Journal of the Korean
Statistical Society
Vol. 21, No. 1, 1992

A LOWER BOUND ON THE PROBABILITY OF CORRECT SELECTION FOR TWO-STAGE SELECTION PROCEDURE

Soonki Kim[1]

ABSTRACT

This paper provides a method of obtaining a lower bound on the probability of correct selection for a two-stage selection procedure. The resulting lower bound sharpens that by Tamhane and Bechhofer(1979) for the normal means problem with a common known variance. The design constants associated with the lower bound are computed and the results of the performance comparisons are given.

1. INTRODUCTION

Let $\pi_1, \ldots, \pi_k$ denote normal populations with unknown means $\theta_1, \ldots, \theta_k$ and a common known variance $\sigma^2 > 0$, and let $\theta_{[1]} \leq \cdots \leq \theta_{[k]}$ denote the ordered θ_i's where the correct pairing between θ_i and $\theta_{[i]}$ are unknown. For selecting the population associated with $\theta_{[k]}$, which is often regarded as the best population, Bechhofer(1954) introduced the so-called indifference-zone approach which requires the probability of correct selection(CS) of a selection procedure R to satisfy

$$P_\theta\{CS|R\} \geq p^* \quad \text{for all } \theta \in \Omega(\delta^*), \tag{1.1}$$

[1]Department of Statistics, Chŏn-buk National University, Chŏnju, Chŏnbuk, 560-756, Korea.

where

$$\Omega(\delta^*) = \{(\theta_1, \ldots, \theta_k) \mid \theta_{[k]} - \theta_{[k-1]} \geq \delta^*\},$$

and $p^*(1/k < p^* < 1)$ and $\delta^* > 0$ are pre-specified numbers.

Alam(1970) and Tamhane and Bechhofer(1977,1979) studied the following two-stage elimination type selection procedure, which uses Gupta's(1965) and Bechhofer's(1954) at the first and the second stage, respectively :

(stage 1) Take kn independent observations X_{ij} from $\pi_i (i = 1, \ldots, k$; $j = 1, \ldots, n)$ and determine a subset I of $\{1, 2, \ldots, k\}$ by

$$I = \{i | \bar{X}_i^{(1)} \geq \max_{1 \leq j \leq k} \bar{X}_j^{(1)} - d\sigma/\sqrt{n}\}, \tag{1.2}$$

where $\bar{X}_i^{(1)} = \sum_{j=1}^{n} X_{ij}/n$ is the sample mean and $d > 0$ is a design constant to be determined. If I has only one element, then stop sampling and assert the population associated with $\max_{1 \leq j \leq k} \bar{X}_j^{(1)}$ as the best. If I has more than one element, then proceed to the second stage.

(stage 2) Take m additional indepenent observations $X_{i,n+1}, \ldots, X_{i,n+m}$ only from each of the populations $\pi_i, i \in I$ retained at the first stage. Then assert the population associated with $\max_{i \in I} \sum_{j=1}^{n+m} X_{ij}/(n+m)$ as the best.

The design constants n, m and d in this two-stage procedure should be chosen so as to satisfy the probability requirement (1.1). Tamhane and Bechhofer(1977,1979) found easily calculable, but conservative lower bounds on the probability of correct selection(PCS) in (1.1). Lee(1990) extended the result to a general class of distributions. Recently, Bhandari and Chaudhuri(1987) and Sehr(1988) have shown that the minimun PCS is attained when the means are at the slippage configuration $\theta_{[1]} = \ldots = \theta_{[k-1]} = \theta_{[k]} - \delta^*$.

As Tamhane and Bechhofer(1979) pointed out, their result is, however, only of theoretical interest since the computation of the exact PCS at the slippage configuration is extremely difficult and costly due to the randomness of the index set I in (1.2). This paper describes a method of obtaining an easily calculable lower bound on the PCS, which is sharper than that by Tamhane and Bechhofer(1979). The method is, as it is clear from the description, easily adaptable to general location parameters selection problem.

2. A LOWER BOUND ON THE PCS

The idea of obtaining a lower bound on the PCS is to enlarging the random set I in (1.2) to other sets of indices. To describe it more precisely, the following notations are needed ;

$$Z_{ij} = X_{ij} - \theta_i,\ \bar{Z}_i^{(1)} = \bar{X}_i^{(1)} - \theta_i,\ \bar{Z}_i = \bar{X}_i - \theta_i, \tag{2.1}$$

where $\bar{X}_i = \sum_{j=1}^{n+m} X_{ij}/(n+m)$, i.e., the over-all sample mean. Further let us consider the sets of indices defined by

$$J = \{i | \bar{X}_i^{(1)} \geq \bar{X}_{[k]}^{(1)} - d\sigma/\sqrt{n}, i \neq [k]\} \tag{2.2}$$

and

$$J_o = \{i | \bar{Z}_i^{(1)} \geq \bar{Z}_{[k]}^{(1)} + \delta^* - d\sigma/\sqrt{n}, i \neq [k]\}, \tag{2.3}$$

where $\bar{X}_{[k]}^{(1)}$ is the statistic associated with $\theta_{[k]}$.
Then the key idea is the observation of the following inclusion relations from (1.2), (2.1), (2.2) and (2.3) ;

$$I \subset J \cup \{[k]\} \subset J_o \cup \{[k]\}. \tag{2.4}$$

Using this relations, we obtain in the next theorem the following lower bound $LB(\delta^*)$ on the PCS ;

$$\begin{aligned} LB(\delta^*) &= \sum_{r=0}^{k-1} \binom{k-1}{r} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [I(x, y, d, \delta^*, n, m)]^r \\ &\quad \times \Phi^{k-r-1}(x + \sqrt{n}\delta^*/\sigma - d)\, d\Phi_2(x, y|\rho), \end{aligned} \tag{2.5}$$

where $I(x, y, d, \delta^*, n, m)$ in the integrand is given by

$$\begin{aligned} &\Phi_2(x + \sqrt{n}\delta^*/\sigma + d, y + \sqrt{n+m}\,\delta^*/\sigma|\rho) \\ &-\Phi_2(x + \sqrt{n}\delta^*/\sigma - d, y + \sqrt{n+m}\,\delta^*/\sigma|\rho) \end{aligned} \tag{2.6}$$

with $\rho = (n/(n+m))^{1/2}$, Φ and $\Phi_2(\cdot, \cdot|\rho)$ denoting the cumulative distribution functions of standard univariate normal distribution and bivariate normal distribution with correlation ρ, respectively.

Theorem 1. For the two-stage selection procedure in Section 1, we have for all $\theta \in \Omega(\delta^*)$

$$P_\theta\{CS\} \geq LB(\delta^*),$$

where $LB(\delta^*)$ is given by (2.5) and (2.6).

Proof. It follows from (2.1) and (2.2) that, for all $\theta \in \Omega(\delta^*)$, i.e., $\theta_{[k]} - \theta_{[k-1]} \geq \delta^*$,

$$\begin{aligned}
&\{\bar{Z}^{(1)}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}^{(1)}_{[i]} + d\sigma/\sqrt{n}\} \cup \\
&\{\bar{Z}^{(1)}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}^{(1)}_{[i]} - d\sigma/\sqrt{n},\ \bar{Z}_{[k]} + \delta^* \geq \max_{i\in J_o} \bar{Z}_i\} \\
\subset\ &\{\bar{X}^{(1)}_{[k]} \geq \max_{i\neq k} \bar{X}^{(1)}_{[i]} + d\sigma/\sqrt{n}\} \cup \\
&\{\bar{X}^{(1)}_{[k]} \geq \max_{i\neq k} \bar{X}^{(1)}_{[i]} - d\sigma/\sqrt{n},\ \bar{X}_{[k]} \geq \max_{i\in J\cup\{[k]\}} \bar{X}_i\} \\
\subset\ &\{\bar{X}^{(1)}_{[k]} \geq \max_{i\neq k} \bar{X}^{(1)}_{[i]} + d\sigma/\sqrt{n}\} \cup \\
&\{\bar{X}^{(1)}_{[k]} \geq \max_{i\neq k} \bar{X}^{(1)}_{[i]} - d\sigma/\sqrt{n},\ \bar{X}_{[k]} \geq \max_{i\in I} \bar{X}_i\}.
\end{aligned}$$

Furthermore the event on the bottom clearly implies the correct selection, and the event on the top obviously includes the following event :

$$\begin{aligned}
&\{\bar{Z}^{(1)}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}^{(1)}_{[i]} + d\sigma/\sqrt{n}\} \cup \\
&\{\max_{i\neq k} \bar{Z}^{(1)}_{[i]} + d\sigma/\sqrt{n} > \bar{Z}^{(1)}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}^{(1)}_{[i]} - d\sigma/\sqrt{n}, \\
&\qquad \bar{Z}_{[k]} + \delta^* \geq \max_{i\in J_o} \bar{Z}_i\}. \qquad (2.7)
\end{aligned}$$

Therefore we have

$$P_\theta\{CS\} \geq LB(\delta^*)$$

for all $\theta \in \Omega(\delta^*)$.

Note that $(\sqrt{n}\bar{Z}^{(1)}_i/\sigma, \sqrt{n+m}\bar{Z}_i/\sigma)(i = 1, \ldots, k)$ have independent bivariate standard normal distributions with correlation $\rho = (n/(n+m))^{1/2}$. Hence the probability $P\{E\}$ of the event E in (2.7) can be written as $LB(\delta^*)$ in (2.5). This completes the proof.

Tamhane and Bechhofer(1979) provided the following lower bound $LB_{TB}(\delta^*)$ on the PCS ;

$$\begin{aligned}
&LB_{TB}(\delta^*) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Phi_2^{k-1}(x + \sqrt{n}\delta^*/\sigma + d, y + \sqrt{n+m}\delta^*/\sigma|\rho)d\Phi_2(x, y\,|\,\rho), \qquad (2.8)
\end{aligned}$$

with $\rho = (n/(n+m))^{1/2}$. The fact that the lower bound $LB(\delta^*)$ in (2.5) is sharper than $LB_{TB}(\delta^*)$ in (2.8) can be shown as follows ;

$$
\begin{aligned}
LB_{TB}(\delta^*) &= P\{\bar{Z}^{(1)}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}^{(1)}_{[i]} - d\sigma/\sqrt{n}, \bar{Z}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}_{[i]}\} \\
&= P\{\bar{Z}^{(1)}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}^{(1)}_{[i]} + d\sigma/\sqrt{n}, \bar{Z}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}_{[i]}\} \\
&\quad + P\{\max_{i\neq k} \bar{Z}^{(1)}_{[i]} + d\sigma/\sqrt{n} > \bar{Z}^{(1)}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}^{(1)}_{[i]} - d\sigma/\sqrt{n}, \\
&\qquad \bar{Z}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}_{[i]}\} \\
&\leq P\{\bar{Z}^{(1)}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}^{(1)}_{[i]} + d\sigma/\sqrt{n}\} \\
&\quad + P\{\max_{i\neq k} \bar{Z}^{(1)}_{[i]} + d\sigma/\sqrt{n} > \bar{Z}^{(1)}_{[k]} + \delta^* \geq \max_{i\neq k} \bar{Z}^{(1)}_{[i]} - d\sigma/\sqrt{n}, \\
&\qquad \bar{Z}_{[k]} + \delta^* \geq \max_{i\in J_o} \bar{Z}_i\} \\
&= LB(\delta^*).
\end{aligned}
$$

3. DESIGN CONSTANTS AND RESULTS OF PERFORMANCE COMPARISON

For the implementation of the two-stage selection procedure, Tamhane and Bechhofer(1977) proposed the following R-minimax criterion to determine the design constants n, m and d : subject to the probability requirement (1.1),

$$\text{minimize} \sup_{\theta\in\Omega(\delta^*)} E_\theta\{\text{TSS}|R\}, \tag{3.1}$$

where TSS denotes the total sample size. They proved that the maximum of $E_\theta\{\text{TSS}|R\}$ for $\theta \in \Omega(\delta^*)$ is given as follows :

$$
\begin{aligned}
kn + m\Big[&\int_{-\infty}^{\infty} \{\Phi^{k-1}(x + \sqrt{n}\delta^*/\sigma + d) - \Phi^{k-1}(x + \sqrt{n}\delta^*/\sigma - d)\} d\Phi(x) \\
&+ (k-1) \int_{-\infty}^{\infty} \{\Phi^{k-2}(x+d)\Phi(x - \sqrt{n}\delta^*/\sigma + d) - \Phi^{k-2}(x-d) \\
&\qquad \times \Phi(x - \sqrt{n}\delta^*/\sigma - d)\} d\Phi(x) \Big],
\end{aligned} \tag{3.1}
$$

which occurs when $\theta_{[1]} = \cdots = \theta_{[k-1]} = \theta_{[k]} - \delta^*$.
Then their design constants were given as the solutions of the optimization problem :

$$\text{minimize (3.1) subject to } LB_{TB}(\delta^*) \geq p^*.$$

Following Tamhane and Bechhofer(1977), we solve the following optimization problem :

$$\text{minimize (3.1) subject to } LB(\delta^*) \geq p^*, \tag{3.2}$$

where $LB(\delta^*)$ is given by (2.5) and (2.6). It should be noted that the expressions (2.5), (2.6) and (3.1) depend on n, m and d only through

$$c_1 = \sqrt{n}\delta^*/\sigma,\ c_2 = \sqrt{m}\delta^*/\sigma,\ d. \tag{3.3}$$

Table 1 gives the design constants c_1, c_2 and d, i.e., the solutions of the optimization problem (3.2) for $k = 2(1)10, 15, 20, 25$ and $p^* = .99, .95, .90$. The optimization problem (3.2) was solved numerically by the SUMT algorithm of Fiacco and McCormick(1968). In searching for optimal solutions, the design constants of Tamhane and Bechhofer(1979) were used as initial values. The computation of the integrals involved was done by Gauss-Hermite quadrature for $k \leq 7$ and by Monte Carlo method for $k > 7$.

Table 1. Design Constants by Criterion (3.2)

p^*	k	c_1	c_2	d	k	c_1	c_2	d
0.99		2.253	2.748	1.288		3.055	2.881	1.186
0.95	2	1.619	1.962	1.071	8	2.306	2.595	1.237
0.90		1.260	1.551	0.985		1.987	2.250	1.260
0.99		2.783	2.398	1.073		3.063	2.968	1.215
0.95	3	2.107	1.696	1.045	9	2.229	2.667	1.243
0.90		1.661	1.517	1.158		2.021	2.362	1.222
0.99		2.857	2.496	1.166		3.069	3.043	1.240
0.95	4	2.113	2.040	1.370	10	2.350	2.694	1.310
0.90		1.793	1.741	1.150		2.031	2.443	1.258
0.99		2.941	2.630	1.139		3.142	3.282	1.261
0.95	5	2.190	2.231	1.228	15	2.474	2.869	1.303
0.90		1.844	2.033	1.170		2.137	2.751	1.238
0.99		2.973	2.789	1.127		3.157	3.375	1.392
0.95	6	2.243	2.355	1.214	20	2.498	3.109	1.357
0.90		1.880	2.069	1.328		2.148	2.901	1.361
0.99		3.016	2.831	1.176		3.226	3.547	1.313
0.95	7	2.279	2.481	1.232	25	2.531	3.302	1.361
0.90		1.933	2.176	1.282		2.235	3.071	1.290

As by-products of this optimization problem, we can get the values of $\max_{\theta \in \Omega(\delta^*)} E_\theta\{TSS\}$ for the design constants so obtained. Table 2 compares these values according to the lower bounds $LB(\delta^*)$ and $LB_{TB}(\delta^*)$ used to satisfy the probability requirement (1.1). It should be noted that the values in Table 2 are the values of $(\delta^*/\sigma)^2 \max_{\theta \in \Omega(\delta^*)} E_\theta\{TSS\}$ due to the reparametrization (3.3).

Table 2. The Maximum of the Expected Total Sample Size

p^*	k	$LB(\delta^*)$	$LB_{TB}(\delta^*)$	k	$LB(\delta^*)$	$LB_{TB}(\delta^*)$
0.99		13.80	13.82		81.55	83.49
0.95	2	7.71	7.78	8	54.86	56.09
0.90		4.94	5.03		43.20	44.47
0.99		25.05	26.08		93.32	95.55
0.95	3	15.42	16.20	9	63.21	64.57
0.90		10.84	12.32		49.92	51.39
0.99		35.87	36.80		104.60	107.09
0.95	4	23.08	23.75	10	71.11	73.00
0.90		16.93	18.19		56.76	58.42
0.99		47.15	48.19		163.90	168.64
0.95	5	30.71	31.54	15	114.30	116.89
0.90		23.63	24.51		92.37	94.38
0.99		58.49	59.99		223.61	226.82
0.95	6	38.66	39.70	20	158.64	162.01
0.90		30.13	31.14		128.80	132.78
0.99		70.06	71.50		286.80	293.80
0.95	7	46.67	47.80	25	203.75	206.80
0.90		36.52	37.86		166.32	170.25

It can be observed from Table 2 that the savings in $\sup_{\theta \in \Omega(\delta^*)} E_\theta\{TSS\}$ by $\mathrm{LB}(\delta^*)$ are not much. It should be, however, noted that the savings are in the units of $(\delta^*/\sigma)^2$. This means that when δ^*/σ is small the savings can be substantial.

Finally, it should be remarked that the method of obtaining the lower bound $\mathrm{LB}(\delta^*)$ in (2.5) only requires the location parameter setting. Thus the idea can be adopted for similar problems. In particular, the same method can be applied to the two-stage procedure of Gupta and Kim(1984) for the normal means problem with unknown common variance. Of course the difficulty of computing the resulting lower bound remains and it needs further research to get the design constants for the resulting procedures.

REFERENCES

(1) Alam, K. (1970). A two-sample procedure for selecting the population with the largest mean from k normal populations. *Annals of the Institute of Statistical Mathematics.*, 22, 127-136.

(2) Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics.*, 25, 16-39.

(3) Bhandari, S. K. and Chaudhuri, A. R. (1987). On two conjectures about two-stage selection problem. Unpublished manuscript.

(4) Fiacco, A. V. and McCormick, G. P. (1968). *Nonlinear Sequential Unconstrained Minimization Techniques*, John Wiley ans Sons, Inc., New York.

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

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

(7) Lee, S. H. (1990). A two-stage elimination type selection procedure for stochastically increasing distributions : with an application to scale parameters problem. *Journal of the Korean Statistical Society.,* Vol.19, 1, 24-44.

(8) Sehr, J. (1988). On a conjecture concerning the least favorable configuration on a two-stage selection procedure. *Communications in Statistics - Theory and Methods.,* 17(10), 3221-3233.

(9) Tamhane, A. C. and Bechhofer, R. E. (1977). A two-stage minimax procedure with screening for selecting the largest normal mean. *Communications in Statistics - Theory and Methods.,* A6, 1003-1033.

(10) Tamhane, A. C. and Bechhofer, R. E. (1979). A two-stage minimax procedure with screening for selecting the largest normal mean (II): an improved PCS lower bound and associated tables. *Communications in Statistics - Theory and Methods.,* A8, 337-358.