ON FINDING THE LARGEST NORMAL MEAN
AND ESTIMATING THE SELECTED MEAN*

by

Shanti S. Gupta*
Purdue University
West Lafayette, IN, USA

and

Klaus J. Miescke**
University of Illinois at
Chicago, IL, USA

Purdue University
Technical Report #89-23C

PURDUE UNIVERSITY

CENTER FOR STATISTICAL DECISION SCIENCES AND DEPARTMENT OF STATISTICS

ON FINDING THE LARGEST NORMAL MEAN AND ESTIMATING THE SELECTED MEAN*

by

Shanti S. Gupta* and Klaus J. Miescke**
Purdue University University of Illinois at
West Lafayette, IN, USA Chicago, IL, USA

Purdue University
Technical Report #89–23C

Department of Statistics
Purdue University

August 1989

*This research was supported in part by the Office of Naval Research Contract N00014–88–K–0170 and NSF Grants DMS–8606964, DMS–8702620 at Purdue University.

**The research of this author was also supported by an Air Force of Scientific Research Contract AFOSR–85–0347 at the University of Illinois at Chicago.

△

ON FINDING THE LARGEST NORMAL MEAN AND ESTIMATING THE SELECTED MEAN

by

Shanti S. Gupta
Purdue University
West Lafayette, IN, USA

and

Klaus J. Miescke
University of Illinois at Chicago
Chicago, IL, USA

Summary

For $k \geq 2$ independent normal populations with unknown means and a common known variance, the problem of selecting the population with the largest mean and simultaneously estimating the mean of the selected population is considered in the decision theoretic approach following Cohen and Sackrowitz (1988). Under several loss functions with two additive components due to selection and due to estimation, Bayes decision rules are derived and studied. Both, the case of equal sample sizes and the case of unequal sample sizes are treated. The "natural" rule, which selects in terms of the largest sample mean and then estimates with the sample mean of the selected population, is critically examined in all situations considered.

1. Introduction

Let $\pi_1, \ldots, \pi_k$ be $k \geq 2$ given normal populations with unknown means $\theta_1, \ldots, \theta_k \in \mathbf{R}$, and a common known variance $\sigma^2 > 0$. Suppose we want to find the population with the largest mean and simultaneously estimate the mean of the selected population; here the observed data are k independent samples of sizes $n_1, \ldots, n_k$ from $\pi_1, \ldots, \pi_k$ with sample means $X_1, \ldots, X_k$, respectively.

All results in the vast literature on ranking and selection are restricted to one of the two decision problems, except one. Cohen and Sackrowitz (1988) have presented a decision theoretic framework for the combined decision problem and derived results for the case of $k = 2$ and $n_1 = n_2$.

Selecting the population with the largest sample mean $X_{[k]}$, say, is usually called the natural selection rule, since it is the uniformly best permutation invariant selection procedure for a general class of loss functions if the sample sizes $n_1, \ldots, n_k$ are all equal. However, for unequal sample sizes, the natural selection rule loses much of its quality. In fact, under 0-1 loss, it can perform "worse than at random" if $\theta_1, \ldots, \theta_k$ are close together. This is studied in detail in Gupta and Miescke (1988).

Estimating the mean of the selected population has been considered only under the assumption that the natural selection rule is employed. Knowing that the natural estimator X_i for θ_i, in case of π_i being selected, $i = 1, \ldots, k$, overestimates $\theta_{[k]} = \max\{\theta_1, \ldots, \theta_k\}$, and thus overestimates even more so the mean of the selected population, alternative estimators have been studied for the present and for other experimental models by the following authors: Sarkadi (1967), Dahiya (1974), Cohen and Sackrowitz (1982), Sackrowitz

and Samuel-Cahn (1984, 1986), Jeyaratnam and Panchapakesan (1984, 1986, 1988), Vellaisamy and Sharma (1988, 1989), Vellaisamy, Kumar and Sharma (1988), and Venter (1988).

Rather than "estimating after selection", the decision theoretic treatment of the combined selection-estimation problem leads to "selecting after estimation", as it has been pointed out by Cohen and Sackrowitz (1988). So far, however, only a few limited situations have been considered. The purpose of this study is to extend known results in several directions. First, the case of $k > 2$ populations needs to be considered since in selection problems alone, typical features and difficulties do not appear before k is at least equal to three. Second, for the two additive components of the loss function due to selection and due to estimation, alternatives to 0-1 loss and squared error loss, respectively, have to be examined. Zero-One loss for selection has the undesired effect of a stiff penalty for selecting a non-best population even if its mean is close to the largest mean, and absolute error loss is a reasonable alternative to squared error loss for estimation. Third, the case of unequal sample sizes $n_1, \ldots, n_k$ has to be considered. If selection alone is under concern, better rules than the natural selection procedure have been derived in Gupta and Miescke (1988) for this purpose, all of which take into account the precisions with which the sample means $X_1, \ldots, X_k$ represent the unknown means $\theta_1, \ldots, \theta_k$. Thus, in the combined selection-estimation problem, where additionally the precisions of the estimates depend on the respective sample sizes, non-standard decision rules are to be expected. This will be discussed in Section 4, after a general framework has been introduced in Section 2, and the case of equal sample sizes has been treated in Section 3.

2. General Framework

Let $\underline{X} = (X_1, \ldots, X_k)$ be a random vector of observations which has a density (probability function) $\prod_{i=1}^{k} f_i(x_i|\theta_i), \underline{x} \in \mathbf{R}^k(\mathbf{Z}^k), \underline{\theta} \in \Omega^k \subseteq \mathbf{R}^k$, with respect to the Lebesgue measure on $\mathbf{R}^k$ (counting measure on $\mathbf{Z}^k$). Of course $\underline{X}$ may be already a collection of k sufficient statistics for $\theta_1, \ldots, \theta_k$.

The goal is to select the population, i.e. coordinate, which is associated with $\theta_{[k]} = \max\{\theta_1, \ldots, \theta_k\}$, and to simultaneously estimate the θ-value of the selected population. Since Bayes rules are the main topic of this paper, only nonrandomized decision rules need to be considered which are represented as follows.

$$\underline{d}(\underline{x}) = (s(\underline{x}), \ell_{s(\underline{x})}(\underline{x})), \ \underline{x} \in \mathbf{R}^k, \tag{1}$$

where $s(\underline{x}) \in \{1, 2, \ldots, k\}$ is the selection rule, and $\ell_i(\underline{x}) \in \Omega, i = 1, \ldots, k$, is a collection of k estimates for $\theta_i, i = 1, \ldots, k$, respectively, available at selection.

The loss function is assumed to be addive,

$$L(\underline{\theta}, \underline{d}) = A(\underline{\theta}, s) + B(\theta_s, \ell_s), \tag{2}$$

where A represents the loss of selecting population π_s at $\underline{\theta}$, and B the loss of estimating θ_s by ℓ_s. The following examples will be considered.

$$A_0(\underline{\theta}, s) = cI_{\{\theta_{[k]}\}}(\theta_s); \tag{3}$$

$$A_1(\underline{\theta}, s) = c(\theta_{[k]} - \theta_s); \ B_1(\theta_s, \ell_s) = |\theta_s - \ell_s|;$$

$$A_2(\underline{\theta}, s) = c(\theta_{[k]} - \theta_s)^2; \ B_2(\theta_s, \ell_s) = (\theta_s - \ell_s)^2.$$

All combinations of A's and B's are reasonable in one way or another, and $c > 0$ gives relative weights to the two types of losses. However, it seems that the most appealing and

realistic combinations are

$$\mathcal{L}_1(\underline{\theta}, \underline{d}) = c(\theta_{[k]} - \theta_s) + |\theta_s - \ell_s|, \text{ and} \tag{4}$$

$$\mathcal{L}_2(\underline{\theta}, \underline{d}) = c(\theta_{[k]} - \theta_s)^2 + (\theta_s - \ell_s)^2.$$

In the Bayes approach, let the vector of the k unknown parameters be random and denoted by $\underline{\Theta}$. Under a prior distribution of it, the posterior risk at $\underline{X} = \underline{x}$ can be represented as follows.

$$r(\underline{d}(\underline{x})|\underline{x}) = r_A(s(\underline{x})|\underline{x}) + r_B(s(\underline{x}), \ell_{s(\underline{x})}(\underline{x})|\underline{x}), \tag{5}$$

where $r_A(s(\underline{x})|\underline{x}) = E\{A(\underline{\Theta}, s(\underline{x}))|\underline{X} = \underline{x}\}$, and $r_B(s(\underline{x}), \ell_{s(\underline{x})}, (\underline{x})|\underline{x}) = E\{B(\Theta_{s(\underline{x})}, \ell_{s(\underline{x})}(\underline{x}))|\underline{X} = \underline{x}\}$.

As it has been mentioned in the Introduction, the decision theoretic treatment of the combined selection-estimation problem leads to "selecting after estimation". This can be seen now from the following fact which is a straightforward extension of the main result in Cohen and Sackrowitz (1988).

Lemma 1. Let $\ell_i^*(\underline{x})$ minimize $r_B(i, \ell_i(\underline{x})|\underline{x}), i = 1, \ldots, k$. Furthermore, let $s^*(\underline{x})$ minimize $r_A(s(\underline{x})|\underline{x}) + r_B(s(\underline{x}), \ell^*_{s(\underline{x})}(\underline{x})|\underline{x})$. Then the Bayes decision rule, at $\underline{X} = \underline{x}$, is $\underline{d}^*(\underline{x}) = (s^*(\underline{x}), \ell^*_{s^*(\underline{x})}(\underline{x}))$.

It should be pointed out that $\ell_i^*(\underline{x})$ is the usual Bayes estimate of $\theta_i, i = 1, \ldots, k$, if estimation alone is under concern. There is no bias reduction involved which has been the main concern in papers dealing with estimation after selection mentioned in the Introduction.

Under certain circumstances, the problems of selection and estimation can be completely separated. More precisely, the following holds.

Corollary 1. Whenever at some $\underline{X} = \underline{x}$, $r_B(i, \ell_i^*(\underline{x})|\underline{x})$ does not depend on $i \in \{1, 2, \ldots, k\}$, $s^*(\underline{x})$ minimizes $r_A(s(\underline{x})|\underline{x})$.

Let us consider briefly the selection problem by its own, i.e. assume that loss function B in (2) is zero. Then the natural selection rule $s^N(\underline{x})$, which selects in terms of the largest x_i, is known to have strong optimality properties. If the density of $\underline{X}$ is of the form $\prod_{i=1}^{k} f(x_i|\theta_i)$, where f has monotone likelihood ratios, and if loss function A in (2) is permutation invariant and favors selection of larger θ-values, then s^N is the best permutation invariant selection rule, uniformly in $\underline{\theta}$, i.e. it is Bayes selection rule for every permutation symmetric prior. This and further results can be found in Gupta and Miescke (1984).

In combination with estimation of the parameter of the selected population, however, it can occur that in the above situation, where s^N is uniformly best invariant selection rule, s^N is not part of the Bayes rule $\underline{d}^* = (s^*, \ell_{s^*}^*)$, i.e. s^* is different from s^N. More precisely, this happens when the assumption of Corollary 1 are not met. The following example illustrates this fact.

Example 1. Let $X_i \sim N(\theta_i, 1), i = 1, \ldots, k$, be independent. Assume that a priori, $\Theta_1, \ldots, \Theta_k$ are a random sample from an exponential distribution with density $\exp(-\theta)$, $\theta > 0$. Finally, let $L(\underline{\theta}, \underline{d}) = A(\underline{\theta}, s) + (\theta_s - \ell_s)^2$, where A is permutation invariant and favors selection of larger θ-values.

A posterior, at $\underline{X} = \underline{x}$, $\Theta_1, \ldots, \Theta_k$ are independent, and the posterior density of Θ_i is $\varphi(\theta_i - y_i)/\Phi(y_i), \theta_i > 0$, where $y_i = x_i - 1, i = 1, \ldots, k$, and where φ and Φ denote the

density and c.d.f. of $N(0,1)$.

Standard calculations lead to the following results for $i = 1, \ldots, k$.

(6) $\ell_i^*(\underline{x}) = E\{\Theta_i | \underline{X} = \underline{x}\} = y_i + \varphi(y_i)/\Phi(y_i)$, and

$$r_B(i, \ell_i^*(\underline{x}) | \underline{x}) = \text{ Var } \{\Theta_i | \underline{X} = \underline{x}\} = 1 + 2y_i^2 + y_i\varphi(y_i)/\Phi(y_i) - [\varphi(y_i)/\Phi(y_i)]^2.$$

Thus, although $s^N(\underline{x})$ minimizes $r_A(s(\underline{x})|\underline{x}), s^N$ is not equal to s^*, since $r_B(i, \ell_i^*(\underline{x})|\underline{x})$ depends on $i \in \{1, \ldots, k\}$ except for a Lebesgue null set.

At the end of this section, let us briefly justify the choice of a Bayes approach to the given problem by pointing out that the classical (frequentist) approach does not offer a direct analytical solution. The risk function for a decision rule $\underline{d}$ at parameter point $\underline{\theta} \in \mathbf{R}^k$ is given by

(7) $$R(\underline{\theta}, \underline{d}) = E_{\underline{\theta}}[A(\underline{\theta}, s(\underline{X}))] + E_{\underline{\theta}}[B(\theta_{s(\underline{X})}, \ell_{s(\underline{X})}(\underline{X}))].$$

For one fixed given selection rule s, the second term can be optimized, at least approximately, in many circumstances. This has been done in the previous papers dealing with estimation of the parameter of the selected population, where $s = s^N$ has been assumed. However, to optimize $R(\underline{\theta}, \underline{d})$, one has to consider at least some class of possible selection rules for s, which appears to be not feasible. Bayes rules, on the other hand, can be found in a constructive way as it is shown in Lemma 1.

3. Independent Normal Populations With Equal Sample Sizes

Let $X_{i1}, \ldots, X_{in}$ be a sample from $N(\theta_i, \sigma^2), i = 1, \ldots, k$, where $\sigma^2 > 0$ is known, and

let all samples be independent. Let $X_i = n^{-1} \sum_{j=1}^{n} X_{ij}, i = 1, \ldots, k$, be the sample means, which are sufficient for $\underline{\theta}$.

Assume that a priori, $\Theta_1, \ldots, \Theta_k$ is a random sample from $N(\mu, q)$, where $\mu \in \mathbf{R}$ and $q > 0$ are known. This conjugate prior will prove to be useful in several aspects, as it has done so previously in Sackrowitz and Samuel-Cahn (1984), Cohen and Sackrowitz (1988), and Gupta and Miescke (1988). Aposteriori, given $\underline{X} = \underline{x}, \Theta_i \sim N\left(\frac{qx_i + p\mu}{q+p}, \frac{pq}{q+p}\right), i = 1, \ldots, k$, are independent, where $p = \sigma^2/n$. And marginally, $X_1, \ldots, X_k$ is a sample from $N(\mu, q+p)$.

For a slightly more general class of priors, the following result can be shown to hold.

Theorem 1. For the loss function $L = A + B$ in (2), assume that A is permutation symmetric and favors selection of larger θ-values, and that B is either B_1 or B_2 in (3). Then for every exchangeable normal prior, the Bayes rule $\underline{d}^* = (s^*, \ell^*_{s^*})$ satisfies $s^* = s^N$ and $\ell^*_i(\underline{x}) = E\{\Theta_i | \underline{X} = \underline{x}\}, i = 1, \ldots, k$.

Proof: Apriori, let $\underline{\Theta} \sim N(\mu \underline{1}, aI + b\underline{1}\,\underline{1}^T)$, where $a > 0$, $a + kb > 0, \underline{1}^T = (1, 1, \ldots, 1)$, and I is the identity matrix. Then aposteriori, given $\underline{X} = \underline{x}, \underline{\Theta} \sim N(\underline{\ell}^*(\underline{x}), \alpha I + \beta \underline{1}\,\underline{1}^T)$, where $\alpha = ap/(a+p), \beta = bp^2/[(p+a+kb)(p+a)]$, and

$$\underline{\ell}^*(\underline{x}) = E\{\underline{\Theta} | \underline{X} = \underline{x}\} \tag{8}$$

$$= [\alpha I + \beta \underline{1}\,\underline{1}^T][(a+kb)^{-1}\mu \underline{1} + p^{-1}\underline{x}],$$

since $E\{\Theta_i | \underline{X} = \underline{x}\}$ minimizes $r_B(i, \ell_i(\underline{x}) | \underline{x})$ under both $B = B_1$ and $B = B_2$. The minimum values are, respectively,

$$r_{B_1}(i, \ell^*_i(\underline{x}) | \underline{x}) = [2(\alpha + \beta)/\pi]^{1/2}, \text{ and} \tag{9}$$

$$r_{B_2}(i, \ell_i^*(\underline{x})|\underline{x}) = \alpha + \beta,$$

which do not depend on $i \in \{1, 2, \ldots, k\}$ and $\underline{x}$. The latter fact will be utilized later in this section.

Thus, the assumption of Corollary 1 is fulfilled at every $\underline{x} \in \mathbf{R}^k$, and from the discussion following Corollary 1 it follows that $s^N(\underline{x})$ minimizes $r_A(s(\underline{x})|\underline{x})$ at every $\underline{x} \in \mathbf{R}^k$, i.e. $s^* = s^N$. This completes the proof of the theorem.

In the remainder of this section let us consider the natural decision procedure $\underline{d}^N = (s_N, \ell_{s^N}^N)$, which employs the estimates $\ell_i^N(\underline{x}) = x_i, i = 1, \ldots, k$. Although from the frequentist point of view, it has the undesirable features of overestimating the largest mean and thus even more so the selected mean, $\underline{d}^N$ is generalized Bayes rule for the noninformative prior, i.e. the Lebesgue measure on $\mathbf{R}^k$. The i.i.d. normal prior considered at the beginning of this section can be used for further examinations of $\underline{d}^N$, since for q tending to infinity, the posterior distributions tend to the formal posterior distribution associated with the noninformative prior.

As mentioned in the Introduction, typical features and difficulties in selection problems do not appear before k is at least equal to three. The following result may be considered as an illustration of this fact.

Theorem 2. For the loss function $L = A + B$ in (2), assume that $A = A_0$, and that B is either B_1 or B_2 in (3). Then the following holds. The rule $\underline{d}^N = (s^N, \ell_{s^N}^N)$ is minimax if and only if $k = 2$.

Proof: Obviously, $\ell_{s^N(\underline{x})}^N(\underline{x}) = x_{[k]}, \underline{x} \in \mathbf{R}^k$. Cohen and Sackrowitz (1982) have shown that for every loss function L with $L(0) = 0, L(z) = L(-z)$, and $L(|z|)$ increasing in

$|z|, E_{\underline{\theta}}(L(X_{[k]} - \theta_{s^N(\underline{X})}))$ is maximized at $\underline{\theta} = \underline{0}$. Thus, the maximum of $E_{\underline{\theta}}(|X_{[k]} - \theta_{s^N(\underline{X})}|^m)$ is found to be equal to $p(a_k^2 + b_k^2)$ for $m = 2$, and equal to $p^{1/2}c_k$ for $m = 1$, where $a_k^2 =$ Var $(N_{[k]}), b_k = E(N_{[k]})$, and $c_k = E(|N_{[k]}|)$, and $N_{[k]}$ is the maximum of a sample of size k from a standard normal distribution. The first fact has been shown in Sackrowitz and Samuel-Cahn (1986), and the second follows in a similar way.

The maximum of $E_{\underline{\theta}}(A_0(\underline{\theta}, s^N(\underline{X})))$ occurs also at $\underline{\theta} = \underline{0}$, and it is equal to $c(1 - 1/k)$, which has been shown in Gupta and Miescke (1988). Thus,

$$\max_{\underline{\theta}} R(\underline{\theta}, \underline{d}^N) = \begin{cases} c(1 - 1/k) + p^{1/2}c_k, & \text{if } B = B_1 \\ c(1 - 1/k) + p(a_k^2 + b_k^2), & \text{if } B = B_2 \end{cases} . \tag{10}$$

At this point it is convenient to consider the following randomized rule $\underline{d}^0$, say, which uses $\ell_i^N(\underline{x}) = x_i, i = 1, \ldots, k$, and selects, without any consideration of the data, each population with the same probability $1/k$. Obviously, for all $\underline{\theta}$,

$$R(\underline{\theta}, \underline{d}^0) = \begin{cases} c(1 - 1/k) + (2p/\pi)^{1/2}, & \text{if } B = B_1 \\ c(1 - 1/k) + p, & \text{if } B = B_2 \end{cases} \tag{11}$$

which provide upper bounds to the respective minimax values. And since $a_k^2 + b_k^2 > 1$ as well as $c_k > (2/\pi)^{1/2}$ for $k \geq 3$, $\underline{d}^N$ cannot be minimax rule for $k \geq 3$.

Finally, to see that $\underline{d}^N$ is minimax for $k = 2$, one realizes that (10) and (11) are identical in this case, since $c_2 = (2/\pi)^{1/2}$, and $a_2^2 + b_2^2 = 1$. It suffices to find a sequence of priors whose Bayes risk tend to (11), because $\underline{d}^0$ is an equalizer rule. This sequence is provided by the conjugate prior considered at the beginning of this section, by letting q vary, $q = 1, 2, \ldots$. From (9), with $a = q, \alpha = qp/(q + p)$, and $\beta = 0$, it follows that the part of the Bayes risk due to estimation tends, for large q-values, to $(2p/\pi)^{1/2}$ and p under $B = B_1$ and $B = B_2$, respectively. The other part of the Bayes risk due to selection tends

to $c(1-1/k)$, as it has been shown by Gupta and Miescke (1988). This completes the proof of the theorem.

Remark: The "if"-part of Theorem 2 is also valid under the assumption of $A = A_1$ or $A = A_2$. Although the parameter configuration $\theta_1 = \theta_2 = \ldots = \theta_k$ is not least favorable for $E_{\underline{\theta}}[A_i(\underline{\theta}, s^N(\underline{X}))]$, minimaxity holds because the risk due to estimation of $\underline{d}^N$ is constant in $\underline{\theta}$ for $k = 2$. For $k \geq 3$, however, the question of minimaxity remains open since the two risk parts, due to selection and due to estimation, do not assume their maxima at a common parameter configuration.

We conclude this section with the following.

Theorem 3. For every loss function (2), with components A and B taken from (3), the rule $\underline{d}^N = (s^N, \ell^N_{s^N})$ is extended Bayes rule.

Proof: Consider the same sequence of priors which was used at the end of the proof of Theorem 2. Under $B = B_2$, the posterior risk due to estimation of $\underline{d}^N$ is given by

$$
\begin{aligned}
(12) \qquad & E\{(\Theta_{s^N(\underline{x})} - x_{[k]})^2 | \underline{X} = \underline{x}\} \\
& = E\{(\Theta_{s^N(\underline{x})} - \ell^*_{s^N(\underline{x})}(\underline{x}))^2 | \underline{X} = \underline{x}\} + (\ell^*_{s^N(\underline{x})}(\underline{x}) - x_{[k]})^2 \\
& = pq(p+q)^{-1} + [p(p+q)^{-1}]^2 (x_{[k]} - \mu)^2.
\end{aligned}
$$

Since, marginally, $X_1, \ldots, X_k$ is a sample from $N(\mu, p+q)$, the Bayes risk due to estimation of $\underline{d}^N$ turns out to be

$$
(13) \qquad pq(p+q)^{-1} + p^2(p+q)^{-1}(a_k^2 + b_k^2),
$$

which tends to p, as q tends to infinity, where p is also the limit of the Bayes risk due to estimation.

Under $B = B_1$, it follows now immediately that the corresponding Bayes risk due to estimation of $\underline{d}^N$ satisfies

$$E(|\Theta_{s^N(\underline{X})} - X_{[k]}|) \underset{q\to\infty}{\longrightarrow} (2p/\pi)^{1/2}, \tag{14}$$

since, by the facts stated above,

$$E((\ell^*_{s^N(\underline{X})} - X_{[k]})^2) \underset{q\to\infty}{\longrightarrow} 0. \tag{15}$$

The second part of the proof deals with the other part of the Bayes risk due to selection. Since under each $A = A_i, = 0, 1, 2$, s^N is employed by the Bayes rule, i.e. $s^N = s^*$, it remains to be shown that in all three cases the limits of Bayes risks are finite. The case of $A = A_0$ has been treated in Gupta and Miescke (1988), where it is shown that the Bayes risk due to selection of $\underline{d}^N$ tends to $c(1 - 1/k)$ as q tends to infinity.

For $A = A_2$, the risk due to selection of $\underline{d}^N$ satisfies at every fixed $\underline{\theta}$ with, say, $\theta_k = \theta_{[k]}$,

$$\begin{aligned} & E_{\underline{\theta}}[A_2(\underline{\theta}, s^N(\underline{X}))] \\ &= \sum_{i=1}^{k-1} (\theta_k - \theta_i)^2 P_{\underline{\theta}}\{X_i = X_{[k]}\} \\ &\leq \sum_{i=1}^{k-1} (\theta_k - \theta_i)^2 P_{\underline{\theta}}\{X_i \geq X_k\} \\ &= 2p \sum_{i=1}^{k-1} \Delta_i^2 \Phi(-\Delta_i) \\ &\leq 2p \sum_{i=1}^{k-1} \Delta_i \varphi(\Delta_i) \leq 2p(k-1)w, \end{aligned} \tag{16}$$

where $\Delta_i = (\theta_k - \theta_i)/(2p)^{1/2}, i = 1, \ldots, k-1$, and $w = \varphi(1)$. The first inequality is obvious, the second follows from the fact that $\Delta\Phi(-\Delta) < \varphi(\Delta)$ for $\Delta > 0$, and the third holds since the maximum of $\Delta\varphi(\Delta)$ on the positive real line occurs at $\Delta = 1$.

Thus, it follows that the Bayes risk due to selection of $\underline{d}^N$ tends to a finite limit as q tends to infinity. Finally, applying Schwarz' inequality to (16), the same is seen to hold under $A = A_1$. Actually, in this case one can verify that the limit is zero. This completes the proof of the theorem.

4. Independent Normal Populations With Unequal Sample Sizes

Let $X_{i1}, \ldots, X_{in_i}$ be a sample from $N(\theta_i, \sigma^2), i = 1, \ldots, k$, where $\sigma^2 > 0$ is known, and where not all of the sample sizes $n_1, \ldots, n_k$ are equal. The k samples are assumed to be independent. As before, sufficiency leads to considering the sample means $X_i = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}, i = 1, \ldots, k$, which have variances $p_i = n_i^{-1}\sigma^2, i = 1, \ldots, k$, respectively.

By various reasons mentioned before, as well as those discussed in Gupta and Miescke (1988), the 0-1 loss for selection, i.e. A_0, will not be considered any further. Moreover, to keep the analysis in reasonable size, we restrict ourselves in the sequel to the two most appealing and realistic loss combinations $\mathcal{L}_1$ and $\mathcal{L}_2$, given by (4). It should be pointed out that the risk function of every decision rule $\underline{d} = (s, \ell_s)$ is continuous in $\underline{\theta}$ under both, $\mathcal{L}_1$ and $\mathcal{L}_2$. Continuity of the risk due to selection under loss A_1 has been justified in Gupta and Miescke (1988), and the same arguments apply to A_2. Continuity of the risk due to estimation under loss B_1 and B_2 is well known. Thus, all proper Bayes rules derived in the sequel, as well as those considered before with any loss combination from (2) and (3), are admissible.

Since the analysis of Bayes rules $\underline{d}^* = (s^*, \ell^*_{s^*})$ under loss function $\mathcal{L}_1$ is easier to manage, let us deal with it first. The risk function of a decision rule $\underline{d} = (s, \ell_s)$ at

parameter point $\underline{\theta}$ is given by

$$R(\underline{\theta}, \underline{d}) = cE_{\underline{\theta}}(\theta_{[k]} - \theta_{s(\underline{X})}) + E_{\underline{\theta}}(|\theta_{s(\underline{X})} - \ell_{s(\underline{X})}(\underline{X})|). \tag{17}$$

In the present situation of unequal sample sizes $n_1, \ldots, n_k$, it is appropriate to consider, more generally, also non-exchangeable normal priors, $\Theta_i \sim N(\mu_i, q_i), i = 1, \ldots, k$. Independence of $\Theta_1, \ldots, \Theta_k$, however, will be kept as before. Thus, a posteriori, given $\underline{X} = \underline{x}, \Theta_1, \ldots, \Theta_k$ are independent, with $\Theta_i \sim N\left(\frac{q_i x_i + p_i \mu_i}{q_i + p_i}, \frac{p_i q_i}{q_i + p_i}\right), = 1, \ldots, k$, and marginally, $X_1, \ldots, X_k$ are independent with $X_i \sim N(\mu_i, p_i + q_i), i = 1, \ldots, k$.

By Lemma 1, the Bayes rule employs the estimator $\ell_i^*(\underline{x}) = (q_i x_i + p_i \mu_i)/(q_i + p_i)$ for $\theta_i, i = 1, \ldots, k$, and it remains to find $s^*(\underline{x})$. For any decision rule $\underline{d} = (s, \ell_s^*)$, the posterior risk at $\underline{X} = \underline{x}$ turns out to be the following for selection $s(\underline{x}) = i \in \{1, \ldots, k\}$.

$$c\left[E\{\Theta_{[k]} | \underline{X} = \underline{x}\} - \frac{q_i x_i + p_i \mu_i}{q_i + p_i}\right] + \left(\frac{2}{\pi} \frac{q_i p_i}{q_i + p_i}\right)^{1/2}. \tag{18}$$

Thus, the following is seen to hold.

Theorem 4. Under loss function $\mathcal{L}_1$ and the normal prior considered above, the Bayes rule $\underline{d}^* = (s^*, \ell_{s^*}^*)$ employs $\ell_i^*(\underline{x}) = (q_i x_i + p_i \mu_i)/(q_i + p_i), i = 1, \ldots, k$, and $s^*(\underline{x})$ maximizes $c\ell_i^*(\underline{x}) - [2q_i p_i/\pi(q_i + p_i)]^{1/2}, i = 1, \ldots, k$.

There are three special cases which deserve to be studied in more detail. They are as follows.

Case 1: Noninformative prior; or $q_i \to \infty, i = 1, \ldots, k$. In this case, $\ell_i^*(\underline{x}) = x_i, i = 1, \ldots, k$, and $s^*(\underline{x})$ maximizes $x_i - c^{-1}(2p_i/\pi)^{1/2}, i = 1, \ldots, k$.

Case 2: Prior variances proportional to sample variances; i.e. $q_i = \gamma p_i, i = 1, \ldots, k$, for some fixed $\gamma > 0$. In this case, $\ell_i^*(\underline{x}) = (\gamma x_i + \mu_i)/(\gamma + 1), i = 1, \ldots, k$, and $s^*(\underline{x})$ maximizes

$\ell_i^*(\underline{x}) - c^{-1}(2\gamma p_i/(\gamma+1)\pi)^{1/2}, i = 1, \ldots, k$. Especially, for $\mu_1 = \ldots = \mu_k = \mu$, say, $\ell_i^*(\underline{x}) = (\gamma x_i + \mu)/(\gamma+1), i = 1, \ldots, k$, and $s^*(\underline{x})$ maximizes $x_i - c^{-1}(2(\gamma+1)p_i/\gamma\pi)^{1/2}, i = 1, \ldots, k$.

Case 3: Posterior is decreasing in transposition (DT); i.e. $q_i^{-1} + p_i^{-1} = \tau^{-1}, i = 1, \ldots, k$, for some fixed $\tau > 0$. Here, the sum of prior precision and sampling precision is constant across the k populations. Such priors have been considered and justified in Gupta and Miescke (1988). The idea for applications is the following. If the normal priors are not exchangeable, a proper choice of sample sizes $n_1, \ldots, n_k$ in the planning of the experiment can lead, at least approximately, to a posterior which is (DT). This is highly desirable since in that situation usually quite simple Bayes rule are found. In the present case, $\ell_i^*(\underline{x}) = \tau(p_i^{-1}x_i + q_i^{-1}\mu_i), i = 1, \ldots, k$, and $s^*(\underline{x})$ maximizes $\ell_i^*(\underline{x}), i = 1, \ldots, k$. Especially, for $\mu_1 = \ldots = \mu_k = \mu$, say, $\ell_i^*(\underline{x}) = p_i^{-1}\tau(x_i - \mu) + \mu, i = 1, \ldots, k$, and $s^*(\underline{x})$ maximizes $p_i^{-1}(x_i - \mu), i = 1, \ldots, k$.

The decision rule considered last in Case 3 is of a very simple and appealing form. Due to the (DT)-property of the posterior, it can be seen to be Bayes rule under the large class of loss functions assumed in Theorem 1. Without further proof, the following can be stated.

Corollary 2. For the loss function $L = A + B$ in (2), assume that A is permutation symmetric and favors selection of larger θ-values, and that B is either B_1 or B_2 in (3). If the normal prior satisfies $q_i^{-1} + p_i^{-1} = \tau^{-1}, i = 1, \ldots, k$, for some $\tau > 0$, then the Bayes rule $\underline{d}^* = (s^*, \ell_{s^*}^*)$ is of the following form. $\ell_i^*(\underline{x}) = p_i^{-1}\tau(x_i - \mu_i) + \mu_i, i = 1, \ldots, k$, and $s^*(\underline{x})$ maximizes $\ell_i^*(\underline{x})$.

There is one interesting feature of the Bayes rule given by Theorem 4 which is worth

to be pointed out explicitly. Whenever for some $i \in \{1, \ldots, k\}, x_i$ turns out to be larger than μ_i, then a smaller (larger) p_i, i.e. a larger (smaller) n_i, works in favor of (against) population π_i to be selected. And for $x_i < \mu_i$, the reverse is seen to hold true for the rule derived in Case 3. On the other hand the rule of Case 1, as well as that one of Case 2 for $\mu_1 = \ldots = \mu_k$, have the property that at any x_i, a smaller (larger) p_i works in favor of (against) π_i being selected.

It is also interesting to note that, from a frequentist point of view, the decision rule in Case 1, which may be considered as a "natural rule" under unequal sample sizes, selects in terms of lower confidence bounds of $\theta_1, \ldots, \theta_k$ at a common fixed confidence level. Especially for the values $c = 0.485$ and $c = 0.343, x_i - c^{-1}(2p_i/\pi)^{1/2}$ is a lower confidence bound for θ_i with 95% and 99%, respectively, level of confidence. Similar can be said about the rule in Case 2 for $\mu_1 = \ldots = \mu_k$. Finally, the following can be shown.

Theorem 5. Under the loss function $\mathcal{L}_1$, the decision rule of Case 1 is extended Bayes rule.

Proof: This can be shown under Case 2 with $\mu_i = 0$, $i = 1, \ldots, k$, by letting γ tend to infinity. The Bayes posterior risk of the Bayes rule at $\underline{X} = \underline{x}$, in view of (18), is

$$(19) \qquad cE\{\Theta_{[k]}|\underline{X} = \underline{x}\} - \max_{i=1,\ldots,k}\{c\gamma(\gamma+1)^{-1}x_i - [2\gamma p_i/\pi(\gamma+1)]^{1/2}\}.$$

On the other hand, the posterior risk of the rule of Case 1 is, under Case 2,

$$(20) \qquad cE\{\Theta_{[k]}|\underline{X} = \underline{x}\} - \max_{i=1,\ldots,k}\{c\gamma(\gamma+1)^{-1}x_i - (2p_i/\pi)^{1/2}\}$$
$$+ E\{|\Theta_s - x_s||\underline{X} = \underline{x}\} - (2p_s/\pi)^{1/2},$$

where $s = s(\underline{x})$ is that index at which the maximum occurs. The difference of the maxima in

(19) and (20) is bounded by the maximum of the values $(2p_i/\pi)^{1/2} - (2\gamma p_i/\pi(\gamma+1))^{1/2}, i = 1,\ldots,k$, which does not depend on $\underline{x}$, and which tends to zero as γ tends to infinity. Furthermore, the last difference in (20) is bounded by

$$\sum_{j=1}^{k} [E\{|\Theta_j - x_j||\underline{X} = \underline{x}\} - (2p_j/\pi)^{1/2}]. \tag{21}$$

Finally, from the fact that for every $j = 1,\ldots,k$,

$$\begin{aligned} &E\{|\Theta_j - x_j||\underline{X} = \underline{x}\} \\ &= [\gamma p_j/(\gamma+1)]^{1/2} E(|N + [\gamma(\gamma+1)p_j]^{-1/2} x_j|), \end{aligned} \tag{22}$$

where $N \sim N(0,1)$ is an auxiliary random variable, and the fact that marginally, $[(\gamma + 1)p_j]^{-1/2} X_j \sim N(0,1), j = 1,\ldots,k$, it is seen that the integral of (21) with respect to the marginal density of $(X_1,\ldots,X_k)$ tends to zero as γ tends to infinity. To summarize, it has been shown that the integral of the difference of (20) and (19) with respect to the marginal density of $(X_1,\ldots,X_k)$ tends to zero as γ tends to infinity.

To justify the relevance of this result, it remains to be shown that the limit of the Bayes risks is finite. The posterior risk of the Bayes rule can be written in the form

$$\begin{aligned} &c[E\{\Theta_{[k]}|\underline{X} = \underline{x}\} - \gamma(\gamma+1)^{-1} x_{[k]}] \\ &+ \min_{i=1,\ldots,k} \{c\gamma(\gamma+1)^{-1}(x_{[k]} - x_i) + [2\gamma p_i/\pi(\gamma+1)]^{1/2}\}. \end{aligned} \tag{23}$$

It is now easy to see that the following provides an upper bound to (23),

$$\begin{aligned} &cE\left(\max_{i=1,\ldots,k} \left\{[\gamma p_i/(\gamma+1)]^{1/2} N_i\right\}\right) \\ &+ [2\gamma/\pi(\gamma+1)]^{1/2} \max_{i=1,\ldots,k} \left\{p_i^{1/2}\right\}, \end{aligned} \tag{24}$$

where $N_1, \ldots, N_k$ is a sample from $N(0,1)$. This bound does not depend on $\underline{x}$, and it tends to a finite limit as γ tends to infinity. Therefore, the proof of the theorem is completed.

To conclude this section, let us consider how Bayes rules $\underline{d}^* = (s^*, \ell^*_{s^*})$ look like under loss function $\mathcal{L}_2$. As mentioned already before, the analysis is more complicated than under $\mathcal{L}_1$. The risk function of a decision rule $\underline{d} = (s, \ell_s)$ at parameter point $\underline{\theta}$ is given by the following counterpart to (17).

$$R(\underline{\theta}, \underline{d}) = cE_{\underline{\theta}}([\theta_{[k]} - \theta_{s(\underline{X})}]^2) + E_{\underline{\theta}}([\theta_{s(\underline{X})} - \ell_{s(\underline{X})}(\underline{X})]^2). \tag{25}$$

Under the normal prior considered before, the posterior risk at $\underline{X} = \underline{x}$ for any decision $\underline{d} = (s, \ell^*_s)$ with $\ell^*_i(\underline{x}) = (q_i x_i + p_i \mu_i)/(q_i + p_i), i = 1, \ldots, k$, which is the estimate employed by the Bayes rule, turns out to be the following for selection $s(\underline{x}) = i \in \{1, \ldots, k\}$.

$$cE\{[\Theta_{[k]} - \Theta_i]^2 | \underline{X} = \underline{x}\} + \frac{q_i p_i}{q_i + p_i}, \tag{26}$$

which has to be minimized by $s^*(\underline{x})$ for $i = 1, \ldots, k$. What makes this task difficult is the fact that for any i, the conditional distribution of $(\Theta_{[k]}, \Theta_i)$ at $\underline{X} = \underline{x}$ does not allow for simpler representations of the conditional expectation in (26), which in most situations has to be evaluated on a computer.

At the end, let us see how much can be said about the Bayes rule under the three cases considered previously.

Case 1: Noninformative prior; or $q_i \to \infty, i = 1, \ldots, k$. In this case, $\ell^*_i(\underline{x}) = x_i, i = 1, \ldots, k$, and $s^*(\underline{x})$ minimizes

$$cE([\max_{j=1,\ldots,k}\{x_j - x_i + p_j^{1/2} N_j - p_i^{1/2} N_i\}]^2) + p_i, \tag{27}$$

where $N_1, \ldots, N_k$ is a random sample from $N(0,1)$.

Case 2: Prior variances proportional to sample variances; i.e. $q_i = \gamma p_i, i = 1, \ldots, k$. In this case, as before under $\mathcal{L}_1$, we have $\ell_i^*(\underline{x}) = (\gamma x_i + \mu_i)/(\gamma + 1), i = 1, \ldots, k$, but $s^*(\underline{x})$ minimizes now (26) with $q_i p_i/(q_i + p_i) = \gamma p_i/(\gamma + 1), i = 1, \ldots, k$.

Case 3: Prior is decreasing in transposition (DT); i.e. $q_i^{-1} + p_i^{-1} = \tau^{-1}, i = 1, \ldots, k$, for some $\tau > 0$. This case is covered by Corollary 2, and thus the Bayes rule is the same as that one in Case 3 under $\mathcal{L}_1$.

Acknowledgements

The authors are thankful to Professor S. Panchapakesan for some helpful discussions.

References

[1] Cohen, A. and Sackrowitz, H.B. (1982). Estimating the mean of the selected population. *Statistical Decision Theory and Related Topics - III*, Vol. 1, eds. S.S. Gupta and J.O. Berger, Academic Press, New York, 243–270.

[2] Cohen, A. and Sackrowitz, H.B. (1988). A decision theory formulation for population selection followed by estimating the mean of the selected population. *Statistical Decision Theory and Related Topics - IV*, Vol. 2, eds. S.S. Gupta and J.O. Berger, Springer Verlag, New York, 33–36.

[3] Dahiya, R.C. (1974). Estimation of the mean of the selected population. *Journal of the American Statistical Association*, **69**, 226–230.

[4] Gupta, S.S. and Miescke, K.J. (1984). Sequential selection procedures: A decision theoretic approach, *Annals of Statistics*, **12**, 336–350.

[5] Gupta, S.S. and Miescke, K.J. (1988). On the problem of finding the largest normal mean under heteroscedasticity. *Statistical Decision Theory and Related Topics - IV*, Vol. 2, eds. S.S. Gupta and J.O. Berger, Springer Verlag, New York, 37–49.

[6] Jeyaratnam, S. and Panchapakesan, S. (1984). An estimation problem relating to subset selection for normal populations. *Design of Experiments; Ranking and Selection*, eds. T.J. Santner and A.C. Tamhane, Marcel Dekker, New York, 287–302.

[7] Jeyaratnam, S. and Panchapakesan, S. (1986). Estimation after subset selection from exponential populations. *Communications in Statistics - Theory and Methods*, **10**, 1869–1878.

[8] Jeyaratnam, S. and Panchapakesan, S. (1988). Selections from uniform populations

based on sample midranges and an associated estimation after selection. *Communications in Statistics - Theory and Methods*, **17**, 2303–2314.

[9] Sackrowitz, H.B. and Samuel-Cahn, E. (1984). Estimation of the mean of a selected negative exponential population. *Journal of the Royal Statistical Society, Series B*, **46**, 242–249.

[10] Sackrowitz, H.B. and Samuel-Cahn, E. (1986). Evaluating the chosen population: A Bayes and minimax approach. *Adaptive Statistical Procedures and Related Topics*, ed. J. Van Ryzin, IMS Lecture Notes – Monograph Series, Vol. 8, 386–399.

[11] Sarkadi, K. (1967). Estimation after selection. *Studia Scientiarum Mathematicarum Hungarica*, **2**, 341–350.

[12] Vellaisamy, P. and Sharma, D. (1988). Estimation of the mean of the selected gamma population. *Communications in Statistics - Theory and Methods*, **17**, 2797–2817.

[13] Vellaisamy, P., Kumar, S., and Sharma, D. (1988). Estimating the mean of the selected uniform population. *Communications in Statistics - Theory and Methods*, **17**, 3447–3476.

[14] Vellaisamy, P. and Sharma, D. (1989). A note on the estimation of the mean of the selected gamma population. *Communications in Statistics - Theory and Methods*, **18**, 555–560.

[15] Venter, J.H. (1988). Estimation of the mean of the selected population. *Communications in Statistics - Theory and Methods*, **17**, 791–805.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release, distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Report #89-23C	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Purdue University 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Department of Statistics West Lafayette, IN 47907	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Office of Naval Research 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-88-K-0170, DMS-8606964 DMS-8702620
8c. ADDRESS (City, State, and ZIP Code) Arlington, VA 22217-5000	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
ON FINDING THE LARGEST NORMAL MEAN AND ESTIMATING THE SELECTED MEAN (Unclassified)

12. PERSONAL AUTHOR(S)
Shanti S. Gupta and Klaus J. Miescke

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	July 1989	22

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

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

For $k \geq 2$ independent normal populations with unknown means and a common known variance, the problem of selecting the population with the largest mean and simultaneously estimating the mean of the selected population is considered in the decision theoretic approach following Cohen and Sackrowitz (1988). Under several loss functions with two additive components due to selection and due to estimation, Bayes decision rules are derived and studied. Both, the case of equal sample sizes and the case of unequal sample sizes are treated. The "natural" rule, which selects in terms of the largest sample mean and then estimates based on the sample mean of the selected population, is critically examined in all situations considered.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Shanti S. Gupta	22b. TELEPHONE (Include Area Code) (317) 494-6031 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED