In *Computing Science and Statistics* **23** (1991), pp. 421-424.

Bandit Strategies for Ethical Sequential Allocation

Janis P. Hardwick[1] Quentin F. Stout[2]
Statistics Department EECS Department
University of Michigan, Ann Arbor, MI 48109

Abstract

The problem of allocating patients in a two treatment clinical trial with dichotomous response is considered. The trial goal is to determine the better treatment while incurring as few patient losses as possible. Several allocation rules are compared and it is found that *bandit* strategies perform well on both criteria in that they achieve nearly optimal power while keeping expected trial failures nearly minimal. The rules are also evaluated according to their computational complexity.

1 Introduction

Researchers designing clinical trials often encounter difficulties when trying to determine the best way to allocate patients to treatments so that trial goals may be achieved and the costs to all concerned kept at a minimum. Conventional designs, in which subjects are allocated to groups in equal or predetermined proportions, have good decision making properties but lack the flexibility to incorporate other desirable design goals. Adaptive designs, in which allocation strategies may depend on data observed during the trial, have more flexibility. The consideration of adaptive techniques raises the question of what an *optimal* allocation rule is for a problem where statistical merit is not the only measure of the quality of a design. This question is complex and intriguing, and it deserves more attention than it is given here, where only a simple trial set-up is examined. What we can show, however, is that adaptive designs based on optimal strategies for *bandit* problems perform well according to *multiple* criteria, which include but are not restricted to the ability to make a good terminal decision. In particular, these rules are evaluated according to ethical and computational criteria and then compared with standard fixed allocation techniques.

Now, consider a clinical trial in which we wish to compare two treatments and determine, if possible, which has the higher efficacy rate. The patients, who enter the trial sequentially, are to be allocated to one of the two therapies in such a way that trial goals are met as well as possible. While any complete description of a clinical trial design should address all aspects of trial protocol (e.g., eligibility criteria, interpretation of responses, data analysis, etc.), we focus on the effects of changing allocation rules within otherwise fully specified designs.

It is assumed that the sample size for the trial is a fixed number, $n,$ but that the sample sizes for the treatment groups, n_1 for T_1 and n_2 for T_2, may be random. The response variables, X and Y from T_1 and T_2 respectively, are independent Bernoulli random variables such that

$$(1)\quad X_1, X_2, \cdots \sim B(1, P_1); \quad Y_1, Y_2, \cdots \sim B(1, P_2)$$

where $(P_1, P_2) \in \Omega$, for $\Omega = (0,1) \times (0,1)$.

An *allocation rule*, $\gamma,$ is defined to be a sequence $(\gamma_1, \ldots, \gamma_n)$ such that,

$$\gamma_i = \begin{cases} 0, & \text{if } T_1 \text{ is used for patient } i; \\ 1, & \text{if } T_2 \text{ is used at patient } i, \end{cases} \qquad i = 1, ..., n.$$

It is required that the decision, γ_i at stage i, depend only on the information available at that time.

The parameter of interest is the mean difference in responses, $\Delta = P_2 - P_1$, and T_1 is said to be *superior* to T_2 if $\Delta > 0$, and *inferior* if $\Delta < 0$. The *terminal decision rule* depends on the maximum likelihood estimate for Δ which, after n observations, is given by

$$\hat{\Delta}_n = \hat{\Delta}_n(\gamma) = \overline{Y}_{n_2} - \overline{X}_{n_1},$$

where $n_1 = \gamma_1 + \ldots + \gamma_n, \ n_2 = n - n_1,$ and

$$\overline{X}_{n_1} = \frac{1}{n_1} \Sigma_{j=1}^n \gamma_j \, X_j; \quad \overline{Y}_{n_2} = \frac{1}{n_2} \Sigma_{j=1}^n (1 - \gamma_j) \, Y_j.$$

2 Design Characteristics

With the primary goal being to select the better of two competing therapies, the decision rule has been formulated to test the hypothesis

$$(2)\qquad H_0 \, : \, \Delta < 0 \quad vs. \quad H_1 \, : \, \Delta > 0,$$

and it specifies

$$(3)\qquad \begin{array}{lll} \text{Reject } H_0 & \text{if} & \hat{\Delta}_n > 0; \\ \text{No decision} & \text{if} & \hat{\Delta}_n = 0; \\ \text{Fail to reject } H_0 & \text{if} & \hat{\Delta}_n < 0. \end{array}$$

[1] Research supported in part by National Science Foundation under grant DMS-8914328.

[2] Research supported in part by National Science Foundation/DARPA under grant CCR-9004727.

An informative measure of how well a test performs is given by its *power*. For this problem, the power is simply the probability, as a function of $\mathbf{P} \in \Omega$, of correctly identifying the superior treatment. In practice, a rule allowing the *no decision* option should not be used without a null hypothesis of equality and corresponding acceptance region. We would prefer, in fact, a test that not only recognizes similar treatment effects with high probability, but also one that has maximum power at the *smallest clinically* significant difference between the parameters. The testing regions here, however, have been established so that we may study the behavior of the allocation rules over the entire parameter space and obtain lower bounds for the power of (3). In [3], we examine problems incorporating both type I and II errors.

It is not difficult to show that, for any $\mathbf{P} \in \Omega$, the probability of making an incorrect decision based on (3) is minimized by allocating patients to therapies in equal proportions. This may be achieved via alternating assignments or by constrained or blocked randomization. Since an equal allocation rule guarantees that fully half of the patients are assigned to the inferior treatment, designs utilizing them tend to incur more failures than may be necessary for the decision process. Our evaluations of allocation rules are based on three criteria:

1. The probability of making a ‘correct’ decision at the end of the trial,

2. The expected number of failures during the trial,

3. The complexity of the computations required to utilize the design.

Due to space limitations, the manner in which these criteria are assessed is quite simplistic. While each of these items can be viewed from many angles, the results (Section 4) seem to be representative of the behavior of the allocation rules in more general settings as well.

2.1 Bandit Problems

The sampling plans that we propose are based on optimal rules for multi-armed bandit problems. In a bandit problem, the goal is to maximize the sum of weighted outcomes arising from a sequence of experiments from *arms* whose outcomes follow the laws of a specified Bayesian model. A *bandit allocation rule* is thus one that utilizes prior information on unknown parameters together with incoming data to determine optimal selections at each stage of the experiment. The weighting of returns is known as *discounting* and it consists of multiplying the payoff of each observation by the corresponding element of a discount sequence. The properties of any given bandit allocation rule will depend upon the associated discount sequence and prior distribution.

Here we have only a *two-armed bandit* (TAB), but these techniques generalize easily to problems with several arms. Let the outcomes for the two treatment arms be given by (1), and model the prior information on the success rates, p_1, p_2, as independent beta distributions

$$p_1 \sim \mathrm{Be}(a_0, b_0) \quad \text{and} \quad p_2 \sim \mathrm{Be}(c_0, d_0).$$

At any stage $m \le n$, the posteriors for p_1 and p_2 are

$$(4) \quad (p_1 \mid k, i, j) \sim \mathrm{Be}(a, b); \; (p_2 \mid k, i, j) \sim \mathrm{Be}(c, d)$$

where $k = \Sigma_{i=1}^{m} \gamma_i, \;\; i = \Sigma_{i=1}^{k} X_i, \; j = \Sigma_{i=1}^{m-k} Y_i,$ and

$$\begin{aligned} a &= i + a_0, \quad b = k - i + b_0, \\ c &= j + c_0, \quad d = m - k - j + d_0. \end{aligned}$$

The posterior means of p_1 and p_2 at m are simply $\mathbf{E}_m[\,p_1\,] = a/(a+b)$ and $\mathbf{E}_m[\,p_2\,] = c/(c+d)$, where $\mathbf{E}_m$ denotes expectation in the model (4).

Typically, the choice of a prior distribution will depend, somewhat subjectively, on the knowledge of the investigator preceding the trial. We use independent uniform priors here, $a_0 = b_0 = c_0 = d_0 = 1$, because they contain no initial bias and little information, and because the parameters of the beta posteriors concisely summarize the relevant study data to date.

It is worthwhile to note that these allocation rules, which arise within a Bayesian framework, are being evaluated according to frequentist standards. In Section 4, the Bayesian design is seen to have had little effect on the results of the trial from this viewpoint. However, if desired, the design may be set up to impact the trial and its results more heavily, since investigators can strengthen and/or bias the parameters of the beta distributions to reflect a preferred level of information.

2.2 Ethical Criteria

An advantage of using bandit problems to model clinical trials is that elements of the discount sequence can be selected to represent an ethical decision regarding the relative importance of the patient outcomes both during the trial and in the future. At each stage of the sequential decision process, a bandit allocation rule is a function both of the effort to gather information and of the effort to gain immediate reward. Here, we consider two discount sequences, $\{1, \beta_1, \beta_2, \ldots, \beta_n\}$: the *n-horizon uniform* sequence with $\beta_i = 1$, $i = 1, \ldots, n$, and the *geometric* sequence, $\{1, \beta, \beta^2, \beta^3, \ldots\}$, $0 < \beta < 1$.

In the uniform, finite horizon case, the optimal strategy will begin by emphasizing the gathering of information with the result being that the first patients will be treated rather like patients in an equal allocation trial where one assumes throughout that the treatments offer the same prognosis. Toward the end of the study, with a decision imminent, the emphasis on immediate reward is increased until, at the last stage, a completely myopic rule is used. In the geometric case, it is assumed that that there will always be more patients, so the need for information is never completely absent as in the last stage of a finite horizon problem. However, as more and more patients are treated, the need to sacrifice immediate reward to gain information will decrease. Since the sample size in the present problem is fixed at n, we truncate the allocations after n observations. Thus bandit allocation strategies for problems with geometric discounting are not exactly optimal for the truncated case. As we see, however, these rules still provide good model strategies for the problem at hand. See Hardwick [2] for further discussion of the incorporation of geometric bandit strategies in clinical trial designs.

2.3 Computational Criteria

Ethical attributes aside, an experimental design must be straightforward to carry out if it is to be useful. For computational purposes, this means that the rules should use reasonable amounts of time and space (memory), and be sufficiently easy to program. We distinguish here between the computational requirements to set design parameters and those needed to carry out the trial. In general the former will be significantly greater than the latter, but can be carried out on large computers without significant deadline pressure. The latter may require timely response, and may often be performed on personal computers. The latter will be analyzed here in the next section, while the former will be discussed in [3].

3 Allocation Rules

The following three allocation rules were evaluated with respect to the given criteria:

TAA = Truncated Alternating Allocation,
UB = Uniform Bandit, and
TGLB = Truncated Gittins Lower Bound.

The “truncation” in TAA and TGLB refers to a rule whereby, if a state is reached such that the final decision can not be influenced by any further outcomes, then the treatment with the best success rate will be used for all further patients.

3.1 Uniform Bandit

By definition, the n-horizon uniform TAB uses prior and accumulated information to minimize the number of failures during the trial. We can determine the optimal strategy for this bandit problem using dynamic programming. Let $\mathcal{F}_m(i,j,k,l)$ denote the minimal possible expected number of failures remaining in the trial, if m patients have already been treated and there were i successes and j failures on T_1, and k successes and l failures on T_2. (Note that one parameter can be eliminated since $m = i + j + k + l$.) The algorithmic approach is based on the observation that if T_1 were used on the next patient, then the expected number of failures for patients $m+1$ through n would be

$$\begin{aligned}\mathcal{F}_m^{T_1}(i,j,k,l) &= \mathbf{E}_m[p_1]\cdot\mathcal{F}_{m+1}(i+1,j,k,l) + \\ &\quad \mathbf{E}_m[1-p_1]\cdot(1+\mathcal{F}_{m+1}(i,j+1,k,l))\end{aligned}$$

while if T_2 were used then we would get

$$\begin{aligned}\mathcal{F}_m^{T_2}(i,j,k,l) &= \mathbf{E}_m[p_2]\cdot\mathcal{F}_{m+1}(i,j,k+1,l) + \\ &\quad \mathbf{E}_m[1-p_2]\cdot(1+\mathcal{F}_{m+1}(i,j,k,l+1)).\end{aligned}$$

Therefore $\mathcal{F}$ satisfies the recurrence

$$\mathcal{F}_m(i,j,k,l) = \min\{\mathcal{F}_m^{T_1}(i,j,k,l), \mathcal{F}_m^{T_2}(i,j,k,l)\}$$

which can be solved by dynamic programming, starting with patient n and proceeding toward the first patient.

For the m^{th} patient there are $\Theta(m^3)$ possible values of i, j, k, l, so to evaluate all possible combinations of m, i, j, k, and l requires $\Theta(n^4)$ computations. A clever implementation might not evaluate all possible values, but a straightforward implementation, as used here, needs to do so, and empirical evidence indicates that, in fact, $\Theta(n^4)$ values must be computed. The space requirements can be kept at $\Theta(n^3)$ (see [3]).

3.2 Gittins Lower Bound

According to a theorem of Gittins and Jones [1], for bandit problems with geometric discount and independent arms, for each arm there exists an index with the property that, at any given stage, it is optimal to select, at the *next* stage, the arm with the higher index. The index for an arm, the *Gittins Index*, is a function only of the posterior distribution and the discount factor β. While the existence of the Gittins Index removes many computational difficulties associated with other bandit problems, the only known technique for computing the index involves an iterated dynamic programming approach which is computationally intensive when β is close to 1 (see [1]). Unfortunately, these are the β values needed to produce tests of suitable power.

Here we show that very good results can be achieved by utilizing an easily computed approximation. For an arm with posterior distribution Be(a,b), a lower bound for the Gittins Index is given by (see [1, 2])

$$\Lambda_r = \frac{\frac{\Gamma(a+1)}{\Gamma(a+b+1)} - b\sum_{i=1}^r \beta^i \frac{\Gamma(a+i)}{\Gamma(a+b+i+1)}}{\frac{\Gamma(a)}{\Gamma(a+b)} - b\sum_{i=1}^r \beta^i \frac{\Gamma(a+i-1)}{\Gamma(a+b+i)}} .$$

Because Λ_r is a unimodal function of r, the best such lower bound is Λ_{r^*}, where $r^* = \min\{r : \Lambda_r - \Lambda_{r+1} \geq 0\}$. Each Λ_r

Parameters	→	$\Delta = 0.1$			$\Delta = 0.3$		
↓	**Criteria**	TAA	TGLB	UB	TAA	TGLB	UB
n=20	Power	0.671	0.667	0.647	0.913	0.906	0.874
$\beta = .999$	Average Failures	9.947	9.774	9.768	9.505	8.330	8.217
n=50	Power	0.760	0.754	0.708	0.985	0.982	0.947
$\beta = .9999$	Average Failures	24.828	24.148	24.117	23.489	19.673	19.214
n=100	Power	0.841	0.835	0.771	0.999	0.996	0.980
$\beta = .99999$	Average Failures	49.614	47.779	47.642	46.762	38.051	36.984
n=150	Power	0.890	0.885	0.811	1.000	0.998	0.989
$\beta = .999999$	Average Failures	74.393	71.243	70.890	70.031	56.367	54.611

Table 1: Comparisons of Discrimination and Ethical Criteria

can be computed from the previous one in a constant number of steps, so the total time to compute the best lower bound is proportional to $r^* + 1$.

The computational requirements of the TGLB approach are difficult to analyze since they depend upon the value of r^* and upon the successes and failures encountered. In the simplest implementation, the approximate indices for both treatments are computed at each stage and compared to determine the best choice. However, computation can be saved by noting that a "play the winner" property holds, in that if the indices resulted in treatment i being chosen for the previous patient, and the outcome was a success, then they will again choose treatment i. Therefore an index needs to be computed only when a failure has occurred, and then only for the treatment that failed since the posterior distribution of the other treatment is unchanged.

4 Results

The results of our investigations are summarized in Tables 1 and 2. The computational techniques used are explained in [3].

Table 1 shows that TAA, which is optimized to make the correct selection, incurs a large ethical cost, while UB, which is optimized to minimize failures, has a poor discrimination ability. The TGLB rule is a compromise with nearly the power of TAA and nearly the ethical behavior of UB. Note that TGLB has an extra parameter, β, which must be adjusted to optimize its performance. One can show that β must converge to 1 as n increases in order to obtain increasing power. The specific values of β used have been indicated.

Table 2 compares UB and TGLB on computational grounds. TAA was not included since the total computation time is merely proportional to n, i.e., $\Theta(n)$. For UB, the value presented is the number of evaluations of $\mathcal{F}$ which occur, each of which takes a constant amount of time. Thus the computational time for a clinician to utilize UB is proportional to the value presented and may be prohibitive. For TGLB, the value also represents a quantity which is proportional to the total computational time needed to utilize TGLB during a trial. The value presented is the average, over all trials, of the total number of Λ_r values which must be computed for index calculations throughout the trial. While space requirements were not tabulated, recall that UB needs $\Theta(n^3)$ space and TGLB needs only $\Theta(1)$ space.

Parameters		**UB**	**TGLB**	
n	β		$\Delta = 0.1$	$\Delta = 0.3$
20	0.999	8,855	180	174
50	0.9999	292,825	611	597
100	0.99999	4,421,275	1,705	1,687
150	0.999999	21,947,850	4,124	4,109

Table 2: Comparisons of Computational Time

References

[1] Berry, D. and Fristedt, B. (1986), *Bandit Problems: Sequential Allocation of Experiments.* Chapman and Hall, New York.

[2] Hardwick, J. (1986), *The Modified Bandit: an approach to ethical allocation in clinical trials*. Ph.d. thesis, University of California at Los Angeles.

[3] Hardwick, J. and Stout, Q. F. (1991), Computational aspects of sequential allocation for testing with multiple criteria. In progress.