The Annals of Statistics
1979, Vol. 7, No. 3, 686–690

EXPECTED INFORMATION AS EXPECTED UTILITY

BY JOSÉ M. BERNARDO
Universidad de Valencia

The normative procedure for the design of an experiment is to select a utility function, assess the probabilities, and to choose that design of maximum expected utility. One difficulty with this view is that a scientist typically does not have, nor can be normally expected to have, a clear idea of the utility of his results. An alternative is to design an experiment to maximize the expected information to be gained from it. In this paper we show that the latter view is a special case of the former with an appropriate choice of the decision space and a reasonable constraint on the utility function. In particular, the Shannon concept of information is seen to play a more important role in experimental design than was hitherto thought possible.

1. Introduction and notation. Maximization of the expected Shannon information was proposed by Lindley (1956) as a sensible, but ad hoc, criterion for the design of experiments "where the object of experimentation is not to reach decisions but rather to gain knowledge about the world". Stone (1959) and Fedorov (1972, Ch. 7) have investigated the design of regression experiments using such a criterion. By recognising the decision problem underlying a problem of statistical inference, we intend to show that such a procedure is, in fact, another instance of the general principle of maximizing expected utility. The argument lies entirely within the Bayesian framework.

Indeed, let E be an experiment which consists of the observation of a random variable X whose probability measure belongs to a family indexed by a parameter Θ, and let $\Psi = \Psi(\Theta)$ be some function of Θ in whose value we are interested. Moreover, let P_Θ be a probability measure which describes the personal opinions of the scientist about Θ before E is performed. After E has been performed and x obtained, the scientist's opinions about Θ should be described by the corresponding posterior density $p_\Theta(.|x)$. We have assumed, however, that the purpose of the research is to make inferences about the value of $\Psi = \Psi(\Theta)$ rather than about Θ itself; thus, the corresponding 'marginal' distribution $p_\Psi(.|x)$ describes the scientist's final opinions about the quantity of interest and constitutes, therefore, the final product of the investigation.

Consequently, we claim, *statistical inference about a quantity Ψ may be seen as a decision problem, in the presence of uncertainty about Ψ, in which the decision space is the class of (final) distributions of Ψ*. This is by no means a generally accepted view. For a nonBayesian approach see Blyth (1970).

Received August 1976; revised October 1977.

AMS 1970 *subject classifications.* Primary 62A15, 62B15; Secondary 62B10.

Key words and phrases. Bayesian statistics, decision theory, design of experiments, information, scientific inference, utility.

2. Expected useful information. The net result of performing E and obtaining x with regard to the parameter of interest is to modify the scientist's opinions from $p_\Psi(.)$ to $p_\Psi(.|x)$; the expected usefulness of the experiment could then be measured by some appropriate expected 'distance' between these densities. Along the lines of previous work by Lindley (1956), the expected information about $\Psi = \Psi(\Theta)$ to be provided by E when the initial density is $p_\Psi(.)$ could be defined as

$$(1) \qquad I^\Psi\{E, p_\Theta(.)\} = \int p_X(x) \int p_\Psi(\psi|x) \log \frac{p_\Psi(\psi|x)}{p_\Psi(\psi)} d\psi dx$$

provided the integral exists. Clearly, in information theoretical terms, Ψ is a 'garbling' of Θ and, as one would expect,

THEOREM 1. *$I^\Psi \leqslant I^\Theta$, with equality if Ψ is a one-to-one transformation of Θ.*

PROOF. Without loss of generality, one may assume $\theta = \{\psi, \omega\}$ where ω is some nuisance parameter and indeed

$$\begin{aligned} I^\Theta\{E, p_\Theta(.)\} - I^\Psi\{E, p_\Theta(.)\} &= \int p_X(x) \left\{ \int\int p_{\Psi\Omega}(\psi, \omega|x) \log \frac{p_{\Psi\Omega}(\psi, \omega|x)}{p_{\Psi\Omega}(\psi, \omega)} d\psi d\omega \right. \\ &\qquad \left. - \int p_\Psi(\psi|x) \log \frac{p_\Psi(\psi|x)}{p_\Psi(\psi)} d\psi \right\} dx \\ &= \int p_\Psi(\psi) \left\{ \int\int p_X(x|\psi) p_\Omega(\omega|\psi, x) \log \frac{p_\Omega(\omega|\psi, x)}{p_\Omega(\omega|\psi)} d\omega dx \right\} d\psi \geqslant 0 \end{aligned}$$

since the double integral in brackets is the (conditional to ψ) expected information about Ω which is known (Lindley, 1956, Thm. 1) to be nonnegative. Moreover, if Ψ is a one-to-one transform of Θ, then $I^\Psi = I^\Theta$; for, clearly, (1) is invariant under one-to-one transformations of the parameter space.

One way of describing the content of this paper is as a demonstration of how (1) arises naturally in inference.

3. Proper utility functions. We have argued that the task of performing inferences about ψ is a *decision problem* whose decision space is $D = \{p_\Psi(.); p_\Psi(\psi) \geqslant 0, \int p_\Psi(\psi) d\psi = 1\}$. To complete the specification of the problem, a utility function u measuring the desirability of each pair $\{p_\Psi(.), \psi\}$ must be defined.

Consider a scientist about to perform an experiment E in order to make inferences about Ψ. Furthermore, let u be the real function which describes the *utility* $u\{p_\Psi^\dagger(.), \psi\}$ obtained by the scientist if he *reports* the density function $p_\Psi^\dagger(.)$ as his final conclusions after E has been performed, when ψ is the true (unknown) value of the quantity of interest. (The superindex $\dagger$ in $p_\Psi^\dagger(.)$ identifies the *reported* density).

After the experiment has been performed and x obtained, the scientist's opinions about ψ, if coherently updated, will be described by the Bayes' posterior density

$p_\Psi(.|x)$. Thus, if he decides to report $p_\Psi^\ddagger(.)$ his expected utility will be

$$\int u\{p_\Psi^\ddagger(.), \psi\} p_\Psi(\psi|x) d\psi. \tag{2}$$

To maximize his expected utility, the scientist should report that density maximizing in D the integral (2), which will not necessarily be $p_\Psi(.|x)$, the one describing his final opinions. Hence to be *honest*, the scientist's (posterior) expected utility (2) must be maximized at, and only at, $p_\Psi(.|x)$. For, otherwise, his optimal policy could be to lie.

DEFINITION 2. A real function u, defined on $D \times \Psi$ is a proper utility function if, for each density $p_\Psi(.)$,

$$\sup_{p_\Psi^\ddagger(.) \in D} \int u\{p_\Psi^\ddagger(.), \psi) p_\Psi(\psi) d\psi = \int u\{p_\Psi(.), \psi\} p_\Psi(\psi) d\psi$$

and the supremum is only attained at $p_\psi(.)$.

Mostly in their discrete version, proper utility functions have been used, under the name of *proper scoring rules*, to elicit personal opinions (Winkler, 1969; Savage, 1971). Recent survey articles of the area are Hogarth (1975) and Spetzler and Stael von Holstein (1975). We prefer to use *utility* rather than *score* in order to underline that we are just following the general Bayesian principle of maximizing the expected utility.

Buehler (1971) and Good (1971) mention a number of examples of proper utility functions. We shall next provide an argument which suggests that the logarithmic proper utility function,

$$u\{p_\Psi^\ddagger(.), \psi\} = A \log p_\Psi^\ddagger(\psi) + B(\psi) \tag{3}$$

where A is an arbitrary constant and $B(.)$ an arbitrary function of ψ is often the more appropriate description for the preferences of a scientist facing an inference problem.

4. A characterization of the utility function. A property of the utility function (3) consists of the fact that the utility of reporting $p_\Psi^\ddagger(.)$ only depends on the probability density attached to the true value.

DEFINITION 3. The function u is a local utility function if $u\{p_\Psi^\ddagger(.), \psi\} = u\{p_\Psi^\ddagger(\psi), \psi\}$ for all values of ψ.

A referee pointed out to me that the requirement of locality could be viewed as a likelihood principle for utility functions in that it requires the utility of the probabilistic influence to depend only upon the probability density of the true state and not upon the density of the states which could have obtained but did not.

Locality reduces the first variable of the utility function to a real variable. A standard calculus of variations argument will be used to prove that, if u is *smooth* enough for such an argument to apply (for precise conditions see e.g., Jeffreys and Jeffreys, 1972, Ch. 10), the logarithmic is the only proper, local utility function.

To avoid difficulties owing to the fact that a density is only defined up to a set of measure zero, we shall assume that given P_Θ and an underlying measure μ with respect to which P_Θ is absolutely continuous, the unique version of the probability density at the point θ is defined as $p_\Theta(\theta) = \lim_{\rho\to 0} P_\Theta\{S_\rho(\theta)\}/\mu\{S_\rho(\theta)\}$ where $S_\rho(\theta)$ is a sphere of radius ρ centered on θ.

THEOREM 2. *If u is a smooth, proper, local utility function and $p_\Psi^\ddagger(.)$ is a density in L_2 then, for some constant A and function B,*

$$u\{p_\Psi^\ddagger(.), \psi\} = A \log p_\Psi^\ddagger(\psi) + B(\psi).$$

PROOF. Since u is local, $u\{p_\Psi^\ddagger(.), \psi\} = u\{p_\Psi^\ddagger(\psi), \psi\}$. Thus, to maximize $\int u\{p_\Psi^\ddagger(.), \psi\} p_\Psi(\psi) d\psi$ subject to the condition $\int p_\Psi^\ddagger(\psi) d\psi = 1$, one must obtain an extreme of

$$F\{p_\Psi^\ddagger(.)\} = \int u\{p_\Psi^\ddagger(\psi), \psi\} p_\Psi(\psi) d\psi - A\left[\int p_\Psi^\ddagger(\psi) d\psi - 1\right]. \tag{4}$$

Moreover, for $p_\Psi^\ddagger(.)$ to make (4) stationary, it is necessary that

$$\frac{\partial}{\partial\alpha} F\{p_\Psi^\ddagger(.) + \alpha\tau_\psi(.)\}|_{\alpha=0} = 0$$

for any function $\tau_\Psi(.)$ of sufficiently small norm. This reduces to the differential equation

$$\mathrm{D}_1 u\{p_\Psi^\ddagger(\psi), \psi\} p_\Psi(\psi) - A = 0$$

where $\mathrm{D}_1 u$ denotes the first partial derivative of u. But, by definition, if u is proper then the maximum of (4) must be attained precisely at $p_\Psi(.)$, so that a proper local utility function u must satisfy the partial differential equation

$$\mathrm{D}_1 u\{p_\Psi(\psi), \psi\} p_\Psi(\psi) - A = 0$$

whose solution is (3) as stated.

The discrete version of Theorem 2 was proved by Good (1952) for the binomial case, mentioned by McCarthy (1956) and proved by Aczel and Pfanzagl (1966) in a different context. It may be argued that the preferences of a scientist faced with a 'pure' inference problem should be described by a local utility function. For, in such a situation, one is only interested in the true value of ψ so that, when assessing the worthiness of a scientist's final conclusions, only the probability he attaches to a small interval containing the true value should be taken into account. For nonpathological posteriors, such probability mainly depends on the probability density attached to the true value of ψ.

We finally turn to the problem of selecting the best available experiment. The utility one may expect from a coherent choice of a decision $d \in D$ in the presence of uncertainty about ψ is $\sup_d \int u(d, \psi) p_\Psi(\psi) d\psi$. If an experiment E were performed and x obtained, one could similarly expect to obtain a utility $\sup_d \int u(d, \psi) p_\Psi(\psi|x) d\psi$. Thus, the increase in utility to be expected if one performs the experiment, the expected utility of the experiment, $u^*(E)$, is the expected value, over x, of the difference between these two expressions. Thus, using Theorem 2 we have that, *if preferences are described by a proper, local utility function then the expected utility of an experiment E intended to make inferences about Ψ is* $u^*(E) = gI^{\Psi}\{E, p_{\Theta}(.)\}$, where g is the expected utility of one unit of information about Ψ.

Acknowledgments. I would like to thank my supervisor, Professor Dennis V. Lindley for many helpful discussions, the editors and the referees for their comments and the British council for their financial support.

REFERENCES

ACZEL, J. and PFANZAGL, J. (1966). Remarks on the measurement of subjective probability and information. *Metrika* **11** 91–105.

BERNARDO, J. M. (1976). The use of information in the design and analysis of scientific experimentation. Ph.D. thesis, Univ. London.

BLYTH, C. R. (1970). On the inference and decision models in statistics. *Ann. Math. Statist.* **41** 1034–1058.

BUEHLER, R. J. (1971). Measuring information and uncertainty. In *Foundations of Statistical Inference.* (Godambe and Sprott, eds.) 330–341. Holt, Rinehart & Winston, Toronto.

FEDOROV, V. V. (1972). *Theory of Optimal Experiments.* Academic Press, New York.

GOOD, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. Ser. B* **14** 107–114.

GOOD, I. J. (1971). In discussion of R. J. Buehler (1971).

HOGARTH, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *J. Amer. Statist. Assoc.* **70** 271–294.

JEFFREYS, H. and JEFFREYS, B. (1972). *Methods of Mathematical Physics.* University Press, Cambridge.

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

MCCARTHY, J. (1956). Measurements of the value of information. *Proc. Nat. Acad. Sci. U.S.A.* **42** 654–655.

SAVAGE, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783–801.

SPETZLER, C. S. and STAEL VON HOLSTEIN, C. A. S. (1975). Probability encoding in decision analysis. *Management Sci.* **22** 340–358.

STONE, M. (1958). Studies with a measure of information. Ph.D. thesis. Univ. Cambridge.

STONE, M. (1959). Application of a measure of information to the design and comparison of regression experiments. *Ann. Math. Statist.* **30** 55–70.

WINKLER, R. L. (1969). Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.* **64** 1073–1078.

DEPARTAMENTO DE BIOESTADISTICA
FACULTAD DE MEDICINA
AVENIDA BLASCO IBANEZ 17
VALENCIA-10, SPAIN