Indian J. pure appl. Math., 13(5): 522–527, May 1982

SELECTION PROCEDURES FOR NONPARAMETRIC FAMILIES OF PROBABILITY DISTRIBUTIONS

GOBIND P. MEHTA

Department of Statistics, Panjab University, Chandigarh

(*Received* 15 *June* 1981; *after revision* 9 *December* 1981)

Let $\pi_1, \ldots, \pi_k$ be k populations with associated distribution functions (d.f.) $F_1, \ldots, F_k$, respectively. We assume that F_i are absolutely continuous and are such that $F_{(1)}(x) \leqslant \ldots \leqslant F_{(k)}(x) \; \forall \; x$, where $((1), \ldots, (k))$ is some unknown permutation of $(1, \ldots, k)$. Let F be a known control d.f. which is assumed to be absolutely continuous and to be such that $F_{(k-t)}(x) < F(x) < F_{(k-t+1)}(x) \; \forall \; x$. Populations associated with $F_{(k-t+1)}, \ldots, F_{(k)}$ are defined to be the t best populations. The problem considered is to select a subset of fixed size s from the k populations which contains atleast c of the t best populations. Let $\Delta(r) = \Delta_{G,H}(r)$, $0 \leqslant r \leqslant 1$, be an appropriately defined shift function, for all absolutely continuous d.f. G and H, $G(x) \leqslant H(x) \; \forall \; x$. Let $\Delta^*(\cdot) = \Delta^*_{F,}(\cdot)$ be such a specified function. Let F^* be the d. f. such that $\Delta_{F,F^*}(\cdot) = \Delta^*(\cdot)$. For a preassigned probability P^*, a procedure R is studied which satisfies the condition $P\{\text{Correct Selection} \mid R\} \geqslant P^*$ whenever $\Delta_{F_{(k-t)},F_{(k-t+1)}}(r) \geqslant \Delta^*(r) \; \forall \; r$. It is seen that the infimum of the probability of correct selection (PCS) occurs when the configuration of the d.f. is $F_{(1)}(x) = \ldots = F_{(k-t)}(x) = F(x)$ and $F_{(k-t+1)}(x) = \ldots = F_{(k)}(x) = F^*(x) \; \forall \; x$. For the case when the statistics which define the selection procedure, depend upon the observations only through the control d.f. the procedure R is proved to be distribution-free in the same sense as are the goodness-of-fit-tests. The infimum of the PCS of the procedure based on the statistics $T_{in} = -2 \sum_{\alpha=1}^{n} \log F(X_{i\alpha})$, $i = 1, \ldots, k$ is obtained explicitly. The procedure in general has many desirable properties.

1. INTRODUCTION

Let $\pi_1, \ldots, \pi_k$ be k populations. Let the distribution function (d.f.) of an observation from the population π_i be denoted by F_i, $i = 1, 2, \ldots, k$. We assume that F_i are absolutely continuous and are such that $F_{(1)}(x) \leqslant \ldots \leqslant F_{(k)}(x) \; \forall \; x$, where $((1), (2), \ldots, (k))$ is an unknown permutation of $(1, 2, \ldots, k)$. Let $\pi_{(i)}$ be the population associated with the d.f. $F_{(i)}$. Let F be a known absolutely continuous d.f. such that $F_{(k-t)}(x) \leqslant F(x) \leqslant F_{(k-t+1)}(x) \; \forall \; x$. $\pi_{(k-t+1)}, \ldots, \pi_{(k)}$ are defined to be the t best populations. The problem considered is that of developing statistical procedures for some nonparametric families of probability distributions to select a subset of fixed size s from the k populations such that it will include at least c of the t best populations. A selection is called correct selection (CS) if the selected subset in fact contains at least c of the t best population. This problem is feasible

and non-trivial, when $\max(s, t) \leqslant k - 1$ and $\max(1, s+t-k+1) \leqslant c \leqslant \min(s, t)$. This formulation of the selection problem is the generalization due to Mahamunulu (1967) of the Bechhofer (1954) approach. He has solved this problem for the parametric situation where the d.f. associated with the k populations are identical except for values of some parameter θ. In his case the populations are ranked according to the values of θ.

Many statisticians have considered selection procedures based on ranks of the observations in the combined order of all the observations. But Rizvi and Woodworth (1970) have shown that unlike the parametric situations, in the case of rank-sum procedures the infimum of the probability of correct selection (PCS) is not easily obtained. The configuration of the parameters at which the infimum of the PCS is obtained is not yet known. The reason for this difficulty seems to be that the rank statistics involved are not independent. Rizvi and Sobel (1967), Sobel (1967), Rizvi, Sobel and Woodworth (1968) and Barlow and Gupta (1969) have developed distribution-free procedures based on the sample quantiles for nonparametric families. Here we take a different approach to propose procedures for which the infimum of the PCS may be obtained when the preference zone is specified in terms of a suitable shift function. It is to be noted that the families of probability distributions considered are nonparametric families.

Section 2 contains our formulation of the selection problem. In section 3 a selection procedure is proposed and the infimum of the PCS over the preference zone is obtained under some assumptions on the statistics to be used. In section 4, it has been shown that our procedure is a distribution-free procedure in the sense of goodness-of-fit tests, i.e., the statistics depend upon the observations through a known d.f. And then the infimum of this PCS is independent of this known d.f. In the last section some general remarks are made.

2. Formulation of the Problem

Let ω denote the vector $(F_{(1)}, \ldots, F_{(k)})'$ and let $\Omega(F) = \{\omega \mid F_{(k-t)}(x) \leqslant F(x) \leqslant F_{(k-t+1)}(x) \ \forall\ x\}$. Let $\Delta(r) = \Delta_{G,H}(r)$, $0 \leqslant r \leqslant 1$, called shift function, be defined for all absolutely continuous d.f. G and H such that $G(x) \leqslant H(x) \ \forall\ x$, as follows :

$$\Delta(r) = \Delta_{G,H}(r) = HG^{-1}(r) - r.$$

The shift function Δ satisfies the following properties:

(i) $\Delta(\cdot)$ is continuous, non-negative and real-valued,

(ii) $\Delta_{G,H}(r) \geqslant \Delta_{G,K}(r) \ \forall\ r$ iff $K(x) \leqslant H(x) \ \forall\ x$,

(iii) $\Delta_{G,H}(r) \geqslant \Delta_{K,H}(r) \ \forall\ r$ iff $K(x) \geqslant G(x) \ \forall\ x$,

(iv) $\Delta_{G,H}(r) = 0 \ \forall\ r$ iff $G(x) = H(x) \ \forall\ x$.

For a given d.f. F let $\Delta^*(\cdot) = \Delta^*_F(\cdot)$ be such a specified shift function and let F^* be the d.f. such that $\Delta_{F,F^*}(r) = \Delta^*(r) \ \forall\ r$. It is seen that $F^*(x) = F(x) + \Delta^*(F(x)) \geqslant F(x) \ \forall\ x$. For a specified shift function Δ^*, let

$$\Omega(F, \Delta^*) = \{\omega \in \Omega(F) \mid F_{(k-t)}(x) \leqslant F(x) \ \forall\ x \text{ and } F^*(x) \leqslant F_{(k-t+1)}(x) \ \forall x\}$$

denote the preference zone.

We wish to take the indifference-zone approach to propose a selection procedure R such that the PCS satisfies the condition

$$P\{CS \mid \omega, R\} \geq P^* \ \forall\ \omega \in \Omega(F, \Delta^*). \qquad \ldots(2.1)$$

Here P^* ($P^* < 1$) is a specified real number and is greater than

$$P(k, t, s, c) = \binom{k}{s}^{-1} \sum_{\alpha=c}^{\min(s,t)} \binom{t}{\alpha} \binom{k-t}{s-\alpha}.$$

3. Proposed Selection Procedure and the Infimum of the PCS

Let $X_{i1}, \ldots, X_{in}$, $i = 1, 2, \ldots, k$ be k independent random samples from the given populations. Let $T_i = T(Xi_1, \ldots, X_{in})$ be a suitable statistic defined for the population π_i, $i = 1, 2, \ldots, k$. Let $T_{(1)} \leq \ldots \leq T_{(k)}$ denote the ordered T_i's. The proposed selection procedure is based on these statistics.

Procedure R—Select the set of s populations corresponding to $T_{(k-s+1)}, \ldots, T_{(k)}$. Once the common sample size n is specified the selection procedure R is completely defined. In remark 5.3 a sufficient condition for the existence of the smallest common sample size n is given such that the requirement (2.1) is met. We shall now determine the configuration of the d.f. at which the PCS attains its infimum over the perference zone.

Let Y_i be the statistic based on the sample from the population $\pi_{(i)}$, $i = 1, 2, \ldots, k$. Thus the set $(Y_1, \ldots, Y_k)$ is some (unknown) permutation of the set $(T_1, \ldots, T_k)$ and $Y_{k-t+1}, \ldots, Y_k$ are the statistics associated with the t best populations. Let G_i be the d.f. of the statistic T_i. Also let $G_{(i)}$ be the d.f. of the statistic Y_i if the statistic Y_i is based on the random sample from the population with d.f. $F_{(i)}$ anp $G^*_{(i)}$ if it is based on the random sample from the population with d.f. $F^*_{(i)}$. We make the following two assumptions about the statistics Y_i.

Assumption I—The statistics Y_i are absolutely continuous random variables.

Assumption II—If $F_{(i)}(x) \leq F^*_{(i)}(x)\ \forall\ x$, then $G_{(i)}(x) \geq G^*_{(i)}(x)\ \forall\ x$.

It is easy to see that

$$P\{CS \mid \omega, R\} = P\{c^{th} \text{ largest of } (Y_{k-t+1}, \ldots, Y_k) > (s-c+1)^{st} \text{ largest of } (Y_1, \ldots, Y_{k-t}) \mid \omega, R\} \qquad \ldots(3.1)$$

where $Y_1, \ldots, Y_k$ are independent statistics with d.f. $G_{(1)}, \ldots, G_{(k)}$ respectively. The following theorem proves a monotonicity property of the PCS.

Theorem 3.1—Let G and H be two d.f. For a fixed integer i between 1 and $k-t$ let $G(x) \geq F_{(i)}(x)\ \forall x$ and for a fixed integer j between $k-t+1$ and k let $H(x) \leq F_{(j)}(x)\ \forall\ x$. If the selection procedure R is based on the statistics Y_i, which satisfy assumptions I and II, then

$$P\{CS \mid (F_{(1)}, \ldots, F_{(i)}, \ldots, F_{(k)})\} \geq P\{CS \mid (F_{(1)}, \ldots, F_{(i-1)}, G, F_{(i+1)}, \ldots, F_{(k)})\}$$

and $P\{CS \mid (F_{(1)}, \ldots, F_{(j)}, \ldots, F_{(k)})\} \geq P\{CS \mid (F_{(1)}, \ldots, F_{(j-1)}, H, F_{(j+1)}, \ldots, F_{(k)})\}$.

The proof of the theorem is straight forward and hence omitted.

Infimum of the PCS over the preference zone—From the above theorem it follows that, for any control d.f. F and for any specified shift function $\Delta^*(\,.\,) = \Delta^*_F(\,.\,)$, the PCS over the preference zone attains its infimum when the configuration of the d.f. is

$$F_{(1)}(x) = \ldots = F_{(k-t)}(x) = F(x) \;\forall\; x \qquad \ldots(3.2)$$

$$F_{(k-t+1)}(x) = \ldots = F_{(k)}(x) = F^*(x) \;\forall\; x.$$

This configuration may be called the least favourable configuration in this context. Let $Q(F, \Delta^*, n)$ denote the PCS at the configuration (3.2). Thus

$$\inf_{\omega \in \Omega(F, \Delta^*)} P\{\text{CS} \mid \omega, R\} = Q(F, \Delta^*, n).$$

Under the configuration (3.2), $Y_1, Y_2, \ldots, Y_{k-t}$ are *i. i. d.*, say, with d.f. G and $Y_{k-t+1}, \ldots, Y_k$ are *i.i.d.*, say with d.f. G^*. We have the following theorem.

Theorem 3.2—The infimum of the PCS over the preference zone is given by

$$Q(F, \Delta^*, n) = \frac{t!}{(t-c)!(c-1)!} \sum_{\alpha=0}^{s-c} \binom{k-t}{\alpha} \int_{-\infty}^{\infty} [G(x)]^{k-t-\alpha} [1-G(x)]^{\alpha}$$

$$[G^*(x)]^{t-c} [1-G^*(x)]^{c-1} \, d\, G^*(x).$$

The proof is again simple and hence omitted

4. Infimum of the PCS when the Statistics depend upon the Observations only through the Control d.f. F

In general the infimum of the PCS, $Q(F, \Delta^*, n)$, will depend upon the control d.f. F but we shall see below that with a proper choice of the statistics T_i and the shift function Δ^* as a function of F, the infimum of the PCS becomes independent of F. Since the family of d.f. satisfying (3.2) is a nonparametric family, the selection procedure R in that case may be said to be a distribution-free procedure.

Let the statistic T_i depend upon the observations $X_{i1}, \ldots, X_{in}$ only through the control d.f. F, i.e.,

$T_i = T(F(x_{i1}), \ldots, F(X_{in}))$, $i = 1, 2, \ldots, k$. Denote by

$X_{(i)1}, \ldots, X_{(i)n}$ the sample from the population $\pi_{(i)}$, $i = 1, 2, \ldots, k$.

So we have $Y_i = T(F(X_{(i)1}), \ldots, F(X_{(i)n}))$, $i = 1, 2, \ldots, k$. In order to obtain $Q(F, \Delta^*, n)$ we need to obtain the distribution of Y_i when $F_{(i)}(x) = F(x) \;\forall\; x$ and also when $F_{(i)}(x) = F^*(x) \;\forall\; x$. Let these distributions be denoted by G and G^* respectively. Obviously, if $X_{(i)}(x) = F(x) \;\forall\; x$, the distribution of $F(X_{(i)j})$, $j = 1, 2, \ldots, n$, is uniform and hence the distribution of Y_i would not depend upon F. If the distribution of $X_{(i)j}$ is F^*, i.e., $F_{(i)}(x) = F^*(x) \;\forall\; x$, then from the definition of F^* it follows that

$$P\{F(X_{(i)j}) \leqslant r\} = F^* F^{-1}(r) = r + \Delta^*(r), \; 0 \leqslant r \leqslant 1.$$

So again the distribution of $F(X_{(i)j})$, $j = 1, 2, \ldots, n$., and hence that of Y_i does not depend upon F and depends only upon the shift function Δ^*.

Since G and G^* are independent of F so is $Q(F, \Delta^*, n)$.

Example 4.1—Let us take $Y_i = -2 \sum_{\alpha=1}^{n} \log F(X_{(i)\alpha})$,

$i = 1, 2, \ldots, k$. For a given real number d^* ($d^* \geqslant 1$) let Δ^* be satisfied as

$$\Delta^*(r) = r^{(1/d^*)} - r, \quad 0 \leqslant r \leqslant 1. \qquad \ldots(4.1)$$

For any absolutely continuous d.f. F and for Δ^* satisfied by (4.1), it is seen that under the configuration (3,2), the d.f. of Y_i is

$$G(x) = \int_0^x [(n-1)!]^{-1} \left(\tfrac{1}{2}\right)^n e^{-y/2} y^{n-1}\, dy$$

for $i = 1, 2, \ldots, k-t$ and

$$G^*(x) = \int_0^x ((n-1)!)^{-1} (1/2d^*)^n e^{-(1/2d^*)y} y^{n-1}\, dy$$

for $i = k-t+1, \ldots, k$. Since G and G^* are independent of F so is $Q(F, \Delta^* n)$. Moreover, $Q(F, \Delta^*, n)$ now depends upon Δ^* only through d^*.

5. Remarks

Remark 5.1 : It might be thought that the requirement that a known control distribution is available for comparing the k distributions is too restrictive. However in practice this situation is often observed. For example, consider a disease for which no drug was available in the past. But a large data regarding the duration of illness due to this disease is available. So we may assume that the distribution F of the duration of illness, without the application of any drug, is known. Suppose that a number of researchers now claim to have developed drugs that can cure this disease. We would naturally like to choose those drugs by which the duration of illness (random) is reduced, without knowing their distributions. Such drugs may be said to be those for which the distribution of the duration of illness, say K, is such that $K(x) \geqslant F(x)\ \forall\ x$.

Remark 5.2 : In the preceding section we may also consider the selection procedure R based on the Kolmogorov's statistic $T_i = D_{in}^+ = \sup_x (F_{in}(x) - F(x))$, where F_{in} is the empirical d.f. corresponding to F_i. The infimum of the PCS in this case also would not depend upon the control d.f. F. Since the distribution of the statistics D_{in}^+ when the sample is from a distribution different from F, is not known in a closed form, the infimum of the PCS cannot be obtained easily.

Remark 5.3 : In the special case considered in section 4, it should be noted that although the knowledge of the d.f. F is necessary to obtain the values of the statistics and to carry out the selection procedure, yet the selection procedure is a distribution-free procedure in the sense that the infimum of the PCS does not depend upon F. Hence in this case the procedure R is distribution-free in the same sense as are the tests for goodness-of-fit

Remark 5.4 : The required common sample size is the smallest value of n for which $Q(F, \Delta^*, n) \geqslant P^*$ (5.1)

where $Q(F, \Delta^*, n)$ is defined by (3.3). The required sample size exists provided the left hand side of (5.1) tends to unity as $n \to \infty$. Using arguments of Mahamunulu

(1967) we can easily see that a sufficient condition for the existence of the required sample size is that $\lim_{n\to\infty} (\text{var}(Y_k) - \text{var}(Y_1))/(E(Y_k) - E(Y_1))^2 = 0$.

The statistics of Example 4.1 are seen to satisfy this condition.

Remark 5.5: Gupta and Nagel (1971) and Santner (1975) have given some desirable properties of a selection procedure, viz., unbiasedness, monotonicity, strong monotonicity and justness. These properties have been defined for procedures dealing with parametric families. It is seen that with the necessary modifications in the definitions these properties also hold for our procedure.

Acknowledgement

The author would like to thank Dr Jayant V. Deshpande for suggesting the above approach to the problem, and for all his subsequent help. Thanks are also due to the referee for his useful comments which led to the revised version of this paper.

References

Barlow, R. E., and Gupta, S. S. (1969). Selection procedures for restricted families of distributions. *Ann. Math. Statist.*, 40, 905–17.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.*, 25, 16–39.

Gupta, S. S., and Nagel, K. (1971). On some contributions to multiple decision theory. In *Statistical Decision Theory and Related Topics* (ed. S. S. Gupta and J. Yackel). Academic Press, New York, pp. 79–102.

Mahamunulu, D. M. (1967). Some fixed-sample ranking and selection problems. *Ann. Math. Statist.*, 38, 1079–91.

Rizvi, M. H., and Sobel, M. (1967). Nonparametric procedures for selecting a subset containing the population with the largest α-quantile. *Ann. Math. Statist.*, 38, 1788–1803.

Rizvi, M. H., Sobel M., and Woodworth, G. G. (1968). Nonparametric ranking procedures for comparison with a control. *Ann. Math. Statist.*, 39, 2075–93.

Rizvi, M. H., and Woodworth, G. G. (1970). On selection procedures based on ranks: counter examples concerning least favourable configuration. *Ann. Math. Statist.*, 41, 1942–51.

Santner, T. J. (1975). A restricted subset selection approach to ranking and selection problems. *Ann. Statist.*, 3, 334–49.

Sobel, M. (1967). Nonparametric procedures for selecting the t populations with the largest α-quantile. *Ann. Math Statist.*, 38, 1804–16.