TRABAJOS DE ESTADISTICA Y
DE INVESTIGACION OPERATIVA
Vol. 36. Núm. 1, 1985, pp. 24 a 30

ON A FAMOUS PROBLEM OF INDUCTION

J. M. Bernardo
Departamento de Bioestadística
Facultad de Medicina
Valencia

RESUMEN

Se ofrece una nueva solución bayesiana al problema de contrastar la hipótesis de que todos los miembros de una población tienen una determinada característica, cuando se ha observado que la tienen todos los elementos de un subconjunto suyo escogido al azar.

Palabras clave: Inferencia bayesiana, poblaciones finitas, inducción, distribuciones de refrencia.

Clasificación AMS (1980)
Primaria: 62A15; secundaria: 62B10

SUMMARY

A Bayesian solution is provided to the problem of testing whether an entire finite population shows a certain characteristic, given that all the elements of a random sample are observed to have it. This is obtained as a direct application of existing theory and, it is argued, improves upon Jeffreys's solution.

Key words: Bayesian inference, finite populations, induction, reference distributions.

1. Introducción

Suppose that n individuals have been observed, randomly chosen from a finite population of size N, and suppose moreover that all the

elements of the sample possess a certain property. The problem of testing a law which states that all the N individuals in the population have such a property has a very long history and its solution finds applications in fields ranging from philosophy of science to quality control.

To the best of our knowledge, no convincing solution is available from a frequentist point of view. The problem was studied by Wrinch and Jeffreys (1921/23) an Jeffreys (1939/61, pp. 128-132) from a Bayesian point of view, as Geisser (1980) recently reminded us. Jeffreys's solution is intuitively reasonable, but it is obtained through an *ad hoc* argument. In this note we revise Jeffreys's proposal and we provide a normative Bayesian solution which has the same intuitive behaviour as Jeffreys', but is obtained through a systematic use of existing theory.

2. The problem and Jeffreys' solution

Let R be the unknown number of individuals in the population which possess the property under study and let T be the number of individuals which have such property in a random sample of size n. Clearly,

$$\Pr(T = t \mid R, n, N) = \frac{\binom{R}{t}\binom{N-R}{n-t}}{\binom{N}{n}}, \quad t = 0, 1, \min(R, n)$$

$$= 0, \quad \text{elsewhere} \tag{1}$$

The, rather natural, Bayesian solution to the problem posed obviously consists in computing the value of $p = \Pr(R = N \mid T = n, n, N)$ i.e. the (posterior) probability of the whole population having the property given that all the elements in the sample have it.

To have a posterior one needs a prior. The use of the «standard» uniform prior

$$\Pr(R = r \mid N) = (N + 1)^{-1}, \quad r = 0, 1, ..., N \tag{2}$$

produces $p = (n + 1)/(N + 1)$ which is obviously small as N grows. As Jeffreys points out, this is clearly untenable. Indeed, if N is large

compared to n, which is the most often encountered case, p will be small and yet, common sense dictates that if a long series of trials were all of one kind, a feeling that this phenomenon would persist should be induced. Jeffreys (1961, p. 129) recognizes that «an adequate theory of scientific investigation must leave it open for any hypothesis whatever that can be clearly stated to be accepted on a moderate amount of evidence», and thus that «*any clearly stated law has a positive prior probability and therefore an appreciable posterior probability until there is definite evidence agains it*». This leads him to the choice of a prior of the form

$$\begin{aligned} &\Pr(R = 0) = \Pr(R = N) = k \\ &\Pr(R = r) = (1 - 2k)/(N - 1)\,,\ r = 1, 2, ..., N - 1 \end{aligned} \tag{3}$$

Note that, in his own terms, this implies that the law to be tested is «either none or all the elements in the population have the property studied».

This is not approproate in our case. Indeed, the problem posed, to induce a property from n observations to the whole population, logically excludes $R = 0$. (The dual problem, to induce the absence of a property in a population from its absence in a sample of size n is, of course, mathematically identical, and excludes logically $R = N$.)

Jeffreys goes on to choose what he considers to be a reasonable value for k. As he recognizes (p. 130) «the best value to take for k is not clear». Ressorting to a clearly *ad hoc* argument he proposes $k = (1/4) + 1/(2N + 2)$ which yields, using Jeffreys (1961, p. 130, eq. 19) to

$$p = \frac{(N + 3)(n + 1)}{(N + 3)(n + 1) + 2(N - n)} = \frac{(N + 3)(n + 1)}{(N + 1)(n + 3)} \tag{4}$$

Thus, as N grows, p tends to $(n + 1)(n + 3)$ a behaviour which «seems satisfactory» to him. Note however that for $n = 0$ this gives $(N + 3)/\{3(N + 1)\}$ so that (4) does not reduce to its prior value

$$p(R = N \mid N) = k = (N + 3)/\{4(N + 1)\} \tag{5}$$

as one would expect.

Technically, this apparent contradiction is due to the fact that the values $R = 0, 1, ..., n - 1$ become logically impossible after observing $T = n$, and, to us, suggests that the prior structure (3) is not reasonable.

If one is prepared to recognize that the question of interest is whether or not $R = N$, then Jeffreys' argument may be restated by considering priors of the form

$$\begin{aligned} &\Pr(R = N) = k \\ &\Pr(R = r) = (1 - k)/N \ , \ r = 0, 1, ..., N - 1 \end{aligned} \tag{6}$$

It seems likely that Jeffreys' choice of k would have then been $k = 1/2$. In the next section we show that this is indeed the prior suggested by the general method proposed the author (Bernardo, 1979) which do not require of any *ad hoc* assumptions.

3. The alternative solution

The basic idea underlying the construction of what we call a *reference* posterior distribution, i.e. a posterior which only use the information provided by the model and the data, may be described as follows. Considerar the amount of information, in Shannon's (1948) and Lindley's (1956) sense, written $I^0\{\varepsilon(k), p(\theta)\}$, that may be expected about the quantity of interest θ from k independent replications of an experiment ε when the prior distribution of θ is $p(\theta)$, and let C be the class of adminissible priors, i.e. those compatible with whatever agreed «objective» information one is willing to assume. By performing infinite replications of ε one would get to know precisely the value of θ. Thus, $I^0\{\varepsilon(\infty), p(\theta)\}$ measures the amount of *missing* information about θ when the prior is $p(\theta)$. It seems natural to define «vague initial knowledge» about θ as that described by the density $\pi(\theta)$ which maximizes the missing information in the class C. The reference posterior distribution of θ after the result x of the experiment ε has been observed, denoted $\pi(\theta \mid x)$, is then obtained *via* Bayes' theorem.

In the presence of nuisance parameters ω, the reference prior $\pi_0(\theta, \omega)$ for the parameter of interest θ is defined as $\pi_0(\theta, \omega) = \pi(\theta)\pi(\omega \mid \theta)$ that is the reference prior $\pi(\omega \mid \theta)$ for the nuisance parameter ω in the conditional model given θ, times the reference prior $\pi(\theta)$ for the para-

meter θ in the reduced model obtained integrating out ω with $\pi(\omega|\theta)$. It must be stressed that *the method proposed gives different reference distributions for different parameters of interest within the same model.*

Moreover, the following may be shown (Bernardo, 1979): (i) In the continuous case, and under regularity conditions which guaranteee asymtotic normality, the reference prior turns out to be Jeffreys's «invariant rule». (ii) In the discrete case the reference prior is that maximizing prior entropy, a procedure strongly recommended by Jaynes (1968); in the absence of other conditions this is achieved by the uniform distribution. Finally, it should be mentioned that the whole procedure may be viewed as a minimax choice when (quasi) utilities are measured by weight of evidence (Good, 1968, 1969).

In the problem we have been discussing the parameter of interest is obviously:

$$\begin{aligned} \theta &= \theta_0 \quad \text{if } R = N \\ &= \theta_1 \quad \text{otherwise} \end{aligned}$$

and *not* whether R equals N or 0; the nuisance parameter is the concrete value of R if different of N. Thus, according to the *general* methodoly just outlined, one would have

$$\begin{aligned} &\pi(\theta_0) = \pi(\theta_1) = 1/2 \\ &\pi(\omega \mid \theta) = 1/N \ , \ \omega = 0, 1, ..., N - 1 \end{aligned}$$

and, therefore, the reference prior when the parameter of interest is θ turns out to be:

$$\begin{aligned} \Pr(R = r \mid N) &= 1/2 \quad , \ r = N \\ &= 1/(2N) \ , \ r = 0, 1, ..., N - 1 \end{aligned}$$

The posterior probability of $R = N$ given $T = n$ is then, by Bayes theorem

$$p = \frac{\Pr(T = n \mid R = N, n, N) \Pr(R = N)}{\sum_{r=n}^{N} \Pr(T = n \mid R = r, n, N) \Pr(R = r)} =$$

$$= \frac{1/2}{\left(\frac{1}{2N}\right) \sum_{r=n}^{N-1} \left\{\binom{r}{n} / \binom{N}{n}\right\} + \frac{1}{2}} = \frac{N(n+1)}{N(n+1) + (N-n)}, \tag{8}$$

using the identity

$$\sum_{r=n}^{N-1} \left\{\binom{r}{n} / \binom{N}{n}\right\} = \frac{N-n}{n+1}. \tag{9}$$

As one would expect, and in contrast with Jeffreys's solution, if $n = 0$ then $p = 1/2$, wahtever the value of N. Moreover, $p = 1$ if $n = N$ and, as N grows, p tends to $(n + 1)/(n + 2)$ giving, as Jeffreys wanted, a high probability for the law been true when n is moderate, even if n is small compared with N.

4. Discussion

The problem considered has been the center of a number of discussions in the philosophy of science and has important applications in taxonomy, sociology and quality control. However, despite of the fact of having a rather controversial solution, it has received relatively little attention from modern statisticians. With this note we try to draw attention to this important, if mathematically simple, example, where different approaches to statistical inference may be tested and compared, thereby providing an interesting pedagogical exercise.

REFERENCES

BERNARDO, J. M. (1979): «Reference posterior distributions for Bayesian inference», *J. Roy. Statist. Soc. B 41*, 113-147 (with discussion).

GEISSER, S. (1980): «The contributions of Sir Harod Jeffreys to Bayesian Inference», in *Bayesian Analysis in Econometrics and Statistics* (A. Zellner, ed.), 13-20, Amsterdam: North-Holland.

GOOD, I. J. (1968): «The utility of a distribution», *Nature, 219*, 1392.

GOOD, I. J. (1969): «What is the use of a distribution?», in *Multivariate Analysis II* (P. R. Krishnaiah, ed.), 183-203, New York: Academic Press.

JAYNES, E. T. (1968): «Prior probabilities», *IEEE Trans. Systems, Science and Cybernetics SCC-4*, 227-291.

JEFFREYS, H. (1939/61): *Theory of Probability,* Oxford: University Press.

LINDLEY, D. V. (1956): «On a measure of the information provided by an experiment», *Amer. Math. Statist., 27*, 986-1005.

SHANNON, C. E. (1948): «A mathematical theory of communications», *Bell System Tech. J., 27*, 379-423, 623-656.

WRINCH, D. M., and JEFFREYS, H. (1921/23): «On certain fundamental principles of scientific inquiry», *Philos. Magazine, 42*, 369-390; *45*, 368-374.