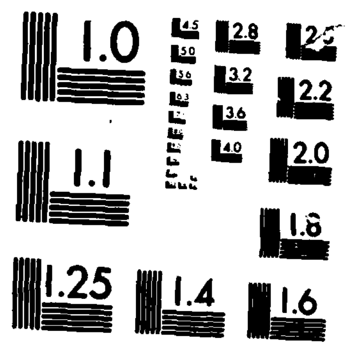
AD-A168 645 SELECTING THE BEST UNKNOWN MEAN FROM NORMAL POPULATIONS 1/1
HAVING A COMMON U..(U) PURDUE UNIV LAFAYETTE IN DEPT OF
STATISTICS S S GUPTA ET AL. MAY 86 TR-86-16

UNCLASSIFIED N00014-84-C-0167 F/G 12/1 NL



MICROCOP CHART

AD-A168 645

(1)

PURDUE UNIVERSITY

DEPARTMENT OF STATISTICS

(1)

Selecting the Best Unknown Mean from Normal Populations Having A Common Unknown Coefficient of Variation*

by

Shanti S. Gupta
Purdue University

TaChen Liang
Southern Illinois University
Technical Report#86-16

Department of Statistics
Purdue Univ rsity

May 1986

* This research was supported by the Office of Naval Research Contract N00014-84-C-0167 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

SELECTING THE BEST UNKNOWN MEAN FROM NORMAL POPULATIONS HAVING A COMMON UNKNOWN COEFFICIENT OF VARIATION*

Shanti S. Gupta
Purdue University
West Lafayette, Indiana, USA

TaChen Liang
Southern Illinois University
Carbondale, Illinois, USA

Abstract

This paper deals with the problem of selecting the population associated with the largest unknown mean from several normal populations having a common unknown coefficient of variation. Both subset selection and indifference zone approaches are studied. Based on the observed sample means and sample standard deviations, a subset selection rule is proposed. Some properties related to this selection rule are discussed. For the indifference zone approach, a two-stage elimination type selection rule is considered. If the experimenter has some prior knowledge about an upper bound on the unknown means, a modification is introduced to reduce the size of the selected subset at the first stage and also to reduce the sample size at the second stage. An example is provided which indicates that the saving of total sample size is quite significant if this prior knowledge is taken into consideration in designing the selection rule. It is shown how to implement the above selection rules by using several existing tables.

*This research was supported by the Office of Naval Research Contract N00014-84-C-0167 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

1. Introduction

The problem of selecting the population associated with the largest unknown mean has been extensively studied in the literature for normal populations. References may be found in Gupta and Panchapakesan (1979). These procedures are not applicable when the population standard deviations are proportional to the population means, a situation that is quite common in physical and biological applications. In the literature, among others, Gleser and Healy (1976) have discussed several best asymptotically normal estimators of a normal mean with a known coefficient of variation. In the area of ranking and selection, Tamhane (1978) used estimators developed by Gleser and Healy (1976) and proposed selection rules (through both subset selection and indifference zone approach) for normal populations having a common known coefficient of variation, and provided tables for implementing the rules in the large sample case. Gupta and Singh (1983) proposed a subset selection rule, based on sample variances, for selecting the population associated with the largest mean. Their selection rule is independent of the value of the common coefficient of variation, and therefore, can be applied to the situation when the value of common coefficient of variation is unknown.

In this paper, the problem of selecting the population with the largest mean from several normal populations having a common unknown coefficient of variation is studied. The statistical formulation of the problem is given in Section 2. In Section 3, a subset selection rule, based on the observed values of the sample means and sample standard deviations, is proposed. Some properties related to this selection rule are also discussed. In Section 4, a two-stage elimination type rule is given based on the indifference zone approach. If the experimenter has some prior knowledge about an upper bound on the unknown means, a modification is introduced which reduces the sample size at the second stage. To implement the above selection rules, the related tables are available from Gupta (1963), and Gupta, Panchapakesan and Sohn (1985).

2. Formulation of the Selection Problems

Let $\pi_1,\ldots,\pi_k$ be k (≥ 2) normal populations with positive means $\theta_1,\ldots,\theta_k$ and a common unknown coefficient of variation b. Let $\theta_{[1]} \leq \ldots \leq \theta_{[k]}$ be the ordered values of the means. We shall assume that the experimenter has no prior knowledge concerning the correct pairing between π_i and $\theta_{[j]}$ ($1 \leq i, j \leq k$). The population corresponding to $\theta_{[k]}$ will be referred as the best population. In the following, we will let $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_k)$ be the parameter vector and Ω be the parameter space.

Subset Selection Approach

According to this approach, the goal of the experimenter is to select a (small) subset of populations which contains the best population. The selection of any subset containing the best population is called a correct selection (CS). The decision-maker restricts attention only to those rules which guarantee the probability requirement (the so-called P*-condition) that

$$P_{\underset{\sim}{\theta}}\{CS\} \geq P^* \quad \text{for all } \underset{\sim}{\theta} \in \Omega, \tag{2.1}$$

where P^*, $k^{-1} < P^* < 1$, is a preassigned constant.

Indifference Zone Approach

According to this approach, the goal of the experimenter is to select the best population. The selection of the best population is called the correct selection (CS). For this approach, in order to specify the probability requirement, it is first necessary to define a measure of distance between two populations. We consider the following measure of distance: $\delta(\theta_i,\theta_j) = \theta_i-\theta_j$, the difference between the two population means of π_i and π_j. Note that this measure of distance is different from that considered by Tamhane (1978). For a preassigned value $\delta^* > 0$, let $\Omega(\delta^*) = \{\underset{\sim}{\theta} \in \Omega \mid \theta_{[k]} \geq \theta_{[k-1]} + \delta^*\}$. $\Omega(\delta^*)$ is known as the preference zone and its complement as indifference zone. The assignment of δ^* value will be based on the experimenter's prior knowledge. The experimenter restricts consideration only to those rules which guarantee the probability requirement that

$$P_{\underset{\sim}{\theta}}\{CS\} \geq P^* \quad \text{for all } \underset{\sim}{\theta} \in \Omega(\delta^*), \tag{2.2}$$

where P^*, $k^{-1} < P^* < 1$, is a preassigned constant.

3. Subset Selection Approach

The goal here is to select, on the basis of an independent random sample X_{ij}, $j = 1,\ldots,n$, from each π_i $(1 \leq i \leq k)$, a subset containing the best population with the minimum probability P^* of a correct selection.

Subset Selection Rule R_1

For each i, let $\bar{X}_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij}$ and $S_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$.

Based on $(\bar{X}_i, S_i)$, $1 \leq i \leq k$, we propose a subset selection rule R_1 as follows:

$$R_1: \quad \text{Select } \pi_i \text{ if } \bar{X}_i \geq \bar{X}_j - \sqrt{\frac{2}{n}}\, cS_i \quad \text{for } j \neq i. \tag{3.1}$$

The constant $c > 0$ is determined by

$$P\{Z_j \leq cW \quad \text{for } 1 \leq j \leq k-1\} = P^* \tag{3.2}$$

where

$$\begin{cases} Z_j \sim N(0,1), \quad 1 \le j \le k-1; \\ \mathrm{cov}(Z_i, Z_j) = 1/2, \quad \text{for } i \neq j; \\ (n-1)W^2 \sim \chi^2(n-1); \quad \text{and} \\ (Z_1, \ldots, Z_{k-1}) \text{ and } W \text{ are independent.} \end{cases} \tag{3.3}$$

The value of c satisfying (3.2) is available from Gupta and Sobel (1957), and Gupta, Panchapakesan and Sohn (1985).

Probability of a Correct Selection Associated with R_1

Let $P_{\underset{\sim}{\theta}}\{CS|R\}$ be the probability of correct selection when $\underset{\sim}{\theta}$ is the true parameter vector and the selection rule $\underset{\sim}{R}$ is applied. Then, we have

<u>Theorem 3.1</u>. $P_{\underset{\sim}{\theta}}\{CS|R_1\} \ge P^*$ for all $\underset{\sim}{\theta} \in \Omega$.

Proof: Without loss of generality, we can assume that $\theta_k = \theta_{[k]}$. Following a straightforward computation,

$$\begin{aligned} & P_{\underset{\sim}{\theta}}\{CS|R_1\} \\ &= P_{\underset{\sim}{\theta}}\{\bar{X}_k \ge \bar{X}_j - \sqrt{2/n}\, cS_k \quad \text{for } j \neq k\} \\ &= P_{\underset{\sim}{\theta}}\{Y_j \le \sqrt{n}\, b^{-1}\alpha_j(\underset{\sim}{\theta}) + \sqrt{2}\, c\beta_j(\underset{\sim}{\theta})W \quad \text{for } 1 \le j \le k-1\}, \end{aligned} \tag{3.4}$$

where

$$\begin{cases} Y_j = \sqrt{n}(\bar{X}_j - \bar{X}_k - \theta_j + \theta_k)b^{-1}(\theta_k^2 + \theta_j^2)^{-\frac{1}{2}} \sim N(0,1), \; 1 \le j \le k-1 \\ \mathrm{cov}(Y_i, Y_j) = \theta_k^2[(\theta_k^2+\theta_i^2)(\theta_k^2+\theta_j^2)]^{-\frac{1}{2}} \equiv \rho_{ij}(\underset{\sim}{\theta}), \; i \neq j \\ W = S_k/(b\theta_k), (n-1)W^2 \sim \chi^2(n-1) \text{ and} \\ (Y_1, \ldots, Y_{k-1}) \text{ and } W \text{ are independent;} \end{cases} \tag{3.5}$$

and

$$\begin{cases} \alpha_j(\underset{\sim}{\theta}) = (\theta_k-\theta_j)(\theta_k^2+\theta_j^2)^{-\frac{1}{2}}, & 1 \le j \le k-1; \\ \beta_j(\underset{\sim}{\theta}) = \theta_k(\theta_k^2+\theta_j^2)^{-\frac{1}{2}}, & 1 \le j \le k-1. \end{cases} \tag{3.6}$$

Note that $\alpha_j(\underset{\sim}{\theta}) \ge 0$, $\beta_j(\underset{\sim}{\theta}) \ge 1/\sqrt{2}$ and $\rho_{ij}(\underset{\sim}{\theta}) \ge \frac{1}{2}$ since $\theta_k \ge \theta_j$ for $j \ne k$. Then by Slepian's inequality, we conclude that the smallest value of the constant c is given by

$$P_{\underset{\sim}{\theta}}\{CS|R_1\} \ge P\{Z_j \le cW \text{ for } 1 \le j \le k-1\} = P^* \quad \text{for } \underset{\sim}{\theta} \in \Omega,$$

where Z_j $(1 \le j \le k-1)$ are defined in (3.3).

Least Favorable Configuration

Let $\Omega_0 = \{\underset{\sim}{\theta}|\theta_1 = \ldots = \theta_k > 0\}$. For each $\underset{\sim}{\theta} \in \Omega$, $\theta_k = \theta_{[k]}$ being fixed, both $\alpha_j(\underset{\sim}{\theta})$ and $\beta_j(\underset{\sim}{\theta})$ are decreasing in θ_j and $\alpha_j(\underset{\sim}{\theta}) = 0$, $\beta_j(\underset{\sim}{\theta}) = 1/\sqrt{2}$ when $\theta_j = \theta_k$. Then by Slepian's inequality, we can see that

$$\inf_{\underset{\sim}{\theta}\in\Omega} P_{\underset{\sim}{\theta}}\{CS|R_1\} = \inf_{\underset{\sim}{\theta}\in\Omega_0} P_{\underset{\sim}{\theta}}\{CS|R_1\} = P_{\underset{\sim}{\theta}_0}\{CS|R_1\}$$

for any $\underset{\sim}{\theta}_0 \in \Omega_0$. Note that $P_{\underset{\sim}{\theta}_0}\{CS|R_1\}$ does not depend on $\underset{\sim}{\theta}_0$.

Some Properties of R_1

Property 1. For fixed $\underset{\sim}{\theta} \in \Omega$, $P_{\underset{\sim}{\theta}}\{CS|R_1\}$ is decreasing in b for $b > 0$.

This is obvious from (3.4) since $\alpha_j(\underset{\sim}{\theta}) \ge 0$ and the distributions of $(Y_1,\ldots,Y_{k-1})$ and W are independent of the parameter b.

For fixed b, suppose that $(\theta_1,\ldots,\theta_k) = (ac_1 + d_1,\ldots,ac_k + d_k)$, where $c_i > 0$, a, $d_i \geq 0$ and $ac_i + d_i > 0$ for each $i = 1,\ldots,k$. Also, we assume that $c_k \geq c_j$, $d_k \geq d_j$ for $j \neq k$. Thus, $\theta_k \geq \theta_j$ for $j \neq k$. Then,

$$P_{\underset{\sim}{\theta}}\{CS|R_1\} = P_{\underset{\sim}{\theta}}\{Y_j \leq \sqrt{n}\, g_j(a)b^{-1} + \sqrt{2}ch_j(a)W \quad \text{for } j \neq k\}, \tag{3.7}$$

where $g_j(a) = [a(c_k-c_j)+(d_k-d_j)][(ac_k+d_k)^2+(ac_j+d_j)^2]^{-\frac{1}{2}}$;

$$h_j(a) = (ac_k+d_k)[(ac_k+d_k)^2+(ac_j+d_j)^2]^{-\frac{1}{2}}.$$

<u>Lemma 3.1</u>. The following three statements are equivalent.

i) $\frac{d_k}{c_k} \leq \frac{d_j}{c_j}$.

ii) $g_j(a)$ is increasing in a for $a > 0$.

iii) $h_j(a)$ is increasing in a for $a > 0$.

From (3.7) and Lemma 3.1, we have the following results.

<u>Property 2</u>. Let b be a fixed constant, then

a) if $\frac{d_k}{c_k} \geq \frac{d_j}{c_j}$ for all $j \neq k$, then $P_{\underset{\sim}{\theta}}\{CS|R_1\}$ is decreasing in a; and

b) if $\frac{d_k}{c_k} \le \frac{d_j}{c_j}$ for all $j \ne k$, then $P_{\underset{\sim}{\theta}}\{CS|R_1\}$ is increasing in a.

The following two situations are special cases of Property 2.

c) When $c_1 = c_2 = \ldots = c_k$ and $d_1,\ldots,d_k$ are fixed constants, then $P_{\underset{\sim}{\theta}}\{CS|R_1\}$ is decreasing in a.

d) When $d_1 = \ldots = d_k = 0$ and $c_1,\ldots,c_k$ are fixed constants, then $P_{\underset{\sim}{\theta}}\{CS|R_1\}$ is a constant, which is independent of the parameter a.

Property 3. Monotonicity of R_1

Let $P_{\underset{\sim}{\theta}}\{\pi_i|R_1\}$ be the probability of including π_i in the selected subset when rule R_1 is applied and $\underset{\sim}{\theta}$ is the true parameter vector. Then,

Theorem 3.2. If $\theta_i \ge \theta_j$, then $P_{\underset{\sim}{\theta}}\{\pi_i|R_1\} \ge P_{\underset{\sim}{\theta}}\{\pi_j|R_1\}$.

Proof: Without loss of generality, let $\theta_2 \ge \theta_1$, and we will prove that $P_{\underset{\sim}{\theta}}\{\pi_2|R_1\} \ge P_{\underset{\sim}{\theta}}\{\pi_1|R_1\}$. Straightforward computations show that

$$P_{\underset{\sim}{\theta}}\{\pi_i|R_1\}$$

$$= P_{\underset{\sim}{\theta}}\left\{\begin{array}{l} Y_{ij} \le \sqrt{n}(\theta_i-\theta_j)b^{-1}(\theta_i^2+\theta_j^2)^{-\frac{1}{2}}+\sqrt{2}c\theta_i(\theta_i^2+\theta_j^2)^{-\frac{1}{2}}W \text{ and} \\ \\ Y_{im} \le \sqrt{n}(\theta_i-\theta_m)b^{-1}(\theta_i^2+\theta_m^2)^{-\frac{1}{2}}+\sqrt{2}c\theta_i(\theta_i^2+\theta_m^2)^{-\frac{1}{2}}W \text{ for} \end{array}\right\}$$

$$3 \le m \le k$$

for $i,j = 1,2$, $i \neq j$, where $\mathrm{cov}(Y_{im},Y_{i\ell}) = \theta_i^2[(\theta_i^2+\theta_m^2)(\theta_i^2+\theta_\ell^2)]^{-\frac{1}{2}}$, $1 \leq \ell, m \leq k$, $\ell \neq m$, $\ell \neq i$, $m \neq i$.

We then conclude the proof of this theorem by an application of Slepian's inequality and by noting the following facts:

i) $(\theta-\theta_m)(\theta^2+\theta_m^2)^{-\frac{1}{2}}$ is increasing in θ for $\theta > 0$ for each m;

ii) $\theta(\theta^2+\theta_m^2)^{-\frac{1}{2}}$ is increasing in θ for $\theta > 0$ for each m; and

iii) $\mathrm{cov}(Y_{2m}, Y_{2\ell}) \geq \mathrm{cov}(Y_{1m},Y_{1\ell})$ for $3 \leq m, \ell \leq k$, $m \neq \ell$ and $\mathrm{cov}(Y_{21}, Y_{2m}) \geq \mathrm{cov}(Y_{12}, Y_{1m})$ for $3 \leq m \leq k$.

Expected Size of Selected Subset

Note the selection rule R_1 selects a non-empty subset, the size of which is not fixed in advance but depends on the outcome of the experiment. Hence, as a measure of performance of the selection rule R_1, we can consider the expected size of the selected subset, say $E_{\underset{\sim}{\theta}}(S|R_1)$. We have the following expression:

$$
\begin{aligned}
E_{\underset{\sim}{\theta}}(S|R_1) &= \sum_{i=1}^{k} P_{\underset{\sim}{\theta}}\{\pi_i|R_1\} \\
&= \sum_{i=1}^{k} P_{\underset{\sim}{\theta}}\{V_i \geq V_j\theta_j\theta_i^{-1}+\sqrt{n}b^{-1}(\theta_j\theta_i^{-1}-1)-\sqrt{2}cW \text{ for } j \neq i\},
\end{aligned} \tag{3.8}
$$

where V_i, $1 \leq i \leq k$, are iid having standard normal distribution and W is as defined in (3.3). It is often of interest to identify the parameter configuration where the supremum of (3.8) occurs. For $k = 2$, let $\theta_2 = \theta$ and $\theta_1 = \Delta\theta$, $0 < \Delta \leq 1$. Then,

$$E_{\underset{\sim}{\theta}}(S|R_1) = \int_0^\infty \Phi(\sqrt{n}b^{-1}(\Delta-1)(1+\Delta^2)^{-\frac{1}{2}}+\sqrt{2}\Delta c(1+\Delta^2)^{-\frac{1}{2}}w)\,dF_W(w)$$

$$+ \int_0^\infty \Phi(\sqrt{n}b^{-1}(1-\Delta)(1+\Delta^2)^{-\frac{1}{2}}+\sqrt{2}c(1+\Delta^2)^{-\frac{1}{2}}w)\,dF_W(w),$$

where $\Phi(\cdot)$ is the standard normal distribution and $F_W(\cdot)$ is the distribution of random variable W.

We see that $\sup_{\underset{\sim}{\theta}\in\Omega} E_{\underset{\sim}{\theta}}(S|R_1) = 2P^*$ and the supremum occurs when $\Delta = 1$. For $k > 2$, it appears difficult to obtain such a result. We can only state that $\sup_{\underset{\sim}{\theta}\in\Omega} E_{\underset{\sim}{\theta}}(S|R_1) \geq \sup_{\underset{\sim}{\theta}\in\Omega_0} E_{\underset{\sim}{\theta}}(S|R_1) = kP^*$.

Remark

For the case when b is known, Tamhane (1978), proposed a subset selection rule for the best population using estimators of Gleser and Healy (1976). He also provided tables for implementing the rule in the large sample case. However, for small sample sizes, tables for implementing Tamhane's rule are not available. It has been pointed out by Tamhane (1978) that for certain values of b, k, n and P*, Tamhane's rule does not exist. Based on sample variances, Gupta and Singh (1983) also proposed a subset selection rule for the problem of selecting the best population. Their selection rule and the associated probability of correct selection are independent of the value of the common coefficient of variation b, and hence, can be applied to the situation when this value is unknown. They made some comparison between Tamhane's and their rules. It is found that in terms of expected size of selected subset, the performance of Gupta and Singh's rule is a little inferior to that of Tamhane's rule. For k = 2 and large sample size, there is not much difference between these two rules.

Note that the proposed rule R_1 is independent of the value of the common coefficient of variation b. However, from (3.4), we can see that the associated probability of a correct selection depends on b for each fixed $\underset{\sim}{\theta} \in \Omega$. Therefore, it is interesting to compare the performance of rule R_1 with Gupta and Singh's rule. Further study of this comparison is to be carried out.

4. Indifference Zone Approach

The goal here is to derive selection rules which will select the best population with a guaranteed probability P*. On $\Omega(\delta^*)$, the associated measure of distance between populations π_i and π_j is $\delta(\theta_i,\theta_j) = \theta_i-\theta_j$, which is different from the one considered by Tamhane (1978). Since both $\underset{\sim}{\theta}$ and the common coefficient of variation b are unknown, on $\Omega(\delta^*)$, it is impossible to construct a single-stage selection rule which guarantees the probability requirement of (2.2). In the following, a two-stage elimination type selection rule is proposed.

Two-Stage Elimination Type Selection Rule R_2

Stage 1: Take $n_0(\geq 2)$ independent observations $X_{ij}(j = 1,\ldots,n_0)$ from each π_i $(i = 1,\ldots,k)$, and compute the sample mean $\bar{X}_i^{(1)} = \frac{1}{n_0}\sum_{j=1}^{n_0} X_{ij}$ and sample variance $S_i^2 = \frac{1}{n_0-1}\sum_{j=1}^{n_0}(X_{ij}-\bar{X}_i^{(1)})^2$. Then determine the set

$$A = \{i \mid \bar{X}_i^{(1)} \geq \bar{X}_j^{(1)} - (\sqrt{2}dS_i/\sqrt{n_0} - \delta^*)^+ \text{ for } j \neq i\}, \quad (4.1)$$

where $y^+ = \max(y,0)$ and d is a positive constant chosen to satisfy (2.2).

If A contains only one element, then stop sampling and assert that the population associated with $\max_{1 \le j \le k} \bar{X}_j^{(1)}$ is the best.

If A has more than one element, then proceed to the second stage.

Stage 2: Let $S_A = \max_{i \in A} S_i$. Determine

$N = \max\{n_0, [2(dS_A/\delta^*)^2]^*\}$, where $[y]^*$ denotes the smallest integer $\geq$ y. Take $N-n_0$ additional observations X_{ij} from each π_i $(i \in A)$ if necessary. Then compute the overall sample means $\bar{X}_i = \frac{1}{N} \sum_{j=1}^{N} X_{ij}$ $(i \in A)$ and assert that the population associated with $\max_{i \in A} \bar{X}_i$ is the best.

Note: This selection rule is essentially of the same type as that of Gupta and Kim (1984), even though we have a different type of screening procedure and different way to determine the value N. The difference is due to the fact that in this paper, the concerned population standard deviations are proportional to respective population means while Gupta and Kim (1984) considered a common unknown standard deviation.

Probability of Correct Selection

Without loss of generality, we still assume that π_k is the best population. Therefore, $\theta_k \geq \theta_j + \delta^*$ for $j \neq k$. Let B be any subset of $\{1,2,\ldots,k\}$ containing the element k; and let E(B) be the event that subset B is selected at the first stage. Also, let $\mathcal{C} = \{B \subset \{1,\ldots,k\} | k \in B\}$. Then,

$$\bigcup_{B\in C} E(B) = \{\bar{X}_k^{(1)} \geq \bar{X}_j^{(1)} - (\frac{\sqrt{2}dS_k}{\sqrt{n_0}} - \delta^*)^+ \text{ for } j \neq k\}. \quad (4.2)$$

Also,

$$\begin{aligned} P_{\underset{\sim}{\theta}}\{CS|R_2\} &= \sum_{B\in C} P_{\underset{\sim}{\theta}}\{CS \cap E(B)\} \\ &= \sum_{B\in C} P_{\underset{\sim}{\theta}}\{CS|E(B)\}P_{\underset{\sim}{\theta}}\{E(B)\}. \end{aligned} \quad (4.3)$$

Here $P_{\underset{\sim}{\theta}}\{CS|E(B)\}$ denotes the conditional probability of CS given E(B).

Since $\theta_k \geq \theta_j + \delta^*$ for $j \neq k$, for each $B \in C$,

$$\begin{aligned} &P_{\underset{\sim}{\theta}}\{CS|E(B)\} \\ &= P_{\underset{\sim}{\theta}}\{\bar{X}_k \geq \bar{X}_j \text{ for } j \in B-\{k\}|E(B)\} \quad (4.4) \\ &\geq P_{\underset{\sim}{\theta}}\{Y_j \leq \sqrt{N}\delta^* b^{-1}(\theta_j^2+\theta_k^2)^{-\frac{1}{2}} \text{ for } j \in B-\{k\}|E(B)\}, \end{aligned}$$

where

$$Y_j = \sqrt{N}[(\bar{X}_j-\bar{X}_k)-(\theta_j-\theta_k)](b(\theta_j^2+\theta_k^2)^{\frac{1}{2}})^{-1} \sim N(0,1), \quad 1 \leq j \leq k-1;$$

$\text{cov}(Y_i,Y_j) = \rho_{ij}(\underset{\sim}{\theta})$ which is defined in (3.5) for $i \neq j$.

Conditional on E(B), $\sqrt{N}\delta^* \geq \sqrt{2}dS_B \geq \sqrt{2}dS_k$. Then, it follows that

$$\begin{aligned} &P_{\underset{\sim}{\theta}}\{CS|E(B)\} \\ &\geq P_{\underset{\sim}{\theta}}\{Y_j \leq \sqrt{2}\beta_j(\underset{\sim}{\theta})dW \text{ for } j \in B-\{k\}|E(B)\} \quad (4.5) \\ &\geq P_{\underset{\sim}{\theta}}\{Y_j \leq \sqrt{2}\beta_j(\underset{\sim}{\theta})dW \text{ for } 1 \leq j \leq k-1|E(B)\} \end{aligned}$$

where $W = S_k/(b\theta_k)$ and $\beta_j(\underset{\sim}{\theta})$ is defined in (3.6).

Let $Y_j^{(1)} = \sqrt{n_0}[(\bar{X}_j^{(1)}-\bar{X}_k^{(1)})-(\theta_j-\theta_k)](b(\theta_j^2+\theta_k^2)^{\frac{1}{2}})^{-1}$,

$1 \leq j \leq k-1.$

Then, $Y_j^{(1)} \sim N(0,1)$, $\mathrm{cov}(Y_i^{(1)},Y_j^{(1)}) = \rho_{ij}(\underset{\sim}{\theta})$, $i \neq j$. Also, $\mathrm{cov}(Y_i,Y_j^{(1)}) \geq 0$, $1 \leq i, j \leq k-1$. It should be pointed out that $(Y_1,\ldots,Y_{k-1})$ and $(Y_1^{(1)},\ldots,Y_{k-1}^{(1)})$ are identically distributed, and $(Y_1,\ldots,Y_{k-1})$ and $(Y_1^{(1)},\ldots,Y_{k-1}^{(1)})$ are independent of W.

For fixed $\theta_k(=\theta_{[k]})$, let $\underset{\sim}{\theta}^k = (\theta_1^k,\ldots,\theta_k^k)$, where $\theta_k^k = \theta_k$ and $\theta_j^k = \theta_k-\delta^*$ for $j \neq k$. Then,

$$\begin{cases} \rho_{ij}(\underset{\sim}{\theta}) \geq \rho_{ij}(\underset{\sim}{\theta}^k); \\ \beta_j(\underset{\sim}{\theta}) \geq \beta_j(\underset{\sim}{\theta}^k). \end{cases} \tag{4.6}$$

Note that both $\rho_{ij}(\underset{\sim}{\theta}^k)$ and $\beta_j(\underset{\sim}{\theta}^k)$ are decreasing in θ_k; $\rho_{ij}(\underset{\sim}{\theta}^k) \to \frac{1}{2}$, $\beta_j(\underset{\sim}{\theta}^k) \to \frac{1}{\sqrt{2}}$ as $\theta_k \to \infty$.

From (4.2) to (4.6) together with the facts that $\theta_k \geq \theta_j + \delta^*$ for $j \neq k$, $\mathrm{cov}(Y_i,Y_j^{(1)}) \geq 0$, $\sqrt{n_0}\,\delta^* + (\sqrt{2}dS_k - \sqrt{n_0}\delta^*)^+ \geq \sqrt{2}dS_k$ and repeated applications of Slepian's inequality, straightforward computations lead to the following result:

$$P_{\underset{\sim}{\theta}}\{CS|R_2\}$$

$$\geq P_{\underset{\sim}{\theta}}\{Y_j \leq \sqrt{2}\beta_j(\underset{\sim}{\theta})dW,\ Y_j^{(1)} \leq \sqrt{2}\beta_j(\underset{\sim}{\theta})dW \text{ for } 1 \leq j \leq k-1\}$$

$$= \int_0^\infty P_{\underset{\sim}{\theta}}\{Y_j \leq \sqrt{2}\beta_j(\underset{\sim}{\theta})dw,\ Y_j^{(1)} \leq \sqrt{2}\beta_j(\underset{\sim}{\theta})dw \text{ for } 1 \leq j \leq k-1\}dF_W(w)$$

$$\geq \int_0^\infty [P_{\underset{\sim}{\theta}}\{Y_j \leq \sqrt{2}\beta_j(\underset{\sim}{\theta})dw \text{ for } 1 \leq j \leq k-1\}]^2 dF_W(w)$$

$$\geq \int_0^\infty [P_{\underset{\sim}{\theta}^k}\{Y_j \leq \sqrt{2}\beta_j(\underset{\sim}{\theta}^k)dw \text{ for } 1 \leq j \leq k-1\}]^2 dF_W(w)$$

$$\geq [\int_0^\infty P_{\underset{\sim}{\theta}^k}\{Y_j \leq \sqrt{2}\beta_j(\underset{\sim}{\theta}^k)dw \text{ for } 1 \leq j \leq k-1\}dF_W(w)]^2$$

$$\geq [\int_0^\infty P\{Z_j \leq dw \text{ for } 1 \leq j \leq k-1\}dF_W(w)]^2.$$

$$= [P\{Z_j \leq dW \text{ for } 1 \leq j \leq k-1\}]^2$$

where $(Z_1,\ldots,Z_{k-1})$ are defined in (3.3), and distributed independently of W.

For given P*, we can choose the value d so that

$$P\{Z_j \leq dW \text{ for } 1 \leq j \leq k-1\} = \sqrt{P^*}. \qquad (4.8)$$

Therefore, the probability requirement of (2.2) will be satisfied. For some specified values of P*, k and n_0, the corresponding d value can be found in Gupta and Sobel (1957) and Gupta, Panchapakesan and Sohn (1985).

A Modified Two-Stage Selection Rule $R_2(\theta^*)$

In practical situations, the experimenter sometimes may have some prior knowledge about an upper bound on $\theta_{[k]}$, say θ^*. This knowledge can be used to reduce the sample size taken at the second stage.

Let $\underset{\sim}{\theta}^*(\delta^*) = (\theta_1^*,\ldots,\theta_k^*)$ where $\theta_{[k]}^* = \theta^*$ and $\theta_{[j]}^* = \theta^*-\delta^*$ for $j \neq k$. Let $\rho^* = \theta^{*2}(\theta^{*2} + (\theta^*-\delta^*)^2)^{-1}$ and $\beta^* = \theta^*(\theta^{*2} + (\theta^*-\delta^*)^2)^{-\frac{1}{2}}$. Let $Y_i^*(1 \leq i \leq k-1)$ be standard normal random variables having $\mathrm{cov}(Y_i^*,Y_j^*) = \rho^*$ for $i \neq j$, and W be a random variable distributed independently of $(Y_1^*,\ldots,Y_{k-1}^*)$ with $(n_0-1)W^2$ having a χ^2 distribution with (n_0-1) degrees of freedom.

Let $d(\theta^*)$ be chosen to satisfy

$$P\{Y_i^* \leq d(\theta^*)\sqrt{2}\beta^*W \text{ for } 1 \leq j \leq k-1\} = \sqrt{P^*}. \quad (4.9)$$

Then a modified two-stage elimination type selection rule, say $R_2(\theta^*)$, can be defined. This selection rule $R_2(\theta^*)$ is similar to rule R_2. The only difference is that now the value $d(\theta^*)$ is used instead of the value d. We denote the corresponding N by $N(\theta^*)$.

Following (4.7), one can see that for each $\underset{\sim}{\theta} \in \Omega(\delta^*,\theta^*)$ where $\Omega(\delta^*,\theta^*) = \{\underset{\sim}{\theta} \in \Omega(\delta^*)|\theta_{[k]} \leq \theta^*\}$,

$$P_{\underset{\sim}{\theta}}\{CS|R_2(\theta^*)\}$$

$$\geq [P\{Y_i^* \leq d(\theta^*)\sqrt{2}\beta^*W \text{ for } 1 \leq j \leq k-1\}]^2 = P^*.$$

Since $\rho^* > \frac{1}{2}$ and $\beta^* > \frac{1}{\sqrt{2}}$, it follows from Slepian's inequality, (4.8) and (4.9) that $d(\theta^*) < d$, and hence $N(\theta^*) \leq N$. Also, if we let $A(\theta^*)$ denote the subset selected at the first stage by applying rule $R_2(\theta^*)$, that is,

$$A(\theta^*) = \{i \mid \bar{X}_i^{(1)} \geq \bar{X}_j^{(1)} - (\frac{\sqrt{2}d(\theta^*)S_i}{\sqrt{n_0}} - \delta^*)^+ \text{ for } j \neq i\}, \quad (4.10)$$

then, since $d(\theta^*) < d$, from (4.1) and (4.10), we can see that $A(\theta^*) \subset A$; hence, $|A(\theta^*)| \leq |A|$ where $|B|$ denote the size of set B.

Generally speaking, the modified selection rule $R_2(\theta^*)$ reduces the size of the subset selected at the first stage and reduce the sample size needed at the second stage.

An Example to Illustrate the Use of R_2 and $R_2(\theta^*)$

Suppose that a consumer has to decide on buying one lot of bolts from among five lots that are available. The tensile strength (in pound per square inch; psi) of the ith ($1 \leq i \leq 5$) lot is normally distributed with mean θ_i and standard deviation $b\theta_i$, where both b and θ_i are positive and unknown. Suppose that $\delta^* = 200$ psi and $P^* = 0.90$ have been specified by the consumer. Further, suppose that $n_0 = 16$ bolts are randomly sampled from each of the five lots and that the sample means and the sample standard deviations are:

$$(\bar{x}_1^{(1)}, \bar{x}_2^{(1)}, \bar{x}_3^{(1)}, \bar{x}_4^{(1)}, \bar{x}_5^{(1)}) = (350, 380, 470, 600, 650),$$

$$(s_1, s_2, s_3, s_4, s_5) = (360, 420, 500, 580, 600).$$

Now $\sqrt{P^*} = 0.9486833 \approx 0.95$, so from Table IV ($\rho = 0.5$) in Gupta, Panchapakesan and Sohn (1985), using interpolation, it is found that $d \approx 2.34582$. Then, $A = \{3, 4, 5\}$. Therefore, we proceed to the second stage and find that $N = 100$. Further additional 84 observations are taken from each of the selected lot and the sample mean $\bar{x}_i$ is computed. Finally, the consumer selects the lot of bolts associated with the largest sample mean among $\bar{X}_i$, $3 \leq i \leq 5$, as the best.

Suppose now that the consumer, from past experience, has some knowledge that $\theta_{[5]} \leq 1000 = \theta^*$. Therefore, he prefers to apply the modified selection rule $R_2(\theta^*)$ for his selection problem. Now,

$$\rho^* = \theta^{*2}(\theta^{*2}+(\theta^*-\delta^*)^2)^{-1} = 0.609756;$$

$$\beta^* = \theta^*(\theta^{*2}+(\theta^*-\delta^*)^2)^{-\frac{1}{2}} = 0.7808688.$$

Let $d_1 = d(\theta^*)\sqrt{2}\beta^*$. From (4.9) and Table IV ($\rho = 0.6$) in Gupta, Panchapakesan and Sohn (1985), using interpolation, it is found that $d_1 = 2.30289$. Therefore $d(\theta^*) \approx 2.0853556$. Note that $\rho^* \approx 0.609756 > 0.6$. So the value $d(\theta^*)$ obtained in this way will be a little conservative since the exact value of $d(\theta^*)$ will be a little less than that used here. We then find that $A(\theta^*) = \{4,5\}$ and $N(\theta^*) = 79$. Therefore the consumer needs to take additional 63 observations from each of 4th and 5th lots to accomplish the selection process. Note that the total sample size by applying rule $R_2(\theta^*)$ is 206 while the total sample size by applying rule R_2 is 332. The saving of the total sample size is quite significant.

References

1. Gleser, L. J. and Healy, J. D. (1976). Estimating the mean of a normal distribution with known coefficient of variation. J. Amer. Statist. Assoc. 71, 977-81.

2. Gupta, S. S. (1963). Probability integrals of multivariate normal and multivariate t. Ann. Math. Statist. 34, 792-828.

3. Gupta, S. S. and Kim, W. C. (1984). A two-stage elimination type procedure for selecting the largest of several normal means with a common unknown variance. Design of Experiments: Ranking and Selection, (Eds. T. J. Santner and A. C. Tamhane), Marcel Dekker, New York, 77-94.

4. Gupta, S. S. and Panchapakesan, S. (1979). Multiple Decision Procedures: Methodology of Selecting and Ranking Populations, John Wiley, New York.

5. Gupta, S. S., Panchapakesan, S. and Sohn, J. K. (1985). On the distribution of the studentized maximum of equally correlated normal random variables. Commun. Statist.-Simula. Computa. 14(1), 103-135.

6. Gupta, S. S. and Singh, A. K. (1983). On subset selection procedures for the largest mean from normal populations having a common known coefficient of variation. Technical Report #83-6, Dept. of Statist., Purdue Univ., W. Lafayette, IN.

7. Gupta, S. S. and Sobel, M. (1957). On a statistic which arises in selection and ranking problems. Ann. Math. Statist. 28, 957-967.

8. Tamhane, A. C. (1978). Ranking and selection problems for normal populations with common known coefficient of variation. Sankhyā B39, 344-361.

BEFORE COMPLETING FORM

1. REPORT NUMBER Technical Report #86-16	2. GOVT ACCESSION NO	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Selecting the Best Unknown Mean from Normal Populations Having A Common Unknown Coefficient of Variation		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER Technical Report #86-16
7. AUTHOR(s) Shanti S. Gupta and TaChen Liang		8. CONTRACT OR GRANT NUMBER(s) N00014-84-C-0167
9. PERFORMING ORGANIZATION NAME AND ADDRESS Purdue University Department of Statistics West Lafayette, IN 47907		10. PROGRAM ELEMENT, PROJECT, TASK, AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Washington, DC		12. REPORT DATE May 1986 13. NUMBER OF PAGES 19
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Coefficient of variation; subset selection approach; correct selection; elimination type two-stage selection rule; expected size of selected subset; indifference-zone approach; least favorable configuration.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper deals with the problem of selecting the population associated with the largest unknown mean from several normal populations having a common unknown coefficient of variation. Both subset selection and indifference zone approaches are studied. Based on the observed sample means and sample standard deviations, a subset selection rule is proposed. Some properties related to this selection rule are discussed. For the indifference zone approach, a two-stage elimination type selection rule is considered. If the experimenter has some

(over)

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

prior knowledge about an upper bound on the unknown means, a modification is introduced to reduce the size of the selected subset at the first stage and also to reduce the sample size at the second stage. An example is provided which indicates that the saving of total sample size is quite significant if this prior knowledge is taken into consideration in designing the selection rule. It is shown how to implement the above selection rules by using several existing tables.

END

DTIC

7-86