ON A MEASURE OF THE INFORMATION PROVIDED BY AN EXPERIMENT[1, 2]

By D. V. Lindley

University of Cambridge and University of Chicago

1. Summary. A measure is introduced of the information provided by an experiment. The measure is derived from the work of Shannon [10] and involves the knowledge prior to performing the experiment, expressed through a prior probability distribution over the parameter space. The measure is used to compare some pairs of experiments without reference to prior distributions; this method of comparison is contrasted with the methods discussed by Blackwell. Finally, the measure is applied to provide a solution to some problems of experimental design, where the object of experimentation is not to reach decisions but rather to gain knowledge about the world.

2. Introduction. Shannon has introduced two important ideas into the theory of information in communications engineering. The first idea is that information is a statistical concept. The statistical frequency distribution of the symbols that make up a message must be considered before the notion can be discussed adequately. The second idea springs from the first and implies that on the basis of the frequency distribution, there is an essentially unique function of the distribution which measures the amount of the information. It is the purpose of the present paper to apply these two ideas to statistical theory by discussing the notion of information in an experiment, rather than in a message. The second of Shannon's ideas has been applied to statistical theory by Kullback and Leibler [6], [7], [8]; but our application is quite distinct from theirs. The interpretation of Shannon's ideas in current statistical theory has been given by McMillan [9]. The discussion in that paper is related to, and partly inspired, that given here. A referee has kindly pointed out that Shannon's theory has been applied in psychometric problems by L. J. Cronbach in an unpublished report [14]. Definition 2, in particular, is used by Cronbach.

The situation in communications engineering is that there is a transmitted message, x, which is received as a message, y. By considerations of the informations in x and y it is possible to discuss the rate at which information has been transmitted along the channel. The analogous description in statistical theory is provided by replacing x by the knowledge of the state of nature, usually expressed by the knowledge of a finite number of parameters, prior to an experiment, and by replacing y by the knowledge after the experiment. The com-

Received August 2, 1955.

[1] Research carried out at the Statistical Research Center, University of Chicago, under sponsorship of the Office of Naval Research.

[2] This is a revised version of papers presented at the Chapel Hill and Berkeley meetings of the Institute of Mathematical Statistics in April and July, 1955.

parison of the knowledge before and after the experiment makes it possible to discuss the amount of information provided by the experiment. The average of this, for fixed prior knowledge, determines the average amount of information. The measure of information is given by Shannon's function. But, just as it is essential to consider the statistical character of the message x, so it is necessary to consider the statistical character of the knowledge of the state of nature. Prior probability distributions are therefore basic to the study. It seems obvious to the author that prior distributions, though usually anathema to the statistician, are essential to the notion of experimental information. To take an extreme case, if the prior distribution is concentrated on a single parameter value, that is, if the state of nature is known, then no experiment can be informative.

It may happen that, whatever the prior knowledge, one experiment is more informative than another. We shall meet such examples below. In this case it is possible to compare the two experiments absolutely, without reference to prior knowledge. Methods of comparing experiments have been suggested by Bohnenblust, Shapley, and Sherman (described by Blackwell in [2]) and by Blackwell [2]. These methods of comparison are contrasted with the one presented here, and it is shown that if one experiment is more informative than another by Blackwell's criterion, then it is also true of that used here; the converse is false.

The Bohnenblust method of comparison is formulated in decision theory language and involves considerations of losses. These notions are not used here; the concepts used are perhaps more related to the inference problem than to the decision problem (see Barnard [1]). In this paper it is suggested that although indisputably one purpose of experimentation is to reach decisions, another purpose is to gain knowledge about the state of nature (that is, about the parameters) without having specific actions in mind. The knowledge is measured by the amount of information, as described above. The following rule of experimentation is therefore suggested: perform that experiment for which the expected gain in information is the greatest, and continue experimentation until a preassigned amount of information has been attained. The consequences of this rule are explored and shown, for example, to lead to sequential probability ratio tests. Binomial and normal sampling are also considered as special cases.

3. The experiment will result in an observation, x, belonging to a space, $\mathbf{X}$. The space $\mathbf{X}$ has a σ-field, $\mathfrak{B}$, of subsets, X. For every θ belonging to a space Θ is defined a probability measure on $\mathfrak{B}$. We shall suppose that as θ ranges through Θ the probability measures on $\mathfrak{B}$ are all absolutely continuous with respect to a fixed measure on $\mathfrak{B}$. This permits us to describe each probability measure by a probability density function $p(x \mid \theta)$, such that the probability measure of a subset X is given by $\int_X p(x \mid \theta)\, dx$, where, for simplicity of notation, we have denoted integration with respect to the dominating measure by

dx. The ordered quadruple[3] $\mathcal{E} = \{\mathbf{X}, \mathfrak{B}, \Theta, P\}$, where P is the set of $p(x \mid \theta)$, characterizes an experiment, $\mathcal{E}$. Again, for simplicity in notation, we shall not distinguish between random variables and the values assumed by them, nor shall we attempt to be specific in describing the density functions. Thus, $p(x)$ will denote the density function of the random variable x; similarly, $p(\theta)$ will denote the density function of θ, without any suggestion that the random variables x and θ have the same density. These devices avoid such clumsy notation as $p_x(y)$ for the density of the random variable x when x assumes the numerical value y.

We shall suppose that Θ is endowed with a σ-field of subsets; usually, Θ will be a subset of n-dimensional Euclidean space and the σ-field will be the Borel field. A prior distribution for θ will be a probability measure on this field, and again we shall suppose it to be described by a probability density function $p(\theta)$ with respect to a measure denoted by $d\theta$. Thus, in accord with the notational conventions described above, we have, for example,

$$p(x) = \int_\Theta p(x \mid \theta)p(\theta)\, d\theta, \tag{1}$$

and Bayes' theorem reads

$$p(\theta \mid x) = p(x \mid \theta)p(\theta)/p(x). \tag{2}$$

The ranges of integration in the following formulas will always be the whole space, either $\mathbf{X}$ or Θ, and will be omitted.

For a prior distribution $p(\theta)$, the amount of information with respect to $d\theta$ is defined to be

$$\mathcal{I}_0 = \int p(\theta) \log p(\theta)\, d\theta \tag{3}$$

whenever the integral exists. For any θ for which $p(\theta) = 0$, define $p(\theta) \log p(\theta)$ to be zero. A useful alternative notation is

$$\mathcal{I}_0 = E_\theta \log p(\theta), \tag{4}$$

where E_θ denotes the expectation operator with respect to θ.

The reasons for the introduction of this function have been given by Shannon. Translated into the language of experimentation, the basic reason is this: Consider the case where Θ is finite; then the amount of information, I, in a prior distribution can be measured by how much information it is necessary to provide before the value of θ is known. This latter information could be provided in two stages. For the first, let Θ_1 be a non-empty proper subset of Θ with $P = \int_{\Theta_1} p(\theta)\, d\theta \neq 0$ or 1, and suppose the experimenter is told whether $\theta \in \Theta_1$ or its complement. This provides amount I_1, say; the prior distribution being $(P, 1 - P)$. In the second stage, suppose the experimenter is told the value of

[3] Strictly, the quadruple should be a quintuple and should include the dominating measure; for convenience, it will be omitted.

θ; the information provided is I_2 or I_3, say, according as he knew $\theta \varepsilon \Theta_1$ or its complement. (The necessary distributions are $p(\theta)/P$ and $p(\theta)/(1 - P)$, respectively.) Then Shannon requires that the information provided in the first stage and the average amount provided in the second stage add up to the total information; that is,

$$I = I_1 + PI_2 + (1 - P)I_3 .$$

This additivity requirement is the fundamental postulate. It finds its general form in Theorem 2, below. Shannon then shows ([10], Appendix 2) that $I = \sum p(\theta) \log p(\theta)$, apart from an arbitary multiplying constant, is the only function having this property together with a mild continuity property.

We note that the amount of information, so defined, is not invariant under a change of description of the parameter space. This lack of invariance need cause no concern, as it will disappear when the expression is used to define the average information in the experiment. The minus sign introduced by Shannon in front of the integral is not used. The reason for this is as follows: the maximum information, in a statistician's sense, will be obtained when the probability distribution is concentrated on a single value of θ, and the information will be reduced as the distribution of θ "spreads"; this is exactly the reverse of the situation faced by a communications engineer, where the concentration on a single value would allow no choice in his messages. The two scales are therefore reversed.

After the experiment has been performed and the value x observed, the posterior distribution of θ is $p(\theta \mid x)$, given by (2), and the amount of information is

$$\mathcal{I}_1(x) = \int p(\theta \mid x) \log p(\theta \mid x)\, d\theta. \tag{5}$$

(If $p(\theta \mid x) = 0$, define the integrand to be zero.)

DEFINITION 1. The amount of information provided by the experiment $\mathcal{E}$, with prior knowledge $p(\theta)$, when the observation is x, is

$$\mathcal{I}(\mathcal{E}, p(\theta), x) = \mathcal{I}_1(x) - \mathcal{I}_0 . \tag{6}$$

This expression is also not invariant under a change of description of the parameter space.

The quantity $\mathcal{I}(\mathcal{E}, p(\theta), x)$ depends on x; some results are more informative than others. However, since θ is regarded as a random variable, this quantity may be averaged with respect to x according to the probability density given by (1). Hence, we have

DEFINITION 2. The average amount of information provided by the experiment $\mathcal{E}$, with prior knowledge $p(\theta)$, is

$$\mathcal{I}(\mathcal{E}, p(\theta)) = E_x[\mathcal{I}_1(x) - \mathcal{I}_0]. \tag{7}$$

Alternative forms for $g(\mathcal{E}, p(\theta))$ are

$$E_x E_\theta \log \{p(\theta \mid x)/p(\theta)\} \qquad \text{(from (3) and (5))}, \tag{8}$$

$$E_x E_\theta \log \{p(x \mid \theta)/p(x)\} \qquad \text{(from (2))}, \tag{9}$$

and, in full, if $p(x, \theta)$ is the joint density for x and θ,

$$\iint p(x, \theta) \log \{p(x, \theta)/p(x)p(\theta)\}\, dx\, d\theta. \tag{10}$$

The expression (10) shows the symmetry between x and θ and also exhibits the fact that $g(\mathcal{E}, p(\theta))$ is invariant under a $1 - 1$ transformation of the parameter space, Θ. The expression occurs in Shannon's theory ([10], Section 24) for the rate of transmission of information along a channel.[4]

Yet another expression for $g(\mathcal{E}, p(\theta))$ which is useful in calculation is obtained by introducing the information operator, I, along with the expectation operator, E. For a density function $p(y)$, we define

$$I_y p(y) = \int p(y) \log p(y)\, dy.$$

It is easy to verify that

$$g(\mathcal{E}, p(\theta)) = E_\theta I_x p(x \mid \theta) - I_x E_\theta p(x \mid \theta). \tag{11}$$

4. The results that we now proceed to establish involve only the use of Bayes' theorem and the two facts that the logarithm of a product is the sum of the two logarithms (in the combination of equations (12) and (13) for example) and that the function $x \log x$ is convex (in Theorem 1). We shall often denote the average information by $g(\mathcal{E})$ when the particular prior distribution does not have to be stressed.

THEOREM 1. $g(\mathcal{E}) \geqq 0$, *with equality if, and only if,* $p(x \mid \theta)$ *does not depend on* Θ, *except possibly in a null set for* θ.

This follows immediately from a well-known inequality (see, for example, Hardy, Littlewood, and Pólya [5], Theorem 205) on writing

$$g(\mathcal{E}) = \iint f(x, \theta) \log f(x, \theta) \cdot p(x)p(\theta)\, dx\, d\theta,$$

where

$$f(x, \theta) = p(x, \theta)/p(x)p(\theta).$$

The inequality says that

$$g(\mathcal{E}) \geqq \iint f(x, \theta)p(x)p(\theta)\, dx\, d\theta \cdot \log \left\{ \frac{\iint f(x, \theta)p(x)p(\theta)\, dx\, d\theta}{\iint p(x)p(\theta)\, dx\, d\theta} \right\}$$

[4] In the particular case of the "experiment" involved in radar work, the above ideas are already contained in a paper by P. M. Woodward [12], and are repeated in [13]. The author is indebted to M. S. Bartlett for these references.

with equality if, and only if, $f(x, \theta)$ equals a constant, except possibly on a null set. The logarithm is zero.

The theorem says that, provided the density of x varies with θ, any experiment is informative, on the average. Note that $\mathcal{I}(\mathcal{E}, p(\theta), x)$ is not necessarily nonnegative. Although the expectation is positive, the experimental result may reduce the amount of information. This can happen when a "surprising" value of x occurs; granted the correctness of the experimental technique, the "surprise" may result in our being less sure about θ than before the experiment.

Suppose that the observations x in an experiment $\mathcal{E}$ consist of a pair of observations x_1, x_2. That is, every $x \in \mathbf{X}$ is an ordered pair (x_1, x_2) with $x_i \in \mathbf{X}_i$ $(i = 1, 2)$. Let $\mathfrak{B}_i$ be the σ-field over $\mathbf{X}_i$ induced from $\mathfrak{B}$ by the transformation $x_i = x_i(x)$, and let P_i be the set of probability densities $p(x_i \mid \theta)$ of the observations x_i $(i = 1, 2)$. (It is again supposed that the measures are, for all θ, dominated by a measure so that the probability distributions can be characterized by densities.) Then, $\mathcal{E}_i = \{\mathbf{X}_i, \mathfrak{B}_i, \Theta, P_i\}$ $(i = 1, 2)$ are two experiments and $\mathcal{E}$ is said to be the sum of the experiments $\mathcal{E}_1$ and $\mathcal{E}_2$, written $\mathcal{E} = (\mathcal{E}_1, \mathcal{E}_2)$. We shall also have to consider the experiment $\mathcal{E}_2(x_1) = \{\mathbf{X}_2, \mathfrak{B}_2, \Theta, P_2(x_1)\}$, where $P_2(x_1)$ is the set of densities $p(x_2 \mid \theta, x_1)$.

Consider $\mathcal{I}(\mathcal{E}_2(x_1), p(\theta \mid x_1))$. Since $p(\theta \mid x_1)$ is the posterior distribution of θ after x_1 has been observed, this quantity is the average information provided by an observation on x_2 after $\mathcal{E}_1$ has been performed and x_1 observed. The average of it over x_1 is defined to be the average information provided by $\mathcal{E}_2$ after $\mathcal{E}_1$ has been performed. We write it $\mathcal{I}(\mathcal{E}_2 \mid \mathcal{E}_1)$, again supressing $p(\theta)$. A proof along the lines of that for Theorem 1 establishes that $\mathcal{I}(\mathcal{E}_2 \mid \mathcal{E}_1) \geqq 0$, with equality if, and only if, $p(x_2 \mid \theta, x_1)$ does not involve θ, except possibly on a null set.

THEOREM 2. $\mathcal{I}(\mathcal{E}_1) + \mathcal{I}(\mathcal{E}_2 \mid \mathcal{E}_1) = \mathcal{I}(\mathcal{E})$.

We have, using the form (9),

$$\begin{aligned} \mathcal{I}(\mathcal{E}_1) &= E_{x_1} E_\theta \log \{p(x_1 \mid \theta)/p(x_1)\} \\ &= E_{x_1} E_{x_2} E_\theta \log \{p(x_1 \mid \theta)/p(x_1)\}. \end{aligned} \tag{12}$$

Also, from the definitions immediately before the statement of the theorem,

$$\begin{aligned} \mathcal{I}(\mathcal{E}_2 \mid \mathcal{E}_1) &= E_{x_1}[\mathcal{I}(\mathcal{E}_2(x_1), p(\theta \mid x_1))] \\ &= E_{x_1} E_{x_2} E_\theta \log \{p(x_2 \mid \theta, x_1)/p(x_2 \mid x_1)\}. \end{aligned} \tag{13}$$

Addition of (12) and (13) gives

$$E_{x_1} E_{x_2} E_\theta \log \left\{ \frac{p(x_2 \mid \theta, x_1) p(x_1 \mid \theta)}{p(x_2 \mid x_1) p(x_1)} \right\} = E_{x_1} E_{x_2} E_\theta \log \left\{ \frac{p(x_1, x_2 \mid \theta)}{p(x_1, x_2)} \right\},$$

which is $\mathcal{I}(\mathcal{E})$, and the theorem is proved.

COROLLARY. *If x_1 is sufficient for x in the Neyman-Fisher sense, then $\mathcal{I}(\mathcal{E}_1) = \mathcal{I}(\mathcal{E})$.*

For if x_1 is sufficient for x, the factorization theorem shows that $p(x_2 \mid \theta, x_1)$

does not involve θ. Hence, by the remark immediately before the statement of the theorem, $g(\mathcal{E}_2 \mid \mathcal{E}_1) = 0$, and the corollary is established.

The corollary establishes that there is no loss in information if attention is confined to observation on a sufficient statistic. Conversely, if a statistic is considered which is not sufficient (in the sense that it does not satisfy the factorization theorem), then information will be lost since $g(\mathcal{E}_2 \mid \mathcal{E}_1) > 0$. Theorem 2 generalizes to a finite number of experiments with common Θ in an obvious manner.

DEFINITION 3. Two experiments, $\mathcal{E}_1$ and $\mathcal{E}_2$, with $\Theta_1 = \Theta_2 = \Theta$, are independent if $p(x_1, x_2 \mid \theta) = p(x_1 \mid \theta)p(x_2 \mid \theta)$ for all $\theta \in \Theta$.

Of course it by no means follows that if $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent, then x_1 and x_2 are independent; i.e., it is not usually true that $p(x_1, x_2) = p(x_1)p(x_2)$.

If $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent, the experiments $\mathcal{E}_2(x_1)$ and $\mathcal{E}_2$, defined above, are equivalent (in the sense that the four pairs of defining elements are all equal when we write $\mathcal{E}_1 \equiv \mathcal{E}_2$), and we have the result

$$g(\mathcal{E}_2 \mid \mathcal{E}_1) = E_{x_1} g(\mathcal{E}_2, p(\theta \mid x_1)). \tag{14}$$

THEOREM 3. *If $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent*

$$g(\mathcal{E}_2 \mid \mathcal{E}_1) \leqq g(\mathcal{E}_2),$$

with equality if, and only if, x_1 and x_2 are independent.

From (13) and the independence, we have

$$\begin{aligned} g(\mathcal{E}_2) - g(\mathcal{E}_2 \mid \mathcal{E}_1) &= E_{x_2} E_\theta \log \{p(x_2 \mid \theta)/p(x_2)\} \\ &\quad - E_{x_1} E_{x_2} E_\theta \log \{p(x_2 \mid \theta)/p(x_2 \mid x_1)\} \\ &= E_{x_1} E_{x_2} E_\theta \log \{p(x_2 \mid x_1)/p(x_2)\} \\ &= E_{x_1} E_{x_2} \log \{p(x_2 \mid x_1)/p(x_2)\}. \end{aligned}$$

The last expression is identical with (9) when x_2, x_1 are replaced by x, θ, respectively. By Theorem 1 it is therefore nonnegative, and is zero if, and only if, $p(x_2 \mid x_1) = p(x_2)$.

Again, the definition and theorem could be generalized to any finite number of independent experiments. The theorem says that if $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent experiments, either one is more informative, on the average, if performed first than if performed second. In particular, if $\mathcal{E}_1 \equiv \mathcal{E}_2$, the theorem says that an independent repeat of the same experiment is less informative, on the average, than the original experiment. This is a property which agrees with the common belief in the diminishing marginal utility of independent equidistributed observations.

COROLLARY. *If $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent experiments, then*

$$g(\mathcal{E}_1) + g(\mathcal{E}_2) \geqq g(\mathcal{E}),$$

with equality if, and only if, x_1 and x_2 are independent.

For

$$g(\mathcal{E}_1) + g(\mathcal{E}_2) \geqq g(\mathcal{E}_1) + g(\mathcal{E}_2 \mid \mathcal{E}_1) \qquad \text{(by the theorem)}$$
$$= g(\mathcal{E}) \qquad \text{(by Theorem 2).}$$

The corollary is not necessarily true for experiments which are not independent. It is easy to construct an example where $\mathcal{E}_1$ or $\mathcal{E}_2$ separately provide no information, but jointly they are completely informative in the sense that the posterior distribution is necessarily concentrated on a single value of θ.

In the case of repetition of identical experiments, more than the result of Theorem 3 can be said about the reduction of information on repetition. Let $\mathcal{E}^{(1)} \equiv \mathcal{E}_1$ be any experiment and let $\mathcal{E}_2, \mathcal{E}_3, \cdots$ be independent identical experiments. Let $\mathcal{E}^{(2)} = (\mathcal{E}_1, \mathcal{E}_2)$ and generally $\mathcal{E}^{(n)} = (\mathcal{E}_n, \mathcal{E}^{(n-1)})$. Let $g(\mathcal{E}^{(n)}) = j_n$; the prior distribution can remain unspecified.

THEOREM 4. *j_n is a concave, increasing function of n.*

It will be enough to establish that

$$0 \leqq j_{n+1} - j_n \leqq j_n - j_{n-1}.$$

The first inequality follows from Theorem 2, for by that theorem

$$j_{n+1} - j_n = g(\mathcal{E}_{n+1} \mid \mathcal{E}^{(n)}) \geqq 0.$$

The second reads:

$$g(\mathcal{E}_{n+1} \mid \mathcal{E}^{(n)}) \leqq g(\mathcal{E}_n \mid \mathcal{E}^{(n-1)}).$$

Since $\mathcal{E}_n \equiv \mathcal{E}_{n+1}$, it will be enough to show that

$$g(\mathcal{E}_{n+1} \mid \mathcal{E}^{(n-1)}, \mathcal{E}_n) \leqq g(\mathcal{E}_{n+1} \mid \mathcal{E}^{(n-1)}).$$

This follows as a slight generalization of Theorem 3, saying that the additional experiment $\mathcal{E}_n$ reduces the average information provided by $\mathcal{E}_{n+1}$, even after $\mathcal{E}^{(n-1)}$.

Consider the following experiment: With probability λ (for all values of θ), perform experiment $\mathcal{E}_1$; with probability $1 - \lambda$ (for all values of θ), perform $\mathcal{E}_2$, where $\mathcal{E}_1$ and $\mathcal{E}_2$ have $\Theta_1 = \Theta_2 = \Theta$. The observation will consist in the observation obtained, according to whichever experiment is performed, and the knowledge of which experiment was performed. Denote this experiment by $(\lambda\mathcal{E}_1 + (1 - \lambda)\mathcal{E}_2)$. In mathematical terms $(\lambda\mathcal{E}_1 + (1 - \lambda)\mathcal{E}_2) = \{\mathbf{X} = \mathbf{X}_1 \cup \mathbf{X}_2, \mathfrak{B} = \mathfrak{B}_1 \cup \mathfrak{B}_2, \Theta, P\}$, where P is the set of densities, $p(x \mid \theta)$, defined as follows: If $x \in \mathbf{X}_1$, then $p(x \mid \theta) = \lambda p(x_1 \mid \theta)$ with $x = x_1$; if $x \in \mathbf{X}_2$, then $p(x \mid \theta) = (1 - \lambda)p(x_2 \mid \theta)$ with $x = x_2$. It is easy to verify that

$$g(\lambda\mathcal{E}_1 + (1 - \lambda)\mathcal{E}_2) = \lambda g(\mathcal{E}_1) + (1 - \lambda)g(\mathcal{E}_2). \tag{15}$$

In this terminology the concavity property established in Theorem 4 says that

$$g(\lambda\mathcal{E}^{(k)} + (1 - \lambda)\mathcal{E}^{(m)}) \leqq g(\mathcal{E}^{(n)}),$$

with $n = \lambda k + (1 - \lambda)m$. The last equality ensures that the average "sample sizes" are the same, and the inequality says that rather than "mixing" two sample sizes, it is better to take a sample of fixed "size" equal to the average size of the mixture. We discuss the result again below.

THEOREM 5. *For fixed* $\mathcal{E}$, $g(\mathcal{E}, p(\theta))$ *is a* concave *function of* $p(\theta)$.

We have to show that if $p_1(\theta)$ and $p_2(\theta)$ are two prior probability densities and $0 \leqq \lambda \leqq 1$, then

$$g(\mathcal{E}, \lambda p_1(\theta) + (1 - \lambda)p_2(\theta)) - \lambda g(\mathcal{E}, p_1(\theta)) - (1 - \lambda)g(\mathcal{E}, p_2(\theta)) \geqq 0.$$

The left-hand side is

$$\iint p(x \mid \theta)(\lambda p_1(\theta) + (1 - \lambda)p_2(\theta)) \log \{p(x \mid \theta)/p(x)\}\, dx\, d\theta$$

$$- \lambda \iint p(x \mid \theta)p_1(\theta) \log \{p(x \mid \theta)/p_1(x)\}\, dx\, d\theta$$

$$- (1 - \lambda) \iint p(x \mid \theta)p_2(\theta) \log \{p(x \mid \theta)/p_2(x)\}\, dx\, d\theta,$$

where $p_i(x) = \int p(x \mid \theta)p_i(\theta)\, d\theta$ $(i = 1, 2)$ and $p(x) = \lambda p_1(x) + (1 - \lambda)p_2(x)$. This simplifies to give

$$\lambda \iint p(x \mid \theta)p_1(\theta) \log \{p_1(x)/p(x)\}\, dx\, d\theta$$

$$+ (1 - \lambda) \iint p(x \mid \theta)p_2(\theta) \log \{p_2(x)/p(x)\}\, dx\, d\theta.$$

Performing the integrations with respect to θ, we have

$$\lambda \int p_1(x) \log \{p_1(x)/p(x)\}\, dx + (1 - \lambda) \int p_2(x) \log \{p_2(x)/p(x)\}\, dx,$$

and these integrals are positive by the inequality used to establish Theorem 1.

THEOREM 6. *Let* $\mathcal{E}_i = \{\mathbf{X}, \mathcal{B}, \Theta, P_i\}$ $(i = 1, 2)$. *Let* $\mathcal{E} = \{\mathbf{X}, \mathcal{B}, \Theta, P\}$, *where* P *is the set of densities*

$$p(x \mid \theta) = \lambda p_1(x \mid \theta) + (1 - \lambda)p_2(x \mid \theta),$$

with $0 \leqq \lambda \leqq 1$. *Then*

$$g(\mathcal{E}) \leqq \lambda g(\mathcal{E}_1) + (1 - \lambda)g(\mathcal{E}_2). \tag{16}$$

(An alternative statement of this theorem reads: For fixed $\mathbf{X}$, $\mathcal{B}$, Θ, and $p(\theta)$, $g(\mathcal{E})$ is a *convex* function of P.)

The experiment $\mathcal{E}$, described in the statement of the theorem, can be thought of as being performed as follows: With probability λ, a value x is obtained according to the density $p_1(x \mid \theta)$; with probability $1 - \lambda$, x is obtained according to $p_2(x \mid \theta)$. The experimenter is informed only of x and not of which event,

of probability λ or $1 - \lambda$, took place. Let the experiment $\mathcal{E}^*$, on the other hand, inform him about this event but not about the value of x. Then, clearly, using the notation developed above,

$$(\mathcal{E}, \mathcal{E}^*) \equiv (\lambda\mathcal{E}_1 + (1 - \lambda)\mathcal{E}_2).$$

Hence,

$$g(\mathcal{E}) + g(\mathcal{E}^* \mid \mathcal{E}) = \lambda g(\mathcal{E}_1) + (1 - \lambda)g(\mathcal{E}_2)$$

and the result follows since $g(\mathcal{E}^* \mid \mathcal{E}) \geqq 0$.

Note that we have a convexity property here and a concavity property in the previous theorem.

5. The previous development assigns to an experiment $\mathcal{E}$ and a prior distribution $p(\theta)$ a numerical measure of the average information provided by $\mathcal{E}$. In particular, this permits a comparison to be made between the amounts of information provided by any two experiments $\mathcal{E}_1$, $\mathcal{E}_2$, with the same Θ, with respect to a prior distribution. It also allows $\mathcal{E}_1$ and $\mathcal{E}_2$ to be compared absolutely, that is, without reference to a prior distribution, in certain cases. To do this we introduce

DEFINITION 4. Let $\mathcal{E}_1$, $\mathcal{E}_2$ be two experiments with $\Theta_1 = \Theta_2 = \Theta$. $\mathcal{E}_1$ is more informative than $\mathcal{E}_2$ if

$$g(\mathcal{E}_1, p(\theta)) \geqq g(\mathcal{E}_2, p(\theta)) \tag{17}$$

for all $p(\theta)$,[5] and strict inequality holds for some $p(\theta)$. We write $\mathcal{E}_1 > \mathcal{E}_2$ or $\mathcal{E}_2 < \mathcal{E}_1$. If equality holds in (17) for all $p(\theta)$, we say $\mathcal{E}_1$ and $\mathcal{E}_2$ are equally informative and write $\mathcal{E}_1 = \mathcal{E}_2$. We write $\mathcal{E}_1 \leqq \mathcal{E}_2$, or $\mathcal{E}_2 \geqq \mathcal{E}_1$, to mean either $\mathcal{E}_1 < \mathcal{E}_2$ or $\mathcal{E}_1 = \mathcal{E}_2$.

There exist pairs of experiments for which neither $\mathcal{E}_1 \geqq \mathcal{E}_2$ nor $\mathcal{E}_1 \leqq \mathcal{E}_2$. The merits of such experiments can only be judged by reference to a prior distribution. An example is given in the discussion of the binomial dichotomy after Theorem 9, below.

THEOREM 7. *If $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$ are three experiments with the same Θ and if $\mathcal{E}_3$ is independent of both $\mathcal{E}_1$ and $\mathcal{E}_2$, then $\mathcal{E}_1 > \mathcal{E}_2$ implies $(\mathcal{E}_1, \mathcal{E}_3) > (\mathcal{E}_2, \mathcal{E}_3)$.*

For any $p(\theta)$, by Theorem 2

$$\begin{aligned} g(\mathcal{E}_1, \mathcal{E}_3) &= g(\mathcal{E}_3) + g(\mathcal{E}_1 \mid \mathcal{E}_3) \\ &= g(\mathcal{E}_3) + E_{x_3} g(\mathcal{E}_1, p(\theta \mid x_3)), \end{aligned}$$

by (14), since $\mathcal{E}_1$ and $\mathcal{E}_3$ are independent. But $\mathcal{E}_1 > \mathcal{E}_2$ implies, in particular, that

$$g(\mathcal{E}_1, p(\theta \mid x_3)) \geqq g(\mathcal{E}_2, p(\theta \mid x_3))$$

[5] That is, for all prior distributions not merely for all prior distributions which are dominated by a fixed measure.

for any x_3. Consequently,

$$g(\mathcal{E}_1, \mathcal{E}_3) \geqq g(\mathcal{E}_3) + E_{x_3} g(\mathcal{E}_2, p(\theta \mid x_3))$$
$$= g(\mathcal{E}_3) + g(\mathcal{E}_2 \mid \mathcal{E}_3),$$

again by (14), since $\mathcal{E}_2$ and $\mathcal{E}_3$ are independent. A further application of Theorem 2 establishes the result.

THEOREM 8. *If* $\mathcal{E}_i (i = 1, 2, 3, 4)$ *are four experiments with the same* Θ *and if* $\mathcal{E}_1 > \mathcal{E}_2$, $\mathcal{E}_3 > \mathcal{E}_4$, $\mathcal{E}_1$ *is independent of* $\mathcal{E}_3$, *and* $\mathcal{E}_2$ *of* $\mathcal{E}_4$, *then* $(\mathcal{E}_1, \mathcal{E}_3) > (\mathcal{E}_2, \mathcal{E}_4)$.

Let x_i be the random variable observed in $\mathcal{E}_i$. Then, for any value of θ, x_1 is independent of x_3 and x_2 of x_4. Consider a new set of random variables (y_1, y_2, y_3, y_4), where, for any value of θ, y_i has the same density as x_i, y_1 is independent of y_3, y_2 of y_4, and, in addition, y_2 is independent of y_3. Let $\mathcal{E}_1'$ denote the experiment in which y_i is observed. Clearly, for any $p(\theta)$, $g(\mathcal{E}_i) = g(\mathcal{E}_i')$ and

$$g(\mathcal{E}_1, \mathcal{E}_3) = g(\mathcal{E}_1', \mathcal{E}_3') \geqq g(\mathcal{E}_2', \mathcal{E}_3') \geqq g(\mathcal{E}_2', \mathcal{E}_4') = g(\mathcal{E}_2, \mathcal{E}_4).$$

Both inequalities follow from Theorem 7, and the result is established.

Two other methods of comparing experiments have been introduced. The first, due to Bohnenblust, Shapley, and Sherman, says that $\mathcal{E}_1$ is more informative than $\mathcal{E}_2$ if every loss function attainable with $\mathcal{E}_2$ is also attainable with $\mathcal{E}_1$. The second, due to Blackwell, says that $\mathcal{E}_1$ is sufficient for $\mathcal{E}_2$ if an experimenter performing $\mathcal{E}_1$ can, by a random device, obtain a result equivalent to performing $\mathcal{E}_2$. (For a precise definition of these two relations, see Blackwell [2].) To avoid confusion, we shall speak of the relation introduced here as "more informative (S)" and the relation in terms of loss as "more informative (B)". For the latter, following Blackwell, we write $\mathcal{E}_1 \supset \mathcal{E}_2$, and for $\mathcal{E}_1$ is sufficient for $\mathcal{E}_2$, we write $\mathcal{E}_1 > \mathcal{E}_2$. We remark that Theorems 7 and 8 above are the same as two theorems of Blackwell's (see [4], p. 332) with $>$ replacing $\supset$. We now discuss the connections between these three relations.

THEOREM 9. *If* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *are two experiments with the same* Θ, *and if* $\mathcal{E}_1$ *is sufficient for* $\mathcal{E}_2$, *then* $\mathcal{E}_1$ *is not less informative* (S) *than* $\mathcal{E}_2$. *In other words,* $\mathcal{E}_1 > \mathcal{E}_2$ *implies* $\mathcal{E}_1 \geqq \mathcal{E}_2$.

Let x_i be the random variable observed in $\mathcal{E}_i$ $(i = 1, 2)$. $\mathcal{E}_1 > \mathcal{E}_2$ implies that there exists a stochastic transformation of x_1, say x_2', such that $x_2' \varepsilon \mathbf{X}_2$ and x_2' and x_2 are identically distributed for each $\theta \varepsilon \Theta$. Let $\mathcal{E}_2'$ be the experiment in which x_2' is observed. Clearly, $g(\mathcal{E}_2) = g(\mathcal{E}_2')$. Consider the experiment $\mathcal{E} = (\mathcal{E}_1, \mathcal{E}_2')$. Then, x_1 is sufficient for (x_1, x_2') in the Neyman-Fisher sense, and, hence, by the Corollary to Theorem 2, $g(\mathcal{E}) = g(\mathcal{E}_1)$. But by Theorem 2, $g(\mathcal{E}) \geqq g(\mathcal{E}_2')$; consequently, $g(\mathcal{E}_1) \geqq g(\mathcal{E}_2)$ for all $p(\theta)$, as required.

Conditions are known under which the relations $\supset$ and $>$ are equivalent (see, for example, Blackwell [3]). Under these conditions, it will follow from Theorem 9 that not less informative (B) implies not less informative (S). That the converse of these results is not true can be illustrated by an example.

Consider the case of a binomial dichotomy. Here $\mathbf{X}$ contains two elements (0, 1), Θ contains two elements with $0 \leqq \theta_1 \leqq \theta_2 \leqq 1$ and $p(x = 1 \mid \theta_i) = \theta_i = 1 - p(x = 0 \mid \theta_i)$. This experiment will be denoted by $\mathcal{E}(\theta_1, \theta_2)$. Denote the prior distribution over (θ_1, θ_2) by $(\lambda, 1 - \lambda)$ with $0 \leqq \lambda \leqq 1$. It follows immediately from (11) that

$$(18) \qquad g[\mathcal{E}(\theta_1, \theta_2), \lambda] = S(\lambda\theta_1 + (1 - \lambda)\theta_2) - \lambda S(\theta_1) - (1 - \lambda)S(\theta_2),$$

where

$$S(\theta) = -\theta \log \theta - (1 - \theta)\log (1 - \theta).$$

Consider a fixed experiment $\mathcal{E}(p_1, p_2)$ and compare it with $\mathcal{E}(\theta_1, \theta_2)$ as θ_1 and θ_2 vary. To do this it is necessary to consider the right-hand side of (18) as a function of λ for (p_1, p_2) and for (θ_1, θ_2): $\mathcal{E}(p_1, p_2) \geqq \mathcal{E}(\theta_1, \theta_2)$ if, and only if,

$$g[\mathcal{E}(p_1, p_2), \lambda] \geqq g[\mathcal{E}(\theta_1, \theta_2), \lambda]$$

for all λ. It does not seem possible to describe the results analytically, and we therefore content ourselves with summarizing the results of some computations in the case $p_1 = \frac{1}{4}$, $p_2 = \frac{3}{4}$. The discussion is carried out with reference to Fig. 1, where P is the point $(\frac{1}{4}, \frac{3}{4})$. It is known (see [2]) that the points (θ_1, θ_2) in the areas with horizontal hatching correspond to experiments which can be compared with $\mathcal{E}(p_1, p_2)$ by either the relation $\supset$ or $>$, which are, in this case, identical. For points in the triangular area, $\mathcal{E}(\theta_1, \theta_2) \subset \mathcal{E}(p_1, p_2)$; for points in the quadrilateral, $\mathcal{E}(\theta_1, \theta_2) \supset \mathcal{E}(p_1, p_2)$; the remaining experiments are not comparable with $\mathcal{E}(p_1, p_2)$. Theorem 9 implies that the relation $\supset$ may be replaced by $>$, but computation shows that the points in the areas with vertical hatching correspond to additional experiments which can be compared with $\mathcal{E}(p_1, p_2)$ by the relation $>$. Those adjacent to the triangular area have $\mathcal{E}(\theta_1, \theta_2) < \mathcal{E}(p_1, p_2)$ and those adjacent to the quadrilateral have $\mathcal{E}(\theta_1, \theta_2) > \mathcal{E}(p_1, p_2)$. The points in the unhatched areas correspond to experiments which cannot be compared by the relation $>$. The points in the area of vertical hatching show that the converse of Theorem 9 is false.

The smallness of the unhatched region is a satisfactory feature of the comparison by the relation $>$, for ideally all experiments would be comparable.

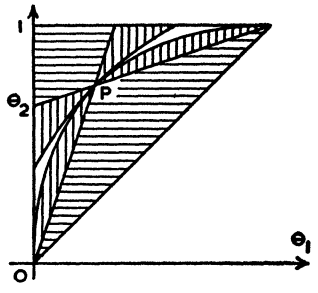

FIG. 1

The following considerations support the view that the relation $>$ holds in substantially more cases than does the relation $\supset$. Blackwell [2] remarks that the binomial trichotomy $\mathcal{E}_1 \equiv \mathcal{E}(0, \frac{1}{2}, 1)$, in an obvious extension of the previous notation, is not more informative than $\mathcal{E}_2 \equiv \mathcal{E}(0, \frac{1}{2}, \frac{1}{2})$. We shall show that $\mathcal{E}_1 > \mathcal{E}_2$.

Let $p(\theta) = \lambda p_1(\theta) + (1 - \lambda)p_2(\theta)$ and let $p_1(\theta) = (p, 0, q)$ and $p_2(\theta) = (p, q, 0)$. Then $p(\theta) = (p, q(1 - \lambda), q\lambda)$, a general prior distribution. From considerations of binomial dichotomies we have $g(\mathcal{E}_1, p_2(\theta)) = g(\mathcal{E}_2, p_2(\theta))$ and $g(\mathcal{E}_1, p_1(\theta)) \geqq g(\mathcal{E}_2, p_1(\theta))$. From Theorem 5

$$\begin{aligned} g(\mathcal{E}_1, p(\theta)) &\geqq \lambda g(\mathcal{E}_1, p_1(\theta)) + (1 - \lambda)g(\mathcal{E}_1, p_2(\theta)) \\ &\geqq \lambda g(\mathcal{E}_2, p_1(\theta)) + (1 - \lambda)g(\mathcal{E}_2, p_2(\theta)) \\ &= \lambda g(\mathcal{E}_2, p(\theta)) + (1 - \lambda)g(\mathcal{E}_2, p(\theta)) \\ &= g(\mathcal{E}_2, p(\theta)). \end{aligned}$$

Since $p(\theta)$ is arbitrary and the inequality is clearly strict for some $p(\theta)$, the result is established.

Another difference between the two methods of comparison is provided by our Theorem 4. We deduced a result which we can now express as

$$(\lambda\mathcal{E}^{(k)} + (1 - \lambda)\mathcal{E}^{(m)}) \leqq \mathcal{E}^{(n)},$$

where $n = \lambda k + (1 - \lambda)m$. W. H. Kruskal has pointed out that the same result is not necessarily true with $\leqq$ replaced by $\subset$. His example involves taking $\mathcal{E}_1$ to be a normal dichotomy, i.e., $\Theta = (\theta_1, \theta_2)$, $\mathbf{X}$ is the real line, and

$$p(x \mid \theta_i) = (2\pi)^{-1/2} \exp [-\tfrac{1}{2}(x - \theta_i)^2].$$

Let $d_i (i = 1, 2)$ be two decisions, with d_i correct when $\theta = \theta_i$. Let $p_{ij_n}(\delta)$ be the probability of saying d_i when $\theta = \theta_j$ on the evidence of the experiment $\mathcal{E}^{(n)}$, using the decision function δ. The relation $\supset$ can be expressed in terms of $p_{12_n}(\delta)$ and $p_{21_n}(\delta)$, and for some values of c, the function

$$\inf_\delta \{p_{12_n}(\delta) + cp_{21_n}(\delta)\}$$

is not concave. Thus it may, to quote an extreme case, produce a smaller loss to do no experimentation with probability $(1 - \lambda)$ and to perform $\mathcal{E}^{(k)}$ with probability λ than to do $\mathcal{E}^{(n)}$ with $n = \lambda k$.

We conclude this section by discussing another example of Blackwell's (see [2]) which demonstrates the techniques of the present theory. Each member of a large population of individuals has or has not each of two characteristics H, S. The proportions, h and s, of individuals with characteristics H, S are known. The proportion w of individuals having both characteristics is not known. Let $\mathcal{E}(H)$ denote the experiment in which a random individual from the population of individuals having characteristic H is observed; use $\mathcal{E}(\sim H)$, $\mathcal{E}(S)$, and $\mathcal{E}(\sim S)$ similarly, where $\sim H$ denotes the absence of the characteristic H. Suppose, with-

out loss of generality, that the characteristics are so named that $0 \leqq h \leqq s \leqq 1 - s \leqq 1 - h \leqq 1$. We proceed to show that $\mathcal{E}(H)$ is not less informative (S) than any of the other three experiments; that is, the best experiment is that in which individuals with the rarest characteristic are observed. Blackwell established the same result for not less informative (B) when w is known to be either hs or some specific alternative $\delta \neq hs$. Our result holds for any prior distribution of w.

Each of the four experiments is binomial with the following probabilities attached:

$$\mathcal{E}(H): \quad pr(S) = w/h = \theta,$$

$$\mathcal{E}(\sim H): pr(\sim S) = \frac{1 - h - s + w}{1 - h} = \frac{1 - h - s}{1 - h} + \frac{h}{1 - h}\theta,$$

$$\mathcal{E}(S): \quad pr(H) = w/s = h\theta/s,$$

$$\mathcal{E}(\sim S): pr(\sim H) = \frac{1 - h - s + w}{1 - s} = \frac{1 - h - s}{1 - s} + \frac{h}{1 - s}\theta,$$

where $\theta = w/h$. The permissible range for θ is $0 \leqq \theta \leqq 1$. Consider an arbitrary prior distribution for θ.

Now each of the four experiments is binomial with probability of the form $\lambda c + (1 - \lambda)\theta$, with $0 \leqq \lambda \leqq 1$, $0 \leqq c \leqq 1$. Alternatively, by introducing a random variable which is 1 or 0 according as the event, indicated above, does or does not occur, the probability density is

$$\begin{aligned} p(1 \mid \theta) &= \lambda c + (1 - \lambda)\theta \\ &= \lambda p_1(1 \mid \theta) + (1 - \lambda)p_2(1 \mid \theta), \end{aligned}$$

where $p_1(1 \mid \theta) = c$, $p_2(1 \mid \theta) = \theta$. Let $\mathcal{E}_1$, $\mathcal{E}_2$ be experiments with $P_1 = \{p_1\}$, $P_2 = \{p_2\}$. Then if $\mathcal{E}$ is any of the four experiments considered above, we have by Theorem 6

$$\mathcal{I}(\mathcal{E}) \leqq \lambda\mathcal{I}(\mathcal{E}_1) + (1 - \lambda)\mathcal{I}(\mathcal{E}_2) = (1 - \lambda)\mathcal{I}(\mathcal{E}_2) \leqq \mathcal{I}(\mathcal{E}_2),$$

since $\mathcal{I}(\mathcal{E}_1) = 0$ as p_1 does not depend on θ. But $\mathcal{E}(H)$ has $\lambda = 0$, so that $\mathcal{E}_2 \equiv \mathcal{E}_H$. This establishes the result, since the prior distribution is arbitrary.

6. Since Wald's introduction of decision theory, many statisticians, the present author included, have identified the theory with statistical theory and have argued that modern statistics *is* decision theory. Some statisticians, for example, Barnard [1] and Fisher [15], have not supported this view; they have contended, for example, that the purpose of a significance test is different from the purpose of a Wald decision problem with two decisions, reject or accept. It is therefore contended that different mathematical models are needed for the two purposes. This latter view is supported by the fact that significance levels do not occur in decision theory. If the purpose of modern statistics is not to come

to decisions, we may ask what is its purpose? Without wishing to take sides in the issue we propose in this section of the paper to investigate some elementary consequences of the attitude that the purpose of *some* statistical experimentation is to gain and measure information about the state of nature.

The first consequence of this attitude is that the statistician, faced with a choice of one among several experiments that he might perform, will choose that one for which the average amount of information is the greatest. The choice will, in general, depend on his prior knowledge, but it may happen that the experiments will be absolutely comparable by the methods of Section 5 and the prior knowledge will be irrelevant. Examples have already been given, but there is one further case worth considering. Let $\mathcal{E}(\sigma)$ denote the experiment in which $\mathbf{X}$ and Θ are the real lines and

$$p(x \mid \theta) = (\sqrt{2\pi}\,\sigma)^{-1}\exp\,[-(x - \theta)^2/2\sigma^2],$$

where $\sigma > 0$. Here x is normally distributed about θ with known variance σ^2. We shall show that $\mathcal{E}(\sigma_1) > \mathcal{E}(\sigma_2)$ if $\sigma_1 < \sigma_2$; that is, the experiment with smaller variance is the more informative (S). To prove the result, we show that $\mathcal{E}(\sigma_1) \succ \mathcal{E}(\sigma_2)$ and then apply Theorem 9, with the additional remark that there obviously exists a $p(\theta)$ such that $g(\mathcal{E}(\sigma_1),\ p(\theta)) > g(\mathcal{E}(\sigma_2),\ p(\theta))$. Let x_i be the random variable observed in $\mathcal{E}(\sigma_i)$. Then

$$x_2' = x_1 + u, \tag{19}$$

(where u is a random variable, independent of x_1, and having a normal distribution with zero mean and variance $\sigma_2^2 - \sigma_1^2$) has, for each θ, the same distribution as x_2. Equation (19) is thus a stochastic transformation from x_1 to x_2 and hence $\mathcal{E}(\sigma_1) \succ \mathcal{E}(\sigma_2)$.

A measure of the information can only be provided by assuming a particular form for $p(\theta)$. Suppose that

$$p(\theta) = (\sqrt{2\pi}\tau)^{-1}\,\exp\,[-(\theta - \mu)^2/2\tau^2] \equiv p_\tau$$

for some μ and $\tau > 0$. It is easy to establish that $p(x)$ is a normal density with mean μ and variance $\sigma^2 + \tau^2$. Also,

$$I_\theta p(\theta) = -\log\,(2\pi e)^{1/2}\tau.$$

Consequently, by equation (11), we have

$$g(\mathcal{E}(\sigma),\ p_\tau) = \tfrac{1}{2}\log\,(1 + \tau^2/\sigma^2).$$

This result provides an illustration of the truth of Theorem 4. If we use the notation of that theorem, with $\mathcal{E}_1 = \mathcal{E}(\sigma)$, we have that

$$j_n = \tfrac{1}{2}\log\,(1 + n\tau^2/\sigma^2),$$

which can be contrasted with the usual measure n/σ^2. Notice that j_n increases without limit in this situation.

Consider now the k-dimensional extension of these results. Let $\mathbf{X}$ and Θ be

k-dimensional Euclidean spaces and let $x = \{x_1, \cdots, x_k\}$ have a multivariate normal density with mean $\theta = \{\theta_i, \cdots, \theta_k\}$ and dispersion matrix C, which is known. Let θ have a prior density, p_A, which is multivariate normal with mean μ and dispersion matrix A. Denote this experiment by $\mathcal{E}(C)$; then, calculation along the same lines as in the univariate case gives

$$g(\mathcal{E}(C), p_A) = \tfrac{1}{2} \log \{|A + C| / |C|\}$$

where $|C|$ is the determinant of C. Clearly, even for this limited class of prior distributions, the two experiments $\mathcal{E}(C_1)$ and $\mathcal{E}(C_2)$ will not be absolutely comparable since their relative average informations depend critically on A. However, in some circumstances there is a possible simplification. Generally, we have that

$$g(\mathcal{E}(C_1), p_A) > g(\mathcal{E}(C_2), p_A)$$

if

$$|A + C_1|\,|C_2| > |A + C_2|\,|C_1|$$

or

$$|1 + A^{-1}C_1|\,|C_2| > |1 + A^{-1}C_2|\,|C_1|.$$

If the elements of $A^{-1}C_i$ are small in comparison with the unit matrix, this is approximately

$$|C_2| > |C_1|.$$

Hence, an approximate basis of comparison in this case, which corresponds to considerable ignorance about θ, is through the determinant of the dispersion matrix. The use of the determinant criterion has been used by Wald [11] in a slightly different context.

A second consequence of the view that one purpose of statistical experimentation is to gain information will be that the statistician will stop experimentation when he has enough information. Such a sequential method does not involve considerations of risks or cost of experimentation, but does involve a statement of prior knowledge. We consider next the sequential methods that this idea results in, for some special cases. In each case we shall consider a sequence $\mathcal{E}_1, \mathcal{E}_2, \cdots$ of independent, identical experiments which are to be performed until enough information about θ has been obtained. It is therefore a question of how much repetition of a given experiment should be performed.

We first take the dichotomy, with $\Theta = (\theta_1, \theta_2)$. $\mathbf{X}$, $\mathfrak{B}$, and P are quite general. Let δ be some preassigned number. Then experimentation will proceed; after n repetitions we shall have observations $(x_1, x_2, \cdots, x_n)$ and the amount of information will be

$$\sum_i p_n(\theta_i) \log p_n(\theta_i), \tag{20}$$

where

$$p_n(\theta_i) = p(\theta_i \mid x_1, \cdots, x_n),$$

the posterior distribution of θ. According to the idea introduced above, experimentation will continue until (20) is not less than δ. It is supposed that δ is chosen so that this sequential scheme will terminate with probability one; in this case δ must be negative. Since (20) is a convex function of $p_n(\theta_1) = 1 - p_n(\theta_2)$, the scheme corresponds to continuing sampling if, and only if,

$$1 - A < p_n(\theta_1) < A, \tag{21}$$

where

$$A \log A + (1 - A) \log (1 - A) = \delta.$$

Expression (21) may be written in terms of the ratio of posterior probabilities for θ_1 and θ_2, and by use of Bayes' theorem, it may be put in the form

$$\frac{1 - A}{A} \frac{p(\theta_2)}{p(\theta_1)} < \frac{p(x_1, \cdots, x_n \mid \theta_1)}{p(x_1, \cdots, x_n \mid \theta_2)} < \frac{A}{1 - A} \frac{p(\theta_2)}{p(\theta_1)}.$$

It is now apparent that the sampling scheme is equivalent to a scheme used in a Wald sequential probability ratio test of θ_1 against θ_2.

The generalization to the case where Θ has n elements will be sufficiently illustrated by the trichotomy $n = 3$. The argument is as with the dichotomy up to the sentence before that in which (21) appears. Now the posterior distribution $p_n(\theta_i)$ may be represented by a point in an equilateral triangle of unit altitude, the distances of the point from the sides being $p_n(\theta_i)(i = 1, 2, 3)$. Since (20) is again a convex function of the distribution, it follows that for sufficiently large values of δ, but $\delta < 0$, the regions of values of $p_n(\theta_i)$ for which sampling will cease will be three congruent convex regions at the three corners of the triangle. The calculation of the exact shapes of the regions would be a simple matter. It is interesting to note that regions of similar convex structure are obtained for termination in an optimum sequential scheme for deciding between three simple hypotheses with given loss function and prior distribution (see, for example, Blackwell and Girshick [4], p. 262).

We now leave the case where Θ is finite and suppose Θ to be an interval on the real line. It is now necessary to remark that as Shannon's measure of information is not invariant under a change of description of the parameter space, a different sequential scheme will be obtained if the description is changed. This unpleasant feature need not bother us unduly since sequential schemes based, for example, on the variance will have a similar feature. A sampling scheme in which sampling is continued until the variance of the estimator of θ is less than some prescribed number will differ from one designed for the variance of the estimator of $f(\theta)$. It is possible to find invariant sequential schemes by the device of sampling until the *average* amount of information to be gained by taking a further sample falls below a prescribed limit. It can then be argued that the further sample is not worth taking and sampling can therefore cease. We shall not investigate such schemes here; they will be invariant since the expression for the average amount of information is invariant.

First consider repetitions of the normal experiment $\mathcal{E}(\sigma)$, above, with prior distribution p_τ. After n observations with mean $\bar{x}$, which is a sufficient statistic, it is easy to verify that the posterior distribution of θ is normal with mean $(n\tau^2\bar{x} + \sigma^2\mu) / (n\tau^2 + \sigma^2)$ and variance $\sigma^2\tau^2 / (n\tau^2 + \sigma^2)$. The posterior information will therefore be $-\frac{1}{2} \log 2\pi e\sigma^2\tau^2 / (n\tau^2 + \sigma^2)$ and sampling will continue as long as this quantity is less than δ, or, equivalently, until

$$n \geqq \frac{2\pi e\sigma^2\tau^2 - \sigma^2 e^{-2\delta}}{\tau^2 e^{-2\delta}}.$$

Thus the optimum sequential scheme is of fixed sample size, given by the above expression. For large τ^2, corresponding to small prior knowledge, the fixed sample size is approximately $n = K\sigma^2$, where $K = 2\pi e^{2\delta+1}$. Thus the scheme is equivalent to sampling until the variance of the sample mean is sufficiently small.

As a final example, consider the case of repeated binomial trials. In the experiment to be considered $\mathbf{X} = (0, 1)$, Θ is the unit interval $0 \leqq \theta \leqq 1$, and $p(1 \mid \theta) = \theta$. The situation where the prior distribution is concentrated on a finite number of points is covered by the results above. We therefore consider densities over the whole interval of θ and, to simplify the calculations, confine attention to the family

$$p_{ab}(\theta) = \theta^{a-1}(1 - \theta)^{b-1}\Gamma(a + b) / \Gamma(a)\Gamma(b), \tag{22}$$

with a and b positive. This family of densities has the property that if the prior distribution is $p_{ab}(\theta)$, then the posterior distribution after a single binomial trial has been performed is $p_{a+1,b}(\theta)$ or $p_{a,b+1}(\theta)$, according as $x = 1$ or 0, respectively (a fact which the reader can easily verify). Simple calculation shows that

$$\begin{aligned} I_\theta p_{ab}(\theta) = \ln \Gamma(a + b) / \Gamma(a)\Gamma(b) + (a - 1)[\Psi(a) - \Psi(a + b)] \\ + (b - 1)[\Psi(b) - \Psi(a + b)], \end{aligned} \tag{23}$$

where

$$\Psi(x) = d \ln \Gamma(x) / dx.$$

This complicated expression can be simplified for large values of both a and b by use of the asymptotic formulas

$$\ln \Gamma(x) \sim \tfrac{1}{2}\ln 2\pi - x + (x - \tfrac{1}{2}) \ln x$$

and

$$\Psi(x) \sim \ln x - 1/2x.$$

We obtain

$$I_\theta p_{ab}(\theta) \sim \tfrac{1}{2} \ln (a + b)^3 / ab - \tfrac{1}{2}\ln 2\pi - \tfrac{1}{2}.$$

It follows that the curve in the plane of a and b along which $I_\theta p_{ab}(\theta)$ is constant is given approximately, for large values of a and b, by the curve

$$(a + b)^3 = \lambda ab$$

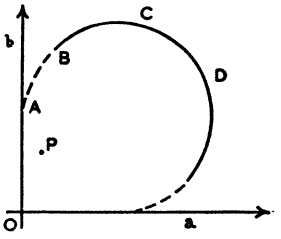

FIG. 2

for some constant λ. The general form of this curve is shown in Fig. 2 by a continuous line for $a, b > 10$; the broken extension shows the general form of the curve outside this range, as found by numerical computation.

Suppose the prior distribution has $a = a_0$, $b = b_0$. Then, after a sample of size n has produced r values of $x = 1$ and $n - r$ of $x = 0$, the posterior distribution will have $a = a_0 + r$, $b = b_0 + n - r$. The experimentation can be represented in the (a, b)-plane by starting at $P = (a_0, b_0)$ and forming a path by moving one unit along the a-axis for each value $x = 1$ and one unit along the b-axis for each value $x = 0$. Sampling will cease when the path intersects the curve corresponding to the amount of information required. If prior knowledge suggests that θ is small, then presumably one would take a_0 to be small and b_0 large, in comparison (for example, the point P in the figure). Ignorance about θ presumably corresponds to $a_0 = b_0 = 1$, or, at least a point with small a_0 and b_0.

We conclude by making a few comments on the boundary curve shown in Fig. 2, based on the assumption that $a_0 = b_0 = 1$. The most prominent feature is perhaps the sharp decrease in the critical value of b as a approaches one—the curve AB in the figure. Repetitions of one value of x, in this case $x = 0$, result in a greater accumulation of information than a mixture of both values. To cite a numerical instance: 6 occurrences of the value $x = 0$ are about as informative as 11 occurrences of $x = 0$ with one occurrence of $x = 1$, or 14 of $x = 0$ with two of $x = 1$. (The sample sizes are 6, 12, and 16, respectively.) This agrees with the "common-sense" feeling adduced by the consideration that if the same thing continually happens, say the sun rises each morning, then we are much better informed than we would be if there was known to be even a single non-occurrence. In the contrary case, when θ is about $\frac{1}{2}$, the part CD of the curve is relevant, and is approximated to by the fixed sample size scheme with boundary $a + b = \text{constant}$. The part BC of the curve can also be approximated to by the straight-line boundary $b = \text{constant}$. This would be appropriate if θ were about $\frac{1}{5}$ (but not too small so that the sharp curve AB was relevant) and would correspond to sampling until b values of $x = 0$ had been observed. If $x = 1$ corresponds to a "defective," this is the same as sampling, when defectives are rare, until the number of nondefectives has reached a preassigned number, and may be contrasted with inverse binomial sampling where the situation is similar but the rule is in terms of defectives.

7. Acknowledgments. I am much indebted to R. R. Bahadur for some valuable discussion, particularly in connection with Theorem 9, and to W. H. Kruskal for introducing me to the example in Section 5 and allowing me to include it in the present paper.

REFERENCES

[1] G. A. BARNARD, "Simplified decision functions," *Biometrika*, Vol. 41 (1954), pp. 241–251.

[2] DAVID BLACKWELL, "Comparison of experiments," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951, pp. 93–102.

[3] DAVID BLACKWELL, "Equivalent comparisons of experiments," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 265–272.

[4] DAVID BLACKWELL AND M. A. GIRSHICK, *Theory of Games and Statistical Decisions*, John Wiley and Sons, New York, 1954.

[5] G. H. HARDY, J. E. LITTLEWOOD, AND G. PÓLYA, *Inequalities*, Cambridge University Press, 1952.

[6] S. KULLBACK, "An application of information theory to multivariate analysis," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 88–102.

[7] S. KULLBACK, "Certain inequalities in information theory and the Cramér-Rao inequality," *Ann. Math. Stat.*, Vol. 25 (1954), pp. 745–751.

[8] S. KULLBACK AND R. A. LEIBLER, "Information and sufficiency," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 79–86.

[9] BROCKWAY MCMILLAN, "The basic theorems of information theory," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 196–219.

[10] C. E. SHANNON, "A mathematical theory of communication," *Bell System Technical Journal*, Vol. 27 (1948), pp. 379–423, 623–656.

[11] ABRAHAM WALD, "On the efficient design of statistical investigations," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 134–140.

[12] P. M. WOODWARD, "Theory of Radar Information" *Proceedings of a Symposium on Information Theory*, Ministry of Supply (London) (1950), pp. 108–113.

[13] P. M. WOODWARD, *Probability and Information Theory, with Applications to Radar*, Pergamon Press Ltd., London, 1953.

[14] LEE J. CRONBACH, "A consideration of information theory and utility theory as tools for psychometric problems" (Technical Report No. 1, Contract N6ori-07146, Urbana, Illinois, 1953).

[15] R. A. FISHER, "Statistical methods and scientific induction." *J. Roy. Stat. Soc.* (B), Vol. 17 (1955), pp. 69–78.