{33}, \beta^{rc}_{34}, \beta^{rc}_{35})$$

The log-odds are written as

$$\eta_{ijk} \equiv \log \frac{p_{ijk}}{p_{11k}} = \text{logit}(p_{ijk}) = \beta^r_i + \beta^c_j + \beta^{rc}_{ij} + u_{ijk}$$

Including the interactions, we have

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^r \\ \boldsymbol{\beta}^c \\ \boldsymbol{\beta}^{rc} \end{pmatrix}$$

and $\boldsymbol{x}$ has $(R-1)(C-1)$ extra columns that are constructed by multiplying each of the existing $r-$columns with each of the $c-$columns from the independence model $\boldsymbol{x}$ matrix.

The more general matrix-form is

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

where $\boldsymbol{X} = \mathbf{1}_M \otimes \boldsymbol{x}$, $\boldsymbol{u}^T = (\boldsymbol{u}^T_1, ..., \boldsymbol{u}^T_M)$ and $\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^*)$ with $\boldsymbol{\Sigma}^* = \boldsymbol{I}_M \otimes \boldsymbol{\Sigma}$. After assigning multivariate Normal prior distributions to fixed parameters and an Inverse Wishart to the covariance matrix of random effects, we summarise the model

$$\boldsymbol{y}_k \sim Multinomial(\boldsymbol{p}_k; n_k)$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{a}_0, \boldsymbol{D})$$

$$\boldsymbol{u}_k \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma} \sim IW(d, \boldsymbol{S}_0)$$

In terms of the original probabilities we have

$$p_{ijk} = \frac{\exp(\eta_{ijk})}{1 + \sum_{i=2}^R \sum_{j=2}^C \exp(\eta_{ijk})} \quad \text{for } i = 2, ..., R \quad \text{and } j = 2, ..., C$$

$$p_{11k} = \frac{1}{1 + \sum_{i=2}^R \sum_{j=2}^C \exp(\eta_{ijk})} \quad \text{for } i = j = 1$$

and the likelihood becomes

$$f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}) \propto \prod_{k=1}^{M}\prod_{i=2}^{R}\prod_{j=2}^{C}\left(\frac{\exp(\eta_{ijk})}{1+\sum_{i=2}^{R}\sum_{j=2}^{C}\exp(\eta_{ijk})}\right)^{y_{ijk}}$$

The full conditional distributions are written

$$\begin{aligned}f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{u},\boldsymbol{\Sigma}) &\propto f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u})f(\boldsymbol{\beta})\\ &= f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u})\exp(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{a}_0)^T\boldsymbol{D}^{-1}(\boldsymbol{\beta}-\boldsymbol{a}_0)) \qquad (5.6)\end{aligned}$$

$$\begin{aligned}f(\boldsymbol{u}_k|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\Sigma}) &\propto f(\boldsymbol{y}_k|\boldsymbol{u}_k)f(\boldsymbol{u}_k|\boldsymbol{\Sigma})\\ &= \frac{\prod_{i=2}^{R}\prod_{j=2}^{C}\exp(\eta_{jk})^{y_{jk}}}{(1+\sum_{i=2}^{R}\sum_{j=2}^{C}\exp(\eta_{ijk}))^{n_k}}\exp(-\frac{1}{2}\boldsymbol{u}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{u}_k) \qquad (5.7)\end{aligned}$$

and

$$\boldsymbol{\Sigma}|\boldsymbol{u} \sim IW(M+d,\boldsymbol{S}_0+\boldsymbol{U}^T\boldsymbol{U}) \qquad (5.8)$$

Next, we give an example of fitting the above model to our dataset and compare between independence and interaction model.

5.3.2 Example

We wish to examine our model under cluster sampling, therefore we use the sample taken in Section 3.5, where cluster sampling with PPS was performed. *Health status* and *marital status* are the two cross-classifying variables of the table. *Marital status* has 3 categories, single, married/couple and divorced/widowed/separated. The response variable is the health status for individuals in different marital statuses and we are also interested in the relation between these two variables across districts. The whole contingency table is modelled as one multinomial response and the first cell is chosen to represent the baseline category.

In Table 5.1, we give the contingency tables created for *district 11* and *district 18* to illustrate the apparent differences between districts. Note that the sample is the sample as Section 3.5 and see there the description of the sampling procedure. The number of sampled districts is $m = 9$ out of $M = 43$ and the number of sampled individuals is $n = 725$ out of $N = 1742$.

Table 5.1: Health status for single, married/couple, divorced/widowed/separated in districts 11 and 18

		Health status				
District	Marital status	1	2	3	4	5
11	1	3	4	2	1	0
	2	12	7	4	0	1
	3	4	6	2	2	0
18	1	16	30	17	5	1
	2	24	38	17	5	3
	3	9	11	12	10	2

The prior distribution for $\boldsymbol{\beta}$ is assumed to be Normal with zero mean and large variance-covariance matrix and the hyperparameters for $\boldsymbol{\Sigma}$ are $d = 17$ and $\boldsymbol{S}_0 = diag(1, ..., 1)$. This way we reflect our ignorance about the parameters. In Table 5.2 we present posterior inference for $\boldsymbol{\beta}$ and Figure 5.1 shows achieved convergence of the chains for the independence model. The draws from the posterior distribution of the parameters are then used to obtain inference about population cell counts in following Section.

Posterior inference for the fixed parameters of the model including interactions is given in Table 5.3. Observing the Table 5.3, we can see that most of the interactions parameter C.I. include zero which may suggest that they are not significant. This gives some evidence in favour of the independence model and before we proceed with estimating the population counts, we need to determine which model fits best our data. In the following Section we develop *bridge sampling* in practice in order to approximate the marginal likelihood

Table 5.2: Posterior inference for $\boldsymbol{\beta}$ under independence model

	Mean	st.dev.	95%C.I..
$\boldsymbol{\beta}_2^r$	0.7884	0.1762	(0.4450, 1.1415)
$\boldsymbol{\beta}_3^r$	-0.0190	0.1770	(-0.3577, 0.3401)
$\boldsymbol{\beta}_2^c$	0.1930	0.1626	(-0.1308, 0.5117)
$\boldsymbol{\beta}_3^c$	-0.6286	0.1937	(-1.0283, -0.2585)
$\boldsymbol{\beta}_4^c$	-1.7395	0.2589	(-2.280, -1.2560)
$\boldsymbol{\beta}_5^c$	-2.9214	0.3479	(-3.6490, -2.2790)

of both models.

Table 5.3: Posterior inference for $\boldsymbol{\beta}$ under interaction model

	Mean	st.dev.	95%C.I..
$\boldsymbol{\beta}_2^r$	0.9067	0.2240	(0.4713, 1.3543)
$\boldsymbol{\beta}_3^r$	-0.4425	0.2748	(-0.9926, 0.0937)
$\boldsymbol{\beta}_2^c$	0.3308	0.2418	(-0.1504, 0.8031)
$\boldsymbol{\beta}_3^c$	-0.7608	0.2982	(-1.3613, -0.1843)
$\boldsymbol{\beta}_4^c$	-2.0503	0.4171	(-2.9350, -1.2760)
$\boldsymbol{\beta}_5^c$	-3.1983	0.6834	(-4.7010, -2.0350)
$\boldsymbol{\beta}_{22}^{rc}$	-0.2201	0.3455	(-0.8960, 0.4845)
$\boldsymbol{\beta}_{23}^{rc}$	0.0457	0.3953	(-0.7340, 0.8300)
$\boldsymbol{\beta}_{24}^{rc}$	-0.3672	0.5513	(-1.4526, 0.7177)
$\boldsymbol{\beta}_{35}^{rc}$	-0.1652	0.8502	(-1.7634, 1.5943)
$\boldsymbol{\beta}_{32}^{rc}$	0.0379	0.4041	(-0.7513, 0.8431)
$\boldsymbol{\beta}_{33}^{rc}$	0.6502	0.4557	(-0.2421, 1.5534)
$\boldsymbol{\beta}_{34}^{rc}$	1.5518	0.5535	(0.4713, 2.6538)
$\boldsymbol{\beta}_{35}^{rc}$	1.1949	0.8670	(-0.4126, 2.9743)

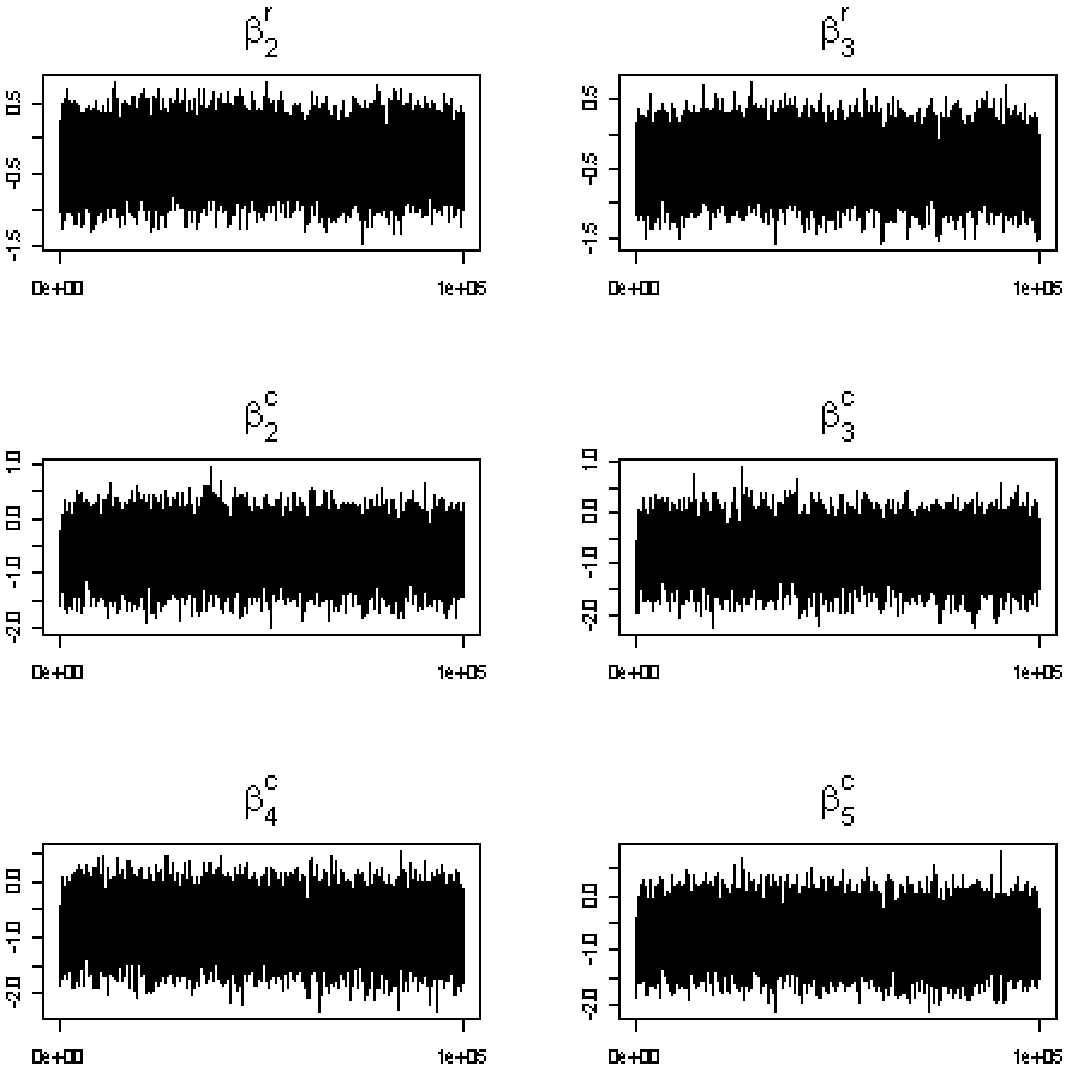

Figure 5.1: Trace plots for $\boldsymbol{\beta}$ assuming independence in contingency table

5.3.3 Bridge sampling in practice

In the previous Section we described both the independence and interactions model. We call them *Model 1* (M1) and *Model 2* (M2) respectively during the following analysis. To use the notation of Section 1.3 for GLMM, let $p_1 = R + C - 1$ be the fixed effects dimension for *Model 1* and $p_2 = RC - 1$

for *Model 2*. The rest of the parameter dimensions are the same across the two models, $q = RC - 1 = 14$, $M = 9$. We see that only parameter $\boldsymbol{\beta}$ changes between the two models while $\boldsymbol{u}$ and $\boldsymbol{\Sigma}$ remain the same and are assigned the same prior distributions. Hence, we can write the marginal likelihood of *Model 1* as

$$\begin{aligned} f(\boldsymbol{y}|M1) &= \int\int\int f(\boldsymbol{y}|\boldsymbol{\beta}_1, \boldsymbol{u}, M1) f(\boldsymbol{u}|\boldsymbol{\Sigma}) f(\boldsymbol{\Sigma}) f(\boldsymbol{\beta}_1) \mathrm{d}\boldsymbol{\beta}_1 \mathrm{d}\boldsymbol{u} \mathrm{d}\boldsymbol{\Sigma} \\ &= \int\int f(\boldsymbol{y}|\boldsymbol{\beta}_1, \boldsymbol{u}, M1) f(\boldsymbol{\beta}_1) \int f(\boldsymbol{u}|\boldsymbol{\Sigma}) f(\boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\Sigma} \mathrm{d}\boldsymbol{\beta}_1 \mathrm{d}\boldsymbol{u} \end{aligned}$$

and the marginal likelihood for *Model 2* as

$$\begin{aligned} f(\boldsymbol{y}|M2) &= \int\int\int f(\boldsymbol{y}|\boldsymbol{\beta}_2, \boldsymbol{u}, M2) f(\boldsymbol{u}|\boldsymbol{\Sigma}) f(\boldsymbol{\Sigma}) f(\boldsymbol{\beta}_1) \mathrm{d}\boldsymbol{\beta}_1 \mathrm{d}\boldsymbol{u} \mathrm{d}\boldsymbol{\Sigma} \\ &= \int\int f(\boldsymbol{y}|\boldsymbol{\beta}_2, \boldsymbol{u}, M2) f(\boldsymbol{\beta}_2) \int f(\boldsymbol{u}|\boldsymbol{\Sigma}) f(\boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\Sigma} \mathrm{d}\boldsymbol{\beta}_2 \mathrm{d}\boldsymbol{u} \end{aligned}$$

Since the prior distribution for $\boldsymbol{\Sigma}$ is Inverse Wishart, $\mathrm{IW}(d, \boldsymbol{S}_0)$ for both models, the integral

$$\int f(\boldsymbol{u}|\boldsymbol{\Sigma}) f(\boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\Sigma}$$

is analytically tractable as

$$\int f(\boldsymbol{u}|\boldsymbol{\Sigma}) f(\boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\Sigma} = \frac{\Gamma_q(\frac{d+M}{2})}{\Gamma_q(\frac{d}{2})} \frac{1}{\pi^{(Mq)/2}} \frac{|\boldsymbol{S}_0|^{d/2}}{|\boldsymbol{S}_0 + \sum_{k=1}^{M} \boldsymbol{u}_k \boldsymbol{u}_k^T|^{(d+M)/2}}$$

where

$$\Gamma_q(a) = \pi^{\frac{1}{4}q(q-1)} \prod_{i=1}^{q} \Gamma\left(a + \frac{1-i}{2}\right)$$

is the multivariate Gamma function. Hence, the marginal likelihoods are reduced to

$$f(\boldsymbol{y}|M1) = \int\int f(\boldsymbol{y}|\boldsymbol{\beta}_1, \boldsymbol{u}, M1) f(\boldsymbol{\beta}_1) \frac{\Gamma_q(\frac{d+M}{2})}{\Gamma_q(\frac{d}{2})} \frac{1}{\pi^{(Mq)/2}}$$
$$\frac{|\boldsymbol{S}_0|^{d/2}}{|\boldsymbol{S}_0 + \sum_{k=1}^{M} \boldsymbol{u}_k \boldsymbol{u}_k^T|^{(d+M)/2}} \mathrm{d}\boldsymbol{u} \mathrm{d}\boldsymbol{\beta}_1$$

and

$$f(\boldsymbol{y}|M2) = \int_{\int} f(\boldsymbol{y}|\boldsymbol{\beta}_2, \boldsymbol{u}, M2) f(\boldsymbol{\beta}_2) \frac{\Gamma_q(\frac{d+M}{2})}{\Gamma_q(\frac{d}{2})} \frac{1}{\pi^{(Mq)/2}}$$

$$\frac{|\boldsymbol{S}_0|^{d/2}}{|\boldsymbol{S}_0 + \sum_{k=1}^{M} \boldsymbol{u}_k \boldsymbol{u}_k^T|^{(d+M)/2}} \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{\beta}_2$$

Now, we have to approximate the two integrals with bridge sampling, where $g(\boldsymbol{\beta}_1, \boldsymbol{u}) = f(\boldsymbol{y}|M1)$ and $g(\boldsymbol{\beta}_2, \boldsymbol{u}) = f(\boldsymbol{y}|M2)$. Hence, for *Model 1* we have

$$g(\boldsymbol{\beta}_1, \boldsymbol{u}) = \frac{\prod_{i=2}^{R} \prod_{j=2}^{C} \exp(\beta_i^r + \beta_j^c + u_{ijk})^{y_{jk}}}{(1 + \sum_{i=2}^{R} \sum_{j=2}^{C} \exp(\beta_i^r + \beta_j^c + u_{ijk}))^{n_k}}$$

$$(2\pi)^{-p_1/2} |D_1|^{-1/2} \exp(-\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{a}_0)^T \boldsymbol{D}_{\boldsymbol{1}}^{-1} (\boldsymbol{\beta}_1 - \boldsymbol{a}_0))$$

$$\frac{\Gamma_q(\frac{d+M}{2})}{\Gamma_q(\frac{d}{2})} \frac{1}{\pi^{(Mq)/2}} \frac{|\boldsymbol{S}_0|^{d/2}}{|\boldsymbol{S}_0 + \sum_{k=1}^{M} \boldsymbol{u}_k \boldsymbol{u}_k^T|^{(d+M)/2}}$$

and for *Model 2*

$$g(\boldsymbol{\beta}_2, \boldsymbol{u}) = \frac{\prod_{i=2}^{R} \prod_{j=2}^{C} \exp(\beta_i^r + \beta_j^c + \beta_{ij}^{rc} + u_{ijk})^{y_{jk}}}{(1 + \sum_{i=2}^{R} \sum_{j=2}^{C} \exp(\beta_i^r + \beta_j^c + \beta_{ij}^{rc} + u_{ijk}))^{n_k}}$$

$$(2\pi)^{-p_2/2} |D_2|^{-1/2} \exp(-\frac{1}{2}(\boldsymbol{\beta}_2 - \boldsymbol{a}_0)^T \boldsymbol{D}_{\boldsymbol{2}}^{-1} (\boldsymbol{\beta}_2 - \boldsymbol{a}_0))$$

$$\frac{\Gamma_q(\frac{d+M}{2})}{\Gamma_q(\frac{d}{2})} \frac{1}{\pi^{(Mq)/2}} \frac{|\boldsymbol{S}_0|^{d/2}}{|\boldsymbol{S}_0 + \sum_{k=1}^{M} \boldsymbol{u}_k \boldsymbol{u}_k^T|^{(d+M)/2}}$$

We have already a sample of size N_π from the joint posterior distribution of $(\boldsymbol{u}, \boldsymbol{\beta}_1)$ and $(\boldsymbol{u}, \boldsymbol{\beta}_2)$ for both models from Section 5.3.2. We also set $N_h = N_\pi$ and $n_h = n_\pi = \frac{1}{2} N_\pi$. Then, we use the algorithm described in page 98 and we get the values for the log marginal likelihoods shown in Table 5.4. Bayes factor for these two models gives strong evidence in favour of the independence model. Therefore, we conclude that *health status* is independent of *marital status* given *district* effect. However, classical chi-square test suggests interactions between the two variables with $X_8^2 = 37.1$ that rejects independence model. Also when fitting and comparing classical log-linear models, independence model has a deviance of 32.715 with 8 degrees of freedom that again gives evidence in favour of the full model (interactions model).

Table 5.4: Approximated log marginal likelihoods by bridge sampling

Model	log Marginal likelihood
Independence	-1645.341
Interactions	-1654.882

5.3.4 Inference for the contingency tables counts

In this Section we return to the main interest of this thesis, estimating population counts for categorical data. Hence, we want to estimate the population cell counts for the contingency table of *health status* and *marital status*. As the sample is obtained through one stage PPS sampling, we can use the algorithm described in page 84 in order to draw from the posterior predictive distribution of new data. Then, we use the the Equations (4.11)-(4.12) to get the posterior mean and variance for the population counts.

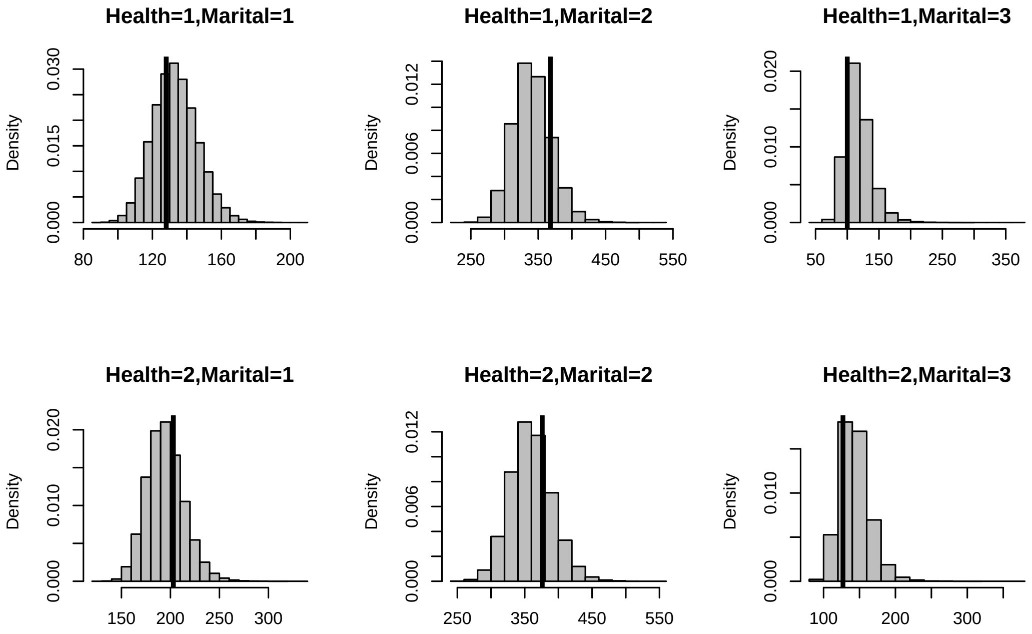

Figure 5.2: Distribution of posterior predictive counts for two way contingency tables under the independence model

We select the independence model as it has the largest marginal distribution

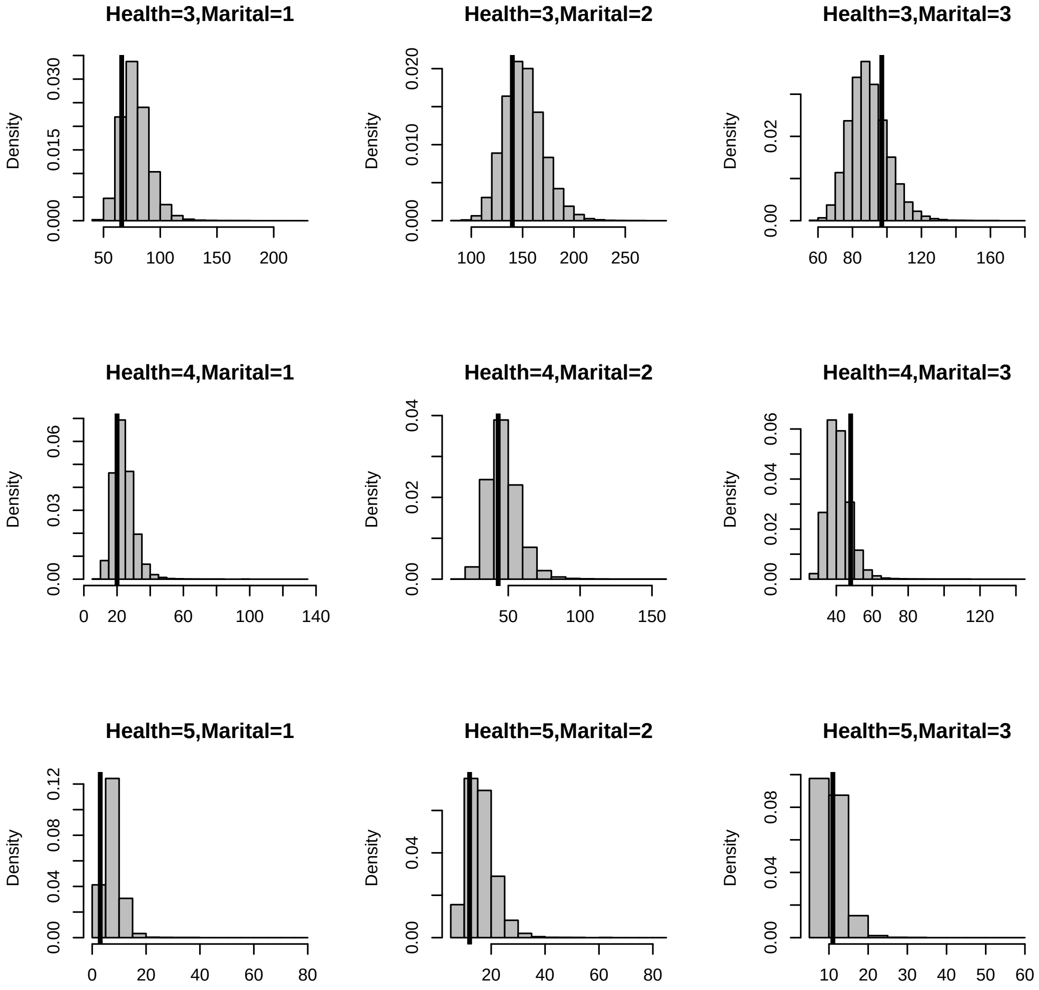

Figure 5.3: Distribution of posterior predictive counts for two way contingency tables under the independence model

and gives a log Bayes factor of 9.541 in favour of the independence model and its posterior probability almost equal to 1. This makes that model a strongly dominant model and model averaging not useful in this case. The histograms show the posterior predictive distribution for the population count in each cell of the contingency table. The black line represents the true population counts in each cell. We see that our model of choice tends to yield accurate predictions of the true values.

5.4 Three way contingency tables

5.4.1 Introduction

The motivation for this Section is Clogg and Eliason (1987) approach where three way contingency tables under stratification are analysed. They suggest a method of analysis to take into account the survey design when fitting a log-linear model. This analysis assumes that sampling weights exist for the dataset. In general, when dealing with a weighted dataset the two common approaches are: use of the unweighted data and ignore completely the weights, or use the weighted data as if they were unweighted. Clogg and Eliason (1987) claim that both strategies can be incorrect since they may give biased estimates, wrong standard errors and fit statistics. Their method works effectively when we have equal probability selection or the stratification variable is one of the variables in the table. However, it fails when we have unequal probabilities of selection within cells, as Skinner and Vallet (2010) investigate.

To describe Clogg and Eliason (1987) method, we introduce the log-linear model for a contingency table with G cells, as described by Clogg and Eliason (1987) and Skinner and Vallet (2010)

$$\log(y_s) = X\beta_s$$

where y_s is the $G \times 1$ vector with the expected frequencies of the cells of the table, X is the $G \times p$ model matrix and β_s is the $p \times 1$ vector of the parameters. Now, y_s and β_s correspond to sample values and thus, the parameters of the model depend on the sampling design. The corresponding population model can be expressed as

$$\log(Y) = X\beta$$

Assuming all units within a specific cell have the same probability of inclusion and that π denotes the $G \times 1$ vector of these probabilities, then

$$\log(y_s) = \log(\pi) + \log(Y)$$

that gives us

$$\log(y_s) = \log(\pi) + X\beta \tag{5.9}$$

The cell inclusion probabilities are assumed to be the inverse of the cell weights, $\pi_g = 1/w_g$ for a cell g. In the case where weights are different within cells, Clogg and Eliason (1987) introduce the *average cell weight*. Let y^w be the weighted frequencies

$$y_g^w = \sum_{t=1}^{n} I_{gt} w_t \tag{5.10}$$

where

$$I_{gt} = \begin{cases} 1 & \text{if unit } t \text{ falls in cell } g \\ 0 & \text{otherwise} \end{cases}$$

and w_t is the weight of unit t. Note that w_t here are normalised weights to sum up to the sample size n. Then, the *average cell weight* is

$$w_g = y_g^w / y_g$$

and $1/w_g$ is an estimator of the π_g. Clogg and Eliason (1987) then fit this model using $\log(1/w_g)$ as an offset. Skinner and Vallet (2010) claim that this method is appropriate only in the case where sampling weights are constant within cells. Also, it is not valid for more complicated sampling schemes such as cluster sampling.

5.4.2 Model description

To extend the notation from previous Section to three-way tables, we must change the notation in order to introduce another index for the the third variable. Let $i = 1, ..., R$, $j = 1, ..., C$ and $k = 1, ..., A$ denote the categories for the three variables constructing the contingency table. Let $m = 1, ..., M$ denote the stratum indicator. Let y_{mt} be the response t in stratum m that can take one of the discrete values for each cross-classified variable, with probabilities $\boldsymbol{p}_m = (p_{111m}, p_{112m}, ..., p_{RCAm})$ in stratum m. The likelihood of the model is written

$$f(\boldsymbol{y}|\boldsymbol{p}) \propto \prod_{m=1}^{M}\prod_{i=1}^{R}\prod_{j=1}^{C}\prod_{k=1}^{A} p_{ijkm}^{y_{ijkm}} \tag{5.11}$$

where y_{ijkm} is the number of units in stratum m that fall into cell (i, j, k). Also, $\sum_i \sum_j \sum_k p_{ijkm} = 1$ and $\sum_i \sum_j \sum_k y_{ijkm} = n_m$ which is the sample size of stratum m. The baseline cell is the first cell, where all variable indexes take value 1. Again, β_i^r, β_j^c and β_k^a can be thought of as coefficients of dummy variables for the last $R-1$, $C-1$ and $A-1$ categories of each variable respectively.

In a three-way table we see main effects, first order interactions and second order interactions. Different combinations of these parameters produce different models. For the independence model the constraints are

$$\beta_1^r = \beta_1^c = \beta_1^a = 0$$

for first order interactions

$$\beta_{1j}^{rc} = \beta_{i1}^{rc} = \beta_{i1}^{ra} = \beta_{1k}^{ra} = \beta_{1k}^{ca} = \beta_{j1}^{ca} = 0$$

and for second order a final constraint is added to the previous

$$\beta_{1jk}^{rca} = \beta_{i1k}^{rca} = \beta_{ij1}^{rca} = 0$$

Finally, we have 9 different models with various combinations of the contingency table variables. These models are given in Table 5.5.

Assuming again a Multinomial model with random effects we can write the general form of log-odds as

$$\boldsymbol{\eta} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{u}_m$$

where vector $\boldsymbol{\beta}$ contains the fixed parameters corresponding to the model we want to analyse. We summarise

$$\boldsymbol{y}_m \sim Multinomial(\boldsymbol{p}_m; n_m)$$

Table 5.5: Models of interest in a three-way contingency table

	Model	Number of parameters
1	$[H][M][S]$	R+C+A-3
2	$[H][M][S][HM]$	A+RC-2
3	$[H][M][S][MS]$	C+RA-2
4	$[H][M][S][HS]$	R+CA-2
5	$[H][M][S][HM][MS]$	RC+RA-R-1
6	$[H][M][S][HM][HS]$	RC+CA-C-1
7	$[H][M][S][MS][HS]$	RA+CA-A-1
8	$[H][M][S][HM][MS][HS]$	R(C-1)+A(R-1)+C(A-1)
9	$[H][M][S][HM][MS][HS][HMS]$	ARC-1

$$\boldsymbol{\beta} \sim N(\boldsymbol{a_0}, \boldsymbol{D})$$

$$\boldsymbol{u}_l \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma} \sim IW(d, \boldsymbol{S}_0)$$

At this point, we need to choose the hyper-parameter values and extra care is required for $\boldsymbol{\beta}$. The choice of prior distribution is important when it comes to model comparison. Forster and O'Hagan (2004) discuss this effect when comparing nested models. The marginal posterior probability of a model is proportional to the product of the prior probability and the marginal likelihood and marginal likelihood is sensitive to the prior distribution of the parameters. Therefore, also model posterior probability is sensitive to the prior distribution of the parameters, except the prior distribution of the common parameters in all models. The parameters present in all models here are the main effect parameters. Assume $\boldsymbol{\beta}_0$ is the parameters present in all models, that are the main effects parameters and $\boldsymbol{\beta}_1$ the additional parameters in any augmented model. Then, assigning a diffuse prior for $\boldsymbol{\beta}_1$ can produce a large Bayes factor against the augmented model. This is another example of *Lindley's paradox* (Forster and O'Hagan, 2004). To avoid this happening, we can assign less vague prior distributions when prior information about $\boldsymbol{\beta}_1$ is weak. Therefore, we assign

$$\boldsymbol{\beta}_0 \sim N(\mathbf{0}, diag(10^5, ..., 10^5)$$

and

$$\boldsymbol{\beta}_1 \sim N(\mathbf{0}, diag(25, ..., 25)$$

Next, we present an example for a three way contingency table under stratified sampling, obtain posterior inference for the parameters and compare with Clogg and Eliason (1987) approach.

5.4.3 Example

We wish to examine cases where the sample is obtained through stratified sampling and the strata are not included as a variable in the contingency table. Our method is to assign the random effects to strata and then make inference as previously. Suppose *age* is the stratification variable and a number of units is sampled from each stratum. Then, we cross-classify the sampled units according to *health status*, *marital status* and *sex*. The target is to examine the relationships between these three variables, to decide which model is appropriate and finally to predict for non-sampled units. *Health status*, *marital status* and *sex* compose the three-way contingency table while *age* is not one of the classifying variables but is assigned the random effects. Inference is made then through the Multinomial model with random effects as described in the previous Section.

All models shown in Table 5.5 are examined and posterior inference is obtained as in the previous Section. Next, bridge sampling is applied for every model and approximations of the marginal log likelihood are given in Table 5.6 together with the posterior model probabilities. It is evident that Model 1 (independence model) has the highest posterior probability and we wish to examine if posterior inference for population counts is better under this model or under model averaging. The model with the second higher posterior probability is Model 2 which includes the main effects plus the interaction between *health status* and *marital status*.

Table 5.6: Approximated log marginal likelihoods by bridge sampling and posterior model probabilities

Model	log Marginal likelihood	Posterior model probabilities
1	-1752.863	0.7037
2	-1753.596	0.2960
3	-1761.744	0.0001
4	-1760.727	0.0002
5	-1764.480	0.0000
6	-1775.161	0.0000
7	-1777.885	0.0000
8	-1779.671	0.0000
9	-1794.914	0.0000

In order to get estimators for the population counts, we draw new data $\tilde{\boldsymbol{y}}$ from their posterior predictive distribution, a process similar to the one described in Section 4.3.1:

1. Use the draws from the joint posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{u})$ to calculate $\boldsymbol{\eta}_m^t = \boldsymbol{\beta}^t + \boldsymbol{u}_m^t$, for $t = 1, ..., T$ number of draws and $m = 1, ..., M$ strata.

2. Draw from the posterior distribution $f(\boldsymbol{p}_m | bmy_m)$ by calculating

$$p_{ijkm}^t = \frac{\exp(\eta_{ijkm}^t)}{1 + \sum_i \sum_j \sum_k \exp(\eta_{ijkm}^t)}$$

   for $i = 2, ...R$, $j = 2, ..., C$, $k = 2, ..., A$ and

$$p_{111ml}^t = \frac{1}{1 + \sum_i \sum_j \sum_k \exp(\eta_{ijkm}^t)}$$

   $t = 1, ..., T$ obtained draws. Finally, we have T random arrays of $R \times C \times A$ containing the draws for $\boldsymbol{p}_m$ for $m = 1, ..., M$ strata.

3. Obtain the stratum sizes N_m as described in Section 3.2 and calculate the $N_m - n_m$ sizes of non-sampled units for every stratum.

4. Get T simulations of new $\tilde{\boldsymbol{y}}_m^t$ from

$$\tilde{\boldsymbol{y}}_m^t \sim Multinom(\boldsymbol{p}_m^t; N_m - n_m)$$

We then plot the histograms of the posterior predictive distributions for each cell of the contingency table (given in Figures 5.4 and 5.5) under Model 1. True population counts are represented in the plots by the black vertical line. Table 5.7 presents the posterior mean and standard deviation under model averaging. In the following Section we implement classical approaches, evaluate and compare estimators under different models using the mean squared error (MSE).

Table 5.7: Posterior mean and standard deviation for population counts under model averaging

		Posterior means					Posterior st.dev.				
		Health status									
Sex	Marital status	1	2	3	4	5	1	2	3	4	5
1	1	66	104	43	5	1	6.885	10.586	5.112	2.904	1.490
	2	154	190	52	12	6	9.431	10.383	6.541	3.606	2.617
	3	33	33	34	14	4	4.941	5.385	4.816	3.742	1.467
2	1	75	126	42	5	2	7.213	10.747	5.209	2.347	1.591
	2	220	184	70	22	12	11.406	10.450	7.409	4.300	3.212
	3	69	80	52	35	1	6.673	8.393	5.870	4.857	1.141

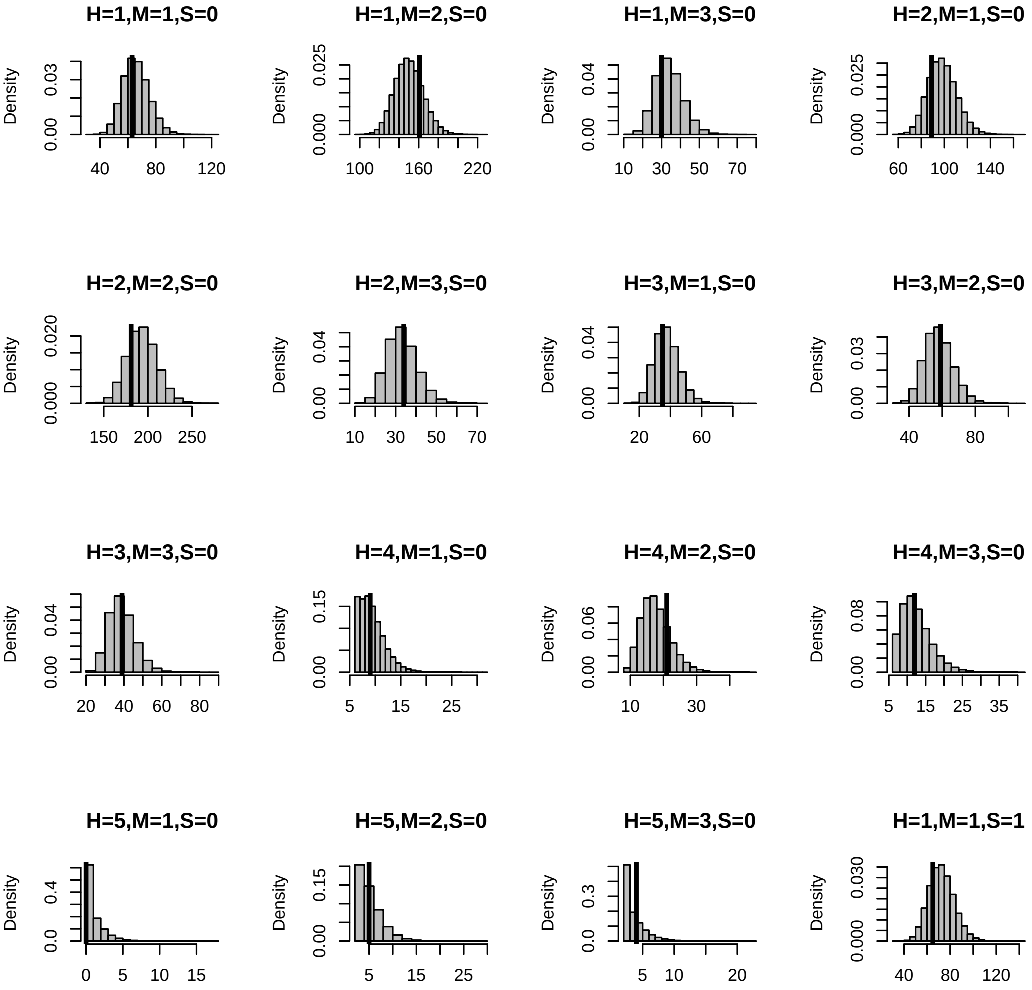

Figure 5.4: a) Distribution of posterior predictive counts for three way contingency tables under the independence model

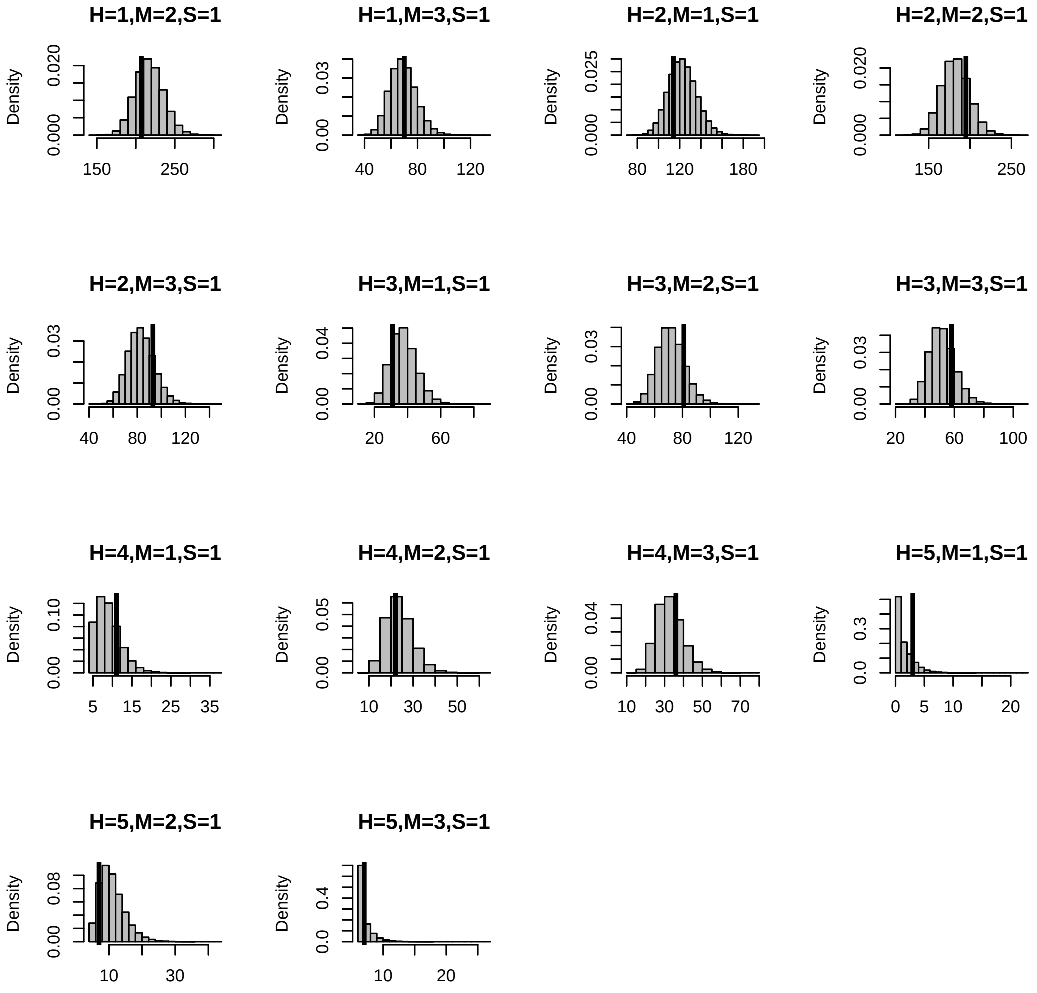

Figure 5.5: b) Distribution of posterior predictive counts for three way contingency tables under the independence model (continued)

5.4.4 Classical analysis

In order to compare Clogg and Eliason (1987) approach with our model for contingency tables, we apply stratification, create the weights and obtain unweighted frequencies. The contingency table of the weighted frequencies is shown in Table 5.8.

Table 5.8: Table of *health status*, *marital status* and *sex weighted* frequencies

		Health status				
Sex	Marital status	1	2	3	4	5
1	1	19	36	12	1	0
	2	50	70	16	4	2
	3	11	9	15	6	2
2	1	22	39	13	1	0
	2	79	62	25	8	5
	3	27	28	21	15	0

Next, we fit all models shown in Table 5.5 using the unweighted data, weighted data and Clogg and Eliason (1987) method. Table 5.9 gives the deviances, degrees of freedom and p-values for all three models and methods. We see that although there are discrepancies between the three different methods, all of them point to Model 5 (model with main effects plus *marital-health status* and *marital status-sex* interactions) as the one that fits better.

In order to provide classical estimates for the population counts in the various categories of health status, social status and sex we use two methods. The first is simply the method described in Section 2.4, where the estimated cell counts are

$$\hat{y}_{ijk} = \sum_{t=1}^{n} w_t I_{ijkt}$$

where

$$I_{ijkt} = \begin{cases} 1 & \text{if unit } t \text{ falls in cell } (ijk) \\ 0 & \text{otherwise} \end{cases}$$

Table 5.9: Deviance and p-value for all models under the three classical methods of analysis

Model		Unweighted		Weighted		CE		df
		Deviance	p-value	Deviance	p-value	Deviance	p-value	
1	$[M][H][S]$	79.512	0.000	79.841	0.000	73.178	0.000	22
2	$[M][H][S][MH]$	22.205	0.074	22.861	0.063	20.582	0.113	14
3	$[M][H][S][MS]$	71.882	0.000	71.852	0.000	65.309	0.000	20
4	$[M][H][S][HS]$	75.204	0.000	74.492	0.000	68.231	0.000	18
5	$[M][H][S][MH][MS]$	14.576	0.265	14.873	0.248	12.284	0.423	12
6	$[M][H][S][MH][HS]$	17.898	0.057	17.513	0.064	15.410	0.118	10
7	$[M][H][S][MS][HS]$	67.575	0.000	66.503	0.000	60.267	0.000	16
8	$[M][H][S][MH][MS][HS]$	11.772	0.162	11.259	0.187	8.843	0.356	8

and which gives the following table of estimated counts.

Table 5.10: Estimated cell counts using sampling weights

		Health status				
Sex	Marital status	1	2	3	4	5
1	1	56	104	35	3	0
	2	145	203	47	13	6
	3	32	27	43	18	6
2	1	64	113	38	2	0
	2	230	179	72	25	16
	3	78	83	60	44	0

Finally, in order to compare our model selection through bridge sampling approximation of the marginal likelihood with the classical approaches, we calculate posterior means of the population counts under Model 5. Model 5 is the model including the main effects plus the health-marital and marital-sex interactions and chosen by all classical methods.

Next, we compare the different estimates we obtained through different models and model averaging as described in Section 1.2.4. Mean squared error is used for this purpose and the results are given in Table 5.11, where we observe that the independence model (Model 1) has the least MSE. This was also the model with the highest posterior probability. Model selection with all three classical approaches under stratified sampling seems to fail, as Model 5 has the highest MSE. Since Bayesian model selection suggested the independence model as the most suitable model, we can conclude that *health status*, *marital status* and *sex* are independent given the *age* stratum.

Table 5.11: MSE under different methods of estimation

Model	MSE
$[H][M][S]$	2580
$[H][M][S][HM][MS]$	3491
Model averaging	2838
Weighted estimates	2746

5.5 Discussion

In this Chapter, we implemented the Multinomial model for contingency tables under cluster sampling and stratified sampling. The underlying problem when analysing contingency tables coming from surveys is the effect of the sampling design in inference. We try to address this problem with the use of random effects corresponding to the design variables.

In Section 5.3, we examined a two way contingency table under cluster sampling. In classical tests, cluster sampling might present associations when they are not actually present and give high X^2 values and small p-values. We see this happening when analysing the sample in Section 5.3 with classical methods, where X^2 test rejects the hypothesis of independence. Using the Bayesian Multinomial model, we fit the independence model and interaction model, both including random effects to account for the cluster effect. We find that the independence model is the dominant model, hence we conclude that *health status* is independent of *marital status* given the *district* effect.

In Section 5.4.1 we discussed three way contingency tables analysis under stratified sampling. In this case, a stratified sample is taken and the stratification variable *age* is not included as one of the variables in the table as this would yield a high dimensional table of 180 cells. To incorporate the stratification variable in analysis, we assign random effects to *age* strata. Next, we implement all models that the contingency table produces and compare them by calculating the posterior model probabilities. The model with the highest posterior probability is the independence model. Sampling weights are also calculated to be used when calculating weighted estimators and when applying the Clogg and Eliason (1987) method. Three different classical approaches are applied for model selection and suggest Model 5, which is the model including the main effects plus the *health-marital* and *marital-sex* interactions. Finally, we obtain posterior means for our model of choice (Model 1), Model 5 and under model averaging. Using the mean squared error, the independence model is the best when it comes to estimating the population counts. We conclude that Bayesian model selection does a bit better than straight weighting and Bayesian model averaging and much better than classical model selection.

Our approach offers a unified methodology to modelling contingency tables, comparing between models, predicting for non-sampled units and finally obtaining efficient estimators for population counts. We use a Mulinomial generalised linear mixed model, for which we perform model selection and averaging and apply for survey data. Incorporating random effects gives a way to account for the effect of the design variables and can be used under any sampling scheme. Various examples are presented in the following Chapter.

Chapter 6

Examples

In this Chapter we examine the relations between various variables of our dataset under stratification and post-stratification by performing two and three way contingency table analysis as described in Chapter 5. Our estimators under the model of choice or under Bayesian model averaging are then compared to classical weighted estimates and estimates under the model chosen by Clogg and Eliason (1987) method.

6.1 Example 1

In the first example, *health region* represents the stratification variable and a sample is taken from each stratum. Then, selected units are cross- classified according to *health status*, *smoker* and *illness*. We want to examine the associations between these three variables and to select the appropriate model under stratified sampling. The Multinomial model with random effects corresponding to *health regions* is applied and draws from the posterior distribution of the parameters are obtained as described in Section 5.4.1. Next, we approximate the marginal likelihood for all suggested models using bridge sampling and we calculate the posterior model probabilities. Clogg and Eliason (1987) approach is used as the frequentist approach to choose between models and all results are given in Table 6.1. There, we see that Bayesian model selection suggests Model 4 (model including the main effects plus *health status-illness* interaction) as the model with the highest posterior probability. With the Clogg and Eliason (1987) approach, Model 6 is preferred which is the model including the *health status-illness* and *health status-smoker* interactions. Both models perform well as far as the MSE is concerned, but Model 4 has the lowest MSE that suggests it predicts best the population counts in each cell. Also, the weighted estimates do not perform well in this example.

Table 6.1: Numerical results for Example 1

	Model	log-Marg.	Post.model prob.	p-values	MSE
1	$[H][S][I]$	-1076.441	0.000	0.000	4910
2	$[H][S][I][HS]$	-1082.171	0.000	0.000	5001
3	$[H][S][I][SI]$	-1077.251	0.0000	0.000	5026
4	$[H][S][I][HI]$	-1056.587	**0.9929**	0.189	**3379**
5	$[H][S][I][HS][SI]$	-1082.582	0.0000	0.000	5084
6	$[H][S][I][HS][HI]$	-1067.788	0.0000	**0.865**	3689
7	$[H][S][I][SI][HI]$	-1061.523	0.0071	0.136	4190
8	$[H][S][I][HS][SI][HI]$	-1071.536	0.0000	0.933	4151
9	$[H][S][I][HS][SI][HI][SHI]$	-1075.194	0.0000	0.000	4101
	Weighted estimates	-	-	-	4496

6.2 Example 2

This example involves post-stratification together with stratification. A stratified sample is taken with *marital status* as the stratification variable and then post-stratification is performed on *social status*. New post-strata are created for all the combinations of the levels of *marital status* and *social* *status*. Next, two way contingency tables within each new post-stratum are created according *health status* and *exercise*. We wish to examine the relation between *health status* and *exercise* in every post-stratum. The Multinomial model with random effects assigned to each post-stratum is applied again and posterior model probabilities are obtained too.

Table 6.2: Approximated log marginal likelihoods and posterior probabilities for Example 2

Model	log Marginal likelihood	posterior model probabilities
Independence	-697.718	0.4314
Interactions	-697.441	0.5686

The alternative models have posterior model probabilities very close to each other and model averaging seems appropriate in this case. We can obtain estimates of the population counts in each cell based not in a single model, since it seems uncertain which one to choose, but averaging over both models. We see that, obtaining estimates under model averaging produces more accurate estimates than any single model but not than weighted estimators, as shown in Table 6.3.

Table 6.3: MSE for Example 2

Model	MSE
Independence	11289
Interactions	11040
Model averaging	9976
Weighted estimates	9396

6.3 Example 3

This last example involves stratification according to *social status* and then cross-classification of the sampled units according to *health status*, *lifestyle* and *alcohol*. The Multinomial model with random effects corresponding to *social statuses* is fitted again. Similar methodology is applied in order to get approximations of the marginal likelihoods, posterior model probabilities and estimates for the population counts in each cell of the table. Finally, the mean squared error is again calculated for all models, model averaging and weighted estimates. In this Example, we note that model averaging provides estimates with the least MSE. The model with the highest posterior probability is the model with the main effects and the interaction between *lifestyle* and *health status*. The model that is chosen through classical analysis is the one with *lifestyle-health status* and *lifestyle-alcohol* interactions. Again, Bayesian model selection method provides a model that performs better than the one suggested by classical model selection. Moreover, Bayesian model averaging gives more accurate estimators than any single model and weighted estimators.

Table 6.4: Numerical results for Example 3

	Model	log-Marg.	Post.model prob.	p-values	MSE
1	$[H][L][A]$	-854.087	0.0000	0.000	5101
2	$[H][L][A][LH]$	-842.892	**0.6671**	0.474	**5050**
3	$[H][L][A][LA]$	-846.084	0.0274	0.001	5179
4	$[H][L][A][HA]$	-849.434	0.0010	0.000	5093
5	$[H][L][A][LH][LA]$	-843.676	0.3044	**0.902**	**5123**
6	$[H][L][A][LH][HA]$	-851.644	0.0001	0.516	5608
7	$[H][L][A][HA][LA]$	-854.855	0.0000	0.000	5363
8	$[H][L][A][LH][LA][HA]$	-853.166	0.0000	0.953	5345
9	$[H][L][A][LH][LA][HA][LHA]$	-857.613	0.0000	0.000	5795
	Weighted estimates	-	-	-	5823
	Model averaging	-	-	-	**5047**

Chapter 7

Discussion

In this thesis we discussed and developed Bayesian methodology for finite population categorical responses under limited information on the design variables. We addressed the problem of not knowing the design variables in stratification, post-stratification and cluster sampling.

In Chapter 3 we suggested ways of dealing with the unknown design variables under stratification and cluster sampling. In the first case, strata sizes can be calculated straightforwardly using the sampling weights. Cluster sampling is more complicated since it requires predicting for non-sampled clusters. Therefore, we proposed two different models, one for simple random cluster sampling and the other for cluster sampling with probability proportional to size. We also compared our model for probability proportional to size cluster sampling with the existing non-parametric model proposed by Little and Zheng (2007).

In Chapter 4 we dealt with the main interest of this thesis, estimation of population counts for univariate and multivariate categorical variables. We applied the Multinomial model with random effects to account for the effect of the design variable in a categorical response with five categories (health status). The model was used in three different sampling designs, stratified sampling, cluster sampling with SRS and cluster sampling with PPS. For all the above designs, posterior means of the population counts were more accurate than the classical estimators in real data applications.

In Chapter 5 we extended the model of Chapter 4 to both two and three way contingency tables. We also addressed the problem of choosing between alternative models for contingency tables by calculating their posterior model probabilities. This was done through approximation of their marginal likelihoods using bridge sampling. In cases where averaging over all plausible models may offer improvement in prediction, we applied this method to get better estimates of population counts in the contingency table cells. Finally, we compare with the approach of Clogg and Eliason (1987) which is a frequentist approach to account for the design effect on contingency table inference.

We conclude that our methodology provides a unified approach for categorical responses from finite populations. It takes into account survey design and provides a method for prediction that classical approaches do not. Our motivation was the lack of this methodology for categorical survey data and the challenge of assuming unknown design variables for non-sampled cases. Moreover, our approach incorporates model comparison or averaging if appropriate that gives more efficient estimates than classical methods.

The work done in Chapter 5 can be extended to higher dimension contingency tables and to a larger number of design variables. Naturally, this makes analysis more difficult, MCMC slower to converge and may produce a model which is hard to interpret. Moreover, modelling the design variable sizes becomes troublesome when the weighting scheme is complex. In order to model the size variable in these cases and be able to predict for the non-sampled groups, we need to obtain first the population sizes for the sampled groups. Gelman (2007) mentions that demographics from previous surveys can be used in these cases or iterative proportional fitting, see Deming and Stephan (1940). This is an issue to be adressed in the future.

Another potential problem for future work is the role of sampling weights in the modelling procedure. As mentioned previously this is a controversial issue. So far, we assign random effects to the design variables used in the weighting procedure and this helps to account for their effect. We also used the weights to calculate group sizes, since weights were the inverse of selection probabilities in the designs we examined. However, in more complex surveys weights are not equal to inverse probabilities of selection. They are constructed by multiplying a series of factors that depend on the design variables and the sampling mechanism. Incorporating them in the modelling process has always been a challenge.

To visualise a model for survey outcomes we provide the following graphical model,

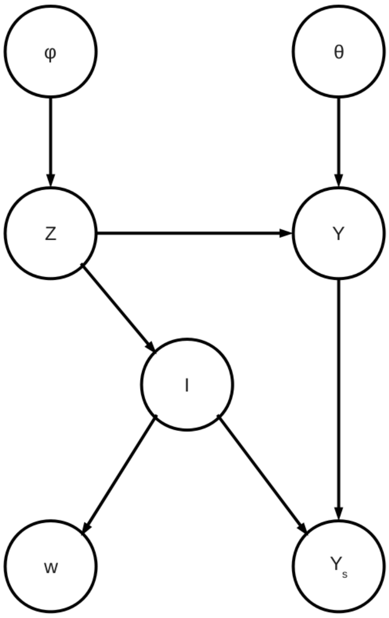

Figure 7.1: Graphical representation for survey models

where

- $\boldsymbol{Y}$ is the survey variable of interest

- $\boldsymbol{Y}_s$ are the sampled part of $\boldsymbol{Y}$
- $\boldsymbol{Z}$ is the set of the design variables
- $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are the parameters for $\boldsymbol{Y}$ and $\boldsymbol{Z}$ respectively
- $\boldsymbol{I}$ the inclusion indicators
- $\boldsymbol{w}$ represents the sampling weights

This graph shows that $\boldsymbol{Y}$ is independent of $\boldsymbol{I}$ given the design variables $\boldsymbol{Z}$ and that the sampling weights $\boldsymbol{w}$ are a product of the design variables and the sampling mechanism. Therefore, the sampling weights can be thought of as "surrogates" of the design variables (Pfeffermann, 1993). Their advantage is that they provide information in a more compact way, since $\boldsymbol{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_N)$ is a set of design variables and $\boldsymbol{w} = (w_1, ..., w_n)$ is just a vector of sampling weights. Thus, sampling weights can be used as a summary of the design variables, especially if modelling $\boldsymbol{Y}$ given $\boldsymbol{Z}$ is too complicated. Also, Rubin (1985) proposes to use the inclusion probabilities (inverse of sampling weights) to replace the design variables but this method requires knowledge of the inclusion probabilites for all the population units. One can note that the weights are only available for the sampled units and cannot be obtained for the non-sampled ones if the design variables are not known for them.

Assuming design variables and hence weights unknown for non-sampled units seems to make modelling impossible. One solution could be to model the weights, then predict for the non-sampled units and finally, model $\boldsymbol{Y}$ conditioning on $\boldsymbol{w}$. This is another challenging problem we wish to investigate in the future.

In general the Multinomial model with random effects can be used for many applications, such as market surveys and transportation modelling where Multinomial modelling is quite popular (Washington et al., 2009). Different types of covariate can be added in the model, such as subject specific, group specific or choice specific covariates. Evidently, our approach seems to produce estimators closer to the true ones than the classical estimators, especially for small samples with low cell frequencies and despite its computational effort.

Bibliography

A. Agresti. *Categorical data analysis, second edition.* John Wiley & Sons, 2002.

F. Jay Breidt and Jean D. Opsomer. Comment on "Struggles with survey weighting and regression modelling" by A. Gelman. *Statistical Science*, 22 (2):168–170, 2007.

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421): 9–25, 1993.

K. R. W. Brewer. A simple procedure for sampling πpswor. *Australian Journal of Statistics*, 17(3):166–172, 1975.

SP. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1997.

W. Browne, F. Steele, M. Golalizadeh, and M. Green. The use of reparameterizations to improve the efficiency of MCMC estimation for multilevel models with applications to discrete time survival models. *Journal of Royal Statistical Society, Series A*, 172:579–598, 2009.

C. C. Clogg and S. R. Eliason. Some common problems in log-linear analysis. *Sociological Methods and Research*, 16:8–44, 1987.

W. E. Deming and F.F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11(4):427–444, 1940.

Economic and Social Data Service. `http://www.esds.ac.uk`.

E. K. Foreman. *Survey sampling principles*, volume 120. Statistics, textbooks and monographs, 1991.

J. J. Forster and A. O'Hagan. *Kendall's Advanced Theory of Statistics. Volume 2B:Bayesian Inference.* Arnold, member of the Hodder Headline Group, London, 2004.

A. E. Gelfand, S. K. Sahu, and B. P. Carlin. Efficient parameterizations for generalized linear mixed models. *Bayesian Statistics 5*, pages 165–180, 1996.

A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper. *Bayesian analysis*, 1(3):515–533, 2006.

A. Gelman. Struggles with survey weighting and regression modelling. *Statistical Science*, 22(2):153–164, 2007.

A Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.

A. Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis.* CHAPMAN & HALL/CRC, Texts in Statistical Science series, 2004.

A. Gelman, A. Jakulin, M. G. Pittau, and Y. S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

M. Hanif and K. R. W. Brewer. Sampling with unequal probabilities without replacement:A review. *International Statistical Review*, 48:317–335, 1980.

J. Hartzel, A. Agresti, and B. Caffo. Multinomial logit random effects model. *Statistical Modelling*, 1:81–102, 2001.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

NHS Health Scotland. `http://www.healthscotland.com/scotlands-health`.

D. Hedeker. A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, 22:1433–1446, 2003.

J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian model averaging:a tutorial. *Statistical Science*, 14(4):382–417, 1999.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

L. N. Kazembe and J. J. Namangale. A Bayesian multinomial model to analyse spatial patterns of childhood co-morbidity in Malawi. *European Journal of Epidemiology*, 22:545–556, 2007.

R. Little. To model or not to model? Competing models of inference for finite population sampling. *Journal of the American Statistical Association*, 99 (466):546–556, 2004.

R. Little and T. Raghunathan. The Bayesian approach to finite population sampling. Short course in Conference:Sample surveys and Bayesian Statistics, Southampton,UK, 2008.

R. Little and H. Zheng. Penalized spline model-based estimation of the finite population total from probability proportional to size samples. *Journal of Official Statistics*, 19:99–107, 2003.

R. Little and H. Zheng. The Bayesian approach to the analysis of finite population surveys. *Bayesian Statistics 8*, pages 283–302, 2007.

S. L. Lohr. *Sampling:Design and Analysis.* Duxbury Press, 1999.

Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park. *MCMCpack: Markov chain Monte Carlo (MCMC) Package*, 2010. URL `http://CRAN.R-project.org/package=MCMCpack`. R package version 1.0-8.

P McCullach and J. Nelder. *Generalized linear models.* Chapman and Hall, 2nd ed., 1989.

X. Meng and S. Shilling. Warp Bridge sampling. *Journal of Computational and Graphical Statistics*, 11:552–586, 2002.

X. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machine. *Journal of chemical Physics*, 21:1087–1091, 1953.

I. Mollina, A. Saei, and M.J. Lombardia. Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society, series A*, 170:975–1000, 2007.

A. Overstall. *Default Bayesian model determination for generalised linear mixed models.* PhD thesis, University of Southampton, School of Mathematics, 2009.

A. M. Overstall and J.J. Forster. Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics and Data Analysis*, 54(12):3269–3288, 2010.

D. Pfeffermann. The role of sampling weights when modelling survey data. *International Statistical Review*, 61:317–337, 1993.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

J.N.K. Rao and A.J. Scott. The analysis of categorical data from complex sample surveys:Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76(374): 221–230, 1981.

D. B. Rubin. Comment on "An evaluation of model-dependant and probability-sampling inferences in sample surveys" by M.H. Hansen, W.G. Madow, and B.J. Tepping. *Journal of the American Statistical Association*, 78(384):803–805, 1983.

D.B. Rubin. The use of prospensity scores in applied Bayesian inference. *Bayesian Statistics 2*, pages 463–472, 1985.

C. Skinner and L. A. Vallet. Fitting Log-linear models to contingency tables from surveys with complex sampling designs:an investigation of the Clogg-Eliason approach. *Sociological methods and Research*, 2010.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64:583–639, 2002.

Yves Till and Alina Matei. *sampling: Survey Sampling*, 2009. URL `http://CRAN.R-project.org/package=sampling`. R package version 2.3.

S. Washington, P. Congdon, M.G. Karlaftis, and F.L. Mannering. Bayesian multinomial logit: theory and route choice example. *Trasportation research record: Journal of the Transportation Research Board*, 2136(4):28–36, 2009.