Simultaneous Selection and Estimation of the Largest Normal Mean

Takada, Yoshikazu
Kumamoto University, Department of Mathematics and Engineering
2-39-1 Kurokami
Kumamoto (860-8555), Japan
E-mail: takada@kumamoto-u.ac.jp

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 $\mu_{[k]}$. They proposed a two-stage procedure to meet the probability requirement for each problem. See also Mukhopadhyay and Solanky (1994) and the references therein. After selection, it would be desirable to get an estimate of $\mu_{[k]}$. In this paper we give a procedure for selecting the population and simultaneously estimating its mean with a fixed-width confidence interval under the indifference zone approach.

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 $\bar{X}_{[k]} = \max(\bar{X}_{1(n)}, \cdots, \bar{X}_{k(n)})$ as the one associated with $\mu_{[k]}$, and estimate $\mu_{[k]}$ by a confidence interval $I_n = (\bar{X}_{[k]} - d, \bar{X}_{[k]} + d)$ with a length $2d(> 0)$. Then for specified $\delta^*(> 0)$ and $P^*(1/k < P^* < 1)$, we want to determine the sample size n such that

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

where CS stands for the correct selection, which is said to be made if the selected population has the mean $\mu_{[k]}$.

In Section 1 we shall propose a two-stage procedure to meet the probability requirement (1). An artificial example is given to see the implementation of the procedure in Section 2. In Section 3 the asymptotic efficiency of the sample size is examined. We also conduct simulations to see moderate sample performances of the two-stage procedure.

1. Two-stage procedure

Without loss of generality, let $\mu_{[k]} = \mu_k$. Then

$$\begin{aligned} P(CS, \mu_{[k]} \in I_n) &= P\left(\bar{X}_{k(n)} > \bar{X}_{i(n)}, i = 1, \ldots, k-1, \left|\bar{X}_{k(n)} - \mu_k\right| < d\right) \\ &= P\Big(n^{1/2}(\bar{X}_{k(n)} - \mu_k)/\sigma + n^{1/2}(\mu_k - \mu_i)/\sigma > n^{1/2}(\bar{X}_{i(n)} - \mu_i)/\sigma, \\ &\qquad i = 1, \ldots, k-1, n^{1/2}\left|\bar{X}_{k(n)} - \mu_k\right|/\sigma < n^{1/2}d/\sigma\Big) \\ &= \int_{|y|<n^{1/2}d/\sigma} \prod_{i=1}^{k-1} \Phi\left(y + n^{1/2}(\mu_k - \mu_i)/\sigma\right) d\Phi(y), \end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Since $\mu_k - \mu_i \geq \delta^*$, it follows that

$$\begin{aligned} P(CS, \mu_{[k]} \in I_n) &\geq \int_{|y|<n^{1/2}d/\sigma} \Phi^{k-1}\left(y + n^{1/2}\delta^*/\sigma\right) d\Phi(y) \\ &= \int_{|y|<\tau\eta} \Phi^{k-1}(y + \tau) d\Phi(y), \end{aligned} \tag{2}$$

where $\eta = d/\delta^*$ and $\tau = n^{1/2}\delta^*/\sigma$. We determine τ^* such that

$$\int_{|y|<\tau^*\eta} \Phi^{k-1}(y + \tau^*) d\Phi(y) = P^*. \tag{3}$$

Then it follows from (2) that if the sample size n is chosen such that

$$n \geq \tau^{*2}\sigma^2/\delta^{*2} = n^* \quad \text{(say)},$$

the probability requirement (1) is satisfied. The n^* is called the optimal fixed sample size.

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 optimal fixed sample size is not available. We propose a two-stage procedure to meet the probability requirement (1).

Take an initial sample $X_{i1}, \cdots, X_{im}$ of size $m(\geq 2)$ from Π_i, $i = 1, \ldots, k$ and calculate

$$\hat{\sigma}_\nu^2 = \frac{1}{\nu}\sum_{i=1}^{k}\sum_{j=1}^{m}\left(X_{ij} - \bar{X}_{i(m)}\right)^2$$

where $\nu = k(m-1)$. Let $\tau_\nu(>0)$ be such a constant that

$$(4) \qquad \int_0^\infty \left\{\int_{|y|<\tau_\nu \eta z} \Phi^{k-1}\left(y + \tau_\nu z\right) d\Phi(y)\right\} g_\nu(z)dz = P^*,$$

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. We define the total sample size N by

$$(5) \qquad N = \max\left\{m, \left[\tau_\nu^2\hat{\sigma}_\nu^2/\delta^{*2}\right] + 1\right\}.$$

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

Theorem 1 The two-stage procedure satisfies

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

2. Example

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 $\tau_\nu = 3.494$ in (4). 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 that

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

Then from (5) the total sample size becomes

$$\begin{aligned} N &= \max\left\{10, \left[\frac{3.494^2 \times 25.42}{4.0^2}\right] + 1\right\} \\ &= \max\{10, 20\} \\ &= 20. \end{aligned}$$

Hence we need additional 10 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 at confidence level $P^* = 0.9$, we can assert that Π_4 has the largest mean, which is contained in

$$I_{20} = (20.1 - 2,\, 20.1 + 2) = (18.1,\, 22.1).$$

3. Asymptotic efficiency

We shall investigate an asymptotic property of the sample size N in (5) as the first sample size m is chosen such that $m \to \infty$ as $d \to 0$ and $\delta^* \to 0$ with $\eta = d/\delta^*$ fixed. The following lemma gives the asymptotic expansion of τ_ν in (4) as ν is large.

Lemma 1

$$\tau_\nu^2 = \tau^{*2} + \frac{b}{\nu} + o\left(\frac{1}{\nu}\right) \quad \text{as } \nu \to \infty,$$

where

$$b = -\frac{\tau^{*4} G''(\tau^{*2})}{G'(\tau^{*2})}$$

with

$$G(z) = \int_{|y|<\eta\sqrt{z}} \Phi^{k-1}(y + \sqrt{z})\, d\Phi(y)$$

and τ^* in (3).

It is difficult to analytically show that the value of b in Lemma 1 is positive, but the numerical analysis supports that its value is positive. In the followings, we suppose that the value of b is positive.

Theorem 2 If the first sample size m is chosen such that

$$m \to \infty \quad \text{and} \quad m\delta^{*2} \to 0$$

as $d \to 0$ and $\delta^* \to 0$ with $\eta = d/\delta^*$ fixed, then

$$\frac{E(N)}{n^*} \to 1,$$

but

$$E(N - n^*) \to \infty.$$

Theorem 2 tells us that the two-stage procedure is asymptotically first-order efficient, but is not asymptotically second-order efficient in the sense of Ghosh and Mukhopadhyay (1981). However, it is possible to make the two-stage procedure asymptotically second-order efficient if we can assume a known lower bound $\sigma_L^2 (> 0)$ for unknown variance σ^2. See Mukhopadhyay and Duggan (1999).

Theorem 3 Suppose that $\sigma^2 > \sigma_L^2$. If the first sample size m is chosen such that

$$m\delta^{*2} \to \tau^{*2}\sigma_L^2 \tag{6}$$

as $d \to 0$ and $\delta^* \to 0$ with $\eta = d/\delta^*$ fixed, then

$$E(N - n^*) \to \frac{b\sigma^2}{k\tau^{*2}\sigma_L^2} + \frac{1}{2}. \tag{7}$$

We conducted simulations to see moderate sample size performances of the two-stage procedure. We chose $k = 5$, $P^* = 0.9$, and $\sigma_L^2 = 1.0$ as the lower bound for unknown variance σ^2. We also chose $\eta = 0.5, 1.0, 2.0$, $m = 10(10)40$, and $\delta^* = \tau^*\sigma_L/\sqrt{m}$ in order that the first sample size m meets the

Table 1: $k = 5$, $P^* = 0.9$, $\sigma_L^2 = 1.0$, $\sigma^2 = 1.5^2$

	$\eta = 0.5$		$\eta = 1.0$		$\eta = 2.0$	
m	CP	$E(N - n^*)$	CP	$E(N - n^*)$	CP	$E(N - n^*)$
10	0.907	1.665	0,906	1.629	0.905	1.430
20	0.901	1.623	0.902	1.578	0.899	1.405
30	0.903	1.604	0.903	1.510	0.900	1.405
40	0.902	1.473	0.904	1.409	0.903	1.402
		1.483		1.434		1.335

requirement (6). The simulation results are based on $10,000$ replications under $\mu_1 = \cdots = \mu_4 = \mu_5 - \delta^*$ and $\sigma^2 = 1.5^2$. We evaluated $E(N - n^*)$ and $CP = P(CS, \mu_5 \in I_N)$ in Table 1. The values in the last row of the table provide the asymptotic ones obtained from the right-hand side of (7). One will observe that the value of $E(N - n^*)$ are well approximated by these asymptotic ones.

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.

Ghosh, M. and Mukhopadhyay, N. (1981). Consistency and asymptotic efficiency of two-stage and sequential estimation procedures. *Sankhya, A*, **43**, 220–227.

Mukhopadhyay, N. and Duggan, W.T. (1999). On a two-stage procedure having second-order properties with applications. *Ann. Inst. Statist.*, **51**, 621-636.

Mukhopadhyay, N. and Solanky, T.K.S. (1994). *Multistage Selection and Ranking procedures: Second-Order Asymptotics*, Marcel Dekker, New York.

Tong, Y.L. (1970). Multi-stage interval estimation of the largest mean of k normal populations. *J. Roy. Statist. Soc., B*, **32**, 272-277.