SIMULTANEOUS LOWER CONFIDENCE BOUNDS FOR PROBABILITIES OF CORRECT SELECTIONS*

by

Shanti S. Gupta and Ta Chen Liang
Department of Statistics, Purdue University
Department of Mathematics, 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, Purdue University
Department of Mathematics, 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 $t = 1, ..., k-1$, 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 $X_{ij}, j = 1,...n$, arising from population π_i with continuous cumulative distribution function $G(x - \theta_i), i = 1,...,k$. Let $\underset{\sim}{\theta} = (\theta_1,...,\theta_k)$ and let $\theta_{(1)} \leq ... \leq \theta_{(k)}$ denote the ordered values of the parameters $\theta_1,...,\theta_k$. It is assumed that the exact pairing between the ordered and the unordered parameters is unknown. For each $t = 1,...,k-1$, the t best populations are those associated with the t largest parameters $\theta_{(k)},...,\theta_{(k-t+1)}$. Assume that the experimenter is interested in the selection of the t best populations. For this purpose, one may choose appropriate statistics $Y_i = Y(X_{il},...,X_{in})$ for inference regarding θ_i with continuous cumulative distribution function $F(y-\theta_i), i = 1,...,k$. Let $Y_{[1]} \leq ... \leq Y_{[k]}$ denote the ordered statistics of $Y_1,...,Y_k$. One then applies the natural selection rule that selects those populations yielding the t largest $Y_{[k]},...,Y_{[k-t+1]}$ 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 $\underset{\sim}{\theta}$ by applying the natural selection rule is:

$$
\begin{aligned}
&PCS_t(\underset{\sim}{\theta}) \\
&= P\{\max_{1\leq i\leq k-t} Y_{(i)} < \min_{k-t+1\leq j\leq k} Y_{(j)}\} \\
&= \int \prod_{i=1}^{k-t} \bar{F}(y-\theta_{(i)}) d\{1 - \prod_{j=k-t+1}^{k} F(y-\theta_{(j)}\} && (1.1a) \\
&= \int \prod_{j=k-t+1}^{k} \bar{F}(y-\theta_{(j)}) d \prod_{i=1}^{k-t} F(y-\theta_{(i)}) && (1.1b)
\end{aligned}
$$

where $\bar{F} = 1 - F$.

In general, to guarantee the $PCS_t(\underset{\sim}{\theta})$ at least at a prespecified probability level, one needs to specify a positive number δ^* 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 $PCS_t(\underset{\sim}{\theta})$ can be written as:

$$PCS_t(\underset{\sim}{\theta}) = \sum_{j=k-t+1}^{k} P_{tj}(\underset{\sim}{\theta}) \tag{2.1}$$

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

$$P_{tj}(\underset{\sim}{\theta}) = \int \prod_{i=1}^{k-t} F(y + \Delta_{tji}(1)) \prod_{m=k-t+1}^{j-1} \bar{F}(y + \Delta_{tjm}(2)) \prod_{l=j+1}^{k} \bar{F}(y + \Delta_{tjl}(3)) dF(y), \tag{2.2}$$

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

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

$$PCS_t(\underset{\sim}{\theta}) = \sum_{i=1}^{k-t} Q_{ti}(\underset{\sim}{\theta}), \tag{2.3}$$

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

$$Q_{ti}(\underset{\sim}{\theta}) = \int \prod_{m=1}^{i-1} F(z + \delta_{tim}(1)) \prod_{l=i+1}^{k-t} F(z + \delta_{til}(2)) \prod_{j=k-t+1}^{k} \bar{F}(z + \delta_{tij}(3)) dF(z) \tag{2.4}$$

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 $i = 1, ..., k-t, Q_{ti}(\underset{\sim}{\theta})$ 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 $\delta_{til}(2)$, $1 \leq m \leq i \leq l \leq k-t, m \neq i, l \neq i$, and upper confidence bounds for $\delta_{til}(3), i \leq k-t < j \leq k$, can be obtained, then based on (2.3)-(2.4), simultaneous lower confidence bounds for the $PCS_t(\underset{\sim}{\theta})$, 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 $\Delta_{tjm}(2), \Delta_{tjl}(3)$ and $\delta_{til}(3)$ for all $t = 1, ..., k-1$.

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

$$P_{\underset{\sim}{\theta}}\{\max_{l \leq i \leq k}(Y_i - \theta_i) - \min_{l \leq j \leq k}(Y_j - \theta_j) \leq c(k, n, d)\} = 1 - \alpha. \tag{2.5}$$

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 $\underset{\sim}{\theta}$. Let

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

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

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 $y^- = min(0, y)$.

<u>*Lemma 2.1*</u>

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

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

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

$$\begin{cases} \hat{\Delta}_{tji}(1) = (Y_{[j]} - Y_{[i]} - c)^+ & \text{for } 1 \leq i \leq k-t; \\ \hat{\Delta}_{tjm}(2) = Y_{[j]} - Y_{[m]} + c & \text{for } k-t+1 \leq m < j; \\ \hat{\Delta}_{tjl}(3) = (Y_{[j]} - Y_{[l]} + c)^- & \text{for } j < l \leq k. \end{cases} \tag{2.6}$$

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

$$\begin{cases} \hat{\delta}_{tim}(1) = (Y_{[i]} - Y_{[m]} - c)^+ & \text{for } 1 \leq m \leq i-1; \\ \hat{\delta}_{til}(2) = Y_{[i]} - Y_{[l]} - c & \text{for } i+1 \leq l \leq k-t; \\ \hat{\delta}_{tij}(3) = (Y_{[i]} - Y_{[j]} + c)^- & \text{for } k-t+1 \leq j \leq k. \end{cases} \tag{2.7}$$

The following lemma is a direct result of Lemma 2.1.

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

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

$$\Delta_{tji}(1) \geq \hat{\Delta}_{tji}(1) \text{ for all } i = 1, ..., k-t;$$

$$\Delta_{tjm}(2) \leq \hat{\Delta}_{tjm}(2) \text{ for all } k-t+1 \leq m < j;$$

$$\Delta_{tjl}(3) \leq \hat{\Delta}_{tjl}(3) \text{ for all} j < l \leq k.$$

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

$$\delta_{tim}(1) \geq \hat{\delta}_{tim}(1) \text{ for all } 1 \leq m \leq i-1;$$

$$\delta_{til}(2) \geq \hat{\delta}_{til}(2) \text{ for all } i+1 \leq l \leq k-t;$$

$$\delta_{tij}(3) \leq \hat{\delta}_{tij}(3) \text{ for all } k-t+1 \leq j \leq k.$$

Now, for each $t = 1, ..., k-1$ and each $j = k-t+1, ..., k$, define

$$\hat{P}_{tj} = \int \prod_{i=1}^{k-t} F(y + \hat{\Delta}_{tji}(1)) \prod_{m=k-t+1}^{j-1} \bar{F}(y + \hat{\Delta}_{tjm}(2)) \prod_{l=j+1}^{k} \bar{F}(y + \hat{\Delta}_{tjl}(3)) dF(y) \quad (2.8)$$

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

$$\hat{P}_t = \sum_{j=k-t+1}^{k} \hat{P}_{tj}. \quad (2.9)$$

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

$$\hat{Q}_{ti} = \int \prod_{m=1}^{i-1} F(z + \hat{\delta}_{tim}(1)) \prod_{l=i+1}^{k-t} F(z + \hat{\delta}_{til}(2)) \prod_{j=k-t+1}^{k} \bar{F}(z + \hat{\delta}_{tij}(3)) dF(z) \quad (2.10)$$

and

$$\hat{Q}_t = \sum_{i=1}^{k-t} \hat{Q}_{ti} \quad (2.11)$$

Define

$$P_{tL} = max(\hat{P}_t, \hat{Q}_t). \quad (2.12)$$

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

<u>*Theorem 2.1*</u> $P_{\underset{\sim}{\theta}}\{PCS_t(\underset{\sim}{\theta}) \geq P_{tl}$ for all $t = 1, ..., k-1\} \geq 1 - \alpha$ for all $\underset{\sim}{\theta}$.

Proof: Note that $P_{tj}(\underset{\sim}{\theta})$ is increasing in $\Delta_{tji}(1)$ and decreasing in $\Delta_{tjm}(2)$ and $\Delta_{tjl}(3)$. Also, $Q_{ti}(\underset{\sim}{\theta})$ is increasing in $\delta_{tim}(1)$ and $\delta_{til}(2)$ 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

$$P_{\underset{\sim}{\theta}} \left\{ \begin{array}{ll} P_{tj}(\underset{\sim}{\theta}) \geq \hat{P}_{tj} & \text{for all } j = k-t+1, ..., k; \\ & \qquad\qquad \text{for all } t = 1, k-1 \\ Q_{ti}(\underset{\sim}{\theta}) \geq \hat{Q}_{ti} & \text{for all } i = 1, ..., k-t; \end{array} \right\} \geq 1 - \alpha. \quad (2.13)$$

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

$$\begin{aligned} 1 - \alpha &\leq P\{PCS_t(\underset{\sim}{\theta}) \geq \hat{P}_t, PCS_t(\underset{\sim}{\theta}) \geq \hat{Q}_t \ \text{ for all } \ t = 1, ..., k-1\} \\ &= P\{PCS_t(\underset{\sim}{\theta}) \geq P_{tL} \ \text{ for all } \ t = 1, ..., k-1\}. \end{aligned}$$

Hence the proof of the theorem is complete. □

Remark 2.1. In the literature, the problem of finding lower confidence bounds for the $PCS_t(\underset{\sim}{\theta})$ 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 $PCS_1(\underset{\sim}{\theta})$, 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 θ_i and variance $\sigma^2, i = 1, ..., k$, where the common variance σ^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:

$$\begin{aligned} PCS_t(\underset{\sim}{\theta}) &= \sum_{j=k-t+1}^{k} P_{tj}(\underset{\sim}{\theta}) \\ &= \sum_{i=1}^{k-t} Q_{ti}(\underset{\sim}{\theta}) \end{aligned} \tag{3.1}$$

where

$$P_{tj}(\underset{\sim}{\theta}) = \int \prod_{i=1}^{k-t} \Phi(y + \frac{\sqrt{n}\Delta_{tji}(1)}{\sigma}) \prod_{m=k-t+1}^{j-1} \bar{\Phi}(y + \frac{\sqrt{n}\Delta_{tjm}(2)}{\sigma}) \prod_{l=j+1}^{k} \bar{\Phi}(y + \frac{\sqrt{n}\Delta_{tjl}(3)}{\sigma}) d\Phi(y) \tag{3.2}$$

$$Q_{ti}(\underset{\sim}{\theta}) = \int \prod_{m=1}^{i-1} \Phi(y + \frac{\sqrt{n}\delta_{tim}(1)}{\sigma}) \prod_{l=i+1}^{k-t} \Phi(y + \frac{\sqrt{n}\delta_{til}(2)}{\sigma}) \prod_{j=k-t+1}^{k} \bar{\Phi}(y + \frac{\sqrt{n}\delta_{tij}(3)}{\sigma}) d\Phi(y) \tag{3.3}$$

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 σ^2 is either known or unknown.

3.1 SIMULTANEOUS CONFIDENCE BOUNDS FOR PCS_t : σ^2 KNOWN CASE

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

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

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

$$\hat{P}_{tj}=\int\prod_{i=1}^{k-t}\Phi(y+\frac{\sqrt{n}\hat{\Delta}_{tji}(1)}{\sigma})\prod_{m=k-t+1}^{j-1}\bar{\Phi}(y+\frac{\sqrt{n}\hat{\Delta}_{tjm}(2)}{\sigma})\prod_{l=j+1}^{k}\bar{\Phi}(y+\frac{\sqrt{n}\hat{\Delta}_{tjl}(3)}{\sigma})d\Phi(y) \tag{3.4}$$

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

$$\hat{Q}_{ti}=\int\prod_{m=1}^{i-1}\Phi(y+\frac{\sqrt{n}\hat{\delta}_{tim}(1)}{\sigma})\prod_{l=i+1}^{k-t}\Phi(y+\frac{\sqrt{n}\hat{\delta}_{til}(2)}{\sigma})\prod_{j=k-t+1}^{k}\bar{\Phi}(y+\frac{\sqrt{n}\hat{\delta}_{tij}(3)}{\sigma})d\Phi(y). \tag{3.5}$$

where $\hat{\Delta}_{tji}(1), \hat{\Delta}_{tjm}(2)$, and $\hat{\Delta}_{tjl}(3)$ 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 $c \equiv c(k,n,\alpha) = \frac{\sigma}{\sqrt{n}} q^{\alpha}_{k,\infty}$.

For each $t = 1, ..., k-1$, let

$$\hat{P}_t=\sum_{j=k-t+1}^{k}\hat{P}_{tj} \tag{3.6}$$

$$\hat{Q}_t=\sum_{i=1}^{k-t}\hat{Q}_{ti} \tag{3.7}$$

Then by Theorem 2.1, we can conclude the following:

Theorem 3.1. $P_{\underset{\sim}{\theta}}\{PCS_t(\underset{\sim}{\theta}) \ge max(\hat{P}_t, \hat{Q}_t)$ for all $t = 1, ..., k-1\} \ge 1-\alpha$ for all $\underset{\sim}{\theta}$.

3.2. SIMULTANEOUS LOWER CONFIDENCE BOUNDS FOR PCS_t : σ^2 UNKNOWN CASE

When the value of the common variance σ^2 is unknown, Theorem 2.1 can not be applied directly. In the following, it is assumed that the original selection goal of the experimenter 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 $\bar{X}_i = \frac{1}{n_0} \sum_{j=1}^{n_0} X_{ij}, i = 1, ..., k$, and $S^2 = \frac{1}{k(n_0-1)} \sum_{i=1}^{k} \sum_{j=1}^{n_o} (X_{ij} - \bar{X}_i)^2$. Define $N = max(n_0, [\frac{S^2 h^2}{\delta^{*2}}])$, where $[y]$ denotes the smallest integer not less than y and h is a positive value such that

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

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 $\bar{X}_i(N) = \frac{1}{N} \sum_{j=1}^{N} X_{ij}, i = 1, ..., k$.

It should be noted that in the preceding two-stage sampling scheme, the values of P^*, δ^* 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 $\bar{X}_1(N), ..., \bar{X}_k(N)$. According to the natural selection rule, for each $t = 1, ..., k-1$, the populations which yield $\bar{X}_{[k]}(N), ..., \bar{X}_{[k-t+1]}(N)$ are selected as the t best populations. For each $t = 1, ..., k-1$, let $A_t = \{(i,j) | k - t + 1 \leq i \leq k, 1 \leq j \leq k - t\}$. Then, the corresponding $PCS_t(\underset{\sim}{\theta})$ is:

$$\begin{aligned}
&PCS_t(\underset{\sim}{\theta}) \\
=&P_{\underset{\sim}{\theta}}\{\bar{X}_{(i)}(N) > \bar{X}_{(j)}(N) \text{ for all } (i,j)\epsilon A_t\} \\
=&P_{\underset{\sim}{\theta}}\{\frac{\sqrt{N}(\bar{X}_{(i)}(N) - \theta_{(i)})}{\sigma} + \frac{\sqrt{N}(\theta_{(i)} - \theta_{(j)})}{\sigma^*} \cdot \frac{S}{\sigma} > \frac{\sqrt{N}(\bar{X}_{(j)}(N) - \theta_{(j)})}{\sigma} \text{ for all } (i,j)\epsilon A_t\} \\
\geq&P_{\underset{\sim}{\theta}}\{\frac{\sqrt{N}(\bar{X}_i(N) - \theta_{(i)})}{\sigma} + \frac{h(\theta_{(i)} - \theta(j))}{\delta^*} \cdot \frac{S}{\sigma} > \frac{\sqrt{N}(\bar{X}_{(j)}(N) - \theta_{(j)})}{\sigma} \text{ for all } (i,j)\epsilon A_t\} \\
=&P_{\underset{\sim}{\theta}}\{Z_i + \frac{h(\theta_{(i)} - \theta_{(j)})W}{\delta^*} > Z_j \text{ for all } (i,j)\epsilon A_t\} \\
=&\int_0^{\infty} P_{\underset{\sim}{\theta}}\{Z_i + \frac{h(\theta_{(i)} - \theta_{(j)})W}{\delta^*} > Z_j \text{ for all } (i,j)\epsilon A_t\} dF_W(w) \\
=&\int_0^{\infty} P_t(\underset{\sim}{\theta}, w) dF_W(w), \qquad (3.8)
\end{aligned}$$

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 $N \geq [\frac{S^2h^2}{\delta^{*2}}]$ and $\theta_{(i)} - \theta_{(j)} \geq 0$ for all $(i,j)\epsilon A_t$.

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

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

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

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

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

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

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

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

and

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

Let $c^* = Sq^{\alpha}_{k,k(n_0-1)}/\sqrt{N}$, where $q^{\alpha}_{k,k(n_0-1)}$ is the $100(1-\alpha)-th$ percentile of Tukey's studentized range statistics with parameter $(k, k(n_0 - 1))$. Then,

$$P\{\max_{1\leq i\leq k}(\bar{X}_i(N) - \theta_i) - \min_{1\leq j\leq k}(\bar{X}_j(N) - \theta_j) \leq c^*\} = 1 - \alpha, \tag{3.14}$$

see Gupta and Liang (1991). Also, a result similar to that of Lemma 2.1 can be obtained as follows.

Let

$$\begin{cases} E_1^* = \{(\bar{X}_{[i]}(N) - \bar{X}_{[j]}(N) - c^*)^+ \leq \theta_{(i)} - \theta_{(j)} \leq \bar{X}_{[i]}(N) - \bar{X}_{[j]}(N) + c^* & \text{for } 1 \leq j < i \leq k\} \\ E_2^* = \{\bar{X}_{[i]}(N) - \bar{X}_{[j]}(N) - c^* \leq \theta_{[i]} - \theta_{[j]} \leq (\bar{X}_{[i]}(N) - \bar{X}_{[j]}(N) + c^*)^- & \text{for } 1 \leq j < i \leq k\} \end{cases} \tag{3.15}$$

Then,

$$P_{\underset{\sim}{\theta}}\{E_1^* \cap E_2^*\} \geq 1 - \alpha \text{ for all } \underset{\sim}{\theta}. \tag{3.16}$$

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

$$\begin{aligned} P_{tj}^*(w) = \int \prod_{i=1}^{k-t} \Phi(y + \frac{hw\hat{\Delta}_{tji}(1)}{\delta^*}) \prod_{m=k-t+1}^{j-1} \bar{\Phi}(y + \frac{hw\hat{\Delta}_{tjm}(2)}{\delta^*}) \\ \times \prod_{l=j+1}^{k} \bar{\Phi}(y + \frac{hw\hat{\Delta}_{tjl}(3)}{\delta^*}) d\Phi(y), \end{aligned} \tag{3.17}$$

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

$$\begin{aligned} Q_{ti}^*(w) = \int \prod_{m=1}^{i-1} \Phi(y + \frac{hw\hat{\delta}_{tim}(1)}{\delta^*}) \prod_{l=i+1}^{k-t} \Phi(y + \frac{hw\hat{\delta}_{til}(2)}{\delta^*}) \\ \times \prod_{j=k-t+1}^{K} \bar{\Phi}(y + \frac{hw\hat{\delta}_{tij}(3)}{\delta^*}) d\Phi(y) \end{aligned} \tag{3.18}$$

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

Let

$$P_t^* = \sum_{j=k-t+1}^{k} \int_0^{\infty} P_{tj}^*(w) dF_W(w), \tag{3.19}$$

and

$$Q_t^* = \sum_{i=1}^{k-t} \int_0^{\infty} Q_{ti}^*(w) dF_W(w). \tag{3.20}$$

From (3.16)-(3.18), we see that for each $w > 0$,

$$P\left\{\begin{array}{l} P_{tj}(\underset{\sim}{\theta}, w) \geq P^*_{tj}(w) \text{ and } Q_{ti}(\underset{\sim}{\theta}, w) \geq Q^*_{ti}(w) \\ \text{for all } (j,i) \in A_t, \text{ for all } t = 1, \ldots, k-1 \end{array}\right\} \geq 1 - \alpha \tag{3.21}$$

for all $\underset{\sim}{\theta}$.

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

Theorem 3.2 $P_{\underset{\sim}{\theta}}\{PCS_t(\underset{\sim}{\theta}) \geq max(P^*_t, Q^*_t)$ for all $t = 1, ..., k-1\} \geq 1 - \alpha$ for all $\underset{\sim}{\theta}$.

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 1. Shear Strength of Six Types of Glue

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

*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 $\bar{X}_1 < \bar{X}_2 < \bar{X}_3 < \bar{X}_4 < \bar{X}_5 < \bar{X}_6$. Hence, according to the natural selection rule, for each $t = 1, ..., 5$, 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 $\hat{Q}_t, t = 1, ..., k-1$, are computed and the result is as follows.

t	1	2	3	4	5
$\hat{P}_t$	0.5000	0.4785	0.7231	0.1969	0.2211
$\hat{Q}_t$	0.4785	0.7231	0.1974	0.2211	0.2662
$max(\hat{P}_t, \hat{Q}_t)$	0.5000	0.7231	0.7231	0.2211	0.2662

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

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., 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. (1988). "On the non-existence of consistent estimators for P(CS)". *Amer. J. Math. Manag. Sci.* 5, 63-76.

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

Gupta, S.S., Leu, L.Y. and Liang, T. (1990). "On lower confidence bounds for PCS in truncated location paramenter models". *Commun. Statist. Theory Method* 19(2), 527-546.

Gupta, S.S. and Liang, T. (1991). "On a lower confidence bound for the probability of a correct selection: analytical and simulation studies". In *The Frontiers of Statistical Scientific Theory & Industrial Applicaitons (Volume II of the Proceedings of ICOSCO-I, The First International Conference on Statistical Computing, 30 March- 2 April, 1987).* American Science Press, Inc., Columbus, Ohio, 77-95.

Gupta, S.S., Liao, Y., Qiu, C. and Wang, J. (1994). "A new technique for improved confidence bounds for the probability of correct selection." *Statistical Sinica, 4*, 715-727.

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.

Gutmann, S. and Maymin, Z. (1987). "Is the selected population the best?" *Ann. Statist. 15*, 456-461.

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.

Jeong, G.J., Kim, W.C. and Jeon, J.W. (1989). "A lower confidence bound on the probability of a correct selection of the t best populations." *J. Korea Statist. Soc., 18*, 26-37.

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

Lam, K. (1986). "A new procedure for selecting good populations." *Biometrika 73*, 201-206.

Lam, K. (1989). "The multiple comparison of ranked parameters." *Commun. Statist. Theory Method, 18(4)*, 1217-1237.

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 IV, Vol. 2* (Eds. S.S. Gupta and J.O. Berger), Academic Press, New York, 193-212.