SEQUENTIAL METHODS IN STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 16
PWN-POLISH SCIENTIFIC PUBLISHERS
WARSAW 1985

SOME PROBLEMS IN SEQUENTIAL ANALYSIS

JOHN A. BATHER

*Department of Mathematics, University of Sussex,
Brighton, United Kingdom*

First of all, I thank you for the honour of inviting me to the International Centre. My lectures will be based mainly on a paper presented to the Royal Statistical Society in London, earlier this year, entitled: *Randomised allocation of treatments in sequential experiments.* However, since I will be giving five lectures, this gives me the opportunity to discuss related work more fully and to extend the paper in several directions.

1. Introduction

1.1. The multi-armed bandit problem. The multi-armed bandit problem derives its name from a gambling situation in which there are $k \geqslant 2$ different machines or "bandits", each capable of producing a reward 1 or 0 in a single trial. The probabilities p_i, $1-p_i$, $i = 1, 2, \ldots, k$, of these events are unknown and the gambler seeks to maximise his total reward over a sequence of T trials, using any combination of the machines. More formally, let $p_1, p_2, \ldots, p_k$ be the unknown probabilities of success in k different sequences of independent Bernoulli trials. It is required to maximise the expected number of successes in T trials.

The association with gambling is perhaps unfortunate, since the problem is equivalent to one concerned with clinical trials and it is worth noting that "bandits" did not appear in the original formulation by Robbins (1952). We can think of a single sequence of patients and k possible treatments for any one of them. The aim is to specify a rule for allocating a treatment to each patient depending, at each stage, only on previously observed successes and failures. The only difference in this alternative view of the problem is that it seems unrealistic to assume that T is known in advance, when it is interpreted as the total number of patients, both inside and outside the period of the experiment. For the moment, let us assume that T is given, but we must bear in mind the question of sensitivity to the choice of T in any comparisons between allocation rules.

A recent paper by Gittins (1979) discussed a wide range of allocation problems, including multi-armed bandits, using an optimality criterion based on a discount factor a, $0 < a < 1$, and an infinite sequence of decisions. There is a close, but rather deceptive relationship between the present formulation ($a = 1$) when T is large and the discounted version of the problem ($T = \infty$) as $a \uparrow 1$. The main advantage of using a discount factor is that, for any choice of independent prior distributions on the parameters $p_1, p_2, \ldots, p_k$, the optimal policy can be expressed in terms of dynamic allocation indices. In effect, the Gittins index can be constructed separately for each of the unknown probabilities p_i, by computing the solution of a one-armed bandit problem. General results of this type have given us a more penetrating view through the complexities of sequential analysis and a powerful stimulus to further research: see the references in Gittins (1979) and also the more recent paper by Whittle (1980). Strictly speaking, the simplification of allocation rules by using a separate index for each unknown probability is not justified by exact optimality when T is finite, but as we shall see the principle can still be effective.

Bayesian models are attractive from the mathematical point of view because they provide a clear specification of the optimisation problem and there has been a great deal of work on the form of optimal policies under various special conditions: see, for example, Berry and Fristedt (1979).

On the other hand, such policies are often complicated to construct and apply. The Bayesian approach also leads to dependence on prior distributions and the effect of parameters like the discount factor may be even more important. In particular, it can be shown that the limiting form of the dynamic allocation index as $a \uparrow 1$ does not determine a good policy. The reasons for this will be explained in Section 2, in the case of the one-armed bandit problem: $k = 2$, but p_2 is given, and the limiting form of the general policy has been established recently by Kelly (1980). Roughly speaking, it will be argued that the best is the enemy of the good and the main purpose in this paper will be to suggest policies that perform reasonably well for any values of $p_1, p_2, \ldots, p_k$ and all except small values of T.

Let us introduce some notation and a criterion for the comparison of decision rules. Suppose that, after a total of t trials, using a sequential rule Δ to allocate the k treatments, we have observed $r_i = r_i(t)$ successes in $n_i = n_i(t)$ trials with treatment i. The proportion of successes achieved so far is r/t, where $r = \sum r_i$ and $t = \sum n_i$. At the end of a sequence of T trials, we are mainly interested in the total number of successes, which will be denoted by $R = r(T)$. Similarly, let us distinguish the final values $R_i = r_i(T)$ and $N_i = n_i(T)$, $i = 1, 2, \ldots, k$. We seek to maximise the expected number of successes or, equivalently, to minimise the *expected successes lost* (e.s.l.):

$$L_A(p_1, p_2, \ldots, p_k, T) = T\max(p_1, p_2, \ldots, p_k) - E_A(R).$$

This is a convenient form for the loss function, since it represents the effective cost of ignorance about which treatment is best. For any given integer $T > 0$, a useful measure of performance is

$$M_A(T) = \sup L_A(p_1, p_2, \ldots, p_k, T),$$

where the supremum is taken over all possible values of the unknown probabilities: $0 \leqslant p_i \leqslant 1$, $i = 1, 2, \ldots, k$.

On the other hand, if T is unknown and thought to be large, it is reasonable to consider the long-term behaviour of decision rules. In this case, an obvious requirement is that the proportion of successes R/T should converge to $\max(p_1, p_2, \ldots, p_k)$ as $T \to \infty$. However, most of the policies investigated previously do not satisfy this condition. The difficulty can be illustrated by considering a naive "play the favourite" rule. Suppose we start the procedure by using each treatment once and then, for all $t \geqslant k$, allocate treatment i in the next trial if and only if i is the smallest integer such that

$$\frac{r_i}{n_i} = \max_{1 \leqslant j \leqslant k} \frac{r_j}{n_j}. \tag{1.1}$$

This is clearly a bad policy, because it might reject the best treatment permanently after its first trial. Even if we start with $m > 1$ applications of each treatment and always use the current favourite for $t \geqslant km$, there is a positive probability of eventually settling on an inferior treatment.

1.2. Asymptotic optimality. A decision rule is said to be *asymptotically optimal* (a.o.) if it leads to the result that, with probability 1, as $T \to \infty$,

$$R/T \to \max(p_1, p_2, \ldots, p_k). \tag{1.2}$$

This definition was introduced in the fundamental paper by Robbins (1952). He described a simple, but artificial policy for the two-armed bandit problem. Let $a_1 = 1 < a_2 < a_3 \ldots$ and $b_1 = 2 < b_2 < b_3 \ldots$ be two disjoint sequences of positive integers with $a_s/s \to \infty$ and $b_s/s \to \infty$ as $s \to \infty$. Suppose that $k = 2$ and consider the situation after t trials: if $t+1 = a_s$ or $t+1 = b_s$ for some $s \geqslant 1$, then the next trial must use p_1 or p_2, if $t+1$ is not a member of either sequence, the next trial is assigned to p_1 or p_2 according as $r_1/n_1 \geqslant r_2/n_2$ or $r_1/n_1 < r_2/n_2$. In other words, most of the trials are allocated sequentially to the favourite, but certain trials are reserved in advance for each of the treatments. The asymptotic optimality of the procedure can be verified easily, by using the strong law of large numbers.

The investigation of a.o. policies has not received much attention in the literature, perhaps because the example given by Robbins involves an artificial distinction between neighbouring trials. Another possible disadvantage is that the property (1.2) can only be achieved by ensuring that an infinite sequence of trials is associated with each of the unknown probabilities. However, it is not necessary to prescribe the sequences in advance: in fact, more sensitive procedures can be defined by generating them from the data as it develops.

We now introduce another class of policies based on a simple modification of the "play the favourite" rule. Roughly speaking, the idea is to add small random perturbations to the observed proportions r_j/n_j, at each stage, obtaining a set of indices which can be used as in (1.1). Let $\{\lambda(n), n \geqslant 1\}$ be a sequence of strictly positive constants such that $\lambda(n) \to 0$ as $n \to \infty$ and let $X_j(t)$, $j = 1, 2, \ldots, k$, $t \geqslant k$, be i.i.d. random variables which are positive and unbounded. The common distribution function will be denoted by F and we suppose that $F(0) = 0$ and $F(x) < 1$, for all $x > 0$. The allocation procedure uses k of these random variables at each stage and it will be assumed that all of them and the results of all the Bernoulli trials are independent of one another. Initially, each treatment is used once and then, for $t \geqslant k$, the allocations depend on the record of successes and failures according to the following rule.

Use treatment i in the $(t+1)$th trial if and only if i is the smallest integer such that $Q_i(t) = \max_{1 \leqslant j \leqslant k} Q_j(t)$, where

$$Q_j(t) = \frac{r_j}{n_j} + \lambda(n_j) X_j(t), \qquad j = 1, 2, \ldots, k. \tag{1.3}$$

The effectiveness of these randomised allocation procedures is not easy to evaluate, because of the complicated mechanism generating the total number of successes after T trials: $R(T) = \sum R_j(T)$, where $R_j = r_j(T)$, $j = 1, 2, \ldots, k$. However, the assumption that the perturbations $\lambda(n_j) X_j(t)$ are positive and unbounded guarantees that (1.3) determines a policy which is a.o.: we note the asymptotic properties established in an earlier paper (Bather (1980)).

THEOREM 1. *Consider the randomised allocation procedure defined by (1.3) and suppose that $\max(p_1, p_2, \ldots, p_k) = p_i > p_j$ for $j \neq i$. Then, with probability 1, the random variables $R_j(T)$, $N_j(T)$ have the following properties as $T \to \infty$:*

$$N_j(T) \to \infty, \qquad R_j(T)/N_j(T) \to p_j, \qquad j = 1, 2, \ldots, k,$$

$N_i(T)/T \to 1$ and $R(T)/T \to p_i$.

1.3. An empirical decision rule. It is clear that a great many a.o. policies can be devised. Indeed, Theorem 1 has already been extended by

Glazebrook (1980) to show that policies based on dynamic allocation indices can be modified, by adding random perturbations, to produce similar asymptotic properties. Another method of ensuring good performance over long sequences of trials was suggested by Poloniecki (1978). However, a more important question is whether we can select from the large class of a.o. procedures a decision rule that performs well over finite sequences of trials: this will be the main objective in what follows.

The randomised allocation procedures we have introduced depend on an arbitrary sequence $\{\lambda(n)\}$ and a distribution function F. This allows a wide range of possibilities and it must be admitted that no single procedure will emerge which is uniquely preferable to all the others. The particular decision rule suggested here is a result of many comparisons and minor adjustments, using trial and error methods. The empirical evidence will be reported in Sections 2,3 and 5, and this is restricted to the case $k = 2$. However, the form of the policy and the close relation between one-armed and multi-armed bandit problems established by Gittins, for discounted models, permits some anticipation of its general behaviour when $k \geqslant 3$. There are strong indications that the following policy, or one similar to it, is preferable to other methods of allocating the treatments, at least when $T \geqslant 50$.

Policy (i). Define $\lambda(n) = (4+n^{1/2})/(15n)$, $n = 1, 2, \ldots$, and let $X_j(t) = 2+ + Y_j(t)$, where the independent random variables $Y_j(t)$ have the common probability density e^{-y}, $y > 0$.

As a preliminary to the detailed comparison of performance between this and other policies, it may be helpful to look at the conditional probabilities in various situations with $k = 2$. Let $\pi_j = \pi_j(r_1, n_1, r_2, n_2)$ denote the conditional probability associated with p_j in the next trial and let $\lambda_j = \lambda(n_j)$, $j = 1, 2$,

$$q = \frac{r_2}{n_2} - \frac{r_1}{n_1} + 2(\lambda_2 - \lambda_1).$$

It is easily verified that, for policy (i),

$$\pi_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2} \exp\left(-\frac{q}{\lambda_1}\right). \tag{1.4}$$

This formula holds only if $q \geqslant 0$ but, otherwise, we can rely on the symmetry of the policy: in general, $\pi_1(r_1, n_1, r_2, n_2) = \pi_2(r_2, n_2, r_1, n_1)$. Table 1 shows that randomisation has little effect on preference for the "favourite" in the next trial, except in situations where the data would be regarded as inconclusive by most statisticians.

The next section of the paper is concerned with a special case of the one-armed bandit problem in which the optimal Bayes procedures are easy to compute and compare with simple randomised allocation procedures.

Table 1

Policy (i): conditional probabilities π_1 associated with p_1, $k = 2$

r_1	0	0	0	0	0	1	0	1	3	4	11	12
n_1	1	1	2	1	2	2	2	5	10	10	25	25
r_2	1	1	1	3	3	3	50	50	50	50	50	50
n_2	1	2	2	5	5	5	100	100	100	100	100	100
π_1	0.025	0.362	0.031	0.594	0.072	0.899	0.397	0.144	0.063	0.515	0.201	0.897

Section 3 discusses the two-armed bandit problem and compares the performance of policy (i) with several others, including the well-known "play the winner" rule. Unfortunately, most of the evaluations of our criterion, e.s.l., are based on simulations. However, the asymptotic results can be strengthened and, as will be seen in Section 4, there are mathematical reasons for choosing the sequence $\{\lambda(n)\}$ so that $\lambda(n)$ is of order $n^{-1/2}$ as $n \to \infty$. Section 5 contains a summary of the properties of policy (i), showing that its maximum loss $M_1(T)$ is about $0.36\ T^{1/2}$ for all $T \geqslant 50$. Such results will enable us to draw some conclusions about the effective cost of using decision procedures with equal allocations of the treatments under comparison. The final section includes some remarks on the possible theoretical and practical implications of the work.

Consider the case $k = 2$ and compare the fact that $M_1(T)$ is of order $T^{1/2}$ with the corresponding result for any finite expermient which relies on equal samples. Suppose we take S observations on each of p_1 and p_2:S might be fixed in advance or determined by a sequential stopping rule. Let P_{in} be the probability that $S = n$ and the terminal decision chooses p_i for all remaining $T-2n$ patients. We assume that

$$E(S) = \sum n(P_{1n}+P_{2n}) < \infty.$$

Then, for a total of T patients, the e.s.l. is

$$L_0(p_1, p_2, T) = (p_1-p_2) \sum_{2n \leqslant T} \{nP_{1n}+(T-n)P_{2n}\}.$$

This holds when $p_1 \geqslant p_2$, otherwise, there is a similar formula. It follows that

$$L_0(p_1, p_2, T)/T \to |p_1-p_2|\,\varepsilon_0, \qquad T \to \infty$$

where $\varepsilon_0 = \sum P_{2n}$ if $p_1 > p_2$, $\varepsilon_0 = \sum P_{1n}$ if $p_1 \leqslant p_2$. In other words ε_0 is the error probability of the decision procedure. Hence, the corresponding maximum e.s.l. $M_0(T)$ is of order T.

2. One-armed bandits

2.1. Bayes procedures. The case when $k = 2$ and p_2 is known was first studied by Bellman (1956), using a discount factor, and by Bradt, Johnson and Karlin (1956), for finite sequences of trials. Both papers established the form of the optimal policies for any given prior distribution on p_1, leaving substantial computations to produce explicit solutions. We shall concentrate on a special formulation of the one-armed bandit problem in order to simplify and extend some of the results obtained in the paper by Bradt et al.

Suppose that $p_2 = \frac{1}{2}$ and let α be a prescribed constant in the range $\frac{1}{2} < \alpha < 1$. We admit two simple hypotheses about p_1:

$$H_1\colon\ p_1 = \alpha, \qquad H_2\colon\ p_1 = 1-\alpha,$$

each with prior probability $\frac{1}{2}$. Thus, H_1 means that $p_1 > p_2$. Under these conditions, it is clear the posterior probabilities associated with H_i, $i = 1, 2$, after any sequence of trials, depend only on r_1 and n_1. In fact, the posterior probability of H_1 is

$$\theta_j = \frac{\alpha^j}{\alpha^j + (1-\alpha)^j}, \qquad j = 2r_1 - n_1. \tag{2.1}$$

We now have a convenient state variable j which changes by ± 1 whenever a trial is allocated to p_1, but remains constant when p_2 is used. The Bayes procedure can be determined by investigating a Markov decision process in which the essential variables are j and s, the number of trials that remain to be allocated.

Consider starting from an arbitrary state $j = 0, \pm 1, \pm 2, \ldots$, with s trials remaining, and let s_i be the number allocated to p_i, $i = 1, 2$, according to a decision rule Δ. Under H_1, s_2 represents the number of "mistakes" and the e.s.l. is $(\alpha - \frac{1}{2}) E_\Delta(s_2 \mid H_1)$. Similarly, under H_2, the e.s.l. is $(\alpha - \frac{1}{2}) E_\Delta(s_1 \mid H_2)$. The average loss with respect to the initial state j is

$$(\alpha - \tfrac{1}{2})\{\theta_j E_\Delta(s_2 \mid H_1) + (1-\theta_j) E_\Delta(s_1 \mid H_2)\}.$$

Clearly, we can omit the factor $(\alpha - \frac{1}{2})$ and define an optimal policy as one that attains the minimum expected number of mistakes. Let $B_j(s)$ denote this minimum expectation so that, in general, the minimum e.s.l. is $(\alpha - \frac{1}{2}) B_j(s)$.

We now obtain the dynamic programming equation for the function $B_j(s)$ by considering the effect of the next trial, given the information represented by the pair (j, s). If p_1 is used next, followed by an optimal sequence of decisions, the expected number of mistakes is

$$(1-\theta_j) + \{\theta_j \alpha + (1-\theta_j)(1-\alpha)\} B_{j+1}(s-1) + \{\theta_j(1-\alpha) + (1-\theta_j)\alpha\} B_{j-1}(s-1).$$

On the other hand, if p_2 is used, the total expectation must be replaced by $\theta_j + B_j(s-1)$. We consider a randomised choice between the two alternatives and associate conditional probabilities φ and $(1-\varphi)$ with p_1 and p_2, respectively. Then $B_j(s)$ must be determined by minimising with respect to φ:

$$B_j(s) = \min_{0 \leqslant \varphi \leqslant 1} [\varphi(1-\theta_j) + (1-\varphi)\theta_j + \varphi\{\theta_j \alpha + (1-\theta_j)(1-\alpha)\} B_{j+1}(s-1) + \\ + \varphi\{\theta_j(1-\alpha) + (1-\theta_j)\alpha\} B_{j-1}(s-1) + (1-\varphi) B_j(s-1)]. \tag{2.2}$$

The first conclusion to be drawn from this relation is that randomisation is never strictly needed. This follows from the fact that the expression on the right is linear in φ. In general, the minimum can be attained by setting $\varphi = 0$ or 1 and, where there is no difference, we may choose $\varphi = 0$ for definiteness. It is worth remarking that randomised allocation procedures cannot be justified by exact optimality for any choice of prior distribution on the unknown probabilities.

In principle, any desired value of the function $B_j(s)$ can be computed by successive applications of (2.2). By definition, $B_j(0) = 0$ for all j, so we can use the relation with $s = 1$ to determine $B_j(1)$; then set $s = 2$ and so on. Notice that, having computed $B_j(1)$ for $j = 0, \pm 1, \ldots, \pm(T-1)$, the values of $B_j(2)$ can be found for $j = 0, \pm 1, \ldots, \pm(T-2)$ and hence, after T iterations, $B_0(T)$ can be determined. Table 2 below gives some examples of the minimum e.s.l., $(\alpha - \frac{1}{2})B_0(T)$, obtained in this way. Policy (ii) in the table refers to the corresponding Bayes procedures: it is not a single policy because of its dependence on the parameters T and α.

Results about the form of the Bayes procedures can be inferred directly from relation (2.2), but all the properties we shall need are special cases of results obtained previously. The prior distribution on p_1 is represented here by the initial state $j = 0$ and the optimal policy, given a total of T trials, is determined by assigning the appropriate value $\varphi = \varphi_j(s)$ to every pair (j, s) that is accessible from the position $(0, T)$. When the properties established in Section 4 of the paper by Bradt et al. are expressed in this notation, they can be summarised as follows.

LEMMA 1. *There is a sequence* $\{\sigma(s)\}$ *with* $\sigma(1) = 0$, $\sigma(2) = -1$, $\sigma(s+1) = \sigma(s)$ *or* $\sigma(s+1) = \sigma(s) - 1$, *in general, and* $\sigma(s) \to -\infty$ *as* $s \to \infty$. *The Bayes procedure, policy* (ii), *is specified by the rule*:

$$\varphi_j(s) = 1 \quad \textit{if} \quad j > \sigma(s), \qquad \varphi_j(s) = 0 \quad \textit{if} \quad j \leqslant \sigma(s).$$

Thus, whenever $T \geqslant 2$, the policy involves a sequence of trials using the unknown probability p_1, followed by a switch to the known probability $p_2 = \frac{1}{2}$ for all remaining trials if the random process $\{j(t), t \geqslant 0\}$ reaches the boundary where $j(t) = \sigma(T-t)$, for some $t < T$. In practice, the sequence $\{\sigma(s)\}$ is best determined by finding the integers $s(\nu) = \min\{s\colon \sigma(s) = -\nu\}$ for $\nu = 1, 2, \ldots$ For example, when $\alpha = 0.6$, computations based on (2.2) show that $s(1) = 2$, $s(2) = 6$, $s(3) = 17$, $s(4) = 34$, $s(5) = 60$, $s(6) = 99$. The fact that $\sigma(s) \to -\infty$ as $s \to \infty$ implies that the limiting form of the optimal policy as $T \to \infty$ corresponds to the rule: allocate all the trials to p_1.

Similar results hold when the model includes a discount factor $a < 1$ and $T = \infty$. In this case, the Bayes procedure is stationary and the sequence $\{\sigma(s)\}$ is replaced by a negative integer σ, depending on α, with the general rule: use p_1 in the next trial if and only if the current state $j(t) > \sigma$. However, the same difficulty occurs when the discount factor is near 1: $\sigma \to -\infty$ as $a \uparrow 1$.

These results do not mean that there are no stationary policies with good long-term behaviour, but it is no use relying on the limiting form of relation (2.2). We know that the long-term average e.s.l. of an a.o. policy must be zero, so the theory of Markov decision processes indicates a stationary form of the dynamic programming equation. Unfortunately, this has no solution.

LEMMA 2. *The equation*

$$B_j = \min_{0 \leqslant \phi \leqslant 1} (\phi(1-\theta_j)+(1-\phi)\theta_j+\phi\{\theta_j\alpha+(1-\theta_j)(1-\alpha)\}B_{j+1}+ \\ +\phi\{\theta_j(1-\alpha)+(1-\theta_j)\alpha\}B_{j-1}+(1-\phi)B_j] \qquad (2.3)$$

has no finite solution with $B_j \geqslant 0$ *for all* j.

Proof. Suppose that $\{B_j, j = 0, \pm 1, \ldots\}$ is a non-negative solution. The expression on the right of (2.3) is linear in ϕ and, when $\phi = 0$, it reduces to $\theta_j + B_j > B_j$. It follows that the minimum must be attained when $\phi = 1$ and we have

$$B_j = 1-\theta_j+\{\theta_j\alpha+(1-\theta_j)(1-\alpha)\}B_{j+1}+\{\theta_j(1-\alpha)+(1-\theta_j)\alpha\}B_{j-1}. \qquad (2.4)$$

Hence, $B_j \geqslant 1-\theta_j$ always and we can strengthen this inequality by substituting the corresponding lower bounds for B_{j+1} and B_{j-1} on the right of (2.4). But equation (2.1) shows that

$$\{\theta_j\alpha+(1-\theta_j)(1-\alpha)\}(1-\theta_{j+1})+\{\theta_j(1-\alpha)+(1-\theta_j)\alpha\}(1-\theta_{j-1}) = 1-\theta_j$$

and we now have $B_j \geqslant 2(1-\theta_j)$. Then, by repeating the argument, it follows that $B_j \geqslant m(1-\theta_j)$ for any positive integer m and, since $\theta_j < 1$, B_j cannot be finite.

2.2. Stationary policies. A stationary allocation rule for the one-armed bandit problem with simple hypotheses H_1 and H_2 is defined by a sequence $\{\phi_j\}$ where ϕ_j is the conditional probability of using p_1 in the next trial, given $j = 2r_1 - n_1$. It is easily shown that randomisation plays an essential part in the conditions for such a policy to be a.o.

LEMMA 3. (a) *No deterministic stationary policy can be a.o.*

(b) *A stationary policy* $\{\phi_j\}$ *is a.o. provided that* $\phi_j > 0$ *for all* j *and that* $\phi_j \to 1$ *as* $j \to \infty$, $\phi_j \to 0$ *as* $j \to -\infty$.

Proof. (a) The policy defined by setting $\phi_j = 1$ for all j cannot be a.o. because its proportion of successes is $1-\alpha < \frac{1}{2}$, under H_2. On the other hand, a stationary policy with $\phi_j = 0$ for some j has at least one absorbing state, which means that the proportion of successes converges to $\frac{1}{2}$ with positive probability, under H_1.

(b) Strictly speaking, the proof of Theorem 1 given in Bather (1980) does not apply here, but it only needs minor modifications. Consider the random variables $n_1(t)$ and $j(t) = 2r_1(t) - n_1(t)$ as $t \to \infty$. Since ϕ_j is bounded away from zero on every bounded set of states, it is easily shown that $n_1(t) \overset{\text{a.s.}}{\to} \infty$ as $t \to \infty$, under either hypothesis. Then the strong law of large numbers guarantees that $r_1(t)/n_1(t) \overset{\text{a.s.}}{\to} p_1$ and it follows that $j(t) \overset{\text{a.s.}}{\to} \infty$, when $p_1 = \alpha > \frac{1}{2}$, and $j(t) \overset{\text{a.s.}}{\to} -\infty$, when $p_1 = 1 - \alpha$. Hence, the conditional probability $\pi_1(t) = \phi_{j(t)} \overset{\text{a.s.}}{\to} 1$, under H_1, and $\pi_1(t) \overset{\text{a.s.}}{\to} 0$, under H_2, because of the asymptotic behaviour of the sequence $\{\phi_j\}$. Finally, it can be established by considering a suitable martingale, as in the proof of Theorem 1, that $n_1(t)/t \overset{\text{a.s.}}{\to} 1$, under H_1, and $n_1(t)/t \overset{\text{a.s.}}{\to} 0$, under H_2. This leads to the required result.

The class of randomised allocation procedures described in Section 1.2 is easily adapted to the one-armed bandit problem: we simply define $Q_2(t) = p_2$ for all t, when p_2 is known. Thus, when $p_2 = \frac{1}{2}$, the general rule allocates the next trial to p_1 whenever $2r_1 - n_1 + 2n_1\lambda(n_1)X_1(t) \geqslant 0$. In particular, by setting $\lambda(n) = n^{-1}$ for each $n \geqslant 1$, we obtain a stationary policy with $\phi_j = 1$, $j > 0$, and $\phi_j = P(X_1(t) \geqslant -j/2)$, $j \leqslant 0$. The stationary policies arising in this way have two properties, both of which are intuitively sensible: $\{\phi_j\}$ is non-decreasing in j and $\phi_j = 1$ when j is positive. The first of these is not easy to justify precisely, but the second is enough to guarantee admissibility within the class of stationary policies. This will be made clear in Lemma 4.

Let us consider the e.s.l. for an arbitrary stationary policy $\{\phi_j\}$. It is convenient to work with the expected number of mistakes, denoted by $C_j(s)$, under H_1, and by $D_j(s)$, under H_2. These functions satisfy recurrence relations similar to (2.2), except that $\phi = \phi_j$ is prescribed for every state j, and it is a straightforward matter to compute values of the e.s.l., $L(p_1, p_2, T)$, by using the equations:

$$L(\alpha, \tfrac{1}{2}, T) = (\alpha - \tfrac{1}{2})C_0(T), \qquad L(1-\alpha, \tfrac{1}{2}, T) = (\alpha - \tfrac{1}{2})D_0(T).$$

We have $C_j(0) = D_j(0) = 0$ for all j and, in general:

$$C_j(s) = \phi_j\{\alpha C_{j+1}(s-1) + (1-\alpha)C_{j-1}(s-1)\} + (1-\phi_j)\{1 + C_j(s-1)\}, \quad (2.5)$$

$$D_j(s) = \phi_j\{1 + (1-\alpha)D_{j+1}(s-1) + \alpha D_{j-1}(s-1)\} + (1-\phi_j)D_j(s-1). \quad (2.6)$$

In what follows, the total number of trials will be treated as an unknown parameter and we need an appropriate definition of admissibility. An allocation procedure is said to be admissible for the one-armed bandit problem with p_2 given, if its risk function $L(p_1, p_2, T)$ cannot be reduced uniformly in $p_1 \in [0, 1]$ and $T \geqslant 1$. In general, the property is difficult to establish because of the wide range of possible decision rules, but weaker results can be obtained, for example, in the case $p_2 = \frac{1}{2}$, by restricting to comparisons between stationary policies.

LEMMA 4. *Let $\{\phi_j\}$ be a stationary policy with $0 < \phi_j \leqslant 1$, for all j, and $\phi_j = 1$, for $j \geqslant 1$. Let $\{\phi'_j\}$ be another stationary policy and suppose that the corresponding risk functions satisfy $L'(p_1, \frac{1}{2}, T) \leqslant L(p_1, \frac{1}{2}, T)$, for $p_1 = \alpha$, $p_1 = 1-\alpha$ and $T \geqslant 1$, where α is a constant in the range $\frac{1}{2} < \alpha < 1$. Then the two policies must be identical.*

Proof. In the notation of (2.5) and (2.6), we have $C'_0(T) \leqslant C_0(T)$ and $D'_0(T) \leqslant D_0(T)$, for all T. When $T = 1$, these inequalities reduce to $1-\phi'_0 \leqslant 1-\phi_0$ and $\phi'_0 \leqslant \phi_0$, so that $\phi'_0 = \phi_0$. Let us assume, inductively, that $\phi'_j = \phi_j$ for $|j| \leqslant T-1$. Relation (2.5) shows that $C_0(T+1)$ and $C'_0(T+1)$ depend only on ϕ_j and ϕ'_j, for $|j| \leqslant T$, and their difference is completely determined by contributions associated with changes of state leading to $j = \pm T$ in exactly T steps. It follows that

$$C_0(T+1)-C'_0(T+1) = \alpha^T \phi_0 \phi_1 \ldots \phi_{T-1}(\phi'_T-\phi_T)+$$
$$+(1-\alpha)^T \phi_0 \phi_{-1} \ldots \phi_{1-T}(\phi'_{-T}-\phi_{-T}).$$

Similarly, (2.6) shows that

$$D_0(T+1)-D'_0(T+1) = (1-\alpha)^T \phi_0 \phi_1 \ldots \phi_{T-1}(\phi_T-\phi'_T)+$$
$$+\alpha^T \phi_0 \phi_{-1} \ldots \phi_{1-T}(\phi_{-T}-\phi'_{-T}).$$

In particular,

$$\alpha^T \{C_0(T+1)-C'_0(T+1)\}+(1-\alpha)^T \{D_0(T+1)-D'_0(T+1)\} \geqslant 0$$

and, since $\alpha^{2T} > (1-\alpha)^{2T}$, we must have $\phi'_T \geqslant \phi_T$. But $\phi_T = 1$, so that $\phi'_T = 1$ also. Then the fact that both $C_0(T+1)-C'_0(T+1)$ and $D_0(T+1)-D'_0(T+1)$ are non-negative implies that $\phi'_{-T} = \phi_{-T}$. This completes the induction.

2.3. Some comparisons. Tables 2 and 3 illustrate the effects of various policies for the one-armed bandit problem with p_1 unknown and $p_2 = \frac{1}{2}$, given. Table 2 gives the Bayes risk for the optimal policy (ii), when the prior distribution assigns equal probabilities to H_1: $p_1 = \alpha$ and H_2: $p_1 = 1-\alpha$. The entries in this column were based on equation (2.2). The other entries give the corresponding average e.s.l. obtained from computations using equations (2.5) and (2.6). The policies (iii), ..., (vi) are all stationary and they can be regarded as special cases of the general allocation procedure described in Section 1.2, with the same sequence $\{\lambda(n)\}$. In each case, the randomisation is defined by setting $X_j(t) = bY_j(t)+c$, where b, c are constants and the independent random variables $Y_j(t)$ all have the probability density e^{-y}, $y > 0$.

Policies (iii) *and* (iv). $\lambda(n) = n^{-1}$, $n \geqslant 1$, and $b = 0$. Both policies are deterministic, with $c = 1.25$ in case (iii) and $c = 1.75$ in case (iv). For the one-armed bandit problem, the corresponding rules are of the form: allocate the next trial to p_1 so long as $j(t) > \sigma$, but switch to p_2 for all the remaining trials if the state σ is reached. The switch occurs at $\sigma = -3$ in policy (iii) and $\sigma = -4$ in policy (iv).

Policy (v). $\lambda(n) = n^{-1}$, $n \geqslant 1$, $b = 1/\log 100 = 0.2171$, and $c = 1 - \log 2/\log 100 = 0.8495$. This defines a sequence of conditional probabilities for the one-armed bandit problem with $\phi_j = 1$, for $j \geqslant -1$, $\phi_{-2} = 1/2$, $\phi_{-3} = 1/20$, $\phi_{-4} = 1/200$, etc.

Policy (vi). $\lambda(n) = n^{-1}$, $n \geqslant 1$, $b = 1/\log 4 = 0.7213$, $c = 1$. Thus, $\phi_j = 1$ for $j \geqslant -2$, $\phi_{-3} = 1/2$, $\phi_{-4} = 1/4$, $\phi_{-5} = 1/8$, etc.

Both policies (v) and (vi) are a.o.

The final column in Table 2 represents the best that can be done by prescribing a fixed sample of n_1 trials with p_1, followed by a simple test to decide whether to switch to p_2 for the remaining $T - n_1$ trials. The sample size (n_1) is shown in each case, next to the minimum average e.s.l.

Policy (vii). If $j = 2r_1 - n_1 < 0$ after n_1 trials with p_1, switch to p_2 for the other $T - n_1$ trials; otherwise, use p_1 throughout.

Table 2

Policies (ii),..., (vii): e.s.l. for symmetric prior, $p_1 = \alpha$ or $1 - \alpha$, $p_2 = \frac{1}{2}$

	α	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	
$T = 20$	0.55	0.408	0.412	0.427	0.417	0.441	0.430	(7)
	0.6	0.645	0.660	0.714	0.680	0.768	0.726	(5)
	0.7	0.760	0.814	0.967	0.876	1.15	0.989	(5)
	0.8	0.655	0.783	1.00	0.885	1.33	0.980	(3)
	0.9	0.543	0.754	1.00	0.882	1.43	0.790	(3)
$T = 50$	0.55	0.889	0.900	0.894	0.909	0.946	0.981	(15)
	0.6	1.19	1.22	1.21	1.25	1.39	1.50	(15)
	0.7	1.08	1.08	1.13	1.19	1.61	1.71	(11)
	0.8	0.855	0.855	1.03	1.04	1.69	1.48	(7)
	0.9	0.617	0.763	1.00	1.00	1.77	1.13	(3)
$T = 100$	0.55	1.51	1.64	1.55	1.61	1.60	1.76	(31)
	0.6	1.71	2.00	1.77	1.92	1.96	2.40	(25)
	0.7	1.30	1.48	1.30	1.47	1.90	2.35	(15)
	0.8	0.973	0.973	1.05	1.15	1.95	1.89	(9)
	0.9	0.741	0.776	1.00	1.08	2.02	1.33	(5)

We note that policies (ii) and (vii) depend on both the parameters α and T, whereas the other policies do not. Table 2 shows that policies (iii) and (iv) are quite effective over the range of parameters considered, except when α is near 1, and the randomised procedure (v) is almost as good. Policy (vi) is not so effective, but even this is comparable with the best fixed sample procedure (vii). However, the advantage of the extra randomisation used in policy (vi) becomes much clearer when the range of T is extended.

Table 3 gives a rough idea of the performance of the stationary policies (iii),...,(vi), for values of T up to 500. For these policies, the e.s.l. was estimated from only 100 simulations, so the standard errors are of order $10^0/_0$. Nevertheless, the pattern is reasonably clear. Policy (i) is the empirical rule described in Section 1.3 and, in this case, the e.s.l. is based on 1000 simulations. It is not so surprising that such a rule can be more effective than the others, when the range of parameter values is considered as a whole: the state variable j is sufficient when α is known, but not otherwise.

Table 3

Policies (i), (iii),..., (vi): e.s.l. from simulations, $p_2 = \frac{1}{2}$

	p_1	(i)	(iii)	(iv)	(v)	(vi)
$T = 50$	0.2	2.59	1.4	1.7	2.2	3.5
	0.4	2.18	1.1	1.9	1.4	2.3
	0.45	1.45	0.9	0.9	1.0	1.4
	0.55	0.45	1.1	0.8	0.8	0.5
	0.6	0.50	1.3	0.7	1.1	0.5
$T = 100$	0.2	2.99	1.5	1.8	2.3	4.1
	0.4	2.94	1.3	2.3	1.8	2.9
	0.45	2.29	1.3	1.1	1.3	1.9
	0.55	0.96	2.5	1.9	1.9	1.0
	0.6	0.88	2.9	1.5	2.3	0.8
$T = 200$	0.2	3.38	1.5	1.9	2.4	4.6
	0.4	3.72	1.5	2.5	2.1	3.7
	0.45	3.36	1.6	1.3	1.8	2.6
	0.55	1.87	5.4	4.2	3.9	2.2
	0.6	1.47	6.1	3.2	4.5	1.3
$T = 500$	0.2	3.92	1.5	1.9	2.6	5.3
	0.4	4.67	1.5	2.5	2.3	4.4
	0.45	5.09	1.7	1.5	2.2	3.7
	0.55	4.07	14.2	11.8	9.6	5.5
	0.6	2.64	15.4	8.3	10.6	2.5

2.4. Minimax policy for a diffusion model. Explicit results are hard to find, even for the one-armed bandit problem, but *I* have recently discovered the minimax allocation rule for a discounted, continuous version of the problem. For comparison with the discrete case, suppose that $p_2 = \frac{1}{2}$ and p_1 is unknown, but not too far from $\frac{1}{2}$: this is the important case in the long run. We write $s = n_1$ and $x = j = 2r_1 - n_1$, treating these as continuous variables. Then $x(s)$ is approximately $N(\mu s, s)$, where $\mu = 2p_1 - 1$. Now introduce a discount factor a, $0 < a < 1$, and write $\alpha = -\log a$, not to be confused with the previous notation in which α was a fixed probability. It will be shown later that the minimax policy, given α, for sharing the time sequentially between the process

$$x(s) = \mu s + w(s) \tag{2.7}$$

with μ unknown and the zero process is to continue observing process (2.7) as long as

$$x(s) > -b\alpha^{-\frac{1}{2}} - c\alpha^{\frac{1}{2}} s$$

and stop whenever this linear boundary is reached. Here $\{w(s)\}$ is standard Brownian motion and b, c are constants which turn out to be: $b = 0.320$, $c = 0.584$.

This policy suggests that, for the discrete model, we should continue observations on p_1 so long as

$$j > -b\left(\log\frac{1}{a}\right)^{-\frac{1}{2}} - c\left(\log\frac{1}{a}\right)^{\frac{1}{2}} n_1. \tag{2.8}$$

Unfortunately the result does not help much when $a \uparrow 1$, since the limiting policy is degenerate. One interesting feature is that the expression on the right of (2.8) has an upper bound with respect to the discount factor $a < 1$. It is easily shown that the least upper bound is $-2(bcn_1)^{\frac{1}{2}} = -0.865 n_1^{\frac{1}{2}}$. Thus, the policy indicates that we should certainly continue observing p_1 if $j > -0.865 n_1^{\frac{1}{2}}$. Note that policy (i) means: continue with p_1 as long as

$$j > -\frac{2}{15}(4 + n_1^{\frac{1}{2}})(2 + Y_1(t)).$$

This is very roughly comparable if we think of the random perturbation here as an attempt to smooth over ignorance of the discount factor a in (2.8).

I will now sketch the proof of the minimax policy for process (2.7). Consider the stopping time τ determined by the line $x = -B - Cs$, where $B > 0$. Thus,

$$\tau = \inf\{s\colon\ w(s) \leqslant -B-(C+\mu)s\}.$$

A standard result from the theory of Brownian motion shows that

$$E\{e^{-\alpha\tau}\} = \exp\{-B[C+\mu+\sqrt{2\alpha+(C+\mu)^2}]\}. \tag{2.9}$$

The stopping time τ determines an allocation rule: the zero process is used for all $s > \tau$. To find the corresponding discounted e.s.l. note that if μ is known and $\mu > 0$ the maximum discounted reward is

$$E\left\{\int_0^\infty e^{-\alpha s}\,dx(s)\right\} = \mu\int_0^\infty e^{-\alpha s}\,ds = \frac{\mu}{\alpha}.$$

Hence, the discounted e.s.l. associated with τ is

$$L^{+}\,(\mu,\alpha)=\frac{\mu}{\alpha}-\mu E\left\{\int_{0}^{\tau}e^{-\alpha s}ds\right\}=\frac{\mu}{\alpha}E\left\{e^{-\alpha\tau}\right\}.$$

Similarly, when $\mu<0$ the corresponding formula is

$$L^{-}\,(\mu,\alpha)=\frac{-\mu}{\alpha}E\left\{1-e^{-\alpha\tau}\right\}$$

Hence, the loss function can be evaluated by using (2.9).

It is convenient to transform the parameters so that α is eliminated. Write $B=b\alpha^{-\frac{1}{2}}$, $C=c\alpha^{\frac{1}{2}}$ and $\mu=(\nu-c)\alpha^{\frac{1}{2}}$. Then it turns out that $L^{+}\,(\mu,\alpha)=\alpha^{-\frac{1}{2}}K^{+}(\nu)$, $L^{-}\,(\mu,\alpha)=\alpha^{-\frac{1}{2}}K^{-}(\nu)$, where

$$K^{+}(\nu)=(\nu-c)\exp\left\{-b(\nu+\sqrt{2+\nu^{2}})\right\},\qquad(\nu>c),$$

$$K^{-}(\nu)=(c-\nu)\left[1-\exp\left\{-b(\nu+\sqrt{2+\nu^{2}})\right\}\right],\qquad(\nu<c).$$

In effect, we can take $a=1$ and use these as the loss functions, with $\mu=\nu-c$.

Now consider two simple hypotheses.

$$H^{+}\colon\ \mu=(\eta-c)\alpha^{\frac{1}{2}},\qquad H^{-}\colon\ \mu=-(\eta+c)\alpha^{\frac{1}{2}}.$$

It will turn out that the constants b, c and η can be chosen with b and c positive, $\eta>c$. The likelihood ratio for H^{+} versus H^{-}, given observations in the interval $[0,s]$ is easily shown to be

$$\exp\left\{2\eta\alpha^{\frac{1}{2}}(x+c\alpha^{\frac{1}{2}}s)\right\}.$$

Hence, the posterior probabilities of H^{+} and H^{-} depend only on the corresponding prior probabilities Π^{+} and Π^{-} and the sufficient statistic $x(s)+c\alpha^{\frac{1}{2}}s$. Then it can be verified that the Bayes solution has the form: continue process (2.7) so long as $x(s)+c\alpha^{1/2}s\geqslant-B$. Here, B is a suitable constant; we shall take $B=b\alpha^{-\frac{1}{2}}$ and then determine b.

We demand that b, c and η satisfy the following conditions:

$$\sup_{\nu>c}K^{+}(\nu)=K^{+}(\eta),\qquad\sup_{\nu<c}K^{-}(\nu)=K^{-}(-\eta),\qquad K^{+}(\nu)=K^{-}(-\eta).\tag{2.10}$$

It is an awkward matter to verify that suitable values can be found but, in fact, they are as follows: $b=0.320$, $c=0.584$, $\eta=2.275$. Further, it can be shown that, for these values, the stopping rule is Bayes when the prior probabilities on H^{+} and H^{-} are $\Pi^{+}=0.372$ and $\Pi^{-}=0.628$. In other words, the Bayes risk $\Pi^{+}K^{+}(\eta)+\Pi^{-}K^{-}(-\eta)$ is a minimum with respect to the choice of b.

The allocation rule we have chosen has loss function $L(\mu, \alpha) = L^+(\mu, \alpha)$ or $L^-(\mu, \alpha)$ according as $\mu \geqslant 0$ or $\mu < 0$. This has the properties $\sup_\mu L(\mu, \alpha) = \alpha^{-\frac{1}{2}} K$, where

$$K = K^+(\eta) = K^-(-\eta) = 0.3465.$$

Further, there is a prior distribution on μ for which the minimum value of the Bayes risk is also $\alpha^{-\frac{1}{2}} K$. It follows that any other allocation rule Δ must have

$$\sup_\mu L_\Delta(\mu, \alpha) \geqslant \alpha^{-\frac{1}{2}} K.$$

Otherwise, it produces a Bayes risk

$$\Pi^+ L_\Delta(\alpha^{\frac{1}{2}}(\eta - c), \alpha) + \Pi^- L_\Delta(-\alpha^{\frac{1}{2}}(\eta + c), \alpha) < \alpha^{-\frac{1}{2}} K.$$

This contradicts the fact that the above policy is Bayes for this prior distribution. Hence, the policy is minimax.

3. Two-armed bandits

3.1. Feldman's rule. We now turn to the case $k = 2$ when both p_1 and p_2 are unknown. The aim here is to compare some of the policies derived from our study of the one-armed bandit with several others. Most Bayes procedures are difficult to evaluate, but there is a special class of prior distributions for which an optimal policy is determined by the following simple rule: always maximise the probability of success in the next trial. The rule is known to be optimal whenever the prior distribution is restricted to a pair of related hypotheses:

$$H_1\colon\ p_1 = \alpha,\ p_2 = \beta, \qquad H_2\colon\ p_1 = \beta,\ p_2 = \alpha,$$

where α and β are given probabilities. This result is due to Feldman (1962). In fact, the same rule produces policies which are optimal for the multi-armed bandit problem with $k \geqslant 3$, when there are k hypotheses of the form H_i: $p_i = \alpha$, $p_j = \beta$ for $j \neq i$, (see Rodman (1978)). Of course, Feldman's rule leads to different policies, depending on the values of α and β and the prior probabilities, and, unfortunately, they do not perform well if their special assumptions are violated.

In order to illustrate the rule, let us assume that $\beta = 1 - \alpha$, with $\frac{1}{2} < \alpha < 1$, and assign equal probabilities to H_1 and H_2. Then it is easy to see that the posterior probability of H_1 after t trials is at least $\frac{1}{2}$ if and only if $2r_1 - n_1 \geqslant 2r_2 - n_2$, Hence, Feldman's rule is equivalent to

Policy (viii). Use p_1 in the next trial if $2r_1-n_1 > 2r_2-n_2$ and p_2 if the reverse inequality holds; choose p_1 or p_2 at random in case of equality.

This procedure does not depend on the parameters α and T and, provided that $p_1+p_2 = 1$, its performance cannot be improved by any symmetric policy. The optimality property is, in this case, a simple consequence of the fact that information about p_1 and p_2 derived from any sequence of trials does not depend on the allocation rule: since $p_1 = 1-p_2$, a success observed in a trial with p_1 is equivalent to a failure with p_2, and vice-versa. If the prior distribution on the line $p_1+p_2 = 1$ is symmetric about its midpoint, policy (viii) maximises the probability of success at each stage and, hence, it produces the maximum expected number of successes in a sequence of T trials. In other words, the common value of its e.s.l. evaluated for $(p_1, p_2) = (\alpha, 1-\alpha)$ and $(p_1, p_2) = (1-\alpha, \alpha)$ is also the Bayes risk for the corresponding symmetric prior distribution. This value sets a lower bound on the e.s.l. for every decision rule Δ with $L_\Delta(\alpha, 1-\alpha, T) = L_\Delta(1-\alpha, \alpha, T)$, which includes all symmetric rules.

The special case of Feldman's result, with $p_1+p_2 = 1$, was first obtained by Bradt et al. (1956) and they also established that the rule is a.o. under the same condition. The asymptotic property can be extended, but it does not hold in general, contrary to the claim in Section 5 of the paper by Berry (1978). Table 4 below shows that the performance of policy (viii) depends critically on whether the condition $p_1+p_2 = 1$ is satisfied or not. Its long-term behaviour can be explained by considering the Markov chain with states $j = (2r_1-n_1)-(2r_2-n_2)$. This has transitions from j to $j+1$ or $j-1$ and the corresponding probabilities are: p_1 and $1-p_1$ if $j \geqslant 1$, $1-p_2$ and p_2 if $j \leqslant -1$, $\frac{1}{2}(1+p_1-p_2)$ and $\frac{1}{2}(1-p_1+p_2)$ if $j = 0$. We may suppose that $p_1 > p_2$. If $p_1 > \frac{1}{2} \geqslant p_2$, the stochastic process $\{j(t)\}$ has the property that $j(t) \overset{\text{a.s.}}{\to} \infty$ as $t \to \infty$, which is enough to guarantee that the policy is a.o. On the other hand, if $p_1 > p_2 > \frac{1}{2}$, $j(t) \to -\infty$ with probability

$$\frac{p_1(1-p_1+p_2)(2p_2-1)}{(p_1+p_2)(p_1+p_2-1)}.$$

For example, this probability is 0.3 when $p_1 = 0.9$ and $p_2 = 0.7$, which suggests that policy (viii) has an e.s.l. of order $0.06T$ when T is large: see Table 4. Finally, if $p_2 < p_1 < \frac{1}{2}$, the process $\{j(t)\}$ converges and a calculation of the limiting distribution shows that the long-term proportion of successes is

$$\frac{p_1(\frac{1}{2}-p_2)+p_2(\frac{1}{2}-p_1)}{1-p_1-p_2} < p_1.$$

Hence, the policy is not a.o. if $(p_1-\frac{1}{2})(p_2-\frac{1}{2}) > 0$, except in the trivial case when $p_1 = p_2$.

3.2. Other policies. Table 4 includes two other policies, both of which have the virtue of simplicity, but neither of them is very good. The first of these is the well-known "play the winner" rule introduced by Robbins (1952) and since investigated by several authors: see, for example, Zelen (1969) and Hoel, Sobel and Weiss (1972).

Policy (ix). If a success is observed in a trial with p_i, use the same p_i in the next trial, but switch to the other one whenever a failure occurs.

This rule determines a simple Markov chain whose state always corresponds to the current choice of p_i, $i = 1,2$. It is easily shown that, in the long run, it yields a proportion of successes

$$\frac{p_1(1-p_2)+p_2(1-p_1)}{(1-p_2)+(1-p_1)} < \max(p_1, p_2),$$

so the policy is not a.o., except when $p_1 = p_2$. The entries in the table were obtained, in this case, by assuming that the chain is stationary throughout.

A slight modification of policy (ix) is the "least failures" rule: use p_i in the next trial if it has produced the smallest number of failures in the past. In Section 1, we referred to optimal procedures determined by dynamic allocation indices and their dependence on the choice of a discount factor a. It has been established by Kelly that, when $a \uparrow 1$, the limiting form of any such policy for the multi-armed bandit problem is the least failures rule. This is in some respects a surprising result, since the limiting operation might be expected to produce a policy with good asymptotic properties.

The final policy in Table 4 illustrates the effect of using fixed samples of equal size, in order to test whether $p_1 > p_2$ or not.

Policy (x). Observe the results of 25 trials on each of p_1 and p_2; then use p_1 or p_2 in all the remaining $T-50$ trials, according to which sample produces the larger number of successes. The e.s.l. given in the table is based on a normal approximation of the error probabilities for the test. The performance cannot be much improved, even by choosing a different sample size $n_1 = n_2$, for each value of T.

Policies (i), (iv) and (vi) were defined in Sections 1 and 2. They are special cases of the general allocation rule (1.3), but only (i) and (vi) are a.o. The entries in Table 4 are averages of 1000 simulations for policy (i), but only 100 were used for policies (iv) and (vi). This means that the figures should be treated with some caution: for example, the theoretical e.s.l. in the cases $p_1 = 0.525$, $p_2 = 0.475$ and $p_1 = 0.55$, $p_2 = 0.5$ must be very similar, for all the policies considered. Simulations of policy (viii) showed that its behaviour is highly variable as well as being sensitive to the choice of p_1 and p_2. Because of this, the first four entries for each T were based on 200 runs.

Table 4

Two-armed bandit problem: e.s.l. for various policies

	p_1	p_2	(i)	(iv)	(vi)	(viii)	(ix)	(x)
$T = 50$	0.6	0.4	2.53	2.5	2.9	2.0	4.00	5.00
	0.9	0.7	2.18	1.9	2.4	2.8	2.50	5.00
	0.55	0.45	1.81	1.8	2.0	1.7	2.25	2.50
	0.6	0.5	1.84	1.9	2.0	1.6	2.22	2.50
	0.525	0.475	1.07	1.1	1.1	1.0	1.19	1.25
	0.55	0.5	1.08	0.9	1.1	0.9	1.18	1.25
$T = 100$	0.6	0.4	3.56	3.3	4.0	2.4	8.00	5.74
	0.9	0.7	2.86	2.5	3.2	5.5	5.00	5.34
	0.55	0.45	3.07	3.3	3.3	2.8	4.50	3.69
	0.6	0.5	3.13	3.1	3.4	2.9	4.44	3.69
	0.525	0.475	1.97	2.2	1.9	1.6	2.38	2.15
	0.55	0.5	2.01	1.6	2.1	1.7	2.37	2.15
$T = 200$	0.6	0.4	4.59	4.8	4.6	2.5	16.00	7.23
	0.9	0.7	3.51	3.0	4.0	10.8	10.00	6.02
	0.55	0.45	4.82	5.5	5.2	4.0	9.00	6.08
	0.6	0.5	4.95	4.9	5.3	4.6	8.89	6.06
	0.525	0.475	3.55	4.1	3.3	2.8	4.75	3.96
	0.55	0.5	3.54	2.7	3.8	3.0	4.74	3.96
$T = 500$	0.6	0.4	5.76	8.4	5.2	2.7	40.00	11.70
	0.9	0.7	4.27	3.7	4.8	26.7	25.00	8.06
	0.55	0.45	7.73	10.7	8.1	4.9	22.50	13.24
	0.6	0.5	7.79	9.2	7.9	8.5	22.22	13.19
	0.525	0.475	7.14	9.6	6.1	5.6	11.88	9.39
	0.55	0.5	7.21	5.7	7.8	5.9	11.84	9.38

The comparisons indicate that the empirical rule, policy (i), is effective over a wide range of parameter values and the results are similar to those obtained for the one-armed bandit problem. It is worth adding that a special case of the a.o. rule devised by Robbins (see Section 1.2 of this paper) has been investigated by Fox (1974), using simulations of $T = 50$, 100 and 1000 trials. Policy (i) seems to be a substantial improvement on this, especially when T is large. It would be interesting to include further Bayes procedures, in addition to Feldman's rule, policy (viii). Some computational studies have been carried out by Wahrenberger et al. (1977), using Beta priors, but their results are limited to the case $T = 50$. For example, they found that the Bayes procedure for a uniform prior distribution is comparable in performance with the rule studied by Fox.

4. Diffusion approximations

4.1. An invariance principle. It is intuitively clear that the long term behaviour of randomised allocation procedures can be approximated by suitable diffusion models, but we shall not attempt a detailed investigation of the approximations. The aim is to set up a normal version of the multiarmed bandit problem and, by using a well-known invariance property of Brownian motion, try to draw some tentative conclusions about the asymptotic behaviour of randomised procedures in discrete time.

A sequence of independent Bernoulli trials, each with probability p of success, is analogous to the stochastic process in continuous time given by

$$x(s) = ps + \sigma w(s), \tag{4.1}$$

where $\{w(s),\, s \geqslant 0\}$ is a standard Brownian motion and $\sigma^2 = p(1-p)$. The dependence of σ^2 on p here is an awkward feature because, in principle σ^2 can be determined without estimation error by continuous observation of $x(s)$ over any short period. In what follows, we shall be mainly interested in probabilities near $\frac{1}{2}$, where σ^2 attains its maximum, so let us assume that $\sigma = \frac{1}{2}$ in (4.1). For convenience, we write $\mu = p - \frac{1}{2}$, $y = x - \frac{1}{2}s$, so that the diffusion model becomes

$$y(s) = \mu s + \tfrac{1}{2} w(s). \tag{4.2}$$

Now consider a scale transformation

$$y' = cy, \qquad s' = c^2 s, \qquad \mu' = c^{-1}\mu, \tag{4.3}$$

where c is any positive constant. It is easily verified that the process $\{w'(s'),\, s' \geqslant 0\}$ defined by $w'(s') = cw(c^{-2}s')$ is again a standard Brownian motion and (4.2) is replaced by

$$y'(s') = \mu' s' + \tfrac{1}{2} w'(s'). \tag{4.4}$$

We can interpret this invariance property in the following way: data consisting of the observations $\{y(s),\, 0 \leqslant s \leqslant T\}$, when the drift parameter is μ, are precisely equivalent to the observations $\{y'(s'),\, 0 \leqslant s' \leqslant T'\}$, when the drift is μ', provided that the relations (4.3) hold and $T' = c^2 T$ for some constant c. Similar (1:1) mappings can be defined from one data set to another, over a different time interval, when we have k independent processes with unknown drift parameters.

Suppose that observation time must be shared between the processes

$$y_i(s_i) = \mu_i s_i + \tfrac{1}{2} w_i(s_i), \tag{4.5}$$

where the drift parameters μ_i are unknown and $\{w_i(s_i),\, s_i \geqslant 0\}$, $i = 1, 2, \ldots, k$, are independent Brownian motions. Each variable s_i records the observation time for process i after a period of length t and the control procedure must ensure that $\sum s_i = t$ at all times. It turns out that, in continuous time, there is no need to consider randomised allocation. The conditional probabilities $\pi_i = \pi_i(r_1, n_1, \ldots, r_k, n_k)$ can be replaced by control functions $\psi_i = \psi_i(y_1, s_1, \ldots, y_k, s_k)$ such that $\psi_i \geqslant 0$, $i = 1, 2, \ldots, k$, and $\sum \psi_i = 1$ always. Observation of the stochastic processes (4.5) is then governed by a set of differential equations: $ds_i = \psi_i\, dt$, $i = 1, 2, \ldots, k$. In order to avoid technical difficulties, let us restrict attention to a class K of well-behaved procedures $\psi = (\psi_1, \psi_2, \ldots, \psi_k)$. In particular, it will be assumed that, for each $\psi \in K$, the corresponding joint process $\{y_i(s_i), i = 1, 2, \ldots, k, t = \sum s_i \geqslant 0\}$ is well-defined.

Consider the results of applying an allocation procedure $\psi \in K$ over the period $[0, t]$. We are mainly interested in the e.s.l.

$$\mathscr{L}_\psi(\mu_1, \mu_2, \ldots, \mu_k, T) = T \max(\mu_1, \mu_2, \ldots, \mu_k) - E_\psi(Y),$$

where $Y = \sum y_i(S_i)$ is the sum of the final values, $\sum S_i = T$. The procedure ψ can be translated to any other period $[0, T']$ in the following way. Define $c > 0$ by $c^2 = T'/T$ and let

$$y_i'(s_i') = \mu_i' s_i' + \tfrac{1}{2} w_i'(s_i'), \tag{4.6}$$

where the $\{w_i'(s_i'), s_i' \geqslant 0\}$ are independent Brownian motions, $i = 1, 2, \ldots, k$. The transformed procedure ψ' is defined by setting

$$\psi_i'(y_1', s_1', \ldots, y_k', s_k') = \psi_i(c^{-1} y_1', c^{-2} s_1', \ldots, c^{-1} y_k', c^{-2} s_k') \tag{4.7}$$

for all values of the variables such that $\sum s_j' \leqslant T'$. Now suppose that $\mu_i' = c^{-1} \mu_i$, $i = 1, 2, \ldots, k$, and compare the joint distribution of the processes (4.5) under ψ with the corresponding joint distribution of the processes (4.6) under ψ'. We can set up a (1:1) correspondence between realisations $\{y_i(s_i), \sum s_i \leqslant T\}$ and $\{y_i'(s_i'), \sum s_i' \leqslant T'\}$, in which $y_i' = c y_i$ and $s_i' = c^2 s_i$, $i = 1, 2, \ldots, k$, always hold. This means that, in general,

$$w_i'(s_i') = c w_i(s_i) = c w_i(c^{-2} s_i'),$$

which shows that the correspondence is consistent with the assumption that both models are generated by Brownian paths. It follows that the joint distributions determined by (4.5) and (4.6) differ only by changes of scale. In particular, we may conclude that

$$\mathscr{L}_{\psi'}(\mu_1', \mu_2', \ldots, \mu_k', T') = c \mathscr{L}_\psi(\mu_1, \mu_2, \ldots, \mu_k, T)$$

and the definition of c shows that

$$T'^{-\frac{1}{2}} \mathscr{L}_{\psi'}(\mu_1', \mu_2', \ldots, \mu_k', T') = T^{-\frac{1}{2}} \mathscr{L}_\psi(\mu_1, \mu_2, \ldots, \mu_k, T). \tag{4.8}$$

We define

$$\mathscr{M}_\psi(T) = \sup_{\mu_1, \ldots, \mu_k} \mathscr{L}_\psi(\mu_1, \mu_2, \ldots, \mu_k, T), \qquad \mathscr{M}(T) = \inf_{\psi \in K} \mathscr{M}_\psi(T).$$

Then it follows immediately from (4.8) that

$$T'^{-\frac{1}{2}}\mathscr{M}_{\psi'}(T') = T^{-\frac{1}{2}}\mathscr{M}_{\psi}(T). \tag{4.9}$$

The mapping: $\psi \to \psi'$ associates each allocation procedure for the interval $[0, T]$ with a similar one for $[0, T']$ and vice-versa. Hence

$$T'^{-\frac{1}{2}}\mathscr{M}(T') = T^{-\frac{1}{2}}\mathscr{M}(T)$$

and, by setting $T' = 1$, we obtain

$$\mathscr{M}(T) = T^{\frac{1}{2}}\mathscr{M}(1), \tag{4.10}$$

for all $T > 0$. The relation (4.9) also shows that, if we can find a minimax procedure $\psi_T \in K$ which attains the infimum $\mathscr{M}(T)$, then it can be used to determine minimax procedures for all other intervals, simply by using the transformation (4.7). This does not mean that the same policy is optimal for all time intervals: for example, if $T' < T$, we have no reason to suppose that $\psi_{T'} = (\psi_T)'$ coincides with the restriction of ψ_T to $[0, T']$. In other words, ψ_T may depend on the given value of T. However, we are concerned with sequential allocation rules which can be applied without a knowledge of the time horizon and this indicates a study of invariant procedures.

An allocation procedure $\psi \in K$ will be called *invariant* if it is unaffected by the transformation (4.7). Thus, we say $\psi \in I$ if

$$\psi_i(y_1, s_1, \ldots, y_k, s_k) = \psi_i(cy_1, c^2 s_1, \ldots, cy_k, c^2 s_k), \tag{4.11}$$

for all $c > 0$ and any $y_i, s_i \geqslant 0$, $i = 1, 2, \ldots, k$. In this case, by setting $T' = 1$ and $\mu_i' = T^{\frac{1}{2}}\mu_i$ in (4.8), we obtain

$$\mathscr{L}_{\psi}(\mu_1, \mu_2, \ldots, \mu_k, T) = T^{\frac{1}{2}}\mathscr{L}_{\psi}(T^{\frac{1}{2}}\mu_1, \ldots, T^{\frac{1}{2}}\mu_k, 1). \tag{4.12}$$

Again, for $\psi \in I$, relation (4.9) can be expressed in the form

$$\mathscr{M}_{\psi}(T) = T^{\frac{1}{2}}\mathscr{M}_{\psi}(1). \tag{4.13}$$

Finally, there is an obvious analogue of (4.10) for the infimum $\mathscr{M}^I(T)$ with respect to $\psi \in I$:

$$\mathscr{M}^I(T) = T^{\frac{1}{2}}\mathscr{M}^I(1). \tag{4.14}$$

Since $I \subset K$, it is clear that $\mathscr{M}^I(T) \geqslant \mathscr{M}(T)$ and we know from (4.10) and (4.14) that the ratio of these two quantities is $\mathscr{M}^I(1)/\mathscr{M}(1)$, for all $T > 0$. The levels of minimax e.s.l. denoted by $\mathscr{M}^I(1)$ and $\mathscr{M}(1)$ are fundamental constants of the k-armed bandit problem. If we could determine these constants, their ratio would provide a very useful measure of the effective cost of ignorance about the time horizon.

4.2. Minimax decision rules. Our discussion of the normal version of the multi-armed bandit problem suggests the possibility of designing allocation procedures for Bernoulli trials so that the maximum e.s.l. is of order $T^{1/2}$ as $T \to \infty$. In view of relation (4.13), a decision rule Δ which is asymptotically invariant, in some sense, should also have the property that $M_\Delta(T)/T^{1/2}$ has a finite limit. We shall proceed heuristically, without attempting to prove the existence of such limits, but some explanation is due for the choice of the sequence $\{\lambda(n)\}$ in policy (i).

Consider the class of randomised allocation procedures defined by (1.3). Any such decision rule will be called *asymptotically invariant* if

$$n^{\frac{1}{2}}\lambda(n) \to d \qquad \text{as } n \to \infty, \tag{4.15}$$

for some $d > 0$. This definition can be explained by examining a simple diffusion model. The conditional probabilites $\pi_i(r_1, n_1, \ldots, r_k, n_k)$ are determined, at any stage, by positive i.i.d. random variables X_j through the quantities Q_j, $j = 1, 2, \ldots, k$. Let us replace the integers r_j and n_j, as in (4.2) and (4.5), by the continuous variables $y_j + \frac{1}{2}s_j$ and s_j, respectively. This leads to $Q_j = \frac{1}{2} + s_j^{-1}y_j + \lambda(s_j)X_j$ and each probability π_i is replaced by a function $\psi_i(y_1, s_1, \ldots, y_k, s_k)$. Clearly, the constant $\frac{1}{2}$ in every Q_j may be omitted in calculating the allocation probabilities, now to be interpreted as local proportions of continuous time assigned to the k processes. The invariance condition (4.11) means that the proportions ψ_i should remain constant when each Q_j depends on a scale parameter $c > 0$:

$$Q_j = c^{-1}s_j^{-1}y_j + \lambda(c^2 s_j)X_j, \qquad j = 1, 2, \ldots, k.$$

The question of which of these quantities is largest is unaffected by the value of c if and only if $\lambda(c^2 s_j) = c^{-1}\lambda(s_j)$. This is equivalent to demanding that $\lambda(s) = \lambda(1)s^{-\frac{1}{2}}$, for all $s > 0$. However, since the diffusion model could not be justified, for short sequences of trials, as an approximation to the randomised allocation procedure, it is more appropriate to apply the asymptotic condition (4.15) to the choice of $\{\lambda(n)\}$.

The empirical decision rule, policy (i), emerged from comparisons based on simulations but, since it satisfies condition (4.15), we also have some idea of its asymptotic behaviour. It is reasonable to conjecture that, for this policy, the maximum e.s.l. $M_1(T)$ is such that

$$\lim_{T \to \infty} M_1(T)/T^{\frac{1}{2}} = c_1, \tag{4.16}$$

where the limit depends only on k. In the case $k = 2$, the empirical evidence supports this (see Table 5 below) and indicates that c_1 is about 0.36. Further simulations, not reported in detail, suggest that c_1 is about 0.22 for the one-armed bandit with $p_2 = \frac{1}{2}$ and about 0.56 when $k = 3$.

For comparison, we can refer to some results about minimax decision rules for the two-armed bandit problem. Vogel (1960b) established that the minimax e.s.l.

$$M(T) = \inf_{\Delta} M_{\Delta}(T)$$

must satisfy

$$c_0 \leqslant M(T)/T^{\frac{1}{2}} \leqslant c_2 \qquad \text{as} \qquad T \to \infty, \tag{4.17}$$

where $c_0 = 0.187$ and $c_2 = 0.376$. His lower bound was improved by Fabius and van Zwet (1970) who showed that (4.17) still holds with $c_0 = 0.265$. They also proved the existence of minimax policies. For each integer $T > 0$, there is a minimax decision rule which is admissible and symmetric: $\pi_1(r_1, n_1, r_2, n_2) = \pi_2(r_2, n_2, r_1, n_1)$. Such rules can be chosen as Bayes procedures with respect to suitable prior distributions, depending on T, and they may involve randomised allocation. The examples given by Fabius and van Zwet show that minimax policies are highly sensitive to the value of T and they become very difficult to determine when $T \geqslant 5$.

The asymptotic bounds in (4.17) were obtained by using diffusion approximations to evaluate special policies. In each case, the approximation can be regarded as a control procedure for the normal version of the two-armed bandit problem and, because of this, we can infer that $c_0 \leqslant \mathscr{M}(1) \leqslant c_2$, where $\mathscr{M}(1)$ is the constant in the exact relation (4.10). The point is that there is no difficulty in justifying diffusion approximations for the special policies concerned. Strictly speaking, a much more detailed analysis would be required to establish whether the ratio $M(T)/T^{\frac{1}{2}}$ has the limit $\mathscr{M}(1)$, as $T \to \infty$.

Leaving aside such mathematical questions, it is worth examining what can be achieved when T is large, by applying policies of various different types. For example, if the procedure is based on a test with samples of equal size, fixed in advance, then the best that can be achieved is a maximum e.s.l. $\sim 0.515\, T^{\frac{1}{2}}$ as $T \to \infty$. The upper bound given by $c_2 = 0.376$ in (4.17) represents the level attained by using a sequential probability ratio test with equal sample sizes and a stopping rule determined by the total T. Thus, policy (i) does slightly better than this without depending on T. A more detailed comparison of these two policies will be given in the next section.

Lower bounds on the ratio $M(T)/T^{\frac{1}{2}}$ have been obtained by restricting or modifying the two-armed bandit problem in such a way that optimal policies can be recognised. In particular, Fabius and van Zwet made use of Feldman's rule, policy (viii), which is optimal provided that $p_1 + p_2 = 1$. It is a straightforward matter to approximate the e.s.l. over a long sequence of trials and find its maximum, subject to the restriction that $p_1+p_2=1$, which leads to the value $c_0=0.265$ in (4.17). In spite of the improvement on Vogel's lower bound, this does not give a realistic idea of what might be achieved by minimax policies. A further improvement can be obtained by considering a modification of the two-armed bandit problem. This permits a more realistic comparison and leads, after a rather complicated argument, to the conclusion that

$$M(T)/T^{\frac{1}{2}} \geqslant 0.283 \qquad \text{as} \quad T \to \infty. \tag{4.18}$$

5. Performance of the empirical rule

5.1. Maximum losses. Table 5 gives a more comprehensive range of values of the e.s.l. for policy (i), obtained by averaging the results of 1000 simulations of the two-armed bandit. The last two columns provide estimates of the maximum loss

$$M_1(T) = \sup L_1(p_1, p_2, T)$$

and the ratio $M_1(T)/T^{\frac{1}{2}}$. In each case, the maximum was determined by examining the results for 15 pairs of probabilities. For example, when $T=20$ and $T=50$, the relevant pair, $p_1=0.7$, $p_2=0.3$, is not included in the table. The asymptotic formula (4.16) appears to be quite effective, provided that $T \geqslant 50$.

Table 5

Two-armed bandit problem: e.s.l. and maximum loss for policy (i)

p_1 p_2	0.6 0.4	0.2 0.1	0.5 0.4	0.6 0.5	0.55 0.5	$M_1(T)$	$\dfrac{M_1(T)}{T^{\frac{1}{2}}}$
$T=20$	1.40	0.83	0.85	0.85	0.46	1.86	0.416
$T=50$	2.53	1.66	1.87	1.84	1.08	2.60	0.368
$T=100$	3.56	2.59	3.15	3.13	2.01	3.56	0.356
$T=200$	4.59	3.73	4.91	4.95	3.54	4.95	0.350
$T=300$	5.12	4.38	6.18	6.16	4.87	6.18	0.357
$T=400$	5.49	4.80	7.15	7.08	6.10	7.15	0.358
$T=500$	5.76	5.09	7.94	7.79	7.21	7.94	0.355
$T=600$	5.98	5.31	8.59	8.36	8.19	8.59	0.351
$T=700$	6.15	5.49	9.17	8.85	9.10	9.17	0.347
$T=800$	6.31	5.64	9.65	9.27	9.95	9.95	0.352
$T=900$	6.45	5.77	10.10	9.66	10.74	10.74	0.358
$T=1000$	6.58	5.87	10.54	10.02	11.48	11.48	0.363

5.2. Error probabilities. So far, we have concentrated on the comparison of allocation procedures by using a criterion based on the number of successes lost. The tables suggest the maximum e.s.l. as the most convenient guide and, from this point of view, randomised allocation procedures such as policy (i) have substantial advantages over policies which depend on prescribing the total number of trials in advance. In particular, we remarked in Section 4.2 that, when T is large, no sequential probability ratio test with sample sizes constrained to be equal can produce a better performance for the two-armed bandit problem. Roughly speaking, the cost of the constraint that both treatments must be used equally often during the period of the test is enough to eliminate the extra advantage of knowing the total number of all the trials before the test is carried out. Since the error probabilities associated with sequential probability ratio tests are minimal, it is also of interest to see whether policy (i) is comparable from that point of view.

The simulations used to produce Table 5 also provided estimates of the error probabilities for policy (i) and some of these are reproduced in the lower part of Table 6. For example, when $p_1 = 0.6$ and $p_2 = 0.5$, 125 out of 1000 simulations led to a position with $r_2/n_2 > r_1/n_1$ after 200 trials, but the observed proportion of similar errors after 1000 trials was 0.029. These estimates are not very accurate when the error probability is small, but they will serve the present purpose. The asymptotic invariance property of the policy could be used to obtain an idea of the general pattern of error probabilities and more accurate estimates.

For comparison, we consider a sequential probability ratio test of the type used by Vogel (1960b) to obtain his upper bound on the minimax e.s.l. for the two-armed bandit. The test consists of using p_1 and p_2 alternately and observing the difference $r_1 - r_2$ after each pair of trials. A decision is reached as soon as $r_1 - r_2 = \pm D$, where D is a fixed positive integer: if this occurs after $2N < T$ trials, all the remaining $T - 2N$ trials are allocated to p_1 or p_2 according to whether the upper or lower boundary is attained.

Policy (xi). The sequential probability ratio test with $D = 5$. This is illustrated in Table 6.

It is a straightforward matter to determine the error probability of this test, if we neglect the effects of truncation. This depends on a parameter γ which is symmetric in p_1 and p_2 and is defined, for $p_1 \geqslant p_2$, by

$$\gamma = p_1(1-p_2)(1-p_1)^{-1}p_2^{-1}.$$

Thus, $\gamma > 1$ except when $p_1 = p_2$. The error probability is

$$\varepsilon_D(p_1, p_2) = (\gamma^D + 1)^{-1} \tag{5.1}$$

and the e.s.l. after $2N$ trials is given by

$$L_D(p_1, p_2) = D(\gamma^D - 1)(\gamma^D + 1)^{-1}. \tag{5.2}$$

Both these formulae are based on the standard "no overshoot" approximation, which is exact here, because $r_1 - r_2 = \pm D$ at the end of the test. They were used to obtain the entries for policy (xi) in two rows of Table 6. The sequential probability ratio test, without truncation, is comparable in performance with policy (i) after $T = 200$ trials. In particular, (5.2) shows that the maximum e.s.l. is $D = 5$. The comparison is slightly misleading in the sense that the expected sample size for the sequential test is unbounded: to see this, consider the case when both p_1 and p_2 are near zero.

Now suppose that the test is truncated after $T = 1000$ trials. This has little effect on the error probabilities given in the table, but it must be taken into account in assessing the e.s.l. The entries for policy (xi), $T = 1000$, were obtained from an inequality established by Vogel (1960a):

$$L_D(p_1, p_2, T) \leqslant T|p_1 - p_2|(\gamma^D+1)^{-1} + D(\gamma^D-1)^2(\gamma^D+1)^{-2}. \tag{5.3}$$

This yields a good approximation, provided that T is large in comparison with the expected sample size of the test. The table indicates a clear preference for policy (i), when $T = 1000$.

Table 6

Policies (i) and (xi): e.s.l. and error probabilities (lower part)

p_1	0.9	0.6	0.2	0.6	0.55	0.51
p_2	0.7	0.4	0.1	0.5	0.5	0.49
(i) $T = 200$	3.51	4.59	3.73	4.95	3.54	1.68
(xi)	4.99	4.83	4.83	3.84	2.32	0.99
(i) $T = 1000$	4.89	6.58	5.87	10.02	11.48	7.35
(xi) $T = 1000$	5.21	8.07	6.37	14.58	14.49	8.22
(i) $T = 200$	0.002	0.015	0.051	0.125	0.274	0.385
(xi)	0.001	0.017	0.017	0.116	0.268	0.401
(i) $T = 1000$	0.001	0.003	0.001	0.029	0.138	0.301

6. General remarks

I referred in the introduction to practical implications, rather than applications, because a great deal still remains to be done. The study of asymptotically optimal policies involves an idealistic assumption that one can defer, for ever, a definite choice of one from amongst the alternative treatments. But in practice, a sequence of randomised allocations will eliminate treatments as they fall into disuse and this raises the question of stopping rules, which has not been seriously considered so far.

The empirical allocation rule, policy (i), was designed to give a reasonably good performance for arbitrary values of the unknown probabilities and for a wide range of values of T. However, it is more realistic to suppose that a scientific experiment must reach a conclusion at some time. In this sense, T can be prescribed in advance. Then the protection given to future patients: those affected by the treatment it is decided to prefer at the end of the experiment, must be measured by the error probabilities. Randomised allocation offers considerable advantages over conventional experiments with equal allocations. This is true even if the experiment has a sequential stopping rule but, for simplicity, let us illustrate the point by considering an experiment with fixed sample sizes.

Policy (xii). Take samples of size S on p_1 and p_2, then choose p_1 or p_2 for all future patients according as $R_1 > R_2$ or $R_1 < R_2$.

We can use a normal approximation to evaluate the error probabilities. For p_1, p_2 not too far from $\frac{1}{2}$, the probability of choosing the inferior p_i after $2S$ trials is about

$$\varepsilon_s(p_1, p_2) = \Phi(-|p_1 - p_2|(2S)^{\frac{1}{2}}),$$

where Φ is the standard normal distribution function.

The effectiveness of the information provided by the experiment can be measured by

$$A_s = \sup_{p_1 \cdot p_2} |p_1 - p_2| \varepsilon_s(p_1, p_2) = \sup_{\xi > 0} \xi\Phi(-\xi)(2S)^{-\frac{1}{2}}$$

where $\xi = |p_1 - p_2|(2S)^{\frac{1}{2}}$. The coefficient here is easily evaluated: $\sup \xi\Phi(-\xi) = 0.170$, with $\xi = 0.75$. Then $A_s = 0.170(2S)^{-\frac{1}{2}}$.

Of course, the cost of the experiment is measured by the e.s.l. which is $L_s(p_1, p_2) = |p_1 - p_2| S$. This has a maximum $M_s = S$.

One of the results of Abdel Hamid's thesis was concerned with finding a reasonable approximation to the error probabilities associated with policy (i). He showed, in effect, that the supremum corresponding to A_s, above, is approximately

$$A_T = 0.21\, T^{-\frac{1}{2}}.$$

Thus, $As \div A_T$ if we choose $T = 3S$. For illustration, we take $S = 50$, $T = 150$. The fact that $A_S \div A_T \div 0.017$ means that the information provided by policy (xii) after $2S = 100$ trials can be matched only by using policy (i) for 150 trials. In other words, the randomised allocation rule lakes longer to produce on equivalent level of precision. On the other hand, there is a very substantial reduction in e.s.l. even if we restrict attention to values of p_1 and p_2 near $\frac{1}{2}$. Tables 7 and 8 below give the error probabilities and e.s.l. for both policies: $S = |p_1 - p_2|$. It is worth remarking that the e.s.l. for the last 50 patients in the sequential allocation procedure does not exceed 1, which is the maximum e.s.l. for any pair of patients in the conventional experiment.

Table 7

Error probabilities

δ	0.025	0.05	0.075	0.10	0.15	0.20
policy (i) $T = 150$	0.389	0.298	0.224	0.166	0.08*	0.04*
policy (ii) $S = 50$	0.401	0.309	0.227	0.159	0.067	0.023

* not reliable estimates

Table 8

Expected successes lost

δ	0.025	0.05	0.075	0.10	0.15	0.20
policy (i) $T = 150$	1.62	2.80	3.67	4.10	4.40	4.28
policy (ii) $S = 50$	1.25	2.50	3.75	5.0	7.5	10.0

Acknowledgments. I am very grateful to Mr. A. Rehim Abdel Hamid for his skill and patience in carrying out all the computations. This research was supported in part by a grant from the Leverhulme Trust.

References

[1] A. R. Abdel Hamid (1981), D. Phil. Thesis, University of Sussex.

[2] J. A. Bather (1980), *Randomised allocation of treatments in sequential trials*, Adv. in Appl. Probab. **12**, 174–182.

[3] – (1981), *Randomised allocation of treatments in sequential experiments*, J. Roy. Statist. Soc. Ser. B **43**, 265–292.

[4] R. E. Bellman (1956), *A problem in the sequential design of experiments*, Sankhyā Ser. A **16**, 221–229.

[5] D. A. Berry (1978), *Modified two-armed bandit strategies for certain clinical trials*, J. Amer. Statist. Assoc. **73**, 339–345.

[6] D. A. Berry and B. Fristedt (1979), *Bernoulli one-armed bandits-arbitrary discount sequences*, Ann. Statist. **7**, 1086–1105.

[7] R. N. Bradt, S. M. Johnson and S. Karlin (1956), *On sequential designs for maximising the sum of n observations*, Ann. Math. Statist. **27**, 1060–1074.

[8] J. Fabius and W. R. van Zwet (1970), *Some remarks on the two-armed bandit*. Ann. Math. Statist. **41**, 1906–1916.

[9] D. Feldman (1962), *Contributions to the two-armed bandit problem*, Ann. Math. Statist. **33**, 847–856.

[10] B. L. Fox (1974), *Finite horizon behaviour of policies for two-armed bandits*, J. Amer. Statist. Assoc. **69**, 963–965.

[11] J. C. Gittins (1979), *Bandit processes and dynamic allocation indices*, J. Roy. Statist. Soc. Ser. B **41**, 148–177.

[12] K. D. Glazebrook (1980), *On randomised dynamic allocation indices*, J. Roy. Statist. Soc. Ser. B **42**, 342–346.

[13] D. G. Hoel, M. Sobel and G. H. Weiss (1972), *A two-stage procedure for choosing the better of two binomial populations*, Biometrica **59**, 317–322.

[14] F. P. Kelly (1981), *Multi-armed bandits with discount factor near one: the Bernoulli case*, Ann. Statist. **9**, 987–1001.

[15] J. D. Poloniecki (1978), *The two-armed bandit and the controlled clinical trial*, The Statistician **27**, 97–102.

[16] H. Robbins (1952), *Some aspects of the sequential design of experiments*, Bull. Amer. Math. Soc. **58**, 527–535.

[17] L. Rodman (1978), *On the many-armed bandit problem*, Ann. Probab. **6**, 491–498.

[18] W. Vogel (1960a), *A sequential design for the two-armed bandit*, Ann. Math. Statist. **31**, 430–443.

[19] – (1960b), *An asymptotic minimax theorem for the two-armed bandit problem*, Ann. Math. Statist. **31**, 444–451.

[20] D. L. Wahrenberger, C. E. Antle and L. A. Klimko (1977), *Bayesian rules for the two-armed problem*, Biometrika **64**, 172–174.

[21] P. Whittle (1980), *Multi-armed bandits and the Gittins index*, J. Roy. Statist. Soc. Ser. B **42**, 143–149.

[22] M. Zelen (1969), *Play the winner rule and the controlled clinical trial*, J. Amer. Statist. Assoc. **64**, 131–146.

Presented to the semester
Sequential Methods in Statistics
September 7–December 11, 1981