University of Southampton

Bayesian Model Determination for Categorical Data Survey

by

Abdulhakim A. Al-Babtain

Thesis submitted for the degree of Doctor of Philosophy

Faculty of Mathematical Studies

May 2001

University of Southampton

ABSTRACT

Faculty of Mathematical Studies

Mathematics

Doctor of Philosophy

Bayesian Model Determination for Categorical Data Survey

by Abdulhakim A. Al-Babtain

Inference for survey data needs to take account of the survey design. Failing to consider the survey design in inference may lead to misleading results. The standard analysis of categorical data, developed under the assumption of multinomial sampling, is inadequate as the commonly used sampling schemes clearly violate this assumption. Since, Kish (1965) introduced the idea of a design effect, many classical solutions have been proposed, such as, first- and second-order corrections to Pearson chi-squared, likelihood-ratio chi-squared, and Wald tests.

Our objective in this thesis is to present an investigation of a Bayesian approach to the analysis of categorical survey data, arising from designs including simple random sampling, finite population sampling, stratification, and cluster sampling. We focus on Bayesian methods for model selection and model averaging, where Bayes factors and the Bayesian Information Criterion (BIC) approximation have been offered as alternative approaches. These Bayesian methods are reviewed, and comparisons made between their performance. The effect of ignoring the complex sampling design is investigated. Moreover, adjustments to the multinomial-based Bayes factor and BIC are produced and evaluated.

With stratification, our results indicate that there is little effect in ignoring the sampling design on the inferences using BIC, if the strata are homogeneous. When the strata are highly inhomogeneous, BIC is affected by this sampling scheme. On the other hand, the Bayes factor is sensitive and affected by stratification.

We investigate the effect of cluster sampling on Bayesian model selection. In a goodness-of-fit test, the results of both the Bayes factor and BIC have the potential to provide a misleading result. However, if the BIC is based on corrected statistics, which consider the sampling design, the results are acceptable. Moreover, we present two simple adjustments to the multinomial-based Bayes factor which perform well in simulations.

For testing of independence, our results indicate that the effect of the sampling design is negligible. Therefore, there is no justification in practice for the use of a more complex test statistic, unless the number of observations in each cluster is very small.

Finally, we demonstrate using risk analysis, how the Bayesian approach for estimating the cell probabilities, using a model averaged estimator, can be better than a pretest estimate.

ACKNOWLEDGMENTS

"Thy Lord has decreed that you worship none but Him, and that you be kind to parents. Whether one or both of them attain old age in your life, say not to them a word of contempt, nor repel them, but address them in terms of honor."

al-Qur'an 17:24.

I want to start by thanking the almighty God who gave me this blessing. I am forever indebted to my parents to whom I owe much more than words can express.

I would like to extend my sincere gratitude to Dr. Jonathan J. Forster not only for supervising this research, but also for providing fine opportunities for my development.

My thanks to Mr. B. J. R. Bailey and Professor T. M. F. Smith for invaluable discussions and advice. I wish to thank Professor S. M. Lewis for her encouragement and support. I would also like to thank Professor P. Prescott and Dr. Sujit K. Sahu for their support, and Ray Brown for his computing assistance. Many thanks to all my fellow students, especially Nan, Mark, and Ralph who made difficult times pass easier.

I would like to express my deep gratitude to my wife without whose endurance, encouragement, motivation, and patience the mission would not have been accomplished. I would, also, like to express my deep gratitude to my father in-law, and my mother in-law for their love and support.

Finally, I cannot forget my dear children Hadeel, Mohsin, and Aljohara, and I dedicate this thesis to them for their patience and tolerance throughout my study.

Lastly I would like to thank Dr. Abdulaziz Al-Nasralla, and Mr. Abdulrahman Al-Babtain for their support.

This research was supported by a grant from the government of the Kingdom of Saudi Arabia through my employer King Saud University.

Contents

Chapter 1

Introduction

1.1 Motivation

"More and more researchers are able to obtain data from complex samples, and to write computer programs for complex analytical statistics. We need methods for dealing properly with complex statistics from complex samples" Kish and Frankel (1974).

When using classical hypothesis tests for categorical survey data, it is now widely recognized that survey design can have a substantial impact on the results of multinomial-based methods. This leads to misleading results if we fail to make adjustments. Commonly used sample surveys for categorical data use stratification and cluster sampling or both. The standard analysis of categorical data, developed under the assumption of multinomial sampling, is inadequate as the commonly used sampling schemes clearly violate this assumption. The well known classical Pearson chi-squared test statistic is too liberal or too conservative, depending on the sampling scheme, when applied to survey data (Holt, Scott and Ewings, 1980, and Wilson and Warde, 1991).

Since, Kish (1965, p.258) introduced the idea of a design effect, many researchers have tried to overcome this problem. As a result, many classical solutions have been proposed, such as first- and second-order corrections to Pearson chi-squared, X^2, likelihood-ratio chi-squared, G^2, and Wald tests (see Kish and Frankel, 1974, Shuster and Downing, 1976, Fellegi, 1980, Holt, Scott and Ewings, 1980, Rao and Scott, 1981, 1984, 1987, Scott and Rao, 1981, Molina C. and Smith, 1986, Wilson and Warde, 1991, Graubard and Korn, 1993), and for a review see Thomas, Singh and Roberts (1996).

The majority of these solution procedures belong to one of four main classes. The first consists of methods based on the Wald statistic, as in Koch, Freeman and Freeman (1975), and Binder (1983). The second is the class of Rao and Scott tests consisting of correction to the classical X^2 and G^2 tests. The methodology was developed by Rao and Scott (1981, 1984, 1987). Thirdly Brier (1980) provided an alternative approach for the simple model for within-cluster dependence for clusters that was proposed by Cohen (1976) and Altham (1976). The fourth main class consists of methods based on Jackknifing the classical X^2 and G^2 tests, as proposed by Fay (1985).

These alternative test procedures, which take account of the complexity of design, are harder to apply. First-order corrections, using estimates of parameters of the population distribution, have the advantage that they only require knowledge of the design effect, deff, for the individual cells and margins of a contingency table. However, the second-order corrections, measures of variation of the first-order estimates, and Wald tests require knowledge of the full covariance matrix of the estimated cell proportions (Skinner, Holt and Smith, 1989).

Rao and Scott, in a series of papers together, (1981, 1984, 1987) presented a classical approach to overcoming this problem. Together with Holt, Scott and Ewings (1980) they illustrated the importance of assessing the impact of design effect upon multinomial-based methods. Rao and Scott proposed new measures

(or correction factors) for the Pearson chi-squared and the likelihood ratio chi-squared test statistics for different applications, including two-way tables and multi-way contingency tables.

The need for modified statistics is mainly due to the fact that summarized data usually do not include the necessary information for constructing Wald statistics and even second-order corrections. Unfortunately, for second-order corrections we need access to the full covariance matrix, which may not be possible, since researchers usually do not have access to primary data. The first-order corrections have the advantage that minimum information is needed, but the tests based upon them can have a serious distortion if the correction factor is fairly small and the size of the degrees of freedom are large (Rao and Scott, 1981, Fellegi, 1980). The effect of the design on the nominal significance level will depend on the size of the cell design effects and on the degrees of freedom (Holt, Scott and Ewings, 1980). In addition, corrections proposed in the classical approach are mainly conservative, in the sense that the actual type I error rate is less than the nominal level α; i.e. loss of power. In Rao and Scott's approach, the correction to Pearson chi-squared, for example, tends to underestimate the upper percentage points of the true asymptotic distribution.

In our research we argue that using Bayesian tests of hypotheses to overcome this problem may gave researchers useful tools as a Bayesian approach seems to offer potential benefits. Some of these advantages are as follows; see Kass and Raftery (1995) and Wasserman (1997);

- When several models are considered initially, a Model Averaged Estimate yields composite estimates or predictions that take account of model uncertainty. We will show how model averaged estimates may have good properties for estimating cell probabilities.
- Bayesian test of hypotheses are very general. Model-building can involve the comparison of more than two models. Also, the models being compared

(tested) do not require to be nested.

- The Bayes factor is evidence in favour of one scientific theory, represented by a statistical model, which can be evidence in favour of a null hypothesis, as first proposed by Jeffreys (1961, Appendix B).
- Algorithms, such as Occam's Window, have been proposed that allow model uncertainty to be taking into account when the class of models initially considered is very large (Raftery, 1995).
- The Bayesian Information Criterion (BIC) gives a simple approximation to the Bayes factor, which is easy to use for assessing competing models and does not require evaluation of prior distributions.
- It is often easier to apply under statistical models that do not satisfy common regularity conditions, see Raftery (1996b).

Under the exact multinomial sampling scheme, we will examine the performance (or behaviour) of model averaged estimation compared with estimation based on a pretest using the classical Pearson chi-squared test statistic. Our objective is to present a comprehensive treatment of Bayesian model selection approach to categorical survey data, encompassing simple random, stratified, and cluster sampling. The ideas are implicit in existing literature, but a full Bayesian treatment has been lacking.

1.2 Outline of the thesis

Our objective in this thesis is to present a comprehensive treatment of Bayesian model selection approach to categorical survey data. Chapters 4, 6, 7 are the core chapters. In these chapters, we demonstrate the effects of a complex sampling scheme on Bayesian model selection, and somewhat on the classical

hypotheses testing. In addition, we feel that chapters 2, 3, and largely 5 contain review material which are essential to understand the thesis.

Chapter 2 reviews previous work on classical approaches. It describes some basic ideas and concepts used in the survey context for categorical data and the basic concepts of design effects. In addition, the effect of complex survey design on the asymptotic distribution is discussed. Then, we review modifications to the standard Pearson chi-squared test statistic for goodness of fit and independence in a two-way contingency table; mainly Rao and Scott's (1981) first and second-order corrections. This survey considers the design effect under three sampling schemes, finite population sampling, stratification, and clustering.

Chapter 3 describes the basic theory of Bayesian model selection. The Bayes factor is described for two competing models. Unfortunately, sometimes it is difficult to evaluate the Bayes factor analytically, thus, the Savage-Dickey density ratio is presented as an approximation method for the Bayes factor. Estimating a parameter of interest via Model Averaged Estimation is also discussed. Finally, the Bayesian Information Criterion approximation is introduced as a criterion for model selection.

The Bayesian approach to model selection for categorical survey data will be considered in chapter 4 for three sampling schemes, simple random samples, a finite population, and stratified samples. We show the effect of these sampling schemes on the Bayesian model selection.

Reliable point density estimates are required to calculate the Savage-Dickey density ratio, for estimating a Bayes factor. Therefore, we dedicate a complete chapter, chapter 5, focussing on kernel density estimation. Virtually all non-parametric algorithms for density estimation are asymptotically kernel methods (Walter and Blum, 1979, and Scott, 1992).

Chapter 6 considers the complexity of the cluster sampling design, or two-

stage sample. This chapter describes a Bayesian treatment for this sampling scheme. The effect of ignoring the sampling design is also discussed. Markov chain Monte Carlo (MCMC) algorithms are developed for sampling from a posterior distribution. Two simple adjustments to the multinomial-based Bayes factor are presented and evaluated.

In chapter 7 we discuss the effect of the survey design on tests of independence in a 2×2 contingency table, and on the corresponding Bayesian model selection procedure. The other objective of this chapter is to investigate the behaviour of point estimates both in the Bayesian approach and in the classical approach with respect to risk.

Finally, we summarize our results in the first section of chapter 8. In the second section, we give some recommendations for possible extensions and development to this work.

1.3 Technical details

Almost all the results presented in this thesis rely heavily on numerical computation. All the routines for computing Bayes factor, BIC's, first- and second-order corrections, Markov chain Monte Carlo sampling algorithm and any associated computations were written in Pascal. In addition, routines for generating random numbers, matrix inversion and calculating determinants of a matrix are adapted from Press *et al.* (1988), and Ahrens and Dieter (1974, 1982). The results of all routines have been validated mainly by Maple and MINITAB. All graphs and plots in the thesis have been produced using MINITAB and S-PLUS.

1.4 Notation and terminology

Various notation conventions and abbreviations are used in the text. Throughout this thesis we denote vectors by bold lower case letters, matrices by bold capitals, and most greek letters represent parameters.

The most frequently occurring quantities are cell probabilities and cell frequencies p_i and n_i respectively, which will usually take just one subscript i ($i = 1, \ldots, K$, where K is the total number of cells). Chapters 2 and 3 are a review chapters, thus we will try to be consistent with the notation in the sample survey literature. In chapter 2, $\mathbf{p} = (p_1, \ldots, p_K)'$ represents the population proportions or probabilities, $\mathbf{p}_l = (p_{l1}, \ldots, p_{lK})'$ is the vector of the population proportions for stratum l, and $\mathbf{p}_t = (p_{t1}, \ldots, p_{tK})'$ is the vector of the population proportions for cluster t. The variance-covariance matrix of $\mathbf{p}$ for the simple random sample design is equal to $\mathbf{P} = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'$. Also, we used the standard sample survey notation $\mathbf{n} = (n_1, \ldots, n_K)'$ for the observed sample, where $n = \sum_{i=1}^{K} n_i$.

In chapters 4, 6, 7, and 8, $\mathbf{p}$ represents the collection of population probabilities, i.e. $\mathbf{p} = (p_1, \ldots, p_K)'$ for the simple random sample, $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_l, \ldots, \mathbf{p}_L)'$, where $\mathbf{p}_l$ is the vector of the population proportions for stratum l for stratification, and $\mathbf{p} = (\mathbf{p}^1, \ldots, \mathbf{p}^t, \ldots, \mathbf{p}^C)$, where $\mathbf{p}^t$ is the vector of the population proportions for cluster t for cluster sample. For stratification and clustering, in these chapters, we denote the marginal (over strata or clusters) probabilities by $\mathbf{q} = (p_1, \ldots, p_K)$.

Finally, the writer would like to apologise in advance if this terminology may confuse the reader. Nevertheless, he hopes that the context will make it clear what he is referring to.

Chapter 2

The effect of complex sampling in surveys

We are first going to discuss the effect of complex survey design on the asymptotic distribution of Pearson chi-squared test statistics for goodness of fit X^2 of prespecified multinomial probabilities and for independence X_I^2 in a two-way contingency table, from the classical point of view.

In this review we are going to consider the design effect under three sampling schemes, finite population sampling, stratification, and two-stage sampling. This chapter reviews previous work on classical approaches to this problem, particularly the work of Rao and Scott (1981), Scott and Rao (1981), and Holt, Scott, and Ewings (1981).

2.1 Multinomial distribution

Assume that Y is a discrete random variable taking K categories with probability of occurrence p_i; $i = 1, \ldots, K$. In a total sample of n independent observations

of Y, the probability that the i^{th} category is observed n_i times for cells $i = 1, ..., K$ is

$$pr(\mathbf{n}) = \begin{pmatrix} n \\ n_1, \ldots, n_K \end{pmatrix} \prod_{i=1}^{K} p_i^{n_i}. \tag{2.1}$$

where $\sum_{i=1}^{K} p_i = 1$. This distribution is known as the multinomial distribution.

2.2 Assumption of multinomial sampling

We assume that we have n independent observations concentrated on K categories and the observations are taken from a probability distribution with cell probabilities $p_1, ..., p_K$. In this sampling scheme the observations are independent and identically distributed, iid. Then the vector of counts $\mathbf{n} = (n_1, \ldots, n_K)'$ of the number of observations in each of the K categories is a random variable having a multinomial distribution.

We refer to the sampling scheme for those counts as multinomial sampling. Therefore, the assumptions of this sampling scheme are:

a) Each of the observations are independent.

b) The total sample size $n = \sum_{i=1}^{K} n_i$ is not a random variable.

Statistical inference based on multinomial sampling is valid only if those assumptions are fulfilled (Agresti, 1990). Otherwise, if at least one of the assumptions is violated, the inferential statistical procedure may not be valid. In sample surveys the violated assumption is usually the independence of the observations (Kish, 1965).

2.3 Multinomial sample

Let us define a variable Y_{ij} $(i = 1, \ldots, K, j = 1, \ldots, n)$, where

$$Y_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ sample element } \epsilon\ i^{th} \text{ class} \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

Then, we can write p_i, the cell proportion, as a population mean of the variable Y_{ij}. Now, assuming that the Y_{ij} are independently drawn, with replacement, from this population, then the cell counts $(n_1, \ldots, n_K)$ have a multinomial distribution with cell probabilities $\mathbf{p} = (p_1, \ldots, p_K)'$. Let $n = n_1 + \cdots + n_K$ and $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_K)'$ denote the sample proportions, where $\hat{p}_i = \sum_{j=1}^{n} Y_{ij}/n = n_i/n$. Then,

$$\begin{aligned} E(Y_{ij}) &= pr(Y_{ij} = 1) = p_i \\ E(Y_{ij}^2) &= pr(Y_{ij} = 1) = p_i \\ &\textit{and} \\ E(Y_{ij}Y_{tj}) &= 0 \qquad \text{if } i \neq t. \end{aligned} \tag{2.3}$$

It follows in the multivariate context that, if $\mathbf{Y}_j = (Y_{1j}, \ldots, Y_{Kj})'$

$$E(\mathbf{Y}_j) = \mathbf{p} \qquad \text{and} \qquad cov(\mathbf{Y}_j) = \mathbf{P} \qquad \forall j = 1, \ldots, n$$

where $\mathbf{P} = [v_{it}]$, with

$$\begin{aligned} v_{ii} &= var(Y_{ij}) = E(Y_{ij}^2) - (E(Y_{ij}))^2 = p_i - p_i^2 = p_i(1 - p_i) \\ v_{it} &= cov(Y_{ij}, Y_{tj}) = E(Y_{ij}Y_{tj}) - E(Y_{ij})E(Y_{tj}) = -p_i p_t \qquad \forall i \neq t. \end{aligned}$$

So, the matrix $\mathbf{P}$ has the form

$$\mathbf{P} = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}' \tag{2.4}$$

where $diag(\mathbf{p})$ is the diagonal matrix with elements of $\mathbf{p}$ on the main diagonal. Therefore,

$$[\mathbf{P}]_{it} = \begin{cases} p_i(1 - p_i) & \forall \quad i = t \\ -p_i p_t & \forall \quad i \neq t. \end{cases}$$

Since $\hat{\mathbf{p}}$ is a sample mean of independent observations, $\hat{\mathbf{p}} = \sum_{j=1}^{n} \mathbf{Y}_j/n$, we have $cov(\hat{\mathbf{p}}) = \mathbf{P}/n$. Using the multivariate central limit theorem

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{d} N(\mathbf{0}, \mathbf{P}) \qquad \text{for sufficiently large } n.$$

2.4 Goodness of fit test

Suppose we have a population split into K categories, classes, with population proportions $p_1, ..., p_K$, where $\sum_{i=1}^{K} p_i = 1$. Let $n_1, \ldots, n_K$ be the observed cell frequencies in a sample, s, of n units drawn with replacement according to a specified sampling design, $pr(s)$, from N elements. The general Pearson chi-squared statistic for testing,

$$H_0 : p_i = p_{0i} \qquad (i = 1, \ldots, K) \quad \text{where } p_{0i} \text{ is specified}$$

against the saturated alternative is given by

$$X^2 = n \sum_{i=1}^{K} (\hat{p}_i - p_{0i})^2 / p_{0i} \tag{2.5}$$

where $\hat{p}_i$ is an unbiased (or consistent) estimator of p_i under $pr(s)$ and $\sum_{i=1}^{K} \hat{p}_i = 1$. This can be expressed as,

$$X^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)' \mathbf{P}_0^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0)$$

where $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_{K-1})'$, $\mathbf{p}_0 = (p_{0_1}, \ldots, p_{0_{K-1}})'$, $\mathbf{P}_0$ is the value of $\mathbf{P} = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'$ for $\mathbf{p} = \mathbf{p}_0$, and $\mathbf{P}/n$ is the covariance matrix of $\mathbf{n}/n$ for multinomial sampling where $\mathbf{n} = (n_1, \ldots, n_{K-1})$.

Note that X^2 is a special case of the generalized Wald statistic X_W^2 for testing H_0, which is given by

$$X_W^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)' \hat{\mathbf{V}}^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0) \tag{2.6}$$

where $\hat{\mathbf{V}}/n$ is an estimate of the covariance matrix, $\mathbf{V}/n$, of $\hat{\mathbf{p}}$. The generalized Wald statistic X^2_W is approximately distributed as a χ^2_{K-1} random variable under H_0, for large n.

2.5 Pearson chi-squared, as a Wald statistic

The Pearson chi-squared statistic, for testing H_0, can be expressed in the multivariate context by,

$$X^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{P}_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0). \tag{2.7}$$

The inverse of $\mathbf{P}$ is of the form,

$$\mathbf{P}^{-1} = diag(\frac{1}{p_i}) + \frac{1}{p_K}\quad \mathbf{J} \qquad , where \quad \mathbf{J} = \begin{pmatrix} 1 & . & . & . & 1 \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ 1 & . & . & . & 1 \end{pmatrix}. \tag{2.8}$$

To check this assumption, we have to compute $\mathbf{PP}^{-1}$,

$$\begin{aligned}
\mathbf{PP}^{-1} &= (diag(p_i) - \mathbf{pp}')\left(diag(\frac{1}{p_i}) + \frac{1}{p_K}\mathbf{J}\right) \\
&= diag(p_i)diag(\frac{1}{p_i}) + \frac{1}{p_K}diag(p_i)\mathbf{J} - \mathbf{pp}'diag(\frac{1}{p_i}) - \frac{1}{p_K}\mathbf{pp}'\mathbf{J} \\
&= \mathbf{I} + \frac{1}{p_K}\begin{pmatrix} p_1 & 0 & .. & . & 0 \\ 0 & p_2 & & & : \\ : & & & & . \\ . & & & & 0 \\ 0 & . & .. & 0 & p_{K-1} \end{pmatrix}\mathbf{J} - \mathbf{p}\begin{pmatrix} p_1 & p_2 & \cdots & p_{K-1} \end{pmatrix}\begin{pmatrix} \frac{1}{p_1} & 0 & .. & . & 0 \\ 0 & \frac{1}{p_2} & & & : \\ : & & & & . \\ . & & & & 0 \\ 0 & . & .. & 0 & \frac{1}{p_{K-1}} \end{pmatrix} \\
&\quad -\frac{1}{p_K}\mathbf{p}\begin{pmatrix} p_1 & p_2 & \cdots & p_{K-1} \end{pmatrix}\mathbf{J}
\end{aligned}$$

$$\mathbf{PP}^{-1} = \mathbf{I} + \frac{1}{p_K}\begin{pmatrix} p_1 & p_1 & \cdots & p_1 \\ p_2 & p_2 & & p_2 \\ \vdots & \vdots & & \vdots \\ p_{K-1} & p_{K-1} & \cdots & p_{K-1} \end{pmatrix} - \mathbf{p1}' - \frac{\sum_{i=1}^{K-1} p_i}{p_K}\,\mathbf{p1}'$$

$$= \mathbf{I} + \frac{1}{p_K}\begin{pmatrix} \mathbf{p} & \cdots & \mathbf{p} \end{pmatrix} - \begin{pmatrix} \mathbf{p} & \cdots & \mathbf{p} \end{pmatrix} - \frac{1-p_K}{p_K}\mathbf{p1}' \qquad , \textit{where} \quad \sum_{i=1}^{K-1} p_i = 1 - p_K$$

$$= \mathbf{I} + \frac{1}{p_K}\begin{pmatrix} \mathbf{p} & \cdots & \mathbf{p} \end{pmatrix} - \begin{pmatrix} \mathbf{p} & \cdots & \mathbf{p} \end{pmatrix} - \frac{1-p_K}{p_K}\begin{pmatrix} \mathbf{p} & \cdots & \mathbf{p} \end{pmatrix}$$

$$= \mathbf{I} + \left(\frac{1-1+p_K}{p_K} - 1\right)\begin{pmatrix} \mathbf{p} & \cdots & \mathbf{p} \end{pmatrix}$$

$$= \mathbf{I}.$$

Therefore, we can conclude that equation (2.8) gives the inverse of $\mathbf{P}$.

Now to prove the multivariate form of the Pearson chi-squared statistic (2.7), let $\mathbf{p}' = (p_1, \ldots, p_{K-1})$ and $(\hat{\mathbf{p}} - \mathbf{p}_0) = \mathbf{a}$. Then,

$$\left[\mathbf{P}_0^{-1}\right]_{it} = \begin{cases} \frac{1}{p_{0i}} + \frac{1}{p_{0K}} & \textit{when} \quad i = t \\ \frac{1}{p_{0K}} & \textit{when} \quad i \neq t \end{cases}.$$

Hence,

$$\mathbf{a}'\mathbf{P}^{-1}\mathbf{a} = \sum_{i=1}^{K-1}\sum_{t=1}^{K-1} a_i \left[\mathbf{P}^{-1}\right]_{it} a_t$$

$$= \sum_{i=1}^{K-1} a_{ii}^2 \left[\mathbf{P}^{-1}\right]_{ii} + \sum_{i=1}^{K-1}\sum_{t\neq i}^{K-1} a_i \left[\mathbf{P}^{-1}\right]_{it} a_t$$

$$= \sum a_i^2 \left[\mathbf{P}^{-1}\right]_{ii} + 2\sum_{i=1}^{K-2}\sum_{t>i}^{K-1} a_i \left[\mathbf{P}^{-1}\right]_{it} a_t$$

$$= \sum a_i^2 \left(\frac{1}{p_{0i}} + \frac{1}{p_{0K}}\right) + 2\sum_{i=1}^{K-2}\sum_{t>i}^{K-1}\left(\frac{1}{p_K}\right) a_i a_t$$

$$= \frac{1}{p_{0i}}\sum a_i^2 + \frac{1}{p_{0K}}\left(\sum a_i^2 + 2\sum_{i=1}^{K-2}\sum_{t>i}^{K-1} a_i a_t\right)$$

Now,

$$\mathbf{a}'\mathbf{P}^{-1}\mathbf{a} = \frac{1}{p_{0i}}\sum_{i=1}^{K-1} a_i^2 + \frac{1}{p_{0K}}\left(\sum_{i=1}^{K-1} a_i\right)^2$$

substitute the value of $a_i = (\hat{p}_i - p_{0i})$, we get

$$= \frac{1}{p_{0i}}\sum(\hat{p}_i - p_{0i})^2 + \frac{1}{p_{0K}}\left(\sum(\hat{p}_i - p_{0i})\right)^2$$

$$= \sum_{i=1}^{K-1}\frac{(\hat{p}_i - p_{0i})^2}{p_{0i}} + \frac{1}{p_{0K}}\left(\sum_{i=1}^{K-1}\hat{p}_i - \sum_{i=1}^{K-1} p_{0i}\right)^2$$

but $\sum_{i=1}^{K-1}\hat{p}_i = 1 - \hat{p}_K$ and $\sum_{i=1}^{K-1} p_{0i} = 1 - p_{0K}$,

$$= \sum_{i=1}^{K-1}\frac{(\hat{p}_i - p_{0i})^2}{p_{0i}} + \frac{1}{p_{0K}}\left((1 - \hat{p}_K) - (1 - p_{0K})\right)^2$$

$$= \sum_{i=1}^{K-1}\frac{(\hat{p}_i - p_{0i})^2}{p_{0i}} + \frac{1}{p_{0K}}\left(-\hat{p}_K + p_{0K}\right)^2.$$

Therefore,

$$(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{P}_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0) = \sum_{i=1}^{K}\frac{(\hat{p}_i - p_{0i})^2}{p_{0i}}. \tag{2.9}$$

Thus,

$$\begin{aligned} X^2 &= n(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{P}_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0) \\ &= n\sum_{i=1}^{K}\frac{(\hat{p}_i - p_{0i})^2}{p_{0i}} \\ &= \sum_{i=1}^{K}\frac{(n\hat{p}_i - np_{0i})^2}{np_{0i}}. \end{aligned} \tag{2.10}$$

■

2.6 Likelihood-Ratio statistics and asymptotic distribution

The Likelihood-ratio test is a general way of testing a null hypothesis H_0 against an alternative hypothesis H_1. It is based on maximizing the likelihood under

H_0, and also under the alternative. It assumed that the two hypotheses are nested. Let Λ denote the ratio of the maximized likelihoods. Then the value of the ratio never exceeds one. In 1935, Wilks showed that $-2\ln(\Lambda)$ has a limiting chi-squared distribution under H_0, as the sample size, n, gets larger. The degrees of freedom is equal to the difference in the dimensions of the parameter spaces under the alternative hypothesis and the null hypothesis.

For multinomial sampling in a contingency table, with K cells, the likelihood is,

$$C\prod_{i=1}^{K} p_i^{n_i} \tag{2.11}$$

where C is a constant and $p_i \geq 0$ and $\sum_{i=1}^{K} p_i = 1$. Under H_0, consider $\hat{p}_i$ to be the maximum likelihood estimate for p_i. For a general saturated alternative, the likelihood is maximized when $p_i = \frac{n_i}{n}$. Therefore,

$$\begin{aligned}\Lambda &= \frac{\sup_{p\in\Omega_0} L(\mathbf{p},\mathbf{n})}{\sup_{p\in\Omega} L(\mathbf{p},\mathbf{n})} \\ &= \prod_{i=1}^{K}\left(\frac{\hat{p}_i}{n_i/n}\right)^{n_i}\end{aligned} \tag{2.12}$$

where Ω_0 is the null space and Ω is the alternative space for p. Now $-2\ln(\Lambda)$, which is denoted by G^2, is called the likelihood-ratio chi-squared statistic. For testing goodness-of-fit, it is equal to,

$$G^2 = 2\sum_{i=1}^{K} n_i \ln\left(\frac{n_i/n}{\hat{p}_i}\right). \tag{2.13}$$

In this test the evidence favours the alternative hypothesis, as G^2 get larger. For a general result for any null model, the likelihood-ratio statistic G^2 has an asymptotic Chi-square distribution with $(K - v - 1)$ degrees of freedom, where v is the number of free parameters of the null model which can be estimated efficiently from the data. To prove that G^2 is asymptotically equivalent to the Pearson chi-squared statistic, X^2, let K, the number of cells, be fixed, and

assume all $p_i \geq 0$. Then we can express G^2 as

$$G^2 = 2\sum_{i=1}^{K} n_i \ln\left(\frac{(n_i/n)}{\hat{p}_i}\right) \tag{2.14}$$
$$= 2\sum_{i=1}^{K} n_i \ln\left(1 + \frac{(n_i/n) - \hat{p}_i}{\hat{p}_i}\right).$$

Now, we apply the expansion

$$\ln(1+x) = x - \frac{x^2}{2} + O(x^3) \qquad \text{for } |x| < 1. \tag{2.15}$$

considering $x = \frac{(n_i/n)-\hat{p}_i}{\hat{p}_i}$, we get

$$G^2 = 2\sum_{i=1}^{K} n_i \left[\frac{(n_i/n) - \hat{p}_i}{\hat{p}_i} - \frac{1}{2}\left(\frac{(n_i/n) - \hat{p}_i}{\hat{p}_i}\right)^2 + ...\right]$$
$$= 2n\sum_{i=1}^{K}\left(\frac{n_i}{n}\right)\left[\frac{(n_i/n) - \hat{p}_i}{\hat{p}_i} - \frac{1}{2}\left(\frac{(n_i/n) - \hat{p}_i}{\hat{p}_i}\right)^2 + ...\right]$$

adding and subtracting $\hat{p}_i$,

$$\begin{aligned}
G^2 &= 2n\sum_{i=1}^{K}[\hat{p}_i + ((n_i/n) - \hat{p}_i)]\left[\frac{(n_i/n) - \hat{p}_i}{\hat{p}_i} - \frac{1}{2}\left(\frac{(n_i/n) - \hat{p}_i}{\hat{p}_i}\right)^2 + ...\right] \\
&= 2n\sum_i\left[((n_i/n) - \hat{p}_i) - \frac{1}{2}\frac{((n_i/n) - \hat{p}_i)^2}{\hat{p}_i} + ...\right] \\
&\qquad + 2n\sum_i\left[\frac{((n_i/n) - \hat{p}_i)^2}{\hat{p}_i} - \frac{1}{2}\frac{((n_i/n) - \hat{p}_i)^3}{\hat{p}_i^2} + ...\right] \\
&= 2n\sum_i((n_i/n) - \hat{p}_i) - n\sum_i\frac{((n_i/n) - \hat{p}_i)^2}{\hat{p}_i} + 2n\sum_i\frac{((n_i/n) - \hat{p}_i)^2}{\hat{p}_i} \\
&\qquad + 2n\, O_p((n_i/n) - \hat{p}_i)^3 \\
&= 2n\sum_i\frac{((n_i/n) - \hat{p}_i)^2}{\hat{p}_i} - n\sum_i\frac{((n_i/n) - \hat{p}_i)^2}{\hat{p}_i} + 2n\, O_p((n_i/n) - \hat{p}_i)^3
\end{aligned}$$

since $\sum_i((n_i/n) - \hat{p}_i) = 0$. Now, we can write $((n_i/n) - \hat{p}_i) = ((n_i/n) - p_i) - (\hat{p}_i - p_i)$, both of which are $O_p\left(n^{-\frac{1}{2}}\right)$. Thus

$$G^2 = n\sum_{i=1}^{K}\frac{((n_i/n) - \hat{p}_i)^2}{\hat{p}_i} + 2n\, O_p\left(n^{-\frac{3}{2}}\right)$$
$$= X^2 + O_p\left(n^{-\frac{1}{2}}\right). \tag{2.16}$$

This implies, that under the null hypotheses, the difference between the Pearson chi-squared statistic X^2 and the likelihood-ratio statistics G^2 converges in probability to zero. In fact, the limiting result for multinomial sampling also apply to any sampling scheme (Agresti, 1990). For further details see Cressie and Read (1989) and Bishop, Fienberg and Holland (1975).

2.7 The design effects

The design effect, deff, is a concept, used repeatedly since Kish (1965), defined as "the ratio of the actual variance of a quantity to the variance of that quantity under a simple random sample, srs, of the same number of elements". This definition is a natural measure of relative efficiency of a survey. It is a measure of the inflation (deflation) in variance of the simple random sample due to the design. Kish's design effect, deff, for $\hat{\theta}$ is a measure comparing the variance of the (randomization) distribution of $\hat{\theta}$ induced by the true complex design, $Var_{true}(\hat{\theta})$, and the variance of the distribution of $\hat{\theta}$ induced by a hypothetical srs design of the same sample size n, $Var_{srs}(\hat{\theta})$,

$$deff(\hat{\theta}) = Var_{true}(\hat{\theta})/Var_{srs}(\hat{\theta}). \tag{2.17}$$

For categorical data, where srs with replacement is multinomial sampling

$$deff(\hat{p}_i) = Var(\hat{p}_i)/p_i(1-p_i), \quad \forall \quad i = 1, ..., K-1. \tag{2.18}$$

It follows in the multivariate context that,

$$deff(\hat{\mathbf{p}}) = \mathbf{P}^{-1}\mathbf{V}. \tag{2.19}$$

The matrix $\mathbf{D} = \mathbf{P}^{-1}\mathbf{V}$ can be thought of as the natural multivariate extension of the design effect. Here, $\mathbf{D}$ represents the inflation factor needed to transform $\mathbf{P}$, the covariance matrix under multinomial sampling, to $\mathbf{V}$, the true covariance matrix for the sampling scheme actually employed.

Theorem 1 *(Rao and Scott, 1981)*
Under the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}_0$, X^2 *may be written as* $\sum_{i=1}^{K-1} \tau_{0_i} Z_i^2$, *where* $Z_1, \ldots, Z_{K-1}$ *are asymptotically independent* $N(0,1)$ *random variables and the* τ_i*'s are the eigenvalues of* $\mathbf{D} = \mathbf{P}_0^{-1}\mathbf{V}_0$ $(\tau_{0_1} \geq \tau_{0_2} \geq \ldots \geq \tau_{0_{K-1}} > 0)$ *where* $\mathbf{V}_0/n$ *denotes the covariance matrix* $\mathbf{V}/n$ *for* $\mathbf{p} = \mathbf{p}_0$.

Proof:

$$X^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{P}_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0)$$

$\mathbf{P}_0^{-1}$ is symmetric and p.d. matrix. So, we can write $\mathbf{P}_0^{-1} = \mathbf{M}\mathbf{M}'$

$$= n(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{M}\mathbf{M}'(\hat{\mathbf{p}} - \mathbf{p}_0).$$

If we denote $\mathbf{D}_0 = \mathbf{M}'\mathbf{V}_0\mathbf{M}$, then there is an orthogonal matrix Γ, such that $\Gamma'\mathbf{D}_0\Gamma$ is diagonal. Therefore,

$$\begin{aligned} X^2 &= n(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{M}\Gamma\Gamma'\mathbf{M}'(\hat{\mathbf{p}} - \mathbf{p}_0) \\ &= (\sqrt{n}\Gamma'\mathbf{M}'(\hat{\mathbf{p}} - \mathbf{p}_0))'(\sqrt{n}\Gamma'\mathbf{M}'(\hat{\mathbf{p}} - \mathbf{p}_0)). \end{aligned}$$

Now, let

$$\begin{aligned} \mathbf{T} &= (\sqrt{n}\Gamma'\mathbf{M}'(\hat{\mathbf{p}} - \mathbf{p}_0)) \overset{L}{\sim} N(0, \Gamma'\mathbf{M}'\mathbf{V}_0\mathbf{M}\Gamma) \\ &\overset{L}{\sim} N(0, \Gamma'\mathbf{D}_0\Gamma) \\ &\overset{L}{\sim} N(0, diag(\tau_{0i})) \end{aligned}$$

here, $\Gamma'\mathbf{M}'\mathbf{V}_0\mathbf{M}\Gamma = \Gamma'\mathbf{D}_0\Gamma$ is a diagonal matrix, with entries equal to the eigenvalues of $\mathbf{D}_0$, $\tau_{0_i} (i = 1, \ldots, K-1)$. Thus, T_i's are independent. Suppose that $Z_i = \frac{T_i}{\sqrt{\tau_{0_i}}}$, then we can write

$$X^2 = \sum_{i=1}^{K-1} \tau_{0_i} Z_i^2$$

where $Z_i \overset{L}{\sim} N(0,1)$. Now,

$$\begin{aligned}
\mathbf{D}_0\mathbf{e}_i &= \tau_{0_i}\mathbf{e}_i \qquad \text{where } \mathbf{e}_i \text{ is an eigenvector} \\
\mathbf{M}'\mathbf{V}_0\mathbf{M}\mathbf{e}_i &= \tau_{0_i}\mathbf{e}_i \qquad \text{then, multiplying by } \mathbf{M} \\
\mathbf{M}\mathbf{M}'\mathbf{V}_0(\mathbf{M}\mathbf{e}_i) &= \tau_{0_i}(\mathbf{M}\mathbf{e}_i) \qquad \text{but} \quad \mathbf{M}\mathbf{M}' = \mathbf{P}_0^{-1} \\
\mathbf{P}_0^{-1}\mathbf{V}_0(\mathbf{M}\mathbf{e}_i) &= \tau_{0_i}(\mathbf{M}\mathbf{e}_i).
\end{aligned}$$

Thus, τ_{0_i} is an eigenvalue of $\mathbf{P}_0^{-1}\mathbf{V}_0$ and $\mathbf{M}\mathbf{e}_i$ is the corresponding eigenvector of $\mathbf{P}_0^{-1}\mathbf{V}_0$. With that, we have the following,

- $Z_i \overset{L}{\sim} N(0,1)$ and Z_i's are independent.
- τ_{0_i}'s are the eigenvalue of $\mathbf{P}_0^{-1}\mathbf{V}_0$, under H_0.
- $\mathbf{P}_0^{-1}$ is p.d. and $\mathbf{V}_0$ is also p.d. $\Rightarrow$ all τ_{0_i}'s > 0.

We can arrange the orthogonal matrix to give us

$$\tau_{0_1} \geq \tau_{0_2} \geq \ldots \geq \tau_{0_{K-1}} > 0.$$

Finally,

$$X^2 \approx \sum_{i=1}^{K-1} \tau_{0_i}[N(0,1)]^2 \approx \sum_{i=1}^{K-1} \tau_{0_i}\chi_1^2. \tag{2.20}$$

So, X^2 is distributed asymptotically as a weighted sum of independent χ_1^2 random variables.■

Corollary 2 *The correct asymptotic distribution of the Pearson chi-squared statistic, X^2, under multinomial sampling is χ_{K-1}^2.*

Proof: by using the proof of the theorem, where $\mathbf{V} = \mathbf{P}$, under the multinomial case. The eigenvalues τ_{0i} will be the eigenvalues of $\mathbf{I}$, i.e.

$$\Rightarrow \tau_{0i} = 1 \qquad \forall\, i = 1, \ldots, K-1.$$

Thus

$$X^2 = \sum_{i=1}^{K-1} \tau_{0_i} Z_i^2 = \sum_{i=1}^{K-1} Z_i^2 \sim \sum_{i=1}^{K-1} \chi_1^2$$

since

$$Z_i \overset{L}{\sim} N(0,1)$$

$$\Rightarrow X^2 = \sum_{i=1}^{K-1} Z_i^2 \overset{L}{\sim} \chi^2_{K-1}. \tag{2.21}$$

Hence X^2 is distributed asymptotically as chi-square distribution with $(K-1)$ d.f. for $\mathbf{p} = \mathbf{p}_0$.■

Corollary 3

$$X^2/\tau_{0_1} \leq \sum_{i=1}^{K-1} Z_i^2 \tag{2.22}$$

where $\sum_{i=1}^{K-1} Z_i^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{V}_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0)$ *is distributed asymptotically as* χ^2_{K-1} *under* H_0.

Proof:

$$X^2 = \sum_{i=1}^{K-1} \tau_{0_i} Z_i^2; \quad \text{we know that } \tau_{0_1} \geq \tau_{0_2} \geq \ldots \geq \tau_{0_{K-1}} > 0$$

$$X^2/\tau_{0_1} = \sum_{i=1}^{K-1} (\frac{\tau_{0_i}}{\tau_{0_1}}) Z_i^2 \qquad \text{where } 0 < (\frac{\tau_{0_i}}{\tau_{0_1}}) \leq 1$$

$$\leq \sum_{i=1}^{K-1} Z_i^2. \tag{2.23}$$

■

Using Corollary 3 we can construct a conservative test if we can find an upper bound for the design effect of any linear combination of $\hat{p}_i$'s, $\mathbf{c}'\hat{\mathbf{p}}$, or obtain a

consistent estimate $\hat{\tau}_1$ of τ_1, where we can write τ_1 as

$$\begin{aligned}\tau_1 &= \sup_{\mathbf{c}}[\mathbf{c}'\mathbf{V}\mathbf{c}/\mathbf{c}'\mathbf{P}\mathbf{c}] \\ &= \sup_{\mathbf{c}}\left[\frac{V\left(\sum_{i=1}^{K-1} c_i\hat{p}_i\right)}{V_{srs}\left(\sum_{i=1}^{K-1} c_i n_i/n\right)}\right]\end{aligned} \tag{2.24}$$

where, V_{srs} denotes the variance operator under srs with replacement (multinomial sampling).

Corollary 4

$$X^2/\tau \approx \chi^2_{K-1} \tag{2.25}$$

for $\mathbf{p}_0$ *if and only if* $\mathbf{V} = \tau\mathbf{P}$ *for some constant* τ, *that is,* $var(\hat{p}_i) = \tau p_i(1-p_i)/n$ *and* $cov(\hat{p}_i, \hat{p}_t) = -\tau p_i p_t/n$.

Proof: by using the proof of the theorem, where $\mathbf{V} = \tau\mathbf{P}$, we get the result of Corollary 4.■

Corollary 4 implies that for $X^2 \approx \tau\chi^2_{K-1}$, we have to satisfy the following; all individual cells have the same design effect τ, and the design effect for each of the covariance terms must also be equal to τ.

2.7.1 Simple random sample without replacement

Consider an observed, simple random sample, srs, of n counts without replacement from a finite population of size N units. Here, $\hat{\mathbf{p}}$ is approximately $(K-1)$-variate normal with mean $\mathbf{p}$ and covariance matrix $\mathbf{V}/n$. The covariance matrix $\mathbf{V}/n$ under this design is equal to,

$$\mathbf{V} = \frac{N}{(N-1)}\left(1 - \frac{n}{N}\right)\mathbf{P}. \tag{2.26}$$

Proof:

Let a_j be a random variable which takes the value 1 if the j^{th} unit in the population is in the sample and the value zero otherwise. Using the random variable a_j, we can rewrite $\hat{p}_i$ as,

$$\hat{p}_i = \frac{\sum_{j=1}^{N} a_j Y_{ij}}{n}$$

where the sum extends over all finite population units N. With this expression the a_j's are random variables and the Y_{ij} are a set of fixed numbers equal to zero or one, i.e.

$$Y_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ population element } \epsilon\ i^{th} \text{ class} \\ 0 & \text{otherwise.} \end{cases}$$

This implies, that a_j has a Bernoulli distribution with probability of success $\frac{n}{N}$. Therefore,

$$E(a_j) = \frac{n}{N} \qquad \& \qquad V(a_j) = \frac{n}{N}\left(1 - \frac{n}{N}\right). \tag{2.27}$$

To find $V(\hat{p}_i)$ we need the covariance of a_j and a_t, where the product $a_j a_t$ is equal to 1 if the j^{th} and t^{th} units are in the sample and zero otherwise. Here,

$$\begin{aligned} Cov(a_j a_t) &= E(a_j a_t) - E(a_j)E(a_t) \\ &= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 \\ &= -\frac{n}{N(N-1)}\left(1 - \frac{n}{N}\right) \end{aligned} \tag{2.28}$$

If we consider $\mathbf{Y}$ to be a $(K-1)$ by N matrix, then

$$\hat{\mathbf{p}} = \frac{1}{n}\mathbf{Y}\ \mathbf{a}$$

where $\mathbf{a} = (a_1, \ldots, a_N)'$. Hence,

$$\begin{aligned} Cov(\hat{\mathbf{p}}) &= Cov\left(\frac{1}{n}\mathbf{Y}\ \mathbf{a}\right) \\ &= \frac{1}{n^2}\mathbf{Y}\ \mathbf{V(a)}\ \mathbf{Y}' \end{aligned}$$

where $\mathbf{V(a)}$ is equal to,

$$\mathbf{V(a)} = \frac{n}{N}\left(1-\frac{n}{N}\right)\left[\frac{N}{N-1}\mathbf{I}-\frac{1}{N-1}\mathbf{J}\right]. \tag{2.29}$$

Hence,

$$\begin{aligned} Cov(\hat{\mathbf{p}}) &= \mathbf{Y}\,\frac{1}{nN}\left(1-\frac{n}{N}\right)\left[\frac{N}{N-1}\mathbf{I}-\frac{1}{N-1}\mathbf{J}\right]\,\mathbf{Y}' \\ &= \frac{1}{nN(N-1)}\left(1-\frac{n}{N}\right)[N\mathbf{YY}'-\mathbf{YJY}'] \\ &= \frac{N}{n(N-1)}\left(1-\frac{n}{N}\right)\left[\frac{1}{N}\mathbf{YY}'-\frac{1}{N^2}\mathbf{YJY}'\right] \end{aligned}$$

where,

$$\begin{aligned} \frac{1}{N}\mathbf{YY}' &= \begin{pmatrix} \frac{\sum_j^N Y_{1j}^2}{N} & \frac{\sum_j^N Y_{1j}Y_{2j}}{N} & \cdots & \frac{\sum_j^N Y_{1j}Y_{(K-1)j}}{N} \\ \frac{\sum_j^N Y_{1j}Y_{2j}}{N} & \frac{\sum_j^N Y_{2j}^2}{N} & \cdots & \frac{\sum_j^N Y_{2j}Y_{(K-1)j}}{N} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\sum_j^N Y_{1j}Y_{(K-1)j}}{N} & & \cdots & \frac{\sum_j^N Y_{(K-1)j}^2}{N} \end{pmatrix} \\ &= diag(\mathbf{p}) \end{aligned}$$

since $Y_{ij}Y_{tj} = 0$ for all $i \neq t$, and $\frac{\sum_j^N Y_{ij}^2}{N} = \frac{\sum_j^N Y_{ij}}{N} = p_i$. Also,

$$\begin{aligned} \frac{1}{N^2}\mathbf{YJY}' &= \frac{1}{N^2}\begin{pmatrix} \left(\sum_j^N Y_{1j}\right)^2 & \sum_j^N Y_{1j}\sum_j^N Y_{2j} & \cdots & \sum_j^N Y_{1j}\sum_j^N Y_{(K-1)j} \\ \sum_j^N Y_{1j}\sum_j^N Y_{2j} & \left(\sum_j^N Y_{2j}\right)^2 & \cdots & \sum_j^N Y_{2j}\sum_j^N Y_{(K-1)j} \\ \vdots & \vdots & & \vdots \\ \sum_j^N Y_{1j}\sum_j^N Y_{(K-1)j} & & \cdots & \left(\sum_j^N Y_{(K-1)j}\right)^2 \end{pmatrix} \\ &= \begin{pmatrix} p_1^2 & p_1p_2 & \cdots & p_1p_{k-1} \\ p_1p_2 & p_2^2 & \cdots & p_2p_{k-1} \\ \vdots & \vdots & & \vdots \\ p_1p_{k-1} & & \cdots & p_{K-1}^2 \end{pmatrix} \\ &= \mathbf{pp}'. \end{aligned}$$

Therefore,

$$Cov(\hat{\mathbf{p}}) = \frac{N}{(N-1)}\left(1-\frac{n}{N}\right)\left[\frac{1}{n}\left(diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'\right)\right]$$
$$= \frac{N}{(N-1)}\left(1-\frac{n}{N}\right)\frac{\mathbf{P}}{n}. \tag{2.30}$$

■

The factor $\frac{N}{(N-1)}$ is redundant for a large population. Therefore, using Corollary 2, where the finite population correction, $\tau = (1 - n/N)$,

$$X^2 \approx (1 - n/N)\chi^2_{K-1}$$

as both N and $n \to \infty$ in such a way $N - n \to \infty$. So, the corrected Pearson chi-squared statistic for finite population sampling will be,

$$X_c^2 = (1 - n/N)^{-1}X^2 \sim \chi^2_{K-1} \tag{2.31}$$

using, $\tau_i = \tau_{0i} = 1 - n/N \quad \forall i = 1, \ldots, K-1$ for any $\mathbf{P}$.

2.7.2 Stratified random sampling (proportional allocation)

More often, sampling is done independently within several subpopulations. For designs involving stratification, we consider the subpopulations to be indexed by $l = 1, ..., L$. Therefore, the population of N units is divided into subpopulations of $N_1, N_2, ..., N_L$ units, respectively, where $\sum_{l=1}^{L} N_l = N$. These subpopulations are not overlapping, and called strata. A simple random sample s_l of size m_l is drawn with replacement from the l^{th} stratum , where $\sum_l m_l = n$; For more detail see Cochran (1977) and Hansen *et al.* (1953). Consider n_{li} $(i = 1, \ldots, K)$ to be the observed cell frequency in stratum l and let p_{li} be the proportion of elements from stratum l belonging to category i. If the sampling design is stratification with proportional allocation W_l of the m_l, then

$$p_i = \sum_{l=1}^{L} W_l p_{li}, \text{ where } \quad W_l = \frac{m_l}{n} = \frac{N_l}{N}.$$

Therefore,

$$\hat{p}_i = \sum_{l=1}^{L} W_l \frac{n_{li}}{n_l} = n_i/n, \text{ where } \quad n_i = \sum_{l=1}^{L} n_{li}.$$

		1	2	...	i	...	K	$Total$
s	1	n_{11} $\hat{p}_{11}$	n_{12} $\hat{p}_{12}$	...	n_{1i} $\hat{p}_{1i}$	...	n_{1K} $\hat{p}_{1K}$	m_1
t	2	n_{21} $\hat{p}_{21}$	n_{22} $\hat{p}_{22}$	...	n_{2i} $\hat{p}_{2i}$	...	n_{2K} $\hat{p}_{2K}$	m_2
r	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
a	l	n_{l1} $\hat{p}_{l1}$	n_{l2} $\hat{p}_{l2}$	...	n_{li} $\hat{p}_{li}$	...	n_{lK} $\hat{p}_{lK}$	$m_l = nW_l$
t	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
u	L	n_{L1} $\hat{p}_{L1}$	n_{L2} $\hat{p}_{L2}$	...	n_{Li} $\hat{p}_{Li}$	...	n_{LK} $\hat{p}_{LK}$	m_L
m	$Total$	n_1 $\hat{p}_1$	n_2 $\hat{p}_2$	...	n_i $\hat{p}_i$	...	n_K $\hat{p}_K$	$n = \sum_i \sum_l n_{li}$

With this design $\hat{\mathbf{p}}$ is approximately $(K-1)$-variate normal, when m_l is large with mean $\mathbf{p}$ and covariance matrix $\mathbf{V}/n$. The covariance matrix $\mathbf{V}/n$ under this design is equal to,

$$\mathbf{V} = \mathbf{P} - \sum_{l=1}^{L} W_l(\mathbf{p}_l - \mathbf{p})(\mathbf{p}_l - \mathbf{p})', \ (= \mathbf{P} - \mathbf{H}). \tag{2.32}$$

Proof,

$$\begin{aligned} cov(\hat{p}_i, \hat{p}_j) &= cov[\sum_{l=1}^{L} W_l \hat{p}_{li}, \sum_{r=1}^{L} W_r \hat{p}_{rj}] \\ &= \sum_{l=1}^{L} \sum_{r=1}^{L} W_l W_r \ cov[\hat{p}_{li}, \hat{p}_{rj}] \end{aligned}$$

but, correlation only exists within a stratum $\Rightarrow l = r$

$$cov(\hat{p}_i, \hat{p}_j) = \sum_{l=1}^{L} W_l^2 cov(\hat{p}_{li}, \hat{p}_{lj})$$
$$= \sum_{l=1}^{L} W_l^2 (v_l)_{ij}$$

where $v_l = \frac{1}{n_l}(diag(\mathbf{p}_l) - \mathbf{p}_l\mathbf{p}_l')$ and $n_l = nW_l$

$$cov(\hat{p}_i, \hat{p}_j) = \sum_{l=1}^{L} W_l^2 \left(\frac{1}{nW_l}(diag(\mathbf{p}_l) - \mathbf{p}_l\mathbf{p}_l')\right)_{ij}$$
$$= \frac{1}{n}\left(\sum_l W_l diag(\mathbf{p}_l) - \sum_l W_l\mathbf{p}_l\mathbf{p}_l'\right)_{ij}$$
$$= \frac{1}{n}\left(diag(\mathbf{p}) - \sum_l W_l\mathbf{p}_l\mathbf{p}_l'\right)_{ij}$$
$$= \frac{1}{n}\left(diag(\mathbf{p}) - \mathbf{p}\mathbf{p}' - \sum_l W_l\mathbf{p}_l\mathbf{p}_l' + \mathbf{p}\mathbf{p}'\right)_{ij}$$
$$= \frac{1}{n}\left((diag(\mathbf{p}) - \mathbf{p}\mathbf{p}') - \sum_l W_l(\mathbf{p}_l - \mathbf{p})(\mathbf{p}_l - \mathbf{p})'\right)_{ij}$$

since $\mathbf{p} = \sum_{l=1}^{L} W_l\mathbf{p}_l$,

$$cov(\hat{p}_i, \hat{p}_j) = \frac{1}{n}\left(\mathbf{P} - \sum_{l=1}^{L} W_l(\mathbf{p}_l - \mathbf{p})(\mathbf{p}_l - \mathbf{p})'\right)_{ij}$$
$$= \frac{1}{n}(v)_{ij}.$$

Hence,

$$\mathbf{V} = \mathbf{P} - \sum_{l=1}^{L} W_l(\mathbf{p}_l - \mathbf{p})(\mathbf{p}_l - \mathbf{p})'$$

where $\mathbf{p}_l = (p_{l_1}, \ldots, p_{l_{K-1}})'$.■

Now, consider the eigenvalues of the generalized design effect, and the asymptotic distribution of X^2, for goodness of fit.

$$\mathbf{c}'\mathbf{V}\mathbf{c}/\mathbf{c}'\mathbf{P}\mathbf{c} \geq 0 \qquad \text{since } \mathbf{P} \text{ is p.d. and } \mathbf{V} \text{ is p.d.}$$

So,

$$\begin{aligned}\mathbf{c}'\mathbf{V}\mathbf{c}/\mathbf{c}'\mathbf{P}\mathbf{c} &= \frac{\mathbf{c}'[\mathbf{P} - \sum_{l=1}^{L} W_l(\mathbf{p}_l - \mathbf{p})(\mathbf{p}_l - \mathbf{p})']\mathbf{c}}{\mathbf{c}'\mathbf{P}\mathbf{c}} \\ &= \frac{\mathbf{c}'\mathbf{P}\mathbf{c}}{\mathbf{c}'\mathbf{P}\mathbf{c}} - \frac{\mathbf{c}'[\sum_l W_l(\mathbf{p}_l - \mathbf{p})(\mathbf{p}_l - \mathbf{p})']\mathbf{c}}{\mathbf{c}'\mathbf{P}\mathbf{c}} \\ &= 1 - \frac{\sum_{l=1}^{L} W_l[\mathbf{c}'(\mathbf{p}_l - \mathbf{p})]^2}{\mathbf{c}'\mathbf{P}\mathbf{c}} \leq 1 \end{aligned} \tag{2.33}$$

or $\tau_{0_1} \leq 1$ for any $\mathbf{p}_0$ and

$$0 \leq X^2 = \sum_{i=1}^{K-1} \tau_{0_i} Z_i^2 \leq \sum_{i=1}^{K-1} Z_i^2 \simeq \chi^2_{K-1}. \tag{2.34}$$

Thus, the Pearson statistic X^2 is always asymptotically conservative in the case of stratified random sampling. Rao and Scott (1981) consider the extent to which X^2 could be conservative;

1. If $L \geq K$ and the stratification is perfect, that is, all elements in a stratum belong to the same category, then $X^2 = 0$.

l	1	2	...	i	...	K	$Total$
1	n_{11} $\hat{p}_{11}$	0	...	0	...	0	$m_1 = n_{11}$
2	0	n_{22} $\hat{p}_{22}$	...	0	...	0	$m_2 = n_{22}$
⋮	0	0	...	n_{li} $\hat{p}_{li}$	...	0	$m_l = n_{li}$
L	0	0	...	0	...	n_{LK} $\hat{p}_{LK}$	$m_L = n_{LK}$
$Total$	$n_1 = n_{11}$	$n_2 = n_{22}$	...	$n_i = n_{li}$	...	$n_K = n_{LK}$	n

2. If $L < K$, then

$$\begin{aligned} X^2 &= \sum_{i=1}^{K} \tau_i Z_i^2 = \sum_{i=1}^{K-L} \tau_i Z_i^2 + \sum_{i=K-L+1}^{K} \tau_i Z_i^2 \\ &\geq \sum_{i=1}^{K-L} Z_i^2 \simeq \chi^2_{K-L} \end{aligned}$$

since the rank of $\mathbf{H}$, equation (2.32), is at most $L-1$. For a proof, consider a column vector $\mathbf{v}$, then the vector $\mathbf{Hv}$ lies in a space of dimension equal to the rank of $\mathbf{H}$. Now,

$$\begin{aligned}\mathbf{Hv} &= \sum_{l=1}^{L} W_l(\mathbf{p}_l-\mathbf{p})(\mathbf{p}_l-\mathbf{p})'\mathbf{v} \\ &= \sum_{l=1}^{L}\{W_l(\mathbf{p}_l-\mathbf{p})'\mathbf{v}\}(\mathbf{p}_l-\mathbf{p}).\end{aligned}$$

So, $\mathbf{Hv}$ is a linear combination of $(\mathbf{p}_l-\mathbf{p})$. Therefore, rank of $\mathbf{H}$ is equal to the dimension of the space spanned by $(\mathbf{p}_l-\mathbf{p})$, which is at most L. Because there are only L vectors of the form $(\mathbf{p}_l-\mathbf{p})$. Furthermore, these vectors are not linearly independent, since $\sum_{l=1}^{L} W_l(\mathbf{p}_l-\mathbf{p}) = 0$. Hence, they span a space of dimension at most $L-1$.■

Since the rank of $\mathbf{H}$ is at most $L-1$, and therefore at least $K-L$ of the τ_i's must be equal to one, see equation (2.33). Thus X^2 is asymptotically well approximated by χ^2_{K-1} if K is large and L be relatively small.

As an example, when $L = 2$

$$\begin{aligned}\mathbf{D} &= \mathbf{P}^{-1}\mathbf{V} \\ &= \mathbf{P}^{-1}\left(\mathbf{P} - \sum_{l=1}^{L} W_l(\mathbf{p}_l-\mathbf{p})(\mathbf{p}_l-\mathbf{p})'\right) \\ &= \mathbf{I} - W_1W_2\mathbf{P}^{-1}(\mathbf{p}_1-\mathbf{p}_2)(\mathbf{p}_1-\mathbf{p}_2)' \\ &= \mathbf{I} - \mathbf{A}. \qquad (2.35)\end{aligned}$$

We can prove this using $\mathbf{p} = \sum_l W_l\mathbf{p}_l = W_1\mathbf{p}_1 + W_2\mathbf{p}_2$, and $W_2 = 1 - W_1$. Since $\mathbf{A}$ is of rank one $(2-1)$, $K-2$ of its eigenvalues are zero and the remaining non zero eigenvalue is,

$$tr(\mathbf{A}) = W_1W_2\sum_{i=1}^{K}(p_{1i} - p_{2i})^2/p_i = \delta^*$$

where $0 \leq \delta^* \leq 1$. Hence $\tau_1 = \ldots = \tau_{K-2} = 1$ and $\tau_{K-1} = 1 - \delta^*$. This implies

$$X^2 = \sum_{i=1}^{K-2} Z_i^2 + (1 - \delta_0^*) Z_{K-1}^2$$
$$\approx \chi_{K-2}^2 + (1 - \delta_0^*)\chi_1^2 \qquad (2.36)$$

where δ_0^* is the value of δ^* under H_0 (for $\mathbf{p} = \mathbf{p}_0$). Unless K is small, X^2 is asymptotically well approximated by χ_{K-1}^2 in the two strata case.

2.7.3 Two stage sampling

Suppose we have C primary sampling unit, psu's, with M_t secondary units in the t^{th} psu ($t = 1, \ldots, C; \sum_t M_t = N$). Consider the following sample, c ($c < C$) psu's selected with replacement, with probability W_t proportional to size M_t, as a first-stage. Then as a second-stage, considering subsamples each of size $n_t = m; t = 1, \ldots, c$, ($m \subset M_t$) drawn as simple random samples with replacement independently from each selected psu. Hence, the sample size is $n = mc$; For more detail see Cochran (1977) and Hansen *et al.* (1953).

Let n_{ti} be the observed cell frequencies in sampled psu t and let p_{ti} be the proportion of the t^{th} psu belonging to category i, where $i = 1, \ldots, K$. Under this design, we have

$$\hat{\mathbf{p}} = \sum_{t=1}^{c} \frac{\hat{\mathbf{p}}_t}{c} = \mathbf{n}/n \quad \text{, where} \quad \hat{\mathbf{p}}_t' = (\hat{p}_{t1}, \hat{p}_{t2}, \ldots, \hat{p}_{tK-1}) \quad ; E(n_i) = np_i \quad ;$$

$$\hat{p}_{ti} = \frac{n_{ti}}{m} \quad ; \quad p_i = \sum_t^C \frac{M_t}{N} p_{ti} = \sum_t^C W_t p_{ti}.$$

	1	...	i	...	K	*Total*
1	$n_{11}\ \hat{p}_{11}$	...	$n_{1i}\ \hat{p}_{1i}$	...	$n_{1K}\ \hat{p}_{1K}$	m
⋮	⋮	⋮	⋮	⋮	⋮	⋮
t	$n_{t1}\ \hat{p}_{t1}$	...	$n_{ti}\ \hat{p}_{ti}$	...	$n_{tK}\ \hat{p}_{tK}$	m
⋮	⋮	⋮	⋮	⋮	⋮	⋮
c	$n_{c1}\ \hat{p}_{c1}$	...	$n_{ci}\ \hat{p}_{ci}$	...	$n_{cK}\ \hat{p}_{cK}$	m
Total	$n_1\ \hat{p}_1$	...	$n_i\ \hat{p}_i$	...	$n_K\ \hat{p}_K$	n

With this design, $\hat{\mathbf{p}}$ is approximately $(K-1)$-variate normal, for large c, with mean $\mathbf{p}$ and covariance matrix $\mathbf{V}/n$, where

$$\begin{aligned}\mathbf{V} &= \mathbf{P} + (m-1)\sum_{t=1}^{C} W_t(\mathbf{p}_t - \mathbf{p})(\mathbf{p}_t - \mathbf{p})' \\ &= \mathbf{P} + (m-1)\mathbf{A}.\end{aligned} \tag{2.37}$$

Proof,

We know that,

$$V(\hat{\mathbf{p}}) = V[E(\hat{\mathbf{p}}|\text{selection of psus})] + E[V(\hat{\mathbf{p}}|\text{selection of psus})]. \tag{2.38}$$

Starting with the first-stage sampling,

$$E(\hat{\mathbf{p}}) = \frac{1}{c}\sum_{t=1}^{c} E(\hat{\mathbf{p}}_t)$$

$$E(\hat{\mathbf{p}}|\text{selection of psus}) = \frac{1}{c}\sum_{t=1}^{C} K_t\mathbf{p}_t$$

where K_t = Number of times t^{th} psu is selected.

$$= \frac{1}{c}\tilde{\mathbf{P}}\mathbf{K}$$

where $\tilde{\mathbf{P}} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_C)$, and $\mathbf{K}' = (K_1, K_2, \ldots, K_C)$.

$$V(\hat{\mathbf{p}}|\text{selection of psus}) = \frac{1}{c^2}\sum_{t=1}^{C} K_t V(\hat{\mathbf{p}}_t)$$

$$= \frac{1}{c^2}\sum_{t=1}^{C} K_t[\frac{1}{m}(diag(\mathbf{p}_t) - \mathbf{p}_t\mathbf{p}_t')].$$

Now, for the second-stage,

$$
\begin{aligned}
V(E(\hat{\mathbf{p}}|\text{selection of psus})) &= V(\frac{1}{c}\tilde{\mathbf{P}}\mathbf{K}) \\
&= \frac{1}{c^2}\tilde{\mathbf{P}}V(\mathbf{K})\tilde{\mathbf{P}}' \\
&\mathbf{K} \sim \text{multinomial}(c, \mathbf{W}) \text{ where } W_t = \frac{M_t}{N} \\
&\Rightarrow V(\mathbf{K}) = c\,(diag(\mathbf{W}) - \mathbf{W}\mathbf{W}') \\
V(E(\hat{\mathbf{p}}|\text{selection of psus})) &= \frac{1}{c}\tilde{\mathbf{P}}(diag(\mathbf{W}) - \mathbf{W}\mathbf{W}')\tilde{\mathbf{P}}' \\
&= \frac{1}{c}\tilde{\mathbf{P}}(diag(\mathbf{W}))\tilde{\mathbf{P}} - \frac{1}{c}\tilde{\mathbf{P}}\mathbf{W}\mathbf{W}'\tilde{\mathbf{P}}' \\
&= \frac{1}{c}\sum_{t=1}^{C} W_t\mathbf{p}_t\mathbf{p}_t' - \frac{1}{c}\mathbf{p}\mathbf{p}' \\
&\text{since } \tilde{\mathbf{P}}diag(\mathbf{W})\tilde{\mathbf{P}}' = \sum_{t=1}^{C} W_t\mathbf{p}_t\mathbf{p}_t' \text{ and } \tilde{\mathbf{P}}\mathbf{W} = \mathbf{p}.
\end{aligned}
$$

Also,

$$
\begin{aligned}
E(V(\hat{\mathbf{p}}|\text{selection of psus})) &= E(\frac{1}{c^2}\sum_{t=1}^{C} K_t\frac{1}{m}(diag(\mathbf{p}_t) - \mathbf{p}_t\mathbf{p}_t')) \\
&= \frac{1}{c^2}\sum_{t=1}^{C} E(K_t)\frac{1}{m}(diag(\mathbf{p}_t) - \mathbf{p}_t\mathbf{p}_t') \quad ;\text{where } E(K_t) = cW_t \\
&= \frac{1}{cm}\sum_{t=1}^{C} W_t diag(\mathbf{p}_t) - \frac{1}{cm}\sum_{t=1}^{C} W_t\mathbf{p}_t\mathbf{p}_t'.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
V(\hat{\mathbf{p}}) &= (\frac{1}{c}\sum_{t=1}^{C} W_t\mathbf{p}_t\mathbf{p}_t' - \frac{1}{c}\mathbf{p}\mathbf{p}') \\
&\quad + (\frac{1}{cm}\sum_{t=1}^{C} W_t diag(\mathbf{p}_t) - \frac{1}{cm}\sum_{t=1}^{C} W_t\mathbf{p}_t\mathbf{p}_t') \\
&= (\frac{1}{cm}(m\sum_{t=1}^{C} W_t\mathbf{p}_t\mathbf{p}_t' - m\mathbf{p}\mathbf{p}' + \sum_{t=1}^{C} W_t diag(\mathbf{p}_t) - \sum_{t=1}^{C} W_t\mathbf{p}_t\mathbf{p}_t')
\end{aligned}
$$

where, $\sum_{t=1}^{C} W_t diag(\mathbf{p}_t) = diag(\mathbf{p})$ and $n = mc$.

$$\begin{aligned} V(\hat{\mathbf{p}}) &= \frac{1}{n}((m-1)\sum_{t=1}^{C} W_t \mathbf{p}_t \mathbf{p}_t' - m\mathbf{p}\mathbf{p}' + diag(\mathbf{p})) \\ &= \frac{1}{n}(diag(\mathbf{p}) - \mathbf{p}\mathbf{p}' + (m-1)\sum_{t=1}^{C} W_t \mathbf{p}_t \mathbf{p}_t' - (m-1)\mathbf{p}\mathbf{p}') \\ &= \frac{1}{n}\left(\mathbf{P} + (m-1)(\sum_{t=1}^{C} W_t \mathbf{p}_t \mathbf{p}_t' - \mathbf{p}\mathbf{p}')\right) \\ &= \frac{1}{n}\left(\mathbf{P} + (m-1)\sum_{t=1}^{C} W_t (\mathbf{p}_t - \mathbf{p})(\mathbf{p}_t - \mathbf{p})'\right) \quad , \text{since } \mathbf{p} = \sum_{t=1}^{C} W_t \mathbf{p}_t \\ &= \mathbf{V}/n \end{aligned}$$

So,

$$\mathbf{V} = \mathbf{P} + (m-1)\sum_{t=1}^{C} W_t (\mathbf{p}_t - \mathbf{p})(\mathbf{p}_t - \mathbf{p})'.$$

■

The generalized design effect matrix $\mathbf{D} = \mathbf{P}^{-1}\mathbf{V}$ is

$$\mathbf{D} = \mathbf{P}^{-1}[\mathbf{P} + (m-1)\mathbf{A}]$$

where $\mathbf{A} = \sum_{t=1}^{C} W_t (\mathbf{p}_t - \mathbf{p})(\mathbf{p}_t - \mathbf{p})'$

$$\begin{aligned} \mathbf{D} &= \mathbf{P}^{-1}\mathbf{P} + (m-1)\mathbf{P}^{-1}\mathbf{A} \\ &= \mathbf{I} + (m-1)\mathbf{P}^{-1}\mathbf{A}. \end{aligned}$$

Let $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_{K-1} \geq 0$ be the eigenvalues of $\mathbf{P}^{-1}\mathbf{A}$, which implies, that

$$\tau_i = 1 + (m-1)\rho_i \ (i = 1, \ldots, K-1).$$

This gives,

$$X^2 = \sum_{i=1}^{K-1} [1 + (m-1)\rho_{0_i}] Z_i^2$$

where, ρ_{0_i} is the value of ρ_i for $p = p_0$. Now, for the asymptotic distribution,

$$0 \leq \frac{\mathbf{c'Vc}}{\mathbf{c'Pc}} = 1 + (m-1)\frac{\mathbf{c'Ac}}{\mathbf{c'Pc}}$$

and

$$\frac{\mathbf{c'Ac}}{\mathbf{c'Pc}} \leq 1 \Rightarrow \rho_{0_1} \leq 1.$$

Using this, we obtain

$$\begin{aligned}
\sum_{i=1}^{K-1} Z_i^2 &\leq [1 + (m-1)\rho_{0_{K-1}}] \sum_{i=1}^{K-1} Z_i^2 \\
&\leq X^2 \\
&\leq [1 + (m-1)\rho_{0_1}] \sum_{i=1}^{K-1} Z_i^2 \text{ , since } \rho_{0_1} \text{ is the largest.} \\
&\leq m \sum_{i=1}^{K-1} Z_i^2 \text{ , since } 0 \leq \rho_{0_1} \leq 1.
\end{aligned}$$

Thus X^2/m gives an asymptotically conservative test whatever the values of the ρ_{0_i}'s.

$$X^2 \leq m \sum Z_i^2 \Rightarrow X^2/m \leq \sum Z_i^2 \cong \chi^2_{K-1}.$$

If we can specify ρ_{0_1} or a consistent estimator $\hat{\rho}_1$, a better conservative test can be obtained. Rao and Scott (1981) called ρ_i's generalized measures of homogeneity.

2.8 Modifications to X^2

To modify X^2 we need to know the τ_{0_i}'s or consistent estimates for τ_i's, which is almost equivalent to knowing the full covariance matrix $\mathbf{V}_0$ under H_0 or $\hat{\mathbf{V}} = (\hat{v}_{ij})$, since the τ_{0_i}'s are the eigenvalues of $\mathbf{D} = \mathbf{P}_0^{-1}\mathbf{V}_0$. Moreover, if we had $\mathbf{V}_0$, then we could construct a generalized Wald statistic. Knowledge of τ_{0_i}'s

or $\hat{\tau}_i$'s are often not available in practice. Therefore, a simple approximation to the asymptotic distribution of X^2 is needed, which requires only very limited information about $\hat{\mathbf{V}}$. Rao and Scott's approach (Rao and Scott, 1981) is to construct a modified statistic as,

$$X_c^2 = X^2/\hat{\tau}_. \tag{2.39}$$

where

$$\hat{\tau}_. = \sum_{i=1}^{K-1} \frac{\hat{\tau}_i}{(K-1)}$$

and under H_0

$$\hat{\tau}_{0.} = \sum_{i=1}^{K-1} \frac{\hat{\tau}_{0_i}}{(K-1)}. \tag{2.40}$$

X_c^2, under H_0, is distributed asymptotically as

$$T = \sum_{i=1}^{K-1} (\tau_{0_i}/\tau_{0.}) Z_i^2 \sim \chi^2_{K-1}. \tag{2.41}$$

The random variable T has the same expectation as χ^2_{K-1}, but the variance

$$Var(T) = 2(K-1) + 2\sum_{i=1}^{K-1} \left(\frac{\tau_{0_i} - \tau_{0.}}{\tau_{0.}} \right)^2,$$

is larger than $Var(\chi^2_{K-1}) = 2(K-1)$, unless all τ_{0_i}'s are equal. To prove this, we compute the expected value and variance of T ,

$$\begin{aligned} E(T) &= E\Big(\sum_{i=1}^{K-1} \frac{\tau_{0_i}}{\tau_{0.}} Z_i^2\Big) = \frac{1}{\tau_{0.}} \sum_{i=1}^{K-1} \tau_{0_i} E(Z_i^2) \\ &= \frac{1}{\tau_{0.}} \sum_{i=1}^{K-1} \tau_{0_i} \quad ; \text{since } Z_i^2 \sim \chi_1^2 \Longrightarrow E(Z_i^2) = 1 \text{ and } V(Z_i^2) = 2 \\ &= \frac{(K-1)\sum \tau_{0_i}}{\sum \tau_{0_i}} \\ &= (K-1). \end{aligned} \tag{2.42}$$

The variance is

$$\begin{aligned} Var(T) &= V(\sum_{i=1}^{K-1} \frac{\tau_{0_i}}{\tau_{0.}} Z_i^2) \\ &= \frac{\sum \tau_{0_i}^2}{\tau_{0.}^2} V(Z_i^2) \qquad \text{, since } Z_i^2 \text{ 's are independent} \\ &= 2\frac{\sum \tau_{0_i}^2}{\tau_{0.}^2} \\ &= 2(K-1) + \frac{2}{\tau_{0.}^2} \sum_{i=1}^{K-1} \tau_{0_i}^2 - 2(K-1) \\ &= 2(K-1) + \frac{2}{\tau_{0.}^2} [\sum_{i=1}^{K-1} \tau_{0_i}^2 - (K-1)\tau_{0.}^2] \\ &= 2(K-1) + \frac{2}{\tau_{0.}^2} \sum_{i=1}^{K-1} (\tau_{0_i} - \tau_{0.})^2. \end{aligned}$$

Thus,

$$Var(T) = 2(K-1) + 2\sum_{i=1}^{K-1} \left(\frac{\tau_{0_i} - \tau_{0.}}{\tau_{0.}} \right)^2.$$

■

So, treating X_c^2 as χ_{K-1}^2 under H_0 tends to underestimate the upper percentage point of true asymptotic distribution, since

$$Var(T) > Var(\chi_{K-1}^2). \tag{2.43}$$

If the coefficients of variation, τ_i, are not large, the effect will be small. With this approach, the modification X_c^2 depends on $\hat{\tau}_0$ which depends on the cell estimated variance $\hat{v}_{ii}$ or equivalently the estimated cell design effects $\hat{d}_1 \ldots, \hat{d}_K$.

$$\begin{aligned} \hat{\tau}_{.} &= tr(\hat{\mathbf{P}}^{-1}\hat{\mathbf{V}})/(K-1) \qquad (X_c^2 = X^2/\hat{\tau}_{0.}) \\ &= \sum_{i=1}^{K} [\hat{v}_{ii}/\hat{p}_i(K-1)] \\ &= \sum_{i=1}^{K} \frac{(1-\hat{p}_i)\hat{d}_i}{(K-1)} \; ; \quad \text{where} \quad \hat{d}_i = \frac{\hat{v}_{ii}}{[\hat{p}_i(1-\hat{p}_i)]}. \end{aligned} \tag{2.44}$$

In empirical results for large study, 13000 households, analyzed by Rao and Scott, they found the modified test based on X_c^2 gives good results, for nominal size $\alpha = 0.05$. Also, if we use the ordinary X^2 test with a desired significance level of five percent, the estimated (asymptotic) significance level can be as high as 41 percent. With a fairly small value for $\hat{\tau}_. = \sum_i \frac{\hat{\tau}_i}{(K-1)}$, there can be a serious distortion in size if the degrees of freedom, $K-1$, are large.

2.9 Dirichlet-Multinomial distribution

Another classical approach for data from cluster samples was presented by Cohen (1976), Altham (1976), and Brier (1980). Cohen (1976) proposed a simple model for within-cluster dependence for c psus or clusters of size two. Cohen (1976) considered the following model for correlated responses within the same clusters for cluster of size 2,

$$p_{ij} = \begin{cases} \rho p_i + (1-\rho)p_i^2 & \text{if } i = j \\ (1-\rho)p_i p_j & \text{if } i \neq j \end{cases} \tag{2.45}$$

where $p_{ij} = pr(\text{first sibling is in category } i \text{ and the second sibling of the same cluster is in category } j)$ and ρ is the measure of within-cluster dependence, $p_{ij} \geq 0$ $\forall i, j = 1, \ldots, K$, since $\rho < 0$ might not be logical in practice. Altham (1976) extended the Cohen results for a larger cluster size, $n_t = m$. Altham obtained

$$\begin{aligned} \mathbf{V} &= m(1+\rho(m-1))\mathbf{P} \\ &= ma\mathbf{P} \end{aligned} \tag{2.46}$$

where m is equal to the cluster size, $a = (1+\rho(m-1))$, and $\mathbf{P}$ is the conventional covariance matrix for $\mathbf{n}$ under multinomial sampling. Hence

$$\begin{aligned} (c\mathbf{V})^{-1} &= \frac{1}{cma}\mathbf{P}^{-1} \\ &= \frac{1}{a}(n\mathbf{P})^{-1}. \end{aligned}$$

Therefore,

$$\begin{aligned}(\mathbf{n}-n\mathbf{p})'(c\mathbf{V})^{-1}(\mathbf{n}-n\mathbf{p}) &= \frac{1}{a}(\mathbf{n}-n\mathbf{p})'(n\mathbf{P})^{-1}(\mathbf{n}-n\mathbf{p}) \\ &= \frac{1}{a}\sum_{i=1}^{K}\frac{(n_i-np_i)^2}{np_i} \\ &= \frac{1}{a}X^2.\end{aligned}$$

Therefore, with simple correction, a, the standard Pearson chi-squared test statistic, X^2, computed under multinomial sampling, can be use for testing a simple null hypothesis for $\mathbf{p}$ under cluster sampling. Applying this requires that (2.45) holds, and ρ is known.

On the other hand, Brier (1980) provided an alternative justification for some results of Altham (1976) and extended the results to the case of unequal cluster sizes. Brier assumed that $\mathbf{p}_t$ are independent and identically distributed with $\mathbf{p}$ ~Dirichlet($d\boldsymbol{\pi}$), see (4.1), where $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_K)$ lies in the $K-1$ dimensional simplex defined by $\Psi_K = \{\ (p_1, ..., p_K) : p_i > 0, \sum_{i=1}^{K} p_i = 1\}$. He, also, assumed that $\mathbf{n}_t$, for any psu, is multinomially distributed, conditional on $\mathbf{p}_t$ for that psu. That is for any cluster, $\mathbf{n}$ ~Multin(m; $\mathbf{p}$), see (2.1). Thus the unconditional distribution of $\mathbf{n}$ is then,

$$\begin{aligned}f(\mathbf{n}|\boldsymbol{\pi},d) &= \int pr(\mathbf{n}|\mathbf{p}) \times pr(\mathbf{p}|\boldsymbol{\pi},d)\ d\mathbf{p} \\ &= \int \binom{m}{n_1, \ldots, n_K} \prod_{i=1}^{K} p_i^{n_i} \times \frac{\Gamma[d]}{\prod_i \Gamma(d\pi_i)} \prod_i p_i^{d\pi_i - 1}\ d\mathbf{p} \\ &= \binom{m}{n_1, \ldots, n_K} \left(\frac{\Gamma[d]}{\prod_i \Gamma(d\pi_i)}\right) \int \prod_{i=1}^{K} p_i^{n_i + d\pi_i - 1}\ d\mathbf{p} \\ &= \binom{m}{n_1, \ldots, n_K} \left(\frac{\Gamma[d]}{\Gamma[m+d]}\right) \prod_{i=1}^{K} \left(\frac{\Gamma(n_i + d\pi_i)}{\Gamma(d\pi_i)}\right) \\ &= \mathrm{DM}_K(m, \boldsymbol{\pi}, d). \end{aligned} \tag{2.47}$$

Brier referred to this distribution as the Dirichlet-Multinomial distribution, and denoted by $\mathrm{DM}_K(m, \boldsymbol{\pi}, d)$. Mosimann (1962) shows that the mean of $\mathrm{DM}_K(m, \boldsymbol{\pi}, d)$

is $m\boldsymbol{\pi}$ and the covariance matrix is $mb(diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}')$, where $b = \frac{m+d}{1+d}$. Thus the covariance matrix is a constant, b, times the corresponding multinomial covariance matrix, $\mathbf{P}$. This means that the constant, b, will play an important role in the design effect function. Also, d is the structural parameter representing the cluster effect. He also considers the intraclass correlation coefficient for this distribution, $\rho = \frac{1}{1+d}$. This provides a justification for Altham's (1976) results, since by substituting the value of ρ in b will give

$$\begin{aligned} b &= \frac{m+d}{1+d} \\ &= \frac{m + \frac{1-\rho}{\rho}}{1 + \frac{1-\rho}{\rho}} \\ &= 1 + \rho(m-1) = a. \end{aligned}$$

Brier's (1980) Dirichlet-multinomial model is fairly restrictive since it implies the same design effect for all individual cells. For the Dirichlet-multinomial model Koehler and Wilson (1986) compared the results of using the exact covariance matrix with the Pearson chi-squared, Rao and Scott (1981) corrected Pearson chi-squared, and Wald statistics. Using the exact covariance matrix gives similar result to Rao and Scott (1981) corrected Pearson chi-squared, and both are very close to the Wald statistic.

2.10 Testing independence

Let X and Y denote two categorical dependent discrete random variables, where X has r levels and Y has c levels, with $K = r \times c$ possible combinations of classifications. The responses (x, y) are assumed randomly chosen from population having a joint probability distribution. The K categories are presented in a rectangular table having r rows for the category of X and c columns for Y. If the table contains frequency counts of outcomes, the table is called a contingency

table. A contingency table having r rows and c columns is referred to as r-by-c (or $r \times c$) table.

The two variables X and Y are said to be stochastically independent if the conditional distributions of Y are identical at each level of X. If the variables are response variables, their relationship can be described by the conditional distribution of one variable given the other, or by their joint distribution function. When the property of all joint probabilities equals the product of their marginal probabilities, i.e. $p_{ij} = p_{i+}p_{+j}$, where $p_{i+} = \sum_{j=1}^{c} p_{ij}$ and $p_{+j} = \sum_{i=1}^{r} p_{ij}$, then the two variables are called independent.

2.10.1 Chi-squared test in two-way table

For testing independence, in a two-way contingency table ($r \times c$), the null hypothesis of interest is the independence of rows and columns,

$$H_0 : p_{ij} = p_{i+}p_{+j}, \qquad (i = 1, \ldots, r; j = 1, \ldots, c)$$

or

$$H_0 : h_{ij}(p) = 0. \tag{2.48}$$

where $h_{ij}(p) = p_{ij} - p_{i+}p_{+j}$ $(i = 1, \ldots, r-1; j = 1, \ldots, c-1)$.

Then, the usual Pearson chi-squared statistic for testing H_0 is

$$X_I^2 = n \sum_{i=1}^{r} \sum_{j=1}^{c} (\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2 / (\hat{p}_{i+}\hat{p}_{+j}). \tag{2.49}$$

If we define the vectors of marginal probabilities,

$$\mathbf{p}_r = (p_{1+}, p_{2+}, \ldots, p_{(r-1)+})'; \qquad \text{where } p_{i+} = \sum_{j=1}^{c} p_{ij}$$

$$\mathbf{p}_c = (p_{+1}, p_{+2}, \ldots, p_{+(c-1)})'; \qquad \text{where } p_{+j} = \sum_{i=1}^{r} p_{ij}$$

is $m\boldsymbol{\pi}$ and the covariance matrix is $mb(diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}')$, where $b = \frac{m+d}{1+d}$. Thus the covariance matrix is a constant, b, times the corresponding multinomial covariance matrix, $\mathbf{P}$. This means that the constant, b, will play an important role in the design effect function. Also, d is the structural parameter representing the cluster effect. He also considers the intraclass correlation coefficient for this distribution, $\rho = \frac{1}{1+d}$. This provides a justification for Altham's (1976) results, since by substituting the value of ρ in b will give

$$\begin{aligned} b &= \frac{m+d}{1+d} \\ &= \frac{m + \frac{1-\rho}{\rho}}{1 + \frac{1-\rho}{\rho}} \\ &= 1 + \rho(m-1) = a. \end{aligned}$$

Brier's (1980) Dirichlet-multinomial model is fairly restrictive since it implies the same design effect for all individual cells. For the Dirichlet-multinomial model Koehler and Wilson (1986) compared the results of using the exact covariance matrix with the Pearson chi-squared, Rao and Scott (1981) corrected Pearson chi-squared, and Wald statistics. Using the exact covariance matrix gives similar result to Rao and Scott (1981) corrected Pearson chi-squared, and both are very close to the Wald statistic.

2.10 Testing independence

Let X and Y denote two categorical dependent discrete random variables, where X has r levels and Y has c levels, with $K = r \times c$ possible combinations of classifications. The responses (x, y) are assumed randomly chosen from population having a joint probability distribution. The K categories are presented in a rectangular table having r rows for the category of X and c columns for Y. If the table contains frequency counts of outcomes, the table is called a contingency

table. A contingency table having r rows and c columns is referred to as r-by-c (or $r \times c$) table.

The two variables X and Y are said to be stochastically independent if the conditional distributions of Y are identical at each level of X. If the variables are response variables, their relationship can be described by the conditional distribution of one variable given the other, or by their joint distribution function. When the property of all joint probabilities equals the product of their marginal probabilities, i.e. $p_{ij} = p_{i+}p_{+j}$, where $p_{i+} = \sum_{j=1}^{c} p_{ij}$ and $p_{+j} = \sum_{i=1}^{r} p_{ij}$, then the two variables are called independent.

2.10.1 Chi-squared test in two-way table

For testing independence, in a two-way contingency table ($r \times c$), the null hypothesis of interest is the independence of rows and columns,

$$H_0 : p_{ij} = p_{i+}p_{+j}, \qquad (i = 1, \ldots, r; j = 1, \ldots, c)$$

or

$$H_0 : h_{ij}(p) = 0. \tag{2.48}$$

where $h_{ij}(p) = p_{ij} - p_{i+}p_{+j}$ $(i = 1, \ldots, r-1; j = 1, \ldots, c-1)$.

Then, the usual Pearson chi-squared statistic for testing H_0 is

$$X_I^2 = n \sum_{i=1}^{r} \sum_{j=1}^{c} (\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2 / (\hat{p}_{i+}\hat{p}_{+j}). \tag{2.49}$$

If we define the vectors of marginal probabilities,

$$\mathbf{p}_r = (p_{1+}, p_{2+}, \ldots, p_{(r-1)+})'; \qquad \text{where } p_{i+} = \sum_{j=1}^{c} p_{ij}$$

$$\mathbf{p}_c = (p_{+1}, p_{+2}, \ldots, p_{+(c-1)})'; \qquad \text{where } p_{+j} = \sum_{i=1}^{r} p_{ij}$$

then, the Pearson chi-squared statistic can be generalized to,

$$X_I^2 = n\,\mathbf{h}(\hat{\mathbf{p}})'(\hat{\mathbf{P}}_r^{-1}\otimes\hat{\mathbf{P}}_c^{-1})\,\mathbf{h}(\hat{\mathbf{p}})$$

where $\hat{p}_{ij}$ is the estimate of p_{ij} under the sampling design, $\mathbf{h}(\hat{\mathbf{p}})$ is the column vector of $h_{ij}(p)$'s, and $\hat{\mathbf{P}}_r$ and $\hat{\mathbf{P}}_c$ are the values of $\mathbf{P}_r = diag(\mathbf{p}_r) - \mathbf{p}_r\mathbf{p}_r'$ and $\mathbf{P}_c = diag(\mathbf{p}_c) - \mathbf{p}_c\mathbf{p}_c'$ for $\mathbf{p} = \hat{\mathbf{p}}$.

The Wald statistic for testing H_0 is given by

$$X_{I(w)}^2 = n\,\mathbf{h}(\hat{\mathbf{p}})'\hat{\mathbf{V}}_h^{-1}\mathbf{h}(\hat{\mathbf{p}}) \xrightarrow{L} \chi^2_{(r-1)(c-1)}$$

but, as we noted before it may be difficult to gain access to the full covariance matrix. As before, the asymptotic behaviour of X_I^2 under a general sampling scheme, can be represented by,

$$X_I^2 = \sum_{i=1}^{(r-1)(c-1)} \delta_i Z_i^2 \tag{2.50}$$

where Z_i's are asymptotically independent $N(0,1)$ under H_0, and δ_i's are eigenvalues of the matrix $\mathbf{D} = (\mathbf{P}_r^{-1}\otimes\mathbf{P}_c^{-1})\mathbf{V}_h$.

A modified statistic similar to X_c^2 for the goodness-of-fit problem is given by Rao and Scott (1981),

$$X_{I(c)}^2 = X_I^2/\hat{\delta}., \tag{2.51}$$

where,

$$\begin{aligned}\hat{\delta}. &= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\hat{v}_{ij}(h)}{(r-1)(c-1)\hat{p}_{i+}\hat{p}_{+j}} \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(1-\hat{p}_{i+})(1-\hat{p}_{+j})\hat{\delta}_{ij}}{(r-1)(c-1)}.\end{aligned} \tag{2.52}$$

Here $\hat{v}_{ij}(h)/n$ are the estimators of variance of $h_{ij}(\hat{p})$ and $\hat{\delta}_{ij}$ is the estimated design effect of $h_{ij}(\hat{p})$, that is

$$\hat{\delta}_{ij} = \frac{\hat{v}_{ij}(h)}{\hat{p}_{i+}(1-\hat{p}_{i+})\hat{p}_{+j}(1-\hat{p}_{+j})}.$$

$X^2_{I(c)}$ requires the knowledge of the design effects of $h_{ij}(\hat{p})$'s. However, such information is seldom available in published reports. As an alternative, Rao and Scott (1981) defined $\hat{\tau}_{\cdot} = \sum_i^r \sum_j^c (1 - \hat{p}_{ij})\hat{d}_{ij}/(rc - 1)$ in place of $\hat{\delta}_{\cdot}$. Empirical results for X^2_I for the same large scale survey study, analyzed by Rao and Scott, show that $\hat{\delta}_{\cdot}$ is smaller than $\hat{\tau}_{\cdot}$ in all cases. The distortion of significance level tends to be less for the chi-squared test of independence than it was in goodness-of-fit case. The distortion can still be severe and some modification is necessary. The test based on $X^2/\hat{\delta}_{\cdot}$ works well in every case, but if $\hat{\delta}_{\cdot}$ is not available, using $X^2/\hat{\tau}_{\cdot}$ seems reasonable. Using $\hat{\tau}_{\cdot}$ gives conservative and sometimes very conservative results, so a considerable loss of power may be found. But with large-scale surveys (large samples) this effect will not be serious; see Rao and Scott (1981) and Holt, Scott and Ewings (1980).

2.10.2 Log odds ratio test for 2×2 table

For testing independence in a 2×2 table, the log odds ratio converges to a normal distribution more rapidly than the odds ratio (Agresti, 1996). Thus, we will consider the log odds ratio, ϕ, equal to $\ln(\frac{p_{11}p_{22}}{p_{12}p_{21}})$. When the two variables are independent the odds ratio is equal to one. Thus, in case of independence $\phi = \ln(1) = 0$. Consider testing hypotheses having the following form

$$H_0 : \phi = \phi_0 = 0$$
$$\text{VS}$$
$$H_1 : \phi \neq \phi_0. \tag{2.53}$$
$$\text{where} \quad \phi = \ln\left(\frac{p_{11}p_{22}}{p_{12}p_{21}}\right).$$

The test statistic for testing this hypothesis is

$$X^2 = \frac{\hat{\phi}^2}{\widehat{var}(\hat{\phi})} \tag{2.54}$$

since $E(\phi) = 0$, under the null hypothesis. Now, if our sampling design is a multinomial sample, then the asymptotic estimator of $var(\hat{\phi})$ is equal to, Altham

$X^2_{I(c)}$ requires the knowledge of the design effects of $h_{ij}(\hat{p})$'s. However, such information is seldom available in published reports. As an alternative, Rao and Scott (1981) defined $\hat{\tau}_{\cdot} = \sum_i^r \sum_j^c (1 - \hat{p}_{ij})\hat{d}_{ij}/(rc - 1)$ in place of $\hat{\delta}_{\cdot}$. Empirical results for X^2_I for the same large scale survey study, analyzed by Rao and Scott, show that $\hat{\delta}_{\cdot}$ is smaller than $\hat{\tau}_{\cdot}$ in all cases. The distortion of significance level tends to be less for the chi-squared test of independence than it was in goodness-of-fit case. The distortion can still be severe and some modification is necessary. The test based on $X^2/\hat{\delta}_{\cdot}$ works well in every case, but if $\hat{\delta}_{\cdot}$ is not available, using $X^2/\hat{\tau}_{\cdot}$ seems reasonable. Using $\hat{\tau}_{\cdot}$ gives conservative and sometimes very conservative results, so a considerable loss of power may be found. But with large-scale surveys (large samples) this effect will not be serious; see Rao and Scott (1981) and Holt, Scott and Ewings (1980).

2.10.2 Log odds ratio test for 2×2 table

For testing independence in a 2×2 table, the log odds ratio converges to a normal distribution more rapidly than the odds ratio (Agresti, 1996). Thus, we will consider the log odds ratio, ϕ, equal to $\ln(\frac{p_{11}p_{22}}{p_{12}p_{21}})$. When the two variables are independent the odds ratio is equal to one. Thus, in case of independence $\phi = \ln(1) = 0$. Consider testing hypotheses having the following form

$$H_0 : \phi = \phi_0 = 0$$

$$\text{VS}$$

$$H_1 : \phi \neq \phi_0. \tag{2.53}$$

$$\text{where} \quad \phi = \ln\left(\frac{p_{11}p_{22}}{p_{12}p_{21}}\right).$$

The test statistic for testing this hypothesis is

$$X^2 = \frac{\hat{\phi}^2}{\widehat{var}(\hat{\phi})} \tag{2.54}$$

since $E(\phi) = 0$, under the null hypothesis. Now, if our sampling design is a multinomial sample, then the asymptotic estimator of $var(\hat{\phi})$ is equal to, Altham

(1976),

$$\widehat{var}_{srs}(\hat{\phi}) = (\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}). \tag{2.55}$$

The statistic X^2, using the log odds ratio, is approximately distributed as a χ^2_1 random variable under H_0 and for sufficiently large sample size, n. If the sample is gathered by a more complex sampling design, then $var(\hat{\phi})$ can be estimated by using, for example, the Jackknife method if full sample information is available.

Chapter 3

Bayesian Statistics

In this chapter, the basic theory of Bayesian model selection is developed. The Bayes factor will be derived for two competing models, $M_k, k = 1, 2$. Unfortunately, sometimes it is difficult to evaluate the Bayes factor. Thus, the Savage-Dickey density ratio is discussed as a way of approximating the Bayes factor. Model averaged estimation is also presented as a Bayesian method for estimating a parameter of interest. Finally, the Bayesian information criterion approximation will be derived as a criterion for model selection.

3.1 Bayesian model selection methods

The standard Bayesian approach to the hypothesis testing and model selection problem has three main features. First, we represent both the null and alternative hypothesis as parametric probability models. Second, we calculate the key quantity, which is the marginal likelihood for each model, also known as the integrated likelihood, or the marginal probability of the data. Third, we use the marginal likelihood function to compute the Bayes factors for comparing two or more competing models and hence posterior probabilities of models, using the

prior probabilities. For a survey see Raftery (1995).

3.2 Bayes factor

Suppose that we are going to use the data, D, to compare two competing hypotheses, which represent the statistical models M_1 and M_2, with parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The posterior probability that $M_k, k = 1, 2$ is the correct model (using Bayes' theorem):

$$\begin{aligned} pr(M_k|D) &= \frac{pr(M_k, D)}{pr(D)} \\ &= \frac{pr(D|M_k)pr(M_k)}{pr(D)}. \end{aligned} \tag{3.1}$$

From that we get,

$$pr(M_k|D) \propto pr(D|M_k) \quad pr(M_k). \tag{3.2}$$

We can write this as

$$\text{Posterior} \propto \text{Marginal likelihood} \times \text{Prior}.$$

When we compare two models, the posterior odds for model M_2 against M_1 are,

$$\frac{pr(M_2|D)}{pr(M_1|D)} = \frac{pr(D|M_2)}{pr(D|M_1)} \times \frac{pr(M_2)}{pr(M_1)}. \tag{3.3}$$

On the right-hand side of equation (3.3), the first factor is the ratio of the marginal likelihoods of the two models and is known as the Bayes factor for model M_2 against M_1, denoted by B_{21}. The second factor is the prior odds. If the prior odds are equal to 1, which is often the case, as it represents the absence of a prior preference for either model, then the posterior odds are equal to the Bayes factor.

The Bayes factor B_{21} for comparing model M_2 against M_1 for observed data D is the ratio of the marginal likelihood functions under the two models being

compared,

$$B_{21} = \frac{pr(D|M_2)}{pr(D|M_1)}. \tag{3.4}$$

The Bayes factor was first introduced by Jeffreys (1935) and in Jeffreys (1961, Appendix B) he introduced an interpretation of B_{21} for testing hypotheses. The calculation of a Bayes factor depends only on the marginal likelihoods, which can be computed by the integration of the likelihood function multiplied by the prior over the parameter space,

$$pr(D|M_k) = \int_{\Theta_k} pr(D|\boldsymbol{\theta}_k, M_k) pr(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k, \;\; k = 1, 2, ..., T \tag{3.5}$$

where $\boldsymbol{\theta}_k$ is the parameter under model M_k, $pr(D|\boldsymbol{\theta}_k, M_k)$ the likelihood function of $\boldsymbol{\theta}_k$, and $pr(\boldsymbol{\theta}_k|M_k)$ is the prior density for $\boldsymbol{\theta}_k$ under model M_k. When $B_{21} > 1$, the data favour M_2 over M_1; and when $B_{21} < 1$ the data favour M_1. The marginal likelihoods yield posterior probabilities for all models, by using Bayes' theorem (Raftery, 1995):

$$pr(M_k|D) = \frac{pr(D|M_k)pr(M_k)}{\sum_k pr(D|M_k)pr(M_k)} \qquad k = 1, 2, ..., T. \tag{3.6}$$

If we consider model M_2 in equation (3.4) to be the null hypotheses against the alternative hypothesis, M_1, then we can see the close relation between the Bayes factor and the likelihood ratio statistic. The only difference in the two is the method of eliminating $\boldsymbol{\theta}_k$, maximization is used in the likelihood ratio, while integration is used in Bayes factor. In applications there are some differences. For example the null model must usually be nested within the alternative for a likelihood ratio test, while for a Bayes factor they do not have to be nested.

Careful consideration of the prior specification is needed when constructing the Bayes factor, as it will be always sensitive to the prior. Throughout this thesis we use non-informative proper priors. Raftery in the discussion of O'Hagan (1995) suggested that the approach of using proper priors that are fairly flat over the region, where the likelihood could be substantial, is more promising than the

compared,

$$B_{21} = \frac{pr(D|M_2)}{pr(D|M_1)}. \tag{3.4}$$

The Bayes factor was first introduced by Jeffreys (1935) and in Jeffreys (1961, Appendix B) he introduced an interpretation of B_{21} for testing hypotheses. The calculation of a Bayes factor depends only on the marginal likelihoods, which can be computed by the integration of the likelihood function multiplied by the prior over the parameter space,

$$pr(D|M_k) = \int_{\Theta_k} pr(D|\boldsymbol{\theta}_k, M_k) pr(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k, \;\; k = 1, 2, ..., T \tag{3.5}$$

where $\boldsymbol{\theta}_k$ is the parameter under model M_k, $pr(D|\boldsymbol{\theta}_k, M_k)$ the likelihood function of $\boldsymbol{\theta}_k$, and $pr(\boldsymbol{\theta}_k|M_k)$ is the prior density for $\boldsymbol{\theta}_k$ under model M_k. When $B_{21} > 1$, the data favour M_2 over M_1; and when $B_{21} < 1$ the data favour M_1. The marginal likelihoods yield posterior probabilities for all models, by using Bayes' theorem (Raftery, 1995):

$$pr(M_k|D) = \frac{pr(D|M_k)pr(M_k)}{\sum_k pr(D|M_k)pr(M_k)} \qquad k = 1, 2, ..., T. \tag{3.6}$$

If we consider model M_2 in equation (3.4) to be the null hypotheses against the alternative hypothesis, M_1, then we can see the close relation between the Bayes factor and the likelihood ratio statistic. The only difference in the two is the method of eliminating $\boldsymbol{\theta}_k$, maximization is used in the likelihood ratio, while integration is used in Bayes factor. In applications there are some differences. For example the null model must usually be nested within the alternative for a likelihood ratio test, while for a Bayes factor they do not have to be nested.

Careful consideration of the prior specification is needed when constructing the Bayes factor, as it will be always sensitive to the prior. Throughout this thesis we use non-informative proper priors. Raftery in the discussion of O'Hagan (1995) suggested that the approach of using proper priors that are fairly flat over the region, where the likelihood could be substantial, is more promising than the

approaches that 'trick' improper priors into giving reasonable Bayes factors, such as the fractional Bayes factor, FBF, (O'Hagan, 1995) and intrinsic Bayes factor, IBF, (Berger and Pericchi, 1996). Furthermore, our priors are compatible with 'unit prior information'(Kass and Wasserman, 1995), where the amount of information in the prior equals to the amount of information in one observation. However, sensitivity analysis remains important. It is more satisfactory if we evaluate the Bayes factor over a range of specified priors under the competing models.

Unfortunately, some times it is difficult to evaluate the Bayes factor, because of the integral in equation (3.5), in such case we need to resort to approximation methods, see Kass and Raftery (1995), Verdinelli and Wasserman (1995), and for a review DiCiccio *et al.* (1997).

3.3 Savage-Dickey density ratio

The Savage-Dickey density ratio is one of the approximation methods to estimates the Bayes factor when it is possible to simulates observation from the posterior distributions via Markov chain Monte Carlo (MCMC) algorithms or other techniques. It was developed by Dickey (1971) and generalized by Verdinelli and Wasserman (1995). Spiegelhalter and Smith (1982) illustrate its application for linear and log-linear models. The Savage-Dickey density ratio provides a particular convenient way of calculating the Bayes factor when the two models being tested are nested. It reduces computing the Bayes factor to the problem of estimating the marginal posterior density at a specified point.

Suppose that the two models being tested are nested, $M_2 \subset M_1$ and the sample distribution of data D under M_k $(k = 1, 2)$ depends only on $\boldsymbol{\theta}_k$, where

$\boldsymbol{\theta}_1 = [\boldsymbol{\theta}_2' : \boldsymbol{\theta}']$ such that

$$pr(D|M_2, \boldsymbol{\theta}_2) = pr(D|M_1, \boldsymbol{\theta}_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0) \tag{3.7}$$

where $\boldsymbol{\theta}_0$ is a specified value. Also, consider prior densities $pr(\boldsymbol{\theta}_k|M_k)$ for $k = 1, 2$, in such a way that

$$pr(\boldsymbol{\theta}_2|M_2) = pr(\boldsymbol{\theta}_1|M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0). \tag{3.8}$$

Then, the Bayes factor for model M_2 against M_1 is

$$\begin{aligned} B_{21} &= \frac{pr(D|M_2)}{pr(D|M_1)} \\ &= \frac{\int pr(D|M_2, \boldsymbol{\theta}_2) pr(\boldsymbol{\theta}_2|M_2) d\boldsymbol{\theta}_2}{\int pr(D|M_1, \boldsymbol{\theta}_1) pr(\boldsymbol{\theta}_1|M_1) d\boldsymbol{\theta}_1} \end{aligned}$$

from (3.7) and (3.8), we get

$$\begin{aligned} B_{21} &= \frac{\int pr(D|M_1, \boldsymbol{\theta}_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0) pr(\boldsymbol{\theta}_1|M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0) d\boldsymbol{\theta}_2}{\int pr(D|M_1, \boldsymbol{\theta}_1) pr(\boldsymbol{\theta}_1|M_1) d\boldsymbol{\theta}_1} \\ &= \frac{pr(D|M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0)}{pr(D|M_1)}. \end{aligned}$$

Using the Bayes theorem,

$$\begin{aligned} pr(D|M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0) &= \frac{pr(D, M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0)}{pr(M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0)} \\ &= \frac{pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|D, M_1) pr(D, M_1)}{pr(M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0)} \\ &= pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|D, M_1) \frac{pr(D|M_1) pr(M_1)}{pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|M_1) pr(M_1)} \\ &= pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|D, M_1) \frac{pr(D|M_1)}{pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|M_1)}. \end{aligned}$$

Therefore, the Bayes factor becomes

$$B_{21} = \frac{pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|D, M_1)}{pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|M_1)}. \tag{3.9}$$

Thus, applying the Savage-Dickey density ratio reduces computing the Bayes factor to the problem of estimating the marginal posterior density $pr(\boldsymbol{\theta}|D, M_1)$, at point $\boldsymbol{\theta}_0$.

$\boldsymbol{\theta}_1 = [\boldsymbol{\theta}_2' : \boldsymbol{\theta}']$ such that

$$pr(D|M_2, \boldsymbol{\theta}_2) = pr(D|M_1, \boldsymbol{\theta}_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0) \tag{3.7}$$

where $\boldsymbol{\theta}_0$ is a specified value. Also, consider prior densities $pr(\boldsymbol{\theta}_k|M_k)$ for $k = 1, 2$, in such a way that

$$pr(\boldsymbol{\theta}_2|M_2) = pr(\boldsymbol{\theta}_1|M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0). \tag{3.8}$$

Then, the Bayes factor for model M_2 against M_1 is

$$\begin{aligned} B_{21} &= \frac{pr(D|M_2)}{pr(D|M_1)} \\ &= \frac{\int pr(D|M_2, \boldsymbol{\theta}_2) pr(\boldsymbol{\theta}_2|M_2) d\boldsymbol{\theta}_2}{\int pr(D|M_1, \boldsymbol{\theta}_1) pr(\boldsymbol{\theta}_1|M_1) d\boldsymbol{\theta}_1} \end{aligned}$$

from (3.7) and (3.8), we get

$$\begin{aligned} B_{21} &= \frac{\int pr(D|M_1, \boldsymbol{\theta}_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0) pr(\boldsymbol{\theta}_1|M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0) d\boldsymbol{\theta}_2}{\int pr(D|M_1, \boldsymbol{\theta}_1) pr(\boldsymbol{\theta}_1|M_1) d\boldsymbol{\theta}_1} \\ &= \frac{pr(D|M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0)}{pr(D|M_1)}. \end{aligned}$$

Using the Bayes theorem,

$$\begin{aligned} pr(D|M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0) &= \frac{pr(D, M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0)}{pr(M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0)} \\ &= \frac{pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|D, M_1) pr(D, M_1)}{pr(M_1, \boldsymbol{\theta} = \boldsymbol{\theta}_0)} \\ &= pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|D, M_1) \frac{pr(D|M_1) pr(M_1)}{pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|M_1) pr(M_1)} \\ &= pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|D, M_1) \frac{pr(D|M_1)}{pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|M_1)}. \end{aligned}$$

Therefore, the Bayes factor becomes

$$B_{21} = \frac{pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|D, M_1)}{pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|M_1)}. \tag{3.9}$$

Thus, applying the Savage-Dickey density ratio reduces computing the Bayes factor to the problem of estimating the marginal posterior density $pr(\boldsymbol{\theta}|D, M_1)$, at point $\boldsymbol{\theta}_0$.

3.4 Model Averaged Estimation

Model averaging refers to the process of estimating some quantity under each model and then averaging the estimates according to how likely each model is. We are going to use a model averaging approach to estimate the values of the parameters $\hat{\theta}_{ij}$. This requires the posterior distribution $pr(M_k|D), k = 1, 2, ..., T$; and the expected value of θ_{ij} under the marginal posterior distribution of $\boldsymbol{\theta}$ given model M_k.

Suppose that $\boldsymbol{\theta}$ is a parameter of main interest, which is well defined for each model. Then for any model, M_k, Bayesian inference about $\boldsymbol{\theta}$ is based on its conditional posterior distribution given M_k, $pr(\boldsymbol{\theta}|D,M_k)$. For overall inference, we may use the marginal posterior density of $\boldsymbol{\theta}$ given by

$$pr(\boldsymbol{\theta}|D) = \sum_{k=1}^{T} pr(\boldsymbol{\theta}|D,M_k)pr(M_k|D). \tag{3.10}$$

Thus the full posterior distribution of $\boldsymbol{\theta}$ is a weighted average of its marginal posterior distributions under each of the models, where the weights are the posterior model probabilities, $pr(M_k|D)$. Here, $pr(\boldsymbol{\theta}|D)$ takes full account of the model uncertainty. Our interest is in estimating the parameter $\boldsymbol{\theta}$, using the posterior mean, where the posterior mean is a Bayes estimator for $\boldsymbol{\theta}$ under squared-error loss (Raftery, 1995, Kass and Raftery, 1995).

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= E(\boldsymbol{\theta}|D) \\ &= \sum_{k=1}^{T} E(\boldsymbol{\theta}|D,M_k)pr(M_k|D). \end{aligned} \tag{3.11}$$

Statistical inference based on selection of a single model M_k will not effectively take account of uncertainty about the parameter of interest $\boldsymbol{\theta}$, and will underestimate the uncertainty associated with an estimator of $\boldsymbol{\theta}$; For a review see Hoeting *et al.* (1999).

3.5 Bayesian Information Criterion approximation

The Bayesian Information Criterion, BIC, approximation was first introduced by Schwarz (1978). Raftery (1995) presents BIC as a measure for model selection that sidesteps many of the problems with P-value and classical hypotheses testing. For example, with a large enough sample, researchers almost always favour the alternate hypothesis over the null. Raftery (1995) argues that BIC overcomes these difficulties, since it estimates the probability that any given model is the correct model given the data. Also, researchers can assign a BIC to the null model, or a saturated model, or to any other model and each BIC can then be compared; more likely models, those with lower BICs, are preferred. Since the null model is always considered just as any other model, the BIC does not exhibit the traditional preference for the null model with smaller samples or against the null for large sample sizes, for example see Raftery (1986). In addition, BIC gives a simple approximation to the Bayes factor, which is easy to use for assessing competing models and does not require evaluation of prior distributions. Thus, BIC may be utilised relatively easily to adjudicate between models.

To derive the BIC approximation we apply Laplace's method for integrals to the marginal likelihood function for equation (3.5), e.g. see De Bruijn (1961) Section 4.4. The Taylor series expansion of the function $g(\boldsymbol{\theta}) = \ln\{pr(D|\boldsymbol{\theta},M)pr(\boldsymbol{\theta}|M)\}$ about $\tilde{\boldsymbol{\theta}}$, the value of $\boldsymbol{\theta}$ that maximizes $g(\boldsymbol{\theta})$, i.e. the posterior mode, is

$$g(\boldsymbol{\theta}) = g(\tilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g'(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g''(\tilde{\boldsymbol{\theta}})\,(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + O(\| \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}} \|^2) \quad (3.12)$$

here, the superscript ^{T} denotes matrix transpose, $g'(\boldsymbol{\theta}) = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1} \quad . \quad . \quad . \quad \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_d} \right)^T$ is the vector of first partial derivatives of $g(\boldsymbol{\theta})$, and $g''(\boldsymbol{\theta}) = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right)$ is the matrix of second partial derivatives of $g(\boldsymbol{\theta})$, called the Hessian matrix. Now $g(\boldsymbol{\theta})$ is maximized at $\tilde{\boldsymbol{\theta}}$, this implies that the first derivative of $g(\boldsymbol{\theta})$ at point $\tilde{\boldsymbol{\theta}}$ is equal

to zero, i.e. $g'(\tilde{\boldsymbol{\theta}}) = 0$. Thus

$$g(\boldsymbol{\theta}) \approx g(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g''(\tilde{\boldsymbol{\theta}})\,(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}). \tag{3.13}$$

This approximation will be good only if $\boldsymbol{\theta}$ is close to $\tilde{\boldsymbol{\theta}}$. But for equation (3.5), when n is large the likelihood function $pr(D|\boldsymbol{\theta}_k, M_k)$ is concentrated around its maximum $\tilde{\boldsymbol{\theta}}$. This means the only values which contribute much to the integral are those which are close to the maximum $\tilde{\boldsymbol{\theta}}$, where the effect of the prior is minor for large n; see Tierney and Kadane (1986).

Using equation (3.13) we get,

$$\begin{aligned} pr(D|M) &= \int_{\theta\epsilon\Theta} \exp\{g(\boldsymbol{\theta})\}d\boldsymbol{\theta} \\ &\approx \int_{\theta\epsilon\Theta} \exp\{g(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g''(\tilde{\boldsymbol{\theta}})\,(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\}d\boldsymbol{\theta} \\ &= \exp[g(\tilde{\boldsymbol{\theta}})] \int_{\theta\epsilon\Theta} \exp\{\frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T g''(\tilde{\boldsymbol{\theta}})\,(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\}d\boldsymbol{\theta}. \end{aligned} \tag{3.14}$$

The integrand is proportional to a multivariate normal density, which gives us

$$pr(D|M) \approx (2\pi)^{\frac{d}{2}} \exp[g(\tilde{\boldsymbol{\theta}})] \left|-g''(\tilde{\boldsymbol{\theta}})\right|^{-\frac{1}{2}} \tag{3.15}$$

where d represent the number of parameters in the model M. Tierney and Kadane (1986) show that in regular statistical models the relative error is of order $O(n^{-1})$ in equation (3.15), where n is the sample size. Here $O(n^{-1})$ represents any quantity such that $|nO(n^{-1})| <$ constant as $n \longrightarrow \infty$. So, by taking the logarithm of $pr(D|M)$, we get

$$\begin{aligned} \ln[pr(D|M)] \;=\; & \ln[pr(D|\tilde{\boldsymbol{\theta}}, M)] + \ln[pr(\tilde{\boldsymbol{\theta}}|M)] + \frac{d}{2}\ln(2\pi) \\ & -\frac{1}{2}\ln\left|-g''(\tilde{\boldsymbol{\theta}})\right| + O(n^{-1}). \end{aligned} \tag{3.16}$$

In large samples, $\tilde{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}$ where $\hat{\boldsymbol{\theta}}$ is the MLE, and $-g''(\tilde{\boldsymbol{\theta}}) = n\mathbf{i}$, where $\mathbf{i}$ is the expected Fisher information matrix for one observation. Thus, $\left|-g''(\tilde{\boldsymbol{\theta}})\right| \approx n^d\,|\mathbf{i}|$. These two approximations, under the large samples, introduce an $O(n^{-\frac{1}{2}})$ error

into equation (3.16),

$$\begin{aligned}\ln[pr(D|M)] &= \ln[pr(D|\hat{\boldsymbol{\theta}}, M)] + \ln[pr(\hat{\boldsymbol{\theta}}|M)] + \frac{d}{2}\ln(2\pi) \\ &\quad -\frac{d}{2}\ln(n) - \frac{1}{2}\ln|\mathbf{i}| + O(n^{-\frac{1}{2}}).\end{aligned} \tag{3.17}$$

$\ln[pr(D|\hat{\boldsymbol{\theta}}, M)]$ is of order $O(n)$ and $\frac{d}{2}\ln(n)$ of order $O(\ln n)$, while the other four terms are of order $O(1)$ or less, thus

$$\ln[pr(D|M)] = \ln[pr(D|\hat{\boldsymbol{\theta}}, M)] - \frac{d}{2}\ln(n) + O(1). \tag{3.18}$$

So, the log-marginal likelihood, $\ln[pr(D|M)]$, is equal to the maximized log-likelihood, $\ln[pr(D|\hat{\boldsymbol{\theta}}, M)]$, minus a correction term; see Raftery (1995).

The BIC approximation is based on equation (3.18), and the $O(1)$ error indicates that the approximation is somewhat crude, because the error does not vanish even with an infinite amount of data. Nevertheless, Kass and Wasserman (1995) show that under a particular choice of prior distribution, consistent with 'unit prior information', the error is of order $O(n^{-\frac{1}{2}})$ rather than $O(1)$. Their definition of 'unit prior information' is equivalent to a normal prior with variance-covariance matrix given by the inverse Fisher Information matrix for a single observation. Empirical results, Raftery (1996), show that equation (3.18) is more accurate in practice than the $O(1)$ error term would suggest, under a reasonable choice of prior. Because the right-hand side of equation (3.18) tends to infinity as n goes to infinity, and so will eventually dominate, this ensures that the relative error in $\ln[pr(D|M)]$ will tend toward zero, and will have no effect on the conclusion reached as n gets large.

Now to approximate the Bayes factor in equation (3.4), we get

$$\begin{aligned}\ln B_{21} &= \ln[pr(D|M_2)] - \ln[pr(D|M_1)] \\ &\approx \left(\ln[pr(D|\hat{\boldsymbol{\theta}}, M_2)] - \frac{d_2}{2}\ln(n)\right) - \left(\ln[pr(D|\hat{\boldsymbol{\theta}}, M_1)] - \frac{d_1}{2}\ln(n)\right).\end{aligned}$$

Thus,

$$2\ln B_{21} \approx 2\left(\ln[pr(D|\hat{\boldsymbol{\theta}}, M_2)] - \ln[pr(D|\hat{\boldsymbol{\theta}}, M_1)]\right) - (d_2 - d_1)\ln(n). \tag{3.19}$$

If M_1 is nested within M_2 , then the first two terms in the right-hand side of equation (3.19) is the standard likelihood ratio test (LRT) statistic for testing M_1 against M_2 , and equation (3.19) can be rewritten

$$2\ln B_{21} \approx L_{21} - d_{21}\ln(n) \tag{3.20}$$

where L_{21} is the LRT statistic for testing M_1 against M_2 , and $d_{21} = d_2 - d_1$ is the number of degrees of freedom associated with the test. Now BIC is the approximation of $2\ln B_{21}$, which is given by equation (3.20). BIC can be used when several model are of interest, via a comparison of each of them in turn with a baseline model, usually a saturated model, M_S, in which each data point is fitted exactly. In general, suppose that the baseline model is saturated, M_S, then the LRT statistic in equation (3.20) is often called the deviance. In this case, BIC for model M_k denoted by BIC_k is the approximation to $2\ln B_{Sk}$ given by equation (3.20), where B_{Sk} is the Bayes factor for model M_S against model M_k, i.e.

$$BIC_k \approx L_k - d_k\ln(n) \tag{3.21}$$

where $L_k = L_{Sk}$, is the deviance for model M_k, and d_k is the number of degrees of freedom of the test. For interpreting the BIC evidence in favour of a model, M_S against M_k, the following table includes a rounded scale for interpreting B_{Sk}, based on Jeffreys (1961, Appendix B), and $2\ln B_{Sk}$ used by Raftery (1996).

B_{Sk}	$2\ln B_{Sk}$	Evidence for M_S
< 1	< 0	Negative (Support M_k)
1 to 3	0 to 2.2	Not worth more than a bare mention
3 to 20	2.2 to 6	Positive
20 to 150	6 to 10	Strong
> 150	> 10	Very strong

Thus, when $\text{BIC}_k > 0$, the saturated model is preferred to M_k, i.e. M_k does not fit the data well. But if $\text{BIC}_k < 0$, M_k is preferred to the saturated model, M_S. The smaller BIC_k, the more negative, the better the fit of M_k. The BIC can be considered as a Bayesian way of evaluating evidence in favour of an alternative, model M_S. It takes sample size, n, directly into account, and imposes a penalty for increasing the number of parameters in the model. For further details see Kass and Wasserman, (1992), Kass and Raftery, (1995), and Raftery (1996).

Chapter 4

Bayesian inference for sample surveys

The Bayesian approach using model selection for categorical survey data will be presented in this chapter for three sampling schemes, simple random samples, a finite population, and stratified samples. The sampling design for these samples will be discussed, and the Bayes factor will be evaluated for a competing models.

In the simple random sample or multinomial sample, we show how the Bayesian approach can compete with the classical approach. This will be demonstrated by comparing the results of estimating the population proportion, $\mathbf{p}$, using pretest estimator, for the classical approach, and using a model averaged estimator, for the Bayesian approach, in a simulation study.

In addition, for the finite population and stratification sampling schemes, we will discuss the design effects on the inference based on $\mathbf{p}$. In this study we will show when the design effects can be ignored.

4.1 Simple random sample

Consider a sample of n independent observations concentrated on K $(r \times c)$ different categories. Let $n_{11}, \ldots, n_{rc}$ be the sampled observations which fall in these categories. If we consider the $p_{ij}(i = 1, 2, ..., r$ and $j = 1, 2, ...c)$ to be the probability that the $(ij)^{th}$ category is selected n_{ij} times, where we assume that each unit has the same probability of selection, then $\mathbf{n} = (n_{11}, \ldots, n_{rc})'$ will have a multinomial distribution with parameters $n = \sum_{ij}^{r,c} n_{ij}$ and $\mathbf{p} = (p_{11}, \ldots, p_{rc})'$. This sampling design was explored and discussed in sections (2.1), (2.2), and (2.3).

For this design, we postulate three competing models. Model M_S will propose a saturated model, whereas model M_I will be the independence model and M_G will be a model which specifies all cell probabilities for a 2 way table.

4.1.1 Bayes factor

Suppose that we want to use the data $\mathbf{n} = (n_{11}, \ldots, n_{rc})'$ to compare three models, which are represented by model M_S, M_I and M_G.

- Model M_S (Saturated model),

Let $\mathbf{n}$ have a multinomial distribution with parameters n, and $\mathbf{p}$, see equation (2.1). The natural conjugate prior for the multinomial distribution is a Dirichlet distribution with density function,

$$pr(\mathbf{p}) = \frac{\Gamma\left[\sum_{ij}^{r,c} \alpha_{ij}\right]}{\prod_{ij} \Gamma(\alpha_{ij})} \prod_{ij}^{r,c} p_{ij}^{\alpha_{ij}-1}. \tag{4.1}$$

To calculate the Bayes factor, we have to compute the marginal likelihood, see equation (3.5),

$$\begin{aligned} pr(\mathbf{n}|M_S) &= \int pr(\mathbf{n}|\mathbf{p}, M_S) pr(\mathbf{p}|M_S) d\mathbf{p} \\ &= \int \frac{n!}{\prod_{ij} n_{ij}!} \prod_{ij}^{r,c} p_{ij}^{n_{ij}} \frac{\Gamma(\sum_{ij} \alpha_{ij})}{\prod_{ij} \Gamma(\alpha_{ij})} \prod_{ij}^{r,c} p_{ij}^{\alpha_{ij}-1} dp_{ij} \\ &= \frac{n!}{\prod_{ij} n_{ij}!} \frac{\Gamma(\sum_{ij} \alpha_{ij})}{\prod_{ij} \Gamma(\alpha_{ij})} \int \prod_{ij} p_{ij}^{n_{ij}+\alpha_{ij}-1} dp_{ij}. \end{aligned} \tag{4.2}$$

If we consider $n = \sum_{ij}^{r,c} n_{ij}$ and $\alpha = \sum_{ij}^{r,c} \alpha_{ij}$, then

$$pr(\mathbf{n}|M_S) = \left(\frac{n!}{\prod_{ij} n_{ij}!} \right) \left(\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \right) \prod_{ij}^{r,c} \left(\frac{\Gamma(n_{ij}+\alpha_{ij})}{\Gamma(\alpha_{ij})} \right). \tag{4.3}$$

- Model M_I (independence model),

Let $\mathbf{n}$ have a multinomial distribution with parameters n and $\mathbf{p}_{i+}$ $\mathbf{p}_{+j}$; $i = 1, 2, ..., r$ and $j = 1, 2, ...c$. Also using an independent Dirichlet distribution as a prior for both parameters $\mathbf{p}_{i+}$ and $\mathbf{p}_{+j}$, we can compute the marginal (or integrated) likelihood function by,

$$pr(\mathbf{n}|M_I) = \int \int pr(\mathbf{n}|\mathbf{p}, M_I)\ pr(\mathbf{p}_{i+}|M_I)\ pr(\mathbf{p}_{+j}|M_I) d\mathbf{p}_{i+}\ d\mathbf{p}_{+j}.$$

Therefore,

$$\begin{aligned} pr(\mathbf{n}|M_I) = \int \int & \frac{n!}{\prod_{ij} n_{ij}!} \prod_{ij}^{r,c} (p_{i+} p_{+j})^{n_{ij}} \frac{\Gamma(\sum_i \alpha_{i+})}{\prod_i \Gamma(\alpha_{i+})} \prod_i^r p_{i+}^{\alpha_{i+}-1} \\ & \frac{\Gamma(\sum_j \alpha_{+j})}{\prod_j \Gamma(\alpha_{+j})} \prod_j^c p_{+j}^{\alpha_{+j}-1} dp_{i+} dp_{+j} \\ = & \frac{n!}{\prod_{ij} n_{ij}!} \left[\frac{\Gamma(\alpha)}{\prod_i \Gamma(\alpha_{i+})} \right] \left[\frac{\Gamma(\alpha)}{\prod_j \Gamma(\alpha_{+j})} \right] \\ & \int \prod_i p_{i+}^{n_{i+}+\alpha_{i+}-1} dp_{i+} \int \prod_j p_{+j}^{n_{+j}+\alpha_{+j}-1} dp_{+j}. \end{aligned}$$

Thus,

$$pr(n|M_I) = \frac{n!}{\prod_{ij} n_{ij}!}\left[\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\right]^2 \frac{\prod_i \Gamma(n_{i+}+\alpha_{i+})}{\prod_i \Gamma(\alpha_{i+})} \frac{\prod_j \Gamma(n_{+j}+\alpha_{+j})}{\prod_j \Gamma(\alpha_{+j})}. \tag{4.4}$$

Now the Bayes factor, B_{IS}, for comparing M_I against M_S for observed data $\mathbf{n}$ is,

$$B_{IS}(\mathbf{n}) = \frac{pr(\mathbf{n}|M_I)}{pr(\mathbf{n}|M_S)}$$

$$B_{IS}(\mathbf{n}) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_i^r \left(\frac{\Gamma(n_{i+}+\alpha_{i+})}{\Gamma(\alpha_{i+})}\right) \prod_j^c \left(\frac{\Gamma(n_{+j}+\alpha_{+j})}{\Gamma(\alpha_{+j})}\right) \prod_{ij}^{r,c} \left(\frac{\Gamma(\alpha_{ij})}{\Gamma(n_{ij}+\alpha_{ij})}\right) \tag{4.5}$$

To use equation (3.11) to compute the model averaged estimate, MAE, for the competing models, we have to know the posterior distributions, equation (3.10), for the parameters of model M_S and M_I and the prior model probabilities. Here, we assume that $pr(M_S) = pr(M_I) = \frac{1}{2}$, which represents the absence of a prior preference for either model.

For model M_S,

$$\mathbf{p}|\mathbf{n} \sim Dir(\mathbf{n}+\boldsymbol{\alpha}).$$

For model M_I independently,

$$\mathbf{p}_{i+}|\mathbf{n} \sim Dir(\mathbf{n}_{i+}+\boldsymbol{\alpha}_{i+})$$

$$\mathbf{p}_{+j}|\mathbf{n} \sim Dir(\mathbf{n}_{+j}+\boldsymbol{\alpha}_{+j}).$$

So, to compute the estimate of p_{ij}'s using the MAE,

$$\begin{aligned}\hat{p}_{ij} &= pr(M_S|\mathbf{n})E(p_{ij}|\mathbf{n}, M_S) + pr(M_I|\mathbf{n})E(p_{i+}p_{+j}|\mathbf{n}, M_I) \\ &= \frac{1}{1+B_{IS}(\mathbf{n})}\left(\frac{n_{ij}+\alpha_{ij}}{n+\alpha}\right) + \frac{B_{IS}(\mathbf{n})}{B_{IS}(\mathbf{n})+1}\left(\frac{n_{i+}+\alpha_{i+}}{n+\alpha}\right)\left(\frac{n_{+j}+\alpha_{+j}}{n+\alpha}\right).\end{aligned} \tag{4.6}$$

- Model M_G (completely specified model),

Under this model , M_G, the parameter $\mathbf{p} = \mathbf{p}_0 = (p_{0_{11}}, \ldots, p_{0_{rc}})'$ is known. Hence, the marginal likelihood $pr(\mathbf{n}|\mathbf{p}_0, M_G)$ is the usual multinomial likelihood,

$$pr(\mathbf{n}|\mathbf{p}_0, M_G) = \frac{n!}{\prod_{ij} n_{ij}!} \prod_{ij}^{r,c} p_{0_{ij}}^{n_{ij}}. \tag{4.7}$$

Then, the Bayes factor, the ratio of the marginal likelihoods, is,

$$\begin{aligned} B_{GS}(\mathbf{n}) &= \frac{pr(\mathbf{n}|M_G)}{pr(\mathbf{n}|M_S)} \\ &= \frac{\frac{n!}{\prod_{ij} n_{ij}!} \prod_{ij} p_{0_{ij}}^{n_{ij}}}{\frac{n!}{\prod_{ij} n_{ij}!} \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \frac{\prod_{ij} \Gamma(n_{ij}+\alpha_{ij})}{\prod_{ij} \Gamma(\alpha_{ij})}} \\ &= \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)} \prod_{ij}^{r,c} \left(\frac{p_{0_{ij}}^{n_{ij}} \Gamma(\alpha_{ij})}{\Gamma(n_{ij}+\alpha_{ij})} \right). \end{aligned} \tag{4.8}$$

Hence, the Bayes factor, B_{GS}, for comparing M_G against M_S for observed data $\mathbf{n}$ is,

$$B_{GS}(\mathbf{n}) = \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)} \prod_{ij}^{r,c} \left(\frac{p_{0_{ij}}^{n_{ij}} \Gamma(\alpha_{ij})}{\Gamma(n_{ij}+\alpha_{ij})} \right). \tag{4.9}$$

4.1.2 Pretest estimate

In two-way contingency tables with multinomial sampling, a test of hypotheses is often performed to test independence between the two variables. For example, if the null hypothesis is of stochastic independence, $H_0 : p_{ij} = p_{i+}p_{+j}$ for all i and j, then to test H_0 we can use the statistic,

$$X_I^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \tag{4.10}$$

where $\hat{m}_{ij} = n\hat{p}_{i+}\hat{p}_{+j}$. This test statistic has an asymptotic chi-squared distribution with d.f. $= (r-1)(c-1)$. We can use this to evaluate a pretest estimate.

The main idea of a pretest estimate is performing the estimation after performing a test of hypotheses. In this research the estimators depend upon the result of a Pearson chi-squared test for independence. If we do not reject the null hypothesis $H_0 : p_{ij} = p_{i+}p_{+j}$, then we consider $\hat{p}_{ij} = \frac{n_{i+}n_{+j}}{n}$, otherwise, $\hat{p}_{ij} = \frac{n_{ij}}{n}$. Usually, we consider $\chi^2_1 = 3.84$ as the critical value for a chi-squared test in a 2 by 2 contingency table.

4.1.3 Kullback-Liebler distance

Finally, we can compare the two estimates, MAE and the pretest estimate. For this we are going to use the Kullback-Liebler distance, D, between the true values and the estimates (Bishop, Fienberg and Holland, 1975),

$$D = \sum_{ij}^{r,c} p_{ij} \ln \frac{p_{ij}}{\hat{p}_{ij}}. \tag{4.11}$$

So,

$$D(B) = \sum_{ij}^{r,c} p_{ij} \ln \frac{p_{ij}}{\hat{p}_{ij}(B)} \quad \text{and} \quad D(X^2_I) = \sum_{ij}^{r,c} p_{ij} \ln \frac{p_{ij}}{\hat{p}_{ij}(X^2_I)}$$

$$Diff = D(B) - D(X^2_I) = \sum_{ij}^{r,c} p_{ij} \ln \frac{\hat{p}_{ij}(X^2_I)}{\hat{p}_{ij}(B)}. \tag{4.12}$$

If the difference $(Diff)$ is less than zero we favour the model averaged estimate, if it is equal to zero then the two methods are equal, otherwise the pretest estimate performs better. The result indicates which of the estimation methods is performing better. In general, p_{ij} are unknown, but for our simulation study they are known, so $Diff$ can be calculated.

4.1.4 Simulation study

We are going to consider the straightforward case of testing independence in a two-way table. There is no violation of the multinomial sampling assumption,

since we are going to generate samples from a multinomial distribution.

The program algorithms:

We wrote a Pascal program in order to do this simulation. This program does the following;

- Generate N samples from a Multinomial distribution.
- Compute matrix inverse (Press *et al.*, 1988).
- Evaluate the Pearson chi-squared statistic, equation (4.10).
- Evaluate the Pretest estimate of $\mathbf{p}$.
- Evaluate the Bayes factor, equation (4.5).
- Evaluate the model averaged estimate, MAE, of $\mathbf{p}$, equation (4.6).

Data

We generate 1000 realisations from three cases of the Multinomial distribution, with three different sample sizes, $n = 40, 200$, and 1000. First, the independent case, where $\mathbf{p}' = (0.63, 0.07, 0.27, 0.03)$, and the odds ratio $\frac{p_{11}p_{22}}{p_{12}p_{21}}$ is equal to 1. Second, not far from independence, where $\mathbf{p}' = (0.64, 0.06, 0.26, 0.04)$, and the odds ratio is equal to 1.64. The third case is a highly dependent case, where $\mathbf{p}' = (0.68, 0.02, 0.22, 0.08)$, and the odds ratio is equal to 12.

4.1.5 Result

For the independent case, the following table presents a measure of central tendency for the difference, $Diff$, in the three sample sizes, where $\mathbf{p}'$=$(0.63, 0.07, 0.27, 0.03)$,

Variable	n	Mean	Median	StDev	Percentage of values < 0
$Diff$	40	-0.00410	-0.00026	0.01293	51
$Diff$	200	-0.00062	0.00000	0.00272	50.2
$Diff$	1000	-0.00009	0.00000	0.00045	49

Thus, for the independent case the performance of the pretest estimate, PTE, and the model averaged estimate, MAE, are very similar, figure(4.1). The extreme observations in figure(4.1-c) correspond erroneous rejection of H_0, in PTE.

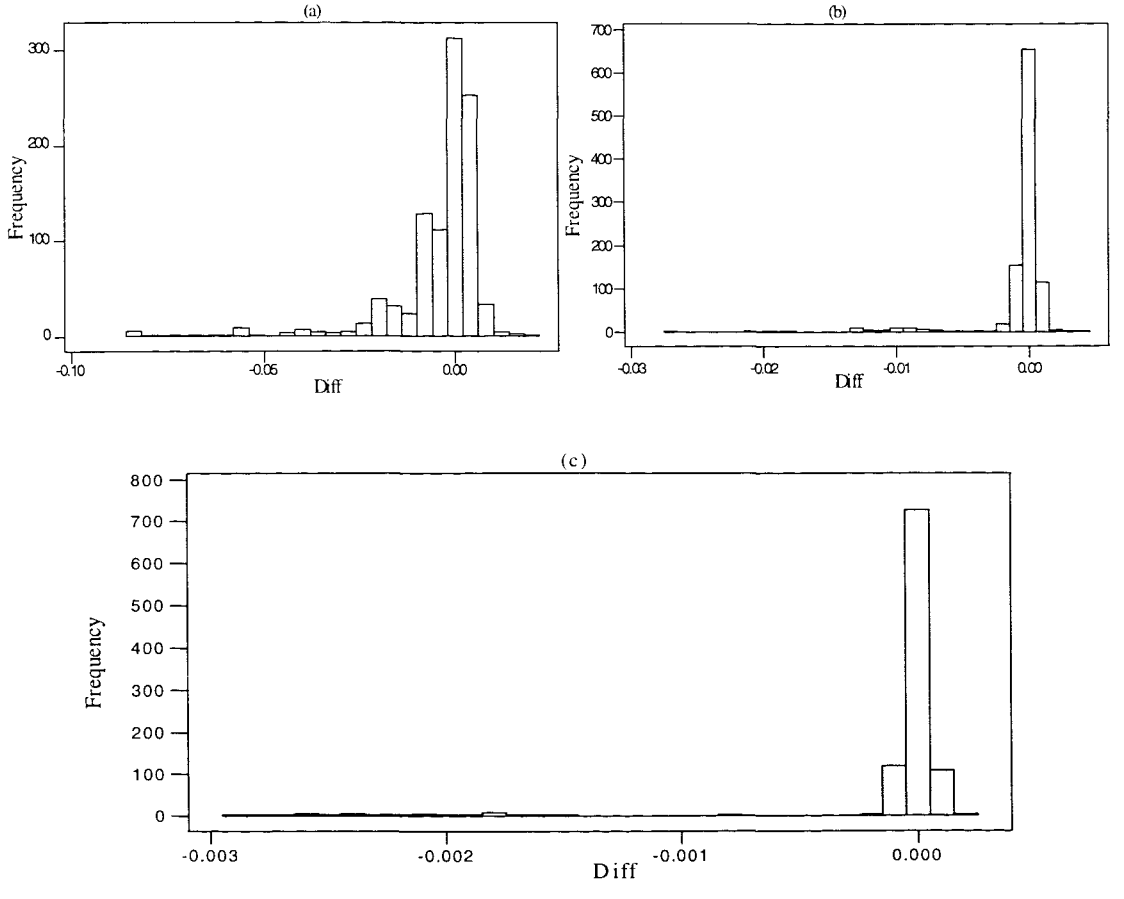

Figure 4.1: Histogram for the difference, $Diff$, between the pretest estimate, PTE, and the model averaged estimate, MAE, for the independent case.(a) For n=40, (b) For n=200, and (c) For n=1000.

If we are not far from independence, then the table of the measure of central tendency for distribution of the difference, $Diff$, in the three sample sizes, where $\mathbf{p}'$=(0.64, 0.06, 0.26, 0.04), is

Variable	n	Mean	Median	StDev	Percentage of values < 0
$Diff$	40	-0.00673	-0.00195	0.01522	61
$Diff$	200	-0.00121	-0.00076	0.00203	80
$Diff$	1000	0.00016	-0.00010	0.00066	63

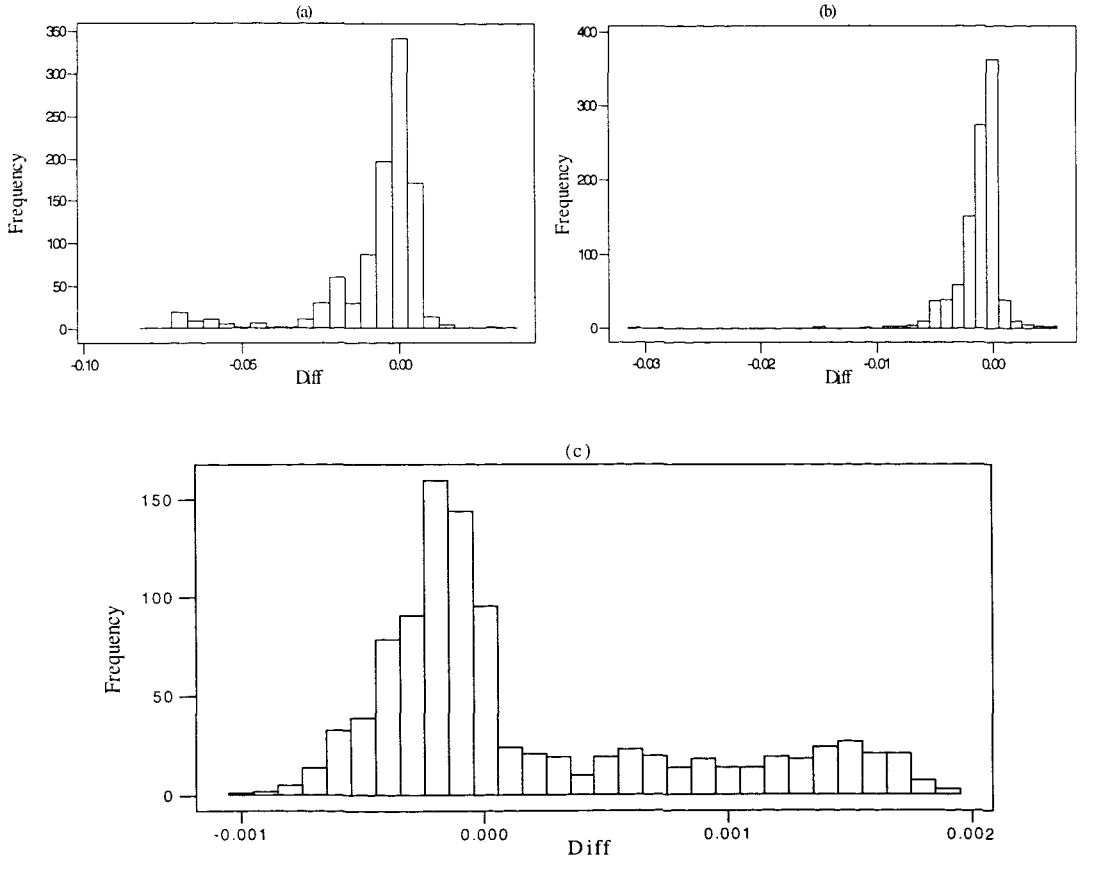

Figure 4.2: Histogram for the difference, $Diff$, between the pretest estimate, PTE, and the model averaged estimate, MAE, for the close to independent case.(a) For n=40, (b) For n=200, and (c) For n=1000.

In this case, there is more evidence that MAE performs better than PTE, see figure(4.2), especially, when the sample size is small. Figure(4.2-c) shows

two different curves, one like the normal curve for the value less than zero and values greater than zero have almost uniform shape.

In the case of high dependence, the measures of central tendency for the difference, $Diff$, in the three sample sizes, where $\mathbf{p}'=(0.68, 0.02, 0.22, 0.08)$, is

Variable	n	Mean	Median	StDev	Percentage of values < 0
$Diff$	40	-0.01392	-0.00900	0.02099	79
$Diff$	200	-0.00017	-0.00004	0.00281	52
$Diff$	1000	-0.00001	-0.00001	0.00006	56

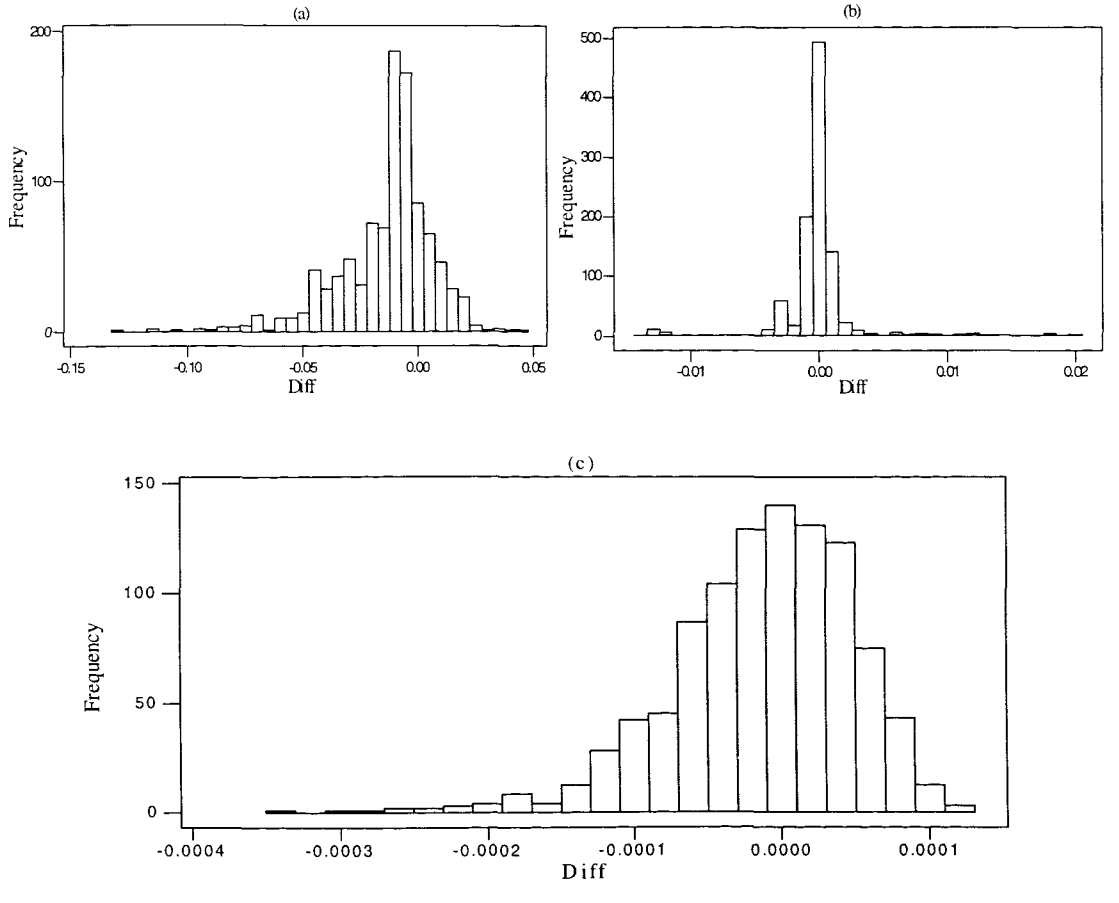

Figure 4.3: Histogram for the difference, $Diff$, between the pretest estimate, PTE, and the model averaged estimate, MAE, for the dependent case.(a) For n=40, (b) For n=200, and (c) For n=1000.

In the highly dependent case, clearly we again have evidence supporting MAE as a better estimate than PTE, as shown in the measure of central tendency and figure(4.3), especially, when the sample size is small.

4.1.6 Conclusion

The results of our simulation study, indicate that in cases of moderate and highly dependence, the MAE may perform better. This approach considers the uncertainty of the model by averaging the estimates according to how likely each model is, especially, when the sample size is small.

Now after this encouraging result, we can proceed to generalize this process. We are going to consider the design effect for more complex sample designs in two-way tables, where there is violation of the multinomial assumption. Those sampling designs are:

- Finite population.
- Stratification.
- Cluster sample.

4.2 Finite population case

Consider an observed sample of $\mathbf{n}$ counts, where $\mathbf{n} = (n_1, \ldots, n_K)'$, without replacement from a finite population of elements $\mathbf{N} = (N_1, N_2, ..., N_K)'$, where $0 \leq n_i \leq N_i$, $N = \sum_{i=1}^{K} N_i$, and $n = \sum_{i=1}^{K} n_i$. It is common, where appropriate, to assume that individuals who make up N are, a priori, exchangeable, see Ericson (1969, 1988), and O'Hagan (1994). For categorical data, Ericson (1969) considered an exchangeable prior by first giving each unit $\mathbf{Y}_i$ an independent Multinomial distribution with total one and probabilities $\mathbf{p}$,

$$\mathbf{Y}_i \overset{ind}{\sim} \text{Multin}(1, \mathbf{p}) = p_i \quad \forall i = 1, \ldots, K \tag{4.13}$$

where $\mathbf{p} = (p_1, \ldots, p_K)'$ and $\sum_{i=1}^{K} p_i = 1$. That is, given the vector $\mathbf{p}$, the finite population of size N, $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$, can be assumed to be the realisation of independent and identically distributed observations each with distribution (4.13). Thus, $\mathbf{N}$ has a Multinomial distribution, with total N and cell probabilities $\mathbf{p}$, i.e.

$$\mathbf{N} = \sum_{j=1}^{N} \mathbf{Y}_j \sim \text{Multin}(N, \mathbf{p}) \tag{4.14}$$

and the sampled units $\mathbf{n}$ have a Multinomial distribution, with total n and cell probabilities $\mathbf{p}$, i.e.

$$\mathbf{n} \sim \text{Multin}(n, \mathbf{p}). \tag{4.15}$$

The cell probabilities $\mathbf{p}$ are then given a second stage prior, or hyperprior. For any hyperprior on $\mathbf{p}$ this yields an exchangeable prior for the finite population units, $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$. Ericson (1969) points out that for discrete data this approach does not have the disadvantage that strong prior information regarding the shape of the finite population distribution is assumed, see also Binder (1982). As in the exact multinomial case, we know that Dirichlet distribution is a convenient choice of prior distribution for $\mathbf{p}$, because, it is conjugate to the multinomial distribution. Therefore, we will take as the prior on $\mathbf{p}$ the $(K-1)$-dimensional Dirichlet distribution (4.1), with parameters, $\boldsymbol{\alpha}$.

We now proceed to develop the posterior distribution of the $N-n$ unsampled units, with corresponding cell counts $\mathbf{N}-\mathbf{n}$. Given our prior specification and any sample, no matter how selected, consisting of n distinct population units, the posterior distribution of $\mathbf{N}-\mathbf{n}$ will be

$$pr(\mathbf{N}-\mathbf{n}|\mathbf{n}) = \int_{p\in\Omega} pr(\mathbf{N}-\mathbf{n},\mathbf{p}|\mathbf{n})\ d\mathbf{p}$$

but given $\mathbf{p}$, $\mathbf{N}-\mathbf{n}$ are independent of $\mathbf{n}$, thus

$$pr(\mathbf{N}-\mathbf{n}|\mathbf{n}) = \int_{p\in\Omega} pr(\mathbf{N}-\mathbf{n}|\mathbf{p})pr(\mathbf{p}|\mathbf{n})\ d\mathbf{p}.$$

Evaluating the marginal posterior density $pr(\mathbf{p}|\mathbf{n})$,

$$\begin{aligned} pr(\mathbf{p}|\mathbf{n}) &\propto pr(\mathbf{n}|\mathbf{p}) \times pr(\mathbf{p}|\boldsymbol{\alpha}) \\ &= \left(\frac{n!}{\prod_i^K n_i!}\prod_i^K p_i^{n_i}\right)\left(\frac{\Gamma\left(\sum_i^K \alpha_i\right)}{\prod_i^K \Gamma(\alpha_i)}\prod_i^K p_i^{\alpha_i-1}\right) \\ &\propto \prod_i p_i^{n_i} p_i^{\alpha_i-1} \\ &= \prod_{i=1}^K p_i^{n_i+\alpha_i-1}. \end{aligned}$$

This implies

$$pr(\mathbf{p}|\mathbf{n}) = \text{Dirichlet}(\mathbf{n}+\boldsymbol{\alpha}). \tag{4.16}$$

Now, $pr(\mathbf{N}-\mathbf{n}|\mathbf{p}) = \text{Multin}(N-n,\mathbf{p})$, and $pr(\mathbf{p}|\mathbf{n}) = \text{Dirichlet}(\mathbf{n}+\boldsymbol{\alpha})$ hence, we get

$$\begin{aligned} pr(\mathbf{N}-\mathbf{n}|\mathbf{n}) &= \int_{p\in\Omega} \frac{(N-n)!}{\prod_i (N_i-n_i)!}\prod_i p_i^{N_i-n_i}\frac{\Gamma(n+\alpha)}{\prod_i \Gamma(n_i+\alpha_i)}\prod_i p_i^{n_i+\alpha_i-1}\ d\mathbf{p} \\ &= \frac{(N-n)!}{\prod_i (N_i-n_i)!}\frac{\Gamma(n+\alpha)}{\prod_i \Gamma(n_i+\alpha_i)}\int_{p\in\Omega}\prod_i p_i^{N_i+\alpha_i-1}\ d\mathbf{p} \\ &= \frac{(N-n)!}{\prod_i (N_i-n_i)!}\frac{\Gamma(n+\alpha)}{\prod_i \Gamma(n_i+\alpha_i)}\frac{\prod_i \Gamma(N_i+\alpha_i)}{\Gamma(N+\alpha)} \\ &= \left(\frac{(N-n)!}{\prod_i^K (N_i-n_i)!}\right)\left(\frac{\Gamma(n+\alpha)}{\Gamma(N+\alpha)}\right)\prod_{i=1}^K\left(\frac{\Gamma(N_i+\alpha_i)}{\Gamma(n_i+\alpha_i)}\right). \end{aligned}$$

If we reparamterise the model by letting $m_i = N_i - n_i$ ($\Rightarrow m = N - n$), and $d_i = n_i + \alpha_i (\Rightarrow d = n + \alpha)$, then

$$pr(\mathbf{m}|\mathbf{n}) = \binom{m}{m_1, \ldots, m_K} \left(\frac{\Gamma(d)}{\Gamma(m+d)}\right) \prod_i^K \left(\frac{\Gamma(m_i + d_i)}{\Gamma(d_i)}\right). \tag{4.17}$$

This distribution is referred to as the Dirichlet-Multinomial distribution, and denoted by $\mathrm{DM}_K(m, \mathbf{d})$, see section (2.9). Mosimann (1962) shows that the mean of $\mathrm{DM}_K(m, \mathbf{d})$, or $\mathrm{DM}_K(N, \boldsymbol{\alpha})$, is

$$E(M_i) = m\frac{d_i}{d}.$$

Therefore

$$E(N_i - n_i|\mathbf{n}) = (N - n)\left(\frac{n_i + \alpha_i}{n + \alpha}\right). \tag{4.18}$$

Also, the covariance matrix is

$$\mathbf{V} = m\left(\frac{m+d}{1+d}\right)\left(diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'\right).$$

Therefore

$$Var(\mathbf{N} - \mathbf{n}|\mathbf{n}) = (N - n)\left(\frac{N + \alpha}{n + \alpha + 1}\right)\left(diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'\right). \tag{4.19}$$

Now, for model selection, we will derive the Bayes factor for three competing models the saturated, independence, and completely specified models. The marginal likelihood for the finite population under any model, M, is equal to,

$$pr(\mathbf{n}|M) = \int_{p\in\Omega} pr(\mathbf{n}|\mathbf{p}) \times pr(\mathbf{p}|M)\, d\mathbf{p} \tag{4.20}$$

where $pr(\mathbf{n}|\mathbf{p}) = \mathrm{Multin}(n, \mathbf{p})$, see (4.15), and $pr(\mathbf{p}|M) = \mathrm{Dirichlet}(\boldsymbol{\alpha})$. This is equal to the marginal likelihood for M under the exact multinomial sampling scheme, see equation (4.3).

- Model M_S , saturated model,

$$pr(\mathbf{n}|M_S) = \int pr(\mathbf{n}|\mathbf{p})pr(\mathbf{p}|M_S)d\mathbf{p}.$$

Therefore,

$$pr(\mathbf{n}|M_S) = \frac{n!}{\prod_{ij}^{r,c} n_{ij}!} \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{ij}^{r,c} \left(\frac{\Gamma(n_{ij}+\alpha_{ij})}{\Gamma(\alpha_{ij})} \right).$$

This result is equal to the marginal likelihood function under the exact multinomial case equation (4.3).

- Model M_I , independence model for 2-way contingency table

$$pr(\mathbf{n}|M_I) = \frac{n!}{\prod_{ij}^{r,c} n_{ij}!} \left(\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \right)^2 \prod_{i}^{r} \left(\frac{\Gamma(n_{i+}+\alpha_{i+})}{\Gamma(\alpha_{i+})} \right) \prod_{j}^{c} \left(\frac{\Gamma(n_{+j}+\alpha_{+j})}{\Gamma(\alpha_{+j})} \right).$$

This is, also, equal to the marginal likelihood for M_I under the exact multinomial sampling scheme, see equation (4.4). This indicates that the Bayes factor B_{IS} for the independence model, M_I, against the saturated model, M_S, in the finite population case is equal to the Bayes factor for the exact multinomial sampling, or srs, equations (4.5).

Also, for the completely specified model, M_G, against the saturated model, M_S, the Bayes factor B_{GS} is equal to the Bayes factor for exact multinomial sampling in equation (4.9). This indicates no inference effect in the finite population case.

This result is equivalent to the standard model comparison result, which has been stated by Rao and Thomas (1989), where the finite population is regarded as a random sample from an infinite superpopulation. Then the Pearson statistic X^2 is asymptotically correct, i.e. no finite population correction is necessary.

4.3 Stratified sample

One Bayesian view of stratification is partitioning the population units in such a way, that within each partition (stratum) the elements are roughly exchangeable a priori. Prior exchangeability represents homogeneity within stratum. For designs involving this, detailed in section (2.7.2), we will treat each stratum independently as assumed in the design.

We will derive the Bayes factor for the competing models, the saturated model, M_S, and the model of interest, which specifies all cell probabilities for a 2 way table, M_G.

4.3.1 Uncorrected Bayes factor

If the researcher ignores the stratification, and considers $\mathbf{n}$ to be drawn from a multinomial distribution with parameters n, and the marginal (over strata) probabilities $\mathbf{q} = (p_1, p_2, ..., p_K)'$, see equation (2.1), and a Dirichlet distribution with probability function (4.1) as prior for $\mathbf{q}$. Then, the marginal likelihood $pr(\mathbf{n}|\mathbf{q},M_S)$ for the saturated model, M_S, will be equal to (4.3).

If we consider the model of interest to be the completely specified model M_G, i.e. $\mathbf{q} = \mathbf{p}_0 = (p_{0_1}, p_{0_2}, ..., p_{0_K})'$, then the marginal likelihood $pr(\mathbf{n}|\mathbf{q},M_G)$ is equal to (4.7). Therefore, the Bayes factor, B_{GS}, the ratio of the marginal likelihoods, is exactly (4.9).

Unfortunately, this Bayes factor B_{GS} is not taking account of the true sampling design, stratification, but instead it consider the multinomial sampling scheme as the sampling design.

4.3.2 Bayes factor, under stratification

- Model M_S, saturated model;

Since strata are independent, we can consider each stratum as a separate finite population. Consider a sample of $\mathbf{n}_l' = (n_{l1}, n_{l2}, ..., n_{lK})$ units sampled without replacement from $\mathbf{N}_l' = (N_{l1}, N_{l2}, ..., N_{lK})$ units, where n_{li} denotes the i^{th} cell total in stratum l and $n = \sum_{l=1}^{L}\sum_{i=1}^{K} n_{li}$ and $N = \sum_{l=1}^{L}\sum_{i=1}^{K} N_{li}$, then we may think of the class of prior distributions of the individuals who make up $N_l, = \sum_{i=1}^{K} N_{li}$, to be, a priori, independent and reflect exchangeability, see Ericson (1969) and section (4.2). As in the finite population case, we consider a $(K-1)$-dimensional Dirichlet distribution (4.1) with parameters $\boldsymbol{\alpha}_l$ as the prior on $\mathbf{p}_l$, where $\mathbf{p}_l$ is the vector of the population proportions for stratum l, and $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_L)$ are independent. Thus, $pr(\mathbf{p}|\boldsymbol{\alpha}) = \prod_{l=1}^{L} pr(\mathbf{p}_l|\boldsymbol{\alpha}_l)$ and the likelihood of $\mathbf{n}$ will be $pr(\mathbf{n}|\mathbf{p}) = \prod_{l=1}^{L} pr(\mathbf{n}_l|\mathbf{p}_l)$, since strata are independent. Moreover, the marginal likelihood for the saturated model, M_S, under stratification is equal to,

$$\begin{aligned} pr(\mathbf{n}|M_S) &= \int_{p\in\Omega} pr(\mathbf{n}|\mathbf{p}, M_S) \times pr(\mathbf{p}|M_S)\, d\mathbf{p} \\ &= \int_{p\in\Omega} \prod_{l=1}^{L} pr(\mathbf{n}_l|\mathbf{p}_l)\, pr(\mathbf{p}_l|\boldsymbol{\alpha}_l) d\mathbf{p} \\ &= \prod_{l=1}^{L} \int pr(\mathbf{n}_l|\mathbf{p}_l)\, pr(\mathbf{p}_l|\boldsymbol{\alpha}_l) d\mathbf{p}_l \\ &= \prod_{l=1}^{L} \int \frac{n_l!}{\prod_i n_{li}!} \prod_i^K p_{li}^{n_{li}} \frac{\Gamma(\alpha_l)}{\prod_i \Gamma(\alpha_{li})} \prod_i^K p_{li}^{\alpha_{li}-1} d\mathbf{p}_l. \end{aligned}$$

Therefore

$$pr(\mathbf{n}|M_S) = \prod_{l=1}^{L} \left(\frac{n_l!}{\prod_i n_{li}!} \frac{\Gamma(\alpha_l)}{\Gamma(n_l + \alpha_l)} \prod_i^K \left(\frac{\Gamma(n_{li} + \alpha_{li})}{\Gamma(\alpha_{li})} \right) \right).$$

If we consider $\mathbf{q}$ to be the marginal (over strata) probabilities under the saturated model. Then, $\mathbf{q}$ are defined as

$$p_i = \sum_{l=1}^{L} w_l p_{li} \tag{4.21}$$

where $w_l = \frac{n_l}{n}$ and p_{li} denote the probability of cell i in stratum l, for all $i = 1, ..., K-1$. This means, that p_i refers to a subclass across strata, of a population.

- Model of interest M_G, where $\mathbf{q} = \mathbf{p}_0$.

Suppose Model M_G is defined by $\mathbf{q} = \mathbf{p}_0 = (p_{0_1}, p_{0_2}, ..., p_{0_K})'$, this is given by the constraint (4.21), i.e. $p_{0_i} = \sum_{l=1}^{L} w_l p_{li}$.

In most cases the variables defining the domain cannot be observed before sampling and so cannot be incorporated into the sampling design as a separate stratum. Hence the subclasses usually cut across strata and contain an unknown number of elements within each stratum. With this design, it is quite difficult to evaluate the marginal likelihood function. Therefore, it is hard to compute a Bayes factor.

Nevertheless, one possibility is to approximate the Bayes factor without ever computing the marginal likelihood directly. Since, we are considering a prior on $\mathbf{p}$ for the saturated model, M_S, conditioning on the constraint (4.21) in such a way to be equal to the prior of our nested model, M_G, we can apply the Savage-Dickey density ratio (Dickey, 1971), see section (3.3). Verdinelli and Wasserman (1995) generalized the Savage-Dickey density ratio for general prior choice. Using the Savage-Dickey density ratio, we can approximate the Bayes factor without computing the marginal likelihood $pr(\mathbf{n}|M_G)$. The Savage-Dickey density ratio reduces computing the Bayes factor to the problem of estimating the marginal posterior density $pr(\mathbf{p}|\mathbf{n},M_S)$, at point $\mathbf{p}_0$. Under the saturated model, the Bayes factor will be,

$$B_{GS} = \frac{pr(\mathbf{p}_0|\mathbf{n},M_S)}{pr(\mathbf{p}_0|M_S)}. \tag{4.22}$$

- Evaluating the marginal posterior density $pr(\mathbf{p}|\mathbf{n},M_S)$,

$$\begin{aligned} pr(\mathbf{p}|\mathbf{n},M_S) &\propto pr(\mathbf{n}|\mathbf{p},M_S) \times pr(\mathbf{p}|\boldsymbol{\alpha}) \\ &= \prod_{l=1}^{L} \frac{n_l!}{\prod_i^K n_{li}!} \prod_i^K p_{li}^{n_{li}} \prod_{l=1}^{L} \frac{\Gamma\left[\sum_i \alpha_{li}\right]}{\prod_i^K \Gamma(\alpha_{li})} \prod_i^K p_{li}^{\alpha_{li}-1} \\ &\propto \prod_l \prod_i p_{li}^{n_{li}+\alpha_{li}-1} \\ &= \left(\prod_i^K p_{1i}^{n_{1i}+\alpha_{1i}-1}\right) \times ... \times \left(\prod_i p_{Li}^{n_{Li}+\alpha_{Li}-1}\right) \end{aligned}$$

this implies,

$$pr(\mathbf{p}|\mathbf{n},M_S) = \prod_{l=1}^{L} \text{Dirichlet}(\mathbf{n}_l + \boldsymbol{\alpha}_l)$$

where $\mathbf{n}_l' = (n_{l1}, n_{l2}, ..., n_{lK})$, and $\boldsymbol{\alpha}_l' = (\alpha_{l1}, \alpha_{l2}, ..., \alpha_{lK})$, for all $l = 1, .., L$.

- Computing $pr(\mathbf{p}|\mathbf{n},M_S)$ and $pr(\mathbf{p}|\boldsymbol{\alpha})$ at point $\mathbf{p}_0$.

  - Sample, using a Monte Carlo Method,

    Using a Monte Carlo sampling method, we sample $J, = 1000$, observations of $\mathbf{p}_j (j = 1, ..., J)$ from the marginal posterior density, $pr(\mathbf{p}|\mathbf{n},M_S)$, where in the real data the observation $\mathbf{n}_l$ are given. Also we sample J observations of $\mathbf{p}_j (j = 1, ..., J)$ from the prior $pr(\mathbf{p}|\boldsymbol{\alpha}) = \prod_{l=1}^{L} \text{Dirichlet}(\boldsymbol{\alpha}_l)$, independently. Then, we estimate the marginal densities of $pr(\mathbf{p}|\mathbf{n},M_S)$ and $pr(\mathbf{p}|\boldsymbol{\alpha})$ at point $\mathbf{p}_0$.

  - Estimating the multivariate densities of $pr(\mathbf{p}|\mathbf{n},M_S)$ and $pr(\mathbf{p}|\boldsymbol{\alpha})$ at point $\mathbf{p}_0$, using a kernel method. In the examples in this chapter, we use a crude uniform kernel; for more details see section (5.5.3).

4.3.3 Simulation study

We are going to consider the case where we have two strata , $L = 2$, and three parameters, $K = 3$, where the vector of parameters of the prior distribution,

Example	w_1	$\mathbf{p}_1$	$\mathbf{p}_2$
1	0.5	$(\frac{1}{3}\ \frac{1}{3}\ \frac{1}{3})$	$(\frac{1}{3}\ \frac{1}{3}\ \frac{1}{3})$
2	0.5	(0.2 0.3 0.5)	(0.5 0.3 0.2)
3	0.5	(0.2 0.3 0.5)	(0.4 0.3 0.3)
4	0.5	(0.89 0.1 0.01)	(0.01 0.1 0.89)
5	0.5	(0.2 0.2 0.6)	(0.6 0.2 0.2)
6	0.5	(0.6 0.32 0.08)	(0.1 0.24 0.66)
7	0.5	(0.8 0.11 0.09)	(0.05 0.1 0.85)

Table 4.1: Examples considered in the stratification case.

$pr(\mathbf{p}|\boldsymbol{\alpha})$, $\boldsymbol{\alpha}_l'=(\frac{1}{6},\frac{1}{6},\frac{1}{6})$, $l=1,2$, is equivalent to a single prior observation evenly distributed over cells of both strata. For the posterior distribution, $pr(\mathbf{p}|\mathbf{n},M_S)$, we consider cases where the strata are homogeneous and where they are not. These cases are presented as seven examples in table (4.1).

In each example, of $\mathbf{p}_l$, $l=1,2$, we generate 1000 samples using a Multinomial distribution, where $n=1000$, and $n_l = w_l n$. Then, using a Monte Carlo sampling method, for each stratum sample parameter $\mathbf{n}_l + \boldsymbol{\alpha}_l$, we generate another 1000 samples from the marginal posterior density $pr(\mathbf{p}_l|\mathbf{n}_l,M_S)$, which is equal to a Dirichlet$(\mathbf{n}_l + \boldsymbol{\alpha}_l)$ distribution. For the prior distribution case, we sample 1000 observations.

In this simulation we restricted the sample size to 1000 observations from two strata, and $w_l = 0.5$, where $l=1,2$. We compared the Bayesian Information Criterion, BIC, see section (3.5), based on four statistics with 2 ln(Bayes factor). The statistics used are the Pearson chi-squared, X^2, the first- and second-order corrections to Pearson chi-squared (Rao and Scott, 1981), X^2_{RS1} and X^2_{RS2}, and the Wald statistic, X^2_W. In section (2.6), we showed the well-known result that the difference between Pearson chi-squared, X^2, and the likelihood-ratio chi-squared, G^2, statistics converges in probability to zero, and this result can be apply to any sampling scheme (Agresti, 1990). Thus, we can consider

$$\begin{aligned} 2\ln B_{21} &\approx BIC(X^2) \\ &= X^2 - d_{21}\ln(n) \end{aligned} \tag{4.23}$$

see equation (3.20). Unfortunately, using G^2 or equivalently X^2 will not take account of the design affect. Therefore, we consider BIC approximations based on the first- and second-order corrections of Rao and Scott (1981) to the Pearson chi-squared statistic

$$BIC(X^2_{RS1}) = X^2_{RS1} - d_{21}\ln(n) \quad (4.24)$$

$$BIC(X^2_{RS2}) = X^2_{RS2} - d_{21}\ln(n) \quad (4.25)$$

and the Wald statistic

$$BIC(X^2_W) = X^2_W - d_{21}\ln(n). \quad (4.26)$$

Finally, we compare two times the logarithm of the uncorrected, multinomial-based, Bayes factor with 2 ln(Bayes factor).

4.3.4 The program algorithms

We wrote two programs in order to do this simulation. The first program does the following;

- Generate s samples of $\mathbf{n}_l$from a multinomial distribution with parameter $n_l = w_l n$ and $\mathbf{p}'_l = (p_{l1}, p_{l2}, ..., p_{lK})$.
- Compute matrix inverse and eigenvalues (Press *et al.*,1988).
- Evaluate the Pearson chi-squared statistic, equation (2.7).
- Evaluate the corrected Pearson chi-squared statistic by Rao and Scott(1981), equation (2.39).
- Evaluate the Wald statistic, equation (2.6).
- Computes uncorrected, i.e. multinomial-based, Bayes factor, equation (4.9).

- Computes the BIC based on a Pearson chi-squared statistic, corrected Pearson chi-squared statistic and Wald statistic (see sections 2.8 and 3.5).

The second program is mainly to evaluate the Savage-Dickey density ratio, as an approximation to the Bayes factor, for $s = 1000$ samples. Using the Monte Carlo sampling method, we sampled from $pr(\mathbf{p}|\mathbf{n},M_S) = \prod_{l=1}^{L} \text{Dirichlet}(\mathbf{n}_l + \boldsymbol{\alpha}_l)$. An efficient way to sample from Dirichlet($\boldsymbol{\theta}$) is to draw $x_1, \ldots, x_K$ from independent gamma distributions with common scale (in our examples the scale is equal to 1) and parameters $\theta_{1,}, \ldots, \theta_K$, and for each j, let $\theta_j = \frac{x_j}{\sum_{i=1}^{K} x_i}$, see Gelman *et al.* (1996). For sampling from Gamma distributions see Ahrens and Dieter (1974, 1982), and Press *et al.* (1988). Also, to evaluate the approximation, we estimated the multivariate density of $pr(\mathbf{p}|\mathbf{n},M_S)$ and $pr(\mathbf{p}|\boldsymbol{\alpha})$ at $\mathbf{p}_0$.

4.3.5 Result

All the results we are going to discuss in this section are based on 1000 samples. We ran the simulation for example (1), and as we expected, the values of all the approximations of BIC are very similar to the values of 2 ln(Bayes factor). This can be seen in figure (4.4). These results for example (1) indicate that the true sampling design has no effect, if we ignore it and assume multinomial sampling, as seen in figure (4.4-a), where the multinomial-based BIC is a good approximation to 2 ln(Bayes factor).

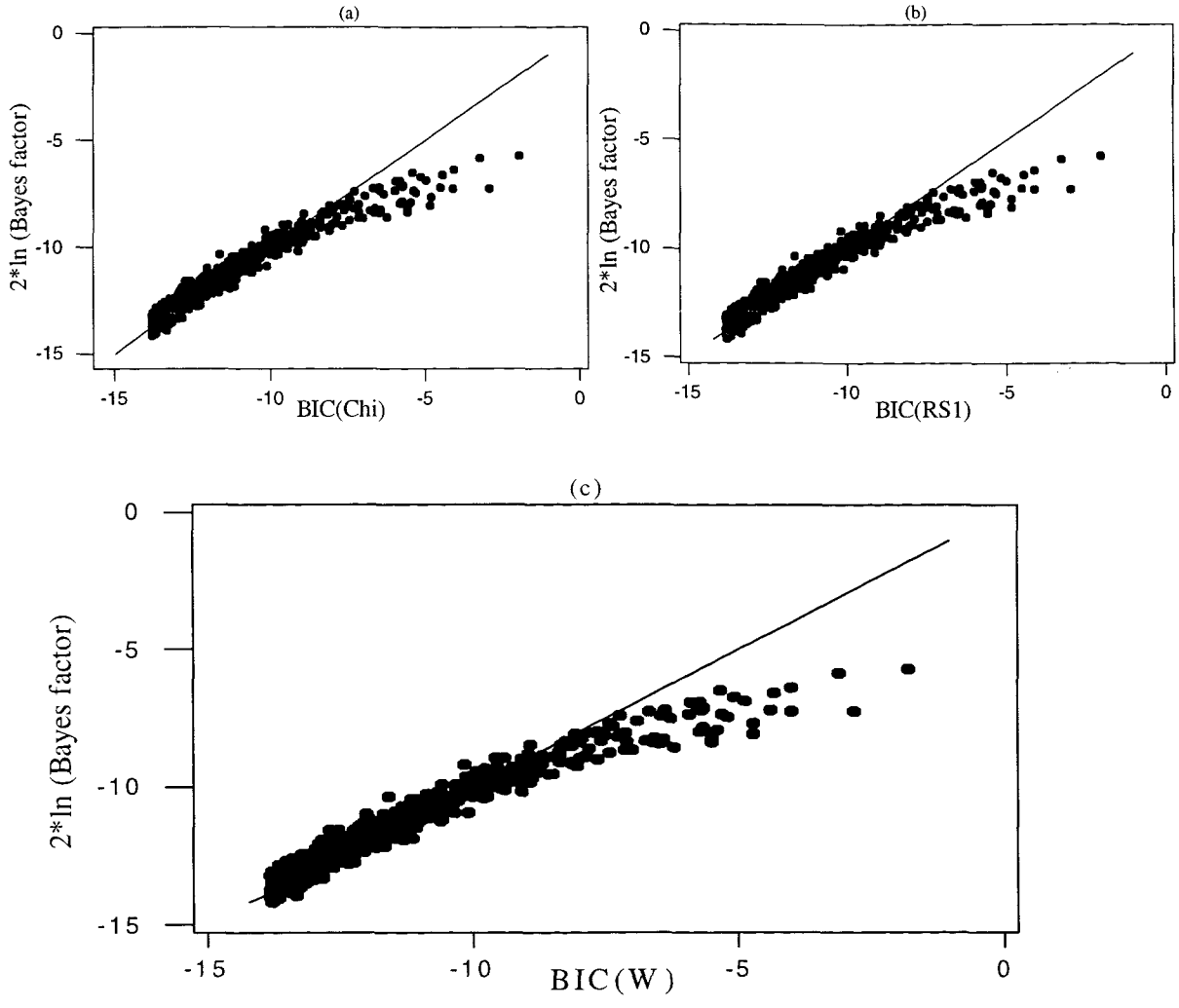

Figure 4.4: Comparison of $2\ln$(Bayes factor) and the BIC's for example (1), where $\mathbf{p}_1' = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, $\mathbf{p}_2' = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, and $w_1 = 0.5$.

If we consider more variable strata, compared with example (1), such as example (2), the results are very similar to the results of example (1). Figure (4.5-a) indicates that if a researcher ignores the stratification by using BIC based on the Pearson chi-squared statistic, the results will be very similar to considering the true sampling design.

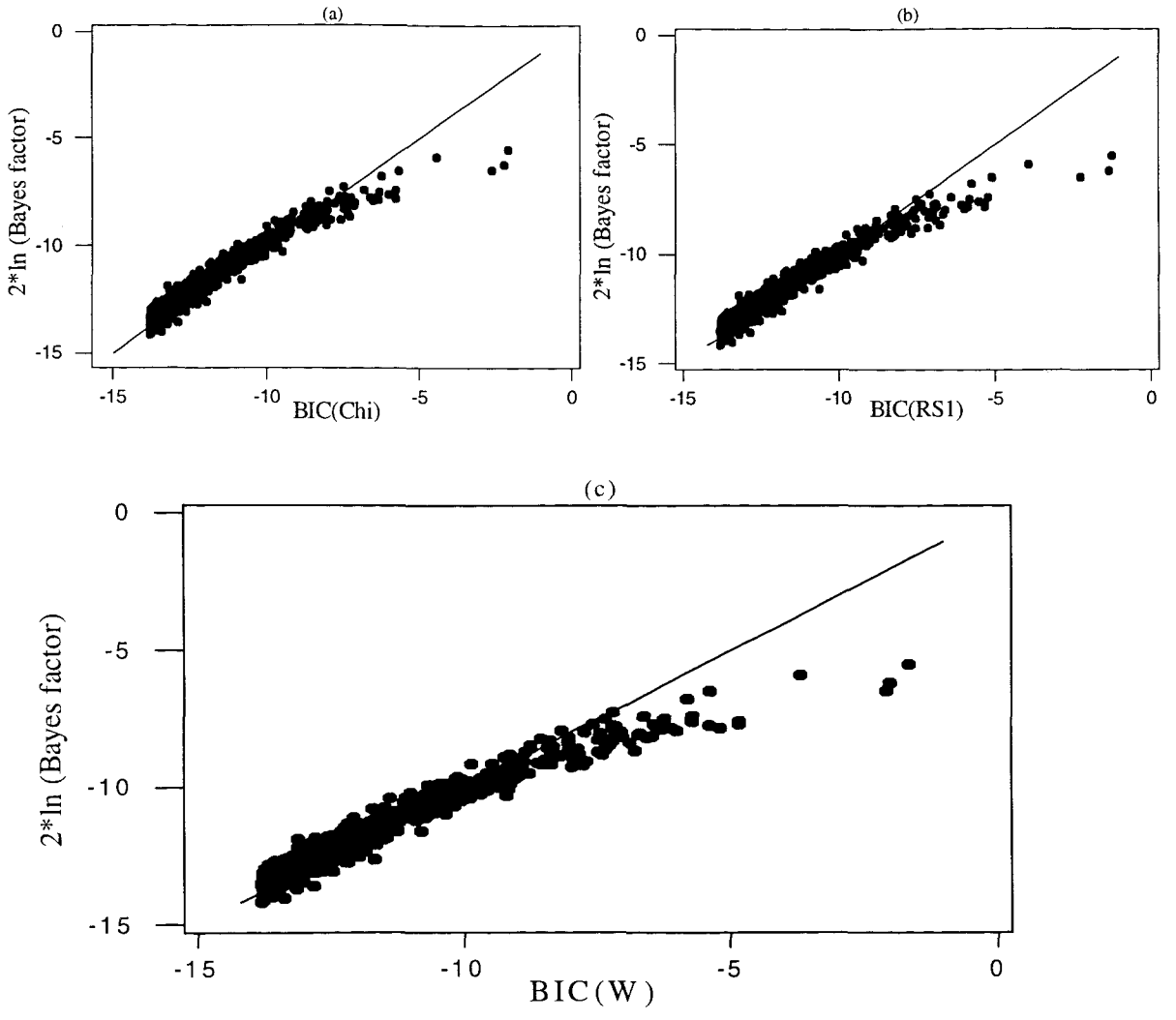

Figure 4.5: Comparison of $2\ln$(Bayes factor) and the BIC's for example (2), where $\mathbf{p}_1' = (0.2, 0.3, 0.5)$, $\mathbf{p}_2' = (0.5, 0.3, 0.2)$, and $w_1 = 0.5$.

For example (3) the results in figure (4.6) are also similar to both results in examples (1), and (2). In all three examples the estimated values of $2\ln$(Bayes factor) never exceed the value -5; This may be related to the crude density estimator used to approximate the Bayes factor.

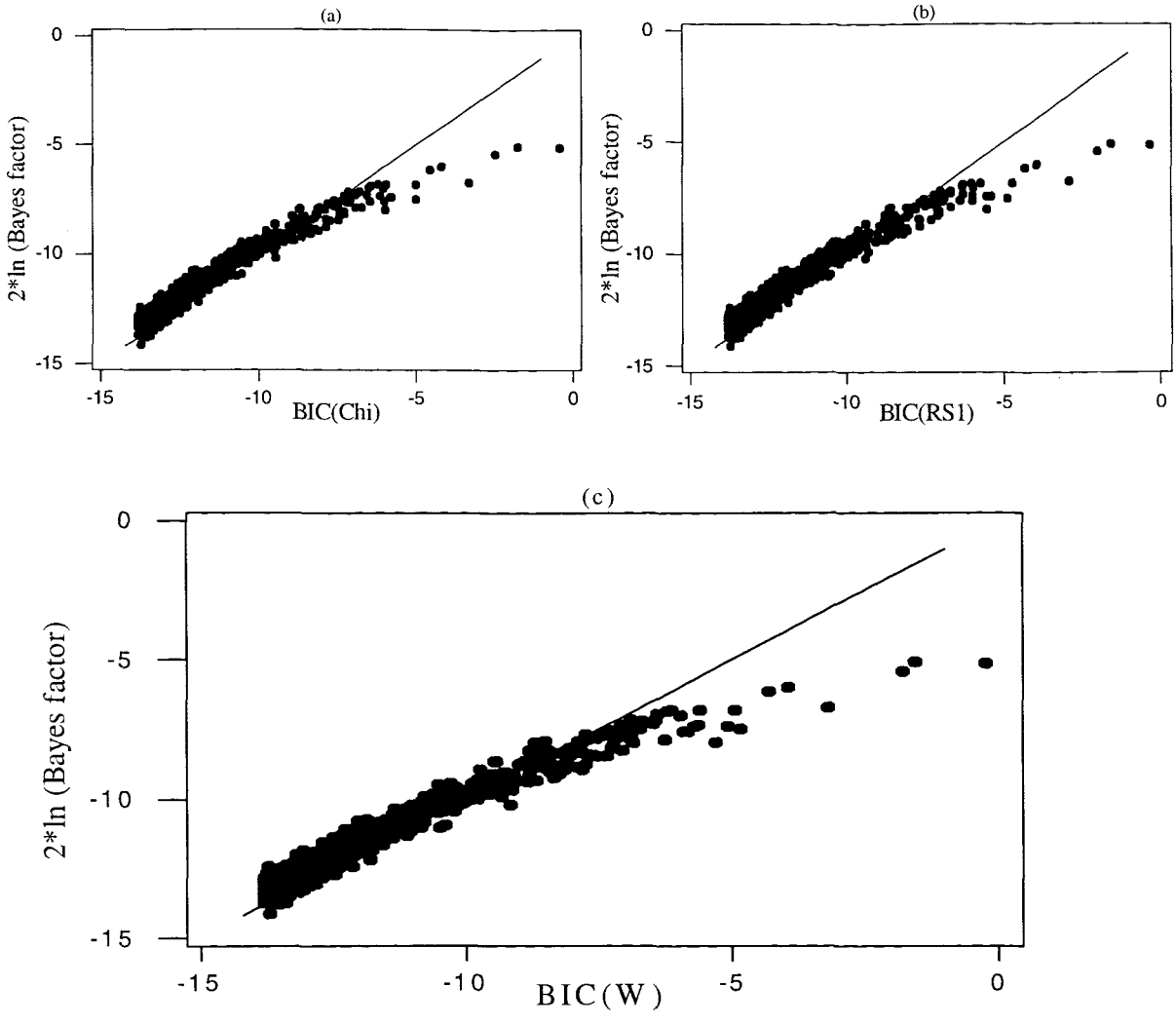

Figure 4.6: Comparison of $2\ln$(Bayes factor) and the BIC's for example (3), where $\mathbf{p}_1' = (0.2, 0.3, 0.5)$, $\mathbf{p}_2' = (0.4, 0.3, 0.3)$, and $w_1 = 0.5$.

For example (4), where the strata are extremely inhomogeneous, the effect of the design is clear in figure (4.7-a). The values of BIC(Chi), which is based on the multinomial sampling design, underestimates the values of $2\ln$(Bayes factor). Using the Rao and Scott first-order correction for the Pearson chi-squared statistic as a base for the approximation of BIC provides better results than BIC(Chi), see figure (4.7-b). In figure (4.7-c) the values of the approximated BIC based on Wald statistic are similar to the values of $2\ln$(Bayes factor).

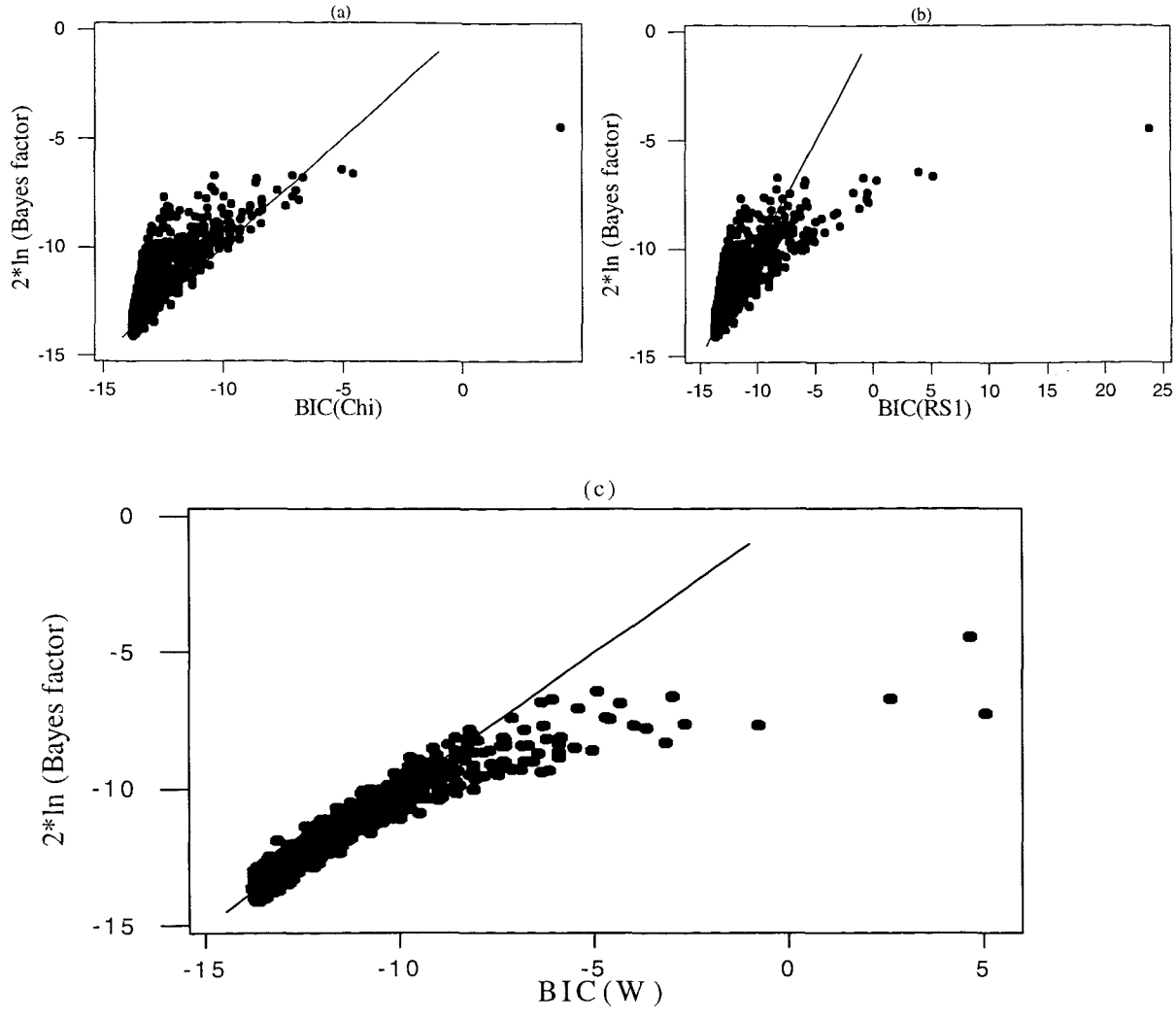

Figure 4.7: Comparison of $2\ln$(Bayes factor) and the BIC's for example (4), where $\mathbf{p}_1' = (0.89, 0.1, 0.01)$, $\mathbf{p}_2' = (0.01, 0.1, 0.89)$, and $w_1 = 0.5$.

For the comparison of the values of $2\ln$(uncorrected Bayes factor), based on multinomial sampling, with the values of $2\ln$(Bayes factor), which considers the true sampling design, figure (4.8) shows that $2\ln$(uncorrected Bayes factor) takes values smaller than $2\ln$(Bayes factor), i.e. underestimates the true values. This suggests that estimation of the Bayes factor is sensitive to this sampling design.

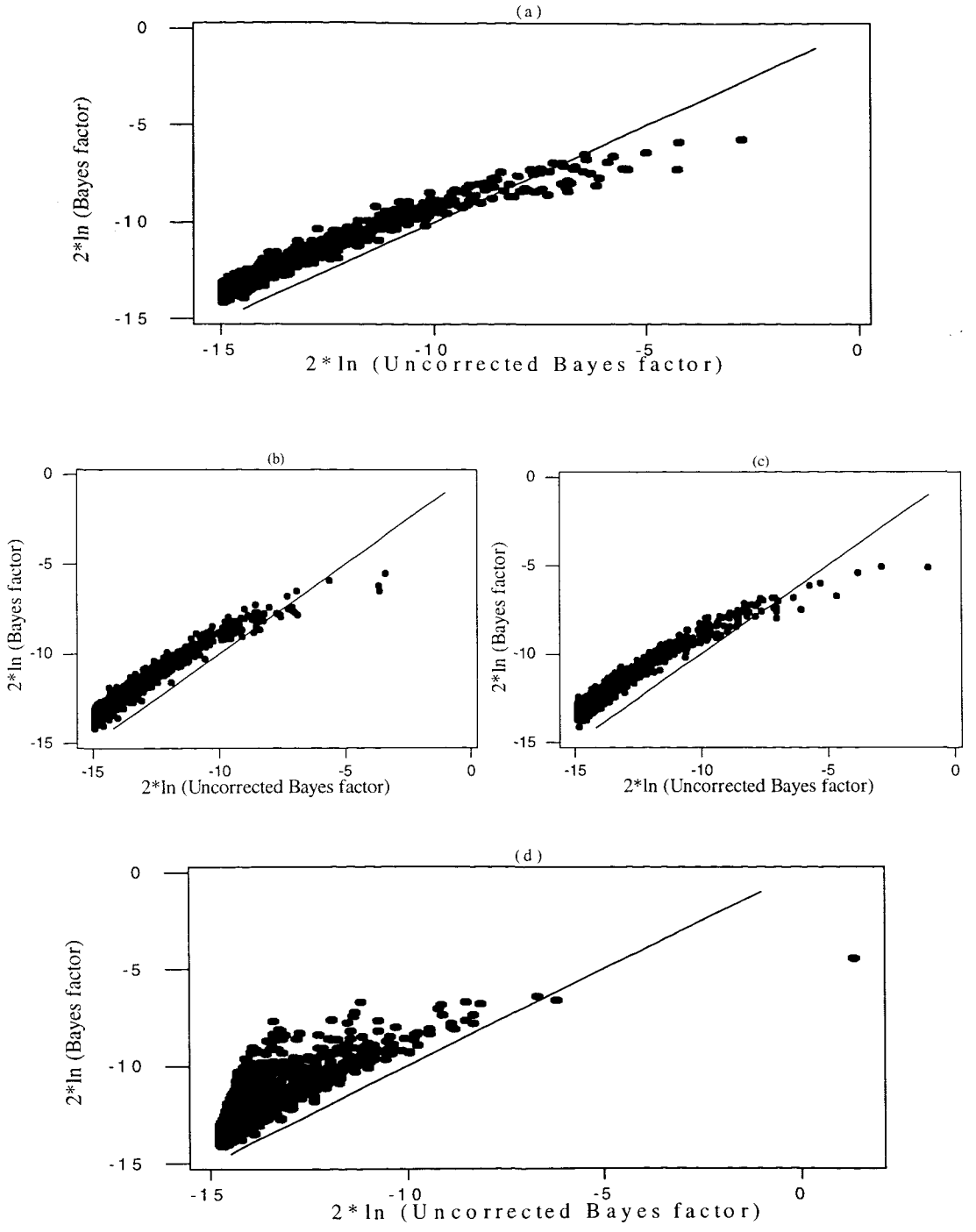

Figure 4.8: Comparison of values of the Bayes factor, which take account of proportional stratification and uncorrected Bayes factor, (a) for example (1), (b) for example (2), (c) for example (3), and (d) for example (4).

4.3.6 Discussion

In the simulation we ran all the seven examples presented in table (1). The results of all those examples are supporting the model of interest, M_G, against the saturated model, M_S. In the results we presented four of those examples including homogeneous strata and inhomogeneous strata. For the homogeneous strata the values of the design effect, $\hat{\tau}_.$, averaged 1 for first example, 0.94 for the second example, and 0.97 for the third example, see figure (4.9). This implies that ignoring the true sampling design, proportional stratification, will not have any serious effect on the analyses of the data.

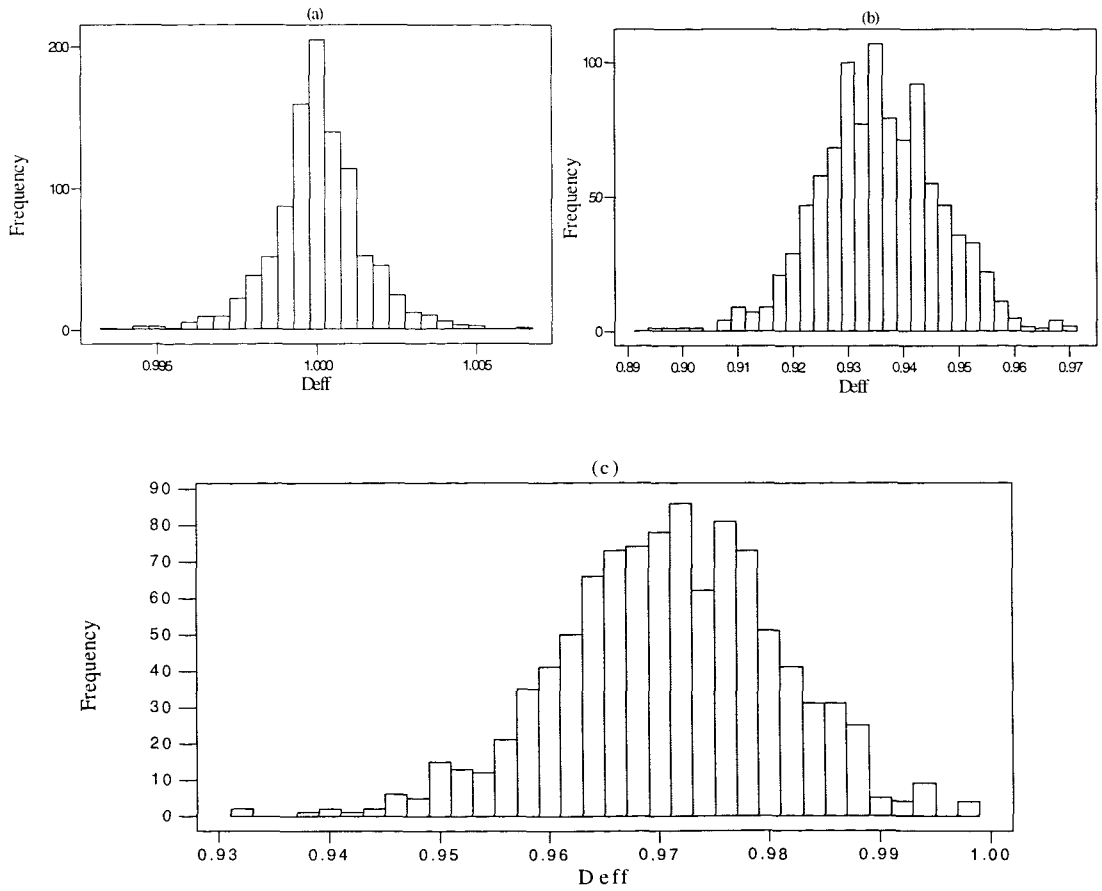

Figure 4.9: A histogram for the values of design effect, $\hat{\tau}_.$, in the 1000 samples of the simulation, (a) for example 1, (b) for example 2, and (c) for example 3.

Nevertheless, for inhomogeneous strata, ignoring this sampling design may affect the results. This is clear in example 4 where the values of design effect, $\hat{\tau}_.$, averaged 0.57, see figure (4.10). Thus , using the Rao and Scott correction is recommended. If the full covariance matrix is known, then using a Wald statistic as a base for computing the BIC approximation will give similar results to $2\ln$(Bayes factor). The sensitivity of the estimation of the Bayes factor to the sampling design is clear even under homogeneous strata and under example (1). The effect of the crude density estimate on the large values of $2\ln$(Bayes factor) is visible. Since the values of $2\ln$(Bayes factor) never exceed the value -5, this will be more visible in the cluster sampling section (6.4.2).

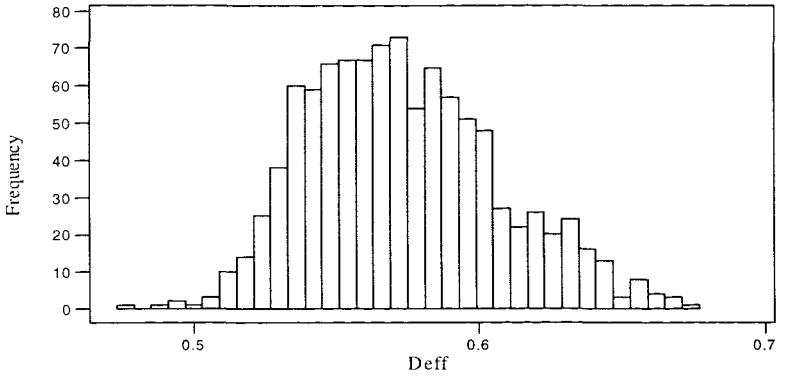

Figure 4.10: A histogram for the values of design effect, $\hat{\tau}_.$, in the 1000 samples of the simulation for example 4.

4.3.7 Conclusion

In all cases, we compared the Bayes factor under stratification with three BIC values, each of which is based on different statistics, namely Pearson chi-squared, corrected chi-squared (Rao and Scott, 1981), and Wald statistics. The results are almost identical if the strata are homogeneous. The various BICs and $2\ln$(Bayes factor) perform almost equally. This indicates that there is no effect in ignoring the sampling design, proportional stratification, on the analysis of

the data if the strata are homogeneous. This result has been mentioned in the classical approach by Kish and Frankel (1974). In the simulations that we have done, for proportional stratification, the values of the design effects, equation (2.18), are almost equal to 1. This can also be seen in the empirical results of Holt, Scott and Ewings (1980).

But if strata are highly inhomogeneous, such as example 4, ignoring this sampling design may affect the results. Using BIC based on corrected Pearson chi-squared statistics will adjust the result, but with large variation. On the other hand, the BIC based on Wald statistic is close to 2 ln(Bayes factor). In this case the design effects averaged 0.57. Therefore, the BICs based on Pearson chi-squared, and corrected chi-squared statistics did not perform as well as for homogeneous strata. In all cases, the Pearson statistic X^2 is always asymptotically conservative under stratified random sampling, see equation (2.34). The result of using the second-order correction by Rao and Scott (1981) for the Pearson chi-squared statistic is always slightly better than first order correction.

Nevertheless, in comparing values of 2 ln(Bayes factor), which take account of proportional stratification equation (4.22), and multinomial-based, uncorrected, Bayes factor, equation (4.9), the 2 ln(uncorrected Bayes factor) underestimates the true values, 2 ln(Bayes factor), i.e. is conservative; see figures (4.8). This indicates that estimation of the Bayes factor is sensitive to the effect of stratification.

Finally, the effect of using a crude density estimate on the large values of 2 ln(Bayes factor) is visible; This will be even more visible in the cluster sampling scheme section (6.4.2).

Chapter 5

Kernel Density Estimator

Reliable point density estimates are required to implement the Savage-Dickey method for estimating Bayes factors. In general, a good density estimate should not only be close to the true density function, but should also reflect important features of interest of the underlying density function. General features of interest often include modes, or local maxima; antimodes, or local minima; the number and location of modes and bumps, or regions where the second derivative is negative; points of inflection, or the regions where the first and the second derivative equal to zero. Our interest is estimating the density function at a certain point. These features will affect our estimate.

The basic idea of nonparametric density estimation is to relax the parametric assumptions about the data, typically replacing these assumptions with ones about the smoothness of the density. The most common and familiar nonparametric estimator is the histogram, which, unfortunately, produces a nondifferentiable estimate, but which is still useful. Here the assumption is that the density is fairly smooth. Unlike the histogram estimate the kernel density estimate produces a smooth differentiable estimate of the density. Nonparametric methods eliminate the need for model specification. The loss of efficiency

need not be too large and is balanced by reducing the risk of misinterpreting data due to incorrect model specification. For a survey see Simonoff (1996), Wand and Jones (1995), Scott (1992), and Tapia and Thompson (1978).

5.1 Simple density estimator

A simple density estimator, such as a Histogram or Frequency Polygon, can give an informative result. Unfortunately, they do not represent most of the important density features, discussed above. In order to solve this problem, consider data $\{x_1, \cdots, x_n\}$ of size n. Then the definition of the density function is,

$$f(x) = \frac{d}{dx}F(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}. \tag{5.1}$$

where $F(x)$ is the cdf. The histogram estimate is defined by considering h as bin width, dividing the line into bins, and then replacing $F(x)$ with its empirical value. Alternatively, this derivative may be estimated separately at each point x by

$$\hat{f}(x) = \frac{\#\{x_i \in (x-h, x+h]\}}{2nh}. \tag{5.2}$$

This can be rewritten as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{5.3}$$

where

$$K(u) = \begin{cases} \frac{1}{2} & \text{if } |u| \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

The form (5.3) is that of a kernel density estimator, with uniform kernel function K on $(-1, 1)$.

Thus the kernel estimator originated as a numeric approximation to the derivative of the cumulative distribution function. It is a very popular example

of a nonparametric density estimation technique (Rosenblatt, 1956). Parzen (1962) explored these ideas in more detail, and established the basic theory of kernel estimation. In fact, virtually all nonparametric algorithms are asymptotically kernel methods, a fact demonstrated empirically by Walter and Blum (1979) and proved rigorously by Terrell and Scott (1992). The kernel density estimate at x is computed as

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - x_i\right), \tag{5.5}$$

where $K_h(t) = K(t/h)/h$ and K is known as the kernel. The kernel estimator is equivalent to a mixture density. It places a probability mass of size $\frac{1}{n}$, i.e. equally weighted, in the shape of the kernel, which has been scaled by the smoothing parameter h, centered on each data point x_i. $\hat{f}_h(x)$ is nonnegative and integrable to one if the kernel K is assumed to be a density function, such as normal density function. The kernel K is generally taken to be symmetric with mean zero and finite variance. After K has been chosen, the remaining element to be specified is the bandwidth, h, also referred to as a smoothing parameter. The bandwidth is a rescaling factor which determines the extent of the region over which the probability mass for point x_i is spread.

5.2 Bandwidth selection

The choice of the smoothing parameter is quite crucial. If we consider a histogram estimate, as an example, then the variance may be controlled by making h large so that the bins are wide and of relatively stable height; However, the bias is then large. On the other hand, the bias may be reduced by making h small so that the bins are narrow; However, the variance is then large. The bias and variance may be controlled simultaneously by choosing an intermediate value of the bin width, and allowing the bin to slowly decrease as the sample size increases. The trade-off of bias versus variance that results from choosing

the amount of smoothing can be quantified through a measure of accuracy of $\hat{f}$, such as mean integrated squared error, MISE.

For the kernel density estimate, define the kernel K to be a bounded probability density function having finite fourth moment and symmetry about the origin. Also, assume that the underlying density is sufficiently smooth, f'' being continuous, square integrable, and ultimately monotone, i.e. monotone over $(-\infty, -b)$ and (b, ∞) for some $b > 0$. If $h \to 0$ with $nh \to \infty$ as $n \to \infty$, i.e. h approaches zero but at a rate slower than n^{-1}, then by Taylor series expansions, Wand and Jones (1995) show that

$$\text{Bias}[\hat{f}(x)] = \frac{1}{2}h^2\sigma_K^2 f''(x) + o(h^2). \tag{5.6}$$

Clearly, $\hat{f}(x)$ is asymptotically unbiased, since the bias is of order h^2, and depends on the true density function $f(x)$, through $f''(x)$. Thus, if the absolute value of $f''(x)$ is large, such as in peaks where $f''(x)$ is negative or valleys where $f''(x)$ is positive, the bias may be large.

The choice of optimal fixed bandwidth, h^*, involves a bias-variance trade-off. The resulting optimum will be relatively wider not only in the tails but also near the mode where $f'(x)$ is small , i.e. small changes in the density function. In the regions where the density is rapidly changing, the optimal bandwidth will be narrower. Thus, even assuming the best global choice for h, the fact remains that no single value of h will perform well for all points x. For more details see Jones (1990) and Terrell and Scott (1992).

Unfortunately, the smoothing parameter, h, must be selected by the user, and no completely satisfactory method of doing so has been found. In practice, given real data from an unknown density, the smoothing parameter chosen will not be an optimal bandwidth, h^*, but instead be of the form $h = rh^*$. Scott (1992) shows that the departures of h from h^* should be measured in a multiplicative rather than additive fashion.

5.3 Choice of kernel

The kernel density estimate is superior at recovering interesting structure, and highly intuitive compared with other density estimates. Much of the first decade of theoretical work focused upon various aspects of estimation properties relating to the characteristics of a kernel. The quality of a density estimate is now widely recognized to be primarily determined by the choice of smoothing parameter, and only in a minor way by the choice of kernel.

The general advice on choosing a kernel based on the observations is to choose a symmetric kernel that is based on a low-order polynomial, since higher order kernels do not have much impact in practice, for a variety of reasons, such as the need for huge sample size and the bandwidth being more complex (Simonoff, 1996). There is no single kernel that can be recommended for all circumstances. However, ordinary kernel estimates behave consistently if the kernel is sufficiently differentiable and if wider bandwidths are selected. Larger smoothing parameters are required as the derivative of a function is noisier than the function itself.

One serious candidate, for the kernel function, is the normal kernel. Another is the Epanechnikov kernel, but the optimal Epanechnikov kernel is not continuously differentiable and cannot be used to estimate derivatives.

In practice, the ability to switch between different kernels without having to reconsider the bandwidth at every turn is convenient. Therefore, it is straightforward to convert a rule based on one kernel function, such as the normal kernel, to any other kernel by using a simple multiplier. For example to switch from normal kernel to Epanechnikov kernel the bandwidth is 2.214 times the normal bandwidth (Simonoff, 1996). This task is easy to accomplish, but only for kernels of the same polynomial order.

5.4 Problem with kernel density estimation

Despite the apparent benefits of using a kernel density estimate, there are several weaknesses, which are currently being investigated by researchers. The most important shortcomings of this estimate are the use of a single bandwidth h and its performance if the region of the data at hand is bounded.

The ordinary kernel estimator does not allow for different levels of smoothing at different parts of the density, as it is controlled by the single bandwidth h. This is related to the asymptotic MISE, which can not be uniformly minimized. When the bandwidth is chosen to be appropriate to the narrowest feature, where f is rapidly changing, the other features will be undersmoothed, with false modes and bumps. On the other hand, if the bandwidth is chosen to be appropriate to the widest feature, then narrower features may be completely lost or severely biased downwards.

Currently, there is much research on density estimation, where the smoothing parameter selection is one of the most intensively studied subjects. At present, no method can claim to be the best. On the other hand, there are some candidates for a general approach, such as adaptive, and spline density estimates; For details see Simonoff (1996), and Scott (1992). Unfortunately, these are computationally intensive and, for our examples, this would require complex computations with long running times. Gelfand and Smith (1990) explored a Rao-Blackwellised approach to estimate a density function, as they suggested that density estimation such as kernel density estimation ignores the known form of conditional distribution of the variable of interest. Where the conditional distributions are available, this would be an interesting approach to explore.

5.5 Multivariate kernel density estimator

Let the data x_{ij} ($j = 1 \cdots, d$, and $i = 1, \cdots, n$) be defined in a $n \times d$ matrix form. In a product kernel, the same (univariate) kernel is used in each dimension but with a different smoothing parameter for each dimension.

$$\hat{f}(\mathbf{x}) = \frac{1}{n\left(h_1 \times \cdots \times h_d\right)} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{d} K\left(\frac{x - x_{ij}}{h_j}\right) \right\}. \tag{5.7}$$

The estimate is defined pointwise, where $\mathbf{x} = (x_1, \cdots, x_n)'$. As in the univariate case, the estimate places a probability mass of size $\frac{1}{n}$ centered on each sample point. Epanechnikov (1969) proved that as $h_j \to 0$ and $n \prod_{j=1}^{d} h_j \to \infty$ the empirical probability density (5.7) is a consistent estimator of the true probability density $f(\mathbf{x})$ at each point of $\mathbf{X}$. A more general approach is the general multivariate kernel estimator, where we use a $d \times d$ nonsingular matrix $\mathbf{H}$, a bandwidth matrix, to scale the kernel

$$\hat{f}(\mathbf{x}) = \frac{1}{n\left|\mathbf{H}\right|} \sum_{i=1}^{n} K\left(\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_i)\right) \tag{5.8}$$

where $|\mathbf{H}|$ is the absolute value of the determinant of the matrix $\mathbf{H}$. Thus, a general multivariate kernel estimator will include not only an arbitrary multivariate density as a kernel, but also an arbitrary linear transformation of the data. Here, $K : R^d \to R^1$ is a kernel satisfying three moment conditions

$$\begin{aligned} \int_{\Re^d} K(\mathbf{u}) d\mathbf{u} &= \mathbf{1} \\ \int_{\Re^d} \mathbf{u} K(\mathbf{u}) d\mathbf{u} &= \mathbf{0} \\ \int_{\Re^d} \mathbf{u}\mathbf{u}' K(\mathbf{u}) d\mathbf{u} &= \mathbf{I}_d. \end{aligned} \tag{5.9}$$

If K is indeed a multivariate probability density, then the last two conditions summarize many assumptions about the marginal kernels, $\{K_j(u_j), j = 1, \cdots, d\}$. The second condition says that the means of the marginal kernels are all zero.

The third condition states that the marginal kernels are pairwise uncorrelated and that each has unit variance. Thus any simple dependence is assumed to be captured entirely in the matrix $\mathbf{H}$ and not in the kernel.

The transformation matrix $\mathbf{H}$ in equation (5.8) can be incorporated into the kernel definition. For example, it is equivalent to choose K to be $N(\mathbf{0}, \Sigma)$ with $\mathbf{H} = \mathbf{I}_d$, or to choose K to be $N(\mathbf{0}, \mathbf{I}_d)$ with $\mathbf{H} = \Sigma^{\frac{1}{2}}$. Thus, it is possible to choose a multivariate kernel with a simple covariance structure without loss of generality; for details see Simonoff (1996), and Scott (1992).

The general multivariate kernel estimator equation (5.8) requires specification of the bandwidth matrix $\mathbf{H}$. If the underlying data distribution is multivariate normal with covariance matrix Σ, then the asymptotic optimal bandwidth matrix is (Simonoff, 1996)

$$\mathbf{H} = \left(\frac{4}{d+2}\right)^{\frac{1}{(d+4)}} \Sigma^{\frac{1}{2}} n^{-\frac{1}{(d+4)}} \tag{5.10}$$

As the dimension, d, varies, the constant in equation (5.10) ranges over the interval $(0.924, 1.059)$, with a limit equal to 1. The constant is exactly 1 in the bivariate case and smallest when $d = 11$. Hence, an easy-to-remember data-based rule, based on Scott's rule (Scott, 1992), is

$$\mathbf{H} = \Sigma^{\frac{1}{2}} n^{-\frac{1}{(d+4)}} \tag{5.11}$$

It follows that for univariate normal data $h = \sigma n^{-\frac{1}{5}}$ is optimal, for large n. The data based choice of $\mathbf{H}$ will be

$$\hat{\mathbf{H}} = \mathbf{S}^{\frac{1}{2}} n^{-\frac{1}{(d+4)}} \tag{5.12}$$

where $\mathbf{S}$ is an estimate of the covariance matrix of $\mathbf{x}$. For the product kernel this is equivalent to

$$h_j = \sigma_j n^{-\frac{1}{(d+4)}} \ , \ j = 1, \ldots, d. \tag{5.13}$$

5.5.1 Multivariate Normal Kernel

If we consider the multivariate normal distribution as a reference distribution, and bandwidth matrix $\mathbf{H}$ as in equation (5.10). The Normal kernel density estimator will be,

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \frac{1}{n\,|\mathbf{H}|}\sum_{i=1}^{n} K\left(\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)\right) \\ &= \frac{1}{n\,|\mathbf{H}|}\sum_{i=1}^{n}(2\pi)^{-\frac{d}{2}}\exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)'\left(\mathbf{H}^{-1}\right)'\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)\} \\ &= \frac{(2\pi)^{-\frac{d}{2}}}{n\left|a\Sigma^{\frac{1}{2}}\right|}\sum_{i=1}^{n}\exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)'\frac{\left(\Sigma^{-\frac{1}{2}}\right)'\Sigma^{-\frac{1}{2}}}{a^2}(\mathbf{x}-\mathbf{x}_i)\}\end{aligned}$$

where, $a = \left(\frac{4}{d+2}\right)^{\frac{1}{(d+4)}} n^{-\frac{1}{(d+4)}}$,

$$\hat{f}(\mathbf{x}) = \frac{(2\pi)^{-\frac{d}{2}}}{na^d\,|\Sigma|^{\frac{1}{2}}}\sum_{i=1}^{n}\exp\{-\frac{1}{2a^2}(\mathbf{x}-\mathbf{x}_i)'\Sigma^{-1}(\mathbf{x}-\mathbf{x}_i)\} \tag{5.14}$$

where $\mathbf{x}_i = (x_{i1},\ldots,x_{id})'$ and Σ is the true covariance matrix, estimated in practice by $\mathbf{S}$.

5.5.2 Epanechnikov Kernel

We also consider the spherically symmetric version of the Epanechnikov kernel, using the same bandwidth matrix $\mathbf{H}$ as in equation (5.10). The Epanechnikov kernel is

$$K(\mathbf{u}) = \begin{cases} \left[\frac{d(d+2)}{4}\right]\Gamma(\frac{d}{2})\pi^{-\frac{d}{2}}\left(1-\mathbf{u}'\mathbf{u}\right), & \text{if } \mathbf{u}'\mathbf{u} \leq 1 \\ 0, & \text{otherwise.}\end{cases}$$

Thus, the resulting Epanechnikov kernel estimator is

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \frac{\left[\frac{d(d+2)}{4}\right]\Gamma(\frac{d}{2})}{n\pi^{\frac{d}{2}}\,|\mathbf{H}|}\sum_{i=1}^{n}\left(1-\left(\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)\right)'\left(\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)\right)\right) \\ &= \frac{\left[\frac{d(d+2)}{4}\right]\Gamma(\frac{d}{2})}{n\pi^{\frac{d}{2}}a^d\,|\Sigma|^{\frac{1}{2}}}\sum_{i=1}^{n}\left(1-\frac{1}{a^2}(\mathbf{x}-\mathbf{x}_i)'\Sigma^{-1}(\mathbf{x}-\mathbf{x}_i)\right)\end{aligned} \tag{5.15}$$

if $\frac{1}{a^2}(\mathbf{x}-\mathbf{x}_i)'\Sigma^{-1}(\mathbf{x}-\mathbf{x}_i)\leq 1$, and 0 otherwise. Unfortunately, the result using the spherically symmetric version of the Epanechnikov kernel was not very promising in our simulations.

5.5.3 Spherical Uniform Kernel

Consider a random sample $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ from a multivariate density $f(\mathbf{x})$, and we want to estimate this function at point $\mathbf{x}_0 \in B(\mathbf{x}_0)$, where $B(\mathbf{x}_0)$ is a hypersphere with radius r, then

$$pr(\mathbf{x} \in B(\mathbf{x}_0)) = \int_{B(\mathbf{x}_0)} f(\mathbf{x})d\mathbf{x} \approx f(\mathbf{x}_0) \times \text{volume of } B(\mathbf{x}_0)$$

if r is small. Thus

$$\hat{f}(\mathbf{x}_0) \approx \frac{\widehat{pr}(\mathbf{x} \in B(\mathbf{x}_0))}{\text{volume of } B(\mathbf{x}_0)} \tag{5.16}$$

where

$$\widehat{pr}(\mathbf{x} \in B(\mathbf{x}_0)) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{x}_i)$$

and n is the total number in the sample and

$$I(\mathbf{x}_i) = \begin{cases} 1 & \text{if} \quad \mathbf{x}_i \in B(\mathbf{x}_0) \\ 0 & \text{if} \quad \mathbf{x}_i \notin B(\mathbf{x}_0). \end{cases}$$

The volume of the hypersphere $B(\mathbf{x}_0)$ is given by

$$\text{Volume of } B(\mathbf{x}_0) = \frac{r^d \times \pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}.$$

Substituting this in equation (5.16) we have,

$$\hat{f}(\mathbf{x}_0) = \frac{\Gamma(\frac{d}{2}+1)\sum_{i=1}^{n} I(\mathbf{x}_i)}{nr^d\pi^{\frac{d}{2}}} \tag{5.17}$$

where $\sum_{i=1}^{n} I(\mathbf{x}_i)$ equal to the total number of points in $B(\mathbf{x}_0)$. This provides a crude estimate of the density at $\mathbf{x}_0$, for small values of r.

This crude estimator can be viewed as kernel estimator, if we consider the multivariate kernel estimator equation (5.8). Let the bandwidth matrix $\mathbf{H} = r\mathbf{I}$, where $\mathbf{I}$ is a d by d identity matrix, then

$$\begin{aligned}\hat{f}(\mathbf{x}_0) &= \frac{\Gamma(\frac{d}{2}+1)}{nr^d\pi^{\frac{d}{2}}}\sum_{i=1}^{n} I(\|\mathbf{x}_i - \mathbf{x}_0\| < r) \\ &= \frac{\Gamma(\frac{d}{2}+1)}{nr^d\pi^{\frac{d}{2}}}\sum_{i=1}^{n} I\left(\left\|\frac{\mathbf{x}_i - \mathbf{x}_0}{r}\right\| < 1\right) \quad \text{, since } r > 0.\end{aligned}$$

If $\mathbf{H} = r\mathbf{I}$, then $|\mathbf{H}| = r^d$ and $\mathbf{H}^{-1} = \frac{1}{r}\mathbf{I}$ and

$$\begin{aligned}K(\frac{\mathbf{x}_i - \mathbf{x}_0}{r}) &= K(\frac{1}{r}\mathbf{I}(\mathbf{x}_i - \mathbf{x}_0)) \\ &= K\left(\mathbf{H}^{-1}(\mathbf{x}_i - \mathbf{x}_0)\right).\end{aligned}$$

The crude estimator will be,

$$\hat{f}(\mathbf{x}_0) = \frac{1}{n|\mathbf{H}|}\sum_{i=1}^{n} K\left(\mathbf{H}^{-1}(\mathbf{x}_i - \mathbf{x}_0)\right)$$

where

$$K(\mathbf{u}) = \frac{\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}}} I(\mathbf{u})$$

and

$$I(\mathbf{u}) = \begin{cases} 1 & \text{if} \quad \|\mathbf{u}\| < 1 \\ 0 & \text{if} \quad \text{otherwise.} \end{cases}$$

5.6 Simulation study

A simulation study was performed for estimating the values of the Dirichlet density function (4.1) at two points in three different cases corresponding to different levels of variation. We estimate the density function at the mode,

$$pr(\mathbf{p}_0|\boldsymbol{\alpha}) = \frac{\Gamma\left[\sum_j^d \alpha_j\right]}{\prod_j \Gamma(\alpha_j)} \prod_j^d p_{0_j}^{\alpha_j - 1}$$

where $\mathbf{p}_0 = (\frac{1}{d}, \cdots, \frac{1}{d})'$, and $\boldsymbol{\alpha}$ takes three values indicating three different levels of variation. The largest variation of the three settings, case 1, is $\boldsymbol{\alpha} = (\frac{20}{d}, \cdots, \frac{20}{d})'$, then, in case 2, $\boldsymbol{\alpha} = (\frac{100}{d}, \cdots, \frac{100}{d})'$, and the smallest, case 3, is $\boldsymbol{\alpha} = (\frac{1000}{d}, \cdots, \frac{1000}{d})'$. The second point, $\mathbf{p}_{1/2}$, at which we estimate the Dirichlet density, is one whose density value is half the density of the mode, i.e.

$$f(\tfrac{1}{d} + \epsilon, \tfrac{1}{d} + \epsilon, \tfrac{1}{d}, \cdots, \tfrac{1}{d}|\boldsymbol{\alpha}) = \frac{1}{2} f(\tfrac{1}{d}, \cdots, \tfrac{1}{d}|\boldsymbol{\alpha}).$$

In the simulation, we used different dimensions, 2, 3, 4, and 7. For the bandwidth, we used $\hat{\mathbf{H}}$ based on Scott's rule, see equation (5.12). The bandwidth matrix, $\mathbf{H}$, is varied in a multiplicative fashion, i.e. $\mathbf{H} = c\hat{\mathbf{H}}$ where c is a real number.

Unfortunately, the simulation is controlled by many variables, such as the bandwidth matrix, $\mathbf{H}$, dimensions, d, interaction between $\mathbf{H}$ and d, (as d gets larger the elements of $\mathbf{H}$ get smaller), and the sample size. Thus, in order to have an overall analysis, we need to discuss each case and this is not our aim in this research. Our aim of this simulation is simply to know how reliable is the pointwise multivariate kernel density estimator based on the multivariate normal distribution as a reference. We use a Dirichlet density function for investigation, since this distribution is used as a prior for $\mathbf{p}$ in our research.

5.7 Result

In the first case, the variation is large, $\boldsymbol{\alpha} = (\frac{20}{d}, \cdots, \frac{20}{d})'$. The estimates of the Dirichlet density function depend on the dimension, d. When the dimension $d < 4$, our estimates are good. As the bandwidth gets wider, the estimated value of the Dirichlet density function at the mode is becoming smaller than it should be, i.e. we underestimate the true density value, see table (5.1). On the other hand, when the bandwidth gets smaller we sometimes overestimate

the true density value. This result cannot be a systematic error and it may be related to simulation error.

The situation of underestimate and overestimates is reversed, table (5.2), when we estimate the Dirichlet density function at the second point, $\mathbf{p}_{1/2}$. As the dimension gets higher, $d > 4$, our estimated values have a large bias toward underestimating the true values.

In the second and third cases, where the variation is smaller, $\boldsymbol{\alpha} = (\frac{100}{d}, \cdots, \frac{100}{d})'$, and $\boldsymbol{\alpha} = (\frac{1000}{d}, \cdots, \frac{1000}{d})'$, we have a similar case to case one. However, as the variation gets smaller the estimates are more accurate in smaller dimensions. On the other hand, with higher dimension the error increases as the variation decreases, see tables (5.3) and (5.4), and tables (5.5) and (5.6).

In conclusion, the multivariate normal kernel density estimator, with normal based bandwidth (5.12), is reasonably reliable for two- and three-dimensional data. For four-dimensional data, the estimated values are sensitive to the bandwidth. Unfortunately, for higher-dimensional data the error seems to be large.

Sample size	r	Dim 2 $\hat{f}(.)$	Diff %	Dim 3 $\hat{f}(.)$	Diff %	Dim 4 $\hat{f}(.)$	Diff %	Dim 7 $\hat{f}(.)$	Diff %
500		2.8	-21.8	8.0	-50.1	24.0	-71.9	894.9	-96.3
2000	3.00	3.0	-14.3	10.1	-37.1	32.8	-61.6	1365.2	-94.3
10000		3.2	-9.1	11.8	-26.4	45.5	-46.7	2384.5	-90.1
500		3.2	-8.3	13.0	-18.5	57.5	-32.7	8208.1	-65.8
2000	1.50	3.4	-3.4	14.7	-7.9	66.5	-22.1	9462.3	-60.5
10000		3.4	-3.2	14.6	-8.7	72.4	-15.2	12150.6	-49.3
500		3.3	-6.1	15.0	-6.1	74.3	-13.0	12976.6	-45.9
2000	1.00	3.5	-0.7	16.8	5.1	83.4	-2.3	13657.3	-43.1
10000		3.5	-1.6	15.3	-4.7	79.0	-7.4	17485.3	-27.1
500		3.4	-4.7	16.1	0.3	86.8	1.6	11774.7	-50.9
2000	0.75	3.5	0.5	18.0	12.7	93.9	10.0	16070.4	-33.0
10000		3.5	-0.6	15.5	-3.1	79.6	-6.8	19254.1	-19.7
500		3.5	-1.2	16.5	3.4	110.6	29.5	4551.6	-81.0
2000	0.50	3.6	1.8	19.1	19.5	108.5	27.1	31616.4	31.8
10000		3.6	1.0	16.0	-0.2	76.7	-10.2	17530.4	-26.9
500		3.7	4.9	14.5	-9.4	165.6	93.9	14.1	-99.9
2000	0.30	3.7	3.9	18.7	16.7	136.9	60.3	47511.8	98.1
10000		3.6	3.3	17.0	6.2	73.7	-13.7	5999.3	-75.0
Value of	$f(.)$	3.5	0.0	16.0	0.0	85.4	0.0	23982.0	0.0

Table 5.1: Estimates of the Dirichlet density function at the mode, using multivariate normal kernel density estimator and the percentage difference, under case 1.

Sample size	r	Dim 2 $\hat{f}(.)$	Diff %	Dim 3 $\hat{f}(.)$	Diff %	Dim 4 $\hat{f}(.)$	Diff %	Dim 7 $\hat{f}(.)$	Diff %
500		1.9	3.9	5.5	-28.2	16.8	-60.5	697	-94.2
2000	3.00	1.9	3.9	6.4	-16.6	21.5	-49.4	1025	-91.5
10000		1.8	3.6	6.8	-10.9	28.0	-34.1	1654	-86.2
500		1.9	5.5	7.2	-6.0	28.9	-32.0	5769	-51.9
2000	1.50	1.8	3.4	7.4	-3.1	32.8	-22.9	6183	-48.5
10000		1.8	2.9	7.5	-1.2	38.2	-10.2	6971	-41.9
500		1.9	4.1	7.7	0.5	31.1	-26.9	12172	1.5
2000	1.00	1.8	2.7	7.7	0.9	36.1	-15.1	10469	-12.7
10000		1.8	1.9	7.5	-1.5	39.9	-6.1	9789	-18.4
500		1.8	3.5	8.1	5.7	31.5	-25.9	14537	21.2
2000	0.75	1.8	1.3	8.1	5.8	38.0	-10.6	13228	10.3
10000		1.8	0.7	7.3	-3.9	40.4	-4.9	12756	6.3
500		1.8	3.3	8.8	15.6	25.2	-40.7	6615	-44.9
2000	0.50	1.7	-3.1	8.9	16.1	37.1	-12.7	29078	142.4
10000		1.7	-2.1	6.9	-9.2	41.8	-1.6	30139	151.3
500		1.8	1.9	9.8	28.5	8.7	-79.5	39.3	-99.7
2000	0.30	1.6	-9.9	10.0	30.8	22.2	-47.7	50259	319.0
10000		1.7	-5.7	5.8	-23.4	51.1	20.2	62990	425.1
Value of	$f(.)$	1.8	0.0	7.6	0.0	42.5	0.0	11995	0.0

Table 5.2: Estimates of the Dirichlet density function at $\mathbf{p}_{1/2}$ using multivariate normal kernel density estimator and the percentage difference, under case 1.

Sample size	r	Dim 2 $\hat{f}(.)$	Diff %	Dim 3 $\hat{f}(.)$	Diff %	Dim 4 $\hat{f}(.)$	Diff %	Dim 7 $\hat{f}(.)$	Diff %
500		6.1	-23.2	36.8	-55.2	258.3	-74.3	97839	-97.2
2000	3.00	6.9	-13.3	48.0	-41.6	375.6	-62.6	153134	-95.6
10000		7.3	-8.3	58.7	-28.5	484.7	-51.7	266240	-92.4
500		7.3	-8.0	59.4	-27.7	577.7	-42.4	1056592	-69.9
2000	1.50	8.0	0.2	66.5	-19.0	733.3	-26.9	1168588	-66.8
10000		7.9	-1.2	72.9	-11.3	810.4	-19.2	1441841	-59.0
500		7.6	-4.7	64.2	-21.8	650.0	-35.2	2239512	-36.3
2000	1.00	8.2	2.6	67.4	-17.9	839.3	-16.3	1891871	-46.2
10000		8.0	0.1	75.5	-8.0	899.4	-10.4	2055874	-41.5
500		7.7	-3.8	61.8	-24.8	636.1	-36.6	2874423	-18.2
2000	0.75	8.1	2.1	65.2	-20.7	855.3	-14.8	1990340	-43.4
10000		8.0	0.6	77.8	-5.3	928.4	-7.5	1953706	-44.4
500		7.7	-2.8	55.5	-32.5	491.9	-51.0	2289380	-34.9
2000	0.50	7.9	-0.8	61.2	-25.6	818.3	-18.4	1174313	-66.6
10000		8.1	1.6	83.2	1.3	953.6	-5.0	968979	-72.4
500		7.8	-1.4	55.1	-33.0	219.5	-78.1	106324	-97.0
2000	0.30	7.6	-3.9	54.7	-33.4	647.3	-35.5	20411	-99.4
10000		8.2	3.5	91.3	11.1	923.7	-7.9	24097	-99.3
Value of	$f(.)$	8.0	0.0	82.2	0.0	1003.3	0.0	3515075	0.0

Table 5.3: Estimates of the Dirichlet density function at the mode, using multivariate normal kernel density estimator and the percentage difference, under case 2.

Sample size	r	Dim 2 $\hat{f}(.)$	Diff %	Dim 3 $\hat{f}(.)$	Diff %	Dim 4 $\hat{f}(.)$	Diff %	Dim 7 $\hat{f}(.)$	Diff %
500		4.0	3.4	28.4	-34.4	196.6	-60.7	74531	-95.7
2000	3.00	4.0	2.3	33.0	-23.9	260.1	-48.0	117182	-93.3
10000		4.1	4.1	36.6	-15.4	321.6	-35.7	196146	-88.7
500		4.1	4.7	39.6	-8.7	405.9	-18.9	561814	-67.7
2000	1.50	3.9	-0.3	41.0	-5.4	409.0	-18.3	703860	-59.6
10000		4.1	3.9	40.9	-5.5	460.5	-8.0	870200	-50.0
500		4.2	6.9	41.3	-4.6	487.9	-2.5	974886	-44.0
2000	1.00	3.9	0.1	44.0	1.5	414.6	-17.2	1315401	-24.4
10000		4.1	4.0	41.6	-4.0	486.8	-2.7	1310107	-24.7
500		4.3	10.2	39.9	-7.8	521.1	4.1	1154098	-33.7
2000	0.75	3.9	0.3	45.4	4.8	377.4	-24.6	1877198	7.9
10000		4.1	3.7	41.3	-4.6	487.2	-2.7	1434532	-17.6
500		4.5	15.9	35.6	-17.8	555.7	11.0	728789	-58.1
2000	0.50	3.9	0.2	45.7	5.6	297.1	-40.6	1687188	-3.1
10000		4.0	2.5	40.7	-6.2	473.0	-5.5	1118028	-35.8
500		4.8	21.6	31.4	-27.6	645.8	29.0	7174	-99.6
2000	0.30	3.8	-1.6	45.5	5.0	228.3	-54.4	158367	-90.9
10000		3.9	0.3	40.3	-7.0	386.1	-22.8	67729	-94.4
Value of	$f(.)$	3.9	0.0	43.3	0.0	500.5	0.0	1740535	0.0

Table 5.4: Estimates of the Dirichlet density function at $\mathbf{p}_{1/2}$ using multivariate normal kernel density estimator and the percentage difference, under case 2.

Sample size	r	Dim 2 $\hat{f}(.)$	Diff %	Dim 3 $\hat{f}(.)$	Diff %	Dim 4 $\hat{f}(.)$	Diff %	Dim 7 $\hat{f}(.)10^7$	Diff %
500		19.1	-24.4	387.4	-53.1	7780	-75.8	8.26	-97.7
2000	3.00	21.6	-14.5	488.1	-40.9	11970	-62.7	13.96	-96.2
10000		22.8	-9.8	593.7	-28.2	15131	-52.8	25.27	-93.1
500		23.5	-6.9	624.8	-24.4	18470	-42.4	85.39	-76.6
2000	1.50	25.0	-0.9	681.2	-17.6	23648	-26.3	106.03	-70.9
10000		24.4	-3.1	765.1	-7.4	24537	-23.5	140.47	-61.5
500		24.5	-2.9	702.6	-15.0	22734	-29.1	174.0	-52.2
2000	1.00	25.7	1.9	713.3	-13.7	26426	-17.6	188.15	-48.4
10000		25.0	-1.1	815.7	-1.3	26374	-17.8	224.30	-38.4
500		24.7	-2.1	749.1	-9.4	23832	-25.7	236.2	-35.2
2000	0.75	25.8	2.1	703.6	-14.9	25591	-20.2	256.66	-29.6
10000		25.2	-0.2	840.5	1.7	26347	-17.9	269.38	-26.1
500		24.5	-2.8	822.0	-0.5	21694	-32.4	392.11	7.6
2000	0.50	25.5	1.1	652.8	-21.0	22597	-29.6	542.00	48.7
10000		25.3	0.3	854.0	3.3	25815	-19.5	286.24	-21.4
500		23.8	-5.6	877.7	6.2	16318	-49.1	228.46	-37.3
2000	0.30	25.0	-0.8	556.4	-32.7	26530	-17.3	866.91	137.9
10000		25.1	-0.3	852.0	3.1	31225	-2.7	112.68	-69.1
Value of	$f(.)$	25.2	0.0	826.4	0.0	32085	0.0	364.39	0.0

Table 5.5: Estimates of the Dirichlet density function at the mode, using multivariate normal kernel density estimator and the percentage difference, under case 3.

Sample size	r	Dim 2 $\hat{f}(.)$	Diff %	Dim 3 $\hat{f}(.)$	Diff %	Dim 4 $\hat{f}(.)$	Diff %	Dim 7 $\hat{f}(.)10^7$	Diff %
500		12.4	-2.9	287.3	-29.7	5744	-63.9	6.76	-96.3
2000	3.00	12.5	-1.6	326.7	-20.1	8445	-47.0	11.01	-93.9
10000		13.1	2.7	354.1	-13.4	10083	-36.7	18.78	-89.7
500		11.6	-8.6	391.7	-4.2	11542	-27.5	53.88	-70.3
2000	1.50	11.7	-8.2	392.3	-4.0	14435	-9.4	65.06	-64.2
10000		13.0	1.7	396.9	-2.9	14442	-9.3	83.67	-53.9
500		10.4	-18.1	411.7	0.7	13269	-16.7	93.21	-48.7
2000	1.00	11.3	-11.4	383.3	-6.2	15911	-0.1	94.2	-48.1
10000		12.9	1.3	404.3	-1.1	15702	-1.4	124.17	-31.6
500		9.4	-25.8	391.7	-4.2	13656	-14.3	100.24	-44.8
2000	0.75	11.2	-12.3	350.6	-14.2	16204	1.7	78.71	-56.6
10000		12.9	1.4	403.3	-1.3	16444	3.3	160.91	-11.4
500		8.5	-32.9	294.9	-27.8	17613	10.6	46.90	-74.2
2000	0.50	11.5	-9.5	270.3	-33.9	16782	5.4	14.41	-92.1
10000		12.9	1.7	393.3	-3.8	16828	5.7	320.16	76.4
500		9.1	-28.3	105.4	-74.2	30636	92.4	0.28	-99.8
2000	0.30	12.4	-2.2	141.4	-65.4	17252	8.3	0.01	-100.0
10000		12.8	0.9	378.4	-7.4	14241	-10.6	847.2	366.7
Value of	$f(.)$	12.7	0.0	408.8	0.0	15926	0.0	181.53	0.0

Table 5.6: Estimates of the Dirichlet density function at $\mathbf{p}_{1/2}$ using multivariate normal kernel density estimator and the percentage difference, under case 3.

Chapter 6

Cluster Sampling

In this chapter, we consider cluster samples obtained by a two-stage sampling process, as described in section (2.7.3), with some changes in notation. The clusters partition individuals into, C, mutually exclusive groups (clusters) where cluster t $(t = 1, ..., C)$ has $\mathbf{N}^t = (N_{t1}, N_{t2}, ..., N_{tK})'$ units and $N_t = \sum_{i=1}^{K} N_{ti}$. Consider a sample of c clusters sampled from C. Then, as a second-stage, a sample of $\mathbf{n}^t = (n_{t1}, n_{t2}, ..., n_{tK})'$ units is drawn with replacement from the N^t within each selected cluster, t $(t = 1, ..., c)$. We consider equal cluster sample sizes $n_t = \sum_{i=1}^{K} n_{ti} = m$, although our inferences can be obtained equally easily for non equal cluster sizes.

Our primary interest is estimating the marginal (over clusters) probabilities $\mathbf{q} = (p_1, ..., p_K)'$. Let $\mathbf{p}^t = (p_{t1}, p_{t2}, ..., p_{tK})'$, where $\sum_{i=1}^{K} p_{ti} = 1$, be a vector of cell probabilities for the t^{th} cluster. A priori, we assume that $\mathbf{N}^t$ are multinomially distributed with cell probabilities $\mathbf{p}^t$, and the vectors $\mathbf{p}^t; t = 1, \ldots, c$, are independent and identically distributed with $\mathbf{p}^t \sim$Dirichlet$(\lambda\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}' = (\alpha_1, \alpha_2, ..., \alpha_K)$ and $\sum_{i=1}^{K} \alpha_i = 1$, and λ is a parameter which represents the cluster effect.

If the value of λ is large, then the variance of $\mathbf{p}^t$ is small. This leads to a smaller cluster effect. However, for a small value of λ, the variance of $\mathbf{p}^t$ will be large and the cluster, design, effect will be also large. Then, we assume that the vector of counts, $\mathbf{n}^t$, for any cluster is a multinomial distribution conditional on the vector $\mathbf{p}^t$ and $n_t = \frac{n}{c}$.

6.1 Bayes factor

- Model M_S , saturated model;

Consider the parameters $\mathbf{p}^t, \boldsymbol{\alpha}$ and λ. In the real data, the observations $\mathbf{n}^t$ are given. If we consider the $\mathbf{n}^t$ are sampled from $\mathbf{N}^t$, and assuming the class of prior distributions of $\mathbf{N}^t$ are, a priori, independent and exchangeable within cluster, then $\mathbf{n}^t$ are distributed as independent Multinomial$(n_t, \mathbf{p}^t)$; see section (2.1). As before, we know that the Dirichlet distribution is a convenient choice of prior distribution for $\mathbf{p}^t$, because, it is conjugate to the multinomial distribution. Therefore, we will take the prior on $\mathbf{p}^t$ to have a $(K-1)$-dimensional Dirichlet distribution (4.1), with parameters $\lambda\boldsymbol{\alpha}$.

To complete the prior specification at the second stage, a distribution needs to be assigned to the hyperparameters $\boldsymbol{\alpha}$, and λ. Various forms of informative prior could be assigned. For computational convenience, let us assume a hyperprior for $\boldsymbol{\alpha}$ to be Dirichlet$(\boldsymbol{\beta})$. Also, a general hyperprior for λ is denoted $pr(\lambda)$, for now. For the saturated model, we define the marginal (over cluster) probabilities, $\mathbf{q}$,

$$p_{\cdot i} = \sum_{t=1}^{C} w_t p_{ti}$$

where $w_t = \frac{N_t}{N}$ and p_{ti} denote the probability of cell i in family t, for all $i = 1, ..., K-1$. Therefore, $p_{\cdot i}$ is probability of a subclass across clusters. As $C \to \infty$, $\mathbf{q}$ and $\boldsymbol{\alpha}$ coincide.

- Model M_0, where $\mathbf{q} = \mathbf{p}_0 = (p_{01}, p_{02}, ..., p_{0K})'$;

Unfortunately, for computing the Bayes factor for model M_0 against M_S, we face the same problem as in stratification see section (4.3.2). With this design, it is quite difficult to evaluate the marginal likelihood function. Therefore, it is hard to compute a Bayes factor. Nevertheless, one possibility is to approximate the Bayes factor without ever computing the marginal likelihood, as in the stratification case. Since, we are considering a prior for $\mathbf{p}^t$ in the saturated model, M_S, conditioning on the constraint (4.21) to give the prior for our nested model, M_0, we can apply the Savage-Dickey density ratio (Dickey, 1971), see section (3.3). Using the Savage-Dickey density ratio, we can approximate a Bayes factor without computing the marginal likelihood $pr(\mathbf{n}|M_0)$. The Savage-Dickey density ratio reduces computing the Bayes factor to the problem of estimating the marginal posterior density $pr(\mathbf{p}|\mathbf{n},M_S)$, at point $\mathbf{p}_0$. Under the saturated model, the Bayes factor will be,

$$B_{0S} = \frac{pr(\mathbf{p}_0|\mathbf{n},M_S)}{pr(\mathbf{p}_0|M_S)}. \tag{6.1}$$

As the marginal density $pr(\mathbf{p}|\mathbf{n},M_S)$ is intractable and can not be written down in closed form, we propose to sample from $pr(\mathbf{p}|\mathbf{n},M_S)$, and then, use a multivariate density estimator to estimate the marginal posterior density $pr(\mathbf{p}|\mathbf{n},M_S)$ at point $\mathbf{p}_0$.

Now, consider the following random variables $\mathbf{p}^t$, $\boldsymbol{\alpha}$ and λ. Under the saturated model, M_S, the joint posterior distribution of $\mathbf{p}$, $\boldsymbol{\alpha}$, and λ will be

$$\begin{aligned} pr(\mathbf{p}^t, \boldsymbol{\alpha}, \lambda|\mathbf{n}) &\propto \left[\prod_{t=1}^{c}\left(\prod_{i=1}^{K} p_{ti}^{n_{ti}}\right) \times \left(\prod_{i=1}^{K} p_{ti}^{\tau\alpha_i - 1}\frac{\Gamma(\lambda)}{\Gamma(\lambda\alpha_i)}\right)\right] \times \left(\prod_{i=1}^{K}\alpha_i^{\beta_i - 1}\right) \times pr(\lambda) \\ &= (\Gamma(\lambda))^c \times \left(\prod_{t=1}^{c}\prod_{i=1}^{K}\frac{p_{ti}^{n_{ti}+\lambda\alpha_i-1}}{\Gamma(\lambda\alpha_i)}\right) \times \left(\prod_{i=1}^{K}\alpha_i^{\beta_i-1}\right) \times pr(\lambda). \end{aligned} \tag{6.2}$$

From this we can see that $pr(\mathbf{p}^t, \boldsymbol{\alpha}, \lambda|\mathbf{n})$ is awkward to generate from. Thus, we are going to use a Markov chain Monte Carlo (MCMC) algorithm to generate a

random sample for $\mathbf{p}^t$, and to estimate the marginal posterior density $pr(\mathbf{p}|\mathbf{n},M_S)$ at point $\mathbf{p}_0$.

6.2 Markov chain Monte Carlo algorithms

Iterative simulation, especially Markov chain Monte Carlo (MCMC) algorithms, have been increasingly popular in statistical simulation, most notably for drawing simulations from Bayesian posterior distributions. See Tierney (1994), Tanner (1996), and Gilks *et al.* (1996) for examples. One of the main reasons for the popularity of MCMC is that integrations of the joint posterior distribution are often extremely difficult to perform, either analytically or numerically. MCMC methods involve simulating Markov Chains with particular stationary distributions, in order to sample indirectly from posterior distribution. The techniques are based only on elementary properties of Markov chains. These elementary properties, that the chain needs to satisfy, are

- It has to be irreducible, that is, from all starting points, the Markov chain can reach any non-empty set with positive probability, in some finite number of iterations.
- It has to be aperiodic, this stops the Markov chain from oscillating between different sets of states in a regular periodic movement.
- Most importantly, it must be positive recurrent, this can be expressed in terms of the existence of a stationary distribution $f(.)$, say, such that if the initial value $\mathbf{X}_0$ is sampled from $f(.)$, then all subsequent iterates, I, will also be distributed according to $f(.)$.

If the chain satisfies these conditions, then the distribution of $\mathbf{X}_I$ converges to the required stationary distribution (Gilks *et al.*, 1996).

We are going to generate a set of random numbers $\mathbf{p}^t$ using MCMC algorithms. Several Markov chain methods are available for sampling from a posterior distribution. After a suitable burn-in, we can use successive iterations from a MCMC scheme as a dependent sample from the marginal posterior distribution for any function of interest. The two most important MCMC methods are the Gibbs sampler and the Metropolis-Hastings algorithm. They can be used as techniques for generating random variables from a marginal distribution indirectly, without having to calculate the density. Although most applications of MCMC have been in Bayesian models, they can also be extremely useful in a classical approach; see Tanner (1996).

6.2.1 Gibbs Sampler

An algorithm for extracting marginal distributions from the full conditional distribution was formally introduced as the Gibbs sampler in Geman and Geman (1984), although its essence dates at least to Hastings (1970); For more discussion see Gelfand and Smith (1990).

The Gibbs sampler is the most common of the MCMC algorithms. It is an MCMC algorithm that requires all the full conditional distributions to be ready to sample from. For the random variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, suppose we are given a joint density $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$, and our interest is in obtaining a sample from $f(\mathbf{x})$. Rather than compute or approximate $f(\mathbf{x})$ directly, the Gibbs sampler allows us effectively to generate a sample $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_m\}$ from $f(\mathbf{x})$ without requiring $f(\mathbf{x})$.

Gibbs sampling is a Markovian updating scheme that proceeds as follows. Given initial values of the variables $\mathbf{Y} = \mathbf{y}_0$ and $\mathbf{Z} = \mathbf{z}_0$, we generate $\mathbf{x}_0$ from $f(\mathbf{x}|\mathbf{y}_0, \mathbf{z}_0)$. Then, we draw $\mathbf{y}_1$ from $f(\mathbf{y}|\mathbf{x}_0, \mathbf{z}_0)$, also using $\mathbf{y}_1$, we draw $\mathbf{z}_1$ from $f(\mathbf{z}|\mathbf{x}_0, \mathbf{y}_1)$. This will complete one iteration of the scheme. The I^{th} iteration

would be

$$\begin{aligned} \mathbf{X}_I &\sim f(\mathbf{x}|\mathbf{Y}_I = \mathbf{y}_I, \mathbf{Z}_I = \mathbf{z}_I) \\ \mathbf{Y}_{I+1} &\sim f(\mathbf{y}|\mathbf{X}_I = \mathbf{x}_I, \mathbf{Z}_I = \mathbf{z}_I) \\ \mathbf{Z}_{I+1} &\sim f(\mathbf{z}|\mathbf{X}_I = \mathbf{x}_I, \mathbf{Y}_{I+1} = \mathbf{y}_{I+1}). \end{aligned} \tag{6.3}$$

The iteration scheme of (6.3) produces a Gibbs sequence

$$\mathbf{Y}_0, \mathbf{Z}_0, \mathbf{X}_0, \mathbf{Y}_1, \mathbf{Z}_1, \mathbf{X}_1, \ldots, \mathbf{Y}_I, \mathbf{Z}_I, \mathbf{X}_I, \ldots \tag{6.4}$$

After I such iterations we would arrive at $\mathbf{X}_I$. Gelfand and Smith (1990) showed that under mild conditions $\mathbf{X}_I \xrightarrow{d} \mathbf{X} \sim f(\mathbf{x})$ as $I \to \infty$. Thus for large enough I, we can consider $\mathbf{X}_I$ as a simulated observation from $f(\mathbf{x})$.

6.2.2 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a MCMC method for generating samples from arbitrary multivariate distributions; see Metropolis *et al.* (1953), Hastings (1970), Tierney (1994), Chib and Greenberg (1995), and Gilks *et al.* (1996).

The objective of the Metropolis-Hastings algorithm is to generate samples from the target density $f(\mathbf{x}) = \frac{\phi(\mathbf{x})}{a}$, where $\mathbf{x} \in \Re^K$, $\phi(\mathbf{x})$ is the unnormalized density, and a is the normalizing constant, possibly unknown. The Metropolis-Hastings algorithm is an updating scheme, as the Gibbs sampler, but at each step the proposed value is accepted with specified probability. Given the current vector, state, $\mathbf{x}_I$, a sampled value $\mathbf{y}$ for the next state, $I+1$, is generated from a proposal distribution, $g(.|\mathbf{x}_I)$. The sampled value $\mathbf{y}$ is then accepted with probability

$$\alpha(\mathbf{x}_I, \mathbf{y}) = \min\left\{1, \frac{f(\mathbf{y})g(\mathbf{x}_I|\mathbf{y})}{f(\mathbf{x}_I)g(\mathbf{y}|\mathbf{x}_I)}\right\}. \tag{6.5}$$

Otherwise, the sampled value is rejected and the process remains at $\mathbf{x}_I$. The choice of the proposal distribution, $g(.|\mathbf{x}_I)$, is essentially arbitrary. This proposal

distribution may or may not depend on the current vector $\mathbf{x}_I$. If $g(.|\mathbf{x}_I)$ satisfies the condition

$$f(\mathbf{x}) \times g(\mathbf{y}|\mathbf{x}) = f(\mathbf{y}) \times g(\mathbf{x}|\mathbf{y})$$

or if $f(\mathbf{x}) = g(\mathbf{x}|\mathbf{y})$, then we automatically accept every proposed new value. This is clear from equation (6.5). Thus, the Gibbs sampler can be viewed as a special case of the Metropolis-Hastings method where we accept every proposed new value.

6.2.3 Generation of $\mathbf{p}^t$ from $pr(\mathbf{p}^t|\mathbf{n}^t, \boldsymbol{\alpha}, \lambda)$

The Gibbs sampler will be applied for generating the $\mathbf{p}^t$'s from $pr(\mathbf{p}^t|\mathbf{n}^t, \boldsymbol{\alpha}, \lambda)$. The Gibbs sampling method will sample iteratively from $pr(\mathbf{p}^t|\mathbf{n}^t, \boldsymbol{\alpha}, \lambda)$, $pr(\boldsymbol{\alpha}|\mathbf{p}^t, \lambda)$, and $pr(\lambda|\mathbf{p}^t, \boldsymbol{\alpha})$, i.e. the I^{th} iteration will be

$$\begin{aligned} \mathbf{p}_I^t &\sim pr(\mathbf{p}^t|\mathbf{n}^t, \boldsymbol{\alpha}_I, \lambda_I) \\ \boldsymbol{\alpha}_{I+1} &\sim pr(\boldsymbol{\alpha}|\mathbf{n}^t, \mathbf{p}_I^t, \lambda_I) \\ \lambda_{I+1} &\sim pr(\lambda|\mathbf{n}^t, \mathbf{p}_I^t, \boldsymbol{\alpha}_{I+1}). \end{aligned} \tag{6.6}$$

This iteration scheme will produces what is called a Gibbs sequence

$$\boldsymbol{\alpha}_0, \lambda_0, \mathbf{p}_0^t, \boldsymbol{\alpha}_1, \lambda_1, \mathbf{p}_1^t, \ldots$$

with the property that, for large I, $\mathbf{p}_I^t$ is effectively a sample point from $pr(\mathbf{p}^t|\mathbf{n}^t, \boldsymbol{\alpha}, \lambda)$ (Gelfand and Smith, 1990, and Casella and George, 1992). These conditional distributions can be driven from equation (6.2) as

$$\begin{aligned} pr(\mathbf{p}^t|\mathbf{n}^t, \boldsymbol{\alpha}, \lambda) &\propto \prod_{t=1}^{c} \prod_{i=1}^{K} p_{ti}^{n_{ti}+\lambda\alpha_i-1} \\ &= \text{Dirichlet}\left(\mathbf{n}^t + \lambda\boldsymbol{\alpha}\right). \end{aligned} \tag{6.7}$$

Thus, the conditional distribution $pr(\mathbf{p}^t|\boldsymbol{\alpha},\lambda)$ is a Dirichlet$(\mathbf{n}^t+\lambda\boldsymbol{\alpha})$. The conditional distribution $pr(\boldsymbol{\alpha}|\mathbf{p}^t,\lambda)$ is

$$\begin{aligned} pr(\boldsymbol{\alpha}|\mathbf{p}^t,\lambda) &\propto \left(\prod_{t=1}^{c}\prod_{i=1}^{K}\frac{p_{ti}^{\lambda\alpha_i}}{\Gamma(\lambda\alpha_i)}\right)\times\left(\prod_{i=1}^{K}\alpha_i^{\beta_i-1}\right) \\ &= \left(\prod_{i=1}^{K}\frac{\alpha_i^{\beta_i-1}}{\Gamma(\lambda\alpha_i)^c}\right)\times\left(\prod_{t=1}^{c}\prod_{i=1}^{K}p_{ti}^{\lambda\alpha_i}\right). \end{aligned} \tag{6.8}$$

For $pr(\lambda|\mathbf{p}^t,\boldsymbol{\alpha})$,

$$pr(\lambda|\mathbf{p}^t,\boldsymbol{\alpha}) \propto \left(\prod_{t=1}^{c}\prod_{i=1}^{K}p_{ti}^{\lambda\alpha_i-1}\frac{\Gamma(\lambda)}{\Gamma(\lambda\alpha_i)}\right)\times pr\,(\lambda)\,. \tag{6.9}$$

Let us assume that λ has a Pareto prior distribution with parameters $\lambda_0=1$ and $\omega=1$, i.e. $pr\,(\lambda)=\frac{1}{\lambda^2}I(\lambda>1)$, which is long tailed and therefore fairly 'non-informative' (Mood *et al.*, 1974). Then,

$$\begin{aligned} pr(\lambda|\mathbf{p}^t,\boldsymbol{\alpha}) &\propto \frac{1}{\lambda^2}\prod_{t=1}^{c}\prod_{i=1}^{K}p_{ti}^{\lambda\alpha_i-1}\frac{\Gamma(\lambda)}{\Gamma(\lambda\alpha_i)} \\ &= \left(\frac{\Gamma(\lambda)}{\prod_{i=1}^{K}\Gamma(\lambda\alpha_i)}\right)^c\left(\frac{1}{\lambda^2}\prod_{t=1}^{c}\prod_{i=1}^{K}p_{ti}^{\lambda\alpha_i}\right). \end{aligned} \tag{6.10}$$

Unfortunately, we can see that the forms of $pr(\boldsymbol{\alpha}|\mathbf{p}^t,\lambda)$ and $pr(\lambda|\mathbf{p}^t,\boldsymbol{\alpha})$ are not as easy to sample from as $pr(\mathbf{p}^t|\boldsymbol{\alpha},\lambda)$. In this case, we will use a hybrid MCMC strategy. A hybrid strategy is a combination of two or more methods. Thus, we will combine the Gibbs sampler with a Metropolis-Hastings algorithm. The Metropolis-Hastings algorithm, can solve the difficulties of dealing with the forms of $pr(\boldsymbol{\alpha}|\mathbf{p}^t,\lambda)$ and $pr(\lambda|\mathbf{p}^t,\boldsymbol{\alpha})$. For convenience and simplicity in the program, we are going to consider $\ln pr(\boldsymbol{\alpha}|\mathbf{p}^t,\lambda), \ln pr(\lambda|\mathbf{p}^t,\boldsymbol{\alpha})$. Thus,

$$\begin{aligned} \ln pr(\boldsymbol{\alpha}|\mathbf{p}^t,\lambda) &\propto \ln\left[\left(\prod_{i=1}^{K}\frac{\alpha_i^{\beta_i-1}}{\Gamma(\lambda\alpha_i)^c}\right)\times\left(\prod_{t=1}^{c}\prod_{i=1}^{K}p_{ti}^{\lambda\alpha_i}\right)\right] \\ &= \sum_{i=1}^{K}\{(\beta_i-1)\ln(\alpha_i)-c\ln(\Gamma(\lambda\alpha_i))\}+\lambda\sum_{t=1}^{c}\sum_{i=1}^{K}\alpha_i\ln(p_{ti}) \\ &= \sum_{i=1}^{K}\left(\{(\beta_i-1)\ln(\alpha_i)-c\ln(\Gamma(\lambda\alpha_i))\}+\lambda\alpha_i\sum_{t=1}^{c}\ln(p_{ti})\right). \end{aligned} \tag{6.11}$$

$$\begin{aligned}
\ln pr(\lambda|\mathbf{p}^t, \boldsymbol{\alpha}) &\propto \ln\left[\left(\frac{\Gamma(\lambda)}{\prod_{i=1}^{K}\Gamma(\lambda\alpha_i)}\right)^c \left(\frac{1}{\lambda^2}\prod_{t=1}^{c}\prod_{i=1}^{K} p_{ti}^{\lambda\alpha_i}\right)\right] \\
&= \ln\left(\frac{\Gamma(\lambda)^c}{\lambda^2}\right) - c\sum_{i=1}^{K}\ln\left(\Gamma(\lambda\alpha_i)\right) + \lambda\sum_{t=1}^{c}\sum_{i=1}^{K}\alpha_i \ln\left(p_{ti}\right) \\
&= \ln\left(\frac{\Gamma(\lambda)^c}{\lambda^2}\right) + \sum_{i=1}^{K}\left(\lambda\alpha_i\sum_{t=1}^{c}\ln\left(p_{ti}\right) - c\ln\left(\Gamma(\lambda\alpha_i)\right)\right). \quad (6.12)
\end{aligned}$$

For the proposal distribution for $\boldsymbol{\alpha}$, we can use a Dirichlet $(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ may or may not depend on the current value of $\boldsymbol{\alpha}$,

$$\begin{aligned}
g(\boldsymbol{\alpha}) &= \frac{\Gamma(\sum_{i=1}^{K}\theta_i)}{\prod_{i=1}^{K}\Gamma(\theta_i)}\prod_{i=1}^{K}\alpha_i^{\theta_i-1} \quad (6.13) \\
\ln g(\boldsymbol{\alpha}) &= \ln(\Gamma(\sum_{i=1}^{K}\theta_i)) + \sum_{i=1}^{K}\{(\theta_i-1)\ln(\alpha_i) - \ln(\Gamma(\theta_i))\}
\end{aligned}$$

Also, for λ, we use Uniform $(\lambda+u, \lambda-u)$, where u is a constant to be specified. The ratio of the proposal densities does not depend on λ, thus it will cancel out in the probability of acceptance, $\alpha(\lambda_I, \lambda_{I+1})$, in equation (6.5).

6.3 MCMC analysis

We have described MCMC simulation algorithms, which generate the random sample $\mathbf{p}^t$ from $pr(\mathbf{p}^t|\mathbf{n}^t, \boldsymbol{\alpha}, \lambda)$. The $\mathbf{p}^t$ generated using a Gibbs sampler where the updating of both parameters λ and $\boldsymbol{\alpha}$ uses the Metropolis-Hastings algorithm.

In this MCMC analysis, we will discuss the results of our MCMC simulations, for three different sizes of the primary sampling unit, psu, c. Finally, we consider conclusions based on our MCMC simulations.

6.3.1 Simulation study 1

As discussed above, we used a hybrid MCMC strategy, which consists of a combination of two algorithms, the Gibbs sampler and the Metropolis-Hastings algorithm. In this simulation, indeed for all our simulations, we generate 1000 samples. In each, we have a sample size of $n = 1000$ observations, selected with probability proportional to size from c psus. The number of psus, c, varies between 10, 50 and 200, with equal numbers of units in each psu. For the MCMC we consider 1500 iterations for each sample.

In the real data, the observations $\mathbf{n}^t$ are given. For the proposed simulation study, we assume $K = 3$, and we simulate a cluster population, i.e. $\mathbf{p}^t$, for the psu t, from Dirichlet($\lambda\boldsymbol{\alpha}$), assuming the values $\boldsymbol{\alpha} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$, and $\lambda = 1000$. Then, we simulate $\mathbf{n}^t$ from Multinomial($n_t, \mathbf{p}^t$), where $n_t = \frac{n}{c}$. We chose $\lambda = 1000$ to give a design effect, $\hat{\tau}_.$, almost equivalent to the design effect of the large empirical study, of two United Kingdom surveys, by Holt, Scott and Ewings (1980). The following table presents the average design effect for different values of λ, where $\boldsymbol{\alpha} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$,

λ	10	30	60	100	150	300	1000
$\hat{\tau}_.$	10.1	4.7	3.36	2.8	2.5	2.3	2.1

6.3.2 The MCMC program algorithms

For each sample, we simulate $\mathbf{n}^t$ from Multinomial($n_t, \mathbf{p}^t$), where $\mathbf{p}^t$ is generated from Dirichlet($\lambda\boldsymbol{\alpha}$). Then,

1) Start iteration $I = 0$.

2) Consider the initial values of $\boldsymbol{\alpha}_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$ and $\lambda_0 = 800$. Then generate a set of random numbers $\mathbf{p}^t$, where $t = 1, \ldots, c$, from Dirichlet($\mathbf{n}^t + \lambda_I \boldsymbol{\alpha}_I$)

equation(6.7), using a Gibbs sampler.

3) Generate a set of random numbers $\boldsymbol{\alpha}_{I+1}$ from the proposal distribution, Dirichlet($\boldsymbol{\theta}$), where $\boldsymbol{\theta} = (\frac{1}{3} + n_1, \frac{1}{3} + n_2, \frac{1}{3} + n_3)'$ for the first simulation and $\boldsymbol{\theta} = \gamma \boldsymbol{\alpha}_I$ in the second simulation where γ is a real number.

4) Compute the acceptance probability, for this first M-H.

5) Update $\boldsymbol{\alpha}_I$ to $\boldsymbol{\alpha}_{I+1}$, if the value of $\boldsymbol{\alpha}_{I+1}$ is accepted. Otherwise set $\boldsymbol{\alpha}_{I+1} = \boldsymbol{\alpha}_I$.

6) Generate random number λ_{I+1} from the proposal distribution, Uniform $(\lambda_I - u, \lambda_I + u)$ where $u = 500$ to allow large jumps in the values of λ for the second step of M-H.

7) Compute the acceptance probability, for this second M-H.

8) Update λ_I to λ_{I+1}, if the value of λ_{I+1} is accepted. Otherwise set $\lambda_{I+1} = \lambda_I$.

9) Finish one iteration, see (6.6). Update the iteration $I = I + 1$, then go to step (2).

10) After a fixed number of iterations stop, discard the burn-in iterations ($z = 500$ here) and consider the remaining iterations as a dependent sample from the marginal posterior distribution.

6.3.3 Results 1

To illustrate the MCMC results, we select one sample size of a 1000 observations. The MCMC output will be presented by plotting the sampled values for λ, one α_i and one p_i where $i = 1, 2, 3$. Then, we will discuss the plots.

For 10 primary sampling units, psus, there are 100 observations in each. The sample we selected had marginal cell totals of $\mathbf{n} = (314, 328, 358)'$. In figure (6.1-a), the sampled values for λ, using M-H algorithm, are not stable and support

some high values of λ. For the first cell, where n_1 is equal to 314, we would expect the values of α_1 and p_1 to be distributed around 0.314. For the sampled values of α_1, using M-H algorithm, the sample quickly settled down. Thus, the values of α_1 are almost stable, around a target distribution, see figure (6.1-b). Nevertheless, the algorithm is not mixing well. It is rejecting many proposed values. The aim of this simulation is to generate a set of random numbers $\mathbf{p}^t$ from it's posterior distribution $pr(\mathbf{p}^t|\mathbf{n}^t, \boldsymbol{\alpha}, \lambda)$. After applying the constraint in equation (4.21) for computing values of $\mathbf{q}$, we can see from figure (6.1-c) that p_1 seems settled down quickly to a stationary distribution.

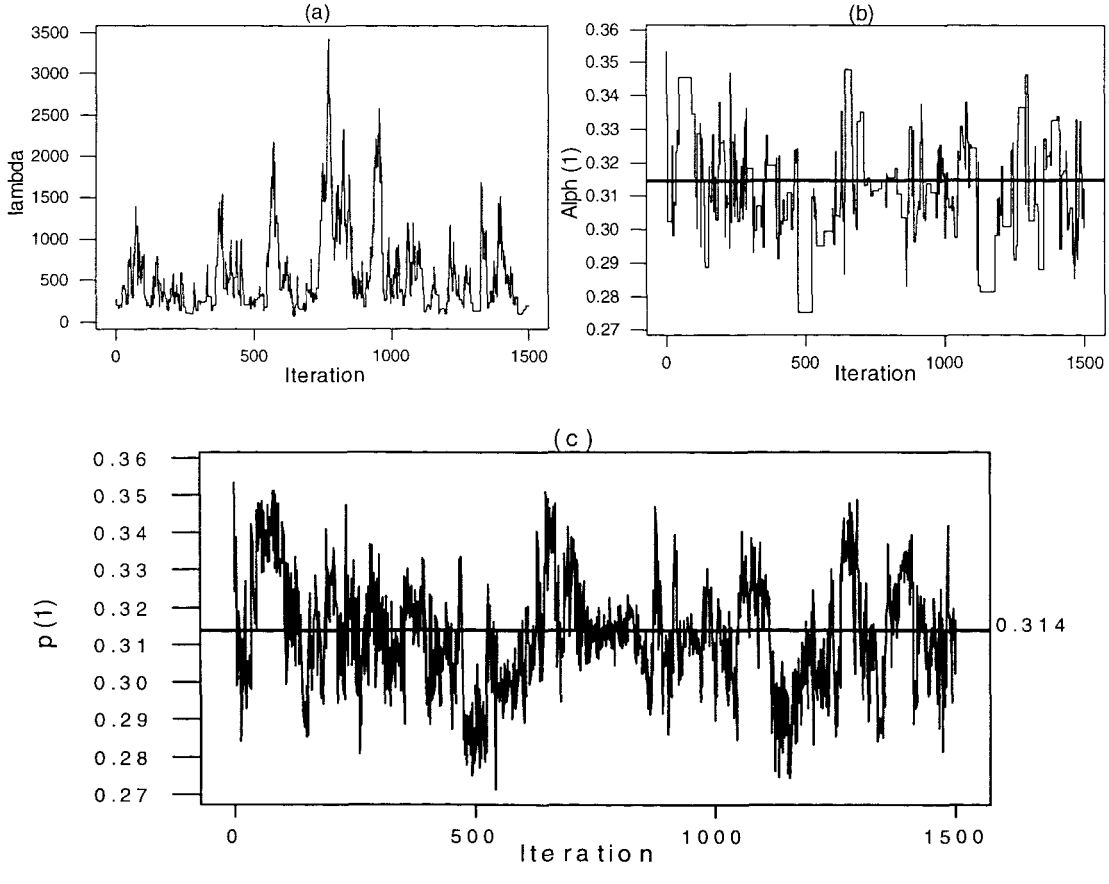

Figure 6.1: Results of 1500 Iterations using a hybrid MCMC strategy, Metropolis-Hastings within Gibbs sampler, based on three components λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)); for sample size of 1000 in 10 psu's: (a) and (b) are sampled values for λ and α_1 using Metropolis-Hastings algorithm. (c) sampled values for p_1 using Gibbs sampler.

For the case where there are 50 primary sampling units, each will have 20 observations. We selected an arbitrary sample. The sample has marginal cell totals of $\mathbf{n} = (329, 341, 330)'$. As in the previous case, the sampled values for λ, figure (6.2-a), are not stable and with more smaller values of λ. Again, the method often rejects for smaller values of λ. If we choose the first cell, where n_1 is equal to 329, then we expect the values of α_1 and p_1 to be distributed around 0.329.

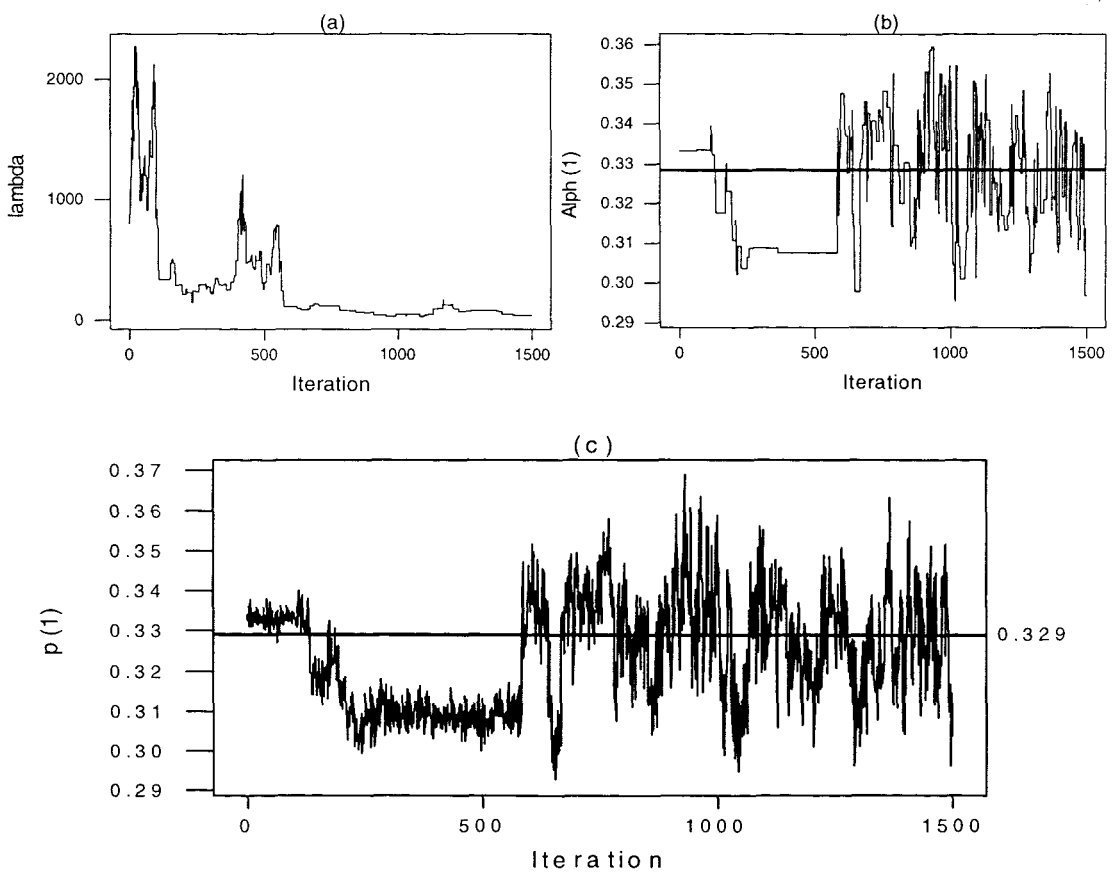

Figure 6.2: Results of 1500 Iterations using a hybrid MCMC strategy, Metropolis-Hastings within Gibbs sampler, based on three components λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)); for sample size of 1000 in 50 psu's: (a) and (b) are sampled values for λ and α_1 using Metropolis-Hastings algorithm. (c) sampled values for p_1 using Gibbs sampler.

For the sampled values of α_1, using M-H algorithm, the sampled values settled

down after 600 iterations (got stuck at one point for several iterations). Thus, the values of α_1 are almost stationary around a target distribution, see figure (6.2-b). The effect of the values α_1 on the sampled values of p_1 is visible in figure (6.2-c).

When we have 200 primary sampling units, each psu has only 5 observations, with $\mathbf{n} = (311, 364, 325)'$. Unlike the previous cases, the sampled values for λ in figure (6.3-a) are small, if compared with $\lambda = 1000$.

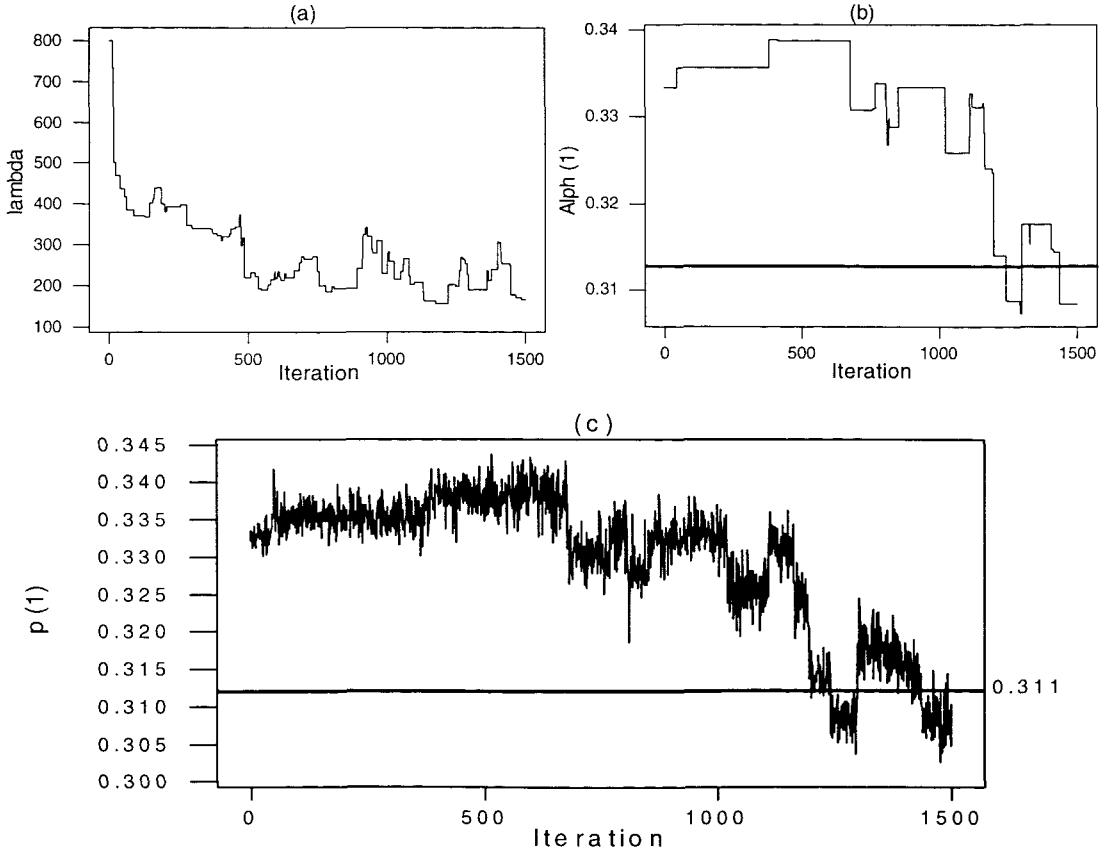

Figure 6.3: Results of 1500 Iterations using a hybrid MCMC strategy, Metropolis-Hastings within Gibbs sampler, based on three components λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)); for sample size of 1000 in 200 psu's: (a) and (b) are sampled values for λ and α_1 using Metropolis-Hastings algorithm. (c) sampled values for p_1 using Gibbs sampler.

We can see in figure (6.3-b) that α_1 has few jumps. In both cases, we can see

that our MCMC algorithm is not working properly. Since, the M-H algorithm is mixing slowly, i.e. the algorithm almost always rejects the sample values of λ and α_1. This may be caused by the small numbers of observations in each psu, 5 observations only, or the large number of psus. The effect of poor application of M-H in the sample values of p_1 can be seen in figure (6.3-c). The sample values of p_1 have been effected strongly by the sampled values of λ and α_1.

6.3.4 Discussion 1

The sampled values of λ, in all cases, represent the uncertainty about the knowledge of the distribution of λ. In both cases, where the number of psus are equal to 10 and 50, the values of λ capture the true value, $\lambda = 1000$, in it's domain. This may indicate that the chain is moving slowly toward the target distribution. Thus, the need is for a longer run for λ to converge to its posterior distribution $pr(\lambda)$. If the M-H assumptions are satisfied with a reasonable proposal distribution, we should see the sampled values of λ distributed according to its stationary distribution. For more detailed discussion of this see Gilks *et al.* (1996).

Unfortunately, when the number of psus is 200, the sampled values of λ are small. This may indicate great uncertainty about the distribution of λ due to the small numbers of observations in each psu, n_t, or large number of psus, c.

The first suggestion is supported by figure (6.4-a). The M-H algorithm captures the target distribution of λ directly when we increase the sample size to 200000, 1000 observations in each psu. The same is true of the sampled values of α_1. In the case when there are 200 psus, for small n_t, the sampled values of α_1 are almost always rejected, see figure (6.3-b). Thus the M-H algorithm is mixing badly, since the accepted sampled values of α_1 are very small relative to the number of iterations.

Figure (6.4-b), again, shows the validity of the first suggestion. Thus, we may say that the small numbers of observation in each psu caused the slow mixing of the algorithm. A solution of this kind of problem is to change the proposal distribution for $pr(\boldsymbol{\alpha}|\mathbf{p}^t, \lambda)$; see simulation 2.

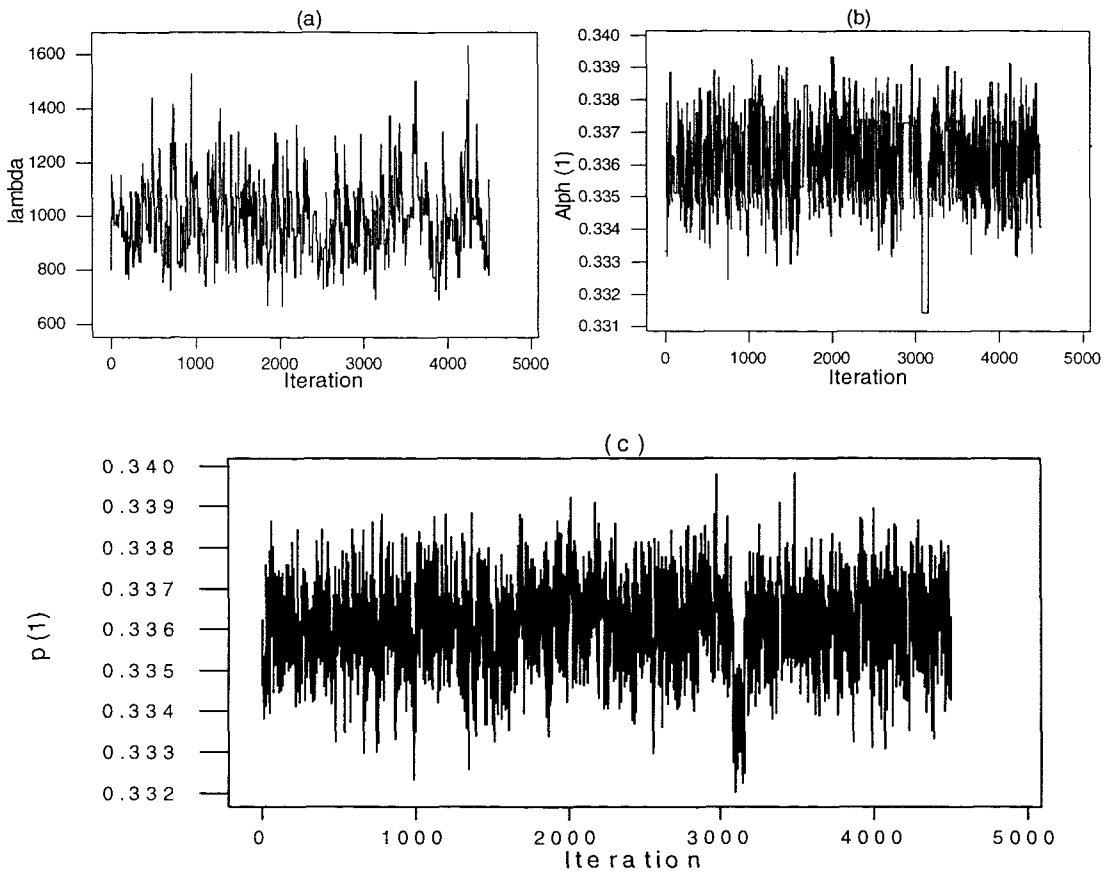

Figure 6.4: Results of 4500 Iterations based on 200000 observation in 200 psu's. For three components λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)). (a) and (b) are sampled values for λ and α_1 using Metropolis-Hastings algorithm. (c) sampled values for p_1 using Gibbs sampler.

For the sample values of $\mathbf{q}$, we can see that they are highly correlated with $\boldsymbol{\alpha}$ in all cases. In general, we can conclude that, when n_t is large, the behaviour of MCMC process are similar in all cases. In these cases, the sample values of $\mathbf{q}$ converge to the target distribution, that is the posterior distribution of $\mathbf{q}$, directly.

6.3.5 Simulation study 2

In this simulation, we try to overcome the weaknesses of our MCMC application in the first simulation, which are slow mixing and the need for a large sample size. This can be done by changing the M-H algorithm that generates the sample values of $\boldsymbol{\alpha}$ from an independence M-H to a M-H algorithm with proposal centered at the current realisation. Thus, we are going to change the proposal distribution of $\boldsymbol{\alpha}$, equation (6.13), from Dirichlet($\boldsymbol{\theta}$) where $\boldsymbol{\theta} = (\frac{1}{3}+n_1, \frac{1}{3}+n_2, \frac{1}{3}+n_3)'$ to $\boldsymbol{\theta}_I = \gamma\boldsymbol{\alpha}_{I-1}$, where I stands for iteration index and γ is a positive real number.

6.3.6 Results 2

Before we start this simulation we have to determine a rough estimate for the value of γ. The value of γ is important in this simulation, since it is a determinant of the size of variation in the proposal distribution, roughly speaking, similar to the value n_i, $i = 1, \ldots, K$, in the previous proposal.

From our research we found out that, first, if the value of γ is small, say $\gamma = 100$, then the M-H algorithm for λ is mixing well, and supporting large values of λ. On the other hand, the M-H algorithm for $\boldsymbol{\alpha}$ is mixing slowly, see figure (6.5). Second, if the value of γ is large, say $\gamma = 50000$, then the M-H algorithm for λ is not stable, not working properly and rejecting many values of λ, while, for $\boldsymbol{\alpha}$ the process is moving very fast with small jumps, which is related to the very small variation in the proposal distribution. Thus, we have a slow mixing chain, see figure (6.6).

The dependence on λ and $\boldsymbol{\alpha}$ of the sample values of $\mathbf{p}$, through $\mathbf{p}^t$, are related to the sample size; see the marginal posterior distribution $pr(\mathbf{p}^t|\boldsymbol{\alpha}, \lambda)$ equation (6.7).

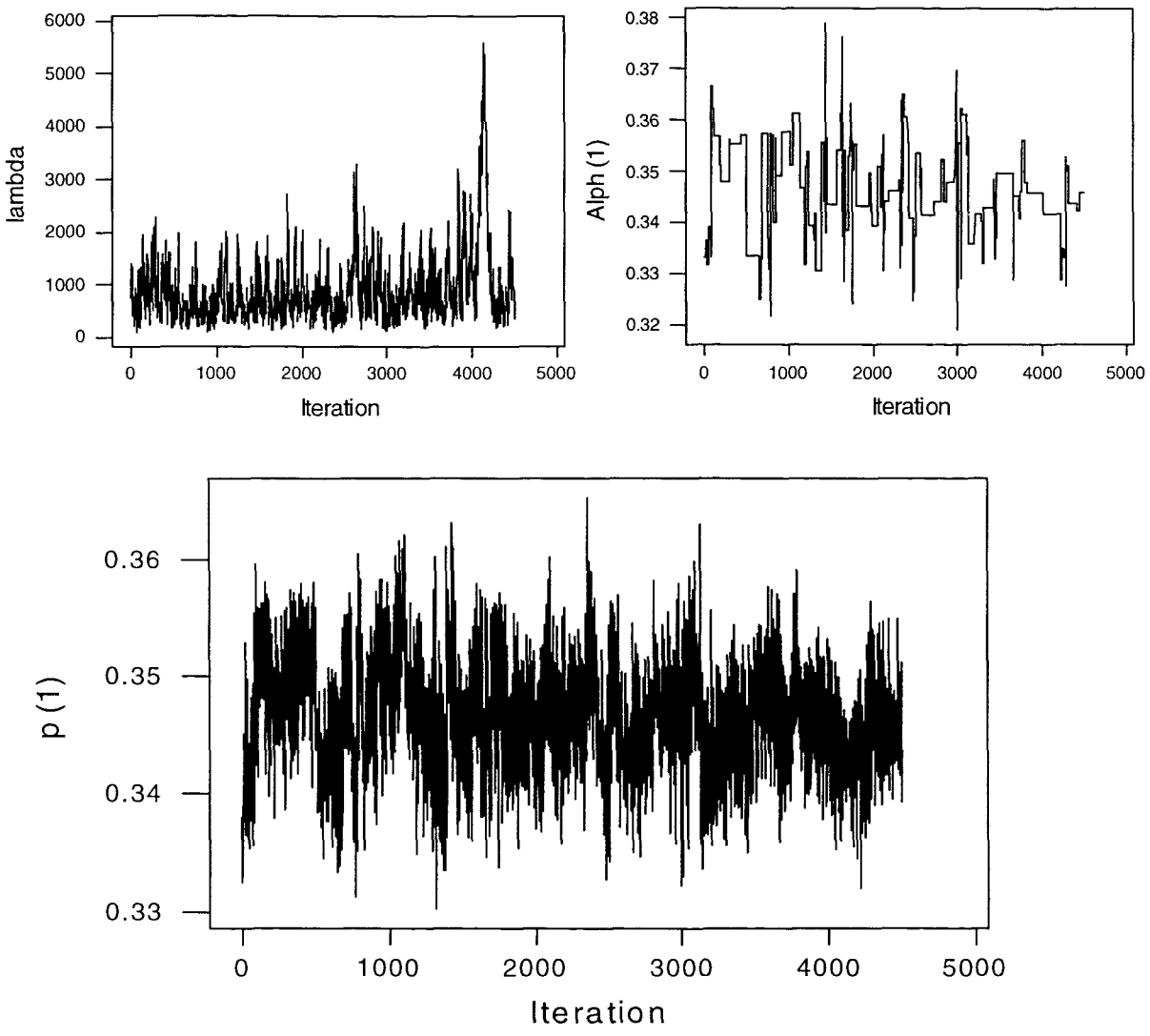

Figure 6.5: Results of the second MCMC simulation , with 4500 Iterations. It is based on γ =100 and sample size of 10000 observation in 10 psu's; For three parameters λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)).

When the number of observations in each psu is small, say $n_t = 5$, the correlation is strong. Thus, the Gibbs sampler will be affected and will not converge to the target distribution in a reasonable number of iterations. On the other hand, if the sample size is large, say $n_t = 1000$, the sample size gives greater support to the sample estimate of $\mathbf{p}$ as in figures (6.5) and (6.6). The mixing of λ and $\boldsymbol{\alpha}$ is also effected by the sample size.

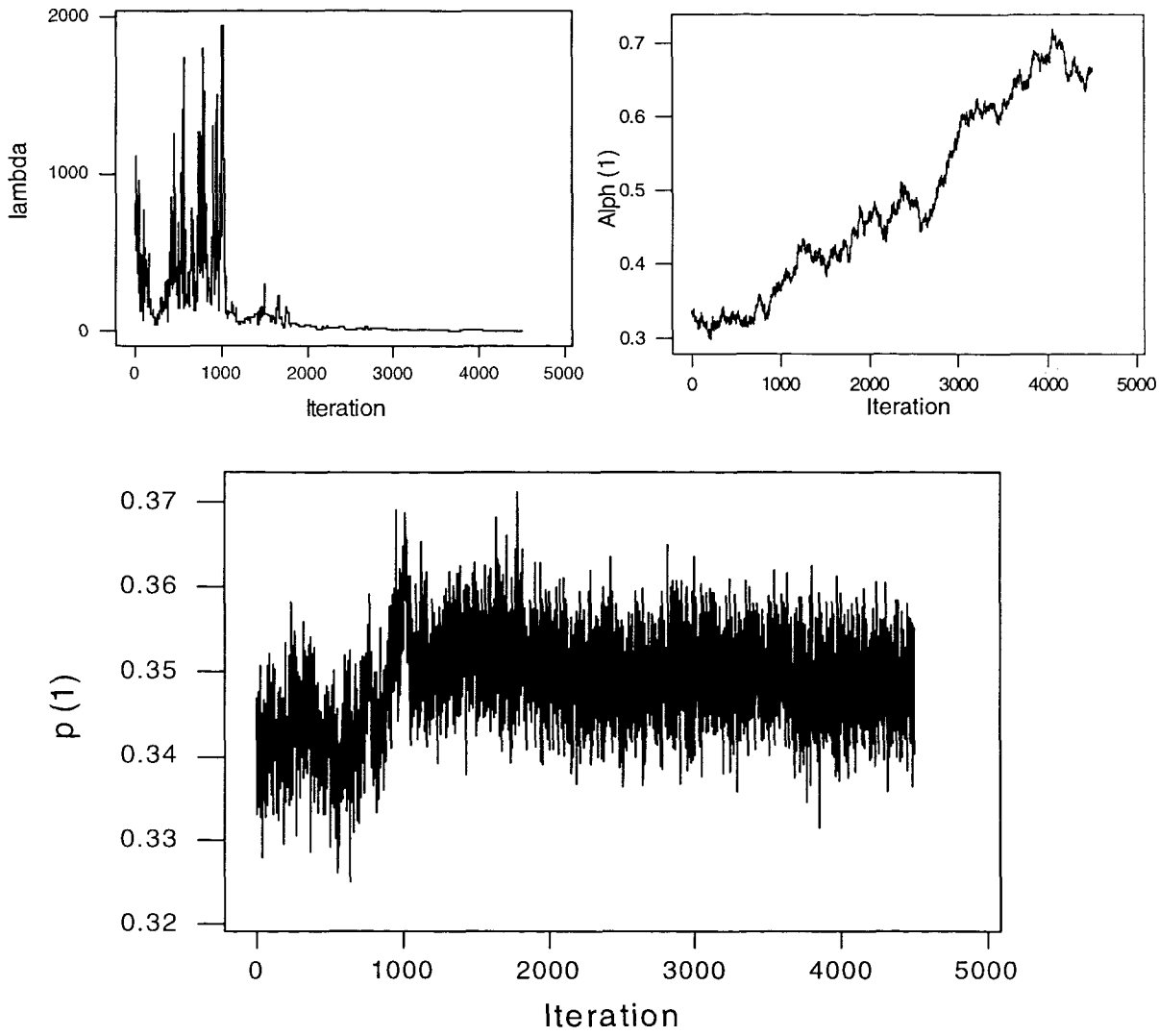

Figure 6.6: Results of the second MCMC simulation , with 4500 Iterations. It is based on γ =50000 and sample size of 10000 observation in 10 psu's; For three components λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)).

This search for a value for γ, leads us to a simple, ad-hoc, rule which seems to work reasonably in our cases, that is

$$\gamma = 50000\left(\frac{c}{200}\right) \tag{6.14}$$

where c refers to the number of psu. The results of our second simulation using equation (6.14) are considered for three situations, all with different psus, as in

simulation one. The number of psus, c, varies between 10, 50 and 200, with equal units in each psu. The first situation considered is a sample size of 1000 units, i.e. if $c = 10$ psus then $n_t = 100$, or 50 psus then $n_t = 20$, and if $c = 200$ psus then $n_t = 5$.

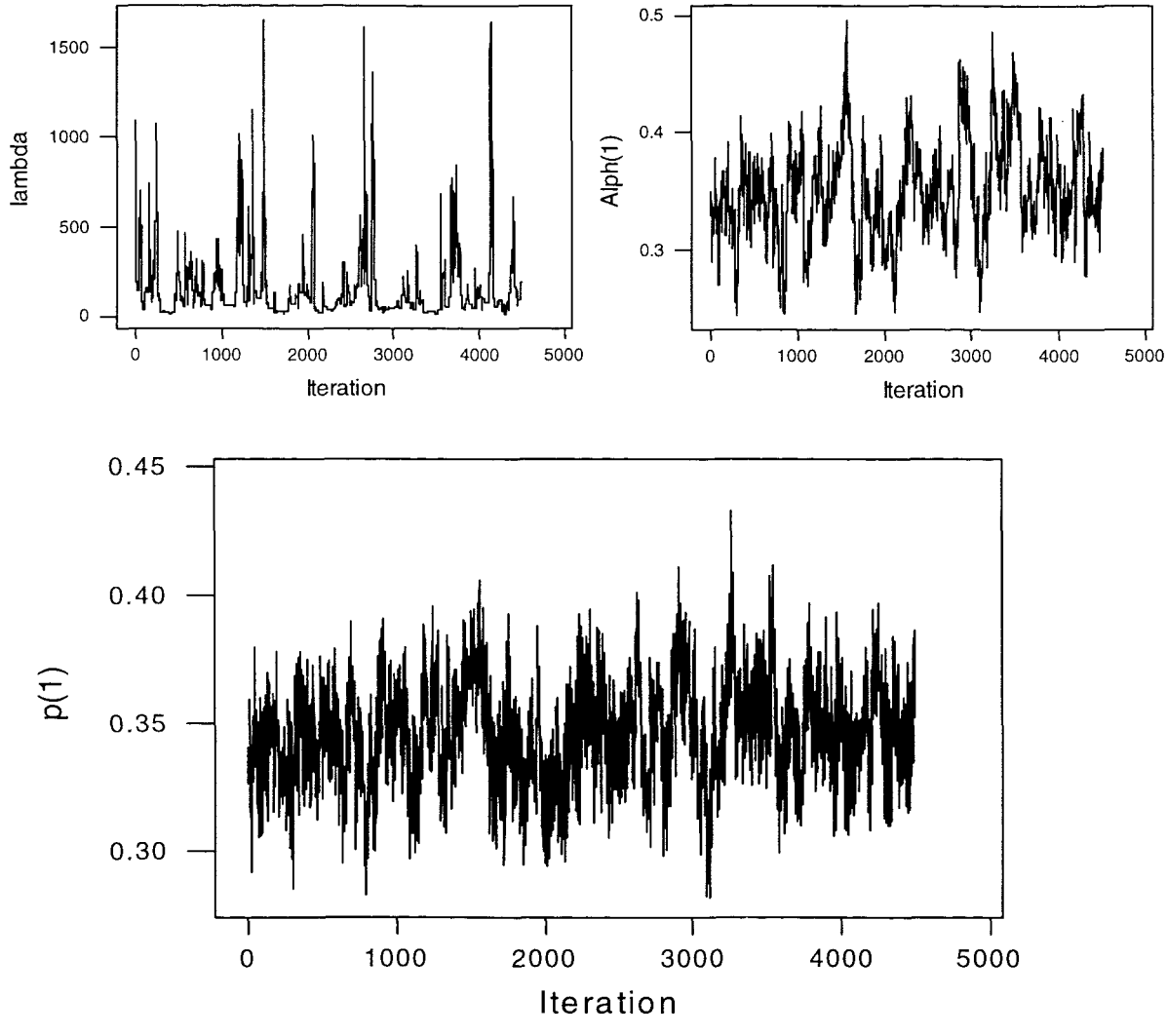

Figure 6.7: Results of the second MCMC simulation , with 4500 Iterations. It is based on 1000 observation in 10 psu's; For three components λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)).

Second, we considered all three cases where the psus are of size 100, i.e. $n_t = 100$, and for the third situation we increase the number of units in all three

cases of the psus to 1000, i.e. $n_t = 1000$. For the sample size of 1000, the first case has 10 psus, with $n_t = 100$. Figure (6.7) shows that the sampled values of λ almost capture the true value, $\lambda = 1000$, in it's domain. Also, the sampled values of α_1 are mixing reasonably in the M-H algorithm. For λ, it supports smaller values, and for α_1, it is mixing a lot better compared with the first simulation.

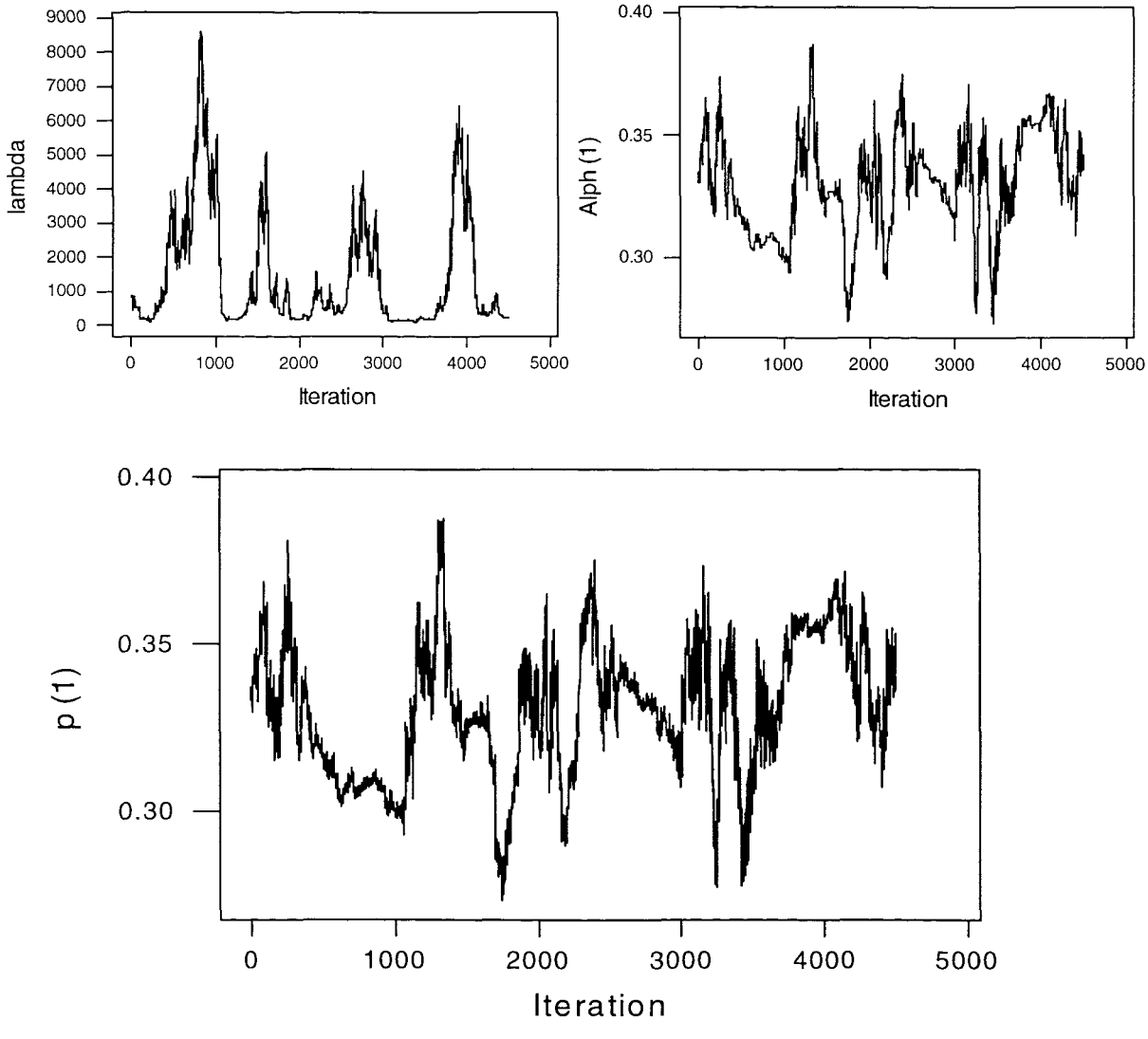

Figure 6.8: Results of the second MCMC simulation , with 4500 Iterations. It is based on 1000 observation in 50 psu's; For three components λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)).

As the number of observations in each psu, n_t, gets smaller, such as the case with 50 psus, figure (6.8), the mixing process start to become slower for α_1. The distribution of λ supports larger values.

In figure (6.9), where the number of psus is 200 and $n_t = 5$, it is clear that the sampled values of λ and α_1 are mixing slowly. In addition, the sample values of $\boldsymbol{\alpha}$ and $\mathbf{p}$, through $\mathbf{p}^t$, are very highly correlated.

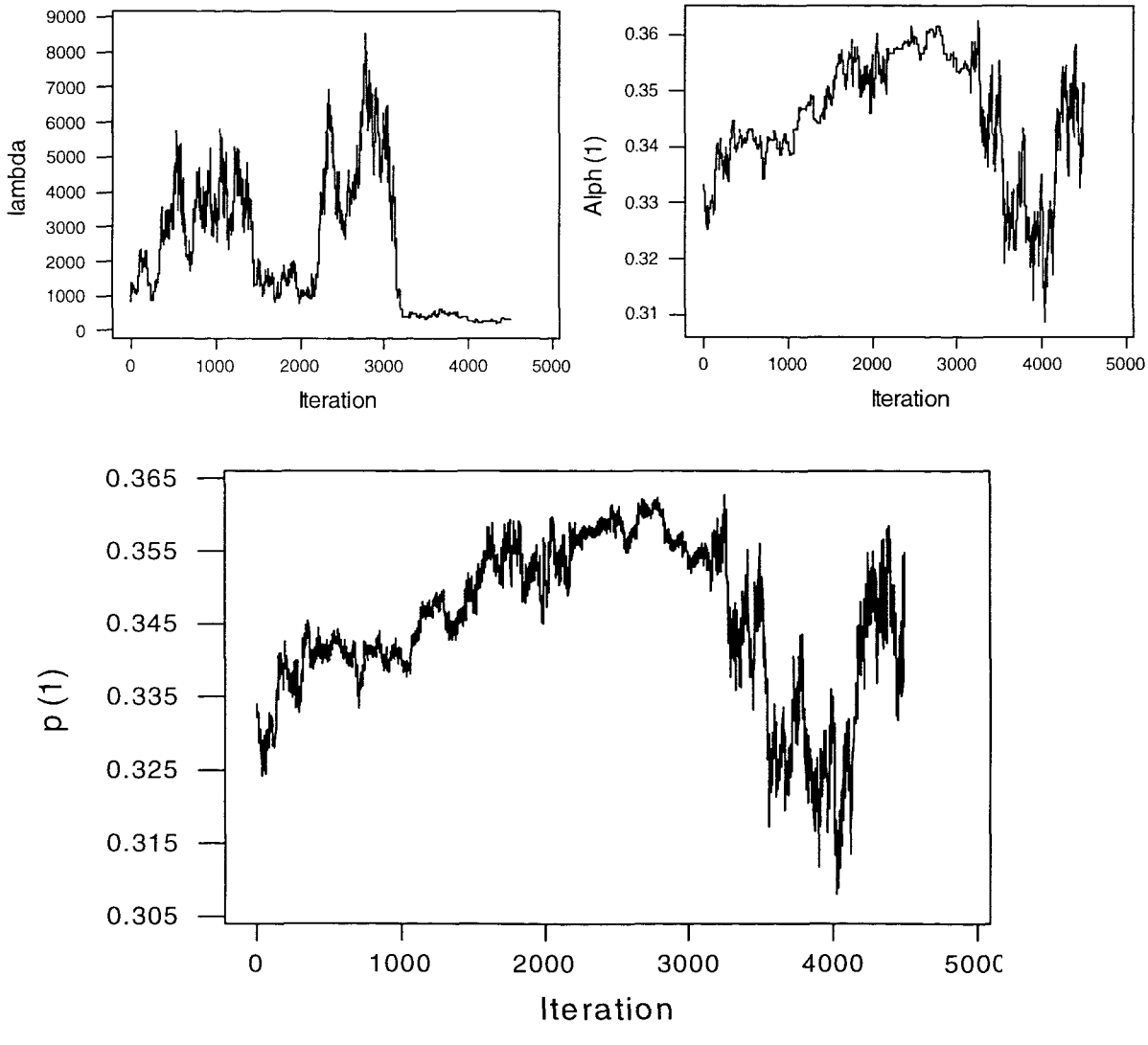

Figure 6.9: Results of 4500 Iterations based on 1000 observation in 200 psu's; For three components λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)).

If we increase the number of observation in each psu, n_t, to 100 for the cases of 50, figure (6.10), and 200 psus, figure (6.11), then we have similar behaviour of the MCMC algorithm as figure (6.7), with 10 psus. Actually, the mixing process for the sample values of λ is better.

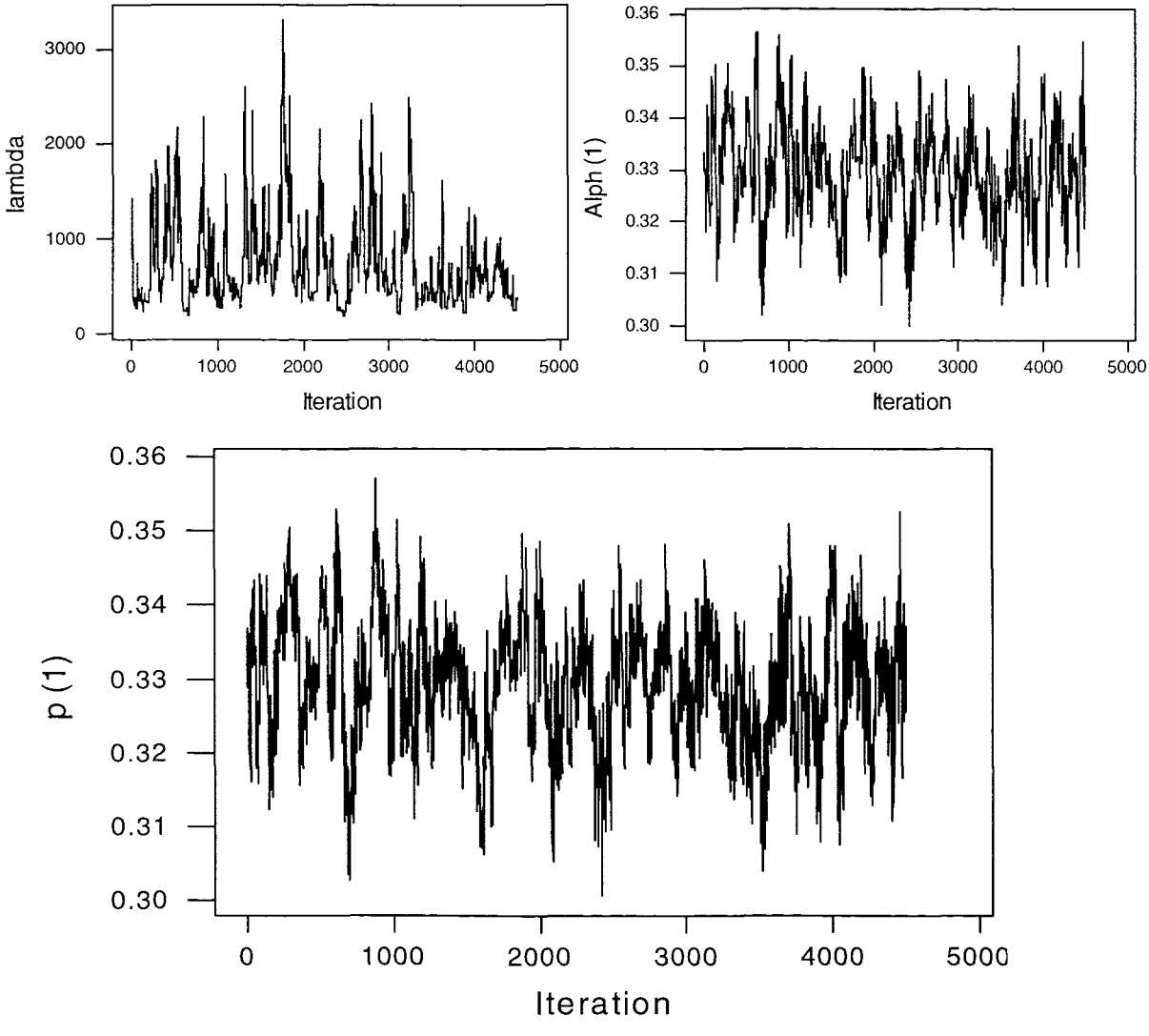

Figure 6.10: Results of the second MCMC simulation , with 4500 Iterations based on 5000 observation in 50 psu's, $n_t = 100$; For three components λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)).

As a final illustration of the MCMC algorithm we increase the number of observations in each psu, n_t, to 1000 for all cases. In all cases, the behaviour of

the MCMC process is similar.

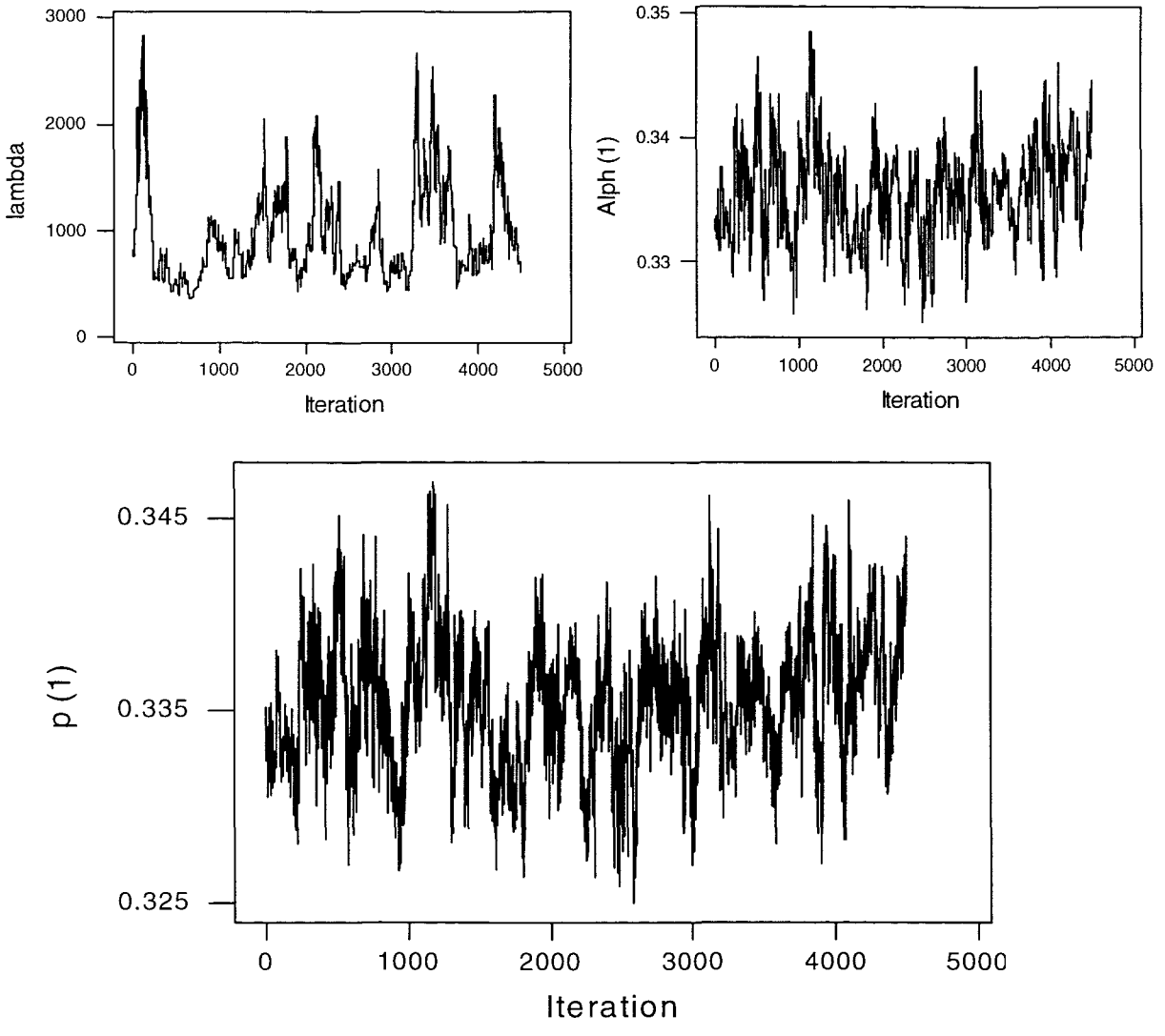

Figure 6.11: Results of the second MCMC simulation , with 4500 Iterations. It is based on 20000 observation in 200 psu's, $n_t = 100$; For three components λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)).

The MCMC algorithms are mixing well for λ, $\boldsymbol{\alpha}$, and also for $\mathbf{p}$. In addition, they all converge to their target distribution directly. The results are represented for the case with 50 psus and $n_t = 1000$ in figure (6.12).

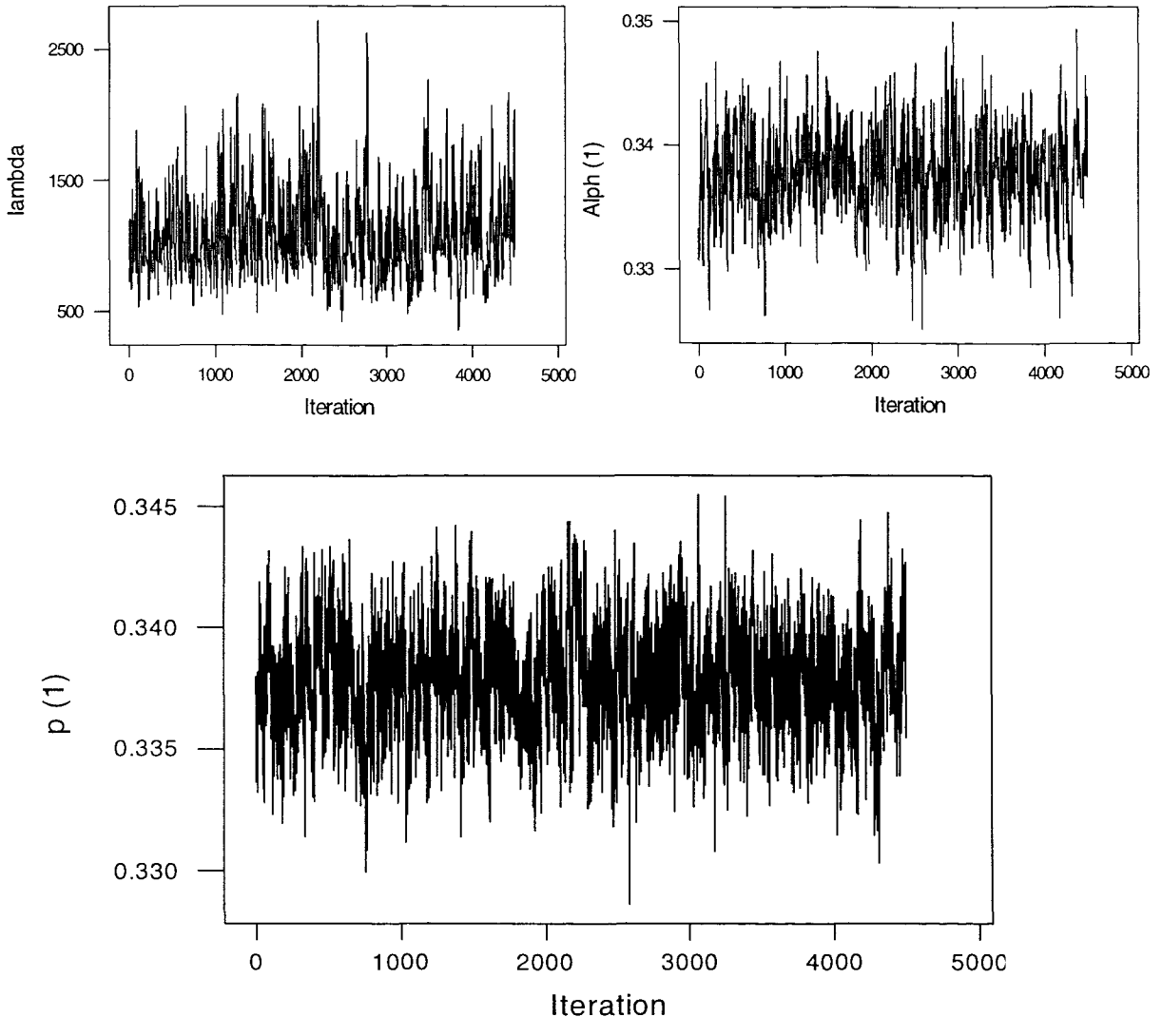

Figure 6.12: Results of the second MCMC simulation , with 4500 Iterations. It is based on 50000 observation in 50 psu's, $n_t = 1000$; For three components λ (Lambda), α_1 (Alph(1)) and p_1 (p(1)).

6.3.7 Discussion 2

We have illustrated the effect of different inputs of the sample size and the value of γ in our simulation. It appears that the major factor that affects our MCMC simulation is the number of observations in each psu, n_t. If it is small then the

high correlation of λ and $\boldsymbol{\alpha}$ affects the sample values of $\mathbf{p}^t$, generated using the Gibbs sampler. Also, due to poor mixing $\mathbf{q}$ does not converge to its posterior distribution. Then, the distribution of $\mathbf{q}$ is a multimodal distribution as in figure (6.13). On the other hand, if n_t is large enough then, we have a moderate

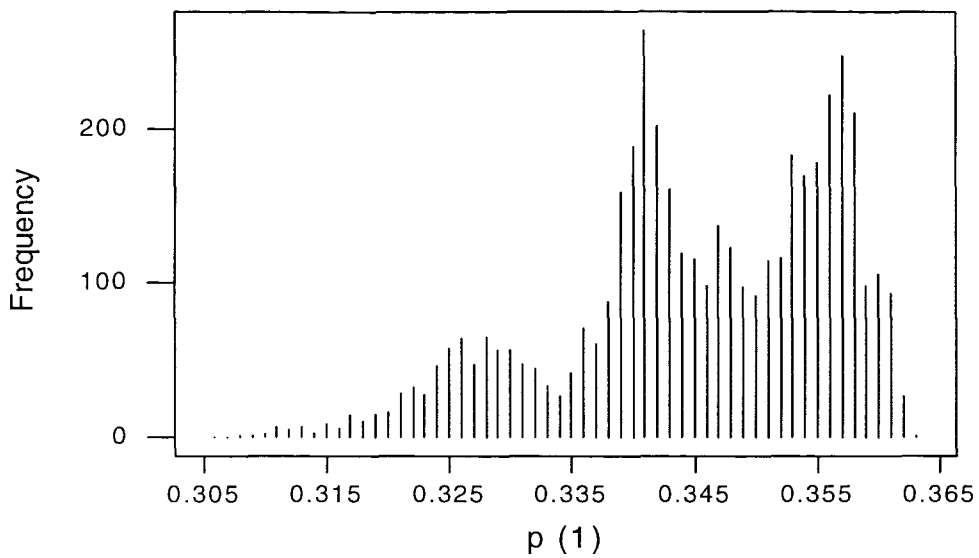

Figure 6.13: A histogram for the sample values of p_1 (p(1)) using the second MCMC simulation, with sample size of 1000 and 200 psu, $n_t = 5$.

correlation. The distribution of $\mathbf{q}$ is approximately normal as in figure (6.14). From this MCMC simulation, the behaviour of the process is heavily dependent on the number of observations in each psu, n_t. When, n_t is small the chain of λ or $\boldsymbol{\alpha}$ is mixing slowly. The effect of γ on the mixing is strong. Thus, the Gibbs sampler is affected and may well not converge to the target distribution within a reasonable number of runs as in figure (6.9). On the other hand, if n_t is large enough, then the posterior distribution supports the sample values of $\mathbf{q}$ as in figures (6.5) and (6.6). Eventually, the effect of γ on $\mathbf{p}^t$ weakens and the sample values of $\mathbf{p}^t$ converge to the target distribution, that is the posterior distribution $pr(\mathbf{p}^t|\mathbf{n}^t, \boldsymbol{\alpha}, \lambda)$, after burn-in.

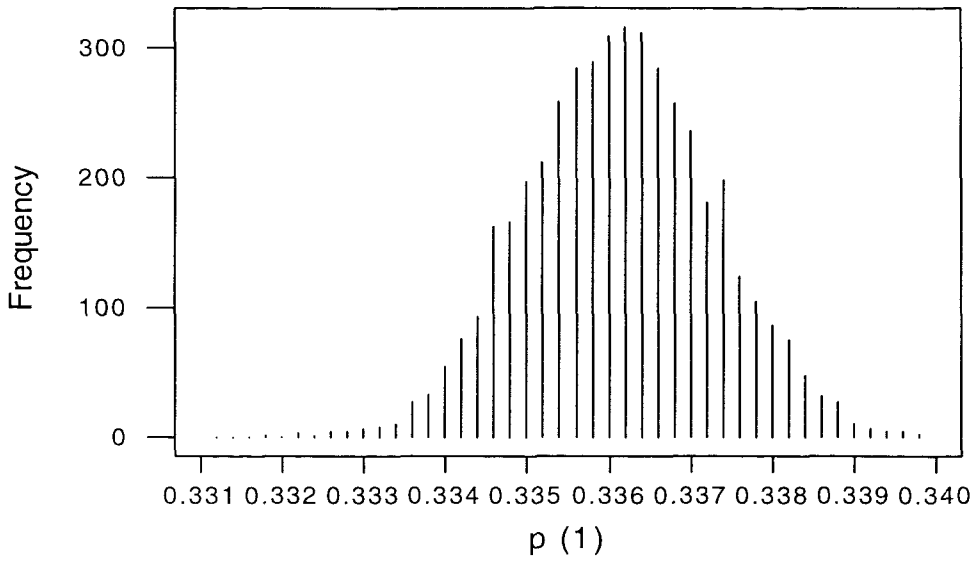

Figure 6.14: A histogram for the sample values of p_1 (p(1)) using the second MCMC simulation, with sample size of 200000 and 200 psu, $n_t = 1000$.

6.3.8 Conclusion

We have illustrated how the Bayesian approach for generating from a complex Bayesian posterior distribution, $pr(\mathbf{p}^t|\mathbf{n}^t, \boldsymbol{\alpha}, \lambda)$, is performed via a MCMC algorithm. We have discussed one way of using a hybrid MCMC strategy, which consists of a combination of two algorithms, the Gibbs sampler and the Metropolis-Hastings algorithm.

Finally, we have shown one strategy for improving the MCMC algorithm which enables it to converge to the target distribution quickly, with large sample sizes. For small sample sizes, the need for more iterations is obvious even in the second strategy. Unfortunately, if the computer program has complex computations with long running times, such as in our case, this will not be a plausible solution.

Now, after sampling from the marginal posterior density $pr(\mathbf{p}|\mathbf{n},M_S)$, and the prior density $pr(\mathbf{p}|M_S)$, we have to estimate them. We are going to use a multi-

variate density estimator to estimate the marginal posterior density $pr(\mathbf{p}|\mathbf{n},M_S)$, and the prior density $pr(\mathbf{p}|M_S)$, at point $\mathbf{p}_0$, in order to apply the Savage-Dickey density ratio. Two estimators will be used; a crude density estimator, section (5.5.3), and the multivariate normal kernel density estimator, section (5.5.1).

6.4 The simulation study

For computing the approximation to the Bayes factor, we wrote a Pascal program. We used a cluster sampling scheme and 1500 iterations in each simulation. The cluster size in each run is variable. We used 10, 50 and 200 psus.

6.4.1 The program algorithms

- For the prior

  1) Generate λ as the reciprocal of a $U(0,1)$ random variable. We are assuming that λ has a Pareto distribution with parameters $\lambda_0 = 1$ and $\omega = 1$, which has advantage of being long-tailed, and hence fairly non-informative.

  2) Generate $\boldsymbol{\alpha}$ from the Dirichlet distribution, with parameters $\boldsymbol{\alpha}'_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, representing unit prior information.

  3) Generate a random observation from a Dirichlet distribution, with parameters $\lambda\boldsymbol{\alpha}$, which we consider as the multinomial parameters $\mathbf{p}^t$, for each cluster.

  4) Then apply the function (4.21) on $\mathbf{p}$ to calculate $\mathbf{q}$.

  5) Compute the sample covariance matrix of $\mathbf{q}$, its inverse and determinant.

  6) Finally use the multivariate density estimator, either the crude or the multivariate normal kernel, to estimate the prior density at $\mathbf{p}_0$.

- For the posterior

  1) Set $\lambda = 1000$, $\boldsymbol{\alpha}' = \boldsymbol{\alpha}'_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

  2) Set sampling counter $s = 1$.

  3) Generate a random observation from a Dirichlet distribution, with parameters $\lambda\boldsymbol{\alpha}$, which we consider as the multinomial parameters $\mathbf{p}^t$, for each cluster t.

  4) Generate a random observation, $\mathbf{n}^t$, from the multinomial distribution, based on the random observation from the Dirichlet distribution $\mathbf{p}^t$, for each cluster t.

  5) Compute the design effect, $\hat{\tau}_.$.

  6) Using the MCMC algorithm, generate the random vectors $\mathbf{p}^t$ from the posterior, see the MCMC program algorithms in page (112).

  7) Then apply the function (4.21) on $\mathbf{p}$ to calculate $\mathbf{q}$.

  8) Compute the estimated covariance matrix for $\mathbf{q}$, to derive the bandwidth, $\hat{\mathbf{H}}$; We are considering $\hat{\tau}_.\hat{\mathbf{H}}$ to be the bandwidth in the multivariate normal kernel.

  9) Estimate the marginal posterior density, by either the crude or the multivariate normal kernel at point $\mathbf{p}_0$.

  10) Compute the approximation to the Bayes factor, then calculate $2\ln$(Bayes factor).

  11) Update the sample counter $s = s + 1$, then go to step 3.

In addition to this program, we wrote a Pascal program for computing Pearson chi-squared, Rao and Scott first and second-order corrections for Pearson chi-squared, Wald statistics, uncorrected (multinomial-based) Bayes factor, and two adjustments for the uncorrected Bayes factor, BFRS1 and BFRS2, details are in section (6.5), based on our samples. The cluster sizes used are 10, 50 and 200.

The program algorithms:

1) Start sample $s = 1$.

2) Use the simulated $\mathbf{n}^t$, generated in the previous program at step 4.

3) Compute the covariance-variance matrix under cluster sampling, equation (2.37).

4) Compute the Pearson chi-squared statistic for testing $\mathbf{q} = \mathbf{p}_0$.

5) Compute the design effect for each sample.

6) Compute the Rao and Scott first and second-order corrections, then correct the Pearson chi-squared statistic for testing $\mathbf{q} = \mathbf{p}_0$.

7) Compute the Wald statistics, using the true covariance-variance matrix under cluster sampling, for testing $\mathbf{q} = \mathbf{p}_0$.

8) Compute the uncorrected, multinomial-based, Bayes factor, UBF, then calculate 2 lnUBF.

9) Compute the adjusted Bayes factors, BFRS1 and BFRS2, details are in section (6.5), then calculate 2 lnBFRS1 and 2 lnBFRS2.

10) Compute the BIC values, based on the Pearson chi-squared, Rao and Scott corrected chi-squared, and Wald statistics.

11) Update the sample counter $s = s + 1$, then go to step 2.

6.4.2 Results 1

We ran the simulation first using the crude density estimator to estimate the marginal posterior density $pr(\mathbf{p}|\mathbf{n},M_S)$, and the prior density $pr(\mathbf{p}|M_S)$, at point $\mathbf{p}_0$. This estimator produces sensible estimates when the point $\mathbf{p}_0$ is close to the

mode of the distribution. Unfortunately, when the point $\mathbf{p}_0$ is faraway from the mode, the crude density estimator is poor. The results of our simulation should support model M_0, when we consider $\mathbf{p}_0' = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, and reject model M_0, when we consider $\mathbf{p}_0$ to be any point other than the mode.

When we test our model M_0 at $\mathbf{p}_0' = (0.32, 0.3, 0.38)$, the result of using the crude density estimator is completely misleading as we see in figure (6.15), when compared to the multivariate normal kernel estimator in figure (6.16). Figure (6.15) supports model M_0 at $\mathbf{p}_0' = (0.32, 0.3, 0.38)$, since all the values of $2\ln$(Bayes factor) are negative. However, figure (6.16) does not support model M_0 at $\mathbf{p}_0' = (0.32, 0.3, 0.38)$, since most of the values of $2\ln$(Bayes factor) are positive. Thus, using the crude density estimator is not trustworthy, for this cluster sampling case. As a result, we will use the multivariate normal kernel estimator.

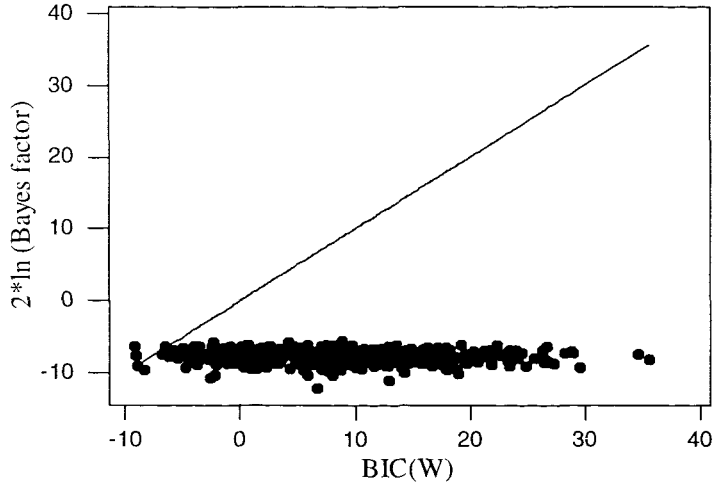

Figure 6.15: Plot of $2\ln$(Bayes factor), using the crude density estimator, and BIC based on Wald statistics, BIC(W), for the model M_0 at $\mathbf{p}_0 = (0.32, 0.3, 0.38)$. The plot is based on 500 samples, each has 50 psus and a sample size of 5000 observations.

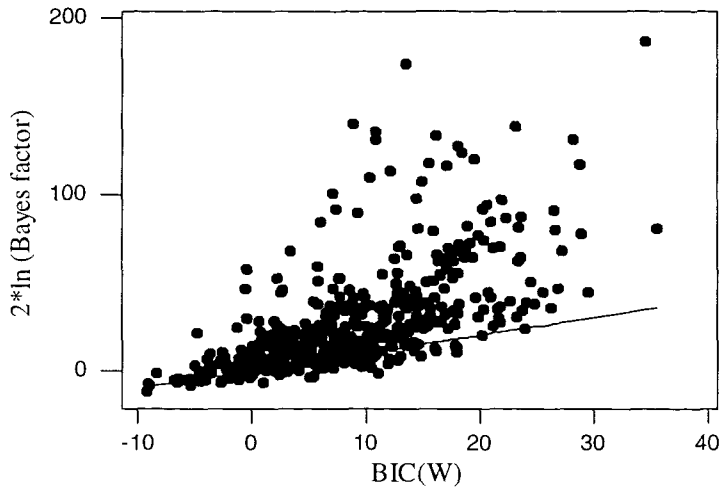

Figure 6.16: Plot of $2\ln$(Bayes factor), using the multivariate normal kernel estimator, and BIC based on Wald statistics, BIC(W), for the model M_0 at $\mathbf{p}_0 = (0.32, 0.3, 0.38)$. The plot is based on 500 samples, each has 50 psus and a sample size of 5000 observations.

6.4.3 Results 2

All the results we are going to discuss in this section are based on 1000 samples. We compare the Bayesian Information Criterion, BIC, see section (3.5), based on four statistics, with $2\ln$(Bayes factor). Those statistics are the Pearson chi-squared, X^2, first-order correction for the Pearson chi-squared by Rao and Scott (1981), X^2_{RS1}, second-order correction for the Pearson chi-squared, also, by Rao and Scott (1981), X^2_{RS2}, and Wald statistic, X^2_W; see section (4.3.3). In the first simulation we restricted the sample size to 1000 observations and we ran the program for 10, 50 and 200 psus, with equal sizes in each psu.

The results of using 10 psus are shown in figure (6.17), 10 observations with extreme positive values of $2\ln$(Bayes factor) have been omitted from the plot. The figure indicates that all the BIC's and $2\ln$(Bayes factor) have the same conclusion supporting M_0, where $\mathbf{q} = \mathbf{p}_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$. The effect of the design is clearly apparent. The approximation of $2\ln$(Bayes factor) is better for the

Wald, and the corrected Pearson chi-squared statistics, than for BIC(X^2).

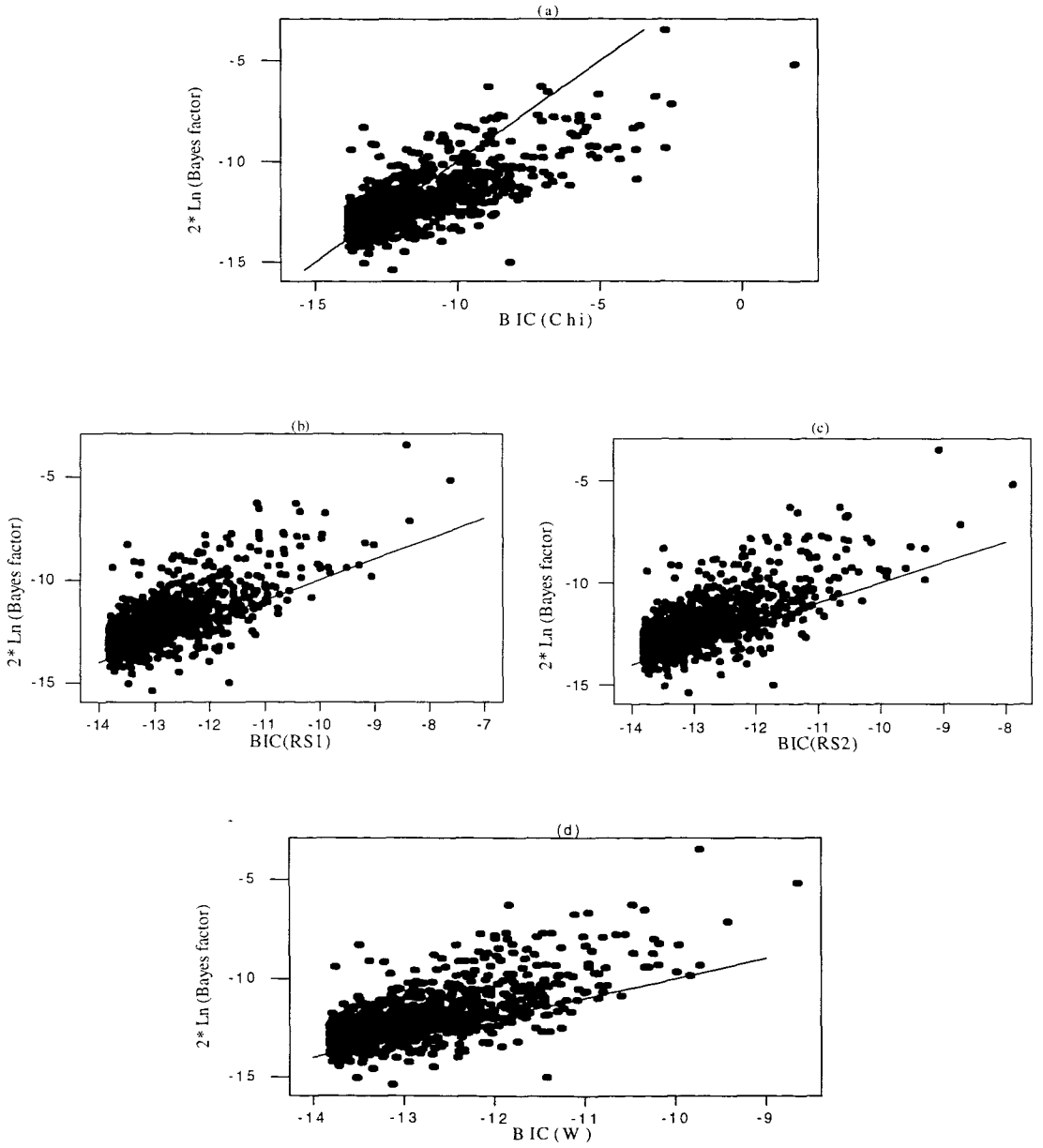

Figure 6.17: Comparison of 2 ln(Bayes factor) values with BIC for sample size of 1000 in 10 psus. (a) BIC based on the Pearson Chi-squared, (b) BIC based on first-order correction for the Pearson Chi-squared , (c) BIC based on second-order correction for the Pearson Chi-squared, and (d) BIC based on Wald statistic.

As the number of psus gets larger, the number of observations in each psu gets smaller. In this case the approximation of 2 ln(Bayes factor) gets worse. This clearly can be seen in figure (6.18) and (6.19), when we have 50 and 200 psus, where $n_t = 20$ and 5. The number of outliers increases as the number of observations in each psu decreases; This is caused by either the normal kernel density estimator or the MCMC algorithm or both.

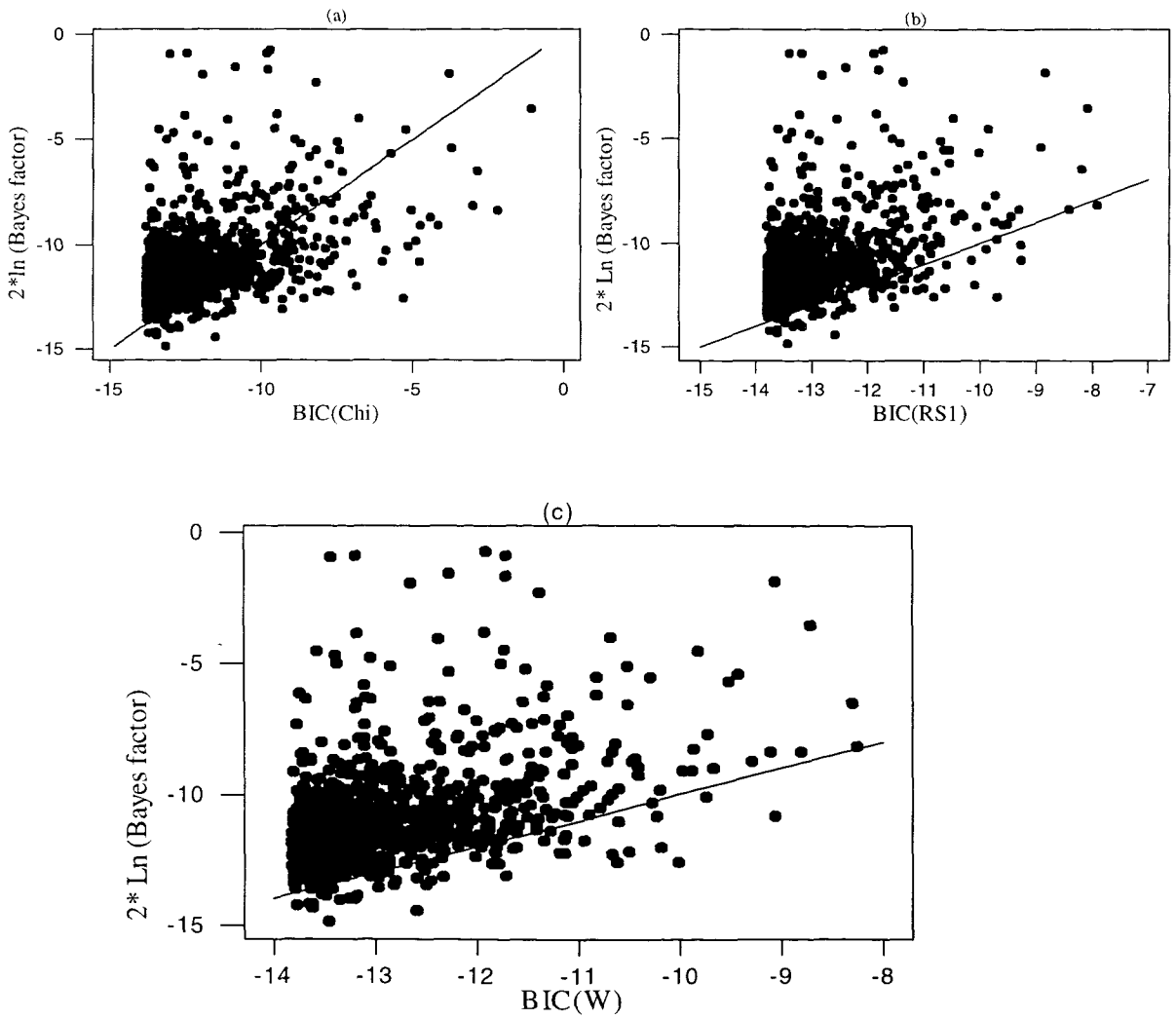

Figure 6.18: Comparison of 2 ln(Bayes factor) values with BIC for sample size of 1000 in 50 psus. (a) BIC based on the Pearson Chi-squared, (b) BIC based on first-order correction for the Pearson Chi-squared, and (c) BIC based on Wald statistic.

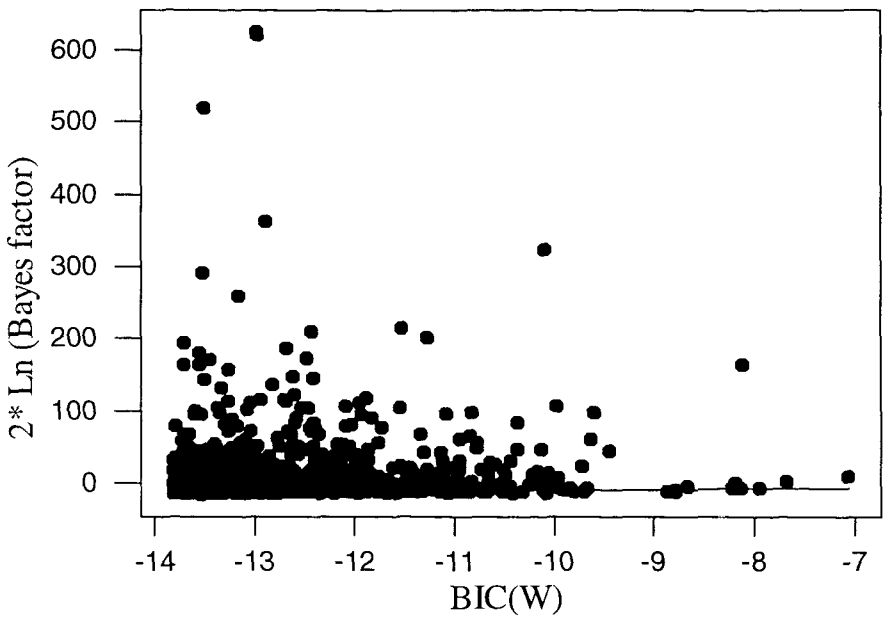

Figure 6.19: Plot of 2 ln(Bayes factor) values with BIC based on Wald statistic, for sample size of 1000 in 200 psus.

When n_t is increased to 100 for 50 psus , the BIC values are reasonably close to 2 ln(Bayes factor). Except, as the values of 2 ln(Bayes factor) get larger, BIC(X^2) tend to overestimate these values. This can be seen in figure (6.20), (26 positive values of 2 ln(Bayes factor) are omitted).

When the number of psus increases to 200, with $n_t = 100$, the BIC approximations are similar to the case with 50 psus, but fewer values lie away from the line; see figure (6.21). In this figure 13 positive values are omitted from 2 ln(Bayes factor).

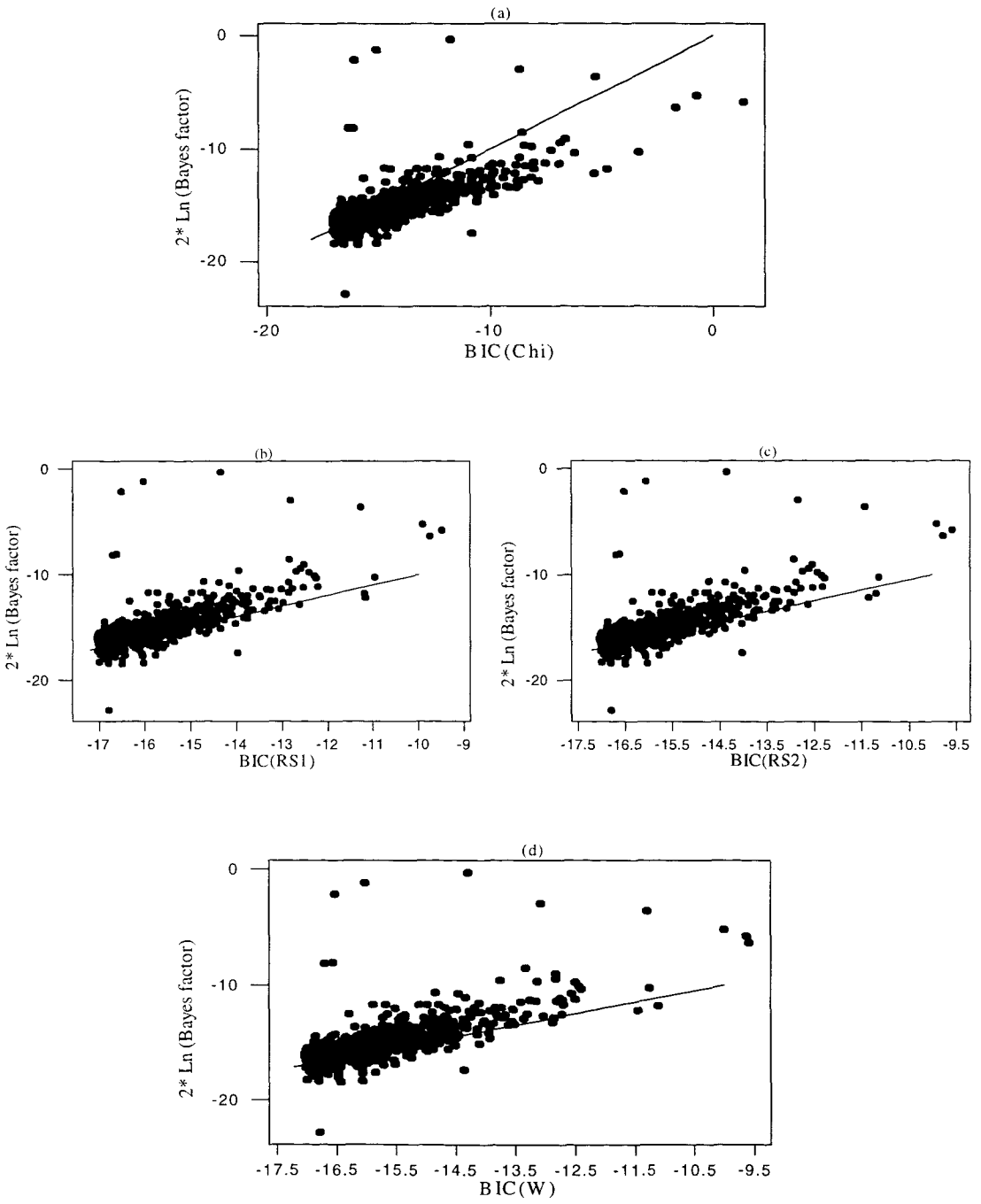

Figure 6.20: Comparison of $2\ln$(Bayes factor) values with BIC for sample size of 5000 in 50 psus, $n_t = 100$. (a) BIC based on the Pearson Chi-squared, (b) BIC based on first-order correction for the Pearson Chi-squared, (c) BIC based on second-order correction for the Pearson Chi-squared, and (d) BIC based on Wald statistic.

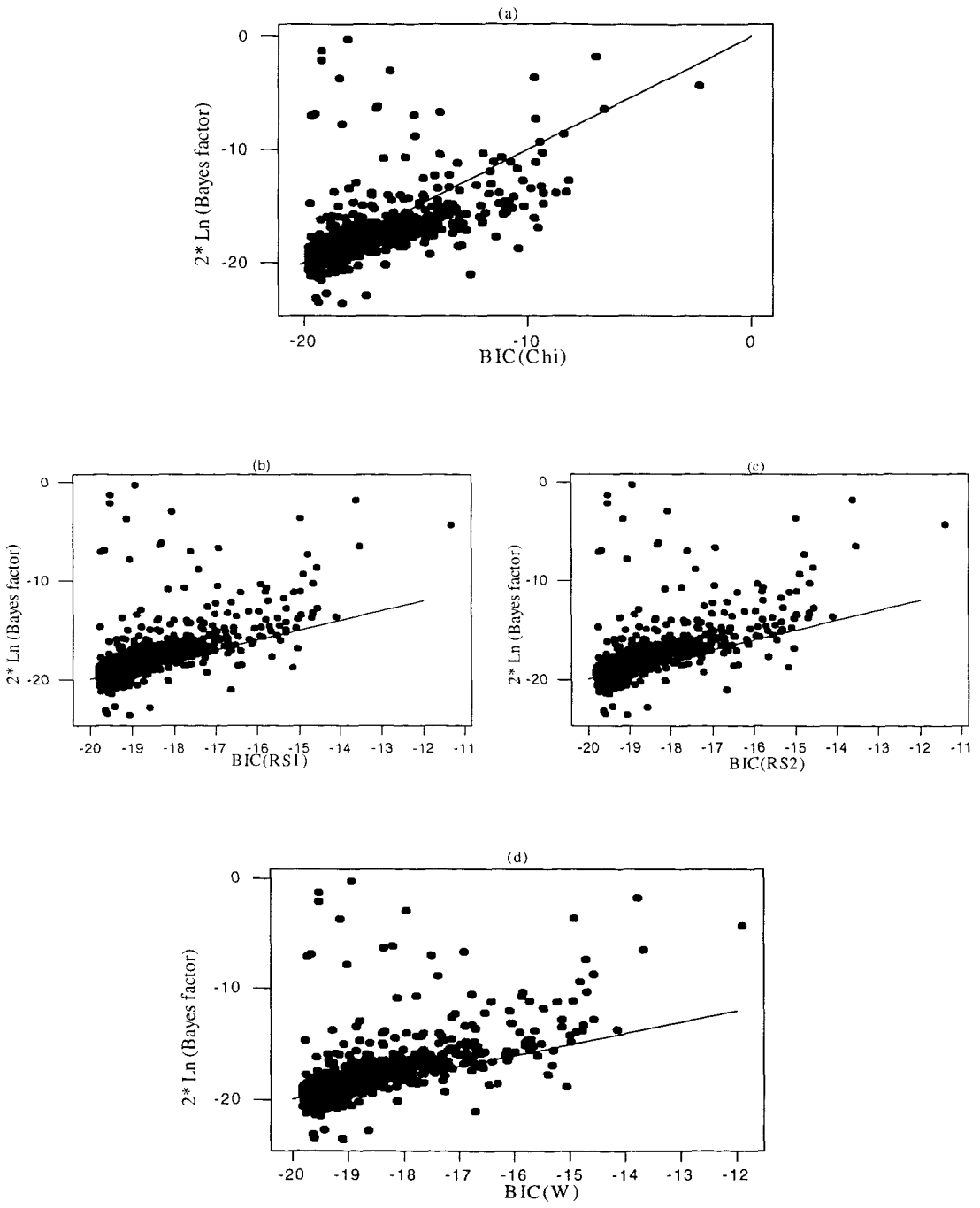

Figure 6.21: Comparison of $2\ln$(Bayes factor) values with BIC for sample size of 20000 in 200 psus, $n_t = 100$. (a) BIC based on the Pearson Chi-squared, (b) BIC based on first-order correction for the Pearson Chi-squared, (c) BIC based on second-order correction for the Pearson Chi-squared, and (d) BIC based on Wald statistic.

If we increase n_t to 1000, i.e. we have a vary large sample size, the asymptotic relationship between 2 ln(Bayes factor) and its approximation is clear in figures (6.22), (6.23) and (6.24). That is, the values of BIC(X^2) tend to overestimate the true values, i.e. 2 ln(Bayes factor) values. While the BIC(X^2_{RS1}), BIC(X^2_{RS2}) and BIC(X^2_W) tend to underestimate the true values, as they get larger and the support for model M_0 is weaker.

When we have 10 psus, with $n_t = 1000$, figure (6.22-a) shows that BIC(X^2) severely overestimates the values of 2 ln(Bayes factor), as they get larger. For large values of 2 ln(Bayes factor), the end result of using BIC(X^2), as determinant of the model selection, is not supporting M_0, giving a misleading conclusion. On the other hand, the BIC(X^2_{RS1}), BIC(X^2_{RS2}) and BIC(X^2_W) underestimate those large values of 2 ln(Bayes factor); BIC(X^2_W) has the best approximation of all.

For 50 psus, with $n_t = 1000$, the relationship is clear. The values of BIC(X^2) tend to overestimates most the true values, i.e. 2 ln(Bayes factor) values. While the BIC(X^2_{RS1}), BIC(X^2_{RS2}) and BIC(X^2_W) tend roughly to underestimate all the true values, figure (6.23).

When we have a large number of psus, such as 200 with $n_t = 1000$, figure (6.24) shows that BIC(X^2) often severely overestimates the values of 2 ln(Bayes factor). On the other hand, the BIC(X^2_{RS1}), BIC(X^2_{RS2}) and BIC(X^2_W) have approximately the same performance. They underestimate nearly all values of 2 ln(Bayes factor).

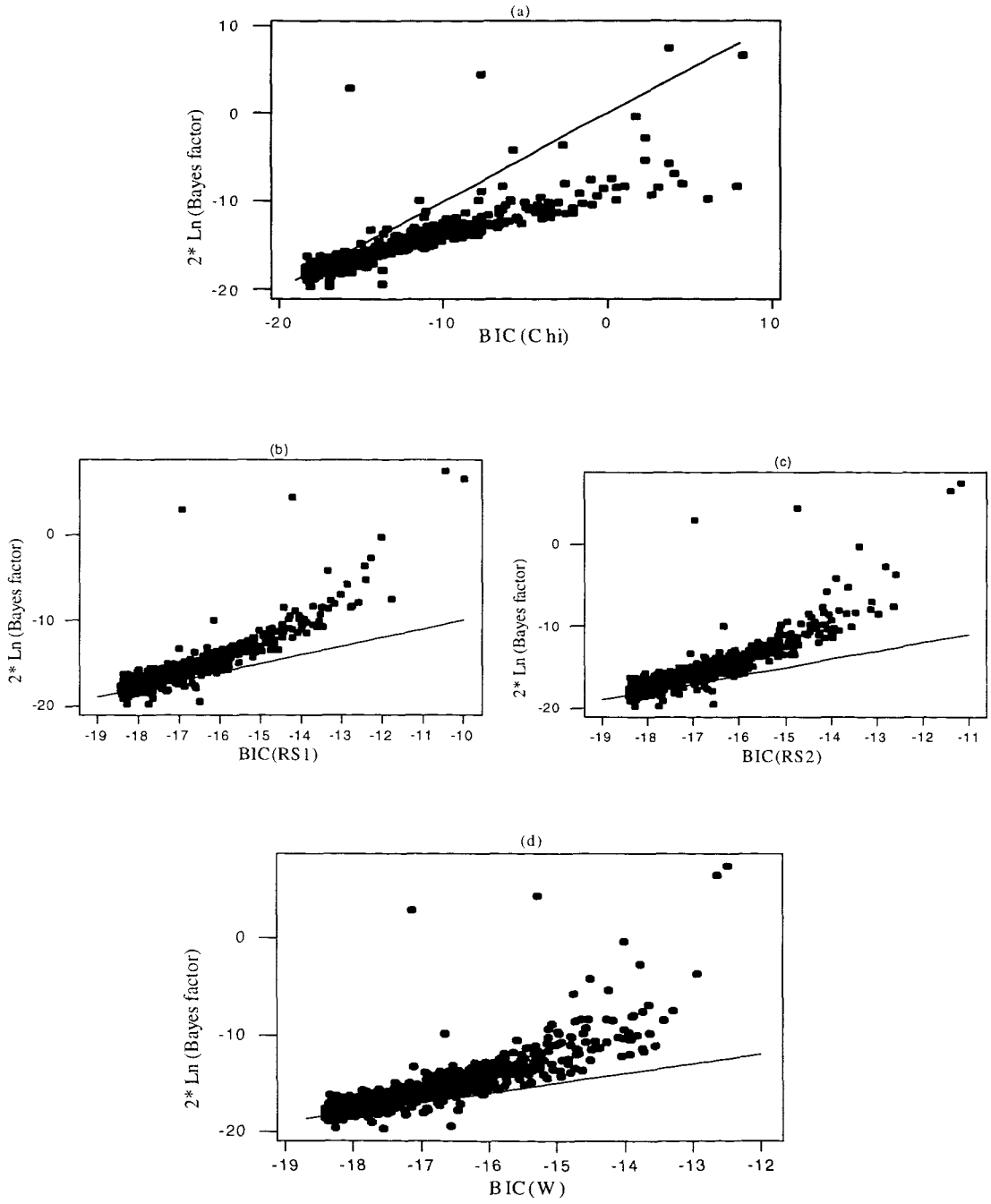

Figure 6.22: Comparison of $2\ln$(Bayes factor) values with BIC for sample size of 10000 in 10 psus, $n_t = 1000$. (a) BIC based on the Pearson Chi-squared, (b) BIC based on first-order correction for the Pearson Chi-squared, (c) BIC based on second-order correction for the Pearson Chi-squared, and (d) BIC based on Wald statistic.

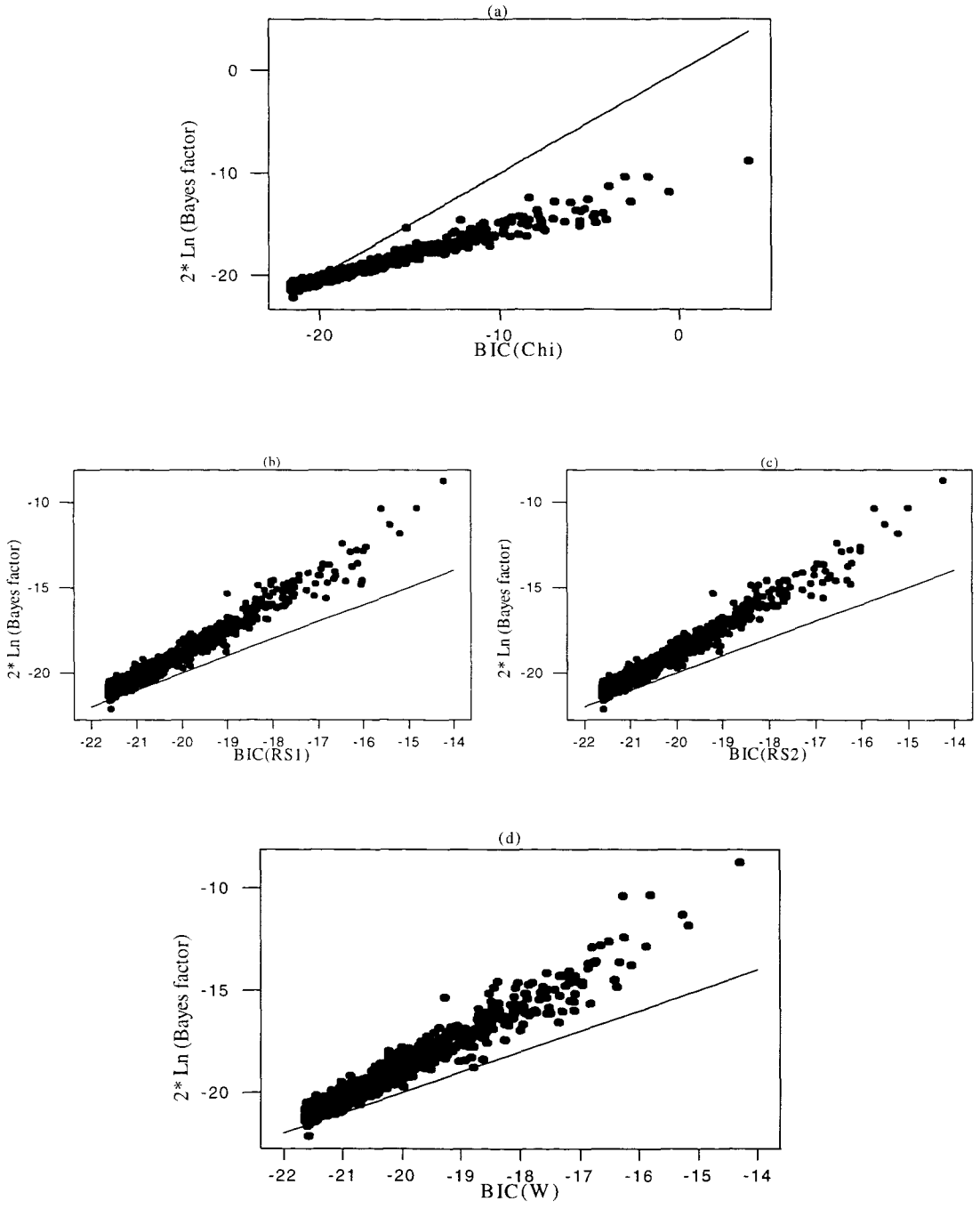

Figure 6.23: Comparison of $2\ln$(Bayes factor) values with BIC for sample size of 50000 in 50 psus, $n_t = 1000$. (a) BIC based on the Pearson Chi-squared, (b) BIC based on first-order correction for the Pearson Chi-squared, (c) BIC based on second-order correction for the Pearson Chi-squared, and (d) BIC based on Wald statistic.

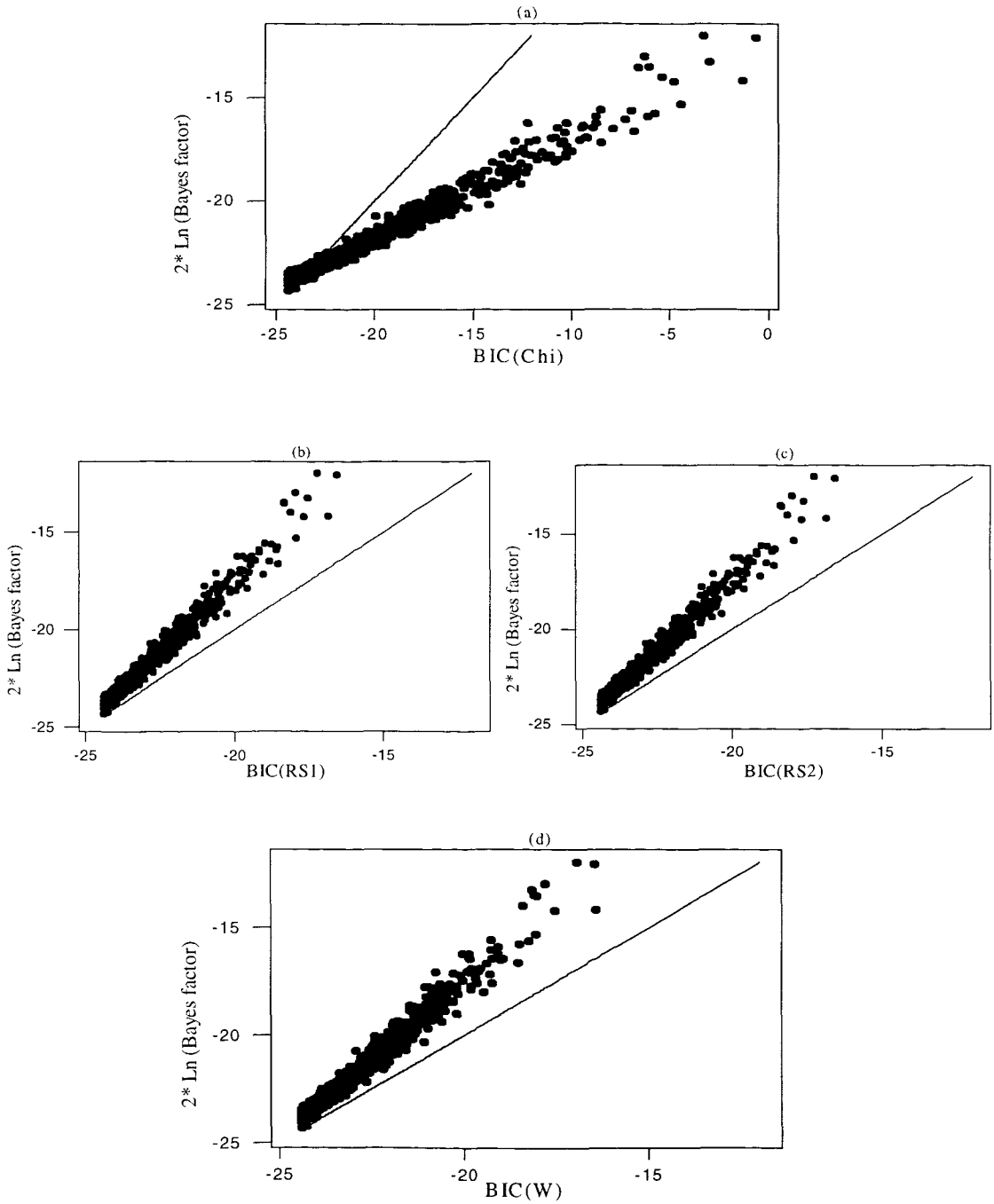

Figure 6.24: Comparison of $2\ln$(Bayes factor) values with BIC for sample size of 200000 in 200 psus, $n_t = 1000$. (a) BIC based on the Pearson Chi-squared, (b) BIC based on first-order correction for the Pearson Chi-squared, (c) BIC based on second-order correction for the Pearson Chi-squared, and (d) BIC based on Wald statistic.

6.4.4 Discussion

In almost all cases $2\ln$(Bayes factor) and BIC support the null model M_0, the exception being figure (6.22-a). In this case, where we have 10 psus and $n_t = 1000$, using BIC(X^2) for model selection may produce the misleading result of rejecting M_0, when it is true. This does not indicate that under cluster sampling the corrected BICs have a good result. In fact all BICs are affected by the sampling design. In general, BIC(X^2) overestimates the values of $2\ln$(Bayes factor) in all cases. This is related to the unsatisfied assumption that X^2 asymptotically has a χ^2 distribution with $K-1$ degrees of freedom, as discussed before. This can be seen in figure (6.25-a) if compared with the χ^2 distribution with 2 degrees of freedom in figure (6.25-c).

On the other hand, the BIC(X^2_{RS1}), BIC(X^2_{RS2}) and BIC(X^2_W) tend to underestimate the large values of $2\ln$(Bayes factor), but not with the same magnitude as BIC(X^2). This is caused by the conservative values of the corrected X^2_{RS1}, X^2_{RS2}, and X^2_W, as shown in figure(6.25-b) for the Wald statistic, X^2_W, sample values. In the statistical literature, there is evidence to suggest that the Wald statistic may exhibit a poor behaviour in sample surveys (Thomas and Rao, 1987, Skinner, Holt and Smith, 1989). Fay (1985) presents his concerns regarding the use of the Wald statistic, X^2_W, and Thomas and Rao (1987), in a Monte Carlo study, confirm Fay's concerns.

The results we present show the effect of cluster sampling on Bayesian model selection. In this simulation two variables affect our results, the sample size, or the number of observations in each psu, and the number of psus. These variables affect our computation of the Bayes factor, through the number of iterations needed in MCMC, and they affect our results in the approximation of $2\ln$(Bayes factor).

In conclusion, BIC(X^2) severely overestimates the true values of $2\ln$(Bayes

factor), as the number of psus, c, and n get larger. However, in this case, even with BIC(X^2_W) the approximation to the true value of 2 ln(Bayes factor) is not as we expected; The cluster sampling design affects its values. In fact, for large c and n, BIC(X^2_W) underestimates almost all values of 2 ln(Bayes factor).

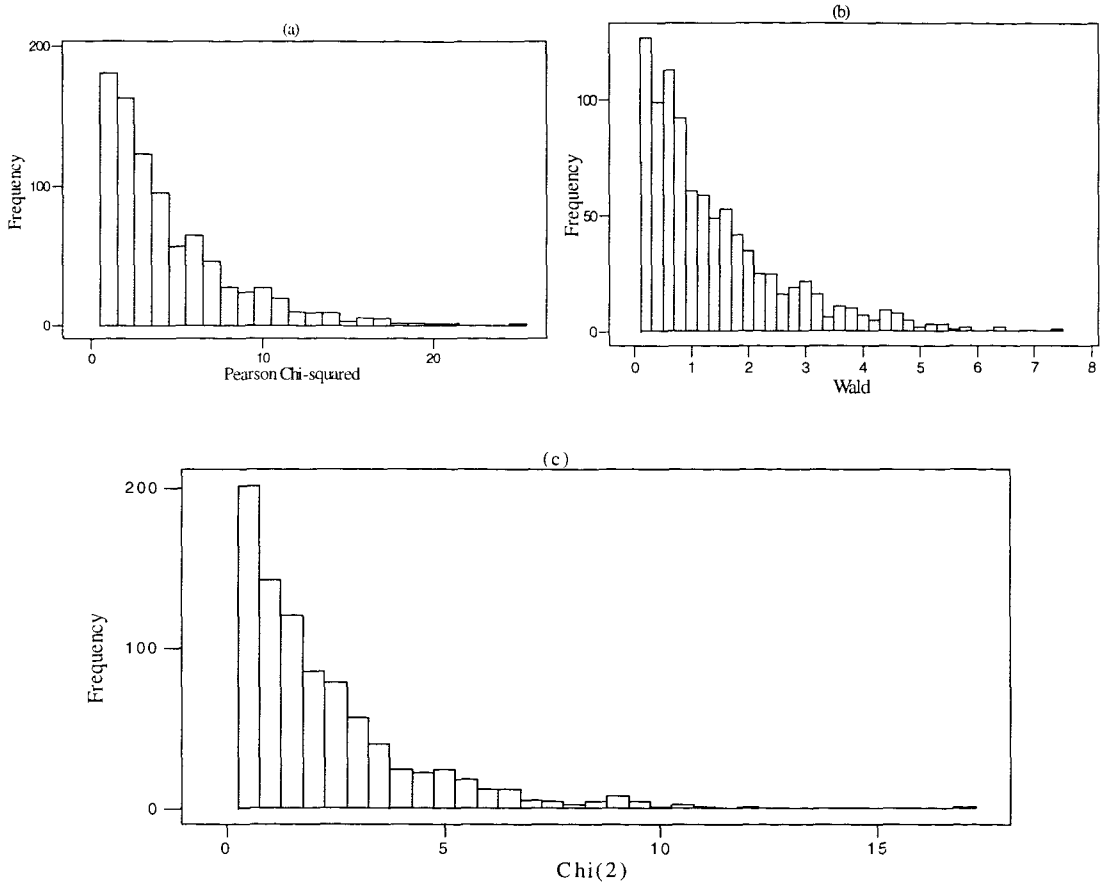

Figure 6.25: A histogram for the sample values of (a) Pearson Chi-squared statistic, X^2, where we have 50 psus and $n_t = 1000$, (b) Wald statistic, X^2_W, where we have 50 psus and $n_t = 1000$, and (c) χ^2 distribution, with 2 degrees of freedom,

Finally, when n_t is small, we are uncertain about the values of 2 ln(Bayes factor). Because we have serious doubts that our MCMC algorithm is successfully generating from the marginal posterior distribution $pr(\mathbf{p}|\mathbf{n},M_S)$, the need for further runs is obvious.

6.5 Adjusting the multinomial-based Bayes factor

To this point we have focussed our discussion on the approximation of 2 ln(Bayes factor) using BIC, based on different classical statistics. In this section, we will present and discuss results using two simple adjustments to the multinomial-based Bayes factor. The first adjustment is dividing the cell total, $\mathbf{n}$, by the first-order correction of Rao and Scott (1981), $\hat{\tau}_.$, see section (2.8). The adjusted cell total will be $\mathbf{n}_{Adj} = (\frac{n_1}{\hat{\tau}_.}, \cdots, \frac{n_K}{\hat{\tau}_.})'$. The second adjustment is dividing the cell total, $\mathbf{n}$, by the second-order correction of Rao and Scott (1981). Then we compute the Bayes factor by considering the multinomial-based Bayes factor (see equation (4.9)), using the adjusted cell total, $\mathbf{n}_{Adj}$ and compare its result with the true values of 2 ln(Bayes factor).

The results of these adjustments are promising. We consider the sample design with 10 psus, as a first case. When our sample has 1000 observations, the 2 ln(uncorrected Bayes factor) values tend to underestimate the small values of 2 ln(Bayes factor), where it supports the model, M_0, and overestimate the larger values of 2 ln(Bayes factor) otherwise, see figure (6.26-a). When we use the first-order adjustment to the uncorrected Bayes factor, BFRS1, we have slightly more reliable estimated values, as seen in figure (6.26-b). Using the second-order adjustment to the uncorrected Bayes factor, BFRS2, gives us a similar result to the first adjustment. This can be seen in figure (6.26-c).

As we increase the sample size to 10000 observations, $n_t = 1000$, 2 ln(uncorrected Bayes factor) has the potential to introduce a misleading result, of rejecting model M_0 while it is true, since it severely overestimates the true values as they get larger. On the other hand, using the first-order adjustment, BFRS1, will yield reliable estimated values. If we consider the second-order adjustment, the result is very similar to the result of the first-order adjustment, figure (6.27).

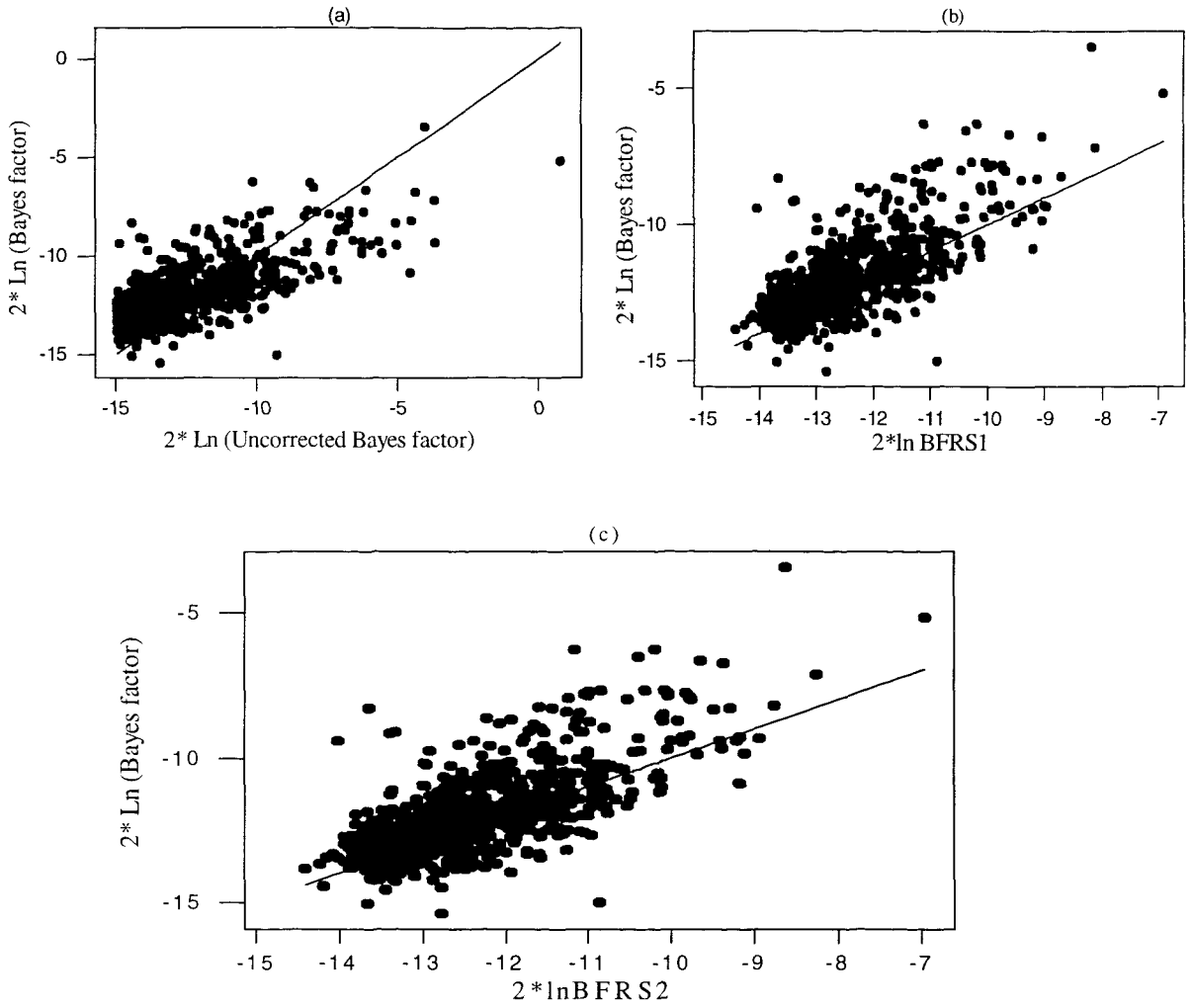

Figure 6.26: Comparison of 2 ln(Bayes factor) with uncorrected and adjusted multinomial-based Bayes factors, for sample size of 1000 in 10 psus. (a) 2 ln(Bayes factor) with 2 ln(uncorrected Bayes factor), (b) 2 ln(Bayes factor) with 2 ln(BFRS1) using first adjustment of the multinomial-based Bayes factor, and (c) 2 ln(Bayes factor) with 2 ln(BFRS2) using second adjustment.

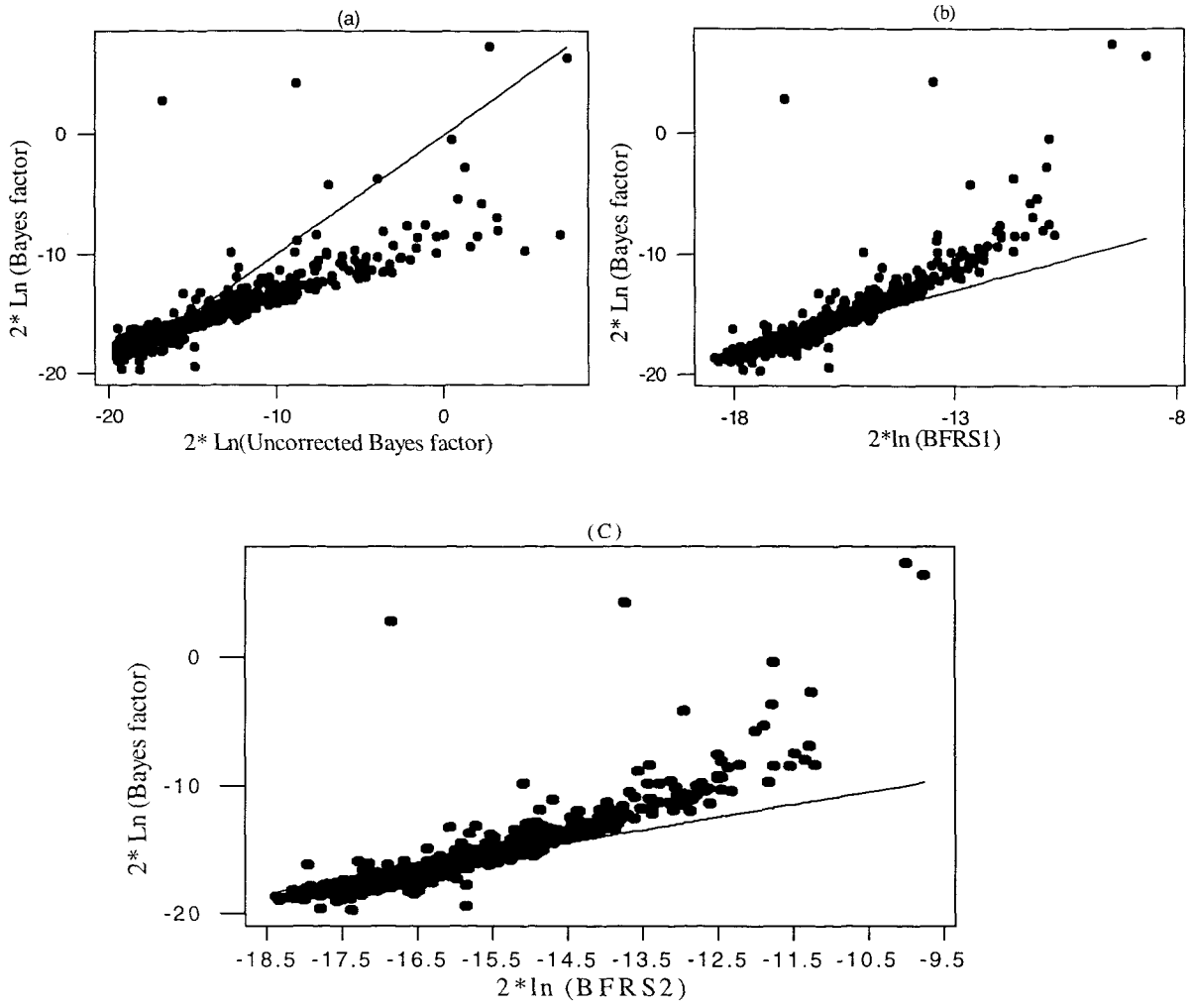

Figure 6.27: Comparison of $2\ln$(Bayes factor) with uncorrected and adjusted multinomial-based Bayes factors, for sample size of 10000 in 10 psus, $n_t = 1000$. (a) $2\ln$(Bayes factor) with $2\ln$(uncorrected Bayes factor), (b) $2\ln$(Bayes factor) with $2\ln$(BFRS1) using first adjustment of the multinomial-based Bayes factor, and (c) $2\ln$(Bayes factor) with $2\ln$(BFRS2) using second adjustment.

If we consider the case where we have 50 psus, when the sample size, n, is equal to 1000, the values of $2\ln$(uncorrected Bayes factor) mainly underestimate the values of $2\ln$(Bayes factor). Also, both adjustment for the multinomial-based Bayes factor, BFRS1 and BFRS2, do not approximate the values of $2\ln$(Bayes factor) reasonably. The cause of this may relate to the small number of observations, $n_t = 20$, in each psu and its direct effect on the actual values of

2 ln(Bayes factor), figure (6.28).

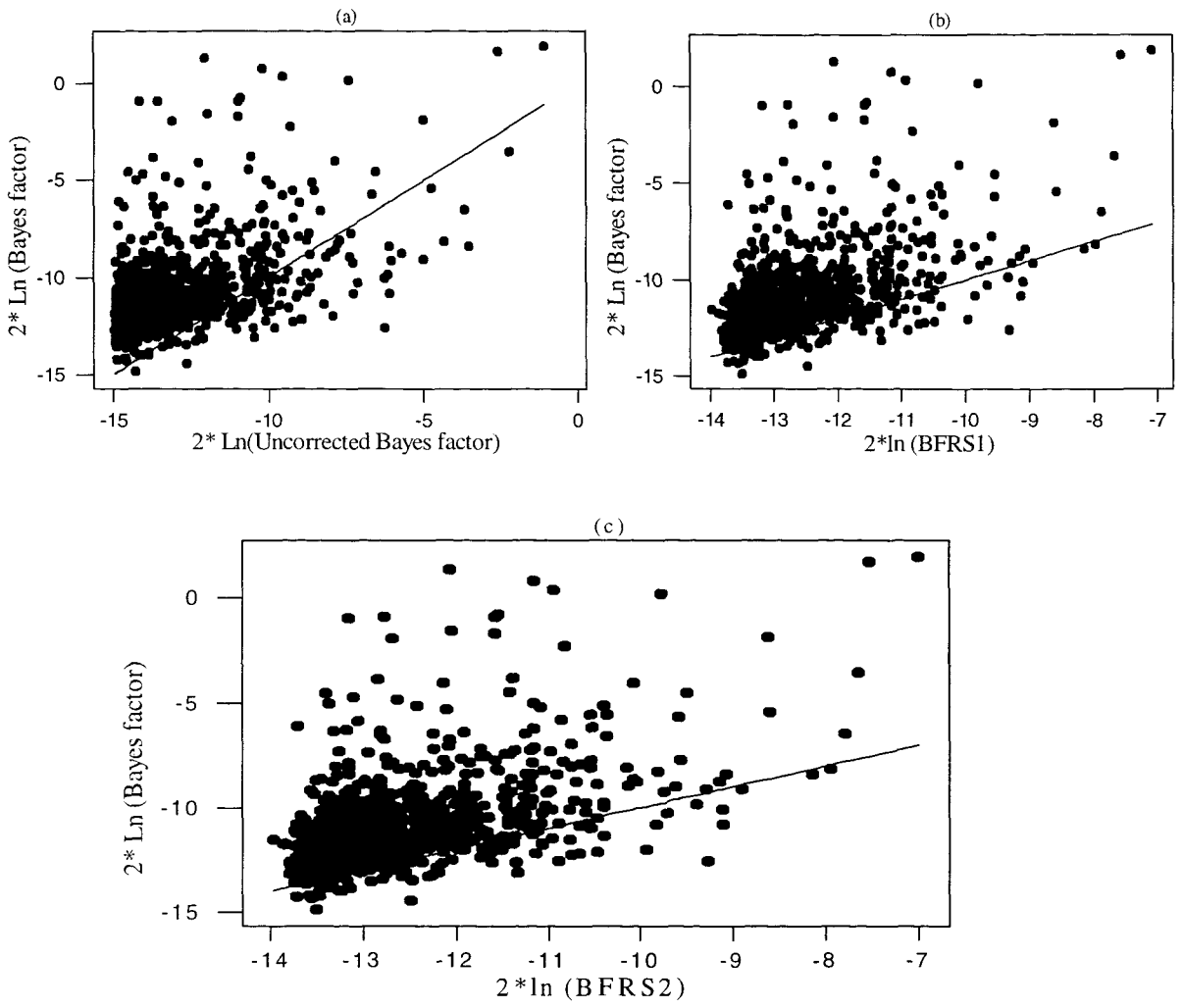

Figure 6.28: Comparison of 2 ln(Bayes factor) with uncorrected and adjusted multinomial-based Bayes factors, for sample size of 1000 in 50 psus, $n_t = 20$. (a) 2 ln(Bayes factor) with 2 ln(uncorrected Bayes factor), (b) 2 ln(Bayes factor) with 2 ln(BFRS1) using first adjustment of the multinomial-based Bayes factor, and (c) 2 ln(Bayes factor) with 2 ln(BFRS2) using second adjustment.

However, when we increase the sample size to 5000, $n_t = 100$, figure (6.29) shows that the adjusted Bayes factor approximates the true values of 2 ln(Bayes factor) sensibly.

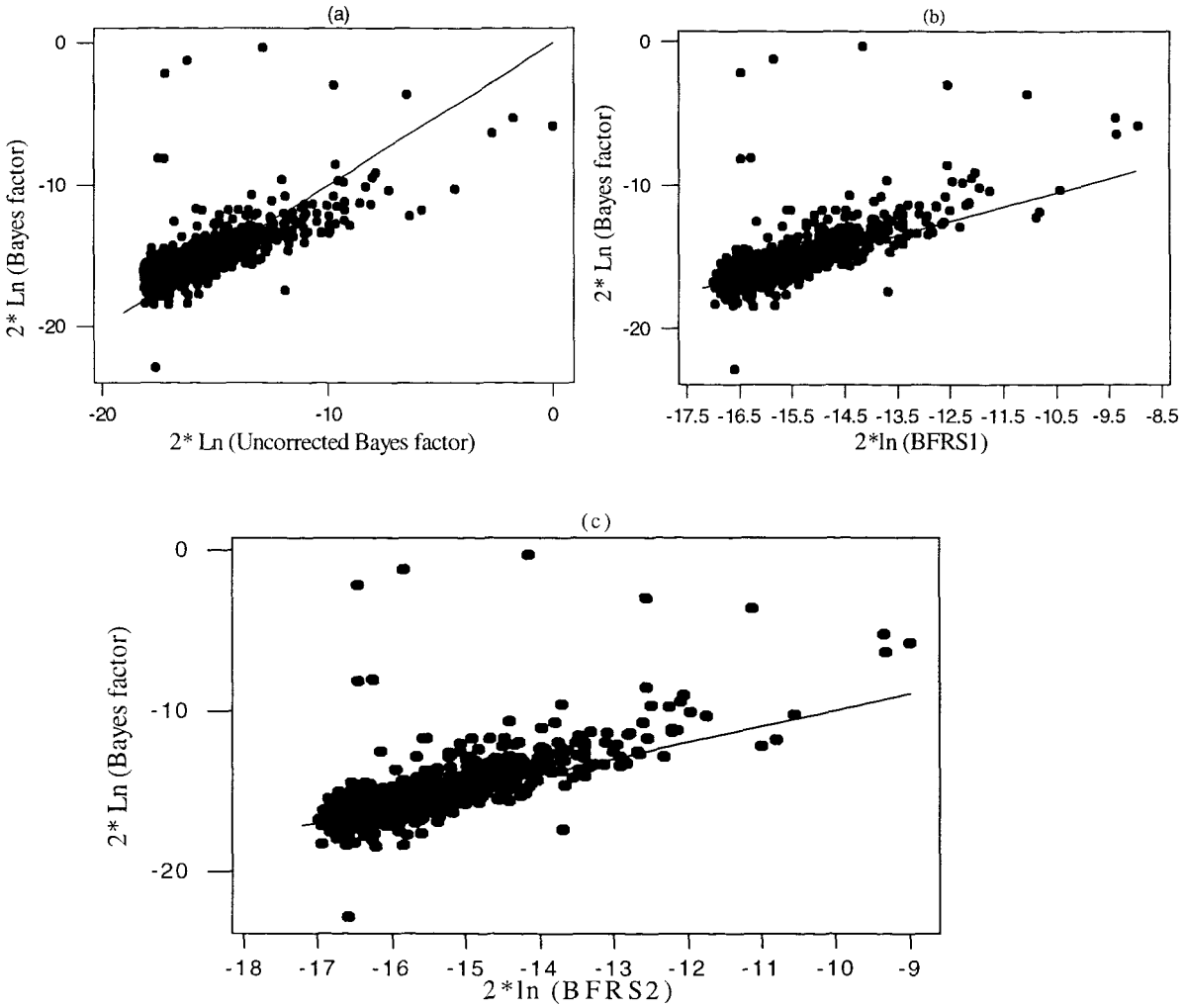

Figure 6.29: Comparson of 2 ln(Bayes factor) with uncorrected and adjusted multinomial-based Bayes factors, for sample size of 5000 in 50 psus, $n_t = 100$. (a) 2 ln(Bayes factor) with 2 ln(uncorrected Bayes factor), (b) 2 ln(Bayes factor) with 2 ln(BFRS1) using first adjustment of the multinomial-based Bayes factor, and (c) 2 ln(Bayes factor) with 2 ln(BFRS2) using second adjustment.

As n gets larger, with $n_t = 1000$, for 50 psus, the values of 2 ln(uncorrected Bayes factor) start underestimating the values of 2 ln(Bayes factor). As those values get larger, 2 ln(uncorrected Bayes factor) severely overestimates the true values. On the other hand, both adjustments, BFRS1 and BFRS2, approximate the values of 2 ln(Bayes factor) more accurately; see figure (6.30).

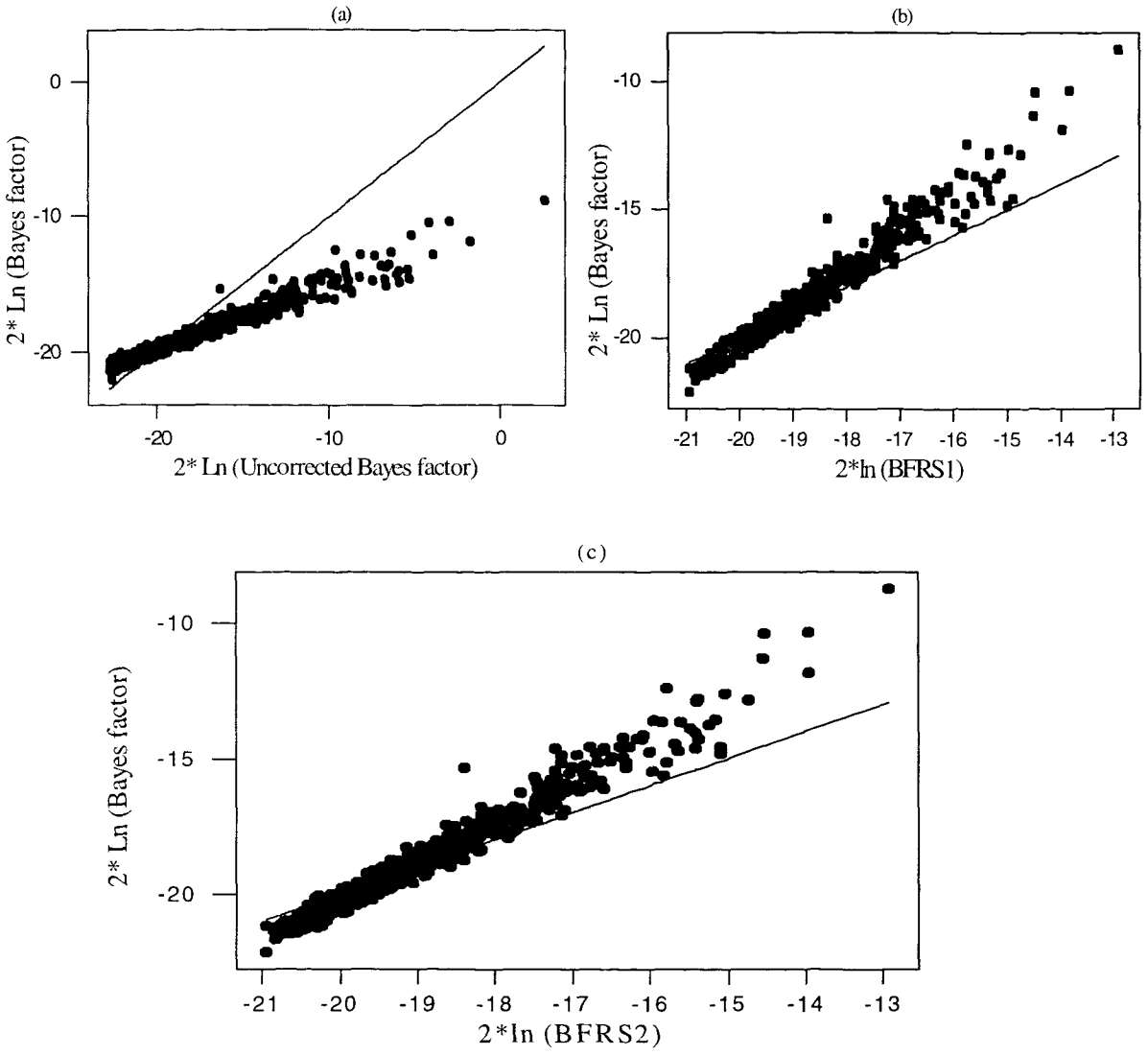

Figure 6.30: Comparison of 2 ln(Bayes factor) with uncorrected and adjusted multinomial-based Bayes factors, for sample size of 50000 in 50 psus, $n_t = 1000$. (a) 2 ln(Bayes factor) with 2 ln(uncorrected Bayes factor), (b) 2 ln(Bayes factor) with 2 ln(BFRS1) using first adjustment of the multinomial-based Bayes factor, and (c) 2 ln(Bayes factor) with 2 ln(BFRS2) using second adjustment.

Finally, if we have 200 psus, the status will be almost identical to the case with 50 psus. Figure (6.31) shows the effect of small n_t on the values of 2 ln(Bayes factor), such as figure (6.19), when $n = 1000$. When we increase the sample size, n, to 20000, $n_t = 100$, the adjusted Bayes factor approximates the true values

of 2 ln(Bayes factor) sensibly, see figure (6.32), with slightly bigger variation compared with the 50 psus case.

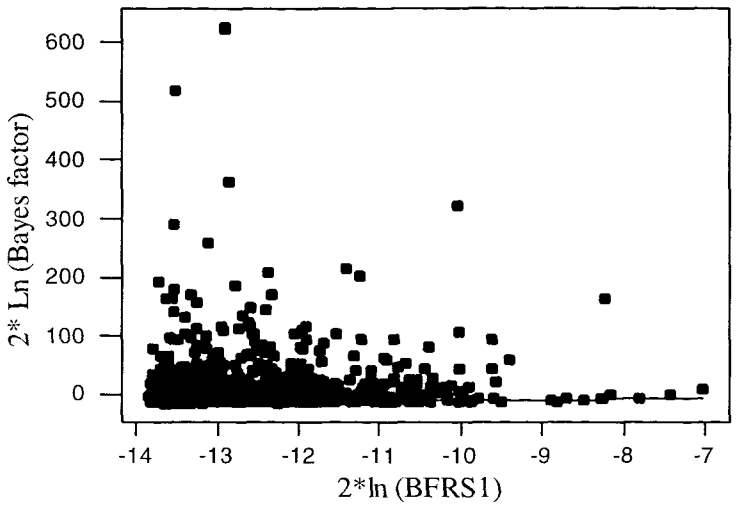

Figure 6.31: Comparison of 2 ln(Bayes factor) with first adjusted multinomial-based Bayes factor, 2 ln(BFRS1), for sample size of 1000 in 200 psus, $n_t = 5$.

For a very large sample size, $n = 200000$, the asymptotic relationship is visible. The 2 ln(uncorrected Bayes factor) underestimates the smaller values of 2 ln(Bayes factor) and badly overestimates the true values as it get larger. Nevertheless, both adjusted Bayes factors insignificantly overestimate the smaller true values and slightly underestimate the large values of 2 ln(Bayes factor), as shown in figure (6.33). Nevertheless, they produce accurate estimates of the true values of 2 ln(Bayes factor).

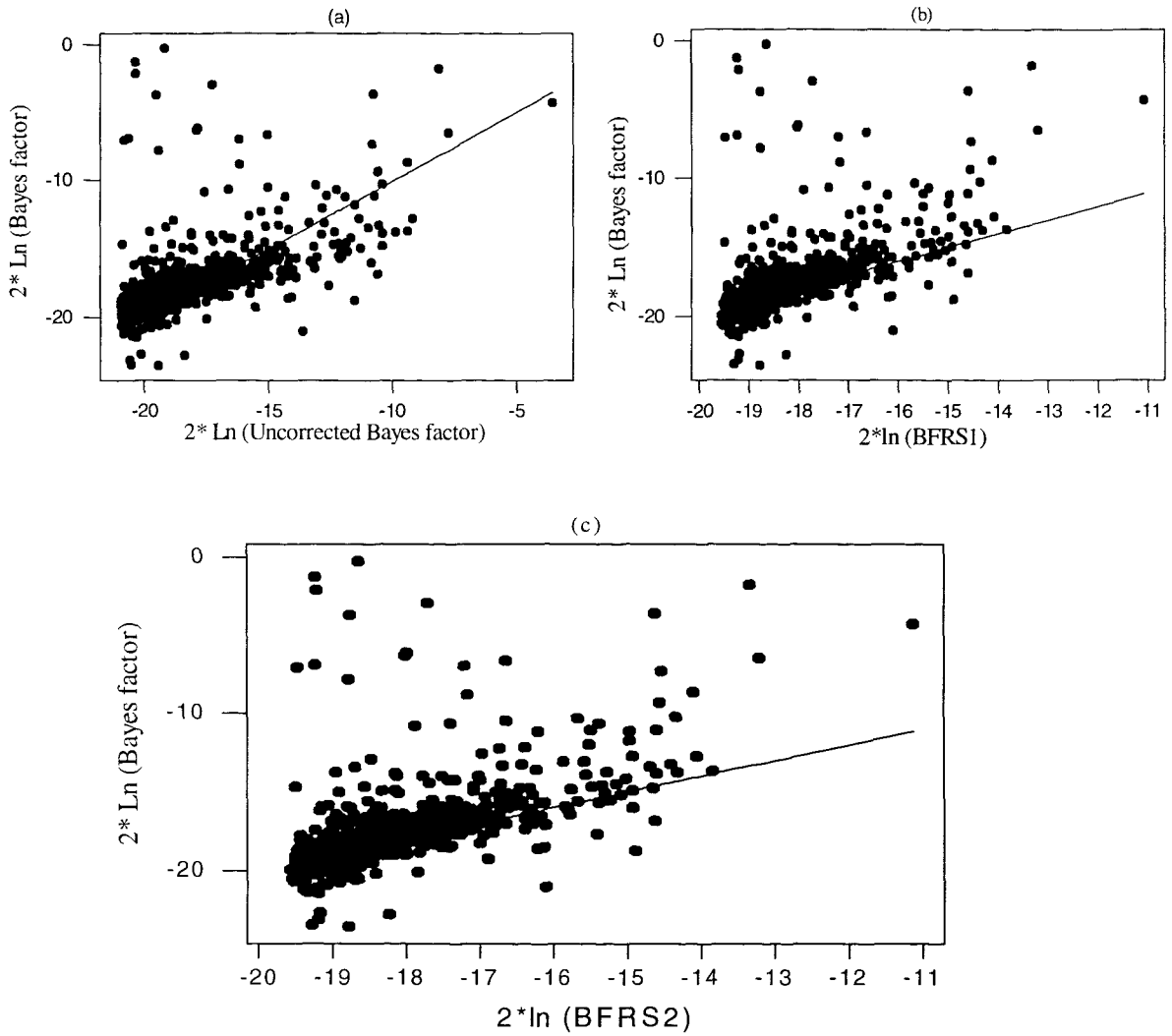

Figure 6.32: Comparison of $2\ln$(Bayes factor) with uncorrected and adjusted multinomial-based Bayes factors, for sample size of 20000 in 200 psus, $n_t = 100$. (a) $2\ln$(Bayes factor) with $2\ln$(uncorrected Bayes factor), (b) $2\ln$(Bayes factor) with $2\ln$(BFRS1) using first adjustment of the multinomial-based Bayes factor, and (c) $2\ln$(Bayes factor) with $2\ln$(BFRS2) using second adjustment.

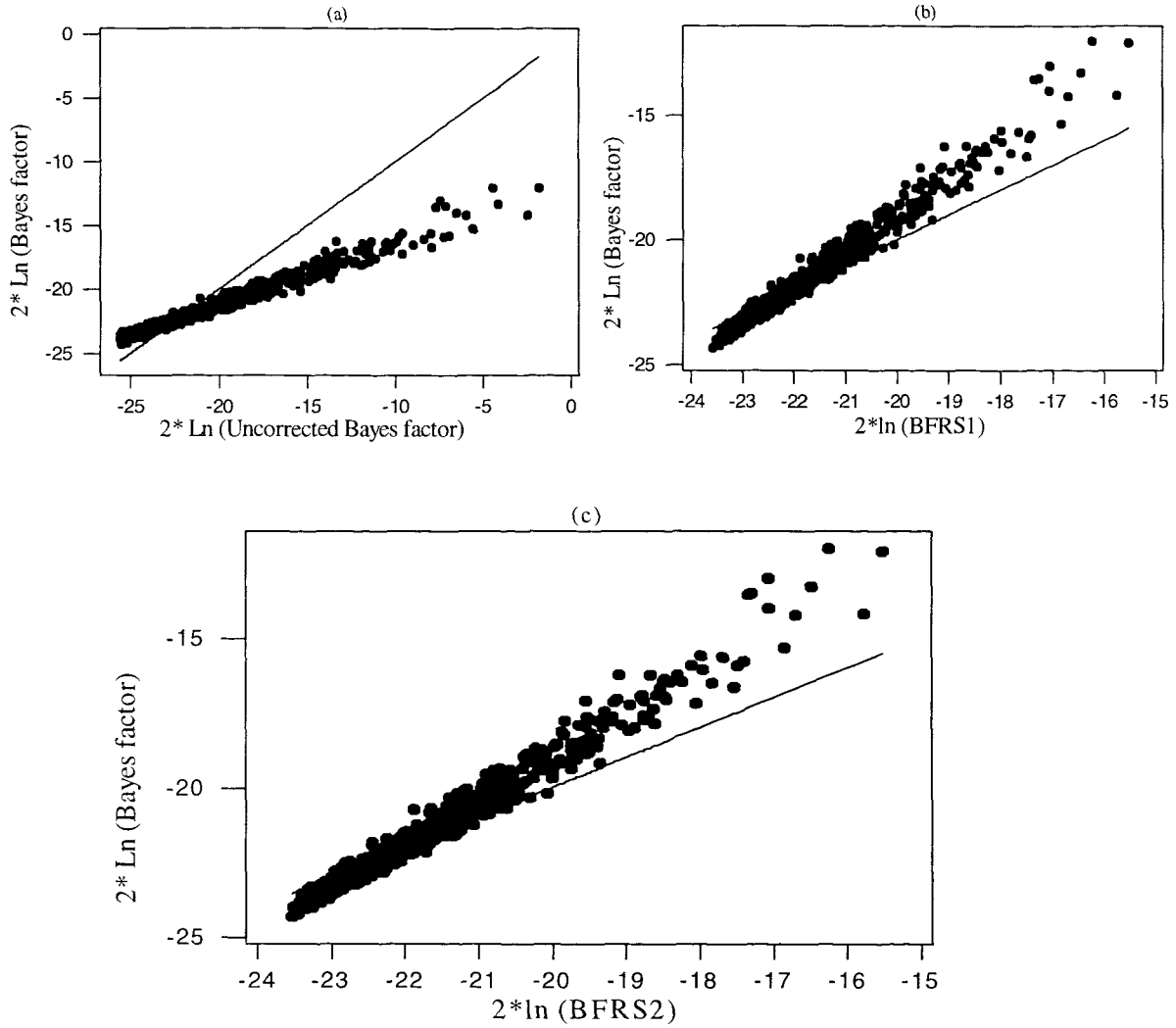

Figure 6.33: Comparison of $2\ln$(Bayes factor) with uncorrected and adjusted multinomial-based Bayes factors, for sample size of 200000 in 200 psus, $n_t = 1000$. (a) $2\ln$(Bayes factor) with $2\ln$(uncorrected Bayes factor), (b) $2\ln$(Bayes factor) with $2\ln$(BFRS1) using first adjustment of the multinomial-based Bayes factor, and (c) $2\ln$(Bayes ffactor) with $2\ln$(BFRS2) using second adjustment.

6.6 Comparison of adjusted Bayes factor with BIC(X^2_W)

In this section, we are going to present the first-order adjustment to the multinomial-based Bayes factor, BFRS1. We will compare it with BIC(X^2_W), as they are both approximations to the true values of 2 ln(Bayes factor). Results in sections (6.4.3) and (6.5), show the adjustment to the multinomial-based Bayes factor, BFRS1 and BFRS2, are superior to BIC(X^2_W) in estimating the values of 2 ln(Bayes factor), if the number of units in each psu is large. This result is illustrated by comparing figure (6.22) with figure (6.27) for 10 psus, figure (6.23) with figure (6.30) for 50 psus, and figure (6.24) with figure (6.33) for 200 psus. Unfortunately, if each psu has small numbers of units, the results in these sections may not be considered, for reasons discussed previously. These factors are the motivation behind this section.

When the number of observations in each psu, n_t, is small, the behaviour of both factors is identical. They are linearly distributed along the line of equality in figure (6.34), particularly, in the case of 200 psus with $n_t = 5$. If $n_t = 100$, figure (6.35) shows clearly as the number of psus get larger, the values have smaller variation between them and are concentrated slightly above the line of equality, i.e. the values of BIC(X^2_W) are slightly smaller than 2 ln(BFRS1). As n_t gets larger the difference between the two factors get systematically larger. The maximum difference is in the case of 200 psus. This difference is roughly equal for all values, i.e. BIC(X^2_W) $-$ 2 ln(BFRS1) = constant ≈ -1. This error may be related to the error associated with BIC approximation, which can be of order $O(1)$.

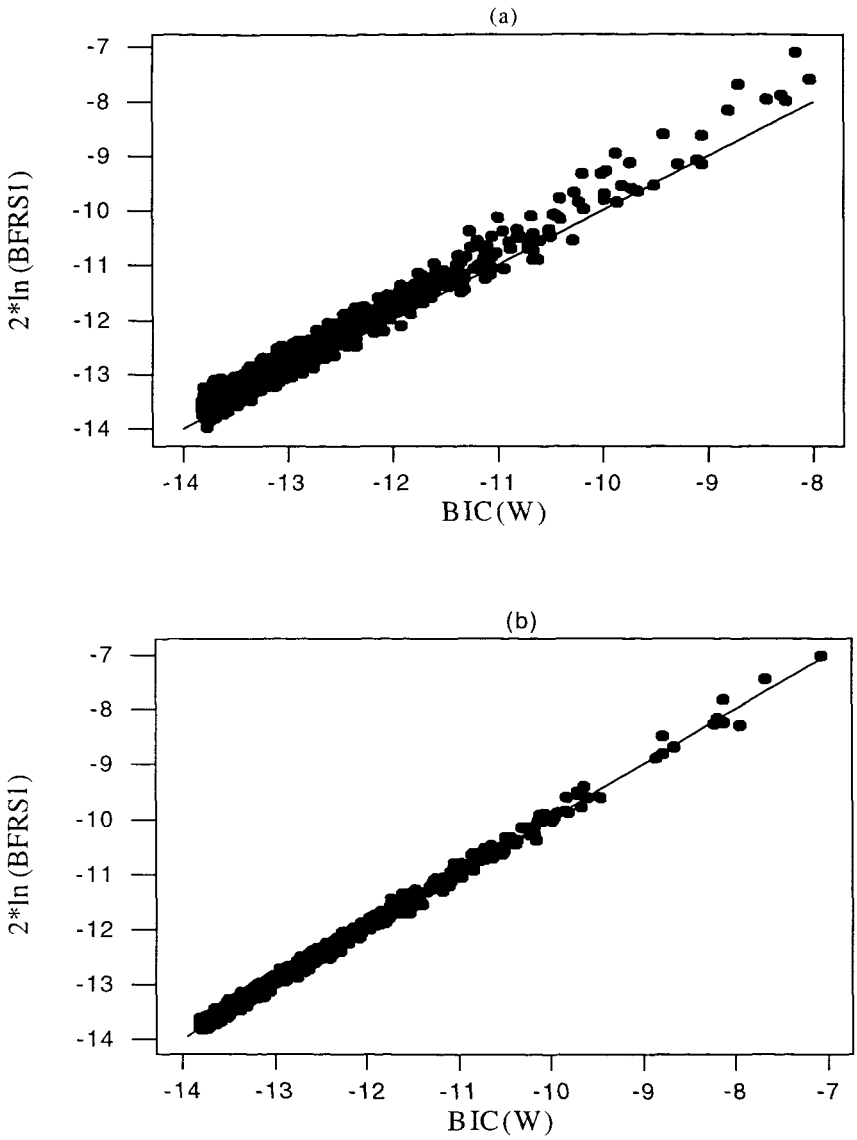

Figure 6.34: Comparison of using the first-order adjustment to the multinomial-based Bayes factor, BFRS1, for approximating the true values of $2\ln$(Bayes factor) with BIC(X^2_W) , where n_t is small. (a) under 50 psus and $n_t = 20$, (b) 200 psus and $n_t = 5$.

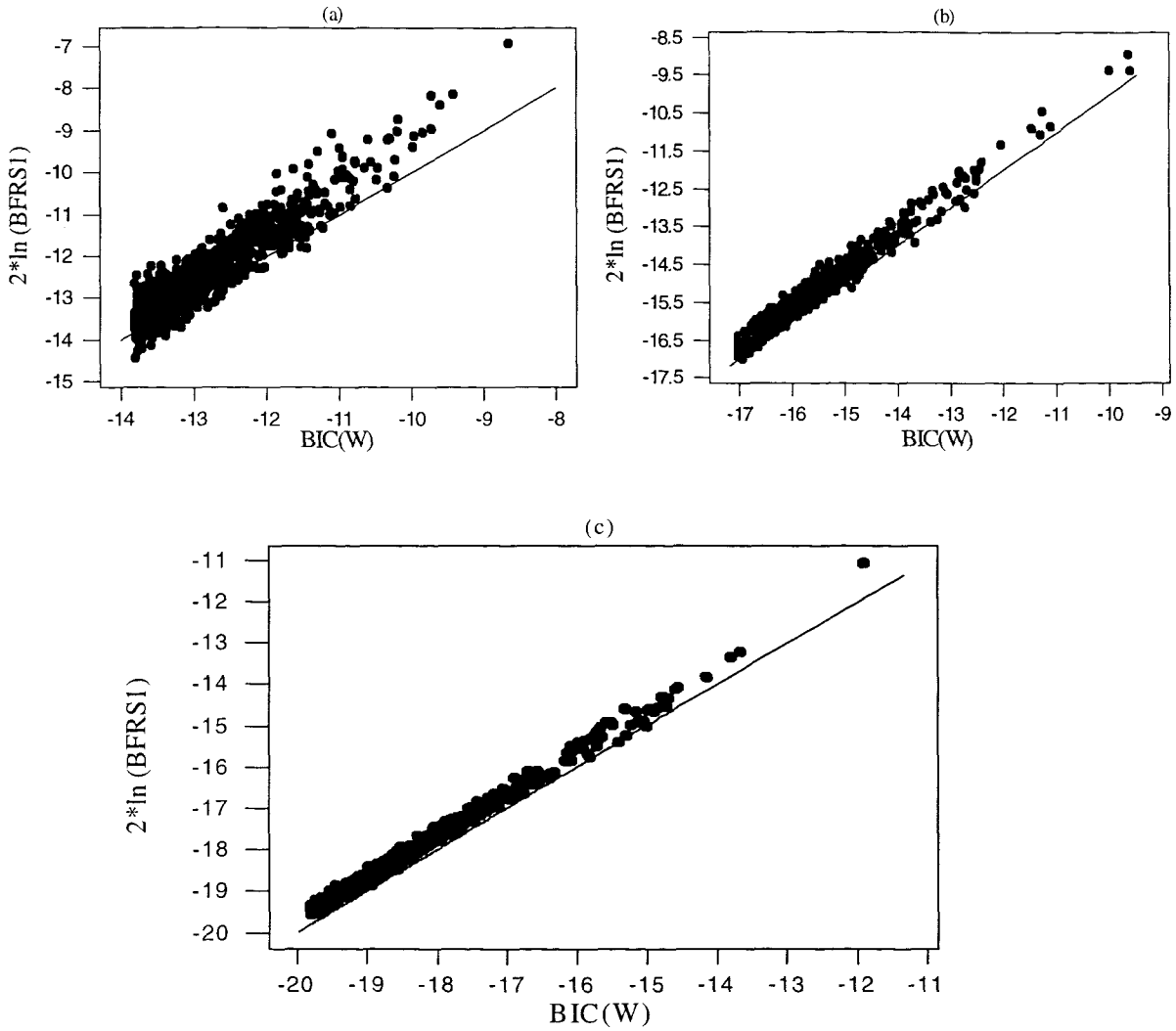

Figure 6.35: Comparison of using the first-order adjustment to the multinomial-based Bayes factor, BFRS1, for approximating the true values of $2\ln$(Bayes factor) with BIC(X^2_W), where $n_t = 100$. (a) under 10 psus, (b) 50 psus, and (c) 200 psus.

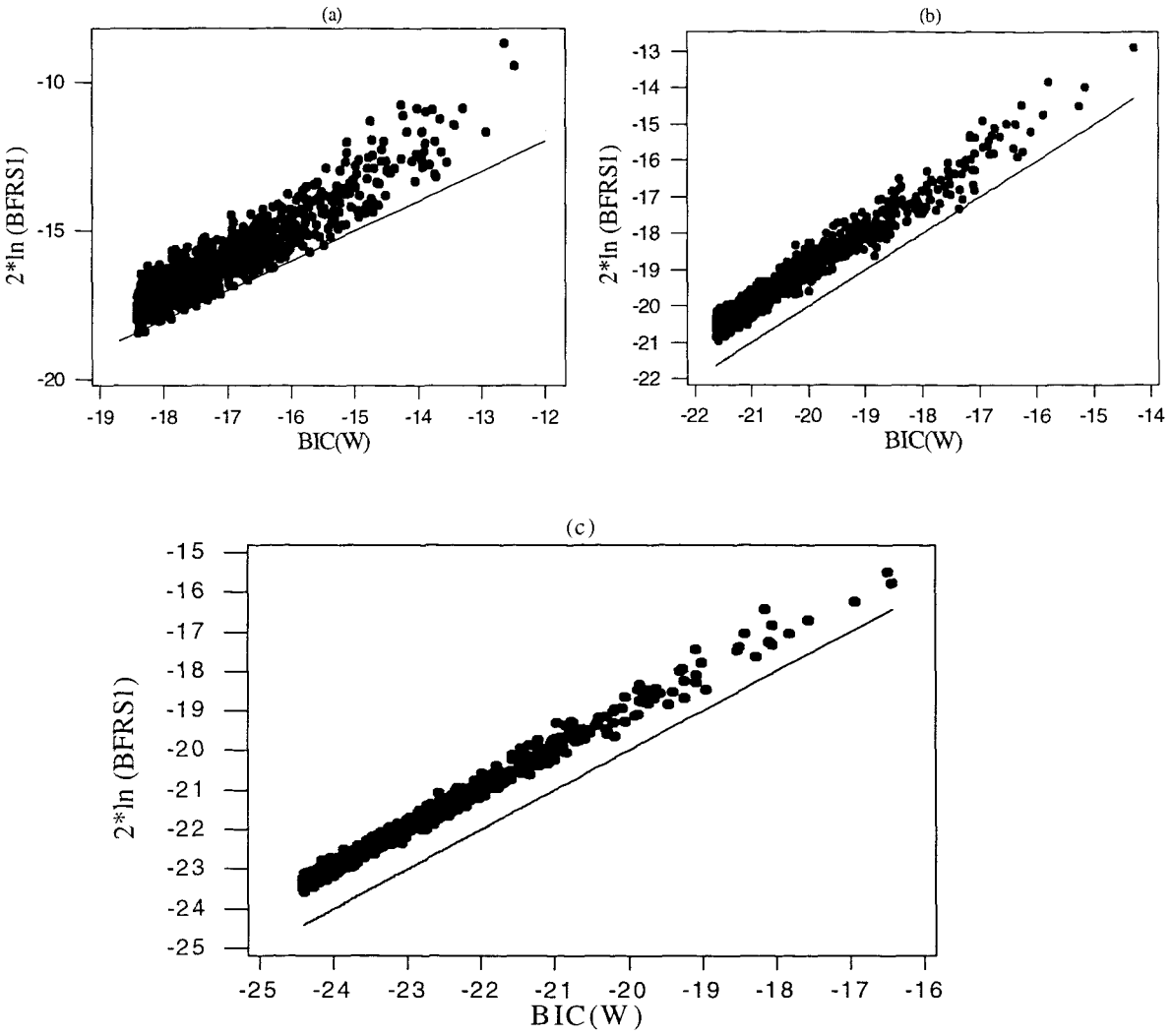

Figure 6.36: Comparison of using the first-order adjustment to the multinomial-based Bayes factor, BFRS1, for approximating the true values of $2\ln$(Bayes factor) with BIC(X^2_W), where $n_t = 1000$. (a) under 10 psus, (b) 50 psus, and (c) 200 psus.

6.7 Conclusion

In this chapter, we have demonstrated the effect of the cluster sampling scheme on Bayesian model selection, and somewhat on the classical hypotheses testing,

through the Pearson chi-squared and the Wald statistics.

We have, also, demonstrated how the Bayesian approach for drawing simulations from a complex Bayesian posterior distribution, $f(\mathbf{p}^t|\mathbf{n}^t, \boldsymbol{\alpha}, \lambda)$, can be performed via a MCMC algorithm. We have used a hybrid MCMC strategy, which consists of a combination of two algorithms, the Gibbs sampler and the Metropolis-Hastings algorithm. Moreover, we have advocated one particular strategy for improving the MCMC algorithm which enables it to converge to the target distribution more quickly. This is very important, especially, if the computer program has complex computations with long running times, such as the examples considered.

As discussed in section (3.5), the BIC approximation has error of magnitude $O(1)$. This indicates that the approximation is somewhat crude, because the error does not vanish even with infinite amount of data. Kass and Wasserman (1995) indicate that if the prior choice is consistent with a particular interpretation of "unit prior information", then the error is of order $O(n^{-\frac{1}{2}})$ rather than $O(1)$. This does not appear to be the case with our definition of "unit prior information", under the cluster sampling scheme, since the error does not vanish or get smaller as n get larger. In fact, the error seems to get larger as n increases. This argument is supported by figures (6.20) and (6.21), if compared with figures (6.23) and (6.24). Nevertheless, our results seem to be consistent with Raftery's (1996) conclusion from his empirical result that BIC is more accurate in practice than the $O(1)$ error term would suggest.

We have presented two adjustments to the multinomial-based Bayes factor. The first adjustment is obtained by dividing the cell total, $\mathbf{n}$, by the first-order correction of Rao and Scott (1981), $\hat{\tau}_.$, and the second adjustment is dividing the cell total, $\mathbf{n}$, by the second-order correction of Rao and Scott (1981). These adjustments yield reliable estimated values for the true values of $2\ln$(Bayes factor). In fact, they seem to produce asymptotically better approximation of the values

of $2\ln$(Bayes factor) than BIC(X^2_W). The performance of both adjusted Bayes factors are nearly identical.

In our simulations we have encountered some difficulties when we have small numbers of observations in each psu, n_t, such as the case where we have 50 and 200 psus with sample size of 1000, $n_t = 20$ and 5 consequently. When n_t is small, we are uncertain about our computation for the values of $2\ln$(Bayes factor). We have some doubt that our MCMC algorithm is adequately sampling the marginal posterior distribution $pr(\mathbf{p}|\mathbf{n},M_S)$, and the need for more runs is visible. Unfortunately, even with 10000 iterations it seems that our MCMC algorithm does not reach our target distribution, the marginal posterior distribution $pr(\mathbf{p}|\mathbf{n},M_S)$. Even, if the MCMC algorithm does reach our target distribution (which we know ultimately it will) after 10000 iterations, that means our program will takes approximately one month to run once, in the case of 200 psus, using Personal Computer with Pentium processor 120MHz and 32 RAM. This lends further support to using BFRS1 for the Bayesian model selection.

We showed in section (6.5), that the result of using the adjustments to the multinomial-based Bayes factor introduce improvement to the approximation of the values of $2\ln$(Bayes factor). Moreover, those adjustments are superior to BIC(X^2_W), when each psu, n_t, has a moderate or large number of observations, since they seem to produce an asymptotically more reliable approximation of the values of $2\ln$(Bayes factor). On the other hand, when n_t is small, both approximations badly approximate the values of $2\ln$(Bayes factor). This, as we discussed above, may be related to our MCMC algorithm not reaching the target distribution. Does our MCMC algorithm reach the marginal posterior distribution $pr(\mathbf{p}|\mathbf{n},M_S)$, when n_t is small? The answer for this question is probably no. Thus, we can not consider our result, in these cases, as trustworthy. Nevertheless, if we compare the values of both $2\ln$(BFRS1) and BIC(X^2_W) for the same cases, n_t is small, figure (6.34) shows that they are almost identical, which adds further support for using BFRS1 for the Bayesian model selection,

as it is easier to calculate.

We will conclude by giving some comparisons for the requirements to compute the true values of 2 ln(Bayes factor), BIC(X^2_W), and 2 ln(BFRS1). Both 2 ln(Bayes factor) and BIC(X^2_W) require full knowledge of the cluster samples. In addition, 2 ln(Bayes factor) requires a multivariate density estimator to estimate the marginal posterior density $pr(\mathbf{p}|\mathbf{n},M_S)$, and the prior density $pr(\mathbf{p}|M_S)$, at point $\mathbf{p}_0$. Also, it requires a MCMC algorithm for drawing samples from the marginal posterior distribution, $f(\mathbf{p}^t|\mathbf{n}^t,\boldsymbol{\alpha},\lambda)$. Last but not least, it requires a good knowledge of programing skills to write and tackle floating point problems, in approximately a thousand line program. This program has a running time approximately between 6hr to 72hr, depending on the number of psus, in a Personal Computer with Pentium processor 120MHz and 32 RAM. One run is including 1000 samples and 1500 iterations, 500 of them is burn-in for the MCMC algorithm.

Now, what does the first-order adjustment to the multinomial-based Bayes factor BFRS1 require? Since, the performance of both adjusted Bayes factors are nearly identical, choosing BFRS1 is more sensible. BFRS1 requires computation of a simple correction factor, which was proposed by Rao and Scott (1981). This correction factor requires only the knowledge of variance estimates, or design effects, for individual cells. Programing wise, it requires a simple program, with about one minute running time.

Chapter 7

Tests of independence and estimation in a 2×2 table

In this chapter we discuss the effect of the survey design on tests of independence in a 2×2 contingency table, and on the corresponding Bayesian model selection procedure. First, we discuss the well known simple random sample case. Then, we discuss the effect of both stratified and cluster sampling on the results of testing independence. The second objective of this chapter is to investigate the behaviour of point estimates both in the Bayesian approach, using model averaging, and in the classical approach, using an estimate based on a pretest. We carry out a simulation to study the performance of those estimators with respect to risk.

7.1 Test of independence

In section (2.10) we presented two approaches for testing independence, a chi-squared test and a test based on log odds ratios. When the sampling design is

multinomial, then we can directly test the independence hypothesis using the chi-squared statistic in equation (2.49) or using the statistic in equation (2.54). If we have a more complex sampling design, then these statistics may not be the appropriate ones. We should try to adjust these statistics to take account of the complexity of the sampling design. For the log odds ratio test in a 2×2 table, we have estimated the variance of the log odds ratio, under the true sampling design, using a Jackknife estimate, see for example Skinner, Holt and Smith (1989, p. 53). The Jackknife method for estimating $var(\hat{\phi})$ is

$$\widehat{var}(\hat{\phi}) = \sum_{t=1}^{T} \frac{n_t - 1}{n_t} \sum_{l=1}^{n_t} (\hat{\phi}_{-tl} - \bar{\phi}_t)^2$$

where $\hat{\phi}_{-tl}$ is the estimate of ϕ, based on the data with individual l from the sampling design partition t (such as strata or cluster) removed and $\bar{\phi}_t$ is the average of $\hat{\phi}_{-tl}$ over all individuals in partition t,

$$\bar{\phi}_t = \frac{1}{n_t} \sum_{l=1}^{n_t} \hat{\phi}_{-tl}$$

For each partition, we have K different possible values of $\hat{\phi}_{-tl}$ depending on which cell the removed individual is in. So, if we now use $\hat{\phi}_{-t1}, \ldots, \hat{\phi}_{-tK}$, to denote these K possibilities, $\widehat{var}(\hat{\phi})$ becomes

$$\widehat{var}(\hat{\phi}) = \sum_{t=1}^{T} \frac{n_t - 1}{n_t} \sum_{d=1}^{K} n_{td} (\hat{\phi}_{-td} - \bar{\phi}_t)^2 \tag{7.1}$$

where $\bar{\phi}_t = \frac{1}{n_t} \sum_{d=1}^{K} n_{td} \hat{\phi}_{-td}$. As an illustration, in stratified sampling with two strata, $\hat{\phi}_{-td}$ is the value of the log odds ratio for stratum t, but with one observation l excluded from cell d, i.e. in a 2×2 table

$$\hat{\phi}_{-11} = \ln \left(\frac{[(n_{11} - 1)w_1 + n_{21}w_2][n_{14}w_1 + n_{24}w_2]}{[n_{12}w_1 + n_{22}w_2][n_{13}w_1 + n_{23}w_2]} \right),$$

for stratum $t = 1$ and cell $d = 1$, where w_1 and w_2 are stratum weights. For the log odds ratio test the design effect will be denoted by ξ, where

$$\hat{\xi} = \frac{\widehat{var}(\hat{\phi})}{var_{srs}(\hat{\phi})} \tag{7.2}$$

For the chi-squared test, Rao and Scott (1981) present a correction, $\hat{\delta}_\cdot$, to the chi-squared statistics, see equation (2.52). To compute this correction an estimate for $v_{ij}(h); i = 1, \ldots, r, j = 1, , \ldots, c$, the estimated variance of $h_{ij}(\hat{p}) = \hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j}$, is required. A Jackknife method for estimating $v_{ij}(h(\hat{p}))$ is applied. Using the same terminology for computing $\widehat{var}(\hat{\phi})$, we get

$$\begin{aligned} \hat{v}_{ij}(h(\hat{p})) &= \sum_{t=1}^{T} \frac{n_t - 1}{n_t} \sum_{l=1}^{n_t} \left(h_{-tl}(\hat{p}) - \bar{h}_t(\hat{p})\right)^2 \\ &= \sum_{t=1}^{T} \frac{n_t - 1}{n_t} \sum_{d=1}^{K} n_{td} \left(h_{-td}(\hat{p}) - \bar{h}_t(\hat{p})\right)^2 \end{aligned} \tag{7.3}$$

where $\bar{h}_t(\hat{p}) = \frac{1}{n_t}\sum_{l=1}^{n_t} h_{-tl}(\hat{p}) = \frac{1}{n_t}\sum_{d=1}^{K} n_{td} h_{-td}(\hat{p})$.

For Bayesian model selection in a 2×2 table, we will compute the approximation to BIC, using the classical Pearson chi-squared statistic, X_I^2, the corrected Pearson chi-squared statistic, by the Rao and Scott correction, X_{IRS}^2, the log odds ratio statistic, $X_{I(\phi)}^2$, and the Wald statistic using the log odds ratio, $X_{W(\phi)}^2$. We will compare them with the values of 2 ln(Bayes factor), using the multinomial Bayes factor and 2 ln(BFRS), using the multinomial Bayes factor adjusted by the Rao and Scott (1981) correction, $\hat{\delta}_\cdot$, see equation (2.52), and also adjusted by the design effect for ϕ.

The Bayes factor was computed for comparing model M_0, the independence, null, model against M_S, the saturated, alternative, model under each sampling design. This Bayes factor was calculated using the MCMC sample from the prior and posterior distributions of $\mathbf{p}$, where $p_i = \sum_{t=1}^{T} w_t p_{ti}, i = 1, \ldots, K$, extracted from a full sample of the prior and posterior distributions of $\mathbf{p}^t$. The Savage-Dickey density ratio at point ϕ_0 was then use to approximate the Bayes factor.

7.1.1 Simulation study

For computing the Bayes factor and the other classical statistics described above, we adjusted our Pascal programs that were used in chapters 4 and 6. The estimation of $var(\hat{\phi})$ and $\hat{v}_{ij}(h)$ using the Jackknife method was added as a procedure in the Pascal programs. Three sampling schemes were used; multinomial sampling or simple random sampling, stratified sampling, and cluster sampling. In all these simulations 1000 samples were drawn. In all sampling schemes, we initially considered the case of $\mathbf{p} = (0.25, 0.25, 0.25, 0.25)$, where our results should support the null, independence, model. For stratification, we consider three cases, all with two strata. These cases differ in the degree of homogeneity between strata. We concluded in section (4.3.7) that the design effect increases if the strata are inhomogeneous for testing goodness-of-fit. For cluster sampling the numbers of psus used are 10, 50, and 200.

7.1.2 Results

The results presented are based on a 1000 simulations. In the results, we discuss testing for independence. Two approaches are used for computing the approximation to the Bayesian information criterion, BIC, the Pearson chi-squared statistic, and a chi-squared statistic based on the log odds ratio. Moreover, for each approach we used uncorrected, assuming simple random sample design, and corrected, given the true sampling design, statistics. As expected both design effects give equivalent results.

7.1.3 Simple random sample

For multinomial sampling, there is no design effect, since the inference is based on this design. Therefore, we will compare the values of $2\ln$(estimated Bayes

factor), using the Savage-Dickey density ratio, and the approximation of BIC, using the classical Pearson chi-squared statistic, with 2 ln(Bayes factor). Figure (7.1-a) shows that using the Savage-Dickey density ratio will produce similar values to the 2 ln(Bayes factor). On the other hand, the error associated with the approximation of BIC is systematic and can be seen in figure (7.1-b) if compared with 2 ln(Bayes factor). From the results discussed in section (3.5) we know that this error can be of magnitude $O(1)$.

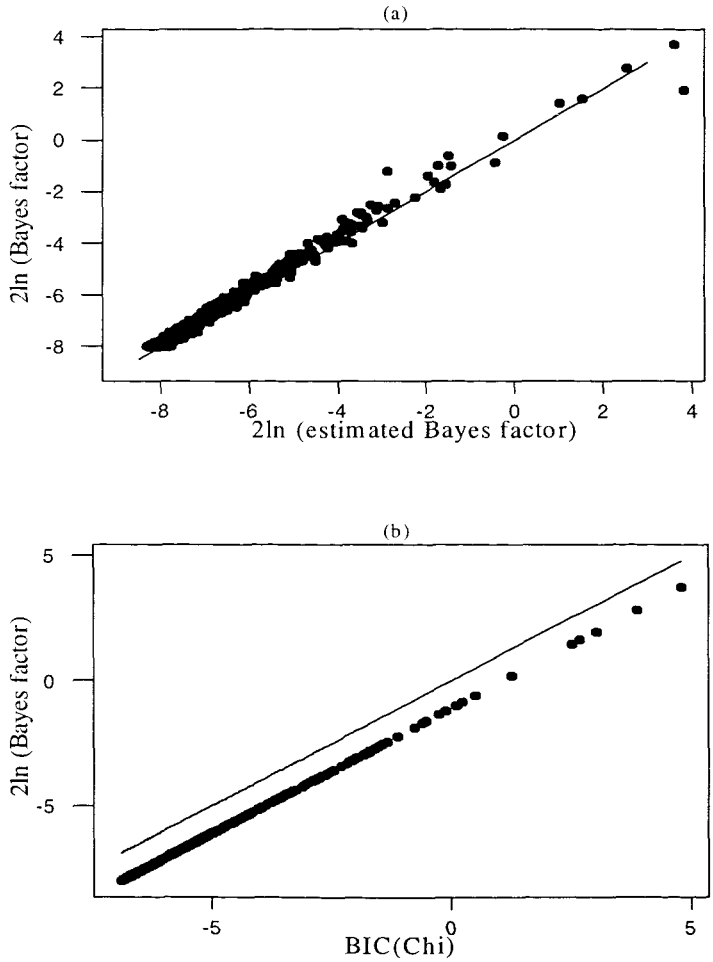

Figure 7.1: Comparison of 1000 samples for 2 ln(Bayes factor) with 2 ln(estimated Bayes factor) in (a), and with BIC(X_I^2), based on Pearson chi-squared, in (b) for a sample size of 1000.

7.1.4 Stratified sample

In the stratified sampling scheme, we consider three cases of cell probabilities, $\mathbf{p}$, each in two strata, $L = 2$. The first case is $\mathbf{p}_1 = (0.25, 0.25, 0.25, 0.25)$ and $\mathbf{p}_2 = (0.25, 0.25, 0.25, 0.25)$ with $w_l = 0.5; l = 1, 2$, the model selection method in this case should support the null, independence, model. This sampling design is effectively a multinomial sampling scheme. Thus, we expect the results to be similar, also.

Figures (7.2-a) and (7.2-b) show that the BIC(X_I^2), based on the Pearson chi-squared, and BIC(X_{IRS}^2), based on the Rao and Scott correction, have identical results if compared with 2 ln(Bayes factor). These results are equivalent with BIC based on the log odds ratio statistic using the multinomial variance, and BIC($X_{W(\phi)}^2$) based on the estimated true variance of the log odds ratio using the Jackknife method. Figure (7.2-d) shows that BIC(X_I^2) and BIC($X_{W(\phi)}^2$) are very similar. In this sampling design the Pearson chi-squared and Wald statistics are distributed as χ_1^2. On the other hand, the approximated BIC overestimates the true value of 2 ln(Bayes factor). As in the multinomial case, this error is associated with the BIC approximation. Figure (7.2-c) shows 2 ln(Uncorrected Bayes factor) has similar values to 2 ln(Bayes factor).

For the second case, the sample size used is 100 and the cell probabilities for each stratum are $\mathbf{p}_1 = (0.3, 0.2, 0.2, 0.3)$ and $\mathbf{p}_2 = (0.5, 0.1, 0.2, 0.2)$ with $w_l = 0.5; l = 1, 2$. Thus, the marginal cell probabilities are $\mathbf{q} = (0.4, 0.15, 0.2, 0.25)$. The odds ratio is equal to 3.33. This means that we should support the saturated model, M_S, and the true model is not the independence model, M_0. In this case, the BIC approximations are very similar, as illustrated by figure (7.3). The limited effect of the design is shown in figure (7.3-d).

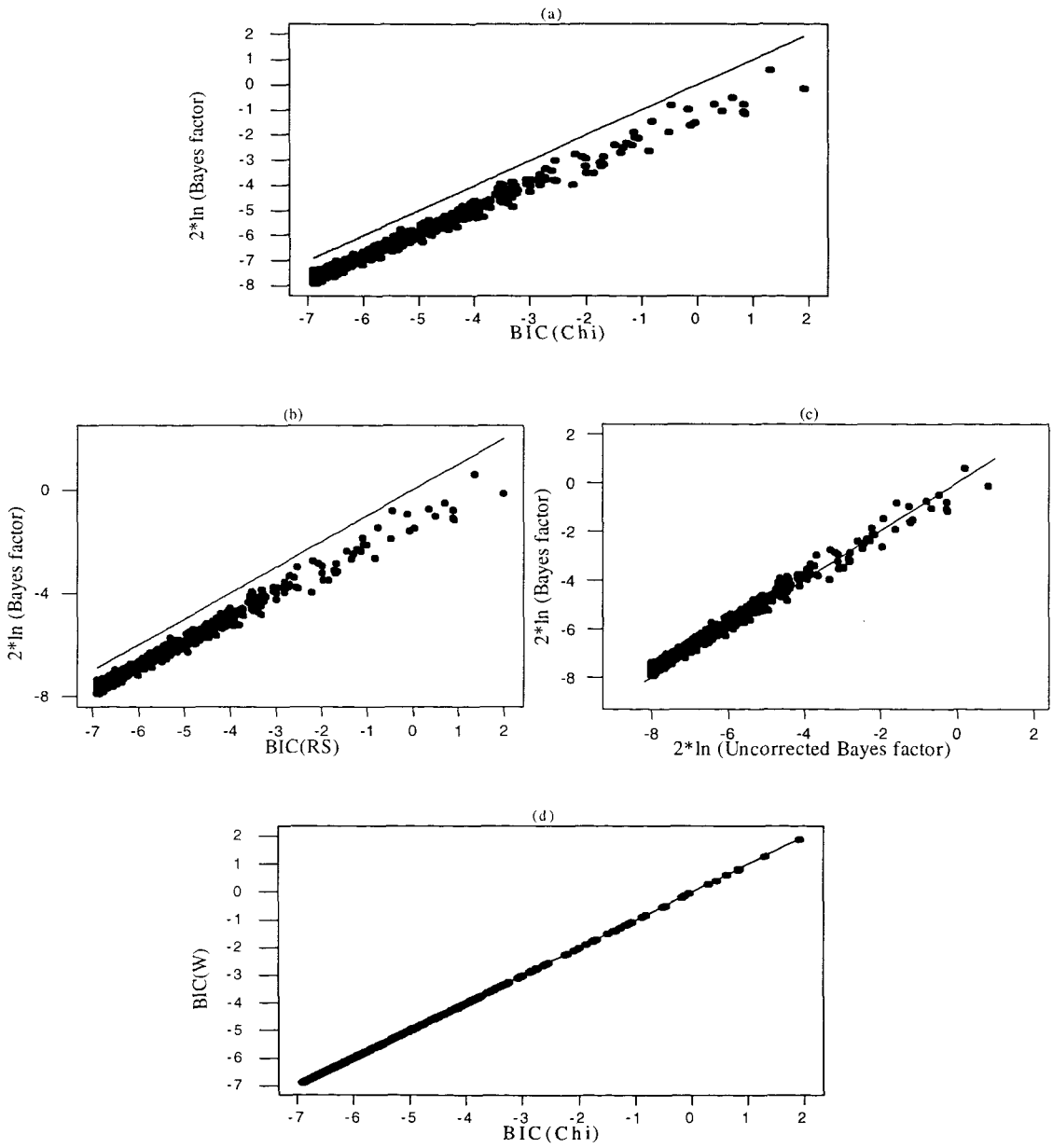

Figure 7.2: Comparison of $2\ln$(Bayes factor) values with (a) BIC(X^2_I), based on the Pearson Chi-squared, (b) BIC(X^2_{IRS}), based on Rao and Scott correction, (c) $2\ln$(uncorrected Bayes factor), and (d) BIC($X^2_{W(\phi)}$) with BIC(X^2_I), for stratified sampling design, with $p_1 = (0.25, 0.25, 0.25, 0.25)$, $p_2 = (0.25, 0.25, 0.25, 0.25)$ in a sample size of 1000.

Moreover, using $2\ln$(Uncorrected Bayes factor) or $2\ln$(BFRS), adjusted by Rao and Scott (1981) correction, will yield equivalent results, see figure (7.4).

But, if we compare them with the values of 2 ln(Bayes factor), there is a noticeable difference for the large values. The difference for these large values is probably related to the small design effect values, see figure (7.4-c) and the density estimator. The normal kernel density estimators produces sensible estimates when the point $\mathbf{p}_0$ is close to the mode of the distribution. Unfortunately, when the point $\mathbf{p}_0$ is faraway from the mode, the density estimator is not performing sensibly. This is caused by the use of a single bandwidth. In Both figures (7.3) and (7.4), as the values of 2 ln(Bayes factor) get larger than 10, where the evidence is very strongly supporting the saturated model M_S, the discrepancy between 2 ln(Bayes factor) and BIC gets considerably larger. In figure (7.4-c), the histogram presents the wide range of the design effect values. For this small sample size there is a large variation in design effect values and the Bayes factor favours the null model approximately 300 times out of 1000 samples indicating uncertainty in the model selection. If we increase the sample size to 1000, then they all favour the alternative model.

In the third case, the strata are highly inhomogeneous, and the cell probabilities for each stratum are $\mathbf{p}_1 = (0.89, 0.05, 0.05, 0.01)$ and $\mathbf{p}_2 = (0.05, 0.89, 0.05, 0.01)$ with $w_l = 0.5; l = 1, 2$. This will produce marginal cell probabilities $\mathbf{q} = (0.47, 0.47, 0.05, 0.01)$. The odds ratio is equal to 0.2. This means, as in the second case, we should supported the alternative model, that the true model is not the independent model. Nevertheless, the strata in this case are highly inhomogeneous. Thus, we expect a large design effect. Our results in figure (7.5) and figure (7.6) show the effect of the sampling design in all statistics used. In fact, if we compare the highly inhomogeneous case in figure (7.5-d) with figure (7.3-d) and the homogeneous case in figure (7.2-d), clearly the degree of design effect is increasing as the strata become more inhomogeneous. Moreover, in this case BIC(X^2_{IRS}) and 2 ln(BFRS) gave a slightly better approximation, figure (7.5-b) and figure (7.6-b). If we compare BIC(X^2_I) and BIC(X^2_{IRS}) with 2 ln(BFRS), the values of BIC(X^2_I) generally underestimate the values of 2 ln(BFRS), see figure (7.6-d), while the values of BIC(X^2_{IRS}) underestimate the large values of

2 ln(BFRS), see figure (7.6-c).

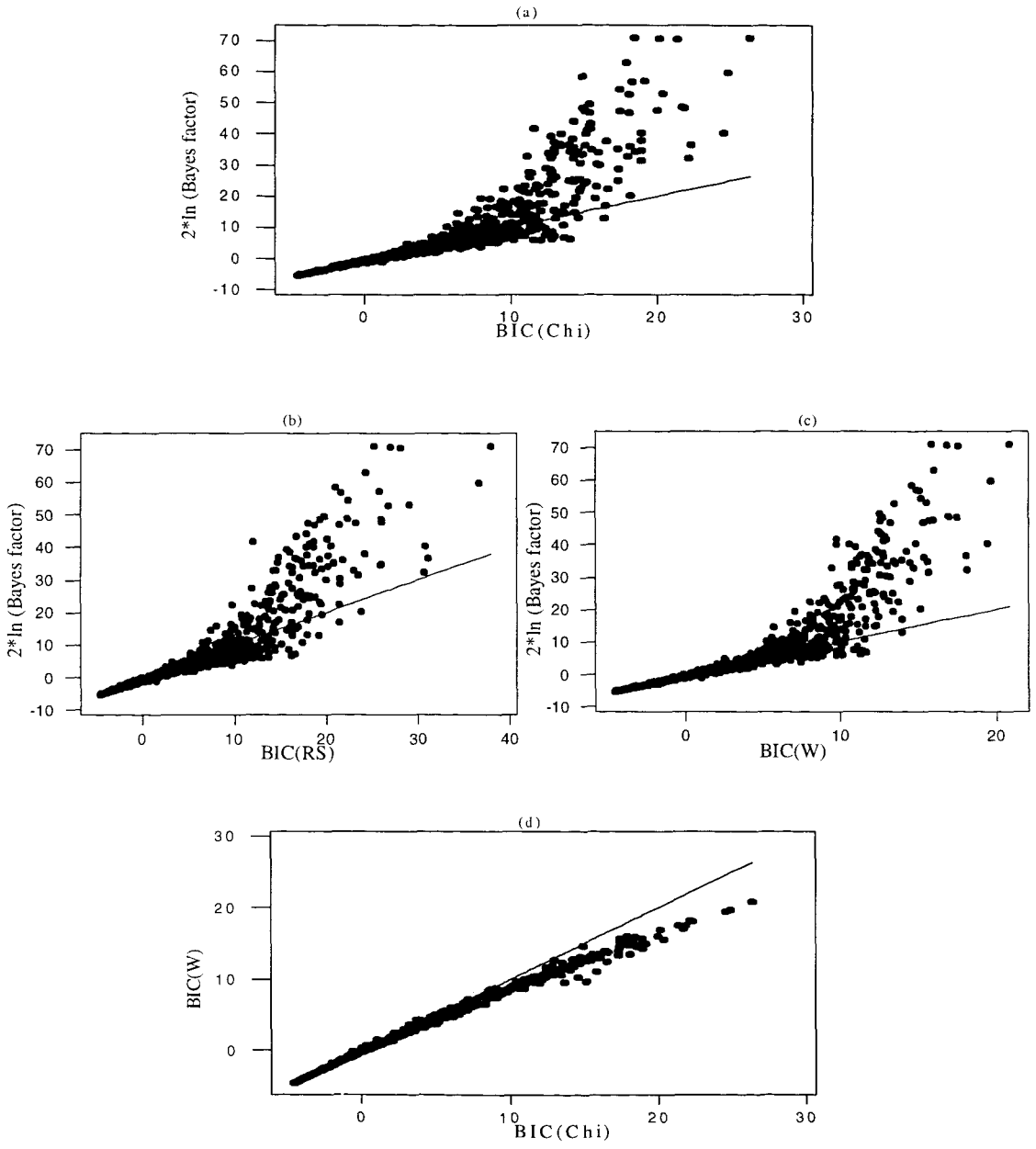

Figure 7.3: Comparison of 2 ln(Bayes factor) values with (a) BIC(X_I^2), based on the Pearson Chi-squared, (b) BIC(X_{IRS}^2), based on Rao and Scott correction, (c) BIC($X_{W(\phi)}^2$, and (d) BIC($X_{W(\phi)}^2$) with BIC(X_I^2), for stratified sampling design, with $\mathbf{p}_1 = (0.3, 0.2, 0.2, 0.3)$, $\mathbf{p}_2 = (0.5, 0.1, 0.2, 0.2)$ and a sample size of 100.

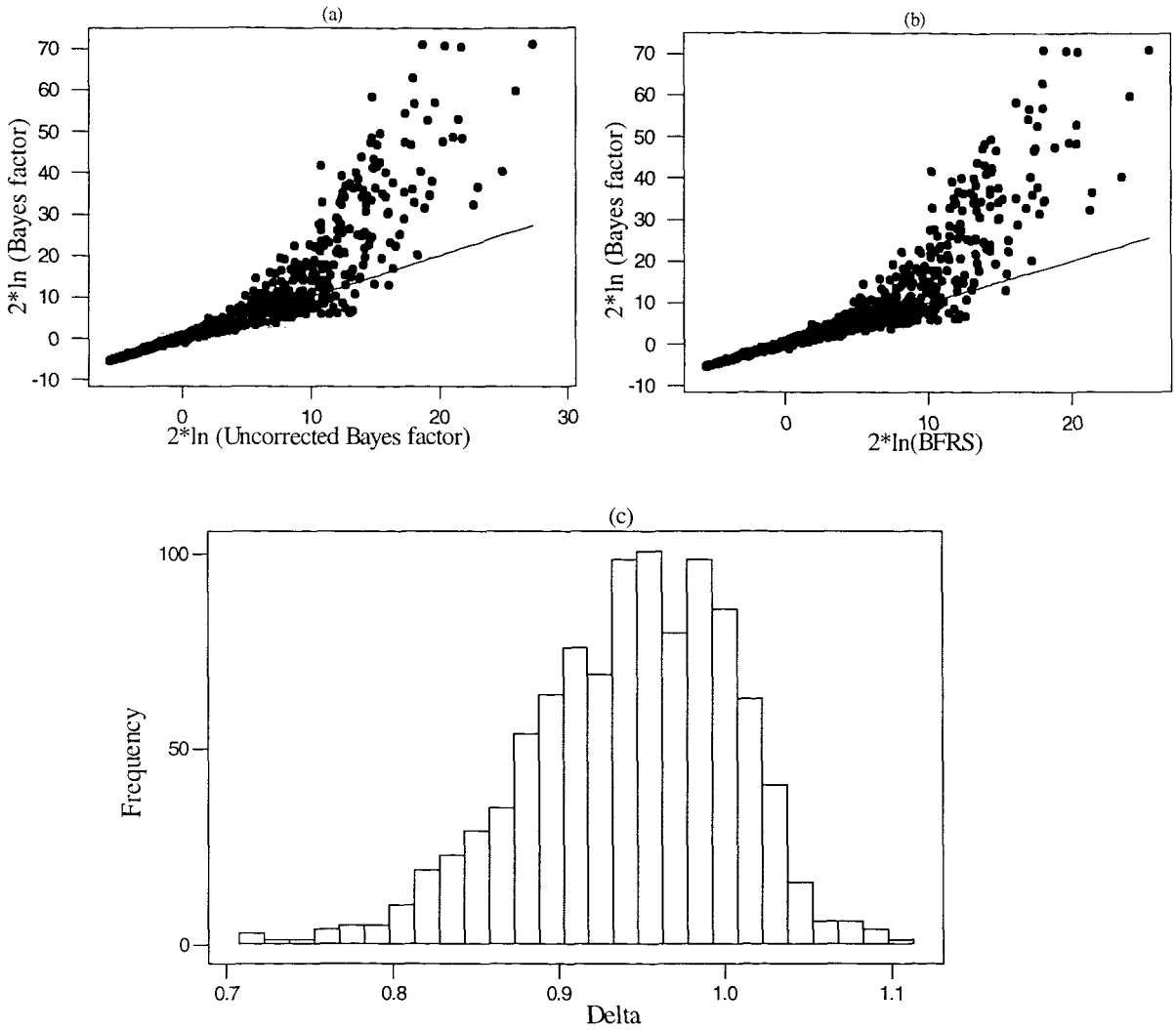

Figure 7.4: Comparison of $2\ln$(Bayes factor) values with (a) $2\ln$(Uncorrected Bayes factor), multinomial-based Bayes factor, (b) $2\ln$(BFRS), first adjustment of the multinomial-based Bayes factor, and (c) Histogram for the values of the first-order Rao and Scott correction, $\hat{\delta}_.$, for stratified sampling design, with $\mathbf{p}_1 = (0.3, 0.2, 0.2, 0.3)$, $\mathbf{p}_2 = (0.5, 0.1, 0.2, 0.2)$ and a sample size of 100.

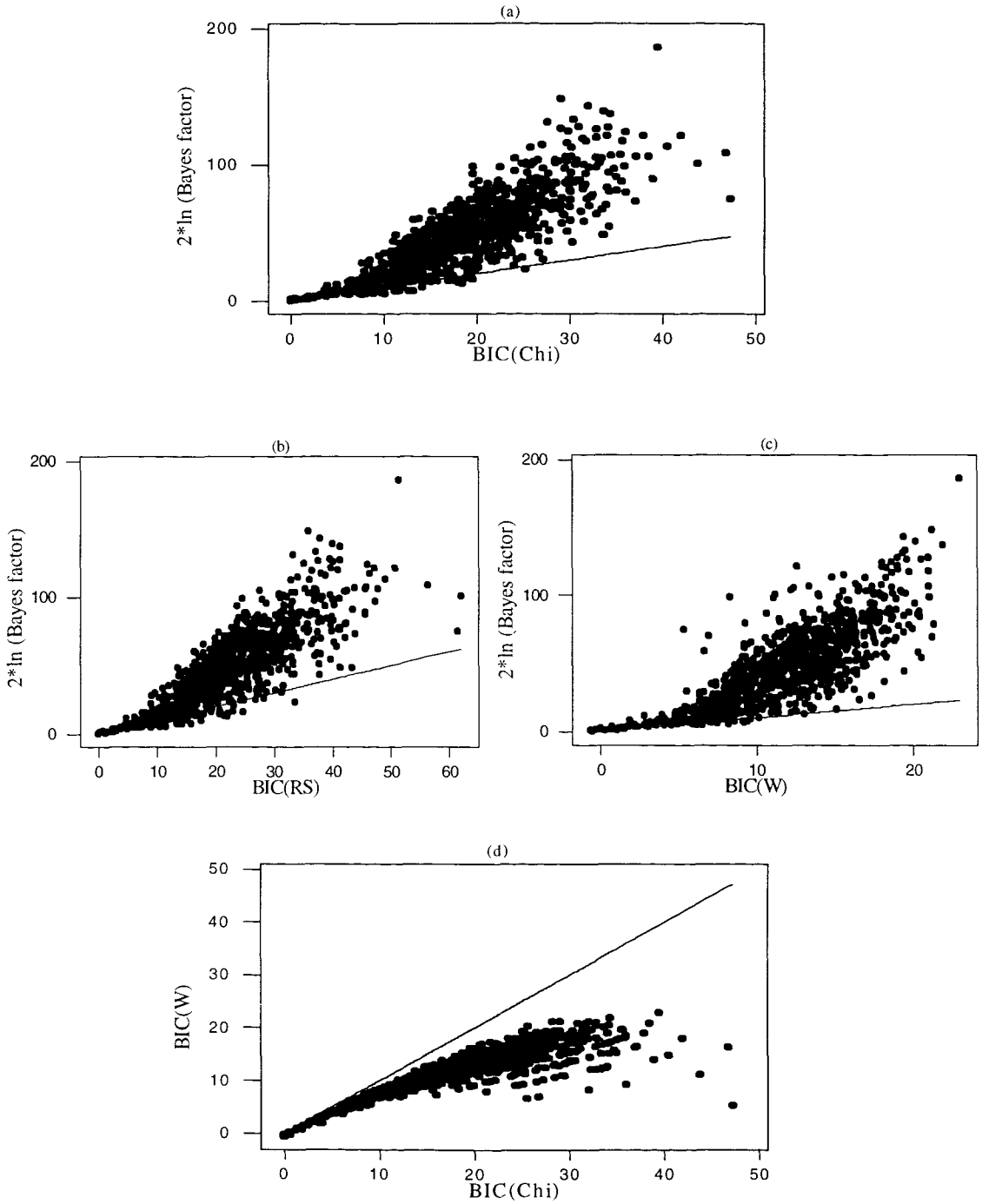

Figure 7.5: Comparison of $2\ln$(Bayes factor) values with (a) BIC(X^2_I), based on the Pearson Chi-squared, (b) BIC(X^2_{IRS}), based on Rao and Scott correction, (c) BIC($X^2_{W(\phi)}$), and (d) BIC($X^2_{W(\phi)}$) with BIC(X^2_I), for stratified sampling design, with $\mathbf{p}_1 = (0.89, 0.05, 0.05, 0.01)$, $\mathbf{p}_2 = (0.05, 0.89, 0.05, 0.01)$ and a sample size of 1000.

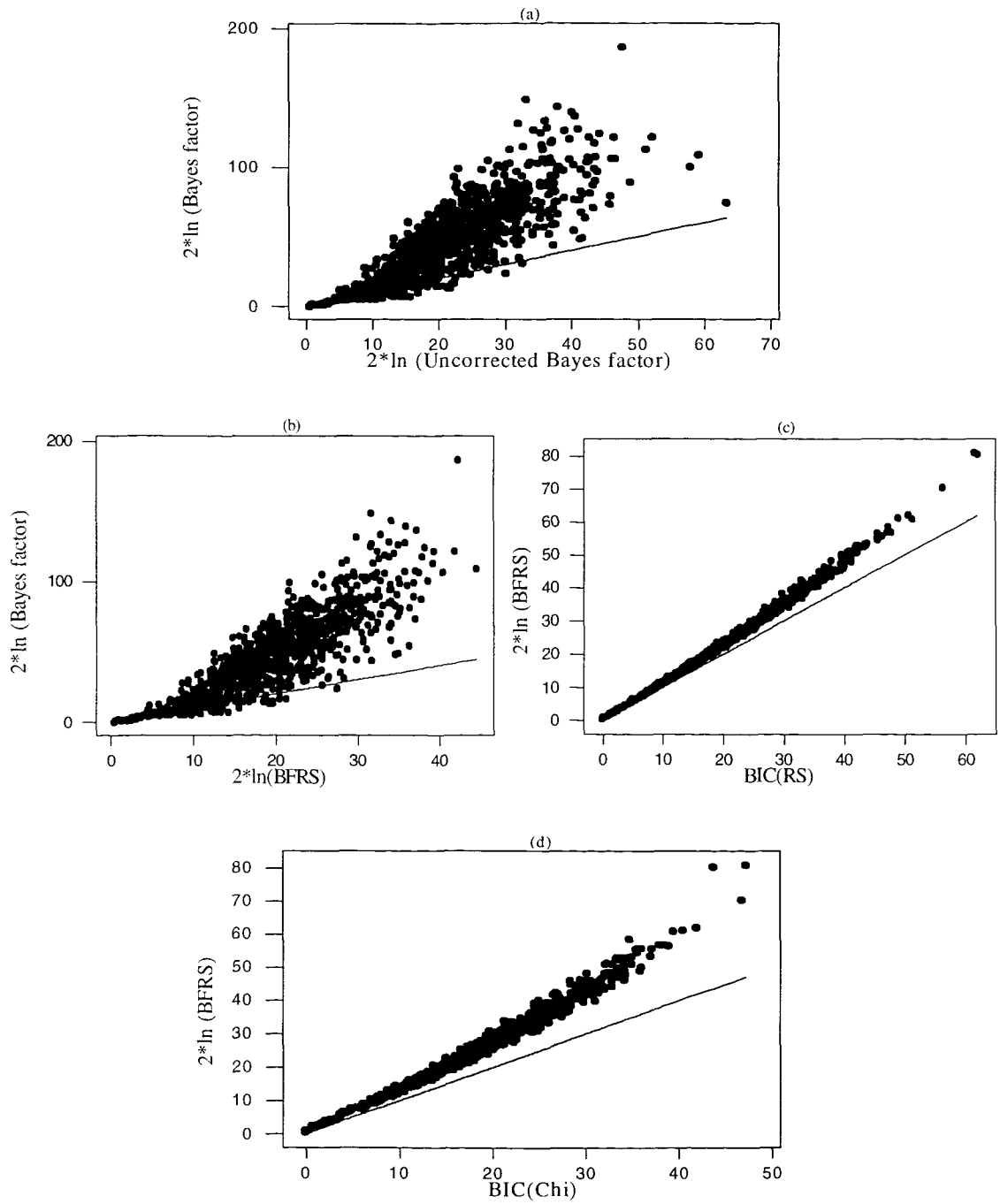

Figure 7.6: Comparison of $2\ln$(Bayes factor) values with (a) $2\ln$(Uncorrected Bayes Factor), multinomial-based Bayes factor, (b) $2\ln$(BFRS), first adjustment of the multinomial-based Bayes factor, (c) $2\ln$(BFRS) with BIC(X^2_{IRS}), and (d) $2\ln$(BFRS) with BIC(X^2_I), for stratified sampling design, with $\mathbf{p}_1 = (0.89, 0.05, 0.05, 0.01)$, $\mathbf{p}_2 = (0.05, 0.89, 0.05, 0.01)$ and a sample size of 1000.

7.1.5 Cluster sampling

For cluster sampling, we will consider the same numbers of psus as before, i.e. $c = 10, 50$, and 200. We, also, consider $\boldsymbol{\alpha} = (0.25, 0.25, 0.25, 0.25)$. Thus, the outcome of all statistics used should support the null , independence, model. In the case where we have 10 psus and a sample size of 1000, $n_t = 100$, the effect on the MCMC is apparent. Nevertheless, the error associated with the approximation of BIC can be seen in figure (7.7-a) if compared with figure (7.7-b) and figure (7.7-d). Figure (7.7-c) shows the values of the design effect distributed around 0.983. Therefore, ignoring the sampling design will not have strong effect on the inference.

This can be demonstrated by increasing the sample size to 10000, i.e. $n_t = 1000$, see figure (7.8). Figure (7.8-d), shows the similarity between the values of the Rao and Scott correction for test of independence, $\hat{\delta}_.$, and the design effects, $\hat{\xi}$, for the log odds ratio, ϕ.

If the numbers of psus is equal to 50 with a total sample size equal to a 1000, the MCMC method does not work properly. Thus, we will consider $2\ln$(BFRS) to be our baseline instead of $2\ln$(Bayes factor); This is supported by our previous results. Figure (7.9) shows the design effects on BIC(X_I^2), and on $2\ln$(Uncorrected Bayes factor). This effect is not serious, since $\hat{\delta}_.$ values are distributed around 0.91, which is still close to one, see figure (7.9-d).

If we increase the sample size to 5000, the histogram for the values of $\hat{\delta}_.$ in figure (7.10), where the average values of $\hat{\delta}_.$ equal 0.98, indicates no noticeable design effect.

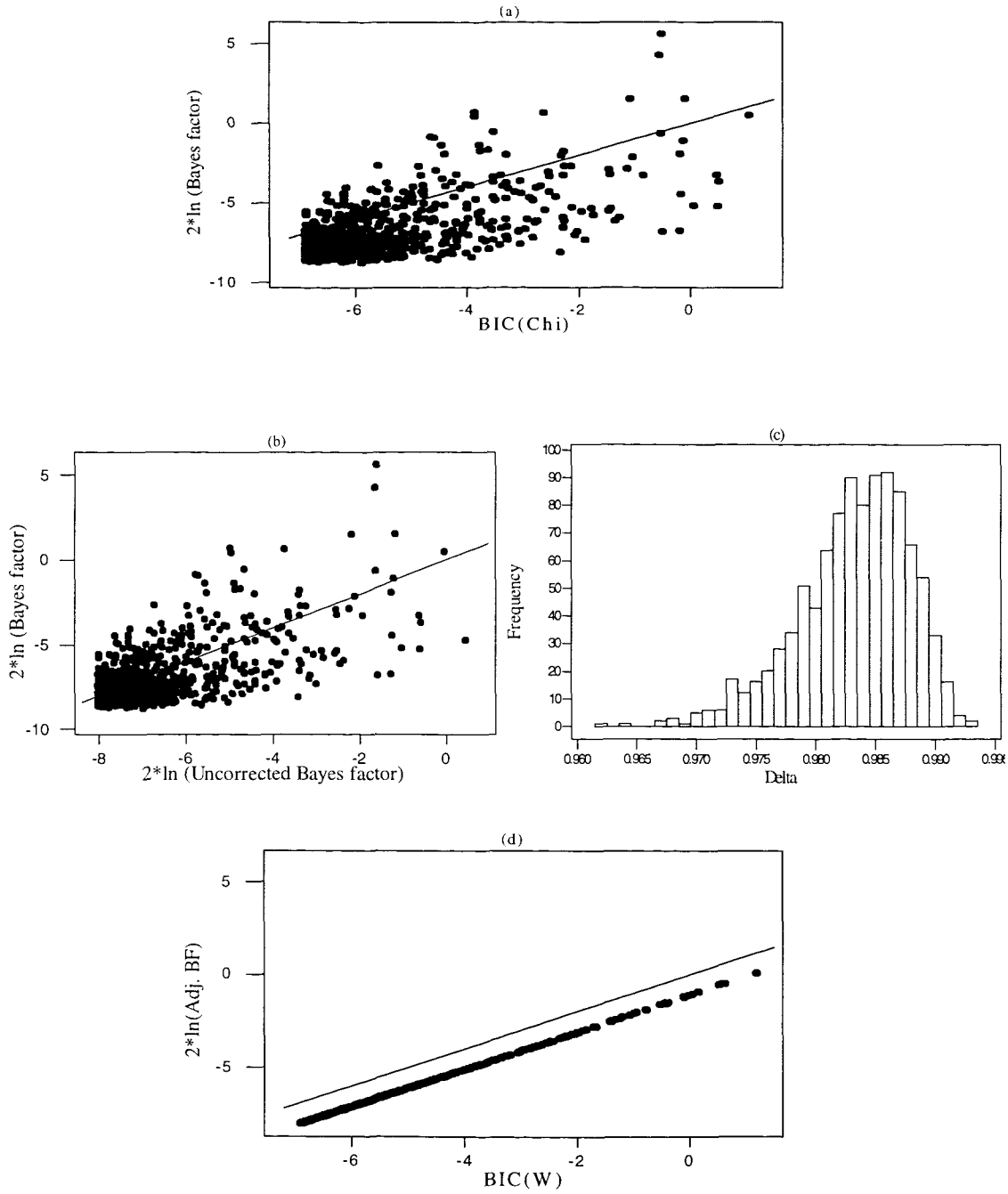

Figure 7.7: Comparison for cluster sampling design where we have 10 psus and sample size of 1000, (a) $2\ln$(Bayes factor) values with BIC(X_I^2), based on the Pearson Chi-squared, (b) $2\ln$(Bayes factor) values with BIC($X^2_{W(\phi)}$), based on Wald statistic for the odd ratio, (c) Histogram for values of $\hat{\delta}$ for the 1000 samples, and (d) $2\ln$(Adjusted Bayes factor) values, based on $\hat{\xi}$, with BIC($X^2_{W(\phi)}$), based on Wald statistic.

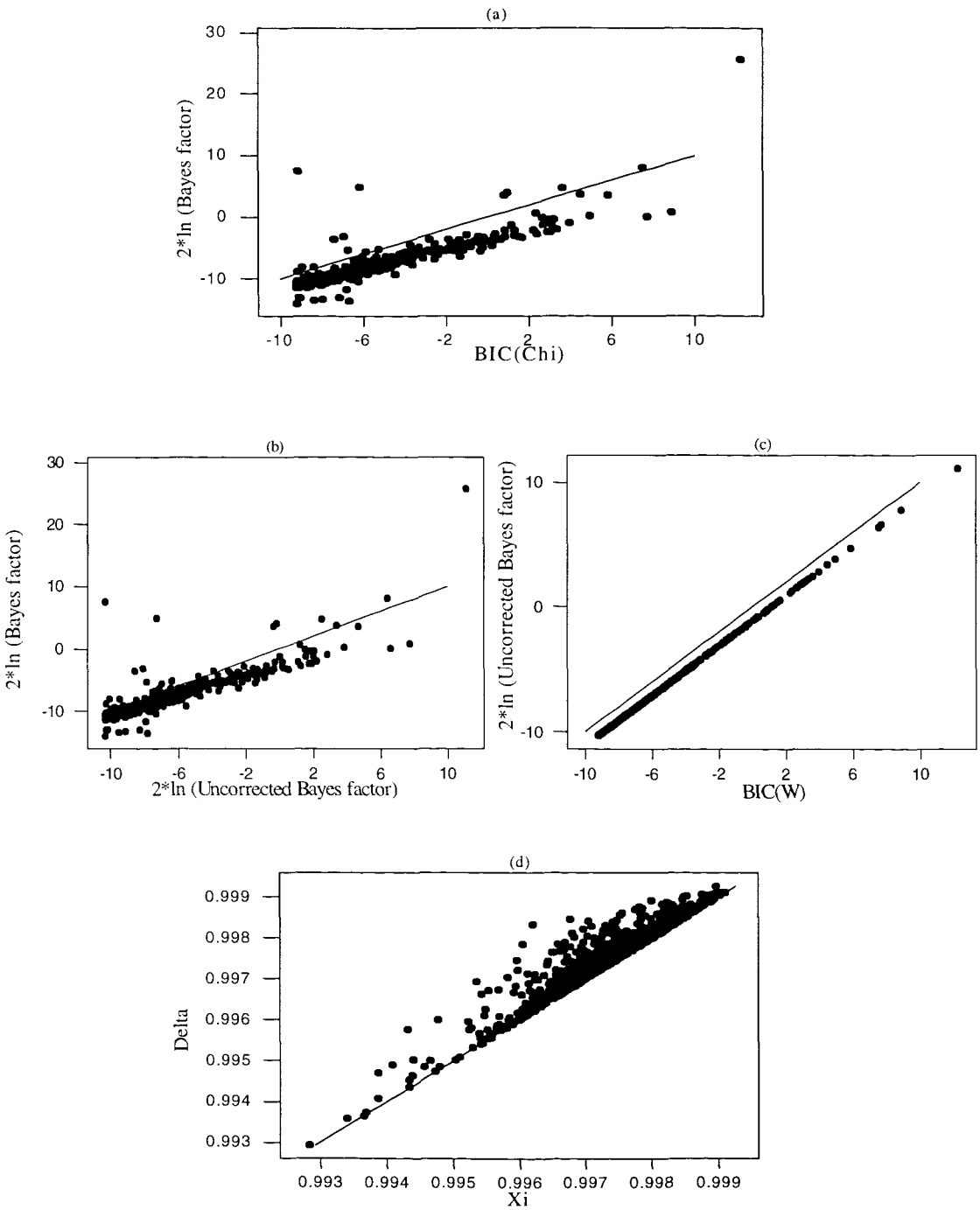

Figure 7.8: Comparison for cluster sampling design where we have 10 psus each with 1000 observations, (a) $2\ln$(Bayes factor) values with BIC(X_I^2), (b) $2\ln$(Bayes factor) values with $2\ln$(Uncorrected Bayes factor) values, (c) $2\ln$(Uncorrected Bayes factor) values with BIC($X^2_{W(\phi)}$), and (d) values of $\hat{\delta}_.$ for the 1000 samples with $\hat{\xi}$ values.

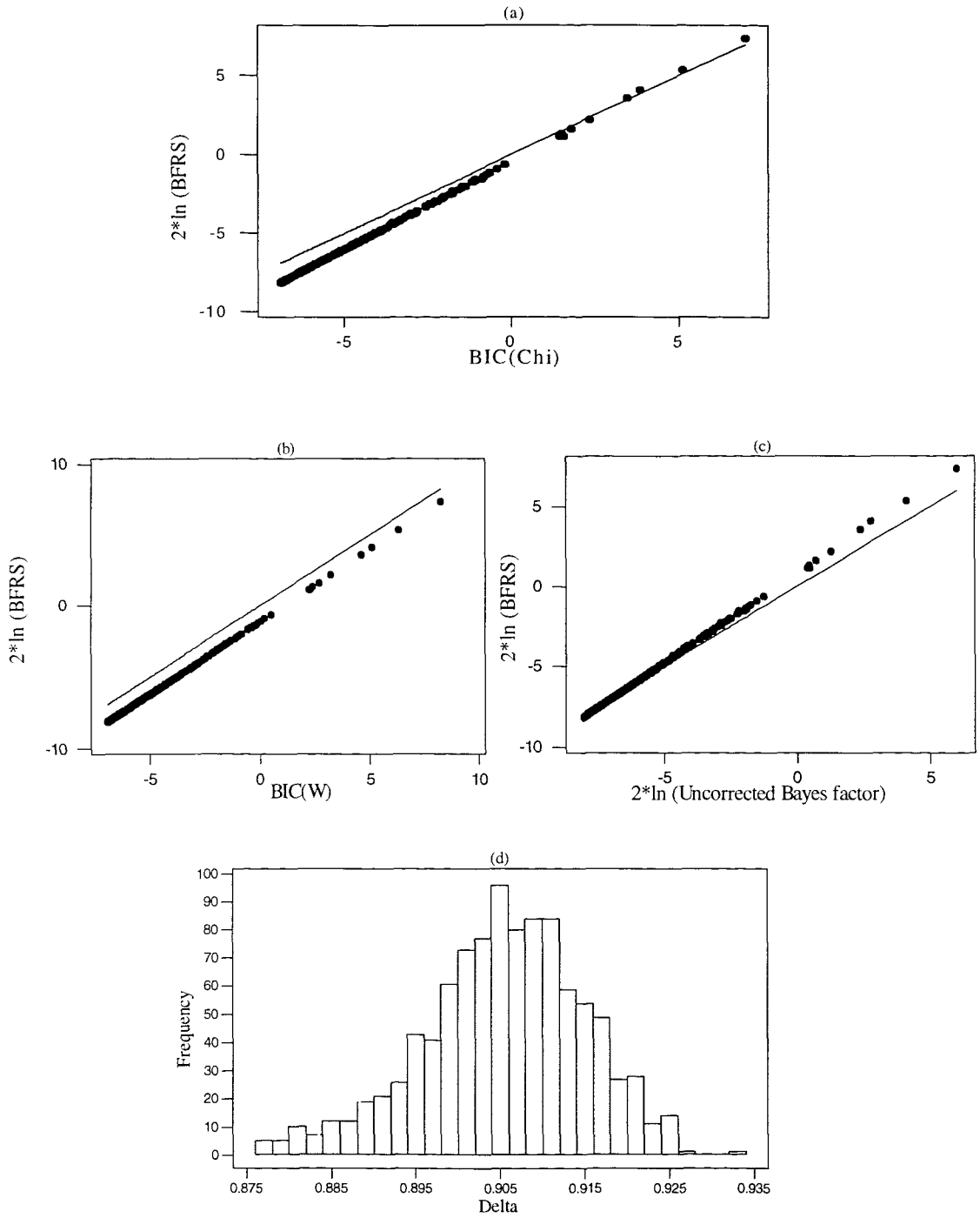

Figure 7.9: Comparison for cluster sampling design where we have 50 psus for a sample size of 1000 observations, (a) $2\ln(\text{BFRS})$ values with $\text{BIC}(X_I^2)$, (b) $2\ln(\text{BFRS})$ values with $\text{BIC}(X^2_{W(\phi)})$ values, (c) $2\ln(\text{BFRS})$ values with $2\ln(\text{Uncorrected Bayes factor})$ values, and (d) a histogram for the values of $\hat{\delta}$ for the 1000 samples.

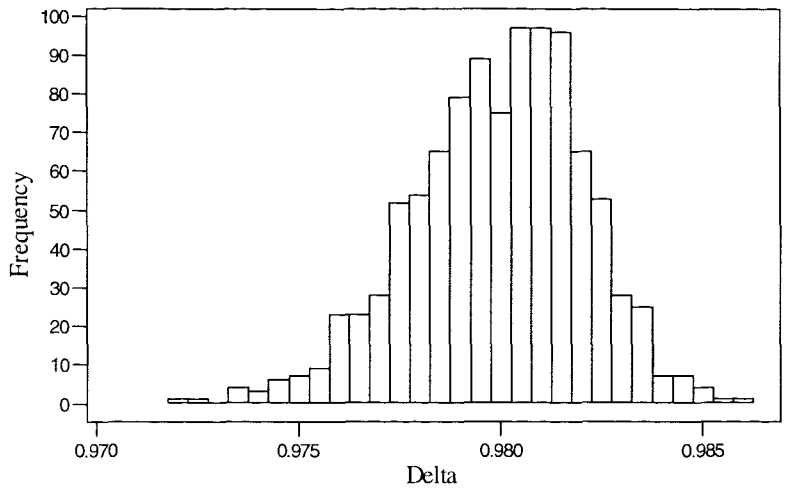

Figure 7.10: A histogram of the values of $\hat{\delta}_{.}$ in the 1000 samples, where number of psus are 50 and a sample size of 5000.

As the sample size gets even larger (50000) it seems that 2 ln(Uncorrected Bayes factor) has almost identical values to 2 ln(Bayes factor), see figure (7.11-a). In this case the values of $\hat{\delta}_{.}$ are distributed around 0.997. Therefore, asymptotically it seems that ignoring the sampling design will not have effect in the statistical inference, except the natural error associated with the approximation of BIC, figure (7.11-b). With a large sample size, such as 50000, the Pearson chi-squared statistic and the Wald statistic, which have identical values, do not seem to be distributed as χ^2_1. Figure (7.11-c) shows how the Wald statistic is greater than would be predicated by the χ^2_1 distribution in figure (7.11-d).

When we increase the number of psus to 200 with a sample size of 1000, i.e. $n_t = 5$, both values of $\hat{\delta}_{.}$ and $\hat{\xi}$ are distributed around 0.64, see figure (7.12-d) for $\hat{\delta}_{.}$. This sampling design affects the values of BIC(X^2_I), since it overestimates the small values of 2 ln(BFRS) and underestimates the large values, figure (7.12-a). In figure (7.12-b) the values of 2 ln(Uncorrected Bayes factor) underestimate the large values of 2 ln(BFRS). Using the Rao and Scott (1981) correction to the Pearson chi-squared will improve the approximation of BIC, but the error associated with it remains visible, see figure (7.12-c).

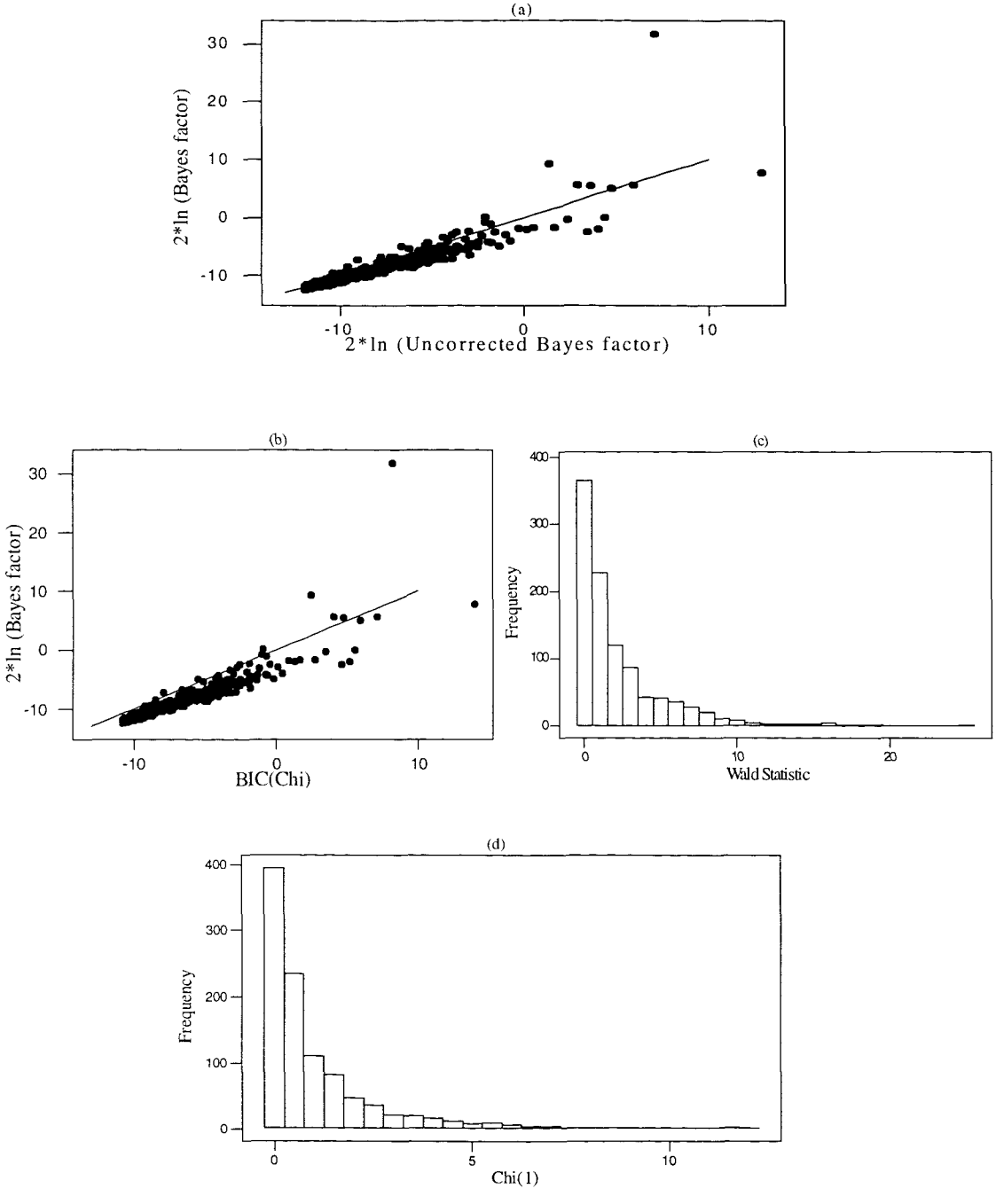

Figure 7.11: Comparison for cluster sampling design where we have 50 psus with a sample size of 50000, (a) $2\ln$(Bayes factor) values with $2\ln$(Uncorrected Bayes factor), (b) $2\ln$(Bayes factor) values with BIC(X_I^2), (c) a histogram for the values of the Wald statistic, $X^2_{W(\phi)}$, in the 1000 samples, and (d) a histogram for a 1000 observation from the χ^2 distribution with one d.f.

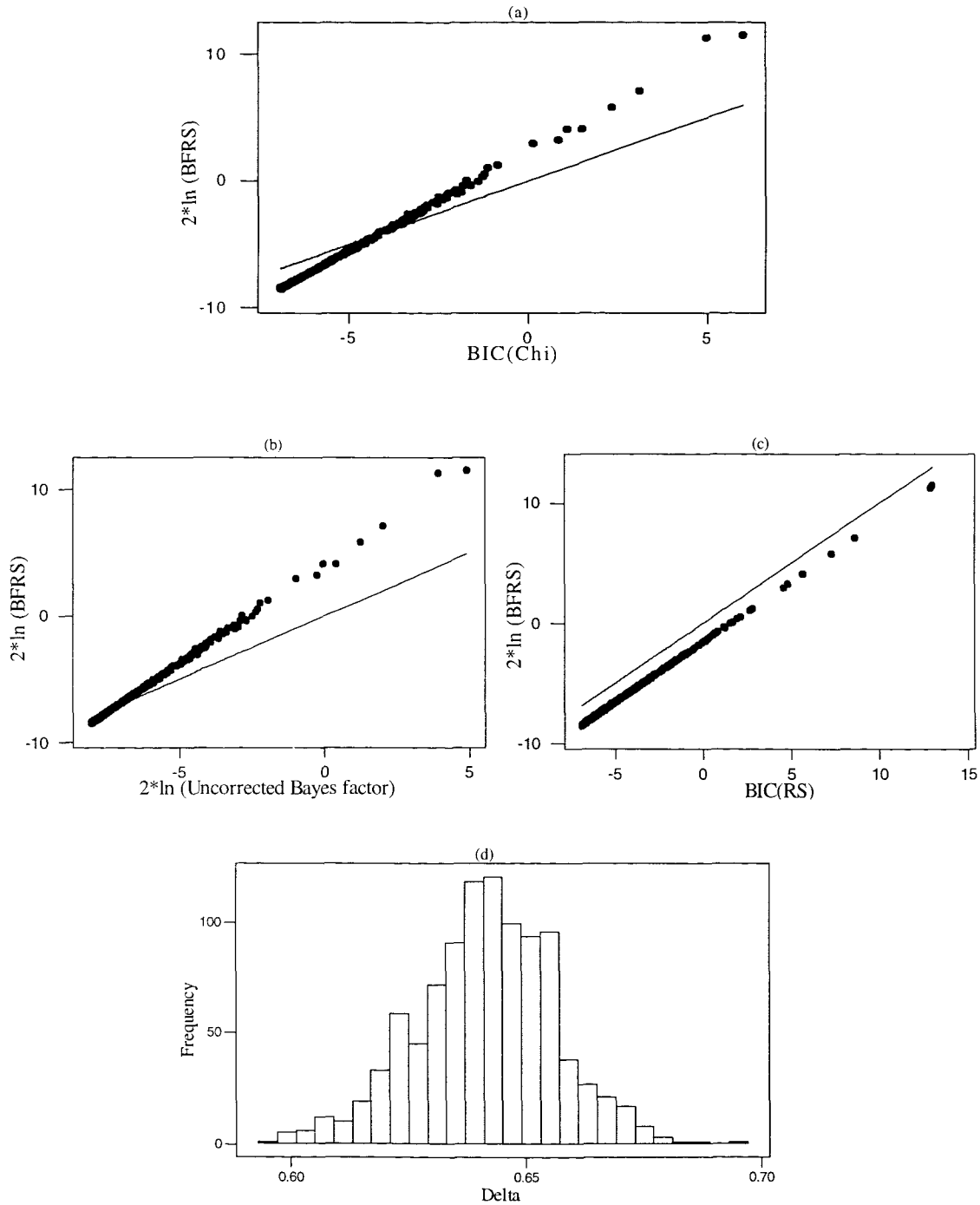

Figure 7.12: Comparison for cluster sampling design where we have 200 psus with a sample size of 1000, (a) $2\ln(\text{BFRS})$ values with $\text{BIC}(X_I^2)$, (b) $2\ln(\text{BFRS})$ values with $\text{BIC}(X^2_{W(\phi)})$ values, (c) $2\ln(\text{BFRS})$ values with $2\ln(\text{Uncorrected Bayes factor})$ values, and (d) a histogram for the values of $\hat{\delta}$. for the 1000 samples.

As the sample size increases to 20000, $n_t = 100$, the result will improve, since the values of $\hat{\delta}_.$ are now distributed close to one, with mean 0.98, see figure (7.13).

If the sample size is equal to 200000, $n_t = 1000$, $2\ln$(Uncorrected Bayes factor) has almost identical values to $2\ln$(Bayes factor), see figure (7.14-a). In this case both values of $\hat{\delta}_.$ and $\hat{\xi}$ are distributed around 0.997, see figures (7.14-c) and (7.14-d). Thus, we have an asymptotically identical result with the case of 50 psus, figure (7.14-b). Also, the values of the Pearson chi-squared statistic and the Wald statistic, which are identical, do not seem to be distributed as χ^2_1.

Figure (7.15) shows how the Pearson chi-squared statistic is larger than would be predicted by the χ^2_1 distribution in figure (7.11-d).

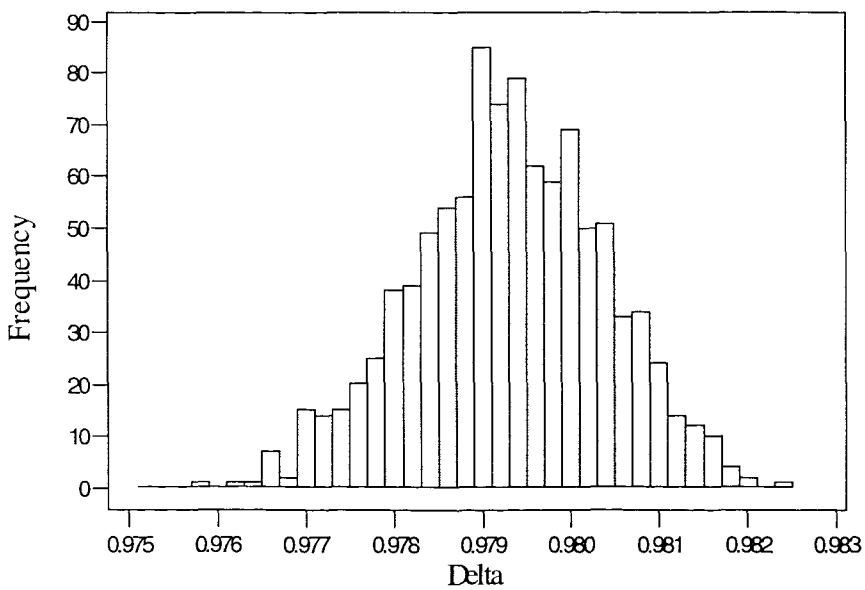

Figure 7.13: A histogram for the values of $\hat{\delta}_.$ in the 1000 samples, where number of psus are 200 and sample size of 20000.

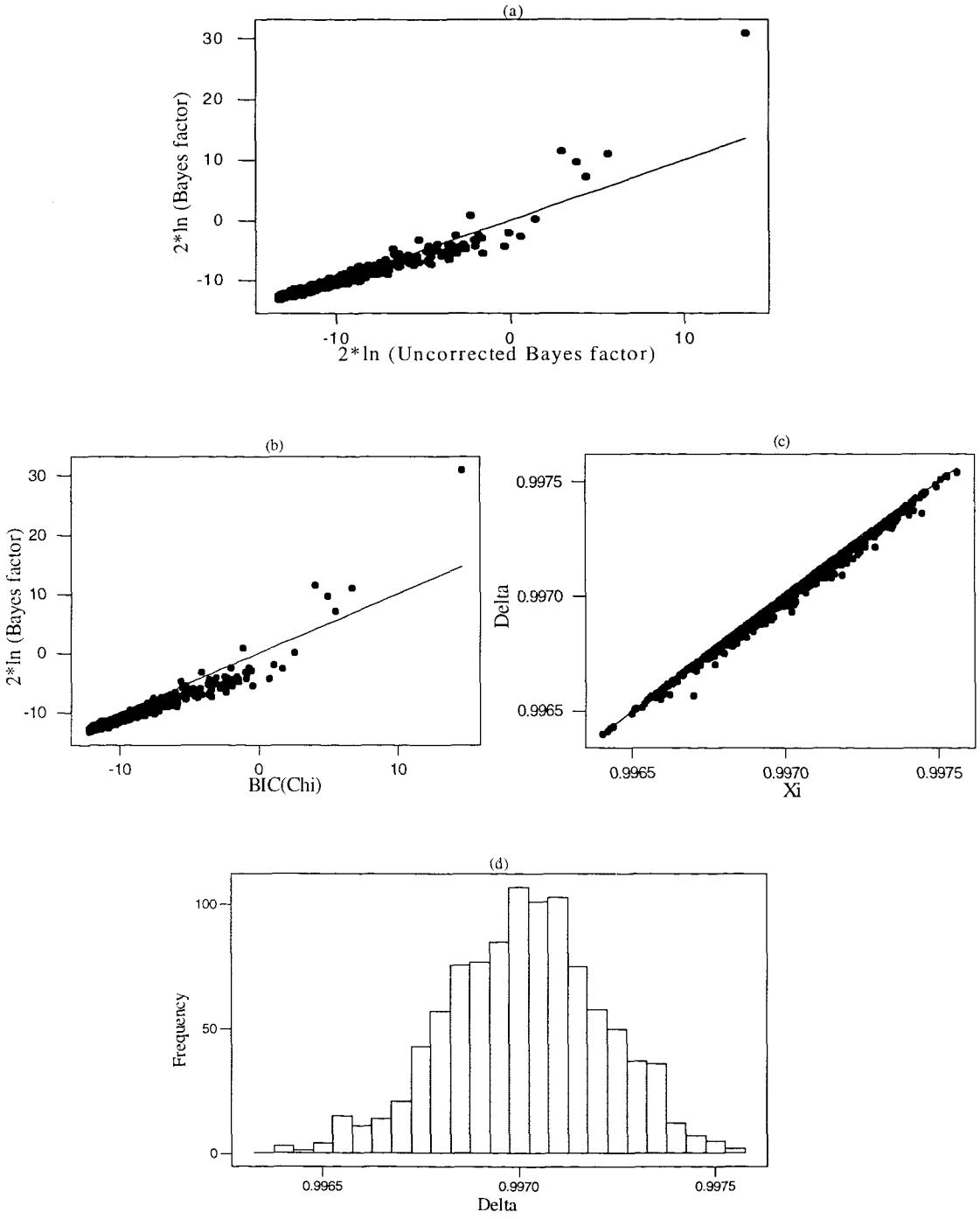

Figure 7.14: Comparison for cluster sampling design where we have 200 psus each with 1000 observations, (a) $2\ln$(Bayes factor) values with $2\ln$(Uncorrected Bayes factor) values, (b) $2\ln$(Bayes factor) values with BIC(X_I^2), (c) values of $\hat{\delta}$ for the 1000 samples with $\hat{\xi}$ values, and (d) a histogram for the values of $\hat{\delta}$ in the 1000 samples.

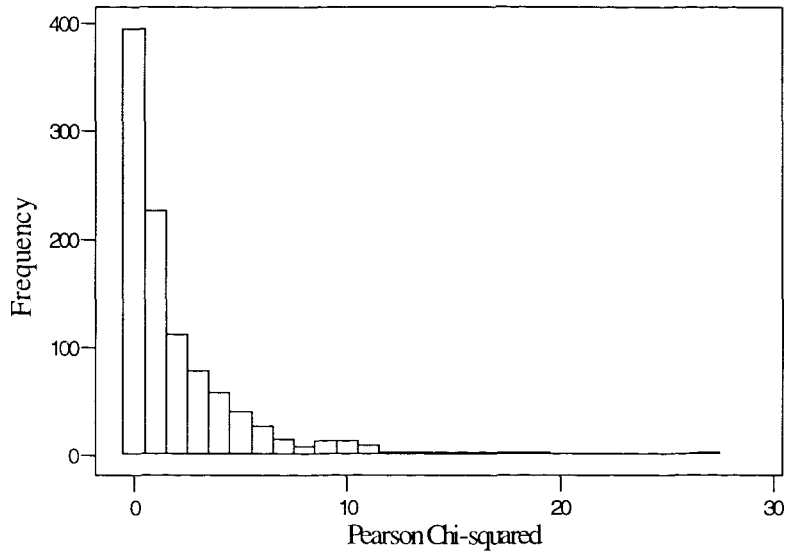

Figure 7.15: A histogram of the values of the Pearson chi-squared statistic, X_I^2, in the 1000 samples, where we have 200 psus and a sample size of 200000.

7.1.6 Discussion

In our simulation, two approaches are used for computing the approximation of Bayesian information criterion, BIC. First, we used the standard way of testing independence, based on the Pearson chi-squared statistic, X_I^2, where we apply the Jackknife method to estimate the variance $v_{ij}(h)$ for obtaining the Rao and Scott correction. Second, we used the log odds ratio, where we used the same method to estimate $var(\hat{\phi})$. Both tests, as expected, have equivalent results, in all sampling schemes considered. In those schemes, we have illustrated the effect of the sampling design on the model selection problem.

For multinomial sampling, we examined the performance of the estimated Bayes factor, using the Savage-Dickey density ratio, and the approximation based on the Bayesian information criterion, BIC. In this sampling scheme the observations are independent and identically distributed. The Savage-Dickey density ratio produced very similar results to the actual Bayes factor.

On the other hand, there are errors associated with the BIC approximation. The error associated with the BIC approximation leads to underestimates of the value of $2\ln(\text{BFRS})$ in the goodness-of-fit test and overestimates the values of $2\ln(\text{BFRS})$ in the independence test. Contrary to the results in the goodness-of-fit test, where the error gets larger as n_t increases in a cluster sample, in the independence test the error get smaller as n_t increases, the following table shows the average difference of $2\ln(\text{BFRS})$ and $\text{BIC}(X^2_{W(\phi)})$ when the number of psus is equal to 10, 50, and 200 for difference sizes of n_t,

psu\n_t	5	20	100	1000
10	_	_	-1.123	-1.115
50	_	-1.208	-1.134	-1.117
200	-1.555	_	-1.135	-1.118

Figure (7.16) shows a histogram of the difference as an example. When $n_t = 1000$, the Pearson chi-squared, X^2_I, and the Wald, $X^2_{W(\phi)}$, statistics are not distributed as χ^2_1, but nearly $2\chi^2_1$. This is a worrying result, but Nathan (1975) has similar results, in table (2) of his paper, for the log likelihood ratio and the Wald statistics. Furthermore, in the statistical literature, there is some evidence suggesting that as the sample size increase, the asymptotic size of the χ^2-test of independence, under multinomial sampling, can be larger than its nominal level (Loh, 1989). Also, there is evidence suggesting that the Wald statistic has a poor behaviour in sampling surveys, which mentioned in section (6.4.4).

In stratified sampling, if the strata are homogenous, the error associated with the approximation of BIC is visible. In this case, both the Pearson chi-squared, X^2_I, and the Wald, $X^2_{W(\phi)}$, statistics are distributed as χ^2_1, under the null model. If the strata are inhomogeneous, the error associated with the approximation of BIC is not so visible. Figure (7.17) if compared with figure (7.6-c) shows the effect of the sampling design in highly inhomogeneous strata, where the Rao and Scott correction, $\hat{\delta}_.$, is distributed around 0.88. In this case, our true model is

not the null, independence, model.

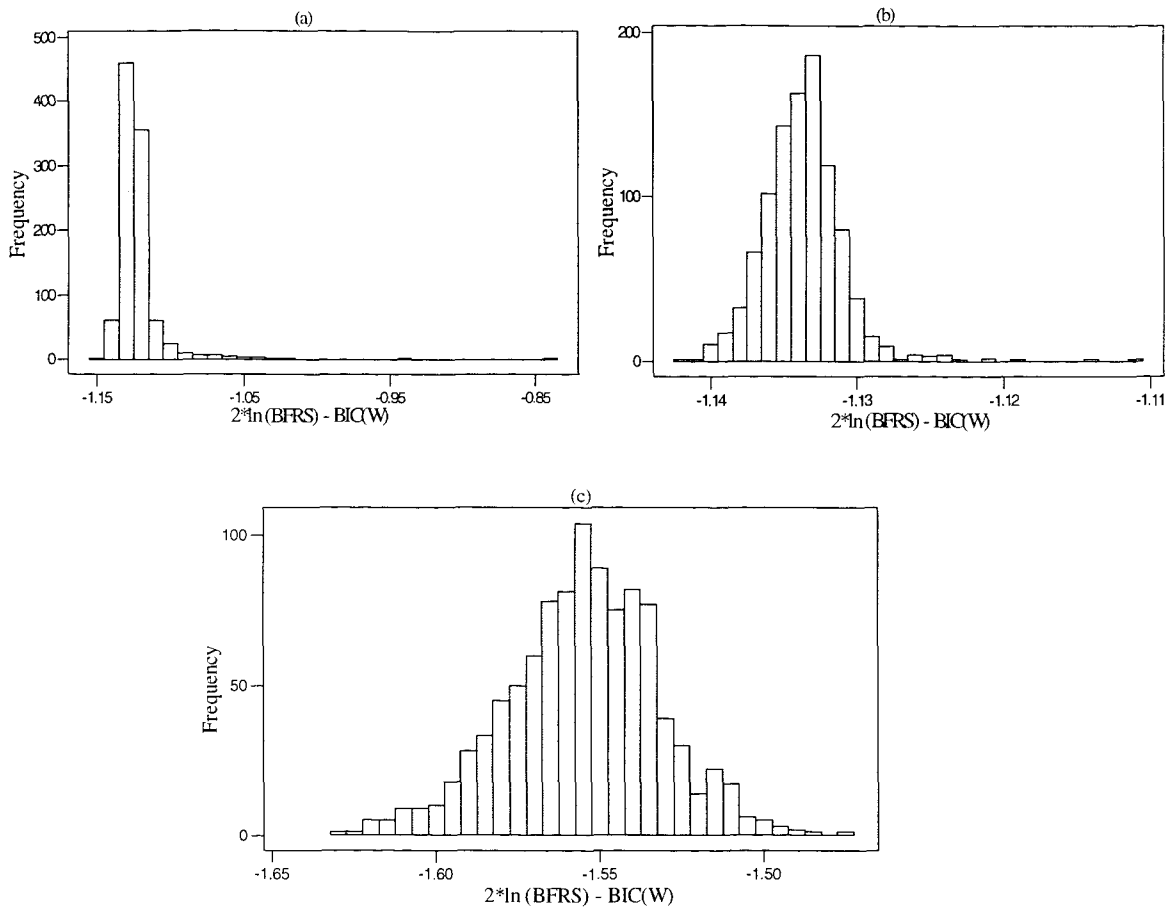

Figure 7.16: Histogram for the difference between $2\ln(\text{BFRS})$ and $\text{BIC}(X^2_{W(\phi)})$ in three cases in cluster sample, (a) for sample size of a 1000 in 10 psus, (b) for sample size of a 5000 in 50 psus, and (c) for sample size of a 1000 in 200 psus.

For the cluster sampling scheme, asymptotically, with a large number of observations in each psu, ignoring the sampling design will not have much effect in testing independence, except in the error associated with the approximation using BIC. Nevertheless, the Pearson chi-squared statistic and the Wald statistic, which have identical values for large psu, do not seem to be distributed exactly as χ^2_1, and they have inflated type I error. For a small number of observations in each psu, the MCMC method does not work properly. Thus, we consider $2\ln(\text{BFRS})$ as a baseline instead of $2\ln(\text{Bayes factor})$. In this case, the effect of

the cluster sample is visible and it is more serious if n_t is very small.

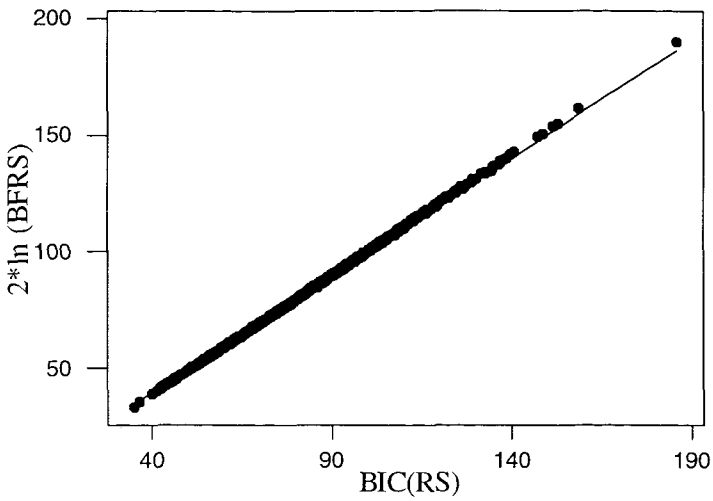

Figure 7.17: Comparison of $2\ln(\text{BFRS})$ values with $\text{BIC}(X^2_{IRS})$, for stratified sampling design, with $p_1 = (0.3, 0.2, 0.2, 0.3)$, $p_2 = (0.5, 0.1, 0.2, 0.2)$ and a sample size of 1000.

7.1.7 Conclusion

We have investigated the effect of the sampling design on a test of independence. The sampling design effect on the ordinary chi-squared test of independence is less serious, in general, than in the goodness-of-fit. This result is consistent with the evidence that the design effect for the difference between two sub-class means tends to be much smaller than that for individual means, i.e. $\hat{\delta}_.$ is smaller than $\hat{\tau}_.$, see Kish and Frankel (1974), Holt, Scott and Ewings (1980), Rao and Scott (1981). In fact, unlike the goodness-of-fit test the sampling design effect almost vanishes as the sample size gets larger.

On the other hand, for the cluster sampling design, the distribution of both the Pearson chi-squared, X^2_I, and the Wald, $X^2_{W(\phi)}$, statistics are distributed almost as $2\chi^2_1$, under the null model for large sample sizes, $n_t = 1000$, in line with the previous work, discussed earlier. This will not effect our results, since

the approximation of BIC depends on the test statistic not on its asymptotic distribution.

Nevertheless, concerning the error associated with the approximation of BIC in a test of independence, our results are consistent with the observations of Kass and Wasserman (1995), who indicate that if the prior choice is consistent with "unit prior information", then the error is of order $O(n^{-\frac{1}{2}})$ rather than $O(1)$, and, also, with Raftery's (1996) conclusion from his empirical research, that BIC is more accurate in practice than the $O(1)$ error term would suggest.

Finally, the results indicate that although there is an effect of sampling design in the Bayesian information criterion (BIC) approximation based on the ordinary Pearson chi-squared test statistic, X_I^2, of independence, the difference between using X_I^2 and any corrected statistic is negligible. Therefore, in testing of independence, there is no justification in practice for the use of a more complex test statistic, unless the number of observations in each psu is very small.

7.2 Risk function

In this section, we determine the behaviour of point estimation in both the Bayesian approach, using a model averaged estimate, MAE, and in the classical approach, using an estimate based on a pretest, PTE. We will carry out a simulation to study the risk performance of these estimators in a 2×2 contingency table. In this simulation three sampling schemes were considered, simple random sampling, stratified sampling, and cluster sampling.

The true cell probabilities, $\mathbf{p}$, are known in our simulations. We will use the Kullback-Liebler distance, described in section (4.1.3), as our loss function

$$L(\mathbf{p}, \hat{\mathbf{p}}) = \sum_{i=1}^{2}\sum_{j=1}^{2} p_{ij} \ln \frac{p_{ij}}{\hat{p}_{ij}}. \tag{7.4}$$

Then, averaging the loss function over the simulation sample (1000 simulations), gives an estimate of the risk function. The three main estimation procedures considered in this simulation are,

- Model averaged estimation, MAE,

$$\begin{aligned}\hat{p}_{ij} &= E(p_{ij}|\mathbf{n}) \\ &= pr(M_S|\mathbf{n})E(p_{ij}|\mathbf{n}, M_S) + pr(M_I|\mathbf{n})E(p_{i+}p_{+j}|\mathbf{n}, M_I) \\ &= \frac{1}{1+B_{IS}(n)}\left(\frac{n_{ij}+\alpha_{ij}}{n+\alpha}\right) + \frac{B_{IS}(n)}{B_{IS}(n)+1}\left(\frac{n_{i+}+\alpha_{i+}}{n+\alpha}\right)\left(\frac{n_{+j}+\alpha_{+j}}{n+\alpha}\right)\end{aligned}$$

where $\alpha_{ij} = 0.25, i, j = 1, 2$, and $B_{IS}(n)$ is the multinomial Bayes factor for the independence model, M_I, against the saturated model, M_S. For stratified and cluster samples, we consider this estimator, and two other approximations using an adjustment to the multinomial Bayes factor. The first adjustment used is the Rao and Scott (1981) correction factor to the chi-squared statistic for testing independence, equation (2.52). The second adjustment used is based on the actual design effect for the log odds ratio, ξ, see equation (7.2).

- Pretest estimator, PTE,

  For the simple random sample, we consider the pretest estimator

$$\hat{p}_{ij} = \begin{cases} \frac{n_{i+}n_{+j}}{n} & \text{if } X^2 \leq 3.84 \\ \frac{n_{ij}}{n} & \text{if } X^2 > 3.84. \end{cases} \tag{7.5}$$

  For the stratified and cluster sampling schemes, in addition, the Rao and Scott (1981) correction factor to the chi-squared statistic for testing independence, and the Wald statistics for the log odds ratio were used.

- Maximum likelihood estimator for the saturated model, Sat, (Bickel and Doksum, 1977)

$$\hat{p}_{ij} = \frac{n_{ij}}{n}. \tag{7.6}$$

7.2.1 Simulation study

We ran a simulation for three sampling designs, simple random sampling, stratified sampling, and cluster sampling and for a range of different dependence structures. This simulation considered three sample sizes, small, moderate, and large. To determine the true values of the cell probabilities, $\mathbf{p}$, we assume that the cell probabilities have the following structure

$\frac{1}{4}(1+\theta)$	$\frac{1}{4}(1-\theta)$	0.5
$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1+\theta)$	0.5
0.5	0.5	1

If ϕ is the log odds ratio, then

$$\theta = \frac{e^{\frac{\phi}{2}} - 1}{e^{\frac{\phi}{2}} + 1}.$$

Given a value of ϕ, and hence θ, we can evaluate $p_{ij}, i, j = 1, 2$. In general, we consider the values $\phi \in [-10, 10]$; This will produce the values of $\mathbf{p}$ in Table (7.1).

7.2.2 Simple random sample

For the simple random sample, the risk function is computed for three estimates. First, the maximum likelihood estimate for the saturated model, Sat. Second, the pretest estimate, PTE, and third the model averaged estimate, MAE.

For a small sample size, $n = 40$, the MAE has a small risk and it is superior to the other estimators of ϕ between -1 and 1, and it is better than PTE elsewhere,

ϕ	p_1	p_2	p_3	p_4
-10.0	0.003	0.497	0.497	0.003
-7.0	0.015	0.485	0.485	0.015
-5.0	0.038	0.462	0.462	0.038
-4.0	0.060	0.440	0.440	0.060
-3.0	0.091	0.409	0.409	0.091
-2.0	0.134	0.366	0.366	0.134
-1.0	0.189	0.311	0.311	0.189
0.0	0.250	0.250	0.250	0.250
1.0	0.311	0.189	0.189	0.311
2.0	0.366	0.134	0.134	0.366
3.0	0.409	0.091	0.091	0.409
4.0	0.440	0.060	0.060	0.440
5.0	0.462	0.038	0.038	0.462
7.0	0.485	0.015	0.015	0.485
10.0	0.497	0.003	0.003	0.497

Table 7.1: The values of ϕ considered in the simulation and the corresponding cell probabilities.

where Sat is the best estimator. For extreme values of ϕ all estimators are equivalent; see figure (7.18-a).

When we increase the sample size to $n = 200$, the MAE is still superior to the other estimators in the cases of independence and not far from independence, i.e. $\phi \in (-0.55, 0.55)$. If the values of ϕ is far from independence then Sat is the best estimate and PTE is better than MAE, figure (7.18-b).

Figure (7.18-c) presents the case for a large sample size, $n = 1000$; In this case, the behaviour of all estimators is similar to the case of $n = 200$, where the MAE is still superior to other estimators in the cases of independence and not far from independence, i.e. $\phi \in (-0.2, 0.2)$. If $\phi < -0.9$ and $\phi > 0.9$ all estimators have the same risk, i.e. they are equivalent. The turning points in the risk function for the different estimator are getting smaller as the sample size gets larger.

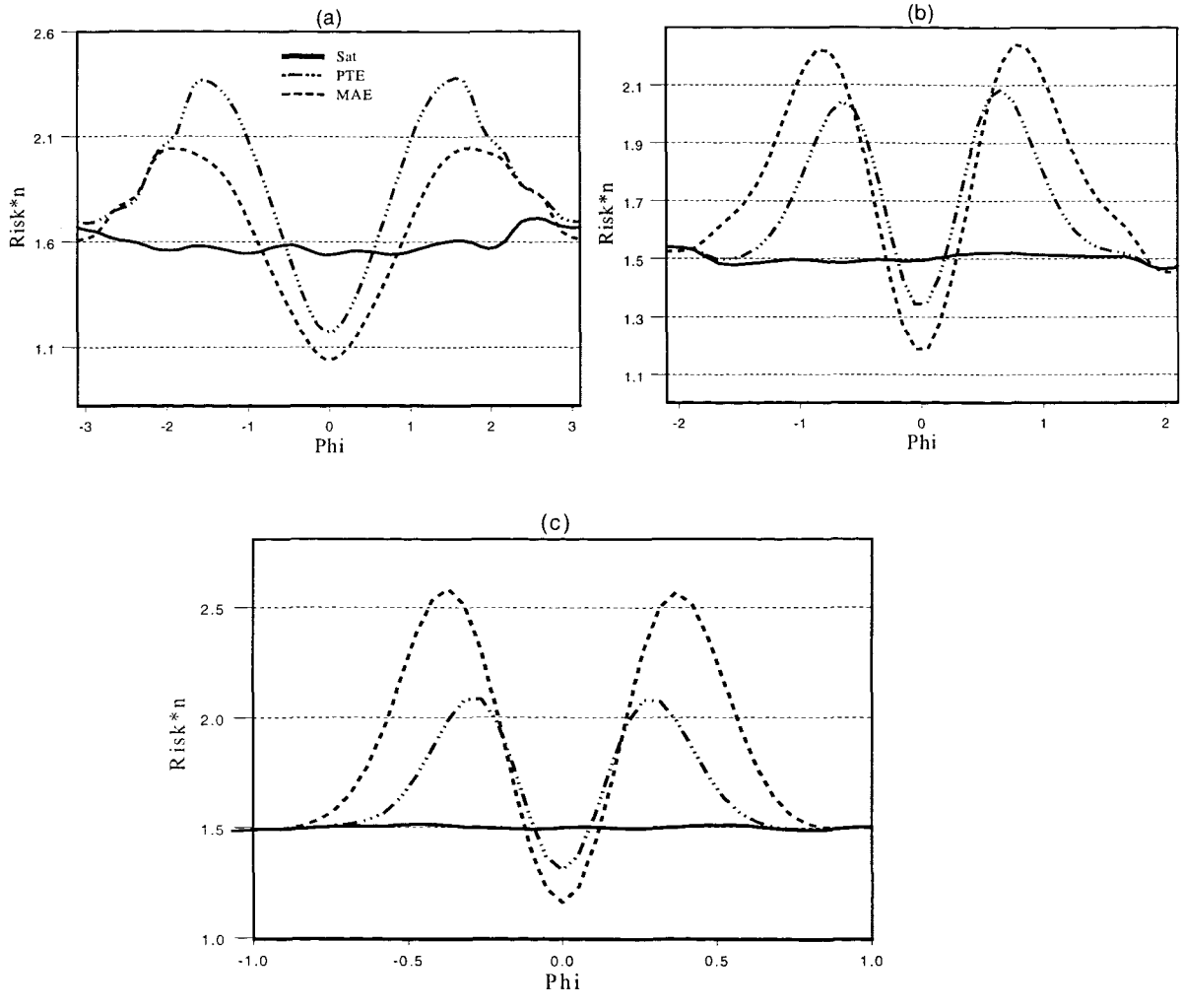

Figure 7.18: Risk function of the estimates of $\mathbf{p}$ in a simple random sample simulation using the maximum likelihood estimate for the saturated model, Sat, pretest estimate, PTE, and model averaged estimate, MAE, in three different sample sizes. (a) $n = 40$, (b) $n = 200$, and (c) $n = 1000$.

7.2.3 Stratification

For a stratified sampling design, our design has two strata. The first stratum, $\mathbf{p}_1$, is computed directly from the assumed values of ϕ, now denoted by ϕ', as in the simple random sample. The second strata, $\mathbf{p}_2$, is always equal to

$(0.25, 0.25, 0.25, 0.25)$, where $w_1 = w_2 = 0.5$. As an example, assume that $\phi' = 2$, this will produce $\mathbf{p}_1 = (0.366, 0.134, 0.134, 0.366)'$, and $\mathbf{p}_2$ is always $(0.25, 0.25, 0.25, 0.25)$, then $\mathbf{q}$ =$(0.308, 0.192, 0.192, 0.308)'$ with this the value of the log odds ratio, ϕ, equal to 1.29. Five estimators are considered in this case. Two are model averaged estimates, two are pretest estimates, and one is the maximum likelihood estimate for the saturated model. The risk functions are presented for the estimates of $\mathbf{q}$, in a stratified sampling design, for three different sample sizes. The two model averaged estimators are based on the multinomial Bayes factor, PBF, and on adjusted multinomial Bayes factor using the Rao and Scott correction, PAdBF. The pretest estimators are based on the chi-squared statistic, PChi, and on the Rao and Scott corrected chi-squared statistic, PRS.

When the sample size is small, $n = 40$, figure (7.19-a) shows that the MAE is superior to other estimators in the cases of independence and not far from independence, i.e. $\phi \in (-0.8, 0.8)$. In this interval the difference between the corrected or uncorrected estimator for both MAE and PTE is very small. The risk of PBF is always smaller than PChi. Also, the risk of PAdBF is always smaller than PRS. The behaviour of PAdBF and PRS near the boundary deserves further comment. PAdBF and PRS dominate the estimators that do not take account of the sampling design, i.e. PChi and PBF. In these cases the strata are inhomogeneous, and as we show in the previous section and in figure (7.19-a) there is a design effect.

If we increase the sample size to 200, figure (7.19-b) shows the similarly between PBF and PAdBF, also PChi with PRS. The model averaged estimator is still superior with minimum risk in the cases where the values of ϕ represent independence and not far from independence, i.e. $\phi \in (-0.31, 0.31)$. If ϕ is between –1.53 and -0.6 or between 0.6 and 1.53 the PBF and PAdBF have a large risk compared with other estimators. At the boundary, all estimators behave equally.

When the sample size is equal to 1000, figure (7.19-c) shows, again, the similarly between PBF and PAdBF, and also PChi and PRS, indicating that the design effect asymptotically vanishes. As in the previous cases MAE is superior in the case of independence and around it, i.e. $\phi \in (-0.23, 0.23)$. On the other hand, if $\phi \in (-0.86, -0.23)$ or $\phi \in (0.23, 0.8)$ the MAE has a large risk compared with the other estimators. At the boundary, all estimators have equal risk.

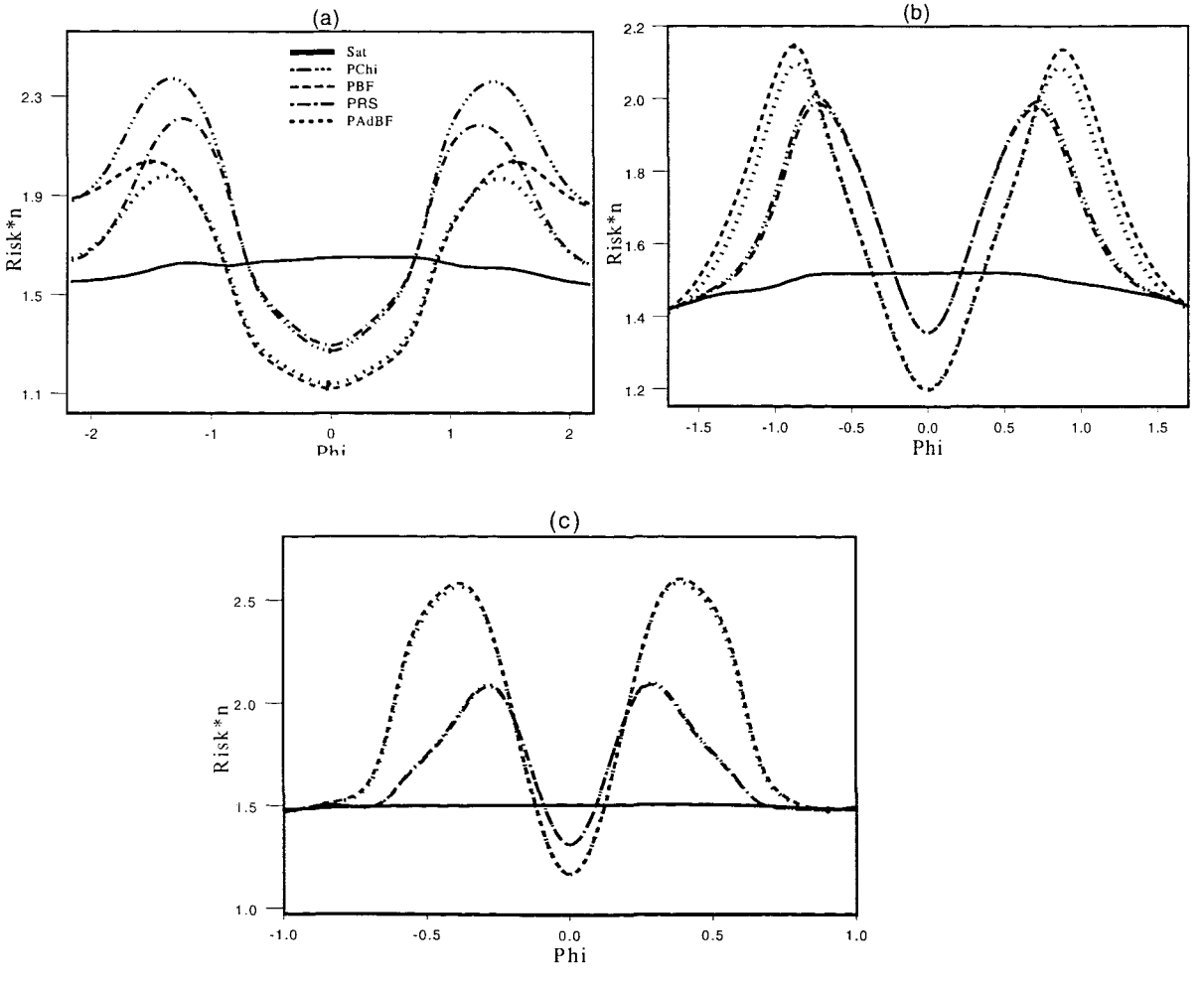

Figure 7.19: Risk function for the estimates of **p** in a stratified sampling design for the maximum likelihood estimate to the saturated model, Sat, pretest estimate, PTE, using chi-squared statistic, PChi, and Rao and Scott corrected chi-squared statistic, PRS, and model averaged estimate, MAE, using the multinomial Bayes factor, PBF, and adjusted multinomial Bayes factor, PAdBF, in three different sample sizes. (a) $n = 40$, (b) $n = 200$, and (c) $n = 1000$.

7.2.4 Clustering

For the cluster sampling design, we consider three numbers of psu, $c = 10, 50,$ and 200, each with a small number of observations in each psu, n_t, and, also, with a total sample size of 1000. The estimators considered for estimating $\mathbf{q}$ are the same five estimators considered in the stratification case.

For the first case, small n_t, when the design has 10 psus, we consider the case of $n = 40$, i.e. $n_t = 4$. The result is shown in figure (7.20-a). The PBF has the smallest risk among all estimators, then comes PAdBF as the second best estimator, when the values of ϕ represent independence and not far from independence, i.e. $\phi \in (-0.85, 0.85)$. The maximum likelihood estimate for the saturated model, Sat, has the lowest risk if $\phi \in (-2.0, -0.85)$ or $\phi \in (0.85, 2.0)$. The design effect is clear in those intervals. At the boundary, the curves do not seem symmetric. This may be caused by the simulation, i.e. Monte Carlo error.

As the number of psus increases to 50 and $n_t = 2$, i.e. $n = 100$, the PBF is still superior with minimum risk when the values of ϕ are between -0.5 and 0.5, figure (7.20-b). Beyond that, the design affects the results of both PBF and PChi, where they have large risk. At the tails, where $\phi < -2.2$ and $\phi > 2.2$, all estimators have almost the same risk.

When we consider 200 psus with $n_t = 2$, i.e. $n = 400$, figure (7.20-c) shows clearly the effect of the design on both PBF and PChi, where they have large risk when $\phi < -0.25$ or $\phi > -0.25$. If $\phi \in (-0.25, 0.25)$ the PBF still has the smallest risk among all estimators.

If we consider a total sample size of 1000, figure (7.21-a) shows the results of using 10 psus with $n_t = 100$. The similarly between PBF and PAdBF, also PChi with PRS is visible. The model averaged estimator, MAE, is still superior with minimum risk in the cases where the values of ϕ represent independence and not far from independence, i.e. $\phi \in (-0.1, 0.1)$. If ϕ is between -0.8 and

-0.25 or between 0.25 and 0.8 the PBF and PAdBF have a large risk compared with other estimators. Beyond that, at the boundary, all estimators behave the same.

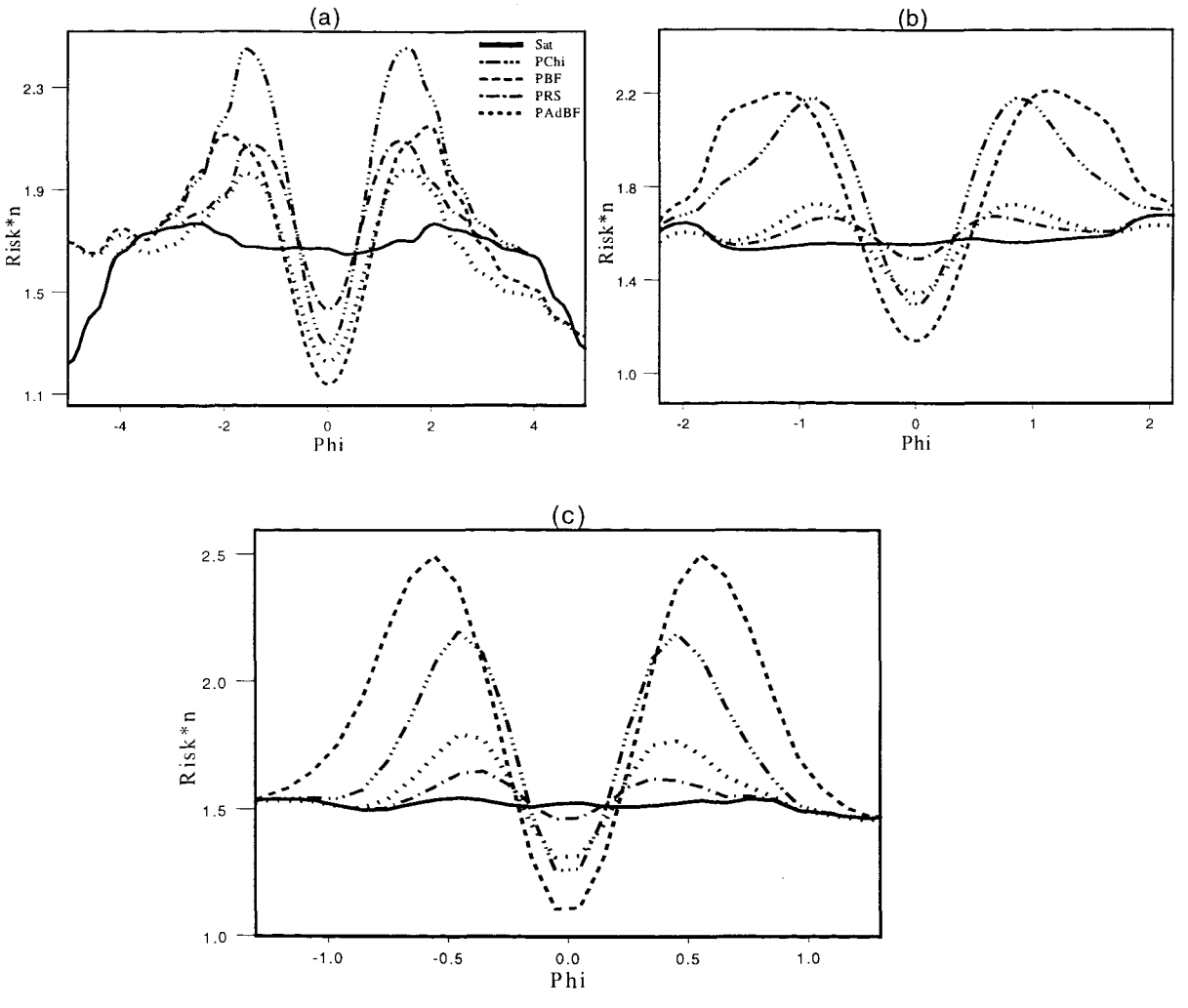

Figure 7.20: Risk function for the estimates of $\mathbf{p}$ in a cluster sampling design for the maximum likelihood estimate to the saturated model, Sat, pretest estimate, PTE, using chi-squared statistic, PChi, and Rao and Scott corrected chi-squared statistic, PRS, and model averaged estimate, MAE, using the multinomial Bayes factor, PBF, and adjusted multinomial Bayes factor, PAdBF, in a small sample size and three different numbers of psus. (a) 10 psus with $n = 40$, (b) 50 psus with $n = 100$, and (c) 200 psus with $n = 400$.

Figure (7.21-b) shows the result when the design has 50 psus with $n_t = 20$,

approximately PBF and PAdBF, also PChi and PRS behave roughly the same. The result in this case is not far from the design with 10 psus and $n_t = 100$, i.e. figure (7.21-a). When we consider 200 psus with $n_t = 5$, the design has an effect on the estimation if we consider PBF and PChi away from the center, where $\phi \notin (-0.2, 0.2)$, see figure (7.21-c). At the boundary, all estimators have equal risk.

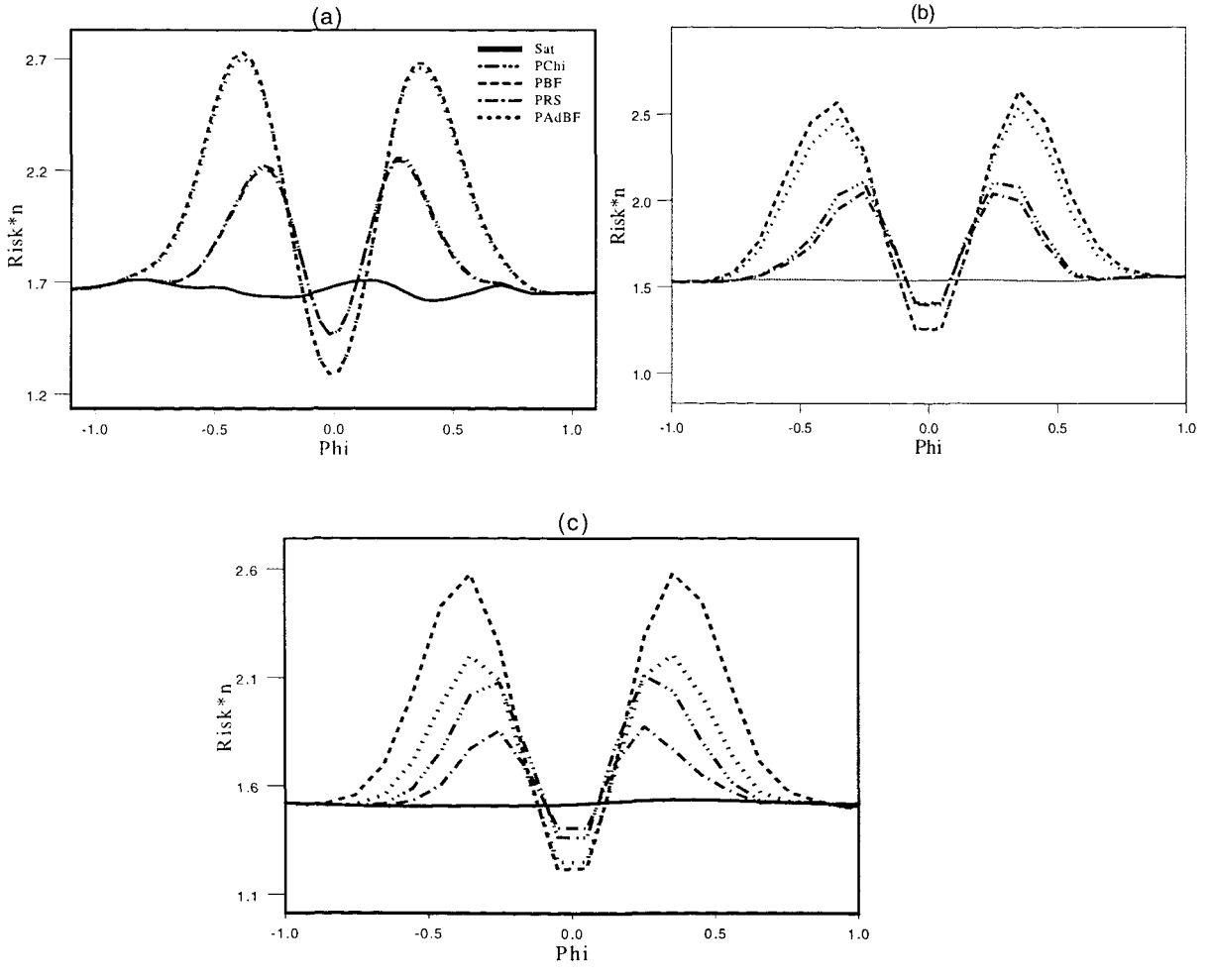

Figure 7.21: Risk function for the estimates of **p** in a cluster sampling design for the maximum likelihood estimate to the saturated model, Sat, pretest estimate, PTE, using chi-squared statistic, PChi, and Rao and Scott corrected chi-squared statistic, PRS, and model averaged estimate, MAE, using the multinomial Bayes factor, PBF, and adjusted multinomial Bayes factor, PAdBF, for three different numbers of psu and a total sample size of 1000. (a) 10 psus, (b) 50 psus, and (c) 200 psus.

If we consider a very large sample size, $n = 50000$, where the design has 50 psus, i.e. $n_t = 1000$, the design effects vanish. Thus, all estimators appear to be similar.

7.2.5 Discussion

We have demonstrated that, for estimating the cell probabilities, $\mathbf{p}$, the model averaged estimator, MAE, is superior to the pretest estimate, PTE, when the sample size considered is small. In addition, when the sample size is moderate or large, the central interval for the values of ϕ, where the values of ϕ represent independence and not far from independence, is dominated by MAE. In this case the MAE has the smallest risk compared with PTE and the maximum likelihood estimate for the saturated model, Sat.

7.2.6 Conclusion

We have illustrated how the Bayesian approach for estimating the cell probabilities, $\mathbf{p}$, using the model averaged estimator, MAE, can be better than the pretest estimate, PTE. Estimation based on selection of a single model, say M_k, will not effectively take account of uncertainty about the parameter of interest, the cell probabilities, $\mathbf{p}$. However, the MAE considers the uncertainty of the model by averaging the estimates according to how likely each model is.

When the values of ϕ are reasonably far from independence and the sample size is not small, the MAE has greater risk than PTE and the maximum likelihood estimate for the saturated model. For small sample size and when the sampling design is complex, using the adjustment of the multinomial Bayes factor or the correction of the Pearson chi-squared will make a significant difference in the risk of the estimator. As the sample size becomes larger the

difference disappears. Thus, the design effect in this case is negligible and the use of multinomial Bayes factor or Pearson chi-squared for estimating $\mathbf{p}$ is more justified than using complex methods, unless the sample size is small.

Finally, the result of the risk function of MAE, using the design effect of log odds ratio, ξ, as an adjustment for the multinomial Bayes factor, is always between PBF and PAdBF, with smaller difference than PAdBF. Also, the risk using the Wald log odds ratio for PTE is between PChi and PRS.

Chapter 8

Summary and recommendations for further research

8.1 Summary of conclusions

Our objective in this thesis was to present a comprehensive investigation of a Bayesian approach to model selection for categorical survey data, encompassing simple random, finite population, stratified, and cluster sampling. We have demonstrated the effects of a complex sampling scheme on Bayesian model selection, and somewhat on classical hypothesis testing (through the Pearson chi-squared, X^2, the Rao and Scott (1981) first- and second-order corrections, X^2_{RS1}, X^2_{RS2}, and the Wald statistic, X^2_W, and the log odds ratio statistics for testing independence). We have illustrated how the well known classical Pearson chi-squared test statistic is either too liberal or too conservative, depending on the sampling scheme, when applied to survey data. Moreover, we have presented two simple adjustments to the multinomial-based Bayes factor. Using these adjustments gave better estimation of $2\ln(\text{Bayes factor})$, compared with using the Wald statistic, X^2_W, in the Bayesian information criterion approximation (BIC).

We have shown that effect of sample design on inference using the ordinary chi-squared test of independence is less serious, in general, than on a test of goodness-of-fit of specified probabilities. This result is consistent with the evidence that the design effect for the difference between two sub-class means tend to be much smaller than that for individual means, see Kish and Frankel (1974), Holt, Scott and Ewings (1980), Rao and Scott (1981).

In the finite population case, we found that there is no effect on inference when this design is ignored, if we consider the individuals who make up the population to be, a priori, exchangeable, as suggested by Ericson (1969, 1988). Equivalent results are stated by Rao and Thomas in Skinner, Holt and Smith (1989, p. 92). Furthermore, for stratification the results indicate that there is no effect in ignoring the sampling design, proportional stratification, on the analysis of the data if the strata are homogeneous. This result has been mentioned in the classical approach by Kish and Frankel (1973). But if strata are highly inhomogeneous ignoring the sampling design may affect the results. In a test of goodness-of-fit of a model M_G where all cell probabilities are specified, the Pearson chi-squared statistic, X^2, is always asymptotically conservative. On the other hand, the BIC based on Wald statistic performs similarly to $2\ln$(Bayes factor). In Bayesian model selection, when we compared the estimated true values of $2\ln$(Bayes factor) and the values of $2\ln$(uncorrected, multinomial-based, Bayes factor), it underestimated the true values in all cases, i.e. with homogeneous and inhomogeneous strata. This indicates that estimation of the Bayes factor is sensitive to the effect of stratification. Moreover, examples for testing independence, in section (7.1.4), show existence of a design effect in highly inhomogeneous strata, even for the corrected Pearson chi-squared statistic.

We have shown the effect of cluster sampling on Bayesian model selection, if the sampling design is ignored. In a goodness-of-fit test, the BIC(X^2) for model selection may give a misleading result of rejecting the null model M_0, when it is true, since it overestimates the true value, $2\ln$(Bayes factor) in all examples

considered. However, the BIC(X^2_{RS1}), BIC(X^2_{RS2}) and BIC(X^2_W) tend to underestimate the large values of 2 ln(Bayes factor), but not with the same magnitude as BIC(X^2). This is caused by the conservative values of the corrected X^2_{RS1}, X^2_{RS2}, and X^2_W statistics. On the other hand, ignoring the sampling design asymptotically will not have any effect on the statistical inference when testing independence. The effect is visible with small sample sizes and more serious if the number of the psus is large.

In this sampling scheme, asymptotically, the values of X^2_W for testing goodness-of-fit are conservative. Moreover, for testing independence, the distribution of both the Pearson chi-squared, X^2_I, and the Wald statistics, $X^2_{W(\phi)}$, are distributed almost as $2\chi^2_1$, under the null model. This is a worrying result, but Nathan (1975) has similar results, in table (2) of his paper, for the log likelihood ratio and the Wald statistics. Furthermore, in the statistical literature, there is some evidence to suggest that as the sample size increases, the asymptotic size of the χ^2-test of independence, under the multinomial sampling, can be larger than its nominal level (Loh, 1989). Also, there is evidence to suggest that the Wald statistic has poor behaviour in sampling survey inference (Thomas and Rao, 1987, Skinner, Holt and Smith, 1989). Fay (1985) presents his concerns regarding the use of the Wald statistic, and Thomas and Rao (1987), in a Monte Carlo study, confirm Fay's worries. Nevertheless, this will not effect our results, since the approximation using BIC depends on the likelihood test statistic not its asymptotic distribution.

In a test of independence, unlike goodness-of-fit test, the results indicate that although there is an effect of sampling design on the Bayesian information criterion (BIC) approximation based on the ordinary Pearson chi-squared test statistic, X^2_I, of independence, the difference between using X^2_I or any corrected statistic is negligible. Therefore, in testing of independence, there is no justification in practice for the use of more complex test statistic, such as one based on the Rao and Scott (1981) first- and second-order corrections, and the Wald

statistic, unless the number of observations in each psu is very small.

We have presented two adjustments to the multinomial-based Bayes factor, BFRS1 and BFRS2. These adjustments, $2\ln(\text{BFRS1})$ and $2\ln(\text{BFRS2})$, yield very reliable estimated values for the true values of $2\ln(\text{Bayes factor})$. Moreover, these adjustments are superior to $\text{BIC}(X^2_W)$, when each psu has a moderate or large number of observations and equivalent, when each psu has a small number of observations. In fact, they seem to produce asymptotically better approximations of the values of $2\ln(\text{Bayes factor})$ than $\text{BIC}(X^2_W)$. The performance of both adjusted Bayes factors are nearly identical.

In both stratification and cluster sampling schemes, we applied the Savage-Dickey density ratio. In order to evaluate the Savage-Dickey density ratio, we have to draw a sample from the marginal posterior density $pr(\mathbf{p}|\mathbf{n},M_S)$, and the prior density $pr(\mathbf{p}|M_S)$, and then, estimate those densities at point $\mathbf{p}_0$. In stratification, we use a Monte Carlo sampling method, and a crude density estimator. The crude density estimate does not estimate the large values of $2\ln(\text{Bayes factor})$ well for this sampling scheme or (simulation 1) for the cluster sampling scheme. For most examples in chapter (6) for cluster sampling and in chapter (7), we use the multivariate normal kernel density estimator. We illustrated, in chapter (5), that the multivariate normal kernel density estimator, with normal based bandwidth, is very reliable for two- and three-dimensional data. For four-dimensional data, the estimated values are very sensitive to the bandwidth. Unfortunately, for higher-dimensional data the bias seems to be large.

In the cluster sampling scheme, we encountered some difficulty in drawing a sample from the marginal posterior density. We used a hybrid MCMC strategy, which consists of a combination of two algorithms, the Gibbs sampler and the Metropolis-Hastings algorithm. This MCMC strategy worked well when each psu has a moderate or large number of observations. However, when the number of observations in each psu is small, we are uncertain about the MCMC

algorithm reaching the marginal posterior, target distribution, even with 10000 iterations. The possibility of considering more iterations is not a plausible solution, since our program has complex computations with long running times; For 1000 samples each with 200 psus and 10000 iterations the estimated running time is about a month! Thus, when each psu has small number of observations, we cannot consider our estimates for the true values of the Bayes factor under cluster sampling to be trustworthy. Nevertheless, as the two adjustments to the multinomial Bayes factor produced reliable estimates, we have compared the values of both $2\ln(\text{BFRS1})$, using adjustments to the multinomial-based Bayes factor, and $\text{BIC}(X^2_W)$, when the number of observations in each psu is small. We show that they are almost identical. This adds further support for using BFRS1 for the Bayesian model selection, as it is easier to calculate. This adjusted Bayes factor, BFRS1, requires only the knowledge of variance estimates, or design effects, for individual cells. Programing wise, it requires a simple program, with fast running time.

We have shown mathematically the well known result that the BIC approximation has error of magnitude $O(1)$. This indicates that the approximation is somewhat crude, because the error does not vanish even with infinite amount of data. Kass and Wasserman (1995) indicates that if the prior choice is consistence with "unit prior information", then the error is of order $O(n^{-\frac{1}{2}})$ rather than $O(1)$. This does not appear to be the case with our definition of "unit prior information", under the cluster sampling scheme, since, the error does not vanish or get smaller as n gets larger in the goodness-of-fit test. In fact, the error seems to get relatively larger as n increases. Nevertheless, our results are consistent with Raftery's (1996) conclusion from his empirical result that BIC is more accurate in practice than the $O(1)$ error term would suggest. In the independence test the error gets smaller as n_t increases and is therefore consistent with Kass and Wasserman (1995), and Raftery (1996).

Finally, in section (4.1), we examined the performance of model averaged

estimation (MAE) against pretest estimation based on the classical Pearson chi-squared test statistic (PTE). In section (7.2), we generalised our result by considering all three sampling schemes. We demonstrated using risk analysis how the Bayesian approach for estimating the cell probabilities, $\mathbf{p}$, using MAE, can be superior to the pretest estimate, PTE. Estimation based on selection of a single model, say M_k, will not effectively take account of uncertainty about the parameter of interest, the cell probabilities, $\mathbf{p}$. However, the MAE considers the uncertainty of the model by averaging the estimates according to how likely each model is. Model averaged estimates based on adjusted Bayes factors seem to work well for complex sample designs.

8.2 Recommendations for further research

Adjusting the multinomial-based Bayes factor provides an improved approximation to the value of the true Bayes factor. Moreover, these adjustments are superior to approximation based on BIC(X^2_W), when each psu has a moderate or large number of observations and identical to it for small numbers of observations. Nevertheless, a comprehensive study is needed to determine how good is the adjusted Bayes factor. Further, theoretical justification for the adjusted Bayes factor would be useful. The adjusted Bayes factor can be computed for more models, such as three-way tables, using Rao and Scott's (1984, 1987) correction factors for nested log-linear models.

In our research we considered three cells ($K = 3$) in the goodness-of-fit and a 2 by 2 contingency table in tests of independence. Thus considering more dimensions is an important extension of this work. Since our programs are generalized for 2 way tables, this extension will be easy in practice. Unfortunately, in computing the true values of $2\ln$(Bayes factor), the precision of the multivariate normal kernel density estimator will be poor for $K \geq 4$, so considering more

advanced density estimators is essential. Possible examples include adaptive, and spline density estimates, for details see Simonoff (1996), and Scott (1992). We are uncertain about the MCMC algorithm reaching the marginal posterior, target distribution, even with 10000 iterations, when the number of observation in each psu is small. Thus, the MCMC algorithm needs to be developed. Both of these problems are related to computing the Savage-Dickey density ratio. Therefore, an alternative for the approximation of the Bayes factor other than the Savage-Dickey density ratio is a useful idea, for details see DiCiccio *et al.* (1997) or Dellaportas, Forster, and Ntzoufras (2000).

Green (1995) introduced Reversible Jump MCMC. This method makes it possible to construct reversible Markov chain samplers that jump between parameter subspaces of different dimensionality, where the traditional MCMC methods are restricted to situation where the dimensionality of the parameter vector is fixed. Bridge sampling (DiCiccio, 1997) is another way of estimating the Bayes factor via posterior simulation using MCMC. However, both methods require generating samples from the reduced model, which in our case is defined by the constraint (4.21), and this is a difficult task. Thus, using the Savage-Dickey density ratio in our point of view, is preferable. On the other hand, it is a potential interesting avenue of further research to find a way to apply these other MCMC methods.

Careful prior specification is needed when constructing the associated Bayes factor for model selection. In our thesis we applied non-informative proper priors. If we consider the argument that the use of a non-informative improper prior is more appropriate than using a non-informative proper prior for model selection, then alternatives other than the conventional Bayes factor should be considered. There is increasing interest in developing statistical procedures that allow Bayes factors to be calculated when using improper priors. Several alternative Bayes factor have been introduced to address this problem. Among these alternatives, the fractional Bayes factor (O'Hagan, 1995) and intrinsic

Bayes factor (Berger and Pericchi, 1996) are the most promising. It would be interesting to investigate these approaches. However, Kass and Wasserman (1995) believe that using unit prior information is preferable to these alternatives.

Finally, this research can be explored for unequal strata and psus. However, we expect the results to be broadly similar. Another possible extension to this work is to study further sampling schemes, for example multi-stage sampling.

References

- Agresti, A. (1990). *Categorical Data Analysis.* London: Wiley.
- Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* London: Wiley.
- Ahrens, J. H. and Dieter, U. (1974). Computer Methods for Sampling from Gamma, Beta, Poisson and Binomial Distributions. *Computing*, **12**, 223-246.
- Ahrens, J. H., and Dieter, U. (1982). Generating Gamma Variates by a Modified Rejection Technique. *Communications of the ACM*, **25**, 47-54.
- Altham, P. M. E. (1976). Discrete variable analysis for individuals grouped into families. *Biometrika*, **63**, 263-269.
- Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109–122.
- Bhapkar,V. P. (1966). A Note on the Equivalence of Two Test Criteria for Hypotheses in Categorical Data. *Journal of the American Statistical Association*, **61**, 228-235.
- Bickel, P. J. and Doksum, K. A. (1977). *Mathematical statistics: Basic ideas and selected topics.* Oakland: Holden-Day.

- Binder, D. A. (1982). Non-parametric Bayesian Models for Samples from Finite Populations. *Journal of Royal Statistical Society*, **B44**, 388-393.

- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279–292.

- Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* The Massachusetts Institute of Technology.

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

- Casella, G., and George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, **46**, 167-174.

- Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hasting Algorithm. *The American Statistician*, **49**, 327-335.

- Cochran W. G. (1977). *Sampling techniques.* 3^{rd} edition, New York: Wiley.

- Cohen, J. E. (1976). The distribution of the chi-squared statistic under clustered sampling from contingency tables. *Journal of the American Statistical Association* , **71**, 665-670.

- Cressie, N. and Read, T. R. C. (1989). Pearson's X^2 and the Loglikelihood Ratio Statistic G^2: A Comparative Review. *International Statistical Review*, **57**, 19-43.

- De Bruijn, N. G. (1961). *Asymptotic Methods in Analysis.* Amsterdam: North-Holland.

- Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2001). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, to appear

- Dickey, J. (1971). The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters. *Annals of Mathematical Statistics*, **42**, 204-223.

- DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997). Computing Bayes factor by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, **92**, 903-915.

- Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of probability and Its Applications*, **14**, 153-158.

- Ericson, W. A. (1969). Subjective Bayesian models in sampling finite population (with discussion). *Journal of Royal Statistical Society*, **B31**, 195-233.

- Ericson, W. A. (1988). Bayesian Inference in Finite Population. *Handbook of Statistics*, V. 6, edited by Krishnaiah, P. R. and Rao, C. R., Elsevier Science Publishers B. V., **6**, 213-246.

- Fay, R. E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, **80**, 148-157.

- Fellegi, I. P. (1980). Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples. *Journal of the American Statistical Association*, **75**, 261-268.

- 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. (1996). *Bayesian Data analysis.* London: Chapman and Hall.

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

- Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice.* London: Chapman and Hall.

- Graubard, B. I. and Korn, E. L. (1993). Hypothesis testing with complex survey data: The use of classical quadratic test statistics with particular reference to regression problems. *Journal of the American Statistical Association*, **88**, 629-641.

- Gray, G. (1994). Bias in Misspecified Mixtures. *Biometrics,* **50**, 457-470.

- Graybill, F. A. (1983). *Matrices with applications in statistics.* 2^{ed} edition, Belmont, California: Wadsworth.

- Green, P. J.(1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

- Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample survey methods and theory.* Volume I, New York: Wiley.

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

- Hoeting, J. A., Madigan, D. , Raftery, A. E., Volinsky, C. T. (1999). Bayesian model averaging (with discussion). *Statistical Science*, **14**, 382–417.

- Holt, D., Scott, A. J. and Ewings, P. D. (1980). Chi-squared Test with Survey Data. *Journal of Royal Statistical Society*, **A143**, 303-320.

- Jeffreys, H. (1935). Some tests of Significance, Treated by the Theory of Probability. *Proceeding of the Cambridge Philosophical Society*, **31**, 203-222.

- Jeffreys, H. (1961). *Theory of Probability.* 3^{rd} edition, Oxford: Oxford University Press.

- Jones, M. C. (1990). Variable Kernel Density Estimates and Variable Kernel Density Estimates. *Australian Journal of Statistics*, **32**, 361-371.
- Kass, R. E. and Raftery, A. E.(1995). Bayes Factor. *Journal of the American Statistical Association*, **90**, 773-795.
- Kass, R. E. and Wasserman, L. (1992). Improving the Laplace Approximation using Posterior Simulation. *Technical Report 566*, Dept. of Statistics, Carnegie Mellon University.
- Kass, R. E. and Wasserman, L. (1995). A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, **90**, 928-934.
- Kish, L. (1965). *Survey Sampling*. New York: Wiley.
- Kish, L. and Frankel, M. R.. (1974). Inference from Complex Samples. *Journal of Royal Statistical Society*, **B63**, 1-37.
- Koch, G. G., Freeman, D. H. and Freeman, J. L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, **43**, 59–78.
- Koehler, K. J. and Wilson, J. R. (1986). Chi-square test for comparing vectors of proportions for several cluster samples. *Communications in Statistics, Theory and methods*, **A15**, 2977-2990.
- Loh, Wei-Yin (1989). Bounds on the Size of the χ^2-Test of Independence in a Contingency Table. *Annals of Statistics*, **17**, 1709-1722.
- Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's Window. *Journal of the American Statistical Association*, **89**, 1535-1545.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, **21**, 1087-1092.

- Molina C., E. A. and Smith, T. M. F. (1986). The effect of sample design on the comparison of associations. *Biometrika*, **73**, 23-33.

- Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the theory of statistics.* 3^{ed}, Singapore: McGraw-Hill.

- Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlation among proportions. *Biometrika*, **49**, 65-82.

- Nathan, G. (1975). Tests of Independence in Contingency Tables from Stratified Proportional Samples. *Sankhyā*, **C37**, 77-87.

- O'Hagan, A. (1994). *Bayesian Inference*; Kendall's Advanced Theory of Statistics, 2B, London: Edward Arnold.

- O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of Royal Statistical Society*, **B57**, 99–138.

- Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics,* **33**, *1065-1076.*

- Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1988). *Numerical Recipes: The Art of Scientific Computing.* Cambridge: Cambridge University Press.

- Raftery, A. E. (1986). Choosing Models for Cross-Classifications. *American Sociological Review*, **51**, 145-146.

- Raftery, A. E. (1995). Bayesian Model Selection in Social Research (with discussion). *Sociological Methodology*, **25**, edited by Peter V. Marsden, Oxford, U.K.: Blackwells, 111-196.

- Raftery, A. E. (1996). Approximation Bayes Factor and Accounting for Model Uncertainty in Generalized linear Models. *Biometrika*, **83**, 251-266.

- Raftery, A. E. (1996b). Hypothesis testing and model selection via posterior simulation. *Markov Chain Monte Carlo in Practice*, edited by Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., London: Chapman and Hall.

- Rao, C. R. (1973). *Linear Statistical Inference and its Applications.* 2^{ed}, New York: Wiley.

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

- Rao, J. N. K. and Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, **12**, 46-60.

- Rao, J. N. K. and Scott, A. J. (1987). On sample adjustments to chi-square tests with sample survey data. *Annals of Statistics*, **15**, 385-397.

- Rao, J. N. K. and Thomas, D. R. (1989). Chi-squared test for contingency table. *Analysis of Complex Surveys*, edited by Skinner, C. J., Holt, D. and Smith, T. M. F., New York: Wiley.

- Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832-837.

- Schwarz, G. (1978). Estimating The Dimension of a Model. *Annals of Statistics*, **6**, 461-464.

- Scott, A. J. and Rao, J. N. K. (1981). Chi-squared tests for contingency tables with proportions estimated from survey data. In *Current Topics in Survey Sampling*, edited by Krewski, D., Platek, R. and Rao, J. N. K. Academic Press, 247-266.

- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York: Wiley.

- Shuster, J. J. and Downing, D. J. (1976). Two-way Contingency Tables for Complex Sampling Schemes. *Biometrika*, **63**, 271-276.

- Skinner, C. J. (1989). Domain means, regression and Multivariate analysis. *Analysis of Complex Surveys*, edited by Skinner, C. J., Holt, D. and Smith, T. M. F., New York: Wiley.

- Skinner, C. J., Holt, D. and Smith, T. M. F. (1989). *Analysis of Complex Surveys.* New York: Wiley.

- Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* New York: Springer.

- Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factor for linear and log-linear models with vague prior information. *Journal of Royal Statistical Society*, **B44**, 377-387.

- Stroud, T. W. F. (1973). Noncentral Convergence of Wald's Large-Sample Test Statistic in Exponential Families. *Annals of Statistics*,**1**, 161-165.

- Tanner, M. A. (1996). *Tools for Statistical Inference : Methods for the Exploration of Posterior Distributions and Likelihood Functions.* 3^{rd} edition, New York: Springer.

- Tapia, R. A. and Thompson, J. R. (1978). *Nonparametric Probability Density Estimation.* Maryland: Johns Hopkins.

- Terrell, G. R. and Scott, D. W. (1992). Variable Kernel Density Estimation. *Annals of Statistics*, **20**, 1236-1265.

- Thomas, D. R. and Rao, J. N. K. (1987). Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics Under Cluster Sampling. *Journal of the American Statistical Association*, **82**, 630-636.

- Thomas, D. R., Singh, A. C. and Roberts, G. R. (1996). Tests of Independence on Two-way tables Under Cluster Sampling: An Evaluation. *International Statistical Review*, **64**, 295-311.

- Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions (in Bayesian Computation). *Annals of Statistics,* **22**, 1701-1728.

- Tierney, L. (1996). Introduction to general state-space Markov chain theory. *Markov Chain Monte Carlo in Practice*, edited by Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., London: Chapman and Hall.

- Tierney, L. and Kadane, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, **81**, 82-86.

- Verdinelli, I. and Wasserman, L. (1995). Computing Bayes Factors Using a Generalization of The Savage-Dickey Density Ratio. *Journal of the American Statistical Association*, **90**, 614-618.

- Walter, G. and Blum, J. (1979). Probability Density Estimation Using Delta Sequences. *Annals of Statistics*, **7**, 328-340.

- Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* London: Chapman and Hall.

- Wasserman, L. (1997). Bayesian Model Selection and Model Averaging. Presented at the *Mathematical Psychology Symposium* in Bloomington, Indiana.

- Wilks, S. S. (1935). The Likelihood Test of Independence in Contingency Tables. *Annals of Mathematical Statistics*, **6**, 190-196.

- Wilson, J. R. and Warde, W. D. (1991). An Adjusted Test for Cross-Classified Cluster Sample Data. *Communications in Statistics*, **20**, 3029-3050.