The Annals of Statistics
1981, Vol. 9, No. 3, 670–672

A BAYESIAN CRITERION FOR SAMPLE SIZE

BY MICHAEL GOLDSTEIN

University of Hull

A criterion is suggested for the size of a sample from a distribution of unknown form, and an upper bound for this quantity is obtained, involving only certain simple aspects of the prior measure.

An i.i.d. sample $\mathbf{X}_m = (X_1, \cdots, X_m)$ is to be drawn from a distribution F of unknown form, where the sample size m must be determined before sampling. We wish to estimate $\theta(F)$, the mean of F. The prior measure for F is $P(F)$. The purposes of this note are, firstly, to suggest an intuitive criterion for the sample size, and secondly, to derive an upper bound for this quantity, involving only certain simple aspects of P.

The motivation for the criterion is as follows: Suppose we are able to take observations sequentially. Having observed a sample $\mathbf{X}_n = \mathbf{x}_n$, we might stop sampling if we believed it to be unlikely that any further observations would cause us to substantially revise our current estimate. Thus, a possible stopping criterion might be to stop sampling having observed $\mathbf{x}_n$, provided that, for specified positive constants a, b,

$$P\{\sup_{m>n} | E_m(\theta) - E_n(\theta) | > a \,|\, \mathbf{X}_n = \mathbf{x}_n\} < b, \tag{1}$$

where $E_r(\theta) = E(\theta \,|\, X_1, \cdots, X_r)$. By varying the choice of a and b, we may produce relatively weak or stringent criteria. Using criterion (1), we are demanding that we should stop when we believe that it is unlikely to make much difference to our final estimate if we continue to sample. However, we may produce a more stringent criterion if we demand not only that (1) should be satisfied, but also that with high probability any future observations we might make would also satisfy criterion (1). Thus, denote by A_n the set of $\mathbf{x}_n$ for which relation (1) is satisfied. The modified stopping criterion is that having observed $\mathbf{x}_n$, we should stop if

$$P\{\mathbf{X}_m \in A_m \quad \text{for all} \quad m \geqslant n \,|\, \mathbf{X}_n = \mathbf{x}_n\} \geqslant 1 - c, \tag{2}$$

for some specified value of c. Again, varying the choice of c determines the stringency of the criterion. Thus, criterion (2) demands that we should stop sampling when we believe that we are unlikely to obtain any future samples for which further sampling might cause us to revise our estimate.

However, we must assign a fixed sample size n_0, so our suggested criterion is to choose n_0 as the smallest value satisfying

$$P\{\mathbf{X}_m \in A_m \qquad \text{for all} \quad m \geqslant n_0\} \geqslant 1 - c. \tag{3}$$

Note that this criterion does not involve the relation between $E_n(\theta)$ and the unobservable quantity θ, but instead considers the relation between $E_n(\theta)$ and the observable quantities $E_m(\theta)$, $m > n$. These are the natural quantities of interest if we dispense with the quantities F and $\theta(F)$ (which are mainly introduced for notational simplicity), and instead consider the X values as members of an infinite exchangeable sequence, so that $E_m(\theta)$ is the prediction of X_{m+1} having observed $\mathbf{X}_m$, and $P(\cdot)$ is the mixing distribution derived as in de Finetti's representation theorem for exchangeable sequences.

Received August, 1979; revised February, 1980.

AMS 1970 *subject classifications.* Primary 62F15; secondary 62G05.

Key words and phrases. Bayesian estimation, posterior mean, sample size, stopping rule.

We will derive a simple upper bound for n_0. Thus, we may ascertain a range of values of a, b and c which may in practice be achieved in the specific problem. If these values are small, then as the criterion is rather stringent, n_0 will be a practical upper bound for general stopping criteria. Conversely, for any proposed sample size, we may use the upper bound to give an intuitive insight into the behaviour of the estimate with respect to the criterion we have outlined above, having specified only a small number of aspects of the prior measure.

Specifically, we shall require the two prior quantities $V(\theta)$, $V(x)$ defined by

$$V(\theta) = E(\theta(F) - E(\theta))^2 = \int \left\{\theta(F) - \int \theta(F)\, dP(F)\right\}^2 dP(F)$$

$$V(x) = E(X - \theta(F))^2 = \int \int (x - \theta(F))^2\, dF(x)\, dP(F).$$

$V(\theta)$ is the variance of our prior measure for θ and $V(x)$ is the prior expected variance of X both of which we assume to be finite. These quantities are sufficient to derive an upper bound for n_0, which involves the ratio of $V(x)$ to a^2. The bound is high for stringent choices of b, c, such as 0.05, but is useful for more moderate choices, of order about 0.2.

THEOREM. *For specified positive values a, b and c, and n_0 as defined by* (3),

$$n_0 \leq V(x)\left(\frac{1}{a^2bc} - \frac{1}{V(\theta)}\right).$$

(*If the right-hand side is negative, then $n_0 = 0$.*)

PROOF. $E_n(\theta)$ is a martingale. Thus given that $\mathbf{X}_n = \mathbf{x}_n$, $(E_{m+n}(\theta) - E_n(\theta))^2$ is a submartingale in m. (A reference for the martingale results used in the proof is Karlin and Taylor (1975) Chapter 6.). From the maximal inequality for submartingales, we have

$$\begin{aligned} &P\{\max_{m \leq M} (E_{m+n}(\theta) - E_n(\theta))^2 > a^2 \mid \mathbf{X}_n = \mathbf{x}_n\} \\ (4) \qquad &\leq \frac{1}{a^2} E((E_{M+n}(\theta) - E_n(\theta))^2 \mid \mathbf{X}_n = \mathbf{x}_n) \\ &\leq \frac{1}{a^2} V(\theta \mid \mathbf{X}_n = \mathbf{x}_n), \end{aligned}$$

the second step in inequality (4) following because

$$V(\theta \mid \mathbf{X}_n) = E(V(\theta \mid \mathbf{X}_{M+n}) \mid \mathbf{X}_n) + V(E_{M+n}(\theta) \mid \mathbf{X}_n).$$

From inequality (4), a sufficient condition to ensure that $\mathbf{x}_n \in A_n$ is $V(\theta \mid \mathbf{X}_n = \mathbf{x}_n) \leq a^2 b$. Thus,

$$(5) \qquad P(\mathbf{X}_m \in A_m \quad \text{for all} \quad m \geq n) \geqq P(\max_{m \geq n} V_m(\theta) \leq a^2 b),$$

where $V_m(\theta) = V(\theta \mid \mathbf{X}_m) = E((\theta - E(\theta \mid \mathbf{X}_m))^2 \mid \mathbf{X}_m)$.

But $V_m(\theta)$ is a supermartingale, as $(E_m(\theta))^2$ is a submartingale. Thus, by the maximal inequality for nonnegative supermartingales,

$$(6) \qquad P(\max_{m \geq n} V_m(\theta) > a^2 b) \leq \frac{1}{a^2 b} E V_n(\theta).$$

But, from Finucan (1971), we have

$$(7) \qquad E V_n(\theta) \leq \frac{V(\theta) V(x)}{n V(\theta) + V(x)}.$$

Thus, from (5), (6) and (7), we have that n will be an upper bound for n_0, as defined by (3), provided that

$$1 - c \leqslant 1 - \frac{1}{a^2 b} \frac{V(\theta) V(x)}{nV(\theta) + V(x)},$$

and the result follows.

REFERENCES

FINUCAN, H. M. (1971). Posterior Precision for non-normal distributions. *J. Roy. Statist. Soc. B* **33** 95–97.

KARLIN, S. and TAYLOR, H. M. (1975). *A First Course in Stochastic Processes.* Second Ed. Academic, New York.

DEPARTMENT OF MATHEMATICAL STATISTICS
THE UNIVERSITY OF HULL
HULL, ENGLAND