ENVIRONMENTAL SCIENCE RESEARCH

Recent Volumes in this Series

Volume 14 THE BIOSALINE CONCEPT
An Approach to the Utilization of Underexploited Resources
Edited by Alexander Hollaender, James C. Ailer, Emanuel Epstein, Anthony San Pietro, and Oskar R. Zaborsky

Volume 15 APPLICATION OF SHORT-TERM BIOASSAYS IN THE FRACTIONATION AND ANALYSIS OF COMPLEX ENVIRONMENTAL MIXTURES
Edited by Michael D. Waters, Stephen Nesnow, Joellen L. Huisingh, Shahbeg S. Sandhu, and Larry Claxton

Volume 16 HYDROCARBONS AND HALOGENATED HYDROCARBONS IN THE AQUATIC ENVIRONMENT
Edited by B. K. Afghan, D. Mackay, H. E. Braun, A. S. Y. Chau, J. Lawrence, O. Meresz, J. R. W. Miles, R. C. Pierce, G. A. V. Rees, R. E. White, and D. T. Williams

Volume 17 POLLUTED RAIN
Edited by Taft Y. Toribara, Morton W. Miller, and Paul E. Morrow

Volume 18 ENVIRONMENTAL EDUCATION: Principles, Methods, and Applications
Edited by Trilochan S. Bakshi and Zev Naveh

Volume 19 PRIMARY PRODUCTIVITY IN THE SEA
Edited by Paul G. Falkowski

Volume 20 THE WATER ENVIRONMENT: Algal Toxins and Health
Edited by Wayne W. Carmichael

Volume 21 MEASUREMENT OF RISKS
Edited by George G. Berg and H. David Maillie

Measurement of Risks

Edited by

George G. Berg and H. David Maillie

The University of Rochester School of Medicine and Dentistry
Rochester, New York

PLENUM PRESS • NEW YORK AND LONDON

Library of Congress Cataloging in Publication Data

Rochester International Conference on Environmental Toxicity (13th : 1980 : University of Rochester) Measurement of risks.

(Environmental science research ; v. 21)
Proceedings of the Thirteenth Rochester International Conference on Environmental Toxicity held June 2-4, 1980, at the University of Rochester, Rochester, N.Y.
Includes bibliographical references and index.
1. Environmental health—Evaluation—Congresses. 2. Environmental health—Statistical methods—Congresses. 3. Environmentally induced diseases—Congresses.
I. Berg, George G., 1919- . II. Maillie, H. David. III. Title. IV. Series.
[DNLM: I. Probability—Congresses. 2. Environmental exposure—Congresses. 3. Environmental pollutants—Toxicity—Congresses. W L EN986F v. 21 / WA 671 159 1980m]
RA565.A2R6 1980 616.9'8 81-13969
ISBN 0-306-40818-X AACR2

Proceedings of the Thirteenth Rochester International Conference on Environmental Toxicity, entitled Measurement of Risks, held June 2–4, 1980, at the University of Rochester, Rochester, New York. This was conference number CONF-800601.

A Division of Plenum Publishing Corporation
233 Spring Street, New York, N.Y. 10013

Printed in the United States of America

ACKNOWLEDGEMENT

We gratefully acknowledge the support provided for this conference by the U.S. Department of Energy, Environmental Protection Agency, Nuclear Regulatory Commission, and Department of Health, Education and Welfare (including the National Institute of Occupational Safety and Health, the National Institute of Environmental Health Sciences, the F.D.A. Bureau of Drugs and the N.I.H. Fogarty International Center), and by the University of Rochester.

We are especially indebted to Dorris Nash for efficient administrative and secretarial service in preparation for and support of the conference, to Florence Marsden for assistance during the conference, and to Rose Gering for her high-quality typing of the discussion sections.

The Conference Committee

George G. Berg, Co-Chairman
H. David Maillie, Co-Chairman
George W. Casarett
James R. Coleman
Solomon M. Michaelson
Morton W. Miller
Paul E. Morrow
Gunter Oberdorster

REMARKS ON SEQUENTIAL DESIGNS IN RISK ASSESSMENT

Teddy Seidenfeld

University of Pittsburgh
Department of Philosophy
Pittsburgh, PA 15260

Part 0 - Preliminaries

My primary aim in this talk is to review some of what I consider to be the special merits of sequential designs in light of particular challenges that attend risk assessment for human populations. In advance of a discussion of sequential experimentation, let me remind you of a distinction that I think is especially important given the title of this morning's session "Statistical Inference of Risks." There are two kinds of "inference" that are commonly called "statistical inference," and we must take care to distinguish them if we are to avoid unnecessary confusion about the relevance of values (as opposed to facts) in statistical inference of risks. First, we may understand a statistical inference to be an argument whose conclusion is a statement of, what philosophers tend to call, "a rational degree of belief," i.e. a statement of evidential support. For example, we can think of statistical conclusions, inference, of the form: on the basis of data E, the probability of H: that $\text{quantity}_1 \geq \text{quantity}_2$, is roughly .9; that is, $p(H;E) \approx .9$. Here the quantities may be place holders for risk levels (or risk indicators), e.g. a quantity may be the chance of premature death (to agents of a given type) due to increased exposure to chemical X. In Bayesian terms, the "inference," then is to a statement of posterior probability for H, given E; where H may, itself, refer to chances (objective probability). Those who try to follow "orthodox" statistics, yet who wish to retain this sense of "inference" (contra Neyman's own warnings) attempt to cite the size and power (or confidence level) in lieu of a posterior probability. However, this attempt is known to suffer from "after-trial" deficiencies. That is, the Neyman-Pearson standards (of low size and high power, say) do not

apply after-trial, once the data are fixed.

I shall not be involved, here, with this sense of "statistical inference," as the problem of design I hope to address does not fall under this sense of "inference." We will, however, become involved (in Part 2) with the clash between Neyman-Pearson and Bayesian programs of sequential design--so we will not avoid the foundational disputes altogether.

A second sense of "inference," the sense I shall use, treats a statistical inference as a decision. In the orthodox statistical parlance (of Neyman-Pearson theory), the inference is whether to accept or to reject the hypothesis H (or, as we shall include, whether to postpone that decision in order to experiment first). Here too we must be cautious and distinguish a pair of senses for "acceptance." The "orthodox" sense, paradigmatically given in problems of quality control, sees the statistical acceptance (or rejection) of an hypothesis as shorthand standing for selection of a limited course of action, e.g. send the items on for sale or send them back for recyling. That is, the options (accept/reject) are non-cognitive acts: the investigator chooses between two courses of action, neither of which involves coming to believe a hypothesis is true or false (let alone coming to believe to some degree that it is true). In Neyman's terminology, the investigator faces a problem of inductive behavior (what to do), not inductive inference (what to believe). The decision is with respect to practical consequences (payoffs), e.g. sending out a defective batch of goods, and the concern with type$_1$ and type$_2$ errors is the "orthodox" way of expressing the concern with carrying out the inappropriate act (inappropriate given the agent's practical goals and preferences). We shall (part 2) examine the propriety of using the probabilities of type$_1$ and type$_2$ errors to express such preferences in sequential designs. In short, this first sense of "acceptance" is one in which a statistical inference is a decision taken with respect to practical goals.

As an alternative, some philosophers (notably Levi [9] and Hempel [7]) have suggested theories of cognitive acceptance: deciding what to believe, decisions taken with purely cognitive goals as payoffs, e.g. Levi trades off truth of hypotheses against their cognitive content (including informativeness, simplicity, etc.). These philosophers adopt a Bayesian form of decision making. (We shall review Bayesian decision theory in the next section.) But, as I will suggest, I do not think that the problem of experimental design can be dealt with decision theoretically if only cognitive goals are recognized. (Basically, I can make no sense of "cost" with respect to cognitive goals that would

justify terminating inquiry in order to decide what to believe. Note: I do not dispute the possibility of identifying cognitive goals that would serve to rank alternative experiments according to respectable cognitive standards. Only I do not see how to identify a cognitive "cost" that would justify ceasing further inquiry prior to deciding what to believe. This concern addresses Levi's program as developed in [9].)

In outline, then, my talk contains three parts. First, I want to rehearse with you the decision theoretic grounds for choosing a sequential design (where possible) over a fixed sample size design. I shall begin with a Bayesian decision theory for this purpose. Throughout, I shall treat statistical inference as decision theoretic with at least some non-cognitive costs associated with the decision. Second, I want to review a comparison of Bayesian and orthodox N-P sequential designs--the application is to sequential medical trials (and the comparison is taken from the recommended plans of P. Armitage [orthodox] and F. Anscombe [bayesian]). I hope to shake your confidence in the N-P recommended tests (if you retain conviction in orthodox statistics) by reminding you that the before-trial/after-trial problem, which plagues orthodox statistical inference when inference is understood in the first (of the two) sense(s), i.e. as evidential, also surfaces when N-P theory is interpreted in the recommended fashion, i.e. when inductive behavior is at stake.

Third, and last, I want to sound a warning about a danger which arises in the general case of sequential decision theory, and to speculate that this danger sets up a dilemma for the philosophical debate between utilitarians and non-utilitarians, as these two parties might approach the sequential medical trials problem, discussed in part 2.

Part 1- The value of sequential designs.

A canonical decision presents a choice among a (finite) list of (terminal) options $A_1,\ldots,A_m$. There is uncertainty about which relevant state of nature $S_1,\ldots,S_n$ obtains. The agent represents this uncertainty by a subjective probability $p(S_j)$. (I assume, for simplicity, that the states are probabilistically independent of the acts, i.e. $p(S_j;A_i) = p(S_j)$, all i,j.) Outcomes of each act, A_i for a given state of nature, S_j, are known: o_{ij}. Moreover, it is assumed that there is a well defined (von Neuman-Morgenstern) utility $U(o_{ij})$, defined over the outcomes, $U(o_{ij}) = u_{ij}$. This information is conveniently summarized in the standard decision matrix:

Table I Decision Matrix

	$p(S_1)$	$p(S_2)$....	$p(S_j)$....	$p(S_n)$
A_1	u_{11}	u_{12}		u_{1n}
A_2	u_{21}	u_{22}		u_{2n}
.				
A_i			u_{ij}	
.				
A_m	u_{m1}	u_{m2}		u_{mn}

The expected utility of option A_i is just the sum: $\sum_j u_{ij}.p(S_j)$, and (Bayes') policy of maximizing expected utility is to declare as admissible any option A* whose expected utility is maximum (among the declared options).

Suppose that, in addition to the m terminal options (above), there is the option to perform a cost-free experiment E, with outcome in the sample space $\Omega = \{e_1,\ldots,e_k\}$, and then choose from among the terminal options after observing the experimental outcome. Let us assume, for simplicity, that the unknown states S_j each provide a simple statistical model for the experiment, i.e. $p(e_i;S_j)$ is well defined. Also, let us suppose that the experiment E is not trivially irrelevant to the uncertainty about the unknown states, i.e. for at least one possible outcome e*, $p(e^*;S_j)$ is, as a function of j, not a constant. (That is, the likelihood is not unity for at least one possible experimental outcome.). This insures that $p(S_j;e_i) \neq p(S_j)$ for some experimental outcome and some unknown states. After the experiment is run Bayes' rule advises choosing an option $A^*_{e_i}$ that maximizes expected utility against the posterior probability $p(S_j;e_i)$, where e_i is the experimental outcome observed. The value of $A^*_{e_i}$ will depend (usually) upon which outcome is observed. For some outcomes it will be above the value of A* (the best option available without experimenting first), for other possible experimental outcomes it may be below the value of A*. However, if we calculate the expected value of the new option to postpone decision until after performing the experiment by averaging the value of $A^*_{e_i}$ against the probability of the outcome e_i, so that the expected value of waiting to see the outcome and then choosing the best (Bayes') terminal option is:

$$\text{value of waiting: } \sum_i u(A^*_{e_i}).p(e_i) = u(A^*_E),$$

A^*_E is strictly preferred to A* (see Good [5]). That is, it is an interesting mathematical fact that decision theoretically, it always pays to postpone a decision in order to acquire cost-free information. (Here, I assume that there is no cost in "processing" the data as well.)

The point is clear. If we are to take advantage of decision theory in deciding when to "stop looking" and decide the question, there must be some cost to looking--otherwise the advice is to procrastinate. In the next section I want to take advantage of this forced concern over costs in inquiry by letting "cost for looking" include typically eithical costs when "looking" involves experimenting with human subjects. But first, let me rehearse (with the aid of a simple example) the argument in favor of sequential design over fixed sample design (see deGroot [6]).

Suppose we are faced with a decision between two options, with only two relevant states of uncertainty, and payoffs as follows:

	$p(S_1) = \pi$	$p(S_2) = 1-\pi$
A_1	0	-b
A_2	-b	0

$(b > 0)$

The expected gain from A_1 is $-(1-\pi)b$ and from A_2 is $-\pi b$, which are equal iff $\pi=1/2$. Otherwise, choose A_1 iff $\pi > 1/2$ and choose A_2 iff $\pi < 1/2$.

Next, consider the opportunity to perform an experiment with one of three possible outcomes: e_1, e_2 and e_3. If e_1 occurs we know for certain that S_1 obtains, i.e. $p(S_1;e_1) = 1$. Similarly, if e_2 occurs we learn that S_2 obtains, i.e. $p(S_2;e_2) = 1$. Finally, let e_3 be irrelevant to S_1, S_2, i.e. $p(e_3;S_1) = p(e_3;S_2) = \alpha$. Let the cost per trial of this experiment be constant, c, and we have the opportunity to repeat (independent) trials, subject to the constant cost per trial, stop when we want, and then choose one of the two terminal options A_1, A_2, (There are no added costs for delaying the decision, by assumption.)

Examine the alternative designs that involve running the experiment n times and then deciding between A_1 and A_2. That is, consider the choice of a fixed sample size experiment. After n trials either an "e_1" or "e_2" will have resulted (but not both), or else all n trials will have outcomes e_3. In the former, we will have learned which of S_1, S_2 obtains and we will lose nothing from the choice between A_1 and A_2, though we will have paid out c.n units in experimental fees. If all outcomes are "e_3" our posterior probability, $p(S_j;e_3^n)$, equals our prior probability,

$p(S_j;e_3^n)$, equals our prior probability, $p(S_j)$ and the best decision after-trial is just the same as our best decision before-trial, with the same expected return less the c·n units for experimental fees. For simplicity assume that $\pi \leq 1/2$ so that A_2 is then "best" with expected value $-\pi b$. As before, the value (pre-trial) of this experiment is obtained by weighting the values after-trial by the probability of the appropriate experimental outcome. There are only two "kinds" of outcomes to consider. With probability $(\alpha)^n$ all outcomes will be "e_3". In which case the pre-trial value of the experiment with constant result e_3 is:

$$-(\alpha)^n[\pi b + cn]. \tag{1}$$

If at least one outcome is other than an "e_3", which has probability $1 - (\alpha)^n$, the loss is just the fee.[3] Thus, the pre-trial value of the experiment with a decisive result is:

$$-[1 - (\alpha)^n](cn). \tag{2}$$

Summing these two values, the total pre-trial value of the design with fixed sample size n is:

$$-(\pi b\alpha^n + cn). \tag{3}$$

We can solve for the best fixed sample design by minimizing (3) over choices of n. Let us assume that b, c, and π are such that it is better to take at least one observation instead of deciding without any experimentation. Then the n* that minimizes (3) is such that the expected value of the design with n* observations prior to deciding between A_1 and A_2 equals

$$-([c/\log(1/\alpha)] + cn^*) = -(\pi b\alpha^{n^*} + cn^*) \tag{4}$$

where $$n^* = [\log(\pi b \log[1/\alpha]/c)]/\log(1/\alpha) \tag{5}.$$

Suppose, next, that the investigator has the opportunity to perform an instance of the experiment, observe the result, and then decide whether or not to continue experimenting in advance of a terminal decision between A_1 and A_2, with n* an upper bound on the total number of trials affordable. Clearly, once an experiment leads to a decisive result, i.e. if any trial has "e_1" or "e_2" as its outcome, it is a waste of resources to continue experimenting. In a sequential design of this sort the expected value is strictly greater than the value of the best fixed sample design, as can be seen from the following considerations.

Because of the symmetry in the statistical models, i.e. in $p(e_i/S_j)$, the probability of stopping the experiment on the m-th trial ($m \leq n^*$) is independent of the true unknown state, S_j. The pre-trial value of the sequential design is merely the sum of the expected cost for looking, i.e. $c \cdot E(N)$ where $N \leq n^*$, plus the added risk of performing the maximum n* trials where all n* results are "e_3", i.e. $\pi b\alpha^{n^*}$. Thus, the total value of this sequential design is:

$$-[\pi b\alpha^{n^*} + cE(n)]. \tag{6}$$

But since $E(N) < n^*$ (in fact, $E(N) = (1-\alpha^{n^*})/(1-\alpha)$), (6) is more (a smaller negative number) than (4) and the sequential design with bound at n* is preferred to the best fixed sample size design, if at least one observation is worth taking. The general result is easily stated: fixed sample size designs do not offer advantages over sequential designs and typically the sequential designs are strictly preferred (if available).

Before I turn to a comparison of sequential designs using Neyman-Pearson considerations (type$_1$ and type$_2$ errors) against the Bayesian styled sequential designs (as illustrated above), let me point out that the sequential design described above (with value given by (6)) is not optimal amongst sequential designs. For the problem just examined, (6) represents an improvement over the optimal fixed sample design, but (6) is a bounded design: a maximum of n* observations was permitted. If we drop the constraint that there is an upper bound to the total purchasable observations and require merely that observations have a constant cost of c units each, there is an optimal design: experiment until a decisive result(e_1 or e_2) obtains, stop experimenting and choose between A_1, A_2--which choice carries no risk since we then know which state S_j is real. The cost for this design is solely the fee for "looking" and, since with probability 1 "looking" terminates, the pre-trial value is cE(N) (where N is the number of observations taken):

$$-c \cdot E(N) = -c(1/[1-\alpha]) \tag{7}$$

Note, however, that this optimal design can, with bad luck, lead the investigator into "ruin," for there is no upper bound on the number of observations that may be taken! At each stage the investigator queries, "Is it better to decide now or to try another instance of the experiment first?" Until a decisive result is seen (and assuming that even the first trial was worth making), the answer is always, "Try again." Of course, in this case, the probability of ruin is decreasing fast enough to make the open-ended strategy best among all designs. In the final section of this talk I shall return to this matter and point out a danger lurking in open-ended designs.

Part 2: Neyman-Pearson vs. Bayesian sequential designs.

In section 1 we examined an artifical statistical decision (inference) in order to highlight the advantages of sequential designs over fixed sample designs. However, with reckless abandon I introduced the full Bayesian machinery, "prior probabilities" π in particular, so that Bayes' theorem was available to fix posterior probabilities, after-trial, and to permit post-trial evaluations of expected gains with these posterior probabilities. I am certain that you are familiar with the "orthodox" objections to this Bayesian approach; specifically, the charge that Bayesians introduce "prior" probabilities over statistical hypotheses (i) at the expense of a conceptual error--all probabilities must have objective bases and there just is no chance process underlying the determination of which state S_j obtains-- or (ii) the "prior" is ad hoc -- merely a subjective preference expressed by the investigator which has no place in scientific inference.

The program of statistical inference tracing back to the work of Neyman-Pearson attempts to avoid this "mistake" by relying on "objective" probabilities of type$_1$ and type$_2$ errors to gauge the merits of statistical tests. That is, for a test of a given size (probability of type$_1$ error--rejecting the null hypothesis when true) one attempts to maximize the power (1 - probability of type$_2$ error -- accepting the null hypothesis when false).

Let us review the application of size and power considerations to sequential tests. The following example is borrowed from P. Armitage's Sequential Medical Trials [2]. Imagine an investigation into the relative merits of two treatments, T_1 and T_2. The treatments may stand for most any contrast where, as usual, it is assumed that presence of a treatment is causally relevant to the observed quantity measured by the test. For simplicity, assume also that there is a linear model connecting treatments with "effects," so that in "like" individuals difference in observed effects is a sum of two components: treatment difference -- which is constant across individuals, plus "random" effects (uncorrelated with treatment effects) -- which, for simplicity, is of known constant variance $\sigma^2 = 1$. Armitage adopts a Normal statistical model for paired differences. That is, each trial consists of a pair of readings (x_i, y_i)--x getting T_1 and y getting T_2-- and $z_i = (x_i - y_i)$ is Normally distributed with mean μ (treatment difference) and known variance σ^2. [Depending upon the correlation due to matching factors, precision will improve with positive correlation.]

Armitage considers a context in which the investigator has three terminal options: declare (">" for "is preferable to") $T_1 > T_2$, $T_2 > T_1$, or decline the judgment of preference, i.e.

suspend judgment regarding preference. He takes the "null hypothesis" to be the state where $\mu = 0$, under which it would be "wrong" to express a preference between the treatments. In order to fix concern with the power of a test, Armitage requires the investigator identify a minimal critical difference (expressed in units of $\delta_1 = \mu/\sigma$) that is worth identifying. Once the size of a test is fixed (he uses .05--see Table II) the challenge is to find the best sequential plan whose power is at least .95 if a critical difference exists. (Note: power here measures the chance of a correct preference.) For example, if $\delta_1 = .4$ (row three of Table II) then a bounded (closed) design with a maximum of 111 trials is available--whose graph has the shape of the region in figure 1.

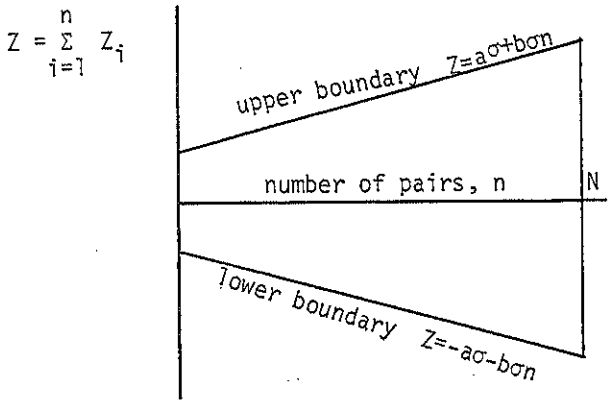

Fig. 1 Schematic representation of restricted design, with "continuation region" and "termination boundaries".

Table II

Critical value of μ/σ δ_1	Coefficients in upper and lower boundaries a	b	Maximum number of pairs allowed N
0.2	18.19	0.10	445
0.3	12.13	0.15	198
0.4	9.09	0.20	111
0.5	7.28	0.25	71
0.6	6.06	0.30	49
0.7	5.20	0.35	36

What is the underlying motive for considering sequential designs? As we noted before, unless there is some "cost" associated with making an observation, there is no decision theoretic ground for stopping experimentation prior to making a terminal choice. In medical trials there are two kinds of "costs" that spring to mind: a practical cost associated with running and processing the trials, but second (in the case of human subjects and where the terminal decision involves a recommended course of action for treating people) there is an "ethical" cost to be identified with using human subjects as a source of information.

Let us follow Neyman's advice and understand the terminal decision, whether or not to declare T_1 preferable to T_2, as a shorthand for adopting a practical course of action, say, deciding to administer (adopt) T_1 over T_2. Moreover, let us suppress (as negligible) the practical costs and focus attention on what might count as "ethical" costs. With regard to those involved in the study, i.e. with regard to those observed in the n trials taken prior to a terminal decision, half (n of them) will have received an inferior treatment (unless the null hypothesis is true). Adopting a suggestion of F.J. Anscombe's [1], let us approximate this "ethical" cost by a moral regret function $n \cdot |\delta_1|$. Moreover, if our terminal decision really amounts to recommending a treatment, then let our payoff (loss) be

$$-k \cdot \max\,(0, -\delta_1 \cdot \operatorname{sgn} \bar{z}),$$

where k is the number of individuals affected by the decision, and $\bar{z}$ is the average of paired observations.

At any stage in the sequential decision we do not know the value δ_1, so our expected payoff depends upon what has been observed. That is, at any stage in the sequential design, using these payoffs we find that our expected gain from stopping the inquiry and deciding is

$$-(n \cdot E(|\delta_1|) + k \cdot E(\max[0, -\delta_1 \cdot \operatorname{sgn} \bar{z}]). \qquad (3)$$

(3) depends, in addition to what has been observed ($\bar{z}$), on k and the prior $\pi(\delta_1)$ through the expectations taken over the posterior probability for δ_1. Since my purpose here is to compare the sequential design when Neyman-Pearson standards are used in place of Bayesian considerations, the most charitable approach is to use a prior that duplicates standard N-P inference in this example. That prior is a familiar "ignorance" prior--which is relatively flat, i.e. the "uniform" prior. Finally, in accord with Armitage's closed plans, let us consider closed (bounded) designs where at most $N = n + k/2$ pairs can be sampled. Thus, k, those affected by the decision decrease (by 2) each time a new pair is sampled.

Given all these assumptions, one can fix an optimal Bayesian sequential decision as a function of N. (The solution is obtained by a "backwards induction" argument--which is analytically quite intractable.) Anscombe provides a respectable guess at the shape (and size) of the stopping boundary: see Figure 2.

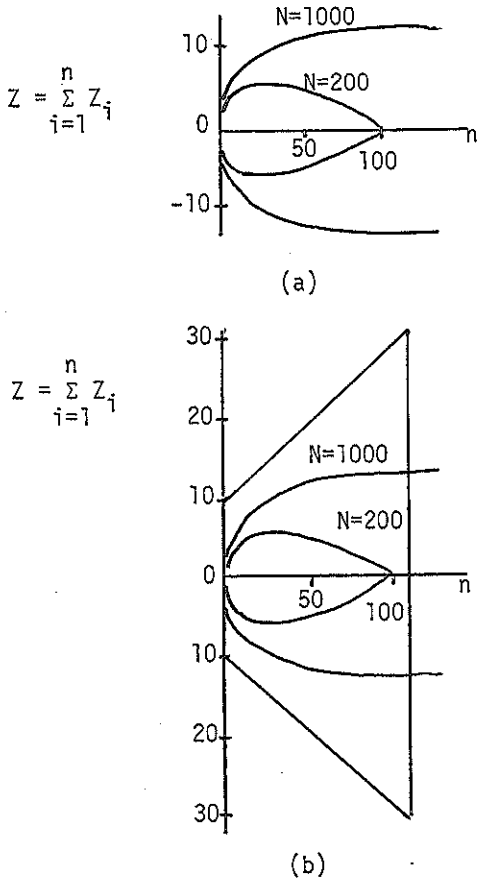

Fig. 2 Stopping boundaries for the comparison of two treatments. The abscissa is n, the number of pairs of patients. The ordinate is Z. (a) Two optimal stopping boundaries. (b) Armitage's stopping boundaries, from line 3 Table II, superimposed on (a).

How does Armitage's "orthodox" plan compare with the Bayesian solution from Anscombe's analysis? I have superimposed Anscombe's diagrams to show the great disparity of the two methods. To paraphrase Anscombe's interpretation: Armitage's boundary for a restricted sample (of at most 111 pairs) increases as though tens of thousands of individuals might be involved in the decision, but then terminates abruptly as though at most a few hundred might be involved.

If we assume a "payoff" function and "cost" function at all similar to Anscombe's "ethical" judgments, then it is quite evident that the Neyman-Pearson styled sequential tests of Armitage do not maximize expected "utility" under the standard "ignorance" prior. (Armitage's plan may be reasonable if the null hypothesis: $\delta_1 = 0$, has a high prior probability.) With these two sequential plans we confront the well known conflict between the "before-trial" concerns of Neyman-Pearson and the "after-trial" concerns of Bayesian expected utility. Prior to experimentation, Armitage's plan has the advertized low size and high power. But after experimentation, the investigator is not entitled to retain confidence in the pre-trial benefits of the plan. This is just the point of difference with the Bayesian sequential plans. True, with poor luck the optimal sequential plan may terminate well beyond the expected (pre-trial) stopping point. However, under the Bayesian plan, whenever one finally stops inquiry and makes a terminal decision it is in light of all the evidence acquired up to that point. One stops looking because, given the data actually acquired, it no longer pays to continue looking.

Part 3: Two warnings with sequential designs.

I would like to close my collection of remarks with related warnings (alerts) about the use of sequential designs. The first danger is more a mathematical point, namely, that optimal designs do not always exist and if, ignoring this fact, one proceeds with what otherwise would be a good strategy certain ruin may await. The second warning is more a philosophical point on the controversial status of the ethical theory that stands behind the kind of argument offered in the preceding section of this paper, e.g. Anscombe's "ethical" payoff functions.

(a) On dangers with unbounded designs. If the design problem has constraints that fix the options so that there is an upper bound on the number of observations that may be taken prior to a terminal decision, e.g. if the funding runs dry after n-trials, or if there is a time constraint preventing postponement beyond a given deadline, the existence of an optimal design is assured. (Solving for it, however, may be a formidable task.) Nevertheless, as we noted in part 1, the optimal design may lie outside the

class of bounded plans. In the example discussed earlier, the assumption of a constant cost per observation led to the optimal open-ended design: sample until an "e_1" or "e_2" results. Do such optimal solutions always exist? No! And the following example (due to Chow, Robbins and Siegmund [4]) reflects the seriousness of the problem.

Suppose, for simplicity, at stage n of the decision problem one has a choice between stopping at n and receiving the reward x_n, or continuing on with another trial, where the expected reward of stopping after $n+1$ (one more) trial is $E_n(x_{n+1})$. Define the set A_n as follows: $A_n = \{E_n(x_{n+1}) \leq x_n\}$ (1)
That is A_n is the set of n-fold sequences for which it does not pay to take another trial before deciding. In a choice between deciding now (at stage n) and deciding at $n+1$, it is not better to delay. Define the monotone case as one where

$$A_1 \subset A_2 \subset \ldots \tag{2}$$

That is, in the monotone case once one proceeds beyond the point where it pays to stop, one continues to face ever non-increasing expectations for terminating after another trial. A most natural candidate for an optimal stopping plan in the monotone case (one that works in bounded monotone cases) is the rule, stop at the first n such that the sequence enters A_n.

Consider, however, the following stopping problem. We are to flip a fair coin until we decide to stop, at which point we receive one of two rewards: at stage n receive $2^n(2n/n+1)$ if all n-flips have landed "heads," and receive nothing otherwise. (This problem is something like "double-or-nothing" with incentive for continuing.) At any stage n, the expectation of playing one more flip before collecting the reward is:

$$2^{n+1}(n+1)/(n+2) > x_n, \text{ if } x_n > 0 \tag{3}$$
and
$$0 = x_n, \text{ if } x_n = 0.$$

Thus, we have the monotone case since it no longer pays to play once more before stopping only when we have already "lost." But the natural candidate (identified above) for the monotone case leads us to certain ruin, as we are directed to play until a "tail" shows, i.e. until we lose! [That there is no optimal strategy here is evident as the pre-trial value of the strategy: play n times and stop is just

$$2n/(n+1), \tag{4}$$

which is increasing (in n) with limit value 2--but the limiting strategy (letting $n \to \infty$) is identical in value to what the natural candidate for the monotone case led to.]

What is the point in raising this riddle-example for your consideration? Simply this: if one approaches sequential decision problems with myopic vision; "Shall I stop here or continue on for one more trial?" the upshot of a sequence of reasonable choices may be unreasonable. That is, it is not enough to be aware of the sequential nature of experimental design problems. One must also face the difficult question of fixing the horizon of choices that exists-or else risk the unpleasant consequences that attend the strategy of looking only one step ahead.

(b) On sequential decision theory as a tool for the moral philosopher (moral investigator): Welfarism is hardly the accepted ethical banner of contemporary moral philosophy. The strategy of resolving ethical decisions involving groups of persons which we find in, say, Anscombe's sequential design is not merely welfarish but straightforwardly utilitarian. His idea is to take account of the "ethical" rewards (or, more accurately, the expected "ethical" rewards) each member of the group receives and to attempt to maximize the sum of these expectations by a clever choice of stopping rule for the experiment. However, whether or not the n-th person is exposed to what,at stage n,is thought to be the less desirable treatment depends crucially on k, the estimated patient horizon (that number waiting to be treated or to be affected by the choice of treatments). Where k is large (yet), the subject next in line has two roles to play. He represents one of the N subjects in the total population; hence, he contributes his fair share to the total "ethical" reward. However, he also represents a potential source of information about the treatment difference--which is of upmost importance for correctly treating the k agents waiting in the wings. Those fortunate ones who stand at the back of the line (of N), do not represent as important a source of information since, by the time they are observed it will be too late to alter the treatments of those who have preceded. Thus, though all N individuals are seen as "ends," each contributes an equal amount to the total "ethical" reward, those coming first are also seen as a "means" of acquiring data valuable to securing the desired "ends" for those remaining, i.e. the k next in line.

A simple answer to the question of how to temper this ethical utilitarianism is to shrink k, to make the estimate of the patient horizon modestly small. The extreme case is to abandon all utilitarian standards, to recommend what is best for that agent, as if $k=0$ and to take due note of what has been learned to that point. But now you see the dilemma that strikes me (at least) as unavoidable. The simple solution to excessive utilitarianism amounts to adopting the myopic stand and that risks falling into the worst strategy over the long run. The ethical dilemma arising from the choice of sequential decision procedures is the conflict between utilitarian and non-utilitarian standards translated into the conflict between strategic and myopic strategies. It would appear,

then, that the very machinery of sequential decision theory presupposes a commitment to utilitarianism. I will not venture here to offer a verdict on the plausibility of sequential decision theory as a neutral tool for the moral investigator. Instead I will rest content with having alerted you to the dilemma and then stop!

References

1. Anscombe, F.J., "Sequential Medical Trials," J.A.S.A., 58 (1963), 365-383.
2. Armitage, P., Sequential Medical Trials, Charles C. Thomas: [Springfield, IL., 1960].
3. _____,Statistical Methods in Medical Research, Blackwell Scientific Publications: Oxford, 1974.
4. Chow, Y.S., Robbins, H., & Siegmund, D., Great Expectations: The Theory of Optimal Stopping, Houghton Mifflin Co.: Boston, 1971.
5. Good, I.J., "On the Principle of Total Evidence," B.J.P.S., 17, #4 (1966), 319-321.
6. DeGroot, M., Optimal Statistical Decisions, McGraw-Hill Book Co.: New York, 1970.
7. Hempel, C.G., "Deductive-Nomological vs. Statistical Explanation," Minn. Studies in the Phil. of Sci., 3 (1962), 98-169.
8. Hoel, D.G., Sobel, M., & Weiss, G.H., "A Survey of Adaptive Sampling for Clinical Trials," in Perspectives in Biometrics, Vol. 1, R.M. Elashoff, ed., Academic Press: New York, 1975.
9. Levi, I., Gambling With Truth, A. Knopf: New York, 1967.