ON TWO-STAGE BAYES SELECTION PROCEDURES*

by

Shanti S. Gupta
Purdue University

Klaus J. Miescke
University of Illinois at Chicago

Technical Report #83-22

Department of Statistics
Purdue University

June 1983

*This research was supported by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

ON TWO-STAGE BAYES SELECTION PROCEDURES[1,2]

by

Shanti S. Gupta
Purdue University

Klaus J. Miescke
University of Illinois at Chicago

1. Introduction

Let $\pi_1,\ldots,\pi_k$ be k given normal populations with unknown means $\theta_1,\ldots,\theta_k \in \mathbb{R}$ and a common known variance $\sigma^2 > 0$. For finding the population with the largest mean, two-stage procedures with screening (elimination) at the first stage are studied in a decision-theoretic framework. The procedures are based on k samples of common size n_1 drawn at Stage 1, and on a random number of samples of common size n_2 drawn at Stage 2 from all those populations which have been selected (not eliminated) at Stage 1. If only one single population is selected at Stage 1, Stage 2 will not be entered. In particular, the stopping rule is thus determined by the size of the selected subset at Stage 1.

Let $\underline{X} = (X_1,\ldots,X_k)$ and $\underline{Y} = (Y_1,\ldots,Y_k)$ denote the vectors of sample means (which are sufficient statistics) at Stages 1 and 2, respectively, and let $\underline{Z} = (n_1\underline{X} + n_2\underline{Y})/(n_1+n_2)$ denote the vector of the k overall means. Although not all of the Y_i's and Z_i's are actually always observed, it will prove to be convenient to consider $\underline{Y}$ and $\underline{Z}$ in the derivations to come. Also for notational convenience, let $p = \sigma^2/n_1$ and $q = \sigma^2/n_2$.

Due to the complexity of the problem, optimality results on elimination type multi-stage procedures are rather scarce in the literature. For an overview and references, see Gupta and Panchapakesan (1979) and Miescke (1982). On the other hand, such procedures are highly desirable for

---

[1]This research was supported by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[2]This paper was delivered at a Statistical Research Conference dedicated to the memory of Jack Kiefer and Jacob Wolfowitz held at Cornell University, July 6-9, 1983 as a special meeting of the Institute of Mathematical Statistics cosponsored by the American Statistical Association.

AMS 1980 Subject Classification: Primary 62F07; Secondary 62C10, 62L99

Key words and phrases: Bayes procedure; two-stage; selecting largest mean; screening.

ON TWO-STAGE BAYES SELECTION PROCEDURES*

by

Shanti S. Gupta
Purdue University

Klaus J. Miescke
University of Illinois at Chicago

Abstract

Let $\pi_1,\ldots,\pi_k$ be normal populations with unknown means and a common known variance. The goal is to find the population with the largest mean. Two-stage procedures with screening at the first stage are studied in a Bayesian approach. They are based on k samples of common size n_1 drawn at Stage 1, and on samples of common size n_2 drawn at Stage 2 from all those populations which have not been screened out at Stage 1. If only one population is selected at Stage 1, the procedure stops at Stage 1.

Under the assumption of a specific loss function which includes costs of sampling, a Bayes procedure is derived with respect to i.i.d. normal priors. Its properties are discussed and several approximations are considered. The expected value of the maximum of k independent normals with known but distinct means and a common known variance plays a crucial rule in the determination of the Bayes procedure.

*This research was supported by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

ON TWO-STAGE BAYES SELECTION PROCEDURES

by

Shanti S. Gupta
Purdue University

Klaus J. Miescke
University of Illinois at Chicago

1. Introduction

Let $\pi_1,\ldots,\pi_k$ be k given normal populations with unknown means $\theta_1,\ldots,\theta_k \in \mathbb{R}$ and a common known variance $\sigma^2 > 0$. For finding the population with the largest mean, two-stage procedures with screening (elimination) at the first stage are studied in a decision-theoretic framework. The procedures are based on k samples of common size n_1 drawn at Stage 1, and on a random number of samples of common size n_2 drawn at Stage 2 from all those populations which have been selected (not eliminated) at Stage 1. If only one single population is selected at Stage 1, Stage 2 will not be entered. In particular, the stopping rule is thus determined by the size of the selected subset at Stage 1.

Let $\underline{X} = (X_1,\ldots,X_k)$ and $\underline{Y} = (Y_1,\ldots,Y_k)$ denote the vectors of sample means (which are sufficient statistics) at Stages 1 and 2, respectively, and let $\underline{Z} = (n_1\underline{X} + n_2\underline{Y})/(n_1+n_2)$ denote the vector of the k overall means. Although not all of the Y_i's and Z_i's are actually always observed, it will prove to be convenient to consider $\underline{Y}$ and $\underline{Z}$ in the derivations to come. Also for notational convenience, let $p = \sigma^2/n_1$ and $q = \sigma^2/n_2$.

Due to the complexity of the problem, optimality results on elimination type multi-stage procedures are rather scarce in the literature. For an overview and references, see Gupta and Panchapakesan (1979) and Miescke (1982). On the other hand, such procedures are highly desirable for

*This research was supported by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

experimenters because of their economical use of observations. An intuitively appealing procedure proposed and studied by Tamhane and Bechhofer (1979), which employs Gupta's maximum means subset selection procedure first and then the natural terminal decision, deserves to be revisited from the optimality point of view. Even though Gupta's rule has been shown by many authors to perform well as a single-stage procedure, its performance has not been studied in the multi-stage context with respect to optimality. This then was one of our motivations to prepare the present paper to find an answer to the interesting question: What type of subset selection rules are used in optimal two-stage selection procedures? In a first step towards an answer, we shall derive a Bayes solution for i.i.d. normal priors under a specific loss function which takes into account costs of sampling.

Assumption (P):

We restrict ourselves to procedures of the following type: At Stage 1, after $\underline{X}$ has been observed, a non-empty subset $s(\underline{X})$ of $\{1,\ldots,k\}$ of random size is selected where, obviously, i is associated with π_i, $i = 1,\ldots,k$. If its size $|s(\underline{X})|$, say, is equal to one, then the procedure stops and selects the corresponding population. Otherwise, for each $i \in s(\underline{X})$, Y_i is observed and then a final selection is made from $s(\underline{X})$ based on $\underline{X}$ and Y_i, $i \in s(\underline{X})$. Furthermore it is assumed that the procedures are permutation invariant.

The restriction of the final selections to populations π_i, with $i \in s(\underline{X})$ is actually crucial for the feasibility of a solution to the given problem. Under a fairly general loss structure which is permutation invariant and which favors, at all stages, selections of populations with

large means, Gupta and Miescke (1982, 1983) have derived two optimality results which can be stated in the present context as follows.

Fact 1: The natural final decision at Stage 2, which selects the population associated with the largest Z_i, $i \in s(\underline{X})$, is optimum in terms of the risk function, uniformly in $\underline{\theta} = (\theta_1,\ldots,\theta_k) \in \mathbb{R}^k$. Due to the restriction mentioned above, this remains true even if the complete vector $\underline{Y}$ had been observed. Thus, for convenience, we can assume in the sequel without loss of generality that all observations $\underline{Y}$ are taken at Stage 2, provided the procedure has not stopped already at Stage 1.

Actually, the above result holds for all exponential families whereas the next result has been proved only for strongly unimodal (log concave) exponential families. However, the underlying distributions of this paper are clearly of the latter type and, therefore, both results can be applied in the present setting.

Fact 2: The class $\mathfrak{D}$ of two-stage procedures which at Stage 1 make subset selections in terms of the largest X_i's and which employ the natural final decision at Stage 2 constitutes an essentially complete class.

The only characteristic with respect to which members of $\mathfrak{D}$ differ from each other is $|s(\underline{X})|$, i.e. the decision of how many populations to select at Stage 1 based on the observations $\underline{X}$. Apparently, optimality of a particular subset size decision is now closely related to the choice of a specific loss structure. But even then after such a choice has been made, no procedure can be expected to be optimum with respect to the risk, uniformly in $\underline{\theta} \in \mathbb{R}^k$. Therefore in a first approach we shall study the Bayes solution with respect to i.i.d. normal priors under the following loss structure.

Assumption (L): Let $g: \mathbb{R}^k \to \mathbb{R}$ be a fixed function. If the procedure stops at Stage 1 and selects $i \in \{1,\ldots,k\}$ then the loss function is given by

(1) $\quad L_1(\underline{\theta},i) = c_1 n_1 + g(\underline{\theta})-\theta_i,\ \underline{\theta} \in \mathbb{R}^k,\ i \in \{1,\ldots,k\}.$

If at Stage 1 the procedure selects $s \subseteq \{1,\ldots,k\}$ with $|s| \geq 2$ and then, at Stage 2, makes a final decision in favor of $i \in \{1,\ldots,k\}$ then the loss function is given by

(2) $\quad L_2(\underline{\theta},s,i) = c_1 n_1 + c_2 n_2 |s| + g(\underline{\theta})-\theta_i,\ \underline{\theta} \in \mathbb{R}^k.$

Under the loss assumptions made above, $c_1 n_1$ and $c_2 n_2 |s|$ represent the respective costs of sampling at the two stages whereas $\theta_i - g(\underline{\theta})$ can be considered as a measure of the quality of the finally selected population. A reasonable choice for g for example is $g(\underline{\theta}) = \max\{\theta_1,\ldots,\theta_k\},\ \underline{\theta} \in \mathbb{R}^k$. Obviously, the Bayes procedure cannot depend on c_1 since $\underline{X}$ has to be observed at Stage 1 anyway. Less obviously, it will turn out that it also does not depend on the choice of the function g. In Section 2, we shall derive the Bayes solution explicitly and finally, in Section 3, approximations to this solution will be considered.

2. The Bayes Procedure

From now on let us assume that the unknown population means are random variables $\Theta_1,\ldots,\Theta_k$, say, which are independently and identically distributed with a common known mean $\theta_0 \in \mathbb{R}$ and a common known variance $r > 0$. It can be anticipated that the Bayes rule does not depend on θ_0 since we are considering a location parameter model. As pointed out before, all components of the Bayes rule are already known except the decision of how many populations to be selected at Stage 1. Let d_s, $s \subseteq \{1,\ldots,k\}$, $|s| \geq 2$, denote the natural final decision at Stage 2,

i.e. $d_s(\underline{Z}) = i$ if $Z_i = \max_{j \in s} Z_j$ and $i \in s$. Here and in the following, the case of ties can be ignored since it occurs only with probability zero. Moreoever, let s_t denote the natural subset selection with fixed size $t \in \{1,\ldots,k\}$ at Stage 1, i.e. the selection of populations associated with the t largest X_i's. Working backwards from Stage 2 to Stage 1, the Bayes rule can now be determined by comparisons of posterior expected losses.

At Stage 2, given $\underline{X} = \underline{x}$ and $\underline{Y} = \underline{y}$ (or $\underline{Z} = \underline{z}$), respectively, for a natural subset selection rule s_t, $t \in \{2,\ldots,k\}$, the posterior expected loss is the following.

$$(3) \qquad E\{L_2(\underline{\theta}, s_t(\underline{x}), d_{s_t(\underline{x})}(\underline{z})) | \underline{Z} = \underline{z}, \underline{X} = \underline{x}\}$$
$$= c_1 n_1 + c_2 n_2 t + E\{g(\underline{\theta}) - \theta_{j_0} | \underline{Z} = \underline{z}\}$$

where j_0 is determined by $z_{j_0} = \max\{z_j | j \in s_t(\underline{x})\}$. Since at Stage 2, $\underline{Z}$ is a sufficient statistic for $\underline{\theta}$, the conditional distribution of $\underline{\theta}$, given $\underline{Z}$ and $\underline{X}$, depends only on $\underline{Z}$. This fact will also be utilized in (5) below.

Therefore, at Stage 1, given $\underline{X} = \underline{x}$, the posterior expected loss for a natural subset selection rule s_t, $t \in \{2,\ldots,k\}$ is given by

$$(4) \qquad E\{L_2(\underline{\theta}, s_t(\underline{x}), d_{s_t(\underline{x})}(\underline{Z})) | \underline{X} = \underline{x}\}$$
$$= c_1 n_1 + c_2 n_2 t + E\{g(\underline{\theta}) | \underline{X} = \underline{x}\} - E\{\theta_{j_0} | \underline{X} = \underline{x}\}$$

where j_0, now being a random index, is determined by $Z_{j_0} = \max\{Z_j | j \in s_t(\underline{x})\}$. At this point, the semigroup property of the normal distribution has to be utilized to evaluate $E\{\theta_{j_0} | \underline{X} = \underline{x}\}$.

$$(5)\qquad E\{\theta_{j_0} | \underline{X} = \underline{x}\} = E\{E\{\theta_{j_0} | \underline{Z}\} | \underline{X} = \underline{x}\}$$

$$= E\{\max_{j \in s_t(\underline{x})} \{(pq\theta_0 + qrx_j + prY_j)/(pq+qr+pr)\} | \underline{X} = \underline{x}\}$$

$$= p(p+r)^{-1}\theta_0 + E(\max_{j \in s_t(\underline{x})} \{ax_j + bN_j\})$$

where $a = r(p+r)^{-1}$, $b = pr((p+r)(pq+pr+qr))^{-\frac{1}{2}}$, and where $N_1,\ldots,N_k$ are auxiliary i.i.d. standard normally distributed random variables which will be used throughout the sequel.

On the other hand, the posterior expected loss for the natural subset selection rule s_1 is given by

$$(6)\qquad E\{L_1(\underline{\theta}, i_0) | \underline{X} = \underline{x}\}$$

$$= c_1 n_1 + E\{g(\underline{\theta}) | \underline{X} = \underline{x}\} - E\{\theta_{i_0} | \underline{X} = \underline{x}\},$$

where i_0 is determined by $x_{i_0} = \max\{x_i | i = 1,\ldots,k\}$. By a similar argument as before, it can be seen that

$$(7)\qquad E\{\theta_{i_0} | \underline{X} = \underline{x}\} = p(p+r)^{-1}\theta_0 + a \max_{i=1,\ldots,k} \{x_i\}.$$

At Stage 1, given $\underline{X} = \underline{x}$, the Bayes procedures decides in favor of the subset $s_i(\underline{x})$, $i = 1,\ldots,k$, if the associated posterior expected loss is the smallest of those given in (4) and (6), respectively. To simplify its representation, let us assume from now on that $x_1 < x_2 < \ldots < x_k$. This can be done without loss of generality since the problem under consideration is permutation invariant. The Bayes procedure can now be described as follows. For notational convenience, let

$$(8)\qquad \varepsilon_t(\underline{x}) = E(\max_{j \geq k-t+1} \{a(x_j - x_k) + bN_j\}),\ t = 1,\ldots,k,$$

and let $t^* \in \{2,\ldots,k\}$ be determined by

$$(9) \qquad \varepsilon_{t^*}(\underline{x}) - c_2 n_2 t^* = \max_{t=2,\ldots,k} \{\varepsilon_t(\underline{x}) - c_2 n_2 t\}.$$

Theorem 1. If $\varepsilon_{t^*}(\underline{x}) - c_2 n_2 t^* \leq 0$, then the Bayes procedure stops at Stage 1 and selects the population which is associated with x_k. Otherwise, the Bayes procedure selects the t^* populations which are associated with $x_{k-t^*+1}, \ldots, x_k$.

From an applied point of view, it can be seen readily that the Bayes procedure can be used without too much computational effort. The functions $\varepsilon_t(\underline{x})$, $t = 2,\ldots,k$, are simply expectations of extreme independent normals with given means and a common known variance. They can be determined either via simulations or, more precisely, numerically since they are one-dimensional integrals given below in (17). Several useful approximations will be derived in the next section. These are considered not only to simplify the application but also to gain further insight into the structure of the Bayes procedure. In the remainder of this section, we shall derive some basic results which will prove to be useful for these considerations.

First we point out that functions of the type $\varepsilon_t(\underline{x})$ have been considered already previously by Dunnett (1960), Chernoff and Yahav (1977) and Miescke (1979). Let

$$(10) \qquad T(\xi) = \int_{-\infty}^{\xi} \Phi(\eta) d\eta = \varphi(\xi) + \xi\Phi(\xi), \ \xi \in \mathbb{R},$$

where φ and Φ denote the standard normal density and cumulative distribution function, respectively. Then it can be shown that

$$(11) \qquad \varepsilon_2(\underline{x}) = 2^{\frac{1}{2}} b T(-2^{-\frac{1}{2}} b^{-1} a(x_k - x_{k-1})).$$

Therefore, if $k = 2$ or if $k \geq 3$ and the experimenter is not willing to select more than two populations at Stage 1, then the optimum procedure

is of the form given below. Let T^{-1} denote the inverse function to T. Then the procedure is as follows:

"Stop at Stage 1 and select π_k if

$$(12) \qquad x_{k-1} \leq x_k + 2^{\frac{1}{2}} b a^{-1} T^{-1}(2^{\frac{1}{2}} b^{-1} c_2 n_2).$$

Otherwise, select π_{k-1} and π_k, proceed to Stage 2, and make the final decision in terms of the larger of the two populations' overall sample means."

Thus in the case of k = 2, the screening rule of our Bayes procedure at Stage 1 turns out to be of the form of Gupta's single stage subset selection procedure, and therefore we can state that in this case the two-stage procedure proposed by Tamhane and Bechhofer (1979) turns out to be a proper Bayes rule with respect to the loss function (1) and (2) for appropriately chosen r and c_2.

Though in general Gupta's rule cannot be expected to be used by the Bayes procedure, it will be shown in the next section that this is true at least in the case of k = 3 if the means are equidistant, i.e. if $x_3 - x_2 = x_2 - x_1$. At first, however, some basic results concerning the Bayes procedure will be derived.

Theorem 2. At Stage 1, given $\underline{X} = \underline{x}$ with $x_1 < x_2 < \ldots < x_k$,

$$(13) \qquad \varepsilon_k(\underline{x}) - \varepsilon_{k-1}(\underline{x}) < \varepsilon_{k-1}(\underline{x}) - \varepsilon_{k-2}(\underline{x}) < \ldots < \varepsilon_2(\underline{x}) - \varepsilon_1(\underline{x}).$$

Proof: To simplify the notation, let from now on be $\mu_j = a(x_j - x_k)$, j = 1,...,k. Thus we have in particular $\mu_1 < \mu_2 < \ldots < \mu_k = 0$. As in the proof of Lemma 6 in Miescke (1979), it can be shown that for t = 2,...,k,

$$(14)\quad \varepsilon_t(\underline{x}) = \mu_{k-t+1} + bE(T(b^{-1}(\max_{j\geq k-t+2}\{\mu_j+bN_j\}-\mu_{k-t+1}))).$$

From (14), $\varepsilon_1(\underline{x}) = 0$, and then from the identity $T(\xi) = \xi+T(-\xi)$, $\xi \in \mathbb{R}$, it follows that

$$(15)\quad \varepsilon_t(\underline{x})-\varepsilon_{t-1}(\underline{x}) = bE(T(b^{-1}(\mu_{k-t+1}-\max_{j\geq k-t+2}\{\mu_j+bN_j\}))).$$

Since T is an increasing function, the assertion (13) can now be seen to be correct.

In view of the above result, the screening procedure at Stage 1 can be simplified as follows. First note that for $t = 2,\ldots,k$,

$$(16)\quad \varepsilon_t(\underline{x})-c_2n_2t = \varepsilon_2(\underline{x})-2c_2n_2 + \sum_{r=3}^{t}(\varepsilon_r(\underline{x})-\varepsilon_{r-1}(\underline{x})-c_2n_2).$$

Therefore, at the beginning one has to compute $\varepsilon_2(\underline{x}) - 2c_2n_2$. Then one has to evaluate and to add, successively, $\varepsilon_3(\underline{x})-\varepsilon_2(\underline{x})-c_2n_2$, $\varepsilon_4(\underline{x}) - \varepsilon_3(\underline{x})-c_2n_2,\ldots$ as long as these terms are positive. If finally the total sum on the right-hand side of (16) with $t = t^*$, turns out to be positive, then one selects $\pi_{k-t^*+1},\ldots,\pi_k$ and proceeds to Stage 2. Otherwise, one stops and selects π_k. It should be pointed out clearly that it may happen that $\varepsilon_2(\underline{x}) < 2c_2n_2$ but nevertheless the total sum mentioned before is positive.

3. Approximations

As noted in the preceding section, the functions $\varepsilon_t(\underline{x})$, $t = 2,\ldots,k$, play a crucial role in the determination of the Bayes procedure. Therefore, we shall study them now in more detail. We shall also develop several bounds which may be used for approximations of the procedure. Throughout the following, we assume that we are at Stage 1 where $\underline{X} = \underline{x}$ has been observed.

Because of the permutation invariance of the procedure we can assume without loss of generality that $x_1 < x_2 < \ldots < x_k$. As before, for convenience, let $\mu_j = a(x_j - x_k)$, $j = 1,\ldots,k$. We start with the following two well-known identities. For $t = 1,\ldots,k$,

$$(17) \qquad \varepsilon_t(\underline{x}) = E\Big(\max_{j \geq k-t+1} \{\mu_j + bN_j\}\Big)$$

$$= -\int_{-\infty}^{0} \prod_{j \geq k-t+1} \Phi(b^{-1}(\xi - \mu_j))d\xi + \int_{0}^{\infty}[1 - \prod_{j \geq k-t+1} \Phi(b^{-1}(\xi - \mu_j))]d\xi$$

and for $t = 2,\ldots,k$,

$$(18) \qquad \varepsilon_t(\underline{x}) - \varepsilon_{t-1}(\underline{x}) = \int_{\mathbb{R}} \prod_{j \geq k-t+2} \Phi(b^{-1}(\xi - \mu_j))[1 - \Phi(b^{-1}(\xi - \mu_{k-t+1}))]d\xi.$$

These results have been derived previously by Chernoff and Yahav (1977). It should be mentioned that (17) and (14) as well as (18) and (15) are related to each other through integration by parts.

Let us now consider the special case of $k = 3$ populations where the means x_1, x_2, x_3 are equidistant. Here a simple expression for $\varepsilon_3(\underline{x})$ can be given using the following result.

Lemma 1.

$$(19) \quad E(\max\{N_1, \alpha + N_2, 2\alpha + N_3\}) = 2^{\frac{1}{2}} T(2^{-\frac{1}{2}}\alpha) + 2^{-\frac{1}{2}} T(2^{\frac{1}{2}}\alpha)$$

$$= E(\max\{N_1, \alpha + N_2\}) + 2^{-1} E(\max\{N_1, 2\alpha + N_3\}), \quad \alpha \in \mathbb{R}.$$

Proof. By using (17), the left-hand side of (19) can be seen to be

$$(20) \qquad H(\alpha) = \alpha - \int_{-\infty}^{0} \Phi(x)\Phi(x-\alpha)\Phi(x+\alpha)dx + \int_{0}^{\infty}[1 - \Phi(x)\Phi(x-\alpha)\Phi(x+\alpha)]dx.$$

Differentiation with respect to α and some standard manipulations lead to

$$H'(\alpha) = \Phi(2^{-\frac{1}{2}}\alpha) + \Phi(2^{\frac{1}{2}}\alpha), \quad \alpha \in \mathbb{R}.$$

Therefore, the first equation in (19) follows now by integration with respect to α and by using (10). The second equation is a consequence of (11).

The expression for $\varepsilon_3(\underline{x})$ in the case of equidistant means now is

$$(21) \quad \varepsilon_3(\underline{x}) = 2^{\frac{1}{2}} bT(-2^{-\frac{1}{2}} b^{-1}\Delta) + 2^{-\frac{1}{2}} bT(-2^{\frac{1}{2}} b^{-1}\Delta)$$

where $\Delta = a(x_3-x_2) = a(x_2-x_1) > 0$. On the other hand, by (11), we have

$$(22) \quad \varepsilon_2(\underline{x}) = 2^{\frac{1}{2}} bT(-2^{-\frac{1}{2}} b^{-1}\Delta).$$

Here we have $\varepsilon_3(\underline{x})-\varepsilon_2(\underline{x}) \leq 2^{-1}\varepsilon_2(\underline{x})$, an inequality which does not hold true in general. Therefore, the difficulty described at the end of the last section cannot occur. If $\varepsilon_2(\underline{x}) \leq 2c_2n_2$, then also $\varepsilon_3(\underline{x})-\varepsilon_2(\underline{x}) \leq c_2n_2$ holds and thus the stopping rule depends only on $\varepsilon_2(\underline{x})$. The optimum subset selection rule now turns out to be the following.

"Select π_3. Furthermore, select π_i, $i = 1,2$, if and only if

$$(23) \quad x_i > x_3 + 2^{\frac{1}{2}} a^{-1} bT^{-1}(2^{\frac{1}{2}} b^{-1} c_2 n_2)."$$

This rule is of the Gupta type. Note that, in view of $T(0) = (2\pi)^{-\frac{1}{2}}$, if $c_2n_2 \geq 2^{-1}\pi^{-\frac{1}{2}} b$ then (23) cannot occur and Stage 2 will not be performed in this case. We shall see below that this actually holds in the general case of $k \geq 2$ and $\mu_1 < \mu_2 < \ldots < \mu_k = 0$.

Returning to the general case, a similar no-data check, before entering Stage 1, can be done as follows. Let $a_t = E(\max\{N_1,N_2,\ldots,N_t\})$, $t = 1,\ldots,k$. Various properties and tables of the a_t's can be found in David (1981). From (17) and (18) it follows that for $t = 2,\ldots,k$,

$$(24) \quad 0 < \varepsilon_t(\underline{x}) < ba_t, \text{ and}$$

$$0 < \varepsilon_t(\underline{x}) - \varepsilon_{t-1}(\underline{x}) < b(a_t-a_{t-1}).$$

Therefore, if for a certain $t \in \{3,\ldots,k\}$, $a_t - a_{t-1} \leq b^{-1}c_2n_2$ then the Bayes procedure selects at most t-1 populations. And since $a_3 - a_2 = 2^{-1}a_2 = 2^{-1}\pi^{-\frac{1}{2}}$, Stage 2 will never be performed if $c_2n_2 \geq 2^{-1}\pi^{-\frac{1}{2}}b$.

Next, several properties of $\varepsilon_t(\underline{x})$, $t = 2,\ldots,k$, will be described below from which bounds and approximations of the Bayes rule can be derived later on.

Lemma 2. For every $t \geq 2$, $\varepsilon_t(\underline{x})$ is a strictly increasing function of b and μ_j, $j = k-t+1,\ldots,k-1$.

Proof: The partial derivative of $\varepsilon_t(\underline{x})$ with respect to b in view of (17) is

$$(25)\quad b^{-2}\sum_{i=k-t+1}^{k}\int_{\mathbb{R}}(\xi-\mu_i)\prod_{\substack{j=k-t+1\\ j\neq i}}^{k}\Phi(b^{-1}(\xi-\mu_j))\varphi(b^{-1}(\xi-\mu_i))d\xi$$

$$= b^{-1}(\varepsilon_t(\underline{x}) - \sum_{i=k-t+1}^{k}\mu_i P\{\mu_i+bN_i = \max_{j\geq k-t+1}\{\mu_j+bN_j\}\}).$$

Since by (24) $\varepsilon_t(\underline{x}) > 0 = \max\{\mu_j \mid j \geq k-t+1\}$, the first assertion follows. The second one is obvious.

Lemma 3. For every $t \geq 2$, $\varepsilon_t(\underline{x})$ considered as a function of $(\mu_{k-t+1},\ldots,\mu_k)$ has the following Taylor expansion of first order at $\mu_{k-t+1} = \ldots = \mu_k = 0$.

$$(26)\quad \varepsilon_t(\underline{x}) = ba_t + t^{-1}\sum_{i=k-t+1}^{k}\mu_i + o(|\mu_{k-t+1}|).$$

Proof: For $i \in \{k-t+1,\ldots,k-1\}$, the partial derivative of $\varepsilon_t(\underline{x})$ with respect to μ_i, in view of (17), can be seen to be equal to

$$(27)\quad \frac{\partial}{\partial\mu_i}\varepsilon_t(\underline{x}) = P\{\mu_i+bN_i = \max_{j\geq k-t+1}\{\mu_j+bN_j\}\}$$

which at $\mu_{k-t+1} = \ldots = \mu_k = 0$ is equal to t^{-1}. It is thus seen that (26) holds.

Remark. The second order term of the Taylor expansion can be shown to be $(2b(t-1))^{-1}ta_tv_t$, where v_t denotes the ordinary sample variance of $\mu_{k-t+1},\ldots,\mu_k$. However, since we will not make any use of it in the sequel, its derivation is omitted.

Lemma 4. For every $t \geq 2$, $\varepsilon_t(\underline{x})$ is a Schur-convex function of $(\mu_{k-t+1},\ldots,\mu_k)$, and thus in particular,

$$(28) \qquad \varepsilon_t(\underline{x}) \geq t^{-1} \sum_{i=k-t+1}^{k} \mu_i + ba_t.$$

Proof: In (17), $bN_{k-t+1},\ldots,bN_k$ are exchangeable multivariate normal random variables with expectations 0, variances b^2 and covariances 0. The function

$$h(u_1,\ldots,u_t) = \max_{j=1,\ldots,t} u_j, \quad \underline{u} \in \mathbb{R}^t,$$

is obviously Schur-convex. Therefore, by Marshall and Olkin (1979), Ch. 11 E.9.,

$$\varepsilon_t(\underline{x}) = E(h(\mu_{k-t+1}+bN_{k-t+1},\ldots,\mu_k+bN_k))$$

is a Schur-convex function of $(\mu_{k-t+1},\ldots,\mu_k)$. Then the inequality (28) follows immediately from the fact that $(\mu_{k-t+1},\ldots,\mu_k)$ majorizes $t^{-1}(\mu_{k-t+1}+\ldots+\mu_k)(1,1,\ldots,1)$.

The results of Lemma (3) and (4) have interesting consequences. Suppose we replace in our Bayes procedure, i.e. in (8) and (9), for $t \geq 2$, $\varepsilon_t(\underline{x})$ by, say,

$$(29) \qquad \tilde{\varepsilon}_t(\underline{x}) = t^{-1} \sum_{j=k-t+1}^{k} \mu_j + ba_t.$$

In view of (26), this can be justified as a reasonable approximation as long as $|\mu_{k-t+1}| = a(x_k - x_{k-t+1})$ is small. In all other cases, because of (28), it can be considered as an approximation to the Bayes procedure which is conservative with respect to costs of sampling. Let us take a brief look at this approximate procedure. Since the functions $\tilde{\varepsilon}_t(\underline{x})$ do not have the property of the functions $\varepsilon_t(\underline{x})$ given in (13), (16) and the process described after (16) is not applicable. Therefore, let us consider the original form of the Bayes procedures as described in Theorem 1, with $\varepsilon_t(\underline{x})$ replaced by $\tilde{\varepsilon}_t(\underline{x})$, $t = 2,\ldots,k$. It is of the following form.

"Select $\pi_{k-\tilde{t}+1},\ldots,\pi_k$ and proceed to Stage 2, if

$$(30) \quad \tilde{t}^{-1} \sum_{j=k-\tilde{t}+1}^{k} x_j > x_k - a^{-1}ba_{\tilde{t}} + a^{-1}c_2n_2\tilde{t},$$

where for $\tilde{t}$, $at^{-1} \sum_{j=k-t+1}^{k} x_j + ba_t - c_2n_2t$, $t = 2,\ldots,k$, assumes its largest value. Otherwise, stop and select π_k."

Here, the interesting feature is that the averages of populations associated with large means are compared with the maximum mean at Stage 1. Considered as a one-stage subset selection procedure, this rule can be seen also to be an approximate Bayes solution for the one-stage subset selection problem under the same distributional assumptions as before but now with a loss function of the type

$$(31) \quad L(\underline{\theta},s) = \max_{j=1,\ldots,k} \theta_i - |s|^{-1} \sum_{j\in s} \theta_j + \gamma_{|s|},$$

where $\gamma_1,\ldots,\gamma_k$ are appropriately chosen constants which may represent, for example, the costs of using the selected populations in the future. In Chernoff and Yahav (1977), a Bayes single stage subset selection procedure has been studied where one of the two components of the loss function is equal to $L(\underline{\theta},s) - \gamma_{|s|}$. It should be pointed out that a

subset selection rule of the form given in (30) has not been considered in the literature till now. It would be interesting to study its performance, in a non-Bayesian approach, for suitably chosen constants α_t replacing $a^{-1}(c_2 n_2 t - b a_t)$ in (30), $t = 1,\ldots,k$.

The approximation of our two-stage Bayes procedure at Stage 1 considered above can of course also be performed partially, i.e. for specific t-values. In a similar way other lower bounds, and upper bounds as well, can be used to approximate the optimum procedure. For this purpose, several bounds will be given below. Some of them have been derived already in Miescke (1979), but for the sake of making the present paper self-contained they will be included in the list presented below, partly accompanied by shorter proofs.

Lemma 5. The following functions listed below are lower bounds of $\varepsilon_t(\underline{x})$, $t = 2,\ldots,k$.

$$2^{\frac{1}{2}} bT(-2^{-\frac{1}{2}} b^{-1} a(x_k - x_{k-1})) \tag{32}$$

$$\varepsilon_{t-1}(\underline{x}) + bT(-b^{-1}(\varepsilon_{t-1}(\underline{x}) + a(x_k - x_{k-t+1}))) \tag{33}$$

$$bT(a_t - b^{-1} at(t-1)^{-1}(x_k - t^{-1} \sum_{j=k-t+1}^{k} x_j)) \tag{34}$$

$$ba_t + a(t^{-1} \sum_{j=k-t+1}^{k} x_j - x_k). \tag{35}$$

Proof: From $\varepsilon_t(\underline{x}) \geq \varepsilon_2(\underline{x})$, (32) follows immediately since by (11), for $t = 2$, (32) is equal to $\varepsilon_2(\underline{x})$.

Since $T(\xi)$, $\xi \in \mathbb{R}$, is a strictly convex function, Jensen's inequality can be applied to (14). This, together with the identity $T(\xi) = \xi + T(-\xi)$, $\xi \in \mathbb{R}$, leads to (33).

To prove (34), consider the following identity which is of the same type as (14).

$$(36)\quad \varepsilon_t(\underline{x}) = bE(T(b^{-1}\max\{\mu_j + bN_j \mid k-t+1 \le j \le k-1\})).$$

Since T is an increasing function, $T(\max\{u_1,\ldots,u_{t-1}\})$, $\underline{u} \in \mathbb{R}^{t-1}$, is Schur-convex. By the same argument as used in the proof of Lemma 4, a lower bound on $\varepsilon_t(\underline{x})$ is thus given by

$$(37)\quad bE(T(\max\{N_j \mid k-t+1 \le j \le k-1\} + b^{-1}(t-1)^{-1}\sum_{j=k-t+1}^{k-1}\mu_j)).$$

Therefore, by applying Jensen's inequality, it can be seen that (34) is a lower bound of $\varepsilon_t(\underline{x})$. Note that (35) has been derived in Lemma (4). In the theorem above, it is only mentioned for the sake of completeness.

It should be pointed out that (33) can be iterated and then (32), (34) or (35) can be applied to the result to get further lower bounds for $\varepsilon_t(\underline{x})$, which then, of course are weaker than the previous ones.

Useful upper bounds for $\varepsilon_t(\underline{x})$ are harder to find. Besides the obvious upper bound ba_t, the following can be established which can be considered as a counterpart to (32).

Lemma 6. For $t = 2,\ldots,k$,

$$(38)\quad \varepsilon_t(\underline{x}) \le 2^{\frac{1}{2}}b\sum_{j=k-t+1}^{k-1} T(2^{-\frac{1}{2}}b^{-1}a(x_j-x_k)).$$

Proof: Consider the inequality

$$\max_{j\ge k-t+1} u_j \le u_k + \sum_{j=k-t+1}^{k-1}\max\{0,u_j-u_k\}.$$

Applying it to (8), and using (11), the result follows immediately.

References

[1] Chernoff, H. and Yahav, J. A. (1977). A subset selection problem employing a new criterion. Statistical Decision Theory and Related Topics II. (Gupta, S. S. and Moore, D. S. eds.). Academic Press, 93-119.

[2] David, H. A. (1981). Order Statistics. Second edition. J. Wiley.

[3] Dunnett, C. W. (1960). On selecting the largest of k normal population means. J. Roy. Statist. Soc., Ser. B22, 1-40.

[4] Gupta, S. S. and Miescke, K. J. (1983). An essentially complete class of two-stage procedures with screening at the first stage. Statistics and Decisions 1 issue 5, to appear.

[5] Gupta, S. S. and Miescke, K. J. (1982). Sequential selection procedures: A decision theoretic approach. Mimeo. Ser. 82-6, Dept. of Statist., Purdue University, W. Lafayette.

[6] Gupta, S. S. and Panchapakesan, S. (1979). Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations. J. Wiley.

[7] Marshall, A. W. and Olkin, I. (1979). Inequalities: Theory of Majorization and Its Applications. Academic Press.

[8] Miescke, K. J. (1979). Bayesian subset selection procedures for additive and linear loss functions. Commun. Statist. A8, 1205-1226.

[9] Miescke, K. J. (1982). Recent results on multi-stage selection procedures. Proceedings of the 7th Conference on Probability Theory in Brasov, Romania, to appear.

[10] Tamhane, A. and Bechhofer, R. E. (1979). A two-stage minimax procedure with screening for selecting the largest normal mean (II): An improved PCS lower bound and associated tables. Commun. Statist. A8, 337-358.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report #83-22
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): On Two-Stage Bayes Selection Procedures
5. TYPE OF REPORT & PERIOD COVERED: Technical
6. PERFORMING ORG. REPORT NUMBER: Technical Report #83-22
7. AUTHOR(s): Shanti S. Gupta and Klaus J. Miescke
8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0455
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Purdue University, Department of Statistics, W. Lafayette, IN 47907
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Washington, DC
12. REPORT DATE: June 1983
13. NUMBER OF PAGES: 17
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release, distribution unlimited.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES:
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Bayes procedure; two-stage procedures; normal populations; unknown means; largest mean; first screening; approximate Bayes procedure.
20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

Let $\pi_1, \ldots, \pi_k$ be normal populations with unknown means and a common known variance. The goal is to find the population with the largest mean. Two-stage procedures with screening at the first stage are studied in a Bayesian approach. They are based on k samples of common size n_1 drawn at Stage 1, and on samples of common size n_2 drawn at Stage 2 from all those populations which have not been screened out at Stage 1. If only one population is selected at Stage 1, the procedures stops at Stage 1. (over)

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Under the assumption of a specific loss function which includes costs of sampling, a Bayes procedure is derived with respect to i.i.d. normal priors. Its properties are discussed and several approximations are considered. The expected value of the maximum of k independent normals with known but distinct means and a common known variance plays a crucial rule in the determination of the Bayes procedure.