Journal of the KSQC
Vol. 17, No. 1 1989

Selecting Populations Close to a Control Based on Sample Medians

Joong Kweon Sohn*
Min Soo Kang*

ABSTRACT

In this paper we study some procedures selecting all populations close to a control based on sample medians for several double exponential populations. The cases of known and unknown control are considered. Tables needed to use the proposed rules are provided and an illustrative example is also included.

1. Introduction

Selection and ranking problems for k populations have been considered by many authors since the early works of Bechhofer(1954) and Gupta(1956) (see Gupta and Panchapakesan(1979) for further references).

As one of many different settings of the problems, the problem of selecting all populations close to a control is closely related to quality control problems. For example, suppose there are several brands of ball bearings for a shaft in a bicycle. In this situation it is important to insure that the shafts will be capable of assembly at random into a bearing. Hence the diametral clearance, which is the difference between the inside diameter of the bearing and the outside diameter of the shaft, should be within some specification limits. Therefore one may be

* Department of Statistics, Kyungpook National University

interested in choosing some brands which meet the quality specification. Gupta and Singh(1977) have considered this problem based on sample means for normal and gamma populations.

It is well-known that for a symmetric distribution the sample median is an unbiased estimator of the location parameter and is robust in the presence of contaminations from heavy-tailed distributions. Hence selection procedures based on sample medians under the formulation of the subset selection approach have been developed for several distrbutions. Gupta and Leong(1979) have proposed and studied a procedure for selecting the largest of location parameters for the case of double exponential distributions. Gupta and Singh(1980) have investigated the case of normal distributions and Lorenzen and McDonald(1981) have considered the case of logistic distributions.

The double exponential distributions have tails which are heavier than those of normal and logistic distributions but not as heavy as that of a Cauchy distribution. Thus for some applications which primarily concerned with exponential tails, it would seem that the double exponential model would be a useful model. For example, it has been suggested as a model for the distribution of the strength of flaws in materials.(Epstein(1948))

In this paper, we propose and study some selection procedures which contain all populations close to a control based on sample medians for double exponential populations.

In Setion 2, we formulate the problem and propose some procedures for the cases of known and unkonwn control. We also investigate some properties of the proposed procedures.

In Setion 3, we provide an illustrative example. The design constants are computed and tabulated in Table I and II.

2. Framework and the proposed rules R_1 and R_2

In this section we formulate the problem for selecting a subset which contains all populations close to a control and propose selection procedures R_1 and R_2.

2.1. Framework

Let $\pi_0, \pi_1, \cdots, \pi_k$ be $k+1(\geq 2)$ independent double exponential populations with unknown location parameters $\theta_0, \theta_1, \cdots, \theta_k$ and common known variance σ^2, respectively. Since σ^2 is assumed to be known, it is assumed that $\sigma^2=1$ without loss of generality. Here π_0 is a control population and θ_0 may be known or unknown.

Let $\Omega=\{\underline{\theta}=(\theta_0, \theta_1, \cdots, \theta_k) \mid -\infty<\theta_i<\infty, i=0, 1, \cdots, k\}$ be the parameter space, where $\Omega \subseteq R^{k+1}$. Note that for θ_0 known, θ_0 is dropped out from $\underline{\theta}$ and thus $\Omega \subseteq R^k$. It is said that π_i is close to a control π_0 if and only if $|\theta_i-\theta_0| \leq \delta$, where $\delta\,(\geq 0)$ is a contant determined by an experimenter a prior to an experiment. Then our goal is to select a subset including all populations close to a control π_0 with the probabilistic requirement which is so-called the P^*-condition, $i.e.$, $\inf_{\underline{\theta} \in \Omega} P(CS|R) \geq P^*$, $0 < P^* < 1$, where CS stands for a correct selection which includes all populations close to a control π_0.

Let $X_{ij}, j=1, 2, \cdots, n$ be n independent random samples from π_i and the pdf $f(\cdot)$ and the cdf $F(\cdot)$ of X are given by

$$f(x) = \frac{\sqrt{2}}{2} \exp\{-\sqrt{2}|x-\theta|\}, \ -\infty < x < \infty$$

and

$$F(x) = \begin{cases} \frac{1}{2} \exp\{\sqrt{2}(x-\theta)\}, & x < \theta \\ 1 - \frac{1}{2} \exp\{-\sqrt{2}(x-\theta)\}, & x \geq \theta. \end{cases}$$

Let $\tilde{X}_i$ be its sample medians $i=1, 2, \cdots, k$, respectively. For convenience, let $n=2m+1, m \geq 0$ and let $F_i(x) \equiv F(x-\theta_i)$ be a distribution function of $\pi_i, i=1, 2, \cdots, k$, respectively. Then the cdf of $\tilde{X}_i - \theta_i$, denoted by $G_m(x)$, is given by

$$G_m(x) = I_{F_i(x)}(m+1, m+1),$$

where $I_x(\alpha, \beta)$ is an incomplete beta function with parameters α and β.

Therefore the pdf $g_m(\cdot)$ and the cdf $G_m(\cdot)$ of $\tilde{X}_i - \theta_i$ are given by

$$g_m(x) = \frac{a \cdot (2m+1)!}{(m!)^2} [\frac{1}{2} e^{-a|x|}]^{m+1} [1 - \frac{1}{2} e^{-a|x|}]^m, \ |x| < \infty$$

and

$$G_m(x) = \begin{cases} 1 - \sum_{j=0}^{m} \binom{2m+1}{j} \left(\frac{1}{2} e^{ax}\right)^j \left(1 - \frac{1}{2} e^{ax}\right)^{2m+1-j}, & x < 0 \\ 1 - \sum_{j=0}^{m} \binom{2m+1}{j} \left(\frac{1}{2} e^{-ax}\right)^{2m+1-j} \left(1 - \frac{1}{2} e^{-ax}\right)^j, & x \geq 0, \end{cases}$$

respectively, where $a = \sqrt{2}$.

2.2. The proposed rules R_1 and R_2

Now we propose selection rules R_1 for known θ_0 and R_2 for unknown θ_0 as follows.

(A) θ_0 known

First, we consider the case that θ_0 is known. In this case, no samples are needed to be taken from the control population π_0. Thus we propose the rule R_1 as follows :

$\boldsymbol{R_1}$: Select π_i if and only if $|\tilde{X}_i - \theta_0| \leq \delta + d_1$,

where $d_1(\geq 0)$ is chosen to satisfy the P^*-condition. Then the following theorem holds.

Theorem 2.1. For given $P^*(0<P^*<1)$, $\delta>0$ and the proposed rule R_1,

$$\inf_{|\theta_i - \theta_0| \leq \delta} P(CS|R_1) = [G_m(2\delta + d_1) + G_m(d_1) - 1]^k.$$

Proof. Let k_1 be the number of populations satisfying $|\theta_i - \theta_0| \leq \delta$. Hence without loss of generality, $\pi_1, \pi_2, \cdots, \pi_{k_1}$ are assumed to be k_1 populations close to a control π_0. Then

$$\begin{aligned} P(CS|R_1) &= P(\theta_0 - \delta - d_1 \leq \tilde{X}_i \leq \theta_0 + \delta + d_1,\ i = 1, 2, \cdots, k_1) \\ &= \prod_{i=1}^{k_1} P(\theta_0 - \theta_i - \delta - d_1 \leq \tilde{X}_i - \theta_i \leq \theta_0 - \theta_i + \delta + d_1) \\ &= [G_m(\theta_0 - \theta_i + \delta + d_1) - G_m(\theta_0 - \theta_i - \delta - d_1)]^{k_1}. \end{aligned}$$

Now consider a function

$$T(u) = G_m(\theta_0 - u + \delta + d_1) - G_m(\theta_0 - u - \delta - d_1).$$

It is easy to see that the function $T(u)$ is symmetric about θ_0 and is increasing(decreasing) in u for $u<\theta_0(u>\theta_0)$. It follows that

$$\inf_{|u - \theta_0| \leq \delta} T(u) = T(\theta_0 - \delta) = T(\theta_0 + \delta),$$

and thus

$$\inf_{|\theta_i - \theta_0| \leq \delta} P(CS|R_1) = [G_m(2\delta + d_1) + G_m(d_1) - 1]^k.$$

This completes the proof.

From Theorem 2.1., one can easily get the following corollary.

Corollary 2.2. For given $P^*(0<P^*<1)$ and $\delta>0$, the design constant d_1 for the rule R_1 is the solution of the equation

$$G_m(2\delta+d_1)+G_m(d_1)-1=(P^*)^{1/k}.$$

Proof. It directly follows from Theorem 2.1.

The values of d_1 are computed and tabulated in Table I for $k=1(1)6$, $m=1(1)6$, $\delta=0.2$ and $P^*=.75, .90, .95, .99$. The proposed rule R_1 has the following property which is regarded as *monotonicity property*.

Theorem 2.3. For $|\theta_i-\theta_0|<|\theta_j-\theta_0|$,

$$P\{\pi_i \text{ is being selected } |R_1\}\geq P\{\pi_j \text{ is being selected } |R_1\}.$$

Proof. Let $|\theta_i-\theta_0|=c_i$ and $|\theta_j-\theta_0|=c_j$. Then $c_i<c_j$. Let P_i be the probability which π_i is being selected. Then

$$\begin{aligned}P_i-P_j&=[G_m(\theta_0-\theta_i+\delta+d_1)-G_m(\theta_0-\theta_i-\delta-d_1)]\\&\quad-[G_m(\theta_0-\theta_j+\delta+d_1)-G_m(\theta_0-\theta_j-\delta-d_1)]\\&=[G_m(-c_i+\delta+d_1)-G_m(-c_i-\delta-d_1)]\\&\quad-[G_m(-c_j+\delta+d_1)-G_m(-c_j-\delta-d_1)]\end{aligned}$$

By using the notation $T(u_i)$ defined in the proof of Theorem 2.1., one can see that

(i) if c_i, $c_j>0$, then $T(c_i)-T(c_j)\geq 0$,
(ii) if c_i, $c_j<0$, then $T(-c_i)-T(-c_j)\geq 0$,
(iii) if $c_i>0$, $c_j<0$, then $T(c_i)-T(-c_j)\geq 0$,
(iv) if $c_i<0$, $c_j>0$, then $T(-c_i)-T(c_j)\geq 0$.

It follows $P_i\geq P_j$ from (i) to (iv). Thus the proof is complete.

(B) θ_0 unknown

Next, we consider the case that θ_0 is unknown. Since θ_0 is unknown, $2m+1(m\geq 0)$ independent random samples $X_{01}, X_{02}, \cdots, X_{02m+1}$ are taken from the control population π_0 and let $\tilde{X}_0$ be its sample median. Then we propose another rule R_2 as follows :

$\boldsymbol{R_2}$: Select π_i if and only if $|\tilde{X}_i-\tilde{X}_0|\leq\delta+d_2$,

where $d_2(\geq 0)$ is chosen to satisfy the P^*-condition. Now similar to the case of known θ_0, the following theorem and corollary hold.

Theorem 2.4. For given $P^*(0<P^*<1)$, $\delta>0$ and the proposed rule R_2,

$$\inf_{|\theta_i-\theta_0|\leq\delta} P(CS|R_2)=\int_{-\infty}^{0}[G_m(h+d_2)-G_m(h-2\delta-d_2)]^k g_m(h)\,dh$$

$$+\int_{0}^{\infty}[G_m(h+2\delta+d_2)-G_m(h-d_2)]^k g_m(h)\,dh.$$

Proof. Let k_1 be the number of populations satisfying $|\theta_i-\theta_0|\leq\delta$ which is defined in the case A. Then

$$P(CS|R_2)=P(\tilde{X}_0-\delta-d_2\leq\tilde{X}_i\leq\tilde{X}_0+\delta+d_2,\ i=1,2,\cdots,\ k_1)$$

$$=\int_{-\infty}^{\infty}[G_m(h+\theta_0-\theta_i+\delta+d_2)$$

$$-G_m(h+\theta_0-\theta_i-\delta-d_2)]^{k_1}g_m(h)\,dh.$$

Here $T(\theta_i,\ h)$ is defined by

$$T(\theta_i,\ h)=G_m(h+\theta_0-\theta_i+\delta+d_2)-G_m(h+\theta_0-\theta_i-\delta-d_2).$$

Then one can see that, for each fixed $h\in R$,

(i) $T(\theta_i,\ h)$ is continuous function of θ_i,

(ii) $T(\theta_i,\ h)$ is symmetric about $\theta=h+\theta_0$, *i.e.*, $T(h+\theta_0-\theta_i,\ h)=T(h+\theta_0+\theta_i,\ h)$,

(iii) $T(\theta_i,\ h)$ increases (decreases) in θ_i if $\theta_i<h+\theta_0\ (\theta_i>h+\theta_0)$.

Thus for each fixed h, it follows from (i) to (iii),

$$\inf_{|\theta_i-\theta_0|\leq\delta} T(\theta_i,\ h)=\begin{cases}T(\theta_0+\delta,\ h) & \text{if } h<0\\ T(\theta_0-\delta,\ h) & \text{if } h>0.\end{cases}$$

Hence

$$P(CS|R_2)\geq\int_{-\infty}^{0}\prod_{i=1}^{k_1}[G_m(h+d_2)-G_m(h-2\delta-d_2)]g_m(h)\,dh$$

$$+\int_{0}^{\infty}\prod_{i=1}^{k_1}[G_m(h+2\delta+d_2)-G_m(h-d_2)]g_m(h)\,dh.$$

Therefore

$$\inf_{|\theta_i-\theta_0|\leq\delta} P(CS|R_2)=\int_{-\infty}^{0}[G_m(h+d_2)-G_m(h-2\delta-d_2)]^k g_m(h)\,dh$$

$$+\int_0^{\infty}[G_m(h+2\delta+d_2)-G_m(h-d_2)]^k g_m(h)\,dh.$$

Thus the proof is complete.

Corollary 2.5. For given $P^*(0<P^*<1)$, $\delta>0$ and the rule R_2, the design constant d_2 is the solution of the equation

$$\inf_{|\theta_i-\theta_0|\leq\delta} P(CS|R_2)=P^*.$$

Proof. It directly follows from Theorem 2.4.

The values of d_2 are computed and tabulated in Table II for $k=1(1)6$, $m=1(1)6$, $\delta=0.2$ and $P^*=.75, .90, .95, .99$. The Gauss-Laguerre quadrature based on 15 points was used to perform the numerical integration.

Similar to the rule R_1 the proposed rule R_2 has also *monotonicity property* as follows.

Theorem 2.6. For $|\theta_i-\theta_0|<|\theta_j-\theta_0|$,

$$P\{\pi_i \text{ is being selected } |R_2\}\geq P\{\pi_j \text{ is being selected } |R_2\}.$$

Proof. The proof is analogous to that of Theorem 2.3. and hence is being omitted.

Remark : All computations have been carried out by Cyber 170/835 at the Kyungpook National University.

3. An illustrative example

In this section we provide an example for the illustrative purpose with imaginary data used by Gupta and Leong(1979). There are 5 populations $\pi_1, \pi_2, \cdots, \pi_5$ with location parameters θ_i to be 0, 2.5, 3.4, −2.0, −0.65. Here θ_0 is assumed to be known as $\theta_0=1.8$. Also δ is chosen to be $\delta=0.2$. Now one wishes to select all the populations which are close to a control $\theta_0=1.8$. From each population 9 observations were taken as follows :

Then the sample medians of $\pi_1, \pi_2, \cdots, \pi_5$ are $\tilde{X}_1=-0.1761$, $\tilde{X}_2=2.3239$, $\tilde{X}_3=3.2239$, $\tilde{X}_4=-2.1761$, $\tilde{X}_5=-0.8261$, respectively. For $P^*=0.95$, $d_1=0.7676$ from Table I and hence the rule R_1 selects all populations whose medians are in $[0.8324, 2.7676]$. Thus only π_2 is selected. For θ_0 unknown, π_2 is regarded as a control π_0. Hence there are 4 populations π_1, π_3, π_4, π_5.

π_1	π_2	π_3	π_4	π_5
−3.4839	−9.839	−.0839	−5.4839	−4.1339
−2.6762	−.1762	.7238	−4.6762	−3.3262
−.3129	2.1871	3.0871	−2.3127	−.9629
−.2264	2.2736	3.1736	−2.2264	−.8764
−.1761	2.3239	3.2239	−2.1761	−.8261
.1462	2.6462	3.5462	−1.8538	−.5038
.3033	2.8033	3.7033	−1.6967	−.3467
.6160	4.1160	5.0160	−.3840	.9660
5.6924	8.1924	9.0924	3.6924	5.0424

Therefore the sample medians of the control population π_0, π_1, π_3, π_4 and π_5 are $\tilde{X}_0 = 2.3239$, $\tilde{X}_1 = -0.1761$, $\tilde{X}_3 = 3.2239$, $\tilde{X}_4 = -2.1761$, $\tilde{X}_5 = -0.8261$, respectively. For $P^* = 0.95$, $d_2 = 1.07$[illegible]0 from Table II and hence the rule R_2 selects all populations whose medians are in [1.0529, 3.5949]. Thus only π_3 is selected.

Table I. Values of d_1 for the case of the double exponential distribution with unit variance when $\delta = 0.2$ and θ_0 is known.

m	k	P^*	.75	.90	.95	.99
1	4		.9101	1.2696	1.5273	2.1087
	5		.9898	1.3495	1.6071	2.1881
2	4		.6509	.9045	1.0837	1.4829
	5		.7076	.9603	1.1389	1.5370
3	4		.5216	.7231	.8640	1.1744
	5		.5669	.7671	.0972	1.2162
4	4		.4434	.6137	.7316	.9893
	5		.4818	.6506	.7676	1.0238
5	4		.3905	.5397	.6424	.8649
	5		.4243	.5719	.6736	.8945
6	4		.3522	.4860	.5776	.7749
	5		.3826	.5148	.6053	.8010

Table II. Values of d_2 for the case of the double exponential distribution with unit variance when $\delta = 0.2$ and θ_0 is unknown.

m	k	P^*	.75	.90	.95	.99
1	4		1.3300	1.7833	2.0988	2.8096
	5		1.4098	1.8637	2.1793	2.8896
2	4		.9753	1.3180	1.5521	2.0202
	5		1.0332	1.3747	1.6088	2.0818
3	4		.8132	1.0634	1.2519	1.6707
	5		.8599	1.1085	1.2961	1.7149
4	4		.7144	.9260	1.0710	1.4486
	5		.7544	.9643	1.1080	1.4854
5	4		.6390	.8352	.9602	1.2762
	5		.6743	.8691	.9927	1.3070
6	4		.5728	.7634	.8770	1.1297
	5		.6044	.7942	.9065	1.1567

REFERENCES

1. Bechhofer, R.E. (1954), *"A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations With Known Varinces"*, Ann. Math. Statist., 25, 16-39.
2. Epstein, B. (1948), *"Statistical Aspects of Fracture Problems"*, J. Applied Physics, 19, 140-147.
3. Gupta, S.S. (1956), *"On a Decision Rule for a Problem in Ranking Means"*, Mimeo. Ser. No. 150. Inst. of Statis. North Carolina Univ. Chapel Hill, NC.
4. Gupta, S.S., and Singh, A.K. (1977), *"On Selection of Populations Close to a Control or Standard"*, Mimeo. Ser. No. 508. Dept. of Statist. Purdue Univ. West Lafayette, Indiana.
5. Gupta, S.S., and Singh, A.K. (1980), *"On Rules Based on Sample Medians for Selection of the Largest Location Parameter"*, Commu. Statist. -Theor. Meth. A9(12), 1277-1298.
6. 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.
7. Gupta, S.S., and Panchapakesan, S. (1979), *"Multiple Decision Procedures-Theory and Methodology of Selecting and Ranking Populations"*, New York ; John Willey and Sons.
8. Lorenzen, T.J., and McDonald, G.C. (1981), *"Selecting Logistic Populations Using the Sample Medians"*, Commu. Statist. -Theor. Meth. A10(2), 101-124.