SUBSET SELECTION PROCEDURES FOR THE MEANS OF NORMAL POPULATIONS WITH UNEQUAL VARIANCES: UNEQUAL SAMPLE SIZES CASE *

BY

Shanti S. Gupta and Wing-Yue Wong **
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #473

October 1976

* This research was supported by the Office of Naval Research contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.
** Now at the University of Malaya, Kuala Lumpur, Malaysia.

SUBSET SELECTION PROCEDURES FOR THE MEANS OF NORMAL POPULATIONS WITH UNEQUAL VARIANCES: UNEQUAL SAMPLE SIZES CASE

by

Shanti S. Gupta and Wing-Yue Wong

Purdue University

ABSTRACT

This paper deals with some subset selection procedures for the largest unknown means of k normal populations with unequal variances. The procedures are based on unequal numbers of observations from each of k normal populations. Some properties of the proposed selection rules are investigated. The problem of selecting all the normal populations with means better than a standard or control is also considered. Again, the proposed procedures are based on unequal number of observations from each of the populations. An application to testing the equality of k normal means with unequal variances, as in Behrens-Fisher problem, is described.

SUBSET SELECTION PROCEDURES FOR THE MEANS OF NORMAL POPULATIONS WITH UNEQUAL VARIANCES: UNEQUAL SAMPLE SIZES CASE *

by

Shanti S. Gupta and Wing-Yue Wong **

Purdue University

1. Introduction.

Let $\pi_1, \pi_2, \ldots, \pi_k$ be k independent normal populations with unknown means $\mu_1, \mu_2, \ldots, \mu_k$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$, respectively. Our goal is to select a nonempty subset of the k populations containing the population with the largest mean. In most of the earlier work (see for example Gupta [8]) it is assumed that either the number of observations from each population is the same or all the populations have a common variance. Very little work has been done in the case of unequal sample sizes and different variances. Sitek [14] proposed a procedure for the normal means; however, her result was shown to be in error by Dudewicz [5]. Recently Gupta and W. T. Huang [10] and Gupta and D. Y. Huang [9] proposed some subset selection procedures for selecting a subset of the unknown normal means. However, all the works mentioned above are based on the assumption that the given k normal populations have a common variance. For the case of unequal variances, Dudewicz [4] and Dudewicz, E. J. and Dalal, S. R. [6] proposed a two-sample procedure for the normal means problem.Their procedure is based on a linear combination of the first stage sample mean and the second stage sample. Chen, Dudewicz and Lee [1] have also made some contributions to this problem.

---

* This research was supported by the Office of Naval Research contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.
**Now at the University of Malaya, Kuala Lumpur, Malaysia.

In this paper some procedures based on the overall sample means are proposed and studied. In Section 2, two procedures are proposed and studied under the assumption that the variances are all known. When the variances are unknown to the experimenter, the problem is more difficult than the one above. In this case, another subset selection procedure is proposed and investigated. In Section 3, selection procedures for treatments better than a standard or control are discussed. A test of homogeneity is proposed which is based on the range of sample means and is given in Section 5.

2. Selecting the Normal Population with the Largest Mean.

Let $\pi_1, \pi_2, \ldots, \pi_k$ be k independent normal populations with unknown means $\mu_1, \mu_2, \ldots, \mu_k$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$, respectively. The ordered μ_i are denoted by $\mu_{[1]} \leq \mu_{[2]} \leq \ldots \leq \mu_{[k]}$. Here we assume that there is no prior knowledge of the correct pairing of the ordered and unordered μ_i's. Let $X_{i1}, X_{i2}, \ldots, X_{in_i}$ be n_i independent observations from population π_i, $i = 1,2,\ldots,k$. Based on these observations, our goal is to select a nonempty subset of the k populations so as to include the population associated with $\mu_{[k]}$. A correct selection (CS) is the selection of any subset containing the population associated with $\mu_{[k]}$. The object is to define a (non-trivial) procedure R so that $P(CS|R)$, the probability of a correct selection, is at least a preassigned number $P^*(\frac{1}{k} < P^* < 1)$ and which has some desirable properties. We shall refer to this requirement as P^*-condition. We shall discuss the two cases: (a) $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$ unequal but known, and (b) $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$ unequal and unknown.

Case (a): $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$ unequal but known.

Let $X_{i1}, X_{i2}, \ldots, X_{in_i}$ be n_i independent random samples drawn from population π_i, $i = 1,2,\ldots,k$. Let $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ denote the sample mean. We define the following rule R_1 based on these sample means.

R_1: Select the population π_i if and only if

$$\bar{X}_i \geq \max_{1 \leq j \leq k} \left(\bar{X}_j - c_1 \sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}}\right) \tag{2.1}$$

where $c_1 = c_1(k, P^*, (\sigma_1, n_1), (\sigma_2, n_2), \ldots, (\sigma_k, n_k))$ is the smallest nonnegative number chosen so as to satisfy the P*-condition.

Let $\bar{X}_{(i)}$, $n_{(i)}$ and $\sigma_{(i)}^2$ be the sample mean, sample size and variance associated with the population $\pi_{(i)}$ with mean $\mu_{[i]}$, $i = 1,2,\ldots,k$. It should be pointed out that $\bar{X}_{(i)}$, $n_{(i)}$ and $\sigma_{(i)}^2$ are all unknown. For the evaluation of the infimum of $P(CS|R_1)$, we need the following lemma which is proved in [2].

Lemma 2.1. If $\underline{X}' = (X_1, \ldots, X_m)$ has density $|\Sigma|^{-\frac{1}{2}} f(\underline{x}'\Sigma^{-1}\underline{x})$, then for any two positive definite (symmetric) mxm matrices $\Gamma_1 = (r_{ij})$ and $\Gamma_2 = (s_{ij})$ such that $r_{ii} = s_{ii}$, $1 \leq i \leq m$ and $r_{ij} \geq s_{ij}$, $1 \leq i < j \leq m$, then

$$P_{\Gamma_1}\{X_1 < a_1, \ldots, X_m < a_m\} \geq P_{\Gamma_2}\{X_1 < a_1, \ldots, X_m < a_m\}$$

for any real numbers $a_1, \ldots, a_m$.

Let Φ and φ denote the cdf and pdf of a standard normal variate. Now we prove the following theorem regarding the infimum of $P(CS|R_1)$.

Theorem 2.1. For the rule R_1 defined in (2.1),

$$\min_{(\sigma_1,n_1),\ldots,(\sigma_k,n_k)} \inf_{\Omega_1} P(CS|R_1) = \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \Phi\left(\frac{c_1 - \alpha_i x}{\sqrt{1-\alpha_i^2}}\right) d\Phi(x) \tag{2.2}$$

where $\Omega_1 = \{\underline{\mu}: \underline{\mu} = (\mu_1,\mu_2,\ldots,\mu_k),\ -\infty < \mu_i < \infty,\ i = 1,\ldots,k\}$ and

$$(2.3) \qquad \alpha_i = \{1 + \frac{(\frac{\sigma^2}{n})_{[i+1]}}{(\frac{\sigma^2}{n})_{[1]}}\}^{-\frac{1}{2}}, \qquad i = 1,\ldots,k-1;$$

where $(\frac{\sigma^2}{n})_{[1]} \leq \ldots \leq (\frac{\sigma^2}{n})_{[k]}$ denote the ordered values of $\frac{\sigma_i^2}{n_i}$.

Proof.

$$P(CS|R_1) + P\{\bar{X}_{(k)} \geq \max_{1\leq j\leq k-1} (\bar{X}_{(j)} - c_1\sqrt{\frac{\sigma^2_{(k)}}{n_{(k)}} + \frac{\sigma^2_{(j)}}{n_{(j)}}})\}$$

$$= P\left\{\frac{\bar{X}_{(j)} - \bar{X}_{(k)}}{\sqrt{\frac{\sigma^2_{(j)}}{n_{(j)}} + \frac{\sigma^2_{(k)}}{n_{(k)}}}} \leq c_1, \qquad j = 1,2,\ldots,k-1\right\}$$

$$(2.4) \qquad = P\{Z_{jk} \leq c_1 + \frac{\mu_{[k]} - \mu_{[j]}}{\sqrt{\frac{\sigma^2_{(k)}}{n_{(k)}} + \frac{\sigma^2_{(j)}}{n_{(j)}}}}, \quad j = 1,2,\ldots,k-1\}$$

where for $j = 1,2,\ldots,k-1$, Z_{jk} is given by

$$(2.5) \qquad Z_{jk} = \frac{\bar{X}_{(j)} - \bar{X}_{(k)} - \mu_{[j]} + \mu_{[k]}}{\sqrt{\frac{\sigma^2_{(j)}}{n_{(j)}} + \frac{\sigma_{(k)}}{n_{(k)}}}}.$$

Thus $(Z_{1k},Z_{2k},\ldots,Z_{k-1\ k})$ is a $(k-1)$-variate normal random vector with zero mean vector, unit variances and correlation matrix $(\rho_{ij}^{(k)})$ where

$$(2.6) \qquad \rho_{ij}^{(k)} = \{(1 + \frac{\sigma^2_{(i)}}{\sigma^2_{(k)}} \cdot \frac{n_{(k)}}{n_{(i)}})(1 + \frac{\sigma^2_{(j)}}{\sigma^2_{(k)}} \cdot \frac{n_{(k)}}{n_{(j)}})\}^{-\frac{1}{2}},$$

$$i,j = 1,2,\ldots,k-1,\ i \neq j.$$

We see from (2.4) that for fixed $(\sigma_1,n_1),\ldots,(\sigma_k,n_k)$, the infimum of $P(CS|R_1)$ will be attained when $\mu_{[1]} = \ldots = \mu_{[k]}$. Thus the infimum we seek in (2.2) is reduced to

$$\min_{(\sigma_1,n_1),\ldots,(\sigma_k,n_k)} P\{Z_{jk} \leq c_1,\ j = 1,\ldots,k-1\}$$

where Z_{jk} are defined by (2.5). For $1 \leq \ell \leq k$, if we let

$$\kappa_{ij}^{(\ell)} = \{(1 + \frac{(\frac{\sigma^2}{n})[i]}{(\frac{\sigma^2}{n})[\ell]})(1 + \frac{(\frac{\sigma^2}{n})[j]}{(\frac{\sigma^2}{n})[\ell]})\}^{-\frac{1}{2}}$$

$$i,j = 1,2,\ldots,k;\ i,j \neq \ell,\ i \neq j, \tag{2.8}$$

$$\kappa_{ii}^{(\ell)} = 1, \qquad i = 1,2,\ldots,k,\ i \neq \ell,$$

it follows from the fact that $(\frac{\sigma^2}{n})[1] \leq (\frac{\sigma^2}{n})[\ell]$ for all $\ell = 2,3,\ldots,k$, we have

$$\kappa_{ij}^{(\ell)} \geq \kappa_{i\,j+1}^{(\ell)} \geq \kappa_{ij}^{(1)} \qquad \text{for all } i,j = 1,2,\ldots,k-1, \tag{2.9}$$

$i \neq \ell,k$. By Lemma 2.1, this implies that

$$\min_{(\sigma_1,n_1),\ldots,(\sigma_k,n_k)} P\{Z_{jk} \leq c_1,\ j = 1,\ldots,k-1\} \tag{2.10}$$
$$= P\{Y_j \leq c_1,\ j = 1,\ldots,k-1\}$$

where $(Y_1,Y_2,\ldots,Y_{k-1})$ represents a $(k-1)$-variate normal random vector with zero mean vector, unit variances and correlation matrix (ζ_{ij}), where

$$\zeta_{ij} = \{(1 + \frac{(\frac{\sigma^2}{n})[i+1]}{(\frac{\sigma^2}{n})[1]})(1 + \frac{(\frac{\sigma^2}{n})[j+1]}{(\frac{\sigma^2}{n})[1]})\}^{-\frac{1}{2}}, \tag{2.11}$$

$$i,j = 1,2,\ldots,k-1,\ i \neq j.$$

Let

$$\alpha_i = (1 + \frac{(\frac{\sigma^2}{n})[i+1]}{(\frac{\sigma^2}{n})[1]})^{-\frac{1}{2}}, \qquad i = 1,\ldots,k-1. \tag{2.12}$$

It is well-known that $Y_1,Y_2,\ldots,Y_{k-1}$ can be generated from k independent

standard normal variates $Z_1, Z_2, \ldots, Z_k$ by the transformation

$$Y_i = (1-\alpha_i^2)^{\frac{1}{2}} Z_i + \alpha_i Z_k, \quad i = 1,\ldots,k-1. \tag{2.13}$$

Hence the right hand side of (2.10) can be rewritten as follows:

$$\int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \Phi\left(\frac{c_1 - \alpha_i x}{\sqrt{1-\alpha_i^2}}\right) d\Phi(x). \tag{2.14}$$

This completes the proof of the theorem.

It should be pointed out that when $\sigma_{[1]} = \sigma_{[2]} = \ldots = \sigma_{[k]} = \sigma$, say, the expression (2.14) is independent of σ. This reduces to the result obtained by Gupta and Huang [9].

Consistent with the basic probability requirement, we would like the size of the selected subset to be small. Now, S, the size of the selected subset is a random variable which takes values $1,2,\ldots,k$. Hence one can use, as a criterion of the efficiency of the procedure R_1, the expected value of the size of the selected subset.

Theorem 2.2. For the rule R_1 defined in (2.1)

$$\max_{(\sigma_1,n_1),\ldots,(\sigma_k,n_k)} \sup_{\underline{\mu}\in\Omega_1} E_{\underline{\mu}}(S|R_1) \leq k\, \Phi(c_1). \tag{2.15}$$

Proof.

$$E_{\underline{\mu}}(S|R_1) = \sum_{i=1}^{k} P(\pi_{(i)} \text{ is selected}|R_1) \tag{2.16}$$

$$= \sum_{i=1}^{k} P\left\{ \max_{\substack{1\leq j\leq k \\ j\neq i}} \left(\frac{\bar{X}_{(j)} - \bar{X}_{(i)}}{\sqrt{\frac{\sigma^2_{(j)}}{n_{(j)}} + \frac{\sigma^2_{(i)}}{n_{(i)}}}}\right) \leq c_1 \right)$$

$$\leq \sum_{i=1}^{k-1} P\left\{\frac{\bar{X}_{(k)} - \bar{X}_{(i)}}{\sqrt{\frac{\sigma^2_{(k)}}{n_{(k)}} + \frac{\sigma^2_{(i)}}{n_{(i)}}}} \leq c_1\right\} + P\left\{\frac{\bar{X}_{(k-1)} - \bar{X}_{(k)}}{\sqrt{\frac{\sigma^2_{(k-1)}}{n_{(k-1)}} + \frac{\sigma^2_{(k)}}{n_{(k)}}}} \leq c_1\right\}$$

$$\leq \sum_{i=1}^{k-2} \Phi\left(c_1 - \frac{\mu_{[k]} - \mu_{[i]}}{\sqrt{\frac{\sigma^2_{(k)}}{n_{(k)}} + \frac{\sigma^2_{(i)}}{n_{(i)}}}}\right) + \Phi\left(c_1 - \frac{\mu_{[k]} - \mu_{[k-1]}}{\sqrt{\frac{\sigma^2_{(k)}}{n_{(k)}} + \frac{\sigma^2_{(k-1)}}{n_{(k-1)}}}}\right)$$

$$+ \Phi\left(c_1 + \frac{\mu_{[k]} - \mu_{[k-1]}}{\sqrt{\frac{\sigma^2_{(k)}}{n_{(k)}} + \frac{\sigma^2_{(k-1)}}{n_{(k-1)}}}}\right).$$

It is easy to verify that if Z is a standard normal random variate, then for any nonnegative real numbers a and b,

(2.17) $\quad P\{a \leq Z \leq a + b\} \leq P\{a-b \leq Z \leq a\}.$

It follows from (2.17) that the right hand member of (2.16) is less than $k\,\Phi(c_1)$.

Remark 2.1. For $k = 2$,

$$\int_{-\infty}^{\infty} \Phi\left(\frac{c_1 - \alpha_1 x}{\sqrt{1-\alpha_1^2}}\right) d\Phi(x) = \Phi(c_1).$$

It follows that the constant c_1 obtained to satisfy the P*-condition is given by $\Phi(c_1) = P^*$. Thus in this case the upper bound of $E(S|R_1)$ is $2P^*$, which is the exact upper bound in the case of equal sample size and equal known variance.

Let $p_i(\underline{\mu}|R_1)$ denote the probability that the population π_i is included in the selected subset.

$$p_i(\underline{\mu}|R_1) = P\{\bar{X}_i \geq \max_{\substack{1\leq j\leq k \\ j\neq i}} (\bar{X}_j - c_1\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}}\}$$

$$= P\{\bar{X}_j - \bar{X}_i \leq c_1\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}}, \; j = 1,\ldots,k, \; j \neq i\}$$

$$= P\{Y_{ji} \leq c_1 + \frac{\mu_i - \mu_j}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}}}, \quad j = 1,\ldots,k,\ j \neq i\}$$

where

$$Y_{ji} = \frac{\bar{X}_j - \bar{X}_i - \mu_j + \mu_i}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}}}.$$

Thus $(Y_{1i},\ldots,Y_{i-1,i},Y_{i+1,i},\ldots,Y_{k_i})$ is a (k-1)-variate normal random vector with zero mean vector, unit variances and correlation matrix (ξ^i_{rs}) where

$$\xi^i_{rs} = \{(1 + \frac{\sigma_r^2}{\sigma_i^2} \cdot \frac{n_i}{n_r})(1 + \frac{\sigma_s^2}{\sigma_i^2} \cdot \frac{n_i}{n_s})\}^{-\frac{1}{2}} \cdot \quad r \neq s,\ r,s \neq i$$

$$= 1 \qquad\qquad r = s.$$

It follows that $p_i(\underline{\mu}|R_1)$ is increasing in μ_i when $\frac{\sigma_\ell^2}{n_\ell}$, $\ell = 1,\ldots,$ and all other components of $\underline{\mu}$ are kept fixed; also $p_i(\underline{\mu}|R_1)$ is decreasing in $\mu_j (j \neq i)$ when $\frac{\sigma_\ell^2}{n_\ell}$, $\ell = 1,\ldots,k$ and all other components of $\underline{\mu}$ are kept fixed. In particular, if $\frac{\sigma_i^2}{n_i} = \frac{\sigma_j^2}{n_j}$ and $\mu_i \leq \mu_j$ then $p_i(\underline{\mu}|R_1) \leq p_j(\underline{\mu}|R_1)$. Moreover, if $\frac{\sigma_1^2}{n_1} = \ldots = \frac{\sigma_k^2}{n_k}$, it follows from Theorem 1 of [8] that $\sup_{\Omega_1} E(S|R_1) = \sup_{\Omega_0} E(S|R_1) = kP^*$, where $\Omega_0 = \{\underline{\mu}: \underline{\mu} = (\mu,\ldots,\mu),\ -\infty < \mu < \infty\}$.

We define below an invariance or symmetry property used by Gupta and Studden [11].

Let $X_1,\ldots,X_k$ be a set of independent observations from k populations $\pi_1,\ldots,\pi_k$, respectively, and let R be a procedure which selects π_i with probability $\delta_i(X_1,..,X_k)$. Then the procedure R is said to be invariant or symmetric if

$$\delta_i(X_1,\ldots,X_i,\ldots,X_j,\ldots,X_k) = \delta_j(X_1,\ldots,X_j,\ldots,X_i,\ldots,X_k)$$

for all i and j.

It follows from the result of Gupta and Studden [11] that when $\frac{\sigma_1^2}{n_1} = \ldots = \frac{\sigma_k^2}{n_k}$, the procedure R_1 is minimax in the class of invariant rules in the sense that for any other procedure R' in the class such that $\inf_{\underline{\mu} \in \Omega_1} P_{\underline{\mu}}(CS|R') \geq P^*$, we have

$$\sup_{\underline{\mu} \in \Omega_1} E_{\underline{\mu}}(S|R') \geq \sup_{\underline{\mu} \in \Omega_1} E_{\underline{\mu}}(S|R_1).$$

Next we consider another selection procedure as follows:

R_1': Select the population π_i if and only if

$$\bar{X}_i \geq \frac{1}{k-1} \sum_{j \neq i} \bar{X}_j - c_1' \sqrt{\frac{\sigma_i^2}{n_i} + \frac{1}{(k-1)^2} \sum_{j \neq i} \frac{\sigma_j^2}{n_j}} \tag{2.18}$$

where $c_1' = c_1'(P^*)$ is the smallest nonnegative number chosen so as to satisfy the P*-condition.

As is clear from the following derivations the constant c_1' associated with R_1' does not depend on k, nor does it depend on $(\sigma_1, n_1), \ldots, (\sigma_k, n_k)$.

$$P(CS|R_1') = P\{\bar{X}_{(k)} \geq \frac{1}{k-1} \sum_{j=1}^{k-1} \bar{X}_{(j)} - c_1' \sqrt{\frac{\sigma_{(k)}^2}{n_{(k)}} + \frac{1}{(k-1)^2} \sum_{j=1}^{k-1} \frac{\sigma_{(j)}^2}{n_{(j)}}}\} \tag{2.19}$$

$$= P\{Z \leq c_1' + \frac{\mu_{[k]} - \frac{1}{k-1} \sum_{j=1}^{k-1} \mu_{[j]}}{\sqrt{\frac{\sigma_{(k)}^2}{n_{(k)}} + \frac{1}{(k-1)^2} \sum_{j=1}^{k-1} \frac{\sigma_{(j)}^2}{n_{(j)}}}}\}$$

where Z denote the standard normal variate. It follows from (2.19) that the infimum of $P(CS|R_1')$ takes place when $\mu_{[1]} = \mu_{[2]} = \ldots = \mu_{[k]}$ and for any preassigned number P^*, $\frac{1}{k} < P^* < 1$, the selection constant c_1' is given by $c_1' = \phi^{-1}(P^*)$. Thus we have shown the following theorem.

Theorem 2.3. For the rule R_1',

$$\min_{(\sigma_1,n_1),\ldots,(\sigma_k,n_k)} \inf_{\Omega_1} P(CS|R_1') = \Phi(c_1'). \tag{2.20}$$

Let $p_i(\underline{\mu}|R_1')$ denote the probability that the population π_i is included in the selected subset. Then

$$p_i(\underline{\mu}|R_1') = P\{\bar{X}_i \geq \frac{1}{k-1}\sum_{j\neq i}\bar{X}_j - c_1'\sqrt{\frac{\sigma_i^2}{n_i} + \frac{1}{(k-1)^2}\sum\frac{\sigma_j^2}{n_j}}\} \tag{2.21}$$

$$= \Phi\left(c_1' + \frac{\mu_i - \frac{1}{k-1}\sum_{j\neq i}\mu_j}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{1}{(k-1)^2}\sum_{j\neq i}\frac{\sigma_j^2}{n_j}}}\right).$$

It follows from (2.15) that $p_i(\underline{\mu}|R_1')$ is increasing in μ_i when $\frac{\sigma_\ell^2}{n_\ell}$, $\ell = 1,\ldots,k$ and all other components of $\underline{\mu}$ are kept fixed; also $p_i(\underline{\mu}|R_1')$ is decreasing in μ_j $(j \neq i)$ when $\frac{\sigma_\ell^2}{n_\ell}$, $\ell = 1,\ldots,k$, and all other components of $\underline{\mu}$ are kept fixed. In particular, if $\frac{\sigma_i^2}{n_i} = \frac{\sigma_j^2}{n_j}$ and $\mu_i \leq \mu_j$, then $p_i(\underline{\mu}|R_1') \leq p_j(\underline{\mu}|R_1')$.

As before, let S denote the subset size of the selected subset, and let

$$\bar{\mu} = \frac{1}{k}\sum_{i=1}^{k}\mu_i$$

Theorem 2.4. If $\frac{\sigma_i^2}{n_i} = \frac{\sigma_j^2}{n_j}$ and $\bar{\mu} - \frac{k-1}{k}c_1'\sqrt{\frac{\sigma_i^2}{n_i} + \frac{1}{(k-1)^2}\sum_{\ell\neq i}\frac{\sigma_\ell^2}{n_\ell}} \leq \mu_i \leq \mu_j$, then the rate of change of $E_{\underline{\mu}}(S|R_1')$ with respect to μ_j is smaller than that with respect to μ_i. Or more precisely,

$$\frac{\partial}{\partial\mu_j}E_{\underline{\mu}}(S|R_1') \leq \frac{\partial}{\partial\mu_i}E_{\underline{\mu}}(S|R_1').$$

Proof. Since $E_{\underline{\mu}}(S|R_1') = \sum_{i=1}^{k} P(\pi_i \text{ is selected}|R_1')$

$$(2.22) \qquad = \sum_{i=1}^{k} \Phi\left(c_1' + \frac{\mu_i - \frac{1}{k-1}\sum_{\ell\neq i}\mu_\ell}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{1}{(k-1)^2}\sum_{\ell\neq i}\frac{\sigma_\ell^2}{n_\ell}}}\right)$$

Differentiating $E_{\underline{\mu}}(S|R_1')$ with respect to μ_j, we have

$$(2.23) \qquad \frac{\partial}{\partial\mu_j} E_{\underline{\mu}}(S|R_1') = -\sum_{\substack{\ell\neq j\\ \ell\neq i}} \frac{1}{k-1} \cdot \frac{1}{\sqrt{\frac{\sigma_\ell^2}{n_\ell} + \frac{1}{(k-1)^2}\sum_{r\neq\ell}\frac{\sigma_r^2}{n_r}}} \cdot \varphi\left(c_1' + \frac{\mu_\ell - \frac{1}{k-1}\sum_{r\neq\ell}\mu_r}{\sqrt{\frac{\sigma_\ell^2}{n_\ell} + \frac{1}{(k-1)^2}\sum_{r\neq\ell}\frac{\sigma_r^2}{n_r}}}\right)$$

$$+ \frac{1}{\sqrt{\frac{\sigma_j^2}{n_j} + \frac{1}{(k-1)^2}\sum_{\ell\neq j}\frac{\sigma_\ell^2}{n_\ell}}} \cdot \varphi\left(c_1' + \frac{\mu_j - \frac{1}{k-1}\sum_{\ell\neq j}\mu_\ell}{\sqrt{\frac{\sigma_j^2}{n_j} + \frac{1}{(k-1)^2}\sum_{\ell\neq j}\frac{\sigma_\ell^2}{n_\ell}}}\right)$$

$$- \frac{1}{k-1} \cdot \frac{1}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{1}{(k-1)^2}\sum_{\ell\neq i}\frac{\sigma_\ell^2}{n_\ell}}} \cdot \varphi\left(c_1' + \frac{\mu_i - \frac{1}{k-1}\sum_{\ell\neq i}\mu_\ell}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{1}{(k-1)^2}\sum_{\ell\neq i}\frac{\sigma_\ell^2}{n_\ell}}}\right)$$

where φ represents the pdf of the standard normal variate. Similarly, one can evaluate $\frac{\partial}{\partial\mu_j} E_{\underline{\mu}}(S|R_1')$. It follows that

$$\frac{\partial}{\partial\mu_j} E_{\underline{\mu}}(S|R_1') - \frac{\partial}{\partial\mu_i} E_{\underline{\mu}}(S|R_1')$$

$$= \frac{k}{k-1} \frac{1}{\sqrt{\frac{\sigma_j^2}{n_j} + \frac{1}{(k-1)^2}\sum_{\ell\neq j}\frac{\sigma_\ell^2}{n_\ell}}} \left\{\varphi\left(c_1' + \frac{\mu_j - \frac{1}{k-1}\sum_{\ell\neq j}\mu_\ell}{\sqrt{\frac{\sigma_j^2}{n_j} + \frac{1}{(k-1)^2}\sum_{\ell\neq j}\frac{\sigma_\ell^2}{n_\ell}}}\right)\right.$$

$$\left. - \varphi\left(c_1' + \frac{\mu_i - \frac{1}{k-1}\sum_{\ell\neq i}\mu_\ell}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{1}{(k-1)^2}\sum_{\ell\neq i}\frac{\sigma_\ell^2}{n_\ell}}}\right)\right\}$$

$$\leq 0.$$

This completes the proof of the theorem.

Recall that if $\underline{x} = (x_1,\ldots,x_k)$ and $\underline{y} = (y_1,\ldots,y_k)$ are such that

$$\sum_{i=1}^{m} x_{[k-i+1]} \geq \sum_{i=1}^{m} y_{[k-i+1]} \qquad \text{for } m = 1,\ldots,k-1$$

and

$$\sum_{i=1}^{k} x_{[k-i+1]} = \sum_{i=1}^{k} y_{[k-i+1]}$$

where $x_{[1]} \leq \ldots \leq x_{[k]}$ denotes the ordered values of x_i, then we say vector $\underline{x}$ majorizes vector $\underline{y}$ and write $\underline{x} \succ \underline{y}$ or equivalently, $\underline{y} \prec \underline{x}$. A real-valued function ϕ defined on the k-dimensional Euclidean space E^k is said to be Schur-convex (Schur-concave) if

$$\phi(\underline{x}) \geq (\leq)\phi(\underline{y}) \qquad \text{whenever } \underline{x} \succ \underline{y}. \tag{2.24}$$

We state without proof the following result which is due to Ostrowski [12].

Lemma 2.2. Let ϕ be a differentiable function defined on E^k. ϕ is Schur-concave if and only if

$$\frac{\partial}{\partial x_{[i]}}\phi(\underline{x}) - \frac{\partial}{\partial x_{[j]}}\phi(\underline{x}) \leq 0 \quad \text{for all } i > j, \text{ where } \underline{x} = (x_1,\ldots,x_k)$$

and $x_{[1]} \leq \ldots \leq x_{[k]}$ represents the ordered values of x_i.

Let $\Omega' = \{\underline{\mu}: \underline{\mu} = (\mu_1,\ldots,\mu_k),\ \mu_i \geq \bar{\mu} - \frac{k-1}{k} c_1' \sqrt{\frac{\sigma_i^2}{n_i} + \frac{1}{(k-1)^2}\sum_{\ell\neq i}\frac{\sigma_\ell^2}{n_\ell}},\ 1\leq i\leq k\}$.

Corollary 2.1. For the rule R_1', if $\frac{\sigma_1^2}{n_1} = \ldots = \frac{\sigma_k^2}{n_k}$

$$\sup_{\Omega'} E_{\underline{\mu}}(S|R_1') = kP^*.$$

Proof. Combining the results of Theorem 2.4 and Lemma 2.2, it follows that $E_{\underline{\mu}}(S|R_1')$ is a Schur-concave function in $\underline{\mu}$ over Ω'. Thus the supremum of $E_{\underline{\mu}}(S|R_1')$ over R_1' takes place when $\mu_1 = \ldots = \mu_k = \mu$. Moreover if

$\underline{\mu}_0 = (\mu,\ldots,\mu)$ $E_{\underline{\mu}_0}(S|R_1') = kP^*$. Thus $E_{\underline{\mu}_0}(S|R_1')$ does not depend on the common unknown μ.

Case (b): $\sigma_1^2,\ldots,\sigma_k^2$ unknown.

As pointed out earlier, this case presents more difficulty than the case in which $\sigma_1^2, \sigma_2^2,\ldots,\sigma_k^2$ are assumed known. We propose and investigate a selection procedure for this case as described below. For this problem it is necessary to require that $n_i \geq 2$ for all n_i where n_i is the total number of independent observations from π_i, $i = 1,2,\ldots,k$. We now define the selection procedure as follows:

Let $X_{i1}, X_{i2},\ldots,X_{in_i}$ be $n_i (\geq 2)$ independent random observations drawn from population π_i, $i = 1,2,\ldots,k$. Based on these observations, we calculate for $i = 1,2,\ldots,k$, the sample means and variances

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

$$S_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij}-\bar{X}_i)^2$$

where X_{ij} represents the jth observation from π_i. For any preassigned P^* $(\frac{1}{k} < P^* < 1)$, let $c_2 = c_2(k,P^*,n_1,\ldots,n_k)$ be the constant determined by the equation

$$\int_0^\infty \{\prod_{i=1}^{k-1} \int_0^\infty \Phi\left(\frac{c_2}{\sqrt{\frac{1}{x_i} + \frac{1}{y}}}\right) dG_{n_{[i]}}(x_i)\} dG_{n_{[1]}}(y) = P^* \tag{2.25}$$

where $G_i(\cdot)$ denotes the cdf of a chi-square random variate with (i-1) degrees of freedom. By using the fact that for a given integer n and any $\epsilon > 0$, there exist $\eta(\epsilon) > 0$ such that

$$\int_0^{\eta(\epsilon)} dG_n(x) < \epsilon,$$

and for $a > 0$,

$$\lim_{c\to\infty} \Phi(ac) = 1,$$

it follows that for any $P^*(\frac{1}{k} < P^* < 1)$, there always exist c_2 such that (2.25) holds. We propose a subset selection procedure as follows:

R_2: Retain the population π_i in the selected subset if and only if

$$\bar{X}_i \geq \max_{1\leq j\leq k} \bar{X}_j - c_2 \max_{1\leq j\leq k} S_j. \tag{2.26}$$

To obtain a lower bound for the infimum of $P(CS|R_2)$ we proceed as follows. Let $D = c_2 \max_{1\leq j\leq k} S_j$. Then

$$\begin{aligned} P(CS|R_2) &= P\{\bar{X}_{(k)} \geq \max_{1\leq j\leq k-1} \bar{X}_{(j)} - D\} \\ &= P\{Z_{ik} \leq (D + \mu_{[k]} - \mu_{[i]})(\frac{\sigma^2_{(k)}}{n_{(k)}} + \frac{\sigma^2_{(i)}}{n_{(i)}})^{-\frac{1}{2}}, \\ &\qquad i = 1,\ldots,k-1\}, \end{aligned} \tag{2.27}$$

where Z_{ik} is given by (2.5). Let $H(\cdot)$ denote the cdf of D and $a_i(t) = (t + \mu_{[k]} - \mu_{[i]})(\frac{\sigma^2_{(k)}}{n_{(k)}} + \frac{\sigma^2_{(i)}}{n_{(i)}})^{-\frac{1}{2}}$, $i = 1,\ldots,k$, then $P(CS|R_2)$ can be rewritten as

$$P(CS|R_2) = \int_0^\infty \{\int_{-\infty}^{a_{k-1}(t)} \cdots \int_{-\infty}^{a_1(t)} \frac{1}{(2\pi)^{\frac{k-1}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}\underline{x}'\Sigma^{-1}\underline{x}} dx_1 \ldots dx_{k-1}\} dH(t).$$

where $\underline{x}' = (x_1,\ldots,x_{k-1})$ and $\Sigma = (\rho^{(k)}_{ij})$. It follows from (2.27) that

$$\begin{aligned} P(CS|R_2) &\geq P\{Z_{ik} \leq D(\frac{\sigma^2_{(k)}}{n_{(k)}} + \frac{\sigma^2_{(i)}}{n_{(i)}})^{-\frac{1}{2}}, \ i = 1,2,\ldots,k-1\} \\ &\geq P\{Z_{ik} \leq c_2(\frac{\sigma^2_{(k)}}{n_{(k)}S^2_{(k)}} + \frac{\sigma^2_{(i)}}{n_{(i)}S^2_i})^{-\frac{1}{2}}, \ i = 1,2,\ldots,k-1\}. \end{aligned} \tag{2.28}$$

Denote the right hand member of (2.28) by T. Since the entities of the correlation matrix of $(Z_{1k}, Z_{2k},\ldots,Z_{k-1\ k})$ are all nonnegative, it follows that for given $S^2_{(1)}, S^2_{(2)},\ldots,S^2_{(k)}$, T is minimized when $\sigma_{(k)}$ approaches zero. Or more precisely

$$T \geq Pr\{Z^*_{ik} \leq c_2(\frac{\sigma^2_{(k)}}{n_{(k)}S^2_{(k)}} + \frac{\sigma^2_{(i)}}{n_{(i)}S^2_{(i)}})^{-\frac{1}{2}}, \ i = 1,2,\ldots,k-1\}$$

where $Z^*_{1k}, Z^*_{2k},\ldots,Z^*_{k-1\ k}$ are iid standard normal variates and are independent of $S_{(1)}, S_{(2)},\ldots,S_{(k)}$. Hence

$$(2.29)\qquad T \geq \int_0^\infty \{\prod_{i=1}^{k-1} \int_0^\infty \Phi\left(\frac{c_2}{\sqrt{\frac{1}{x_i}+\frac{1}{y}}}\right) dG_{n_{[i]}}(x_i)\} dG_{n_{[1]}}(y).$$

Thus we have proved the following theorem.

Theorem 2.3. If c_2 is defined by (2.25), then

$$\min_{n_1,\ldots,n_k} \inf_{\Omega_2} P(CS|R_2) \geq P^*$$

where $\Omega_2 = \{(\mu_1,\ldots,\mu_k;\ \sigma_1,\ldots,\sigma_k) \mid -\infty < \mu_i < \infty,\ \sigma_i > 0,\ i = 1,2,\ldots,k\}$.

Next we consider the expected subset size $E(S|R_2)$. It is given by

$$E(S|R_2) = \sum_{i=1}^{k} P(\pi_{(i)} \text{ is selected}|R_2)$$

$$= \sum_{i=1}^{k} P\{\max_{\substack{1\leq j\leq k \\ j\neq i}} (\bar{X}_{(j)} - \bar{X}_{(i)}) \leq D\}.$$

It is easy to see that

$$(2.30)\qquad E(S|R_2) \leq \frac{1}{k-1} \sum_{i-1}^{k} \sum_{\substack{j=1 \\ j\neq i}}^{k} \int_0^\infty \Phi\left(\frac{x+\mu_{[i]}-\mu_{[j]}}{\sqrt{\xi_{ij}}}\right) dH(x)$$

where $H(\cdot)$ represents the cdf of D, and $\xi_{ij} = \frac{\sigma^2_{(i)}}{n_{(i)}} + \frac{\sigma^2_{(j)}}{n_{(j)}}$, $i, j = 1,2,\ldots,k$, $i \neq j$. Denote the right hand member of (2.30) by $\frac{1}{k-1}\int_0^\infty Q_2 dH(x)$, and consider the configuration $\mu_{[1]}=\ldots=\mu_{[m]}=\mu \leq \mu_{[m+1]} \leq \ldots \leq \mu_{[k]}$, $1 \leq m \leq k-1$. Then Q_2 is given by

$$(2.31)\qquad Q_2 = \sum_{i=1}^{m} \{\sum_{\substack{j=1 \\ j\neq i}}^{m} \Phi\left(\frac{x}{\sqrt{\xi_{ij}}}\right) + \sum_{j=m+1}^{k} \Phi((x + \mu-\mu_{[j]})\ \xi_{ij}^{-\frac{1}{2}})\}$$

$$+ \sum_{i=m+1}^{k} \{\sum_{j=1}^{m} \Phi((x+\mu_{[i]}-\mu)\ \xi_{ij}^{-\frac{1}{2}}) + \sum_{\substack{j=m+1 \\ j\neq i}}^{k} \Phi((x+\mu_{[i]}-\mu_{[j]})\ \xi_{ij}^{-\frac{1}{2}})\}.$$

Keeping all parameters but μ fixed and differentiating Q_2 with respect to μ and interchanging the labels i and j in the sum $\sum_{i=1}^{m}\sum_{j=m+1}^{k}$, we obtain

$$(2.32)\qquad \frac{\partial Q_2}{\partial \mu} = \sum_{i=1}^{m}\sum_{j=m+1}^{k} \xi_{ij}^{-\frac{1}{2}}\{\varphi((x+\mu-\mu_{[j]})\xi_{ij}^{-\frac{1}{2}}) - \varphi((x+\mu_{[j]}-\mu)\xi_{ij}^{-\frac{1}{2}})\}$$

$$\geq 0.$$

This shows that

$$(2.33)\qquad Q_2 \leq \sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i}}^{k} \Phi\left(\frac{x}{\sqrt{\xi_{ij}}}\right),$$

and hence

$$E(S|R_2) \leq \frac{1}{k-1}\sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i}}^{k}\int_0^\infty\int_0^\infty \Phi\left(\frac{c_2}{\sqrt{\frac{1}{x}+\frac{1}{y}}}\right)dG_{n_i}(x)dG_{n_j}(y).$$

Thus we have shown the following theorem.

Theorem 2.4. For the rule R_2,

$$(2.34)\qquad \max_{n_1,\ldots,n_k}\sup_{\Omega_2} E(S|R_2) \leq \frac{1}{(k-1)}\sum_{1\leq i<j\leq k}\int_0^\infty\int_0^\infty \Phi\left(\frac{c_2}{\sqrt{\frac{1}{x}+\frac{1}{y}}}\right)dG_{n_i}(x)dG_{n_j}(y).$$

3. Selecting a Subset Which Contains All Populations Better Than a Standard.

In this section, we discuss a related selection problem.

Let $\pi_0, \pi_1,\ldots,\pi_k$ be k+1 independent normal populations with means $\mu_0, \mu_1,\ldots,\mu_k$ and variances $\sigma_0^2, \sigma_1^2,\ldots,\sigma_k^2$, respectively. It is assumed that $\mu_0, \mu_1,\ldots,\mu_k$ are unknown. The procedure described in this section controls the probability that the selected subset contains all those populations better than the standard ($\mu_i \geq \mu_0$), with the probability of a correct decision to be at least P*. Again, we discuss separately the following cases:

Case A. $\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2$ are known.

Let $\bar{X}_i$ denote the mean based on n_i independent observations taken from population π_i, $i = 0,1,\ldots,k$. We propose a procedure as follows:

R_3: Retain in the selected subset those and only those populations π_i $(i = 1,2,\ldots,k)$ for which

$$\bar{X}_i \geq \bar{X}_0 - c_3\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_i^2}{n_i}} \tag{3.1}$$

Let r_1 and r_2 denote the unknown number of populations with $\mu \geq \mu_0$ $\mu < \mu$, respectively, so that $r_1 + r_2 = k$. The probability of a correct decision (CD) is given by

$$P(CD|R_3) = P\left\{\bar{X}_{(i)} \geq \bar{X}_0 - c_3\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_{(i)}^2}{n_{(i)}}},\quad i = r_2+1,\ldots,k\right\} \tag{3.2}$$

$$= P\left\{Z_i \leq c_3 + \frac{\mu_{[i]} - \mu_0}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_{(i)}^2}{n_{(i)}}}},\quad i = r_2+1,\ldots,k\right\}$$

where $(Z_{r_2+1},\ldots,Z_k)$ is a r_1-variate normal random vector with zero mean vector, unit variances and correlation matrix (ρ_{ij}) where

$$\rho_{ij} = \left\{\left(1 + \frac{n_0}{n_{(i)}} \cdot \frac{\sigma_{(i)}^2}{\sigma_0^2}\right)\left(1 + \frac{n_0}{n_{(j)}} \cdot \frac{\sigma_{(j)}^2}{\sigma_0^2}\right)\right\}^{-\frac{1}{2}}, \quad i, j = r_2+1,\ldots,k, \; i \neq j,$$

(3.3)

$$\rho_{ii} = 1 \qquad , \quad i = r_2+1,\ldots,k.$$

By using the transformation (2.15), it follows that the infimum of $P(CD|R_3)$ is

$$\int_{-\infty}^{\infty} \prod_{i=1}^{k} \Phi\left(\frac{c_3 - \alpha_i x}{\sqrt{1-\alpha_i^2}}\right) d\Phi(x) \tag{3.4}$$

where

$$\alpha_i = \left(1 + \frac{n_0}{n_i}\frac{\sigma_i^2}{\sigma_0^2}\right)^{-\frac{1}{2}}, \qquad i = 1,2,\ldots,k. \tag{3.5}$$

Hence,

Theorem 3.1. For rule R_3,

$$\min_{(\sigma_1,n_1),\ldots,(\sigma_k,n_k)} \inf_{\Omega_3} P(CD|R_3) = \int_{-\infty}^{\infty} \prod_{i=1}^{k} \Phi\left(\frac{c_3 - \alpha_i x}{\sqrt{1-\alpha_i^2}}\right) d\Phi(x) \tag{3.6}$$

where $\Omega_3 = \{(\mu_0, \mu_1,\ldots,\mu_k): -\infty < \mu_i < \infty, i = 0,1,\ldots,k\}$.

Let S_3 denote the number of populations with means less than μ_0 that are included in the selected subset.

$$\begin{aligned} E(S_3|R_3) &= \sum_{i=1}^{r_2} P\,(\pi_{(i)} \text{ is selected}|R_3) \\ &= \sum_{i=1}^{r_2} P\left\{\bar{X}_{(i)} \geq \bar{X}_0 - c_3\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_{(i)}^2}{n_{(i)}}}\right\} \\ &= \sum_{i=1}^{r_2} P\left\{\frac{\bar{X}_0 - \bar{X}_{(i)} - \mu_0 + \mu_{[i]}}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_{(i)}^2}{n_{(i)}}}} \leq c_3 - \frac{\mu_0 - \mu_{[i]}}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_{(i)}^2}{n_{(i)}}}}\right\} \\ &\leq r_2\Phi(c_3) \\ &\leq k\Phi(c_3). \end{aligned}$$

Case B. $\sigma_0, \sigma_1,\ldots,\sigma_k$ unknown.

In this case, the problem is more complicated. We assume that $n_i \geq 2$ for $i = 0,1,\ldots,k$, as in case (b) of Section 2. Let $\bar{X}_i$ and S_i^2 denote the sample mean and sample variance based on n_i independent observations from population π_i, $i = 0,1,\ldots,k$. For any preassigned number P^* $(\frac{1}{k} < P^* < 1)$, let c_4 be the number determined by the following equation

$$\int_0^{\infty} \left\{\prod_{i=1}^{k} \int_0^{\infty} \Phi\left(\frac{c_4}{\sqrt{\frac{1}{x_i} + \frac{1}{y}}}\right) dG_{n_i}(x_i)\right\} dG_{n_0}(y) = P^*. \tag{3.6}$$

Let D be the random number such that $\max_{0\le i\le k} \frac{S_i c_4}{D} = 1$. We propose a selection procedure as follows:

R_4: Include population π_i in the selected subset if and only if

$$\bar{X}_i \ge \bar{X}_0 - D \tag{3.7}$$

The probability of a correct decision is given by

$$P(CD|R_4) = P\{\bar{X}_0 - \bar{X}_{(i)} \le D,\ i = r_2+1,\ldots,k\} \tag{3.8}$$

where r_2 denote the number of populations whose means are less than μ_0. Now $P(CD|R_4)$ can be expressed as follows:

$$P(CD|R_4) = P\{Z_i \le (D + \mu_{[i]} - \mu_0)\xi_i^{-\frac{1}{2}},\ i = r_2+1,\ldots,k\} \tag{3.9}$$

where

$$Z_i = (\bar{X}_0 - \bar{X}_{(i)} + \mu_{[i]} - \mu_0)\xi^{-\frac{1}{2}}$$

$$\xi_i = \frac{\sigma_0^2}{n_0} + \frac{\sigma_{(i)}^2}{n_{(i)}}. \tag{3.10}$$

Using the similar argument as in the proof of Theorem 2.2, we obtain that

$$P(CD|R_4) \ge \int_0^\infty \{\prod_{i=r_2+1}^{k} \int_0^\infty \Phi\left(\frac{c_4}{\sqrt{\frac{1}{x_i} + \frac{1}{y}}}\right) dG_{n_{(i)}}(x_i)\}dy. \tag{3.11}$$

Hence we have shown the following theorem.

Theorem 3.2. For procedure R_4, if c_4 is determined by (3.6), then

$$\min_{n_1,\ldots,n_k} \inf_{\Omega_4} P(CD|R_4) \ge P^*.$$

where $\Omega_4 = \{(\mu_0, \mu_1,\ldots,\mu_k; \sigma_0, \sigma_1,\ldots,\sigma_k):\ -\infty < \mu_i < \infty,\ \sigma_i > 0,\ i = 0,1,\ldots,k\}$.

Let S_4 be the number of populations with means less than μ_0 that are included in the selected subset.

$$E(S_4|R_4) = \sum_{i=1}^{r_2} P\{\bar{X}_{(i)} \geq \bar{X}_0 - D\}$$

$$= \sum_{i=1}^{r_2} P\left\{\frac{\bar{X}_0 - \bar{X}_{(i)} - \mu_0 + \mu_{[i]}}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_{(i)}^2}{n_{(i)}}}} \leq \frac{D - \mu_0 + \mu_{(i)}}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_{(i)}^2}{n_{(i)}}}}\right\}$$

$$\leq \sum_{i=1}^{r_2} P\left\{\frac{\bar{X}_0 - \bar{X}_{(i)} - \mu_0 + \mu_{[i]}}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_{(i)}^2}{n_{(i)}}}} \leq \frac{c_4}{\sqrt{\frac{\sigma_0^2}{n_0 S_0^2} + \frac{\sigma_{(i)}^2}{n_{(i)} S_{(i)}^2}}}\right\}.$$

This implies that if $r_1 + r_2 = k$,

$$E(S_4|R_4) \leq \sum_{i=1}^{r_2} \int_0^\infty \int_0^\infty \Phi\left(\frac{c_4}{\sqrt{\frac{1}{x_i} + \frac{1}{y}}}\right) dG_{n_{[r_1+i]}}(x_i) dG_{n_0}(y)$$

$$\leq \sum_{i=1}^{k} \int_0^\infty \int_0^\infty \Phi\left(\frac{c_4}{\sqrt{\frac{1}{x_i} + \frac{1}{y}}}\right) dG_{n_i}(x_i) dG_{n_0}(y).$$

4. Some comparisons of the procedures.

This section deals with some comparisons between the selection procedures R_1 and R_1'. Consider the configuration $\underline{\mu}$: $\mu_{[1]} = \mu - \delta < \mu_{[2]} = \ldots = \mu_{[k-1]} = \mu < \mu_{[k]} = \mu + \delta$. Without loss of generality, we may assume that $\mu = 0$. In this case, it is easy to show that

$$(4.1) \qquad E_{\underline{\mu}}(S|R_1') = (k-2)P^* + \Phi\left(c_1' - \frac{k}{k-1} \frac{\delta}{\sqrt{\frac{\sigma_{(1)}^2}{n_{(1)}} + \frac{1}{(k-1)^2} \sum_{i=2}^{k} \frac{\sigma_{(i)}^2}{n_{(i)}}}}\right)$$

$$+ \Phi(c_1' + \frac{k}{k-1} \frac{\delta}{\sqrt{\frac{\sigma^2_{(k)}}{n_{(k)}} + \frac{1}{(k-1)^2} \sum_{i=1}^{k-1} \frac{\sigma^2_{(i)}}{n_{(i)}}}}).$$

On the other hand,

$$E_{\underline{\mu}}(S|R_1) \leq 2 + (k-2)\Phi(c_1 - \frac{\delta}{\sqrt{(\frac{\sigma^2}{n})_{[k]} + (\frac{\sigma^2}{n})_{[k-1]}}})$$

Hence if $\delta > \sqrt{(\frac{\sigma^2}{n})_{[k]} + (\frac{\sigma^2}{n})_{[k-1]}} \{c_1 - \Phi^{-1}(P^* - \frac{2}{k-2})\}$, then $E_{\underline{\mu}}(S|R_1) < E_{\underline{\mu}}(S|R_1')$.

Next we consider the equally-spaced configuration:

$$\mu_{[i]} - \mu_{[1]} = \delta(i-1), \quad i = 2,\ldots,k.$$

Under this configuration,

$$(4.2) \qquad E_{\underline{\mu}}(S|R_1') = \sum_{i=1}^{k} \Phi(c_1' + \frac{\frac{k-2}{k-1}(i-1)\delta - \frac{k}{2}\delta}{\sqrt{\frac{\sigma^2_{(i)}}{n_{(i)}} + \frac{1}{(k-1)^2} \sum_{j \neq i} \frac{\sigma^2_{(j)}}{n_{(j)}}}}).$$

It is easy to see that

$$E_{\underline{\mu}}(S|R_1) \leq \sum_{i=1}^{k-2} \Phi(c_1 - \frac{(k-i)\delta}{\sqrt{\frac{\sigma^2_{(k)}}{n_{(k)}} + \frac{\sigma^2_{(i)}}{n_{(i)}}}}) + 2\,\Phi(c_1).$$

Hence $\delta > \sqrt{(\frac{\sigma^2}{n})_{[k]} + (\frac{\sigma^2}{n})_{[k-1]}} \{c_1 - \Phi^{-1}(\frac{k^2+k-4}{2(k-2)^2} P^* - \frac{2}{k-2} \Phi(c_1))\}$

then one can verify that $E_{\underline{\mu}}(S|R_1) < E_{\underline{\mu}}(S|R_1')$. In this sense the rule R_1 is better than rule R_1'.

Now suppose $\delta \to 0$, then $E_{\underline{\mu}}(S|R_1') \to k\Phi(c_1') = kP^*$. Whereas it follows from Section 2 that

$$P\{\pi_{(i)} \text{ is selected}|R_1\} \geq \inf_{\underline{\mu} \in \Omega_1} P_{\underline{\mu}}(CS|R_1).$$

If c_1 is determined by the equation

$$\min_{(\sigma_1,n_1),\ldots,(\sigma_k,n_k)} \inf_{\Omega_1} P(CS|R_1) = P^*$$

and $c_1' = \Phi^{-1}(P^*)$, then $E_{\underline{\mu}}(S|R_1) \geq E_{\underline{\mu}}(S|R_1')$ for $\underline{\mu} \in \Omega_0$.

For $n_1 = n_2 = \ldots = n_k = n$ and $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$ where σ^2 is assumed known, the rule R_1 reduces to the rule proposed by Gupta [7], [8]; namely

R_1: Select π_i if and only if

$$\bar{X}_i \geq \max_{1 \leq j \leq k} \bar{X}_j - \frac{d\sigma}{\sqrt{n}}$$

where $d = \sqrt{2}\, c_1$.

Also in this case R_1' can be written as

R_1': Select π_i if and only if

$$\bar{X}_i \geq \frac{1}{k-1} \sum_{j \neq i} \bar{X}_j - \frac{d'\sigma}{\sqrt{n}}$$

where $d' = c_1' \sqrt{\frac{k}{k-1}}$.

In his Purdue Ph.D. thesis Deverman [3] computed the efficiencies of R_1 and R_1' as defined by the ratio

$$\text{Eff}_{\underline{\mu}}(R) = \frac{P_{\underline{\mu}}(CS|R)}{E_{\underline{\mu}}(S|R)}.$$

Using the slippage configuration

$$\mu_{[1]} = \ldots = \mu_{[k-1]} = \mu_{[k]} - \delta$$

we give some of these numbers (excerpted from Table 6 of [3]) in the following table.

Efficiency of R_1 (top) and R_1' (bottom) in the case of equal sample sizes and a common known variance σ.

$P^* = .90$

k \ $\frac{\delta\sigma}{\sqrt{n}}$	.5	1.0	1.5	2.5	5.0
3	.3575	.3849	.4216	.5437	.9523
	.3724	.3778	.3991	.4533	.6909
4	.2680	.2873	.3138	.4107	.9032
	.2677	.2809	.2926	.3187	.4334
5	.2143	.2288	.2487	.3258	.8482
	.2142	.2236	.2311	.2463	.3068
6	.1784	.1900	.2055	.2680	.7918
	.1784	.1858	.1911	.2009	.2370
7	.1529	.1623	.1947	.2266	.7368
	.1529	.1589	.1663	.1698	.1934
8	.1337	.1416	.1521	.1957	.6847
	.1338	.1388	.1420	.1471	.1636

Since a larger value of the efficiency as defined above are desirable, the table seems to indicate that R_1 is more efficient than R_1' whenever δ or $\frac{\delta\sigma}{\sqrt{n}}$ is large (larger than .5 for $P^* = .90$), which is the same sort of conclusion that we found in the unequal sample sizes case, although there we looked at efficiency in terms of the expected size only.

5. k-Sample Behrens-Fisher Problem.

The Behrens-Fisher problem in its original simple version can be formulated as follows: Given two samples $X_{11},X_{12},\ldots,X_{1n_1}$ and $X_{21},X_{22},\ldots,X_{2n_2}$, it is assumed that the first sample comes from a normal distribution with mean μ_1 and variance σ_1^2 and that the second sample has arisen from a normal distribution with mean μ_2 and variance σ_2^2. The true values of μ's and σ's are not known and the sample sizes are not equal in general. The problem consists in describing with the inference about the actual value of the difference $\mu_1-\mu_2$ of the means. So far, no entirely satisfactory test for the Behrens-Fisher problem has been derived. When k = 2, several solutions to this problem were provided (see Pfanzagl [13]). Unfortunately none of these methods is applicable for the case when $k \geq 3$.

In this section, we demonstrate that the procedure given in Section 2 provides a solution for the Behren-Fisher problem when $k \geq 3$.

Now let $\pi_1, \pi_2,\ldots,\pi_k$ be k independent normal populations with means $\mu_1, \mu_2,\ldots,\mu_k$ and variances $\sigma_1^2, \sigma_2^2,\ldots,\sigma_k^2$, respectively. Suppose we are allowed to take $n_i (\geq 2)$ observations from each normal population π_i, $i = 1,2,\ldots,k$. Based on these observations, we wish to know whether μ_i are significantly different or not. The problem is to test the homogeneity of the means of the k normal populations. Let

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

$$S_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij}-\bar{X}_i)^2, \quad i = 1,2,\ldots,k.$$

Then our test rejects the hypothesis

$$H_0: \ \mu_1 = \mu_2=\ldots=\mu_k$$

when $\max\limits_{1\leq i\leq k} \bar{X}_i - \min\limits_{1\leq j\leq k} \bar{X}_j \geq D$, where D is given by $D = \max\limits_{1\leq j\leq k} S_i c$ and c is a constant such that the hypothesis of homogeneity will be rejected at

level α if the observed value of $\max_{1\le i\le k} \bar{X}_i - \min_{1\le j\le k} \bar{X}_j$ is greater than D.

Now using Theorem 2.3, we have the following theorem.

Theorem 5.1. For any α, $0 < \alpha < 1$, let $P^* = 1 - \frac{\alpha}{k}$ and let c be the constant determined by (2.25), then

$$(5.1) \qquad \max_{n_1,\ldots,n_k} \sup_{\Omega_0} P(H_0 \text{ is rejected}) \le \alpha$$

where $\Omega_0 = \{(\mu,\ldots,\mu;\sigma_1^2,\ldots,\sigma_k^2): -\infty<\mu<\infty, \sigma_i>0,\ i = 1,\ldots,k\}$

Proof:

$$\sup_{\Omega_0} P(H_0 \text{ is rejected})$$

$$= \sup_{\Omega_0} P\left\{\max_{1\le i\le k} \bar{X}_i - \min_{1\le j\le k} \bar{X}_j > D\right\}$$

$$= \sup_{\Omega_0} P\left\{\bar{X}_j < \max_{1\le i\le k} \bar{X}_i - D \text{ for some } 1 \le j \le k\right\}$$

$$\le k \sup_{\Omega_0} P\left\{\bar{X}_k < \max_{1\le i\le k} - D\right\}$$

$$= k\left\{1 - \inf_{\Omega_0} Pr\left\{\bar{X}_k \ge \max_{1\le i\le k} - D\right\}\right\}$$

$$\le k\left(1 - \left(1 - \frac{\alpha}{k}\right)\right)$$

$$= \alpha.$$

Special case k = 2:

Let Z denote the standard normal variate. Since under H_0,

$$P\ (H_0 \text{ is rejected})$$

$$= P\ (|\bar{X}_1 - \bar{X}_2| > D)$$

$$\le P\ \left(|Z| > \frac{c}{\sqrt{\frac{\sigma_1^2}{n_1 s_1^2} + \frac{\sigma_2^2}{n_2 s_2^2}}}\right)$$

$$(5.2) \qquad = 1 - \int_0^\infty \int_0^\infty \{2\Phi(\frac{c}{\sqrt{\frac{1}{x} + \frac{1}{y}}}) - 1\} dG_{n_1}(x) dG_{n_2}(y)$$

$$= \alpha(c) \text{ (say)}.$$

Under the alternatives

$$(5.3) \qquad \mu_1 = \mu_0,\ \mu_2 = \mu_0 + t(\frac{\sigma_{10}^2}{n_1} + \frac{\sigma_{20}^2}{n_2})^{\frac{1}{2}},\quad \sigma_i^2 = \sigma_{i0}^2,\ i = 1,2$$

The power of the test is

$$1 - P\{|\bar{X}_1 - \bar{X}_2| \leq D\}$$

$$\geq 1 - P\{|Z + t| \leq \frac{c}{\sqrt{\frac{\sigma_{10}^2}{n_1 S_1^2} + \frac{\sigma_{20}^2}{n_2 S_2^2}}}\}$$

where Z is a standard normal variate which is independent of S_1 and S_2. Hence the power under (5.3) exceeds

$$(5.4) \qquad 1 - \int_0^\infty \int_0^\infty \{\Phi(-t + \frac{c}{\sqrt{\frac{1}{x} + \frac{1}{y}}}) - \Phi(-t - \frac{c}{\sqrt{\frac{1}{x} + \frac{1}{y}}})\} dG_{n_1}(x) dG_{n_2}(x)$$

$$= \alpha(c,t) \text{ (say)}.$$

It is easy to see that $\alpha(c,t) \geq \alpha(c)$ for all nonnegative value of t. In other words, the test is unbiased.

5. Acknowledgement.

The authors wish to thank Dr. D. Y. Huang for some helpful conversations, and also Dr. R.L. Berger for his critical readirg and comments.

References

[1] Chen, H. J., Dudewicz, E. J., and Lee, Y. J. (1976). Subset selection procedures for normal means under unequal sample sizes. Sankhyā, Series B, 38, 249-255.

[2] Das Gupta, S., Eaton, M. L., Olkin, I., Perlman, M., Savage, L. J., and Sobel, M. (1970). Inequalities on the probability content of convex regions for elliptically contoured distributions. Proc. of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. II, 241-265.

[3] Deverman, J. N. (1969). A general selection procedure relative to the t-best populations. Ph.D. thesis, Department of Statistics, Purdue University.

[4] Dudewicz, E. J. (1972). Statistical inference with unknown and unequal variances. Trans. Ann. Qual. Contr. Conf., Rochester Quality Control, 28, 71-85.

[5] Dudewicz, E. J. (1974). A note on selection procedures with unequal observation numbers. Zastosowania Matematyki, 14, 31-35.

[6] Dudewicz, E. J., and Dalal, S. R. (1975). Allocation of observations in ranking and selection with unequal variance. Sankhyā, Ser. B, 37, 28-78.

[7] Gupta, S. S. (1956). On a decision rule for a problem in ranking means. Mimeograph Series No. 150, Inst. of Statist., University of North Carolina, Chapel Hill, North Carolina.

[8] Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. Technometrics, 7, 225-245.

[9] Gupta, S. S., and Huang, D. Y. (1976). Selection procedures for the means and variances of normal populations when the samples sizes are unequal. Sankhyā, Series B, 38, 111-129.

[10] Gupta, S. S., and Huang, W. T. (1974). A note on selecting a subset of normal populations with unequal sample sizes. Sankhyā, Ser. A, 36, 389-396.

[11] Gupta, S. S. and Studden, W. J. (1966). Some aspects of selection and ranking procedures with applications. Mimeograph Series No. 81, Dept. of Statistics, Purdue University, Lafayette, Ind. 47907.

[12] Ostrowski, A. (1952). Sur quelques applications des fonctions convexes et concaves au sens de I. Schur. J. Math. Pure Appl., 31, 253-292.

[13] Pfanzagl, J. (1974). On the Behrens-Fisher problem. Biometrika, 61, 39-47.

[14] Sitek, M. (1972). Application of the selection procedure R to unequal observation numbers. Zastosowania Matematyki, 12, 355-371.