CHAPTER-II

SELECTION PROCEDURES BASED ON TWO-SAMPLE U-STATISTICS USING MAXIMA OF SUB-SAMPLES

2.1: INTRODUCTION

Selection procedures for ordering k populations according to their location parameters have been proposed and studied extensively in literature. The subset selection procedures proposed by Gupta (1965) for ranking the given k populations in terms of their location parameters are based on sample means

\[ \bar{X}_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij}, \quad i=1, \ldots, k, \tag{2.1.1} \]

where \( X_{11}, \ldots, X_{1n} \) is a random sample of size \( n \) from population \( \Pi_1, \quad i=1, \ldots, n. \)

Gupta and Huang (1976), while restricting to normal populations, proposed selection procedures for location parameters using sample means (2.1.1), in case of unequal sample sizes. Lehmann (1963b) proposed a class of nonparametric selection procedures based on joint ranking of all the observations for arbitrary location parameter models. He employed the indifference-zone approach with the selection decision based on ranks, and claimed that the slippage configuration was least favourable. Puri and Puri (1968, 1969),
Bartlett and Govindarajulu (1968) among others proposed nonparametric selection procedures based on joint ranking under both the indifference zone and subset selection approaches. However, in a paper Rizvi and Woodworth (1970) gave counter examples to show that the procedures based on joint ranking may not control PCS over the least favourable configuration of parametric space (i.e. equal parameter in subset selection and slippage configuration under indifference zone). Later, Hsu (1980a) used pairwise ranking to propose subset-selection procedures, based on two-sample linear rank statistics, for selecting a subset containing the best population associated with the largest location parameter and proved that these procedures control the PCS over the entire parametric space.

In this chapter we consider k populations which differ only in their location parameters. A class of subset selection procedures has been proposed, using two-sample U-statistics, to select a subset of the given k populations which includes the population with the largest location parameter. The rank representation of these two-sample U-statistics, under pairwise ranking, is also given. Some of the important properties of the proposed procedures have been established. A modification of the proposed selection procedures for selecting a subset of the given k populations better than the unknown control population is discussed.
2.2: 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 \( \theta_j > \theta_i \). 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 the largest location parameter. Any such selection will be called a correct selection (CS). The problem is to find a selection procedure \( R \) such that for a pre-assigned probability \( P^* \), this satisfies the probability requirement

\[
P_\theta[CS|R] \geq P^*, \quad \ldots (2.2.1)
\]

regardless of the true unknown value of the population location parameters. The probability requirement (2.2.1) is also called \( P^* \)-condition.
The selection procedures proposed here are based on two-sample $U$-statistics. The cdf of the $i^{\text{th}}$ population is

$$F_i(x) = F(x - \theta_i) = H(x), \text{ (say)}$$

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

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

where

$$\Delta_{ij} = \theta_j - \theta_i, \ i \neq j, i, j = 1, \ldots, k.$$ 

Let $X_i = (x_{i1}, \ldots, x_{in_i})^t$ be the vector of observations from the $i^{\text{th}}$ population, $i = 1, \ldots, k$ and $X = (x_1, \ldots, x_k)^t$ be the vector of all the observations. Consider two fixed integers $r_1$ and $r_2$ belonging to the set $\{1, 2, \ldots, \min(n_1, \ldots, n_k)\}$. Define the indicator function $\tilde{g}(.)$ as

$$\tilde{g}_{ij}(x_{i\alpha_{r_1}}, \ldots, x_{i\alpha_{r_2}}; x_{j\beta_{r_1}}, \ldots, x_{j\beta_{r_2}}) = \begin{cases} 1 \text{ if } \max(x_{i\alpha_{r_1}}, \ldots, x_{i\alpha_{r_2}}) \leq \max(x_{j\beta_{r_1}}, \ldots, x_{j\beta_{r_2}}) \\ 0 \text{ otherwise.} \end{cases} \quad \text{...(2.3.1)}$$

Let $U_{ij}$ $\quad \text{(i \neq j, i, j = 1, \ldots, k)}$ be the two-sample $U$-statistic corresponding to the kernel (2.3.1). Thus we can write

$$(r_1, r_2) \quad \text{\{2.3.2\}}$$

$$U_{ij} = \left[\begin{array}{c} \binom{n_i}{r_1} \\ \binom{n_j}{r_2} \end{array}\right]^{-1} \sum \tilde{g}_{ij}(x_{i\alpha_{r_1}}, \ldots, x_{i\alpha_{r_2}}; x_{j\beta_{r_1}}, \ldots, x_{j\beta_{r_2}}), \quad \text{(2.3.2)}$$

$$\sum \tilde{g}_{ij}(x_{i\alpha_{r_1}}, \ldots, x_{i\alpha_{r_2}}; x_{j\beta_{r_1}}, \ldots, x_{j\beta_{r_2}}), \quad \text{(2.3.2)}$$
where $\Sigma^*$ denotes the summation extended over all combinations of $r_1$ integers $(\alpha_1, \ldots, \alpha_{r_1})$ chosen without replacement from $(1, \ldots, n_1)$ and all combinations of $r_2$ integers $(\beta_1, \ldots, \beta_{r_2})$ chosen without replacement from $(1, \ldots, n_j)$. Clearly $U_{ij}^{(1,1)}$ is the Mann-Whitney statistic computed for the $i$th and $j$th $(r_1, r_2)$ samples. The statistic $U_{ij}$ was proposed by Stephenson and Ghosh (1985) in the context of two-sample location problem. It can be easily seen that the expected value of $U_{ij}(r_1, r_2)$, under $\theta_i = \theta_j$ is $r_2/(r_1 + r_2)$, $i \neq j$. Define

$$T_{ij} = U_{ij} - r_2/(r_1 + r_2).$$

... (2.3.3)

Remark 2.3.1: The statistic $U_{ij}$ can also be written in terms of ranks as follows:

Let $Q_{ij} = \binom{n_i}{r_1} \binom{n_j}{r_2} U_{ij}$

= number of combinations of $r_1 X_i$ observations and $r_2 X_j$ observations with the largest of these $(r_1 + r_2)$ observations as one of the $r_2 X_j$ observations.

Therefore,

$$Q_{ij} = \sum_{s=1}^{n_j} \binom{a_s}{r_1 (r_2 - 1)} \binom{b_s}{r_1}$$

where,
Here,

\[ a_s = \text{number of } X_j \text{ observations less than } X_j(s) = s - 1 \]

\[ b_s = \text{number of } X_1 \text{ observations less than } X_j(s) = R_j(s) - s. \]

Here, 

\[ X_j(s) = s^{th} \text{ order statistic of the } X_j \text{ sample} \]

and

\[ R_j(s) = \text{rank of } X_j(s) \text{ in the combined increasing arrangement of } X_1 \text{ observations and } X_j \text{ observations.} \]

Thus,

\[ Q_{ij}^{(r_1', r_2')} = \frac{n_j^{s-1}}{s!} \binom{R_j(s)-s}{r_1 - 1}. \]

Therefore,

\[ U_{ij}^{(r_1', r_2')} = \left( \begin{array}{c} n_j \end{array} \right)^{-1} \frac{n_j^{s-1}}{s!} \binom{R_j(s)-s}{r_1 - 1}. \]

Let \( n = (n_1, ..., n_k)^t \) and \( P^* \) be a pre-assigned probability \((k^{-1} < P^* < 1)\). Choose the constants \( t_{ij}(n, P^*, r_1, r_2) \) such that

\[ P_0[T_{ij}^{(r_1', r_2')} \geq t_{ij}(n, P^*, r_1, r_2)] \forall i, i \neq j \geq P^*, \]

where \( P_0 \) indicates that the probability is computed under the configuration \( \theta_1 = ... = \theta_k \).

The proposed class of selection procedures based on the \( (r_1', r_2') \) statistics \( T_{ij} \) is:
\( R_1(r_1, r_2) \): Select \( \Pi_j \) in the subset if and only if

\[
T_{ij} \geq t_{ij}(n, p^*, r_1, r_2) \quad \forall i, i \neq j.
\]

For fixed values of \( r_1 \) and \( r_2 \), in their domain of definition, the selection procedure \( R_1(r_1, r_2) \) is a member of the proposed class. In the following section we shall obtain the expressions for the PCS and expected subset size.

2.4: PROBABILITY OF A CORRECT SELECTION AND EXPECTED SUBSET SIZE

Let \( \theta[k] \) be the unique component of \( \theta, \theta \in \Omega \), which corresponds to the best population \( \Pi(k) \). Let \( A \) be the action space of the subset selection problem which is the set of all non-empty subsets of \( \{1, \ldots, k\} \), where taking action \( a \in A \) means the selection of those populations whose indices are in \( a \). For any \( a \in A \), let \( |a| \) denote the cardinality of the set \( a \). Define the indicator function as

\[
CS(\theta, a) = \begin{cases} 
1 \quad \text{if } \theta[k] \in \{\theta_i, i \in a\} \\
0 \quad \text{otherwise}. 
\end{cases} \quad \text{...(2.4.1)}
\]

For any subset selection rule \( R \), let \( Z_R(X, a) \) be the probability assigned to \( a \) by \( R \) having observed \( X \).

In order to obtain the probability of a correct selection, define \( A = \{a \in A | CS(\theta, a) = 1\} \). The probability of correct selection is then
There may be different subsets in $A$ which contain the best population. All these subsets form a. However on the basis of available observations one and only one subset from $A$ can be chosen, that is, only one action at a time is possible. Thus the above integral becomes

$$\mathbb{E}_{\Theta} \left[ \sum_{a \in A} \mathbb{P}_{\Theta} (\text{action } a \text{ is taken} | X, R, R_{1}(r_{1}, r_{2})) \right] = \sum_{a \in A} \mathbb{E}_{\Theta} \left[ \mathbb{C}(\Theta, a) Z_{R, 1}(X, a) \right]. \quad \ldots (2.4.2)$$

Let $S$ denote the size of the selected subset. Then the expected subset size is

$$\mathbb{E}_{\Theta} [S | R_{1}(r_{1}, r_{2})] = \mathbb{E}_{\Theta} \left[ \sum_{a \in A} |a| Z_{R, 1}(X, a) \right]. \quad \ldots (2.4.3)$$

In the following section we shall show that the selection procedures proposed in section 2.3 meet the $P^*$-condition (2.2.1).

2.5: $P^*$-CONDITION FOR PROCEDURE $R_{1}(r_{1}, r_{2})$

In order to show that the procedure $R_{1}(r_{1}, r_{2})$ satisfies the $P^*$-condition (2.2.1), for every choice of $r_{1}$ and $r_{2}$ in
their domain of definition, we first show that the family of distributions of \( T_{ij} \) is stochastically increasing (SI).

**Definition 2.5.1:** A distribution function \( F_\theta(.) \) is said to be stochastically increasing (SI) in \( \theta \), where \( \theta \) belongs to some interval \( \Theta \) on the real line, if \( \theta_1 \geq \theta_2 (\theta_1, \theta_2 \in \Theta) \) implies \( F_{\theta_1}(x) \leq F_{\theta_2}(x) \) for all \( x \).

Some examples of SI distribution functions are

1. **(2.5.1):** any location parameter family (that is, \( F_{\theta}(x) = F(x-\theta) \));
2. **(2.5.2):** any scale parameter family with positive support (that is, \( F_{\theta}(x) = F(x/\theta), x>0, \theta>0 \));
3. **(2.5.3):** any monotone likelihood ratio family (that is, \( f_{\theta_1}(x_1)f_{\theta_2}(x_2) - f_{\theta_1}(x_2)f_{\theta_2}(x_1) \geq 0 \) for any \( \theta_1 \geq \theta_2 \) and \( x_1 \geq x_2 \));
4. **(2.5.4):** Lehmann's alternative type distribution function (that is, \( F_{\theta}(x) = [H(x)]^{\theta} \), where \( \theta>0 \) and \( H(x) \) is an arbitrary d.f.).

**Lemma 2.5.1:** The family of distributions of \( T_{ij} \) is stochastically increasing (SI).

In order to show that the family of distributions of \( T_{ij} \) is SI, we need a result stated below:

**Result:** Let \( F \) and \( G \) be two continuous cumulative distribution functions (cdf's) then the random variable (r.v.) \( G^{-1}F(X) \) has distribution function \( G \), that is, if \( Y \) has distribution
function $G$ then $Y = G^{-1}F(X)$, where $F$ is the cdf of the r.v. $X$.

Proof of the lemma 2.5.1:

Let $\theta_i < \theta_j \leq \theta_j'$

$\Rightarrow F_i(x) \geq F_j(x) \geq F_j'(x) \ \forall \ x,$

where $F_i(x) = F(x - \theta_i)$, $i = 1, \ldots, k$. Let $G(x; \Delta_{ij})$ be the cdf of $(r_1', r_2')$.

Now $\Delta_{ij} < \Delta_{ij}'$, and we want to show that $(r_1', r_2') \text{ st } (r_1, r_2')$

$\Rightarrow T_{ij} \leq T_{ij}'$, that is, $G(x; \Delta_{ij}) \geq G(x; \Delta_{ij}')$ for all $x$.

Consider the event

$$\max(X_{i1}, \ldots, X_{ir_1}) \leq \max(X_{j1}, \ldots, X_{jr_2})$$

$\text{ st }$

$$= \max(X_{i1}, \ldots, X_{ir_1}) \leq \max(F_1^{-1}F_i(X_{i1}), \ldots, F_j^{-1}F_i(X_{ir_1}))$$

$\Rightarrow \max(X_{i1}, \ldots, X_{ir_1}) \leq \max(F_1^{-1}F_1(X_{i1}), \ldots, F_j^{-1}F_1(X_{ir_2}))$

(since $F_1^{-1}(x) \geq F_j^{-1}(x) \ \forall \ 0 < x < 1$)

$\text{ st }$

$$= \max(X_{i1}, \ldots, X_{ir_1}) \leq \max(X_{j1}, \ldots, X_{jr_2}).$$

Thus

$$\tilde{\xi}_{ij}(X_{i1}, \ldots, X_{ir_1}; X_{j1}, \ldots, X_{jr_2}) = 1$$

$\Rightarrow \tilde{\xi}_{ij}(X_{i1}, \ldots, X_{ir_1}; X_{j1}, \ldots, X_{jr_2}) = 1.$

$(r_1', r_2')$

As $U_{ij}$ is the average of $\tilde{\xi}$'s it follows that

$$(r_1', r_2') \text{ st } (r_1, r_2') \Rightarrow U_{ij} \leq U_{ij}'$$

or $T_{ij} \leq T_{ij}'$.

Hence the lemma follows.
We shall now use a result of Mahamunulu (1967) (stated below as lemma 2.5.2) to show that the procedure $R_1(r_1, r_2)$ satisfies the $P^*$-condition for every choice of pair $(r_1, r_2)$ with $r_1$ and $r_2$ taking values in the set \{1, 2, ..., \min(n_1, ..., n_k)\}.

**Lemma 2.5.2:** Let $G(x|\theta) = G_{\theta}(x)$ where $\theta \in \Theta$ be an SI family of distribution functions on the real line. Let $X_1, ..., X_k$ be independent random variables, where the distribution function of $X_i$ is $G(x_i|\theta_i)$. For any fixed $i (1 \leq i \leq k)$, if $\psi = \psi(x_1, ..., x_k)$ is a nondecreasing (nonincreasing) function of $x_i$ when all $x_j$ for $j \neq i$ are held fixed, then $E[\psi(X_1, ..., X_k)]$ is a nondecreasing (nonincreasing) function of $\theta_i$.

**Theorem 2.5.1:** For every choice of $r_1$ and $r_2$ in their domain of definition, the selection procedure $R_1(r_1, r_2)$ satisfies the $P^*$-condition.

**Proof:** Assume without loss of generality that $\pi_j$ is the best population. Now $\min_{i \neq j} T_{ij}$ is nondecreasing in $X_{j_1}, ..., X_{j_n_j}$ and nonincreasing in other components of $X$. Moreover, the family of distributions of $T_{ij}$ is SI (by lemma 2.5.1).

Define the indicator function $I_1(.)$ as
By writing the expected value of the indicator function \( I(.) \) as the expectation of conditional expectation and then using lemma 2.5.2, we have for any \( \theta \in \Omega \),

\[
P^* \leq P_0 \left[ \frac{r_1, r_2}{T_{ij}} \geq t_{ij}(n, p^*, r_1, r_2) \ \forall \ i, i \neq j \right]
\]

\[
P^* \leq \frac{r_1, r_2}{T_{ij}} \geq t_{ij}(n, p^*, r_1, r_2) \ \forall \ i, i \neq j
\]

\[
P^* = P|_{\Omega}^{r_1, r_2}[CS|_{R}(r_1, r_2)].
\]

This proves the Theorem.

Gupta and Nagel(1971), and Santner(1975) have defined unbiasedness, monotonicity and strong monotonicity properties of a selection procedure while proposing selection procedures for parametric families of probability distributions. Now we define the unbiasedness, monotonicity and strong monotonicity properties of a selection procedure \( R \) when the parameters of interest are location parameters.

For any \( \theta \in \Omega \), let

\[
P_{\theta}(j) = P_{\theta}^{r_1, r_2}(\Pi_j \ \text{is included in the subset}|R), \ j=1,...,k.
\]

Definition 2.5.1: The selection procedure \( R \) is said to be unbiased if and only if
\[ \theta_j > \theta_i, \text{ for } i=1,\ldots,k \text{ implies that } \]
\[ P_{\theta}(j) > P_{\theta}(i), \text{ for } i=1,\ldots,k \text{ and for all } \theta \in \Omega. \]

**Definition 2.5.2:** The selection procedure \( R \) is said to be monotone if and only if
\[ \theta_j > \theta_1 \text{ implies that } P_{\theta}(j) > P_{\theta}(i) \text{ for all pairs } (j,i) \text{ and for all } \theta \in \Omega. \]

**Definition 2.5.3:** The selection procedure \( R \) is strongly monotone in \( \prod_j \) if and only if
\[ P_{\theta}(j) \text{ is increasing in } \theta_j \text{ when all other components of } \theta \text{ are fixed, and } \]
\[ P_{\theta}(j) \text{ is decreasing in } \theta_i (i \neq j) \text{ when all other components of } \theta \text{ are fixed. } \]

Following Theorem shows the strong monotonicity of procedures \( R^1(r_1,r_2) \) which in turn establishes its monotonicity and unbiasedness.

**Theorem 2.5.2:** The selection procedures \( R^1(r_1,r_2) \) are strongly monotone.

**Proof:** In Theorem 2.5.1 we have seen that the indicator function \( \text{I}_1(T_{1j}^1, \ldots, T_{j-1j}^1, T_{j+1j}^1, \ldots, T_{kj}^1) \) is nondecreasing function of observations from \( j \)th population and nonincreasing function of observations from \( i \)th population \( (i=1,\ldots,k, i \neq j) \). By writing the expectation as expectation of conditional expectation and then using lemma 2.5.2, we get,
$$P_{\hat{\theta}}(j) = E_{\hat{\theta}} \left[ I_{1}(T_{1j}, \ldots, T_{j-1,j}, T_{j+1,j}, \ldots, T_{kj}) \right]$$

$$\leq E_{\hat{\theta}^*} \left[ I_{1}(T_{1j}, \ldots, T_{j-1,j}, T_{j+1,j}, \ldots, T_{kj}) \right]$$

$$= P_{\hat{\theta}^*}(j),$$

where $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_{j-1}, \hat{\theta}_j, \hat{\theta}_{j+1}, \ldots, \hat{\theta}_k)^t$ and

$$\hat{\theta}^* = (\hat{\theta}_1, \ldots, \hat{\theta}_{j-1}, \hat{\theta}_j^*, \hat{\theta}_{j+1}, \ldots, \hat{\theta}_k)^t$$

with $\hat{\theta}_j \leq \hat{\theta}_j^*$. This proves the Theorem.

It is easy to see that strong monotonicity implies monotonicity which in turn implies unbiasedness (see Santner(1975)); as a consequence we have the following corollary.

**Corollary 2.5.1:** The selection procedures $R_{1}(r_1, r_2)$ are monotone and unbiased.

**Remark 2.5.1:** The selection procedures $R_{1}(r_1, r_2)$ can be modified to select a subset of $k$ populations better than an unknown control population as follows:

Let $\Pi_0, \Pi_1, \ldots, \Pi_k$ be $(k+1)$ populations which differ only in their location parameters. Let $F_i(x) = F(x - \theta_i)$ be the cdf of the $i^{th}$ population, $i = 0, 1, \ldots, k$. The population $\Pi_0$ with cdf $F_0(x) = F(x - \theta_0)$ is the control population with unknown location parameter $\theta_0$. The population $\Pi_i$ is said to be 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 $\mathbf{X}_i = (X_{i0}, \ldots, X_{in_i})^t$ and $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})^t$ be
the independent vector of random observations from $\Pi_0$ and $\Pi_i$
respectively, $i=1,\ldots,k$. The proposed subset selection
procedure is

$$R_{c_1}(r_1, r_2): \text{ For any } i \text{ between } 1 \text{ and } k \text{ include the population}
$$

$$\Pi_i \text{ in the subset if and only if}
$$

$$T_{i0} \geq t_{i0}(\mathbf{n}^*, p^*, r_1, r_2),
$$

where $\mathbf{n}^*=(n_0, n_1, \ldots, n_k)^t$ and the constants $t_{i0}(\mathbf{n}^*, p^*, r_1, r_2)$ are
chosen such that for a pre-assigned probability $p^* (2 < p^* < 1)$,
we have

$$P_0[T_{i0} \geq t_{i0}(\mathbf{n}^*, p^*, r_1, r_2) \forall i = 1, \ldots, k] \geq p^*.
$$

Here $P_0$ indicates that the probability is computed under the
parametric configuration $\theta_0 = \theta_1 = \ldots = \theta_k$. 