ELSEVIER

Journal of Statistical Planning and Inference 98 (2001) 239–258

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Generalized subset selection procedures under heteroscedasticity

Yi-Ping Chang[a], Wen-Tao Huang[b,*]

[a] *Department of Business Mathematics, Soochow University, Taipei, Taiwan, ROC*

[b] *Department of Management Sciences, Tamkang University, Tamsui-251, Taiwan, ROC*

Received 17 May 2000; received in revised form 15 September 2000; accepted 17 November 2000

Abstract

In this paper, we propose and study a generalized subset selection procedure for selecting the best population. Based on the concept of generalized subset selection procedure, some selection procedures for normal populations are proposed and studied. They are used, respectively, to select the best population (populations) with respect to the largest mean, the largest pth quantile and the largest signal-to-noise ratio. For the case of common unknown variance, the proposed generalized subset selection procedure for selecting the largest mean becomes exactly the same as that has been given in Hsu (in: T.J. Santner, A.C. Tamhane (Eds.), Design of Experiments: Ranking and Selection, Marcel Dekker, New York, 1984, pp. 179–198). A Monte Carlo study shows that the proposed generalized subset selection procedures behave satisfactorily. An illustration of a set of real data is also given.

MSC: 62F07

Keywords: Ranking and selection; Generalized subset selection procedure; Generalized probability of correct selection; Quantile; Signal-to-noise ratio

1. Introduction

Let $\pi_1,\ldots,\pi_k$ be k ($\geqslant 2$) normal populations where observations X_{ij} from π_i are independently distributed as $N(\mu_i,\sigma_i^2)$ ($j=1,\ldots,n_i$, $i=1,\ldots,k$). All means μ_i and variances σ_i^2 are unknown. When μ_i is the parameter of main interest, the problem of selecting the best population was studied in papers pioneered by Bechhofer (1954) using the indifference zone approach and by Gupta (1956) employing the subset

* Corresponding author. Tel.: +886-2-8631-3221; fax: +886-2-8631-3214.
E-mail address: 005697@mail.tku.edu.tw (W.-T. Huang).

0378-3758/01/$ - see front matter
PII: S0378-3758(00)00305-0

selection. A detailed discussion of these approaches and various related results can be found in Gupta and Panchapakesan (1979) among others.

Let θ_i be some function of μ_i and σ_i^2 $(i=1,\ldots,k)$. The population which is associated with the largest θ_i is called the best. And it is said that the selection criterion is the quantity θ_i. We are interested in selecting the best population.

Selection criterion in most results for selecting the best normal population so far is commonly focused either on some function of μ_i or some function of σ_i^2. However, in many practical situations, the pth quantile of population π_i is an important quantity to be considered. Also, the quantity of signal-to-noise ratio is an important index for practical application, in particular, in industry statistics (e.g. see Box, 1988). On the other hand, the quantity of signal-to-noise ratio is one of the most natural and also most important to be considered for the criterion that controls the mean (to be large) and simultaneously its associated variance (to be small).

In this paper, our selection criterion is considered, respectively, to be the mean, the pth quantile and the signal-to-noise ratio. When selection criterion is population mean, Hsu (1984) proposed a one-stage subset selection procedure for the case that σ_i^2 are all equal but unknown, and Gupta and Wong (1982) proposed approximated subset selections when σ_i^2 are unequal and unknown. Also, Santner and Tamhane (1984) proposed a two-stage procedure for selecting the population associated with the largest mean whose associated variance is under some control.

Based on the concept of generalized test variable (see, e.g., Tsui and Weerahandi, 1989; Zhou and Mathew, 1994; Weerahandi, 1995) and generalized confidence intervals of Weerahandi (1993, 1995), we propose generalized subset selection procedures in Section 2. Procedures for selecting the largest mean for both cases of equal variance and unequal variances are studied in Section 3. Finally, procedures for selecting the largest pth quantile and the largest signal-to-noise ratio are, respectively, studied in Sections 4 and 5. An illustration of a real data set is also given in Section 6.

2. Generalized subset selection procedure

Consider an observable random vector $\boldsymbol{X}_i$ with sample size n_i from population π_i, $i=1,\ldots,k$, with cumulative distribution function F_{ζ_i}, where $\boldsymbol{\zeta}_i=(\theta_i,\boldsymbol{\delta}_i)$ is a vector of unknown parameters, θ_i being the parameter of interest, and $\boldsymbol{\delta}_i$ is a vector of nuisance parameters.

Suppose that $\boldsymbol{x}_i$ is the observed value of $\boldsymbol{X}_i$. For convenience, let $\boldsymbol{X}=(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k)$, $\boldsymbol{x}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k)$, $\boldsymbol{\zeta}=(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_k)$, $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_k)$, $\boldsymbol{\delta}=(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_k)$, $\boldsymbol{n}=(n_1,\ldots,n_k)$, and $\theta_{[1]}\leqslant\theta_{[2]}\leqslant\cdots\leqslant\theta_{[k]}$ be the ordered θ's. Furthermore, let (k) denote some association such that the population $\pi_{(k)}$ is associated with parameter $\theta_{[k]}$. Usually, based on the observations of $\boldsymbol{x}$, the goal is to propose a rule for selecting a non-empty subset from the k populations and it is guaranteed that the probability that the best population being in the subset selected is at least a prefixed number P^* $(1/k<P^*<1)$. Let $P(\mathrm{CS}|R)$ denote the probability of a correct selection (CS) using the procedure R, the

P^*-condition is usually written by

$$\inf_{\boldsymbol{\Omega}} P(\mathrm{CS}|R) \geqslant P^*, \tag{2.1}$$

where $\boldsymbol{\Omega}$ is the parameter space of all k-tuples $(\zeta_1, \ldots, \zeta_k)$. A statistic $Y_i(\boldsymbol{X})$ is said to be a selection variable for θ_i if it is an appropriate estimator of θ_i. A subset selection procedure can be slightly modified to be as follows:

R: For observed $\boldsymbol{x}$ and some given constant d_i, retain π_i in the selected subset if and only if

$$\min_{j \neq i} \{ g(Y_i(\boldsymbol{x}), Y_j(\boldsymbol{x})) + d_i \widehat{\mathrm{Sd}}(g(Y_i(\boldsymbol{X}), Y_j(\boldsymbol{X}))) \} \geqslant 0,$$

where g is some real value function, $\widehat{\mathrm{Sd}}(g(Y_i(\boldsymbol{X}), Y_j(\boldsymbol{X}))$ is some appropriate estimate of the standard deviation of $g(Y_i(\boldsymbol{X}), Y_j(\boldsymbol{X}))$, and $d_i = d_i(\boldsymbol{n}, k, P^*)$ is some constant to be chosen that the P^*-condition is satisfied.

Following the ideas of generalized test variable (Tsui and Weerahandi, 1989; Weerahandi, 1995) and generalized confidence intervals (Weerahandi, 1993, 1995), by relaxing some of the requirements in the selection variables, we define the generalized selection variables as follows:

Definition 2.1. Let $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ be a function of $\boldsymbol{X}$ and possibly $\boldsymbol{x}$, $\boldsymbol{\zeta}$ as well. The random quantity $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ is said to be a generalized selection variable for θ_i if

(a) the value $Y_i(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{\zeta})$ is an appropriate estimate of θ_i,
(b) the joint distribution of $(Y_1(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}), \ldots, Y_k(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}))$ does not depend on nuisance parameter $\boldsymbol{\delta}$,
(c) for fixed $\boldsymbol{x}$, $P(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) > t)$ is an increasing function of θ_i, the selection criterion, for any given t.

Let g be a real value function. To construct a reasonable generalized subset selection procedure for selecting the best population, we need the following condition:

(d) the standard deviation of $g(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}), Y_j(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}))$ is independent of $\boldsymbol{\zeta}$ and depends only on observed data $\boldsymbol{x}$.

For observed $\boldsymbol{x}$ and given constant P^* $(1/k < P^* < 1)$, for each i, define $d_i(\boldsymbol{x})$ to be the constant (depending on $\boldsymbol{x}$) such that

$$P_{\theta_1 = \cdots = \theta_k}(g(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}), Y_j(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})) + d_i(\boldsymbol{x}) \mathrm{Sd}(g(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}), Y_j(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}))) \geqslant 0 \text{ for all } j,\ j \neq i) = P^*. \tag{2.2}$$

Then, a generalized subset selection procedure can be defined as follows:

R_g: For observed $\boldsymbol{x}$ and some given constant $d_i(\boldsymbol{x})$, which may depend on $\boldsymbol{x}$, retain π_i in the selected subset if and only if

$$\min_{j \neq i} \{ g(Y_i(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{\zeta}), Y_j(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{\zeta})) + d_i(\boldsymbol{x}) \mathrm{Sd}(g(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}), Y_j(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}))) \} \geqslant 0. \tag{2.3}$$

Also, the probability of generalized CS (GCS) of R_g is defined by

$$\begin{aligned} P(\mathrm{GCS}|R_g) &= P\left(\min_{j\neq(k)}\{g(Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})) \right. \\ &\qquad \left. + d_{(k)}(\boldsymbol{x})\,\mathrm{Sd}(g(Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})))\} \geqslant 0\right) \\ &= P(g(Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})) \\ &\qquad + d_{(k)}(\boldsymbol{x})\,\mathrm{Sd}(g(Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}))) \geqslant 0 \quad \text{for all } j,\ j\neq(k)). \end{aligned}$$

Therefore, if the minimizer of the probability of GCS, i.e. the least favorable configuration, occurs at $\theta_1 = \cdots = \theta_k$, then

$$\inf_{\boldsymbol{\Theta}} P(\mathrm{GCS}|R_g) = P^*, \tag{2.4}$$

where $\boldsymbol{\Theta}$ is the parameter space of all k-tuples $(\theta_1,\ldots,\theta_k)$. We call condition (2.4) the generalized P^*-condition. It is to be noted that the statement of (2.4) involves observed $\boldsymbol{x}$ and thus it is usually different from the usual P^*-condition (2.1). However, they may be equivalent in some situations (see Remark 3.1).

Remark 2.1. As in the case of conventional selection variables, conditions (a) and (c) of Definition 2.1 are imposed to ensure that $Y_i(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})$ is a reasonable generalized selection variable. Also, condition (a) of Definition 2.1 and condition (d) are imposed to ensure that generalized subset selection procedure generates selection regions which involve observed data $\boldsymbol{x}$ only. Condition (b) of Definition 2.1 is imposed to ensure that a selection region can be defined at desirable level in terms of probability of GCS under presence of nuisance parameters.

Remark 2.2. The probability in (2.2) is defined in terms of the random vector $\boldsymbol{X}$ for given observed value $\boldsymbol{x}$. The quantities $d_i(\boldsymbol{x})$ in the generalized subset selection rule R_g depend on the observed value $\boldsymbol{x}$ and thus it is quite different from the classical subset selection rule. As can be seen that (2.2) is a local property, local at $\boldsymbol{x}$, and it theoretically does not give any guarantee to the experimenter regarding the probability of a correct selection in the frequentist sense (see also Figs. 1–6).

As in the Bayesian treatment, the idea is to do the best with the observed data rather than on all possible samples. As also pointed out in Hamada and Weerahandi (2000), generalized inference is not based on repeated sampling considerations. Nevertheless, generalized procedures have desirable repeated sampling properties, e.g. the generalized tests and confidence intervals are typically procedures with guaranteed size (also see Figs. 1 and 3 for cases (ii) and (iii) in Sections 3 and 4, respectively). For further discussions and details in this direction, it is referred to Weerahandi (1995).

When a subset selection procedure has been applied, conclusion needs to be evaluated in terms of some statistical language. One way of evaluating the results of a subset selection procedure is to state the probability of GCS under least favorable

configuration. In addition to P^* and k, the quantities $d_i(\boldsymbol{x})$ depend on both $\boldsymbol{n}$ and data $\boldsymbol{x}$. Thus, it is not possible to tabulate $d_i(\boldsymbol{x})$ in general.

Another way of reporting the results of subset selection procedure is in terms of the so-called *s-values* (selection value) proposed by Hsu (1984). Note that, when the sample $\boldsymbol{x}$ is observed, for $P^* < P^{**}$, the selection under P^* is a subset of the selection under P^{**}. Hence, for some selection rule and an observed data $\boldsymbol{x}$, the *s-value* of π_i is defined to be the smallest value of P^* such that π_i is in the selected subset determined by the selection rule satisfying the P^*-condition. Therefore, the population π_i is in the selected subset associated with P^*-condition if and only if the s-value of π_i for the observed data $\boldsymbol{x}$ is less than or equal to P^*. Note that, the s-value depends on the observations $\boldsymbol{x}$ and the smaller the s-value of π_i, the stronger the evidence (by means of sample $\boldsymbol{x}$) that π_i is in the selected subset. Accordingly, we can extend the concept of s-value to the generalized s-value by associating with the generalized subset selection procedure.

Definition 2.2. For some generalized subset selection rule and an observed data $\boldsymbol{x}$, the *generalized s-value* of π_i is defined to be the smallest value of P^* for which π_i is in the selected subset determined by the generalized subset selection rule satisfying the generalized P^*-condition.

Theorem 2.1. *For the generalized subset selection procedure R_g defined by* (2.3), *let*

$$d_i^s(\boldsymbol{x}) = -\min_{j\neq i} \frac{g(Y_i(\boldsymbol{x},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{x},\boldsymbol{x},\boldsymbol{\zeta}))}{\mathrm{Sd}(g(Y_i(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})))}.$$

Then, the generalized s-value $P_i^s(\boldsymbol{x})$ of π_i is given by

$$P_i^s(\boldsymbol{x}) = P_{\theta_1=\cdots=\theta_k}(g(Y_i(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})) + d_i^s(\boldsymbol{x})\,\mathrm{Sd}(g(Y_i(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}))) \geqslant 0 \quad \textit{for all } j,\ j\neq i).$$

Proof. Since, for any fixed i and observed $\boldsymbol{x}$,

$$P_{\theta_1=\cdots=\theta_k}(g(Y_i(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})) + d_i(\boldsymbol{x})\,\mathrm{Sd}(g(Y_i(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}))) \geqslant 0 \quad \text{for all } j,\ j\neq i)$$

is increasing in $d_i(\boldsymbol{x})$, and for all j, $j\neq i$,

$$g(Y_i(\boldsymbol{x},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{x},\boldsymbol{x},\boldsymbol{\zeta})) + d_i(\boldsymbol{x})\,\mathrm{Sd}(g(Y_i(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}), Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}))) \geqslant 0$$

if and only if $d_i(\boldsymbol{x}) \geqslant d_i^s(\boldsymbol{x})$. Hence, if $P_i^s(\boldsymbol{x}) > P^*$ then $d_i(\boldsymbol{x}) < d_i^s(\boldsymbol{x})$ which indicates that π_i is not in the selected subset. On the other hand, if $P_i^s(\boldsymbol{x}) \leqslant P^*$ then $d_i(\boldsymbol{x}) \geqslant d_i^s(\boldsymbol{x})$. Therefore, π_i is in the selected subset. □

Obviously, the generalized s-value depends on the observations $\boldsymbol{x}$. Furthermore, the smaller the generalized s-value of π_i, the stronger the sample evidence (evidence under presence of $\boldsymbol{x}$) that π_i is in the selected subset.

3. Selection criterion is population mean

Let the ordered values of the unknown μ_i be denoted by $\mu_{[1]} \leqslant \mu_{[2]} \leqslant \cdots \leqslant \mu_{[k]}$. Furthermore, let (i) denote the association such that the population $\pi_{(i)}$ is associated with parameter $\mu_{[i]}$. Consider selection criterion to be μ_i.

3.1. When $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$ are unknown

Let $n^* = \sum_{i=1}^k (n_i - 1)$, $\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$, $S_p^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2/n^*$, and $\bar{x}_i$ and s_p^2 be the observed values of $\bar{X}_i$ and S_p^2, respectively. Consider the identity

$$\mu_i = \bar{X}_i - \frac{\sigma}{\sqrt{n_i}} Z_i = \bar{X}_i - \sqrt{\frac{n^* S_p^2}{n_i V}} Z_i,$$

where $Z_i = \sqrt{n_i}(\bar{X}_i - \mu_i)/\sigma \sim \mathrm{N}(0,1)$, $V = n^* S_p^2/\sigma^2 \sim \chi_{n^*}^2$.

Obviously, $Z_1, \ldots, Z_k$ and V are independent. We define the generalized selection variable $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ based on the sufficient statistics $\bar{X}_i$ and S_p^2 by

$$Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) = \sqrt{\frac{n^* s_p^2}{n_i V}} Z_i + \mu_i.$$

From the previously mentioned identity, the sample mean from π_i can be rewritten by

$$\bar{X}_i = \sqrt{\frac{n^* S_p^2}{n_i V}} Z_i + \mu_i.$$

Obviously, the generalized selection variable $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ is derived from $\bar{X}_i$ in which S_p^2 is replaced by its observed value s_p^2. More exactly, let $g_1(\boldsymbol{X}) = S_p^2$, $g_{2i}(\boldsymbol{X}) = Z_i/\sqrt{V/n^*}$ and $g_i^*(a,b) = \sqrt{a/n_i}\, b$, where n_i, n^*, S_p^2, V and Z_i are previously defined. Then, it is noted that $\bar{X}_i = g_i^*(g_1(\boldsymbol{X}), g_{2i}(\boldsymbol{X})) + \mu_i$ and $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) = g_i^*(g_1(\boldsymbol{x}), g_{2i}(\boldsymbol{X})) + \mu_i$ for observed $\boldsymbol{x}$. Therefore, the generalized selection variable $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ is a random quantity of $\boldsymbol{X}$ based on a partial data (information) of $\boldsymbol{x}$ (i.e. $g_1(\boldsymbol{x})$). When all data $\boldsymbol{x}$ is utilized, $Y_i(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{\zeta}) = g_i^*(g_1(\boldsymbol{x}), g_{2i}(\boldsymbol{x})) + \mu_i$ becomes $\bar{x}_i$. Here, it is noted that the first and the second variable in $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ denote, respectively, the random sample and the observed data from same experiment.

On the other hand, we can explain briefly the modified term. Note that the main part of $\bar{X}_i$ can be decomposed into a product form

$$\bar{X}_i = \left(\sqrt{\frac{n^* S_p^2}{n_i V}} \right) (Z_i) + \mu_i.$$

The part $\sqrt{n^* S_p^2/n_i V}$ contains the information of scale parameter of $\bar{X}_i$ (since Z_i has variance 1). So, for the case of equal unknown variances, when $\boldsymbol{x}$ is observed, the

"adjusted" quantity $\sqrt{n^* s_p^2/n_i V}$ is "better" than the original $\sqrt{n^* S_p^2/n_i V}$ in the sense that it reduces the variation of the latter. In this sense, instead of considering $\bar{X}_i$, we consider $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ just to "standardize" its scale for selection and leaving the part of statistic relating to information of mean unchanged. In this sense, $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ is more "informative" than $\bar{X}_i$ for the purpose of comparisons of means. Therefore, $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ is obtained from $\bar{X}_i$ through a process of selection-oriented adjustment. This interpretation will be more obvious for other selection criteria of unequal variances, pth quantile and signal-to-noise ratio.

Clearly, the distribution of $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ depends on (μ_i, σ^2) only through μ_i, and when all random quantities are replaced by its observed values,

$$Y_i(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{\zeta}) = (n^* s_p^2/n_i)^{1/2}(n^* s_p^2/\sigma^2)^{-1/2}\sqrt{n_i}(\bar{x}_i - \mu_i)/\sigma + \mu_i = \bar{x}_i$$

which is a naive estimate of μ_i, and for fixed $\boldsymbol{x}$, $P(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) > t)$ is an increasing function of μ_i for any given t. Therefore, $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ is indeed a generalized selection variable of μ_i. By a straightforward calculation, it can be obtained that

$$\lambda_{ij}(\boldsymbol{x}) \equiv \mathrm{Sd}(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) - Y_j(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})) = \sqrt{\frac{n^* s_p^2}{n^* - 2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

For observed $\boldsymbol{x}$, given constant P^*, and for each i, define $d_i(\boldsymbol{x})$ to be the constant such that

$$\int_0^\infty \int_{-\infty}^\infty \prod_{j \neq i} \Phi\left(\sqrt{\frac{n_j}{n_i}} z + \sqrt{\frac{n_j v}{n^* s_p^2}} d_i(\boldsymbol{x}) \lambda_{ij}(\boldsymbol{x})\right) \phi(z) p_{\chi^2_{n^*}}(v)\, \mathrm{d}z\, \mathrm{d}v = P^*, \tag{3.1}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are, respectively, the standard normal distribution and density function, and $p_{\chi^2_{n^*}}(\cdot)$ is the density function of $\chi^2_{n^*}$. In fact, Eq. (3.1) is equivalent to

$$\int_0^\infty \int_{-\infty}^\infty \prod_{j \neq i} \Phi\left(\sqrt{\frac{n_j}{n_i}} z + \sqrt{\frac{n_j v}{n^* - 2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} d_i(\boldsymbol{x})\right) \phi(z) p_{\chi^2_{n^*}}(v)\, \mathrm{d}z\, \mathrm{d}v = P^*. \tag{3.2}$$

Indeed, here the quantities $d_i(\boldsymbol{x})$ in fact are independent of $\boldsymbol{x}$. For convenience, we denote d_i instead of $d_i(\boldsymbol{x})$. We can define the generalized selection procedure R_{M} as follows:

R_{M}: For observed $\boldsymbol{x}$ and some given constant d_i, retain π_i in the selected subset if and only if

$$\min_{j \neq i}\{\bar{x}_i - \bar{x}_j + d_i \lambda_{ij}(\boldsymbol{x})\} \geqslant 0. \tag{3.3}$$

Theorem 3.1. *The $P(\mathrm{GCS}|R_{\mathrm{M}})$ satisfies the generalized P^*-condition if d_i satisfies* (3.2).

Proof. The probability of GCS applying the generalized selection procedure R_{M} is given by

$$
\begin{aligned}
&P(\mathrm{GCS}|R_{\mathrm{M}})\\
&\quad = P\left(\min_{j\neq(k)}\{Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) - Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) + d_{(k)}\lambda_{(k)j}(\boldsymbol{x})\} \geqslant 0\right)\\
&\quad = P(Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) - Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) + d_{(k)}\lambda_{(k)j}(\boldsymbol{x}) \geqslant 0 \quad \text{for all } j,\ j\neq(k))\\
&\quad = P\left(Z_j \leqslant \sqrt{\frac{n_j}{n_{(k)}}}Z_{(k)} + \sqrt{\frac{n_j V}{n^* s_p^2}}(\mu_{[k]} - \mu_j + d_{(k)}\lambda_{(k)j}(\boldsymbol{x})) \quad \text{for all } j,\ j\neq(k)\right)\\
&\quad = \int_0^\infty \int_{-\infty}^\infty P\left(Z_j \leqslant \sqrt{\frac{n_j}{n_{(k)}}}z + \sqrt{\frac{n_j v}{n^* s_p^2}}(\mu_{[k]} - \mu_j + d_{(k)}\lambda_{(k)j}(\boldsymbol{x}))\right.\\
&\qquad \left.\text{for all } j,\ j\neq(k)\right)\phi(z)\,p_{\chi^2_{n^*}}(v)\,\mathrm{d}z\,\mathrm{d}v\\
&\quad = \int_0^\infty \int_{-\infty}^\infty \prod_{j\neq(k)} \Phi\left(\sqrt{\frac{n_j}{n_{(k)}}}z + \sqrt{\frac{n_j v}{n^* s_p^2}}(\mu_{[k]} - \mu_j + d_{(k)}\lambda_{(k)j}(\boldsymbol{x}))\right)\\
&\qquad \times\phi(z)\,p_{\chi^2_{n^*}}(v)\,\mathrm{d}z\,\mathrm{d}v.
\end{aligned}
$$

Hence, the minimum of $P(\mathrm{GCS}|R_{\mathrm{M}})$ occurs at $\mu_1=\cdots=\mu_k$. Therefore, by the definition of $d_{(k)}$, $P(\mathrm{GCS}|R_{\mathrm{M}})\geqslant P^*$. □

Remark 3.1. Note that Gupta and Huang (1974, 1976), and Hsu (1984) have considered this problem. The procedure R_{H} given by Hsu (1984) is defined as follows:

R_{H}: Retain π_i in the selected subset, if and only if, for observed $\boldsymbol{x}$,

$$\min_{j\neq i}\left\{\bar{x}_i - \bar{x}_j + D_i s_p\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right\} \geqslant 0,$$

where D_i satisfies

$$\int_0^\infty \int_{-\infty}^\infty \prod_{j\neq i} \Phi\left(\sqrt{\frac{n_j}{n_i}}z + D_i v\sqrt{1+\frac{n_j}{n_i}}\right)\phi(z)\,p_{\chi_{n^*}/\sqrt{n^*}}(v)\,\mathrm{d}z\,\mathrm{d}v = P^*.$$

Take $d_i=\sqrt{(n^*-2)/n^*}D_i$, then the generalized subset selection procedure R_{M} defined by (3.3) is equivalent to that given in Hsu (1984). Also, by Hsu (1984), the procedure R_{GH} given by Gupta and Huang (1976) can be defined as follows:

R_{GH}: Retain π_i in the selected subset, if and only if, for observed $\boldsymbol{x}$,

$$\min_{j\neq i}\left\{\bar{x}_i - \bar{x}_j + D s_p\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right\} \geqslant 0,$$

where $D=\max_{1\leqslant i\leqslant k} D_i$. Note that, when $n_1=\cdots=n_k=n$, the procedure becomes the classical rule which was firstly proposed and studied by Gupta (1956, 1965).

It is to be noted that since the proposed generalized subset selection rule R_{M} is exactly equivalent to that of Hsu (1984), the generalized P^*-condition under present situation is just the same as that of the usual P^*-condition.

3.2. When σ_i^2 are unequal and unknown

Consider the situation that variances are unequal. Let $\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$, $S_i^2 = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2/n_i$, and $\bar{x}_i$ and s_i^2 be the observed values of $\bar{X}_i$ and S_i^2, respectively. Consider the identity

$$\mu_i = \bar{X}_i - \frac{\sigma_i}{\sqrt{n_i}} Z_i = \bar{X}_i - \frac{S_i}{\sqrt{V_i}} Z_i,$$

where

$$Z_i = \sqrt{n_i}(\bar{X}_i - \mu_i)/\sigma_i \sim \mathrm{N}(0,1), \tag{3.4}$$

$$V_i = n_i S_i^2/\sigma_i^2 \sim \chi^2_{n_i-1}, \tag{3.5}$$

and Z_i and V_i, $i = 1,\ldots,k$, are independent. For observed $\boldsymbol{x}$ (and thus an estimate s_i^2), we define the generalized selection variable $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ based on the sufficient statistics $\bar{X}_i$ and S_i^2 by

$$Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) = \frac{s_i}{\sqrt{V_i}} Z_i + \mu_i.$$

Clearly, the distribution of $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ depends on (μ_i, σ_i^2) only through μ_i,

$$Y_i(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{\zeta}) = s_i(n_i s_i^2/\sigma_i^2)^{-1/2}\sqrt{n_i}(\bar{X}_i - \mu_i)/\sigma_i + \mu_i = \bar{x}_i$$

which is a naive estimate of μ_i, and for fixed $\boldsymbol{x}$, $P(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) > t)$ is an increasing function of μ_i for any given t. Therefore, $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ is indeed a generalized selection variable of μ_i. By a straightforward calculation, it can be obtained that

$$\tilde{\lambda}_{ij}(\boldsymbol{x}) \equiv \mathrm{Sd}(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) - Y_j(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})) = \sqrt{\frac{s_i^2}{n_i - 3} + \frac{s_j^2}{n_j - 3}}\,. \tag{3.6}$$

For observed $\boldsymbol{x}$, given constant P^*, and for each i, define $\tilde{d}_i(\boldsymbol{x})$ to be constant such that

$$\int_0^\infty \int_{-\infty}^\infty \prod_{j \neq i} P\left(t_{n_j-1} \leqslant \sqrt{\frac{n_j - 1}{v}} \frac{s_i}{s_j} z + \frac{\sqrt{n_j - 1}}{s_j} \tilde{d}_i(\boldsymbol{x}) \tilde{\lambda}_{ij}(\boldsymbol{x}) \right) \times \phi(z)\, p_{\chi^2_{n_i-1}}(v)\, \mathrm{d}z\, \mathrm{d}v = P^*, \tag{3.7}$$

where t_{n_j-1} denotes a random variable of t-distribution with $n_j - 1$ degrees of freedom. Define the generalized selection procedure $\tilde{R}_{\mathrm{M}}$ as follows:

$\tilde{R}_{\mathrm{M}}$: For observed $\boldsymbol{x}$ and some given constant $\tilde{d}_i(\boldsymbol{x})$, retain π_i in the selected subset if and only if

$$\min_{j \neq i} \{\bar{x}_i - \bar{x}_j + \tilde{d}_i(\boldsymbol{x}) \tilde{\lambda}_{ij}(\boldsymbol{x})\} \geqslant 0.$$

Theorem 3.2. *The $P(\mathrm{GCS}|\tilde{R}_{\mathrm{M}})$ satisfies the generalized P^*-condition if $\tilde{d}_i(\boldsymbol{x})$ satisfies* (3.7).

Proof. The probability of GCS applying the generalized selection procedure $\tilde{R}_{\mathrm{M}}$ is given by

$$
\begin{aligned}
&P(\mathrm{GCS}|\tilde{R}_{\mathrm{M}})\\
&\quad= P\left(\min_{j\neq(k)}\{Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})-Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})+\tilde{d}_{(k)}(\boldsymbol{x})\tilde{\lambda}_{(k)j}(\boldsymbol{x})\}\geqslant 0\right)\\
&\quad= P(Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})-Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})+\tilde{d}_{(k)}(\boldsymbol{x})\tilde{\lambda}_{(k)j}(\boldsymbol{x})\geqslant 0 \quad \text{for all } j,\ j\neq(k))\\
&\quad= P\left(\frac{Z_j}{\sqrt{V_j/(n_j-1)}}\leqslant\sqrt{\frac{n_j-1}{V_{(k)}}}\frac{s_{(k)}}{s_j}Z_{(k)}+\frac{\sqrt{n_j-1}}{s_j}\right.\\
&\qquad\left.\times(\mu_{[k]}-\mu_j+\tilde{d}_{(k)}(\boldsymbol{x})\tilde{\lambda}_{(k)j}(\boldsymbol{x}))\quad \text{for all } j,\ j\neq(k)\right)\\
&\quad= \int_0^\infty\int_{-\infty}^\infty P\left(\frac{Z_j}{\sqrt{V_j/(n_j-1)}}\leqslant\sqrt{\frac{n_j-1}{v}}\frac{s_{(k)}}{s_j}z\right.\\
&\qquad\left.+\frac{\sqrt{n_j-1}}{s_j}(\mu_{[k]}-\mu_j+\tilde{d}_{(k)}(\boldsymbol{x})\tilde{\lambda}_{(k)j}(\boldsymbol{x}))\text{ for all } j,\ j\neq(k)\right)\\
&\qquad\times\phi(z)p_{\chi^2_{n_{(k)}-1}}(v)\,\mathrm{d}z\,\mathrm{d}v\\
&\quad= \int_0^\infty\int_{-\infty}^\infty\prod_{-\infty j\neq(k)}P\left(t_{n_j-1}\leqslant\sqrt{\frac{n_j-1}{v}}\frac{s_{(k)}}{s_j}z+\frac{\sqrt{n_j-1}}{s_j}(\mu_{[k]}-\mu_j+\tilde{d}_{(k)}(\boldsymbol{x})\tilde{\lambda}_{(k)j}(\boldsymbol{x}))\right)\\
&\qquad\times\phi(z)p_{\chi^2_{n_{(k)}-1}}(v)\,\mathrm{d}z\,\mathrm{d}v.
\end{aligned}
$$

Hence, the minimum of $P(\mathrm{GCS}|\tilde{R}_{\mathrm{M}})$ occurs at $\mu_1=\cdots=\mu_k$. Therefore, by the definition of $\tilde{d}_{(k)}$, $P(\mathrm{GCS}|\tilde{R}_{\mathrm{M}})\geqslant P^*$. □

For present situation, the generalized P^*-condition in general is not the same as in the usual frequentist sense, therefore, simulation study is important. To study the empirical PCS performance of the proposed procedure $\tilde{R}_{\mathrm{M}}$, suppose that $k=3$ and π_3 is the best population. Fig. 1 is based on 10,000 simulations. For the ℓth simulation, suppose the observations $\boldsymbol{x}$ are generated. Then the generalized s-values of π_i, $i=1,\ldots,3$, are calculated according to

$$
\begin{aligned}
P^s_{\ell i}(\boldsymbol{x})=&\int_0^\infty\int_{-\infty}^\infty\prod_{j\neq i}P\left(t_{n_j-1}\leqslant\sqrt{\frac{n_j-1}{v}}\frac{s_i}{s_j}z+\frac{\sqrt{n_j-1}}{s_j}\tilde{d}_{\ell i}(\boldsymbol{x})\tilde{\lambda}_{ij}(\boldsymbol{x})\right)\\
&\times\phi(z)p_{\chi^2_{n_i-1}}(v)\,\mathrm{d}z\,\mathrm{d}v,
\end{aligned}
$$

where

$$
\tilde{d}_{\ell i}(\boldsymbol{x})=-\min_{j\neq i}\{(\bar{x}_i-\bar{x}_j)/\tilde{\lambda}_{ij}(\boldsymbol{x})\},
$$

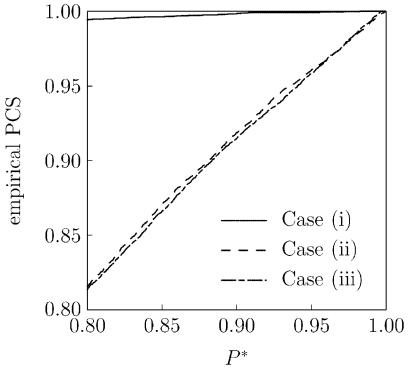

Fig. 1. Empirical PCS for selecting the largest mean.

and $\tilde{\lambda}_{ij}(\boldsymbol{x})$ is defined in (3.6). If the generalized s-value of π_3 is less than or equal to P^*, i.e. $P^s_{\ell 3}(\boldsymbol{x}) \leqslant P^*$, it means that π_3 is in the selected subset by $\tilde{R}_{\mathrm{M}}$ and thus the generalized subset selection procedure selects the population associated with the largest mean correctly. Therefore, for given value of P^*, the empirical PCS of $\tilde{R}_{\mathrm{M}}$ is given by $\sum_{\ell=1}^{10,000} I(P^s_{\ell 3}(\boldsymbol{x}) \leqslant P^*)/10,000$, where $I(\cdot)$ is the indicator function. By this way the empirical PCS can thus be plotted. Here, we consider three situations in the simulations: Case (i): $\mu_i = \sigma_i^2 = i,\ n_i = 10,\ i = 1,\ldots,3$, case (ii): $\mu_i = 3,\ \sigma_i^2 = 4 - i,\ n_i = 10,\ i = 1,\ldots,3$, and case (iii): $\mu_i = 3,\ \sigma_i^2 = 4 - i,\ n_i = 20,\ i = 1,\ldots,3$. Note that π_1, π_2 and π_3 all have the same mean for cases (ii) and (iii). Therefore, cases (ii) and (iii) belong to "least favorable" situations, but case (i) does not. From Fig. 1, it is obvious to see that the empirical PCS is greatly larger than its associated P^* under the case (i), but the empirical PCS is approximately equal to (no less than) its associated P^* under cases (ii) and (iii), respectively. Therefore, the generalized selection procedure $\tilde{R}_{\mathrm{M}}$ behaves conservative in the frequentist sense for case (i).

A naive procedure selects a single population randomly and its PCS achieves $1/k$. Therefore, we define efficiency of the generalized subset selection procedure with respect to the naive procedure by

$$\mathrm{eff} = \frac{\text{empirical PCS}}{(\mathrm{ES})^a} \Big/ \frac{1}{k},$$

where ES denotes empirical average size of subset selected and a is some constant in $(0, 1]$ which is to be adjusted by experimenter. For its simplicity, here we take $a = 1$. For given value of P^*, the size of subset selected for the ℓth simulation is just the number of those generalized s-values of π_i $(i = 1,\ldots,3)$ which are less than or equal to P^*, i.e. $\sum_{i=1}^{3} I(P^s_{\ell i}(\boldsymbol{x}) \leqslant P^*)$. Hence, for any given value P^*, based on 10,000 simulations, the efficiency of $\tilde{R}_{\mathrm{M}}$ can be obtained and Fig. 2 shows the efficiency for selecting the largest mean under case (i). On the other hand, the efficiencies are very close to 1 for cases (ii) and (iii). The reason is that π_1, π_2, and π_3 are all the best populations in these situations.

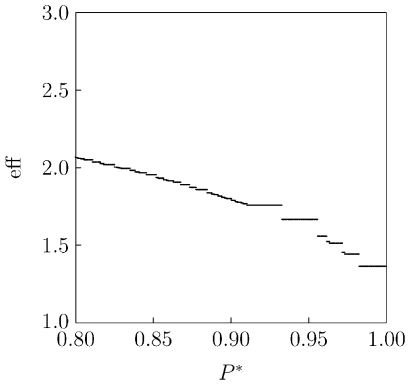

Fig. 2. Efficiency for selecting the largest mean under case (i).

4. Selection criterion is the *p*th quantile

Let the pth quantile corresponding to π_i be denoted by θ_i^p, then $\theta_i^p = \mu_i + \sigma_i \Phi^{-1}(p)$ where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function. Let the ordered values of the unknown θ_i^p be denoted by $\theta_{[1]}^p \leqslant \theta_{[2]}^p \leqslant \cdots \leqslant \theta_{[k]}^p$. Consider the identity

$$\theta_i^p = \bar{X}_i - \frac{S_i}{\sqrt{V_i}}(Z_i - \sqrt{n_i}\Phi^{-1}(p)),$$

where Z_i and V_i are defined, respectively, in (3.4) and (3.5). For observed $\boldsymbol{x}$ and estimated s_i, we define the generalized selection variable $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ based on the sufficient statistics $\bar{X}_i$ and S_i^2 as

$$Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) = \frac{s_i}{\sqrt{V_i}}(Z_i - \sqrt{n_i}\Phi^{-1}(p)) + s_i\Phi^{-1}(p) + \theta_i^p.$$

Clearly, the distribution of $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ depends on (θ_i^p, σ_i^2) only through θ_i^p, also,

$$\begin{aligned} Y_i(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{\zeta}) &= s_i(n_i s_i^2/\sigma_i^2)^{-1/2}\{\sqrt{n_i}(\bar{X}_i - \mu_i)/\sigma_i - \sqrt{n_i}\Phi^{-1}(p)\} + s_i\Phi^{-1}(p) + \theta_i^p \\ &= \bar{x}_i + s_i\Phi^{-1}(p) \end{aligned}$$

which is a naive estimate of θ_i^p, and for fixed $\boldsymbol{x}$, $P(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) > t)$ is an increasing function of θ_i^p for any given t. Therefore, $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ is thus a generalized selection variable of θ_i^p according to Definition 2.1. Also, it can be obtained that

$$\begin{aligned} \lambda_{ij}^Q(\boldsymbol{x}) &\equiv \mathrm{Sd}(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) - Y_j(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})) \\ &= \left(s_i^2 \left[\frac{1}{n_i - 3} + n_i(\Phi^{-1}(p))^2 \left\{ \frac{1}{n_i - 3} - \frac{1}{2}\left(\frac{\Gamma(\frac{n_i-2}{2})}{\Gamma(\frac{n_i-1}{2})} \right)^2 \right\} \right] \right. \\ &\quad \left. + s_j^2 \left[\frac{1}{n_j - 3} + n_j(\Phi^{-1}(p))^2 \left\{ \frac{1}{n_j - 3} - \frac{1}{2}\left(\frac{\Gamma(\frac{n_j-2}{2})}{\Gamma(\frac{n_j-1}{2})} \right)^2 \right\} \right] \right)^{1/2}. \end{aligned}$$

For observed $\boldsymbol{x}$, given constant P^*, and for each i, define $d_i^Q(\boldsymbol{x})$ to be the constant such that

$$\int_0^\infty \int_{-\infty}^\infty \prod_{j\neq i} P\left(t_{n_j-1}(\delta_j^Q) \leqslant \frac{\sqrt{n_j-1}}{s_j} \left\{ (s_i - s_j)\Phi^{-1}(p) + \frac{s_i}{\sqrt{v}}(z - \sqrt{n_i}\Phi^{-1}(p)) + d_i^Q(\boldsymbol{x})\lambda_{ij}^Q(\boldsymbol{x}) \right\} \right) \phi(z)\, p_{\chi^2_{n_i-1}}(v)\, \mathrm{d}z\, \mathrm{d}v$$

$$= P^*, \tag{4.1}$$

where $t_{n_j-1}(\delta_j^Q)$ denotes the random variable of non-central t-distribution with $n_j - 1$ degrees of freedom and noncentrality parameter $\delta_j^Q = -\sqrt{n_j}\Phi^{-1}(p)$. Now, define the generalized selection procedure R_Q as follows:

R_Q: For observed $\boldsymbol{x}$ and some given constant $d_i^Q(\boldsymbol{x})$, retain π_i in the selected subset if and only if

$$\min_{j\neq i}\{\bar{x}_i + s_i\Phi^{-1}(p) - (\bar{x}_j + s_j\Phi^{-1}(p)) + d_i^Q(\boldsymbol{x})\lambda_{ij}^Q(\boldsymbol{x})\} \geqslant 0.$$

Theorem 4.1. *The $P(\mathrm{GCS}|R_Q)$ satisfies the generalized P^*-condition if $d_i^Q(\boldsymbol{x})$ satisfies* (4.1).

Proof. Let (i) denote some association such that the population $\pi_{(i)}$ is associated with parameter $\theta_{[i]}^p$. The probability of GCS applying the generalized selection procedure R_Q is given by

$$\begin{aligned}
&P(\mathrm{GCS}|R_Q)\\
&= P\left(\min_{j\neq(k)}\{Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) - Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) + d_{(k)}^Q(\boldsymbol{x})\lambda_{(k)j}^Q(\boldsymbol{x})\} \geqslant 0 \right)\\
&= P(Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) - Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) + d_{(k)}^Q(\boldsymbol{x})\lambda_{(k)j}^Q(\boldsymbol{x}) \geqslant 0 \quad \text{for all } j,\ j\neq(k))\\
&= P\left(\frac{Z_j - \sqrt{n_j}\Phi^{-1}(p)}{\sqrt{V_j/(n_j-1)}} \leqslant \frac{\sqrt{n_j-1}}{s_j} \left\{ (s_{(k)} - s_j)\Phi^{-1}(p) + \theta_{[k]}^p - \theta_j^p \right.\right.\\
&\quad \left.\left. + \frac{s_{(k)}}{\sqrt{V_{(k)}}}(Z_{(k)} - \sqrt{n_{(k)}}\Phi^{-1}(p)) + d_{(k)}^Q(\boldsymbol{x})\lambda_{(k)j}^Q(\boldsymbol{x}) \right\} \text{ for all } j,\ j\neq(k) \right)\\
&= \int_0^\infty \int_{-\infty}^\infty P\left(\frac{Z_j - \sqrt{n_j}\Phi^{-1}(p)}{\sqrt{V_j/(n_j-1)}} \leqslant \frac{\sqrt{n_j-1}}{s_j} \left\{ (s_{(k)} - s_j)\Phi^{-1}(p) + \theta_{[k]}^p - \theta_j^p \right.\right.\\
&\quad \left.\left. + \frac{s_{(k)}}{\sqrt{v}}(z - \sqrt{n_{(k)}}\Phi^{-1}(p)) + d_{(k)}^Q(\boldsymbol{x})\lambda_{(k)j}^Q(\boldsymbol{x}) \right\} \text{ for all } j,\ j\neq(k) \right)\\
&\quad \times \phi(z)\, p_{\chi^2_{n_{(k)}-1}}(v)\, \mathrm{d}z\, \mathrm{d}v
\end{aligned}$$

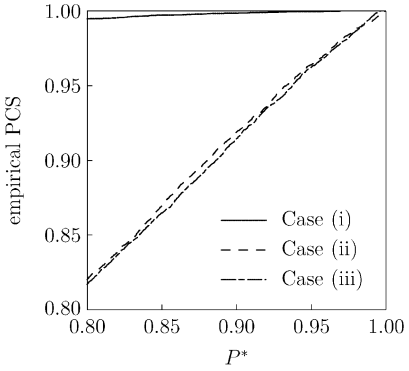

Fig. 3. Empirical PCS for selecting the largest 0.9th quantile.

$$= \int_0^\infty \int_{-\infty}^\infty \prod_{j\neq(k)} P\left(t_{n_j-1}(\delta_j^Q) \leqslant \frac{\sqrt{n_j-1}}{s_j} \left\{ (s_{(k)} - s_j)\Phi^{-1}(p) + \theta_{[k]}^p - \theta_j^p \right.\right.$$

$$\left.\left. + \frac{s_{(k)}}{\sqrt{v}}(z - \sqrt{n_{(k)}}\Phi^{-1}(p)) + d_{(k)}^Q(\boldsymbol{x})\lambda_{(k)j}^Q(\boldsymbol{x}) \right\} \right) \phi(z)\, p_{\chi^2_{n_{(k)}-1}}(v)\, \mathrm{d}z\, \mathrm{d}v,$$

where $\delta_j^Q = -\sqrt{n_j}\Phi^{-1}(p)$. Hence, the minimum of $P(\mathrm{GCS}|R_Q)$ occurs at $\theta_1^p = \cdots = \theta_k^p$. Therefore, by the definition of $d_{(k)}^Q(\boldsymbol{x})$, $P(\mathrm{GCS}|R_Q) \geqslant P^*$. □

Remark 4.1. If $p = 1/2$, then the generalized selection procedure R_Q is equivalent to the generalized selection procedure $\tilde{R}_{\mathrm{M}}$.

Remark 4.2. In general, if the selection criterion is the quantity $\theta_i = a\mu_i + b\sigma_i$, where a and b are given constants, the generalized selection procedure can be obtained by the same argument.

To study the empirical PCS performance of the proposed generalized subset selection procedure for the largest pth quantile, suppose that $p = 0.9$, $k = 3$ and π_3 is the best population. For given value of P^*, the empirical PCS and the efficiency of R_Q can be obtained by a similar argument in Section 4. Figs. 3 and 4 are based on 10,000 simulations, and the following three cases are considered, case (i): $\mu_i = \sigma_i^2 = i$, $n_i = 10$, $i = 1,\ldots,3$, case (ii): $\mu_i = 3 - \Phi^{-1}(p)\sigma_i$, $\sigma_i^2 = 4 - i$, $n_i = 10$, $i = 1,\ldots,3$, and case (iii): $\mu_i = 3 - \Phi^{-1}(p)\sigma_i$, $\sigma_i^2 = 4 - i$, $n_i = 20$, $i = 1,\ldots,3$. Note that π_1, π_2 and π_3 all have the same 0.9th quantile for cases (ii) and (iii). Therefore, cases (ii) and (iii) belong to the "least favorable" situations, however, case (i) does not. From Fig. 3, it is obvious to see that the empirical PCS is greatly larger than its associated P^* under case (i), but it is approximately equal to (but not less than) its associated P^* under

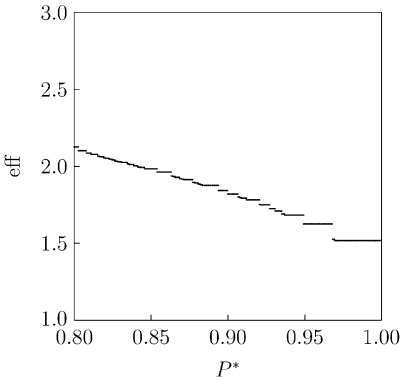

Fig. 4. Efficiency for selecting the largest 0.9th quantile under case (i).

cases (ii) and (iii), respectively. Therefore, the generalized selection procedure R_Q is conservative in the frequentist sense for case (i).

Fig. 4 shows the efficiency for selecting the largest 0.9th quantile under case (i). However, the efficiencies are very close to 1 for cases (ii) and (iii). The reason is that π_1, π_2, and π_3 are all the best populations in these cases.

5. Selection criterion is signal-to-noise ratio

Signal-to-noise ratio is an important quantity to be measured, particularly in industrial statistics. Let $\theta_i = \mu_i/\sigma_i$ denote the signal-to-noise ratio of π_i and the ordered values of the unknown θ_i be denoted by $\theta_{[1]} \leqslant \theta_{[2]} \leqslant \cdots \leqslant \theta_{[k]}$. Consider the identity

$$\theta_i = \frac{\bar{X}_i}{S_i}\sqrt{\frac{V_i}{n_i}} - \frac{Z_i}{\sqrt{n_i}},$$

where Z_i and V_i are defined in (3.4) and (3.5), respectively. For observed $\boldsymbol{x}$ and estimated $\bar{x}_i$ and s_i, we define the generalized selection variable $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ based on the sufficient statistics $\bar{X}_i$ and S_i^2 by

$$Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) = \frac{\bar{x}_i}{s_i} - \left(\frac{\bar{x}_i}{s_i}\sqrt{\frac{V_i}{n_i}} - \frac{Z_i}{\sqrt{n_i}}\right) + \theta_i.$$

Clearly, the distribution of $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ depends on (θ_i, σ_i^2) only through θ_i, and

$$Y_i(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{\zeta}) = \frac{\bar{x}_i}{s_i} - \left(\frac{\bar{x}_i}{s_i}\sqrt{\frac{n_i s_i^2/\sigma_i^2}{n_i}} - \frac{\sqrt{n_i}(\bar{x}_i - \mu_i)/\sigma_i}{\sqrt{n_i}}\right) + \theta_i = \bar{x}_i/s_i,$$

which is a naive estimate of θ_i, and for fixed $\boldsymbol{x}$, $P(Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta}) > t)$ is an increasing function of θ_i for any given t. Therefore, $Y_i(\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\zeta})$ is a generalized selection variable

of θ_i. Again,

$$\begin{aligned}\lambda_{ij}^{\mathrm{SN}}(\boldsymbol{x}) &\equiv \mathrm{Sd}(Y_i(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) - Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta})) \\ &= \left(\frac{1}{n_i}\left[\left(\frac{\bar{x}_i}{s_i}\right)^2\left\{n_i - 1 - 2\left(\frac{\Gamma(n_i/2)}{\Gamma((n_i-1)/2)}\right)^2\right\} + 1\right]\right. \\ &\quad \left. + \frac{1}{n_j}\left[\left(\frac{\bar{x}_j}{s_j}\right)^2\left\{n_j - 1 - 2\left(\frac{\Gamma(n_j/2)}{\Gamma((n_j-1)/2)}\right)^2\right\} + 1\right]\right)^{1/2}.\end{aligned}$$

For observed $\boldsymbol{x}$, given constant P^*, and for each i, define $d_i^{\mathrm{SN}}(\boldsymbol{x})$ to be the constant such that

$$\int_0^\infty \int_{-\infty}^\infty \prod_{j\neq i} P\left(t_{n_j-1}(\delta_j^{\mathrm{SN}}(\boldsymbol{x})) \leqslant \sqrt{n_j - 1}\left(\frac{\bar{x}_j}{s_j}\right)\right)\phi(z)\, p_{\chi^2_{n_i-1}}(v)\,\mathrm{d}z\,\mathrm{d}v = P^*, \tag{5.1}$$

where

$$\delta_j^{\mathrm{SN}}(\boldsymbol{x}) = -\sqrt{\frac{n_j}{n_i}}z - \sqrt{n_j}\left(\frac{\bar{x}_i}{s_i} - \frac{\bar{x}_j}{s_j}\right) + \frac{\bar{x}_i}{s_i}\sqrt{\frac{n_j}{n_i}}v - \sqrt{n_j}d_i^{\mathrm{SN}}(\boldsymbol{x})\lambda_{ij}^{\mathrm{SN}}(\boldsymbol{x}).$$

We can define the generalized selection procedure R_{SNR} as follows:

R_{SNR}: For observed $\boldsymbol{x}$ and some given constant $d_i^{\mathrm{SN}}(\boldsymbol{x})$, retain π_i in the selected subset if and only if

$$\min_{j\neq i}\left\{\frac{\bar{x}_i}{s_i} - \frac{\bar{x}_j}{s_j} + d_i^{\mathrm{SN}}(\boldsymbol{x})\lambda_{ij}^{\mathrm{SN}}(\boldsymbol{x})\right\} \geqslant 0.$$

Theorem 5.1. *The $P(\mathrm{GCS}|R_{\mathrm{SNR}})$ satisfies the generalized P^*-condition if $d_i^{\mathrm{SN}}(\boldsymbol{x})$ satisfies* (5.1).

Proof. As before, let (i) denote some association such that the population $\pi_{(i)}$ is associated with parameter $\theta_{[i]}$. The probability of GCS applying the generalized selection procedure R_{SNR} is then given by

$$\begin{aligned}&P(\mathrm{GCS}|R_{\mathrm{SNR}}) \\ &= P\left(\min_{j\neq(k)}\{Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) - Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) + d_{(k)}^{\mathrm{SN}}(\boldsymbol{x})\lambda_{(k)j}^{\mathrm{SN}}(\boldsymbol{x})\} \geqslant 0\right) \\ &= P(Y_{(k)}(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) - Y_j(\boldsymbol{X},\boldsymbol{x},\boldsymbol{\zeta}) + d_{(k)}^{\mathrm{SN}}(\boldsymbol{x})\lambda_{(k)j}^{\mathrm{SN}}(\boldsymbol{x}) \geqslant 0 \quad \text{for all } j,\ j\neq(k)) \\ &= P\left(\left\{Z_j - \sqrt{\frac{n_j}{n_{(k)}}}Z_{(k)} - \sqrt{n_j}\left(\frac{\bar{x}_{(k)}}{s_{(k)}} - \frac{\bar{x}_j}{s_j}\right) + \frac{\bar{x}_{(k)}}{s_{(k)}}\sqrt{\frac{n_j}{n_{(k)}}V_{(k)}}\right.\right. \\ &\quad \left.\left. - \sqrt{n_j}(\theta_{[k]} - \theta_j + d_{(k)}^{\mathrm{SN}}(\boldsymbol{x})\lambda_{(k)j}^{\mathrm{SN}}(\boldsymbol{x}))\right\} \Big/ \sqrt{V_j/(n_j-1)} \leqslant \sqrt{n_j - 1}\left(\frac{\bar{x}_j}{s_j}\right)\right.\end{aligned}$$

$$\text{for all } j,\ j \neq (k)\Bigg)$$

$$= \int_0^\infty \int_{-\infty}^\infty P\left(\left\{Z_j - \sqrt{\frac{n_j}{n_{(k)}}}z - \sqrt{n_j}\left(\frac{\bar{x}_{(k)}}{s_{(k)}} - \frac{\bar{x}_j}{s_j}\right) + \frac{\bar{x}_{(k)}}{s_{(k)}}\sqrt{\frac{n_j}{n_{(k)}}}v \right.\right.$$

$$\left.\left. - \sqrt{n_j}(\theta_{[k]} - \theta_j + d^{\mathrm{SN}}_{(k)}(\boldsymbol{x})\lambda^{\mathrm{SN}}_{(k)j}(\boldsymbol{x}))\right\} \Big/ \sqrt{V_j/(n_j-1)} \leqslant \sqrt{n_j - 1}\left(\frac{\bar{x}_j}{s_j}\right)$$

$$\text{for all } j,\ j \neq (k)\Bigg)\phi(z)\, p_{\chi^2_{n_{(k)}-1}}(v)\,\mathrm{d}z\,\mathrm{d}v$$

$$= \int_0^\infty \int_{-\infty}^\infty \prod_{j\neq(k)} P\left(t_{n_j-1}(\delta^{\mathrm{SN}}_j(\boldsymbol{x})) \leqslant \sqrt{n_j - 1}\left(\frac{\bar{x}_j}{s_j}\right)\right)\phi(z)\, p_{\chi^2_{n_{(k)}-1}}(v)\,\mathrm{d}z\,\mathrm{d}v,$$

where

$$\delta^{\mathrm{SN}}_j(\boldsymbol{x}) = -\sqrt{\frac{n_j}{n_{(k)}}}z - \sqrt{n_j}\left(\frac{\bar{x}_{(k)}}{s_{(k)}} - \frac{\bar{x}_j}{s_j}\right) + \frac{\bar{x}_{(k)}}{s_{(k)}}\sqrt{\frac{n_j}{n_{(k)}}}v$$

$$- \sqrt{n_j}(\theta_{[k]} - \theta_j + d^{\mathrm{SN}}_{(k)}(\boldsymbol{x})\lambda^{\mathrm{SN}}_{(k)j}(\boldsymbol{x})).$$

It is readily seen that the minimum of $P(\mathrm{GCS}|R_{\mathrm{SNR}})$ occurs at $\theta_1 = \cdots = \theta_k$. Therefore, according to the requirement of $d^{\mathrm{SN}}_{(k)}(\boldsymbol{x})$, $P(\mathrm{GCS}|R_{\mathrm{SNR}}) \geqslant P^*$. □

To study the empirical PCS performance of the proposed generalized subset selection procedure for present case, suppose that $k = 3$ and π_3 is the best population. Figs. 5 and 6 are based on 10,000 simulations, and the following cases are considered, case (i): $\mu_i = \sigma_i^2 = i,\ n_i = 10,\ i = 1,\ldots,3$, case (ii): $\mu_i = 3\sigma_i,\ \sigma_i^2 = 4 - i,\ n_i = 10,\ i = 1,\ldots,3$, and case (iii): $\mu_i = 3\sigma_i,\ \sigma_i^2 = 4 - i,\ n_i = 20,\ i = 1,\ldots,3$. Note that π_1, π_2 and π_3 all have the same signal-to-noise ratio for cases (ii) and (iii). Therefore, both cases (ii) and (iii) belong to the "least favorable" situations, but case (i) does not. From Fig. 5, it is seen that R_{SNR} behaves rather conservative for case (i) in the same sense as noted previously. Under cases (ii) and (iii), although, for some values of P^*, the empirical PCS is smaller than its associated P^*, but it is quite close to its associated P^*.

Fig. 6 shows the efficiency for selecting the largest signal-to-noise ratio under case (i). It is to be noted that the efficiency is less than that in Sections 3 and 4. For other cases, the efficiencies are very close to 1, due to π_1, π_2 and π_3 all being the best populations.

6. Illustration of a set of real data

The data in Table 1 originated from the study by Ott (1993, p. 840) of the nitrogen contents (X_{ij}) of red cover plants inoculated with $k = 3$ strains of Rhizobium (3dok1, 3dok5, 3dok7). It is seen that the sample standard deviation is rather large when the sample mean is large. Applying the Bartlett's test for homogeneity of variances, the

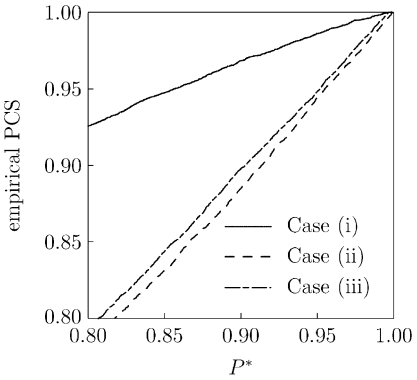

Fig. 5. Empirical PCS for selecting the largest signal-to-noise ratio.

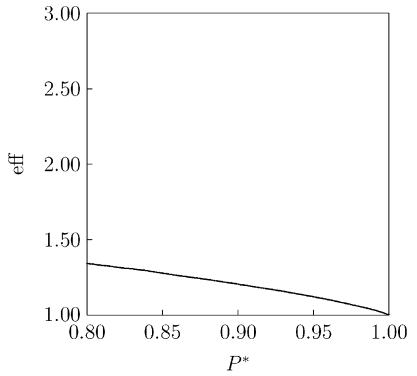

Fig. 6. Efficiency for selecting the largest signal-to-noise ratio under case (i).

Table 1
Rhizobium data

Term	3dok1	3dok5	3dok7
	19.4	18.2	20.7
	32.6	24.6	21.0
	27.0	25.5	20.5
x_{ij}	32.1	19.4	18.8
	33.0	21.7	18.6
		20.8	20.1
			21.3
n_i	5	6	7
$\bar{x}_i$	28.820	21.700	20.143
s_i	5.188	2.620	0.978
$\bar{x}_i + s_i\Phi^{-1}(0.9)$	35.468	25.058	21.396
$\bar{x}_i/s_i$	5.555	8.281	20.593

Table 2
Generalized s-value of each population for respective generalized subset selection procedure $\tilde{R}_{\mathrm{M}}$, R_Q, and R_{SNR}

Subset selection procedure	3dok1	3dok5	3dok7
$\tilde{R}_{\mathrm{M}}$	0.02392	0.94295	0.98086
R_Q	0.00006	0.91675	0.99924
R_{SNR}	0.98176	0.95955	0.00001

p-value is equal to 0.0037 and thus it shows that there are statistically significant differences in variances at 0.05 significance level. This fact is also supported by other tests.

Applying the proposed generalized subset selection procedure $\tilde{R}_{\mathrm{M}}$, R_Q, and R_{SNR}, Table 2 tabulates the generalized s-value of each population. For given value of P^*, a population is selected if its generalized s-value is less than or equal to P^*. For example, if P^* is 0.95, then the selected subset is $\{3\mathrm{dok}1, 3\mathrm{dok}5\}$ for selecting the largest mean, the selected subset is $\{3\mathrm{dok}1, 3\mathrm{dok}5\}$ for selecting the largest 0.9th quantile and the selected subset is $\{3\mathrm{dok}7\}$ for the largest signal-to-noise ratio.

Acknowledgements

The authors would like to thank two referees for helpful comments and suggestions which have improved the presentation of this paper. This research was partially supported by NSC 88-2118-M-001-004 and NSC 89-2118-M-031-002 from National Science Council, ROC.

References

Bechhofer, R.E., 1954. A single-sample multiple-decision procedure for ranking means of normal populations with known variances. Ann. Math. Statist. 25, 16–39.

Box, G., 1988. Signal-to-noise ratios, performance criteria and transformations. Technometrics 30, 1–17.

Gupta, S.S., 1956. On a Decision Rule for a Problem in Ranking Means. Mimeograph Series No. 150, Institute of Statistics, University of North Carolina, Chapel Hill, NC.

Gupta, S.S., 1965. On some multiple decision (selection and ranking) rules. Technometrics 7, 225–245.

Gupta, S.S., Huang, D.Y., 1976. Selection procedures for the means and variance of normal population: unequal sample size case. Sankhyā Ser. A 38, 153–173.

Gupta, S.S., Huang, W.T., 1974. A note on selecting a subset of normal populations with unequal sample sizes. Sankhyā Ser. A 36, 389–396.

Gupta, S.S., Panchapakesan, S., 1979. Multiple Decision Procedures. Wiley, New York.

Gupta, S.S., Wong, W.Y., 1982. Subset selection procedures for the means of normal populations with unequal variances: unequal sample sizes case. Sel. Statist. Can. 6, 109–149.

Hamada, M., Weerahandi, S., 2000. Measurement system assessment. J. Qual. Tech. 32, 241–253.

Hsu, J.C., 1984. Ranking and selection and multiple comparisons with the best. In: Santner, T.J., Tamhane, A.C. (Eds.), Design of Experiments: Ranking and Selection. Marcel Dekker, New York, pp. 179–198.

Ott, R.L., 1993. An Introduction to Statistical Methods and Data Analysis. Wadsworth Publishing Company Inc., Belmont, CA.

Santner, T.J., Tamhane, A.C., 1984. Designing experiments for selecting a normal population with a largest mean and small variance. In: Santner, T.J., Tamhane, A.C. (Eds.), Design of Experiments: Ranking and Selection. Marcel Dekker, New York, pp. 179–198.

Tsui, K.W., Weerahandi, S., 1989. Generalized *p*-values in significance testing of hypotheses in the presence of nuisance parameters. J. Amer. Statist. Assoc. 84, 602–607.

Weerahandi, S., 1993. Generalized confidence intervals. J. Amer. Statist. Assoc. 88, 899–905 (correction in 89 (1994) 726).

Weerahandi, S., 1995. Exact Statistical Methods for Data Analysis. Springer, New York.

Zhou, L., Mathew, T., 1994. Some tests for variance components using generalized *p*-values. Technometrics 36, 394–403.