Simultaneous Selection and Estimation of the Largest Normal Mean by Confidence Statement Approach

Yoshikazu Takada
Kumamoto University, Kumamoto, JAPAN
e-mail: takada@kumamoto-u.ac.jp

Abstract

Ranking and selection studies have focused primary on selecting the best population. After selecting a population as the best one, it would be desirable to estimate its mean with a confidence interval. We (2011) proposed a procedure of simultaneously selecting the best population and estimating its mean by the indifference-zone approach. However, the procedure tells us nothing when the parameter lies in the indifference zone. In order to overcome its drawback, this presentation will propose a procedure based on the confidence statement approach. An artificial example will be used to illustrate the procedure.

Keywords: ranking and selection, two-stage procedure, indifference-zone approach, confidence statement approach.

1. Introduction

Let Π_i be a normal population with unknown mean μ_i and unknown common variance σ^2, $i = 1, \ldots, k(\geq 2)$. Let $\mu_{[1]} \leq \cdots \leq \mu_{[k]}$ denote the ordered μ_i-values. Bechhofer et al. (1954) considered the problem of selecting the population with $\mu_{[k]}$ under the indifference zone approach. Tong (1970) considered the problem of constructing a fixed-width confidence interval of $\mu_{[k]}$. After selection, it would be desirable to get an estimate of $\mu_{[k]}$.

Let $X_{i1}, \cdots, X_{in}$ be n observations from Π_i and let $\bar{X}_{i(n)} = \sum_{j=1}^{n} X_{ij}/n$ be the sample mean, $i = 1, \ldots, k$. We select the population that yields $\bar{X}_{[k]} = \max(\bar{X}_{1(n)}, \ldots, \bar{X}_{k(n)})$ as the one associated with $\mu_{[k]}$, and estimate $\mu_{[k]}$ by an interval $I_n = (\bar{X}_{[k]} - d, \bar{X}_{[k]} + d)$ with a length $2d(> 0)$. Applying the indifference zone approach, we (2011) discussed the determination of the sample size n such that for specified $\delta^*(> 0)$ and $P^*(1/k < P^* < 1)$,

$$P(CS, \mu_{[k]} \in I_n) \geq P^* \quad \text{whenever } \mu_{[k]} - \mu_{[k-1]} \geq \delta^*, \tag{1}$$

where CS stands for the correct selection, which is said to be made if the selected population has the mean $\mu_{[k]}$. However, one may be concerned with the performance of the procedure when $\mu_{[k]} - \mu_{[k-1]} < \delta^*$. In order to meet such a requirement, we apply the confidence statement approach. That is, we want to determine the sample size n such that for specified $\delta^*(> 0)$ and $P^*(1/k < P^* < 1)$,

$$P(\mu_{[k]} < \mu_S + \delta^*, \mu_S \in I_n) \geq P^*, \tag{2}$$

where μ_S is the mean of the selected population. It is easy to see that the condition (2) implies the condition (1). However, contrary to the usual selection problem (Fabian, 1962), the opposite is not true. In this paper we discuss the determination of the sample size satisfying the condition (2).

In Section 2 we shall propose a two-stage procedure to meet the probability requirement (2). An artificial example is given to see the implementation of the procedure. In Section 3 we shall consider the case that the variances are unknown and unequal. The example in Section 2 is reconsidered.

2. Two-stage procedure

Letting $U_i = \left(\max_{j\neq i} \bar{X}_{j(n)} - \bar{X}_{i(n)} + \delta^*\right)^+, i = 1, \ldots, k$, where $a^+ = \max(a, 0)$, then

$$U_S = \left(\max_{j\neq S} \bar{X}_{j(n)} - \bar{X}_{[k]} + \delta^*\right)^+ \leq \delta^*.$$

Hence

$$\begin{aligned} P(\mu_{[k]} < \mu_S + \delta^*, \mu_S \in I_n) &\geq P\left(\mu_{[k]} < \mu_S + U_S, \mu_S \in I_n\right) \\ &\geq P\left(\mu_{[k]} < \mu_i + U_i, |\mu_i - \bar{X}_{i(n)}| < d, i = 1, \ldots, k\right) \\ &= P\left(\mu_{[k]} < \mu_{[i]} + U_{(i)}, |\mu_{[i]} - \bar{X}_{(i)n}| < d, i = 1, \ldots, k-1,\right. \\ &\quad \left.|\mu_{[k]} - \bar{X}_{(k)n}| < d\right) \end{aligned}$$

where (i) denotes the population number corresponding $\mu_{[i]}, i = 1, \ldots, k$. Then it follows that

$$\begin{aligned} P(\mu_{[k]} < \mu_S + \delta^*, \mu_S \in I_n) &\geq P\left(\mu_{[k]} < \mu_{[i]} + \bar{X}_{(k)n} - \bar{X}_{(i)n} + \delta^*, |\mu_{[i]} - \bar{X}_{(i)n}| < d,\right. \\ &\quad \left. i = 1, \ldots, k-1, |\mu_{[k]} - \bar{X}_{(k)n}| < d\right) \\ &= G\left(\frac{\sqrt{n}\delta^*}{\sigma}\right), \end{aligned}$$

where

$$G(\tau) = \int_{|x|<\tau\eta} F^{k-1}(\tau, x) d\Phi(x)$$

with $\eta = d/\delta^*$, $F(\tau, x) = \Phi\left(\max(\min(x + \tau, \tau\eta), -\tau\eta)\right) - \Phi(-\tau\eta)$, and Φ is the cumulative distribution function of the standard normal distribution. We determine τ^* such that $G(\tau^*) = P^*$. Then if the sample size n is chosen such that $n \geq \tau^{*2}\sigma^2/\delta^{*2} = n^*$ (say), the probability requirement (2) is satisfied.

If σ were known, one would take $[n^*] + 1$ observations from each population and implement the corresponding selection and estimation procedure, where $[n^*]$ denotes the largest integer less than n^*. Unfortunately, σ^2 is unknown, and hence the procedure is not available. We propose a two-stage procedure to meet the probability requirement (2).

Take an initial sample $X_{i1}, \ldots, X_{im}$ of size $m(\geq 2)$ from Π_i and calculate $V_i^2 = \frac{1}{m-1}\sum_{j=1}^m (X_{ij} - \bar{X}_{i(m)})^2$, $i = 1, \ldots, k$. Let $\hat{\sigma}_\nu^2 = \frac{1}{k}\sum_{i=1}^k V_i^2$, where $\nu = k(m-1)$. Let $\tau_\nu(> 0)$ be such a constant that

$$\int_0^\infty G(\tau_\nu z) g_\nu(z) dz = P^*, \tag{3}$$

where $g_\nu(z)$ is the density function of $\sqrt{\chi_\nu^2/\nu}$ with a chi-squared random variable χ_ν^2 with ν degrees of freedom. Then we define the total sample size N from each population by

$$N = \max\left\{m, \left[\frac{\tau_\nu^2 \hat{\sigma}_\nu^2}{\delta^{*2}}\right] + 1\right\}. \tag{4}$$

If $N > m$, take $N - m$ additional observations $X_{im+1}, \ldots, X_{iN}$ from Π_i, $i = 1, \ldots, k$. Calculate $\bar{X}_{i(N)} = \frac{1}{N}\sum_{j=1}^{N} X_{ij}$, $i = 1, \ldots, k$. Then the selection and estimation procedure is implemented by using $\bar{X}_{1(N)}, \ldots, \bar{X}_{k(N)}$.

Theorem 1 The two-stage procedure satisfies

$$P(\mu_{[k]} < \mu_S + \delta^*, \mu_S \in I_N) \geq P^*.$$

An artificial example is given to see how the two-stage procedure is implemented. We choose $k = 5$, $\delta^* = 4.0$, $d = 2.0$, $P^* = 0.9$, and $m = 10$. Then we find $\tau_\nu = 4.774$ in (3). Suppose that the sample variances based on the first 10 observations are given by

$$22.3\,(\Pi_1), \quad 30.4\,(\Pi_2), \quad 25.2\,(\Pi_3), \quad 21.4\,(\Pi_4), \quad 27.8\,(\Pi_5).$$

We have

$$\hat{\sigma}_\nu^2 = \frac{1}{5}\,(22.3 + 30.4 + 25.2 + 21.4 + 27.8) = 25.42.$$

Then from (4) the total sample size becomes

$$\begin{aligned} N &= \max\left\{10, \left[\frac{4.774^2 \times 25.42}{4.0^2}\right] + 1\right\} \\ &= \max\{10, 37\} \\ &= 37. \end{aligned}$$

Hence we need additional 27 observations from each population. Suppose that the following cumulative sample means after the additional sampling are obtained

$$13.2\,(\Pi_1), \quad 10.4\,(\Pi_2), \quad 18.2\,(\Pi_3), \quad 20.1\,(\Pi_4), \quad 16.2\,(\Pi_5).$$

Then we select Π_4 and at confidence level $P^* = 0.9$, we can assert that the difference between μ_4 and the largest mean is less than 4.0 and μ_4 is contained in

$$I_{37} = (20.1 - 2,\ 20.1 + 2) = (18.1,\ 22.1).$$

3. Unequal variances

In this section we assume that the variances of each population are unknown and unequal. We apply the heteroscedastic method due to Dudewicz and Dalal (1975) to the problem. Take an initial sample $X_{i1}, \ldots, X_{im}$ of size m from Π_i and calculate $V_i^2, i = 1, \ldots, k$. Let $\tilde{\tau}_\nu$ be such a constant that

$$\int_{|y|<\tilde{\tau}_\nu \eta} \tilde{F}_\nu^{k-1}(\tilde{\tau}_\nu, x) dT_\nu(y) = P^* \tag{5}$$

with $\nu = m-1$, in which T_ν is the cumulative distribution function of Student's t-distribution with ν degrees of freedom and $\tilde{F}(\tau, x) = T_\nu\,(\max(\min(x + \tau, \tau\eta), -\tau\eta)) - T_\nu(-\tau\eta)$. We define the total sample size N_i from Π_i by

$$N_i = \max\left\{m + 1, \left[\frac{\tilde{\tau}_\nu^2 V_i^2}{\delta^{*2}}\right] + 1\right\}, \quad i = 1, \ldots, k. \tag{6}$$

Take $N_i - m$ additional observations $X_{im+1}, \ldots, X_{iN_i}$ from Π_i, $i = 1, \ldots, k$ and define

$$\tilde{X}_{i(N_i)} = \sum_{j=1}^{N_i} a_{ij} X_{ij}, \quad i = 1, \ldots, k.$$

The a_{ij}'s are to be chosen so that

$$\sum_{j=1}^{N_i} a_{ij} = 1, \quad a_{i1} = \cdots = a_{im}, \quad V_i^2 \sum_{j=1}^{N_i} a_{ij}^2 = \frac{\delta^{*2}}{\tilde{\tau}_\nu^2}, \quad i = 1, \ldots, k.$$

Denoting the largest among $\tilde{X}_{1(N_1)}, \ldots, \tilde{X}_{k(N_k)}$ by $\tilde{X}_{[k]}$, we select the population that yields $\tilde{X}_{[k]}$ and construct an interval $\tilde{I} = (\tilde{X}_{[k]} - d, \tilde{X}_{[k]} + d)$.

Theorem 2 The two-stage procedure satisfies

$$P(\mu_{[k]} \leq \mu_S + \delta^*, \mu_S \in \tilde{I}) \geq P^*,$$

where μ_S is the mean of the selected population.

Dudewicz, Ramberg and Chen (1975) recommended that

$$\tilde{X}_{i(N_i)} = c_i \bar{X}_{i(m)} + (1 - c_i) \bar{Y}_{i(N_i - m)}, \quad i = 1, \ldots, k,$$

where $\bar{X}_{i(m)}$ and $\bar{Y}_{i(N_i-m)}$ are the means of the first and second samples, and

$$c_i = \frac{m}{N_i} \left(1 + \sqrt{1 - \frac{N_i}{m} \left(1 - \frac{(N_i - m)z}{V_i^2} \right)} \right), \quad i = 1, \ldots, k \tag{7}$$

with $z = \delta^{*2}/\tilde{\tau}_\nu^2$.

We reconsider the example given in Section 2 under the condition that the variances are unequal. We find $\tilde{\tau}_\nu = 5.593$ in (5). Then from (6) the total sample size from Π_1 becomes

$$\begin{aligned} N_1 &= \max \left\{ 10 + 1, \left[\frac{5.593^2 \times 22.3}{4.0^2} \right] + 1 \right\} \\ &= \max \{11, 44\} \\ &= 44. \end{aligned}$$

Hence we need additional 34 observations from Π_1. Likewise, we have $N_2 = 60$, $N_3 = 50$, $N_4 = 42$, and $N_5 = 55$. Suppose that the means of the first and second sample from each population are given in the third and fourth column in Table 1. Then c_i's in (7) and $\tilde{X}_{i(N_i)}$'s are obtained in the fifth and last column in Table 1. From the last column in Table 1, we select Π_4 and at confidence level $P^* = 0.9$, we can assert that the difference between μ_4 and the largest mean is less than 4.0 and μ_4 is contained in

$$\tilde{I} = (24.6 - 2,\ 24.6 + 2) = (22.6,\ 26.6).$$

Table 1: Calculation of the two-stage procedure

	Π_1	Π_2	Π_3	Π_4	Π_5
N_i	44	60	50	42	55
$\bar{X}_{i(m)}$	18.7	16.2	20.3	22.4	21.0
$\bar{Y}_{i(N_i-m)}$	19.7	17.2	22.3	25.4	19.8
c_i	0.267	0.203	0.249	0.264	0.224
$\tilde{X}_{i(N_i)}$	19.4	17.0	21.8	24.6	20.1

References

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.

Dudewicz, E.L. and Dalal, S.R. (1975) " Allocation of observations in ranking and selection with unknown variances," *Sankhyā*, B37, 28–78.

Dudewicz, E.L., Ramberg, J.S., and Chen, H.J. (1975) "New tables for multiple comparisons with a control (unknown variances)," *Biometrsche Zeitschrift*, 17, 13–26.

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

Takada, Y. (2011) "Simultaneous selection and estimation of the largest normal mean," *Proc. 58th Session of the ISI, Dublin.*

Tong, Y.L. (1970) "Multi-stage interval estimation of the largest mean of k normal populations," *J. Roy. Statist. Soc.*, B32, 272-277.