THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT #263

SUBSET SELECTION – EXTENDING RIZVI-SOBEL

BY

CONSTANCE VAN EEDEN
JAMES V ZIDEK

February 2011

SUBSET SELECTION - EXTENDING RIZVI-SOBEL

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

Abstract

This paper presents a non-parametric procedure for selecting a subset of a set of k populations, which contains the one with the largest (L) or smallest (S) α^{th} quantile, as specified by the user, when independent samples are available from each and one population is the uniformly correct choice whatever be α. The result, an extension of a method previously proposed for the case of equal sample sizes, includes population i, if its α^{th} sample quantile exceeds (in the case of L) the largest of all the other sample $(\alpha - \beta)^{\text{th}}$ quantiles for the other populations, where $0 < \beta < \alpha$. The selection index β is specified by the user. An obvious adaptation of this rule covers S. The paper includes an asymptotic theory for the method, which gives a practical way of selecting the selection index by optimizing an objective criterion, a linear combination of the probability of correct selection, which ideally should be large, and the expected subset size, which ideally should be small. Furthermore, the criterion provides a way of selecting the sample sizes in practical situations where the cost of obtaining the samples will differ for the different populations.

Keywords and Phrases: Nonparametric subset selection, unequal sample sizes, quantiles
AMS 2000 classifications: 62F07, 62G99

1 Introduction

This paper presents an extension of a subset selection method proposed by Rizvi and Sobel (1967). It also includes strategies for implementing the method. Before introducing the work in this paper below, we describe its genesis, a problem encountered by the second author in work performed under a nondisclosure agreement involving proprietary data. That work concerned the grouping of populations of manufactured lumber for the purpose of marketing them under a single engineering design standard. These populations could for example represent different geographical regions. Creating such groups is done according to protocols specified in a standards document published by the American Society for Testing and Measurement and labelled ASTM D 1990. The protocols stipulate that a subset of the populations, the so–called "controlling species", be formed. Engineering design values are then based on the controlling species, the idea being that they represent the most conservative choice from a reliability perspective. This approach builds robustness into the published design values against the possibility of species being withdrawn from the overall collection of populations.
The protocols prescribe statistical methods for finding the controlling species. These are nonparametric in nature to assure that ASTM D 1990 is broadly applicable. Moreover, they are based on combinations of statistical testing procedures such as the chi-squared test, the Kruskal - Wallis nonparametric method for the analysis of variance and Tukey's method of paired comparisons. Which combination is used depends on whether the design values are based on the median or the 5^{th} percentile. Although ASTM D 1990 has served effectively for a long time, it was recently found to produce unexpected results, leading to a search for alternatives.

The subset selection procedure of Rizvi and Sobel (1967) (RS from now on) is based on $k \geq 2$ independent samples $X_{1,i}, \ldots, X_{n_i,i}$ from distributions F_i, $i = 1, \ldots, k$. They give a subset selection procedure for selecting the F_i with the largest α^{th} quantile, as well as a procedure for selecting the one with the smallest α^{th} 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 $P^* \in (1/k, 1)$. Here "correct selection" means selecting a subset containing the population with the largest α^{th} quantile (resp. the smallest α^{th} quantile). However the RS requirement of equal samples sizes is a serious practical limitation, which is removed in this paper.

Several other procedures for subset selection with unequal sample sizes have been proposed. Sitek (1972) generalizes, to unequal sample sizes, a procedure of Gupta (1956) (see also Gupta and Sobel (1957)) for selecting a subset containing the population with the largest θ_i when $X_{j,i} \sim \mathcal{N}(\theta_i, \sigma^2)$ and σ is known. However, Dudewicz (1974) shows Sitek's derivation to be incorrect. Then 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) further consider the Gupta-Huang (1974) case. They propose a procedure which is different from the Gupta-Huang procedure and compare the two

methods with respect to average subset size and ease of implementation. However, these methods are parametric in nature and hence unsuitable for the situation confronted in this paper.
In contrast, several non-parametric procedures have been proposed, 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 consider models that resemble the one on which RS is based although Hsu unlike Gupta and McDonald fails to refer to RS. These complexity of these procedures relative to RS led to their rejection in the application addressed in the paper.
The paper is organized as follows. Section 2 describes our extension of the RS procedure. Section 3 lays out the asymptotic foundations we need and these are implemented in Section 4, which leads to a practical way of implementing our extension of the RS method. The next section (5) demonstrates through numerical examples how the asymptotic theory can be used. Lessons learned from the numerical results are summarized in Section 6. The paper wraps up with Section 7 save for some technical details in the Appendix.

2 Procedure for the largest quantile

This section describes our extensions of the RS procedures to the case where the sample sizes are not necessarily equal, and it presents their needed properties. Only partial proofs are given. Complete proofs can be found in van Eeden (2009).

The basic conditions on the distribution functions F_i are the same as those of RS. In particular, we assume

Assumption 2.1 *The F_i are continuous and strictly increasing and there exists a $\tau \in \{1, \ldots, k\}$ such that, for all y,*

$$F_\tau(y) < F_j(y), j \neq \tau. \tag{2.1}$$

For continuous, strictly increasing F_i, Assumption 2.1 holds if and only if for all $u \in (0, 1)$ and $j \neq \tau$,

$$F_\tau(F_j^{-1}(u)) \ < \ u \ < \ F_j(F_\tau^{-1}(u)). \tag{2.2}$$

In the notation of RS, $F_\tau = F_{[k]}$, where for, $i = \{1, \ldots, k\}$, $F_{[i]}$ is the distribution with the i^{th} smallest α^{th} quantile.
Further, r_i and c_i, $i = 1, \ldots, k$, are integers satisfying

$$1 \leq r_i \leq (n_i + 1)\alpha < r_i + 1 \leq n_i + 1 \text{ and } 0 \leq c_i \leq r_i - 1 \tag{2.3}$$

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 described as follows:

$$R_1: \text{ put } F_i \text{ in the subset } \Leftrightarrow Y_{r_i,i} \geq \max_{1\leq j\leq k,\, j\neq i} Y_{r_j-c_j,j}, \tag{2.4}$$

so that, when $F_{[k]} = F_\tau$ for some (unknown) $\tau \in \{1,\ldots,k\}$, the probability of correct selection is given by given by

$$\left.\begin{array}{l} P_{\tau,d_\tau}(CS|R_1) \\ = P(Y_{r_\tau,\tau} \geq \max_{j\neq\tau} Y_{r_j-c_j,j}) = P(Y_{r_j-c_j,j} \leq Y_{r_\tau,i}, j \neq \tau) \\ = \int_{-\infty}^{\infty} \prod_{j\neq\tau} I_{F_j(y)}(r_j - c_j, n_j - r_j + c_j + 1) dI_{F_\tau(y)}(r_\tau, n_\tau - r_\tau + 1). \end{array}\right\} \tag{2.5}$$

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

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

The problem now is to find $c_1,\ldots,c_k$ and all possible $P^* \in (0,1)$, such that $P_{\tau,d_\tau}(CS|R_1) \geq P^*$ for all $(F_1,\ldots,F_k)$ satisfying Assumption 2.1 and all $\tau \in \{1,\ldots,k\}$, i.e. we need

$$L_{i,d_i}(CS|R_1) = \min_{(F_1,\ldots,F_k)} P_{i,d_i}(CS|R_1) \geq P^*, \tag{2.7}$$

where the minimum is taken over all $(F_1,\ldots,F_k)$ satisfying Assumption 2.1. That assumption implies for all $(F_1,\ldots,F_k)$ that

$$\left.\begin{array}{l} P_{i,d_i}(CS|R_1) \geq L_{i,d_i}(CS|R_1) = \\ \int_{-\infty}^{\infty} \prod_{j\neq i} I_{F_j(y)}(r_j - c_j, n_j - (r_j - c_j) + 1) dI_{F_i(y)}(r_i, n_i - r_{i+1}) = \\ \int_0^1 \prod_{j\neq i} I_u(r_j - c_j, n_j - (r_j - c_j) + 1) dI_u(r_i, n_i - r_i + 1). \end{array}\right\} \tag{2.8}$$

Now note that, for fixed $u \in (0,1)$ and fixed $n \geq 1$, $I_u(r, n-r+1)$ is strictly increasing in r. Also for each $i \in \{1,\ldots,k\}$, $0 \leq c_i \leq r_i - 1$, so that

$$A_i \leq L_{i,d_i}(CS|R_1) \leq B_i, i = 1,\ldots,k, \tag{2.9}$$

where

$$A_i = \int_0^1 \prod_{j\neq i} I_u(r_j, n_j - r_j + 1) dI_u(r_i, n_i - r_i + 1) \tag{2.10}$$

and

$$B_i = \int_0^1 \prod_{j\neq i} I_u(1, n_j) dI_u(r_i, n_\tau - r_i + 1). \tag{2.11}$$

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

$$\min_{1\leq i\leq k} L_{i,d_i}(CS|R_1) \geq P^*. \tag{2.12}$$

To find these quantities note that the subset size is increasing in each of the c's, so they should be chosen as small as possible. Further note that $\min_{1\le i\le k} B_i < 1$ and

$$\sum_{1}^{k} A_i = 1 \text{ for all } n_i \text{ and } r_i \text{ satisfying (2.3)}, \tag{2.13}$$

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

Theorem 2.1 *The results above imply 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 $L_{i,d_i}(CS|R_1)$ $\ge A_{i_0}$ for all (i, d_i);

2) when $P^* > B_{i_1}$, then (by (2.9)

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

3) when $A_{i_0} \le P^* \le B_{i_1}$, the monotonicity of $L_{i,d_i}(CS \mid R_1)$ in each of $c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_k$ and (2.9), imply that there exists, for each i, a d_i such that $L_{i,d_i}(CS \mid R_1) \ge P^*$.

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 (2.9) 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 $P(CS|R_1) \ge P^*$. And this is essentially the RS solution for this case - they take $P^* \in (1/k, 1)$. But who would use $P^* = 1/k$?

Some of the questions that need looking into when designing an experiment to be analyzed using RS, or using the RS procedure with possibly unequal sample sizes, are:

1. In cases where the cost of getting observations varies between the populations and we know these costs, which set of possibly unequal sample sizes is the cheapest way to get the needed P^*? This kind of question will be looked into in Section 4.

2. Is the probability of correct selection $P(CS|R_1)$ larger than $P(ICS|R_1)$, the probability of an incorrect selection, that is a selection containing only populations whose α^{th} quantile is not the largest? RS showed that for equal sample sizes $P(CS|R_1) \ge P(ICS|R_1)$. However, for unequal sample sizes this is not the case. An example can be found in van Eeden (2009). She looked at the case where $k = 2$ with (see Table 1) $n_1 = 4$, $n_2 = 6$, $c_1 = 1$, $c_2 = 2$, $P^* = .8333$ and

   $$F_1(x) = 1 - e^{-\theta_1 x} \text{ and } F_2(x) = 1 - e^{-\theta_2 x}.$$

   Then F_2 has, for all α^{th}, the largest α quantile if and only if $\theta_1 > \theta_2$ and she shows that, for $\alpha = .5$, $P(F_1 \text{ in the subset}) > P(CS|R_1)$ if and only if $\theta_1/\theta_2 \le 1.089$, i.e. $P(ICS|R_1) > P(CS|R_1)$ if and only if $1 < \theta_1/\theta_2 < 1.089$.

3. For a given $N = kn$, which is "better": equal sample sizes or unequal ones - and if unequal ones, which unequal ones? Of course, this depends on what one means by "better". For instance, one could ask for more and (or) larger possible values of P^* or one could ask the lower bound $L_{i,d_i}(CS|R_1)$ on $P(CS|R_1)$ to be as close as possible to one's preferred P^*. In the case where $k = 2$ and $N = 10$, Table 1 uses Theorem 2.1 for $n_1 = n_2 = 5$ and $n_1 = 4$ and $n_2 = 6$ to make such a comparison when $\alpha = .5$. From this table we see that, if our $P^* = .90$, then we have to use $n_1 = n_2 = 5$ which gives us, with $c_1 = c_2 = 2$, $P(CS|R_1) \geq .9167$, while for $n_1 = 4$, $n_2 = 6$ the largest possible $P(CS|R_1)$ one can get is .8333 by using $c_1 = 1$, $c_2 = 2$. In the case where we want to use $P^* = .80$, one can use either $n_1 = n_2 = 5$ or $n_1 = 4$, $n_2 = 6$, giving, respectively, $P(CS|R_1) \geq .9167$ with $c_1 = c_1 = 2$ or $n_1 = 4$, $n_2 = 6$, giving $P(CS|R_1) \geq .8333$ with $c_1 = 1$, $c_2 = 2$. However, if one wants to use $P^* = .80$ and asks that $P(CS|R_1)$ be as close as possible to .80, then $n_1 = 4$, $n_2 = 6$ with $c_1 = 1$, $c_2 = 2$ is the only solution.

Table 1: Some lower bounds in (2.8) with $k = 2$ and $i = 1, 2$

n_1	n_2	c_1	c_2	$L_{1,d_1}(CS\|R_1)$	$L_{2,d_2}(CS\|R_1)$
5	5	0	0	$A_1 = .5$	$A_2 = .5$
		1	1	.7380	.7380
		2	2	$B_1 = .9167$	$B_2 = .9167$
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	$B_1 = .8667$	$B_2 = .8333$

4. Let $S \subset \{F_1, \ldots, F_k\}$ be the subset obtained by our extension of the RS subset-selection procedure for the largest α^{th} quantile. When the sample sizes are large, approximations to $P(F_j \in S)$ or computer programs to compute it are needed to answer the above questions. Note that for $j = \tau$ this probability is the probability of correct selection. Taking all F_i to be uniform distributions on $(0, 1)$ gives the lower bounds in (2.8). Approximations are given in Section 3.

Remark

A procedure for selecting a subset containing the population with the smallest α-quantile can be found in van Eeden (2010). That Technical Report is a corrected version of Section 3 of van Eeden (2009). The procedure described there is, essentially, the procedure for the largest $(1-\alpha)^{\text{th}}$ quantile applied to $-X_{1,i},\ldots,-X_{n_i,i}$, $i=1,\ldots,k$.

3 Approximations for large sample sizes

In this section large-sample-size approximations are given for the probability of correct selection $P_{\tau,d_\tau}(CS|R_1)$ and its lower bound $L_{i,d_i}(CS|R_1)$ and, more generally, for $P(F_j \in S)$. Note that Assumption 2.1 implies that, for all $i=1,\ldots,k$, $F_i(X_{j,i})$, $j=1,\ldots,n_i$, are samples from uniform distributions on $(0,1)$. So, the needed approximations can be obtained from approximations to the distributions of the order statistics from such uniform distributions. Further, the r-th order statistic of a sample of size n then has a $B(r,n-r)$ distribution, i.e. it has density is given by

$$\frac{1}{B(r,n-r+1)}u^{r-1}(1-u)^{n-r}\ 0<u<1.$$

An approximation to the $B(\nu_1,\nu_2)$ distribution for large values of the ν_i is given in the next lemma.

Lemma 3.1 *Let T have a $B(\nu_1,\nu_2)$ distribution, i.e., let T have density*

$$\frac{1}{B(\nu_1,\nu_2)}u^{\nu_1-1}(1-u)^{\nu_2-1}\ 0<u<1.$$

Then, as ν_1 and ν_2 converge to infinity,

$$P(T\le t)\to\Phi\left(\frac{-\nu_1(1-t)+\nu_2 t}{\sqrt{\nu_1(1-t)^2+\nu_2 t^2}}\right)\ 0<t<1. \tag{3.1}$$

Proof. As is well-known, when Γ_{ν_1} and Γ_{ν_2} are independent random variables with densities

$$\frac{1}{\Gamma(\nu_i)}x^{\nu_i-1}e^{-x},\ x>0,\ i=1,2,$$

then

$$Z_{\nu_1,\nu_2}=\frac{\Gamma_{\nu_1}}{\Gamma_{\nu_1}+\Gamma_{\nu_2}}$$

has a $B(\nu_1,\nu_2)$ distribution, i.e. Z_{ν_1,ν_2} has density

$$\frac{1}{B(\nu_1,\nu_2)}z^{\nu_1-1}(1-z)^{\nu_2-1},\ 0<z<1.$$

Using the fact that, for large ν, Γ_ν is approximately $\mathcal{N}(\nu,\nu)$, it follows that, as ν_1 and ν_2 go to ∞ and $z \in (0,1)$

$$P(Z_{\nu_1,\nu_2} \le z) \to \Phi\left(\frac{-[\nu_1(1-z)-\nu_2 z]}{\sqrt{\nu_1(1-z)^2+\nu_2 z^2}}\right). \tag{3.2}$$

♡

The approximation (3.2) can be used to find, for given k, $n_1,\ldots,n_k$, P^* and α, approximations to the c_i such that, approximately, $P(CS|R_1) \ge P^*$. By (2.8), we have

$$L_{i,d_i}(CS|R_1) = P(U_{r_j-c_j,j} \le U_{r_i,i}, j \ne i),$$

where, by Lemma 3.1,

$$P(U_{r_j-c_j,j} \le t) \approx \Phi\left(\frac{-[(r_j-c_j)(1-t)-(n_j-(r_j-c_j)+1)t]}{\sqrt{(r_j-c_j)(1-t)^2+(n_j-(r_j-c_j)+1)t^2}}\right)$$

and

$$P(U_{r_i,i} \le t) \approx \Phi\left(\frac{-[r_i(1-t)-(n_i-r_i+1)t]}{\sqrt{r_i(1-t)^2+(n_i-r_i+1)t^2}}\right),$$

so that

$$\begin{aligned} &L_{i,d_i}(CS|R_1) \approx \\ &\textstyle\int_0^1 \prod_{j\ne i} \Phi\left(\frac{-[(r_j-c_j)(1-u)-(n_j-(r_j-c_j)+1)u]}{\sqrt{(r_j-c_j)(1-u)^2+(n_j-(r_j-c_j)+1)u^2}}\right) d\Phi\left(\frac{-[r_i(1-u)-(n_i-r_i+1)u]}{\sqrt{r_i(1-u)^2+(n_i-r_i+1)u^2}}\right). \end{aligned} \tag{3.3}$$

Remarks:

1. *The exact expression for $L_{i,d_i}(CS|R_1)$ is increasing in each of the c_i. As shown in the Appendix, the approximation (3.3) has this same property.*

2. *Note that the approximation (3.2) to the beta distribution, which is a distribution on the interval $(0,1)$, is a distribution on the interval $(-\epsilon_{\nu_1},\epsilon_{\nu_2})$ for positive ϵ_{ν_i} which converge to 0 as the ν_i converge to infinity.*

In Section 4, an approximation to $P(F_l \in S \mid R_1)$, for $l = 1,\ldots,k$, is needed. First note that this probability is given by (see (2.4))

$$P(F_l \in S|R_1) = P(Y_{r_i-c_i,i} \le Y_{r_l,l}, i \ne l), \tag{3.4}$$

where, for $\nu = 1,\ldots,n_i$ and $i = 1,\ldots,k$, $Y_{\nu,i}$ are the order statistics of a sample of size n_i from the distribution F_i. So, with $H_{il}(u) = F_i(F_l^{-1}(u))$, (3.4) can be written as

$$P(H_{li}(U_{r_i-c_i,i}) \le U_{r_l,l}, i \ne l) = P(U_{r_i-c_i} \le H_{il}(U_{r_l,l})),$$

where, for $i = 1, \ldots, k$ and $\nu = 1, \ldots, n_i$, $U_{\nu,i}$ are the order statistics of a sample of size n_i from a uniform distribution on $(0, 1)$. As a result,

$$P(F_l \in S|R_1) = \int_0^1 \prod_{i \neq l} P(V_i \leq H_{il}(u) dP(W_l \leq u), \tag{3.5}$$

where, for $i = 1, \ldots, k$, V_i are independent $Beta(r_i - c_i, n_i - (r_i - c_i) + 1)$ random variables and for, $l = l, \ldots, k$, W_l are independent $Beta(r_l, n_l + 1)$ random variables. So (3.2) implies that

$$\left.\begin{array}{rcl} P(V_i \leq H_{il}(u)) & \approx & \Phi\left(\frac{-[(r_i-c_i)(1-H_{il}(u))-(n_i-(r_i-c_i)+1)H_{il}(u)]}{\sqrt{(r_i-c_i)(1-H_{il}(u))^2+(n_i-(r_i-c_i)+1)H_{il}(u)^2}}\right) \\ & = & \Phi\left(\frac{\sqrt{n_i+1}(H_{il}(u)-K_{i,n_i})}{\sqrt{K_{i,n_i}(1-H_{il}(u))^2+(1-K_{i,n_i})H_{il}(u)^2}}\right), \end{array}\right\} \tag{3.6}$$

where $K_{i,n_i}(n_i + 1) = r_i - c_i$ for $i = 1, \ldots, k$.
As well, Lemma 3.1 implies

$$P(W_l \leq u) \approx \Phi\left(\frac{-[r_l(1-u) - (n_l - r_l + 1)u]}{\sqrt{r_l(1-u)^2 + (n_l - r_l + 1)u^2}}\right). \tag{3.7}$$

Since $H_{il}(u)$ is increasing

$$P(V_i \leq H_{il}(u)) \approx I\{H_{il}(u) > K_{i,n_i}\}$$

where I denotes the indicator function.
Thus, with $G_l = \max_{i \neq l} H_{il}^{-1}(K_{i,n_i}) = \max_{i \neq l} H_{li}(K_{i,n_i})$,

$$P(F_l \in S) =$$

$$\int_0^1 \Pi_{i \neq l} I\{H_{il}(u) > K_{i,n_i}\} dP(W_l \leq u) =$$

$$\int_0^1 \Pi_{i \neq l} I\{u > H_{il}^{-1}(K_{i,n_i})\} dP(W_l \leq u) =$$

$$\int_0^1 I\{u > G_l\} dP(W_l \leq u) =$$

$$P[W_l > G_l] \approx$$

$$1 - \Phi\left(\frac{-[r_l(1-G_l) - (n_l - r_l + 1)G_l]}{\sqrt{r_l(1-G_l)^2 + (n_l - r_l + 1)G_l^2}}\right) \approx$$

$$1 - \Phi\left(\frac{-\alpha(n_l+1) + (n_l+1)G_l}{\sqrt{\alpha(n_l+1)(1-G_l)^2 + (n_l - \alpha(n_l+1) + 1)G_l^2}}\right) =$$

$$1 - \Phi\left(\frac{\sqrt{(n_l+1)}[-\alpha + G_l]}{\sqrt{\alpha[1-G_l]^2 + (1-\alpha)G_l^2}}\right).$$

The previous expression gives, by (3.5), (3.6) and (3.7), the following approximation to $P(F_l \in S|R_1)$:

$$\left.\begin{array}{l} P(F_l \in S|R_1) \approx \\ 1 - \Phi\left(\frac{\sqrt{(n_l+1)}[-\alpha+\max_{i\neq l} H_{li}(K_{i,n_i})]}{\sqrt{\alpha[1-\max_{i\neq l} H_{li}(K_{i,n_i})]^2+(1-\alpha)[\max_{i\neq l} H_{li}(K_{i,n_i})]^2}}\right). \end{array}\right\} \tag{3.8}$$

Now assume that, for $i = 1, \ldots, k$, $c_i/n_i \to \beta_i$ for $\beta_1, \ldots, \beta_k$ satisfying $0 < \beta_i < \alpha$. Then $(r_i - c_i)/n_i \to \alpha - \beta_i > 0$ for $i = 1, \ldots, k$ and the approximation (3.8) becomes

$$\left.\begin{array}{l} P(F_l \in S|R_1) \approx \\ 1 - \Phi\left(\frac{\sqrt{(n_l+1)}[-\alpha+\max_{i\neq l} H_{li}(\alpha-\beta_i)]}{\sqrt{\alpha[1-\max_{i\neq l} H_{li}(\alpha-\beta_i)]^2+(1-\alpha)[\max_{i\neq l} H_{li}(\alpha-\beta_i)]^2}}\right). \end{array}\right\} \tag{3.9}$$

These approximations are used in Section 4.

4 Optimal subsets

Recall that in our extension of the Rizvi - Sobel method, S is found according to the rule:

$$F_i \in S \;\Leftrightarrow Y_{r_i,i} \geq \max_{1\leq j\leq k,\ j\neq i} Y_{r_j-c_j,j} \tag{4.10}$$

for integers r_i and c_i, satisfying (2.3) that must be specified by the user who may also need to specify the n_i. An ideal subset selection procedure would maximize the probability of correctly selecting the one labelled $\tau \in \{1, \ \ldots, \ k\}$, whose α^{th} quantile ξ_τ is the largest (or smallest depending on the context). That goal could be achieved by selecting $c_j = r_j - 1, \ j \in \{1, \ \ldots, \ k\}$, were it not for the further competing objective of minimizing the expected size of the subset S. This second objective could be met by including just 1 population in S by choosing $c_j = 0, \ j \in \{1, \ \ldots, \ k\}$ but at the cost of minimizing the chances of correctly including τ. A constraint may also arise in some cases due to cost considerations: data may be more difficult to obtain from some of the populations than others.

This section presents a practical approach to optimizing the Rizvi - Sobel selection procedure depending on the number, one or two of the above objectives specified by the user, possibly subject to a sampling cost constraint. To make optimization practical, concessions are needed. Thus the quest for a fully distribution free method has had to be abandoned. Moreover we have had to rely on asymptotic approximation theory using the approximations presented in Section 3. Finally we need a strengthened version of Assumption 2.1 obtained by adding the requirement that the population F_τ be separable at or below ξ_τ when α is small, more precisely we now assume::

Assumption 4.1 *There exists* $\beta_0 \in (0,1)$ *such that for all* $u \in (0,\alpha]$ *and* $j \neq \tau$

$$F_\tau(F_j^{-1}(u)) \;<\; u \;<\; F_j(F_\tau^{-1}(u))(1-\beta_0/\alpha). \tag{4.11}$$

To conclude we simplify the notation by leaving off the R_1 in expressions like $P(j \in S \mid R_1)$ and turn to the specification of the objective function. That function has two components.

Probability of correct selection. For the probability of correct selection we need $P(F_\tau \in S) > LB(n,\beta)$ with $n = (n_1, \ldots, n_k)$ and $\beta = (\beta_1, \ldots, \beta_k)$. Such a bound has already been obtained in Section 2. However, one can use (3.9) to get an approximate lower bound that conforms with the bound we derive below for the expected subset size. Since by assumption, $H_{i\tau}^{-1}(u) = H_{\tau i}(u) = F_\tau(F_i^{-1})(u) < u$ for $u < \alpha$, that approximation gives:

$$\begin{aligned}
P(F_\tau \in S) &\approx 1 - \Phi\left(\frac{\sqrt{(n_\tau+1)}[-\alpha + \max_{i\neq\tau} H_{\tau i}(\alpha - \beta_i)]}{\sqrt{\alpha[1-\max_{i\neq\tau} H_{\tau i}(\alpha - \beta_i)]^2 + (1-\alpha)\max_{i\neq\tau} H_{\tau i}(\alpha - \beta_i)^2}}\right) \\
&> 1 - \Phi\left(\frac{\sqrt{(n_\tau+1)}[-\alpha + \max_{i\neq\tau}(\alpha - \beta_i)]}{\sqrt{\alpha[1 - \max_{i\neq\tau}(\alpha - \beta_i)]^2 + (1-\alpha)\max_{i\neq\tau}(\alpha - \beta_i)^2}}\right) \\
&= 1 - \Phi\left(\frac{-\sqrt{(n_\tau+1)}[\min_{i\neq\tau}\beta_i]}{\sqrt{\alpha[(1-\alpha) - \min_{i\neq\tau}\beta_i]^2 + (1-\alpha)(\alpha - \min_{i\neq\tau}\beta_i)^2}}\right),
\end{aligned}$$

where the inequality follows from the fact that this approximate expression for $P(F_\tau \in S)$ is decreasing in $\max_{i\neq\tau} H_{\tau i}(\alpha - \beta_i)$. In summary asymptotically as $(n_1, \ldots, n_k) \to \infty$

$$\left.\begin{array}{l} P(F_\tau \in S) > \\ 1 - \Phi\left(\frac{-\sqrt{(n_\tau+1)}[\min_{i\neq\tau}\beta_i]}{\sqrt{\alpha[(1-\alpha) - \min_{i\neq\tau}\beta_i]^2 + (1-\alpha)(\alpha - \min_{i\neq\tau}\beta_i)^2}}\right) \end{array}\right\}. \tag{4.12}$$

Note that through (4.12), Assumption 4.1 implies that $P(F_\tau \in S) \to 1$ as the n_i go to infinity. However that bound depends on the unknown τ leaving us with competing objectives, depending on the true τ. The situation is the one confronted in classical statistical decision theory where, since "the true state of nature" is unknown, the objective becomes that of maximally reducing, in some aggregate sense, the risk function over all possibilities for the unknown state. That led in particular cases to the minimax criterion and the Bayes risk criterion. We choose the latter for definiteness and reasons of personal preference. Thus we obtain the objective function:

$$\sum_{\tau=1}^{k} \omega_\tau \left\{1 - \Phi\left(\frac{-\sqrt{(n_\tau+1)}[\min_{i\neq\tau}\beta_i]}{\sqrt{\alpha[(1-\alpha) - \min_{i\neq\tau}\beta_i]^2 + (1-\alpha)(\alpha - \min_{i\neq\tau}\beta_i)^2}}\right)\right\},$$

where $\sum_{\tau=1}^{k} \omega_\tau = 1$ and $\omega \geq 0$.

Remark. In a Bayesian context ω_τ could represent the probability that τ is the correct selection, although we need not think of it that way. Moreover in the context of setting policy $\omega_\tau \equiv 1/k$ would seem a natural choice.

Expected subset size. Let $\mid S \mid$ be the subset size. Then, conditional on the unknown population distributions $\{F_1, \ldots, F_k\}$ including the correct one labelled $i = \tau$,

$$E[\mid S \mid] = \sum_{i=1}^{k} P(F_i \in S). \tag{4.13}$$

We add the additional assumption that $F_\tau(F_i^{-1}u) < u$ for all $u \in (0,1)$. Then, as all $n_i \to \infty$, $\lim P(\ F_\tau \in S\ ) < 1$. For the remaining $i \neq \tau$ we use the approximation (3.9) and the fact that $H_{il}^{-1} = H_{li}$, but now with $\beta_j < \beta_0,\ j\ = 1,\ \ldots,\ k$. This gives

$$\begin{aligned}
&P(F_l\ \in S) \approx \\
&1 - \Phi\left(\frac{\sqrt{(n_l+1)}[-\alpha + H_{l\tau}(\alpha\ -\ \beta_\tau)]}{\sqrt{\alpha[1 - H_{l\tau}(\alpha\ -\ \beta_\tau)]^2 + (1-\alpha)H_{l\tau}(\alpha\ -\ \beta_\tau)^2}}\right) < \\
&1 - \Phi\left(\frac{\sqrt{(n_l+1)}[-\alpha + \alpha(\alpha\ -\ \beta_0)^{-1}(\alpha\ -\ \beta_\tau)]}{\sqrt{\alpha[1 - \alpha(\alpha\ -\ \beta_0)^{-1}(\alpha\ -\ \beta_\tau)]^2 + (1-\alpha)\alpha(\alpha\ -\ \beta_0)^{-1}(\alpha\ -\ \beta_\tau)^2}}\right) = \\
&1 - \Phi\left(\frac{\sqrt{(n_l+1)}[\alpha(\alpha\ -\ \beta_0)^{-1}(\beta_0\ -\ \beta_\tau)]}{\sqrt{\alpha[1-\alpha\kappa_\tau]^2 + (1-\alpha)[\alpha\kappa_\tau]^2}}\right),
\end{aligned}$$

where $\kappa_\tau = (\alpha\ -\ \beta_\tau)/(\alpha\ -\ \beta_0)\ <\ 1$ whatever be τ under our assumption that $\beta_j < \beta_0$ for all j.
Thus asymptotically

$$E[\mid S \mid] < 1 + \sum_{l\ \neq\ \tau} \left\{1 - \Phi\left(\frac{\sqrt{(n_l+1)}[\alpha(\alpha\ -\ \beta_0)^{-1}(\beta_0\ -\ \beta_\tau)]}{\sqrt{\alpha[1-\alpha\kappa_\tau]^2 + (1-\alpha)[\alpha\kappa_\tau]^2}}\right)\right\}.$$

Once again we confront the unknown τ by taking a weighted average to get an upper bound for $E[\mid S \mid]$, that is

$$\begin{aligned}
&\sum_{\tau=1}^{k} \omega_\tau \left[1 + \sum_{l\ \neq\ \tau} \left\{1 - \Phi\left(\frac{\sqrt{(n_l+1)}[\alpha(\alpha\ -\ \beta_0)^{-1}(\beta_0\ -\ \beta_\tau)]}{\sqrt{\alpha[1-\alpha\kappa_\tau]^2 + (1-\alpha)[\alpha\kappa_\tau]^2}}\right)\right\}\right] \\
= \quad & k\ -\ \sum_{\tau=1}^{k} \omega_\tau \sum_{l\ \neq\ \tau} \Phi\left(\frac{\sqrt{(n_l+1)}[\alpha(\alpha\ -\ \beta_0)^{-1}(\beta_0\ -\ \beta_\tau)]}{\sqrt{\alpha[1-\alpha\kappa_\tau]^2 + (1-\alpha)[\alpha\kappa_\tau]^2}}\right) \\
\equiv \quad & UB(n,\beta)].
\end{aligned}$$

Objective function for optimization. Treating our problem as a multicriteria decision problem, we get the following objective function by using the bounds we have established:

$$\max_{n,\beta}\{LB(n,\beta) - \lambda UB(n,\beta)\},$$

subject to a cost contraint ($\text{Cost}_1\ n_1\ +\ \ldots\ +\ \text{Cost}_k\ n_k\ <\ C$), where C is the cost ceiling. A particular choice of interest is $C = n$, the total sample size, and $\text{Cost}_j \equiv 1$, so that the final term expresses the desirability of keeping the sample size under a prescribed sample size.

5 Numerical results

This section demonstrates the use of the extended Rizvi - Sobel method as well as our approach to optimizing it. In the process, we investigate the effects of changing the inputs in various ways. To begin, recall that the user must fix an upper bound β_0 for the range of the available β_j's, as required by the supplementary Assumption 4.1. It constrains the family of allowable population distributions by requiring for $u\ <\ \alpha$ that $F_j(F_\tau(u))\ >\ \alpha/(\alpha - \beta_0)u,\ u \in (0,1)$. That will be a strong constraint when β_0 is close to α, with the effect that the correct population, the one with the (unknown) label by $\tau \in \{1,\ \ldots,\ \}$, is easier to find. That means that if the β_j's are fixed, the upper bound for the expected size of the Rizvi - Sobel subset of populations will decrease towards 1 when β_0 increases towards α. Another way of looking at this situation is that Assumption 4.1 leads to the constraint $\beta_j\ <\ \beta_0$ for all j. Thus, restricting the space of distributions gives us more latitude in our choice of the β_j's. Thus, restricting the domain of applicability by that assumption, increases the number of available Rizvi - Sobel subsets. But recall that while small β values tend to yield small subsets, large values favour high correct selection probabilities. So we see the competition that plays out in the optimization procedure as when we specify β_0. We now turn to our numerical results.

The effect of increasing sample size for varying βs.

The demonstration looks at the effect of increasing the total sample size in the case of $k\ =\ 2$ populations in the unbalanced case where 60% of the sample is allocated to Population 2. Furthermore, for simplicity we look at eight cases where $\alpha \in (0.05, 0.5)$, $\beta_0 \in (0.4\alpha, 0.8\alpha)$ for each α and $\beta_1 = \beta_2 = \beta \in (0.5\beta_0, 0.9\beta_0)$ for each β_0. For each of these cases we calculate the lower bound for the probability of correctly selecting Population 2 as well as our upper bound for the expected subset size. In all cases we make the mixing weights equal, i.e. $w_1\ =\ w_2\ =\ 0.5$ The results are seen in the Table 2.
We discuss the lessons learned from this analysis in Section 6.

The effect of increasing the number of populations

This subsection looks at the effect of increasing the number of populations from $k\ =\ 2$ to $k\ =\ 7$. For the βs and αs, we make the same choices as in the previous subsection

Table 2: This table presents the lower bounds for the probability of correct selection and upper bounds for the expected sample size when the quantile of interest are for $\alpha = 0.05$ and $\alpha = 0.5$. Here we have two populations with selection indices $\beta_1 = \beta_2 = \beta$ and sample sizes $n_1 = 0.4n$ and $n_2 = 0.6n$ for various values of n.

$\alpha = 0.05$					$\alpha = 0.5$			
β_0	β	n	PCS	ESS	β_0	β	PCS	ESS
0.02	0.010	10	0.55	1.43	0.20	0.10	0.73	1.22
		100	0.63	1.29			0.96	1.01
		300	0.72	1.18			1.00	1.00
		600	0.79	1.10			1.00	1.00
0.02	0.018	10	0.58	1.49	0.20	0.18	0.91	1.44
		100	0.73	1.46			1.00	1.32
		300	0.85	1.43			1.00	1.21
		600	0.93	1.40			1.00	1.13
0.04	0.020	10	0.59	1.16	0.40	0.20	0.95	1.02
		100	0.75	1.00			1.00	1.00
		300	0.88	1.00			1.00	1.00
		600	0.95	1.00			1.00	1.00
0.04	0.036	10	0.67	1.41	0.40	0.36	1.00	1.18
		100	0.89	1.26			1.00	1.01
		300	0.98	1.13			1.00	1.00
		600	0.98	1.06			1.00	1.00

thereby making our results more directly comparable with those in the previous section. In other words $\alpha \in (0.05, 0.5)$, $\beta_0 \in (0.4\alpha, 0.8\alpha)$ for each α and $\beta_j = \beta \in (0.5\beta_0, 0.9\beta_0)$, $j \in \{1, \ldots, 7\}$ for each β_0. We distribute the sample over the seven populations in the proportions, $0.04, 0.08, 0.12, 0.14, 0.18, 0.20, 0.24$ and beginning with a sample size of 100 to ensure that there are enough items to cover all the populations at least somewhat realistically.
Once again we defer discussion of these results to the following section.

Optimizing the Rizvi – Sobel procedure.

We amend the general form of the objective function proposed in Section 4 by dividing the expected subset size by the total number of populations k to get a fraction that is then more directly comparable to the probability of correct selection, which is then on the same scale, $(0,\ 1)$. Then without loss of generality we can represent that function as

$$\max_{n,\beta}\{\eta\ [LB(n,\beta)] - (1\ -\ \eta)[\ UB(n,\beta)/k\ ]\}$$

Table 3: This table presents the lower bounds for the probability of correct selection and upper bounds for the expected sample size when the quantile of interest are for $\alpha = 0.05$ and $\alpha = 0.5$. Here we have two populations with selection indices $\beta_j \equiv \beta$ and total sample sizes of $100, 300, 600$ distributed across the seven populations in the successive proportions $0.04, 0.08, 0.12, 0.14, 0.18, 0.20, 0.24$.

$\alpha = 0.05$					$\alpha = 0.5$			
β_0	β	n	PCS	ESS	β_0	β	PCS	ESS
		100	0.57	3.32			0.82	1.74
0.02	0.010	300	0.62	2.88	0.20	0.10	0.93	1.21
		600	0.66	2.49			0.97	1.07
		10	0.63	3.86			0.98	3.02
0.02	0.018	300	0.71	3.77	0.20	0.18	1.00	3.01
		600	0.78	3.67			1.00	2.66
		100	0.64	1.41			0.98	1.02
0.04	0.020	300	0.73	1.08	0.40	0.20	1.00	1.00
		600	0.80	1.02			1.00	1.00
		100	0.75	3.19			1.00	1.54
0.04	0.036	300	0.86	2.68	0.40	0.36	1.00	1.13
		600	0.93	2.26			1.00	1.03

with $\eta \in [0,\ 1]$, subject to the cost constraint $\text{Cost}_1\ n_1\ +\ \ldots\ +\ \text{Cost}_k\ n_k\ <\ C$. Thus $\eta = 1$ would put all the weight on correct selection while $\eta = 0$ would mean putting it on on expected sample size. The ultimate choice would be context-dependent, but a reasonable default might be $\eta = 1/2$. Furthermore, we choose: $k = 7,\ \alpha = 0.05,\ \beta_0 = 0.04,$ and $w_j\ =\ 1/7,\ j\ =\ 1,\ \ldots,\ 7$. Finally we consider two cost scenarios. In the first $\text{Cost}_j = 1$ for each population while in the second, $\text{Cost}_j = 1,\ j\ =\ 1, 2; \text{Cost}_j = 4,\ j\ =\ 3,\ 4,\ 5; \text{Cost}_j = 20,\ j\ =\ 6, 7$. For these two scenarios we explore various total cost limits, beginning with $C = 280$ in the first case and $C = 1200$ in the second. Finally the cases $\eta \in \{0, 0.5, 1\}$ are considered.

Outputs from the optimization are the sample sizes and the subset selection indices, $\beta_j < \beta_0$ and we now turn to the results. These were obtained using constrOptim with the Nelder Mead option in version 2.12.0 of R. Convergence was extremely slow when the maximum cost was high. By the time the convergence criterion (outer.eps = 1e-06) was reached after sometimes hundreds of iterations, the cost constraint was not attained. However, by that time the probability of correct selection and the expected sample size had already attained their optimal values to several decimal places. The tables below report the results at the point where the criterion was reached.

Tables 4, 5, and 6, report the results for three case when respectively, the expected sample size gets all the weight, one - half the weight and none of the weight against the alternative of correct selection.

In each case the optimal values of the $\{\beta_j\}$ turns out to be the same in all cases, so a single entry appears in the table for them.

Table 4: Optimal selection indices and sample sizes for the Rizvi - Sobel method under various scenarios. This first table concerns the case where all the emphasis is put minimizing the expected subset size, in other words where $\eta = 0.0$.

Scenario	Max Cost	Optimal βs	Optimal sample sizes							PCS Bd	ESS Bd
	70	0.00	10			...			10	0.50	1.08
1	140	0.00	20			...			20	0.50	1.00
	350	0.00	44			...			44	0.50	1.00
	540	0.00	18	18	13	13	13	9	9	0.50	1.05
2	1080	0.00	36	36	26	26	26	17	17	0.50	1.00
	2700	0.00	44	44	42	42	42	32	32	0.50	1.00

Table 5: Optimal selection indices and sample sizes for the Rizvi - Sobel method under various scenarios. This table concerns the balanced case where equal weights are put on correct selection and expected subset size, in other words where $\eta = 0.5$.

Scenario	Max Cost	Optimal βs	Optimal sample sizes							PCS Bd	ESS Bd
	70	0.02	10			...			10	0.61	1.41
1	140	0.02	20			...			20	0.70	1.32
	350	0.03	50			...			50	0.83	1.12
	540	0.02	18	18	13	13	13	9	9	0.64	1.40
2	1080	0.03	39	39	26	26	26	17	17	0.74	1.31
	2700	0.03	145	145	70	70	70	39	39	0.88	1.17

Table 6: Optimal selection indices and sample sizes for the Rizvi - Sobel method under various scenarios. Here we see the results for the extreme case all the weight is put on correct selection, in other words where $\eta = 1.0$.

Scenario	Max Cost	Optimal βs	Optimal sample sizes							PCS Bd	ESS Bd
	70	0.04	10			...			10	0.74	4.00
1	140	0.04	20			...			20	0.82	4.00
	350	0.04	50			...			50	0.92	4.00
	540	0.04	17	17	13	13	13	9	9	0.77	4.00
2	1080	0.04	37	37	26	26	26	17	17	0.84	4.00
	2700	0.04	109	109	67	67	67	42	42	0.94	4.00

The results described above are for the relatively restricted case when the maximum value of the selection bias is close to the maximum corresponding to α - a somewhat limited number of population distributions would meet the assumption imposed by our assumptions when $\beta_0 = 0.04$. To complete our analysis, Table 7 presents results for the case where $\beta_0 = 0.02$.

For brevity we examine only the case when the probability of correct selection gets all the weight. However the two scenarios are the same as those above. The lessons learned from all these numerical results will be discussed in the Section 6.

Table 7: Optimal selection indices and sample sizes for the Rizvi - Sobel method under various scenarios. This table again puts full weight on correct selection where $\eta = 1.0$ but now we now $\beta_0 = 0.02$ Note that the upper bound for the probability of correct selection is smaller than in the previous table. At the same time, the bound for expected subset size suggests the latter has been reduced.

Scenario	Max Cost	Optimal βs	Optimal sample sizes							PCS Bd	ESS Bd
	70	0.02	10			...			10	0.61	1.50
1	140	0.02	20			...			20	0.67	1.16
	350	0.02	50			...			50	0.75	1.00
	540	0.02	22	22	14	14	14	8	8	0.64	1.35
2	1080	0.02	47	47	28	28	28	16	16	0.70	1.10
	2700	0.02	136	136	74	74	74	38	38	0.80	1.00

6 Discussion

This paper has investigated an extension of the Rizvi – Sobel procedure stated in (4.10). Users need to specify what might be called the "selection indices", i.e. the $\{c_j\}$ and possibly even the sample sizes $\{n_j\}$. Setting those indices close to zero will make getting into the subset more difficult and hence will make it smaller, all at the cost of reducing the probability of correct selection. While the user must make specific choices, how to do so does not seem to have been investigated previously. We will make make our recommendations below.

But first recall the ideals embraced by the method. First to maximize its domain of applicability, it should be non–parametric, i.e. the population distributions cannot have parametric form, for example be Gaussian with parameters μ and σ. The ideal is achieved in its formulation (2.4), which relies on the minimal sufficient statistics for the class of non–parametric distributions, i.e. the order statistics.

Second, the properties of the method should ideally be distribution free to enable the user to select such things as the sample sizes and selection indices, without needing to know the true but unknown population distribution. However that ideal cannot be realized exactly. Rizvi and Sobel (1967) recognized this fact and found the bypass route we follow in this paper. First they add a fairly weak Assumption 2.1, which restricts the class of possible distributions somewhat. Then they find the lower bound in (2.8) for the probability of correct selection that is the same for all members of the class. The user can then raise that bound by manipulating the $\{c_j\}$ and $\{n_j\}$ until a satisfactory level is attained, one that ensures a sufficiently large probability of correct selection.

That route does not however lead us to an analogous upper bound for the expected subset size, that would force this size down by suitably manipulating the $\{c_j\}$ and $\{n_j\}$. To achieve that bound requires the further restriction on the class of population distributions imposed by Assumption 4.1. As well, we need a more complex analysis beginning with the asymptotic theory presented in Section 3. To understand why, we need to revisit (4.13) for the expected sample size and (3.5) on which it relies. The latter involves $H_{ij}(u) = F_i(F_j^{-1}(u))$, which when used in the former becomes $F_j(F_\tau^{-1}(u))$ where $j \neq \tau$, the correct population choice. To get an upper bound for the expected sample size, we would therefore like to replace $F_j(F_\tau^{-1}(u))$ by something larger that is independent of the these true population distributions. But now the seemingly reasonable assumption of separability around α fails - it gives us a lower, not upper bound. So how can we turn this around?

Asymptotics provides both insight as well as the answer. For the insight, we examine (2.6) with the generic parameters there replaced by the ones required for the Rizvi –Sobel method $n_j\ (\alpha - \beta_j)$ and $n_j - [n_j\ (\alpha - \beta_j) + 1]$. Then, if we apply Stirling's approximation for the gamma functions there, we find that the density is approximately proportional to:

$$\left[\left(\frac{t}{K_j}\right)^{K_j}\left(\frac{1-t}{1-K_j}\right)^{1-K_j}\right]^{n_j},$$

with $K_j = \alpha - \beta_j$, which attains its maximum value of 1 at $t = K_j$. Thus for large n_j the probability for this distribution is concentrated largely at the point $t = K_j$ Revisiting (3.5), we see that the j^{th} factor in its integrand $P[V_j \leq F_\tau(F_j^{-1}(u))]$ will be nearly 0 for $j \neq \tau$ unless $F_\tau(F_j^{-1}(u)) > K_\tau$ that is $u > F_j(F_\tau^{-1}(K_\tau))$. In fact the integrand will be approximately 0 unless $u > \max_{j\neq\tau} F_j(F_\tau^{-1}(K_\tau))$. Thus the probability of including j in the subset S, is approximately $P(W_j > \max_{j\neq\tau} F_j(F_\tau^{-1}(K_\tau)))$, where W_j is defined just below Equation (3.5). Finally Assumption 4.1 can now be applied to give us the upper bound we seek by giving us a lower bound for $\max_{j\neq\tau} F_j(F_\tau^{-1}(K_\tau))$.

Although the heuristic asymptotic reasoning might be formalized, say by using a saddlepoint approximation, we proceed instead to use a normal approximation that was developed in Section 3 and applied in Section 4, since it gave us as a convenient byproduct a convenient approximation to the probabilities involved in our theory.

The two key assumptions we needed to obtain the bounds above do limit the class of allowable population distributions and thus make them less than completely distribution free. However, they play the important qualitative role of showing intuitive constraints on the populations needed to make a subset selection method work properly. While it may be possible to weaken them, we believe some sort of separability assumptions like them will be needed for things like the optimal sample sizes. In practice, their validity can be checked given sufficient data from each of the populations, through diagnostic plots and so on. However, in future work we will be examining the Rizvi–Sobel method's robustness against their failure.

We conclude this section by summarizing the lessons learned from the numerical studies in Section 5.

- In agreement with intuition, small β_j's or equivalently c_js produce small expected

subsets of populations with a correspondingly diminished probability of correct selection. The opposite is true when they are large.

- Large values of the upper bound for the selection indices, given by β_0 in this paper's notation, restrict the population of allowable distributions and not surprisingly yield both smaller expected subset sizes and greater probabilities of correct selection. Our optimization results show that the βs tend to approach 0 or β_0 according as more or less weight is attached to minimizing the expected subset size.

- Finding the population with the largest median seems much easier than finding the one with the largest 5^{th} percentile. That is seen quite clearly in Table 2 by comparing the results corresponding to these two cases. For every fixed sample size, no matter how small or how large the βs, the method does better for both the correct selection probability as well as the expected subset size for medians. For medians, surprisingly small samples yield good results for correct selection and expected subset size, which are close to their ultimate large sample limits.

- Table 3 shows that when choosing the subset from a large number of populations as against a small number, setting a large value of β_0 to restrict the class of possible population distributions pays off. In fact it reduces dramatically the expected subset size when the βs are chosen to be small.

- Not surprisingly when the sampling costs are the same for all populations, both the optimal sample sizes and optimal βs are identical for all populations. Convergence is slow in that case, requiring a large number of iterations and suggesting the choices will be quite robust. Indeed when the optimization's number of iterations constraint was set to 1000, and the cost constraint was not always attained. bf The sentence that ends here is not clear to me - do not know what it means The expected subset size as well as the probability of correct selection reached their terminal value to several decimal places early in the iterative sequence. Even with extremely variable sampling costs, the optimal βs are equal across all population distributions.

- The optimal results tend to follow the weights attached to the two criteria. So for example if correct selection gets relatively high weight, the optimal βs tend to be equal to their upper bound β_0. Future work will be needed to find ways of eliciting the relative multi-attribute criteria weights for users in particular contexts. Equal weighting tends to produce optimal β around one–half of β_0, the natural compromise estimator.

- The optimal sample sizes change surprisingly little even when the relative costs are changed dramatically.

7 Conclusions

Overall, our results lead us to recommend the extended Rizvi – Sobel procedure for use in practice. Like the original for equal sample sizes (Rizvi and Sobel 1967), it is extremely simple to describe and use, once the requisite sample sizes and our parameters have been specified. Furthermore, with our normal approximations, these quantities are easily calculated. In fact we needed just a few lines of code to implement our optimization criterion. Unlike the original, it allows the sample sizes to be unequal, making it more practical than the original. It is quite flexible, allowing the user latitude in weighting two primary objectives, a high probability of correct selection (PCS) and expected subset size (ESS). Finally the numerical results we provide, confirm what seem to be desirable heuristic properties. In other words, the method produces results in agreement with our qualitative heuristics while giving a more or less complete quantitative framework for specifying the procedure. However as with any recommendation some caveats are in order. First producing the bounds needed for our implementation of the method led us to resort to asymptotic approximations and inevitably the quality of those approximations is an issue in particular contexts. However the answer to that will depend on the context and on what might be the true population distributions.

The veracity of our assumptions will also be an issue for anyone thinking of implementing the method. As noted earlier, these can be diagnostically assessed in any particular application. At the same time in future work, we intend to explore the robustness of our procedure when the assumptions are violated.

Finally, the procedure provides little scope for expert input in contexts where lots of background knowledge is available. Indeed the nonparametric - distribution free ideals built into its construction rule out such input by design. So we developed a Bayesian nonparametric alternative method that does allow such input and that is the subject of a second manuscript now in preparation.

Acknowledgements. Mr Conroy Lum, FPInnovations introduced the second author to the application that led to the work reported in this paper and we are indebted to him. The work was partially completed as part of a research program funded by a Natural Science and Engineering Research Program Collaborative Research and Development Grant with contributions from FPInnovations.

8 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.

Embrechts, P., Klüppenberg, C. and Mikosch, T. (1997). Modeling extremal events for New York: Springer.

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 University 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.

van Eeden, C. (2009). Rizvi-Sobel subset selection with unequal sample sizes, Technical Report # 253, Department of Statistics, The University of British Columbia, Vancuver, B.C., Canada.

van Eeden, C. (2010). Correction to: Rizvi-Sobel subset selection with unequal sample sizes, Technical Report # 261, Department of Statistics, The University of British Columbia, Vancouver, B.C. Canada.

A Appendix

In this Appendix it is shown that (3.3) is monotone in each of the c_i by showing that

$$\frac{-(r_i - c_i)(1-u) + (n_-(r_i - c_i) + 1)u}{\sqrt{(r_i - c_i)(1-u)^2 + (n_i - (r_i - c_i) + 1)u^2}} \tag{A.1}$$

is, for each $\in \{1, \ldots, k\}$, monotone in c_i.

Proof. For notational convenience, the index i is left off, so it needs to be shown that

$$\frac{-(r-c)(1-u)+(n-(r-c)+1)u}{\sqrt{(r-c)(1-u)^2+(n-(r-c)+1)u^2}} \tag{A.2}$$

is monotone in c for $c \in [0, r-1]$.

The derivative of (A.2) with respect to c multiplied by

$$2\sqrt{(r-c)(1-u)^2+(n-(r-c)+1)u^2}$$

equals

$$\begin{aligned}
&2(r-c)(1-u)^2+2(n-(r-c)+1)u^2+\\
&(1-2u)[-(r-c)(1-u)+(n-(r-c)+1)u]=\\
&(n+1-2(r-c))u+r-c.
\end{aligned}$$

So, the derivative of (A.2) is positive because $r-c>0$ and $n+1-(r-c)>0$. ♡