On the Problem of Finding the Largest Normal Mean under Heteroscedasticity

Shanti S. Gupta[1]
Purdue University

Klaus J. Miescke[2]
University of Illinois at Chicago

Technical Report #86-27

Department of Statistics
Purdue University

July 1986

AMS(1980) Subject Classification: Primary 62F07; Secondary 62F15

KEY WORDS: Selecting the largest mean; Heteroscedasticity in normal populations; Bayes selection rules.

[1] Research supported by the Office of Naval Research Contract N00014-84-C-0167 and NSF Grant DMS-8606964 at Purdue University.

[2] Research supported by the Air Force Office of Scientific Research Contract AFOSR-85-0347 at the University of Illinois at Chicago.

On the Problem of Finding the Largest Normal Mean under Heteroscedasticity

Shanti S. Gupta[1]
Purdue University

Klaus J. Miescke[2]
University of Illinois at Chicago

Abstract

Let $P_1, \ldots, P_k$ be $k \geq 3$ given normal populations with unknown means $\theta_1, \ldots, \theta_k$, and a common known variance σ^2. Let $\overline{X}_1, \ldots, \overline{X}_k$ be the sample means of k independent samples of sizes $n_1, \ldots, n_k$ from these populations. To find the population with the largest mean, one usually applies the natural rule d^N, which selects in terms of the largest sample mean.

In this paper, the performance of this rule is studied under $0-1$ loss. It is shown that d^N is minimax if and only if $n_1 = \ldots = n_k$. d^N is seen to perform weakly whenever the parameters $\theta_1, \ldots, \theta_k$ are close together. Several alternative selection rules are derived in a Bayesian approach which seem to be reasonable competitors to d^N, worth comparing with d^N in a future simulation study.

[1] Research supported by the Office of Naval Research Contract N00014-84-C-0167 and NSF Grant DMS-8606964 at Purdue University.

[2] Research supported by the Air Force Office of Scientific Research Contract AFOSR-85-0347 at the University of Illinois at Chicago.

ON THE PROBLEM OF FINDING THE LARGEST NORMAL MEAN UNDER HETEROSCEDASTICITY

Shanti S. Gupta
Department of Statistics
Purdue University
West Lafayette, IN 47907

Klaus J. Miescke
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
P. O. Box 4348
Chicago, IL 60680

1. INTRODUCTION

Let $P_1, \ldots, P_k$ be $k \geq 3$ given normal populations with unknown means $\theta_1, \ldots, \theta_k \in \mathbb{R}$, and a common known variance $\sigma^2 > 0$. Suppose we want to find the population with the largest mean, where k independent samples of sizes $n_1, \ldots, n_k$ from $P_1, \ldots, P_k$ are available with sample means $\overline{X}_1, \ldots, \overline{X}_k$, respectively.

The natural decision rule d^N, which selects that population which is associated with the largest sample mean, has been studied by many authors since it was introduced in the pioneering paper of Bechhofer (1954). It was found that it is the uniformly best permutation invariant procedure if the sample sizes $n_1, \ldots, n_k$ are all equal. The most general version of this so-called "Bahadur-Eaton-Goodman-Lehmann Theorem" is presented in Gupta and Miescke (1984), where the risk function of multi- stage selection rules with screening is studied under a permutation invariant loss structure.

The situation changes drastically when the assumption of equal sample sizes is dropped. Besides being asymptotically consistent when the sample sizes tend to infinity, no optimum property of the natural rule d^N is known so far. On the contrary, Lam and Chiu (1976), and more generally Tong and Wetzell (1979), have brought to light quite pathological behavior of the probability of a correct selection under d^N, $P(CS|d^N)$, say. If $\theta_1, \ldots, \theta_k$ are sufficiently close together and if $\theta_1, \ldots, \theta_{k-1} < \theta_k$, then its value is strictly decreasing in n_k.

It should be noted that technically there will be no great changes if we assume that $P_1, \ldots, P_k$ have different but known variances. However, we feel that the chosen model provides a better motivation for our considerations. Nevertheless, our analysis will be based on k independent random variables $X_i \sim N(\theta_i, p_i), i = 1, \ldots, k$, where $p_1, \ldots, p_k$ are known, and can be thus applied to the more general case, too.

Whenever comparisons with a control are incorporated into the problem, difficulties caused by heteroscedasticity can be overcome more easily. This has been done for example by Miescke (1981) and Gupta and Miescke (1985). However, the transition to the corresponding problem without a control, as it is described in Miescke (1979), cannot be made in the given situation.

Although some work has been done already to solve the given selection problem, no modification or substitute of d^N has been found so far which can be considered to be better in some reasonable sense. Some insight into the structure of the problem has been gained by Bechhofer and Tamhane (1986), who looked for the best allocations of observations, subject to $n_1 + \ldots + n_k$ being fixed, to maximize $P(CS|d^N)$ for the case of known but unequal variances.

The problem under concern, although being rarely mentioned in the literature, e. g. Berger (1983) and Miescke (1984), is well known to the statistical community. A recent simulation study by Zaher and Heiny (1984), where d^N is compared with two similar rules which are based on medians and rank-sums, respectively, under $n_1 = \ldots = n_k$ but different variances of $P_1, \ldots, P_k$, corroborates this fact. It should be pointed out that the problem of selecting a subset for unequal sample sizes (or unequal variances) has been studied by Gupta and Huang (1976).

In the next section, the minimax approach is used to detect weak points in the performance of d^N. However, no alternative decision rule can be found in this approach. Therefore, Bayes rules with respect to various priors are studied in the subsequent sections to find reasonable modifications of or alternatives to d^N. Similar techniques have been used previously by Ehrman, Krieger and Miescke (1986) in the related subset selection problem. Several promising candidates to be used as alternatives to d^N will be derived and proposed in this paper. Comparisons of the performance characeristics of all rules considered in a simulation study is planned to be made in the future.

2. MINIMAXITY

The problem, which will be considered throughout this paper, can be formulated in a concise form as follows. Given are independent random variables $X_i \sim N(\theta_i, p_i), i = 1, \ldots, k$, where $p_1, \ldots, p_k$ are fixed known positive numbers. To be found is the index i_0, say, with $\theta_{i_0} = max\{\theta_1, \ldots, \theta_k\}$, which we may assume to be unique for the sake of simplicity. Under the $0-1$ loss function, the probability of a correct selection and the risk function of a (possibly randomized) decision rule d at a parameter configuration $\underline{\theta} = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ are connected through

$$P_{\underline{\theta}}(CS|d) = 1 - R(\underline{\theta}, d). \tag{1}$$

Thus all decision theoretic formulations in terms of risk can be translated immediately into the "P(CS)-language" used in the area of ranking and selection.

We begin our study with minimax considerations since this will lead us directly to the weak points in the performance of the natural decision rule d^N, which selects in terms of the largest value among $X_1, \ldots, X_k$. We shall see that the performance of d^N becomes unsatisfactory whenever the parameters $\theta_1, \ldots, \theta_k$ are lying closely together. This complements the findings of Lam and Chiu (1976) and of Tong and Wetzell (1979) and indicates that d^N cannot be considered to be a universally acceptable decision rule.

Let φ and Φ denote the density and cumulative distribution function, respectively, of

$N(0,1)$ in the sequel. The first result is a reformulation of the findings by Tong and Wetzell (1979), presented however in a form which is more suitable for our further considerations, and proved differently.

LEMMA 1. The function

$$H(\gamma_1,\ldots,\gamma_{k-1}) = \int_{\mathbb{R}} \prod_{i=1}^{k-1} \Phi(\gamma_i z)\varphi(z)dz \tag{2}$$

it is strictly increasing in $\gamma_i \geq 0, i = 1,\ldots,k-1$.

Proof: The partial derivative of H with respect to γ_1 is equal to

$$\int_{\mathbb{R}} \prod_{i=2}^{k-1} \Phi(\gamma_i z) z\varphi(\gamma_1 z)\varphi(z)dz. \tag{3}$$

After combining the two φ-functions, and then integrating by parts, it can be seen that (3) equals

$$(2\pi)^{-\frac{1}{2}}(1+\gamma_1^2)^{-1} \int_{\mathbb{R}} M(w)\varphi(w)dw, \tag{4}$$

where

$$M(w) = \frac{\partial}{\partial w} \prod_{i=2}^{k-1} \Phi((1+\gamma_1^2)^{-\frac{1}{2}}\gamma_i w), w \in \mathbb{R}, \tag{5}$$

is clearly positive over the whole real line.

As an immediate consequence, we can state the following.

COROLLARY 1. The function

$$G(\sigma_1,\ldots,\sigma_k) = \int_{\mathbb{R}} \prod_{i=1}^{k-1} \Phi(\sigma_i^{-1} z)\sigma_k^{-1}\varphi(\sigma_k^{-1} z)dz \tag{6}$$

is strictly decreasing in $\sigma_1,\ldots,\sigma_{k-1}$, and strictly increasing in σ_k.

Now we can state the main result of this section. The points of weakness of d^N, which we have mentioned before, will become visible in the course of the proof.

THEOREM 1. *For the given problem, the natural decision rule d^N is minimax if and only if $p_1 = p_2 = \ldots = p_k$. Moreover, the minimax- value of the problem* is $1 - 1/k$.

Proof: Consider the no-data rule d^0, which selects every $i \in \{1,\ldots,k\}$ with the same probability $1/k$. It has clearly the constant risk $1-1/k$.

The risk function of d^N can be represented in a convenient way by using the following notation. For any vector $\underline{a} \in \mathbb{R}^k$, let $a_{[1]} \leq \ldots \leq a_{[k]}$ denote the ordered coordinates. Moreover, whenever $\underline{\theta} = (\theta_1,\ldots,\theta_k)$ and $\underline{X} = (X_1,\ldots,X_k)$ are considered jointly in the sequel, let subscript $(j) = i$, if $\theta_i = \theta_{[j]}, i,j = 1,\ldots,k$. As mentioned before, we may assume that no ties occur among the θ_i's. This simplifies our considerations without losing generality. Introducing generic random variables $N_1,\ldots,N_k$, which are independent standard normals, we can represent the risk of d^N at $\underline{\theta} \in \mathbb{R}^k$ by

$$\begin{aligned} R(\underline{\theta}, d^N) &= 1 - P_{\underline{\theta}}\{X_{(k)} = X_{[k]}\} \\ &= 1 - P\{\theta_{[i]} + p_{(i)}^{\frac{1}{2}} N_i \leq \theta_{[k]} + p_{(k)}^{\frac{1}{2}} N_k, i < k\}. \end{aligned} \tag{7}$$

And since this is an increasing function of $\theta_{[i]}, i < k$, we conclude that

$$\begin{aligned} &\sup_{\underline{\theta}} R(\underline{\theta}, d^N) \\ &\quad = 1 - \inf_{\underline{\theta}} \int_{\mathbb{R}} \prod_{i=1}^{k-1} \Phi((p_{(k)}/p_{(i)})^{\frac{1}{2}} z)\varphi(z)dz \\ &\quad = 1 - \int_{\mathbb{R}} \prod_{i=2}^{k} \Phi((p_{[1]}/p_{[i]})^{\frac{1}{2}} z)\varphi(z)dz, \end{aligned} \tag{8}$$

where the second equation is a consequence of Lemma 1. Moreover, from Lemma 1 we see that

$$\sup_{\underline{\theta}} R(\underline{\theta}, d^N) \geq 1 - 1/k, \tag{9}$$

with equality holding if and only if $p_1 = p_2 = \ldots = p_k$.

Thus, to complete the proof, we have to show that the minimax value of the given problem is equal to $1-1/k$. Since the no-data rule d^0 has constant risk $1-1/k$, it suffices to find a sequence of priors such that the sequence of associated Bayes risks tends to this value. The following class of conjugate priors will be seen to contain such a sequence.

Let, apriori, $\Theta_1,\ldots,\Theta_k$ be independent random variables with $\Theta_i \sim N(\mu_i, r_i), \mu_i \in \mathbb{R}, r_i > 0, i = 1,\ldots,k$. Then, as it is well known, aposteriori, given $\underline{X} = \underline{x}, \Theta_1,\ldots,\Theta_k$ are independent normals with expectations $(p_i\mu_i + r_i x_i)/(p_i + r_i)$ and variances $r_i p_i/(p_i + r_i), i = 1,\ldots,k$, respectively. And marginally, $X_1,\ldots X_k$ are independent normals with expectations μ_i and variances $p_i + r_i, i = 1,\ldots,k$, respectively.

At $\underline{X} = \underline{x}$, the Bayes rule d^B, say, minimizes the posterior expected loss, and it yields the posterior risk

$$\min_{i=1,\ldots,k}(1 - P\{\Theta_i = \Theta_{[k]}|\underline{X} = \underline{x}\})$$

$$= 1 - \max_{i=1,\ldots,k}\int_{\mathbf{R}}\prod_{j\neq i}\Phi((\alpha_j p_j)^{-\frac{1}{2}}(\alpha_i x_i + (1-\alpha_i)\mu_i$$

$$- \alpha_j x_j - (1-\alpha_j)\mu_j + (\alpha_i p_i)^{\frac{1}{2}} z))\varphi(z)dz, \tag{10}$$

where $\alpha_s = r_s/(p_s + r_s), s = 1,\ldots,k$.

For the special case of $r_s = 1/n$ and $\mu_s = 0, s = 1,\ldots,k$, we see that (10) tends to $1 - 1/k$, if n tends to infinity. And since the marginal densities of $X_1,\ldots,X_k$ are bounded by a constant, a routine application of Lebesgue's dominated convergence theorem shows that the sequence of Bayes risks tend in fact to $1-1/k$, if n tends to infinity. This completes the proof of the theorem.

From (8) in the last proof, we can see now clearly what might go wrong in the performance of the natural rule d^N. If the parameters $\theta_1,\ldots,\theta_k$ are close together, and if the variance $p_{(k)}$ of $X_{(k)}$, which is associated with $\theta_{[k]}$, is relatively small in comparison with $p_{(i)}, i \neq k$, then the rule d^N performs "worse than at random".

One natural way out of this dilemma, and to possibly save the reputation of d^N, is to look at the average risk over all $k!$ permutations of a given parameter vector $\underline{\theta}$, rather than taking the risk function as a measure of performance. The average risk of a rule d at $\underline{\theta} \in \mathbb{R}^k$ would be

$$\tilde{R}(\underline{\theta}, d) = (1/k!)\sum_{\pi} R(\pi(\underline{\theta}), d), \tag{11}$$

where $\pi(\underline{\theta}) = (\theta_{\pi(1)},\ldots,\theta_{\pi(k)})$, and the summation being taken over $k!$ permutations π of $(1,2,\ldots,k)$. The average risk reflects perhaps better the prevailing attitude of researchers in the area of ranking and selection, which states that "the pairing between the $\theta_i's$ and the P_i's is completely unknown."

It can be shown that with respect to the average risk $\tilde{R}, d^N$ is in fact minimax. This result, however, is not of great support for d^N, since it shares this property with a large class of monotone decision rules, as we shall see in the next theorem.

THEOREM 2. *Let d^h be the decision rule which selects in terms of the largest $h_i(X_i), i = 1,\ldots,k$, where $h_1,\ldots,h_k$ are strictly increasing functions. Then d^h is minimax with respect to the average risk $\tilde{R}$, and the minimax value of the problem is again $1 - 1/k$.*

Proof: For every decision rule d, and for every permutation symmetric prior with density ρ w. r. t. the Lebesgue measure on $\mathbb{R}^k$, the Bayes risk satisfies

$$r(\rho,d) = \int_{\mathbf{R}^k} R(\underline{\theta}, d)\rho(\underline{\theta})d\underline{\theta}$$

$$= \int_{\{\underline{\theta}|\theta_1<\ldots<\theta_k\}} \tilde{R}(\underline{\theta}, d)\rho(\underline{\theta})d\underline{\theta} \leq \sup_{\underline{\theta}} \tilde{R}(\underline{\theta}, d). \tag{12}$$

Since the sequence of priors chosen below of (10) consists of such symmetric priors, it follows similarly as in the proof of Theorem 1 that the no-data rule d^0 is minimax w. r. t. $\tilde{R}$, and that the minimax value is again $1 - 1/k$. It remains thus to show that every rule of type d^h has supremal average risk $1 - 1/k$. Let d^h be any such rule, and let $\underline{\theta} \in \mathbb{R}^k$ be fixed, where we may assume without loss of generality that $\theta_1 < \theta_2 < \ldots < \theta_k$ holds. Then similarly as before in (7),

$$\begin{aligned}
&(1/k!)\sum_{\pi} P_{\pi(\underline{\theta})}(CS|d^h) \\
&\quad = (1/k!)\sum_{\pi} P_{\pi(\underline{\theta})}\{h_{\pi^{-1}(k)}(X_{\pi^{-1}(k)}) > h_{\pi^{-1}(j)}(X_{\pi^{-1}(j)}), j < k\} \qquad (13) \\
&\quad = (1/k!)\sum_{\pi} P\{h_{\pi^{-1}(k)}(\theta_k + \beta_{\pi^{-1}(k)}N_k) > h_{\pi^{-1}(j)}(\theta_j + \beta_{\pi^{-1}(j)}N_j), j < k\},
\end{aligned}$$

where $N_1, \ldots, N_k$ are independent standard normals, and $\beta_i = p_i^{\frac{1}{2}}, i = 1, \ldots, k$. A lower bound of (13) is attained if all $\theta_1, \ldots, \theta_{k-1}$ are put equal to θ_k, because of the monotonicity of $h_1, \ldots, h_k$. Doing so, and then splitting the sum into a suitable double sum, we see that the lower bound is

$$\begin{aligned}
&(1/k!)\sum_{i=1}^{k} \sum_{\pi, \pi(i)=k} P\{h_i(\theta_k + \beta_i N_k) \\
&\qquad > h_{\pi^{-1}(j)}(\theta_k + \beta_{\pi^{-1}(j)}N_j), j < k\} \\
&\quad = (1/k)\sum_{i=1}^{k} P\{h_i(\theta_k + \beta_i N_i) > h_j(\theta_k + \beta_j N_j), j \neq i\} = 1/k. \qquad (14)
\end{aligned}$$

Thus, in view of (1), the supremal average risk of d^h is equal to $1 - 1/k$, and the proof of the theorem is completed.

Our conclusions of this section are (1) that the natural rule d^N cannot be accepted as a universally good decision rule, and (2) that the minimax principle does not lead to a convincing alternative to d^N. Therefore it seems to be reasonable to study the form of Bayes rules with respect to various priors in more detail, in the hope to learn more about how such good decision rules act in different situations. Our main interest thereby will focus on permutation symmetric (exchangeable) and on conjugate priors. This will be done in the subsequent sections.

3. BAYES RULES FOR EXCHANGEABLE PRIORS

Permutation invariant (exchangeable) priors appear to be the suitable priors to adopt if there is no initial knowledge available as to how the ordered parameters $\theta_{[1]}, \ldots, \theta_{[k]}$ are

associated with the populations $P_1, \ldots, P_k$. They reflect the prior opinion that each of the k populations may equally likely be the one which has the largest mean.

Since we are considering Bayes rules, we may restrict considerations to nonrandomized decision rules d, which can be represented simply by measurable functions $d: \mathbb{R}^k \to \{1, 2, \ldots, k\}$, where $d(\underline{x}) = i$ means that at $\underline{X} = \underline{x}$, d selects population $P_i, i = 1, \ldots, k, \underline{x} \in \mathbb{R}^k$.

Now, for any prior τ, after $\underline{X} = \underline{x}$ has been observed, the Bayes rule selects that population which is associated with the smallest posterior expected loss. This decision process consists thus of pairwise comparisons of the k competing posterior risks. Ignoring a common factor, which depends on $\underline{x}$ and $\underline{p}$, the Bayes rule d^B can be written as

$$d^B(\underline{x}) = i \text{ if } \mathcal{G}(i|\underline{x}) = \max_{j=1,\ldots,k} \mathcal{G}(j|\underline{x}), \tag{15}$$

where

$$\mathcal{G}(s|\underline{x}) = \int_{\{\underline{\theta}|\theta_s = \theta_{[k]}\}} \prod_{j=1}^{k} \varphi((x_j - \theta_j)/p_j^{\frac{1}{2}}) d\tau(\underline{\theta}), \quad i, s = 1, \ldots, k.$$

To find out under which conditions one population is preferred over another one if τ is symmetric, let us compare without loss of generality $\mathcal{G}(2|\underline{x})$ and $\mathcal{G}(1|\underline{x})$, say, to keep the notation simple. After exchanging the variables θ_1 and θ_2 in the integral representation of $\mathcal{G}(2|\underline{x})$, and some standard calculations, we see that

$$\begin{aligned} &\mathcal{G}(2|\underline{x}) - \mathcal{G}(1|\underline{x}) \\ &\quad = \int_{\{\underline{\theta}|\theta_1 = \theta_{[k]}\}} [M_{2,1}(\underline{x}, \underline{\theta}) - 1] \prod_{i=1}^{k} \varphi((x_i - \theta_i)/p_i^{\frac{1}{2}}) d\tau(\underline{\theta}), \end{aligned} \tag{16}$$

where

$$\begin{aligned} &M_{2,1}(\underline{x}, \underline{\theta}) \\ &\quad = exp\{(\theta_1 - \theta_2)[(x_2 - (\theta_1 + \theta_2)/2)/p_2 - (x_1 - (\theta_1 + \theta_2)/2)/p_1]\}. \end{aligned}$$

Although the Bayes rules may have in general very complicated forms, several conclusions can be drawn from (16). The first one is

THEOREM 3. *Under a symmetric prior τ, suppose that for two populations P_a and P_b, say, the variances p_a and p_b are equal. Then the Bayes rule relatively ranks P_a and P_b in the same way as the natural rule d^N, namely according to the larger of the two values x_a and x_b, no matter what $x_i, i \neq a, b$, might actually be.*

Another finding is the following. Suppose we know a constant lower (upper) bound d to $\theta_1, \ldots, \theta_k$. Then if $p_a > (<) p_b$ and if $(x_a - d)/p_a > (x_b - d)/p_b$, every Bayes rule w.r.t. a symmetric prior prefers P_a to P_b.

If the prior knowledge asserts that the parameters $\theta_1, \ldots, \theta_k$ are in a slippage configuration $\theta_1 = \ldots = \theta_{i-1} = \theta_{i+1} = \ldots = \theta_k = \delta$, and $\theta_i = \delta + \Delta$, where $\delta \in \mathbb{R}$ and $\Delta > 0$ are known, and where apriori each $i \in \{1, \ldots, k\}$ may be, with the same probability $1/k$, the index of the slipped population, then from (16) it follows that the Bayes rule is given by

$$d^{\delta,\Delta}(\underline{x}) = i \;\; if \;\; (x_i - \delta - \Delta/2)/p_i = \max_{j=1,\ldots,k} (x_j - \delta - \Delta/2)/p_j. \tag{17}$$

It should be noted that it is quite different from the decision rule d^t, say, which selects in terms of the smallest p-value of the best 1-sample tests for $H_i: \;\; \theta_i = \delta$ versus $K_i: \;\; \theta_i = \delta + \Delta$, and thus selects in terms of the largest $(x_i - \delta)/p_i^{\frac{1}{2}}, i = 1, \ldots, k$.

Exchangeable normal priors give Bayes rules which are in general quite complicated in their structure. Although we know that the Bayes rule is determined by

$$\begin{aligned} d^B(\underline{x}) = i \;\; &\text{if} \;\; P\{\Theta_i = \Theta_{[k]} | \underline{X} = \underline{x}\} \\ &= \max_{j=1,\ldots,k} P\{\Theta_j = \Theta_{[k]} | \underline{X} = \underline{x}\}, \end{aligned} \tag{18}$$

and a prior $\underline{\Theta} \sim N(\underline{\mu}, A)$ with $\underline{X}|\underline{\Theta} = \underline{\theta} \sim N(\underline{\theta}, \Sigma)$ would result in $\Theta|\underline{X} = \underline{x} \sim N(\underline{x} - \Sigma(\Sigma + A)^{-1}(\underline{x} - \underline{\mu}), (\Sigma^{-1} + A^{-1})^{-1})$ where, marginally, $\underline{X} \sim N(\underline{\mu}, \Sigma + A)$, there is not much simplification to gain if we assume that $\underline{\mu} = \mu_0 \underline{1}$, and $A = aI + b\underline{1}\,\underline{1}^T$, where $\underline{1} = (1, \ldots, 1)^T, I$ is the $k \times k$ identity matrix, $\mu_0, b \in \mathbb{R}, a > 0$, and $a + kb > 0$ to have A positive definite, even if, as in the present setting, Σ is diagonal with diagonal elements $p_1, \ldots, p_k$.

One limiting case, however, the noninformative prior case, is of natural interest and leads in fact to an interesting decision rule. Suppose we are in the situation which led to (10), but now letting $r_1, \ldots, r_k$ tend to infinity. Then the generalized Bayes rule d^∞, say, can be seen to be based, formally, on $\Theta_i \sim N(x_i, p_i), i = 1, \ldots, k$, independent, at $\underline{X} = \underline{x}$, and to be given by

$$d^\infty(\underline{x}) = i \;\; if \;\; \mathcal{H}(i|\underline{x}) = \max_{j=1,\ldots,k} \mathcal{H}(j|\underline{x}), \tag{19}$$

where

$$\mathcal{H}(s|\underline{x}) = \int_{\mathbb{R}} \prod_{j \neq s} \Phi(p_j^{-\frac{1}{2}}(x_s - x_j + p_s^{\frac{1}{2}} z)) \varphi(z) dz, \;\; i, s = 1, \ldots, k.$$

One interesting feature of d^∞ is that it selects in terms of the largest variance among $p_1, \ldots, p_k$, whenever $x_1, \ldots, x_k$ are lying closely together. This is an immediate consequence of Lemma 1. We conclude this section by proposing two other type of decision rules which seem to be reasonable alternatives to d^N, worth to be studied in more detail in the future. The first is given by

$$\begin{aligned} &d^\epsilon(\underline{x}) = d^N(\underline{x}), \;\; \text{if} \;\; \underline{x} \notin B, \text{ and} \\ &d^\epsilon(\underline{x}) = i, \;\; \text{if} \;\; \underline{x} \in B \text{ and } p_i = \max_{j=1,\ldots,k} p_j, \end{aligned} \tag{20}$$

where $B \subseteq \mathbb{R}^k$ is an area where the coordinates of the vectors are close to each other, e. g. where $\max_{i,j} |x_i - x_j| < \epsilon$ for some $\epsilon > 0$. The other type of decision rule is of the form (19), where $\mathcal{H}$ is replaced by $\tilde{\mathcal{H}}$, say, with

$$\tilde{\mathcal{H}}(s|\underline{x}) = (x_s - \tilde{x})/p_s, \quad s = 1, \ldots, k, \tag{21}$$

and where $\tilde{x}$ is an average of $x_1, \ldots, x_k$, e. g. the weighted average with weights $p_1^{-1}, \ldots, p_k^{-1}$.

4. BAYES RULES FOR POSTERIORS WITH (DT).

One of the basic facts which lead to the "Bahadur *et. al.* Theorem" mentioned in the introduction is the following. Suppose that at every $\underline{X} = \underline{x}$, the posterior depends on $\underline{x}$ through $\underline{g}(\underline{x}) = (g_1(\underline{x}), \ldots, g_k(\underline{x}))$, where $g_1, \ldots, g_k$ are given functions. Then if the posterior is (DT) in $(\underline{\theta}, \underline{g}(\underline{x}))$, and if the loss function is permutation invariant and favors selection of larger parameters, then the posterior risk acts like the loss function where $\underline{g}(\underline{x})$ plays the role of $\underline{\theta}$. For details see Gupta and Miescke (1984). Thus the Bayes rule selects here in terms of the largest $g_i(\underline{x}), i = 1, \ldots, k$.

Under a normal prior $\underline{\Theta} \sim N(\underline{\mu}, A)$, as considered after the statement of (18), the posterior is (DT) if and only if the covariance matrix associated with it is of the form

$$(\Sigma^{-1} + A^{-1})^{-1} = a^2[(1-\rho)I + \rho \underline{1}\,\underline{1}^T], \tag{22}$$

where $a \in \mathbb{R}$ and $-(k-1)^{-1} < \rho < 1$ are necessary and sufficient for this matrix to be positive definite.

If (22) holds true, the conditional expectation of $\underline{\Theta}$, given $\underline{X} = \underline{x}$, can be seen to be

$$\begin{aligned} E\{\underline{\Theta}|\underline{X} = \underline{x}\} &= \underline{\mu} + \gamma(\underline{x})\underline{1} + \\ &+ a^2(1-\rho)((x_1 - \mu_1)/p_1, \ldots, (x_k - \mu_k)/p_k), \end{aligned} \tag{23}$$

where $\gamma(\underline{x})$ is a certain function which is, as we shall see, of no relevance for the Bayes rule. Namely, if we set $g_i(\underline{x}) = E\{\Theta_i|\underline{X} = \underline{x}\}, i = 1, \ldots, k, \underline{x} \in \mathbb{R}^k$, then the posterior is (DT) in $(\underline{\theta}, \underline{g}(\underline{x}))$, and the Bayes rule is given by

$$d^B(\underline{x}) = i \text{ if } \mathcal{F}(i|\underline{x}) = \max_{j=1,\ldots,k} \mathcal{F}(j|\underline{x}), \tag{24}$$

where $\mathcal{F}(s|\underline{x}) = \mu_s + a^2(1-\rho)(x_s - \mu_s)/p_s, \quad i, s = 1, \ldots, k.$

Of special interest hereby is the case of $\mu_1 = \ldots = \mu_k = \mu$, where the Bayes rule d^μ, say, assumes the simple form

$$d^\mu(\underline{x}) = i \text{ if } (x_i - \mu)/p_i = \max_{j=1,\ldots,k} (x_j - \mu)/p_j, \tag{25}$$

$i = 1, \ldots, k$, which is almost the same as that one of the Bayes rule for the slippage situation, given by (17).

The interesting feature of the rule d^{∞}, discussed just after the statement of (19), has an analog in the rules given by (25) and (17). If there are (almost) tied x_i's, which are smaller (larger) than μ or $\delta + \Delta/2$, respectively, then the Bayes rules prefer the population with the larger (smaller) variance.

The choice of a prior, which results in a posterior with the (DT)- property and ultimately in a Bayes rule of simple structure, is made not only for convenience. It has also a statistical justification since it leads to a posterior situation where the information about the unknown parameters $\theta_1, \ldots, \theta_k$ is equally and thus fairly balanced. This can be seen perhaps most easily in the case where $\rho = 0$ in (22). Then A is diagonal with diagonal elements $r_1, \ldots, r_k$, say, which brings us back to the situation considered at (10), where now we have

$$p_i^{-1} + r_i^{-1} = a^{-2}, \quad i = 1, \ldots, k. \tag{26}$$

Calling, as usual, the inverse of a variance "precision," the sum of the prior precision and the sampling precision is constant across the k populations, if (26) holds.

Returning to the original form of the problem, as it was presented in the introduction, we can state the following interesting fact. Suppose that the prior is known, as it should be, before the sampling is performed. Suppose further that the sample sizes from the populations $P_1, \ldots, P_k$ can be chosen in such a way that (22) holds, which means that in case of a diagonal A, the condition (26) is fulfilled. Then the information about the unknown parameters $\theta_1, \ldots, \theta_k$ is fairly balanced, and the Bayes decision rule assumes the simple form given by (24) or (25), respectively. It should be pointed out clearly, that in this case the Bayes rule is the same under every loss which is permutation invariant and favors selection of larger parameters.

5. CONCLUDING REMARKS.

It can be seen easily that the natural rule d^N is an extended Bayes rule. Since if apriori, $\Theta_1, \ldots, \Theta_k$ are i. i. d. $N(0, n)$, then the Bayes risk of d^N with respect to this prior tends to 0 if n tends to infinity. This is not surprising, as we know that the performance of d^N is only unsatisfactory if the parameters $\theta_1, \ldots, \theta_k$ are close together. On the other hand, we saw that d^N cannot be the Bayes rule for any normal prior $\underline{\Theta} \sim N(\underline{\mu}, A)$.

We could not settle, however, the interesting question of whether or not d^N is admissible under the $0-1$ loss function on the parameter space $\Omega = \{\underline{\theta} \in \mathrm{R}^k | \theta_{[k]} \text{is unique}\}$. The restriction of parameters to Ω is made for simplicity, and does not cause any loss of generality. For other loss functions, however, this restriction may not simplify matters and may not be made, as e. g. in the example given below.

The $0-1$ loss function was adopted in our study because it connects the risk function in a natural way through (1) with the probability of a correct selection, which is the performance characteristic of decision rules considered primarily in the area of ranking and selection. With respect to other loss functions, however, the natural rule d^N may in

fact be a proper Bayes rule and admissible on $\mathbb{R}^k$, as the following example demonstrates.

Assume that the loss for selecting population P_i at parameters $\theta_1, \ldots, \theta_k$ is of the form

$$L(\underline{\theta}, i) = \theta_{[k]} - \theta_i, \quad i = 1, \ldots, k, \quad \underline{\theta} \in \mathbb{R}^k. \tag{27}$$

Then the Bayes rule at $\underline{X} = \underline{x}$ is given by

$$d^*(\underline{x}) = i \text{ if } E\{\Theta_i | \underline{X} = \underline{x}\} = \max_{j=1,\ldots,k} E\{\Theta_j | \underline{X} = \underline{x}\}, \quad i = 1, \ldots, k. \tag{28}$$

Therefore, if apriori, $\Theta_i \sim N(\mu_i, r_i), i = 1, \ldots, k$, independent, the Bayes rule at $\underline{X} = \underline{x}$ turns out to be

$$d^b(\underline{x}) = i \text{ if } M(i|\underline{x}) = \max_{j=1,\ldots,k} M(j|\underline{x}), \tag{29}$$

where $M(s|\underline{x}) = (p_s \mu_s + r_s x_s)/(p_s + r_s), s = 1, \ldots, k$. And it can be seen now that d^b is the natural rule d^N if $\mu_i = \mu$ and $r_i = c\ p_i, i = 1, \ldots, k$, for some fixed $\mu \in \mathbb{R}$ and $c > 0$.

The admissibility of all Bayes rules considered in this paper, those under $0-1$ loss on Ω, as well as those under the loss function (27) on $\mathbb{R}^k$, follows from the fact that the risk function of every selection rule in these problems is continuous in $\underline{\theta}$. This is an immediate consequence of the well known fact that the expectation of a bounded function under a multi-parameter exponential family is continuous in these parameters.

ACKNOWLEDGEMENTS

The research of the first author was supported by the Office of Naval Research Contract N00014–84–C–0167 and NSF Grant DMS–860694 at Purdue University. The research of the second author was supported by the Air Force Office of Scientific Research Contract AFOSR–85–0347 at the University of Illinois at Chicago.

REFERENCES

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

[2] Bahadur, R. and Goodman, L. (1952). Impartial decision rules and sufficient statistics. *Ann. Math. Statist.* **23**, 553–562.

[3] 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.

[4] Bechhofer, R. E. , and Tamhane, A. C. (1986). Private Communication.

[5] Berger, J. O. (1983). Comment on [16]. *J. Amer. Statist. Assoc.* **78**, 55–57.

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

[7] Ehrman, C. M., Krieger, A., and Miescke, K. J. (1986). Subset selection towards optimizing the best performance at a second stage. 25 pp. in ms. *Journal of Business and Economic Statistics.* To appear.

[8] Gupta, S. S. and Huang, D. Y. (1976). Subset selection procedures for the means and variances of normal populations: unequal sample size case. *Sankhyā*, Series B, **38**, 112–128.

[9] Gupta, S. S. and Miescke, K. J. (1984). Sequential selection procedures: A decision theoretic approach. *Ann. Statist.* **12**, 336–350.

[10] Gupta, S. S., and Miescke, K. J. (1985). Minimax multiple t- tests for comparing k normal populations with a control. *J. Statist. Plann. Inference* **12**, 161–169.

[11] Lam, K., and Chiu, W. K. (1976). On the probability of correctly selecting the best of several normal populations. *Biometrika* **63**, 410–411.

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

[13] Miescke, K. J. (1979). Identification and selection procedures based on tests. *Ann. Statist.* **7**, 207–219.

[14] Miescke, K. J. (1981). Gamma-minimax selection procedures in simultaneous testing problems. *Ann. Statist.* **9**, 215–220.

[15] Miescke, K. J. (1984). Two-stage selection procedures based on tests. *Design of Experiments: Ranking and Selection (Essays in honor of R. E. Bechhofer)*, T. J. Santner, A. C. Tamhane eds., Marcel Dekker, 1984, 165–178.

[16] Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78**, 47–65.

[17] Tong, Y. L., and Wetzell, D. E. (1979). On the behavior of the probability function for selecting the best normal population. *Biometrika* **66**, 174–176.

[18] Zaher, A. M., and Heiny, R. L. (1984). A comparison of selection procedures for selecting the best normal population under heterogeneity of variance. *Commun. Statist.-Simula. Computa.* **13**, 635–654.

BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
Technical Report #86-27		

4. TITLE (and Subtitle)
ON THE PROBLEM OF FINDING THE LARGEST NORMAL MEAN UNDER HETEROSCEDASTICITY

5. TYPE OF REPORT & PERIOD COVERED
Technical

6. PERFORMING ORG. REPORT NUMBER
Technical Report #86-27

7. AUTHOR(s)
Shanti S. Gupta and Klaus J. Miescke

8. CONTRACT OR GRANT NUMBER(s)
N00014-84-C-0167

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Purdue University
Department of Statistics
West 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
July 1986

13. NUMBER OF PAGES
13

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)
Selecting the largest mean; Heteroscedasticity in normal populations; Bayes selection rules.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Let $\pi_1,\ldots,\pi_k$ be $k \geq 3$ given normal populations with unknown means $\theta_1,\ldots,\theta_k$, and a common known variance σ^2. Let $\bar{X}_1,\ldots,\bar{X}_k$ be the sample means of k independent samples of sizes $n_1,\ldots,n_k$ from these populations. To find the population with the largest mean, one usually applies the natural rule d^N, which selects in terms of the largest sample mean.

In this paper, the performance of this rule is studied under 0-1 loss. It is shown

cont.

DD FORM 1 JAN 73 1473

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

that d^N is minimax if and only if $n_1 = \ldots = n_k$. d^N is seen to perform weakly whenever the parameters $\theta_1, \ldots, \theta_k$ are close together. Several alternative selection rules are derived in a Bayesian approach which seem to be reasonable competitors to d^N, worth comparing with d^N in a future simulation study.