SIMULTANEOUS LOWER CONFIDENCE BOUNDS
FOR PROBABILITIES OF CORRECT SELECTIONS*

by
Shanti S. Gupta and Ta Chen Liang
Department of Statistics Department of Mathematics
Purdue University Wayne State University

Technical Report #95-27C

Department of Statistics
Purdue University

June 1995

*This research was supported in part by US Army Research Office, Grant DAAH04-95-1-0165.
SIMULTANEOUS LOWER CONFIDENCE BOUNDS
FOR PROBABILITIES OF CORRECT SELECTIONS*

Shanti S. Gupta and Ta Chen Liang
Department of Statistics Department of Mathematics
Purdue University Wayne State University

Abstract

In this paper, we are dealing with the problem of constructing lower confidence bounds for the $PCS_t$, simultaneously, for all $t = 1, \ldots, k - 1$, for the general location-parameter models, where $k$ is the number of populations involved in a selection problem and $PCS_t$ denotes the probability of correctly selecting the $t$ best populations. The result is then applied to the selection of the $t$ best means of normal populations for the two cases where the common variance may be known or unknown. An example is provided to illustrate the implementation and interpretation of the lower confidence bounds of the $PCS_t$.

Short Title: Lower Confidence Bounds For PCS.

AMS 1991 Subject Classification: 62F25; 62F07.

Keywords and phrases: Correct selection, lower confidence bound, $t$ best populations, probability of correct selection.

*This research was supported in part by US Army Research Office, Grant DAAH04-95-1-0165.
1. INTRODUCTION

Consider independent observations \( X_{ij}, j = 1, \ldots, n \), arising from population \( \pi_i \) with continuous cumulative distribution function \( G(x - \theta_i), i = 1, \ldots, k \). Let \( \theta = (\theta_1, \ldots, \theta_k) \) and let \( \theta(1) \leq \ldots \leq \theta(k) \) denote the ordered values of the parameters \( \theta_1, \ldots, \theta_k \). It is assumed that the exact pairing between the ordered and the unordered parameters is unknown. For each \( t = 1, \ldots, k - 1 \), the \( t \) best populations are those associated with the \( t \) largest parameters \( \theta(k), \ldots, \theta(k-t+1) \). Assume that the experimenter is interested in the selection of the \( t \) best populations. For this purpose, one may choose appropriate statistics \( Y_i = Y(X_{i1}, \ldots, X_{in}) \) for inference regarding \( \theta_i \) with continuous cumulative distribution function \( F(y - \theta_i), i = 1, \ldots, k \). Let \( Y_{[1]} \leq \ldots \leq Y_{[k]} \) denote the ordered statistics of \( Y_1, \ldots, Y_k \). One then applies the natural selection rule that selects those populations yielding the \( t \) largest \( Y_{[k]}, \ldots, Y_{[k-t+1]} \) as the \( t \) best populations. Thus, a question which arises naturally is: What kind of confidence statement can be made about this selection result?

Let \( CS_t \) (a correct selection of the \( t \) best populations) denote the event that the \( t \) best populations are actually selected. Also, let \( Y_{(i)} \) denote the \( Y \) statistic associated with the \( i \)-th ordered parameter \( \theta_{(i)} \). Thus, the probability of correctly selecting the \( t \) best populations (\( PCS_t \)) at \( \theta \) by applying the natural selection rule is:

\[
PCS_t(\theta) = P\left\{ \max_{1 \leq i \leq k-t} Y_{(i)} < \min_{k-t+1 \leq j \leq k} Y_{(j)} \right\} \\
= \int \prod_{i=1}^{k-t} \bar{F}(y - \theta_{(i)}) d\left\{ 1 - \prod_{j=k-t+1}^{k} F(y - \theta_{(j)}) \right\} \\
= \int \prod_{j=k-t+1}^{k} \bar{F}(y - \theta_{(j)}) d\left\{ \prod_{i=1}^{k-t} F(y - \theta_{(i)}) \right\} \quad (1.1a)
\]

where \( \bar{F} = 1 - F \).

In general, to guarantee the \( PCS_t(\theta) \) at least at a prespecified probability level, one needs to specify a positive number \( \delta^* \) such that \( \theta_{(k-t+1)} - \theta_{(k-t)} \geq \delta^* \); see Bechhofer (1954). Clearly, this indifference zone approach is formulated on the basis of designing an experiment.

Recently, retrospective analyses regarding the \( PCS_t \) have been studied by several
authors. Anderson, Bishop and Dudewicz (1977) and Lam (1989) have, respectively, given lower confidence bounds on the $PCS_1$ for normal distribution models. Kim (1986) has presented a lower confidence bound on the $PCS_1$ for the location-parameter models where the underlying density functions have the monotone likelihood ratio property. Gupta, Leu and Liang (1990) have constructed a lower confidence bound for the $PCS_1$ in the truncated location-parameter models following Kim's approach. Gupta and Liang (1991) have derived a lower confidence bound for $PCS_1$ for the general location-parameter models. Recently, Gupta, Liao, Qiu and Wang (GLQW) (1994) have proposed a new method for constructing a lower confidence bound for the $PCS_1$ under the case considered by Kim (1986). The lower confidence bound of GLQW is better than that of Kim in the sense that for the same confidence probability, the value of the lower confidence bound of GLQW is larger than that of Kim.

Note that in the previously referenced works, the investigations are made only for $t = 1$ case. Recently, Jeong, Kim and Jeon (1989) have developed a lower confidence bound on the $PCS_t$ for a fixed $t$, $1 \leq t \leq k - 1$, for the location-parameter models having the monotone likelihood ratio property. Also, the reader is referred to Olkin, Sobel and Tong (1976, 1982), Bofinger (1985) and Gutmann and Maymin (1987) for certain related problems regarding the $PCS_1$.

In this paper, we are concerned with the problem of deriving simultaneous lower confidence bounds for the $PCS_t$, $t = 1, ..., k - 1$, for the general location-parameter models. The result is then applied to the selection of the $t$ best means of normal populations. An example is provided to illustrate the implementation of the procedure.

2. SIMULTANEOUS LOWER CONFIDENCE BOUNDS FOR $PCS_t$

From (1.1a), the $PCS_t(\theta)$ can be written as:

$$PCS_t(\theta) = \sum_{j=k-t+1}^{k} P_{tj}(\theta)$$

where for each $j = k - t + 1, ..., k$,

$$P_{tj}(\theta) = \int \prod_{i=1}^{k-t} F(y + \Delta_{tji}(1)) \prod_{m=k-t+1}^{j-1} \bar{F}(y + \Delta_{tjm}(2)) \prod_{l=j+1}^{k} \bar{F}(y + \Delta_{tjl}(3)) dF(y),$$

3
and $\Delta_{tji}(1) = \theta(j) - \theta(i) \geq 0$ for $1 \leq i \leq k - t < j$; $\Delta_{tjm}(2) = \theta(j) - \theta(m) \geq 0$ for $k - t + 1 \leq m < j$, and $\Delta_{tji}(3) = \theta(j) - \theta(l) \leq 0$ for $k - t + 1 \leq j < l \leq k$. Here, $\prod_{s}^{t} \equiv 1$ if $t < s$. Note that for each $j, (k - t + 1 \leq j \leq k), P_{tj}(\theta)$ is increasing in $\Delta_{tji}(1)$, and decreasing in $\Delta_{tjm}(2)$ and $\Delta_{tji}(3)$, respectively. Thus if simultaneous lower confidence bounds for $\Delta_{tji}(1), 1 \leq i \leq k - t$, and upper confidence bounds for $\Delta_{tjm}(2)$ and $\Delta_{tji}(3), k - t + 1 \leq m \leq j \leq l \leq k, m \neq j, l \neq j$ for all $t = 1, 2, ..., k - 1$, can be obtained, then simultaneous lower confidence bounds for $PCS_t(\theta)$, for all $t = 1, ..., k - 1$, can also be established.

Also, note that, from (1.1b) the $PCS_t(\theta)$ can be expressed as

$$PCS_t(\theta) = \sum_{i=1}^{k-t} Q_{ti}(\theta),$$

(2.3)

where for each $i = 1, ..., k - t$,

$$Q_{ti}(\theta) = \int \prod_{m=1}^{i-1} F(z + \delta_{tim}(1)) \prod_{l=i+1}^{k-t} F(z + \delta_{til}(2)) \prod_{j=k-t+1}^{k} F(z + \delta_{tij}(3)) dF(z)$$

(2.4)

and $\delta_{tim}(1) = \theta(i) - \theta(m) \geq 0$ for $1 \leq m < i \leq k - t$; $\delta_{til}(2) = \theta(i) - \theta(l) < 0$ for $1 \leq i < l \leq k - t$; and $\delta_{tij}(3) = \theta(i) - \theta(j) < 0$ for $i \leq k - t < j \leq k$. Note that for each $i = 1, ..., k - t, Q_{ti}(\theta)$ is increasing in $\delta_{tim}(1)$ and $\delta_{til}(2)$ and decreasing in $\delta_{tij}(3)$, respectively. Thus if simultaneous lower confidence bounds for $\delta_{tim}(1)$ and $\delta_{til}(2), 1 \leq m \leq l \leq k - t, m \neq i, l \neq i$, and upper confidence bounds for $\delta_{tij}(3), i \leq k - t < j \leq k$, can be obtained, then based on (2.3)-(2.4), simultaneous lower confidence bounds for the $PCS_t(\theta)$, for all $t = 1, ..., k - 1$, can also be established.

In the following, a result of Lam (1986) is used to construct simultaneous lower confidence bounds for all $\Delta_{tji}(1), \delta_{tim}(1), \delta_{til}(2)$ and upper confidence bounds for all $\Delta_{tjm}(2), \Delta_{tji}(3)$ and $\delta_{tij}(3)$ for all $t = 1, ..., k - 1$.

For each $\alpha, 0 < \alpha < 1$, let $c(k, n, \alpha)$ be the value such that

$$P_{\theta}\left\{ \max_{1 \leq j \leq k} (Y_j - \theta_i) - \min_{1 \leq j \leq k} (Y_j - \theta_j) \leq c(k, n, d) \right\} = 1 - \alpha.$$  

(2.5)

Note that since $Y_i$ has a distribution function $F(y - \theta_i), i = 1, ..., k$, the value of $c = c(k, n, \alpha)$ is independent of the parameter $\theta$. Let

$$E = \{ \max_{1 \leq l \leq k} (Y_l - \theta_i) - \min_{1 \leq j \leq k} (Y_j - \theta_j) \leq c \}

$$

$$E_1 = \{ (Y_{[i]} - Y_{[j]} - c)^+ \leq \theta(i) - \theta(j) \leq Y_{[i]} - Y_{[j]} + c, \text{ for all } 1 \leq j < i \leq k \}$$

4
and

\[ E_2 = \{ Y[i] - Y[j] - c \leq \theta(i) - \theta(j) \leq (Y[i] - Y[j] + c)^{-} \text{ for all } 1 \leq i < j \leq k \} \]

where \( y^+ = \max(0, y) \) and \( y^- = \min(0, y) \).

**Lemma 2.1**

(a) \( E \subset E_1 \cap E_2 \), and therefore

(b) \( P_{\theta}(E_1 \cap E_2) \geq P_{\theta}(E) = 1 - \alpha \) for all \( \theta \).

**Proof:** By Theorem 4 of Lam (1986) and noting that \( \theta(i) - \theta(j) \geq 0 \) as \( j < i \) and \( \theta(i) - \theta(j) \leq 0 \) for \( i < j \), we have that \( E \subset E_1 \) and \( E \subset E_2 \). Therefore \( E \subset E_1 \cap E_2 \).

Now, part (b) follows immediately from part (a) and (2.5). \( \square \)

For each \( t = 1, ..., k - 1 \), and \( j = k - t + 1, ..., k \), let

\[
\begin{align*}
\hat{\Delta}_{tji}(1) &= (Y[i] - Y[j] - c)^+ \quad \text{for } 1 \leq i \leq k - t; \\
\hat{\Delta}_{tjm}(2) &= Y[j] - Y[m] + c \quad \text{for } k - t + 1 \leq m < j; \\
\hat{\Delta}_{tjl}(3) &= (Y[j] - Y[l] + c)^- \quad \text{for } j < l \leq k. 
\end{align*}
\]

(2.6)

Also, for each \( t = 1, ..., k - 1 \) and \( i = 1, ..., k - t \), let

\[
\begin{align*}
\hat{\delta}_{tim}(1) &= (Y[i] - Y[m] - c)^+ \quad \text{for } 1 \leq m \leq i - 1; \\
\hat{\delta}_{til}(2) &= Y[i] - Y[l] - c \quad \text{for } i + 1 \leq l \leq k - t; \\
\hat{\delta}_{tij}(3) &= (Y[i] - Y[j] + c)^- \quad \text{for } k - t + 1 \leq j \leq k. 
\end{align*}
\]

(2.7)

The following lemma is a direct result of Lemma 2.1.

**Lemma 2.2** With probability at least \( 1 - \alpha \), the following (S1) and (S2) hold simultaneously.

(S1) For each \( t = 1, ..., k - 1 \) and each \( j = k - t + 1, ..., k \),

\[
\Delta_{tji}(1) \geq \hat{\Delta}_{tji}(1) \quad \text{for all } i = 1, ..., k - t; \\
\Delta_{tjm}(2) \leq \hat{\Delta}_{tjm}(2) \quad \text{for all } k - t + 1 \leq m < j; \\
\Delta_{tlj}(3) \leq \hat{\Delta}_{tjl}(3) \quad \text{for all } j < l \leq k.
\]

(S2) For each \( t = 1, ..., k - 1 \), and each \( i = 1, ..., k - t \),

\[
\begin{align*}
\delta_{tim}(1) &\geq \hat{\delta}_{tim}(1) \quad \text{for all } 1 \leq m \leq i - 1; \\
\delta_{til}(2) &\geq \hat{\delta}_{til}(2) \quad \text{for all } i + 1 \leq l \leq k - t; \\
\delta_{tij}(3) &\leq \hat{\delta}_{tij}(3) \quad \text{for all } k - t + 1 \leq j \leq k.
\end{align*}
\]

\[ 5 \]
Now, for each $t = 1, \ldots, k - 1$ and each $j = k - t + 1, \ldots, k$, define

$$
\hat{P}_{tj} = \int \prod_{i=1}^{k-t} F(y + \hat{\Delta}_{tj'i}(1)) \prod_{m=k-t+1}^{j-1} \bar{F}(y + \hat{\Delta}_{tjm}(2)) \prod_{l=j+1}^{k} \bar{F}(y + \hat{\Delta}_{tjl}(3)) dF(y)
$$

(2.8)

and for each $t = 1, \ldots, k - 1$, define

$$
\hat{P}_t = \sum_{j=k-t+1}^{k} \hat{P}_{tj}.
$$

(2.9)

Also, for each $t = 1, \ldots, k - 1$ and each $i = 1, \ldots, k - t$, define

$$
\hat{Q}_{ti} = \int \prod_{m=1}^{i-1} F(z + \hat{\delta}_{tim}(1)) \prod_{l=i+1}^{k-t} F(z + \hat{\delta}_{tit}(2)) \prod_{j=k-t+1}^{k} \bar{F}(z + \hat{\delta}_{tij}(3)) dF(z)
$$

(2.10)

and

$$
\hat{Q}_t = \sum_{i=1}^{k-t} \hat{Q}_{ti}.
$$

(2.11)

Define

$$
P_{tL} = \max(\hat{P}_t, \hat{Q}_t).
$$

(2.12)

We propose $P_{tL}$ as an estimator of a lower confidence bound of the $PCS_t(\theta)$ for each $t = 1, \ldots, k - 1$. We have the following theorem.

**Theorem 2.1** $P_{\theta}(PCS_t(\theta) \geq P_{tL} \text{ for all } t = 1, \ldots, k - 1) \geq 1 - \alpha$ for all $\theta$.

**Proof:** Note that $P_{tj}(\theta)$ is increasing in $\Delta_{tj'i}(1)$ and decreasing in $\Delta_{tjm}(2)$ and $\Delta_{tjl}(3)$. Also, $Q_{ti}(\theta)$ is increasing in $\delta_{tim}(1)$ and $\delta_{tit}(2)$ and decreasing in $\delta_{tij}(3)$. Then by (2.2) and (2.8), and (2.4) and (2.10), and Lemma 2.2, we have

$$
P_{\theta} \left\{ \begin{array}{l}
P_{tj}(\theta) \geq \hat{P}_{tj} \quad \text{for all } j = k - t + 1, \ldots, k; \\
Q_{ti}(\theta) \geq \hat{Q}_{ti} \quad \text{for all } i = 1, \ldots, k - t;
\end{array} \right\} \geq 1 - \alpha.
$$

(2.13)

Then by (2.1), (2.9), (2.3), (2.11) and (2.13), we have

$$
1 - \alpha \leq P\{PCS_t(\theta) \geq \hat{P}_t, PCS_t(\theta) \geq \hat{Q}_t \text{ for all } t = 1, \ldots, k - 1\}
$$

$$
= P\{PCS_t(\theta) \geq P_{tL} \text{ for all } t = 1, \ldots, k - 1\}.
$$

Hence the proof of the theorem is complete. $\square$
Remark 2.1. In the literature, the problem of finding lower confidence bounds for the \(PCS_t(\hat{\theta})\) for a fixed \(t\) has been studied by several authors, including Anderson, Bishop and Dudewicz (1977), Kim (1986), Lam (1989), Gupta, Liu and Liang (1990), Gupta and Liang (1991), Jeong, Kim and Jeon (1989) and GLQW (1994). Except Jeong, Kim and Jeon (1989), all other authors only considered the case \(t = 1\). For \(t = 1\), the \(\hat{P}_1\) proposed in the present paper is essentially the same as the \(\hat{P}_L\), an estimator of a lower bound of \(PCS_1(\theta)\), proposed by Gupta and Liang (1991). Thus, Theorem 2.2 of Gupta and Liang (1991) can be viewed as a Corollary of Theorem 2.1 of this paper. One can see that without any loss in the guaranteed probability of confidence, say, \(1 - \alpha\), the result of Theorem 2.1 is much stronger than that of Gupta and Liang (1991).

3. SELECTION FOR NORMAL POPULATIONS IN TERMS OF MEANS

Let \(X_{ij}, j = 1, ..., n\), be a sample of size \(n\) arising from a normal population with mean \(\theta_i\) and variance \(\sigma^2, i = 1, ..., k\), where the common variance \(\sigma^2\) may be either known or unknown. For each \(i\), let \(Y_i = Y(X_{i1}, ..., X_{in}) = \frac{1}{n} \sum_{j=1}^{n} X_{ij}\). Then, based on the statistics \(Y_1, ..., Y_k\), by applying the natural selection rule, for each \(t = 1, ..., k - 1\), the associated \(PCS_t\) is:

\[
PCS_t(\theta) = \sum_{j=k-t+1}^{k} P_{ij}(\theta) = \sum_{i=1}^{k-t} Q_{ti}(\theta)
\]

(3.1)

where

\[
P_{ij}(\theta) = \int \prod_{i=1}^{k-t} \Phi(y + \frac{\sqrt{n}\Delta_{tji}(1)}{\sigma}) \prod_{m=k-t+1}^{j-1} \Phi(y + \frac{\sqrt{n}\Delta_{tjm}(2)}{\sigma}) \prod_{l=j+1}^{k} \Phi(y + \frac{\sqrt{n}\Delta_{tlj}(3)}{\sigma}) d\Phi(y)
\]

(3.2)

\[
Q_{ti}(\theta) = \int \prod_{m=1}^{i-1} \Phi(y + \frac{\sqrt{n}\delta_{tim}(1)}{\sigma}) \prod_{l=i+1}^{k-t} \Phi(y + \frac{\sqrt{n}\delta_{til}(2)}{\sigma}) \prod_{j=k-t+1}^{k} \Phi(y + \frac{\sqrt{n}\delta_{tlj}(3)}{\sigma}) d\Phi(y)
\]

(3.3)

and \(\Phi(\cdot)\) is the standard normal distribution function and \(\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)\). We consider two situations according to whether the common variance \(\sigma^2\) is either known or unknown.
3.1 SIMULTANEOUS CONFIDENCE BOUNDS FOR $PCS_t : \sigma^2$ KNOWN CASE

For each $\alpha, 0 < \alpha < 1$, let $c \equiv c(k, n, \alpha) = \frac{\sigma}{\sqrt{n}}q_k^{\alpha, \infty}$ where $q_k^{\alpha, \infty}$ is the 100$(1 - \alpha)$th percentile of Tukey's studentized range statistics with parameter $(k, \infty)$. The value of $q_k^{\alpha, \infty}$ is available from Harter (1969). Then, by the definition of $q_k^{\alpha, \infty}$,

$$P_\theta \{ \max_{1 \leq i \leq k} (Y_i - \theta_i) - \min_{1 \leq j \leq k} (Y_j - \theta_j) \leq c \} = 1 - \alpha \text{ for all } \theta.$$

For each $t = 1, \ldots, k - 1$, and each $j = k - t + 1, \ldots, k$, let

$$\hat{P}_{tj} = \prod_{i=1}^{k-t} \Phi(y + \frac{\sqrt{n} \hat{\Delta}_{tji}(1)}{\sigma}) \prod_{m=k-t+1}^{j-1} \Phi(y + \frac{\sqrt{n} \hat{\Delta}_{tjm}(2)}{\sigma}) \prod_{l=j+1}^{k} \Phi(y + \frac{\sqrt{n} \hat{\Delta}_{tji}(3)}{\sigma})d\Phi(y)$$

and for each $t = 1, \ldots, k - 1$, and each $i = 1, \ldots, k - t$,

$$\hat{Q}_{ti} = \prod_{m=1}^{i-1} \Phi(y + \frac{\sqrt{n} \hat{\delta}_{tim}(1)}{\sigma}) \prod_{l=i+1}^{k-t} \Phi(y + \frac{\sqrt{n} \hat{\delta}_{til}(2)}{\sigma}) \prod_{j=k-t+1}^{k} \Phi(y + \frac{\sqrt{n} \hat{\delta}_{tij}(3)}{\sigma})d\Phi(y).$$

(3.4)

(3.5)

where $\hat{\Delta}_{tji}(1), \hat{\Delta}_{tjm}(2),$ and $\hat{\Delta}_{tji}(3)$ are defined as (2.6) and $\hat{\delta}_{tim}(1), \hat{\delta}_{til}(2),$ and $\hat{\delta}_{tij}(3)$ are defined in (2.7), with $c \equiv c(k, n, \alpha) = \frac{\sigma}{\sqrt{n}}q_k^{\alpha, \infty}$.

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

$$\hat{P}_t = \sum_{j=k-t+1}^{k} \hat{P}_{tj}$$

$$\hat{Q}_t = \sum_{i=1}^{k-t} \hat{Q}_{ti}$$

(3.6)

(3.7)

Then by Theorem 2.1, we can conclude the following:

**Theorem 3.1.** $P_\theta \{ PCS_t(\theta) \geq \max(\hat{P}_t, \hat{Q}_t) \text{ for all } t = 1, \ldots, k - 1 \} \geq 1 - \alpha \text{ for all } \theta.$

3.2. SIMULTANEOUS LOWER CONFIDENCE BOUNDS FOR $PCS_t : \sigma^2$ UNKNOWN CASE

When the value of the common variance $\sigma^2$ is unknown, Theorem 2.1 can not be applied directly. In the following, it is assumed that the original selection goal of the
experiment is to select the best normal population (that is, \( t = 1 \)), and the two-stage sampling scheme of Bechhofer, Dunnett and Sobel (1954) is adopted. For completeness, the two-stage sampling scheme is described as follows.

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

\[
\int_{-\infty}^{\infty} \int_{0}^{\infty} [\Phi(x + wh)]^{k-1} d\Phi(x) dF_W(w) = P^* \]

where \( k^{-1} < P^* < 1 \), and \( F_W(w) \) is the distribution function of the nonnegative random variable \( W \) with \( k(n_0 - 1)W^2 \) following a \( \chi^2(k(n_0 - 1)) \) distribution. Then, take additional \( N - n_0 \) observations from each populations. Compute the overall means \( \bar{X}_i(N) = \frac{1}{N} \sum_{j=1}^{N} X_{ij}, i = 1, \ldots, k \).

It should be noted that in the preceding two-stage sampling scheme, the values of \( P^*, \delta^* \) and \( n_0 \) should be assigned before the selection is made.

Let \( \bar{X}_{(i)}(N) \) denote the random variable associated with the ranked parameter \( \theta_{(i)} \). Also, let \( \bar{X}_{[1]}(N) \leq \cdots \leq \bar{X}_{[k]}(N) \) be the ordered statistics of \( \bar{X}_1(N), \ldots, \bar{X}_k(N) \). According to the natural selection rule, for each \( t = 1, \ldots, k - 1 \), the populations which yield \( \bar{X}_{[k]}(N), \ldots, \bar{X}_{[k-t+1]}(N) \) are selected as the \( t \) best populations. For each \( t = 1, \ldots, k - 1 \), let \( A_t = \{(i, j) | k - t + 1 \leq i \leq k, 1 \leq j \leq k - t \} \). Then, the corresponding \( PCS_t(\theta) \) is:

\[
PCS_t(\theta) = \begin{cases} 
\bar{X}_{(i)}(N) > \bar{X}_{(j)}(N) & \text{for all } (i, j) \in A_t \\
\frac{\sqrt{N} (\bar{X}_{(i)}(N) - \theta_{(i)})}{\sigma} + \frac{\sqrt{N} (\theta_{(i)} - \theta_{(j)})}{\sigma^*} > \frac{\sqrt{N} (\bar{X}_{(j)}(N) - \theta_{(j)})}{\sigma} & \text{for all } (i, j) \in A_t \\ 
\frac{\sqrt{N} (\bar{X}_{(i)}(N) - \theta_{(i)})}{\sigma} + \frac{h(\theta_{(i)} - \theta_{(j)})}{\delta^*} \cdot \frac{\sqrt{N} (\bar{X}_{(j)}(N) - \theta_{(j)})}{\sigma} & \text{for all } (i, j) \in A_t \\
\frac{Z_i + h(\theta_{(i)} - \theta_{(j)})W}{\delta^*} > Z_j & \text{for all } (i, j) \in A_t \\
= \int_{0}^{\infty} P_{\theta} \left\{ Z_i + \frac{h(\theta_{(i)} - \theta_{(j)})W}{\delta^*} > Z_j \right\} dF_W(w) \\
= \int_{0}^{\infty} P_t(\theta, w) dF_W(w), 
\end{cases}
\]
where $Z_i \sim N(0,1), i = 1,...,k, k(n_0 - 1)W^2 \sim \chi^2(k(n_0 - 1))$ and $Z_1,...,Z_k$ and $W$ are mutually independent. Note that in (3.8), the inequality is obtained based on the fact that $N \geq \left[ \frac{S_k^2 k^2}{\delta^2} \right]$ and $\theta_{(i)} - \theta_{(j)} \geq 0$ for all $(i,j) \in A_{t}$.

Analogous to (2.1)-(2.2) and (2.3)-(2.4), respectively, $P_t(\theta, w)$ can be expressed as:

$$P_t(\theta, w) = \sum_{j=k-t+1}^{k} P_{tj}(\theta, w) = \sum_{i=1}^{k-t} Q_{ti}(\theta, w) \quad (3.9)$$

where for each $j = k - t + 1,...,k$,

$$P_{tj}(\theta, w) = \int \prod_{i=1}^{k-t} \Phi(y + \frac{hw\Delta_{tji}(1)}{\delta^*}) \prod_{m=k-t+1}^{j-1} \Phi(y + \frac{hw\Delta_{tjm}(2)}{\delta^*}) \prod_{l=j+1}^{k} \tilde{\Phi}(y + \frac{hw\Delta_{tjl}(3)}{\delta^*}) d\Phi(y), \quad (3.10)$$

and for each $i = 1,...,k - t$,

$$Q_{ti}(\theta, w) = \int \prod_{m=1}^{i-1} \Phi(y + \frac{hw\delta_{tim}(1)}{\delta^*}) \prod_{l=i+1}^{k-t} \Phi(y + \frac{hw\delta_{tli}(2)}{\delta^*}) \prod_{j=k-t+1}^{k} \tilde{\Phi}(y + \frac{hw\delta_{tij}(3)}{\delta^*}) d\Phi(y). \quad (3.11)$$

Combining (3.8)-(3.11) yields that for each $t = 1,...,k - 1$,

$$PCS_t(\theta) \geq \sum_{j=k-t+1}^{k} \int_{w=0}^{\infty} P_{tj}(\theta, w) dF_W(w) \quad (3.12)$$

and

$$PCS_t(\theta) \geq \sum_{i=1}^{k-t} \int_{w=0}^{\infty} Q_{ti}(\theta, w) dF_W(w). \quad (3.13)$$

Let $c^* = S_k^{a_{k,k(n_0-1)/2}}$, where $q_{k,k(n_0-1)}^{\alpha}$ is the $100(1 - \alpha) - th$ percentile of Tukey's studentized range statistics with parameter $(k, k(n_0 - 1))$. Then,

$$P\{ \max_{1 \leq i \leq k} (\bar{X_i}(N) - \theta_i) - \min_{1 \leq j \leq k} (\bar{X_j}(N) - \theta_j) \leq c^* \} = 1 - \alpha, \quad (3.14)$$
see Gupta and Liang (1991). Also, a result similar to that of Lemma 2.1 can be obtained as follows.

Let

\[
E_1^* = \{(\bar{X}_{[i]}(N) - \bar{X}_{[j]}(N) - c^*)^+ \leq \theta_{(i)} - \theta_{(j)} \leq \bar{X}_{[i]}(N) - \bar{X}_{[j]}(N) + c^* \text{ for } 1 \leq j < i \leq k \}
\]

\[
E_2^* = \{\bar{X}_{[i]}(N) - \bar{X}_{[j]}(N) - c^* \leq \theta_{[i]} - \theta_{[j]} \leq (\bar{X}_{[i]}(N) - \bar{X}_{[j]}(N) + c^*)^- \text{ for } 1 \leq j < i \leq k \}
\]

(3.15)

Then,

\[
P_{\tilde{\theta}}\{E_1^* \cap E_2^* \} \geq 1 - \alpha \text{ for all } \tilde{\theta}.
\]

(3.16)

Now, for each \(j = k - t + 1, \ldots, k\), let

\[
P_{t_j^*}(w) = \int \prod_{i=1}^{k-t} \Phi(y + \frac{hw'\hat{\Delta}_{tji}(1)}{\delta^*}) \prod_{m=k-t+1}^{j-1} \Phi(y + \frac{hw'\hat{\Delta}_{tjm}(2)}{\delta^*})
\]

\[
\times \prod_{l=j+1}^{k} \Phi(y + \frac{hw'\hat{\Delta}_{tji}(3)}{\delta^*})d\Phi(y),
\]

(3.17)

where \(\hat{\Delta}_{tji}(1), \hat{\Delta}_{tjm}(2)\) and \(\hat{\Delta}_{tji}(3)\) are defined as (2.6) with \(Y_{[i]}\) being replaced by \(\bar{X}_{[i]}(N)\) and \(c = c^* = S_{q_k,k(n_0-1)}/\sqrt{N}\). Also, for each \(i = 1, \ldots, k - t\), let

\[
Q_{i}^*(w) = \int \prod_{m=1}^{i-1} \Phi(y + \frac{hw'\hat{\delta}_{tim}(1)}{\delta^*}) \prod_{l=i+1}^{k-t} \Phi(y + \frac{hw'\hat{\delta}_{iti}(2)}{\delta^*})
\]

\[
\times \prod_{j=k-t+1}^{K} \Phi(y + \frac{hw'\hat{\delta}_{tij}(3)}{\delta^*})d\Phi(y)
\]

(3.18)

where \(\hat{\delta}_{tim}(1), \hat{\delta}_{iti}(2)\) and \(\hat{\delta}_{tij}(3)\) are defined in (2.7) with \(c = c^*\) and \(Y_{[i]}\) being replaced by \(\bar{X}_{[i]}(N)\).

Let

\[
P_t^* = \sum_{j=k-t+1}^{k} \int_0^\infty P_{t_j^*}(w)dF_W(w),
\]

(3.19)

and

\[
Q_t^* = \sum_{i=1}^{k-t} \int_0^\infty Q_{i}^*(w)dF_W(w).
\]

(3.20)

From (3.16)-(3.18), we see that for each \(w > 0\),
\[
    P \left\{ P_{ij}(\theta, w) \geq P^*_j(w) \text{ and } Q_{ii}(\theta, w) \geq Q^*_i(w) \right\} \geq 1 - \alpha 
\]

for all $\theta$.

Combining (3.17)-(3.21), we conclude the following Theorem.

**Theorem 3.2** \[ P_\theta(PCS_t(\theta)) \geq \max(P^*_t, Q^*_t) \text{ for all } t = 1, \ldots, k - 1 \] \[ \geq 1 - \alpha \text{ for all } \theta. \]

4. **AN ILLUSTRATIVE EXAMPLE**

In the following example, the data is taken from Problem 3.1, page 97, of Gibbons, Olkin and Sobel (1977) with some modification.

The experimenter wants to compare dry shear strength of $k = 6$ different resin glues for bonding yellow birch plywood. Assume that the distribution of the strength for each glue are normal with common variance $\sigma^2 = 400$. From each kind of glue, a sample of size 10 is taken. The data is given in Table 1. The observations (readings) are taken to measure the strength of the glue. Thus, large values are more desirable in their application.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>102</td>
<td>70</td>
<td>100</td>
<td>120</td>
<td>151</td>
<td>220</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>83</td>
<td>102</td>
<td>125*</td>
<td>156</td>
<td>243</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>78</td>
<td>80</td>
<td>182</td>
<td>192</td>
<td>189</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>93</td>
<td>119</td>
<td>130</td>
<td>162</td>
<td>176</td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>98</td>
<td>59</td>
<td>130*</td>
<td>166</td>
<td>176</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>66</td>
<td>99</td>
<td>143</td>
<td>158</td>
<td>181</td>
<td></td>
</tr>
<tr>
<td>117</td>
<td>92</td>
<td>100</td>
<td>113</td>
<td>173</td>
<td>206</td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>79</td>
<td>109</td>
<td>140</td>
<td>157</td>
<td>233</td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>134</td>
<td>117*</td>
<td>123</td>
<td>233</td>
<td>162</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>131</td>
<td>100*</td>
<td>132</td>
<td>238</td>
<td>179</td>
<td></td>
</tr>
</tbody>
</table>

*Indicates that the entry has been changed and is different from the one in Gibbons, Olkin and Sobel (1972).

We have the sample mean values: $\bar{X}_1 = 78.8, \bar{X}_2 = 92.4, \bar{X}_3 = 98.5, \bar{X}_4 = 133.8, \bar{X}_5 = 178.6$, and $\bar{X}_6 = 196.5$. Note that $\bar{X}_1 \leq \bar{X}_2 \leq \bar{X}_3 \leq \bar{X}_4 < \bar{X}_5 < \bar{X}_6$. Hence, according to the natural selection rule, for each $t = 1, \ldots, 5$, Glue $j'$s, $6 - t + 1 \leq j \leq 6$, are selected as
the \( t \) best glues. For \( t = 1 \), Glue 6 which yields the largest sample mean value, is selected as the best. However, it is possible that the selected one may not be the best. Hence, a reasonable question is: What kind of confidence statement can be made regarding the \( PCS_1 \)? A more common question may be: How many populations should be selected according to the data? The result of Theorem 3.1 may shed some light on this aspect.

Based on the data, for \( \alpha = 0.1, \hat{P}_t \) and \( \hat{Q}_t, t = 1, \ldots, k - 1 \), are computed and the result is as follows.

<table>
<thead>
<tr>
<th>( t )</th>
<th>( \hat{P}_t )</th>
<th>( \hat{Q}_t )</th>
<th>( \max(\hat{P}_t, \hat{Q}_t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5000</td>
<td>0.4785</td>
<td>0.5000</td>
</tr>
<tr>
<td>2</td>
<td>0.4785</td>
<td>0.7231</td>
<td>0.7231</td>
</tr>
<tr>
<td>3</td>
<td>0.7231</td>
<td>0.1974</td>
<td>0.7231</td>
</tr>
<tr>
<td>4</td>
<td>0.1969</td>
<td>0.2211</td>
<td>0.2211</td>
</tr>
<tr>
<td>5</td>
<td>0.2211</td>
<td>0.2662</td>
<td>0.2662</td>
</tr>
</tbody>
</table>

Therefore, we can state, with at least 90% confidence, that simultaneously \( PCS_1(\theta) \geq 0.5000, PCS_2(\theta) \geq 0.7231, PCS_3(\theta) \geq 0.7231, PCS_4(\theta) \geq 0.2211, PCS_5(\theta) \geq 0.2662 \). Hence one may like to select the two best instead of the best.
References


Tests Based on Range and Studentized Range of Samples from a Normal Population, Aerospace Research Laboratories.


