

Exact conditional inference on h -sample problems for categorical data using algebraic algorithms

Inferenza condizionale esatta su problemi ad h campioni per dati categorici tramite algoritmi algebrici

Daniele De Martini
Dipartimento SEMEQ
Università del Piemonte Orientale
demartin@eco.unipmn.it

Fabio Rapallo
Dipartimento di Matematica
Università di Genova
rapallo@dima.unige.it

Riassunto: Nella recente letteratura il problema del calcolo del p -value nell'inferenza condizionata è stato trattato diffusamente e sono state fornite diverse soluzioni computazionali. In questo lavoro si propone l'utilizzo di un algoritmo algebrico per il calcolo approssimato del p -value esatto per test di uguaglianza tra distribuzioni categoriche. Tale algoritmo è basato sulla teoria dell'algebra computazionale ed è implementato usando un metodo MCMC.

Keywords: p -value computation; independence model; MCMC methods.

1. Introduction

This paper deals with the computation of the exact p -values for tests on equality of h multinomial distributions and its links with the test of independence on $h \times k$ tables. It is worth noting that when the p -value is not exactly computed, classical approximations based on chi-squared asymptotic distribution of the test statistic are used. Unfortunately, asymptotic approximations in many situations are not accurate, indeed the problem of computing exact p -values has been widely considered in recent literature.

Mehta and Patel (1983) provide a network algorithm to inspect the orbit of $h \times k$ contingency tables with fixed marginals. But, the entire set of tables is often too wide and consequently approximation methods have been proposed.

The algorithm proposed by Metha *et al.* (1988) provides an approximation of the exact p -value and it is based on a complicated Monte Carlo (MC) method which generates random tables providing statistics belonging to the critical region of the test. Recently, Strawderman and Wells (1998) used saddlepoint approximations to the exact distribution of the conditional maximum likelihood estimate (MLE) for several 2×2 tables, providing approximated p -values, power calculations and confidence intervals. Booth and Butler (1999) proposed a MC importance sampling for conducting exact conditional tests in log-linear models. Moreover, Metha *et al.* (2000) proposed a MC method for approximating the exact p -values in conditional logistic regression and foresee the applicability of algebraic-based methods in this framework.

The main problem in Metha *et al.* (1988), and in most cited works, was how to generate random tables with fixed marginals. In this paper we give a new solution to the h -sample problem applying the technique described in Diaconis and Sturmfels (1998), and the computer program source can be included in few lines of statements. Moreover, we provide a solution not only for $2 \times k$ tables but, more generally, for $h \times k$ tables.

The h -sample problem can be stated as follows: let $X_1, \dots, X_h$ be random variables with

these groups: $(X_{1,1}, \dots, X_{1,n_1})$ from the multinomial distribution $D_1, \dots, (X_{h,1}, \dots, X_{h,n_h})$ from the multinomial distribution D_h . Let $N = \sum_{i=1}^h n_i$ be the total sample size. The distribution D_i has parameters $p_{i1}, \dots, p_{ik}$, for $i = 1, \dots, h$ with the constraints $p_{ij} \geq 0$ and $\sum_{j=1}^k p_{ij} = 1$ for all $i = 1, \dots, h$. The usual test in this situation is the test for equality of the proportions, where the null hypothesis is

$$H_0 : p_{1j} = \dots = p_{hj} \quad \text{for } j = 1, \dots, k$$

against the composite alternative hypothesis of different proportions.

The components of the sufficient statistic are the sums of the observations in each group and the sums of observations for each possible value of the variables, that is

$$S = \left(\sum_{u,j} \mathbb{I}_{\{X_{1,u}=j\}}, \dots, \sum_{j,u} \mathbb{I}_{\{X_{h,u}=j\}}, \sum_{i,u} \mathbb{I}_{\{X_{i,u}=1\}}, \dots, \sum_{i,u} \mathbb{I}_{\{X_{i,u}=k\}} \right)$$

for $i = 1, \dots, h$ and $j = 1, \dots, k$, where $\mathbb{I}_A$ is the indicator function of A . Of course, the sample can be summarized in a contingency table with h rows and k columns and we denote this table by F ; in other words $F_{ij} = \sum_{u=1}^{n_i} \mathbb{I}_{\{X_{i,u}=j\}}$. The table F has row parameters $p_{ij}^* = p_{ij}n_i/N$. Moreover, we denote by f_{i+} and by f_{+j} the row sums and the column sums, respectively. The sufficient statistic S is represented by the margins of this table and the MLE of the parameters is

$$\hat{p}_{ij}^* = f_{i+}f_{+j}/N^2 \quad \text{for } i = 1, \dots, h, j = 1, \dots, k.$$

For this hypothesis the most commonly used test statistics are the Pearson statistic $C = \sum_{i,j} (F_{ij} - N\hat{p}_{ij}^*)^2 / N\hat{p}_{ij}^*$ and the log-likelihood ratio $L = 2 \sum_{i,j} F_{ij} \log(F_{ij}/N\hat{p}_{ij}^*)$, see for example Agresti (1996). In classical asymptotic theory it is well known that, under H_0 , both C and L have a limit chi-squared distribution with $(k-1)(h-1)$ degrees of freedom.

When the sample size is small, chi-squared approximations may be not adequate, see Agresti (1996). There are many nonparametric algorithms, based on MC approximations of the p -value which do not involve limit theorems.

The algorithm described here is still valid when the test statistic is of the form

$$T(F) = \sum_{i,j} a_{ij}(F_{ij})$$

where the a_{ij} 's are real valued functions on the set $\mathbb{N}$ of the nonnegative integers.

As usual in this framework, we consider a problem of conditional inference, where only the tables with the same margins of the original table are relevant. Then, we define the reference set

$$\mathcal{F} = \left\{ f' \mid f'_{i+} = f_{i+}, f'_{+j} = f_{+j} \right\}$$

i.e., the set of all tables with fixed margins. The MC approach to this problem is given by generating an *iid* sample of tables $f^{(1)}, \dots, f^{(B)}$ in $\mathcal{F}$ with the appropriate distribution $\mathcal{H}$ and computing the value of the test statistic T for each table $T(f^{(1)}), \dots, T(f^{(B)})$. The approximate MC p -value is simply the ratio of the number of tables whose value of the test statistic is greater than or equal to the one of the original table, over the number B of sampled tables. In the case of independence, $\mathcal{H}$ follows the hypergeometric distribution:

$$\mathcal{H}(f) = \frac{\prod_{j=1}^k \binom{f_{+j}}{f_{1j} \dots f_{hj}}}{\binom{N}{f_{1+} \dots f_{h+}}},$$

for further details see Rapallo (2001).

For the two-sample problem, there is an algorithm proposed by Metha *et al.* (1988), which is an importance sampling algorithm and is based on the network representation of the reference set $\mathcal{F}$. It were also used in order to approximate the power of the test, but is difficult to rewrite it for the h -sample problem.

2. Algebraic algorithm

Recently, techniques of commutative algebra has been used for statistical applications, see Pistone *et al.* (2000). More precisely, Diaconis and Sturmfels (1998) used the algebraic theory of toric ideals for giving a method for sampling from conditional distributions, where the sample space is finite.

The algebraic method is based on the Metropolis–Hastings algorithm and performs a Markov chain on $\mathcal{F}$ with the appropriate hypergeometric stationary distribution. The Markov chain is a random-walk-like chain on $\mathcal{F}$ with moves in a set $\mathcal{M} = \{m_1, \dots, m_L\}$, which must be a Markov basis to allow the connectedness of the Markov chain. A set of tables $\mathcal{M}$ with integer entries (even negative) is a Markov basis if its sufficient statistic is equal to $\mathbf{0}_{h+k}$ and such that for every pair of tables f and f' in $\mathcal{F}$ there is a path between f and f' through nonnegative tables.

In the first step, at time 0, the algorithm provides for the Markov chain to be in f ; second, it draws a move m from $\mathcal{M}$ and a random variable ϵ from $\{\pm 1\}$, both uniformly; then, if $f + \epsilon m \geq 0$ it moves the chain from f to $f + \epsilon m$ with probability $\min\{\mathcal{H}(f + \epsilon m)/\mathcal{H}(f), 1\}$, and in all other cases it remains at f .

Note that this method allows to approximate the p -value as stated before, instead of considering also the hypergeometric probability weights of the sampled tables. Moreover, in the last step of the algorithm, the hypergeometric probabilities are not calculated entirely, but only the ratio between two successive table probabilities is needed, with a considerable computational saving, see Rapallo (2001) for details.

Also Metha *et al.* (2000) considered the possibility of applying algebraic-based methods: “An investigation is needed to establish conditions under which the probability of connectness of the Markov chain is high”. But, there is a theorem (see Diaconis and Sturmfels (1998)) stating that the algebraic algorithm described above defines a connected Markov chain. Hence, algebraic-based methods can actually be used to compute the exact p -value also for conditional logistic regression and log-linear models.

Following Diaconis and Sturmfels (1998), the computation of the Markov basis is made through computational algebra, using a polynomial representation of the tables.

In the independence case, no algebraic computations are needed, because the Markov basis can be theoretically characterized, by computational algebra, see Rapallo (2001): the

Markov basis for the independence case is given by all the moves of the form $\begin{matrix} +1 & -1 \\ -1 & +1 \end{matrix}$

for all two-by-two minors of the table.

For example, the three moves for the 2×3 tables are

$$\begin{array}{ccc} +1 & -1 & 0 \\ -1 & +1 & 0 \end{array} \quad \begin{array}{ccc} +1 & 0 & -1 \\ -1 & 0 & +1 \end{array} \quad \begin{array}{ccc} 0 & +1 & -1 \\ 0 & -1 & +1 \end{array}$$

Finally, we emphasize that classical methods, for example Metha *et al.* (1988), proposed the solution by sampling few well chosen tables with great computational effort. On the

other hand, our method easily generates tables from the reference set. Future works will consider comparisons among different methods.

References

- Agresti A. (1996) *An Introduction to Categorical Data Analysis*, Wiley, New York.
- Booth J.G. and Butler R.W. (1999) An importance sampling algorithm for exact conditional tests in log-linear models, *Biometrika*, 86, 2, 321–331.
- Diaconis P. and Sturmfels B. (1998) Algebraic algorithms for sampling from conditional distributions, *Ann. Statist.*, 26, 1, 363–397.
- Mehta C.R. and Patel N.R. (1983) A network algorithm for performing Fisher’s exact test in $r \times c$ contingency tables, *J. Amer. Statist. Assoc.*, 78, 382, 427–434.
- Metha C.R., Patel N.R. and Senchaudhuri P. (1988) Importance sampling for estimating exact probabilities in permutational inference, *J. Am. Statist. Assoc.*, 83, 404, 999–1005.
- Metha C.R., Patel N.R. and Senchaudhuri P. (2000) Efficient monte carlo methods for conditional logistic regression, *J. Am. Statist. Assoc.*, 95, 449, 99–108.
- Pistone G., Riccomagno E. and Wynn H.P. (2000) *Algebraic Statistics: Computational Commutative Algebra in Statistics*, Chapman&Hall, New York.
- Rapallo F. (2001) Algebraic Markov bases and MCMC for two-way contingency tables, submitted to Scand. J. Statist.
- Strawderman R.L. and Wells M.T. (1998) Approximately exact inference for the common odds ratio in several 2×2 tables, *J. Am. Statist. Assoc.*, 93, 444, 1294–1307.