Bayesian analysis of hierarchical models for polychotomous data from a multi-stage cluster sample

by

Michael Edward Schuckers

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Major Professor: Hal S. Stern

Iowa State University
Ames, Iowa
1999

UMI Number: 9940238

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of

Michael Edward Schuckers

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College

For Dan and Sara

and

For Stephanie

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

First and foremost, I must thank Dean Isaacson, who took a chance on me when few others would have.

Thanks to Hal Stern, my advisor and mentor. For never letting me slide by and for reminding how far I really have to go. Hal put up with more than an advisor should have to and still managed to keep a sense of direction as well as a sense of humor. Many thanks.

Thanks to my committee members: Professors Breidt, Bushman, Kaiser, Koehler and Opsomer. Their suggestions and comments were always helpful and insightful. Additionally, Professor Vardeman time and again proved to be an invaluable source of wisdom.

Thanks to my old family: My old grey mère for encouraging and supporting me along the way. And for never failing to be impressed. My father for his stochastic outlook on life. The larch. I don't know how many times I've heard, "If these two teams played 10 times, team A would win 7." It definitely took its toll. The larch. My sister Lisa — and now her family, Marty and Kara and Kate — who always remind me of what really matters. Lisa, you wear me out. My sister Carly, who makes us all laugh. Most of the time, it's with her. You are special and I love you, ***sugar-pie honey-bunch chickadee.*** The larch.

A uncountably large set of thank you's to John Gabrosek and William Christensen, To John Gabrosek, for his sense of humor and his friendship during many long nights and many early mornings. You have been a welcome companion on this journey. To my

squash partner William, for his energy and for his will I greatly admire you my friend. You are one of the best people I know. To Deanne "I'm too sexy for my dissertation" Reber and Pam "I am SO sick of soils" Abbitt, another set of special thank you's. Along with John and William, we shared many late nights, many early Saturday mornings and more m&m's than any of us care to admit. I couldn't have done it without any of you. Also thanks to Kari Azevedo, Ian Carlson, Jenny Schumi and Barb Mock and lots more who made the trip more bearable. Non-Iowa Stat-ers people that I need to thank include Aliki Compos, Christopher Koeppen, Scott Saul and Raymond Sedwick. Thanks for all your love and support.

Finally, to my beautiful wife, Stephanie. You have made so many days, halycon days. Everyday is brighter knowing that you are a part of my life. I am so lucky. I love you very much. Now let's go down to the jubilee...

CHAPTER 1 INTRODUCTION

1.1 Sampling

Statistical inference is concerned with drawing conclusions about a population based on data collected from a subset of the population. The techniques that are used to obtain the subset from which data are collected fall under the general heading sampling. Sampling plays an integral part in modern life. Public policy is often swayed or altered based upon the opinions of the U. S. population. Samples are undertaken because a census of every element in the population is a costly and time-consuming endeavor. The information in the sample, a subset of the population, is then used to make decisions about the population.

There are several classic texts on sampling; these include Kish (1965) and Cochran (1977). The remainder of this section is a brief review of key concepts of sampling. To describe simple random sampling, we consider selecting a sample of size n from a population of size N. Under simple random sampling each subset of size n has the same probability of being selected. The resulting sample is known as a simple random sample (SRS). Simple random sampling is also known as random sampling without replacement (WOR), since once an element of the population is selected it is removed from further consideration. Random sampling with replacement (WR) is possible. In that case a single element is chosen from the population with all elements having equal probability of being selected. The information from that element is recorded and that element is returned to the list of possible elements for the next sample. This allows for the possibility that an element could appear in the sample more than once. As an example, consider an urn of balls of various colors where the color of the balls is the variable of interest. A SRS-WR is obtained by repeatedly selecting one ball out of the urn noting the color and returning the ball to the urn. A SRS is obtained by obtained by selecting one ball out of the urn, noting the color and setting the ball aside before drawing another ball from the urn. Sampling without replacement is a much more natural idea. Sampling with replacement leads to simpler formulae for computing variance estimates and consequently is occasionally used in more complex designs. We will use SRS to denote a sample obtained using a traditional simple random sample (without replacement). We use SRS-WR to denote a sample obtained by simple random sampling with replacement.

There are many other sampling techniques besides simple random sampling that are used in practice for making inference. Examples include stratified random sampling, systematic sampling and cluster sampling. This dissertation is concerned with cluster sampling. To obtain a cluster sample, one first divides the population into subsets, known as clusters. Then a simple random sample of clusters is obtained. A census of the elements in the chosen clusters is then carried out. The advantage of this methodology is that is often saves time and money. This is especially true when the population covers a broad geographic region. It is then cheaper to take samples of congregated elements than it is to take a SRS. The following is an example of a cluster sample. Suppose that we are interested in estimating the salaries of orchestra performers from the main orchestra's in each of the 50 largest cities in the U. S. One method of cluster sampling would be to carry out a SRS-WOR of 10 of the 50 cities. The sample would then be composed of all the main orchestra performers in those 10 selected cities. The information collected from these individuals would then be used to make information about all orchestra performers in the 50 largest cities.

In this thesis, we will focus on a particular type of cluster sampling, the multi-stage cluster sample. The first-stage proceeds, as above, a SRS of clusters is selected from the population of clusters. Following that, at the second stage, a SRS of individuals (or perhaps subclusters) is selected from the population of individuals (or subclusters) in the cluster. In large complicated studies there can be several stages. Each level partitions the population into more manageable pieces. At the final stage, we obtain a SRS of the elements of interest from the most recent subclusters. As an example, suppose we take a cluster sample of individuals from the U.S. by treating the states as clusters and the counties within each state as subclusters. As the final stage, we select a SRS of the individuals within the selected counties. Since there are three levels of sampling in this design— states, counties, and individuals — this is referred to as a three-stage cluster sample. Särndal et al. (1992) refer to this methodology as three-stage element sampling, since the final stage involves the sampling of individuals in the population, instead of clusters. Again the reasons for using such a sampling design are savings in time and money. All of the methods in this thesis are developed for two-stage and three-stage cluster samples.

Sampling is carried out on many different types of populations including people, cereal boxes, agricultural land or income tax returns. Typically, many different types of data are collected on the elements of a sample. These types include continuous variables, such as the percentages of clay in soil or daily household alcohol consumption; as well as discrete variables such as a yes-no response for approval of a government initiative or a categorization of federal land based on its usage. The focus in this work will be on a particular type of categorical variable, the polychotomous response, sometimes known as the polytomous response. Here, we take polychotomous to mean divided into more than two groups or classes. A polychotomous variable has more than two possible responses and the order of those responses is irrelevant. An example of this would be the favorite professional soccer team of a respondent. The variable is clearly a category and there is no natural ordering to the possible responses.

1.2 Inference for Finite Populations

Making inference about a finite population is the ultimate goal of nearly every sampling exercise. Consequently a great deal of work has been done on inferential techniques samples from a finite population. These techniques tend to fall into two categories. The first approach is known as design-based inference. This non-parametric methodology uses weights (numbers assigned to the selected elements to give certain elements more elements than others) and selection probabilities to estimate quantities of interest, such as population means, totals and proportions and their variances. Cochran (1977) and Kish (1965) are two classic references describing design-based methods. The second category of analysis techniques is known as model-based inference. This methodology assumes a model for the population, often called a superpopulation model. The superpopulation is an infinite population from which the finite population is assumed to have been sampled. The parameters of the model represent quantities of interest for the infinite population. The model parameters are estimated using the sampled data and the resulting estimates are used for inference about the entire population (Thomsen and Tesfu (1988)). Both Bayesian and non-Bayesian analyses have been undertaken for model-based procedures, see Ericson (1988) and Royall and Cumberland (1978). In this dissertation, we will use a Bayesian model-based approach to analyze the data collected from a multi-stage cluster survey. An excellent comparison of design-based and model-based methods is given by Särndal (1978).

1.3 Missing Data

Missing data is a common problem in statistical analyses of data. Sampling techniques intentionally create missing data by not sampling all of the elements of the population. Since not all variables are observed for each member of the population, the unobserved values can be thought of as missing data. Naturally, because these values are missing by design they do not represent a serious problem. Design-based or model-based inferential techniques allow us to assess the uncertainty introduced by these intentionally missing data. What can be more problematic is that often data for members of the sample are not recorded. Individuals may refuse to respond to a particular question, or a sampled individual may decline to participate. Such "unintentional" missing data may mean that the observed data is no longer representative of the population from which it was selected.

There are many possible circumstances that could lead to unrecorded data. Statisticians have developed methods to handle various types of unintentional missing data. In particular, Little and Rubin (1987) describe three categories for mechanisms that lead to unintentional missing data. The missing data mechanism is missing completely at random (MCAR) if the probability that a values is missing is independent of the observed as well as the unobserved responses. The missing data mechanism is called missing at random (MAR) if the probability that the response is missing does not depend on the unobserved value though it may depend on other observed covariates. Under fairly general conditions, MAR and MCAR mechanisms are ignorable, in the sense that valid inferences can be carried out using only observed responses.

Missing data mechanisms for which the probability of a missing response depends on the value that would have been observed are referred to as non-ignorable. In the non-ignorable case, additional information, perhaps, in the form of modeling assumptions, is needed to draw valid inferences. In this thesis we show how information from unintentional missing observations can be incorporated into the data analyses assuming the missing data mechanism is ignorable.

1.4 Slovenian Public Opinion Survey

The motivation for the methods developed in this thesis is the 1990 Slovenian Public Opinion Survey (SPO). The SPO is a three-stage cluster survey that is conducted every year or every other year. It is a general opinion survey carried out to gauge public opinion on a variety of issues. In 1990 the people of Slovenia were preparing for a vote on independence from Yugoslavia. Included in the SPO, along with the usual demographic and attitudinal questions, that year were three questions concerning the upcoming plebiscite.

1. Are you in favor of Slovenian independence?

2. Are you in favor of Slovenia's secession from Yugoslavia?

3. Will you attend the plebiscite?

For the plebiscite on independence, the outcome would be determined not by the percentage of actual voters that voted for independence, but by the percentage of eligible voters that voted for independence. Thus not attending the plebiscite was implicitly a vote against independence from Yugoslavia. In this thesis we will focus on analyzing these three questions concerning independence.

1.5 Thesis Outline

In Chapter 2 we present a hierarchical model for analyzing polychotomous data from a two-stage cluster sample. We use a Bayesian approach to carrying out these analyses. The observed data in each cluster are modeled as coming from a multinomial model. Then the parameters of the multinomial model are modeled as a sample from a Dirichlet distribution. We develop methodologies for analyzing the data when the observations are fully observed and when some of the observations are only partially observed (in a sense that is made clear in Section 2.5). We extend that model to handle three-stage cluster samples in Chapter 3. In both of these chapters we use the proposed methods to analyze the 1990 Slovenian Public Opinion Survey. Chapter 3 also uses simulations to demonstrate that the hierarchical model can be used in a variety of scenarios. In Chapter 4 we turn our attention to making use of some of the other variables that were measured as part of the SPO to improve prediction. With complete data on the questions of interest, there is little reason to worry about other responses. However when there are unintentional missing data, the other variables can be used to improve inference about the missing responses. Finally, Chapter 5 provides conclusions and a discussion of future work.

CHAPTER 2 TWO-STAGE MODEL

2.1 Introduction

In this chapter we consider polychotomous data from a two-stage cluster sample. We use the Bayesian approach to analyze data under a hierarchical superpopulation model. Section 2.2 reviews the relevant literature on design-based and superpopulation inference for two-stage cluster samples. Section 2.3 introduces notation for two-stage survey data and specifies the hierarchical model. As the model includes an improper prior distribution for some parameters we also describe conditions required to obtain a proper posterior distribution. Section 2.4 reviews the Bayesian approach to inference and provides computational details for our approach to sampling from the posterior distribution. By design, sample surveys create intentionally missing data. Occasionally there are also unintentional missing values. Section 2.5 shows how unintentional missing data can be accommodated in the model. Finally, Section 2.6 applies the models of this chapter to the 1990 Slovenian Public Opinion Survey.

2.2 Literature Review

The traditional design-based approach to analyzing cluster samples is described, for example, by Cochran (1977) and Kish (1965). There sample proportions or sample means are used to estimate population quantities and estimated standard errors are derived that account for the correlations within clusters. Specifically, the design effect gives the ratio of the variance for the cluster sample of a given total size to that of a simple random sample (SRS) of the same size. Cochran gives results for both continuous and binary responses. Ghosh and Meeden (1997) presented what could be considered a design-based Bayesian approach. Their methodology is to use a non-parametric approach and a non-informative prior to making inference from a population.

Several frequentist approaches have been outlined for analyzing polychotomous data from a cluster sample using a model-based approach. Brier (1980) analyzes data of this type using a Multinomial-Dirichlet model where survey responses in a cluster are assumed to have a multinomial distribution and the multinomial parameters for the clusters are modeled as draws from a Dirichlet distribution. We consider a Bayesian analysis of the same model. Brier derives a method of moments estimator for the cluster effect using a log-linear model. Rao and Scott (1981) develop a methodology for estimating the effect of several complex sample designs, including cluster sampling, on χ^2 statistics for testing hypotheses about a vector of proportions. Additionally they develop several estimators for the design effect under a Dirichlet-Multinomial model. Wilson and Koehler, in a series of papers — Wilson (1984), Wilson (1986), Koehler and Wilson (1986) and Wilson (1987) — present techniques for estimating the design effect using a regression estimator, for comparing vectors of proportions taken from independent two-stage cluster samples, and for modeling multinomial data with extra variation via a Dirichlet-Multinomial. More recently, Morel and Nagaraj (1993) use a finite mixture of multinomial random variables to model polychotomous data that exhibit levels of variation that are higher than would be expected under the a multinomial model. Morel and Koehler (1995) model both underdispersion and overdispersion in categorical data ; they develop a "sandwich" estimator of the variance-covariance matrix that corrects for extreme levels of variability.

Hierarchical Bayesian models have been used to analyze data from cluster samples for some time. Scott and Smith (1969) describe a hierarchical Bayesian approach to

Gaussian data collected from a two-stage cluster sample. They treat the observations as coming from a superpopulation as we do here. Malec and Sedransk (1985) extend the methodology of Scott and Smith to multi-stage cluster sampling when the data is Gaussian. Nandram and Sedransk (1993b) develop a model for longitudinal surveys when the observed data is from a Gaussian distribution or is transformable to a Gaussian distribution. Some recent work has focused on developing models for analyzing data from non-Gaussian distributions. Nandram and Sedransk (1993a) develop a hierarchical Bayesian model for binary data from a two-stage cluster sample. The model they used is a Beta-binomial model where the binary counts in a cluster a modeled as coming from a binomial distribution and the cluster probabilities are modeled as draws from a beta distribution. Stroud (1991) analyzed binary data from a survey sample using a Beta-binomial model, while Stroud (1994) utilized a logit transformation to analyze the same data set. Another example of the use of a hierarchical analysis for binary data is Stasny (1991). In this article Stasny presented an empirical Bayes approach for analyzing binary data with a hierarchical model that also models non-response probabilities as well as the probability of interest. Nandram (1998) described a Bayesian hierarchical model for multinomial data from a two-stage cluster sample. His model is quite similar to the one considered here; we describe it more later in this chapter. Several authors have used Bayesian generalized linear models to analyze multinomial data. Recently, Ghosh et al. (1998) used a generalized linear model to analyze multinomial data for the problem of small area estimation in sampling.

2.3 Probability Model

We take the total population to be N and assume that the population is divided into M clusters, with N_j individuals in the j^{th} cluster. Suppose that we sample J of the M clusters, and that within the j^{th} cluster we sample n_j of the N_j individuals. We denote the number of unsampled clusters by $J' = M - J$. Our focus in this thesis is polychotomous responses. We let I represent the number of possible responses and let Y_{ij} denote the number of individuals in the j^{th} cluster with response i, $i = 1, 2, \ldots, I$. To pursue a model-based approach to the analysis of a two-stage cluster sample, it seems natural to model the hierarchical structure implicit in the design with a hierarchical model. Subjects are assumed to be randomly sampled within clusters, and clusters are taken to be randomly sampled from the population of clusters.

We propose the following hierarchical superpopulation model. Given the vector $\boldsymbol{\theta}_j = (\theta_{1j}, \ldots, \theta_{Ij})^T$ of probabilities summing to 1 in cluster j, the data $\mathbf{Y}_j$ are modeled as multinomial random variables,

$$\mathbf{Y}_j \mid \boldsymbol{\theta}_j, n_j \sim \text{ Multinomial}(\boldsymbol{\theta}_j, n_j) \tag{2.1}$$

with

$$\text{p}(\mathbf{Y}_j \mid \boldsymbol{\theta}_j, n_j) = \frac{n_j!}{Y_{1j}! \ldots Y_{Ij}!} \prod_{i=1}^{I} \theta_{ij}^{Y_{ij}}, \tag{2.2}$$

for $j = 1, \ldots, J$. We also assume the vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_J$ are independent given the collection of probability vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J$. Usually samples are taken without replacement from the population of interest. Strictly speaking this invalidates the multinomial model, which is appropriate for sampling with replacement. As long as n_j is small compared to N_j the multinomial is a convenient and good approximation. We define $\boldsymbol{\Theta}=(\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_J^T)^T$ as the vector consisting of all the cluster level probabilities concatenated. This part of the model implicitly assumes that individuals within a cluster are exchangeable. That is, no information is recorded to distinguish individuals. If the sample were stratified by education level or occupation, for example, this could easily be accommodated by modifying the above portion of the model.

The cluster level probability vectors, $\boldsymbol{\theta}_j$, are modeled as exchangeable draws from a Dirichlet distribution,

$$\boldsymbol{\theta}_j \mid \boldsymbol{\alpha} \sim \text{ Dirichlet}(\boldsymbol{\alpha}) \tag{2.3}$$

with

$$\mathrm{p}(\boldsymbol{\theta}_j \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum\limits_{k=1}^{I} \alpha_k\right)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_I)} \prod_{i=1}^{I} \theta_{ij}^{\alpha_i - 1}, \quad j = 1, \ldots, J. \tag{2.4}$$

The vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)^T$ are parameters describing the population of $\boldsymbol{\theta}_j$'s with

$$\begin{aligned} \mathrm{E}[\boldsymbol{\theta}_j \mid \boldsymbol{\alpha}] &= \frac{\boldsymbol{\alpha}}{\sum\limits_{k=1}^{I} \alpha_k} \\ \mathrm{Var}[\boldsymbol{\theta}_j \mid \boldsymbol{\alpha}] &= \frac{\boldsymbol{\alpha}(\kappa\mathbf{1} - \boldsymbol{\alpha})}{\kappa^2(\kappa + 1)}, \end{aligned}$$

where $\mathbf{1} = (1, \ldots, 1)^T$ and $\kappa = \sum\limits_{k=1}^{I} \alpha_k$. The assumption of exchangeability here implies no information is available to discern clusters. The Dirichlet distribution is the conjugate prior distribution for the multinomial distribution. This is convenient for computation later. Though it is the conjugate prior distribution, the Dirichlet is somewhat restrictive. It would not be difficult to use a mixture of Dirichlet distributions in its place. Finally the prior distribution of $\boldsymbol{\alpha}$ is taken as an improper distribution.

$$\mathrm{p}(\boldsymbol{\alpha}) \propto \left(\sum_{k=1}^{I} \alpha_k\right)^{-\frac{2I+1}{2}} I_{(\alpha_i > 0, \forall i)}. \tag{2.5}$$

This form of non-informative prior distribution is motivated by results for the normal-normal and especially the Beta-binomial hierarchical models, see e. g. Gelman et al. (1995). There the non-informative hyperprior distribution is flat on the standard deviation of the prior distribution. For the Dirichlet we take the hyperprior distribution to be flat on the mean and flat on $\left(\sum\limits_{k=1}^{I} \alpha_k\right)^{-1/2}$, the latter quantity representing a quantity similar in magnitude to the standard deviation,

$$\mathrm{p}\left(\frac{\alpha_1}{\sum\limits_{k=1}^{I} \alpha_k}, \ldots, \frac{\alpha_{I-1}}{\sum\limits_{k=1}^{I} \alpha_k}, \frac{1}{\sqrt{\sum\limits_{k=1}^{I} \alpha_k}}\right) \propto 1. \tag{2.6}$$

After taking account of the Jacobian of the transformation, we get (2.5) or the prior distribution for $\boldsymbol{\alpha}$. Hyperprior distributions, like (2.5), are often called "diffuse" or "vague" meaning that the density places probability somewhat evenly throughout the parameter space. The advantage of this type of hyperprior distribution is that it allows the data to shape the posterior distribution and, consequently, to shape the inferences that are made. Since this prior distribution is clearly improper, it is necessary to find sufficient conditions that yield a proper posterior distribution. Theorem 1 at the end of this section provides such conditions.

At this point it should be noted that the model developed in (2.2), (2.4) and (2.5) is quite similar to the model of Nandram (1998). Though developed independently, both Nandram's approach the approach described here model the cluster-level counts as draws from a Multinomial distribution. The cluster-level probability vectors are then modeled as draws from a Dirichlet distribution in both methods. The difference between the two models is in the hyperprior distribution that is placed on $\boldsymbol{\alpha}$, the population-level vector or proportions. Nandram's model places a flat prior on the mean of the Dirichlet distribution and a gamma prior distribution on $\kappa = \sum_{k=1}^{I} \alpha_k$. If we take the gamma density or $\mathrm{p}(\kappa \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \kappa^{\alpha-1} e^{-\beta\kappa}$, then the Nandram model is equivalent to the model considered here when the shape parameters α is chosen to be 1/2 and the scale parameter β is taken to be zero.

Let $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_J^T)^T$ represent the observed cluster counts concatenated into a single column vector and let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)^T$ represent the cluster probability vectors concatenated into a single column vector. Then the posterior distribution, up to a normalizing constant, is

$$
\begin{aligned}
\mathrm{p}(\boldsymbol{\Theta}, \boldsymbol{\alpha} \mid \mathbf{Y}) &\propto \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\Theta})\mathrm{p}(\boldsymbol{\Theta} \mid \boldsymbol{\alpha})\mathrm{p}(\boldsymbol{\alpha}) \\
&\propto \left(\sum_{k=1}^{I} \alpha_k\right)^{-(2I+1)/2} \prod_{j=1}^{J} \left\{ \Gamma\left(\sum_{k=1}^{I} \alpha_k\right) \prod_{i=1}^{I} \left\{ \frac{\theta_{ij}^{Y_{ij}+\alpha_i-1}}{\Gamma(\alpha_i)} \right\} \right\}
\end{aligned}
$$

The posterior distribution can be factored as

$$
\mathrm{p}(\boldsymbol{\Theta}, \boldsymbol{\alpha} \mid \mathbf{Y}) = \mathrm{p}(\boldsymbol{\Theta} \mid \boldsymbol{\alpha}, \mathbf{Y})\mathrm{p}(\boldsymbol{\alpha} \mid \mathbf{Y}). \tag{2.7}
$$

The first term, $p(\boldsymbol{\Theta},\boldsymbol{\alpha} \mid \mathbf{Y})$, is obtained easily because of the choice of the conjugate Dirichlet prior. It is

$$p(\boldsymbol{\Theta} \mid \boldsymbol{\alpha},\mathbf{Y}) = \prod_{j=1}^{J} \text{Dirichlet}(\mathbf{Y}_j + \boldsymbol{\alpha}). \tag{2.8}$$

It follows that

$$\begin{aligned} p(\boldsymbol{\alpha} \mid \mathbf{Y}) &= p(\boldsymbol{\Theta},\boldsymbol{\alpha} \mid \mathbf{Y})/p(\boldsymbol{\Theta} \mid \boldsymbol{\alpha},\mathbf{Y}) \\ &\propto \left[\prod_{j=1}^{J}\left\{\frac{\Gamma\left(\sum\limits_{k=1}^{I}\alpha_k\right)}{\Gamma\left(\sum\limits_{k=1}^{I}\alpha_k+n_j\right)}\prod_{i=1}^{I}\frac{\Gamma(Y_{ij}+\alpha_I)}{\Gamma(\alpha_i)}\right\}\right]\left(\sum_{k=1}^{I}\alpha_k\right)^{-\frac{2I+1}{2}}. \end{aligned} \tag{2.9}$$

We can also derive $p(\boldsymbol{\alpha} \mid \mathbf{Y})$ by first calculating the marginal distribution of $\mathbf{Y}$ given $\boldsymbol{\alpha}$,

$$\begin{aligned} p(\mathbf{Y} \mid \boldsymbol{\alpha}) &= \int p(\mathbf{Y} \mid \boldsymbol{\Theta})p(\boldsymbol{\Theta} \mid \boldsymbol{\alpha})d\boldsymbol{\Theta} \\ &= \prod_{j=1}^{J}\left\{\binom{n_j}{Y_{1j},\ldots,Y_{Ij}}\frac{\Gamma\left(\sum\limits_{k=1}^{I}\alpha_k\right)}{\Gamma\left(\sum\limits_{k=1}^{I}\alpha_k+n_j\right)}\prod_{i=1}^{I}\frac{\Gamma(Y_{ij}+\alpha_I)}{\Gamma(\alpha_i)}\right\}, \end{aligned}$$

and then multiplying by the hyperprior on $\boldsymbol{\alpha}$,

$$p(\boldsymbol{\alpha} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\alpha})p(\boldsymbol{\alpha}). \tag{2.10}$$

Each $\mathbf{Y}_j$ in (2.10) is distributed as a Dirichlet-Multinomial with parameters $\boldsymbol{\alpha}$ and n_j. The distribution in (2.10) is a product of Dirichlet-Multinomial distributions.

The Dirichlet-Multinomial distribution plays a large role in the algorithm that we develop for sampling from the posterior distribution. For future reference we note here that we can simulate draws from the Dirichlet-Multinomial distribution with parameters $\boldsymbol{\alpha}$ and m in a straightforward manner. We first draw a realization $\boldsymbol{\theta}$ from a Dirichlet distribution with parameters $\boldsymbol{\alpha}$. We then simulate a vector of counts $\mathbf{Y}$ from a multinomial with parameters $\boldsymbol{\theta}$ and m. The realization $\mathbf{Y}$ is then a draw from the Dirichlet-Multinomial distribution and $\boldsymbol{\theta}$ can be discarded.

We close this section with a theorem identifying conditions required for the posterior distribution (2.9) to be a proper distribution.

Theorem 1 *For the model defined by (2.2), (2.4), (2.5), the posterior distribution is proper if there exists at least one cluster that has responses in at least two different cells.*

Proof: The proof is included as the final section of this chapter, Section 2.7.

2.4 Posterior Inference

Having shown that the posterior distribution is proper under the conditions of Theorem 1, we consider statistical inference for the parameters in the model. From the Bayesian perspective, the posterior distribution $\mathrm{p}(\mathbf{\Theta}, \boldsymbol{\alpha} \mid \mathbf{Y})$ describes the uncertainty in the parameters after observing the data. Recall that the conditional posterior distribution of $\mathbf{\Theta} \mid \boldsymbol{\alpha}, \mathbf{Y}$ is simply a product of Dirichlet distributions, $\boldsymbol{\theta}_j \mid \boldsymbol{\alpha}, \mathbf{Y}_j \sim \text{Dirichlet}(\mathbf{Y}_j + \boldsymbol{\alpha})$. This means that we can obtain simulations from the posterior distribution by first simulating $\boldsymbol{\alpha}$ from $\mathrm{p}(\boldsymbol{\alpha} \mid \mathbf{Y})$ and then simulating $\mathbf{\Theta}$ from $\mathrm{p}(\mathbf{\Theta} \mid \boldsymbol{\alpha}, \mathbf{Y})$. We use a Markov chain Monte Carlo (MCMC) procedure to obtain samples from the posterior distribution of $\boldsymbol{\alpha}$ given the data.

2.4.1 Review of MCMC

The use of MCMC methodology has become quite common in recent years for Bayesian data analysis. In general terms, we simulate from a given distribution by identifying a Markov chain that has the desired distribution as its stationary distribution and then simulating the Markov chain. In the Bayesian context the distribution from which we wish to sample is the posterior distribution. The Markov chain is run until it converges to the stationary distribution. There are a number of MCMC algorithms. These are reviewed for example by Gilks et al. (1996). We describe several of these in this section. The first we consider is the Metropolis algorithm. We describe the algorithm by considering the t^{th} iteration. Suppose that the current state of the Markov chain is $\boldsymbol{\alpha}^{(t-1)}$. The Metropolis algorithm uses a symmetric jumping distribution $\mathrm{J}(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^{(t-1)})$ to generate a candidate state, $\boldsymbol{\alpha}^*$. Define the importance ratio as

$$r = \frac{\mathrm{p}(\boldsymbol{\alpha}^* \mid \mathbf{Y})}{\mathrm{p}(\boldsymbol{\alpha}^{(t-1)} \mid \mathbf{Y})}. \tag{2.11}$$

Then the Metropolis Markov chain transition rule is

$$\boldsymbol{\alpha}^{(t)} = \begin{cases} \boldsymbol{\alpha}^* & \text{with probability } min(r, 1) \\ \boldsymbol{\alpha}^{(t-1)} & \text{otherwise.} \end{cases} \tag{2.12}$$

If the prospective state has higher posterior density than the present state, then the chain moves to the new state. If the prospective state has lower posterior density, then the chain moves to the prospective state with probability equal to the importance ratio. Thus the algorithm is completely specified by giving a starting point $\boldsymbol{\alpha}^{(0)}$ and a symmetric jumping distribution, J($\cdot \mid \cdot$). A standard symmetric jumping distribution is the Gaussian distribution J($\boldsymbol{\alpha}^* \mid \boldsymbol{\alpha}^{(t-1)}$) = N($\boldsymbol{\alpha}^* \mid \boldsymbol{\alpha}^{(t-1)}, V$) where V is a fixed variance matrix. Note that we use $N(\mu, \sigma^2)$ to denote the Gaussian distribution and $N(x \mid \mu, \sigma^2)$ to denote the Gaussian probability density function with argument x. Non-symmetric jumping distributions can also be accommodated. In that case the algorithm is known as the Metropolis-Hastings algorithm (Hastings (1970)), and the importance ratio becomes

$$r = \frac{\mathrm{p}(\boldsymbol{\alpha}^* \mid \mathbf{Y})/J(\boldsymbol{\alpha}^* \mid \boldsymbol{\alpha}^{(t-1)})}{\mathrm{p}(\boldsymbol{\alpha}^{(t-1)} \mid \mathbf{Y})/J(\boldsymbol{\alpha}^{(t-1)} \mid \boldsymbol{\alpha}^*)}. \tag{2.13}$$

Under fairly general conditions on the Markov chain, we are assured that the chain will eventually converge to the distribution from which we wish to sample, see for example, Tierney (1996). However, assessing when convergence has occurred is not simple. There is a growing literature on this topic, e. g. Cowles and Carlin (1996), or Gilks et al. (1996). We apply the multiple chain methodology of Gelman and Rubin (1992) to assess convergence. Their approach utilizes several independent sequences with starting points chosen to be overdispersed relative to the target distribution. The independent chains are run from these starting points for a fixed number of iterations. After some initial iterations — presumed to correspond to transient behavior of the Markov chain

— are deleted, each chain is analyzed. The potential scale reduction (PSR), (Gelman and Rubin (1992)), measures the degree to which posterior inference would improve with repeated simulation. The PSR is the ratio of two estimates of the posterior variance of a parameter. One estimate examines variability between the independent chains. This will overestimate the variability until the chain has converged, as long as the starting points are overdispersed. The second estimate uses variability within each chain. This is likely to underestimate the posterior variance until the chains have sampled the entire posterior distribution. If the PSR is close to unity, for all parameters of interest, then we conclude that the samples can be treated as having come from the target density.

The choice of a jumping distribution is important because it can affect the speed of convergence. Gelman et al. (1996) give guidelines for selecting V, the variance of a normal jumping distribution, when the target distribution is normal or nearly so. They find that acceptance rates, the percentage of time that the proposed candidate is accepted, between 25% and 50% are optimal. Consequently, they suggest the variance matrix, V, be chosen to achieve these rates.

2.4.2 Gibbs sampling

Another MCMC algorithm is the Gibbs algorithm, sometimes simply referred to as Gibbs sampling. It is most often used when the joint distribution is high dimensional or when sampling from the full conditionals is relatively straightforward. Suppose that $p(\Psi \mid \mathbf{Y})$ is a multivariate distribution from which we would like to obtain samples. Also suppose that we can break Ψ into D univariate or at least lower-dimensional components, $\Psi = (\psi_1, \ldots, \psi_D)$. We assume the full conditional posterior distributions $p(\psi_d \mid \psi_1, \ldots, \psi_{d-1}, \psi_{d+1}, \ldots, \psi_D)$ are available. Under mild conditions, these conditional distributions uniquely define the full joint distribution, Besag (1974). For each iteration of the Markov chain we cycle through the conditional posterior distribution generating a realization for each element of Ψ from its conditional posterior distribution using the most recently generated value for each element of Ψ. . This is done as follows for the t^{th} iteration:

1. Simulate $\psi_1^{(t)}$ from $p(\psi_1 \mid \psi_2^{(t-1)}, \ldots, \psi_D^{(t-1)})$

$\vdots$

d. Simulate $\psi_d^{(t)}$ from $p(\psi_d \mid \psi_1^{(t)}, \ldots, \psi_{d-1}^{(t)}, \psi_{d+1}^{(t-1)}, \ldots, \psi_D^{(t-1)})$

$\vdots$

D. Simulate $\psi_D^{(t)}$ from $p(\psi_D \mid \psi_1^{(t)}, \ldots, \psi_{D-1}^{(t)})$.

This algorithm is easy to implement when each of the full conditionals is a known distribution. However, it is often true that one or more of the full conditionals are not standard distributions and, thus, drawing samples from these distributions is not simple. In that situation, a Metropolis or a Metropolis-Hastings algorithm can be utilized to draw a realization for that particular step of the Gibbs sampler. The initial development of the Gibbs sampler was done by Geman and Geman (1984) on Gibbs distributions, hence its name. Gelfand and Smith (1990) demonstrated its applicability to Bayesian computation. Roberts and Smith (1993) proved the convergence of the Markov chain generated by the Gibbs sampler under general conditions.

2.4.3 Posterior inference in the two-stage model

With the two-stage model for complete data, we use a Metropolis algorithm with a multivariate normal jumping distribution. Because the normal distribution assigns probability to the entire real line, we introduce a transformation of $\boldsymbol{\alpha}$. Define $\boldsymbol{\gamma}$ as follows:

$$\begin{aligned} \gamma_i &= log\left(\frac{\alpha_i}{\sum_{k=1}^{I} \alpha_k - \alpha_i}\right), \quad i = 1, \ldots, I-1 \\ \gamma_I &= log\left(\sum_{k=1}^{I} \alpha_k\right) \end{aligned} \tag{2.14}$$

The first $I-1$ elements of $\boldsymbol{\gamma}$ are logit transformations of the corresponding elements mean of the Dirichlet and the last element γ_I is the log of the sum of the α's. A normal jumping distribution is more suitable for $\boldsymbol{\gamma}$. The mean of the normal jumping distribution is taken to be the previous state in the Markov chain. The variance, V, of the multivariate Normal jumping distribution is taken proportional to the inverse of the estimated negative second derivative matrix of the posterior distribution of $\mathrm{p}(\boldsymbol{\alpha} \mid \mathbf{Y})$ at the posterior mode. Formally,

$$V = \left\{ -\frac{\partial^2 \log \mathrm{p}(\boldsymbol{\gamma} \mid \mathbf{Y})}{\partial \boldsymbol{\gamma}^2} \right\}\Bigg|_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} \tag{2.15}$$

with $\hat{\boldsymbol{\gamma}}$ equal to the posterior mode. Then the jumping distribution is $\mathrm{N}(\boldsymbol{\gamma}^{(t-1)}, \mathrm{cV})$, where $\boldsymbol{\gamma}^{(k)}$ is the k^{th} iteration of the Markov chain and c is a constant value chosen to make the jumping distribution efficient. Once posterior simulations are obtained, we transform the parameters back to the original scale using the inverse transformation

$$\alpha_i = \left(\frac{e^{\gamma_i}}{e_i^{\gamma}+1}\right) e^{\gamma_I}, \quad i = 1, \ldots, I-1; \tag{2.16}$$

$$\alpha_I = e^{\gamma_I} - \sum_{i=1}^{I-1} \alpha_i. \tag{2.17}$$

The Bayesian framework allows for inference about any function of the parameters, $\phi = \phi(\boldsymbol{\Theta}, \boldsymbol{\alpha})$, that might be of interest. In other words, we can calculate summaries of the posterior distribution for any ϕ, e. g. posterior quantiles. To do this we draw a sample $\boldsymbol{\alpha}$ from $\mathrm{p}(\boldsymbol{\alpha} \mid \mathbf{Y})$ and then draw a sample $\boldsymbol{\Theta}$ from $\mathrm{p}(\boldsymbol{\Theta} \mid \boldsymbol{\alpha}, \mathbf{Y})$. From the sampled parameters we calculate a value of ϕ. Repeating this process, we get a collection of values of ϕ that can then be used to compute the summaries of interest. Because the $(\boldsymbol{\Theta}, \boldsymbol{\alpha})$ realizations are sampled from a MCMC algorithm the samples are not independent. If we want Monte Carlo standard errors for a specific quantity, like $\mathrm{E}(\phi \mid \mathbf{Y})$, then we must do additional analyses like those described by (Geyer (1992)).

In the remainder of this section we identify various quantities that might be of interest in the sample survey context. With multinomial data in a finite population context, it is natural to focus on the proportion of the population with responses in some subset of the multinomial categories. Of course, this is not a deterministic function of the model parameters. We construct a stochastic realization of the finite population. Let $\boldsymbol{\lambda}$ be a vector of zero's and one's that identifies which cells are of interest. Recall that $\mathbf{Y}_j$ is the vector of observed data for cluster j and n_j as the total number of sample individuals in cluster j. Now define

$$\mathbf{Y}_j \equiv 0, \text{ if } j > J \quad \text{and} \quad n_j \equiv 0 \text{ if } j > J. \tag{2.18}$$

These definitions explicitly specify that there were no observed respondents in clusters $J+1, \ldots, M$. Now let $\mathbf{Y}_j^*$ represents a realization of the unobserved responses in cluster j. There are two cases: for $j = 1, \ldots, J$, there are $N_j - n_j$ unobserved responses; for $j = J+1, \ldots, M$, the entire population of size N_j is not sampled and consequently all are unobserved. For the first case we have simulations from the posterior distribution of $\boldsymbol{\theta}_j$ as a result of our posterior simulations. We draw a vector of observations $\mathbf{Y}_j^*$ as follows,

$$\mathbf{Y}_j^* \sim \text{Multinomial}(N_j - n_j, \boldsymbol{\theta}_j^*). \tag{2.19}$$

For the second case, the unsampled clusters, we have no direct information about the population in the cluster. However, we do have information about the population of all clusters. That information is encapsulated in $\text{p}(\boldsymbol{\alpha} \mid \mathbf{Y})$. We use that information to generate a vector of probabilities, $\boldsymbol{\theta}_j^*$, for the unseen cluster, j_u from the Dirichlet distribution,

$$\boldsymbol{\theta}_j^* \sim \text{p}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}). \tag{2.20}$$

Finally we simulate responses for the entire population of cluster j,

$$\mathbf{Y}_j \sim \text{Multinomial}(N_j, \boldsymbol{\theta}_j^*) \tag{2.21}$$

Combining all of the clusters and taking a weighted average based upon the number in each cluster, we get the population proportion

$$\phi_1 = \frac{\sum_{j=1}^{M} \lambda^T \left(\mathbf{Y}_j + \mathbf{Y}_j^*\right)}{\sum_{j=1}^{M} N_j}. \tag{2.22}$$

This quantity simplies to

$$\frac{\sum_{j=1}^{M} \lambda^T \left(\mathbf{Y}_j + \mathbf{Y}_j^*\right)}{MN^*} \tag{2.23}$$

when all of the clusters are of equal size, $N_j = N^*$ for each j. This quantity, (2.22), represents the percentage of individuals in a realization of the entire population that would choose the cells specified in $\boldsymbol{\lambda}$. The variability in ϕ_1 would be the variability under the model for the population proportion. Gelman et al. (1995) show that a quantity analogous to ϕ_1 exhibits the same variability as measured by the usual design-based estimate of the repeated sampling variance for the traditional survey estimate in the Gaussian case. Consequently, the posterior interval for ϕ_1 should be approximately the same as the design-based confidence interval.

We also define a number of model-based quantities that are related to the population proportion. The first is a weighted average of the cluster proportions,

$$\phi_2 = \frac{\sum_{j=1}^{M} N_j \boldsymbol{\lambda}^T \boldsymbol{\theta}_j^*}{\sum_{j=1}^{M} N_j} \tag{2.24}$$

One again there are two cases: for sampled clusters we have available a sample of draws $\boldsymbol{\theta}_j^*$, but for unsampled clusters we sample $\boldsymbol{\theta}_j^*$ from the prior distribution $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$. We generate $\boldsymbol{\theta}_j^*$ for each cluster j by taking realizations $\mathrm{p}(\boldsymbol{\Theta} \mid \boldsymbol{\alpha}, \mathbf{Y})$ according to (2.20). As we did with (2.22), we can simplify (2.24) to when the number of individuals, N_j is the same in each cluster.

$$\frac{\sum_{j=1}^{M} \boldsymbol{\lambda}^T \boldsymbol{\theta}_j^*}{M} \tag{2.25}$$

Another model-based quantity is the population-weighted average of the expected cluster proportions. The posterior expected proportions for cluster j is $E(\boldsymbol{\theta}_j^* \mid \mathbf{Y})$. This quantity is defined below for the two cases:

$$E[\boldsymbol{\theta}_j^*] = \begin{cases} \dfrac{\boldsymbol{\alpha}+\mathbf{Y}_j}{\sum\limits_{k=1}^{I} \alpha_k + n_j} & j = 1, \ldots, J \qquad \text{(sampled clusters)} \\ \dfrac{\boldsymbol{\alpha}}{\sum\limits_{k=1}^{I} \alpha_k} & j = J+1, \ldots, M \quad \text{(unsampled clusters)} \end{cases} \tag{2.26}$$

Then the population weighted average of the posterior expected proportions is

$$\phi_3 = \frac{\sum\limits_{j=1}^{M} N_j \boldsymbol{\lambda}^T E[\boldsymbol{\theta}_j^*]}{\sum\limits_{j=1}^{M} N_j}. \tag{2.27}$$

Another model-based quantity of interest is the superpopulation proportion

$$\phi_4 = \frac{\boldsymbol{\lambda}^T \boldsymbol{\alpha}}{\sum\limits_{k=1}^{I} \alpha_k}. \tag{2.28}$$

This quantity is the proportion in the underlying superpopulation that would choose the cells determined by $\boldsymbol{\lambda}$. The superpopulation model is a construct that posits an infinite population from which the finite population has been sampled. Thus, we interpret (2.28) as the proportion of interest in the superpopulation.

Of the quantities defined here, ϕ_1 is the primary goal of sample survey inference. The model-based quantities are primarily of interest for assessing the model. We expect less variability as we move our consideration from ϕ_1 to ϕ_2 to ϕ_3 to ϕ_4. Thus the relative sizes of the posterior intervals are of interest.

Inference need not be limited to estimates of the population proportion. The cluster effect is of great interest in traditional design-based inference. As defined by Cochran (1977) the design effect or deff is the ratio of the estimated variance in the quantity of interest under the present sampling design to the estimated variance if the sample had been collected under SRS. For the Dirichlet- Multinomial models, several authors, including Altham (1976), Brier (1980) and Koehler and Wilson (1986), have shown that the design effect can be expressed as a function of the superpopulation parameter $\boldsymbol{\alpha}$,

$$\text{PDE} = \frac{\sum_{k=1}^{I} \alpha_k + \overline{n}}{\sum_{k=1}^{I} \alpha_k + 1} \tag{2.29}$$

where $\overline{n}$ is the average number of individuals per cluster. We will refer to this quantity as the posterior design effect (PDE). Under the Bayesian approach we obtain not just a point estimate but the complete posterior distribution for the PDE. The PDE is of interest in that is allows us to assess the degree to which the clusters are homogeneous. If $\sum_{k=1}^{I} \alpha_k$ is large, the design effect is approximately one and all clusters have the same proportions. If $\sum_{k=1}^{I} \alpha_k$ is very small, then observations are homogeneous within a cluster (all responses are similar) and heterogeneous between clusters. The deff is then approximately $\overline{n}$ which indicates that the additional observations from within a cluster do not help with inference.

2.5 Missing Data

Sample surveys of finite populations implicitly involve missing data, namely the responses of the members of the population that were not selected. These are a form of "intentional" missing data or designed missing data. In addition, it is possible to have unintentional missing data when survey respondents refuse to answer an item. These are often treated as a separate response, NA, for not answered, rather than as missing data. In the example that motivated this thesis, however, the failure to answer can be considered a form of missing data because of the rules of the plebiscite with which the data are concerned, (Rubin et al. (1995)). Consequently, we extend our approach to accommodate unintentional missing values.

Little and Rubin (1987) introduce a categorization of the mechanisms that produce missing data. The easiest to deal with are data mechanisms that are missing completely at random (MCAR). Under MCAR, the probability that an individual response is observed is completely independent of observed or missing data (i. e. unrelated to demographic features, answers to the specific question). A more general notion is missing at random (MAR). MAR assumes that the probability that a response to a specific question is observed, i. e. , is not missing, may depend on the observed data (possibly including other variables that were observed), but not on the response in question. This is crucial in that it presupposes no tendency for an individual with a specific response on a question fail to respond to the question. Under MCAR we can restrict our attention to those individuals with complete data since they are a random sample of the original population. Under MAR we can carry out traditional model-based analysis, conditional only on the observed values. The final category described by Little and Rubin covers the case when the actual missing data mechanism is nonignorable (NI). In that case the probability that a response is missing depends on the value that would have been observed. The only way to proceed is to build a model for the response mechanism. A key point is that the observed data themselves do not allow us to determine if MAR or NI is a more accurate description. This is because the issue is whether the probability of being missing depends on the value that would have been observed which is, of course, not available. Several authors, including Gelman et al. (1995), have advocated proceeding under the MAR assumption and then possibly assessing the sensitivity of the conclusions to alternative NI models. Rubin et al. (1995) describe one methodology for carrying this out.

In this section we will assume the observations are missing at random (MAR) in the terminology of Little and Rubin (1987) or Rubin (1976). Again, MAR assumes that the probability that a response is observed depends on the parameters that generated the data and possibly on other variables that were observed but not on the response of interest.

2.5.1 Notation

The data that motivated this work is based upon a trivariate binary response which we analyze as a $2^3 = 8$-dimensional multinomial random variable. A consequence of this setup is that when one or more of the questions is answered with a Don't Know, there is a subset of the 8 cells into which the actual response might fall. In effect a missing binary response corresponds to partial information about the multinomial random variable. For example, for a respondent that answered Yes to the question about independence, Don't Know to the question about Attendance and Yes to the question about Secession, their response could be in one of only two possible cells (Yes, No, Yes) or (Yes, Yes, Yes). Similar patterns can be derived for each partial pattern of incomplete responses. We will follow Rubin et al. (1995) in using the term "pattern of missingness" when referring to these patterns.

Let ν be a pattern of missingness and let P be the set of all such patterns. Next, let A_ν be the set of all possible cells in missingness pattern ν. For example, if $I = 8$ and ν_1 has possible categories 1, 4, and 6, then $A_{\nu_1} = \{1, 4, 6\}$. This would correspond to an individual whose response is known to lie in one of the three categories but for which further refinement is not possible due to a lack of information.

Let n_j^ν be the number of individuals in the j^{th} cluster with pattern of missingness ν. Also let Y_{ij}^ν is the unobserved number of responses that actually fall in category i from cluster j from among the n_j^ν individuals with missingness pattern ν. It is these Y_{ij}^ν that are the unintentional missing data. For completeness, define $Y_{ij}^\nu \equiv 0$ if $n_j^\nu = 0$ or $i \notin A_\nu$. Let $\mathbf{Y}_j^\nu = (Y_{1j}^\nu, Y_{2j}^\nu, \ldots, Y_{Ij}^\nu)$ and let $\mathbf{Y}^{mis}$ represent $\{\mathbf{Y}_j^\nu : j = 1, \ldots, J\}$. Finally, let $\mathbf{Y}^{obs}$ encompass not only the completely observed multinomial counts $\mathbf{Y}_j$'s but the marginal totals for the patterns of missingness in each cluster, n_j^ν.

2.5.2 Probability model

The basic probability model that we developed in Section 2.4 remains unchanged except that now we wish to incorporate information from both the fully observed and the partially observed data. Assuming MAR, we are interested in the posterior distribution of $\mathrm{p}(\mathbf{\Theta}, \boldsymbol{\alpha} \mid \mathbf{Y}^{obs})$. In practice it is convenient to augment the problem by incorporating the missing values $\mathbf{Y}^{mis}$ as unobserved random variables. The joint posterior distribution of all the unknown quantities is

$$
\begin{aligned}
&p(\mathbf{\Theta}, \boldsymbol{\alpha}, \mathbf{Y}^{mis} \mid \mathbf{Y}^{obs}) \\
&\propto \left[\prod_{j=1}^{J}\left\{\prod_{\nu \in P}\binom{n_j^{\nu}}{\mathbf{Y}_j^{\nu}}\right\}\prod_{i=1}^{I}\theta_{ij}^{Y_{ij}+\sum_{\nu\in P} Y_{ij}^{\nu}}\right] \\
&\quad \times \left[\prod_{j=1}^{J}\left\{\Gamma\left(\sum_{k=1}^{I}\alpha_k\right)\prod_{i=1}^{I}\frac{\theta_{ij}^{\alpha_i-1}}{\Gamma(\alpha_i)}\right\}\right]\left(\sum_{k=1}^{I}\alpha_k\right)^{-\frac{2I+1}{2}}
\end{aligned}
\tag{2.30}
$$

If we can study the above distribution by obtaining simulations of$(\mathbf{\Theta}, \boldsymbol{\alpha}, \mathbf{Y}^{mis})$, then we merely ignore the simulated $\mathbf{Y}^{mis}$ to study the marginal posterior distribution $\mathrm{p}(\mathbf{\Theta}, \boldsymbol{\alpha} \mid \mathbf{Y}^{obs})$.

As in (2.7) we can proceed by factoring the posterior distribution as the product of a marginal and a conditional distribution,

$$
\mathrm{p}(\mathbf{\Theta}, \boldsymbol{\alpha}, \mathbf{Y}^{mis} \mid \mathbf{Y}^{obs}) = \mathrm{p}(\mathbf{\Theta} \mid \boldsymbol{\alpha}, \mathbf{Y}^{mis}, \mathbf{Y}^{obs})\mathrm{p}(\boldsymbol{\alpha}, \mathbf{Y}^{mis} \mid \mathbf{Y}^{obs}). \tag{2.31}
$$

One advantage of this factorization is that the conditional distribution of $\mathbf{\Theta}$ given $(\boldsymbol{\alpha}, \mathbf{Y}^{mis}, \mathbf{Y}^{obs})$ is the complete data conditional, $\mathrm{p}(\mathbf{\Theta} \mid \boldsymbol{\alpha}, \mathbf{Y})$ from Section 2.4. Recall that it is the product of Dirichlet distributions by conjugacy and hence it is easy to simulate from this piece of the posterior distribution. The second piece of the right hand side of (2.31) is more complicated. We can simulate from this distribution, once again using MCMC. However, the Metropolis algorithm will not suffice. Instead we use the Gibbs sampling algorithm described in Section 2.4.2. The following two-step Gibbs sampling algorithm allows us to draw samples from the distribution $\mathrm{p}(\boldsymbol{\alpha}, \mathbf{Y}^{mis} \mid \mathbf{Y}^{obs})$

1. Simulate $\boldsymbol{\alpha}$ from $\mathrm{p}(\boldsymbol{\alpha} \mid \mathbf{Y}^{mis}, \mathbf{Y}^{obs})$

2. Simulate $\mathbf{Y}^{mis}$ from $\mathrm{p}(\mathbf{Y}^{mis} \mid \boldsymbol{\alpha}, \mathbf{Y}^{obs})$

To draw from the first of these two Gibbs steps, we simply condition on $\mathbf{Y}^{mis}$ and $\mathbf{Y}^{obs}$ which is equivalent to conditioning on the complete data set. This can be sampled from using the Metropolis algorithm of the previous section. This is an example of using a Metropolis step within a Gibbs algorithm. There is flexibility in determining how many Metropolis steps to perform in each Gibbs step. We use a single step.

To draw samples from the second step of our Gibbs algorithm turns out to be a simple task, but this is not immediately obvious from the form of the distribution.

$$\mathrm{p}(\mathbf{Y}^{mis} \mid \boldsymbol{\alpha}, \mathbf{Y}^{obs}) \propto \prod_{j=1}^{J} \left\{ \left\{ \prod_{\nu \in P} \binom{n_j^{\nu}}{\mathbf{Y}_j^{\nu}} \right\} \prod_{i=1}^{I} \frac{\Gamma\left(Y_{ij} + \alpha_i + \sum_{\nu \in P} + Y_{ij}^{\nu} \right)}{\Gamma(\alpha_i)} \right\} \tag{2.32}$$

where factors depending on $\boldsymbol{\alpha}$ have been omitted. For a particular pattern of missingness, ν^* in cluster j, we have

$$\binom{n_j^{\nu^*}}{\mathbf{Y}_j^{\nu^*}} \prod_{i=1}^{I} \Gamma\left(Y_{ij}^{\nu^*} + Y_{ij} + \alpha_i + \sum_{\substack{\nu \in P \\ \nu \neq \nu^*}} Y_{ij}^{\nu} \right). \tag{2.33}$$

This distribution is the kernel of a Dirichlet-Multinomial distribution. This can be seen by looking at the density of an S-dimensional Dirichlet-Multinomial random variable, $\mathbf{Z}$, with parameters m and $\phi_1, \ldots, \phi_S$,

$$\mathrm{p}(\mathbf{Z} \mid \phi, m) = \binom{m}{z_1, \ldots, z_S} \frac{\Gamma\left(\sum_{s=1}^{S} \phi_i \right)}{\Gamma\left(\sum_{s=1}^{S} \phi_i + m \right)} \prod_{s=1}^{S} \frac{\Gamma(z_s + \phi_s)}{\Gamma(\phi_s)}, \tag{2.34}$$

and noting that the kernel of this distribution (the terms involving $\mathbf{Z}$) is simply

$$\binom{m}{z_1, \ldots, z_S} \prod_{s=1}^{S} \Gamma(z_s + \phi_s) \tag{2.35}$$

Then (2.33) is recognized as a Dirichlet-Multinomial distribution with I categories, sample size $n_j^{\nu^*}$ and parameters $Y_{ij} + \alpha_i + \sum_{\substack{\nu \in P \\ \nu \neq \nu^*}} Y_{ij}^{\nu}$. Generating realizations from a Dirichlet-Multinomial(ϕ, m) distribution is straightforward. As described in Section 2.3, to generate a realization of $\mathbf{Z}$ from above we draw a sample $\boldsymbol{\theta}$ from a Dirichlet with parameters ϕ. Then we generate $\mathbf{Z}$ from a multinomial with a vector of proportions, $\boldsymbol{\theta}$, and count m. Thus, the second step of our two-step Gibbs algorithm actually consists of a distinct Gibbs sampling algorithm that cycles through each pattern of missingness/cluster combination. The algorithm described here yields samples from the posterior distribution $p(\boldsymbol{\Theta}, \boldsymbol{\alpha} \mid \mathbf{Y}^{obs})$. Those samples can be used to draw inference as described in Section 2.4. We now turn our attention to applying these methodology to a data set.

2.6 Application: The Slovenian Public Opinion Survey

Today, Slovenia is a small nation in southeastern Europe. From 1945 to 1991, Slovenia was one of six republics in Yugoslavia. In December of 1990, a plebiscite was held regarding the possibility of independence from Yugoslavia. The adult citizens of Slovenia overwhelmingly voted for independence. On June 25, 1991 the Slovenian parliament voted to declare itself independent from Yugoslavia. What followed was a 10-day conflict between the Yugoslav army and the Slovene territorial defense forces (Silber and Little (1997)). On October 8, 1991, Slovenia became a wholly independent nation.

In the month preceding the December 1990 plebiscite, the Slovenian Public Opinion Survey (SPO), a regular survey on a variety of subjects, was conducted. Included in the 1990 SPO were the following three questions concerning independence:

1. Are you in favor of Slovenian independence?

2. Are you in favor of Slovenia's secession from Yugoslavia?

3. Will you attend the plebiscite?

Table 2.1 Survey Totals for the SPO

		Independence		
Secession	Attendance	Yes	No	Don't Know
	Yes	1191	8	21
Yes	No	8	0	4
	Don't Know	107	3	9
	Yes	158	68	29
No	No	7	14	3
	Don't Know	18	43	31
	Yes	90	2	109
Don't Know	No	1	2	25
	Don't Know	19	8	96

For each question a Yes (Y), No (N) or Don't Know (DK) response was recorded. The last question was especially important, since under the rules of the plebiscite an individual not attending the plebiscite would be treated as an voting NO on the question of independence. Table 2.6 provides the survey responses to these questions. The survey included many other questions on demographics, attitudes toward other nations, etc. For this chapter and the next, we ignore the other survey questions, though these might be helpful for improving model-based inference with missing data. In Chapter 4 we present a model-based methodology for incorporating additional covariates such as those described above.

The SPO was carried out via a three-stage sampling design. The entire nation was divided into 1000 primary sampling units (PSU's) of approximately equal size. Within each PSU they were further broken into 16 secondary sampling units (SSU's). A subset consisting of 139 of the 1000 clusters, or PSU's, were sampled and then 3 of 16 SSU's were chosen from each sampled PSU. Finally 5 individuals were chosen from the approximately 100 individuals within each SSU. As in most large scale surveys, not all of the selected individuals responded. Nonresponding individuals were replaced by substitutes. We treat these substitutes as if they were the original respondents. In total, there were 2074 respondents. Note that this is just under the $139 \times 3 \times 5 = 2085$ that were intended. Apparently, in 11 cases they could not obtain a respondent even in exhausting the list of substitutes. For the purposes of this chapter we treat the SPO as a two-stage cluster sample and ignore the SSU's. Thus we will treat the data as being comprised of respondents from 139 PSU's, with approximately 15 individuals per SSU. Chapter 3 describes a methodology for analyzing the full three-stage design.

The fully observed responses, i. e., those without a single DK response, form a trivariate binomial random variable. For the purposes of this section and the remainder of the thesis, we will treat this data as a $2^3 = 8$ dimensional multinomial distribution. In doing so, we follow Rubin et al. (1995) who originally analyzed this data set. Their main focus was taking account of the DK responses, treating them as missing data. This treatment is different from the common practice in the U. S. There a DK is often treated as a separate response but here because all individuals would ultimately vote (perhaps by not attending) it seems appropriate to think of the DK's as hiding the respondents intentions. Rubin et al. (1995) ignored the clustering; they treat the responses as if they came from a SRS and then use the classical design effect, deff, to increase the size of the resulting confidence (or posterior) intervals to reflect the clustering. The primary contribution of this reanalysis is to incorporate the effect of clustering directly into the model.

2.6.1 Complete data results

In this section we restrict attention to the 1454 individuals that answered yes or no to each of the questions listed in the previous subsection. In other words, we will eliminate from our analysis all respondents with at least one DK response from the three questions.

Table 2.2 Summary of the Posterior Distribution for Model Parameters

	Corresponding	Posterior		Percentiles	
	Response	Mean	2.5^{th}	50^{th}	97.5^{th}
p_1	NNN	0.010	0.006	0.010	0.016
p_2	NNY	0.001	0.001	0.001	0.003
p_3	NYN	0.006	0.003	0.006	0.010
p_4	NYY	0.006	0.003	0.006	0.012
p_5	YNN	0.048	0.038	0.048	0.060
p_6	YNY	0.006	0.003	0.006	0.012
p_7	YYN	0.108	0.092	0.109	0.126
p_8	YYY	0.814	0.793	0.814	0.834
$\sum_{k=1}^{I} \alpha_k$		898.643	34.805	124.617	5867.530

The order of the questions is Attendance, Independence, Succession.

Using the Metropolis MCMC algorithm of Section 2.4, we ran six chains of length 4000 each. All of the parameters had PSR's, see Section 2.4.1, under 1.2, so we conclude that the sample are representative of the posterior distribution. In Table 2.2 we summarize the posterior distribution of $\boldsymbol{\alpha}$, by giving summaries of the posterior distribution of $p_i = \alpha_i / \sum_{k=1}^{I} \alpha_k$ and of $\sum_{k=1}^{I} \alpha_k$. The proportions agree well with the data. The vector of sample proportions from the population for the eight categories is $(0.010, 0.000, 0.005, 0.006, 0.0470, 0.006, 0.109, 0.819)^T$. The responses that correspond to these categories are $(NNN, NNY, NYN, NYY, YNN, YNY, YYN, YYY)^T$ for the questions regarding Attendance, Independence and Succession, respectively. The posterior means and medians in Table 2.2 differ very little from these totals. Histograms of the marginal posterior distributions of the p_i's are given in Figure 2.1.

In practice, the individual parameters, $p_i, i = 1, \ldots, I$, are not of most interest. Table 2.3 gives posterior inference for some of the key quantities defined in Section 2.4. In terms of the notation that we defined, $\boldsymbol{\lambda} = (0, \ldots, 0, 1, 1)^T$ which identifies the two

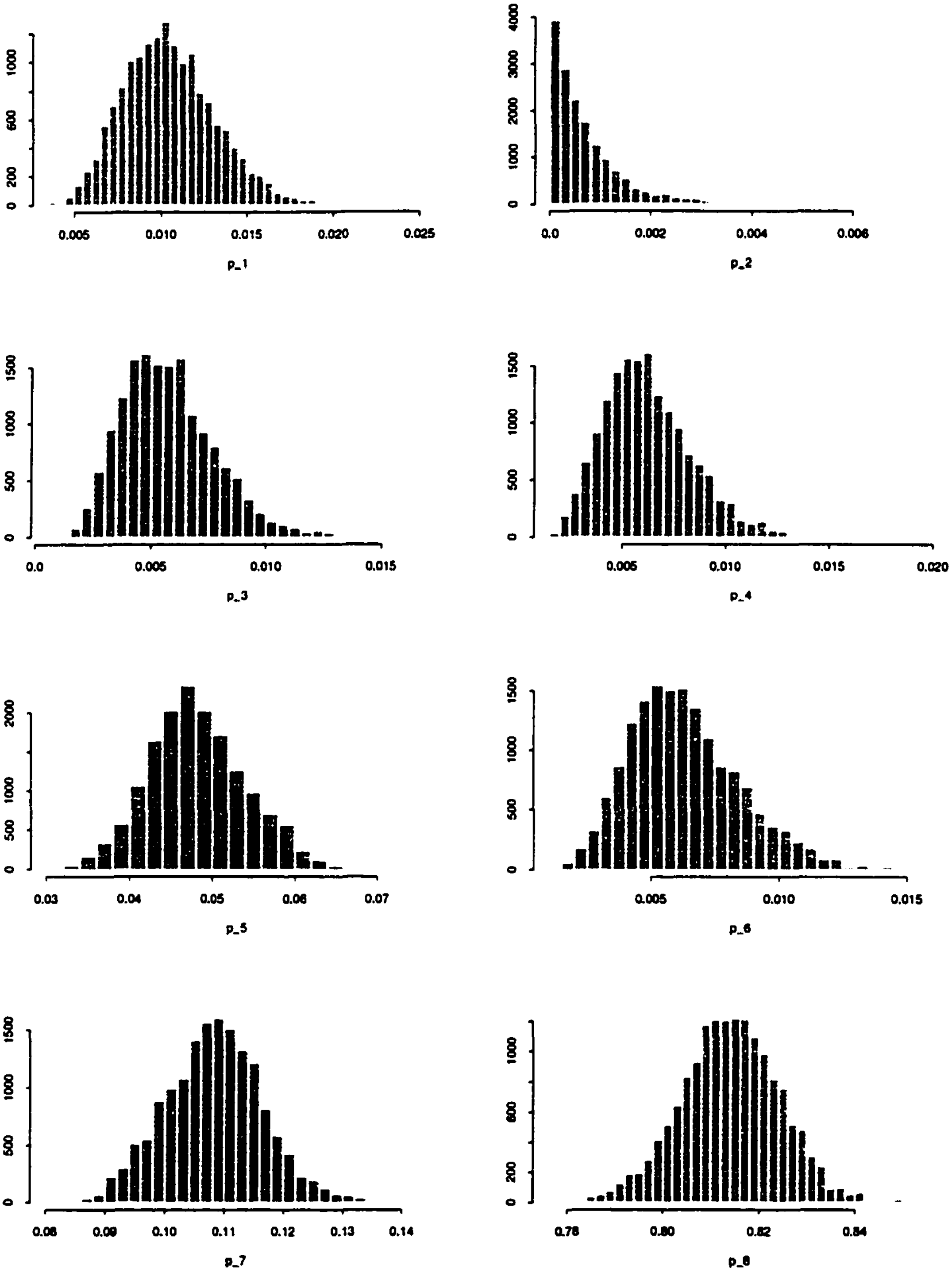

Figure 2.1 Histograms of Samples from the Posterior Distributions for the Proportions in Table 2.2

Table 2.3 Proportions of Slovenian Voters Who Intend to Attend the Plebiscite and to Vote for Independence

	Posterior		Percentiles	
	Mean	2.5^{th}	50^{th}	97.5^{th}
ϕ_1	0.921	0.903	0.921	0.935
ϕ_2	0.921	0.903	0.921	0.935
ϕ_3	0.924	0.910	0.925	0.937
ϕ_4	0.922	0.907	0.923	0.936
Design-based	0.928	0.913	0.928	0.943
$PDE = \frac{\sum_{k=1}^{I} \alpha_k+\bar{n}}{\sum_{k=1}^{I} \alpha_k+1}$	1.089	1.002	1.075	1.264

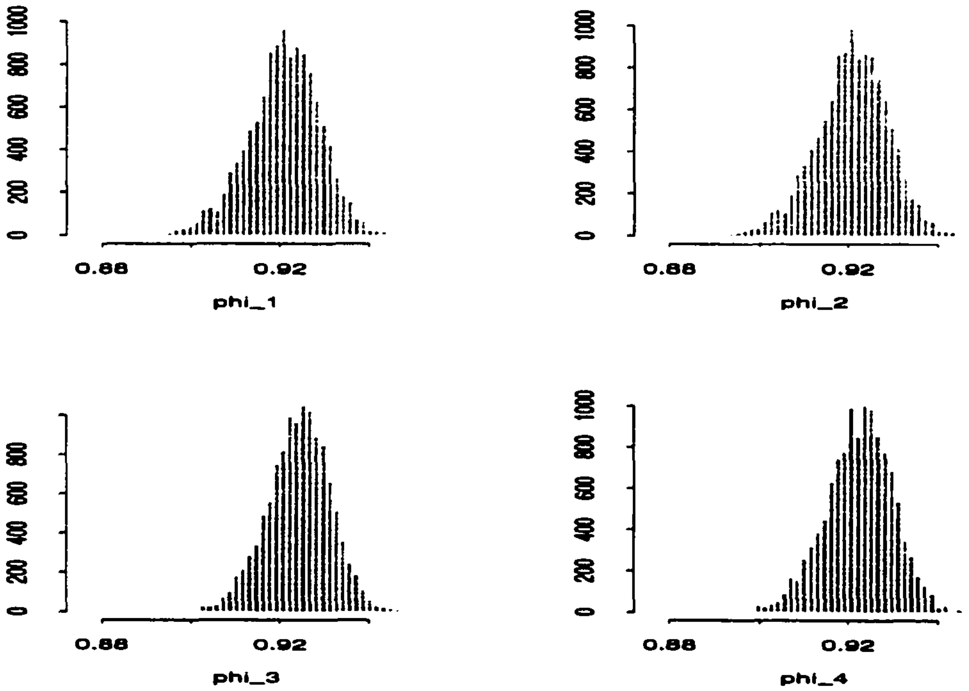

Figure 2.2 Histograms of Samples from the Posterior Distributions for the Results in Table 2.3

responses consisting with attending the plebiscite and voting yes. We include in this table, the ϕ's that we discussed in Section 2.4: ϕ_1, the proportion of the population attending and voting yes; ϕ_2, the mean of the cluster proportions; ϕ_3, the mean of the expected cluster proportions; ϕ_4, the superpopulation proportion. Table 2.3 also gives a 95% confidence interval for the proportion of interest using the design-based approach to two-stage cluster sampling (Equation (10.24) in Cochran (1977)).

We first note that the $\phi's$ all seem to be centered about 0.922. This is slightly lower than the sample proportion that is the center of the design-based confidence interval, 0.928. This may be a result of the flat prior distribution assumed on the proportions, since the posterior is a weighted average of the sample proportions and the proportions assigned by the prior distribution. The prior proportion in this case is 0.250 which may slightly lower the population proportion. It is also noteworthy that the distributions are skewed. This can be contrasted with the designed based interval which is based on the assumption of asymptotic normality, which yields a symmetric distribution. We can see the skewness of the posterior distribution for these quantities in Figure 2.2. It is also worth commenting on the amount of variability in each of the posterior quantities. Focusing on the finite population quantities of interest, we note that ϕ_2 and ϕ_1 have less variability than ϕ_3 which we should expect. Recall that ϕ_1 is the posterior realization of the proportion of the vote for the entire population, ϕ_2 is the mean of the cluster proportions, and ϕ_3 is value of the mean of the cluster proportions. One surprising result is that ϕ_1 and ϕ_2 have the same amount of variability. One likely explanation is that since N_j is 1600 for each cluster, the simulated cluster counts, $\mathbf{Y}_j^*$, closely match the expected counts $\boldsymbol{\theta}_j$ in each cluster. Finally, the posterior distribution of ϕ_4, the superpopulation proportion, is more variable than the posterior distribution for the finite population quantities. We also note that the posterior interval for ϕ_1 has approximately the same width as quantity is was designed to mimic the design-based confidence interval, 0.032 versus 0.030. This show that the Bayesian analysis of the hierarchical model reproduces the traditional result for the complete data case.

The last row of Table 2.3 is the posterior design effect (PDE). It is worth noting that the range of possible values for the PDE is 1 to the average cluster sample size, $\overline{n}$. First we note that the posterior interval for ϕ_1 has approximately the same width as quantity is was designed to mimic the design-based confidence interval, 0.032 versus 0.030. This discrepancy may in part be due to the skewness of the posterior distribution of ϕ_1. The design-based 95% confidence interval from a two-stage cluster sample has a width that is the 1.154 times that of the equivalent confidence interval assuming an SRS. Thus the posterior 95% credible set is about 1.154 times as large as the 95% confidence interval analyzing the same data as if it came from an SRS.

2.6.2 Missing data results

In this section we present the results obtained by analyzing the SPO data incorporating the responses of those who had at least one DK response. The patterns of missingness that were discussed in Section 2.5 are listed in Table 2.4. For example, the first pattern of missingness describes those individuals that answered Y to the question about Independence, Y to the question about Attendance and DK to the question about Secession or (Y,Y,DK). For these individuals their true, unobserved response is either (Y,Y,Y) or (Y,Y,N). In all there are 12 patterns with one question missing and each of these has two elements in the corresponding set A_ν. There are 6 patterns with two questions missing, each with four elements in A_ν. There are also respondents that answered DK to each question. There are 8 different elements in A_ν but, these respondents provide no information to improve our estimate of the model parameters and consequently are omitted.

Our approach to the partially observed responses requires that the missing data mechanism is MAR. In this context we can provide further intuition about the assumption by considering a pool of individuals identical on all observed variables. Then MAR

Table 2.4 Patterns of Missingness

	Questions		
Independence	Attendence	Secession	A_ν
Y	Y	DK	{7, 8}
Y	N	DK	{5, 6}
Y	DK	Y	{6, 8}
Y	DK	N	{5, 7}
Y	DK	DK	{5, 6, 7, 8}
N	Y	DK	{3, 4}
N	N	DK	{1, 2}
N	DK	N	{1, 3}
N	DK	Y	{2, 4}
N	DK	DK	{1, 2, 3, 4}
DK	Y	Y	{4, 8}
DK	Y	N	{3, 7}
DK	Y	DK	{3, 4, 7, 8}
DK	N	Y	{2, 6}
DK	N	N	{1, 5}
DK	N	DK	{1, 2, 5, 6}
DK	DK	Y	{2, 4, 6, 8}
DK	DK	N	{1, 3, 5, 7}
DK	DK	DK	{1, 2, 3, 4, 5, 6, 7, 8}

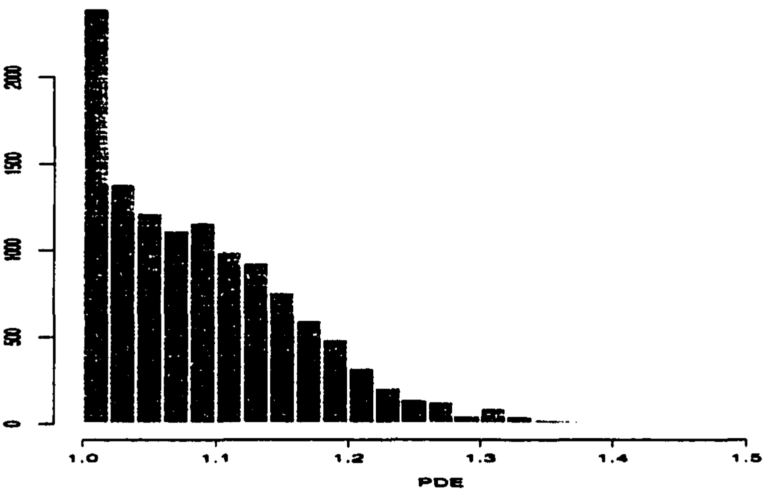

Figure 2.3 Histogram of Samples from the Posterior Distributions for the Posterior Design Effect

Table 2.5 Summaries of the Posterior Distribution for Model Parameters

	Corresponding	Posterior		Percentiles	
	Responses	Mean	2.5^{th}	50^{th}	97.5^{th}
p_1	NNN	0.023	0.015	0.023	0.035
p_2	NNY	0.002	0.000	0.002	0.007
p_3	NYN	0.009	0.004	0.009	0.016
p_4	NYY	0.011	0.006	0.011	0.018
p_5	YNN	0.069	0.054	0.069	0.084
p_6	YNY	0.010	0.006	0.010	0.016
p_7	YYN	0.121	0.104	0.121	0.139
p_8	YYY	0.754	0.731	0.755	0.776
$\sum_{k=1}^{I} \alpha_k$		224.063	32.048	72.415	809.238

The order of the questions is Attendance, Independence, Succession.

Table 2.6 Proportions of Slovenian Voters Who Intend to Attend the Plebiscite and to Vote for Independence Incorporating Data from Partially Observed Response

	Posterior		Percentiles	
	Mean	2.5^{th}	50^{th}	97.5^{th}
ϕ_1	0.876	0.857	0.876	0.892
ϕ_2	0.876	0.857	0.876	0.892
ϕ_3	0.880	0.864	0.881	0.895
ϕ_4	0.875	0.857	0.876	0.892
Rubin, Stern and Vehovar		0.863	0.883	0.900
Actual			0.885	

implies that for individuals in this pool the probability of a DK response to a question is the same for those who would have answered Yes and for those that would have answered No. A priori one might expect that individuals planning to answer No, an unpopular response, are more likely to answer DK. The MAR assumption requires that any such tendency is completely explained by observed variables, (Rubin et al. (1995)). We carried out the analyses of Section 2.4 with 6 independent Gibbs sampling sequences of length 5000. Based on the PSR, the last 4000 draws from each of the 6 chains appear to be sufficiently well mixed together that we can treat them as draws from the posterior distribution, $p(\boldsymbol{\Theta}, \boldsymbol{\alpha} \mid \mathbf{Y}^{obs})$. Table 2.5 contains posterior distribution summaries for the parameters, $p_i = \alpha_i / \sum_{k=1}^{I} \alpha_k$. Figure 2.4 contains histograms for the proportions.

We can compare the results of Table 2.5, with those from using only the completely observed data. The most striking observation to be made is that the first seven proportions all increase while the eighth proportion is the only one to decrease. This is interesting because the first seven proportions have at least one No response, while the eighth proportion corresponds is all Yes responses. This suggests that those who an-

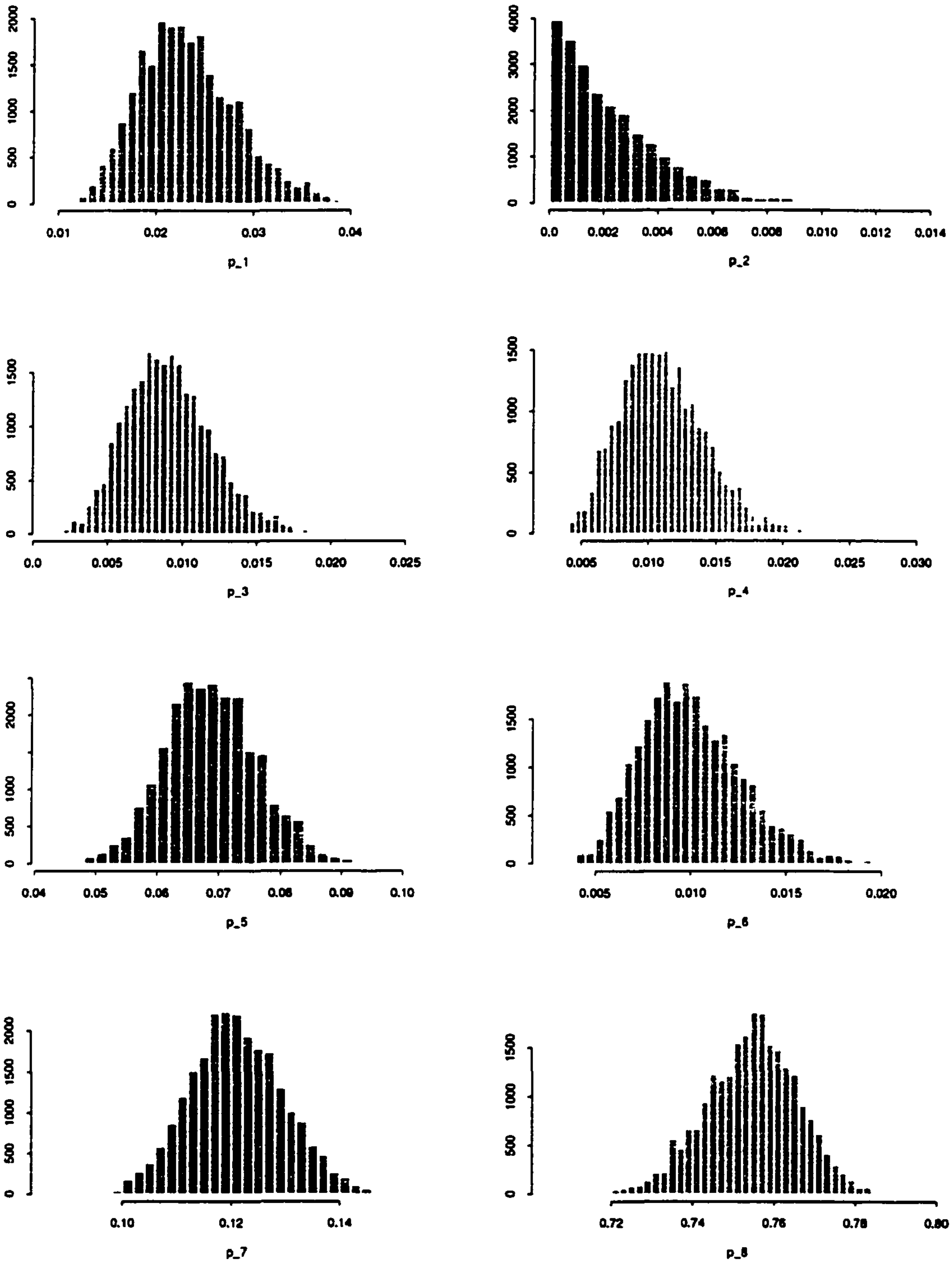

Figure 2.4 Histograms of Samples from the Posterior Distributions for the Proportions in Table 2.5

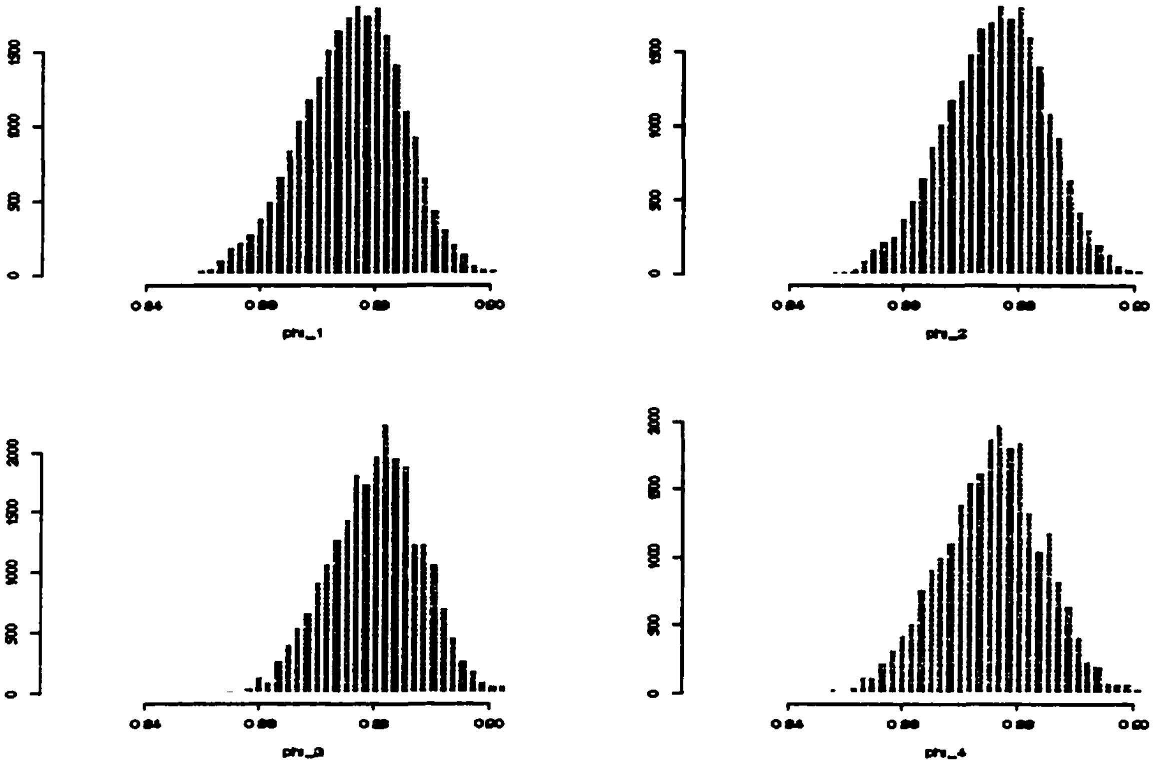

Figure 2.5 Histograms of Samples from the Posterior Distributions of the Proportions in Table 2.6

swered with at least one DK were probably hiding a No response. The MAR assumption allows for conditioning on covariates, here the questions that were answered by the respondent. Hence the covariates for the cases with missing responses provide additional information beyond that contained in the completely observed responses. Note that these results imply that MCAR is not a valid assumption for this data set. Under MCAR we would have expected that the results from the complete case would be the same as those for the missing data case. The results of Table 2.5 clearly contradict this. The first proportion, p_1, corresponding to the proportion of the population giving all No responses drastically from the complete analysis to the missing analysis. Although the proportion is small in both cases, less than 2%, it is worth noting that the percentage doubles when we incorporated the missing observations.

The quantities described in Section 2.4 are summarized in Table 2.6 and Figure 2.5.

The posterior intervals for the missing data analysis are slightly larger than the posterior intervals for the same proportions for the completely observed data analysis. There are two forces acting here. The inclusion of additional data (50% more cases though they are not fully observed) responses yields more information about these quantities which implies smaller intervals. However, the difference in the distribution of the responses among complete and incomplete cases adds variability to the estimates.

Finally, while it is nice to note that all of the posterior quantities engulf the actual plebiscite total, it is not entirely relevant. The goal of the SPO was to gauge public opinion in Slovenia about a month before the plebiscite, not to predict the actual vote. It is certainly natural to expect that the plebiscite and the survey would be similar but had any major events occurred during the time interval between the survey and the plebiscite in between there results could be very different.

2.7 Proof of Theorem 1

Theorem 1 *For the model (2.2), (2.4) and (2.5), the posterior distribution is proper if there exists at least one cluster that has responses in at least two different cells.*

A more convenient parameterization for the proof is

$$\kappa = \sum_{k=1}^{I} \alpha_k \tag{2.36}$$

$$\gamma_i = \frac{\alpha_i}{\sum\limits_{k=1}^{I} \alpha_k} \quad \forall i = 1, \ldots, I-1 \tag{2.37}$$

with γ_I defined as $1 - \sum\limits_{i=1}^{I-1} \gamma_i$. This parameterization focuses on the first $I-1$ elements of the mean of the Dirichlet distribution, the γ_i's, and the sum of the Dirichlet parameters, κ. The posterior distribution of (γ, κ) given $\mathbf{Y}$ is obtained by applying a change of variables to (2.9) and incorporating the Jacobian, $\kappa^{(I-1)}$. The transformed posterior distribution (up to a normalizing constant) is

$$\left[\prod_{j=1}^{J}\left\{\frac{\Gamma(\kappa)}{\Gamma(\kappa+n_j)}\prod_{i=1}^{I}\frac{\Gamma(Y_{ij}+\gamma_i\kappa)}{\Gamma(\gamma_i\kappa)}\right\}\right]\kappa^{-\frac{3}{2}}. \tag{2.38}$$

We need to show that the posterior is integrable. To do this we examine the limiting behavior of the posterior distribution to determine if it is integrable for the following scenarios.

I. κ is fixed and one or more of the elements of γ_i go to 0.

II. γ is fixed and κ goes to 0.

III. γ is fixed and κ goes to ∞.

Before we begin note that

$$\varepsilon\Gamma(\varepsilon) = \Gamma(1+\varepsilon) \to 1 \text{ as } \varepsilon \to 0. \tag{2.39}$$

and we assume that $n_j \geq 1$ for each cluster j.

Consider limit **I.**

Let $Z \subset \{1, 2, \ldots, I\}$ denote the indices of the proportions tending to 0, with $1 \leq\mid Z \mid\leq I-1$. We assume that $\gamma_i \to 0$ for all $i \in Z$. In addition we assume that the remaining proportions maintain constant ratio's $\frac{\gamma_j}{\gamma_k}$ for all $j,k \notin Z$. Then the unnormalized posterior is,

$$\begin{aligned}
&\left[\prod_{j=1}^{J}\left\{\frac{\Gamma(\kappa)}{\Gamma(\kappa+n_j)}\prod_{i=1}^{I}\frac{\Gamma(Y_{ij}+\gamma_i\kappa)}{\Gamma(\gamma_i\kappa)}\right\}\right]\kappa^{-3/2} \\
&= \left[\kappa^{-3/2}\prod_{j=1}^{J}\frac{\Gamma(\kappa)}{\Gamma(\kappa+n_j)}\right]\left[\prod_{j=1}^{J}\prod_{i\notin Z}\frac{\Gamma(Y_{ij}+\gamma_i\kappa)}{\Gamma(\gamma_i\kappa)}I_{(Y_{ij}>0)}\right]\left[\prod_{j=1}^{J}\prod_{i\in Z}\frac{\Gamma(Y_{ij}+\gamma_i\kappa)}{\Gamma(\gamma_i\kappa)}I_{(Y_{ij}>0)}\right] \\
&< \left[\kappa^{-3/2}\prod_{j=1}^{J}\frac{\Gamma(\kappa)}{\Gamma(\kappa+n_j)}\right]\left[\prod_{j=1}^{J}\prod_{i\notin Z}\frac{\Gamma(Y_{ij}+1+\gamma_i\kappa)}{\Gamma(\gamma_i\kappa)}I_{(Y_{ij}>0)}\right] \\
&\qquad\qquad \times\left[\prod_{j=1}^{J}\prod_{i\in Z}\frac{\Gamma(Y_{ij}+1+\gamma_i\kappa)}{\Gamma(\gamma_i\kappa)}I_{(Y_{ij}>0)}\right]
\end{aligned} \tag{2.40}$$

where the equality follows because $\frac{\Gamma(Y_{ij}+\gamma_i\kappa)}{\Gamma(\gamma_i\kappa)} = 1$ if $Y_{ij} = 0$, and the inequality holds because $\Gamma(x) < \Gamma(x+1)$ for any $x \in (1,\infty)$.

Now let $\delta = {}^{argmin}_{x\in(0,\infty)}\Gamma(x)$, and use the fact that $\Gamma(2+px) < \Gamma(2+x)$ if $0 < p < 1$ and $x \in (0,\infty)$. Then (2.40)

$$\begin{aligned} &< \left[\kappa^{-3/2}\prod_{j=1}^{J}\frac{\Gamma(\kappa)}{\Gamma(\kappa+n_j)}I_{(Y_{ij}>0)}\right] \\ &\quad\times\left[\prod_{j=1}^{J}\prod_{i\notin Z}\frac{\Gamma(Y_{ij}+1+\kappa)}{\Gamma(\delta)}I_{(Y_{ij}>0)}\right]\left[\prod_{j=1}^{J}\prod_{i\in Z}\frac{\Gamma(Y_{ij}+1+\kappa)(\gamma_i\kappa)}{\gamma_i\kappa\Gamma(\gamma_i\kappa)}I_{(Y_{ij}>0)}\right]. \end{aligned} \tag{2.41}$$

Using (2.39), this is then asymptotically equivalent to

$$\begin{aligned} &\left[\kappa^{-3/2}\prod_{j=1}^{J}\frac{\Gamma(\kappa)}{\Gamma(\kappa+n_j)}I_{(Y_{ij}>0)}\right] \\ &\quad\times\left[\prod_{j=1}^{J}\prod_{i\notin Z}\frac{\Gamma(Y_{ij}+1+\kappa)}{\Gamma(\delta)}I_{(Y_{ij}>0)}\right]\left[\prod_{j=1}^{J}\prod_{i\in Z}\Gamma(Y_{ij}+1+\kappa)(\gamma_i\kappa)^{I_{(Y_{ij}>0)}}\right] \\ &< \left[\kappa^{-3/2}\prod_{j=1}^{J}\frac{\Gamma(\kappa)}{\Gamma(\kappa+n_j)}I_{(Y_{ij}>0)}\right] \\ &\quad\times\left[\prod_{j=1}^{J}\prod_{i\notin Z}\frac{\Gamma(Y_{ij}+1+\kappa)}{\Gamma(\delta)}I_{(Y_{ij}>0)}\right]\left[\prod_{j=1}^{J}\prod_{i\in Z}\Gamma(Y_{ij}+1+\kappa)(\kappa)^{I_{(Y_{ij}>0)}}\right] \end{aligned} \tag{2.42}$$

Thus we have a finite bound for the unnormalized posterior. Since γ_i ranges over a finite space, the posterior is integrable and, hence, proper.

Consider limit **II**.

Assume that γ_i's are fixed and let κ tend to 0.

The unnormalized posterior is

$$= \kappa^{-3/2}\prod_{j=1}^{J}\left\{\frac{\Gamma(\kappa)}{\Gamma(\kappa+n_j)}\prod_{i=1}^{I}\frac{\Gamma(Y_{ij}+\gamma_i\kappa)}{\Gamma(\gamma_i\kappa)}\right\} \tag{2.43}$$

$$= \kappa^{-3/2}\prod_{j=1}^{J}\left\{\frac{\Gamma(\kappa)\kappa}{\Gamma(\kappa+n_j)\kappa}\prod_{i=1}^{I}\frac{\Gamma(Y_{ij}+\gamma_i\kappa)(\gamma_i\kappa)}{\gamma_i\kappa\Gamma(\gamma_i\kappa)}I_{(Y_{ij}>0)}\right\} \tag{2.44}$$

which, using (2.39) is asymptotically equivalent to the following

$$\kappa^{-3/2}\prod_{j=1}^{J}\left\{\frac{\kappa^{-1}}{\Gamma(n_j)}\prod_{i=1}^{I}\left\{\Gamma(Y_{ij})I_{(Y_{ij}>0)}(\gamma_i\kappa)^{I_{(Y_{ij}>0)}}\right\}\right\}$$

Then under the condition of limit **II**, the posterior is integrable if

$$-3/2 - J + \sum_{j=1}^{J}\sum_{i=1}^{I} I_{(Y_{ij}>0)} > -1 \tag{2.45}$$

which occurs if there exists a cluster that has at least one response in at least two different cells.

Consider limit **III**.

Assume now that γ_i's are fixed and let κ tend to ∞. The unnormalized posterior is again,

$$\kappa^{-3/2}\prod_{j=1}^{J}\left\{\frac{\Gamma(\kappa)}{\Gamma(\kappa+n_j)}\prod_{i=1}^{I}\frac{\Gamma(Y_{ij}+\gamma_i\kappa)}{\Gamma(\gamma_i\kappa)}\right\} \tag{2.46}$$

$$= \kappa^{-3/2}\prod_{j=1}^{J}\left\{\frac{1}{(n_j-1+\kappa)\dots(\kappa)}\prod_{i=1}^{I}(Y_{ij}-1+\gamma_i\kappa)\dots(\gamma_i\kappa)\right\} \tag{2.47}$$

$$< \kappa^{-3/2}\prod_{j=1}^{J}\left\{\kappa^{-n_j}\prod_{i=1}^{I}(Y_{ij}-1+\gamma_i\kappa)\dots(\gamma_i\kappa)\right\} \tag{2.48}$$

$$= \kappa^{-3/2}\prod_{j=1}^{J}\prod_{i=1}^{I}\left\{\left(\frac{Y_{ij}-1+\gamma_i\kappa}{\kappa}\right)\dots\left(\frac{\gamma_i\kappa}{\kappa}\right)\right\} \tag{2.49}$$

This last product results from the following equality $\sum_{i=1}^{I} Y_{ij} = n_j$. Then as κ tend to ∞, (2.49) becomes

$$\kappa^{-3/2}\prod_{j=1}^{J}\prod_{i=1}^{I}\gamma_i^{Y_{ij}} \tag{2.50}$$

Then the posterior is integrable in the limit as κ goes to ∞ since $\kappa^{-3/2}$ is integrable.

To summarize for the given hyperprior distribution, the posterior distribution is proper as long as there exist at least one cluster with responses in more than one cell. We can also note that a flat prior distribution on $\sum_{k=1}^{I}\alpha_k$, which corresponds to $p(\gamma,\kappa) \propto \kappa^{(I-1)}$ would fail to produce a integrable posterior under limit **III**.

CHAPTER 3 THREE-STAGE MODEL

This chapter we will presents a model for analyzing polychotomous data from a three-stage cluster sample. In three-stage cluster sample the population is divided into primary sampling units (PSU's) and each PSU is divided into secondary sampling units (SSU's), with each SSU consisting of a subset of the original population. A three-stage cluster sample is obtained by selecting a simple random sample (SRS) of primary sampling units or clusters, a simple random sample of SSU's or subclusters within each chosen clusters, and a SRS or elements from each of the chosen subclusters. The organization of this chapter parallels that of the previous chapter. Section 3.1 introduces notation for three-stage cluster sampling and specifies a hierarchical superpopulation model for analyzing data collected in this manner. Section 3.2 describes our approach for making posterior inferences based on the model. Once again MCMC algorithms are used to generate a sample from the posterior distribution. The incorporation of missing data into the analysis is described in Section 3.3. In Section 3.4 we construct and analyze simulated data sets. These assist in the interpretation of the parameters of the three-stage model, as well as demonstrate the feasibility of using the model. Finally, Section 3.5 presents results for the case of the Slovenian Public Opinion Survey.

3.1 The Model

Suppose the population of interest is divided into M clusters or primary sampling units (PSU's) and the j^{th} PSU is further divided into N_j subclusters or secondary sampling units (SSU's), $j = 1, \ldots, M$. Finally we suppose that N_{jk} elements of the population are in the k^{th} SSU of the j^{th} PSU, $j = 1, \ldots, M$, $k = 1, \ldots, N_j$. A simple random sample of J primary sampling units (PSU's) is selected; $J' = M - J$ clusters are not selected. Then a SRS of the size K_j is selected from the N_j SSU's in the j^{th} PSU, $j = 1, \ldots, M$ with $K_j' = N_j - K_j$ subclusters remaining. Finally, a SRS of n_{jk} individuals are selected from the k^{th} SSU of the j^{th} PSU. Often the number of SSU's chosen are the same for each selected PSU, i.e., $N_j = N$ for all j. Likewise, the number of individuals chosen within each selected SSU is often constant, $n_{jk} = n$ for all j, k. Each individual's response is one of I possible responses patterns.

We propose the following hierarchical superpopulation model for analyzing polychotomous data from a three-stage cluster sample. Let Y_{ijk} represent the number of individuals in the k^{th} subcluster of the j^{th} cluster with response i and let $\mathbf{Y}_{jk} = (Y_{1jk}, \ldots, Y_{Ijk})^T$. The data $\mathbf{Y}_{jk}$ are modeled as multinomial given $\boldsymbol{\theta}_{jk}$ and n_{jk},

$$\mathbf{Y}_{jk} \mid \boldsymbol{\theta}_{jk}, n_{jk} \sim \text{Multinomial}(n_{jk}, \boldsymbol{\theta}_{jk}). \tag{3.1}$$

where $\boldsymbol{\theta}_{jk} = (\theta_{1jk}, \ldots, \theta_{Ijk})$ is the vector of response probabilities for an individual in the k^{th} SSU of the j^{th} PSU. As in Chapter 2, the multinomial model is not totally correct but can be a useful approximation if n_{jk} is small relative to N_{jk}.

The subcluster probability vectors, $\boldsymbol{\theta}_{jk}$, are modeled as exchangeable draws from a Dirichlet distribution. As in Chapter 2, this distribution is a natural choice because it is the conjugate prior distribution for the multinomial distribution. We parameterize the Dirichlet distribution for the probability vector in cluster (PSU) j in terms of a vector or proportions, $\boldsymbol{\gamma}_j = (\gamma_{1j}, \gamma_{2j}, \ldots, \gamma_{Ij})^T$, with $\sum_{i=1}^{I} \gamma_{ij} = 1$ and a prior sample size of η_j,

$$\boldsymbol{\theta}_{jk} \mid \eta_j, \boldsymbol{\gamma}_j \sim \text{Dirichlet}(\eta_j \boldsymbol{\gamma}_j). \tag{3.2}$$

Thus the SSU proportions, $\boldsymbol{\theta}_{jk}$, have mean , $\boldsymbol{\gamma}_j$, which represents the average vector of proportions for the SSU's in PSU j. The similarity of the collection $(\boldsymbol{\theta}_{j1}, \ldots, \boldsymbol{\theta}_{jN_j})$ is measured by the quantity η_j which determines the variability of the Dirichlet distribution. A large value of η_j means little variance among the vectors of proportions within a PSU. A small values of η_j implies that there is considerable variability among SSU's within a PSU.

The PSU proportion means, $\gamma_1, \ldots, \gamma_J$, are modeled as exchangeable draws from a Dirichlet distribution

$$\gamma_j \mid \kappa, \boldsymbol{\alpha} \sim \text{Dirichlet}(\kappa \boldsymbol{\alpha}), \quad j = 1, \ldots, J, \tag{3.3}$$

with mean $\boldsymbol{\alpha}$. The parameter κ governs the variability of the cluster level proportions. The η_j's are treated as exchangeable draws from a gamma distribution. The gamma distribution is used because η_j must be positive and because the gamma is a flexible family of distributions. The particular parameterization of the gamma distribution that we use here is one that was formulated by Morris (1982), Morris (1983).

$$\text{p}(\eta_j \mid a, b/a) = \left(\frac{\eta_j}{b/a}\right)^{a-1} \frac{e^{a\eta_j/b}}{b/a\Gamma(a)}. \tag{3.4}$$

With this parameterization, $E[\eta_j] = b$ and $V[\eta_j] = b^2/a$. Using this setup, we model the η_j's as

$$\eta_j \mid m, \mu \sim \text{Gamma}(m, \mu/m), j = 1, \ldots, J, \tag{3.5}$$

where the mean of each η_j is equal to μ and the variance of each η_j is μ^2/m. As at the previous level of the model, large κ implies that all of the γ_j's will be similar, whereas small κ implies great variability among the PSU means.

The hierarchical structure that is inherent in the data collection methodology is exploited by this model. The individual observations yield information about the SSU level probability vectors. These, in turn, contain information about the PSU level proportions. Finally, the PSU level proportions provide information about the Dirichlet distribution that describes the population of PSU's.

It remains to place prior distributions on the parameters $\boldsymbol{\alpha}, \kappa, \mu, m$. We assume that $\boldsymbol{\alpha}$ has a Dirichlet prior distribution with parameters equal to 1. This is a uniform distribution over the simplex of I-dimensional probability vectors. We model the mean of the η_j's, the parameter μ, as an Inverse-gamma(c_1, c_2) random variable. We chose the inverse-gamma prior distribution since it is the conjugate prior distribution for μ, (Morris (1983)). We model the parameter m using a gamma distribution with parameters c_3 and c_4. A traditional gamma parameterization is used as the prior distribution for m because m must be positive. For the remaining parameter κ we also use a traditional gamma distribution as the prior distribution with parameters c_5 and c_6. The constants c_1 through c_6 are constants chosen to make the prior distributions flat or diffuse, e. g. $c_1 = 0.005, c_2 = 3, c_3 = 0.1, c_4 = 0.01, c_5 = 0.1, c_6 = 0.01$. The reason for doing this is to allow the data to shape the posterior. When a prior distribution is specified with large variance, data is given precedence over the prior distribution, since the prior is quite "diffuse."

We now provide additional discussion about some features of the model. Note that it is only possible to construct a hierarchical model using the Dirichlet parameterization that includes a probability vector and an "effective" sample size. A hierarchy using this parameterization was first suggested for a two-stage cluster sample by Nandram (1998). It would be difficult to create a hierarchy using the standard Dirichlet parameterization. Another interesting feature of the model is the assumption of prior independence of $\boldsymbol{\eta}$ and κ. We make this assumption because it allows for small or great variability among the PSU probability vectors for SSU's within a PSU, and small or great variability among the PSU probability vectors.

For example, consider a national sample of the United States population with states serving as PSU's and counties within a state as the SSU's. For some responding variables we might find little variability among counties within a state, but a large amount of variability across states. This would imply large η's and small κ. On the other hand, it is useful to also allow for the possibility of little variability among states and great variability among counties (small η's and large κ).

The complete model is given below,

$$\begin{aligned}
\mathbf{Y}_{jk} \mid \boldsymbol{\theta}_{jk}, n_{jk} &\sim \text{Multinomial}(n_{jk}, \boldsymbol{\theta}_{jk}) \\
\boldsymbol{\theta}_{jk} \mid \boldsymbol{\gamma}_{jk}, \eta_j &\sim \text{Dirichlet}(\eta_j \boldsymbol{\gamma}_{jk}) \\
\boldsymbol{\gamma}_j \mid \boldsymbol{\alpha}, \kappa &\sim \text{Dirichlet}(\kappa \boldsymbol{\alpha}) \\
\eta_j \mid m, \mu &\sim \text{Gamma}(m, \mu) \\
\boldsymbol{\alpha} &\sim \text{Dirichlet}(1, 1, \ldots, 1) \\
\mu &\sim \text{Inverse-gamma}(c_1, c_2) \\
m &\sim \text{Gamma}(c_3, c_4) \\
\kappa &\sim \text{Gamma}(c_5, c_6)
\end{aligned} \tag{3.6}$$

3.2 Posterior Inference

3.2.1 The posterior distribution

For the three-stage cluster sample all of the prior distributions are proper distributions which guarantees that the posterior distribution is a proper distribution. Before constructing the posterior distribution we introduce some convenient notation for referring to subsets of the parameters and the data. Let $\mathbf{Y} = (\mathbf{Y}_{11}^T, \ldots, \mathbf{Y}_{1K_1}^T, \ldots, \mathbf{Y}_{J1}^T, \ldots, \mathbf{Y}_{JK_J})^T$ denote a single vector that is the concatenation of all of the SSU response vectors, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_{11}^T, \ldots, \boldsymbol{\theta}_{1K_1}^T, \ldots, \boldsymbol{\theta}_{J1}^T, \ldots, \boldsymbol{\theta}_{JK_J}^T)^T$, a similar vector for the SSU probability vectors, $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_J^T)^T$ a concatenation of the PSU probability vectors, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_J)^T$ The posterior distribution is easily derived (up to a normalizing constant) from a product of the sampling distribution of the data, $\mathbf{Y}$, and the prior distributions on the parameters,

$$\mathrm{p}(\boldsymbol{\Theta},\boldsymbol{\Gamma},\boldsymbol{\alpha},\boldsymbol{\eta},\kappa,\mu,m \mid \mathbf{Y}) \propto \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\Theta},\boldsymbol{\Gamma},\boldsymbol{\alpha},\boldsymbol{\eta},\kappa,\mu,m) \times \mathrm{p}(\boldsymbol{\Theta},\boldsymbol{\Gamma},\boldsymbol{\alpha},\boldsymbol{\eta},\kappa,\mu,m). \tag{3.7}$$

Implicit in these equations is the the number of individuals sampled in each subcluster, n_{jk}. Taking account of the hierarchical structure we can write, the joint posterior distribution (again, up to a normalizing constant) as

$$\begin{aligned}
&\mathrm{p}(\boldsymbol{\Theta},\boldsymbol{\Gamma},\boldsymbol{\alpha},\boldsymbol{\eta},\kappa,\mu,m \mid \mathbf{Y}) \qquad (3.8)\\
\propto\ & \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\Theta})\mathrm{p}(\boldsymbol{\Theta} \mid \boldsymbol{\eta},\boldsymbol{\Gamma})\mathrm{p}(\boldsymbol{\eta} \mid m,\mu)\mathrm{p}(\boldsymbol{\Gamma} \mid \kappa,\boldsymbol{\alpha})\mathrm{p}(\boldsymbol{\alpha})\mathrm{p}(\kappa)\mathrm{p}(m)\mathrm{p}(\mu)\\
=\ & \prod_{j=1}^{J}\prod_{k=1}^{K_j}\left[\binom{n_{jk}}{\mathbf{Y}_{jk}}\prod_{i=1}^{I}\theta_{ijk}^{Y_{ijk}}\right]\\
\times\ & \prod_{j=1}^{J}\prod_{k=1}^{K_j}\left\{\Gamma(\eta_j)\prod_{i=1}^{I}\frac{\theta_{ijk}^{\gamma_{ij}\eta_j-1}}{\Gamma(\gamma_{ij}\eta_j)}\right\}\\
\times\ & \prod_{j=1}^{J}\left\{\Gamma(\kappa)\prod_{i=1}^{I}\frac{\gamma_{ij}^{\alpha_i\kappa-1}}{\Gamma(\alpha_i\kappa)}\right\}\\
\times\ & \prod_{j=1}^{J} \mathit{left}\{\left(\frac{\eta_j m}{\mu}\right)^{m-1}\frac{e^{-m\eta_j/\mu}}{\mu/m\Gamma(m)}\mathit{right}\}\\
\times\ & \left(\frac{1}{\mu c_1}\right)^{c_2-1}\frac{e^{-1/c_1\mu}}{c_1\Gamma(c_2)}\\
\times\ & m^{c_3-1}e^{-c_4 m}\\
\times\ & \kappa^{c_5-1}e^{-c_6\kappa}
\end{aligned}$$

3.2.2 A MCMC algorithm

To sample from the posterior distribution, it is convenient to factor the posterior distribution,

$$\begin{aligned}
\mathrm{p}(\boldsymbol{\Theta},\boldsymbol{\Gamma},\boldsymbol{\eta},\boldsymbol{\alpha},\kappa,\mu,m \mid \mathbf{Y}) &= \mathrm{p}(\boldsymbol{\Theta} \mid \boldsymbol{\Gamma},\boldsymbol{\eta},\boldsymbol{\alpha},\kappa,\mu,m)\mathrm{p}(\boldsymbol{\Gamma},\boldsymbol{\eta},\boldsymbol{\alpha},\kappa,\mu,m \mid \mathbf{Y}) \qquad (3.9)\\
&= \mathrm{p}(\boldsymbol{\Theta} \mid \boldsymbol{\Gamma},\boldsymbol{\eta},\mathbf{Y})\mathrm{p}(\boldsymbol{\Gamma},\boldsymbol{\eta},\boldsymbol{\alpha},\kappa,\mu, \mid \mathbf{Y})
\end{aligned}$$

The first quantity on the right-hand side of (3.9) is easily identified as a product of Dirichlet distributions.

$$\mathrm{p}(\boldsymbol{\Theta} \mid \boldsymbol{\Gamma}, \boldsymbol{\eta}, \mathbf{Y}) = \prod_{j=1}^{J} \prod_{k=1}^{K_j} \binom{n_{jk}}{Y_{1jk}, \ldots, Y_{Ijk}} \left[\Gamma(\eta_j) \prod_{i=1}^{I} \frac{\theta_{ijk}^{Y_{ijk}+\gamma_{ij}\eta_j-1}}{\Gamma(\gamma_{ij}\eta_j)} \right]. \tag{3.10}$$

Since drawing realizations from Dirichlet distributions can be done quite easily, we focus on the second term on the right-hand side of (3.9).

The second distribution in (3.9), $\mathrm{p}(\boldsymbol{\Gamma}, \boldsymbol{\eta}, \boldsymbol{\alpha}, \kappa, \mu, m \mid \mathbf{Y})$, can be found by first integrating out $\boldsymbol{\Theta}$ to obtain the marginal distribution

$$\int \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\Theta})\mathrm{p}(\boldsymbol{\Theta} \mid \boldsymbol{\Gamma}, \boldsymbol{\eta}, \boldsymbol{\alpha}, \kappa, \mu, m)\, d\boldsymbol{\Theta} = \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\Gamma}, \boldsymbol{\eta}, \boldsymbol{\alpha}, \kappa, \mu, m) \tag{3.11}$$

and then multiplying (3.11) by the prior distribution $\mathrm{p}(\boldsymbol{\Gamma}, \boldsymbol{\eta}, \boldsymbol{\alpha}, \kappa, \mu, m)$. The resulting distribution (up to a normalizing constant) is

$$\begin{aligned}
&\mathrm{p}(\boldsymbol{\Gamma}, \boldsymbol{\eta}, \boldsymbol{\alpha}, \kappa, \mu, m \mid \mathbf{Y}) && (3.12)\\
&\propto \prod_{j=1}^{J} \prod_{k=1}^{K_j} \left[\binom{n_{jk}}{\mathbf{Y}_{jk}} \frac{\Gamma(\eta_j)}{\Gamma(\eta_j + n_{jk})} \prod_{i=1}^{I} \frac{\Gamma(Y_{ijk} + \gamma_{ij}\eta_j)}{\Gamma(\gamma_{ij}\eta_j)} \right] \\
&\times \prod_{j=1}^{J} \left[\Gamma(\kappa) \prod_{i=1}^{I} \frac{\gamma_{ij}^{\alpha_i\kappa-1}}{\Gamma(\alpha_i\kappa)} \right] \\
&\times \prod_{j=1}^{J} \left(\frac{\eta_j m}{\mu} \right)^{m-1} \frac{e^{-m\eta_j/\mu}}{\mu/m\Gamma(m)} && (3.13)\\
&\times \left(\frac{1}{\mu c_1} \right)^{c_2-1} \frac{e^{-1/c_1\mu}}{c_1\Gamma(c_2)} \\
&\times m^{c_3-1} e^{-c_4 m} \\
&\times \kappa^{c_5-1} e^{-c_6\kappa}
\end{aligned}$$

Because (3.13) is a high-dimensional non-standard multivariate distribution, it is easiest to sample from this distribution using MCMC methodology. Specifically we sample from the posterior distribution using Gibbs sampling (see Section 2.4.1). To describe the Gibbs sampling algorithm, we introduce $\boldsymbol{\beta} = (\boldsymbol{\Gamma}^T, \boldsymbol{\eta}^T, \boldsymbol{\alpha}^T, \kappa, \mu, m)^T$ as notation for the entire parameter vector and denote its dimension by D. It turns out that none of the full conditional distributions, p($\beta_d \mid \beta_1, \ldots, \beta_{d-1}, \beta_{d+1}, \ldots, \beta_D$), is of standard form. As a consequence we will take each element of $\boldsymbol{\beta}$, β_d, to be a scalar and use a one-dimensional Metropolis algorithm to draw a univariate sample from the full conditional posterior distribution. We choose a univariate Gaussian distribution for the Mctropolis jumping distribution. Since the Gaussian distribution places mass on the entire real number line, we transform the parameters as follows before carrying out the algorithm. For the proportions $\boldsymbol{\alpha}$ and $\boldsymbol{\Gamma}$, we use a logit transformation. For the remaining elements, $\boldsymbol{\eta}$, κ, μ, m we use a log transformation. More explicitly we let $\boldsymbol{\phi}=f(\boldsymbol{\beta})$ then f is defined as follows $\boldsymbol{\beta}$.

$$\begin{aligned}
f(\alpha_i) &= logit(\alpha_i), \quad i = 1, \ldots, I \\
f(\kappa) &= log(\kappa) \\
f(\gamma_{ij}) &= logit(\gamma_{ij}), \quad i = 1, \ldots, I, j = 1, \ldots, J \\
f(\eta_j) &= log(\eta_j), \quad j = 1, \ldots, J \\
f(\mu) &= log(\mu) \\
f(m) &= log(m).
\end{aligned}$$

For the remainder of this section, we use $\boldsymbol{\phi}$ to be the vector representing the transformed elements of $\boldsymbol{\beta}$.

Consider the Metropolis step for ϕ_d. For the mean of the Gaussian jumping distribution at the t^{th} iteration we use $\phi_k^{(t-1)}$. For the variance, we use a quantity that is proportional to an estimate of the marginal posterior variance of ϕ_d. To estimate the marginal posterior variance, we begin with an initial estimate of the variance, σ_d. We carry out a pilot run of the MCMC algorithm and calculate the percentage of times, p_d, that we accept a proposed candidate, ϕ_d^*, based upon the Metropolis transition rule. If p_d is significantly lower than 0.25 (the optimal rate recommended by Gelman et al. (1996)) then we decrease our estimate of σ_d. If p_d is significantly higher than 0.25, then we increase σ_d. Then we run another pilot run. Ultimately, we find an estimate of $\hat{\sigma}_d$ such that the Metropolis acceptance rate is approximately 0.25. Then. the Metropolis jumping distribution that we use to generate a candidate for ϕ_d is

$$\phi_d^* \sim \text{Normal}(\phi_d^{(t-1)}, \hat{\sigma}_d). \tag{3.14}$$

Our Gibbs sampling algorithm is a series of univariate Metropolis steps with a jumping distribution of this form.

3.2.3 Quantities of interest

Although we may be interested in any of the model parameters, we proceed as in Chapter 2 to define a series of quantities related to the finite population proportion of interest. Let $\boldsymbol{\lambda}$ be an I-dimensional column vector of zero's and one's which determines the proportion of interest from among the I categories. For notational convenience, define

$$\begin{aligned} \mathbf{Y}_{jk} &\equiv \mathbf{0} \quad \text{if} j > J \quad \text{or} \quad j \leq J,\ k > K_j \\ &\text{and} \\ n_{jk} &\equiv 0 \quad \text{if} j > J \quad \text{or} \quad j \leq J,\ k > K_j \end{aligned} \tag{3.15}$$

These definitions take the observed vector of counts to be 0 and the sum of those counts to be 0 for the unsampled PSU's and for the unsampled SSU's within sampled PSU's define the observed vector of counts to be 0 and the sum of those counts to be 0. We assume in this section that we have draws from the full posterior distribution

$$\mathrm{p}(\boldsymbol{\Theta}, \boldsymbol{\beta} \mid \mathbf{Y}) \tag{3.16}$$

where $\boldsymbol{\Theta}$ and $\boldsymbol{\beta}$ are defined in Section 3.2.2.

The first and primary quantity of interest is the percentage of people in the entire population who would choose a particular set of categories of interest as defined by $\boldsymbol{\lambda}$,

$$\xi_1 = \frac{\sum\limits_{j=1}^{J+J'} \sum\limits_{k=1}^{N_j} \lambda^T \left(\mathbf{Y}_{jk} + \mathbf{Y}^*_{jk}\right)}{\sum\limits_{j=1}^{J+J'} \sum\limits_{k=1}^{N_j} N_{jk}}, \tag{3.17}$$

where $\mathbf{Y}_{jk}$ is the observed vector of counts in PSU j and SSU k, and $\mathbf{Y}^*_{jk}$ is a realization of the unobserved counts in PSU j and SSU k. Recall that $\mathbf{Y}^*_{jk}$ is intended to accommodate two circumstances: the case in which SSU k of PSU j is not sampled at all. and the case in which SSU k of PSU j is selected. We generate each $\mathbf{Y}^*_{jk}$ according to the multinomial distribution in each case,

$$\mathbf{Y}^*_{jk} \sim \begin{cases} \text{Multinomial}(N_{jk} - n_{jk}, \boldsymbol{\theta}_{jk}) & j \leq J,\ k \leq K_j \\ \text{Multinomial}(N_{jk}, \boldsymbol{\theta}^*_{jk}) & \text{otherwise.} \end{cases} \tag{3.18}$$

For the sampled subclusters, we have seen n_{jk} observations already, and we simulate the unobserved responses of the remaining $N_{jk} - n_{jk}$ individuals. For those sampled SSU's we have posterior draws of $\boldsymbol{\theta}_{jk}$ from (3.16). The parameter vectors $\boldsymbol{\theta}^*_{jk}$ for the unsampled subclusters must be generated from the prior Dirichlet distribution. Once again, there are two cases: if the SSU in question is from a PSU that was sampled then we have some information about η_j and $\boldsymbol{\gamma}_j$; if the unsampled SSU is from an unsampled PSU the we sample $\boldsymbol{\theta}^*_{jk}$ from a Dirichlet distribution with parameters $\eta^*\boldsymbol{\gamma}^*$ that must themselves be generated from their posterior distributions. We generate posterior draws for $\boldsymbol{\theta}^*_{jk}$ as follows

$$\boldsymbol{\theta}^*_{jk} \sim \begin{cases} \text{Dirichlet}(\eta_j\boldsymbol{\gamma}_j) & j \leq J,\ k > K_j \\ \text{Dirichlet}(\eta^*\boldsymbol{\gamma}^*) & j > J \end{cases} \tag{3.19}$$

with

$$\begin{aligned} \eta^* &\sim \text{gamma}(m, \mu/m) \\ \boldsymbol{\gamma}^* &\sim \text{Dirichlet}(\kappa\boldsymbol{\alpha}) \end{aligned} \tag{3.20}$$

We also describe quantities related to the finite population proportion that are of some interest in understanding the model. The quantities ξ_2, ξ_3, ξ_4 are defined by analogy with the two-stage sample quantities of Chapter 2. The population weighted average of the probability assigned to the proportions of interest in each SSU is,

$$\xi_2 = \frac{\sum_{j=1}^{J+J'} \sum_{k=1}^{N_j} N_{jk} \boldsymbol{\lambda}^T \left(\boldsymbol{\theta}^*_{jk}\right)}{\sum_{j=1}^{J+J'} \sum_{k=1}^{N_j} N_{jk}}, \tag{3.21}$$

where we let $\boldsymbol{\theta}^*_{jk} = \boldsymbol{\theta}_{jk}$ for the sampled SSU's. Another quantity is obtained by replacing each SSU probability vector with its expectation. The expectation for the probability vector in an SSU depends on whether the SSU was sampled, unsampled but part of a sampled PSU or unsampled and in an unsampled PSU. Formally,

$$\xi_3 = \frac{\sum_{j=1}^{J+J'} \sum_{k=1}^{N_j} N_{jk} \boldsymbol{\lambda}^T E[\boldsymbol{\theta}^*_{jk}]}{\sum_{j=1}^{J+J'} \sum_{k=1}^{N_j} N_{jk}} \tag{3.22}$$

where

$$E[\boldsymbol{\theta}^*_{jk}] = \begin{cases} \left(\mathbf{Y}_{jk} + \eta_j \boldsymbol{\gamma}_j\right) / (n_{jk} + \eta_j) & \text{if } j \le J,\ k \le K_j \\ \boldsymbol{\gamma}_j & \text{if } j \le J,\ k > K_j \\ \boldsymbol{\gamma}^* & \text{if } j > J \end{cases} \tag{3.23}$$

Again $\boldsymbol{\gamma}^*$ is generated as in (3.20).

The final quantity of interest is that probability assigned to the proportion of interest by the probability vector at the highest level of the hierarchy,

$$\xi_4 = \boldsymbol{\lambda}^T \boldsymbol{\alpha}. \tag{3.24}$$

We compute these quantities later in reviewing the results obtained by analyzing the SPO data under the three-stage model.

3.3 Missing Data

In this section we extend our approach to accommodate unintentional missing data. The Slovenian Public Opinion Survey data consist of three Yes/No responses that are modeled as an eight cell multinomial response. One or more missing Yes/No responses corresponds to a partially observed response in the multinomial setting. As in Section 2.5 we assume that the unobserved part of the responses are missing at random (MAR). That is, we assume that the probability of a Yes/No response is unavailable doesn't depend upon the value that would have been observed. See Section 2.5 for a discussion of MAR, as well as alternatives to it.

Let ν be a pattern of missingness. A pattern of missingness identifies the value of the observed responses to the Yes/No questions and an indication of which questions were not answered (the Don't Know's). Next, let A_ν be the set of possible multinomial cells for observations with missingness pattern ν. For example, if $\mathrm{I} = 8$ and ν_1 has possible responses 1, 3, 5, 7, then $\mathrm{A}_{\nu_1} = \{1, 3, 5, 7\}$. Let n_{jk}^{ν} be the number of individuals in the k^{th} subcluster of the j^{th} cluster with pattern of missingness ν. Then Y_{ijk}^{ν} is the (unobserved) number of responses from the n_{jk}^{ν} individuals with missingness pattern ν from subcluster k of cluster j that (actually) fall in category i. Define $\mathrm{Y}_{ijk}^{\nu} = 0$ if $n_{jk}^{\nu} = 0$ or $i \notin \mathrm{A}_\nu$ and let $\mathbf{Y}_{jk}^{\nu} = (Y_{1jk}^{\nu}, Y_{2jk}^{\nu}, \ldots, Y_{Ijk}^{\nu})$. We take P to be the set of all patterns of missingness. Then we let $\mathbf{Y}^{mis} = \{\mathbf{Y}_{jk}^{\nu} : j = 1, \ldots, J, k = 1, \ldots, N_j, \nu \in P\}$ denote all of the missing data. Following the notation used in Section 2.5, we let $\mathbf{Y}^{obs}$ represent the marginal missing data totals, n_{jk}^{ν}, as well as the observed counts for each subcluster.

Under the MAR assumption for the missing data mechanism, we obtain the joint distribution

$$\begin{aligned}
&\mathrm{p}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\eta}, \boldsymbol{\alpha}, \kappa, \mu, m, \mathbf{Y}^{mis}, \mathbf{Y}^{obs}) \\
=\ & \mathrm{p}(\mathbf{Y}^{obs}, \mathbf{Y}^{mis} \mid \boldsymbol{\Theta})\mathrm{p}(\boldsymbol{\Theta} \mid \boldsymbol{\Gamma}, \boldsymbol{\eta})\mathrm{p}(\boldsymbol{\Gamma} \mid \boldsymbol{\alpha}, \kappa) \\
\times\ & \mathrm{p}(\boldsymbol{\alpha})\mathrm{p}(\kappa)\mathrm{p}(\boldsymbol{\eta} \mid \mu, m)\mathrm{p}(\mu)\mathrm{p}(m)
\end{aligned}$$

where individual distributions are given in Section 3.1. Having observed $\mathbf{Y}^{obs}$, we obtain the posterior distribution,

$$
\begin{aligned}
\mathrm{p}(\mathbf{Y}^{mis}, \mathbf{\Theta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \kappa, \mu, m \mid \mathbf{Y}^{obs}) \propto\ & \prod_{j=1}^{J}\prod_{k=1}^{K_j}\left\{\left\{\prod_{\nu\in P}\binom{n^{\nu}_{jk}}{\mathbf{Y}^{\nu}_{jk}}\right\}\prod_{i=1}^{I}\theta_{ijk}^{\left(Y_{ijk}+\sum_{\nu\in P}Y^{\nu}_{ijk}\right)}\right\} \\
\times\ & \prod_{j=1}^{J}\prod_{k=1}^{K_j}\left\{\Gamma(\eta_j)\prod_{i=1}^{I}\frac{\theta_{ijk}^{\gamma_{ij}\eta_j-1}}{\Gamma(\gamma_{ij}\eta_j)}\right\} \\
\times\ & \prod_{j=1}^{J}\left\{\Gamma(\kappa)\prod_{i=1}^{I}\frac{\gamma_{ij}^{\alpha_i\kappa-1}}{\Gamma(\alpha_i\kappa)}\right\} \\
\times\ & \prod_{j=1}^{J}\left(\frac{\eta_j m}{\mu}\right)^{m-1}\frac{e^{-m\eta_j/\mu}}{\mu/m\Gamma(m)} \\
\times\ & \left(\frac{1}{\mu c_1}\right)^{c_2-1}\frac{e^{-1/c_1\mu}}{c_1\Gamma(c_2)} \\
\times\ & m^{c_3-1}e^{-c_4 m} \\
\times\ & \kappa^{c_5-1}e^{-c_6\kappa}
\end{aligned}
\tag{3.25}
$$

with $(c_1, c_2, \ldots, c_6)$ constants.

As in the previous section, let $\boldsymbol{\beta}=(\mathbf{\Gamma}^T, \boldsymbol{\eta}^T, \boldsymbol{\alpha}^T, \kappa, \mu, m)^T$ represent all of the parameters other than $\mathbf{\Theta}$. Then we can factor (3.25) as

$$
\mathrm{p}(\mathbf{\Theta}, \boldsymbol{\beta}, \mathbf{Y}^{mis} \mid \mathbf{Y}^{obs}) = \mathrm{p}(\mathbf{\Theta} \mid \boldsymbol{\beta}, \mathbf{Y}^{mis}, \mathbf{Y}^{obs})\mathrm{p}(\boldsymbol{\beta}, \mathbf{Y}^{mis} \mid \mathbf{Y}^{obs}). \tag{3.26}
$$

Simulation from the first term is straightforward because we have complete data and conjugate priors. We focus on the second term and describe an MCMC approach to simulating from the marginal posterior distribution of $\boldsymbol{\beta}$ and $\mathbf{Y}^{mis}$.

The MCMC algorithm that we use for drawing samples from the posterior distribution, p($\boldsymbol{\beta}$, $\mathbf{Y}^{mis} \mid \mathbf{Y}$) can be thought of as a Gibbs sampling algorithm comprising two steps:

1. Simulate $\boldsymbol{\beta}$ from $\mathrm{p}(\boldsymbol{\beta} \mid \mathbf{Y}^{mis}, \mathbf{Y}^{obs})$;

2. Simulate $\mathbf{Y}^{mis}$ from $\mathrm{p}(\mathbf{Y}^{mis} \mid \boldsymbol{\beta}, \mathbf{Y}^{obs}) = \mathrm{p}(\mathbf{Y}^{mis} \mid \boldsymbol{\gamma}, \boldsymbol{\eta}, \mathbf{Y}^{obs})$.

In practice each of the steps itself requires one or more steps of a MCMC algorithm be carried out. To generate a sample for the first step we use a single step of the complete data MCMC algorithm of Section 3.2 because we are conditioning on the complete data $\mathbf{Y} = (\mathbf{Y}^{obs}, \mathbf{Y}^{mis})$. This means that within step 1 we carry out one-step of the random walk Metropolis algorithm for each component of $\boldsymbol{\beta}$. The conditional distribution in the second step, that of $\mathbf{Y}^{mis}$ given $\boldsymbol{\beta}$ and $\mathbf{Y}^{obs}$, actually depends only on $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$ and $\mathbf{Y}^{obs}$. The posterior of $\mathbf{Y}^{mis}$ given $\mathbf{Y}^{obs}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ is obtained, up to a constant of proportionality, by examining the relevant terms of (3.8)

$$
\begin{aligned}
&\mathrm{p}(\mathbf{Y}^{mis} \mid \boldsymbol{\gamma}, \boldsymbol{\eta}, \mathbf{Y}^{obs}) \\
&= \prod_{j=1}^{J} \prod_{k=1}^{K_j} \left\{ \left\{ \prod_{\nu \in P} \binom{n_{jk}^{\nu}}{\mathbf{Y}_{jk}^{\nu}} \right\} \frac{\Gamma(\eta_j)}{\Gamma\left(n_{jk} + \eta_j + \sum_{\nu \in P} n_{jk}^{\nu}\right)} \right. \\
&\qquad\qquad \left. \times \prod_{i=1}^{I} \frac{\Gamma\left(Y_{ijk} + \gamma_{ij}\eta_j + \sum_{\nu \in P} Y_{ijk}^{\nu}\right)}{\Gamma(\gamma_{ij}\eta_j)} \right\} \\
&\propto \prod_{j=1}^{J} \prod_{k=1}^{K_j} \left\{ \left\{ \prod_{\nu \in P} \binom{n_{jk}^{\nu}}{\mathbf{Y}_{jk}^{\nu}} \right\} \prod_{i=1}^{I} \Gamma\left(Y_{ijk} + \gamma_{ij}\eta_j + \sum_{\nu \in P} Y_{ijk}^{\nu}\right) \right\} \qquad (3.27)
\end{aligned}
$$

The final expression for the posterior distribution of $\mathbf{Y}^{mis}$, (3.27), is a product kernels of Dirichlet-Multinomial distributions for each pattern of missingness/subcluster/cluster combination. For example, consider missingness pattern ν^* for cluster j, and subcluster k. Generation of the vector $\mathbf{Y}_{jk}^{\nu^*}$ is equivalent to allocating the $n_{jk}^{\nu^*}$ observations to the cells in A_{ν^*}. Recall $Y_{ijk} = 0$ for any i not in A_{ν^*} by definition. Then $\mathrm{p}(\mathbf{Y}_{jk}^{\nu^*} \mid \mathbf{Y}^{obs}, \gamma_j, \eta_j)$ is

$$
\binom{n_{jk}^{\nu^*}}{\mathbf{Y}_{jk}^{\nu^*}} \prod_{i \in A_{\nu^*}} \Gamma\left(Y_{ijk}^{\nu^*} + Y_{ijk} + \gamma_{ij}\eta_j + \sum_{\substack{\nu \in P \\ \nu \neq \nu^*}} Y_{ijk}^{\nu^*}\right) \qquad (3.28)
$$

which is the kernel of Dirichlet-Multinomial distribution with parameters $n_{jk}^{\nu^*}$ and $Y_{ijk} + \gamma_{ij}\eta_j + \sum_{\substack{\nu \in P \\ \nu \neq \nu^*}} Y_{ijk}^{\nu^*}$. We sample from this distribution by generating a probability vector of dimension $\mid A_{\nu^\bullet} \mid$ from a Dirichlet distribution with parameters $Y_{ijk} + \gamma_{ij}\eta_j + \sum_{\substack{\nu \in P \\ \nu \neq \nu^\bullet}} Y_{ijk}^{\nu^\bullet}$ for $i \in A_{\nu^\bullet}$. Then we sample from a multinomial distribution with sample size n_{jk}^{ν} using the probability vector generated from the Dirichlet distribution. We can draw samples from the posterior given in (3.27) by cycling through all the patterns of missingness for each subcluster of each cluster. We then have a realization of fully observed responses for each cluster and each subcluster, which allows us to return to the first step of our basic algorithm.

Summarizing we use a MCMC procedure to get samples from the joint posterior distribution of $\boldsymbol{\beta}$ and $\mathbf{Y}^{mis}$ given the fully observed $\mathbf{Y}^{obs}$. Once realizations from that distribution have been drawn, realizations from the posterior distribution of $\boldsymbol{\Theta}$ can be obtained. Simulation based inference proceeds as in the complete data case.

3.4 Simulations

In order to better understand the nature of the three-stage hierarchical model, we created five simulated data sets. We simulated multinomial data with four possible responses. Each simulated data set contained four PSU's and three SSU's within each PSU. This small size facilitated looking at posterior distributions later on. The number of observations within each SSU was fixed at 40. For each of five simulated data sets we used the same vector of superpopulation proportions, $\boldsymbol{\alpha} = (0.15, 0.25, 0.20, 0.40)^T$. We varied $\boldsymbol{\eta}$ and κ to create different scenarios for within and between PSU variation. In each scenario, we generated a single sample following the hierarchical model described in Section 3.1.

Each simulated data set is analyzed using the MCMC approach of Section 3.2.2. For purposes of analysis we fixed $c_1 = 0.005, c_2 = 3, c_3 = 0.1, c_4 = 0.01, c_5 = 0.1, c_6 = 0.01$ in the prior distributions for m, μ, κ. This gives a prior distribution for m with mean 10 and variance 1000, for μ with mean 10 and variance 1000 and for κ with mean 10 and variance 1000. In fact the distribution used to generate the values of κ and $\boldsymbol{\eta}$ which generate the simulated data were not always consistent with this prior distribution. We return to this point later. The results given for each simulated data set are based on the last 3000 iterations of four MCMC chains of total length 4000 iterations each: convergence was determined based using the PSR discussed in Section 2.4.

We analyze five different simulated data sets. The first four simulations represent a form of factorial design on the parameters κ and $\boldsymbol{\eta}$. We considered high and low values of κ and $\boldsymbol{\eta}$ in each possible combination. High levels of κ (or η_j) indicate that the cluster (or subcluster) proportions are similar to each other. Likewise, low levels of κ (or η_j) indicate that there is considerable variability among the proportion vectors for different clusters (or subclusters). The final simulation considers the extreme case in which all SSU's are identical. That case, corresponding to infinitely large κ and $\boldsymbol{\eta}$, is equivalent to sampling data from a single multinomial distribution — the clustering contains no information.

Table 3.1 Data for Simulation 1

		Counts			
PSU	SSU	Y_{1jk}	Y_{2jk}	Y_{3jk}	Y_{4jk}
1	1	0	9	7	24
1	2	0	4	7	29
1	3	0	18	0	22
2	1	20	0	4	16
2	2	25	6	9	0
2	3	9	12	9	10
3	1	3	9	2	26
3	2	7	11	5	17
3	3	3	9	2	26
4	1	0	15	16	9
4	2	1	11	16	12
4	3	1	12	13	14

Table 3.2 Results for Simulation 1

					Posterior	
	Simulated		Posterior	Percentiles		
Parameter	Value	Data	Mean	2.5^{th}	50^{th}	97.5^{th}
α_1	0.150	0.144	0.115	0.037	0.108	0.243
α_2	0.250	0.242	0.271	0.129	0.269	0.417
α_3	0.200	0.188	0.208	0.096	0.212	0.351
κ	8.785		7.452	2.624	7.417	22.442
γ_{11}	0.025	0.000	0.006	0.001	0.009	0.080
γ_{12}	0.075	0.258	0.254	0.136	0.256	0.423
γ_{13}	0.142	0.117	0.125	0.046	0.127	0.292
η_1	9.995		12.052	1.762	11.343	90.133
γ_{21}	0.340	0.450	0.351	0.121	0.368	0.584
γ_{22}	0.206	0.150	0.160	0.052	0.167	0.334
γ_{23}	0.253	0.183	0.215	0.089	0.220	0.388
η_2	11.440		5.656	1.589	5.695	9.890
γ_{31}	0.098	0.108	0.103	0.050	0.105	0.181
γ_{32}	0.289	0.242	0.245	0.165	0.246	0.340
γ_{33}	0.143	0.075	0.086	0.038	0.088	0.165
η_3	10.061		89.185	16.873	77.857	995.196
γ_{41}	0.046	0.017	0.023	0.009	0.024	0.079
γ_{42}	0.323	0.317	0.306	0.190	0.310	0.414
γ_{43}	0.377	0.375	0.356	0.240	0.360	0.476
η_4	10.505		87.716	13.948	82.666	715.771
μ	10.000		69.007	25.808	61.537	377.862
m	100.000		1.499	0.135	0.629	3.700

Simulation 1

The first simulated data set is show in in Table 3.1. For this data set we chose both κ and η to be small; the elements of κ and η are random draws from a gamma distribution with mean 10 and variance 1 (equivalent to choosing $\mu = 10$ and $m = 100$). Note that the PSU proportions appear quite variable as expected with a small κ. The first PSU has large entries for the fourth column and no observations in the first column. Whereas the second PSU has a reasonably high frequency in the first column. Table 3.2

presents the results of the posterior analysis of the first simulated data set. The first column is the simulated value that was used to create the data set. Naturally, $\boldsymbol{\eta}$ and κ are all close to 10. The second column labeled "Data" is gives the sample proportions for each PSU (which correspond to the parameters $\boldsymbol{\gamma}$) and for the entire population (which corresponds to the parameter $\boldsymbol{\alpha}$). For the most part the posterior distribution summarizes the data well for this simulation. In particular the simulated proportions, with the exceptions of γ_{12} , are all inside the central 95% posterior interval. This is the interval formed by the 2.5^{th} and 97.5^{th} percentiles. The posterior credible sets for η_2, η_3, and η_4 do not contain the simulated values for these parameters. Finally the model does a poor job of modeling the hyperparameters μ and m.

We first comment on γ_{12}. The data that we observed is quite surprising for the first PSU. For a Beta-binomial distribution with $n = 120$, $\alpha = 0.74$ and $\beta = 0.92$, using a Gaussian approximation, we find that we should expect about 95% of all realized proportions to fall between 0.000 and 0.125 or counts between 0 and 15. The proportion that was observed falls quite far outside this interval, 0.258 or 31 observations. Consequently, it is not surprising that the 95% posterior credible set did not contain the simulated value. Turning to the elements of $\boldsymbol{\eta}$, we first note that the data in PSU's 3 and 4 are quite similar across SSU's. Thus the data are consistent with larger values for η_3 and η_4. The posterior intervals are wide reflecting the fact that inference for each η_j is based only on three SSU's. It is likely that having large estimated values for η_3 and η_4 force the posterior inference for μ to concentrate on larger values and the resulting variability in the distribution of η_j's leads to small values of m. A key point appears to be that given the small number of PSU's and SSU's the "vague" prior distribution may be leading to some of the poor performance observed here. This will be addressed in future work.

The results also demonstrate the typical behavior for parameter estimates under the hierarchical model. Posterior means for the $\boldsymbol{\gamma}$'s are a compromise between the data from the specific PSU, (listed under Data) and the total population (information contained in α). Thus the posterior mean for γ_{21} is a compromise between the sample proportion for that PSU (0.450) and the overall proportion in category 1 (0.144). Because the model suggests considerable variability among PSU's the compromise is weighted toward the data from PSU 2.

Table 3.3 Data for Simulation 2

		Counts			
PSU	SSU	Y_{1jk}	Y_{2jk}	Y_{3jk}	Y_{4jk}
1	1	7	5	7	21
1	2	0	12	2	26
1	3	2	7	9	22
2	1	3	12	10	15
2	2	8	12	9	11
2	3	2	6	12	20
3	1	5	9	9	17
3	2	6	11	6	17
3	3	3	14	12	11
4	1	6	11	6	17
4	2	8	6	8	18
4	3	6	12	17	5

Simulation 2

The data for the second simulation, found in Table 3.3, is quite different from the first. In the first, we simulated small values for the superpopulation parameters κ and the η_j's. Here we simulate larger values for these parameters, all are approximately equal to 100. The data are much more consistent when we consider the SSU's within a single PSU or when we compare PSU's. The posterior analysis reflects this change. The 95% posterior intervals contain the simulated values for all of the model parameters except γ_{31} and m. For γ_{31}, the posterior distribution matches the simulated data quite closely but the data are far from the true parameters. This is not unexpected with so many

Table 3.4 Results for Simulation 2

					Posterior	
	Simulated		Posterior		Percentiles	
Parameter	Value	Data	Mean	2.5^{th}	50^{th}	97.5^{th}
α_1	0.150	0.117	0.118	0.073	0.118	0.181
α_2	0.250	0.243	0.250	0.185	0.249	0.326
α_3	0.200	0.223	0.223	0.160	0.224	0.294
κ	104.685		74.183	17.319	74.704	313.856
γ_{11}	0.128	0.075	0.088	0.041	0.091	0.151
γ_{12}	0.205	0.200	0.230	0.154	0.232	0.313
γ_{13}	0.179	0.150	0.187	0.117	0.188	0.272
η_1	100.107		51.459	14.076	54.707	139.923
γ_{21}	0.124	0.108	0.110	0.062	0.112	0.177
γ_{22}	0.270	0.250	0.248	0.172	0.249	0.329
γ_{23}	0.220	0.258	0.240	0.175	0.239	0.325
η_2	99.774		65.168	23.274	65.453	179.586
γ_{31}	0.059	0.117	0.116	0.069	0.117	0.179
γ_{32}	0.247	0.283	0.266	0.193	0.267	0.354
γ_{33}	0.260	0.225	0.223	0.151	0.225	0.300
η_3	96.608		69.607	26.000	68.213	214.595
γ_{41}	0.169	0.167	0.141	0.087	0.142	0.219
γ_{42}	0.252	0.242	0.247	0.177	0.247	0.342
γ_{43}	0.247	0.258	0.238	0.166	0.238	0.324
η_4	96.631		54.405	15.488	56.721	154.882
μ	100.000		64.886	28.920	65.009	144.178
m	1000.000		14.859	0.806	15.183	225.576

parameters. For the posterior distribution of m, recall that

$$Var[\eta_j \mid \mu, m] = \mu^2/m. \tag{3.29}$$

Consequently, the value of m models the amount of variability in the η_j's. Since the posterior distributions for the η_j's reflect much more variability than we would expect, it is, perhaps, not surprising that the posterior distribution of m does not contain the simulated value 1000. Once again the small number of PSU's is likely responsible.

Simulation 3

Table 3.5 Data for Simulation 3

		Counts			
PSU	SSU	Y_{1jk}	Y_{2jk}	Y_{3jk}	Y_{4jk}
1	1	3	12	4	21
1	2	0	8	3	29
1	3	3	22	5	10
2	1	7	9	11	13
2	2	7	8	11	14
2	3	17	6	3	14
3	1	0	5	11	24
3	2	1	13	0	26
3	3	0	6	3	31
4	1	3	14	11	12
4	2	4	13	8	15
4	3	8	6	8	18

For the third simulation we chose value for κ to be small, (a sample from a distribution with mean 10) and the η_j's to be large (sampled from a distribution with mean 100). The data are given in Table 3.5. The data exhibit great variability across PSU's but little variability among SSU's within a single PSU. The results of this analysis can be found in Table 3.6. The posterior 95% intervals for all but one of the parameters include the simulated values. The only parameter that was not included in the 95% credible set was η_3. These results highlight the benefit of having independent prior distributions on κ and the η_j's. In this case there is a clear difference between the small value of κ and the larger values of the η's. A common joint distribution for κ and the η's might not allow for this possibility. Once again posterior inference for m is poor as the posterior interval is wide, though for this data set it does contain the true value. which tend to be smaller and those for the η's which are generally larger.

Simulation 4

For the fourth simulation we reversed the scenario of simulation 3. We chose a large

Table 3.6 Results for Simulation 3

					Posterior	
	Simulated		Posterior		Percentiles	
Parameter	Value	Data	Mean	2.5^{th}	50^{th}	97.5^{th}
α_1	0.150	0.110	0.112	0.053	0.112	0.203
α_2	0.250	0.254	0.256	0.161	0.257	0.376
α_3	0.200	0.162	0.173	0.096	0.176	0.272
κ	8.600		23.473	6.738	22.667	103.207
γ_{11}	0.059	0.050	0.068	0.025	0.070	0.151
γ_{12}	0.316	0.350	0.316	0.214	0.318	0.430
γ_{13}	0.061	0.100	0.130	0.068	0.131	0.227
η_1	41.158		27.009	5.986	29.340	94.543
γ_{21}	0.208	0.258	0.206	0.119	0.210	0.312
γ_{22}	0.232	0.192	0.213	0.132	0.214	0.319
γ_{23}	0.178	0.208	0.195	0.120	0.197	0.292
η_2	98.564		43.962	15.390	40.954	193.569
γ_{31}	0.011	0.008	0.036	0.006	0.039	0.137
γ_{32}	0.203	0.200	0.229	0.128	0.228	0.378
γ_{33}	0.108	0.117	0.121	0.051	0.124	0.228
η_3	116.549		18.609	2.268	23.899	67.538
γ_{41}	0.157	0.125	0.122	0.063	0.123	0.207
γ_{42}	0.321	0.275	0.264	0.182	0.265	0.367
γ_{43}	0.207	0.225	0.210	0.129	0.213	0.306
η_4	116.758		49.379	17.223	43.278	292.841
μ	100.000		45.887	22.631	43.336	116.693
m	100.000		4.812	0.254	3.672	150.847

value for κ and small values for the elements of $\boldsymbol{\eta}$. This allows for great variability among SSU's within a PSU, but somehow little variation among PSU totals. Table 3.7 contains the data for this simulation. The results of the posterior analysis can be found in Table 3.8. In this instance, 95% posterior intervals contain the true simulated values for all of the proportions $\boldsymbol{\alpha}$ and $\boldsymbol{\Gamma}$, as well as for $\boldsymbol{\eta}$ and κ. However, the model does not do well in estimating the simulated values for μ and m. This is especially surprising for μ, since μ is the expected value of the η's. Here it is not close to the average of

Table 3.7 Data for Simulation 4

			Counts		
PSU	SSU	Y_{1jk}	Y_{2jk}	Y_{3jk}	Y_{4jk}
1	1	17	0	7	16
1	2	3	2	20	15
1	3	1	6	22	11
2	1	3	13	19	5
2	2	2	9	2	27
2	3	3	23	11	3
3	1	3	17	7	13
3	2	0	11	6	23
3	3	10	4	3	23
4	1	12	3	1	24
4	2	1	6	5	28
4	3	11	18	3	8

the posterior mean of the η's. Once again the best current explanation is that the prior distribution is not completely dominated by the small number of PSU's providing data.

Simulation 5

For the final simulation we chose to simulate and analyze data in which all 12 SSU's have exactly the same underlying probability vector. This corresponds to infinite values of κ and $\boldsymbol{\eta}$ which yield Dirichlet distributions with no variability. We analyzed the data as if it were generated in the same manner as the other simulations, via a three-stage cluster sample with 4 PSU's and 3 SSU's within each PSU. The data for this simulation is found in Table 3.9. The results of the posterior analysis are found in Table 3.10. We first note that the true values for proportions are contained in the 95% posterior intervals for $\boldsymbol{\alpha}$ and the $\boldsymbol{\gamma}_j$'s. The posterior means for κ and the η's is larger than we observed in any of the other simulations. The large posterior values for κ and for the η's clearly suggest that there is great consistency within clusters. If we increase the sample

Table 3.8 Results for Simulation 4

					Posterior	
	Simulated		Posterior		Percentiles	
Parameter	Value	Data	Mean	2.5^{th}	50^{th}	97.5^{th}
α_1	0.150	0.110	0.148	0.078	0.150	0.248
α_2	0.250	0.233	0.232	0.123	0.236	0.342
α_3	0.200	0.220	0.207	0.117	0.209	0.333
κ	105.615		42.309	9.799	40.624	212.801
γ_{11}	0.131	0.175	0.146	0.069	0.148	0.265
γ_{12}	0.205	0.067	0.165	0.060	0.174	0.316
γ_{13}	0.248	0.408	0.270	0.142	0.271	0.444
η_1	10.718		6.596	1.757	6.445	25.148
γ_{21}	0.097	0.067	0.132	0.055	0.136	0.247
γ_{22}	0.273	0.375	0.285	0.123	0.292	0.448
γ_{23}	0.257	0.267	0.218	0.114	0.219	0.368
η_2	9.309		6.914	2.708	6.817	18.258
γ_{31}	0.089	0.108	0.116	0.047	0.120	0.222
γ_{32}	0.265	0.267	0.241	0.112	0.248	0.364
γ_{33}	0.239	0.133	0.179	0.094	0.179	0.311
η_3	10.850		14.546	3.815	14.472	58.304
γ_{41}	0.210	0.200	0.160	0.072	0.164	0.290
γ_{42}	0.250	0.225	0.226	0.115	0.230	0.363
γ_{43}	0.171	0.075	0.161	0.073	0.163	0.295
η_4	10.378		7.490	2.269	7.662	20.682
μ	10.000		39.021	16.417	38.085	103.908
m	100.000		0.723	0.145	0.732	3.463

size within each SSU, then we would expect the posterior means for these quantities to increase even further.

To summarize these simulations we note that the posterior distribution for each simulation modeled well the superpopulation from which it was generated. In particular the posterior summaries of the proportions were accurate reflections of the data sets they were meant to describe. That is, the model performed well in balancing the hyperprior distribution for the population and the PSU level data when making inference for an

Table 3.9 Data for Simulation 5

		Counts			
PSU	SSU	Y_{1jk}	Y_{2jk}	Y_{3jk}	Y_{4jk}
1	1	7	13	6	14
1	2	7	9	10	14
1	3	2	13	9	16
2	1	5	14	6	15
2	2	4	13	7	16
2	3	5	7	12	16
3	1	8	11	7	14
3	2	9	7	7	17
3	3	8	9	6	17
4	1	5	13	6	16
4	2	6	8	11	15
4	3	4	12	9	15

observed PSU.

One major drawback to these simulations is the small number of PSU's four. That small number of PSU's provides little information about the hyperparameters of the model. Consequently the posterior intervals for these quantities often did not include the simulated value. Additional simulations with more PSU's is one way to address this. An alternative is to consider a more informative prior on m when the number of PSU's is small.

3.5 Application: The Slovenian Public Opinion Survey

In this section, we apply the methodology of this chapter to the 1990 Slovenian Public Opinion (SPO) survey that was described in Section 2.6. Recall that the SPO is a three-stage cluster sample. Of the 1000 clusters, or primary sampling units (PSU), 139 were sampled and then 3 of 16 secondary sampling units (SSU) were chosen within each PSU. Finally 5 individuals were chosen from the approximately 100 individuals within

Table 3.10 Results for Simulation 5

					Posterior	
	Simulated		Posterior		Percentiles	
Parameter	Value	Data	Mean	2.5^{th}	50^{th}	97.5^{th}
α_1	0.150	0.145	0.149	0.107	0.149	0.197
α_2	0.250	0.268	0.267	0.212	0.268	0.329
α_3	0.200	0.200	0.198	0.148	0.199	0.258
κ			128.722	33.525	131.940	419.757
γ_{11}		0.133	0.140	0.093	0.141	0.199
γ_{12}		0.292	0.274	0.207	0.275	0.348
γ_{13}		0.208	0.203	0.147	0.203	0.276
η_1			171.537	54.764	167.773	660.918
γ_{21}		0.117	0.134	0.088	0.135	0.189
γ_{22}		0.283	0.274	0.214	0.274	0.344
γ_{23}		0.208	0.199	0.142	0.200	0.267
η_2			167.753	52.651	164.125	650.558
γ_{31}		0.208	0.171	0.120	0.172	0.234
γ_{32}		0.225	0.248	0.185	0.250	0.315
γ_{33}		0.167	0.183	0.129	0.185	0.249
η_3			179.377	57.831	173.183	793.118
γ_{41}		0.125	0.140	0.093	0.140	0.201
γ_{42}		0.275	0.267	0.206	0.267	0.338
γ_{43}		0.217	0.205	0.144	0.205	0.270
η_4			174.213	54.409	166.285	692.557
μ			174.615	65.821	169.783	568.676
m			14.429	0.781	17.353	172.238

each SSU. Like most large scale surveys, not all of the selected individuals responded. However, 2074 of 2085 or over 99% did; this number does include some substitutes. More details on the SPO can be found in Section 2.6 of this dissertation. In the three-stage analysis, only 138 of the 139 sample PSU's were considered because the SSU's for one of the PSU's were not correctly labeled. Since represents a small proportion of the data, deleting that PSU should have little effect on the overall results.

We report results for the population level quantities ξ_1, ξ_2, ξ_3, and ξ_4 (defined in

Section 3.2.3) as we did in Section 2.6. Here we also consider the cluster-level parameters of the model as quantities of interest. For simplicity, we focus on a subset of the 138 PSU's. The data for the subset of clusters that we consider are listed in Table 3.11. Table 3.12 contains brief explanations for why these particular clusters were chosen.

3.5.1 Complete data results

We ran 4 MCMC chains of length 2000 using the algorithm of Section 3.2.2. We then used the methodology of Gelman and Rubin to assess whether the last half of these chains could be taken as draws from the posterior distribution of p($\boldsymbol{\beta} \mid \mathbf{Y}$), where $\boldsymbol{\beta}$ contains all of the parameters except $\boldsymbol{\Theta}$. To complete the Monte Carlo draws from the posterior distribution, we next sampled from p($\boldsymbol{\Theta} \mid \mathbf{Y}, \boldsymbol{\beta}$). For these realizations, we used each $\boldsymbol{\beta}$ draw to generate a complete set of $\boldsymbol{\theta}_{jk}$'s. Summaries of the posterior distributions for the population proportions $\boldsymbol{\alpha}$ are contained in Table 3.13 along with results for the parameters κ, μ and m. As we might expect they are quite similar to the results in Section 2.6. Focusing on the posterior mean we notice only slight differences in the proportions for the two-stage and three-stage models. The variability of the posterior distribution as measured by the widths of the 95% posterior intervals is also quite similar, though there seems to be slightly more variability in the three-stage model. The survey proportions for the eight categories are $(0.010, 0.000, 0.005, 0.006, 0.047, 0.006, 0.109, 0.819)^T$, which correspond to the following responses $(NNN, NNY, NYN, NYY, YNN, YNY, YYN, YYY)^T$ to the questions concerning Attendance, Independence and Succession, respectively.

The posterior mean for each element of $\boldsymbol{\alpha}$ is within a half percent of the survey total. This discrepancy is likely due choosing a non-informative prior distribution for the proportions. The posterior distribution for κ is concentrated on large values suggesting that the overall proportions for the PSU's are similar. The posterior distribution of μ which measures the expected values of the η_j's is also concentrated on large values. This means that SSU's within a single PSU tend to be relatively homogeneous. But the

Table 3.11 SSU Totals for Several PSU's

		Results for SPO questions concerning Attendance, Independence, Secession							
PSU	SSU	NNN	NNY	NYN	NYY	YNN	YNY	YYN	YYY
1	1	0	0	0	0	0	0	1	3
1	2	0	0	0	0	0	0	1	3
1	3	0	0	1	0	0	0	1	3
3	1	0	0	0	0	0	0	0	4
3	2	0	0	0	0	0	0	0	4
3	3	0	0	0	0	0	0	0	2
5	1	0	0	0	0	1	0	0	2
5	2	0	0	0	0	1	0	1	2
5	3	0	0	0	0	0	0	1	2
8	1	0	0	0	0	0	0	1	2
8	2	0	0	0	0	0	1	0	2
8	3	0	0	0	0	0	0	0	4
9	1	0	0	0	0	0	0	1	2
9	2	0	0	0	0	0	0	0	3
9	3	0	0	0	0	0	0	0	4
28	1	0	0	0	0	0	0	0	5
28	2	0	0	0	0	0	0	0	6
28	3	0	0	0	0	0	0	0	4
40	1	0	0	0	0	1	0	2	1
40	2	1	0	0	1	0	0	0	2
40	3	0	0	1	0	0	1	0	2

Note: the last two columns represent the number of individuals who said that they would attend and vote yes.

Table 3.12 Reasons for Choosing Selected PSU's

PSU	Reasoning
1	Two of the three subclusters are identical
3	Only YYY responses
5	Two responses that would count as No's in the plebiscite
8	PSU totals similar to overall population totals
9	All responses either YYN or YYY
28	All YYY responses(and no nonrespondents)
40	At least one response in each category except NNY

small value of m indicates that there is a large amount of variability in the η_j's with some exhibiting considerable heterogeneity. For several of these parameters the PSR was larger than 1.2. A large value of PSR for a single parameter indicates that the MCMC algorithm may not have converged to the "target" posterior distribution. Consequently, we treat the results presented here as tentative.

We now focus on the cluster level parameters. We report only the posterior means for these parameters though full posterior distributions are available. Table 3.14 displays the posterior means for γ_j and η_j for the seven PSU's identified earlier. The first row of Table 3.14 gives the posterior means for the population proportion $\boldsymbol{\alpha}$. The last entry of the first row is the posterior mean for κ. The remaining rows correspond to several PSU's identified in Tables 3.11 and 3.12. The key point about the results in Table 3.14 is that each row of the table is a weighted averages of population proportions ($\boldsymbol{\alpha}$) and the data for that particular PSU. For example the sample proportions for the data in PSU 1, see Table 3.11, were larger in the 3^{rd} and 7^{th} responses than for the population as a whole. Specifically, one of the thirteen responses (0.077) gave the response NYN compared to

Table 3.13 Summaries of the Posterior Distributions for Model Parameters

	Posterior	Percentiles		
	Mean	2.5^{th}	50^{th}	97.5^{th}
α_1	0.012	0.008	0.012	0.017
α_2^c	0.001	0.000	0.001	0.002
α_3^c	0.007	0.004	0.007	0.013
α_4^c	0.008	0.005	0.008	0.012
α_5^c	0.050	0.037	0.052	0.061
α_6^c	0.007	0.005	0.007	0.011
α_7	0.113	0.097	0.113	0.129
α_8	0.802	0.790	0.810	0.832
κ^c	181.235	125.647	180.939	236.344
μ	156.990	67.704	143.416	431.160
${}^c m$	0.546	0.376	0.544	0.816

^{c} indicates a PSR > 1.2.

less than 0.01 in the entire sample. As a consequence the posterior mean for γ_{13}, the posterior proportion of respondents choosing NYN, is higher in those categories than the population level proportions, α_3. The actual posterior mean is heavily weighted toward the population parameter because the PSU 1 estimate is based on only 13 observations.

Across the seven PSU's we find that the posterior inference at the PSU level is a balance between the population as a whole and the data in each PSU, with the balance heavily weighted towards the population. There are two contributing factors. First the small sample size within each PSU means there is considerable uncertainty about the proportions in the PSU if we rely only on data from that PSU. Instead the hierarchical model allows the inference to "borrow strength" from data in other PSU's. Second, as noted earlier it appears that κ is large and that PSU's are quite similar. This also supports the notion of each PSU borrowing strength from the others.

Table 3.14 Posterior Means of Probability Vectors for Selected Clusters

	Posterior Proportions by Category								
	γ_{1j}	γ_{2j}	γ_{3j}	γ_{4j}	γ_{5j}	γ_{6j}	γ_{7j}	γ_{8j}	
PSU	NNN	NNY	NYN	NYY	YNN	YNY	YYN	YYY	η_j
α	0.011	0.001	0.006	0.007	0.047	0.006	0.108	0.814	181.235
1	0.007	0.000	0.008	0.004	0.042	0.004	0.115	0.820	^{c}130.408
3	0.008	0.000	0.003	0.004	0.042	0.002	0.096	0.838	^{c}52.615
5	0.008	0.000	0.003	0.004	0.052	0.004	0.111	0.818	^{c}170.379
8	0.008	0.000	0.003	0.004	0.047	0.004	0.106	0.828	^{c}98.228
9	0.008	0.000	0.003	0.004	0.042	0.004	0.106	0.833	^{c}45.126
28	0.008	^{c}0.000	0.003	0.004	0.041	0.004	0.100	0.840	^{c}133.929
40	0.013	^{c}0.000	0.008	0.009	0.047	0.008	0.109	0.805	^{c}90.577

^{c} indicates a PSR > 1.2

Finally as the last analysis of this section we consider some of the finite population quantities of interest for the SPO. These quantities are defined and described in Section 3.2.3. Table 3.15 gives summaries for the posterior distributions of these quantities. The 95% posterior intervals for the quantities ξ_1, ξ_2, ξ_3 and ξ_3 are almost identical, as are their posterior means. These intervals are all slightly smaller than the design-based confidence interval for the population proportion. The design-based 95% confidence interval has a center closer to that for the population as a whole, 0.927. We attribute this difference to the "non-informative" prior distribution placed on the population proportion, α. The Dirichlet (1,...,1) prior distribution has a prior mean of 0.250 on the response categories of interest, so some modest shrinkage occurs. This prior is also likely part of the explanation for the smaller variability of the ξ's relative to the design-based inference.

3.5.2 Missing data results

Here we present the results for the three-stage model which incorporates the partially observed responses. As with the analysis that included only the fully observed responses, we ran three MCMC chains for 2000 iterations and assessed the convergence of these

Table 3.15 Proportions of Slovenian Voters Who Intend to Attend the Plebiscite and to Vote for Independence: Three-stage Results

	Posterior		Percentiles	
Parameter	Mean	2.5^{th}	50^{th}	97.5^{th}
ξ_1	0.917	0.902	0.918	0.931
ξ_2	0.918	0.904	0.918	0.931
ξ_3	0.919	0.905	0.919	0.932
ξ_4	0.921	0.908	0.921	0.933
Design-based	0.927	0.900	0.927	0.954

chains using the methodology of Gelman and Rub'n. The difference is that each MCMC step includes a draw from the posterior distribution of $\mathbf{Y}^{mis}$, the partially observed responses. Once samples from p($\mathbf{Y}^{mis}, \boldsymbol{\beta} \mid \mathbf{Y}^{obs}$) are obtained, we generate samples from the posterior distribution of the remaining parameter $\boldsymbol{\Theta}$ using the distribution p($\boldsymbol{\Theta} \mid \mathbf{Y}^{mis}, \boldsymbol{\beta}, \mathbf{Y}^{obs}$). Using these sampled values from the posterior, we can make inference about the quantities of interest defined in Section 3.2 of this chapter.

Table 3.16 contains the results for the parameter $\boldsymbol{\alpha}$ which represents the mean of the PSU level probability vectors, and the associated hyperparameters κ, μ, and m. As in the two-stage model there are differences between the analysis of the completely observed data and the analysis that also incorporates the partially observed data. The most noticeable of these is the difference in the population level proportions, $\boldsymbol{\alpha}$. As with the two-stage model there is an upward shift of the posterior intervals for the first seven proportions which include No as one of their responses. The distribution for the 8^{th} probability which corresponds to Yes responses for all three questions decreases. It should also be noted that for the analysis incorporating the missing data there is a slight increase in the posterior mean (and quantiles of) κ and a slight decrease in the posterior mean of μ, and a slight increase in the posterior mean for m. As for the changes in

Table 3.16 Summaries of Posterior Distribution for Model Parameters

	Posterior		Percentiles	
	Mean	2.5^{th}	50^{th}	97.5^{th}
α_1	0.027	0.022	0.027	0.033
α_2^c	0.005	0.004	0.005	0.008
α_3^c	0.008	0.005	0.009	0.011
α_4^c	0.010	0.007	0.010	0.014
α_5^c	0.068	0.051	0.075	0.084
α_6^c	0.008	0.005	0.009	0.014
α_7	0.128	0.117	0.127	0.142
α_8	0.744	0.717	0.744	0.775
κ^c	214.090	137.550	214.556	403.328
μ	50.543	27.150	48.769	117.626
m^c	1.626	0.674	1.603	3.649

^{c} indicates a PSR > 1.2

the distributions of κ, μ and m, we can infer that the cluster-level probability vectors are more similar (larger κ), while the subcluster-level probability vectors are less similar (smaller μ) for the missing data analysis than in the complete data analysis.

The cluster-level parameter posterior means are given in Table 3.17. As we saw with the complete data case, the cluster-level proportions are weighted averages of the data in that cluster and the population proportions. In this case since κ is larger relative to the η's, the estimated cluster-level proportions are weighted quite heavily toward the population proportions. The variability that is incorporated through the partially observed data heavily affects the values for each η_j; the posterior mean of η_j is lower for each of the seven PSU's when the incomplete cases are incorporated. Recall that a small η implies more variability among the subcluster proportions within a cluster. One possible explanation is that each partially observed response can add additional variability to the subcluster estimates.

Table 3.17 Results for Selected Clusters

Posterior Means for Probability Vectors for Selected Clusters

	γ_{1j}	γ_{2j}	γ_{3j}	γ_{4j}	γ_{5j}	γ_{6j}	γ_{7j}	γ_{8j}	
PSU	NNN	NNY	NYN	NYY	YNN	YNY	YYN	YYY	η_j
α	0.027	0.005	0.008	0.010	0.068	0.008	0.128	0.744	214.090
1	0.024	0.004	0.010	0.010	0.062	0.007	0.132	0.751	47.562
3	0.024	0.003	0.006	0.007	0.062	0.006	0.121	0.771	24.707
5	0.024	0.004	0.006	0.008	^{c}0.075	0.007	0.130	0.746	40.665
8	0.024	0.003	0.006	0.007	0.067	0.006	0.124	0.763	35.373
9	0.023	0.003	0.005	0.007	0.063	^{c}0.008	0.125	0.766	35.006
28	0.024	0.003	0.007	0.006	0.070	^{c}0.006	0.129	0.755	25.752
40	0.026	0.003	^{c}0.011	0.013	0.069	^{c}0.010	0.127	0.741	43.856

^{c} indicates a PSR > 1.2

Turning to the quantities, ξ_1, ξ_2, ξ_3 and ξ_4, we observe that the posterior distribution for each is nearly identical. This was also the case for the analysis of the fully observed data. The 95% posterior intervals for these quantities exhibit more variability than the equivalent quantities for the complete data. The addition of the partially observed responses shifts the center of the finite population proportions lower. As mentioned above this is likely due to the DK's representing more No than Yes responses. Finally, we note that all of finite population 95% posterior credible sets contain the actual plebiscite vote. As we noted in Chapter 2, this is a positive outcome, but not a direct confirmation of the validity of the model. It is not possible to easily construct a design-based estimate that accounts for the missing data so no comparison is made.

To conclude this section we note that the model appears to fit the data from the SPO quite well. Additionally the results indicate, based on the large values for κ and the η_j's, that the cluster and subcluster probability vectors are fairly homogeneous with little intracluster correlation. This is not surprising given some additional information concerning the methodology for the SPO. In the SPO the frame is the voting registry for the population. PSU's were selected systematically. Then within a PSU the SSU's

Table 3.18 Proportions of Slovenian Voters Who Intend to Attend the Plebiscite and to Vote for Independence Incorporating Data from Partially Observed Response: Three-stage Results

	Posterior		Percentiles	
Parameter	Mean	2.5^{th}	50^{th}	97.5^{th}
ξ_1	0.869	0.853	0.868	0.892
ξ_2	0.870	0.854	0.869	0.893
ξ_3	0.870	0.854	0.870	0.893
ξ_4	0.872	0.855	0.871	0.895
Actual Vote			0.885	

are deterministically sampled from the beginning, middle and end of the selected PSU. Finally, the respondents are sampled systematically from the individuals in the selected SSU's, (Vehovar (1998)).

CHAPTER 4 TWO-STAGE MODEL WITH ADDITIONAL COVARIATES

The 1990 Slovenian Public Opinion Survey (SPO) that was analyzed in the preceding chapters actually includes many more survey items than the three that were analyzed in Chapters 2 and 3. Under the assumption that the unobserved responses are missing at random (MAR), values of observed variables are used to infer the likely answer of nonrespondents. It is natural to wonder if covariates other than responses to the three related questions would help draw more accurate inferences. The hierarchical approach of Chapters 2 and 3 could be extended to accommodate additional categorical covariates by enlarging the number of multinomial cells. However, with only five observations per subcluster the resulting data would be sparse in the multinomial. Here we consider an alternative model for analyzing multivariate binary responses from multi-stage cluster samples using latent variables. The model that is outlined in this chapter is an extension of that described by Chib and Greenberg (1998). In Section 4.1 we introduce the model. Section 4.2 outlines an approach to simulating samples from the posterior distribution.

4.1 Probability Model

The model that we propose is an extension of one that was developed by Chib and Greenberg (1998), hereafter abbreviated as CG. CG proposed a multivariate probit model for analyzing multivariate binary responses and their relationship to covariates. For each of the responses, a Gaussian latent variable is hypothesized. These latent variables can be thought of as measuring the strength of each respondent's opinion regarding that particular question. We assume that if the latent variable is positive, then the binary response is 1; if the latent variable is negative, the binary response is 0. The covariates are linked to the binomial responses through these latent variables. Specifically, the latent variables are assumed to depend on the covariates through a Gaussian linear regression. Under the CG model, the regression coefficients are the same for each individual in the population. That is, the same relationship between the covariates and the latent variables exists for every member of the population. Given the covariates, CG treat the individuals as independent from each other, while allowing for the possibility that the binary responses are correlated. We extend this formulation to accommodate multi-stage cluster sampling. To do this we allow for the possibility that there will be different relationships between the covariates and the latent variables within each cluster. We specify a model that takes the the vector of slopes for the Gaussian linear regression in each cluster as a random draw from a population of cluster level slopes. Additionally, we extend the CG model to incorporate missing data. In this we describe only the model for a two-stage cluster sample. The extension to multi-stage cluster sampling is straightforward.

We repeat the two-stage cluster sample notation of Chapter 2. Here specifically we assume that the population is divided into M clusters, and we sample J clusters from among the M. The number of unsampled clusters is J'. A sample of size n_j is selected from the population of N_j individuals in the j^{th} cluster. Each respondent is asked a series of I questions with binary responses. Note that this is a change from Chapter 2 where I represented the number of multinomial cells. Let Y_{ijk} be the binary response for the i^{th} variable of the k^{th} respondent in the j^{th} cluster, where $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, n_j$. Now let $\mathbf{Y}_{jk} = (Y_{1jk}, \ldots, Y_{Ijk})^T$ be the vector of responses for the k^{th} individual in the j^{th} cluster. Note that here if we construct a multinomial distribution for the binary responses then it will have 2^I cells. The methods of Chapters 2 and 3 allow for an arbitrary number of multinomial cells. Additionally we assume that for each member of the population there is a p_i−dimensional vector of covariates, $\mathbf{X}_{ijk}$ relevant to the i^{th} binary response for the k^{th} respondent in the j^{th} cluster. There are two main uses for covariates. First, from a modeling perspective, we may be interested in the relationship between the covariates, $\mathbf{X}_{ijk}$, and the response, Y_{ijk}, for each question i. The second potential use of covariates, which is more relevant for our example, is that they may provide additional insight for dealing with missing values. Recall that under the missing at random (MAR) assumption, the probability of a variable's being missing can depend on values of other variables, but not on the value of the variable of interest. By adding covariates into the model we make the assumption of MAR more plausible. By further conditioning the unobserved responses on covariate information, we may improve our prediction of these responses. See Section 2.5 for a more detailed discussion of MAR. Throughout this chapter we assume that the covariates are completely observed.

The probit model is most easily motivated by introducing latent variables. Moreover, the latent variable formulation provides advantages in computation as well. We introduce $\mathbf{w}_{jk}$, a vector of latent variables associated with $\mathbf{Y}_{jk}$. The elements of $\mathbf{w}_{jk}$ can be thought of as measures of the intensity of feeling for the k^{th} individual in the j^{th} cluster toward the questions. The binary responses are completely determined by the w_{ijk}'s. If $w_{ijk} > 0$ then $Y_{ijk} = 1$, whereas if $w_{ijk} < 0$ then $Y_{ijk} = 0$. We model the w_{ijk}'s as Gaussian random variables and allow their mean to depend on the covariates. Let $p = \sum_{i=1}^{I} p_i$ represent the total number of covariates for all questions. This can include some duplicates if the same covariates are relevant for more than one response. We define $\mathbf{X}_{jk}$ as a $p \times I$ matrix of covariates with the i^{th} column equal to

$$(\mathbf{0}_{p_1}^T, \ldots, \mathbf{0}_{p_{i-1}}^T, \mathbf{X}_{ijk}^T, \mathbf{0}_{p_{i+1}}^T, \ldots, \mathbf{0}_{p_I}^T)^T \tag{4.1}$$

where $\mathbf{0}_k$ is a k-dimensional vector of zero's. Thus each element of $\mathbf{X}_{jk}$ contains the covariates relevant to a single response. The Gaussian assumptions for $\mathbf{w}_{jk}$ imply

$$\mathbf{w}_{jk} \sim N_I(\mathbf{X}_{jk}^T \boldsymbol{\beta}_j, \mathbf{V}_j) \tag{4.2}$$

where $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{Ij})$ and $\mathbf{V}_j$ is a correlation matrix. Then the discrete distribution for $\mathbf{Y}_{jk}$ is just the relevant probability computed from the multivariate distribution for $\mathbf{w}_{jk}$.

$$\mathrm{P}(\mathbf{Y}_{jk} = \mathbf{y}_{jk} \mid \boldsymbol{\beta}_j, \mathbf{V}_j) = \int\limits_{H_{1jk}} \ldots \int\limits_{H_{Ijk}} \phi_I(t \mid \mathbf{X}_{jk}^T \boldsymbol{\beta}_j, \mathbf{V}_j) dt \tag{4.3}$$

where

$$\text{where} \quad H_{ijk} = \begin{cases} (0, \infty) & \text{if} \quad y_{ijk} = 1 \\ (-\infty, 0] & \text{if} \quad y_{ijk} = 0. \end{cases} \tag{4.4}$$

Again following CG, we let $H_{jk} = H_{1jk} \times H_{2jk} \times \ldots \times H_{Ijk}$.

The reason that $\mathbf{V}_j$ is a correlation matrix rather than a covariance matrix is that the multivariate normal probability (4.3) is unchanged if $\mathbf{V}_j$ and $\boldsymbol{\beta}_j$ are replaced by $\mathbf{C}\mathbf{V}_j\mathbf{C}^T$ and $\mathbf{C}\boldsymbol{\beta}_j$ for any diagonal matrix $\mathbf{C}$. Thus the coefficients $\boldsymbol{\beta}_j$ and a variance matrix cannot be uniquely identified. The resolution favored by CG is to take $\mathbf{V}_j$ as a correlation matrix.

We now take the basic CG model and incorporate hierarchical structure to account for the cluster sampling. For the two-stage cluster sample, we model the relationship between the covariates, $\mathbf{X}_{jk}$, and the latent variable, $\mathbf{w}_{jk}$, as the same for each individual within a cluster. This relationship is described by the vector of regression coefficients $\boldsymbol{\beta}_j$. We then model the p-dimensional vector of regression coefficients for the j^{th} cluster, $\boldsymbol{\beta}_j$, as coming from a population of cluster-level regression coefficients,

$$\boldsymbol{\beta}_j \sim N_p(\mathbf{b}, \boldsymbol{\Omega}). \tag{4.5}$$

To complete the Bayesian treatment of this model we specify prior distributions for $\mathbf{V}_j$, $\mathbf{b}$ and $\boldsymbol{\Omega}$. For convenience we parameterize $\mathbf{V}_j$ in terms of the upper triangular off-diagonal elements, $\boldsymbol{\sigma}_j = (\sigma_{(j)12}, \ldots, \sigma_{(j)1I}, \ldots, \sigma_{(j)I-1,I})$. The dimension of $\boldsymbol{\sigma}_j$ is q

$= \frac{I(I-1)}{2}$. Note that this completely characterizes $\mathbf{V}_j$, since $\mathbf{V}_j$ is a correlation matrix. The prior distribution for $\mathbf{b}, \boldsymbol{\Omega}$, and $\boldsymbol{\sigma}_j$ are as follows,

$$\begin{aligned} &\mathrm{p}(\mathbf{b}) \propto 1 \\ &\boldsymbol{\Omega} \sim \text{Inverse-Wishart}_{\delta}(\mathbf{I}_p) \\ &\mathrm{p}(\boldsymbol{\sigma}_j) I(\boldsymbol{\sigma}_j \in C) \end{aligned} \tag{4.6}$$

where C is a convex set in the hypercube $[-1,1]^q$ that yields a proper correlation matrix, $\mathbf{I}_p$ is the $p-$dimensional identity matrix and δ is a degrees of freedom parameter. The prior on $\mathbf{b}$ is an improper distribution that is flat on the entire real line for each of the p elements. The Inverse-Wishart distribution is chosen as the prior distribution for $\boldsymbol{\Omega}$ because it is the conjugate prior distribution for the variance matrix of a Gaussian distribution. As a consequence the conditional posterior distribution of $\boldsymbol{\Omega}$ that arises in the Gibbs sampling algorithm of the next section will be Inverse-Wishart. Finally the choice of a uniform prior on each of the $\boldsymbol{\sigma}_j$'s is made to allow the greatest flexibility in the correlations between the latent variables within a cluster, i. e. to let the data dictate the correlations. This is a proper distribution since we are placing a uniform distribution on a finite space, C. Note that the improper prior distribution on $\mathbf{b}$ is not a problem; it is easy to show that the joint posterior distribution of all parameters is proper.

4.2 Posterior Inference for Complete Data

Section 4.1 describes a multivariate probit model for analyzing multivariate binary responses from a two-stage cluster sample. It should be noted that the logit is another popular transformation for binary data, see for example McCullagh and Nelder (1983) or Agresti (1990). Here, however, the Gaussian latent variables that generate the probit model make it convenient to identify full conditional distributions for use in a Gibbs sampling algorithm to sample from the posterior distribution of the model parameters. Gibbs sampling draws samples from the posterior distribution by cycling through a sequence of the full conditional distributions. A more detailed review of Gibbs sampling can be found in Section 2.4.1. We include the latent variables as unknown parameters in the posterior distribution.

The full posterior distribution is

$$\begin{aligned}
&\mathrm{p}(\boldsymbol{\beta}, \boldsymbol{\sigma}, \mathbf{b}, \boldsymbol{\Omega}, \mathbf{w} \mid \mathbf{Y}, \mathbf{X}) \\
&\quad \propto \; p(\boldsymbol{\beta}, \boldsymbol{\sigma}, \mathbf{b}, \boldsymbol{\Omega}, \mathbf{w}, \mathbf{Y} \mid \mathbf{X}) \\
&\qquad = \mathrm{p}(\mathbf{Y} \mid \mathbf{w})\mathrm{p}(\mathbf{w} \mid \boldsymbol{\beta}, \mathbf{X}, \boldsymbol{\sigma})\mathrm{p}(\boldsymbol{\beta} \mid \mathbf{b}, \boldsymbol{\Omega})\mathrm{p}(\mathbf{b})\mathrm{p}(\boldsymbol{\Omega})\mathrm{p}(\boldsymbol{\sigma})
\end{aligned}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$, and $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_J)$ are matrices of the the respective cluster-level quantities, and $\mathbf{X}$, $\mathbf{w}$, and $\mathbf{Y}$ are introduced as notation for all of the covariates, latent variables and responses, respectively. Then,

$$\begin{aligned}
&\mathrm{p}(\boldsymbol{\beta}, \boldsymbol{\sigma}, \mathbf{b}, \boldsymbol{\Omega}, \mathbf{w} \mid \mathbf{Y}, \mathbf{X}) \\
&= \prod_{j=1}^{J}\prod_{k=1}^{n_j}(2\pi)^{-I/2} \mid \mathbf{V}_j \mid^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{w}_{jk} - \mathbf{X}_{jk}^T\boldsymbol{\beta}_j)^T\mathbf{V}_j^{-1}(\mathbf{w}_{jk} - \mathbf{X}_{jk}^T\boldsymbol{\beta}_j)\right\} \\
&\times \prod_{j=1}^{J}(2\pi)^{-p/2} \mid \Omega \mid^{-1/2} exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_j - \mathbf{b})^T\Omega^{-1}(\boldsymbol{\beta}_j - \mathbf{b})\right\} && (4.7) \\
&\times \left(2^{\nu p/2}\pi^{k(k-1)/4}\prod_{r=1}^{p}\Gamma\left(\frac{\nu+1-r}{2}\right)\right)^{-1} \\
&\times \mid \Psi_p \mid^{\nu/2}\mid \Omega \mid^{-(\nu+p+1)/2} exp\left(-\frac{1}{2}tr(\Psi_p\Omega^{-1})\right) \times \prod_{j=1}^{J} I_{(\boldsymbol{\sigma}_j \in C)} \\
&\times \prod_{j=1}^{J}\prod_{k=1}^{n_j}\prod_{i=1}^{I}\left(I_{(w_{ijk}>0)}\right)^{Y_{ijk}}\left(I_{(w_{ijk}\leq 0)}\right)^{1-Y_{ijk}} && (4.8)
\end{aligned}$$

From (4.7) we can derive the full posterior conditional distributions. They are as follows

$$\begin{aligned}
\mathbf{w}_{jk} \mid \mathbf{Y}_{jk}, \mathbf{X}_{jk}, \boldsymbol{\beta}_j, \mathbf{V}_j &\sim \text{Gaussian}(\mathbf{X}_{jk}\boldsymbol{\beta}_j, \mathbf{V}_j)I(\mathbf{Y}_{jk} \in H_{jk}) \\
\boldsymbol{\beta}_j \mid \mathbf{w}_{jk}, \mathbf{X}_{jk}, \mathbf{b}, \mathbf{V}_j, \boldsymbol{\Omega} &\sim \text{Gaussian}(\mu, D)
\end{aligned}$$

$$\begin{aligned} &\text{where} \quad D = \left(\boldsymbol{\Omega}^{-1} + \sum_{k=1}^{n_j} \mathbf{X}_{jk}^T \mathbf{V}_j^{-1} \mathbf{X}_{jk}\right)^{-1} \\ &\text{and} \quad \mu = D \left(\boldsymbol{\Omega}^{-1}\mathbf{b} + \sum_{k=1}^{n_j} \mathbf{X}_{jk}^T \mathbf{V}_j^{-1} \mathbf{w}_{jk}\right) \end{aligned}$$

$$\mathbf{b} \mid \boldsymbol{\beta}, \boldsymbol{\Omega}, \sim \text{Gaussian}\left(1/J \sum_{j=1}^{J} \boldsymbol{\beta}_j, (\boldsymbol{\Omega}/J)\right)$$

$$\boldsymbol{\Omega} \mid \boldsymbol{\beta}, \mathbf{b}, \nu \sim \text{Inverse-Gamma}_{\delta+J}\left(I_p + \sum_{j=1}^{J} \left(\boldsymbol{\beta}_j - \mathbf{b}\right)\left(\boldsymbol{\beta}_j - \mathbf{b}\right)^T\right)$$

and

$$\begin{aligned} \mathrm{p}(\boldsymbol{\sigma}_j \mid \mathbf{w}, \mathbf{X}, \boldsymbol{\beta}_j) \propto \prod_{k=1}^{n_j} \mid \mathbf{V}_j \mid^{-1/2} \\ \times exp\left\{-1/2\left(\mathbf{w}_{jk} - \mathbf{X}_{jk}^T\boldsymbol{\beta}_j\right)^T \mathbf{V}_j^{-1}\left(\mathbf{w}_{jk} - \mathbf{X}_{jk}^T\boldsymbol{\beta}_j\right)\right\} I_{(\boldsymbol{\sigma}_j \in C)}. \end{aligned}$$

We then draw samples from the posterior distribution following the Gibbs sampling algorithm outlined in Section 2.4.

Several aspects of this Gibbs sampling algorithm require elaboration. First, the posterior distribution of $\mathbf{V}_j$ is not a well-known form. So we cannot sample directly from the distribution $\mathrm{p}(\boldsymbol{\sigma}_j \mid \mathbf{b}_j, \mathbf{w}_{jk}, \mathbf{X}_{jk})$. As a consequence we use a Metropolis algorithm for this step. We generate candidates for $\boldsymbol{\sigma}_j$ via

$$\boldsymbol{\sigma}_j^* \sim N_q(\boldsymbol{\sigma}_j^{(t-1)}, \mathrm{c}\mathbf{I}_q) \tag{4.9}$$

where $\mathbf{I}_q$ is the q-dimensional identity matrix. As before c is chosen to achieve an efficient jumping rate; this approach to sampling non-standard distributions is described in Section 2.4.

A second noteworthy aspect of the Gibbs sampling algorithm is that the conditional distribution of $\mathbf{w}_{jk}$ is a truncated multivariate Gaussian distribution. The region to which $\mathbf{w}_{jk}$ is restricted, H_{jk}, is determined by the values of $\mathbf{Y}_{jk}$. To generate observations for this distribution we modify an algorithm of Geweke (1991) for generating realizations from univariate truncated Gaussian densities. Specifically, we generate a realization for each $\mathbf{w}_{jk}$ conditional on the remaining elements of $\mathbf{w}_{jk}$. Thus, we can generate a truncated Gaussian variate for each of the q elements of $\mathbf{w}_{jk}$. In this way we get a realization of $\mathbf{w}_{jk}$ that is restricted to H_{jk}. These restrictions also play a role in missing data.

4.3 Posterior Inference with Missing Data

The Gibbs sampling algorithm of Section 4.2 considers the case where $\mathbf{Y}_{jk}$ is completely observed for each individual. It is possible that one or more of elements of $\mathbf{Y}_{jk}$ is missing but we continue to assume that $\mathbf{X}_{jk}$ is always observed completely. As in the preceding chapters we assume a MAR mechanism for the missing data. Under the MAR assumption, the probability that a response Y_{ijk} is missing may depend on observed variables (such as $\mathbf{X}_{jk}$ or observed elements of $\mathbf{Y}_{jk}$) but not on the value (one or zero) that would have been observed. It is straightforward to incorporate missing responses into the Gibbs sampling algorithm of Section 2.4. Unlike Chapters 2 and 3, we need not formally include a step for the missing binary responses because they are completely determined by the corresponding latent variables. Note that all of the latent variables w_{ijk} are "missing." For the latent variable corresponding to observed responses, we use a univariate truncated Gaussian distribution to simulate from their conditional posterior distribution with the relevant intervals equal to H_{ijk} where

$$H_{ijk} = \begin{cases} (0,\infty) & \text{if} \quad y_{ijk} = 1 \\ (-\infty,0) & \text{if} \quad y_{ijk} = 0. \end{cases} \tag{4.10}$$

All that is required to incorporate the missing responses is to modify the definition of H_{ijk} to

$$H_{ijk} = \begin{cases} (0,\infty) & \text{if} \quad y_{ijk} = 1 \\ (-\infty,0) & \text{if} \quad y_{ijk} = 0 \\ (-\infty,\infty) & \text{if} \quad y_{ijk}\text{is unobserved.} \end{cases} \tag{4.11}$$

So w_{ijk} is generated from an ordinary untruncated Gaussian distribution if y_{ijk} is unobserved. Thus we simulate a complete set of latent variables and proceed with the remaining Gibbs sampling steps as in Section 4.2

4.4 Final Comments

The multivariate binary model allows us to easily introduce covariates and perhaps thereby improve inference when there are unobserved responses. As future goal is to apply the models of this Chapter to the SPO survey data.

CHAPTER 5 CONCLUSIONS AND FUTURE WORK

The goal of this work was to develop hierarchical models for analyzing polytomous data from multi-stage cluster samples. In Chapter 2 we created such a model for polytomous data collected from a two-stage cluster sample. This model takes observed responses within a PSU to follow a multinomial distribution and then models the PSU-level probability vectors as draws from a Dirichlet population. At the top level of the hierarchy a hyperprior distribution placed on the Dirichlet parameters. The hyperprior distribution that was suggested is an improper distribution. We derived conditions under which this improper hyperprior yields a proper posterior distribution. Additionally, we showed how to incorporate unintentional missing data assuming that the data are missing at random. This two-stage model was then applied to data from the 1990 Slovenian Public Opinion survey (SPO). The polychotomous response that arose there was the result of transforming three binary questions into a $2^3 = 8-$dimensional multinomial response.

The two-stage model was extended to a three-stage model in Chapter 3. The three-stage model is appropriate for data collected via a three-stage cluster sample. As in the two-stage model, we showed how to incorporate missing data into our analyses. A notable feature of both models is that they can be used to simulate data for the unseen members of the population. That is, we can find the distribution of the intentionally missing (unsampled) observations, $\mathbf{Y}^{unsamp}$, given the observations that we have seen (sampled) , $\mathbf{Y}^{samp}$. For both the two-stage and the three-stage models the primary quantity of interest is the finite population proportion that includes both $\mathbf{Y}^{samp}$ and

$\mathbf{Y}^{unsamp}$. As with the two-stage model we analyzed the SPO data using the three-stage model. As with any model-based approach the question of how well the model fits the data is an important one. Though we have not formally addressed it here, other than to note that the model estimates match the observed plebiscite outcomes, we recognize the importance of model assessment and hope to return to it at a later date.

The SPO contained additional information besides the three questions that we focused on in Chapters 2 and 3. Chapter 4 treats the responses of interest as a trivariate binary vector and includes other SPO items as covariates. We expanded the multivariate probit model of Chib and Greenberg (1998), which relates binary response variables to covariates, to allow for data collected via a multi-stage cluster sample and for data that was unintentionally missing. We built a hierarchical model to accomplish this and describe an MCMC algorithm for simulation from the posterior distribution. This method has not yet been implemented for the SPO. This is left as future work.

The basic methodology that is presented here can be extended to other types of non-normal data. Specifically, count data would be easily amenable to this approach. The basis for an analysis of that kind might be a Poisson-gamma model. The approach for other types of non-normal data is not as obvious but should be workable under the framework outlined in this thesis.

BIBLIOGRAPHY

Agresti, A. (1990). *Categorical Data Analysis.* John Wiley & Sons, New York.

Altham, P. M. E. (1976). Discrete variable analysis for individuals grouped into families. *Biometrika*, 63:263–269.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36:192–236.

Brier, S. S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika*, 67:591–596.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85:347–361.

Cochran, W. G. (1977). *Sampling Techniques.* John Wiley & Sons, New York, third edition.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91:883–904.

Ericson, W. A. (1988). Bayesian inference in finite populations. In Krishnaiah, P. R. and Rao, C. R., editors, *Handbook of Statistics*, volume 6, pages 213–246. Elsevier Science Publishers B. V., New York.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall, New York.

Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient Metropolis jumping rules. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics V*, pages 599–607. Oxford University Press.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linears constraints and the evaluation of constraint probabilities. In Keramidas, E. and Kaufman, S., editors, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 571–578. Interface Foundation of America, Fairfax Station, VA.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7:473–483.

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, New York.

Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93:273–282.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*, pages 1-20. Chapman & Hall, New York.

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

Kish, L. (1965). *Survey Sampling.* John Wiley & Sons, New York.

Koehler, K. J. and Wilson, J. R. (1986). Chi-square tests for comparing vectors of proportions for several cluster samples. *Communications in Statistics*, 15:2977–2990.

Little, R. J. A. and Rubin, D. J. (1987). *Statistical Analysis with Missing Data.* John Wiley & Sons, New York.

Malec, D. and Sedransk, J. (1985). Bayesian inference for finite population parameters in multistage cluster sampling. *Journal of the Americal Statistical Association*, 80:897–902.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models.* Chapman & Hall, New York.

Morel, J. G. and Koehler, K. J. (1995). A one-step Gauss-Newton estimator for modelling categorical data with extraneous variation. *Applied Statistics*, 44:187–200.

Morel, J. G. and Nagaraj, N. K. (1993). A finite mixture distribution for modelling multinomial extra variation. *Biometrika*, 80:363–371.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10:65–80.

Morris, C. N. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics*, 11:515–529.

Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61:97–126.

Nandram, B. and Sedransk, J. (1993a). Bayesian predictive inference for a finite population proportion: Two-stage cluster sampling. *Journal of the Royal Statistical Society, Series B*, 55:399–408.

Nandram, B. and Sedransk, J. (1993b). Bayesian predictive inference for longitudinal sample survey. *Biometrics*, 49:1045–1055.

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

Roberts, G. O. and Smith, A. F. M. (1993). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, 49:207–216.

Royall, R. M. and Cumberland, W. G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73:351–358.

Rubin, D. B. (1976). Inference and missing data. *Biometrics*, 63:581–592.

Rubin, D. B., Stern, H. S., and Vehovar, V. (1995). Handling "Don't Know" survey responses: The case of the Slovenian plebiscite. *Journal of the American Statistical Association*, 90:822–828.

Särndal, C.-E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, 5:27–52.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Scott, A. and Smith, T. M. F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 64:830–840.

Silber, L. and Little, A. (1997). *Yugoslavia: Death of a Nation.* Penquin Books, New York.

Stasny, E. A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse: An example from the national crime survey. *Journal of the American Statistical Association*, 86:296–303.

Stroud, T. W. F. (1991). Bayesian inference from categorical survey data. Mathematical preprint, Department of Mathematics and Statistics, Queen's University, Kingston, Canada.

Stroud, T. W. F. (1994). Bayesian analysis of binary survey data. *The Canadian Journal of Statistics*, 22:33–45.

Thomsen, I. and Tesfu, D. (1988). On the use of models in sampling from finite populations. In Krishnaiah, P. R. and Rao, C. R., editors, *Handbook of Statistics*, volume 6, pages 369–397. Elsevier Science Publishers B. V., New York.

Tierney, L. (1996). Introduction to general state-space Markov chain theory. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, pages 59–88. Chapman & Hall.

Vehovar, V. (1998). Personal correspondence.

Wilson, J. R. (1984). *Statistical Methods for Frequency Data from Complex Sampling Schemes.* PhD dissertation, Department of Statistics, Iowa State University, Ames, IA.

Wilson, J. R. (1986). Approximate distribution and test of fit for the clustering effect in the Dirichlet multinomial model. *Communications in Statistics: Theory and Methods*, 15:1235–1249.

Wilson, J. R. (1987). A generalized Dirichlet multinomial model for categorical data with extra variation. In *Proceedings of the ASA Section on Survey Research Method*, pages 353–355. American Statistical Association, Alexandria, VA.