JOURNAL OF APPLIED MATHEMATICS AND DECISION SCIENCES, 6(1), 23–42

An Integrated Selection Formulation for the Best Normal Mean: The Unequal and Unknown Variance Case

PINYUEN CHEN†
Department of Mathematics, Syracuse University, Syracuse, NY 13244-1130

JUN-LUE ZHANG
Department of Mathematics, Indiana Univ. of Pennsylvania, Indian, PA 15705- 1072

Abstract. This paper considers an integrated formulation in selecting the best normal mean in the case of unequal and unknown variances. The formulation separates the parameter space into two disjoint parts, the preference zone (PZ) and the indifference zone (IZ). In the PZ we insist on selecting the best for a correct selection (CS_1) but in the IZ we define any selected subset to be correct (CS_2) if it contains the best population. We find the least favorable configuration (LFC) and the worst configuration (WC) respectively in PZ and IZ. We derive formulas for $P(CS_1|LFC)$, $P(CS_2|WC)$ and the bounds for the expected sample size $E(N)$. We also give tables for the procedure parameters to implement the proposed procedure. An example is given to illustrate how to apply the procedure and how to use the table.

Keywords: Integrated formulation; two-stage selection procedure

1. Introduction

This paper studies the integrated approach in selecting the best normal mean among k normal populations with unequal and unknown variances. Unlike the case of common and unknown variance studied in Chen and Zhang (1997), we can not use the pooled sample variance to estimate the unknown variances in this case. One important change, compared to the case of common and unknown variance, is that in the case of unequal and unknown variances we use weighted averages as the estimators for the population means. Such change enables us to effectively evaluate the lower bounds of the probability of a correct selection.

Historically, many have studied multiple decision procedures in the case of unequal and unknown variances using the classical approaches. In the indifference zone approach, Bechhofer, Dunnett, and Sobel (1954) had men-

† Requests for reprints should be sent to Pinyuen Chen Department of Mathematics, Syracuse University, Syracuse, NY 13244.

tioned the possibility of a two-stage procedure in selecting the best population among k normal populations with unknown means and unequal and unknown variances. Dudewicz (1971) showed that under the indifference zone approach of Bechhofer (1954), a single-stage procedure is not appropriate in the case of unequal and unknown variances. Dudewicz and Dalal (1975) proposed a generalized Stein-type two-stage procedure using the indifference zone approach. In subset selection approach, Gupta and Huang (1974) have proposed a single-stage procedure based on unequal sample sizes for selecting a subset which would contain the best population when the variances are unknown and possibly unequal.

Chen and Sobel (1987) was the first article that proposed the integrated selection formulation. They studied a single-stage procedure for the common known variance case. The integrated formulation approach to the selection problem in the case of unequal and unknown variances has not been studied. However, such a case is important in applications since variances are often unknown and unequal in most of the real world problems. The objective of this paper is to develop a two-stage procedure, using the integrated approach, to select the best normal mean from k normal populations with unequal and unknown variances.

In section 2 we state our goal, assumptions and the probability requirements. We propose a two-stage procedure in section 3. In section 4 we derive lower bounds for the probability of a correct selection. These bounds will enable us to effectively compute the unknown parameters in our selection procedure and to guarantee the procedure to satisfy a given probability requirement (P_1^*, P_2^*). The experimenter can allocate sample sizes according to these parameters. In section 5, we develop bounds for the expected sample size for the proposed procedure. The integrated formulation requires our procedure to satisfy two probability requirements simultaneously. Therefore, it is reasonable that the expected sample size in our procedure is larger than the expected sample size in the indifference zone approach. Section 6 discusses the computation of the tables. Section 7 gives an illustrative example.

2. Assumptions, Goal, and The Probability Requirements

Suppose that we have k normal populations $\pi_1, \ldots, \pi_k$ with unknown means and unequal and unknown variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$. We denote the ordered means as $\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[k]}$ and denote $\pi_{(i)}$ as the population which corresponds to $\mu_{[i]}$. We also define the best population to be $\pi_{(k)}$, the population corresponding to the largest population mean $\mu_{[k]}$.

Our goal is to derive a two-stage selection procedure P_E which would

$$\begin{aligned} &\text{select } \pi_{(k)} \text{ if } \mu_{[k]} \geq \mu_{[k-1]} + \delta^*, \\ &\text{or,} \\ &\text{select a subset containing } \pi_{(k)} \text{ if } \mu_{[k]} < \mu_{[k-1]} + \delta^*, \end{aligned} \tag{1}$$

where $\delta^* > 0$ is a specified constant.

We first define the parameter space as follows:

$$\Omega = \{(\mu, \sigma^2) | -\infty < \mu_i < \infty,\ 0 < \sigma_i < \infty;\ i = 1, \ldots, k\}, \tag{2}$$

where $\mu = (\mu_1, \ldots, \mu_k)$ and $\sigma^2 = (\sigma_1^2, \ldots, \sigma_k^2)$.

We divide the parameter space into preference zone (PZ) and indifference zone (IZ), PZ and IZ are defined as follows, respectively.

$$PZ = \{(\mu, \sigma) \in \Omega | \mu_{[k]} - \mu_{[k-1]} \geq \delta^*\}, \tag{3}$$

$$IZ = \{(\mu, \sigma) \in \Omega | \mu_{[k]} - \mu_{[k-1]} < \delta^*\}, \tag{4}$$

where $0 < \delta^*$ is a prespecified constant.

We define CS_1 to be the event that our procedure selects the one best population when $\mu \in PZ$ and CS_2 to be the event that our procedure selects a subset that contains the best population when $\mu \in IZ$. We require that our two-stage selection procedure, P_E, which will be defined formally in Section 3, for a given (P_1^*, P_2^*), would satisfy the following probability requirements:

$$\begin{aligned} &P(CS_1 | P_E) \geq P_1^*, \text{ and} \\ &P(CS_2 | P_E) \geq P_2^*. \end{aligned} \tag{5}$$

3. Procedure P_E

We propose a Dudewicz-Dalal-type two-stage selection procedure.

Procedure P_E:

(i) Take an initial sample $X_{i1}, X_{i2}, \ldots, X_{in_0}$ of size n_0 (≥ 2) from population π_i $I = 1, 2, \ldots, k$.

Compute:

$$\bar{X}_i(n_0) = \sum_{j=1}^{n_0} \frac{X_{ij}}{n_0}, \tag{6}$$

$$S_i^2(n_0) = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} \left(X_{ij} - \bar{X}_i(n_0)\right)^2.$$

(ii) Define

$$n_i = \max\left\{n_0 + 1, \left[\left(\frac{h^* S_i}{\delta^* - c}\right)^2\right]\right\}. \tag{7}$$

$[y]$ denotes the smallest integer greater than or equal to y. Here $h^* = \max\{h_1^*, h_2^*\}$ and h_1^*, h_2^*, h_3^*, and c are chosen to satisfy the probability requirement (5). They are the solutions of the following integral equations: When $k = 2$, for given n_0 and specification $(\delta^*, P_1^*, P_2^*, a)$, the h_1^*, and h_2^* values simultaneously satisfy:

$$\int_{-\infty}^{+\infty} G\left(t + h_1^*\right) g(t)\, dt = P_1^*, \textit{ and} \tag{8}$$

$$\int_{-\infty}^{+\infty} G\left(t + \frac{h_2^*}{a-1}\right) g(t)\, dt = P_2^* \tag{9}$$

Here G and g are Student's t-distribution and density function, respectively. For any $k \geq 3$ and any n_0 and specification $(\delta^*, P_1^*, P_2^*, a)$, the h_1^*, h_2^* and h_3^* values simultaneously satisfy:

$$\int_{-\infty}^{+\infty} G^{k-1}\left(t + h_1^*\right) g(t)\, dt = P_1^*, \tag{10}$$

and

$$\begin{aligned} \frac{1}{k} + \quad & (k-1)\int_{-\infty}^{+\infty} G^{k-2}(t)\left[G\left(t + \frac{h_2^*}{(a-1)}\right) - G\left(t\right)\right] g(t)\, dt \\ + \quad & (k-1)(k-2)\int_{-\infty}^{+\infty} G^{k-3}\left(t\right)\left[G\left(t + \frac{h_2^*}{(a-1)}\right) - G\left(t\right)\right] \\ \times \quad & \left[G\left(t\right) - G\left(t - h_3^*\right)\right] g(t)\, dt = P_2^* \end{aligned} \tag{11}$$

Here G and g are Student's t-distribution and density function, respectively.

(iii) Take $n_i - n_0$ additional observations from the i^{th} population. Denote the observations by X_{ij}, where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$. Compute:

$$\widetilde{X}_i = \sum_{j=1}^{n_i} a_{ij} X_{ij} \quad i = 1, 2, \ldots, k, \tag{12}$$

where a_{ij}'s are to be chosen so that the following conditions are satisfied:

$$\sum_{j=1}^{n_i} a_{ij} = 1,\ a_{i1} = a_{i2} =, \ldots, = a_{in_0}, \tag{13}$$

and

$$S_i^2 \sum_{j=1}^{n_i} a_{ij}^2 = (\frac{\delta^* - c}{h^*})^2, \tag{14}$$

where $i = 1, 2, \ldots, k$, and we use $\widetilde{X}_{[1]} \leq \widetilde{X}_{[2]} \leq, \ldots, \leq \widetilde{X}_{[k]}$ to denote the ranked $\widetilde{X}$'s.

(iv) If $\widetilde{X}_{[k]} \geq \widetilde{X}_{[k-1]} + c$, we select the population associated with $\widetilde{X}_{[k]}$. If $\widetilde{X}_{[k]} < \widetilde{X}_{[k-1]} + c$, we select a random sized subset which contains all populations π_i with $\widetilde{X}_i \geq \widetilde{X}_{[k-1]} - d$.

Here $\delta^* > c$, $\delta^* = ac$, and $a > 1$ is given; $h^* = \max\{h_1^*, h_2^*\}$, $d = \frac{h_3^*}{\max\{h_1^*, h_2^*\}}(\delta^* - c)\widetilde{X}_i$ is the weighted average associated with population π_i.

The previous procedure would be meaningful only if the a_{ij} exist. One can show the existence of the a_{ij}'s through simple, but extended lines of algebra. Essentially what is being done on a_{ij}'s here is an adjustment to allow for the fact that sample size must be a whole number, and that therefore a standard error estimate based on the preliminary sample takes only discrete values if all observations are equally weighted. By allocating unequal weights, the estimated standard error can be equated to a specific quantity.

Result: There exist a_{ij}'s which satisfy:

$$\begin{aligned} \sum_{j=1}^{n_i} a_{ij} &= 1, \\ a_{i1} = a_{i2} &= \cdots = a_{in_0}, \\ S_i^2 \sum_{j=1}^{n_i} a_{ij}^2 &= (\frac{\delta^* - c}{h^*})^2, \end{aligned} \tag{15}$$

where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$.

4. Lower Bounds for $P(CS_1)$ and $P(CS_2)$

To derive lower bounds for the probability of a correct selection, one needs to find the least favorable configuration as well as the worst configuration. We first define the least favorable configuration in the PZ and the worst configuration in the IZ.

Definition 1 *For any $\sigma^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$, the least favorable configuration in PZ is defined to be:*

$$LFC|P_E = \left\{(\mu_0, \sigma^2)|P\left(CS_1|(\mu_0, \sigma^2), P_E\right) = \inf_{\mu \in PZ} P\left(CS_1|(\mu, \sigma^2), P_E\right)\right\}. \tag{16}$$

Definition 2 *For any $\sigma^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$, the worst configuration in IZ is defined to be:*

$$WC|P_E = \left\{(\mu_1, \sigma^2)|P\left(CS_2|(\mu_1, \sigma^2 P_E\right) = \inf_{\mu \in IZ} P\left(CS_2|(\mu, \sigma^2 P_E\right)\right)\}. \tag{17}$$

To derive lower bounds for $P(CS_1)$ and $P(CS_2)$ on the parameter space Ω we first show that (for any σ^2),

$$LFC|P_E = \{(\mu, \sigma^2)|\delta_{ki} = \delta^* \quad \forall i \neq k\}, \tag{18}$$

where $\delta_{ki} = \mu_{[k]} - \mu_{[i]}$ and

$$WC|P_E = \{(\mu, \sigma^2)|\delta_{ki} = 0 \quad \forall i \neq k\}.$$

Lemma 1 *Let $T_i = \dfrac{\widetilde{X}_{(i)} - \mu_{[i]}}{\frac{\delta^* - c}{h^*}}$, then T_i's have independent student's t-distribution with $n_0 - 1$ degrees of freedom, $i = 1, 2, \ldots, k$.*

Proof: The proof can be found in Stein (1945). □

As the denominator $(s^* - c)/h^*$ is a constant, this lemma can only be true because the additional sample sizes n_i are random.

Theorem 1 *Under procedure P_E the LFC for $P(CS_1|PZ)$ is given by the slippage configuration, i.e. by $\mu_{[1]} = \cdots = \mu_{[k-1]} = \mu_k - \delta^*$ and the WC for $P(CS_2|IZ)$ is given by the equal parameter configuration, i.e. by $\mu_{[1]} = \cdots = \mu_{[k]}$.*

Proof: From (18), we find that the random variable $T_i (i = 1, 2, \ldots, k)$ has a t- distribution with $n_0 - 1$ degrees of freedom.
Rewrite $\widetilde{X}_I$ as

$$\widetilde{X}_{(i)} = \left(\frac{\delta^* - c}{n^*}\right) T_i + \mu_{[i]} \tag{19}$$

and consider the family of distribution function $\{(G_n(X|\mu))\}$ where G_n is the distribution of the random variable $\left(\frac{\delta^* - c}{n^*}\right) \cdot t_{n-1} + \mu$ where $\frac{\delta^* - c}{n^*}$ is a constant, μ is the parameter of interest, and t_{n-1} is the random variable which has t distribution with $n - 1$ degrees of freedom. Then it is clear

that $\{(G_n(X|\mu))\}$ is a stochastically increasing family in μ. We now show that the LFC for $P(CS_1|PZ)$ is given by $\mu_{[1]} = \cdots = \mu_{[k-1]} = \mu_{[k]} - S^*$. The proof of the WC for $P(CS_2|IZ)$ is similar. We start with an arbitrary configuration in the IZ

$$\mu_{[1]} \leq \mu_{[1]} \leq \cdots \leq \mu_{[k]} \text{ with } \mu_{[k]} - \mu_{[k_1]} S^*$$

Letting $\bar{X}(i)$ denote the sample mean associated with $\mu_{[i]}$, we have

$$P(CS_1|PZ) = P(\widetilde{X}_{(k)} > \max_{1\leq\beta\leq k-1} \widetilde{X}_{(\beta)} + C).$$

Define the function $\psi = \psi(y_1, y_2, \ldots, y_k)$ by

$$\psi = \begin{cases} 1 & \text{if } Y_k > \max\limits_{l\leq\beta\leq k-t} y_\beta + C \\ 0 & \text{otherwise} \end{cases}$$

Then we have $P(CS_1|PZ) = E\psi\big(\widetilde{X}_{(1)}, \widetilde{X}_{(2)}, \ldots, \widetilde{X}_{(k)}\big)$. It is clear that $\psi(y_1, y_2, \ldots, y_k)$ is non-increasing in Y_i (for $i = 1, \ldots, k-1$) when all the y_i for $j \neq i$ are held fixed. Since $\widetilde{X}$'s are from a stochastically increasing family, we use Lemma 5.1 by Chen and Sobel (1987) to conclude that $P(CS_1|PZ)$ is non-increasing in $\mu_{[i]}$ for $i = 1, 2, \ldots, k-1$ and it is non-decreasing in $u_{[k]}$. This completes the proof of the Theorem. □

Lemma 2 *Under procedure P_E, the probability of a correct selection in the PZ and the IZ are, respectively:*

$$P(CS_1|P_E) = P(\widetilde{X}_{(k)} \geq \widetilde{X}_{(i)} + c; \quad i = 1, 2, \ldots, k-1), \tag{20}$$

$$P(CS_2|P_E) = H_0 + H_1 + H_2, \tag{21}$$

where

$$H_0 = P\left(\widetilde{X}_{(k)} \geq M_0\right) = P\left(\widetilde{X}_{(k)} \geq \widetilde{X}_{(i)}; \quad i = 1, 2, \ldots, k-1\right); \tag{22}$$

$$H_1 = P\left(M_i \leq \widetilde{X}_{(k)} < \widetilde{X}_{(i)} < \widetilde{X}_{(k)} + c, \quad i = 1, 2, \ldots, k-1\right) \tag{23}$$

$$= \sum_{i=1}^{k-1} P\left(\widetilde{X}_{(i)} > \widetilde{X}_{(k)} > \widetilde{X}_{(j)},\ \widetilde{X}_{(k)} + c > \widetilde{X}_{(i)},\ j = 1, 2, \ldots, k-1,\ j \neq i\right);$$

$$H_2 = P\left(M_i - d \leq \widetilde{X}_{(k)} \leq M_i \leq \widetilde{X}_{(i)} \leq M_i + c, \quad i = 1, 2, \ldots, k-1\right)$$

$$= \sum_{i=1}^{k-1} \sum_{j=1,\, j\neq i}^{k-1} P(\widetilde{X}_{(i)} > \widetilde{X}_{(j)} > \widetilde{X}_{(m)},\ m = 1, 2, \ldots, k-1,\ m \neq i, j;$$

$$\widetilde{X}_{(j)} > \widetilde{X}_{(k)} > \widetilde{X}_{(j)} - d;\ \widetilde{X}_{(j)} + c > \widetilde{X}_{(i)}). \tag{24}$$

and

$$M_0 = \max\{\widetilde{X}_{(\alpha)} | \alpha = 1, 2, \ldots, k-1\},$$
$$M_i = \max\{\widetilde{X}_{(\alpha)} | \alpha = 1, 2, \ldots, k-1,\ \alpha \neq i\}. \tag{25}$$

Proof: The result is clear for $P(CS_1|P_E)$. For $P(CS_2|P_s)$, H_0, H_1 and H_2 correspond to the cases of $\widetilde{X}_{(k)}$ being the largest, the second longest, and neither the largest nor the second largest, respectively.

The following theorems give lower bounds for $P(CS_1|P_E)$ and $P(CS_2|P_E)$.

Theorem 2 *When $k = 2$, for given n_0 and specification $(\delta^*, P_1^*, P_2^*, a)$, the h_1^*, and h_2^* values which simultaneously satisfy:*

$$\int_{-\infty}^{+\infty} G\left(t + h_1^*\right) g(t)\, dt = P_1^*, \quad \textit{and} \tag{26}$$

$$\int_{-\infty}^{+\infty} G\left(t + \frac{h_2^*}{a-1}\right) g(t)\, dt = P_2^* \tag{27}$$

are the values for procedure P_E to satisfy the probability requirement (5)*. Here G and g are Student's t-distribution and density function, respectively.*

Remark: When $k = 2$, $d > 0$ can be arbitrarily chosen since if we did not select the one best population, we would select two populations regardless the value of d.

Proof: Denote $\frac{\delta^* - c}{h^*}$ by e^*. By Lemma 2,

$$\begin{aligned} P(CS_1|P_E) &= P(\widetilde{X}_{(2)} \geq \widetilde{X}_{(1)} + c) \\ &= P(T_1 \leq T_2 + \frac{\delta_{21} - c}{e^*}) \\ &\geq P(T_1 \leq T_2 + h_1^*) \\ &= \int_{-\infty}^{\infty} G(t + h_1^*) g(t) dt = P_1^*. \end{aligned} \tag{28}$$

By Lemma 2, $P(CS_2|P_E) = H_0 + H_1 + H_2$. When $k = 2$, the term H_2 does not exist. Thus

$$H_0 = P(\widetilde{X}_{(2)} > \widetilde{X}_{(1)}) = P(T_1 \leq T_2 + \frac{\delta_{21}}{e^*}), \tag{29}$$

$$\begin{aligned} H_1 &= P(\widetilde{X}_{(2)} < \widetilde{X}_{(1)}, \widetilde{X}_{(1)} < \widetilde{X}_{(2)} + c) \\ &= P(\widetilde{X}_{(2)} < \widetilde{X}_{(1)} < \widetilde{X}_{(2)} + c) \\ &= P(T_2 + \frac{\delta_{21}}{e^*} \leq T_1 < T_2 + \frac{\delta_{21} + c}{e^*}). \end{aligned} \tag{30}$$

Therefore,

$$\begin{aligned} P(CS_2|P_E) &= H_0 + H_1 \\ &= P(T_1 < T_2 + \frac{\delta_{21}}{e^*}) + P(T_2 + \frac{\delta_{21}}{e^*} < T_1 < T_2 + \frac{\delta_{21}+c}{e^*}) \\ &= P(T_1 < T_2 + \frac{\delta_{21}+c}{e^*}) \qquad (31) \\ &\geq \int_{-\infty}^{\infty} G(t + \frac{c}{e^*}) g(t) dt \\ &\geq \int_{-\infty}^{\infty} G(t + \frac{h_2^*}{a-1}) g(t) dt = P_2^*. \end{aligned}$$

The first inequality follows from the fact that T_1 and T_2 both have students' t distributions and $\delta_{21} = \mu_{[2]} - \mu_{[1]}$ □

From Theorem 2, it is clear that as h_1^*, $h_2^* \to \infty$, the left hand sides of (26) **and** (27) **approach** 1**.**

Theorem 3 *For any $k \geq 3$ and any n_0 and specification $(\delta^*, P_1^*, P_2^*, a)$, the h_1^*, h_2^* and h_3^* values which simultaneously satisfy:*

$$\int_{-\infty}^{+\infty} G^{k-1}(t + h_1^*)\, g(t)\, dt = P_1^*, \text{ and} \qquad (32)$$

$$\begin{aligned} &\frac{1}{k} + (k-1) \int_{-\infty}^{+\infty} G^{k-2}(t) \left[G\left(t + \frac{h_2^*}{(a-1)} \right) - G(t) \right] g(t)\, dt \\ &\quad + (k-1)(k-2) \int_{-\infty}^{+\infty} G^{k-3}(t) \left[G\left(t + \frac{h_2^*}{(a-1)} \right) - G(t) \right] \\ &\quad \times \left[G(t) - G(t - h_3^*) \right] g(t)\, dt = P_2^* \qquad (33) \end{aligned}$$

are the values for procedure P_E to satisfy the probability requirement (5)*. Here G and g are Student's t-distribution and density function, respectively.*

Proof: The proof of Theorem 3 is lengthy. It is omitted here. The readers may contact the first author for a full version of the manuscript which contains the proof.

The left hand side of the integral equations in (32) and in (33) in Theorem 3 are increasing in h_1^*, h_2^* and h_3^*. Indeed, when h_1^* approaches infinity, the left hand side of (32) increases to 1. When h_2^* and h_3^* approach infinity, the left hand side of (33) also increases to 1. Thus we can always find h_1^*, h_2^*, and h_3^* that satisfy the probability requirements P_1^* and P_2^*.

One should note that it is necessary to let h_1^*, h_2^* and h_3^* vary freely so that our procedure will be applicable for any given probability requirements. Otherwise, the integral equations in (32) and in (33) might not have a solution, and in such a case, procedure P_E is not applicable. For instance, if one requires $h_1^* = h_2^*$, then for some (P_1^*, P_2^*) the integral equations in (32) and in (33) might not have a solution.

In procedure P_E, we let $\delta^* = ac$, $a > 1$. Such a requirement has the advantage that the lower bounds of the probability of a correct selection do not involve c. Instead of letting $\delta^* = ac$, $a > 1$, one can require that $\delta^* = a + c$, $a > 0$. In such a case, (32) in Theorem 3 is unchanged. But (33) is changed to:

$$\begin{aligned}&\frac{1}{k} + (k-1)\int_{-\infty}^{+\infty} G^{k-2}(t)\left[G\left(t + \frac{h_2^* c}{a}\right) - G(t)\right] g(t)\,dt \\ &\quad + (k-1)(k-2)\int_{-\infty}^{+\infty} G^{k-3}(t)\left[G\left(t + \frac{h_2^* c}{a}\right) - G(t)\right] \\ &\quad \times \left[G(t) - G(t - h_3^*)\right] g(t)\,dt = P_2^*.\end{aligned}$$

5. The Expected Sample Sizes and The Expected Subset Size

The total sample size n_i from population π_i $(i = 1, 2, \ldots, k)$ in procedure P_E can be calculated from (7),

$$n_i = \max\left\{n_0 + 1, \left[\left(\frac{S_i h^*}{\delta^* - c}\right)^2\right]\right\}.$$

It is clear that n_i, $i = 1, 2, \ldots, k$, are random variables. The expected values of the sample sizes are often valuable to the experimenter. In our case, studying the expected sample size is especially important since there are two unknowns in the integral equation (11) with only one constraint. Thus we have infinitely many solutions. It is clear that we need some additional guidelines to choose h_2^* and h_3^*. The expected sample size, which is a function of h^*, will give us some idea about how h^* relates to $E(n_i)$. It is reasonable to choose h_2^* and h_3^* to minimize the expected sample sizes in addition to satisfying the probability requirements. To evaluate the expected sample sizes, we use the method of Stein (1945).

Theorem 4 *For any $i \in \{1, 2, \ldots, k\}$, the expected sample size $E(n_i)$ for procedure P_E satisfies the following inequality:*

$$(n_0+1)F_{n_0-1}\left(\frac{(n_0^2-1)e^{*^2}}{\sigma_i^2}\right) + \frac{\sigma_i^2}{e^{*^2}}\left[1 - F_{n_0+1}\left(\frac{(n_0^2-1)e^{*^2}}{\sigma_i^2}\right)\right] \leq E(n_i)$$
$$< (n_0+1)F_{n_0-1}\left(\frac{(n_0^2-1)e^{*^2}}{\sigma_i^2}\right) + \frac{\sigma_i^2}{e^{*^2}}\left[1 - F_{n_0+1}\left(\frac{(n_0^2-1)e^{*^2}}{\sigma_i^2}\right)\right]$$
$$+ \left[1 - F_{n_0-1}\left(\frac{(n_0^2-1)e^{*^2}}{\sigma_i^2}\right)\right], \tag{34}$$

where $F_i(x)$ is a chi-squared probability distribution function with i degrees of freedom and $e^{*^2} = \left(\frac{\delta^* - c}{h^*}\right)^2$.

Proof: The proofs follow the ideas of Stein (1945). It is omitted here. Readers are recommended to contact the first author for a full version of the transcript which contains the proof.

Corollary 1 *For each i, $i = 1, 2, \ldots, k$, the expected sample size $E(n_i)$ has the following properties:*

1. *For fixed e^{*^2}, $E(n_i) \longrightarrow \infty$ as $\sigma_i^2 \longrightarrow \infty$ (the lower bound of $E(n_i)$ goes to $+\infty$).*
2. *For fixed e^{*^2}, $E(n_i) \longrightarrow n_0 + 1$ as $\sigma_i^2 \longrightarrow 0$ (the upper bound of $E(n_i)$ goes to $n_0 + 1$).*
3. *For fixed σ_i^2, $E(n_i) \longrightarrow \infty$ as $e^{*^2} \longrightarrow 0$ (the lower bound of $E(n_i)$ goes to $+\infty$).*
4. *The difference between the upper bounds and the lower bounds of $E(n_i)$ is at most 1 since $\left[1 - F_{n_0-1}\left(\frac{(n_0^2-1)e^{*^2}}{\sigma_i^2}\right)\right]$ is less than 1.*

Proof: These properties are immediate by Theorem 4.

6. Tables

To carry out procedure P_E, one needs the values of h_1^*, h_2^*, and h_3^*. In Table 1, we provide a table of the h_1' value, for the cases $k = 3, 4$, which

satisfies the following integral equation:

$$\int_{-\infty}^{+\infty} G\left(t + h_1'\right) g(t)\, dt = P^*, \tag{35}$$

for $P^* = .5, .75, .90, .95, .99$.

As discussed in section 4, there are infinitely many solutions for the integral equation (33). Therefore, it is impossible to provide tables which would cover all the practical situations. A particular solution of the integral equation (33) might be good for one objective yet might not be suitable for another goal.

Table 1. This table provides some h_1' values for procedure P_E.

Number of populations: $k = 3$					
n_0	Probability (P)				
	.50	.75	.90	.95	.99
3	.7620	2.1560	4.0560	5.8750	13.1800
4	.6820	1.8650	3.2110	4.2840	7.4000
5	.6515	1.7390	2.8960	3.7500	5.9330
6	.6312	1.6700	2.7360	3.4810	5.2400
7	.6180	1.6260	2.6340	3.3180	4.8500
8	.6090	1.5960	2.5680	3.2100	4.6330
9	.6022	1.5740	2.5200	3.1340	4.4600
10	.5970	1.5578	2.4850	3.0746	4.3500
11	.5930	1.5445	2.4550	3.0370	4.2580
12	.5890	1.5318	2.4320	3.0060	4.1781
13	.5860	1.5240	2.4160	2.9800	4.1480
14	.5840	1.5180	2.4010	2.9560	4.0800
15	.5820	1.5128	2.3850	2.9360	4.0400
20	.5760	1.4900	2.3440	2.8560	3.8600
25	.5720	1.4770	2.3180	2.8260	3.8000
30	.5690	1.4700	2.3000	2.8000	3.7600

Table 1. Continuation.

Number of populations: $k = 4$					
n_0	Probability (P)				
	.50	.75	.90	.95	.99
3	1.1860	2.6615	4.7800	6.8200	15.1000
4	1.0540	2.2500	3.6328	4.8000	8.2500
5	.9940	2.0810	3.2630	4.1360	6.0960
6	.9570	1.9880	3.0600	3.8150	5.6400
7	.9390	1.9310	2.9360	3.6160	5.2000
8	.9240	1.8920	2.8560	3.4980	4.9300
9	.9130	1.8630	2.7940	3.4100	4.7310
10	.9040	1.8410	2.7520	3.3450	4.6170
11	.8960	1.8230	2.7200	3.2960	4.5190
12	.8910	1.8100	2.6920	3.2580	4.4400
13	.8860	1.7970	2.6690	3.2280	4.3600
14	.8820	1.7890	2.6500	3.1980	4.3310
15	.8790	1.7820	2.6350	3.1760	4.2760
20	.8680	1.7530	2.5820	3.0880	4.0700
25	.8610	1.7380	2.5560	3.0450	4.0100
30	.8570	1.7280	2.5360	3.0220	3.960

We tabulate in Table 2 the values of h_2' and h_3' for $k = 3, 4$, $P_2^* = .50, .75, .90, .99$, where h_2' and h_3' satisfy the following integral equation:

$$\frac{1}{k} + (k-1)\int_{-\infty}^{+\infty} G^{k-2}(t)\left[G(t+h_2') - G(t)\right] g(t)\,dt + (k-1)(k-2)\int_{-\infty}^{+\infty} G^{k-3}(t)\left[G(t+h_2') - G(t)\right] \times \left[G(t) - G(t-h_3')\right] g(t)\,dt = P_2^*. \tag{36}$$

The relationship between h_2^*, h_3^* and h_2', h_3' are as follows:

$$h_2^* = (a-1)h_2', \quad h_3^* = h_3'. \tag{37}$$

The computation of Table 2 follows the following assumptions:

1. We take $a = 2$ (thus, $c = \frac{1}{2}\delta^*$).
2. We take $h_1^* = h_2^* = h_1' = h_2'$ where h_1' is the value corresponding to $P_1^* = P_2^* = .50, .75, .90, .95, .99$ in Table 1, respectively.
3. The probability is accurate to $\pm .0003$.

Table 2. This table provides some $(\prime h_2, \prime h_3)$ values for procedure P_E.

Number of populations: $k = 3$					
n_0	Probability (P)				
	.50	.75	.90	.95	.99
3	.7620	2.1560	4.0560	5.8750	13.1800
	.3860	1.2180	2.4850	3.8500	9.9600
4	.6820	1.8650	3.2110	4.2840	7.4000
	.3550	1.0200	1.9450	2.7160	5.3000
5	.6515	1.7390	2.8960	3.7500	5.9330
	.3260	.9580	1.7480	2.3560	3.9100
6	.6312	1.6700	2.7360	3.4810	5.2400
	.3210	.9180	1.6380	2.1800	3.5000
7	.6180	1.6260	2.6340	3.3180	4.8500
	.3160	.8940	1.5830	2.1000	3.3600
8	.6090	1.5960	2.5680	3.2100	4.6330
	.3100	.8745	1.5320	2.0160	3.1300
9	.6022	1.5740	2.5200	3.1340	4.4600
	.3060	.8570	1.5030	1.9650	3.0300
10	.5970	1.5578	2.4850	3.0746	4.3500
	.3050	.8500	1.4800	1.9460	2.9200
11	.5930	1.5445	2.4550	3.0370	4.2580
	.3020	.8420	1.4620	1.9160	2.8800
12	.5890	1.5318	2.4320	3.0060	4.1781
	.2990	.8360	1.4500	1.8830	2.7700
13	.5860	1.5240	2.4160	2.9800	4.1480
	.2970	.8300	1.4360	1.8680	2.8000
14	.5840	1.5180	2.4010	2.9560	4.0800
	.2950	.8260	1.4250	1.8400	2.7600
15	.5820	1.5128	2.3850	2.9360	4.0400
	.2930	.8210	1.4180	1.8390	2.7400
20	.5760	1.4900	2.3440	2.8560	3.8600
	.2900	.8080	1.3900	1.8060	2.6800
25	.5720	1.4770	2.3180	2.8260	3.8000
	.2880	.8030	1.3810	1.7900	2.6150
30	.5690	1.4700	2.3000	2.8000	3.7600
	.2879	.8000	1.3660	1.7730	2.6140

Note: Here we let $h_2' = h_1'$.

Table 2. Continuation.

Number of populations: $k = 4$					
n_0	Probability (P)				
	.50	.75	.90	.95	.99
3	1.1860	2.6615	4.7800	6.8200	15.1000
	.6500	1.6220	3.1200	4.5800	12.0000
4	1.0540	2.2500	3.6328	4.8000	8.2500
	.5760	1.3620	2.3750	3.2000	12.9000
5	.9940	2.0810	3.2630	4.1360	6.0960
	.5410	1.2560	2.1030	2.7720	4.6500
6	.9570	1.9880	3.0600	3.8150	5.6400
	.5260	1.1990	1.9920	2.5170	3.9950
7	.9390	1.9310	2.9360	3.6160	5.2000
	.5100	1.1658	1.9200	2.4640	3.7960
8	.9240	1.8920	2.8560	3.4980	4.9300
	.5000	1.1380	1.8660	2.3780	3.5300
9	.9130	1.8630	2.7940	3.4100	4.7310
	.4935	1.1190	1.8260	2.3190	3.4000
10	.9040	1.8410	2.7520	3.3450	4.6170
	.4880	1.1050	1.8000	2.2800	3.2900
11	.8960	1.8230	2.7200	3.2960	4.5190
	.4810	1.0950	1.7800	2.2480	3.2500
12	.8910	1.8100	2.6920	3.2580	4.4400
	.4780	1.0860	1.7560	2.2200	3.2050
13	.8860	1.7970	2.6690	3.2280	4.3600
	.4776	1.0780	1.7500	2.1900	3.1450
14	.8820	1.7890	2.6500	3.1980	4.3310
	.4760	1.0370	1.7380	2.1860	3.1400
15	.8790	1.7820	2.6350	3.1760	4.2760
	.4730	1.0680	1.7300	2.1750	3.1130
20	.8680	1.7530	2.5820	3.0880	4.0700
	.4670	1.0520	1.6980	2.1300	3.0330
25	.8610	1.7380	2.5560	3.0450	4.0100
	.4660	1.0430	1.6840	2.1030	3.0200
30	.8570	1.7280	2.5360	3.0220	3.9600
	.4640	1.0380	1.6720	2.1000	2.9800

We use Fortran77 to program the double integrals. Integration is carried out by the Romberg numerical method (Burden and Faires (1988)) in which

Neville's algorithm (Burden and Faires (1988)) is used for extrapolation. We modified the subroutines provided by Press, Teukolsky, Vetterling, and Flannery (1992) for our program. The upper limits of the integration for the student's t-density functions depend on the degree of freedom of the density function. All real variables are declared as double precision. Programs are executed under a UNIX environment using SUN4 600 Series and SUN4 Sparc 2000 machines.

We also provide a table (Table 3) of the approximation for the expected sample sizes using the $h = h_1'$ value obtained in Table 1 and for $r = \frac{(\delta^* - c)^2}{\sigma_i^2}$. Mathematica was used to perform the calculation. We compute the approximation of the expected sample sizes using the lower bound formula for $E(n_i)$ in Theorem 4. The formula is:

$$(n_0 + 1)F_{n_0-1}\left(\frac{(n_0^2 - 1)r}{h^2}\right) + \frac{h^2}{r}\left[1 - F_{n_0+1}\left(\frac{(n_0^2 - 1)r}{h^2}\right)\right]. \tag{38}$$

By (38), it is clear that $E(n_i)$ is dominated by $\frac{h^2}{r}$ when r is small, h is large, and n_0 is not very large. Indeed, from Table 3 one sees that the change of $E(n_i)$ is proportional to the change of r for a fixed h and when r is small, h is large, and n_0 is not very large. In fact $\frac{h^2}{r}$ is a precise estimate of $E(n_i)$ when r is small, h is large, and n_0 is not very large.

Table 3. This table provides some approximations of the expected sample sizes for procedure P_E.

$k = 3, \quad P_1^* = .90$									
n_0	r								
	.05	.10	.30	.45	.60	.75	1.0	1.25	1.5
3	329.047	164.560	54.980	36.769	27.697	22.278	16.900	13.712	11.616
4	206.210	103.105	34.374	22.930	17.222	13.814	10.445	8.472	7.205
6	149.714	74.857	24.961	16.681	12.606	10.252	8.040	7.046	6.514
8	131.892	65.946	22.002	14.794	11.429	9.705	8.493	8.129	8.031
10	123.505	61.760	20.993	14.961	12.610	11.648	11.136	11.027	11.005
15	113.765	56.882	19.474	15.541	15.036	15.001	15.000	15.000	15.000
20	109.887	54.944	21.006	20.010	20.000	20.000	20.000	20.000	20.000
25	107.462	53.732	25.069	25.000	25.000	25.000	25.000	25.000	25.000

Table 3. Continuation.

$k = 3, \quad P_1^* = .95$									
n_0	r								
	.05	.10	.30	.45	.60	.75	1.0	1.25	1.5
3	690.324	345.179	115.121	76.804	57.662	46.190	34.739	27.889	23.339
4	367.053	183.527	61.177	40.787	30.596	24.485	18.384	14.739	12.324
6	242.347	121.174	40.392	26.934	20.217	16.208	12.261	9.990	8.582
8	206.082	103.041	34.349	22.914	17.242	13.922	10.853	9.341	8.606
10	189.113	94.558	31.621	21.397	16.632	14.119	12.169	11.430	11.155
15	172.402	86.201	28.763	19.641	16.249	15.254	15.010	15.000	15.000
20	163.135	81.567	27.456	20.932	20.050	20.001	20.000	20.000	20.000
25	159.726	79.863	28.024	25.060	25.000	25.000	25.000	25.000	25.000

$k = 3, \quad P_1^* = .99$									
n_0	r								
	.05	.10	.30	.45	.60	.75	1.0	1.25	1.5
3	3474.250	1737.130	579.055	386.048	289.548	231.651	173.758	143.027	115.877
4	1095.200	547.600	182.533	121.689	91.267	73.014	54.761	43.811	36.512
6	549.152	274.576	91.525	61.017	45.763	36.612	27.464	21.981	18.337
8	429.294	214.647	71.549	47.700	35.776	28.625	21.488	17.240	14.464
10	378.450	189.225	63.083	42.086	31.639	25.446	19.448	16.107	14.124
15	326.432	163.216	54.405	36.274	27.248	21.986	17.425	15.862	15.155
20	297.992	148.996	49.666	33.160	25.349	21.748	20.152	20.007	20.000
25	288.800	144.400	48.138	32.441	26.546	25.178	25.002	25.000	25.000

$k = 4, \quad P_1^* = .90$									
n_0	r								
	.05	.10	.30	.45	.60	.75	1.0	1.25	1.5
3	456.985	228.519	76.265	50.928	38.284	30.716	23.179	18.686	15.714
4	263.974	131.987	43.999	29.339	22.017	17.633	13.273	10.686	8.993
6	187.272	93.636	31.216	20.827	15.664	12.615	9.693	8.108	7.206
8	163.135	81.567	27.196	18.179	13.786	11.330	9.287	8.467	8.158
10	151.470	75.739	25.461	17.561	14.130	12.502	11.433	11.120	11.032
15	138.865	69.432	23.292	16.844	15.270	15.027	15.000	15.000	15.000
20	133.334	66.667	23.264	20.143	20.002	20.000	20.000	20.000	20.000
25	130.663	65.331	25.670	25.001	25.000	25.000	25.000	25.000	25.000

Table 3. Continuation.

$k = 4, \quad P_1^* = .95$									
n_0	r								
	.05	.10	.30	.45	.60	.75	1.0	1.25	1.5
3	930.257	465.141	155.092	103.437	77.622	62.143	46.680	37.417	31.256
4	460.800	230.400	76.801	51.202	38.404	30.728	23.057	18.463	15.410
6	291.085	145.542	48.515	32.346	24.267	19.431	14.630	11.808	9.998
8	244.720	122.360	40.787	27.198	20.423	16.401	15.529	10.441	9.288
10	223.781	111.891	37.353	25.091	19.196	15.924	13.250	11.935	11.401
15	201.740	100.870	33.631	22.600	17.771	15.831	15.072	15.004	15.000
20	190.715	95.357	31.858	22.549	20.299	20.020	20.000	20.000	20.000
25	185.441	92.720	31.389	25.386	25.006	25.000	25.000	25.000	25.000

$k = 4, \quad P_1^* = .99$									
n_0	r								
	.05	.10	.30	.45	.60	.75	1.0	1.25	1.5
3	4560.200	2280.100	760.044	605.705	380.038	304.040	228.045	182.452	152.059
4	1361.250	680.625	226.875	151.250	113.438	90.750	68.063	54.452	45.378
6	636.192	318.096	106.032	70.688	53.016	42.414	31.813	25.456	21.224
8	486.098	243.049	81.016	54.011	40.509	32.409	24.317	19.482	16.293
10	426.334	213.167	71.606	47.393	35.595	28.568	21.685	17.758	15.348
15	365.684	182.842	60.947	40.633	30.492	24.482	18.911	16.384	15.412
20	331.298	165.649	55.217	36.828	27.845	23.166	20.442	20.035	20.002
25	321.602	160.801	53.601	35.865	28.145	25.577	25.012	25.000	25.000

7. An Illustrative Example

Now we present an example to illustrate the procedure P_E.

Example: Suppose that we are given three normal populations with unequal and unknown variances. Suppose that we wish to use the integrated formulation to select the population having the largest population mean if $\mu_{[3]} - \mu_{[2]} \geq 1$, and to select a subset that contains the longest mean if $\mu_{[3]} - \mu_{[2]} < 1$.

Suppose that for certain practical reasons, the experimenter decides to take a initial sample of size $n_0 = 15$. We use Fortran to generate three random samples of size 15 from population $N(4, .9^2)$, $N(4.5, 1^2)$, and $N(5.5, 1.5^2)$.

We obtain:

$$\sum_{j=1}^{15} X_{1j} = 57.4729, \sum_{j=1}^{5} X_{2j} = 63.6917, \sum_{j=1}^{5} X_{3j} = 89.5628, \tag{39}$$
$$S_1(15) = .76247, \quad S_2(15) = .82931, \quad S_3(15) = 1.2974.$$

Now we suppose that the experimenter has specified $P_1^* = P_2^* = .95$ and $\delta^* = 1$. Suppose that the experimenter also specified $a = 2$ (i.e. $c = \frac{1}{2}$). From Table 1 with $k = 3$, $n_0 = 15$, and $P_1^* = .95$, the experimenter finds $h_1^* = 2.9360$. From Table 2 with $k = 3$, $n_0 = 15$ and $P_2^* = .95$, the experimenter finds that $h_2' = 2.9360$, $h_3' = 1.839$. Therefore, $h_2^* = (a-1)h_2' = 2.9360$ and $h_3^* = h_3' = 1.839$. Thus the experimenter finds that $h^* = \max\{h_1^*, h_2^*\} = 2.9360$ (here h_1^* and h_2^* are the same since we choose them to be the same (when $a = 2$) in the calculation of Table 2), and

$$n_i = \max\left\{16, \left[\left(\frac{S_i \times 2.9360}{1 - \frac{1}{2}}\right)^2\right]\right\}. \tag{40}$$

We obtain $n_1 = 21$, $n_2 = 24$, and $n_3 = 59$. Hence 6, 9, and 44 additional observations must be taken from population one, two, and three, respectively. The experimenter also computes $d = 1.839 \times \frac{\frac{1}{2}}{2.9360} = 0.3132$. Therefore, the selection rule is:

select the population associated with $\bar{X}_{[3]}$ if $\bar{X}_{[3]} \geq \bar{X}_{[2]} + .5$,
or
select the populations which satisfy $\bar{X}_{(i)} \geq \bar{X}_{[2]} - .3132$ if $\bar{X}_{[3]} < \bar{X}_{[2]} + .5$. (41)

The Fortran program generates the second samples of appropriate size from populations $N(4, .9^2)$, $N(4.5, 1^2)$, and $N(5.5, 1.5^2)$, respectively. In order to compute the weighted averages, one needs to specify the weights $a_{ij}, i = 1, 2, 3; j = 1, 2, \ldots, n_i$ which would satisfy the conditions (13) and (14). To specify the a_{ij}'s, we first compute:

$$c_i = \frac{(n_i - 1) + \sqrt{(n_i - 1)\left[(n_i - 1) - n_i(1 - \frac{e^2}{S_i^2})\right]}}{(n_i - 1)n_i}. \tag{42}$$

By letting $a_{ij} = c_i, i = 1, 2, 3; j = 1, 2, \ldots, n_i - 1$ and $a_{in_i} = 1 - c_i(n_i - 1), i = 1, 2, 3$, we are guaranteed that the conditions (13) and (14) are satisfied. Our program computes $c_1 = .0499423$, $c_2 = .0426206$, and $c_3 = .0172361$. Therefore, $a_{1j} = .0499423, j = 1, 2, \ldots, 20$, $a_{1,21} = .00154$; $a_{2j} = .0426206, j = 1, 2, \ldots, 23$, $a_{2,24} = .0197262$; $a_{3j} = .0172361, j =$

$1, 2, \ldots, 58$, $a_{3,59} = .0003062$. One can easily check that $S_i^2 \sum_{j=1}^{n_i} a_{ij} = (\frac{1}{2h_2^*})^2 = .0290$, for $i = 1, 2, 3$. The weighted averages are:

$$\bar{X}_1 = 3.95310, \quad \bar{X}_2 = 4.37875, \quad \bar{X}_3 = 5.44820. \tag{43}$$

Since $\bar{X}_{[2]} + .5 = 4.37875 + .5 = 4.87875$ and $\bar{X}_{[3]} = 5.44820 > 4.87875$, the experimenter will select only the population number three and claim that its weight is the largest.

References

1. R. E. Bechhofer. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics*, 25:16–39, 1954.
2. R. E. Bechhofer, C. W. Dunnett, and M. Sobel. A two sample multiple decision procedure for ranking means of normal populations with a common unknown variance. *Biometrika*, 41:170-176, 1954.
3. R. L. Burden, and J. D. Faires. *Numerical Analysis, 4th Edition.* PWS-Kent Publishing Co., 1988.
4. P. Chen and M. Sobel. An integrated formulation for selecting the t best of k normal populations. *Communications in Statistics, Theory and Methods*, 16(1):121-146, 1987.
5. P. Chen and J. Zhang. An integrated formulation for selecting the best normal populations: the common and unknown variance case. *Communications in Statistics, Theory and Methods*, 26(11): 2701- 2724, 1997.
6. E. J. Dudewicz. Non-existence of a single-sample selection procedure whose $P(CS)$ is independent of the variances. *South Africa Statistics Journal*, 5:37–39, 1971.
7. E. J. Dudewicz and S. R. Dalal. Allocation of observations in ranking and selection with unequal variances. *Sankhya*, Series A:28–78, 1975.
8. S. S. Gupta and W. T. Huang. A note on selecting a subset of normal populations with unequal sample sizes. *Sankhya*, Series A:389–396, 1974.
9. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in FORTRAN.* Cambridge University Press, 1992.
10. C. M. Stein. A two sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16:243–258, 1945.