SOME FIXED-SAMPLE RANKING AND SELECTION PROBLEMS[1]

By D. M. Mahamunulu

University of Minnesota and State University of New York at Buffalo.

1. Introduction and Summary. In recent years considerable research has been devoted to a class of problems which is concerned with ranking and/or selecting a subset of k given populations where the ranking or the selection is defined in terms of a (scalar) parameter of the populations. In these problems the interest is often centered on the populations having large (small) values of the ranking parameter. One usually refers to the $t(< k)$ populations with largest (smallest) values of the ranking parameter as the t best populations. In this paper we consider a problem of selecting a subset of specified size, from a given set of k populations, which contains a subset of the t best populations.

Suppose that $\Pi_1, \Pi_2, \cdots, \Pi_k$ is a given set of k populations, where the distribution function of each observation from Π_i is $F(\cdot \mid \theta_i)$. The parameter θ_i is unknown, but it belongs to the interval, Θ, of the real line $(1 \leqq i \leqq k)$. We assume that the functional form of F is known. Let $\theta_{[1]} \leqq \theta_{[2]} \leqq \cdots \leqq \theta_{[k]}$ be the ranked θ_i; we assume that it is not known with which population $\theta_{[i]}$ is associated $(1 \leqq i \leqq k)$. The t populations with largest θ-values are defined as the t best populations and we refer to θ as the ranking parameter.

The problem of selecting the t best populations in an unordered manner has been studied extensively from the sampling point of view, in relation to several distributions. The usual formulation of the problem is the following: the experimenter's goal is to select the t best populations in an unordered manner. He specifies two positive constants d^* and P^*, where $\binom{k}{t}^{-1} < P^* < 1$. He desires to have a fixed-sample procedure which has a probability of at least P^* of selecting the t best populations whenever $\theta_{[k-t+1]}$ is at a distance not less than d^* from $\theta_{[k-t]}$.

Bechofer [3] developed a procedure based on predetermined number of observations from each population when F is the normal distribution function with unknown mean θ and known variance. Bechhofer and Sobel [4] considered a similar problem in relation to normal populations where the ranking parameter is the variance. Sobel and Huyett [9], Sobel [8], Rizvi [7], Baar and Rizvi [2] considered similar problems for other distributions.

Here we consider a generalized version of the above selection problem. We solve the problem in broad generality by not considering a specific family of distributions $F(\cdot \mid \theta)$, but by assuming certain properties of F. Thus many of the

---

Received 7 June 1966; revised 6 December 1966.

[1] This paper forms a part of author's doctoral dissertation submitted to the University of Minnesota. This research was supported by National Science Foundation under Grant No. NSF-GP-3813.

available results (concerning the above selection problem) can be obtained as special cases of our results.

Let c, s, t be integers such that $\max(1, s + t + 1 - k) \leqq c \leqq \min(s, t)$, which implies that $\max(s, t) \leqq k - 1$. The experimenter's goal, which is referred as Goal I, is to select a subset of size s which contains at least c of the t best populations. The experimenter specifies two positive constants d^* and $P^*(<1)$. He desires to have a fixed-sample procedure for which the probability of selecting the desired type of subset of populations is not less than P^* whenever the distance between $\theta_{[k-t]}$ and $\theta_{[k-t+1]}$ is at least d^*.

Two particular cases of this goal are of special interest: (1) Selection of a subset of size $s(\geqq t)$ which contains the t best populations, (2) Selection of a subset of size $s(\leqq t)$ which includes any s of the t best populations. Sobel pointed (see the footnote on page 22 of [3]) out that sometimes Case 2 is of interest. These two cases correspond to $c = t$, $s \geqq t$ and $c = s$, $s \leqq t$. Further when $c = s = t$, Goal I reduces to the goal of selecting the t best in an unordered manner.

The proposed selection procedure, R_s, is based on suitable statistics $T_1, T_2, \cdots, T_k$, where T_i is computed from a random sample of size n from $\Pi_i (1 \leqq i \leqq k)$. The procedure R_s selects the subset of populations which corresponds to the s largest T-values. Having specified the procedure as above, the problem is to determine the common sample size n so that the probability requirement imposed on the procedure is satisfied. This problem has been solved under the assumption that T_i is an absolutely continuous random variable and its distribution function is stochastically increasing in θ_i for each value of n. This has been done in Section 5.

It should be noted (by considering the subset not selected) that the above selection problem is logically equivalent to similar selection problem where the experimenter's goal is to select a subset of size $(k - s)$ which contains at least $(k - t) - (s - c)$ of the $(k - t)$ populations with smallest θ-values. Thus solutions to the above problem for all admissible values of c, s and t (with fixed k) will provide solutions to the selection problem where the goal is the selection of a subset of size s, which contains at least c of those t populations with smallest values for the ranking parameter.

Section 6 gives results for the two above mentioned particular cases of Goal I. A theorem, which relates the sample sizes necessary to achieve Goal I, and its particular cases, is given in that section. Section 7 deals with an easily verifiable sufficient condition for the existence of the required sample size. In Section 8, the general results have been applied to normal distributions with unknown mean and known common variance.

2. Formulation of the problem. Let $\boldsymbol{\theta}$ denote the vector $(\theta_{[1]}, \theta_{[2]}, \cdots, \theta_{[k]})$ and let Ω stand for the parameter space, which is the set of all admissible vectors $\boldsymbol{\theta}$. Further let $d(x, y)$ be a continuous non-negative real-valued function defined for $x \geqq y$, where x and y are both real, such that $d(x, y) = 0$ if and only if $x = y$. For fixed y, it is a strictly increasing function of x and for fixed x, it is a strictly

decreasing function of y. We also assume that $d(x, y)$ can take on indefinitely large values. We shall call such a function a distance measure. Let d^* be a specified position number. The parameter space Ω is partitioned into a "preference zone" $\Omega(d^*)$ defined by

$$\Omega(d^*) = \{\boldsymbol{\theta}: d(\theta_{[k-t+1]}, \theta_{[k-t]}) \geqq d^*\} \tag{2.1}$$

and its complement $\bar{\Omega}(d^*)$, the "indifference zone." The choice of the distance measure depends on the class of distribution functions $\mathfrak{F} = \{F(\cdot \mid \theta), \theta \varepsilon \Theta\}$ under consideration in a specific example.

Let B be a subset of size t of the set $\mathfrak{s} = (1, 2, \cdots k)$ with the property that $\alpha \varepsilon B$ implies $\theta_\alpha \geqq \theta_\beta$ for all $\beta \varepsilon \mathfrak{s} - B$. Let B_s be a subset of size s, of $\mathfrak{s}$ such that $B \cap B_s$ contains at least c elements. A correct selection (CS) is defined to be the selection of a subset, $\Pi(s)$, of populations where $\Pi(s) = \{\Pi_i: i \varepsilon B_s\}$. The experimenter desires to have a fixed-sample selection procedure for which the probability of a CS satisfies the condition

$$P(CS \mid \boldsymbol{\theta}) \geqq P^* \quad \text{for all} \quad \boldsymbol{\theta} \varepsilon \Omega(d^*). \tag{2.2}$$

(Here P^* is a specified positive number less than 1.)

Without loss of generality we can assume that P^* is not less than $P(c, k, s, t)$ where

$$P(c, k, s, t) = \binom{k}{s}^{-1} \sum_{i=c}^{\min(s,t)} \binom{t}{i}\binom{k-t}{s-i}. \tag{2.3}$$

If P^* were less than $P(c, k, s, t)$, we can meet the requirement (2.2) by choosing the subset at random. So, to make the problem non-trivial, we will assume that P^* is not less than $P(c, k, s, t)$. We also need the restriction that $P(c, k, s, t)$ is less than one, which is satisfied by our choice of the integers c, s, t in relation to k.

3. Proposed procedure R_s. Let $\{X_{ij}\}$ $(1 \leqq j \leqq n)$ be independent observations from Π_i $(1 \leqq i \leqq k)$ and let $T_i = T(X_{i1}, X_{i2}, \cdots, X_{in})$ $(1 \leqq i \leqq k)$ where $T_1, T_2, \cdots, T_k$ is an independent set of statistics and T_i's have density functions. Let $G_n(\cdot \mid \theta_i)$ be the distribution function of T_i $(1 \leqq i \leqq k)$. The choice of the function T will depend upon the distribution function $F(\cdot \mid \theta)$; in general $T_1, T_2, \cdots, T_k$ are statistics relevant to the estimation of $\theta_1, \theta_2, \cdots, \theta_k$ respectively. The proposed procedure is based on the statistics T_i.

PROCEDURE R_s: Let $T_{[1]} \leqq T_{[2]} \leqq \cdots \leqq T_{[k]}$ be the ordered T_i. The set of populations corresponding to $T_{[k-s+1]}, \cdots, T_{[k]}$ is the set to be selected.

REMARK. In practice we do encounter situations in which two or more T_i may be equal, even when T is a continuous random variable. In such cases the equal T-values should be ranked by using a randomized procedure which assigns equal probability to each possible ordering of those values.

Once the common sample size n is specified, the procedure R_s is completely defined; our problem will be that of determining this sample size so that the probability requirement (2.2) is satisfied. As to the existence of the required n-value one can argue heuristically as follows: if T is a consistent estimator of θ, then the

largest T-values will come from the populations with largest θ-values with a probability that tends to one as n tends to infinity. Thus the probability of a CS, under the procedure R_s, will tend to one as n tends to infinity. Hence by choosing n sufficiently large we can meet the probability requirement (2.2).

Remarks about choice of T_i. Whenever a sufficient statistic (which has fixed dimensionality for all n) for θ_i exists, then the proper choice of T_i is some appropriate function of the sufficient statistic. The choice of T_i becomes a problem only when such a sufficient statistic does not exist. The results to be obtained will be applicable only when the chosen statistics T_i are such that the family $\mathcal{G} = \{G_n(\cdot \mid \theta) : \theta \in \Theta\}$ is a stochastically increasing (SI) family of distribution functions (for the definition of an SI family of distribution functions one may refer to [5] p. 73). It may be pointed out that in some cases of interest F contains two or more unknown parameters. The results to be proved will also be applicable to such cases provided the distribution of T_i depends only on θ_i (the ranking parameter), but not on the nuisance parameters in addition to the above mentioned property. For simplicity (with a slight loss of generality) we have assumed that F involves only one unknown parameter θ.

In the next section we determine the infimum of the probability of a CS (PCS) over the preference zone $\Omega(d^*)$. The sample size is then determined as the smallest integer for which this infimum is not less than P^*.

4. Probability of a correct selection and its infimum. As a first step towards obtaining the infimum of PCS we prove a theorem dealing with its monotone properties. To prove this theorem we need the following results on an SI family of distribution functions.

LEMMA 4.1. *Let $F(x \mid \theta) = F_\theta(x)$ where $\theta \in \Theta$, be an* SI *family of distribution functions on the real line. If ψ is any non-decreasing (non-increasing) function of x, then $E_\theta\psi(X)$ is a non-decreasing (non-increasing) function of θ.*

This result is a simple consequence of a problem given by Lehmann ([5], p. 112, #11). So the proof is ommitted.

LEMMA 4.2[2]. Let $F(x \mid \theta) = F_\theta(x)$ *where $\theta \in \Theta$ be an* SI *family of distribution functions on the real line. Let $X_1, X_2, \cdots, X_k$ be independent random variables, where the distribution function of X_i is $F(x_i \mid \theta_i)$. For any fixed i $(1 \leqq i \leqq k)$, if $\psi = \psi(x_1, x_2, \cdots, x_k)$ is a non-decreasing (non-increasing) function of x_i when all x_j for $j \neq i$ are held fixed, then $E\psi(X_1, X_2, \cdots, X_k)$ is a non-decreasing (non-increasing) function of θ_i.*

PROOF.

$$(4.1) \quad E\psi(X_1, X_2, \cdots, X_k) = \int \psi \prod_{i=1}^{k} dF(x_i \mid \theta_i)$$
$$= \int [\int \psi \, dF(x_i \mid \theta_i)] \prod_{j=1, j\neq i}^{k} dF(x_j \mid \theta_j).$$

Since ψ is a non-decreasing (non-increasing) function of x_i when all x_j for $j \neq i$

[2] After obtaining this lemma, I learned that Alam and Rizvi [1] have independently derived a similar lemma.

are held fixed, from the Lemma 4.1 it follows that

$$(4.2)\quad E\{\psi(X_1, X_2, \cdots, X_k) \mid x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_k\} = \int \psi \, dF(x_i \mid \theta_i)$$

is a non-decreasing (non-increasing) function of θ_i. Since this holds for each value $(x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_k)$, the right hand member of (4.1) and hence $E\psi$ is a non-decreasing (non-increasing) function of θ_i. Since this holds for each fixed i, the lemma follows.

Let Y_i be the statistic based on the sample from the population with the parameter $\theta_{[i]}$ $(1 \leqq i \leqq k)$. That is, the set $(Y_1, Y_2, \cdots, Y_k)$ is same as the set $(T_{j_1}, T_{j_2}, \cdots, T_{j_k})$ where $(j_1, j_2, \cdots, j_k)$ is some unknown permutation of $(1, 2, \cdots, k)$. Our procedure R_s is defined in terms of the statistics T_j and hence it is based on the statistics Y_i. We make the following assumptions on the statistics Y_i.

ASSUMPTION 4.1: The statistics Y_i are absolutely continuous random variables.

ASSUMPTION 4.2: The family $\mathcal{G} = \{G_n(\cdot \mid \theta) : \theta \in \Theta\}$ of distribution functions is an SI family for each positive integer n.

It is easy to see that the following two events are equivalent:

$$\{CS\} = \{c\text{th largest of } (Y_{k-t+1}, Y_{k-t+2}, \cdots, Y_k) > (s - c + 1)\text{st largest of } (Y_1, Y_2, \cdots, Y_{k-t})\}$$

so that the PCS at the parameter point $\boldsymbol{\theta}$ is

$$(4.3)\quad P(CS \mid \boldsymbol{\theta}) = P[c\text{th largest of } (Y_{k-t+1}, Y_{k-t+2}, \cdots, Y_k) > (s - c + 1)\text{st largest of } (Y_1, Y_2, \cdots, Y_{k-t})]$$

where $Y_1, \cdots, Y_k$ is a set of independent random variables such that the distribution function of Y_i is $G_n(\cdot \mid \theta_{[i]})$, $(1 \leqq i \leqq k)$.

THEOERM. *Under the Assumptions* 4.1 *and* 4.2 *the* $P(CS \mid \boldsymbol{\theta})$ *is a non-increasing function of* $\theta_{[\alpha]}$ $(\alpha = 1, 2, \cdots, k - t)$ *and a non-decreasing function of* $\theta_{[\beta]}$ $(\beta = k - t + 1, k - t + 2, \cdots, k)$.

PROOF. By (4.3) the set of points in R^k where a CS occurs is the set $\{(y_1, y_2, \cdots, y_k) : u < v\}$ where u and v are, respectively, the $(s - c + 1)$st largest of $(y_1, y_2, \cdots y_{k-t})$ and the cth largest of $(y_{k-t+1}, y_{k-t+2}, \cdots, y_k)$. If ψ is the indicator function of this set, then

$$(4.4)\qquad P(CS \mid \boldsymbol{\theta}) = E\psi(Y_1, Y_2, \cdots, Y_k).$$

It is easy to see that u is a non-decreasing function of y_α $(\alpha = 1, 2, \cdots, k - t)$ when all y_i for $i \neq \alpha$ are held fixed and also that v is a non-decreasing function of y_β $(\beta = k - t + 1, k - t + 2, \cdots, k)$ when all y_m for $m \neq \beta$ are held fixed. Hence ψ is a non-increasing function of y_α $(\alpha = 1, 2, \cdots, k - t)$ when all other y's are held fixed and it is a non-decreasing function of y_β $(\beta = k - t + 1, k - t + 2, \cdots, k)$ when all other y's are held fixed. By applying the Lemma 4.2 to the function ψ we obtain the desired result.

This theorem represents a valuable tool in obtaining the infimum of $P(CS \mid \boldsymbol{\theta})$. It forms one of the key results of this investigation.

Infimum of $P(CS \mid \boldsymbol{\theta})$ *over the preference zone* $\Omega(d^*)$ (*see* 2.1). From the theorem it follows that any distance measure d, we have

$$\inf_{\boldsymbol{\theta} \varepsilon \Omega(d^*)} P(CS \mid \boldsymbol{\theta}) = \inf_{\boldsymbol{\theta} \varepsilon \omega(\theta, \theta_0)} P(CS \mid \boldsymbol{\theta}) \tag{4.5}$$

where $\omega(\theta, \theta_0)$ is that set of points $\boldsymbol{\theta} \varepsilon \Omega(d^*)$, for which

$$\theta_{[1]} = \theta_{[2]} = \cdots = \theta_{[k-t]} = \theta_0 \text{ (say)}, \tag{4.6}$$

$$\theta_{[k-t+1]} = \theta_{[k-t+2]} = \cdots = \theta_{[k]} = \theta \text{ (say)}.$$

Here θ and θ_0 are arbitrary values belonging to Θ such that $d(\theta, \theta_0) \geqq d^*$. A configuration of the parameters $\theta_1, \theta_2, \cdots, \theta_k$ for which (4.6) holds is, sometimes, called a generalized least favorable (GLF) configuration. The $P(CS \mid \theta)$ for the GLF configuration (4.6) is given by

$$P(\theta, \theta_0) = \int_{-\infty}^{\infty} U(x \mid \theta_0)\, dV(x \mid \theta) = \int_{-\infty}^{\infty} [1 - V(x \mid \theta)]\, dU(x \mid \theta_0), \tag{4.7}$$

where

$$U(x \mid \theta_0) = \sum_{\alpha=0}^{s-c} \binom{k-t}{\alpha} G_n^{k-t-\alpha}(x \mid \theta_0)[1 - G_n(x \mid \theta_0)]^{\alpha} = I[G_n(x \mid \theta_0); c', s - c + 1] \tag{4.8}$$

and

$$V(x \mid \theta) = \sum_{\alpha=0}^{t-c} \binom{t}{\alpha} G_n^{\alpha}(x \mid \theta)[1 - G_n(x \mid \theta)]^{t-\alpha} = I[G_n(x \mid \theta); t - c + 1, c]; \tag{4.9}$$

here

$$c' = k - t - s + c \quad \text{and} \quad I(x; p, q) = I_x(p, q) = [\beta(p, q)]^{-1} \int_0^x t^{p-1}(1 - t)^{q-1}\, dt. \tag{4.10}$$

Since $G_n(x \mid \theta_0)$ is a non-increasing function of θ_0 for each x, from (4.8) it follows that $U(x \mid \theta_0)$ is a non-increasing function of θ_0 for each x. Thus $P(\theta, \theta_0)$ is a non-increasing function of θ_0 for fixed θ. From (4.5) and (4.7), for any distance measure d we have

$$\inf_{\boldsymbol{\theta} \varepsilon \Omega(d^*)} P(CS \mid \boldsymbol{\theta}) = \inf_{\{(\theta, \theta_0): \theta, \theta_0 \epsilon \Theta, d(\theta, \theta_0) \geqq d^*\}} P(\theta, \theta_0). \tag{4.11}$$

From the monotone properties of the distance measure d (see Section 2) and of the function $P(\theta, \theta_0)$, it follows that for fixed θ

$$\inf_{\theta_0, d(\theta, \theta_0) \geqq d^*} P(\theta, \theta_0) = P(\theta, \theta') = Q(\theta, n) \text{ (say)}, \tag{4.12}$$

where θ' is that function of θ determined by $d(\theta, \theta') = d^*$. Hence

$$\inf_{\boldsymbol{\theta} \varepsilon \Omega(d^*)} P(CS \mid \boldsymbol{\theta}) = \inf_{\theta \varepsilon \Theta} Q(\theta, n). \tag{4.13}$$

Using (4.8) and (4.9) in the first expression for $P(\theta, \theta')$ (see 4.7), we obtain

$$
\begin{aligned}
Q(\theta, n) &= (t!/(t-c)!(c-1)!)\sum_{\alpha=0}^{s-c}\binom{k-t}{\alpha}\int_{-\infty}^{\infty} G_n^{\,k-t-\alpha}(x \mid \theta') \\
&\quad \cdot [1 - G_n(x \mid \theta')]^{\alpha} G_n^{\,t-c}(x \mid \theta)[1 - G_n(x \mid \theta)]^{c-1}\, dG_n(x \mid \theta) \qquad (4.14) \\
&= \int_{-\infty}^{\infty} I[G_n(x \mid \theta'); c', s-c+1]\, dI[G_n(x \mid \theta); t-c+1, c].
\end{aligned}
$$

Using the second expression of (4.7) for $P(\theta, \theta')$ one can obtain two other equivalent expressions for $Q(\theta, n)$.

The determination of the infimum of $Q(\theta, n)$ over the admissible values of θ calls for the exact knowledge of the distribution function $G_n(\cdot \mid \theta)$. We need separate analysis to obtain this infimum for each particular distribution function $G_n(\cdot \mid \theta)$. But when θ happens to be either a location parameter or a scale parameter for the family $\mathcal{G}$, this infimum can be obtained "automatically", i.e., without any further analysis by adopting a suitable distance measure.

Infimum of $P(CS \mid \boldsymbol{\theta})$ over the entire parameter space Ω:

From the theorem

$$
\inf_{\boldsymbol{\theta}\in\Omega} P(CS \mid \boldsymbol{\theta}) = \inf_{\{(\theta,\theta_0):\theta,\theta_0\in\Theta,\theta\geqq\theta_0\}} P(\theta, \theta_0). \qquad (4.15)
$$

Since $P(\theta, \theta_0)$ is a non-increasing function of θ_0 for fixed θ we have

$$
\begin{aligned}
\inf_{\boldsymbol{\theta}\in\Omega} P(CS \mid \boldsymbol{\theta}) &= \inf_{\theta\in\Theta} P(\theta, \theta) = \int_0^1 I_y(c', s-c+1)\, dI_y(t-c+1, c) \qquad (4.16) \\
&= J(c, k, s, t) \text{ (say)}.
\end{aligned}
$$

LEMMA. *$J(c, k, s, t) = P(c, k, s, t)$, where $P(c, k, s, t)$ is defined by (2.3).*

PROOF. Expressing $I(y; c', s-c+1)$ as a finite series, we obtain

$$
\begin{aligned}
J(c, k, s, t) &= \sum_{j=0}^{s-c}(t!/(t-c)!(c-1)!)\binom{k-t}{s-c-j}\int_0^1 y^{k-s+j}(1-y)^{s-j-1}\, dy \qquad (4.17) \\
&= \binom{k}{t}^{-1}\sum_{j=c}^{s}\binom{j-1}{c-1}\binom{k-j}{t-c}.
\end{aligned}
$$

Let X denote the number of red balls in a random sample of size s chosen, without replacement, from an urn containing k balls of which t are red. Also let Y denote the number of balls needed to be drawn without replacement from the above urn so as to include exactly c red balls in the sample. Then, from (4.17), we have

$$
\begin{aligned}
J(c, k, s, t) &= P(Y \leqq s) \qquad (4.18) \\
&= P(X \geqq c) = P(c, k, s, t).
\end{aligned}
$$

This completes the proof of the lemma.

In view of the lemma, from (4.16) we obtain

$$
\inf_{\boldsymbol{\theta}\in\Omega} P(CS \mid \boldsymbol{\theta}) = P(c, k, s, t). \qquad (4.19)
$$

5. Determination of the required sample size. The required sample size is the smallest value of n for which

$$
\inf_{\boldsymbol{\theta}\in\Omega(d^*)} P(CS \mid \boldsymbol{\theta}) = \inf_{\theta\in\Theta} Q(\theta, n) \geqq P^*, \qquad (5.1)
$$

where $Q(\theta, n)$ is given by (4.14). Let us denote the infimum of $Q(\theta, n)$ by

$H(n; d^*)$. If H is a non-decreasing function of n, then the required sample size is the smallest integer not less than the solution of the equation

$$H(n; d^*) = P^*. \tag{5.2}$$

In such a case the required sample size is unique. Further if the limit of H as $n \to \infty$, is one then the above equation has a solution for any specified $P^* < 1$.

Remarks on the need and definition of the preference zone. Now we can answer the question—why we restrict our attention to the preference zone in writing the probability requirement (2.2)?

If there were no such restriction, then the sample size necessary is the smallest integer value of n for which

$$\inf_{\theta \in \Omega} P(CS \mid \theta) \geqq P^*. \tag{5.3}$$

We have shown that regardless of the sample size the infimum of the $P(CS \mid \theta)$ over Ω is $P(c, k, s, t)$, which is the lower bound for P^*. Thus without the restriction to the preference zone, we cannot achieve our goal however large our sample may be.

The choice of the preference zone is equivalent to the choice of the d-function. This choice is governed by the behavior of $P(\theta, \theta')$ which depends on the form of $G_n(\cdot \mid \theta)$. It should be noted that in some problems, it is sufficient to define the preference zone through one restriction such as $d(\theta_{[k-t+1]}, \theta_{[k-t]}) \geqq d^*$, whereas in other problems it may be desirable to introduce more than one restriction. One such example is the problem where θ is the mean of a Poisson population [8]. The particular definition of the preference zone given in any specific case enables us to determine explicitly the infimum of $P(\theta, \theta')$. In some cases obtaining this infimum may not be a simple matter and may even have to be obtained by numerical methods. One such example is the problem where θ is the probability of success for a Bernoulli variable; this problem, for the case $c = s = t = 1$, is considered by Sobel and Huyett [9].

We shall now see how the Equation (5.1) simplifies in the cases when θ is either a location or a scale parameter for the family $\mathcal{G}$.

Case (i) *θ is a location parameter for the family $\mathcal{G}$:*

In this case we have that for all x

$$G_n(x \mid \theta) = G_n(x - \theta), \quad \text{where} \quad G_n(x) = G_n(x \mid 0). \tag{5.4}$$

Using (5.4) in (4.7) and transforming the variable of integration, we have

$$P(\theta, \theta_0) = H_L(n; \delta) = \int_{-\infty}^{\infty} I[G_n(x + \delta); k - t - s + c, s - c + 1] \cdot dI[G_n(x); t - c + 1, c] \tag{5.5}$$

where $\delta = \theta - \theta_0$. Since $H_L(n; \delta)$ depends on θ, θ_0 only through δ, we define the "natural" distance measure for such a problem as

$$d(a, b) = a - b. \tag{5.6}$$

It is easy to see that

$$\inf_{\theta \in \Omega(d^*)} P(CS \mid \theta) = \inf_{\delta \geqq d^*} H_L(n; \delta) = H_L(n; d^*). \tag{5.7}$$

Hence the Equation (5.1) reduces to

$$H_L(n; d^*) \geqq P^*. \tag{5.8}$$

$H_L(n; d^*)$ can be expressed in any one of the following equivalent forms.

$$\begin{aligned} H_L(n; d^*) &= (t!/(t-c)!(c-1)!) \sum_{\alpha=0}^{s-c} \binom{k-t}{\alpha} \int_{-\infty}^{\infty} G_n^{k-t-\alpha}(x + d^*) \\ &\quad \cdot [1 - G_n(x + d^*)]^{\alpha} G_n^{t-c}(x)[1 - G_n(x)]^{c-1} dG_n(x) \\ &= \int_{-\infty}^{\infty} I[G_n(x + d^*); c', s - c + 1]\, dI[G_n(x); t - c + 1, c] \\ &= ((k-t)!/(s-c)!(c'-1)!) \sum_{\alpha=0}^{t-c} \binom{t}{\alpha} \int_{-\infty}^{\infty} G_n^{\alpha}(x - d^*) \\ &\quad \cdot [1 - G_n(x - d^*)]^{t-\alpha} G_n^{c'-1}(x)[1 - G_n(x)]^{s-c} dG_n(x) \\ &= \int_{-\infty}^{\infty} \{1 - I[G_n(x - d^*); t - c + 1, c]\} \\ &\quad \cdot dI[G_n(x); c', s - c + 1]. \end{aligned} \tag{5.9}$$

Case (ii) θ *is a scale parameter for the family* $\mathcal{G}$:
In this case we have that for all x.

$$G_n(x \mid \theta) = G_n(x/\theta), \quad \text{where} \quad G_n(x) = G_n(x \mid 1) \quad \text{and} \quad G_n(0) = 0. \tag{5.10}$$

By transforming the variable of integration we obtain

$$\begin{aligned} P(\theta, \theta_0) = H_S(n; \delta_1) &= \int_0^{\infty} I[G_n(x\delta_1); k - t - s + c, s - c + 1] \\ &\quad \cdot dI[G_n(x); t - c + 1, c] \end{aligned} \tag{5.11}$$

where $\delta_1 = \theta/\theta_0$. Here we define the distance measure as

$$d(a, b) = a/b. \tag{5.12}$$

Arguing as in the case (i) one can see that the Equation (5.1) reduces to

$$H_S(n; d^*) \geqq P^*. \tag{5.13}$$

One can obtain the various (equivalent) expressions for $H_S(n; d^*)$ from those of $H_L(n, d^*)$ by changing $x + d^*$ to xd^*, $x - d^*$ to x/d^* and changing the lower limit of integration from $-\infty$ to 0.

6. Particular cases of goal I which are of special interest. Two particular cases of Goal I, corresponding to $c = t$ when $s \geqq t$ and $c = s$ when $s \leqq t$, are of special interest. These are:

GOAL 1. Selection of a subset of size s which includes the t best populations, where $s \geqq t$.

GOAL 2. Selection of a subset of size s which includes any s of the t best populations, where $s \leqq t$.

It should be noted that these two goals coincide when $s = t$. Then the common goal is the selection of the t best populations (without ordering). The solutions

to the selection problem in relation to the above goals have been mentioned earlier by the author in an abstract [6]. We shall now give the final results for these particular cases.

GOAL 1. Here the lower bound for P^* is $\binom{k-t}{s-t}/\binom{k}{s}$. Selection of a subset which includes the t best populations (those with parameter values $\theta_{[k-t+1]}$, $\theta_{[k-t+2]}$, $\cdots$, $\theta_{[k]}$) is a correct selection. Now the sample size needed to achieve this goal, when the procedure R_s is used, is the smallest value of n for which

$$\inf_{\theta\epsilon\Theta} Q_1(\theta, n) \geqq P^* \tag{6.1}$$

where

$$\begin{aligned} Q_1(\theta, n) &= \binom{k-t}{k-s} \int_{-\infty}^{\infty} [1 - G_n(x \mid \theta)]^t [1 - G_n(x \mid \theta')]^{s-t} d[G_n^{k-s}(x \mid \theta')] \\ &= \int_{-\infty}^{\infty} \{1 - I[G_n(x \mid \theta); 1, t]\}\, dI[G_n(x \mid \theta'); k - s, s - t + 1]. \end{aligned} \tag{6.2}$$

Here θ', as a function of θ, is determined by $d(\theta, \theta') = d^*$.

GOAL 2. In this case the lower bound to P^* is $\binom{t}{s}/\binom{k}{s}$. Selecting any subset of size s of the t best populations constitutes a correct selection. The sample size necessary is the smallest value of n for which

$$\inf_{\theta\epsilon\Theta} Q_2(\theta, n) \geqq P^* \tag{6.3}$$

where

$$\begin{aligned} Q_2(\theta, n) &= (t!/s!\,(t - s - 1)!) \int_{-\infty}^{\infty} G_n^{k-1}(x \mid \theta') G_n^{t-s}(x \mid \theta) \\ &\quad \cdot [1 - G_n(x \mid \theta)]^{s-1} dG_n(x \mid \theta), \\ &= \int_{-\infty}^{\infty} I[G_n(x \mid \theta'); k - t, 1]\, dI[G_n(x \mid \theta); t - s + 1, s]. \end{aligned} \tag{6.4}$$

Here also θ' is determined by the relation $d(\theta, \theta') = d^*$.

It is easy to see that Goal I is less "stringent" than both Goal 1 and Goal 2. So one expects that, for fixed c, k, t, P^* and d^*, the sample size necessary to achieve Goal I will be smaller than the sample size necessary to achieve Goal 1 (if $s \geqq t$) or Goal 2 (if $s \leqq t$). This result follows directly from the theorem given below. Let $n(c, s)$ denote the sample size necessary to achieve Goal I.

THEOREM. *For fixed k, t, s, P^*, d^*, and for any distance measure*

$$n(c + 1, s) \geqq n(c, s), \tag{6.5}$$

provided $c + 1 \leqq \min(s, t)$, i.e., *provided Goal I is meaningful with c replaced by $c + 1$.*

PROOF. Let c_0 be an arbitrary integer such that c_0 and $c_0 + 1$ are admissible values of c. The result directly follows from the fact that, when the procedure R_s is used, a correct selection for Goal I with $c = c_0 + 1$ implies a correct selection for Goal I with $c = c_0$.

From the theorem, it follows that $n(c, s) \leqq n_1(s)$ and $n(c, s) \leqq n_2(s)$ where $n_i(s)$ is the sample size necessary to achieve Goal i ($i = 1, 2$).

7. A sufficient condition for the existence of the required sample size. It

has been shown that the required common sample size is the smallest value of n for which

$$\inf_{\theta \epsilon \Theta} Q(\theta, n) \geqq P^*, \tag{7.1}$$

where $Q(\theta, n)$ is given by (4.14). The required sample exists provided the left side of (7.1) tends to one as n tends to infinity. We will now obtain a sufficient condition for the same, under the assumption that the infimum of $Q(\theta, n)$ is its value at $\theta = \theta_0$. This assumption is satisfied in many cases of interest; in particular this is true when θ is either a location parameter or a scale parameter for the family $\mathcal{G}$. Thus we need to find a sufficient condition for the limit of $Q(\theta_0, n)$, as $n \to \infty$, to be one. Now

$$\begin{aligned} Q(\theta_0, n) &= P[c\text{th largest of } (Y_{k-t+1}, \cdots, Y_k) > (s - c + 1)\text{st largest of} \\ &\qquad (Y_1, \cdots, Y_{k-t})] \\ &\geqq P[\min (Y_{k-t+1}, \cdots, Y_k) > \max (Y_1, \cdots, Y_{k-t})] \end{aligned} \tag{7.2}$$

where $Y_1, \cdots, Y_k$ is a set of independent random variables such that the cdf of $Y_i (1 \leqq i \leqq k - t)$ is $G(\cdot \mid \theta_0')$ and the cdf of $Y_i (k - t + 1 \leqq i \leqq k)$ is $G(\cdot \mid \theta_0)$. The constant θ_0' is given by the relation $d(\theta_0, \theta_0') = d^*$. From (7.2), we obtain

$$\begin{aligned} 1 - Q(\theta_0, n) &\leqq 1 - P\left[\bigcap_{i,j} \{Y_i > Y_j\}, \begin{matrix} i = k - t + 1, \cdots, k \\ j = 1, \cdots, k - t \end{matrix}\right] \\ &= P\left[\bigcup_{i,j} \{Y_i < Y_j\}\right] \\ &\leqq \sum_{i,j} P[Y_i < Y_j[= t(k - t)\, P[Y_k < Y_1]. \end{aligned} \tag{7.3}$$

From (7.3), a sufficient condition for the existence of the required sample size is

$$\lim_{n \to \infty} P[Y_k < Y_1] = 0 \tag{7.4}$$

$$\text{i.e., } \lim_{n \to \infty} \int_{-\infty}^{\infty} G_n(x \mid \theta_0)\, dG_n(x \mid \theta_0') = 0.$$

In some particular cases it may not be easy to verify this condition. So we will obtain an easily verifiable sufficient condition for (7.4) to be true, on the assumption that the variances of the distributions defined by $G_n(\cdot \mid \theta_0)$ and $G_n(\cdot \mid \theta_0')$ are finite. Let

$$Z = [(Y_k - Y_1) - E(Y_k - Y_1)]/[\text{Var } (Y_k - Y_1)]^{\frac{1}{2}}, \tag{7.5}$$

$$a = E(Y_k - Y_1)/[\text{Var } (Y_k - Y_1)]^{\frac{1}{2}}.$$

From Lemma 4.1, it follows that a is non-negative. Using Chebyschev's inequality, we obtain

$$P[Y_k < Y_1] = P[Z < -a] \leqq P[|Z| < a] \leqq 1/a^2. \tag{7.6}$$

Thus a sufficient condition for (7.4) to be true is

$$\lim_{n\to\infty} [\text{Var } Y_k + \text{Var } Y_1]/[EY_k - EY_1]^2 = 0. \tag{7.7}$$

8. An example. In this section we apply the general results to the case where the distributions $F(\cdot \mid \theta)$ (which characterize the populations) are normal distributions. Let $\phi(x)$ and $\Phi(x)$ be the density and the distribution functions of the standard normal distribution. Here

$$F(x \mid \theta) = \Phi[(x - \theta)/\sigma]. \tag{8.1}$$

We assume that the variances of all the k populations are equal and the common value σ^2 is known. Clearly one chooses $T_i = \bar{X}_i$ (the mean of the sample from Π_i), so that

$$G_n(x \mid \theta_i) = \Phi[\{(x - \theta_i)n^{\frac{1}{2}}\}/\sigma]. \tag{8.2}$$

Since $\mathcal{G} = \{G_n(\cdot \mid \theta) : \theta \in R\}$ is a location parameter family we define the distance measure as $d(a, b) = a - b$. Here $H_L(n; d^*)$ (see (5.9)) reduces to

$$H(\lambda) = \int_{-\infty}^{\infty} I[\Phi(x + \lambda); c', s - c + 1]\, dI[\Phi(x); t - c + 1, c] \tag{8.3}$$

where $\lambda = (d^* n^{\frac{1}{2}})/\sigma$ and $c' = k - t - (s - c)$. It is easy to see that H is an increasing function of n and it tends to one as n tends to infinity. Using results of Section 5, we obtain that the required common sample size is the smallest integer not less than

$$n_0 = (\lambda\sigma/d^*)^2 \tag{8.4}$$

where λ is given by

$$H(\lambda) = P^*. \tag{8.5}$$

This equation has a unique solution since H is an increasing function of λ. Tables giving λ-values (the solutions of (8.5)) for various c, k, s, t and P^* values are under preparation and will be published in the near future. When $t \leqq s$ and $s = c$, $H(\lambda)$ reduces to

$$(t!/(t - s)!(s - 1)!) \int_{-\infty}^{\infty} \Phi^{k-t-1}(x + \lambda)\Phi^{t-s}(x)[1 - \Phi(x)]^{s-1}\phi(x)\, dx.$$

This expression is given by Bechhofer (see (25) in [3]). We have to find the value λ for which the above expression is P^* and use that value in (8.4) to obtain the sample size when Case 2 of Goal I is the experimenter's goal. More details of this case and other particular cases of the distributions F will be reported in a separate paper.

Acknowledgment. My sincere thanks are due to Professor Milton Sobel, under whose guidance this research was carried.

REFERENCES

[1] Alam, K. and Rizvi, M. H. (1965). Selection from multivariate normal populations. Technical Report No. 65-1. Mathematics Department. The Ohio State University.

[2] BARR, D. R. and RIZVI, M. H. (1964). Ranking and selection problems of uniform distributions. *Ann. Math. Statist.* **35** 1842. Abstract #16.

[3] BECHHOFER, R. E. (1954). A single-sample multiple decision procedure for ranking means normal populations with known variances. *Ann. Math. Statist.* **25** 16–39.

[4] BECHHOFER, R. E. and SOBEL, M. (1954). A single-sample multiple decision procedure for ranking variances of normal populations. *Ann. Math. Statist.* **25** 273–289.

[5] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. John Wiley and Sons, New York.

[6] MAHAMUNULU, D. M. (1965). A class of ranking and selection procedures (preliminary report). *Ann. Math. Statist.* **36** 728. Abstract #7.

[7] RIZVI, M. H. (1963). Ranking and selection problems of normal populations using the absolute values of their means: fixed sample size case. Technical Report No. **31**, Department of Statistics, Univ. of Minnesota.

[8] SOBEL, M. (1963). Single sample ranking problems with Poisson populations. Technical Report No. 19, Department of Statistics, Univ. of Minnesota.

[9] SOBEL, M. and HUYETT, M. J. (1957). Selecting the best one of several binomial populations. *Bell System Technical J.* **36** 537–576.