Selecting the Best Normal Population Better Than a Standard Under the Unequal Variance Case

Yoshikazu Takada

Department of Applied Mathematics, Kumamoto University,
2-39-1 Kurokami,
Kumamoto, 860-8555, Japan
takada@kumamoto-u.ac.jp

Abstract. We consider the problem of selecting the best normal population that is better than a standard when the variances are unequal. A single-stage and two-stage selection procedures are proposed by using the methods of Dudewicz and Dalal (1975), Rinott (1978), and Lam (1988). A comparison is made between these selection procedures.

Keywords. Correct selection, indifference zone approach, normal means, unequal variances.

1 Introduction

We consider the problem of selecting the normal population provided that its associated mean is greater than a standard μ_0. The constant μ_0 is called a standard when μ_0 is known. Let Π_i be a normal population with unknown mean μ_i and known or unknown variance σ_i^2, $i = 1, \cdots, k (\geq 2)$. Let

$$\mu_{[1]} \leq \cdots \leq \mu_{[k]}$$

denote the ordered μ_i-values. The goal is to select the population associated with $\mu_{[k]}$ if $\mu_{[k]} > \mu_0$, or to select no population if $\mu_{[k]} \leq \mu_0$. Bechhofer and Turnbull (1978) used an indifference-zone approach to study the problem. Let $\delta_0^*, \delta_1^*, \delta_2^*, P_0^*$ and P_1^* be five constants determined by an experimenter, in which $0 < \delta_1^*, \delta_2^* < \infty, -\delta_1^* < \delta_0^* < \infty, 2^{-k} < P_0^* < 1$, and $(1 - 2^{-k})/k < P_1^* < 1$. We require

$$P(\Pi_0) \geq P_0^* \quad \text{whenever } \mu_{[k]} \leq \mu_0 - \delta_0^* \tag{1}$$

and

$$P(\Pi_{[k]}) \geq P_1^* \quad \text{whenever } \mu_{[k]} \geq \mu_0 + \delta_1^* \text{ and } \mu_{[k]} \geq \mu_{[k-1]} + \delta_2^*, \tag{2}$$

where $\Pi_0 (\Pi_{[k]})$ denotes the event of selecting no population (the population associated with $\mu_{[k]}$).

Let $X_{i1}, \cdots, X_{in_i}$ be n_i observations from Π_i and let $\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$ be the sample mean $(i = 1, \cdots, k)$. The selection procedure is as follows. Denoting the largest of the sample means by $\bar{X}_{[k]}$, we select the population that yields $\bar{X}_{[k]}$ as the one associated with $\mu_{[k]}$ if $\bar{X}_{[k]} > \mu_0 + c$, otherwise, we select no population. The problem is to determine the sample sizes n_i's and the constant c so as to satisfy (1) and (2).

When the variances are known and equal, Bechhofer and Turnbull (1978) proposed $n_1 = \cdots = n_k (= n)$ and

$$n = \left[\frac{g^2 \sigma^2}{\delta_2^{*2}} \right] + 1, \qquad c = \frac{h}{g} \delta_2^*, \tag{3}$$

where σ^2 is the common variance and $[x]$ is the largest integer less than x. The constants h and g are the solutions of the simultaneous equations

$$\Phi^k \left(h + \delta_0^* g / \delta_2^* \right) = P_0^* \tag{4}$$

and

$$\int_{h - \delta_1^* g / \delta_2^*}^{\infty} \Phi^{k-1}(y + g) d\Phi(y) = P_1^*. \tag{5}$$

Here Φ is the standard normal distribution function. They showed that the probability requirements (1) and (2) are satisfied. The values of h and g are available in their paper when $\delta_0^* = 0$ and $\delta_1^* = \delta_2^*$.

When the variances are known and unequal, the form (3) suggests us to consider the following sample sizes and the constant c

$$n_i = \left[\frac{g^2 \sigma_i^2}{\delta_2^{*2}} \right] + 1, \quad i = 1, \cdots, k, \qquad c = \frac{h}{g} \delta_2^*. \tag{6}$$

We consider the problem of selecting h and g so as to meet the probability requirements (1) and (2). In particular, we explore if the constants h and g, which are determined by the simultaneous equations (4) and (5), satisfy the probability requirements.

When the variances are unknown and unequal, we propose two-stage selection procedures by using the methods in Dudewicz and Dalal (1975), Rinott (1978), and Lam (1988). See Wilcox (1984) for other two-stage selection procedure. Comparisons are made between these procedures.

Similar problems are discussed in Mukhophadhyay (1979) and Takada (2007) for selecting the best normal population.

2 Known variances

In this section we suppose that the values of σ_i^2's are known. Throughout this section, we assume that the constants h and g in (6) satisfy (4). We propose two other equations than (5) to determine the constants h and g along with (4):

$$\Phi^{k-1}\left(\frac{g}{\sqrt{2}}\right) \Phi\left(\frac{g\delta_1^*}{\delta_2^*} - h\right) = P_1^* \tag{7}$$

and

$$\frac{1}{2}\Phi^{k-1}(g) + \int_{-a}^{0} \Phi^{k-1}(y+g) d\Phi(y) = P_1^*, \tag{8}$$

where $a = \min\{g, g\delta_1/\delta_2^* - h\}$. Let $h_0(>0)$ be a constant such that $\Phi^k(h_0) = P_0^*$. Then from (4)

$$h = h_0 - \frac{\delta_0^*}{\delta_2^*} g, \tag{9}$$

and hence

$$\frac{g\delta_1^*}{\delta_2^*} - h = \frac{g(\delta_0^* + \delta_1^*)}{\delta_2^*} - h_0.$$

We define three functions of x;

$$\begin{aligned} F_1(x) &= \int_{-f(x)}^{\infty} \Phi^{k-1}(y+x) d\Phi(y), \\ F_2(x) &= \Phi^{k-1}\left(x/\sqrt{2}\right) \Phi(f(x)), \\ F_3(x) &= \frac{1}{2}\Phi^{k-1}(x) + \int_{-b(x)}^{0} \Phi^{k-1}(y+x) d\Phi(y), \end{aligned}$$

where $f(x) = (\delta_0^* + \delta_1^*)x/\delta_2^* - h_0$ and $b(x) = \min(x, f(x))$. Then it follows from (5) ((7), (8)) that the solutions h and g of the simultaneous equations (4) and (5) ((4) and (7), (4) and (8)) are such that $F_1(g) = P_1^*$ ($F_2(g) = P_1^*, F_3(g) = P_1^*$) and h is determined by (9).

Now we consider the probability requirements when one of the equations (5), (7) and (8) is used with the equation (4) to determine the constants h and g in (6). First we consider the equations (7) and (8).

Theorem 1. *Suppose $P_1^* \geq 1/2$. If the equation (7) or (8) is used with the equation (4) to determine the constants h and g in (6), then the probability requirements (1) and (2) are satisfied.*

Unlike the equations (7) and (8), it is generally difficult to see if the probability requirements are satisfied when the equation (5) is used with the equation (4) to determine the constants h and g in (6). So we consider the special case $\delta_0^* = 0$ and $\delta_1^* = \delta_2^*$. Let $\phi(x) = \Phi'(x)$ and $D_+ = \sup_{x>0} x\phi(x)/\Phi(x)$.

Theorem 2. *Suppose $P_1^* \geq 1/2$, $\delta_0^* = 0$ and $\delta_1^* = \delta_2^*$. If the equation (5) is used with the equation (4) to determine the constants h and g in (6) and the constants satisfy*

$$g > \max\left(h + \frac{(k-1)D_+}{h},\, 2\sqrt{(k-2)\phi(1)}\right), \tag{10}$$

then the probability requirements (1) and (2) are satisfied.

A numerical calculation shows $D_+ = 0.29453$. See Bonfiger (1979, p.153). When $\delta_0^* = 0$ and $\delta_1^* = \delta_2^*$, the simultaneous equations (4) and (5) become

$$\Phi^k(h) = P_0^*, \quad \int_{h-g}^{\infty} \Phi^{k-1}(y+g)d\Phi(y) = P_1^*.$$

Since $h = \Phi^{-1}(P_0^{*-1/k})$, the inequality (10) is equivalent to

$$P_1^* > \int_{h-d}^{\infty} \Phi^{k-1}(y+d)d\Phi(y), \tag{11}$$

where $d = \max\left(h + (k-1)D_+/h, 2\sqrt{(k-2)\phi(1)}\right)$. So the probability requirements are satisfied if we choose the value of P_1^* greater than that of the right side of (11). Table 1 gives the values of the right side of (11) when $P_0^* = 0.5(0.05)0.95$ and $k = 2(1)9$.

		k							
		2	3	4	5	6	7	8	9
	.50	.638214	.685497	.733918	.778208	.817279	.851047	.879785	.903914
	.55	.617899	.672842	.723835	.769453	.809452	.844014	.873498	.898350
	.60	.603926	.662697	.715081	.761455	.802048	.837188	.867275	.892756
	.65	.593799	.654191	.707141	.753837	.794758	.830305	.860883	.886923
P_0^*	.70	.586063	.646699	.699596	.746260	.787285	.823090	.854067	.880612
	.75	.579794	.639724	.692063	.738379	.779296	.815222	.846511	.873519
	.80	.574338	.632809	.684121	.729771	.770361	.806257	.837769	.865197
	.85	.569137	.625425	.675199	.719812	.759803	.795484	.827101	.854898
	.90	.563547	.616743	.664285	.707327	.746316	.781491	.813029	.841102
	.95	.556284	.604735	.648724	.689136	.726293	.760340	.791372	.819476

Table 1. Lower bounds of P_1^* when $\delta_0^* = 0$ and $\delta_1^* = \delta_2^*$

We denote by $n_{iD}\ (n_{iR}, n_{iL}), i = 1, \cdots, k$, the sample sizes (6) in which the constant h and g are the solutions of the simultaneous equations (4) and (5) ((4) and (7), (4) and (8)). Letting g_D, g_R and g_L be the constants satisfying $F_1(g_D) = F_2(g_R) = F_3(g_L) = P_1^*$, the sample sizes $n_{iD}\ (n_{iR}, n_{iL}), i = 1, \cdots, k$, are n_i's in (6) with g replaced by $g_D(g_R, g_L)$. We can show that

$$g_R \geq g_D, \quad g_L \geq g_D. \tag{12}$$

The inequalities yield the following result.

Theorem 3. $n_{iR} \geq n_{iD}, \quad n_{iL} \geq n_{iD}, \quad i = 1, \cdots, k.$

This theorem shows that the sample sizes determined by the simultaneous equations (4) and (5) are better than those by the simultaneous equations (4) and (7) or (4) and (8), but it is not known that these sample sizes always guarantee the probability requirements. See Theorem 2. However, we can show that it is possible to construct a selection procedure which guarantees the probability requirements if we use other estimates of μ_i's instead of the sample means (see Dudewicz and Dalal, 1975).

Let $a_{ij}, j = 1, \cdots, n_{iD}, i = 1, \cdots, k$ be constants such that

$$\sum_{j=1}^{n_{iD}} a_{ij} = 1, \quad \sigma_i^2 \sum_{j=1}^{n_{iD}} a_{ij}^2 = \delta_2^{*2}/g^2, \quad i = 1, \cdots, k.$$

Such constants exist because $n_{iD} \geq g^2\sigma_i^2/\delta_2^{*2}, i = 1, \cdots, k$. Let $\tilde{X}_i = \sum_{j=1}^{n_{iD}} a_{ij}X_{ij}$ for n_{iD} observations $X_{i1}, \cdots, X_{in_{iD}}$ from $\Pi_i, i = 1, \cdots, k$. The selection procedure is the same as the previous one except that $\tilde{X}_i$'s are used instead of the sample means.

Theorem 4. *If the solutions h and g of the simultaneous equations (4) and (5) are used to determine the sample sizes in (6), then the above selection procedure satisfies the probability requirements (1) and (2).*

3 Unknown variances

In this section we suppose that the values of σ_i^2's are unknown. We propose two-stage selection procedures which satisfy (1) and (2) by using the methods of Dudewicz and Dalal (1975), Rinott (1978) and Lam (1988).

Let $X_{i1}, \cdots, X_{im}$ be the initial sample of size $m(\geq 2)$ from Π_i and let $\bar{X}_{i(m)} = \frac{1}{m}\sum_{j=1}^{m} X_{ij}$ and

$$V_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(X_{ij} - \bar{X}_{i(m)})^2, \quad i = 1, \cdots, k. \tag{13}$$

First we propose a two-stage selection procedure S_R related to that of Rinott (1978). Let $(h, g) = (h_{mR}, g_{mR})$ be the solutions of the simultaneous equations

$$F_{m-1}^k(h + \delta_0^* g/\delta_2^*) = P_0^* \tag{14}$$

and

$$\int_0^\infty \left\{ \int_0^\infty \Phi\left(\frac{g}{\sqrt{(m-1)(\frac{1}{x}+\frac{1}{y})}}\right) g_{m-1}(x)dx \right\}^{k-1} \Phi\left(\left(\frac{g\delta_1^*}{\delta_2^*} - h\right)\sqrt{\frac{y}{m-1}}\right) g_{m-1}(y)dy = P_1^*,$$

where F_ν is the distribution function of a t-distribution with ν degrees of freedom and g_ν is the density function of a chi-squared random variable with ν degrees of freedom. Then the total sample size R_i from Π_i is determined by

$$R_i = \max\left\{m, \left[\frac{g_{mR}^2 V_i^2}{\delta_2^{*2}}\right] + 1\right\}, \quad i = 1, \cdots, k.$$

If $R_i > m$, take $R_i - m$ additional observations $X_{im+1}, \cdots, X_{iR_i}$ from Π_i. Calculating the sample mean $\bar{X}_{i(R_i)} = \sum_{j=1}^{R_i} X_{ij}/R_i, i = 1, \cdots, k$ and denoting the largest sample mean by $\bar{X}_{[k]}$, we select the population which yields $\bar{X}_{[k]}$ as the one associated with $\mu_{[k]}$ if $\bar{X}_{[k]} > \mu_0 + c_R$, otherwise, we select no population, where $c_R = h_{mR}\delta_2^*/g_{mR}$.

Theorem 5. *If $P_1^* \geq 1/2$, then the selection procedure S_R satisfies the probability requirements (1) and (2).*

The selection procedure S_L related to that of Lam (1978) is different from S_R only in choosing the constants h_{mR} and g_{mR}. Let $(h, g) = (h_{mL}, g_{mL})$ be the solutions of the simultaneous equations (14) and

$$\frac{1}{2}F_{m-1}^{k-1}(g) + \int_{-a}^{0} F_{m-1}^{k-1}(x+g)dF_{m-1}(x) = P_1^*,$$

where $a = \min\{g, g\delta_1^*/\delta_2^* - h\}$. Then the total sample size M_i from Π_i is determined by

$$M_i = \max\left\{m, \left[\frac{g_{mL}^2 V_i^2}{\delta_2^{*2}}\right] + 1\right\}, \quad i = 1, \cdots, k,$$

where V_i^2's are (13).

Theorem 6. *If $P_1^* \geq 1/2$, then the selection procedure S_L satisfies the probability requirements* (1) *and* (2).

The selection procedure S_D related to that of Dudewicz and Dalal (1975) is different from S_R and S_L. Let $(h, g) = (h_{mD}, g_{mD})$ be the solutions of the simultaneous equations (14) and

$$\int_{h - g\delta_1^*/\delta_2^*}^{\infty} F_{m-1}^{k-1}(x+g)dF_{m-1}(x) = P_1^*.$$

Then the total sample size N_i from Π_i is determined by

$$N_i = \max\left\{m+1, \left[\frac{g_{mD}^2 V_i^2}{\delta_2^{*2}}\right] + 1\right\}, \quad i = 1, \cdots, k,$$

where V_i^2's are (13). Take $N_i - m(\geq 1)$ additional observations $X_{im+1}, \cdots, X_{iN_i}$ from Π_i $(i = 1, \cdots, k)$ and choose such constants $\{a_{ij}, j = 1, \cdots, N_i, i = 1, \cdots, k\}$ as

$$\sum_{j=1}^{N_i} a_{ij} = 1, \quad a_{i1} = \cdots = a_{im}, \quad V_i^2 \sum_{j=1}^{N_i} a_{ij}^2 = \delta_2^{*2}/g_{mD}^2, \quad i = 1, \cdots, k.$$

Letting $\tilde{X}_{i(N_i)} = \sum_{j=1}^{N_i} a_{ij}X_{ij}, i = 1, \cdots, k$, the selection procedure S_D is the same as S_R and S_L except that $\tilde{X}_{i(N_i)}$'s are used instead of the sample means.

Theorem 7. *The selection procedure S_D satisfies the probability requirements* (1) *and* (2).

Now we compare the selection procedures S_R, S_L and S_D in terms of their expected sample sizes.

Theorem 8. *Suppose $\delta_0^* = 0$ and $\delta_1^* = \delta_2^* (= \delta^*)$. If the initial sample size m is chosen such that $m \to \infty$ and $m\delta^{*2} \to 0$ as $\delta^* \to 0$. Then*

$$\lim_{\delta^* \to 0} \frac{E(R_i)}{E(N_i)} = \frac{g_R^2}{g_D^2}, \quad \lim_{\delta^* \to 0} \frac{E(M_i)}{E(N_i)} = \frac{g_L^2}{g_D^2}, \quad i = 1, \cdots, k,$$

where g_D, g_R and g_L are determined by $F_1(g_D) = F_2(g_R) = F_3(g_L) = P_1^*$.

From (12) it turns out that the selection procedure S_D is asymptotically more efficient than S_D and S_L in terms of the expected sample size.

When $\delta_0 > 0$ and $\delta_1^* \geq \delta_2^* > 0$, Wilcox (1984) proposed the following two-stage selection procedure. The total sample size Q_i from Π_i is determined by

$$Q_i = \max\left\{m, \left[\frac{h_{m1}^2 V_i^2}{\delta_0^2}\right] + 1, \left[\frac{h_{m2}^2 V_i^2}{\delta_2^2}\right] + 1\right\}, \quad i = 1, \cdots, k,$$

where V_i^2's are (13), $F_{m-1}^k(h_{m1}) = P_0^*$ and

$$\int_0^\infty \Phi\left(h_{m2}\sqrt{\frac{y}{m-1}}\right)\left(\int_0^\infty \Phi\left(\frac{h_{m2}}{\sqrt{(m-1)(\frac{1}{x}+\frac{1}{y})}}\right) g_{m-1}(x)dx\right)^{k-1} g_{m-1}(y)dy = P_1^*.$$

If $Q_i > m$, take $Q_i - m$ additional observations $X_{im+1}, \cdots, X_{iQ_i}$ from Π_i. Calculating $\bar{X}_{i(Q_i)} = \sum_{j=1}^{Q_i} X_{ij}/Q_i, i = 1, \cdots, k$ and letting $\bar{X}_{[k]}$ be the largest sample mean, we select the population that yields $\bar{X}_{[k]}$ as the one associated with $\mu_{[k]}$ if $\bar{X}_{[k]} > \mu_0$, otherwise, we select no population. Then the selection procedure satisfies the probability requirements (1) and (2). However, the selection procedure is improved by S_R in terms of the sample sizes.

Theorem 9. *Suppose $\delta_0 > 0$ and $\delta_1^* \geq \delta_2^* > 0$. Then*

$$Q_i \geq R_i, \quad i = 1, \cdots, k.$$

References

Bechhofer, R. E. and Turnbull, B. W. (1978). Two (k+1)-decision selection procedures for comparing k normal means with a specified standard. *J. Am. Stat. Assoc.*, **73**, 385–392.

Bofinger, E. (1979). Two stage selection problem for normal populations with unequal variances. *Austal. J. Statist.*, **21**, 149–156.

Dudewicz, E. L. and Dalal, S. R. (1975). Allocation of observations in ranking and selection with unknown variances. *Sankhā*, **B 37**, 28–78.

Lam, K. (1988). An improved two-stage selection procedure. *Commun. Statist. -Simula.*, **17**, 995–1006.

Mukhopadhyay N. (1979). Some comments on two-stage selection procedures. *Commun. Statist. -Theor. Meth.*, **8**, 671–683.

Rinott, Y. (1978). On two-stage selection procedures and related probability-inequalities. *Commun. Statist. -Theor. Meth.*, **7**, 799-811.

Takada, Y. (2007). Selection problem from normal populations with unequal variances. *Far East J. Theor. Stat.*, **23**, 165-176.

Wilcox, R. R. (1984) Selecting the best population, provided it is better than a standard: the unequal variance case. *J. Am. Stat. Assoc.*, **79**, 887–891.