The Full Bayesian Significance Test for marginal homogeneity in two-way contingency tables

Helton Graziadei de Carvalho*
University of Sao Paulo, Sao Paulo, Brazil - heltongc@ime.usp.br

Luís Gustavo Esteves
University of Sao Paulo, Sao Paulo, Brazil - lesteves@ime.usp.br

Abstract

Matched sample studies have become increasingly popular in a wide range of areas especially those dealing with categorical variables. For instance, suppose that for a group of individuals, the quality of the vision in each eye is classified into four categories. In this context, a two-way contingency table with the same row and column categories summarizes the data and it is usual to investigate whether the qualities of left and right eyes are the same or, equivalently, if the marginal distributions of the qualities of vision are the same, the so-called Marginal Homogeneity (MH) hypothesis. Classical approaches to the problem of testing MH rely on the asymptotic distribution of the test statistics which may yield imprecise results in certain situations. To overcome these limitations, we develop the Full Bayesian Significance Test (FBST) for MH in two-dimensional contingency tables. The FBST is a procedure that has some important features such as: (i) it obeys the likelihood principle (ii) it does not rely on asymptotic distributions (iii) it does not depend on the elimination of nuisance parameters. Furthermore, we calculate p-values and compare them with the FBST (iv) it is coherent in simultaneous hypothesis testing. To summarize, we propose a coherent measure of evidence to test MH and compare it with classical approaches to the problem.

Keywords: Marginal homogeneity test; Full Bayesian Significance Test; Matched studies.

1. Introduction

The problem of comparing the marginal discrete distributions for two paired-samples plays an important role in a variety of subjects such as: genetics, demography, politics and psychology (Agresti, 2002). In genetics, for example, the Transmission Disequilibrium Test (TDT) is an application of the MH hypothesis. We next present another example to illustrate the MH test as well as the techniques to be presented.

Example 1.1: Table 1 presents frequencies regarding the vision quality of the left and right eye for a group of 7477 women during the Second World War (Stuart, 1953). Let $\mathbb{C} = \{$Highest , Second, Third, Lowest$\}$ be the set of possible accuracies for each eye. In this context, considering a multinomial model we have that the parameter space is: $\Theta = \{\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \ldots, \theta_{14}, \ldots, \theta_{41}, \theta_{42}, \ldots, \theta_{44}) \in \mathbb{R}_+^{16}, \sum_{i=1}^{4} \sum_{j=1}^{4} \theta_{ij} = 1\}$, where θ_{ij} is the probability that the individual's right eye vision is classified in the j-th category of $\mathbb{C}$ and the left eye quality is classified in the i-th category of $\mathbb{C}$, $\forall i, j = 1, \ldots, 4$. The sample space is: $\chi = \{(n_{11}, n_{12}, n_{13}, n_{14}, \ldots, n_{41}, \ldots, n_{44}) \in \mathbb{N}^{16} : \sum_{i=1}^{4} \sum_{j=1}^{4} n_{ij} = 7477\}$, where n_{ij} represents the count of cell (i, j), $i, j = 1, \ldots, 4$.

In this scenario, one hypothesis of interest may be formulated as $H : \theta_{i+} = \theta_{+i}, i = 1, 2, 3$, where $\theta_{i+} = \sum_{j=1}^{k} \theta_{ij}$ and $\theta_{+i} = \sum_{j=1}^{k} \theta_{ji}$, that is, H means that individuals have the same individuals have the same quality in both eyes (marginal homogeneity). A further hypothesis that is common to be investigated in matched-sample studies is $H' : \theta_{ij} = \theta_{ji}, (i, j) \in \mathbb{C} \times \mathbb{C}$ such that $i \neq j$, which is called symmetry. The p-value obtained from the likelihood ratio test for H (Madansky, 1956) is equal to 0.009. Additionally, using a generalization of McNemar test (Bowker, 1948), the p-value for H' is 0.080. Hence, adopting a 5% (or even 1%) significance level, we conclude simultaneously that the distributions of the qualities of vision are different, while the corresponding joint distribution is symmetric, which seem to be inconsistent because $H' \subset H$ (Agresti, 2002). Hence, aiming to overcome these logical flaws usually met by p-values, Bayes

Tabela 1: Unaided distance vision of 7455 women in Britain (Stuart, 1953)

	Highest	Second	Third	Lowest	Total
Highest	1520	266	124	66	1976 (0.264)
Second	234	1512	432	78	2256 (0.302)
Third	117	362	1772	205	2456 (0.328)
Lowest	36	82	179	492	789 (0.106)
Total	1907 (0.256)	2222 (0.297)	2507 (0.335)	841(0.112)	7477

factors, among others, we develop along the next sections a genuinely bayesian measure of evidence to test H which is also coherent according to Lavine and Schervish (1999).

2. The Full Bayesian Significance Test (FBST)

Testing sharp hypothesis has been a controversial problem in Statistical Inference. In this context, the FBST (Pereira and Stern, 1999) was developed as an alternative to overcome some difficulties usually met by frequentist and bayesian tests. For instance, the FBST does not require the adoption of a positive prior probability for the subset defined by the sharp hypothesis, denoted by Θ_0, as Jeffrey's tests do. Next, we briefly review the construction of the FBST.

Suppose a bayesian statistical model, i.e., $\Theta \subset \mathbb{R}^k$ is the parameter space and $\chi \subset \mathbb{R}^k$ is the sample space. Also, $f(\theta)$ is a prior probability density over Θ and $L_x(\theta)$ is the likelihood function generated by an observation $x \in \chi$. Consider that a sharp hypothesis $H : \theta \in \Theta_0$ (that is, $dim(\Theta_0) < dim(\Theta)$) is to be tested. The FBST is based on the measure of evidence, called e-value, described in the sequel. To calculate the e-value, let $f(\theta|x)$ be the posterior density function for θ given by

$$f(\theta|x) \propto f(\theta)L_x(\theta).$$

Let $T_x = \{\theta \in \Theta : f(\theta|x) > \sup_{\theta \in \Theta_0} f(\theta|x)\}$ be the tangential (to Θ_0) set which is composed of the points in parameter space that are more consistent with x than the posterior mode under the null hypothesis. The e-value in favour of H is defined as

$$ev(\Theta_0; x) = 1 - P(\theta \in T_x|x). \tag{1}$$

As defined by Pereira and Stern (1999), the FBST is the procedure that rejects H whenever $ev(\Theta_0; x)$ is small. In addition, it should be emphasized that the posterior density is the only requirement to calculate e-values. In the next section, we develop the FBST for MH in two-dimensional contingency tables.

3. Marginal Homogeneity for two-dimensional contingency tables

Suppose a random vector $\boldsymbol{X}$ that conditionally on θ is distributed according to a Multinomial distribution. In an analogous way with Section 2, the parameter space is $\Theta = \{\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \ldots, \theta_{1k}, \theta_{21}, \theta_{22}, \ldots, \theta_{2k}, \ldots, \theta_{k1}, \theta_{k2}, \ldots, \theta_{kk}) \in \mathbb{R}_+^{k^2}, \sum_{i=1}^k \sum_{j=1}^k \theta_{ij} = 1\}$, whereas the sample space is $\chi = \{(n_{11}, n_{12}, \ldots, n_{1k}, n_{21}, n_{22}, \ldots, n_{2k}, \ldots, n_{k1}, n_{k2}, \ldots, n_{kk}) \in \mathbb{N}^{k \times k} : n_{++} = n\}$. The marginal homogeneity hypothesis is written as

$$H : \theta_{i+} = \theta_{+i}, i = 1, \ldots, k - 1. \tag{2}$$

Notice that this is sharp hypothesis since $dim(\Theta_0) = k(k - 1) < k^2 - 1 = dim(\Theta)$. Furthermore, because $\boldsymbol{X}|\boldsymbol{\theta} \sim \text{Multinomial}(n; \boldsymbol{\theta})$, the likelihood function generated by $\boldsymbol{X} = \boldsymbol{x} = (n_{11}, \ldots, n_{kk}) \in \chi$ is given by

$$L_x(\boldsymbol{\theta}) \propto \left[\prod_{i=1}^k \prod_{j=1}^k \theta_{ij}^{n_{ij}} \right] \mathbb{1}_\Theta(\boldsymbol{\theta}) \tag{3}$$

where Θ is a $(k^2 - 1)$-simplex. Suppose that $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{a})$. Then, using Bayes Theorem, we have that the kernel of the posterior density is

$$f(\boldsymbol{\theta}|\boldsymbol{x}) \propto \left[\prod_{i=1}^{k} \prod_{j=1}^{k} \theta_{ij}^{n_{ij}+a_{ij}-1} \right] \mathbb{1}_{\Theta}(\boldsymbol{\theta}). \tag{4}$$

Thus, $\boldsymbol{\theta}|\boldsymbol{x} \sim \mathrm{Dirichlet}(\boldsymbol{a}+\boldsymbol{x})$, where $\boldsymbol{a}+\boldsymbol{x} = (n_{11}+a_{11}, n_{12}+a_{12}, \ldots, n_{1k}+a_{1k}, \ldots, n_{k1}+a_{k1}, \ldots, n_{kk}+a_{kk})$. To obtain comparisons with frequentists results, we define $a_{ij} = 1, \forall\, i,j \in \{1, \ldots, k\}$ as long as the posterior mode is equal to the maximum likelihood estimator of $\boldsymbol{\theta}$ in this case.

4. FBST for marginal homogeneity

In this section, we describe some numerical optimization techniques and Monte Carlo methods to calculate the e-value in favour of the marginal homogeneity hypothesis for two-dimensional contingency tables, namely $\Theta_0 = \{\boldsymbol{\theta} \in \Theta : \theta_{i+} = \theta_{+i}, i = 1, \ldots, k-1\}$. Firstly, it is necessary to specify the tangential set T_x for Θ_0. In order to do so, we first need to maximize the kernel of the log-posterior density that is

$$\mathrm{Maximize} \sum_{i=1}^{k} \sum_{j=1}^{k} (n_{ij} + a_{ij} - 1) \log \theta_{ij}$$

subject to k constraints

$$\theta_{i+} = \theta_{+i}, \quad i = 1, \ldots, k-1.$$

$$\sum_{i=1}^{k} \sum_{j=1}^{k} \theta_{ij} = 1$$

Using a vector of Lagrange Multipliers $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_{k-1})$, we need to maximize $\mathcal{L}$,

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{i=1}^{k} \sum_{j=1}^{k} (n_{ij} + a_{ij} - 1) \log \theta_{ij} - \lambda_0 \left(\sum_{i=1}^{k} \sum_{j=1}^{k} \theta_{ij} - 1 \right) - \sum_{l=1}^{k-1} \lambda_l (\theta_{l.} - \theta_{.l})$$

The derivatives are given by:

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\theta_{ij}} = \frac{(n_{ij} + a_{ij} - 1)}{\theta_{ij}} - \lambda_0 - (\lambda_i - \lambda_j) \quad \forall i,j = 1, \ldots, k$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\lambda_l} = (\theta_{l.} - \theta_{.l}) \quad \forall l = 1, \ldots, k-1$$

$$\frac{\partial f(\boldsymbol{\theta}|\boldsymbol{x})}{\lambda_0} = \left(\sum_{i=1}^{k} \sum_{j=1}^{k} \theta_{ij} - 1 \right).$$

Equating them to zero, we obtain the following system of equations:

$$(n_{ij} + a_{ij} - 1) = \theta_{ij}(\lambda_0 + \lambda_i - \lambda_j) \quad \forall i,j = 1, \ldots, k.$$

$$\theta_{i+} = \theta_{+i} \quad \forall i = 1, \ldots, k-1$$

$$\sum_{i=1}^{k} \sum_{j=1}^{k} \theta_{ij} = 1$$

It is easy to see that $\lambda_0 = n$, $\tilde{\theta}_{ii} = \frac{n_{ii}+a_{ii}-1}{n}$, $i = 1, \ldots, k$, which is equal to the corresponding coordinate of the posterior mode, and

$$\tilde{\theta}_{ij} = \frac{n_{ij} + a_{ij} - 1}{n + \lambda_i - \lambda_j} \tag{5}$$

subject to

$$\sum_{\substack{j=1 \\ j \neq i}}^{k} \frac{n_{ij} + a_{ij} - 1}{n + \lambda_i - \lambda_j} = \sum_{\substack{j=1 \\ j \neq i}}^{k} \frac{n_{ji} + a_{ji} - 1}{n + \lambda_j - \lambda_i}, i = 1, \ldots, k - 1. \tag{6}$$

From Equation 6, it is possible to obtain the Lagrange Multipliers regardless of $\boldsymbol{\theta}$. Next, we use Equation 5 to determine the estimator of $\boldsymbol{\theta}$ under H.

After finding the maximum of the posterior density under H, we need to calculate the posterior probability of the tangential set, $\mathbb{P}(\theta \in T_x | x)$. We perform this calculation by means of Monte Carlo method: we generate $\theta_1, \theta_2, \ldots, \theta_M$ of the posterior density and compare their densities with the maximum density under H. Let $\boldsymbol{\theta^*} = \arg\max\{\pi(\boldsymbol{\theta}|\boldsymbol{x}) : \boldsymbol{\theta} \in \Theta_0\}$ and define

$$\mathbb{1}_A(\boldsymbol{\theta_i}) = \left\{ \begin{array}{ll} 1, & \text{if} \quad f(\boldsymbol{\theta_i}|\boldsymbol{x}) \geq f(\boldsymbol{\theta^*}|x), \\ 0, & \text{otherwise,} \end{array} \right.$$

$i = 1, \ldots, M$. By the Strong Law of Large Numbers,

$$\frac{\sum_{i=1}^{M} \mathbb{1}_A(\boldsymbol{\theta_i})}{M} \xrightarrow{a.s.} P(\boldsymbol{\theta} \in T_x | x). \tag{7}$$

Thus, we approximate the evidence $ev(\Theta_0; \boldsymbol{x})$ in favor of the marginal homogeneity hypothesis for a $k \times k$ contingency table using the Monte Carlo Method, that is,

$$\tilde{ev}(\Theta_0; x) = 1 - \frac{\sum_{i=1}^{M} \mathbb{1}_A(\boldsymbol{\theta_i})}{M}. \tag{8}$$

In Example 1.1, the evidence for marginal homogeneity is $ev(\Theta_0; x) = 0.68$ whereas for symmetry is $ev(\Theta_0'; x) = 0.20$, leading us to conclude that the data give more evidence to support MH hypothesis than the Symmetry hypothesis.

5. Conclusions

In this work, we proposed the FBST for marginal homogeneity hypothesis in two-dimensional contingency tables. To calculate the e-value, the evidence measure under the FBST, simple numerical and Monte Carlo techniques were implemented with high performance computing without reference to asymptotic distributions for the test statistics or to resampling methods. We showed that it is possible to obtain the posterior mode under MH hypothesis from lagrange multipliers to decrease the computational cost. From an example, we illustrated that e-values are coherent for nested hypotheses (Lavine and Schervish, 1999) whereas p-values may yield inconsistent conclusions.

References

Agresti, A. (2002). Categorical Data Analysis. John Wiley & Sons.

Bowker, A.H. (1948). A Test for Symmetry in Contingency Tables. Journal of the American Statistical Association.

Lavine, M and Schervish, M.J. (1999). Bayes Factors: What They Are and What They Are Not? The American Statistician.

Madansky, A. (1963). Tests of Homogeneity for Correlated Samples. Journal of the American Statistical Association. Taylor & Francis.

Stuart, A. (1953). The Estimation and Comparison of Strengths of Association in Contingency Tables. Biometrika.

Pereira et al. Can a significance test be genuinely bayesian? Bayesian Analysis.

Pereira, C.A.B., & Stern, J.M. (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypothesis. Entropy.