Available online at www.sciencedirect.com

ScienceDirect

Statistics and Probability Letters 78 (2008) 1328–1338

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Multiple comparisons of several heteroscedastic multivariate populations

Yoshihide Kakizawa*

Faculty of Economics, Hokkaido University, Nishi 7, Kita 9, Kita-ku, Sapporo 060–0809, Japan

Received 19 June 2007; accepted 11 December 2007
Available online 28 December 2007

Abstract

This paper presents several statistics appearing in multiple comparisons of heteroscedastic multivariate populations. Due to the very slow convergence of these statistics to their limiting distributions, the large sample Bonferroni or DL-based procedures reveal poor coverage probabilities even in the normal case. Thus, the second-order asymptotic expansions with estimated cumulants are applied to improve their coverage probabilities. A large simulation study illustrates the performance of the second-order corrected procedures.

MSC: primary 62J15; 62H10; secondary 62E20; 62J10

1. Introduction

Given $q(\geq 2)$ levels, let $\boldsymbol{X}_i^{(a)} = (X_{1i}^{(a)}, \ldots, X_{pi}^{(a)})'$ be the ith observation on the ath level, and assume the linear model (one-way layout model)

$$\boldsymbol{X}_i^{(a)} = \boldsymbol{\theta}^{(a)} + \boldsymbol{U}_i^{(a)}, \quad a = 1, \ldots, q;\ i = 1, \ldots, N_a, \tag{1}$$

where $\boldsymbol{U}_i^{(a)}$'s are (unobservable) independent $p \times 1$ random vectors with mean zero vector and positive definite (unknown) covariance matrix $\Sigma^{(a)} = (\sigma_{jk}^{(a)})_{j,k=1,\ldots,p}$. The total number of such vectors is $\sum_{a=1}^{q} N_a = N$ (say). In the model (1), the least squares estimates of the $\boldsymbol{\theta}^{(a)}$'s are given by the sample mean vector $\overline{\boldsymbol{X}}^{(a)} = N_a^{-1} \sum_{i=1}^{N_a} \boldsymbol{X}_i^{(a)} = (\overline{X}_j^{(a)})_{j=1,\ldots,p}$, $a = 1, \ldots, q$. Let $S_X^{(a)} = (N_a - 1)^{-1} \sum_{i=1}^{N_a} (\boldsymbol{X}_i^{(a)} - \overline{\boldsymbol{X}}^{(a)})(\boldsymbol{X}_i^{(a)} - \overline{\boldsymbol{X}}^{(a)})'$, $a = 1, \ldots, q$, denote the sample covariance matrix for the ath level based on N_a observations. We will use the same notation for the (unobservable) sample $\boldsymbol{U}_i^{(a)}$, $i = 1, \ldots, N_a$.

We consider multiple comparisons among mean vectors of multivariate populations, especially, comparisons with a control (the qth level is now regarded as a control) and all pairwise comparisons. That is, we aim at constructing simultaneous confidence intervals of (I) $\boldsymbol{\ell}'(\boldsymbol{\theta}^{(a)} - \boldsymbol{\theta}^{(q)})$, $\boldsymbol{\ell} \in \mathbf{R}^p - \{\mathbf{0}\}$, $a = 1, \ldots, q-1$, and (II) $\boldsymbol{\ell}'(\boldsymbol{\theta}^{(a)} - \boldsymbol{\theta}^{(b)})$,

* Fax: +81 011 706 4947.
E-mail address: kakizawa@econ.hokudai.ac.jp.

0167-7152/$ - see front matter
doi:10.1016/j.spl.2007.12.017

$\boldsymbol{\ell} \in \mathbf{R}^p - \{\mathbf{0}\}$, $a, b = 1, \ldots, q$; $a < b$. This problem is a nonnormal and heteroscedastic extension of Roy and Bose (1953, (4.3.1)) and Siotani (1960, (12) and (13)), who originally considered the normal and homogeneous covariance matrix case.

Throughout this paper, $G_\nu(\cdot)$ denotes the distribution function of the central chi-square distribution with ν degrees of freedom; χ^2_ν, whose density function and upper $100\alpha\%$ point are $g_\nu(\cdot)$ and $\chi^2_{\nu,\alpha}$, respectively.

2. Background

2.1. Bonferroni-based solution

We define the correlated Behrens–Fisher statistics

$$T^2_{BF,ab} = (\overline{\boldsymbol{U}}^{(a)} - \overline{\boldsymbol{U}}^{(b)})' \left(\frac{S_U^{(a)}}{N_a} + \frac{S_U^{(b)}}{N_b} \right)^{-1} (\overline{\boldsymbol{U}}^{(a)} - \overline{\boldsymbol{U}}^{(b)}), \quad a, b = 1, \ldots, q;\ a < b, \tag{2}$$

by taking account of the multivariate Behrens–Fisher problem (e.g. Siotani et al. (1985, page 212)). In line with Siotani (1960, (12) and (13)) under the homogeneous covariance matrix case, one may have

$$\begin{aligned} \Pr[\mathrm{SCI_I}(A)] &= \Pr\left[\boldsymbol{\ell}'(\boldsymbol{\theta}^{(a)} - \boldsymbol{\theta}^{(q)}) \in \boldsymbol{\ell}'(\overline{\boldsymbol{X}}^{(a)} - \overline{\boldsymbol{X}}^{(q)}) \pm A \left\{ \boldsymbol{\ell}' \left(\frac{S_X^{(a)}}{N_a} + \frac{S_X^{(q)}}{N_q} \right) \boldsymbol{\ell} \right\}^{1/2} \right. \\ &\qquad \left. \text{for all } \boldsymbol{\ell} \in \mathbf{R}^p - \{\mathbf{0}\}, a = 1, \ldots, q-1 \right] \\ &= \Pr(T^2_{BF,\max,\mathrm{I}} \le A^2) \end{aligned} \tag{3.I}$$

for comparisons with a control and

$$\begin{aligned} \Pr[\mathrm{SCI_{II}}(A)] &= \Pr\left[\boldsymbol{\ell}'(\boldsymbol{\theta}^{(a)} - \boldsymbol{\theta}^{(b)}) \in \boldsymbol{\ell}'(\overline{\boldsymbol{X}}^{(a)} - \overline{\boldsymbol{X}}^{(b)}) \pm A \left\{ \boldsymbol{\ell}' \left(\frac{S_X^{(a)}}{N_a} + \frac{S_X^{(b)}}{N_b} \right) \boldsymbol{\ell} \right\}^{1/2} \right. \\ &\qquad \left. \text{for all } \boldsymbol{\ell} \in \mathbf{R}^p - \{\mathbf{0}\}, a, b = 1, \ldots, q; a < b \right] \\ &= \Pr(T^2_{BF,\max,\mathrm{II}} \le A^2) \end{aligned} \tag{3.II}$$

for all pairwise comparisons, respectively. But, the distributions of $T^2_{BF,\max,\mathrm{I}} = \max_{a=1,\ldots,q-1}(T^2_{BF,aq})$ and $T^2_{BF,\max,\mathrm{II}} = \max_{1\le a<b\le q}(T^2_{BF,ab})$, hence, tables of their upper $100\alpha\%$ points $T^2_{BF,\max,\mathrm{I}}(\alpha)$ and $T^2_{BF,\max,\mathrm{II}}(\alpha)$, are not available even in the case of q multivariate normal populations.

The usual Bonferroni procedure uses the upper bounds $T^2_{BF,\mathrm{Bon,I}}(\alpha)$ and $T^2_{BF,\mathrm{Bon,II}}(\alpha)$ for the percentiles $T^2_{BF,\max,\mathrm{I}}(\alpha)$ and $T^2_{BF,\max,\mathrm{II}}(\alpha)$ by equating the right-hand sides of the following Bonferroni inequalities to $1-\alpha$;

$$\Pr(T^2_{BF,\max,\mathrm{I}} \le x) \ge 1 - \sum_{a=1}^{q-1} \Pr(T^2_{BF,aq} > x) \quad \text{and} \quad \Pr(T^2_{BF,\max,\mathrm{II}} \le x) \ge 1 - \sum_{1\le a<b\le q} \Pr(T^2_{BF,ab} > x).$$

However, the exact computation of each distribution function $\Pr(T^2_{BF,ab} \le x)$ is also complicated even for the multivariate normal population (see Nel et al. (1990)) and it is hopeless for the multivariate nonnormal populations. Fortunately, if the N_a's are large, for possibly multivariate nonnormal populations, the central limit theorem and Slutsky's theorem state that the Behrens–Fisher statistic $T^2_{BF,ab}$ is asymptotically chi-squared with p degrees of freedom. In Section 3, we use the asymptotic expansion formula (see Kakizawa and Iwashita (in press)) to see the effect of the nonnormality upon an approximation $T^2_{BF,\mathrm{Bon,I}}(\alpha) \approx \chi^2_{p,\alpha_\mathrm{I}}$ and $T^2_{BF,\mathrm{Bon,II}}(\alpha) \approx \chi^2_{p,\alpha_\mathrm{II}}$, where $\alpha_\mathrm{I} = \alpha/(q-1)$ and $\alpha_\mathrm{II} = \alpha/\{q(q-1)/2\}$.