ON CERTAIN ASPECTS OF SEQUENTIAL CLINICAL TRIALS

JEROME CORNFIELD and SAMUEL W. GREENHOUSE
NATIONAL INSTITUTES OF HEALTH

1. Introduction

A clinical trial can become considerably more complicated than elementary accounts would lead one to expect. Often, despite simple initial objectives, unforeseen complications develop and the questions that one started out to answer become modified. We shall give an example to illustrate this in the next section, but merely emphasize here that, because of the possibility of unforeseen complications, one needs a good deal of *flexibility* to engage in and learn from a clinical trial. Yet much of the theory that biostatisticians are accustomed to use as guides in the planning and analysis of such trials makes for inflexibility. One must specify an exact hypothesis, an exact alternative, a criterion or end point for choosing between them, numerical values for type I and type II errors, probability models, and stopping rules. If the trial is sequential and uses binomial responses patients must be paired. If it is a fixed sample size trial, some biostatisticians will not release preliminary results, even to the participating clinicians, for fear of upsetting the significance level.

The conflicting pressures towards and away from flexibility create a problem that each biostatistician actively involved in such trials tends to resolve in his own way. A solution favored by clinicians is to consider each problem on its own scientific merits and without reliance upon theoretical rules as guides. Alternatively, one can accept the inflexibility that theory seems to impose, and resist any compromise with experimental pressures for flexibility. As statisticians interested in the probabilistic basis of methodology used in practice, we clearly cannot be happy with the first solution. At the same time if biostatisticians are to be of help to biomedical scientists, we feel that a relaxation of current restrictive attitudes is essential. In what follows we describe our own efforts to resolve this conflict.

2. An example of unforeseen complications

To illustrate the effect of unforeseen complications we shall consider the results of a recently published trial concerned with the effectiveness of an estrogen, premarin, in the secondary prevention of coronary disease [1]. The trial included 275 men, each with a diagnosis of definite clinical coronary disease. All patients entering the study were assigned either to the treatment or the

placebo groups at random, and treatment was subsequently administered on a double blind basis. It consisted of the daily oral administration for the full five years of the study of either the estrogen, or the placebo. The daily dosage of estrogen administered was determined experimentally during the course of the study with a view to achieving certain desired estrogenic effects—particularly breast enlargement and depression of libido and potency. The daily dosage finally decided upon was 10 mg. All patients admitted to the study after this decision was reached were placed upon this dosage immediately, but patients admitted earlier built up to it over a period of months.

It shortly became clear, however, that a 10 mg dose, given without buildup within three months of the initial infarct had a deleterious effect. As shown in table I the number of new cardiovascular-renal events, both fatal and non-fatal, in the first two months of study was considerably higher for those receiving 10 mg of the estrogen within three months of a previous infarct.

TABLE I

PER CENT NEW CARDIOVASCULAR-RENAL EVENTS

	Estrogen	Placebo
10 mg without buildup within three months of infarct	$\frac{13}{60} = 22\%$	$\frac{2}{31} = 4\%$
All other	$\frac{4}{96} = 4\%$	$\frac{4}{68} = 6\%$

Clearly, something more than a routine analysis of a single test of a pre-specified null hypothesis was called for. What was in fact done was to analyze the results separately for those with and without the buildup and both including and excluding the first two months. The five year mortality rate on each of those axes of classification is shown in table II.

TABLE II

DEATHS PER 100 PARTICIPANTS, PLACEBO AND ESTROGEN TREATED PATIENTS

	Including 1st Two Months		Excluding 1st Two Months	
	Placebo	Estrogen	Placebo	Estrogen
All patients	34 ± 5	24 ± 4	31 ± 5	17 ± 4
All patients, exluding those receiving 10 mg without buildup within 3 months of infarct	34 ± 6	15 ± 4	32 ± 6	15 ± 4

The overall mortality rate is lower for the estrogen treated group, but the magnitude of the effect and indeed the certainty with which it is established is dependent on which of the comparisons is regarded as most appropriate. The

authors conclude that these results (and others not given here) "strongly indicate that the hormone exerted a marked beneficial effect . . . [but that] the conclusion . . . must be tempered by a note of uncertainty and caution."

Whether one shares this conclusion or not, it seems clear that the information on the findings to be communicated to anyone who wishes to form his own opinion on the effectiveness of estrogens in the secondary prevention of coronary disease is far too complex to be usefully summarized in a single global test of a hypothesis.

3. Data condensation, the likelihood principle and behaviorism

There are several morals to trials like that described in the previous section. (a) The analysis contemplated at the beginning of the study may be crucially altered as the study proceeds. (b) As a corollary, continuing analysis of the results as they accumulate is desirable, whether or not the study is formally regarded as sequential. (c) The statistician's most important function may also reduce to his most elementary one—data condensation.

This attitude has some strong theoretical implications. When data of the type shown in table II are presented as a summary of findings, the clear suggestion is that any conclusions drawn or actions taken should depend only on them and not on the stopping rules which led to them. Classical analysis, whether sequential or fixed sample size, insists, however, that conclusions should also depend on the data that were not observed but might have been—otherwise the calculation of error probabilities becomes impossible. But it is precisely this concern with the entire sample space, rather than with the single point observed, that introduces undesired inflexibility.

If inferences drawn or decisions taken were dependent only on the sample point observed and not on the remainder of the sample space, a more flexible attitude towards data analysis would result. The irrelevance of the remainder of the sample space to inference or decision is formalized by the likelihood principle [2], which asserts that all observations leading to the same likelihood function should lead to the same conclusion. The irrelevance follows from the principle because the likelihood function does not depend on the stopping rule that led to the data [3]. The biostatistician seeking a theoretical basis for flexibility in the analysis of data arising in a clinical trial is compelled therefore to consider this inferential principle with care. Most discussions of its application to sequential experimentation [4] have turned on the question of sampling to a foregone conclusion, a question we consider in sections 4 and 5.

It seems useful to consider first, however, the relation between this principle and the principle that inferential rules should be selected with due regard to their consequences, or more graphically, that rules which minimize errors of inference are to be preferred. If these two principles were in fundamental conflict, it would be difficult to accept the likelihood principle or to make it the basis of any scientifically acceptable system of data analysis. It does not seem

to us, however, that any such conflict exists. An apparent conflict arises from the traditional textbook treatment of hypothesis testing, in which rejection regions are selected so as to minimize β, the probability of a type II error, for a standard α, like 0.05 or 0.01. Clearly, if the sample space is poorly defined, so are α and β. But even when the sample space is well defined the use of a single standard significance level for all sample sizes and problems lacks behavioristic justification and leads to anomalies [5]. One must therefore balance small α against large β. If one chooses to balance them by seeking a rejection region which minimizes $\lambda\alpha + \beta$, where λ measures the undesirability of an error of the first kind relative to that of the second kind, one is led directly to the likelihood principle.

More specifically, consider a simple null hypothesis H_0, a simple alternative H_1, and k possible experimental designs (that is, different sample sizes, different stopping rules, different endpoints) by which data permitting a choice between them might be obtained. For each design the sample space can be divided into acceptance and rejection regions, with probabilities of type I and type II errors for the ith design of α_i and β_i. We then ask for a rejection region for the ith design which minimizes $\lambda\alpha_i + \beta_i$ for $i = 1, 2, \cdots, k$ where λ is the same for all designs. Consider any sample point t and the ratio of the likelihood of H_1 to H_0, given t, and denote the ratio for the ith design by $R_i(t)$. Then it is easy to show that the rejection region for the design must consist of all points for which $R_i(t) > \lambda$. Therefore, as long as λ is the same for all designs the value of the likelihood ratio is a constant yardstick, that is, it means the same thing in each design. But the likelihood principle would also lead to this conclusion and cannot therefore be considered in conflict with the principle of minimizing a linear combination of type I and type II errors.

Why minimize a linear combination? Savage and Lindley give the following justification [6], starting with a suggestion of Lehmann's [7] that we consider a set of indifference curves in the $\alpha - \beta$ plane, such that any two (α, β) points lying on the same curve are equally desirable. Suppose the points (α_1, β_1) and (α_2, β_2) corresponding to rejection regions W_1 and W_2 are two such points. Then it seems reasonable to postulate that selecting W_1 with probability p and W_2 with probability $1 - p$ would lead to a third equally desirable point, for any p between 0 and 1. If this postulate is accepted, all (α, β) points such that $\alpha = p\alpha_1 + (1 - p)\alpha_2$ and $\beta = p\beta_1 + (1 - p)\beta_2$ lie on the same indifference curve, defined by the straight line L such that $\beta = \beta_2 - \lambda(\alpha - \alpha_2)$, where

$$\lambda = (\beta_2 - \beta_1)/(\alpha_1 - \alpha_2) \tag{3.1}$$

and, because the two points are equally desirable, is positive. It is now easy to see that all other indifference curves must also be linear and parallel. Thus, consider two other points, which are also equally desirable, say (α_1, β_1^*) and (α_2, β_2^*), where $\beta_i^* < \beta_i$ for $i = 1, 2$. Since the β are smaller and the α the same, these points are to be preferred to the points (α_1, β_1) and (α_2, β_2) and hence to any point on L. We may now, as before, construct another linear indifference

curve L^*. If $(\beta_2^* - \beta_1^*)/(\alpha_1 - \alpha_2) \neq \lambda$, the L^* and L will intersect (and by making $\beta_1^* - \beta_1$ sufficiently small the intersection can always be made to occur in the positive quadrant) and some of the points on L^* will be preferable to L while others will not. This contradicts the assumption that each is an indifference curve, and L and L^* must therefore be parallel. In choosing among various possible rejection regions we will choose that with an α, β point lying on the southwesternmost indifference line, and this will be the line with the smallest value of $\lambda\alpha_i + \beta_i$.

Thus, a given sample point can be regarded as imbedded in many different possible sample spaces, and in applications like those associated with clinical trials it is not possible to specify realistically which sample space is most appropriate. What the likelihood principle offers is a balancing of the two types of error which, for a given sample point, leads to the same answer for all possible sample spaces, and in this sense combines behaviorism and realism in a way that seems appropriate for clinical trials.

4. On sampling to a foregone conclusion

Much of the material in this and the next section is a summary of material previously presented by one of us [8].

Concern that the likelihood principle may entail a disregard for the consequences of the inferential procedure used is most forcibly expressed in the problem of sampling to a foregone conclusion [9], [10]. We imagine an investigator (hypothetical, of course) with a strong prejudice against the hypothesis H_0, that the mean of a normal population θ has value zero, and who wishes to amass evidence that will support this prejudice. He decides to make sequential observations on the normal random variable x, where $Ex = 0$ and $Ex^2 = 1$, and to compute

$$t_n = \frac{1}{n}(x_1 + x_2 + \cdots + x_n), \tag{4.1}$$

for $n = 1, 2, \cdots$. He adopts the following stopping rule:

(4.2) if $|t_n| < k/n^{1/2}$, he takes another observation;

but

(4.3) if $|t_n| \geqq k/n^{1/2}$, he stops,

where $k > 0$.

After stopping, he interprets the inequality (4.3) as providing strong evidence against H_0—the larger the value of k, the stronger the evidence. His argument is that the likelihood of H_0 relative to the likelihood of the hypothesis $\theta = t_n$ is low, in fact no greater than $e^{-k^2/2}$, and that H_0 is therefore inconsistent with the evidence. Alternatively, if a normal prior with variance well in excess of unity is assigned to θ, then the posterior probability that θ is zero or has a sign

opposite to that of t_n will be small for large k, and again the inequality (4.3) might be considered as providing strong evidence against H_0.

But as several authors have pointed out, the probability that the inequality (4.3) will eventually be realized is, by the law of the iterated logarithm, essentially unity, even when H_0 is true. Thus, if one accepts the likelihood principle and the irrelevance of the stopping rule, one must also accept as evidence against H_0 observations that one is virtually certain to obtain when H_0 is true. If this is not an actual logical contradiction, it is uncomfortably close to one. Nor can it be dismissed by reference to the asymptotic nature of the law of the iterated logarithm, since, as Armitage has shown [11], the probability of realizing the inequality (4.3) can be substantial even for finite n. In view of this result, one must again ask whether the likelihood principle can be reconciled with concern for probabilities of incorrect conclusions. If it cannot, the flexibility provided by the likelihood principle will, in the opinion of most biostatisticians, have been purchased at too high a price.

Up to this point the likelihood principle has been considered without reference to prior probabilities. But if that principle and concern for error probabilities are to be reconciled, such probabilities must now, in our opinion, be introduced. If one is concerned about the high probability of rejecting H_0 if observation is continued until (4.3) is satisfied, it must be because some possibility of the truth of H_0 is being entertained. An appropriate formal expression of such concern is provided by assignment of a nonzero prior probability to H_0. But the use of any smooth and bounded prior probability distribution function is equivalent to the assignment of zero probability to H_0 and does not provide an expression of this concern. In fact the apparent justification for sampling to a foregone conclusion provided by the likelihood principle is entirely a consequence of assignment of zero probability to H_0 and is eliminated by assignment of a nonzero probability, no matter how small.

To show this we introduce a prior probability function for θ, namely,

$$P\{\Theta = 0\} = p,$$

$$P\{\theta \leqq \Theta \leqq \theta + d\theta\} = \frac{1 - p}{\sigma} \varphi[\theta/\sigma]\, d\theta \qquad \text{for} \quad \theta \neq 0, \tag{4.4}$$

where

$$\varphi(x) = (2\pi)^{-1/2} e^{-x^2/2}. \tag{4.5}$$

Then, by Bayes' theorem the posterior probability that $\theta = 0$, given t_n, which we denote by $P\{0|t_n\}$, is given by

$$P\{0|t_n\} = \left[1 + \frac{1 - p}{p(n\sigma^2 + 1)^{1/2}} e^{n^2\sigma^2 t_n^2/2(n\sigma^2 + 1)}\right]^{-1}. \tag{4.6}$$

We see from (4.6) that differing from zero by k standard errors need no longer imply strong evidence against H_0. In fact, setting $n^{1/2}t_n = k$ in (4.6) we note that for any $p > 0$, we have $P\{0|t_n\}$ approaches unity as n increases. Thus, although there is unit probability that $|n^{1/2}t_n|$ will eventually exceed k, it no

longer follows that there is also unit probability that $P\{0|t_n\}$ will eventually be less than any preassigned quantity. Another way of looking at this is to consider as an alternative to the stopping rules (4.2) and (4.3) the rules

(4.7) if $\alpha_1 < P\{0|t_n\} < \alpha_2$ take another observation,

(4.8) but if $P\{0|t_n\} \leqq \alpha_1$
or $P\{0|t_n\} \geqq \alpha_2$ stop,

where $0 \leqq \alpha_1 < p < \alpha_2 \leqq 1$.
The inequality (4.7) is by (4.6) equivalent to

$$(4.9) \qquad \frac{2(n\sigma^2+1)}{n\sigma^2}\log\left[\frac{p}{1-p}\frac{1-\alpha_2}{\alpha_2}(n\sigma^2+1)^{1/2}\right] < nt^2 < \frac{2(n\sigma^2+1)}{n\sigma^2}\log\left[\frac{p}{1-p}\frac{1-\alpha_1}{\alpha_1}(n\sigma^2+1)^{1/2}\right].$$

The limits on nt^2 are not independent of n, so that for $p > 0$ the rules (4.7), (4.8) and (4.2) are not consistent. For large n the upper limit of (4.9) is of the form $A + \log n$, where A does not depend on n. Since $[A + \log n]^{1/2}$ eventually becomes and remains greater than $[2 \log \log n]^{1/2}$ the probability that H_0 will be rejected, when true, by repeated tests after a sufficient accumulation of material is essentially zero.

An even stronger conclusion is possible. Suppose an effort is made to disprove H_0 by setting α_2 in (4.7) equal to unity. If one stops at all, then it will be only when $P\{0|t_n\} \leqq \alpha_1$ and H_0 hence is improbable. It is easy to show [8] that the probability of ever stopping, when H_0 is true, is less than or equal to $[\alpha_1/(1-\alpha_1)]/[p/(1-p)]$. Therefore, if one set out to disprove H_0 by continuing observation until the posterior odds in favor of H_0 are only 100α per cent of the prior odds, the chance of succeeding is at most α.

The suggestion that H_0 be assigned nonzero prior probability is not new. It was originally proposed by Jeffreys [12], has been considered by Lindley [13], and further studied by Edwards, Lindman, and Savage [14]. Our results suggest to us that this form of hypothesis testing rather than the classical form is most appropriate for situations in which flexibility is an important requirement.

5. Acceptability in practice

Intellectually satisfying though the argument of the preceding section may be, it is dependent upon the prior parameters p and σ. These are subjective, only vaguely specifiable for any one individual, and likely to vary from one individual to another. It is natural to ask whether a theoretical basis for clinical trials can be built on such a soft foundation. In this section we shall consider this question under three general headings (a) comparison with classical sequential trials (b) sensitivity of $P\{0|t_n\}$ to the choice of p and σ and (c) some experience in applying these ideas to the planning of a trial.

5.1. *Comparison with classical sequential trials.* Our major point here is that

all the arbitrary elements involved in the assignment of p and σ have their formal counterpart in classical sequential trials and that in addition, classical trials, as actually used, assume very special parameter spaces which need not, and often do not, correspond to realistic alternatives to H_0.

Consider first the following assignment of prior probabilities for the θ of the preceding section

$$P\{\theta = 0\} = \frac{1}{2}, \qquad P\{\theta = \Delta\} = P\{\theta = -\Delta\} = \frac{1}{4}, \tag{5.1}$$

and let the three possible values of the likelihood function, given t_n, be given by $n^{1/2}\varphi[n^{1/2}(t_n - \Delta)]$, $n^{1/2}\varphi[n^{1/2}t_n]$ and $n^{1/2}\varphi[n^{1/2}(t_n + \Delta)]$. No confusion will arise from referring to them as φ_{-1}, φ_0, and φ_1. Then

$$P\{0|t_n\} = \frac{\varphi_0}{\varphi_0 + \frac{1}{2}[\varphi_1 + \varphi_{-1}]}. \tag{5.2}$$

If the stopping rule provided by (4.7) and (4.8) is followed, observation is continued as long as

$$\frac{1 - \alpha_2}{\alpha_2} < \frac{\varphi_1 + \varphi_{-1}}{2\varphi_0} < \frac{1 - \alpha_1}{\alpha_1} \tag{5.3}$$

and is discontinued when either inequality is infringed.

The scheme originally proposed by Wald [15] for testing H_0 against the composite alternative $\Theta = \pm\Delta$ also involves the use of the statistic $(\varphi_1 + \varphi_{-1})/2\varphi_0$ and lower and upper limits of $\beta/(1 - \alpha)$ and $(1 - \beta)/\alpha$, where α and β are approximate probabilities of type I and type II errors. The Bayes and the classical scheme are thus identical when

$$(1 - \alpha_1)/\alpha_1 = (1 - \beta)/\alpha, \qquad (1 - \alpha_2)/\alpha_2 = \beta/(1 - \alpha). \tag{5.4}$$

But the three point parameter space assumed is quite restrictive and would rarely be appropriate. Several anomalous features of the classical sequential test are most easily understood as a consequence of this special parameter space. Thus, if Δ is set equal to 0.25 and α and β both set equal to 0.05, the limit on $|n^{1/2}t_n|$ provided by (5.3) exceeds 14 for $n = 1$ and even for $n = 10$ exceeds 5. But the reason for not regarding an observation which differs from the hypothesized value by 14 sigma as evidence against it, is that it differs by 13.75 sigma from the only alternative admitted. For the case of unknown σ the situation becomes even more anomalous. The sequential t test, as tabulated by the National Bureau of Standards [16], will not permit the rejection of H_0 until after 14 observations, for these values of Δ, α, and β, no matter what the values of the first 13 observations, even if they were all $10^{10^{10}}$, $10^{10^{10}}$, $\cdots$, $10^{10^{10}}$. This remarkable result is again understandable in the context of the special parameter space assumed for θ.

The sensitivity of the sequential boundaries to the choice of Δ is a result of the nonexistence in sequential, as contrasted with fixed sample size, trials of a uniformly most powerful one sided test of H_0 ([17], p. 101). The practical consequences of this are illustrated by a recent clinical trial on the objective efficacy of prayer [18]. The first six pairs of patients tested gave results in favor of prayer and the sequential path almost touched, but did not cross the upper boundary at $n = 6$. Subsequent results were less favorable to the treatment, and eventually the lower boundary, leading to a conclusion of no difference, was reached. The authors comment: "It is worth pointing out that, had slightly wider differences in the success rates of the two treatments been insisted on, resulting in an even larger value of θ_1 [equivalent to our Δ] than that used, the sixth result would have attained the upper boundary, and a statistically significant result claimed. This anomalous situation incidentally suggests a defect in the sequential method that does not seem to have received much attention."

Consider now modifying the assumption of a three point parameter space by the assignment of the prior probabilities given by (4.4). It is easy to see that this is equivalent to specifying as the alternative to H_0 the hypothesis that t_n has the probability density function

$$\int_{-\infty}^{\infty} \varphi(\theta/\sigma)n^{1/2}\varphi[n^{1/2}(t_n - \theta)]\, d\theta/\sigma. \tag{5.5}$$

To use this probability density function one must specify a value for the prior parameter σ. Such specification seems no more arbitrary than selecting a value for Δ, and more realistic. If σ is determined by the condition

$$\int_0^{\infty} \theta\varphi(\theta/\sigma)\, d\theta \Big/ \int_0^{\infty} \varphi(\theta/\sigma)\, d\theta = \Delta, \tag{5.6}$$

a test approximately comparable to the classical test results, but without the restriction to a three point parameter space. Using, as before, $\alpha = \beta = 0.05$, this test permits the rejection of H_0 for $n = 1$ when $t_1 \geqq 8.2$ and for $n = 10$ when $t_{10} \geqq 3.8$, these being somewhat more reasonable than the limits of 14+ and 5+ yielded by the three point parameter space. More dramatically, however, for the case of unknown variance if the prior probabilities given by (4.4) are assigned to θ, and a uniform prior is assigned to the logarithm of the unknown variance the classical inability to reject H_0 before some minimum number of observations have been made, no matter what the value of t_n, disappears. For every α_1, α_2, and σ there is a finite value of t_2 which infringes the inequality (4.7).

It thus appears to us that arbitrary elements are present in the problem no matter how it is formulated, and that the specification of p and σ, rather than introducing them, permits one to control them in a way that is not grossly inconsistent with what is actually known prior to the conduct of the trial.

5.2. *Sensitivity to choice of p and σ.* In many applications the conditional prior distribution of θ, given the falsity of H_0, will be quite diffuse, that is, σ may be quite large. But from equation (4.6) it is clear that

$$\lim_{\sigma \to \infty} P\{0|t_n\} = 1, \qquad \text{for all} \quad p > 0 \text{ and } t_n, \tag{5.7}$$

so that we cannot simply assign σ an arbitrarily large value. The ratio of the prior probability assigned to H_0 to the prior probability density assigned to θ near H_0 is $p\sigma/(1-p)$, and for large n the values of p and σ assigned affect $P\{0|t_n\}$ only through their effect on the quantity $p\sigma/(1-p)$. As a preliminary to investigating sensitivity to choice of p and σ, it seems useful, therefore, to permit σ to become infinite, and p to go to zero, but in such a way that $p\sigma/(1-p)$ remains fixed and equal to a constant c^{-1}. We then have from (4.6)

$$P^*\{0|t_n\} = \lim_{\substack{p\to 0 \\ \sigma\to\infty}} P\{0|t_n\} = \left[1 + \frac{c}{n^{1/2}} e^{nt_n^2/2}\right]^{-1}. \tag{5.8}$$

In situations in which such an assignment is appropriate the number of prior constants is then reduced from two to one. It is instructive to consider the interpretation of the prior constant, c. The limits on nt^2 implied by the stopping rule (4.7) are, using (5.8),

$$\begin{aligned} &\text{lower: } \log n + 2\log\frac{1-\alpha_2}{\alpha_2 c} \\ &\text{upper: } \log n + 2\log\frac{1-\alpha_1}{\alpha_1 c}. \end{aligned} \tag{5.9}$$

But nt_n^2 is nonnegative, so that the smallest n for which the lower limit could possibly be reached is that value for which the lower limit has value zero. For smaller values of n, the lower limit is negative and $P^*\{0|t_n\} < \alpha_2$, even for $t_n = 0$. Denoting the value of n which makes the lower limit zero by n_0, we find

$$\frac{1-\alpha_2}{\alpha_2} n_0^{1/2} = c. \tag{5.10}$$

The constant n_0 is thus the smallest number of observations leading to zero mean that will lead to the acceptance of H_0 at posterior probability level α_2. It thus provides a convenient, and operationally interpretable, way of quantifying vague prior beliefs about H_0, when the conditional prior distribution of θ, given the falsity of H_0, can be taken as diffuse.

For a clinical trial of some new form of therapy an appropriate value of n_0 for $\alpha_2 = 0.95$, say, would rarely fall below 10 or rarely exceed 1000, the lower value being perhaps reserved for forms of therapy suggested by others, and the upper for testing one's own ideas. The maximum effect of vagueness in prior beliefs on both upper and lower limits for nt_n^2 is thus, from (5.9) $\log 1000 - \log 10$ or 4.6 for all α_1 and α_2. By contrast, the effect on the upper limit of raising $1-\alpha_1$ from 0.95 to 0.99 is to raise it by 3.3 for all n. The arbitrariness introduced by the vagueness of prior beliefs about H_0 can thus reasonably be argued to be of the same magnitude as that involved in selecting a posterior probability level and a good deal less than that involved in specifying alternatives. The effect of n_0 relative to the effect of selecting a posterior probability level on the lower limit is larger. Nevertheless, the ambiguity introduced by vague prior probabilities does not, it appears to us, introduce any more qualitatively serious

vagueness than is introduced by various uncertainties about the statistical model used or by various scientific uncertainties, such as choice of treatment schedule, type of patient, or stage of the disease.

5.3. *Experience in an application.* The question of whether subjective elements can play a role in the interpretation of scientific data is of at least as much importance to scientists as it is to statisticians. The reaction of scientists, particularly those who have absorbed the frequentist tradition in statistics, to an unabashedly subjective interpretation is consequently crucial to the general applicability of the results we have presented. This reaction can be found out only by experience in a variety of situations over a considerable period of time. We should like to report our experience in the planning of a trial designed to test the efficacy of a new agent, which we refer to as U, in the treatment of myocardial infarcts shortly after the attack. But before we turn to it, we call attention to a short discussion of objectivity and subjectivity in science by George Beadle ([19], pp. 4–5), which will be of interest to statisticians of all philosophical persuasions.

One of us has been serving as a member of a group planning the trial of U. The other four members are physicians, one with considerable experience in the conduct of clinical trials, the others with a primary interest in the scientific basis of the new therapy, but none with more than a casual acquaintance with statistical methods and theory. We proposed a design and analysis based essentially on the computation of $P^*\{0|t_n\}$ (equation 5.8). We quote from a relevant portion of the planning document submitted to the committee.

"A desirable way to conduct a trial is to analyze the data periodically, say, every month, and to stop only when it can be asserted with high probability either that

(a) Some positive therapeutic effect exists or
(b) That the therapeutic effect is either non-existent or negative.

In the past it has been customary to specify the stopping rules in advance in great detail and to abide by them. Armitage gives a comprehensive discussion of such plans. Once one has specified the values of the probabilities, the magnitude of the effects to be detected and the upper limit to the number of patients to be studied, exact stopping rules can be computed.

"This is not an entirely realistic way of proceeding, however, since it is impossible to foresee all contingencies. In Stamler's study of estrogens in the long-term therapy of myocardial infarction, for example, the dosage of estrogen given was increased several times during the course of the study after balancing observed therapeutic and side-effects. For the present trial if a marked therapeutic effect appears early, the investigators might wish to continue with the hope of comparing effects among subgroups, i.e., in severe versus mild cases or in first attacks versus recurrences. If observation continues for some time, however, with the data leading to no clear-cut choice between (a) and (b), the investigators might wish to substitute for (b) the less ambitious

(b′) The therapeutic effect, if any, is to reduce mortality by no more than 10 percent

and to decide that they were not interested in continuing further if all that could be accomplished was to decide between a 0 and 10 percent reduction.

"In the past it has been necessary to use inflexible stopping rules because of limitations in statistical theory. Essentially, no method of analysis was known except for the case of a small number of pre-specified stopping rules. The situation has changed drastically and analyses which depend only on the results and not at all on the stopping rules, are now possible. This does not mean that it is undesirable to consider in advance what might be done if certain results are obtained. This is still worth doing, if only to make sure that our resources and objectives are in balance and that we have a reasonably good chance of choosing between (a) and (b) with the amount of patient material likely to be at our disposal. Several such stopping rules are given in Tables 1, 2, 3 and 4 but they are to be considered as guidelines and not strait jackets.

"Before considering these plans and their implications, however, a brief discussion of the thinking that lies behind them may be helpful. The older and the new theory differ in the questions they ask. Given a set of observations, the older theory asks: if there is no real therapeutic effect what is the probability of obtaining at least as favorable an apparent effect as that observed? The answer to this question depends on the stopping rule used. Two different investigators who obtained identical results with different stopping rules would obtain different answers to the question. Despite the identity of results one might conclude that the therapy was effective, while the other could draw no such conclusion. Even worse, unless the stopping rules were of a special kind, it would be mathematically difficult and often impossible to compute the required probabilities.

"The question asked by the newer theory is: in the light of the observations what is the probability that there is no therapeutic effect (or that it is x percent, or greater than y percent, or less than z percent, where x, y and z can be chosen at will). The older theory asks about the probability of the observations, given some true but unknown therapeutic effect, while the newer theory asks about the probability of the true but unknown therapeutic effect, given the observations. This somewhat subtle difference in concept has important practical consequences—the most immediate of which for present purposes is that the answer to the question asked by the newer theory does not depend on the stopping rule. Two investigators who obtain identical results must change in identical ways their prior opinions on what the true but unknown therapeutic effect is, no matter how different their stopping rules. The irrelevance of the stopping rule to the conclusion to be drawn appears to correspond more closely with the attitude of most experimental scientists than does the older theory's insistence on its prime importance.

"Tables 1, 2 and 3 give stopping rules for three different circumstances: a strong initial presumption that U is without therapeutic effect (Table 2), a

weak initial presumption (Table 3), and an intermediate presumption (Table 1). It has not been customary for the older theory to talk about initial presumptions, although they were there. Thus, the naive procedure of performing repeated tests of significance as data accumulate until significance is obtained is equivalent to assigning zero initial presumption to the hypothesis of no therapeutic effect, while the Wald three decision sequential procedure is equivalent to assigning a prior probability of one-half to this hypothesis. The major effect of the initial presumption on the stopping rule is on the minimum number of observations required before an observed therapeutic effect of zero would lead one to stop and conclude that the agent is ineffective or harmful—the stronger the initial presumption the sooner a zero effect would lead one to stop. The tables correspond to a minimum number of 22 deaths (strong prior presumption), 225 deaths (intermediate) and 2250 (weak). This would seem to more than blanket the range of reasonable opinion. The decision on when to give up in the face of no therapeutic effect is clearly the investigator's and any form of mathematics that presumes to make this decision for him is in principle wrong. Comparison of the three tables shows that the amount of evidence required to assert the existence of a positive therapeutic effect is largely independent of the initial presumption and comparison with Table 4 which gives the values for $P = 0.99$ shows that a change in significance level has only a slightly more pronounced effect on the critical values than a change in prior presumption."

Although there is obviously some special pleading in the document, we believe that the subjective component to the proposal is honestly stated, and that no wool was being pulled over anyone's eyes. The committee found the proposal interesting, welcomed the flexibility it provided, understood the subjective elements involved, and accepted it as the basis for further planning.

6. Choice of stopping rule

At any given point in a sequential trial it is useful to have a decision rule, which indicates under what circumstances one would stop and accept or reject H_0 (or act as if one did)—even though further observations might introduce unforeseen complications that led to a subsequent change in the decision rule, and even though the terminal conclusions or decisions depend only on the terminal observations and not on the decision rule or rules used along the way. In considering possible decision rules we note first that the rules (4.7) and (4.8) have a certain intuitive appeal, but appear to command no general support in decision theory. More specifically, consider a choice between H_0 and H_1 and define the continuation region after n observations by

$$b_n < R(t_n) < a_n, \tag{6.1}$$

where $R(t_n)$ is the likelihood ratio of section 3. Kiefer and Weiss [20], [21] have shown that for densities from the exponential family, and for max $n = N$, that a minimum cost solution implies that the sequence of a_n and b_n will usually

satisfy $b_n < b_{n+1}$ and $a_n > a_{n+1}$. This is in contradiction to the constant probability limits given by (4.7) and (4.8).

In attempting to understand the reasons for this result we have found it helpful to consider a two stage sequential scheme, for which the Kiefer and Weiss result can be exhibited in an elementary way. For the second or terminal stage we consider the three decisions, accept H_0, accept H_1, and suspend judgment and stop. The cost of accepting H_0 when H_1 is true is 1, of accepting H_1 when H_0 is true is c, of suspending judgment when H_0 or H_1 is true is d_0 or d_1. Then the expected costs of the three decisions are respectively $1 - P^*\{0|t_{n_2}\}$, $cP^*\{0|t_{n_2}\}$, and $(d_0 - d_1)P^*\{0|t_{n_2}\} + d_1$. Hypothesis H_0 will be accepted if the expected cost of its acceptance is the smallest, that is, if

$$P^*\{0|t_{n_2}\} > \max\left[\frac{1}{c+1}, \frac{1-d_1}{d_0 - d_1 + 1}\right], \tag{6.2}$$

and rejected if the expected cost of accepting H_1 is the smallest, that is, if

$$P^*\{0|t_{n_2}\} < \min\left[\frac{1}{c+1}, \frac{d_1}{c + d_1 - d_0}\right]. \tag{6.3}$$

At stage 1 we consider the same three terminal decisions, with the same costs, but consider a fourth decision—to suspend judgment and go on to stage 2. The expected cost of this is nonnegative and denoted by k. Then at stage 1, the result will be H_0 accepted if

$$P^*\{0|t_{n_1}\} > \max\left[\frac{1}{c+1}, \frac{1-d_1}{d_0 - d_1 + 1}, 1 - k\right], \tag{6.4}$$

and rejected if

$$P^*\{0|t_{n_1}\} < \min\left[\frac{1}{c+1}, \frac{d_1}{c + d_1 - d_0}, \frac{k}{c}\right]. \tag{6.5}$$

But clearly

$$\max\left[\frac{1}{c+1}, \frac{1-d_1}{d_0 - d_1 + 1}, 1 - k\right] \geqq \max\left[\frac{1}{c+1}, \frac{1-d_1}{d_0 - d_1 + 1}\right] \tag{6.6}$$

and

$$\min\left[\frac{1}{c+1}, \frac{d_1}{c + d_1 - d_0}, \frac{k}{c}\right] \leqq \min\left[\frac{1}{c+1}, \frac{d_1}{c + d_1 - d_0}\right] \tag{6.7}$$

and for sufficiently small k both inequalities will be strict. Constant probability limits can therefore by no means be taken for granted. When k is expressed as an explicit function of the elementary costs and the cost of additional observations, preliminary calculations based on an exponential density indicate that the limits for the two stages can differ considerably.

The major problem in applying this kind of thinking is that the cost of additional observations tends to be incommensurable with the cost of the three terminal decisions. An alternative formulation of stopping rules independently proposed by Anscombe [22] and Colton [23] avoids the incommensurability

problem by considering that the major cost of a trial depends upon both the number of patients assigned to the inferior treatment and the treatment difference, while the major cost of stopping too soon is the cost of perhaps assigning a finite number of future patients to be treated to what is in fact the inferior treatment. They both propose to stop when the sum of these expected costs is a minimum. Their solution is unfortunately quite sensitive to the value assumed for the future number of patients to be treated. Several authors have remarked that for an infinite number of future patients the Anscombe and Colton solution requires indefinite continuation.

A generalization of their procedure may avoid this dependence, however. Using Colton's notation, he proposes that of N patients, n are assigned to each of two treatments and $N - 2n$ to the apparently superior one. The value of n is determined by minimizing the expected loss, which is given as

$$c\delta[n + (N - 2n)P\{\text{select inferior}\}], \tag{6.8}$$

where c is a proportionality constant, and δ is the true but unknown difference in means between the treatments. Colton assumes a prior distribution for δ with zero mean, and finds the n which minimizes the average of (6.8) over this prior. As a generalization, consider that at any point in the trial we assign n_1 patients to the treatment which then appears superior and n_2 to that which appears inferior. The expected cost is then

$$\begin{aligned} &c\delta[n_2 + (N - n)P\{\text{select inferior}\}], && \delta > 0 \\ &c\delta[n_1 + (N - n)P\{\text{select inferior}\}], && \delta < 0 \end{aligned} \tag{6.9}$$

where $n = n_1 + n_2$. At this point in the trial a posterior distribution for δ, whose mean in general will not be zero, is available. Let (6.9) be averaged over this posterior and select n_1 and n_2 so as to minimize this average cost. Call the values which minimize this average $\hat{n}_1$ and $\hat{n}_2$. Then assign the next patient to the treatment which appears superior with probability $\hat{n}_1/(\hat{n}_1 + \hat{n}_2)$. Repeat this calculation and random assignment after each new result. The ethical advantages of such a rule as well as its relative insensitivity to the value of N selected need no elaboration.

Finally, we mention the famous two armed bandit problem [24], whose general solution, would, in medical terminology, permit one to assign each new patient in a trial in such a way as to maximize the number of patients assigned to the superior treatment.

Clearly, much work remains to be done in the theoretical study of possible decision rules and their consideration from the point of view of application. The abstract ideas of statisticians have had important effects on clinical experimentation in the past, and there is every indication that they will continue to do so in the future. But even short of the future advances that we may expect, there is a simple application of existing ideas, not now used in clinical trials, which can be of assistance in considering whether to stop. This application depends on nothing more profound than the idea that in considering whether to stop it

would be well to have some idea of what information is likely to be provided by future observations. More specifically, if t_m is the mean of m future observations, we need the posterior distribution of t_m, given t_n. Roberts [25] has recently reviewed this problem, and termed this posterior distribution the predictive distribution.

If $p(\theta|t_n)$ denotes the posterior distribution of θ, given t_n, then the predictive distribution of t_m, given t_n is, using the notation of the two previous sections,

$$p(t_m|t_n) = \int_{-\infty}^{\infty} m^{1/2}\varphi[m^{1/2}(t_m - \theta)]p(\theta|t_n)\, d\theta. \tag{6.10}$$

But

$$\begin{aligned} p(\theta|t_n) &= P^*\{0|t_n\} \text{ for } \theta = 0 \\ &= [1 - P^*\{0|t_n\}]n^{1/2}\varphi[n^{1/2}(\theta - t_n)], \qquad \theta \neq 0, \end{aligned} \tag{6.11}$$

so that

$$\begin{aligned} p(t_m|t_n) = {} & P^*\{0|t_n\}m^{1/2}\varphi[m^{1/2}t_m] \\ & + [1 - P^*\{0|t_n\}]\left(\frac{mn}{m+n}\right)^{1/2}\varphi\left[\left(\frac{mn}{m+n}\right)^{1/2}(t_m - t_n)\right]. \end{aligned} \tag{6.12}$$

The predictive distribution of t_m is therefore a mixture of two normal densities with means zero and t_n, variances m^{-1} and $m^{-1} + n^{-1}$ and mixing coefficients $P^*\{0|t_n\}$ and $1 - P^*\{0|t_n\}$. If one now wishes the predictive distribution of $P^*\{0|t_m, t_n\}$, we have

$$P^*\{0|t_m, t_n\} = \left[1 + \frac{c}{(m+n)^{1/2}}\, e^{(m+n)t^2_{m+n}/2}\right]^{-1}, \tag{6.13}$$

where

$$t_{m+n} = \frac{nt_n + mt_m}{n+m} \tag{6.14}$$

and

$$p[P^*\{0|t_m, t_n\}|t_n] = \frac{p(t_m|t_n)}{\partial P^*\{0|t_m, t_n\}/\partial t_m}, \tag{6.15}$$

where (6.13) is used to express t_m in (6.15) as a function of $P^*\{0|t_m, t_n\}$. If this predictive distribution is calculated for various values of n_0 and m for the entire body of experience, and subclassifications of it, then there is provided a summary of what has been learned, and what future observations will add, which must be taken into account in even the most informal assessment of whether to stop or to continue.

REFERENCES

[1] J. STAMLER, R. PICK, L. N. KATZ, B. M. KAPLAN, D. M. BERKSON, and D. CENTURY, "Effectiveness of estrogens for therapy of myocardial infarction in middle-aged men," *J. Amer. Med. Assoc.*, Vol. 183 (1963), pp. 632–638.

[2] A. BIRNBAUM, "On the foundations of statistical inference," *J. Amer. Statist. Assoc.*, Vol. 57 (1963), pp. 269–326.

[3] F. J. ANSCOMBE, discussion to D. V. Lindley, "Statistical inference," *J. Roy. Statist. Soc. Ser. B*, Vol. 15 (1953), pp. 30–76.
[4] P. ARMITAGE, "Sequential medical trials: some comments on F. J. Anscombe's paper," *J. Amer. Statist. Assoc.*, Vol. 58 (1963), 384–387.
[5] J. CORNFIELD, "Sequential trials, sequential analysis and the likelihood principle," *Amer. Statist.*, Vol. 20 (1966), pp. 18–23.
[6] L. J. SAVAGE, "The foundation of statistics reconsidered," *Studies in Subjective Probability* (edited by H. E. Kyburg, Jr. and H. E. Smokler), New York, Wiley, 1964.
[7] E. L. LEHMANN, "Significance level and power," *Ann. Math. Statist.*, Vol. 29 (1958), pp. 1167–1176.
[8] 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.
[9] H. ROBBINS, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, Vol. 58 (1952), pp. 527–536.
[10] F. J. ANSCOMBE, "Fixed sample-size analysis of sequential observations," *Biometrics*, Vol. 10 (1954), pp. 89–100.
[11] P. ARMITAGE, "Some developments in the theory and practice of sequential medical trials," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, 1967, Vol. 4, pp. 791–804.
[12] H. JEFFREYS, *Theory of Probability*, Oxford, Clarendon Press, 1948 (2nd ed.).
[13] D. V. LINDLEY, "A statistical paradox," *Biometrika*, Vol. 44 (1957), pp. 187–192.
[14] W. EDWARDS, H. LINDMAN, and L. J. SAVAGE, "Bayesian statistical inference for psychological research," *Psychol. Rev.*, Vol. 70 (1963), pp. 193–242.
[15] A. WALD, *Sequential Analysis*, New York, Wiley, 1947.
[16] NATIONAL BUREAU OF STANDARDS, *Tables to Facilitate Sequential t-Tests*, Applied Mathematics Series (NBS-AMS-7), Washington, U. S. Government Printing Office, 1951.
[17] E. L. LEHMANN, *Testing Statistical Hypotheses*, New York, Wiley, 1959.
[18] 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. 367–378.
[19] G. W. BEADLE, *Genetics and Modern Biology*, Philadelphia, American Philosophical Society, 1963.
[20] J. KIEFER and L. WEISS, "Some properties of generalized sequential probability ratio tests," *Ann. Math. Statist.*, Vol. 28 (1957), pp. 57–69; pp. 73–74.
[21] E. L. LEHMANN, "A theory of some multiple decision problems, I," *Ann. Math. Statist.*, Vol. 28 (1957), pp. 14–17.
[22] F. J. ANSCOMBE, "Sequential medical trials," *J. Amer. Statist. Assoc.*, Vol. 58 (1963), pp. 365–384.
[23] T. COLTON, "A model for selecting one of two medical treatments," *J. Amer. Statist. Assoc.*, Vol. 58 (1963), pp. 388–401.
[24] D. FELDMAN, "Contributions to the 'two-armed' bandit problem," *Ann. Math. Statist.*, Vol. 33 (1962), pp. 847–856.
[25] H. V. ROBERTS, "Probabilistic prediction," *J. Amer. Statist. Assoc.*, Vol. 60 (1965), pp. 50–62.