<Research Paper>

Journal of the Korean
Statistical Society
Vol. 11, No. 2, 1982

Subset Selection Procedures Based on Some Robust Estimators

Moon Sup Song*, Han Yeong Chung* and Wha Soo Bae*

ABSTRACT

In this paper, a preliminary study is performed on the subset selection procedures which are based on the trimmed means and the Hodges-Lehmann estimator derived from the Wilcoxon test. The proposed procedures are compared to the Gupta's rule through a small sample Monte Carlo study. The results show that the procedures based on the robust estimators are successful in terms of efficiency and robustness.

1. Introduction

Consider a set of k independent populations $\pi_1, \pi_2, \cdots, \pi_k$ with unknown location parameters $\theta_1, \theta_2, \cdots, \theta_k$, respectively. The ordered location parameters are denoted by

$$\theta_{[1]} \leq \theta_{[2]} \leq \cdots \leq \theta_{[k]}.$$

The population with the largest location parameter $\theta_{[k]}$ is called the "best" population. Here, we are interested in selecting a nonempty subset of populations containing the best one. Such a selection is called a correct selection(CS).

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

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

where $P^* \in (1/k, 1)$. We thus need the information of the configuration of θ_i's for

*This work was supported in part by the Research Fund of the Ministry of Education, Korean Government, 1982. The authors are grateful to Dr. Woo-Chul Kim for his helpful comments throughout the preparation of this work.

*Department of Computer Science & Statistics, Seoul National University.

which the P^*-condition (1. 1) is satisfied. This configuration is called the least favorable configuration(LFC).

The subset selection rules in terms of the sample means have been developed by Gupta(1956, 1965) and Gupta and Huang(1976), among others. Gupta and Huang (1974) investigated selection rules based on the Hodges-Lehmann(H-L) estimators of location for the problem of selecting the t $(1\leq t<k)$ best populations, assuming that the populations have a common known variance.

Gupta and Leong(1979) considered a subset selection procedure based on the sample medians for double exponential populations. Gupta and Singh(1980) investigated the selection rules based on the sample medians for normal and double exponential populations. Lorenzen and McDonald(1981) also studied the subset selection rules based on the sample medians of logistic populations. All of these procedures are investigated under the assumption of common known variance.

In the problem of subset selection, relatively little work is done in terms of robust procedures. This is partly due to the complexity of the distributions of robust estimators. In this paper we propose the subset selection procedures based on the trimmed means and the H-L estimator derived from the Wilcoxon signed-rank test, using a rather heuristic approach. The trimmed means and the H-L estimator are chosen because of their simplicity in computation and their robustness with respect to the heaviness of distribution tails. A parallel selection rule based on an H-L type estimator in regression problems was considered by Song and Oh(1981).

In Section 2 brief descriptions of the trimmed means and the H-L estimator are given. Section 3 contains the formulation of the subset selection rules and Section 4 deals with a preliminary Monte Carlo study to compare the selection rules. The results show that the heuristic selection rules proposed in this paper are quite robust with respect to the heaviness of distribution tails and worthy to study further.

2. Location Parameter Estimators

Let $X_1, X_2, \cdots, X_n$ be an independent sample of size n from a population with cumulative distribution function(cdf) $F(x-\theta)$ and density function $f(x-\theta)$. We assume that $f(x)$ is continuous and symmetric about 0.

The α-trimmed mean as an estimator of θ is defined by

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

where $p=1+[n\alpha]-n\alpha$, $h=n-2n\alpha$, and $X_{(1)}<X_{(2)}<\cdots<X_{(n)}$ are the order statistics for the random sample (see, for example, Andrews et al.(1972) and Stigler(1977) for the behavior of the trimmed means). For computational convenience, we assume that $g=n\alpha$ is an integer. To Studentize the trimmed means, Tukey and McLaughlin(1963) suggested the estimator

$$S_\alpha'=\sqrt{SS(\alpha)/(h(h-1))}$$

for the standard deviation of $\overline{X}_\alpha$, where $h=n-2g=n-2n\alpha$ and $SS(\alpha)$ is the Winsorized sum of squares given by

$$\begin{aligned}SS(\alpha)=&(g+1)\ (X_{(g+1)}-\overline{X}_\alpha)^2+(X_{(g+2)}-\overline{X}_\alpha)^2+\cdots\\&+(X_{(n-g-1)}-\overline{X}_\alpha)^2+(g+1)(X_{(n-g)}-\overline{X}_\alpha)^2\end{aligned}\qquad(2.1)$$

Through a small sample experiment they showed that a t-distribution with $h-1$, or perhaps slightly less, degrees of freedom gives a good approximation to the distribution of $(\overline{X}_\alpha-\theta)/S_\alpha'$. Huber(1970) also confirmed that the Studentized trimmed mean has both excellent small sample and excellent large sample properties.

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

$$\hat{\theta}=\underset{i\le j}{\text{med}}\left\{\frac{X_i+X_j}{2}\right\}.$$

Hodges and Lehmann(1963) showed that $\sqrt{n}(\hat{\theta}-\theta)$ has a limiting normal distribution with mean 0 and variance

$$h^2(f)=\frac{1}{12[\int f^2(x)dx)]^2}$$

Note that $h^2(f)=\pi\sigma^2/3$ in normal distribution case.

A typical robust estimator of the scale parameter σ is the median absolute deviation (MAD) defined by

$$\hat{\sigma}=1.48\ \underset{i}{\text{med}}|X_i-\underset{j}{\text{med}}(X_j)|,$$

where the value of $\Phi\left(\frac{3}{4}\right)=1.48$ is used to make the estimator consistent under the normal distribution. Thus the asymptotic variance of $\sqrt{n}(\hat{\theta}-\theta)$ can be estimated by $\pi\hat{\sigma}^2/3$.

3. Selection Procedures

We again consider the set of k indepedent populations π_1, π_2, $\cdots$, π_k with *cdf*'s

$F(x-\theta_1)$, $F(x-\theta_2)$, $\cdots$, $F(x-\theta_k)$, respectively. It is assumed that the k populations have a common unknown variance σ^2. Let X_{i1}, X_{i2}, $\cdots$, X_{in} be an independent sample of size n from π_i, $i=1, 2, \cdots, k$. Here we are interested in selecting a subset which contains the "best" population associated with the largest location parameter $\theta_{[k]}$.

Gupta(1956, 1965) has considered the following subset selection rule R based on the sample mean:

R : Select π_i if and only if

$$\overline{X}_i \geq \overline{X}_{[k]} - \frac{dS}{\sqrt{n}}, \tag{3.1}$$

where $\overline{X}_i$ is the sample mean of the ith population, $\overline{X}_{[k]}$ is the largest sample mean, $d=d(k, n, P^*)>0$ is to be determined subject to the P^*-condition (1.1), and S^2 is the usual pooled sample unbiased estimator for the common variance σ^2 with $\nu=k(n-1)$ degrees of freedom. Assuming normality, the constant d is a solution of

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

where Φ and ϕ are the cdf and density function of standard normal, and $q_\nu(y)$ is the density of $\chi_\nu/\sqrt{\nu}$.

The values of d have been tabulated by Gupta and Sobel (1957) for various combinations of k, ν and P^*. Gupta and Huang (1976) has also considered the case of unequal sample sizes.

The rule R in (3.1) is based on the sample mean and the sample variance which are known to be too sensitive to gross errors. We thus want to use some robust estimators for the selection rules.

The first rule we propose is based on the trimmed means. The selection rule is defined by

R_1 : Select π_i if and only if

$$\overline{X}_{i\alpha} \geq \overline{X}_{[k]\alpha} - \frac{d_1 S_\alpha}{\sqrt{h}}, \tag{3.2}$$

where $\overline{X}_{i\alpha}$ is the α-trimmed mean associated with the population π_i, $\overline{X}_{[k]\alpha}$ is the largest α-trimmed mean, $d_1=d_1(k, n, P^*, \alpha)$ is to be chosen to satisfy the P^*-condition (1.1), and $h=n-2g=n-2n\alpha$. $S_\alpha/\sqrt{h}$ is the pooled sample estimated standard error of the α-trimmed mean, i.e.,

$$S_\alpha = \sqrt{SS(\alpha)/(k(h-1))}$$

with

$$SS(\alpha)=\sum_{i=1}^{k} SS_i(\alpha)$$

and $SS_i(\alpha)$ is the Winsorized sum of squares defined by (2.1) for the *ith* sample. Here, we intuitively suggest the use of d in (3.1) for d_1 in (3.2). This intuitive idea may be justified by the t-distribution approximation discussed in Section 2. The results of a small sample Monte Carlo study presented in Section 4 also agree with this conjecture.

The second rule we propose is based on the H-L estimator. We now consider the selection rule

R_2 : Select π_i if and only if

$$\hat{\theta}_i \geq \hat{\theta}_{[k]} - \frac{d_2\hat{\sigma}}{\sqrt{n}}, \tag{3.3}$$

where $\hat{\theta}_i$ is the H-L estimator of θ_i based on the Wilcoxon signed-rank test, $\hat{\theta}_{[k]}$ is the largest of $\hat{\theta}_i$'s, $\hat{\sigma}$ is the pooled sample MAD estimator of σ defined by

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

where the median is taken over the $k(n-1)$ largest median absolute deviations. We consider only the $k(n-1)$ largest absolute deviation to exclude k zero deviations (when n is odd) or k smallest absolute deviations which appear twice (when n is even). A small sample experiment, which is not reported in this paper, supports this adjustment.

The value of d_2 in (3.3) 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 $\pi\sigma^2/3$. We thus suggest the use of $\sqrt{\pi/3}d$, where d is defined in (3.1), for d_2 in (3.3).

Note also that, since the trimmed means and the H-L estimator have the location invariance property, the infimum of the probability of CS for the rules R_1 and R_2 occurs when $\theta_1=\theta_2=\cdots=\theta_k$.

4. An Empirical Study on the Procedures

A small sample Monte Carlo study was performed to compare the selection procedures discussed in Section 3. The procedures considered are the Gupta's rule R based on the normal theory, the rule R_1 based on the trimmed means with $\alpha=1/9$ (denoted by R_1) and $\alpha=2/9$ (denoted by R_1'), and the rule R_2 based on the H-L estimator.

In our simulation study we compared the four rules on the uniform (0, 1), standard normal, double exponential, and Cauchy distributions. The uniform random numbers were generated by using the subroutine RANDU in PDP 11/70. The normal variates were generated by standardizing the sum of twelve random numbers, and the inverse integral transformation was applied to generate double exponential and Cauchy samples.

To investigate the performance of the rules we considered the case when the location parameters are equally spaced, i.e.,

$$\theta_i=\theta_0+(i-1)\delta\sigma,\ i=1,2,\ \cdots,\ k,$$

where $\delta>0$ is a given constant and σ is a standard deviation of each distribution, and $\sigma=2$ is used for Cauchy distribution. The constants used in our simulation study are $k=5$, $n=9$, and $\delta\sqrt{n}=0,2,4$. 500 simulations were performed for each distribution and for each value of δ.

When $\delta\sqrt{n}=0$, the average number of selected population divided by 500 can be interpreted as the empirical P^*, the empirical probability of CS for LFC. These values are given in Table 1.

Table 1. Empirical P^* Based on 500 Replications

distribution	rule	P^* .75	.90	.95	.975	.99
uniform	R	.74	.90	.94	.97	.99
	R_1	.75	.90	.95	.97	.99
	R_1'	.75	.90	.94	.97	.99
	R_2	.78	.91	.95	.97	.98
normal	R	.75	.91	.95	.97	.99
	R_1	.76	.91	.95	.98	.99
	R_1'	.77	.90	.95	.98	.99
	R_2	.75	.89	.94	.98	.99
doulbe exponential	R	.76	.90	.95	.97	.99
	R_1	.76	.90	.95	.97	.99
	R_1'	.76	.91	.95	.97	.98
	R_2	.71	.87	.92	.95	.97
Cauchy	R	.68	.92	.97	.99	1.00
	R_1	.78	.92	.96	.97	.99
	R_1'	.78	.92	.96	.97	.99
	R_2	.65	.81	.87	.91	.94

R : Gupta's rule, R_1 : trimmed mean rule with $\alpha=1/9$
R_1' : trimmed mean rule with $\alpha=2/9$, R_2 : H-L estimator rule

For uniform and normal distributions, the four rules seem to satisfy the P^*-conditions.

In double exponential case the empirical P^* values of R_2 are slightly lower than the required values. For Cauchy distribution, the empirical P^* of R_2 violates the P^*-condition. This may imply that the MAD estimator used in the rule R_2 slightly underestimates the standard deviation of the H-L estimator for long-tailed distributions.

To compare the efficiencies of the selection rules, we use the definition of the relative efficiency suggested by Song and Oh(1981). The relative efficiency of the procedure R' relative to the procedure R is defined by

$$e(R', R)=\frac{E(S|R)}{E(S|R')}\times\frac{P(CS|R')}{P(CS|R)},$$

where $E(S|R)$ is the expected number of populations to be selected with a given rule R. Note that the bounds of the relative efficiency are $1/k\leq e(R', R)\leq k$, where k is the number of populations. Thus, in our simulation study, the upper bound of

Table 2. Empirical Relative Efficiencies Based on 500 Replciations

efficiency	$\delta\sqrt{n}$	P^* .75	.90	.95	.975	.99
				uniform		
$e(R_1, R)$	2	.915	.909	.892	.887	.895
	4	.928	.887	.888	.892	.909
$e(R_1', R)$	2	.864	.828	.810	.800	.785
	4	.854	.819	.820	.821	.815
$e(R_2, R)$	2	.892	.879	.859	.856	.852
	4	.919	.838	.865	.862	.881
				normal		
$e(R_1, R)$	2	.980	.947	.955	.962	.938
	4	.978	.988	.971	.952	.964
$e(R_1', R)$	2	.941	.909	.906	.909	.906
	4	.962	.930	.944	.927	.916
$e(R_2, R)$	2	.978	.963	.966	.975	.979
	4	.985	.970	.976	.967	.983
				double exponential		
$e(R_1, R)$	2	1.050	1.074	1.074	1.091	1.074
	4	1.027	1.043	1.048	1.069	1.050
$e(R_1', R)$	2	1.085	1.095	1.106	1.099	1.075
	4	1.039	1.072	1.080	1.090	1.067
$e(R_2, R)$	2	1.117	1.150	1.154	1.182	1.184
	4	1.035	1.080	1.122	1.131	1.141
				Cauchy		
$e(R_1, R)$	2	1.623	1.549	1.474	1.426	1.372
	4	1.726	1.821	1.804	1.769	1.698
$e(R_1', R)$	2	2.040	1.932	1.839	1.780	1.656
	4	2.094	2.313	2.322	2.280	2.180
$e(R_2, R)$	2	2.142	2.176	2.104	2.053	1.957
	4	2.120	2.427	2.512	2.523	2.588

$e(R', R)$ is 5.

To estimate the relative efficiencies, empirical relative efficiencies of R_1, R_1' and R_2 relative to R are computed from the number of times that each population is selected for the configuration $(\theta_0, \theta_0+\delta\sigma, \cdots, \theta_0+(k-1)\delta\sigma)$ in 500 replications. These relative efficiencies are summarized in Table 2.

The results in Table 2 show that the Gupta's rule R which is based on the normal theory performs slightly better than the proposed rules in normal case. For long-tailed distributions, the proposed rules are significantly better than the Gupta's rule. The number of times that each population is selected, which is not reported in this paper, also shows that the proposed rules are quite robust with respect to the heaviness of distribution tails.

As a conclusion, this preliminary and heuristic study shows that the selection rules based on the robust estimators are successful. The results indicate that it is worthy to study further on this subject in terms of theory and extended simulation experiment.

REFERENCES

(1) Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. (1972) *Robust Estimates of Location*, Princeton: Princeton University Press.

(2) Gupta, S.S. (1956) On a decision rule for a problem in ranking means, *Memeo. Ser.* No. 150, Inst. of Statistics, Univ. of North Carolina, Chapel Hill.

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

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

(6) Gupta, S.S. and Huang, D.Y. (1976) Subset selection procedures for the means and variances of normal populations: unequal sample sizes case, *Sankhyā* Ser. B 38, 112-128.

(7) Gupta, S.S. and Leong. Y.K. (1979) Some results on subset selection procedures for double exponential populations, *Decision Information* (ed. C.P. Tsokos and R.M. Thrall), New York: Academic Press, 277-305.

(8) Gupta, S.S. and Singh, A.K. (1980) On rules based on sample medians for selection of the largest location parameter, *Commun. Statist.-Theor. Meth.* A9, 1277-1298.

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

(10) Huber, P.J. (1970) Studentizing robust estimates, *Nonparametric Techniques in Statistical Inference* (ed. M.L. Puri), New York: Cambridge University Press, 453-463.

(11) Lorenzen, T.J. and McDonald G.C. (1981) Selecting logistic populations using the sample medians, *Commun. Statist.-Theor. Meth.* A10, 101-124.

(12) Song, M.S. and Oh, C.H. (1981) On a robust subset selection procedure for the slopes of regression equations, *J. Korean. Statist. Soc.* 10, 105-121.

(13) Stigler, S.M. (1977) Do rubust estimators work with real data?, *Ann. Statist.* 5, 1055-1089.

(14) Tukey, J.W. and McLaughlin, D.H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: trimming/Winsorization 1, *Sankhyā* Ser. *A* 25, 331-352.