The Annals of Statistics
1977, Vol. 5, No. 5, 969-983

ON SELECTING A SUBSET CONTAINING THE BEST POPULATION—A BAYESIAN APPROACH

BY PREM K. GOEL[1] AND HERMAN RUBIN[2]

Purdue University

The problem of selecting a subset of k populations $\pi_1, \cdots, \pi_k$, which contains the "best" population, is considered. The unknown values $\theta_1, \cdots, \theta_k$ are the characteristics associated with $\pi_1, \cdots, \pi_k$ and the unknown population associated with $\theta_{[k]} = \max_i \theta_i$ is called the "best." It is assumed that, given $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$ the pdf of the independent random variables $X_1, \cdots, X_k$ belong to a monotone likelihood ratio family, the prior distribution of $\boldsymbol{\theta}$ is exchangeable, and the loss function is a linear combination of two components, namely the subset size $|s|$ and the distance between the "best" and the "best" in the selected subset s, i.e., $L(\boldsymbol{\theta}, s) = c|s| + [\theta_{[k]} - \max_{j:\pi_j \in s} \theta_j]$.

It is shown that the Bayes rule depends on *at most* $(k-1)$ computable expressions. Some lower and upper bounds on the differences of Bayes risks are given to help reduce the amount of computation for the Bayes rule. If X_i has a normal distribution with mean θ_i and known variance σ^2, then it is shown that (i) for $k = 2$, the Bayes rule with vague prior knowledge and the classical rule are the same if the probability of correct selection, P^*, is chosen as a suitable function of c, and (ii) if $c/\sigma \geqq 1/\pi^{\frac{1}{2}}$, then the Bayes rule selects only one population and if $.2821 \leqq c/\sigma < 1/\pi^{\frac{1}{2}}$, then it selects at most two populations. The tables for implementing the Bayes rule for normal populations are also given.

1. Introduction and summary. During the early fifties, it was pointed out by several researchers, e.g., Bahadur (1950), that testing the homogeneity of population means or variances is not a satisfactory solution to a comparison of the performance of several populations. One would, generally, want to either rank them according to their performance or select one or more from among them for future use or further evaluation. These problems are known as ranking and selection problems.

Let $X_1, \cdots, X_k$ be random variables representing the k populations $\pi_1, \cdots, \pi_k$ respectively, with X_i having the pdf $f(\cdot, \theta_i)$, $\theta_i \in \Theta \subset R$. In many cases X_i are sufficient statistics for θ_i based on a random sample from π_i, $i = 1, \cdots, k$. The population π_i is characterized by the unknown parameter θ_i and the ranking or selection is based on the information contained in $X_1, \cdots, X_k$. The population $\pi_{[k]}$ associated with the largest θ value will be called the best.

Received November 1975; revised February 1977.

[1] Research was supported in part by the National Science Foundation under grant # SOC74-02071 A01 at Carnegie-Mellon University.

[2] Research supported in part by the National Science Foundation under grant # 74-07836 at Purdue University.

AMS 1970 *subject classifications.* Primary 62F07; Secondary 62G30.

Key words and phrases. Subset selection, nonlinear loss function, normal populations, monotone likelihood ratio family, exchangeable prior distribution, Bayes rules

Two formulations to these problems have been suggested in the classical framework. The first one, proposed by Bechhofer (1954) and known as the indifference zone formulation, allows the experimenter to select one population which is the best with a fixed probability P^*, whenever the unknown parameters are outside a zone of indifference. The second one, proposed by Gupta (1956) and known as the subset selection formulation, allows the experimenter to select a subset of random size which contains the best population with a probability P^* or more. The event that the subset contains the best population is denoted by CS. The idea underlying the classical subset selection procedures is to choose a rule satisfying $P_{\theta}(CS) \geqq P^*$ for all $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k) \in \Omega = \Theta^k$, such that the size of the subset has some desirable distributional properties. Extensive tables are available to implement these procedures for various parametric and nonparametric families of distributions.

In the decision theoretic framework, most of the literature in this area is devoted to (i) ranking the populations, (ii) selecting the "best" or selecting the t "best" populations and (iii) comparing all the population with a standard. The pioneer work in this direction is by Bahadur and Robbins (1950), where the problem of selecting a population with greater mean from two normal populations is discussed. Bahadur and Goodman (1952) assumed that the loss function is a linear combination of $L_i(\boldsymbol{\theta})$, $i = 1, \cdots, k$, where $L_i(\boldsymbol{\theta})$ is the loss if the ith population is selected in the subset. Similar loss functions were assumed by Lehmann (1966), Eaton (1967) and Alam (1973). Dunnett (1960) compared the operating characteristics of various rules for selecting the largest of k normal population means in the Bayesian setup with the loss $L_i(\boldsymbol{\theta}) = c'(\theta_{[k]} - \theta_i)$, $i = 1, \cdots, k$. The problem of choosing a single population, when the utility function is equal to the probability of coverage in a specified interval, is discussed in Guttman and Tiao (1964).

It is generally agreed that there are two components of loss in the subset selection formulation. The first depends only on the populations selected in the subset and the second depends on whether or not the selected subset contains the best population. We believe that the classical subset selection rules are hard to justify in a decision theoretic framework because these procedures are chosen such that the second component of the loss is fixed at a value which is difficult to link with the loss in real problems, and the first component of the loss is minimized. The decision theoretic procedures mentioned above do not seem to be appropriate in situations where a subset is selected for further evaluation of the selected populations mainly because they ignore the second component of the loss and secondly because all these procedures specify the subset size in advance, whereas it should depend on the information available from the sample.

Deeley and Gupta (1968) proved that the Bayes rule for selecting a subset of k normal populations selects only one population if the loss functions is a linear combination of $L_i(\boldsymbol{\theta}) = \theta_{[k]} - \theta_i$, $i = 1, \cdots, k$. It follows that a linear loss function does not represent the loss of an experimenter who wants to examine one

or more populations. This leads us to believe that one has to consider nonlinear loss functions which consider both the components of loss simultaneously. The optimal procedures corresponding to this kind of loss functions will not be as easy to implement in practice as the ad hoc rules. But this should be a secondary consideration in this age of computers, since one should be willing to use "optimal" procedures if they depend on "computable" expressions.

Studden (1967) assumed that $\theta_{[1]} \leqq \cdots \leqq \theta_{[k]}$ are k fixed *known* values but the correct pairing of the populations and the parameters are unknown. For the loss function which is a linear combination of two components, namely (i) the sum of $L_i(\theta)$ over $i \ni \pi_i \in s$, and (ii) the function $I_{\{\theta_{[k]} \notin s\}}$, he obtained the best invariant rule which depends on a computable expression.

We assume that the loss function is a linear combination of the two components, the first one being the size of the selected subset s and the second one being $[\theta_{[k]} - \max_{\{j:\pi_j \in s\}} \theta_j]$. This loss function is reasonable when the cost c of further evaluating each population is equal, and the loss due to the incorrect selection is proportional to the difference between the "best" and the "best" in the selected subset. We shall now present a summary of the results obtained in this paper.

In Section 2, it is assumed that the density function $f(x, \theta)$ possesses a monotone likelihood ratio property and that, a priori, the random variables $\theta_1, \cdots, \theta_k$ are exchangeable. In Theorem 1 it is proved that the Bayes rule is obtainable by computing at most $(k - 1)$ differences, Δ_m, $m = 1, \cdots, k - 1$ between the posterior risks of the decision rules d_{m+1} and d_m, where d_m selects the subset containing the m largest observations. An easily computable lower bound for Δ_m is given, which is useful in obtaining the Bayes rule. Next it is assumed that the prior distribution is a mixture of i.i.d. random variables and a simplified version of the Bayes rule is given in Theorem 2.

In Section 3, it is assumed that the observation vector $\mathbf{x}$ is a location parameter in the posterior distribution and the Bayes rule is given in Theorem 3. The relationship between the cost c and the maximum size of the subset is given in Theorem 4.

In Section 4, it is assumed that X_i has a normal distribution with mean θ_i and known variance σ^2 and that the prior of $\boldsymbol{\theta}$ is exchangeable multivariate normal. Some more simplification of the Bayes rule is done. If γ denotes the posterior standard deviation of θ_i, then it is shown that (i) for $c/\gamma \geqq 1/\pi^{\frac{1}{2}}$, the Bayes rule selects the population corresponding to the largest observation and (ii) for $.2821 \leqq c/\gamma \leqq 1/\pi^{\frac{1}{2}}$, the Bayes rule selects the populations corresponding to the largest two observations. Furthermore, for $k = 2$ the classical selection rule is the same as the Bayes rule with vague prior knowledge provided c and P^* satisfy (4.6).

2. Notation and formulation of the decision problem. Let $\theta_{[1]} \leqq \cdots \leqq \theta_{[k]}$ denote the ordered values of θ_i's and $\pi_{[i]}$ denote the unknown population associated with $\theta_{[i]}$. It is assumed that, given $\boldsymbol{\theta} \in \Omega = \Theta^k$, the random variables

$X_1, \cdots, X_k$ are independently distributed and that the pdf $f(x, \theta)$ possesses the monotone likelihood ratio (MLR) property.

Let $x_{(1)} \leqq x_{(2)} \leqq \cdots \leqq x_{(k)}$ denote the ordered observations where the ties for a label are broken at random, and $\pi_{(i)}$ and $\theta_{(i)}$ denote the π and the θ associated with $x_{(i)}$, $i = 1, \cdots, k$.

The elements of the action space $\mathscr{A}$ are all possible nonempty subsets of $\pi_1, \cdots, \pi_k$. Let these subsets be denoted by s_j, $j = 1, 2, \cdots, 2^k - 1$. For $\boldsymbol{\theta} \in \Omega$ and $s \in \mathscr{A}$, the loss function is assumed to be of the form

$$L(\boldsymbol{\theta}, s) = c|s| + [\theta_{[k]} - \theta_{\{s\}}], \tag{2.1}$$

where $|s|$ denotes the number of populations in the subset s, $\theta_{\{s\}}$ denotes $\max_{j:\pi_j \in s} \theta_j$ and $c > 0$ is to be interpreted as the relative cost of further evaluating a population versus being one unit away from the best population. We believe that one can determine the cost c more realistically than the P^*-value for a classical procedure in most practical problems.

Since only Bayes decision rules will be used in this paper, one only needs to consider nonrandomized decision rules; see DeGroot (1970), Section 8.5. The set $\mathscr{D}$ of all nonrandomized decision (selection) rules containing the functions $d_j(x_1, \cdots, x_k)$, $j = 1, \cdots, 2^k - 1$, is defined as follows:

For $j = 1, 2, \cdots, k$, the decision rule d_j chooses the subset s_j^* with probability 1 where $s_j^* = \{\pi_{(k)}, \pi_{(k-1)}, \cdots, \pi_{(k-j+1)}\}$ and for $j = k + 1, \cdots, 2^k - 1$, the d_j's and the remaining subsets in $\mathscr{A}$ are associated one to one arbitrarily.

Given $\mathbf{x} = (x_1, \cdots, x_k)$, the posterior risk of a decision $d \in \mathscr{D}$, which selects the subset $s \in \mathscr{A}$ with probability 1, is denoted by

$$r(d, \mathbf{x}) = c|s| + E[\theta_{[k]} - \theta_{\{s\}} \mid \mathbf{x}]. \tag{2.2}$$

The following result reduces the number of decision rules to be compared for the Bayes rule from $2^k - 1$ to k.

LEMMA 1. *If the prior distribution of $\boldsymbol{\theta}$ is symmetric on Ω then the Bayes rule d^* is given by*

$$r(d^*, \mathbf{x}) = \min_{j=1,\cdots,k} r(d_j, \mathbf{x}). \tag{2.3}$$

PROOF. Let us partition the action space $\mathscr{A}$ into k components $\mathscr{A}_m$, $m = 1, 2, \cdots, k$, where $\mathscr{A}_m$ contains all the subsets of size m. Now the mth decision problem with the observation vector $\mathbf{x}$, the action space $\mathscr{A}_m$, and the loss function $L(\boldsymbol{\theta}, s) = cm + (\theta_{[k]} - \theta_{\{s\}})$ for $s \in \mathscr{A}_m$, is equivalent to subdividing $\pi_1, \cdots, \pi_k$ into two subsets (γ_1, γ_2) where γ_1 is of size m and γ_2 is of size $k - m$.

Given the above structure, the mth decision problem is invariant under the permutation group and the loss function satisfies the monotonicity and invariance conditions (3.4) and (3.5) in Eaton (1967). It follows from his Theorem 4.1 that the rule which assigns the populations $\pi_{(k)}, \cdots, \pi_{(k-m+1)}$ to γ_1 and $\pi_{(k-m)}, \cdots, \pi_{(1)}$ to γ_2 is Bayes against a symmetric prior distribution of $\boldsymbol{\theta}$ on Ω, i.e.,

$$r(d_m, \mathbf{x}) \leqq r(d, \mathbf{x}) \qquad \text{for all} \quad d \in \mathscr{D}_m,$$

where $\mathscr{D}_m$ is the set of all nonrandomized decision rules associated with $\mathscr{A}_m$. Hence one only needs to compare the decision rules $d_1, \cdots, d_k$ to obtain the Bayes rule in this problem.

REMARK 1. This result holds for every loss function which satisfies the conditions (3.4) and (3.5) in Eaton (1967). However, this specific loss function is being considered to get some insight into the nature of the Bayes rule.

REMARK 2. For Lemma 1 to hold, the assumption of MLR property of $f(x, \theta)$ can be replaced by stochastically increasing property (SIP). However, one needs an extra condition that the posterior distribution of $\boldsymbol{\theta}$ has the SIP property. This proof of Lemma 1 is valid, if we use Theorem 2.2 in Alam (1973) in place of Theorem 4.1 in Eaton (1967).

REMARK 3. After this paper was submitted for publication, two more papers came to our attention that consider nonlinear loss functions. Bickel and Yahav (1977) assume that $\theta_{[1]}, \cdots, \theta_{[k]}$ are known and consider the loss function of the form

$$\frac{1}{|S|} \sum_{\{j:\pi_j \in S\}} (\theta_{[k]} - \theta_j) + rI_{\{\theta_{[k]} \notin S\}} . \tag{2.4A}$$

They obtain the best invariant rule for the normal pdf and then digress from the decision theoretic approach to simplify this rule as $k \to \infty$. Chernoff and Yahav (1977) consider the loss function of the form

$$r(\theta_{[k]} - \theta_{\{s\}}) - \frac{1}{|S|} \sum_{\{j:\pi_j \in S\}} \theta_j . \tag{2.4B}$$

They obtain the result in Lemma 1 for the normal means problem with an exchangeable normal prior distribution of $\boldsymbol{\theta}$ and then compare the operating characteristics of the Bayes procedure with those of fixed subset size procedures of Desu and Sobel (1968) and Gupta (1965) by means of a Monte Carlo study.

Since the computation of the risk $r(d_j, \mathbf{x})$, $j = 1, \cdots, k$ is a difficult task, we further analyze the problem to simplify the computation of the Bayes procedure in the remainder of this section and Section 3. It should be noted that the techniques used here may not work for loss functions that are more complicated.

Unless otherwise mentioned, all the expectations below are with respect to the posterior distribution of $\boldsymbol{\theta}$ on Ω, given the observation vector $\mathbf{x} = (x_1, \cdots, x_k)$.

LEMMA 2. *For $m = 1, \cdots, k-1$, let Δ_m denote*

$$\Delta_m = r(d_{m+1}, \mathbf{x}) - r(d_m, \mathbf{x}) . \tag{2.5}$$

If the prior distribution, $\xi(\boldsymbol{\theta})$, of $\boldsymbol{\theta}$ is symmetric on Ω, then

$$\Delta_m \geqq \Delta_{m-1} , \qquad m = 2, 3, \cdots, k-1 . \tag{2.6}$$

PROOF. Let $m \in \{1, 2, \cdots, k-1\}$ and let z^+ denote the positive part of z. It follows from (2.2) that Δ_m can be written as

$$\Delta_m = c - E[(\theta_{(k-m)} - \theta_{\{s_m^*\}})^+] . \tag{2.7}$$

Let $\eta_m^* = (\theta_{(k-m)} - \theta_{\{s_{m-1}^*\}})^+$ and $\eta_{m-1} = (\theta_{(k-m+1)} - \theta_{\{s_{m-1}^*\}})^+$. Since $\theta_{\{s_m^*\}} \geqq \theta_{\{s_{m-1}^*\}}$, it follows from (2.7) that

$$\begin{aligned}\Delta_m - \Delta_{m-1} &\geqq E\{\eta_{m-1} - \eta_m^*\}, \\ &= b \int_\Omega (\eta_{m-1} - \eta_m^*) f(\mathbf{x}, \boldsymbol{\theta})\, d\xi(\boldsymbol{\theta}), \\ &= b \int_B (\eta_{m-1} - \eta_m^*)(f(\mathbf{x}, \boldsymbol{\theta}) - f(\mathbf{x}, \boldsymbol{\theta}'))\, d\xi(\boldsymbol{\theta}),\end{aligned}$$

where b is a normalizing factor, $B = \{\boldsymbol{\theta} : \theta_{(k-m)} \leqq \theta_{(k-m+1)}\}$ and $\boldsymbol{\theta}'$ is obtained from $\boldsymbol{\theta}$ by interchanging the components $\theta_{(k-m)}$ and $\theta_{(k-m+1)}$. For all $\boldsymbol{\theta} \in B$, $\eta_{m-1} \geqq \eta_m^*$ and $f(\mathbf{x}, \boldsymbol{\theta}) - f(\mathbf{x}, \boldsymbol{\theta}')$ is nonnegative by the MLR property of $f(\mathbf{x}, \boldsymbol{\theta})$. Therefore, $\Delta_m - \Delta_{m-1} \geqq 0$.

REMARK 4. It follows from (2.7) that the selection of $\pi_{(k-m)}$ in the optimal subset depends on a symmetric function of the distances between $\pi_{(k-m)}$ and $\pi_{(i)}$, $i = k - m + 1, \cdots, k$. This property is also satisfied by the rules obtained in Studden (1967) and Chernoff and Yahav (1977). This suggests that the ad hoc subset selection rules, in which the selection of $\pi_{(i)}$ in the subset depends on the distance between $x_{(i)}$ and $x_{(k)}$ only, cannot be optimal in a decision theoretic framework for any loss function in which both the components of loss are considered.

The next result now follows from Lemmas 1 and 2.

THEOREM 1. *Let the prior distribution of $\boldsymbol{\theta}$ be symmetric on Ω. If the set $A = \{j : \Delta_j \geqq 0\}$ is empty then the Bayes rule $d^* = d_k$, otherwise $d^* = d_m$ where m is the smallest number in the set A.*

It follows from Theorem 1 that one needs to compute *at most* $k - 1$ expressions Δ_m, $m = 1, \cdots, k - 1$ for a complete specification of the Bayes rule. Since these expressions involve multiple integrals, it will be worthwhile to obtain some easily computable lower and upper bounds on Δ_m which will eliminate the computation of some of these integrals. One lower bound on Δ_m, which will also be used in obtaining an "approximate" Bayes rule is given below.

LEMMA 3. *For $m \in \{1, 2, \cdots, k - 1\}$, Δ_m satisfies the relationship*

$$\Delta_m \geqq c - E\{(\theta_{(k-m)} - \theta_{(k)})^+\}. \tag{2.8}$$

PROOF. The expression (2.8) follows from (2.7) and the fact that $\theta_{\{s_m^*\}} \geqq \theta_{(k)}$.

No further simplification of the Bayes rule is possible unless we make some more assumptions about the prior distribution of $\boldsymbol{\theta}$.

In the remainder of this paper, we assume that, given $W = \omega$, $\theta_1, \cdots, \theta_k$ are i.i.d. random variables with density $\xi(\cdot, \omega)$ and the distribution of W is known. We shall call this prior as a mixture of i.i.d. random variables.

It follows that given $\mathbf{X} = \mathbf{x}$ and $W = \omega$, the random variables $\theta_1, \cdots, \theta_k$, are a posteriori, independently distributed and we denote the pdf and the cdf of $\theta_{(i)}$ given $x_{(i)}$ and ω by $g_{(i)}(\theta_{(i)}) = g(\theta_{(i)} \mid x_{(i)}, \omega)$ and $G_i(\theta_{(i)}) = G(\theta_{(i)} \mid x_{(i)}, \omega)$. Furthermore let $Q(\omega \mid \mathbf{x})$ denote the conditional cdf of W given $\mathbf{x}$. With this specification of the prior, we have the following result.

LEMMA 4. *Let* $I_m(\omega)$ *denote the integral*

$$I_m(\omega) = \int_{-\infty}^{\infty} \prod_{i=k-m+1}^{k} G_i(z)(1 - G_{k-m}(z))\, dz\,, \tag{2.9}$$

then Δ_m *is given by*

$$\Delta_m = c - \int_{-\infty}^{\infty} I_m(\omega)\, dQ(\omega \mid \mathbf{x})\,, \quad m = 1, \cdots, k-1\,. \tag{2.10}$$

PROOF. Let Z denote the random variable $(\theta_{(k-m)} - \theta_{[s_m^*]})^+$ and let $H(\lambda)$ denote the posterior probability of the event $Z > \lambda$. We have

$$H(\lambda) = P[(\theta_{(k-m)} - \theta_{(i)})^+ > \lambda \text{ for } i = k-m+1, \cdots, k]\,.$$

It follows that, for $\lambda > 0$

$$H(\lambda) = P[\theta_{(k-m)} > \lambda + \theta_{(i)}, \text{ for } i = k-m+1, \cdots, k]\,. \tag{2.11}$$

Let $H(\lambda \mid \omega) = P[Z > \lambda \mid W = \omega]$, it follows from (2.11) that

$$H(\lambda \mid \omega) = \int_{-\infty}^{\infty} \prod_{i=k-m+1}^{k} G_i(z) g_{k-m}(z + \lambda)\, dz\,. \tag{2.12}$$

Since Z is a nonnegative random variable, we have

$$E(Z \mid \omega) = \int_0^{\infty} H(\lambda \mid \omega)\, d\lambda\,. \tag{2.13}$$

Assuming that the posterior risk is finite for every $\mathbf{x}$, it follows from (2.12), (2.13) and a change in the order of integration that

$$E(Z \mid \omega) = \int_{-\infty}^{\infty} \prod_{i=k-m+1}^{k} G_i(z)(1 - G_{k-m}(z))\, dz\,. \tag{2.14}$$

The result now follows from (2.7) and (2.14).

Since, it can be proven that, for $i < j$

$$1 - G_i(z) < 1 - G_j(z) \qquad \text{for all } z\,, \tag{2.15}$$

the next result follows from (2.8) and (2.10).

LEMMA 5. *For* $m \in \{1, 2, \cdots, k-1\}$, Δ_m *satisfy the inequalities*

$$c - \min(a_m, v_m) \leqq \Delta_m \leqq c - u_m\,, \tag{2.16}$$

where

$$v_m = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G_{k-m+1}^m(z)(1 - G_{k-m}(z))\, dz\, dQ(\omega \mid \mathbf{x})\,, \tag{2.17}$$

$$u_m = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G_k{}^m(z)(1 - G_{k-m}(z))\, dz\, dQ(\omega \mid \mathbf{x})\,, \tag{2.18}$$

and

$$a_m = E[(\theta_{(k-m)} - \theta_{(k)})^+]\,.$$

It can be proved that

$$a_m = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G_k(z)(1 - G_{k-m}(z))\, dz\, dQ(\omega \mid x)\,. \tag{2.19}$$

Some interesting properties of a_m, u_m, and v_m are (i) $a_1 = u_1 = v_1$ and therefore $\Delta_1 = c - a_1 = c - v_1 = c - u_1$, (ii) for $m \geqq 2$, $u_m < a_m$ and $u_m < v_m$, however, there is no clear relationship between a_m and v_m, and (iii) a_m and u_m decrease as m increases.

We let r_i denote the min (a_i, v_i), and $i = 1, \cdots, k-1$ and let $r_k = 0$. The next result now follows from Theorem 1, Lemma 4 and Lemma 5.

THEOREM 2. *If the prior distribution of* $\boldsymbol{\theta}$ *is a mixture of* i.i.d. *random variables, then the Bayes rule* d^* *is given by*

(i) $c \geqq r_1 \Rightarrow d^* = d_1$,

(ii) $r_2 \leqq c < r_1 \Rightarrow d^* = d_2$, *and*

(iii) *for* $i = 3, \cdots, k$, $r_i \leqq c < r_{i-1}$ *and* $c \leqq u_j$ *for some* $j \in \{2, \cdots, i-1\} \Rightarrow d^* = d_m$, *where* m *is the smallest number* $\in \{j+1, \cdots, i\}$ *which satisfies* $\Delta_m \geqq 0$.

REMARK 5. Since the computational time for Δ_i increases as i increases, one should start the computations with Δ_{j+1} and continue it until m is found.

COROLLARY 1. *For* $k = 2$, *the Bayes rule is given by*

$$d^* = d_1 \quad \text{if} \quad c \geqq a_1\,, \quad \text{otherwise} \quad d^* = d_2\,.$$

An "approximate" Bayes rule R, which will select bigger subsets than d^*, but is the Bayes rule for $k = 2$, is suggested by Theorem 2.

Rule R. If $c \geqq a_1$, select s_1^*; if $a_m \leqq c < a_{m-1}$, select s_m^*, $m = 2, \cdots, k$.

The next result follows from (2.10) and (2.11).

COROLLARY 2. *If* ω *is a location parameter in the posterior distribution of* $\theta_{(i)}$, *given* $x_{(i)}$ *and* $W = \omega$, *then* Δ_m *is given by*

$$\Delta_m = c - I_m(0)\,. \tag{2.20}$$

In this case the integrals in (2.9) and (2.17) to (2.19) are univariate. Furthermore, since u_m, v_m, and a_m are functions of only two variables, they can be tabulated in order to reduce the computational time needed to obtain the Bayes rule via Theorem 2. However Δ_m is still a function of $m+1$ variables and hence it cannot be tabulated.

3. Location parameter posterior distribution. In this section, it is assumed that the posterior cdf of $\theta_{(i)}$ can be written in the form

$$G_i(\theta_{(i)} \mid x_{(i)}, \omega) = G[(\theta_{(i)} - b_1 x_{(i)} - b_0\omega)/b_2]\,, \quad i = 1, \cdots, k\,, \tag{3.1}$$

for some $b_1, b_2 > 0$. We say that $x_{(i)}$ and ω are location parameters in the posterior distribution of $\theta_{(i)}$. Since ω is a location parameter in this posterior distribution, it follows from Corollary 2 that the Bayes rule does not depend on $Q(\omega)$. However, the Bayes risk will certainly depend on $Q(\omega)$.

LEMMA 6. *If the posterior distribution of* $\boldsymbol{\theta}$ *satisfies* (3.1) *then*

(a) Δ_m *is a symmetric function of* $\nu_{k-m,j} = x_{(k-m)} - x_{(j)}$, $j = k-m+1, \cdots, k$, *which is increasing in each argument.*

(b) *The expressions* u_m *and* a_m, *defined by* (2.18) *and* (2.19), *are increasing functions of* $\nu_{k-m,k}$ *and furthermore* v_m, *defined in* (2.17), *is an increasing function of* $\nu_{k-m,k-m+1}$.

PROOF. It follows from (2.9), (3.1) and a change of variable that

$$I_m(0) = b_2 \int_{-\infty}^{\infty} \prod_{i=k-m+1}^{k} G\left(y + \frac{1}{b}\nu_{k-m,i}\right)(1 - G(y))\,dy\,, \tag{3.2}$$

where $b = b_2/b_1$.

Now, let $t_m(z)$ denote

$$t_m(z) = \int_{-\infty}^{\infty} G^m(u + z)(1 - G(u))\,du\,, \tag{3.3}$$

then it follows from (2.17) to (2.20), (3.1) and a change of variable that

$$a_m = b_2 t_1\left(\frac{1}{b}\nu_{k-m,k}\right), \qquad v_m = b_2 t_m\left(\frac{1}{b}\nu_{k-m,k-m+1}\right), \qquad \text{and} \tag{3.4}$$

$$u_m = b_2 t_m\left(\frac{1}{b}\nu_{k-m,k}\right).$$

The results in (a) and (b) now follow from (2.10), (3.2) to (3.4).

COROLLARY 3. *$t_m(z)$ is a nonnegative increasing function of z, and a decreasing function of m.*

The following representation of the Bayes rule follows from Theorem 2 and part (b) of Lemma 6.

THEOREM 3. *If the prior distribution of $\boldsymbol{\theta}$ is a mixture of* i.i.d. *random variables and the posterior distribution of $\boldsymbol{\theta}$ satisfies* (3.1), *then the Bayes rule d^* is given as follows*:

For $i = 2, \cdots, k$, let $\alpha_i = \max [x_{(k)} - b\delta_1, x_{(k-i+1)} - b\delta_i]$, where $\delta_i = -t_i^{-1}(c/b_2)$, and let $x_{(0)}$ denote $-\infty$, then

(i) $x_{(k-1)} \leqq x_{(k)} - b\delta_1 \Rightarrow d^* = d_1$,

(ii) $x_{(k-2)} \leqq \alpha_2 < x_{(k-1)} \Rightarrow d^* = d_2$, *and*

(iii) $x_{(k-i)} \leqq \alpha_i < x_{(k-i+1)}$ *and* $x_{(k)} - b\delta_j \leqq x_{(k-j)}$ *for some* $j \in \{2, \cdots, i-1\}$, *and let m be the smallest integer* $\in \{j+1, \cdots, i\}$ *satisfying* $\Delta_m \geqq 0 \Rightarrow d^* = d_m$; $i = 3, \cdots, k$.

COROLLARY 4. *For $k = 2$, the Bayes rule d^* is given by*:

$$\textit{If} \quad x_{(1)} \leqq x_{(2)} - b\delta_1\,, \quad \textit{then} \quad d^* = d_1\,, \quad \textit{otherwise} \quad d^* = d_2\,.$$

The rule R can be simplified to

Rule R. Select π_i if $x_i > x_{\max} - \delta b$, where $\delta = \delta_1$.

This rule is the classical subset selection procedure for the location parameter problem, Gupta (1965), where the value of δ is selected such that $P(CS) \geqq P^*$. Therefore δ depends on k and P^* both in the classical case, whereas δ_1 depends only on the *cost per population*. In effect, the user of the classical rule assumes that either the cost c depends on the number of populations or else the P^* depends on k. Both of these assumptions seem unreasonable to us. This representation of the "approximate" Bayes rule, part (a) of Lemma 6 and the results

in Chernoff and Yahav (1976) strengthen our belief that the classical selection rules are not optimal with respect to nonlinear loss function, even though they are close to the optimal rule provided P^* is chosen suitably.

An important consequence of Theorem 3 is the following result, which relates the maximum possible size of the subset to the cost c.

THEOREM 4. *If $t_j^{-1}(c/b_2) \geqq 0$, for some $j \in (1, \cdots, k)$ then the maximum subset size corresponding to the Bayes rule d^* is equal to j.*

4. The normal distribution. Let us now assume that, given $\theta_1, \cdots, \theta_k$, the random variables $X_1, \cdots, X_k$ are independently and normally distributed with mean θ_i and known common variance σ^2, and that the prior distribution of $\boldsymbol{\theta}$ is an exchangeable normal distribution. One way to specify this symmetric multivariate prior distribution is as follows.

Given $W = \omega$, $(\theta_1, \cdots, \theta_k)$ are i.i.d. with θ_i having a normal distribution with mean ω and variance β^2, and W has a known distribution function $Q(\omega)$.

Another way of choosing an exchangeable multivariate normal prior is to assume that, given $W^* = \omega^*$, $(\theta_1, \cdots, \theta_k)$ have a multivariate normal distribution with $E(\theta_i) = \omega^*$, $V(\theta_i) = \beta^2$, $\operatorname{cor}(\theta_i, \theta_j) = \rho$ where $\rho > -1/(k-1)$, and W^* has a known distribution. If $\rho \geqq 0$, then this prior can be reduced to the first type.

If the prior distribution of $\boldsymbol{\theta}$ is the mixture of i.i.d. normal random variables then given $W = \omega$, the random variables $\theta_1, \cdots, \theta_k$ are, a posteriori, independently distributed with $\theta_{(i)}$ having a normal distribution with mean $\alpha_i = (x_{(i)}/\sigma^2 + \omega/\beta^2)\gamma^2$ and the variance $\gamma^2 = 1/(1/\sigma^2 + 1/\beta^2)$ [see DeGroot (1970), Section 9.5]. Clearly the posterior cdf satisfies (3.1) with $b_2 = \gamma$ and $b = \sigma^2/\gamma$. If the prior knowledge is assumed to be vague, i.e., $\beta^2 \to \infty$, then $\gamma = \sigma$. It follows from (3.2) that

$$\Delta_m = c - \gamma \int_{-\infty}^{\infty} \prod_{i=k-m+1}^{k} \Phi\left(z + \frac{1}{b}\nu_{k-m,i}\right) \Phi(-z)\, dz \tag{4.1}$$

where Φ denotes the standard normal cdf. Furthermore, it follows from (3.3) that $t_m(x)$ can be written as

$$t_m(x) = \int_{-\infty}^{\infty} \Phi^m(z + x)\Phi(-z)\, dz\,. \tag{4.2}$$

We will determine the function $t_1(x)$ by using the following result.

LEMMA 7. *Let Z be a normally distributed random variable with mean μ and variance δ^2. Then*

$$E[Z^+] = \delta t(\mu/\delta)\,, \tag{4.3}$$

where the function t is defined by

$$t(x) = x\Phi(x) + \varphi(x)\,. \tag{4.4}$$

Here φ denotes the standard normal pdf.

Since, a posteriori, the $\theta_{(i)}$'s are independently and normally distributed, it

follows from Lemma 7 and the definition of $t_1(x)$, that

$$t_1(x) = 2^{\frac{1}{2}}t(x/2^{\frac{1}{2}}) . \tag{4.5}$$

The Bayes rule d^* is given by Theorem 3 with Δ_m and t_m, $m = 2, \cdots, k$ defined in (4.1) and (4.2) respectively and $t_1(x)$ defined in (4.5).

Clearly $t_m(x)$, $m = 1, 2, \cdots, k$ are increasing functions of x, and it can be verified that

(i) $t'(x) = \Phi(x)$,
(ii) $t''(x) = \varphi(x)$,
(iii) $t(0) = 1/(2\pi)^{\frac{1}{2}}$ and
(iv) $t_m(-\infty) = 0$, $t_m(\infty) = \infty$ for all m.

The next result follows from Theorem 4 and the above properties of $t(x)$.

COROLLARY 5. *If $c/\gamma \geqq 1/\pi^{\frac{1}{2}} = .56419$, then the Bayes rule d^* selects $s_1^* = \{\pi_{(k)}\}$, and if $t_m(0) < c/\gamma < .56419$ for some m, then the maximum size of the selected subset is m.*

The values of $t_m(0)$, $m = 1(1)30$ are given in Table 2.

It should be noted that, for $k = 2$, the Bayes rule with vague prior knowledge and the classical rule are the same except for the choice of δ. In the classical case, given a value of P^*, the value of δ satisfies

$$P^* = \int_{-\infty}^{\infty} \Phi(z + \delta)\varphi(z)\, dz = \Phi(\delta/2^{\frac{1}{2}}) .$$

Hence, for $k = 2$, the two rules are same if c and P^* satisfy

$$P^* = 1 - \Phi[t^{-1}(c/\sigma 2^{\frac{1}{2}})] . \tag{4.6}$$

If a user insists on using the classical procedure and the choice of P^* does not depend on k, then expression (4.6) will help determine the value of P^* for a given value of c/σ. The values of c/σ satisfying (4.6) for $P^* = .75, .90, .95$, and $.99$ are given in the following table. For other values of c/σ, the implementation of the classical rule will also require some computations.

c/σ	.210774	.066932	.028312	.005684
P^*	.75	.90	.95	.99

The expression (4.6) makes sense intuitively in that for large (small) values of c/σ, one should choose a small (large) P^*.

At this point, some comments about the computational aspect of δ_i's are in order. Since t_1 satisfies (4.5), the values of δ_1 were obtained by using the bisection method on the function $t(x)$. For $m \geqq 2$, the values of c/γ which are of interest to us satisfy $t_m^{-1}(c/\gamma) < 0$. It follows that

$$\int_{-\infty}^{a} \Phi^m(z + d)\Phi(-z)\, dz < \Phi^{m-1}(a) \int_{-\infty}^{a} \Phi(z)\, dz = \Phi^{m-1}(a)t(a) , \tag{4.7}$$

and

$$\int_{b^*}^{\infty} \Phi^m(z + d)\Phi(-z)\, dz < \int_{b^*}^{\infty} \Phi(-z)\, dz = t(-b^*) . \tag{4.8}$$

For $\lambda < 0$, it can be proved that

$$t_m(\lambda) - \int_{a(m)}^{b^*} \Phi^m(z + \lambda)\Phi(-z)\, dz \leqq 1.0 * 10^{-18}, \tag{4.9}$$

if $b^* = 9.0$ and $a(m)$ are given by:

m	2	3	4	5	6–7	8–11	12–17	18–23	24–30
$a(m)$	−6.0	−5.0	−4.0	−3.5	−3.0	−2.5	−2.0	−1.5	−1.0

These limits of integration are given to facilitate the computation of Δ_m if required by Theorem 3. The function $t_m(\lambda)$ for $m \geqq 2$ was evaluated by Gaussian quadrature method over intervals of length $D = 0.25$ starting from $a(m)$ to 9.0. The bisection method was used to find the solution to

$$t_m(\lambda) = c/\gamma \tag{4.10}$$

in such a way that the two final values of λ were within $1.12 * 10^{-7}$ *or* else

$$|t_m(\lambda) - c/\gamma| < 1.0 * 10^{-8}. \tag{4.11}$$

For one case, $t_m(\lambda)$ was within $1.5 * 10^{-8}$ of c/γ. The computation was done for $c/\gamma = .001, .005, .01(.01)t_m(0)$. The final values of δ_i are tabulated to 5 decimal places in Table 1. The next result now follows from Theorem 3, Corollary 5 and Table 2.

COROLLARY 6. *For* $.2821 \leqq c/\gamma < 1/\pi^{\frac{1}{2}}$, *the Bayes rule* d^* *is given by*:

$$\textit{if} \quad x_{(k-1)} \leqq x_{(k)} - \delta_1 b \quad \textit{then} \quad d^* = d_1, \quad \textit{otherwise} \quad d^* = d_2.$$

The values of δ_1 for $c/\gamma = .29(.01).56$ are also given in Table 1.

Finally if the choice of P^* is independent of k and the cost c is related to P^* according to (4.6), then it follows from Corollary 6 that

(i) $d^* = d_1$ if $P^* \leqq \frac{1}{2}$, and
(ii) $d^* = d_1$ or d_2 if $\frac{1}{2} < P^* < .8389$.

whereas the classical procedure may select larger subsets for these P^*-values.

Acknowledgment. We would like to thank three referees, an associate editor and the editor for valuable comments and suggestions which have greatly improved the presentation of the material.

TABLE 1

Table of $\delta_m = -t_m^{-1}(c/\gamma)$, *where,* $t_m(x) = \int_{-\infty}^{\infty} \Phi^m(z + x)\Phi(-z)\, dz$

m	c/γ							
	.001	.005	.01	.02	.03	.04	.05	.06
1	3.99091	3.26957	2.92398	2.55085	2.31718	2.14330	2.00325	1.88514
2	3.00589	2.35515	2.04207	1.70298	1.49002	1.33125	1.20315	1.09497
3	2.56281	1.93821	1.63728	1.31101	1.10590	.95285	.82930	.72491
4	2.29134	1.68076	1.38640	1.06706	.86621	.71628	.59522	.49290
5	2.10063	1.49896	1.20877	.89387	.69575	.54784	.42837	.32739
6	1.95583	1.36039	1.07314	.76137	.56518	.41868	.30035	.20031
7	1.84025	1.24946	.96441	.65497	.46023	.31480	.19732	.09799
8	1.74471	1.15755	.87421	.56662	.37301	.22842	.11161	.01284
9	1.66368	1.07944	.79750	.49139	.29870	.15479	.03852	
10	1.59359	1.01178	.73098	.42611	.23419	.09084		
11	1.53201	.95224	.67243	.36859	.17732	.03445		
12	1.47723	.89921	.62023	.31730	.12658			
13	1.42797	.85149	.57324	.27108	.08086			
14	1.38330	.80816	.53055	.22909	.03930			
15	1.34249	.76855	.49151	.19066	.00125			
16	1.30496	.73209	.45557	.15527				
17	1.27025	.69836	.42230	.12250				
18	1.23801	.66699	.39135	.09201				
19	1.20792	.63770	.36245	.06352				
20	1.17973	.61025	.33535	.03681				
21	1.15322	.58443	.30986	.01167				
22	1.12823	.56007	.28580					
23	1.10459	.53702	.26304					
24	1.08218	.51516	.24144					
25	1.06089	.49438	.22091					
26	1.04061	.47458	.20134					
27	1.02126	.45568	.18266					
28	1.00276	.43761	.16480					
29	.98505	.42030	.14768					
30	.96806	.40370	.13126					

m	c/γ							
	.07	.08	.09	.10	.11	.12	.13	.14
1	1.78248	1.69135	1.60915	1.53410	1.46491	1.40061	1.34046	1.28389
2	1.00083	.91716	.84162	.77257	.70886	.64961	.59413	.54191
3	.63403	.55321	.48022	.41349	.35188	.29457	.24090	.19036
4	.40380	.32455	.25296	.18750	.10726	.07081	.01814	
5	.23940	.16120	.09052	.02587				
6	.11317	.03566						
7	.01147							

TABLE 1 (cont.)

m	c/γ							
	.15	.16	.17	.18	.19	.20	.21	.22
1	1.23042	1.17968	1.13136	1.08520	1.04098	.99850	.95762	.91819
2	.49253	.44563	.40093	.35820	.31724	.27787	.23996	.20337
3	.14255	.09713	.05384	.01244				

m	c/γ					
	.23	.24	.25	.26	.27	.28
1	.88009	.84321	.80747	.77277	.73904	.70622
2	.16800	.13374	.10051	.06824	.03687	.00630

c/γ	δ_1	c/γ	δ_1	c/γ	δ_1	c/γ	δ_1
.29	.67425	.36	.47017	.43	.29241	.50	.13340
.30	.64307	.37	.44339	.44	.26868	.51	.11191
.31	.61263	.38	.41710	.45	.24531	.52	.09070
.32	.58289	.39	.39130	.46	.22229	.53	.06975
.33	.55381	.40	.36595	.47	.19960	.54	.04906
.34	.52535	.41	.34103	.48	.17723	.55	.02861
.35	.49748	.42	.31652	.49	.15516	.56	.00840

TABLE 2

Values of $t_m(0) = \int_{-\infty}^{\infty} \Phi^m(z)\Phi(-z)\,dz$

m	$t_m(0)$	m	$t_m(0)$	m	$t_m(0)$	m	$t_m(0)$	m	$t_m(0)$	m	$t_m(0)$
1	.5642	6	.0850	11	.0428	16	.0280	21	.0205	26	.0161
2	.2821	7	.0714	12	.0388	17	.0261	22	.0195	27	.0154
3	.1831	8	.0614	13	.0354	18	.0245	23	.0185	28	.0148
4	.1336	9	.0537	14	.0325	19	.0230	24	.0176	29	.0142
5	.1042	10	.0477	15	.0301	20	.0217	25	.0168	30	.0137

REFERENCES

[1] ALAM, K. (1973). On a multiple decision rule. *Ann. Statist.* **1** 750–755.

[2] BAHADUR, R. R. (1950). On a problem in the theory of k populations. *Ann. Math. Statist.* **21** 362–375.

[3] BAHADUR, R. R. and GOODMAN, L. A. (1952). Impartial decision rules and sufficient statistics. *Ann. Math. Statist.* **23** 553–562.

[4] BAHADUR, R. R. and ROBBINS, H. (1950). The problem of the greater mean. *Ann. Math. Statist.* **21** 469–487.

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

[6] BICKEL, P. and YAHAV, J. A. (1977). On selecting a set of good populations. *Statistical Decision Theory and Related Topics II* (S. S. Gupta and D. Moore, eds.). Academic Press, New York.

[7] CHERNOFF, H. and YAHAV, J. A. (1977). A subset selection problem employing a new criterion. *Statistical Decision Theory and Related Topics II* (S. S. Gupta and D. Moore, eds.). Academic Press, New York.

[8] DEELEY, J. J. and GUPTA, S. S. (1968). On the properties of subset selection procedures. *Sankhyā Ser. A* **30** 37–50.

[9] DEGROOT, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

[10] DESU, M. M. and SOBEL, M. (1968). A fixed subset-size approach to a selection problem. *Biometrika* **55** 401–410.

[11] DUNNETT, C. W. (1960). On selecting the largest of k normal population means. *J. Roy. Statist. Soc. Ser. B* **22** 1–40.

[12] EATON, M. L. (1967). Some optimum properties of ranking procedures. *Ann. Math. Statist.* **38** 124–137.

[13] GUPTA, S. S. (1956). On a decision rule for a problem in ranking means. Mimeo Series # 150, Institute of Statistics, Univ. of North Carolina, Chapel Hill.

[14] GUPTA, S. S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics* **7** 222–245.

[15] GUTTMAN, I. and TIAO, G. C. (1964). A Bayesian approach to some best population problems. *Ann. Math. Statist.* **35** 825–835.

[16] LEHMANN, E. L. (1966). On a theorem of Bahadur and Goodman. *Ann. Math. Statist.* **37** 1–6.

[17] STUDDEN, W. J. (1967). On selecting a subset of k populations containing the best. *Ann. Math. Statist.* **38** 1072–1078.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907