SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK

TECHNICAL REPORT NO. 463
(Preliminary Report)

June 1980

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

by

Chris Jennison, Iain M. Johnstone
and
Bruce W. Turnbull

This research supported in part by HEW PHS Grant No. R23 ES01714.

Presented at 173rd Meeting of the Institute of Mathematical Statistics, Davis, California, June 16-18, 1980.

§ 1 INTRODUCTION Suppose we have k ($\geq$ 2) normal populations with common known variance and unknown means. Without loss of generality we can take the common variance to be 1. We wish to select the population with the highest mean, call this the best population. We want to guarantee the probability of correct selection (PCS) to be at least P* ($\frac{1}{k} < P^* < 1$) whenever the mean of the best population exceeds that of the second best by at least δ. This is an example of the indifference zone approach to selection problems, where the experimenter gives a small difference in population means, δ, which it is not worthwhile trying to detect.

The problem was first formulated and solved by Bechhofer (1954). He gave a fixed sample size procedure in which an equal number of observations is taken from each population. Just as in one population problems, a reduction in average sample size can be achieved by using a sequential procedure. It also seems feasible to reduce sample size by sampling unequally from the k populations. Such sampling is called data dependent or adaptive sampling and the reductions it can achieve in sample size will become apparent.

Paulson (1964) gave a sequential procedure which has the feature of elimination. During the course of an experiment populations are successively eliminated from consideration and observations are allocated equally among the remaining populations. Fabian (1974), using a likelihood ratio method employed in Lawing and David (1966), gave a better lower bound for the guaranteed PCS than that proposed by Paulson. Swanepoel and Geertsema (1976) use a result of Robbins (1970) to give a modified version of Paulson's procedure with a smaller average sample size.

Bechhofer, Kiefer and Sobel (1968, chapter 3) 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. As long as observations are taken equally from all k populations the identification problem approach leads to a solution of our selection problem. We call such sampling "vector at a time sampling" (VT). If adaptive sampling is used it is difficult to determine the PCS guaranteed since the least favorable configuration may not be the usual one, and may depend on the sampling rule employed.

Turnbull, Kaspi and Smith (1977) investigated various sampling rules for the identification problem. However, their procedures do not give a solution to our problem for anything other than vector at a time sampling since they do not satisfy the PCS requirements. As in the case of Bechhofer, Kiefer and Sobel's procedure it is difficult to determine the least favorable configuration when adaptive sampling is used.

The procedures we consider are similar to Paulson's procedure, which latter we now describe. The populations are compared in pairs. When the observations on one population are sufficiently smaller than those on another the first one is eliminated. Uneliminated populations are sampled equally. Paulson noted that to guarantee a PCS of P* it is sufficient to ensure that P{A fixed inferior population eliminates the best population} $\leq$ $(1 - P^*)/(k - 1)$, and this is a relatively simple requirement as it involves observations on only two populations. Denote observations on two populations X and Y by $\{X_i\}$ and $\{Y_i\}$, $(i=1,2,\ldots)$.

While neither population has been eliminated there will be an equal number of observations from each. When this number is n, define the statistic

$$S_n = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i. \qquad (1.1)$$

If (S_n, n) lies outside a triangular continuation region we eliminate the population with the smaller sum of observations. Fabian proposed the following method of calculating an approximation to the error probability. Let δ_{XY} = mean of population X - mean of population Y. The joint distribution of the $\{S_n\}$ is the same as that of a Brownian motion with drift $\frac{1}{2}\delta_{XY}$ per unit time, observed at times $\{2n\}$. Hence the behaviour of S_n can be approximated by a Brownian motion with continuous time parameter and error probabilities may be calculated accordingly.

Our first generalization of Paulson's procedure is to allow an arbitrary continuation region, as long as the procedure terminates with probability one for any configuration of the population means. The square root region of Schwarz (1962) and a region proposed by Swanepoel and Geertsema (1976) are both asymptotically optimal in the sense of minimizing total expected sample size (ASN) as $P^* \to 1$. This remains true for the procedures with adaptive sampling which we develop.

In order to deal with adaptive sampling we need a statistic to replace S_n. At a particular point in the experiment denote the number of X and Y observations by m and n respectively. Set

$$\bar{X}(m) = \frac{1}{m} \sum_{i=1}^{m} X_i$$

$$\bar{Y}(n) = \frac{1}{n} \sum_{i=1}^{n} Y_i .$$

A natural statistic to consider is

$$Z(m,n) = \frac{mn}{m+n} (\bar{X}(m) - \bar{Y}(n)) . \qquad (1.2)$$

Robbins and Siegmund (1974) show that for the two population problem with certain sampling rules the random sequences $\{Z(m,n)\}$ have the same joint distribution as a Brownian motion with drift δ_{XY} observed at times $\{\frac{mn}{m+n}\}$. The restriction on the sampling rules is that at each stage the choice of which population to sample next must depend only on the differences between observations, $x_i - y_j$. Equivalently the choice must be independent of

$$T = \sum_{i=1}^{m} X_i + \sum_{i=1}^{n} Y_i .$$

In view of the translation invariant structure, this is a reasonable restriction for the two population problem and there are plenty of adaptive procedures allowed. However, roughly speaking, in the k population problem ($k > 2$) a sampling rule which satisfies the condition for all pairs of populations must be independent of the sample mean for each population. Since the motivation for adaptive sampling rules 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 of statistics $\{Z(m,n)\}$ corresponding to a comparison of a given pair of the populations is not simple, nor is $Z(m,n)$ a sufficient statistic for δ_{XY}. For any observation, X_i say, future values of its coefficient in the expression for $Z(m,n)$, (1.1), depend on the sampling process, which in turn depends on X_i. For explicit examples of the difficulties that can arise in the k-population case, see Appendices 1 and 2.

In view of the difficulty in rigorously applying a Brownian approximation to statistics of the form $Z(m,n)$, two approaches are adopted in this report. The first, described in Sections 2 to 5, proposes a class of statistics for which the Brownian approximation is much easier to justify, although the degree of "adaptiveness" permitted for the sampling rules is restricted somewhat. A more satisfactory mathematical asymptotic theory is then possible, but only preliminary simulation results are available at this stage. The second approach, discussed in Sections 6 and 7, uses the "natural statistics" of the form $Z(m,n)$ suggested by analogy with the two population case. In the absence of a convincing Brownian approximation, PCS guarantees are not available. Fully adaptive sampling rules are allowed however, and heuristic mean path approximations can be used to compare the efficiencies of various appealing sampling rules for any given stopping region. Simulation results bear out the conclusions of these calculations and in view of the attained PCS values, suggest that the Brownian approximation may in fact be reasonable for the sampling rules used in practice.

The procedures defined in Section 2 may be briefly described as follows. We divide the experiment into stages, the sampling mechanism and overall length of each stage is determined at the end of the previous

stage. Statistics of the same form as the $\{Z(m,n)\}$ are calculated separately from observations in each stage. Summing these over all the stages so far gives a sequence of statistics whose joint distribution is approximately that of a Brownian motion with drift δ_{XY} observed at appropriate times. The coefficient of any observation in this new statistic is determined before the observation is taken and it is therefore independent of the value of the observation.

We consider the class of tests based on this multistage statistic with arbitrary stopping region and stage-wise adaptive sampling, as long as the test terminates with probability one for all configurations of the population means. Included as special cases are procedures with deterministic (i.e. specified in advance) sampling rules, in which case there is only one stage and our statistics are the original $\{Z(m,n)\}$. Within this sub-class are procedures using vector at a time sampling such as Paulson's procedure or Swanepoël and Geertsema's procedure.

Another interesting sub-class 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 thus obtained to decide on the proportions in which to sample during the main part of the experiment. These two stage experiments have the advantage of being simple to apply, and some preliminary simulation results are presented in Section 5. A point on which some work remains to be done is the choice of initial sample size. There is a loose analogy between these two stage procedures and the more classical two stage procedures proposed for estimation and testing of normal means when the variance is unknown (e.g., Stein (1945), Dudewicz and Dalal (1975), Rinott (1978), Mukhopadhyay (1979). Indeed it is likely that our approach could be extended to give adaptive

sampling rules for selecting the best normal mean in the case of unknown variances.

In Section 3 we obtain an asymptotic lower bound for the efficacy ([Average Sample Number] ÷ [-log(error probability)]) of any stopping region and stagewise adaptive sampling rule as $P^* \to 1$. This is based on a corresponding result for the exit time of a Brownian motion from an arbitrary region (with prescribed error probability $\frac{1-P^*}{k-1}$) symmetric about the horizontal axis.

In Section 4 it is shown that this lower bound is sharp: it may be attained asymptotically by appropriate two-stage sampling rules in combination with the stopping region proposed by Schwarz (1962), which is well known to be optimal in various sequential Bayesian settings. For this argument, we modify a result of Berk (1978) on asymptotic error probabilities for random walks to the Brownian motion case. Section 5 contains discussion of some other aspects of multistage procedures and some simulation results.

In Section 6, a fully adaptive sampling rule is proposed for use with the "natural statistics" $Z_{ij}(m_i,m_j)$, $1 \leq i \leq j \leq k$, which is simple to use and has good efficiency properties. Essentially, it tries to ensure for any epoch N at which k_N populations are uneliminated, that the numbers of observations so far on each of the surviving populations are in the ratio $\sqrt{k_N-1}$: 1 ... :1, with the largest number of observations belonging to the population currently estimated as best. This rule is motivated by optimal allocation results for comparison of k normal treatments with a control (Dunnett (1955), Bechhofer (1969)). It is shown for any given region that this sampling rule is asymptotically optimal in slippage

configurations of the population means, and that for any configuration in the preference zone it dominates the corresponding procedure based on vector at a time sampling. An asymptotic lower bound is derived for the efficacy of various sampling rules for a given stopping region, and preliminary computations indicate that the $\sqrt{k_N - 1}$ rule achieves greater than 90% efficiency relative to this lower bound for reasonable configurations of means. It is also of interest to consider (and to try to minimise) the total number of observations on the 'inferior' populations (i.e. all except the one selected as 'best') -- this is referred to as the inferior treatment number (ITN). We show that the above "ASN" results have analogues for the ITN problem. Finally, Section 7 reports on the encouraging simulation results obtained using the $\sqrt{k_N - 1}$ rule and the "natural" statistics.

§ 2 GENERAL FORM OF THE MULTISTAGE PROCEDURES Observations X_{ip} ($1 \leq i \leq k$, $p \geq 1$) are available from each of k ($k \geq 2$) populations $\pi_1, \pi_2, \ldots, \pi_k$. If $k = 2$ the extra structure necessary to accommodate adaptive sampling is not required (for the purpose of minimising ASN) and the procedures can be simplified. The observations $\{X_{ip}\}$ are independently distributed normal random variables with unknown means μ_i and common known variance which we take to be 1. When a total of N observations have been taken we say we are at epoch N. The experiment is divided into stages and we denote the current stage by s(N), or for simplicity just by s. We relabel the observations as X_{irp} ($1 \leq i \leq k$, $r \geq 1$, $1 \leq p \leq M_{ir}$) where r denotes the stage in which an observation is taken and p its order in that stage. Let $m_{is}(N)$ be the number of observations taken on population π_i in the current stage s, at epoch N. For $r < s$ let M_{ir} be the total number of observations taken from π_i in a previous stage r. If π_i and π_j have not been eliminated at epoch N define

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

$$\bar{X}_{ir} = \frac{1}{M_{ir}} \sum_{p=1}^{M_{ir}} X_{irp} , \qquad (r < s)$$

$$Z_{ij}(N) = \sum_{r=1}^{s-1} \frac{M_{ir} M_{jr}}{M_{ir}+M_{jr}} (\bar{X}_{ir}-\bar{X}_{jr}) + \frac{m_{is}(N)\, m_{js}(N)}{m_{is}(N)+m_{js}(N)} (\bar{X}_{is}(N)-\bar{X}_{js}(N)) ,$$

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

Elimination Rule At epoch N-1 let $I_{N-1} = \{i: \pi_i \text{ not yet eliminated}\}$. At epoch N eliminate all populations π_j, $j \in I_{N-1}$ for which there is an $i \in I_{N-1}$ with

$$Z_{ij}(N) \geq g\{t_{ij}(N)\}, \tag{2.1}$$

where g is a non-negative function depending on P*. 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.

Stopping Region Plotting $z_{ij}(N)$ against $t_{ij}(N)$ we see that (2.1) holds if $Z_{ij}(N)$ lies above $g\{t_{ij}(N)\}$. If $Z_{ij}(N) \leq -g\{t_{ij}(N)\}$ 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 $Z = g(t)$, $Z = -g(t)$, $t \geq 0$ as the continuation region and its complement as the stopping region. Denote the continuation region 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. 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 Let

$$\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[k-1]} \leq \mu_{[k]}$$

denote the ordered means. We consider only the case where $\mu_{[k-1]} \leq \mu_{[k]}-\delta$ as only then do we have to guarantee a PCS. Now

$$P\{\text{Incorrect selection}\} = P\{\pi_{[k]} \text{ is eliminated at some point}\}$$

$$= P\{\bigcup_{i\neq[k]} (\pi_i \text{ eliminates } \pi_{[k]})\}$$

$$\leq \sum_{i\neq[k]} P\{Z_{i[k]}(N) \text{ exits C upwards}\}. \quad (2.2)$$

Here $\pi_{[k]} = \pi_j$, $Z_{i[k]} = Z_{ij}$, $t_{i[k]} = t_{ij}$, where $\mu_j = \mu_{[k]}$. By exiting C upwards we mean that the first point at which $(Z_{i[k]}(N), t_{i[k]}(N))$ lies outside C is in the upper half plane $\{Z_{i[k]}(N) \geq 0\}$. Since π_i or $\pi_{[k]}$ may be eliminated by a third population there is a positive probability that $Z_{i[k]}(N)$ remains in C throughout the experiment.

Fix i and j and let $\delta_{ij} = \mu_i-\mu_j$. It follows, as in Robbins and Siegmund (1974), that $\{Z_{ij}(N); N=1,2,\ldots\}$ have the same joint distribution

as a Brownian motion with drift δ_{ij} (per unit time), $B_{ij}(t)$ say, observed at times $\{t_{ij}(N)\}$. We say that Z_{ij} can be embedded in B_{ij} and we can regard the $Z_{ij}(N)$ as being generated by observing B_{ij} at the random times $\{t_{ij}(N)\}$.

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 Brownian motion with drift $-\Delta$ ($\Delta > 0$) and $\{Z_{-\Delta}(t_\alpha)\}$ are the values of a Brownian motion with drift $-\Delta$ observed at the sequence of times $\{t_\alpha;\ \alpha = 1,2,\ldots\}$. If the increments $\{t_\alpha - t_{\alpha-1}\}$ are small, then

$$P\{Z_{-\Delta}(t_\alpha) \text{ exits C upwards}\} \simeq P\{B_{-\Delta}(t) \text{ exits C upwards}\}. \quad (2.3)$$

Since the exit time of the discrete process is stochastically larger than that of the continuous process and the drift is negative, it is reasonable that

$$P\{Z_{-\Delta}(t_\alpha) \text{ exits C upwards}\} \simeq P\{B_{-\Delta}(t) \text{ exits C upwards}\}. \quad (2.4)$$

This was noted by Anderson (1960) and we give a proof in Appendix 3 for certain types of random sequence $\{t_\alpha\}$. In our context the proof does not apply, and we use the approximation (2.3). Let $-\Delta = \delta_{i[k]}$, then $\Delta \geq \delta$. Since $Z_{i[k]}$ may not exit C at all, (2.3) implies

$$P\{Z_{i[k]}(n) \text{ exits C upwards}\} \lesssim P\{B_{-\Delta}(t) \text{ exits C upwards}\}$$

$$\leq P\{B_{-\delta}(t) \text{ exits C upwards}\}. \quad (2.5)$$

From (2.2) and (2.5) it follows that PCS $\geq$ P* is guaranteed approximately if

$$P\{B_{-\delta}(t) \text{ exits } C \text{ upwards}\} \leq \frac{1-P^*}{k-1} , \qquad (2.6)$$

where $B_{-\delta}(t)$ is a Brownian motion with drift $-\delta$.

Since the smallest value t_{ij} can take is $\frac{1}{2}$ we need only consider the behaviour of $B_{-\delta}(t)$ on $[\frac{1}{2},\infty)$ and (2.6) becomes

$$P\{B_{-\delta}(\tau) \geq g(\tau)\} \leq \frac{1-P^*}{k-1} , \qquad (2.7)$$

where $\tau_{-\delta} = \{\inf t \geq \frac{1}{2}: B_{-\delta}(t) \notin C\}$ and τ_{Δ} is assumed finite a.s. for all $\Delta \in R$. We shall make use of this fact later to avoid difficulties near t = 0.

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.

In the next three sections we shall consider multi-stage procedures of the form described in this section. Such procedures are defined by a sampling rule and a non-negative function g(t) for which (2.7) holds. Let the class of all such procedures be C and let $C(\delta,P^*)$ be the sub-class of procedures guaranteeing PCS $\geq$ P* for indifference parameter δ.

§ 3 AN ASYMPTOTIC LOWER BOUND FOR EFFICACY For a fixed $\delta > 0$ we investigate the minimum expected sample sizes as $P^* \to 1$. Let $B_\Delta(t)$ be a standard Brownian motion with drift Δ and let P_Δ denote the probability measure on the space of paths $B(t)$, generated by $B_\Delta(t)$. Set $\varepsilon = \frac{1-P^*}{k-1}$. For each P^* we have a non-negative function $g_\varepsilon(t)$, $t \geq 0$, and the region bounded by $B = g_\varepsilon(t)$, $-B = -g_\varepsilon(t)$ and $t = 0$ is the continuation region C_ε. Setting $\tau_\varepsilon = \inf\{t \geq \frac{1}{2}: |B_\Delta(t)| \geq g_\varepsilon(t)\}$, g_ε must be such that $P_\Delta\{\tau_\varepsilon = \infty\} = 0$ for all $\Delta \in R$ and $P_\delta\{B(\tau_\varepsilon) \leq -g\} \leq \varepsilon$. The first lemma refers to a general property concerning Brownian motion.

Lemma 3.1 Let $B_\Delta(t)$, $g_\varepsilon(t)$, C_ε and τ_ε be as above, $\Delta \in R$. Then given any $T > 0$,

$$P_\Delta(\tau_\varepsilon < T) \to 0 \quad \text{as } \varepsilon \to 0.$$

Proof Case 1, $\Delta = \delta$. For any $u > 0$

$$P_\delta\{\tau_\varepsilon < T\} = P_\delta(\Omega_1) + P_\delta(\Omega_2) + P_\delta(\Omega_3),$$

where

$$\Omega_1 = \{B(t): \tau_\varepsilon < T,\ B(\tau_\varepsilon) \geq g,\ B(\tau_\varepsilon) \geq u\},$$

$$\Omega_2 = \{B(t): \tau_\varepsilon < T,\ B(\tau_\varepsilon) \geq g,\ B(\tau_\varepsilon) < u\},$$

and $$\Omega_3 = \{B(t): \tau_\varepsilon < T,\ B(\tau_\varepsilon) \leq -g\}.$$

Now,

$$P_\delta(\Omega_1) \le P_\delta(\sup\{B(t): t \le T\} \ge u),$$

$$P_\delta(\Omega_2) = \int_{\Omega_2} dP_\delta = \int_{\Omega_2} \frac{dP_\delta}{dP_{-\delta}}\, dP_{-\delta}$$

$$= \int_{\Omega_2} \exp\{2\delta B(\tau_\varepsilon)\}\, dP_{-\delta} < \exp(2\delta u) P_{-\delta}\{B(\tau_\varepsilon) \ge g\}$$

$$= \exp(2\delta u) P_\delta\{B(\tau_\varepsilon) \le -g\} \le \varepsilon \exp(2\delta u) \text{ , and}$$

$$P_\delta(\Omega_3) \le P_\delta\{B(\tau_\varepsilon) \le -g\} \le \varepsilon.$$

Hence,

$$P_\delta\{\tau_\varepsilon < T\} \le P_\delta(\sup\{B(t): t \le T\} \ge u) + (1 + \exp(2\delta u))\, \varepsilon .$$

Choose $u = u^\circ$ such that the first term $< \frac{\eta}{2}$, then choose ε° such that the second term $< \frac{\eta}{2}$ for $u = u^\circ$ and $\varepsilon \le \varepsilon^\circ$. Then

$$P_\delta\{\tau_\varepsilon < T\} < \eta \text{ for } \varepsilon \le \varepsilon^\circ.$$

Case 2, $\Delta \neq \delta$. For any $u > 0$

$$P_\Delta\{\tau_\varepsilon < T\} = P_\Delta(\Omega_4) + P_\Delta(\Omega_5) ,$$

where

$$\Omega_4 = \{B(t): \tau_\varepsilon < T,\ |B(\tau_\varepsilon)| \ge u\} ,$$

$$\Omega_5 = \{B(t): \tau_\varepsilon < T, \ |B(\tau_\varepsilon)| < u\}.$$

Now,

$$P_\Delta(\Omega_4) \leq P_\Delta(\sup\{B(t): t < T\} \geq u)$$

$$P_\Delta(\Omega_5) = \int_{\Omega_5} dP_\Delta = \int_{\Omega_5} \frac{dP_\Delta}{dP_\delta} \, dP_\delta$$

$$= \int_{\Omega_5} \exp\{(\Delta-\delta)B(\tau_\varepsilon) - \frac{(\Delta^2-\delta^2)}{2}\tau_\varepsilon\} \, dP_\delta$$

$$\leq \exp\{|\Delta-\delta| \ u + \tfrac{1}{2}|\Delta^2-\delta^2| \ T\} \ P_\delta(\tau_\varepsilon < T).$$

Choose $u = u^\circ$ such that $P_\Delta(\Omega_4) < \frac{\eta}{2}$. The result for case 1 allows us to choose ε° such that $P_\Delta(\Omega_5) < \frac{\eta}{2}$ for $u = u^\circ$, $\varepsilon \leq \varepsilon^\circ$. Then

$$P_\Delta\{\tau_\varepsilon < T\} < \eta \text{ for } \varepsilon \leq \varepsilon^\circ. \qquad \square$$

We now derive an asymptotic property of the stopping time, τ_ε, as $\varepsilon \to 0$.

Theorem 3.2 Let $B_\Delta(t)$, $g_\varepsilon(t)$, C_ε and τ_ε be as above, $\Delta \in R$. Then

$$P_\Delta\{\frac{\tau_\varepsilon}{-\log\varepsilon} \leq \frac{2}{(|\Delta|+\delta)^2} - \nu\} \to 0 \text{ as } \varepsilon \to 0 \ , \tag{3.1}$$

where $\nu > 0$ is as small as we please.

Proof Case 1, $\Delta > 0$. Let $\Omega_1 = \{B(t): B(\tau_\varepsilon) \geq g\}$, then

$$\varepsilon \geq P_{-\delta}(\Omega_1) = \int_{\Omega_1} dP_{-\delta} = \int_{\Omega_1} \frac{dP_{-\delta}}{dP_\Delta} \, dP_\Delta.$$

Hence

$$\int_{\Omega_1} \exp\{[\frac{-B(\tau_\varepsilon)}{\tau_\varepsilon} \quad (\Delta+\delta) + \frac{(\Delta^2-\delta^2)}{2}]\cdot[\frac{\tau_\varepsilon}{-\log\varepsilon}]\cdot[-\log\varepsilon]\ dP_\Delta \leq \varepsilon. \quad (3.2)$$

Suppose there is a sequence $\{\varepsilon_i\}$, $\varepsilon_i \to 0$ and $\eta > 0$ such that

$$P_\Delta\{\frac{\tau_{\varepsilon_i}}{-\log\varepsilon_i} < A\} > \eta \text{ for all } i \ . \quad (3.3)$$

Given $\lambda > 0$, $\xi > 0$ $\exists t^o(\lambda,\xi)$ such that

$$P_\Delta\{\frac{B(t)}{t} \quad < \Delta+\lambda\} \geq 1 - \frac{\xi}{2} \quad \text{for } t \geq t^o \ .$$

By Lemma 3.1, $\exists \varepsilon^o(\lambda,\xi)$ such that

$$P_\Delta\{\tau_\varepsilon \geq t^o\} \geq 1 - \frac{\xi}{2} \qquad \text{for } \varepsilon \leq \varepsilon^o,$$

hence

$$P_\Delta\{\frac{B(\tau_\varepsilon)}{\tau_\varepsilon} < \Delta+\lambda\} \geq 1 - \xi \quad \text{for } \varepsilon \leq \varepsilon^o \ . \quad (3.4)$$

Since $\Delta > 0$ it follows from the lemma that $P_\Delta\{B(\tau_\varepsilon) \geq g\} \to 1$ as $\varepsilon \to 0$. Take $\varepsilon^1(\eta)$ such that

$$P_\Delta\{B(\tau_\varepsilon) \geq g\} \geq 1 - \frac{\eta}{3} \qquad \text{for } \varepsilon \leq \varepsilon^1 \ . \quad (3.5)$$

Putting $\xi = \frac{\eta}{3}$, (3.3), (3.4) and (3.5) imply

$$P_\Delta\{B(t): \frac{\tau_{\varepsilon_i}}{-\log\varepsilon_i} < A, \frac{B(\tau_{\varepsilon_i})}{\tau_{\varepsilon_i}} < \Delta+\lambda, B(\tau_{\varepsilon_i}) \geq g\} > \frac{\eta}{3}$$

for $\varepsilon_i \leq \min(\varepsilon^o, \varepsilon^1)$.

With (3.2) this implies

$$\frac{1}{3}\eta \exp\{[-(\Delta+\lambda)(\Delta+\lambda) + \frac{(\Delta^2-\delta^2)}{2}] A [-\log\varepsilon_i]\} \leq \varepsilon_i$$

for $\varepsilon_i \leq \min(\varepsilon^o, \varepsilon^1)$,

and letting $\varepsilon_i \to 0$ we see that

$$A \geq \frac{2}{(\Delta+\delta)^2+2\lambda(\Delta+\delta)} .$$

But λ was arbitrary, so

$$A \geq \frac{2}{(\Delta+\delta)^2} = \frac{2}{(|\Delta|+\delta)^2} ,$$

and (3.1) follows.

Case 2, $\Delta < 0$. Since the regions C_ε are symmetric about $B = 0$ the result follows from Case 1.

Case 3, $\Delta = 0$. As before

$$\int_{\Omega_1} \exp\left[\frac{-B(\tau_\varepsilon)\delta}{\tau_\varepsilon} - \frac{\delta^2}{2}\right]\cdot\left[\frac{\tau_\varepsilon}{-\log\varepsilon}\right]\cdot[-\log\varepsilon]\} \, dP_o \leq \varepsilon .$$

Under P_o the distribution of $B(t)$ is symmetric with respect to reflections in the axis $B = 0$. So

$$P_\Delta\{\frac{\tau_{\varepsilon_i}}{-\log\varepsilon_i} < A\} > \eta \quad \Rightarrow \quad P_\Delta\{B(t): \frac{\tau_{\varepsilon_i}}{-\log\varepsilon_i} < A,\ B(\tau_{\varepsilon_i}) \geq g\} > \frac{\eta}{2}\ .$$

For $\nu > 0$, $P\{\frac{-B(\tau_\varepsilon)}{\tau_\varepsilon} > -\nu\} \to 1$ as $\varepsilon \to 0$, and arguing as in case 1 gives the result. □

We now fix a value for δ, the indifference parameter and use Theorem 3.2 to derive an asymptotic lower bound for the average sample number (ASN) of the multistage procedures described in Section 2. In the asymptotic setting we consider sequences of procedures indexed by $\varepsilon = (1-P^*)/(k-1)$, each procedure guaranteeing PCS equal to the corresponding P^*. For a procedure in the class C of section 2, index the outcome of an experiment by ω and let Ω be the space of all possible outcomes. Define $h_i(\omega)$ to be the total number of observations taken on population π_i, over the whole experiment, and let $H(\omega)$ be the overall total. (Note that $H < \infty$ a.s. for all parameter configurations.) We shall frequently suppress the relation of random variables to ω. If π_i is eliminated or selected as best at epoch N_i, let $s(i) = s(N_i)$ be the stage during which this occurs $(i=1,2,\ldots,k)$. Then

$$h_i(\omega) = \sum_{r=1}^{s(i)-1} M_{ir} + m_{i,s(i)}(N_i) \qquad (i=1,2,\ldots,k)$$

and

$$H(\omega) = \sum_{i=1}^{k} h_i(\omega).$$

Let μ denote the vector of population means $(\mu_1,\mu_2,\ldots,\mu_k)$, with $\delta_{ij} = \mu_i-\mu_j$ as before. The average sample number of a procedure is defined to be $E_\mu(H)$. Indexing a sequence of procedures by ε, $E_{\mu,\varepsilon}(H)$ is the ASN for the procedure guaranteeing a PCS of P*, where $\varepsilon = (1-P^*)/(k-1)$. 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

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

Theorem 3.3 With the above notation, let the indifference parameter, δ, and the vector of population means, μ, be fixed. Then for any sequence of procedures indexed by ε

$$\lim_{\varepsilon\to 0} \{\frac{E_{\mu,\varepsilon}(H)}{-\log\varepsilon}\} \geq f_\delta(\mu) \; . \tag{3.6}$$

Before proving the theorem we note this is true for all μ, not just for those outside of the indifference region.

Proof Throughout the proof a subscript ε refers to the procedure which gives a PCS of P*, where $\varepsilon = (1-P^*)/(k-1)$. For instance, Ω_ε is the space of all possible outcomes of an experiment when the procedure indexed by ε is used. Recall that we say π_i eliminates π_j if π_j is eliminated at epoch N and $Z_{ij}(N) \geq g\{t_{ij}(N)\}$. If this does not uniquely identify an eliminating population then select one of the candidates arbitrarily. For a fixed value of $\eta > 0$, let $\Omega_{1\varepsilon}$ be the set of all

outcomes $\omega \in \Omega_\varepsilon$ such that

(i) If π_j eliminates π_i then $\mu_j \geq \mu_i$,

(ii) If π_j eliminates π_i then the time $T_{ji,\varepsilon}$ at which $(Z_{ji,\varepsilon}(N), t_{ji,\varepsilon}(N))$ exits C_ε satisfies

$$\frac{T_{ji,}}{-\log\varepsilon} > \frac{2}{(|\delta_{ji}|+\delta)^2+2\eta}.$$

Note that condition (i) implies a correct selection is made.

Now, if $\mu_j < \mu_i$

$$P\{Z_{ji,\varepsilon}(N) > 0\} \to 0 \text{ as } t_{ji,\varepsilon}(N) \to \infty,$$

and with Lemma 3.1 this implies $P\{\text{(i) holds}\} \to 1$ as $\varepsilon \to 0$.

For a fixed pair (i,j) we follow the discussion of section 2 and embed $\{Z_{ji,\varepsilon}\}$ in a Brownian motion with drift δ_{ji}. We are interested in the distribution of the exit time

$$\tau_{ji,\varepsilon} = \inf\{t \geq \frac{1}{2} : B_{\delta_{ji}}(t) \notin C_\varepsilon\}.$$

Since the exit time of the discrete process is stochastically larger than that of the continuous process, it follows from Theorem 3.2 that

$$P\{\pi_j \text{ eliminates } \pi_i \text{ and } \frac{T_{ji,\varepsilon}}{-\log\varepsilon} \leq \frac{2}{(|\delta_{ij}|+\delta)^2+2\eta}\} \to 0 \text{ as } \varepsilon \to 0.$$

Summing over pairs (i,j) we get P{(ii) holds} → 1 as ε → 0. Hence

$$P\{\Omega_{1\varepsilon}\} \to 1 \text{ as } \varepsilon \to 0.$$

Choose ε° such that $P\{\Omega_{1\varepsilon}\} > 1-\eta$ for $\varepsilon \leq \varepsilon^{\circ}$. We have

$$E_{\mu,\varepsilon}(H) \geq \int_{\Omega_{1\varepsilon}} H(\omega)\, P_{\mu,\varepsilon}\,(d\omega) \tag{3.7}$$

and we shall bound $H(\omega)$ below on $\Omega_{1\varepsilon}$. Suppose at epoch N we have a total of $\ell_i(N)$ observations on π_i (i=1,2,...,k), then

$$t_{ji,\varepsilon}(N) \leq \frac{\ell_i(N)\,\ell_j(N)}{\ell_i(N)+\ell_j(N)} \qquad (i\neq j) \tag{3.8}$$

Hence, if π_j eliminates π_i

$$T_{ji,\varepsilon}(\omega) \leq \frac{h_i(\omega)\,h_j(\omega)}{h_i(\omega)+h_j(\omega)} ,$$

or equivalently

$$\frac{1}{h_i} + \frac{1}{h_j} \leq \frac{1}{T_{ji,\varepsilon}}$$

and on $\Omega_{1\varepsilon}$ we have

$$\frac{1}{h_i} + \frac{1}{h_j} \leq (-\log\varepsilon)^{-1}\{\frac{(|\delta_{ji}|+\delta)^2}{2} + \eta\}.$$

But on $\Omega_{1\varepsilon}$, $\pi_{[k]}$ is selected as best (or one of the other populations with mean $\mu_{[k]}$ is selected if there are several). Also, condition (i) implies $\delta_{ji} \geq 0$ if π_j eliminates π_i, so $|\delta_{ji}| = \delta_{ji}$ and we can use the result of Appendix 4 to conclude

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

$$= f_\delta(\mu;\eta), \text{ say, for } \omega \in \Omega_{1\varepsilon},\ \varepsilon \leq \varepsilon^0.$$

With (3.7) this gives

$$\frac{E_{\mu,\varepsilon}(H)}{-\log\varepsilon} \geq (1-\eta)\, f_\delta(\mu;\eta) \qquad \text{for } \varepsilon \leq \varepsilon^0,$$

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

We note that we have proved a stronger result than that stated in the theorem, namely that for any $\nu > 0$, $\underline{\lim}\ P_\varepsilon[\{ASN/(-\log\varepsilon)\} \geq f_\delta(\mu)-\nu] = 1$. The proof of the theorem suggests how we should try to achieve the lower bound asymptotically. The $\{d_i\}$ corresponding to the infimum for $f_\delta(\mu)$ give optimal ratios in which to sample from the k populations. So, if we knew the means $\{\mu_i\}$ we would know how to minimise the ASN (although of course there would be no need to run an experiment). By estimating the means early in an experiment and sampling so as to mimic the procedure which would be optimal if these were the true means we can achieve asymptotic optimality.

§ 4 ASYMPTOTICALLY OPTIMAL TWO STAGE PROCEDURES We first derive an asymptotic lower bound for the probability of a Brownian motion exiting Schwarz's region through the lower boundary. The method of proof works for more general regions. The theorem is a continuous time version of a theorem by Berk (1978). When the indifference parameter is δ there is a family of Schwarz regions indexed by a,

$$g_a(t) = \begin{cases} \sqrt{2at} - \delta t & 0 \leq t \leq \frac{2a}{\delta^2} \\ 0 & t > \frac{2a}{\delta^2} \end{cases} .$$

In this section we retain the notation of sections 2 and 3, although we shall usually index by a instead of ε. Let $B_{-\delta}(t)$ be a Brownian motion with drift $-\delta$ and define

$$\tau_a = \inf\{t \geq \tfrac{1}{2}: |B_{-\delta}(t)| \geq g_a(t)\}.$$

Let $\varepsilon(a) = P\{B_{-\delta}(\tau_a) \geq g_a(\tau_a)\}$. To satisfy (2.7) we choose a such that $\varepsilon(a) = \varepsilon = (1-P^*)/(k-1)$. We note for later reference that $\log\varepsilon(a) \sim \log(1-P^*)$ as $a \to \infty$.

Theorem 4.1 For Schwarz's region, as described above, $\varepsilon(a)$ satisfies

$$\varepsilon(a) = \exp\{-a+o(a)\} \qquad \text{as } a \to \infty .$$

Proof $\varepsilon(a) = P\{B_{-\delta}(\tau_a) \geq g_a(\tau_a)\}$

$$\leq P\{B_{-\delta}(t) \geq g(t) \text{ for some } t \in [\tfrac{1}{2}, \tfrac{2a}{\delta^2}]\}$$

$$= P\{B_o(t) \geq \sqrt{2at} \text{ for some } t \in [\tfrac{1}{2}, \tfrac{2a}{\delta^2}]\},$$

where $B_o(t)$ is a Brownian motion with zero drift. For $0 < \gamma < 1$ we cover $[\frac{1}{2}, \frac{2a}{\delta^2}]$ with intervals of the form $[\gamma^q \frac{2a}{\delta^2}, \gamma^{q-1} \frac{2a}{\delta^2}]$, $q=1,2,\ldots,Q$, such that $[\gamma^Q 2a/\delta^2, \gamma^{Q-1} 2a/\delta^2]$ contains $\frac{1}{2}$. Hence

$$Q \leq \frac{\log a}{-\log \gamma} - \frac{\log(\delta^2/4)}{-\log \gamma} + 1.$$

Now,

$$P\{B_o(t) \geq \sqrt{2at} \quad \text{for some } t \in [\gamma^q \tfrac{2a}{\delta^2}, \gamma^{q-1} \tfrac{2a}{\delta^2}]\}$$

$$\leq P\{B_o(t) \geq \sqrt{2a\, \gamma^q \tfrac{2a}{\delta^2}} \quad \text{for some } t \in [0, \gamma^{q-1} 2a/\delta^2]\}$$

$$\leq 2P\{B_o(\gamma^{q-1} \tfrac{2a}{\delta^2}) \geq \tfrac{2a}{\delta} \sqrt{\gamma^q}\} \quad \text{(by the reflection principle)}$$

$$\leq 2P\{W \geq \sqrt{2a\gamma}\} \qquad (W \sim N(0,1))$$

$$\leq 2e^{-\gamma a},$$

since $\int_u^\infty e^{-w^2/2}\, dw \leq e^{-u^2/2}$ for $u \geq 0$. Summing over the Q intervals

$$\varepsilon(a) \leq (\frac{\log a}{-\log \gamma} - \frac{\log(\delta^2/4)}{-\log \gamma} + 1)\, 2e^{-\gamma a}$$

so for a fixed value of γ, $\log(\varepsilon(a)) \leq -\gamma a + o(a)$. But $\gamma \in (0,1)$ was arbitrary and hence

$$\log(\varepsilon(a) \leq -a + o(a).$$

For the reverse inequality, we note simply that

$$\varepsilon(a) \geq P\{B_{-\delta}(\tfrac{1}{2}) \geq \sqrt{a} - \delta/2\} \sim e^{-a+o(a)}.$$

Berk proved this result for random walk (i.e. when the Brownian motion is only observed at integral values of t). This agreement between the continuous and discrete time results supports our approximation (2.3). To achieve the asymptotic lower bound of Theorem 3.3 we propose the following two-stage procedures.

Procedures are in the class C of section 2 and the stopping regions are given by Schwarz's region with a value of a chosen to give the required PCS. The sampling rule is as follows: sample equally from all k populations during the first stage (unless, of course, a population is eliminated), then calculate optimal sampling ratios based on the sample means at the end of the first stage and use these during the second stage. Indexing procedures by a we have $M_{il} \leq n^{o}$ with equality unless π_i is eliminated during the first stage. Here n^{o} denotes the common first stage sample size. If π_i and π_j are uneliminated then $t_{ij}(N) = n^{o}/2$ at the end of stage one. Since n^{o} depends on a, set $\lambda_1(a) = \frac{n^{o}}{2}$ for use in Theorem 4.2. Let

$$\hat{\mu}_i = \frac{1}{M_{il}} \sum_{p=1}^{M_{il}} X_{ilp}$$

and let J be the set {i: π_i not eliminated during stage one}. Unless the experiment terminates during stage one, let $\hat{\mu}_{(k)} = \max(\hat{\mu}_i,\ i\epsilon J)$, and set $\hat{\delta}_{(k)i} = \hat{\mu}_{(k)} - \hat{\mu}_i$. Define

$$f_\delta(\hat{\mu};J) = \inf_{d_i>0} \left\{ \sum_{i\epsilon I} d_i : \frac{1}{d_i} + \frac{1}{d_{(k)}} \leq \frac{(\hat{\delta}_{(k)i} + \delta)^2}{2},\ i\epsilon J\backslash\{k\} \right\} \quad (4.0)$$

Remark 4.1. It might seem more efficient to take the length of the first stage into account in defining $f_\delta(\hat{\mu},J)$ and hence the "optimal" sampling ratios for the second stage. One would replace

$$\frac{(\hat{\delta}_{(k)i} + \delta)^2}{2}$$

by

$$\left[\frac{2a}{(\hat{\delta}_{(k)i} + \delta)^2} - \lambda_1(a)\right]^{-1}$$

in the definition of $f_\delta(\hat{\mu};J)$, but the proof is correct and simpler in the present formulation. (Note that the above expression is positive for populations i ϵ I).

Suppose $\{\overline{d}_i;\ i\epsilon J\}$ are d_i's which achieve the infimum, then during stage two we sample from the populations $\{\pi_i\}$ proportionally. If I(N) is the set of uneliminated populations and i ϵ I(N), then the proportion of future observations to be given to π_i is

$$R_i = \frac{\bar{d}_i}{\sum_{j \in I(N)} \bar{d}_j} .$$

Of course these proportions cannot be achieved exactly since the numbers of observations are integral. Asymptotically this effect is negligible so we shall simplify the notation by proceeding as if the numbers of observations are continuous.

Theorem 4.2 Let the indifference parameter, δ, and the vector of population means, μ, be fixed, and suppose that there is a unique best population. Consider a sequence of two-stage procedures of the above form with lengths of stage one given by $\{\lambda_1(a)\}$. If $\lambda_1(a) \to \infty$ and $\lambda_1(a)/a \to 0$ as $a \to \infty$, then

$$\overline{\lim_{\varepsilon \to 0}} \left\{ \frac{E_{\mu,\varepsilon}(H)}{-\log \varepsilon} \right\} \leq f_\delta(\mu).$$

Proof. We note that by Theorem 4.1 it is sufficient to prove

$$\overline{\lim_{a \to \infty}} \left\{ \frac{E_{\mu,a}(H)}{a} \right\} \leq f_\delta(\mu).$$

Denote the outcome of an experiment by ω and let Ω_a be the space of all possible outcomes for the procedure indexed by a. For a fixed value of $\xi > 0$ let Ω_{1a} be the set of all outcomes $\omega \in \Omega_a$ such that

(i) No eliminations take place in stage one.

(ii) $\pi_{[k]}$ has the largest sample mean at the end of stage one, so $\pi_{(k)} = \pi_{[k]}$ and $\hat{\mu}_{(k)} = \hat{\mu}_{[k]}$.

(iii) All populations are eliminated by $\pi_{[k]}$.

(iv) $f_\delta(\hat{\mu}) \leq (1+\xi) f_\delta(\mu)$.

(v) If π_i is eliminated by $\pi_{[k]}$ when $t_{[k]i,a}(N) = T_{[k]i,a}$, then

$$T_{[k]i,a} \leq \frac{2a}{(\hat{\delta}_{[k]i,a}+\delta)^2} \quad (1+\xi)$$

Using an argument similar to that of Theorem 4.1, the condition $\frac{\lambda_1(a)}{a} \to 0$ implies P{(i) holds} $\to 1$ as $a \to \infty$. Since $\lambda_1(a) \to \infty$ as $a \to \infty$ P{(ii) holds} $\to 1$. $\pi_{[k]}$ is selected with probability at least P*, the sampling rules are such that the largest number of observations is allocated to $\pi_{(k)}$, unless it is eliminated, and hence P{(iii) holds} $\to 1$. The function $f_\delta(\mu)$ is continuous in μ, $\hat{\mu}_a \to \mu$ in probability as $\lambda_1(a) \to \infty$ and $\frac{\lambda_1(a)}{a} \to 0$ and therefore P{(iv) holds} $\to 1$. When

$$t_{[k]i,a}(N) = \frac{2a}{(\delta_{[k]i,a}+\delta)^2} \quad (1 + \tfrac{1}{2}\xi)$$

we have

$$P\{Z_{[k]i,a}(N) \geq \sqrt{2a t_{[k]i,a}(N)} - \delta t_{[k]i,a}(N)\}$$

$$= P\{W \geq \sqrt{2a}\,(1 - \sqrt{1 + \tfrac{1}{2}\xi}\,)\},$$

where $W \sim N(0,1)$, and this probability $\to 1$ as $a \to \infty$. $\hat{\delta}_{[k]i,a} \to \delta_{[k]i}$ in probability as $\lambda_1(a) \to \infty$ and hence P{(v) holds} $\to 1$. Thus

$$P\{\Omega_{1a}\} \to 1 \quad \text{as } a \to \infty.$$

On Ω_{1a} the total number of observations in stage one is at most

$$2k\lambda_1(a) = o(a). \tag{4.1}$$

Since

$$\frac{M_{i2}M_{[k]2}}{M_{i2}+M_{[k]2}} < T_{[k]i,a} \leq \frac{2a(1+\xi)}{(\hat{\delta}_{[k]i,a}+\delta)^2},$$

the number of observations in stage two is no greater than

$$\inf_{d_i>0} \{ \sum_{i=1}^{k} d_i : \frac{1}{d_i} + \frac{1}{d_{[k]}} \leq \frac{(\hat{\delta}_{[k]i}+\delta)^2}{2a(1+\xi)} \} = (1+\xi)a\, f_\delta(\hat{\mu})$$

$$\leq (1+\xi)^2 a\, f_\delta(\mu), \qquad \text{by (iv)}. \tag{4.2}$$

This is so because on Ω_{1a}, $\hat{\delta}_{[k]i} = \hat{\delta}_{(k)i}$ and the terms $(\hat{\delta}_{[k]i}+\delta)^2/2a(1+\xi)$ are proportional to the terms $(\hat{\delta}_{(k)i}+\delta)^2/2$ used to find the $\{\bar{d}_i\}$, from which the sampling ratios were determined.

Let $\Omega_{2a} = \Omega_a \backslash \Omega_{1a}$. Recall that when the set of uneliminated populations is I and for $i \in I$

$$R_i = \frac{\bar{d}_i}{\sum_{j\in I} \bar{d}_j} \tag{4.3}$$

where the $\{\bar{d}_i\}$ minimise $\sum_{i\in J} d_i$ subject to

$$\frac{1}{d_i} + \frac{1}{d_{(k)}} \leq \frac{(\hat{\delta}_{(k)i}+\delta)^2}{2}, \quad i \in J\backslash\{(k)\}$$

$$d_i > 0, \quad i \in J. \tag{4.4}$$

J is the set of populations not eliminated during stage one. Hence

$$\overline{d}_i \geq \frac{2}{(\hat{\delta}_{(k)i}+\delta)^2}, \quad i \in J\backslash\{(k)\}. \tag{4.5}$$

Now $d_i = \frac{4}{\delta^2}$, $i \in J$, satisfies the conditions (4.4), and then $\sum_{i\in J} d_i \leq \frac{4k}{\delta^2}$. Therefore

$$\sum_{j\in I} \overline{d}_j \leq \frac{4k}{\delta^2},$$

and together with (4.3) and (4.5) this implies

$$R_i \geq \frac{2}{2k(\hat{\delta}_{(k)i}+\delta)^2}.$$

For $\omega \in \Omega_{2a}$ the total number of observations on any population is at most

$$\frac{2a}{\delta^2} \cdot 2\{\min_{i\in J} R_i\}^{-1} \leq \frac{8k}{\delta^4} a \max_{i\in J}\{(\hat{\delta}_{(k)i}+\delta)^2\}$$

$$\leq \frac{8k}{\delta^4} a \sum_{i,j\in J; i\neq j} (\hat{\delta}_{ji}+\delta)^2.$$

So denoting the probability distribution on Ω_a by P_a,

$$\int_{\Omega_{2a}} H(\omega)dP_a(\omega) \leq \frac{8k}{\delta^4}\, a \sum_{i,j} \int_{\Omega_{2a}} \Big[\frac{1}{2\lambda_1(a)} \Big(\sum_{p=1}^{2\lambda_1(a)} X_{jlp} - \sum_{p=1}^{2\lambda_1(a)} X_{ilp} \Big) + \delta\Big]^2 dP_a(\omega),$$

where the first summation is over all pairs (i,j); $i,j \in \{1,2,\ldots,k\}$, $i \neq j$. For the pair (i,j) the integrand has the same distribution as W^2 where $W \sim N(\mu_j - \mu_i + \delta, \frac{1}{2\lambda_1(a)})$. Since $\lambda_1(a) \to \infty$ and $P_a(\Omega_{2a}) \to 0$ as $a \to \infty$, the integrals $\to 0$. So

$$\int_{\Omega_{2a}} H(\omega)\, dP_a(\omega) = o(a). \tag{4.6}$$

From (4.1) and (4.2)

$$\int_{\Omega_{1a}} H(\omega)\, dP_a(\omega) \leq (1+\xi)^2 a\, f_\delta(\mu) + o(a)\ . \tag{4.7}$$

But $\Omega_a = \Omega_{1a} \cup \Omega_{2a}$ and ξ was arbitrary. Letting $\xi \to 0$, (4.6) and (4.7) imply

$$\overline{\lim_{a\to\infty}} \Big\{\frac{E_{\mu,a}(H)}{a}\Big\} \leq f_\delta(\mu)$$

and as we remarked earlier this establishes the result. □

Combining the results of Theorems 3.3 and 4.2 we obtain

$$\overline{\lim_{\varepsilon\to 0}} \Big\{\frac{E_{\mu,\varepsilon}(H)}{-\log\varepsilon}\Big\} = f_\delta(\mu)$$

for the two-stage procedures. This demonstrates that the bound of Theorem 3.3 is sharp. The procedures considered in Theorem 4.2 are not the only ones to attain this bound. Clearly, there are asymptotically optimal multi-stage procedures and for these the choice of how long to make the first stage is not so crucial. Continuation regions other than those of Schwarz are also possible. In fact 'asymptotic shape' is the relevant property for a sequence of regions. This is discussed, for example, in Schwarz (1962) and Berk (1978). Swanepoel and Geertsema (1976) use regions given by

$$g_a(t) = \sqrt{2(a+\log t)t} - \delta t, \quad t \leq t_a^o,$$

where t_a^o solves $2(a+\log t) = \delta^2 t$. These regions have the same asymptotic shape as the Schwarz regions and hence they can give asymptotically optimal procedures. A difficulty arises when Schwarz regions are used, namely the calculation of error probabilities, although asymptotic approximations (Woodroofe (1976)) are available. The Swanepoel and Geertsema regions have advantage in this respect as good bounds on the error probabilities are given by a result of Robbins (1970).

§ 5 FURTHER COMMENTS ON MULTI-STAGE PROCEDURES

In practical use of the procedures proposed in Section 2, the first question to arise is: how many stages? Naturally, two stage procedures are simplest, but the choice of first-stage sample size is then of considerable importance (see below). Multi-stage procedures are less arbitrary and more flexible--as the number of stages increases, they approximate the flexibility of the fully adaptive procedures of Section 6. Of course, the calculation of the optimal sampling proportions at the end of each stage may prove burdensome without a computer. In such cases a reasonable approach would seem to be to allocate in the proportions $\sqrt{k_N - 1} : 1 : \ldots : 1$ amongst the k_N remaining populations, with the currently 'best' population receiving the higher number of observations. This approximate mathod of allocation coincides with the optimal one in slippage configurations (see Section 6), and appears to perform well in other cases also.

For the simulations reported in Section 7, two two-stage procedures based on the Schwarz region were used--one using the 'optimal' allocation (denoted TWOP) and the other the $\sqrt{k_N - 1}$ allocation (TWOA) described above at the beginning of the second stage. In order to decide on the length n° of the first stage, a sensitivity study was conducted, and the results are presented in Table I. The first number in each cell is the mean total number (ASN) of observations on all $k = 10$ populations in both stages. The figure in parentheses is the standard error of these estimates. (For further details concerning the simulations and the parameter values used, see Section 7). The values for n° tested are fractions of 56, which

is one quarter of the length required for the fixed sample size procedure of Bechhofer (1954) to achieve a PCS of .90. The number 56 was chosen because savings of up to 75% in ASN are obtained by passing from fixed sample size to fully sequential procedures in hypothesis testing (Wald, 1974, §3.6, Woodroofe, 1976, §8). Table I indicates that in general, one quarter of the fixed sample size might be a good choice for first stage length. This is reasonable because an incorrect selection at the first stage increases the length of the experiment dramatically, and this is less likely to occur if the first stage is longer. Recall also that populations are being eliminated throughout the first stage (especially in the equally spaced configuration), so a long first stage need not be overly inefficient.

The question of first stage sample size clearly requires further study. One interesting possibility is to make the length of the first stage data dependent--stop sooner if there is a clear candidate for best sample mean. Since the sampling is VT until the end of the first stage, the Brownian approximation remains valid. Another interesting question is whether James-Stein-type shrinkage of the sample mean estimates at the end of the first stage will lead to appreciable improvements in ASN.

A problem seems to arise with the 'optimal allocation' of Section 4 when three or more opoulations are simultaneously "best". The current allocation scheme (in this situation unjustly) emphasises the population with the largest sample mean at the conclusion of the first stage. If this population were to be eliminated by one of the other 'equal best' populations, an unnecessarily long experiment would result.

It would be interesting to derive expressions for the asymptotic relative efficiency of the optimal procedures of Section 4 to VT (or

$\sqrt{k-1}$: 1 : ... : 1) sampling for various configuration of means. This could perhaps be done by extending the mean path approximation method of Section 6.

Table I

	Slippage configuration $\mu_{[1]} = \mu_{[9]} = \mu_{[10]} - 0.2$.		Equally spaced configuration $\mu_{[i+1]} - \mu_{[i]} = 0.2 \quad 1 \leq i \leq 9$.	
n°	TWOP	TWOA	TWOP	TWOA
56	1283* (24)	1193* (20)	384* (7)	378* (7)
42	1336 (53)	1171 (46)	381 (17)	375 (17)
28	1561 (78)	1246 (49)	383 (17)	372 (18)
14	1851 (97)	1195 (43)	458* (12)	420** (14)

Based on 100 replications, $k = 10$, $P^* = 0.9$, $\delta^* = 0.2$, $\sigma^2 = 1$ except where indicated

* Based on 500 replications.
**Based on 200 replications.

§ 6 HEURISTIC RESULTS FOR FULLY ADAPTIVE SEQUENTIAL PROCEDURES

General form of the procedures. As in Section 2, observations now labelled X_{ij} $(1 \leq i \leq k,\ j \geq 1)$ are available sequentially from each of $k(\geq 2)$ populations $\pi_1,\ldots,\pi_k$. The observations $\{X_{ij}\}$ are independently distributed normal random variables with unknown means μ_i and common known variance σ^2 (taken equal to 1).

If a total of N observations have been taken from all the populations, we shall say that we are at epoch N, and denote the number of observations taken so far on π_i by $m_i(N)$. Let

$$\overline{X}_i(N) = \frac{1}{m_i(N)} \sum_{j=1}^{m_i(N)} X_{ij},$$

and let $\overline{X}_{[1]} \leq \overline{X}_{[2]} \leq \cdots \leq \overline{X}_{[k]}$ be the ordered sample means.

Elimination rule. After epoch N, let $I_N = \{i: \pi_i \text{ not yet eliminated}\}$, and eliminate from all further consideration all populations π_i, $i \in I_N$ for which $\exists j \in I_N$ with

$$Z_{ji}(N) = t_{ji}(N)(\overline{X}_j(N) - \overline{X}_i(N) \geq g(t_{ji}(N)), \tag{6.1}$$

where

$$t_{ji}(N) = \frac{m_i(N)\ m_j(N)}{m_i(N)+m_j(N)}$$

and $g(t)$, $t \geq 0$ is a non-negative function depending on the PCS desired. The elimination leaves the set of populations $I_{N+1} \subseteq I_N$. The experiment terminates when a single population remains, and it is selected as best.

Sampling rule. Begin by taking one observation on each population. For epochs k+1, k+2, ... the sampling rule gives the next population to observe as a function of the observed values already taken. For a given function g we allow any sampling rule for which the procedure terminates in finite time with probability one.

Comparing these procedures with the multistage ones proposed in Section 2, we note firstly that here the sampling rule may be fully adaptive. The population to be sampled at epoch N need not be decided until after epoch N-1. The multistage procedures have the restriction that the sampling rule for an entire stage must be specified at the beginning of that stage.

Secondly, the statistics $Z_{ji}(N)$ are simpler than their multistage analogues. $Z_{ji}(N)$ is the statistic that would be used in the two population problem obtained by restricting attention to π_i and π_j. In the two population problem, $Z_{ji}(N)$ enjoys many appealing properties (Robbins and Siegmund (1974)). It is the invariant likelihood ratio test statistic; and after a reduction by invariance it is sufficient for $\delta_{ji} = \mu_j - \mu_i$ if the sampling is not adaptive. Even if the sampling is adaptive (and invariant under translations), its distribution coincides with that of a Brownian motion with drift δ_{ji}, observed at times $t_{ji}(N)$. Furthermore, it is a martingale with respect to the σ-fields generated by previous Z_{ji}'s. As mentioned in the introduction, these properties do not persist in the k-population case: in particular the embedding of $Z_{ji}(N)$ into Brownian motion is not always justifiable, and the martingale property fails in general (Appendices 1 and 2).

The situation is not hopeless, however. For translation invariant sampling rules there do exist martingales of a relatively simple form for k populations which appear to be the appropriate analogues of the two population case. These martingales (discussed in Appendix 5) are linear combinations of the pairwise statistics $\{Z_{ji}(N)\}_{i,j=1}^{k}$, and thus may be regarded as linear contrasts of the sample means. Although it is not clear how to exploit these martingales, they do suggest that invariant sampling rules in the k-population problem have some stability properties, and lend extra credibility to the mean path approximations which are the main tool of this section.

Some Notation. For convenience, we shall suppose that $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_{k-1} < \mu_k$, and write $\delta_{ji} = \mu_j - \mu_i$; and δ_i for $\delta_{k,i}$.

In doing the asymptotics it will be convenient to use families of regions with an "asymptotic shape". Following Berk (1978), the family will be described by functions $\{g_a(t)\}_{a \in \mathbb{R}}$ which are related to the elimination rule (6.1) as follows: π_j eliminates π_i at time $t = t_{ji}(N)$ if

$$Z_{ji}(N) \geq a\, g_a(t_{ji}(N)/a).$$

Let $R_a(t) = g_a(t)/t$. We assume also that as $a \to \infty$, the functions g_a (equivalently R_a) converge pointwise to limits $g(t)$ (resp. $R(t)$), which correspond to the "asymptotic shape".

Examples:

1) Paulson (1964). Now (6.1) has the form "reject π_i at t if

$$Z_{ji}(t) \geq a - \lambda t \quad \lambda \geq 0."$$

Thus $g(t) \equiv g_a(t) = 1 - \lambda t$ and $R(t) = 1/t - \lambda$.

2) Schwarz (1962). $g(t) \equiv g_a(t) = \sqrt{2t} - \delta t$.

3) Swanepoel and Geertsema (1976). $g_a(t) = \sqrt{2t(1 + \frac{\log t}{a})} - \delta t$. Thus $g(t) = \sqrt{2t} - \delta t$ so this family is asymptotically equivalent to that of Schwarz.

The mean path approximation involves replacing the statistic $Z_{ji}(N)$ by $\tilde{Z}_{ji}(N) = t_{ji}(N)\delta_{ji}$. The asymptotic validity (as $a \to \infty$) of the approximation is guaranteed by the strong law of large numbers as long as $m_i(N) \to \infty$, $i = 1,\ldots,k$.

For a given region $g(t) = a\, g_a(t/a)$, the exit time τ_{ji} of the mean path can be determined by solving the equation

$$\tau_{ji}\,\delta_{ji} = a\, g_a(\frac{\tau_{ji}}{a}),$$

which yields

$$\tau_{ji} = a\, R_a^{-1}(\delta_{ji}). \tag{6.2}$$

To simplify the analysis, we shall make the assumption that at all times (actually only at times of elimination of various populations)

$$m_k(N) \geq m_i(N), \quad 1 \leq i \leq k-1. \tag{6.3}$$

This is again reasonable asymptotically, as one would certainly wish to

use sampling rules which gave the most observations to the best population. In addition, one would expect asymptotically that all eliminations would be done by the best population (see the proof of Theorem 4.2 in this regard), and the theory of multiple comparisons with a control (Dunnett, Bechhofer and others) then suggests favouring the best population.

The assumption (6.3) implies in the mean path approximation that all eliminations are in fact performed by π_k. We shall therefore write m_i for the number of observations on π_i at the time of its elimination, and $m_{k,i}$ for the number of observations on π_k at that time.

Let m_k be the number of observations on π_k at the end of the experiment and note that $m_k = m_{k,k-1}$. We are interested in the variation with stopping region and sampling rule of the average total sample size ("ASN", for average sample number), which in the mean path approximation may be represented by

$$ASN = \sum_{i=1}^{k} m_i . \tag{6.4}$$

For a given region the τ_{ji} values are fixed by (6.2), so the m_i values are subject to the constraints (writing τ_i for τ_{ki})

$$\frac{1}{m_i} + \frac{1}{m_{k,i}} = \frac{1}{\tau_i} = \frac{1}{a\, R_a^{-1}(\delta_i)} . \tag{6.5}$$

In order to study, for a given stopping region, the effects on ASN of various sampling rules, it is useful to separate the two factors by means of the following combination of (6.4) and (6.5),

$$ASN = \sum_{i=1}^{k-1} (1 + \frac{m_i}{m_{k,i}}) \tau_i + (1 + \frac{m_k}{m_{k-1}})\tau_{k-1}. \qquad (6.6)$$

Here the ratios $m_i/m_{k,i}$ are specified by the sampling rule; whilst the τ_i depend only on the region and the configuration $\underset{\sim}{\mu}$ of population means.

In comparing the ASN performance of two families of regions $\{R_a\}$ and $\{R_a^1\}$, some method of matching their error probabilities must be employed. In the selection problem, the natural criterion is the probability of incorrect selection in the slippage configuration at the vertex of the preference zone $P_\delta = \{\mu: \mu_k - \mu_i \geq \delta \quad i = 1,\ldots,k-1\}$ (called the least favourable configuration in VT sampling). This will be denoted by $\varepsilon(a) = \varepsilon(a,R_a)$. Thus we shall use

$$\lim_{a\to\infty} \frac{ASN(\mu)}{-\log\varepsilon(a)}$$

in comparing two regions.

If the regions $\{R_a\}$ satisfy the conditions of Berk (1978) (essentially that R_a decrease from ∞ to 0 on $[0,\infty]$ and $R_a \downarrow R$), then it is natural to conjecture that Berk's result would extend to give

$$\frac{a}{-\log\varepsilon(a)} \sim \frac{1}{\kappa_R} \quad \text{as } a \to \infty, \qquad (6.7)$$

where $\kappa_R = \inf_{x>0} \frac{x}{2}(R(x) + \delta)^2$. This conjecture may be verified using methods similar to those of Theorem 4.1 if the Brownian approximation is assumed to hold for the statistics $Z_{ij}(N)$. Even in the absence of this approximation, (6.7) is still plausible, since it is an asymptotic relation, and one would expect the sampling rule to stabilise as information about the population means increases.

Combining (6.2) and (6.7) with (6.6) we then obtain

$$\frac{ASN}{-\log\varepsilon(a)} \sim \frac{1}{\kappa_R}\left[\sum_{i=1}^{k-1}\left(1+\frac{m_i}{m_{k,i}}\right)R^{-1}(\delta_i) + \left(1+\frac{m_k}{m_{k-1}}\right)R^{-1}(\delta_{k-1})\right] \qquad (6.8)$$

Proceeding analogously to Section 3, we may write down a lower bound for the ASN attainable with a given region by various sampling rules. Indeed,

$$ASN \geq \min\{\sum_1^k m_i : \frac{1}{m_i} + \frac{1}{m_{k,i}} \leq \frac{1}{\tau_i} \quad i=1,\ldots,k-1\}$$

$$= \min\{\sum_1^k m_i : \frac{1}{m_i} + \frac{1}{m_k} \leq \frac{1}{\tau_i} \quad i=1,\ldots,k-1\} = f(\underset{\sim}{\tau}), \qquad (6.9)$$

since $m_{k,i} \leq m_k$ for all i. It is easily checked that the minimum is uniquely attained for any given $\underset{\sim}{\tau}$, and that $f(\underset{\sim}{\tau})$ is increasing in each of its co-ordinates. Further, f is homogeneous of order 1: for any positive scalar c, $f(c\underset{\sim}{\tau}) = c\, f(\underset{\sim}{\tau})$. It is easy to compute values of f numerically.

A glance at (6.8) then yields for arbitrary regions $\{R_a\}$,

$$\lim_{a\to\infty} \frac{ASN}{-\log\varepsilon(a)} \geq f\left(\frac{R^{-1}(\delta_1)}{\kappa_R}, \ldots, \frac{R^{-1}(\delta_{k-1})}{\kappa_R}\right).$$

To minimise the right side, we need a region R which minimises

$$\frac{R^{-1}(\tilde{\delta})}{\kappa_R} = \frac{2R^{-1}(\hat{\delta})}{\inf_{x>0} x(R(x)+\delta)^2} \quad \text{for all } \tilde{\delta} > 0. \qquad (6.10)$$

For simplicity only, let us restrict to continuous R. Then (6.10) is clearly bounded below by $2/(\delta+\tilde{\delta})^2$. On the other hand, solving

$$x(R(x)+\delta)^2 \equiv c \quad \text{leads to} \quad R(x) = \sqrt{\frac{c}{x}} - \delta,$$

which is seen to attain the lower bound for all $\tilde{\delta}$. This is just the Schwarz region described above (with c=1) and in Section 4!

It is now evident how to compute the efficiency of a sampling rule given a particular region R and configuration of means $\underset{\sim}{\mu}$. Simply calculate the ratio of the ASN in (6.6) (with τ_i replaced by $R^{-1}(\delta_i)$) to $f(R^{-1}(\delta_1),\ldots,R^{-1}(\delta_{k-1}))$. Of course the relative efficiencies of two sampling rules may be easily compared using (6.6) alone.

Results for slippage configurations. Some more explicit results are available when the means are in a "slippage configuration": $\delta_1 = \ldots = \delta_{k-1} = \tilde{\delta}$. In this case all $m_{k,i} \equiv m_k$ and $m_i \equiv m$, say. Letting $\beta = m_k/m$, (6.8) becomes

$$\frac{ASN}{-\log\varepsilon(a)} = \frac{R^{-1}(\tilde{\delta})}{\kappa_R}\left[\kappa + \frac{\kappa-1}{\beta} + \beta\right] \tag{6.11}$$

which is minimised by taking $\beta = \sqrt{k-1}$. Hence $f(\tau\underset{\sim}{1}) = \tau(\sqrt{k-1}+1)^2$. From (6.11) we get

$$\frac{ASN}{-\log\varepsilon} \sim \frac{R^{-1}(\delta)}{\kappa_R}(\sqrt{k-1}+1)^2 \tag{6.12}$$

Hence in slippage configurations, the optimal sampling rule allocates

in the ratio $\sqrt{k-1}: 1 : \ldots : 1$. This is also the asymptotically optimal allocation rule in multiple comparisons of k-1 normal treatments with a control (Dunnett (1955), Bechhofer (1969)).

When using VT sampling, $\beta = 1$ in (6.11), so that the relative efficiency of VT to the optimal sampling rule is

$$e_k = \frac{ASN_{VT}}{ASN_{\sqrt{k-1}}} = \frac{2k}{(\sqrt{k-1}+1)^2} .$$

For $k = 2$, VT is the optimal rule; whilst $e_3 = 1.029$, $e_{10} = 1.25$ and $e_\infty = 2$. It is important to note that these efficiencies are independent of the stopping region.

It is also easy to compare stopping regions in slippage configurations. For example, from (6.8) (using P for Paulson, S for Schwarz),

$$\left(\frac{ASN}{-\log\varepsilon}\right)_P \Big/ \left(\frac{ASN}{-\log\varepsilon}\right)_S \sim \frac{R_P^{-1}(\tilde{\delta})}{K_P} \Big/ \frac{R_S^{-1}(\tilde{\delta})}{K_S}$$

$$= \frac{(\delta+\tilde{\delta})^2}{4(\delta-\lambda)(\tilde{\delta}+\lambda)} \geq 1 \qquad \forall \delta, \tilde{\delta}, \text{ and } 0 < \lambda < \delta,$$

which was to be expected in view of the optimality of Schwarz's region.

In summary, it can be seen from (6.11) that in slippage configurations the relative efficiency of two procedures is the product of two independent components - the relative efficiencies of stopping rules and of sampling rules. Thus

$$\left(\frac{ASN}{-\log\varepsilon}\right)_{P,\sqrt{k-1}} \Big/ \left(\frac{ASN}{-\log\varepsilon}\right)_{S,VT} = \frac{2k}{(\sqrt{k-1}+1)^2} \; \frac{(\delta+\tilde{\delta})^2}{4(\delta-\lambda)(\tilde{\delta}+\lambda)} > 1.$$

Arbitrary configurations of means. The sampling ratios which minimise $f(\underset{\sim}{\tau})$ naturally depend on the unknown vector $\underset{\sim}{\delta} = (\delta_1,\ldots,\delta_{k-1})$. An obvious sampling rule (whose asymptotic optimality is strongly indicated by the results of Section 4) is to calculate the optimal allocation in (6.9) using the current sample means after each observation and to allocate the next observation by picking a population at random according to the resulting probabilities. This rule would be complicated to use in practice, so we shall study instead a much simpler rule, motivated by the slippage configuration results above.

Definition of the "compensating $\sqrt{k_N-1}$ " sampling rule: $(C\sqrt{k_N-1})$

Take one observation from each population initially. Let I_N = {populations not eliminated after epoch N}, and let $k_N = |I_N|$. Let $\overline{X}_{i_{max}}(N) = \max_{i\in I_N} \overline{X}_i(N)$ and let i_N be the index of the population sampled at epoch N. Beginning with $i_N+1 \pmod k$, search for the next $j \in I_N \setminus \{i_{max}\}$ with $m_j(N) < m_{i_{max}}(N)/\sqrt{k_N-1}$. If such a j is found, take the next observation from π_j, otherwise sample from $\pi_{i_{max}}$.

In summary, the rule tries to ensure that the number of observations on the population currently regarded as best bears the ratio $\sqrt{k_N-1}$: 1 to the sample size on each of the other non-eliminated populations. We shall now show analytically using the mean path approximation that for any stopping region this rule dominates VT sampling. Simulation results are presented in Section 7.

In the mean path approximation, the $C\sqrt{k_N-1}$ rate satisfies assumption (6.3). If $\mu_i - \mu_{i-1}$ is small, it may happen, however, that there is insufficient time between the elimination of π_{i-1} and π_i for the sampling

rule to adjust the sampling ratio from $\sqrt{k-i+1} : 1$ to $\sqrt{k-i} : 1$. We shall therefore introduce constants $\alpha_i \geq 1$ (which will depend on the μ_i) defined by

$$\frac{m_{k,i}}{m_i} = \alpha_i\sqrt{k-i}. \tag{6.13}$$

Notice that

$$1 \leq \sqrt{k-i}\ \alpha_i \leq \sqrt{k-i+1}\ \alpha_{i-1} \tag{6.14}$$

From (6.6),

$$ASN_{C\sqrt{k_N-1}} = \sum_1^{k-1} (1 + \frac{1}{\alpha_i\sqrt{k-i}})\tau_i + (1 + \alpha_{k-1})\tau_{k-1},$$

while

$$ASN_{VT} = \sum_1^{k-1} 2\tau_i + 2\tau_{k-1}.$$

$$\text{Hence } \Delta = ASN_{VT} - ASN_{C\sqrt{k_N-1}} = \sum_1^{k-1} \tau_i(1 - \frac{1}{\alpha_i\sqrt{k-1}}) + \tau_{k-1}(1 - \alpha_{k-1}). \tag{6.15}$$

That $\Delta > 0$ is obvious if $\alpha_{k-1} = 1$. So suppose $\alpha_{k-1} > 1$ and let k_o be the smallest integer such that $\alpha_{k-k_o} = 1$. Then $m_{k,k-k_o} = \ldots = m_{k,k-1} = m_k$, and substituting (6.13) into (6.5) it follows that

$$\tau_i = \frac{\alpha_{k-1} + 1}{\alpha_i\sqrt{k-i} + 1}\ \tau_{k-1} \text{ for } i \geq k-k_o,$$

and hence that

$$\Delta > \tau_{k-1}[(\alpha_{k-1} + 1) \sum_{k-k_o}^{k-1} \frac{1}{\alpha_i\sqrt{k-i} + 1} (\frac{\alpha_i\sqrt{k-i} - 1}{\alpha_i\sqrt{k-i}}) + (1 - \alpha_{k-1})]$$

$$> \tau_{k-1}(\alpha_{k-1} - 1)[\sum_{k=k_o}^{k-1} \frac{1}{\alpha_i\sqrt{k-i}} - 1] > 0.$$

The second inequality follows from (6.14) and the third because $\alpha_{k-k_o} = 1$ and $k_o/\sqrt{k_o} > 1$.

In slippage configurations, the mean path approximation to $C\sqrt{k_N - 1}$ yields the $\sqrt{k-1}$ allocation discussed earlier, so all the optimality conclusions apply. Another configuration of means in which it is of interest to do more explicit comparisons of the $C\sqrt{k_N - 1}$ and VT rules is the "equally spaced means" (ES) case in which $\mu_i = \mu_{i-1} + \delta$, so that $\delta_i = (k-i)\delta$. To carry out the computations, we need to assume that the α_i's corresponding to the ES configuration equal one. A simple sufficient condition on a region R which ensures this is that $R^{-1}(j\delta)$ can be put in the form

$$R^{-1}(j\delta) = \frac{h(j)g(\delta)}{1 + \sqrt{j}}, \tag{6.16}$$

where g is arbitrary and h is decreasing in j. It is easily checked that Paulson's and Schwarz's regions satisfy this condition. When (6.16) obtains,

$$\frac{ASN}{\frac{C\sqrt{k_N - 1}}{a}} = \sum_{j=1}^{k-1} (1 + \frac{1}{\sqrt{j}}) R^{-1}(j\delta) + 2R^{-1}(\delta)$$

and

$$\frac{ASN}{a} VT = 2 \sum_{1}^{k-1} R^{-1}(j\delta) + 2R^{-1}(\delta).$$

Using Schwarz's region $R^{-1}(j\delta) = [\frac{2}{(j+1)\delta}]^2$, and for k = 10,

$$\frac{ASN}{\frac{C\sqrt{k_N - 1}}{ASN_{VT}}} \tilde{=} 1.09.$$

For Paulson's region $R^{-1}(j\delta) = \frac{1}{j\delta+\lambda}$, and for $\lambda = 0.25\delta$, k = 10,

$$\frac{ASN}{\frac{C\sqrt{k_N - 1}}{ASN_{VT}}} \tilde{=} 1.14.$$

Finally, we shall briefly discuss some of the similar results which hold when one considers the total number of observations on all populations except the one selected as best, called the inferior treatment number (ITN). Notice firstly that

$$m_i \geq \frac{m_i\, m_{k,i}}{m_i + m_{k,i}} = \tau_i$$

for any sampling rule, whilst for VT sampling, $\tau_i = \frac{m_i}{2}$. Hence the observation of Robbins and Siegmund (1974) extends to the k-population case: for any parameter configuration $ITN_{\text{any rule}} \geq \frac{ITN}{2}$ VT. Thus it is impossible to achieve a saving of greater than 50% over VT sampling.

By analogy with the ASN discussion,

$$\frac{ITN}{-\log\varepsilon(a)} = \frac{1}{-\log\varepsilon(a)} \sum_{1}^{k-1} (1 + \frac{m_i}{m_{k,i}})\tau_i \sim \frac{1}{\kappa_R} \sum_{1}^{k-1} (1 + \frac{m_i}{m_{k,i}})R^{-1}(\delta_i).$$

It follows as before that the Schwarz region is asymptotically optimal. Appropriate analogues of the results for slippage and ES configurations may easily be derived. Since the last term drops out of (6.15) it is trivial that $C\sqrt{k_N - 1}$ dominates VT; but it is likely that much better rules for purposes of minimising ITN could be developed. We have not pursued this topic because it has been argued that minimisation of ASN is a more relevant problem in contexts such as that of clinical trials (Byar et al. (1976)).

§ 7 SIMULATION RESULTS

A Monte Carlo study was performed in order to compare performances with regard to ASN, ITN and PCS of the procedures described in Sections 2 and 6. The values $P^* = 0.9$, $\delta^* = 0.2$, $\sigma^2 = 1$ and $k = 10$ were chosen along with two configurations of populations, namely the δ^*-slippage and equally spaced (ES) configurations described in Section 6. The results are displayed in Tables II, III. These values and configurations are chosen so that our results would be directly comparable with those contained in the tables in Chapter 18 of Bechhofer, Kiefer and Sobel (BKS) (1968) and in Turnbull, Kaspi and Smith (TKS) (1978). Nine combinations of stopping and elimination rules were considered, namely (A) FIXED sample size (Bechhofer (1954)); (B) the BKS likelihood based stopping rule with vector at a time (VT) sampling; (C) the BKS rule together with the adaptive RAND Q sampling rule described in TKS; (D),(E) Paulson's procedure with the Fabian (1974) modification with $\lambda/\delta^* = 0.25$ and the VT and $C\sqrt{k_n - 1}$ sampling rules (Section 6), respectively; (F),(G) the Schwarz stopping region together with the VT and $C\sqrt{k_N - 1}$ sampling rules; and (H),(I) Schwarz's region with two two-stage sampling rules. The first allocates in the ratio $\sqrt{k_I - 1} : 1 : \ldots : 1$ at the end of the first stage, with the most observations going to the population currently estimated as best, and k_I equal to the number of non-eliminated populations at the end of the first stage. The second allocates in the optimal ratios corresponding to the unique $\underline{d}_i$'s which achieve the infimum in expression (4.0). (Whilst the modification suggested in Remark 4.1 would be more correct in theory, in practice it produced large increases in the mean (ASN) and standard deviation of the total sample size over those presented in Tables II and III. The modification seems therefore to be

more sensitive to inaccuracies in estimates of the sample means at the end of the first stage.)

For procedures (H), (I), the length of the first stage was chosen to be 56--approximately one quarter of the number of observations required in the FIXED procedure. This point was discussed in greater detail in Section 5.

In all cases involving the use of Schwarz's region, the asymptotic error probability approximation given in Woodroofe (1976, Formulas (5.1) and (5.3) with r = 0) were used to determine the value of the parameter a (see Section 4) corresponding to $P^* = 0.9$ with k = 10 populations is a = 5.31.

Remark: Woodroofe's asymptotic approximation might be circumvented in the following ways: (1) calculate the error probabilities numerically either by finite difference methods or by simulation or (2) substitute the (asymptotically equivalent) Swanepoel-Geertsema (1976) region, for which good bounds are available (Robbins, 1970).

In the tables, the FIXED sample sizes of the nonsequential single stage procedure were taken from Table I of Bechhofer (1954). The results for procedures (B), (C), and (D) are taken from Tables I and III of TKS. The TKS RAND Q rule is included only in the slippage configuration case, and then purely for the sake of comparison. Although it is more efficient in slippage configurations than our rules based on pairwise comparisons, TKS found that it did not guarantee PCS $\geq$ .9 throughout the preference zone. Hence it cannot be regarded as a legitimate solution to the selection problem, and it is not presented in Table III.

The tables display $\{N_i;\ 1 \leq i \leq k\}$ where N_i is the mean number of observations taken from the population associated with $\mu_{[i]}$. Also displayed is the average total sample size ASN (N_{TOT} in the tables) and the observed proportion of correct selections. The second entry in any cell is the standard error of the estimated mean above it. The results are based on 500 replications of each procedure.

The following points are clear from the tables. The $C\sqrt{k_N - 1}$ rule achieves savings of 10-12% in ASN over the VT rule in all cases except the Schwarz region in the ES configuration, where they are comparable. The Schwarz region achieves savings between 12% and 25% over Paulson-Fabian in the ES case, whilst the two are roughly equivalent in the slippage configuration. Both two-stage procedures are dominated in both configurations by the non-adaptive VT Schwarz rule--the problem seems to be that N_{10} is inflated by the huge number of observations required to terminate the experiment in the cases when the wrong population is selected as best at the end of the first stage.

Table II

Simulation Results

$k = 10 \quad P^* = .9 \quad \delta^* = .2 \quad \sigma = 1$

δ^* - slippage configuration $\mu_{[1]} = \mu_{[9]} = \mu_{[10]} - \delta^*$

(First four columns are from Turnbull et al. (1978) Table II, remaining column entries are based on 500 replications.)

Stop. and Elim. rule / Sampling rule	FIXED	BKS VT	TKS	PF(.25) VT	PF(.25) $\sqrt{k_N - 1}$	Schwarz $\sqrt{k_N - 1}$	Schwarz VT	Schwarz 2-stage† $\sqrt{k_I - 1}$	Schwarz 2-stage† "opt"
N_{10}	223	145.3	179.9	175.9	200.6	207.4	193.7	247.3	335.5
					(4.1)	(3.9)	(3.8)	(5.8)	(10.9)
N_9	223	145.3	62.9	107.3	96.0	97.4	110.3	110.5	107.1
					(4.1)	(3.6)	(3.8)	(4.4)	(4.3)
N_8	223	145.3	59.8	112.4	95.3	96.2	105.7	100.6	109.7
					(4.0)	(3.8)	(3.7)	(4.0)	(5.3)
N_7	223	145.3	55.7	104.1	93.0	91.0	106.4	101.8	106.0
					(3.7)	(3.4)	(3.6)	(3.8)	(5.0)
N_6	223	145.3	61.5	107.3	95.0	94.9	106.0	109.0	96.0
					(4.0)	(3.4)	(3.8)	(4.3)	(3.7)
N_5	223	145.3	63.1	114.0	96.5	91.0	103.2	104.8	113.8)
					(4.0)	(3.5)	(3.7)	(4.0)	(6.1)
N_4	223	145.3	63.2	106.3	93.9	97.5	107.8	104.0	112.4
					(3.6)	(3.6)	(3.8)	(3.9)	(5.0)
N_3	223	145.3	59.7	110.4	94.8	90.7	103.3	98.5	100.3
					(3.9)	(3.4)	(3.5)	(3.7)	(4.5)
N_2	223	145.3	65.9	105.9	86.3	91.8	106.9	108.1	103.1
					(3.5)	(3.4)	(3.7)	(4.6)	(4.1)
N_1	223	145.3	60.1	111.0	95.3	93.7	105.6	108.5	99.3
					(4.1)	(3.5)	(3.6)	(4.0)	(3.8)
N_{TOT}	2230	1453.3	731.9	1154.7	1046.7	1051.7	1148.8	1193.3	1283.2
		(28.3)	(32.5)	(24.0)	(21.4)	(18.2)	(19.2)	(19.7)	(24.0)
Proportion of Correct Selections	-	0.911	0.910	0.925	0.956	0.912	0.918	0.922	0.918
		(.010)	(.020)	(.019)	(.013)	(.013)	(.012)	(.012)	(.012)

†First stage sample size is $56 \approx 223 \div 4$

Table III

Simulation Results

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

Equally spaced configuration $\mu_{[i+1]} - \mu_{[i]} = \delta^*$ $(1 \leq i \leq 9)$

(First three columns** are from Turnbull et al. (1978) Table III, remaining column entries are based on 500 replications.)

Stop. and Elim. Rule / Sampling Rule	FIXED	BKS VT	PF(.25) VT	PF(.25) $\sqrt{k_N - 1}$	Schwarz $\sqrt{k_N - 1}$	Schwarz VT	Schwarz 2-stage† $\sqrt{k_I - 1}$	Schwarz 2-stage† "opt"
N_{10}	223	64.8	121.6	122.9	126.0	118.1	126.3	130.0
				(2.6)	(3.3)	(3.2)	(3.4)	(3.6)
N_9	223	64.8	120.1	116.2	117.3	115.2	114.2	119.2
				(2.8)	(3.6)	(3.3)	(3.3)	(3.2)
N_8	223	64.8	65.6	56.0	45.9	50.3	50.4	49.0
				(1.0)	(1.4)	(1.5)	(1.4)	(1.4)
N_7	223	64.8	44.8	35.5	24.9	27.4	29.5	28.5
				(0.5)	(0.8)	(0.9)	(0.8)	(0.8)
N_6	223	64.8	34.9	26.2	15.3	17.7	18.2	18.0
				(0.3)	(0.4)	(0.5)	(0.5)	(0.5)
N_5	223	64.8	28.3	20.2	9.9	12.1	13.0	12.1
				(0.2)	(0.3)	(0.3)	(0.4)	(0.3)
N_4	223	64.8	23.1	16.9	7.4	8.3	9.7	9.7
				(0.2)	(0.2)	(0.3)	(0.3)	(0.3)
N_3	223	64.8	19.8	14.2	5.4	6.7	7.1	6.9
				(0.1)	(0.1)	(0.2)	(0.2)	(0.2)
N_2	223	64.8	17.6	12.3	4.4	5.4	5.4	5.8
				(0.1)	(0.1)	(0.1)	(0.1)	(0.2)
N_1	223	64.8	15.8	10.8	3.5	4.0	4.5	4.6
				(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
N_{TOT}	2230	648.2	491.5	431.4	360.2	365.3	378.3	383.7
		(27.5)	(9.3)	(6.2)	(7.7)	(7.5)	(7.5)	(7.5)
Proportion of Correct Selections	-	0.950	0.975	0.992	0.988	0.978	0.990	0.992
		(.015)	(.011)	(.004)	(.005)	(.007)	(.004)	(.004)

** TKS RAND Q is not relevant here as it only guarantees the PCS-requirement in slippage configurations--see Turnbull et al. (1978).

† First stage sample size was chosen to be $56 \approx 223 \div 4$

REFERENCES

Anderson, T. W. (1960). A modification of the sequential probability ratio test to reduce the sample size. Ann. Math. Statist. 31, 165-197.

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.

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

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

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

Byar, D. P., Simon, R. M., et al. (1976). Randomised clinical trials, perspectives on some recent ideas. New Engl. J. of Med. 295, July, 74-80.

Dudewicz, E. J., and Dalal, S. R. (1975). Allocation of observations in ranking and selection with uneuqal variances. Sankhya, Ser. B, 37, 28-78.

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

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

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

Mukhopadhyay, N. (1979). Some comments on two stage selection procedures. Commun. Statist.-Theor. Meth. A8(7), 671-683.

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

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.

Rinott, Y. (1978). On two-stage selection procedures and related probability inequalities. Commun. Statist.-Theor. Meth., A7(8), 799-811.

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

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

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

Stein, C. (1945). A two sample test for a linear hypothesis whose power is independent of the variance. Ann. Math. Statist. 16, 243-58.

Swanepoel, J. W. H., and Geertsema, J. C. (1976). Sequential procedures with elimination for selecting the best of k normal populations. S. Afr. Statist. J. 10, 9-36.

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

Wald, A. (1947). Sequential Analysis. Dover Reprint (1973). New York.

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

Appendix 1. INVALIDITY OF BROWNIAN APPROXIMATION TO THE "NATURAL STATISTIC"

We begin with the two population case, so as to exhibit more clearly where the extension to three populations breaks down.

Let observations on the populations π_x and π_y be denoted $x_1, x_2, \ldots$; $y_1, y_2, \ldots$.

Suppose that x_1 and y_1 have been observed and that we wish to determine from which population to take the third observation. Introduce indicator variables

$$\delta_x(\delta_y) = \begin{cases} 1 & \text{if next observation is on } \pi_x \ (\pi_y) \\ 0 & \text{otherwise.} \end{cases}$$

Thus $\delta_y = 1 - \delta_x$. The natural statistic, after taking the third observation, $z_{xy}(3)$, may be written as follows

$$z_{xy}(3) = \frac{(1+\delta_x)(1+\delta_y)}{1+\delta_x+1+\delta_y}\left[\frac{x_1+\delta_x x_2}{1+\delta_x} - \frac{y_1+\delta_y y_2}{1+\delta_y}\right]$$

$$= \frac{1}{2}(x_1-y_1) - \frac{\delta_x-\delta_y}{6}(x_1+y_1) + \frac{1}{3}\{\delta_x x_2 - \delta_y y_2\}. \qquad (A.1.1)$$

Thus the increment

$$\Delta_{xy}(2) = z_{xy}(3) - z_{xy}(2) = \frac{\delta_y-\delta_x}{6}(x_1+y_1) + \frac{1}{3}\{\delta_x x_2 - \delta_y y_2\}. \qquad (A.1.2)$$

In order to embed the $z_{xy}(N)$ process in Brownian motion, the crucial property required is that the increments Δ_{xy} be normally distributed with appropriate means and variances.

In the two population case, invariance considerations imply that one should use sampling rules which depend only on the differences $\{x_i-x_1, y_j-x_1\}$ (c.f. Robbins-Siegmund (1974)). In our case, this forces $\delta_x \in F(x_1-y_1)$. Consider the distribution of $\Delta_{xy}(2)|\delta_x,\delta_y$. All terms x_1+y_1, x_2 and y_2 are uncorrelated with x_1-y_1 and hence independent of it. Hence the conditional distribution of the increment is normal with parameters

$$N\left(\frac{\delta_y-\delta_x}{6}(\mu_x+\mu_y) + \frac{1}{3}(\delta_x\mu_x-\delta_y\mu_y), \frac{(\delta_y-\delta_x)^2}{18} + \frac{1}{9}(\delta_x^2+\delta_y^2)\right),$$

that is,

$$N\left(\frac{1}{6}(\mu_x-\mu_y), \frac{1}{6}\right).$$

Thus the increment corresponds to the increment $W(3/2) - W(1)$ of a B.M. with drift $\mu_x-\mu_y$.

In the three population case, let observations on the third population π_z be denoted $z_1,z_2,\ldots$. Suppose that x_1, y_1 and z_1 have been observed and introduce indicators δ_x, δ_y and δ_z as above to describe which population receives the next observation. Notice now that $\delta_x + \delta_y = 1 - \delta_z$ may be either 0 or 1. The natural statistic $z_{xy}(3)$ is now

$$z_{xy}(3) = \frac{1}{2}(x_1-y_1) - \frac{\delta_x-\delta_y}{2(2+\delta_x+\delta_y)}(x_1+y_1) + \frac{1}{2+\delta_x+\delta_y}\{\delta_x x_2-\delta_y y_2\},$$

which is in fact equivalent to the form (A.1.1).

Again, invariance considerations require that sampling rules depend only on pairwise differences, so that $\{\delta_x,\delta_y,\delta_z\} \in F(x_1-y_1,x_1-z_1)$. However, it is no longer true that the distribution of the increment $\Delta_{xy}(2)|\delta_x,\delta_y$ is normal. Indeed it is clear that x_1+y_1 is correlated with x_1-z_1. Suppose we used a rule which set $\delta_x=1$ if $(x_1-z_1) + (y_1-z_1) > 0$ and $\delta_z = 1$ otherwise. The correlation between $x_1+y_1-2z_1$ and x_1+y_1 is $1/\sqrt{2}$, so in this case $L(\Delta_{xy}|\delta_x,\delta_y)$ is certainly not normal, and no Brownian approximation can apply. For similar reasons it follows also that $z_{xy}(1),z_{xy}(2),\ldots$ need not be a martingale.

Appendix 2. A SECOND EXAMPLE CONCERNING THE BROWNIAN APPROXIMATION

This example further illustrates how we can underestimate error probabilities if we use the "natural" statistic discussed in sections 6 and 7. At the moment the conclusion is not complete; a more exact statement should be possible if we use Woodroofe's (1976) methods.

We consider two-stage procedures in the class C of section 2. The procedures use Schwarz's region and are indexed by a

$$g_a(t) = \begin{cases} \sqrt{2at} - \delta t & 0 \leq t \leq \dfrac{2a}{\delta^2} \\ 0 & t > \dfrac{2a}{\delta^2} \end{cases} .$$

We concentrate our attention on two populations, π_x and π_y, say. Suppose π_y is the best population so that exiting C_a through the upper boundary constitutes an error. During the first stage sampling is vector at a time on uneliminated populations. The length of stage one is such that if π_x and π_y are uneliminated they will both have $2\phi(a)$ observations at the end of it. In the second stage there are two possible values for the sampling ratio

$$R_{xy} = \text{Number of observations on } \pi_x : \text{Number of observations on } \pi_y .$$

At the end of stage one we set $R_{xy} = \rho$ or $\frac{1}{\rho}$, $\rho > 1$, and we sample in this ratio until one of π_x and π_y is eliminated. Define

$u_1,\ldots,u_{2\phi}$: observations on π_x in stage I

$v_1,\ldots,v_{2\phi}$: observations on π_y in stage I.

If we choose $R_{xy} = \rho$, then at a typical point in stage II we have

$X_1,\ldots,X_{\rho n}$: observations on π_x in stage II

$Y_1,\ldots,Y_n$: observations on π_y in stage II.

Let $\bar{u}_{2\phi} = \frac{1}{2\phi}\sum_{i=1}^{2\phi} u_i$, etc. The so-called natural statistic is

$$Z(t) = \frac{(2\phi+\rho n)(2\phi+n)}{4\phi+(\rho+1)n}\left\{\frac{2\phi\bar{u}_{2\phi}+\rho n\bar{X}_{\rho n}}{2\phi+\rho n} - \frac{2\phi\bar{v}_{2\phi}+n\bar{Y}_n}{2\phi+n}\right\}$$

where

$$t = \frac{(2\phi+\rho n)(2\phi+n)}{4\phi+(\rho+1)n}\,.$$

Consider a procedure for which $R_{xy} = \rho$ always. Then the process $Z(t)$ corresponds to a deterministic sampling rule and can therefore be embedded in a Brownian motion. Also, $P\{\pi_x \text{ eliminates } \pi_y\} \leq \varepsilon(a)$. The same holds if $R_{xy} = \frac{1}{\rho}$ always. In a sense the choice of sampling ratio allows us to choose one of two Brownian motions. This is seen by re-expressing $Z(t)$. For the general procedure suppose we choose $R_{xy} = \rho$, then

$$Z(t) = -\frac{\frac{1}{2}(\rho-1)n.2\phi}{4\phi+(\rho+1)n}\{(\bar{u}_{2\phi}-\mu_x) + (\bar{v}_{2\phi}-\mu_y)\}$$

$$+ \frac{(2\phi+\rho n)(2\phi+n)}{4\phi+(\rho+1)n}\left\{\frac{\rho n(\bar{X}_{\rho n}-\mu_x)}{2\phi+\rho n} - \frac{n(\bar{Y}_n-\mu_y)}{2\phi+n}\right\} +$$

$$+ \frac{(2\phi + \frac{1}{2}(\rho+1)n)2\phi}{4\phi+(\rho+1)n} \{(\bar{u}_{2\phi}-\mu_x) - (\bar{v}_{2\phi}-\mu_y)\}$$

$$+ \frac{(2\phi+\rho n)(2\phi+n)}{4\phi+(\rho+1)n} (\mu_x-\mu_y)$$

$$= T_1 + T_2 + T_3 + T_4, \text{ say.}$$

We can write

$$T_1(t) = \frac{-\frac{1}{2}(\rho-1)n}{4\phi+(\rho+1)n} \{W^u(2\phi) + W^v(2\phi)\},$$

$$T_2(t) = t\{\frac{W^x(\rho n)}{2\phi+\rho n} - \frac{W^y(n)}{2\phi+n}\},$$

$$T_3(t) = \frac{(2\phi + \frac{1}{2}(\rho+1)n)}{4\phi+(\rho+1)n} \{W^u(2\phi) - W^v(2\phi)\},$$

$$T_4(t) = t(\mu_x-\mu_y). \qquad \text{(A.2.1)}$$

where W^u, W^v, W^x, and W^y are independent, zero drift Brownian motions. The choice of R_{xy} depends only on W^u and W^v. By comparing the corresponding formulae when $R_{xy} = \frac{1}{\rho}$ we see that the joint distributions of $T_2(t)$ and $T_3(t)$ are the same in both cases. $T_1(t)$, which is determined at the start of stage II, is independent of $T_2(t)$ and $T_3(t)$, and for a given value of $W^u(2\phi) + W^v(2\phi)$ the choice of R_{xy} is equivalent to a choice of sign for $T_1(t)$. For $R_{xy} = \rho$, write

$$Z(t) = T_1(t) + (T_2(t) + T_3(t)) + T_4(t),$$

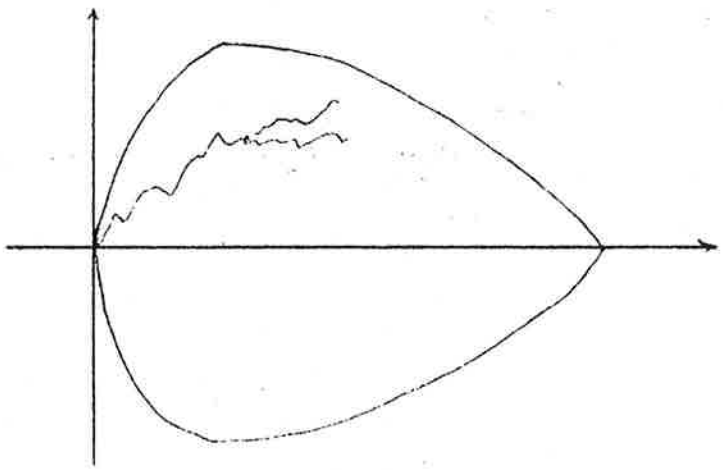
where $T_1(t) \sim N(0, \phi[\frac{\rho-1}{\rho+1+4\phi/n}]^2)$

$\sim N(0, \phi\beta^2(n))$, say, and

$$T_2(t) + T_3(t) \sim N(0, t-\phi\beta^2(n)). \qquad (A.2.2)$$

To generate Z(t) with the correct distribution we can proceed as follows. Observe a Brownian motion with drift $\mu_x-\mu_y$ at appropriate times in $[0,\phi]$. Take a value for $W^u(2\phi) + W^v(2\phi)$ from a normal distribution $N(0,4\phi)$, independent of the Brownian motion. Choose $R_{xy} = \rho$ or $\frac{1}{\rho}$, with regard to present knowledge. From two independent Brownian motions, W^x and W^y, construct values of Z(t) in stage II using $Z(t) = T_1 + T_2 + T_3 + T_4$ for $R_{xy} = \rho$ or the corresponding formula if $R_{xy} = \frac{1}{\rho}$.

The above formulation yields two results concerning the error probability $P\{\pi_x$ eliminates $\pi_y\}$.



If we regard the choice of R_{xy} as a choice between two possible paths, each of which is a Brownian motion with the correct drift, an

error can only occur if one of the paths exits C_a upwards so

$$P\{\pi_x \text{ eliminates } \pi_y\} \leq 2\varepsilon(a).$$

The method of generating Z(t) just described allows a correspondence between pairs of paths. The two paths for $R_{xy} = \rho, \frac{1}{\rho}$ coincide during stage I and differ by $2T_1(t)$ during stage II. If we can show that P{both paths give errors} << $\varepsilon(a)$, then by choosing R_{xy} appropriately we can have

$$P\{\pi_x \text{ eliminates } \pi_y\} \simeq 2\varepsilon(a).$$

Consider $P\{\pi_x \text{ eliminates } \pi_y \mid W^u(2\phi) + W^v(2\phi) = 0\}$. If this is of a smaller order of magnitude than $\varepsilon(a)$ then clearly P{both paths give errors} << $\varepsilon(a)$. By Berk's method we can check this criterion. For $\phi(a) = O(a^\nu)$, $\nu > 0$ we get

$$P\{\pi_x \text{ eliminates } \pi_y\} \simeq 2\varepsilon(a)$$

but for $\phi(a) \sim \log a$ this method is inconclusive. Maybe we can get a result using Woodroofe's techniques.

Appendix 3. DISCRETE AND CONTINUOUS ERRORS

We have a continuation region C of the usual form, given by the non-negative function $g(t)$, $t \geq 0$. The symmetry of C is essential. $B_\Delta(t)$ denotes a Brownian motion with drift Δ. We consider two processes. The continuous process is a Brownian motion $B_{-\delta}(t)$, $t \geq 0$ with $\delta > 0$. Let $\tau_C = \inf\{t > 0: B_{-\delta}(t) \notin C\}$. The discrete process is obtained by observing $B_{-\delta}(t)$ at a random, increasing sequence of times $\{t_i; i=1,2,\ldots\}$ taking values in a given countable set. The value of t_i depends on $B_{-\delta}(t)$ only through its values in the period $[0,t_{i-1}]$. We define $\tau_D = \inf\{t_i: B_{-\delta}(t_i) \notin C\}$ and assume that $\tau_D < \infty$ a.s. Note that $\tau_D \geq \tau_C$. The error probabilities are

$$\varepsilon_C = P\{B_{-\delta}(\tau_C) \geq g(\tau_C)\} = P\{B_{-\delta}(\tau_C) > 0\},$$

$$\varepsilon_D = P\{B_{-\delta}(\tau_D) \geq g(\tau_D)\} = P\{B_{-\delta}(\tau_D) > 0\}.$$

We wish to show $\varepsilon_D \leq \varepsilon_C$. Suppose the sequence $\{t_i\}$ depends on $B_{-\delta}(t)$ in the following symmetric way. Consider an outcome $\{(b(t); t \geq 0), \{t_i\}\}$ where $b(t)$ is the path of a Brownian motion. The conditional distribution of $\{t_i\}$ given $B_{-\delta}(t) = b(t)$, $t \geq 0$ is the same as that given $B_{-\delta}(t) = -b(t)$, $t \geq 0$. Under these conditions $\varepsilon_D \leq \varepsilon_C$.

Proof. Let $P_\Delta(\omega)$ denote the probability distribution on the space with elements $\omega = \{(b(t); 0 \leq t \leq \tau_D), \{t_i; t_i \leq \tau_D\}\}$, where $b(t)$ is the realisation of a Brownian motion $B_\Delta(t)$. Then

$$\frac{dP_\delta}{dP_{-\delta}}(\omega) = \exp\{2\delta b(\tau_D)\}.$$

(The derivative is taken relative to the σ-field $\sigma\{b(t);\ 0 \leq t \leq \tau_D\}$).

Let $\Omega_1 = \{\omega: b(\tau_C) > 0,\ b(\tau_D) < 0\}$ and $\Omega_2 = \{\omega: b(\tau_C) < 0,\ b(\tau_D) > 0\}$.

Then

$$\varepsilon_C - \varepsilon_D = \int_{\Omega_1} dP_{-\delta}(\omega) - \int_{\Omega_2} dP_{-\delta}(\omega).$$

We note that Ω_2 can be obtained from Ω_1 by replacing $b(t)$, $t \geq 0$ by $-b(t)$, $t \geq 0$. By the symmetry property of the t_i's

$$\varepsilon_C - \varepsilon_D = \int_{\Omega_1} dP_{-\delta}(\omega) - \int_{\Omega_1} dP_{\delta}(\omega).$$

$$= \int_{\Omega_1} \left(1 - \frac{dP_\delta}{dP_{-\delta}}(\omega)\right) dP_{-\delta}(\omega).$$

On Ω_1, $b(\tau_D) < 0$ so

$$\frac{dP_\delta}{dP_{-\delta}}(\omega) = \exp\{2\delta b(\tau_D)\} < 1,$$

and hence $\varepsilon_C - \varepsilon_D \geq 0$, as required.

Appendix 4. A COMBINATORIAL RESULT

We use the following combinatorial result in proving Theorem 3.3. Suppose we are given pairs $\{(\mu_i,h_i)\colon h_i > 0,\ i = 1,\ldots,k\}$ and $\mu_k \geq \mu_i$ for $i = 1,\ldots,k-1$. There is a function $\theta\colon \{1,2,\ldots,k-1\} \to \{1,2,\ldots,k\}$ with the property

(1) $\theta(i) \neq i$; $\mu_{\theta(i)} \geq \mu_i$ and for each $i \neq k$ there exists n such that $\theta_{(i)}^n = k$.

Let $G(x)$, $x \geq 0$ be an increasing function of x such that

$$(2) \qquad \frac{1}{h_{\theta(i)}} + \frac{1}{h_i} \quad G(\mu_{\theta(i)} - \mu_i).$$

Then

$$(3) \qquad \sum_{i=1}^{k} h_i \geq \inf_{d_i>0} \{\sum_{1}^{k} d_i\colon \frac{1}{d_i} + \frac{1}{d_k} \leq G(\mu_k - \mu_i)\}.$$

Remarks. In the context of Theorem 3.3, $\theta(i)$ is the index of the population that eliminates π_i, and G is given by

$$G(x) = (-\log \varepsilon)^{-1} \{\frac{(x+\delta)^2}{2} + \eta\}$$

and we note that for these purposes it is sufficient to take $[k] = k$.

Proof. The symmetric function $\sum_{1}^{k} h_i$ is unchanged by permutations $\{\bar{h}_i\}$ of the numbers $\{h_i\}$. Hence conclusion (3) will follow if we can find a

permutation which satisfies

$$(4) \qquad \frac{1}{h_i} + \frac{1}{h_k} \leq G(\mu_k - \mu_i).$$

This in turn becomes obvious if we can find a permutation $\{h_i\}$ and a modification $\tilde{\theta}$ of θ such that properties (1) and (2) are preserved and in addition

$$(5) \qquad h_{\tilde{\theta}(i)} \geq h_i.$$

Indeed, suppose that (1), (2), and (5) obtain, and that $k = \theta^n(i)$. It follows that $h_k \geq h_{\theta(i)}$ and $\mu_k \geq \mu_{\theta(i)}$; and we recover (4) from (2) by recalling that G is increasing.

To obtain (5) we argue as follows. For any set of $\theta(i)$'s we can construct a tree with nodes labeled $i = 1,2,\ldots,k$. The parent of node i is $\theta(i)$ and the root is node k. The tree satisfies the following (see Figure A.4.1):

(i) The root is node k.

(ii) The parent of node i is $\theta(i)$, $i = 1,2,\ldots,k-1$.

(iii) Associated with node i is the pair (μ_i, h_i).

(iv) $\mu_{\theta(i)} \geq \mu_i$.

(v) $\frac{1}{h_{\theta(i)}} + \frac{1}{h_i} \leq G(\mu_{\theta(i)} - \mu_i)$.

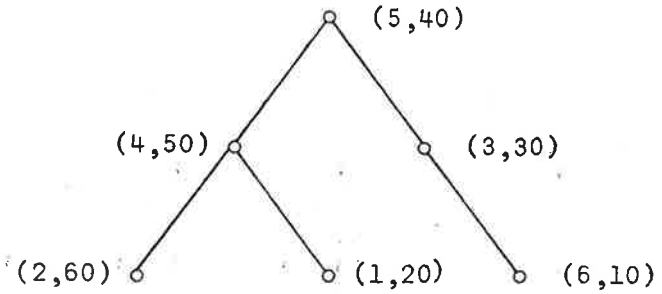
Consider the following transformation of the tree. For an i such that $h_{\theta(i)} < h_i$,

(a) Interchange the values $h_{\theta(i)}$ and h_i, so now $h_{\theta(i)} > h_i$,

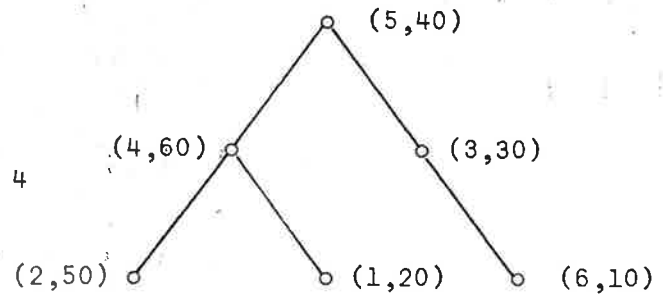
(b) If $\theta(j) = i$ for any $j \in \{1,2,\ldots,k-1\}$, set $\theta(j) = \theta(i)$.

This transformation preserves properties (i) to (v). By repeated applications we arrive at a tree with the additional property (5), as desired.

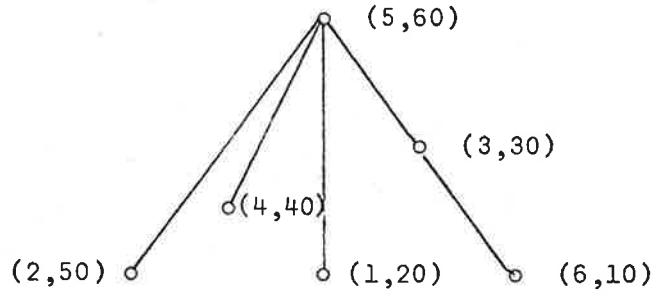
Figure A.4.1. Examples of trees.



swap observations
on populations 2 & 4



swap observations
on π_4 and π_5



Appendix 5. EXTENSION OF THE MARTINGALE PROPERTY TO $k \geq 3$ POPULATIONS

Consider first the two population case. Let the observations on π_x and π_y be denoted $x_1, x_2, \ldots, x_m, \ldots$ and $y_1, y_2, \ldots$ respectively. Write $\bar{x}_m = \frac{1}{m} \sum_{i=1}^{m} x_i$, $\bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_j$, $\delta = \mu_x - \mu_y$ and define

$$Z_{m,n} = \frac{mn}{m+n} (\bar{y}_n - \bar{x}_m).$$

To see that for translation invariant sampling rules, $Z_{m,n} - \frac{mn}{m+n}\delta$ is a martingale with respect to σ-fields $\mathcal{F}_{m,n} = \sigma\{Z_{ij},\ 1 \leq i \leq m,\ 1 \leq j \leq n\}$, consider the decomposition (Robbins and Siegmund (1974)):

$$Z_{m+j,n+k} = Z_{m,n} + \frac{(m+j)\sum_{n+1}^{n+k} y_i - (n+k)\sum_{m+1}^{m+j} x_i}{m+n+j+k}$$

$$+ \frac{(nj-mk)(\sum_{1}^{n} y_i + \sum_{1}^{m} x_i)}{(m+n)(m+n+j+k)}.$$

It is then apparent that the second term on the right side is independent of $\mathcal{F}_{m,n}$ since it consists of "future observations" and the third term is independent because it is orthogonal to the differences $\{x_i - x_1, y_j - x_1,\ 1 \leq i \leq m,\ 1 \leq j \leq n\}$ which generate $\mathcal{F}_{m,n}$.

A similar argument is possible to exhibit martingales in the k population case. This will be done here for $k = 3$, but the argument is completely general. Denote observations on π_z by $z_1, z_2, \ldots, z_\rho, \ldots$ and let the mean of π_z be μ_z and $\bar{z}_p = \frac{1}{p} \sum_{k=1}^{p} z_k$. Define

$$Z_{m,n,p} = \frac{1}{m+n+p}\,[(n+p)m\bar{x}_m - mn\bar{y}_n - mp\bar{z}_p]$$

$$= \frac{1}{m+n+p}\,[mn(\bar{x}_m-\bar{y}_n) + mp(\bar{x}_m-\bar{z}_p)]\,.$$

The following decomposition shows (by an argument analogous to the one above) that $Z_{m,n,p} - EZ_{m,n,p}$ is a (one-dimensional) martingale with respect to the σ-fields generated by the differences $\{x_i-x_1, y_j-x_1, z_k-x_1, 1 \le i \le m, 1 \le j \le n, 1 \le k \le p\}$, for any sampling rule which depends only on these differences (i.e. is translation invariant):

$$Z_{m+j,n+k,p+\ell} = Z_{m,n,p} + \frac{(n+p+k+\ell)\sum_{m+1}^{m+j} x_i - (m+j)(\sum_{n+1}^{n+k} y_i + \sum_{p+1}^{p+k} z_i)}{m+n+p+j+k+\ell}$$

$$+ \frac{[m(k+\ell) - j(n+p)](\sum_{1}^{m} x + \sum_{1}^{n} y + \sum_{1}^{p} z)}{(m+n+p)(m+n+p+j+k+\ell)}\,.$$

One consequence of the martingale property may be obtained from the optional sampling theorem. In the two population case, if M and N are the values of m and n on termination of the experiment, and $E\,\frac{MN}{M+N} < \infty$ then

$$E\,Z_{M,N} = \delta E(\frac{MN}{M+N}).$$

This result provides supporting evidence for the validity of the mean path approximation, in which one approximates the random path $(\frac{mn}{m+n}\,,\, Z_{m,n})$ by $(\frac{mn}{m+n}\,,\, \frac{mn}{m+n}\,\delta)$. Similar results hold for general k, and in particular for $Z_{m,n,p}$.

Two further martingales ($W_{m,n,p}$ and $V_{m,n,p}$, say) may be obtained by permuting the labels m,n,p and populations X, Y and Z, and one finds that $Z + W + V \equiv 0$. The "natural statistics" $\frac{mn}{m+n}(\bar{x}_m - \bar{y}_n)$, $\frac{np}{n+p}(\bar{y}_n - \bar{z}_p)$ and $\frac{mp}{m+p}(\bar{x}_m - \bar{z}_p)$ used in sections 6 and 7 are linear combinations of Z, W and V in which the coefficients depend on m, n and p. Thus the martingale property is not preserved, but it is still possible to feel that the mean path approximation is reasonable, and so this approximation forms the basis of the heuristic analysis of section 6.