CHAPTER-III

A CLASS OF SELECTION PROCEDURES BASED ON TWO-SAMPLE U-STATISTICS USING MAXIMA AND MINIMA OF SUB-SAMPLES

In the preceding chapter we have proposed a class of subset-selection procedures for selecting a subset containing the 'best population' from a set of given $k$ populations, when the populations differ only in their location parameters. There the proposed selection procedures were based on a statistic which makes use of only right extrema (i.e. maxima) of the sub-samples. However, selection procedures based on a statistic which takes into account upper as well as lower extrema (i.e. maxima and minima) of the sub-samples is expected to perform better. With this motivation, in this chapter we propose another class of subset-selection procedures for location parameters based on both maxima and minima of the sub-samples. The rank representation of the two-sample U-statistic used to propose the subset selection procedures in this chapter is given and some desirable properties of these procedures are also discussed. A modification of these procedures for selecting the populations better than an unknown control population is presented. A comparison of both the classes of subset-selection procedures will be made in chapter-IV. The contents of this chapter along with its ARE's and simulation
results given in chapter-IV have been accepted for publication (see Kumar, N. et al).

3.1 : STATEMENT OF THE PROBLEM

Let \( \Pi_1, \ldots, \Pi_k \) be \( k \) independent populations and \( F_i(x) = F(x - \theta_i) \) be the absolutely continuous cumulative distribution function (cdf) of the \( i \)th population indexed by the location parameter \( \theta_i \), \( i = 1, \ldots, k \). Let \( \Omega = (\theta, \theta = (\theta_1, \ldots, \theta_k)^t, -\infty < \theta_i < \infty, i = 1, \ldots, k) \) denote the parametric space. For any two populations \( \Pi_i \) and \( \Pi_j \), \( \Pi_j \) is considered to be better than \( \Pi_i \) if \( \theta_j \geq \theta_i \). Thus the population with the largest \( \theta \) is labelled as the best. We assume that there is a unique best population. If more than one populations are tied for the best then arbitrarily one of them is chosen to be the best. We first consider the goal of selecting a subset of \( k \) populations which contains the population corresponding to largest location parameter. Any such selection will be called a correct selection (CS). The problem is to find a rule \( R \) such that for a pre-assigned probability \( P^* \), this satisfies the \( P^* \)-condition, that is,

\[
P_\theta[\text{CS}\mid R] \geq P^*, \quad \ldots (3.1.1)
\]

regardless of the true unknown value of the population location parameters.
3.2 : PROPOSED SELECTION PROCEDURES

The selection procedures proposed here are based on two-sample U-statistics. These two-sample U-statistics make use of more information from sample data as compared to the two-sample U-statistics of second chapter. Fix \( i \) in the discussion below and define

\[
H(x) = F_i(x) = F(x - \theta_i),
\]

where \( H(.) \) is any continuous distribution function. Now

\[
F_j(x) = F(x - \theta_j) = H(x - \Delta_{ij}),
\]

where

\[
\Delta_{ij} = \theta_j - \theta_i, \text{ } i \neq j, \text{ } i,j = 1,...,k.
\]

Let \( n_i \) be the number of observations taken from the \( i \)th population, \( i=1,...,k \) and let \( \bar{z}=(n_1,...,n_k)^{t} \). For each \( i \), let \( \bar{z}_i=(x_{i1},...,x_{in_i})^{t} \) be the vector of observations from the \( i \)th population and let \( \bar{z}=(x_1,...,x_k)^{t} \) be the vector of observations from all the populations. For a subset selection procedure \( R \), let \( Z_R(\bar{z},g) \) be the probability assigned to \( g \) by \( R \) having observed \( \bar{z} \). Consider two fixed integers \( r_1 \) and \( r_2 \) belonging to the set \( \{1,2,...,\min(n_1,...,n_k)\} \). For \( i,j=1,...,k(i \neq j) \), define the two-sample statistics as

\[
\phi_{ij}(x_{i1},...,x_{ir_1};x_{j1},...,x_{jr_2}),...(3.2.1)
\]

where
The statistic $V_{ij}$ proposed by Deshpande and Kochar (1982) for the two-sample location problem is obviously a U-statistic (Hoeffding, 1948) corresponding to the kernel (3.2.2). Clearly $V_{ij}$ is the Mann-Whitney statistic computed for the $i^{th}$ and $j^{th}$ samples. It can be seen that the expected value of $V_{ij}$, under $\theta_i = \theta_j$ is unity for all $i \neq j$.

Define

$$S_{ij} = V_{ij} - 1$$

For the rank representation of the statistic $V_{ij}$, define the following kernels.
(I)

\[ \phi_{ij}(X_{i\alpha_1}, \ldots, X_{i\alpha_r}; X_{j\beta_1}, \ldots, X_{j\beta_r}) \]

\[ = \begin{cases} 
1 & \text{if } \max(X_{i\alpha_1}, \ldots, X_{i\alpha_r}) \leq \max(X_{j\beta_1}, \ldots, X_{j\beta_r}) \\
0 & \text{otherwise,} 
\end{cases} \]

(2)

\[ \phi_{ij}(X_{i\alpha_1}, \ldots, X_{i\alpha_r}; X_{j\beta_1}, \ldots, X_{j\beta_r}) \]

\[ = \begin{cases} 
1 & \text{if } \min(X_{i\alpha_1}, \ldots, X_{i\alpha_r}) \leq \min(X_{j\beta_1}, \ldots, X_{j\beta_r}) \\
0 & \text{otherwise,} 
\end{cases} \]

so that

\[ \phi_{ij}(X_{i\alpha_1}, \ldots, X_{i\alpha_r}; X_{j\beta_1}, \ldots, X_{j\beta_r}) \]

\[ = \phi_{ij}(X_{i\alpha_1}, \ldots, X_{i\alpha_r}; X_{j\beta_1}, \ldots, X_{j\beta_r}) \\
+ \phi_{ij}(X_{i\alpha_1}, \ldots, X_{i\alpha_r}; X_{j\beta_1}, \ldots, X_{j\beta_r}) \\
+ \phi_{ij}(X_{i\alpha_1}, \ldots, X_{i\alpha_r}; X_{j\beta_1}, \ldots, X_{j\beta_r}) \\
+ \phi_{ij}(X_{i\alpha_1}, \ldots, X_{i\alpha_r}; X_{j\beta_1}, \ldots, X_{j\beta_r}) \]

and hence

\[ \binom{r_1}{r_2} \sum_{r_1}^{r_2} \phi_{ij}(X_{i\alpha_1}, \ldots, X_{i\alpha_r}; X_{j\beta_1}, \ldots, X_{j\beta_r}) \]

\[ = \binom{n_i}{r_1} \binom{n_j}{r_2}^{-1} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \phi_{ij}(X_{i\alpha_1}, \ldots, X_{i\alpha_r}; X_{j\beta_1}, \ldots, X_{j\beta_r}) \]

\[ \cdots (3.2.3) \]

Let \( X_{i(s)}(X_{j(t)}) \) be the \( s^{th} (t^{th}) \) order statistic of the \( X_i(X_j) \) sample and \( R_{i(s)}(R_{j(t)}) \) be its rank in the combined increasing arrangement of \( X_i \) observations and \( X_j \) observations. Then by using the arguments of Remark 2.3.1, for the kernels (1) (2) \( \phi_{ij} \) and \( \phi_{ij} \), we have
Our class of selection procedures is based on the statistics $\mathbf{r}_1', \mathbf{r}_2'$ and is defined as follows:

$$V_{ij} = \left[\begin{array}{c} n_i \\ n_j \end{array} \right]^{-1} \left( \sum_{t=1}^{n_j} \left( \begin{array}{c} t-1 \\ r_1 \end{array} \right) \left( \begin{array}{c} R_j(t) - t \\ r_1 \end{array} \right) + \sum_{s=1}^{n_i} \left( \begin{array}{c} n_i - s \\ r_2 \end{array} \right) \mathbf{r}_1(s) \mathbf{r}_2 \right).$$

For fixed values of $r_1'$ and $r_2'$ the procedure $R_2(r_1', r_2')$ is a member of the proposed class, where $r_1'$ and $r_2'$ take values in the set $\{1, 2, \ldots, \min(n_1, \ldots, n_k)\}$.

### 3.3: Probability of a Correct Selection and Expected Subset Size

For any $\mathbf{r} = (\mathbf{r}_1', \ldots, \mathbf{r}_k')$, let $\mathbf{r}^{[k]}$ be the unique component of $\mathbf{r}$ which corresponds to the best population $\Pi^{(k)}$. Let $G$ be the set of all nonempty subsets of the set $\{1, \ldots, k\}$. We call $G$ as the action space of the subset selection problem where an action $g \in G$ means selecting those populations whose indices
are in \( g \). For any \( g \in G \), let
\[
CS(\theta, g) = \begin{cases} 
1 & \text{if } \theta[k] \in [\theta_1, \theta_2] \subseteq g \\
0 & \text{otherwise}
\end{cases}
\]
and \( |g| \) = cardinality of the set \( g \).

In order to obtain the probability of a correct selection, define
\[
\tilde{G} = \{ g \in G | CS(\theta, g) = 1 \}
\]
The probability of a correct selection is
\[
P_{\theta}[CS|R_2(r_1, r_2)] = P_{\theta}[\prod (k) \text{ is in the selected subset}|R_2(r_1, r_2)]
\]
\[
= P_{\theta}\left[ \bigcup_{g \in G} \text{ (X is observed and action g is taken}|R_2(r_1, r_2) \right]
\]
\[
= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} P_{\theta}\left[ \bigcup_{g \in G} \text{ (action g is taken}|X=x, R_2(r_1, r_2) \right] \prod_{i=1}^{k} dF_i(X_{i\alpha}).
\]
Using the same argument as in section 2.4 the above integral is equal to
\[
E_{\theta}\left[ \sum_{g \in G} P_{\theta}(\text{action g is taken}|X, R_2(r_1, r_2)) \right]
\]
\[
= E_{\theta}\left[ \sum_{g \in G} CS(\theta, g) Z_{R_2(r_1, r_2)}(X, g) \right],
\]
where \( Z_{R_2(r_1, r_2)}(X, g) \) is the probability assigned to action \( g \in G \) by subset selection procedure \( R_2(r_1, r_2) \) having observed \( X \). The expected subset size is
In the following section we shall show that the selection procedures proposed in section 3.3 meet the \( P^* \)-condition (3.1.1).

3.4 : \( P^* \)-CONDITION FOR PROCEDURE \( R_2(\mathbf{r}_1, \mathbf{r}_2) \)

In order to find the infimum of the PCS and to show that the \( P^* \)-condition holds, we first show that the family of distributions of \( S_{ij}(\mathbf{r}_1, \mathbf{r}_2) \) is stochastically increasing (SI). Let

\[
\theta_i < \theta_j < \theta_{j'}.
\]

Since \( F_i(x) = F(x-\theta_i), \ i = 1, \ldots, k, \)

\[
\theta_i < \theta_j < \theta_{j'},
\]

\[
F_i(x) \geq F_j(x) \geq F_{j'}(x), \ \forall \ x.
\]

Suppose that \( G(x, \Delta_{ij}) = P[S_{ij}(\mathbf{r}_1, \mathbf{r}_2) \leq x] \) be the cumulative distribution function (cdf) of \( S_{ij}(\mathbf{r}_1, \mathbf{r}_2) \), where \( \Delta_{ij} = \theta_{j'} - \theta_i \). Now \( \theta_i < \theta_j < \theta_{j'} \) if and only if \( \Delta_{ij} < \Delta_{ij'} \) and we want to show that

\[
(r_1, r_2) \text{ st } (r_1, r_2) \leq (r_1, r_2), \text{ that is, } G(x; \Delta_{ij}) \geq G(x; \Delta_{ij'}).
\]

Consider the event

\[
\max(X_{i_1}, \ldots, X_{i_{r_1}}) \leq \max(X_{j_1}, \ldots, X_{j_{r_2}})
\]

\[
= \max(X_{i_1}, \ldots, X_{i_{r_1}}) \leq \max(F_j^{-1}F_i(X_{i_1}), \ldots, F_j^{-1}F_i(X_{i_{r_2}}))
\]
\[ \max(x_{i_1}, \ldots, x_{i_1}) \leq \max(F_{i_1}^{-1}(x_{i_1}), \ldots, F_{i_1}^{-1}(x_{i_1})) \]

\[ \text{st} \]
\[ \max(x_{i_1}, \ldots, x_{i_1}) \leq \max(x_{j_1}, \ldots, x_{j_1}). \]

Thus

\[ (1) \]
\[ x_{i_1} \leq x_{j_1} \]

Similarly

\[ \min(x_{i_1}, \ldots, x_{i_1}) \leq \min(x_{j_1}, \ldots, x_{j_1}) \]

\[ \text{st} \]
\[ \min(x_{i_1}, \ldots, x_{i_1}) \leq \min(F_{i_1}^{-1}(x_{i_1}), \ldots, F_{i_1}^{-1}(x_{i_1})) \]

Thus

\[ (2) \]
\[ x_{i_1} \leq x_{j_1} \]

and hence in view of

(3.2.3), we have

\[ (r_{1}, r_{2}) \leq (r_{1}, r_{2}) \]

or

\[ (r_{1}, r_{2}) \leq (r_{1}, r_{2}) \]
Theorem 3.4.1: The selection procedures $R_2(r_1, r_2)$ satisfy the $P^*$-condition for every choice of $r_1$ and $r_2$.

Proof: Assume without loss of generality that $\Pi_j$ is the best population. Now $\min_{i \neq j} S_{ij}$ is nondecreasing in $X_{j1}, \ldots, X_{jn_j}$ and nonincreasing in other components of $X$. Furthermore the family of distributions of $S_{ij}$ is $S_1$. Define the indicator function $I_2(.)$ as

$$I_2(S_{ij}, \ldots, S_{j-1,j}, S_{j+1,j}, \ldots, S_{kj}) = \begin{cases} 1 & \text{if } S_{ij} \geq d_{ij}(n, P^*, r_1, r_2) \forall i, i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

By writing the expectation of indicator function $I_2(.)$ as the expectation of conditional expectation and then using lemma 2.5.2, we have for any $\theta \in \Omega$.

$$P^* \leq P_0 [S_{ij} \geq d_{ij}(n, P^*, r_1, r_2) \forall i, i \neq j]$$

$$\leq P_\theta [S_{ij} \geq d_{ij}(n, P^*, r_1, r_2) \forall i, i \neq j]$$

$$= P_\theta [CS|R_2(r_1, r_2)].$$

This proves that the selection procedures $R_2(r_1, r_2)$ satisfy the $P^*$-condition.

Remark 3.4.1: Using Theorem (2.5.2) it can be easily verified that selection procedures $R_2(r_1, r_2)$ are strongly monotone, and
Remark 3.4.2: The selection procedures \( R_2(r_1, r_2) \) can also be modified to select a subset of the \( k \) populations better than the unknown control population as follows:

Let \( \Pi_0 \) be the control population and \( \Pi_1, \ldots, \Pi_k \) denote the treatment populations. Let the population \( \Pi_i \) has continuous distribution function \( F_i(x) = F(x - \Theta_i) \), \( i = 0, 1, \ldots, k \). The population \( \Pi_0 \) with cdf \( F_0(x) = F(x - \Theta_0) \) is the control population with unknown location parameter \( \Theta_0 \). A population \( \Pi_i \) is said to better than \( \Pi_0 \) if \( \Theta_i \geq \Theta_0 \), \( i = 1, \ldots, k \). The goal is to select a subset of \( k \) populations \( \Pi_1, \ldots, \Pi_k \) which are better than the control population \( \Pi_0 \). Let \( n_i \) be the number of observations taken from the \( i \)th population, \( i = 0, 1, \ldots, k \) and let \( n^* = (n_0, n_1, \ldots, n_k)^t \). For each \( i \), let \( X_i = (X_{i1}, \ldots, X_{in_i})^t \) be vector of observations from population \( \Pi_i \), \( i = 0, 1, \ldots, k \). The proposed subset selection procedure is

\[
R_{C2}(r_1, r_2) : \text{For any } i \text{ between } 1 \text{ and } k \text{ include the population } \Pi_i \text{ in the subset if and only if}
\]

\[
S_{i0} \geq d_{10}(n^*, P^*, r_1, r_2).
\]

Here the constants \( d_{10}(n^*, P^*, r_1, r_2) \) are chosen such that for a pre-assigned probability \( P^*(2^{-k} < P < 1) \), we have

\[
P_0[S_{i0} \geq d_{10}(n^*, P^*, r_1, r_2) \forall i = 1, \ldots, k] \geq P^*.
\]
where $P_0$ indicates that the probability is computed under the parametric configuration $\theta_0 = \theta_1 = \ldots = \theta_k$. 