A BAYESIAN TEST OF ASSOCIATION IN A TWO-WAY CATEGORICAL TABLE WITH INTRA-CLASS CORRELATION

Balgobin Nandram[1] and Jai Won Choi[2]
National Center for Health Statistics, 3311 Toledo Road, Hyattsville MD 20782

Abstract:

It is straight forward to analyze data from a single multinomial table. Specifically, for the analysis of a two-way categorical table, the common chi-squared test of independence between the two variables and maximum likelihood estimators are readily available. When the counts in the two-way categorical table are formed from familial data (clusters of correlated data), the common chi-squared test no longer applies. We note that there are several approximate adjustments to the common chi-squared test. However, our main contribution is the construction and analysis of a Bayesian model which removes all analytical approximations. This is an extension of a standard multinomial-Dirichlet model to include the intra-class correlation associated with the individuals within a cluster. This intra-class correlation varies with the size of the cluster, but we assume that it is the same for all clusters of the same size for the same variable. We use Markov chain Monte Carlo methods to fit our model, and to make posterior inference about the intra-class correlations and the cell probabilities. We use data from the National Health Interview Survey to show how our alternative test performs and to obtain the posterior density of the cell probabilities. Also, using Monte Carlo integration, we obtain the Bayes factor for a test of no association.

Key words: Bayes factor, Gibbs sampler, Monte Carlo integration, Multinomial-Dirichlet.

1. Introduction

It is a common practice to use two-way categorical tables to present survey data. In this situation it is assumed that the cell counts in the $r \times c$ table follow a multinomial distribution. However, because of stratification and clustering the joint distribution of the cell counts is no longer multinomial. Thus, the standard chi-squared statistic no longer has a chi-squared distribution, and therefore the test based on the multinomial distribution may be inadequate. It is standard practice to make an adjustment to the standard chi-squared statistic, but in general the accuracy of this adjustment is not well understood, and one can not estimate the cell probabilities based on this adjustment. We propose a Bayesian alternative which is based on the Bayes factor to obtain a test for association between the two categorical variables. Our Bayesian method also provides posterior distributions for the cell probabilities.

Several authors have recognized inaccuracy in the analysis when the usual chi-squared test is applied to correlated "multinomial" data. Efforts to correct for spurious inflation in such tests have been based on two approaches. The design-based approach provides inference with respect to the asymptotic sampling distribution of estimates over repetitions of the sample design (Fellegi 1980, Holt, Scott and Ewings 1980, Rao and Scott 1981, 1984, Bedrick 1983, and Fay 1985). For example, Rao and Scott (1981) investigate the effects of stratification and clustering on the asymptotic distribution of Pearson's chi-squared statistic for goodness of fit and independence. They propose new measures called generalized design effects. See also Rao and Scott (1984) who generalized the results of Rao and Scott (1981) to multi-way categorical tables. The model-based approach postulates a probability distribution to model the sample data (Altham 1976, Cohen 1976, Brier 1980, Fienberg 1979, and Choi and McHugh 1989). For example, Choi and McHugh (1989), applying the probabilistic development in Altham (1976), shows how to adjust the standard chi-squared test statistic when there is an intra-class correlation and data are weighted. Of less relevance, we also note a recent activity on the adjustment of the chi-squared test statistic when there is missing data under stratified random sampling.

Let n_{jk} denote the number of individuals in the j^{th} row and k^{th} column of the $r \times c$ categorical table. Also let $n_{j.} = \sum_{k=1}^{c} n_{jk}$, $j = 1, \ldots, r$, $n_{.k} = \sum_{j=1}^{r} n_{jk}$, $k = 1, \ldots, c$, $n = \sum_{j=1}^{r} \sum_{k=1}^{c} n_{jk}$ and $e_{jk} = n_{j.}n_{.k}/n$, $j = 1, \ldots, r$, $k = 1, \ldots, c$. Then, Pearson's chi-squared statistic, under inde-

[1] Balgobin Nandram is on sabbatical leave from the Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute road, Worcester, MA 01609.

[2] Contact author.

pendence of the row and column classification, is

$$X_u = \sum_{j=1}^{r}\sum_{k=1}^{c}(n_{jk} - e_{jk})^2/e_{jk}.$$

If the responses from the individual members are independent and identically distributed, then asymptotically (as $n \to \infty$) $X_u \to \chi^2_{(r-1)(c-1)}$, a chi-squared random variable with $(r-1)(c-1)$ degrees of freedom. In practice, the validity of the chi-squared test depends on (a) the magnitude of the expected values e_{jk} and (b) whether the cell counts $(n_{jk},\ j=1,\ldots,r,\ k=1,\ldots,c)$ follow a multinomial distribution given the sample size n (i.e., the individual responses are independent and identically distributed). In (a) the test is valid if the e_{jk} are larger than 5, and clearly the only way to achieve this is to increase the sample size subject to cost. In (b) when there is correlation among the members (e.g., familial correlation), the asymtotic distribution of X_u is no longer $\chi^2_{(r-1)(c-1)}$, and the estimates of the cell proportions can be inaccurate. The problem about the asymtotic distribution has received much attention, but the problem about the inaccuracy of the estimates of the cell proportions has received virtually no attention. We address both problems simultaneously within a Bayesian framework in this paper when there are familial count data.

We describe one solution that has been proposed for the problem about the asymptotic distribution. Let n_t denote the number of members in all families of the same size $t = 1,\ldots,T$, and let θ_t denote the intra-class correlation for clusters of size t $(\theta_1 \equiv 0)$. Motivated by Rao and Scott (1981), Choi and McHugh (1989) derive the following adjusted chi-squared statistic

$$X_a = X_u\{1 + n^{-1}\sum_{t=1}^{T}(t-1)n_t\theta_t\}^{-1}$$

which is more accurately $\chi^2_{(r-1)(c-1)}$. The p-value corresponding to the adjusted chi-squared statistic will be larger.

We provide a Bayesian analysis of this problem. This is a direct extension of the probabilistic development in Altham (1976) which is used to provide a likelihood function. Then proper but noninformative priors are assigned to the parameters to provide a full Bayesian approach. The model includes a nonnegative intra-class correlation which varies according to the number of individuals in a cluster (i.e., all clusters of the same size have the same intra-class correlation). In this framework we can provide (a) the posterior densities of the cell probabilities and (b) a test of association between the two categorical variables. For weighted data they further adjust $\chi^2_{(r-1)(c-1)}$ appromiately by the average weight.

In (b) we use the Bayes factor to quantify the difference between a model with association and one without. This is the ratio of the prior odds of one model to the other to their posterior odds (obtained through the use of Bayes' theorem), and it is the same as the ratio of the marginal likelihoods of the data under two models, one without association and the other with association. See Kass and Raftery (1995) for a very informative discussion about the Bayes factor and a rule of thumb for quantifying the degree of evidence. There are several methods to compute the marginal likelihood (e.g., see Section 1 of Chib and Jeliazkov 2001), and we note that one standard method is Monte Carlo integration using an importance function.

In this paper, we introduce a Bayesian method to analyze data from an $r \times c$ categorical table. We consider the situation in which there are no missing data, but one in which the table is built up by aggregating clustered multinomial data. In Section 2, we describe the methodology to obtain estimates of the cell probabilities, and to obtain the Bayes factor for a test of no association between the two categorical variables. We also show how to use Markov chain Monte Carlo methods to fit the models. We show how to use Monte Carlo integration with an importance function to compute the marginal likelihoods under different models. In Section 3 we illustrate our method using data from the National Health Interview Survey. In Section 4, we perform several simulated examples to compare inference using our model with another model which does not incorporate the intra-class correlation. Finally, Section 5 has concluding remarks.

2. Bayesian Methodology

We describe the methodology to fit "multinomial" data when there is an intra-class correlation. We build our model based on the work of Altham (1976).

2.1 Model

Suppose that there are s_i individuals in the i^{th} cluster, $i = 1,\ldots,\ell$, and s_{ijk} individuals fall in the j^{th} row and k^{th} column in the r x c table, j=1,...r, k=1,...,c. Here $\sum_{j=1}^{r}\sum_{k=1}^{c} s_{ijk} = s_i$, $s_{ijk} \geq 0$. Altham (1976) shows that the probability that all s_i individuals fall in the j^{th} row and k^{th} column is

$$\theta_{s_i}\pi_{jk} + (1-\theta_{s_i})\pi_{jk}^{s_i} \qquad (1)$$

and the probability that the individuals are in different *specified* cells is

$$(1-\theta_{s_i})\prod_{j=1}^{r}\prod_{k=1}^{c}\pi_{jk}^{s_{ijk}} \qquad (2)$$

where we allow the intraclass correlation θ_{s_i}, $0 \leq \theta_{s_i} \leq 1$, to depend on the cluster size s_i. Note that (1) can be interpreted as a mixture of two distributions. Let w_{s_i} be the latent variable

$$w_{s_i} = \begin{cases} 1, & \text{perfect dependence} \\ 0, & \text{perfect independence} \end{cases}$$

where $p(w_{s_i}=1 \mid \theta_{s_i}) = 1 - p(w_{s_i}=0 \mid \theta_{s_i}) = \theta_{s_i}$, and dependence/independence refers to the intraclass correlation. Note also that the notation in Choi and McHugh (1989) is slightly different.

This model of clustering permits only positive association or independence among the individuals within a cluster, and this is typically the case for many demographic, social and economic characteristics.

Note that $\theta_{s_i}\pi_{jk} + (1-\theta_{s_i})\pi_{jk}^{s_i}$ is strictly increasing in θ_{s_i}. When $\theta_{s_i}=0$, the probability that all individuals in the i^{th} cluster belong to cell (j,k) is $\pi_{jk}^{s_i}$, and when $\theta_{s_i}=1$, the probability that all individuals in the i^{th} cluster belong to cell (j,k) is π_{jk}, which can be much larger. In addition, $(1-\theta_{s_i})\prod_{j=1}^{r}\prod_{k=1}^{c}\pi_{jk}^{s_{ijk}}$ is a strictly decreasing function in θ_{s_i}. When $\theta_{s_i}=0$, the probability that the individuals in the i^{th} cluster belong to different specified cells is $\pi_{jk}^{s_i}$, and when $\theta_{s_i}=1$, the probability that the individuals in the i^{th} cluster belong to different specified cells is 0. Thus, the intra-class correlation has an important role when inference is made about the π_{jk} and the association between the two categorical variables.

Let $\mathcal{C}$ denote the set of clusters in which all individuals fall in a single cell of the r x c table. Then letting $\underset{\sim}{s_i} = \{s_{ijk}\}$,

$$p(\underset{\sim}{s_i} \mid s_i, \theta_{s_i}, \underset{\sim}{\pi}) = \begin{cases} \theta_{s_i}\pi_{jk} + (1-\theta_{s_i})\pi_{jk}^{s_i}, & i \in \mathcal{C} \\ (1-\theta_{s_i})s_i!\prod_{j,k}^{r,c}\pi_{jk}^{s_{ijk}}/s_{ijk}! & i \notin \mathcal{C}. \end{cases} \qquad (3)$$

Assuming independence over clusters, we have

$$p(\underset{\sim}{s} \mid s_1,\ldots,s_\ell,\underset{\sim}{\theta},\underset{\sim}{\pi}) = \prod_{i\in\mathcal{C}}\prod_{j,k}^{r,c}\{\theta_{s_i}\pi_{jk} + (1-\theta_{s_i})\pi_{jk}^{s_i}\}$$

$$\times\prod_{i\notin\mathcal{C}}\{(1-\theta_{s_i})s_i!\prod_{j,k}^{r,c}\pi_{jk}^{s_{ijk}}/s_{ijk}!\}.$$

Observe that if $\theta_{s_i}=0,\ i=1,\ldots,\ell$,

$$p(\underset{\sim}{s} \mid s_1,\ldots,s_\ell,\underset{\sim}{\theta},\underset{\sim}{\pi}) = \prod_{i=1}^{\ell}\{s_i!\prod_{j,k}^{r,c}\pi_{jk}^{s_{ijk}}/s_{ijk}!\},$$

which is a product of multinomial probability functions and the statistics $\sum_{i=1}^{\ell}s_{ijk} = n_{jk}$ are sufficient as in regular multinomial sampling (i.e., observations are from a simple random sample) and each individual belongs to cell (j,k) with probability $\pi_{jk}\geq 0$, $\sum_{j=1}^{r}\sum_{k=1}^{c}\pi_{jk}=1$.

Suppose that each cluster has size t, t=1,...,T; in applications T is typically 2 to 4 or so. Then letting g_{tjk} denote the number of clusters in $\mathcal{C}$ of size t with all individuals in cell (j,k) and $\tilde{g}_t$ the number of clusters of size t in $\tilde{\mathcal{C}}$ (i.e., outside $\mathcal{C}$),

$$p(\underset{\sim}{s} \mid s_1,\ldots,s_\ell,\underset{\sim}{\theta},\underset{\sim}{\pi}) \propto \prod_{t=1,j=1,k=1}^{T,r,c}(\theta_t\pi_{jk}+(1-\theta_t)\pi_{jk}^{t})^{g_{tjk}}$$

$$\times\{\prod_{t=1}^{T}(1-\theta_t)^{\tilde{g}_t}\}\prod_{i\notin\mathcal{C}}\{s_i!\prod_{j=1}^{r}\prod_{k=1}^{c}\pi_{jk}^{s_{ijk}}/s_{ijk}!\}.$$

Finally for a full Bayesian approach, noting that $\theta_1=0$, we assume

$$\theta_t \overset{iid}{\sim} \text{Uniform(0,1)},\ t=2,\ldots,T$$

and independently

$$\underset{\sim}{\pi} \sim \text{Dirichlet}(\underset{\sim}{1}).$$

These are noninformative but proper prior densities.

Observe that

$$(\theta_t\pi_{jk}+(1-\theta_t)\pi_{jk}^{t})^{g_{tjk}} =$$

$$\sum_{z_{tjk}=0}^{g_{tjk}}\binom{g_{tjk}}{z_{tjk}}(\theta_t\pi_{jk})^{z_{tjk}}\{(1-\theta_t)\pi_{jk}^{t}\}^{g_{tjk}-z_{tjk}}.$$

The latent variables z_{tjk} simplify the computation because they replace the mixture with a product and they provide more accessible conditional densities (see Robert and Casella 1999 for the demarginalization trick). Thus, incorporating the latent variables into our model, the joint posterior density of $\underset{\sim}{\theta},\underset{\sim}{\pi},\underset{\sim}{z}$ given $\underset{\sim}{s}$ is $p(\underset{\sim}{\theta},\underset{\sim}{\pi},\underset{\sim}{z} \mid \underset{\sim}{s}) \propto$

$$\prod_{t=1}^{T}(1-\theta_t)^{\tilde{g}_t}\prod_{j=1}^{r}\prod_{k=1}^{c}\binom{g_{tjk}}{z_{tjk}}$$

$$\times(\theta_t\pi_{jk})^{z_{tjk}}\{(1-\theta_t)\pi_{jk}^{t}\}^{g_{tjk}-z_{tjk}}\prod_{j,k}^{r,c}\pi_{jk}^{\tilde{s}_{jk}}$$

where $\tilde{s}_{jk} = \sum_{i \notin C} s_{ijk}$.

2.2 Computation

The joint posterior density is complex, so we use the Gibbs sampler to draw samples which are used to make inference about π_{jk} and θ_t.

To run the Gibbs sampler we need starting values for $\underset{\sim}{\theta}$ and $\underset{\sim}{\pi}$, and these are easy to obtain. Letting $n_{jk} \tilde{=} \sum_{i=1}^{\ell} \tilde{} s_{ijk}$ and $n = \sum_{j=1}^{r} \sum_{k=1}^{c} n_{jk}$, we take $\hat{\pi}_{jk} = n_{jk}/n$ and $\theta_t = 1/t,\ t = 2, ..., T$. We also take $z_{tjk} = g_{tjk}[\theta_t \pi_{jk}/\{\theta_t \pi_{jk} + (1-\theta_t)\pi_{jk}^t\}]$.

The conditional posterior densities (cpd's) of each parameter given the others are needed to implement the Gibbs sampler. Note that $z_{1jk} = \theta_1 = 0$. Specifically, the cpd for $\underset{\sim}{\theta}$ is

$$\theta_t \mid \underset{\sim}{\pi}, \underset{\sim}{z}, \underset{\sim}{s} \overset{ind}{\sim} \text{Beta} \left\{ 1 + \sum_{j,k}^{r,c} z_{tjk}, \right.$$

$$\left. 1 + \tilde{g}_t + \sum_{j,k}^{r,c} (g_{tjk} - z_{tjk}), \right\} \ t = 2, \ldots, T,$$

the cpd for $\underset{\sim}{\pi}$ is

$$\underset{\sim}{\pi} \mid \underset{\sim}{\theta}, \underset{\sim}{z}, \underset{\sim}{s} \sim \text{Dirichlet} \left\{ 1 + g_{1jk} + \sum_{t=1}^{T} [z_{tjk} + \right.$$

$$t(g_{tjk} - z_{tjk})] + \tilde{s}_{jk},\ j = 1, \ldots, r,\ k = 1, \ldots, c\}$$

and the cpd for $\underset{\sim}{z}$ is, $t = 2, \ldots, T,\ j = 1, \ldots, r,\ k = 1, \ldots, c$,

$$z_{tjk} \mid \underset{\sim}{\theta}, \underset{\sim}{\pi}, \underset{\sim}{s} \overset{ind}{\sim} \text{Binomial} \left\{ g_{tjk}, \frac{\theta_t \pi_{jk}}{\theta_t \pi_{jk} + (1-\theta_t)\pi_{jk}^t} \right\}.$$

We "burn in" 1000 iterates, and took every tenth to get 1000 iterates which we use for inference. These choices are very conservative, and the algorithm runs very quickly.

2.3 Inference

To test for association between the two categorical variables, we use the Bayes factor, the ratio of the two marginal likelihoods. A problem of the slightly less interest is to test for no intra-class correlation.

Consider our problem with intra-class correlation. For the model with association, taking $\theta_1 = 0$, the marginal likelihood is

$$p\text{as}(\underset{\sim}{s}) = (rc-1)! \int \int \prod_{t,j,k}^{T,r,c} \{\theta_t \pi_{jk} + (1-\theta_t)\pi_{jk}^t\}^{g_{tjk}}$$

$$\times \{\prod_{t=1}^{T} (1-\theta_t)^{\tilde{g}_t}\} \prod_{i \notin C} \{s_i! \prod_{j,k}^{r,c} \frac{\pi_{jk}^{s_{ijk}}}{s_{ijk}!}\} \mathrm{d}\underset{\sim}{\theta} \mathrm{d}\underset{\sim}{\pi}$$

and for the model without association the marginal likelihood is

$$p\text{nas}(\underset{\sim}{s}) = (r-1)!(c-1)! \int \int \int \prod_{t,j,k}^{T,r,c} \{\theta_t q_j^{(1)} q_k^{(2)} +$$

$$(1-\theta_t)(q_j^{(1)} q_k^{(2)})^t\}^{g_{tjk}} \{\prod_{t=1}^{T} (1-\theta_t)^{\tilde{g}_t}\}$$

$$\times \prod_{i \notin C} \{s_i! \prod_{j,k}^{r,c} \frac{(q_j^{(1)} q_k^{(2)})^{s_{ijk}}}{s_{ijk}!}\} \mathrm{d}\underset{\sim}{q}_j^{(1)} \mathrm{d}\underset{\sim}{q}_k^{(2)} \mathrm{d}\underset{\sim}{\theta}$$

where $\pi_{jk} = q_j^{(1)} q_k^{(2)}$ and $\sum_{j=1}^{r} q_j^{(1)} = \sum_{k=1}^{c} q_k^{(2)} = 1$.

Then, it is easy to show that

$$p\text{as}(\underset{\sim}{s}) = (rc-1)! \prod_{i \notin C} \{s_i! / \prod_{j,k}^{r,c} s_{ijk}!\}$$

$$\times \prod_{t=2}^{T} B\{\sum_{j,k}^{r,c} g_{tjk} + 1, \hat{g}_t + 1\}\ D(\underset{\sim}{a})$$

$$\times \int_{\underset{\sim}{\theta}} \int_{\underset{\sim}{\pi}} \prod_{t=2,j,k}^{T,r,c} [1 + \frac{1-\theta_t}{\theta_t} \pi_{j,k}^{t-1}]^{g_{tjk}}$$

$$\times \prod_{t=2}^{T} \{\theta_t^{g_{t++}} (1-\theta_t)^{\hat{g}_t} / B(\sum_{j,k}^{r,c} g_{tjk} + 1, \hat{g}_t + 1)\}$$

$$\times \prod_{j,k}^{r,c} \pi_{jk}^{(a_{jk}-1)} / D(\underset{\sim}{a}) d\underset{\sim}{\theta} d\underset{\sim}{\pi}, \qquad (4)$$

where $a_{jk} = \sum_{t=1}^{T} g_{tjk} + \sum_{i \notin C} s_{ijk} + 1, j = 1, \ldots, r,\ k = 1, \ldots, c$, and $g_{t++} = \sum_{j,k}^{r,c} g_{tjk}$. It is also easy to show that and

$$p\text{nas}(\underset{\sim}{s}) = (r-1)!(c-1)! \prod_{i \notin C} \{s_i! / \prod_{j=1}^{r} \prod_{k=1}^{c} s_{ijk}!\}$$

$$\times \prod_{t=2}^{T} B\{\sum_{j,k}^{r,c} g_{tjk} + 1, \tilde{g}_t + +1\}\ D_1(\underset{\sim}{a}^{(1)}) D_2(\underset{\sim}{a}^{(2)})$$

$$\times \int_{\underset{\sim}{\theta}} \int_{\underset{\sim}{q}^{(1)} \underset{\sim}{q}^{(2)}} \prod_{t=2,j,k}^{T,r,c} [1 + \frac{1-\theta_t}{\theta_t} (\underset{\sim}{q}^{(1)} \underset{\sim}{q}^{(2)})^{t-1}]^{g_{tjk}}$$

$$\times \prod_{t=2}^{T} \{ \frac{\theta_t^{g_{t++}} (1-\theta_t)^{\hat{g}_t}}{B(\sum_{j,k}^{r,c} g_{tjk} + 1, \hat{g}_t + 1)} \}$$

$$\times \prod_{j=1}^{r} \frac{q_j^{(1)(a_{j.}^{(1)}-1)}}{D(\underset{\sim}{a}^{(1)})} \frac{q_k^{(2)(a_{k.}^{(2)}-1)}}{D(\underset{\sim}{a}^{(2)})} d\underset{\sim}{\theta} d\underset{\sim}{q}^{(1)} d\underset{\sim}{q}^{(2)}, \qquad (5)$$

where $a_j^{(1)} = \sum_{k=1}^{c} a_{jk}$ and $a_k^{(2)} = \sum_{j=1}^{r} a_{jk}$, $D_1(\underset{\sim}{a}) = D(a_{1.} + 1, \ldots, a_{r.} + 1)$ and $D_2(\underset{\sim}{a}) = D(a_{.1} + 1, \ldots, a_{.k} + 1)$.

We use Monte Carlo integration to execute (4) and (5) with previous importance functions.

We have chosen the Monte Carlo sample size to M=10,000. Numerical standard error can be obtained in an obvious manner.

3. Examples from the NHIS

The National Health Interview Survey (NHIS) has been conducted every year since 1957 by the National Center for Health Statistics (NCHS) to measure an aspect of health status of the U.S. population (see Adams and Marano 1995). Through this sample survey, NCHS conducts surveys on chronic and acute conditions, doctor visits, hospital episodes, disability, household and personal information, and other special aspects of health of the U.S. population. One of the variables we use in the NHIS is activity limitation status, and the relation between age and activity limitation status is of interest.

Activity limitation status (ALS) is a measure of long-term disability resulting from chronic conditions. It is defined as inability to carry out the major activity for one's age-sex group such as working, keeping house or going to school; restriction in the amount or kind of major activity; or restriction in relation to other activities such as recreational, church and civic interests. ALS has served as a measure of long-term disability since the inception of the Health Interview Survey in 1957. ALS is typically classified into three categories: "unable to perform major activity", "limited in kind/amount major activity and in other activities" and "not limited (includes unknowns)" ranging from severe individuals to individuals unnecessary to classify.

In the health interview survey, information (i.e., chronic disease and impairment) for each household member about the major activity he usually performed during the 12 months prior to interview is requested by the interviewer. Age is an important determinant of ALS; there is a positive association between ALS and age. To study the relation between age and ALS three age groups (under 56 years, 56-70 years and more than 70 years) are used.

The households are poststratified by states and there are data from all 51 states (including the District of Columbia). For some states there are extremely small numbers of sampled households (e.g., Iowa, Idaho, Wyoming) and for some states there are extremely large numbers of sampled households (e.g., California, New York, Texas). We study these states individually and we report results for Iowa, Maryland (medium size state) and California to illustrate how our procedure performs. It is of general interest to test the hypothesis that age and ALS are independent and to estimate the proportion of individuals in each cell of the 3×3 table. We present the 3×3 tables for Iowa, Maryland and California in Table 1.

In Table 2 we compare the correlations. We can see that the θ_k are moderately large and the 95% credible intervals narrow down from Iowa to California (i.e., inference is much sharper for California). At least for Maryland and California one needs to evaluate the impact of these θ_k.

In Table 3, there is very strong evidence of a positive association between age and ALS for Maryland and California using both the Bayes factor and the chi-squared tests. However, the difference between the two models does not matter for California, but there is a small difference for Maryland. For Iowa it appears that there is no association between age and ALS; the chi-squared test shows the contrary with some expected cell counts smaller than 5 (defective chi-squared test).

In Table 5 we present summaries of the posterior distribution of the π_{jk}. In general, inference is very sharp even for Iowa. It is very interesting that inference about the π_{jk} can differ under the model with intra-class correlation and the one without. For example, the 95% credible intervals of π_{31} are (.32, .39) versus (.63, .72) for Iowa, (.32, .36) versus (.67, .73) for Maryland and (.64, .66) versus (.75, .77) for California.

4. Simulated Example

We have simulated data from our model to assess how changes in the intra-class correlation affect (a) inference about the π_{jk}.

We have chosen the π_{jk} to represent association between the categorical variables in a 3×3 table. Specifically, we choose $\pi_{11} = .05$, $\pi_{12} = .10$, $\pi_{13} = .15$, $\pi_{21} = .15$, $\pi_{22} = .10$, $\pi_{23} = .05$, $\pi_{31} = .20$, $\pi_{32} = .15$, and $\pi_{33} = .05$. Letting c_k denote the number of clusters of size k, $k = 1, \ldots, T = 5$, we take $c_1 = 50$, $c_2 = 150$, $c_3 = 100$, $c_4 = 50$ and $c_5 = 50$ to get a total of 1100 observations. We have taken $\theta_k = \theta$, $k = 2, \ldots, T(T = 5)$, and we study 5 values of θ (.1, .3, .5, .7, .9). Thus, we study the effect of our choice of θ on inference about π_{jk} and the association between the two categorical variables. We note that when θ is small (large), there is a tendency for the simulated individuals to be in different (same) cell(s) of the $r \times c$ table.

We have simulated a single data set of size 1100 for each value of θ. We fit our model to each data set using the Gibbs sampler as described in Section 2.2.

We have presented results for our simulated examples in Tables 6 and 4. As for the π_{jk} the posterior summaries indicate that the model with the intra-class correlation is more in concordant with the design values than the regular multinomial, and the regular multinomial degrades as θ increases. We also note that apart from $\theta = 10$, the posterior sumaries are concordant with the design values for θ_2 and θ_5.

5. Concluding Remarks

We have shown how to analyze multinomial data from $r \times c$ categorical tables when there is an intra-class correlation. We have also shown that by using the Bayes factor (ratio of the marginally likelihoods of two models) we can test for association between the two categories.

We have analyzed 3×3 categorical data of age and activation limitation status from the 1996 National Health Interview Survey. We have found moderately large intra-class correlations, and these correlations have small effects on tests of hypothesis (both the standard chi-squared test and our Bayesian alternative). We have reported results for Iowa, Maryland and California, and shown that inference for Iowa is problematic due to the sparseness of the data.

In future we can extend our methodology to accommodate (a) small areas (b) nonresponse and (c) an intra-class correlation coefficient corresponding to each categorical variable. In (a) we can consider the states (including the District of Columbia) as small areas. There are very sparse data from some of the states (e.g., Iowa, Idaho, Wyoming), and to make reliable inference about one of these states, one needs to "borrow strength" across the states. In (b) there is a non-negligible number of nonrespondents from each state, and one would need to construct a model that can adjust for nonignorable nonresponse. Finally, in (c) we can replace θ_{s_i} in Altham's formula by $\frac{1}{2}(\kappa_{s_i} + 1)\theta_{s_i}$, $0 \leq \theta_{s_i} \leq 1$, $0 < \kappa_{s_i} \leq \theta_{s_i}^{-1}$, $i = 1, \ldots, \ell$ with an appropriate joint prior density on $(\theta_{s_i}, \kappa_{s_i})$.

BIBLIOGRAPHY

Adams, P. F. and Marano, M. A. (1995), "Current estimate from the National Health Interview Survey, 1994," Series **10** (193), National Center for Health Statistics, CDC.

Altham, P. M. (1976), "Discrete variable analysis for individuals grouped into families," Biometrika, 63, 263-269.

Bedrick, E. J. (1983), "Adjusted chi-squared tests for cross-classified tables of survey data," Biometrika, 70, 591-595.

Brier, S. E. (1980), "Analysis of contingency tables under cluster sampling," Biometrika, 67, 591-596.

Chen, T. and Fienberg, S. E. (1974), "Two-dimensional contingency tables with both completely and partially cross-classified data," Biometrics, 30, 629-642.

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

Choi, J. W. and McHugh, R. B. (1989), "A reduction factor in goodness-of-fit and independence for clustered and weighted observations," Biometrics, 45, 979-996.

Cohen, J. E. (1976), "The distribution of the chi-squared under cluster sampling from contingency tables," Journal of the American Statistical Association, 71, 591-596.

Fellegi, I. P. (1980), "Approximate tests of independence and goodness of fit based upon stratified multistage samples," Journal of the American Statistical Association, 75, 261-268.

Fay, R. E. (1985), "Complex samples," Journal of the American Statistical Association, 80, 148-157.

Fienberg, S. E.(1979), "The use of chi-squared statistics for categorical data problems," Journal of the Royal Statistical Society, Series B, 41, 54-64.

Holt, D. and Scott, A. J. and Ewings, P. O. (1980), "Chi-squared tests with survey data," Journal of the Royal Statistical Society, Series A, 143, 302-320.

Kass, R. and Raftery, A. (1996), "Bayes factors," Journal of the American Statistical Association, 90, 773-795.

Rao, J. N. K. and Scott, A. J. (1981), "The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables," Journal of 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," The Annals of Statistics, 12, 46-60.

Robert, C. P. and Casella, G. (1999), Monte Carlo Statistical Methods, Springer-Verlag: New York.

Wang, H. (2001), Two-way contingency tables with marginally and conditionally imputed nonrespondents, Ph.D. Dissertation, Department of Statistics, University of Wisconsin-Madison.

Table 1: Classification of sampled individuals in 1996 NHIS for Iowa, Maryland and California by age and activity limitation status (ALS)

			ALS		
State	Age	1	2	3	Total
Iowa	1	7	21	264	292
	2	6	3	42	51
	3	4	10	27	41
	Total	17	34	333	384
Maryland	1	13	60	633	706
	2	13	20	81	114
	3	3	23	51	77
	Total	29	103	765	897
California	1	176	378	5638	6192
	2	106	123	490	719
	3	51	120	278	449
	Total	333	621	6406	7360

NOTE: Age (1: under 56 years; 2: 56-70 years; 3: over 70 years) and ALS (1: unable to perform major activity; 2: limited in kind/amount major activity and limited in other activities; 3: not limited (includes unknowns).

Table 2: Comparison of the posterior mean (PM), and 95% credible intervals (CI) for the intra-class correlations (i.e., $\theta_k,\ k = 2, \ldots, 5$) for Iowa, Maryland and California

	IA	MD	CA
k	PM(CI)	PM(CI)	PM(CI)
2	.43(.2, .6)	.46(.3, .6)	.33(.3, .4)
3	.58(.4, .8)	.61(.5, .7)	.37(.3, .4)
4	.56(.4, .7)	.63(.5, .8)	.60(.5, .7)
5	.82(.6, 1)	.44(.2, .7)	.58(.5, .6)

Table 3: Comparison of the Bayes factors (BF) and chi-squared tests with or without intra-class correlation for Iowa, Maryland and California

Test	BF_1	BF_2	χ^2_u	χ^2_a
IA	35	2	27(0.00)	19(0.00)
MD	7	16	67(0.00)	46(0.00)
CA	205	218	606(0.00)	425(0.00)

NOTE: The Bayes factors are on the log-scale. The p-value of the chi-squared tests are in parentheses. BF (1=with intra-class correlation; 2=without intra-class correlation) and χ^2(u=unadjusted; a=adjusted).

Table 4: Logarithm of Bayes factors under two models, the model with intra-class correlation and the regular multinomial model, and the adjusted chi-squared statistic for five values of intra-class correlation (.1, .3, .5, .7, .9)

	Log Bayes Factor			
θ	IC	No IC	χ^2_u	χ^2_a
0.1	30	76	169	165
0.3	28	101	208	183
0.5	31	101	212	159
0.7	20	66	145	102
0.9	22	48	111	72

NOTE: The Bayes factor is the ratio of the marginal likelihood for a model with association (i.e., no restriction on π_{ij}, $\sum_{j=1}^{r}\sum_{k=1}^{c}\pi_{jk} = 1$) to the marginal likelihood for a model with no association (i.e.,$\pi_{ij} = \pi_j\ \pi_k$, $\sum_{j=1}^{r}\pi_j = \sum_{k=1}^{c}\pi_k = 1$). χ^2_u and χ^2_a are respectively the unadjusted and adjusted χ^2 tests. The p-values are zeros for both χ^2_u and χ^2_a. IC=Intra-cluster correlation.

Table 5: Comparison of the posterior means (PM), and 95% credible intervals (CI) for π_{jk} with and without intra-class correlation for Iowa, Maryland and California

	Intra-class		No intra-class	
Cell	PM	CI	PM	CI
Iowa				
(1, 1)	.082	(.06, .10)	.020	(.01, .04)
(1, 2)	.044	(.03, .06)	.017	(.01, .03)
(1, 3)	.038	(.03, .05)	.012	(.00, .02)
(2, 1)	.119	(.10, .14)	.056	(.04, .08)
(2, 2)	.065	(.05, .08)	.010	(.00, .02)
(2, 3)	.060	(.05, .08)	.028	(.01, .05)
(3, 1)	.351	(.32, .39)	.674	(.63, .72)
(3, 2)	.139	(.12, .17)	.110	(.08, .14)
(3, 3)	.100	(.08, .12)	.072	(.05, .10)
Maryland				
(1, 1)	.092	(.08, .10)	.015	(.01, .02)
(1, 2)	.059	(.05, .07)	.016	(.01, .03)
(1, 3)	.030	(.02, .04)	.004	(.00, .01)
(2, 1)	.128	(.11, .14)	.067	(.05, .08)
(2, 2)	.070	(.06, .08)	.023	(.01, .03)
(2, 3)	.062	(.05, .07)	.027	(.02, .04)
(3, 1)	.340	(.32, .36)	.700	(.67, .73)
(3, 2)	.135	(.12, .15)	.090	(.07, .11)
(3, 3)	.084	(.07, .10)	.058	(.04, .07)
California				
(1, 1)	.041	(.04, .05)	.024	(.02, .03)
(1, 2)	.023	(.02, .03)	.015	(.01, .02)
(1, 3)	.013	(.01, .02)	.007	(.005, .009)
(2, 1)	.083	(.08, .09)	.051	(.05, .06)
(2, 2)	.029	(.02, .03)	.017	(.01, .02)
(2, 3)	.026	(.02, .03)	.016	(.01, .02)
(3, 1)	.650	(.64, .66)	.765	(.75, .77)
(3, 2)	.090	(.08, .10)	.067	(.06, .07)
(3, 3)	.046	(.04, .05)	.038	(.03, .04)

NOTE: For 9 cells

Table 6: Posterior means (PM) of cell probabilities, and 95% credible intervals (CI) for π_{jk} under two models: with and without intra-class correlation for three values of intra-class correlation (0.1, 0.5, 0.9)

			IC correlation		NIC correlation	
θ	Cell	π	PM	CI	PM	CI
.1	1	.05	.05	(.04, .07)	.04	(.03, .06)
	2	.10	.11	(.09, .13)	.11	(.09, .12)
	3	.15	.14	(.12, .17)	.16	(.04, .18)
	4	.15	.15	(.13, .17)	.15	(.13, .17)
	5	.10	.09	(.07, .11)	.10	(.08, .12)
	6	.05	.05	(.04, .07)	.05	(.04, .06)
	7	.20	.21	(.18, .23)	.21	(.19, .24)
	8	.15	.15	(.13, .17)	.14	(.12, .16)
	9	.05	.05	(.04, .07)	.04	(.03, .06)
.5	1	.05	.05	(.04, .06)	.06	(.04, .07)
	2	.10	.11	(.10, .13)	.11	(.09, .13)
	3	.15	.14	(.12, .15)	.14	(.12, .16)
	4	.15	.14	(.13, .16)	.15	(.13, .17)
	5	.10	.11	(.09, .12)	.11	(.09, .12)
	6	.05	.06	(.05, .07)	.06	(.04, .07)
	7	.20	.19	(.17, .21)	.18	(.16, .20)
	8	.15	.15	(.13, .17)	.16	(.13, .18)
	9	.05	.06	(.05, .07)	.06	(.05, .07)
.9	1	.05	.07	(.05, .09)	.05	(.04, .07)
	2	.10	.08	(.06, .11)	.08	(.07, .10)
	3	.15	.14	(.11, .17)	.13	(.11, .15)
	4	.15	.13	(.10, .16)	.15	(.12, .17)
	5	.10	.09	(.07, .12)	.10	(.08, .12)
	6	.05	.06	(.04, .08)	.04	(.03, .06)
	7	.20	.23	(.19, .26)	.22	(.19, .24)
	8	.15	.15	(.11, .18)	.17	(.15, .19)
	9	.05	.07	(.05, .09)	.07	(.05, .08)

NOTE: No intra-class correlation (NIC) refers to the standard Multinomial-Dirichlet model.