On a Bayesian Approach to Selecting the Best among Good Populations*

by

Shanti S. Gupta** and Joong K. Sohn
Purdue University

Technical Report #83-17

Department of Statistics
Purdue University

October 1983

(Revised)

*Research 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.

**Paper presented at the Second Valencia Conference on Bayesian Statistics, Spain, September 6-10, 1983.

On a Bayesian Approach to Selecting the Best among Good Populations

By Shanti S. Gupta and Joong K. Sohn

Purdue University, U.S.A.

SUMMARY

The problem of selecting the best among a set of populations (treatments) which are better than a control (or standard) using an elimination type two-stage procedure for the case of normal populations is studied. After retaining good populations based on a Bayes decision rule, at Stage 2 one takes additional samples from selected populations according to the stopping rule N_i which provides the $100(1-2\alpha)\%$ Highest Posterior Density (HPD) credible region with a common width 2d for each selected population. Then one decides on the choice of the best population based on the overall sample means. The proposed stopping rule N_i is shown to be asymptotically efficient.

Keywords: ELIMINATION TYPE PROCEDURES; CREDIBLE REGION; STOPPING RULE; ASYMPTOTIC EFFICIENCY

1. INTRODUCTION

Since the early work of Bechhofer, Dunnett and Sobel (1954) on the two-sample (two-stage) problem for selecting the normal population associated with the largest unknown mean from $k(\geq 2)$ normal populations, several different types of two-stage procedures have been studied for the case of known variances, common unknown variance and unknown and unequal variances. Among them, elimination type procedures which use subset selection approach at Stage 1 and use indifference zone approach at Stage 2 are important and frequently studied. Alam (1970) studied the known variances case and Tamhane and Bechhofer (1977, 1979), using a minimax criterion, also studied the known variances case. Gupta and Kim (1982) and Tamhane (1975) considered the common unknown variance case. Recently Gupta and Miescke (1981, 1982), among others, have studied the problem under a decision-theoretic Bayesian framework.

In this paper, we propose an elimination type procedure under the Bayesian setting, which retains good populations based on a Bayes decision rule for a certain loss function and a noninformative prior for unknown parameters. We also use a stopping rule to construct the $100(1-2\alpha)\%$ Highest Posterior Density (HPD) credible region with a common width 2d to decide on the selection of the best based on the overall sample means. The proposed stopping rule is shown to be asymptotically efficient.

2. PRELIMINARIES: NOTATIONS AND DEFINITIONS

Let $\pi_i(i = 1,2,\ldots,k)$ be $k(\geq 2)$ independent normal populations with unknown means θ_i and unknown variances $\sigma_i^2(0 < \sigma_i^2 < \infty)$. Also let X_i be the (observable) characteristic corresponding to π_i and let $(X_{i1},X_{i2},\ldots,X_{in})$ $(i = 1,2,\ldots,k)$ be n independent samples from π_i. We denote by x_{ij} a realization of X_{ij}. Assuming that very little is known about the prior of (θ_i,σ_i^2), we may use a noninformative prior density $\tau(\theta_i,\sigma_i^2)$, where

$$\tau(\theta_i,\sigma_i^2) = \sigma_i^{-2} I_{(0,\infty)}(\sigma_i^2), \tag{2.1}$$

and $I_A(x)$ is the usual indicator function. Here we denote by $\tau_1(\theta_i,\sigma_i^2|\underset{\sim}{x}_i)$ and $\tau_1(\theta_i|\underset{\sim}{x}_i)$ the posterior density of (θ_i,σ_i^2) and the marginal posterior density of θ_i, respectively, where $\underset{\sim}{x}_i = (x_{i1},x_{i2},\ldots,x_{in})$.

Remark: One can use a specific prior density if one has a reasonable amount of information about the prior. However, in many practical cases, there is very little knowledge about the prior and in those cases noninformative priors work fairly well and could provide robust solutions; (For further discussion, refer to Berger (1980, 1982)).

A population π_i is said to be 'good' ('bad') if $\theta_i \geq \theta_0$ $(\theta_i < \theta_0)$, where θ_0 refers to the value associated with the control population π_0 (in some problems θ_0 may be a constant specified a priori by the experimenter). Let our loss function be as follows:

$$L(\theta_i,a_p) = \begin{cases} 0 & \text{if } \theta_i \in \Theta_p \quad (p = 0,1), \\ k_p & \text{if } \theta_i \in \Theta-\Theta_p \ (p = 0,1), \end{cases}$$

where $\mathcal{A} = \{a_0,a_1\}$ is the action space and where $\Theta = R'$, $\Theta_0 = [\theta_0,\infty)$ and $\Theta_1 = \Theta-\Theta_0$. Here the action accepts π_i as a good population and the action a_1 rejects π_i as being a bad population. Also the definition of the $100(1-2\alpha)\%$ HPD credible region which we will use at Stage 2 is as follows.

<u>Definition</u> (see Berger (1980)). The $100(1-2\alpha)\%$ HPD credible region for θ_i is the subset $C_{(i,1-2\alpha)}$ of Θ of the form

$$C_{(i,1-2\alpha)} = \{\theta_i \in \Theta: \tau_1(\theta_i|\underset{\sim}{x}_i) \geq p(2\alpha)\}, \tag{2.3}$$

where $p(2\alpha)$ is the largest constant such that

$$Pr(C_{(i,1-2\alpha)}|\underset{\sim}{X} = x_i) \geq 1-2\alpha. \tag{2.4}$$

Note that the HPD credible region $C_{(i,1-2\alpha)}$ are intervals of the form (a_i,b_i) on $\mathbb{R}^1$ for this problem.

3. GOAL AND A PROPOSED PROCEDURE $R(\theta_0,\alpha,d)$

Assume that no knowledge is available concerning the correct pairing between populations and the ordered values of θ_i. Our goal is to select the population associated with the largest unknown mean among the subset of good populations. The procedure $R(\theta_0,\alpha,d)$ is designed to meet the goal.

3.1. Definition of the Procedure $R(\theta_0,\alpha,d)$

Stage 1. Take $n_0 = \max\{2,[Z_{(1-\alpha)}/d]+1\}$ observations from each population π_i, where 2d is the common width of the 100(1-2α)%HPD credible region and $Z_{(1-\alpha)}$ is a 100(1-α) upper percentile of the standard normal distribution. Then select a subset S by the following rule:

At Stage 1, retain π_i if and only if $Pr(\Theta_1|\underset{\sim}{x}_i) \leq k_1/(k_0+k_1)$, (3.1)

where $Pr(\Theta_1|\underset{\sim}{x}_i) = \int_{\Theta_1} \tau_1(\theta_i|\underset{\sim}{x}_i)d\theta_i$.

Note that the value of $Pr(\Theta_1|\underset{\sim}{x}_i)$ is evaluated explicitly in Corollary 1 of Section 3.2. Note also that the rule (3.1) is a Bayes rule. This follows from the fact that the expected posterior losses of actions a_0 and a_1 are $k_0Pr(\Theta_1|\underset{\sim}{x}_i)$ and $k_1Pr(\Theta_0|\underset{\sim}{x}_i)$, respectively and $Pr(\Theta_0|\underset{\sim}{x}_i)+Pr(\Theta_1|\underset{\sim}{x}_i) = 1$.

Let s be the size of the subset S. Then

(i) if s = 0, we decide that none of the populations are good and stop,

(ii) if $s = 1$, we decide that the population selected is the only good and the best at the same time and stop,

(iii) if $s \geq 2$, we proceed to Stage 2.

Stage 2. Take $N_i - n_0$ additional samples from each selected population π_i such that

$$N_i = \inf\{n_i: n_i \geq n_0 \text{ and } n_i \geq [(1/c-1)\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2/d^2]+1\}, \quad (3.2)$$

where c is the 100α lower percentile point of the beta distribution with parameters $\frac{1}{2}(n_i-1)$ and 1/2, and [a] is the largest integer less than or equal to a. The stopping rule N_i is set up to provide a $100(1-2\alpha)\%$ HPD credible region with a common width 2d for each selected population π_i. Then our final decision at Stage 2 is that the population with the largest overall sample mean is the best.

3.2. Discussion of the Procedure $R(\theta_0,\alpha,d)$

It is easily seen that at Stage 1, $\tau_1(\theta_i|\underset{\sim}{x}_i)$ is a Student's t-distribution density with (n_0-1) degrees of freedom, the location parameter $\bar{x}_i = \sum_j x_{ij}/n_0$, and the scale parameter $\sum_j(x_{ij}-\bar{x}_i)^2/\{n_0(n_0-1)\}$.

Lemma 1. For a random variable T having a Student's t-distribution with (m-1) degrees of freedom, the location parameter $\bar{x}_i$, and the scale parameter S_i^2,

$$\Pr(T \leq t_0) = \begin{cases} 1-\frac{1}{2}\, I_u\{\frac{1}{2}(m-1),\frac{1}{2}\} & \text{if } t_0-\bar{x}_i \geq 0, \\ \\ \frac{1}{2}\, I_u\{\frac{1}{2}(m-1),\frac{1}{2}\} & \text{if } t_0-\bar{x}_i < 0, \end{cases} \quad (3.3)$$

where $I_x(a,b)$ is an incomplete beta function with parameters a and b, $u = (m-1)/(m-1+t^2)$ and $t = (t_0-\bar{x}_i)/\ S_i$.

Corollary 1. From Lemma 1, $Pr(\Theta_1|\underset{\sim}{x}_i)$ can be obtained by substituting for m, $\bar{x}_i$ and S_i^2 by n_0, $\bar{x}_i = \sum_j x_{ij}/n_0$ and $\sum_j(x_{ij}-\bar{x}_i)/\{n_0(n_0-1)\}$, respectively.

Theorem 1. Let $C_{(i,1-2\alpha)} = (a_i,b_i)$. The the stopping rule N_i provides the $100(1-2\alpha)\%$ HPD credible region with a common width 2d for each selected population π_i.

Proof. Denote by n_i the overall sample size for selected population π_i, the following two equations provide the $100(1-2\alpha)\%$ HPD credible region $C_{(i,1-2\alpha)}$:

$$\tau_1(a_i|\underset{\sim}{x}_i) = \tau_1(b_i|x_i) \tag{3.4}$$

and

$$\int_{a_i}^{b_i} \tau_1(\theta_i|\underset{\sim}{x}_i)d\theta_i = 1-2\alpha. \tag{3.5}$$

After we stop sampling at Stage 2, $\tau_1(\theta_i|\underset{\sim}{x}_i)$ is still a Student's t-distribution. Hence $\tau_1(\theta_i|\underset{\sim}{x}_i)$ is unimodal and symmetric about the location parameter $\bar{x}_i = \sum_j x_{ij}/n_i$.

Therefore by Lemma 1, a_i and b_i of the credible region $C_{(i,1-2\alpha)}$ are given by

$$a_i = \bar{x}_i - \{(\frac{1}{c}-1)(\frac{\sum_j(x_{ij}-\bar{x}_i)^2}{n_i})\}^{\frac{1}{2}} \tag{3.6}$$

and

$$b_i = \bar{x}_i + \{(\frac{1}{c}-1)(\frac{\sum_j(x_{ij}-\bar{x}_i)^2}{n_i})\}^{\frac{1}{2}}. \tag{3.7}$$

Thus the width 2d of the credible region $C_{(i,1-2\alpha)}$ is

$$2d = 2\{(\frac{1}{c} - 1)(\frac{\sum_j (x_{ij}-\bar{x}_i)^2}{n_i})\}^{\frac{1}{2}} \tag{3.8}$$

and this implies that

$$n_i = \frac{(1/c-1)\sum_j (x_{ij}-\bar{x}_i)^2}{d^2}. \tag{3.9}$$

This completes the proof of the theorem.

Lemma 2. For c as defined in Stage 2, $\{(1/c-1)(n_i-1)\}^{\frac{1}{2}} \to Z_{(1-\alpha)}$ as $n_i \to \infty$.

Proof. The proof follows from Lemma 1 and the central limit theorem.

Finally, we want to show that the proposed stopping rule N_i is asymptotically efficient.

Theorem 2. Let $n = \sigma_i^2 Z_{(1-\alpha)}^2/d^2$. Then for a fixed $\sigma_i^2 (0 < \sigma_i^2 < \infty)$ and the stopping rule N_i,

(a) $N_i/n \to 1$ a.s. as $d \to 0$

and

(b) $\lim_{d\to 0} E(N_i/n) = 1$ (asymptotic efficiency).

Proof. From the definitions of n_0 and N_i, one gets the following inequalities;

$$\frac{(1/c-1)(N_i-1)S_i^2}{d^2} \le N_i \le \frac{(1/c-1)(N_i-1)S_i^2}{d^2} + \frac{Z_{(1-\alpha)}}{d} + 4, \tag{3.10}$$

where $S_i^2 = \sum_j (x_{ij}-\bar{x}_i)^2/(N_i-1)$ and $\bar{x}_i = \sum_j x_{ij}/N_i$. Since $n_0 \to \infty$ and $N_i \to \infty$ as $d \to 0$ hence $S_i^2 \to \sigma_i^2$ a.s.; Using Lemma 2, one gets the results (a) and (b).

ACKNOWLEDGEMENT

This research was supported by the United States Navy through the Office of Naval Research Contract N00014-75-C-0455 at Purdue University, W. Lafayette, Indiana, U.S.A.

REFERENCES

ALAM, K. (1970). A two-sample procedure for selecting the population with the population with the largest mean from k normal populations. Ann. Inst. Statist. Math. 22, 127-136.

BECHHOFER, R. E., DUNNETT, C. W., and SOBEL, M. (1954). A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. Biometrika 41, 170-176.

BERGER, J. O. (1980). Statistical Decision Theory: Foundations, Concepts, and Methods. Springer-Verlag, New York.

BERGER, J. O. (1982). Robust Bayesian viewpoint. Technical Report 82-9, Dept. of Statist., Purdue University, W. Lafayette, IN.

GUPTA, S. S. and KIM, W. C. (1982). A two-stage elimination type procedure for selecting the largest of several normal means with a common unknown variance, Technical Report 82-27, Dept. of Statist., Purdue University, W. Lafayette, IN.

GUPTA, S. S. and MIESCKE, K. J. (1981). An essentially complete class of two-stage procedures with screening at the first stage. Mimeo. Ser. No. 81-53, Dept. of Statist., Purdue University, W. Lafayette, IN. To appear in Statistics and Decisions.

GUPTA, S. S. and MIESCKE, K. J. (1982). On the problem of finding a best population with respect to a control in two stages. Statistical Decision Theory and Related Topics III (S. S. Gupta and J. O. Berger eds.), Vol. I, Academic Press, New York, 473-496.

TAMHANE, A. C. (1975). On minimax multistage elimination type rules for selecting the largest normal mean. Technical Report 259, Dept. of Operations Research, Cornell University, Ithaca, New York.

TAMHANE, A. C. and BECHHOFER, R. E. (1977). A two-stage minimax procedure with screening for selecting the largest normal mean. Commun. Statist.-Theor. Meth. A6, 1003-1033.

TAMHANE, A. C. 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 #83-17	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): ON A BAYESIAN APPROACH TO SELECTING THE BEST AMONG GOOD POPULATIONS	5. TYPE OF REPORT & PERIOD COVERED: Technical
	6. PERFORMING ORG. REPORT NUMBER: Technical Report #83-17
7. AUTHOR(s): Shanti S. Gupta and Joong K. Sohn	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0455
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: (Revised October 1983)
	13. NUMBER OF PAGES: 8
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)

Elimination type procedures; Credible region; Stopping rule; Asymptotic efficiency.

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

The problem of selecting the best among a set of populations (treatments) which are better than a control (or standard) using an elimination type two-stage procedure for the case of normal populations is studied. After retaining good populations based on a Bayes decision rule, at Stage 2 one takes additional samples from selected populations according to the stopping rule N_i which provides the $100(1-2\alpha)\%$ Highest Posterior Density (HPD) credible region with a common width 2d for each selected population. Then one decides on the choice of the best population based on the overall sample means. The proposed stopping rule N_i is shown to be asymptotically efficient.

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)