Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A Bayesian test of independence in a two-way contingency table using surrogate sampling

Balgobin Nandram [a,*], Dilli Bhatta [a], Joe Sedransk [b], Dhiman Bhadra [c]

[a] Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, United States
[b] Department of Statistics, Case Western Reserve University, 335 Euclid Avenue, Cleveland, OH 44106, United States
[c] Production and Quantitative Methods Area, Indian Institute of Management Ahmedabad, Gujarat 380015, India

ARTICLE INFO

Article history:
Received 9 May 2012
Received in revised form
22 December 2012
Accepted 6 March 2013

Keywords:
Bayes factor
Chi-squared test
Cluster table
Rao–Scott approximations
Surrogate samples
Sampling-based method
Total table

ABSTRACT

We consider a Bayesian approach to the study of independence in a two-way contingency table which has been obtained from a two-stage cluster sampling design. If a procedure based on single-stage simple random sampling (rather than the appropriate cluster sampling) is used to test for independence, the *p*-value may be too small, resulting in a conclusion that the null hypothesis is false when it is, in fact, true. For many large complex surveys the Rao–Scott corrections to the standard chi-squared (or likelihood ratio) statistic provide appropriate inference. For smaller surveys, though, the Rao–Scott corrections may not be accurate, partly because the chi-squared test is inaccurate. In this paper, we use a hierarchical Bayesian model to convert the observed cluster samples to simple random samples. This provides surrogate samples which can be used to derive the distribution of the Bayes factor. We demonstrate the utility of our procedure using an example and also provide a simulation study which establishes our methodology as a viable alternative to the Rao–Scott approximations for relatively small two-stage cluster samples. We also show the additional insight gained by displaying the *distribution* of the Bayes factor rather than simply relying on a summary of the distribution.

1. Introduction

We consider a Bayesian test of independence for a two-way contingency table which arises with data from a two-stage cluster sampling design. Due to the resulting intracluster correlation, the usual multinomial sampling scheme is no longer appropriate. Specifically, the standard chi-squared or likelihood ratio test can fail. Various adjustments to these tests have been proposed to account for a cluster sampling design. One of these is the Rao–Scott corrections which are now implemented in various statistical packages such as SAS. These corrections are based on normal approximations and moment-matching principles, and they perform quite well for large complex surveys. However, for a two-stage cluster sampling design with not too many clusters and small expected cell counts, the performance of these adjusted tests may be sub-optimal (i.e., they can result in misleading *p*-values). We propose a hierarchical Bayesian model to provide more accurate tests of independence in two-way contingency tables. While we consider a fairly simple sample design, the methodology we propose is general and can be extended to more complex survey designs and contingency tables.

Let $\{n_{jk}, j=1,\ldots,r, k=1,\ldots,c\}$ denote the cell counts in a $r\times c$ contingency table and let $n=\sum_{j=1}^{r}\sum_{k=1}^{c} n_{jk}$ denote the total sample size. The marginal totals for the jth row and kth column are, respectively, $n_{j\cdot}=\sum_{k=1}^{c} n_{jk}, j=1,\ldots,r$, and $n_{\cdot k}=\sum_{j=1}^{r} n_{jk}$,

[*] Corresponding author.
E-mail addresses: balnan@wpi.edu (B. Nandram), drb122@wpi.edu (D. Bhatta), jxs123@case.edu (J. Sedransk), dhiman@iimahd.ernet.in (D. Bhadra).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.03.011

$k=1,\ldots,c$. Let $S=rc$ denote the total number of cells and π_{jk} the cell probability for the (j,k)th cell, where $\sum_{j=1}^{r}\sum_{k=1}^{c}\pi_{jk}=1$, $p_j=\sum_{k=1}^{c}\pi_{jk}$ and $q_k=\sum_{j=1}^{r}\pi_{jk}$. The independence hypothesis states that $\pi_{jk}=p_jq_k$, $j=1,\ldots,r$, $k=1,\ldots,c$, where $\sum_{j=1}^{r}p_j=\sum_{k=1}^{c}q_k=1$. Assuming random sampling, the Pearson chi-squared and the likelihood ratio statistics are, respectively,

$$X^2=n\sum_{jk}\left(n_{jk}-\frac{n_{j\cdot}n_{\cdot k}}{n}\right)^2\Big/n_{j\cdot}n_{\cdot k},\quad G^2=2n\sum_{jk}(n_{jk}/n)\log\left\{\frac{n_{jk}}{n_{j\cdot}n_{\cdot k}/n}\right\},$$

where $n_{j\cdot}$, $j=1,\ldots,r$ and $n_{\cdot k}$, $k=1,\ldots,c$, are positive. It is well known that both X^2 and G^2 have the same asymptotic chi-squared distributions with $(r-1)(c-1)$ degrees of freedom (as $n\to\infty$ with S fixed). For a complex sample design (e.g., two-stage cluster sampling, stratified multistage cluster sampling, etc.), both X^2 and G^2 have 'skewed' distributions, and alternative methods are needed.

When there is a clustering effect, the units in a cluster are, in general, positively correlated leading to a smaller effective sample size and therefore larger variability in the estimates of the cell probabilities. This will evidently result in larger p-values than what would be obtained under simple random sampling (e.g., see Brier, 1980; Bedrick, 1983; Holt et al., 1980; Scott and Holt, 1982).

Rao and Scott (1981, 1984) have studied this problem very carefully and obtained simple corrections to the standard X^2 and G^2 statistics, not only for the test of independence for two-way contingency tables arising from two-stage cluster sampling but essentially for any complex sampling design. Their procedure is asymptotic and nonparametric in nature, and therefore, very general. While we do not attempt such generality in this paper, we present a method to overcome limitations in the Rao–Scott methodology. Moreover, our approach can be extended to more complex scenarios. The Rao–Scott corrections are obtained through *design effects*. A design effect is the ratio of the variance of a statistic under a complex sampling design to that under simple random sampling. For two-stage cluster sampling, these design effects can be much larger than one, thereby having a large impact on the standard chi-squared statistic. As a by-product of our methodology we obtain a Bayesian analogue of these design effects.

Rao and Scott (1981, 1984) show that, under very general complex designs, X^2 and G^2 have the same asymptotic distribution. Let $\theta=(\theta_{11},\ldots,\theta_{r-1,c-1})'$, where $\theta_{jk}=\pi_{jk}-\pi_{j\cdot}\pi_{\cdot k}$, $P_r=(\pi_{1\cdot},\ldots,\pi_{r-1,\cdot})'$ and $P_c=(\pi_{\cdot 1},\ldots,\pi_{\cdot,c-1})'$. Let V denote the covariance matrix of $\hat{\theta}$ under the null hypothesis of independence and let $\hat{V}$ denote an estimator of V; $\hat{V}$ can be complex as it can involve survey weights and other design features. If the entire data set is available, $\hat{V}$ can be obtained using linearization or a resampling method (e.g., bootstrap or jackknife). Let $\hat{\pi}_{jk}=n_{jk}/n_{j\cdot}n_{\cdot k}$ denote the MLE of π_{jk} under multinomial sampling with corresponding notations such as $\hat{\theta}$, $\hat{P}_r$, $\hat{P}_c$ and $\hat{V}$. Let $\delta_g, g=1,\ldots,\kappa=(r-1)(c-1)$ denote the eigenvalues of $(P_r^{-1}\oplus P_c^{-1})V$ where $\oplus$ is the usual direct matrix product; the δ_g are known as *generalized design effects*, a phrase originally coined by Rao and Scott (1981). Assuming that the central limit theorem holds, Rao and Scott (1981, 1984) show that, asymptotically, $X^2=\sum_{g=1}^{\kappa}\delta_gZ_g^2=G^2$, where the Z_g are independent standard normal random variables. Let $\hat{\delta}_g$ be the consistent estimators of δ_g, $g=1,\ldots,\kappa$, and $\hat{\overline{\delta}}$ be the same for $\overline{\delta}=\sum_{g=1}^{\kappa}\delta_g/\kappa$. Then, the effective sample size in the complex survey equivalent to simple random sampling is $\tilde{n}=n/\overline{\delta}$, and the Rao and Scott (1981) adjusted X^2 and G^2 are

$$\overline{X}^2=\tilde{n}\sum_{jk}\left(n_{jk}-\frac{n_{j\cdot}n_{\cdot k}}{n}\right)^2\Big/n_{j\cdot}n_{\cdot k},\quad \tilde{G}^2=2\tilde{n}\sum_{jk}(n_{jk}/n)\log\left\{\frac{n_{jk}}{n_{j\cdot}n_{\cdot k}/n}\right\}.$$

For a two-stage cluster sampling design, $\tilde{n}$ can be much smaller than n depending on the intra-cluster correlation. Rao and Scott (1981) obtained a first order approximation by matching first moments and a second order approximation by matching the first two moments using Satterthwaite's procedure, both ignoring the sampling variation in $\hat{V}$.

A third approximation, an adjustment which uses the degrees of freedom in the variance estimate to account for sampling variation in $\hat{V}$ and other parameters, is more accurate than the first two methods; see Thomas and Rao (1987), Rao and Thomas (1989) and Thomas et al. (1996). However, the first order approximation is typically used in practice (e.g., SAS Proc Surveyfreq Version 9.2) and can be calculated using information on the standard errors of the cell probabilities and marginal proportions which are generally available (e.g., see Bedrick, 1983). The Rao–Scott corrections are very useful and practical for large complex surveys.

However, for smaller complex surveys (i.e., when expected cell counts are less than 5), the asymptotic distributions of both $\overline{X}^2$ and $\overline{G}^2$ can be grossly incorrect and hence their applicability is questionable. The Rao–Scott corrections are not constructed to deal with small expected cell counts. Our objective is to provide a methodology to permit appropriate analyses for two-way tables arising from a two-stage cluster sampling design.

In two-stage cluster sampling, a sample of ℓ clusters (primary sampling units or psu's) is selected and within the ith sampled cluster, a sample of n_i units (secondary sampling units or ssu's) is selected. Let n_{ijk} denote the counts in the (j,k)th cell of the $r\times c$ table constructed from the ith cluster; we call this table the ith *cluster table*. Analogously, let $n_{jk}=\sum_{i=1}^{\ell}n_{ijk}$ be the cell counts for the (j,k)th cell of the table of total counts. We will call the table of total counts the *total table* and in our method we assume that we have information from all the cluster tables. Interest is on a test of independence of two categorical variables in the $r\times c$ total table. However, analysis of only the total table will not account for the clustering effect and the test of independence will be misleading. Although the first order Rao–Scott approximation requires only the total table and the average design effect, the (preferred) second order Rao–Scott approximation essentially requires information from all ℓ tables.

Please cite this article as: Nandram, B., et al., A Bayesian test of independence in a two-way contingency table using surrogate sampling. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.03.011

There are two approaches to obtain a test of independence of two categorical variables when data are available from a complex survey. First, one can use the effective sample size under simple random sampling. This is exactly what the Rao–Scott approximations do. Second, one can adopt a model that is appropriate when there is random sampling by adjusting the parameters to account for clustering. Here, we adjust the usual multinomial model for the cell counts in the ith cluster by multiplying the population cell probabilities, π_s, by quantities, α_{is}, representing the clusters (i) and cells (s). We use a hierarchical model for the α_{is} to accommodate intracluster correlation. We then use this model to make inference for the population cell probabilities (i.e., the π_s). This, in turn, permits us to draw samples which are surrogates for the counts of the observed total table. That is, each of these surrogate samples is a data set that can be regarded as a simple random sample from the superpopulation and is part of the *equivalence class* associated with the *observed* cluster sample. Dong et al. (2011) has a related idea for converting a complex sample to a simple random sample when multiple surveys are combined.

The idea is to simulate a large sample of total tables under simple random sampling, and compute the Bayes factor for a test of independence from each simulated table. A summary of the distribution of the Bayes factor is used to perform the Bayesian test of independence. Surrogate sampling has been used by Nandram (2007) to convert data obtained through a selection bias mechanism to provide equivalent data chosen using simple random sampling. While we present methodology for two stage cluster sampling, a special case, the approach is quite general. We start with a model appropriate for simple random sampling and elaborate it to accommodate the more complex sample design. Next, we make inference for the (population) parameter, θ, of the initial model (i.e., draw M' samples from the posterior distribution of θ). Then we use $\theta^{(1)},\ldots,\theta^{(M')}$ to draw simple random samples consistent with the observed data. The data from these simple random samples are then used to make the required inferences (e.g., to test independence in a contingency table using a Bayes factor).

Finally, we note that the Bayes factor (Kass and Raftery 1995) is sensitive to prior specifications. Nandram and Choi (2007) discussed this issue and reverted to a Bayesian estimation procedure to do the test since it is well known that estimation is relatively less sensitive to moderate changes in the specifications of the hyperprior distribution. However, in our case the Bayes factor will not be sensitive to small changes in the uniform prior because the cell counts of the total table are expected to be much larger than zero (although one or two cells can have zero counts).

We provide a Bayesian test of independence when data are available from the cluster tables. Our main objective is to obtain a Bayesian method that maintains the simplicity of the Rao–Scott method and is more accurate. In Section 2, we describe the Bayesian test of independence for the two categorical variables. In Section 3, we show how to compute the Bayes factor and its distribution from surrogate samples. In Section 4, we present a real example and a simulation study. We also compare our method to the Rao–Scott approximations. Section 5 has concluding remarks.

2. Hierarchical Bayesian model

We study independence of two categorical variables when data are obtained from a clustered superpopulation in which each unit has exactly one of the S characteristics. In Brier's (1980) model, given the cell probabilities indexed by the cluster indicators, the cell counts are assumed to have a multinomial distribution. To accommodate the cluster effects, these cell probabilities are assigned the same Dirichlet distribution with independence over the clusters. Thus, the standard multinomial–Dirichlet model provides the same design effect for the estimator of each cell probability of the two-way table (Brier, 1980). We make an adjustment to the standard multinomial–Dirichlet model to get different design effects for each of the estimators of the cell probabilities of the $r\times c$ table and, therefore, any linear combination of these cell probabilities (e.g., marginal probabilities), if required.

A sample of the clusters is taken and, in turn, a sample of the units within each sampled cluster is taken. We assume that the number ℓ of clusters sampled is small compared to the number of clusters in the population, and the total number of units in each cluster is much larger than the cluster sample size. Our results hold for any ignorable two-stage sample design where the data are concordant with the model in (2)–(5) below; see Sugden and Smith (1984). We assume that there are data for all ℓ cluster tables. The probability that a unit has the sth characteristic within the ith cluster of the superpopulation is assumed to be $\alpha_{is}\pi_s$, $i=1,\ldots,\ell$, $s=1,\ldots,S$. Here, $\pi_s, s=1,\ldots,S$, are the cell probabilities and the α_{is} correct for cluster effects.

We use a hierarchical Bayesian model to obtain the surrogate samples. This model is used to convert the total table obtained from the two-stage cluster sampling design to a surrogate total table which behaves as a table obtained from a simple random sampling design. Our method gives a large number of replicates of the total table from an output analysis. Then, we can calculate the Bayes factor for each total table, thereby providing a distribution of the Bayes factor from which we can use a summary statistic such as the mode.

We string out the counts in the total table to an array of S cells (i.e., $n_s, s=1,\ldots,S$). If we assume simple random sampling, our Bayesian model is

$$\underset{\sim}{n}\mid\underset{\sim}{\pi}\sim\text{Multinomial}(n,\underset{\sim}{\pi}),$$

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

where $\underset{\sim}{n}=(n_1,\ldots,n_S)$, $\underset{\sim}{\pi}=(\pi_1,\ldots,\pi_S)$, $n=\sum_{s=1}^{S}n_s$ and $\underset{\sim}{1}$ is a vector of S ones. We call this model with simple random sampling MSRS. Typically, the total table will have large counts relative to the cluster tables, so that the uniform prior is approximately noninformative (i.e., the posterior mode is the same as the maximum likelihood estimator). It is possible to have a few cells with zero counts, but most of the cell counts are expected to be larger than zero.

We take care of the clustering by assuming that

$$\underset{\sim}{n}_i \mid \underset{\sim}{a}_i \overset{ind}{\sim} \text{Multinomial}(n_i, \underset{\sim}{a}_i), \tag{2}$$

where $\underset{\sim}{n}_i = (n_{i1},\ldots,n_{iS})$, $n_{i.} = \sum_{s=1}^{S} n_{is}$ and $a_{is} = \alpha_{is}\pi_s$, $i=1,\ldots,\ell$, $s=1,\ldots,S$. In (2) we have the constraints $\{\sum_{s=1}^{S}\alpha_{is}\pi_s = 1,\ i=1,\ldots,\ell, \sum_{s=1}^{S}\pi_s = 1, \alpha_{is} > 0, \pi_s > 0\}$. Here, the α_{is} are used to adjust for the clustering. We need a test of independence based on the π_s.

A priori we take,

$$\alpha_{is} \mid \tau_s, \nu \overset{ind}{\sim} \text{Gamma}(\tau_s, \tau_s\nu), \quad s=1,\ldots,S \tag{3}$$

and

$$\underset{\sim}{\pi} \sim \text{Dirichlet}(\underset{\sim}{1}). \tag{4}$$

Note that in $a_{is} = \alpha_{is}\pi_s$, neither the α_{is} nor the π_s are identifiable. This is true because the number of cells in the ith cluster table is S while the number of parameters corresponding to the ith cluster is $2(S-1)$. Thus, we specify the τ_s to allow both α_{is} and π_s to be identifiable.

We note two important features of this model. First, a model for simple random sampling is a special case of ours. This is easily seen by setting $\alpha_{is}\equiv 1$. Second, by construction, the model gives a positive correlation among the units in a cluster and this correlation varies with the cell of the contingency table. To show this, let $I_{isj} = 1$ if a jth ssu falls in the sth cell and $I_{isj} = 0$ otherwise. Then, given α_{is} and π_s, $I_{isj} \overset{iid}{\sim} \text{Bernoulli}(\alpha_{is}\pi_s)$. After some algebraic manipulation, it follows that $\text{var}(I_{isj}) = (\nu S-1)/\nu^2 S^2$, $\nu > S^{-1}$, independent of s, and $\text{cov}(I_{isj}, I_{isj'}) = \{2\tau_s^{-1} + (S-1)/S\}/S(S+1)\nu^2$, $j \neq j'$, positive. Therefore, $\text{cor}(I_{isj}, I_{isj'}) = \{S(2\tau_s^{-1}+1)-1\}/(S+1)(\nu S-1)$, $j \neq j'$, and by the Cauchy–Schwarz inequality the intracluster correlation lies in (0,1) provided that $\nu > S^{-1}$. Because the correlation varies with the cell of the contingency table, we have different design effects for the estimators of the cell probabilities of the total table. Henceforth, we let $\nu_o = S^{-1}$; so that $\nu > \nu_o$.

Finally, for ν, we assume a standard noninformative prior,

$$p(\nu) \propto 1/\nu, \quad \nu > \nu_o. \tag{5}$$

Note that the joint prior density of the α_{is}, π_s and ν must satisfy the constraints, $\{\sum_{s=1}^{S}\alpha_{is}\pi_s = 1,\ i=1,\ldots,\ell, \sum_{s=1}^{S}\pi_s = 1, \alpha_{is} > 0, \pi_s > 0\}$. This is our model for a two-stage cluster sampling design and we will call it MCSD.

It is easy to fit MSRS. In fact, under MSRS,

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

However, the Bayesian model under cluster sampling is much more complex partly because of the constraints and the complexity of the a_{is}.

Letting $t_{is} = \alpha_{is}\pi_s$ and $\underset{\sim(S)}{\pi} = (\pi_1,\ldots,\pi_{S-1})$, the joint posterior density, obtained from Appendix A, is

$$p(\underset{\sim}{t}, \underset{\sim(S)}{\pi} \mid \underset{\sim}{n}) \propto \{1 - F_{\ell b}(A\nu_o)\} A^{-\ell b}$$
$$\times \prod_{i=1}^{\ell}\left[\left(\prod_{s=1}^{S-1} t_{is}^{n_{is}+\tau_s-1}\right)\left(1-\sum_{s=1}^{S-1} t_{is}\right)^{n_{iS}+\tau_S-1}\left\{\left(\prod_{s=1}^{S-1}\pi_s^{\tau_s}\right)\left(1-\sum_{s=1}^{S-1}\pi_s\right)^{\tau_S}\right\}^{-1}\right], \quad (\underset{\sim}{t}, \underset{\sim(S)}{\pi}) \in \tilde{T}^*, \tag{6}$$

where $F_{\ell b}(a) = \int_0^a t^{\ell b-1} e^{-t}/\Gamma(\ell b)\, dt$ is the cdf of a Gamma random variable,

$$\tilde{T}^* = \left\{(\underset{\sim}{t}, \underset{\sim(S)}{\pi}) : 0 < \sum_{s=1}^{S-1} t_{is}, t_{is} > 0, \sum_{s=1}^{S-1}\pi_s < 1, t_{is}, \pi_s > 0,\ i=1,\ldots,\ell, s=1,\ldots,S-1\right\},$$

and

$$A = \sum_{i=1}^{\ell}\left\{\sum_{s=1}^{S-1}\tau_s\frac{t_{is}}{\pi_s} + \tau_S\left(\frac{1-\sum_{s=1}^{S-1} t_{is}}{1-\sum_{s=1}^{S-1}\pi_s}\right)\right\}.$$

In Appendix A we also show that the joint posterior density is proper.

3. Computations, Bayes factor and specifications

Letting $\underset{\sim}{n}$ denote the observed data from the total table and $\underset{\sim}{\hat{n}}$ the vector of surrogate sample counts for the total table, we need to generate samples from

$$f_{SRS}(\underset{\sim}{\hat{n}} \mid \underset{\sim}{n}) = \int f_{SRS}(\underset{\sim}{\hat{n}} \mid \underset{\sim}{\pi}, \underset{\sim}{n}) f_{CL}(\underset{\sim}{\pi} \mid \underset{\sim}{n})\, d\underset{\sim}{\pi}. \tag{7}$$

Please cite this article as: Nandram, B., et al., A Bayesian test of independence in a two-way contingency table using surrogate sampling. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.03.011

In (7) f_{SRS} indicates that $\underset{\sim}{\hat{n}}$ are the surrogate cell counts appropriate to simple random sampling and f_{CL} is the posterior density of $\underset{\sim}{\pi}$ using the model for the observed cluster data (MCSD). In Section 3.1 we show how to generate samples from $f_{CL}(\underset{\sim}{\pi}|\underset{\sim}{n})$ using (6). We then show how to obtain the Bayes factor, and we also show how to specify the parameters τ_s, $s=1,\ldots,S$. We use a computational method which ensures that our method is more accurate and at least as fast as the methods of Rao and Scott (1981).

3.1. Outline of computations

As is apparent, the joint posterior density is complicated, and so we need a sampling based method to draw samples from it. We obtain random draws from an approximation of the joint posterior density and then use the sampling importance resampling (SIR) algorithm (Gelman et al., 2004, Chapter 12) to subsample these draws to obtain samples from $\underset{\sim}{\pi}|\underset{\sim}{n}$; this gives us the required samples of $\underset{\sim}{\pi}$. Note that we are not using Markov chain Monte Carlo methods because we want to avoid monitoring and make our algorithm at least as fast as the Rao–Scott methods.

Having obtained samples from the posterior density of $\underset{\sim}{\pi}$, we can now obtain samples from the distribution of the Bayes factor. Let $\underset{\sim}{\pi}^{(h)}$, $h=1,\ldots,M$, denote the M samples from our MCSD (i.e., cluster model). Then, we draw $\underset{\sim}{\hat{n}}^{(h)}$ from the total table,

$$\underset{\sim}{\hat{n}}^{(h)} \overset{ind}{\sim} \text{Multinomial}\{n,\underset{\sim}{\pi}^{(h)}\}, \quad h=1,\ldots,M.$$

Here, $\underset{\sim}{\hat{n}}^{(h)}$ is surrogate data because the original total table (observed data) has now been converted and a model for simple random sampling is appropriate. Thus, we have M surrogates for the total table. Now, to compute M values of the Bayes factor, we fit a model of association and a model of no association to the surrogate data, $\underset{\sim}{\hat{n}}^{(h)}$, $h=1,\ldots,M$, each surrogate in turn. We take the model of association to be

$$\underset{\sim}{n}^{(h)} \sim \text{Multinomial}(n,\underset{\sim}{\pi}), \quad \underset{\sim}{\pi} \sim \text{Dirichlet}(\underset{\sim}{u}), \quad h=1,\ldots,M, \tag{8}$$

where $u_s=.5$, $s=1,\ldots,S$, for Jeffreys' prior (proper prior). Letting $\pi^*_{jk}=\pi^{(1)}_j\pi^{(2)}_k$, $j=1,\ldots,r$, $k=1,\ldots,c$, the model with no association is

$$\begin{gathered}\underset{\sim}{n}^{(h)}|\underset{\sim}{\pi}^{(1)},\underset{\sim}{\pi}^{(2)} \sim \text{Multinomial}(n,\underset{\sim}{\pi}^*),\\ \underset{\sim}{\pi}^{(1)} \sim \text{Dirichlet}(\underset{\sim}{v}) \text{ and independently } \underset{\sim}{\pi}^{(2)} \sim \text{Dirichlet}(\underset{\sim}{w}),\end{gathered} \tag{9}$$

where $v_j=.5$, $j=1,\ldots,r$ and $w_k=.5$, $k=1,\ldots,c$. It is worth noting that, while the computation of the Bayes factor requires proper prior distributions, proper priors are not required in MCSD as long as the posterior density (6) is proper (as we have shown in Appendix A). However, we do need a proper prior in (8). Inference should not be sensitive to moderate departures from the Jeffreys' prior because the cell counts of the total table are expected to be large.

In Appendix D we present the Bayes factor for a test of independence for the total table. Matching notation with Appendix D, the Bayes factor is given by

$$BF^{(h)}=p_{\text{as}}(\underset{\sim}{n}^{(h)})/p_{\text{nas}}(\underset{\sim}{n}^{(h)}), \quad h=1,\ldots,M,$$

where $p_{\text{as}}(\underset{\sim}{n}^{(h)})$ and $p_{\text{nas}}(\underset{\sim}{n}^{(h)})$ are, respectively, the marginal likelihoods under the models with association (as) and without association (nas). (Note that larger values of $BF^{(h)}$ give stronger evidence for association relative to no association (independence).) In Appendix C, we show how to obtain the mode of the posterior distribution of the Bayes factor. It is straightforward to obtain other summaries of the Bayes factor.

Thus, our method obtains M estimates of the Bayes factor and these estimates, in turn, provide an estimate of the empirical distribution of the true Bayes factor. Our computations show that the entire procedure to obtain the M estimates of the Bayes factor and its distribution takes less than 5 s on our 850 MHz computer for data from small two-stage cluster sampling designs. Henceforth, we will mostly work with the log-Bayes factor. We use log-Bayes (base e) factors because the marginal likelihoods can be large; see Kass and Raftery (1995) for a discussion of the log-Bayes factor.

We use two rules of thumb for our comparisons. The standard rule of thumb for p-values is as follows: (.05–.10), borderline; (.025–.05), reasonably strong; (.01–.025) strong, (0–.01) very strong. Looking for evidence of association, the rule of thumb of the log-Bayes factor is as follows: (0–1), not worth more than a bare mention (same as borderline); (1–3), positive (same as reasonably strong); (3–5) strong, 5+, very strong; see Kass and Raftery (1995).

Details about the computations are in Sections 3.2 and 3.3 while the numerical analysis is in Section 4.

3.2. Computational details

First, we need an approximation for (6). Using a heuristic argument we conjecture that an approximation which satisfies four properties may be useful. First, the approximation should have some dependence between the t_{is} and the π_s; see (6). Second, t_{is} and π_s should have similar forms. Third, the distributions of t_{is} and π_s should be functions of the data (i.e., the cell counts of the cluster tables) to allow the data to have a direct influence on these distributions. Fourth, the computations of the approximation must be fast and should not require any monitoring. To approximate the joint density of $\underset{\sim}{t}$ and $\underset{\sim}{\pi}$, we take

$\underset{\sim i}{t}=(t_{i1},\ldots,t_{iS})$, $i=1,\ldots,\ell$, given $\underset{\sim}{\pi}$ and $\underset{\sim}{n}$ to be independent, giving

$$p_a(\underset{\sim}{t},\underset{\sim}{\pi}|\underset{\sim}{n})=\left\{\prod_{i=1}^{\ell}p_a(\underset{\sim i}{t}|\underset{\sim}{\pi},\underset{\sim}{n})\right\}p_a(\underset{\sim}{\pi}|\underset{\sim}{n}), \tag{10}$$

where $p_a(\underset{\sim i}{t}|\underset{\sim}{\pi},\underset{\sim}{n})$ and $p_a(\underset{\sim}{\pi}|\underset{\sim}{n})$ are determined next.

First, to obtain the approximation, $p_a(\underset{\sim}{\pi}|\underset{\sim}{n})$, we consider the posterior density under simple random sampling. Here,

$$p^*(\underset{\sim}{\pi}|\underset{\sim}{n})\propto\prod_{s=1}^{S}\pi_s^{n_{\cdot s}},\qquad\sum_{s=1}^{S}\pi_s=1.$$

Our intuition is that the correct posterior density under cluster sampling should be related to this posterior density under simple random sampling. However, it should reflect the clustering through the design effects. Thus, we make two additional adjustments to $p^*(\pi|\underset{\sim}{n})$. First, by penalizing $n_{\cdot s}$, $s=1,\ldots,S$, we replace $n_{\cdot s}$ by $n_{\cdot s}/\delta_s$ where δ_s are design effects, possibly all the same as in Brier's method. A method for choosing the δ_s is given in Section 3.3. Second, to make this dependent on τ_s (suggested by the term in π_s in (A.5)), we add τ_s to $n_{\cdot s}/\delta_s$ to get the approximate posterior density, $p_a(\underset{\sim}{\pi}|\underset{\sim}{n})$

$$\underset{\sim}{\pi}|\underset{\sim}{n}\sim\text{Dirichlet}(\underset{\sim}{d}), \tag{11}$$

where $d_s=n_{\cdot s}/\delta_s+\tau_s+1$, $s=1,\ldots,S$.

Second, note that ignoring the term $(1-F(A\nu_o))A^{-\ell b}$ and the constraints, the conditional posterior density in (6) is of the form

$$p^{**}(\underset{\sim}{t}|\underset{\sim}{\pi},\underset{\sim}{n})\propto\prod_{i=1}^{\ell}\prod_{s=1}^{S}t_{is}^{n_{is}+\tau_s-1},\ t_{is}>0,\ s=1,\ldots,S,\qquad\sum_{s=1}^{S}t_{is}=1,\ i=1,\ldots,\ell.$$

That is, approximately, $\underset{\sim i}{t}|\underset{\sim}{n}\overset{ind}{\sim}\text{Dirichlet}(\underset{\sim i}{n}+\underset{\sim}{\tau})$. We allow this to be dependent on $\underset{\sim}{\pi}$ by replacing n_{is} with $n_{i\cdot}\pi_s$. Thus, approximately, $\underset{\sim i}{t}|\underset{\sim}{\pi},\underset{\sim}{n}\overset{ind}{\sim}\text{Dirichlet}(n_{i\cdot}\underset{\sim}{\pi}+\underset{\sim}{\tau})$. Adding unity to the Dirichlet parameters to increase computational stability, the final approximation, $p_a(\underset{\sim i}{t}|\underset{\sim}{\pi},\underset{\sim}{n})$, of the conditional posterior distribution of $\underset{\sim i}{t}|\underset{\sim}{\pi},\underset{\sim}{n}$ is

$$\underset{\sim i}{t}|\underset{\sim}{\pi},\underset{\sim}{n}\overset{ind}{\sim}\text{Dirichlet}(\underset{\sim i}{b}), \tag{12}$$

where $b_{is}=n_{i\cdot}\pi_s+\tau_s+1$, $i=1,\ldots,\ell$, $s=1,\ldots,S$. Observe that (11) and (12) have similar forms.

We now show how to carry out the SIR algorithm. To obtain the probability of selecting each sampled iterate, we need to study the ratio

$$R(\underset{\sim}{t},\underset{\sim}{\pi}|\underset{\sim}{n})=\frac{p(\underset{\sim}{t},\underset{\sim}{\pi}|\underset{\sim}{n})}{p_a(\underset{\sim}{t},\underset{\sim}{\pi}|\underset{\sim}{n})},$$

where $p(\underset{\sim}{t},\underset{\sim}{\pi}|\underset{\sim}{n})$ and $p_a(\underset{\sim}{t},\underset{\sim}{\pi}|\underset{\sim}{n})$ are given, respectively, in (A.5) and (10). Simplifying, we get

$$R(\underset{\sim}{t},\underset{\sim}{\pi})=C\frac{\{1-F_{\ell b}(A\nu_o)\}\prod_{i=1}^{\ell}[\{\prod_{s=1}^{S}t_{is}^{n_{is}-n_{i\cdot}\pi_s-1}\}D(n_{i\cdot}\underset{\sim}{\pi}+\underset{\sim}{\tau}+\underset{\sim}{1})]}{\{\prod_{s=1}^{S}\pi_s^{n_s/\delta_s+(\ell+1)\tau_s}\}A^{b\ell}}, \tag{13}$$

where strictly $0<\pi_s<1$, $0<t_{is}<1$, $D(\cdot)$ is the Dirichlet function and C is a proportionality constant. Note that, by construction, $R(\underset{\sim}{t},\underset{\sim}{\pi})$ is bounded because both $p(\underset{\sim}{t},\underset{\sim}{\pi}|\underset{\sim}{n})$ and $p_a(\underset{\sim}{t},\underset{\sim}{\pi}|\underset{\sim}{n})$ are bounded.

We use 10% subsampling. We draw $\tilde{M}=10{,}000$ samples from the approximate joint posterior density in (10). This is obtained using the composition rule by first drawing $\underset{\sim}{\pi}$ from (11) and, in turn, drawing $\underset{\sim i}{t}$ from (12). Letting $\Omega^{(h)}=(\underset{\sim}{t}^{(h)},\underset{\sim}{\pi}^{(h)}),h=1,\ldots,\tilde{M}$, the subsampling probabilities are $W_h=R(\Omega^{(h)})/\sum_{h'=1}^{\tilde{M}}R(\Omega^{(h')}),h=1,\ldots,\tilde{M}$, where $R(\cdot)$ is given in (13). Then we sample 10% of the $\tilde{M}$ samples without replacement to get $M=.10\tilde{M}$ samples. Thus, we finally have samples from the posterior density of $(\underset{\sim}{t},\underset{\sim}{\pi})$ in (6). We have checked that the largest W_h is not too close to unity, anecessity for good performance of the SIR algorithm, in all our examples.

3.3. *Specifications*

We now show how to specify the design effects δ_s and the τ_s. Note that while the τ_s are part of MSCD, the δ_s only affect the computations.

We state and prove an important lemma about the maximum likelihood estimator (MLE) of the parameters of a Gamma distribution in Appendix B. We will use this lemma repeatedly to specify the hyperparameters and the tuning constants.

Let n'_{is}, $i=1,\ldots,\ell$, $s=1,\ldots,S$ denote past data or data from a similar survey. We obtain estimates of α_{is} from the cluster tables with cell counts n'_{is}, $i=1,\ldots,\ell$, $s=1,\ldots,S$, adding 0.5 because of some zero cell counts. First, define $\hat{p}_{is}=(n'_{is}+.5)/(n'_{i\cdot}+.5S)$, $\hat{\pi}_s=(n'_{\cdot s}+.5)/(n'+.5S)$ and $\hat{\alpha}_{is}=\hat{p}_{is}/\hat{\pi}_s$, $i=1,\ldots,\ell$, $s=1,\ldots,S$. We use this form for the $\hat{\alpha}_{is}$ because under (2) only, $E(n'_{is}/n'_{i\cdot})=\alpha_{is}\pi_s$, $i=1,\ldots,\ell$, $s=1,\ldots,S$. Therefore, removing the expectation on the left-hand side, we get $\hat{p}_{is}\approx\hat{\alpha}_{is}\hat{\pi}_s$. Then, we take

$$\hat{\alpha}_{is}\overset{iid}{\sim}\text{Gamma}(\tau_s,\tau_s\nu)$$

Please cite this article as: Nandram, B., et al., A Bayesian test of independence in a two-way contingency table using surrogate sampling. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.03.011

as in (3). Second, assuming momentarily that the τ_s are equal and letting A denote the arithmetic mean of the $\hat{\alpha}_{is}$, the MLE of ν is $\hat{\nu}=A^{-1}$ as in Appendix B. Then, for τ_s a 'profile' log-likelihood is obtained by replacing ν in the log-likelihood function by A^{-1}. For each τ_s with ν fixed at A^{-1}, we obtain the MLE of τ_s by maximizing the profile log-likelihood function,

$$\tau_s \ln(\tau_s)-\tau_s \ln(A)+(\tau_s-1)\ln(G_s)-\tau_s A_s/A-\ln\Gamma(\tau_s), \quad s=1,\ldots,S,$$

where A_s and G_s are the arithmetic and geometric means of $\hat{\alpha}_{is}$. By an argument similar to Appendix B, the MLE exists and is unique. We use the Nelder–Mead algorithm to do the maximization.

We now show how to obtain the design effects for the computation. We consider the following simpler model for cluster sampling:

$$\underset{\sim}{n}_i|\underset{\sim}{\pi}_i \overset{ind}{\sim} \text{Multinomial}(n_{i\cdot},\underset{\sim}{\pi}_i) \quad \text{and} \quad \underset{\sim}{\pi}_i \overset{iid}{\sim} \text{Dirichlet}(\underset{\sim}{\mu}\phi),$$

where $n_{i\cdot}$ is the number of ssu's in the ith cluster, $\underset{\sim}{\pi}_i=(\pi_{i1},\ldots,\pi_{iS})$, $n_{i\cdot}=\sum_{s=1}^{S} n_{is}$ and μ and ϕ are to be specified. Note that simple random sampling occurs in the limit as ϕ goes to infinity. The covariance matrix of $\underset{\sim}{n}$ under cluster sampling is a constant times the covariance matrix under simple random sampling; see Brier (1980). This constant is the design effect and, letting $n=\sum_{i=1}^{\ell} n_{i\cdot}$, it is $(1/n)\sum_{i=1}^{\ell} n_{i\cdot}\left((n_{i\cdot}+\phi)/(1+\phi)\right)$, a weighted average of $(n_{i\cdot}+\phi)/(1+\phi)$, $i=1,\ldots,\ell$.

To specify ϕ, we start by using a method of moments estimator for μ (i.e., $\hat{\mu}_s=\sum_{i=1}^{\ell} n_{is}/n$, $s=1,\ldots,S$). These are reasonably efficient estimators because they are formed from the total table. We obtain ϕ by maximizing the profile log-likelihood of the multinomial–Dirichlet model,

$$\sum_{i=1}^{\ell}\left[\sum_{s=1}^{S}\{\ln\Gamma(n_{is}+\hat{\mu}_s\phi)-\ln\Gamma(\hat{\mu}_s\phi)\}-\{\ln\Gamma(n_i+\phi)-\ln\Gamma(\phi)\}\right]$$

over $\phi>0$. We denote the MLE of ϕ by $\hat{\phi}$ and it is easily obtained using the Nelder–Mead algorithm. Thus we take $\delta_s=(1/n)\sum_{i=1}^{\ell} n_{i\cdot}((n_i+\hat{\phi})/(1+\hat{\phi}))$, $s=1,\ldots,S$, all equal.

4. Numerical analysis

We discuss an illustrative example. This example suggests certain features which are investigated further in a simulation study.

We use the mode of the distribution of the Bayes factors, obtained from the surrogate total tables, for testing and the interquartile range of these Bayes factors for gauging this evidence. We interpret the mode using the rule of thumb of Kass and Raftery (1995), as discussed earlier. However, we share the philosophy that evidence cannot be measured by a single test and other tests (e.g., Rao–Scott test) should also be used. It is not sensible to look at a single p-value or just the mode of the distribution of the Bayes factor.

4.1. Illustrative example

To illustrate our methodology, we use data from the Third International Mathematics and Science Study (TIMMS). The data consist of 2477 students (see Valliant et al., 2000, Appendix B.6). Here, the clusters are schools while the units are the students. There are four strata, the Northeast, South, Central and West regions of the US. We consider three of the variables in the survey, mathematics test scores (below average, average and above average), science test scores (below average, average, above average) and the communities the students come from (village or rural area, outskirts of a town or city and close to the center of a town or city). Within each stratum, we study the association between mathematics test scores (MTS) and communities (COM) and science test scores (STS) and communities (COM), so there are eight examples. We assume that the finite population is a sample from a superpopulation.

In Table 1 we present the total tables for the eight examples (E1–E4 for MTS versus COM and E5–E8 for STS versus COM in each of the four regions). The number (ℓ) of clusters changes considerably over regions as does the number of

Table 1
Features of the total table for each of the eight examples.

Example	n	ℓ	ρ	deff	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
E1	469	37	.56	11.8	44	57	5	83	71	5	63	136	5
E2	663	24	.33	6.85	49	74	1	107	151	13	93	164	11
E3	438	23	.34	7.39	44	47	8	54	44	3	56	167	15
E4	857	51	.33	6.62	25	17	0	157	134	13	205	294	12
E5	469	24	.54	11.46	63	38	5	105	47	7	70	124	10
E6	663	37	.31	6.45	61	56	7	117	141	13	117	145	6
E7	438	23	.35	7.67	53	44	2	67	30	4	95	133	10
E8	857	51	.33	6.63	34	7	1	181	112	11	226	272	13

Note: These are all 3 × 3 contingency tables; n is the number of observations; ℓ is the number of schools; ρ is the intracluster correlation and deff stands for design effect. E4 has a zero cell and E2, E3, E7, E8 have some cell counts near zero.

observations. The intra-class correlations are moderately large and they change considerably over examples. The design effects (deffs), obtained from Brier's model, are considerably larger than one. Thus, in all the examples, the cluster effect is substantial. Some of the observed counts in the total tables do not exceed 5. This is noticeable in cell (1,3) (below average in a town or city) in all examples except E3 and E6. In E4, cell (1,3) is 0 and so standard X^2 and G^2 tests are not really applicable. In fact, for E4 the Rao–Scott first order test cannot be computed using SAS because it uses linearization or the jackknife to estimate the covariance matrix. We are able to compute the Rao–Scott test because we use the bootstrap method. Rao–Scott methods do not provide a sensible adjustment because in our case they correct X^2 and G^2 only for clustering, not for tables with small cell counts.

The 'posterior' deffs for the individual cells are presented in Table 2. These are different from those in Brier's method (see Section 3.3) and are computed using the diagonals of the posterior variance of $\underset{\sim}{\pi}$ under the hierarchical Bayesian model specified by (2)–(5) and the posterior variance under the model for simple random sampling specified by (1). These deffs are considerably larger than 1. The average deffs, 9.01, 6.75, 5.91, 7.80, 8.34, 5.67, 6.71, 7.10 for E1–E8, are very similar to the design effects obtained from Brier's method given in Table 1. But what is more important is that the deffs vary quite a bit over the cells for all examples except E3 and E6. Thus, Brier's method is inappropriate except, perhaps, for E3 and E6. With such large variations in deffs across the cells the Rao–Scott approximations are not expected to work so well. This is particularly true in E4 in which cell (1, 3) has a design effect of 25.65 corresponding to the zero count. For completeness we have also calculated the effective sample size (ESS), the sum of the ratios of the original cell counts of the total table divided by the corresponding design effects. As can be seen in the last row of Table 2, these are considerably smaller than the original sample size (see Table 1).

In Table 3, we present summaries of the Bayes factor obtained from our model. Our decision rule is the one presented in Section 3.1 and applied to the mode of the distribution of the Bayes factor. For example, in example E1 the mode is 5.7 (the Bayes factor is $e^{5.7} \approx 299$) and according to Kass and Raftery (1995) this is a 'very strong' evidence against independence. For comparison, we also present the p-values obtained from the standard chi-squared test and Rao–Scott first order (RSF) and second order (RSS) approximations.

In example E1, RSF and RSS do not reject independence, while the chi-squared test and Bayes factor test show evidence against independence. The very strong evidence against independence shown by the chi-squared test may be due to ignoring the large cluster effect ($\rho = .56$, see Table 1); the effective sample size (67 in Table 2) indicates a degree of sparseness. Except for E2 and E6, the Bayes factors show that there is evidence for a strong dependence between

Table 2
Bayesian design effects for each cell by example.

Example	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)	ESS
E1	7.42	6.99	12.10	7.94	6.88	14.04	6.23	6.54	12.97	67
E2	5.17	5.09	17.05	5.31	5.18	5.85	5.20	5.35	6.56	126
E3	4.75	4.92	7.35	4.81	5.47	9.99	5.23	4.85	5.80	87
E4	5.32	5.69	25.65	5.22	5.04	6.14	5.37	5.14	6.60	164
E5	6.88	6.66	12.69	7.65	6.37	11.40	6.45	6.59	10.37	68
E6	4.93	4.59	7.51	5.07	5.06	6.13	5.02	5.30	7.40	130
E7	4.72	5.25	12.91	4.83	5.35	9.43	5.92	5.27	6.70	82
E8	5.12	7.45	17.59	5.31	5.13	6.59	5.39	5.22	6.09	161

Note: The cells are (j,k), $j,k = 1,2,3$. ESS stands for the effective sample size and it is the sum of the cell counts divided by the design effects, taken for the total table.

Table 3
Comparison of the log-Bayes factors with the p-values by example.

Example	p-Values			log-Bayes factor							
	χ^2	RSF	RSS	Min	Q_1	Q_2	Q_3	Max	Mode	P	N
E1	.001	.17	.14	−7.1	3.6	12.5	23.3	105	5.7	.81	9
E2	.247	.58	.66	−7.9	−2.8	1.1	7.2	60	−1.6	.89	9
E3	.000	.04	.02	−7.4	7.9	17.0	28.0	94	10.8	.69	7
E4	.001	.04	.02	−8.6	1.0	8.0	16.9	84	4.5	.76	9
E5	.000	.02	.01	−6.4	9.3	20.1	33.7	109	14.6	.66	7
E6	.240	.60	.69	−7.9	−1.3	3.4	10.2	55	−0.7	.95	9
E7	.000	.03	.01	−7.2	2.5	9.6	18.5	77	6.05	.71	9
E8	.000	.01	.00	−8.1	7.2	16.2	26.6	108	10.6	.66	7

Note: RSF and RSS denote, respectively, the first and second order Rao–Scott approximations; a bootstrap method is used to estimate the covariance matrix in the Rao–Scott approximations.

Please cite this article as: Nandram, B., et al., A Bayesian test of independence in a two-way contingency table using surrogate sampling. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.03.011

mathematics test scores and community and science test scores and community. It is interesting that the tests based on chi-squared, RSF, RSS and Bayes factor agree in all examples except E1.

We also obtained the proportion, P, of estimated Bayes factors in the 1000 runs that are larger than the observed Bayes factor under the (incorrect) simple random sampling in the observed total table. If the cluster sampling design was a simple random sampling design, it seems reasonable that these Bayes factors should have a distribution symmetric around the observed Bayes factor obtained from simple random sampling. Thus, under simple random sampling these P's should be around .5. These are shown in the penultimate column of Table 3. However, these P's are significantly larger than .5, showing that the clustering effect our model accounts for is substantial.

Because Bayesian estimation procedures are much less sensitive to prior specifications than Bayesian hypothesis tests, we have considered an estimation procedure as well. Based on the hierarchical Bayesian model we have obtained 95% credible intervals for the ratios, $\pi_{jk}/p_j q_k$, $j=1,\ldots,r$, $k=1,\ldots,c$, where $p_j=\sum_{k=1}^{c}\pi_{jk}$ and $q_k=\sum_{j=1}^{r}\pi_{jk}$. Note that there are $S=rc=9$ credible intervals. Then, we have computed the number, N, of 95% credible intervals of $\pi_{jk}/p_j q_k$ containing 1 (e.g., see Nandram, 2007, for a similar procedure). If some of these intervals do not contain 1, this provides some evidence against independence. The values of N, presented in the last column of Table 3, show some evidence of dependence in examples E3, E5 and E8. Of course, these intervals are much too wide for this latter procedure to be particularly useful. Nevertheless it is sensible to consider it as well.

In Fig. 1, we present the distributions of the Bayes factor obtained from the 1000 estimates of the Bayes factor for each of the eight examples. Looking at where most of the distribution lies, it shows that in E2 and E6 there is little evidence against independence and in the other examples there is much stronger evidence against independence. Note that calculating the distribution provides substantially more information than reporting a single summary but it is not done in practice.

In Table 4, we study the issue of sensitivity of the Bayes factor to the specification of $\tau_s, s=1,\ldots,S(S=9)$. We set $\tau_s=\eta\hat{\tau}_s$ where we take $\eta=.5,1,2$ and $\hat{\tau}_s$ are the maximum likelihood estimates. The mode, median, the first and third quartiles and P all decrease as η changes from 0.5 to 2. However, the evidence against independence does not change markedly. This is true in all eight examples. We have also looked at sensitivity to the specification of the uniform prior for the model based on simple random sampling applied to the surrogate total tables. Small variations in the Jeffreys' prior show very small changes in the Bayes factor (e.g., changing .5 in the Jeffreys' prior to .10 or 1).

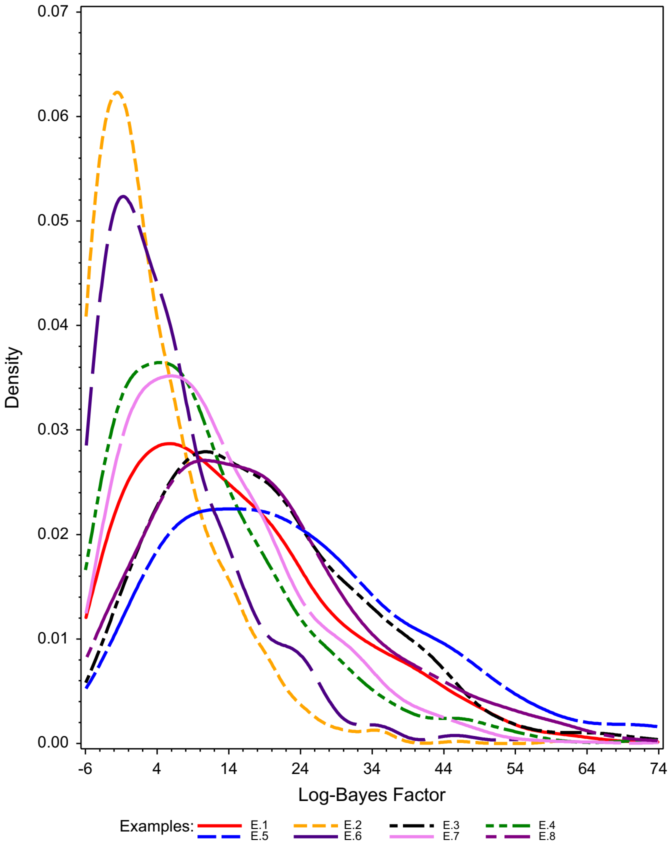

Fig. 1. Plots of the empirical densities of the log-Bayes factors for the eight strata in the third grade example.

Table 4
Sensitivity analysis of the log-Bayes factor with respect to τ_s, $s=1,\ldots,9$, by region (reg) and example.

reg	η	MTS vs. COM			STS vs. COM		
		.5	1	2	.5	1	2
1	Mode	9.3	6.3	4.9	24.5	12.6	9.4
	Median	13.8	12.4	10.3	25.6	21.1	15.6
	IQR	(4.8,25.4)	(4.1,23.5)	(2.9,19.0)	(11.9,37.8)	(9.9, 33.7)	(6.9, 26.9)
	P	.84	.82	.79	.73	.66	.56
2	Mode	−1.5	−1.7	−2.6	−0.9	−0.5	−0.5
	Median	2.1	1.1	1.0	3.6	3.5	2.8
	IQR	(−2.3,7.7)	(−2.5,7.3)	(−3.1,6.9)	(−1.5,10.8)	(−1.5,10.7)	(−1.7,9.1)
	P	.90	.90	.88	.95	.94	.94
3	Mode	14.0	11.1	10.1	5.8	4.9	2.5
	Median	17.5	16.5	13.9	10.5	9.6	6.8
	IQR	(8.3,28.6)	(7.5,27.0)	(6.1,23.1)	(3.6,21.2)	(3.0,18.7)	(1.0,14.8)
	P	.70	.68	.62	.74	.72	.62
4	Mode	3.7	4.2	1.8	12.6	12.1	9.2
	Median	8.8	8.2	7.0	17.0	15.7	14.0
	IQR	(1.8,19.0)	(1.6,16.9)	(0.4,15.9)	(8.3,28.4)	(6.6,28.0)	(5.1,24.1)
	P	.78	.78	.73	.70	.66	.62

Note: Each region has two examples (i.e., MTS vs. COM and STS vs. COM). We have used $\tau_s=\eta\hat{\tau}_s$, $s=1,\ldots,9$, where the $\hat{\tau}_s$ are maximum likelihood estimates and $\eta=.5,1,2$.

4.2. Simulation study

We have performed a small simulation study to help understand these tests further. We consider three factors, dependence between the two categorical variables (weak, strong), the table density of the cluster tables (low, medium) and intracluster correlation (very small, small, moderate). The density of a total table is the total number of observations divided by the product of the number of clusters and the number of cells ($S=9$ for a 3×3 table). We have set the number of clusters, ℓ at 35, and the table density, Δ at 2 and 4 giving total numbers of observations 630 and 1260 respectively.

Corresponding to the nine cells (3×3 contingency table), let $\psi_s=\text{ind}$, $s=1,5,9$ (i.e., the diagonal cells (1, 1), (2, 2), (3, 3)) and $\psi_s=1$, otherwise (off-diagonal cells) where ‘ind’ is to be specified. The cell probabilities are $\psi_s/\sum_{s=1}^{S}\psi_s$, $s=1,\ldots,S$. When the ψ_s are roughly the same (ind=1), there will be independence and when the diagonal ψ_s are larger than 1, there will be dependence (ind=2). For a 3×3 table with large cell counts, if the diagonal probabilities are twice the off diagonals, there will be strong dependence (ind=2). With intracluster correlation ρ, we set $\alpha_s=\{(1-\rho)/\rho\}\psi_s/\sum_{s=1}^{S}\psi_s$, $s=1,\ldots,S$. For $i=1,\ldots,\ell$ we generate $\tilde{\pi}_i \overset{iid}{\sim} \text{Dirichlet}(\tilde{\alpha})$ to get the cell probabilities for the 35 cluster tables. We divide the total number of observations into the clusters with sizes, n_i, $i=1,\ldots,\ell$, based on a multinomial distribution with equal cell probabilities. Finally, the cluster tables are generated independently from multinomial distributions with total counts n_i and cell probabilities $\tilde{\pi}_i$. We choose $\rho=.01$,. 10 ,. 30.

Thus, there are twelve ($2\times2\times3$) design points, and 100 cluster samples are generated at each design point. We perform our computations exactly as for the Third Grade population and obtain both the p-values and the Bayes factors from our model. We ‘average’ various quantities over the 100 replications at each design point. For example, in Table 5 the mode is the average of the 100 modes.

In Table 5 we present numerical summaries from the simulation study. For these choices of Ind, Δ and ρ the conclusions from using RSS or BF are similar. With Ind=1 and $\rho=.01$ or .10 we do not expect any differences between the RSS test and the log-Bayes test because these intra-class correlations are not large enough to offset the independence assumption. Besides the sample sizes at each design point are relatively large. Perhaps some differences are expected at $\rho=.30$ but should also be small because the sample sizes are relatively large. If ind=2, we do not expect any difference between the RSS test and the log-Bayes test because this is a relatively large dependence. Looking at the modes under dependence and using the Kass and Raftery (1995) criteria (see Section 3.1) there is ‘very strong’ evidence for dependence which is essentially the same inference under the RSS test or even the incorrect chi-squared test. As expected, the p-value of the RSS test is at least as large as the p-value of the chi-squared test. But note that at ind=2, $\Delta=2$ and $\rho=.01$, the p-value of the RSS test is smaller than that for the chi-squared test but the difference is small (.005 vs. .001).

In addition, we observe a few interesting things. Under independence as ρ increases, both p-values decrease, but under dependence these p-values increase (note the minor aberration at (Ind=2, $\Delta=2$)). However, there is a clear advantage in using the mode because it is the most plausible value, there is a measure of uncertainty (e.g., the interquartile range) and symmetry between the ‘association’ and ‘no association’ cases. Unlike the behavior of the p-values, the evidence against independence increases as ρ increases (for fixed Ind and δ) for both cases. For the six design points under independence, the interquartile ranges are much narrower than their counterparts under dependence. While there are changes in the

Please cite this article as: Nandram, B., et al., A Bayesian test of independence in a two-way contingency table using surrogate sampling. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.03.011

Table 5
Simulation: comparison of *p*-values and log-Bayes factors.

Ind	Δ	ρ	Deff	*p*-Values		log-Bayes factor			
				χ^2	RSS	Mode	Q_1	Q_2	Q_3
1	2	.01	1.00	.964	.993	−5.86	−6.14	−4.50	−2.31
1	2	.10	1.50	.760	.899	−4.66	−5.35	−3.36	−0.51
1	2	.30	3.74	.350	.816	−1.48	−2.78	0.85	6.06
1	4	.01	1.00	.991	.999	−7.50	−7.61	−6.08	−3.95
1	4	.10	2.32	.700	.883	−5.62	−5.97	−3.25	0.38
1	4	.30	6.71	.280	.884	−0.73	−1.76	3.46	11.39
2	2	.01	1.15	.005	.001	6.16	2.47	8.44	15.58
2	2	.10	2.70	.006	.010	7.60	4.73	12.09	23.47
2	2	.30	6.12	.001	.044	11.19	8.23	19.87	35.79
2	4	.01	1.29	.000	.000	15.40	11.83	20.33	31.35
2	4	.10	4.38	.000	.002	25.39	15.33	30.57	47.62
2	4	.30	11.26	.001	.033	29.50	19.14	39.87	66.92

Note: Ind is the degree of dependence, Δ is table density, ρ is the intracluster correlation, deff is the design effect from Brier's method, RSS is the second order Rao–Scott correction.

magnitudes of the *p*-values and the log-Bayes factors, the changes in inference over the design points are small whether the modes or the *p*-values are used.

In Fig. 2, we show the distributions of the estimated Bayes factors for the twelve design points. The distributions are essentially unimodal; the locations of the mode tell us about the strength of the evidence against independence. We expect the evidence to be weak under independence but the intra-cluster correlation blurs this vision. As the intra-cluster correlation increases, we expect more spread, of course, and what we see is that the distributions move over to the right. There is not much change in the distributions with the table density. Also, as we go from independence to dependence, the distributions of the Bayes factor tend to be flatter with more spread.

Finally, we have performed an additional simulation study for a small sample size and a moderately large intracluster correlation. Specifically, we have taken $n=50$ and $\rho=.50$ with $\ell=25$ and different values of ind. For the total table and ind$=3$ most of the counts will be on the diagonal of the 3×3 categorical table with the off-diagonal elements tending to be less than 5 and sometimes zero. For ind$=1$ in the total table some cells will have counts less than 5 because of the strong cluster effect within the cluster tables. With $\rho=.50$ there is an overall deff of 2 so that the effective sample size is just 25 with some cluster tables having an effective sample size of less than 2 observations.

One obvious problem that RSS faces is respect to the intracluster correlation. It is true that for cluster sampling the *p*-value of the RSS test must be larger than that of the chi-squared test but this is false for two of the four examples (summarized in Table 6), providing further evidence that RSS is inappropriate here. In all examples, the average deffs under Brier's model are larger than 1 (just around 2). Using the criteria given at the end of Section 3.1, the inferences using RSS or BF are different for ind$=1$ and ind$=1.25$. Here, the BF indicates no evidence for association (or borderline evidence for independence) while RSS concludes that there is reasonably strong evidence for association (ind$=1.25$) and borderline evidence for association (ind$=1.00$). The conclusions are similar for ind$=2.75$ and ind$=3.00$, somewhat stronger for RSS (i.e., positive for BF and very strong for RSS).

5. Concluding remarks

We have proposed a method to test for independence in a $r\times c$ contingency table which is obtained from a two-stage cluster sampling design. We have used a hierarchical Bayesian model and a sampling-based method to fit it. By making close approximations to several densities we avoid using Markov chain Monte Carlo methods for inference. Specifically, we use random samples from the approximate posterior density and subsample them using the SIR algorithm. Although ours is a sampling based method it is at least as fast as the Rao–Scott methods. We use the Bayes factor to make inference about independence. Relative to standard methods our approach provides additional insight by displaying the distribution of the Bayes factor rather than simply relying on a single summary measure.

Our most important contribution may be the provision of surrogate random samples. This permits the analyst to use standard software to carry out any analysis.

The Rao–Scott methods were developed to correct for design effects such as cluster effects, i.e., by correcting the standard X^2 and G^2 statistics. They are 'large sample' methods and work well when there are large cell counts. However, they are less successful when there are small cell counts. An extreme case is a table with zero counts, in which case the X^2 and G^2 tests are not applicable. Consequently, the Rao–Scott methods do not apply either (since they are adjustments of the X^2 and G^2 tests for design effects, not sparse tables). Our procedure can get around this problem when there are a few cells having zero counts. However, by doing a sophisticated analysis, we have validated RSS for two-stage cluster sampling with many examples, but there are examples where the use of RSS is inappropriate. Moreover, using the Bayes factor for inference

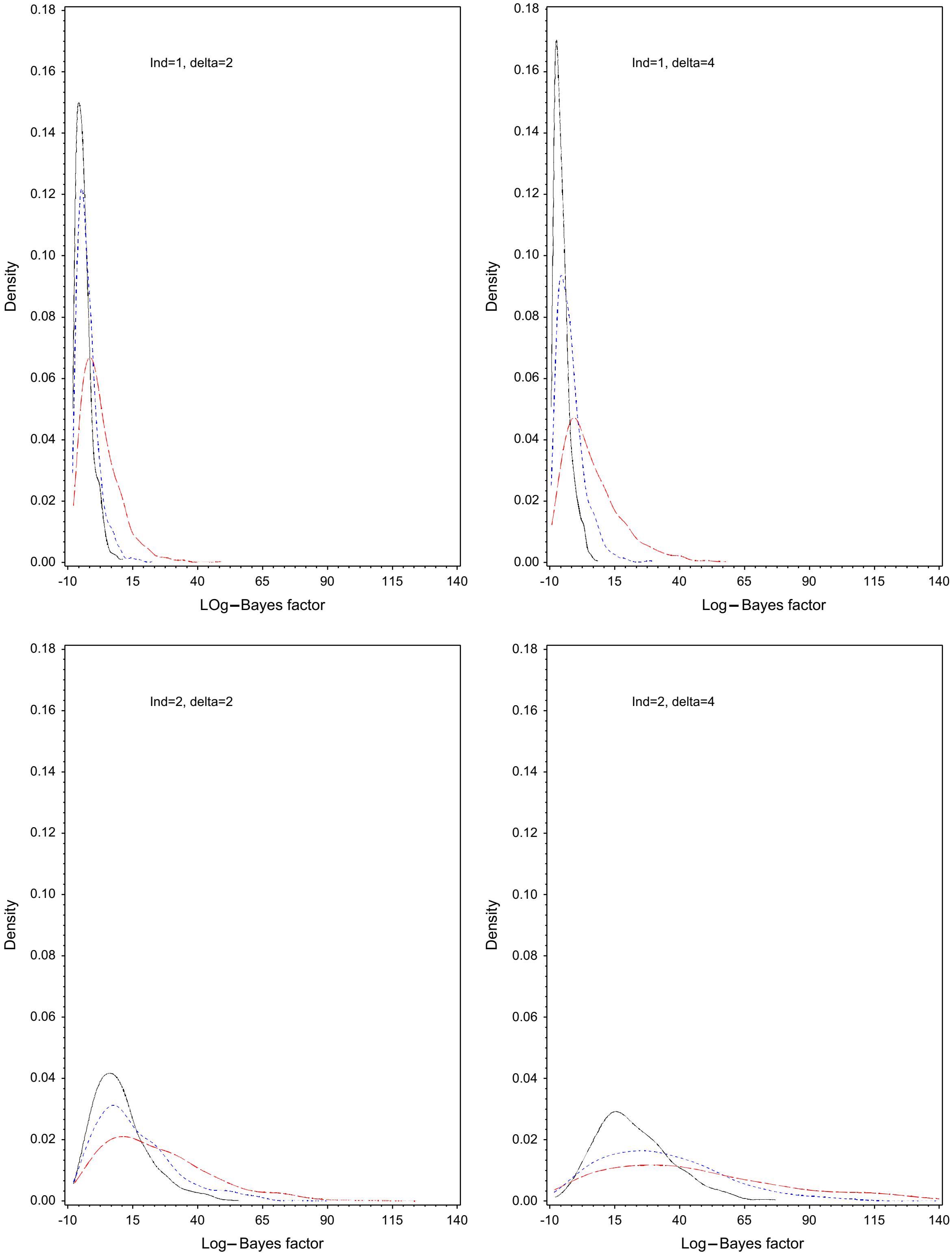

Fig. 2. Simulation: plots of the empirical densities of the log-Bayes factors at twelve design points. The symbols are correlation (solid: $\rho=.01$, dotted: $\rho=.10$, long dashed: $\rho=.30$), association (independence: ind=1 and dependence: ind=2) and table density (delta).

Table 6
Simulation: comparison of the log-Bayes factors and the *p*-values by ind.

Ind	*p*-Values		log-Bayes factor			
	χ^2	RSS	Mode	Q_1	Q_2	Q_3
1.00	.0943	.0633	−0.63	−1.37	0.16	2.35
1.25	.0348	.0281	−0.49	−0.84	1.03	3.61
2.75	.0038	.0045[a]	1.11	0.37	2.73	5.80
3.00	.0032	.0042[a]	1.62	0.84	3.45	6.61

Note : See notes to Tables 3, 5. At each value of ind the number of clusters is $\ell=25$ and the total number of observations is $n=50$. The intraclass correlation is $\rho=0.50$ giving an equivalent sample size of less than 25.

[a] $\chi^2 < RSS$.

permits the analyst to give the strength of the evidence both 'for' and 'against' independence unlike the frequentist decision approach which is directed to rejection of the null hypothesis of independence.

The methodology, described in this paper, is appropriate if there is an ignorable two-stage design and the model in (2)–(5) fits the observed data. If the survey weights provide additional information, one may include them by modifying the likelihood in (A.3) to

$$L_1=\prod_{j=1}^{n_i}\left[\left\{\prod_{s=1}^{S-1}(\alpha_{is}\pi_s)^{I^*_{ijs}}\right\}\left\{1-\sum_{s=1}^{S-1}(\alpha_{is}\pi_s)\right\}^{1-\sum_{s=1}^{S-1}I^*_{ijs}}\right]^{w_{ij}}$$

for the *i*th cluster, where the survey weights, w_{ij}, are rescaled to sum to n, the total sample size, and j indexes the second stage units with $I^*_{ijs}=1$, if unit ij is in cell s, and $I^*_{ijs}=0$, otherwise. Since

$$L_1=\left\{\prod_{s=1}^{S-1}(\alpha_{is}\pi_s)^{\sum_{j=1}^{n_i}w_{ij}I^*_{ijs}}\right\}\left\{1-\sum_{s=1}^{S-1}(\alpha_{is}\pi_s)\right\}^{\sum_{j=1}^{n_i}w_{ij}(1-\sum_{s=1}^{S-1}I^*_{ijs})}$$

has the same form as in (A.3), the analysis is essentially the same as that provided in the paper. If the sample design is informative, Pfeffermann and Sverchkov (2009) and Pfeffermann (2011) provide good coverage of most methods, but these are limited for Bayesian analysis.

With an additional step the methodology is also appropriate if there is a stratified two-stage design. Considering H strata, we simply need to add the subscript $h,h=1,\ldots,H$, to the variables in (2)–(5).

$$\underset{\sim}{n}_{hi}|\underset{\sim}{a}_{hi}\overset{ind}{\sim}\text{Multinomial}(n_{hi},\underset{\sim}{a}_{hi}),\tag{14}$$

where $\underset{\sim}{n}_{hi}=(n_{hi1},\ldots,n_{hiS})$, $n_{hi}=\sum_{s=1}^{S}n_{his}$ and $a_{his}=\alpha_{his}\pi_s$, $h=1,\ldots,H$, $i=1,\ldots,\ell$, $s=1,\ldots,S$. In (14) we have the constraints, $\sum_s\alpha_{his}\pi_s=1$, $\sum_s\pi_s=1$, $\alpha_{his}\pi_s>0$ and $\pi_s>0$. Again the α_{his} are used to adjust for the clustering. A priori we take

$$\alpha_{his}|\tau_{hs},\nu_h\overset{ind}{\sim}\text{Gamma}(\tau_{hs},\tau_{hs}\nu_h),\underset{\sim}{\pi}\sim\text{Dirichlet}(\underset{\sim}{1}),\tag{15}$$

and the ν_h are independent with

$$p(\nu_h)\propto 1/\nu_h,\quad h=1,\ldots,H,\tag{16}$$

where τ_{hs}, $h=1,\ldots,H$, $s=1,\ldots,S$, are to be specified. To perform the surrogate sampling, we need the posterior density of $\underset{\sim}{\pi}$. With minimal changes, our computational procedure will go through.

Finally, we note that in small complex surveys, most cluster tables will have many zero cells (e.g., contingency tables with categorical variables having many levels). As noted above the problem of sparse total tables cannot be accommodated within the Rao–Scott framework. However, it may be possible to do so within our framework. For example, a likelihood ratio test of independence in a single contingency table with many sampling zeros is given by Nandram et al. (2012) assuming simple random sampling. It will be useful to extend this work to complex surveys.

Appendix A. Joint posterior density

Letting $S=rc$, the set of constraints is

$$T=\left\{(\underset{\sim}{\alpha},\underset{\sim}{\pi},\nu):\sum_{s=1}^{S}\alpha_{is}\pi_s=1,\ \sum_{s=1}^{S}\pi_s=1,\alpha_{is}>0,i=1,\ldots,\ell,\pi_s>0,s=1,\ldots,S,\nu>\nu_o\right\}.$$

Letting $b=\sum_{s=1}^{S}\tau_s$, the joint prior density is

$$p(\underset{\sim}{\alpha},\underset{\sim}{\pi},\nu|\underset{\sim}{\tau})\propto\nu^{\ell b-1}\prod_{i=1}^{\ell}\prod_{s=1}^{S}\alpha_{is}^{\tau_s-1}e^{-\nu\tau_s\alpha_{is}},(\underset{\sim}{\alpha},\underset{\sim}{\pi},\nu)\in T.\tag{A.1}$$

In (A.1) we want to accommodate the constraints, $\sum_{s=1}^{S}\alpha_{is}\pi_s=1, i=1,\ldots,\ell$, and $\sum_{s=1}^{S}\pi_s=1$. We have a convenient way of doing so.

We transform α_{iS}, $i=1,\ldots,\ell$, to ϕ_i and π_S to ϕ_0 keeping all other random variables untransformed such that

$$\sum_{s=1}^{S}\alpha_{is}\pi_s=1+\phi_i,\ i=1,\ldots,\ell \quad\text{and}\quad \sum_{s=1}^{S}\pi_s=1+\phi_0.$$

Our idea is to remove π_S and α_{iS}, $i=1,\ldots,\ell$ when ϕ_i, $i=0,1,\ldots,\ell$ are set to zero. Then, $\pi_S=1+\phi_0-\sum_{s=1}^{S-1}\pi_s$ and $\alpha_{iS}=(1+\phi_i-\sum_{s=1}^{S-1}\alpha_{is}\pi_s)/(1+\phi_0-\sum_{s=1}^{S-1}\pi_s)$, $i=1,\ldots,\ell$. Note that π_S and α_{iS} are all kept in (0,1).

The Jacobian of the transformation is $(|1+\phi_0-\sum_{s=1}^{S-1}\pi_s|)^{-\ell}$ and the joint prior density is

$$p(\underset{\sim}{\alpha}_{(S)},\underset{\sim}{\pi}_{(S)},\underset{\sim}{\phi},\nu)\propto\nu^{\ell b-1}\prod_{i=1}^{\ell}\left[\prod_{s=1}^{S-1}\alpha_{is}^{\tau_s-1}e^{-\nu\tau_s\alpha_{is}}\right.$$
$$\left.\times\left(\frac{1+\phi_i-\sum_{s=1}^{S-1}\alpha_{is}\pi_s}{1+\phi_0-\sum_{s=1}^{S-1}\pi_s}\right)^{\tau_S-1}e^{-\nu\tau_S\left((1+\phi_i-\sum_{s=1}^{S-1}\alpha_{is}\pi_s)/(1+\phi_0-\sum_{s=1}^{S-1}\pi_s)\right)}\left(|1+\phi_0-\sum_{s=1}^{S-1}\pi_s|\right)^{-1}\right],$$
$$0<\sum_{s=1}^{S-1}\alpha_{is}\pi_s<1,\ i=1,\ldots,\ell,\ 0<\sum_{s=1}^{S-1}\pi_s<1,\ \alpha_{is}\pi_s>0,\ \pi_s>0,\ \nu>\nu_o.$$

Then, letting

$$\tilde{T}=\left\{(\underset{\sim}{\alpha}_{(S)},\underset{\sim}{\pi}_{(S)},\nu):0<\sum_{s=1}^{S-1}\alpha_{is}\pi_s<1,\ i=1,\ldots,\ell,\ 0<\sum_{s=1}^{S-1}\pi_s<1,\alpha_{is}\pi_s>0,\pi_s>0,\right.$$
$$\left.s=1,\ldots,S-1,\nu>\nu_o\right\},$$

$$p(\underset{\sim}{\alpha}_{(S)},\underset{\sim}{\pi}_{(S)},\nu|\underset{\sim}{\phi}=\underset{\sim}{0})\propto\nu^{\ell b-1}\prod_{i=1}^{\ell}\left[\prod_{s=1}^{S-1}\alpha_{is}^{\tau_s-1}e^{-\nu\tau_s\alpha_{is}}\right.$$
$$\left.\times\left(\frac{1-\sum_{s=1}^{S-1}\alpha_{is}\pi_s}{1-\sum_{s=1}^{S-1}\pi_s}\right)^{\tau_S-1}e^{-\nu\tau_S((1-\sum_{s=1}^{S-1}\alpha_{is}\pi_s)/(1-\sum_{s=1}^{S-1}\pi_s))}\left(1-\sum_{s=1}^{S-1}\pi_s\right)^{-1}\right],\quad(\underset{\sim}{\alpha}_{(S)},\underset{\sim}{\pi}_{(S)},\nu)\in\tilde{T}.\tag{A.2}$$

Henceforth, for convenience, we will denote this prior distribution by $p(\underset{\sim}{\alpha}_{(S)},\underset{\sim}{\pi}_{(S)},\nu)$ which, we note, is improper.

Now, the conditional distribution of $\underset{\sim}{n}$ given $(\underset{\sim}{\alpha}_{(S)},\underset{\sim}{\pi}_{(S)},\nu)\in\tilde{T}$ is

$$p(\underset{\sim}{n}|\underset{\sim}{\alpha}_{(S)},\underset{\sim}{\pi}_{(S)},\nu)=\prod_{i=1}^{\ell}\left[n_i!\left(\prod_{s=1}^{S-1}(\alpha_{is}\pi_s)^{n_{is}}/n_{is}!\right)\left(1-\sum_{s=1}^{S-1}\alpha_{is}\pi_s\right)^{n_{iS}}\Big/n_{iS}!\right]\tag{A.3}$$

$n_{is}\geq0,\sum_{s=1}^{S}n_{is}=n_i,i=1,\ldots,\ell.$

Then, using Bayes' theorem, the joint posterior density is

$$p(\underset{\sim}{\alpha}_{(S)},\underset{\sim}{\pi}_{(S)},\nu\Big|\underset{\sim}{n})\propto\prod_{i=1}^{\ell}\left[n_i!\left(\prod_{s=1}^{S-1}(\alpha_{is}\pi_s)^{n_{is}}/n_{is}!\right)\left(1-\sum_{s=1}^{S-1}\alpha_{is}\pi_s\right)^{n_{iS}}/n_{iS}!\right]$$
$$\times\nu^{\ell b-1}\prod_{i=1}^{\ell}\left[\prod_{s=1}^{S-1}\alpha_{is}^{\tau_s-1}e^{-\nu\tau_s\alpha_{is}}\right.$$
$$\left.\times\left(\frac{1-\sum_{s=1}^{S-1}\alpha_{is}\pi_s}{1-\sum_{s=1}^{S-1}\pi_s}\right)^{\tau_S-1}e^{-\nu\tau_S((1-\sum_{s=1}^{S-1}\alpha_{is}\pi_s)/(1-\sum_{s=1}^{S-1}\pi_s))}\left(1-\sum_{s=1}^{S-1}\pi_s\right)^{-1}\right],\quad(\underset{\sim}{\alpha}_{(S)},\underset{\sim}{\pi}_{(S)},\nu)\in\tilde{T}.\tag{A.4}$$

Note that in (A.4) $\alpha_{iS}=(1-\sum_{s=1}^{S-1}\alpha_{is}\pi_s)/(1-\sum_{s=1}^{S-1}\pi_s)$ and $\pi_S=1-\sum_{s=1}^{S-1}\pi_s$. Finally, because the prior density in (A.2) is improper, the joint posterior density in (A.4) may be improper.

We next show that the joint posterior density in (A.4) is proper. We make the transformation $t_{is}=\alpha_{is}\pi_s$, $s=1,\ldots,S-1$, $i=1,\ldots,\ell$, keeping the π_s untransformed. The Jacobian of the transformation is $(\prod_{s=1}^{S-1}\pi_s)^{-\ell}$ and the joint posterior density becomes

$$p(\underset{\sim}{t},\underset{\sim}{\pi}_{(S)},\nu|\underset{\sim}{n})\quad\propto\nu^{\ell b-1}e^{-\nu\sum_{i=1}^{\ell}(\sum_{s=1}^{S-1}\tau_s(t_{is}/\pi_s)+\tau_S(1-\sum_{s=1}^{S-1}t_{is})/(1-\sum_{s=1}^{S-1}\pi_s))}\prod_{i=1}^{\ell}\left[\left(\prod_{s=1}^{S-1}t_{is}^{n_{is}+\tau_s-1}\right)\left(1-\sum_{s=1}^{S-1}t_{is}\right)^{n_{iS}+\tau_S-1}\right.$$
$$\left.\times\left\{\left(\prod_{s=1}^{S-1}\pi_s^{\tau_s}\right)\left(1-\sum_{s=1}^{S-1}\pi_s\right)^{\tau_S}\right\}^{-1}\right],\quad(\underset{\sim}{t},\underset{\sim}{\pi}_{(S)},\nu)\in T^*,$$

Please cite this article as: Nandram, B., et al., A Bayesian test of independence in a two-way contingency table using surrogate sampling. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.03.011

where

$$T^* = \left\{ (\underset{\sim}{t},\underset{\sim}{\pi}_{(S)},\nu) : 0 < \sum_{s=1}^{S-1} t_{is}, \sum_{s=1}^{S-1} \pi_s < 1, t_{is},\pi_s > 0, i=1,\ldots,\ell, s=1,\ldots,S-1, \nu > \nu_o \right\}.$$

Now, assuming $\ell b > 1$ and letting $F_{\ell b}(a) = \int_0^a t^{\ell b-1} e^{-t}/\Gamma(\ell b)\, dt$, the cdf of a gamma random variable and integrating out ν, we get

$$p(\underset{\sim}{t},\underset{\sim}{\pi}_{(S)}|\underset{\sim}{n}) \propto \{1-F_{\ell b}(A\nu_o)\}A^{-\ell b}$$
$$\times \prod_{i=1}^{\ell} \left[\left(\prod_{s=1}^{S-1} t_{is}^{n_{is}+\tau_s-1} \right) \left(1-\sum_{s=1}^{S-1} t_{is}\right)^{n_{iS}+\tau_S-1} \left\{ \left(\prod_{s=1}^{S-1} \pi_s^{\tau_s} \right) \left(1-\sum_{s=1}^{S-1} \pi_s\right)^{\tau_S} \right\}^{-1} \right], (\underset{\sim}{t},\underset{\sim}{\pi}_{(S)}) \in \tilde{T}^*, \tag{A.5}$$

where

$$\tilde{T}^* = \left\{ (\underset{\sim}{t},\underset{\sim}{\pi}) : 0 < \sum_{s=1}^{S-1} t_{is}, \sum_{s=1}^{S-1} \pi_s < 1, t_{is},\pi_s > 0, i=1,\ldots,\ell, s=1,\ldots,S-1 \right\},$$

and

$$A = \sum_{i=1}^{\ell} \left\{ \sum_{s=1}^{S-1} \tau_s \frac{t_{is}}{\pi_s} + \tau_S \left(\frac{1-\sum_{s=1}^{S-1} t_{is}}{1-\sum_{s=1}^{S-1} \pi_s} \right) \right\}.$$

Since $p(\underset{\sim}{t},\underset{\sim}{\pi}_{(S)}|\underset{\sim}{n})$ is finite on any compact subset of $\tilde{T}^*$, the integral of $p(\underset{\sim}{t},\underset{\sim}{\pi}_{(S)}|\underset{\sim}{n})$ over any compact subset of $\tilde{T}^*$ is finite. Thus, the joint posterior density $p(\underset{\sim}{t},\underset{\sim}{\pi}_{(S)},\nu|\underset{\sim}{n})$ is proper.

Appendix B. A property of the gamma distribution

Let $d_1,\ldots,d_n \overset{iid}{\sim} \text{Gamma}(e,ef)$. Let $A = \sum_{i=1}^n d_i/n$ and $G = (\prod_{i=1}^n d_i)^{1/n}$ denote, respectively, the arithmetic and the geometric mean of the d_i.

Lemma. *The maximum likelihood estimator (MLE) of f is $\hat{f} = A^{-1}$ which is the unique solution of*

$$\ln(\hat{f}) - \psi(\hat{f}) = \ln(A/G), \tag{B.1}$$

where $\psi(\cdot)$ is the digamma function.

Proof of Lemma. The log-likelihood function is

$$\Delta(e,f) = n\{e \ln(f) + e \ln(e) + (e-1)\ln(G) - efA - \ln(\Gamma(e))\}.$$

Differentiating, we have

$$\frac{\partial \Delta(e,f)}{\partial f} = ne\left(\frac{1}{f} - A\right) \quad \text{and} \quad \frac{\partial^2 \Delta(e,f)}{\partial f^2} = -\frac{ne}{f^2}. \tag{B.2}$$

Using (B.2) it follows that the MLE of f is unique and is given by $\hat{f} = A^{-1}$.

Thus, the profile log-likelihood is

$$\Delta(e,\hat{f}) = n\{e \ln(\hat{f}) + e \ln(e) + (e-1)\ln(G) - e - \ln(\Gamma(e))\}.$$

Differentiating, we have

$$\frac{\partial \Delta(e,\hat{f})}{\partial e} = n\{\ln(e) - \psi(e) + \ln(G/A)\} \quad \text{and} \quad \frac{\partial^2 \Delta(e,\hat{f})}{\partial e^2} = \frac{1}{e} - \psi'(e), \tag{B.3}$$

where $\psi'(\cdot)$ is the trigamma function.

Then, because $e\psi'(e) > 1$ for all positive real numbers e (Abramowitz and Stegun, 1972, Chapter 6), it follows from (B.3) that the MLE of e is the unique solution of (B.1).

Appendix C. Mode of a kernel density estimator

Let $x_1,\ldots,x_n \overset{iid}{\sim} f(x)$, where $f(x)$ is an unknown density function. We need the mode of this density function based on a large sample of size n. We use the Parzen–Rosenblatt kernel density estimator with a standard normal kernel and optimal window width (Silverman, 1986), where

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \phi\left(\frac{x-x_i}{h}\right), \quad -\infty < x < \infty, \tag{C.1}$$

and h is the optimal window width.

Please cite this article as: Nandram, B., et al., A Bayesian test of independence in a two-way contingency table using surrogate sampling. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.03.011

Using differentiation,

$$\hat f'(x) = -\frac{1}{nh^3}\sum_{i=1}^{n}(x-x_i)\phi\left(\frac{x-x_i}{h}\right)$$

and

$$\hat f''(x) = -\frac{1}{nh^3}\sum_{i=1}^{n}\left\{1-\left(\frac{x-x_i}{h}\right)^2\right\}\phi\left(\frac{x-x_i}{h}\right).$$

A necessary condition for a mode x^* is that $\hat f'(x^*)=0$ which gives

$$x^* = \sum_{i=1}^{n} w\{(x^*-x_i)\}x_i, \tag{C.2}$$

where $w\{(x^*-x_i)\} = \phi((x^*-x_i)/h)\{\sum_{i=1}^{n}\phi((x^*-x_i)/h)\}^{-1}$, $i=1,\ldots,n$ (i.e., x^* is a weighted average).

We use a simple iterative procedure to solve (C.2). Starting with the sample mean on the right side of (C.2), we update x^* and iterate the procedure. This procedure is very fast even though it can take a large number of iterations for convergence. We need to check that $\hat f''(x^*)<0$. This is approximately true because $\{1-((x-x_i)/h)^2\} \approx \exp\{-((x-x_i)/h)^2\}$ which is positive. In fact, it is easy to show that $\hat f''(x^*) \ge -h^{-1}$; so it can be negative.

Alternatively, the global mode can be found by drawing samples from (C.1) and then finding the maximum of the values of $\hat f(x)$ over these samples; this procedure is easy and fast. We have performed both procedures and they give virtually the same answer; but the latter procedure is expected to work always (Robert and Casella, 1999, Chapter 5) for more complex optimization procedures.

Appendix D. Bayes factor for a test of independence

For the $r \times c$ contingency table, we can consider two multinomial–Dirichlet models, one with association and the other with no association.

The model with association is

$$\underset{\sim}{n}|\underset{\sim}{\pi} \sim \text{Multinomial}(n,\underset{\sim}{\pi}) \quad \text{and} \quad \underset{\sim}{\pi} \sim \text{Dirichlet}(\underset{\sim}{u}), \tag{D.1}$$

where $\underset{\sim}{u}$ is specified.

Letting $\pi_{jk}^{*} = \pi_j^{(1)}\pi_k^{(2)}$, $j=1,\ldots,r$, $k=1,\ldots,c$, the model with no association is

$$\underset{\sim}{n}|\underset{\sim}{\pi}^{(1)},\underset{\sim}{\pi}^{(2)} \sim \text{Multinomial}(n, \underset{\sim}{\overset{*}{\pi}}),$$

$$\underset{\sim}{\pi}^{(1)} \sim \text{Dirichlet}(\underset{\sim}{v}) \text{ and independently } \underset{\sim}{\pi}^{(2)} \sim \text{Dirichlet}(\underset{\sim}{w}), \tag{D.2}$$

where $\underset{\sim}{\pi}^{(1)}$ and $\underset{\sim}{\pi}^{(2)}$ have r and c components, respectively, and $\underset{\sim}{v}$ and $\underset{\sim}{w}$ are specified.

Therefore, integrating out $\underset{\sim}{\pi}^{(1)}$ and $\underset{\sim}{\pi}^{(2)}$ from (D.2) and $\underset{\sim}{\pi}$ from (D.1), it is easy to show that the marginal likelihood with association (as) is $p_{\text{as}}(\underset{\sim}{n}) = (n!/\prod_{j=1}^{r}\prod_{k=1}^{c}n_{jk}!)D(\underset{\sim}{n}+\underset{\sim}{u})/D(\underset{\sim}{u})$ and with no association (nas) is

$$p_{\text{nas}}(\underset{\sim}{n}) = p_{\text{as}}(\underset{\sim}{n})\left\{\frac{D(\underset{\sim}{n}^{(1)}+\underset{\sim}{v})}{D(\underset{\sim}{v})}\frac{D(\underset{\sim}{n}^{(2)}+\underset{\sim}{w})}{D(\underset{\sim}{w})}\bigg/\frac{D(\underset{\sim}{n}+\underset{\sim}{u})}{D(\underset{\sim}{u})}\right\}, \tag{D.3}$$

where $\underset{\sim}{n}^{(1)} = (n_{1.},\ldots,n_{r.})'$ and $\underset{\sim}{n}^{(2)} = (n_{.1},\ldots,n_{.c})'$. Thus, using (D.3) the Bayes factor (BF) is given by

$$BF = p_{\text{as}}(\underset{\sim}{n})/p_{\text{nas}}(\underset{\sim}{n}), \tag{D.4}$$

which provides evidence for association relative to no association. With Jeffreys' prior (i.e., elements of $\underset{\sim}{u}$, $\underset{\sim}{v}$ and $\underset{\sim}{w}$ are all 0.5) there is no simplification to (D.3) or (D.4).

However, for the special case where $\underset{\sim}{u}=\underset{\sim}{1}$, $\underset{\sim}{v}=\underset{\sim}{1}$ and $\underset{\sim}{w}=\underset{\sim}{1}$ (i.e., uniform priors), we have $p_{\text{as}}(\underset{\sim}{n}) = (rc-1)!n!/(n+rc-1)!$ and with no association (nas),

$$p_{\text{nas}}(\underset{\sim}{n}) = p_{\text{as}}(\underset{\sim}{n})\frac{(r-1)!(c-1)!}{(rc-1)!}\frac{(n+rc-1)!}{(n+r-1)!(n+c-1)!}\frac{\prod_{j=1}^{r}n_{j\cdot}!\prod_{k=1}^{c}n_{\cdot k}!}{\prod_{j=1}^{r}\prod_{k=1}^{c}n_{jk}!}. \tag{D.5}$$

References

Abramowitz, M., Stegun, I.A., 1972. Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. Dover Publications, New York.

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

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

Dong, Q., Elliott, M., Raghunathan, T., 2011. Combining information from multiple complex surveys. In: Proceedings of the Joint Meetings of the American Statistical Association.

Gelman, A., Carlin, J., Stern, H., Rubin, D., 2004. Bayesian Data Analysis, 2nd ed. Chapman & Hall, New York.

Holt, D., Scott, A.J., Ewings, P.D., 1980. Chi-squared tests with survey data. Journal of the Royal Statistical Society, Series A 143, 303–320.

Kass, R.E., Raftery, A.E., 1995. Bayes factor. Journal of the American Statistical Association 90, 773–795.

Nandram, B., Bhatta, D., Bhadra, D., 2012. A Likelihood Ratio Test of Quasi-independence for Sparse Two-way Contingency Tables. Technical Report, Department of Mathematical Sciences, Worcester Polytechnic Institute.

Nandram, B., 2007. Bayesian predictive inference under informative sampling via surrogate samples. In: Upadhyay, S.K., Umesh Singh, Dipak K. Dey (Eds.), Bayesian Statistics and its Applications. Anamaya, New Delhi, pp. 356–374 (Chapter 25).

Nandram, B., Choi, J.W., 2007. Alternative tests of independence in two-way categorical tables. Journal of Data Science 5, 217–237.

Pfeffermann, D., Sverchkov, M., 2009. Inference under informative sampling. In: Pfeffermann, D., Rao, C.R. (Eds.), Handbook of Statistics 29; Survey Sampling: Design, Methods and Applications, vol. 29A. . North Holland, Amsterdam, pp. 455–487. (Chapter 39).

Pfeffermann, D., 2011. Modelling complex survey data: Why model? Why it is a problem? How can we approach it. Survey Methodology 37 (2), 115–136.

Rao, J.N.K., 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., Scott, A.J., 1984. On chi-squared tests for multi-way tables with cell proportions estimated from survey data. Annals of Statistics 12, 46–60.

Rao, J.N.K., Thomas, D.R., 1989. Chi-squared tests for contingency tables. In: Holt, D., Skinner, C.J., Smith, T.M.F. (Eds.), The Analysis of Complex Surveys. Wiley, New York.

Robert, C.P., Casella, G., 1999. Monte Carlo Statistical Methods. Springer, New York.

Scott, A.J., Holt, D., 1982. The effect of two-stage sampling on ordinary least squares methods. Journal of the American Statistical Association 77, 848–854.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, New York.

Sugden, R.A., Smith, T.F.M., 1984. Ignorable and informative designs in survey sampling inference. Biometrika 71, 495–506.

Thomas, D.R., 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., Roberts, G.R., 1996. A simple method for the analysis of clustered data. International Statistical Review 64, 295–311.

Valliant, R., Dorfman, A.H., Royall, R.M., 2000. Finite Population Sampling and Inference: A Prediction Approach. Wiley, New York.

Please cite this article as: Nandram, B., et al., A Bayesian test of independence in a two-way contingency table using surrogate sampling. Journal of Statistical Planning and Inference (2013), http://dx.doi.org/10.1016/j.jspi.2013.03.011