SEQUENTIAL MEDICAL TRIALS*

F. J. ANSCOMBE
Princeton University

In an extended review of *Sequential Medical Trials* by P. Armitage, the statistical principles which should govern the analysis of experimental observations, and the planning of experiments, are discussed. It is suggested that the operating-characteristic concepts of the Neyman-Pearson theory of tests are inappropriate to the analysis and interpretation of experimental data; the likelihood principle should be followed instead. The planning of medical trials under an ethical injunction against unnecessary continuance of inferior treatments is studied in some detail. The propriety of such trials is considered.

THIS article has arisen out of an invitation to review the book, *Sequential Medical Trials*, by Peter Armitage [4]. The reviewer soon found his attention focused on certain questions of general statistical principle, that are much "in the air" at the present time. The book is written from an "orthodox" statistical point of view, based on the Neyman-Pearson theory of tests. It seemed to the reviewer that this point of view was inappropriate, and that a rather careful attempt to say why would be worth undertaking. Professor Armitage's book seemed a particularly suitable starting point for such a discussion, as it is a well-written account of sequential experimentation in a real field; though at a pinch a great number of other items in the recent literature could have served instead. Thus this article might be described as statistical polemic thinly disguised as a book review.

Professor Armitage's book is concerned with clinical experiments to compare the effectiveness of therapeutic or prophylactic treatments. There is some general discussion of the planning and conduct of such trials—their objectives, conditions and difficulties. The main purpose of the book is to make available some sequential procedures, mostly developed by the author himself, which he suggests are particularly suitable for certain kinds of medical trials. He writes: "Sequential analysis has an immediate appeal in clinical research. Patients normally enter an investigation serially and a continuous scrutiny of the results is usually quite feasible. Rather more important is the ethical consideration, which requires that any unnecessary use of inferior treatments should be avoided. The investigator will, therefore, frequently wish to bring a trial to an early close if an important difference can, at that early stage, be established." The book is short and clear, and is likely to be widely read by the medical research workers for whom it is intended. It should also prove interesting to nonmedical statisticians, as an original and thoughtful account of a little-known field.

Before any adverse criticisms are developed, it is proper to make two observations in defense of the book. First, its net effect on medical research will almost certainly be good, partly because the concise but comprehensive general discussion of medical trials will provoke and clarify thought, partly because

* Prepared in connection with research supported by the Office of Naval Research.

any consideration of sequential designs encourages flexibility. A planned experiment is not necessarily, and sometimes ought not to be, of fixed size. Even if the sequential designs given are not as good as they might have been, an intelligent user will do fairly well with most of them. In any case, efficiency in the statistical planning and analysis of an experiment, though important, is not all-important. A far more serious matter is the relative merit of planned experiments as compared with uncontrolled case studies—on which there is still much controversy.

The other observation in defense is that the book is addressed to readers who are not professional statisticians, and it represents what has been the dominant school of statistical thought during the past quarter century. If the author had had any radical doubt concerning the statistical concepts, he would presumably not have written such a book at all. Today many statisticians, sympathetic with the tradition, will feel that what Professor Armitage has done is perfectly reasonable and satisfactory. It seems to this reviewer, however, that although Professor Armitage has obviously thought a lot about his subject, still longer thought and relentless pursuit of ideas already present in the book would have led him to reject the tradition in favor of something different. As a matter of fact, just after the book was completed Professor Armitage suggested to T. Colton a (to me) more satisfactory type of study [9], which will be discussed below. It is rather unfortunate that the results of this study could not be incorporated in the book.

Finally, apropos of defenses, a brief one for this paper is perhaps excusable, namely that this appears, not in a medical journal, but in a statistical journal, addressed to readers interested in the current debate on the foundations of statistics. Like the practitioners of any other discipline, statisticians must be concerned to have their house in order, whether or not a united front is always presented to the outside world.

COMPARISON OF A PAIR OF TREATMENTS

Armitage's book is concerned mainly with the following type of problem. Two alternative treatments, say A and B, are to be compared. Observations are made in pairs, one member of each pair for each treatment. If the ailment being treated is a recurrent one, such as headache or seasickness or the common cold, the two observations of each pair may be made on the same patient (at different times and in random order). For more serious illnesses, the pairs of observations will usually refer to different patients. The latter situation will be presumed here, for definiteness. The pairing of patients may be merely fortuitous, or there may be some stratification, so that (for example) patients in the same pair are alike in age and sex.

Armitage suggests that at the end of the trial the investigator will normally wish to draw one of the three conclusions: that A is preferable to B, that B is preferable to A, or a noncommital statement about inadequate evidence, likely to be taken to mean that A and B are about equally effective.

Armitage regards this type of trial as coming within the scope of the Neyman-Pearson theory of tests, and he postulates control of the probabilities of "errors of the first and second kinds" as a primary requirement. The error probability

of the first kind is the chance, given that A and B are in fact equally effective, that a conclusion that one treatment is preferable to the other will be reached (either A preferred to B or B preferred to A). The error probability of the second kind is the chance, given that one treatment is preferable to the other in some stated degree, that the conclusion reached will not be the correct statement of preference. In tables the author considers error probabilities of the first kind equal to 0.05 and 0.01, and error probabilities of the second kind equal to 0.05.

Two types of sequential plan are suggested, open plans and closed plans. The open plans amount (very nearly) to the simultaneous application of two linear sequential tests of the type developed by A. Wald (and W. Bartky and G. A. Barnard). The closed plans are a sort of truncated version of the open plans. In detail, the plans depend on the type of readings that are obtained from the patients. Several possibilities are considered: ranked pairs, all-or-nothing responses, measured responses following a normal distribution of error, with known or unknown variance. In each case, the parametric specification of the chance distribution of the observations, or (as is often said) the "model," is supposed given; its reasonableness or goodness of fit is not in dispute.

If it is agreed that control of the above probabilities of error of the first and second kinds at assigned values is a proper requirement, the author's designs will surely be judged eminently satisfactory. When there is a large difference in effectiveness between the treatments, the trial will soon be terminated, and then the ethical objection to persisting in the application of a clearly inferior treatment is (more or less) allayed. The expected number of observations, and so the costs of running the experiment, are kept somewhere near the minimum possible with the assigned error rates.

ROUTINE DECISION PROCEDURES

Well, what about controlling the errors of the first and second kinds, and what about the associated three-decision structure for the problem? I shall argue, first, that whereas the concept of error probabilities of the first and second kinds, of power curve or operating characteristic, has some relevance to the design of impartial routine decision procedures, such as industrial inspection plans, it has no direct relevance to experimentation; secondly, that the three-decision formulation of the inference problem is inappropriate, and that a two-decision formulation would provide a better basis for determining the sequential stopping rule.

Very briefly, the objection to the concept of error probabilities of the two kinds is that these are expectations of something (of making a wrong decision), taken conditionally on the parameter values but *unconditionally over the whole sample space.* If such expectations are used as a means towards drawing inferences from some observations, the consequent inferences, beliefs and actions will perhaps be much affected by what was *not* observed, by all the rest of the sample space besides the one observed point in it. Absurdity can (and in the present case certainly will sometimes) result.

Consider the problem of designing a decision procedure for routine use, such as a sampling inspection plan. Whatever the precise objectives of the plan may

be, they can always (I believe) be adequately expressed in terms of performance on the average or in the long run, that is, in terms of expectations over the sample space. The purpose will usually be to maximize as nearly as possible the average net profitability or utility of the plan. Two plans are equally good if, after due allowance has been made for the costs of operating the plans, their average profitabilities are indistinguishable. One plan is to be judged better than another if the average net profitability of the first is perceptibly higher than that of the second. How the plans function in special cases is irrelevant.

Usually it is desirable that the inspection procedure should require simple observations, not too extensive, and the rules for reaching a decision should be clear and easy to apply. Importance is generally attached to impartial administration, especially when the plan is agreed to by two parties. The inspector should follow the stated procedure impartially, and pay no attention to rumors or private hunches in reaching his decisions.

There are two reasons why a precise optimization of the inspection plan is not attainable. One is that how well a plan functions, how nearly it attains the objectives set, depends on the quality or properties of the material to which it is applied. In choosing the plan it is necessary to guess at the relative frequency of occurrence of material of various qualities, or else to avoid such guessing by some arbitrary device such as Wald's minimax rule. In the end, in retrospect, the plan may be seen to have been unfortunate, because the frequency distribution of qualities experienced was different from that catered for.

The other impediment to precise optimization is the following. In order to assess the net profitability of a plan, or to compare the profitabilities of two alternative plans, an economic assessment must be made of the operating costs and of the consequences of the possible decisions; and these economic assessments are likely to be rather rough. A step in the calculations is to determine the "operating characteristic," showing the chances of the various possible outcomes when the plan is applied to material of given quality. If it should happen that there is serious dispute over the economic valuation of consequences of actions, or if such valuation requires too great a mental effort, a fully economic analysis may be by-passed by arbitrarily fixing a point (or two) on the operating characteristic. In that case the operating characteristic is of direct and primary interest as a measure of the performance of the plan.

The concepts of Wald's theory of decision functions constitute (as far as they go) a perfect formalization of the problem of choosing impartial routine decision procedures, in accordance with a fully economic analysis. The Neyman-Pearson theory of tests is the corresponding formalization of the problem in terms of operating characteristics.

INFERENCE FROM AN EXPERIMENT

When we turn to consider scientific experiments, we recognize two rather different main phases of statistical activity, (i) planning an experiment, (ii) interpreting and reporting the observations of an experiment. Although these activities may involve the use of established procedures and the making of decisions, the phrase "routine decision procedure" does not seem to fit well

anywhere. Let us consider particularly the second phase, the interpretation of scientific observations.

There is indeed a repetitiveness in science, but it is more subtle than that of industrial inspection. If an experiment shows some interesting result, that result will no doubt be confirmed by later experiments, but quite possibly not by any simple repetition of the first experiment. Typically, the confirmation will be indirect; consequences of the first result will be investigated, and the result will be set in a larger pattern. Often, much individual thought goes into the planning of an experiment, and much care and trouble in its execution. A detailed factual report of the observations is highly desirable, but does not constitute the whole interpretation of the experiment. The formation of opinions, decisions concerning further experimentation and other required actions, are not dictated in a simple prearranged way by the formal analysis of the experiment, but call for judgment and imagination. If a particular statistical procedure for reaching decisions or forming beliefs is under consideration, the performance it would have on the average if it were applied to an unlimited sequence of "similar" experiments is not obviously relevant to its suitability for use on this occasion. The reason is that the "similar" experiments may be dissimilar in informativeness, and we are specially interested in this one actual experiment.

Suppose that a class of admissible hypotheses is given, so that a chance distribution for the observations is fully specified in terms of a finite number of parameters. Then anyone who adopts a probabilistic approach to inference, by way of Bayes' theorem, will necessarily consider that the likelihood function constitutes a complete summary of the observations. Inferences should be based on the one observed point in sample space, without any averaging over unobserved points. (This is true even for the "weak" probabilistic approach of Smith [18].) The same principle has been reached from other points of view. In particular, Fisher has emphasized the full informativeness of the likelihood function [11], even though his fiducial argument is not solely dependent on the likelihood function [2]. Elsewhere, as in [12], he has stressed the need for recognizing appropriate reference sets; and that is a closely related matter. The likelihood principle has been discussed and defended by Lindley [14], Barnard and others [5, 6], and Birnbaum [7]. The arguments in favor of the principle are indeed weighty. Although much statistical practice violates the principle, no serious attempt at direct refutation seems ever to have been made. The principle will accordingly be accepted here.

The difference between (a) devising a routine decision procedure that will work sufficiently well in the long run and (b) interpreting a particular body of data as well as possible, can be illustrated by reference to life insurance. A company that issues thousands of policies a day cannot give much thought to each one. Premiums must be based on a small amount of information in each case, according to simple rules. For example, some policies are issued for members of the teaching profession, based on information concerning the insured person's age and the fact that he is a teacher. In principle, a juster assessment of such a person's life expectancy could be arrived at by considering

much other information about him—a thorough medical examination, and information about where and what and whom he teaches, his personality and interests and activities. But all this additional information, even if it cost nothing to obtain, would be more expensive to process accurately, and though it might occasionally lead to an entirely different assessment, might well not lead to a noticeably better assessment *on the average*. The situation would be different if one single case were of peculiar importance; say, if someone wished to insure his life for a very large sum indeed, or if his survival were a contingency affecting many policies, in various ways, by reason of his high office. An insurance company might well give such a case far more careful consideration than usual.

Unfortunately the difference between the two approaches to inference has been blurred by the fact that in some very common cases both methods seem to lead to substantially the same result. Suppose it is given that a certain reading (say, x) is an observation of a chance variable having a normal distribution with unit variance and unknown mean θ; and suppose we are interested in whether θ is positive. Someone habituated to the operating-characteristic approach to statistics will no doubt make a one-sided test of the null hypothesis that $\theta=0$, calculating as "significance level" the chance that such an observation will exceed x if $\theta=0$,

$$\int_x^\infty \phi(y)\,dy, \tag{1}$$

where $\phi(y)$ is the standard normal density function. On the other hand, someone interested in the individual case will remark that the likelihood function for θ is (proportional to)

$$\phi(\theta - x), \tag{2}$$

and if his prior opinion about θ is diffuse near $\theta=x$ he will reckon the probability that $\theta<0$ to be roughly

$$\int_{-\infty}^0 \phi(\theta - x)\,d\theta. \tag{3}$$

By a mathematical coincidence, expressions (1) and (3) are equal, and it is likely that both persons will form the same opinion as to whether it is reasonable to conclude that θ is positive (yes if (1) and (3) are small, no if large). A skeptic will comment that the difference in the approaches is merely "philosophical" and of no practical importance. He will be wrong. If in repeated trials under the same conditions the informativeness of the observations varies (the spread of the likelihood function varies), a significance level calculated as at (1) may possibly be quite different from a probability calculated as at (3), and there is a most practical necessity to decide which is wanted. In particular, this situation arises when a sequential rule is followed such that the chance distribution of the number of observations, for given parameter values, has a large spread.[1] The sequential plans given in Armitage's book have this prop-

[1] It may be objected that the virtue of Wald's likelihood ratio sequential procedure for choosing between two simple hypotheses is just that all possible sets of observations have (very nearly) the same informativeness. That is true provided the parameter space contains only the two points referred to—a rare situation. More usually, it would be unreasonable to assert that one or other of the two simple hypotheses must be true; they are merely representatives, and the statement in the text is correct.

erty. Thus if it is agreed that averages over the sample space are not required for the just interpretation of experimental observations, we are bound to judge Armitage's five-percents to be strictly irrelevant and possibly misleading.

FORMULATION OF THE ANALYSIS PROBLEM

Let us consider now explicity the other contentious aspect of Armitage's treatment, his formulation of the statistical analysis of the observations as a choice between three possible conclusions (as already indicated above).

We note first that sequential rules such as Armitage's are simultaneously two things, stopping rules and decision rules; that is, they indicate how long the observation process should continue, and they also indicate what verdict should then be given. When the experiment has been completed, the number of observations taken is an unalterable fact. The verdict, on the other hand, is no better than an opinion of the experimenter, and if anyone considers it to be a mistaken opinion he can form a different opinion of his own. The experimenter had his decision rule in mind when he chose his stopping rule; they may have seemed inseparable, but they are not.

Ideally, what should we like to get from an experiment to compare two treatments? Presumably an exact statement of the relative effectiveness of the treatments. The treatments may very well differ in side effects and in cost. In any particular case, if the physician knew exactly the relative effectiveness of the treatments, he could weigh this against the other factors, in relation to his patient, and make an informed choice. In reality, the relative effectiveness of the treatments will not be determined exactly by the experiment—especially will this be so if the difference seems to be large and the ethical objection to continuing to test a poorer treatment is present. It would seem, therefore, that the primary aim of the statistical analysis of the experiment should be to present as clearly and accurately as possible the evidence concerning relative effectiveness—that is, given the class of admissible hypotheses, to quote the likelihood function or some abbreviation of it.

To see more clearly some of the issues involved, let us now make every possible simplifying assumption. Let us suppose that the two treatments do not differ in side effects nor in cost, that they can be assumed to have the same relative effectiveness for all patients, and that they are the only two treatments available that need to be considered. Let us suppose further that the differences in response of pairs of patients (the response of the patient treated by A minus the response of the patient treated by B, for each pair) are realizations of independent chance variables normally distributed with known variance—let us say with unit variance—and with unknown mean θ. Suppose that the higher a response is the better, so that A will be preferred to B if θ is positive, B to A if θ is negative. When the trial stops, let n be the number of pairs of patients that have been observed, and let y denote the sum of the n response differences. Then y and n are jointly sufficient for θ, the likelihood function being (proportional to)

$$\phi\left(\frac{y - n\theta}{\sqrt{n}}\right). \tag{4}$$

Let us suppose that the initial opinion about θ held by the experimenter and

other persons concerned is diffuse relative to the likelihood function. The ensuing calculations are made easier if a normal prior probability distribution for θ is postulated, and are easiest when the distribution is uniform (a limiting case of the normal). Let us therefore postulate the uniform. The final probability distribution for θ, given y and n, is then

$$\sqrt{n}\,\phi(\sqrt{n}\theta - y/\sqrt{n})\,d\theta. \tag{5}$$

If $\theta>0$, the verdict "A better than B" is preferred; if $\theta<0$, the verdict "B better than A" is preferred. If the wrong verdict is given, and if somehow this fact could be known, the wrong verdict would be regretted to a degree depending on the size of θ. If the regret would depend on $|\theta|$ but not on the sign of θ, then it is easy to see that the decision rule minimizing the expected regret would be to give the verdict "A better than B" if $y>0$, "B better than A" if $y<0$.

Thus the choice facing a physician in treating a patient is easy. He will choose whichever treatment appears the more effective, whether the difference is large or small, whether it is "significant" or not. The only thing he needs to know is whether it seems more plausible that A is better than B, or that B is better than A. If he thinks the odds are a million to one, or merely 6:5, that A is better than B, he will choose A. He has to choose something.

Now it is only too likely in practice that several of the above simplifying assumptions will be false, and the best choice of treatment for a physician to make, in the light of the experiment, for a given patient, will perhaps not depend simply on y. In any case, the choice is the responsibility of the physician—he alone *decides* which treatment to use. It is possible that the experimenter, when he planned and carried out the experiment, did not have all the information concerning side effects and other relevant matters available to physicians after the experiment has been published, when they make their decisions in treating patients. It is therefore unwise for the experimenter to view himself seriously as a decision maker, although admittedly the report he gives on his experiment may possibly carry great weight. The experimenter at most *recommends*.

THE DESIGN PROBLEM

We see here a sort of paradox in scientific experimentation. An experiment is usually planned in connection with a particular line of investigation, to throw light on some specific questions, against a particular background. In planning an experiment, the experimenter will attempt to judge the value for his investigation of the results that can be expected, weighing this against the probable cost of execution. If this planning process is to be formalized, one will naturally turn to statistical decision theory, and think in terms of maximizing a well-defined expected utility. (See Raiffa and Schlaifer [17] for an excellent review of the concepts.) But when the experiment is completed and the report is published, it ceases to belong simply to the original decision problem. It may well be viewed by others in relation to other decision problems. Therefore the reporting should be (in part, at least) open and undirected. The original problem need not

be forgotten, but it is not necessarily the same as the problem the reader will be interested in. So an experimenter will quite properly appear to narrow his purposes at the planning stage and broaden them again at the reporting stage. The experimenter pays the piper and calls the tune he likes best; but the music is broadcast so that others may listen. With this warning in mind, let us consider the planning stage for medical trials.

There are two quite different sorts of cost associated with carrying out a medical trial. One is the usual sort of cost in any experiment—the money spent, and the use of time and energy of the personnel. The other is peculiar to medical experiments—the ethical objection to treating patients suffering from a serious illness by what appears to be an inferior treatment. No doubt in some trials only the second kind of cost is of much importance, so let us begin by ignoring the first altogether.

Under our previous simplifying assumptions, let us suggest a measure of the ethical cost. One patient in each of the n pairs tested in the experiment was given the less effective treatment. It seems reasonable to assess the regret at n multiplied by some function of $|\theta|$. Sometimes the function could reasonably be just $|\theta|$ itself, so let us assume that now. This is the regret conditional on θ, which of course is unknown. The expected regret at the end of the trial that n patients have been less effectively treated is thus

$$n\mathcal{E}(|\theta|), \tag{6}$$

where the expectation is over the probability distribution (5) for θ, which is normal with mean y/n and variance $1/n$.

Against this must be weighed the regret that, after the experiment has been published, future patients will perhaps be less effectively treated, because perhaps the experiment gave a misleading result, that is, because y had the opposite sign to θ. Let us make a guess at the number, k, of future patients whose treatment will be decided in accordance with the outcome of the experiment, and so will be given A or B according to whether y was positive or negative.Then the regret caused by this possibility of inferior treatment is assessed at

$$k\mathcal{E}[\max(0, -\theta \operatorname{sgn} y)]. \tag{7}$$

Adding (6) and (7) and calculating the expectations, we readily obtain the following expression for the total regret $R(n, y)$ at the end of the trial, due to the wrong treatment of patients in the trial and to the possible wrong treatment of future patients:

$$R(n, y) = |y| + \frac{k+2n}{\sqrt{n}}\left\{\phi\left(\frac{y}{\sqrt{n}}\right) - \frac{|y|}{\sqrt{n}}\,\Phi\left(-\frac{|y|}{\sqrt{n}}\right)\right\}, \tag{8}$$

where $\Phi(x)$ denotes the standard normal integral from $-\infty$ to x.

The design problem is (as nearly as possible) to minimize $R(n, y)$; that is, to find a sequential stopping rule such that the trial is stopped as soon as further continuance is expected to lead to an increase in R. If an explicit expression for $R(n, y)$ is adopted, such as (8) with a numerical value for k, the determination

of the optimum stopping rule is a matter of computation—possible in principle, if difficult in practice, requiring a formidable "backwards induction."

Assigning a numerical value for k is obviously something of a puzzle. It is a genuine puzzle inherent in the ethics of the situation. No attempt should be made to suppress it by sleight of mathematician's hand. Less effective treatment of some patients now is morally justifiable if there is reasonable hope that through the knowledge gained many more patients will be better treated in future. But how many more? Some estimate can be formed by considering how many cases of the sort arise per year, and how long the results of previous experiments in the same field have remained decisive, tempering these considerations by observing how research interest in the field is developing. There will be a temptation to choose k much too high, from an exaggerated idea of the attention that will be paid to the experiment, and lack of knowledge about research being done elsewhere, destined to supersede the work at hand. In any case, regular clinical experience provides an ultimate check on some of the results of experiments. Hence it would be wise to apply a deflating factor to the number first thought of for k, say one-tenth or less. Another good reason for modesty in assessing k is that there may well be an interaction between the relative effectiveness of the treatments and the place of administration. I understand that results obtained at one hospital are sometimes not supported at first by similar trials at other hospitals, because some detail in the first trial, thought to be unimportant and not mentioned, or perhaps even not known at all, turns out to be critical, and only when it is recognized is confirmation obtained. Hence if the results of an experiment seem to be important, independent confirmatory trials elsewhere are highly desirable. This interaction is distinct from the other sort of interaction that will commonly occur, between the effectiveness of the treatments and certain easily recognized characteristics of the patients; A may be better than B for patients with normal hearts, but worse for those with some kinds of heart disease.

Perhaps k should be assessed, not as a constant, but as an increasing function of $|y|/n$, since the more striking the treatment difference indicated the more likely it is that the experiment will be noticed and so will affect the treatment given to future patients. One way of introducing such a dependence of k on $|y|/n$ is to assess $k+2n$ as a constant, as though the sum of the number of patients in the trial and of the number of later patients directly affected by the trial were fixed. At any rate, the two possibilities that k is fixed and that k is a fixed number (N, say) minus $2n$ are easy to consider, and seem both to be worth thinking about.

It should be noticed that numerous assumptions have been made in deriving the expression (8) above. What assumptions are reasonable to make must be considered carefully in any particular instance. Meanwhile, we may hope that (8) is plausible enough to give some insight into the nature of the problem. A few comments about assumptions may be helpful at this point. Our assumption that the treatment differences have a normal chance distribution with known (unit) variance will not often seem plausible. Normality with weak prior information about the variance would more often be appropriate. Another rela-

tively simple but interesting assumption would be that responses to both treatments were all-or-nothing and we were comparing two binomial populations. Both these types of assumption have a nuisance parameter, and the exact expressions are more difficult to handle. Results for normality with known variance may be expected to approximate the other cases when n is large.

In expression (6) there is a tacit assumption that the two treatments are on a par. Suppose that treatment A is the standard treatment in general use and B is a new treatment. There will be more pressure to close the trial and announce the result if B appears better than A than if it does not. For if B is indeed better than A, patients elsewhere are being given, during the period of the trial, the less effective treatment, and the longer the trial continues the more such patients there will be. If this sort of asymmetry in the treatments is present, a corresponding asymmetry should be introduced into the expression for $R(n, y)$.

We have been ignoring the time and money cost of experimentation. It is unlikely that if the ethical type of cost is present, the more usual time and money cost can be included in the regret function without grave dispute. But the way in which time and money affect an experiment is usually (I believe) to impose an upper limit on the number of observations or on the duration of the experiment or on both—and duration is related to the number of observations through the rate of admission of patients to the trial. This limit will not be exceeded until the results so far obtained have been reviewed in the light of other current knowledge, and the prospects of alternative programs of research have been assessed. Thus a rough allowance for the time and money cost can be made by adding an extra condition to the design problem as already described, that n should not exceed a stated limit.

APPROXIMATE SOLUTIONS

The design problem posed above at (8) has been considered by Maurice [15], in the context, not of medical trials, but of industrial production, and from a slightly different point of view. Expectations are taken in the first instance over the sample space, for fixed θ, and then the introduction of a prior probability distribution for θ is avoided by Wald's minimax device. Now to avoid any reference to prior judgments about θ by following the minimax rule on principle seems an unnecessary and inadvisable handicap, at least in the present context. Good experimenters usually look for the things that, rightly or wrongly, they expect to find; they plan according to their expectations. Let us therefore proceed directly from expression (8) for $R(n, y)$. We shall consider the two possibilities, (i) that k is given, (ii) that $N = k+2n$ is given. We assume now that the duration of treatment and observation of a patient is shorter than the interval between admissions. Thus when the decision is made to terminate the trial, we ignore the possibility that some patients have already been admitted to the trial and undergone treatment, but have not yet been completely observed. We can therefore think of choosing a stopping boundary in the (n, y)-plane, such that when it is reached the trial ceases suddenly and no further observations accrue.

$R(n, y)$ is of course symmetric about the line $y=0$. It is easy to see that, for fixed n, $R(n, y)$ has a local maximum at $y=0$ and minima where y satisfies

$$\Phi\left(-\frac{|y|}{\sqrt{n}}\right) = \frac{n}{k+2n}. \tag{9}$$

At the minima we have

$$R(n, y) = \frac{k+2n}{\sqrt{n}}\,\phi\left(\frac{y}{\sqrt{n}}\right) \tag{10}$$

If k is fixed, the right-hand side of (10) increases with n. Then if, at any stage in the process of observation, n (the number of patient pairs observed so far) and y (the cumulative sum of the response differences) should satisfy (9), no more observations should be taken, because R must necessarily be increased by further observations. The optimum stopping boundary must therefore lie within the boundary defined by (9).[2] It seems reasonable to conjecture that the optimum boundary is quite close to (9), at any rate for the smaller values of n, and that if (9) is used to define a stopping boundary the expected regret will scarcely exceed the minimum possible. The conjecture could be checked by computation, but because of the vagueness of the data of the problem such checking hardly seems worth-while; anyway, it has not been done. (The boundary (9) is open, in the sense that it imposes no limit to the attainable values of n. But it can easily be shown that because n can assume only integer values the optimum boundary is closed. This difference in character between the boundary (9) and the optimum boundary is of no practical importance.)

If, alternatively, $N=k+2n$ is fixed, the right-hand side of (10) increases with n up to $n=0.27N$, about, and then decreases slowly until n reaches its greatest possible value of $0.5N$. It is no longer obvious that the optimum stopping boundary must lie within the boundary defined by (9), but it still seems reasonable to conjecture that the latter boundary almost minimizes the expected regret.

In Table 1 some ordinates of boundaries defined by (9) are given: first, two boundaries with k fixed ($k=100, 200$); and then three boundaries with N fixed ($N=200, 1000, 10000$). What are tabulated are ordinates of the upper part of the boundary. The ordinates of the lower part are the same with the sign changed. The first, third, and fourth of the tabulated boundaries are also shown graphically in Figure 1(a).

In view of the foregoing considerations regarding the choice of k, a simplified type of boundary as shown in Figure 1(b) would be reasonable. This is formed by a pair of horizontal lines, together with a truncation. The place and manner of truncation might well be influenced by time and money considerations.

If the uniform prior probability distribution for θ, assumed above at (5), is changed to a proper normal distribution, that is equivalent in informativeness to supplying some extra observations *free of regret*. The above type of calculation can easily be modified to allow for this situation. So long as the

[2] The optimum boundary lies *outside* the boundary yielded by a procedure recently studied by Amster [1].

TABLE 1. ORDINATES OF BOUNDARIES

n	k constant		$N=k+2n$ constant		
	$k=100$	$k=200$	$N=200$	$N=1000$	$N=10,000$
1	2.33	2.58	2.57	3.09	3.72
2	2.93	3.30	3.29	4.07	5.01
4	3.57	4.14	4.11	5.30	6.71
10	4.37	5.35	5.20	7.36	9.77
25	4.84	6.41	5.75	9.80	14.04
50	4.77	6.84	4.77	11.63	18.21
75	4.54	6.86	2.76	12.47	21.07
100	4.31	6.74	0.00	12.82	23.26
125	4.09	6.59	—	12.86	25.06
150	3.90	6.42	—	12.69	26.58

prior information is weak, the effect on the boundaries is nearly (not quite) expressed by saying that the boundaries are unaltered but the starting point for the sample path is displaced a little from the origin. Insofar as the boundaries are roughly horizontal lines, any horizontal displacement of the starting point, corresponding to a zero prior expectation for θ, makes almost no difference to the plan. Strong prior information, on the other hand, may greatly alter the plan, and even call for no experimentation at all.

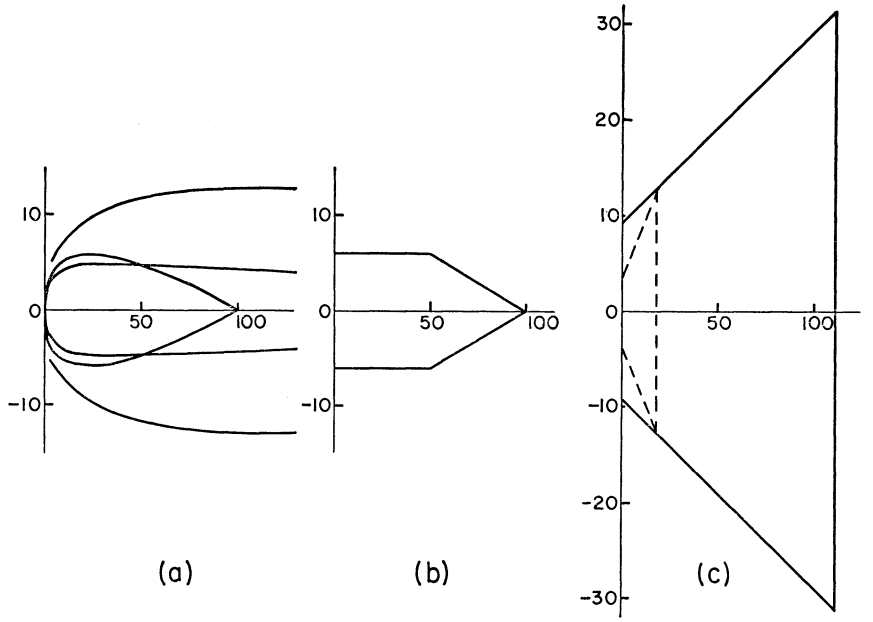

FIGURE 1. Stopping boundaries for the comparison of two treatments. The abscissa is n, the number of pairs of patients. The ordinate is y, the cumulative sum of response differences. (a) Three boundaries from Table 1. (b) A simplified boundary. (c) Two boundaries given by Armitage.

A paper by Colton [9], come to hand just as the above investigation was completed, has much in common with it.[3] In particular, taking N to be fixed, Colton has studied the optimum choice of a sequential plan defined (arbitrarily) by a pair of horizontal boundary lines. A possible choice for the latter, derived by a certain maximin argument, namely (in the present notation)

$$y = \pm \sqrt{\frac{N}{6}}, \tag{11}$$

is shown to be nearly optimum for a considerable range of prior distributions for θ. This finding is in excellent agreement with the present investigation, because the lines (11) are almost the horizontal tangents to the boundary (9) when N is fixed. For example, if $N=200$, the boundary (11) is the pair of lines $y=\pm 5.77$, and if $N=1000$ it is the pair $y=\pm 12.91$. The argument given above could be summarized by suggesting that for smaller values of n the boundary should be Colton's horizontal lines (11), where if k rather than N is thought of as fixed k may be substituted for N; and then for larger n there should be a truncation, possibly in the style of Figure 1(b), guided by the behavior of the regret function (8) and the locus (9), and perhaps influenced by the availability of time and money.

COMPARISON WITH ARMITAGE'S PLANS

Comparison is invited with the boundaries given by Armitage in his Tables 5.1 (open plans) and 5.2 (closed plans). He suggests that the closed plans are more suitable, so let us consider those now. In order to choose one of Armitage's plans, the experimenter must decide on the "significance level" or chance of error of the first kind, and also must decide on the "critical value" of $|\theta|$, the treatment effect, such that the correct preference has 95 per cent chance of being reached. But in these tables just one significance level, 5 per cent, is catered for, so effectively the critical value of $|\theta|$ is the only thing he has to worry about. Now if the experimenter feels that even a small treatment effect is worth detecting, he will be tempted to choose a low critical value, and so take one of the higher entries in Armitage's Table 5.2. The third plan is shown in Figure 1(c); this is the earliest plan in the table that can be shown conveniently on the same scale as Figures 1(a) and (b). Another of Armitage's plans is also shown (broken lines), from lower down the table.

Evidently Armitage's plans are quite different from those illustrated in Figures 1(a) and (b) and in Table 1 of this paper. According to the interpretation given here of the ethical problem, the sloping part of the larger boundary in Figure 1(c) would be ethically justified only if the number k of patients whose treatment would be decided by the outcome of the trial were in the tens of thousands. But if k were really as large as that, it would surely be quite unreasonable to insist on terminating the trial after only 111 pairs of patients had been treated. The smaller boundary is unreasonable in the same way. Its sloping part seems to suggest that k is of the order of 5000; then why prohibit continuation of the trial beyond 18 pairs?

[3] See also Dunnett [10].

These remarks concern Armitage's plans viewed merely as stopping rules. It has already been pointed out that experimental designs are not properly to be regarded as decision rules, and the correct statistical analysis of some observations does *not* consist merely of announcing that the treatment difference is or is not "significant." Suppose that the larger boundary in Figure 1(c) is used and observation ceases when $n=111$, $y=30$. The likelihood function for θ is

$$\exp\,[-55.5(\theta - 0.27)^2], \tag{12}$$

and the experimental evidence thus points strongly to positive values for θ. Armitage would have us conclude that θ does not differ significantly from 0. Now this conclusion might be reasonable in some circumstances. Suppose that the treatments A and B were intended to be exactly equivalent, one being perhaps a synthetic version of the other, and the purpose of the trial was to verify that this intention had been achieved. If no unexpected side effect is important, the treatments will be equivalent, and θ will be (very close to) 0. But possibly there is some real difference in effectiveness, and the equivalence is only partial. Against a background such as this it might be reasonable to conclude from the likelihood function (12) that probably $\theta=0$. That would depend on (i) the prior probability attached to $\theta=0$ and on (ii) the way the rest of the prior probability was distributed over other values of θ. (This type of inference situation has been studied by H. Jeffreys, D. V. Lindley, and L. J. Savage.) The conclusion may be reasonable and convincing, but the reasons need to be carefully explained, and in any case the likelihood function (12) or its equivalent should be quoted. Although a number of actual experiments are referred to in Armitage's book, none appears to have this character. In no case does there seem to be any ground for assigning a sizable lump of prior probability to the hypothesis that $\theta=0$. Conclusions of no significant difference are therefore unsatisfactory.

Oldham [16] has remarked: "There are peculiar difficulties that arise when a clinical trial ends without significant differences between treatments, or when a sequential trial reaches the boundary which, somewhat curiously, requires one to decide not to make a decision. In the medical field the temptation to regard the null hypothesis as then proved seems particularly strong, yet in this field a small difference, say in case-fatality rates, which could be shown to be real by a large enough trial, can seldom be said to be of no material significance."

ETHICAL CONSIDERATIONS

(a) Clinical experimentation in the treatment of serious illnesses has been subject to controversy. There is a well-known legal principle that ignorance of the law is not a valid reason for breaking it. It would be not unjust to adopt an equally severe attitude towards medical practice, that ignorance of an effective method of treatment is not a valid reason for ineffective treatment. It is not sufficient that a physician should give the best treatment he knows about; he is responsible for his own knowledge. Anything he can do to help improve his knowledge, and that of other physicians, is as much his duty as the correct application of his knowledge in treating patients.

From this point of view, there can be no doubt as to the propriety of clinical

experimentation as such. In the improvement of knowledge some patients will be ill treated, but if knowledge is not improved all patients will be ill treated. Let something be learned from the mistakes.

Consider the following imaginary situation. An unprecedented accident has occurred, and some fifty-seven persons arrive together in a hospital in serious condition. The hospital staff judge that a day or two may properly be devoted to careful discussion and thought about the course of treatment to be adopted. It is not clear what relevance previous experience has to the present malady. After much discussion, two methods of treatment, A and B, stand out as worthy of consideration. They are quite different, and once undertaken must be followed through; their effectiveness can only be judged some time after treatment ends. It is to be expected that, because the two treatments are so different in method, they may differ considerably in effectiveness. Which treatment should be adopted, or should both treatments be tried out in a controlled experiment?

If the doctors are confident that treatment A offers a substantially better prospect of success than treatment B, for all patients, then ordinary medical ethics will require that treatment A be used for all patients.

But suppose that the doctors are far from confident about the relative merits of A and B, and they merely think that on balance A has a slightly better prospect than B. It is now proper to consider the improvement of knowledge. If the accident that gave rise to this condition were thought to be unique, never to be repeated, there would be little case for running an experiment to compare A and B. It might even be suggested that the less anyone could subsequently discover about how these patients ought to have been treated the better. But if other accidents of the same sort can reasonably be expected in future, the case for experimentation is surely strong. At present, medical opinion sees little to choose between A and B. To experiment now has little effect on the prospects of these patients, but may greatly affect the prospects of future patients. A doctor's responsibility to his future patients, and to all patients everywhere, should no doubt be discounted in weight, in comparison with his immediate responsibility, but not denied. (In terms of the foregoing theory, the discounting is effected by choosing a modest value for k.)

(b) Decisive experiments designed to test what promise to be major innovations in the treatment of serious illnesses are liable to require much effort, and can hardly be an everyday activity for most of the medical profession. There must first, of course, be new treatments proposed that warrant testing, on the basis of experiments on animals and possibly of theoretical considerations. There must be a sufficient supply of suitable patients during some reasonable time available for the experiment. The actual conduct of the trial usually involves cooperation by many persons, a great deal of thought in planning and care in execution. Terms must be precisely defined, biases must be avoided, results of tests and assessments must be carefully recorded and processed, and a full report must be published. A broad survey of these matters is given in Hill [13].

However, it seems possible that the moral obligation to improve knowledge by experimentation could be discharged also by a humbler but more wide-

spread type of investigation, analogous to the evolutionary operation that has proved so valuable in industry [8]. Within any established method of treatment there are many points of detail that are open to question and on which there is a variation in practice. Of a drug or radiation treatment one may ask, how large should the dose be, how often given, and for how long? It would be possible for a single physician or group of physicians to vary deliberately one or more of these details, keeping at first within the range of accepted practice, and later going outside that range if results so indicated. All identifiable illness conditions that are commonly encountered could be the object of continual experimentation, leading in course of time to improvements in effectiveness of the standard treatment if any improvement should be possible. If it happened that the standard treatment was ineffectual or even harmful, the fact would eventually come to light and the treatment would be dropped. A remarkable feature of this type of experimentation is that it is so free from even the most perverse ethical objections, since it is experimentation with apparently unimportant details only.[4]

If such informal experimentation indicated that a substantial departure from usual practice should be made, that finding could be tested by a more formal, more carefully controlled trial of the type previously mentioned.

These brief remarks on the ethics of experimentation have been made in order to define more clearly the point of view from which the preceding analysis of the design problem was developed.

SUMMARY

The criticisms of Armitage's book advanced above are of course controversial, and have therefore been made at some length. To sum up, here are the most essential points affecting the conduct and reporting of a sequential medical trial. They are expressed as emphatically as possible (some will say, dogmatically), for the sake of clarity and because the controversy is important. As in Armitage's book, it has been supposed here that the "model" (specification of the admissible chance hypotheses) is given and unquestioned, so that from the observations a well-defined likelihood function can be calculated, involving a finite number of parameters.

1. *Analysis.* "Sequential analysis" is a hoax. The correct statistical analysis of the observations consists primarily of quoting the likelihood function. So long as all observations made are fairly reported, the sequential stopping rule that may or may not have been followed is irrelevant. The experimenter should feel entirely uninhibited about continuing or discontinuing his trial, changing his mind about the stopping rule in the middle, etc., because the interpretation of the observations will be based on what was observed, and not on what might have been observed but wasn't.

2. *Design.* The experimenter does not decide the treatment of future patients.

[4] Professor L. J. Savage has argued this case cogently in private conversation. From another source I have the comment that once a drug is on the market, to prescribe less than the commonly accepted dose (or in other than the commonly accepted manner) has been held by some courts to be malpractice. Thus there could be legal difficulties. In any case this type of experimentation would require care and discipline, with random allocation of treatments to patients and other precautions, if it were to have much scientific value. The suggestion is made here tentatively.

A medical trial is not, in any clearcut fashion, a decision procedure. (Hence the above remarks on analysis.) But in experimentation with serious illnesses, where the ethical consideration is controlling, the experimenter should assess *as best he can* how many future patients will be directly affected in the treatment they receive by the outcome of his trial, and in what way they will be affected, so that he may balance the possible ill effect of a misleading outcome against the ill effect of the inferior treatment (or treatments) tested in the trial. The balancing will necessarily be rather rough.

It may be noted that the first item above applies to any sort of sequential experiment, while the second relates specifically to medical experiments. Especially in item 2, the phrase "the experimenter" may possibly refer to a group of persons in concert rather than to a single individual.

A topic not discussed by Armitage is the goodness of fit of the assumed "model," or the advisability of substituting a different "model." The basic notions of testing goodness of fit or acceptability seem to constitute a very obscure part of statistical theory. The obscurity is especially felt when such questions are considered in regard to a sequential experiment. In fact, sequential experiments are a most stimulating and provoking topic for the statistical theorist to meditate on. Too little attention of the right sort has been paid to them. The first draft of this article concluded with some general discussion of tests of goodness of fit, which in revision has become a separate paper [3].

ACKNOWLEDGMENTS

I have benefited greatly from the comments and suggestions offered by many persons, especially the following: P. Armitage, J. Berkson, I. D. J. Bross, A. P. Dempster, W. J. Hall, P. D. Oldham, L. J. Savage, M. A. Schneiderman, and the referees.

REFERENCES

[1] Amster, S. J., "A modified Bayes stopping rule," submitted to *Annals of Mathematical Statistics*.

[2] Anscombe, F. J., "Dependence of the fiducial argument on the sampling rule," *Biometrika*, 44 (1957), 464–9.

[3] Anscombe, F. J., "Tests of goodness of fit," *Journal of the Royal Statistical Society, Series B*, 25 (1963) (in press).

[4] Armitage, P. *Sequential Medical Trials*. Springfield, Illinois: Thomas, 1960.

[5] Barnard, G. A., "Statistical inference," *Journal of the Royal Statistical Society, Series B*, 11 (1949), 115–49.

[6] Barnard, G. A., Jenkins, G. M. and Winsten, C. B. "Likelihood inference and time series," *Journal of the Royal Statistical Society, Series A*, 125 (1962), 321–72.

[7] Birnbaum, A., "On the foundations of statistical inference," *Journal of the American Statistical Association*, 57 (1962), 269–306.

[8] Box, G. E. P., "Evolutionary operation: a method for increasing industrial productivity," *Applied Statistics*, 6 (1957), 81–101.

[9] Colton, T., "A model for selecting one of two medical treatments," *Bulletin de l'Institut International de Statistique*, 39 (in press); *Journal of the American Statistical Association*, 58 (1963), 388–400.

[10] Dunnett, C. W., "Approaches to some problems in drug screening and selection." (Gordon Research Conference on Statistics in Chemistry and Chemical Engineering, August 7, 1961.)

[11] Fisher, R. A., "Theory of statistical estimation" (reprinted from *Proceedings of the Cambridge Philosophical Society*, 22 (1925), 700–25, with prefatory note added), *Contributions to Mathematical Statistics*. New York: John Wiley, 1950. Paper 11.
[12] Fisher, R. A., *Statistical Methods and Scientific Inference*. Edinburgh and London: Oliver and Boyd, 1956.
[13] Hill, A. B., (Editor), *Controlled Clinical Trials*. Oxford: Blackwell, 1960.
[14] Lindley, D. V., "Statistical inference," *Journal of the Royal Statistical Society, Series B*, 15 (1953), 30–76.
[15] Maurice, R., "A different loss function for the choice between two populations," *Journal of the Royal Statistical Society, Series B*, 21 (1959), 203–13.
[16] Oldham, P. D., Review of [13], *Journal of the Royal Statistical Society, Series A*, 124 (1961), 105–6.
[17] Raiffa, H., and Schlaifer, R., *Applied Statistical Decision Theory*. Boston: Harvard University Graduate School of Business Administration, 1961. Chapter 1.
[18] Smith, C. A. B., "Consistency in statistical inference and decision," *Journal of the Royal Statistical Society, Series B*, 23 (1961), 1–37.