THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT #253

RIZVI-SOBEL SUBSET SELECTION WITH UNEQUAL SAMPLE SIZES

BY

CONSTANCE VAN EEDEN

November 2009

SECTION 3 OF THIS TECHNICAL REPORT CONTAINS A MISTAKE

A CORRECTED VERSION OF THIS SECTION CAN BE FOUND IN

Technical Report # 261, December 2010

by

Constance van Eeden

Department of Statistics, The University of British Columbia
Vancouver, B.C., Canada

RIZVI-SOBEL SUBSET SELECTION WITH UNEQUAL SAMPLE SIZES

by Constance van Eeden
Department of Statistics, The University of British Columbia
Vancouver, B.C., Canada

Abstract

Rizvi and Sobel (1967) give procedures for selecting, from k distribution functions $F_1, \ldots, F_k$, a subset containing the one with the largest α-quantile. The sizes of the samples taken from these F_i are equal and the authors give conditions on the F_i under which the probability of correct selection is at least P^* for a given $P^* \in (1/k, 1)$. In this 1967-paper, Rizvi and Sobel also give a procedure for selecting a subset containing the F_i with the smallest α-quantile. In the present paper these Rizvi-Sobel procedures are extended to the case where the sample sizes are not necessarily equal.

Keywords: Subset selection; unequal sample sizes; quantiles.
AMS 2000 classification: 62F07.

1 Introduction

Let, for $k \geq 2$ and $i = 1, \ldots, k$, $X_{1,i}, \ldots, X_{n_i,i}$ be independent samples from distributions F_i. Rizvi and Sobel (1967) (RS from now on) give a subset selection procedure for selecting the F_i with the largest α-quantile, as well as a procedure for selecting the one with the smallest α-quantile. Their sample sizes are equal and under conditions on the F_i their procedures have a probability of correct selection $\geq P^*$ for a given $P^* \in (1/k, 1)$. In the present paper the RS procedures are extended to the case where the sample sizes are not necessarily equal.

Several results on subset selection with unequal sample sizes have been obtained for models other than the RS one. For $X_{j,i} \sim \mathcal{N}(\theta_i, \sigma^2)$ with σ known and the goal of a subset containing the population with the largest θ_i, Sitek (1972) generalizes, to unequal sample sizes, a procedure of Gupta (1956) (see also Gupta and Sobel (1957)). However, Dudewicz (1974) shows that Sitek's derivation is incorrect. But Gupta and Huang (1974) give a procedure for this normal-mean problem with unequal sample sizes, as well as one for the case when σ is unknown. Chen, Dudewicz and Lee (1976) also consider the Gupta-Huang (1974) case. They study a procedure which is different from the Gupta-Huang procedure and make comparisons between the two methods with respect to average subset size and ease of implementation. They find that their procedure is the better one on both points.
There are several non-parametric procedures, namely by Gupta and Mcdonald (1970), by Hsu (1981) and by Kumar, Mehta and Kumar (2002). Gupta and McDonald consider a stochastically increasing family and base their procedure on rank statistics where the ranks are from the pooled samples. Kumar, Mehta and Kumar consider the location problem and use U-statistics, while Hsu considers a stochastically increasing family and uses two-sample rank statistics - one for each pair of samples. Only Gupta-McDonald and Hsu look at models that are not too far from the RS one, where Hsu does not refer to RS while Gupta and McDonald do.

The two procedures extending the two RS-procedures as well as their formulas for the probability of correct selection are given in the sections 2 and 3. Section 4 gives numerical values for the probability of correct selection for the case of the largest α-quantile for several values of k, the n_i and α and in Section 5 some properties of the probability of correct selection of our extensions are compared with those of the RS procedures. Section 6 contains some comments on the fact that the RS procedures as well as our extensions of them are not sign-invariant. Our proofs are generalizations of those of RS.

2 The procedure for the largest quantile

The conditions on the distribution functions F_i are the same as those of RS. So the F_i satisfy

$$\min_{1\leq i\leq k} F_{[i]}(y) = F_{[k]}(y) \text{ for all } y, \tag{2.1}$$

where $F_{[i]}$ is the distribution with the i-th smallest α-quantile, the F_i are continuous and have a unique α-quantile. Further, r_i and c_i, $i = 1, \ldots, k$, are integers satisfying

$$1 \leq r_i \leq (n_i + 1)\alpha < r_i + 1 \leq n_i + 1 \text{ and } 0 \leq c_i \leq r_i - 1 \tag{2.2}$$

and $Y_{j,i}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$ is the j-th order statistics of the sample from F_i.

The proposed procedure is

$$R_1 : \text{ put } F_i \text{ in the subset } \Leftrightarrow Y_{r_i,i} \geq \max_{1\leq j\leq k,\, j\neq i} Y_{r_j - c_j, j}. \tag{2.3}$$

Then, when $F_i = F_{[k]}$, the probability of correct selection given by

$$P_{i,d_i}(CS|R_1) = P(Y_{r_i,i} \geq \max_{j\neq i} Y_{r_j-c_j,j}) = P(Y_{r_j-c_j,j} \leq Y_{r_i,i}, j \neq i), \tag{2.4}$$

where, for $i = 1, \ldots, k$, $d_i = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_k)$. We now need, for $U_1, \ldots, U_N$ independent with continuous distribution function G, the distribution of the ν-th order statistic Z_ν which is given by

$$P(Z_\nu \leq z) = \sum_{l=\nu}^{N} \binom{N}{l} G(z)^l (1 - G(z))^{N-l} = I_{G(z)}(\nu, N - \nu + 1), \tag{2.5}$$

where, for $u \in (0, 1)$ and positive a and b, $I_u(a, b)$ is the standard incomplete beta function given by

$$I_u(a, b) = \frac{1}{B(a, b)} \int_0^u t^{a-1}(1 - t)^{b-1} dt. \tag{2.6}$$

This gives, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$,

$$P(Y_{j,i} \leq x) = \sum_{l=j}^{n_i} \binom{n_i}{l} F_i(x)^l (1 - F_i(x))^{n_i - l} = I_{F_i(x)}(j, n_i - j + 1), \tag{2.7}$$

implying that (2.4) can be written as

$$\left.\begin{aligned} & P_{i,d_i}(CS|R_1) \\ & = P(Y_{r_i,i} \geq \max_{j \neq i} Y_{r_j - c_j, j}) = P(Y_{r_j - c_j, j} \leq Y_{r_i,i}, j \neq i) \\ & = \textstyle\int_{-\infty}^{\infty} \prod_{j \neq i} I_{F_j(y)}(r_j - c_j, n_j - (r_j - c_j + 1) dI_{F_i(y)}(r_i, n_i - r_i + 1). \end{aligned}\right\} \tag{2.8}$$

Further, (see (2.6)) $I_u(a,b)$ is, for all (a,b) with $a > 0$ and $b > 0$, strictly increasing in u, which implies by (2.1) that, when $F_i = F_{[k]}$, $F_[j](x) \geq F_i(x)$ for all j and all x. So, with $L_{i,d_i}(CS|R_1) = \min_{(F_1,\ldots,F_k)} P_{i,d_i}(CS|R_1)$, where the minimum is taken over all $(F_1, \ldots, F_k)$ satisfying (2.1),

$$\left.\begin{aligned} & P_{i,d_i}(CS|R_1) \geq L_{i,d_i}(CS|R_1) = \\ & \textstyle\int_{-\infty}^{\infty} \prod_{j \neq i} I_{F_i(y)}(r_j - c_j, n_j - (r_j - c_j + 1) dI_{F_i(y)}(r_i, n_i - r_i + 1) = \\ & \textstyle\int_0^1 \prod_{j \neq i} I_u(r_j - c_j, n_j - (r_j - c_j) + 1) dI_u(r_i, n_i - r_i + 1). \end{aligned}\right\} \tag{2.9}$$

Next note that for $u \in (0,1)$

$$I_u(r, n - r + 1) - I_u(r + 1, n - r) = \binom{n}{r} u^r (1 - u)^{n-r} > 0 \tag{2.10}$$

implying that $I_u(r, n - r + 1)$ is decreasing in r for fixed $u \in (0,1)$ and fixed $n \geq 1$. Then, using (2.10) and the fact that $0 \leq c_i \leq r_i - 1$, gives, for $i = 1, \ldots, k$,

$$A_i \leq L_{i,d_i}(CS|R_1) \leq B_i, i = 1, \ldots, k, \tag{2.11}$$

where

$$A_i = \int_0^1 \prod_{j \neq i} I_u(r_j, n_j - r_j + 1) dI_u(r_i, n_i - r_i + 1) \tag{2.12}$$

and

$$B_i = \int_0^1 \prod_{j \neq i} I_u(1, n_j) dI_u(r_i, n_i - r_i + 1). \tag{2.13}$$

Now we need to find $(c_1, \ldots, c_k)$ and the possible P^*'s such that

$$\min_{1 \leq i \leq k} L_{i,d_i}(CS|R_1) \geq P^*. \tag{2.14}$$

To solve this problem, first note that the subset size is increasing in each of the c_i, so the c's should be chosen as small as possible. Further note (from (2.13)) that $\min_{1 \leq i \leq k} B_i < 1$ and from (2.12) (see the Appendix for a proof) that

$$\sum_1^k A_i = 1 \text{ for all } n_i \text{ and } r_i \text{ satisfying (2.2)}, \tag{2.15}$$

implying that $\min_{1 \leq i \leq k} A_i \leq 1/k$.
Now let $A_{i_0} = \min_{1 \leq k} A_i$ and let $B_{i_1} = \min_{1 \leq i \leq k} B_i$. Then $A_{i_0} < B_{i_1}$ and the following theorem holds

Theorem 2.1 *From the above it follows that*

1) when $P^* < A_{i_0}$, $L_{i,d_i}(CS|R_1) > P^*$ for all (i, d_i), but with $P^* = A_{i_0}$ one gets $L_{i,d_i}(CS|R_1) \geq A_{i_0}$ for all (i, d_i);

2) when $P^* > B_{i_1}$, then (by (2.11)

$$P_{i_1,d_{i_1}}(CS|R_1) < P^* \textit{ for all } d_{i_1};$$

3) when $A_{i_0} \leq P^* \leq B_{i_1}$ there exists (by (2.11) and the monotonicity in c_i of $L_{i,d_i}(CS|R_1)$) a d_i such that $P_{i_1,d_{i_1}}(CS|R_1) \geq P^*$ for all d_{i_1}.

For the case where the n_i are equal, we have $r_i = r$, $L_{i,d_i}(CS|R_1) = L(CS|R_1)$, $A_i = A = 1/k$ and $B_i = B < 1$. So, the interval (2.11) is the interval $[1/k, B] \subset [1/k, 1)$ and Theorem 2.1 tells us that in this case there exists, for each $P^* \in [1/k, 1)$ at least one c such that $P(CS|R_1) \geq P^*$. And this is essentially the RS solution for this case - they take $P^* \in (1/k, 1)$. But who would use $P^* = 1/k$?

3 The procedure for the smallest quantile

As in Section 2, the conditions on the F_i are the ones used by RS, i.e. the F_i satisfy

$$\max_{1 \leq i \leq k} F_{[i]}(y) = F_{[1]}(y) \text{ for all } y. \tag{3.1}$$

Further, as before, the F_i are continuous and have a unique α-quantile, but in this case the integers r_i and c_i satisfy

$$1 \leq r_i \leq (n_i+1)\alpha < r_i + 1 \leq n_i + 1 \text{ and } 0 \leq c_i \leq n_i - r_i. \tag{3.2}$$

The proposed procedure is then

$$R_2 : \text{put } F_i \text{ in the subset } \Leftrightarrow Y_{r_i,i} \leq \min_{1\leq j\leq k,\, j\neq i} Y_{r_j+c_j,j} \tag{3.3}$$

and when $F_i = F_{[1]}$ the probability of correct selection is

$$\left.\begin{array}{rcl} P_{i,d_i}(CS|R_2) & = & P(Y_{r_i,i} \leq \min_{1\leq j\leq k, j\neq i} Y_{r_j+c_j,j}) \\ & = & 1 - P(Y_{r_j+c_j,j} \leq Y_{r_i,i}, j \neq i), \end{array}\right\} \tag{3.4}$$

where, as for the case of the largest quantile, the probability of correct selection when $F_i = F_{[1]}$ depends on the c's only through $d_i = (c_1, \ldots, c_{i-1}, c_{i+1}, c_k)$. Using (2.5) now gives, for the case where $F_i = F_{[1]}$,

$$\left.\begin{array}{l} 1 - P_{i,d_i}(CS|R_2) = \\ \int_{-\infty}^{\infty} \prod_{j\neq i} I_{F_j(y)}(r_j + c_j, n_j - r_j - c_j + 1) dI_{F_i(y)}(r_i, n_i - r_i + 1) \end{array}\right\} \tag{3.5}$$

and (3.1) then gives

$$\left.\begin{array}{l} 1 - P_{i,d_i}(CS|R_2) \leq \\ \int_{-\infty}^{\infty} \prod_{j\neq i} I_{F_i(y)}(r_j + c_j, n_j - r_j - c_j + 1) dI_{F_i(y)}(r_i, n_i - r_i + 1) = \\ \int_0^1 \prod_{j\neq i} I_u(r_j + c_j, n_j - r_j - c_j + 1) dI_u(r_i, n_i - r_i + 1). \end{array}\right\} \tag{3.6}$$

Calling the lower bound on $P_{i,d_i}(CS|R_2)$ in (3.6) $L_{i,d_i}(CS|R_2)$, using (2.10) and the fact that $0 \leq c_i \leq n_i - r_i$, gives, for $i = 1, \ldots, k$,

$$\left.\begin{array}{l} \int_0^1 \prod_{j\neq i} I_u(r_j, n_j - r_j + 1) dI_u(r_i, n_i - r_i + 1) \\ \leq 1 - L_{i,d_i}(CS|R_2) \leq \\ \int_0^1 \prod_{j\neq i} I_u(n_j, 1) dI_u(r_i, n_i - r_i + 1). \end{array}\right\} \tag{3.7}$$

Now let, for $i = 1, \ldots, k$,

$$A_i^* = 1 - \int_0^1 \prod_{j \neq i} I_u(n_j, 1) dI_u(r_i, n_i - r_i + 1) \tag{3.8}$$

and

$$B_i^* = 1 - \int_0^1 \prod_{j \neq i} I_u(r_j, n_j - r_j + 1) dI_u(r_i, n_i - r_i + 1), \tag{3.9}$$

then we have, by the same reasoning as in Section 2,

$$\sum_{i=1}^{k} B_i^* = k - 1, \tag{3.10}$$

implying that $\max B_i^* \geq 1 - (1/k)$. Further, by the same reasoning as the one that gives (2.11),

$$A_i^* \leq L_{i,d_i}(CS|R_2) \leq B_i^*, i = 1, \ldots, k. \tag{3.11}$$

Finally, Theorem 2.1 with, for $i = 1, \ldots, k$, (A_i, B_i) replaced by (A_i^*, B_i^*), gives $(c_1, \ldots, c_k)$ and the possible P^* so that $\min_{1 \leq i \leq k} P_{i,d_i}(CS|R_2) \geq P^*$.

4 Numerical results

In this section some numerical results are presented for $k = 2$ for the case of the largest α-quantile.

We take $\alpha = 1/2$ with

1. $n_1 = n_2 = 4$ where $0 = c_1 = c_2 \leq 1$
2. $n_1 = 4$, $n_2 = 5$ where $0 \leq c_1 \leq 1$, $0 \leq c_2 \leq 2$
3. $n_1 = 4$, $n_2 = 6$ where $0 \leq c_1 \leq 1$, $0 \leq c_2 \leq 2$.

Table 1 gives, for these n_1, n_2 and α, the values of $L_{i,d_i}(CS|R_1)$ for all possible values of the c_i.

Table 1: The lower bounds in (2.9) with $k = 2$ and $i = 1, 2$

n_1	n_2	c_1	c_2	$L_{1,d_1}(CS)$	$L_{2,d_2}(CS)$
4	4	0	0	$A_1 = .5$	$A_2 = .5$
		1	1	$B_1 = .7857$	$B_2 = .7857$
4	5	0	0	$A_1 = .3572$	$A_2 = .6428$
		0	1	.5952	$A_2 = .6428$
		0	2	$B_1 = .8333$	$A_2 = .6428$
		1	0	$A_1 = .3572$	$B_2 = .8810$
		1	1	.5952	$B_2 = .8810$
		1	2	$B_1 = .8333$	$B_2 = .8810$
4	6	0	0	$A_1 = .4524$	$A_2 = .5476$
		0	1	.6667	$A_2 = .5476$
		0	2	$B_1 = .8667$	$A_2 = .5476$
		1	0	$A_1 = .4524$	$B_2 = .8333$
		1	1	.6667	$B_2 = .8333$
		1	2	$B_1 = .8667$	$B_2 = .8333$

Using Theorem 2.1 for these cases gives

1. When $n_1 = n_2 = 4$, $c = 0$ does not give anything interesting, but $c = 1$ gives a probability of correct selection = .7857.

2. When $n_1 = 4$ and $n_2 = 5$, we have $A_{i_0} = .3572$ and $B_{i_1} = .8333$, so for $P^* \in [.3572, .8333]$ there exist (c_1, c_2) such that $\min_{1\le i\le k} P_{i,d_i}(CS|R_1) \ge P^*$. For example: for $.6428 \le P^* < .8333$, $P_{1,d_1}(CS|R_1) = .8333 > .6428$ when $c_2 = 2$ and $P_{2,d_2}(CS|R_1) = .6428$ when $c_1 = 1$ - and these are the smallest c's for which the inequality holds.

3. When $n_1 = 4$ and $n_2 = 6$, we have $A_{i_0} = .4524$ and $B_{i_1} = .8333$. So, for $P^* \in [.4524, .8333]$ there exists (c_1, c_2) such that $\min_{1\le i\le k} P_{i,d_i}(CS|R_1) \ge$

P^*. For example: when (see Table 1) $P^* = .6667$, using $c_2 = 1$, $c_1 = 1$ gives $P_{1,d_1}(CS|R_1) = .6667$ and $P_{2,d_2}(CS|R_1) = .8333$. Here, again, these are the smallest c's possible.

REMARK:

For $k = 2$ we have (a proof is in the Appendix)

$$B_i = 1 - \frac{n_i!(n_1 + n_2 - r_i)!}{(n_i - r_i)!(n_1 + n_2)!}, i = 1, 2. \tag{4.1}$$

5 Some properties of the procedure R_1.

In this section we consider three questions concerning some of the properties of the procedure R_1 with unequal sample sizes as compared with those of the RS procedure.

Question 1:

Which is "better":

1) unequal sample sizes $n_1, \ldots, n_k$
or
2) equal sample sizes $n = \sum_{1 \le i \le k} n_i/k$?

As an example, take $k = 2$, $\alpha = 1/2$ and $n_1 = 4$, $n_2 = 6$ and $n = 5$.

For $n_1 = n_2 = 5$, we have $r = 3$, $0 \le c \le 2$ and (see 2.9)

$$P(CS) \ge \int_0^1 I_u(r - c, n - (r - c) + 1) dI_u(r, n - r + 1).$$

This gives

- $P(CS) \ge \frac{1}{2}$ when $c = 0$;
- $P(CS) \ge .7381$ when $c = 1$;
- $P(CS) \ge .9167$ when $c = 2$

obtained from

- for $c = 0$ the result is obvious;
- $c = 1$ gives

$$P(CS) \geq \int_0^1 I_u(2,4) dI_u(3,3) = \frac{\int_0^1 u^2(1-u)^2 \int_0^u t(1-t)^3 dt du}{B(2,4)B(3,3)},$$

  with $\int_0^u t(1-t)^3 dt = \frac{1}{2}u^2 - u^3 + \frac{3}{4}t^4 - t^5$, so that (with a little algebra) we get for this case $P(CS) \geq .7381$;

- for $c = 2$ we have (I have a formula for this one, but have not (yet) put it in the file);

$$P(CS) \geq 1 - \frac{n!(2n-r)!}{(n-r)!(2n)!} = .9167.$$

In summary, for $k = 2$, $n_1 = n_2 = 5$ and $\alpha = 1/2$, we have

- $P(CS) \geq 1/2$ when $c = 0$;
- $P(CS) \geq .7381$ when $c = 1$
- $P(CS) \geq .9167$ when $c = 2$.

For the case where $n_1 = 4$, $n_2 = 6$ and $\alpha = 1/2$ we get from Table 1 that the possible P^* are .4524, .5476, .6667 and .8333 with

- $\min_i P_{i,d_i}(CS) \geq .4524$ when $c_1 = c_2 = 0$ and when ($c_1 = 1$, $c_2 = 0$);
- $\min_i P_{i,d_i}(CS) \geq .5476$ when ($c_1 = 0$, $c_2 = 1$) when ($c_1 = 1$, $c_2 = 2$);
- $\min_i P_{i,d_i}(CS) \geq .6667$ when $c_1 = c_2 = 1$;
- $\min_i P_{i,d_i}(CS) \geq .8333$ when $(c_1 = 1, c_2 = 2)$.

Now: which of these two is “better”?

Among the reasonable P^* we have

for $n_1 = n_2 = 5$:

$$P(CS) \geq .7381 \text{ when } c = 1 \text{ and } P(CS) \geq .9167 \text{ when } c = 2.$$

for $n_1 = 4$ and $n_2 = 6$:

$P(CS) \geq .6667$ when $c_1 = c_2 = 1$ and $P(CS) \geq .8333$ when $(c_1 = 1, c_2 = 2)$.

If we want $\min_i P_{i,d_i}(cs) \geq .9$ we can not get this with $n_1 = 4$, $n_2 = 6$, but we can with $n_1 = n_2 = 5$. If we want $\min_i P_{i,d_i}(CS) \geq .8$, we have a choice between $n_1 = n_2 = 5$ with $P^* = .9167$ and $n_1 = 4$, $n_2 = 6$ with $P^* = .8333$. But for the $n_1 = n_2 = 5$ case we then have to take a large value of c and get a $P(CS)$ much larger then what we asked for. So, in this case the average subset size for $n_1 = n_2 = 5$ might well be larger than the one for $n_1 = 4$, $n_2 = 6$.

QUESTION 2:

For each of their procedures, RS prove that, if F_i is the distribution with the largest α-quantile, the probability of including it in the subset is not smaller than the probability of including any of the other distributions in the subset.

Question: is this also the case for unequal sample sizes?

For the case where $k = 2$, suppose that F_2 has the largest α-quantile. Then $F_1(x) \geq F_2(x)$ for all x,

$$\left.\begin{array}{l} P(CS) = P(F_2 \text{ is in the subset}) = \\ \int_{-\infty}^{\infty} I_{F_1(x)}(r_1 - c_1, n_1 - (r_1 - c_1) + 1) dI_{F_2(x)}(r_2, n_2 - r_2 + 1) \geq \\ \int_0^1 I_u(r_1 - c_1, n_1 - (r_1 - c_1) + 1) dI_u(r_2, n_2 - r_2 + 1) \end{array}\right\} \quad (5.1)$$

and c_1 is chosen such that

$$\int_0^1 I_u(r_1 - c_1, n_1 - (r_1 - c_1) + 1) dI_u(r_2, n_2 - r_2 + 1) \geq P^*. \quad (5.2)$$

Further,

$$\left.\begin{array}{l} P(F_1 \text{ is in the subset}) = \\ \int_{-\infty}^{\infty} I_{F_2(x)}(r_2 - c_2, n_2 - (r_2 - c_2) + 1) dI_{F_1(x)}(r_1, n_1 - r_1 + 1) \leq \\ \int_0^1 I_u(r_2 - c_2, n_2 - (r_2 - c_2) + 1) dI_u(r_1, n_1 - r_1 + 1), \end{array}\right\} \quad (5.3)$$

Finally, c_2 is chosen such that, if F_1 were the distribution with the largest α-quantile, $P(F_1$ is in the subset $) \geq P^*$, i.e. such that

$$\int_0^1 I_u(r_2 - c_2, n_2 - (r_2 - c_2) + 1) dI_u(r_1, n_1 - r_1 + 1) \geq P^*. \tag{5.4}$$

So, we get P(correct one in) $\geq P$(incorrect one in) when we have equality in (5.4). But that means that we need equality in both (5.4)and (5.2), because we do not know which one is the one with the largest α-quantile - and from the numerical results that is something that does not seem to be possible. In the case of equal sample sizes, the integrals in (5.2) and (5.4) are equal. So for that case we have P(correct one in) $\geq P$(incorrect one in) - as proved in RS for the general case of $k \geq 2$ samples of equal size.

An example with $k = 2$ where the probabilities $P(F_1$ is in the subset) and $P(F_2$ is in the subset) can be explicitly obtained is the case where

$$F_1(x) = 1 - e^{\theta_1 x} \text{ and } F_2(x) = 1 - e^{-\theta_2 x},\ 0 < x < \infty,\ \theta_1 > 0,\ \theta_2 > 0.$$

For $n_1 = 4$, $n_2 = 6$, we have (see Table 1)

$$\min A_i = \min(.4524, .5476) = .4576 \text{ and } \min B_i = \min(.8333, .8667),$$

so, by Theorem 2.1, for each $P^* \in [.4576, .8333]$ there exist (c_1, c_2) such that $\min P_{i,d_i}(CS) \geq P^*$. Choosing $P^* = .8333$, it is seen from Table 1 that we need to choose $c_1 = 1$ and $c_2 = 2$ for this P^*.

To find the values of $P_{i,d_i}(CS)$, note that (see (2.3))

$$P(F_1 \text{ is in the subset}) = P(Y_{2,1} \geq Y_{1,2}),$$

where $Y_{2,1}$ and $Y_{1,2}$ are independent with

$$Y_{2,1} \sim I_{F_1(x)}(2, 3) \text{ and } Y_{1,2} \sim I_{F_2(x)}(1, 6).$$

So (see the Appendix for details)

$$\left.\begin{aligned} P(F_1 \text{ is in the subset}) &= \textstyle\int_{-\infty}^{\infty} I_{F_2(x)}(1, 6) dI_{F_1(x)}(2, 3) \\ &= \frac{\theta_2(7\theta_1 + 6\theta_2)}{(\theta_1 + 2\theta_2)(2\theta_1 + 3\theta_2)}. \end{aligned}\right\} \tag{5.5}$$

Further,

$$P(F_2 \text{ is in the subset}) = P(Y_{3,2} \geq Y_{1,1}),$$

where $Y_{3,2}$ and $Y_{1,1}$ are independent with

$$Y_{3,2} \sim I_{F_2(x)}(3,4) \text{ and } Y_{1,1} \sim I_{F_1(x)}(1,4)$$

so that

$$\left.\begin{aligned} P(F_2 \text{ is in the subset}) &= \textstyle\int_{-\infty}^{\infty} I_{F_1(x)}(1,4) dI_{F_2(x)}(3,4) \\ &= 1 - \frac{15\theta_2^3}{(\theta_1+\theta_2)(4\theta_1+5\theta_2)(2\theta_1+3\theta_2)}. \end{aligned}\right\} \quad (5.6)$$

Now let $\lambda = \theta_1/\theta_2$ and let $\lambda > 1$. Then F_2 has, for all α, the larger α-quantile. Moreover (see Table 1 or use (5.5)) and (5.6)), when $\lambda = 1$, $P(F_1$ is in the subset$) = .8667$ and $P(F_2$ is in the subset$) = .8333$. So, as we already saw above, when $F_1(x) = F_2(x)$ for all x, F_1 is more likely to get in the subset than F_2. Question: what about $\lambda > 1$? How large does λ have to be to get F_2 more likely to get in the subset than F_1? Or, for which λ is

$$\frac{7\lambda+6}{(\lambda+2)(2\lambda+3)} > 1 - \frac{15}{(\lambda+1)(4\lambda+4)(2\lambda+3)}? \quad (5.7)$$

First note that (5.7) is equivalent to

$$h(\lambda) = 8\lambda^4 + 18\lambda^3 + 10\lambda^2 - 15\lambda - 30 > 0, \quad (5.8)$$

which is an increasing function of λ because its derivative $32\lambda^3 + 54\lambda^2 + 20\lambda - 15$ is positive for $\lambda = 1$ and increasing in λ. So, there exists exactly one $\lambda_0 > 1$ such that (see (5.8)) $h(\lambda_0) = 0$ and $h(\lambda) < 0$ for $\lambda < \lambda_0$ and positive for $\lambda > \lambda_0$. One easily finds that $\lambda_0 \approx 1.089$.
Further, from the above it follows that, when $\lambda \leq 1$ (i.e. when F_1 has the largest α-quantile), the probability that F_1 gets into the subset is larger than the probability that F_2 gets into the subset. So, the RS result that probability that the F with the largest α-quantile gets into the subset is not smaller than the probability that any of the other F_i get into the subset, does not necessarily hold when the sample sizes are not equal.

The question now arises whether, for models other than the RS one, the "best" population gets in the subset with a probability no less than the probability that any particular "non-best" gets in.

Among the results quoted in the Introduction, neither Gupta and Huang (1974), nor Chen, Dudewicz and Lee (1976), nor Hsu (1981) mention the question, but Gupta and McDonald (1970) as well as Kumar, Mehta and Kumar (2002) present models and procedures for which the inequalities hold. And it turns out that it has nothing to do with the (in)equality of the sample sizes - it is a question of the properties of the F_i. Gupta and McDonald suppose that the $X_{i,j}$ have distribution function F_{θ_i}, $\theta_i \in \Theta$, where Θ is an interval on the real line and that the family of distribution functions $\{F_{\theta_1}, \ldots, F_{\theta_k}\}$ is a stochastically increasing family. That is: $\theta_1 < \theta_2$ implies that F_{θ_1} and F_{θ_2} are distinct and $F_{\theta_2}(x) \leq F_{\theta_1}(x)$ for all x. They discuss three subset selection procedures, all of them based on linear rank statistics, for selecting a subset containing the population with the largest θ_i and show that each has the above property of putting the best population in the subset with a probability that is not smaller than putting any other particular F_{θ_i} in the subset. Note that this result implies that, for the RS model with $k = 2$, there exists a procedure with this property.
Kumar, Mehta and Kumar (2002) assume that the $X_{i,j}$ have distribution function $F(x - \theta_i)$ and present a subset selection procedure for selecting the population with the largest θ_i. Their procedure is based on U-statistics and they show that this procedure also has the property of putting the best population in the subset with a probability that is not smaller than putting any other particular F_i in the subset.

QUESTION 3:

Let S be the size of the subset. Then RS show, for their procedure for the largest α-quantile, that

$$\mathcal{E}S \leq kP(F_{[k]} \text{ is in the subset}) \tag{5.9}$$

with equality in this inequality when the F_i are identical. For their procedure for the smallest α-quantile, they show that (5.9) holds with $F_{[k]}$ replaced by $F_{[1]}$.

Question: do these inequalities also hold for our procedures when the sample sizes are not equal?

RS prove their result for the largest α-quantile as follows: first note that

$$S = \sum_{i=1}^{k} I(F_i \text{ is in the subset}).$$

So, $\mathcal{E}S = \sum_{i=1}^{k} P(F_i \text{ is in the subset})$. Further, as we saw above, for equal sample sizes we have, for $i = 1, \ldots, k$

$$P(F_i \text{ is in the subset}) \leq P(F_{[k]} \text{ is in the subset}), \tag{5.10}$$

which proves their result. But, as we also saw above, the inequality (5.10) does not necessarily hold for unequal sample sizes. For example, in the example above where $n_1 = 4$, $n_2 = 6$ and $\alpha = 1/2$, we find from (5.5) and (5.6) for the case where $\lambda = \theta_1/\theta_2 \geq 1$, i.e. when F_2 has the larger α-quantile, that the RS inequality does not hold for $1 \leq \theta_1/\theta_2 \leq 1.089$, but it does for $\lambda > 1.089$.

6 Some comments on the procedures

The comments on the above subset-selection procedures are concerned with the fact that they are not sign-invariant. That is: changing the sign of the $X_{j,i}$ should change the procedure for the largest α-quantile based on the $X_{j,i}$ into the procedure for the smallest $(1-\alpha)$-quantile based on the $-X_{i,j}$. That this is not necessarily true for the two RS procedures can be seen as follows. Suppose the n_i are equal, then the procedure for the largest α-quantile based on the $X_{j,i}$ has (see (2.2))

$$1 \leq r \leq (n+1)\alpha < r+1 \leq n+1 \text{ and } 0 \leq c \leq r-1 \tag{6.1}$$

and (see (2.1))

$$\min_{1 \leq i \leq k} F_{[i]}(y) = F_{[k]}(y) \text{ for all } y.$$

Further, the procedure is (see(2.3))

$$\text{put } F_i \text{ in the subset } \Leftrightarrow Y_{r,i} \geq \max_{1 \leq j \leq k,\, j \neq i} Y_{r-c,j}. \tag{6.2}$$

Now let, for $j = 1, \ldots, n$ and $i = 1, \ldots, k$, $U_{j,i} = -X_{j,i}$ and let, for $i = 1, \ldots, k$, $V_{1,i}, \ldots, V_{n,i}$ be the order statistics of $U_{1,i} \ldots U_{n,i}$. Then, for $i = 1, \ldots, k$,

$$H_i(x) = P(U_{j,i} \leq x) = 1 - F_i(x), \; -\infty < x < \infty, \; j = 1, \ldots, n$$

is the distribution function of the $U_{j,i}$, it satisfies

$$\max_{1\le i\le k} H_i(y) = H_{[1]}(y) \text{ for all } y$$

and the procedure for a subset containing the smallest $(1-\alpha)$-quantile based on the the $U_{j,i}$ has, for the case of equal sample sizes (see (3.2)):

$$1 \le r^* \le (n+1)(1-\alpha) < r^*+1 \le n+1 \text{ and } 0 \le c^* \le n-r^*, \qquad (6.3)$$

or, equivalently,

$$0 \le n-r^* < (n+1)\alpha \le n+1-r^* \le n \text{ and } 0 \le c^* \le n-r^*. \qquad (6.4)$$

Further, the procedure is (see(3.3))

$$\text{put } H_i \text{ in the subset } \Leftrightarrow V_{r^*,i} \le \min_{1\le j\le k,\, j\ne i} V_{r^*+c^*,j}, \qquad (6.5)$$

or, equivalently,

$$\text{put } F_i \text{ in the subset } \Leftrightarrow Y_{n-r^*+1,i} \ge \max_{1\le j\le k,\, j\ne i} Y_{n-r^*-c^*+1,j}. \qquad (6.6)$$

So, the two procedures (the one based on the $X_{i,j}$ and the one based on the $-X_{i,j}$) are the same if and only if $n-r^*+1=r$ and $n-r^*+1=r-c$, or, equivalently,

$$r^* = n-r+1 \text{ and } c^* = n-r^*+1-r+c = c.$$

But, by (6.4) we then have

$$1 \le r < (n+1)\alpha+1 \le r+1 \le n+1 \qquad (6.7)$$

which is equivalent to (6.1) if and only if $(n+1)\alpha$ is an integer in which case $r=(n+1)\alpha$. In case $(n+1)\alpha$ is not an integer, $r=[(n+1)\alpha)]$ by (6.1) and $r-1=[(n+1)\alpha)]$ by (6.7).

ACKNOWLEDGMENT

The author thanks her good friend and co-author James V. Zidek for proposing she work on this problem and for his encouragement during the preparation of this paper. Many thanks also for the help with the numerical results.

REFERENCES

Chen, H. J., Dudewicz, E.J. and Lee, Y.J. (1976). Subset selection procedures for normal means under unequal sample sizes. Sankhyā, B, 38, 249-255.

Dudewicz, E.J. (1974). A note on selection procedures with unequal observation numbers. Zastosow. Matem., XIV, 32-35.

Gupta, S.S. (1956). On a decision rule for a problem in ranking means. Institute of Statistics, Mineograph Series No. 150, University of North Carolina, Chapel Hill, North Carolina.

Gupta, S.S. and Huang, W-T. (1974). A note on selecting a subset of normal populations with unequal sample sizes. Sankhyā, A, 36, 389-396.

Gupta, S.S. and McDonald, G.C. (1970). On some classes of selection procedures based on ranks. In: "Nonparametric Techniques in Statistical Inference", Conference Proceedings, (M.L. Puri, Editor), Cambridge U niversity Press, p. 491-514.

Gupta, S.S. and Sobel, M. (1957). On a statistics which arises in selection and ranking problems. Ann. Math. Statist., 28, 957-967.

Hsu, J.C. (1981). A class of nonparametric subset selection procedures. Sankhyā, B, 43, 235-244.

Kumar, N., Mehta, G.P. and Kumar, V. (2002). Two new classes of subset selection procedures for location parameters. Statist. Decisions, 20, 415-427.

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.

Sitek, M. (1972). Application of the selection procedure R to unequal observations numbers. Zastosow. Matem., XII, 355-363.

A Appendix

Proof of (2.15):

Let, for $0 \leq u \leq 1$, $G(u) = \prod_{j=1}^{k} I_u(r_j, n_j - r_j + 1)$. Then

$$\sum_{1 \leq i \leq k} \int_0^1 \prod_{j \neq i} I_u(r_j - 1, n_j - r_j + 1) dI_u(r_i, n_i - r_i + 1) =$$

$$\sum_{1 \leq i \leq k} \int_0^1 G(u) \frac{dI_u(r_i, n_i - r_i + 1)}{I_u(r_i, n_i - r_i + 1)} =$$

$$\int_0^1 G(u) d \log G(u) = 1,$$

because $G(u)$ is an absolutely continuous distribution function. ♡

Proof of (4.1)

From (2.13) it follows that, for $k = 2$, B_1 is given by

$$\int_0^1 I_u(1, n_2) dI_u(r_1, n_1 - r_1 + 1) =$$

$$\frac{\int_0^1 \left(1 - (1-u)^{n_2}\right) u^{r_1 - 1}(1-u)^{n_1 - r_1}}{B(1, n_2) B(r_1, n_1 - r_1)} =$$

$$\frac{n_1!}{(r_1 - 1)!(n_1 - r_1)!} \left(B(r_1, n_1 - r_1 + 1) - B(r_1, n_1 + n_2 - r_1 + 1)\right) =$$

$$\frac{n_1!}{(r_1 - 1)!(n_1 - r_1)!} \left(\frac{\Gamma(r_1)\Gamma(n_1 - r_1 + 1)}{\Gamma(n_1 + 1)} - \frac{\Gamma(r_1)\Gamma(n_1 + n_2 - r_1 + 1)}{\Gamma(N - 1 + n_2 + 1)} \right) =$$

$$1 - \frac{n_1!}{(n_1 - r_1)!} \frac{(n_1 + n_2 - r_1)!}{(n_1 + n_2)!}.$$

And this implies that

$$B_2 = 1 - \frac{n_2!}{(n_2 - r_2)!} \frac{(n_1 + n_2 - r_2)!}{(n_1 + n_2)!}.$$

♡

Proof of (5.5)

$$P(F_1 \text{ is in the subset}) = \int_{-\infty}^{\infty} I_{F_2(x)}(1,6) dI_{F_1(x)}(2,3) =$$

$$\frac{\Gamma(5)}{\Gamma(2)\Gamma(3)} \int_{-\infty}^{\infty} F_1(x)(1-F_1(x))^2 \left(1-(1-F_2(x))^6\right) dF_1(x) =$$

$$12 \int_{-\infty}^{\infty} (1-e^{-\theta_1 x}) e^{-2\theta_1 x} \theta_1 e^{-\theta_1 x} (1-e^{-6\theta_2}) dx =$$

$$12\theta_1 \int_{-\infty}^{\infty} \left(e^{-3\theta_1 x} - e^{-4\theta_1 x} - e^{-(3\theta_1+6\theta_2)x} + e^{(4\theta_1+6\theta_2)x}\right) dx =$$

$$12\theta_1 \left(\frac{1}{3\theta_1} - \frac{1}{4\theta_1} - \frac{1}{3(\theta_1+2\theta_2)} + \frac{1}{2(2\theta_1+3\theta_2)}\right) =$$

$$\frac{\theta_2(7\theta_1+6\theta_2)}{(\theta_1+2\theta_2)(2\theta_1+3\theta_2)}.$$

♡

Proof of (5.6)

$$P(F_2 \text{ is in the subset}) = \int_{-\infty}^{\infty} I_{F_1(x)}(1,4) dI_{F_2(x)}(3,4) =$$

$$\frac{\Gamma(5)}{\Gamma(4)} \frac{\Gamma(7)}{\Gamma(3)\Gamma(4)} \int_{-\infty}^{\infty} F_2(x)^2 (1-F_2(x))^3 \int_0^{F_1(x)} (1-t)^3 dt =$$

$$\frac{6!}{2!3!} \int_{-\infty}^{\infty} (1-(1-F_1(x))^4) F_2(x)^2 (1-F_2(x))^3 dF_2(x) =$$

$$60\theta_2 \left(\frac{1}{60\theta_2} - \frac{1}{4(\theta_1 + \theta_2)} + \frac{2}{4\theta_1 + 5\theta_2} - \frac{1}{4\theta_1 + 6\theta_2} \right)$$

$$1 - \frac{15\theta_2^3}{(\theta_1 + \theta_2)(4\theta_1 + 5\theta + 2)(2\theta_1 + 3\theta_2)}.$$