MORRIS H. DEGROOT

CHANGES IN UTILITY AS INFORMATION*

1. INTRODUCTION

The central topic of this paper is the measurement of the amount of information about some parameter θ that is present in a set of data or an observation $X = x$. The parameter θ can be any quantity such that a decision maker (DM) is uncertain about its value. We follow a Bayesian approach and assume that the DM can represent his uncertainty at any stage of the learning process in terms of a subjective probability distribution over the parameter space Ω of all possible values of θ. This distribution, in turn, will be represented by a generalized probability density function (gpdf) ξ with respect to some fixed σ-finite measure λ on Ω.

Furthermore, we assume that the observation X is a random variable or random vector taking values in some sample space X. Uncertainty about X is represented, as in the usual statistical models, in terms of a family of conditional distributions indexed by the parameter θ. Again, these conditional distributions are represented by their gpdf's $\{f(\bullet \mid \theta), \theta \in \Omega\}$ with respect to some fixed σ-finite measure ν on X.

We shall let $I(X = x)$ denote the (amount of) information about θ in the observation $X = x$. When it is understood that the particular random variable X was observed, we shall write simply $I(x)$. Before X has been observed, $I(X)$ is a random variable with a well-defined predictive distribution and mean $E[I(X)]$. Thus, $E[I(X)]$ is the expected information to be gained from observing X. In this paper, we shall review various methods that have been proposed for defining $I(x)$ and $E[I(X)]$.

Some authors have distinguished between measures of information that are based on both the DM's probability distribution for θ and his utility function, and measures that are based only on the DM's probability distribution [see, e.g., Good (1951, 1969), and Goel (1983)]. In this paper, we shall try to demonstrate that this distinction is not as sharp as it may at first

Theory and Decision 17 (1984) 287–303. 0040–5833/84/0173–0287$01.70

seem. In Section 2, the basic definition of expected information and its properties are introduced. In Section 3, the observed or actual information in an observation is discussed, a graphical interpretation is presented, and the additivity of expected information is established. In Section 4, the distribution of information is studied and this distribution is discussed in an example in which one experiment is sufficient for another. Finally, in Section 5 the concept of retrospective information is defined, its role in sequential experimentation is described, and the expectations of retrospective information at various stages of the sequential process are determined.

2. EXPECTED INFORMATION

We shall begin with a basic definition of the expected information $E[I(X)]$. Contrary to intuition, it seems to be more natural to develop this definition rather than the definition of the more basic concept $I(x)$ itself.

Consider a decision problem involving θ in which a DM must choose a decision d from some given set D. For each $\theta \in \Omega$ and $d \in D$, let $U(\theta, d)$ denote the utility of the DM if he chooses decision d when the value of the parameter is θ. For any gpdf $\xi(\theta)$, define

$$V(\xi) = \sup_{d \in D} \int_{\Omega} U(\theta, d)\, \xi(\theta)\, d\lambda(\theta) =$$

$$= \sup_{d \in D} E_{\theta}[U(\theta, d)]. \tag{2.1}$$

Throughout this paper we shall assume that all required integrals and expectations exist.

It follows from (2.1) that $V(\xi)$ can be regarded as the utility to the DM of his having the distribution ξ. Now let ξ_0 denote the prior gpdf of θ and let $\xi_0(\bullet \mid x)$ denote the posterior gpdf given $X = x$. Then we can determine $V[\xi_0(\bullet \mid x)]$ for each $x \in \mathrm{X}$. The expected information $E[I(X)]$ is defined as follows:

$$E[I(X)] = [E_x\{V[\xi_0(\bullet \mid X)]\}] - V(\xi_0). \tag{2.2}$$

In words, $E[I(X)]$ is the expected gain in utility from X. It can be shown that $E[I(X)] \geqslant 0$.

A basic property of the function V is that it is convex; i.e., for any two distributions ξ_1 and ξ_2 of θ and $0 < a < 1$, the distribution $a\xi_1 + (1-a)\xi_2$ of θ will have the property that

$$V[a\xi_1 + (1-a)\xi_2] \leqslant aV(\xi_1) + (1-a)V(\xi_2). \tag{2.3}$$

Some examples of such convex functions V are

$$V_1(\xi) = -\mathrm{Var}_\xi(\theta), \quad \text{for a real-valued parameter } \theta; \tag{2.4}$$

$$V_2(\xi) = \int_\Omega \xi(\theta) \log \xi(\theta)\, d\lambda(\theta); \tag{2.5}$$

$$V_3(\xi) = \max_i \xi(\theta_i), \quad \text{for } \Omega = \{\theta_1, \theta_2, \ldots\}. \tag{2.6}$$

Each of the functions V_1, V_2, and V_3 arises from a specific type of decision problem. To obtain V_1, suppose that a real-valued parameter θ must be estimated with squared-error loss. Since loss is simply negative utility, this assumption is that $U(\theta, d) = -(\theta - d)^2$. Then it is well-known that the Bayes decision with respect to any distribution ξ is $d = E_\xi(\theta)$. It follows from (2.1) that $V(\xi) = V_1(\xi)$, as defined by (2.4).

To obtain V_2, consider a decision problem in which the DM must specify a gpdf ϕ with respect to the measure λ, and $U(\theta, \phi) = \log \phi(\theta)$. Then the Bayes decision will be the gpdf ϕ such that the expectation

$$\xi(\theta) \log \phi(\theta)\, d\lambda(\theta) \tag{2.7}$$

is maximized. It is well-known that this maximization occurs for $\phi \equiv \xi$. Hence, $V(\xi) = V_2(\xi)$, as defined by (2.5). This particular example has been discussed by Bernardo (1979) and earlier by Good (1969).

To obtain V_3, suppose that the DM must choose one of the finite or countable possible values of θ, and that

$$U(\theta, d) = \begin{cases} 1 & \text{if } d = \theta \\ 0 & \text{if } d \neq \theta. \end{cases} \tag{2.8}$$

Then the Bayes decision is to choose a value of θ with the highest probability, i.e., a mode of ξ, and it follows that $V(\xi) = V_3(\xi)$, as defined by (2.6).

It follows from (2.2) and this discussion that expected information can be defined directly in terms of a convex utility function V on the space of

distributions of θ, and that it is not necessary to begin with the specification of a decision problem with a decision space D and utility function $U(\theta, d)$. On the other hand, every suitably regular convex V arises from some decision problem. However, we shall not pursue this topic further here. It is discussed for finite Ω in DeGroot (1962).

3. OBSERVED INFORMATION

In a sense, we have put the cart before the horse by defining the expected information in X before we have defined the observed or actual information in a realization $X = x$. One natural approach to defining the observed information $I(x)$ is simply to consider the observed change in utility.

$$I(x) = V[\xi_0(\bullet \mid x)] - V(\xi_0). \tag{3.1}$$

It follows immediately that the expected information $E[I(X)]$ will then be as we have defined it in (2.2). However, the information $I(x)$, as defined by (3.1), might well be negative. It is quite possible, and common to our experience, for the posterior distribution to leave the DM with more uncertainty and smaller expected utility than he had under his prior distribution.

How should the DM react to such an observation? Should he regret that he obtained it? Should he throw it away and act as though he had not seen it? Does it in fact contain negative information? The answer to these last three questions is "No". The information in any set of data is always nonnegative. An observation that spreads out the DM's posterior distribution is just as informative as one that makes his posterior distribution more concentrated. It tells him that his prior distribution may have been inappropriate or misleading in that it was probably concentrated around an incorrect value of θ. The correct definition of $I(x)$ proceeds as follows.

For any distribution ξ and any decision d in a given decision problem, let

$$U(\xi, d) = E_\xi[U(\theta, d)] = \int_\Omega U(\theta, d)\, \xi(\theta)\, d\lambda(\theta). \tag{3.2}$$

Also, let d_0 denote the Bayes decision with respect to the prior distribution ξ_0; i.e.,

$$U(\xi_0, d_0) = \sup_{d \in D} U(\xi_0, d) = V(\xi_0). \tag{3.3}$$

For simplicity, we assume the existence and uniqueness of d_0. Then

$$I(x) = V[\xi_0(\bullet \mid x)] - U[\xi_0(\bullet \mid x), d_0]. \tag{3.4}$$

In words, $I(x)$ is the expected difference, calculated with respect to the posterior distribution, between the utility from the Bayes decision and the utility from the decision d_0 that would have been chosen if the observation x had not been available. If the decision d_0 does not exist or is not unique then Eq. (3.4) must be modified by means of some convention, but we shall not consider such modifications in the paper.

It follows from the definition (3.4) that $I(x) \geqslant 0$ for every value of x. Furthermore, since

$$E\{U[\xi_0(\bullet \mid X), d]\} = U(\xi_0, d) \quad \text{for } d \in D. \tag{3.5}$$

where the expectation is taken with respect to the prior predictive distribution of X, it follows that $E[I(X)]$ will satisfy Eq. (2.2). Raiffa and Schlaifer (1961, Chapter 4), call $I(x)$, as defined by (3.4), the conditional value of sample information and call $E[I(X)]$ the expected value of sample information.

There is a helpful geometric interpretation of $I(x)$. Suppose that we want to define the information $I(\xi_0 \to \xi_1)$ in going from one distribution ξ_0 to another ξ_1, based on a given convex utility function $V(\xi)$. For simplicity, assume for the moment that Ω contains just a finite number of possible values of θ, so each distribution ξ on Ω can be regarded as a point in a finite-dimensional Euclidean space. Let $L(\xi \mid \xi_0)$ be the supporting hyperplane to the function $V(\xi)$ at the point $\xi = \xi_0$, which we assume to exist and be unique. Then

$$I(\xi_0 \to \xi_1) = V(\xi_1) - L(\xi_1 \mid \xi_0) \geqslant 0. \tag{3.6}$$

This expression is illustrated graphically in Figure 1 for a problem in which Ω contains just two points θ_1 and θ_2, so each distribution on Ω can be represented by the single number $\xi = \Pr(\theta = \theta_1)$.

Note that the definition of $I(\xi_0 \to \xi_1)$ is given in (3.6) directly in terms of the function V without any reference to an underlying decision problem or utility function U. In effect, the choice of a function V as a reference curve for measuring information, as illustrated in Figure 1, is tantamount to the choice of a utility function U in some decision problem. Thus, as

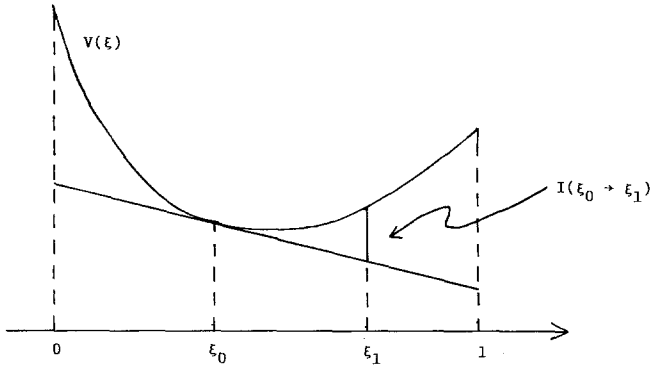

Fig. 1. The information $I(\xi_0 \to \xi_1)$.

stated in Section 1, the distinction made by some authors between measures of information that are based on both the DM's distribution ξ and his utility function U, and measures that are based only on ξ is not clear cut. Measures that are ostensibly based only on ξ require the choice of a function V or some other type of distance function that reflects how far, in terms of information gained, ξ_1 is from ξ_0.

Thus, the function V serves as the basis of a kind of "distance" measure in the space of distributions ξ. It is necessary to use quotation marks around the term "distance" because, in general, there are distributions ξ_0, ξ_1, and ξ_2 such that $I(\xi_0 \to \xi_1) \neq I(\xi_1 \to \xi_0)$ and such that $I(\xi_0 \to \xi_1) + I(\xi_1 \to \xi_2) < I(\xi_0 \to \xi_2)$. In other words, this function is not symmetric and does not satisfy the triangle inequality.

In general, we shall replace the definition of $I(x)$ given in (3.4) by the more general definition

$$I(\xi_0 \to \xi_1) = V(\xi_1) - U(\xi_1, d_0). \tag{3.7}$$

where d_0 is the Bayes decision with respect to the distribution ξ_0. We now reconsider the examples V_i $(i = 1, 2, 3)$ given by (2.4)–(2.6) and the corresponding utility functions.

PROPOSITION 1. If $V(\xi) = V_1(\xi)$, as given by (2.4), then

$$I_1(\xi_0 \to \xi_1) = (\mu_1 - \mu_0)^2. \tag{3.8}$$

where μ_i is the mean of the distribution ξ_i $(i = 0, 1)$.

Proof. In this problem,

$$I_1(\xi_0 \to \xi_1) = -\mathrm{Var}_{\xi_1}(\theta) - U(\xi_1, \mu_0)$$

and

$$U(\xi_1, \mu_0) = -E_{\xi_1}[(\theta - \mu_0)^2] = -[\mathrm{Var}_{\xi_1}(\theta) + (\mu_1 - \mu_0)^2].$$

The result (3.8) now follows. ∎

It should be noted that $I_1(\xi_0 \to \xi_1) = I_1(\xi_1 \to \xi_0)$.

PROPOSITION 2. If $V(\xi) = V_2(\xi)$, as given by (2.5), then

$$I_2(\xi_0 \to \xi_1) = \int_\Omega \xi_1(\theta) \log \frac{\xi_1(\theta)}{\xi_0(\theta)} \, d\lambda(\theta). \tag{3.9}$$

Proof. The result (3.9) follows immediately from (3.7) and the relations

$$V_2(\xi_1) = \int_\Omega \xi_1(\theta) \log \xi_1(\theta) \, d\lambda(\theta),$$

$$U(\xi_1, d_0) = \int_\Omega \xi_1(\theta) \log \xi_0(\theta) \, d\lambda(\theta). \quad ∎$$

The function I_2 is called the expected weight of evidence in favor of ξ_1 against ξ_0 (Good, 1950) or the Kullback–Leibler information for discriminating between ξ_1 and ξ_0 [see, e.g., Kullback (1968), Chap. 1, or Goel and DeGroot (1979)]. An important feature of I_2 is that it is invariant under any one-to-one differentiable transformation of the parameter θ. This property is not shared by the measure I_1.

PROPOSITION 3. If $V(\xi) = V_3(\xi)$, as given by (2.6), then

$$I_3(\xi_0 \to \xi_1) = \xi_1(\theta^1) - \xi_1(\theta^0), \tag{3.10}$$

where θ^i is the mode of the distribution ξ_i $(i = 0, 1)$.

Proof. The result (3.10) follows immediately from (3.7) and the relations

$$V_3(\xi_1) = \xi_1(\theta^1),$$

$$U(\xi_1, d_0) = \Pr(\theta = \theta^0 \mid \xi_1) = \xi_1(\theta^0). \quad ∎$$

For future use, we record here the value of $I_2(\xi_0 \to \xi_1)$ for normal distributions of θ.

PROPOSITION 4. Suppose that ξ_i is a normal distribution with mean μ_i and variance σ_i^2 $(i = 0, 1)$. Then

$$I_2(\xi_0 \to \xi_1) = \frac{1}{2}\left[\log\frac{\sigma_0^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_0^2} - 1 + \frac{(\mu_1 - \mu_0)^2}{\sigma_0^2}\right]. \quad (3.11)$$

Proof. In this example,

$$\log\frac{\xi_1(\theta)}{\xi_0(\theta)} = \log\frac{\sigma_0}{\sigma_1} + \frac{1}{2\sigma_0^2}(\theta - \mu_0)^2 - \frac{1}{2\sigma_1^2}(\theta - \mu_1)^2.$$

Since

$$E_{\xi_1}[(\theta - \mu_0)^2] = \sigma_1^2 + (\mu_1 - \mu_0)^2,$$

$$E_{\xi_1}[(\theta - \mu_1)^2] = \sigma_1^2.$$

and

$$I_2(\xi_0 \to \xi_1) = E_{\xi_1}\left[\log\frac{\xi_1(\theta)}{\xi_0(\theta)}\right],$$

the result (3.11) follows. ∎

We conclude this section with a basic result regarding the additivity of *expected* information.

THEOREM 1. Let ξ_0 denote the prior distribution of θ, let X and Y be any observations, and let $\xi_1(\bullet) = \xi_0(\bullet \mid X)$ and $\xi_2(\bullet) = \xi_0(\bullet \mid X, Y)$. Then

$$E[I(\xi_0 \to \xi_2)] = E[I(\xi_0 \to \xi_1)] + E[I(\xi_1 \to \xi_2)], \quad (3.12)$$

where each of the expectations is taken with respect to the prior predictive distribution of X and Y.

Proof. By (3.7),

$$I(\xi_1 \to \xi_2) = V[\xi_0(\bullet \mid X, Y)] - U[\xi_0(\bullet \mid X, Y), d_1(X)],$$

where $d_1(X)$ is the Bayes decision with respect to $\xi_0(\bullet \mid X)$. Furthermore,

$$E\{U[\xi_0(\bullet \mid X, Y), d_1(X)] \mid X\} = U[\xi_0(\bullet \mid X), d_1(X)] =$$
$$= V[\xi_0(\bullet \mid X)].$$

Hence,

$$E[I(\xi_1 \to \xi_2)] = E\{V[\xi_0(\bullet \mid X, Y)]\} - E\{V[\xi_0(\bullet \mid X)]\}. \tag{3.13}$$

Next, by (3.7),

$$I(\xi_0 \to \xi_1) = V[\xi_0(\bullet \mid X)] - U[\xi_0(\bullet \mid X), d_0]. \tag{3.14}$$

where d_0 is the Bayes decision with respect to ξ_0. Furthermore, it follows from (3.5), that

$$E\{U[\xi_0(\bullet \mid X), d_0]\} = V(\xi_0). \tag{3.15}$$

Hence, from (3.13)–(3.15), we obtain the relation

$$E[I(\xi_0 \to \xi_1)] + E[I(\xi_1 \to \xi_2)] = E\{V[\xi_0(\bullet \mid X, Y)]\} - V(\xi_0). \tag{3.16}$$

But also, by (3.7),

$$E[I(\xi_0 \to \xi_2)] = E\{V[\xi_0(\bullet \mid X, Y)]\} - E\{U[\xi_0(\bullet \mid X, Y), d_0]\}. \tag{3.17}$$

It now follows from (3.5) that the right-hand side of (3.17) is the same as that of (3.16). ∎

It has been stated in the literature [see, e.g., Lindley (1956)] that the relation (3.12) characterizes the information measure I_2 as given by (3.9). In fact, however, as Theorem 1 states, (3.12) holds for all information measures.

4. THE DISTRIBUTION

Essentially all previous work on the subject of statistical information has been restricted to the study of expected information. The reason for this is basically that the maximization of expected utility is equivalent to the maximization of expected information. However, there are at least two circumstances under which the DM is interested in the entire distribution of information: (i) He sometimes finds that he has obtained much more or much less information than he expected. (ii) He may have a choice among

different experiments to be performed at varying costs, where the overall utility of each experiment is not simply a linear function of the utility of the decision problem and the cost of experimentation, as is usually assumed in statistical decision theory. In case (i), the DM must study the distribution of information in order to evaluate how unusual his data are and decide whether his model is reasonable. In case (ii), he must study the distribution in order to choose an appropriate experiment.

We shall now present an example based on normal distributions in which the calculation of the distribution of $I_1(\xi_0 \to \xi_1)$ and of $I_2(\xi_0 \to \xi_1)$, as defined by (3.8) and (3.9), is essentially the same. Suppose that the prior distribution ξ_0 of θ is normal with mean μ_0 and precision τ_0, where the precision of a normal distribution is the reciprocal of its variance. Suppose also that an observation X is to be obtained such that the conditional distribution of X given θ is normal with mean θ and known precision r. In this example, X might be the sample mean of a random sample of n observations. Then it is well known [see, e.g., DeGroot (1970), Section 9.5] that the posterior distribution ξ_1 of θ given X is normal with mean

$$\mu_1 = \frac{\tau_0 \mu_0 + rX}{\tau_0 + r} \tag{4.1}$$

and precision

$$\tau_1 = \tau_0 + r. \tag{4.2}$$

It now follows from (3.8) that

$$I_1(\xi_0 \to \xi_1) = (\mu_1 - \mu_0)^2 \tag{4.3}$$

and from (3.11) that

$$I_2(\xi_0 \to \xi_1) = a_1 + a_2 I_1(\xi_0 \to \xi_1), \tag{4.4}$$

where a_1 and a_2 are constants depending on τ_0 and r, but not depending on either μ_0 or the observation X. Thus, studying the distribution of either I_1 or I_2 reduces to studying the distribution of $(\mu_1 - \mu_0)^2$.

It follows from the conditions of this example that the prior predictive distribution of X is normal with mean μ_0 and variance $(\tau_0 + r)/\tau_0 r$. Hence, from (4.1), the prior predictive distribution of $\mu_1 - \mu_0$ is found to be normal

with mean 0 and variance $r/[\tau_0(\tau_0 + r)]$. Finally, therefore, the distribution of the random variable

$$\frac{\tau_0(\tau_0 + r)}{r} (\mu_1 - \mu_0)^2 \tag{4.5}$$

is the χ^2 distribution with one degree of freedom.

In particular, as might have been anticipated the distribution of $I_1(\xi_0 \to \xi_1)$ and of $I_2(\xi_0 \to \xi_1)$ does not depend on μ_0. Also,

$$E[I_1(\xi_0 \to \xi_1)] = \frac{r}{\tau_0(\tau_0 + r)} . \tag{4.6}$$

An unusually small value or large value of I_1 reflects a value of X that was unusually close to or far from μ_0.

It is also instructive to study the conditional distribution of I_1 for a given value of θ. It follows from (4.1) that the conditional distribution of $\mu_1 - \mu_0$ given θ is normal with mean $r(\theta - \mu_0)/(\tau_0 + r)$ and variance $r/(\tau_0 + r)^2$. Thus, the conditional distribution of $I_1(\xi_0 \to \xi_1)$ given θ will be a suitably scaled noncentral χ^2 distribution, with

$$E[I_1(\xi_0 \to \xi_1) \mid \theta] = \frac{r}{(\tau_0 + r)^2} [1 + r(\theta - \mu_0)^2]. \tag{4.7}$$

Thus, the DM expects to obtain the most information when θ is far from μ_0.

We shall conclude this section with some comments regarding the relationship between sufficient experiments in the sense of Blackwell (1951, 1953) and the distribution of information. It is known that if some observation or experiment X is sufficient for another observation or experiment Y, then

$$E[I(X)] \geqslant E[I(Y)], \tag{4.8}$$

as defined by (2.2), for every convex function V and every prior distribution ξ_0 [see, e.g., De Groot (1962)].

It might be anticipated from this result that if X is sufficient for Y, then the random variable $I(X)$ must be stochastically larger than the random variable $I(Y)$ for every convex function V and prior distribution ξ_0. However, that conclusion is not correct. We now present a simple example in which $I(X)$ is not stochastically larger than $I(Y)$ for any strictly convex V or any ξ_0.

Suppose that Ω contains just two possible values θ_1 and θ_2. Suppose that the conditional distribution of X given $\theta = \theta_1$ is uniform on the interval $0 \leqslant x \leqslant 3$, and given $\theta = \theta_2$ it is uniform on the interval $1 \leqslant x \leqslant 4$. Then an observed value $X = x$ in the interval $1 \leqslant x \leqslant 3$ yields no information about θ, since the posterior distribution of θ will be the same as the prior distribution. On the other hand, an observed value $x < 1$ or $x > 3$ yields maximum information, because the posterior distribution will assign probability 1, to either $\theta = \theta_1$ or $\theta = \theta_2$. Thus,

$$\Pr\,[I(X) = 0] = 2/3 \tag{4.9}$$

for every function V and every prior distribution ξ_0.

Now define the random variable Y as follows:

$$Y = \begin{cases} 0 & \text{if } X \leqslant 2, \\ 1 & \text{if } X > 2, \end{cases} \tag{4.10}$$

Since Y is simply a function of X, it follows that X is sufficient for Y. However, for any prior distribution ξ_0 that does not assign probability 1 to either $\theta = \theta_1$ or $\theta = \theta_2$, the posterior distribution of θ will be different from ξ_0 for both of the possible observed values $Y = 0$ and $Y = 1$. Hence, if V is strictly convex, it can be seen from Figure 1 that $I(Y)$ will be positive for both $Y = 0$ and $Y = 1$. Thus,

$$\Pr\,[I(Y) > 0] = 1. \tag{4.11}$$

It follows from (4.10) and (4.11) that $I(X)$ is not stochastically larger than $I(Y)$.

5. RETROSPECTIVE INFORMATION

With just a single observation X or a single change in the distribution of θ from ξ_0 to ξ_1, it is not possible to determine whether an unusual value of $I(\xi_0 \to \xi_1)$ is due to an "inappropriate" prior distribution, i.e., an unlikely value of θ, or an "inappropriate" likelihood function, i.e., an incorrect model of the sampling process. In this section we shall extend the notion of information to new concepts that are relevant in problems of sequential analysis.

Consider a prior distribution ξ_0 of θ and a finite sequence of observations $X_1, X_2, \ldots, X_n$ leading successively to the sequence of posterior distributions

$\xi_1, \ldots, \xi_n$. Thus, ξ_j is the posterior distribution of θ given $X_1, \ldots, X_j$ $(j = 1, \ldots, n)$. For a given decision problem, let d_j denote the Bayes decision with respect to ξ_j $(j = 0, 1, \ldots, n)$. As before, we assume that d_j exists and is unique.

Now, for $i < j \leqslant k$, we define the information in changing from ξ_i to ξ_j, evaluated from the perspective of ξ_k, to be

$$I(\xi_i \to \xi_j \mid \xi_k) = U(\xi_k, d_j) - U(\xi_k, d_i). \tag{5.1}$$

We refer to the information defined by (5.1) at retrospective information because it represents information that we seem to have obtained at an earlier stage of the sequential process as evaluated with respect to a posterior distribution that we have reached at a later stage of the process. It is possible that $I(\xi_i \to \xi_j \mid \xi_k) < 0$. Roughly speaking, a negative value of this retrospective information will be obtained if, viewed from our current posterior distribution, the change from ξ_i to ξ_j moved us away from the values of θ that we now regard to be the most likely.

Retrospective information is an extension of the concept of information defined by (3.7) since

$$I(\xi_i \to \xi_j) = I(\xi_i \to \xi_j \mid \xi_j) \tag{5.2}$$

in our present notation. The following additivity property of retrospective information follows immediately from (5.1):

$$\sum_{i=0}^{n-1} I(\xi_i \to \xi_{i+1} \mid \xi_n) = I(\xi_0 \to \xi_n). \tag{5.3}$$

The sequential pattern of information that is obtained from a sequential sample can be analyzed in a variety of ways. For example, each of the following sequences can be studied:

$$I(\xi_i \to \xi_{i+1}), \qquad i = 0, 1, \ldots, n-1; \tag{5.4}$$

$$I(\xi_i \to \xi_{i+1} \mid \xi_n), \qquad i = 0, 1, \ldots, n-1; \tag{5.5}$$

$$I(\xi_0 \to \xi_1 \mid \xi_j), \qquad j = 1, 2, \ldots, n. \tag{5.6}$$

The analysis of these sequences and their relevance to decision making will be discussed in a future paper.

We shall now present the exact form of retrospective information for the functions V_i $(i = 1, 2, 3)$ given by (2.4)–(2.6).

PROPOSITION 5. If $V(\xi) = V_1(\xi)$, as given by (2.4), then

$$I_1(\xi_i \to \xi_j \mid \xi_k) = (\mu_k - \mu_i)^2 - (\mu_k - \mu_j)^2, \tag{5.7}$$

where μ_i, μ_j, and μ_k are the means of the distributions ξ_i, ξ_j, and ξ_k.

Proof. In this example, $d_r = \mu_r$ for $r = i, j, k$, and as shown in the proof of Proposition 1,

$$U(\xi_s, d_r) = -\mathrm{Var}_{\xi_s}(\theta) - (\mu_s - \mu_r)^2.$$

The result (5.7) now follows from (5.1). ∎

PROPOSITION 6. If $V(\xi) = V_2(\xi)$, as given by (2.5), then

$$I_2(\xi_i \to \xi_j \mid \xi_k) = \int_\Omega \xi_k(\theta) \log \frac{\xi_j(\theta)}{\xi_i(\theta)} \, d\lambda(\theta). \tag{5.8}$$

Proof. As shown in the proof of Proposition 2,

$$U(\xi_s, d_r) = \int_\Omega \xi_s(\theta) \log \xi_r(\theta) \, d\lambda(\theta).$$

The result (5.7) again follows from (5.1). ∎

PROPOSITION 7. If $V(\xi) = V_3(\xi)$, as given by (2.6), then

$$I_3(\xi_i \to \xi_j \mid \xi_k) = \xi_k(\theta^j) - \xi_k(\theta^i), \tag{5.9}$$

where θ^r is the mode of ξ_r $(r = i, j)$.

Proof. The proof again follows directly from the relations presented in the proof of Proposition 3. ∎

We conclude the paper with the calculation of three different types of expectations of retrospective information. For $r = 1, \ldots, n$, we shall use the notation E_r to denote a conditional expectation that is calculated given the observations $X_1, \ldots, X_r$.

THEOREM 2. For $i < j \leqslant k \leqslant n$,

$$E_k[I(\xi_i \to \xi_j \mid \xi_n)] = I(\xi_i \to \xi_j \mid \xi_k). \tag{5.10}$$

Proof. The expectation on the left-hand side of (5.10) is $E_k[U(\xi_n, d_j) - U(\xi_n, d_i)]$. At stage k, both d_i and d_j are fixed. Also,

$$E_k[U(\xi_n, d)] = U(\xi_k, d) \quad \text{for } d \in D. \tag{5.11}$$

It follows that the left-hand side of (5.10) is

$$U(\xi_k, d_j) - U(\xi_k, d_i),$$

which by definition is $I(\xi_i \to \xi_j \mid \xi_k)$. ∎

In words, Theorem 2 states that if we ask at a given stage k how we expect to evaluate, at some future stage n, a past change in information from ξ_i to ξ_j, the answer is that we expect to evaluate that change exactly as we presently evaluate it at stage k. The answer does not depend on n.

THEOREM 3. For $i \leqslant j \leqslant k < n$,

$$E_j[I(\xi_i \to \xi_k \mid \xi_n)] = E_j[V(\xi_k)] - U(\xi_j, d_i). \tag{5.12}$$

Proof. The expectation on the left-hand side of (5.12) is $E_j[U(\xi_n, d_k) - U(\xi_n, d_i)]$. At stage j, the decision d_i is fixed and

$$E_j[U(\xi_n, d_i)] = U(\xi_j, d_i). \tag{5.13}$$

Also,

$$E_j[U(\xi_n, d_k)] = E_j E_k[U(\xi_n, d_k)] = E_j[U(\xi_k, d_k)] = \\ = E_j[V(\xi_k)]. \tag{5.14}$$

Eqn. (5.12) now follows from (5.13) and (5.14). ∎

Theorem 3 has an interpretation in words that is not unlike that given for Theorem 2. Again it can be seen that the expectation in (5.12) does not depend on n.

THEOREM 4. For $i \leqslant j < k < n$,

$$E_i[I(\xi_j \to \xi_k \mid \xi_n)] = E_i[I(\xi_j \to \xi_k)]. \tag{5.15}$$

Proof. The expectation on the left-hand side of (5.15) is

$$E_i[U(\xi_n, d_k) - U(\xi_n, d_j)] = E_i E_k[U(\xi_n, d_k) - U(\xi_n, d_j)]$$
$$= E_i[V(\xi_k) - U(\xi_k, d_j)]$$
$$= E_i[I(\xi_j \to \xi_k)]. \quad \blacksquare$$

In words, Theorem 4 states that if we ask at the beginning of the process how we expect to view a future change from ξ_j to ξ_k from the perspective of the final stage n of the process, the answer is that we expect our final retrospective evaluation to be precisely the same as our prior expectation of that information. Again, the result does not depend on n.

Finally, there is a simple alternate expression for the result in Theorem 4.

COROLLARY 1. For $i \leqslant j < k < n$,

$$E_i[I(\xi_j \to \xi_k \mid \xi_n)] = E_i[V(\xi_k) - V(\xi_j)]. \tag{5.16}$$

Proof. It was shown in the proof of Theorem 4 that the expection on the left-hand side of (5.16) is equal to $E_i[V(\xi_k) - U(\xi_k, d_j)]$. But

$$E_i[U(\xi_k, d_j)] = E_i E_j[U(\xi_k, d_j)] = E_i[V(\xi_j)]. \quad \blacksquare$$

NOTE

Presented at the Second International Conference on Foundations of Utility and Risk Theory, Venice, June 5–9, 1984. This research was supported in part by the National Science Foundation under grants SES-8207295 and DMS-8320618. I am indebted to John Bacon-Shone of the University of Hong Kong and Richard Barlow of the University of California, Berkeley, with whom the concept of retrospective information introduced in this paper was jointly developed.

REFERENCES

Bernado, J. M.: 1979, 'Expected information as expected utility', *Ann. Statist.* 7, 686–690.

Blackwell, D.: 1951, 'Comparison of experiments', *Proc. Second Berkeley Symp. Math. Statist. Probability*, University of California Press, Berkeley 93–102.

Blackwell, D.: 1953, 'Equivalent comparison of experiments', *Ann. Math. Statist.* **24**, 265–272.

DeGroot, M. H.: 1962, 'Uncertainty, information and sequential experiments', *Ann. Math. Statist.* **33**, 404–419.

DeGroot, M. H.: 1970, *Optimal Statistical Decisions*, McGraw-Hill, New York.

Goel, P. K.: 1983, 'Information measures and Bayesian hierarchical models', *J. Amer. Statist. Assoc.* 78, 408–410.

Goel, P. K. and DeGroot, M. H.: 1979, 'Comparison of experiments and information measures', *Ann. Statist.* 7, 1066–1077.

Good, I. J.: 1950, *Probability and the Weighting of Evidence*, Charles Griffin, London.

Good, I. J.: 1951, 'Discussion of a paper by G. A. Barnard', *J. Royal Statist. Soc. Series B* **13**, 61–62.

Good, I. J.: 1969, 'What is the use of a distribution?', *Multivariate Analysis II* (ed. by P. R. Krishnaiah), Academic Press, New York pp. 183–203.

Kullback, S.: 1968, *Information Theory and Statistics*, Dover Publications, New York.

Lindley, D. V.: 1956, 'On a measure of the information provided by an experiment', *Ann. Math. Statist.* **27**, 986–1005.

Raiffa, H., and Schlaifer, R.: 1961, *Applied Statistical Decision Theory*, Division of Research, Graduate School of Business Administration, Harvard University, Boston.

Department of Statistics,
Carnegie-Mellon University,
Pittsburgh, PA 15213,
U.S.A.