UNCERTAINTY, INFORMATION, AND SEQUENTIAL EXPERIMENTS[1]

By M. H. DeGroot

Carnegie Institute of Technology

1. Introduction and summary. Consider a situation in which it is desired to gain knowledge about the true value of some parameter (or about the true state of the world) by means of experimentation. Let Ω denote the set of all possible values of the parameter θ, and suppose that the experimenter's knowledge about the true value of θ can be expressed, at each stage of experimentation, in terms of a probability distribution ξ over Ω.

Each distribution ξ indicates a certain amount of uncertainty on the part of the experimenter about the true value of θ, and it is assumed that for each ξ this *uncertainty* can be characterized by a non-negative number. The *information* in an experiment is then defined as the expected difference between the uncertainty of the prior distribution over Ω and the uncertainty of the posterior distribution.

In any particular situation, the selection of an appropriate uncertainty function would typically be based on the use to which the experimenter's knowledge about θ is to be put. If, for example, the actions available to the experimenter and the losses associated with these actions can be specified as in a statistical decision problem, then presumably the uncertainty function would be determined from the loss function. In Section 2 some properties of uncertainty and information functions, and their relation to statistical decision problems and loss functions, are considered.

In Section 3 the sequential sampling rule whereby experiments are performed until the uncertainty is reduced to a preassigned level is studied for various uncertainty functions and experiments. This rule has been previously studied by Lindley, [8], [9], in special cases where the uncertainty function is the Shannon entropy function.

In Sections 4 and 5 the problem of optimally choosing the experiments to be performed sequentially from a class of available experiments is considered when the goal is either to minimize the expected uncertainty after a fixed number of experiments or to minimize the expected number of experiments needed to reduce the uncertainty to a fixed level. Particular problems of this nature have been treated by Bradt and Karlin [6]. The recent work of Chernoff [7] and Albert [1] on the sequential design of experiments is also of interest in relation to these problems.

2. Uncertainty and information functions. Let Ω denote the set of all possible values of some parameter θ. For simplicity, it is assumed that Ω is finite and

Received September 12, 1961.

[1] This research was supported in part by the National Science Foundation under grant NSF-G9662, and was done in part at the University of California at Los Angeles.

consists of the k distinct points $\theta_1, \cdots, \theta_k$, $k \geqq 2$. Our approach throughout the paper is Bayesian in that it is assumed that an experimenter can, at all times, assign a probability distribution over Ω that represents his beliefs about the true value of θ.

Let Ξ denote the space of all probability distributions ξ over Ω. That is, Ξ is the $(k-1)$-dimensional simplex of all vectors $\xi = (\xi_1, \cdots, \xi_k)$ such that $\xi_i \geqq 0$, $i = 1, \cdots, k$, and $\sum_{i=1}^{k} \xi_i = 1$.

An *uncertainty function* U is a non-negative measurable function defined on Ξ. Intuitively, the value $U(\xi)$ is meant to represent the uncertainty of an experimenter (measured in some appropriate units) about the true value of θ when his probability distribution over Ω is ξ. Thus, a typical uncertainty function U would take the value 0 at those distributions $\xi = (\xi_1, \cdots, \xi_k)$ for which $\xi_i = 1$ for some value of i, and might attain its maximum value at or near the distribution $(1/k, \cdots, 1/k)$. In general, however, the maximum value could be attained at any distribution in Ξ. Examples are given below.

An *experiment* X is a random variable (not necessarily real-valued), defined on some probability space, with specified conditional probability distribution given each possible value of θ. Since every finite set of probability distributions is dominated by an appropriate σ-finite measure, say μ, there is no loss of generality in representing the conditional distributions by their probability density functions $f_1, \cdots, f_k$ with respect to μ. Performing the experiment X and taking an observation on the random variable X are two ways of saying the same thing.

If the distribution over Ω prior to performing the experiment X is $\xi = (\xi_1, \cdots, \xi_k)$, then after having performed the experiment and observing the value $X = x$, the posterior distribution, $\xi(x) = (\xi_1(x), \cdots, \xi_k(x))$, over Ω is, from Bayes' Theorem,

$$\xi_i(x) = \frac{\xi_i f_i(x)}{\sum_{j=1}^{k} \xi_j f_j(x)}, \qquad i = 1, \cdots, k. \tag{2.1}$$

When ξ is the distribution over Ω, the marginal density function of X is $\sum_{j=1}^{k} \xi_j f_j$, and hence, the denominator of (2.1) vanishes with probability 0. The definition of $\xi_i(x)$ on this null set is irrelevant.

The *information* $I[X, \xi; U]$ in an experiment X when the prior distribution over Ω is ξ, relative to the uncertainty function U, is defined as

$$I[X, \xi; U] = U(\xi) - E[U(\xi(X)) \mid \xi], \tag{2.2}$$

where the random vector $\xi(X)$ is defined by (2.1) and the expectation is computed under the marginal distribution of X determined by the prior distribution ξ. Thus, $I[X, \xi; U]$ is the difference between the uncertainty prior to observing X and the expected uncertainty after having observed X.

It is commonly felt, and often stated, that an experiment can, at worst, contain no information about the problem at hand and that, typically, an experiment does contain some information. This feeling is expressed in the requirement that,

under any reasonable definition of an uncertainty function U, the information $I[X, \xi; U] \geqq 0$ for all conceivable experiments X and all prior distributions $\xi \in \Xi$.

THEOREM 2.1. *Let U be a given real-valued measurable function defined on Ξ. Then $I[X, \xi; U] \geqq 0$ for all experiments X and all $\xi \in \Xi$ if and only if U is concave*; i.e., *if and only if*

$$U(\alpha\xi + (1 - \alpha)\nu) \geqq \alpha U(\xi) + (1 - \alpha) U(\nu) \tag{2.3}$$

for all $\xi, \nu \in \Xi$ and all $0 < \alpha < 1$.

PROOF. Suppose that U is concave. Then, by the familiar Jensen's inequality, for any experiment X and any $\xi \in \Xi$,

$$E[U(\xi(X)) \mid \xi] \leqq U(E[\xi(X) \mid \xi]), \tag{2.4}$$

where $E[\xi(X) \mid \xi]$ denotes the vector $(E[\xi_1(X) \mid \xi], \cdots, E[\xi_k(X) \mid \xi])$. (It should be noted that although Jensen's inequality is often stated under the assumption that U is continuous, this assumption is not needed here.) But

$$\begin{aligned} E[\xi_i(X) \mid \xi] &= \int \xi_i(x) \left[\sum_{j=1}^{k} \xi_j f_j(x)\right] d\mu \\ &= \int \xi_i f_i(x)\, d\mu = \xi_i, \qquad i = 1, \cdots, k. \end{aligned} \tag{2.5}$$

Hence, $E[\xi(X) \mid \xi] = \xi$ and, from (2.4),

$$E[U(\xi(X)) \mid \xi] \leqq U(\xi). \tag{2.6}$$

It follows from (2.2) that $I[X, \xi; U] \geqq 0$.

Conversely, suppose that $I[X, \xi; U] \geqq 0$ for all conceivable experiments X and all $\xi \in \Xi$. Let ξ and ν be any two distributions in Ξ and let $\pi = \alpha\xi + (1 - \alpha)\nu$, where $0 < \alpha < 1$.

Consider an experiment X in which X can take only the two values 0 and 1. Let $P_j(x)$ denote the conditional probability that $X = x$ given that $\theta = \theta_j$, for $x = 0, 1$ and $j = 1, \cdots, k$. Suppose that $P_j(0) = \alpha\xi_j/\pi_j$ and $P_j(1) = 1 - P_j(0) = (1 - \alpha)\nu_j/\pi_j, j = 1, \cdots, k$. (If $\pi_j = 0$ for some value of j then $\xi_j = \nu_j = 0$ and $P_j(0)$ can be defined arbitrarily.) If the prior distribution over Ω is π then the posterior distribution after observing X is $\pi(0) = \xi$ if $X = 0$, and it is $\pi(1) = \nu$ if $X = 1$. Also $\sum_{j=1}^{k} \pi_j P_j(0) = \alpha$ and $\sum_{j=1}^{k} \pi_j P_j(1) = 1 - \alpha$. Hence,

$$E[U(\pi(X)) \mid \pi] = \alpha U(\xi) + (1 - \alpha) U(\nu). \tag{2.7}$$

Since, by assumption, $I[X, \pi; U] \geqq 0$ it follows from (2.2) that U is concave.

Theorem 2.1 indicates that it might not be unreasonable to restrict one's attention in any particular problem to concave uncertainty functions. Examples of such functions are $U(\xi) = 1 - \max\{\xi_1, \cdots, \xi_k\}$ and the famous Shannon entropy function

$$U(\xi) = -\sum_{j=1}^{k} \xi_j \log \xi_j . \tag{2.8}$$

The function (2.8) has been proposed and studied by Lindley, [8], [9] in the context being discussed here. Lindley restricts his considerations to the uncertainty function (2.8) because of certain additive properties of the resulting information function for composite experiments (Theorem 2 of [8]). However, it is not difficult to verify that these additive properties hold quite generally for information functions derived from any uncertainty functions.

The following theorem yields further examples of concave uncertainty functions.

THEOREM 2.2. *Let U be a non-negative, concave function on Ξ, and let X be an experiment. Then the function V on Ξ defined by*

$$V(\xi) = E[U(\xi(X)) \mid \xi], \qquad \xi \in \Xi, \tag{2.9}$$

is also a non-negative, concave function.

PROOF. Let ξ and ν be any two distributions in Ξ and let $\pi = \alpha\xi + (1 - \alpha)\nu$, where $0 < \alpha < 1$. It must be shown that

$$V(\pi) \geqq \alpha V(\xi) + (1 - \alpha) V(\nu). \tag{2.10}$$

Let $q(x) = \sum_{j=1}^{k} \xi_j f_j(x)$ and $r(x) = \sum_{j=1}^{k} \nu_j f_j(x)$ for all values of x. Then for each x such that neither $q(x) = 0$ nor $r(x) = 0$,

$$\begin{aligned} \pi_i(x) &= \frac{\pi_i f_i(x)}{\sum_{j=1}^{k} \pi_j f_j(x)} = \frac{[\alpha\xi_i + (1 - \alpha)\nu_i] f_i(x)}{\alpha q(x) + (1 - \alpha) r(x)} \\ &= \frac{\alpha q(x)}{\alpha q(x) + (1 - \alpha) r(x)} \left[\frac{\xi_i f_i(x)}{q(x)}\right] \\ &\qquad + \frac{(1 - \alpha) r(x)}{\alpha q(x) + (1 - \alpha) r(x)} \left[\frac{\nu_i f_i(x)}{r(x)}\right] \\ &= \frac{\alpha q(x)}{\alpha q(x) + (1 - \alpha) r(x)} \xi_i(x) + \frac{(1 - \alpha) r(x)}{\alpha q(x) + (1 - \alpha) r(x)} \nu_i(x). \end{aligned} \tag{2.11}$$

Note that the final expression for $\pi_i(x)$ in (2.11) still holds if either $q(x) = 0$ or $r(x) = 0$ (but not both). The quantities $q(x)$ and $r(x)$ vanish simultaneously only on a null set under the density $\alpha q(x) + (1 - \alpha) r(x)$. Thus, (2.11) says that for almost all x, the distribution $\pi(x)$ is a convex combination of $\xi(x)$ and $\nu(x)$. It follows from the concavity of U that for almost all x,

$$\begin{aligned} U(\pi(x)) \geqq{}& \frac{\alpha q(x)}{\alpha q(x) + (1 - \alpha) r(x)} U(\xi(x)) \\ &+ \frac{(1 - \alpha) r(x)}{\alpha q(x) + (1 - \alpha) r(x)} U(\nu)(x)). \end{aligned} \tag{2.12}$$

Hence,

$$E[U(\pi(X)) \mid \pi] = \int U(\pi(x))\,[\alpha q(x) + (1-\alpha)r(x)]\,d\mu$$

$$\geqq \alpha \int U(\xi(x))q(x)\,d\mu + (1-\alpha)\int U(\nu(x))r(x)\,d\mu \tag{2.13}$$

$$= \alpha E[U(\xi(X)) \mid \xi] + (1-\alpha)E[U(\nu(X)) \mid \nu].$$

From the definition (2.9) it is seen that this is the desired result.

Finally, an important class of concave uncertainty functions can be derived from standard statistical decision problems. In a statistical decision problem there is given a decision space A and a loss function L, assumed to be non-negative and bounded, on $\Omega \times A$. Let

$$U(\xi) = \inf_{a \in A} \sum_{j=1}^{k} \xi_j L(\theta_j, a), \qquad \xi \in \Xi. \tag{2.14}$$

Thus $U(\xi)$ is the risk from the optimal decision. It is known ([5], p. 147) that U is a continuous, concave function on Ξ. Furthermore, for any experiment X, $E[U(\xi(X)) \mid \xi]$ is the risk resulting from the Bayes decision procedure using the observation X. Hence, in this context, $I[X, \xi; U]$ is simply the reduction in risk that can be attained by performing the experiment X.

The class of uncertainty functions of the form (2.14) is quite large. Indeed, every continuous, concave U can be thought of as being of the form (2.14) for some appropriately defined A and L. The proof of this is as follows.

Every linear function $l(\xi) = \sum_{j=1}^{k} \alpha_j \xi_j + \beta$, $\xi \in \Xi$, can be thought of as defining an action a for which the loss function L takes the values $L(\theta_j, a) = \alpha_j + \beta$, $j = 1, \cdots, k$. Then $l(\xi) = \sum_{j=1}^{k} \xi_j L(\theta_j, a)$. Let F be the class of all linear functions l such that $U(\xi) \leqq l(\xi)$ for all $\xi \in \Xi$. It is well-known that $U(\xi) = \inf_{l \in F} l(\xi)$. The result follows, since to each $l \in F$ there corresponds an action and a loss vector.

The above discussion indicates that the distinction sometimes made, as in [8], between decision problems and problems in which the experimenter simply wants to gain knowledge may not be very sharp.

3. The reduction of uncertainty through sequential sampling. Let X be an experiment that can be replicated independently indefinitely. In other words, it is assumed that a random sequential sample of observations $X_1, X_2, \cdots$, each X_i having the same distribution as X, can be obtained. (By independent experiments, $X_1, X_2, \cdots$, we mean here that the joint conditional distribution of $X_1, X_2, \cdots$, given the true value of θ, is a product distribution. The joint marginal distribution of $X_1, X_2, \cdots$, when θ is considered a random variable will generally not be a product distribution.)

Consider the sequential sampling rule whereby observations are taken as long as $U(\xi(x_1, \cdots, x_n)) > \epsilon$, for some $\epsilon > 0$, and sampling stops as soon as $U(\xi(x_1, \cdots, x_n)) \leqq \epsilon$ for some value of n. This sampling rule would be of interest, for example, in those situations where the uncertainty function arises, as

in (2.14), from the loss function of a decision problem. If an experimenter wishes to control his risk from a wrong decision and is unable to assign an explicit cost per observation the above rule might be appropriate with a suitably chosen value of ϵ. Lindley has studied this rule in [8] and [9] for the uncertainty function given by (2.8) and we will now give some examples for a variety of uncertainty functions. These examples will reveal that many of the sampling plans derived by Lindley using the uncertainty function given by (2.8) can also be derived from other uncertainty functions. In some of the examples we dispense with the assumption that the parameter space Ω is finite.

EXAMPLE 3.1. Suppose that Ω contains only two points. In this case, each $\xi \varepsilon \Xi$ can be represented by its first component ξ_1 whose domain is the closed unit interval. Suppose that U is a continuous, concave function of ξ_1 such that $U(0) = U(1) = 0$. Then for any experiment X, sampling as long as $U(\xi_1(x_1, \cdots, x_n)) > \epsilon$ is equivalent to sampling as long as $\gamma < \xi_1(x_1, \cdots, x_n) < \delta$ for some γ and δ. For any prior probability ξ_1, this in turn, is equivalent to sampling as long as $A < [f_2(x_1) \cdots f_2(x_n)]/[f_1(x_1) \cdots f_1(x_n)] < B$, for some A and B. Thus, the sampling rule is a Wald sequential probability ratio test. Lindley [8], derived this result for the special case when U is given by (2.8).

EXAMPLE 3.2. Suppose that Ω is the real line and consider the family Ξ of all normal distributions over Ω. For each $\xi \varepsilon \Xi$, define $U(\xi)$ to be the variance of the distribution ξ. If the conditional distribution of X given θ is normal with mean θ and known variance, then each observation on X yields a posterior distribution over Ω with a reduced variance independent of the observed value of X. Thus, for a given prior distribution ξ, sampling until $U(\xi(x_1, \cdots, x_n)) \leqq \epsilon$ is equivalent to taking a sample of fixed size. This example can obviously be extended to include any U that is an increasing function of the variance when applied to normal distributions. It is interesting to note that Lindley [8] derived the fixed sample size plan in this context with U given by (2.8).

EXAMPLE 3.3. Consider a decision problem in which we are interested in estimating the mean of a binomial distribution with loss equal to squared error. Thus, the parameter space Ω and the decision space A are both the closed unit interval and $L(\theta, a) = (a - \theta)^2$, $a \varepsilon A$, $\theta \varepsilon \Omega$. The random variable X takes only the values 0 and 1, and the conditional probability that $X = 1$ given θ is θ. Let Ξ consist of all beta distributions over Ω. For each pair of positive constants (α, β), let $\xi_{\alpha,\beta}$ denote that distribution in Ξ with density function

$$\{[\Gamma(\alpha + \beta)]/[\Gamma(\alpha)\Gamma(\beta)]\}\theta^{\alpha-1}(1 - \theta)^{\beta-1}, \qquad 0 \leqq \theta \leqq 1. \tag{3.1}$$

As is well-known, if the prior distribution over Ω is $\xi_{\alpha,\beta}$ then the posterior distribution, after having observed the value x is either $\xi_{\alpha+1,\beta}$ or $\xi_{\alpha,\beta+1}$, according as $x = 1$ or $x = 0$. The optimal estimate of θ when the distribution over Ω is $\xi_{\alpha,\beta}$ is $\hat{\theta} = \alpha/(\alpha + \beta)$ and the resulting Bayes risk is $\alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]$. Thus, if the uncertainty function U is defined by the integral analogue of (2.14), then

$$U(\xi_{\alpha,\beta}) = \alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)], \qquad \alpha > 0, \beta > 0. \tag{3.2}$$

The sampling rule whereby observations are taken as long as

$$U(\xi(x_1, \cdots, x_n)) > \epsilon$$

can be described graphically as follows. Suppose the prior distribution over Ω is ξ_{α_0,β_0}. Then, in the $\alpha\beta$-plane, start at the point (α_0, β_0) and after each observation move either one unit in the positive α direction or one unit in the positive β direction, according as the observation is 1 or 0. Stop sampling as soon as the curve $\alpha\beta = \epsilon(\alpha + \beta)^2(\alpha + \beta + 1)$ is crossed.

Again, it is very interesting to note that through use of the uncertainty function (2.8) and some approximations, Lindley [8], [9], also arrived at sampling regions of this same form.

Further examples of this sampling rule are, of course, easily generated. When the uncertainty function is derived from a decision problem with a decision space and loss function appropriate to deciding between two composite hypotheses the resulting shape of the sampling region will typically be different from the regions derived in Examples 3.2 and 3.3. Although it was suggested above that the sampling rules discussed here might be useful in decision problems where it is difficult to assign a precise sampling cost, it is also true that these rules were used by Schwarz [10] in determining the asymptotic shapes of the optimal sequential sampling regions where the sampling cost is explicitly given.

4. The sequential design of experiments. We now turn our attention to sequential experiments such that at each stage of experimentation the experimenter is free to choose the random variable that he will observe from a given class of random variables.

Specifically, let $\Omega = \{\theta_1, \cdots, \theta_k\}$ be the finite set of all possible values of the parameter θ, let Ξ be the space of all probability distributions ξ over Ω, and let U be a given non-negative uncertainty function defined on Ξ. Let $\mathcal{C}$ be a given class of experiments X; i.e., each $X \varepsilon \mathcal{C}$ is a random variable with a known conditional density function $f_i(x)$ given $\theta = \theta_i$, $i = 1, \cdots, k$. At each stage of some overall sequential experiment, the experimenter is free to select any one of the experiments $X \varepsilon \mathcal{C}$ and observe a value of X. At each stage the selection of the experiment to be performed can depend on the outcomes of all experiments that have been performed at earlier stages. All observations are assumed to be independent in the sense that given the experiment that is to be performed at some stage, and given the true value of θ, the outcome of the experiment is independent of all previous observations.

Consider now the following problem. A fixed number, say n, of experiments are to be performed, and it is desired to select the experiments $X_1, \cdots, X_n$ sequentially so as to minimize $E[U(\xi(X_1, \cdots, X_n)) \mid \xi]$, where ξ is the prior distribution over Ω. Bradt and Karlin, [6], have shown how complicated the optimal sequential design can be, even in problems of quite simple appearance. However, by the familiar dynamic programming technique of working backward from the last stage of experimentation (see, e.g., [2] or [5], Chap. 9), an explicit rule for the construction of the optimal design can be given. In the following

derivation it is assumed that all minima taken over the class $\mathcal{C}$ are actually attained at some $X \in \mathcal{C}$.

Suppose that after having performed the first $n-1$ experiments, the posterior distribution over Ω is ξ^{n-1}. Then, clearly, the best choice for the final experiment X_n is an experiment $X_n^* \in \mathcal{C}$ such that

$$E[U(\xi^{n-1}(X_n^*)) \mid \xi^{n-1}] = \min_{X \in \mathcal{C}} E[U(\xi^{n-1}(X)) \mid \xi^{n-1}]. \tag{4.1}$$

For each $\xi^{n-1} \in \Xi$, (4.1) defines the optimal choice $X_n^*(\xi^{n-1})$ for the nth experiment.

For each $\xi^{n-1} \in \Xi$, let $U_1(\xi^{n-1})$ denote the expression (4.1). Suppose that after having performed the first $n-2$ experiments, the posterior distribution over Ω is ξ^{n-2}. Since $\xi^{n-1} = \xi^{n-2}(x_{n-1})$, where x_{n-1} is the outcome of the $(n-1)$th experiment, we want to choose X_{n-1} so as to minimize $E[U_1(\xi^{n-2}(X_{n-1}))|\xi^{n-2}]$. Thus the best choice for X_{n-1} is an experiment $X_{n-1}^* \in \mathcal{C}$ such that

$$E[U_1(\xi^{n-2}(X_{n-1}^*)) \mid \xi^{n-2}] = \min_{X \in \mathcal{C}} E[U_1(\xi^{n-2}(X)) \mid \xi^{n-2}]. \tag{4.2}$$

For each $\xi^{n-2} \in \Xi$, (4.2) defines the optimal choice $X_{n-1}^*(\xi^{n-2})$ for the $(n-1)$th experiment.

By continuing in this fashion (denoting the expression (4.2) by $U_2(\xi^{n-2})$, etc.), the optimal sequential design can be obtained. The entire procedure can be summarized as follows.

Let

$$U_0(\xi) = U(\xi), \qquad \xi \in \Xi, \tag{4.3}$$

and define $U_j(\xi), j = 1, \cdots, n-1$; recursively by the relation

$$U_j(\xi) = \min_{X \in \mathcal{C}} E[U_{j-1}(\xi(X)) \mid \xi], \qquad \xi \in \Xi. \tag{4.4}$$

Let $X_n^*(\xi)$ be defined implicitly by the relation

$$E[U_0(\xi(X_n^*)) \mid \xi] = \min_{X \in \mathcal{C}} E[U_0(\xi(X)) \mid \xi], \qquad \xi \in \Xi, \tag{4.5}$$

and, in general, let $X_{n-j}^*(\xi), j = 0, 1, \cdots, n-1$, be defined implicitly by the relations

$$E[U_j(\xi(X_{n-j}^*)) \mid \xi] = \min_{X \in \mathcal{C}} E[U_j(\xi(X)) \mid \xi], \qquad \xi \in \Xi. \tag{4.6}$$

If ξ^0 is the prior distribution over Ω then the optimal choice for the first experiment is $X_1^*(\xi^0)$. If, after observing $X_1^*(\xi^0)$, the posterior distribution over Ω is ξ^1 then the optimal choice for the second experiment is $X_2^*(\xi^1)$. In general, if the posterior distribution after the first j experiments, $j = 0, 1, \cdots, n-1$ is ξ^j, then the optimal choice for the $(j+1)$th experiment is $X_{j+1}^*(\xi^j)$. Call this rule for selecting the experiments R^*. We then know

THEOREM 4.1. *For any prior distribution, ξ^0, the rule R^* minimizes $E[U(\xi^0(X_1, \cdots, X_n)) \mid \xi^0]$ among the class of all rules for selecting the experiments $X_1, \cdots, X_n$.*

The main trouble with the rule R^* is that it is often extremely difficult to

compute for moderate values of n. The simplest form that the sequential design can take is, of course, when it specifies n replications of a single experiment in $\mathcal{C}$. The following theorem provides a sufficient condition for this to occur.

THEOREM 4.2. *If there exists an experiment $X^* \in \mathcal{C}$ such that $E[U(\xi(X^*)) \mid \xi] \leqq E[U(\xi(X)) \mid \xi]$ for all $X \in \mathcal{C}$ and all $\xi \in \Xi$, then for all values of n and all prior distributions ξ, $E[U(\xi(X_1, \cdots, X_n)) \mid \xi]$ is minimized by performing n replications of X^*.*

PROOF. It follows directly from the hypothesis of this theorem and the derivation of Theorem 4.1 that the nth experiment to be performed should always be X^*. As in (4.4), let $U_1(\xi) = E[U(\xi(X^*)) \mid \xi]$, $\xi \in \Xi$. If it can be shown that

$$E[U_1(\xi(X^*)) \mid \xi] \leqq E[U_1(\xi(X)) \mid \xi], \quad \xi \in \Xi, \qquad X \in \mathcal{C}, \tag{4.7}$$

then it would follow that the $(n-1)$th experiment should always be X^*, and, by induction, repeated use of X^* would be optimal.

Let X by any experiment in $\mathcal{C}$ other than X^* and define $V(\xi)$ by the equation $V(\xi) = E[U(\xi(X)) \mid \xi]$, $\xi \in \Xi$. If we consider an experiment in which first X^* and then X is performed, it follows by computing the posterior distribution $\xi(X^*, X)$ in two stages as $\xi(X^*)(X)$ and using the usual properties of conditional expectation that

$$\begin{aligned} E[U(\xi(X^*, X)) \mid \xi] &= E\{E[U(\xi(X^*)(X) \mid \xi(X^*)] \mid \xi\} \\ &= E\{V(\xi(X^*)) \mid \xi\}, \qquad \xi \in \Xi. \end{aligned} \tag{4.8}$$

Similarly,

$$\begin{aligned} E[U(\xi(X, X^*)) \mid \xi] &= E\{E[U(\xi(X)(X^*) \mid \xi(X)] \mid \xi\} \\ &= E\{U_1(\xi(X)) \mid \xi\}, \qquad \xi \in \Xi. \end{aligned} \tag{4.9}$$

But $\xi(X, X^*) = \xi(X^*, X)$ since the order in which the observations are taken is irrelevant to the posterior distribution. Hence, (4.8) and (4.9) are equal for all $\xi \in \Xi$.

By hypothesis, $U_1(\xi) \leqq V(\xi)$, $\xi \in \Xi$, and therefore $U_1(\xi(X^*)) \leqq V(\xi(X^*))$ for all values of X^*, and $E[U_1(\xi(X^*)) \mid \xi] \leqq E[V(\xi(X^*)) \mid \xi]$. The desired result (4.7) follows from this inequality and the equality of (4.8) and (4.9).

Theorem 4.2 leads naturally to a study of the conditions under which, for two experiments X and Y, $E[U(\xi(X)) \mid \xi] \leqq E[U(\xi(Y)) \mid \xi]$ for all $\xi \in \Xi$. Bradt and Karlin, [6], have derived many interesting results concerning this question in problems where the parameter θ can take only two values. The following Theorem 4.3 deals with the more general situation where Ω contains k points. It is a straightforward adaptation of a theorem of Blackwell, [3], [4]. It should be noted that it is now necessary to assume that U is concave, but that this assumption was not needed earlier in the section.

We first give some notation and a lemma. Let A be the set of all vectors $\mathbf{a} = (a_1, \cdots, a_k)$ with $a_i \geqq 0$, $i = 1, \cdots, k$. For any vectors $a \in A$ and $b \in A$,

let $\mathbf{a}\cdot\mathbf{b} = \sum_{j=1}^{k} a_j b_j$ and define $\mathbf{a} \otimes \mathbf{b}$ to be the vector

$$\mathbf{a} \otimes \mathbf{b} = \begin{cases} (1/\mathbf{a}\cdot\mathbf{b})(a_1 b_1, \cdots, a_k b_k) & \text{if } \mathbf{a}\cdot\mathbf{b} > 0, \\ (1, 0, \cdots, 0)\,(\text{say}) & \text{if } \mathbf{a}\cdot\mathbf{b} = 0. \end{cases} \tag{4.10}$$

LEMMA 4.1. *Let U be a concave function on Ξ. (Ξ is the subset of A containing all $\mathbf{a}$ such that $\sum_{j=1}^{k} a_j = 1$.) For any fixed $\mathbf{v} \varepsilon A$, define W on A by the expression*

$$W(\mathbf{a}) = (\mathbf{v}\cdot\mathbf{a})\,U(\mathbf{v} \otimes \mathbf{a}), \qquad \mathbf{a} \varepsilon A. \tag{4.11}$$

Then W is concave on A.

PROOF. Consider any $\mathbf{a} \varepsilon A$ and $\mathbf{b} \varepsilon A$, and any constants α and β such that $0 < \alpha < 1$, $\alpha + \beta = 1$. It must be shown that $W(\alpha\mathbf{a} + \beta\mathbf{b}) \geqq \alpha W(\mathbf{a}) + \beta W(\mathbf{b})$. If $\mathbf{v}\cdot\mathbf{a} > 0$ and $\mathbf{v}\cdot\mathbf{b} > 0$, then a simple computation yields

$$\begin{aligned} W(\alpha\mathbf{a} + \beta\mathbf{b}) &= [\alpha(\mathbf{v}\cdot\mathbf{a}) + \beta(\mathbf{v}\cdot\mathbf{b})] \\ &\quad U\left[\frac{\alpha(\mathbf{v}\cdot\mathbf{a})}{\alpha(\mathbf{v}\cdot\mathbf{a}) + \beta(\mathbf{v}\cdot\mathbf{b})}\,(\mathbf{v} \otimes \mathbf{a}) + \frac{\beta(\mathbf{v}\cdot\mathbf{b})}{\alpha(\mathbf{v}\cdot\mathbf{a}) + \beta(\mathbf{v}\cdot\mathbf{b})}\,(\mathbf{v} \otimes \mathbf{b})\right]. \end{aligned} \tag{4.12}$$

Since U is concave, it follows from (4.12) that

$$\begin{aligned} W(\alpha\mathbf{a} + \beta\mathbf{b}) &\geqq \alpha(\mathbf{v}\cdot\mathbf{a})\,U(\mathbf{v} \otimes \mathbf{a}) + \beta(\mathbf{v}\cdot\mathbf{b})\,U(\mathbf{v} \otimes \mathbf{b}) \\ &= \alpha W(\mathbf{a}) + \beta W(\mathbf{b}). \end{aligned} \tag{4.13}$$

It is easily checked that if $\mathbf{v}\cdot\mathbf{a} = 0$ or $\mathbf{v}\cdot\mathbf{b} = 0$ then $W(\alpha\mathbf{a} + \beta\mathbf{b}) = \alpha W(\mathbf{a}) + \beta W(\mathbf{b})$.

THEOREM 4.3. *Let X be an experiment taking values in the set $\mathfrak{X}$, on which is defined the σ-field $\mathfrak{A}$, and having conditional density function f_i given that $\theta = \theta_i$, $i = 1, \cdots, k$, with respect to the σ-finite measure μ on $(\mathfrak{X}, \mathfrak{A})$. Let Y be another experiment taking values in the set $\mathfrak{Y}$, on which is defined the σ-field $\mathfrak{B}$, and having conditional density function g_i given that $\theta = \theta_i$, $i = 1, \cdots, k$, with respect to the σ-finite measure ν on $(\mathfrak{Y}, \mathfrak{B})$. Let h be a non-negative measurable $(\mathfrak{A} \times \mathfrak{B})$ function on $\mathfrak{X} \times \mathfrak{Y}$ such that*

(i) $$g_i(y) = \int_{\mathfrak{X}} h(x, y) f_i(x)\, d\mu(x) \qquad \text{a.e.}(\nu), \qquad i = 1, \cdots, k;$$

(ii) $$\int_{\mathfrak{Y}} h(x, y)\, d\nu(y) = 1, \qquad x \varepsilon \mathfrak{X};$$

(iii) $$\int_{\mathfrak{X}} h(x, y)\, d\mu(x) < \infty, \qquad y \varepsilon \mathfrak{Y}.$$

Then, for any concave uncertainty function U and any $\xi \varepsilon \Xi$,

$$E[U(\xi(X)) \mid \xi] \leqq E[U(\xi(Y)) \mid \xi]. \tag{4.14}$$

PROOF. Let

$$\varphi_i(y) = \frac{\int_{\mathfrak{X}} f_i(x)h(x, y)\, d\mu(x)}{\int_{\mathfrak{X}} h(t, y)\, d\mu(t)}, \qquad y \in \mathfrak{Y}, \quad i = 1, \cdots, k, \tag{4.15}$$

and let $\varphi(y) = (\varphi_1(y), \cdots, \varphi_k(y))$. If W is defined as in (4.11), with ξ playing the role of $\mathbf{v}$, then it is readily verified that

$$\begin{aligned} E[U(\xi(Y)) \mid \xi] &= \int_{\mathfrak{Y}} W(\mathbf{g}(y))\, d\nu(y) \\ &= \int_{\mathfrak{Y}} W(\varphi(y)) \left[\int_{\mathfrak{X}} h(t, y)\, d\mu(t)\right] d\nu(y). \end{aligned} \tag{4.16}$$

Since W is concave, then for each $y \in \mathfrak{Y}$,

$$W(\varphi(y)) = W\left(\frac{\int_{\mathfrak{X}} \mathbf{f}(x)h(x, y)\, d\mu(x)}{\int_{\mathfrak{X}} h(t, y)\, d\mu(t)}\right) \geqq \frac{\int_{\mathfrak{X}} W(\mathbf{f}(x))h(x, y)\, d\mu(x)}{\int_{\mathfrak{X}} h(t, y)\, d\mu(t)}. \tag{4.17}$$

Hence,

$$\begin{aligned} E[U(\xi(Y)) \mid \xi] &\geqq \int_{\mathfrak{Y}} \int_{\mathfrak{X}} W(\mathbf{f}(x))h(x, y)\, d\mu(x)\, d\nu(y) \\ &= \int_{\mathfrak{X}} \left[\int_{\mathfrak{Y}} h(x, y)\, d\nu(y)\right] W(\mathbf{f}(x))\, d\mu(x) \\ &= \int_{\mathfrak{X}} W(\mathbf{f}(x))\, d\mu(x) = E[U(\xi(X)) \mid \xi]. \end{aligned} \tag{4.18}$$

The import of Theorem 4.3 is that if X is sufficient for Y, in the sense that a random variable with the same distributions as Y can be generated from X and an auxiliary randomization, then the information in X is no smaller than the information in Y relative to any concave uncertainty function.

Together Theorems 4.2 and 4.3 yield

COROLLARY 4.4. *If there exists an experiment $X^* \in \mathcal{C}$ that is sufficient for every other experiment in $\mathcal{C}$, then for any concave uncertainty function U, any prior distribution ξ, and any positive integer n, $E[U(\xi(X_1, \cdots, X_n)) \mid \xi]$ is minimized by performing n replications of X^*.*

5. Minimizing the expected sample size. Let U be a given uncertainty function and let $\epsilon > 0$ be fixed. Consider the sampling rule whereby experiments X_1, X_2, $\cdots$, chosen from some class $\mathcal{C}$ of possible experiments, are performed sequentially until $U(\xi(X_1, \cdots, X_n)) \leqq \epsilon$. For any sequential design S specifying the choice to be made from $\mathcal{C}$ at each stage of experimentation, let $N(S)$ a

random variable, denote the total number of experiments that must be performed. The problem to be considered in this section is that of choosing S so as to minimize $E[N(S) \mid \xi]$ for any prior distribution ξ.

Contrary to the problem considered in the preceding section, in which the sample size is fixed and the expected terminal uncertainty is to be minimized, no general rule is known for computing the solution to the problem now being considered, in which the terminal uncertainty is fixed and the expected sample size is to be minimized. One example was considered in [6]. In the remainder of this section we will consider in detail a simple example whose solution illustrates the peculiarities that can arise.

Suppose that the parameter θ can take only two values, θ_1 and θ_2. Then each $\xi \in \Xi$ is of the form $\xi = (\xi_1, \xi_2)$ with $0 \leqq \xi_1 \leqq 1$, $\xi_1 + \xi_2 = 1$. Suppose that the uncertainty function U is given by

$$U(\xi) = \min \{\xi_1, \xi_2\}, \qquad \xi \in \Xi. \tag{5.1}$$

(This uncertainty function is appropriate if there are two terminal actions and errors of each kind are equally costly.) Thus if sampling terminates as soon as $U(\xi(X_1, \cdots, X_n)) \leqq \epsilon$, for some fixed value of ϵ, $0 < \epsilon < \frac{1}{2}$, then sampling terminates as soon as either $\xi_1(X_1, \cdots, X_n) \leqq \epsilon$ or $\xi_1(X_1, \cdots, X_n) \geqq \bar{\epsilon}$, where $\bar{\epsilon} = 1 - \epsilon$.

Suppose that the class $\mathcal{C}$ from which a choice must be made at each stage of experimentation contains only the two experiments X and Y defined as follows.

If $\theta = \theta_1$, X is uniformly distributed on the interval $[0, 1]$; if $\theta = \theta_2$, X is uniformly distributed on the interval $[1 - \alpha, 2 - \alpha]$, where α is a given constant, $0 < \alpha < 1$. Thus, the conditional density functions of X (with respect to Lebesgue measure) are

$$f_1(x) = \begin{cases} 1, & 0 \leqq x \leqq 1, \\ 0, & \text{otherwise}; \end{cases}$$
$$f_2(x) = \begin{cases} 1, & 1 - \alpha \leqq x \leqq 2 - \alpha, \\ 0, & \text{otherwise}. \end{cases} \tag{5.2}$$

Y takes only the values 0 and 1, and $p_i(y)$, the probability that $Y = y$ when $\theta = \theta_i (y = 0, 1; i = 1, 2)$ is given by

$$p_1(0) = \epsilon, \quad p_1(1) = \bar{\epsilon}; \qquad p_2(0) = \bar{\epsilon}, \quad p_2(1) = \epsilon. \tag{5.3}$$

We first investigate the experiment X. It is easy to see that for any prior probability ξ_1 and any observed value x of X, the posterior probability $\xi_1(x)$ is

$$\xi_1(x) = \begin{cases} 1 & \text{if } 0 \leqq x < 1 - \alpha, \\ \xi_1 & \text{if } 1 - \alpha \leqq x \leqq 1, \\ 0 & \text{if } 1 < x \leqq 2 - \alpha. \end{cases} \tag{5.4}$$

Let S_X denote the sampling plan whereby the experiment X is replicated re-

peatedly. It follows from (5.4) that under the plan S_X observations on X are taken as long as each observation falls in the interval $1 - \alpha \leqq x \leqq 1$. But $\Pr\{1 - \alpha \leqq X \leqq 1 \mid \xi\} = \alpha$ for all $\xi \varepsilon \Xi$. Let Ξ_0 be the subset of Ξ containing all ξ such that $\epsilon < \xi_1 < \bar{\epsilon}$. Then for any prior distribution $\xi \varepsilon \Xi_0$ (and these are the only prior distributions that need be considered since, otherwise, no sampling is needed), $\Pr\{N(S_X) > k \mid \xi\} = \alpha^k$, $k = 0, 1, \cdots$. Hence

$$E(N(S_X) \mid \xi) = 1/(1 - \alpha), \qquad \xi \varepsilon \Xi_0. \tag{5.5}$$

Furthermore, it follows from (5.1) and (5.4) that

$$E[U(\xi(X)) \mid \xi] = \alpha U(\xi) = \alpha \min\{\xi_1, \xi_2\}, \qquad \xi \varepsilon \Xi. \tag{5.6}$$

Now let us consider the experiment Y. For any prior probability ξ_1, let $\xi_1(y)\,(y = 0, 1)$ denote the posterior probability after having observed $Y = y$, and let $\xi_1(y_1, y_2) = \xi_1(y_1)(y_2)$ $(y_1 = 0, 1; y_2 = 0, 1)$; i.e., $\xi_1(y_1, y_2)$ is the posterior probability after having observed two values of Y. It is readily computed that for any $\xi \varepsilon \Xi$,

$$\begin{aligned} \xi_1(0) &= \epsilon\xi_1/(\epsilon\xi_1 + \bar{\epsilon}\xi_2), \\ \xi_1(1) &= \bar{\epsilon}\xi_1/(\bar{\epsilon}\xi_1 + \epsilon\xi_2), \end{aligned} \tag{5.7}$$

and

$$\begin{aligned} \xi_1(0, 0) &= \epsilon^2\xi_1/(\epsilon^2\xi_1 + \bar{\epsilon}^2\xi_2), \\ \xi_1(1, 1) &= \bar{\epsilon}^2\xi_1/(\bar{\epsilon}^2\xi_1 + \epsilon^2\xi_2), \\ \xi_1(0, 1) &= \xi_1(1, 0) = \xi_1. \end{aligned} \tag{5.8}$$

Let S_Y denote the sampling plan whereby the experiment Y is replicated repeatedly. Suppose that the prior probability $\xi_1 = \frac{1}{2}$. Then $\xi_1(0) = \epsilon$ and $\xi_1(1) = \bar{\epsilon}$. Thus, under the plan S_Y sampling always terminates after the first observation.

Suppose that $\epsilon < \xi_1 < \frac{1}{2}$. The following relations are all easily derived from (5.7) and (5.8): $\xi_1(0) < \epsilon$, $\epsilon < \xi_1(1) < \bar{\epsilon}$, $\xi_1(1, 1) \geqq \bar{\epsilon}$, $\epsilon < \xi_1(1, 0) < \frac{1}{2}$. Together these relations imply that under the plan S_Y, sampling continues as long as the observations follow the pattern $1, 0, 1, 0, 1, 0, \cdots$ and sampling ceases as soon as the pattern is violated. Hence for $\epsilon < \xi_1 < \frac{1}{2}$,

$$\begin{aligned} \Pr\{N(S_Y) > 2k + 1 \mid \xi\} &= \bar{\epsilon}^{k+1}\epsilon^k\xi_1 + \epsilon^{k+1}\bar{\epsilon}^k\xi_2 \\ &= (\epsilon\bar{\epsilon})^k(\bar{\epsilon}\xi_1 + \epsilon\xi_2), \qquad k = 0, 1, \cdots; \\ \Pr\{N(S_Y) > 2k \mid \xi\} &= \bar{\epsilon}^k\epsilon^k, \qquad k = 0, 1, \cdots. \end{aligned} \tag{5.9}$$

It follows from a simple computation that

$$\begin{aligned} E(N(S_Y) \mid \xi) &= \sum_{i=0}^{\infty} \Pr\{N(S_Y) > i \mid \xi\} \\ &= (1 + \bar{\epsilon}\xi_1 + \epsilon\xi_2)/(1 - \epsilon\bar{\epsilon}), \qquad \epsilon < \xi_1 < \tfrac{1}{2}. \end{aligned} \tag{5.10}$$

Suppose that $\frac{1}{2} < \xi_1 < \bar{\epsilon}$. In this case, sampling continues as long as the observations follow the pattern 0, 1, 0, 1, 0, 1, $\cdots$ and sampling ceases as soon as the pattern is violated. A computation similar to the one just given shows that

$$E(N(S_Y) \mid \xi) = (1 + \epsilon\xi_1 + \bar{\epsilon}\xi_2)/(1 - \epsilon\bar{\epsilon}), \qquad \tfrac{1}{2} < \xi_1 < \bar{\epsilon}. \tag{5.11}$$

Equations (5.10) and (5.11), together with the fact that $E(N(S_Y) \mid \xi) = 1$ when $\xi_1 = \frac{1}{2}$, completely describe the function $E(N(S_Y) \mid \xi)$ for all $\xi \in \Xi_0$.

Furthermore, a straightforward computation using (5.1), (5.3), and (5.7) shows that

$$E[U(\xi(Y)) \mid \xi] = \begin{cases} \xi_1 & \text{if } 0 \leqq \xi_1 \leqq \epsilon, \\ \epsilon & \text{if } \epsilon \leqq \xi_1 \leqq \bar{\epsilon}, \\ \xi_2 & \text{if } \bar{\epsilon} \leqq \xi_1 \leqq 1. \end{cases} \tag{5.12}$$

Let us now compare the experiments X and Y. It follows from (5.10), (5.11) and the comment following (5.11), that $\sup_{\xi \in \Xi_0} E(N(S_Y) \mid \xi) = 3/[2(1 - \epsilon\bar{\epsilon})]$. From (5.5), $E(N(S_X) \mid \xi) = 1/(1 - \alpha)$ for all $\xi \in \Xi_0$, and it follows that if α and ϵ satisfy the inequality

$$3/[2(1 - \epsilon\bar{\epsilon})] \leqq 1/(1 - \alpha), \tag{5.13}$$

then

$$E(N(S_Y) \mid \xi) < E(N(S_X) \mid \xi), \qquad \xi \in \Xi_0. \tag{5.14}$$

That is, if (5.13) holds, then the sequential plan S_Y yields universally (in ξ) smaller expected sample size than the plan S_X. In fact, if (5.13) holds then the plan S_Y yields universally smaller expected sample size than any other sequential plan that specifies the observation of either X or Y at each stage of experimentation. This can be seen as follows. Suppose that for some prior distribution $\xi \in \Xi_0$, the optimal plan specified the observation of X at some stage. Then, by (5.4), after having observed X either sampling ceases or else the posterior distribution is precisely what it was before X was observed. In this case, the optimal plan must again specify the observation of X. In other words, if at some stage the optimal plan specifies the observation of X, then from that stage on the plan must specify repeated observation of X until sampling ceases. But it follows from (5.14) that, from that stage on, repeated observation on Y yields a smaller expected sample size. Hence, the optimal sequential plan never specifies the observation of X; i.e., the optimal plan is S_Y.

A comparison of (5.6) and (5.12) shows that if α and ϵ satisfy the inequality

$$\alpha/2 \leqq \epsilon, \tag{5.15}$$

then

$$E[U(\xi(X)) \mid \xi] \leqq E[U(\xi(Y)) \mid \xi], \qquad \xi \in \Xi. \tag{5.16}$$

It follows from Theorem 4.2 that if (5.15) holds, then for any fixed number n

of experiments and any prior distribution $\xi \in \Xi$, the expected uncertainty after n experiments is minimized by taking n observations of X.

The intriguing feature of this example is that there are values of α and ϵ that simultaneously satisfy both (5.13) and (5.15) (e.g., the values $\alpha = \frac{1}{2}$, $\epsilon = \frac{1}{4}$). For such values it is true that for any fixed number of experiments the expected uncertainty is minimized by repeated observation of X, whereas if sampling is continued until the uncertainty is reduced to ϵ, the expected sample size is minimized by repeated observation of Y.

The following property of this example is also of interest.

Suppose that only the experiment X is available to the experimenter and he wishes to take observations on X until the uncertainty is reduced to ϵ or below. In other words, by (5.1), he wishes to take observations until he arrives at a posterior distribution $\xi = (\xi_1, \xi_2)$ for which the likelihood ratio ξ_2/ξ_1 satisfies either $\xi_2/\xi_1 \geqq \bar{\epsilon}/\epsilon$ or $\xi_2/\xi_1 \leqq \epsilon/\bar{\epsilon}$. (The sampling plan is a Wald sequential probability ratio test.)

Suppose that $\alpha = \frac{1}{2}$ and $\epsilon = \frac{1}{4}$. Define the random variable Z in terms of X as follows:

$$Z = \begin{cases} 0 & \text{if } X \geqq \frac{3}{4}, \\ 1 & \text{if } X < \frac{3}{4}. \end{cases} \tag{5.17}$$

It is readily checked that Z has the same conditional distribution, (5.3), as Y given each value of the parameter θ. It then follows from the above discussion that if instead of computing the likelihood ratio at each stage from the observed values of X, it is computed from the corresponding values of Z and the distributions (5.3), the expected sample size is reduced.

It is somewhat surprising that by means of a transformation of the values of X, the expected sample size needed to reach the boundaries can be lowered. There is, however, a satisfactory intuitive explanation. With each observation on X, the likelihood ratio either jumps outside of the boundaries (to 0 or ∞) or else it remains the same as it was before the observation. With each observation on Z, the likelihood ratio moves by a fixed factor toward one of the boundaries. It is plausible that a boundary is reached more quickly through the constant jumps resulting from the Z's than through the "all or nothing" jumps resulting from the X's.

REFERENCES

[1] Albert, Arthur E. (1961). The sequential design of experiments for infinitely many states of nature. *Ann. Math. Statist.* **32** 774–799.

[2] Bellman, Richard (1957). *Dynamic Programming*. Princeton Univ. Press.

[3] Blackwell, David (1951). Comparison of experiments, 93–102. *Proc. Second Berkeley Symp. Math. Statist. Prob.* Univ. of California Press.

[4] Blackwell, David (1953). Equivalent comparisons of experiments. *Ann. Math. Statist.* **24** 265–272.

[5] Blackwell, David and Girshick, M. A. (1954). *Theory of Games and Statistical Decisions*. Wiley, New York.

[6] BRADT, RUSSELL N. AND KARLIN, SAMUEL (1956). On the design and comparison of certain dichotomous experiments. *Ann. Math. Statist.* **27** 390–409.

[7] CHERNOFF, HERMAN (1959). Sequential design of experiments. *Ann. Math. Statist.* **30** 755–770.

[8] LINDLEY, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27** 986–1005.

[9] LINDLEY, D. V. (1957). Binomial sampling schemes and the concept of information. *Biometrika.* **44** 179–186.

[10] SCHWARZ, GIDEON (1960). Asymptotic shapes of optimal sampling regions in sequential testing. unpubl. memo CU-20-60-Nonr-266(59)MS, Dept. of Math. Statist., Columbia University.