A Note on Selecting a Subset of Normal Populations with Unequal Sample Sizes

by

Shanti S. Gupta and Wen-Tao Huang

Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #302

August 1972

*This research was supported by the Office of Naval Research Contract N00014-67-A-0226-00014 at Purdue University Reproduction in whole or in part is permitted for any purpose of the United States Government.

A Note on Selecting a Subset of Normal Populations with Unequal Sample Sizes*

by

Shanti S. Gupta and Wen-Tao Huang
Purdue University

1. Introduction and Summary

Let $\pi_1, \pi_2, \ldots, \pi_k$ be k normal populations such that π_i has normal cumulative distribution function $\Phi(x; \theta_i, \sigma^2)$ with unknown mean θ_i and common variance σ^2 for i = 1, 2,...,k. Based on an equal number n observations from each population, Gupta [2] gives a subset selection rule which selects a non-empty subset such that the probability of correct selection is at least P*, a preassigned value. For special values of k,n and P*, the constant in the selection rule is tabulated in Gupta [1] for the case when σ^2 is known. When σ^2 is unknown, the constants for the selection rule for special given values of k,n and P* are tabulated in Gupta and Sobel [4].

In this note we study the subset selection problem for the normal means for both known and unknown variance when the sample sizes are not necessarily equal. Tables are given for the constants of the rule for various special values.

*This research was supported by the Office of Naval Research Contract N00014-67-A-0226-00014 at Purdue University Reproduction in whole or in part is permitted for any purpose of the United States Government.

2. Notation and Formulation of the Problem

Let $\theta_{[1]} \leq \theta_{[2]} \leq \cdots \leq \theta_{[k]}$ be the ordered means $\theta_1, \theta_2, \ldots, \theta_k$. Let $\pi_{(i)}$ be the population (unknown) associated with $\theta_{[i]}$ for $i = 1, 2, \ldots, k$. Let $n_1, n_2, \ldots, n_k$ be k given positive integers. Suppose n_i observations are drawn from π_i and let $\bar{x}_{(i)}$ and $n_{(i)}$ denote, respectively, the sample mean and sample size of $\pi_{(i)}$, $i = 1, 2, \ldots, k$. We do not know correct pairings between π_i and $\theta_{[j]}$ or π_i and $n_{(\ell)}$ or $\theta_{[j]}$ and $n_{(\ell)}$ for all i, j and ℓ. Without loss of generality, we assume $n_k = \max(n_1, n_2, \ldots, n_k)$. Let $r_{ij}^2 = n_{(i)}/n_{(j)}$, $s_{ij}^2 = n_i/n_j$ and $t_i^2 = n_i/n_k$, $i, j = 1, 2, \ldots, k$.

Based on $n_1, n_2, \ldots, n_k$ observations from the k populations, we are required to select a non-empty subset of normal populations such that the probability that the subset selected includes $\pi_{(k)}$ is at least P^* ($\frac{1}{k} < P^* < 1$) a preassigned number.

3. The Subset Selection Rule

A. When σ^2 is known

Without loss of generality, we assume $\sigma^2 = 1$. Let x_{ij} be the jth observation from π_i, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$. Let $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$. A subset selection rule is defined by R_1: Select π_i if, and only if, $\bar{x}_i \geq \max_{1 \leq j \leq k} \bar{x}_j - \frac{c}{\sqrt{n_i}}$, $(c > 0)$. In general, the constant c depends on k, $n_1, n_2, \ldots, n_k$ and P^* and is to be computed so that $P(CS|R_1) \geq P^*$ for all possible configurations of $\theta_1, \theta_2, \ldots, \theta_k$. In theorem 1 we discuss the evaluation of c; the constant actually depends only on k, $t_1, t_2, \ldots, t_{k-1}$ and P^*. Before studying the constant c, we need the following lemma.

Let $I(n_i) = \int_{-\infty}^{\infty} \prod_{j \neq i} \Phi(s_{ji}(x+b))d\Phi(x)$ for some constant $b > 0$. Then, the following holds.

Lemma 1. $I(n_k) = \min_{1 \leq i \leq k} I(n_i)$ where $n_k = \max(n_1, n_2, \ldots, n_k)$.

Proof: Let $m = \min\{n_1, n_2, \ldots, n_k\}$. Let u be any real value in $[m, n_k]$. Define $I(u) = \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \Phi(v_i(x+b))d\Phi(x)$ where $v_i^2 = n_i/u$, $i = 1, 2, \ldots, k-1$. It suffices to show that $I(u)$ is monotone decreasing in u. Make a transformation $y = x+b$. We have $I(u) = \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \Phi(v_i y)d\Phi(y-b)$.

Let $\eta_i(y) = \prod_{j \neq i} \Phi(v_j y)$ for $i, j = 1, 2, \ldots, k-1$. Then,

$$\frac{dI(u)}{du} = \sum_{i=1}^{k-1} \int_{-\infty}^{\infty} \eta_i(y) \frac{d}{dv_i} \Phi(v_i y) \frac{dv_i}{du} d\Phi(y-b)$$

since the regularity conditions for differentiation under an integral hold. It suffices to show $\frac{dI(u)}{du} < 0$ for $u \in [m, n_k]$. Let

$$L_i = \int_{-\infty}^{\infty} \eta_i(y) \frac{d}{dv_i} \Phi(v_i y)\, d\Phi(y-b) \text{ and } w_i = \sqrt{n_i}/2\sqrt{u^3}$$

for $i = 1, 2, \ldots, k$. Then, we note that $\frac{dI(u)}{du} = -\sum_{i=1}^{k-1} w_i L_i$. Since $w_i > 0$ for each i, it suffices to show $L_i > 0$ for $i = 1, 2, \ldots, k-1$. Let $\varphi(x)$ denote the standard normal density function, we have then

$$\begin{aligned} L_i &= \left(\int_0^{\infty} + \int_{-\infty}^{0}\right)\eta_i(y) \frac{d}{dv_i} \Phi(v_i y)d\Phi(y-b) \\ &= \int_0^{\infty} \eta_i(y)\varphi(v_i y)y\, d\Phi(y-b) - \int_0^{\infty} \eta_i(-y)\varphi(v_i y)y\, d\Phi(-y-b) \\ &= \int_0^{\infty} \left[\prod_{j \neq i} \Phi(v_j y)\varphi(y-b) - \prod_{j \neq i} (1-\Phi(v_j y))\varphi(y+b)\right] y\varphi(v_i y)dy \\ &> 0 \end{aligned}$$

by noting that $\Phi(v_i y) \geq 1-\Phi(v_i y)$ for $y \geq 0$ and $v_j > 0$ and $\varphi(y-b) > \varphi(y+b)$ for $b > 0$ and $y > 0$.

This completes the proof.

Theorem 1. For given $P^*(\frac{1}{k} < P^* < 1)$, we have

$\inf_{\Omega} P_{\underline{\theta}}(CS|R_1) \geq P^*$ for $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ in parameter space Ω if c satisfies $\int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \Phi(t_i(x+c))d\Phi(x) = P^*$.

Proof: $P(CS|R_1) = P(\bar{x}_{(k)} \geq \max_{1\leq j\leq k-1} \bar{x}_{(j)} - \frac{c}{\sqrt{n_{(k)}}})$

$$= \int_{-\infty}^{\infty} \prod_{j=1}^{k-1} \Phi(r_{jk}(x+c)+\sqrt{n_{(j)}}\ (\theta_{[k]}-\theta_{[j]}))d\Phi(x).$$

Since $\theta_{[k]} - \theta_{[i]} \geq 0$ for $i \neq k$, it follows then

$$\inf_{\Omega} P(CS|R_1) = \int_{-\infty}^{\infty} \prod_{j=1}^{k-1} \Phi(r_{jk}(x+c))d\Phi(x).$$

Since r_{jk} is unknown for each $j = 1, 2, \ldots, k-1$, the problem reduces to finding $\min_{1\leq i\leq k} \int_{-\infty}^{\infty} \prod_{j\neq i} \Phi(s_{ji}(x+c))d\Phi(x)$.

By Lemma 1, we have

$$\min_{1\leq i\leq k} \int_{-\infty}^{\infty} \prod_{j\neq i} \Phi(s_{ji}(x+c))d\Phi(x) = \int_{-\infty}^{\infty} \prod_{j=1}^{k-1} \Phi(t_j(x+c))d\Phi(x).$$

Hence, we have $\inf_{\Omega} P(CS|R_1) \geq \int_{-\infty}^{\infty} \prod_{j=1}^{k-1} \Phi(t_j(x+c))d\Phi(x) = P^*$.

This completes the proof.

A cruder lower bound for the probability of a correct selection has been pointed out by Santner (1972). The result is as follows:

$$\inf_{\Omega} P(CS|R_1) \geq \int_{-\infty}^{0} \Phi^{k-1}(\frac{n_{max}}{n_{min}}u)d\Phi(u-d) + \int_{0}^{\infty} \Phi^{-k-1}(\frac{n_{min}}{n_{max}}u)d\Phi(u-d).$$

The above is easier to compute than the sharper bound given in this paper.

B. When σ^2 is unknown

Let S_ν^2 denote the usual pooled sample variance. Then, it is well-known that $\nu S_\nu^2/\sigma^2$ is chi-square distributed with d.f. $\nu = \sum_{i=1}^{k}(n_i-1)$. We define a selection rule as follows. R_2: Select π_i if, and only if, $\bar{x}_i \geq \max_{1\leq j\leq k} \bar{x}_j - a\, S_\nu/\sqrt{n_i}$, $(a > 0)$.

Let $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k, \sigma^2)$ and let Ω be the parameter space. Let $Q_\nu(y)$ denote the cdf of $\chi_\nu/\sqrt{\nu}$. Then, the following holds.

Theorem 2. For given $P^*(\frac{1}{k} < P^* < 1)$, we have $\inf_{\Omega} P_{\underline{\theta}}(CS|R_2) \geq P^*$ if a satisfies

$$\int_0^\infty \int_{-\infty}^{\infty} \prod_{j=1}^{k-1} \Phi(t_j(x+ay))d\Phi(x)dQ_\nu(y) = P^*.$$

Proof: By similar argument as in Theorem 1, we have

$$P(CS|R_2) = \int_0^\infty \int_{-\infty}^{\infty} \prod_{j=1}^{k-1} \Phi(r_{ik}(x+ay)+\sqrt{n_{(j)}}\,(\theta_{[k]}-\theta_{[j]})/\sigma)d\Phi(x)dQ_\nu(y).$$

Since the right hand side is monotone increasing in $(\theta_{[k]}-\theta_{[j]})$ (≥ 0), hence, we have

$$\inf_{\Omega} P_{\underline{\theta}}(CS|R_2) = \int_0^\infty \int_{-\infty}^{\infty} \prod_{j=1}^{k-1} \Phi(r_{jk}(x+ay))d\Phi(x)dQ_\nu(y).$$

However, r_{jk} is unknown so we reduce the problem to finding

$\min_{1\leq i\leq k} \int_0^\infty \int_{-\infty}^{\infty} \prod_{j\neq i} \Phi(s_{ji}(x+ay))d\Phi(x)dQ_\nu(y)$. For each fixed $y > 0$, it follows from Lemma 1 that $\int_{-\infty}^{\infty} \prod_{j=i} \Phi(s_{ji}(x+ay))d\Phi(x)$ attains its minimum when $n_i = \max(n_1, n_2, \ldots, n_k)$, i.e. $s_{ji} = t_j$. This concludes that

$$\min_{1\le i\le k} \int_0^\infty \int_{-\infty}^\infty \prod_{j=i} \Phi(s_{ji}(x+ay))d\Phi(x)dQ_\nu(y)$$

$$= \int_0^\infty \int_{-\infty}^\infty \prod_{j=1}^{k-1} \Phi(t_j(x+ay))d\Phi(x)dQ_\nu(y)$$

This completes to proof.

4. Computations of c-value and Examples.

(i) Suppose $n_1 = n_2 = \ldots = n_{k-1} = \alpha n_k$, $0 < \alpha < 1$, i.e. $t_j = \alpha$, $j = 1, 2, \ldots, k-1$. For some special values of $\alpha = \frac{1}{\sqrt{7}}, \frac{1}{3}, \frac{1}{\sqrt{3}}, \frac{1}{2}, \frac{1}{\sqrt{2}}, \frac{\sqrt{3}}{\sqrt{7}}; \frac{\sqrt{3}}{\sqrt{5}}, \frac{\sqrt{2}}{\sqrt{3}}$, c-value of Theorem 1 can be obtained from tables of Gupta, Nagel and Panchapakesan [3] through simple calculations. We note that

$$\int_{-\infty}^\infty \prod_{j=1}^{k-1} \Phi(t_j(x+c))d\Phi(x) = \int_{-\infty}^\infty \Phi^{k-1}(\alpha x+\alpha c)d\Phi(x) = \int_{-\infty}^\infty \Phi^{k-1}\left(\frac{x\rho^{\frac{1}{2}}+H}{\sqrt{1-\rho}}\right)d\Phi(x)$$

with $c = \sqrt{1+\alpha^2}\, H/\alpha$ and $\rho = \alpha^2/(1+\alpha^2)$. For given values of k, ρ and P*, the H-value is tabulated in [3] and thus c is obtained. For example, $\alpha = 0.5$, $k = 5$, $P^* = 0.90$, we have $\rho = (0.5)^2/(1 + 0.5^2) = 0.2$ and $H = 1.9167$. Hence, $c = \sqrt{1+0.5^2} \times 1.9167/0.5 = 4.2859$.

For some values of $\alpha = \frac{1}{4}, \frac{2}{3}, \frac{3}{4}$, the c-value is tabulated in Table 1.

(ii) When $k = 2\ell$ and $n_1 = n_2 = \ldots = n_\ell = \alpha n_{\ell+1} = \alpha n_{\ell+2} = \ldots = \alpha n_k$.

$$\int_{-\infty}^\infty \prod_{j=1}^{k-1} \Phi(t_j(x+c))d\Phi(x) = \int_{-\infty}^\infty \Phi^{\ell-1}(x+c)\Phi^{\ell}(\alpha x+\alpha c)d\Phi(x).$$

For special values of $k = 4, 6, 8, 10$ and $\alpha = \frac{1}{2}, \frac{1}{4}, \frac{2}{3}, \frac{3}{4}$, the c-value is tabulated in Table 2.

(iii) When $k = 3\ell$ and $n_1 = n_2 = \ldots = n_\ell = \alpha n_{\ell+1} = \alpha n_{\ell+2} = \ldots = \alpha n_{2\ell} = \beta n_{2\ell+1}$ $= \beta n_{2\ell+2} = \ldots = \beta n_k$ $(\beta < \alpha)$, we have

$$\int_{-\infty}^{\infty} \prod_{j=1}^{k-1} (t_j(x+c))d\Phi(x) = \int_{-\infty}^{\infty} \Phi^{\ell-1}(x+c)\Phi^{\ell}(\beta x+\beta c)\Phi^{\ell}(\frac{\beta}{\alpha} x+ \frac{\beta}{\alpha} c)d\Phi(x).$$

For special values of $k = 3, 6, 9,$ and $\alpha = \frac{1}{2}, \frac{1}{4}$, $\beta = \frac{1}{2}, \frac{3}{4}$, the c-value is tabulated in Table 3.

REFERENCES

[1] Gupta, S. S. (1963). Probability integrals of the multivariate normal and multivariate t. Ann. Math. Statist. 34, 792-828.

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

[3] Gupta, S. S., Nagel, Klaus and Panchapakesan, S. (1972). On the order statistics from equally correlated normal random variables. Mimeo Ser. No. 290, Dept. of Statistics, Purdue University, W. Lafayette, Indiana.

[4] Gupta, S. S. and Sobel, M. (1957). On a statistic which arises in selection and ranking problems. Ann. Math. Statist. 28, 957-967.

[5] Santner, T. J. (1972). Private communication.

TABLE 1

c-value of rule R_1 for special values of k, α and P*

k	P* \ α	$\frac{1}{4}$	$\frac{2}{3}$	$\frac{3}{4}$
2	0.75	2.781	1.217	1.125
	0.90	5.286	2.311	2.136
	0.95	6.777	2.966	2.742
	0.99	9.600	4.201	3.884
3	0.75	4.538	1.909	1.746
	0.90	6.711	2.896	2.666
	0.95	8.042	3.495	3.224
	0.99	10.662	4.647	4.292
5	0.75	6.046	2.506	2.283
	0.90	7.978	3.417	3.138
	0.95	9.198	3.977	3.661
	0.99	11.678	5.068	4.675
10	0.75	7.582	3.115	2.829
	0.90	9.324	3.967	3.634
	0.95	10.456	4.494	4.130
	0.99	12.831	5.536	5.099

The entry is the smallest value c (to 3 decimals of accuracy) satisfying

$$\int_{-\infty}^{\infty} \Phi^{k-1}(\alpha x+\alpha c)d\Phi(x) \geq P^*.$$

TABLE 1

c-value of rule R_1 for special values of k, α and P*

k	P* \ α	$\frac{1}{4}$	$\frac{2}{3}$	$\frac{3}{4}$
2	0.75	2.781	1.217	1.125
	0.90	5.286	2.311	2.136
	0.95	6.777	2.966	2.742
	0.99	9.600	4.201	3.884
3	0.75	4.538	1.909	1.746
	0.90	6.711	2.896	2.666
	0.95	8.042	3.495	3.224
	0.99	10.662	4.647	4.292
5	0.75	6.046	2.506	2.283
	0.90	7.978	3.417	3.138
	0.95	9.198	3.977	3.661
	0.99	11.678	5.068	4.675
10	0.75	7.582	3.115	2.829
	0.90	9.324	3.967	3.634
	0.95	10.456	4.494	4.130
	0.99	12.831	5.536	5.099

The entry is the smallest value c (to 3 decimals of accuracy) satisfying

$$\int_{-\infty}^{\infty} \Phi^{k-1}(\alpha x+\alpha c)d\Phi(x) \geq P^*.$$

TABLE 2

c-value of rule R_1 for special values of k, α and P*

k	P* \ α	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{4}$
4	0.75	4.542	2.524	2.085	1.946
	0.90	6.711	3.652	2.998	2.803
	0.95	8.042	4.365	3.564	3.328
	0.99	10.662	5.776	4.675	4.349
6	0.75	5.447	2.962	2.432	2.268
	0.90	7.468	4.042	3.305	3.088
	0.95	8.728	4.727	3.849	3.593
	0.99	11.260	6.092	4.924	4.581
8	0.75	6.046	3.251	2.651	2.469
	0.90	7.978	4.305	3.505	3.270
	0.95	9.198	4.973	4.037	3.765
	0.99	11.678	6.311	5.094	4.736
10	0.75	6.490	3.466	2.810	2.613
	0.90	8.361	4.502	3.652	3.404
	0.95	9.553	5.159	4.178	3.893
	0.99	11.999	6.478	5.223	4.852

The entry is the smallest value of c (3 decimals of accuracy) satisfying

$$\int_{-\infty}^{\infty} \Phi^{\frac{k}{2}-1}(x+c)\Phi^{\frac{k}{2}}(\alpha x+\alpha c)d\Phi(x) \geq P^*.$$

TABLE 3

c-value of rule R_1 for special values of k, α, β and P*

k	(α,β) / P*	$\frac{1}{4}$ $\frac{1}{2}$	$\frac{1}{4}$ $\frac{3}{4}$	$\frac{1}{2}$ $\frac{3}{4}$
3	0.75	3.417	3.071	2.074
	0.90	5.441	5.301	3.180
	0.95	6.824	6.779	3.879
	0.99	9.604	9.600	5.280
6	0.75	4.783	4.581	2.811
	0.90	6.758	6.712	3.829
	0.95	8.057	8.042	4.484
	0.99	10.663	10.662	5.820
9	0.75	5.572	5.547	3.182
	0.90	7.492	7.468	4.176
	0.95	8.737	8.73	4.816
	0.99	11.262	11.260	6.126

The entry is the smallest value of c (3 decimals of accuracy) satisfying

$$\int_{-\infty}^{\infty} \Phi^{\frac{k}{3}-1}(x+c)\Phi^{\frac{k}{3}}(\alpha x+\alpha c)\Phi^{\frac{k}{3}}(\beta x+\beta c)d\Phi(x) \geq P^*.$$

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Purdue University

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
A Note on Selecting a Subset of Normal Populations with Unequal Sample Sizes

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report, August 1972

5. AUTHOR(S) (Last name, first name, initial)
Gupta, Shanti S. and Huang, Wen-Tao

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
August 1972	10	5

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-A-0226-00014	Mimeo Series #302
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. AVAILABILITY/LIMITATION NOTICES
Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research Washington, D. C.

13. ABSTRACT

Let $\pi_1, \pi_2, \ldots, \pi_k$ be k normal populations with common variance σ^2 and let π_i have mean θ_i. Let $n_1, n_2, \ldots, n_k$ be k given positive integers. Suppose n_i observations are drawn from π_i for each $i = 1, 2, \ldots, k$, we require to select a non-empty subset of populations so that the probability that the subset selected includes a population associated with the largest mean is no less than $P^*(\frac{1}{k} < P^* < 1)$, a preassigned value. Two procedures of subset selection are given for the cases of known and unknown σ^2, respectively. Tables for the constants in the rule are given for some special choices of k, P^* and α, the ratio of n_i over the maximum n_j.

DD FORM 1 JAN 64 1473

Unclassified
Security Classification