Chapter 6

Subset Selection Procedure for Quantile Function based on Two-Sample Statistic

6.1 Introduction

Selection and ranking problems have been considered by using either the indifference zone approach introduced by Bechhofer (1954) or the subset selection approach suggested by Gupta (1956). In indifference zone approach, the “best” population is chosen out of $k$ populations, such that the probability of it being the “best” is at least a prespecified number $P^*$ where $1/k < P^* < 1$. In subset selection, a non empty subset of random size containing the best population is chosen. In this approach, the probability of inclusion of the best population in the subset, is at least $P^*$ and the size of selected subset is a random number. The selection procedure based on a rule $R$ is called correct selection (CS) if the selected subset contains the desired population. The probability of correct selection (PCS) using rule $R$ should satisfy the requirement that $P(CS|R) \geq P^*$. The configuration of parameters which yields the infimum of the probability of correct selection (PCS) is referred to as the least favourable configuration (LFC).
Gupta (1956) proposed a subset selection procedure to select the population with the largest mean from \( k \) normal populations where common variance \( \sigma^2 \) may be known or unknown. Gupta (1962) considered a procedure for selecting a gamma population with the largest scale parameter when the shape parameters in \( k \) gamma populations are same. Gupta and Sobel (1960) proposed a procedure for selecting a Binomial population with the largest success probability. A restricted subset selection rule for selecting at least one of the \( t \) best normal populations in terms of means with known variance was suggested by Chen et al. (2014).

Gupta (1965) proposed a general procedure for the families of distributions differing in location or scale parameter and suggested a procedure for subset selection in terms of location and scale parameters. Gupta and McDonald (1970) considered families of distributions with different scale or location parameters and proposed nonparametric procedures based on ranks for subset selection with the largest scale or location parameter. A nonparametric procedure to select a subset of \( k \) populations consisting of the largest \( u^{th} \) order quantile was given by Rizvi and Sobel (1967) under the assumption of equal sample sizes. van Eeden and Zidek (2012) extended the procedure given by Rizvi and Sobel (1967) to unequal sample sizes.

Gill and Mehta (1991) proposed subset selection procedures based on two-sample U-statistic for selecting a subset of \( k \) populations containing the population associated with the smallest scale parameter. Kumar et al. (1992) also suggested a procedure based on U-statistic to select a subset containing the population with the largest location parameter.

Gill and Mehta (1989) proposed a subset selection procedure based on a two-sample statistic to select a subset containing the population associated with uniformly smaller hazard rate. They developed the procedure using the two-sample test for hazard rate given by Cheng (1985). These kinds of procedures, where we have to select a subset containing the population with uniformly smaller or larger function viz survival function, mean residual life function, quantile function, have not got much attention in literature.

Quantile functions are alternative to distribution functions and are used for modelling lifetime.
6.2 Preliminaries

The comparison of the quantile functions of several populations is important. For example, if we want to compare per capita income of several countries, then the country with uniformly larger quantile will be considered rich. Another situation where this comparison is important is when we need to compare $k$ different brands of hypertension medicines. A medicine will be considered effective if it brings the blood pressure to normal level for majority of patients. So in such situations, population with the smallest quantile function will be considered as the best.

In this chapter, we propose a subset selection procedure for selecting a population with uniformly larger quantile function. This procedure is based on a two-sample statistic proposed for testing the equality of quantile functions of two populations. The empirical type estimators of the quantile function have been used in the test statistic for selection procedure.

In Section 6.2, preliminaries have been defined. In Section 6.3, a two-sample test for testing equality of quantile functions has been proposed. The subset selection procedure has been described in Section 6.4. Simulations have been carried out to find the expected subset size and probability of correct selection in Section 6.5. Conclusions are given in Section 6.6.

6.2 Preliminaries

Let $X_{ij}, j = 1, 2, \cdots, n$ be independent and identically distributed (i.i.d) random variables from population $\pi_i, i = 1, 2, \cdots, k$. Let $F_i(x)$ be the cdf of the $i^{th}$ population with $f_i(x)$ as the corresponding pdf.

Let $X = (X_{11}, \cdots, X_{1n}, X_{21}, \cdots, X_{2n}, \cdots, X_{k1}, X_{k2}, \cdots, X_{kn})$ be the vector of all observations with joint cdf as $F(x) = \prod_{i=1}^{k} F_i(x)$. The quantile function of $i^{th}$ population is defined in (4.2.1). The quantile density function for $i^{th}$ population is $q_i(u) = Q'_i(u)$ and is given in (4.2.3).

An empirical estimator of the quantile function is defined in (4.2.4).
6.3 Two - Sample Statistic

We consider the problem of testing of equality of quantile functions. In particular, we consider $i^{th}$ and $j^{th}$ populations where $i \neq j$ and wish to test the hypothesis:

$H_0 : Q_i(u) = Q_j(u) \quad \forall u$ versus

$H_1 : Q_i(u) > Q_j(u) \quad \forall u \in (0, 1)$.

We consider first the following measure of departure

$W_{ij} = \sup_{0 < u < 1} T_{ij}(u)$ \hspace{1cm} (6.3.1)

where

$T_{ij}(u) = \frac{\sqrt{n} \left( \hat{Q}_i(u) - \hat{Q}_j(u) - (Q_i(u) - Q_j(u)) \right)}{\sigma_T}$ \hspace{1cm} (6.3.2)

In (6.3.2), $\sigma_T = \max_u \sqrt{\sigma_i^2(u) + \sigma_j^2(u)}$, where $\sigma_i^2(u)$ and $\sigma_j^2(u)$ are the variances of $\hat{Q}_i(u)$ and $\hat{Q}_j(u)$, the empirical estimators of quantile functions of $i^{th}$ and $j^{th}$ populations respectively.

Next, we find distribution of $W_{ij}$ under $H_0$. Let $D[0, 1]$ be the space of cadlag functions on $[0, 1]$ and $D = D[0, 1] \times D[0, 1]$. $D$ and $D[0, 1]$ are equipped with the product norm and the uniform norm $||.||$ respectively. Csörgö (1983) proved the weak convergence of normed quantile process under certain conditions and this result is stated as Lemma 6.3.1.

Lemma 6.3.1. If

(i) $F_i(x)$ is twice differentiable on $(a, b)$ where $a = \sup \{ x : F_i(x) = 0 \}$ and $b = \inf \{ x : F_i(x) = 1 \}$, $-\infty \leq a < b \leq \infty$;

(ii) $F_i'(x) = f_i(x) > 0$ on $(a, b)$;
(iii) for some $\gamma > 0$, we have

$$
\sup_{a < x < b} F_i(x)(1 - F_i(x)) \frac{|f'_i(x)|}{f'_i(x)} = \sup_{0 < u < 1} u(1 - u) \frac{f'_i(\tilde{Q}_i(u))}{f'_i(Q_i(u))} \leq \gamma;
$$

(iv) $A := \lim_{x \to a} f_i(x) < \infty$, $B := \lim_{x \to b} f_i(x) < \infty$;

(v) one of

$(v, \alpha)\min(A, B) > 0$

$(v, \beta)$ if $A = 0$ (resp $B = 0$), then $f_i$ is nondecreasing (nonincreasing) on $(a, b)$;

then $\sup_{0 < u < 1} |\rho_n(u) - B_n(u)| \to 0$ and $h(\rho_n(u); 0 < u < 1) \xrightarrow{d} h(B(u)); 0 < u < 1$ for every continuous functional $h : D(0, 1) \to \mathbb{R}$ where \{B(u); 0 < u < 1\} is a Brownian bridge and $\rho_n(u) = \sqrt{n} f_i(\tilde{Q}_i(u))(\tilde{Q}_i(u) - Q_i(u))$.

Theorem 6.1. Under $H_0$, $U_{ij} = \sup_{0 < u < 1} \sqrt{n}(\tilde{Q}_i(u) - \hat{Q}_j(u))$ converges in distribution to $\sup_{0 < u < 1} (B_i(Q_i(u)) - B_j(Q_j(u)))$ as $n \to \infty$.

Proof. From Lemma 6.3.1, we have

$$
\sqrt{n}\{f_i(\tilde{Q}_i(u))(\tilde{Q}_i(u) - Q_i(u)), f_j(\tilde{Q}_j(u))(\tilde{Q}_j(u) - Q_j(u))\} \xrightarrow{d} \{B_i(Q_i(u)), B_j(Q_j(u))\}
$$

where $B_i$ and $B_j$ are Brownian bridge processes with zero means. Under $H_0$, we have

$Q_i(u) = Q_j(u) = Q(u) \forall u$.

This gives $f_i(Q_i(u)) = f_j(Q_j(u)) = f(Q(u))$.

Using continuous mapping theorem, it is seen under $H_0$ that
6.3 Two-Sample Statistic

\[ U_{ij} \text{ converges to } \sup_u \left( \frac{B_i(Q_i(u))}{f(Q(u))} - \frac{B_j(Q_j(u))}{f(Q(u))} \right) \text{ as } n \to \infty \text{ for } 0 < u < 1. \]

Let \[ \frac{B_i(Q_i(u)) - B_j(Q_j(u))}{f(Q(u))} = h_{ij}(u) \text{ with } \]

\[ V(h_{ij}(u)) = \frac{V(B_i(Q_i(u))) + V(B_j(Q_j(u)))}{f^2(Q(u))} = \sigma_k^2(u). \]

\[ \Rightarrow \frac{\sqrt{n}(\hat{Q}_i(u) - \hat{Q}_j(u))}{\sigma_T} \to W\left(\frac{\sigma_k^2(u)}{\sigma_T^2}\right) \text{ where } \sigma_T^2 = \max_t[\sigma_k^2(t)] \]

and \( \{ W(t) : t \geq 0 \} \) is a standard Brownian motion (Wiener process). Under \( H_0 \),

\[ W_{ij} = \frac{U_{ij}}{\sigma_T} = \sup_{0 < u < 1} \frac{\sqrt{n}(\hat{Q}_i(u) - \hat{Q}_j(u))}{\sigma_T} \text{ and hence } \]

\[ \lim_{n \to \infty} P[W_{ij} > b] = P[\sup_{0 < u < 1} T_{ij}(u) > b] = 2(1 - \Phi(b)), \tag{6.3.3} \]

where \( \Phi(b) \) is the cdf of standard Normal distribution at \( b \).

\( \sigma_T \) depends on the unknown quantile density functions. Soni et al. (2012) proposed a smooth consistent estimator of the quantile density function as given by (4.2.9). The test statistic proposed to test \( H_0 \) against \( H_1 \) is given by

\[ \hat{W}_{ij} = \sup_{0 < u < 1} \frac{\sqrt{n}(\hat{Q}_i(u) - \hat{Q}_j(u))}{\hat{\sigma}_T}. \tag{6.3.4} \]

We reject \( H_0 \) for large values of the test statistic. Note that under \( H_0 \), \( W_{ij} \) and \( \hat{W}_{ij} \) have the same limiting distribution.
6.4 Procedure

In this section, the subset selection procedure is described. Consider
\[ \Omega = \{ Q : Q = (Q_1(u), Q_2(u), \ldots, Q_k(u)) \}, \]
where \( Q_i(u) < \infty, i = 1, 2, \ldots, k \) as the space of
the quantile functions of \( k \) populations. For two populations \( \pi_i \) and \( \pi_j \), we consider \( \pi_i \) to be
better than \( \pi_j \) if \( Q_i(u) > Q_j(u) \ \forall \ u \in (0,1) \). This means that population with uniformly larger
quantile function is considered to be better. We assume that there exists a best population \( Q \in \Omega \)
such that \( Q_i(u) > Q_j(u) \ \forall \ u \) and \( j \neq i \). Let \( Q |_{[1]}(u) < Q |_{[2]}(u) < \ldots < Q |_{[k]}(u) \) be the exact
ordering of the quantile functions \( \forall \ u \in (0,1) \). The aim is to select a subset of \( k \) populations
which contains \( Q |_{[k]}(u) \), the largest quantile function. Any such selection will be the correct
selection (CS). The problem is to find a rule \( R \) such that for a preassigned \( P^* \) (\( \frac{1}{k} < P^* < 1 \)),
the following probability statement holds

\[ P[CS|R] > P^* \ \forall \ \Omega(u) \in \Omega. \] (6.4.1)

Let \( A \), the set of all non-empty subsets of \( \{1, 2, \ldots, k\} \) be the action space of the subset selection
problem. Taking action \( a \in A \) means that the selected subset contains the population whose
quantile function has index in set \( a \). For any \( a \in A \), let

\[ CS(Q, a) = \begin{cases} 1 & \text{if } Q |_{[a]}(u) \in \{Q_i(u); \ i \in a\}, \\ 0 & \text{otherwise}. \end{cases} \] (6.4.2)

\( CS(Q, a) \) is an indicator function which takes value 1 if the selected subset contains the largest
quantile function. If \( a \) is the selected subset, then \( |a| \) denotes its size.
6.4 Procedure

6.4.1 Selection Procedure

The selection procedure proposed to select a subset with the largest quantile function is given by the following rule:

$R$: Select $\pi_i$ in the subset iff $\hat{W}_{ij} > C_j^*(n, P^*) \ \forall \ j \neq i$

where the constants $C_j^*(n, P^*)$ are chosen such that

$$P_0(\hat{W}_{ij} > C_j^*(n, P^*) \ \forall \ j, j \neq i) \geq P^*$$

and $P_0$ indicates that the probability is computed under the assumption that

$$Q_1(u) = Q_2(u) = \cdots = Q_k(u). \quad (6.4.3)$$

Now

$$P^* = P_0\left(\hat{W}_{ij} > C_j^*(n, P^*) \ \forall j, j \neq i\right)$$

$$= P_0\left(Z_{ij} > B_j^*(n, P^*) \ \forall j, j \neq i\right)$$

$$\leq P_0\left(Z_{ij} > C^*(n, P^*) \ \forall j, j \neq i\right)$$

where $Z_{ij} = \frac{\hat{W}_{ij} - \sqrt{\frac{2}{n}}}{\sqrt{1 - \frac{2}{n}}}$, is the standardised version of $\hat{W}_{ij}$,

$$B_j^*(n, P^*) = \frac{C_j^*(n, P^*) - \sqrt{\frac{2}{n}}}{\sqrt{1 - \frac{2}{n}}}$$

and $C^*(n, P^*) = \min_j B_j^*(n, P^*)$.

Then

$$P^* \leq P_0\left(\min_j Z_{ij} > C^*(n, P^*)\right)$$

$$= P_0\left(\max_j Z_{ij} < -C^*(n, P^*)\right). \quad (6.4.4)$$
It is well known that the joint distribution of standardised correlated variables tends asymptotically to multivariate normal distribution (MVN). Thus, under assumption (6.4.3), as \( n \to \infty \), the distribution of the vector \((Z_{i1}, Z_{i2}, \cdots, Z_{ik})\) tends to MVN with mean vector \( \mathbf{0} \), variance \( \sigma_{ii} = 1 \) and covariance as \( \sigma_{ij} = \rho \) for \( i \neq j \). Simulations show that components of the vector are equally correlated. Hence, we can use the tables given by Gupta et al. (1973) for finding \(-C^n(n, P^*)\) by using (6.4.4).

### 6.4.2 Probability of Correct Selection and Expected Subset Size

In this subsection, we show that our selection procedure satisfies \( P^* \) condition and also find an expression for the expected subset size.

Let \( \bar{A} = \{ a \in A | CS(Q, a) = 1 \} \) be the collection of all subsets of the action space which leads to correct selection. Then the probability of CS using rule \( R \) is

\[
P(CS|R) = P(Q_{[k]} \text{ is in the selected subset}|R)
= P(\bigcup_{a \in \bar{A}}(X \text{ is observed and action } a \text{ is taken}|R)
= \int P(\bigcup_{a \in \bar{A}}\{ \text{action } a \text{ is taken}|X = x, R\})dF(x).
\]

There will be several subsets in \( \bar{A} \) containing the best population. However, we choose only one subset from \( \bar{A} \) and only one action is taken at a time. So (6.4.5) can be written as

\[
P(CS|R) = \mathbb{E}\left[ \sum_{a \in \bar{A}} P(\text{action } a \text{ is taken}|X, R) \right]
= \mathbb{E}\left[ \sum_{a \in \bar{A}} CS(Q, a)Z_R(X, a) \right],
\]

(6.4.6)

where \( Z_R(X, a) \) is the probability assigned to \( a \) by \( R \) after having observed \( X \). Let \( S \) denote the selected subset size which is a random variable. Then expected subset size is
We have to show that the selection procedure based on R ensures

\[ \inf_{Q \in \mathcal{B}} P(CS|R) \geq P^*. \]  

(6.4.7)

This is called the \( P^* \) condition and to prove this, we need to ensure that test based on two-sample statistic is unbiased.

**Lemma 6.4.1.** Under \( H_0 \), \( P_{H_0}(W_{ij} > w) \leq P_{H_1}(W_{ij} > w) \) for all \( w \) where,

\[ W_{ij} = \sup_{0 < u < 1} \sqrt{n} \left( \frac{\hat{Q}_i(u) - \hat{Q}_j(u)}{\hat{\sigma}} \right) \]

is the test statistic for testing

\( H_0 : Q_i(u) = Q_j(u) \) versus \( H_1 : Q_i(u) \geq Q_j(u) \).

**Proof.** It is clear that the test statistic \( \hat{W}_{ij} \) under \( H_0 \) is stochastically smaller than test statistic \( \hat{W}_{ij} \) under \( H_1 \), that is

\[ \hat{W}_{ij}|H_0 \leq \hat{W}_{ij}|H_1 \]  

(6.4.8)

\[ \Rightarrow P_{H_0}(\hat{W}_{ij} > w) \leq P_{H_1}(\hat{W}_{ij} > w) \text{ for } w > 0. \]  

(6.4.9)

This proves that the proposed test is unbiased.

**Theorem 6.2.** For procedure \( R \), \( P^* \) condition is satisfied.

**Proof.** For \( i = 1, 2, \ldots, k \), using Lemma 6.4.1,

\[ P^* \leq P_{H_0}(\hat{W}_{ij} > w, j \neq i, j = 1, 2, \ldots, k) \]
6.5 Simulations

Simulations have been carried out to check the efficiency of the procedure described in Section 6.4. We consider the cases when \( k \), the number of populations under consideration is 3, 4, 8. The considered sample sizes are \( n = 50 \) and 100, \( P^* = 0.05 \). To estimate the quantile density function, we use bandwidth \( h(n) = 0.15 \) and the optimal Epaneknikov kernel given by

\[
K(u) = 0.75(1 - u^2)I(|u| \leq 1).
\]

The observations have been generated from Generalized Lambda and Normal distributions. The quantile function \( Q_{GL}(u) \) for GLD(\( \lambda_1, \lambda_2, \lambda_3, \lambda_4 \)) is given in (1.3.1). We can see from Figure 6.1 that quantile function corresponding to the largest \( \lambda_1 \), the location parameter, is uniformly larger than the quantile functions corresponding to smaller \( \lambda_1 \). The quantile function for Normal distribution with mean \( \mu \) and variance \( \sigma^2 \) is given by

\[
Q_N(u) = \mu + \sigma Q_S(u)
\]

where \( Q_S(u) \) denotes the quantile function for standard normal distribution. If there is increase in \( \mu \), then new quantile function will be uniformly larger than the previous one.

The expected subset size has been calculated for GLD and normal distributions. Table 6.1 shows the expected subset size for GLD when location parameter is varied by 0.2 and 1, respectively and other parameters \( \lambda_2 = 1, \lambda_3 = 2 \) and \( \lambda_4 = 1 \) are kept fixed. In the body of tables, values
6.5 Simulations

Table 6.1: Expected subset size for GLD when $\lambda_2 = 1$, $\lambda_3 = 2$, $\lambda_4 = 1$

<table>
<thead>
<tr>
<th>$k$</th>
<th>n</th>
<th>$\lambda_{1,k} = 3.6 + 0.2(k - 1)$</th>
<th>$\lambda_{1,k} = k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>50</td>
<td>1.905</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.574</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>2.052</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.648</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>2.035</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.647</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>2.182</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.741</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>2.220</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.771</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>50</td>
<td>2.265</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.796</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 6.1: Plot of quantile functions for GLD ($k = 5$)

give expected subset size when the procedure is based on empirical estimator.
From Table 6.1, we see that

(i) the expected subset size is less when the difference in location parameter is 1 as compared to the case when the difference is 0.2. This means that larger the difference in location parameter, smaller the expected subset size;

(ii) the expected subset size decreases with an increase in sample size.

On calculation of PCS, it is observed that it is equal to 1 in all cases.

The following table gives the expected subset size in case of \( k \) Normal distributions with \( \mu_i = i \) for \( i = 1, 2, \ldots, k \). The variances of all populations are assumed to be equal, that is \( \sigma = 1 \) in first case and \( \sigma = 2 \) in other case.

<table>
<thead>
<tr>
<th>( k )</th>
<th>( n )</th>
<th>( \sigma = 1 )</th>
<th>( \sigma = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>50</td>
<td>1.092</td>
<td>1.629</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.051</td>
<td>1.508</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>1.100</td>
<td>1.624</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.065</td>
<td>1.506</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>1.114</td>
<td>1.581</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.106</td>
<td>1.710</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>1.128</td>
<td>1.738</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.065</td>
<td>1.600</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>1.124</td>
<td>1.733</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.064</td>
<td>1.631</td>
</tr>
<tr>
<td>8</td>
<td>50</td>
<td>1.131</td>
<td>1.768</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.071</td>
<td>1.645</td>
</tr>
</tbody>
</table>
Table 6.2 shows the effect of scale parameter on expected subset size, when the observations are generated from two sets of normal distributions with common variance. Expected sample size is more for observations from normal distribution with larger scale parameter; In this situation also, PCS for all cases is found to be equal to 1.

From all the tables, we notice that expected subset size decreases with an increase in sample size. Values have been reported upto three decimal places as otherwise the expected sample size would be same.

6.6 Conclusions

In this chapter, a two-sample test statistic for testing equality of quantile functions of two independent populations is proposed. The test statistic is based on an empirical estimator of the quantile function. There is little work available for selection the population with uniformly largest function, here an attempt has been made. A subset selection procedure to select a population with uniformly larger quantile function is proposed. The subset selection procedure is based on two-sample statistic. The probability of correct selection is always 1 and expected subset size reduces as departure in the quantile functions increases.