韓國數學敎育學會誌<數學敎育>
987. 6. Vol. XXV, No. 3

Journal of the Korea Society of Mathematical Education
Jun. 1987. Vol. XXV, No. 3

A Small-Sample Comparative Study on Subset Selection Procedures

by Moon Sup Song and Sung Suck Chung

Seoul National University, Seoul, Korea

——Dedicated to Professor Han Shick Park on his 60th birthday——

1. Introduction

We often want to select one or more populations out of several populations. For example, we ıay want to choose a teaching method associated with the largest mean among several methods. 'here are basically two approaches to this kind of selection problems, namely, the ***indifference ɔne*** approach and the ***subset selection*** approach. In this paper, we compare the small-sample pro-ɛrties of the subset selection procedures for location problem through a Monte Carlo study.

Consider a set of k independent populations $\pi_1, ..., \pi_k$ with unknown location parameters $\theta_1, ..., \theta_k$, spectively. The ordered parameters are denoted by $\theta_{[1]} \leq \cdots \leq \theta_{[k]}$. The population with the largest ırameter $\theta_{[k]}$ is called the "best" population. In case several populations possess the largest ırameter, one of them is tagged at random and called the best. Here we are interested in selecting nonempty subset of populations containing the best one. Such a selection is called a correct selec-ɔn (CS).

In subset selection procedures it is usually required that for a given rule R the probability of a ɔ is at least a preassigned number P^*, i.e.,

$$\inf_{\Omega} P(\mathrm{CS}\,|\,R) \geq P^*, \tag{1.1}$$

th $P^* \in (1/k,\ 1)$ and $\Omega = \{(\theta_1, ..., \theta_k) : -\infty < \theta_i < \infty, i = 1, ..., k\}$. The probability requirement (1.1) called P^*-condition. The configuration of θ_i's for which the infimum of $P(\mathrm{CS})$ occurs is called a ıst favorable configuration (LFC). It is clear that we prefer procedures which make the size of ɛ selected subset as small as possible subject to the P^*-condition.

The subset selection procedure was introduced by Gupta (1956, 1965). Gupta's procedure is based sample means under the assumption of normality. Gupta and Huang (1974) proposed selection es based on the Hodges-Lehmann (H-L) estimators for selecting the t best populations, assuming ıt the populations have a common "known" variance. Song, Chung and Bae (1982) investigated ɔust selection procedures based on trimmed means and H-L estimators, without the assumption known variances. Gupta and Leong (1979) and Lorenzen and McDonald (1981) considered ɛction procedures based on sample medians.

ɔome nonparametric procedures have also been developed. Bartlett and Govindarajulu (1968) and pta and McDonald (1970) studied nonparametric procedures based on combined ranks. But, in

general, the LFC is not given by the equal parameters configuration in rank procedures. To overcome this difficulty, Hsu (1980) proposed a nonparametric procedure based on pairwise ranks.

We compare these procedures for various underlying distributions in terms of efficiency and robustness through a small-sample simulation study. The results show that the procedures based on trimmed means and pairwise ranks are most successful. The nonparametric procedures based on combined ranks have seriously low efficiencies.

2. The Subset Selection Procedures

Let $X_{i1}, \ldots, X_{in}$ be an independent sample of size n from π_i with a continuous cdf $F(x-\theta_i)$ $i=1, \ldots, k$. We assume that the k populations have a common unknown variance σ^2. In this section we briefly review the subset selection procedures which are included in our simulation study.

The parametric procedure proposed by Gupta (1965) is based on sample means. Let $\bar{X}_i$ be the sample mean from the population π_i and S^2 be the usual pooled sample estimate of σ^2 based on $\nu=k(n-1)$ degrees of freedom. Let the ordered values of the k observed sample means be denoted by $\bar{X}_{[1]} \leq \bar{X}_{[2]} \leq \ldots \leq \bar{X}_{[k]}$. Gupta's procedure R_1 based on sample means is defined by

(2.1) $\quad R_1$: Select π_i if and only if $\bar{X}_i \geq \bar{X}_{[k]} - d_1 S/\sqrt{n}$,

where the constant $d_1=d_1(k, n, P^*)$ is to be chosen so as to satisfy the P^*-condition (1.1). Assuming that π_i is a normal population, it can be shown that the LFG is the equal means configuration (EMC), i.e., the infimum of $P(\mathrm{CS}|R_1)$ occurs when $\theta_1=\cdots=\theta_k$. Thus, the constant d_1 in (2.1) is a solution of the following equation:

$$\int_0^\infty \int_{-\infty}^\infty \Phi^{k-1}(u+d_1 y)\phi(u) g_\nu(y)\, du dy = P^*, \tag{2.2}$$

where Φ and ϕ are the cdf and pdf of standard normal, respectively, and $g_\nu(u)$ is the density $\chi_\nu/\sqrt{\nu}$. The values of d_1 have been tabulated by Gupta and Sobel (1957).

To formulate robust procedures Song, Chung and Bae (1982) proposed subset selection rules based on trimmed means and H-L estimators without the assumption of known variances. The α-trimmed mean as an estimator of θ is defined by

$$\bar{X}_\alpha = \frac{1}{h}\{p(X_{([n\alpha]+1)}+X_{(n-[n\alpha])}) + \textstyle\sum_{i=[n\alpha]+2}^{n-[n\alpha]-1} X_{(i)}\},$$

where $p=1+[n\alpha]-n\alpha$, $h=n-2n\alpha$, and $X_{(1)} < \cdots < X_{(n)}$ are the order statistics. For computational convenience, we assume that $g=n\alpha$ is an integer. To Studentize the trimmed means, Tukey and McLaughlin (1963) suggested the estimator

$$S_\alpha = \{SS(\alpha)/(h(h-1))\}^{\frac{1}{2}}$$

for the standard deviation of $\bar{X}_\alpha$, where $SS(\alpha)$ is the Winsorized sum of squares defined by

$$SS(\alpha) = (g+1)\{(X_{(g+1)}-\bar{X}_\alpha)^2 + (X_{(n-g)}-\bar{X}_\alpha)^2\} + \sum_{i=g+2}^{n-g-1}(X_{(i)}-\bar{X}_\alpha)^2. \tag{2.3}$$

Through a small-sample experiment they showed that a t-distribution with $h-1$ degrees of freedom gives a good approximation to the distribution of $(\bar{X}_\alpha-\theta)/S_\alpha$.

The robust procedure considered by Song et al. is given by

(2.4) $\quad R_2$: Select π_i if and only if $\bar{X}_{i\alpha} \geq \bar{X}_{[k]\alpha} - d_2 S_\alpha/\sqrt{h}$

where $\bar{X}_{i\alpha}$ is the α-trimmed mean associated with the population π_i, $\bar{X}_{[k]\alpha}$ is the largest α-trim

mean, $d_2=d_2(k, n, \alpha, P^*)$ is to be chosen to satisfy the P^*-condition, and $h=n(1-2\alpha)$. $S_\alpha/\sqrt{h}$ is the pooled-sample estimated standard error of the α-trimmed mean, i.e.,

$$S_\alpha=\{SS(\alpha)/(k(h-1))\}^{\frac{1}{2}}$$

with $SS(\alpha)=\sum_{i=1}^{k} SS_i(\alpha)$ and $SS_i(\alpha)$ the Winsorized sum of squares defined by (2.3) for the i-th sample. Here, they intuitively suggested the use of d_1 in (2.1) for d_2 in (2.4).

The H-L estimator of θ, which derived from the Wilcoxon signed-rank test, is given by

$$\hat{\theta}=\underset{i<j}{med}\{(X_i+X_j)/2\}.$$

Hodges and Lehmann (1963) showed that $\sqrt{n}(\hat{\theta}-\theta)$ has a limiting normal distribution with mean 0 and variance $\{12[\int f^2(x)dx]^2\}^{-1}$. In normal case the asymptotic variance of $\sqrt{n}\,\hat{\theta}$ is $\pi\sigma^2/3$.

The selection rule based on H-L estimators is defined by

(2.5) R_3 : Select π_1 if and only if $\hat{\theta}_i \geq \hat{\theta}_{[k]}-d_3\hat{\sigma}/\sqrt{n}$

where $\hat{\theta}_i$ is the H-L estimator of θ_i based on the Wilcoxon signed rank test and $\hat{\theta}_{[k]}$ is the largest $\hat{\theta}_i$'s. Song et al. (1982) used, as an estimator of the common standard deviation σ, the pooled-sample median absolute deviation (MAD) $\hat{\sigma}$ defined by

(2.6) $$\hat{\sigma}=1.48\ \underset{j}{\text{med}}\,|X_{ij}-\underset{j}{\text{med}}\,(X_{ij})|$$

where the median is taken over the $k(n-1)$ largest absolute deviations. The value of d_3 in (2.5) is to be chosen to satisfy the P^*-condition (1.1). But note that under the assumption of normality the asymptotic variance of $\sqrt{n}(\hat{\theta}_i-\theta_i)$ is $\pi\sigma^2/3$, and $\hat{\theta}_i$ is approximately normally distributed, provided the sample size is sufficiently large. Thus they suggested the use of $\sqrt{\pi}/3\,d_1$, where d_1 is defined in (2.1) for d_3 in (2.5).

Subset selection procedures based on sample medians have been considered by Gupta and Leong (1979), Lorenzen and McDonald (1981), and Gupta and Singh (1980), among others. But, in these procedures the common variance σ^2 is assumed to be known. In this paper, without the assumption of known variances, we consider the following selection procedure R_4 based on sample medians.

(2.7) R_4 : Select π_i if and only if $\tilde{X}_i \geq \tilde{X}_{[k]}-d_4\hat{\sigma}/\sqrt{n}$

where $\tilde{X}_i$ is the sample median associated with the population π_i, $\tilde{X}_{[k]}$ is the largest sample median, and $\hat{\sigma}$ is the pooled sample MAD estimater of σ in (2.6). The value of d_4 in (2.7) is to be chosen to satisfy the P^*-condition. Since $\tilde{X}_i$ is asymptotically normal with mean θ_i and variance $\sigma^2/2n$, the constant d_4 satisfying the P^*-condition may be given by $\sqrt{\pi}/2\,d_1$, where d_1 defined in (2.1).

A rank procedure, which is a nonparametric analogue of the Gupta's parametric procedure R_1, was proposed by Bartlett and Govindarajulu (1968). We let R_{ij} denote the rank of X_{ij} in the combined sample, and let R_i be the average rank associated with the i-th population defined by

$$R_i=\frac{1}{n}\sum_{j=1}^{n} R_{ij},\quad i=1, \ldots, k.$$

The procedure suggested by Bartlett and Govindarajulu is defined by

(2.8) R_5 : Select π_i if and only if $R_i \geq R_{[k]}-d_5$,

where $R_{[k]}$ is the maximum of the R_i's and d_5 is to be chosen to satisfy the P*-condition. For the selection rule R_5, the LFC does not occur at the EMC. As a lower bound of the $P(\text{CS}|R_5)$, Gupta

and McDonald (1970) obtained the inequality

$$\inf_{\Omega} P(\mathrm{CS}\,|\,R_5) \geq P(U \leq nd_5), \tag{2.9}$$

where U is the Mann-Whitney statistic associated with sample sizes n and $(k-1)n$. But according to a small-sample Monte Carlo study, which is not reported here, the constant d satisfying $P(U \leq nd) \geq P^*$ is too conservative to be meaningful. We thus in our simulation study used the normal approximation which was stated in Theorem 6.1 of Gupta and McDonald (1970).

To formulate a distribution-free procedure Gupta and McDonald (1970) suggested the following selection rule:

$$R_6\text{: Select } \pi_i \text{ if and only if } R_i \geq d_6, \tag{2.10}$$

where d_6 is to be chosen to satisfy the P^*-condition. The rule R_6 is distribution-free and the LFC occurs at the EMC. The values of d_6 can be obtained from the inequality

$$P\left[U \leq n^2\left(k-\frac{1}{2}\right) - n\left(d_6-\frac{1}{2}\right)\right] \geq P^*$$

The last procedure to be considered in this paper is the pairwise rank procedure proposed by Hsu (1980). To present the Hsu's procedure we introduce some notations. Let $R_{j\beta}^{(i)}$ be the rank of $X_{j\beta}$ among $\{X_{i1}, \ldots, X_{in}; X_{j1}, \ldots, X_{jn}\}$, $i \neq j$, and let $R_j^{(i)}$ be the sum of ranks of $X_{j\beta}$ defined by $R_j^{(i)} = \sum_{\beta=1}^{n} R_{j\beta}^{(i)}$. We also define

$$T_i = \sum_{j=1}^{j} D_{\mathrm{med}}^{(ji)}/k, \quad i = 1, \ldots, k,$$

where $D_{\mathrm{med}}^{(ji)} = \underset{\alpha,\beta}{\mathrm{med}}\{X_{i\alpha} - X_{j\beta}\}$ with $D_{\mathrm{med}}^{(ii)} = 0$. Then the Hsu's procedure can be written as follows:

$$R_7 : \text{Select } \pi_i \text{ if and only if } \max_{j \neq i} R_j^{(i)} < d_7 \text{ and/or } T_i = \max_{1 \leq j \leq k} T_j,$$

where d_7 is the smallest integer such that $P_0\,(\max_{j \neq i} R_j^{(i)} \geq d_7) \leq 1 - P^*$ with P_0 the probability computed at the EMC. The values of d_7 can be obtained in Miller (1966, Table 8) for $P^* = 0.9$ and 0.99. Note that, in rule R_7, the role of "Select π_i if $T_i = \max T_j$" is to ensure a non-empty subset to be selected.

The asymptotic relative efficiencies (ARE) have also been studied for the procedures reviewed in this section. For the procedures based on translation invariant estimators, the ARE is given by the inverse ratio of asymptotic variances of the estimators. (See Theorem 4.3. of Hsu (1980).) For the procedure R_7, Hsu (1980) has shown that the ARE of R_7 relative to R_1 is the same as that of Wilcoxon test to t-test. In the case of two populations (i.e. $k=2$), the procedures R_5 and R_6 are equivalent and the ARE's are the same as that of R_7.

3. Small-Sample Monte Carlo Results

To compare the small-sample properties of the procedures discussed in Section 2, we made some Monte Carlo studies. The underlying distributions considered are normal, double exponential, contaminated normal and Cauchy. Here, the ε-contaminated normal distribuiton has a p.d.f. of the form

$$f(x) = (1-\varepsilon)\phi(x) + \frac{\varepsilon}{\sigma}\phi\left(\frac{x}{\sigma}\right).$$

We use the subroutine GGNML in IMSL(VAX 780) in generating normal samples with a

without contamination. The other samples are generated by using the subroutine GGUBT in IMSL and the inverse integral transformations.

In the simulation study, we consider the equally spaced configuration given by

$$\theta_i=\theta_1+(i-1)\delta\sigma,\quad i=1,\ldots,k,$$

where $\delta>0$ is a given constant and σ is a standard deviation of each population. When the underlying distribution is a Cauchy centered at 0, σ denotes the value such that the probability between $-\sigma$ and σ is the same as that between -1 and 1 for a standard normal distribution. 500 replications were performed for each value of δ ($\delta\sqrt{n}=0.0,\ 0.5,\ 1.0,\ 3.0,\ 5.0$). The constants used in our simulation study are $k=5$, $n=10$, and $P^*=0.90,\ 0.95$. For the contaminated normal samples, $\varepsilon=0.1$ and $\sigma=3,5$ are taken. For the α-trimmed means, $\alpha=0.1$ is used. The constants d_7 for R_7 with $P^*=0.90$ were obtained by simulation using 1000 replications.

When $\delta=0$, the sum of average number of selected populations divided by 500 can be interpreted as the empirical P^*. The values of empirical P^* are tabulated in Table 1. The results show that all procedures except R_3 seem to satisfy the P^*-condition. The rule R_3, which is based on the H-L estimator, does not satisfy the P^*-condition. This may imply that the MAD estimator used in R_3 underestimates the standard error of the H-L estimators.

To compare the efficiencies of selection rules, we use the relative efficiency of the rule R_i to R_1 defined by

$$e(R_i,R_1)=\frac{E(S|R_1)}{E(S|R_i)}\cdot\frac{P(\mathrm{CS}|R_i)}{P(\mathrm{CS}|R_1)},\quad i=2,\ldots,7,$$

where $E(S|R)$ is the expected number of populations to be selected with a given rule R. To estimate the relative efficiencies, empirical relative efficiencies of R_i relative to R_1 are computed from the number of times that each population is selected in the simulation. The results are summarized in Table 2 ($P^*=.90$) and Table 3 ($P^*=.95$).

The results in Table 2 and 3 show that the procedures based on robust estimators are efficient for heavy-tailed distributions. The procedure R_2, which is based on trimmed means, is most successful among the procedures considered in this paper. The high relative efficiencies of R_3 in Table 2 and 3 do not mean that R_3 is most efficient, since it did not satisfy the P^*-condition in Table 1.

Table 1. Empirical P^* Based on 500 Replications ($k=5$, $n=10$)

Rule	Normal		Double Exp.		Contaminated Normal				Cauchy	
					$\sigma=3.0$		$\sigma=5.0$			
	P^*		P^*		P^*		P^*		P^*	
	.90	.95	.90	.95	.90	.95	.90	.95	.90	.95
R_1	.90	.95	.90	.96	.90	.95	.90	.95	.92	.96
R_2	.90	.95	.91	.95	.90	.95	.90	.95	.91	.95
R_3	.86	.94	.84	.92	.84	.93	.84	.93	.80	.87
R_4	.90	.96	.91	.97	.91	.96	.91	.96	.92	.96
R_5	.91	.96	.91	.96	.90	.95	.90	.96	.90	.95
R_6	.90	.95	.90	.95	.90	.95	.91	.96	.90	.95
R_7	.91	.96	.90	.96	.90	.96	.91	.97	.91	.96

Table 2. Empirical Relative Efficiencies with 500 Replications ($P^*=.90$, $k=5$, $n=10$)

Rel. Eff.	$\sqrt{n}\delta$	Normal	Doub. Exp.	Contaminated Normal		Cauchy
				$\sigma=3.0$	$\sigma=5.0$	
$e(R_2, R_1)$	0.5	1.00	1.04	1.02	1.13	1.06
	1.0	.98	1.08	1.06	1.25	1.18
	3.0	1.00	1.08	1.06	1.20	1.69
	5.0	.99	1.04	1.04	1.05	1.97
$e(R_3, R_1)$	0.5	1.02	1.11	1.06	1.16	1.22
	1.0	.99	1.15	1.14	1.30	1.59
	3.0	.98	1.08	1.07	1.21	2.34
	5.0	.97	1.02	1.02	1.05	2.48
$e(R_4, R_1)$	0.5	.93	1.00	.97	1.04	1.07
	1.0	.86	1.00	.95	1.17	1.34
	3.0	.88	1.04	.95	1.13	2.28
	5.0	.90	1.02	1.01	1.05	2.65
$e(R_5, R_1)$	0.5	.99	1.05	1.02	1.12	1.15
	1.0	.95	1.00	1.00	1.12	1.36
	3.0	.72	.70	.72	.70	1.53
	5.0	.53	.55	.55	.54	1.43
$e(R_6, R_1)$	0.5	.95	1.00	.95	1.01	1.11
	1.0	.79	.81	.80	.84	1.15
	3.0	.48	.47	.50	.45	1.09
	5.0	.35	.36	.36	.35	.93
$e(R_7, R_1)$	0.5	.98	1.04	1.02	1.11	1.15
	1.0	.96	1.04	1.04	1.22	1.38
	3.0	.97	1.02	1.02	1.17	1.94
	5.0	.98	.99	1.03	1.04	2.07

Table 3. Empirical Relative Efficiencies with 500 Replications ($P^*=.95$, $k=5$, $n=10$)

Rel. Eff.	$\sqrt{n}\,\delta$	Normal	Doub. Exp.	Contaminated Normal		Cauchy
				$\sigma=3.0$	$\sigma=5.0$	
$e(R_2, R_1)$	0.5	1.01	1.01	1.02	1.09	1.03
	1.0	.99	1.04	1.07	1.20	1.15
	3.0	.99	1.06	1.10	1.25	1.60
	5.0	1.00	1.05	1.07	1.11	1.95
$e(R_3, R_1)$	0.5	1.02	1.09	1.06	1.18	1.20
	1.0	.98	1.15	1.08	1.27	1.50
	3.0	.98	1.11	1.07	1.24	2.29
	5.0	.98	1.04	1.05	1.12	2.57
$e(R_4, R_1)$	0.5	.96	.98	.97	1.04	1.02
	1.0	.87	.96	.95	1.12	1.26
	3.0	.85	.99	.98	1.19	2.24
	5.0	.88	1.03	1.00	1.13	2.64
(R_5, R_1)	0.5	.99	1.02	1.01	1.09	1.08
	1.0	.96	1.01	.99	1.10	1.27
	3.0	.69	.72	.70	.71	1.48
	5.0	.56	.55	.55	.57	1.95
(R_6, R_1)	0.5	.96	.98	.97	1.00	1.06
	1.0	.84	.85	.83	.87	1.13
	3.0	.47	.49	.48	.49	1.09
	5.0	.37	.37	.37	.37	.99
(R_7, R_1)	0.5	.97	.99	1.00	1.07	1.07
	1.0	.93	.99	.99	1.13	1.24
	3.0	.92	.98	.98	1.13	1.77
	5.0	.94	.96	1.00	1.08	2.01

Among the rank procedures, the Hsu's procedure is most successful. The procedures R_5 and R_6, which are based on combined ranks, have poor efficiencies. For large values of δ, the efficiencies of R_5 and R_6 get worse.

References

(1) Bartlett, N.S. and Govindarajulu, Z. (1968), *Some distribution-free statistics and their application to the selection problem*,, Ann. Inst. Statist. Math., 20, 79∼97.

(2) Gupta, S.S. (1956), *On a decision rule for a problem in ranking means*, Tech. Report No. 150, Institute of Statistics, University of North Carolina.

(3) Gupta, S.S. (1965), *On some multiple decision (selection and ranking) rules*, Technometrics, 7, 225∼245.

(4) Gupta, S.S. and Huang, D.Y. (1974), *Nonparametric subset selection procedures for the best populations*, Bull. Inst. Math. Academia Sinica, 2, 377∼386.

(5) Gupta, S.S. and Leong, Y.K. (1979), *Some results on subset selection procedures for doubl exponential populations*, Decision Information (ed. C.P. Tsokos and R.M. Thrall), Academi Press, New York, 277∼305.

(6) Gupta, S.S. and McDonald, G. (1970), *On some classes of selection procedures based o ranks*, Nonparametric Techniques in Statistical Inference (ed. M.L. Puri), Cambridge Univ Press, London, 491∼514.

(7) Gupta, S.S. and Singh, A.K. (1980), *On rules based on sample medians for selection of th largest location parameter*, Commun. Statist. -Theor. Meth. A9, 1277∼1289.

(8) Gupta, S.S. and Sobel, M. (1957), *On a statistic which arises in selection and rankin problems*, Ann. Math. Statist. 28, 857∼867.

(9) Hodges, J.L. Jr. and Lehmann, E.L. (1963), *Estimates of location based on rank tests*, An Math. Statist. 34, 598∼611.

(10) Hsu, J.C. (1980), *Robust and nonparametric subset selection procedures*, Commun. Statis -Theor. Meth. A9, 1439∼1459.

(11) Lorenzen, T.J. and McDonald G.C. (1981), *Selecting logistic population using the samp medians*, Commun. Statist. -Theor. Meth. A9, 101∼124.

(12) Miller, R.G. (1966), *Simultaneous Statistical Inference*, McGraw-Hill, New York.

(13) Song, M.S., Chung, H.Y. and Bae, W.S. (1982), *Subset selection procedures based on so robust estimators*, J. Korean. Statist, Soc. 11. (1982), 109∼117.

(14) Tukey, J.W. and McLaughlin, D.H. (1963), *Less vulnerable confidence and significan procedures for location based on a single sample: trimming/Winsorization* 1, Sankhya S A25, 331∼352.