Computational Statistics and Data Analysis 99 (2016) 25–37

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Testing hypothesis for a simple ordering in incomplete contingency tables

Hui-Qiong Li [a,*], Guo-Liang Tian [b], Xue-Jun Jiang [c], Nian-Sheng Tang [a]

[a] Department of Statistics, Yunnan University, Kunming, Yunnan 650091, PR China

[b] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China

[c] Department of Mathematics, South University of Science and Technology of China, Shenzhen, Guangdong 518055, PR China

ARTICLE INFO

Article history:
Received 26 May 2015
Received in revised form 21 October 2015
Accepted 6 January 2016
Available online 14 January 2016

Keywords:
EM algorithm
Incomplete contingency tables
Ordering
Statistical inferences

ABSTRACT

A test for ordered categorical variables is of considerable importance, because they are frequently encountered in biomedical studies. This paper introduces a simple ordering test approach for the two-way $r \times c$ contingency tables with incomplete counts by developing six test statistics, i.e., the likelihood ratio test statistic, score test statistic, global score test statistic, Hausman–Wald test statistic, Wald test statistic and distance-based test statistic. Bootstrap resampling methods are also presented. The performance of the proposed tests is evaluated with respect to their empirical type I error rates and empirical powers. The results show that the likelihood ratio test statistic based on the bootstrap resampling methods perform satisfactorily for small to large sample sizes. A real example from a wheeze study in six cities is used to illustrate the proposed methodologies.

1. Introduction

In biomedical studies, especially in clinical trials, incomplete ordered categorical data arise quite frequently. Incomplete ordered data can occur for various reasons. For example, Ware et al. (1984), Lipsitz and Fitzmaurice (1996) and Tang et al. (2007a,b) considered a wheeze study from six cities. The data are summarized in Table 1. The columns of Table 1 correspond to the wheezing status ($Y = 1$: no wheeze; $Y = 2$: wheeze with cold; $Y = 3$: wheeze apart from cold) of a child at age 10. The rows represent the smoking status of the child's mother ($X = 1$: none; $X = 2$: moderate; $X = 3$: heavy) during that time. Note that for some individuals the maternal smoking variable is missing whereas for others the child's wheezing status is missing. Thus, the resultant data include two parts: the complete observations and the incomplete observations. Following the arguments of Ware et al. (1984), Lipsitz and Fitzmaurice (1996) and Tang et al. (2007a,b), we assume that the missing mechanism is *missing at random* (MAR; Rubin, 1976).

Under missing at random, one is often interested in investigating whether there is positive association between two ordered variables; that is, as the maternal smoking increases, whether a more effects on respiratory illness in children tends to occur. For this purpose, we can consider testing ordering alternatives. Statistical inference for ordering alternatives has been widely studied in the literature. For example, Robertson and Wright (1981) considered the likelihood ratio statistic to test equality of two multinomial distributions against the simple stochastic ordering, and presented explicit expressions of the *maximum likelihood estimates* (MLEs). Lucas and Wright (1991) considered a similar problem and derived the asymptotic distributions for the simple stochastic ordering in discrete univariate and bivariate cases. For more than two multinomial

* Corresponding author.
E-mail address: lihuiqiong@ynu.edu.cn (H.-Q. Li).

http://dx.doi.org/10.1016/j.csda.2016.01.003
0167-9473/

Table 1
Maternal smoking cross-classified by child's wheezing status.

Maternal smoking	Child's wheezing status			Missing on Y
	No wheeze ($Y = 1$)	Wheezing with cold ($Y = 2$)	Wheezing apart from cold ($Y = 3$)	
None ($X = 1$)	287	39	38	279
Moderate ($X = 2$)	18	6	4	27
Heavy ($X = 3$)	91	22	23	201
Missing on X	59	18	26	

Table 2
Observed counts and cell probabilities for an $r \times c$ table with incomplete ordinal observations.

	$Y = 1$	$\cdots$	$Y = c$	Subtotal	Missing on Y
$X = 1$	$n_{11}(\pi_{11})$	$\cdots$	$n_{1c}(\pi_{1c})$	$n_{1+}(\pi_{1+})$	m_{1x}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$X = r$	$n_{r1}(\pi_{r1})$	$\cdots$	$n_{rc}(\pi_{rc})$	$n_{r+}(\pi_{r+})$	m_{rx}
Subtotal	$n_{+1}(\pi_{+1})$	$\cdots$	$n_{+c}(\pi_{+c})$	n	m_x
Missing on X	m_{y1}	$\cdots$	m_{yc}	m_y	

populations, Wang (1996) developed a test for the equality of distributions against simple stochastic ordering in several populations. In that paper, the limit distribution of likelihood ratio test statistic was characterized by minimization problems and has no closed form. Dardanoni and Forcina (1998) considered the same hypothesis test problem and gave the asymptotic distribution of the likelihood ratio statistic. Their results were cited by Silvapulle and Sen (2005). Feng and Wang (2007) also considered the same hypothesis test and obtained the asymptotic distribution of likelihood ratio test statistic by completely different approach from that in Dardanoni and Forcina (1998). In order to obtain the null asymptotic distribution of the likelihood ratio statistic, Feng and Wang (2007) transformed the simple stochastic ordering constraint into a polyhedral cone constraint, and the restricted MLEs are characterized by maximization problem with a concave objective function and a series of linear inequality constraints. Thus, the desired null asymptotic distribution was obtained by limit theory of the optimization. Klingenberg et al. (2009) presented an alternative bootstrap approach to test marginal homogeneity against stochastic ordering in two-sample multivariate ordinal data. Davidov and Peddada (2011) developed a general methodology for testing the multivariate stochastic ordering.

However, none of the aforementioned works has been generalized to incomplete $r \times c$ tables with correlated ordinal data. Besides, the score statistic and the Wald statistic are not yet discussed in above-mentioned references. Note that the likelihood ratio, score and Wald test statistics are asymptotically equivalent. However, when parameter space is constrained, the likelihood ratio, score, Wald test statistics for a simple ordering restriction with incomplete data have not yet been considered in the literature. Hence, it is the aim of this article to consider the likelihood ratio test statistic, the score test statistic, the global score test statistic, the Hausman–Wald test statistic, the Wald test statistic and distance-based test statistic and to present bootstrap resampling methods to test a simple ordering restriction in incomplete $r \times c$ tables.

The rest of this paper is organized as follows. In Section 2, we transform the problem of testing simple ordering into a polyhedral cone constrained problem. Section 3 presents the likelihood ratio test statistic, the score test statistic, the global score test statistic, the Hausman–Wald test statistic, the Wald test statistic and distance-based test statistic and bootstrap resampling methods for testing a simple ordering in incomplete $r \times c$ tables. Simulation studies are conducted to investigate the performance of various methods in Section 4. A real example from the aforementioned wheeze study in six cities is used to illustrate the proposed methodologies in Section 5. Some concluding remarks are given in Section 6.

2. The formulation of the simple ordering test

Let X and Y be two correlated ordinal variables with the joint distribution $\pi_{ij} = \Pr(X = i, Y = j)$ for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. The observed counts and the corresponding cell probabilities for $n = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$ complete observations and $m_x + m_y = \sum_{i=1}^{r} m_{ix} + \sum_{j=1}^{c} m_{yj}$ partially incomplete observations are listed in Table 2. In this paper, we use

$$\mathbb{T}_n = \left\{(x_1, \ldots, x_n)^\top : x_k \geqslant 0,\ k = 1, \ldots, n,\ \sum_{k=1}^{n} x_k = 1\right\}$$

to denote the n-dimensional hyperplane.

Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_1^\top, \ldots, \boldsymbol{\pi}_r^\top)^\top \in \mathbb{T}_{rc}$ be the unknown parameter vector, where $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ic})^\top$. Assume that n_{ij} denotes the number of subjects who go through both variables with $X = i$ $(1 \leqslant i \leqslant r)$ and $Y = j$ $(1 \leqslant j \leqslant c)$, m_{ix} is the number of subjects who go through only variable X with $X = i$, and m_{yj} is the number of subjects who go through only variable Y with