ON SELECTION OF POPULATIONS CLOSE TO A CONTROL OR STANDARD*

Shanti S. Gupta and Ashok K. Singh
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #508

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

ON SELECTION OF POPULATIONS CLOSE TO A CONTROL OR STANDARD

Shanti S. Gupta and Ashok K. Singh

Purdue University

ABSTRACT

This paper discusses the following problem of selection: Given a set of k ($k \geq 2$) populations, select a subset which contains all populations 'close' to a given control population. A Bayes rule for the case of normal populations and Gupta-type rules for normal and gamma populations are investigated. Applications to problems involving tolerance regions in quality control are described.

1. INTRODUCTION

Let $\Pi_0, \Pi_1, \ldots, \Pi_k$ be $k+1$ independent populations with densities $f(x,\theta_0)$, $f(x,\theta_1), \ldots, f(x,\theta_k)$, respectively, where $\theta_i \in \Theta \subset \mathbb{R}$, the real line, and Π_0 is the control population. Let $E = [a(\theta_0), b(\theta_0)]$ be a given interval in $\mathbb{R}$. The subset of populations $\Pi_1, \ldots, \Pi_k$ with parameters in E is of interest in many practical situations.

The choice of E of course depends on the problem in hand. For instance, the problem of selection of all populations better than Π_0 corresponds to $E = (\theta_0,\infty)$. Gupta and Sobel (1958) have considered this problem for normal, gamma and binomial populations and have investigated procedures for selecting a subset which contains all populations better than a standard with probability at least P^*, where $0 < P^* < 1$ is a preassigned constant. Huang (1975) has derived a Bayes rule for partitioning a set of $k(k\geq 2)$ normal populations with respect to control. In the present paper we have considered the case $E = [a(\theta_0),b(\theta_0)]$, where a and b are known functions of θ_0, $-\infty < a(\theta_0) < b(\theta_0) < \infty$. We give two examples to show how the above problem arises in practice. Both of these examples are adapted from Burr (1976).

In the first example, a diesel engine plant had to make plunger rods for forcing fuel through small holes. The diameter of the plunger rods was to meet certain specification limits. In this situation, given several plunger rods, one may wish to select a subset of the given rods which meets the specification limits.

For the second example, suppose bearings and shafts are being produced for assembly. In this situation it is important to insure that the shafts will be capable of assembly at random into a bearing and hence the diametral clearance, which is the difference between the inside diameter of bearing and the outside diameter of shaft, should be within some specification limits. Here one may be interested in the subset of pairs of bearings and shafts for which diametral clearance meets the specification limits.

In Section 2 the problem of selecting a subset containing all normal populations with means in $[A+\theta_0,B+\theta_0]$ has been investigated from Bayes approach. In Section 3 the problem has been considered in the subset selection framework of Gupta (1965). For normal and gamma populations, procedures have been proposed and investigated which select a subset containing all populations 'close' to control with probability of correct selection at least $P^*(0 < P^* \leq 1)$.

2. A BAYES RULE FOR SELECTING ALL NORMAL POPULATIONS 'CLOSE' TO CONTROL

Let $\Pi_0, \Pi_1, \ldots, \Pi_k$ be k+1 independent normal populations with means $\theta_0, \theta_1, \ldots, \theta_k$, respectively, ($\theta_i \in \Theta \subset \mathbb{R}$, $i=0,1,\ldots,k, k \geq 2$) and a common known variance σ^2. We will say that Π_i is 'close' to Π_0 if $A + \theta_0 \leq \theta_i \leq B + \theta_0$, where A and B ($-\infty < A < B < \infty$) are given constants. The goal is to select a subset containing all populations close to Π_0.

The action space $\mathcal{A}$ consists of all subsets S of $\{1,\ldots,k\}$, including the null set. Assume that the loss function $L: \Theta^{k\times 1} \times \mathcal{A} \to \mathbb{R}$ is given by

$$L(\underline{\theta}, S) = \sum_{i \notin S} I_{[A+\theta_0, B+\theta_0]}(\theta_i) + \sum_{i \in S} I_{[A+\theta_0, B+\theta_0]^C}(\theta_i) \quad (2.1)$$

where $I_D(\cdot)$ is the indicator function of a set $D \subset \{1,\ldots,k\}$, and $D^C = \{1,\ldots,k\} \sim D$.

The mean θ_0 of the control population Π_0 may or may not be known. We consider the two cases separately.

(1) θ_0 known.

Let $\overline{x}_i$ be the mean of n independent observations from $\Pi_i (i=1,\ldots,k)$. Suppose the a priori distribution $g(\underline{\theta})$ of $\underline{\theta} = (\theta_1,\ldots,\theta_k)$ is given by $g(\underline{\theta}) = \prod_{i=1}^{k} [\tau^{-1}\phi(\tau^{-1}(\theta_i - \mu))]$, where $(\cdot)$ is the density function of a standard normal random variable. Then the posterior distribution of $\underline{\theta}$ given the observations is

$$g^*(\underline{\theta}|\overline{x}_1,\ldots,\overline{x}\ ) = \prod_{i=1}^{k} [v\text{-}1\ \phi(v^{-1}(\theta_i - m_i))]$$

where

$$m_i = \frac{(\mu\sigma^2/n) + \overline{x}_i\tau^2}{(\sigma^2/n) + \tau^2}, \quad i=1,\ldots,k$$

$$v = \left[\frac{\sigma^2\gamma^2/n}{(\sigma^2/n) + \tau^2}\right]^{\frac{1}{2}}$$

Let $a = A+ {}_0$, $b = B+ {}_0$, and let $(\cdot)$ denote the cumulative distribution function of a standard normal random variable.

Set $S_1 = \{i: \ a \leq m_i \leq b, \ \Psi(m_i) \geq \frac{1}{2}$ (2.2)

where

$$\Psi(y) = \Phi(\frac{b-y}{v}) - \Phi(\frac{a-y}{v}) \tag{2.3}$$

Let d_1 be the rule which selects S_1 with probability one. We show that the rule d_1 is a Bayes rule for the problem. As only Bayes rules are being considered, it is sufficient to show that

$$r(d_1,\underline{x}) \leq r(d,\underline{x}) \text{ for any nonrandomized rule } d \tag{2.4}$$

where $r(d,\underline{x})$ is the Bayes posterior risk in using the rule d. Let d select a subset $S \in \mathcal{U}$ with probability one. Then

$$r(d,\underline{x}) = \int\ldots\int L(\underline{\theta},S)g^*(\theta_1, \quad ,\theta_k|\overline{x}_1,\ldots,\overline{x}_k)d\theta_1\ldots d\theta_k$$

$$= \sum_{i\notin S} \Psi(m_i) + \sum_{i\in S} [1-\Psi(m_i)] \tag{2.5}$$

We show that, if $S \neq S_1$, then $r(d,\underline{x}) \geq r(d_1,\underline{x})$. The following cases need to be considered:

(i) There exist i and j $(1 \leq i,j \leq k)$ such that

$i \in S^c \cap S_1$, $j \in S \cap S_1^c$

Let $S' = (j,i)S$, where $(j,i)S$ denotes the subset obtained from S by replacing j by i. Letting d' denote the rule which selects the set S' with probability one, we have

$$r(d,\underline{x})-r(d',\underline{x}) = 2(\Psi(m_i)-\Psi(m_j))$$

where $\Psi(\cdot)$ is given by (2.3).

It is easy to see that the function $\Psi(y)$ is symmetric about $\frac{a+b}{2}$ and strictly decreases with $|y - \frac{a+b}{2}|$. Then by (2.2) we have

$m_i \in [a,b]$ and $\Psi(m_i) \geq \frac{1}{2}$

$m_j \notin [a,b]$ or $\Psi(m_j) < \frac{1}{2}$

If $m_j \notin [a,b]$, then $|m_i - \frac{a+b}{2}| < |m_j - \frac{a+b}{2}|$

and hence

$$\Psi(m_i) > \Psi(m_j).$$

It is clear then that $r(d,\underline{x}) \geq r(d',\underline{x})$. Since S_1 can be obtained from S by the operation used above, the inequality (2.4) follows.

(ii) $S \subset S_1$

It is easily seen from (2.5) that

$$r(d,\underline{x})-r(d_1,\underline{x}) = \sum_{i\varepsilon S^c \cap S_1} [2\Psi(m_i)-1] \geq 0$$

by (2.2).

(iii) $S \supset S_1$

In this case we have

$$r(d,\underline{x})-r(d_1,\underline{x}) = \sum_{i\varepsilon S \cap S_1^c} [1-2\Psi(m_i)] \geq 0.$$

It follows that d_1 is a Bayes rule for the problem.

(2) $\underline{\theta_0 \text{ unknown.}}$

Here we are given sample means $\bar{x}_i$ from all k+1 normal populations $\Pi_i (i=0,1,\ldots,k)$. Suppose that the a priori distribution of $(\theta_0,\theta_1,\ldots,\theta_n)$ is

$$g(\theta_0,\theta_1,\ldots,\theta_n) = \prod_{i=0}^{k} [\tau^{-1}\phi(\tau^{-1}(\theta_i-\mu))] \tag{2.6}$$

Then the peterior distribution of $(\theta_0,\theta_1,\ldots,\theta_k)$ is

$$g^*(\theta_0,\theta_1,\ldots,\theta_k\ \bar{x}_0,\bar{x}_1,\ldots,\bar{x}_k) = \prod_{i=0}^{k} [v^{-1}\phi(v^{-1}(\theta_i-m_i))] \tag{2.7}$$

where $m_i (i=0,1,\ldots,k)$ and v are given by (2.2).

Let $S_2 = \{i:\ A \leq m_i-m_0 \leq B,\ \Psi^*(m_i-m_0) \geq \frac{1}{2}\}$ (2.8)

where Ψ^* is obtained from Ψ by replacing v by $v\sqrt{2}$. Also, let d_2 be the rule which selects the subset S_2 with probability one. Then using the fact that the posterior distribution of $\theta_i-\theta_0$ is normal with mean m_i-m_0 and variance $2v^2$ we can show, as in case (1), that d_2 is a Bayes rule for the problem.

3. SELECTION OF ALL THE POPULATIONS CLOSE TO CONTROL FROM SUBSET SELECTION APPROACH

In this section we investigate procedures to relect a subset which contains all populations Π_i which are 'close' to Π_0 with probability at least P^*, where $P^* \varepsilon (0,1)$ is a preassigned constant. Gupta and Sobel (1958) have used this approach for the problem of selecting a subset containing all populations better than a standard.

1. Location parameter - normal populations with common known variance.

Let Π_i be normal with mean θ_i and variance $\sigma^2 (i=0,1,\ldots,k)$. The mean θ_0 of the control population Π_0 may or may not be known. The two cases will be considered separately.

Let $E = [\theta_0 - a, \theta_0 + a]$, where $a > 0$ is a given constant.

Case A. θ_0 known.

A sample of size n_i is taken from $\Pi_i (i=1,\ldots,k)$. For selecting a subset containing all populations with means in $[\theta_0 - A, \theta_0 + A]$, consider the following rule:

$$R_A: \text{ select } \Pi_i \text{ iff } \theta_0 - a - \frac{d\sigma}{\sqrt{n_i}} \le \bar{x}_i \le \theta_0 + a + \frac{d\sigma}{\sqrt{n_i}} \qquad (3.1)$$

where the constant $d \geq 0$ is chosen to satisfy

$$P(CS|R_A) \geq P^*, \quad 0 < P^* < 1.$$

Here CS stands for correct selection, i.e., the selection of all Π_i with $|\theta_i - \theta_0| \leq a$. Let k_1 and k_2 denote the true number of populations with $|\theta_i - \theta_0| \leq a$ and $|\theta_i - \theta_0| > a$, respectively, so that $k_1 + k_2 = k$. If we let primes refer to values associated with the k_1 populations with $|\theta_i - \theta_0| \leq a$, then

$$P(CS|R_A) = \prod_{i=1}^{k_1} P\left(\theta_0 - a - \frac{\sigma d}{\sqrt{n_i'}} \le \bar{x}_i' \le \theta_0 + a + \frac{\sigma d}{\sqrt{n_i'}}\right)$$

$$= \prod_{i=1}^{k_1} \left[\Phi\left(\frac{(\theta_0 - \theta_i')\sqrt{n_i'}}{\sigma} + \frac{a\sqrt{n_i'}}{\sigma} + d\right) - L_i\right] \qquad (3.2)$$

where

$$L = \Phi\left(\frac{(\theta_0-\theta_i')\sqrt{n_i'}}{\sigma} - \frac{a\sqrt{n_i'}}{\sigma} - d\right).$$

Now consider the function

$$h(u) = \Phi\left(\frac{(\theta_0-u)\sqrt{n_i'}}{\sigma} + \frac{a\sqrt{n_i'}}{\sigma} + d\right) - \Phi\left(\frac{(\theta_0-u)\sqrt{n_i'}}{\sigma} - \frac{a\sqrt{n_i'}}{\sigma} - d\right) \quad (3.3)$$

It is easily verified that the function h(u) is symmetric about θ_0, and is increasing (decreasing) if $u < \theta_0 (u > \theta_0)$. It follows that

$$\inf_{|u-\theta_0|\leq a} h(u) = h(\theta_0-a) = h(\theta_0+a)$$

Hence

$$\inf_{|\theta_i'-\theta_0|\leq a} P(CS|R_A) = \prod_{i=1}^{k_1} \left[\Phi\left(\frac{2a\sqrt{n_i'}}{\sigma} + d\right) - \Phi(-d)\right] \quad (3.4)$$

If k_1 is known, then the constant d is obtained by equating the right hand side of (3.4) to P^*. In many situations k_1 is not known and a lower bound for $P(CS|R_A)$ can be obtained by setting $k_1=k$. Then the equation for d is given by

$$\prod_{i=1}^{k} \left[\Phi\left(\frac{2a\sqrt{n_i}}{\sigma} + d\right) - \Phi(-d)\right] = P^*$$

For unequal sample sizes, computation of d is difficult. If $n_i=n$ for all $i=1,\ldots,k$, the equation for d becomes

$$\Phi\left(\frac{2a\sqrt{n}}{\sigma} + d\right) + \Phi(d) - 1 = (P^*)^{1/k} \quad (3.5)$$

For selected values of k, P^* and $\frac{a\sqrt{n}}{\sigma}$, d-values satisfying (3.5) have been computed and are given in Tables I and II.

TABLE I

Values of d satisfying (3.5) for $P^* = .90$

k \ $\frac{a\sqrt{n}}{\sigma}$	.05	.1	.2	.4	.5
2	1.91	1.85	1.79	1.70	1.69
3	2.06	2.02	1.95	1.88	1.84
4	2.18	2.13	2.07	1.99	1.97
5	2.28	2.22	2.16	2.08	2.06

TABLE II

Values of d satisfying (3.5) for $P^* = .95$

$a\sqrt{n}/\sigma$ k	.05	.1	.2	.4	.5
2	2.19	2.15	2.08	2.00	1.98
3	2.34	2.30	2.23	2.16	2.15
4	2.45	2.40	2.34	2.27	2.25
5	2.54	2.50	2.44	2.37	2.35

Expected subset size for R_A

The size of the subset selected by the procedure R_A is a random variable which can take values 0,1,...,k. Gupta (1965) has proposed the expected size of the selected subset as a measure of performance for a selection rule. We have,

$$E(S|R_A) = \sum_{i=1}^{k} P\{\Pi_i \text{ is selected in the subset}\}$$

$$= \sum_{i=1}^{k} [\Phi(\frac{(\theta_0-\theta_i)\sqrt{n_i}}{\sigma} + \frac{a\sqrt{n_i}}{\sigma} + d) - \Phi(\frac{(\theta_0-\theta_i)\sqrt{n_i}}{\sigma} - \frac{a\sqrt{n_i}}{\sigma} - d)] \quad (3.6)$$

Specific Example

Given five normal population $\Pi_i (i=1,...,5)$ with unknown means and a common variance 1, we wish to select all the populations which are 'close' to a standard normal control population with a = .1. Observe that the problem is equivalent to selection of all populations with means in the internal [-.1,.1]. Using a program for generating normal random variables with means $\theta_1 = -.1$, $\theta_2 = .25$, $\theta_3 = -.40$, $\theta_4 = .15$, $\theta_5 = .50$ and variances 1, the following sample means based on n=25 were observed:

$$\bar{x}_1 = -.225, \quad \bar{x}_2 = .278, \quad \bar{x}_3 = -.582, \quad \bar{x}_4 = .246, \quad \bar{x}_5 = .705$$

Here $a\sqrt{n}/\sigma = .5$. For $P^* = .90$, we have, from Table I, $d = d(k, a\sqrt{n}/\sigma, P^*) = 2.06$ and hence the rule R_A selects all populations with means in [-.51, .51]. Thus Π_1, Π_2 and Π_4 are selected in the subset. It can be seen from Table II that for $P^* = .95$, the same three populations get selected.

PROBABILITY THAT EXACTLY ONE POPULATION IS SELECTED

Assume that the k unknown parameters are $\theta,\ldots,\theta,\theta_1$, where θ and θ_1 satisfy $|\theta_1-\theta_0| \leq a < |\theta-\theta_0|$. In this configuration it is meaningful to compute the probability that the rule R_A selects exactly one population. We will consider only the equal sample size case.

We have

P{Rule R_A selects exactly one population}

$$= \sum_{i=1}^{k} P\{\bar{x}_i \in [\theta_0 - a - \frac{d\sigma}{\sqrt{n}}, \theta_0 + a + \frac{d\sigma}{\sqrt{n}}], \bar{x}_j \notin [\theta_0 - a - \frac{d\sigma}{\sqrt{n}}, \theta_0 + a + \frac{d\sigma}{\sqrt{n}}]\}$$

$$= [\Phi(\frac{(\theta_0-\theta_1+a)\sqrt{n}}{\sigma} + d) - \Phi(\frac{(\theta_0-\theta_1-a)\sqrt{n}}{\sigma} - d)]$$

$$\cdot [1-\Phi(\frac{(\theta_0-\theta+a)\sqrt{n}}{\sigma} + d) + \Phi(\frac{(\theta_0-\theta-a)\sqrt{n}}{\sigma} - d)]^{k-1}$$

$$+ (k-1)[\Phi(\frac{(\theta_0-\theta+a)\sqrt{n}}{\sigma} + d) - \Phi(\frac{(\theta_0-\theta-a)\sqrt{n}}{\sigma} - d)]$$

$$\cdot [1-\Phi(\frac{(\theta_0-\theta+a)\sqrt{n}}{\sigma} + d) + \Phi(\frac{(\theta_0-\theta-a)\sqrt{n}}{\sigma} - d)]^{k-2}.$$

$$\cdot [1-\Phi(\frac{(\theta_0-\theta_1+a)\sqrt{n}}{\sigma} + d) + \Phi(\frac{(\theta_0-\theta+a)\sqrt{n}}{\sigma})] \qquad (3.7)$$

It should be noted that the constant d in this case is obtained by taking $k_1=1$ in (3.4) and equating the resulting expression to P^*.

For selected values of k, P^*, $a\sqrt{n}/\sigma$, $(\theta_0-\theta)\sqrt{n}/\sigma$ and $(\theta_0-\theta_1)\sqrt{n}/\sigma$ the expected subset size given by (3.6) and the probability of selecting exactly one population given by (3.7) have been computed. These values are shown in Tables III and IV. For example, if $P^* = .90$, $k=3$, $a\sqrt{n}/\sigma = .4$, $(\theta_0-\theta_1)\sqrt{n}/\sigma = 0.32$ and $(\theta_0-\theta)\sqrt{n}/\sigma = 3.45$ then the expected size of the selected subset is 1.21 and the probability that only one population is selected is .76. It appears from Tables III and IV that the expected subset size and the probability of selecting only one population do not change significantly as θ_1, the mean of the population close to control, varies inside $[\theta_0-a,\theta_0+a]$.

Remarks:

(i) It can easily be seen from expression (3.7) that if a and $|\theta-\theta_0|$ are large and $|\theta_1-\theta_0|$ is small then the probability of selecting exactly one population is close to 1. It should be observed that the probability of selecting the population 'close' to control, i.e., the population with mean θ_1, is at least P^*.

(ii) It is also clear from (3.7) that the probability of selecting exactly one population approaches unify as $n\to\infty$.

TABLE III

For $P^* = .90$ This Table Shows the Values of the Expected Subset Size (top entry) and the Probability of Selecting Exactly One Population (bottom entry) When the Unknown Means of the k Normal Populations are $\theta,\ldots,\theta,\theta_1$, Where $|\theta_0-\theta_1| \leq a < |\theta_0-\theta|$.

	$a\sqrt{n}/\sigma$	.1			.2			.4	
	$(\theta_0-\theta_1)\sqrt{n}/\sigma$ →	0	0.08		0	.16		0	.32
k	$(\theta_0-\theta)\sqrt{n}/\sigma$			$(\theta_0-\theta)\sqrt{n}/\sigma$			$(\theta_0-\theta)\sqrt{n}/\sigma$		
2	1.15	1.74 .24	1.74 .24	1.25	1.72 .26	1.72 .26	1.45	1.71 .28	1.70 .28
	3.15	1.06 .85	1.06 .85	3.25	1.06 .06	1.05 .86	3.45	1.05 .88	1.04 .87
	5.15	.95 .95	.95 .95	5.25	.95 .95	.95 .95	5.45	.96 .96	.96 .95
3	1.15	2.63 .04	2.63 .04	1.25	2.60 .04	2.60 .04	1.45	2.57 .05	2.56 .05
	3.15	1.27 .70	1.27 .70	3.25	1.24 .73	1.24 .73	3.45	1.22 .76	1.21 .76
	5.15	.97 .96	.97 .96	5.25	.97 .97	.97 .97	5.45	.98 .98	.97 .97
4	1.15	3.55 .01	3.55 .01	1.25	3.51 .01	3.51 .01	1.45	3.46 .01	3.46 .01
	3.15	1.51 .55	1.51 .55	3.25	1.47 .58	1.47 .58	3.45	1.42 .62	1.41 .62
	5.15	.98 .97	.98 .97	5.25	.98 .97	.98 .97	5.45	.99 .98	.98 .97
5	1.15	4.49 .00	4.49 .00	1.25	4.45 .00	4.45 .00	1.45	4.38 .00	4.38 .00
	3.15	1.79 .40	1.79 .40	3.25	1.73 .44	1.73 .44	3.45	1.65 .48	1.65 .48
	5.15	.99 .97	.99 .97	5.25	.99 .97	.99 .97	5.45	.99 .98	.99 .98

TABLE IV

For $P^* = .95$ This Table Shows the Values of the Expected Subset Size (top entry) and the Probability of Selecting Exactly One Population (bottom entry) When the Unknown Means of the k Normal Populations are $\theta,\ldots\theta,\theta_1$, Where $|\theta_0-\theta_1| \le a < |\theta_0-\theta|$.

$a\sqrt{n}/\sigma$	.1			.2			.4		
$(\theta_0-\theta_1)\sqrt{n}/\sigma$		0	0.08		0			0	.32
k	$(\theta_0-\theta)\sqrt{n}/\sigma$			$(\theta_0-\theta)\sqrt{n}/\sigma$			$(\theta_0-\theta)\sqrt{n}/\sigma$		
2	1.15	1.04 .15	1.84 .15	1.25	1.83 .17	1.82 .17	1.45	1.81 .18	1.81 .19
	3.15	1.16 .80	1.16 .80	3.25	1.14 .82	1.14 .82	3.45	1.13 .84	1.12 .84
	5.15	.98 .97	.98 .97	5.25	.98 .98	.98 .97	5.45	.98 .98	.98 .98
3	1.15	2.77 .01	2.77 .01	1.25	2.75 .02	2.75 .02	1.45	2.72 .02	2.72 .02
	3.15	1.44 .59	1.44 .59	3.25	1.40 .63	1.40 .63	3.45	1.36 .66	1.36 .66
	5.15	.99 .98	.99 .98	5.25	.99 .98	.99 .98	5.45	.99 .99	.99 .98
4	1.15	3.72 .00	3.72 .00	1.25	3.69 .00	3.69 .00	1.45	3.66 .00	3.66 .00
	3.15	1.76 .41	1.76 .41	3.25	1.71 .44	1.70 .44	3.45	1.65 .48	1.64 .48
	5.15	1.00 .98	1.00 .98	5.25	1.00 .98	1.00 .98	5.45	1.00 .98	1.00 .98
5	1.15	4.70 .00	4.70 .00	1.25	4.66 .00	4.66 .00	1.45	4.62 .00	4.62 .00
	3.15	2.16 .25	2.16 .25	3.25	2.08 .28	2.07 .28	3.45	1.99 .32	1.98 .32
	5.15	1.01 .97	1.01 .97	5.25	1.01 .97	1.01 .97	5.45	1.01 .98	1.01 .98

Case B. θ_0 Unknown.

In this case observations are taken from all of the (k+1) populations. Let $\overline{x}_i$ denote the sample mean of n_i observations from Π_i (i=0,1,...,k). Consider the following selection rule:

$$R_B: \text{ select } \Pi_i \text{ iff } \overline{x}_0 - a - \frac{D\sigma}{\sqrt{n_i}} \leq \overline{x}_i \leq \overline{x}_0 + a + \frac{D\sigma}{\sqrt{n_i}} \tag{3.8}$$

Simple calculation gives

$$P(CS|R_B) = \int_{-\infty}^{\infty} [\prod_{i=1}^{k_1} H(\theta_i', y)]\Phi(y)dy \tag{3.9}$$

where θ_i', n_i' (i=1,...,k) and k_1 are as in Case A, and $H(\theta_i',y)$ is defined by

$$H(\theta_i',y) = \Phi(\frac{\sqrt{n_i}}{\sigma}[\frac{\sigma}{\sqrt{n_0}}y + \theta_0 - \theta_i'] + \frac{a\sqrt{n_i'}}{\sigma} + D) - \Phi(\frac{\sqrt{n_i}}{\sigma}[\frac{\sigma}{\sqrt{n_0}}y + \theta_0 - \theta_i'] - \frac{a\sqrt{n_i'}}{\sigma} - D) \tag{3.10}$$

We can easily verify that, for each fixed $y \in \mathbb{R}$

(i) $H(\theta_i',y)$ is a continuous function of θ_i' and hence attains its minimum in the compact set $[\theta_0-a, \theta_0+a]$,

(ii) $H(\theta_i',y)$ is symmetric about $\theta = \frac{\sigma}{\sqrt{n_0}}y + \theta_0$, i.e.

$$H(\frac{\sigma}{\sqrt{n_0}}y + \theta_0 + \theta_i', y) = H(\frac{\sigma}{\sqrt{n_0}}y + \theta_0 - \theta_i', y)$$

(iii) $H(\theta_i',y)$ increases (decreases) with θ_i' if

$$\theta_i' < \frac{\sigma}{\sqrt{n_0}}y + \theta_0 \quad (\theta_i' > \frac{\sigma}{\sqrt{n_0}}y + \theta_0)$$

It follows from (i) to (iii) above that, for each fixed y,

$$\inf_{|\theta_i'-\theta_0| \leq a} H(\theta_i',y) = \begin{cases} H(\theta_0+a,y) & \text{if } y < 0 \\ H(\theta_0-a,y) & \text{if } y > 0 \end{cases}$$

Hence

$$P(CS|R_B) \geq \int_{-\infty}^{0} \prod_{i=1}^{k_1} [\Phi(\sqrt{\frac{n_i'}{n_0}}\, y + D) - (\sqrt{\frac{n_i'}{n_0}}\, y - \frac{2a\sqrt{n_i'}}{\sigma} - D)]\phi(y)\,dy$$

$$+ \int_{0}^{\infty} \{ \prod_{i=1}^{k_1} [\Phi(\sqrt{\frac{n_i'}{n_0}}\, y + \frac{2a\sqrt{n_i'}}{\sigma} + D) - \Phi(\sqrt{\frac{n_i'}{n_0}}\, y - D)]\} \phi(y)\,dy \qquad (3.11)$$

If k_1 is unknown a lower bound for $P(CS|R_B)$ can be obtained by replacing k_1 by k in (3.11).

2. SCALE PARAMETER-GAMMA POPULATIONS WITH KNOWN SHAPE PARAMETERS

Here $\Pi_i (i=0,1,\ldots,k)$ has density

$$g(x;\theta_i,\alpha) = \frac{\theta_i^{-\alpha_i/2}}{\Gamma(\alpha_{i/2})} x^{\alpha_i/2-1} e^{-x/\theta_i} \quad , \; x,\theta_i > 0$$

where α_i are known positive constants. Let $G_1(x;\theta_i,\alpha_i)$ denote the cumulative distribution function (cdf) of Π_i. In this case we say that Π_i is 'close' to Π_0 if

$$\frac{\theta_0}{\beta} \leq \theta_i \leq \beta\theta_0$$

where $\beta > 1$ is a given constant.

Case A. θ_0 Known.

Let $X_{ij} (j=1,\ldots,n_i)$ be an independent sample of size n_i from $\Pi_i (i=1,\ldots,k)$. Define $T_i = \sum_{j=1}^{n_i} X_{ij}$ and consider the rule

$$R_A': \text{ select } \Pi_i \text{ iff } \frac{\beta\theta_0}{c} \leq \frac{T_i}{\nu_i} \leq \beta\theta_0 c \qquad (3.12)$$

where $\nu_i = n_i\alpha_i$, $i=1,\ldots,k$, and $c > 1$ is chosen so as to satisfy the basic P^*-condition.

Using the fact that the cdf of $\frac{T_i}{\theta_i}$ is $G_1(t;1,\nu_i)$ we obtain

$$P(CS|R_A') = \prod_{i=1}^{k_1} [G(\frac{\beta\theta_0 c\nu_i'}{\theta_i'};1,\nu_i)-G_1(\frac{\beta\theta_0\nu_i'}{c\theta_i'};1,\nu_i)] \quad (3.13)$$

where, as before, k_1 is the number of populations Π_i with $\frac{\theta_0}{p} \leq \theta_i \leq \beta\theta_0$, and the primes refer to values corresponding to the populations close to Π_0.

It is easily verified that each term in the product on the right hand side of (3.13) is increasing in θ_i' if $\theta_i' \leq \beta\theta_0$, and hence

$$\inf_{\frac{\theta_0}{\beta} \leq \theta_i' \leq \beta\theta_0} P(CS|R_A') = \prod_{i=1}^{k_1} [G_1(\beta^2 c\nu_i';1,\nu_i')-G(\frac{\beta^2\nu_i'}{c};i,\nu_i')] \quad (3.14)$$

If k_1 is unknown, a conservative value of c can be obtained by taking $k_1=k$ in (3.14) and equating the expression to P^*.

Case B. θ_0 Unknown.

In this case, consider the rule R_B': select Π_i iff

$$\frac{\beta T_0}{\nu_0 c} \leq \frac{T_i}{\nu_i} \leq \frac{\beta T_0 c}{\nu_0} \quad (3.15)$$

where T_i and $\nu_i (i=0,1,\ldots,k)$ are defined as in Case A, and $C > 1$ is a constant to be determined from the basic P^*-condition.

$$P(CS|R_B') \geq \int_0^\infty \prod_{i=1}^{k} [G_1(\frac{\beta^2 C\nu_i}{\nu_0} u;1,\nu_i)-G_1(\frac{\beta^2\nu_i}{C\nu_0} u;1,\nu_i)] \cdot g(u;1,\nu_0)\,du \quad (3.16)$$

A conservative C can be obtained by equating the right hand side of (3.16) to P^*.

APPLICATION TO SELECTING VARIANCES OF NORMAL POPULATIONS

Let Π_i be a normal population with mean μ_i and variance $\sigma_i^2 (i=0,1,\ldots,k)$. We will say that Π_i is close to Π_0 if

$\frac{\theta_0}{\beta} \leq \theta_i \leq \theta_0\beta$, where $\theta_i = 2\sigma_i^2$ and $\beta>1$ is a given constant. Assume θ_0 is known.

When the means $\mu_i (i=0,1,\ldots,k)$ are known, the statistic

$S_i^2 = \sum_{j=1}^{n_i} (x_{ij}-\mu_i)^2/n_i$ is sufficient for σ_i^2.

For selecting a subset containing all populations with variances lying in $[\theta_0/\beta, \theta_0\beta]$ consider the following rule:

Select Π_i iff $\frac{2\beta\sigma_0^2}{d} \leq S_i^2 \leq 2\beta d\sigma_0^2$, $d>1, i=1,\ldots,k$.

Using the fact that $n_i S_i^2/\sigma_i$ is distributed as $\frac{1}{2}\chi_{n_i}^2$ random variable, we can show that the equation for d is the same as that obtained from (3.14) with $\nu_i = n_i$. If the means $\mu_i (i=0,1,\ldots,k)$ are unknown and $n_i > 1 (i=1,\ldots,k)$, then we use $\bar{x}_i$ in place of μ_i and $n_i - 1$ for n_i.

ACKNOWLEDGEMENT

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

BIBLIOGRAPHY

Burr, I. W. (1976). Statistical Quality Control Methods. Marcel Dekker, Inc., N.Y.

Gupta, S. S. (1965). On Some Multiple Decision (Selection and Ranking) Procedures. Technometrics 7: 225-245.

Gupta, S. S. and Sobel, M. (1958). On Selecting a Subset Which Contains All Populations Better Than a Standard. Ann. Math. Statist. 29: 235-244.

Huang, W. T. (1975). Bayes Approach to a Problem of Partitioning K Normal Populations. Bull. Inst. Math. Acad. Sinica. 3: 87-97.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Mimeograph Series #508	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): On Selection of Populations Close to a Control or Standard		5. TYPE OF REPORT & PERIOD COVERED: Technical
		6. PERFORMING ORG. REPORT NUMBER: Mimeo Series #508
7. AUTHOR(s): Shanti S. Gupta and Ashok K. Singh		8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0455
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, Purdue University, W. Lafayette, IN 47907		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE: September, 1977
		13. NUMBER OF PAGES: 16
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report): UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bayes rule, correct selection, loss function, indicator function

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

In many treatment vs. control situations the experimenter is interested in the populations which are 'close' to control rather than the populations which are 'better' than control. In this paper the problem of selection of all populations which are 'close' to a given standard or control population has been considered from the subset selection approach, and some Bayes and classical rules are investigated.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N-0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)