ON A LOWER CONFIDENCE BOUND FOR THE PROBABILITY OF A CORRECT SELECTION: ANALYTICAL AND SIMULATION STUDIES*

by

Shanti S. Gupta
Purdue University

TaChen Liang
Purdue University
Technical Report#87-5

Department of Statistics
Purdue University

March 1987

Revised March 1988

* The research of this author was supported in part by the Office of Naval Research Contract N00014–84–C–0167, N00014–88–K–0170 at Purdue University and NSF Grant DMS–8606964. Reproduction in whole or in part is permitted for any purpose of the United States Government.

ON A LOWER CONFIDENCE BOUND FOR THE PROBABILITY OF A CORRECT SELECTION: ANALYTICAL AND SIMULATION STUDIES

by

Shanti S. Gupta* and TaChen Liang
Purdue University

Abstract

For the problem of selecting the best of several populations using the indifference (preference) zone formulation, a natural rule is to select the population yielding the largest sample value of an appropriate statistic. For this approach, it is required that the experimenter specify a number δ^*, say, which is a lower bound on the difference (separation) between the largest and the second largest parameter. However, in many real situations, it is hard to assign the value of δ^* and, therefore, in case that the assumption of indifference zone is violated, the probability of a correct selection cannot be guaranteed to be at least P^*, a prespecified value. In this paper, we are concerned with deriving a lower confidence bound for the probability of a correct selection for the general location model $F(x - \theta_i), i = 1, \ldots, k$. First, we derive simultaneous lower confidence bounds on the differences between the largest (best) and each of the other non-best population parameters. Based on these, we obtain a lower confidence bound for the probability of a correct selection. The general result is then applied to the selection of the best mean of k normal populations with both the known and unknown common variances. In the first case one needs a single-stage procedure while in the second case a two-stage procedure is required. Some simulation investigations are described and their results are provided.

* Invited paper to be presented by this author at the First International Conference on Statistical Computing, 30 March–2 April, 1987, to be held at Cesme, Izmir, Turkey.

This research was partially supported by the Office of Naval Research Contract N00014–84–C–0167, N00014–88–K–0170 and NSF Grant DMS–8606964 at Purdue University.

AMS 1980 Subject Classification: Primary 62F25, 62F07
KEY WORDS: Correct selection, Indifference zone, Lower confidence bound, Best population

1. Introduction

Let $X_{ij},\ j = 1, \ldots, n$, be n independent observations from a population π_i, where $\pi_1, \pi_2, \ldots, \pi_k$ are independently distributed with continuous cumulative distribution function $G(x-\theta_i),\ 1 \leq i \leq k$, respectively. Let $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$ and let $\theta_{(1)} \leq \ldots \leq \theta_{(k)}$ denote the ordered values of $\theta_1, \ldots, \theta_k$. It is assumed that the exact pairing between the ordered parameters and the unordered parameters is unknown. The population associated with the largest location parameter $\theta_{(k)}$ is called the best population. Assume that the experimenter is interested in the selection of the best population. For this purpose, we choose an appropriate statistic $Y_i = Y(X_{i1}, \ldots, X_{in})$ with cumulative distribution function $F_n(y-\theta_i)$ and use the natural selection rule that selects the population yielding the largest Y_i as the best population. Let CS (correct selection) denote the event that the best population is selected. Then, the probability of a correct selection (PCS) applying the natural selection rule is:

$$P_{\underset{\sim}{\theta}}\{CS\} = \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} F_n(y + \theta_{(k)} - \theta_{(i)}) dF_n(y). \tag{1.1}$$

To guarantee the probability of a correct selection, Bechhofer (1954) introduced the indifference zone approach in which the experimenter is asked to assign a positive value δ^* such that

$$\theta_{(k)} \geq \theta_{(k-1)} + \delta^*. \tag{1.2}$$

Thus, the subspace $\Omega(\delta^*) = \{\underset{\sim}{\theta} | \theta_{(k)} \geq \theta_{(k-1)} + \delta^*\}$ is called the preference zone and its complement $\Omega^c(\delta^*) = \{\underset{\sim}{\theta} | \theta_{(k)} < \theta_{(k-1)} + \delta^*\}$ is the indifference zone. We also let $\Omega = \Omega(\delta^*) \bigcup \Omega^c(\delta^*)$. On $\Omega(\delta^*)$, we have,

$$\inf_{\underset{\sim}{\theta} \epsilon \Omega(\delta^*)} P_{\underset{\sim}{\theta}}\{CS\} = \int_{-\infty}^{\infty} [F_n(y + \delta^*)]^{k-1} dF_n(y). \tag{1.3}$$

Suppose that the function on the right-hand-side of (1.3) is an increasing function of the common sample size n and tends to one as n tends to infinity. Then, for a given probability $P^*(k^{-1} < P^* <$

1), the minimum common sample size n_0 that is required to guarantee the probability of a correct selection to be at least P^* over the preference zone is determined by

$$n_0 \equiv n_0(\delta^*, P^*) = \min\{n | \int_{-\infty}^{\infty} [F_n(y + \delta^*)]^{k-1} dF_n(y) \geq P^*\}. \tag{1.4}$$

However, in a real situation, it may be hard to assign the value of δ^* such that $\theta_{(k)} \geq \theta_{(k-1)} + \delta^*$ since the parameter values $\theta_{(k)}, \theta_{(k-1)}$ are unknown. So that if the above assumption is not satisfied, then the probability of a correct selection cannot be guaranteed to be at least equal to P^*. Parnes and Srinivasan (1986) have also pointed out certain inconsistencies in the indifference zone formulation of some selection problems. It should be pointed out that the work of Fabian (1962) and Hsu (1981) is of some relevance in indicating a way out of this impasse.

Recently, retrospective analyses regarding the PCS have been studied by some authors. Olkin, Sobel and Tong (1976, 1982) and Gibbons, Olkin and Sobel (1977) have presented estimators of the PCS. Faltin and McCulloch (1983) have studied the small-sample properties of the Olkin-Sobel-Tong's estimator of the PCS for the case when $k = 2$. Bofinger (1985) has discussed the non-existence of consistent estimators of the PCS. Anderson, Bishop and Dudewicz (1977) have given a lower confidence bound on the PCS in the case of normal populations having a common variance which is either known or unknown. Kim (1986) has presented a lower confidence bound on the PCS for the case where the underlying probability density function $f_n(y - \theta)$ of $F_n(y - \theta)$ has the monotone likelihood ratio property in y and θ and studied its application to the case of normal populations with common known or common unknown variances.

In this paper, we are concerned with deriving a lower confidence bound for the probability of a correct selection for the general location model $G(x - \theta_i)$, $i = 1, \ldots, k$. First, we derive simultaneous lower confidence bounds on the differences between the largest (best) and each of the other non-best population parameters. Based on these, we obtain a lower confidence bound for the probability of a correct selection. The general result is then applied to the selection of the best mean of k normal populations with both the known and unknown common variances. In the first case one needs a single-stage procedure while in the second case a two-stage procedure is required. Some simulation investigations are described and their results are provided.

2. A Lower Confidence Bound on PCS

For given δ^* and P^*, let n_0 be the minimum common sample size determined by (1.4). Let $Y_i = Y(X_{i1}, \ldots, X_{in_0})$, be an appropriate statistic for inference regarding θ_i and let us assume that the distribution of $Y_i - \theta_i$ is independent of θ_i, $1 \leq i \leq k$. Let $Y_{[1]} \leq \ldots \leq Y_{[k]}$ denote the order statistics of Y_i, $1 \leq i \leq k$. Also, let $\theta_{[i]}$ denote the (unknown) parameter associated with $Y_{[i]}$. For given α, $0 < \alpha < 1$, let $c(k, n_0, \alpha)$ be the value such that

$$P_{\underset{\sim}{\theta}}\{\max_{1\leq i\leq k}(Y_i - \theta_i) - \min_{1\leq j\leq k}(Y_j - \theta_j) \leq c(k, n_0, \alpha)\} = 1 - \alpha. \tag{2.1}$$

Let $E = \{\max_{1\leq i\leq k}(Y_i - \theta_i) - \min_{1\leq j\leq k}(Y_j - \theta_j) \leq c(k, n_0, \alpha)\}$. Then, we have the following lemma.

Lemma 2.1. $E \subset \{(Y_{[k]} - Y_{[i]} - c(k, n_0, \alpha))^+ \leq \theta_{(k)} - \theta_{(i)},\ 1 \leq i \leq k-1\}$, where $(y)^+ = \max(0, y)$.

Proof: First note that for each $i = 1, \ldots, k$,

$$\begin{aligned} \min_{j\leq i}(Y_{[j]} - \theta_{[j]}) &\leq \min_{j\leq i}(Y_{[i]} - \theta_{[j]}) \\ &= Y_{[i]} - \max_{j\leq i}\theta_{[j]} \\ &\leq Y_{[i]} - \theta_{(i)}. \end{aligned} \tag{2.2}$$

Thus,

$$\begin{aligned}
E &\subset \{\max_{1\le i\le k}(Y_i-\theta_{(k)})-\min_{1\le j\le k}(Y_{[j]}-\theta_{[j]})\le c(k,n_0,\alpha)\}\\
&\subset \{(Y_{[k]}-\theta_{(k)})-\min_{1\le j\le k-1}(Y_{[j]}-\theta_{[j]})\le c(k,n_0,\alpha)\}\\
&= \{(Y_{[k]}-\theta_{(k)})-\min_{j\le i}(Y_{[j]}-\theta_{[j]})\le c(k,n_0,\alpha),\ 1\le i\le k-1\}\\
&\subset \{(Y_{[k]}-\theta_{(k)})-(Y_{[i]}-\theta_{(i)})\le c(k,n_0,\alpha),\ 1\le i\le k-1\}(\text{by } (2.2))\\
&= \{Y_{[k]}-Y_{[i]}-c(k,n_0,\alpha)\le \theta_{(k)}-\theta_{(i)},\ 1\le i\le k-1\}\\
&= \{(Y_{[k]}-Y_{[i]}-c(k,n_0,\alpha))^+\le \theta_{(k)}-\theta_{(i)},\ 1\le i\le k-1\}.
\end{aligned}$$

Note that the last equality follows from the fact that $\theta_{(k)}-\theta_{(i)}\ge 0$ for all $1\le i\le k-1$. Hence, we complete the proof of this lemma.

Note that in (1.1), the probability of a correct selection $P_{\underset{\sim}{\theta}}\{CS\}$ depends on the parameters $\underset{\sim}{\theta}=(\theta_1,\ldots,\theta_k)$ only via the differences $\theta_{(k)}-\theta_{(i)},\ 1\le i\le k-1$. For convenience, we write $P_{\underset{\sim}{\theta}}\{CS\}=P(\delta_1,\ldots,\delta_{k-1})$ where $\delta_i=\theta_{(k)}-\theta_{(i)},\ 1\le i\le k-1$. We see that $P(\delta_1,\ldots,\delta_{k-1})$ is a nondecreasing function of δ_i for each $i=1,2,\ldots,k-1$.

For each $i=1,\ldots,k-1$, let

$$\hat{\delta}_{L,i}=(Y_{[k]}-Y_{[i]}-c(k,n_0,\alpha))^+, \tag{2.3}$$

$$\hat{P}_L=P(\hat{\delta}_{L,1},\ldots,\hat{\delta}_{L,k-1}). \tag{2.4}$$

We propose $\hat{P}_L$ as an estimator of a lower bound of the PCS. We have the following theorem.

Theorem 2.2. $P_{\underset{\sim}{\theta}}\{P_{\underset{\sim}{\theta}}\{CS\}\ge \hat{P}_L\}\ge 1-\alpha$ for all $\underset{\sim}{\theta}\epsilon\Omega$.

Proof: By nondecreasing property of $P(\delta_1,\ldots,\delta_{k-1})$ with respect to $\delta_i,\ 1\le i\le k-1$, from (2.1) and Lemma 2.1, we have, for $\underset{\sim}{\theta}\epsilon\Omega$,

$$
\begin{aligned}
1-\alpha &= P_{\underset{\sim}{\theta}}\{E\} \\
&\le P_{\underset{\sim}{\theta}}\{\hat{\delta}_{L,i} \le \theta_{(k)} - \theta_{(i)},\ 1 \le i \le k-1\} \\
&\le P_{\underset{\sim}{\theta}}\{P(\hat{\delta}_{L,1},\ldots,\hat{\delta}_{L,k-1}) \le P(\delta_1,\ldots,\delta_{k-1})\} \\
&= P_{\underset{\sim}{\theta}}\{\hat{P}_L \le P_{\underset{\sim}{\theta}}\{CS\}\}.
\end{aligned}
$$

This completes the proof of this theorem.

3. Selection of the Best Normal Population in Terms of Means

Let $X_{ij},\ 1 \le j \le n$ be independent observations from $N(\theta_i,\ \sigma^2),\ i = 1,\ldots,k$ where the common variance σ^2 may be either known or unknown. The best population is the one associated with the largest mean $\theta_{(k)}$. We consider two situations according to whether the common variance σ^2 is known or unknown.

3.1. Lower Confidence Bound for PCS : σ^2 Known Case.

When the value of the common variance σ^2 is known, for $\underset{\sim}{\theta}\epsilon\Omega$, the probability of a correct selection applying the natural selection rule is:

$$
P_{\underset{\sim}{\theta}}\{CS\} = \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \Phi(x + \frac{\sqrt{n_0}(\theta_{(k)} - \theta_{(i)})}{\sigma}) d\Phi(x), \tag{3.1}
$$

where $\Phi(\cdot)$ is the standard normal distribution function, and the value of the sample size n_0, for the indifference zone formulation, is determined by

$$
n_0 = min\{n | \int_{-\infty}^{\infty} [\Phi(x + \frac{\sqrt{n}\delta^*}{\sigma})]^{k-1} d\Phi(x) \ge P^*\}. \tag{3.2}
$$

Let $\overline{X}_i = \frac{1}{n_0} \sum_{j=1}^{n_0} X_{ij}$. For given $0 < \alpha < 1$, choose the value $c(k, n_0, \alpha)$ such that

$$
P_{\underset{\sim}{\theta}}\{\max_{1\le i\le k}(\overline{X}_i - \theta_i) - \min_{1\le j\le k}(X_j - \theta_j) \le c(k, n_0, \alpha)\} = 1 - \alpha. \tag{3.3}
$$

Note that here, $c(k, n_0, \alpha) = \frac{\sigma}{\sqrt{n_0}} q^{\alpha}_{k,\infty}$, where $q^{\alpha}_{k,\infty}$ is the $100(1-\alpha)$%th percentile of Tukey's studentized range statistic with parameters (k, ∞). The value of $q^{\alpha}_{k,\infty}$ is available from Harter (1969). Then, we define

$$\hat{\delta}_{L,i} = (\overline{X}_{[k]} - \overline{X}_{[i]} - c(k, n_0, \alpha))^{+} \tag{3.4}$$

and

$$\hat{P}_L = P(\hat{\delta}_{L,1}, \ldots, \hat{\delta}_{L,k-1}) = \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \Phi(x + \frac{\sqrt{n_0}\hat{\delta}_{L,i}}{\sigma}) d\Phi(x). \tag{3.5}$$

Then, by Theorem 2.2, $P_{\underset{\sim}{\theta}}\{P_{\underset{\sim}{\theta}}\{CS\} \geq \hat{P}_L\} \geq 1 - \alpha$ for all $\underset{\sim}{\theta} \epsilon \Omega$.

3.2. Lower Confidence Bound for PCS : σ^2 Unknown Case.

When the common variance σ^2 is unknown, Bechhofer, Dunnett and Sobel (1954) presented a two-stage selection rule, which is briefly described as follows.

Take a first sample of $n_0 (n_0 \geq 2)$ observations from each of the k populations. Compute $\overline{X}_i = \frac{1}{n_0} \sum_{j=1}^{n_0} X_{ij}$, $(1 \leq i \leq k)$, and $S^2 = \frac{1}{k(n_0-1)} \sum_{i=1}^{k} \sum_{j=1}^{n_0} (X_{ij} - \overline{X}_i)^2$. Define $N = \max\{n_0, [\frac{S^2 h^2}{\delta^{*2}}]\}$ where the symbol $[y]$ denotes the smallest integer not less than y, and h is a positive value such that

$$\int_0^{\infty} \int_{-\infty}^{\infty} [\Phi(x + wh)]^{k-1} d\Phi(x) dF_W(w) = P^*, \tag{3.6}$$

$k^{-1} < P^* < 1$, and $F_W(\cdot)$ is the distribution function of the nonnegative random variable W with $k(n_0 - 1)W^2$ following $\chi^2(k(n_0 - 1))$ distribution.

Then, take additional $N - n_0$ observations from each population. Compute the overall mean $\overline{X}_i(N) = \frac{1}{N} \sum_{j=1}^{N} X_{ij}$, $1 \leq i \leq k$. We then select the population yielding the largest observation $\overline{X}_{[k]}(N)$ as the best population.

For this two-stage selection rule, the probability of a correct selection is:

$$
\begin{aligned}
P_{\underset{\sim}{\theta}}\{CS\} &= P_{\underset{\sim}{\theta}}\{\overline{X}_{(k)}(N) > \overline{X}_{(i)}(N),\ i \neq k\} \\
&= P_{\underset{\sim}{\theta}}\{\frac{\sqrt{N}(\overline{X}_{(k)}(N) - \theta_{(k)})}{\sigma} + \frac{\sqrt{N}(\theta_{(k)} - \theta_{(i)})}{\sigma} > \frac{\sqrt{N}(\overline{X}_{(i)} - \theta_{(i)})}{\sigma}, i \neq k\} \\
&\geq P_{\underset{\sim}{\theta}}\{\frac{\sqrt{N}(\overline{X}_{(k)}(N) - \theta_{(k)})}{\sigma} + \frac{h(\theta_{(k)} - \theta_{(i)})}{\delta^*}\frac{S}{\sigma} > \frac{\sqrt{N}(\overline{X}_{(i)} - \theta_{(i)})}{\sigma}, i \neq k\} \\
&\quad (\text{ since } N \geq [\frac{S^2 h^2}{\delta^{*2}}]) \qquad (3.7) \\
&= P\{Z_k + \frac{h(\theta_{(k)} - \theta_{(i)})}{\delta^*} W > Z_i,\ i \neq k\} \\
&= \int_0^\infty \int_{-\infty}^\infty \prod_{i=1}^{k-1} \Phi(z + \frac{h(\theta_{(k)} - \theta_{(i)})}{\delta^*} w) d\Phi(z) dF_W(w),
\end{aligned}
$$

where $Z_1, \ldots, Z_k$ are iid random variables having standard normal distribution, and $W = S/\sigma$ with $k(n_0 - 1)W^2 \sim \chi^2(k(n_0 - 1))$ and $(Z_1, \ldots, Z_k)$ and W are independent.

Thus, to obtain a lower confidence bound for $P_{\underset{\sim}{\theta}}\{CS\}$, it suffices to find simultaneous lower confidence bounds for $\theta_{(k)} - \theta_{(i)}$, $1 \leq i \leq k-1$. Then, replacing the $\theta_{(k)} - \theta_{(i)}$, $1 \leq i \leq k-1$, by the corresponding lower confidence bounds into the function on the right-hand-side of (3.7), we obtain a lower confidence bound for $P_{\underset{\sim}{\theta}}\{CS\}$. For convenience, we let

$$
Q(\delta_1, \ldots, \delta_{k-1}) = \int_0^\infty \int_{-\infty}^\infty \prod_{i=1}^{k-1} \Phi(z + \frac{h(\theta_{(k)} - \theta_{(i)})w}{\delta^*}) d\Phi(z) dF_W(w). \qquad (3.8)
$$

Let $c = S q^{\alpha}_{k,k(n_0-1)}/\sqrt{N}$, where $q^{\alpha}_{k,k(n_0-1)}$ is the $100(1-\alpha)\%$th percentile of Tukey's studentized range statistic with parameters $(k, k(n_0 - 1))$. Define

$$
\hat{\delta}_{L,i} = (\overline{X}_{[k]}(N) - \overline{X}_{[i]}(N) - c)^+, \qquad (3.9)
$$

and

$$
\hat{Q}_L = Q(\hat{\delta}_{L,1}, \ldots, \hat{\delta}_{L,k-1}). \qquad (3.10)
$$

We propose $\hat{Q}_L$ as an estimator of a lower bound of $P_{\underset{\sim}{\theta}}\{CS\}$.

Lemma 3.1. Let $E = \{\max_{1\le i\le k}(\overline{X}_i(N) - \theta_i) - \min_{1\le j\le k}(\overline{X}_j(N) - \theta_j) \le c\}$. Then, $P_{\underset{\sim}{\theta}}\{E\} = 1-\alpha$ for all $\underset{\sim}{\theta}\epsilon\Omega$.

Proof:

$$\begin{aligned} P_{\underset{\sim}{\theta}}(E) &= P_{\underset{\sim}{\theta}}\{\max_{1\le i\le k}(\overline{X}_i(N) - \theta_i) - \min_{1\le j\le k}(\overline{X}_j(N) - \theta_j) \le c\} \\ &= P_{\underset{\sim}{\theta}}\{\max_{1\le i\le k}\sqrt{N}(\overline{X}_i(N) - \theta_i) - \min_{1\le j\le k}\sqrt{N}(\overline{X}_j(N) - \theta_j) \le Sq^{\alpha}_{k,k(n_0-1)}\} \\ &= 1-\alpha, \end{aligned}$$

where the last equality follows from the definition of $q^{\alpha}_{k,k(n_0-1)}$.

Lemma 3.2. $P_{\underset{\sim}{\theta}}\{\hat{\delta}_{L,i} \le \theta_{(k)} - \theta_{(i)},\ 1\le i\le k-1\} \ge 1-\alpha$ for all $\underset{\sim}{\theta}\epsilon\Omega$.

Proof: Following the same argument as in Lemma 2.1, we have $E \subset \{\hat{\delta}_{L,i} \le \theta_{(k)} - \theta_{(i)},\ 1\le i\le k-1\}$. Then using Lemma 3.1 leads to the conclusion of Lemma 3.2.

Lemma 3.2 and the increasing property of the function $Q(\delta_1,\ldots,\delta_{k-1})$ with respect to δ_i, $1\le i\le k-1$, lead to the following main result.

Theorem 3.3. $P_{\underset{\sim}{\theta}}\{P_{\underset{\sim}{\theta}}\{CS\} \ge \hat{Q}_L\} \ge 1-\alpha$ for all $\underset{\sim}{\theta}\epsilon\Omega$.

Proof: Note that $P_{\underset{\sim}{\theta}}\{CS\} \ge Q(\delta_1,\ldots,\delta_{k-1})$ for all $\underset{\sim}{\theta}\epsilon\Omega$. Therefore, $P_{\underset{\sim}{\theta}}\{P_{\underset{\sim}{\theta}}\{CS\} \ge \hat{Q}_L\} \ge P_{\underset{\sim}{\theta}}\{Q(\delta_1,\ldots,\delta_{k-1}) \ge \hat{Q}_L\} \ge 1-\alpha$ for all $\underset{\sim}{\theta}\epsilon\Omega$.

4. Remark and Example

Anderson, Bishop and Dudewicz (1977) and Kim (1986) have also studied the problem of finding a lower confidence bound on PCS. They considered the retrospective analysis to approach a lower confidence bound for PCS. However, our approach is different from theirs. We use the following example to illustrate our procedure and describe the difference between ours and the above mentioned approaches.

Example (The data is taken from Example 3, page 506, of Gupta and Panchapakesan (1979)).

An experimenter wants to compare the glowing time of five different types of phosphorescent coatings of airplane instrument dials. Assume that the distributions of the glowing time for each type of phosphorescent coatings are normal with a common unknown variance σ^2. Based on some past information, the experimenter assigns $\delta^* = 5$. Then, using the indifference zone formulation, a two-stage natural selection rule as described in Section 3.2 is applied. We use $P^* = 0.90$ and the initial sample size n_0 to be 5. The coated dials were then excited with an ultraviolet light. The upper part of Table 1 shows the number of minutes each dial glows after the light source was turned off.

Table 1. Glowing Time of Five Types of Phosphorescent Coatings

	Coatings 1	2	3	4	5
observations taken at the first-stage	45.7	51.7	45.9	54.8	65.9
	48.4	46.4	54.8	55.6	65.4
	51.9	49.8	62.9	63.5	60.0
	57.0	5.27	64.7	61.6	70.1
	41.0	48.1	54.3	55.7	69.5
n_0	5	5	5	5	5
$\overline{X}_i$	48.8	49.74	56.52	58.24	66.18
	$S^2 = \frac{1}{k(n_0-1)} \sum_{i=1}^{k} \sum_{j=1}^{n_0} (X_{ij} - \overline{X}_i)^2 = 26.7305$				
observations taken at the second-stage	61.4	54.8	57.9	59.2	64.0
	47.0	54.0	53.9	53.2	56.0
	51.8	49.1	51.7	56.9	68.1
N	8	8	8	8	8
$\overline{X}_i(N)$	50.4375	50.825	55.7625	57.5625	64.875

For $k = 5,\ n_0 = 5,\ P^* = 0.90$, from Gupta, Panchapakesan and Sohn (1985), $h = 1.92727\sqrt{2}$. Therefore, $N = \max\{n_0, [\frac{S^2h^2}{\delta^{*2}}]\} = 8$ and hence $N - n_0 = 3$ additional observations should be taken from each population. The observations taken at the second-stage are given in the lower part of Table 1.

We then have the overall sample means: $\overline{X}_1(N) = 50.4375, \overline{X}_2(N) = 50.825, \overline{X}_3(N) = 55.7625, \overline{X}_4(N) = 57.5625$ and $\overline{X}_5(N) = 64.875$. According to the two-stage natural selection rule, coating number 5 which yields the largest sample mean is selected as the best.

However we do not know whether the largest and the second largest unknown means differ at least by $\delta^* = 5$ or not. A reasonable question is: what kind of confidence statement can be made regarding the PCS? By the method described in Section 3.2, for $\alpha = 0.10$, we see from Harter (1969) that, $q^{\alpha}_{k,k(n_0-1)} = 3.736$. Then, $c = Sq^{\alpha}_{k,k(n_0-1)}/\sqrt{N} = 6.8290$. Therefore, $\hat{\delta}_{L,1} = 7.6135$, $\hat{\delta}_{L,2} = 7.221$, $\hat{\delta}_{L,3} = 2.2835$ and $\hat{\delta}_{L,4} = 0.4835$. After some computation, we have $\hat{Q}_L = 0.518$. Therefore, we can state with at least 90% confidence that $PCS \geq \hat{Q}_L = 0.518$ for all values of true unknown means.

For different α values, the $100(1-\alpha)\%$ lower confidence bounds $\hat{Q}_L$ of the PCS are also computed and given as follows:

α	0.1	0.2	0.3	0.4	0.5
$\hat{Q}_L$	0.518	0.672	0.759	0.817	0.869

Remark: The procedure used in the above example and in our paper is designed for the data which are collected in two stages. The procedures of Anderson, Bishop and Dudewicz (1977) and Kim (1986) cannot be employed in this example. Thus the procedure of this paper covers the case where neither Anderson, Bishop, and Dudewicz's procedure nor Kim's procedure can be used.

5. Simulation Studies

For the normal means selection problem, for various parameter configurations, the behaviors of $\hat{P}_L$ and $\hat{Q}_L$ were simulated. Two types of parameter configurations were simulated: a slippage configuration $\theta_{(1)} = \ldots = \theta_{(k-1)} = \theta_{(k)} - \Delta$ and an equally spaced configuration $\theta_{(i)} - \theta_{(i-1)} = \Delta$, $i = 2, \ldots, k$. For simulation, we suppose that the assigned value of δ^* is 1 and also the assigned probability levels are $P^* = 0.90$ and $P^* = 0.95$. When the common variance σ^2 is known, the common sample size n_0 is determined by (3.2). When σ^2 is unknown, the initial common sample size is set equal to ten. The simulation process was repeated $M = 1000$ times for the case where σ^2 is known and $M = 400$ times for the σ^2 unknown case. For each simulation, the random observation X_{ij} is generated from $N(\theta_i, \sigma^2)$ with $\sigma^2 = 1$. The values of $\hat{P}_L$ and $\hat{Q}_L$ were computed. The averages of the 1000 $\hat{P}_L$ and 400 $\hat{Q}_L$ are reported in Table 2 and Table 3, respectively. In each table, the numbers in the parentheses are the standard errors of the corresponding estimators.

For convenience, we let $\hat{P}_L(P^*, \Delta, \alpha, T)$ and $\hat{Q}_L(P^*, \Delta, \alpha, T)$ denote the corresponding $\hat{P}_L$ and $\hat{Q}_L$ for given values of P^*, Δ, α and T, where T denotes the type of parameter configuration. The slippage configuration is denoted by S and the equally spaced configuration is denoted by ES.

The simulation results indicate the following:

1. Note that for fixed P^*, α and T, the PCS is a nondecreasing function of Δ. Therefore, it is reasonable to expect that both $\hat{P}_L(P^*, \Delta, \alpha, T)$ and $\hat{Q}_L(P^*, \Delta, \alpha, T)$ be nondecreasing in Δ. The simulation results indicate that this is so.

2. For fixed P^*, α and Δ, the PCS under equally spaced parameter configuration is larger than the PCS under the slippage configuration. The simulation results also indicate that this behavior holds. That is, from the simulation results, we find: $\hat{P}_L(P^*, \Delta, \alpha, ES) > \hat{P}_L(P^*, \Delta, \alpha, S)$ and $\hat{Q}_L(P^*, \Delta, \alpha, ES) > \hat{Q}_L(P^*, \Delta, \alpha, S)$.

3. For fixed values P^*, Δ and T, the simulation results indicate that $\hat{P}_L(P^*, \Delta, 0.2, T) > \hat{P}_L(P^*, \Delta, 0.1, T)$ and $\hat{Q}_L(P^*, \Delta, 0.2, T) > \hat{Q}_L(P^*, \Delta, 0.1, T)$. These results are as expected since $q^{\alpha}_{k,\nu}$ is nondecreasing in α for fixed k and ν.

4. For fixed Δ, α and T, $\hat{P}_L(P^*, \Delta, \alpha, T)$ is nondecreasing in P^*. Note that according to the sampling rule used in this paper, assigning large P^*-value implicitly implies taking more observations. Thus the simulation results seem to indicate that $\hat{P}_L(P^*, \Delta, \alpha, T)$ is nondecreasing

in the sample size. For the σ^2 unknown case, for both $k = 3$ and 5, the corresponding values of $\hat{Q}_L(P^*, \Delta, \alpha, T)$ are also nondecreasing in P^*.

Table 2. Simulated Values of $\hat{P}_L$ for $k = 3$; σ^2 Known

Δ	90% lower confidence bound				80% lower confidence bound			
	Slippage		Equally Spaced		Slippage		Equally Spaced	
	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$
0.5	0.3497 (0.0016)	0.3573 (0.0020)	0.3708 (0.0023)	0.3993 (0.0032)	0.3648 (0.0023)	0.3779 (0.0029)	0.3975 (0.0031)	0.4340 (0.0040)
0.8	0.3703 (0.0026)	0.3989 (0.0036)	0.4389 (0.0037)	0.5037 (0.0047)	0.3995 (0.0035)	0.4400 (0.0047)	0.4846 (0.0045)	0.5553 (0.0053)
1.0	0.3960 (0.0035)	0.4485 (0.0049)	0.5007 (0.0045)	0.5823 (0.0052)	0.4376 (0.0045)	0.5057 (0.0058)	0.5555 (0.0052)	0.6395 (0.0057)
1.5	0.5140 (0.0057)	0.6463 (0.0067)	0.6610 (0.0057)	0.7679 (0.0057)	0.5853 (0.0064)	0.7229 (0.0066)	0.7231 (0.0058)	0.8254 (0.0053)
2.0	0.6876 (0.0066)	0.8539 (0.0052)	0.8060 (0.0054)	0.9151 (0.0039)	0.7635 (0.0062)	0.9036 (0.0042)	0.8597 (0.0049)	0.9458 (0.0030)

Table 2 (continued) Simulated Values of $\hat{P}_L$ for $k = 5$; σ^2 Known

Δ	90% lower confidence bound				80% lower confidence bound			
	Slippage		Equally Spaced		Slippage		Equally Spaced	
	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$
0.5	0.2075 (0.0007)	0.2103 (0.0011)	0.3054 (0.0025)	0.3459 (0.0030)	0.2158 (0.0011)	0.2204 (0.0016)	0.3398 (0.0032)	0.3814 (0.0036)
0.8	0.2237 (0.0016)	0.2382 (0.0025)	0.4371 (0.0038)	0.4874 (0.0039)	0.2419 (0.0023)	0.2651 (0.0033)	0.4786 (0.0045)	0.5310 (0.0046)
1.0	0.2460 (0.0025)	0.2798 (0.0037)	0.5133 (0.0049)	0.5690 (0.0047)	0.2774 (0.0035)	0.3243 (0.0048)	0.5603 (0.0051)	0.6190 (0.0053)
1.5	0.3813 (0.0056)	0.5003 (0.0069)	0.6829 (0.0057)	0.7685 (0.0057)	0.4536 (0.0065)	0.5839 (0.0072)	0.7381 (0.0058)	0.8191 (0.0055)
2.0	0.6231 (0.0072)	0.7907 (0.0061)	0.8430 (0.0057)	0.9228 (0.0038)	0.7076 (0.0069)	0.8541 (0.0051))	0.8855 (0.0046)	0.9489 (0.0030)

Table 2 (continued) Simulated Values of $\hat{P}_L$ for $k = 10$; σ^2 Known

Δ	90% lower confidence bound				80% lower confidence bound			
	Slippage		Equally Spaced		Slippage		Equally Spaced	
	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$
0.5	0.1021 (0.0003)	0.1029 (0.0003)	0.2906 (0.0023)	0.3260 (0.0026)	0.1045 (0.0004)	0.1059 (0.0005)	0.3146 (0.0026)	0.3534 (0.0030)
0.8	0.1090 (0.0007)	0.1144 (0.0010)	0.4185 (0.0032)	0.4668 (0.0036)	0.1164 (0.1111)	0.1251 (0.0016)	0.4526 (0.0037)	0.4981 (0.0041)
1.0	0.1213 (0.0014)	0.1365 (0.0021)	0.4984 (0.0038)	0.5434 (0.0042)	0.1355 (0.0020)	0.1588 (0.0029)	0.5345 (0.0044)	0.5798 (0.0048)
1.5	0.2187 (0.0045)	0.3118 (0.0061)	0.6643 (0.0056)	0.7403 (0.0058)	0.2709 (0.0055)	0.3843 (0.0069)	0.7096 (0.0059)	0.7879 (0.0057)
2.0	0.4773 (0.0076)	0.6713 (0.0073)	0.8397 (0.0052)	0.9167 (0.0039)	0.5647 (0.0078)	0.7508 (0.0067)	0.8797 (0.0046)	0.9423 (0.0032)

Table 3. Simulated Values of $\hat{Q}_L$ for $k = 3$; σ^2 Unknown, $n_0 = 10$.

Δ	90% lower confidence bound				80% lower confidence bound			
	Slippage		Equally Spaced		Slippage		Equally Spaced	
	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$
0.8	0.3467 (0.0021)	0.4007 (0.0037)	0.4141 (0.0043)	0.5158 (0.0035)	0.3689 (0.0037)	0.4757 (0.0062)	0.4633 (0.0005)	0.5791 (0.0058)
1.0	0.3647 (0.0034)	0.4831 (0.0063)	0.4813 (0.0050)	0.5933 (0.0060)	0.4031 (0.0055)	0.5759 (0.0081)	0.5352 (0.0064)	0.6641 (0.0076)
2.0	0.7068 (0.0082)	0.9466 (0.0024)	0.8203 (0.0066)	0.9661 (0.0014)	0.8054 (0.0070)	0.9737 (0.0014)	0.8869 (0.0054)	0.9839 (0.0007)

Table 3 (continued) Simulated Values of $\hat{Q}_L$ for $k = 5$; σ^2 Unknown, $n_0 = 10$.

Δ	90% lower confidence bound				80% lower confidence bound			
	Slippage		Equally Spaced		Slippage		Equally Spaced	
	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$	$P^* = 0.90$	$P^* = 0.95$
0.8	0.2066 (0.0006)	0.2887 (0.0014)	0.4391 (0.0024)	0.6999 (0.0024)	0.2306 (0.0014)	0.3347 (0.0020)	0.5251 (0.0036)	0.8113 (0.0021)
1.0	0.2331 (0.0015)	0.3492 (0.0022)	0.5564 (0.0034)	0.8757 (0.0013)	0.2790 (0.0024)	0.3963 (0.0029)	0.6732 (0.0042)	0.9328 (0.0008)
2.0	0.7844 (0.0050)	0.9430 (0.0017)	0.9717 (0.0015)	0.9994 (0.0003)	0.8701 (0.0040)	0.9717 (0.0011)	0.9860 (0.0013)	0.9998 (0.0000)

6. A Lower Confidence Bound on PCS for Scale Parameter Model

The results in Section 2 are derived for a location parameter model. Similar results for a scale parameter model can also be obtained. For the problem of selecting the population with the largest scale parameter $\theta_{(k)}$, the PCS in (1.1) is replaced by

$$P_{\underset{\sim}{\theta}}\{CS\} = \int_0^\infty \prod_{i=1}^{k-1} F_n(\frac{\theta_{(k)}y}{\theta_{(i)}})dF_n(y), \quad \theta_i > 0, \ \forall_i \tag{6.1}$$

where $F_n(y|\theta_i)$ is the cumulative distribution function of an appropriate nonnegative statistic $Y_i = Y(X_{i1}, \ldots, X_{in})$, and the common sample size n is determined according to some sampling rule. Suppose that the distribution of Y_i/θ_i is independent of θ_i, $i = 1, \ldots, k$. For given α, $0 < \alpha < 1$, let d be the smallest value such that

$$P_{\underset{\sim}{\theta}}\{(\max_{1\le i\le k} Y_i/\theta_i)/(\min_{1\le j\le k} Y_j/\theta_j) \le d\} = 1 - \alpha. \tag{6.2}$$

Note that $d \ge 1$, since $(\max_{1\le i\le k} Y_i/\theta_i)/(\min_{1\le j\le k} Y_j/\theta_j) \ge 1$.

Analogous to the result obtained in Section 2, we have

$$P_{\underset{\sim}{\theta}}\{\theta_{(k)}/\theta_{(i)} \ge Y_{[k]}/(Y_{[i]}d), \ i = 1, \ldots, k-1\} \ge 1 - \alpha, \tag{6.3}$$

for all $\underset{\sim}{\theta}\epsilon\Omega$ where, now, the parameter space $\Omega = \{\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)|\theta_i > 0, \ 1 \le i \le k\}$.

Define

$$\hat{\delta}_{L,i} = (Y_{[k]}/(Y_{[i]}d))^0, \quad i = 1, \ldots, k-1, \tag{6.4}$$

where $(y)^0 = \max(y, 1)$. Replacing $\theta_{(k)}/\theta_{(i)}$ in (6.1) by $\hat{\delta}_{L,i}$, we obtain

$$\hat{P}_L = \int_0^\infty \prod_{i=1}^{k-1} F_n(\hat{\delta}_{L,i} y) dF_n(y). \tag{6.5}$$

We propose $\hat{P}_L$ as an estimator of a lower bound for the PCS. We have

$$P_{\underset{\sim}{\theta}}\{P_{\underset{\sim}{\theta}}\{CS\} \geq \hat{P}_L\} \geq 1 - \alpha \text{ for all } \underset{\sim}{\theta} \epsilon \Omega. \tag{6.6}$$

Acknowledgements

This research was partially supported by the Office of Naval Research Contracts, N00014–84–C–0167, N00014–88–K–0170 and NSF Grant DMS–8606964 at Purdue University. The authors are grateful to two referees for helpful comments and to Dr. T. F. Lu, Purdue University, for implementing the simulation.

References

Anderson, P. O., Bishop, T. A. and Dudewicz, E. J. (1977). Indifference zone ranking and selection: confidence intervals for true achieved P(CD). *Commun. Statist.* **A(6)**, 1121–1132.

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.

Bechhofer, R. E. and Dunnett, C. W. and Sobel, M. (1954). A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. *Biometrika* **41**, 170–176.

Bofinger, E. (1985). On the non-existence of consistent estimators for $P(CS)$. *American Journal of Mathematics and Management Sciences* **5**, 63–76.

Fabian, V. (1962). On multiple decision methods for ranking population means. *Ann. Math. Statist.* **33**, 248–254.

Faltin, F. and McCulloch, C. E. (1983). On the small-sample properties of the Olkin-Sobel-Tong estimator of the probability of correct selection. *J. Amer. Statist. Assoc.* **78**, 464–467.

Gibbons, J. D., Olkin, I. and Sobel, M. (1977). *Selecting and Ordering Populations: A New Statistical Methodology.* New York: John Wiley.

Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedures.* New York: John Wiley.

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.

Harter, H. L. (1969). *Order Statistics and Their Use in Testing and Estimation,* **Vol. 1**, *Tests Based on Range and Studentized Range of Samples from a Normal Population*, Aerospace Research Laboratories.

Hsu, J. C. (1981). Simultaneous confidence intervals for all distances from the "best". *Ann. Statist.* **9**, 1026–1034.

Kim, W.-C. (1986). A lower confidence bound on the probability of a correct selection. *J. Amer. Statist. Assoc.* **81**, 1012–1017.

Olkin, I., Sobel, M. and Tong, Y. L. (1976). Estimating the true probability of a correct selection for location and scale parameter families. Technical Report 110, Stanford University, Department of Statistics.

Olkin, I., Sobel, M. and Tong, Y. L. (1982). Bounds for a k-fold integral for location and scale parameter models with applications to statistical ranking and selection problem. *Statistical Decision Theory and Related Topics III* **Vol. 2**. (Eds. S. S. Gupta and J. O. Berger), New York, Academic Press, pp. 193–212.

Parnes, M. and Srinivasan, R. (1986). Some inconsistencies and errors in the indifference zone formulation of selection. *Sankhya* **48**, 86–97.

BEFORE COMPLETING FORM

1. REPORT NUMBER Technical Report #87-5	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ON A LOWER CONFIDENCE BOUND FOR THE PROBABILITY OF A CORRECT SELECTION: ANALYTICAL AND SIMULATION STUDIES		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER Technical Report #87-5
7. AUTHOR(s) Shanti S. Gupta and TaChen Liang		8. CONTRACT OR GRANT NUMBER(s) N00014-84-C-0167, N00014-88-K-0170 and NSF Grant DMS-8606964
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 March 1987
		13. NUMBER OF PAGES 17
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)

Correct selection, Indifference zone, Lower confidence bound, Best population

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

For the problem of selecting the best of several populations using the indifference (preference) zone formulation, a natural rule is to select the population yielding the largest sample value of an appropriate statistic. For this approach, it is required that the experimenter specify a number δ^*, say, which is a lower bound on the difference (separation) between the largest and the second largest parameter. However, in many real situations, it is hard to assign the value of δ^* and, therefore, in case that the assumption of indifference zone is violated, the probability of a correct selection cannot be guaranteed to be at least P^*, a prespecified value.

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

In this paper, we are concerned with deriving a lower confidence bound for the probability of a correct selection for the general location model $F(x-\theta_i)$, $i = 1,\ldots,k$. First, we derive simultaneous lower confidence bounds on the differences between the largest (best) and each of the other non-best population parameters. Based on these, we obtain a lower confidence bound for the probability of a correct selection. The general result is then applied to the selection of the best mean of k normal populations with both the known and unknown common variances. In the first case one needs a single-stage procedure while in the second case a two-stage procedure is required. Some simulation investigations are described and their results are provided.