SOME DEVELOPMENTS IN THE THEORY AND PRACTICE OF SEQUENTIAL MEDICAL TRIALS

P. ARMITAGE
London School of Hygiene and Tropical Medicine

1. Introduction

It would have seemed unlikely fifteen years ago that the conduct of controlled medical trials should give rise to serious thought about the nature of statistical inference. These trials in preventive or clinical medicine seemed to provide a straightforward example of Fisherian experimentation, the mere application of which to as difficult a subject as medicine was more noteworthy than any subtleties of statistical design or analysis. Yet they have continued to stimulate theoretical study and controversy. Even the reader of our more recondite journals, deep in Radon-Nikodym territory, is apt to find himself without warning in a discussion starting "Consider now a trial of two drugs, A and B. . . ." Theoretical studies of this sort are often concerned particularly with the design and analysis of sequential experiments, which have been used in medicine from time to time over the last ten years or so. In this paper I shall try to review the development of sequential medical trials, consider the extent and propriety of their present use, and summarize some of the recent theoretical discussions on this topic.

The celebrated cooperative trials in preventive and clinical medicine of the 1940's and 1950's, accounts of many of which are contained in [1], set the standard in a number of respects. By the use of random allocation they provided information about the relative merits of different treatments which was not otherwise available and which, in its finer aspects, could not have been otherwise obtained. This information related both to the therapeutic or prophylactic value of the treatments and also to their potentially adverse effects. It was often possible to investigate interactions between treatment effects and particular characteristics of the subjects, such as age or severity of disease. Second, they provided valuable experience in the administrative problems of large scale trials, which have much in common with those of sample surveys. Third, they provided evidence on the nature of the ethical problems which are peculiar to this branch of experimentation and pervade almost all discussion upon it [2].

Since the case for sequential experimentation in medical trials is closely bound up with these ethical problems it is appropriate to discuss them a little further here. The basic point is that the physician will usually be unwilling to allocate rival treatments at random if he believes that a particular one of these treatments is more effective than the other(s). The choice of treatment may depend on factors other than the presumed effectiveness—such as adverse effects, cost, and ease of administration—all of which may affect different subjects to different extents. In most circumstances it will be possible to launch a controlled trial with random allocation only if the physician is satisfied that no subject is to receive a treatment known to be less appropriate than a rival treatment.

There have been, and will continue to be, many situations in which such agnostic attitudes are permissible. In these situations the ethical course is surely to gather information from controlled trials rather than to allot treatments in such an unsystematic fashion that reliable inference is precluded. Suppose, however, that a trial is started under such circumstances, that subjects enter it in sequence (as is often the case in medical investigations) and that the results gradually accumulate. At some stage the evidence in favor of one particular treatment may have become so strong that the initial state of agnosticism no longer prevails, and there will then be a strong incentive to stop the trial. It is this ethical consideration which first led to the proposal that sequential methods should be used in the design and analysis of clinical trials. The argument seems to be most cogent in situations where the outcome may be grave, where the relative merits of different treatments are indicated primarily by a single variable which can provide a stopping rule and where an individual's response to treatment is obtained rapidly after the start of that treatment.

The statistical analysis of a nonsequential trial has usually followed conventional lines. Differences between means or proportions are tested for significance and subjected to interval estimation; interactions between treatment responses and various characteristics of the subjects may be explored. It seemed natural, therefore, to follow a similar approach to the specification of sequential plans. The usual procedure has been to specify a probability of error of the first kind (the null hypothesis being that treatments are equally effective) and a certain power against given alternative hypotheses. Sequential plans satisfying such criteria were required also to have suitable average sample number characteristics: broadly, the average sample number should be small for large departures from the null hypothesis.

This basic approach is the same as that of Wald's *Sequential Analysis* [3], and some of the first proposals and practical examples used Wald's methods with little or no change. Modifications within this general tradition included the combination of two Wald tests to provide a two sided (three decision) test of a null hypothesis [4], [5]; and the use of specially designed closed schemes providing greater reduction in maximum sample size than Wald's form of truncation [6], [7]. A number of more recent proposals for truncation have not, to my knowledge, been used much in practice [8], [9], [10].

Later in this paper I shall consider some recent criticisms of this general framework and some essentially different proposals which have been put forward. Before I do so, however, it may be useful to make some comments on the sort of use now being made of sequential analysis in medical trials.

2. Current practice

A number of published sequential trials are described briefly in [11] and many other papers have appeared more recently. A rough count, without any systematic search of the literature, has yielded about 50 reported sequential trials. These cover a wide range of medical topics with some concentration on cardiovascular, cerebrovascular and respiratory conditions, pain, and mental illness. The most colorful application is perhaps a study of the efficacy of prayer [12]. In about half the total number of trials a significant difference was established between the rival treatments, although a number of negative results have no doubt escaped publication.

As one looks through this literature certain shortcomings become apparent, and it may be useful to comment on some of these under four headings.

(1) *Appropriateness of any form of sequential approach.* The type of stopping rule which seems appropriate in medical trials has one particular statistical disadvantage. Large differences between treatment effects tend to lead to small terminal sample sizes and hence to imprecise estimates. One may, therefore, terminate a sequential trial with some assurance that a difference exists, but be unable to measure this difference at all precisely, even though this is just the situation in which a precise estimate would be valuable. If ethical considerations are serious the difficulty may be unavoidable. If, on the other hand, ethical considerations are slight it would seem foolish to accept the disadvantage incurred by the choice of an unduly small sample size.

Now, in a few of the published sequential trials the ethical problem seemed to be relatively unimportant. For example, in the palliative treatment of a minor chronic condition it may not be very serious if a patient receives less than the optimal treatment during the trial provided he receives the best treatment as soon as possible afterwards. In such circumstances it would usually be preferable to do a nonsequential trial of large enough size to give estimates of adequate precision.

(2) *Use of plans with inadequate power.* There is a general illusion that the virtue of sequential procedures is to economize in observations and that therefore the more observations one can save the better. Perhaps in consequence a number of sequential trials have been conducted with plans which do not permit more than a small number of observations and hence provide tests with very low power. A negative result is then often reported without reference to the fact that estimated differences are subject to very wide confidence limits. If small trials were normally repeated by different investigators this might not matter, but all too often a negative result in a small trial is enough to inhibit any further experimentation.

I feel, therefore, that these small trials should be done only when practical considerations such as time and expense absolutely preclude trials of greater length.

(3) *Oversimplification in analysis.* There has often been a tendency in medical trials to present an unduly simplified statistical analysis consisting, perhaps, of a single significance test for differences between groups in one particular variable. This temptation is increased in sequential medical trials with the result that in some published reports it is impossible to find basic information such as the numbers of patients receiving each treatment.

The report of a sequential trial should present such information, and the analysis should go well beyond a mere statement of which boundary was hit, by giving standard errors for appropriate contrasts and by presenting comparisons of treatments within relevant subgroups of the individuals so that interactions between treatments and patient characteristics can be assessed.

(4) *General misconceptions.* The simplicity of the stopping rules proposed for sequential trials and the availability of graphical procedures have had a certain appeal to physicians with little previous statistical experience. This has meant that sequential procedures have often been applied with inadequate understanding of their statistical properties and, inevitably, misconceptions have arisen and mistakes have been made.

I have mentioned these shortcomings, not to discredit the use of sequential methods, but to counteract any impression which may be current that these methods make irrelevant the principles of design and analysis rightly regarded as essential in other types of experimentation.

To redress the balance a little I shall refer briefly to a series of trials to compare the effectiveness of different doses of tetanus antitoxin in the treatment of clinical tetanus [13] to [16]. The first trial [13] compared a high dose (200,000 I.U.) with zero dose. This comparison was regarded as ethically permissible because the evidence for the efficacy of tetanus antitoxin was contradictory and weak; the use of antitoxin is costly and not without risk. A sequential design was used because of the extreme importance of not extending the trial unnecessarily: the response was life or death. In the event there was a marked advantage in favor of antitoxin and the trial was stopped after 79 patients had been treated. The sequential diagram is shown in figure 1.

There followed three trials [14], [15], [16] with the same plan to compare different dosages. A control group not receiving antitoxin was now ruled out on ethical grounds, but to preserve continuity one group of patients in each trial received 200,000 I.U. (as in the first trial), the other doses being 20,000, 50,000 and 500,000. In none of these trials has it been possible to show a significant difference, although the sample sizes were much larger than in the first trial (about 150 patients in each group).

In all these trials the death rates have been compared separately within various subgroups of patients, the grouping being defined by variables known to be associated with fatality. Where necessary the overall differences between

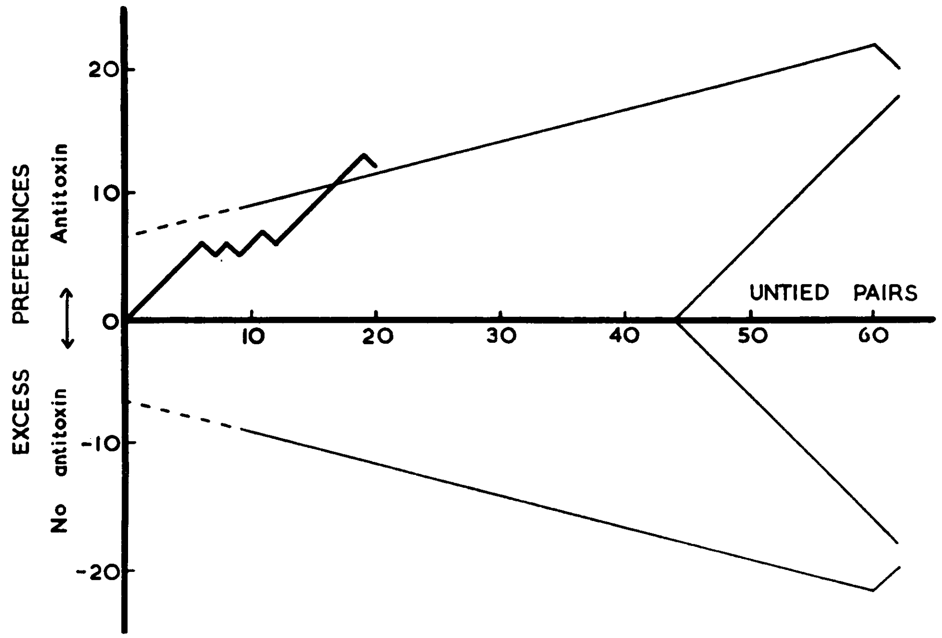

FIGURE 1

Sequential analysis of a trial to compare the treatment of tetanus with and without antitoxin. A pair of patients following the two treatments is called "untied" when only one member of the pair survives, the "preference" then being given to the successful treatment. From [13].

proportions of deaths have been adjusted for differences in these prognostic variables. Figure 2 shows the difference between the death rate at a particular dose and that at 200,000 I.U., with approximate 95 per cent confidence limits.

It appears from figure 2 that the dose-response curve is very flat, except possibly at quite low doses. It is, of course, conceivable that the significant difference between 0 and 200,000 I.U. was due to chance; if not, it seems possible that the dose-response curve changes steeply at some dose less than 20,000 I.U. To investigate this, in what we expect will be the final trial in the series, we are comparing 200,000 with 10,000 I.U., using a sequential plan with much larger maximum sample size and hence greater power. (The plans used in the earlier trials and in the current trial require a maximum of 62 and 191 untied pairs, respectively; the tests have a power of 0.95 when the proportion of untied pairs in one direction is, respectively, 0.75 and 0.65; at the current level of mortality (about 30 per cent) the larger plan would be likely to detect a true difference of 10 to 15 per cent between the proportions of deaths.)

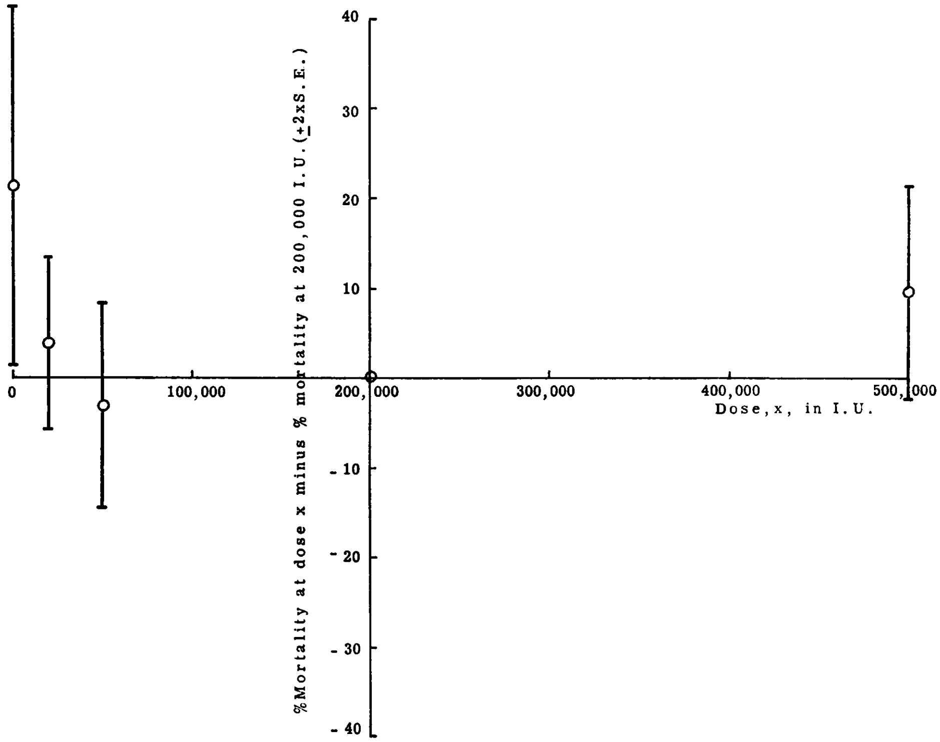

FIGURE 2

Provisional results from a series of therapeutic trials to compare various doses of tetanus antitoxin. In each trial 200,000 I.U. was one of the doses used.

3. Objections and proposals

The general approach in the body of work which I have discussed so far has been to specify certain characteristics of the power function of a significance test and to use a sequential design which achieves these characteristics with sufficiently attractive sample size properties. A certain unease has occasionally been felt among both statisticians and physicians about the propriety of this approach. The use of a rigid stopping rule imposes some inflexibility on the conduct of a trial which may not always be desirable. If the stopping rule is regarded as an indication of a difference between treatments it is a little difficult to see why the rule appropriate at any stage should depend on observations which might have been made but have not in fact been made, particularly when they would have occurred after the stage in question.

The statistical objections, which are summarized by Anscombe [17], are particularly forceful when they are accompanied by constructive proposals developed within the frameworks of likelihood of Bayesian inference and of decision theory. In discussing these proposals it will be convenient to consider first problems of inference.

3.1. *Inference.* Suppose we wish to stop a trial for ethical reasons when we can be fairly certain that the population value of a certain difference is greater than zero (or when we can be equally certain that it is less than zero). How is this certainty to be expressed? It has recently been argued, particularly by Birnbaum [18] and Barnard, Jenkins, and Winsten [19], that if inference about hypotheses is to satisfy certain very reasonable axioms, it must be based on the likelihood function and on no other features of the observed values of the random variables. This conclusion is automatically accepted by advocates of Bayesian inference [20], [21], since the data are used in Bayes's theorem only to provide the likelihood function. An important consequence of the likelihood approach is that the stopping rule is irrelevant for purposes of inference because a change in the stopping rule affects the likelihood only by a multiplying factor and hence does not affect likelihood ratios for the comparison of different hypotheses.

Birnbaum [22] points out, however, that evidential interpretation through the likelihood function is incompatible with another plausible axiom—namely, that if a hypothesis H is true there should be a low probability of outcomes interpreted as strong evidence against H. Birnbaum appears to regard this as an intrinsic anomaly in the concept of statistical evidence. The relevance of this point to the interpretation of sequential experiments may be illustrated by a simple example [23]. Suppose that the difference between the effects of two treatments is measured by a variable x_i distributed as $N(\mu, \sigma^2)$ and evidence is assessed on each cumulative sum $y_n = x_1 + x_2 + \cdots + x_n$. The null hypothesis is that $\mu = 0$. A reasonable stopping rule from the likelihood or Bayesian viewpoint would be to stop after n observations if $|y_n| > k\sigma n^{1/2}$. This would mean that likelihoods of values of μ of zero, or with the opposite sign to that of y_n, were less than some critical fraction of the maximum likelihood; in a Bayesian interpretation with a widely dispersed prior distribution the posterior probability that μ had the same sign as y_n would be greater than some critical level. However, it is well known that if $\mu = 0$ such strong evidence against this value must occur eventually; if μ is a small value $\epsilon < 0$, strong evidence for a positive value of μ will occur with nonnegligible probability. This stopping rule is, of course, equivalent to one based on repeated significance tests of a conventional type, at a constant level of significance, and the recognition of this so called "optional stopping" effect was one of the factors which led to the advocacy of designs in which error probabilities were controlled [5].

Novick and Grizzle [24] present a Bayesian analysis of a clinical trial which was also being analyzed by a restricted sequential plan. The problem is that of the comparison of two binomial variables, and some use is made of the Poisson approximation, since the proportions are small. The authors adopt a "logical probability" approach and advocate the use of natural conjugate Bayes densities to express initial ignorance. Posterior probabilities can be obtained adequately by the appropriate normal approximation (footnote to page 93 of [24]). Except for small sample sizes the analysis is dominated by the likelihood functions and the conclusions are similar to those obtained from a conventional significance test. The authors make the useful point that the marked effect in small samples of a prior distribution centered around the null hypothesis will be to make "rejection" of that hypothesis appreciably less likely, and this may have a substantial effect on the frequency of incorrect rejection in all except very large samples. For example, if the prior distribution for μ is $N(0, \sigma_0^2)$, the boundaries will take the form

$$(3.1) \qquad y_n = \pm k\sigma[n + (\sigma^2/\sigma_0^2)]^{1/2}.$$

The introduction of a finite σ_0^2 is equivalent to a shift of the boundaries parallel to the n-axis and may substantially widen the distance between them at low values of n.

A possible Bayesian reply to the optional stopping dilemma [25], [26], [27] is that difficulties will arise only for values of μ in the neighborhood of zero and these have been given very low prior probability. One should, therefore, either ignore the problem as a red herring, or use a prior distribution more closely concentrated round $\mu = 0$ (with the consequent effect noted above) or with nonzero probability allotted to $\mu = 0$ (a situation explored in some detail recently by Cornfield [27]). The "red herring" view will no doubt satisfy the fully converted, but may be less attractive for those unwilling to put complete faith in their prior distributions. In any case it would be highly desirable to know more about the frequency properties of various stopping rules, a topic to which we return in section 4.

3.2. *Decision.* One of the main purposes of a medical trial is to help to select the best from a group of two or more rival treatments. It has therefore often been suggested that the statistical design and analysis of a trial should be regarded as a problem in decision theory. Given an appropriate formulation in terms of prior probabilities and utilities, it should be possible to obtain a good (and perhaps an optimal) solution.

There have recently been a number of contributions to the theory of sequential decision making. Some authors [28], [29] have been concerned primarily with sequential design; that is, with situations in which at each stage of the investigation a choice may be made between a number of possible experiments. A conceivable trial of this sort would be one in which a patient entering at any stage could receive either of two drugs, A and B. The decision whether to use A or B would depend at least in part on the results obtained earlier, as in the "two armed bandit" problem. One of the practical difficulties of this sort of sequential design is that of reconciling the allocation rule with the requirements of randomization, which is rightly regarded as one of the cornerstones of controlled medical trials.

Another group of papers [30] to [34] is concerned particularly with the specification of stopping rules for investigations with a constant design. (Some authors [35], [36] consider both the design and the stopping rule problem.) The determination of an optimal stopping rule requires the specification of costs of experimentation and terminal decision, the aim being to choose a stopping rule which maximizes the expected utility; analytical solutions appear to be difficult, but certain asymptotic results have been obtained and methods of dynamic programming may be applied to provide explicit solutions.

How should these general concepts be applied to the problem of medical trials? Winsten [37] pointed out that in such a trial financial costs are normally of secondary importance and that a more pertinent measure of cost is the extent to which patients, whether in the trial or not, receive an inappropriate treatment. Colton [38] has supposed that a known number N of patients are to receive one of two treatments. Of these $2n$ will take part in a randomized trial (n on each treatment), and the remaining $N - 2n$ will receive whichever treatment appears to perform better in the trial. The response variable (which can be taken to be a measure of the difference in response between paired individuals receiving the two treatments) is supposed to be normally distributed with known variance and a mean which has a normal prior distribution with zero mean and known variance. The loss due to use of the wrong treatment is proportional to the mean of the response variable. The optimal value of $2n$ is shown to be not greater than $N/3$, this value being taken when the prior distribution is concentrated at zero. Colton also considers an open sequential plan for the trial, with parallel boundaries; similar results are obtained. In a later paper [39] he considers certain two stage designs for the trial. Anscombe [17], with essentially the same model, discusses the problem of determining optimal sequential stopping rules and derives some approximate results.

Studies of this type undoubtedly provide insight into the desirable properties of an ideal system for the selection of medical treatments. Whether, at the present time, they give any detailed help in the planning of specific trials is rather more doubtful [40]. Some of the difficulties are as follows.

(a) Selection is only one of a number of aims of a controlled medical trial, and any selection taking place as a result of trial may be influenced by the additional considerations mentioned in section 1.

(b) The number N should presumably be interpreted as the number of patients who will receive the chosen treatment before it in turn is superseded. This number is clearly difficult to estimate at all precisely, but it seems likely to be so large as to require trials substantially bigger than those undertaken at present. This serves to underline the importance of avoiding very small trials, but it is doubtful whether any organizations at present responsible for the planning of trials can work on the scale required, or indeed have the authority to determine medical treatment in this way.

(c) A conflict between inference and decision will occur if the stopping rule permits the trial to continue well beyond the stage at which the physicians become convinced that a difference exists.

My suggestion, therefore, would be to regard the decision theoretic results as providing general qualitative guidance rather than specific rules of procedure. The boundaries of sequential plans should, I think, accord more closely with likelihood considerations than they do at present. The simplest suggestion would be to use boundaries of the form $y_n = \pm k\sigma n^{1/2}$ instead of the linear boundaries used, for example, in restricted sequential plans. In choosing k, I should have regard to sample space probabilities to avoid optional stopping difficulties; that is, a relatively large value of k would be chosen for sequential plans with a large maximum number of observations. Alternatively, some widening of the boundaries could be achieved by the introduction of a finite σ_0^2, as in (3.1).

Whether the consideration of sample space probabilities clashes logically with the likelihood principle still seems unclear. There are a number of statistical situations in which it is useful to think collectively of a large number of possible inferences, and the interpretation of a particular inference should be judged in the context of the whole situation rather than in isolation. Examples are multiple comparisons (when the comparison being considered has no prior importance) and the assessment of associations suggested by the data. It may be that sequential inference should be regarded in a similar light.

These other situations are different from sequential analysis in that the various inferences are about distinct parameters, whereas in sequential analysis we are concerned with successive inferences about the same parameter. From a Bayesian point of view the problems involved in multiple comparisons or in the interpretation of unforeseen associations may be resolved by appropriate prior distributions. For example, in multiple comparisons the prior distributions of differences may be more concentrated about zero than if the comparisons were solitary, or, following Duncan [41], the observed F ratio may be used to provide prior information. Similarly, unforeseen associations could be given prior distributions more concentrated around zero than those singled out for attention before the data were collected.

These considerations perhaps suggest a Bayesian rationalization of an intuitive desire to allow for optional stopping. If we choose a sequential plan with a large maximum number of observations we are presumably more interested in small differences than if we had chosen a smaller plan. It would seem reasonable in a Bayesian approach to concentrate the prior distribution more and more around zero as the maximum length increased. A stopping rule based on posterior probabilities would then lead to boundaries which were further away from the origin for long than for short plans. In principle a reconciliation would seem possible.

4. Some results on optional stopping

Discussion on the relevance or importance of optional stopping effects have been hampered by a paucity of analytical or numerical information about the size of these effects. No exact solution is known; exact computations are, except in simple cases, prohibitive; approximate results seem difficult to obtain. Mr.

B. C. Rowe and I have therefore started a series of sampling experiments on the London University Atlas computer, some of which I shall describe here.

In the first experiment we had 2000 realizations each consisting of 100 random standardized normal deviates x_i. In each realization the "path" formed by the cumulative sum $y_n = x_1 + x_2 + \cdots + x_n$ was compared for each value of n with the boundaries $\pm k\sqrt{n}$, for values of k corresponding to the two sided normal tail-area probabilities, 2α, of 0.10, 0.05, 0.02 and 0.01. For each k a count was made of the number of paths in which one of the boundaries had been crossed at or before the nth stage. The results are shown in table I.

TABLE I

Cumulative Sums of Random Standardized Normal Deviates

Proportions of paths, out of 2000, in which the absolute value of the cumulative sum y_m, has exceeded $km^{1/2}$ for some $m \leqq n$.
Theoretical probabilities shown in parentheses.

	k:	1.645	1.960	2.326	2.576
n	2α:	0.10	0.05	0.02	0.01
1		.0970 (.100)	.0545 (.050)	.0230 (.020)	.0135 (.010)
2		.1650	.0885 (.083)	.0385	.0235
3		.1980	.1115 (.107)	.0510	.0275
4		.2295	.1260 (.125)	.0610	.0345
5		.2590	.1420 (.141)	.0675	.0390
10		.3425	.1925	.0905	.0525
20		.4300	.2455	.1200	.0695
30		.4765	.2855	.1360	.0815
40		.5045	.3150	.1550	.0905
50		.5260	.3295	.1700	.1000
100		.5975	.3830	.2015	.1195

Some expected numbers are shown in parentheses. Those for $n = 1$ are obvious. For $k = 1.960$ and $n = 1$ to 5, the probabilities were obtained previously [42] by numerical integration. The observed and expected frequencies agree well.

Some comments on table I are as follows.

(1) In many investigations it will be appropriate to examine the data at a number of intermediate points, with about equal numbers of observations in each stage. The number of examinations may then be taken as the value of n in table I. If n is, say, 5 to 10, the effect on the significance level seems to be to multiply by a factor of 2.5 to 5. A rough correction for the optional stopping effect would, for example, be to use repeated tests at the 1 per cent rather than the 5 per cent level.

(2) The probability of crossing the boundary at the nth stage soon becomes low as n increases, although as $n \to \infty$ the cumulative probabilities corresponding to the entries in the columns of table I all tend towards unity. At $n = 100$ the numbers of crossings are disturbingly high. The probability is rather more

than 1/3 that some cumulative sum for $n \leqq 100$ will be significant at the 5 per cent level.

(3) The concentration of the crossings at low values of n suggests that a widening of the boundaries at these low values would have a marked effect on the probabilities of crossing, as conjectured by Novick and Grizzle. As noted above, a modification of this sort will be achieved by concentrating a prior distribution of μ at or around zero. We intend to investigate the frequency properties of boundaries of this type and of some of the boundaries proposed by Cornfield [27].

(4) Although the results of table I relate to normal deviates with known variance, it seems likely that similar results would be obtained for other distributions and perhaps for repeated t tests. Results for cumulative binomial variates (with probability 1/2 for each outcome), given in ([11] table 1.2), are incorrect. The probabilities of crossing the boundaries for $2\alpha < 0.05$ at or before the nth stage ($n \leqq 50$) are as follows:

n	10	20	30	40	50
probability	0.055	0.107	0.134	0.154	0.171.

The discrete nature of the variable permits crossing only at 18 values of $n \leqq 50$. The cumulative probability for $n = 50$ therefore corresponds to 18 repeated tests at $2\alpha < 0.05$ on accumulated data, and although the stages do not involve equal numbers of observations the results accord fairly well with those of table I. (Compare the probability of 0.171 given above with the relative frequencies of 0.2455 for $2\alpha = 0.05$ and 0.1200 for $2\alpha = 0.02$ given in table I for $n = 20$.)

Results similar to those of table I have also been obtained in some provisional experiments with cumulative sums of exponential deviates.

(5) It would be interesting to see results analogous to those of table I for nonzero values of μ. This would provide information about the power functions for sequential plans with boundaries of the square root form, and would permit comparison with other sequential plans. It would also indicate the range of values of μ around zero for which optional stopping is a relevant problem. We hope to carry out investigations of this type.

I have appreciated recent discussions with Mr. J. Cornfield and Dr. M. R. Novick. I am grateful to Mr. B. C. Rowe for programming the investigation referred to in section 4, to Miss J. Bailey and Miss C. Tutton for programming assistance and to the Director of the Institute of Computer Science, University of London, for the use of the Atlas computer.

REFERENCES

[1] A. B. Hill, *Statistical Methods in Clinical and Preventive Medicine*, London, Livingstone, 1962.

[2] ———, "Medical ethics and controlled trials," *Brit. Med. J.*, Vol. 1 (1963), pp. 1043–1049.

[3] A. Wald, *Sequential Analysis*, New York, Wiley, 1947.

[4] M. Sobel and A. Wald, "A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution," *Ann. Math. Statist.*, Vol. 20 (1949), pp. 502–522.

[5] P. Armitage, "Sequential tests in prophylactic and therapeutic trials," *Quart. J. Med.*, Vol. 23 (1954), pp. 255–274.

[6] I. Bross, "Sequential medical plans," *Biometrics*, Vol. 8 (1952), pp. 188–205.

[7] P. Armitage, "Restricted sequential procedures," *Biometrika*, Vol. 44 (1957), pp. 9–26.

[8] T. W. Anderson, "A modification of the sequential probability ratio test to reduce the sample size," *Ann. Math. Statist.*, Vol. 31 (1960), pp. 165–197.

[9] M. A. Schneiderman and P. Armitage, "A family of closed sequential procedures," *Biometrika*, Vol. 49 (1962), pp. 41–56.

[10] ———, "Closed sequential t-tests," *Biometrika*, Vol. 49 (1962), pp. 359–366.

[11] P. Armitage, *Sequential Medical Trials*, Oxford, Blackwell, 1960.

[12] C. R. B. Joyce and R. M. C. Welldon, "The objective efficacy of prayer: a double-blind clinical trial," *J. Chron. Dis.*, Vol. 18 (1965), pp. 367–377.

[13] A. Brown, S. D. Mohamed, R. D. Montgomery, P. Armitage, and D. R. Laurence, "Value of a large dose of antitoxin in clinical tetanus," *Lancet*, Vol. 2 (1960), pp. 227–230.

[14] B. J. Vakil, T. H. Tulpule, P. Armitage, and D. R. Laurence, "A comparison of the value of 200,000 I.U. tetanus antitoxin with 50,000 I.U. in the treatment of tetanus," *Clin. Pharmac. Ther.*, Vol. 4 (1963), pp. 182–187.

[15] ———, "A comparison of the value of 200,000 I.U. tetanus antitoxin (horse) with 20,000 I.U. in the treatment of tetanus," *Clin. Pharmac. Ther.*, Vol. 5 (1964), pp. 695–698.

[16] A. O. Lucas, A. J. Willis, S. D. Mohamed, R. D. Montgomery, H. Steiner, P. Armitage, and D. R. Laurence, "A comparison of the value of 500,000 I.U. tetanus antitoxin (horse) with 200,000 I.U. in the treatment of tetanus," *Clin. Pharmac. Ther.*, Vol. 6 (1965), pp. 592–597.

[17] F. J. Anscombe, "Sequential medical trials," *J. Amer. Statist. Assoc.*, Vol. 58 (1963), pp. 365–383.

[18] A. Birnbaum, "On the foundations of statistical inference," *J. Amer. Statist. Assoc.*, Vol. 57 (1962), pp. 269–306.

[19] G. A. Barnard, G. M. Jenkins, and C. B. Winsten, "Likelihood inference and time series," *J. Roy. Statist. Soc. Ser. A*, Vol. 125 (1962), pp. 321–327.

[20] L. J. Savage, *The Foundations of Statistics*, New York, Wiley, 1954.

[21] D. V. Lindley, *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2: Inference*, Cambridge, Cambridge University Press, 1965.

[22] A. Birnbaum, "The anomalous concept of statistical evidence: axioms, interpretations, and elementary exposition," *Technical Report IMM*, Courant Institute of Mathematical Science, New York University, 1964.

[23] P. Armitage, discussion of [26], *J. Roy. Statist. Soc. Ser. B*, Vol. 23 (1961), pp. 30–31.

[24] M. R. Novick and J. E. Grizzle, "A Bayesian approach to the analysis of data from clinical trials," *J. Amer. Statist. Assoc.*, Vol. 60 (1965), pp. 81–96.

[25] L. J. Savage, M. S. Bartlett, G. A. Barnard, D. R. Cox, E. S. Pearson, and C. A. B. Smith, *The Foundations of Statistical Inference*, London, Methuen, 1962.

[26] C. A. B. Smith, "Consistency in statistical inference and decision," *J. Roy. Statist. Soc. Ser. B*, Vol. 23 (1961), pp. 1–25; pp. 34–37.

[27] J. Cornfield, "A Bayesian test of some classical hypotheses—with applications to sequential clinical trials," *J. Amer. Statist. Assoc.*, Vol. 61 (1966), pp. 577–594.

[28] H. Chernoff, "Sequential design of experiments," *Ann. Math. Statist.*, Vol. 30 (1959), pp. 755–770.

[29] A. E. Albert, "The sequential design of experiments for infinitely many states of nature," *Ann. Math. Statist.*, Vol. 32 (1961), pp. 774–799.

[30] D. V. Lindley, "Dynamic programming and decision theory," *Appl. Statist.*, Vol. 10 (1961), pp. 39–51.

[31] H. Chernoff, "Sequential tests for the mean of a normal distribution," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1961, Vol. 1, pp. 79–91.

[32] G. B. Wetherill, "Bayesian sequential analysis," *Biometrika*, Vol. 48 (1961), pp. 281–292.

[33] G. Schwarz, "Asymptotic shapes of Bayes sequential testing regions," *Ann. Math. Statist.*, Vol. 33 (1962), pp. 224–236.

[34] D. V. Lindley and B. N. Barnett, "Sequential sampling: two decision problems with linear losses for binomial and normal random variables," *Biometrika*, Vol. 52 (1965), pp. 507–532.

[35] J. Kiefer and J. Sacks, "Optimal sequential inference and design," *Ann. Math. Statist.*, Vol. 34 (1963), pp. 705–750.

[36] P. Whittle, "Some general results in sequential design," *J. Roy. Statist. Soc. Ser. B*, Vol. 27 (1965), pp. 371–394.

[37] C. B. Winsten, discussion of P. Armitage, "The comparison of survival curves," *J. Roy. Statist. Soc. Ser. A*, Vol. 122 (1959), pp. 297–298.

[38] T. Colton, "A model for selecting one of two medical treatments," *Bull. Inst. Int. Statist.*, Vol. 39, Part 3 (1962), pp. 185–200. (Also in *J. Amer. Statist. Assoc.*, Vol. 58 (1963), pp. 388–400.)

[39] ———, "A two-stage model for selecting one of two treatments," *Biometrics*, Vol. 21 (1965), pp. 169–180.

[40] P. Armitage, "Sequential medical trials: some comments on F. J. Anscombe's paper," *J. Amer. Statist. Assoc.*, Vol. 58 (1963), pp. 384–387.

[41] D. B. Duncan, "A Bayesian approach to multiple comparison," *Technometrics*, Vol. 7 (1965), pp. 171-222.

[42] P. Armitage, "Sequential methods in clinical trials," *Amer. J. Publ. Health*, Vol. 48 (1958), pp. 1395–1402.