行政院國家科學委員會專題研究計劃成果報告

變異數不等時雙階段常態分配與平均的多重比較程序

Multiple Comparison Procedures with the Average for Normal Means Under Heteroscedasticity by Two Stage Sampling

計劃編號: NSC89-2118-M-032-005
執行期限: 88 年 8 月 1 日 至 89 年 7 月 31 日

主持人: 吳淑妃 淡江大學統計系
研究助理：鄭茂益

一. 中文摘要

在這篇報告裡, 我們提出了雙階段抽樣程序來解決當變異數不等時常態分配平均數與平均的多重比較程序的問題. 我們的多重比較程序包括三部份來討論, 一為上界同時信賴區間, 另一為下界同時信賴區間, 最後為雙邊同時信賴區間. 這些程序應可廣氾的應用在選出比平均好的常態分配平均數上, 或是把 k 個常態分配平均數分成比平均好, 比平均差, 及和平均沒什麼差異的三個群體. 我們用 Bonferroni 不等式, 提出較保守的臨介值, 且我們用 Monte-Carlo 方法, 找出模擬臨介值的表供使用者使用.

關鍵辭:

子集選擇程序, 同時信賴區間, Bonferroni 不等式, Monte-Carlo 方法

ABSTRACT

In this article, multiple comparison procedures with the average for normal means under heterocedasity are under investigation by two stage sampling technique. One-sided and two-sided confidence intervals are proposed. These intervals will have broad applicability in identifying a subset which includes all no-worse-than-the-average treatments in experimental designs and/or in identifying groups of treatments with smaller than the average, larger than the average and not much difference from the average means in various fields. An upper limit of critical values are obtained using Bonferroni inequality. But the approximate values are shown to be too conservative compared with the simulated critical values by Monte-Carlo method. Therefore, simulated critical values should be used in our multiple comparison procedures.

Key Words and Phrases:

subset selection, simultaneous confidence interval, Bonferroni inequality, Monte Carlo technique.

1. 計劃緣由與目的

Wu and Chen (1998) has proposed multiple comparison procedures with the average for normal means when variances are equal. When variances are unequal, we investigate a two-stage sampling procedure developed by Dudewicz and Dalal (1975) for multiple comparison problem in this paper; which was a generalized Stein-type procedure (Stein(1945)).

Consider k independent popularions

$X_{i1},\ldots,X_{in_0}$, where observations from popularion π_i are independent and normally distributed by N(θ_i,σ_i^2), where the normal means $\theta_1,\ldots,\theta_k$ are all unknown and unequal and the variances $\sigma^2{}_1,\ldots\sigma_k^2$ are also unknown and unequal. A two-stage multiple comparison procedures with the average is proposed as follows:

Take an initial sample $X_{i1},\ldots,X_{in_0}$ of size n_0 from π_i, Let S_i^2 be the usual unbiased estimate of σ_i^2 and define

$n_i = \max\{n_0+1,[\frac{S_i^2}{c^2}]+1\}$, where the value of c is an arbitrary positive constant to be chosen to control the width of the confidence intervals for $\theta_i-\overline{\theta}$ and $[x]$ denotes the largest integer smaller than or equal to x. Then take additional n_i-n_0 observations $X_{i,n_0+1},\ldots,X_{in_i}$ from π_i.

For each i, set the coefficients $a_{i1},\ldots,a_{in_0},\ldots,a_{in_i}$ so that

$$a_{i1}=\ldots=a_{in_0}=\frac{1-(n_i-n_0)b_i}{n_0}=a_i$$

$$a_{i,n_0+1}=\ldots=a_{in_i}=\frac{1}{n_i}\{1+\sqrt{\frac{n_0(n_ic^2-S_i^2)}{(n_i-n_0)S_i^2}}\}=b_i$$

and finally compute the weighted mean

$\widetilde{X}_i=a_i\sum_{j=1}^{n_0}X_{ij}+b_i\sum_{j=n+1}^{n_i}X_{ij}$ and the r.v.'s $Y_i=(\widetilde{X}_i-\theta_i)/c$ have i.i.d. student's t distribution with $\nu=n_0-1$ df.

三. 結果與討論

Let $\overline{\widetilde{X}}=\sum_{i=1}^{k}\widetilde{X}_i/k$. Then the one-sided and two-sided confidence intervals for $\theta_i-\overline{\theta}$ are given in the following Theorem by Bonferroni inequality.

Theorem For a given $0<P^*<1$,

(a) $P(\theta_i-\overline{\theta}\in\widetilde{U}_i,i=1,\ldots k)\geq P^*$, where

$\widetilde{U}_i=(-\infty,\widetilde{X}_i-\overline{\widetilde{X}}+ch_U)$, if h_U is the solution of the given equation

$$\int_{-\infty}^{\infty}F(t+d)^{k-1}f(t)dt=\frac{P^*+k-1}{k},$$

$d=(k/(k-1))h_U$. Thus $\widetilde{U}_i=(-\infty,\widetilde{X}_i-\overline{\widetilde{X}}+ch_U)$ is a set of upper confidence intervals for $\theta_i-\overline{\theta}$ with confience coefficient P^*.

(b) $P(\theta_i-\overline{\theta}\in\widetilde{L}_i,i=1,\ldots k)\geq P^*$, where

$\widetilde{L}_i=(\widetilde{X}_i-\overline{\widetilde{X}}-ch_L,\infty)$, if h_L is the solution of the given equation

$$\int_{-\infty}^{\infty}F(t+d)^{k-1}f(t)dt=\frac{P^*+k-1}{k},$$

$d=(k/(k-1))h_L$. Thus $\widetilde{L}_i=(\widetilde{X}_i-\overline{\widetilde{X}}-ch_L,\infty)$ is a set of lower confidence intervals for $\theta_i-\overline{\theta}$ with confience coefficient P^*.

(c) $P(\theta_i-\overline{\theta}\in\widetilde{C}_i,i=1,\ldots k)\geq P^*$, where

$\widetilde{C}_i = (\widetilde{X}_i - \overline{\overline{X}} - ch_t, \widetilde{X}_i - \overline{\overline{X}} + ch_t)$, if h_t is the solution of the given equation

$$\int_{-\infty}^{\infty} (F(t+d)^{k-1} - F(t-d)^{k-1}) f(t) dt = \frac{P^* + k - 1}{k}$$

, $d = (k/(k-1))h_t$. Thus $\widetilde{C}_i = (\widetilde{X}_i - \overline{\overline{X}} - ch_t, \widetilde{X}_i - \overline{\overline{X}} + ch_t)$ is a set of two-sided confidence intervals for $\theta_i - \overline{\theta}$ with confidence coefficient P^*.

Since it is difficult to find the exact joint sampling distribution of the singular k-variate statistic of $T_1, \ldots, T_k$, where $T_i = \frac{k-1}{k} Y_i - \sum_{j \neq i}^{k} Y_j / k, i = 1, \ldots, k$. Therefore, the Monte-Carlo simulation is used here to obtain an approximate sampling distribution of the maximum order statistic of $T_{[k]}$. In order to have the probability of inclusion of $\theta_i - \overline{\theta}$ in $\widetilde{U}_i = (-\infty, \widetilde{X}_i - \overline{\overline{X}} + ch_U)$ or $\widetilde{L}_i = (\widetilde{X}_i - \overline{\overline{X}} - ch_L, \infty)$ being at least P^*, the value of h_U or h_L is given by the P^* th percentile of the approximate sampling distribution of $T_{[k]}$, $P(T_{[k]} < h_U) = P^*$.

Likewise, the $(1-\alpha)100\%$ SCI of $\theta_i - \overline{\theta}$ is given by $\widetilde{C}_i = (\widetilde{X}_i - \overline{\overline{X}} - ch_t, \widetilde{X}_i - \overline{\overline{X}} + ch_t)$, where h_t is given by the upper $\alpha/2$ percentile of the approximate sampling distribution of $T_{[k]}$, $P(T_{[k]} < h_t) = 1 - \alpha/2$.

In the case of equal sample size $n_i = n$, the simulated value of h_U or h_L for subset selection when $P^* = .95$ and the simulated value of h_t for simultaneous confidence interval when $1-\alpha = .90$ are given in the following table. The upper entry is the average of the simulated percentiles, each based on 3000 runs, and the standard error (s.e.) is reported at the lower entry. To do this we only need to generate some random numbers from a student's t distribution with $\nu = n_0 - 1$ degrees of freedom and calculate T_i's, then order these T_i's, and select the maximum value $T_{[k]}$ at each run. After *N* independent runs, the P^* th percentile of the distribution of $T_{[k]}$ can be approximated by selecting the NP^* biggest number out of the ranking list of the *N* largest values of $T_{[k]}$.

四. 計劃結果與自評

我已完成理論近似值的推導以及模擬的研究，所有的程式皆已建立完成. 故此研究已完成百分之百. 此研究結果可有效的解決當變異數不等時常態分配與平均的多重比較程序的問題，相信很快就可在學術學刊發表.

五. 參考資料

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with know variance. *Annals of Mathematical Statistics*, 25, 16-39.

Bishop, T. A. and Dudewicz, E. J. (1978). Exact Analysis of Variance with Unequal Variances: Test Procedures and Tables. *Technometrics*, 20, 419-430.

Chen, H. J. and Dudewicz, E. J. (1976). Procedures for fixed-width interval estimation of the largest normal mean. *Journal of American Statistical Association*, 71, 752-756.

Chen, H. J. and Lam, K. (1989). Single-stage interval estimation of the largest normal mean under heteroscedasticity. *Communications in Statistics - Theory and Methods*, 18(10), 3703-3718.

Dudewicz, E. J. and Dalal, S. R. (1975). Allocation of observations in ranking and selection with unequal variances. *Sankhya, Ser. B.*, 37, 28-78.

Gupta, S. S. (1956). On a decision rule for ranking means. Inst. Statist. Mimeo. Ser. No. 150, University of North Carolina, Chapel Hill, NC.

Gupta, S. S. (1965). On some multiple decision rules. *Technometrics*, 7, 225-245.

Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedure: Theory and Methodology of Selecting and Ranking Populations.* Wiley, New York.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures.* Wiley, New York.

Lam, K. (1992). Subset selection of normal populations under heteroscedasticity. The Frontiers of Modern Statistical Inference Procedures, II} (Eds. Bofinger, E., Dudewicz, E. J., Lewis, G. J., and Mengersen, K.), 307-344, American Sciences Press, Inc., Columbus, Ohio.

Marsaglia, G., Zaman, A., and Tsang, W. W. (1990). Toward a universal random number generator. *Statistics and Probability Letters*, 8, 35-39.

Stein, C. M. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16, 243-58.

Tong, Y. L. (1980). *Probability Inequalities in Multivariate Distributions.* Academic Press, New York.

Wu, S. F. and Chen, H. J. (1998). Multiple Comparisons with the Average for Normal Distributions. *American Journal of Mathematical Management and Sciences*, Vol 18, No. 1 & 2.