ON SOME NONPARAMETRIC SELECTION PROCEDURES

by

Shanti S. Gupta and Sayaji N. Hande
Purdue University

Technical Report #90-26C

Department of Statistics
Purdue University

June 1990

◇

On Some Nonparametric Selection Procedures*

Shanti S. Gupta and Sayaji N. Hande

Department of Statistics, Purdue University

Abstract

In this paper we consider the selection and ranking problem in a nonparametric setup when the populations $\Pi_1, \Pi_2, \ldots, \Pi_k$ are characterized by functionals of the associated distribution functions $\theta(F_1), \theta(F_2), \ldots, \theta(F_k)$, where $\theta(F_i) = \int g_i dF_i$, for $i = 1, 2, \ldots, k$ and $g_1, g_2, \ldots, g_k$ are known bounded functions. The problems of selecting the best population under the indifference zone approach and the subset selection approach are considered. Approximate non-randomized rules are obtained. Finally, some simulation studies concerning these procedures are given.

Key Words: Selection and ranking, nonparametric.

AMS 1985 subject classification: 62G99, 62C20.

*This research was supported in part by the Office of Naval Research Contract N00014-88-K-0170 and NSF Grants DMS-86066964, DMS-8702620 at Purdue University.

1 Introduction

In many practical situations, the experimenter often faces the problem of comparing several competing populations, treatments in clinical trials or processes. The selection and ranking methodology of ranking and selection provides the useful techniques for solving such problems. There have been two main approaches to selection and ranking problems, the indifference zone approach due to Bechhofer (1954) and the subset selection approach due to Gupta (1956). In the indifference zone approach a single population is chosen and is guaranteed to be the best (worst) with probability at least equal to P^*. However, in this formulation it is assumed that the best population is sufficiently apart from the remaining $k-1$ populations. In the subset selection approach no such restriction on the parameter space is assumed. A random size subset of k populations is chosen which is guaranteed to contain the best (worst) population with probability at least equal to P^*. In this approach the data or the outcome of the experiment is used to decide on how many populations to select. For an extensive review of these formulations see Gupta and Panchpakesan (1979) and Gupta and Panchpakesan (1986).

Often in practice, especially for the new treatments, or for expensive products there is not much information (the past data) which could lead us to assume a parametric model. In this paper we consider a ranking and selection problem in a nonparametric setup. Considerable amount of work has been done on the problems of selecting population associated with the largest α-th quantile (or the largest location parameter) or selecting a subset of the populations which contains the population as-

sociated with the largest α-th quantile (or location parameter). Some references are Barlow and Gupta (1969), Gupta and McDonald (1970), Gupta and Huang (1974), Rizvi and Sobel (1967), and Sobel (1967). An extensive review of non-parametric selection and ranking procedures is in Desu and Bristol (1986).

To formulate the problem, let $\Pi_1, \Pi_2, \ldots, \Pi_k$ be the k independent populations. The population Π_i is associated with the cumulative distribution function $F_i(.)$ on R^p, for $i = 1, 2, \ldots, k$. The population Π_i is characterized by the real-valued functional,

$$\theta(F_i) = \int_{R^p} g_i(x) dF_i(x) \ ;$$

where g_i is a known, real-valued bounded function on R^p. In this paper we obtain the "optimal" classical type procedures. Non-randomized procedures are proposed. It is also shown that the proposed non-randomized selection procedures are "close" to the optimal procedures. A lower bound for the probability of a correct selection is also obtained. The non-parametric procedures which are developed in this paper are robust and may also be used to do the preliminary analysis. We believe these procedures would be of use in many selection and ranking problem where the distribution functions associated with the populations do not possess "nice" properties. Some Monte Carlo results are presented in the Section 4.

2 Indifference Zone Approach

In this section we consider the problem of selecting the best (worst) population under the indifference zone approach. The goal is to select the best population with probability at least P^*, provided that the "distance" between the best population and the

remaining $k-1$ populations is at least d, where d is some positive number specified by the experimenter.

As defined before, let $\Pi_1, \Pi_2, \ldots, \Pi_k$ be the k populations. First we consider the problem of selecting the best population among k population when the population Π_i is characterized by the functional $\theta(F_i) = \int g dF_i$, for $i = 1, 2, \ldots, k$ and we are interested in selecting large (small) values of θ. If necessary, we make the transformation

$$g \longrightarrow \frac{g - \inf g}{\sup g - \inf g},$$

and, without any loss of generality assume that $\sup g(x) = 1$ and $\inf g(x) = 0$. Let $\theta_{[1]} \leq \theta_{[2]} \leq \ldots \leq \theta_{[k-1]} \leq \theta_{[k]}$ be the ordered values of $\theta_1, \theta_2, \ldots, \theta_k$. The correct pairing between ordered and unordered θ's is completely unknown. The population corresponding to $\theta_{[k]}$ is called the best population, in case of ties we assume that one of them is tagged to be the best population. Our goal is to select the best population with probability of correct selection at least P^*. We need to define some notations.

Let

$$\mathcal{F} = \{(F_1, F_2, \ldots, F_k) : F_i \text{ is distribution on } R^p \ \}.$$

In general, if we allow F to take any value in $\mathcal{F}$ then there does not exist a procedure which would satisfy the P^* condition, hence we need to restrict the space. Let d be a real number in the interval (0,1) and define, following Bechhofer (1954),

$$\Theta' = \left\{(\theta_1, \theta_2, \ldots, \theta_k) : \theta_{[k]} - \theta_{[k-1]} \geq d\right\}$$

and

$$\mathcal{F}' = \{F : \theta(F) \in \Theta'\}.$$

Correct selection (CS) : Selecting the best population

Goal: For given P^* $(1/k < P^* < 1)$, find a procedure R such that for any n;

$$P_F(CS|R,n) \geq P^* \text{ for every } F \in \mathcal{F}' , \tag{1}$$

where $P_F(CS|R,n)$ denotes the probability of a correct selection for the procedure R. The above condition is called the P^*-condition.

In dealing with the above problem, we need to introduce some notations. Let $p_i = (p_{i1}, p_{i2}, \ldots, p_{in})$; $p = (p_1, p_2, \ldots, p_k)$, where $p_{ij} \geq 0$ for $i = 1, 2, \ldots, k$ and for $j = 1, 2, \ldots, n$. Let Z_{ij} be the independent Bernoulli random variables with parameters p_{ij} for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$, and let $U_1, U_2, \ldots, U_k$ be the k independent uniform random variables on interval $(0, 1/2)$. Let

$$S_i = \sum_{j=1}^{n} Z_{ij} + U_i.$$

Define

$$\psi_i(p) = P(S_i = \max_{1 \leq j \leq k} S_j). \tag{2}$$

Now let $X_{i1}, X_{i2}, \ldots, X_{in}$ be the observable independent random vectors from the population Π_i, for $i = 1, 2, \ldots k$. Let $X = (X_{11}, X_{12}, \ldots\ldots X_{kn})$, and let

$$\tilde{g}(X) = (g(X_{11}), g(X_{12}), \ldots\ldots, g(X_{kn})).$$

Now we propose the following selection procedures.

Procedure R_1:

Select one of the populations $\Pi_1, \Pi_2, \ldots, \Pi_k$ with probabilities $\psi_1(\tilde{g}(X)), \psi_2(\tilde{g}(X)), \ldots, \psi_k(\tilde{g}(X$ respectively.

A natural non-randomized version of this procedure is:

Procedure R_2:

Select the population Π_i for which

$$\psi_i(\tilde{g}(X)) = \max_{1 \le j \le k} \psi_j(\tilde{g}(X)),$$

randomize in case of ties.

Notice that the procedure R_1 is randomized procedure and the procedure R_2 is a non-randomized (randomization for ties considered) version of procedure R_1.

First we prove that the decision rule $\delta(X) = (\psi_1(\tilde{g}(X)), \psi_2(\tilde{g}(X)), \ldots, \psi_k(\tilde{g}(X)))$ is "optimum" decision rule for selecting the best population among k populations.

Theorem 2.1 *:*

The procedure R_1 maximizes the infimum of the probability of a correct selection. i.e. If R' is any other selection procedure then

$$\inf_{F \in \mathcal{F}'} P_F(CS|R') \le \inf_{F \in \mathcal{F}'} P_F(CS|R_1).$$

Proof:

Observe that $\inf g(x) = 0$ and $\sup g(x) = 1$.

Fix $\epsilon > 0$, and get a and b such that $g(a) = \epsilon_1$, $g(b) = 1 - \epsilon_2$ and $0 < \epsilon_1 + \epsilon_2 < \epsilon$.

Let P_i be the counting probability measure induced by a distribution function F_i.

Define

$$\mathcal{F}_0 = \mathcal{F}_{0(\epsilon_1,\epsilon_2)} = \left\{ F : \begin{array}{ll} P_i(\{b\}) = p_i; & P_i(\{a\}) = 1 - p_i \\ 0 \le p_i \le 1; & \text{for } i = 1, 2 \ldots, k \end{array} \right\} \cap \mathcal{F}'. \qquad (3)$$

For $i = 1, 2, \ldots k$, define

$$A_i = \{X_{ij} : \ X_{ij} = b \ \ j = 1, 2, \ldots, n.\} \ \text{ and } \ T_i(X) = |A_i|.$$

Note that for a class of distribution functions $\mathcal{F}_0$, the statistics $T = (T_1, T_2, \ldots, T_k)$ is a complete sufficient statistic.

We also note that $T_1, T_2, \ldots, T_k$ are independent and they have binomial distributions with parameters $(n, p_1), (n, p_2), \ldots, (n, p_k)$, respectively. Since the binomial distribution has the monotone likelyhood ratio property, it is easy to see that for every invariant prior, a rule which selects populations with largest T_i (randomize in case of ties) is a Bayes rule for 0-1 valued loss function. Also notice that the risk function of the procedure R_1 is same as the risk function of the Bayes rule.

Hence

$$\inf_{F \in \mathcal{F}_{0(\epsilon_1, \epsilon_2)}} P_F(CS|R') \leq \inf_{F \in \mathcal{F}_{0(\epsilon_1, \epsilon_2)}} P_F(CS|R_1).$$

Since ϵ is arbitrary, letting $\epsilon \longrightarrow 0$ the result follows.

Remark 2.1 :

From this theorem we see that the procedure R_1 is the "most economical" in the sense that for a given P^* and d there doesn't exist any other procedure which can meet the basic probability requirement with a smaller sample. This was also proved in a special case by Hall (1958)

Theorem 2.2 *:*

[1] $P_F(CS|R_1, n)$ *is increasing in* n *and*

[2] $P_F(CS|R_1, n)$ *is increasing function of* $\theta_{[k]}$ *provided* $\theta_{[1]}, \theta_{[2]}, \ldots, \theta_{[k-1]}$ *held fixed.*

[3] $\inf_{F \in \mathcal{F}'} P_F(CS|R_1, n) \longrightarrow 1$ *as* $n \longrightarrow \infty$.

Proof:

It is straightforward to see that

$$P_F(CS|R_1, n) = P(Y_{kn} + U_1 = \max_{1 \le j \le k}(Y_{jn} + U_j)), \tag{4}$$

where $Y_{1n}, Y_{2n}, \ldots, Y_{kn}$ are independent binomial random variables with parameters $(n, \theta_{[1]}), (n, \theta_{[2]}), \ldots, (n, \theta_{[k]})$ respectively, and $U_1, U_2, \ldots, U_k$ are independent uniform random variables over the interval $(0, 1/2)$. If we consider the problem of selecting the best population among k binomial populations, the procedure which selects the population Π_i for which $Y_{i,n} + U_i = \max_j Y_{j,n} + U_j$ is the best invariant and is a Bayes procedure with respect to every invariant prior on Θ', provided that the underlying loss function is permutation invariant, "monotone" (more loss for selecting bad population) and nonnegative. Hence the Bayes risk of the procedure R_1 decreases as n increases for every permutation invariant prior on Θ'. Thus $P_F(CS|R_1, n)$ is increasing in n. From equation (2) it is clear that $P_F(CS|R_1, n)$ is an increasing function of $\theta_{[k]}$.

The third result is an immediate consequence of the strong law of large numbers. This completes the proof of the theorem.

The above theorem insures that for a given P^*, there exists $n_0(P^*, k, d)$ such that

$$P_F(CS|R_1, n) \ge P^* \text{ for every } F \in \mathcal{F}'.$$

The procedure R_1 has nice properties, however it is a randomized procedure. In practice the experimenter would like to use a non-randomized procedure. The procedure R_2 is a non-randomized version of R_1. The following theorem gives the relationship between $P_F(CS|R_1)$ and $P_F(CS|R_2)$.

Theorem 2.3 *:*

For every $F \in \mathcal{F}$

$$P_F(CS|R_2) \geq 2\, P_F(CS|R_1) - 1. \tag{5}$$

Proof: Let Π_1 be the best population, and I be a indicator function then

$$\begin{aligned}
P(CS|R_2) &= E_F I_{(\psi_1(\tilde{g}(x))=\max_i \psi_i(\tilde{g}(x)))} \\
&\geq \int I_{(\psi_1(\tilde{g}(x))>\max_{j\neq 1} \psi_i(\tilde{g}(x)))} dF \\
&\geq \int \psi_1(\tilde{g}(x)) - \max_{j\neq 1} \psi_i(\tilde{g}(x)) dF \\
&= \int \psi_1(\tilde{g}(x)) dF - \int \max_{j\neq 1} \psi_i(\tilde{g}(x)) dF \\
&= P_F(CS|R_1) - \int \max_{j\neq 1} \psi_i(\tilde{g}(x)) dF \\
&\geq P_F(CS|R_1) - \int \sum_{j\neq 1} \psi_i(\tilde{g}(x)) dF \\
&= P_F(CS|R_1) - \int (1 - \psi_1(\tilde{g}(x))) dF \\
&= P_F(CS|R_1) - 1 + P_F(CS|R_1) \\
&= 2\, P_F(CS|R_1) - 1.
\end{aligned}$$

This proves the theorem.

Remark 2.2 :

From Theorems 2.2 and 2.3 it follows that, for every $F \in \mathcal{F}'$

$$P_F(CS|R_2, n) \longrightarrow 1 \quad \text{as} \quad n \longrightarrow \infty.$$

Remark 2.3 :

Observing the method of the proof of the above theorem, we note that, the above

result holds for any multiple decision problem with 0-1 loss, R_1 is any procedure and the procedure R_2 is a "non-randomized version" of the procedure R_1.

As we noticed before we can generalize the procedures R_1 and R_2 to obtain the procedure for selecting the best population with highest parameter, when the population Π_i is characterized by the functional

$$\Theta(F_i) = \int g_i dF_i;$$

where g_i is a known real-valued function with $\inf_x g_i(x) = 0$ and $\sup_x g_i(x) = 1$, for $i = 1, 2, \ldots, k$. This can be done in the following way.

Define

$$\tilde{\tilde{g}}(x) = (g_1(x_{11}), g_1(x_{12}), \ldots, g_1(x_{1n}), g_2(x_{21}), \ldots, g_2(x_{2n}), \ldots, \ldots g_k(x_{k1}), \ldots, g_k(x_{kn})).$$

Let $\psi_1, \psi_2, \ldots, \psi_k$ as by equation (2).

Procedure R_3 :

Select one of the populations $\Pi_1, \Pi_2, \ldots, \Pi_k$ with probabilities $\psi_1(\tilde{\tilde{g}}(x)), \psi_2(\tilde{\tilde{g}}(x)), \ldots \psi_k(\tilde{\tilde{g}}(x))$, respectively.

A non-randomized version of this procedure is given by

Procedure R_4 :

Select the population Π_i for which

$$\psi_i(\tilde{\tilde{g}}(x)) = \max_{1 \le j \le k} \psi_j(\tilde{\tilde{g}}(x))$$

and randomize in case of ties.

Theorem 2.1 , Theorem 2.2 , Theorem 2.3 and the above remarks hold true for these procedures also.

Theorem 2.3 indicates that the procedure R_2 (R_4) is a "good" approximate non-randomized version of procedure R_1 (R_3), whenever P^* is large, and that is the case in general. For example, if $P_F(CS|R_1) \geq 0.99$ then $P_F(CS|R_2) \geq 0.98$. The procedure is good, in the sense that we lose at most $1 - P^*$ due to non-randomization. We also note that these procedures can be generalized to the problem of selecting the t best populations.

As given by equation (4) the probability of a correct selection can be written in terms of the binomial probabilities. The sample sizes, $n_o(P^*, k, d)$ (exact and approximate) are tabulated by Sobel and Huyett (1957) for $k = 2, 3, 4, 10$, $d = 0.05(0.05)0.5$ and $P^* = 0.5, 0.6, 0.75, 0.90, 0.95, 0.99$. For $k = 2$ they conjectured that the least favorable configuration occurs at $\theta_{[2]} = (1 + d)/2$ and $\theta_{[1]} = (1 - d)/2$. This conjecture is shown to be true by Eaton and Gleser (1989).

3 Subset Selection Approach

In the subset selection approach we select a random size subset of the k populations which contains the best population with probability P^* $(1/k < P^* < 1)$. The main feature of selecting a subset of random size is to allow the size to be determined by the observations themselves. Also in the subset selection approach we need not assume any restriction on the "parameter space".

Now we describe the problem formally, let us assume that there are k populations $\Pi_1, \Pi_2, \ldots, \Pi_k$. The random variable associated with population Π_i has the cumulative distribution function $F_i(.)$ on R^p. Again the characterizing function is real-valued

as defined earlier. Let $\theta_{[1]} \leq \theta_{[2]}, \ldots, \leq \theta_{[k]}$ be the ordered values of $\theta_1, \theta_2, \ldots, \theta_k$. The population associated with $\theta_{[k]}$ is called the best population, in case of ties one of them is tagged as the best population. Our goal is to select a non empty subset of these k populations so that the selected subset includes the population associated with $\theta_{[k]}$ with large probability. Let CS denote the event of correct selection and $P(CS|R)$ denote the probability of correct selection for the procedure R.

CS: Selecting a subset of k populations which contains the best population.

Goal: Find a subsect selection procedure R for which

$$P(CS|R) \geq P^*.$$

Let the decision space $\mathcal{D}$ consists of $2^k - 1$ subsets of the set $\{1, 2, \ldots, k\}$ we write this formally as

$$\mathcal{D} = \{a : a \subset \{1, 2, \ldots, k\} \text{and } |a| \geq 1\}.$$

Action $a = \{i_1, i_2, \ldots, i_r\} \in \mathcal{D}$ corresponds to the selection of the populations $\Pi_{i_1}, \Pi_{i_2}, \ldots, \Pi_{i_r}$. A decision "$a$" is called a correct selection (CS) if the best population is included in the selected subset. We implement the procedures established by Gupta and Sobel (1960) for selecting a subset of k binomial populations containing the best population. To define the procedures we need some notation. Let $p_i = (p_{i1}, p_{i2}, \ldots, p_{in})$, $0 \leq p_{ij} \leq 1$ for $i = 1, 2, \ldots, k$ and for $j = 1, 2, \ldots, n$. Let $p = (p_1, p_2, \ldots, p_k)$.

For every $a \in \mathcal{D}$ define

$$\psi_a(p) = P(\min_{i \in a} S_i \geq \max_{1 \leq j \leq n} S_j - d > \max_{i \notin a} S_i), \tag{6}$$

where $S_i = \sum_{j=1}^{n} Z_{ij}$ for $i = 1, 2, \ldots, k$. For $i = 1, 2, \ldots, k$ and for $j = 1, 2, \ldots, n$ Z_{ij} are independent Bernoulli random variables with $P(Z_{ij} = 1) = p_{ij}$. Let $X_{i1}, X_{i2}, \ldots, X_{in}$ be the observable random vectors form population Π_i for $i = 1, 2, \ldots, k$. Let

$$\tilde{g}(x) = (g(x_{11}), g(x_{12}), \ldots, g(x_{1n}), \ldots, g(x_{k1}), g(x_{k2}), \ldots, g(x_{kn})).$$

Procedure R_s :

Having observed $X = x$, select a subset of populations $\Pi_{i_1}, \Pi_{i_2}, \ldots, \Pi_r$ with probability $\psi_a(\tilde{g}(x))$, where $a = \{i_1, i_2, \ldots, i_r\}$.

Theorem 3.1

$$\inf_{F \in \mathcal{F}} P_F(CS|R_s) = \inf_{0<\theta<1} P_\theta(Y_1 \geq \max_{1 \leq i \leq k} Y_i - d) \tag{7}$$

where $Y_1, Y_2, \ldots, Y_k$ are i. i. d. binomial random variables with parameter (n, θ).

Proof:

Let Π_1 be the best population then

$$\begin{aligned} P_F(CS|R_s) &= E_F \sum_{a:1 \in a} \psi_a(X) \\ &= E_F P(S_1 \geq \max_{1 \leq j \leq k} S_j - d | X = x), \end{aligned}$$

where $S_i = \sum_{j=1}^{n} Z_{ij}$ and for given $X = x$, Z_{ij} 's are independent Bernoulli random variables with $P(Z_{ij} = 1|X = x) = g(x_{ij})$. Hence marginally $S_1, S_2, \ldots, S_k$ are independent binomial random variables with parameters $(n, \theta_1), (n, \theta_2), \ldots, (n, \theta_k)$, respectively.

Hence we have

$$\inf_{F \in \mathcal{F}} P_F(CS|R_s) = \inf_{1 \leq \theta_i \leq \theta_1} P(S_1 \geq \max_{1 \leq j \leq k} S_j - d).$$

From Gupta and Sobel (1960) we know that

$$\inf_{1\leq\theta_i\leq\theta_1} P(S_1 \geq \max_{1\leq j\leq k} S_j - d) = \inf_{\theta_1=\theta_2=\ldots=\theta_k} P(S_1 \geq \max_{1\leq j\leq k} S_j - d).$$

This completes the proof of the theorem.

In the case of $k = 2$, Gupta and Sobel (1960) proved that

$$\inf_{\theta_1=\theta_2=\ldots=\theta_k} P(S_1 \geq \max_{1\leq j\leq k} S_j - d)$$

is attained at $\theta_1 = \theta_2 = 1/2$. For $k > 2$, the common value θ_0 at which infimum takes place is not known. The conservative values of d based on the normal approximation have been tabulated by Gupta and Sobel (1960) for $k = 2(1)20(5)50$, $n = 1(1)20(5)50(10)100(25)200(50)500$ and $P^* = 0.75, 0.90, 0.95, 0.99$. Gupta, Huang and Huang (1976) obtained consurvative values of d when $P^* = 0.75, 0.90, 0.95, 0.99$ and $n = 1(1)4$ when $k = 3(1)15$, and $n = 5(1)10$ when $k = 3(1)5$.

The procedure R_s is randomized, the non-randomized version of this procedure is given by

Procedure R'_s :

After observing $X = x$ select a subset of populations $\Pi_{i_1}, \Pi_{i_2}, \ldots, \Pi_r$ if

$$\psi_a(\tilde{g}(x)) = \max_{a'\in\mathcal{D}} \psi_{a'}(\tilde{g}(x)),$$

where $a = \{i_1, i_2, \ldots, i_r\}$, randomize in case of ties.

As in the previous section we can generalize these subset selection procedures when population Π_i is characterized by the functional $\theta(F_i) = \int g_i dF_i$, for $i = 1, 2, \ldots k$.

As in the indifference zone approach case, we are not been able to get lower bound for the probability of correct selection of procedure R'_s. We feel however that, there

exists a constant $c = c(n)$ and a non trivial subset $\mathcal{F}_0$ of $\mathcal{F}$ such that

$$P_F(CS|R_s') \geq P_F(CS|R_s) - c(1 - P_F(CS|R_s)),$$

$\forall F \in \mathcal{F}_0$.

4 Examples

Let us suppose that there are k populations, $\Pi_1, \Pi_2, \ldots, \Pi_k$, associated with distributions functions, $F_1, F_2, \ldots, F_k$, respectively. In this section we will present some examples and the Monte Carlo results. Standard errors for all the estimates is less than 0.035 .

Example 4.1 :

Let

$$f_i(x) = \frac{1}{2}e^{-|x-\mu_i|} \quad \text{for } i = 1, 2, \ldots, k,$$

where f_i is the density associated with F_i for $i = 1, 2, \ldots, k$. We want to select the population associated with $\mu_{[k]}$. Take g as c.d.f. of double exponential random variable with parameter μ. The problem of selecting the population with the highest location parameter is same as the problem of selecting the population with the highest functional. Now we will use the nonparametric procedures and make comparisons. Let

R_1: The nonparametric rule.

R_2 : Non randomized version of the rule R_1.

R_{median} : Selects the population associated with the highest median.

R_{mean} : Selects the population associated with the highest mean.

$\mu_1 = \mu_2 = \mu_3 = \mu_4 = 0, \ \mu_5 = 1, n = 13,$

$P(CS)$	$\mu = 2$	$\mu = 0.75$	$\mu = 0.50$
R_1	0.4823	0.6815	0.665
R_2	0.690	0.882	0.880
R_{median}	0.887	0.887	0.887

In practice we may not know the configuration; then we estimate $\mu_{[k]}$ and $\mu_{[k-1]}$ by sample medians and set μ = estimate of $\mu_{[k-1]} + 3/4(\mu_{[k]} - \mu_{[k-1]})$. We will take $n = 23$

$$P(CS|R_1) = 0.69$$

$$P(CS|R_2) = 0.89$$

$$P(CS|R_{median}) = 0.93$$

Example 4.2 :

Let

$$F_i(x) = \frac{1}{1 + e^{-}(x - \mu_i)}, \text{ for } i = 1, 2, \ldots, k.$$

$\mu_1 = \mu_2 = \mu_3 = \mu_4 = 0, \ \mu_5 = 1$

g — c.d.f of logistic, $\mu = 0.75$

$$P(CS|R_1) = 0.550$$

$$P(CS|R_2) = 0.806$$

$$P(CS|R_{median}) = 0.748$$

$$P(CS|R_{mean}) = 0.803$$

As in the previous example we will estimate $\mu_{[k]}$ and $\mu_{[k-1]}$ and by sample medians and set μ = estimate of $\mu_{[k-1]} + 3/4(\mu_{[k]} - m_{[k-1]})$. Let $n = 23$

$$P(CS|R_1) = 0.69$$

$$P(CS|R_2) = 0.92$$

$$P(CS|R_{median}) = 0.86$$

$$P(CS|R_{mean}) = 0.91$$

In above examples we observe that $P(CS|R1) < P(CS|R2)$. These results indicate that the nonrandomized version of the nonparametric procedure would be better when the associated distribution functions are stochastically increasing in the parameters.

This can be proved for $k = 2$ and $n = 1$ in the location parameter case. Let $F(.)$ be the associated distribution function, X_1 be the observable random variable from the population Π_1 with location parameter μ_1 and X_2 be the observable random variable from population Π_2 with location parameter μ_2. Let g be the distribution function of X_2. Let Π_1 be the best population.

Then

$$P(CS|R_1) = E[g(X_1)(1 - g(X_2)) + \frac{1}{2}g(X_1)g(X_2) + \frac{1}{2}(1 - g(X_1))(1 - g(X_2))]$$

$$
\begin{aligned}
&= E[g(X_1) - g(X_1)g(X_2) + \frac{1}{2}g(X_1)g(X_2)] \\
&\quad + E[\frac{1}{2} - \frac{1}{2}(g(X_1) + g(X_2)) + \frac{1}{2}g(X_1)g(X_2)] \\
&= \frac{1}{2}E[g(X_1) - g(X_2) + 1].
\end{aligned}
$$

Let Z_1 and Z_2 be the independent random variables with common distribution function $F(.)$. Set $Z = Z_1 - Z_2$. We have the following;

$$
\begin{aligned}
P(CS|R_1) &= \frac{1}{2}[P(Z > -\mu_1 + \mu_2) - P(Z > 0) + 1] \\
&= \frac{1}{2}[P(-\mu_1 + \mu_2 < Z < 0) + 1]
\end{aligned}
$$

Since $P(Z > 0) = P(Z < 0) = 1/2$, we have,

$$
= \frac{1}{2}P(-\mu_1 + \mu_2 < Z < 0) + P(Z > 0).
$$

It is straightforward to see that

$$
P(CS|R_2) = P(Z > -\mu_1 + \mu_2).
$$

Hence

$$
P(CS|R_2) > P(CS|R_1)
$$

if and only if

$$
P(Z > -\mu_1 + \mu_2) > \frac{1}{2}P(-\mu_1 + \mu_2 < Z < 0) + P(Z > 0)
$$

which is true if and only if

$$
P(Z > -\mu_1 + \mu_2) - P(Z > 0) = P(-\mu_1 + \mu_2 < Z < 0) > \frac{1}{2}P(-\mu_1 + \mu_2 < Z < 0),
$$

which is always true.

References

[1] Bahadur, R. R. (1950) *" On a problem in the theory of k populations "*Annals of Mathematical Statistics 2, 362-375.

[2] Barlow, R. E. and Gupta, S. S. (1969) *" Selection procedures for restricted families of probability distribution "* , Annals of Mathematical Statistics 40, 905-917.

[3] Barlow, R. E. Gupta, S. S. and Panchapakesan, S. (1969) *" On the distribution of the maximum and minimum of ratios of order statistics "* Annals of Mathematical Statistics 40, 918-934.

[4] Bechhofer, R. E. (1954) *" A single sample multiple decision procedure for ranking means of normal populations with known variances "* Annals of Mathematical Statistics 25, 16-39.

[5] Desu, M. M. and Bristol, D. R. (1986) *" Rizvi and Sobel type non parametric selection procedures : Review and Extensions "* AJMMS 6, 87-107.

[6] Eaton, M. L. and Gleser L. J. (1989) *"Some results on convolutions and a statistical application"* Contributions to Probability and Statistics, Essays in Honor of Ingram Olkin, (Ed. Gleser et. al.), Springer-Verlag, 75-90.

[7] Gupta, S. S. (1956) *" On a decision rule for a problem of ranking means "* Mimeograph series No. 150 , Institute of statistics, University of North Carolina, Chapel Hill.

[8] Gupta, S. S. and Huang, D. Y. (1974) *" Non parametric subset selection procedures for the 't' best populations "* Bulletin of Mathematics , Academia Sinica 2, 377-386.

[9] Gupta, S. S. and Huang, D. Y., Huang, W. T. (1976) *"On ranking and selection procedures and tests of homogeneity for binomial populations"* Essays in probability and statistics (Eds. S¿ Ikeda et al.), Shinko Tsusho Co. Ltd., Tokyo, Japan.

[10] Gupta, S. S. and McDonald, G. C. (1970) *" On classes of selection procedures based on ranks "* Nonparametric Techniques in statistical Inference , ed M. L. puri , Cambridge University Press, London, 419-514.

[11] Gupta, S. S. and Panchapakesan, S. (1986) *" Subset selection procedures : Review and Assessments "* JAMMS , 5, 235-311.

[12] Gupta, S. S. and Panchapakesan (1979) *" Multiple Decision Procedures "* Wiley, New York.

[13] Hall, W. J. (1958) *"Most economical multiple-decision rules"* Annals of Mathematical Statistics, 29, 1079-1094.

[14] Mosteller, F. (1948) *" A k-sample slippage test for an extreme population "* Annals of Mathematical Statistics , 19, 58-65.

[15] Rizvi, M. H. and Sobel, M. (1967) *" Nonparametric procedures for selecting a subset containing the population with largest α -th quantile "* Annals of Mathematical Statistics , 38, 1788-1803.

[16] Sobel, M. (1967) *" Nonparametric procedures for selecting the 't' populations with the largest α -th quantile "* Annals of Mathematical Statistics, 38, 1804-1816.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release, distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Report #90-26C	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Purdue University 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Department of Statistics West Lafayette, IN 47907	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Office of Naval Research 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-88-K-0170, NSF Grants DMS-86066964, DMS-8702620
8c. ADDRESS (City, State, and ZIP Code) Arlington, VA 22217-5000	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

ON SOME NONPARAMETRIC SELECTION PROCEDURES

12. PERSONAL AUTHOR(S)

Shanti S. Gupta and Sayaji N. Hande

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	June 1990	

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Selection and ranking, nonparametric

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

In this paper we consider the selection and ranking problem in a nonparametric setup when the populations $\Pi_1, \Pi_2, \ldots, \Pi_k$ are characterized by functionals of the associated distribution functions $\theta(F_1), \theta(F_2), \ldots, \theta(F_k)$, where $\theta(F_i) = \int g_i dF_i$, for $i = 1,2,\ldots,k$ and $g_1, g_2, \ldots, g_k$ are known bounded functions. The problems of selecting the best population under the indifference zone approach and the subset selection approach are considered. Approximate non-randomized rules are obtained. Finally, some simulation studies concerning these procedures are given.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Shanti S. Gupta	22b. TELEPHONE (Include Area Code) 317-494-6031 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED