ASYMPTOTICALLY OPTIMAL PROCEDURES FOR SEQUENTIAL ADAPTIVE SELECTION OF THE BEST OF SEVERAL NORMAL MEANS

Christopher Jennison[1], Iain M. Johnstone[2]

and Bruce W. Turnbull[1]

Departments of Operations Research and Mathematics
Cornell University
Ithaca, New York, U.S.A.

I. INTRODUCTION

Suppose we have $k(\geq 2)$ normal populations with common variance σ^2 and unknown means $\{\mu_i;\ 1 \leq i \leq k\}$. We wish to select a population with a "high" mean, the population with the highest mean is called the best population. Let $\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[k]}$ denote the ordered means. Bechhofer [1] formulated this problem with the following probability of correct selection (PCS) requirement:

(PCS 1) Whenever $\mu_{[k]} - \mu_{[k-1]} \geq \delta$,

P(Select the best population) $\geq P^*$,

where $\delta > 0$ and $1/k < P^* < 1$ are to be set by the experimenter. Fabian [6] and Kao and Lai [10] proposed a stronger PCS requirement:

(PCS 2) P(Mean of selected population $> \mu_{[k]} - \delta) \geq P^*$.

For the case of known variance Bechhofer gave a fixed sample size

[1]Research supported in part by grants from the National Institutes of Health.

[2]Research supported by Australian National University Postbachelor Travelling Scholarship. This author is now at the Department of Statistics, Stanford University, Stanford, California.

ISBN 0-12-307502-5

procedure which, in fact, satisfies both PCS requirements. However, some of the later sequential methods designed to satisfy PCS1 may not meet PCS2.

Just as in one population problems, a reduction in average sample size can be achieved by using a sequential procedure. Sequential methods are also necessary when σ^2 is unknown. Paulson [12] proposed a sequential procedure based on elimination. As the experiment progresses, populations with sufficiently small sample means are successively eliminated from consideration and observations are taken equally from the remaining populations. Similar procedures have been studied by Swanepoel and Geertsema [17] and by Kao and Lai [10]. The expected total number of observations is called the average sample number (ASN). We show that, for this form of elimination procedure, if $k, \delta, \{\mu_i\}$ and σ^2 are fixed and we let $P^* \to 1$, there is a sharp asymptotic lower bound for the natural measure of efficiency, namely $ASN/(-\log(1-P^*))$. This bound depends on which of the two PCS requirements is used and it is larger in the case of unknown variance. In Section 2 we derive these lower bounds and give procedures which achieve them.

For the case of known variance, Bechhofer, Kiefer and Sobel [3] (hereafter referred to as BKS) suggested another sequential procedure based on an identification problem. In the identification problem the values of the population means are assumed known, but the correspondence between means and populations is unknown. The problem is to identify the population associated with the highest mean. Observations are taken equally from all k populations, we call this "vector at a time sampling" (VT). BKS (Theorem 6.1.1) prove that their procedure also solves the selection problem for PCS1.

The average sample number may also be reduced by data dependent (or adaptive) sampling. Instead of taking an equal number of observations from each population, observations are allocated preferentially to populations with high sample means. This has the added effect of reducing the expected sample size for inferior populations, an important consideration in comparisons of medical treatments. Turnbull, Kaspi and Smith [18] study a number of adaptive sampling rules for the identification problem, but their procedures, which are based on the BKS procedure, do not provide a solution to the *selection* problem. In fact, the questions posed by their paper stimulated the present study. In Sections 3.1-3.4 we consider a class of elimination procedures with adaptive sampling which do solve the selection problem. Within this class we find a sharp asymptotic lower bound for ASN/(-log(1-P*)) as $P^* \to 1$ and procedures that attain it. If k is large, this bound can be substantially lower than that for VT elimination procedures. Some Monte Carlo simulations, presented in Section 3.5, illustrate the potential savings in sample size which can be achieved by using fairly simple adaptive sampling rules.

II. ELIMINATION PROCEDURES WITH VECTOR AT A TIME SAMPLING

2.1 *General Form of the Procedures*

Observations X_{ip} $(1 \le i \le k,\ p \ge 1)$ are available from k $(k \ge 2)$ populations $\Pi_1, \Pi_2, \ldots, \Pi_k$. The $\{X_{ip}\}$ are independent normal random variables with mean μ_i and variance σ^2. We first consider the selection problem when σ^2 *is known* and requirement PCS1 is used.

In the general elimination procedure observations are taken in groups, one from each uneliminated population. When the number of vectors of observations taken is n we say we are at stage n. Define

$$I_n = \{i:\ \Pi_i \text{ is not eliminated by the end of stage } n\},$$

$$S_{ij}(n) = \sum_{p=1}^{n} X_{ip} - \sum_{p=1}^{n} X_{jp}.$$

Elimination Rule. At stage n eliminate all populations Π_j, $j \in I_{n-1}$, for which there is $i \in I_{n-1}$ with

$$S_{ij}(n) > g(n), \tag{2.1}$$

where g is a nonnegative function. We say Π_i eliminates Π_j. This leaves a new set of uneliminated populations $I_n \subseteq I_{n-1}$. Once a population has been eliminated no more observations are taken on it. When only one population remains, select it as the best population.

Continuation Region. Plotting $S_{ij}(n)$ against $g(n)$ we see that (2.1) holds if $S_{ij}(n)$ lies above $g(n)$. If $S_{ij}(n) < -g(n)$ then (2.1) holds with i and j interchanged, which corresponds to elimination of Π_i by Π_j. We refer to the region inside $S = g(n)$, $S = -g(n)$, $n \geq 0$ as the continuation region and its complement as the stopping region. (More accurate terminology might be "non-elimination" and "elimination" regions, respectively.) Denote the continuation region by C.

Guaranteeing Probability of Correct Selection. For PCS1 we need only consider the case where $\mu_{[k-1]} \leq \mu_{[k]} - \delta$. Then

$$\begin{aligned} P\{\text{Incorrect Selection}\} &= P\{\Pi_{[k]} \text{ is eliminated at some point}\} \\ &= P\{\bigcup_{i \neq [k]} (\Pi_i \text{ eliminates } \Pi_{[k]})\} \qquad (2.2) \\ &\leq \sum_{i \neq [k]} P\{S_{i[k]}(n) \text{ exits C upwards}\}. \end{aligned}$$

Here [k] denotes the subscript of the population with mean $\mu_{[k]}$. By exiting C upwards we mean that the first point at which $\{S_{i[k]}(n), n\}$ lies outside C is in the upper half plane $\{S_{i[k]}(n) \geq 0\}$. Since Π_i or $\Pi_{[k]}$ may be eliminated by a third population there is a positive probability that $S_{i[k]}(n)$ remains in C throughout the experiment. Let $Z_\theta(n) = \sum_{p=1}^{n} Y_p$ where the Y_p

are independent $N(\theta, 2\sigma^2)$. The sequence $\{S_{i[k]}(n)\}$ is stochastically smaller than $\{Z_{-\delta}(n)\}$, hence (2.2) is satisfied if

(2.3) $$P\{Z_{-\delta}(n) \text{ exits C upwards}\} \leq \frac{1-P^*}{k-1}.$$

The choice of g depends on δ, σ^2, k and P^*. We allow any nonnegative function g which meets the requirements PCS1 by satisfying (2.3) and for which the procedure terminates with probability 1.

Denote $x^+ = \max(x,0)$. Paulson [12] proposed functions of the form

$$g(n) = (a_\lambda - n\lambda)^+$$

with $0 < \lambda < \delta$. Fabian [7] used a likelihood ratio method employed by Lawing and David [11] to give a better lower bound for the guaranteed PCS than that proposed by Paulson.

Let $\{B(t);\ t \geq 0\}$ be a standard Brownian notion with drift $-\delta/\sqrt{2}\sigma$ and let $W(t) = \sqrt{2}\ \sigma\ B(t)$. Then $\{Z_{-\delta}(n);\ n = 1,2,\ldots\}$ has the same joint distribution as $\{W(t);\ t = 1,2,\ldots\}$. If g is a smooth function, then whenever $\{W(t);\ t \geq 0\}$ exits C it is very likely that $\{W(t);\ t = 1,2,\ldots\}$ will exit C nearby and in the same direction. Hence

$$P\{Z_{-\delta}(n) \text{ exits C upwards}\} \approx P\{W(t),\ t \geq 0, \text{ exits C upwards}\}.$$

In fact, a likelihood ratio argument in Appendix 3 of Jennison, Johnstone and Turnbull [9], (hereafter referred to as JJT) shows that

$$P\{Z_{-\delta}(n) \text{ exits C upwards}\} \leq P\{W(t),\ t \geq 0, \text{ exits C upwards}\},$$

so choosing g to satisfy $P\{W(t) \text{ exits C upwards}\} \leq (1-P^*)/(k-1)$ will guarantee the PCS requirement.

Swanepoel and Geertsema [17] use

$$g(n) = [(2\sigma^2\ n(b^2 + \log n))^{1/2} - n\delta]^+,$$

and an upper bound on $P\{Z_{-\delta}(n)$ exits C upwards$\}$ is found from a result of Robbins [14].

Kao and Lai [10] do not give an explicit formula for $g(n)$. In their procedure population Π_i eliminates Π_j when

$$g(S_{ij}(n) - n\delta,\ n) > \frac{k-1}{1-P^*},$$

where $g(x,n)$ is a mixture likelihood ratio--see their Equation 7, page 1661.

2.2. *An Asymptotic Lower Bound for the Average Sample Number*

We first derive some preliminary results. Let $Z_\theta(n) = \sum_{p=1}^{n} Y_p$, where the Y_p are independent $N(\theta, 2\sigma^2)$. Consider a size ε sequential test of H_0: $\theta = 0$ based on the sequence $\{Z_\theta(n)\}$ which rejects H_0 if and only if the test stops in finite time--see Robbins ([14], Section 4). Any such test must be open ended and will fail to terminate with probability at least $1-\varepsilon$ if $\theta = 0$. Let N be the number of observations taken before stopping, with $N = \infty$ if the experiment does not terminate. Now consider a collection of such tests, one for each $\varepsilon \in (0,1)$. We shall denote the corresponding probability measures by $P_{\theta,\varepsilon}$ although for notational convenience we shall usually suppress the dependence on ε.

LEMMA 2.1. *If* $\kappa > 0$ *and* θ *are fixed then* $P_{\theta,\varepsilon}(N \leq \kappa) \to 0$ *as* $\varepsilon \to 0$.

Proof. Take $u > 0$ and define $\Omega_1 = \{N \leq \kappa,\ |Z_\theta(N)| \leq u\}$, $\Omega_2 = \{N \leq \kappa,\ |Z_\theta(N)| > u\}$ so $P_\theta(N \leq \kappa) = P_\theta(\Omega_1) + P_\theta(\Omega_2)$. Now

$$P_\theta(\Omega_1) = \int_{\Omega_1} dP_\theta = \int_{\Omega_1} \frac{dP_\theta}{dP_0}(Z_\theta(N), N)\, dP_0$$

$$= \int_{\Omega_1} \exp\left(\frac{\theta Z_\theta(N)}{2\sigma^2} - \frac{\theta^2 N}{4\sigma^2}\right) dP_0$$

$$\leq \exp\left(\frac{|\theta| u}{2\sigma^2}\right) \cdot \int_{\Omega_1} dP_0 \leq \exp\left(\frac{|\theta| u}{2\sigma^2}\right) \cdot \varepsilon.$$

Thus, $P_\theta(\Omega_1) \to 0$ as $\varepsilon \to 0$. Also $P_\theta(\Omega_2) \leq P_\theta(\max_{n \leq \kappa} |Z_\theta(n)| > u)$ so $P_\theta(\Omega_2) \to 0$ as $u \to \infty$. By first taking u large and then taking ε small we see that $P_{\theta,\varepsilon}(N \leq \kappa)$ is arbitrarily small for small enough ε.

THEOREM 2.2. *If* $\zeta > 0$ *is fixed, then for* $\theta \neq 0$

$$P_{\theta,\varepsilon}\left(\frac{N}{-\log\varepsilon} \leq \frac{4\sigma^2}{\theta^2} - \zeta\right) \to 0 \quad \textit{as} \quad \varepsilon \to 0.$$

Proof. Without loss of generality take $\theta > 0$, suppose that $\gamma > 0$, $\xi > 0$ and there is a sequence $\{\varepsilon(i);\ i = 1,2,\ldots\}, \varepsilon(i) \to 0$, such that

$$P_{\theta,\varepsilon(i)}\left(\frac{N}{-\log\ \varepsilon(i)} \leq \gamma\right) \geq \xi.$$

Let $\Omega_1 = \{N/(-\log\ \varepsilon) \leq \gamma\}$ and for $\lambda > 0$ let

$$\Omega_2 = \{\theta/2 \leq \frac{Z_\theta(N)}{N} \leq \theta + \lambda\}.$$

As $\kappa \to \infty$, $P\{\theta/2 \leq \frac{Z_\theta(n)}{n} \leq \theta + \lambda$ for $n \geq \kappa\} \to 1$ and with the lemma this implies $P(\Omega_2) \to 1$ as $\varepsilon \to 0$. Hence, for large i, $P_\theta(\Omega_1 \cap \Omega_2) \geq \xi/2$. Now

$$\varepsilon(i) \geq P_0(\text{reject } H_0) \geq \int_{\Omega_1 \cap \Omega_2} dP_0$$

$$= \int_{\Omega_1 \cap \Omega_2} \frac{dP_0}{dP_\theta}(Z_\theta(N),N)\, dP_\theta$$

$$= \int_{\Omega_1 \cap \Omega_2} \exp\{[\frac{\theta}{2\sigma^2} \cdot \frac{Z_\theta(N)}{N} - \frac{\theta^2}{4\sigma^2}] \cdot [\frac{N}{-\log \varepsilon(i)}] \log \varepsilon(i)\} dP_\theta$$

$$\geq \int_{\Omega_1 \cap \Omega_2} \exp\{(\frac{\theta^2}{4\sigma^2} + \frac{\theta\lambda}{2\sigma^2})\gamma \log \varepsilon(i)\} dP_\theta .$$

So

$$\varepsilon(i) \geq \exp\{(\frac{\theta^2}{4\sigma^2} + \frac{\theta\lambda}{2\sigma^2})\gamma \log \varepsilon(i)\} \cdot \xi/2,$$

and letting $\varepsilon(i) \to 0$ we see that

$$\gamma \geq (\frac{\theta^2}{4\sigma^2} + \frac{\theta\lambda}{2\sigma^2})^{-1}.$$

Since λ was arbitrary $\gamma \geq \frac{4\sigma^2}{\theta^2}$ and the result follows.

We now return to the selection problem. Suppose we have a collection of elimination procedures indexed by $\varepsilon = (1-P^*)/(k-1)$, $(0 < \varepsilon < 1)$. For each procedure there is a nonnegative function $g_\varepsilon(n)$ and a symmetric continuation region C_ε. In the case of known σ^2 with requirement PCS1 the probability of correct selection is guaranteed by satisfying.

(2.4) $\quad P\{Z_{-\delta}(n) \text{ exits } C_\varepsilon \text{ upwards}\} \leq \varepsilon,$

where $Z_\theta(n) = \sum_{p=1}^{n} Y_p$ with the Y_p's independent $N(\theta, 2\sigma^2)$.

THEOREM 2.3. *Let* $N_0 = \inf\{n\colon\ Z_\theta(n)$ *lies outside* $C_\varepsilon\}$. *If* $\zeta > 0$ *is fixed, then*

$$P_{\theta,\varepsilon}\{\frac{N_0}{-\log \varepsilon} \leq \frac{4\sigma^2}{(|\theta|+\delta)^2} - \zeta\} \to 0 \quad \text{as} \quad \varepsilon \to 0.$$

Proof. Consider the following test of H_0: $\theta = \delta$. Reject H_0 after N_0 observations if $Z_\theta(N_0) < 0$; otherwise do not reject H_0. In the notation of Theorem 2.2, $N = N_0$ if $Z_\theta(N_0) < 0$ and $N = \infty$ otherwise. By (2.4), $P\{\text{Reject } H_0 | \theta = \delta\} \leq \varepsilon$ and it follows from the theorem that

$$P_{\theta,\varepsilon}\{\frac{N_0}{-\log \varepsilon} \leq \frac{4\sigma^2}{(\theta-\delta)^2} - \zeta \text{ and } Z_\theta(N_0) < 0\} \to 0 \text{ as } \varepsilon \to 0. \tag{2.5}$$

Similarly, by constructing a test of H_0: $\theta = -\delta$ in the obvious way, we get

$$P_{\theta,\varepsilon}\{\frac{N_0}{-\log \varepsilon} \leq \frac{4\sigma^2}{(\theta+\delta)^2} - \zeta \text{ and } Z_\theta(N_0) > 0] \to 0 \text{ as } \varepsilon \to 0. \tag{2.6}$$

The result follows from (2.5), (2.6) and the fact that $P(Z_\theta(N_0) = 0) = 0$.

COROLLARY 2.4. *Let* $N_{ij} = \inf\{n: S_{ij}(n)$ *lies outside* $C_\varepsilon\}$. *If either* Π_i *or* Π_j *is eliminated by a third population* $N_{ij} = \infty$. *If* $\zeta > 0$ *is fixed, then*

$$P_\varepsilon\{\frac{N_{ij}}{-\log \varepsilon} \leq \frac{4\sigma^2}{(|\mu_i-\mu_j|+\delta)^2} - \zeta\} \to 0 \quad \textit{as} \quad \varepsilon \to 0.$$

Proof. Let $Z_\theta(n) = S_{ij}(n)$, $\theta = \mu_i - \mu_j$.

For a particular experiment let the total number of observations taken on all k populations be H. The average sample number (ASN) is E(H). If there is no unique best population let [k] be the label of any one of the populations with the largest mean.

THEOREM 2.5. *For any collection of elimination procedures which guarantee* PCS1 *in the case of known* σ^2 *by satisfying* (2.3),

(i) *if* $\zeta > 0$ *is fixed then*

(2.7) $$P\{\frac{H}{-\log\varepsilon} \geq \sum_{i\neq[k]} \frac{4\sigma^2}{(\mu_{[k]}-\mu_i+\delta)^2} + \frac{4\sigma^2}{(\mu_{[k]}-\mu_{[k-1]}+\delta)^2} - \zeta\} \to 1$$

as $\varepsilon \to 0$,

(ii)

(2.8) $$\liminf_{\varepsilon\to 0} \left(\frac{ASN}{-\log\varepsilon}\right) \geq \sum_{i\neq[k]} \frac{4\sigma^2}{(\mu_{[k]}-\mu_i+\delta)^2} + \frac{4\sigma^2}{(\mu_{[k]}-\mu_{[k-1]}+\delta)^2}.$$

Proof. Let Ω_1 be the event that

(1) $\frac{N_{ij}}{-\log\varepsilon} \geq \frac{4\sigma^2}{(|\mu_i-\mu_j|+\delta)^2} - \frac{\zeta}{k}$ for all pairs $i \neq j$.

(2) The final elimination takes place between populations with means $\mu_{[k]}$ and $\mu_{[k-1]}$.

It is easily shown that on Ω_1

$$\frac{H}{-\log\varepsilon} \geq \sum_{i\neq[k]} \frac{4\sigma^2}{(\mu_{[k]}-\mu_i+\delta)^2} + \frac{4\sigma^2}{(\mu_{[k]}-\mu_{[k-1]}+\delta)^2} - \zeta.$$

By Corollary 2.4, Lemma 2.1 and the strong law of large numbers $P(\Omega_1) \to 1$ as $\varepsilon \to 0$. This proves (i), and (ii) follows immediately.

For any collection of procedures define

$$R = \limsup_{\varepsilon\to 0} \frac{ASN}{-\log\varepsilon}.$$

This is the reciprocal of Berk's [4] "efficacy." Note that $\lim_{\varepsilon\to 0} -\log\varepsilon/(-\log(1-P^*)) = 1$ so $R = \limsup(ASN/(-\log(1-P^*))$, as $P^* \to 1$. It follows from Theorem 2.5 and the results of the next section that R is a natural measure of a procedure's performance.

2.3 *Asymptotically Efficient Procedures*

Perng [13] calculated $\lim_{\varepsilon \to 0} (\text{ASN}/(-\log \varepsilon))$ for Paulson's procedures. In his method, which can be applied to most closed procedures, the limit is found by calculating the sample size when we observe $X_{ip} = \mu_i$ $(1 \leq i \leq k,\ p \geq 1)$. This is called the mean path approximation. Denoting $\lim (\text{ASN}/-\log \varepsilon)$ for Paulson's procedure with parameter λ by $R(P,\lambda)$ Perng shows that

$$R(P,\lambda) = \sum_{i \neq [k]} \frac{\sigma^2}{(\lambda+\mu_{[k]}-\mu_i)(\delta-\lambda)} + \frac{\sigma^2}{(\lambda+\mu_{[k]}-\mu_{[k-1]})(\delta-\lambda)} \qquad (0 < \lambda < \delta).$$

The procedures of Swanepoel and Geertsema [17] and of Kao and Lai [10] both achieve the lower bound derived in Theorem 2.5, namely

$$\sum_{i \neq [k]} \frac{4\sigma^2}{(\delta+\mu_{[k]}-\mu_i)^2} + \frac{4\sigma^2}{(\delta+\mu_{[k]}-\mu_{[k-1]})^2} = R^*, \quad \text{say}.$$

Thus, these procedures are asymptotically optimal within the class of elimination procedures defined in Section 2.2.

Bechhofer, Kiefer and Sobel ([3], page 161) give $\lim(\text{ASN}/-\log \varepsilon)$ for their procedure:

$$R(\text{BKS}) = \begin{cases} \dfrac{k\sigma^2}{\delta(\mu_{[k]}-\mu_{[k-1]})} & \text{if } \mu_{[k]} \neq \mu_{[k-1]} \\ \infty & \text{if } \mu_{[k]} = \mu_{[k-1]}. \end{cases}$$

Simple calculations show that if $\lambda \in (0,\delta)$ then $R(P,\lambda) > R^*$ although if the means are in the so-called δ-slippage configuration, i.e., $\mu_{[1]} = \ldots = \mu_{[k-1]} = \mu_{[k]} - \delta$, then $R(P,\lambda) \to R^*$ as $\lambda \to 0$. For any set of means $\{\mu_i\}$, $R(\text{BKS}) \geq R^*$ with equality only in the δ-slippage configuration.

Schwarz [16] and others have discussed the asymptotic shape of a sequence of stopping regions. For our optimal procedures the definition becomes

$$g(x) = \lim_{\varepsilon \to 0} \frac{g_\varepsilon(-x \cdot \log \varepsilon)}{-\log \varepsilon},$$

which gives

$$g(x) = 2\sigma\sqrt{x} - x\delta \qquad (0 < x < 4\sigma^2/\delta^2).$$

Schwarz [16] showed that this is the asymptotic shape of a Bayes test when the cost of an observation tends to zero. An optimal procedure could be constructed using regions with exactly this shape,

$$g_\varepsilon(n) = ((4\sigma^2\ a(\varepsilon)n)^{1/2} - n\delta)^+,$$

where $a(\varepsilon)$ is chosen to give $P(Z_{-\delta}(n)$ exits C_ε upwards$) = \varepsilon$.

In small sample experiments the choice of stopping region is difficult, typically a region will perform well for one configuration of means and poorly for another. Kao and Lai [10] show that their procedures have good small sample properties for a variety of configurations of the population means.

2.4. *The Case of a Common but Unknown Variance*

Suppose the variance of each observation, σ^2, is unknown. Define

$$(2.9) \qquad T_{ij}(n;\delta) = \frac{n^{-1/2}(S_{ij}(n) + n\delta)}{(V_{ij}(n)/(2n-2))^{1/2}},$$

where $V_{ij}(n) = \sum_{p=1}^{n} (X_{ip} - \bar{X}_{i\cdot}(n))^2 + \sum_{p=1}^{n} (X_{jp} - \bar{X}_{j\cdot}(n))^2$ and

$\bar{X}_{i\cdot}(n) = \frac{1}{n} \sum_{p=1}^{n} X_{ip}$. The $\{T_{ij}(n;\ \delta)\}$ have non-central Student's t distributions and they may be used to construct elimination procedures. Let population Π_i eliminate Π_j if $S_{ij}(n) > 0$ and $T_{ij}(n;\ \delta) > g(n)$ where $g(n)$ is a non-negative function. As before, PCS1 is guaranteed if

$$P\{T_{i[k]}(n;\ \delta) > g(n) \text{ for some } n \geq 1 \mid \mu_{[k]}-\mu_i=\delta\} \leq \frac{1-P^*}{k-1} = \varepsilon. \tag{2.10}$$

For elimination procedures based on the $\{T_{ij}(n;\ \delta)\}$ which guarantee PCS1 by satisfying (2.10) we can find a lower bound on $\liminf_{\varepsilon \to 0}(\text{ASN}/-\log \varepsilon)$ in the same way as in Section 2.2. Consider a size ε sequential test of H_0: $\mu_i-\mu_j + \delta = 0$, based on the $\{T_{ij}(n;\ \delta)\}$ which rejects H_0 if and only if the test stops in finite time. Let N be the number of observations taken before stopping, with $N = \infty$ if the experiment does not terminate. Jennison [8] shows that, for fixed $\zeta > 0$

$$P\left(\frac{N}{-\log \varepsilon} \leq \left[\log\left(1 + \frac{(\mu_i-\mu_j+\delta)^2}{4\sigma^2}\right)\right]^{-1} - \zeta\right) \to 0 \quad \text{as} \quad \varepsilon \to 0.$$

Proceeding as in Section 2.2 we find

$$\liminf_{\varepsilon \to 0}\left(\frac{\text{ASN}}{-\log \varepsilon}\right) \geq \sum_{i \neq [k]} \left[\log\left(1+ \frac{(\mu_{[k]}-\mu_i+\delta)^2}{4\sigma^2}\right)\right]^{-1} + \left[\log\left(1+ \frac{(\mu_{[k]}-\mu_{[k-1]}+\delta)^2}{4\sigma^2}\right)\right]^{-1}. \tag{2.11}$$

This bound is strictly larger than that for the case of known σ^2, (2.8), so the effect of not knowing σ^2 does not die away asymptotically. The bound is attainable and classes of procedures which attain it are given in Jennison [8].

Kao and Lai [10] considered the problem of unknown variance and their procedures suggested this approach. Procedures for the case of unequal and unknown variances are given by Swanepoel and Geertsema [17].

2.5 *Procedures for Requirement* PCS2

If all we know about a procedure is that it satisfies requirement PCS1 there is no guarantee about what happens when $\mu_{[k-1]} > \mu_{[k]} - \delta$. Fabian [6] and Kao and Lai [10] proposed the stronger PCS requirement

(PCS2) $\quad P(\text{Mean of selected population} > \mu_{[k]} - \delta) \geq P^*$.

This requirement implies PCS1, but it is stronger since the procedure must make a good selection with high probability for all possible values of $\{\mu_i\}$.

In the case of known variance, suppose an elimination procedure satisfying PCS1 is defined by $g(n)$, that is Π_i eliminates Π_j if $S_{ij}(n) > g(n)$. Kao and Lai show how to construct a new procedure which will satisfy PCS2. Let $I_n = \{i\colon \Pi_i$ is not eliminated at the end of stage $n\}$. The new procedure is given by the following elimination rule and stopping rule.

Elimination Rule. At stage n eliminate all populations Π_j, $j \in I_{n-1}$, for which there is $i \in I_{n-1}$ with

$$S_{ij}(n) > g(n) + n\delta.$$

Stopping Rule. If there is $i \in I_{n-1}$ such that $S_{ij}(n) > g(n)$ for all $j \in I_{n-1} \setminus \{i\}$ then stop and select Π_i as the best population.

We can modify procedures for the case of unknown variance in a similar way. For a PCS1 procedure such that Π_i eliminates Π_j if $S_{ij}(n) > 0$ and $T_{ij}(n;\,\delta) > g(n)$ the PCS2 procedure is as follows.

Elimination Rule. At stage n eliminate all populations Π_j, $j \in I_{n-1}$, for which there is $i \in I_{n-1}$ with

$$T_{ij}(n) = \frac{n^{-1/2}\, S_{ij}(n)}{(V_{ij}(n)/(2n-2))^{1/2}} > g(n).$$

Stopping Rule. If there is $i \in I_{n-1}$ such that

$T_{ij}(n;\ \delta) > g(n)$ for all $j \in I_{n-1}$, $j \neq i$,

then stop and select Π_i as the best population.

Using the argument of Kao and Lai([10], Theorem 7) it is easily shown that the new procedures guarantee PCS2.

For PCS2 procedures produced in this way optimality properties are inherited from the PCS1 procedures. Sharp asymptotic lower bounds may be found for (ASN/(-log ε)) and they are attained by those procedures derived from optimal PCS1 procedures.

III. ELIMINATION PROCEDURES WITH ADAPTIVE SAMPLING

3.1 *General Form of the Procedures*

For simplicity we shall restrict attention to the case of a common known variance, which we shall take to be 1, and requirement PCS1. The procedures are generalizations of the VT elimination procedures considered in Section 2 to allow data dependent sampling. At any time during the experiment a population may be eliminated if its sample mean is sufficiently less than that of another. In order to handle adaptive sampling we need a statistic to replace $S_{ij}(n)$. When the numbers of observations on Π_i and Π_j are n_i and n_j respectively a natural statistic is

$$Z_{ij}(n_i,n_j) = \frac{n_i\, n_j}{n_i+n_j}\,(\bar{X}_i(n_i) - \bar{X}_j(n_j)), \tag{3.1}$$

where $\bar{X}_i(n) = \sum_{p=1}^{n} X_{ip}/n$.

If the sequence of pairs (n_i,n_j) is deterministic it is easy to show that $\{Z_{ij}(n_i,n_j)\}$ has the same joint distribution as a standard Brownian motion with drift $\mu_i - \mu_j$ (per unit time) observed at the sequence of times $n_i n_j/(n_i+n_j)$. Robbins and Siegmund [15] show that this is also true for some data dependent sequences, $\{(n_i,n_j)\}$. They require that the allocation of observations between Π_i and Π_j should depend on $\bar{X}_i(n_i)$ and $\bar{X}_j(n_j)$ only through their difference, or equivalently, allocation should be independent of $(\sum_1^{n_i} X_{ip} + \sum_1^{n_j} X_{jp})$. In view of the translation invariant structure this is a reasonable restriction when there are only two populations, however, in the k population problem $(k > 2)$ a sampling rule which satisfies this condition for all pairs of populations must sample independently of each population mean. Since the motivation for adaptive sampling is to reduce sample size by estimating the population means and sampling accordingly this is an unacceptable restriction.

With a general adaptive procedure the joint distribution of the sequence $\{Z_{ij}(n_i,n_j)\}$ is not simple. For any observation, X_{ip} say, future values of its coefficient in the expression for $Z_{ij}(n_i,n_j)$, (3.1), depend on the sampling process which in turn depends on X_{ip}. Some examples of the difficulties that can arise are given in JJT, Appendices 1 and 2.

In this section we shall describe a class of 'multistage' procedures for which there is a sequence of statistics with the same joint distribution as a Brownian motion observed at random times. This allows a rigorous mathematical treatment which is given in Sections 3.1-3.4.

In Section 3.5 we shall discuss heuristic methods that approximate the joint distribution of $\{Z_{ij}(n_i,n_j)\}$ by that of a Brownian motion. Simulation results suggest that the heuristic methods achieve the required PCS and the Brownian motion approximation may in fact be reasonable for the sampling rules used in practice.

For the multistage procedures the experiment is conducted in a number of stages. The length of each stage and the order of sampling within it are fixed at the end of the previous stage.

Relabel the observations as X_{irp} $(1 \le i \le k,\ r \ge 1,\ p \ge 1)$ where r denotes the stage in which an observation is taken and p its order in that stage. At the start of stage r the number of observations to be allocated to Π_i during the stage is fixed, call this number M_{ir}. When a total of ν observations have been taken we say we are at time ν. Denote the current stage by $s(\nu)$, or for simplicity just by s. Let $m_{is}(\nu)$ be the number of observations taken on Π_i in the current stage at time ν. If Π_i and Π_j have not been eliminated before time ν define

$$\bar{X}_{ir\cdot} = \frac{1}{M_{ir}} \sum_{p=1}^{M_{ir}} X_{irp} \qquad 1 \le r \le s-1$$

$$\bar{X}_{is\cdot}(\nu) = \frac{1}{m_{is}(\nu)} \sum_{p=1}^{m_{is}(\nu)} X_{isp}$$

$$Z_{ij}(\nu) = \sum_{r=1}^{s-1} \frac{M_{ir}M_{jr}}{M_{ir}+M_{jr}} (\bar{X}_{ir\cdot} - \bar{X}_{jr\cdot})$$

$$+ \frac{m_{is}(\nu)\ m_{js}(\nu)}{m_{is}(\nu) + m_{js}(\nu)} (\bar{X}_{is\cdot}(\nu) - \bar{X}_{js\cdot}(\nu))$$

$$t_{ij}(\nu) = \sum_{r=1}^{s-1} \frac{M_{ir}\ M_{jr}}{M_{ir}+M_{jr}} + \frac{m_{is}(\nu)m_{js}(\nu)}{m_{is}(\nu)+m_{js}(\nu)}.$$

Elimination Rule. Let $I_\nu = \{i: \Pi_i$ is not eliminated by time $\nu\}$. At time ν eliminate all populations Π_j, $j \in I_{\nu-1}$ for which there is an $i \in I_{\nu-1}$ with

(3.2) $$Z_{ij}(\nu) > g(t_{ij}(\nu)),$$

where g is a nonnegative function. When only one population remains, select it as the best population. We refer to the region inside $Z = g(t)$, $Z = -g(t)$, $t \geq 0$ as the continuation region and denote it by C.

Sampling Rule. The sampling mechanism is determined by stages. The number of observations to be taken on each population during the first stage and the order in which they are to be taken must be determined before the first stage is started. If a population is eliminated the remaining observations due on it are not taken, but observations on the other populations are taken as originally planned. At the end of a stage the sampling for the next stage is determined based on observations currently available. The only restrictions we impose on the sampling rule and stopping region are that they should give the required PCS by the method described below and the experiment should terminate almost surely for any configuration of the population means.

Guaranteeing Probability of Correct Selection. For PCS1 we need only consider the case where $\mu_{[k-1]} \leq \mu_{[k]} - \delta$. Then, as in Section 2,

$$\text{(3.3)} \qquad P\{\text{Incorrect selection}\} \leq \sum_{i \neq [k]} P\{Z_{i[k]}(\nu) \text{ exits C upwards}\}.$$

Suppose we are at the start of stage r and populations Π_i and Π_j have not been eliminated. The sampling rule tells us how to take observations during the next stage so the increments in $t_{ij}(\nu)$ that will occur during stage r are now fixed. The increments in $Z_{ij}(\nu)$ during stage r now depend in a deterministic way on the future random observations X_{irp} and X_{jrp} $(p \geq 1)$. One may check that these increments have the same distribution as increments in a Brownian motion. Let $B_{ij}(t)$, $(t \geq 0)$, be a standard Brownian motion with drift $\mu_i - \mu_j$. We can regard the sequence $\{Z_{ij}(\nu)\}$ as being generated by observing B_{ij} at times $t_{ij}(\nu)$ and we say that Z_{ij} is embedded in B_{ij}.

To obtain an approximate upper bound on the error probability we consider the exit probabilities of a continuous time Brownian motion. Suppose $B_{-\delta}(t)$ is a standard Brownian motion with drift $-\delta$ and $\{W_{-\delta}(t_\alpha)\}$ are the values of a Brownian motion with drift $-\delta$ observed at the sequence of times $\{t_\alpha;\ \alpha \geq 1\}$. If the increments $\{t_\alpha - t_{\alpha-1}\}$ are small and C is smooth

(3.4) $$P\{W_{-\delta}(t_\alpha) \text{ exits C upwards}\} \approx P\{B_{-\delta}(t) \text{ exits C upwards}\}.$$

To satisfy (3.3) it is sufficient that

$$P\{Z_{i[k]}(\nu) \text{ exits C upwards}\} \leq \frac{1-P^*}{k-1} = \varepsilon \text{ for each } i \neq [k]$$

and it follows from (3.4) that this is satisfied approximately as long as

(3.5) $$P\{B_{-\delta}(t) \text{ exits C upwards}\} \leq \varepsilon.$$

Since the smallest value t_{ij} can take is 1/2 we need only consider the behavior of $B_{-\delta}(t)$ on $[1/2,\infty)$ and (3.5) becomes

(3.6) $$P\{B_{-\delta}(\tau) \geq g(\tau)\} \leq \varepsilon,$$

where $\tau = \inf\{t \geq 1/2:\ B_{-\delta}(t) \notin C\}$.

This criterion is a property of the continuation region only, and so approximate PCS is guaranteed independently of the sampling mechanism. This leads to considerable simplifications in the search for optimal procedures.

Our general elimination procedure with adaptive sampling is thus a multistage procedure defined by a sampling rule and a nonnegative function g(t) for which (3.6) holds. Elimination procedures with VT sampling are a special case where the procedures have only one stage. Another interesting subclass consists of two stage procedures. Typically, one might run a preliminary experiment with, say, vector at a time sampling and then use the sample means to decide on the proportions in which to sample during the main part of the experiment.

3.2. *An Asymptotic Lower Bound for the Average Sample Number*

Suppose we have a collection of adaptive elimination procedures indexed by $\varepsilon = (1-P^*)/(k-1)$. For each procedure there is a nonnegative function $g_\varepsilon(t)$ and a symmetric continuation region C_ε. Letting $B_\theta(t)$ denote a standard Brownian motion with drift θ, the probability of correct selection is guaranteed by satisfying (3.6), $P\{B_{-\delta}(\tau_\varepsilon) \geq g_\varepsilon(\tau_\varepsilon)\} \leq \varepsilon$, where $\tau_\varepsilon = \inf\{t \geq 1/2; B_{-\delta}(t) \notin C_\varepsilon\}$.

THEOREM 3.1. *Let* τ_ε *be the exit time from* C_ε *of* $B_\theta(t;\ t \geq 1/2)$ *as defined above. Fix* $\zeta > 0$, *then*

$$P\left\{\frac{\tau_\varepsilon}{-\log\varepsilon} \leq \frac{2}{(|\theta|+\delta)^2} - \zeta\right\} \to 0 \quad \textit{as} \quad \varepsilon \to 0.$$

Proof. The result follows as a continuous time version of the proof of Theorem 2.3.

This result is the basis for deriving a lower bound for the ASN. If Π_i eliminates Π_j then Z_{ij} must exit C_ε, with $t_{ij}(\nu) = T_{ij}$, say. The discrete process exits later than the continuous process, so Theorem 3.1 gives a bound on the T_{ij}'s that might occur. Let the total number of observations on Π_i during the whole experiment be H_i. It follows from the definitions that if Π_i eliminates Π_j then $H_iH_j/(H_i+H_j) \geq T_{ij}$, i.e.

$$\frac{1}{H_i} + \frac{1}{H_j} \leq \frac{1}{T_{ij}}. \tag{3.7}$$

Unless Π_i receives no further observations, the number of observations on Π_i when it eliminates Π_j is less than H_i and the inequality in (3.7) is strict. To reduce ASN we try to sample in such a way that all eliminations occur simultaneously and the sampling proportions remain roughly constant. Then if Π_i eliminates Π_j, $1/H_i + 1/H_j \simeq 1/T_{ij}$. To minimize $H = \sum_{i=1}^{k} H_i$, $\Pi_{[k]}$ should eliminate all the other populations since then the lower bounds on the T_{ij}'s from Theorem 3.1 are as small as possible. The lower bound for the ASN is found by minimizing $\sum_{i=1}^{k} H_i$ subject to $1/H_i + 1/H_{[k]} = 1/T_{[k]i}$ for $i \neq [k]$.

Let $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ denote the vector of population means. Denote the ASN for procedure ε by $E_{\mu,\varepsilon}(H)$. Let $[k]$ be the integer such that $\mu_{[k]}$ is the largest mean (or one of the largest if there is a tie) and define

$$f_\delta(\mu) = \inf_{d_i>0} \{\sum_{i=1}^{k} d_i : \frac{1}{d_i} + \frac{1}{d_{[k]}} = \frac{(\mu_{[k]} - \mu_i + \delta)^2}{2}, \ i \neq [k]\}.$$

THEOREM 3.2. *For any* μ *and* $\delta > 0$

$$\lim \inf\{\frac{E_{\mu,\varepsilon}(H)}{-\log \varepsilon}\} \geq f_\delta(\mu) \quad \textit{as } \varepsilon \to 0.$$

Proof. Take $\eta > 0$. Let Ω_1 be the set of outcomes for which $\mu_i \geq \mu_j$ and

$$(3.8) \qquad \frac{T_{ij}}{-\log \varepsilon} > \frac{2}{(\mu_i - \mu_j + \delta)^2 + 2\eta}$$

whenever Π_i eliminates Π_j. On Ω_1, (3.7) and (3.8) imply

$$\frac{-\log \varepsilon}{H_i} + \frac{-\log \varepsilon}{H_j} < \frac{(\mu_i - \mu_j + \delta)^2}{2} + \eta.$$

It follows by a combinatorial argument, details of which are in JJT Appendix 4, that the smallest sample size in Ω_1 occurs when $\Pi_{[k]}$ eliminates all other populations. On Ω_1

$$\frac{\sum_{i=1}^{k} H_i}{-\log \varepsilon} \geq \inf_{d_i>0} \{\sum_{i=1}^{k} d_i : \frac{1}{d_i} + \frac{1}{d_{[k]}} = \frac{(\mu_{[k]} - \mu_i + \delta)^2}{2} + \eta, \ i \neq [k]\}$$

$= f_\delta(\mu;\ \eta)$, say.

By Theorem 3.1, $P(\Omega_1) \to 1$ as $\varepsilon \to 0$ and thus

$$\lim\inf\{\frac{E_{\mu,\varepsilon}(H)}{-\log\varepsilon}\} \geq f_\delta(\mu;\ \eta).$$

But η was arbitrary and $f_\delta(\mu;\ \eta)$ is a continuous function of η and so the result follows.

As long as $\mu_{[k]} \neq \mu_{[k-1]}$ the bound of Theorem 3.2 is sharp. Procedures that attain this bound are described in the next section.

3.3 *Asymptotically Optimal Procedures*

In this section we exhibit two-stage procedures which attain the asymptotic lower bound of Theorem 3.2 whenever $\mu_{[k]} \neq \mu_{[k-1]}$. The proof of the theorem suggests how to achieve the bound. First we need a family of continuation regions which attains the lower bound of Theorem 3.1. We use Schwarz's regions (C_a) defined by

$$g(t;a) = \begin{cases} \sqrt{2at} - \delta t & 0 \leq t \leq 2a/\delta^2 \\ 0 & t > 2a/\delta^2. \end{cases} \tag{3.9}$$

Let $\varepsilon(a) = P\{B_{-\delta}(t;\ t \geq 1/2)$ exits C_a upwards$\}$. For selection procedure ε choose a so that $\varepsilon(a) = \varepsilon$. We shall frequently index by a instead of ε.

THEOREM 3.3. *For Schwarz's continuation regions*

(i) $\varepsilon(a) = \exp\{-a + o(a)\}$ *as* $a \to \infty$,

(ii) *If* $B_\theta(t;\ t \geq 1/2)$ *exits region* a *at time* τ_a *then*

$$\frac{\tau_a}{a} \overset{P}{\to} \frac{2}{(|\theta|+\delta)^2} \quad as \quad a \to \infty.$$

Proof. Berk [4] proved (i) for a Brownian motion observed only at integer values of t. A modification of Berk's proof to obtain the continuous time result is given in JJT, Theorem 4.1. The proof of (ii) is a simple application of the strong law of large numbers.

The $\{d_i\}$ corresponding to the infimum of $f_\delta(\mu)$ give the optimal sampling ratios, so if, hypothetically, we knew the vector of means $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ we would know how to minimize the ASN. In the first stage of our procedures all k populations are sampled equally (except for those which are eliminated), the estimate of μ at the end of stage one is used to calculate the sampling ratios for stage two. Denote by $\lambda(a)$ the number of observations taken on each uneliminated population at the end of stage one. Since sampling in stage one is not optimal, $\lambda(a)$ should be a small fraction of the total number of observations to be taken.

Let $\{\hat{\mu}_i\}$ be the sample means based on observations in stage one. Let $J = \{i: \Pi_i$ not eliminated during stage one$\}$. Let (k) be the integer such that $\hat{\mu}_{(k)} = \max(\hat{\mu}_i;\ i \in J)$. Using the mean path approximation we expect Z_{ij} to exit the continuation region when $T_{ij}/a \approx 2/(|\hat{\mu}_i - \hat{\mu}_j| + \delta)^2$ and if the first stage is sufficiently long we expect $\Pi_{(k)}$ to be selected as best. To minimize sample size, $\Pi_{(k)}$ should eliminate all the remaining populations simultaneously so $1/H_{(k)} + 1/H_i = 1/T_{(k)i}$ for all $i \in J \setminus \{(k)\}$. This suggests the following sampling rule. Let $\{\bar{d}_i\}$ minimize $\sum_{i \in J} d_i$ subject to $d_i > 0$ and

$$\frac{1}{d_i} + \frac{1}{d_{(k)}} = \frac{(\hat{\mu}_{(k)} - \hat{\mu}_i + \delta)^2}{2} \qquad i \in J \setminus \{(k)\}.$$

During stage two sample from Π_i $(i \in J)$ at a rate proportional to $\bar{d}_i$.

THEOREM 3.4. *For the procedures described above, if* $\lambda(a) \to \infty$ *and* $\lambda(a)/a \to 0$ *as* $a \to \infty$ *then*

$$\lim\left\{\frac{E_{\mu,\varepsilon}(H)}{-\log \varepsilon}\right\} = f_\delta(\mu) \quad as \quad \varepsilon \to 0,$$

whenever $\mu_{[k]} \neq \mu_{[k-1]}$.

Proof. An outline of the proof is given here, further details may be found in JJT, Theorem 4.2.

Fix $\xi > 0$ and let Ω_{1a} be the set of outcomes for procedure a where

(1) No elimination takes place in stage one,

(2) Population $\Pi_{[k]}$ has the largest mean at the end of stage one,

(3) $f_\delta(\hat{\mu}) \leq (1+\xi)\ f_\delta(\mu)$,

(4) $\Pi_{[k]}$ eliminates all other populations and

$$T_{[k]i} \leq \frac{2a}{(\hat{\mu}_{[k]}-\hat{\mu}_i+\delta)^2} \cdot(1+\xi) \quad \text{for all } i \neq [k].$$

Using the strong law of large numbers and Theorem 3.3 it can be shown that $P(\Omega_{1a}) \to 1$ as $a \to \infty$, as long as $\mu_{[k]} \neq \mu_{[k-1]}$.

On Ω_{1a} there are at most $a(1+\xi)^2\ f_\delta(\mu)$ observations in stage two and the number of observations in stage one is $o(a)$ as $a \to \infty$. The contribution to $E(H)$ outside Ω_{1a} is of order $o(a)$, hence

$$(3.10) \qquad \limsup_{a \to \infty}\left\{\frac{E_{\mu,\varepsilon}(H)}{a}\right\} \leq (1+\xi)^2\ f_\delta(\mu).$$

Combining this with Theorem 3.3 (i), the arbitrariness of ξ and Theorem 3.2 gives the result.

If there is not a unique best population the lower bound of Theorem 3.2 is not sharp since it is not possible to predict which population will be selected as best, early on in the experiment. In small sample size experiments the same problem arises if $\mu_{[k]} \simeq \mu_{[k-1]}$; allocating a high proportion of observations to a population which is subsequently eliminated increases the sample size. The two-stage procedures described in this section were constructed to show that the asymptotic lower bound of Theorem 3.2 is attainable. Typically they do not have good small sample size properties. A better procedure would have several stages and a more sophisticated sampling rule, for instance, sampling should be closer to vector at a time in the early stages of the experiment when the estimate of μ is not very accurate.

3.4. *Some Properties of Adaptive Procedures*

Suppose we have a family of continuation regions $\{C_\varepsilon\}$ satisfying equation (3.6) for which

$$g(x) = \lim_{\varepsilon \to 0} \frac{g_\varepsilon(-x \cdot \log \varepsilon)}{-\log \varepsilon} \tag{3.11}$$

exists and is continuous. Such regions are said to have an asymptotic shape. The exit times T_{ij} satisfy

$$\frac{T_{ij}}{-\log \varepsilon} \to \tau_{ij} \quad \text{in probability,} \tag{3.12}$$

where τ_{ij} is the smallest solution to $g(x) = |\mu_i - \mu_j| \cdot x$, $x > 0$. This can be seen from the mean path approximation under which T_{ij} satisfies $g_\varepsilon(T_{ij}) = |\mu_i - \mu_j| \cdot T_{ij}$, or equivalently

$$\frac{g_\varepsilon(-\log \varepsilon \cdot [T_{ij}/-\log \varepsilon])}{-\log \varepsilon} = |\mu_i - \mu_j| \cdot [T_{ij}/-\log \varepsilon],$$

and the result follows by (3.11).

Suppose there is a unique best population and for notational convenience let $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{k-1} < \mu_k$. Under mild regularity conditions on the $\{g_\varepsilon\}$, for VT sampling

$$\frac{E_{\mu,\varepsilon}(H)}{-\log \varepsilon} \to \sum_{i=1}^{k-1} 2\tau_{ki} + 2\tau_{k,k-1} = R_{VT}(\tau), \text{ say.}$$

Using an optimal adaptive sampling rule (OAS), the methods of Section 3.3 give

$$\frac{E_{\mu,\varepsilon}(H)}{-\log \varepsilon} \to \inf_{d_i>0} \{ \sum_{i=1}^{k} d_i : \frac{1}{d_i} + \frac{1}{d_k} = \frac{1}{\tau_{ki}} \} = R_{OAS}(\tau), \quad \text{say.}$$

The asymptotic efficiency of VT sampling is the ratio $\Lambda(\tau) = R_{OAS}(\tau)/R_{VT}(\tau)$. Note that Λ depends on the means and continuation regions only through $\tau = (\tau_{k1},\tau_{k2},\ldots,\tau_{k,k-1})$.

To demonstrate the reductions in sample size that can be achieved by adaptive sampling consider the Δ-slippage configuration, $\mu_1 = \mu_2 = \ldots = \mu_{k-1} = \mu_k-\Delta$, $\Delta > 0$. Here $\tau_{k1} = \tau_{k2} = \ldots = \tau_{k,k-1}$. The optimal allocation is found by minimizing $\sum_{i=1}^{k} d_i$ subject to $1/d_i + 1/d_k = 1/\tau_{ki}$, $1 \le i \le k-1$. The minimization gives $d_k = \tau_{k1}(\sqrt{k-1} + 1)$, $d_i = \tau_{k1}(\sqrt{k-1}+1)/\sqrt{k-1}$ $(i \ne k)$ and

$$\Lambda(\tau) = \frac{(\sqrt{k-1} + 1)^2}{2k}.$$

For $k = 2$, VT is optimal, but for $k = 3$, 10 and ∞, $\Lambda(t) = 97\%$, 80% and 50% respectively. The optimal rule allocates observations to $\Pi_1,\ldots,\Pi_{k-1},\Pi_k$ in the proportions $1: \ldots :1: \sqrt{k-1}$. This is also the asymptotically optimal allocation rule in multiple comparisons of k-1 normal treatments with a control (Dunnett [5], Bechhofer [2]). Note that here $\Lambda(\tau)$ and this optimal allocation rule are independent of the continuation region used.

We have seen that, for a particular family of regions and a given set of means, asymptotic comparisons of sampling rules can be made using the $\{\tau_{ij}\}$. Regions with low τ_{ij}'s give low sample sizes although the exact relationship is not simple. For the

Δ-slippage configuration, however, τ_{ki} is the same for all i and the asymptotic relative efficiency is the product of two independent components - the relative efficiencies of stopping rules and of sampling rules. For instance, comparing Paulson's procedures with parameter λ and VT sampling to procedures using Schwarz's regions and optimal sampling, in the Δ-slippage configuration,

$$\lim_{\varepsilon\to 0}\left(\frac{ASN}{-\log\varepsilon}\right)_{P,VT}\Big/\left(\frac{ASN}{-\log\varepsilon}\right)_{S,\sqrt{k-1}} = \frac{2k}{(\sqrt{k-1}+1)^2}\cdot\frac{(\delta+\Delta)^2}{4(\delta-\lambda)(\Delta+\lambda)} \geq 1.$$

3.5 *Simulation Results for Heuristic Procedures*

So far we have only made asymptotic comparisons; to study small sample properties we use Monte Carlo simulations. Although they are asymptotically efficient, the two-stage procedures of Section 3.3 have poor small sample properties (see JJT, Tables II and III). Instead we consider heuristic procedures based on the statistics $Z_{ij}(n_i,n_j)$ defined in Section 3.1. The sampling rule is motivated by the asymptotically optimal ratio in the slippage configuration, 1: ... :1: $\sqrt{k-1}$. At any point during the experiment we try to ensure that the number of observations on the population currently regarded as best bears the ratio $\sqrt{k_\nu-1}$:1 to the sample size on each of the other non-eliminated populations, where k_ν is the number of non-eliminated populations.

Sampling Rule. Initially, take one observations from each population. At time ν, let $\bar{X}_{i_{max}} = \max_{i\in I_\nu} \bar{X}_i$ and let i_ν be the index of the last population sampled. Beginning with $i_\nu+1$ (mod k) search for the next $j \in I_\nu \setminus \{i_{max}\}$ with $n_j < n_{i_{max}}/\sqrt{k_\nu-1}$. If such a j is found take the next observation from Π_j, otherwise sample from $\Pi_{i_{max}}$.

Probability of Correct Selection. Population Π_i eliminates Π_j when $Z_{ij}(n_i,n_j) \geq g(n_in_j/(n_i+n_j))$ where g satisfies equation

(3.6). If Z_{ij} could be embedded in a Brownian motion PCS would be guaranteed in the usual way. Such an embedding is not possible (see Section 3.1), however it is a reasonable approximation and the PCS requirement should be met in most cases. The simulation results suggest that these heuristic procedures do satisfy PCS1 for the sampling rule used.

Sample Size Calculations. As a consequence of the strong law of large numbers, sample sizes may be found from the mean path approximation. Thus, the optimal sampling rules apply to the heuristic procedures. In particular, the formulae for $R_{VT}(\tau)$ and $R_{OAS}(\tau)$ in Section 3.4 still hold.

The main advantage of the heuristic methods over multistage procedures is flexibility. If it becomes apparent that the sampling rule was based on a poor estimate of μ the rule can be changed immediately. The $\sqrt{k_\nu - 1}$ sampling rule is asymptotically optimal in any slippage configuration. For all configurations of means it is almost as effecent as the optimal rule and it does better asymptotically than VT sampling. If the means are equally spaced at a distance δ apart the asymptotic relative efficiency of $\sqrt{k_\nu - 1}$ sampling versus VT is 1.09 for Schwarz's region and 1.14 for Paulson's region. Details of these results and calculations are given in JJT, Section 6.

In the Monte Carlo study the values $P^* = 0.9$, $\delta = 0.2$, $\sigma^2 = 1$ and $k = 10$ were chosen. Three configurations of means were used, the δ-slippage configuration, equal means, and equally spaced (ES) means with $\mu_i - \mu_{i-1} = 0.2$. The results are displayed in Table I. Seven procedures were considered, namely (A) FIXED sample size (Bechhofer [1]); (B) the BKS likelihood based stopping rule with VT sampling; (C) the BKS rule with the adaptive RAND Q sampling rule described in Turnbull et al. [18] (hereafter referred to as TKS); (D), (E) Paulson's procedure with the Fabian [7] modification and $\lambda/\delta = 0.25$ and the VT and $\sqrt{k_\nu - 1}$ sampling rules respectively; (F), (G) the Schwarz stopping region together with VT and $\sqrt{k_\nu - 1}$ sampling rules. For Schwarz's region the error probability approximation given in Woodroofe ([19], Formulas (5.1) and (5.2) with r = 0) was used to give a = 5.31.

The sample sizes for (A) were taken from Bechhofer [1] Table I. In the equal means configuration, $\mu_{[1]} = \mu_{[10]}$, results for (B), taken from BKS Table 18.4.10, were based on 500 replications; while ASN for the other procedures are based on 100 replications. In the δ-slippage configuration, $\mu_{[1]} = \mu_{[9]} = \mu_{[10]} - \delta$, results for (B), taken from BKS Table 18.4.5, were based on 800 replications; results for (C) and (D), taken from TKS Table II, were based on 200 replications; and results for (E), (F) and (G) are based on 500 replications. In the equally spaced configuration, $\mu_{[i+1]} = \mu_{[i]} + \delta$ $(1 \leq i \leq 9)$, results for (B) and (D), taken from TKS Table III, were based on 200 replications; results for (E), (F) and (G) are based on 500 replications. Details of sample sizes obtained on individual populations can be found in JJT, Table III. The TKS RAND Q rule in column (C) is included only for the slippage configuration, since only in this case does it guarantee the PCS requirement.

The table shows the average sample number (ASN), the average inferior treatment number (ITN) and the proportion of correct selections made. The inferior treatment number is the total number of observations taken on all populations other than the best; in medical applications this is the number of patients who are not given the best treatment so a low ITN is desirable. The entry in parentheses in any cell is the standard error of the estimate above it. The asymptotic mean path approximations are also shown.

In this limited Monte Carlo study, for Paulson's region, $\sqrt{k_\nu - 1}$ sampling appears significantly better than VT in all three configurations. For Schwarz's region, $\sqrt{k_\nu - 1}$ sampling is significantly better in the slippage configuration and has approximately the same ASN as VT in the other two configurations. The ITN is also smaller for the adaptive procedures. Note that the achieved proportion of correct selections is greater than 0.9 in the slippage and the equally spaced configurations.

Table 1.

Simulation Results

$k = 10,\ P^* = 0.9,\ \delta = 0.2,\ \sigma^2 = 1$

	(A)	(B)	(C)	(D)	(E)	(F)	(G)
Stopping and Elimination Rule	FIXED	BKS	TKS	PF(.25)	PF(.25)	Schwarz	Schwarz
Sampling Rule		VT		VT	$\sqrt{k_\nu - 1}$	VT	$\sqrt{k_\nu - 1}$
Equal Means							
ASN: Estd	2230	2906		1643	1528	1708	1701
(s.e.)		(78)		(37)	(37)	(53)	(51)
δ-Slippage Configuration							
ASN: Estd	2230	1453	732	1155	1047	1149	1052
(s.e.)		(25)	(32)	(24)	(21)	(19)	(18)
MPA		1099		1215	972	1328	1062
ITN: Estd	2007	1308	552	979	846	955	844
MPA		989		1094	729	1195	797
Proportion of Correct Selections	0.90	0.911	0.910	0.925	0.956	0.918	0.912
		(0.010)	(0.020)	(0.019)	(0.009)	(0.012)	(0.013)
Equally Spaced Means							
ASN: Estd	2230	648		491	431	365	360
(s.e.)		(28)		(9)	(6)	(7)	(8)
MPA		576		495	435	425	390
ITN: Estd	2007	583		370	308	247	234
MPA		518		374	314	292	257
Proportion of Correct Selections		0.950		0.975	0.992	0.978	0.988
		(0.015)		(0.011)	(0.004)	(0.007)	(0.005)

KEY: s.e., standard error; MPA, mean path approximation.

REFERENCES

[1] 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.

[2] Bechhofer, R. E. (1969). Optimal allocation of observations when comparing several treatments with a control. In *Multivariate Analysis II*, P. R. Krishnaiah, ed. Academic Press, New York.

[3] Bechhofer, R. E., Kiefer, J. and Sobel, M. (1968). *Sequential Identification and Ranking Procedures.* Univ. of Chicago Press, Chicago.

[4] Berk, R. H. (1978). Asymptotic efficiencies of sequential tests II. *Ann. Statist.* 6, 813-819.

[5] Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Assoc.* 50, 1096-1121.

[6] Fabian, V. (1962). On multiple decision methods for ranking population means. *Ann. Math. Statist.* 33, 248-254,

[7] Fabian, V. (1974). Note on Anderson's sequential procedures with triangular boundary. *Ann. Statist.* 2, 170-176.

[8] Jennison, C. (1981). Sequential methods for normal populations when the variance is unknown. Unpublished manuscript.

[9] Jennison, C., Johnstone, I.M. and Turnbull, B. W. (1980). Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. Technical report No. 463, School of Operations Research and Industrial Engineering, Cornell University.

[10] Kao, S. C. and Lai, T. L. (1980). Sequential selection procedures based on confidence sequences for normal populations. *Comm. Statist.-Theory Methods* A9(16), 1657-1676.

[11] Lawing, W. D., and David, H. T. (1966). Likelihood ratio computations of operating characteristics. *Ann. Math. Statist.* 37, 1704-1716.

[12] Paulson, E. (1964). A sequential procedures for selecting the population with the largest mean from k normal populations. *Ann. Math. Statist.* 35, 174-180.

[13] Perng, S. K. (1969). A comparison of the asymptotic expected sample sizes of two sequential procedures for ranking problem. *Ann. Math. Statist. 40*, 2198-2202.

[14] Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Ann. Math. Statist. 41*, 1397-1409.

[15] Robbins, H. and Siegmund, D. O. (1974). Sequential tests involving two populations. *J. Amer. Statist. Assoc. 69*, 132-139.

[16] Schwarz, G. (1962). Asymptotic shapes of Bayes sequential testing regions. *Ann. Math. Statist. 33*, 224-236.

[17] Swanepoel, J. W. H. and Geertsema, J. C. (1976). Sequential procedures with elimination for selecting the best of k normal populations. *South African Statist. J. 10*, 9-36.

[18] Turnbull, B. W., Kaspi, H. and Smith, R. L. (1978). Adaptive sequential procedures for selecting the best of several normal populations. *J. Statist. Comput. Simulation 7*, 133-150.

[19] Woodroofe, M. (1976). Frequentist properties of Bayesian sequential tests. *Biometrika 63*, 101-110.