Departament d'Estadística i I.O., Universitat de València.
Facultat de Matemàtiques, 46100–Burjassot, València, Spain.
Tel. 34.6.363.6048, Fax 34.6.363.6048 (direct), 34.6.386.4735 (office)
Internet: bernardo@uv.es, Web: http://www.uv.es/˜bernardo/
Printed on March 27, 1996

Statistical Inference as a Decision Problem: The Choice of the Sample Size

JOSÉ M. BERNARDO

Universitat de València, Spain

Professor Lindley is to be congratulated for providing another example of the power and versatility of a decision-oriented, fully Bayesian approach. As he stresses, maximisation of the expected utility provides a general method to determine the optimal sample size, whatever the specific preferences of the decision-maker might be; in particular, he points out that inference may be seen as a particular decision problem where the action space consists of the class of possible distributions of the quantity of interest, and the utility function is a logarithmic score. I will extend his comments on this very important particular case, in the hope of making easier to the reader its practical use.

As described in the paper, if one faces a decision problem with alternatives $d \in D$ whose consequences depend on some (unknown) relevant quantity $\theta \in \Theta \subseteq \Re$ with prior distribution $p_\theta(.)$, if one is prepared to make an experiment in order to learn more about θ which consists of observing data $\{x_1, x_2, \ldots\}$, $x_i \in X$, related to θ by $p_x(.\,|\,\theta)$, and if the utility of performing an experiment of size n, obtain $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, and choose d if θ is true, is given by

$$u(n, \boldsymbol{x}, d, \theta) = g(d, \theta) - c(n, \boldsymbol{x}),$$

where the *gain* function g and the *cost* function c are expressed in the same (say monetary) units, then coherence implies that the optimal sample size is that value of n which maximizes

$$u^*(n) = \int_{X^n} p_x(\boldsymbol{x}) \sup_{d \in D} \int_\Theta g(d, \theta) p_\theta(\theta\,|\,\boldsymbol{x})\, d\theta\, d\boldsymbol{x} - c^*(n)$$

where

$$p_\theta(\theta\,|\,\boldsymbol{x}) \propto \prod_{i=1}^n p_x(x_i\,|\,\theta) p_\theta(\theta),$$

$$p_{\boldsymbol{x}}(\boldsymbol{x}) = \int_\Theta \prod_{i=1}^n p_x(x_i\,|\,\theta) p_\theta(\theta)\, d\theta,$$

$$c^*(n) = \int_{X^n} c(n, \boldsymbol{x}) p_{\boldsymbol{x}}(\boldsymbol{x})\, d\boldsymbol{x}.$$

José M. Bernardo is Professor of Statistics at the University of Valencia. Research partially funded with grant PB93-1204 of the DGICYT, Madrid, Spain. Invited discussion to be published in *The Statistician*.

Moreover, scientific inference may appropriately be described as the particular decision problem where $D = \{q_\theta(.)\}$ is the class of possible distributions of θ which could conceivably be *reported* as the final conclusion of the statistical analysis, and the gain function is a proper, local scoring rule, so that, necessarily, (Bernardo, 1979a)

$$g(q_\theta(.), \theta) = A \log q_\theta(\theta) + B(\theta).$$

In a setting designed to analyse the value of an experiment, it is natural to set to zero the gain from reporting the prior distribution. This obviously implies $B(\theta) = -A \log p_\theta(\theta)$ and hence, the expected gain from an experimental result $\boldsymbol{x}$ is

$$\int_\Theta g(d, \theta) p_\theta(\theta \mid \boldsymbol{x})\, d\theta = I^\theta\{\boldsymbol{x}, p_\theta(.)\} = A \int_\Theta p_\theta(\theta \mid \boldsymbol{x}) \log \frac{p_\theta(\theta \mid \boldsymbol{x})}{p_\theta(\theta)}\, d\theta.$$

The function $I^\theta\{\boldsymbol{x}, p_\theta(.)\}$, which is obviously invariant under one-to-one transformations of θ and depends both on the data $\boldsymbol{x}$ and on the prior $p_\theta(.)$, is the appropriate expression for the *amount of information* about θ provided by $\boldsymbol{x}$, *not* the non-invariant minus entropy of equation (9) in the paper. If, furthermore, one chooses $A = v/\log(2)$, where v is the *expected value of one bit of information* about θ, that is, the expected value for the decision maker of the answer to one question about θ with *two* alternative answers with the *same prior probability*, then the expected utility of a sample of size n is given by

$$u^*(n) = vI(n) - c^*(n)$$

where

$$I(n) = \int_{X^n} p_{\boldsymbol{x}}(\boldsymbol{x}) \int_\Theta p_\theta(\theta \mid \boldsymbol{x}) \log_2 \frac{p_\theta(\theta \mid \boldsymbol{x})}{p_\theta(\theta)}\, d\theta\, d\boldsymbol{x}$$

is Shannon's expected information about θ from a sample of size n.

Extending previous results (Bernardo, 1979b), I have found (work in progress) that, under broad regularity conditions,

$$I(n) = \frac{1}{2} \log_2 (1 + \frac{n}{n_0}) + o(1), \quad (n \to \infty),$$

where n_0, which may aptly be termed the *sample size equivalent* of the prior distribution $p_\theta(.)$ with respect to an experiment which consists of a random sample from $p_x(.\mid\theta)$, is given by

$$n_0 = -\frac{1}{i(\theta)} \frac{\partial^2}{\partial\theta^2} \log \frac{p_\theta(\theta)}{\sqrt{i(\theta)}}\bigg|_{\theta=\theta_0}$$

where θ_0 is the solution to the equation

$$\frac{\partial}{\partial\theta} \log \frac{p_\theta(\theta)}{\sqrt{i(\theta)}} = 0,$$

and $i(\theta)$ is Fisher's information function,

$$i(\theta) = \int_X p_x(x \mid \theta) \left\{ -\frac{\partial^2}{\partial\theta^2} \log p_x(x \mid \theta) \right\} dx.$$

It may be verified that, provided that $p_\theta(.)$ is reasonably well behaved, the expression above gives very good approximations to the expected information even for rather moderate values of n.

Very often, the cost structure of an experiment is such that its expected value $c^*(n)$ is approximately linear on n. In this case, the expected utility of a sample of size n may be written as

$$u^*(n) \approx \frac{v}{2} \log_2(1 + \frac{n}{n_0}) - c_0 - cn,$$

where v is the expected value of one bit of information about θ, and c is the expected cost of each observation. It follows that the optimal sample size must be such that

$$n_0 + n = \frac{1}{2 \log 2} \frac{v}{c}\,.$$

In words, the *total* sample, i.e., the sample size equivalent of the prior information plus the experimental sample size, must equal $1/(2 \log 2) \approx 0.72$ of the ratio of the value of one bit of information to the cost of one observation.

I will conclude by outlining a simple example. Suppose that a firm is about to order a survey to learn on the proportion θ of people which would like a particular product, and that their beliefs are such that (i) their most likely value for θ is 0.15 and (ii) they have probability 0.99 that θ is smaller than 0.39. It may be verified that this roughly corresponds to a prior Beta distribution $\mathrm{Be}(\theta \,|\, 4.5, 20.5)$.

Using the definition given above, the sample size equivalent n_0 of a $\mathrm{Be}(\theta \,|\, \alpha, \beta)$ prior distribution with respect to a binomial experiment with parameter θ may found to be $\alpha + \beta - 1$. Incidentally, this is consistent with the fact the appropriate reference distribution in this case is $\pi_\theta(\theta) = \mathrm{Be}(\theta \,|\, 0.5, 0.5)$, with sample size equivalent 0, *not* the uniform distribution. Thus, in our example, the sample size equivalent of the prior knowledge is $4.5 + 20.5 - 1 = 24$.

If, say, the firm is prepared to pay $5,000 for the answer to one bit of information about θ (in order to know, for example, whether θ is inside or outside [0.11, 0.21], which is the 50% HPD interval of the prior), and if the expected cost of each interview is $10, then the firm should require a *total* sample of

$$\frac{1}{2 \log 2} \frac{5000}{10} \approx 360.$$

Since the sample size equivalent of their prior knowledge is 24, they should pay for a survey of size $360 - 24 = 336$.

REFERENCES

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.

Bernardo, J. M. (1979b). Comportamiento asintótico de la información proporcionada por un experimento. *Rev. Real Acad. Ciencias Madrid* **73**, 491–502.