Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT

Journal of Statistical Planning and Inference 127 (2005) 1–21

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Exact null distributions of distribution-free quadratic t-sample statistics

A. Di Bucchianico, M.A. van de Wiel*

Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Received 18 September 2002; accepted 7 September 2003

Abstract

We present new algorithms for computing the exact distributions and p-values of quadratic t-sample distribution-free statistics of Kruskal–Wallis type. These algorithms are presented in terms of generating functions. We show that our algorithm also works for cases with ties and that it is much faster than existing algorithms. Moreover, we show how to use the results for the Kruskal–Wallis type statistics to compute the exact null distribution of the Chacko–Shorack statistic.

MSC: 62E15; 62E30; 62G10; 62J10; 62Q05; 90C20

Keywords: Kruskal–Wallis statistic; Exact p-values; Generating functions; Quadratic optimization; Ties; Branch-and-bound algorithm

1. Introduction

Computing exact null distributions and exact p-values of quadratic t-sample distribution-free statistics (also known as 'k-sample rank statistics') is in general a very time-consuming task. Direct enumeration methods usually only work for very small sample sizes. For the Kruskal–Wallis statistic some exact procedures and tables are available in Iman et al. (1975), while Wallace (1959), Iman and Davenport (1976) and Robinson (1980) contain approximations to the exact distribution that improve upon the standard χ^2-approximations. These approximations are often worse for unbalanced cases and cases with large ties than for balanced and untied cases. Exact

* Corresponding author. Tel.: +31-40-2472499; fax: +31-40-2465995.

E-mail addresses: a.d.bucchianico@tue.nl (A.D. Bucchianico), markvdw@win.tue.nl (M.A. van de Wiel).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2003.09.024

null distributions when ties are present have been calculated for small sample sizes in Klotz and Teng (1977). The network algorithms developed by Mehta and co-workers (see, e.g. Mehta and Patel 1983; Mehta et al., 1988) have proven to be very efficient for computing exact distributions of many nonparametric statistics. They are implemented in the dedicated software package *StatXact*. However, even this fine package has difficulties computing the exact distribution of the Kruskal–Wallis statistics for moderate sample sizes (see Section 8 for details). The goal of this paper is to present new, efficient algorithms for computing exact null distributions and tail probabilities of 'Kruskal–Wallis type' statistics with any rank score function. Our algorithms also work for cases with ties. They turn out to be significantly faster than the network algorithm implemented in *StatXact 4*. Our algorithms can also be used for computing the exact null distribution of the Chacko–Shorack statistic as introduced in Chacko (1963) and Shorack (1967).

Our algorithms belong to the class of recursive polynomial multiplication algorithms as introduced in Hirji and Johnson (1996) in the context of $2 \times K$ contingency tables. Similar to this case, our algorithms and the network algorithms are essentially different representations of the same idea (cf. Hirji and Johnson, 1996, Theorem 2). However, a distinctive feature of our approach is a branch-and-bound technique tailored to the quadratic nature of Kruskal–Wallis type. This brand-and-bound technique involves subtle combinatorial optimization issues involving extreme points of special convex sets. The drastic effect of the branch-and-bound algorithm on computing times is shown in Section 8.

This paper is organized as follows. In Section 2 we introduce the class of Kruskal–Wallis type statistics that we study. In Section 3 we introduce our new recursive way of computing the generating function of the statistics at hand. This recursion is enhanced by using exchangeability in Section 4. A branch-and-bound algorithm for highly efficient calculation of p-values is discussed in detail in Section 5. In Section 6 we summarize the main algorithm and compare approximations with exact results for cases with and without ties. We apply our algorithms to the Chacko–Shorack statistic in Section 7. We discuss performance of our algorithms in Section 8. Section 9 contains information on software that we made available publicly.

2. Rank tests for one-way classification

Suppose we deal with t independent samples, each of size n_j, $j = 1, \ldots, t$. The total number of observations is $N = \sum_{j=1}^{t} n_j$. We assume that all observations X_{ij} ($i = 1, \ldots, n_j$) belonging to treatment j are mutually independent and have unknown distribution function $F(x - \gamma_j)$, $j = 1, \ldots, t$. We assume, without loss of generality, that $\gamma_1 = 0$. We wish to test the null hypothesis

$$\mathrm{H}_0 : \gamma_1 = \cdots = \gamma_t = 0, \qquad (1)$$

against the alternative hypothesis H_1: there exists a j, $j = 2, \ldots, t$, for which $\gamma_j \neq 0$. Before defining the test statistic, we define the sum of scores assigned to