Journal of Statistical Planning and Inference 139 (2009) 2407 – 2419

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A-dependence statistics for mutual and serial independence of categorical variables

M. Bilodeau[a,*], P. Lafaye de Micheaux[b]

[a]*Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, Succursale Centre-ville, Montréal, Canada H3C 3J7*
[b]*Laboratoire Jean Kuntzmann, Department of Statistics, Sagag team, Grenoble University BSHM, 1251 avenue centrale BP 47, 38040 GRENOBLE Cedex 09, France*

ARTICLE INFO

Article history:
Received 10 April 2007
Received in revised form
12 November 2008
Accepted 14 November 2008
Available online 25 November 2008

MSC:
62H15
62G09
60F05

Keywords:
Categorical variables
Chi-square tests
Mutual independence
Serial independence

ABSTRACT

The Möbius transformation of probability cells in a multi-way contingency table is used to partition the Pearson chi-square test of mutual independence into *A*-dependence statistics. A similar partition is proposed for a universal and consistent test of serial independence in a stationary sequence of a categorical variable. The partition proposed can be adapted whether using estimated or theoretical marginal probabilities. With the aim of detecting a dependence of high order in a long sequence, *A*-dependence terms of the partition measuring increasing lagged dependences can be combined in a Box–Pierce type test of serial independence. A real data analysis of a nucleotides sequence using the Box–Pierce type test is provided.

1. Introduction

When the Pearson chi-square test of mutual independence is significant, it gives little indication to the way that the null hypothesis disagrees with the data. This paper uses the Möbius transformation to partition Pearson chi-square into components measuring what was termed by Deheuvels (1979) as the *A*-dependence. The *A*-dependence statistics of the partition are mutually independent and asymptotically distributed as chi-square. They are associated with additive interactions. The dependence structure of a multi-way contingency table can also be investigated by multiplicative interactions in a log-linear model, see Santner and Duffy (1989).

The main contribution of this paper relates to the use of *A*-dependence statistics to test for serial independence of a stationary sequence of a categorical variable. The adaptation of log-linear models to stationary sequences seems unwieldy, whereas the *A*-dependence approach extends very naturally. A universal and consistent test of serial independence is constructed using *A*-dependence statistics. This test is also a chi-square test and it can be partitioned into *A*-dependence statistics which are asymptotically independent and distributed as chi-square. Similar properties are obtained for the chi-square test to be used with theoretical marginal probabilities. A chi-square test of the Box–Pierce type is also proposed to detect large lagged dependences with an application to a nucleotides sequence. The *A*-dependence statistics have a closed form. They can be evaluated by the most common statistical softwares which provide the Pearson chi-square test of independence for multi-way contingency tables.

* Corresponding author.
E-mail addresses: bilodeau@dms.umontreal.ca (M. Bilodeau), Pierre.Lafaye-de-Micheaux@upmf-grenoble.fr (P. Lafaye de Micheaux).

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.11.006

The Möbius transformation sheds new light on the partition of the mutual independence chi-square test in Lancaster (1951) and the serial independence chi-square test in Good (1953).

Tests of serial independence for a stationary quantitative sequence are numerous. Delgado (1996) extends the work of Blum et al. (1961) in the higher dimensional case to obtain a test of serial independence in the continuous case. Genest and Rémillard (2004) propose another test based on the Möbius transformation with a simpler covariance function than that of Delgado (1996). Both of these tests are distribution free in the continuous case. They could also be used in the discrete case with the use of the bootstrap distribution even though they are no longer distribution free, see Beran et al. (2007). Hong (1998) obtains a test for pairwise serial independence applicable in discrete or continuous models with a standard normal approximation for large lagged dependences. All of the above tests are based on the empirical distribution function and should not be applied to categorical time series, especially those of a nominal (or non-ordinal) nature. Nominal categories are mere labels and their quantification can yield different orderings of the labels. The tests based on ranks, which are invariant to monotone transformations of the data, are not invariant to permutation of the labels. In a genetic application, the assignment of nucleotides (A = 1, G = 2, C = 3, T = 4) or (T = 1, A = 2, G = 3, C = 4) would give different values of the test statistic.

2. Chi-square tests

A d-dimensional categorical random vector is denoted $X = (X^{(1)}, \dots, X^{(d)})$. As for the vector $t = (t_1, \dots, t_d)$, it represents a d-dimensional cell in a multi-way contingency table. The joint and marginal cell probabilities are

$$p(t) = \mathrm{pr}\,(X^{(1)} = t_1, \dots, X^{(d)} = t_d),$$

$$p^{(k)}(t_k) = \mathrm{pr}\,(X^{(k)} = t_k), \quad k = 1, \dots, d.$$

The cell counts in the d-dimensional table are based on n independent realizations

$$X_i = (X_i^{(1)}, \dots, X_i^{(d)}), \quad i = 1, \dots, n$$

from the distribution of X. The sampling distribution of the counts is a single multinomial experiment, where only the total n is fixed. The joint and marginal empirical cell probabilities are

$$p_n(t) = \frac{1}{n} \sum_{i=1}^{n} \prod_{k=1}^{d} \mathbb{I}\{X_i^{(k)} = t_k\},$$

$$p_n^{(k)}(t_k) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i^{(k)} = t_k\}.$$

Pearson chi-square

$$X^2_{1,\dots,d} = \sum_t \frac{n[p_n(t) - \prod_{k=1}^d p_n^{(k)}(t_k)]^2}{\prod_{k=1}^d p_n^{(k)}(t_k)} \tag{1}$$

is often used to test the hypothesis of mutual independence among d categorical variables, where $d \geqslant 2$. Assuming there are I_k categories associated with the variable $X^{(k)}$, then the number of cells of the d-dimensional table is $\prod_{k=1}^d I_k$. The asymptotic null distribution of this test is chi-square with

$$f = \prod_{k=1}^{d} I_k - 1 - \sum_{k=1}^{d} (I_k - 1)$$

degrees of freedom.

Sometimes, the chi-square test is used with theoretical marginal probabilities in which case it is defined as

$$\check{X}^2_{1,\dots,d} = \sum_t \frac{n[p_n(t) - \prod_{k=1}^d p^{(k)}(t_k)]^2}{\prod_{k=1}^d p^{(k)}(t_k)} \tag{2}$$

and it has $f = \prod_{k=1}^d I_k - 1$ degrees of freedom.